# Evolutionary and functional analysis of gene expression regulation in *Drosophila melanogaster*

**Winfried Hense**

München 2009

# Evolutionary and functional analysis of gene expression regulation in *Drosophila melanogaster*

Dissertation der Fakultät für Biologie
der Ludwig-Maximilians-Universität München

vorgelegt
von
Dipl.-Biol. Winfried Hense
aus Bobingen

Dissertation eingereicht am: 17. Dezember 2008

Erstgutachter: Professor Dr. John Parsch

Zweitgutachter: Professor Dr. Wolfgang Stephan

Mündliche Prüfung am: 13. Februar 2009

ERKLÄRUNG

Diese Dissertation wurde im Sinne von §12 der Promotionsordnung von Herrn Professor Dr. John Parsch betreut. Ich erkläre hiermit, dass die Dissertation nicht einer anderen Prüfungskommission vorgelegt worden ist und dass ich mich nicht anderweitig einer Doktorprüfung ohne Erfolg unterzogen habe.

EHRENWÖRTLICHE VERSICHERUNG

Ich versichere hiermit ehrenwörtlich, dass die vorgelegte Dissertation von mir selbständig und ohne unerlaubte Hilfe angefertigt wurde.

München, den 17. Dezember 2008

Winfried Hense

*Once more unto the breach, dear friends, once more.*

William Shakespeare, *Henry V*, Act III

*Und wenn auch durch den Nebel nicht viel zu erkennen ist, hat man doch irgendwie das selige Gefühl, in die richtige Richtung zu blicken.*

Vladimir Nabokov (1899 – 1977), russisch-amerikanischer Schriftsteller,
Literaturwissenschaftler und Schmetterlingsforscher

# Table of Contents

# List of Figures

# List of Tables

# List of Abbreviations

| | |
|---|---|
| A | adenine, adenosine |
| *Adh*, ADH | alcohol dehydrogenase (gene, protein/enzyme) |
| ANOVA | analysis of variance |
| bp | base pair(s) |
| C | cytosine, cytidine |
| CV | coefficient of variation |
| *CyO* | *Curly* |
| DNA | deoxyribonucleic acid |
| EDTA | ethylenediaminetetraacetic acid |
| ESE | exonic splicing enhancer |
| Fop | frequency of optimal codon usage |
| G | guanine, guanosine |
| *jan* | *janus* |
| kbp | kilo base pair(s) |
| kcal | kilocalorie(s) |
| mg | milligram(s) |
| miRNA | micro ribonucleic acid |
| ml | milliliter(s) |
| *mle*, MLE | maleless (gene, protein) |
| mM | millimolar |
| mOD | milli optical density |
| nm | nanometer(s) |
| *ocn* | *ocnus* |
| PCR | polymerase chain reaction |
| qRT-PCR | quantitative reverse-transcription PCR |
| *R* | Pearson's correlation coefficient |
| RNA | ribonucleic acid |
| RNAi | RNA interference |
| RSCU | relative synonymous codon usage |
| SAXI | sexually antagonistic X inactivation |
| *Sb* | *Stubble* |
| *Sco* | *Scutoid* |
| *su(Hw)* | *suppressor of Hairy-wing* |
| T | thymine, thymidine |
| tRNA | transfer ribonucleic acid |
| U | uracil, uridine |
| *Ubx* | *Ultrabithorax* |
| UTR | untranslated region |
| *w* | *white* |
| *y* | *yellow* |
| YES | *yellow* enhancers suppressed |

# Zusammenfassung

Die in dieser Dissertation präsentierten Ergebnisse tragen aus dem Blickwinkel der Evolutionsbiologie zu unserem Verständnis der Regulation von Genexpression bei. Ich verwende einen bestens bekannten Modellorganismus, die Fruchtfliege *Drosophila melanogaster*, nicht nur als Objekt der Beobachtung, sondern auch als ein genetisches Manipulationswerkzeug, und untersuche drei verschiedene Aspekte des Prozesses, durch den die in der DNA gespeicherte Information förmlich „entfesselt" oder umgesetzt wird zu biologischem Sinn, letztlich also zu Form und Funktion.

In Kapitel 1 zeige ich zunächst, dass eine Inaktivierung des X-Chromosomes (und somit Genregulation auf chromosomaler Ebene) in der männlichen Keimbahn von *D. melanogaster* stattfindet. Im Gegensatz zur X-Inaktivierung in weiblichen Säugetieren, wo dies in den somatischen Zellen als Mechanismus zur Dosiskompensation auftritt, ist diese Art der Inaktivierung auf die Spermatogenese beschränkt und wurde wahrscheinlich während der Genomevolution als eine Möglichkeit etabliert, schädliche Auswirkungen in Zusammenhang mit Sexualantagonismus zu umgehen. Durch *P*-Element-vermittelte Keimbahntransformation erhielt ich fast 50 unabhängige Insertionen eines testisspezifischen Reportergenkonstrukts und untersuchte die dazugehörigen Reportergenaktivitäten durch Messung der Enzymaktivität und durch quantitative RT-PCR. Autosomale Insertionen dieses Konstrukts zeigten das erwartete Muster hoher männchen- und testisspezifischer Expression. Insertionen auf dem X-Chromosom zeigten dagegen wenig bzw. gar keine Expression des Transgens. Da die X-chromosomalen Insertionen die euchromatischen Abschnitte des Chromosoms abdeckten (bestimmt durch inverse PCR), konnte eine systematische Bevorzugung bestimmter Regionen bei Insertionen, die ein Fehlen von Expression auf dem X-Chromosom hätte erklären können, ausgeschlossen werden. Der Effekt scheint eine globale Eigenschaft des X-Chromosomes zu sein. Lediglich die Testisspezifität des transgenen Konstrukts ist für das Erscheinen des Effekts erforderlich, was somit eine Selektionshypothese für die X-Inaktivierung erhärtet sowie einige Beobachtungen erklären könnte, die im Zusammenhang mit der Verteilung von im Männchen und Testis exprimierten Genen im *Drosophila*-Genom gemacht wurden.

