MOLECULAR EVOLUTION OF SEX DETERMINATION GENES IN CLOSELY RELATED SPECIES OF THE *DROSOPHILA MELANOGASTER* SUBGROUP

By

ROB J. KULATHINAL, B.Sc. (Hons.), M.Sc.

A Thesis

Submitted to the School of Graduate Studies

in Partial Fulfilment of the Requirements

for the Degree

Doctor of Philosophy

McMaster University

© Copyright by Rob Kulathinal, October 2001

EVOLUTION OF SEX DETERMINING GENES IN DROSOPHILA

DOCTOR OF PHILOSOPHY (2001)
(Biology)

McMASTER UNIVERSITY
Hamilton, Ontario

TITLE: Molecular Evolution of Sex Determination Genes In Closely Related Species of the *Drosophila melanogaster* Subgroup

AUTHOR: Rob Kulathinal, B.Sc., M.Sc. (McMaster University, Hamilton)

SUPERVISOR: Professor R.S. Singh

NUMBER OF PAGES: xii, 184

ABSTRACT

Recent studies have highlighted the importance of sex and reproductive related traits in speciation. Hence, an evolutionary examination of the primary regulators of sexual differentiation is an important step in understanding the origins of phenotypic diversity. In this thesis, a molecular analysis of sequence variation is performed on the three interacting sex determining genes, *transformer* (*tra*), *transformer-2* (*tra-2*) and *doublesex* (*dsx*), among closely related species of the *Drosophila melanogaster* subgroup.

At the *dsx* locus, both female- and male-specific DSX isoforms are found to be conserved relative to other loci sequenced from species of the *D. melanogaster* subgroup. However, levels of selective constraints on the male-specific portion of the dsx protein has varied considerably across Dipteran lineages and indicate the apparent heterogeneity in selection pressures across taxa. Meanwhile, *tra* and *tra-2* are demonstrated to be among the most rapidly evolving loci found in the *D. melanogaster* subgroup. The presence of large arginine-serine rich insertions in sibling species indicate that TRA is tolerant to substantial amino acid change. Generally, within and between species patterns of variation in the sibling species do not deviate from a neutral model of evolutionary change. At the *tra-2* locus, the ratios of replacement to synonymous substitutions were significantly different within *vs.* between species indicating the presence of diversifying selection at this locus. Polymorphism data from *D. simulans* corroborates this result and thus, provides evidence that positive selection may be acting directly on a gene involved in the primary sex determination hierarchy.

The high levels of variability found among sex determining genes indicate the evolutionary potential of this important sexual system. The amenability of this sexual system to tolerate large genetic perturbations may expedite the generation of evolutionary novelties. Since sex determination genes control various components of the mating system, this high degree of genetic variability may be an important source of heritable genetic material for sexual selection - an important driver of speciation - to act upon.

ACKNOWLEDGEMENTS

I would like to offer my sincerest thanks to all those who have assisted me on my journey as a graduate student throughout the years. Members of the departmental staff, particularly Pat and the office staff, have made my stay both comfortable and enjoyable. It was a great pleasure to meet and know the multitude of students and postdoctoral fellows (especially members of the soccer team) who have gone through our department. Many thanks to my past roommates, Sen and Randy. And my deepest gratitude to the great friendships that I have acquired in Hamilton, in particular, Alberto, Jawn, Debbie and Cindy. I have learned much through them and have enjoyed even more during our times together. I look forward to our friendships in the future.

I would also like to thank our faculty. I will certainly miss the terrific interactions that I have had with nearly all the professors from our department. In particular, my appreciation goes to Ana and Roger who initially directed me to exciting questions in developmental biology and Susan, Lisle, J.P. and Brian who have selflessly instructed me in the way of evolutionary genetics. A special thanks to Brian for being so patient in my development as a molecular evolutionist.

Members of the Singh lab, past and present, have become a very important component in my life. My gratitude to my longtime mentor, Ling-Wen, for first revealing to me the "spirit" of population genetics. Alberto has become not only a valued colleague to discuss science with but also a wonderful friend with many common interests. Anouk, too, will remain a friend for life. I hope that this trend continues with my current

labmates and friends, Santosh, Dara, Li and (last but certainly not least) Aaron. I also would like to thank my two favourite students, Lara and Jacqui. Much of thesis is the result of their hard work.

I have been extremely fortunate to possess two of the wisest supervisors. Dick has been inspirational in his work ethics and analytical mode of thinking. He has trained me to evaluate anything and everything critically and has constantly challenged me to learn new things. Rama has taught me the importance of philosophy in both scientific pursuits and daily living as well as the importance of stepping back to look at the grander picture. In only a few short years, he has taught me the ways of academia what most people struggle to learn in a lifetime. One could not ask for better teachers and I am greatly indebted to them. Both are scientists to the core yet demonstrate wonderfully caring personalities. Over the years, their roles have evolved from mentors to colleagues to friends - roles which I will cherish for a lifetime.

Of course, I am deeply grateful to my family and I thank them for their unconditional support and patience throughout my impoverished years.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES AND APPENDICES

CHAPTER ONE

General Introduction

1.1 On the Origins of Organic Diversity

1.11 Explaining the variety of life

The origins of the vast diversity of organic forms has been one of mankind's most enduring mysteries. The wealth of organismal variation can be found on many different levels. Obvious differences in size and complexity abound among organisms from different taxonomic kingdoms. Within these groups, a seemingly endless variety of developmental plans, metabolic processes and life strategies exist. Even within sexually reproducing species, which comprise the majority of eukaryotes, the existence of pronounced dimorphisms in the form of males and females adds to the extraordinary diversity of life.

A large amount of effort has been dedicated to explain the diversity of organisms that inhabit our planet at any time. Of course, the first step in understanding these origins was the recognition of observable patterns and the classification of organisms into morphologically distinct groups or *taxa* became a necessary task. Early Greek philosophers, for example, categorized organisms into like forms based on a perceived archetype, or *eidos* (translates to species). This concept originated from Plato's Theory of Forms which first suggested that ideas were based on an ideal type and that any variation represented a distortion from reality and was unimportant (Plato, in *The Republic*).

Although this concept was a typological one, the classical tradition was important in laying the foundation of modern biology. Plato's student, Aristotle, later extended the Theory of Forms to the biological world and formulated the *scala naturae*, or "Great Chain of Being". In this hierarchical ladder, simple (or lower) organisms were found on the bottom rungs while more complex (or higher) animals were situated on top. From this hierarchical categorization, a biological framework for classification materialized.

With the observation of a pattern of simple to complex organisms, the next step was to explain how this pattern came into being. For the Greeks, this Chain of Being was fixed and immutable. As the contemporary philosopher Arthur Lovejoy remarked, "The Chain of Being, in so far as its continuity and completeness were affirmed on the customary grounds, was a perfect example of an absolutely rigid and static scheme of things" (Lovejoy 1964, pg. 242). For over two millennium, this hierarchical concept of life and its nonchanging behaviour was absorbed into Western philosophy and schools of religious thought. Christian notions of Creation and man's "dominion ... over all the wild animals of the earth" (Genesis 1:26) were consistent with *scala naturae*. But with the arrival of the Renaissance in the 16th century, there was a revival in philosophical and empirical interests, especially in the biological sciences. Finally, old Aristotlean dogmas were challenged and ultimately eliminated. For example, the French philosopher Herder, a student of Immanuel Kant, argued that *scala naturae* is not a fixed order of being but rather, a progressive scale of descent (see Lovejoy 1959, pg.220). Extinct species were thought to represent gaps in the ladder thus rejecting the aged dictum, *natura non facit saltum*, or "nature makes no leaps". Other members of the French environmental school

of thought also advocated the mutability of the fixed ladder. Lamarck renamed it, *La marche de la nature*, and championed the revolutionary viewpoint that adaptations to the environment allow species to ascend the hierarchical rungs of ladder (Lamarck 1809, 1984).

1.12 Darwin provides a new framework

With the publication of *On the Origins of Species*, Charles Darwin (1859) introduced a nonessentialist and materialistic theory to explain the diversity of life. While other thinkers had already suggested a changing or *evolutionary* concept of species, Darwin provided a complete synthesis. Here, I describe two of his greatest achievements in dispelling traditional conceptions of species immutability. First, Darwin discarded the concept of a fixed ladder of being. Darwin argued that a branching model of evolution would best explain the discontinuities, or gaps, among presently living species. Related extant species would share common ancestors which were most likely absent or extinct. Hence, the diversity of life would be better represented as a tree, with its terminal branches denoting extant species, and its internal trunks, those species that are ancestral and creating the discontinuities. This fundamental paradigmatic change in viewing organic reality allowed Darwin to free himself from the constraints of *scala naturae*.

Darwin's second revolutionary idea in transforming the static concept of the species into a dynamic one was the idea of descent with modification. Influenced by an essay entitled, "An Essay on the Principle of Population", by the economist Thomas Malthus (1799) and his tenure as a naturalist on the Beagle, Darwin utilized previously

ignored variation found in populations as the basis of his general theory of natural selection. For the first time, a plausible and probable mechanism of evolutionary change was proposed and propelled an evolutionary notion of the species into the forefront of biological thought.

1.13 Two central problems in evolutionary biology

Understanding the processes involved in species formation became an important focus in evolutionary biology long after Darwin published his seminal work in 1859. (I refer to this as the first problem not due to chronological order, but because of its importance.) As observed by many contemporary critics, Darwin's treatise on species origins really had little to do with the process of species formation but simply explained the divergence of species (or as he called them, varietals) through gradual adaptive mechanisms (Coyne 1992). Speciation theory was not advanced until almost a century later (see below). Much of this had to do with the absence of a material basis for heredity which would only be developed with the rediscovery of Mendel's Laws by the turn of the century while another part had to do with the nominalistic views of Darwin himself (Darwin 1859, Chapter 1). And it wasn't as though Darwin lacked the observations necessary to achieve a fully comprehensive theory of speciation. Darwin realized that varieties may be reproductively isolated from each other. For example, in his writings on hybrid sterility (*On the Origin of Species*, Chapter 8), Darwin (1859) explains how such a maladaptive trait as hybrid sterility could evolve (see below). Nonetheless, with Darwin's publication, a new biological framework was constructed which would

inevitable lead to an understanding of how species are formed.

The second problem, which was introduced in the nineteenth century by Darwinian theory, was the relationship between development and evolution. Darwin considered his examples showing the common features of embryos from related species to be among the most important evidence for evolution. "Embryology is to me by far the strongest single class of facts in favour of change of forms, and not one, I think, of my reviews has alluded to this" he wrote in his autobiography (Darwin 1888). Of course, Darwin was not the first to observe the commonality of embryonic forms from phylogenetically similar species. Many philosophers and scientists of the German school of thought, *naturphilosophie*, observed common body plans, or *bauplans*, and similar design was often observed early in development. In the 1820s, the German biologist Meckel and the French anatomist, Serres, independently arrived at the law of parallelism (also known as the transcedentalist law) which proposed that the embryological development of a higher animal recapitulates the adult morphologies of the lower animals beneath it in the order of beings (Raff and Kaufman 1983, pg. 8). For example, Serres stated that "a man becomes a man only after passing through transitional stages of organization in which he is similar first to a fish, then to a reptile, then to birds and mammals" (Serres 1824). The German embryologist, Karl von Baer, was adamantly opposed to this idea of embryonic recapitulation as a result of his observations on vertebrate development. From his careful observations, he recognized simply that embryos of similar kind shared common features (von Baer 1828). This created a misconception, as von Baer pointed out, since the development of complexity among

higher animals superficially makes it seem as though their ontogeny recapitulated that of the lower animals. Unfortunately, in his evolutionary interpretation of the law of parallelism, the German embryologist Haeckel (1879) readopted the mistaken belief that "ontogeny recapitulates phylogeny" (also known as the biogenetic law). These early theories, however, correctly viewed developmental processes to be generally conserved between related species. Recent phylogenetic studies on developmental genes have supported this notion (Nullsein-Volhard 1994, see Carroll *et al.* 2001).

Both speciation and the role of developmental changes in creating evolutionary novelty are central problems in evolutionary biology. These problems are associated with the origin of diversity on two distinct levels. On one level, microevolutionary processes, which include the evolution of reproductive isolation and speciation (Dobzhansky 1937), are the primary causes of the discontinuities we observe in nature. On a higher level, drastic differences between distant taxa, most likely caused by changes in developmental programmes, generate the extraordinary diversity of life. The interconnection between both levels represents a fascinating yet enigmatic facet in modern evolutionary theory.

1.2 Reproductive Isolation and Speciation

1.21 Early theories of species formation

It was not until almost a century after Darwin's initial publication that the process of speciation began to be understood in a unified manner. During this time, evolutionary theory underwent a period of development in order for it to fit current and novel paradigms. The idea of evolutionary change first went through methodical and

sometimes fierce debate (i.e. Huxley 1900). The understanding of a heritable basis of evolutionary transformation was assured with the rediscovery of Mendel's laws. The rift between the naturalists and the geneticists was reconciled by the development of population genetics which advanced the quantitative understanding of gene frequency (Fisher 1930; Wright 1931; Haldane 1932). Yet, it was not until the Evolutionary Synthesis of the 1940s that an acceptable comprehension of the process of speciation was achieved. During this period of time, such diverse scientific disciplines as geology, botany, zoology, genetics and palaeontology combined forces and developed a comprehensive and modern synthesis of evolutionary theory (Mayr 1963; Dobzhansky 1970).

Perhaps the most important contribution was the association of speciation with reproductive isolation (Fisher 1930; Dobzhansky 1937; Mayr 1942). A major focus was placed on geographic modes of reproductive isolation. This focus on physical isolation models can be seen in the numerous debates about the importance of allopatry *vs.* sympatry found in the literature (Maynard Smith 1966; Bush 1969; Kulathinal and Singh 2000a,b; Coyne and Price 2001). Other examples that highlight the importance of geographical models include founder effect models (reviewed in Carson 1968; Templeton 1980) which concern the relocation of a small population to a new environment, and parapatric models which argue for the formation of species along a geographical axis associated with a cline in selection intensity for a particular trait (i.e. Slatkin 1973). This dependence on geography to explain the initial divergence of incipient species may well restrict a complete understanding of the diversity of processes that lead to speciation (see

below).

1.22 The genetics of reproductive isolation

The knowledge that mechanisms of reproductive isolation invariably lead to speciation allowed species formation to become an experimentally addressable problem. Incompatibilities found in the interspecific hybrid (i.e. inviability and sterility) offered the promising potential that one could acquire at least some idea of the nature of the genetic changes that led to reproductive isolation. However, with the discovered importance of geography in speciation theory, initial stages of divergence were thought to take place in allopatry, and allelic incompatibilities were expected to accrue as a byproduct of divergence (Dobzhansky 1937, Muller 1942). Once these diverged species again meet in sympatry, the accumulated incompatibilities are manifested in the hybrid. Thus, the evolution of hybrid incompatibilities and their role in speciation, was finally understood and other patterns of species hybridization could be explained. For example, J.B.S. Haldane (1922) found that "when in the F1 hybrid offspring of two different animal races one sex is absent, rare or sterile, that sex is the heterozygous [heterogametic] sex". Most explanations of Haldane's rule have utilized the fact that the heterogametic sex possesses two different sex chromosomes while the homogametic sex contains autosomes and sex chromosomes from each parent. Interest in Haldane's rule across a wide spectrum of animal taxa lead to the introduction of new tools which allowed reproductive isolation to be finely dissected. Dobzhansky (1936) pioneered some of the first studies of hybrid male sterility in two related species of Drosophila, *D. pseudoobscura* and *D. persimilis*.

Using a series of backcrosses, he demonstrated that the X-chromosome has a large effect on hybrid sterility which lead to the development of various genetic models of species divergence (see Coyne 1992).

Because of its association to early stages of species divergence (Coyne and Orr 1989, 1997), a renewed interest has been found in explaining the genetic basis of Haldane's rule (see Laurie 1997, Orr 1997). One explanation has been termed the dominance theory (Turelli and Orr 1995) and is an updated version of Muller's imbalance theory (Muller 1940, 1942; Zeng 1996). According to the dominance theory, if incompatible alleles act in a recessive manner in the hybrid, the heterogametic sex will always be the first to be inviable or sterile (Orr 1995). Another explanation of Haldane's Rule highlights the importance of sexual selection on male traits (Wu and Davis 1993), a process which has been the focus of extensive research.

1.23 Sexual selection as a factor in speciation

After revolutionizing biological (as well as philosophical and political) thought with his 1859 publication, Darwin produced another less appreciated, until recently, evolutionary theory. His 1871 publication, *The descent of man, and selection in relation to sex*, was a successful attempt to explain the sometimes bizarre diversity within species - between male and female. Exaggerated features found on males such as elaborate colouration and aggressive behaviour could be explained as the result of either female selection on male variability or direct competition between males for females.

Over the last decades, countless studies have demonstrated a remarkable

collection of traits resulting from sexual selective mechanisms in a wide variety of sexually reproducing taxa (Andersson 1994). Sexual selection's association to speciation has also been emphasized. Fisher (1930) first demonstrated how Darwin's theory of sexual selection may cause a trait, even if maladaptive to the male (i.e. natural selection does not favour it), to still evolve rapidly due to its selective advantage in mating. Lande (1981) modelled this process and demonstrated that a maladaptive male trait may evolve rapidly in a runaway process. Thus, the formation of species can be achieved in a very short time period. That sexual selection may be the driver of rapid evolutionary changes in mate signals and preference runs counter to previous models which assume the species must be geographically separated from each other and divergence takes place in a rather gradual manner. Using the newly formed species of Hawaiian Drosophila (over 500 species have evolved within the last five million years), Carson (1997) demonstrated that many of the diverged male traits used in mating, as well as female preferences, have evolved through sexual selection on small founding populations.

Theories of sexual selection have traditionally invoked female preference of male traits (Kirkpatrick 1985). This preference allows female traits to remain more-or-less constant and has been explained by the good-genes model among others (Fisher 1915; Zahavi 1975). Recently, a coevolutionary process between male and female traits has been proposed. Rice (1996) suggested that males and females continuously produce strategies that would increase the fitness of themselves, even if it involves a loss of fitness in its partner. This antagonistic conflict between the sexes will cause both male and female traits to evolve at a rapid rate. Arnquist *et al.* (2000) indeed showed that in taxa

which exhibited sexual conflict, rates of speciation were four times higher than similar (related) taxa in which sexual selection is greatly reduced because they possessed monogamous mating systems.

1.24 Sex genes and their role in speciation

One recent and important development in sexual selection theory has been the extension of sexual selection to traits other than secondary sexual characters (Eberhard 1996; Civetta and Singh 1998a). Previous examples focussed on classical morphological traits involved in mating (precopulatory). This extension increases the number of traits on which sexual selection could act upon. Eberhard (1985) demonstrated that male genitalia, directly involved in copulation, are extremely diverged in a variety of animal taxa and proposed that this diversity was caused by sexual selection. Proteins involved in fertilization such as Drosophila accessory gland proteins (Aguadé *et al.* 1992, Clark *et al.* 1995, Tsaur and Wu 1997) are also highly diverged and sexual selection may be the causative factor. Because sexual conflict involves the coevolution of male and female traits, females traits/genes are also expected to be diverged. Civetta and Singh (1995), using two-dimensional electrophoresis, demonstrated that proteins from male and female reproductive tracts are more diverged between closely related species of Drosophila than are proteins from other sampled tissue. Swanson *et al.* (2001) also showed that three female-specific mammalian egg proteins involved in binding sperm are highly diverged and that divergence is driven by positive selection. In another study, Civetta and Singh (1998b) classified available genes as sex or non-sex and found that sex genes were

significantly more diverged than non-sex genes indicative of the role of sexual selection on these genes. Singh and Kulathinal (2000) review a number of classes of sex and reproduction-related genes (SRR) genes which reveal high divergence.

An accumulating number of examples of SRR genes are found to be rapidly evolving among a wide range of species (see Table 5.1). The use of an extended or 'broadened' concept of sexual selection (Civetta and Singh 1999) allows us to abandon the view of speciation as simply the gradual divergence of allopatric taxa. Sexual selection may represent an impressive force which increases the rate of speciation. The greater fitness component found in sexual systems within species (Prout 1971; Kingsolver *et al.* 2001; Hoekstra *et al.* 2001) may translate to the phenotypic variation we observe between species. Thus, a new and encompassing view of speciation is being formed.

1.3 The Role of Development in Macroevolution

1.31 Importance of gene regulation in generating species diversity

While much of the population genetics of speciation deals with protein divergence and allelic differences in enzymatic activity (Lewontin 1974), differences in the regulation of gene expression have remained relatively unexplored. But regulatory interactions may play a major role in producing evolutionary novelty (Tautz 2000). King and Wilson's (1975) comparison of chimpanzees and humans using DNA hybridization and protein and electrophoretic comparisons, revealed that over 99% of our genes are similar, supporting the idea that the large conspicuous differences between these two

primates are primarily regulatory in nature (Wilson *et al.* 1974). Until recently, biologists could only make inferences about the role of regulatory elements in species diversity. Now, with the readily available tools of molecular biology, the role of regulation in species differences can be experimentally pursued.

Studying the evolution of gene regulation in both non-coding regions and regulatory proteins presents a difficult challenge. Regulatory regions, which may comprise much of the noncoding part of a gene, do not follow a general design and may be located far from the actual gene itself. Furthermore, regulation itself may include a wide range of mechanisms from transcriptional control to post-translational modifications. However, by selecting loci with characterized and tractable regulatory sequences, these problems could be alleviated. Recently, a series of studies have examined the evolutionary dynamics of a well-defined regulatory region in Drosophila. Comparing the enhancer sequence which drives *even-skipped* (*eve*) expression in stripe number two (pair-rule gene expressed in embryogenesis), Ludwig and Kreitman (1995) found the presence of extensive polymorphism at binding sites in the *eve* enhancer region among natural populations of *D. melanogaster* as well as the divergence of enhancer sequence between closely related species. They also demonstrated that a *D. pseudoobscura* enhancer placed in a *D. melanogaster* background drives normal *eve* stripe 2 expression (Ludwig *et al.* 1998). However, most interestingly, they showed that chimeric constructs containing both *D. pseudoobscura* and *D. melanogaster* enhancer sequence caused misexpression (Ludwig *et al.* 2000) and concluded that enhancer regions evolve rapidly but stabilizing selection appears to conserve their function. In a recent

experiment, Kopp *et al.* (2000) revealed that regulatory changes in the *bric-a-brac* (*bab*) gene (which causes differences in *bab* expression) have played an important role in the evolution of adult abdominal segment morphology in Drosophila. *bab* modulates signals from the homeotic and sex determination pathways to create sex-specific pigmentation, most likely evolved by sexual selection (Kopp *et al.* 2000).

Regulatory proteins (i.e. transcription factors) have also been shown to play an important role in the evolution of phenotypic diversity by means of reproductive isolation. Ting *et al.* (1998) isolated the transcription factor, *OdsH*, involved in the reproductive isolation of *Drosophila simulans* and *D. mauritiana*. This gene contains an extremely fast evolving homeobox domain which, when its *D. mauritiana* ortholog is introgressed into a *D. simulans* genetic background, produces a significant male sterile effect.

1.32 The diversity of sex determination mechanisms

One of the most significant patterns identified in evolutionary and developmental biology is the conservation of developmental processes at a molecular level across distant taxa (Carroll 1995; Peterson and Davidson 2000). Sex determination, however, is a notable exception. For most organisms, sexual differentiation is a fundamental process which affects a whole array of sex-specific traits including morphology, physiology, behaviour and reproduction. While the decision to become either male or female represents a major developmental stage which takes place early in ontogeny, a wide variety of mechanisms exists (Hodgkin 1990, 1992). Table 1.1 displays various

mechanisms of sex determination in an assortment of taxa. Not only are there differences between principle determinants of sex (i.e. chromosomal *vs.* environmental), but particular sexes are either heterogametic or homogametic and modes of dosage compensation vary. Mechanisms include dominant X-, Y- or autosomally-linked loci in mammals, temperature or density dependence in amphibians and reptiles, and the ratio of X chromosomes to autosomes in worms and flies. In addition, genetic and molecular analyses have shown that the molecular bases of regulatory systems which ultimately control sexual development are not homologous (Cline and Meyer 1996). Variation found among sex determination systems is particularly evident in the Order Diptera (see Table 1.1).

1.33 Sex determination in Drosophila

Through extensive genetic and molecular analyses, an excellent comprehension of the sex determining genetic pathway has been procured in Drosophila. Early in Drosophila embryonic development, each cell must autonomously decide whether to be male or female. This critical task is accomplished using the X-linked binary switch gene, *Sex-lethal* (*Sxl*), and the principle signal that activates this switch is the ratio of X-chromosomes to autosomes (X:A) (Baker 1989). Females (X/X;A/A) have equal numbers of X-chromosomes and autosomes (i.e. an X:A ratio of 1.0) while males (X/Y;A/A) have an X:A ratio of 0.5 (Bridges 1925) (Figure 1). Each cell determines this ratio by assessing the number of numerator (X-linked) *vs.* denominator (autosomal) elements or expressed genes. A number of genes have been localized on the

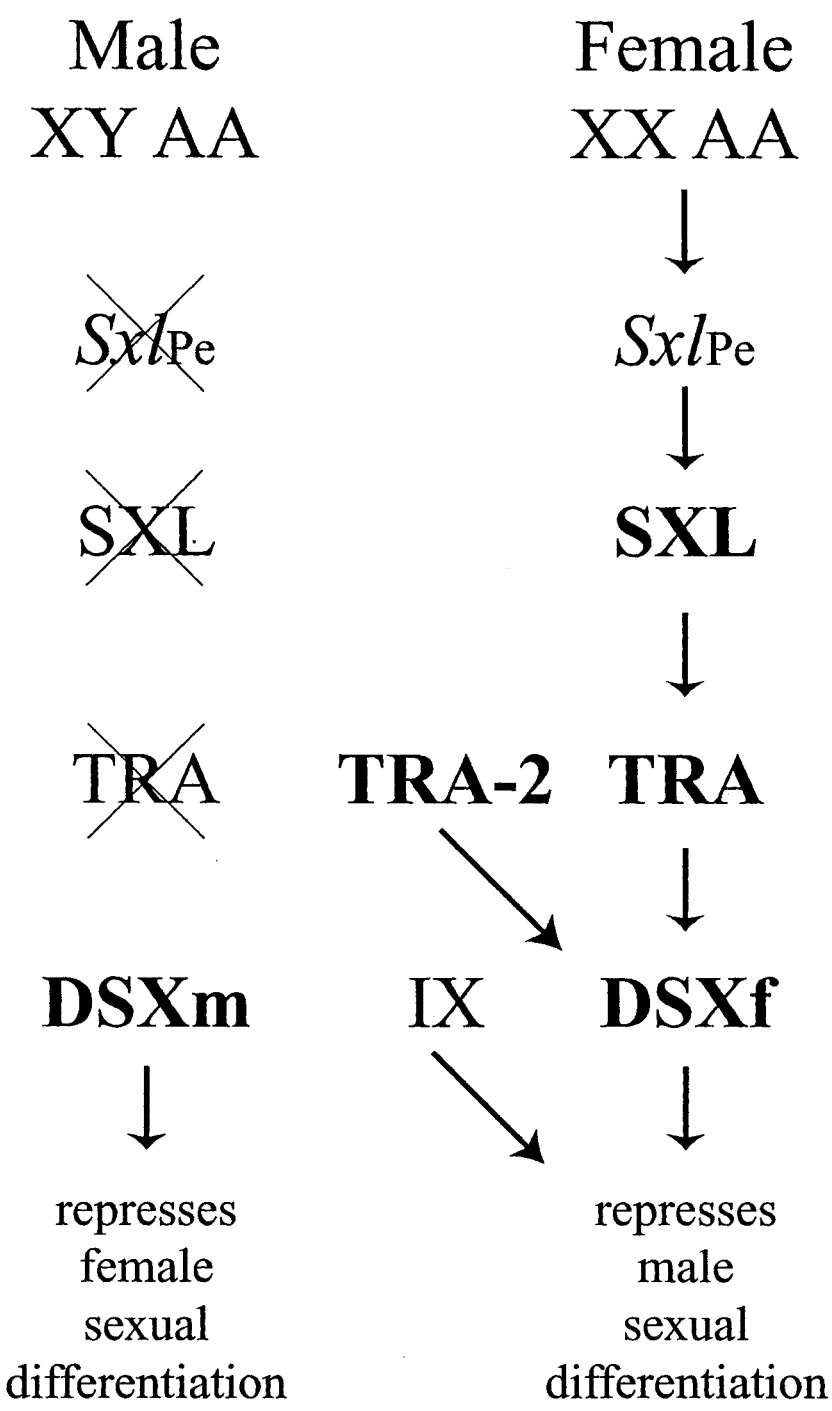
TABLE 1.1

Diversity of sex determination mechanisms

Taxa	Genus	Sex determinant (female, male)	Sex determination mechanism
Mammals		XX, XY	dominant male-determining
Birds		ZW, ZZ	Z:W ratio
Alligators		warm, cool	temperature
Turtles		cool, warm	temperature
Nematodes	*Caenorhabitis*	XX (herm), XO (male)	X:A ratio
	Meliodogyne	sparse, crowded	population density
Insects	*Pales*	XX, XY	dominant male-determining
(Order:	*Sciara*	XX, XO (somatic)	X:A ratio
Diptera)	*Mayetiola*	X_1 X_1 X_2 X_2, X_1 X_2O	genotype of mother
	Culex	homomorphic	dominant male-determining
	Aedes	XX, XY	dominant male-determining
	Anopheles	homomorphic	dominant male-determining
	Eusimulium	homomorphic	dominant male-determining
	Chironomus	homomorphic	dominant male-determining
	Megaselia	homomorphic	dominant male-determining
	Ceratitis	XX, XV	dominant male-determining
	Anastrepha	$X_1X_1X_2X_2$, X_1X_1Y; ZW, ZZ	variable
	Drosophila	XX, XY	X:A ratio
	Musca	XX, XY; homomorphic	variable
	Chrysomya	XX, XY	genotype of mother
	Lucilia	XX, XY	dominant male-determining

Data compiled from Hodgkin (1992) and Marín and Baker (1998); herm=hermaphrodite

Figure 1.1 The sex determination regulatory hierarchy in Drosophila. The ratio between certain genes found on the X-chromosome (numerator elements) and autosomes (denominator elements) directs whether or not the early *Sex-lethal* (Sxl_{Pe}) promoter is activated. In females, functional SXL splices *transformer* (*tra*) *pre*mRNA to form fully functional TRA. TRA, with the transformer-2 protein (TRA-2) activates production of the female-specific doublesex isoform, DSXf. Along with the intersex protein (IX), DSXf represses the activation of downstream male-specific gene expression. In males, the null pathway (i.e. no SXL nor TRA) allows for the default splicing of male-specific doublesex isoforms, DSXm. DSXm represses female differentiation by acting on downstream sex-specific targets.


Male
XY AA
Female
XX AA
SxlPe
SxlPe
SXL
SXL
TRA
TRA-2
TRA
DSXm
IX
DSXf
represses female sexual differentiation
represses male sexual differentiation

X-chromosome and an autosome providing a molecular basis for this X:A ratio (Cline and Meyer 1996). The products of these genes interact transiently during early embryonic development (Cline 1985) to control the transcription of *Sxl* (Figure 1.1), using its early (also known as embryonic or establishment) promoter, *Sxl*-Pe (Keyes *et al.* 1992). Transcription using this promoter occurs in females during the early formation of somatic cells between nuclear cycles 12 and 14 (Erickson and Cline 1991) producing an early SXL protein. Once transcription of the early *Sxl* promoter ceases, constitutive *Sxl* transcription (in both males and females) ensues using the late (i.e. maintenance) promoter, *Sxl*-Pm. The transient production of early SXL serves as a female-specific pulse to alternatively splice late *Sxl* transcripts, produced from the maintenance promoter, into functional SXL (Bopp *et al.* 1991). Late SXL eventually autoregulate their own *Sxl* *pre*mRNA transcripts. Since *Sxl*-Pe is not activated in males, *Sxl*-Pm transcripts are spliced using the default pathway whereby exon 3, which contains an in-frame stop codon, is utilized, thereby producing a non-functional truncated protein. In females, functional SXL can regulate downstream genes such as *transformer* (*tra*), involved in somatic differentiation, to instruct female development. Both female-specific *tra* and the ubiquitously expressed *transformer-2* (*tra-2*) gene products interact with *doublesex* (*dsx*) *pre*mRNA to effect the production of the female isoform, DSXf. Males, without any functional SXL, follow the default pathway to produce the male-specific dsx isoform, DSXm. The sex-specific proteins, DSXf and DSXm, then act upon a cascade of downstream genes involved in somatic sexual differentiation.

1.34 Rapid evolution of sex determination genes

A remarkable diversity of sex determination mechanisms exists in nature (see Table 1.1). Correspondingly, many of the loci involved in sex determination have been shown to be rapidly evolving. O'Neil and Beloté (1992) demonstrated that the *tra* locus was among the most diverged between *D. melanogaster* and *D. virilis*. Both selective (Walthour and Schaeffer 1994) and neutral (McAllister and McVean 2000) modes of selection were supported. In nematodes, the terminal regulators of the sex determining pathway, *tra-1* and *tra-2*, (no homology to their namesakes in Drosophila) were also highly diverged between *Caenorhabditis elegans* and *C. brigassae* (de Bono and Hodgkin 1996; Kuwabara 1996). A large portion of the primary sex determination switch locus, *Sry*, is rapidly evolving in mammals (Whitfield *et al.* 1993; Tucker and Lundrigan 1993; Hawkins 1994). The algal sex determination *mid* locus is found to possess high rates of substitution between *Chlamydomonas reinhardtii* and *C. incerta* (Ferris and Goodenough 1997). Hence, sex determination appears to be a variable genetic system.

1.4 Thesis Overview

1.41 The problem

The molecular evolution of sex determination genes, which are necessary for normal sexual differentiation, presents a fascinating paradox. On the one hand, developmental processes and genes tend to be conserved. But on the other hand, sex determination systems are extremely diverse among different taxa (Hodgkin 1992) and a number of sex determining genes, which primarily direct the regulation of traits involved

in sexual dimorphisms, have been found to be evolving rapidly. This thesis investigates the molecular evolution of sex determination by performing a comprehensive evaluation of sequence variation in the three primary sex determining genes, *doublesex* (Chapter 2), *transformer* (Chapter 3)and *transformer-2* (Chapter 4) in species of the *Drosophila melanogaster* subgroup. Since all differences between taxa ultimately find their roots in the genetic variation found within populations (Purugganan 1998), understanding the evolutionary dynamics among closely related species may direct us to the proximal causes of the observed diversity in sexual systems.

1.42 The species

Sequence variation was quantified using species of the *Drosophila melanogaster* subgroup. This subgroup has a Afrotropical origin (Lachaise *et al.* 1986; Powell 1997). For the highly diverged *tra* and *tra-2* sequences, sibling species of the *D. melanogaster* complex - the cosmopolitan species, *D. melanogaster* and *D. simulans*, and the island endemics, *D. mauritiana* and *D. sechellia* - were sequenced for divergence and polymorphism. These sibling species are morphologically indistinguishable and homosequential. The only diagnostic difference are the shapes of their genial arch's posterior lobes (Coyne and Kreitman 1986). Also, when the latter three species (species of the *D. simulans* clade) are crossed, they produce fertile female hybrids but sterile males (Lachaise *et al.* 1986; Kulathinal and Singh 1998), following Haldane's rule. One advantage of using sibling species of the *D. melanogaster* group is that they possess many genes that have been sequenced in population studies (Hey and Kliman 1993; Kliman and

Hey 1993) and represent "the most thoroughly studied speciation model" (Kliman *et al.* 2000). Of course, one of the other advantages is that sex determination has been extensively characterized at both genetic and molecular levels in *D. melanogaster* (Cline and Meyer 1996). For the relatively conserved gene, *dsx*, seven species of the *D. melanogaster* subgroup - which include the four sibling species and *D. erecta*, *D. tiessieri*, and *D. yakuba* - were assessed for patterns of divergence.

1.43 Objectives and findings

Genes and traits involved in sex and reproduction have been demonstrated to possess higher evolutionary rates indicative of their preferential involvement in the formation of species (Singh and Kulathinal 2000). In particular, sex determination genes have been found to be rapidly evolving across a wide range of taxa. The main objective of this thesis has been to study and understand, in a comprehensive manner, the molecular evolution of this particular subset of genes. I choose to study the evolution of sex determination loci from the *D. melanogaster* species subgroup, a model system in studies of speciation. Genetic variation of three interacting genes of the Drosophila sex determination pathway, *doublesex*, *transformer*, and *transformer-2*, was examined in detail. The evolutionary analysis of sequence variation in these three sex determination genes is expected to provide answers about their rates of evolution and mechanisms of evolutionary change as well as provide valuable information about the possible role of rapidly evolving genes from this sexual system in generating species diversity.

Chapter 2 examines the molecular evolution of the terminal sex determining

switch gene, *doublesex*. *dsx* is directly regulated by the presence or absence of the upstream regulators, TRA and TRA-2. As the furthest downstream locus in the primary sex determination pathway, this gene effects both male and female somatic sexual function by means of its sex-specific gene products, DSXm and DSXf. Hence, an adaptive change in any of its upstream transcription factors, if directed at primary sex determination, should be observed at this point in the genetic pathway. Also, *dsx*'s low expected divergence, given that it contains known conserved domains, can be used as a baseline against estimated divergences of the other rapidly evolving sex determining loci, *tra* and *tra-2*. In an internal comparison of divergence between species of the *D. melanogaster* subgroup, male-specific exons of *dsx* were found to possess higher levels of selective constraint compared to exons common to both males and females. This pattern is reversed, however, across larger phylogenetic distances and may indicate the presence of lineage-specific differences in selection.

Chapter 3 examines *transformer* evolution among sibling species of the *D. melanogaster* complex. TRA is expressed only in females and regulates the *dsx* locus to produce female-specific DSX isoforms. *tra* had previously been demonstrated to be rapidly evolving across Drosophila (O'Neil and Beloté 1992) and different evolutionary processes have been suggested to drive its divergence (Schaeffer and Walthour 1994; McAllister and McVean 2000). Among the sibling species of *D. melanogaster*, replacement changes are rapidly fixed between species and the presence of very large amino acid insertions, which drastically change the biochemical structure of this transcription factor, indicate that this protein can accommodate large disruptive changes.

Except for the significant presence of heterogeneity between amino acid sites, a neutral explanation was not rejected.

Chapter 4 examines the divergence and polymorphism of the ubiquitously expressed gene *transformer-2*, in the *D. melanogaster* complex. Its gene product interacts with TRA in females and is also necessary for male fertility. Using sibling species of the *D. melanogaster* subgroup, it was demonstrated that TRA-2, like TRA, also evolved in a rapid manner. Although arginine-serine (RS) domains seemed to remain constant between species, significant differences in evolutionary rate, likely caused by differences in selective constraint, were found between species. The low levels of replacement polymorphism in both *D. melanogaster* and *D. simulans* compared to the relatively higher levels of fixed replacement substitutions and higher variation at synonymous sites (both within and between species), indicate the possibility that adaptation may be driving the rapid fixation of its amino acids.

