2008

The DM gene family in the parasitoid wasp nasonia vitripennis: identification of a sex-specific homolog of the doublesex gene

Megan Riddle
Western Washington University

THE DM GENE FAMILY IN THE PARASITOID WASP NASONIA VITRIPENNIS: IDENTIFICATION OF A SEX-SPECIFIC HOMOLOG OF THE DOUBLESEX GENE

By

Megan Riddle

Accepted in Partial Completion

of the Requirements of the Degree

Master of Science

Moheb A. Ghali, Dean of the Graduate School

ADVISORY COMMITTEE

Chair, Dr. Carol Trent

Dr. David Leaf

Dr. Sandra Schulze

MASTER'S THESIS

Signature ______________________________
Date __________________________________

THE DM GENE FAMILY IN THE PARASITOID WASP NASONIA VITRIPENNIS: IDENTIFICATION OF A SEX-SPECIFIC HOMOLOG OF THE DOUBLESEX GENE

A Thesis
Presented to
The Faculty of
Western Washington University

In Partial Fulfillment
Of the Requirements for the Degree
Master of Science

by
Megan Riddle

May 2008

ABSTRACT

Sexual dimorphism is the result of a cascade of genes that triggers sex-specific development. The cascade begins with a primary signal that affects a hierarchy of genes and, through their selective activation and repression, results in the development of an individual of a particular sex. The genes in this regulatory hierarchy are very divergent, with little conservation across taxa. However, homologs of doublesex, a master regulator in the sex-determining cascade of *Drosophila melanogaster*, have been found in organisms ranging from nematodes to humans. These homologs are identified by the DM domain, a DNA-binding motif found in genes that function as transcription factors. The DM domain defines an entire family of genes, of which only a select few play a role in sex determination. Here I describe a family of four DM-containing genes in the parasitoid wasp *Nasonia vitripennis*. Using molecular and computational techniques, I completed the sequence of two previously discovered members of this family and identified two new genes that contain the DM domain. One of these new genes, NvDM4, shows sex-specific expression reminiscent of the doublesex gene, suggesting that it is part of the sex-determination cascade in *N. vitripennis*.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

INTRODUCTION

Among the plans laid out for both mice and men in the early embryo is the decision to become male or female. This is not a matter of a single gene, one marked *Boy*, the other *Girl*, but instead a series of inter-connected regulators, each one influencing the next. These genes form a regulatory hierarchy, passing the message of sexual differentiation down the chain until, at the bottom of this cascade, the verdict of which sex to become is passed on to a suite of effectors that specify the differentiation of sexually dimorphic cells and tissues. While the overall plan laid out in the embryo, with a hierarchical cascade of regulators, is seen in a variety of organisms, the specifics of the cascade itself, what genes are used and when and how, varies. However, the basic structure of one gene, *doublesex* discovered in the fruit fly, has been found to be conserved across taxa. Homologs of this gene contain a DNA-binding motif called the DM domain and can be found in the sex determination cascades of organisms ranging from mice to men to *Musca*. My thesis research used computational and molecular techniques to examine a family of DM domain genes in the wasp *Nasonia vitripennis*. This family has four members and I have shown that one of these is a doublesex homolog and likely to be involved in the sex-determination of *N. vitripennis*.

Gene families

A gene family is a group of genes that, because of the high degree of sequence similarity, are thought to have evolved from a single ancestral gene. The original ancestral gene may have duplicated through unequal crossing over during homologous recombination, unequal exchange between sister chromatids, or strand slippage during DNA replication (fig.1) (Brown 2002). Gene families are very common in eukaryotes such as *Drosophila melanogaster*, *Caenorhabditis elegans*, and *Arabidopsis thaliana* (table 1).

With duplicate or multiple copies of a single gene, mutations in a single version of the gene can escape selective pressure that would have eliminated such variation had the gene not been duplicated. This enables different forms of a gene to arise. After duplication, a variety of possible fates exist for the now-redundant genes. The most common is for one gene to maintain its original function under purifying selection while the other gene, not under such constraints, accumulates mutations (Prince and Pickett 2002). This usually results in the loss of function of one gene, forming a pseudogene (fig.2-A). Alternatively, the duplicated genes can acquire a new function in a process referred to as neo-functionalization (fig.2-B). Another possibility is proposed by the duplication-degeneration-complementation (DDC) model in which both genes undergo mutations and there is a loss of function, called sub-functionalization, in each (fig.2-C). Instead of losing function, though, under the DDC model the genes work together to carry out the function of the original gene, thus complementing each other (Prince and Pickett 2002). In this scenario, functional copies of both genes would remain in the genome as both would be under positive selection. Mutations can occur in the regulatory region of the duplicated gene, resulting in differential expression of the genes (Louis 2007). These alterations include changes at cis-acting sites and can result in the sub- or neofunctionalization of the gene as the timing and level of expression is altered.

A classic example of gene duplication is the globin family which arose from a single ancestral gene that duplicated about 1.1 billion years ago (Voet and Voet 2004). Globin-like genes can be found in archaea, bacteria, plants, fungi, and animals, with five distinct types occurring in vertebrates alone (Roesner *et al.* 2005). The primordial gene, which likely served as a simple monomeric protein that stored oxygen, has since evolved into a variety of forms that occur as both a monomer (myoglobin) and a tetramer (hemoglobin). These forms appear in different tissues and at different stages in development. For example, in the first eight weeks after conception, the tetrameric hemoglobin of human embryos is made of two ζ and two ε subunits (Voet and Voet 2004). As the embryo develops, a new form of hemoglobin called fetal hemoglobin, made of two α and two γ subunits, is produced. Both of these tetramers have a higher affinity for oxygen than adult hemoglobin, which consists of two α and two β subunits, enabling the embryonic and fetal hemoglobin to pull oxygen off the mother's hemoglobin, thus providing oxygen for the growing fetus (Voet and Voet 2004). All of these variations have arisen due to evolution acting on duplicated genes.

Table 1. The proportion of genes that are members of gene families in *D. melanogaster*, *C. elegans*, and *A. thaliana* (adapted from Lewin 2004).

Organism	Unique genes	Families with >1 member
Drosophila melanogaster (fruit fly)	72%	28%
Caenorhabditis elegans (nematode)	55%	46%
Arabidopsis thaliana (plant)	35%	65%

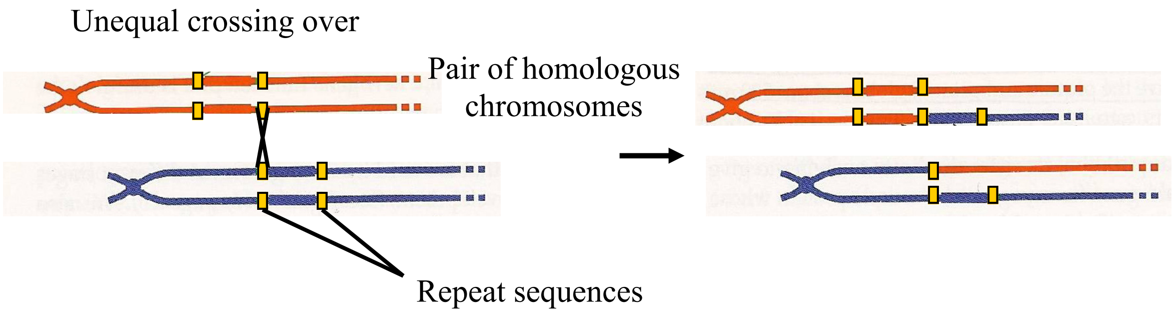


Figure 1. Unequal crossing over of homologous chromosomes. During homologous recombination, if unequal amounts of chromosome are exchanged, this can result in the gene duplication (adapted from Brown 2002).

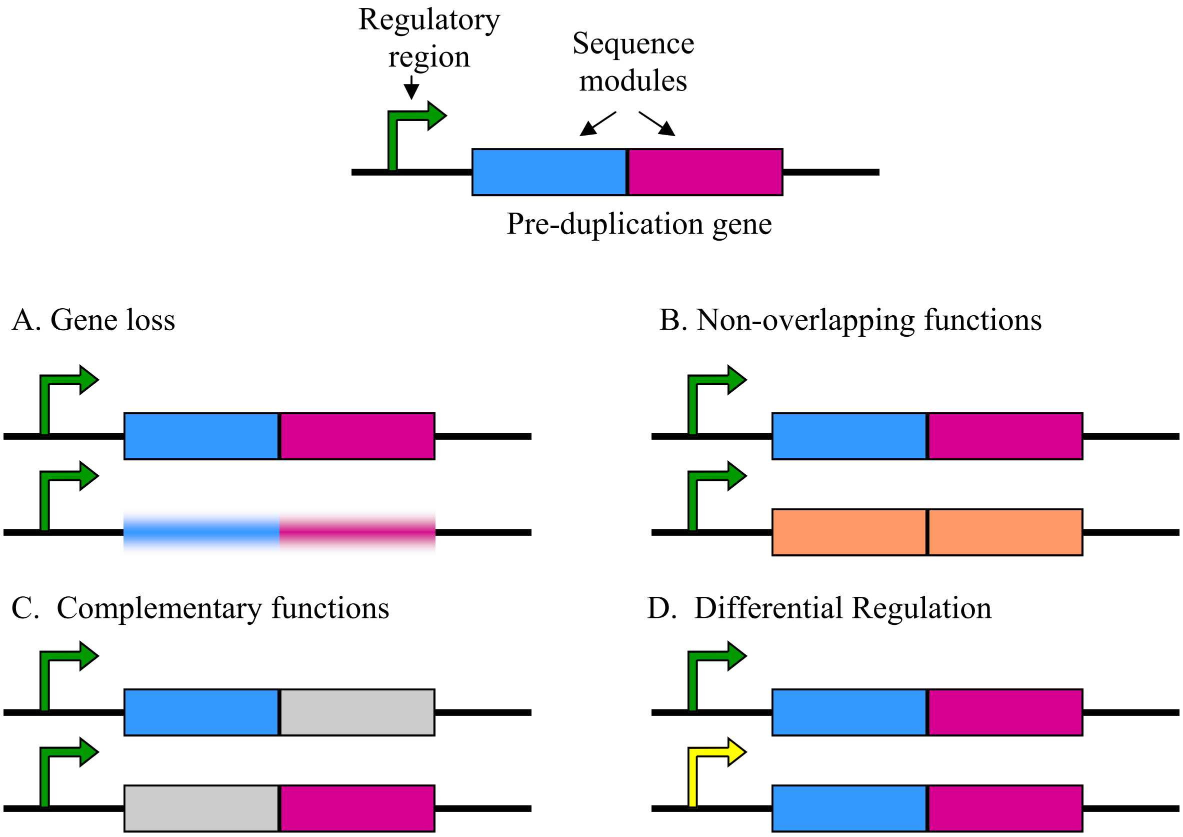

Regulatory region
Sequence modules
Pre-duplication gene
A. Gene loss
B. Non-overlapping functions
C. Complementary functions
D. Differential Regulation

Figure 2. Potential fates of duplicated genes. The duplicated gene may (A) accumulate mutations and degrade, resulting in a pseudogene, (B) evolve non-overlapping functions in which the duplicated gene plays a new role in the organism, (C) develop complementary functions such that both genes are required to fulfill the role of the original gene, referred to as subfunctionalization, (D) or incur mutations in the regulatory region, resulting in differential expression and either sub- or neofunctionalization (adapted from Louis 2007).

Alternative splicing

Another way in which new gene function can arise without interfering with original gene function is through alternative splicing. In order to understand alternative splicing, it is necessary to first understand the process of splicing itself. Eukaryotic genes consist of both coding regions, called exons, and non-coding regions, called introns. When transcribed, this complete sequence of exons and introns is called pre-mRNA. In order for correct translation to occur, the introns must be removed and the exons connected together, creating the mRNA through a process called RNA splicing (Lewin 2004). It is the processed mRNA that is then translated into protein (fig. 4).

The boundary between an intron and exon in the pre-mRNA is called a splice site. The splice site at the 5' end of an intron is referred to as the donor site, while the site at the 3' end is called the acceptor site (fig. 4). Introns also contain a branch-site adenine located 18 to 40 nucleotides upstream of the acceptor site (Lewin 2004). There is a short region of conserved sequence at both ends of the intron, although only the GU at the donor site and AG at the acceptor site are present in virtually all introns.

The process of removing the intron occurs via two transesterfication reactions in which the 2' hydroxyl group of the branch site adenine attacks the phosphodiester bond that occurs at the donor site (fig. 5-A) and then the 3' hydroxyl group of the exon at the donor site attacks the phosphodiester bond at the acceptor site (fig. 5-B). The end product is two exons connected by a phosphodiester bond and a separated intron in the lariat structure (fig. 5-C).

Splicing of transcripts from protein coding genes is catalyzed by a complex of proteins that brings together the donor, acceptor, and branch sites so that the reactions can occur (fig. 6). This protein machinery, called the spliceosome, assembles sequentially onto the pre-mRNA and consists of both proteins and small nuclear RNAs, referred to as snRNP for small nuclear ribonucleoprotein particles (Lewin 2004). The correct assemblage of the spliceosome requires the presence of certain consensus sequences in the pre-mRNA. Variations in these splice site sequences can result in differences of the relative strength of the splice sites, with some being stronger or weaker than others. The strength of a splice site refers to how well the pre-mRNA sequence binds to the spliceosomal machinery (Irimia *et al*. 2007). For example, there are 9 semi-conserved nucleotides at the 5' splice site that base pair with the 5' terminus of the U1 SnRNA (figs. 6 and 9), one of the protein-RNA complexes that make up the spliceosome, and this base-pairing is critical for the splice site detection (Roca *et al*. 2005). Splice site strength plays a direct role in alternative splicing, with introns containing weaker splice sites more likely to undergo alternative splicing and these weaker sites may require exon splice enhancers for splicing to occur (fig. 7) (Zavolan *et al*. 2003).

In the most basic form of splicing, both donor and acceptor sites are recognized independently of any sequence outside the intron. This is called intron definition (Lewin 2004). However, when introns are long or splice sites weak, sequences downstream of the intron are required for correct spliceosome assembly, a mechanism called exon definition (fig. 8). These downstream sites that aid in the binding of the splicing machinery are referred to as splicing enhancers (Lewin 2004).

While in some cases the same donor and acceptor sites are used in the processing of every pre-mRNA from a particular gene, others undergo a process called alternative splicing in which the donor and acceptor sites utilized in the removal of introns may vary.

Alternative splicing can take a variety of forms (fig. 10). These can include different 5' or 3' ends, exon skipping, and mutually exclusive exons.

The effect of alternative splicing of mRNA on resulting protein structure varies. In some cases, there may be no change to the protein if the alternative splicing occurs in the 5' or 3' untranslated region (UTR) of the gene, as can occur when splicing involves alternative initiation or polyadenylation (fig. 10). While this does not affect the resulting protein, variations in the UTR can influence transcript regulation by altering the stability or translatability of the mRNA. For example, alternative splicing of the 3' UTR of a gene called maskin, a protein involved in early development of *Xenopus laevis*, results in the repression of maskin in stage four embryos and its activation in stage six, due to different 3' UTRs interacting with different proteins (Meijer *et al*. 2007). However, studies in humans have indicated that these variations in UTR are in the minority, with 70% - 88% of alternatively spliced transcripts resulting in a change in the protein (Modrek and Lee 2002). These variations can alter the function of the protein by changing the domains for which an mRNA codes. A domain is sequence of amino acids within a protein that forms a self-stabilizing structure that often folds independently from the rest of the protein. These domains, also called protein motifs, are often highly conserved due to pleiotropy, by which one sequence interacts with a number of others, making co-evolution difficult because a change in the conserved sequence will affect all the sequences with which it interacts. When alternative splicing affects the domains of a protein, it can alter the role of the protein within the organism. For example, one member of the FC receptor family, a group of proteins that sit on the outside of some immune cells and are involved in antibody binding, undergoes alternative polyadenylation (fig. 11). This results in the replacement of the transmembrane domain and cytoplasmic tail, involved in signal transduction, with another transmembrane domain and tail, thereby changing the specificity of the protein (Modrek and Lee 2002). It is this modularity, with domains that can often fold and function independently of one another, that enables the great variation in proteins through alternative splicing.

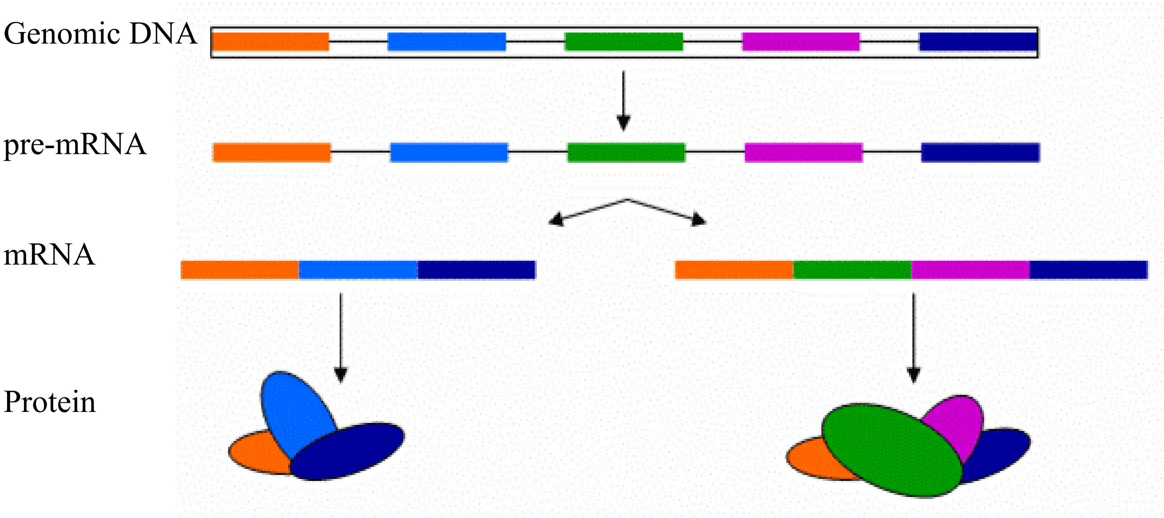


Figure 3. The steps from genomic DNA to protein in an alternatively spliced gene. A gene is first transcribed into pre-mRNA and then processed, removing the exons as well as adding a 5' cap and a poly-A tail (not shown). The processed product is called mRNA, which is then translated into protein.

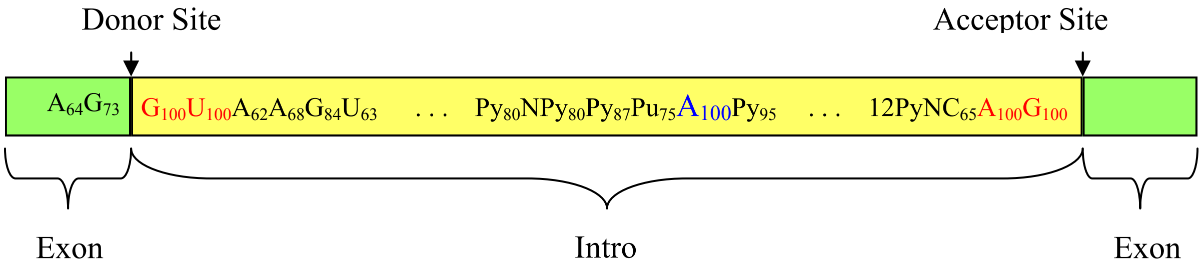

Donor Site
Acceptor Site
A64G73
G100U100A62A68G84U63 . . . Py80NPy80Py87Pu75A100Py95 . . . 12PyNC65A100G100
Exon
Intro
Exon

Figure 4. Splice site consensus sequences at the donor and acceptor sites of pre-mRNA. The subscripts indicate the percent occurrence of a particular nucleotide at that location. Note that the GU at the donor site and AG at the acceptor site are conserved in all introns. The yellow box indicates and intron and the green boxes indicate exons. Please note that, although the GU-AG nucleotides are listed as having 100% conservation, in reality a small fraction of the sites, less than 1%, have a GC-AG sequence and approximately 0.1% have an AU-AC sequence (adapted from Lewin 2004).

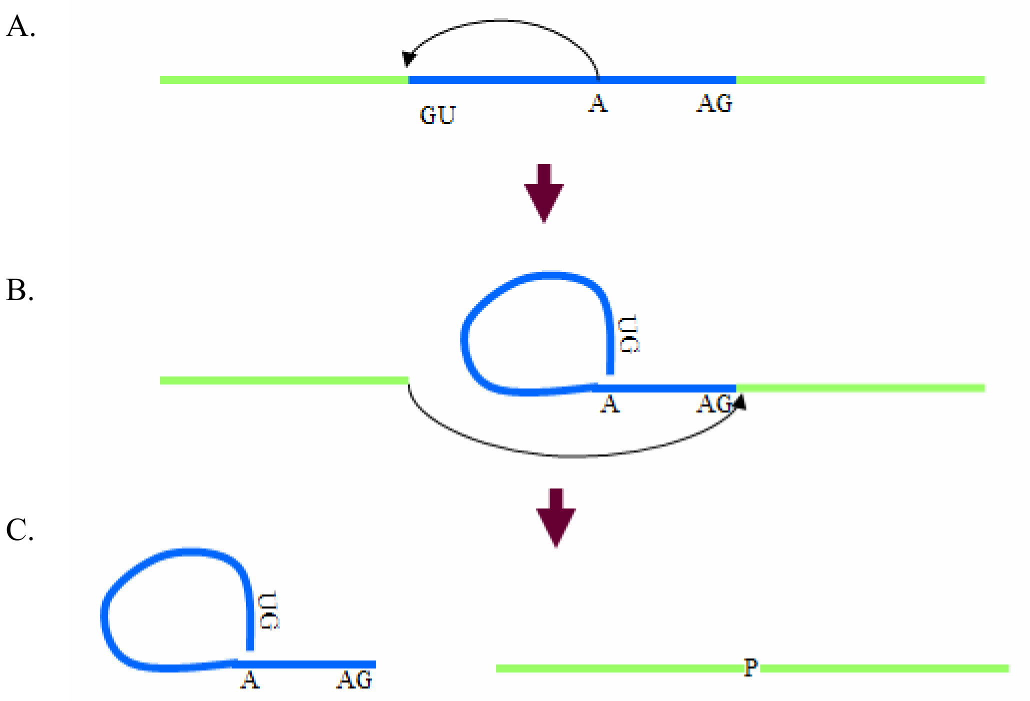


Figure 5. The basic process of intron removal via two transesterfication reactions. Exons are in green and introns in blue. Black arrows represent the nucleophilic attack by a hydroxyl group on a phosphodiester bond. The result of this splicing is two exons connected together by a phosphodiester bond, represented by the P (adapted from Lewin 2004).

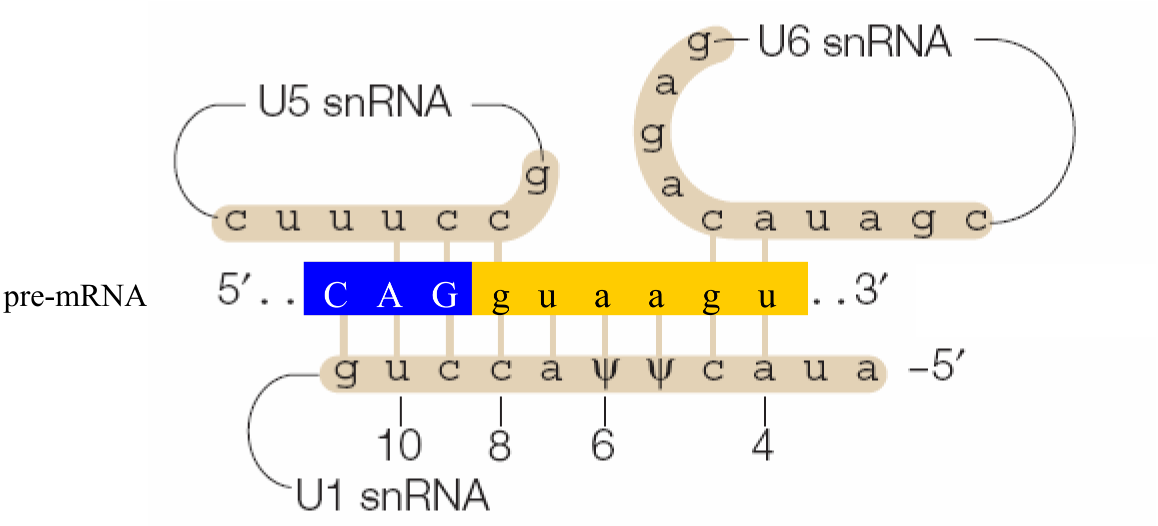
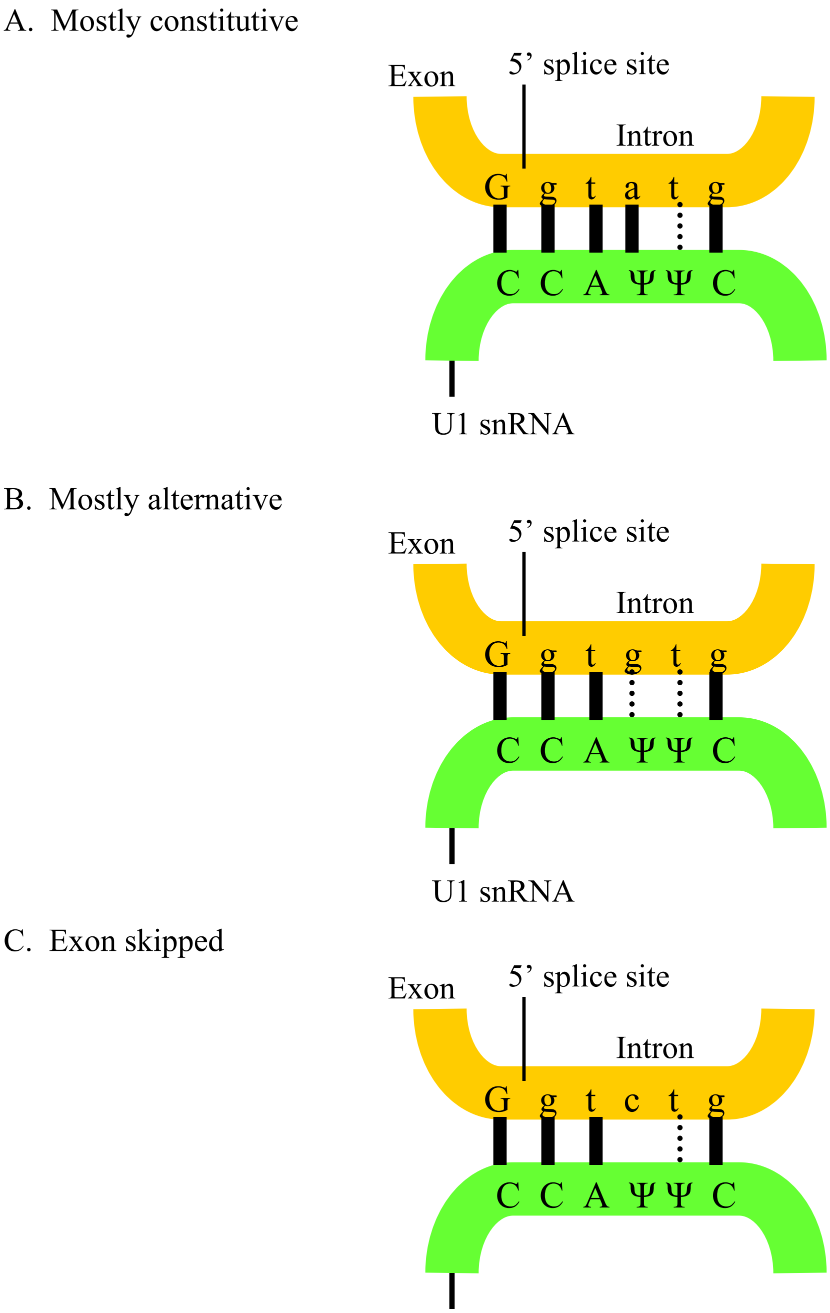


Figure 6. Watson-Crick base pairing between the 5' splice site of pre-mRNA and three members of the spliceosome, U5 snRNA, U6 snRNA, and U1 snRNA. The pre-mRNA is at the center, with capital letters on a blue background indicating the exon and lower-case letters on a yellow background representing the intron. The Ψ indicates a pseudo-uridine (adapted from Ast 2004).



A. Mostly constitutive

B. Mostly alternative

C. Exon skipped

Figure 7. The base pairing of three splice site variants with U1 snRNA. The pre-mRNA is the sequence on top, with position three varying between the transcripts. (A) In sites with A in positions three, the exon is most often included, (B) while sites with a G in that position may or may not be recognized, typically resulting in alternative splicing. In splice sites that have a C at position three, the exon is skipped. Solid lines indicate Watson-Crick base pairing while dashed lines are non-Watson-Crick pairing. The Ψ represents a pseudo-uridine and N indicates non-specific nucleotides (adapted from Ast 2004).

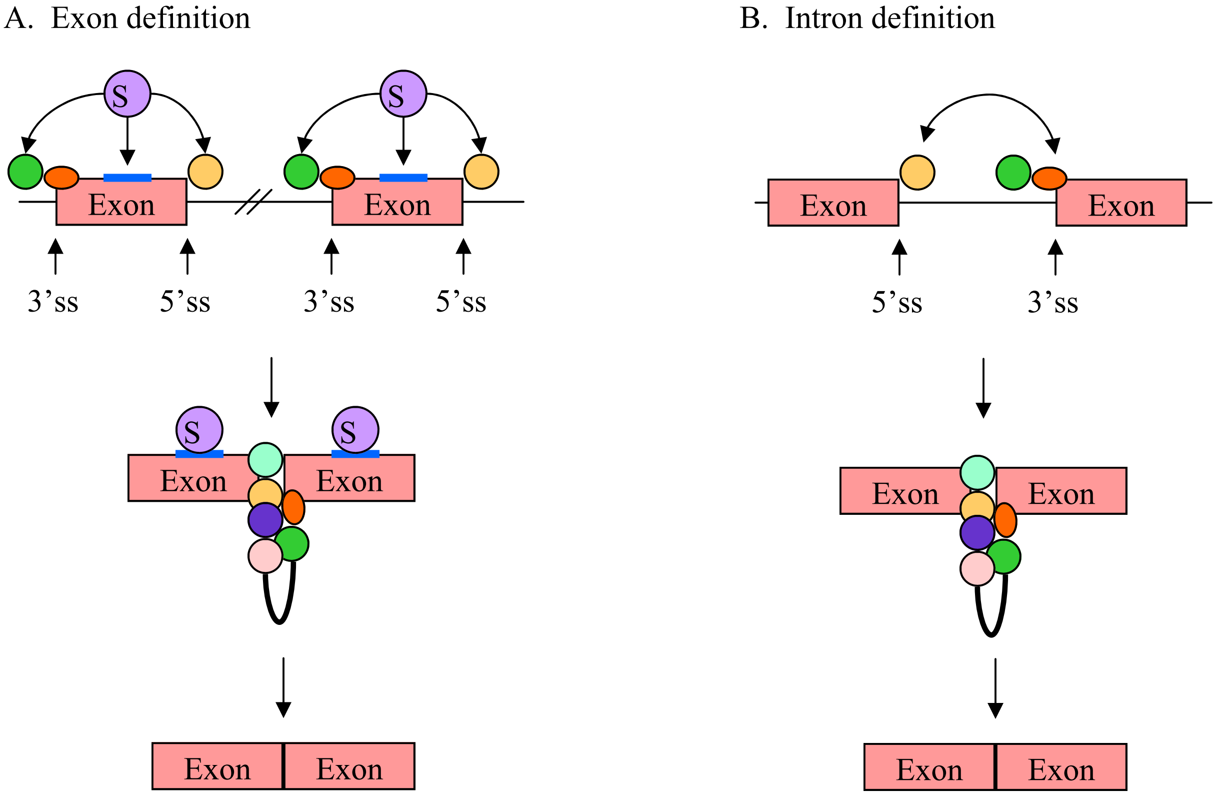


Figure 8. The different ways in which introns defined. (A) Exon definition requires additional information not contained within the intron called exon splice enhancers (ESEs, blue). SR proteins (purple) are conserved serine- and arginine-rich proteins that regulate splicing. These proteins bind to ESEs and instigate a cross-exon recognition complex. (B) While intron definition can also involve SR proteins, all the information for splice site excision is contained within the intron. Pink rectangles are exons and unlabeled colored circles are various proteins involved in splicing (adapted from Ast 2004).

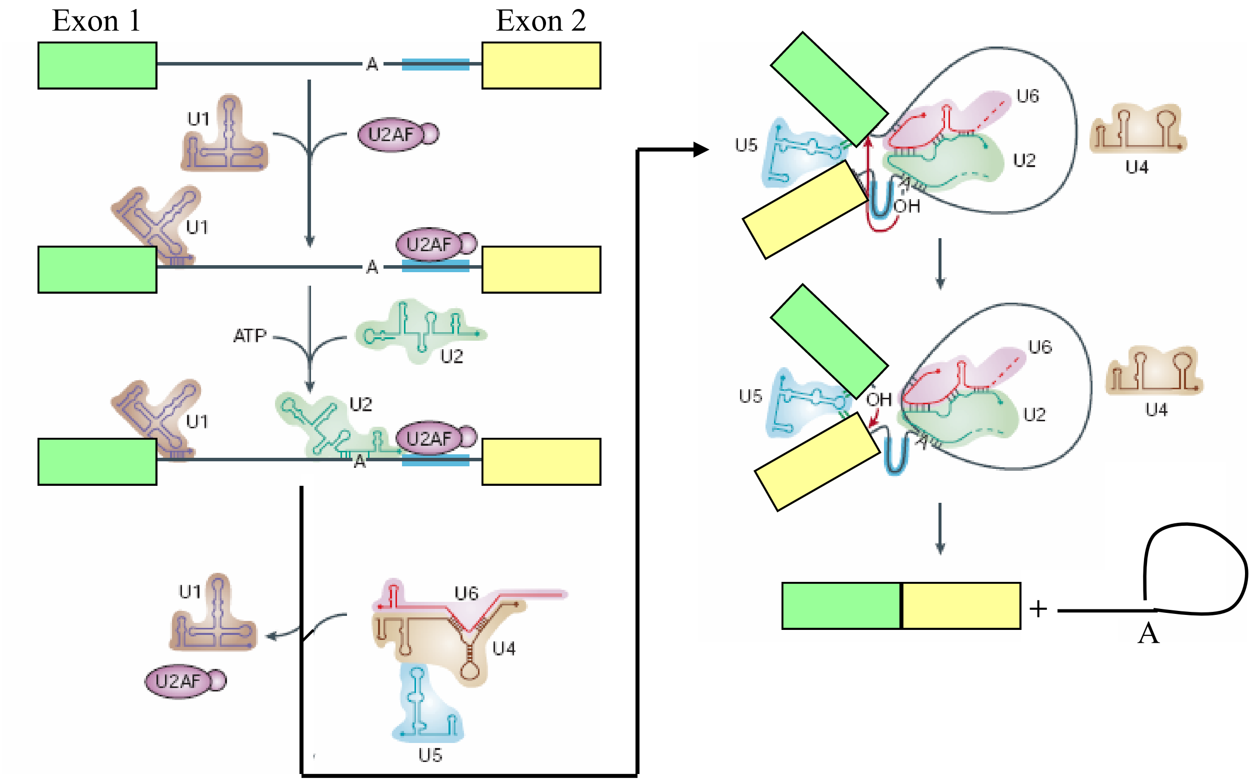


Figure 9. Assemblage of spliceosome on pre-mRNA and removal of intron, illustrating the many different components involved in the splicing process. The two transesterfication reactions are depicted with red arrows. The snRNPs are shown as small nuclear RNA (i.e. U1, U2, etc.) with the surrounding protein represented by the shading. The polypyrimidine region of the intron is shaded in blue (adapted from Patel and Steitz 2003).

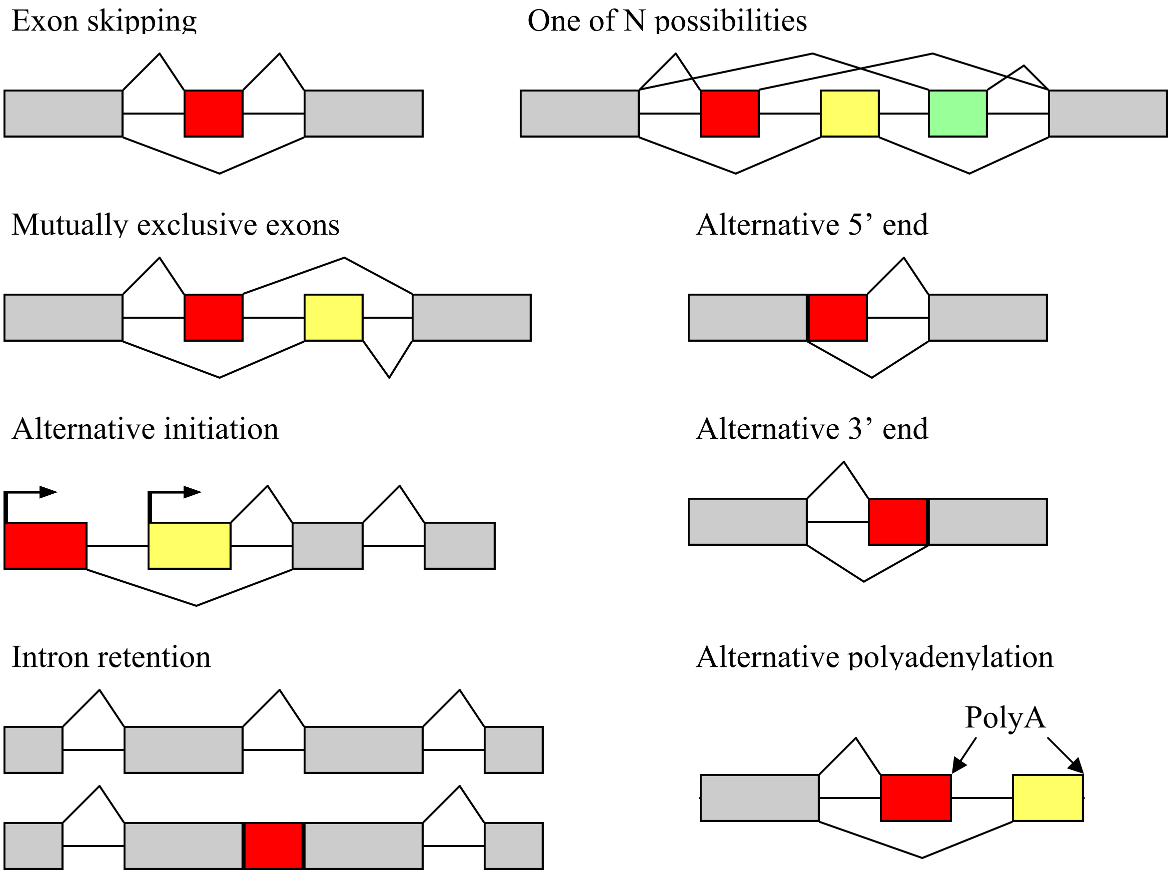


Figure 10. Patterns of alternative splicing. Grey boxes are constitutively spliced exons and colored boxes represent alternatively spliced exons (adapted from Ying and Lee 2006).