In Kapitel 2 untersuche ich dann mutmaßliche *cis*-regulatorische Sequenzen und ihr Vermögen, allelspezifische Genexpression zu steuern. Nachdem Microarray-Studien umfangreiche Variabilität im Primärmerkmal Genexpression in unterschiedlichsten Taxa aufgedeckt haben, ist eine naheliegende Frage, mit der sich Evolutionsbiologen konfrontiert sehen, die nach der dieser Variabilität zugrunde liegenden genetischen Quelle. Neben epigenetischen Mechanismen gibt es einen Disput darüber, ob regulatorische Sequenzen nahe des exprimierten Gens (*cis*-Faktoren) und anderswo im Genom kodierte Faktoren (*trans*-Faktoren) einen qualitativ und quantitativ unterschiedlichen Beitrag zur Variabilität der Genexpression liefern. Hierzu wählte ich ein Gen von *D. melanogaster*, das nachweislich konsistente Expressionsunterschiede zwischen afrikanischen und nicht-afrikanischen („kosmopolitischen“) Stämmen zeigt, und klonierte die entsprechenden stromaufwärts flankierend gelegenen Teile jeweils in ein bakterielles Reportergenkonstrukt, um – nach erfolgreicher Integration ins Fruchtfliegengenom – direkt die von ihnen gesteuerte Auswirkung auf die Genexpression zu vergleichen. Der beobachtete Effekt war klein, jedoch signifikant, und zeigte sich nur in transgenen Fliegen, die ein X-Chromosom des afrikanischen Ausgangsstammes besaßen. Dies legt den Schluss nahe, dass zusätzlich zu den *cis*-regulatorischen Faktoren auch noch *trans*-Faktoren (vor allem auf dem X-Chromosom) zu dem zwischen den Stämmen beobachteten Expressionsunterschied beitragen.

Letztendlich untersuche ich in Kapitel 3 das Phänomen des *Codon bias* durch seinen Zusammenhang mit Genexpression. Aufgrund der Redundanz des genetischen Codes werden viele der proteinogenen Aminosäuren durch mehr als ein Codon kodiert. Dies ermöglicht es, synonyme Codons in einer kodierenden Gensequenz auszutauschen, ohne dabei die Aminosäurensequenz des kodierten Polypeptids zu verändern. Ob dies Konsequenzen für die produzierte Proteinmenge hat (Translationseffizienz) ist Gegenstand dieses Kapitels. Ich verglich dabei die von zwei Allelen des Gens Alkoholdehydogenase (*Adh*) (von *D. melanogaster*) vermittelte Enzymaktivität direkt miteinander, welche sich in sieben Leucin-Codons unterschieden. Es ergab sich nahezu kein Unterschied in der ADH-Enzymaktivität, obwohl eines der Allele aus gänzlich optimalen Leucin-Codons bestand und das andere sieben suboptimale Leucin-Codons enthielt. Da Letzteres die Wildtypform von *Adh* war, legen die Ergebnisse den Schluss nahe, dass das *Adh*-Gen in seiner Leucin-Codonzusammensetzung (und vielleicht auch in seiner Codonzusammensetzung allgemein) bereits ausreichend optimiert ist. Weitere Versuche, die Zahl der optimalen Leucin-Codons zu erhöhen, können sogar einen Negativeffekt hinsichtlich der Enzymproduktion haben; dies

möglicherweise aufgrund einer Sättigung des tRNA-Pools und/oder der Konsquenzen veränderter mRNA-Sekundärstrukturen.

# Introduction

*Kein Ding, kein Ich, keine Form, kein Grundsatz sind sicher, alles ist in einer unsichtbaren, aber niemals ruhenden Wandlung begriffen.*

Robert Musil (1880 – 1942), Austrian writer

WITH Charles Darwin's (1809 – 1882) bicentenary approaching it is worthwhile beginning with the foundations of evolutionary thought that were laid in 1859 with his seminal work *On the Origin of Species by Means of Natural Selection* (DARWIN 1859), which itself will celebrate its 150th anniversary in 2009. After the voyage on board of the *Beagle* from 1831 to 1836 Darwin had gathered a lot of material and ideas to finally begin to develop the groundbreaking novel thought of descent with modification. The hitherto widespread religious belief of the constancy of all species, based on and derived from the Scriptures, was shattered. Briefly, Darwin envisioned all life forms as passively and gradually changing through time simply because evolutionary change cannot not happen. This can be concluded from his main observations and inferences: Organisms can produce far more offspring than the environment could carry; they have the potential to grow exponentially, yet natural populations remain rather constant in size. Because natural resources are always limited, a struggle for existence must have happened to accomplish that. On the other hand, there is variability among members of a population in almost every trait, from morphology to physiology, which is partially genetic, *i.e.* passed on from generation to generation. Therefore, traits that help organisms survive and cope with their biotic and abiotic environment must gradually accumulate in a species, allowing them to adapt to nature. This selection theory, with selection taking place when only slightly more fit organisms produce (in every generation) on average only marginally more offspring, leads to the above claim that evolutionary change caused by selection cannot not happen.