Finally, Chapter 5 is treated as a general discussion of the results from this thesis. In particular, I discuss the rapid evolution of sex determining genes in the context of sexual system diversity and speciation.

REFERENCES

Aguadé, M., Miyashita, N., and C.H. Langley (1992) Polymorphism and divergence in the *Mst26A* male accessory gland gene region in *Drosophila*. Genetics **132:** 755-770

Andersson, M.B. (1994) *Sexual selection*. Princeton University Press, Princeton, New Jersey

Arnqvist, G., Edvardsson, M., Friberg, U., and T. Nilsson (2000) Sexual conflict promotes speciation in insects. Proc. Nat. Acad. Sci. U.S.A. **97:** 10460-10464

Baker, B.S. (1989) Sex in flies: The splice of life. Nature **340:** 521-524

Bridges, C.A. (1925) Sex in relation to chromosomes. Am. Nat. **59:** 127-137

Bopp, D., Bell, L.R., Cline, T.W., and P. Schedl (1991) Developmental distribution of female-specific *Sex-lethal* proteins in *Drosophila melanogaster*. Genes Dev. **5:** 403-415

Bush, G.L. (1969) Sympatric host race formation and speciation in frugivorous flies of the genus *Rhagoletis*. Evolution **23:** 237-251

Carroll, S.B. (1995) Homeotic genes and the evolution of arthropods and chordates. Nature **376:** 479-485

Carroll, S.B., Grenier, J.K., and S.D. Weatherbee (2001) *From DNA to diversity: Molecular genetics and the evolution of animal design*. Blackwell Science, U.S.A.

Carson, H.L. (1968) The population flush and its genetic consequences. In *Population biology and evolution*. (Ed: R.C. Lewontin) pp. 123-137. Syracuse University Press, Syracuse, NewYork

Carson, H.L. (1997) Sexual selection: a driver of genetic change in Hawaiian *Drosophila*. J. Hered. **88:** 343-352

Civetta, A., and R.S. Singh (1995) High divergence of reproductive tract proteins and their association with postzygotic reproductive isolation in *Drosophila melanogaster* and *Drosophila virilis* group species. J. Mol. Evol. **41:** 1085-1095

Civetta, A., and R.S. Singh (1998a) Sex and speciation: Genetic architecture and evolutionary potential of sexual versus non-sexual traits in the sibling species of the *Drosophila melanogaster* complex. Evolution **52:** 1080-1092

Civetta, A., and R.S. Singh (1998b) Sex-related genes, directional sexual selection, and speciation. Mol. Biol. Evol. **15:** 901-909

Civetta, A., and R.S. Singh (1999) Broad-sense sexual selection, sex gene pool evolution, and speciation. Genome 42: 1033-1041

Clark, A.G., Aguadé, M., Prout, T., Harshman, L.G., and C. H. Langley (1995) Variation in sperm displacement and its association with accessory gland protein loci in *Drosophila melanogaster*. Genetics **139:** 189-201

Cline, T.W. (1985) Primary events in the determination of sex in *Drosophila melanogaster*. In *The origin and evolution of sex*. (Ed: H.O. Halvorson, and A. Monroy), pp. 301-327, Liss Publishing, New York

Cline, T.W., and B.J. Meyer (1996) Vive la différence: Males *vs.* females in flies *vs.* worms. Annu. Rev. Genet. **30:** 637-702

Coyne, J.A. (1992) Genetics and speciation. Nature **355:** 511-515

Coyne, J.A., and M. Kreitman (1986) Evolutionary genetics of two sibling species, *Drosophila simulans* and *D. sechellia*. Evolution **40:** 673-691

Coyne, J.A., and H.A. Orr (1989) Patterns of speciation in *Drosophila*. Evolution **43:** 362-381

Coyne, J.A., and H.A. Orr (1997) "Patterns of speciation in *Drosophila*" revisited. Evolution **51:** 295-303

Coyne, J.A., T.D. Price (2000) Little evidence for sympatric speciation in island birds. Evolution **54:** 2166-2171

Darwin, C.R. (1859) *On the origin of species by means of natural selection, or the preservation of the favoured races in the struggle for life*. John Murray, London

Darwin, C.R. (1871) *The descent of man, and selection in relation to sex*. John Murray, London

Darwin, C.R. (1888). *The life and letters of Charles Darwin: Including an autobiographical chapter*. John Murray, London

deBono, M., and J. Hodgkin (1996) Evolution of sex determination in *Caenorhabditis*: Unusually high divergence of *tra-1* and its functional consequences. Genetics **144:** 587-595

Dobzhansky, Th. (1936) Studies on hybrid sterility. II. Localization of sterility factors in *Drosophila pseudoobscura* hybrids. Genetics **21:** 566-570

Dobzhansky, Th. (1937) *Genetics and the origin of species.* Columbia University Press, New York

Dobzhansky, Th. (1970) *Genetics of the evolutionary process.* Columbia University Press, New York

Eberhard, W.G. (1985) *Sexual selection and animal genitalia.* Harvard University Press, Cambridge, MA

Eberhard, W.G. (1996) *Female control: Sexual selection by cryptic female choice.* Princeton University Press, Princeton, New Jersey

Erickson, J.W., and T.W. Cline (1991). Molecular nature of the *Drosophila* sex determination signal and its link to neurogenesis. Science **251:** 1071-7074

Ferris, P.J., and U.W. Goodenough (1997) Mating type in *Chlamydomonas* is specified by *mid*, the minus-dominance gene. Genetics **146:** 859-869

Fisher, R.A. (1915) The evolution of sexual preference. Eugenics Review **7:** 115-123

Fisher, R.A. (1930) *The genetical theory of natural selection.* Clarendon, Oxford

Haeckel, E. (1879) *The evolution of man: A popular exposition of the principal points of human ontogeny and phylogeny.* D. Appleton and Co., New York

Haldane, J.B.S. (1922) Sex-ratio and unisexual sterility in hybrid animals. J. Genetics **12:** 101-109

Haldane, J.B.S. (1932) The time of action of genes and its bearing on some evolutionary problems. Am. Natural. **66:** 643-646

Hawkins, J.R. (1994) Sex determination. Hum. Mol. Genet. **3:** 1463-1467

Hey, J., and R.M. Kliman (1993) Population genetics and phylogenetics of DNA sequence variation at multiple loci within the *Drosophila melanogaster* species complex. Mol. Biol. Evol. **10:** 804-822

Hodgkin, J. (1990) Sex determination compared in *Drosophila* and *Caenorhabditis*. Nature **344:** 721-728

Hodgkin, J. (1992) Genetic sex determination mechanisms and evolution. BioEssays **14:** 253-256

Hoekstra, H.E., Hoekstra, J.M., Berrigan, D., Vignieri, S.N., Hoang, A., Hill, C.E., Beerli. P., and J.G. Kingsolver (2001) Strength and tempo of directional selection in the wild. Proc. Natl. Acad. Sci. U.S.A. **98:** 9157-9160

Huxley, L. (1900) *Life and letters of Thomas Henry Huxley.* Appleton, New York

Huxley, J.S. (1942) *Evolution: The modern synthesis.* Allen and Unwin., London

Keyes, L.N., Cline, T.W., and P. Schedl (1992) The primary sex determination signal of *Drosophila* acts at the level of transcription. Cell **68:** 933-943

King, M.-C., and A.C. Wilson (1975) Evolution at two levels in humans and chimpanzees. Science **188:** 107-116

Kingsolver, J.G., Hoekstra, H.E. Hoekstra, J.M., Berrigan, D., Vignieri, S.N., Hill, C.E., Hoang, A., Gibert, P., and P. Beerli (2001) Strength of phenotypic selection. Am. Nat. **157:** 245-261

Kirkpatrick, M. (1985) Evolution of female choice and male parental investment in polygynous species: The demise of the "sexy son". Am. Nat. **125:** 788-810

Kliman, R. M., and J. Hey (1993) Reduced natural selection associated with low recombination in *Drosophila melanogaster*. Mol. Biol. Evol. **10:** 1239-1258

Kliman, R. M., P. Andolfatto, J. A. Coyne, F. Depaulis, M. Kreitman, A. J. Berry, J. McCarter, J. Wakeley, and J. Hey (2000) The population genetics of the origin and divergence of the *Drosophila simulans* complex species. Genetics **156:** 1913-1931

Kopp, A., I. Duncan, Godt, D., and S. B. Carroll (2000) Genetic control and evolution of sexually dimorphic characters in Drosophila. Nature **408:** 553-559

Kulathinal, R.J. and Singh, R.S. (1998) Cytological characterization of premeiotic versus postmeiotic defects producing hybrid male sterility among sibling species of the *Drosophila melanogaste*r complex. Evolution **52:** 1067-1079

Kulathinal, R.J., and R.S. Singh (2000a) A biogeographic genetic approach for testing

the role of reinforcement: the case of *Drosophila pseudoobscura* and *D. persimilis*. Evolution **54:** 210-217

Kulathinal, R.J., and R.S. Singh (2000b) Reinforcement with gene flow? Evolution **54:** 2176-2177

Kuwabara, P.E. (1996) Interspecies comparison reveals evolution of control regions in the nematode sex-determining gene *tra-2*. Genetics **144:** 597-607

Lachaise, D., David, J.R., Lemeunier, F., Tsacas, L., and M. Ashburner (1986) The reproductive relationships of *Drosophila sechellia* with *D. mauritiana*, *D. simulans*, and *D. melanogaster* from the Afrotropical region. Evolution **40:** 262-271

Lamarck, J.B. (1984) *Zoological philosophy: An exposition with regard to the natural history of animals*. Translation of J.B. Lamarck (1809) *Philosophie zoologique, ou exposition des considérations relatives à l'histoire naturelle des animoux* by Hugh Elliot. University of Chicago Press, Chicago

Lande, R. (1981) Models of speciation by sexual selection on polygenic traits. Proc. Natl. Acad. Sci. U.S.A. **78:** 3721-3725

Laurie, C.C. (1997) The weaker sex is heterogametic: 75 years of Haldane's rule. Genetics **147:** 937-951

Lewontin, R.C. (1974) *The genetic basis of evolutionary change*. Columbia University Press, New York

Lovejoy, A.O. (1959) Herder: Progressionism without transformism. In *Forerunners of Darwin*. (Eds: B. Glass, O. Temkin and W.L. Straus) Pg. 207-221. John Hopkins Press, Baltimore, MD

Lovejoy, A.O. (1964) *The Great Chain of Being*. Harvard University Press, Cambridge, MA

Ludwig, M.Z., and M. Kreitman (1995) Evolutionary dynamics of the enhancer region of *even-skipped* in Drosophila. Mol. Biol. Evol. **12:** 1002-1011

Ludwig, M. Z., N. H. Patel, and M. Kreitman (1998) Functional analysis of *eve stripe 2* enhancer evolution in Drosophila: Rules governing conservation and change. Development **125:** 949-958

Ludwig, M.Z., Bergman, C., Patel, N.H., M. Kreitman (2000) Evidence for stabilizing

selection in a eukaryotic enhancer element. Nature **403:** 564-567

Malthus, T.R. (1799) *An essay on the principle of population as it affects the future improvement of society, etc.* J. Johnson, London; reprinted in 1926, Macmillan, London

Marín, I., and B.S. Baker (1998) The evolutionary dynamics of sex determination. Science **281:** 1990-1994

Maynard Smith, J. (1966) Sympatric speciation. Am. Nat. **100:** 637-650

Mayr, E. (1942) *Systematics and the origin of species.* Columbia University Press, New York

Mayr, E. (1963) *Animal species and evolution.* Belknap Press, Cambridge, MA

McAllister, B.F. and G.A.T. McVean (2000) Neutral evolution of the sex-determining gene *transformer* in Drosophila. Genetics **154:**1711-1720

Muller, H.J. (1940) Bearing of the *Drosophila* work on systematics. In *The new systematics.* (Ed: J.S. Huxley) pp. 185-268. Clarendon Press, Oxford

Muller, H.J. (1942) Isolating mechanisms, evolution, and temperature. Biological Symposium **6:** 71-125

Nusslein-Volhard, C. (1994) Of flies and fishes. Science **266:** 572-574

O'Neil, M.T., and J.M. Beloté (1992) Interspecific comparison of the *transformer* gene of *Drosophila* reveals an unusually high degree of evolutionary divergence. Genetics **131:** 113-128

Orr, H.A. (1995) The population genetics of speciation: The evolution of hybrid incompatibilities. Genetics **139:** 1805-1813

Orr, H.A. (1997) Haldane's rule. Annual Review of Ecology and Systematics **28:** 195-218

Peterson, K.J., and E.H. Davidson (2000) Regulatory evolution and the origin of the bilaterians. Proc. Natl. Acad. Sci. U.S.A. **97:** 4430-4433

Plato (1901) *The Republic*, Translation of Benjamin Jowett, web edition at http://www.mdx.ac.uk/www/study/xpla0.htm

Powell, J.R. (1997) *Progress and prospects in evolutionary biology: The Drosophila model*, Oxford University Press, Oxford

Prout, T. (1971) The relation between fitness components and population prediction in *Drosophila*. I. The estimation of fitness components. Genetics **68:** 127-149

Purugganan, M.D. (1998) The molecular evolution of development. Bioessays **20:** 700-711

Raff, R.A., and T.C. Kaufman (1983) *Embryos, genes and evolution.* Macmillan Publishing Co. Inc., New York

Rice, W.R. (1996) Sexually antagonistic male adaptation triggered by experimental arrest of female evolution. Nature **381:** 232-234

Serres, E.R.A. (1824) Explication du système nerveux des animaux invertébrés. Ann. Sci. Nat. **3:** 377-380

Singh, R.S., and R.J. Kulathinal (2000) Sex gene pool evolution and speciation: A new paradigm. Genes. Genet. Syst. **75:** 119-130

Slatkin, M. (1973) Gene flow and selection in a cline. Genetics **75:** 733-756

Swanson, W.J., Yang, Z., Wolfner, M.F., and C.F. Aquadro (2001) Positive Darwinian selection drives the evolution of several female reproductive proteins in mammals. Proc. Natl. Acad. Sci. U.S.A.. **98:** 2509-2514

Tautz, D. (2000) Evolution of transcriptional regulation. Curr. Opin. Genet. Dev. **10:** 575-579

Templeton, A.R. (1980) The theory of speciation via the founder principle. Genetics **94:** 1011-1038

Ting, C.-T., Tsaur, S.-C., Wu, M.-L., and C.-I. Wu (1998) A rapidly evolving homeobox at the site of a hybrid sterility gene. Science **282:** 1501-1504

Tsaur, S.-C., and C.-I. Wu (1997) Positive selection and the molecular evolution of a gene of male reproduction, *Acp26Aa* of *Drosophila*. Mol. Biol. Evol. **14:** 544-549

Tucker, P.K., and B.L. Lundrigan (1993) Rapid evolution of the sex determining locus in Old World mice and rats. Nature **364:** 715-717

Turelli, M., and H.A. Orr (1995) The dominance theory of Haldane's rule. Genetics **140:**

389-402

von Baer, K.E. (1828) *Über entwickelungsgeschichte der thiere: Beobachtung und reflexion.* Vol. I, Borntrager, Konigsberg, Translation of *Fifth scholium* by T.H. Huxley, published as part of Fragments relating to philosophical zoology: Selected from the works of K.E. von Baer. In *Scientific memoirs, Selected from the transactions of foreign academies of science, and from foreign journals: Natural history.* (Eds: A. Henfrey, and T.H. Huxley, 1853). Taylor and Francis, London

Walthour, C.S., and S.W. Schaeffer (1994) Molecular population genetics of sex determination genes: The *transformer* gene of *Drosophila melanogaster.* Genetics **136:** 1367-1372

Whitfield, L.S., Lovell-Badge, R., and P.N. Goodfellow (1993) Rapid sequence evolution of the mammalian sex-determining gene SRY. Nature **364:** 713-715

Wilson, A.C., L.R. Maxson, and V.M. Sarich (1974) Two types of molecular evolution. Evidence from studies of interspecific hybridization. Proc. Natl. Acad. Sci U.S.A. **71:** 2843-2847

Wright, S. (1931) Evolution in Mendelian populations. Genetics **16:** 97-159

Wu, C.-I., and A.W. Davis (1993) Evolution of postmating reproductive isolation: The composite nature of Haldane's rule and its genetic bases. Am. Natural. **142:** 187-212

Zahavi, A. (1975) Mate selection- a selection for a handicap. J. Theor. Biol. **67:** 603-605

Zeng, L.-W. (1996) Resurrecting Muller's theory of Haldane's rule. Genetics **143:** 603-607

CHAPTER TWO

The terminal regulator of sex determination, *doublesex*, is conserved among species of the *D. melanogaster* subgroup

ABSTRACT

The *doublesex* (*dsx*) locus plays a dual role in the regulation of sex determination in Drosophila. In females, the female-specific isoform, DSXf, represses the expression of male differentiation while in males, the male-specific DSXm isoform represses the expression of genes involved in female differentiation (Burtis and Baker 1989). To function, DSX forms a homodimer via the interaction of its oligomerization domains - a common OD1 domain that is found in both sexes and a sex-specific OD2 domain located at the DSX carboxy-terminus. Although sex determination genes expressed immediately upstream to the *dsx* locus were found in this thesis to be rapidly evolving, DSX appears to be relatively conserved - particularly in its protein oligomerization domains - across distant taxa. Using seven species of the *D. melanogaster* subgroup, sequence divergence was assessed at the *dsx* locus. Overall protein divergence is low. All replacement substitutions are found outside the conserved oligomerization domains, OD1 and OD2. The noncoding *dsx pre*mRNA repeat element (*dsx*RE) region is also relatively conserved between species of the *D. melanogaster* subgroup. This *dsx*RE conservation among these species contrasts to *D. virilis* divergence. Although repeat units were identical, *D. virilis* possessed only four out of the six repeats found in the *D. melanogaster* subgroup. To compare evolutionary rates across the dsx protein, a sex-specific region of the *dsx* locus was compared to a region of *dsx* which is common to both sexes. Using Dipteran sequences from *Megaselia*, *Bactrocera* and *D. melanogaster*, the OD1 domain common to both DSX isoforms were compared to male-specific portions of the OD2 domain.

While OD1 appears moderately conserved across Dipterans, the male-specific carboxy terminus of OD2 has low similarity. In another comparison using the seven species of the *D. melanogaster* subgroup, divergence of the sequenced common exon was compared to the sequenced male-specific exon. In terms of nonsynonymous divergence, the male-specific exon was found to be, in contrast to the Dipteran lineages, less diverged on average than the common exon. Using maximum likelihood, selective constraints were estimated to be larger in the male-specific exon ($d_N/d_S = 0.04$) compared to the common exon ($d_N/d_S = 0.09$). The differences in divergence between sex- and non-sex-specific regions from these two phylogenetic comparisons may suggest the presence of variable selection pressures between these lineages. Intraspecific variation in *D. melanogaster* was found to be lower in the second common exon relative to the male-specific exon. Among six geographically diverse lines of *D. melanogaster*, only five segregating sites were observed (2 being replacement polymorphisms) over a 1332 base pair sequence.

INTRODUCTION

In most eukaryotes, biparental reproduction is a common evolutionary strategy. The determination of sex generally takes place early in the development of the organism and is coordinated by a small number of genes. These genes, in turn, regulate a cascade of downstream genes and genetic pathways which ultimately control sex-specific aspects of development, physiology and behaviour.

Much of our understanding of the genetic mechanisms of sex determination is based on the extensively studied model organism, *Drosophila melanogaster* (for an excellent review, see Cline and Meyer 1996). Genetic and molecular analyses have been used to acquire a precise understanding of the genetic pathways involved in the early decision of sexual identity, particularly somatic sexual differentiation (see Introduction, Figure 1.1). The primary signal is found in the X-chromosome to autosome ratio. In females (X/X;A/A), the double dose of X-chromosomal genes allows for the production of functional SXL protein. SXL, in turn, splices *transformer* (*tra*) *pre*mRNA (Handa *et al.* 1998) into full length transcripts which translate to functional TRA protein. TRA and the ubiquitously expressed TRA-2 then bind to *dsx pre*mRNA to form female-specific isoforms, which regulate a whole cascade of downstream genes involved in female somatic sexual differentiation (Nagoshi *et al.* 1988). In males, the single dose of X-linked genes does not allow for the production of functional SXL and therefore functional TRA is not produced. Without TRA, *dsx pre*mRNA is spliced through default mechanisms and a male DSX isoform is produced (Figure 1.1).

As the final gene in the regulatory hierarchy that controls sexual differentiation in Drosophila, *dsx* plays a pivotal role in the development of male and female somatic sexual identity. Classic genetic screens have uncovered the dichotomous nature of *dsx* expression and subsequent production of sex-specific isoforms (Baker and Ridge 1980). These zinc-finger transcription factors target downstream genes that are involved in various aspects of somatic sexual differentiation including courtship behaviour (Villella and Hall 1996), morphological dimorphisms and developmental differences (Burtis and Baker 1989). Sex-specific targeting of *yolk-protein2* (*yp-2*) has become the best example of *dsx* regulation (Burtis and Wolfner 1992). Male isoforms of DSX repress transcription of this gene involved in yolk protein production during embryogenesis while female isoforms allow for *yp-2* expression (Coschigano and Wensink 1993). Hence, *dsx* acts as a double-switch gene that "selects between two alternative sexual programmes" (Schütt and Nöthiger 2000).

While the primary genetic hierarchy of sex determination is a critically important genetic pathway, sex determination mechanisms appear to be quite variable among taxa (Hodgkin 1992; Marín and Baker 1998). The primary trigger can be either genetic or environmental and a wide diversity of regulatory genes and mechanisms exists (Table 1.1). The observed plasticity in evolutionary strategies employed to decide sexual fate is seen across distant taxa as well as within species. For example, in the common housefly, *Musca domestica*, sex determination is controlled by either a single gene or chromosomal ratio, depending on the strain used (Dübendorfer *et al.* 1992). A similar situation exists in the lemming, *Myopus schisticolor* (Fredga 1994).

In addition to the variation found among sex determination systems, many sex determination genes have been found to be rapidly evolving. In nematodes, the sex determining loci, *tra-1* and *tra-2*, are among the most diverged genes between, *C. elegans* and *C. brigassae* (deBono and Hodgkin 1996; Kuwabara and Hodgkin 1996). *Sry*, the master control gene of sex determination in mammals has also been demonstrated to be rapidly evolving among rodents and primates (Tucker and Lundrigan 1993; Whitfield *et al.* 1993). In Drosophila, O'Neil and Beloté (1992) found that *transformer* (*tra*), one of the sex determining transcription factors that interacts with *dsx pre*mRNA, is among the most diverged genes in Drosophila.

It must be noted that while a handful of sex determination genes have been found to be rapidly evolving, most others are relatively conserved. Interestingly, the sex determination loci that are most diverged are usually situated at the top of the genetic pathway (i.e. upstream regulators). One evolutionary hypothesis to explain this pattern of variation between sex determination loci has been proposed by Wilkins (1995). In his model, downstream regulatory genes in sex determination are suggested to be more ancient and that upstream genes have been recruited more recently. Hence, upstream loci would tend to be less similar to other upstream recruits from different taxa. The tolerance of this part of the sex determination pathway to different regulators suggests that selective constraints on these genes may be quite low. Another explanation of the pattern of rapidly evolving upstream regulators in sex determination is that downstream genes may control a greater repertoire of genes and, therefore, be more constrained (Marín and Baker 1998).

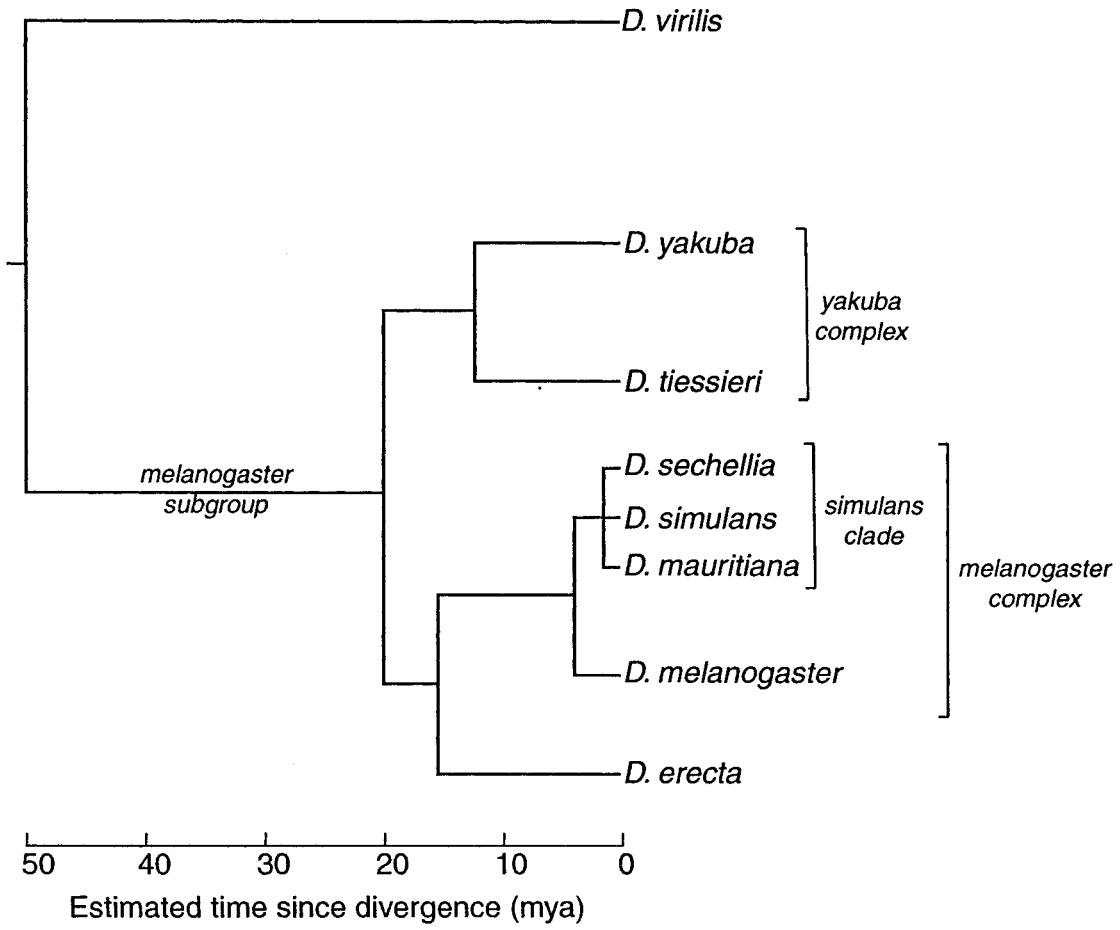
Comparative studies have indicated that regions of the *dsx* locus that code for certain protein domains are conserved across distant taxa (Raymond *et al.* 1998; Smith *et al.* 1999). On the other hand, diverged regions have also been found. Hertel *et al.* (1996) demonstrated structural differences between *D. virilis* and *D. melanogaster* in the *pre*mRNA repeat region used to promote binding of the TRA/TRA-2 mediated spliceosomal complex in females. In order to better understand the evolutionary constraints at the *dsx* locus, seven species of the *D. melanogaster* subgroup were sequenced in selected regions. This comparative approach allows differences in divergence of various regions of the locus to be assessed. I report that *dsx* is indeed a conserved locus in the sex determination pathway. Neither protein nor regulatory regions sequenced revealed much change across the *D. melanogaster* subgroup. However, differences in selective constraints were observed between common and sex-specific regions of the gene and appear to change between evolutionary lineages.

MATERIALS AND METHODS

Isolation of DNA from species of the *D. melanogaster* subgroup: In order to assess DSX divergence in closely related species of the *D. melanogaster* subgroup, the *dsx* locus was partially sequenced in seven species of the *D. melanogaster* subgroup (Figure 2.1). Unless otherwise stated, flies were originally obtained from the Drosophila Species Stock Centre in Bowling Green. This group of species (stock centre designations are indicated in parentheses) includes the more distantly related species, *D. yakuba* (14021-0261.0), *D. tiessieri* (14021-257.0), and *D. erecta* (S-18, kindly provided by John Roote, Cambridge University), as well as the sibling species, *D. sechellia* (14021-0248.3), *D. mauritiana* (0241.2), *D. simulans* (0251.2), and the published sequence of *D. melanogaster* (GenBank accession number M25292). Several lines of *D. melanogaster* were used to determine within species polymorphism in the common second exon and male-specific fifth exon (both exons are separated by over 25 kb) including the laboratory strains Canton-S (provided by Ana Campos, McMaster University), InAB (obtained from UMEA Stock Centre, Sweden) and isofemale lines originating from populations in Hawaii (14021-0231.0), Peru (14021-0231.1), India (14021-0231.6), Pennsylvania (CPA46, kindly provided by Brian Lazzaro, Penn State), and Zimbabwe (Z(H)12 and Z(H)34), provided by B.L.). Fly stocks were maintained at low density between 22-23° in banana medium on a diurnal 12h dark/light cycle.

Extraction of genomic DNA was implemented using a single fly procedure (Gloor and Engels 1992). One fly is ground in 10 mM Tris-Cl (pH 8.2), 25 mM NaCl, 1 mM

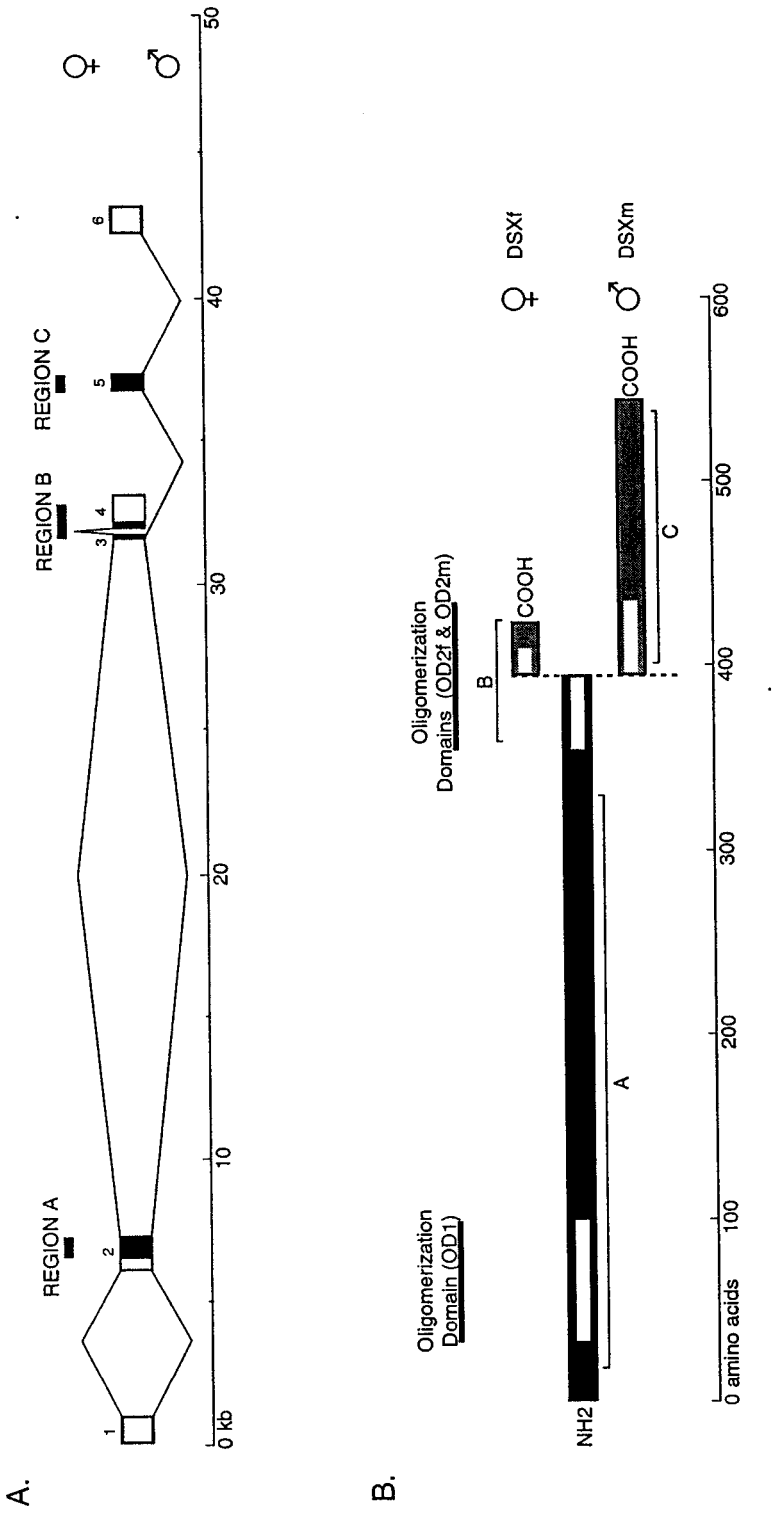
Figure 2.1 Phylogenetic relationship of members from the *Drosophila melanogaster* subgroup. *D. virilis* is also indicated. Divergence times are from Powell (1997).


D. virilis
D. yakuba
D. tiessieri
yakuba complex
melanogaster subgroup
D. sechellia
D. simulans
D. mauritiana
simulans clade
melanogaster complex
D. melanogaster
D. erecta
50
40
30
20
10
0
Estimated time since divergence (mya)

EDTA and 200 mg/ml proteinase-K and placed at room temperature for 30 minutes. As a final step, this mixture is heated at 95° for three minutes.

Amplification and sequencing of the *dsx* locus: Since the *dsx* locus spans over 40 kb (see Figure 2.2A), I employed a strategy to simply sample coding regions which translate to male and female-specific DSX isoforms and designed ubiquitous primers (from the published *D. melanogaster* sequence, Burtis and Baker 1989) which amplified coding regions in all seven species. Only exons with coding sequences were sequenced and amplified, thereby restricting this analysis to just three regions (Figure 2.2A). With respect to the published sequence of *D. melanogaster*, this sequencing strategy effectively covers 86.5% of the total coding region, including 471 out of the total 549 amino acids found in male-specific isoforms and 372 out of 427 amino acids found in female-specific isoforms. Region A includes amino acids 17 to 328 which originate from common exon 2 and is amplified using the external primers, 5'GGAGGAGAACTGGAATAGCG3' and 5'AACTGAGATCGGCAAAATGG3', plus an internal set of primers for sequencing, 5'AGTGTGGGTGGCTTTGGTG3' and 5'GACATGTCCTGCACCACCAG3'. Region B includes amino acids 367-397 from common exon 3 and amino acid sites 398-427 originating from the female-specific exon. Region B also contains intron 3 as well as noncoding portions of the female-specific exon which includes the *dsx*RE region. Its primary sequence is amplified using the flanking primers, 5'GCCAAGACGTTTTCCTAGAC3' and 5'TCGCAAGACATCGATGAAAC3'. Internal primers (for sequencing), 5'GCTGAGATGTCTGGCCTC3' and 5'AGATCCGTTTGCGGATTG3', were also employed in this region. Region C includes

Figure 2.2 Transcribed and translated products of *doublesex* in *D. melanogaster*. A. Gene structure and sex-specific splicing of the *dsx* locus. Exons represented as boxes and numbered above. Solid boxes denote coding portion of the exons. Female and male-specific mRNA transcripts are shown above and below, respectively. The three sequenced regions (A, B, and C) used in this study are indicated above. B. Sex-specific *dsx* products, DSXf and DSXm. Both isoforms have a common amino-terminus and sex-specific carboxy-termini. Protein binding domains (OD1, OD2m/f) are represented as open boxes. Sequenced regions (A, B, C) are also indicated.


A.
REGION A
REGION B
REGION C
1
2
3
4
5
6
♀
♂
0 kb
10
20
30
40
50
B.
Oligomerization Domain (OD1)
Oligomerization Domains (OD2f & OD2m)
B
A
C
COOH
COOH
NH2
DSXf
DSXm
♀
♂
0 amino acids
100
200
300
400
500
600

amino acid sites 410-541 from male-specific exon 5 and amplification of its sequence is accomplished with the primer pair, 5'TCGAGTGGAAATAAATCGCA3' and 5'CGTTGCGATACTGCTACGTG3'. All primers were constructed at the Molecular Biology Institute of McMaster University (MOBIX).

Using flanking primers for each of the particular regions listed above, the *dsx* locus was partially amplified between members of the *D. melanogaster* subgroup using a Perkin-Elmer 480 thermocycler. A hotstart and touchdown protocol was used in all cases with cycles based on a 95°/1 minute dissociation and 72°/1 minute extension, however, annealing conditions varied according to which region was amplified. For Region A, a $MgCl_2$ concentration of 1.5mM and a decremental series of cycles starting at 62° and ending at 54° (2 cycles per degree Celsius except for 3 cycles at both 59° and 58°) was used. A similar touchdown protocol, but $MgCl_2$ concentration of 2.0 mM, was used for Region B. Region C ($MgCl_2$=1.25mM) also employed a touchdown procedure starting at 64° and terminating at 56° (2 cycles per degree Celsius except for 3 cycles each at 60°, 59°, and 58°). 10X PCR buffer, $MgCl_2$ and Taq polymerase were supplied by Fermentas (Burlington, ON). Amplifications performed on populations of *D. melanogaster* to assess levels of intraspecific variation used similar conditions as above, but on a different thermocycler, an MJ Research PT-200.

All amplification products were run on 1% agarose gels with EtBr to check for amplification specificity and purified using a QIAGEN DNA purification kit. Sequencing was performed on an ABI PRISM® 377 DNA sequencer using both flanking and internal primers (as described above) on both strands.

Sequence analyses: Sequence validity was confirmed by sequencing both strands. Sequences were aligned using ClustalX (Thompson *et al.* 1997) and all variable sites were again confirmed. Coding regions were identified in-frame as defined by the *D. melanogaster* published sequence (Burtis and Baker 1989). Variable nucleotide and amino acid sites were found using MEGA ver. 2.0b (Kumar *et al.* 2000). For comparisons of the *dsx*RE region across Diptera, sequences were obtained from the following sources; *D. melanogaster*, Burtis and Baker (1989), Inoue *et al.* (1992); *D virilis*, Hertel *et al.* (1996); *Megaselia scalaris*, Kuhn *et al.* (2000); *Bactrocera tryoni*, Shearman and Frommer (1998). Neighbor-joining trees (Saitou and Nei 1987) were constructed using the Phylip package of programs ver 3.5 (Felsenstein 1993) and visualized with TREEVIEW ver. 1.61 (Page 2000). Two estimates of nucleotide diversity, θ, determined from the number of segregating sites in a sample of genes (Watterson 1975), and π, the average pairwise difference between haplotypes (Nei 1987), were calculated using DnaSP v3.51 (Rozas and Rozas 1999). Variances are calculated as in Nei (1987). For estimates of interspecific divergence, the proportion of synonymous (Ks) and nonsynonymous (Ka) substitutions per site (Nei and Gojobori 1986) were also calculated.

Evolutionary models (clock *vs.* no clock and d_N/d_S heterogeneity between lineages *vs.* constant d_N/d_S) were compared using a maximum likelihood approach implemented by the program CODEML in the PAML package v3.0b (Yang 1997). Likelihood ratio tests were performed to examine whether a particular model produces a significantly better fit to the data. PAML also generates estimates of the parameter, $\omega=d_N/d_S$.