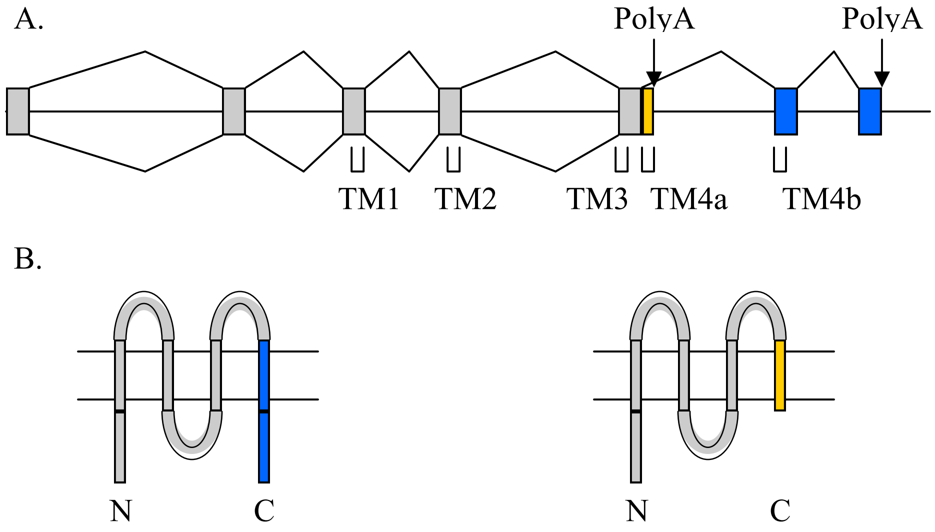


Figure 11. The FC receptor genomic and protein structure. (A) The FC receptor undergoes a form of alternative splicing called polyadenylation. The constitutive exons are in gray while alternatively spliced exons are indicated by colored rectangles. Each Transmembrane domain is indicated by TM and numbered from the amino end of the protein.. (B) The different 3' ends affects the protein structure and function, with different transmembrane domains and cytoplasmic tails. Colors of the regions of the protein correspond with colors of the exons in (A) (adapted from Modrek and Lee 2002).

Transcription factors

For a gene to be transcribed, RNA polymerase II must first bind stably to the promoter region of the gene. However, RNA polymerase will not bind efficiently and precisely without the aid of other proteins. On naked DNA *in vitro*, this suite of helper proteins required for the accurate transcription initiation of RNA polymerase consists of general transcription factors. The general transcription factors bind to specific recognition elements in the core promoter region (Watson *et al.* 2008). The core promoter region is a sequence 40 to 60 nucleotides in length that extends upstream or downstream of the transcriptional start site. These general transcription factors then recruit RNA polymerase to bind to the promoter (fig. 12).

In the cell, additional factors play a regulatory role and can act as repressors or activators, dictating whether or not a gene will be expressed (Taneri *et al*. 2004). Because DNA *in vivo* is complexed into chromatin, additional regulatory sequences are required to ensure efficient transcription. These regulatory regions all bind to proteins that act as activators or repressors, by interacting with proteins associated with the core promoter. Such sequence elements can be located near the promoter or hundreds of kilobases away, interacting with the other proteins via loops in the DNA (Watson *et al.* 2008).

Transcription factors generally contain both a transcription activating (trans-activating) domain and a DNA-binding domain (fig. 13). While the name might imply the activating of transcription, the trans-activating domain is the portion of the protein that effects transcription, and can have either a positive of negative impact on transcriptional levels (Gilbert 2006). The DNA-binding domain serves as the recognition site, binding only to specific DNA sequences. In some cases, these two domains will not be part of the same protein, but instead reside on two separate proteins that work together to bind and activate transcription. Thus transcription factors can provide precise, combinatorial control of genes. This makes these transcription factors critical for organismal development, as they control entire cascades of genes that must be properly regulated to ensure normal growth and specialization of cell and tissue identity.

Transcription factors can be classified by their specific DNA-binding domain, the short sequence of amino acids that are required for the protein to associate with DNA. Classic examples of these domains include the zinc finger, steroid receptor, helix-turn-helix, helix-loop-helix, and leucine zipper, all of which interact with DNA via unique structural motifs. Entire gene families are defined by their specific DNA-binding motifs despite the fact that it comprises a relatively small portion of the protein. In some cases, two members of a gene family may only share the binding domain with little or no homology in the rest of the gene (Carroll *et al*. 2001).

The domains alone do not dictate whether a transcription factor will act as an activator or repressor, however. Because each transcription factor binds in a specific context, interacting with a number of other proteins, the impact any one factor has can vary depending on what other proteins it interacts with. The specific array of regulatory sites associated with a particular gene will determine with which proteins a transcription factor will interact, thus modulating its effects. Therefore, a single transcription factor may act as an activator in one context, but a repressor in another, due to differences in associated proteins.

One family of transcription factors is the Hox gene family that is found in all metazoans and is characterized by a 180 base pair sequence called the homeobox. The homeobox sequence encodes the homeodomain, a DNA-binding domain of 60 amino acids that is a slight variation on the standard helix-turn-helix motif found in *Escherichia coli* and bacterial phage lambda (Brown 2002). Hox genes, which occur in the genome in clusters, have been conserved in species extending from flies to humans (fig. 14). They control the organization of the body along the anterior-posterior axis; different Hox genes are expressed in different segments of the embryo and result in the proper development of that specific region (Lewin 2004). For example, the Hox gene *Antennapedia* (*Antp*) dictates the transcription of a cascade of genes that result in development of thoracic segments. Gain-of-function mutations of *Antp* in the head can result in the fly growing legs where the antennae should be (Carroll *et al*. 2001).

Because transcription factors like *Antp* act as regulators of other genes, their activity in turn must be carefully controlled. The regulation can occur either pre-transcriptionally or post-transcriptionally. One type of post-transcriptional control can be orchestrated by alternative splicing, with differently spliced versions of a single gene being produced at different locations in the organism and changing throughout development. The *Antp* gene undergoes a variety of different splice patterns that are both spatially and temporally specific (Stroeher *et al.* 1988). Such changes in the splice patterns of a transcription factor can influence the regions of DNA to which it binds, alter protein-protein interactions, or change the effect of the transcription factor once bound, all of which would influence which genes a transcription factor regulates. In addition to changes in the effect of a transcription factor due to alternative splicing, the effect of a specific binding protein depends on the sequence context in which it binds. The homeobox protein Ultrabithorax (UBX) can act as either an activator or repressor. UBX activates the transcription of *decapentaplegic* (*dpp*) in the visceral mesoderm of fruit flies, but represses *Antp* (Tour *et al.* 2005). This ability to function as either activator or repressor is the result of binding to different locations in the genome and interacting with different proteins. For example, UBX binds cooperatively with ABD-A to repress *Antennapedia P1*. In other genes, there are multiple UBX binding sites, enabling several UBX proteins to interact via a looping mechanism that enhances transcriptional activation (Brody 1996). Therefore, the effect of a transcription factor can vary greatly depending both on the specific domains it contains as well as the context in which it binds.

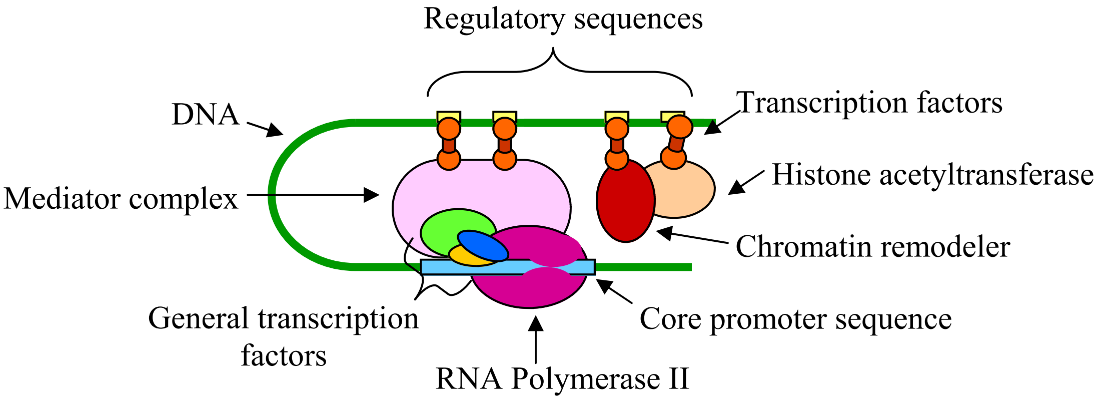


Figure 12. Factors involved in eukaryotic transcription initiation. For transcription of naked DNA *in vitro*, only the general transcription factors are required (dark blue, green, and yellow ovals). These basal transcription factors bind to the core promoter sequence and recruit RNA polymerase II (purple). However, due to careful packaging of the DNA *in vivo*, additional elements are required. These include transcription factors (orange) that bind to regulatory sequences (light yellow) of the DNA (dark green). The effect of transcription factors on the core promoter is facilitated by the mediator complex (pink). Transcription factors also interact with proteins that are involved in the packaging of the DNA, including histone acetyltransferases (peach) and chromatin remodelers (burgundy) (adapted from Watson *et al.* 2008).

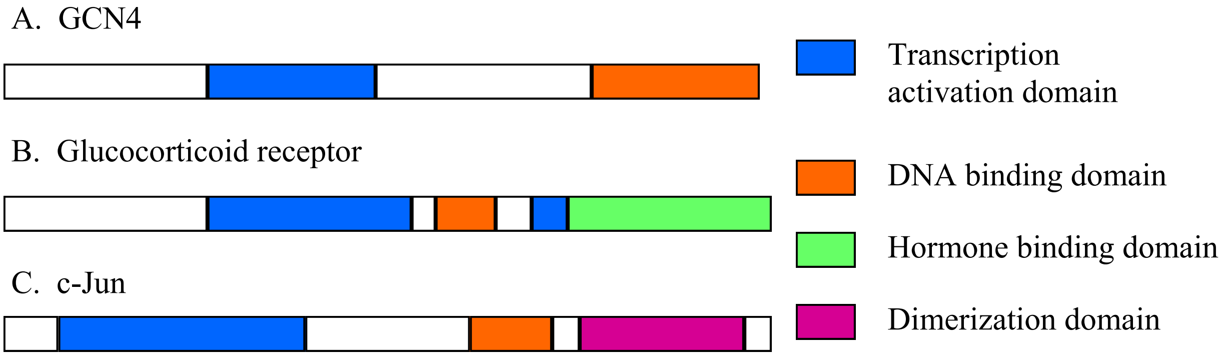


Figure 13. The basic structure of three transcription factors showing the modular nature of the proteins. (A) GCN4, found in *Saccharomyces cerevisiae*, modulates the translation of enzymes that are involved in the biosynthesis of amino acids (Voet and Voet 2004). (B) The glucocorticoid receptor is a transcription factor that becomes active only when bound to glucocorticoid (Lewin 2004). (C) The transcription factor c-Jun is involved in cell division (Kaiser *et al*. 2006). Note that the DNA-binding and trans-activating domains are not the same in the different proteins, despite their identical depiction. Not to scale (adapted from Latchman 1997).

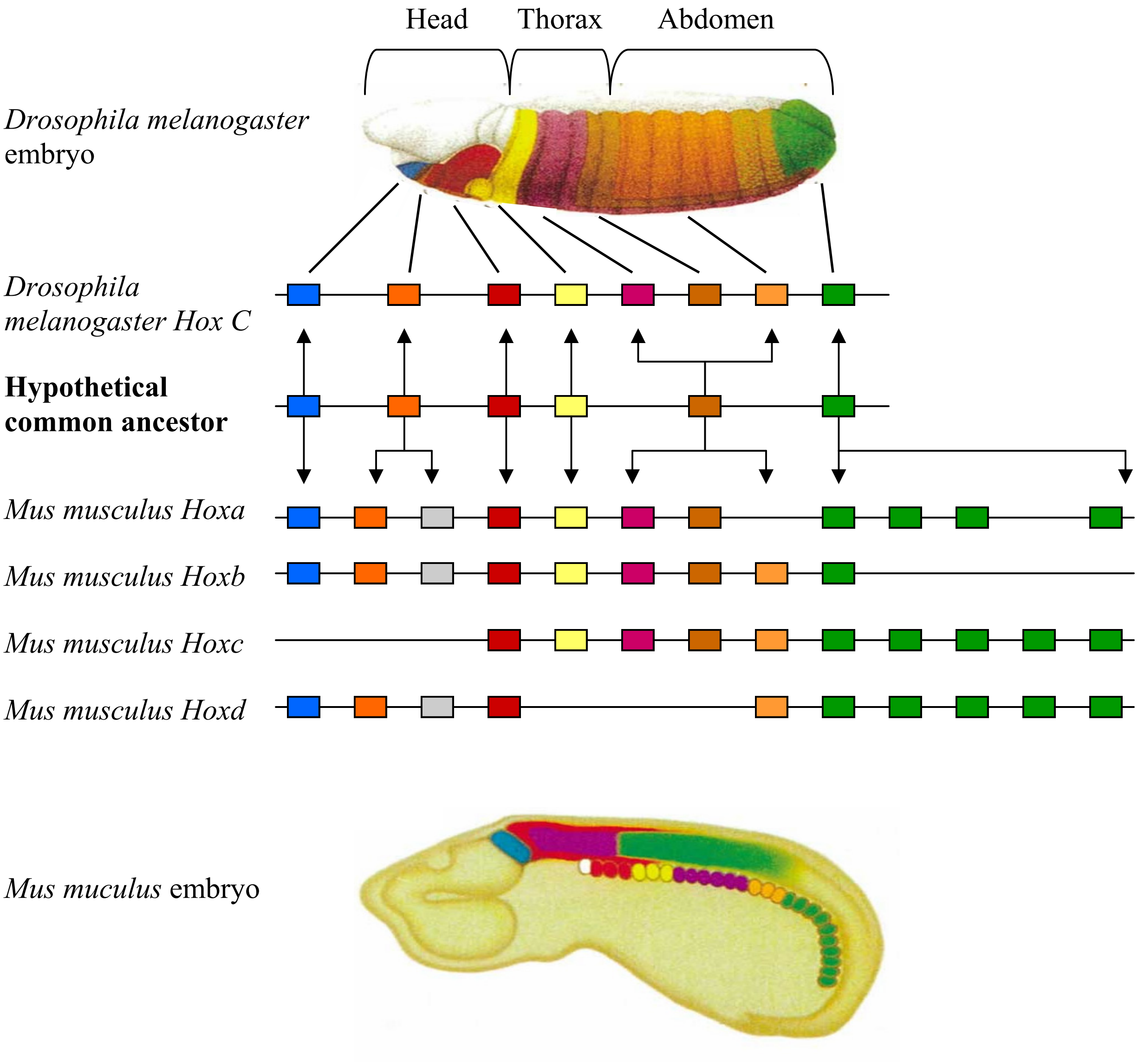


Figure 14. Organization and expression of the *Hox* gene in fruit fly (*D. melanogaster*) and mouse (*Mus musculus*) embryos, including details of the evolutionary relationship between the *Hox* genes in both species, with hypothetical ancestral genes. Mice have four different clusters of *Hox* genes, *Hoxa* – *Hoxd*, while the fruit fly has only one. Colors indicate orthologous genes (adapted from Carroll *et al*. 2001).

Sex determination and the DM domain

Transcription factors function as activators and repressors throughout the life of the organism, ensuring that the proper genes are turned on and off. One area of development which must be carefully regulated is the determination of the sex. In sexually dimorphic organisms, an initial signal is present that triggers the regulation of a whole cascade of genes, resulting in a particular animal developing into a male or female. This initial switch varies greatly. Just among vertebrates, for example, some organisms use a genetic cue as the initial signal, while others are dependent upon the environment. For most land turtles, all crocodilians, and all sea turtles, the sex of an animal is the result of the temperature at which the egg is incubated during the thermosensitive period (Manolakou *et al*. 2006). In contrast, it is the presence or absence of certain sex chromosomes in the embryo that determine its sexual development in mammals, snakes, and birds. Even here, however, the system varies, with mammals following the XX/XY system in which the presence of the Y chromosome results in male development and birds and snakes following the ZW/ZZ system, where ZZ organisms become male (Manolakou *et al*. 2006). It is unclear in the ZW/ZZ system whether it is the numbers of Zs or the presence W that determines the sex.

Significant differences are seen in invertebrates as well. The initial signal in *D. melanogaster* and *C. elegans* is the ratio of sex chromosomes to sets of autosomes. In contrast, the order hymenoptera, which includes bees, ants, sawflies, and wasps, exhibits haplodiploidy whereby fertilized diploid eggs usually become female and unfertilized, haploid eggs develop into males (fig. 15). However, it is not ploidy itself that is the primary signal and the molecular mechanism by which sex determination is accomplished varies

(Manolakou *et al*. 2006). Much research has examined the molecular mechanism by which sex is determined in the honeybee, *Apis mellifera*. In *A. mellifera*, the initial signal depends on a gene called the complementary sex-determiner (*csd*) that codes for an arginine serine-rich (SR) type protein (Beye 2004). It is interesting to note that the SR type proteins also include TRA, a protein involved in sex-determination in *D. melanogaster*, which is not haplodiploid (see below). There are at least 19 different alleles of the *csd* gene and individuals that are heterozygous at the *csd* locus develop into females while those that are hemizygous or homozygous become male (Beye *et al.* 2004). In honeybees, homozygous males are eaten by workers shortly after hatching and, if allowed to mature, are sterile (Cho *et al*. 2006). These results suggest that the protein product of the *csd* allele acts as a dimer, with only the heterodimer being functional. This specific molecular mechanism results in unfertilized, hemizygous eggs developing into males, while fertilized, heterozygous eggs develop into females.

At least 50 different species of hymenopterans use this system of complementary sex determination (Cook and Crozier 1995; Haig 1998). However, in an order consisting of over 200,000 species, that leaves plenty of room for variety (van Wilgenburn 2006). One hymenopteran that appears to have a different master signal is the wasp *Nasonia vitripennis*. Although little is known with respect to the molecular basis of sex determination in *N. vitripennis*, various lines of genetic evidence suggest that the primary sex determining signal in this species involves an imprinted gene, a novel mechanism not previously linked to sex determination (Trent *et al*. 2006). Imprinted genes bear epigenetic instructions that are established in the parental germ cells (Reik and Walter 2007). These epigenetic marks then affect gene expression in the offspring. In the system of sex determination proposed in *N.*

vitripennis, an active paternal copy of a gene is required for female sexual development, while the maternal copy of this gene is silenced by imprinting.

Whatever this initial signal, it triggers a cascade of genes that result in the morphological differences that distinguish the sexes. However, unlike many other signaling pathways that are highly conserved between organisms, the genes involved in sex-determination vary greatly. In fact, none were thought to be conserved between phyla until the discovery of a small domain shared by a gene in the cascades of *D. melanogaster* and *C. elegans* (Hodgkin 2002). At the bottom of the sex-determining cascade in both organisms is a transcription factor, called *doublesex* in *D. melanogaster* and *mab-3* in *C. elegans* (fig. 16). These two proteins share a conserved DNA-binding, named the DM domain.

Since discovering the presence of the DM domain in flies and nematodes, researchers have found that this motif is involved in the sex-determination of species that vary from the Queensland fruit fly (Zarkower 2001) and honey bee (Cho *et al.* 2007) to humans (*Homo sapiens*) (Hodgkin 2002).

The DM domain belongs to a family of transcription factors called zinc fingers. A zinc finger consists of a short sequence of conserved amino acids that bind to a zinc ion in a tetrahedral structure (fig. 17). Zinc fingers are about 23 amino acids in length and commonly occur as part of a series of zinc fingers with seven to eight amino acids linking them together (Lewin 2004). The structure of this DNA-binding motif consists of short double-stranded β sheet followed by an α helix in which the zinc ion is bound. This is followed by another β sheet. Typically, this structure wraps around the DNA in the major groove (fig. 18). Over 1,000 distinct zinc fingers have been identified via sequence analysis (Luscombe *et al*. 2000).

In the DM domain, a series of cysteine and histidine residues interact with two Zn^{2+} atoms (fig. 19). While other regions of DM-containing genes vary widely, these cys and his residues are highly conserved (fig. 20) and define this class of zinc fingers called the DM domain. In addition to the Zn^{2+}-binding site, the DM domain has a disordered tail. When this tail comes in contact with the DNA it forms an α-helix that serves to recognize the proper binding site (Zhu *et al*. 2000). As noted above, a zinc finger typically interacts with the major groove of the DNA. However, the DM domain binds to the minor groove of DNA (Zhu *et al*. 2000).

The way in which the DM containing gene influences the sex-determination varies between organisms. For example, the *mab-3* gene in *C. elegans* is turned on in males, but off in hermaphrodites. In contrast, the *D. melanogaster* gene *dsx* undergoes alternative splicing at the 3’ end, producing one transcript specific to males and another specific to females (fig. 21). While the 5’ ends of both transcripts are identical, including the region that codes for the DM domain, the alternative splicing results in differences in the C-terminus of the protein. This includes variation in the oligomerization domain, a motif present in DSX that enables these proteins to form dimers (Schutt and Nothiger 2006). While both transcripts contain identical DM domains and therefore bind to identical sequences, their effects differ due to differences in the carboxyl region. DSX-M represses the genes required for female differentiation and activates those necessary for male differentiation while DSX-F activates female-specific genes and represses those that are male-specific (Yang *et al*. 2008).

This raises two questions. First, how can one gene act as both a repressor and an activator, and, second, how can two transcription factors bind to the same regulatory sequence but have differing effects on transcription? The answer to both is context.

Specifically, what other regulatory sequences are located nearby and how DSX-M and DSX-F interact with both the proteins that bind to these regulatory sequences and to co-activators influences whether the transcription factor will act as an activator or repressor at a specific site. For example, the regulatory region of the yolk protein gene includes three regulatory sequences, aef1, dsxA, and bzip1, which bind the AEF1 repressor, DSX-M and DSX-F, and, possibly, the DmC/EBP activator, respectively (An and Wensink 1995). The yolk protein gene is turned on in females because DSX-F binds to dsxA and sterically excludes the AEF1 repressor while also working with a protein bound to bzip1 to activate transcription. In contrast, because the DSX-M has a longer carboxyl tail making it larger than DSX-F (fig. 22), it represses transcription by interfering with the protein that binds at the bzip1 site, either though physically obstructing the regulatory sequence or inactivating the protein once bound (An and Wensink 1995). Therefore, although both DSX-M and DSX-F bind to the same regulatory sequence, they have opposite effects, resulting in the sex-specific expression of the yolk protein.

While most work on DM genes focuses on those that are involved in sexual differentiation, the presence of the DM domain does not guarantee that the gene has a role in sex-determination. As is common in gene families, once duplicated, DM genes have diverged in function. For example, *terra*, a well-characterized gene in zebrafish (*Danio rerio*), is involved in somatic mesoderm development, and is expressed identically in both sexes (Meng *et al*. 1999). The teleost Medaka (*Oryzias latipes*) has a group of DM-containing genes called the dmrt family (Winkler *et al*. 2004). One member of this family, *dmrt1by*, regulates male development. However, paralogs of *dmrt1by*, such as *dmrt1a*, *dmrt2*, *dmrt3*, and *dmrt4*, are expressed in a non-sex-specific manner. The genes *dmrt2*, 3, and *4* are involved in embryogenesis and all four genes show differing temporal patterns of expression (Winkler *et al*. 2004).

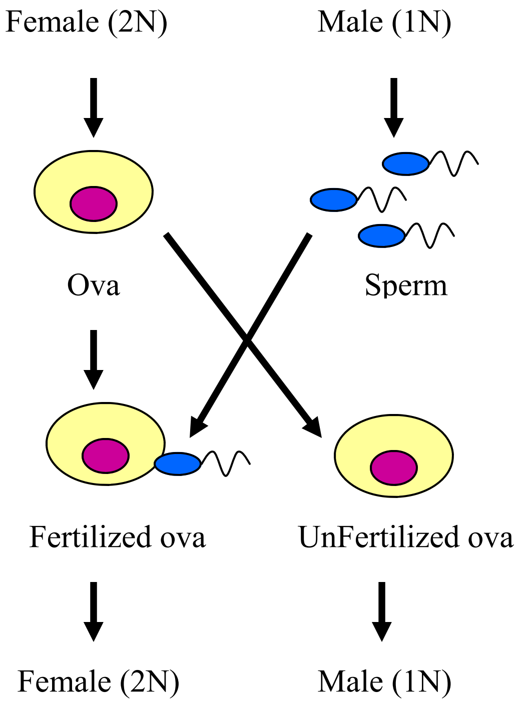


Figure 15. Reproduction in haplodiploid insects. Fertilized eggs develop into females while unfertilized eggs become males (adapted from Bull 1983).

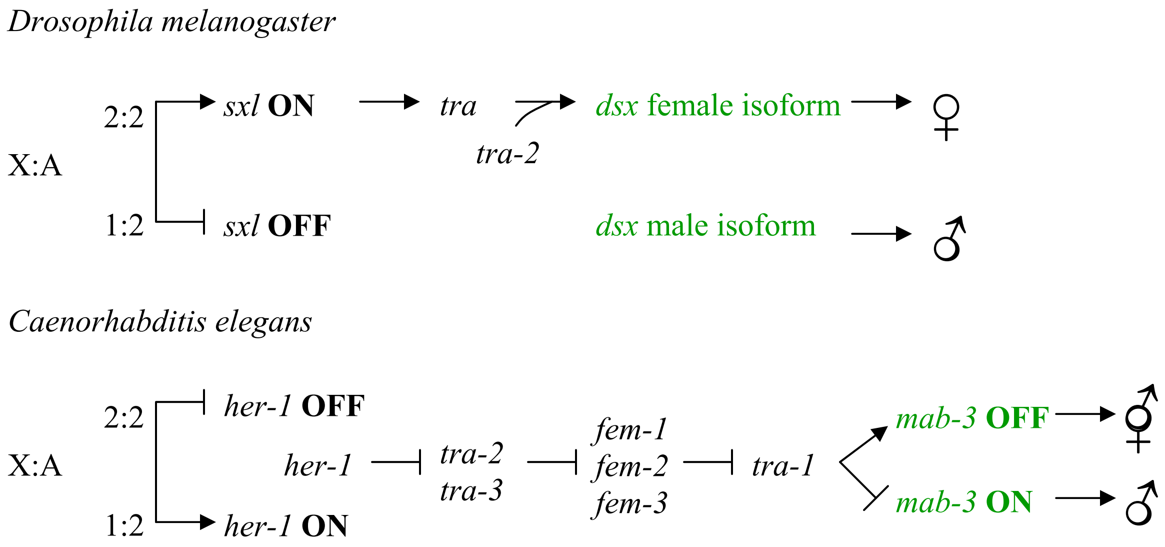


Figure 16. Sex determining cascade of D. melanogaster and C. elegans, both of which include a DM-containing gene (green). Pathways above DM genes are not conserved (adapted from Zhu *et al.* 2000).

Cys - $X_{2\text{-}4}$ - Cys - X_3 - Phe - X_5 - Leu -X_2 – His - X_3 - His

Figure 17. The consensus sequence of the zinc finger motif. The X stands for any amino acid, with subscript indicating the number (adapted from Lewin 2004).

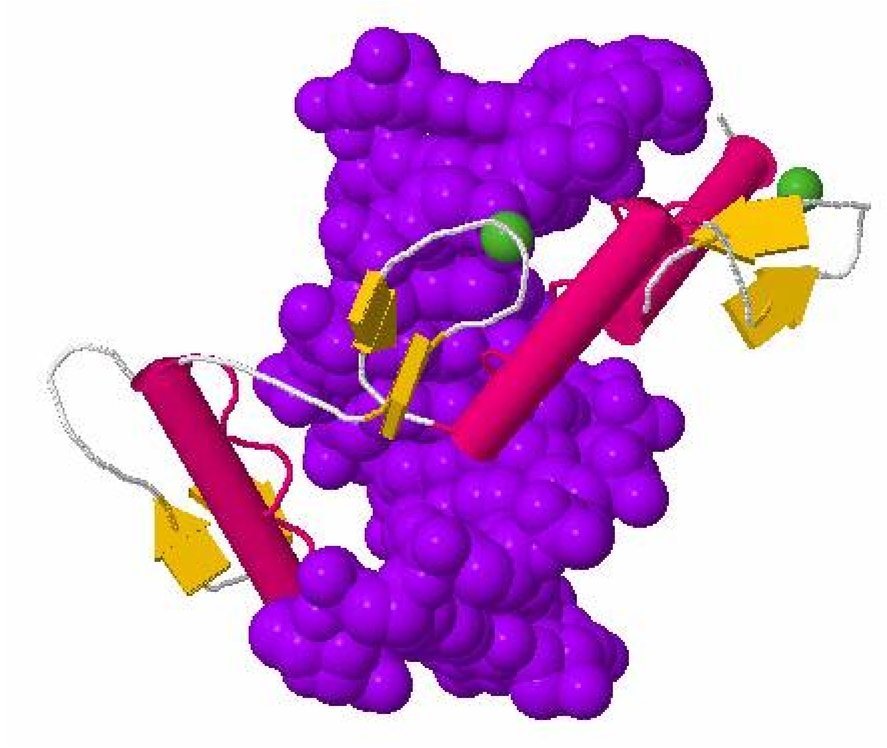


Figure 18. A series of three zinc fingers contacting DNA. The α helix is represented by the magenta cylinders, the β sheets by the yellow arrows, and the associated zinc atoms by green balls. The DNA is in purple (PDB: 1aay).

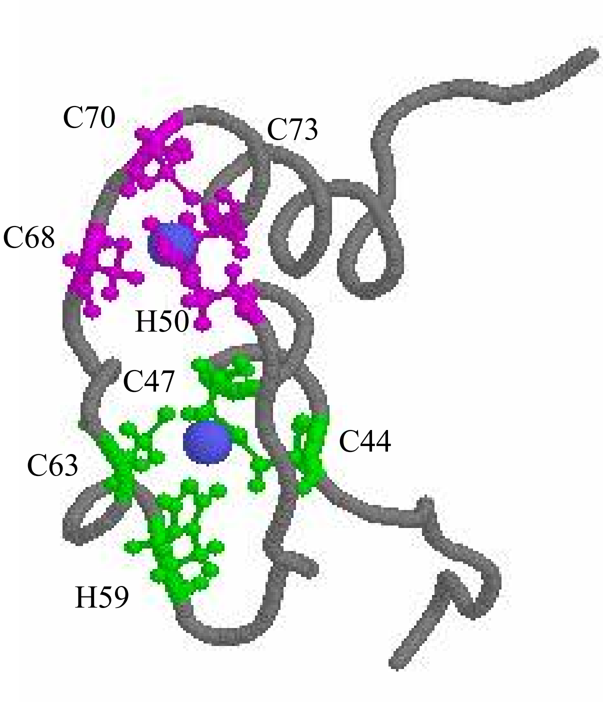
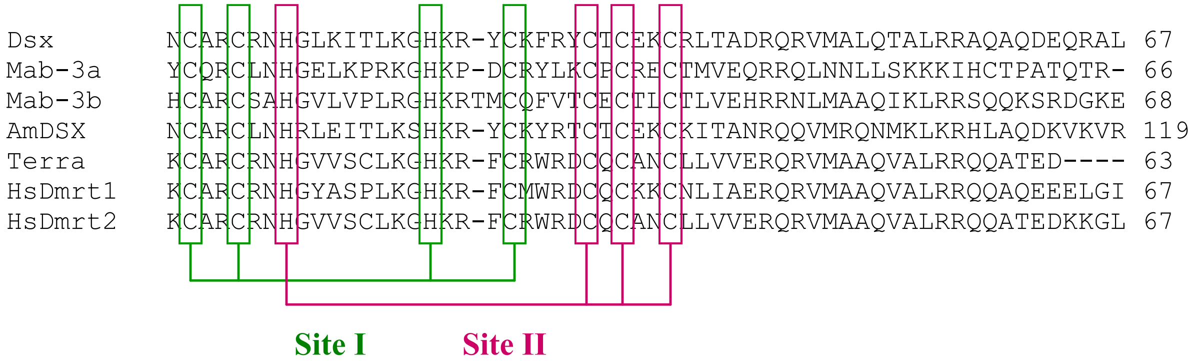


Figure 19. DM domain in DSX. Conserved cysteines and histidines are colored. The green coloration corresponds with the site I Zn-binding sites seen in figure 20 and the magenta corresponds to site II. The zinc ions are in blue (PDB: 1plv).



```
Dsx       NCARCRNHGLKITLKGHKR-YCKFRYCTCEKCRLTADRQRVMALQTALRRAQAQDEQRAL 67
Mab-3a    YCQRCLNHGELKPRKGHKP-DCRYLKCPCRECTMVEQRRQLNNLLSKKKIHCTPATQTR- 66
Mab-3b    HCARCSAHGVLVPLRGHKRTMCQFVTCECTLCTLVEHRRNLMAAQIKLRRSQQKSRDGKE 68
AmDSX     NCARCLNHRLEITLKSHKR-YCKYRTCTCEKCKITANRQQVMRQNMKLKRHLAQDKVKVR 119
Terra     KCARCRNHGVVSCLKGHKR-FCRWRDCQCANCLLVVERQRVMAAQVALRRQQATED---- 63
HsDmrt1   KCARCRNHGYASPLKGHKR-FCMWRDCQCKKCNLIAERQRVMAAQVALRRQQAQEEELGI 67
HsDmrt2   KCARCRNHGVVSCLKGHKR-FCRWRDCQCANCLLVVERQRVMAAQVALRRQQATEDKKGL 67
```

Figure 20. Alignment of the DM domain from *D. melanogaster* (DSX), *C. elegans* (Mab-3a and Mab-3b), *Apis mellifera* (AmDSX), *Danio rerio* (Terra), *Homo sapien* (Dmrt1 and Dmrt2). Conserved cysteine and histidine residues are outlined. Site I and Site II refer to Zn-binding sites (adapted from Zhu *et al.* 2000).

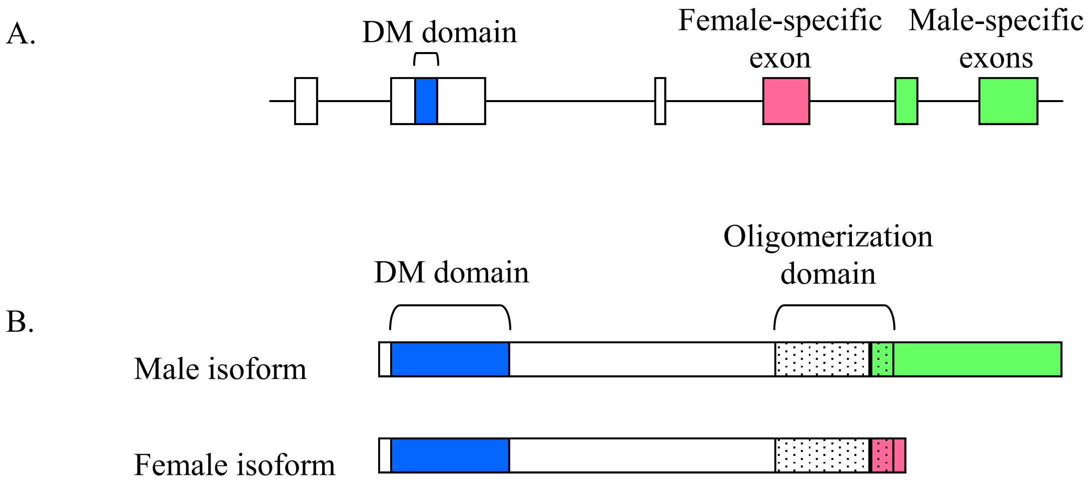


Figure 21. Alternative splice patterns of *doublesex* in *D. melanogaster*. (A) The gene includes three exons that are constitutively spliced (white), followed by a female-specific exon (pink), and two male-specific exons (green). (B) The alternatively splicing results in two different transcripts with different C-termini. The DM domain appears in blue and the oligomerization domain is speckled, with the white background indicating the shared region of this domain present in both sexes and the green and pink backgrounds indicate the region and is male- or female-specific, respectively (adapted from Zhu *et al*. 2000).

```
   LGQDVFLDYCQKLLEKFRYPWELMPLMYVILKDADANIEEASRRIEEARVEINRTVA
♂  QIYYNYYTPMALVNGAPMYLTYPSIEQGRYGAHFTHLPLTQICPPTPEPLALSRSPS
   SPSGPSAVHNQKPSRPGSSNGTVHSAASPTMVTTMATTSSTPTLSRRQRSRSATPTT
   PPPPPPAHSSSNGAYHHGHHLVSSTAAT

♀  LGQDVFLDYCQKLLEKFRYPWELMPLMYVILKDADANIEEASRRIEEGQYVVNEYSR
   QHNLNIYDGGELRNTTRQCG
```

Figure 22. The male and female versions of the oligomerization domains and carboxyl tails in *D. melanogaster*. This region of the protein has different protein-protein interactions in males and females, resulting in the sex-specific activation and repression of other genes. The blue and pink high lights indicate the male- and female-specific region of the oligomerization domain, respectively. Colored text represents the sex-specific carboxyl ends.

Hymenopterans: *Nasonia vitripennis*

My research examines the molecular basis of sex determination in *Nasonia vitripennis*, a diminutive parasitoid wasp from the order hymenoptera. Unlike honeybees, *N. vitripennis* do not use variation at a *csd* locus as their primary signal. Highly inbred strains of *N. vitripennis* still produce males and females, which would not be the case sex-determination required heterozygosity of a specific allele (Trent *et al*. 2006). While both honeybees and wasps are hymenopterans, they vary in life history patterns, and thus, as seen in vertebrates, it is not surprising that their primary sex-determining signal varies as well. As discussed previously, genomic imprinting has been proposed as the primary signal for sex determination in which female development requires the presence of an imprinted gene that is expressed only from the paternal copy (Trent *et al*. 2006).