It must be noted that Darwin's theory contains two main statements: There is change and the mechanism or force enabling this change is natural selection. Some 50 years before Darwin, Jean-Baptiste de Lamarck (1744 – 1829) had already developed a complete theory of evolution including modification through time, yet the mechanism he proposed was different (acquisition of useful traits during an individual's lifetime and inheritance to the next generation) and later rejected (Interestingly, in the recent past this idea has been revived by the field of epigenetics.). Furthermore, Darwin, when developing his thoughts about evolution, was influenced by people like the English political economist Thomas Robert Malthus (1766 – 1834) and the Scottish geologist Charles Lyell (1798 – 1875) about population growth and gradual processes over geological time-scales, respectively. Nevertheless, Darwin's genuine contribution to evolutionary theory was the careful synopsis of all known facts and observations and the idea of the final mechanism that leads to evolutionary change and speciation (The latter, however, not so much: speciation was not the main theme of his 1859 book!). Two more things are noteworthy. At around the same time as Darwin another British naturalist, Alfred Russel Wallace (1823 – 1913), came to the same idea about evolution by natural selection. He sent his ideas to Darwin to get his opinion, which accelerated Darwin's publication of *The Origin*. Although Darwin deservedly received most of the credit for the theory of evolution by natural selection, Wallace also made an important contribution to the foundations of modern biology. Secondly, Darwin had no scientific knowledge whatsoever about what was later to be termed genetics, although it was exactly in those years when Darwin was developing his theory that Gregor Mendel (1822 – 1884) discovered the first rules of inheritance by experimenting with peas.

Mendel's insights were later, at the turn of the century, independently rediscovered by Hugo de Vries, Carl Correns, and Erich Tschermak. However, it was not until the 1940s that a synthesis took place, in which conflicts between the fields of genetics, cytology, systematics, and paleontology were reconciled to create a powerful updated theory of evolution, which is associated with the names of Ernst Mayr, George G. Simpson, and Theodosius Dobzhansky (among others). Included in this modern synthesis was a mathematical theory (population genetics) that describes the temporal dynamics of alleles and their frequencies in the gene pool of a population under the influence of five fundamental evolutionary forces: mutation, recombination, genetic drift, demography, and selection. Here, Ronald A. Fisher, Sewall Wright, and John B.S. Haldane were the founders of and major contributors to that theory. Eventually population genetics culminated in Motoo Kimura's renowned *Neutral Theory of Molecular Evolution* (KIMURA 1968, 1983), which states that in order to explain molecular patterns of polymorphisms observed in protein and nucleotide sequences, positive or directional selection need not to be invoked, leaving random genetic drift and purifying selection (which removes deleterious mutations relatively quickly from the gene pool) as the dominant forces governing (observable) evolutionary change. Indeed, it was somewhat uninspiring to think about organisms (and hence ourselves) as the products of chance (genetic drift and demography) and as leftovers of the removal of deleterious mutations. On the other hand, natural selection acts upon relative fitness, *i.e.* as long as the wild-type performs better than a newly arisen mutant, the latter will consequently be purged from the gene pool. The same or some similar mutant, however, could be advantageous in a different environmental setting and be the variant to survive in the gene pool. Natural selection as an outcome always describes an interaction of genotypes with nature with the final result of better adaptation. This process operates continually, also on a background tendency to higher complexity of life forms. In this sense, one could argue that in the early stages of evolution, when the complexity of organisms was still quite low, it was much easier to improve genetic entities, whereas nowadays after one of the major transitions in evolution (MAYNARD SMITH and SZATHMARY 1995) – multicellularity – has enabled complex life forms to evolve, improvement might have become a much rarer event. Thus, the apparent lack of molecular evidence of positive selection described by Kimura that puzzles and challenges population geneticists up to now might also be the result of its relatively rare occurrence compared to the number of deleterious mutations, making it a daunting task to find signatures of positive selection in a sea of neutral or slightly deleterious mutations or abundant signatures of negative selection. Moreover, even if signatures of positive selection can be found (for example reduced variation in the genomic neighborhood surrounding a fixed beneficial mutation, aka a selective sweep; MAYNARD SMITH and HAIGH 1974), the question is for how long it can be detected, since all characteristics of such a selective sweep (reduced polymorphism, a skew in the site frequency spectrum, and high linkage disequilibrium) are expected to vanish after the advantageous mutation (and thus the causative agent) has become fixed in the gene pool, thereby bringing the process producing those characteristics to a halt. Thus, it is possible that positive selection is acting intensely, but its traces in the genetic material might not be detectable in most instances. Today it is becoming possible to address the above issues in a quantitative way with large-scale population genetic surveys.