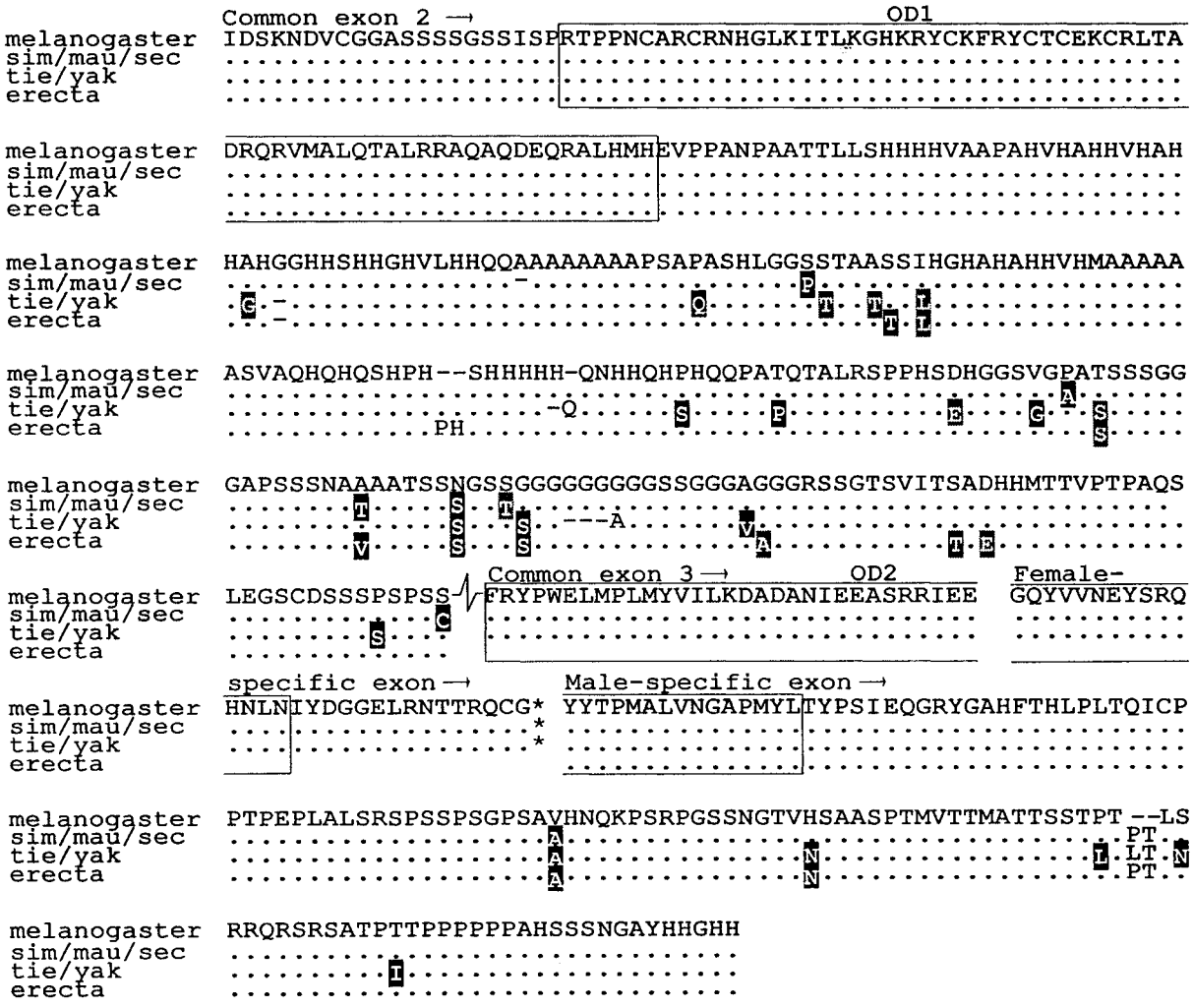
RESULTS AND DISCUSSION

Conservation of the oligomerization domains in the *D. melanogaster* subgroup: As terminal regulators of the sex determination hierarchy in Drosophila, sex-specific isoforms of DSX (Figure 2.2B) play a critical role in sexual differentiation. Genetic analyses using *dsx* mutants have indicated their involvement in controlling many aspects of somatic sexual differentiation including differences in pigmentation of the fifth and sixth abdominal tergites, courtship behaviour (Villella and Hall 1996), yolk protein expression in oogenesis (Bownes 1994), development of the nervous system (Truman 1992) and differences in genitalia and secondary sexual traits such as male sex combs.

In order to study divergence between members of the *D. melanogaster* subgroup at the *dsx* locus, a sequencing strategy was employed that maximized the amount of coding region sequenced. Although the coding region of the *dsx* locus spans over 30 kb (Figure 2.2A), nucleotide sequence encompassing over 86% of the *dsx* protein code from seven species of the *D. melanogaster* subgroup was collected allowing us to observe approximately 20 million years of evolutionary change on synonymous and nonsynonymous variation at this locus.

Figure 2.3 shows an alignment of the partially sequenced dsx protein from species of the *D. melanogaster* subgroup. Its two protein dimerization domains, OD1 and OD2 (both male and female-specific) are completely devoid of substitutions in this subgroup. Recent studies have demonstrated that these domains are important in DSX homodimerization. *In vitro* crosslinking assays and *in vivo* yeast-two-hybrid system

Figure 2.3 DSX protein alignment in the *D. melanogaster* subgroup. Variable amino acid sites which are translated from the three sequenced regions of the *dsx* locus (indicated in Figure 2.1) are shown. Sibling species of the *D. melanogaster* complex, *D. simulans*, *D. mauritiana*, and *D. sechellia*, are grouped together, as are the more distant members of the *yakuba* complex, *D. tiessieri*, and *D. yakuba*. Variable sites are shown if at least one amino acid from a particular group differs from the published *D. melanogaster* sequence. Lightly shaded sites denote similar amino acid substitution. Exons of protein origin are indicated. Oligomerization domains, OD1 and OD2 are represented in boxes. OD2 has a sex-specific carboxy terminus. Astericks denote stop codons. Gaps indicate region of protein not sequenced in this study.


Common exon 2 →
OD1
melanogaster IDSKNDVCGGASSSSGSSISPRTPPNCARCRNHGLKITLKGHKRYCKFRYCTCEKCRLTA
sim/mau/sec
tie/yak
erecta
melanogaster DRQRVMALQTALRRAQAQDEQRALHMHEVPPANPAATTLLSHHHHVAAPAHVHAHHVHAH
sim/mau/sec
tie/yak
erecta
melanogaster HAHGGHHSHHGHVLHHQQAAAAAAAAPSAPASHLGGSSTAASSIHGHAHAHHVHMAAAAA
sim/mau/sec
tie/yak
erecta
melanogaster ASVAQHQHQSHPH--SHHHHH-QNHHQHPHQQPATQTALRSPPHSDHGGSVGPATSSSGG
sim/mau/sec
tie/yak
erecta
melanogaster GAPSSSNAAAATSSNGSSGGGGGGGGGSSGGGAGGGRSSGTSVITSADHHMTTVPTPAQS
sim/mau/sec
tie/yak
erecta
Common exon 3 →
OD2
Female-
melanogaster LEGSCDSSSPSPSSFRYPWELMPLMYVILKDADANIEEASRRIEEGQYVVNEYSRQ
sim/mau/sec
tie/yak
erecta
specific exon →
Male-specific exon →
melanogaster HNLNIYDGGELRNTTRQCG*YYTPMALVNGAPMYLTYPSIEQGRYGAHFTHLPLTQICP
sim/mau/sec
tie/yak
erecta
melanogaster PTPEPLALSRSPSSPSGPSAVHNQKPSRPGSSNGTVHSAASPTMVTTMATTSSTPT--LS
sim/mau/sec
tie/yak
erecta
melanogaster RRQRSRSATPTTPPPPPPAHSSSNGAYHHGHH
sim/mau/sec
tie/yak
erecta

studies (Chen 1995; An *et al.* 1996; Erdman *et al.* 1996; Cho and Wensink 1997) revealed the presence of DSX protein dimers (in both males and females). Also corroborating the hypothesis that DSX isoforms act in dimerized form was the determination that DSX binds to symmetrical sequence on its mRNA targets (Erdman *et al.* 1996). It appears that OD1 and OD2 are necessary to dimerize full length DSX isoforms as well as to form a dimeric DNA binding unit required for the activation or repression of transcription in downstream gene targets (An *et al.* 1996).

In an evolutionary analysis of the *doublesex* locus using distantly related species of non-Drosophilids, Kuhn *et al.* (2000) also demonstrated the conservation of oligomerization domains, relative to the remainder of the protein. The phorid fly, *Megaselia scalaris*, the tephrytid fly, *Bactrocera tryoni*, and the silk moth, *Bombyx mori*, all showed protein homology in OD1 and the common and female-specific portion of OD2 when aligned with *D. melanogaster*. In fact, conservation of this domain extends beyond Diptera. A slightly smaller region of OD1 (termed the "minimal DNA-binding domain") has recently been shown to have high similarity to MAB-3 domains in the nematode, *Caenorhabitis elegans* (Raymond *et al.* 1998). Not only do these genes from *D. melanogaster* and *C. elegans* code for similar protein domains, but both control sex-specific neuroblast differentiation and female-specific yolk protein synthesis. It was further demonstrated that male DSX isoforms can rescue *mab-3* null mutants. A testis-expressed gene, *DMT1*, encoding a similar minimal DNA-binding domain was also isolated in humans (Raymond *et al.* 1998). Recently, Smith *et al.* (1999) isolated genes from reptiles, chickens and mice having similar minimal DNA-binding domain as DSX.

Furthermore, they demonstrated that these genes possess sexually dimorphic expression patterns indicating the remarkable conservation of both structure and function across a wide range of taxa.

The distribution of amino acid substitutions among members of the *D. melanogaster* subgroup reveals much about the levels of selective constraints found across the DSX protein. The immediate regions surrounding the OD1 and OD2 domains are also conserved (Figure 2.3). The sequenced region of common exon 3, the entire female-specific exon and much of the 5' end of the sequenced male-specific exon lack nonsynonymous substitutions which suggests the presence of purifying selection on these particular regions of the dsx protein.

Conservation of *dsx pre*mRNA enhancer region: The *transformer* locus, which encodes an upstream transcription factor that interacts with *dsx pre*mRNA, has been previously found to be rapidly evolving in the genus Drosophila (O'Neil and Beloté 1992). Two contrasting evolutionary mechanisms have been suggested to cause this divergence. Walthour and Schaeffer (1994) argued that the low polymorphism found in *D. melanogaster* implies that positive selection may have been responsible for its rapid evolution. McAllister and McVean (2000), by comparing *tra* evolutionary rates, suggested high rates of neutral evolution. Since the only known function of TRA is in its association to *dsx pre*mRNA via its interaction with other serine-arginine (SR) proteins such as TRA-2, one might expect that the enhancer region may have also diverged as a coevolutionary response. As shown above, the downstream protein target, DSX, is relatively conserved relative to its upstream regulator, *tra*. Among the four sibling

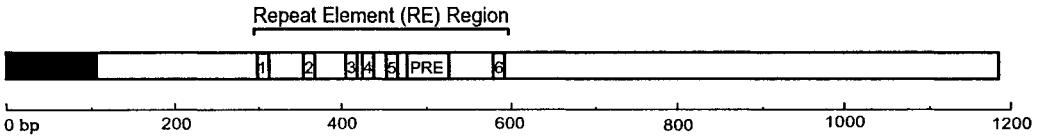
species of the *D. melanogaster* complex, only 1.4% of the total amino acid sites sampled from the DSX protein were variable while 8.7% of amino acid sites in TRA were variable among the same four species (see Chapter 3). In order to detect any coevolutionary process between the rapidly evolving TRA and its regulatory *pre*mRNA target, divergence was assessed in the enhancer region where this binding takes place.

Six repeats (each 13 nucleotides long), comprising the enhancer region, are found in the 3' untranslated part of the female-specific exon (Figure 2.4). Except for a single substitution in *D. sechellia*'s first repeat and two dinucleotide substitutions at precisely the same nucleotide sites in repeat 5 in *D. simulans* and *D. tiessieri* and *D. yakuba*, the enhancer region is identical among species of the *D. melanogaster* subgroup. In terms of enhancer region structure, *dsx pre*mRNA of the distantly related fruitfly, *D. virilis*, is quite different from that found in the *D. melanogaster* subgroup. *D. virilis* contains four identical repeat units compared to six repeat units found in the *D. melanogaster* subgroup. The *D. virilis* repeat likely represents the ancestral sequence since the Dipterans, *Megaselia* and *Bactrocera*, also contain repeats that are identical to the common *D. virilis* repeat (Figure 2.4C). The overall primary conservation of these noncoding repeats over time reveals the importance of proper alternative splicing in the female to produce solely female-specific isoforms.

Comparison of variation between common and male-specific exons: The presence of highly similar domains across vertebrate and invertebrate taxa in genes which have sexually dimorphic expression patterns, suggest that *dsx* has a conserved function (Raymond *et al.* 1998; Smith *et al.* 1999). As female DSX isoforms repress the

Figure 2.4 Conservation of the *dsx* splicing enhancer region (*dsx*RE). A. Structure of female-specific exon 4 which includes the *dsx pre*RE in *D. melanogaster*. Six repeats of 13 nucleotides are found in the female-specific exon. A purine-rich element (PRE) is found between repeats five and six. Solid boxes denote coding portion of exon B. Comparison of repeat elements from species of the *Drosophila melanogaster* subgroup and *D. virilis*. The number of nucleotides found between repeats are indicated. C. Alignment of the *dsx* repeat elements across Diptera. Among species of the *D. melanogaster* subgroup, the published *D. melanogaster* sequence is used as the consensus sequence and sites that are variable between subgroup species are indicated in lowercase. The most similar Diperan repeat elements were compared to repeats of the *D. melanogaster* consensus sequence.

A.



B.

	Repeat 1		Repeat 2		Repeat 3	
D. melanogaster	TCTTCAATCAACA	<44>	TCTTCAATCAACA	<44>	TCTACAATCAACA	<8>
D. simulans	.....	<48>	.....	<44>	.....	<8>
D. mauritiana	.....	<48>	.....	<44>	.....	<8>
D. sechellia	.....C.....	<48>	.....	<44>	.....	<8>
D. tiessieri	.....	<47>	.....	<38>	.....	<5>
D. yakuba	.....	<47>	.....	<38>	.....	<5>
D. virilis	.....	<29>	.....	<34>	----------	

	Repeat 4		Repeat 5		Repeat 6
D. melanogaster	TCTTCAATCAACA	<18>	TCAACAATCAACA	<92>	TCAACGATCAACA
D. simulans	.....	<18>	......CG.....	<93>	.....
D. mauritiana	.....	<18>	.....	<93>	.....
D. sechellia	.....	<18>	.....	<93>	.....
D. tiessieri	.....	<18>	......TG.....	<86>	.....
D. yakuba	.....	<18>	......TG.....	<86>	.....
D. virilis	.....	<24>	..TT.....		----------

C.

melanogaster group	
Repeat 1	TCTTCaATCAACA
Repeat 2	.....
Repeat 3	...A.....
Repeat 4	.....
Repeat 5	..AA..at.....
Repeat 6	..AA.G.....
D. virilis (Repeats 1-4)	.....
Megaselia (Repeat 1)	.....
Bactrocera (Repeat 4)	.....

transcription of downstream male-specific loci and male DSX isoforms repress the transcription of downstream female-specific loci, both DSXf and DSXm are critical to somatic sexual differentiation in Drosophila. An attempt was made to compare the evolutionary rates between domains that have a sex-specific function to those that are common in both sexes. In the first comparison, only the conserved oligomerization domains were compared between the three Dipterans, *Megaselia*, *Bactrocera* and *Drosophila* (the published sequence of *D. melanogaster* was used). Unlike other portions of the *dsx* locus, these protein domains produced unambiguous alignments necessary to calculate a reliable set of identities. Among these three species, a highly conserved OD1 domain (88% identity) and a moderately conserved OD2 domain that is common in both males and females (78% identity) is observed (Table 2.1). Female-specific OD2 domains are 93% identical. However, conservation of the male-specific OD2 domain is, in stark contrast, very low (1 out of 30 aligned amino acids is identical between all three species) and these results indicate that male-specific DSX isoforms possess higher rates of evolutionary change. These result are consistent with literature that have demonstrated a high variance in male sexual traits. Traditional notions of sexual selection have invoked female choice on an arena of male traits with high additive genetic variance (for example, Lande 1981, Kirkpatrick 1982). Examples of highly variable male traits include insect genitalia (Eberhard 1985), sperm morphology (Karr and Pitnick 1996; Joly *et al.* 1997) and accessory gland proteins (Clark *et al.* 1995; Begun *et al.* 2000). Since DSX sex-specific domains (i.e. OD2) are the principle regulators of a variety of downstream targets involved in somatic sexual differentiation, the higher evolutionary rates found in the

TABLE 2.1

Conservation of DSX oligomerization protein domains across Diptera

Species Comparison	Common Domains		Sex-specific Domains	
	OD1	OD2	OD2 female	OD2 male
Drosophila - Megaselia	88% *n=66*	78% *n=47*	93% *n=15*	31% *n=27*
Drosophila - Bactrocera	98% *n=66*	96% *n=48*	100% *n=15*	78% *n=30*
Megaselia - Bactrocera	84% *n=66*	79% *n=47*	93% *n=15*	31% *n=35*
Drosophila - Megaselia - Bactrocera	88% *n=66*	78% *n=47*	93% *n=15*	13% *n=30*

Sequences retrieved from GenBank (see Materials and Methods for accession numbers) and aligned using ClustalX. Oligomerization domains, OD1 and OD2, are as defined by Erdman and Burtis (1993). Percentages indicate proportion of domain region that is identical or has conserved amino acids. Indels are ignored. Sample size, *n*, is in nucleotides.

male-specific domain support a generally higher level of variation in male-specific genes/traits.

Among species of the *D. melanogaster* subgroup, nucleotide divergence between common exon 2 and the male-specific exon 5 was compared (Table 2.2) using synonymous (Ks) and nonsynonymous (Ka) substitutions per site (Nei and Gojobori 1986) and an opposite trend was found. In nearly all pairwise comparisons, both K_a and K_s were greater in the common exon than the male-specific exon. A codon substitution model and maximum likelihood approach (Yang 1997) was used to estimate the parameter, d_N/d_S (Table 2.3). This parameter, which is an estimate of a protein's overall selective constraint, was obtained from the model of best fit (i.e. a molecular clock was not assumed) and revealed that the male-specific exon had greater selective constraints ($d_N/d_S = 0.04$) when compared to common exon 2 ($d_N/d_S = 0.10$). Thus, greater selective constraints are observed in the male-specific portion of DSX relative to the common portion, contrary to the observations across a greater phylogenetic species range (i.e. *Megaselia, Bactrocera, Drosophila*).

Although both common and male-specific exons contain differences in evolutionary rates (Table 2.2) and selective constraints (Table 2.3), the overall level of change at the *dsx* locus was among the lowest found between species of the *D. melanogaster* complex (Appendix A). Despite its low divergence, there is ample evidence of heterogeneity in *dsx* evolutionary rates *between species* of the *D. melanogaster* subgroup as significant differences in likelihoods are obtained between evolutionary models that employ a molecular clock and models that do not (Table 2.3).

TABLE 2.2

Pairwise nucleotide divergence of *dsx* common exon 2 and male-specific exon 5 between species of the *D. melanogaster* subgroup

Species	*mel*	*sim*	*mau*	*sec*	*ere*	*tie*	*yak*
mel	***C2***	0.0044	0.0059	0.0059	0.0192	0.0192	0.0162
	M5	0.0035	0.0035	0.0035	0.0071	0.0143	0.0179
sim	0.1024	***C2***	0.0015	0.0044	0.0192	0.0192	0.0162
	0.0916	***M5***	0	0	0.0036	0.0107	0.0143
mau	0.1431	0.0449	***C2***	0.0059	0.0207	0.0207	0.0177
	0.0809	0.0096	***M5***	0	0.0036	0.0107	0.0143
sec	0.1537	0.0449	0.0635	***C2***	0.0207	0.0207	0.0177
	0.1360	0.0597	0.0494	***M5***	0.0036	0.0107	0.0143
ere	0.1592	0.1705	0.2269	0.2152	***C2***	0.0207	0.0177
	0.1592	0.1243	0.1131	0.1716	***M5***	0.0071	0.0107
tie	0.2338	0.2168	0.2771	0.2646	0.1390	***C2***	0.0147
	0.2090	0.1837	0.1716	0.2086	0.2344	***M5***	0.0107
yak	0.2510	0.2518	0.3149	0.3018	0.1652	0.0736	***C2***
	0.2480	0.2214	0.2086	0.1708	0.1592	0.0810	***M5***

Proportion of nonsynonymous substitutions per nonsynonymous site (K_a) are found in upper triangular. Lower triangular contains proportion of synonymous substitutions per synonymous site (K_s). For each pairwise species comparison, the top value corresponds to common exon 2 while the bottom value corresponds to male-specific exon 5. Sequences from both exons comprise of coding region only. *D. melanogaster* is represented by its previously published sequence (Burtis and Baker 1989). *mel, D. melanogaster*; *sim, D. simulans*; *mau, D. mauritiana*; *sec, D. sechellia*; *ere, D. erecta*; *tie, D. tiessierri*; *yak, D. yakuba*

TABLE 2.3

Likelihood-ratio tests and evolutionary rates of *dsx* coding regions amongst species of the *D. melanogaster* subgroup

Species topology utitlized	Region of *dsx*	Number of codons	Difference in log-likelihood[c]		Estimate of d_N/d_S under no clock model
			Clock *vs.* no clock	Lineage *vs.* constant (d_N/d_S)	
[a]Siblings resolved, *erecta - melanogaster*	common exon 2	305	22.0**	18.1**	0.09
	male exon 5	129	12.6*	5.6	0.04
[b]Sibling trichotomy, *erecta - yakuba*	common exon 2	305	17.7**	14.3*	0.10
	male exon 5	129	12.6*	4.2	0.04

[a](((mel,((sim,mau),sec)),ere),(yak,tie)); [b]((mel,(sim,mau,sec)),(ere,(yak,tie)));
D. melanogaster is represented by its published sequence. [c] Twice the difference in log likelihood between models tested. *95% significance, **99% significance

However, this rate heterogeneity between lineages is found in both common and male-specific exons and was confirmed using different species topologies. Figure 2.5 compares the divergence of both common and male-specific exons. No obvious differences were observed between trees again suggesting that this rate heterogeneity among lineages equally affects both regions of the *dsx* locus. When comparing evolutionary models that possess constant d_N/d_S *vs.* lineage-specific d_N/d_S, it also appears that certain lineages of the *D. melanogaster* subgroup have significantly different levels of selective constraint across the common exon (Table 2.3), indicating the presence of variable selective pressures between these species.

Intraspecific nucleotide variation was also compared in *D. melanogaster* between these two exons (Table 2.4). Although the number of segregating sites are low, thereby reducing the power of this analysis, an opposite trend from *dsx* divergence in the *D. melanogaster* species subgroup emerges. Overall nucleotide diversity is significantly higher in the male-specific exon (π_{male}=0.00554 *vs.* π_{common}=0.00092, P<0.01). Since these two regions of the *dsx* locus are situated relatively close to each other, the effects of differing levels of recombination rates (Begun and Aquadro 1992; Andolfatto 2001) are neglible. While the estimated diversity from the male exon appears to be similar to the average nucleotide diversity calculated from 24 *D. melanogaster* loci (Moriyama and Powell 1996; see Table 3.1), variation in the common exon is low. The low levels of diversity found in the *dsx* common exon, which consists of a larger sampled area than the *dsx* male exon (Table 2.4), can best be explained by its chromosomal location. The *dsx* locus (Drosophila cytological map, 84E1, Lindsley and Zimm 1992) is situated near the

Figure 2.5 Neighbor-joining trees of common *vs.* male-specific *dsx* exons. Bootstrap values (out of 1000 replicates) are indicated. Strains are indicated for *D. melanogaster* in parentheses. Both trees are drawn to the same scale.

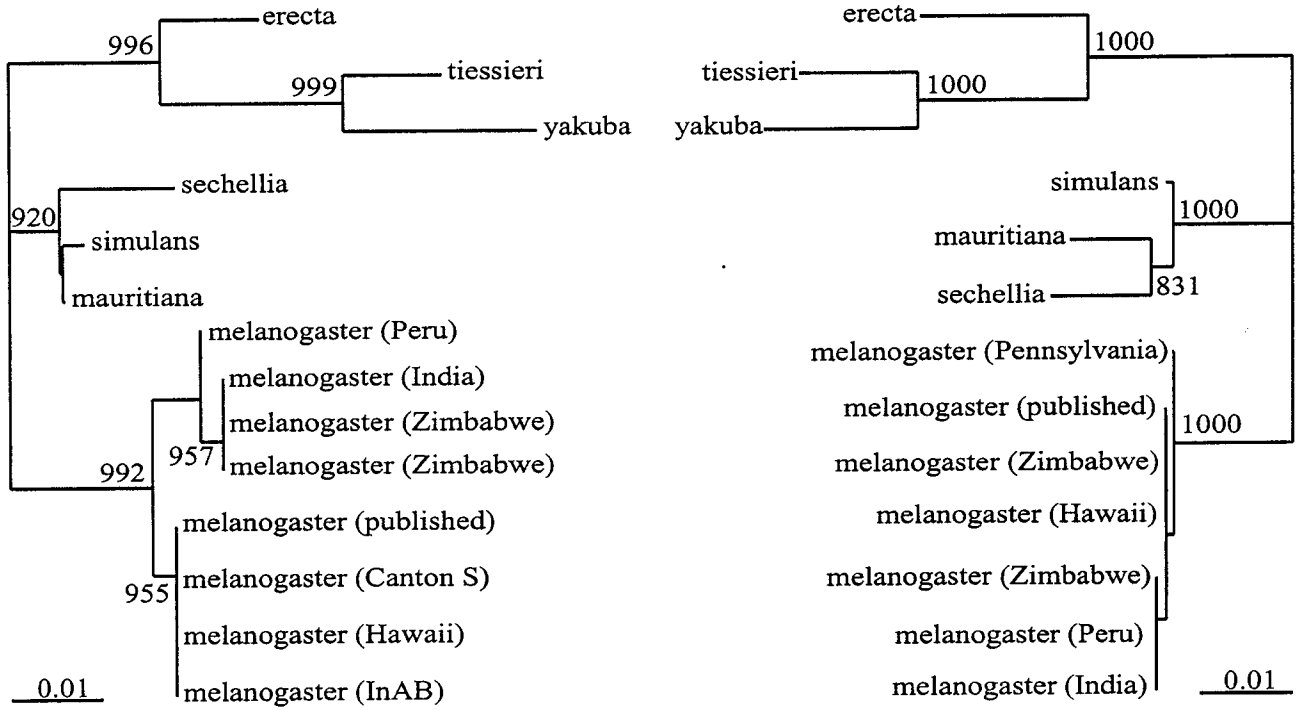

male-specific exon 5
erecta
996
tiessieri
999
yakuba
sechellia
920
simulans
mauritiana
melanogaster (Peru)
melanogaster (India)
melanogaster (Zimbabwe)
957
melanogaster (Zimbabwe)
992
melanogaster (published)
melanogaster (Canton S)
955
melanogaster (Hawaii)
melanogaster (InAB)
0.01
common exon 2
erecta
1000
tiessieri
1000
yakuba
simulans
1000
mauritiana
831
sechellia
melanogaster (Pennsylvania)
melanogaster (published)
1000
melanogaster (Zimbabwe)
melanogaster (Hawaii)
melanogaster (Zimbabwe)
melanogaster (Peru)
melanogaster (India)
0.01

TABLE 2.4

Nucleotide diversity at the *dsx* locus in *D. melanogaster*

Region of *dsx*	Number of sequences *N*	Number of haplotypes *H*	Number of nucleotides *l*	Segregating sites *S*	Replacement			Synonymous		
					r	π	θ	*s*	π	θ
Common exon 2	7	3	939	2	1	0.82	0.59	1	1.21	1.73
Male-specific exon 5	8	3	393	4	1	2.02	1.36	3	15.1	11.1
Combined exons 2 + 5	6	6	1332	5	2	1.16	0.89	3	4.71	3.86

Two different estimates of the neutral population parameter, $4N_e\mu$, were calculated, π (Nei 1987) and θ (Watterson 1975). Nucleotide diversities are multiplied by 10^{-3}. Corrections for multiple hits were not performed. Only coding regions were analyzed.

third chromosome centromere where recombination rates are low. In fact, reported recombination rates of genes located in the immediate vicinity of *dsx* appear to be an order of magnitude lower than the majority of genes found distal to both centromeric and telomeric regions of the *D. melanogaster* genome (see Appendix A1 in Kliman and Hey 1993). Therefore, the low diversity found at the *dsx* common exon (Table 2.4) may be a reflection of the observed correlation between recombination rate and nucleotide diversity (Begun and Aquadro 1992). The higher observed diversity found in the male exon is more difficult to explain. However, its higher variation is based on a small 393 base pair region which may be more sensitive to stochastic error.

Conclusions: The evolution of sex determination loci presents us with a study in contrasts. While some loci are rapidly evolving between closely related taxa, others maintain protein domains which are quite conserved even across large phylogenetic distances. *doublesex* falls in the latter category. Among species of the *D. melanogaster* subgroup, the *dsx* locus is relatively conserved. A closer evaluation using different levels of phylogenetic comparison reveals that selective constraints may have varied across lineages. Relative to the *dsx* common exon, lower levels of divergence are found in male-specific regions between members of the *D. melanogaster* subgroup. Across non-Drosophilid Dipterans, much higher divergence levels are found in male-specific protein domains. Such transformations in selective constraints indicate the variability of selective pressures across lineages in sex determining loci.

REFERENCES

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

Andolfatto, P. (2001) Contrasting patterns of X-linked and autosomal nucleotide variation in *Drosophila melanogaster* and *Drosophila simulans*. Mol. Biol. Evol. **18:** 279-290

Baker, B.S., and K.A. Ridge (1980) Sex and the single cell. I. On the action of major loci affecting sex determination in *Drosophila melanogaster*. Genetics **94:** 383-423

Begun, D.J., and C.F. Aquadro (1992) Levels of naturally occurring DNA polymorphism correlate with recombination rates in *D. melanogaster*. Nature **356:** 519-520

Begun, D.J., Whitley, P., Todd, B.L., Waldrip-Dail, H.M., and A.G. Clark (2000) Molecular population genetics of male accessory gland proteins in Drosophila. Genetics **156:** 1879-1888

Bownes, M. (1994) The regulation of the *yolk protein* genes, a family of sex differentiation genes in *Drosophila melanogaster*. BioEssays **16:** 745-752

Burtis, K.C., and B.S. Baker (1989) *Drosophila doublesex* gene controls somatic sexual differentiation by producing alternatively spliced mRNAs encoding related sex-specific polypeptides. Cell **55:** 997-1010

Burtis, K.C., and M.F. Wolfner (1992) The view from the bottom: Sex-specific traits and their control in *Drosophila*. Semin. Dev. Biol. **3:** 331-340

Chen, H.-J. (1995) Identification and localization of the dimerization domains of the DSX proteins using the yeast two-hybrid system. In *Functional domains of the doublesex proteins in Drosophila melanogaster*, Ph.D. Dissertation, University of California, Davis

Cho, S., and P.C. Wensink (1997) DNA binding by the male and female *doublesex* proteins of *Drosophila melanogaster*. J. Biol. Chem. **272:** 3185-3189

Clark, A.G., Aguadé, M., Prout, T., Harshman, L.G., and C. H. Langley (1995)

Variation in sperm displacement and its association with accessory gland protein loci in *Drosophila melanogaster*. Genetics **139:** 189-201

Cline, T.W., and B.J. Meyer (1996) Vive la différence: Males *vs.* females in flies *vs.* worms. Annu. Rev. Genet. **30:** 637-702

Coschigano, K.T., and P.C. Wensink (1993) Sex-specific transcriptional regulation by the male and female *doublesex* proteins of Drosophila. Genes Dev. **7:** 42-54

deBono, M., and J. Hodgkin (1996) Evolution of sex determination in *Caenorhabditis*: Unusually high divergence of *tra-1* and its functional consequences. Genetics **144:** 587-595

Dübendorfer, A., Hilfiker-Kleiner, D., and R. Nöthiger (1992) Sex determination mechanisms in dipteran insects: The case of *Musca domestica*. Semin. Dev. Biol. **3:** 349-356

Eberhard, W.G. (1985) *Sexual selection and animal genitalia*. Harvard University Press, Cambridge, MA

Erdman, S.E., Chen, H.J., and K.C. Burtis (1996) Functional and genetic characterization of the oligomerization and DNA binding properties of the *Drosophila doublesex* proteins. Genetics **144:** 1639-1652

Erdman, S.E., and K.C. Burtis (1993) The Drosophila *doublesex* proteins share a novel zinc finger related DNA binding domain. EMBO J. **12:** 527-535

Felsenstein, J., (1993) PHYLIP, phylogeny inference package. Version 3.5. Distributed by the author, Department of Genetics, University of Washington, Seattle, U.S.A.

Fredga, K. (1994) Bizarre mammalian sex-determining mechanisms. In *The differences between the sexes*. (Eds: R.V. Short and E. Balaban) Pg. 419-431. Cambridge University Press, Cambridge

Gloor, G., and W. Engels (1992) Single-fly preps for PCR. D.I.S. **71:** 148-149

Handa, N., Nureki, O., Kurimoto, K., Kim, I., Sakamoto, H., Shimura, Y., Muto, Y., and S. Yokoyama (1999) Structural basis for recognition of the *tra* mRNA precursor by the Sex-lethal protein. Nature **398:** 579-585

Hertel, K.J., Lynch, K.W., Hsiao, E.C., Liu, E.H., and T. Maniatis (1996) Structural and functional conservation of the *Drosophila doublesex* splicing enhancer repeat elements. RNA **2:** 969-981

Hodgkin, J. (1992) Genetic sex determination mechanisms and evolution. BioEssays **14:** 253-256

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

Joly, D., Bazin, C., Zeng, L.W., Singh, R.S. (1997) Genetic basis of sperm and testis length differences and epistatic effect on hybrid inviability and sperm motility between *Drosophila simulans* and *D. sechellia.* Heredity **78:** 354-362

Karr, T.L., and S. Pitnick (1996) The ins and outs of fertilization. Nature **379:** 405-406

Kirkpatrick, M. (1982) Sexual selection and the evolution of female choice. Evolution **36:** 1-12

Kliman, R.M., and J. Hey (1993) Reduced natural selection associated with low recombination in *Drosophila melanogaster.* Mol. Biol. Evol. **10:** 1239-1258

Kuhn, S., Sievert, V., and W. Traut (2000) The sex-determining gene *doublesex* in the fly *Megaselia scalaris*: Conserved structure and sex-specific splicing. Genome **43:** 1011-1020

Kumar, S., Tamura, K., Jakobsen, I., and M. Nei (2000) MEGA: Molecular evolutionary genetics analysis (version 2.0)

Kuwabara, P.E. and Hodgkin, J. (1996) Interspecies comparisons reveal evolution of control regions in the nematode sex-determining gene *tra-2.* Genetics **144:** 597-607

Lande, R. (1981) Models of speciation by sexual selection on polygenic traits. Proc. Natl. Acad. Sci. U.S.A.. **78:** 3721-3725

Lindsley, D.L., and G.G. Zimm (1992) The genome of *Drosophila melanogaster.* Academic Press

Marín, I., and B.S. Baker (1998) The evolutionary dynamics of sex determination. Science **281:** 1990-1994

McAllister, B.F., and G.A. McVean (2000) Neutral evolution of the sex-determining gene *transformer* in Drosophila. Genetics **154:** 1711-1720

Moriyama, E.N., and J.R. Powell (1996) Intraspecific nuclear DNA variation in Drosophila. Mol. Biol. Evol. **13:** 261-277

Nagoshi, R.N., McKeown, M., Burtis K.C., Beloté, J.M., and B.S. Baker (1988) The control of alternative splicing at genes regulating sexual differentiation in *D. melanogaster*. Cell **53:** 229-236

Nei, M. (1987) *Molecular Evolutionary Genetics*. Columbia University Press, New York

Nei, M., and T. Gojobori (1986) Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Mol. Biol. Evol. **3:** 418-426

O'Neil, M.T., and J.M. Beloté (1992) Interspecific comparison of the *transformer* gene of *Drosophila* reveals an unusually high degree of evolutionary divergence. Genetics **131:** 113-128

Page, R.D. (2000) TREEVIEW: An application to display phylogenetic trees on personal computers. Comput. Appl. Biosci. **12:** 357-358

Raymond, C.S., Shamu, C.E., Shen, M.M., Siefert, K.J., Hirsch, B., Hodgkin, J., and D. Zarkower (1998) Evidence for evolutionary conservation of sex-determining genes. Nature **391:** 691-695

Rozas, J., and R. Rozas (1999) DnaSP version 3: An integrated program for molecular population genetics and molecular evolution analysis. Bioinformatics **15:** 174-175

Saitou, N., and M. Nei (1987) The neighbor-joining method: A new method for reconstructing phylogenetic trees. Mol. Biol. Evol. **4:** 406-425

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

Shearman, D.C., and M. Frommer (1998) The *Bactrocera tryoni* homogue of the *Drosophila melanogaster* sex-determination gene *doublesex*. Insect. Mol. Biol. **7:** 355-366

Smith, C.A., McClive, P.J., Western, P.S., Reed, K.J., and A.H. Sinclair (1999) Conservation of a sex-determining gene. Nature **402:** 601-602

Thompson, J.D., Gibson, T.J., Plewniak, F., Jeanmougin, F., and D.G. Higgins (1997)

The CLUSTAL_X windows interface: Flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic Acids Res. **25:** 4876-4882

Truman, J.W. (1992) Developmental neuroethology of insect metamorphosis. J. Neurobiol. **23:** 1404-1022

Tucker, P.K., and B.L. Lundrigan (1993) Rapid evolution of the sex determining locus in Old World mice and rats. Nature **364:** 715-717

Villella, A., and J.C. Hall (1996) Courtship anomalies caused by *doublesex* mutations in *Drosophila melanogaster*. Genetics **143:** 331-344

Walthour, C.S., and S.W. Schaeffer (1994) Molecular population genetics of sex determination genes: The *transformer* gene of *Drosophila melanogaster*. Genetics **136:** 1367-1372.