Despite the incredible variation in primary sex-determining signal and resulting cascades, the conservation of DM-containing genes at the bottom of the cascade in organisms as varied as nematodes and humans led Carol Trent and members of her lab at Western Washington University to focus on searching for a DM-containing gene involved in sex-determination in *N. vitripennis*. Studies in genetics, evolutionary biology, development, behavior, and ecology have utilized *N. vitripennis* and, because of its importance as a model organism, this wasp was chosen for complete genomic sequencing and now has its own genome project. However, the search for DM-containing genes in *N. vitripennis* began in the Trent lab before the initiation of the genome project and used a PCR-based approach with degenerate primers that targeted the conserved amino acids of the DM domain. Two *N. vitripennis* genes containing the DM domain were identified by Andrea Llewellyn.

In this thesis, I detail the results of my work in the Trent lab with the DM gene family in *N. vitripennis*. My research, which utilized both genetic and computational techniques, included the addition of two new genes, bringing the total number of genes in the DM gene family of *N. vitripennis* to four. Beginning with the partial sequences of two DM-containing genes found previously by the Trent lab called NvDM1 and NvDM2, I completed the DM domains of both genes using computational methods. I also confirmed the alternative splicing pattern seen previously in NvDM1 and showed that this was not sex-specific in yellow pupae, the first stage at which males and females can be easily differentiated anatomically. By computational analysis of these two genes, I confirmed that neither were homologs of *dsx*. Then, using a combination of computational and molecular techniques including RACE and RT-PCR, I found two more members of the *N. vitripennis* DM gene family, which I named NvDM3 and NvDM4. Further analysis revealed that NvDM4 undergoes extensive alternative splicing that results in four different classes of transcripts, two of which appear to be sex-specifically spliced. I concluded that NvDM4 is the closest homolog to *dsx* in *N. vitripennis* and is likely involved in the sex-determination pathway.

METHODS

Nasonia vitripennis genome

As an important model organism, the complete genome of *Nasonia vitripennis*, in addition to those of humans, bees, flies, and various other organisms, is available to the public through GenBank, which is managed by the National Institute of Health (NIH) through the National Center of Biotechnology Information (NCBI). GenBank provides an annotated collection of DNA, RNA, and protein sequences that can be analyzed using a wide range of publicly available computational tools (Benson *et al*. 2006). The Human Genome Sequencing Center (HGSC) at Baylor College of Medicine sequenced the *N. vitripennis* genome, and version 1.0 of the release was published in April 5, 2007, version 0.5 having been released less than a year earlier on July 9, 2006 (Appendix A). Version 1.0 consists of 5936 scaffolds of genomic DNA that have yet to be placed on chromosomes (NCBI).

In addition to a searchable nucleotide database, the NCBI *N. vitripennis* database includes an official gene set, referred to as RefSeq. Genes included in the RefSeq database were predicted using a gene prediction program called Gnomon that uses a Hidden Markov Model (HMM)-based algorithm to find genes within a genome in a multi-step process (NCBI). Gnomon uses both *ab initio* predictions and sequence homology as part of its predictive process (Rice Annotation Project *et al*. 2008). In the case of the *N. vitripennis* gene set, Gnomon used a set of approximately 80,000 expressed sequence tags (ESTs) from *N. vitripennis* in addition to homology with curated proteins from *Drosophila melanogaster*, other insects, *Homo sapiens*, *Mus musculus*, *Caenorhabditis elegans*, and *Saccharomyces* *cerevisiae*. This gave rise to a dataset of 9254 sequences. Predicted genes that lacked homology or EST support were placed in the *ab initio* database, which includes a total of 27,287 predicted genes. Therefore, while genes in the RefSeq dataset often have proposed biological function, predicted genes in the *ab initio* database lack such information. These databases can also be accessed through the HGSC.

Effective use of a database requires the ability to search through it effectively and, in the case of the genome, this can be done using the Basic Local Alignment Search Tool (BLAST) (Altschul *et al*. 1990). Different types of BLAST programs allow for different searches, varying by query sequence used and database to be searched (table 2). BLAST uses the same basic three stage algorithm regardless of the type of search performed. First, the program looks for short exact matches of length W between the query and the database. These matches are referred to as seeds. Then, BLAST attempts to extend the match in both directions from the seed. Finally, if the first two steps have resulted in a high-scoring un-gapped alignment, a variation on the Smith-Waterman algorithm is used to perform a gapped alignment (Altschul, *et al*. 1990). The resulting alignments are then displayed to the user in order of increasing E-value.

The HGSC also provides access to the *N. vitripennis* genome with corresponding BLAST capabilities. BLAST results from the HGSC are linked directly to a program called Geneboree. Among other capabilities, Geneboree allows the user to retrieve segments of the genome corresponding to the BLAST hit. For example, the user can access the genomic material to the 5' of an alignment for further analysis.

Table 2. Description of BLAST types used in this project.

BLAST Type	Query Type	Database Type
megaBLAST	nucleotide	nucleotide
BLASTP	protein	protein
BLASTX	nucleotide	protein
TBLASTN	protein	nucleotide

Sequence analysis

I used a variety of programs to analyze both sequences retrieved from the databases and the sequences that I acquired through my experiments (see below). These programs gave me the ability to align mRNA transcripts with the genome to establish exon boundaries, to measure the strength of the splice sites at these boundaries, to translate the sequence into amino acids, to scan the resulting protein for known motifs, and to align and compare different transcripts.

For Spidey, a program available through the NCBI, the user provides a number of mRNA transcripts and the corresponding genomic sequence and the program aligns the transcripts with the genome. This enabled me to assess exon boundaries and see how splicing patterns varied between alternatively spliced transcripts.

To measure the relative strengths of the alternative splice sites, I looked at the similarity between each site and the consensus splice site sequence (fig. 4). For this, I used the Splice Site Prediction by Neural Network program (NNSPLICE) available through the Berkley Drosophila Genome Project which runs NNSPLICE version 0.9 (Reese *et al*. 1997) as well as the Alternative Splicing Prediction program (ASPIC) available through the Research Network of Bioinformatics and Comparative Genomics (Bonizzoni *et al*. 2005). In NNSPLICE, splice sites are scored on a scale from 0 to 1 with 1 being the most likely to be an actual splice site. ASPIC also scores each predicted splice site, but on a scale from 0 to 100, with higher numbers again indicating a higher probability of an actual splice site. For NNSPLICE, the input consists of the genomic sequence that contain the exon-intron boundary of interest, with the output providing a list of donor and acceptor sites with their sequence, location, and strength/probability. ASPIC, on the other hand, requires both genomic sequence and my various NvDM4 transcripts for input. The output consists of the inferred full length isoforms, with various viewing options, one of which, "Alignment View," provides scores for splice sites donors and acceptors. NNSPLICE and ASPIC produce results using different algorithms. In ASPIC, the user-submitted transcripts are aligned with the genomic sequence and the program attempts to produce the most parsimonious set of splice sites that correspond to the various sequences. With NNSPLICE, a Hidden Markov Model is used for splice site prediction. This model is based upon neural networks that use dinucleotide frequencies

When first using ASPIC, I copied in all of the NvDM4 transcripts – three from class OD-A, four from OD-B, one from OD-X, two from OD-C.1, and one from OD-C.2 – and the complete genome sequence that incorporates all of these transcripts, which is the reverse complement of base pairs 134000 to 185000 on scaffold 23 as defined by HGSC. This, however, did not result in the splicing scores of introns G, H, and I (fig. 66). In order for it to predict these, I entered only the OD-C transcripts and included only the genome sequence that spans 134000-178000, eliminating all but the OD-C-specific region.

For using NNSPLICE, I pasted in small pieces of the genome that contained my splice site of interest. I chose to do this instead of searching the entire genomic sequence corresponding to a particular transcript at once in order to limit the number of possible splice sites I needed to search through to find the site of interest.

In addition to analyzing the nucleotide sequences, I also examined the resulting proteins. To translate mRNA to protein, I used Translate algorithm from the web-based program Expert Protein Analysis System (ExPASy) (Gasteiger *et al*. 2003). ExPASy translates all six frames of possible amino acids.

I then used Pfam to search the amino acid sequences for protein motifs. Pfam is a database of 9318 protein families, each defined by a specific sequence motif (version 22.0, July 2007). Every protein family contains a seed alignment of representative sequences of a particular domain (Sonhammer *et al*. 1998). This alignment, which is verified manually, serves as the basis for the creation of an HMM-profile. The HMM-profile is then used to search SwissProt, a protein database, for all proteins that align with the HMM-profile, below a certain E-value (the "gathering threshold"). These proteins are then aligned to make a full alignment, which is updated regularly as new proteins are added to SwissProt (Sonhammer *et al*. 1998). The user can search the Pfam database with a novel protein and the program will return a list of domains that are present in that sequence and fall within a statistical threshold that can be set by the user.

Both protein and nucleotide sequences were compared using the alignment programs Malign (Nikolaev *et al*. 1997) and ClustalW (Chenna *et al*. 2003). I used Mobyle to create color-coded alignments (Neron *et al*. 2005).

Primer design

Primers were designed using the web-based Primer3 program (Rozen and Skaletsky 2000). I used default parameters except for those listed in table 3. If no acceptable primers could be found using these stipulations, parameters were further modified. I used these specifications for RT-PCR, RACE PCR, and sequencing experiments. Acceptable primers were ordered from Integrated DNA Technologies. Upon receiving the primers, I re-suspended them in TE to 50 or 100µM concentrations. The TE was 10mM tris at pH of 7.5 and 0.1mM of EDTA in molecular grade water. Aliquots were further diluted in molecular grade water to 10µM working stocks before use.

Agarose gel electrophoresis

I analyzed PCR products using gels cast with out on molecular grade agarose in beds made by Jordan Scientific. I used Fisher Scientific Genetic Analysis Grade Agarose for the majority of the gels. An EC135 from the E-C Apparatus Corporation was used to apply constant voltage. I ran most gels at 80V for about an hour in a 1.2% agarose gel, although these conditions were modified as needed due to expected product size. While initial RT-PCR experiments were stained for half an hour with Invitrogen's SyBr Gold (1µl/10mL or 1:10,000), equally good results were obtained by staining for an equal time in ethidium bromide (0.5µg/ml), a much less expensive option. Gels were illuminated using the Ultra-Lum Electronic UV Transilluminator with the UV intensity set to Max. I captured the image using a 10 Megapixil Canon PowerShot A640 camera set to remote capture with the macros on. Images automatically opened in the Canon Utilities ZoomBrowser EX version 5.7. I took a number of images, choosing the one with the clearest image. I saved the best image and then used Photoshop to modify it, inverting the colors and using the auto-contrast function to better observe the bands. Images were then printed and thoroughly labeled.

Table 3. Parameters used for Primer3 that were not default.

	Minimum	Optimum	Maximum
Primer Size	23	25	27
Primer Melting Temperature (°C)	65	70-72	75
Primer GC	40		60

Rapid amplification of cDNA ends PCR

I conducted all rapid amplification of cDNA ends (RACE) PCR experiments using the Clontech SMART™ RACE cDNA Amplification Kit. RACE PCR allows the user to use a single gene specific primer and amplify the region of the mRNA from that primer to the end of the transcript. Therefore, if only a portion of the transcript is known, primers can be designed for that region and amplification of both 3' and 5' RACE performed (fig. 23). The products of these reactions can then be sequenced, enabling the user to learn the entire sequence of the gene.

Synthesizing cDNA

In all cases, I prepared separate female and male batches of RACE-ready cDNA. The RNA I used was total RNA that had been previously isolated from the B2 strain of Nasonia by Carol Trent from male and female yellow pupae on 7/21/00 and 7/18/00 respectively. This strain had been acquired from Mary Ann Pultz. The male RNA, designated J2.1, had concentration of 2.2μg/μl and the female RNA, called K1, had a concentration of 4.4μg/μl as determined by optical density. An aliquot of female RNA was diluted to 2.2μg/μl on 12/1/06. All RNA had been kept at -80°C prior to conversion to cDNA. I followed the protocol as described in the SMART™ RACE cDNA Amplification Kit User Manual. For both males and females, I used 2.2ng of RNA to make the cDNA. For the preparation, all incubations not on ice were done using the Perkin Elmer GeneAmp 9700 thermocycler. I then diluted the cDNA in 100µl of tricine-EDTA buffer included in the kit. Prepared cDNA was stored in the freezer and -20°C until used for RACE PCR.

RACE

For the Rapid Amplification of cDNA Ends PCR, all reactions were reduced to 2/5th of the recommended volume, resulting in a 20µl of product instead of a 50µl PCR reaction. I used the recommended Advantage 2 Polymerase from Clontech. Unless noted otherwise, gene-specific primers (GSPs) had melting temperatures above 70°C and the touchdown PCR protocol described in the manual was used. Products were run on a gel and visualized. To decrease the presence of background or nonspecific amplification, I then used the initial products as my template for nested gene-specific primers (NGSPs). NGSPs were designed to sit inside the original GSPs (fig. 23). These products were also run out on agarose gels and visualized. The products of the GSPs were then compared to the products of the NGSPs for expected shifts.

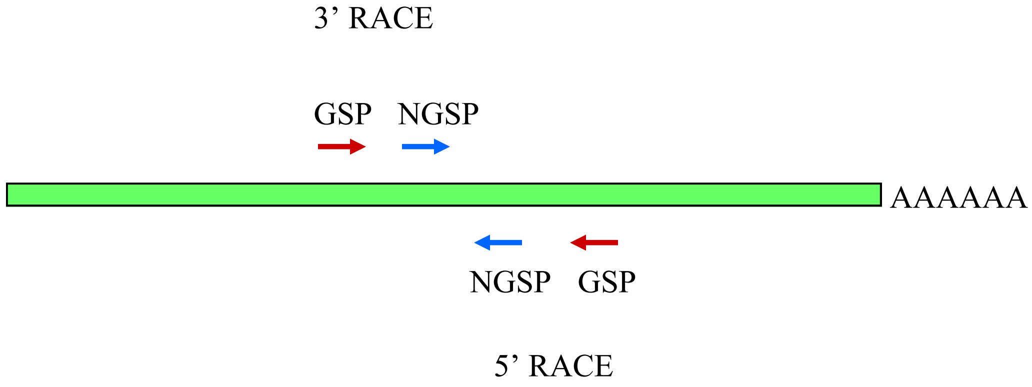


Figure 23. General figure of 5' and 3' RACE primers. The mRNA is in green, with the gene specific primer (GSP) represented as a red arrow and the nested gene specific primer seen as a blue arrow.

RT-PCR

Reverse transcriptase polymerase chain reaction (RT-PCR) uses an enzyme found in retroviruses such as HIV to turn RNA into complementary DNA (cDNA). This mixed pool of cDNA can then be used as a template for PCR, which uses primers designed to amplify a portion of DNA of interest (fig. 24).

Synthesizing cDNA

I made first strand cDNA following the protocol described in Invitrogen's SuperScript First-Strand Synthesis System for RT-PCR. For both males and females, 2.2ng of RNA were used. As in RACE-PCR, the J2.1 and K1 RNA preparations served as the template. All incubations not on ice were done using the Perkin Elmer GeneAmp 9700 thermocycler. Once made, cDNA was stored at -20°C until use. I could then use one batch of first strand cDNA for a number of RT-PCR experiments, with the cDNA thawed and refrozen with each use. Whenever I made cDNA, I also made male and female controls that lacked reverse transcriptase. These were made as described in the protocol, following the exact same steps as done for making the cDNA except for omitting the reverse transcriptase. Because an RNAse was used as part of the cDNA preparation, the resulting control pools contained only genomic DNA and served as a negative control for the PCR experiments, with all resulting bands presumably arising from priming off genomic DNA.

PCR

For PCR, I used Platinum *Taq* DNA Polymerase, a hot-start enzyme from Invitrogen, unless otherwise noted. I used the protocol described in the SuperScript First-Strand Synthesis System for RT-PCR manual, except I reduced all reactions to 2/5th of their original volume, or 20µl. All amplifications were done using the Perkin Elmer GeneAmp 9700 thermocycler. Optimal cycling conditions were determined empirically, with the following protocol providing consistently high amounts of product with minimal background using a wide range of primers:

94ºC 2 min

35 cycles

- 94ºC 30 sec
- 68ºC 3 min

68ºC 3 min

8ºC hold

Any time that I used a different protocol, these variations are noted. Primers were designed using the Primer3 program. This program, along with Sequence Extractor, was used to determine predicted products of both mRNA and genomic DNA. All products were visualized using the gel running protocol described elsewhere in this section.

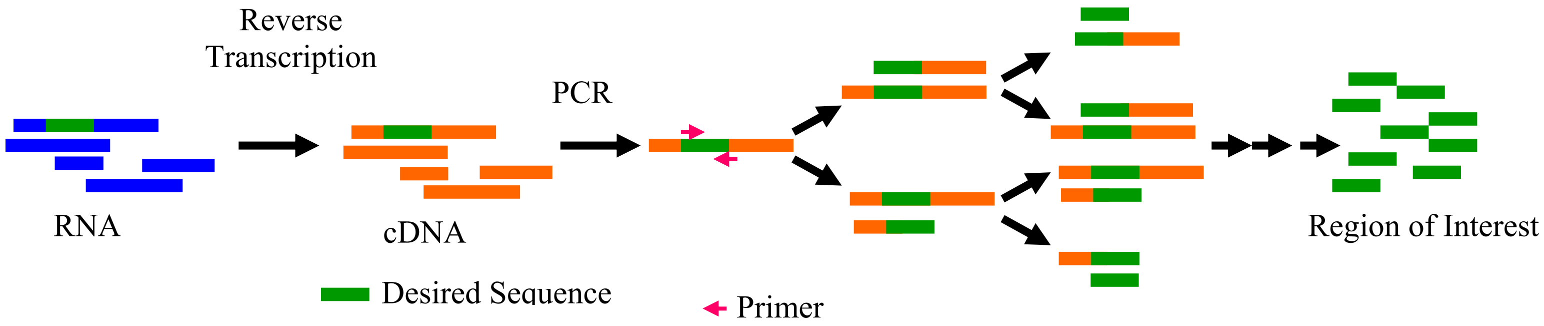


Figure 24. Flow diagram of RT-PCR beginning with the RNA, represented by blue bars. The RNA is converted into cDNA using reverse transcriptase, shown as orange bars. The portion of the transcript to be amplified is in green and the primers appear as pink arrows.

Cloning

All cloning was carried out using pCR®8/GW/TOPO® TA Cloning® Kit from Invitrogen and following the directions in the product manual. I cloned sequences from both RACE PCR and RT-PCR reactions. First, products to be cloned were ligated with vector DNA. To do this, I added 0.5µL of *taq* to each reaction and incubated the reactions at 72°C for 10 minutes in the Perkin Elmer GeneAmp 9700 thermocycler (fig. 25). The *taq* added a non-template adenine to the 3' end of the PCR product, enabling it to base pair with the vector, which had a thymine overhang. Recommended values were reduced by half to 3µl reactions. The amount of PCR product added to the reaction varied from 1-2µl depending on the amount of product as seen previously on a gel. Vector reactions were then kept at -20°C until the next step.

I then transformed One Shot® Competent *E. coli* following the protocol described in the manual, although the volumes of reactions were reduced by half. Competent cells were stored at -80°C and an aliquot was used only once to avoid damage from repeated freeze/thaw cycles. Once transformed, cells were shaken horizontally at approximately 200rpm in the Lab-Line Orbit Environ Shaker at 37°C for 45 minutes to an hour in order to allow the transformed *E. coli* sufficient time to recover and express ampicillin resistance. I then plated the transformed cells onto L agarose plates with ampicillin (100µg/mL). Each product was plated on two plates with different amounts on each, typically between 5µl and 30µl. These plates were grown up overnight at 37°C. Single colonies were dotted on another agar plate in a grid pattern with 50 single colonies to a plate unless otherwise noted.

I then did PCR with plasmid-specific primers (T7 and M13R) to determine the size of the inserts in a particular clone. For the primer-size checks, I typically used BioLabs Taq DNA Polymerase. I used the following cycling protocol on the Perkin Elmer GeneAmp 9700 thermocycler for amplification:

94ºC 8 min

25 cycles

- 94ºC 15 sec
- 45ºC 30 sec
- 72°C 2 min

72ºC 5 min

8ºC hold

From each grid of clones, I checked the size of the inserts in 10 to 20 colonies. Products were run out on an agarose gel and, from these, sizes of interest for sequencing were chosen.

Having determined which clones contained products I wanted to sequence, I then streaked these for single colonies using the *E. coli* colony from the grid plate. While gridding was not always done on plates with ampicillin, streaking for single colonies was. These plates then grew up overnight at 37°C. Using a sterile Fisher bacterial loop, I inoculated 4µl of LB broth that contained ampicillin (100µg/mL) which had been put in sterile 20mL centrifuge tubes. These tubes were then placed on their side and put in the shaking incubator set at 37°C and about 200rpm overnight. The resulting turbid broth would contain many, many *E. coli*, all containing the identical plasmid of interest.

Plasmid purification

To purify the plasmids, I used the QIAprep Spin Miniprep Kit by Qiagen. I collected *E. coli* cells by placing 1.3mL of the culture into a 1.5mL Eppendorf tube, centrifuging at 14,000rpm for 3min and then repeating with another 1.3mL of culture in the same tube. The rest of the purification was carried out as described in the QIAprep Miniprep Handbook. Plasmids eluted from the column using molecular grade water, which is preferred by Nevada Genomics for sequencing.

The concentration of each plasmid was determined by loading ½ to 1µl on an agarose gel and comparing the band intensity after staining with known amounts of plasmid that had been quantified using a Nanodrop that measures A260 absorbance. In preparation for sequencing, DNA of the proper quantity was dried using the Savant Speed Vac Concentrator.

Sequencing

All sequencing was done by the Nevada Genomics Center based at the University of Nevada, Reno. Sequencing at Nevada Genomics is carried out using the ABI BigDye Terminator Cycle Sequencing Ready Reaction Kit v3.1. Reactions are then run on the ABI3730 DNA Analyzer.

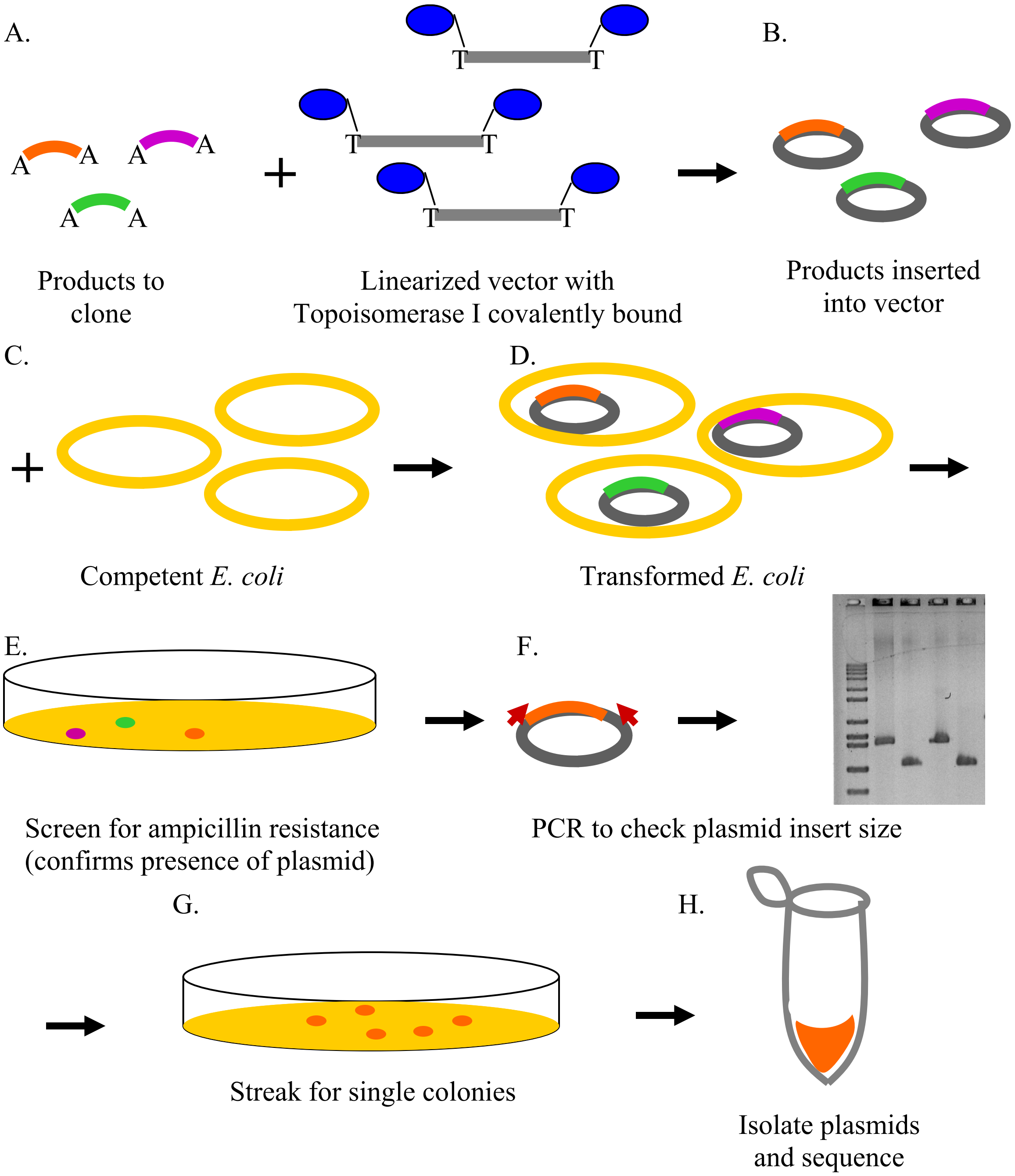

A.
B.
A
A
A
A
A
A
T
T
T
T
T
T
+
Products to clone
Linearized vector with Topoisomerase I covalently bound
Products inserted into vector
C.
D.
+
Competent *E. coli*
Transformed *E. coli*
E.
F.
Screen for ampicillin resistance (confirms presence of plasmid)
PCR to check plasmid insert size
G.
H.
Streak for single colonies
Isolate plasmids and sequence

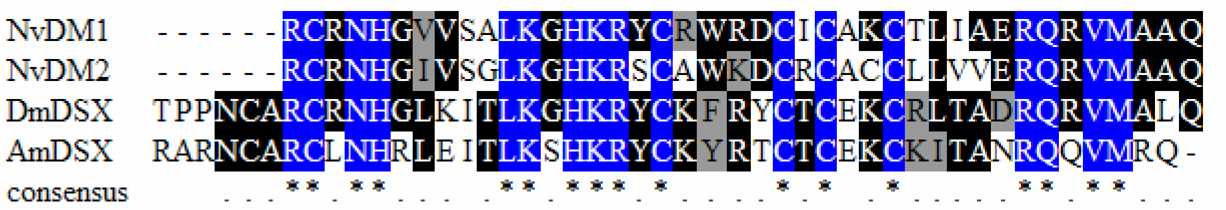
Figure 25. Overview of the cloning process. (A, B) Products and vectors are combined and the products inserted with the help of the topoisomerase enzyme. (C, D) These vectors are put into competent *E. coli* cells and then (E) screened for ampicillin resistance, which is present only in bacteria that have taken up a plasmid insert. After gridding out separate colonies (not shown), 10 to 20 colonies are chosen and (F) PCR performed to check the size of the product inserts using M13R and T7 primers (shown as red arrows) that amplify off the plasmid. (G) Clones with the appropriately sized inserts are then streaked for single colonies using the colony off the grid plate. (H) A single colony is then grown up overnight in a broth and plasmids from the resulting clones are isolated and sequenced.

COMPLETION AND CHARACTERIZATION OF NVDM1 AND NVDM2

Using PCR-based techniques, Andrea Llewellyn in the Trent lab had previously identified two genes in *Nasonia vitripennis* that include a DM domain, naming them NvDM1 and NvDM2 (fig. 26). Typical of DM-containing genes, NvDM1 and NvDM2 show a high degree of homology within the DM domain, but not outside of this motif. All of the sequences were obtained from 3' RACE PCR that used primers complementary to the DM domain to amplify the portion of the transcripts between the DM domain and the 3' end. Attempts to obtain the 5' end of the transcripts via 5' RACE PCR had proved unsuccessful. Therefore, the sequences did not include the 5' ends, including part of the DM domain. Comparing the original sequence of DM1 and DM2 to the sequence of the DM domain defined by Pfam, a domain-finding program that uses Hidden-Markov Models to identify protein motifs, along with the doublesex genes from *D. melanogaster* and *A. mellifera*, one can see that the both DM1 and DM2 are missing the N-terminal of their DM domains. Therefore, one of my goals was to complete the DM domain for both NvDM1 and NvDM2. In addition, I wanted to locate these two genes in the genome using newly available tools of the *N. vitripennis* Genome Project.

As discussed in the introduction, *doublesex* in *D. melanogaster* achieves sex-specificity through alternative splicing, making any evidence of such patterns of particular interest. The Trent lab found that NvDM2 was constitutively spliced, creating only one mRNA transcript in yellow pupae. In contrast, NvDM1 displayed considerable alternative splicing, and the lab documented four different transcripts in yellow pupae (Fig. 27). Of the NvDM1 transcripts, the longest open reading frame (ORF) occurs in transcript 3.4 with 415 amino acids. Despite being the longest transcript, 3.46 has an ORF of only 75 amino acids, slightly less than 3.42 with an ORF of 83 amino acids, but more than 2.9 that has only 59 amino acids in its ORF.

Although no sex-specificity in NvDM1 and NvDM2 transcription was seen in 3' RACE PCR experiments by previous members of the Trent lab, I chose to check this independently using RT-PCR with primers designed to span specific exon splice sites to check for products in both male and female yellow pupae. The results of this experiment could serve two purposes. First, it would be an independent affirmation of the complex splicing pattern found via RACE PCR and, second, it would use a different technique to check for sex-specificity of the transcripts.



```
NvDM1      ------RCRNHGVVSALKGHKRYCRWRDCICAKCTLIAERQRVMAAQ
NvDM2      ------RCRNHGIVSGLKGHKRSCAWKDCRCACCLLVVERQRVMAAQ
DmDSX      TPPNCARCRNHGLKITLKGHKRYCKFRYCTCEKCRLTADRQRVMALQ
AmDSX      RARNCARCLNHRLEITLKSHKRYCKYRTCTCEKCKITANRQQVMRQ-
consensus     ...**.**... .**.***.*....*.* .*.....**.**...
```

Figure 26. Alignment of the DM domain of NvDM1, NvDM2, and *DSX* from *Apis mellifera* and *Drosophila melanogaster*. The blue background indicates identical amino acids in all four sequences, the black background indicates identical amino acids in two to three sequences, and gray indicates similar amino acids.

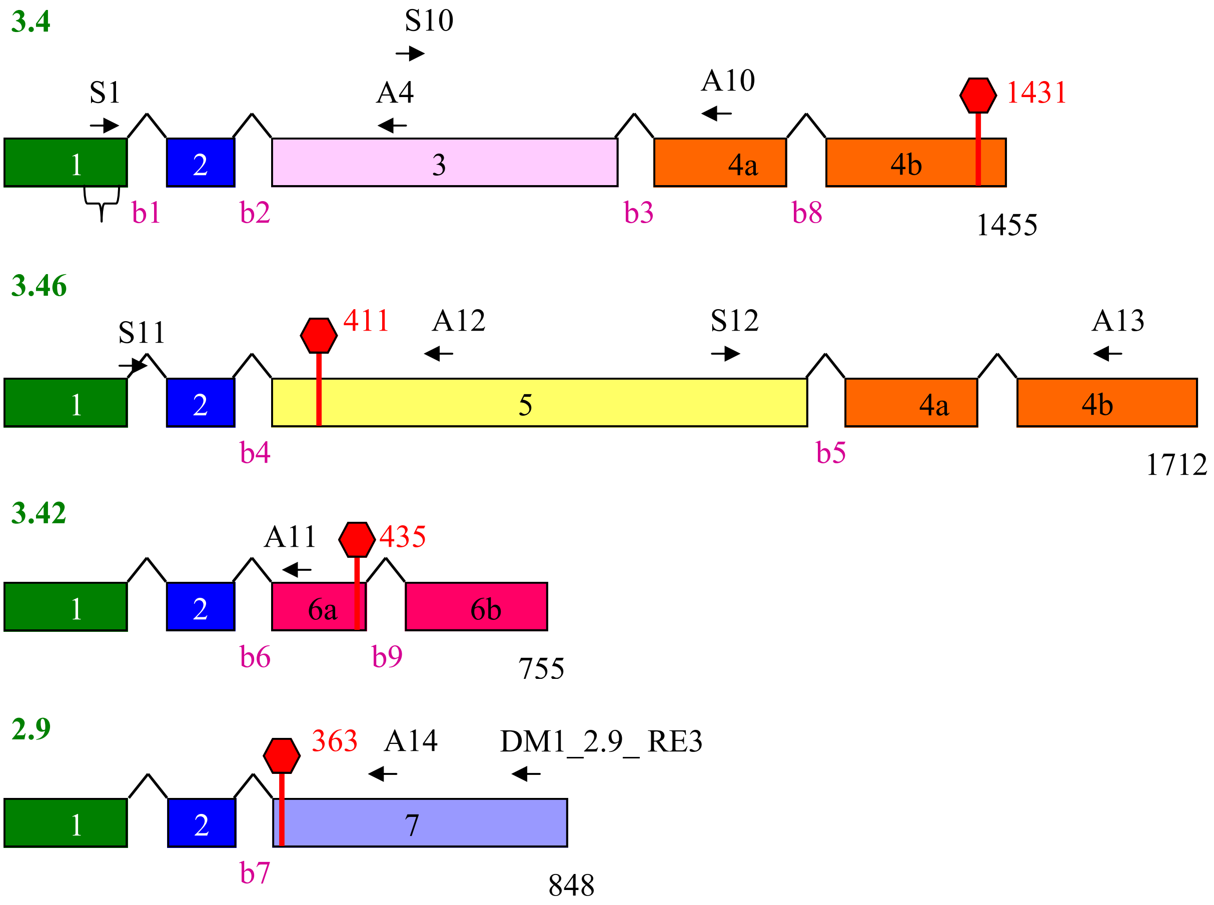


Figure 27. NvDM1 alternative transcripts. Green numbers indicate transcript name. Each colored box represents a unique sequence of mRNA. The DM domain is indicated by a bracket in exon one of transcript 3.4 and is present in all of the transcripts. The stop sign indicates the end of the longest open reading frame for each transcript, with red numbers to right indicating the length of the longest open reading frame. Arrows indicate primer location and direction, with primer names indicated above. The junctions between exons are numbered in purple. Introns are not to scale. Numbers inside exons are arbitrary and for identification purposes only, not to be considered an indication of genomic placement. Please note that this figure includes information gleaned from my work described in this section, including the additional 5' end sequence added to exon 1 and the splitting of exon 4 and 6 into two separate, constitutively spliced, exons.

Locating NvDM1 and NvDM2 in the genome

A great deal about the structure of DM1 and DM2 could be gleaned from the *N. vitripennis* Genome. The database includes a number of different search programs based on the basic BLAST algorithm (see Methods). I used one of the search programs, called megaBLAST, set to default parameters to search version 1.0 of the genome using the nucleotide sequence of DM1 and DM2 that had previously been sequenced by the Trent lab.

From this, I discovered that portion of the NvDM2 sequenced by the Trent lab is located on scaffold 62 and consists of only two exons, which previous research indicated are constitutively spliced. The two exons are separated by an intron 63 nucleotides in length. Previous research relied on alternative splicing to establish the presence of separate exons and so had been unable to establish the number of exons that are constitutively spliced.

In contrast, DM1 is not located on a single scaffold, but instead spans at least six (table 4). Both sequences 3.4 and 3.46 are on at least 4 different scaffolds. In both sequences, the 3rd exon is on scaffold 413. These two different exons on scaffold 413 are separated by an intron of 43,570 base pairs. Exon 5 is slightly problematic as the Blast search provides two different possibilities for its genomic location. The entirety of exon 5 aligns with 100% identity with a 747 nucleotide long stretch of scaffold 413. The same sequence, in two separate portions, align also with scaffolds 313 (E value: 8e-138, 97% identity) and 287 (E value: 6e-139, 87% identity). While scaffold 413 appears to be the better match, the other should not be ignored. Genomic information suggests that exon 4, which 3.4 and 3.46 share, is not one exon, but two that are separated by an intron of 155,076 base pairs. Similar to exon 4, while molecular work indicated exon 6 in transcript 3.42 could be one exon, genomic sequence shows that it is two exons separated by a 76 nucleotide intron. Also noteworthy, exon 2, which appears in all four transcripts, is on the same scaffold as exon 6 of transcript 3.42. There is a 21806 base pair intron between exons 6a and 2. Transcript 2.9 is located on three scaffolds, with exon 7 being a single exon as predicted molecularly.

Table 4. Scaffold locations of various NvDM1 exons. Exon number refers to those given in figure 27. Scaffold numbers are those assigned by NCBI. An * means that what was thought to be a single exon looking at cDNA transcripts is actually two exons that were constitutively spliced in the transcripts sequenced.

Exon Number	Transcript	Scaffold Number	Relative Location to Other Exon
1	All	68	–
2	All	168	21807bp from exon 6a
3	3.4	413	43571 from exon 5
4(a and b)	3.4, 3.46	138*	155075 between 4a and 4b
5	3.46	413 or 287 and 313	43571 from exon 3
6(a and b)	3.42	168*	76bp between 6a and 6b 21806bp between 6a and 2
7	2.9	583	–

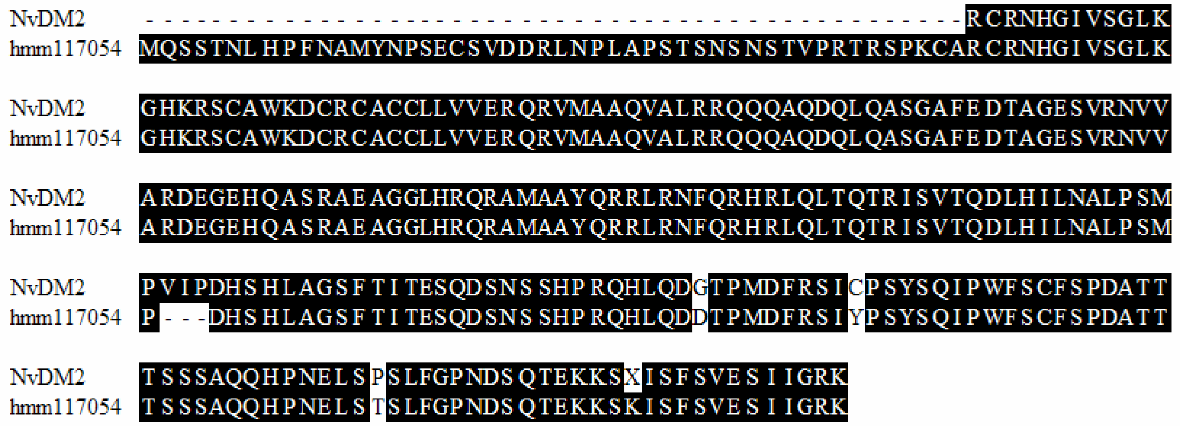
Completing the DM domain of NvDM1 and NvDM2

When I started work on this project, the sequence of the N-terminal regions of the DM domain of DM1 and DM2 had not been established. I first looked for these genes in the set of predicted proteins in both the *ab initio* and RefSeq datasets using BLASTX which compares a nucleotide sequence to a protein database. Neither was present in the RefSeq database. However, DM2 did have a match in the *ab initio* database called hmm117054. The match was highly significant with an e-value of 2e-112. Because of the 96% identity between the two sequences, I concluded this was the computational prediction of DM2 (differences to be addressed in the discussion) (fig. 28).