A specific drawback of population genetics could be seen in its primary focus on the genotype. Certainly, mutations arise in the genetic material in the first place, from where they are transformed into the phenotype. As long as the dynamics of mutant variants in the gene pool and the forces governing them are to be considered and mathematically modeled, this focus is understandable. It can also be justified historically, as the molecular level at which variation could be observed was pushed forward only step-wise due to technical limitations and improvements to overcome them. Being a purely theoretical science in the beginning, it took decades for population genetics to take advantage of technical achievements to be able to quantify polymorphisms at the molecular level of proteins and DNA, with the latter starting at variation at restriction sites, moving further to the sequencing of single genes (*e.g.* in *Drosophila* KREITMAN 1983), and finally arriving at whole genome sequences (*e.g.* ADAMS *et al.* 2000; VENTER *et al.* 2001). Another potential reason for the focus on the genotype is the still poorly understood relationship between genotype and phenotype. Although great progress has been made in the field of developmental genetics, whose genuine task it is to reveal this relationship, in the last two to three decades (NÜSSLEIN-VOLHARD and WIESCHAUS 1980; ST. JOHNSTON and NÜSSLEIN-VOLHARD 1992; LEWIS 1992) and more and more genome sequences of diverse taxa are becoming available, the genetic basis for adaptations, which are phenotypic by nature and were the starting point for Darwin, are unknown in most cases. Moreover, the fitness or adaptive value as a phenotypic outcome is very difficult to measure for most traits or genes, making experimental validation of theoretical findings difficult. Thus, with this limitation to our knowledge, it seems justified for population genetics to restrict its efforts to the genotype. What is needed in the future is a comprehensive functional annotation of genomes in a quantitative genetics framework (with the help of developmental genetics and biochemistry) to elucidate each gene's contribution to the phenotype.

To support this process, some 40 years ago BRITTEN and DAVIDSON (1969) and later KING and WILSON (1975) proposed that gene regulation plays an eminent role in phenotypic evolution. Based on models of gene regulation by JACOB and MONOD (1961) and the observation that primary sequence information in proteins and DNA between closely related species like humans and chimpanzees is very conserved, they argued in favor of differences in the regulation of genes to account for the larger part of phenotypic differences. Despite a particular focus on bacterial gene regulation in the model of Jacob and Monod, research since then has shown that gene regulation can be achieved on several molecular levels. First, the chromatin of each chromosome occupies distinct higher-order regions within the nucleus called chromosome territories (CREMER *et al.* 2006). These are not always fixed in their structure and position, but can depend dynamically on processes taking place in the nucleus (like replication or transcription). Transcriptional activation can lead to relocalization of chromosome territories within the nucleus during interphase. Thus, higher-order nuclear architecture plays a role in gene regulation, as shown at least in mammalian cells (reviewed by LANCTÔT *et al.* 2007; FEDOROVA and ZINK 2008). Second, chromatin allows for gene expression regulation also at the level of nucleosomes. Histone modifications like methylation and acetylation are known to have regulatory potential. Adding or removing those functional chemical groups can flexibly modify the density of DNA packaging to allow the transcription machinery (consisting of transcription factors and RNA polymerase) to access its target region (WANG *et al.* 2004). Furthermore, the methylation of DNA also can have an impact on gene regulation (LEONHARDT and CARDOSO 2000). The patterns of such methylation in regions upstream of genes differ from the surrounding DNA and depend on cell-type, tissue, age, and sex. Quite often this leads to the appearance of CpG islands. A special case of DNA methylation is genomic imprinting, in which case one of two existing alleles – either the maternal or the paternal one – is transcriptionally silenced. The whole field is nowadays called "epigenetics" because changes in gene expression or other traits need not to be caused by changes (mutations) in the DNA sequence itself, but instead on a layer "*on top of*" mere sequence (hence the Greek prefix "*epi*"). Among epigenetic phenomena are two that play a role in this thesis: First, some position effects, which describe the variation in expression of a transgenic reporter gene construct depending on the chromosomal location, may be caused by chromatin structure and/or modification. In transgenic experiments, *e.g.* in the fruit fly *Drosophila melanogaster*, DNA from a different organism (transgene) can be integrated into the genome by injecting it into early embryos. Using the method of *P*-element mediated germline transformation, the location of the integrated transgene can usually not be targeted and thus remains random. However, as the DNA sequence of a specific transgene is identical among all of its insertion sites, and the expression shows considerable variation, the observed differences must be explained by regulatory features that lie outside of the transgenic DNA. In transgenic experiments used throughout this thesis, position effect variation is an issue. Second, the inactivation of the X chromosome during spermatogenesis in *D. melanogaster*, the topic of CHAPTER 1 of this thesis, also must be regarded as epigenetic, as it occurs only in a particular tissue (testis).

Furthermore, research of recent years has demonstrated that transposable elements, which usually make up a large fraction of eukaryotic genomes and therefore contribute to genome architecture, evolution, and the emergence of genetic innovations (FESCHOTTE and PRITHAM 2007), can have an important role in the construction of gene regulatory networks (FESCHOTTE 2008). Finally, when considering the processes of transcription and translation, numerous additional ways to control gene expression have been elucidated. Transcription is initiated by the binding of general and specific transcription factors (the *trans* factors) and RNA polymerase to the specific regions of uncoiled and opened DNA close to the transcription start site (the *cis* factors) (Figure 0.1). Polymorphisms in both the *trans* and the *cis* factors are thought to contribute to intra- and interspecific differences in gene expression