Watterson, G.A. (1975) On the number of segregating sites in genetical models without recombination. Theor. Popul. Biol. **7:** 256-276

Whitfield, L.S., Lovell-Badge, R., and P.N. Goodfellow (1993) Rapid sequence evolution of the mammalian sex-determining gene SRY. Nature **364:** 713-715

Wilkins, A.S. (1992) Moving up the hierarchy: a hypothesis on the evolution of a genetic sex determination pathway. BioEssays **17:** 71-77

Yang, Z. (1997) PAML: A program package for phylogenetic analysis by maximum likelihood. CABIOS **13:** 555-556

CHAPTER THREE

Molecular variation at the sex determining locus, *transformer*: Rapid evolution and varying selective constraints in the *Drosophila melanogaster* subgroup

ABSTRACT

While developmentally regulated genes are generally conserved, *transformer* (*tra*), a key locus involved in the regulation of sexual differentiation, is highly diverged between species of of Drosophila. Coupled to observations of low intraspecific variation in a population of *Drosophila melanogaster*, it has been suggested that positive selection may be the driving force behind *tra*'s rapid evolution. With an aim to understand its divergence between sibling species, *tra* sequence variation was investigated among the four sibling species of the *D. melanogaster* complex. Worldwide samples of *D. melanogaster* revealed only four segregating sites in a 926 base pair region, confirming that a reduction of genetic variation is found throughout the species range. *D. simulans* exhibits much higher levels of nucleotide diversity, but similar to other loci found in *D. simulans*, while *D. mauritiana* has elevated amino acid diversity relative to other conspecific loci. Except for one replacement polymorphism, *D. sechellia* is invariant. In this species complex, *tra* evolution is rapid yet clocklike and exhibits large differences in protein size. In terms of average amino acid divergence, d_N, *tra* ranks second highest out of eighteen loci having sequence available from all four sibling species (d_N=0.042) while its silent divergence, d_S, remains at intermediate levels (d_S=0.13). Rapid *tra* divergence is also seen in large differences in protein size as well as the number and density of arginine-serine (RS) domains, important for protein-protein interaction. *D. melanogaster* possesses a 13 amino acid tandem insertion containing an RS domain while a larger tandem insertion of 72 amino acids, which represents an almost 30 % amino acid

increase, was found in *D. sechellia*. This unusually large insertion introduces at least two additional RS domains and three basic amino acid domains. In the *D. melanogaster* complex, significant heterogeneity in d_N/d_S ratios was shown between sites using maximum likelihood. Statistical significance, however, is lost when the sister species, *D. erecta*, is added to the comparison. Finally, I show that *tra*'s rapid divergence is the result of lower selective constraints in the RS-rich second exon and a significantly higher substitution rate around the site *D. sechellia*'s insertion. The proximity of rapidly diverged regions to sites of insertion suggest that higher local rates of mutation may provide a causal mechanism for *tra*'s rapid divergence in this subgroup. A comparison of *tra* orthologs across Drosophila suggest that *tra* maintains RS domains for proper sex determining function while much of the protein evolves relatively unconstrained.

INTRODUCTION

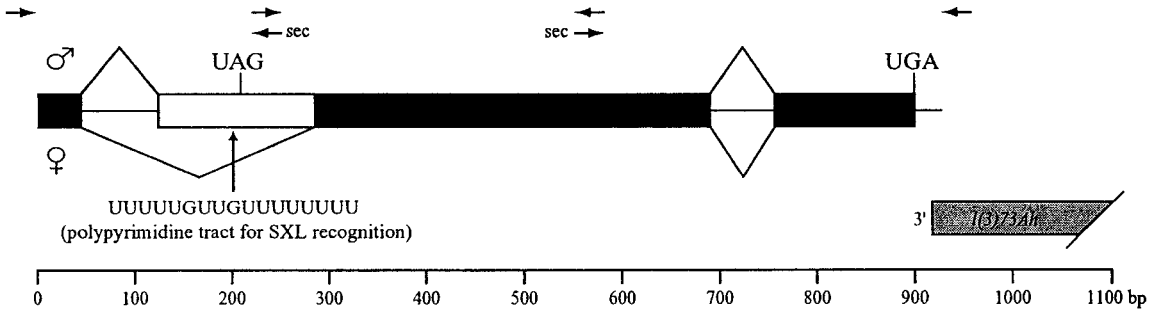
Developmental processes have long been thought to be more conserved, even between distantly related taxa. One consequence of this perception has been the use of comparative embryology in studies of evolutionary biology. As a result many classic phylogenetic inferences have been based on shared morphology from early developmental stages (Gould 1977). However, it was only with the sequencing of orthologous sequences from diverse taxa that biologists began to fully appreciate the extent of homology in development. Examples such as genes involved in embryonic pattern formation and cellular signalling have demonstrated the conservation of structure and function over a broad range of developmental loci (Nusslein-Volhard 1994; Patel 1994; Chan and Jan 1999). Population genetic analyses of developmental loci, although few in number, also reveal low levels of variation between species. For example, *dpp* and *runt*, have been shown to possess a relatively small number of fixed differences between *D. simulans* and *D. melanogaster* when compared to other genes (Richter *et al.* 1997; Labate *et al.* 1999). These loci also have very few amino acid polymorphisms and relatively normal levels of synonymous polymorphism.

It is therefore surprising when an essential gene, expressed early in development, does not evolve in a typically conserved manner. The sex-determination gene, *transformer* (*tra*), has previously been shown to have the lowest sequence identity among known orthologous proteins between *D. melanogaster* and *D. virilis* (O'Neil and Beloté 1992). The primary function of *tra* is to act as a regulatory switch in the determination of

sexual identity in each cell. Depending on the X:autosome ratio, *tra* is expressed as a functional protein (in XX females) or a truncated, non-functional protein (in XY males; see Figure 3.1) via female-specific alternative splicing of *tra pre*mRNA controlled by the Sex-lethal protein (SXL). This switch is regulated by SXL binding to a polypyrimidine site in *tra*'s first intron (Sosnowski *et al.* 1989; Handa *et al.* 1999; see Figure 3.1). In females, TRA associates with another protein, TRA-2, via its arginine-serine (RS) domains, and this complex binds to *doublesex* (*dsx*) *pre*mRNA repeat elements (Inoue *et al.* 1992). *dsx pre*mRNA is then alternatively spliced in females to produce a female-specific product while in males (without functional TRA), default splice sites are used to form a male-specific product. Each sex-specific dsx protein regulates a variety of downstream genes involved in sexual differentiation including body size, genitalia development, mating preference and pheromone production (Burtis and Baker 1989; Ferveur *et al.* 1997). *tra* is also involved in the regulation of *fruitless* which affects almost all components of male courtship behaviour (Ryner *et al.* 1996).

Two different aspects of *transformer* sequence variation - within *vs.* between species - have been studied in the genus Drosophila and have suggested two contrary mechanisms of rapid *tra* divergence. In the first study, Walthour and Schaeffer (1994) observed drastically reduced levels of genetic variation in a natural population of *D. melanogaster*. After comparing levels of within species variation to that between species (Hudson *et al.* 1987), they rejected the hypothesis of neutral evolution. Given that *tra* is located in a region of low to moderate recombination (Kliman and Hey 1993; Walthour and Schaeffer 1994), the authors suggested that directional selection via recent selective

Figure 3.1 Gene organization of the *transformer* locus in *D. melanogaster*. Solid boxes represent exons and horizontal lines indicate the first and second introns and 3' flanking region. The untranslated region in male-specific exon 2 (with premature stop codon) is shown as an open box. Male-specific mRNA transcripts are shown above while female-specific transcripts are below. Primers used in amplification and sequencing are indicated by horizontal arrows. Additional set of primers used in *D. sechellia* are also indicated (denoted as sec) and flank the insert. Known constraints are shown including the polypyrimidine SXL recognition domain in females and neighbouring 3' gene (shaded) which is coded on the opposite strand. Nucleotide numbering (i.e. position 1) commences at the first base pair sequenced which corresponds to the first nucleotide of the start codon in exon 1.


sec
sec
♂
UAG
UGA
♀
UUUUUGUUGUUUUUUUU
(polypyrimidine tract for SXL recognition)
3'
0
100
200
300
400
500
600
700
800
900
1000
1100 bp

sweeps acting in or around the *tra* locus may have caused the pattern of low polymorphism in *D. melanogaster* and high divergence between species. In the second study, McAllister and McVean (2000) studied the species pair, *D. americana* and *D. virilis* using a series of population genetic and phylogenetic approaches. They showed that high rates of neutral evolution may be sufficient to explain the high divergence between these species. They also tested the molecular clock on three previously published sequences of *D. melanogaster*, *D. simulans* and *D. erecta* and found that a neutral model was consistent with the three Sophophoran species.

In this chapter, both aspects of variation will be evaluated using sequence variation in all four sibling species of the melanogaster complex - the cosmopolitan species, *D. melanogaster* and *D. simulans*, and the island endemics, *D. mauritiana* and *D. sechellia*. The latter three species have diverged from each other within the last half million years (Kliman *et al.* 2000). The primary goal is to understand both the pattern and mechanism of *tra*'s rapid evolution in light of its importance in the sex determination pathway and to gain insight into the functional nature of this essential protein. I report that this developmentally regulated gene has undergone rapid divergence under varying selective constraints in this clade producing drastic changes in protein size and conclude that *tra* evolution in Drosophila has been shaped by an overall lack of selective constraints and lineage-specific selective pressures on RS domains.

MATERIALS AND METHODS

Fly strains: Nine strains each of *D. melanogaster* and *D. simulans* and representing various geographical populations, were chosen in order to sample global *transformer* variation. In this study, sequences are denoted according to a three letter species prefix and a five character suffix indicating its original stock number or geographical origin (only ambiguous designations are listed below). *D. melanogaster* strains: A single line from India (14021-0231.6), Hawaii (14021-0231.0) and Malaysia (14021-0231.4) originated from the Drosophila Species Stock Centre at Bowling Green (D.S.S.C.). Two lines from State College, PA (CPA-46 and CPA-129, denoted as mel.penn1and mel.penn2) were kindly provided by Brian Lazzaro (Penn State). One line from Rome, Italy (I-13) was collected by Alberto Civetta (University of Winnipeg). Three homozygous lines from Zimbabwe (Z(H)-12, Z(H)-16 and Z(H)-34, denoted as mel.zimb1, mel.zimb2 and mel.zimb3) were provided from the Andrew Clark lab (originally from David Begun, UC Davis). *D. simulans* strains: Lines from Colombia (0251.2), Florida (0251.166), California (0251.163) and a strain of unknown origin, Solway-Hochman (1088, denoted as sim.unkno), were obtained from the D.S.S.C. John Roote (Cambridge University, UK) supplied flies from Ethiopia (S-23) and Madagascar (S-24). Lines from Kenya (S-144), Italy (S-132) and a strain of unknown origin (S-148) were obtained from the Drosophila Stock Centre at Umea, Sweden. *D. mauritiana* strains: Two lines (0241.1, 0241.3) originated from the D.S.S.C and two lines (S-080, S-081) were obtained from the Umea Stock Centre. *D. sechellia* strains: Three lines

(0241.0, 0241.3, 0241.4) originated from the D.S.S.C and one line (3151) was kindly provided by Alberto Civetta.

DNA extraction, PCR amplification and sequencing: As in Chapter two, a single fly protocol was employed to extract genomic DNA (Gloor and Engels 1992). Briefly, one fly is macerated in 10 mM Tris-Cl (pH 8.2), 1 mM EDTA, 25 mM NaCl and 200 mg/ml proteinase-K and left to stand at room temperature for 30 minutes. The preparation is then heated at 95° for three minutes. A fragment approximately 1 kb long (depending on the species) was amplified in a Perkin-Elmer 480 thermal cycler using the extracted genomic DNA in 1X PCR buffer (Fermentas), 0.2 mM dNTPs, 1.0 pM primers, 2.5 mM $MgCl_2$ and Taq polymerase (Fermentas). For all four species, the same primers flanking the *tra* locus were used for amplification:
5'GTGCATCATTTAATTTCCAGCA3' and
5'TTTTAATGTACAAAACACACGAATG3'. The amplified product contains the full coding sequence along with *tra*'s two introns, untranslated male exon and 53-54 nucleotides (in *D. simulans*) of the 3' flanking region. PCR products were excised from a 2% agarose gel, DNA extracted and then purified using a QIAGEN QIAquick Gel Extraction Kit. Sequencing was performed using an ABI PRISM® 377 DNA sequencer using the same primers as above. Two internal primers were also used, 5'GAGGTTCGAGAACAGGATCGG3' and 5'CGTTCACTGCTGCGACTTCGG3', so that polymorphisms could be confirmed on both strands. Discrepancies between strands were not observed in any cases. Two singletons in *D. melanogaster* (a replacement and silent polymorphism) were reconfirmed through independent DNA extraction,

amplification and sequencing. The large insertion found in *D. sechellia* made necessary the construction and use of a third set of internal primers (located in regions common to all four species) 5'TAATGCGCAGTTGAGAGTCC3' and 5'GAAGTCGCAGCAGTGAACG3' for amplification and sequencing. The unusually large sequence found in *D. sechellia* was verified independently by two other researchers (Lara Skwarek, University of Toronto; Santosh Jagadeeshan, McMaster University) - each extracted, amplified and sequenced different lines of this species. Approximate primer locations are shown in Figure 3.1.

Sequence analysis: In addition to the lines described above, 11 *D. melanogaster* sequences (Walthour and Schaeffer 1994, accession numbers L19464-L19470 and L19618-L19620, O'Neil and Beloté 1992, M17478), and a *D. erecta* sequence (Walthour and Schaeffer 1994, X66527) were obtained from the NCBI database to include in *D. melanogaster* subgroup analyses. Sequences were labelled according to previously published designations. The published sequence of *D. simulans* (O'Neil and Beloté 1992, X66930) was excluded from all analyses because six singletons (three of which cause nonsynonymous changes) were not observed with any of the nine *D. simulans* sequences. Although nucleotide diversity is quite high in *D. simulans*, such a high frequency of unique polymorphic sites stemming from a single haplotype suggests that this sequence may be in error. Nucleotide and protein sequences were aligned using CLUSTALW (Thompson *et al.* 1994). Protein distances, Fitch-Margoliash protein trees, protein parsimony and neighbor-joining algorithms were implemented using the PHYLIP package v3.5c (Felsenstein 1993) and displayed using TREEVIEW ver. 1.6.1

(taxonomy.zoology.gla.ac.uk/rod/rod.html). For intraspecific data, two nucleotide diversity statistics, θ (Watterson 1975) and π (Nei 1987), were calculated using DnaSP ver. 3 (Rozas and Rozas 1997). SITES ver. 1.1 (Hey and Wakeley 1997) was used to confirm polymorphism statistics.

To test various evolutionary models such as constant *vs.* variable rates of substitution, clock *vs.* no clock and the variability of d_N/d_S (a parameter measuring the expected ratio of the proportion of nonsynonymous substitutions per nonsynonymous site to synonymous substitutions per synonymous site according to a codon model of substitution) among lineages as well as sites, I used a maximum likelihood approach implemented by the program CODEML in the PAML package v3.0b (Yang 1997). Models were compared using likelihood ratio tests. Pairwise estimates of d_N and d_S, nonsynonymous and synonymous substitutions per site, were calculated using the program YN00 (Yang and Neilson 2000). A second set of divergence estimates, d_N and d_S, was derived using the approximate method of Nei and Gojobori (1986). To compare *transformer* evolution to that of other proteins in the *melanogaster* species complex, levels of divergence (and polymorphism for *D. mauritiana* and *D. sechellia*, in Table 3.1) were compared in the protein coding regions of loci that were found in the NCBI database for all four species of the *D. melanogaster* complex. These loci are listed in Appendix A. Only the portion aligned in all four species was compared when partial sequences were available.

Positional heterogeneity of amino acid variation was assessed using the broken-stick model of Goss and Lewontin (1996). Clustering of variable sites is measured by

three statistics: L_{max}, the maximum fractional interval length, Var(L), the variance in fractional interval length and Q, a modified variance statistic. A computer program, HET2, was kindly provided by Peter Goss (Harvard University). In order to specify which region(s) possessed significant differences in rates of nucleotide substitution, the permutation method of Hartmann and Golding (1998) was employed (a computer program was kindly provided by Brian Golding, McMaster University). Indels were removed from aligned coding regions and nucleotides were divided into subsets that comprised of sites that are more prone to amino acid change (first and second codon positions, nondegenerate sites) as well as to sites that serve as controls (third codon position, all nucleotide sites). Nondegenerate sites were chosen from a *D. melanogaster* sequence. Topologies with branch lengths were obtained using DNAML (PHYLIP package v3.5c, Felsenstein 1993) for each dataset. Sliding windows of various length then survey the sequence for regional rate heterogeneity using maximum likelihood.

I compared the amino acid profile, particularly the distribution of arginine-serine (RS) dipeptides and basic amino acids (arginine, histidine and lysine), across all known TRA sequences in Drosophila. A computer program was kindly provided by Richard Morton (McMaster University) and used a sliding window approach to calculate the proportion of RS dipeptides / basic amino acids for a given protein sequence. Sequences from members of the subgenus Drosophila were obtained from GenBank including *D. virilis* (O'Neil and Beloté 1992, X66528) and 31 sequences from *D. americana* (McAllister and McVean 2000, AF208127-AF208157). Consensus sequences for each species were produced. The proportion of RS dipeptides and basic amino acids was

calculated over a series of overlapping windows (11 amino acids) across each protein.

RESULTS AND DISCUSSION

Nucleotide variation in *D. melanogaster* and its sibling species: In a previously published study on *tra* variation, Walthour and Schaeffer (1994) found two segregating sites, both silent, in ten sequences (n=1063 base pairs) from a North American population (Pennsylvania). One of these polymorphisms was a singleton while the other was found at an intermediate frequency. In order to test whether this reduced variation was caused by population or sampling artefacts such as local bottlenecks, sweeps, or the sequencing of alleles related by descent, nine sequences from global populations of *D. melanogaster* were sampled. Among the 926 nucleotide sites sequenced in worldwide populations (excluding sequences from the Pennsylvanian population), two polymorphic silent sites and one replacement polymorphism which was only present in an Asian population were found. Thus genetic variation at the *tra* locus is reduced not only in the Pennsylvanian population, but throughout the species range.

In *D. melanogaster*, nucleotide diversities from *tra* are significantly smaller than the average value calculated for a sample of 24 loci by Moriyama and Powell (1996) (Table 3.1, $P<0.01$). These differences are observed for both synonymous and replacement sites, as well as non-coding regions of the *tra* locus. Nucleotide variation is absent in the first and second introns (not including the untranslated region of the male-specific exon) and the 3' flanking region (Table 3.1). The test statistics, Tajima's D (1989) and Fu and Li's (1993) D*, were calculated in order to measure any deviation of haplotype distribution from neutral expectations. If hitchhiking or background selection

TABLE 3.1

Comparison of *tra* nucleotide diversity amongst species of the *Drosophila melanogaster* complex

	Sequence Summary				Coding Regions: Segreg. Sites		Coding Regions: Total Sites		Coding Regions: Nonsynon. Sites		Coding Regions: Synon. Sites		Noncoding Regions: Intron 1 & untrans. male		Noncoding Regions: Intron 2		Noncoding Regions: 3' Flanking Region	
	n	*H*	*l*	*S*	*r*	*s*	π	θ	π	θ	π	θ	π	θ	π	θ	π	θ
D. melanogaster																		
tra (1 pop'tn)[1]	10	3	926	2	0	1	0.79	0.60	0.00	0.00	3.15	2.39	0.81	1.43	0.00	0.00	0.00	0.00
tra (global)	10	5	926	3	1	2	1.92	1.79	0.45	0.79	6.32	4.79	0.00	0.00	0.00	0.00	0.00	0.00
tra (combined)[2]	20	6	926	4	1	2	1.36	1.42	0.22	0.63	4.81	3.81	0.40	1.14	0.00	0.00	0.00	0.00
[3]Average (24 loci)	-	-	-	-	-	-	4.02	4.03	NA	NA	13.5	13.5		[	10.8	10.5	]	
D. simulans tra	9	8	883	29	4	12	9.51	10.6	3.50	3.50	28.2	32.8	11.0	11.9	29.8	40.9	0.00	0.00
[3]Average (16 loci)	-	-	-	-	-	-	7.83	7.99	NA	NA	30.4	31.1		[	18.9	18.8	]	
D. mauritiana tra	4	4	884	13	4	4	7.55	7.91	5.20	5.26	14.8	16.1	8.06	8.80	9.43	10.3	0.00	0.00
[4]Average (4 loci)	-	-	-	-	-	-	4.51	4.37	0.63	0.61	20.1	19.7		[	NA	NA	]	
D. sechellia tra	4	2	1106	1	1	0	0.64	0.70	0.85	0.92	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
[4]Average (4 loci)	-	-	-	-	-	-	0.26	0.30	0.17	0.23	0.55	0.55		[	NA	NA	]	

n, number of sequences; *H*, number of haplotypes; *l*, length of sequence excluding sites with alignment gaps; *S*, number of total segregating sites; r, number of replacement polymorphic sites; *s*, number of synonymous polymorphic sites. π is calculated as in NEI (1987), θ is calculated as in Watterson (1975). Corrections for multiple hits were not performed. Nucleotide diversities have been multiplied by 10^{-3}. [1]Data from previously published Pennsylvanian sample (Walthour and Schaeffer 1994). [2]Includes sequences from this study, the population study and a previously published sequence (O'Neil and Beloté 1992). [3]Values determined by Moriyama and Powell (1996). [4]Calculated from the coding region of the following loci available from GenBank: *nullo*, *period*, *yp-2* and *zeste*. N, number of loci analyzed. Noncoding regions sequenced in *tra* include only intronic, untranslated male exon and 3' regions consisting of 328 nucleotides in *D. simulans* and 332 nucleotides in the other three species. Values in square brackets denote pooled diversity measures in 5', intron and 3' end as measured by Moriyama and Powell (1996). NA, data not available.

caused the underlying decrease in nucleotide variation, estimates of nucleotide diversity, θ and π, should differ since average pairwise differences converge to equilibrium much slower than the number of segregating sites. Both test statistics were negative, indicating an excess of 'new' alleles, but not significantly different from neutral expectations (Tajima's D = -0.557, $P>0.10$; Fu and Li's D* = -0.756, $P>0.10$). However, HKA tests on the extended *D. melanogaster* data set reject a neutral model of *tra* evolution. This test of selective neutrality is based on the premise that under neutrality, polymorphism and divergence should positively correlate (Hudson *et al.* 1987). Using the complete *D. melanogaster* data set (n=20 sequences) against a consensus *D. simulans* sequence, HKA tests revealed significant differences in both silent site comparisons ($\chi^2 = 4.26$, $P=0.04$) and total site comparisons ($\chi^2 = 5.36$, $P=0.02$). This raises the question whether the rapid divergence of *tra*, an essential gene in the sex determination pathway, has been driven by adaptation in the *D. melanogaster* complex.

Levels of variation found in sibling species of *D. melanogaster* suggest that other forces may be a factor in *tra*'s rapid evolution. Compared to *D. melanogaster*, significantly higher levels of variation, almost an order of magnitude greater, were observed at the *tra* locus among nine global strains of *D. simulans*. Such diversity levels are similar to other loci from *D. simulans* (Table 3.1). A total of 29 segregating sites are observed among the 883 nucleotides sequenced. Haplotype diversity is high (H=0.89). In the coding regions, four replacement and twelve synonymous polymorphic sites were found. Singletons comprised almost half of the total polymorphic sites giving negative, but non-significant D statistics (Tajima's D = -0.497, $P>0.10$; Fu and Li's D*=-0.612,

$P>0.10$). However, sequences utilized were not sampled from an "equilibrium population", so the use of these geneaological approaches must be treated with caution. As well, low sample size may cause these tests of neutrality to lack statistical power (Simonsen *et al.* 1995).

Although more extreme, the low levels of diversity found at the *tra* locus in *D. melanogaster* relative to *D. simulans* correspond to previously reported genome-wide patterns (Aquadro *et al.* 1988). In their comprehensive comparison of polymorphism data between these cosmopolitan species, Moriyama and Powell (1996) found average levels of silent and replacement polymorphisms to be twofold larger in loci found in *D. simulans* than in *D. melanogaster*. Demographic and other historical events that may have affected genetic differentiation at a genome-wide level have been used to explain the difference in polymorphism levels between the two species. The reduced diversity in *D. melanogaster* loci relative to levels found in *D. simulans* (Aquadro *et al.* 1988; Moriyama and Powell 1996), may be attributed to recent bottleneck/founder events in *D. melanogaster* (Choudhary and Singh 1987), larger effective population size in *D. simulans* (Aquadro *et al.* 1988; Akashi 1996) or by a more differentiated population substructure in *D. simulans* (Begun and Aquadro 1995; Irvin *et al.* 1998; Hamblin and Veuille 1999). The higher degree of variation in *D. simulans* may in fact be an underestimate given the lack of sequence variation studies from geographically diverse populations compared to the more liberally sampled *D. melanogaster* (Hamblin and Veuille 1999).

Another explanation is that a recent selective sweep (or alternatively, background

selection) has reduced nucleotide polymorphism in or around the *tra* locus in *D. melanogaster*. The recombinational landscape among *D. melanogaster*, *D. simulans*, and *D. mauritiana*, corresponds quite well to the observed levels of species diversity at the *tra* locus. True *et al.* (1996) compared recombination frequency between these three species and revealed high and almost equivalent coefficient of exchanges in *D. mauritiana* and *D. simulans* ($\approx$ 0.1). *D. melanogaster* exhibited the so-called "centromere effect" whereby a suppression of crossovers takes place around the centromere and decreases with distance. In the region of the third chromosome where *tra* resides, recombination rate in *D. melanogaster* (0.00106; from Kliman and Hey 1993) was shown to be an order of magnitude lower than its siblings, *D. simulans* and *D. mauritiana* (also see True *et al.* 1996). Any effect of selection (either positive or negative) will be more evident in *D. melanogaster*, where recombination is low (Hudson and Kaplan 1995; Charlesworth *et al.* 1993).

The other sibling species reveal contrasting levels of polymorphism. Nucleotide diversity is found to be quite high in *D. mauritiana* (Table 3.1). A total of 13 variable sites were distributed among four sequences and each sequence sampled represented a unique haplotype (H=1.0). Variability is comparable between exonic *vs.* intronic regions except for the common absence of variation in the 3' flanking region in all sibling species. In the coding regions, four replacement and four synonymous polymorphisms were observed. Compared to the average level of coding region variation observed from four other loci sequenced in *D. mauritiana*, *tra* locus diversity is almost twice as large, due mainly to replacement variation. In contrast, among four lines of the sibling species, *D.*

sechellia, only one polymorphism was observed at the *tra* locus. This replacement singleton is situated in one of the tandem repeats unique to *D. sechellia* (discussed below). The low levels of polymorphism found at *tra* was similar to the low nucleotide polymorphism found for other loci sequenced in *D. sechellia* such as *asense*, *period*, *yp2* and *zeste* (Hey and Kliman 1993).

Divergence between species of the *D. melanogaster* subgroup: Out of 18 loci which have been sequenced in all four species of the *melanogaster* complex (Appendix A), *tra* ranks second (after *Acp26Ab*) in average pairwise nonsynonymous divergence, d_N, between *D. melanogaster* and its three sibling species using the method of Nei and Gojobori (1986). Synonymous divergence, d_S, on the other hand, ranks sixth. In terms of species pair divergence between the three sibling species of the *D. simulans* clade, *D. sechellia* and *D. mauritiana* possessed the highest divergence, particularly in nonsynonymous substitutions (Table 3.2). *D. simulans-D. mauritiana* reveal a relative deficiency of fixed differences because many of the polymorphisms have remained shared between these species (Kliman *et al.* 2000; Table 3.2).

Using the four sibling species of the *D. melanogaster* subgroup, along with the closely related outgroup, *D. erecta*, the number of amino acid substitutions fixed in each of the *melanogaster* and *simulans* lineages was examined (Figure 3.2). The principle objective was to test whether the rate of *tra* amino acid fixation was similar in all lineages. Both parsimony and distance (Fitch-Margoliash) methods produced identical topologies except for minor differences within the *simulans/mauritiana* cluster. The separation of the *D. melanogaster* protein cluster from sequences of the *D. simulans* clade

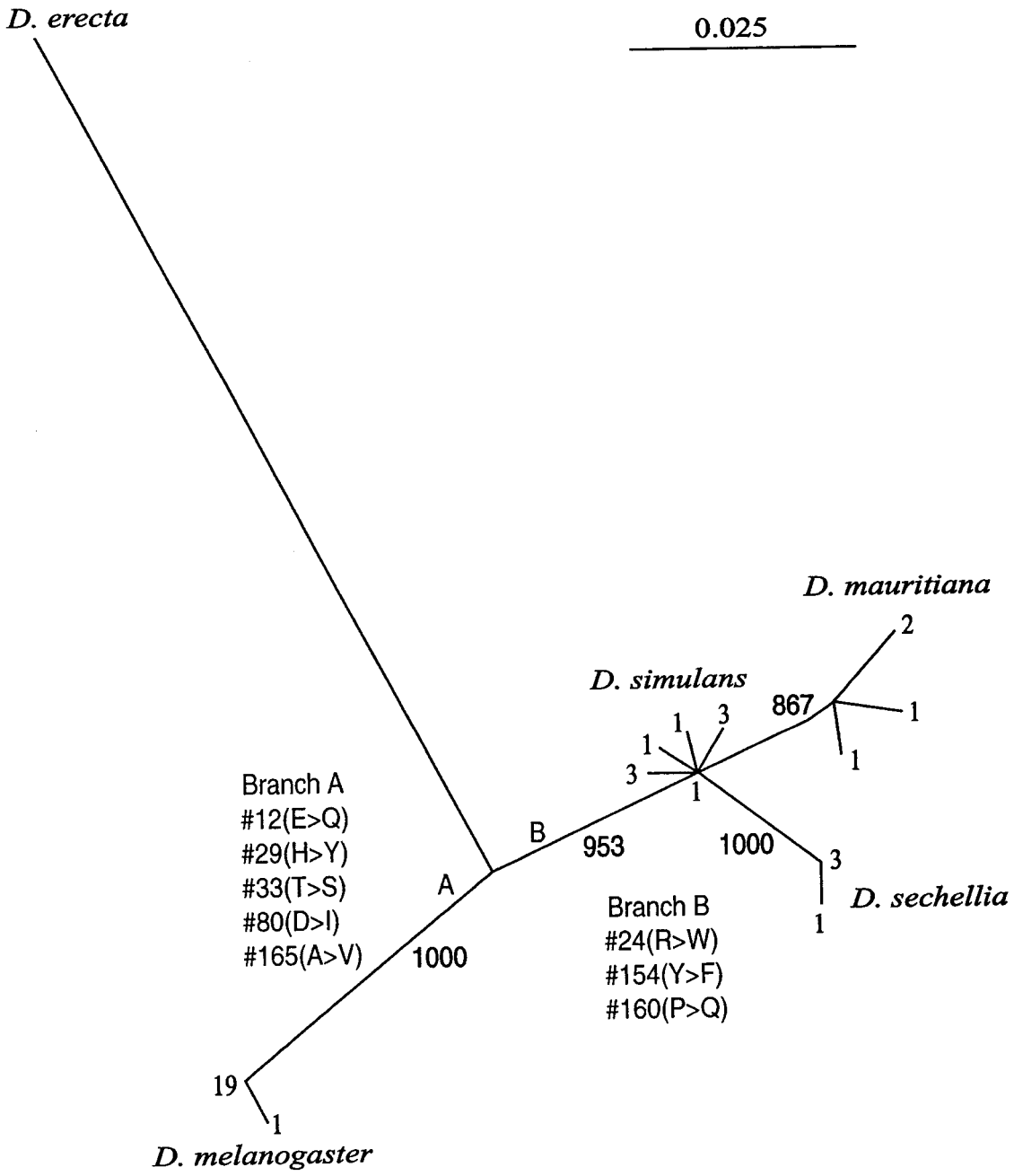
TABLE 3.2

tra polymorphism and divergence amongst species of the _D. melanogaster_ complex

	Within species		Between species		Significance level of MK test		Divergence		
species 1 - species 2	*s*	*r*	*S*	*R*	Fisher's Exact	G-test	d_S	d_N	d_N/d_S
mel–sim	14	5	17	12	0.36	0.28	0.146 (0.189)	0.039 (0.031)	0.27 (0.16)
mel–sec	2	2	21	15	1.00	0.75	0.131 (0.191)	0.043 (0.038)	0.33 (0.20)
mel–mau	6	5	20	13	0.74	0.73	0.125 (0.190)	0.045 (0.041)	0.36 (0.22)
sim–sec	11	5	3	3	0.62	0.42	0.046 (0.041)	0.012 (0.010)	0.26 (0.24)
sim–mau	15	8	1	1	1.00	0.67	0.042 (0.056)	0.011 (0.012)	0.26 (0.21)
sec-mau	4	5	6	4	0.66	0.50	0.039 (0.055)	0.017 (0.019)	0.44 (0.35)

Within species values refer to the total of all polymorphic differences in species 1 and 2. Between species values refer to the number of fixed differences between species 1 and species 2. *s* and *r*, silent and replacement polymorphism; *S* and *R*, silent and replacement fixed substitutions. The Fisher's exact test and G-test were used to assess the significance of the McDonald-Kreitman (1991) test of neutrality. d_S and d_N were calculated using a maximum likelihood approach implemented in the PAML package (Yang and Neilson 2000) on representative sequences from each species (see Appendix A). Values in brackets were calculated using the method of Li *et al.* (1985) on all possible sequence pairs and averaged. mel, *D. melanogaster*; sim, *D. simulans*; sec, *D. sechellia*; mau, *D. mauritiana*.

Figure 3.2 TRA parsimony tree in the *D. melanogaster* subgroup. Branch lengths are calculated by the Fitch-Margolias algorithm. Numbers at nodes represent the number of lines that have an identical protein sequence. Percentages separate species clusters and are for 1000 bootstraps supporting monophyly. Minimum parsimony substitutions are superimposed on the two internal branches, A and B, separating *D. melanogaster* from its sibling species, *D. simulans*, *D. mauritiana* and *D. sechellia*. The amino acid sites (identified by #) correspond to the site number in the outgroup, *D. erecta* in Figure 3.3. Substitutions at sites #58, #82 and #163 were ambiguous as to branch A or B.


D. erecta
0.025
D. mauritiana
2
D. simulans
867
1
1
1
3
1
1
3
1
Branch A
#12(E>Q)
#29(H>Y)
#33(T>S)
#80(D>I)
#165(A>V)
B
953
1000
3
D. sechellia
1
A
Branch B
#24(R>W)
#154(Y>F)
#160(P>Q)
1000
19
1
D. melanogaster

were supported by more than 95% of the bootstraps. Branch lengths along the *D. melanogaster* lineage were similar in length to those along the *simulans/mauritiana/sechellia* lineage. Protein parsimony was applied in order to determine *tra* substitutional differences between lineages and also did not support an increased rate of amino acid substitution in *tra* along the *D. melanogaster* lineage. Five amino acid substitutions could be placed on branch A leading to *D. melanogaster* and three on branch B leading to the *simulans* clade, while three substitutions remained ambiguous (Figures 3.2, 3.3) indicating no differences in evolutionary rates between lineages.

The constancy of evolutionary rate between lineages was also supported using a codon-based maximum likelihood approach. I showed that evolutionary models which utilized a molecular clock were not significantly different from models which allowed rates to vary between lineages (Table 3.3). In other words, models with free rate parameters did not offer a significant improvement in likelihoods to clocklike models. This was the case whether using *D. erecta* or *D. melanogaster* as the outgroup, or whether using representative sequences (n=4) or all sequences (n=37) with the use of local clocks (i.e. a single clock parameter was estimated for each of the species used) in the analysis. Testing a neutral model that uses a molecular clock against one without, McAllister and McVean (2000) also did not reject a neutral model of *tra* evolution on the three Sophophoran species, *D. melanogaster*, *D. simulans* and *D. erecta.*

The distribution of amino acid substitutions among lineages, as inferred by parsimony and maximum likelihood, does not support the fixation of beneficial *tra* alleles

Figure 3.3 TRA protein alignment in the *D. melanogaster* subgroup. Variable sites in the *D. melanogaster* complex (species are in boldface) are shaded in all sequences. Variable sites found solely in *D. erecta* are highlighted only in its sequence. Lightly shaded sites denote similar amino acid substitutions. A consensus sequence from each species was obtained by choosing for each site amino acids which were represented in the majority of lines. Replacement polymorphisms are highlighted, underlined and denoted in lowercase boldface. Indels are indicated by dashes. Both sets of repeats (found in *D. melanogaster* and *D. simulans*) are tandemly arranged and shown above the consensus sequences. Differences between repeats are shaded accordingly. The single region of the *tra* locus showing a significant higher substitution rate by permutation test is shown as an open box with asterisk below alignment. The last number denotes the number of amino acids in TRA for that particular species.