The first amino acid of DM2 aligns with the 49th amino acid of the computational prediction. Thus, the prediction can provide more information about the N-terminal of the DM2 protein. Comparing the new sequence with the model defined by Pfam, I found that this new sequence completed DM domain (fig. 29).

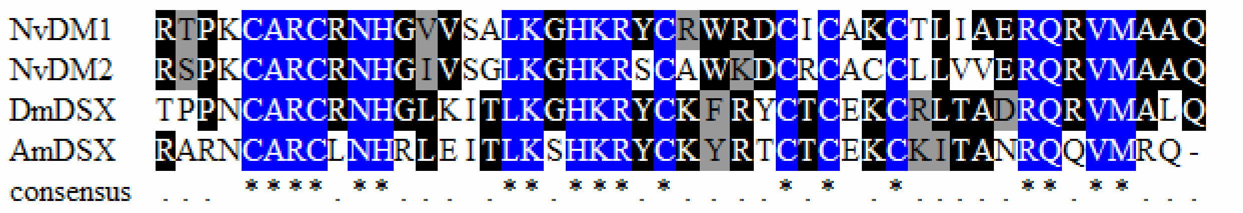
However, different tactics were needed to determine the complete DM domain of DM1 because it was not present in either of the protein databases. Instead, I returned to the genome database. The known region of the DM domain of NvDM1 is on scaffold 68. If the DM domain were not broken up by an intron, it should be possible to complete the DM domain by looking at the 5' end of the aligned sequence. I performed a search of version 1.0 of the *Nasonia* Genome using the BLASTn tool at Baylor College of Medicine's Human Genome Sequencing Center with the DM1 nucleotide sequence as my query. I then used the provided link to Geneboree to fetch sequence to the 5' end of where my sequence aligned with scaffold 68. I took that nucleotide sequence and translated it. Selecting the sequence that was in the same reading frame as the rest of the DM domain on scaffold 68, I found there was an open reading frame of 33 amino acids in addition to the previously determined sequence. Had there been an intron, I would not have expected an open reading frame of any significant length. Furthermore, comparing this additional sequence to the DM domain model using Pfam, I found that the new sequence completed the DM domain of NvDM1 (fig. 29).

It is important to note that, while I was able to complete the DM domain and some additional sequence, I did not identify the 5' end of the gene for any of the transcripts of NvDM1 or NvDM2. While the genes remain incomplete, these added sequences likely contain the N-terminal of the protein. In *A. mellifera*, the closest relative of *N. vitripennis* in which a doublesex ortholog has been documented, the DM domain is in the first coding exon of the gene (Cho *et al*. 2007). Although there is an exon to the 5' end of the one containing the DM domain, it is a non-coding exon. The same is true for the canonical *dsx* in *D. melanogaster* (Zhu *et al*. 2000).



```
NvDM2      ------------------------------------------------RCRNHGIVSGLK
hmm117054  MQSSTNLHPFNAMYNPSECSVDDRLNPLAPSTSNSNSTVPRTRSPKCARCRNHGIVSGLK

NvDM2      GHKRSCAWKDCRCACCLLVVERQRVMAAQVALRRQQQAQDQLQASGAFEDTAGESVRNVV
hmm117054  GHKRSCAWKDCRCACCLLVVERQRVMAAQVALRRQQQAQDQLQASGAFEDTAGESVRNVV

NvDM2      ARDEGEHQASRAEAGGLHRQRAMAAYQRRLRNFQRHRLQLTQTRISVTQDLHILNALPSM
hmm117054  ARDEGEHQASRAEAGGLHRQRAMAAYQRRLRNFQRHRLQLTQTRISVTQDLHILNALPSM

NvDM2      PVIPDHSHLAGSFTITESQDSNSSHPRQHLQDGTPMDFRSICPSYSQIPWFSCFSPDATT
hmm117054  P---DHSHLAGSFTITESQDSNSSHPRQHLQDDTPMDFRSIYPSYSQIPWFSCFSPDATT

NvDM2      TSSSAQQHPNELSPSLFGPNDSQTEKKSXISFSVESIIGRK
hmm117054  TSSSAQQHPNELSTSLFGPNDSQTEKKSKISFSVESIIGRK
```

Figure 28. Alignment of the translated amino acid sequence of DM2 with the predicted gene model hmm117054 found by running a BLASTX with the nucleotide sequence of DM2 as the query against the *ab initio* database.



```
NvDM1      RTPKCARCRNHGVVSALKGHKRYCRWRDCICAKCTLIAERQRVMAAQ
NvDM2      RSPKCARCRNHGIVSGLKGHKRSCAWKDCRCACCLLVVERQRVMAAQ
DmDSX      TPPNCARCRNHGLKITLKGHKRYCKFRYCTCEKCRLTADRQRVMALQ
AmDSX      RARNCARCLNHRLEITLKSHKRYCKYRTCTCEKCKITANRQQVMRQ-
consensus  ... ****.** ....**.***.*....*.* .*.....**.**...
```

Figure 29. Amino acid alignment of complete DM domains from *N. vitripennis* DM1 and DM2 with those of *D. melanogaster* and *A. mellifera*. Blue background indicates conservation in all four proteins, black indicates identical amino acids in two or three proteins, and a gray background means the amino acids are similar.

Searching for other domains in NvDM1 and NvDM2

In my analysis of DM1 and DM2, I wanted to address whether they contained any domains besides the DM domain. As described in the introduction, DM-containing genes that are associated with sex determination contain an oligomerization domain. I used Pfam to search for other domains in both genes, leaving the parameters on default for both.

The resulting search showed that DM2 contained only the DM domain, even when the threshold was raised to an E-value of 1.0. The E-value, or expect value, is the probability of finding the searched for sequence by chance in the database, taking into consideration the size of the sequence used to search and the size of the database. Thus, the higher the E-value, the more likely that any matches are the result of simple chance and do not have biological significance. The DM domain of DM2 has an E value of 4.8e-19.

For NvDM1, I searched Pfam using all four transcripts depicted in figure 27. For all but 3.4, the only significant domain found was the DM domain with an E value of 9.22e-30. In addition to the DM domain, transcript 3.4 has a second domain called DMA or DMRTA. Proteins containing this additional domain are called DMRTA proteins. This domain of unknown function is associated with the DM domain, occurring towards the C-terminal end in the DM-containing proteins in a variety of organisms. The DMRTA motif in transcript 3.4 has an E value of 1.7e-14 (fig. 30).

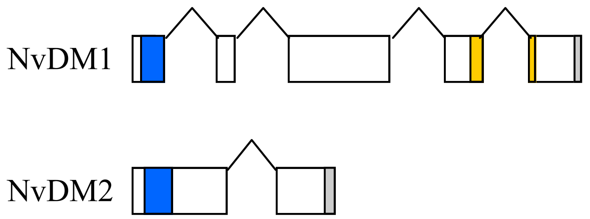


Figure 30. The exon structure of NvDM1 and NvDM2. The NvDM1 transcript shown is of the longest transcript with the longest open reading frame (ORF), number 3.4. The blue box indicates the sequence that codes for the DM domain while the yellow boxes indicates the region that codes for the DMRTA motif. White boxes are ORF and gray boxes indicate 3' untranslated region (UTR). Neither sequence includes 5' UTR information as this could not be determined using available computational methods.

Confirming alternative splicing with RT-PCR

The Trent lab had established the presence of four different mRNA transcripts all coded for by the NvDM1 gene. Due to the alternative splicing seen in *doublesex*, this presented the possibility of sex-specific expression. However, the Trent lab's previous work using 3' RACE PCR had shown no such specificity in yellow pupae. In order to test this, I designed primers to span various exon boundaries (fig. 27) and then did RT-PCR to check for products of the expected sizes following protocol described in Methods. A complete list of primer sequences is located in the appendix. In each run, I did PCR on a single primer pair with separate male and female cDNA templates, including no reverse transcriptase controls for each. When running the products out on a gel, a space was left between male and female reactions in order to prevent contamination due product spilling over into the next lane.

Results of RT-PCR confirmed exon boundaries one through seven (fig. 27) with products of the expected size seen in both male and female pools. Gels from this experiment can be found in the appendix. In some cases, such as with the primers S11 and A12, bands in addition to the expected size were present (Appendix B). However, in all primer pairs checking boundaries one through eight, these products were not sex-specific; the extra bands appeared to be the same in both sexes. Therefore, even if these are actual transcripts and not artifacts, I did not consider them worthy of pursuit. The splice junction between exons two and seven proved problematic (fig. 27). With my first set of primers, I ran the reaction a number of times and had a different result in each. I then decided to design a new right primer and repeated the RT-PCR, but this time had no amplification. I decided not to pursue this further because of the very short ORF provided by transcript 2.9, that only one with this b9 exon junction. Because the ORF codes for only 59 amino acids, it is not likely to result in a functional protein.

Conclusion

While both NvDM1 and NvDM2 are members of the broader DM gene family, neither appears to be a functional ortholog of *doublesex*. Two significant lines of evidence support this conclusion. First, no sex-specific expression was seen in either gene. Although NvDM1 did undergo significant alternative splicing, none of the splice forms examined showed sex-specific expression in yellow pupae. Second, neither NvDM1 nor NvDM2 contain the oligomerization domain that is critical for the recruitment of other proteins that lead to sex-specific expression in *D. melanogaster*. A more detailed analysis of these findings appears in the Discussion.

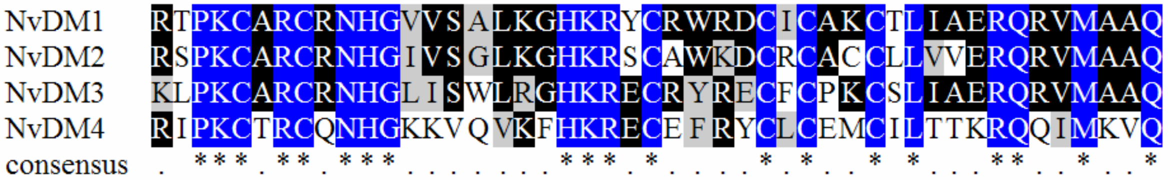
DISCOVERY OF NEW GENES: NVDM3 AND NVDM4

While NvDM1 and NvDM2 did not display sex-specific transcription, this did not preclude another, undiscovered DM-containing gene from acting in a sex-specific manner. Precedent suggested that *N. vitripennis* was likely to contain more members of the DM gene family than the two previously discovered by the Trent lab. The genome of *C. elegans*, for example, has eleven different DM-containing genes and *D. melanogaster* has four (Hodgkin 2002). Another member of the order Hymenoptera, *A. mellifera*, contains at least four genes in the DM family (Cristino 2006). Furthermore, personal correspondence with John Werren suggested the presence of a third DM-containing gene discovered by his lab.

A computational approach to gene discovery

The search for a *doublesex* homolog in *N. vitripennis* began well before the initiation of the genome project. The availability of the sequenced genome enabled me to approach this problem of gene discovery from a new direction. In addition to a the sequenced genome, the genome project also included two sets of predicted genes, the Reference Sequence (RefSeq) database and the *ab initio* database, both of which were predicted using a program called Gnomon (for further details, see Methods). The RefSeq database contains genes that have homologs in curated proteins, while the *ab initio* database includes predicted genes that lack homology or expressed sequence tag (EST) support. Both of gene sets provide predictions only of the open reading frame (ORF) and so do not include either 5' or 3' untranslated regions (UTRs). I used all three databases to search for additional members of the DM family.

Because the DM-containing proteins are poorly conserved outside of the DM domain, I searched the RefSeq and *ab initio* databases using just the amino acid sequence of the DM domain from NvDM1. Using the default parameters for both, I ran a BLASTP search which compares protein sequences. The RefSeq database provided no matches, even when the E-value was raised to 1.0. In contrast, the search of the *ab initio* proteins resulted in three hits: hmm117054, hmm336964, and hmm284494. These hits, relative to the DM1 sequence, had E-values of 3e-18, 1e-17, and 1e-09 respectively. The first hit, hmm117054, corresponded to NvDM2. The two other predicted proteins, hmm336964 and hmm284494, had not been previously identified and represent two additional genes in the DM gene family. I named them NvDM3 (hmm336964) and NvDM4 (hmm284494). An alignment of all four DM-containing genes shows that the two new genes have the highly conserved cysteines and histidines characteristic of the DM domain (fig. 31).



```
NvDM1      RTPKCARCRNHGVVSALKGHKRYCRWRDCICAKCTLIAERQRVMAAQ
NvDM2      RSPKCARCRNHGIVSGLKGHKRSCAWKDCRCACCLLVVERQRVMAAQ
NvDM3      KLPKCARCRNHGLISWLRGHKRECRYRECFCPKCSLIAERQRVMAAQ
NvDM4      RIPKCTRCQNHGKKVQVKFHKRECEFRYCLCEMCILTTKRQQIMKVQ
consensus  . ***.**.***.......***.*....*.*..* *...**..*..*
```

Figure 31. Alignment of the DM domains of the DM-containing genes in *Nasonia vitripennis*, including the two new genes NvDM3 and NvDM4.

CHARACTERIZATION OF NVDM3

The new gene NvDM3 was predicted to have an open reading frame 1053 nucleotides long, resulting in a protein 351 amino acids in length. It was predicted to consist of six exons, although the first contains a very short reading frame of only 5 nucleotides (fig. 35). The gene is located on scaffold 53. In order to learn more about this gene, I searched for domains in the amino acid sequence using Pfam. In addition to the DM domain, it also contained a DMRTA motif, as seen in NvDM1.

The computational prediction is just that, a prediction. Thus, I needed to confirm my findings using molecular techniques. To do this, I designed primers to amplify across the different exon boundaries. The first two exons were too small to design useful primers and thus represent only a computational prediction that, as of yet, has not been confirmed by molecular work. However, all of the exon pairs checked did confirm the gnomon model, resulting in products of the expected sizes in both sexes (fig. 36).

As with DM1 and DM2, it appears that NvDM3 is not sex-specific on the level of the entire organism during the yellow pupa stage. It is possible, as discussed previously, that this gene could show sex-specificity on the tissue-level or at different stages in development.

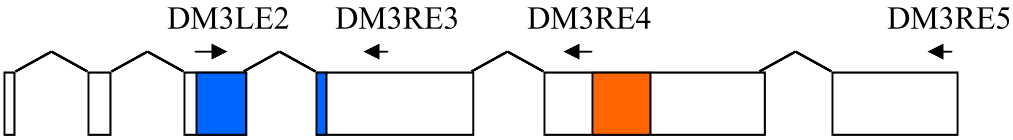


Figure 32. Exon structure of NvDM3 as predicted by Gnomon. Blue indicates DM domain and orange indicates the DMA (Dmrta) motif. Arrows are primer locations, with the names of the primers shown above. The first two exons were too small to design useful primers and thus represent only a computational prediction. All other exons were confirmed using RT-PCR of the primers shown. Because this is a prediction, there are no 3' or 5' UTRs. Thus all exons shown are coding exons with open reading frames. The last three nucleotides of exon six are a stop codon.

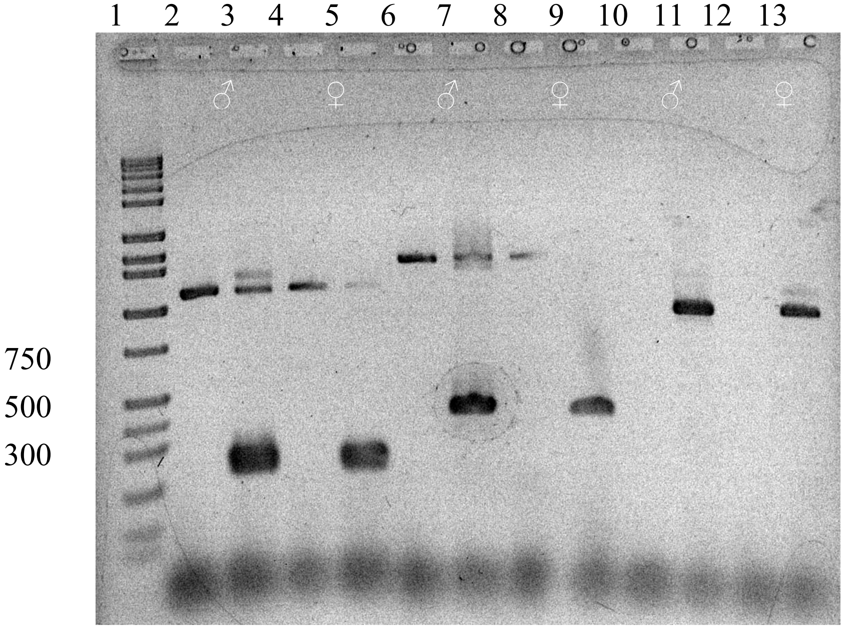


Figure 33. RT-PCR of NvDM3 to confirm prediction found in RefSeq. Note that the larger bands in lanes 2 through 8 correspond with genomic product size. (2 and 4) Primers LE2 and RE3 in males and females with no reverse transcriptase (-RT); (3 and 5) LE2 RE3 m and f (+RT); (6 and 8) LE2 RE4 m and f (-RT); (7 and 9) LE2 RE4 m and f (+RT); (10 and 12) LE2 RE5 m and f (-RT); (11 and 13) LE2 RE5 m and f (+RT).

COMPLETION AND ANALYSIS OF NVDM4

Of the four DM genes in *N. vitripennis*, DM4 contained a DM domain that looked the least like the others (fig. 31). While NvDM4 contained all of the conserved cysteines and histidines that characterize a DM domain, there were eight amino acids conserved in all three of the other *N. vitripennis* DM genes that were absent from NvDM4, only two of which were substitutions of similar amino acids. In contrast, there were only two amino acids conserved in NvDM1, NvDM2, and NvDM4 but not in NvDM3 and, in both cases, the substitution was for a similar amino acid. In addition, the predicted length of NvDM4 was markedly shorter that the rest with only two exons and an open reading frame of just 363 nucleotides that coded for a protein of 120 amino acids. I used molecular techniques to assess whether the prediction of NvDM4 was correct.

5' RACE of NvDM4

In order to obtain the beginning of the gene, I did 5' RACE on cDNA prepared from yellow pupae. For this experiment, I used separate pools of male and female RNA as the templates. In the first set of reactions, I used the gene-specific primer DM4RE1.2 (fig. 34). This primer was designed to sit outside the DM domain to avoid picking up other members of the DM gene family. Analysis of the products by agarose gel electrophoresis revealed a complex series of bands (not shown). To reduce artifacts, I did a second RACE reaction, using a nested gene-specific primer and the products from the first 5’ RACE reaction as the template (fig. 34). Once again, there was a complex pool of products (fig. 35).

I then cloned the products of the nested 5’ RACE reaction (see materials and methods) and checked the insert sizes of 20 male (fig. 36) and 20 female clones (fig. 37). Because there is no precedent for alternative splicing in the 5’ end of DM-containing genes, I chose to sequence just the longest transcripts, one from males and one from females. In support of this decision, there was no evidence of a consistently-sized product of shorter length that might have indicated an alternatively spliced form. Instead, sizes of smaller products varied and were likely artifacts, the result of the enzyme disassociating before the entire transcript had been copied, or the result of slightly degraded RNA transcripts.

Plasmids from the indicated clones were sequenced. To sequence my PCR products inserted in the plasmids, I used primers called M13R and T7 that were located in the plasmid on either side of the inserted sequence (fig. 38). The resulting sequence reads showed that male and female 5’ ends are identical. In both, the first exon does not contain an open reading frame and the second exon contains a short untranslated region, following by the start of the gene (fig. 39). The first methionine of the open reading frame in the second exon corresponds with the beginning of the *ab initio* prediction.

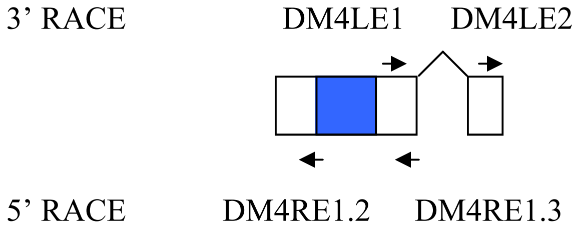


Figure 34. The *ab initio* prediction of NvDM4 transcript with primers used for RACE. Primers for 3’ RACE are on top of the transcript while those for 5’ RACE appear below. The blue box represents the DM domain. A stop codon appears at the end of the second exon; the rest of the sequence is an open reading frame.

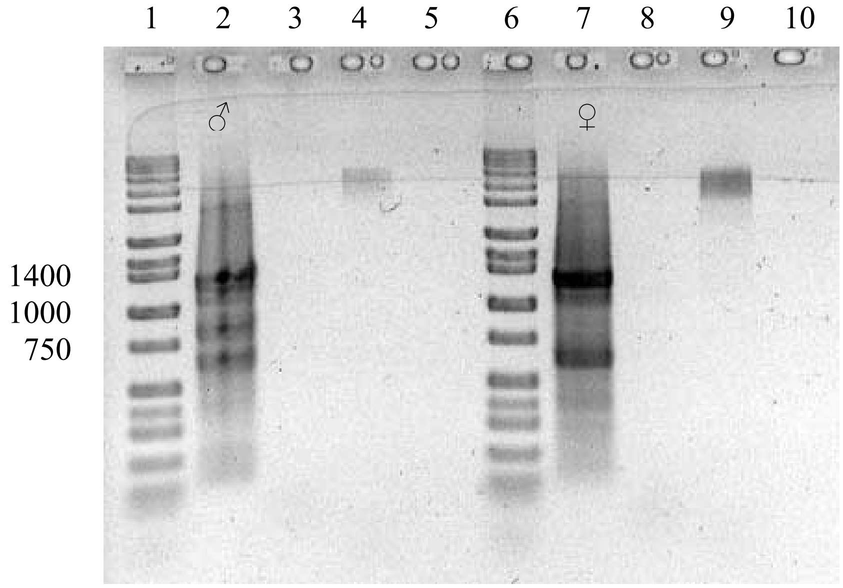


Figure 35. Nested priming of 5' RACE. (2 and 7) 5' RACE DM4RE1.2 (m, f); (3 and 8) DM4RE1.2 only (m, f); (4 and 9) Nested universal primer mix only (m, f); (5and 10) No primers (m, f).

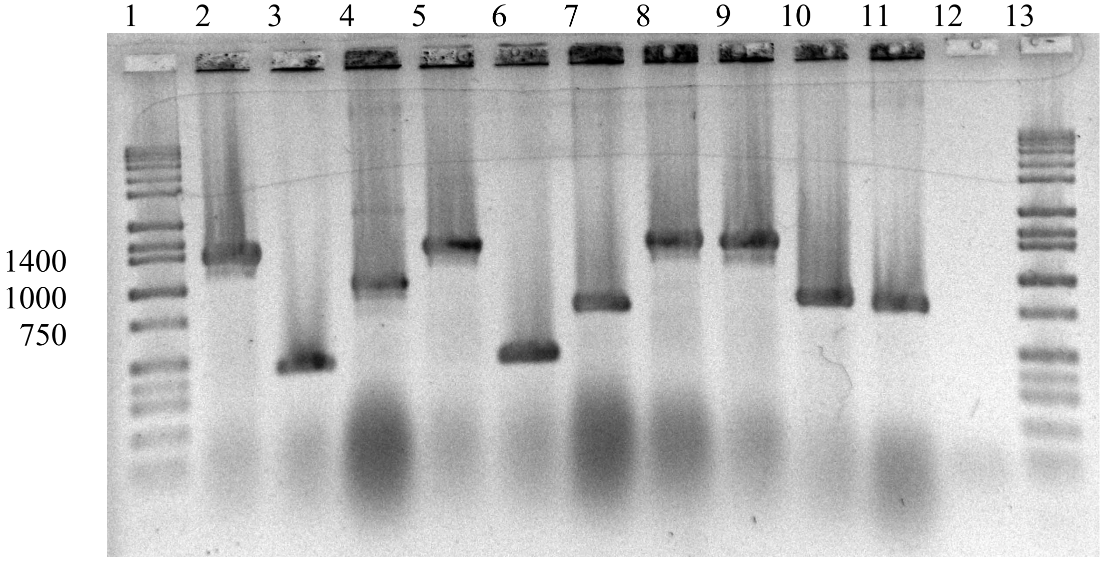


Figure 36. Check of the sizes of plasmids cloned from male products of 5' RACE. Lanes 2, 5, 8, and 9 have plasmids of approximately the same size. The plasmid in lane 8 was sequenced. Lanes 1 and 13 are Hi-Lo and lane 12 is a no *E. coli* control.

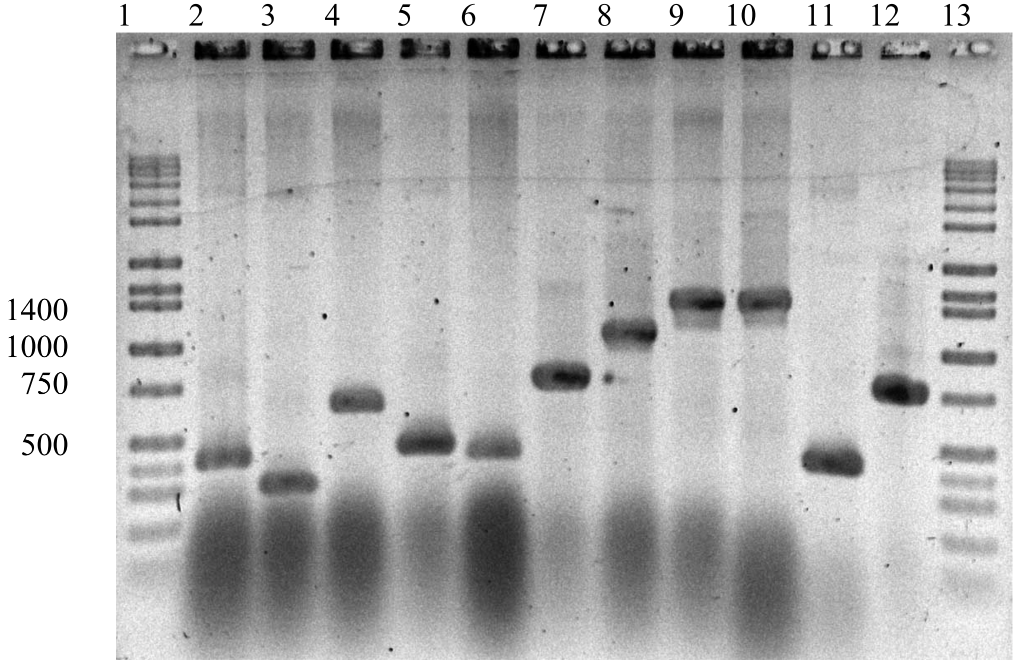


Figure 37. Check of the sizes of plasmids cloned from female products of 5' RACE. Lanes 9 and 10 have plasmids of approximately the same size. The plasmid in lane 9 was sequenced. Lanes 1 and 13 are Hi-Lo.

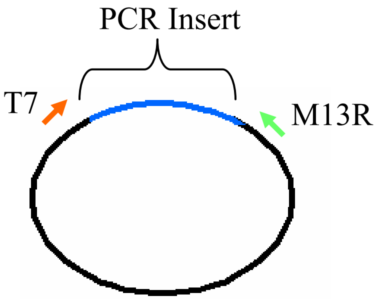


Figure 38. The location of the T7 (orange) and M13R (green) primers on either side of the PCR insert (blue) in the plasmid (black).

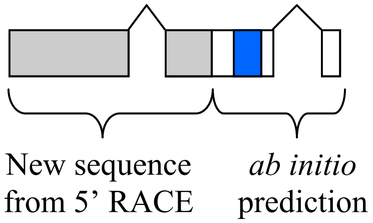


Figure 39. The results of 5' RACE of NvDM4. The untranslated region (UTR) is in gray, the open reading frame in white, and the DM domain in blue. No sex-specificity was expected in the 5' end of the gene.

3’ RACE of NvDM4

For the 3’ RACE, I designed both a gene-specific (GSP) and a nested gene specific primer (NGSP) (fig. 34). To avoid amplifying other members of DM gene family, I used primers that were positioned outside the DM domain. As with 5’ RACE, I used separate pools of male and female total RNA prepared from yellow pupae as template for the primary PCR. I analyzed the products of the primary RACE on an agarose gel and observed PCR products of different sizes in males and females (fig. 40). In an attempt to decrease the presence of artifacts, I then repeated the 3’ RACE using the NGSP and the products from the first 3’ RACE as templates. Again, once run out on a gel, the results showed a mixed pool of products with distinctly different bands in males and females (fig. 41). Both of these gels would suggest sex-specific splicing of NvDM4 at the 3’ end.

Because both males and females had a mixed pool of products, I cloned the nested products (see Methods). I checked the size of 10 clones in each sex using PCR and anaylzed these products as described previously (fig. 42 and 43).

The gels showed that there were a number of different sized clones in each sex (fig. 42 and 43). In some cases, there were two or more amplifications that resulted in products of the same size. I sequenced one clone of each size, excluding the very short male clone C12. From the males, I decided to sequence C2, C4, C5, and C10. Of the female clones, I sequenced C1, C2, C8, and C21. In the first round of sequencing, I was able to complete female clones C1 and C8 and male clones C2 and C4. For the remainder, I then designed internal primers and was able to complete the rest of the clones in the second round of sequencing.

I then aligned these sequences and the *ab initio* prediction with the genomic transcript using Spidey in order to determine the exon structure of the transcripts (fig. 44). Spidey is a program provided by NCBI that aligns genomic sequence with mRNA transcripts, enabling the user to examine exon size and location. Upon examination, I found that, although the second exon of the transcripts appears to be the same as the *ab initio* sequence, on all but the male 5 transcript, the exon actually ends four nucleotides earlier, eliminating the premature stop codon. Another important feature was that, at the 3' end of these transcripts, there was a sequence present in males but absent in females (see box in Fig. 44). I named this short sequence VI for variable intron. Within the VI was a stop codon. In the female transcripts, the stop codon appeared shortly after the spliced out VI. There were also significant differences in the length of the untranslated regions among the transcripts. While male transcripts 10 and 5 and female transcript 21 all end with a base pair or two of the other, the rest vary; these differences in length are not due to variation in the length of the poly-A tail. Also, although it cannot be seen on the resolution provided by Spidey, a closer look at the male clones 2 and 5 revealed that they both have 12 base pairs spliced out of the second exon. However, this is an in-frame deletion so it does not result in a frame-shift and the rest of the protein remains the same.

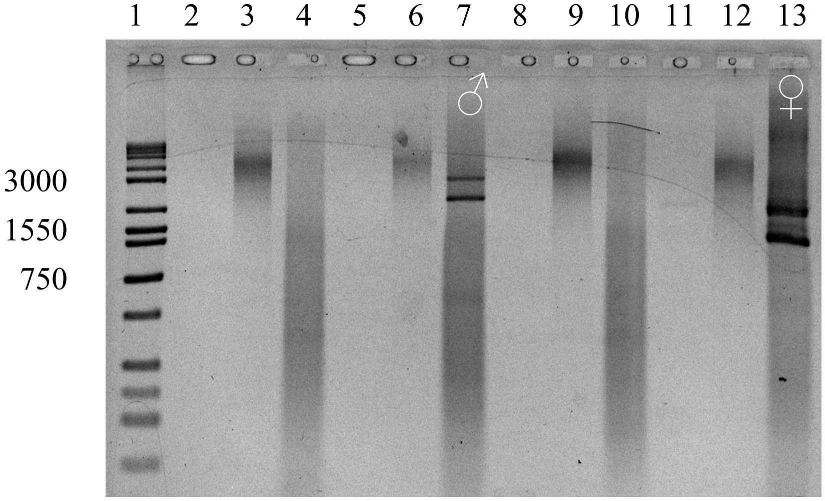


Figure 40. Results of 3’ RACE in males and females. Lane 7 is 3’ RACE products from male RNA and Lane 13 from female RNA. Note the differences in product sizes. The remaining lanes are as follows: (4 and 10) 3’ RACE DM3 male and female; (2 and 8) DM3-specific primer only m and f; (3, 6, 9, 12) Universal Primer Mix only m and f; (5 and 11) DM4-specific primer only m and f; (1) Hi-Lo.

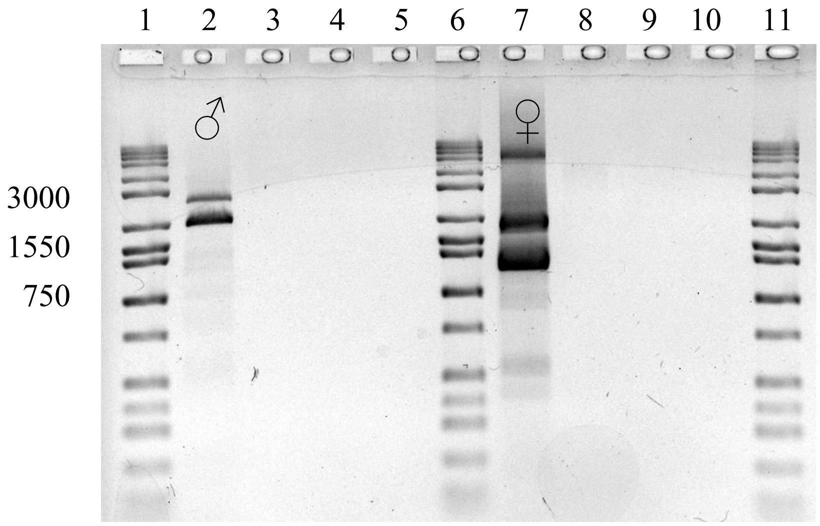


Figure 41. Results of nested 3’ RACE in males and females. Lane 2 is 3’ nested RACE products from males and lane 7 is from females. The remaining lanes are size standards and controls as follows: (3and8) 3’ RACE with Nested Universal Primer only, male and female; (4 and 9) 3’ RACE with DM4LE2 Primer only; (5and10) 3’ RACE without primers; (1, 6, and 11) Hi-Lo.

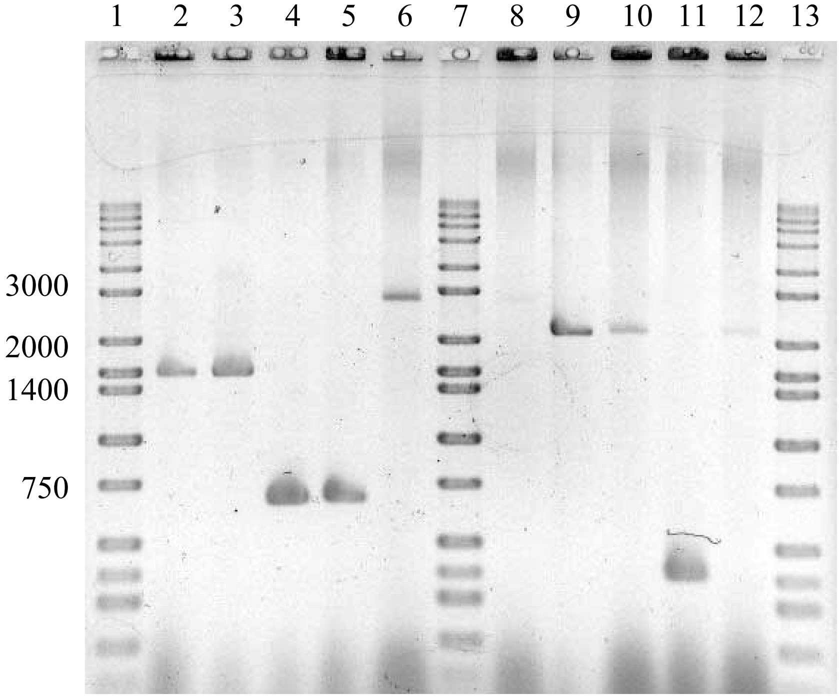


Figure 42. Results of checking the clone sizes in males. Each clone is given a particular number in order to keep track of it and these are indicated as follows with the lane number in parenthesis: (2) C1, (3) C2, (4) C3, (5) C4, (6) C5, (8) C6, (9) C10, (10) C11, (11) C12, (12) C13. One clone of each size was sequenced, except for C12, which was deemed too small. I sequenced clones C2, C4, C5, and C10. Lanes 1, 7, and 13 are Hi-Lo.

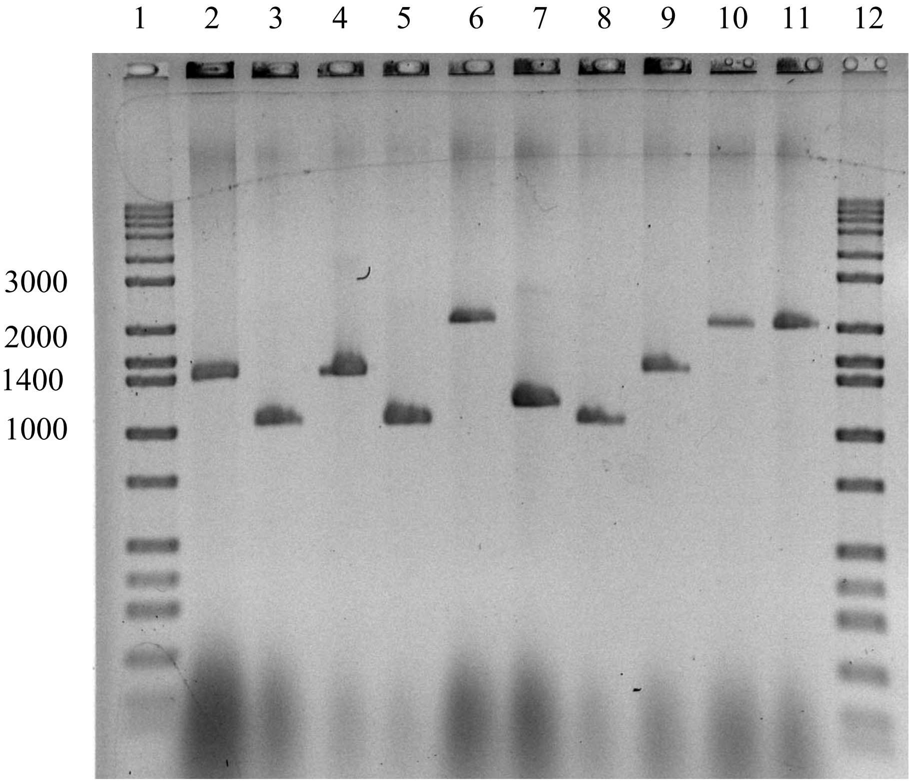


Figure 43. Results of checking the clone sizes in females. Clones: (2) C1, (3) C2, (4) C3, (5) C4, (6) C5, (7) C8, (8) C13, (9) C14, (10) C20, (11) C21. One clone of each size was sequenced. I sequenced clones C1, C2, C8, and C21. Lanes 1 and 12 are Hi-Lo.