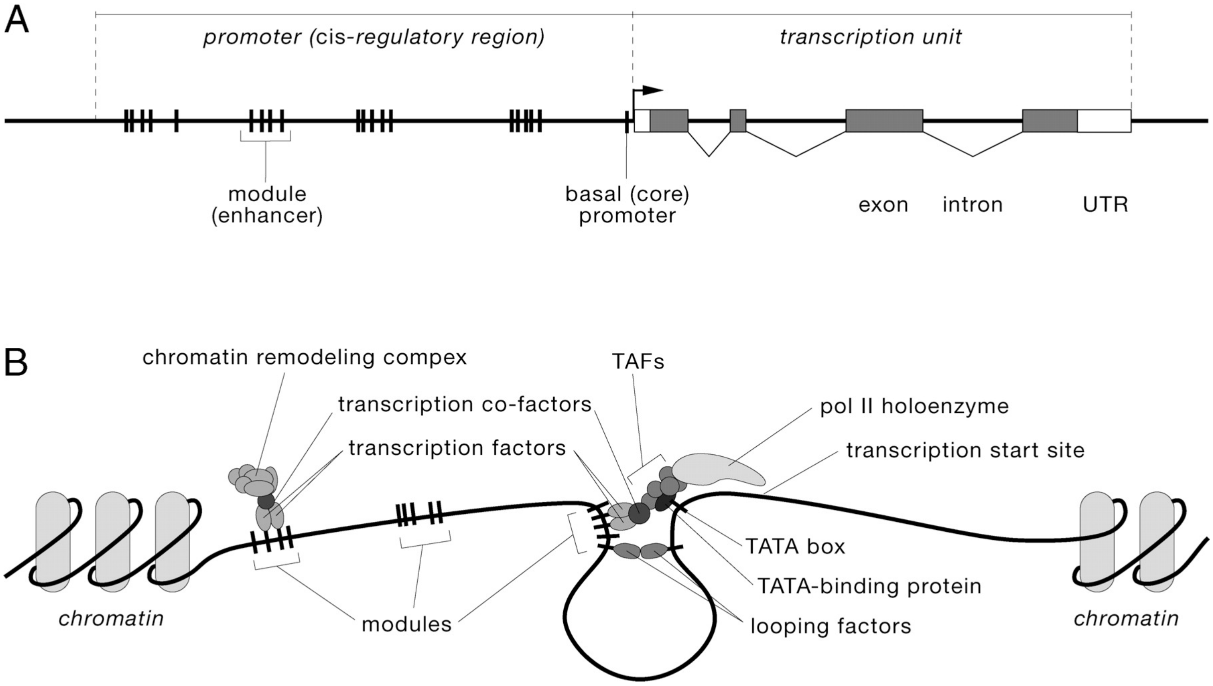


**Figure 0.1 Gene expression regulation.** – A) Shown is the organisation of a typical eukaryotic gene with its exon-intron structure and additional basal regulatory sequences (UTRs and core promoter) and the upsteam cis-regulatory region (promoter) consisting of several modules serving as transcription factor binding sites. B) The promoter at the start of transcription. The chromatin structure has been decondensed to allow the transcription machinery to bind to its respective cis-sequences. Numerous basal and accessory factors are depicted, some of which are only facultative. (taken from Wray *et al.* 2003)

(WRAY *et al.* 2003). In CHAPTER 2 of this thesis I investigate putative *cis*-regulatory polymorphisms and their ability to regulate expression of a bacterial reporter gene inserted into the *Drosophila* genome. After an mRNA transcript is produced it must be further processed by splicing, which offers additional possibilities for regulation. Splicing signals at the sequence level, *e.g.* exonic splicing enhancers or silencers, but also the typical intron boundaries (GU-AG at the 5' and 3' ends of the intron, respectively), can be created or removed by point mutations. Moreover, alternative splicing is a process that has to be regulated by additional signals through genetic or epigenetic mechanisms (WANG and BURGE 2008). The mature mRNA is then transferred from the nucleus to the cytoplasm where translation at the ribosomes is the next step of gene expression. Normally, the mRNA transcript forms secondary structures that stabilize the transcript thermodynamically. Mutations, especially synonymous mutations that do not alter the final amino acid sequence, are thought to influence this stability in an advantageous or deleterious way that may be subject to selection (WADA and SUYAMA 1986; but see also CARLINI *et al.* 2001; STENØIEN and STEPHAN 2005; ECK and STEPHAN 2008). The mature mRNA further contains untranslated regions at the 5' and 3' end (5' and 3' UTRs) that harbor more regulatory sequences, the most prominent of which are binding sites for microRNAs. This type of RNA is known to silence gene expression post-transcriptionally (CHEN and RAJEWSKY 2007; FILIPOWICZ *et al.* 2008). When bound to its appropriate target sequence, it starts its degradation by a process that resembles RNA interference (RNAi). Finally, at the ribosomes, the accuracy and efficiency of translation is also determined by the availability of appropriate tRNAs, whose abundance is regulated on its own. The efficiency of translation is thought to be increased by the use of synonymous codons that match the most abundant tRNAs. This may lead to biased codon usage, especially in highly-expressed genes, which has been observed in many species (BULMER 1987) (Figure 0.2). The influence of synonymous codon usage on translational efficiency is a particular focus of CHAPTER 3.