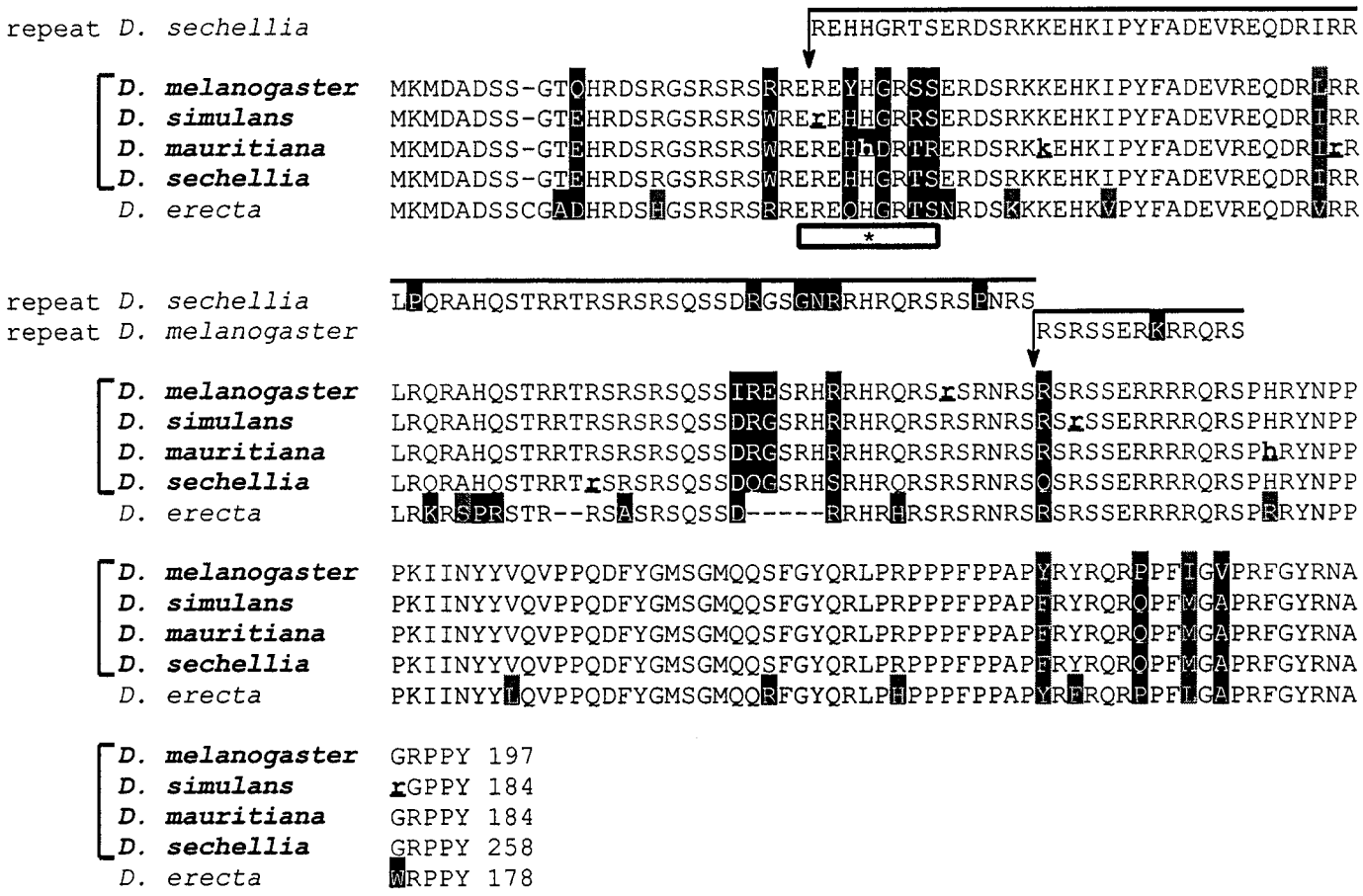

repeat D. sechellia
REHHGRTSERDSRKKEHKIPYFADEVREQDRIRR
D. melanogaster MKMDADSS-GTQHRDSRGSRSRSRRREREYHGRSSERDSRKKEHKIPYFADEVREQDRIRR
D. simulans MKMDADSS-GTEHRDSRGSRSRSWRErEHHGRRSERDSRKKEHKIPYFADEVREQDRIRR
D. mauritiana MKMDADSS-GTEHRDSRGSRSRSWREREHhDRTRERDSRKkEHKIPYFADEVREQDRIrR
D. sechellia MKMDADSS-GTEHRDSRGSRSRSWREREHHGRTSERDSRKKEHKIPYFADEVREQDRIRR
D. erecta MKMDADSSCGADHRDSHGSRSRSRREREQHGRTSNRDSKKKEHKVPYFADEVREQDRVRR
repeat D. sechellia
repeat D. melanogaster
LPQRAHQSTRRTRSRSRSQSSDRGSGNRRHRQRSRSPNRS
RSRSSERKRRQRS
D. melanogaster LRQRAHQSTRRTRSRSRSQSSIRESRHRRHRQRSrSRNRSRSRSSERRRRQRSPHRYNPP
D. simulans LRQRAHQSTRRTRSRSRSQSSDRGSRHRRHRQRSRSRNRSRSrSSERRRRQRSPHRYNPP
D. mauritiana LRQRAHQSTRRTRSRSRSQSSDRGSRHRRHRQRSRSRNRSRSRSSERRRRQRSPhRYNPP
D. sechellia LRQRAHQSTRRTrSRSRSQSSDQGSRHSRHRQRSRSRNRSQSRSSERRRRQRSPHRYNPP
D. erecta LRKRSPRSTR--RSASRSQSSD-----RRHRHRSRSRNRSRSRSSERRRRQRSPRRYNPP
D. melanogaster PKIINYYVQVPPQDFYGMSGMQQSFGYQRLPRPPPFPPAPYRYRQRPPFIGVPRFGYRNA
D. simulans PKIINYYVQVPPQDFYGMSGMQQSFGYQRLPRPPPFPPAPFRYRQRQPFMGAPRFGYRNA
D. mauritiana PKIINYYVQVPPQDFYGMSGMQQSFGYQRLPRPPPFPPAPFRYRQRQPFMGAPRFGYRNA
D. sechellia PKIINYYVQVPPQDFYGMSGMQQSFGYQRLPRPPPFPPAPFRYRQRQPFMGAPRFGYRNA
D. erecta PKIINYYLQVPPQDFYGMSGMQQRFGYQRLPHPPPFPPAPYRFRQRPPFLGAPRFGYRNA
D. melanogaster GRPPY 197
D. simulans rGPPY 184
D. mauritiana GRPPY 184
D. sechellia GRPPY 258
D. erecta WRPPY 178

TABLE 3.3

Likelihood ratio tests and d_N/d_S estimates for *tra* using codon substitution models

Species Sequence subset	Codon Length	$2\Delta l$				Estimate of $\omega=d_N/d_S$ under null model or model of best fit
		Site Variation		Lineage Variation		
		Constant *vs.* Variable rate of substitution	Constant *vs.* Variable d_N/d_S among sites	Clock *vs.* No clock	Constant *vs.* Variable d_N/d_S among lineages	
mel/sim/sec/mau (n=4)						
Coding Region	183	10.6**	2.11	0.81	2.06	0.20
Exon 1	13	1.27	0	2.22	1.06	0.12
Exon 2	120	7.28*	4.91	2.14	6.24	0.24
Exon 3	49	1.97	0	2.41	3.28	0.24
mel/sim/sec/mau (n=20/9/4/4)						
Coding Region	183	37.7**	12.0*	5.44[a]	2.37[a]	0.21
D. sechellia repeat region	74	5.73	4.54	4.74	10.9**	0.30
ere/mel/sim/sec/mau (n=5)						
Coding Region	176	2.75	5.41	0.52	4.50	0.32
Exon 1	13	1.49	2.07	3.97	2.03	0.24
Exon 2	114	2.08	3.91	0.26	6.40	0.35
Exon 3	49	3.89	0.62	1.90	4.71	0.32

A representative sequence from each species was selected (mel=penn2; sim=colom, mau=0241.3; sec=0248.0) except in the one case where all available sequences were used (n_{mel}=20, n_{sim} =9, n_{sec} = 4, n_{mau} =4). [a]In this case, a "local rate" (i.e. a single rate parameter) was estimated for each species. Codon frequencies directly calculated from nucleotide frequencies (F_3X4) and the transition/transversion ratio, κ, was estimated from the data. Each model is tested against its subset null model. $2\Delta l$ is twice the difference of the log-likelihood value used in the likelihood ratio test. *$P<0.05$. **$P<0.01$.

in *D. melanogaster*. Akashi (1996) reported a greater amino acid substitution rate in the *D. melanogaster* lineage, relative to the *D. simulans* lineage, for 8 out of 14 genes. Among these 14 proteins, 68 replacement substitutions were found in the *D. melanogaster* lineage compared to 28 in the *D. simulans* lineage. In his analysis, Akashi (1996) suggested that the overall increased rate of amino acid substitution in *D. melanogaster* could be due to the fixation of slightly deleterious mutations, but could not rule out the possibility of adaptive evolution. Eanes *et al.* (1996) found that 15 out of 21 replacement fixations in the *G6pd* gene between *D. melanogaster* and *D. simulans* occurred in the *D. simulans* lineage. Since they expected the fixation of slightly deleterious mutations to be more frequent in *D. melanogaster* because of its smaller population size, this was taken as evidence for adaptive evolution of *G6pd*, primarily in the early stages of separation from *D. simulans*. They further suggested that a similar burst of replacement fixations might have occurred for *tra*, and that this may be responsible for the reduced nucleotide polymorphism found in *D. melanogaster*. This analysis, however, does not support this hypothesis for *tra* divergence since amino acid substitutions have occurred more or less equally along both lineages and low polymorphism is unique only to *D. melanogaster*.

Heterogeneity in selective constraints between sites: Applying maximum likelihood to a codon-based substitution model (Yang 1997), the parameter, $\omega=d_N/d_S$, was estimated. This parameter, which is a measure of the level of selective constraints imposed on a coding sequence, was calculated using a series of different evolutionary models. Using the four sibling species of the *D. melanogaster* complex, the likelihood

function was maximized at $d_N/d_S = 0.20$ (Table 3.3) with a model which allowed for a variable rate of substitution among sites. In order to localize the region(s) of the gene with the least selective constraints, the coding region was subdivided into its three constituent exons. All exons were found to possess moderate functional constraints, with exons 2 and 3 bearing the least constraints (Table 3.3). However, since the first exon contains only 13 codons, there may not be sufficient power to detect whether differences are significant.

The heterogeneity in substitution rate across the *transformer* locus is also evident using clustering stastistics. In the *tra* coding region, the nonrandom distribution of replacement substitutions found among species of the *D. melanogaster* complex indicates the presence of regions with variable rates of evolution (Table 3.4). A maximum likelihood analysis confirms this result when all 37 sequences with lineage-specific parameters were used (Table 3.3). However, there is evidence that the constraints found in *tra* may also be lineage-specific. When *D. erecta*, a species that has diverged from the *D. melanogaster* lineage between 10 and 15 million years ago, is added to the analyses, a non-random distribution of nonsynonymous sites is not found (Table 3.4) Although many of the amino acid changes introduced by the inclusion of *D. erecta* are conservative in nature, they may reveal the existence of different levels of selection in both lineages. Codon substitution models also support this claim since heterogeneity among sites no longer produces significantly better likelihoods when *D. erecta* is included.

The addition of *D. erecta* sequence differs from similar analyses in the Drosophila subgenus whereby the inclusion of *D. virilis* appears to produce an even more significant

TABLE 3.4

Clustering statistics for variable codon sites among Sophophora and Drosophila subgenera in *tra*

Subgenus	Codon Sites	Variable	Statistic (Probability) Var(L)	Q	L_{max}
Sophophora					
mel/sim/mau/sec					
Replacement	177	21	**0.00313 (0.03)**	**0.0250 (0.09)**	**0.2584 (0.03)**
Synonymous	177	33	0.00062 (0.54)	0.0085 (0.60)	0.1067 (0.54)
ere/mel/sim/mau/sec					
Replacement	177	37	0.00050 (0.52)	0.0079 (0.18)	0.0899 (0.69)
Synonymous	177	49	0.00018 (0.93)	0.0040 (0.47)	0.0569 (0.97)
Drosophila					
americana					
Replacement	187	19	**0.00318 (0.09)**	0.0258 (0.39)	0.2287 (0.12)
Synonymous	187	24	0.00136 (0.38)	0.0181 (0.19)	0.1596 (0.30)
virilis/americana					
Replacement	185	30	**0.00134 (0.05)**	0.0102 (0.59)	**0.1774 (0.06)**
Synonymous	185	28	0.00081 (0.63)	0.0112 (0.74)	0.1183 (0.59)

Protein sequence from each of two subgenera were separately aligned and variable amino acid sites linearly mapped. Insertion and deletions were ignored. Var(L), variance of interval lengths statistic; Q, modified variance statistic; L_{max}, fractional length of longest interval statistic. Significance (in brackets) was assessed using the computer program of Goss and Lewontin (1996). The probabilities are the fraction of replicates for which the replicate statistic was equal to or greater than the sample value. Statistically significant values (α=0.10) are indicated in boldface. mel, *D. melanogaster*; sim, *D. simulans*; sec, *D. sechellia*; mau, *D. mauritiana*; ere, *D. erecta*.

non-random effect (Table 3.4). This does not appear to be a statistical anomaly since both subgenera possess similar amounts of replacement changes. Another piece of evidence which may indicate that functional constraints have evolved in a lineage-specific manner is the location of conserved *vs.* low constrained regions. In their analysis of *D. americana tra* nucleotide variation, McAllister and McVean (2000) showed a very high d_N/d_S in the third exon (d_N/d_S=1.19) and a more moderate ratio in the second exon (d_N/d_S=0.31). This study's results show that among members of the *D. melanogaster* complex, d_N/d_S ratios were equally moderate in the second and third exons (d_N/d_S =0.24).

Distribution of variable sites: Amino acid variation among species of the *D. melanogaster* complex appear to cluster in regions near the arginine-serine (RS) domains and in a C-terminal segment of the coding region (Figure 3.3). To test whether this clustering is significant, the broken stick model of Goss and Lewontin (1996) was employed. The distribution of amino acid sites that are variable within the *D. melanogaster* clade is significantly non-random using three of the statistics recommended by Goss and Lewontin (Table 3.4). The fourth statistic, L_{min} or minimum interval length, is unlikely to be useful when a large fraction of the sites are variable as is the case for *tra*. In contrast to replacement sites, variable synonymous sites are not significantly different from random. Using maximum likelihood and a codon-based substitution model on the *tra* coding region of the four sibling species, a model which incorporated substitution rate heterogeneity among sites was shown to fit the data significantly better than a model with a constant rate of substitution ($2\Delta l$=10.6, d.f.=2, P<0.05). Furthermore, the origins of this heterogeneity can be found in the second exon (Table 3.3). (Alternatively, the other

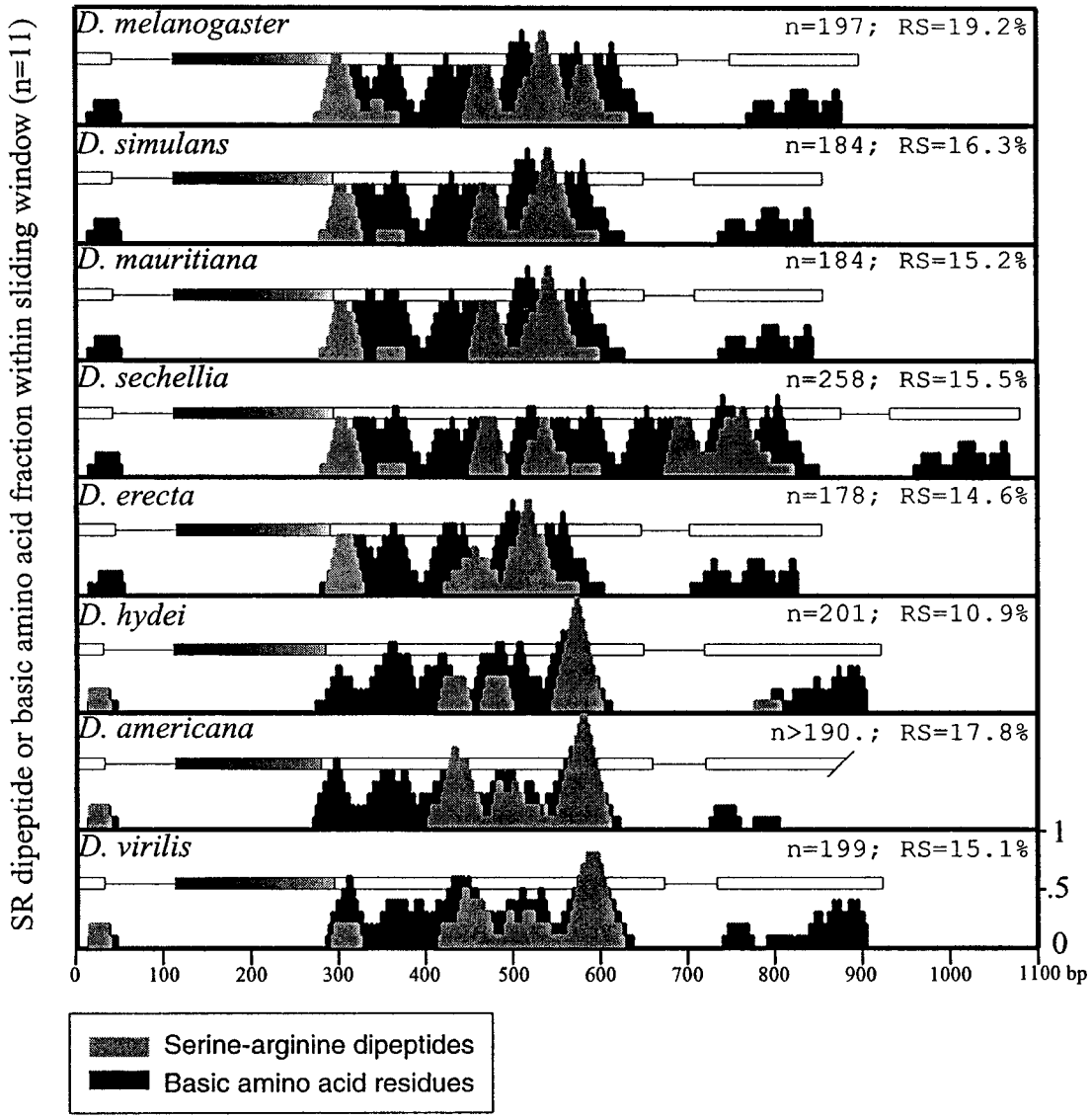
exons may not have enough power for heterogeneity to be detected.) It is evident from the *tra* protein sequence alignment of these five species of the Sophophoran subgenus, that regions which appear well conserved in the *D. melanogaster* complex are substituted in the *D. erecta* protein (Figure 3.3). This does not mean, however, that these domains are not conserved. In fact, substitutions in the C-terminal segment for which amino acid variability was not found within the *D. melanogaster* clade have relatively conservative amino acid replacements in *D. erecta* suggesting that it may be an unidentified functional domain.

Insertions and protein size evolution: The remarkable divergence in *tra* protein size is unprecedented between sibling species of the *D. melanogaster* complex. Figure 3.3 reveals the source of the disparity in amino acid sequence length. A fixed 13 amino acid insertion in *D. melanogaster* as well as a 74 amino acid insertion in *D. sechellia* are found at the *tra* locus in the *D. melanogaster* complex. One interesting feature is that both species-specific insertions are precisely arranged in tandem. The 222 base-pair insertion in *D. sechellia* originates from a single duplication near the N-terminus. Immediately adjacent to the 3' end of *D. sechellia*'s second repeat, are *D. melanogaster*'s two tandemly arranged repeats (each 39 base pairs). Figure 3.3 shows the tandem arrangement of both insertions. A sliding windows approach was employed to scan regions of the locus for significantly higher rates of evolution using maximum likelihood (Hartmann and Golding 2000). Interestingly, the only region which exhibited significantly higher rates of amino acid substitution when a maximum likelihood permutation test was applied is found around the putative *D. sechellia* insertion site

(Figure 3.3). This region may be a mutational hotspot for both nonsynonymous substitutions and duplications thus providing a non-selective causal mechanism for rapid evolutionary change. A similar correlation between repeat structure and rapid divergence is also observed at the *tra* locus in the subgenus Drosophila - a series of indel polymorphisms was found among *D. americana* sequences within its largest RS domain (data from McAllister and McVean 2000). A recent study has shown that the rate of evolution within and adjacent to simple repeat sequences found in the yeast genome was observed to be significantly higher than the surrounding protein (Huntley and Golding 2000). Hence, a correlation between repeated sequences and faster rates of evolution may be part of a general pattern of divergence.

Another interesting feature of *tra* gene evolution in the *D. melanogaster* subgroup is that each of the inserts carry additional arginine-serine (RS) domains, which are important in the recruitment of spliceosome factors. Figure 3.4 reveals the addition of one such domain in *D. melanogaster* and the addition of two large RS regions in *D. sechellia.* In both cases, basic amino acid regions were also added. However, while drastic changes in TRA protein structure are observed in the *D. melanogaster* complex, the proportion of the protein remaining as RS dipeptides appears to be maintained. This maintenance of RS domains may be one of the limiting selective constraints found at the *tra* locus and may play an important role in the proper functioning of TRA as a splicing initiator in sex determination.

Figure 3.4 Differences in gene length and amino acid profile of the *tra* locus across Drosophila. *D. melanogaster* and its closely related species, *D. simulans*, *D. mauritiana*, *D. sechellia* and *D. erecta* as well as the more distant species, *D. hydei*, *D. americana* and *D. virilis* are compared. Open boxes represent coding region. Shaded boxes represent untranslated region in exon 2. Full coding sequence was not available from *D. americana.* Gene structures are superimposed by graphs which indicate the fractional score of arginine-serine (RS) dipeptides (shown in gray) or basic amino acids (arginine, histidine or lysine; shown in black) along the *transformer* locus. The consensus sequence of each species was used to calculate exon length and amino acid scores. Scores are calculated as the fraction of RS dipeptides or basic amino acids within a sliding window of 11 amino acids and are indicated at each window's midpoint. Scores equal or greater than 0.5 represent a region of the protein with RS dipeptides or basic amino acids comprising a majority. n=number of amino acids in TRA; RS=percentage of TRA consisting of arginine-serine dipeptides.


SR dipeptide or basic amino acid fraction within sliding window (n=11)
D. melanogaster
n=197; RS=19.2%
D. simulans
n=184; RS=16.3%
D. mauritiana
n=184; RS=15.2%
D. sechellia
n=258; RS=15.5%
D. erecta
n=178; RS=14.6%
D. hydei
n=201; RS=10.9%
D. americana
n>190.; RS=17.8%
D. virilis
n=199; RS=15.1%
1
.5
0
0
100
200
300
400
500
600
700
800
900
1000
1100 bp
Serine-arginine dipeptides
Basic amino acid residues

Conclusions: TRA is involved in the regulation of somatic sexual differentiation in females by binding (with TRA2) to regulatory elements in *doublesex* and *fruitless* (Hoshijima *et al.* 1991; Steinman-Zwicky 1994). Since these two proteins control aspects of female sexual differentiation, including all somatic and behavioural components, it is possible that subtle changes in the *tra* protein may affect downstream genes involved in reproduction. However, we are far from understanding the full significance of *tra*'s rapid evolution. In a transgenic experiment with *tra*, O'Neil and Beloté (1992) transferred the wild type *tra* gene of *D. virilis* to *D. melanogaster* by P element-mediated germline transformation. The *D. virilis* gene was capable of shifting male structures in *D. melanogaster* flies, which were chromosomally female but homozygous for a *tra* deletion, towards femaleness. Whether the absence of a full recovery to femaleness was caused by positional effects of the transgenes or a difference in the evolution of different RS domain specificity between both species, remains an open question.

While a minimal presence of RS domains is maintained across Drosophila, I suggest that different functional constraints have evolved between species of the *D. melanogaster* and *D. americana* subgroups. The observed high rates of divergence combined with unusual gene structure (indels, duplications of RS domains), site- and lineage-specific heterogeneity in divergence rates, and low but varying functional constraints, identify *tra* as a different class of developmentally regulated genes in contrast to highly conserved developmental genes.

REFERENCES

Akashi, H. (1996) Molecular evolution between *Drosophila melanogaster* and *D. simulans*: reduced codon bias, faster rates of amino acid substitution, and larger proteins in *D. melanogaster*. Genetics **144:** 1297-1307

Aquadro, C.F., Lado, K.M., and W.A. Noon (1988) The *rosy* region of *Drosophila melanogaster* and *Drosophila simulans*. I. Contrasting levels of naturally occurring DNA restriction map variation and divergence. Genetics **119:** 875-888

Begun, D.J., and C.F. Aquadro (1995) Molecular variation at the *vermillion* locus in geographically diverse populations of *Drosophila melanogaster* and *D. simulans*. Genetics **149:** 1019-1032

Burtis, K.C., and B.S. Baker (1989) Drosophila *doublesex* gene controls somatic sexual differentiation by producing alternatively spliced mRNA's encoding related sex-specific polypeptides. Cell **56:** 997-1010

Chan, Y.M., and Y.N. Jan (1999) Conservation of neurogenic genes and mechanisms. Curr Opin Neurobiol. **9:**582-588

Charlesworth, B., Morgan, M.T., and D. Charlesworth (1993) The effect of deleterious mutations on neutral molecular variation. Genetics **134:** 1289-1303

Choudhary, M, and R.S. Singh (1987) A comprehensive study of genic variation in natural populations of *Drosophila melanogaster*. III. Variations in genetic structure and their causes between *Drosophila melanogaster* and its sibling species *Drosophila simulans*. Genetics **117:** 697-710

Eanes, W.F., Kirchner, M., Yoon, J., Biermann, C.H., Wang, I.-N., McCartney, M., and B.C. Verrelli (1996) Historical selection, amino acid polymorphism and lineage-specific divergence at the *G6pd* locus in *Drosophila melanogaster* and *D. simulans*. Genetics **144:** 1027-1041

Felsenstein, J. (1993) PHYLIP, phylogeny inference package. Version 3.5. Distributed by the author, Department of Genetics, University of Washington, Seattle, U.S.A.

Ferveur, J.-F., Savarit, F., O'Kane, C.J., Sureau, G., Greenspan, R.J., and J.-M. Jallon (1997) Genetic feminization of pheromones and its behavioural consequences in Drosophila males. Science **276:** 1555-1558

Fu, Y.-X., and W.-H. Li (1993) Statistical tests of neutrality of mutations. Genetics **133:** 693-709.

Gloor, G., and W. Engels (1992) Single-fly preps for PCR. D. I. S. **71:** 148-149

Gould, S.J. (1977) Ontogeny and Phylogeny. Harvard University Press, Cambridge

Goss, P.J.E., and R.C. Lewontin (1996) Detecting heterogeneity of substitution along DNA and protein sequences. Genetics **143:** 589-602

Hamblin, M.T., and M. Veuille (1999) Population structure among African and derived populations of *D. simulans*: Evidence for ancient subdivision and recent admixture. Genetics **153:** 305-317

Handa, N., Nureki, O., Kurimoto, K., Kim, I., Sakamoto, H., Shimura, Y., Muto, Y., and S. Yokoyama (1999) Structural basis for recognition of the *tra* mRNA precursor by the Sex-lethal protein. Nature **398:** 579-585

Hartmann, M., and G.B. Golding (1998) Searching for substitution rate heterogeneity. Mol. Phylogen. Evol. **9:** 64-71

Hey, J., and R.M. Kliman (1993) Population genetics and phylogenetics of DNA sequence variation at multiple loci within the *Drosophila melanogaster* species complex. Mol. Biol. Evol. **10:** 804-822

Hey, J., and J. Wakeley (1997) A coalescent estimator of the population recombination rate. Genetics **145:** 833-846

Hoshijima, K., Inoue, K., Higuchi, I., Sakamoto, H., and Y. Shimura (1991) Control of *doublesex* alternative splicing by *transformer* and *transformer*-2 in Drosophila. Science **252:** 8333-8336

Hudson, R.R., Kreitman, M., and M. Aguadé (1987) A test of neutral molecular evolution based on nucleotide data. Genetics **116:** 153-159

Hudson, R.R., and N.L. Kaplan (1995) Deleterious background selection with recombination. Genetics **141:** 1605-1617

Huntley, M., and G.B. Golding (2000) Evolution of simple sequence in proteins. J. Mol. Evol.. **51:** 131-140

Inoue K., Hoshijima, K., Higuchi, I., Sakamoto, H., and Y. Shimura (1992) Binding of the Drosophila *transformer* and *transformer*-2 proteins to the regulatory elements

of *doublesex* primary transcript for sex-specific RNA processing. Proc. Natl. Acad. Sci. U.S.A. **89:** 8092-8096

Irvin, S.D., Wetterstrand, K.A. Hutter, C.M., and C.F. Aquadro (1998) Genetic variation and differentiation at microsatellite loci in *Drosophila simulans*: evidence for founder effects in New World populations. Genetics **150:** 777-790

Kliman, R.M., and J. Hey (1993) Reduced natural selection associated with low recombination in *Drosophila melanogaster*. Mol. Biol. Evol. **10:** 1239-1258

Kliman, R.M., Andolfatto, P., Coyne, J.A., Depaulis, F., Kreitman, M., Berry, A.J., McCarter, J., Wakeley, J., and J. Hey (2000) The population genetics of the origin and divergence of the *Drosophila simulans* complex species. Genetics **156:** 1913-1931

Labate, J. A., Biermann, C.H., and W.F. Eanes (1999) Nucleotide variation at the *runt* locus in *Drosophila melanogaster* and *Drosophila simulans*. Mol. Biol. Evol. **16:** 724-731

McAllister, B.F., and G.A. McVean (2000) Neutral evolution of the sex-determining gene *transformer* in Drosophila. Genetics **154:** 1711-1720

Moriyama, E.N., and J.R. Powell (1996) Intraspecific nuclear DNA variation in Drosophila. Mol. Biol. Evol. **13:** 261-277

Nei, M., 1987 Molecular Evolutionary Genetics. Columbia University Press, New York

Nei, M., and T. Gojobori (1986) Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Mol. Biol. Evol. **3:** 418-426

Nusslein-Volhard, C. (1994) Of flies and fishes. Science **266:** 572-574

O'Neil, M.T., and J.M. Beloté (1992) Interspecific comparison of the *transformer* gene of Drosophila reveals an unusually high degree of evolutionary divergence. Genetics **131:** 113-128.

Patel, N.H. (1994) Developmental evolution: Insights from studies of insect segmentation. Science **266:** 581-590

Richter, B., Long, M., Lewontin, R.C., and E. Nitasaka (1997) Nucleotide variation and conservation at the *dpp* locus, a gene controlling early development in Drosophila. Genetics **145:** 311-323

Rozas, J., and R. Rozas (1997) DnaSP version 3: an integrated program for molecular population genetics and molecular evolution analysis. Bioinformatics **15:** 174-175

Ryner, L.C., Goodwin, S.F., Castrillon, D.H., Anand, A., Villella, A., Baker, B.S., Hall, J.C., Taylor, B.J., and S.A. Wasserman (1996) Control of male sexual behavior and sexual orientation in Drosophila by the *fruitless* gene. Cell **87:** 1079-1089

Simonsen, K.L., Churchill, G.A., and C.F. Aquadro (1995) Properties of statistical tests of neutrality for DNA polymorphism data. Genetics **141:** 413-429

Sosnowski, B.A., Beloté, J.M., and M. McKeown (1989) Sex-specific alternative splicing of RNA from the *transformer* gene results from sequence-dependent splice site blockage. Cell **58:** 449-459

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

Tajima, F. (1989) Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics **123:** 585-595

Thompson, J.D., Higgins, D.G., and T.J. Gibson (1994) CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position specific gap penalties and weight matrix choice. Nucleic Acids Res. **22:** 4673-4680

True, J.R., Weir, B.S., and C.C. Laurie (1996) A genome-wide survey of hybrid incompatibility factors by the introgression of marked segments of *Drosophila mauritiana* chromosomes into *Drosophila simulans*. Genetics **142:** 819-837

Walthour, C.S., and S.W. Schaeffer (1994) Molecular population genetics of sex determination genes: the *transformer* gene of *Drosophila melanogaster*. Genetics **136:** 1367-1372

Watterson, G.A. (1975) On the number of segregating sites in genetical models without recombination. Theor. Popul. Biol. **7:** 256-276

Yang, Z. (1997) PAML: a program package for phylogenetic analysis by maximum likelihood. CABIOS **13:** 555-556

Yang, Z., and R. Nielsen (1998) Synonymous and nonsynonymous rate variation in nuclear genes of mammals. J. Mol. Evol. **46:** 409-418

CHAPTER FOUR

Rapid fixation of alleles at the sex determination locus, *transformer-2*, in sibling species of the *D. melanogaster* subgroup

ABSTRACT

In the previous chapter, the sex determining transcription factor, *transformer* (*tra*), was demonstrated to be rapidly and neutrally evolving among closely related species of Drosophila. *tra* regulates female-specific splicing of the terminal control switch, *doublesex* (*dsx*), by binding to the ubiquitously expressed transformer-2 protein (TRA-2) via arginine-serine (RS) domains. Since TRA-2 plays a similar role in sex determination, it too may be evolving in a rapid manner. This study addresses the nature and mechanism of *tra-2* divergence through a comparative sequence analysis of the minimally functional transcript, *tra2-226*, in the four sibling species of the *D. melanogaster* complex. In terms of nonsynonymous substitutions per site, K_a, *tra-2* divergence exceeds that of *tra* in pairwise species comparisons between the three sibling species of the *D. simulans* species complex. Significant heterogeneity in substitution rate was also found among regions of TRA-2 using maximum likelihood. High levels of selective constraint are evident in the RNA recognition motif (RRM) domain (d_N/d_S = 0.06) compared to levels of variability outside this domain (d_N/d_S = 0.74). Within species variation was assessed, using five lines of *D. melanogaster* sampled from worldwide populations, to test levels of polymorphism against neutral expectations. Levels of within and between species variation, as a whole, did not deviate from neutral expectations (HKA test using *D. simulans* as an outgroup, P=0.36). However, when variation was apportioned as synonymous and nonsynonymous, *tra-2* revealed a complete absence of replacement polymorphism, violating neutral expectations that the ratio of

nonsynonymous to synonymous substitutions between species are similar to the ratio of nonsynonymous to synonymous polymorphisms within species (McDonald-Kreitman 1991). A comparison of these ratios using intraspecific variation from *D. melanogaster* and divergence from *D. simulans* rejects a neutral hypothesis of *tra-2* evolution (Fisher's Exact test, P=0.035). Using partial sequences of the *tra2-226* transcript from five global strains of *D. simulans*, this result is corroborated with certain statistical tests (G-test, P=0.029) but not with others (Fisher's Exact test, P=0.080). These results suggest that the rapid accumulation of replacement substitutions between sibling species at the sex determining locus, *tra-2*, as indicated in *D. melanogaster*, may be the result of positive Darwinian selection.

INTRODUCTION

Two aspects of sex determination evolution distinguish it from the evolution of other developmental systems. First, sex determining mechanisms in eukaryotes have evolved in a rather rapid manner (Hodgkin 1992; Marín and Baker 1998). This generalization is becoming increasingly more evident as the biology and development of more organisms become known (see Table 1.1). Even in cases where the molecular mechanisms of determining sex superficially look similar between organisms, these mechanisms may have independently evolved. Fruitflies and nematodes, for example, have common primary mechanisms of determining sex based on X:A signals but these primary genes are not homologous (see Cline and Meyer 1996). In other instances, systems of sex determination have been demonstrated to be fundamentally different even between strains of the same species (i.e. *Musca domestica*, see Dübendorfer *et al.* 1992). The second distinguishing feature of sex determination evolution is that some of the genes involved in the pathway are rapidly evolving. For example, the *Sry* locus, which acts as a primary control gene in mammalian sex determination, has been shown to be highly diverged in primates (Whitfield *et al.* 1993) and rodents (Tucker and Lundrigan 1993). The fungal *mid* locus, part of the sex determining genetic pathway, possesses high rates of substitution between *Chlamydomonas reinhardtii* and *C. incerta* (Ferris and Goodenough 1997). In nematodes, the primary sex determining genes, *transformer-1* and *transformer-2*, are also among the highest diverged genes between *Caenorhabditis elegans* and *C. brigassae* (deBono and Hodgkin 1996; Kuwabara and Hodgkin 1996).

In Drosophila, *transformer* (*tra*) has also been shown to be highly diverged between taxa (O'Neil and Beloté 1992; McAllister and McVean 2000, see Chapter 3). In the previous chapter, it was demonstrated that *tra* is not only rapidly evolving in terms of the amount of amino acid substitutions but that its protein can accommodate unprecedented large insertions of a putatively important protein binding domain in sibling species of the *D. melanogaster* subgroup. It was also found that while significant selective constraints were observed, a large fraction of the gene has been evolving relatively unconstrained. The rapid divergence of TRA is somewhat puzzling given that it is a primary developmental switch and a necessary component of female somatic differentiation.

For proper function to occur, TRA requires an interacting molecular partner. As a female-specific transcription factor, TRA functions only in the presence of the transcription factor, TRA-2, encoded by the *transformer-2* (*tra-2*) gene (Mattox *et al.* 1990). Both proteins contain arginine-serine (RS) domains important for protein-protein binding (Boggs *et al.* 1987, Amrein *et al.* 1988) and are part of a spliceosome complex which alternatively splices *dsx pre*mRNA into a female-specific transcript. When translated, DSXf regulates the repression of male somatic differentiation (Nogoshi *et al.* 1988, Hoshijima *et al.* 1991). In contrast to *tra*, which does not have an apparent function in males as first indicated by studies using null *tra* mutants (Sturtevant 1945) and validated in later molecular analyses (see Cline and Meyer 1996), *tra-2* is constitutively expressed in both female and male tissue. In addition to sexual differentiation in the female soma and germline (Oliver *et al.* 1993), *tra-2* also regulates

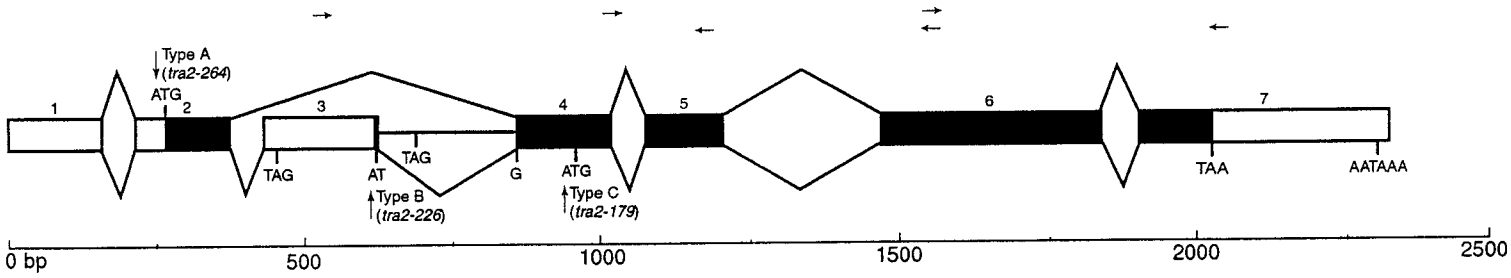
male-specific pathways. TRA-2 is necessary for male fertility (Watanabe 1975). Specifically, it is required in male germline cells for the efficient processing of male-specific *exuperantia pre*mRNA (Hazelrigg and Tu 1994) and male-specific splicing of *alternative testis pre*mRNA transcripts during spermatogenesis (Dauwalder *et al.* 1996).

Consistent with its multiple functions, *tra-2* expresses a variety of sex-specific mRNA transcripts, the products of alternative splicing. Of the four major transcripts, two are expressed mainly in the female soma and germline as well as the male soma (TRA2-254, TRA2-226, see Figure 4.1B) and the other two transcripts are expressed exclusively in the male germline (TRA2-179, TRA2-136) (Mattox *et al.* 1990). Using a transformed line that expresses a *tra2*-226 construct in a *tra-2* null background, Mattox *et al.* (1996) demonstrated that TRA2-226 was both necessary and sufficient for female sexual differentiation and male spermatogenesis.