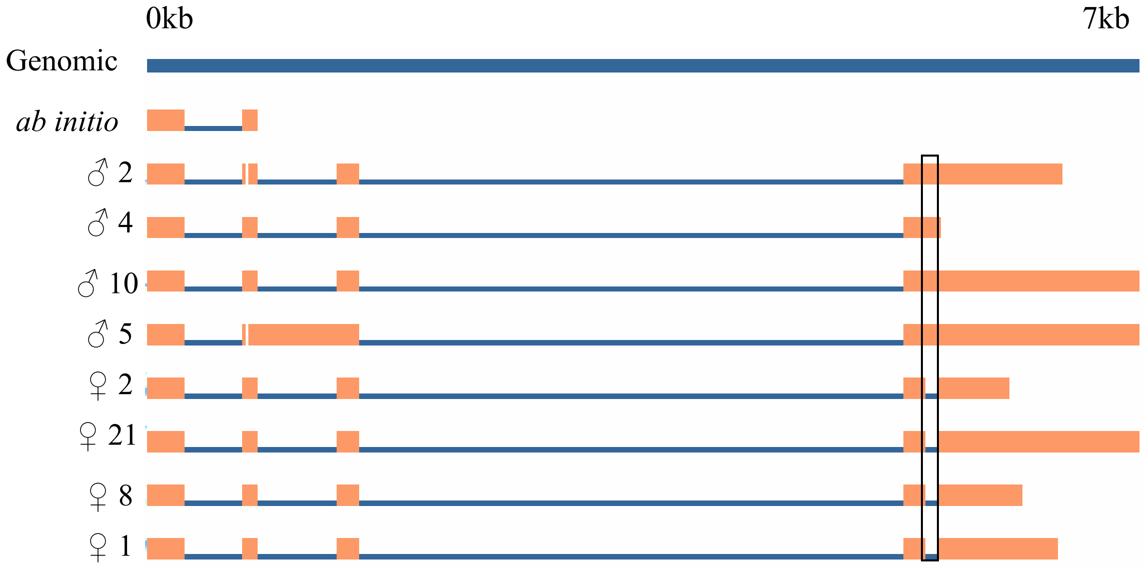


Figure 44. Alignment of male and female NvDM4 transcripts and the original ab initio prediction with genomic sequence. Exons appear in peach and introns in blue. There is a stop codon at the end of the ab initio sequence and at the same location in male clone 5, but not in the remaining sequences, which are spliced just before the stop codon. The black box indicates the VI, or variable intron, which was found excluded from female transcripts and included in male transcripts. A small (12bp) sequence spliced out of male transcripts 2 and 5 indicated by white line.

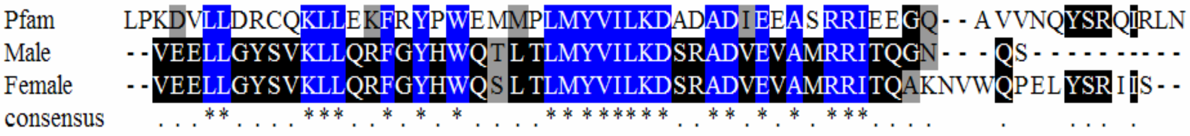
Complete transcripts

I combined information from the 5' and 3' RACE experiments to build complete mRNA transcripts of each clone (fig. 46). For this, I assumed that each transcript had the same 5' end. While I have not tested this assumption, there is no precedent for alternative splicing at the 5' end of *doublesex* as well as its orthologs.

Counting from the first methionine, the male transcripts translate to 219 or 223 amino acids (with or without 12bp) and the female transcripts translate to a protein 235 amino acids in length.

I then put these amino acid sequences into Pfam to look for functional domains. Both male and female transcripts showed two domains (fig. 46). The first was the DM domain, as seen in the *ab initio* prediction. The second was the doublesex oligomerization domain (OD). Unlike the DMA domain seen in NvDM1 and NvDM3, the OD is specifically associated with doublesex and its homologs involved in sex determination. Also, as seen in both *D. melanogaster* and *A. mellifera* along with a number of other doublesex homologs, males and females share the amino end of the OD, but differ in the carboxyl region (fig. 45). This difference contributes to the sex-specificity of the doublesex transcription factors, with the female-specific OD interacting with other proteins that in turn influence transcription (Bayrer *et al*. 2005). Compared to the Pfam Hidden Markov Model of the OD, the male sequence had an E value of 9.96e-6 and the female sequence had an E value of 2.9e-10.

I have categorized the transcripts into 3 classes: OD-A, OD-B and OD-X. Class OD-A refers to the transcripts that code for polypeptides containing the oligomerization domain originally found in males (table 5). Transcripts coding for the oligomerization domain originally found in females are classified as OD-B transcripts and the single transcript found in the males that lacks an oligomerization domain due to an early stop codon is classified as OD-X.



```
Pfam       LPKDVLLDRCQKLLEKFRYPWEMMPLMYVILKDADADIEEASRRIEEGQ--AVVNQYSRQIRLN
Male       --VEELLGYSVKLLQRFGYHWQTLTLMYVILKDSRADVEVAMRRITQGN---QS----------
Female     --VEELLGYSVKLLQRFGYHWQSLTLMYVILKDSRADVEVAMRRITQAKNVWQPELYSRIIS--
consensus    ...**....***..*.*.*....********..**.*.*.***....   .   ... .
```

Figure 45. An alignment of the doublesex oligomerization domains (OD) of the male and female *N. vitripennis* transcripts with the Pfam Hidden Markov Model for the OD. Residues conserved in all three sequences have a blue background, those that are the same in two sequences have a black background, and similar residues have a gray background. Notice that the *N. vitripennis* male and female proteins are different at the carboxyl end due to the alternative splicing, while the amino ends are identical.

Table 5. The classification of transcripts according to the oligomerization domain contained in their polypeptides. The e-value is calculated using the Pfam Hidden Markov Model.

Classification	Clones	OD E-value
OD-A	M2, M4, M10	9.96e-6
OD-B	F1, F2, F8, F21	2.9e-10
OD-X	M5	None

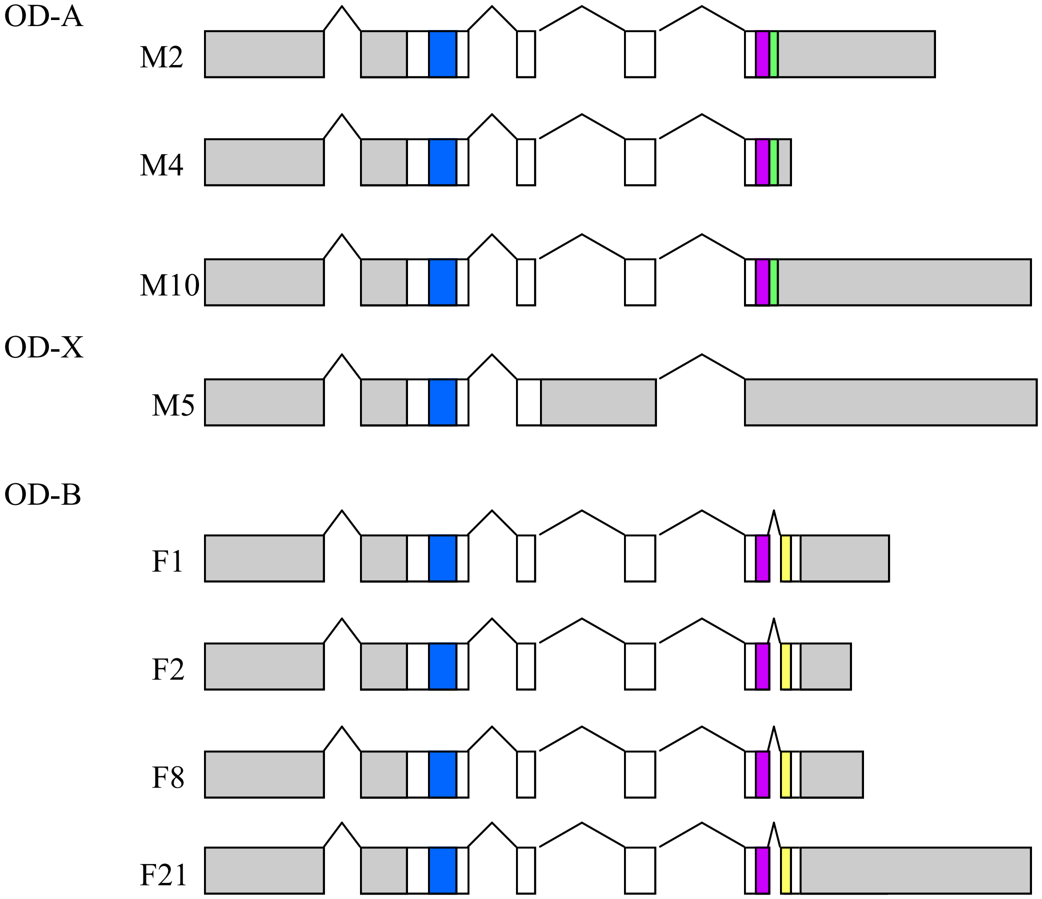


Figure 46. Complete mRNA transcripts of NvDM4 in males (M) and females (F). The untranslated regions are in gray, the sequence coding for DM domains in light blue, the sequence coding for the shared oligomerization domain is in purple, and the sequence of the oligomerization domain unique to OD-A and OD-B are in green and yellow, respectively. Note: Does not show small introns in M5 and M2.

Carboxyl terminal conundrum

One way in which the NvDM4 transcripts differed considerably from the *doublesex* gene its orthologs in other species including *A. mellifera* was the length of the sequence following the oligomerization domain (OD). The *N. vitripennis* male protein, OD-A, had no amino acids following the OD while the female protein, OD-B, had only one (fig. 47). In contrast, other *dsx* orthologs have a number of amino acids after the OD.

The shortness of the male transcript was of particular concern. In *D. melanogaster*, there are 137 amino acids following the oligomerization domain in the DSX-M. The length of this tail has been correlated with male-specific function. For example, research suggests that the long tail of DSX-M interferes with neighboring cis-regulatory sites in the regulation of the yolk protein gene, resulting in its repression, as discussed previously (An and Wensink 1995). With such a truncated C-terminal, OD-A would be unable to act in this way.

The short female transcript was less of a problem biologically. One protein with which DSX-F interacts is INTERSEX (IX). Only DSX-F associates with IX and this seems to be dependent on the female-specific region of the oligomerization domain, but not the C-terminus that follows (Yang *et al*. 2008). Mutagenesis work has shown that the entire female carboxyl tail following the oligomerization domain can be deleted without interfering with the interaction between IX and DSX-F (Yang *et al.* 2008). This would suggest that OD-B would also be able to interact with IX, despite the abbreviated tail. Thus, my main concern was the short male-specific oligomerization domain and C-terminal tail coded for by the OD-A protein.

>DmDSX-M
LGQDVFLDYCQKLLEKFRYPWELMPLMYVILKDADANIEEASRRIEEARVEINRTVAQIYYN
YYTPMALVNGAPMYLTYPSIEQGRYGAHFTHLPLTQICPPTPEPLALSRSPSSPSGPSAVHNQ
KPSRPGSSNGTVHSAASPTMVTTMATTSSTPTLSRRQRSRSATPTTPPPPPPAHSSSNGAYHH
GHHLVSSTAAT

>DmDSX-F
LGQDVFLDYCQKLLEKFRYPWELMPLMYVILKDADANIEEASRRIEEGQYVVNEYSRQHNL
NIYDGGELRNTTRQCG

>AmDSX-M
NVEILLEHSSKLVELFQYPWEALLLMYINLKYAGANPEEVVRRMVDASNEIRNMHFLKAIR
MSQPSRAFRCTAACAAPTGPPTGPPTYEGDVPFIGVGPPPNPIHFRPFLHPENAHIPATRLPSSP
DGPPKHT

>AmDSX-F
VEILLEHSSKLVELFQYPWEALLLMYINLKYAGANPEEVVRRMVDALIIFCSKNFIWNSILNKI
VSFINLLPT

>NvDM4-M4
VEELLGYSVKLLQRFGYHWQTLTLMYVILKDSRADVEVAMRRITQGNQS

>NvDM4-F1
VEELLGYSVKLLQRFGYHWQSLTLMYVILKDSRADVEVAMRRITQAKNVWQPELYSRIISV

Figure 47. Sequence from doublesex homologs beginning at the start of the doublesex oligomerization domain and ending at the end of the protein in *D. melanogaster* (Dm), *A. mellifera* (Am), and *N. vitripennis* (Nv). The text highlighted in green is the oligomerization domain shared between the two sexes of each organism. Text highlighted in blue is the male-specific region of the OD and text highlighted in pink is the female-specific region. Note that the tails succeeding the OD are considerably shorter in *N. vitripennis* than in any other organism for both males and females. Underlined portion of *A. mellifera* male is the sequence used to search for additional carboxyl ends.

Looking for a longer carboxyl terminus[1]

This significant difference between the male and female transcripts of NvDM4 and the other DSX homologs made me wonder if there were more alternatively spliced NvDM4 transcripts that had not been pick up by the 3' RACE and subsequent cloning. I decided to try a computational approach. To do this, I used a portion of the *A. mellifera* male DSX homolog after the oligomerization domain (fig. 47) to search to genome of *N. vitripennis* using a TBLASTN search (amino acids v. nucleotides). This resulted in one hit. Although the E-value was fairly high, 0.002, the newly found sequence was on scaffold 23, the same scaffold as the rest of the NvDM4 transcripts (fig. 48).

I needed to establish whether this hit was oriented in the correct direction to splice with the beginning exons of NvDM4. In order to serve as a potential new 3' end, the sequence had to be located to the 3' end of the *ab initio* sequence and oriented in the same direction. To determine this, I attached the new sequence onto the end of my NvDM4 male protein and repeated a TBLASTN search of the genome.

The results revealed that the new sequence was oriented correctly to be attached to the rest of the NvDM4 sequence. Also, the new sequence, along with the preceding two exons of NvDM4, aligned with another predicted protein, this one in RefSeq (XM_001602495.1) (fig. 49). According to NCBI, the RefSeq hit was predicted to be similar to an elongase, which is involved in fatty acid synthesis (Oh *et al*. 1997). When I took this predicted protein and searched for domains within it using Pfam, I found that it had two domains. In addition to the elongase domain, which had an E value of 1.5e-9 and was

[1] Please note that all references to OD1 and OD2 exon numbers do not include the results of 5' RACE, but instead begin with the coding region that corresponds with the beginning of the ab initio prediction.

located towards the carboxyl terminus, the protein contained a DSX oligomerization domain at the amino end with an E value of 2e-11, a better value than seen in either the OD-A or OD-B transcripts. Also, a search of the literature as well as the Pfam architecture revealed no previous example of a doublesex oligomerization domain being associated with an elongase domain. This raised a tantalizing possibility: could the predicted RefSeq be incorrect and, instead, could the exons that coded for the OD domain be associated with alternative transcripts of NvDM4?

To pursue this possibility, I returned to the lab to do RT-PCR. For this, I used Primer3 to design right primers for each of the exons unique to the putative elongase (PE). The PE consists of 9 transcripts, with the first two exons being shared with the NvDM4 sequence (fig. 50). Exon four of the PE was too small for a primer. I also designed a primer to be used as a positive control that sat in the in the fourth coding exon of NvDM4. I conducted RT-PCR following the protocol described in the methods section, using 35 cycles.

I first conducted the RT-PCR in males, then in females. Both produced the same results. While the primers PERE5 – PERE9 resulted in no strong bands, the combination of DM4LE1.6 with PERE3 produced a thick band in both males and females (fig. 51 and 52). In addition, the size of the products were consistent with the first two coding exons of NvDM4 being connected to the first three exons of the putative elongase. While exon three of the PE codes for the new carboxyl terminal of the OD, it does not include any of the elongase domain. These results would suggest the existence of a third transcript containing both a DM domain and an oligomerization domain distinct from OD-A and OD-B. I named this class of transcripts OD-C.

My next question was whether transcripts with this third OD were present in the pool of 3' RACE products from the first priming with the GSP. To do this, I created a new NGSP, this one in the third exon of the PE. In an attempt to have at least one working primer, I designed four different nested gene-specific primers and tested all of them in both males and females. The results of two of these primers, along with single primer controls for the new NGSPs are shown in figure 53. In all male lanes, there was a faint band between 750 and 1000 nucleotides which was not present in females. This band shifted in a size consistent with the location of the different primers, suggesting it was not an artifact. Both males and females had a larger band at between 3 and 6kb. While not seen in figure 53, at other times the nested universal primer mix has resulted in a band of about this size when alone, so this larger band may have been an artifact.

To determine the sequence of the smaller, male band, I cloned only the male products from the nested priming of PELE3.1 using protocol described in the methods section. I then performed a check of clone sizes on 20 clones (fig. 54). This revealed bands of five different main sizes, and I decided to prepare and sequence one clone from each. These were clones C2, C3, C4, C10, and C11.

With only one round of sequencing, I was able to complete clones C2, C3, and C11, using the sequence from M13R and T7 primers (fig. 38). Before designing internal primers for the remaining two clones, I looked more closely at the partial sequence reads. I found that C10 appeared to be an artifact, with all exons of the sequenced transcript residing on scaffold 62. To be a legitimate transcript, it would have to rest at least partially on transcript 23, where the primer was located. Taking a closer look at C4, I noticed that, although the trimmed sequence read from the T7 primer was extremely short, looking at the quality of the chromatogram, I felt I could trust more of the sequence than would be suggested by the trimmed read. Also, although the untrimmed sequence read of T7 still did not overlap with M13R, I found by aligning both transcripts with the genome using Spidey, that the ends were separated by only five nucleotides. I determined these five nucleotides – AAAAG – by referring to the genomic sequence. I decided to assume that this was not an extremely short intron and glued the two sequence reads together with the AAAAG between them. Also, even if this were an intron, it would not affect the resulting protein as further analysis revealed this region of the transcript was in the UTR. Therefore, I had four new transcripts to investigate further: C2, C3, C4, and C11.

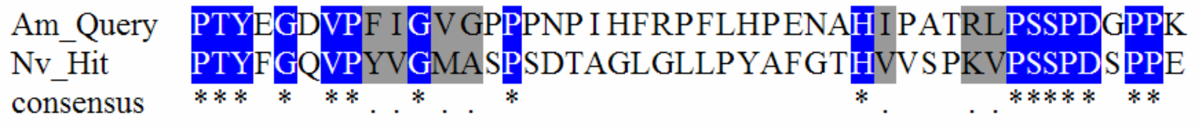
I aligned all four of these sequences against the genome using Spidey (fig. 55). I also used BLAST to search for each sequence in version 1.0 of the *N. vitripennis* genome and found that the last part of clone C11 was on scaffold 10, while the rest of the transcript, along with all the other sequences, was on scaffold 23. As can be seen in figure 55, while the new sequences align with exon three of the putative elongase, they do not align with the succeeding exons four through nine. Transcripts C2, C4, and C11 all have their stop codon in exon B, resulting in the same ORF for all three, but C3 is missing exon B and thus does not have a stop codon until exon C. Therefore, these new clones actually represent two different carboxyl ends (fig. 56). To reflect these differences in the polypeptide, I subdivided class OD-C into OD-C.1, containing C2, C4, and C11, and OD-C.2, consisting of only C3.

This left the question of what was happening to the 5’ end of these transcripts. To assess this, I cloned and sequenced the RT-PCR products from the primer DM4LE1.6 and PERE3 in both males and females (fig. 51 and 52). Checking clone sizes revealed that all products were the same size, so I sequenced one from each sex. After sequencing was complete, I aligned the results, labeled DM4-PE3 ♂ and DM4-PE3 ♀, with each other and with the genome (fig. 57). The sequences are the same in males and females. They correspond to the first four exons of the OD-B transcripts, using the splice site that the OD-A transcripts skip.

With the PE exon 3 appearing in both males and females, it raised the question as to whether exons A, B, C, and D were also in both sexes, despite the fact that I only saw product in the 3' RACE using male RNA as the template. I designed primers for each of the exons, and did RT-PCR using primer DM4RE1.6 as the left primer to check for the presence of these exons in both males and females. I ran the RT-PCR following the protocol described in the methods section, using 35 cycles. The primer in exon D did not result in any product in either males of females, suggesting that this exon, the only exon not on scaffold 23, was an artifact. In contrast, all of the other primers resulted in products of the expected sizes in both sexes (fig. 58 and 60). This would suggest that exons A, B, and C are present in females as well as males (fig. 59).

Amplifications off of both male and female templates result in a high quantity of products. This does not agree with the nested 3' RACE in which a band could be seen only in males, with no product apparent in the female RACE (fig. 53). However, while my protocol of 35 cycles for RT-PCR is excellent for showing presence versus absence of a particular transcript, it does not provide a quantitative look at the transcripts. By the time the reaction has gone through 35 cycles, it is likely that differences in original template quantities would be reduced considerably as other factors, such as beginning to use up nucleotides, becomes more limiting than amount of transcript. In an attempt to make a semi-quantitative examination of these new exons, I varied the number of cycles. To do this, I set up a number of reactions that all contained the primers DM4LE1.6 and DM4RA1. I then took one tube with male template and one tube with female template out of thermocycler after 15 and 20 cycles, leaving in two tubes to run for a full 25 cycles and the 3 minute extension. It should also be noted that, because I was running other experiments in the same thermocycler, I lowered the annealing and extension temperatures to 63°C. Previously, the annealing and extension temperatures were kept at 68°C as described in Methods.

The resulting gel revealed noticeable differences between males and females (fig. 60). While a very light band first appears in males at 15 cycles, no evidence of product is seen in females until 25 cycles have been completed. This would strongly suggest that this transcript, belonging to the OD-C class, is transcribed at a higher level in males than in females in yellow pupae. There is one significant caveat to this observation. Although the RNA was quantified and the same amount used in each cDNA preparation, the resulting cDNA has not been quantified. Therefore, these same differences could be the result of male cDNA being more concentrated than the female. The results of this semi-quantitative RT-PCR coincide well with the 3’ RACE results (fig. 53).



```
Am_Query   PTYEGDVPFIGVGPPPNPIHFRPFLHPENAHIPATRLPSSPDGPPK
Nv_Hit     PTYFGQVPYVGMASPSDTAGLGLLPYAFGTHVVSPKVPSSPDSPPE
consensus  *** * **..*.. *               *.   ..***** ** 
```

Figure 48. Alignment of *A. mellifera* query sequence with resulting hit from *N. vitripennis* that resulted from a TBLASTN search of the genome. The E value was 0.002, but the hit was on scaffold 23, the same scaffold as the male and female NvDM4 transcripts found using 3’ RACE.

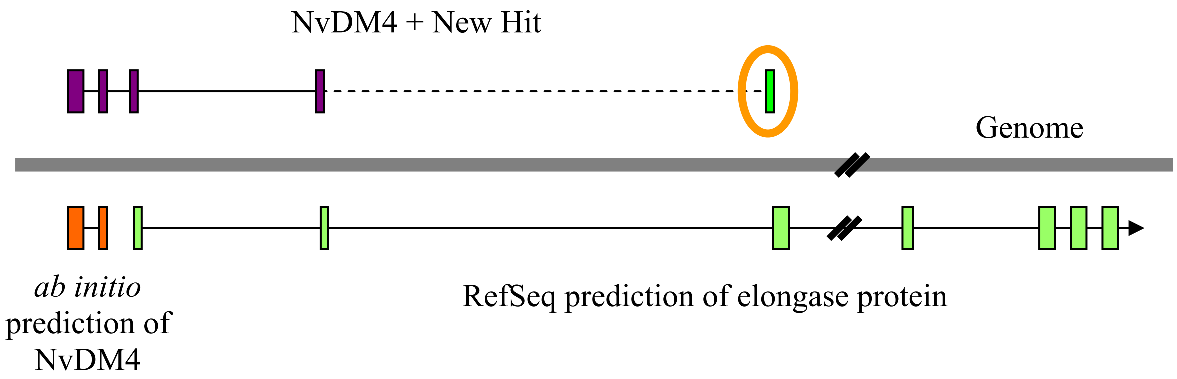

NvDM4 + New Hit
Genome
ab initio prediction of NvDM4
RefSeq prediction of elongase protein

Figure 49. The location of various transcripts relative to the genomic sequence. The genomic sequence is in gray. Predicted sequences are below the genomic sequence. Exons appear as boxes, with introns as connecting lines. The orange exons are the *ab initio* prediction of NvDM4. The light green exons are a RefSeq prediction of an elongase protein. The entire elongase sequence is not shown, but continues farther to the 3' end. The purple exons above the genomic sequence are the male NvDM4 sequence with the new added-on sequence in neon green and circled in orange.

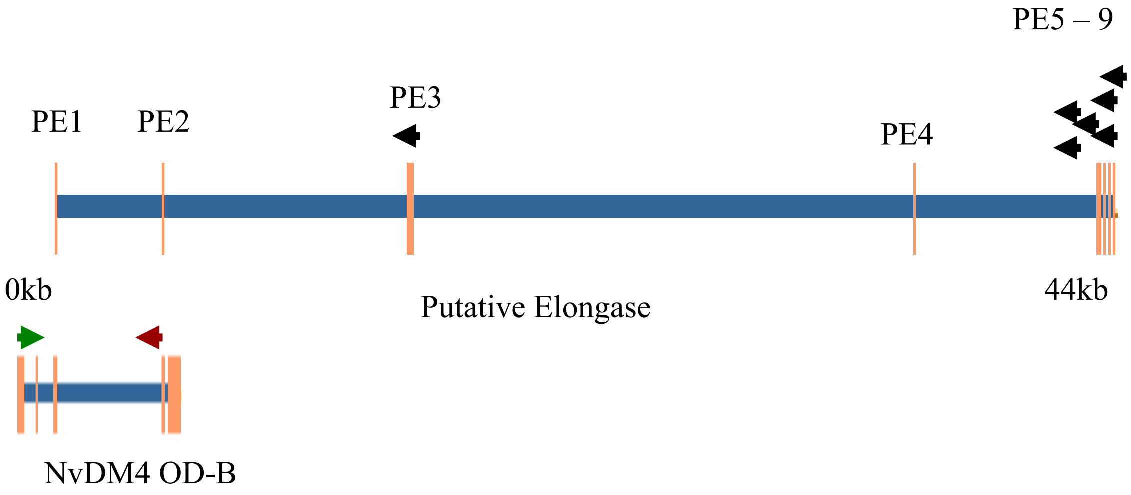


Figure 50. Location of primers for the RT-PCR testing whether any exons of the putative elongase were connected to NvDM4. The left primer, NvDM4LE1.6, is located in the beginning of the ORF of NvDM4. Primers for exons 3, 5, 6, 7, 8, and 9 of the putative elongase are in black. These primers are named PERE3, PERE5, etc. Exon 4 was too small for a primer. Exons are labeled as PE1, PE2, etc. A primer serving as a positive control is located in NvDM4 and is shown in burgundy. Note that the second exon of the PE aligns exactly with the fourth exon of OD-B transcripts. This sequence is also in OD-A and OD-X, but the PE excludes the VI seen in OD-A and OD-X. The genomic length from the beginning of the DM4 prediction to the end of the putative elongase is 44kb.

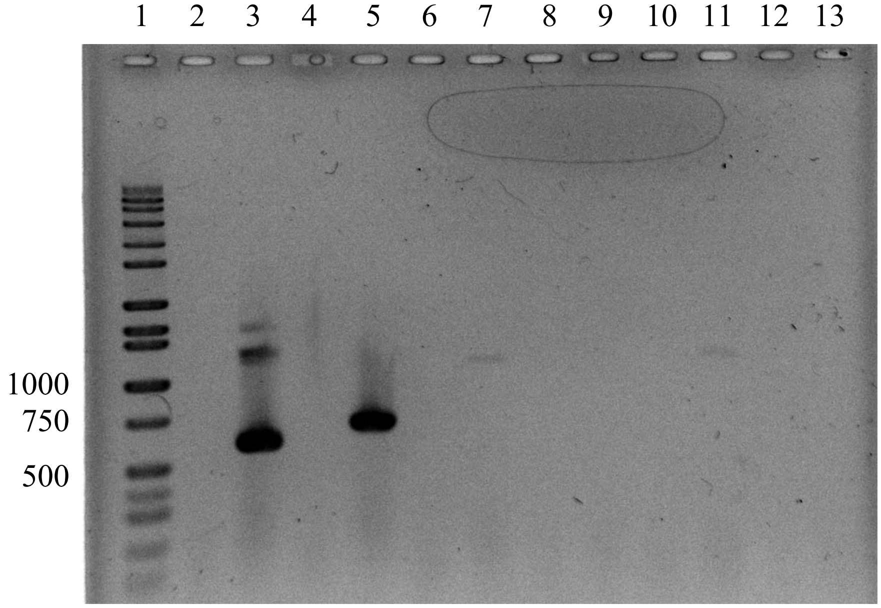


Figure 51. Results of RT-PCR to check whether exons of the predicted elongase were connected to the first coding exon of NvDM4 in males. The primers (2 and 3) DM4LE1.6 with DM4RE4 represent a positive control (no reverse transcriptase [RT], +RT). The rest of the lanes contain the following primers with in each case the lane representing the no RT control: (4 and 5) DM4LE1.6, PERE3; (6 and 7) DM4LE1.6, PERE5; (8 and 9) DM4LE1.6, PERE6; (10 and 11) DM4LE1.6 PERE7; (12 and 13) DM4LE1.6, PERE8. Lane (1) is Hi-Lo. This shows that the third exon of the PE is connected to the first exon of DM4.

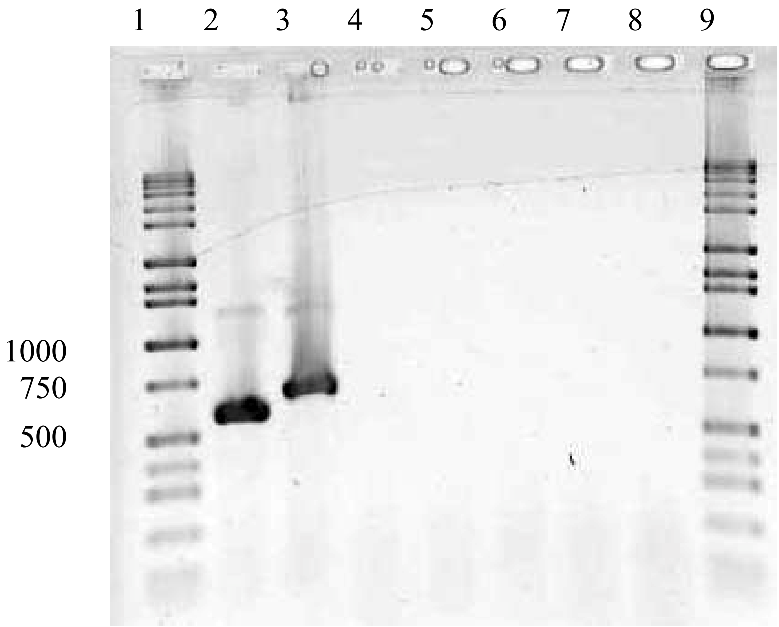


Figure 52. Results of RT-PCR to check whether exons of the predicted elongase were connected to the first exon of NvDM4 in females. (2) DM4LE1.6, DM4RE4 (+RT); (3) DM4LE1.6, PERE3; (4) DM4LE1.6, PERE5; (5) DM4LE1.6, PERE6; (6) DM4LE1.6 PERE7; (7) DM4LE1.6, PERE8; (8) DM4LE1.6, PERE9; (1 and 9) Hi-Lo. Again, the results suggest that the third exon of the PE is connected to the first exon of NvDM4.

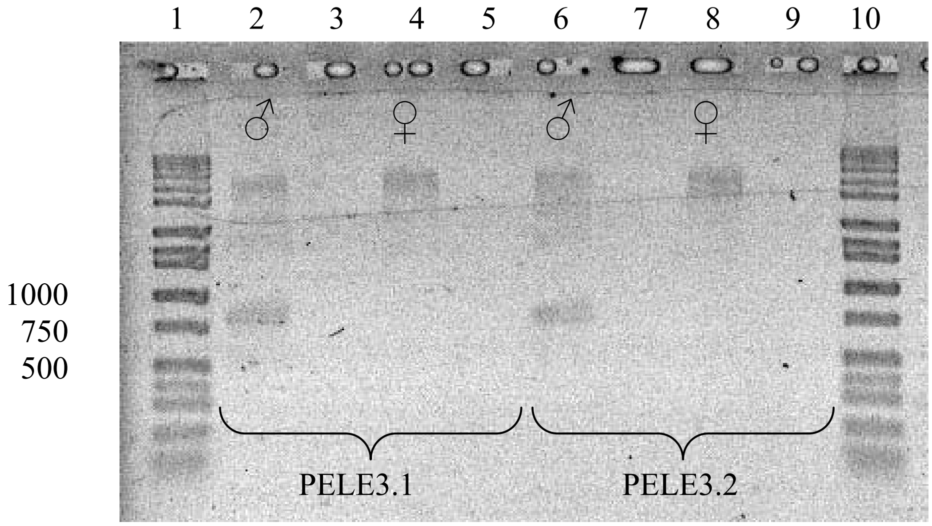


Figure 53. Results of nested gene-specific priming of original 3' RACE. The nested primers are designed to target the third exon of the PE, previously shown to be connected to the first exon of NvDM4 using RT-PCR (fig. 51 and 52). Lanes contain the following: (2) ♂PELE3.1, Nested Universal Primer Mix; (3) ♂ PELE3.1; (4) ♀ PELE3.1, NUPM; (5) ♀ PELE3.1; (6) ♂ PELE3.2, NUPM; (7) ♂ PELE3.2; (8) ♀ PELE3.2, NUPM; (9) ♀ PELE3.2; (1 and 10) Hi-Lo.

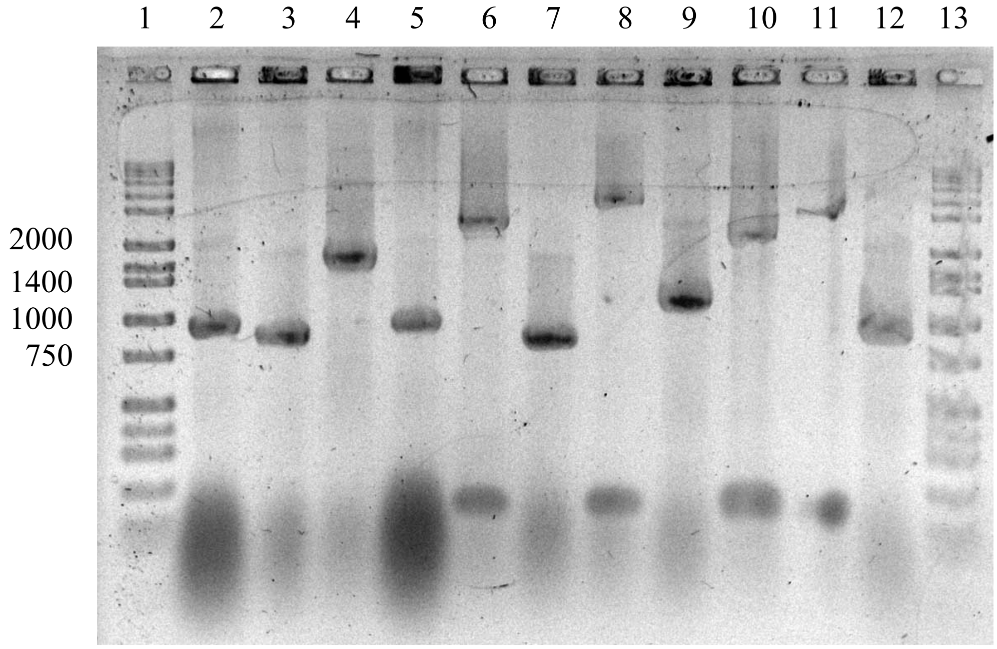


Figure 54. One of the two gels resulting from a check of clone sizes from the NGSP seen in fig. 51. Each clone is given a number identification as follows: (2) C2; (3) C3; (4) C4; (5) C5; (6) C7; (7) C8; (8) C10; (9) C11; (10) C12; (11) C13; (12) C15; (1, 13) Hi-Lo. Clones C2, C3, C4, C10, and C11 were sequenced.

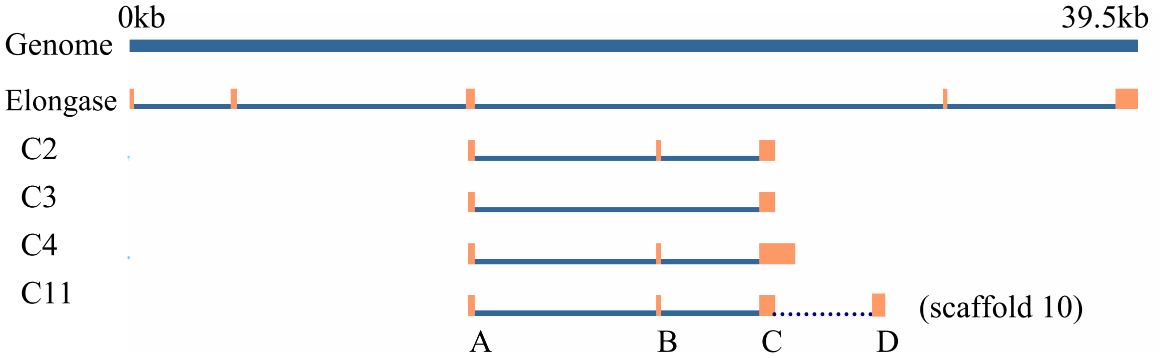


Figure 55. Alignment of clones C2, C3, C4, and C11 with the genome and the putative elongase. Exons are labeled A through D. C3 lacks exon B. Exon B contains a stop codon, as does exon C, resulting in two different ORF, one shared by C2, C4, and C11 and another seen only in C3. Note that exons five through nine of the PE are very close together and represented by the solid orange box at the far right.