In all the above modes of gene expression regulation, from an evolutionary standpoint it is interesting to ask whether the mechanisms rely upon DNA sequence or are epigenetic. Whereas the former is accessible to evolutionary and population genetic analysis (Questions to be addressed include: What are the dynamics of such DNA sequence variants/alleles? What are typical mutation rates for this kind of DNA? What kind of selection is acting upon such sequences?), the latter cannot be addressed in such a simple way, although there must be genetic factors in the end that are responsible for an epigenetic mode of regulation. Transcription factors, for instance, strongly influence gene expression, but they are difficult to identify and map to the genome of an organism. Once found, they are most likely regulated themselves in a complex manner. Methyltransferases, to give another example, have already been identified and analyzed (SPADA *et al.* 2006; SCHERMELLEH *et al.* 2008), linking their activity and performance, however, to specific polymorphisms in their coding sequence on the one hand, and to their overall effects on target DNA on the other is a task beyond current methodology. Moreover, it is quite possible that in many cases the relationship of genotype to phenotype becomes blurred relatively quickly after the first basic step of transcription, since there may be many genes contributing to a specific phenotype, and is thus lost in a kind of statistical noise. An alternative view would be that most traits are governed by the expression of only a few genes, perhaps under the control of one master gene (as is the case for some developmental pathways). This would allow selection to operate more effectively and quickly. Clearly, this belongs to the field of statistical and quantitative genetics.

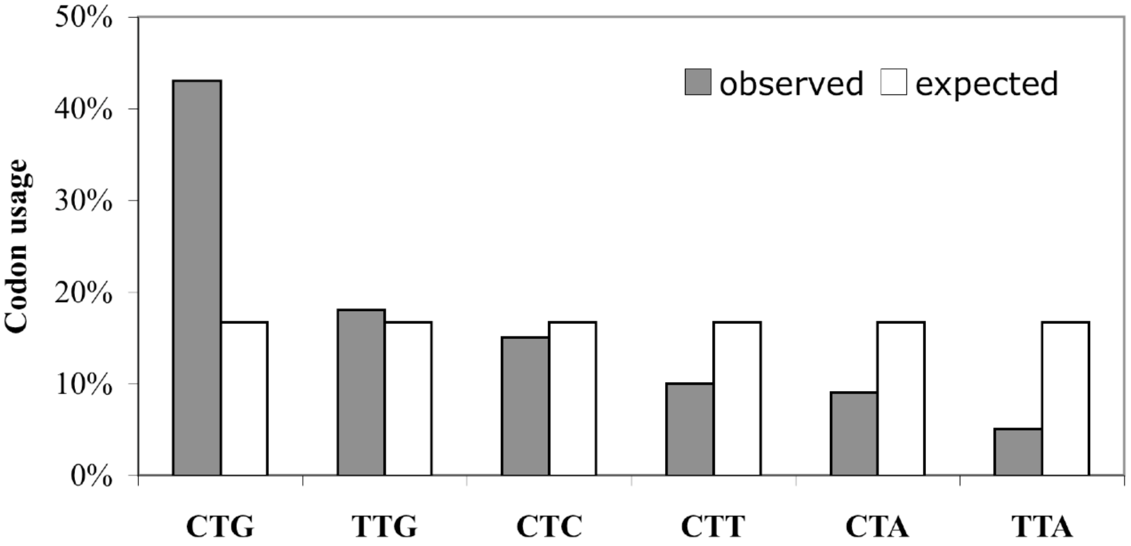


**Figure 0.2 Leucine codon usage in *D. melanogaster.*** – Since there are six different leucine codons, the expected random usage would be 16% per codon (open bars). The real codon usage is biased as shown by the filled bars with one major codon (CTG) used more than 40% of the time in the fruitfly genome. (Data taken from *Codon Usage Database*, http://www.kazsusa.or.jp/codon)

What experimental methods are available to investigate the regulation of gene expression? One approach that was employed several times during the course of my dissertation research is germline transformation to create transgenic organisms. This has become a standard method to analyze gene expression in *Drosophila melanogaster*. The most-used method that has been established in the fruit fly is called *P*-element mediated germline transformation. It makes use of recognition sites derived from the DNA transposon *P*, which was discovered together with a syndrome named hybrid dysgenesis (KIDWELL *et al.* 1977; reviewed by ENGELS 1992). The *P*-element usually transposes itself by expression of the only gene it encodes, a transposase, which cuts out the element (by utilizing *inverted repeat* sequences that flank the element) and reintegrates it somewhere else in the genome. This was later developed into a molecular genetic tool for *Drosophila* transformation by constructing plasmid vectors carrying the *P*-element where the transposase gene is exchanged with a gene or genetic element of interest (not necessarily from the same organism, thereby allowing for transgenesis; RUBIN and SPRADLING 1982; SPRADLING and RUBIN 1982). If the vector construct carries an additional marker gene, *e.g.* a phenotypic marker like the pigmentation gene *yellow*, or a gene responsible for eye color (*e.g. white*), successfully transformed flies can easily be recognized (PIRROTTA 1988). To mobilize the modified *P*-element, an independent source of transposase is required. Often this is done by transforming flies that possess a constitutively-expressed variant of this enzyme already integrated into their genome (ROBERTSON *et al.* 1988). As an alternative, the gene for transposase can be co-injected on a second "helper" plasmid together with the plasmid carrying the genetic element of interest. Germline transformation is accomplished by injecting the plasmid vector(s) into the posterior end of pre-blastoderm embryos, where the precursor cells of future gonads (germline cells) are located. In a few of those cells the modified *P*-element will transpose from the plasmid to a random chromosomal location. As a consequence, the offspring derived from these transformed germline cells will be "transformants" and will carry the transgene in all cells of their body.