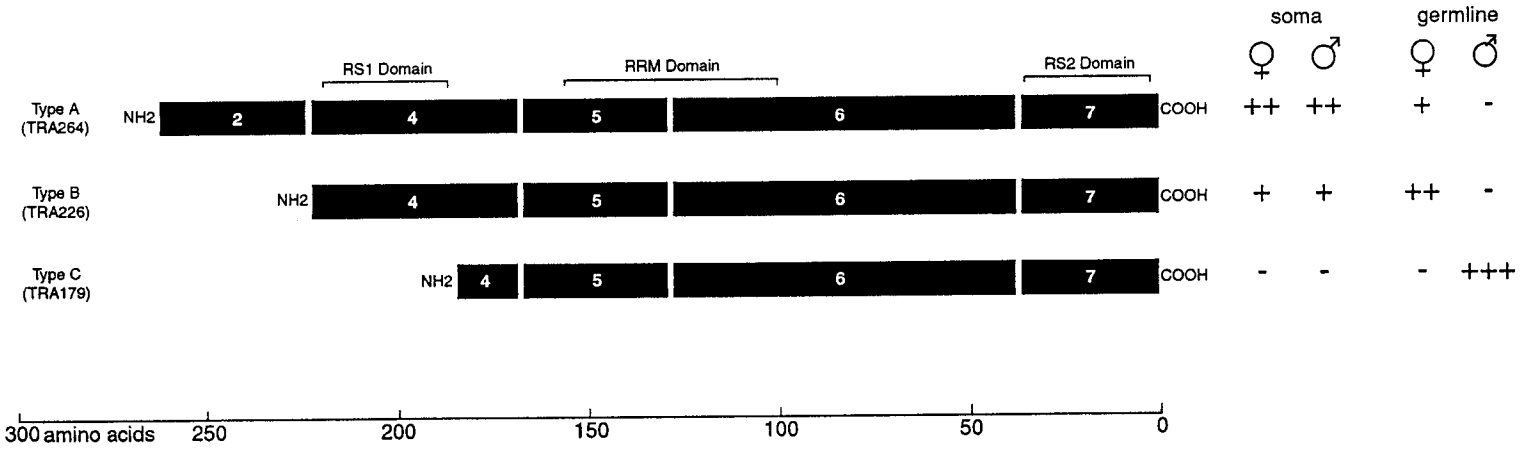
Recently, Chandler *et al.* (1997) compared the *tra-2* locus between *D. melanogaster* and *D. virilis* and found that both species produce a set of similar alternatively spliced mRNAs. Furthermore, they demonstrated that a *D. virilis* homolog could functionally replace the endogenous *D. melanogaster tra-2* gene and restore normal female differentiation and male fertility. Finally, they concluded that similar genetic pathways control sex-specific traits exist in both distant species. However, alignment of tra-2 homologs from both species reveals that this gene, like *tra*, has diverged considerably. Except for a conserved RNA recognition domain found in the middle of the protein and moderately conserved RS domains, much of the two protein homologs are diverged. In this chapter, I compare the nucleotide divergence of *tra-2* to the rapidly

Figure 4.1 Transcribed and translated products of *tra-2* in *D. melanogaster*. A. *tra-2* gene structure. Exons are indicated as boxes and numbered above. Coding portion of exons is filled while untranslated portion is open. Sex- and tissue-enriched transcripts are indicated above and below gene structure and are labelled at their start codon (indicated by a horizontal arrow). Type A transcript, encoded by a protein 264 amino acids long, is indicated above. Type B transcript encodes a protein 226 amino acid long and is indicated below. Also indicated below is a shorter *tra-2* transcript, *tra2-179*. Primers used in this study are shown above. Termination codons and polyadenylation sites are also indicated. B. Protein structure of the three major TRA isoforms. Numbers indicate exon of origin. The arginine-serine domains, RS1 and RS2, and the mRNA recognition motif domain, RRM, are indicated. Abundance data of each mRNA transcript (corresponding to its particular TRA-2 isoform) in male and female soma and germline is taken from Amrein *et al.* (1994).

A.


Type A
(tra2-264)
ATG
1
2
3
4
5
6
7
TAG
TAG
AT
Type B
(tra2-226)
G
ATG
Type C
(tra2-179)
TAA
AATAAA
0 bp
500
1000
1500
2000
2500

B.


soma
germline
RS1 Domain
RRM Domain
RS2 Domain
Type A
(TRA264)
NH2
2
4
5
6
7
COOH
++
++
+
-
Type B
(TRA226)
NH2
4
5
6
7
COOH
+
+
++
-
Type C
(TRA179)
NH2
4
5
6
7
COOH
-
-
-
+++
300 amino acids
250
200
150
100
50
0

evolving *tra* locus (Chapter 3) and the conserved *dsx* gene (Chapter 2). Here, I report that *tra-2* is a rapidly evolving gene with significant regional differences in selective constraints. Furthermore, I find substantive evidence that selection is driving the rapid divergence at the *tra-2* locus. A lower than expected number of replacement polymorphisms in both sibling species, *D. melanogaster* and *D. simulans*, relative to synonymous variation, suggests that this locus may have experienced recent bouts of positive selection.

MATERIALS AND METHODS

Fly stocks: Nucleotide variation at the *tra-2* locus was examined in the *D. melanogaster* species complex by sequencing multiple lines of *D. melanogaster* and *D. simulans* and a representative line from each of the two island endemics, *D. sechellia* and *D. mauritiana*. For all interspecific analyses, the published *D. melanogaster* sequence (accession number M30939, Mattox *et al.* 1990) was compared to a *D. simulans* sequence from Italy (S132, Drosophila Stock Centre at Umea, Sweden) as well as a sequence from *D. mauritiana* (0241.5 from Drosophila Species Stock Centre (D.S.S.C.), Bowling Green) and one from *D. sechellia* (0248.0, D.S.S.C.). To compare levels of intraspecific variation between *D. melanogaster* and *D. simulans*, each species was sampled from a worldwide distribution. Lines from Zimbabwe and Pennsylvania (Z34 and CPA124, respectively, kindly provided by Brian Lazzaro, Penn State), Italy (I13, provided by Alberto Civetta, University of Winnipeg) and Mysore, India (0231.6, D.S.S.C.) were used for *D. melanogaster*. For *D. simulans*, lines from Colombia (0251.2, D.S.S.C.), Florida (0251.166, D.S.S.C.), Madagascar (S24, kindly provided by John Roote, Cambridge University) and a strain of an unknown origin (1088, D.S.S.C.) were used. Only a portion of the coding region was sequenced for *D. simulans* polymorphism (see below). All flies were obtained as isofemale lines and maintained at low density on banana medium on a diurnal 12 hour light/dark cycle between 22° and 23°.

DNA extraction: For each species/line, genomic DNA was isolated from five flies. Briefly, flies were macerated in 100 ul of homogenizing buffer (0.1 M Tris HCl,

0.1 M EDTA, 1% SDS) and incubated at 70° for 30 minutes. 14 ul of 8M potassium acetate was added and left on ice for 30 minutes. Samples were spun at 4° for 20 minutes at 15 krpm and DNA precipitated with 50 µl of 100% isoproponal. After 10 minutes of centrifugation at 15 krpm, the pellet was washed twice with 70% EtOH. Finally, DNA was redissolved in ddH20 and stored at -20°.

Amplification and sequencing of the *tra-2* region: Figure 4.1A shows the genetic structure of the *tra-2* gene. The strategy employed to study nucleotide variation at the *tra-2* locus was to sequence the minimum portion of the gene demonstrated to be required for proper functioning of both sex determination in the female and fertility in the male. Of the three major transcripts, types A, B and C (Figure 4.1) which encode TRA2-264, TRA2-226 and TRA2-179, respectively, Mattox *et al.* (1996) showed that only the intermediate-sized transcript (Figure 4.1B) was needed for proper sex determination in the female and fertility in the male. Therefore, I choose to study the region of the gene corresponding to this transcript. Based on the aligned published sequences of *D. melanogaster* (GenBank accession number M30939, Mattox *et al.* 1990) and *D. virilis* (GenBank accession number U72682, Chandler *et al.* 1997), a series of primers was designed which would amplify the minimal functional coding region, *tra2-226*, both necessary and sufficient for somatic sexual differentiation and male fertility. Flanking (external) primers were devised so that exons three to seven were amplified, thus allowing for all of *tra2*-226's coding region to be sequenced. For most species/lines, the flanking primers, 5'CTCAGCCGATTCAGCTGGTGC3' and 5'GTTAATGAAGTAGCCTCCTCC3', were used to amplify a region corresponding to

1458 nucleotides in *D. melanogaster*. In *D. simulans*, the 3' primer, 5'GCGACCATCCACTTCTATTCC3', was used instead of the previously described 3' primer. In this case, the amplified product contains a smaller region of 1051 base pairs. All primers were constructed at MOBIX, McMaster University.

Since divergent sequences were amplified using the same set of primers, a touchdown PCR protocol was employed (Roux 1995). A Perkin-Elmer 480 thermal cycler was used with PCR buffer, $MgCl_2$ (1.5 mM) and Taq supplied by Fermentas. After a five minute hotstart, annealing temperatures commenced at 62° and every two cycles decreased one degree Celsius (except for three cycles per degree Celsius at 59° and 58°) to 54°. DNA disassociation and extension temperatures were 95° and 72°, respectively. PCR products were run on 1% agarose gels, compared to a 1 kb ladder (Fermentas) and checked for the presence of a single distinct band.

Before sequencing, PCR products were purified using a QIAGEN DNA purification kit. Sequencing was performed at MOBIX using an ABI PRISM® 377 DNA sequencer. Two additional sets of internal primers were used for sequencing. At the 5' region of the locus, the primer pair, 5'GCACAAGTCTCGCGTAAGCC3' and 5'GTCAATCACCATCTGGATGCG3' was used while the 3' portion utilized 5'GGAATAGAAGTGGATGGTCGC3' and its reverse complement, 5'GCGACCATCCACTTCTATTCC3' (see Figure 4.1) .

Sequence analysis: Nucleotide and protein sequences were aligned with CLUSTALX ver. 1.81 (Thompson *et al.* 1997). Synonymous, K_s, and nonsynonymous, K_a, divergence were calculated between sibling species using the method of Nei and

Gojobori (1986). Estimates of nucleotide diversity, θ and π, as well as selective tests of neutrality, were calculated and applied using DnaSP v3.51 (Rozas and Rozas, 1999). Evolutionary models including clock *vs.* no clock, constant *vs.* variable d_N/d_S between lineages, and constant *vs.* variable rates of substitution were compared using a maximum likelihood approach implemented by the program CODEML in the PAML package v3.0b (Yang 1997). These models were compared using likelihood ratio tests (twice the difference in log-likelihoods between models are compared to corresponding critical values of a chi-square distribution). The parameter, $\omega=d_N/d_S$, a measure of a protein's functional constraint, was estimated using various evolutionary models. Arginine-serine (RS) domains (as well as basic amino acid domains) were defined using a sliding window of eleven amino acids (code developed by Richard Morton, McMaster University). Each RS or SR dipeptide was counted as contributing a score of one, and the summed score is plotted at the midpoint of the window. For the RS domain analysis, the *D. virilis* sequence was taken from Chandler *et al.* (1997), GenBank accession number U72682.

Positional heterogeneity of amino acid substitution was assessed using the broken-stick model of Goss and Lewontin (1996). Clustering of variable sites is measured by four statistics: L_{max}, the maximum fractional interval length, L_{min}, the minimum fractional interval length, Var(L), the variance in fractional interval length and Q, a modified variance statistic. The significance of the observed values for each of these statistics was assessed by a Monte Carlo method of 10,000 replicates in which the same number of variable sites were distributed at random along the protein sequence. The probability of the statistic is the fraction of replicates exceeding the observed value.

RESULTS AND DISCUSSION

***tra-2* is a rapidly evolving sex determining gene:** As part of the sex determining pathway, *tra-2* plays a similar role as its regulatory counterpart, *tra.* TRA-2 and TRA bind to SR proteins of a spliceosomal complex and alternatively splice female-specific *dsx pre*mRNA transcripts. Previous studies have indicated that the *tra* locus is highly diverged across Drosophila (O'Neil and Beloté 1992; McAllister and McVean 2000) and in the previous chapter, a comprehensive sequence analysis confirmed *tra*'s rapid evolution among sibling species of the *D. melanogaster* complex. Using the same sibling species as a comparison, I found that *tra-2* is also rapidly evolving.

Considerable nucleotide variation was found at the *tra-2* locus (Appendix B3). The region sampled at this locus spans the untranslated exon 3 to the termination codon in exon 7 (Figure 4.1). Extensive indel variation was evident at the *tra-2* locus (data not shown). Most indels were situated in noncoding regions, including a tetranucleotide CCAA repeat polymorphism in *D. melanogaster* in the untranslated portion of exon 3 (*D. melanogaster* (n=5 lines) had 3-6 repeats of length 12, 16, 20 and 24 nucleotides while *D. simulans* (n=5 lines) was fixed for four repeats, *D. mauritiana* possessed three repeats and *D. sechellia* had four repeats). In coding regions, three small indels were found toward the carboxy terminus in three of the species (Figure 4.2).

A relatively large number of substitutions were found in *tra-2*'s coding region between the four species of the *D. melanogaster* complex. A high proportion of these substitutions were nonsynonymous (Appendix B3). In this study, the region of *tra-2*

Figure 4.2 TRA-2 protein alignment in the *D. melanogaster* species complex. Variable sites are highlighted. Lightly shaded sites indicate substitution of similar amino acids. Indels are indicated by dashes. The conserved RRM domain is boxed.


melanogaster 1 MSDYDYCGSRRHQRSSSRRRSRSRSSSESPPPEPRHRSGRSSRDRERMHKSREHPQASRC
simulans 1 MSDYDYYGSRRCQRSSSRRRSRSRSSSGSPPPEPRLRSGRSPHDRERMHKSREQPKASRC
sechellia 1 MSDYDYYGSRRCQRSSSRRRSRSGSSSGSPPPEPRLRSGRSPHDRERMHKSREQPKASRC
mauritiana 1 MSDYDYYGSRRCQRSSSRRRSRSRSSSGSPPPEPRLRSVRSPQDRDKSHKSREQPKASRC
5′ RRM Domain
melanogaster 61 IGVFGLNTNTSQHKVRELFNKYGPIERIQMVIDAQTQRSRGFCFIYFEKLSDARAAKDSC
simulans 61 IGVFGLNTNTSQHKVRELFNKYGPIERIQMVIDAQTQRSRGFCFIYFEKLSDARAAKDSC
sechellia 61 IGVFGLNTNTSQHKVRELFNKYGPIERIQMVIDAQTQRSRGFCFIYFEKLSDARAAKDNC
mauritiana 61 IGVFGLNTNTSQHKVRELFNKYGPIERIQMVIDAQTQRSRGFCFIYFEKLSDARVAKDSC
RRM Domain 3′
melanogaster 121 SGIEVDGRRIRVDFSITQRAHTPTPGVYLGRQPRGKAPRSFSPRRGRRVYHDRSASPYDN
simulans 121 SGIEVDGRRIRVDFSITQRAHTPTPGVYLGRQQRGKALRSYSPRRGRR--HDRSASPHDN
sechellia 121 SGIEVDGRRIRVDFSITQRAHTPTPGVYLGRQQRGKALRSYSPRRGRRPYHDRSASPHDN
mauritiana 121 SGIEVDGRRIRVDITITQRAHTPTPGVYLGRQQRGKALRSYSPRRGRRPYHDRSASPHDN
melanogaster 181 Y-RDRYDYRNDRYDRNLRRSPSRNRYTRNRSYSRSRSPQLRRTSSRY
simulans 179 YNRDRYNYRNDRNDRNLRRSPSRNRYTRNRSYSRSRSPQLHRTSSRY
sechellia 181 YNRDRDNYRNDRYDRNIRRSPSRNRYTRNRSYSRSRSPQLHRTSSRY
mauritiana 181 YNRDRYNYRNDRYDRNLRRSPNRNRYTRNRS--RSRSPQLHRTSSRY

which encoded the minimally functional transcript (Chandler *et al.* 1997), *tra2-226*, was sequenced and compared among sibling species. An alignment of the TRA2-226 protein between sibling species reveals a distribution of largely nonconservative amino acids outside the protein's conservative RRM domain.

Table 4.1 compares nonsynonymous and synonymous divergence of the sex determining genes, *tra-2*, *tra*, and *dsx*, between the four sibling species of the *D. melanogaster* complex. In the majority of cases, *tra-2* nonsynonymous divergence exceeds *tra* divergence in pairwise species comparisons. *tra-2*'s rapid rate of replacement change is best demonstrated when compared to the more conserved sex determining gene, *dsx*. While the rate of nonsynonymous substitution is an order of magnitude higher in pairwise species comparisons, *tra-2* and *dsx* synonymous divergence is similar. Hence, differences in mutation rate are not the cause of *tra*'s rapid evolution (Table 4.1).

This study demonstrates the existence of another highly diverged sex determination gene in Drosophila. The numerous examples of rapid evolving genes found in sex determining hierarchies (see Introduction) is analogous at the molecular level to the variety of sex determination mechanisms across taxa. The association of this evolutionary plastic developmental system with its rapidly evolving constituents will be further discussed in the next chapter (General Discussion).

Conservation of the RRM and RS domains: Although the overall rate of nonsynonymous substitutions is higher in *tra-2* compared to many other loci in Drosophila, regions of conservation exist. The *pre*mRNA recognition motif (RRM) domain comprises approximately one third of the minimally sufficient TRA2-226 protein

TABLE 4.1

Comparison of divergence between the sex determination loci, *tra-2*, *tra*, and *dsx*, in the *D. melanogaster* subgroup

Species	*mel*	*sim*	*mau*	*sec*
	tra-2	**0.0310**	**0.0504**	**0.0374**
mel	*tra*	0.0357	0.0396	0.0395
	dsx	0.0036	0.0045	0.0045
	0.1356	***tra-2***	**0.0102**	**0.0258**
sim	0.1547	*tra*	0.0084	0.0108
	0.0948	*dsx*	0.0009	0.0027
	0.1309	**0.0418**	***tra-2***	**0.0321**
mau	0.1500	0.0410	*tra*	0.0145
	0.1157	0.0294	*dsx*	0.0036
	0.1565	**0.0265**	**0.0325**	***tra-2***
sec	0.1692	0.0260	0.0374	*tra*
	0.1374	0.0431	0.0514	*dsx*

Coding regions of the *tra2-226* transcript, the complete *transformer* transcript, and exons two and five from *doublesex* are compared. Proportion of nonsynonymous (K_a) and synonymous substitutions (K_s) per site are found, respectively, in the upper and lower triangulars. For each pairwise species comparison, values correspond to *tra-2* (found in boldface), followed below by *tra*, then *dsx*.

yet only two amino acid substitutions are found among the four species sampled (Figure 4.2). This variability in levels of divergence across the tra-2 protein is clearly the result of higher levels of functional constraint in the RRM domain relative to the rest of the protein. Using a codon-based substitution model (Yang 1997), estimates of the parameter, d_N/d_S, a measure of a protein's selective constraint, were very low at the RRM domain ($d_N/d_S = 0.06$) compared to much higher estimates for the rest of the protein ($d_N/d_S = 0.74$) (Table 4.2). These results were consistent no matter which topology was used (Table 4.2). A model using a topology which forms a trichotomy between the three sibling species of the *D. simulans* clade produces similar d_N/d_S estimates as a model which uses the currently accepted phylogenetic relationship in which *D. sechellia* diverges first from a *D. melanogaster* common ancestor (Kliman *et al.* 2000).

Flanking the RNA recognition domain are two protein binding domains. These arginine-serine domains, RS1 and RS2, are found, respectively, at the amino- and carboxy termini. These domains are rich in RS dipeptides which are important in the facilitation of protein-protein interactions with other SR proteins that constitute the spliceosome complex (Fu 1991) as well as the intracellular localization of splicing factors to nuclear speckles (Li and Bingham 1991). TRA-2 and TRA, in particular, are jointly required for the alternative splicing of *dsx pre*mRNA in females and therefore play an essential role in the repression of male sexual differentiation. Protein alignment of TRA-2 homologs between *D. melanogaster* and the distantly related species, *D. virilis*, is quite difficult in both the RS1 and RS2 domains. Using a sliding window approach, I defined RS domains (as well as regions enriched with basic amino acids) by counting the proportion of RS

TABLE 4.2

D. melanogaster nucleotide diversity and likelihood ratio tests amongst its sibling species at the _tra-2_ locus

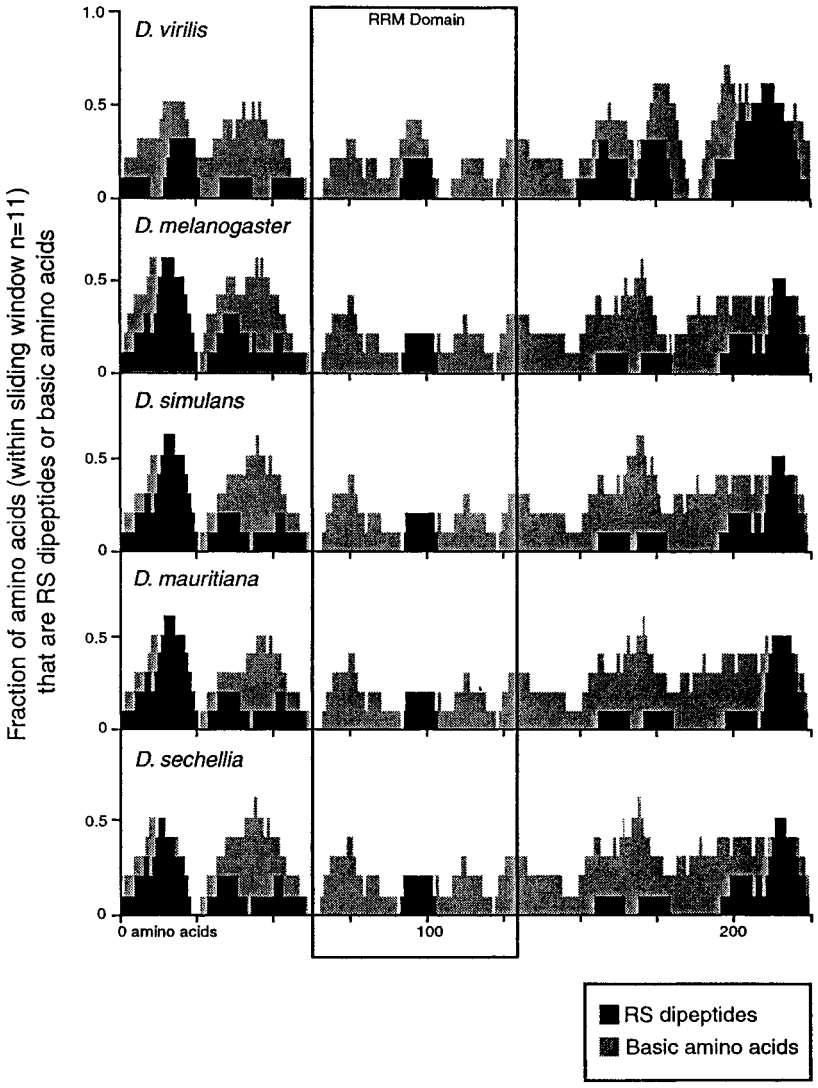
Species topology	Region of *dsx*	*D. melanogaster* coding region diversity: Summary statistics					No. of codons	2Δlog-likelihood: Clock *vs.* no clock	2Δlog-likelihood: Lineage *vs.* constant d_N/d_S	2Δlog-likelihood: Neutral *vs.* model of selection	Estimate of d_N/d_S under null model
		n	*H*	*R+S*	*S*	θ					
[a]Siblings resolved	RRM	5	4	4	4	11.1	72	1.9	7.1	0.0	0.06
	non-RRM	5	3	2	2	2.1	150	6.2*	10.6*	7.1	0.74
	Total *tra*226	5	5	7	7	4.9	222	3.9	17.2**	5.4	0.37
[b]Sibling trichotomy	RRM	-	-	-	-	-	72	1.1	5.7	0.0	0.06
	non-RRM	-	-	-	-	-	150	6.3*	10.6*	7.1	0.74
	Total *tra*226	-	-	-	-	-	222	5.6	14.4**	6.4	0.36

n, number of sequences; *H*, number of haplotypes; *R+S*, number of total segregating sites; *S*, number of synonymous polymorphic sites. θ is calculated for all sites as in Watterson (1975). Corrections for multiple hits were not performed. Values for θ were multiplied by 10^{-3}. [a](mel,((sim,mau),sec)); [b](mel,(sim,mau,sec)); *95% significance, **99% significance

dipeptides that are found in a window of eleven residues. The distribution of RS domains is shown in Figure 4.3 for species of the *D. melanogaster* complex as well as *D. virilis*. It is quite evident among the four sibling species that although substitutions are found in the RS1 and RS2 domains, RS domains of consistent profile are found at both ends of the protein. *D. virilis*, however, reveals large changes in RS profile. In particular, the RS1 domain contains a smaller proportion of RS dinucleotides (compared to other species). A window which contains only 30% RS dipeptides (compared to a window containing 60% RS dipeptides in *D. melanogaster*) is present at the *D. virilis* TRA-2 amino terminus (Figure 4.3). Of the RS1 domain as defined by Chandler *et al.* (1997), only 21% of the *D. virilis* domain consists of RS dipeptides while the four sibling species possess, on average, a domain comprising of 40% RS dipepetides. However, in the RS2 domain (as again defined by Chandler *et al.* 1997; also see Figure 4.1B), there is a greater proportion of RS dipeptides in *D. virilis* (35% RS dipeptides) compared to the *D. melanogaster* species complex (25% RS dipeptides) and may indicate the evolution of compensatory mutations that may aid is stabilizing RS content across taxa.

Evidence of selection at the *tra-2* locus in sibling species: Although numerous loci involved in the primary architecture of sex determination have been found to be rapidly evolving, the cause of their high evolutionary rates remains unknown (Marín and Baker 1998). Patterns of higher nonsynonymous to synonymous ratios in the mammalian *Sry* gene has been concentrated around terminal regions that may not be functionally important (Pamilo and O'Neill 1997). Alternatives to positive selection may explain these observations including the lack of functional constraints and the effects of

Figure 4.3 Amino acid profile of the *tra-2* locus among species of the *D. melanogaster* subgroup and *D. virilis*. *D. virilis* sequence taken from Chandler *et al.* (1997). Black graph indicates the fractional score of arginine-serine (RS) dipeptides (shown in black) or basic amino acids (arginine, histidine or lysine; shown in grey) along the TRA-2 protein. Scores are calculated as the fraction of RS dipeptides or basic amino acids within a sliding window of 11 amino acids and are indicated at each window's midpoint. Scores equal or greater than 0.5 represent a region of the protein with RS dipeptides or basic amino acids comprising a majority. The conserved RRM domain is also indicated.


Fraction of amino acids (within sliding window n=11) that are RS dipeptides or basic amino acids
1.0
0.5
0
D. virilis
RRM Domain
D. melanogaster
D. simulans
D. mauritiana
D. sechellia
0 amino acids
100
200
RS dipeptides
Basic amino acids

hitchiking on other loci in the nonrecombining Y-chromosome (Tucker and Lundrigan 1995). Previous studies have suggested that both selective and neutral mechanisms may explain the rapid replacement of amino acids in *transformer*, another Drosophila sex determining gene (Walthour and Schaeffer 1994; McAllister and McVean 2000). In the previous chapter, I conducted a comprehensive analysis of *tra* evolution in the *D. melanogaster* species clade and concluded that a lack of selective constraints in a large portion of TRA best explains its rapid divergence.

In contrast, this *tra-2* study suggests the presence of differential levels of selective constraint between species of the *D. melanogaster* complex as well as the possible occurrence of positive Darwinian selection on the *tra-2* locus. Using a maximum likelihood approach, significant heterogeneity in d_N/d_S was detected in sibling species (Table 4.2). Thus, significant differences in the level of selective constraints are found among species of the *D. melanogaster* subgroup which indicates the presence of substantial changes in selective pressure across closely related taxa. A test of the molecular clock, which examines whether an evolutionary model with species-specific rates of evolution is significantly better than a model with one common rate, also supports the existence of rate variation between taxa (Table 4.2). Such heterogeneity would be expected under a model of positive selection since selective pressures are expected to act in an independent manner and therefore should be specific to one lineage. However, an alternative explanation of variable d_N/d_S across species is that a relaxation in selection has occurred in at least one lineage. Using the same phylogenetic approach to detect the significant presence of amino acid sites with $d_N/d_S > 1$ (based on a codon model

of substitution), I did not reject a neutral mode of evolution (Table 4.2).

Using another approach to test selective *vs.* neutral evolutionary hypotheses, within species variation was surveyed at the *tra-2* locus in *D. melanogaster*. Four lines sampled from global populations were assayed along with the existing published *D. melanogaster* sequence for polymorphism. Out of 1344 nucleotides, 11 segregating sites were found. Tests of selective neutrality reveal that intraspecific variation at the *tra-2* locus does not deviate from a neutral model (Tajima's D = -0.10905, $P>0.10$; Fu and Li's D* = -0.10905, $P>0.10$). A comparison between *D. melanogaster* polymorphism and divergence (using *D. simulans* as an outgroup) was also conducted to test for deviations from neutral expectations. If *tra-2* variation is selectively neutral, then the ratio of segregating sites to sites with fixed substitutions should be similar to a corresponding neutral gene region (i.e. the 5' end of *Alcohol dehydrogenase*; HKA test, Hudson *et al.* 1987). The HKA test, essentially a goodness-of-fit test, was not significant ($P=0.36$).

An interesting feature of the *tra-2* data set which the HKA ignored was that all of the intraspecific variation found in *D. melanogaster* appears to be synonymous. Therefore, given the high levels of replacement substitution found between species, replacement polymorphism must be rapidly fixed relative to synonymous variation. To detect selection, a McDonald-Kreitman (1991) test of neutrality was performed on the *tra-2* protein coding sequence. Briefly, the ratio of replacement to synonymous polymorphism within species is expected to be similar, under a neutral hypothesis, to the ratio of replacement to synonymous fixed substitutions between species. Using a Fisher's Exact test, I found a significant under-representation of replacement polymorphism in *D.*

melanogaster (Table 4.3, P=0.035) indicating positive selection (G-tests could not be applied using this data set). In order to test whether this pattern is found in a closely related species, five worldwide lines of *D. simulans* was partially sequenced for the *tra2-226* transcript. The combined polymorphism from both species was then compared to their divergence (Table 4.3). Results were significant according to G tests (P = 0.029) indicating the possibility that selection may be driving *tra-2* gene evolution. However, a Fisher's Exact test was not significant in this case (P=0.080).

At least in the case of low *D. melanogaster* polymorphism, *tra-2* possesses patterns of replacement variation which depart from neutral expectations. Relatively high levels of replacement substitution *vs.* low replacement polymorphism indicate that amino acids were rapidly becoming fixed. And as stated previously, selection is not expected to act equally in all lineages. But why didn't other tests of selective neutrality, such as the HKA test, detect selection? In a comparison of genetic variation across nine coding regions of genes with polymorphism data from all three species of the *D. simulans* complex, Kliman *et al.* (2000) also performed the McDonald-Kreiman (1991) test of selective neutrality. They found that three genes, *esterase-6*, *janus*, and *zw*, showed significant deviations from a neutral model. However, in all three cases, HKA tests (Hudson *et al.* 1987) did not reveal differences from neutral expectations. This discordance between tests of selective neutrality is also found in *tra-2* which showed non-significant HKA tests. This discordance may be accounted by the fact that the HKA test may not have enough statistical power to differentiate between neutral and nonneutral models with such low numbers of segregating sites within sibling species. Power

TABLE 4.3

McDonald-Kreitman tests of selective neutrality at the *tra-2* locus

Between Species	Within Species	Number of Sites	Polymorphism		Divergence		Fisher's Exact Test	G Test		
			S	R	S	R		not corrected	Yate's correction	William's correction
D. melanogaster - *D. simulans*	*mel*, n=5	675	7	0	19	15	**0.035***	NA	NA	NA
D. melanogaster - *D. mauritiana*	*mel*, n=5	675	7	0	20	25	**0.010****	NA	NA	NA
D. melanogaster - *D. sechellia*	*mel*, n=5	681	7	0	22	20	**0.032***	NA	NA	NA
D. melanogaster - *D. simulans*	*mel*, n=5 *sim*, n=5	360	8	1	7	8	0.080	**0.029***	**0.036***	0.091

Fisher's exact test and G-tests (with and without corrections) are used to assess significance. *95% significance, **99% significance. A Bonferroni correction for multiple tests was not performed.

analyses have not yet been performed on either of these two selective tests of neutrality (Wayne and Simonsen 1998). Regardless, the HKA test may be a less reliable indicator of selection since assumptions of the HKA test are more prone to be violated. Unlike the HKA test, the McDonald-Kreitman test doesn't assume a constant population size at equilibrium. Thus, the McDonald-Kreitman test may serve as a better indication of the presence of adaptive mechanisms.

The results of the McDonald-Kreitman test using variation assessed from *D. melanogaster* and *D. simulans* suggest that selection may have driven *tra-2*'s evolution. Since TRA-2 and TRA are interacting molecular partners in *dsx pre*mRNA splicing, one might expect similar evolutionary processes to be causing the rapid divergence of these two sex determining proteins. The results from *tra-2* seem to run counter to the high rates of *neutral* evolution at the *transformer* locus (see preceding chapter). However, directional selection on *tra-2* may not be acting specifically in this pathway. TRA-2 has a broader range of functions other than female sexual differentiation. *tra-2* is necessary in male germ cells to process *exu* RNA (Hazelrigg and Tu 1994) as well as to autoregulate the processing of its own primary transcript in spermatogenesis (Mattox and Baker 1991). Hence, *tra-2*'s role in sperm development and maturation may be the target of selection. Presently, there is growing evidence that molecular adaptation is occurring in genes involved in spermatogenesis (Wyckoff *et al.* 2000; Yang *et al.* 2000; Torgerson *et al.* 2001). Therefore, positive selection acting on the fertility component of *tra-2* function may be possible.

Conclusions: *tra-2* represents the second example of a rapidly evolving sex determination gene in Drosophila. Relatively high numbers of replacement substitution have been fixed between closely related species relative to other loci from the *D. melanogaster* species complex. The rapid evolution of the majority of this gene contrasts the conservation of the RRM domain and indicates that both positive and negative selective pressures have been major contributors of TRA-2 divergence. Positive selection is compatible with the high rate of amino acid substitutions and the near absence of replacement, but not synonymous, polymorphism within two sibling species of the *D. melanogaster* complex. Selective signatures are not evident in *tra* evolution so coevolution is not a plausible explanation. Furthermore, the evolution of RS domains in TRA-2 between members of this species domain does not follow the more extreme divergence of RS domains observed in TRA. The absence of any apparent coevolutionary pattern in TRA suggests that the action of selection in other TRA-2 functions such as male fertility may best explain its rapid evolution.

REFERENCES

Amrein, H., Gorman, M., and R. Nöthiger (1988) The sex-determining gene *tra-2* of Drosophila encodes a putative RNA binding protein. Cell **55:** 1025-1035

Amrein H., Maniatis, T., and R. Nöthiger (1990) Alternatively spliced transcripts of the sex-determining gene *tra-2* of Drosophila encode functional proteins of different size. EMBO J. **9:** 3619-3629

Boggs, R.T., Gregor, P., Idriss, S., Beloté, J.M., and M. McKeown (1987) Regulation of sexual differentiation in *D. melanogaster* via alternative splicing of RNA from the *transformer* gene. Cell **50:** 739-747

Chandler, D., McGuffin, M.E., Piskur, J., Yao, J., Baker, B.S., and W. Mattox (1997) Evolutionary conservation of regulatory strategies for the sex determination factor *transformer-2*. Mol. Cell. Biol. **17:** 2908-2919

Cline, T.W., and B.J. Meyer (1996) Vive la différence: Males *vs.* females in flies *vs.* worms. Annu. Rev. Genet. **30:** 637-702

Dauwalder, B., Amaya-Manzanares, F., and W. Mattox (1996) A human homologue of the Drosophila sex determination factor *transformer-2* has conserved splicing regulatory functions. Proc. Natl. Acad. Sci. U.S.A. **93:** 9004-9009

deBono, M., and J. Hodgkin (1996) Evolution of sex determination in *Caenorhabditis*: Unusually high divergence of *tra-1* and its functional consequences. Genetics **144:** 587-595

Ferris, P.J., and U.W. Goodenough (1997) Mating type in *Chlamydomonas* is specified by *mid*, the minus-dominance gene. Genetics **146:** 859-869

Fu, X.-D. (1991) The superfamily of arginine/serine-rich splicing factors. RNA **1:** 663-680

Goss, P.J.E., and R.C. Lewontin (1996) Detecting heterogeneity of substitution along DNA and protein sequences. Genetics **143:** 589-602

Hazelrigg T, and C. Tu (1994) Sex-specific processing of the Drosophila *exuperantia* transcript is regulated in male germ cells by the *tra-2* gene. Proc. Natl. Acad. Sci. U.S.A. **91:** 10752-10756

Hodgkin, J. (1992) Genetic sex determination mechanisms and evolution. BioEssays **14:** 253-256

Hoshijima, K., Inoue, K., Higuchi, I., Sakamoto, H., and Y. Shimura (1991) Control of *doublesex* alternative splicing by *transformer* and *transformer-2* in Drosophila. Science **252:** 833-836

Kliman, R. M., P. Andolfatto, J. A. Coyne, F. Depaulis, M. Kreitman, A. J. Berry, J. McCarter, J. Wakeley, and J. Hey (2000) The population genetics of the origin and divergence of the *Drosophila simulans* complex species. Genetics **156:** 1913-1931

Kuwabara, P.E., and J. Hodgkin (1996) Interspecies comparisons reveal evolution of control regions in the nematode sex-determining gene *tra-2*. Genetics **144:** 597-607

Li, H., and P.M. Bingham (1991) Arginine/serine-rich domains of the *su(wa)* and *tra* RNA processing regulators target proteins to a subnuclear compartment implicated in splicing. Cell **67:** 335-342

Marín, I., and B.S. Baker (1998) The evolutionary dynamics of sex determination. Science **281:** 1990-1994

Mattox, W., Palmer, M.J., B.S. Baker (1990) Alternative splicing of the sex determination gene *transformer-2* is sex-specific in the germ line but not in the soma. Genes Dev. **4:** 789-805

Mattox, W., McGuffin, M.E., and B.S. Baker (1996) A negative feedback mechanism revealed by functional analysis of the alternative isoforms of the Drosophila splicing regulator *transformer-2*. Genetics **143:** 303-314

Mattox, W., and B.S. Baker (1991) Autoregulation of the splicing of transcripts from the *transformer-2* gene of Drosophila. Genes Dev. **5:** 786-796

McAllister, B.F., and G.A. McVean (2000) Neutral evolution of the sex-determining gene *transformer* in Drosophila. Genetics **154:** 1711-1720

McDonald, J.H., and M. Kreitman (1991) Adaptive protein evolution at the *Adh* locus in Drosophila. Nature **351:** 652-654

Nagoshi, R.N., McKeown, M., Burtis, K.C., Beloté, J.M., and B.S. Baker (1988) The control of alternative splicing at genes regulating sexual differentiation in D. melanogaster. Cell **53:** 229-236

Nei, M., and T. Gojobori (1986) Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Mol. Biol. Evol. **3:** 418-426

O'Neil, M.T., and J.M. Beloté (1992) Interspecific comparison of the *transformer* gene of Drosophila reveals an unusually high degree of evolutionary divergence. Genetics **131:** 113-128

Oliver, B., Kim, Y.J., and B.S. Baker (1993) *Sex-lethal*, master and slave: a hierarchy of germ-line sex determination in Drosophila. Development **119:** 897-908

Pamilo, P., R.J. O'Neill (1997) Evolution of the *Sry* genes. Mol. Biol. Evol. 14: 49-55

Roux, K.H. (1995) Optimization and troubleshooting in PCR. PCR Methods Appl. **4:** S185-194

Rozas, J., and R. Rozas (1999) DnaSP version 3: An integrated program for molecular population genetics and molecular evolution analysis. Bioinformatics **15:** 174-175

Sturtevant, A.H. (1945) A gene in *Drosophila melanogaster* that transforms females into males. Genetics **30:** 297-279

Thompson, J.D., Gibson, T.J., Plewniak, F., Jeanmougin, F., and D.G. Higgins (1997) The CLUSTAL_X windows interface: Flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic Acids Res. **25:** 4876-4882

Torgerson, D.G., Kulathinal, R.J., and R.S. Singh (2001) Evidence of positive selection in rapidly evolving mouse-human sperm orthologs. Submitted to Nature (Nov. 2001)

Tucker, P.K., and B.L. Lundrigan (1993) Rapid evolution of the sex determining locus in Old World mice and rats. Nature **364:** 715-717

Tucker, P.K., and B.L. Lundrigan (1995) The nature of gene evolution on the mammalian Y chromosome: Lessons from *Sry*. Philos. Trans. R. Soc. Lond. B Biol. Sci. **350:** 221-217

Walthour, C.S., and S.W. Schaeffer (1994) Molecular population genetics of sex determination genes: The *transformer* gene of *Drosophila melanogaster*. Genetics **136:** 1367-1372

Watanabe, T.K. (1975) A new sex-transforming gene on the second chromosome of

Drosophila melanogaster. Jpn. J. Genet. **50:** 269-271

Wayne, M.L., and K.L. Simonsen (1998) Statistical tests of neutrality in the age of weak selection. TREE **13:** 236-240

Whitfield, L.S., Lovell-Badge, R., and P.N. Goodfellow (1993) Rapid sequence evolution of the mammalian sex-determining gene *SRY* Nature **364:** 713-715

Wyckoff, G.J., Wang, W., and C.-I. Wu (2000) Rapid evolution of male reproductive genes in the descent of man. Nature **403:** 304-309

Yang, Z. (1997) PAML: a program package for phylogenetic analysis by maximum likelihood. CABIOS **13:** 555-556

Yang, Z., Swanson, W.J., and V.D. Vacquier (2000) Maximum-likelihood analysis of molecular adaptation in abalone sperm lysin reveals variable selective pressures among lineages and sites. Mol. Biol. Evol. **17:** 1446-1455

CHAPTER FIVE

General Discussion

5.1 Synopsis

Genetic variation provides the basis for evolutionary change. Therefore, a precise understanding of the variability found in various genetic systems will provide important insight into their past evolutionary history as well as their potential for future change. This approach has already proven successful in understanding the evolution of genes involved in sex and reproduction and their preferential importance in speciation (Singh and Kulathinal 2000). In an attempt to comprehend the evolutionary dynamics of a primary sexual system, this thesis focuses on evaluating variation in genes involved in sex determination - a system known to be mechanistically, quite variable across taxa (Hodgkin 1992; Marín and Baker 1998).