>NvDM4OD-C.1
VEELLGYSVKLLQRFGYHWQSLTLMYVILKDSRADVEVAMRRITQANSEIQATAQF
NATFGGYYRGGYYPPSAFTNSLANIGNLGNPTYFGQVPYVGMASPSDTAGLGLLPY
AFGTHVVSPKVPSSPDSPPERPSSYPGGTSFPRSKQASIHS

>NvDM4OD-C.2
VEELLGYSVKLLQRFGYHWQSLTLMYVILKDSRADVEVAMRRITQANSEIQATAQF
NATFGGYYRGGYYPPSAFTNSLANIGNLGNPTYFGQVPYVGMASPSDTAGLGLLPY
AFGTHVVGPKVPSSPDSPPERPSSYPETSMSHRFKIEKSEDTD

Figure 56. Amino acid sequence of OD-C.1 and OD-C.2 class transcripts beginning at the start of the oligomerization domain and ending at the C-terminal of the protein. The difference between the proteins is underlined. The sequence highlighted in green are the region of the OD shared among OD-A, OD-B, and OD-C transcripts while the blue is the region of the OD specific to OD-C. As can be seen, OD-C.1 and OD-C.2 share the same oligomerization domain.

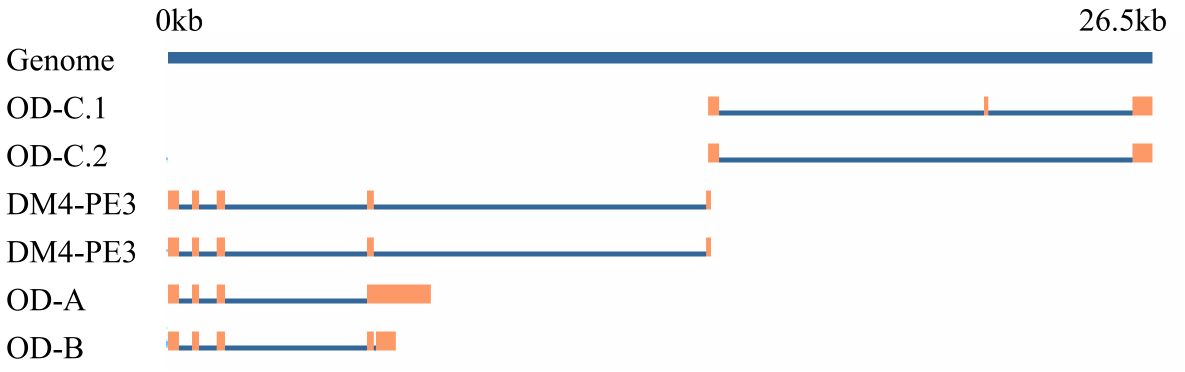


Figure 57. Alignment of the results of sequencing the male and female RT-PCR products with primers DM4LE1.6 and PERE3 (rows 4 and 5). Also included in this alignment the incomplete sequences from ODC.1 and ODC.2 (row 2 and 3) and representative transcripts from the OD-A and OD-B classes (rows 6 and 7). The fourth exon of DM4-PE3 male and female are the same as the fourth exon OD-B.

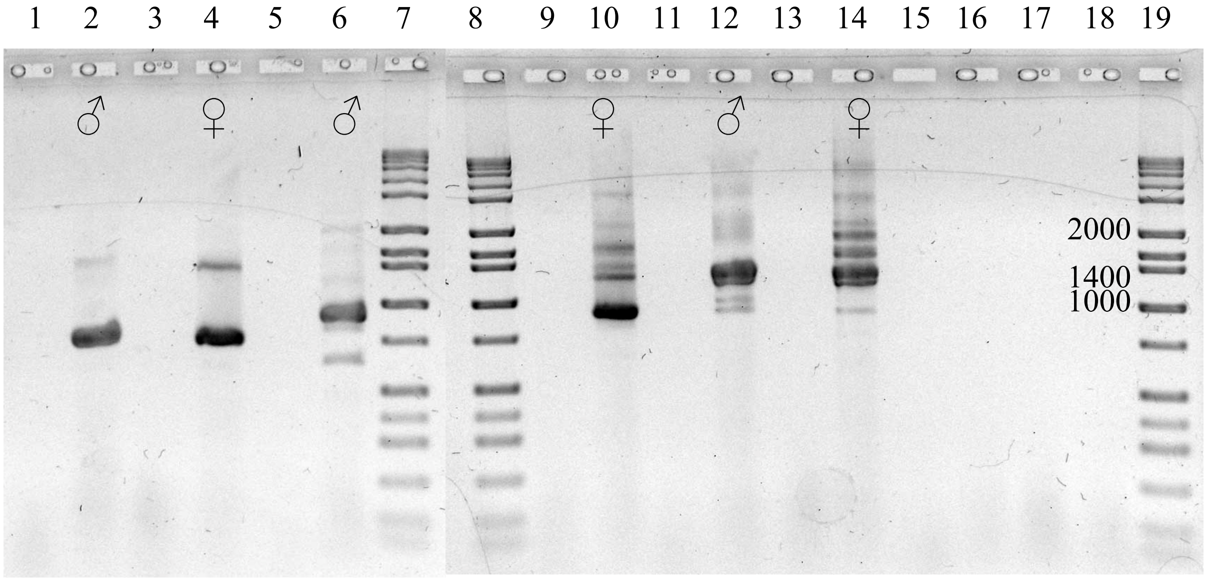


Figure 58. The results of amplification between the first coding exon of NvDM4 (primer NvDM4LE1.6) and exons A, B, C and D in males and females. The lanes are as follows: (1 and 2) male DM4LE1.6 and DM4RA.1 (-reverse transcriptase [RT], +RT); (3 and 4) female DM4LE1.6 and DM4RA.1 (-RT, +RT); (5 and 6) male DM4LE1.6 and DM4RB.1 (-RT, +RT); (9 and 10) female DM4LE1.6 and DM4RB.1 (-RT, +RT); (11 and 12) male DM4LE1.6 and DM4RC.1 (-RT, +RT); (13 and 14) female DM4LE1.6 and DM4RC.1 (-RT, +RT); (15 and 16) male DM4LE1.6 and DM4RD.1 (-RT, +RT); (17 and 18) female DM4LE1.6 and DM4RD.1 (-RT, +RT); (7, 8 and 19) Hi Lo.

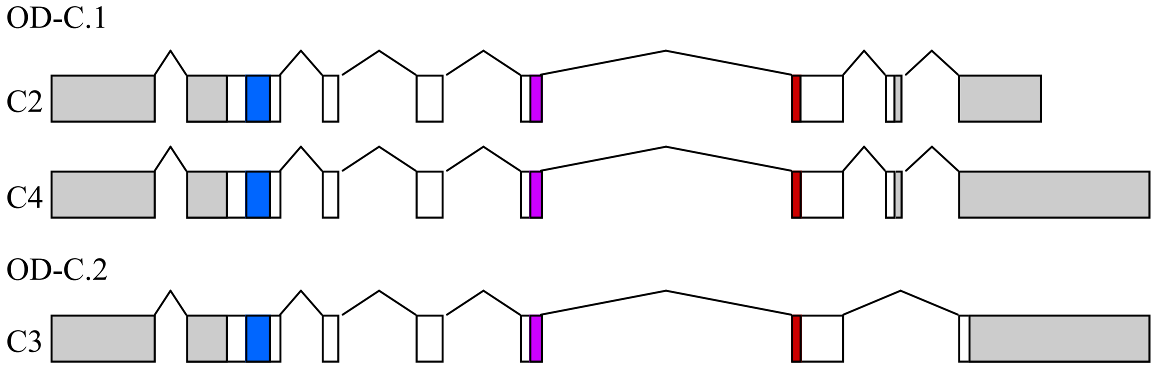


Figure 59. NvDM4 OD-C transcripts. DM domain is in blue, the oligomerization domain shared among the NvDM4 transcripts is in purple, and the carboxyl end of the oligomerization domain unique to the OD-C transcripts is in red. OD-C.1 and OD-C.2 have the same oligomerization domain, but differ at the C-terminal.

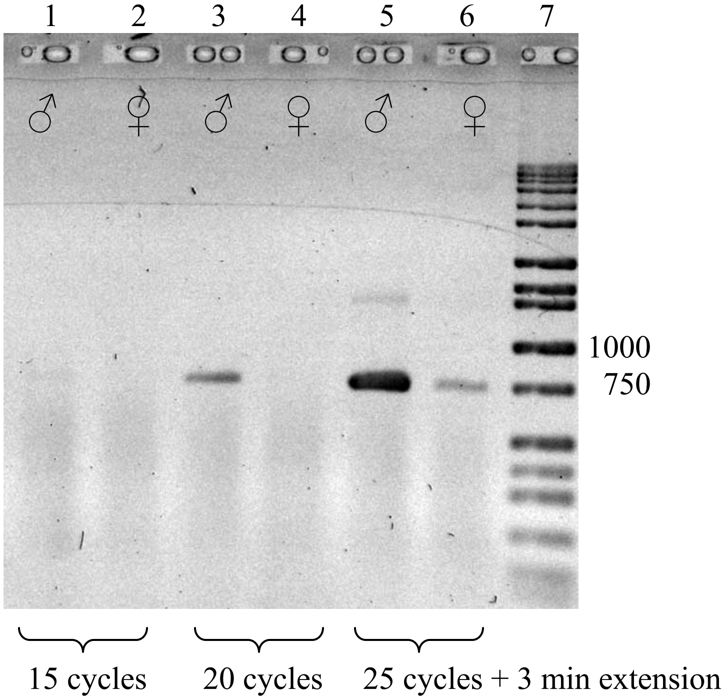


Figure 60. The effect of varying the cycle numbers on the amplification of DM4LE1.6 and DM4RA.1. Lanes 1, 3, and 5 have a male template and lanes 2, 4, and 6 have a female template. Lane 7 is Hi-Lo. A very light band first appeared in males at 15 cycles, but no evidence of product was seen in females until 25 cycles had been completed.

Another look at OD-A and OD-B

My original 3’ RACE resulted in two clearly different classes of doublesex oligomerization domains that divided neatly between the sexes: all males containing an OD fell into the OD-A class while all females were classified as OD-B, lacking the VI. However, especially in light of finding a third class of transcripts that I had missed in the first round of nested RACE, I was concerned that I had only seen a sex-specific pattern by chance, and that the OD-A and OD-B classifications might not be exclusively male or exclusively female as initially observed. Therefore, I attempted to confirm what I had found using RT-PCR. I designed primers to check whether OD-X and OD-A were male-specific and whether OD-B was female-specific.

To test whether OD-A was present exclusively in males, I designed a primer called DM4RMale1 to sit in the VI, which is spliced out of OD-B (fig. 61). Because the primer would amplify off of both OD-X and OD-A, I could expect to see two bands of product in the gel, one at 667 base pairs that would correspond with OD-A transcripts, and the other at 1236 to correspond to the OD-X transcript. The results showed these products in both males and females (fig. 62). This indicates that neither OD-X nor OD-A are male-specific, but instead are expressed in both sexes.

The presence of OD-X in both sexes was confirmed when I used the ODX-specific primer DMR-M5-1 along with DM4LE1.6. A band appears in both sexes at the expected size (fig. 63). A second, larger band is also seen, but it corresponds with the size of amplification off of the genomic DNA.

I then wanted to check whether OD-B was female specific. To do this, I designed primers to sit on either side of the VI. In those transcripts containing the VI, all those in classes OD-A and OD-X, the expected product would be 300bp in length. For transcripts in class OD-B, which do not contain the VI, the product should be 192bp long. At 30 cycles, males have only one product, corresponding with the longer transcript, while females display products of both sizes (fig. 64). This suggests that OD-B is a female-specific splice pattern.

Looking at the results as a whole, it appears that class OD-B transcripts are present exclusively in females and OD-C transcripts are at considerably higher levels in males, making OD-B the female-specific transcript and OD-C the male-specific transcript. On the other hand, OD-A and OD-X, while initially found only in males, are actually present in both sexes.

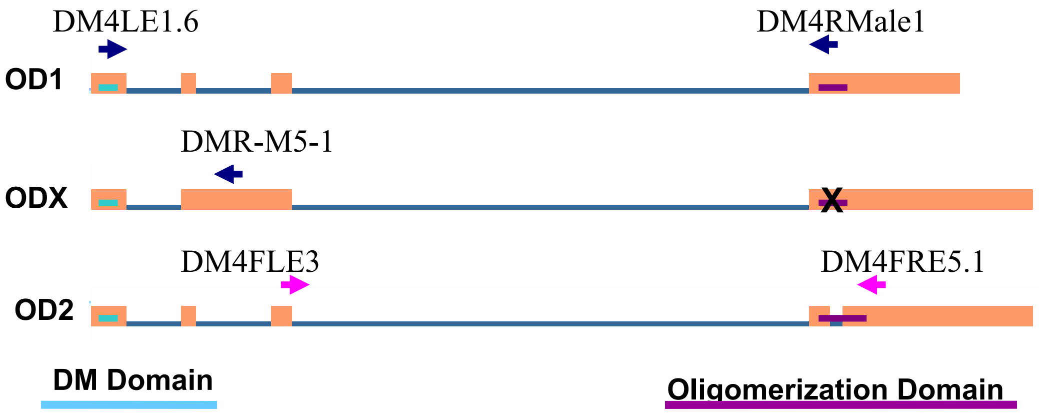


Figure 61. Location of primers to determine whether the splicing patterns seen as sex-specific in 3' RACE are indeed transcribed exclusively in one sex of yellow pupae. DM4RMale1 sits entirely in the VI. Blue arrows indicate primers testing for OD-A and OD-X. Pink arrows indicate primers testing for OD-B. In both cases primer labels appear above arrows.

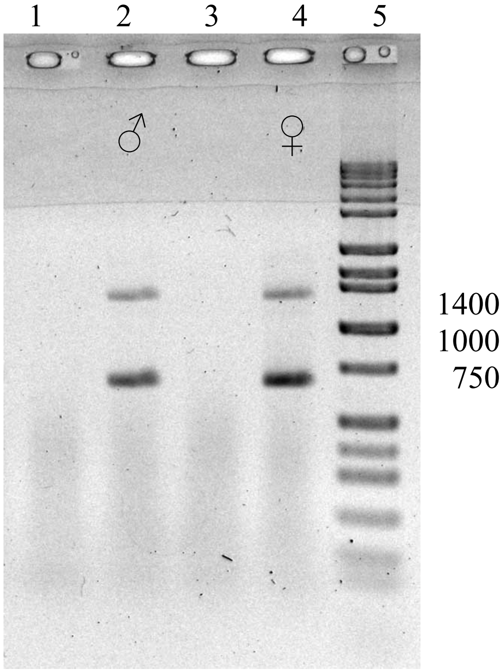


Figure 62. Amplification to check for presence of VI in females. VI had previously been seen only in males. Lanes 1 and 2 are male products (-RT, +RT) and lanes 3 and 4 are from female template (-RT, +RT). Lane 5 is Hi-Lo marker. The shorter product corresponds to OD-A transcripts (expected size: 667) while the larger product is due to amplification off of OD-X (expected size: 1236).

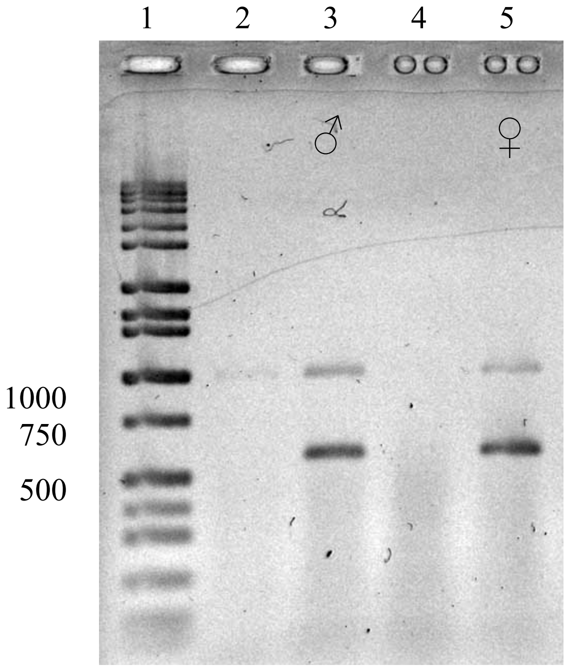


Figure 63. Amplification to check for the presence of the OD-X transcript. Lanes 2 and 3 are amplified off of male cDNA (+RT, -RT). Lanes 4 and 5 are amplified from female cDNA (+RT, -RT). Lane 1 is Hi-Lo marker. The darker band is at the expected product size of 578bp with the larger, fainter band corresponding with the expected size of amplification off of genomic DNA (1011bp).

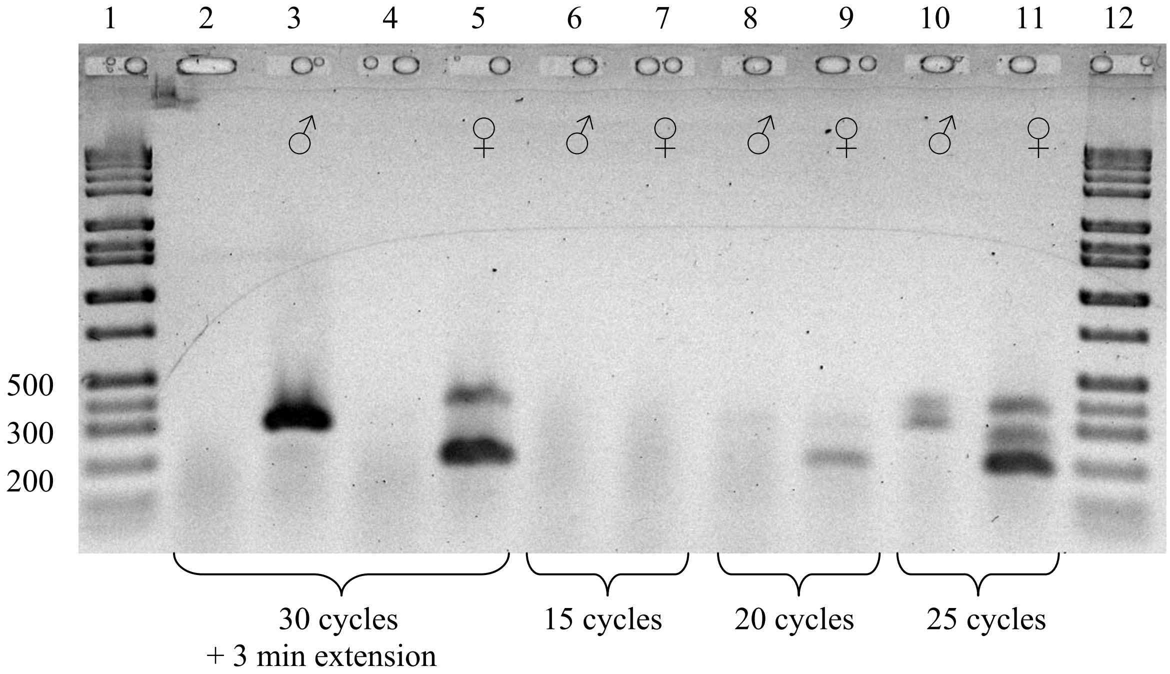


Figure 64. Amplification to check for the presence of OD-B transcripts in males. OD-B transcripts result in a 192bp product while OD-A and OD-X transcripts result in a 300bp product. Lanes 2 and 3 amplify off of male cDNA (-RT, +RT) and lanes 4 and 5 amplify off of female cDNA at 30 cycles, plus a 3 minute extension. Lanes 6, 8, and 10 have male cDNA as a template with 15, 20, and 25 cycles respectively. Lanes 7, 9, and 11 have female cDNA as a template. Lanes 1 and 12 are Hi-Lo marker.

Splice-site analysis

In *N. vitripennis*, I found four different classes of transcripts in NvDM4, OD-A, OD-B, OD-C, and OD-X, that all arise from the same pre-mRNA. Of these, OD-C appears to be male-specific, with significantly higher levels of transcript in males, while OD-B seems to be female-specific, expressed in only females at the yellow pupal stage. Because males and females share an identical pre-mRNA, signals have to be present in the primary transcript to result in the differential splicing that can be seen in doublesex orthologs, including NvDM4. How this alternative splicing is carried out varies between organisms. In *D. melanogaster*, there are two cis-regulatory elements that work in concert with a protein complex that includes the transformer (tra) and transformer-2 proteins and result in the production of the female mRNA. One of these elements is called the dsx repeat element (dsxRE) and it consists of 6 repeats of a 13-nt (Inoue *et al*. 1992). This 13-nt sequence varies slightly (table 6a) and is found in the fourth exon of *D. melanogaster* doublesex (fig. 65). The second element is the purine-rich enhancer (PRE) which sits within the dsxRE between the fifth and sixth repeat elements (for sequence, see table 6b; for location, see Fig. 65). These two elements work together to bind the protein complex and result in the female-specific transcript (Lynch and Maniatis 1995).

Using Microsoft Word's find function, I searched for all four variations of dsxRE in all of the NvDM4 transcripts, but found no examples in any of my sequences. There were also no examples of the complete PRE sequence.

In addition to the presence of cis-regulatory elements, another consideration in splicing is the strength of the splice sites themselves. To measure the relative strengths of the alternative splice sites, I used the Splice Site Prediction by Neural Network program called NNSPLICE version 0.9 (Reese *et al*. 1997) as well as the Alternative Splicing Prediction (ASPIC) program (Bonizzoni *et al*. 2005).

For the construction of each transcript, there are three splicing "decisions" that must be made in terms of alternative splicing. Considering these from 5' to 3', the first is whether or not to splice out intron C (see Fig. 66), which contains donor and acceptor scores of 0.76/74 and 0.62/77 respectively (table 7). Note that the donor site is at the 5' end of an intron and the acceptor site is at the 3' end. While these scores are lower than for some of the other introns, in only one of the transcripts sequenced – the single transcript in the class OD-X – is this intron left in. The second "choice" is following exon 4a, whether to splice out intron E (shorter, seen in OD-B) or F (longer, seen in OD-C) or to not splice out any intron (seen in OD-X and OD-A). For E, the donor and acceptor scores are 0.63/79 and 0.89/96, while for F the donor score is the same and the acceptor score is 0.47/72, making F a weaker splice site according to both programs. Finally, for those transcripts that are destined to fall under the class OD-C, there is a final "decision" to be made as to whether to splice out introns G and H, leaving in exon 6, or to splice out the longer intron I, which removes exon 6. Intron G has donor and acceptor scores of 0.94/85 and 0.99/92 and intron H has scores of 0.87/82 and 0.84/94. Intron I shares the same donor as G and the same acceptor as H, giving it scores of 0.94/85 and 0.84/94 for donor and acceptor sites, respectively.

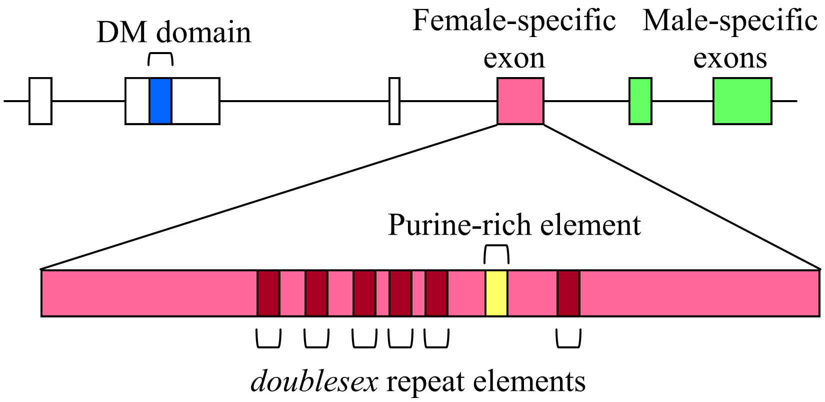


Figure 65. The location of the purine rich enhancer (PRE) and *dsx* repeat elements (dsxRE) in the transcript of *D. melanogaster doublesex* mRNA. The exon structure of doublesex shows the female-specific exon in pink and the male-specific exons in green. The female-specific exon is magnified and the dsxREs are red and the PRE is yellow (adapted from Lynch and Maniatis 1995).

Table 6. (A) 13-nt sequences that are repeated six times in the fourth exon of doublesex in *D. melanogaster* and make up the dsx repeat elements (dsxRE). The dsxRE is involved in binding tra and other proteins that result in female-specific splicing. (B) Sequence of the purine-rich enhancer (adapted from Inoue *et al.* 1992)..

A.

dsxRE Sequence
TCTTCAATCAACA
TCTACAATCAACA
TCAACAATCAACA
TCAACGATCAACA

B.

PRE Sequence
AAAGGAC AAAGGAC AAAA

Table 7. Splice site scores (higher is stronger) of alternatively spliced NvDM4 exons. For names of introns, refer to figure 66.

Intron: d=Donor; a=Acceptor	NNSPLICE (0 – 1)	ASPIC (0 – 100)
C: d	0.76	74
C: a	0.62	77
D: d	0.94	87
D: a	0.95	92
E: d	0.63	79
E: a	0.89	94
F: d	0.63	79
F: a	0.47	72
G: d	0.94	85
G: a	0.99	92
H: d	0.87	82
H: a	0.84	94
I: d	0.94	85
I: a	0.84	94

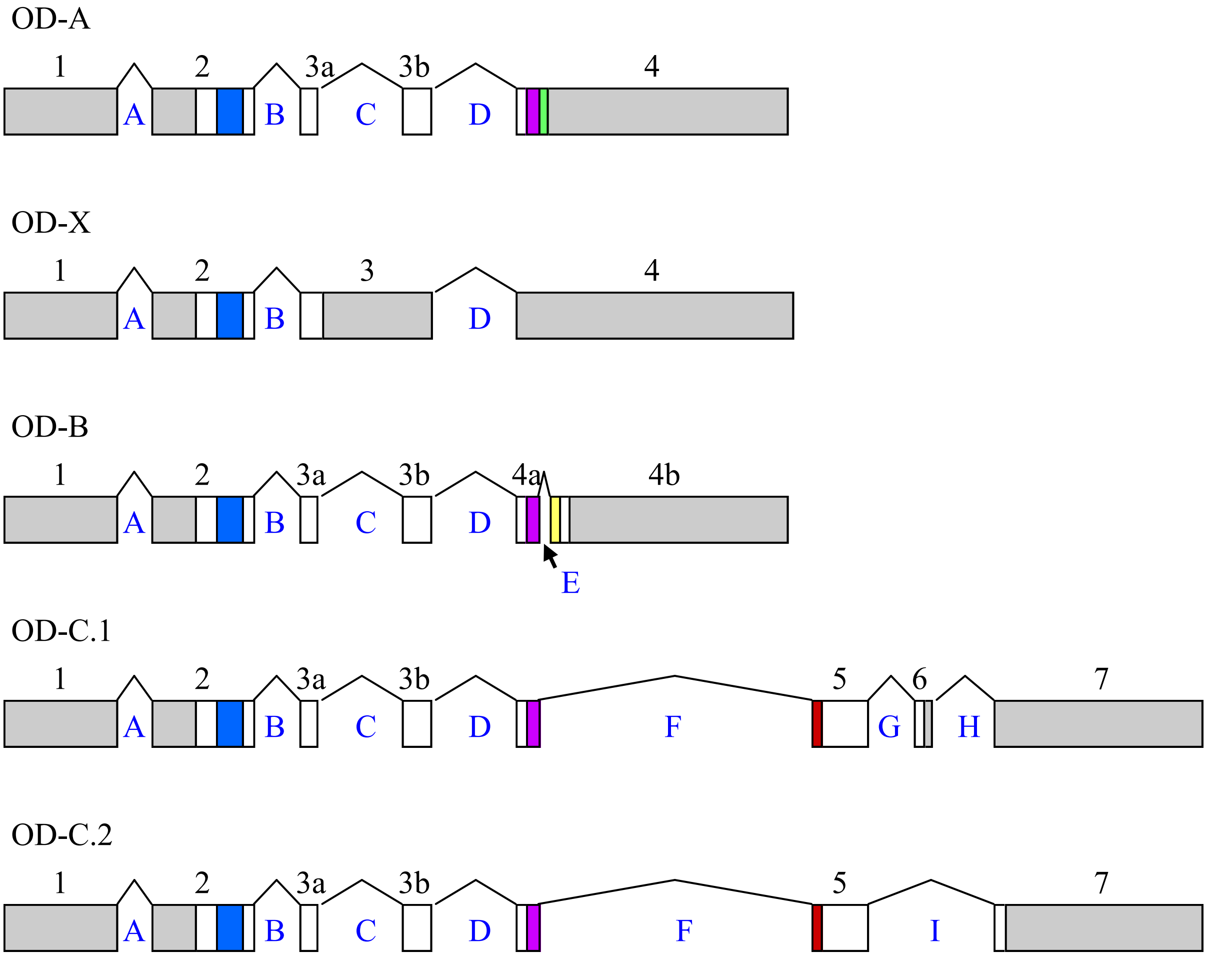


Figure 66. Complete exon/intron structure of the NvDM4 gene. Exons, indicated by rectangles, are numbered in black above each exon and introns, indicated by lines, are lettered in blue below each intron. An intron is considered unique if it has a distinct combination of donor and acceptor site. Exons that occupy a subset of the sequence of a larger exon are denoted by the name of the larger exon, followed by the letters a or b. The blue rectangle represents the DM domain. The purple represents the beginning of the oligomerization domain shared by OD-A, OD-B, and OD-C with the green, yellow, and red representing the C-terminus of the OD-A, OD-B, and OD-C transcripts respectively. The white rectangles indicate the open reading frame that is not part of the two motifs and the gray represents the untranslated region.

Transcript variety

Previously, the highest number of doublesex transcripts had been found in honeybees, with a total of four transcripts (Cho *et al*. 2007). Of these, the two found primarily in females vary only in the 3' UTR, resulting in the same sequence of amino acids. Thus, if one compares merely the polypeptides, the *N. vitripennis doublesex* ortholog appears to code for four different proteins, while the *A. mellifera* ortholog codes for three. In honeybees, one of these proteins is found in both males and females, with amounts varying throughout development (Cho *et al*. 2007). In contrast, in *N. vitripennis*, I found two classes of transcripts, OD-X and OD-A, that are present in both sexes.

When considering differences in 3' UTR, *N. vitripennis* has nearly three times the number of transcripts compared with *A. mellifera*, with 11 versus 4. This raises the question as to whether all 11 represent legitimate transcripts or whether the shorter transcripts from any particular class are simply artifacts. Because the variation among transcripts within a class occurs due to variations in the length of the last exon, this is very difficult to tease out using RT-PCR, as primers cannot be specifically designed for shorter transcripts that will not also amplify off of the longer transcript.

One possibility for addressing this is to look for polyadenylation signals that might help to confirm the 3' ends of the shorter transcripts. For polyadenylation to occur, there are both upstream and downstream signals. The best conserved is the upstream signal AATAAA that occurs 10 to 30 bases upstream of the cleavage site. Note that, in the mRNA, this would read AAUAAA, with thymine replaced by uracil in RNA. In one study of 2084 genes in *D. melanogaster*, this signal appeared in over 47% of the genes, with the second-most common sequence being ATTAAA at a little over 10.3% of the genes (Retelska *et al*. 2006). There is also often a less-well conserved T or TG-rich region within 30 nucleotides downstream of the cleavage site (Retelska *et al*. 2006). Examining the shorter OD-A transcripts for upstream signals, I found an ATTAAA motif 30 base pairs upstream of the polyadenylation site for the DM4-M2 transcript. Also, 15 base pairs after the cleavage site for DM4-M2, there is a TTTGTTT, which is the most common of the T-rich motifs (Retelska *et al*. 2006). None of the other OD-A transcripts has these motifs near to their cleavage sites, although DM4-M4 does have the a TATAAA sequence, the third most-common upstream motif in flies at a little less than 5%, about 60 base pairs upstream of its polyadenylation site. Among the shorter OD-B transcripts, DM4-F1 has an ATTAAA motif less than 20 base pairs upstream of the polyadenylation site and a TTTGTTT site 40 base pairs downstream. The rest of the shorter transcripts do not have these motifs nor does the shorter OD-C.1 transcript. However, a lack of such motifs cannot be seen as certain evidence of artifact. In their analysis of over 2000 fly genes, Reteleska and his colleagues showed that over 22% lacked any previously identified upstream motif (2006).

Another possibility is to look for ways in which artifacts could be formed. For example, in 3’ RACE there are two primers used: one is gene-specific and the second, used in the first round of RACE, consists primarily of a poly-T region that binds to the poly-A tail of mRNA. However, this poly-T primer could feasibly bind at least partially anywhere in a transcript where there is a run of As. If the binding was sufficient, this could result in amplification of an artificially shortened transcript. Thus, I looked at the sequence directly following the cleavage site of the shorter transcripts (fig. 67, table 8). From this table, it is apparent that all polyadenylation sites are followed by at least one A and include poly-A runs up to 8 base pairs in length. The presence of an A itself after a cleavage site is not unusual. At least in humans, there is a significantly higher frequency of adenines at this position than would be expected at random (Retelska *et al*. 2006). While having one or two As in a row is of little concern, a string of 8, as seen in DM4-M2, might be enough to result in aberrant priming and an artificially shortened transcript. Interestingly, as noted previously, DM4-M2 has the best surrounding motifs of the shortened male transcripts to indicate it is legitimate. Thus, while looking more closely at the sequences surrounding the polyadenylation sites of shortened transcripts may provide some insight, it cannot indicate with certainty whether any of these transcripts are artifacts.

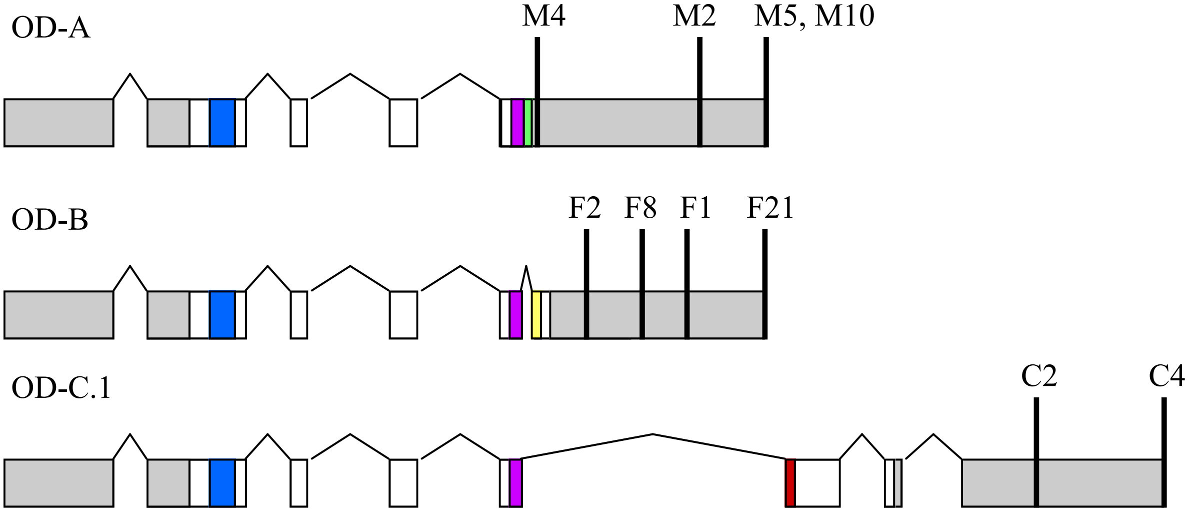


Figure 67. The location of polyadenylation sites on class OD-A, OD-B, and OD-C.1 transcripts. The transcript classes OD-C.2 and OD-X did not show alternative polyadenylation sites.

Table 8. Examination of the 10 base pairs immediately following the cleavage site in shortened transcripts, assuming that the nucleotide preceding the cleavage site is not an adenine.

Transcript	10bp following cleavage site
DM4-M2	AAAAAAAATT
DM4-M4	AAAAAACGTC
DM4-F1	AAACAAAGTG
DM4-F2	AAAAAATTAT
DM4-F8	AGTAAGTTCC
DM4-C2, DM4-C3	AAAAAACGAC

DISCUSSION

The DM gene family

My research shows that the DM gene family of *N. vitripennis* has four members, consistent with the number found in *A. mellifera* and *D. melanogaster* (Cristino *et al*. 2006). In addition to the DM domain, three of these transcripts show additional protein motifs (fig. 57). Both NvDM1 and NvDM3 contain a DMRTA motif at their C-terminal ends while NvDM4 has an alternatively spliced doublesex oligomerization domain, also at the C-terminal.

In some gene families, the genes remain close together. The *Hox* genes, for example, appear in clusters (Carroll *et al*. 2001). However, following duplication, it is possible for a gene to be transferred to another location in the genome via translocation (Lewin 2004). Because there are at present nearly 6000 scaffolds in the *N. vitripennis* genome that have not yet been placed on chromosomes, it is not within the scope of this study to say whether the DM gene family in *N. vitripennis* has remained clustered or has spread throughout the genome. While the chromosomal locations cannot yet be determined using computational techniques, scaffold locations show that NvDM2, NvDM3, and NvDM4 are present on single scaffolds while the sequence of NvDM1 stretches across at least five scaffolds. No two genes reside on the same scaffold.