Among the many vectors that make use of the principal method described above are two that were used in this thesis, 1) the YES vector ("*yellow*, enhancers suppressed"; PATTON *et al.* 1992), and 2) the "waffle" vector (p*P[wFl]*; SIEGAL and HARTL 1996), which both have the advantage of controlling for position effects. As already mentioned above, the chromosomal position where a transgene is inserted into cannot be determined *a priori*. (The random insertion site, however, could be mapped afterwards by inverse PCR methods.) Normally, when doing transgenic experiments, a certain number of independent insertions is obtained, and the transgene's outcome (*e.g.* expression of a reporter gene) is averaged over all insertions. Because this outcome varies considerably depending on chromosomal location (position effect), scientists were soon interested in applying transformation vectors that reduced this problem. The YES vector accomplishes this by adding binding sites for a specific protein (Suppressor of Hairy wing) which, when bound, serves to insulate the transgene from external regulatory elements. The second vector, the "waffle" vector, on the other hand, circumvents position effects differently. Occasionally researchers are interested in comparing two or more versions of genes or genetic elements with each other. In such a case, the "waffle" vector can be applied to first insert a pair of them into a random chromosomal position (as described above; transgene coplacement), and afterwards remove one of them while leaving the other untouched (by utilizing site-specific recombinases). If this is done with each of the two variants to be compared separately, one ends up with a pair of transformant lines, each with one of the transgenes at precisely the same chromosomal position as the other. This means that the chromosomal context in which two variants are

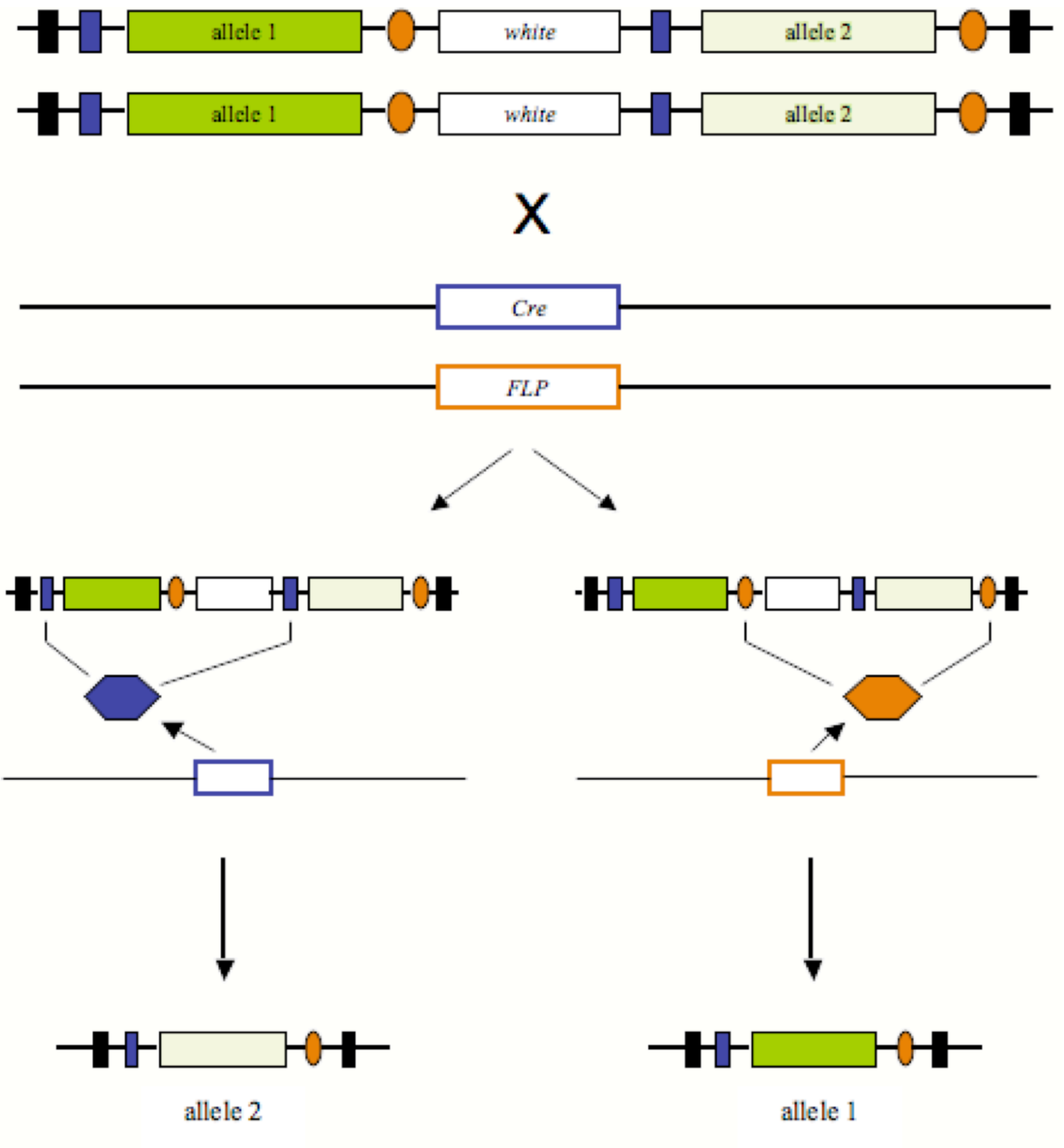


**Figure 0.3 Transgene coplacement and the "waffling" crossing scheme.** – A transgenic fly homozygous for a "waffle" vector double construct on the 3rd chromosome is crossed to a fly strain heterozygous for two recombinase genes (*Cre* and *FLP*), also on the 3rd chromosomes. The resulting offspring will be heterozygous for the "waffle" construct and one of the recombinases. The recombinase will then excise one "allele" from the waffle construct, leaving the other one behind. (Figure completely designed by W. HENSE.)