In Drosophila, sex determination is regulated by a small number of well-defined loci (Cline 1985; Baker 1989; Nagoshi and Baker 1990). The sex determining genetic hierarchy plays a critical role in sexual differentiation and represents one of the best characterized developmental systems in higher eukaryotes (Cline and Meyer 1996). While the developmental genetics of sex determination has been extensively studied, its evolutionary dynamics have been addressed with much less rigour. Previous studies have demonstrated a variety of evolutionary patterns from *transformer*'s high divergence across Drosophila (O'Neil and Beloté 1992) to the conservation of *doublesex* protein

binding domains across large phylogenetic distances (Raymond *et al.* 1998; Smith *et al.* 1999). In this thesis, I investigated the molecular evolution of three loci of the Drosophila sex determination pathway, *transformer* (*tra*), *transformer-2* (*tra-2*) and *doublesex* (*dsx*), among closely related species. My primary objective was to evaluate the genetic variation found among the principle components of this critical sexual system. The transcription factors, *tra* and *tra-2*, were found to be rapidly evolving while the downstream gene in this pathway, *dsx*, was conserved between species of the *D. melanogaster* subgroup. *tra* possessed low levels of functional constraints as well as insertions of sizable proportions in its protein. Outside its RNA binding domain, *tra-2* was also found to be rapidly evolving and the relative lack of replacement polymorphisms in both *D. melanogaster* and *D. simulans* indicates that this locus may have experienced episodes of adaptive evolution. *dsx*, while relatively conserved, showed signs of differences in selective constraints between distant lineages. The relatively high degree of genetic variation found among sex determination loci suggests that this sexual system possesses a greater capacity to generate heritable genetic variation and therefore a greater propensity for rapid evolutionary change. The findings of this thesis support recent studies that have observed greater variation in genes involved in sex and reproduction (Civetta and Singh 1995; Singh and Kulathinal 2000).

5.2 Evolution of Sexual Systems

5.21 Diversity of sex and reproduction-related traits/genes

Sexual dimorphism represents an extraordinary example of phenotypic diversity.

Although utilizing a common genetic complement, males and females in many cases show great differences in development, morphology and behaviour that are not exclusively associated with primary reproduction. As Charles Darwin observed (with his male biases being quite apparent),

> "There are, however, other sexual differences quite disconnected with the primary organs with which we are more especially concerned - such as the greater size, strength, and pugnacity of the male, his weapons of offence or means of defence against rivals, his gaudy colouring and various ornaments, his power of song, and other such characters.
> Darwin (1871, Chapter 8)

This phenotypic variation found *within* species is usually controlled by a defined genetic pathway. As a key sexual system, sex determination regulates the development of all aspects of sexual differentiation. In Drosophila, primary signals from autosomes and X-chromosomes are transduced to direct either male- or female-specific differentiation. Yet despite the critical importance of sex determination, a plurality of sex determination systems exists among various taxa indicating that this particular system has evolved in a rapid manner (Hodgkin 1992; Marín and Baker 1998). Insects, in particular, demonstrate a wide diversity of sex determination mechanisms (see Table 1.1). For example, in the common housefly, *Musca domestica*, different strains possess different primary mechanisms of determining sex. The male determining factor may act either dominantly in the Y-chromosome or the autosome or, in some strains, may altogether be replaced by a dominant female determining locus (Dübendorfer *et al.* 1992).

The diversity of sex determination mechanisms fit the general observation that sexual systems are highly variable in a wide range of taxa. A remarkable diversity of mating systems is found among organisms (reviewed in Smith 1984) and the widespread

diversity of sexual selection mechanisms (Andersson 1994) also indicate the variability of sexual systems.

In addition, evidence is mounting that a variety of traits involved sex and reproduction are evolving rapidly. In a classic example, Eberhard (1985) demonstrated that male genitalia from a wide range of animal taxa are extremely variable between species. This particular pattern is not surprising as entomologists have distinguished related species on the basis of male genitalia for centuries and floral characters were the basis for the Linnean system of classification. Another example demonstrates the origins of the widespread diversity of Hawaiian Drosophila. During a very short period of time, a rapid radiation of both fauna and flora took place on the Hawaiian archipelago. Through multiple founder effects, species of Drosophila have diverged to over 500 species. Much of this divergence has been proposed to be driven by the evolution of preferences for conspecific mates in newly founded allopatric populations (Carson 1997), indicating the importance of the rapid evolution of behavioural mating cues in speciation.

Two-dimensional electrophoresis on reproductive tract proteins first demonstrated the rapid evolution of SRR genes. Such studies using Drosophila found that, on average, testis proteins evolve more rapidly than proteins from other sampled tissues (Coulthart and Singh 1988; Thomas and Singh 1990). In particular, these studies found that between sibling species of Drosophila, which are morphologically identical but produce sterile hybrids when crossed (Lachaise *et al.* 1986; Kulathinal and Singh 1998), 20% of all variable testis protein spots revealed an absence of a detectable homolog in one of the species. These may represent proteins which have diverged significantly, possess

varying levels of gene expression, or are novel. Furthermore, highly diverged testis proteins, for the most part, were found to be less polymorphic among individuals within species (consistent with rapid diversifying selection) and in many cases differed in levels of gene expression between species. In a subsequent study, using a larger sample of proteins, tissue comparisons and species groups, it was found that not only male reproductive tissues were rapidly evolving, but female reproductive tissues (i.e. ovaries) were also highly diverged between sibling species (Civetta and Singh 1995).

A growing number of sex-specific characters involved in mating and fertility have been found to be rapidly evolving. Traits affecting copulation and fertility which include such primary sexual traits as testis and sperm length (Karr and Pitnick 1996; Joly *et al.* 1995) as well as sperm displacement mechanisms (reviewed in Birkhead and Möller 1998) have been shown to be highly diverged between species. Genes that affect mating behaviour such as *period* (involved in Drosophila mate song rhythm; Ritchie and Kyriacou 1994) as well as those involved in sperm/egg interactions in marine invertebrates (Lee *et al.* 1993; Metz and Palumbi 1996; Swanson *et al.* 1998; Hellberg and Vacquier 1999) and sperm competition (male-male interaction), as exemplified by such accessory gland proteins as *Acp26Aa* and *Acp70A* in Drosophila (Cirera and Aguadé 1997; Tsaur and Wu 1997; see Begun *et al.* 2000), have been demonstrated to be rapidly evolving. Swanson *et al.* (2001a) recently demonstrated the rapid evolution of female proteins involved in sperm-egg binding in mammals. Table 5.1 demonstrates the diversity of rapidly evolving sex and reproduction-related (SRR) genes among a wide range of taxa.

Table 5.1

Examples of highly diverged sex and reproduction-related (SRR) genes

Taxa	Species / Groups Compared	Gene	Functional Classification	Relevant observations	Ref.
Mammals	Primates	*SRY*	Sex determination	high Ka/Ks between species	(1)
	Primates	*Prm-1, 2*	Spermatogenesis	high Ka/Ks between species	(2,3)
	Humans, Old World Monkeys	*Acr-tryp inh.*	Fertilization (epididymis)	high Ka/Ks between species	(3)
	Humans, Old World Monkeys	*PSP94*	Fertilization (prostate fluid)	high Ka/Ks between species	(3)
	Humans, Old World Monkeys	*fertilin beta*	Gamete interaction (male)	high Ka/Ks between species	(3)
	Humans, Old World Monkeys	*TSPY*	Spermatogenesis	high Ka/Ks between species	(3)
	Mammals	*Tnp-2*	Spermatogenesis	variable dN/dS among sites	(4)
	Mammals	*ZP2, ZP3*	Fertilization (egg coat)	variable dN/dS among sites	(4)
	Mammals	*OGP*	Gamete interaction (female)	variable dN/dS among sites	(4)
	Rodents	*Sry*	Sex determination	high Ka/Ks between species	(5)
	Rodents	*Pem*	Gametogenesis in male/female	high Ka/Ks between species	(6)
	Musculus subspecies	*Abpa*	Sex peptide (salivary gland)	high Ka/Ks between species	(7)
Drosophila	*D. melanogaster* subgroup species	*Acp26Aa*	Fertilization (seminal fluid)	high Ka/Ks between species	(8)
	D. melanogaster / D. simulans	*Acp29AB*	Fertilization (seminal fluid)	excess of fixed replacements	(9)
	D. melanogaster species complex	*Acp70A*	Fertilization (seminal fluid)	low protein polymorphism	(10)
	D. melanogaster species complex	*OdsH*	Spermatogenesis	high Ka/Ks between species	(11)
	D. melanogaster species complex	*ocn/janA,B*	Spermatogenesis	higher rates of substitution	(12)
	D. virilis / D. melanogaster complex	*tra*	Sex determination	higher rates of substitution	(13,14)
	D. melanogaster species complex	*tra2*	Sex determination	high Ka/Ks between species	(15)
	D. melanogaster / D. simulans	*Sdic*	Sperm structure	evidence of selective sweep	(16)

Table 5.1 (cont'd)

Taxa	Species / Groups Compared	Gene	Functional Classification	Relevant observations	Ref.
Marine Invertebrates	*Haliotis* (abalones)	*lysin*	Fertilization (sperm)	high Ka/Ks between species	(17)
	Haliotis (abalones)	*VERL*	Fertilization (egg)	concerted evolution of repeats	(18)
	Echinometra (sea urchins)	*bindin*	Fertilization (sperm)	fixed replacement substitutions	(19)
	Tegula, Norrisia (snails)	*TMAP*	Fertilization (sperm)	high Ka/Ks between species	(20)
Nematodes	*Caenorhabditis elegans / C. brigassae*	*tra1, tra2*	Sex determination	low identity (out of 15 proteins)	(21,22)
Algae	*Chlamydomonas reinhardtii / C.incerta*	*fus1*	Gamete recognition	no homologs detected	(23)
	Chlamydomonas reinhardtii / C.incerta	*mid*	Sex determination	high Ka, Ks between species	(24)
Plants	*Brassicaceae*	*S-locus*	Gamete recognition	high allelic diversity	(25)
	Arabadopsis thaliana / Brassica nigra	*CO, COL1*	Flower induction response	high Ka/Ks	(26)
	Rosaceae	*S-RNase*	Gamete recognition	high Ka/Ks on surface regions	(27)

(1) L. S. Whitfield, R. Lovell-Badge, P. N. Goodfellow. *Nature* **364**:713 (1993). (2) A. P. Rooney, J. Zhang. *Mol. Biol. Evol.* **16**:70 (1999). (3) G. J. Wyckoff, W. Wang, C.-I. Wu. *Nature* **403**:304 (2000). (4) W. J. Swanson, Z. Yang, M. F. Wolfner, C. F. Aquadro. *Proc. Nat. Acad. Sci. U.S.A.* **98**:2509 (2001). (5) P. K. Tucker, B. L. Lundrigan. *Nature* **364**:715 (1993). (6) K. A. Sutton, M. F. Wilkinson. *J. Mol. Evol.* **45**:579 (1997). (7) R. C. Karn, M. W. Nachman. *Mol. Biol. Evol.* **16**:1192 (1999). (8) S.-C. Tsaur, C.-I. Wu. *Mol. Biol. Evol.* **14**:544 (1997). (9) M. Aguadé. *Genetics* **152**:543 (1999). (10) S. Cirera, M. Aguadé. *Genetics* **147**:189 (1997). (11) C.-T. Ting, S.-C. Tsaur, M.-L. Wu, C.-I. Wu. *Science* **282**:1501 (1998). (12) J. Parsch, C. D. Meiklejohn, E. Hauschteck-Jungen, P. Hunziker, D. L. Hartl. *Mol. Biol. Evol.* **18**:801 (2001). (13) M. T. O'Neil, J. M. Beloté. *Genetics* **131**:113 (1992). (14) R. J. Kulathinal, L. Skwarek, R. A. Morton, R. S. Singh. submitted to Genetics (2001). (15) R. J. Kulathinal, manuscript in preparation. (16) D. Nurminsky, D. D. Aguiar, C. D. Bustamante, D. L. Hartl. *Science* **291**:128 (2001). (17) Y. H. Lee, T. Ota, V. D. Vacquier. *Mol. Biol. Evol.* **12**:231 (1995). (18) W. J. Swanson, V. D. Vacquier. *Science* **281**:710 (1998). (19) E. C. Metz, S. R. Palumbi. *Mol. Biol. Evol.* **13**:397 (1996). (20) M. E. Hellberg, G. W. May, V. D. Vacquier. *Mol. Biol. Evol.* **17**:458 (2000). (21) M. deBono, J. Hodgkin. *J. Genetics* **144**:587 (1996). (22) P. E. Kuwabara, J. Hodgkin. *J. Genetics* **144**:597 (1996). (23) P. J. Ferris, J. P. Woessner, U. W. Goodenough. *Mol. Biol. Cell* **7**:1235 (1996). (24) P. J. Ferris, U. W. Goodenough. *Genetics* **146**:859 (1997). (25) M. K. Uyenoyama. *Genetics* **147**:1389 (1997). (26) U. Lagercrantz, T. Axelsson. *Mol. Biol. Evol.* **17**:1499 (2000). (27) T. Ishimizu, T. Endo, Y. Yamaguchi-Kabata, K. T. Nakamura, E. Sakiyama, S. Norioka.*FEBS Lett.* **440**:337 (1998).

Recent studies in comparative genomics are also demonstrating the rapid evolution of sex and reproduction related genes. In a comparison of over 2400 orthologs between humans and rodents, Duret and Mouchiroud (2000) found that genes expressed in both the uterus and testis possessed a higher than average number of nonsynonymous substitutions. Reproductive proteins, as a functional class, were shown to be highly diverged among primates (Wyckoff *et al.* 2000). Using a functional genomics approach, Mayfield *et al.* (2001) demonstrated that pollen coat proteins of *Arabidopsis thaliana*, important for "mate recognition", have considerably diverged among different ecotypes as well as from *Brassica oleracea*. Comparing expressed sequence tags (ESTs) from accessory glands between the sibling species, *D. melanogaster* and *D. simulans*, Swanson *et al.* (2001b) found that one out of every ten genes had a ratio of nonsynonymous to synonymous substitutions significantly greater than unity, suggesting that their rapid divergence has been driven by positive selection.

5.22 Evolutionary novelty in SRR traits/genes

Novel evolutionary innovations that are involved in sex and reproduction abound in the literature. In many cases, sexual selection has been suggested to direct the evolution of such novel traits. For example, Carson (1997) observed a number of *de novo* sexual characters between closely related species of Hawaiian Drosophila. In members of the *D. planitibia* subgroup, the distribution of foreleg cilia, which are used in copulation, have diverged between species, possibly due to sexual selection. Thus, sexual selection may be a potent force in the generation of morphological innovation. Even classic

examples of evolutionary novelty generated by natural selection have been reevaluated in the context of sexual selection. The variation in beak size among species of Darwin's finches, distributed across different islands of the Galapagos archipelago, had previously been thought to be solely the result of foraging optimization. However, a recent study has shown that beak size variation may in fact be due to the rapid evolution of mate song preferences (Podos 2001). Therefore, the classic example of beak size as a consequence of ecological adaptation may be one of sexual trait novelty.

In addition to the evolution of sexual trait novelty, a growing number of molecular studies are discovering genes which have recently appeared *de novo* and have evolved a novel function in reproductive tissue. Many of these genes show high rates of nucleotide substitution. This accelerated rate of change in genes that have evolved a novel function has been previously observed. Early studies testing the validity of molecular clocks found numerous examples whereby the rates of amino substitution suddenly accelerate after a new function has been acquired. One classic example demonstrated the rapid fixation of 13 amino acids in a testes-specific isoform of cytochrome c in mouse (Carlson *et al.* 1977). This particular isoform is coded by a gene which was duplicated from the relatively conserved, but ubiquitously expressed, cytochrome c gene.

Sdic is an example of a gene that has recently evolved *de novo* in Drosophila (Nurminsky *et al.* 1998). This particular gene has been demonstrated to be a fusion between two neighbouring loci and functions specifically in *D. melanogaster* sperm as an axonemal dynein subunit. *Sdic* possesses a novel testes-specific promoter derived from a protein-coding region and contains a new protein-coding exon derived from an intron. In

Drosophila, other cases of genes evolving a novel testis function have been demonstrated. After identifying *Odysseus,* a factor involved in hybrid male sterility between *D. simulans* and *D. mauritiana*, Ting *et al.* (1998) demonstrated the rapid divergence of its homeodomain, which may coincide with novel function in the *D. mauritiana* testes. *jing-wei* and *ocnus* are two other examples of genes which were demonstrated to result from recent duplications in Drosophila (Long *et al.* 1995; Parsch *et al.* 2001). In both cases, novel function was acquired in the testis and high rates of evolutionary change were evident.

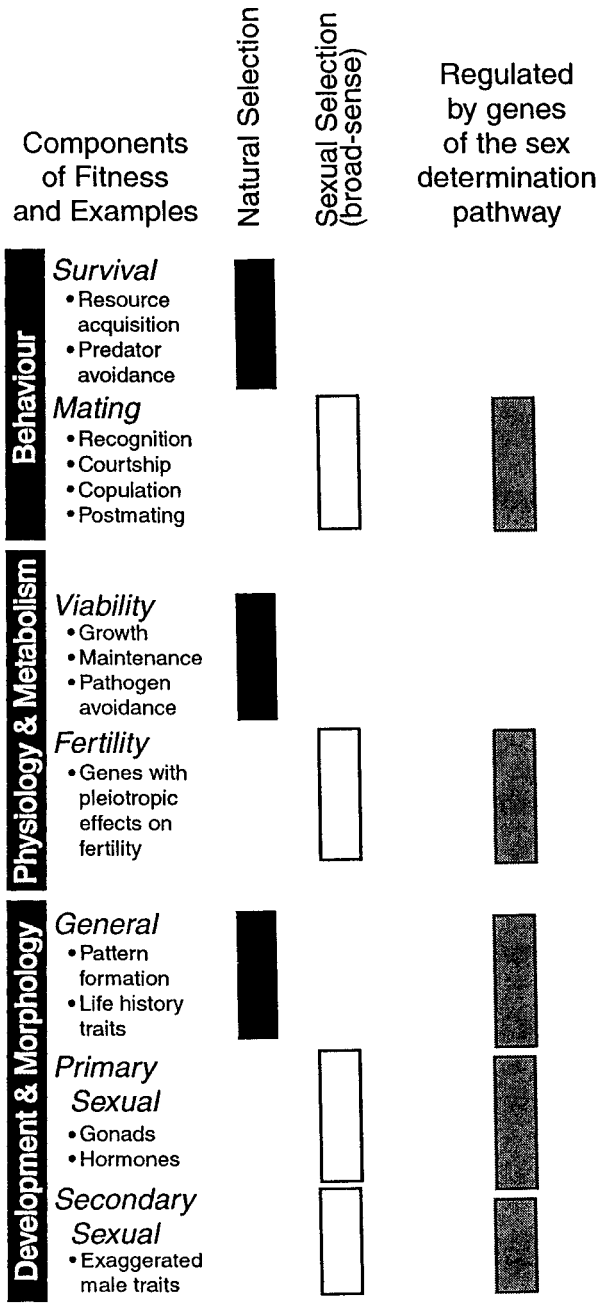
The consistent appearance of novelty in sexual systems indicates that these systems may contain large yet concealed amounts of genetic variation and furthermore, such systems can tolerate large genetic perturbations. While alterations in protein structure may produce no effect on the organism's fitness at their time of introduction, the relative fitnesses of these protein changes may vary over time in a context-specific manner - a possibility which will be further discussed below (section 5.4).

5.23 Sexual selection and rapid evolution

The rapid evolution of SRR genes and traits and discovery of novel sexual genes/traits demonstrate the diversity of sexual systems. A major factor in both generating and operating this observed diversity has been sexual selection. Over the last decade, a broadened definition of sexual selection - one that encompasses selection over all sexual traits (Civetta and Singh 1998) - has greatly advanced our understanding of the evolution of sexual systems. This new definition allows mechanisms of sexual selection

to be extended beyond classical examples of female choice and exaggerated male phenotypes which were limited to male secondary sexual traits and extreme male behaviour (Darwin 1871). Such an emphasis on classic male features can be seen in early models (i.e. runaway selection) which attempted to explain Darwin's original paradoxical observation that maladaptive traits evolve in order to increase male fitness (Fisher 1930; Lande 1981). Later theories emphasized the benefits of female discrimination such as direct resource-based advantages (Maynard Smith 1991) and the indirect appraisal of genes by the good-genes model (Andersson 1994). This broadening of sexual selection theory allowed its mechanisms to be extended to mate guarding, gift giving and a whole range of other mating trickery. The generality of sexual selection mechanisms in a whole range of sex-related traits (Figure 5.1) has identified it as an important force of evolutionary change.

While both *tra* and *tra-2* are found to be rapidly evolving between species of the *D. melanogaster* subgroup, only *tra-2* demonstrates patterns of nucleotide variation consistent with an adaptive model. The absence of replacement (but not synonymous) polymorphisms at the *tra-2* locus supports the rapid fixation of amino acids using a broadened definition of sexual selection. Since *tra-2* has multiple sexual functions in various tissues, the precise selective agent(s) which may have directed its evolution will be difficult to elucidate. However, there are a number of reasons why it may be surmised that selection did not act on the *primary* sex determination pathway (i.e. Figure 1.1). First, *tra-2*'s molecular partner in the sex determination pathway, *tra*, did not reveal any selective signatures. If primary sex determination was indeed a target of positive


Components of Fitness and Examples
Natural Selection
Sexual Selection (broad-sense)
Regulated by genes of the sex determination pathway
Behaviour
Survival
• Resource acquisition
• Predator avoidance
Mating
• Recognition
• Courtship
• Copulation
• Postmating
Physiology & Metabolism
Viability
• Growth
• Maintenance
• Pathogen avoidance
Fertility
• Genes with pleiotropic effects on fertility
Development & Morphology
General
• Pattern formation
• Life history traits
Primary Sexual
• Gonads
• Hormones
Secondary Sexual
• Exaggerated male traits

selection, one would expect that *tra* too would be a target and show evidence of selection. *dsx*, *tra-2*'s downstream target, also did not show any signs of selection and is one of the most conserved loci in the *D. melanogaster* complex. Second, no major changes in the three major domains of the *tra-2* locus were observed between species. The RS domains, RS1 and RS2, remained consistent in placement and size relative to the rapidly evolving RS domains of TRA while the RNA recognition domain was invariant. In fact, transgenic rescues of *tra-2* primary sex determination function were demonstrated using a *D. virilis tra-2* homolog (Chandler *et al.* 1997) as well as human and mouse homologs (Dauwalder *et al.* 1996). The third reason is that there is no other convincing evidence of adaptive mechanisms in primary sex determination loci (see Marín and Baker 1998). An alternative explanation employs selection on other functions of *tra-2*, namely its role in spermatogenesis (Mattox and Baker 1991; Hazelrigg and Tu 1994). For this particular role, there exists a precedent of a rapidly evolving regulatory gene driven by positive selection. In their search for a "speciation gene", Ting *et al.* (1998) isolated *OdsH*, a homeobox protein, which had rapidly diverged in a very short period of time in the *D. mauritiana* lineage. Not only did this gene reveal large number of replacement substitutions relative to synonymous substitutions, but its introgression into a sibling species' genetic background resulted in decreased fertility. Ting *et al.* (1998) suggested that sexual selection had driven the rapid divergence of *OdsH*. Sexual selection on male fertility may be driving *tra-2* divergence as well.

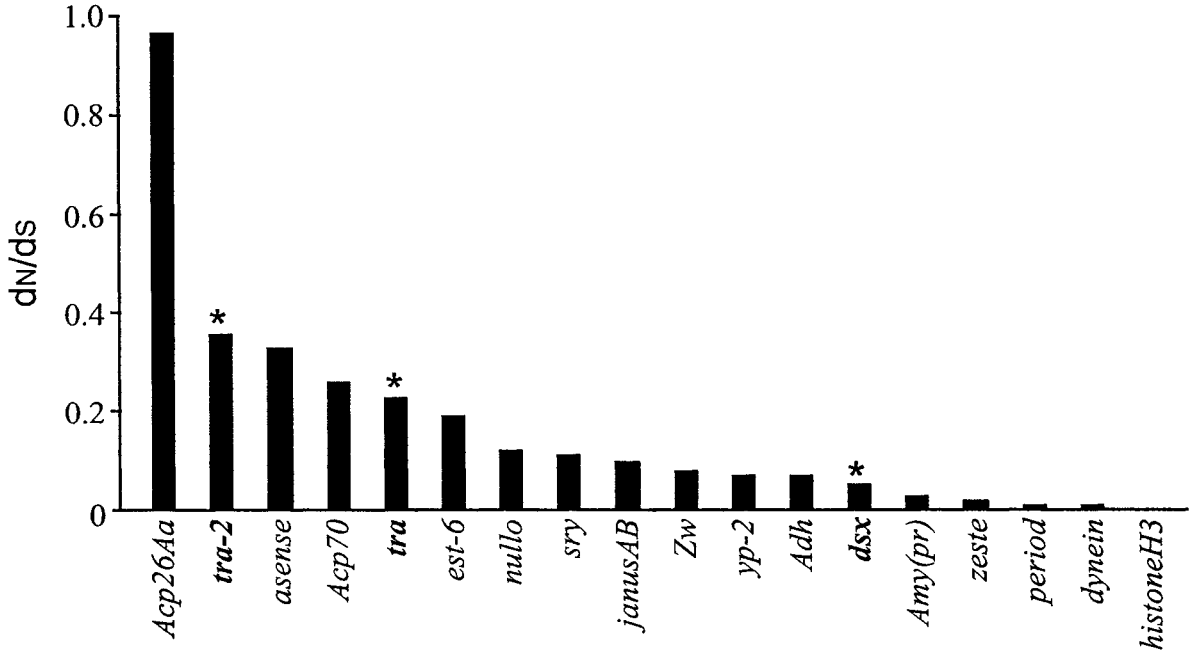
5.3 High Variation Among Genes that Determine Sex

5.31 Tolerance of sex determination to genetic perturbations

The diversity of sex determination mechanisms is a well known evolutionary pattern (Hodgkin 1992; Marín and Baker 1998) and reveals the amenability of this important developmental system to large genetic alterations, particularly in its upstream regulators (Wilkins 1995). The evolution of *tra* in species of the *D. melanogaster* complex best exemplifies this tolerance to heritable change. In addition to the high proportion of nonsynonymous substitutions present in *tra*, large insertions were found in the coding region. A fixed 13 amino acid insertion appears in *D. melanogaster* while a larger 74 amino acid insertion was fixed in *D. sechellia*. This latter insertion represents a 30% increase in the number of amino acids and includes putatively important protein binding domains. Hence, *tra* function can tolerate drastic changes in protein biochemical structure which directly translates to the presence of large differences in genetic structure.

A measurement of a protein's functional constraint, d_N/d_S, was calculated and compared between loci. Among species of the *D. melanogaster* subgroup, *tra* and *tra-2* were found to be among the least constrained genes in terms of overall d_N/d_S (Figure 5.2). (Since high d_N/d_S may also indicate that a significant portion of the gene is under positive selection, this hypothesis was also tested and subsequently rejected.) It is noteworthy that other loci with low selective constraints are also sex-related. The male-specific accessory gland proteins, Acp26Aa and Acp70 (Appendix A), are thought to assist sperm in the fertilization process (Bertram *et al.* 1996) and have demonstrated positive selection (Tsaur and Wu 1997; Cirera and Aguadé 1997).

Experimental evidence also supports the tolerance of sex determination to large



Loci with sequences available from all four sibling species of the *D. melanogaster* complex

genetic changes. The transgenic experiment of Chandler *et al.* (1997) demonstrate that the practically unalignable *D. virilis* ortholog of *tra-2* can rescue *D. melanogaster* (null *tra-2* genetic background) back to normal, including normal male fertility. Remarkably, a human *tra-2* homolog, *htra-2α*, has also been demonstrated to rescue Drosophila sex determination in a null *tra-2* mutant (Dauwalder *et al.* 1996) but male fertility was not rescued. These sometimes extreme demonstrations reveal the relative ease of change that occurs in genes of the sex determination pathway. However, subtle differences on other functions affected by changes in these principle loci must also be investigated. For example, O'Neil and Beloté's (1992) *tra* rescue (see section 5.32) did not *fully* revert femaleness indicating pleiotropic (or epistatic) effects caused by changes in TRA structure. The absence of *tra-2* reversion of male fertility in mammalian transgenic *D. melanogaster tra-2* mutants (Dauwalder *et al.* 1996) also suggest that while sex determination may be rescued, other functions are not.

The tolerance of sex determining loci to large changes in protein structure also suggests that sex determination may be sufficiently flexible for the recruitment of new genes. If large genetic perturbations do not aversely affect the ability to choose sex, than it is not difficult to imagine that novel epistatic interations may be more prone to evolve. This suggestion conforms to Wilkin's (1995) hypothesis that sex determination evolves rapidly in response to the recruitment of novel upstream regulators. This feature of variable sexual systems, in general, may be important for the generation of evolutionary novelty and may play a large role in the production of phenotypic diversity.

5.32 Effect of rapid evolutionary change on downstream targets

Since *tra*'s high divergence is most likely due to lower selective constraints, an adaptive explanation is not likely. However, *tra* does hold a pivotal role in the regulation of Drosophila sex determination and its rapid evolution may indirectly effect the regulation of downstream genes involved in sexual differentiation (Figure 5.1). One way that *tra* could effect regulatory change is by altering the concentrations of its arginine-serine (RS) domains. As part of a protein family containing RS domains, TRA is involved in various aspects of spliceosome assembly and the regulation of alternative splicing (Fu 1991). RS domains are thought to help bridge 5' and 3' splice sites in *pre*mRNA transcripts by interacting with other SR (serine-arginine) proteins. In Drosophila (female) sex determination, TRA interacts with another SR protein, TRA-2, to form part of the spliceosome complex. If the presence of RS domains for spliceosome recruitment represents the only important element of the protein, it is not surprising that TRA has undergone high rates of neutral evolution. Domain-swap experiments have demonstrated the exchangeability of TRA RS domains by *suppressor-of-white-apricot* RS domains (Li and Bingham 1991). In another experiment, SXL, which acts as a splicing suppressor in Drosophila sex determination, was transformed into a splicing activator by the addition of RS domains (Valcárcel *et al.* 1993). While RS domain placement, the concentration of RS amino acid dipeptides per domain, and number of RS–rich regions differ between related species of Drosophila (especially as demonstrated in the *D. melanogaster* complex; see Figure 3.4), there is evidence that a minimum proportion of the protein is maintained to consist of RS domains. The large fixed

insertions of *D. melanogaster* and *D. sechellia* contain RS domains which maintain this minimal proportion. At this stage, we would require many more sequences from other species (and perhaps a structural protein map of TRA) to even begin to understand what proportion of RS dipeptides is required for normal sex determination. Such minimal requirements may allow a large portion of *tra* to be functionally unconstrained and amenable to rapid evolutionary change. More distant lineages may find themselves under different selective constraints depending on the affinity of their RS domains to regulate splicing. Rooney *et al.* (2000) demonstrated a similar phenomenon among primate protamines whereby the proportion of arginine residues, important for DNA binding, remains conserved across distant taxa.

Sexual dimorphism is a consequence of the sex determination genetic hierarchy. Genes involved in dimorphic characters such as gonadal morphology, mating behaviour and other aspects of sex (Civetta and Singh 1995, 1998), have been shown to be more rapidly evolving than non-sex genes indicating their preferential involvement in species differences and speciation (Singh and Kulathinal 2000). Recently, Kopp *et al.* (2000), showed that *bric-a-brac* (*bab*), a gene that modulates signals from both the sex determination and homeotic pathways (i.e. embryonic patterning), has evolved an effect on sex-specific abdominal pigmentation in the *D. melanogaster* lineage, important for mate recognition. They further suggest the evolution of *bab* regulation may have been driven by sexual selection. Since TRA is involved in the regulation of somatic sexual differentiation in females by binding (with TRA2) to regulatory elements in *doublesex* and *fruitless* (Hoshijima *et al.* 1991; Steinman-Zwicky 1994), it indirectly controls

aspects of female sexual differentiation, including all somatic and behavioural components (i.e. Ferveur *et al.* 1997; Arthur *et al.* 1998; Sylvain *et al.* 2000). Thus, changes in the *tra* protein, particularly in RS domain affinity, may indirectly contribute to various aspects of sexual differentiation such as pheromone and accessory gland production, mating behaviour, and secondary sexual characteristics such as abdominal pigmentation and bristle number (see Figure 5.1). In a transgenic experiment with *tra*, O'Neil and Beloté (1992) transferred the wild type *tra* gene of *D. virilis* to *D. melanogaster* by P element-mediated germline transformation. The *D. virilis* gene was capable of shifting male structures in *D. melanogaster* flies, which were chromosomally female but homozygous for a *tra* deletion, towards femaleness Whether the absence of a full recovery to femaleness was caused by positional effects of the transgenes or a difference in the evolution of different RS domain specificity between both species, remains an open question. The elucidation of *tra*'s role in the regulation of sexual dimorphic characters among different species of Drosophila as well as the specificity of particular domains in different species, will be an important step in understanding the rapid divergence of genes involved in development and reproduction.

5.4 Sexual System Variation and Speciation

The evolution of developmental systems presents a fascinating glimpse into how their change may effect phenotypic diversity. The rapid evolution of the two sex determining transcription factors, *tra* and *tra-2*, indicate that this particular sexual system is highly variable and amenable to large genetic changes. Generally, high variability has

been found among sexual systems (see section 5.21). One important consequence of highly variable sexual systems is that they may serve as depots of genetic variation which may allow for the introduction of evolutionary novelties and eventually adaptive mechanisms of evolutionary change. Such flexible genetic systems increase the probability that nonlethal genetic mutations accumulate, thereby increasing the appearance of phenotypic innovations.

The results of this thesis also demonstrate the rapid evolution of key regulators of a critically important sexual system. As principle control genes of sex determination, *tra* and *tra-2* play a key role in somatic female sexual differentiation and male gametogenesis in Drosophila. Various components of fitness, under which both natural and sexual selection have been postulated to act, may directly be affected by genetic changes at these loci (Figure 5.1).

The high degree of genetic variability found among sex determining loci may be attributed to selectively neutral changes. But the classification of this variation as neutral does not imply that these alterations are functionless. They simply represent alternative forms that are (nearly) equally fit or acceptable, in terms of survival and reproduction of the organism (Kimura 1968, 1983). This important qualifier of neutral theory differs from previous misconceptions which suggest that amino acid substitutions that are absolutely impartial to the action of natural selection can be considered "genetic junk" (i.e. Lewontin 1974, pg. 197). This latter term represents a misnomer which prevents us from further understanding the potential importance of transient neutral polymorphisms as a source of heritable genetic material. As Motoo Kimura, founder of the neutral theory

of evolution, stated,

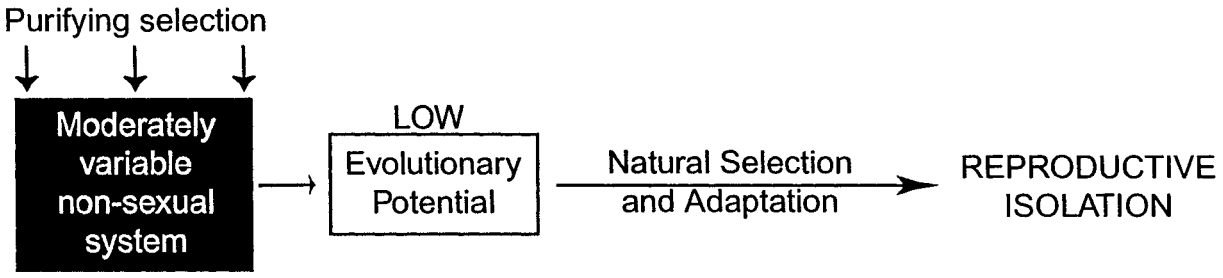
"We should not overlook the possibility that some of the 'neutral' alleles may become advantageous under an appropriate environmental condition or a different genetic background; thus, neutral mutants have a latent potential for selection. This means that polymorphic molecular mutants, even if selectively neutral under prevailing conditions of a species, can be the raw material for future adaptive evolution. To regard random fixation of neutral mutants as 'evolutionary noise' is inappropriate and misleading."
(Kimura 1983; pg. xiii)

The intrinsically variable nature of sex determination may allow the opting in of new genes (or conversely, the opting out) as part of the sex determination pathway. The diversity of sex determination mechanisms may reflect this particular feature. While downstream genes remain conserved in order to maintain primary function, the upstream portion of sex determining pathways seem to be the most variable (Wilkins 1995). Another consequence of the variable nature of sex determining genes is the adaptive evolution of *tra-2*. While *tra-2* may have evolved in a similar fashion to *tra* (i.e. rapidly evolving due to low functional constraints), certain advantageous novelty could have arisen and been selected in another function (i.e. spermatogenesis). Hence, the variability in sex differentiation may have contributed to the adaptation of other sexual traits.