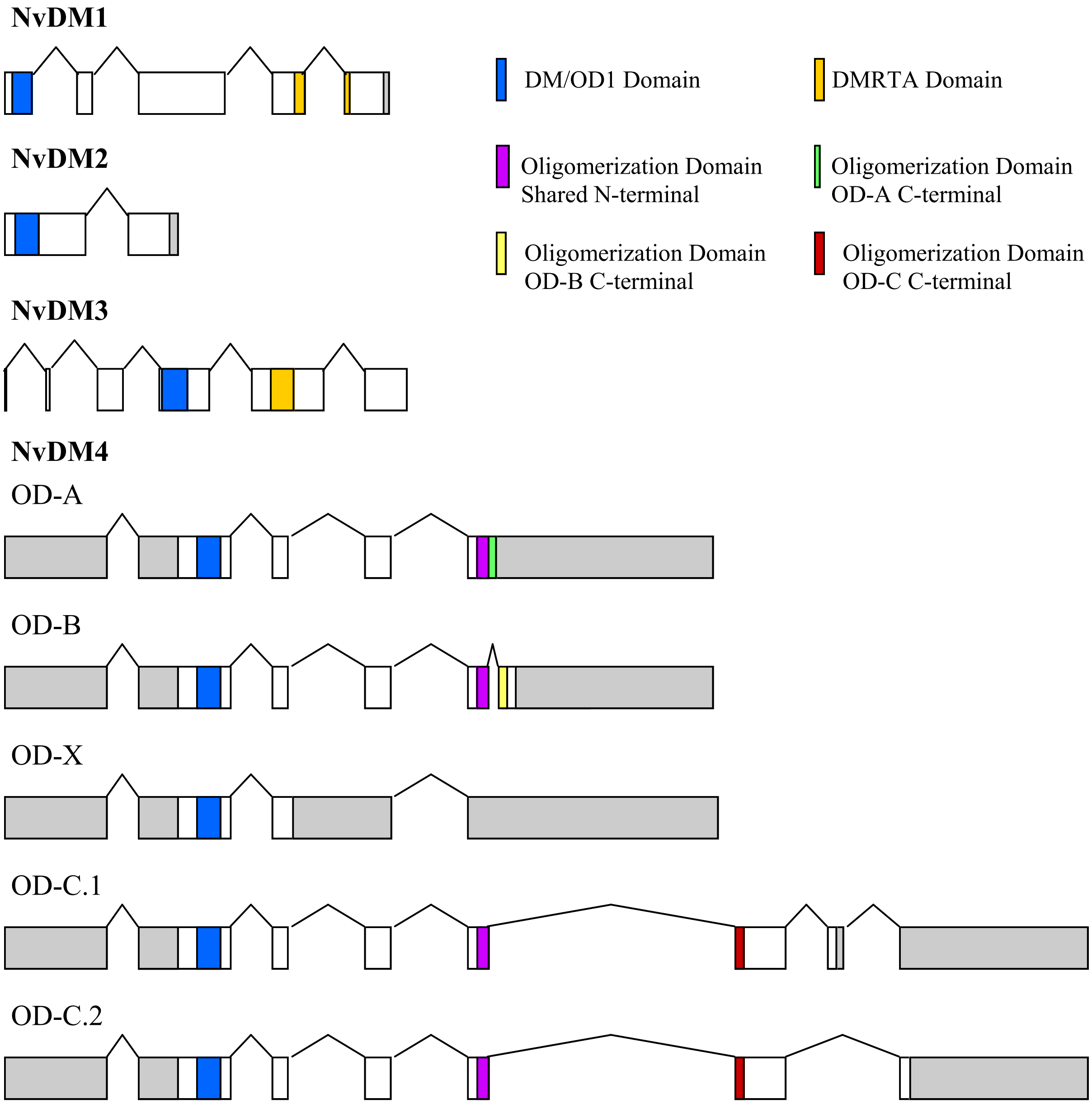

NvDM1
NvDM2
NvDM3
NvDM4
OD-A
OD-B
OD-X
OD-C.1
OD-C.2
DM/OD1 Domain
DMRTA Domain
Oligomerization Domain Shared N-terminal
Oligomerization Domain OD-A C-terminal
Oligomerization Domain OD-B C-terminal
Oligomerization Domain OD-C C-terminal

Figure 68. The exon structure of NvDM1, NvDM2, NvDM3, and NvDM4, including the four splicing classes of NvDM4. White and colored boxes represent open reading frames while grey indicates untranslated regions. Introns are not to scale.

Transcripts from NvDM1, NvDM2, and NvDM3 lack obvious sex-specificity in the yellow pupal stage

NvDM1, NvDM2, and NvDM3 are all members of the DM gene family in *N. vitripennis*. However, my results combined with previous work done in the Trent lab suggest that none of these genes serve as a *doublesex* homolog. Both NvDM2 and NvDM3 are constitutively spliced in yellow pupae and neither show sex-specific expression in this stage. Although NvDM1 is alternatively spliced, molecular evidence showed the same splicing patterns in both males and females during the yellow pupae stage of development.

Computational analysis of the protein domains of each gene product confirmed the molecular evidence that none of these three genes are doublesex homologs. Neither the many alternative transcripts of DM1 nor the single transcripts of DM2 and DM3 contained the doublesex oligomerization domain (OD) (Pfam accession: PF08828). The oligomerization domain serves two roles in sex-determination. First, DSX functions as a dimer *in vivo*, binding to palindromic sequences of DNA, and this domain allows it to form those dimers (Erdman *et al*. 1996). Second, in most DSX orthologs, it is the differences in the C-terminal region of the protein that includes the carboxyl end of the OD that confers the sex-specific properties via protein-protein interactions (Zhu *et al*. 2000).

Instead of an OD, both transcript 3.4 of NvDM1 and the single transcript of NvDM3 contained a DMRTA motif, also referred to as a DMA domain (Pfam accession: PF03474). This domain, named because it was found in the DMRTA proteins in humans, has an unknown function, but has been found in a wide range of organisms in association with the DM domain (Miller *et al*. 2003). One *C. elegans* and two *D. melanogaster* proteins contain both the DM and DMRTA domains. Of particular importance is that none of these genes are homologs of *doublesex* or the *C. elegans* gene *mab-3*. In *A. mellifera*, of the four DM-containing genes, two have DMRTA motifs, but neither of these occurs in the honeybee's *dsx* homolog (Cristino *et al*. 2006).

Interestingly, unpublished work by Wen suggests that one of these DMRTA-containing genes in *D. melanogaster*, Dmdmrt93B, does have a sex-specific function in flies, playing a role in early female development and male fertility (2002). Differences in the expression between the sexes are tissue-specific, with major transcripts found in the heads of males and females but only in the bodies of males (Wen 2002). Similarly, *Daphnia magna*, a tiny crustacean that switches between sexual and asexual reproduction, contains a homolog to dmrt93B that is only transcribed in the testis (Kato *et al*. 2008). Both of these studies suggest that, although it is less well-documented, genes with a DMRTA domain can also play a sex-specific role. In my own work, I used RNA isolated from the entire wasp. Thus, differences in spatial expression of NvDM1, NvDM2, and NvDM3 have not been addressed; if one of the genes is transcribed in a tissue specific manner similar to Dmdmrt93b, my techniques would have not been able to detect such differences. Future research should consider the possibility that sex-specific expression of DM1, DM2, or DM3 could be tissue-specific. Also, because I looked at only one stage in *N. vitripennis* development, these transcripts may show sex-specificity at either earlier or later stages. While it is certainly possible that these genes are showing sex-specificity at earlier stages, in *A. mellifera* sex-specific splicing of the *doublesex* homolog occurs in embryos, small larvae, large larvae, and pupae (Cho *et al.* 2007). In that particular study, adults were not tested. From these results, we would likely expect to see the sex-specific splicing occurring in the yellow pupa stage of

N. vitripennis. However, there is the possibility that differences occur only in early development before the sexes become morphologically distinct.

Alternative splicing and regulation of expression in NvDM1

Of the four NvDM1 transcripts analyzed, three code for proteins less than 150 amino acids in length, suggesting that they may not even be functional (fig. 27). This raises the question as to the biological role of these transcripts. It is possible that these shortened forms, with their early termination codons, have a regulatory role despite not coding for biologically active protein.

There are other examples of alternative splicing with early stop codons serving as a way to turn genes off or on. One example of this is the *transformer* (*tra*) gene in the sex-determination pathway in *D. melanogaster* (Schütt and Nöthiger 2000). In fruit flies, the *tra* gene regulates the alternative splicing of *dsx* and is, itself, alternatively spliced. Both males and females produce identical pre-mRNAs of *tra*. However, the male pattern of alternative splicing, which results from the absence of *sex-lethal* (*sxl*) expression in this sex, includes a stop codon in the second of four exons, which is spliced out of females (Schütt and Nöthiger 2000). This premature stop codon prevents the formation of an active TRA protein in males, resulting in the default male-specific splicing of *dsx*. In females, SXL causes the second exon of *tra* to be skipped, resulting in a full-length, active TRA protein. The female TRA protein then causes female-specific splicing of doublesex (Schütt and Nöthiger 2000).

While *tra* is an example of sex-specific inclusion of a premature termination codon (PTC), this also can occur in a spatial- or temporal-specific manner. Glutamic acid decarboxylase (GAD) is expressed at high levels in the GABAergic neurons of the central nervous system (Bond 1990). In rodents, GAD is alternatively spliced such that in embryos there is an early stop codon that is then spliced out of adults. This results in a truncated protein in embryos. Unlike *tra*, in this instance the shortened protein is also active. It should be noted that whether the shortened transcripts of NvDM1 code for functional proteins has not been addressed.

The cell has a vested interest in assuring that only the proper mRNA resulting in functional protein are translated and those with mistakes are quickly found and destroyed. In eukaryotic organisms, a system called nonsense-mediated decay is capable of detecting transcripts that contain a PTC that would result in a shortened protein (fig. 69) (Wagner and Lykke-Andersen 2002). In order to successfully remove transcripts with PTCs, the cell must have an accurate way of detecting these transcripts. In mammals, exon junction complexes, a multi-protein complex, sit near the exon-exon boundaries of new transcripts. When the transcript undergoes translation for the first time, these complexes are removed (Metzstein and Krasnow 2006). However, if the stop codon appears at least 50 base pairs to the 5' end of the last exon-exon junction, an exon complex will remain attached to the transcript, resulting in the recruitment of other proteins that mark this transcript for decay (Garneau *et al*. 2007).

According to this model, only transcript 3.46 of NvDM1 would be subject to NMD, meeting the requirement of having the stop codon more than 50 nucleotides to the 5' end of the last exon junction. The other transcripts, while short, have their stop codon in the final exon.

However, while the previously described model can explain some of the NMD seen, it is incomplete because some transcripts with early stop codons are translated normally and others that have a stop codon in the very last exon are still subject to NMD. A second model, called the *faux* UTR model, argues that NMD occurs when the translation by the ribosome fails to terminate near a correctly configured 3'-UTR. This model can help explain why, in both yeast and *D. melanogaster*, premature stop codons can be detected by the cell machinery even in transcripts that lack introns, and thus lack exon junction complexes (Metzstein and Krasnow 2006). Our ability to detect the various NvDM1 transcripts at any significant level suggests that they may not be linked closely with NMD as they are present at high enough levels to be identified.

According to work done by Lewis *et al.* premature stop codons may be an underappreciated form of protein expression regulation (2003). NMD is thought to primarily target transcripts that result from genetic mutations, with several human diseases associated with this phenomenon, or those that are the result of transcription errors or abnormal RNA splicing events (Metzstein and Krasnow 2006). However, evidence also suggests that it is involved in the normal regulation of protein levels via alternative splicing (Lewis *et al.* 2003). In a study of alternative splicing in humans, Lewis and colleagues found 1,106 alternatively spliced genes that produce 1,989 transcripts that appear to be targets of NMD (2003). Such significant numbers suggest that combining alternative splicing with NMD allows for further regulation of protein levels. However, more recent work done by Pan *et al.* with alternative splicing microarrays, suggests that, while about 35% of mammalian alternative splicing events result in a splice variant with a PTC, they are transcribed at a very low level to begin with and thus NMD does not play a significant role. Thus, while NMD may play a role in the regulation of NvDM1, is it unclear whether this role could be significant.

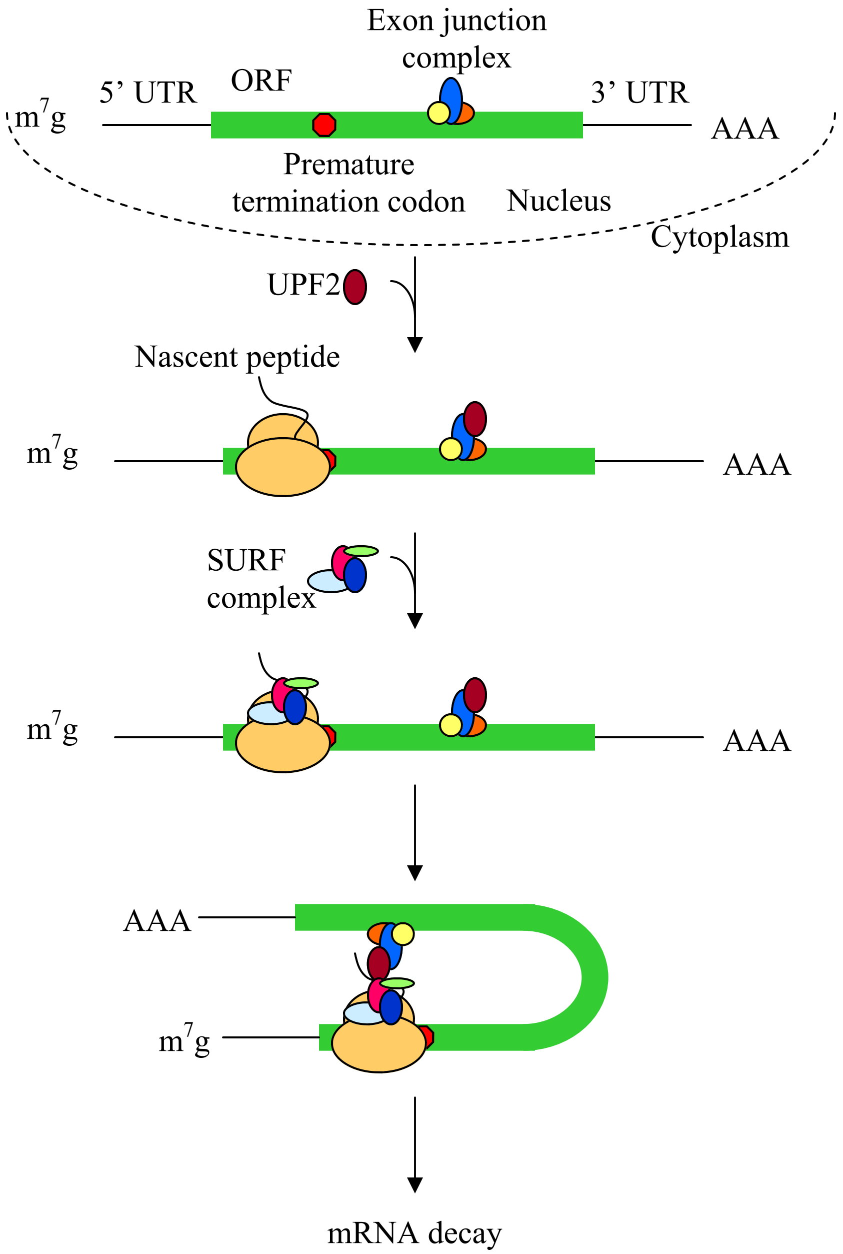


Figure 69. The current model of mammalian nonsense-mediated decay (NMD). After the transcript is spliced in the nucleus, the exon junction complexes (EJCs) associate with each exon-exon junction. Once moved into the cytoplasm, the mRNA is translated for the first time. However, if a premature termination codon (PTC) is present, not all EJCs will be removed and the ribosome will stall, resulting in the recruitment of proteins to the exon EJC and the ribosome. This marks the transcript for decay (adapted from Garneau *et al*. 2007).

Sex-specificity in NvDM4

Among the four genes in the DM family of *N. vitripennis*, NvDM4 is the best candidate for a doublesex ortholog. NvDM4 has a number of different splice forms in yellow pupae that fall into four major classes, differentiated by variations in the oligomerization domain. Of the four classes of transcripts, OD-B appears to be female-specific, with all transcripts from this class isolated from female RACE products and RT-PCR showing products of the correct-sized band only in females. OD-C appears to be male-specific; the transcripts in this class were isolated from male RACE reactions and, while it is expressed in both sexes, crude quantitation with RT-PCR suggests that OD-C transcripts are transcribed at higher levels in males. Two other transcripts, OD-A and OD-X, are transcribed in both sexes and do not appear to be sex-specifically expressed.

I aligned the predicted polypeptides of NvDM4 the *doublesex* homologs of *A. mellifera*, *D. melanogaster*, and the silk moth *Bombyx mori* (fig. 70A, B, C). Looking at the female-specific region, one notes fairly little similarity (fig. 70 B), which is comparable to what is seen in *A. mellifera* (Cho *et al*. 2007). Of the 15 amino acids that make up the female-specific portion of the oligomerization domain, only one residue in the OD-B protein from *N. vitripennis* is identical to any one of the other three (both *N. vitripennis* and *B. mori* share a K at the second amino acid in the female-specific region), and just four similarities. It should be noted that the honeybee has only one identity and just three similarities. Work done by Yang and colleagues suggest that a tyrosine (third residue in the female-specific OD) and asparagine (sixth residue) are critical in *D. melanogaster* for the binding of intersex (*ix*), a protein important for the correct phenotypic development of females (2008). While *N.*

vitripennis has a similar residue at the sixth position, it shows no such similarity at the third position. Yet, neither does *A. mellifera* or *B. mori*.

This suggests a couple of possibilities. It may be that intersex does not interact with these other *dsx* orthologs. However, intersex itself has been shown to be highly conserved, with the *ix* genes from dipterans and lepidopterans able to restore sexual differentiation to *ix* null mutants of *D. melanogaster* (Siegal and Baker 2005). Even an *ix* ortholog from mice was able to partially rescue the mutants, suggesting that intersex is functionally conserved (Siegal and Baker 2005). Furthermore, performing a BLASTP search of the predicted *N. vitripennis* gene sets, I was able to find an intersex ortholog in the *N. vitripennis* RefSeq database (results not shown). Both of these would suggest *ix* would be interacting with the *dsx* homolog in *N. vitripennis*. The second possibility is that *ix* is interacting with other amino acids in the female-specific region of the gene. In fact, Yang and his colleagues note that the alanine scanning mutagenesis technique used to show the importance of the asparagines and tyrosine can underestimate the size of contacts between proteins, which leaves open the opportunity for interactions with other amino acids (2008).

Also of interest in females is the C-terminus region following the oligomerization domain. In the *N. vitripennis* OD-B protein, this consists of only a single residue. While a shortened C-terminal could interfere with the protein's ability to bind *ix*, deletion studies in *D. melanogaster* have shown that *ix* is still able to bind to *dsx* when the entire sequence after the oligomerization domain has been deleted (Yang *et al*. 2008). This would suggest the shortened C-terminal region in the OD-B protein would not affect interactions with *ix*.

Two of the alternatively spliced transcripts, OD-A and OD-X, do not appear to be sex-specific in yellow pupae at the level of the entire organism. OD-A contains a truncated oligomerization domain, while OD-X has only the DM domains. In the alternatively spliced *dsx* homolog of *A. mellifera*, one transcript is not sex-specific. Like OD-X, it contains a DM domain, but lacks an oligomerization domain. Whether these shortened transcripts code for functional proteins has not been determined.

Splice Sites

An examination of how the pre-mRNA of *doublesex* is differentially spliced can help us postulate about what may be occurring in *N. vitripennis*. In the canonical *doublesex* of *D. melanogaster* as well as the sex-specifically spliced ortholog in *A. gambiae*, the male splice pattern is the default transcript (Lynch and Maniatis 1995; Scali *et al*. 2005) and the female-specific splice site is weakened by nearby purine nucleotides (Cho *et al*. 2007). In contrast, both *B. mori* and *A. mellifera* appear to have the female transcript as default (Cho *et al.* 2007). The default splice pattern refers to the pattern of pre-mRNA processing that occurs without the presence of genes known to affect sexual differentiation, requiring only general splicing machinery (Nagoshi and Baker 1990). In *D. melanogaster*, for example, the transformer protein (*tra*) is known to regulate sexual differentiation. If *tra* is lost in genetically female flies due to mutations, the flies will express the male *dsx* splice pattern (Nagoshi and Baker 1990). After examining both computational and molecular evidence as described below, I was unable to propose a specific transcript as the default splicing pattern in *N. vitripennis*.

Female specific-splicing in both fruit flies and mosquitoes is dependent on the interaction of TRA with two doublesex-specific splicing enhancers: the doublesex repeat element (dsxRE) and the purine rich enhancer (PRE) (Lynch and Maniatis 1995). Both of these sequences are located within the exon that codes for the female-specific region of the oligomerization domain (Fig. 65). On the other hand, neither *B. mori* nor *A. mellifera* possess these splicing enhancers (Cho *et al.* 2007). In both, the female-specific splice site is not weakened and the female transcript is the default sequence. *Nasonia vitripennis* appears to fall into this latter category with respect to splice site strength and signaling. It lacks both of the splicing enhancers and the female-specific exon 4b of class OD-B does not seem to have a weakened splice site (table 7). This would suggest that, like the silk moth and honeybee, *N. vitripennis* has the female transcript as default splicing pattern.

However, unlike what is seen in *A. mellifera*, the results of RT-PCR in *N. vitripennis* do not confirm this proposition. In the case of *A. mellifera*, the hypothesis of the female transcript as default is further supported by the fact that female transcripts were detected in male bees at low levels in some stages of development using RT-PCR, suggesting there is incomplete repression of this default splicing pattern. In contrast, I saw no evidence of class OD-B in males using RT-PCR (Fig. 64). Instead, the OD-C transcripts, the male-specific transcript, are present at a low level in female yellow pupae (Fig. 60). This suggests that the default transcript would be male, not female. Therefore, I cannot draw a conclusion as to which form of NvDM4 is the default splice pattern. The RT-PCR work tentatively suggests the male transcript is the default splicing pattern. However, the strength of the female splice site (and relative weakness of the male splice site) along with the lack of the splice enhancers dsxRE and PRE indicate that the female transcript represents the default pattern.

Nagoshi and his colleagues showed that the male transcript of *dsx* of *D. melanogaster* was the default transcript through mutation studies (Nagoshi *et al.* 1988 and 1990). In *B. mori*, the female transcript was shown to be the default version of the splicing pattern by examining the splicing *in vitro* using HeLa cell nuclear extracts (Suzuki *et al*. 2001). No such work has been done in bees. Thus, in the discussion of their results, Cho and his colleagues note that, while circumstantial evidence suggests that the female transcript is the default splicing pattern in *A. mellifera*, either male or female transcripts may serve as default in the honey bee (2007). In order to draw firm conclusions about splicing in *N. vitripennis*, further studies need to be conducted. At present, it is not possible to determine what the default splice pattern is for NvDM4 in *N. vitripennis*.

The male-specific regions of the proteins are also of interest. Previous studies have suggested that the male-specific region of the C-terminal diverged rapidly, with few identities between difference species in the male-specific region of the oligomerization domain (Cho *et al*. 2007). While this may be the case generally, among the insects compared here the male-specific region of the OD shows a higher degree of conservation than the female-specific region. Four of the fifteen amino acids in *D. melanogaster* are identical to those in *N. vitripennis*, and three more amino acids in the wasp are similar to either the silk moth or the fruit fly. Among *N. vitripennis* and *A. mellifera*, there are even some conserved regions in the C-terminal region beyond the oligomerization domain. Any biological importance of this similarity is impossible to decipher without further study; the bulk of the research in protein structure has focused on the female version of doublesex, and so relatively little is known about the male-specific region.

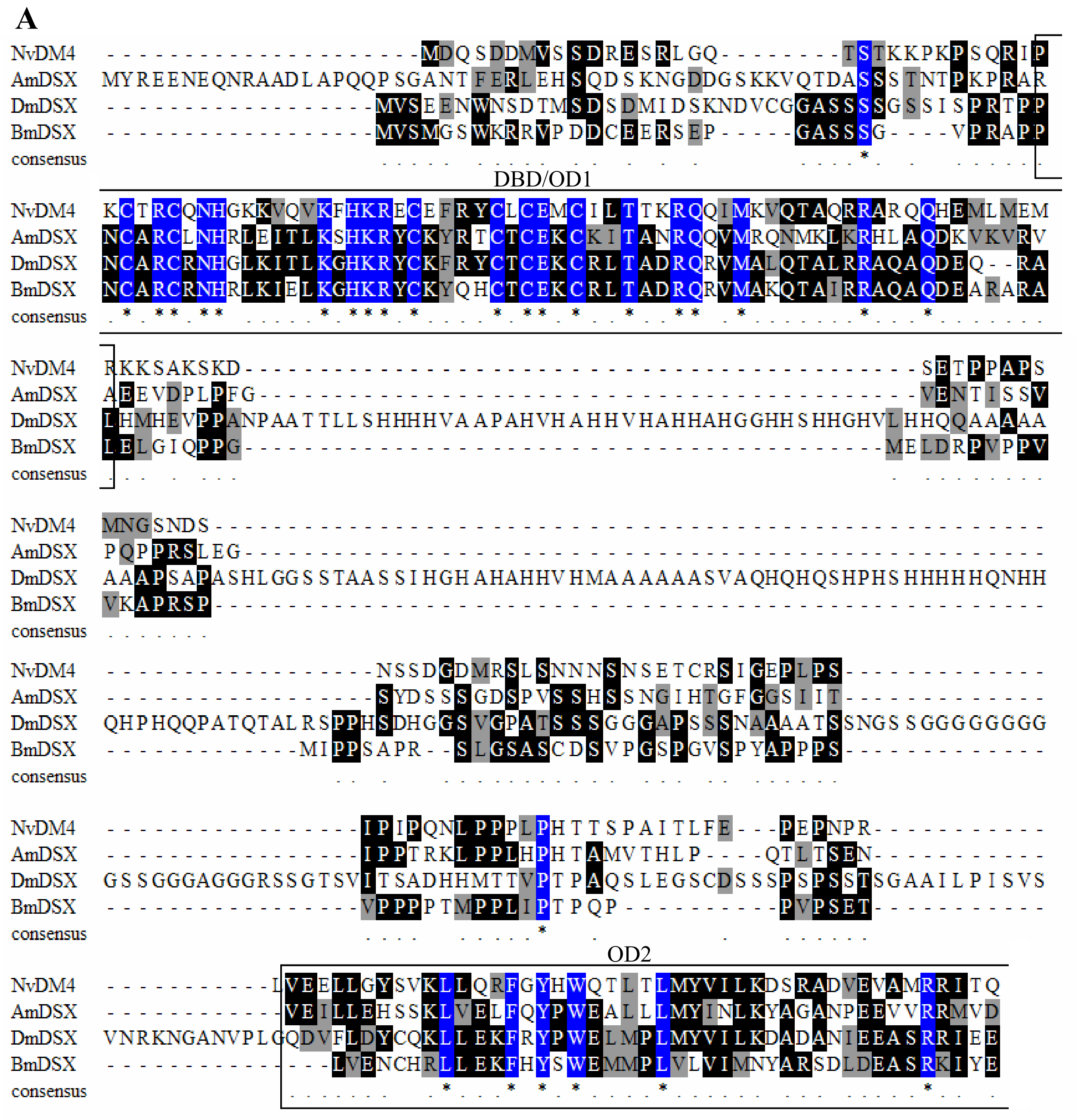


Figure 70A. See following page for description.

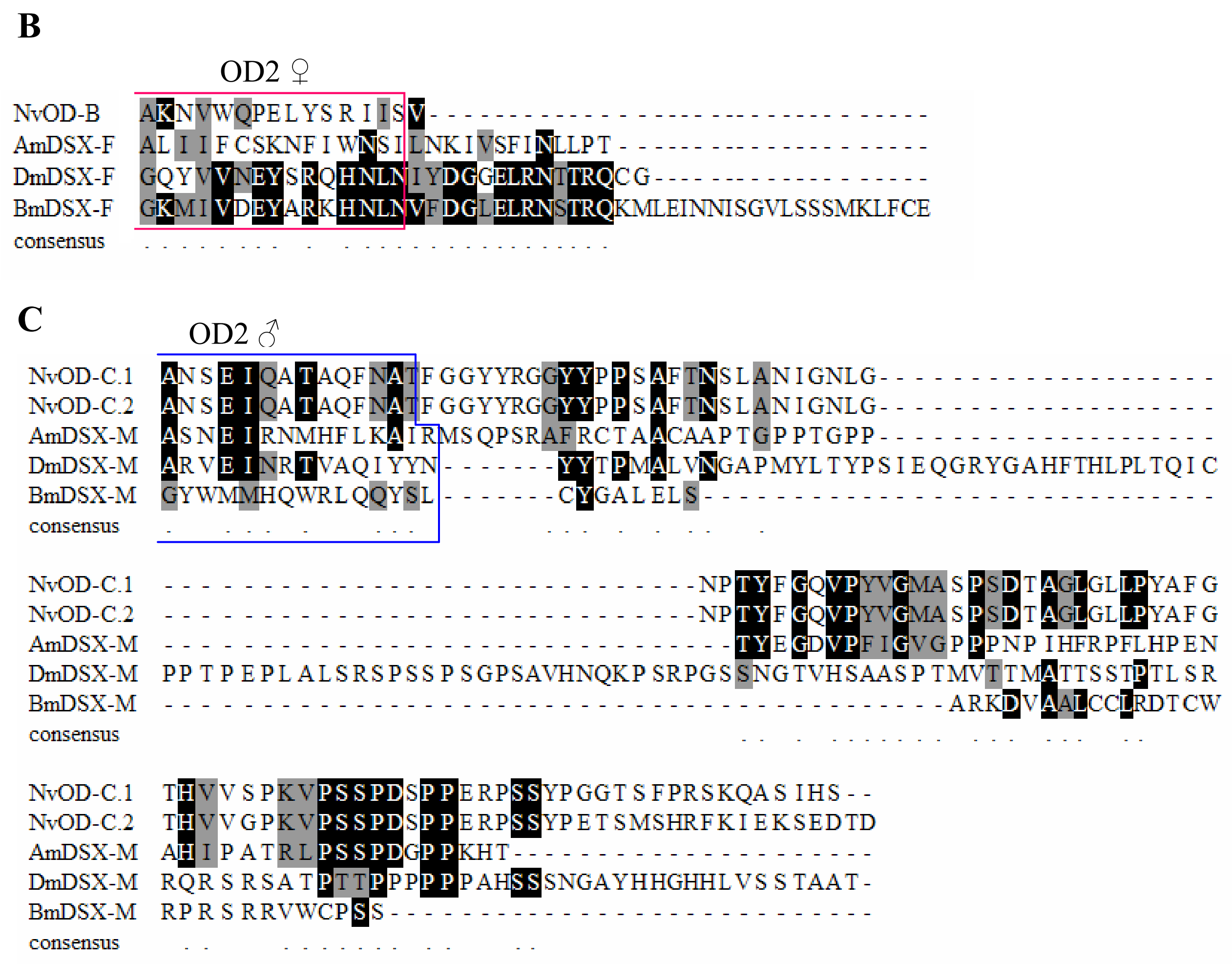

B
OD2 ♀
NvOD-B AKNVWQPELYSRIISV-
AmDSX-F ALIIFCSKNFIWNSILNKIVSFINLLPT-
DmDSX-F GQYVVNEYSRQHNLNIYDGGELRNITRQCG-
BmDSX-F GKMIVDEYARKHNLNVFDGLELRNSTRQKMLEINNISGVLSSSMKLFCE
consensus
C
OD2 ♂
NvOD-C.1 ANSEIQATAQFNATFGGYYRGGYYPPSAFTNSLANIGNLG-
NvOD-C.2 ANSEIQATAQFNATFGGYYRGGYYPPSAFTNSLANIGNLG-
AmDSX-M ASNEIRNMHFLKAIRMSQPSRAFRCTAACAAPTGPPTGPP-
DmDSX-M ARVEINRTVAQIYYN------YYTPMALVNGAPMYLTYPSIEQGRYGAHFTHLPLTQIC
BmDSX-M GYWMMHQWRLQQYSL------CYGALELS-
consensus
NvOD-C.1 -NPTYFGQVPYVGMASPSDTAGLGLLPYAFG
NvOD-C.2 -NPTYFGQVPYVGMASPSDTAGLGLLPYAFG
AmDSX-M -TYEGDVPFIGVGPPPNPIHFRPFLHPEN
DmDSX-M PPTPEPLALSRSPSSPSGPSAVHNQKPSRPGSSNGTVHSAASPTMVTTMATTSSTPTLSR
BmDSX-M -ARKDVAALCCLRDTCW
consensus
NvOD-C.1 THVVSPKVPSSPDSPPERPSSYPGGTSFPRSKQASIHS--
NvOD-C.2 THVVGPKVPSSPDSPPERPSSYPETSMSHRFKIEKSEDTD
AmDSX-M AHIPATRLPSSPDGPPKHT-
DmDSX-M RQRSRSATPTTPPPPPPAHSSSNGAYHHGHHLVSSTAAT-
BmDSX-M RPRSRRVWCPSS-
consensus

Figure 70. Amino acid sequence alignment of NvDM4 with *A. mellifera* (Am), *D. melanogaster* (Dm), and *B. mori* (Bm). (A) is non-sex specific region of each gene, (B) is the female-specific C-terminal, and (C) is the male-specific C-terminal. The DNA binding domain (DBD)/oligomerization domain 1 (OD1) and non-sex-specific OD2 are indicated with black boxes. The sex-specific regions of OD2 are indicated by pink and blue boxes for males and females respectively.

Future Studies

My thesis research provides many opportunities for future studies. My work showed that the DM gene family in *N. vitripennis* consists of four members and one, NvDM4, is a homolog of *doublesex*. The NvDM4 gene is alternatively spliced, with four different classes of transcripts. Two of these show sex-specific expression. My discoveries can provide a basis for furthering out understanding of sex-determination in *N. vitripennis*.

One avenue available for exploration is the examination of these genes at different developmental stages, with particular focus on the four transcript classes of NvDM4. In my own research, I worked with the yellow pupal stage, the first stage at which individuals can be differentiated by sex. Future work should look at transcription at other stages of development ranging from embryo to adult. At stages prior to the yellow pupae, it will be difficult to isolate a pool of exclusively female RNA. However, a pool of just males can be acquired from virgin females due to the haplodiploid system of sex determination and, under proper conditions, mated females will produce mostly female offspring. This would provide two sources of RNA in which to compare transcript expression. Also, tissue-specific expression might be examined. As discussed previously, the *D. melanogaster* gene dmrt93B has shown sex-specific expression in the fly body, but not its head, a nuance that would not have been picked up following my protocol (Wen 2002).

Future research could complete the transcripts of NvDM1, NvDM2, and NvDM3. I did not determine the 5' untranslated regions of these genes or the 3' untranslated region of NvDM3. Completing the 3' end of the transcripts would allow future work to include RNA interference (RNAi) of these genes, which could elucidate the functional roles of these genes.

Because relatively little work has been done on the DM-containing genes not involved in sex-determination, this would provide insight into other roles members of this family can play.

While my work has shown sex-specific splicing of NvDM4, this does not guarantee that the gene plays a role in sex-determination, although the similarities observed between NvDM4 and other *doublesex* homologs strongly support this possibility. In order to show definitively that NvDM4 is a *dsx* ortholog, further experiments are needed. This would require observing the phenotypes in wasps that lack a functional NvDM4 gene. Such experiments might include RNAi, although, because of variability in the 3' UTR where small interfering RNA typically targets, as well as the sheer number of transcripts, this could be a complex proposition.

If NvDM4 expression was inhibited, it is likely that the result would be a wasp with an intersexual phenotype as seen in *D. melanogaster doublesex* (Waterbury *et al*. 1999). Male fruit flies lacking *dsx* have lightly pigmented abdomens, improperly formed male-specific sex-comb teeth, and malformed genitalia in addition to expressing the female-specific yolk protein at low levels. Females that do not have a functional *dsx* also have improperly formed sex-comb teeth, along with a darkened abdomen, malformed genitalia and abnormally low levels of yolk protein (Waterbury *et al*. 1999). In light of these observations in *D. melanogaster*, if NvDM4 is involved in sex-determination, one would expect an intersexual phenotype from a successful knockout. This might include such phenotypic alterations as variation in wing length, as *N. vitripennis* males have markedly shortened wings relative to females, changes in antennae, with females typically have thicker and darker antennae, or genitalia, because wild type males lack the ovipositor seen in females (Kamping *et al.* 2007).

To further establish the level of conservation between NvDM4 and *dsx* one could see if NvDM4 would rescue *dsx* null fruit flies, which would suggest a high degree of conservation. Previous experiments have shown that the male version of DSX from *D. melanogaster* can rescue *mab-3* mutants of *C. elegans* (Waterbury *et al*. 1999). As *D. melanogaster* and *N. vitripennis* are more closely related than *D. melanogaster* and *C. elegans*, it seems plausible that NvDM4 could at least partially rescue mutant *D. melanogaster*.

Future studies are also needed to follow up on the differences I observed in transcript expression between the sexes. By varying the cycling numbers, I was able to make a rudimentary analysis of the differences in level of transcription between the sexes for the OD-C transcripts. It would be useful to do either quantitative PCR (Q-PCR) or Northern blots to determine the transcription levels more accurately. These techniques could also be used to check the levels of transcription of the other NvDM4 transcripts. For example, using a Northern blot to look for the OD-B transcripts would confirm whether this class is female-specific.

Another avenue that should be investigated is molecular mechanism behind the alternative splicing pattern observed in NvDM4. By understanding the regulation of NvDM4, it might be possible to determine the next step up in the sex-determination hierarchy. These studies might include attempting to establish splicing *in vitro* using HeLa as was done in *B. mori* (Cho *et al.* 2007). With this information, it could be determined how NvDM4 is spliced in the absence of sex-specific regulators. For example, if the results indicate the male transcript is the default splice pattern as seen in *D. melanogaster*, it would be worthwhile to search for a *tra* ortholog in *N. vitripennis*, although I failed to find an obvious *tra* ortholog in a preliminary search of the available *N. vitripennis* databases.

The DM gene family in *Nasonia vitripennis* provides many possibilities for future study. Of particular interest is work to be done with NvDM4, the *doublesex* homolog. Work done with this gene will contribute to our understanding of the regulation of sexual dimorphism, both within *N. vitripennis* and hymenopterans and in the broader context of conservation amidst the diversity of hierarchical regulators in the sex determination pathways.

LITERATURE CITED

An W, Cho S, Ishii H, and Wensink P. Sex-specific and non-sex-specific oligomerization domains in both of the doublesex transcription factors from *Drosophila melanogaster*. Molecular and Cellular Biology. 16:6 (1996): 3106 – 3111.

An W and Wensink PC. Integrating sex- and tissue-specific regulation within a single Drosophila enhancer. Genes Dev. 9: 2 (1995): 256-66.

Altschul S, Gish W, Miller W, Myers E, and Lipman D. Basic local alignment search tool. J. Mol. Biol. 215 (1990):403-410.

Ast G. How did alternative splicing evolve? Nat Rev Genet. 5:10 (2004): 773 – 782.

Bayrer JR, Zhang W, and Weiss MA. Dimerization of Doublesex Is Mediated by a Cryptic Ubiquitin-associated Domain Fold: Implications for sex-specific gene regulation. J Biol Chem. 280:38 (2005): 32989 – 96.

Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, and Wheeler DL. GenBank. Nucleic Acids Res. 35 (2007): D21-5.

Beye M. The dice of fate: the *csd* gene and how its allelic composition regulates sexual development in the honey bee, *Apis mellifera*. BioEssays. 26 (2004): 1131 – 1139.

Bond RW, Wyborski RJ, and Gottlieb DI. Developmentally regulated expression of an exon containing a stop codon in the gene for glutamic acid decarboxylase. Proc Natl Acad Sci. 87(1990): 8771 – 8775.

Bonizzoni P, Rizzi R, and Pesole G. ASPIC: a novel method to predict the exon-intron structure of a gene that is optimally compatible to a set of transcript sequences. BMC Bioinformatics. 6 (2005): 244.

Brody TB. *Ultrabithorax*: Protein interactions. The Inter*active* Fly. 1995. http://sdbonline.org/fly/segment/ultrabt4.htm. Accessed on 5/9/08.

Brown TA. Genomes, second ed. John Wiley and Sons, Inc. (2002): 470 – 471.

Bull JJ. Evolution of Sex Determining Mechanisms. The Benjamin/Cummings Publishing Company, Inc. (1983).

Carroll SB, Grenier JK, and Weatherbee SD. From DNA to Diversity: molecular genetics and the evolution of animal design. Blackwell Science. (2001): 18 – 121.

Chenna R, Sugawara H, Koike T, Lopez R, Gibson TJ, Higgins DG, and Thompson JD.

Multiple sequence alignment with the Clustal series of programs. Nucleic Acids Res. 31: 13 (2003): 3497-500

Chessler SD and Lernmark A. Alternative Splicing of GAD67 Results in the Synthesis of a Third Form of Glutamic-acid Decarboxylase in Human Islets and Other Non-neural Tissues. J Biol Chem. 275:7 (2000): 5188-5192.

Cho S, Huang ZY, Green DR, Smith DR, and Zhang J. Evolution of the complementary sex-determination gene of honey bees: Balancing selection and trans-species polymorphisms. Genome Res. 16:11 (2006):1366-75.

Cook JM and Crozier RH. Sex determination and population biology in Hymenoptera. *Trends Ecol. Evol.* 10 (1995): 281 – 286.

Cristino A, Mendes do Nascimiento A, da Fontoura Costa L, and Luz Paulino Simões Z. A comparative analysis of highly conserved sex-determining genes Apis mellifera and Drosophila melanogaster. Genetics and Molecular Research. 5 (2006).

Emani KH, Burke TW, and Smale ST. Sp1 activation of a TATA-less promoter requires a species-specific interaction involving transcription factor IID. 26:3 (1998): 839–846.

Erdman SE, Chen H, and Burtis KC. Functional and genetic characterization of the oligomerization and DNA binding properties of the Drosophila doublesex proteins. Genetics. 144 (1996): 1639 – 1652.

Evans JD, Shearman DCA, and Oldroyd BP. Molecular basis for sex determination in haplodiploids. Trends in Eco and Evo. 19:1 (2004): 1 – 3.

Garneau NL, Wilusz J, and Wilusz CJ. The highways and byways of mRNA decay. Nat Rev Mol Cell Biol. 8:2 (2007): 113 – 26.

Gasteiger E, Gattiker A, Hoogland C, Ivanyi I, Appel RD, and Bairoch A. ExPASy: the proteomics server for in-depth protein knowledge and analysis. Nucleic Acids Res. 31(2003): 3784-3788.

Gilbert SF. Developmental Biology, 8th ed. Sinauer Associates, Inc. (2006).

Greive SJ and von Hippel PH. Thinking quantitatively about transcriptional regulation. Nat. Rev. Mol. Cell Biol. 6 (2005): 221 – 232.

Haig D. Mother's boy or daddy's girl? Sex determination in Hymenoptera. *Trends Ecol. Evol.* 13 (1998): 380 – 381.

Hodgkin J. The remarkable ubiquity of DM domain factors as regulators of sexual phenotype: ancestry or aptitude? Genes and Dev. 16(2002): 2322 – 2326.

Huang J and Brutlag D. The EMOTIF database. Nucleic Acids Research. 29:1 (2001): 202 – 4.

Inoue K, Hoshijima K, Higuchi I, Sakamoto H, and Shimura Y. Binding the *Drosophila* transformer and transformer-2 proteins to the regulatory elements of doublesex primary transcript for sex-specific RNA processing. Proc. Natl. Acad. Sci. 89 (1992): 8092 – 8096.

Irimia M, Penny D, and Roy SW. Coevolution of genomic intron number and splice sites. Trends Genet. 23:7 (2007): 321 – 325.

Kaiser P, Su NY, Yen JL, Ouni I, and Flick K. The yeast ubiquitin ligase SCFMet30: connecting environmental and intracellular conditions to cell division. Cell Div. 1:16 (2006).

Kamping A, Katju V, Beukeboom LW, and Werren JH. Inheritance of gynandromorphism in the parasitic wasp *Nasonia vitripennis*. Genetics. 5:3 (2007): 1321 – 33.

Kato Y, Kobayashi K, Oda S, Colbourn JK, Tatarazako N, Watanabe H, and Iguchi T. Molecular cloning and sexually dimorphic expression of DM-domain genes in *Daphnia magna*. Genomics. 91(2008): 94 – 101.

Lee TI and Young RA. Transcription of eukaryotic protein-coding genes. Annu. Rev. Genet. 34 (2000): 77–137.

Latchman DS. Transcription factors: An overview. Int. J. Biochem. Cell Biol. 29: 12 (1997): 1305 – 1312.

Lewin B. Genes VIII. Pearson Prentice Hall. (2004): 95 – 98.

Lewis BP, Green RE, and Brenner SE. Evidence for the widespread coupling of alternative splicing and nonsense-mediated mRNA decay in humans. Proc. Natl. Acad. Sci. 100: 1 (2003):189-192.

Louis EJ. Evolutionary genetics: Making the most of redundancy. Nature. 449 (2007): 673-674.

Luscombe NM, Austin SE, Berman HM, and Thornton JM. An overview of the structures of protein-DNA complexes. Genome Biol. 1:1 (2000).

Lynch KW and Maniatis T. Synergistic interactions between two distinct elements of a regulated splicing enhancer. Genes Dev. 9: 3 (1995): 284 – 293.

Manolakou P, Lavranos G, and Angelopoulou R. Molecular patterns of sex determination in the animal kingdom: a comparative study of the biology of reproduction. Reproductive Biology and Endocrinology. 4: 59 (2006).

Meijer HA, Radford HE, Wilson LS, Lissenden S, and de Moor CH. Translational control of maskin mRNA by its 3' untranslated region. Biol Cell. 99:5 (2007): 239 – 250.

Meng A, Moore B, Tang H, Yuan B, and Lin S. A Drosophila doublesex-related gene, terra, is involved in somitogenesis in vertebrates. Development. 126 (1999): 1259 – 1268.

Metzstein MM and Krasnow MA. Functions of the Nonsense-Mediated mRNA Decay Pathway in Drosophila Development. PLOS Genet. 2:12 (2006): 2143 – 2154.

Miller SW, Hayward DC, Bunch TA, Miller DJ, Ball EE, Bardwell VJ, Zarkower D, and Brower DL. A DM domain protein from a coral, *Acropora millepora*, homologous to proteins important for sex determination. Evol Dev. 5:3 (2003): 251 – 258.

Modrek B and Lee C. A genomic view of alternative splicing. Nature Genetics. 30 (2002): 13 – 19.

Nagoshi RN and Baker BS. Regulation of sex-specific RNA splicing at the Drosophila doublesex gene: cis-acting mutations in exon sequences alter sex-specific RNA splicing patterns. Genes Dev. 4 (1990): 89 – 97.

Nagoshi RN, McKeown M, Burtis KC, Belote JM, and Baker BS. The control of alternative splicing at genes regulating sexual differentiation in D. melanogaster. Cell. 53 (1988): 229 – 236.

National Center of Biotechnology Information. http://www.ncbi.nlm.nih.gov/

NCBI. Wasp Genome Resource. http://www.ncbi.nlm.nih.gov/projects/genome/guide/wasp/

Nekrutenko A and Li W. Transposable elements are found in a large number of human protein-coding regions. Trends in Genetics. 17: 11 (2001): 619 – 621.

Neron B, Tuffery P, and Letondal C. Mobyle: a Web portal framework for bioinformatics analyses, poster presented at NETTAB 2005.

Nevill-Manning C, Wu T, and Brutlag D. Highly specific protein sequence motifs for genome analysis. Proc. Natl. Acad. Sci. 95 (1998): 5865 – 5871.

Nikolaev VK, Leontovich AM, Drachev VA, and Brodsky LI. Building multiple alignment using iterative analyzing biopolymers structure dynamic improvement of the initial motif alignment. Biochemistry. 62:6 (1997): 578-582.

Oh CS, Toke DA, Mandala S, and Martin CE. ELO2 and ELO3, homologues of the Saccharomyces cerevisiae ELO1 gene, function in fatty acid elongation and are required for sphingolipid formation. J Biol Chem. 272:28 (1997):17376 – 84.

Patel AA and Steitz JA. Splicing double: Insights from the second spliceosome. Nat. Rev. Mol. Cell Biol. 4: 12 (2003): 960 – 970.

Pertea M, Lin X, and Salzberg S. GeneSplicer: a new computational method for splice site prediction. Nucleic Acids Research. 29: 5 (2001): 1185 – 1190.

Prince VE and Pickett B. Splitting pairs: the diverging fates of duplicated genes. Nat. Rev. Genet. 3 (2002): 827 – 837.

Raymond S, Shamu C, Shen M, Seifert K, Hirsch B, Hodgkin J, and Zarkower D. Evidence for evolutionary conservation of sex-determining genes. Nature. 39 (1998): 691 – 695.

Reese MG, Eeckman FH, Kulp D, and Haussler D. Improved splice site detection in Genie. J Comput Biol. 4:3(1997): 311 – 323.

Reik W and Walkter J. Genomic imprinting: parental influence on the genome. Nat. Rev. Genet. 2: (2001): 21 – 32.

Retelska D, Iseli C, Bucher P, Jongeneel CV, and Naef F. Similarities and differences of polyadenylation signals in human and fly. BMC Genomics. 7: 176 (2006).

Rice Annotation Project, Tanaka T, Antonio BA, Kikuchi S, Matsumoto T, Nagamura Y, Numa H, Sakai H, Wu J, Itoh T, Sasaki T, Aono R, Fujii Y, Habara T, Harada E, Kanno M, Kawahara Y, Kawashima H, Kubooka H, Matsuya A, Nakaoka H, Saichi N, Sanbonmatsu R, Sato Y, Shinso Y, Suzuki M, Takeda J, Tanino M, Todokoro F, Yamaguchi K, Yamamoto N, Yamasaki C, Imanishi T, Okido T, Tada M, Ikeo K, Tateno Y, Gojobori T, Lin YC, Wei FJ, Hsing YI, Zhao Q, Han B, Kramer MR, McCombie RW, Lonsdale D, O'Donovan CC, Whitfield EJ, Apweiler R, Koyanagi KO, Khurana JP, Raghuvanshi S, Singh NK, Tyagi AK, Haberer G, Fujisawa M, Hosokawa S, Ito Y, Ikawa H, Shibata M, Yamamoto M, Bruskiewich RM, Hoen DR, Bureau TE, Namiki N, Ohyanagi H, Sakai Y, Nobushima S, Sakata K, Barrero RA, Sato Y, Souvorov A, Smith-White B, Tatusova T, An S, An G, OOta S, Fuks G, Fuks G, Messing J, Christie KR, Lieberherr D, Kim H, Zuccolo A, Wing RA, Nobuta K, Green PJ, Lu C, Meyers BC, Chaparro C, Piegu B, Panaud O, and Echeverria M. The Rice Annotation Project Database (RAP-DB): 2008 update. Nucleic Acids Res. 2008 Jan;36(Database issue):D1028-33.

Roca X, Sachidanandam R, and Krainer AR. Determinants of the inherent strength of human 5' splice sites. RNA. 11:5 (2005): 683 – 698.

Rong YS, Titen SW, Xie HB, Golic MM, Bastiani M, Bandyopadhyay P, Olivera BM, Brodsky M, Rubin GM, and Golic KG. Targeted mutagenesis by homologous recombination in D. melanogaster. Genes Dev. 16: 12 (2002): 1568 – 81.

Roesner A, Fuchs C, Hankeln T, and Burmester T. A globin gene of ancient origin in lower vertebrates: evidence of two distinct globin families in animals. Molec Biol and Evo. 22:1 (2005): 667 – 677.

Rozen S and Skaletsky HJ. Primer3 on the WWW for general users and for biologist programmers. In: Krawetz S, Misener S (eds) Bioinformatics Methods and Protocols: Methods in Molecular Biology. Humana Press, Totowa, NJ. (2000) 365-386

Scali C, Catteruccia F, Li Q, and Crisanti A. Identification of sex-specific transcripts of the Anopheles gambiae doublesex gene. J Exp Biol. 208(2005): 3701 – 9.

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

Sonhammer E, Eddy S, Birney D, Bateman A, and Durbin R. Pfam: multiple sequence alignments and HMM-profiles of protein domains. Nuc Acid Res. 26:1 (1998): 320-322.

Steinmann-Zwicky M. Sex determination of the Drosophila germ line: *tra* and *dsx* control somatic inductive signals. Development 120: 3 (1994): 707 – 716.

Stroeher VL, Gaiser C, and Garber RL. Alternative RNA splicing that is spatially regulated: generation of transcripts from the antennapedia gene of *Drosophila melanogaster* with different protein-coding regions. Molec and Cell Biol. 8:10 (1988): 4143 – 4154.

Taneri B, Snyder B, Novoradovsky A, and Gaasterland T. Alternative splicing of mouse transcription factors affects their DNA-binding domain architecture and is tissue specific. Genome Biol. 5: 10 (2004): R75.

Trent C, Crosby C, and Eavey J. Additional evidence for the genomic imprinting model of sex determination in the haplodiploid wasp *Nasonia vitripennis*: isolation of biparental diploid males after X-ray mutagenesis. Heredity. (2006).

Tour E, Hittinger CT, and McGinnis W. Evolutionarily conserved domains required for activation and repression functions of the *Drosophila* Hox protein Ultrabithorax. Development. 132: 23 (2005): 5271 – 81.

van Wilgenburn E, Driessen G, and Beukeboom L. Single locus complementary sex determination in Hymenoptera: an "unintelligent" design? Frontiers in Zoology. 3: 1 (2006)

Voet D and Voet JG. Biochemistry. John Wiley and Sons, Inc. (2004): 314 – 315, 1269 – 1271.

Wagner E and Lykke-Andersen J. mRNA surveillance: the perfect persist. J of Cell Sci. 115: 15 (2002): 3033 – 3038.

Waterbury JA, Jackson LL, and Schedl P. Analysis of the doublesex female protein in *Drosophila melanogaster*: Role in sexual differentiation and behavior and dependence on intersex. Genetics. 152:4 (1999): 1653 – 67.

Watson JD, Baker TA, Bell SP, Gann A, Levine M and Losick R. Molecular Biology of the Gene, 6th ed. Pearson (2008).

Wen S. Analysis of *dmrt3B*, a *doublesex* related gene, suggests that it play a role in sexual differentiation. Unpublished Master's thesis. University of Texas Health Science Center at Houston. (2002).

Werren J. http://www.rochester.edu/College/BIO/labs/WerrenLab/nasonia/development time.html

Werren J. Personal communication. (2006)

Winkler C, Hornung U, Kondo M, Neuner C, Duschk J, Shima A, and Schartl M. Developmentally regulated and non-sex-specific expression of autosomal *dmrt* genes in embryos of Medaka fish (*Oryzias latipes*). Mechanisms of Development. 121 (2004): 997 – 1005.

Yang Y, Zhang W, Bayrer JR, and Weiss MA. Doublesex and the regulation of sexual dimorphism in *Drosophila melanogaster*: Structure, function, and mutagenesis of a female-specific domain. J Biol Chem. 283:11 (2008): 7280 – 7292.

Ying X and Lee C. Alternative splicing and RNA selection pressure – evolutionary consequences for eukaryotic genomes. Nature Reviews Genetics. 7 (2006): 499 – 509.

Zarkower D. Establishing sexual dimorphism: conservation amidst diversity. Nature Reviews: Genetics. 2 (2001): 175 – 185.

Zavolan M, Kondo S, Schonbach C, Adachi J, Hume DA, Hayashizaki Y, Gaasterland T, RIKEN GER Group, and GSL Members. Impact of alternative initiation, splicing, and termination on the diversity of the mRNA transcripts encoded by the mouse transcriptome. Genome Res. 13:6B (2003): 1290 – 1300.

Zhu L, Wilken J, Phillips N, Narendra U, Chan G, Stratton S, Kent S, and Weiss M. Sexual dimorphism in diverse metazoans is regulated by a novel class of intertwined zinc fingers. Genes and Development. 14 (2000): 1750 – 1764.

APPENDIX A

Internet sources

Translation
ExPASy Translate Tool:
http://www.expasy.ch/tools/dna.html

Sequence Alignment
ClustalW:
http://www.ebi.ac.uk/Tools/clustalw2/index.html

Malign – AliBee - Multiple Alignment :
http://www.genebee.msu.su/services/malign_reduced.html

Mobyle:
http://mobyle.pasteur.fr/cgi-bin/MobylePortal/portal.py?form=boxshade

Wasp Genome Sites
NCBI:
http://www.ncbi.nlm.nih.gov/genome/guide/wasp/

HGSC:
http://www.hgsc.bcm.tmc.edu/projects/nasonia/

Geneboree:
http://www.genboree.org/java-bin/index.jsp

Protein Domains
Pfam:
http://pfam.janelia.org/

SMART:
http://smart.embl-heidelberg.de/

Exon/Intron Analysis
Spidey:
http://www.ncbi.nlm.nih.gov/IEB/Research/Ostell/Spidey/

Splice Site Analysis
NNSPLICE v. 0.9:
http://www.fruitfly.org/seq_tools/splice.html

ASPic:
http://t.caspur.it/ASPIC/use.php

Primer Design and Alignment
Primer3:
http://frodo.wi.mit.edu/

Sequence Extractor:
http://www.bioinformatics.org/seqext/

APPENDIX B

RT-PCR of NvDM1

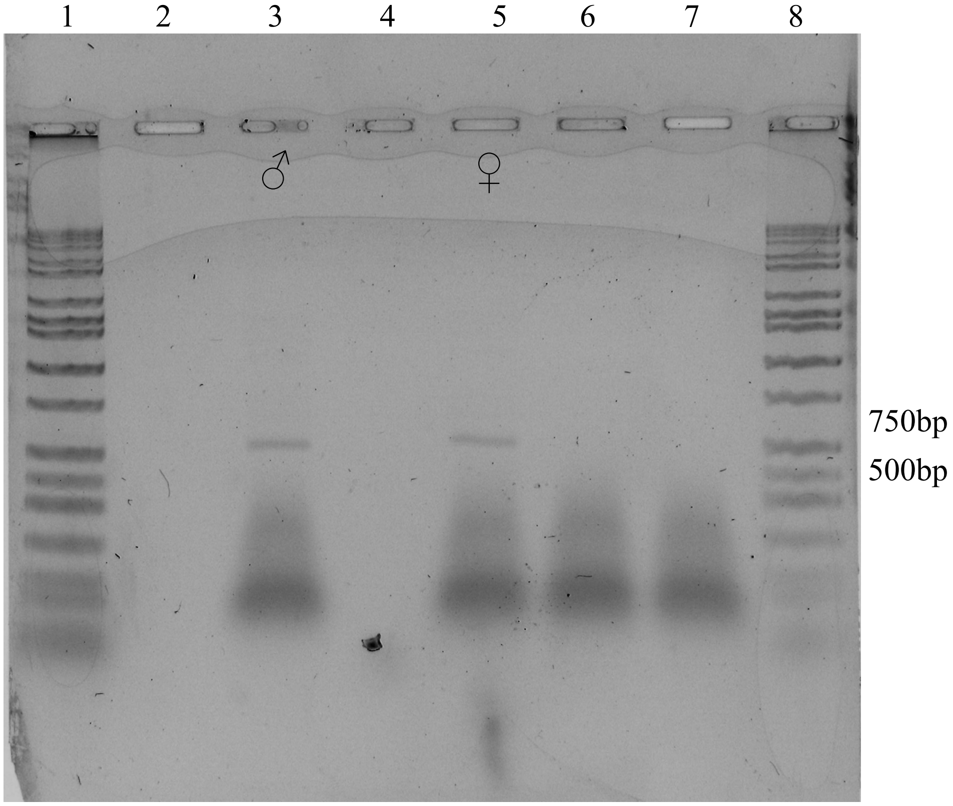


Figure 1. Primers S1 and A4 to amplify across junction b1 and b2. Expected product size: 524. The lanes are as follows: (3) male template (+RT); (5) female template (+RT); (6) male template (-RT); (7) female template (-RT); (2, 4) Empty; (1, 8) Hi-Lo.

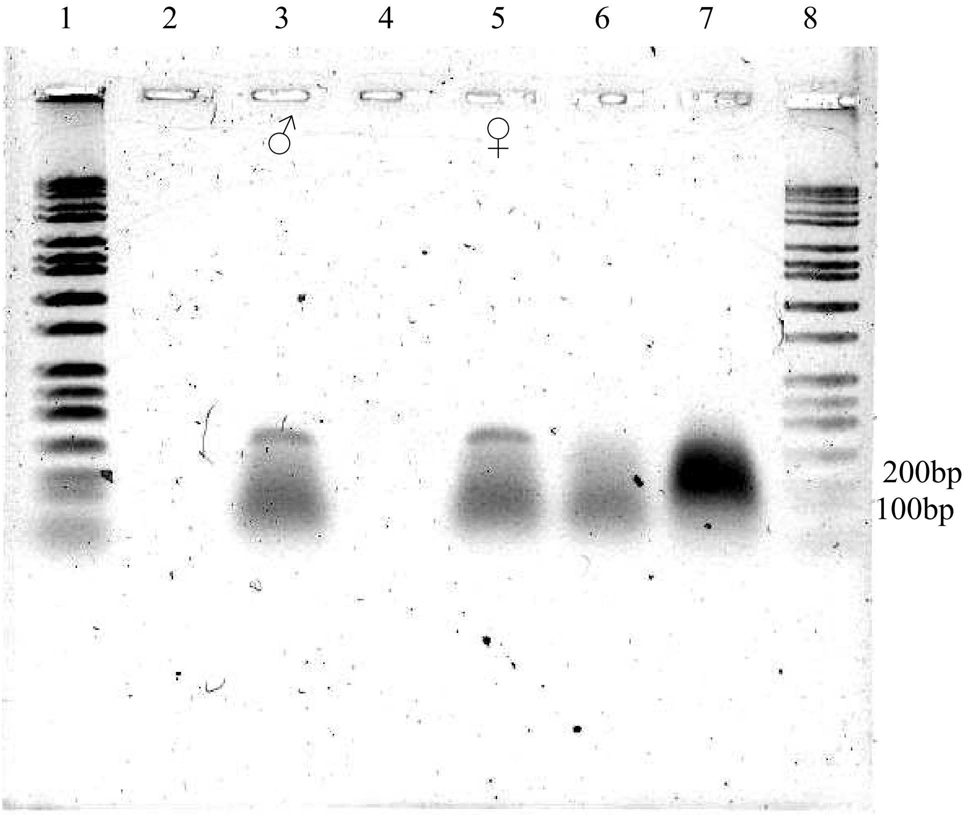


Figure 2. Primers S11 and A11 to amplify across junction b1 and b6. Expected product size: 204. The lanes are as follows: (3) male template (+RT); (5) female template (+RT); (6) male template (-RT); (7) female template (-RT); (2, 4) Empty; (1, 8) Hi-Lo.

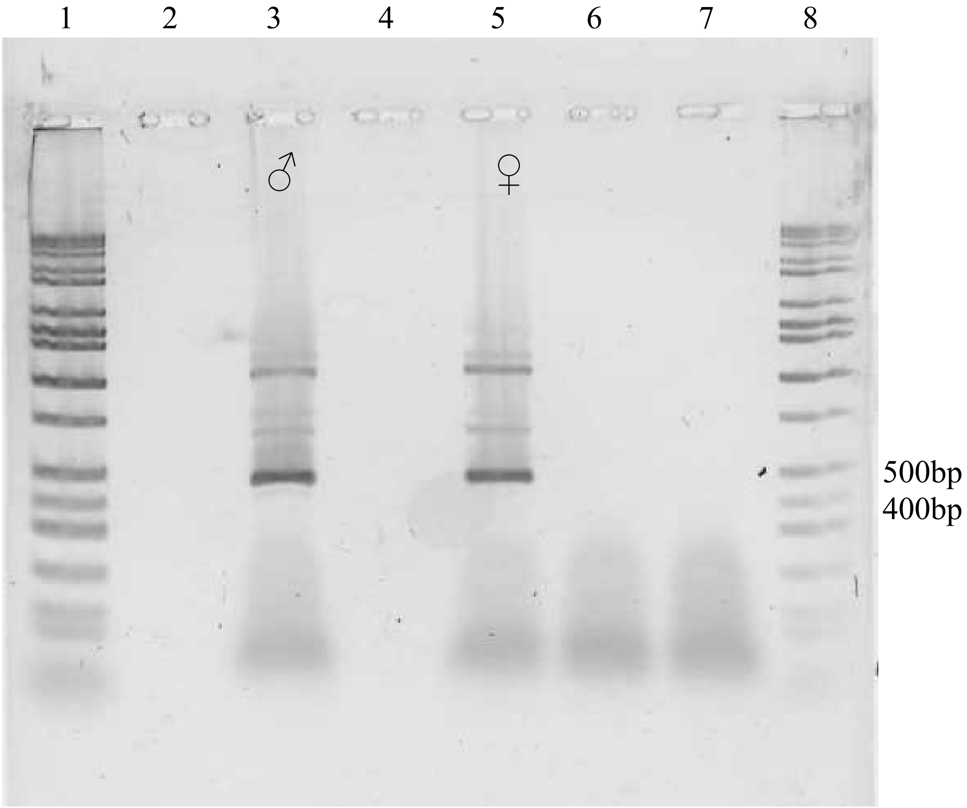


Figure 3. Primers S11 and A12 to amplify across junction b1 and b4. Expected product size: 460. The lanes are as follows: (3) male template (+RT); (5) female template (+RT); (6) male template (-RT); (7) female template (-RT); (2, 4) Empty; (1, 8) Hi-Lo.

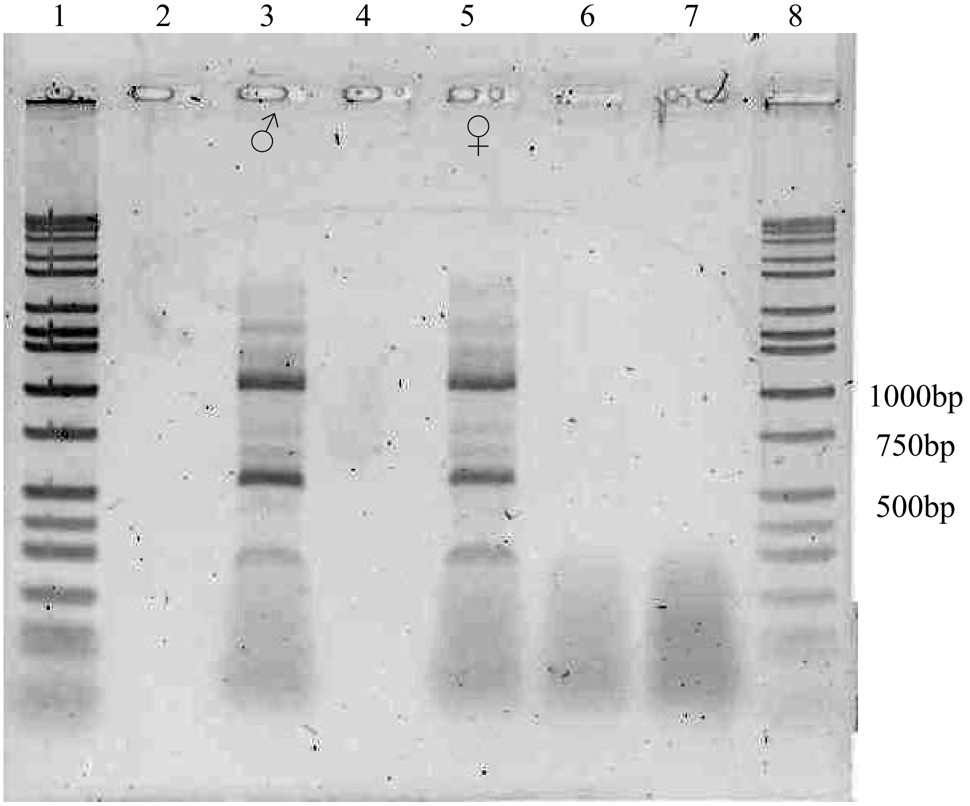


Figure 4. Primers S12 and A13 to amplify across junction b5. Expected product size: 549. The lanes are as follows: (3) male template (+RT); (5) female template (+RT); (6) male template (-RT); (7) female template (-RT); (2, 4) Empty; (1, 8) Hi-Lo.

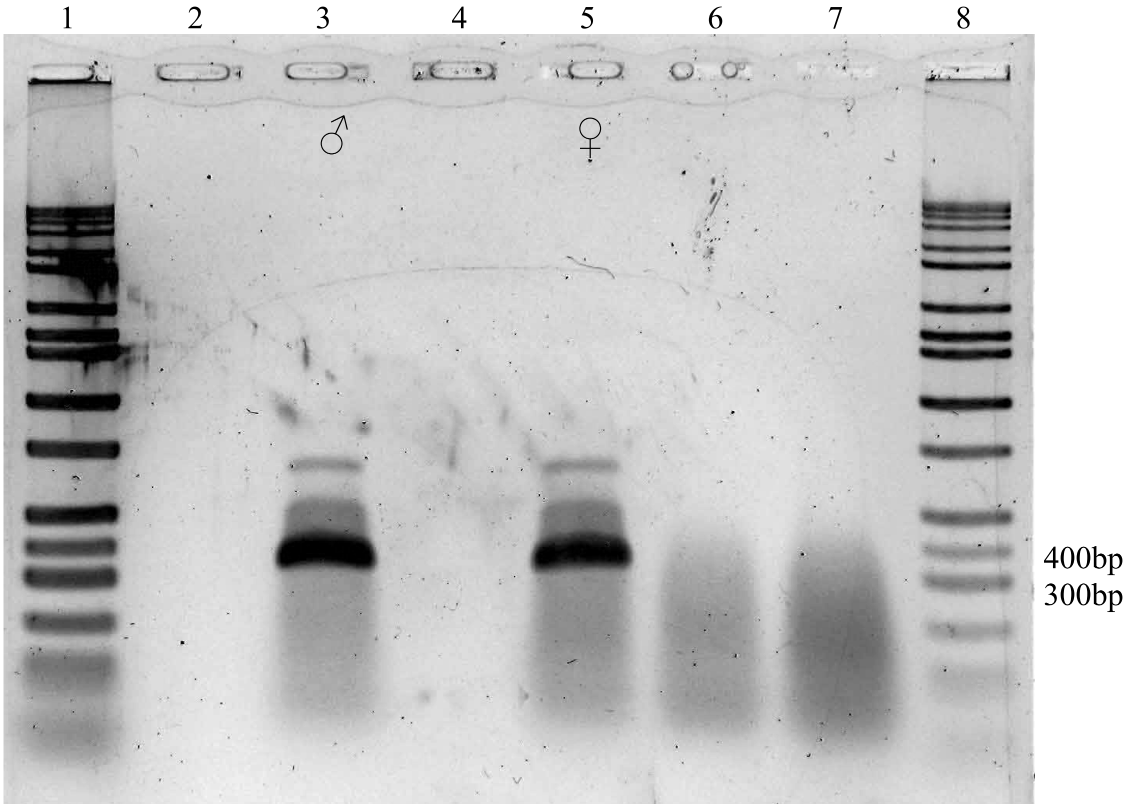


Figure 5. Primers S10 and A10 to amplify across junction b3. Expected product size: 328. The lanes are as follows: (3) male template (+RT); (5) female template (+RT); (6) male template (-RT); (7) female template (-RT); (2, 4) Empty; (1, 8) Hi-Lo.

APPENDIX C

Primer sequences

NvDM1

Name	L	R	Sequence	Use	Designer
DM1 S1	X		CTGCCGATGGCGTGACTGTATATGTGCAA	RT-PCR cDNA 3.4	CT
DM1 S10	X		AAGCCCATCTCATACCCCACCCATT	RT-PCR cDNA 3.4	MR
DM1 S11	X		CGAGTTATGGCAGCACAGGTTGCTT	RT-PCR cDNA 3.42	MR
DM1 S12	X		TCTTCGTGAGGGAGAAGCCAAGGTC	RT-PCR cDNA 3.46	MR
DM1 A4			TGGGTGGGGTATGAGATGGGCTTGATG	RT-PCR cDNA 3.4	CT
DM1 A10		X	CGGATGCTGAATCTGGTGTCCATTG	RT-PCR cDNA 3.4	MR
DM1 A11		X	AAACCATGGCATTGCAAACGAACAG	RT-PCR cDNA 3.42	MR
DM1 A12		X	CCACTCGCAGAGCAATCTCCTCAGA	RT-PCR cDNA 3.46	MR
DM1 A13		X	GTGCCGTGTTAGTGGTGACGGAGAG	RT-PCR cDNA 3.46	MR
DM1 A14		X	AAAGTTGTATGCTTGCCGGTTTTGA	RT-PCR cDNA 2.9	MR

NvDM3

Name	L	R	Sequence	Use	Designer
DM3LE4	X		GGTCGAGTGCAGATGTGGGACTTGG	3′ RACE	MR
DM3LE5	X		CTTGAGCAAGCCGACGACGTACGAA	3′ RACE	MR
DM3LE5.2	X		AGGCCGATTTTTCACCTGCCTTCGT	3′ RACE	MR
DM3LE2	X		ACCACGGGCTGATATCCTGGCTGAG	RT-PCR	MR
DM3RE3		X	CGGCTCCGTCACTGTCATACCGAAA	RT-PCR	MR
DM3RE4		X	GGTGGCGAGTGATGTCGAGGCTTTT	RT-PCR	MR
DM3RE5		X	AGCAAACAGTTCGGTTGGACGCAAG	RT-PCR	MR

NvDM4

Name	L	R	Sequence	Use	Designer
PERE3		X	AGTTCGTGAAGGCCGAGGGTGGATAGT	RT-PCR	MR
PERE5		X	CATACGGTGGACGATCTCGCATGTACC	RT-PCR	MR
PERE6		X	GCAACCCCAGGATAAGTTCGTCGTGAA	RT-PCR	MR
PERE7		X	GAGCTCCGCGATTTTCAGCAGCATT	RT-PCR	MR
PERE8		X	GTAGGGCTTCCACGGTGCGATCTTC	RT-PCR	MR
PERE9		X	TGTGGATCACCATCACGACGAACTG	RT-PCR	MR
PELE3.1	X		CGGCTACTATCCACCCTCGGCCTTC	3′ RACE	MR
PELE3.2	X		GCAACCTCGGCAACCCGACCTACTT	3′ RACE	MR
PELE3.3	X		GACCTACTTCGGCCAGGTCCCCTAC	3′ RACE	MR
PELE3.4	X		GCACACACGTCGTCAGTCCGAAGGT	3′ RACE	MR
DM4LE1.7	X		CTCAGAGGAGGGCGAGACAGCAACA	RT-PCR	MR
DM4LE1.6	X		CATCGACGAAAAAGCCCAAGCCAAG	RT-PCR	MR
DM4LE2	X		GGAAGAAATCGGACTCGGAGACACCAC	RT-PCR	MR
DM4RE4		X	TACGCCGCATGGCTACTTCCACATC	RT-PCR	MR
DM4RMale1		X	TGCCAAAAATACTTGAACTTTTGACGAT	ODX, ODA	MR
DM4RMale2		X	TGCAATGCCAAAAATACTTGAACTT	ODX, ODA	MR
DM4R-M5-1		X	TGGTGTAACTTCAATACACTGCTTCATCTG	ODX	MR
DM4R-M5-2		X	GCCTTTAGTTCGAATACATAATTCCGAAAA	ODX	MR
DM4RA.1		X	CGTAGGGGACCTGGCCGAAGTAGGT	ODC.1,2	MR

DM4RA.2		X	CACCTTCGGACTGACGACGTGTGTG	ODC.1,2	MR
DM4RA.3		X	GACTGGAGGGCACCTTCGGACTGAC	ODC.1,2	MR
DM4RB.1		X	TCGACGCTTGCTTACTCCGTGGAAA	ODC.1	MR
DM4RB.2		X	CACTCTCATGAATGAATCGACGCTTGC	ODC.1	MR
DM4RB.3		X	CCTGAAGTGAGGGGAATTGAGAAGTGC	ODC.1	MR
DM4RC.1		X	CAACAAAGAGAGGCGCACGACGAGA	ODC.1,2	MR
DM4RC.2		X	TTGCAAACGTGGCAACAAAGAGAGG	ODC.1,2	MR
DM4RC.3		X	GCTCGAAATCCTCGGCCGGAAATAG	ODC.1,2	MR
DM4RD.1		X	CTCGCGCTTCCCTCTCGTTCTCTC	ODC.1	MR
DM4RD.2		X	AAAACGCGGATCTCTCGGGGAATTA	ODC.1	MR
DM4FLE3	X		AACCCGAACCAAATCCACGACTTGT	OD-B	MR
DM4FRE5.1		X	TCGAATAAAGTTCTGGTTGCCAGACG	OD-B	MR
DM4**R**E1.1		X	CTGGACCTTCTTGCCGTGATTCTGACA	5′ RACE	MR
DM4**R**E1.2		X	ATTTTGGTATGCGCTGACTTGGCTTGG	5′ RACE	MR
DM4RE1.3		X	TCCTTGGATTTGGCCGATTTCTTCCTC	5′ RACE	MR
DM4RE1.4		X	GTCACCGTCGCTGCTATTGCTGTCATT	5′ RACE	MR
DM4RE1.5		X	CTGCACGTCTCGCTGTTGCTGTTGTTA	5′ RACE	MR
DM4LE1	X		CTCAGAGGAGGGCGAGACAGCAACA	3′ RACE	MR
DM4LE2	X		GGCCAAATCCAAGGACTCGGAGACA	3′ RACE	MR
DM4F-21-M13R			CCTTGAAGATCAAAAGTTCTGCCAATC	Sequencing	MR
DM4F-21-T7			CTAACAATCGTTGATGCGAATGACA	Sequencing	MR
DM4F-2-M13R			TGGCTGTGAATTCTTGTACCTGATGA	Sequencing	MR
DM4M-10-M13R			TCTGGCAACCAGAACTTTATTCGAGA	Sequencing	MR
DM4M-10-T7			AATATTCGTTCTACGTTACCCCCTATCAAA	Sequencing	MR
DM4M-5-M13R			AGATAAGTCGCGCTGCACACTGCGATA	Sequencing	MR
DM4M-5-T7			ACCCGCAGTGACATGCGTAGTTTGA	Sequencing	MR