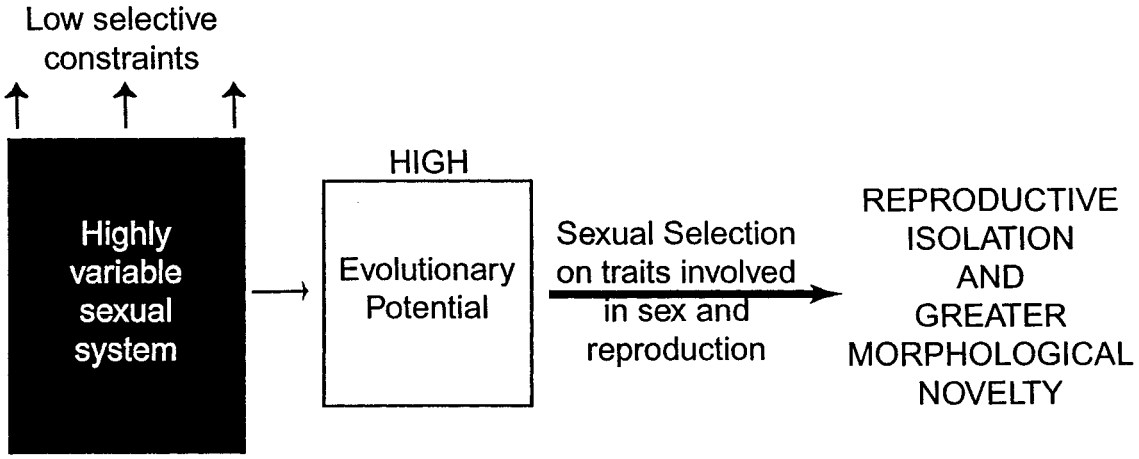
The presence of greater genetic variation in sexual systems offer a variety of important consequences for the development of phenotypic diversity. Less selective constraints on genes involved in sex and reproduction will generate a larger pool of genetic variation. And the addition of other selective devices, such as sexual selection, may drive the rapid divergence of sexual traits (Figure 5.3). The rapid fixation of alleles between species, whether caused by selective or neutral mechanisms, is an important factor in reproductive isolation. According to the Dobzhansky-Muller incompatibility

Figure 5.3 Modelling speciation. A. Classic adaptation model of speciation. Natural selection acts upon the existing pool of genetic variation in the population. Adaptation slowly drives population divergence. Reproductive isolation results as a byproduct of a phenotypic character shift. B. Sexual selection utilizes the extensive genetic variation from a large pool of sex and reproduction-related genes. The intrinsically larger levels of variation found in this system allows for the rapid evolution of phenotypic novelty. Rapid selection on these sexual traits results in the rapid evolution of isolating mechanisms. Magnitude of figure components indicate relative significance in both models.

A. Adaptation via Natural Selection



B. Selection on Variable Sexual System



model of speciation (Dobzhansky 1937; Muller 1942), any acceleration of the fixation process within populations will accelerate the production of incompatibilities in the hybrid. This model demonstrates the importance of variable sexual systems in the generation of organic diversity at both micro- and macroevolutionary levels.

5.5 Conclusions

The molecular evolutionary study of the three regulatory genes controlling Drosophila sexual differentiation, *transformer*, *transformer-2*, and *doublesex*, offers valuable insight into the evolutionary dynamics of an important sexual system. In this thesis, *tra* and *tra-2* were both found to be among the most rapidly evolving genes in the *D. melanogaster* subgroup, in terms of fixed nonsynonomous substitutions. Heterogeneity in selective constraints between different lineages at the *tra* locus as well as the *dsx* locus also reveal the variability of selective pressures that act upon these genes. Furthermore, the presence of large insertions in the *tra* coding region among sibling species and the high d_N/d_S ratios found outside the conserved RRM domains in *tra-2* indicate that the genetic architecture of sex determination can accommodate large allelic changes, which in turn, may become rapidly fixed by sexual selection.

One of the major problems that Charles Darwin faced with his theory of natural selection concerned the transitional states through which complex characters evolved. While Darwin sufficiently answered such difficulties with clever morphological and anatomical arguments, a more up-to-date molecular explanation can be formulated. Depending on the flexibility of a particular genetic system to genetic perturbations,

transitional novelties may be readily produced. In highly variable genetic systems, there may be an increased probability that viable evolutionary innovations will be generated and subsequently fixed. Organic complexities may become the eventual outcome of repeated episodes of this process. The high degree of genetic variability observed in Drosophila sex determination loci, as indicated by the results of this thesis, suggest that this sexual system may contain a high capacity for evolutionary change. In conjunction with sexual selection, the rapid evolution of genes and traits constituting variable sexual systems is argued to be an important generator of the vast diversity of life.

REFERENCES

Aguadé, M., Miyashita, N., and C.H. Langley (1992) Polymorphism and divergence in the *Mst26A* male accessory gland gene region in *Drosophila*. Genetics **132:** 755-770

Andersson, M.B. (1994) *Sexual selection.* Princeton University Press, Princeton, New Jersey

Arthur, B.I., Jallon, J.M., Caflisch, B., Choffat, Y., and R. Nöthiger (1998) Sexual behaviour in Drosophila is irreversibly programmed during a critical period. Curr. Biol. **8:** 1187-1190

Baker, B.S. (1989) Sex in flies: The splice of life. Nature **340:** 521-524

Begun, D.J., Whitley, P., Todd, B.L., Waldrip-Dail, H.M., and A.G. Clark (2000) Molecular population genetics of male accessory gland proteins in Drosophila. Genetics **156:** 1879-1888

Bertram, M.J., Neubaum, D.M., and M.F. Wolfner (1996) Localization of the Drosophila male accessory gland protein Acp36DE in the mated female suggests a role in sperm storage. Insect Biochem. Mol. Biol. **26:** 971-980

Birkhead, T.R., and A.P. Möller (1998) *Sperm competition and sexual selection.* Academic Press, San Diego, CA

Carlson, S.S., G.A. Mross, A.C. Wilson, R.T. Mead, L.D. Wolin, S.F. Bowers, N.T. Foley, A.O. Muijsers, and E. Margoliash (1977) Primary structure of mouse, rat, and guinea pig cytochrome c. Biochemisty **16:** 1437-1442

Carson, H.L. (1997) Sexual selection: a driver of genetic change in Hawaiian *Drosophila*. J. Hered. **88:** 343-352

Chandler, D., McGuffin, M.E., Piskur, J., Yao, J., Baker, B.S., and W. Mattox (1997) Evolutionary conservation of regulatory strategies for the sex determination factor *transformer-2*. Mol. Cell. Biol. **17:** 2908-2919

Cirera, S., and M. Aguadé (1997) Evolutionary history of the sex-peptide (*Acp70A*) gene region in *Drosophila melanogaster*. Genetics **147:** 189-197

Civetta, A., and R.S. Singh (1995) High divergence of reproductive tract proteins and

their association with postzygotic reproductive isolation in *Drosophila melanogaster* and *Drosophila virilis* group species. J. Mol. Evol. **41:** 1085-1095

Civetta, A., and R.S. Singh (1998) Sex-related genes, directional sexual selection, and speciation. Mol. Biol. Evol. **15:** 901-909

Clark, A.G., Aguadé, M., Prout, T., Harshman, L.G. and C.H. Langley (1995) Variation in sperm displacement and its association with accessory gland protein loci in *Drosophila melanogaster*. Genetics **139**, 189-201

Cline, T.W. (1985) Primary events in the determination of sex in *Drosophila melanogaster*. In *The origin and evolution of sex*. (Ed: H.O. Halvorson, and A. Monroy), pp. 301-327, Liss Publishing, New York

Cline, T.W., and B.J. Meyer (1996) Vive la différence: Males *vs.* females in flies *vs.* worms. Annu. Rev. Genet. **30:** 637-702

Coschigano, K.T., and P.C. Wensink (1993) Sex-specific transcriptional regulation by the male and female *doublesex* proteins of Drosophila. Genes Dev. **7:** 42-54

Coulthart, M.B. and R.S. Singh (1988) High level of divergence of male-reproductive-tract proteins between *Drosophila melanogaster* and its sibling species, *D. simulans*. Mol. Biol. Evol. **5**, 182-191

Darwin, C. (1871) *The descent of man, and selection in relation to sex*. John Murray, London

Dauwalder, B., Amaya-Manzanares, F., and W. Mattox (1996) A human homologue of the Drosophila sex determination factor *transformer-2* has conserved splicing regulatory functions. Proc. Natl. Acad. Sci. U.S.A. **93:** 9004-9009

Dobzhansky, Th. (1937) *Genetics and the origin of species*. Columbia University Press, New York

Dübendorfer, A., Hilfiker-Kleiner, D., and R. Nöthiger (1992) Sex determination mechanisms in dipteran insects: The case of *Musca domestica*. Semin. Dev. Biol. **3:** 349-356

Duret, L. and D. Mouchiroud (2000) Determinants of substitution rates in mammalian genes: Expression pattern affects selection intensity but not mutation rate. Mol. Biol. Evol. **17**, 68-74

Eberhard, W.G. (1985) *Sexual selection and animal genitalia*. Harvard University Press,

Cambridge, MA

Ferveur, J.-F., Savarit, F., O'Kane, C.J., Sureau, G., Greenspan, R.J., and J.-M. Jallon (1997) Genetic feminization of pheromones and its behavioural consequences in Drosophila males. Science **276:** 1555-1558

Fisher, R.A. (1930) *The genetical theory of natural selection.* Clarendon, Oxford

Fu, X.-D. (1991) The superfamily of arginine/serine-rich splicing factors. RNA **1:** 663-680

Hazelrigg T, and C. Tu (1994) Sex-specific processing of the Drosophila *exuperantia* transcript is regulated in male germ cells by the *tra-2* gene. Proc. Natl. Acad. Sci. U.S.A. **91:** 10752-10756

Hellberg, M.E., and V.D. Vacquier (1999) Rapid evolution of fertilization selectivity and lysin cDNA sequences in teguline gastropods. Mol. Biol. Evol.**16:** 839-848

Hodgkin, J. (1992) Genetic sex determination mechanisms and evolution. BioEssays **14:** 253-256

Hoshijima, K., Inoue, K., Higuchi, I., Sakamoto, H., and Y. Shimura (1991) Control of *doublesex* alternative splicing by *transformer* and *transformer-2* in Drosophila. Science **252:** 833-836

Joly, D., Bressac, C., and D. Lachaise (1995) Disentangling giant sperm. Nature **377:** 202

Karr, T.L., and S. Pitnick (1996) The ins and outs of fertilization. Nature **379:** 405-406

Kimura, M. (1968) Evolutionary rate at the molecular level. Nature **217:** 624-626

Kimura, M. (1983) *The neutral theory of molecular evolution.* Cambridge University Press, New York

Kopp, A., Duncan, I., Godt, D., and S.B. Carroll (2000) Genetic control and evolution of sexually dimorphic characters in Drosophila. Nature **408:** 553-559.

Kulathinal, R.J. and Singh, R.S. (1998) Cytological characterization of premeiotic versus postmeiotic defects producing hybrid male sterility among sibling species of the *Drosophila melanogaste*r complex. Evolution **52:** 1067-1079

Lachaise, D., David, J.R., Lemeunier, F., Tsacas, L., and M. Ashburner (1986) The

reproductive relationships of *Drosophila sechellia* with *D. mauritiana*, *D. simulans*, and *D. melanogaster* from the Afrotropical region. Evolution **40:** 262-271

Lande, R. (1981) Models of speciation by sexual selection on polygenic traits. Proc. Natl. Acad. Sci. U.S.A. **78**, 3721-3725.

Lee, Y.H., Ota, T., and V.D. Vacquier (1995) Positive selection is a general phenomenon in the evolution of abalone sperm lysin. Mol. Biol. Evol. **12:** 231-238

Lewontin, R.C. (1974) *The genetic basis of evolutionary change*. Columbia University Press, New York

Li, H., and P.M. Bingham (1991) Arginine/serine-rich domains of the *su(wa)* and *tra* RNA processing regulators target proteins to a subnuclear compartment implicated in splicing. Cell **67:** 335-342

Long, M., Wang, W., and J. Zhang (1995) Origin of new genes and source for N-terminal domain of the chimerical gene, *jingwei*, in Drosophila. Gene **238:** 135-141

Marín, I., and B.S. Baker (1998) The evolutionary dynamics of sex determination. Science **281:** 1990-1994

Mattox, W., and B.S. Baker (1991) Autoregulation of the splicing of transcripts from the *transformer-2* gene of Drosophila. Genes Dev. **5:** 786-796

Mayfield, J.A., Fiebig, A., Johnstone, S.E., and D. Preuss (2001) Gene families from the *Arabidopsis thaliana* pollen coat proteome. Science **292:** 2482-2485

Maynard Smith, J. (1991) Theories of sexual selection. TREE **6:** 146-151

Metz, E.C., and S.R. Palumbi (1996) Positive selection and sequence rearrangements generate extensive polymorphism in the gamete recognition protein bindin. Mol. Biol. Evol. **13:** 397-406

Muller, H.J. (1942) Isolating mechanisms, evolution, and temperature. Biological Symposium **6:** 71-125

Nagoshi, R.N., B.S. Baker (1990) 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:** 89-97

Nurminsky, D.I., Nurminskaya, M.V., DeAguiar, D. and D.L. Hartl (1998) Selective sweep of a newly evolved sperm-specific gene in Drosophila. Nature **396**: 572-575

O'Neil, M.T., and J.M. Beloté (1992) Interspecific comparison of the *transformer* gene of Drosophila reveals an unusually high degree of evolutionary divergence. Genetics **131:** 113-128

Parsch, J., Meiklejohn, C.D., Hauschteck-Jungen, E., Hunziker, P., and D.L. Hartl (2001) Molecular evolution of the *ocnus* and *janus* genes in the *Drosophila melanogaster* species subgroup. Mol. Biol. Evol. **18:** 801-811

Podos, J. (2001) Correlated evolution of morphology and vocal signal structure in Darwin's finches. Nature **409:** 185-188

Raymond, C.S., Shamu, C.E., Shen, M.M., Siefert, K.J., Hirsch, B., Hodgkin, J., and D. Zarkower (1998) Evidence for evolutionary conservation of sex-determining genes. Nature **391:** 691-695

Ritchie, M.G., and C.P. Kyriacou (1994) Reproductive isolation and the *period* gene of Drosophila. Mol. Ecol. **3:** 595-599

Rooney, A.P., Zhang, J. and M. Nei (2000) An unusual form of purifying selection in a sperm protein. Mol. Biol. Evol. **17:** 278-283

Singh, R.S., and R.J. Kulathinal (2000) Sex gene pool evolution and speciation: A new paradigm. Genes. Genet. Syst. **75:** 119-130

Smith, C.A., McClive, P.J., Western, P.S., Reed, K.J., and A.H. Sinclair (1999) Conservation of a sex-determining gene. Nature **402:** 601-602

Smith, R.L. (1984) *Sperm competition and the evolution of animal mating systems.* Academic Press, New York.

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

Sylvain, G., Jean-Francois, G., and F. Jean-Rene (2000) Genetic identification of neurons controlling a sexually dimorphic behaviour. Curr. Biol. **10:** 667-670

Swanson, W.J., and V.D Vacquier (1998) Concerted evolution in an egg receptor for a rapidly evolving abalone sperm protein. Science **281:** 710-712

Swanson, W.J., Yang, Z., Wolfner, M.F., and C.F. Aquadro (2001a) Positive Darwinian selection drives the evolution of several female reproductive proteins in mammals. Proc. Natl. Acad. Sci. U.S.A.. **98:** 2509-2514

Swanson, W.J., Clark, A.G., Waldrip-Dail, H.M., Wolfner, M.F., and C.F. Aquadro (2001b) Evolutionary EST analysis identifies rapidly evolving male reproductive proteins in Drosophila. Proc. Natl. Acad. Sci. U.S.A. **98:** 7375-7379

Thomas, S., and R.S. Singh (1992) A comprehensive study of genic variation in natural populations of *Drosophila melanogaster*. VII. Varying rates of genic divergence as revealed by two-dimensional electrophoresis. Mol. Biol. Evol. **9:** 507-525

Ting, C.-T., Tsaur, S.-C., Wu, M.-L. and C.-I. Wu (1998) A rapidly evolving homeobox at the site of a hybrid sterility gene. Science **282**, 1501-1504

Tsaur, S.-C., and C.-I. Wu (1997) Positive selection and the molecular evolution of a gene of male reproduction, *Acp26Aa* of *Drosophila*. Mol. Biol. Evol. **14:** 544-549

Valcárcel, J., Singh, R., Zamore, P.D., and M.R. Green (1993) The protein Sex-lethal antagonizes the splicing factor U2AF to regulate alternative splicing of *transformer* pre-mRNA. Nature **362:** 171-175

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

Wolfner, M.F. (1997) Tokens of love: functions and regulation of Drosophila male accessory gland products. Insect. Biochem. Mol. Biol. **27:** 179-192

Wyckoff, G.J., Wang, W., and C.-I. Wu (2000) Rapid evolution of male reproductive genes in the descent of man. Nature **403:** 304-309

APPENDICES

APPENDIX A

Likelihood ratio tests and d_N/d_S estimates of loci from all four species of the *D. melanogaster* subgroup

Gene	GenBank Accession Number / Strain				L_{codon}	$2\Delta\ell$				Estimate of $\omega=d_N/d_S$ under null model (d_N, d_S)[b]
	D. mel	*D. sim*	*D. mau*	*D. sec*		Constant *vs.* Variable rate of substitut'n	[a]Constant *vs.* Variable d_N/d_S among sites	Clock *vs.* No clock	Constant *vs.* Variable d_N/d_S among lineages	
Acp26Aa	X70891	X70899	X70898	X72630	247	**19.9****	**14.2****	1.96	1.89	0.97 (0.188, 0.165)
Acp70	X99415	X99417	X99412	X99414	55	0.00	0.00	4.34	3.11	0.26 (0.034, 0.129)
Adh	M36580	M36581	M19264	X04672	256	2.08	0.00	4.52	6.84	0.07 (0.005, 0.085)
Amy(pr)	L22733	D17734	D17730	D17732	494	**7.86***	**11.0***	1.91	**7.87***	0.03 (0.012, 0.281)
asense	X52892	Hilton *et al.*(1994)			352	3.86	2.98	4.70	0.27	0.33 (0.015, 0.047)
dsx	M25292	0251.2	0241.2	0248.3	**439**	2.69	3.72	**10.4****	1.49	0.05 (0.005, 0.121)
dynein	AF136252	AF136263	AF136265	AF136264	543	1.98	2.17	**8.87***	1.20	0.01 (0.003, 0.118)
est6	M15961	L34264	AF284482	AF284489	489	**8.23***	**18.5****	5.38	1.91	0.19 (0.025, 0.146)
histoneH3	AB019400	AB019401	AB019403	AB019402	135	0.00	0.00	0.00	0.00	0.00 (0.002, 0.104)
janusAB	M27033	AF284453	AF284457	AF284459	171	2.66	2.25	2.39	5.33	0.10 (0.016, 0.122)
nullo	X65444	U44733	U64710	U64711	173	0.00	0.00	0.50	5.23	0.12 (0.020, 0.131)
period	L07819	L07826	L07816	L07820	558	3.22	0.00	1.10	1.10	0.01 (0.002, 0.259)
sry	X03121	U64718	U64715	U64717	493	**8.11***	**25.8****	3.79	0.84	0.11 (0.013, 0.135)
tra	penn2	colom	0241.3	0248.0	**196**	**13.9****	3.47	0.29	1.58	0.23 (0.042, 0.134)
tra-2	M30939	S-132	0241.5	0248.0	**222**	1.09	8.42	5.64	**14.4****	0.36 (0.039, 0.147)
yp2	L14423	L14428	L14418	L14424	348	0.00	0.00	5.56	0.62	0.07 (0.007, 0.101)
zeste	L13044	L13049	L13059	L13062	264	**9.21****	0.00	0.66	2.14	0.02 (0.003, 0.126)
Zw	Ah002543	AF148150	AF284490	AF284496	388	**19.0****	**22.8****	**11.1****	**7.27***	0.08 (0.032, 0.133)

Likelihood ratio tests and d_N/d_S estimates of loci from all four species of the *D. melanogaster* subgroup. [a]M0 ($\omega_{constant}$) vs. M3 ($\omega_{1,2,3}$), as denoted in Yang and Neilson (1998). [b]d_N, d_S was calculated using the method of Nei and Gojobori (1986) between *D. melanogaster* and each of its three sibling species and averaged. A sim/sec/mau trichotomy was modelled in this analysis. The sex determining genes, *dsx*, *tra*, and *tra-2*, are indicated in boxes and divergence estimates are boldfaced. Significant values for the likelihood ratio tests are in boldface, $*P<0.05$, $**P<0.01$.

APPENDIX B1

Variable nucleotide sites at the *doublesex* locus in the *D. melanogaster* subgroup

Region A (common exon 2)

```
                1 1111111111 2222222222 2233333333 4444444444 4444555555 5666666666
       2234466790 1113455679 0334667999 9903556689 2466667788 8899134578 8000123455
       1702569505 1470403215 1179173145 7890473519 6925892313 4602013530 2039220015

mel-publ  CCACCGAGCA CCACCCCGAG TGTGCGTGCC CCCTCCCCCC CCGTCTAGGT TCACGTCGCT GCGCCCGACC
mel-02310 .......... .......... .......... .......... .......... .......... ..........
mel-02311 .......... .......... .......... .......... .......... .......... ..........
mel-02316 .......... .......... .......... .......... .......... .......... ..........
mel-cpa46 .......... .......... .......... .......... .......... ...T...... ..........
mel-z12   .......... .......... .......... .......... .......... .......... ..........
mel-z34   .......... .......... .......... .......... .......... .......... ..........
simulans  ......GC.G ..G..... C..A..... GTT.T..... A.T..... .T.T.C.... ..........
mauritiana......GCTG ..G.T....A CC.A.A.... GTT.T.T.T. A.T..... .T.T.C.... ..........
sechellia ...A.AGC.G ..G..... CC.A.AC... GTT.T..... A.T.GC.... .T.T.C..T. ...T......
erecta    .A....GC.. T.C.....G. ..C....ATT G..CT..... ..TCG.TC.C A.C.AC.A.C A.C.......
teissieri ....T.GC.G TAGA..T.G. ..CA...ATT G..CTT.G.G .AT.G..C.C ..C..CT... .G-...ACTA
yakuba    A.T.T..C.G TAGA.TTAG. ..CAT..ATT G..CT..... ..T.G..CAC ..C..CT... .TA.TTA.T.
REPLMNT                                        R     R   RRRR  R R      R  R   R R

       6666666777 7777777777 7777788888 8888888888 899999
       7788999000 1244455677 7789900011 1122335667 922223
       5819034128 1312336102 4701514713 5828174149 114578

mel-publ  CTTTGTCCCC CAGGCCGAGG CTAGGCGTGG CGTGAAGTCG CGGCCC
mel-02310 .......... .......... ........A. .......... ......
mel-02311 .......... .......... ........A. .......... ......
mel-02316 .......... .......... .......... .......... ......
mel-cpa46 .......... .......... .......... .......... ......
mel-z12   .......... .......... ........A. .......... ......
mel-z34   .......... .......... .......... .......... ......
simulans  TCC..... ..TA...G.. T.....T..A .......... .....G
mauritianaT.C.A..... ..TA...GC. T.....T..A ..C..... .....G
sechellia TCC...G... ..TA..TG.. T.....T..A ..C..... ......
erecta    .CC....GT. ....T..G.A TA...T.A.A .C..CGCAT. AA....
teissieri GCG....G.T T....T.G.A TA.CTT.A.A ...ACG..TT .ATTA.
yakuba    .CGG.C.G.. TC...T.G.A TACC.T.A.A T..ACG..T. .AT.A.
REPLMNT   R  R  RR      RR  R    RRR    r  RR    RR      R R
```

Region B (common exon 3 to female exon 4)

```
          1111111 1111111111 1111112333 3333334444 4444444445 5555566666 6666666667
        240000111 1123333444 4556684223 3447780111 2345668991 4455500111 5667788992
       9753489123 4571239156 8890434795 7234986145 8985390172 4625889138 8683567391

mel-publ  TAAGTTCGTG ATTTCCCAAA CTGAGCAAGT CCCTTCTTTA GAGAAACAAA AATACCTATA AGAGTATTTT
simulans  .GG..C.... .......G.. ..AT..GT.. ...C...... .C...CT... GC......AT .CG..CG..C
mauritiana.GGC.C.... .......G.. .C....GT.. .......... CC...CT... .C....C.AT CCG..CG.C.
sechellia .GG..C.... .......G.. ..AT..GT.. ....C..... .C...CT..C .C......AT .CG..TG...
tiessieri CGG.C.AC.. TAGCTTT.CT A...AAG.AC .A-..TCG.G CCA.T..CC. .CACTAGCAT .CCTATGG..
yakuba    .GG.C.ACCT TAGCTT..CT A...A.G.AC TAA..T.GC. .C.T....C. .C.CT.GCAT .CCTATGG..
REPLMNT

                                                    1111 1111111111 1111111111 1111
          7777777777 7888888888 8888889999 9999990000 0000000000 0111111111 1111
          2344455777 8111112334 5569991233 5558890134 4566778899 9012223334 5669
          2625767246 3156898574 4931495556 3565675164 6314270645 8180563564 2493

mel-publ  CCGGAT-AAG ATATACGTGT CTAACTGCAT TCTATTAAAT GCTGAACCTT TCTGGC-ATT GACA
simulans  ..TCT.ATGC TC....A... .C........ .....C..G. A..ATG.A.A C.....CC.. .C..
mauritianaT.TCT.TTGC .C...GA... .....A.... ........G. A..ATG...A C.....CC.. .C..
sechellia T.TCTCTTGC .C....A... ...T...... ........G. A..ATG...A C.....CC.. .C..
tiessieri -ATC.C-..C ..TAC.ACAA T.G.T.AT.. .TCGC.T..A A.GAT..... CT..ATAGGA .CTT
yakuba    -.TC.C-..C ..TAC.ACA. T.G.T..AGG CTCGC.TTGA ATAAT.T.G. CTGAATAGG. ACT.
REPLMNT
```

APPENDIX B1 (continued)

Region C (Male exon 5)

```
                111111 1111111112 2222222222 2333333333 3333
           1689011233 4577799990 1244568899 9012334445 5789
           2676547628 7017924584 6606271715 6359692581 4510

mel-publ   GGCGATAAGC CCGTTGCCCG TCAACGC--A GTCCGAAGAG GCCC
mel-CanS   .......... .......... .......--. .......... ....
mel-IAB    .......... .......... .......--. .......... ....
mel-02310  .......... .......... .......--. .......... ....
mel-02311  .......... .T....TG.. .......--. .......... ....
mel-02316  .......... .T....TG.. .......--. .......... ....
mel-z12    ....C..... .T....TG.. .......--. .......... ....
mel-z34    ....C..... .T....TG.. .......--. .......... ....
simulans   ....C..... ..A.C..AG. ..G..C.CA. .G...CG... ....
mauritiana....C..... ..A.C...G. ..G..C.CA. .G...CG... ....
sechellia  ....C..... TTA.C...GT ..G..C.CA. .G...CG..A .G..
erecta     ....CCGG.. ...CC...G. CAGT.C.CA. .GA..C.A.. C...
tiesieri   ..T.C.G.AT ...CCA..G. CAG.GA.TG. AGTTAC.A.. C..T
yakuba     CATTC.G.AT ...CC...G. CAG..TTCGC AGT.AC.AG. C.TT
REPLACEMNT               R r     R    R R R  R
```

Variable sites are separated into the three regions of *dsx* sequenced in this study. Nucleotide positions are numbered relative to the following points of reference from Burtis and Baker (1988): Region A - nucleotide 1307; Region B - nucleotide 2354; Region C - nucleotide 2492^{m}. Sites from coding regions are highlighted in boldface. Within species variation was only assessed in Regions A and C. Replacement substitutions are denoted below as an **R**. Replacement polymorphisms are denoted in lowercase and underlined as r. Indels are not included as variable sites.

APPENDIX B2

Variable nucleotide sites of *transformer* in the *Drosophila melanogaster* subgroup

Nucleotide Position

```
                  111111111111111122222222222222222222222222222222233333333  355555555555566666666666666
            122333456666678911345566677778990111223444455566666777778889911233344  466677788899001122344445556
            21514601123468623908260490126649245636325892360139134790486818835701  278923514616381709224670173

mel.pub     TC0ACGCctaacgtccctggggg5acagaaatcacgctcactacgagctacg2atctaCGTCGTCAGG  0TCACGGCCG3GAGCACAATGGAATCG
mel.(x2)    ..0...A...............5.........................2..............  0.........3................
mel.india   ..0...................5.........................2..............  0.........3................
mel.(x9)    ..0...................5.........................2..............  0.........3................
mel.(x5)    ..0...................5.........................2..............  0.........3................
mel.italy   ..0...A...............5.........................2..............  0.........3................
mel.ST05    ..0...................5...............a.........2..............  0.........3................
sim.(x2)    C.0.G................c5.....g............gt.t...t.c2t.tct..GTAC....  0A.....T..3....G...AA..GC..
sim.flori   C.0.G................c5.....................gt.g...t.c2t.tct..GT.C....  0A.....T..3....G...AA..GC..
sim.colom   C.0.G................c5.....g......t.....gt.g...t.c2t.t.tT.GTAC...C  0A.T...T..3....G...AA..GC..
sim.italy   C.0.G................c5............t......t.g...t.c2t.t.tT.GT.C...C  0A.T...T..3....G...AA..GC..
sim.kenya   C.0.G..t.............c5.....g............gt.g...t.c2t.tct..GT.C....  0A.....T..3....G...AA..GC..
sim.ethio   C.0.G................c5.....g.....a......gt.g...t.c2t.tct..GT.C....  0A.....T..3..T.G...AA..GC..
sim.madag   C.0.G..............a.c5.....g.....a......gt.g...t.c2t.tct..GT.C....  0A.....TT.3....G...AA..GC..
sim.S148    C.0.G................c5.....g.....a......gt.g...t.c2t.tct..GT.C....  0AG....T..3....G...AA..GC..
sec.(x3)    C.0.G...c.......g.....c5..................gt.gt..t.c2t.t.t..GT.C....222A.....T..3A...G...AA..GC..
sec.02483   C.0.G...c.......g.....c5..................gt.gt..t.c2t.t.t..GT.C....222A.....T..3A...G...AA..GC..
mau.S081    C.0.G....g.a..........c5..................t.g.....c2tat....GT.C...C  0A.TA..T..3....G...AA..GC..
mau.S080    C.0.G......a..........c5..................t.g.....c2t.t....GT.C..AC  0A.TA..T..3....G...AA..GC..
mau.02411   C.0.G......a..........c5..................t.g.....c2t.t....GT.C..AC  0A.TA..T..3....G..GAA..GC..
mau.02413   C.0.GA.....g..........c5.......c..........t.g.....c2t.t....GT.C.C.C  0A.TA..T..0....G...AAA.GC..
ere.pub     CT3GGT....t.taaa.gaata.0gagcctg.ttt..cat.g.gcg.gg.tc0t.....AG..CG...  0A.G.AC..A3.GTTTTG.G..GGCAT

FIXED       S   R   S       S                          S  S  S S S S S  SR R     R  R  S   S   S   RS  SS
POLYMORPH       sss s s          s s      s s   ss   s s  s        s s s   r  rrs   rs      s   s   s r
```

APPENDIX B2 (continued)

Nucleotide Position

```
                                                                      1111111111111111111111111111
          666666666777 77777777777 788888888889999999999999999999999990000000000000000000001111111111
          667788899111 13445667789 944555566890223333344445556666677790112334455667788890111223344
          790757829045 74364472821 238123506115083457901395690348913524797146924012602740378194725

mel.pub   ACACCCGC5GAT15CCGGTCGGCA39CCGCACTTCGAt1tgttgaac3aactctatgttTTAGGTGGCCAACCGATTACGGTCGtt1a
mel.(x2)  ........5...15.........39...........1.......3...................................1.
mel.india ........5...15...T.....39...........1.......3...................................1.
mel.(x9)  ........5...15.........39...........1.......3...................................1.
mel.(x5)  ........5...15.........39...........1.......3..........C........................1.
mel.italy ........5...15.........39...........1.......3..........C........................1.
mel.ST05  ........5...15.........39...........1.......3...................................1.
sim.(x2)  ........5.GA15TT...T.... 0A......A...a1.......a0..........aC.C..G..TAT.TA..GC....C...g0.
sim.flori ........5.GA15TT...T.... 0A......A...a1....a..a0..........aC.C....TAT.TA..GC...AC...g0.
sim.colom ........5.GA15TT...T.... 0A....A.A....1.......a0..t.......aC.C..G..TAT.TA..GC....C...g0.
sim.italy ........5.GA15TT...T.... 0A....A.A....1.......a0..ta......aC.C..G..TAT.TA..GC....C...g0.
sim.kenya ........5.GA15TT...T.... 0A......A...a1.......a0..........aC.CT....TAT.TA..GC...AC...g0.
sim.ethio ........5.GA15TT...T.C.. 0A......A...a1.......a0.ta.......aC.C....TAT.TA..GC....CA..g0.
sim.madag ........5.GA15TT...T.... 0A......A...a1.......a0..........aC.C....TAT.TA..GC...AC...g0.
sim.S148  ........5.GA15TT...T.... 0A......AT..a1.......a0..........aC.CT......T.TA..GC....C...g0.
sec.(x3)  ........5.GA15TT...TA... 0A.....CA....1.......a0..ta......aC.C....TAT.TA..GC....C...g1.
sec.02483 .....T..5.GA15TT...TA... 0A.....CA....1.......a0..ta......aC.C....TAT.TA..GC....C...g1.
mau.S081  ........5.GA15.T..CT.... 0A..G..CA...a1.......a0..........aC.C.......T.TA..GC....C...g1.
mau.S080  ........5.GA15.T...T.... 0A.....CA...a0.......a0..........aC.C.......T.TA..GC....C...g1.
mau.02411 ........5.GA15.T...T..T. 0A.....CA...a1......ta0..........aC.C.......T.TA..GC....C...g1.
mau.02413 ........5.GA15.T...T.... 0A.....CA...a1.......a0..........aC.C.......T.TA..GC....C...g1.
ere.pub   CAGTGGCT0TGA 0..T.C....G 0ATT.G..A.TT.3ataaat.t7g.a.tctccc.CAC.A.TT...TT.ATGCGAT.C.Aag1t

FIXED               RR  RS   SR     S     SS            S           SS S       R SR  RR    S   S
POLYMORPH      r           rs  rs      r s  s  s     s s   sss          s s  ss           r s
```

APPENDIX B2 (continued)

Nucleotide positions are numbered relative to the largest assembled consensus sequence (n=1157 nucleotides) using these five species. Only unique haplotypes were listed. mel.(x2) = mel.malay, mel.zimb3; mel.(x9) = mel.hawai, mel.zimb1, mel.zimb2, mel.ST03, mel.ST06, mel.ST09, mel.ST15, mel.F283, mel.F296; mel.(x5) = mel.penn1, mel.penn2, mel.ST08, mel.ST12, mel.ST16. Position 1 corresponds to the first position of the start codon. Nucleotides identical to the first sequence are indicated by dots. Numbers indicate size of indel. Variable sites within indels were excluded. Coding regions are shown in uppercase letters. The nature of each variable site is also summarized amongst the four species of the *D. melanogaster* subgroup (*D. erecta* was excluded from the summary). Fixed substitutions between species of the subgroup are denoted in uppercase letters (S, silent substitution; R, replacement substitution) while polymorphisms are indicated in lowercase (s, silent substitution; r, replacement substitution). Sites that were fixed in a species but polymorphic in another were counted as polymorphic. Coding regions are shown in bold. mel, *D. melanogaster*; sim, *D. simulans*; sec, *D. sechellia*; mau, *D. mauritiana*; ere, *D. erecta*.

APPENDIX B3

Variable nucleotide sites at the *transformer-2* locus amongst sibling species of *D. melanogaster*

```
                                        Exon4                    Exon 5

                 111111-2222222222 2222333334 4444444444 4444555555 5556666666 6677888888
            1224037888 8011111223 3478126780 0112233333 3778011445 5770005568 8901023357
            4127700056 8123568092 3221393690 9671212345 6245437362 8363690195 7733284791

mel-publ    GACGAACCGC TCCGAAAACT TTCCGCCAGA GATGCACGAA TGCCTTCTAC TGACCACGAC TCCCGAGCGG
mel-02316   ......A... .......... .......... .......... .......C.. ......T... ...A......
mel-ital13  .......... .......... .......... .......... .A.....C.. CA..T.T... ...A......
mel-cpal29  .......... .......... .......... .......... .......... ...T...... ..........
mel-z34     .......... .......... .......... .......... .......C.. ......T... ...A......
simulans    TGA.CG.GA. AAAAT.TCGG GG.TAA.G.T .GCA....T. ..TTCGACGT C.C..GTA.G A.ATTC.AAT
sechellia   TGA....GAA ..AATTTCGG GGATAAGGTT .GCA....T. ..TTCGAC.T C.C..GTATG A.A.TC.AA.
mauritiana  TGAAC..GAA ..AATTTCGG GG.TAA.G.T TGCAACAAGT C.TTCGAC.T ..C..GTA.G ATA.TCTAA.

REPLACEMENT                          RRRRSR RSRRRRSSSR R    RRSSS SSSSSS
```

```
                Exon 6                                                  Exon 7

                          11111111 1111111111 1111111111 1111111111 11
            88899999999 9900000000 1111111111 2222222233 3333333333 33
            7990136888 9903445579 1223447999 1223459911 1112344456 68
            6241039137 5929126947 2151451458 6183440315 6890526744 86

mel-publ    GGACATCCCT GCCTCTGTGC CTATGTTCTG TCCCGTTGTC CCCGCCTACG AA
mel-02316   .......... .......... .......... .....C.... .......... ..
mel-ital13  -......... .......... .......... .....C.... .......... ..
mel-cpal29  .......... .......... .......... .......... .......... ..
mel-z34     .......... ....T..... .......... .....C.... .......... ..
simulans    AATAT.TT.C ..T...T.CA TAGA--CG.A AT...CCCAT TTATT.A..A CG
sechellia   AA.ATATT.C ATT...TC.A TAGACCCGGA .TAG.CCC.T T...TTA.AA CG
mauritiana  AA.AT.TTTC ..TA.ATC.A TAGACCCG.A .T..ACCC.. ..T.T.AG.A CG

REPLACEMENT      SSSRS RSSRSRSSSR RRSSSSRSRR RSRSRSSS           R SS
```

The first nucleotide position in this study corresponds to nucleotide position 704 from Amrein *et al.* (1990) and is situated in the noncoding third exon. Sites from coding regions are highlighted in boldface. Exons are indicated as described in Figure 4.1. Substitutions were classified as replacement, R, or silent, S. Indel variation not included.