embedded in is identical, and hence the outcome or effect of the transgenes can be compared directly. Details of this method are illustrated in Figure 0.3.

With this background theory and technical knowledge at hand, in CHAPTER 1 I investigate the global silencing of a chromosome during *Drosophila* spermatogenesis. The X chromosome, which only exists as a single copy in males (the heterogametic sex) is shown to be inactivated early in the process of sperm maturation in the male germline. This was demonstrated by the integration of a bacterial reporter gene construct that exhibits testis-specific expression into the genome of *D. melanogaster*. When inserted on the autosomes, expression of the reporter gene was measured at medium to high levels and was specific to testis. In contrast, X-chromosomal insertions of the reporter gene showed only very low levels of expression. These observations hold for 50 different chromosomal insertions, with both the YES and the "waffle" transformation vectors. The expression difference was confirmed at the level of gene transcription (in addition to enzymatic activity of the reporter gene product) by quantitative measurement of transcript abundance by qRT-PCR. These results are in accordance with and support a selective hypothesis in genome evolution that states that male- and testis-expressed genes are selectively favored to "escape" the X chromosome during the course of evolution to avoid inactivation in the male germline.

In CHAPTER 2, I examine gene expression at the level of an individual gene, with the focus on upstream regulatory elements. By performing transgenic experiments with the "waffle" transformation vector, I investigate the ability of putative *cis*-regulatory sequences to drive allele-specific gene expression. Focusing on the gene *CG13360* of *D. melanogaster*, which shows a consistent expression difference between African and non-African ("cosmopolitan") strains, I sequenced the upstream region to identify sequence polymorphisms that are associated with the respective expression states. These were then functionally analyzed through experimentation by transgene coplacement. For this, the upstream regions of two *D. melanogaster* strains, one African and one cosmopolitan, were cloned in front of a reporter gene, coplaced into the genome, and their reporter activities were compared. I found a small, yet significant, expression difference between the two putative upstream promoters, which interestingly appears only in transgenic flies, and only if an X chromosome of the African strain is present. This suggests that, in addition to the *cis*-regulatory polymorphisms present in the cloned upstream region, there are also unlinked regulatory factors that act in *trans*. These *trans*-factors appear to be located on the X chromosome and contribute to the expression difference of *CG13360* observed between the two original strains.

Finally, in CHAPTER 3, I examine the role of synonymous codon usage in post-transcriptional gene regulation. In contrast to the random expectation that the synonymous codons within a given codon family be used with equal frequency, many species show a strong bias in their codon usage. A previous study by CARLINI and STEPHAN (2003) showed that replacement of optimal leucine codons in the *D. melanogaster* alcohol dehydrogenase gene (*Adh*), one of the most highly expressed genes in the fruit fly genome, with sub-optimal codons resulted in decreased ADH enzymatic activity. This suggested that translational efficiency was reduced, because the amino acid sequences of both the wild-type and the mutated *Adh* alleles were identical. In CHAPTER 3 I describe the reverse experiment, in which seven sub-optimal leucine codons in the *Adh* gene were replaced with the optimal codon. The resulting ADH activities were measured *in vivo* using the method of transgene coplacement. The introduction of these optimal codons did not lead to an increase in ADH enzymatic activity. Instead, transformants with the optimized *Adh* allele showed slightly less ADH activity than those with the wild-type allele. These results can be explained within the scope of the translational selection hypothesis of codon bias, which postulates that optimal codons increase the accuracy and/or efficiency of translation, if one assumes that there are diminishing returns to increasing optimal codon usage. For example, codon bias in the wild-type *Adh* gene may already be sufficiently optimized to match the species' tRNA pool and further increases in codon bias may have little or no phenotypic effect.

# Chapter 1

# X chromosome inactivation during *Drosophila* spermatogenesis

SEX chromosomes, such as the X and Y chromosomes of *Drosophila*, are thought to have evolved from a pair of homologous autosomes that lost their ability to recombine with each other (CHARLESWORTH 1996; RICE 1996). Over evolutionary time, the sex chromosome that is present only in the heterogametic sex (the Y) tends to degenerate, losing most of its gene complement and accumulating transposable elements (GANGULY *et al.* 1992; STEINEMANN and STEINEMANN 2000, 2001; BACHTROG 2005). The X chromosome, which is still able to recombine within the homogametic sex, maintains a fully functional complement of genes and resembles an autosome in its size, cytogenetic appearance, repetitive element content, and gene density. Recent genomic studies, however, have revealed a number of more subtle differences in gene content, expression pattern, and molecular evolution between the X chromosome and the autosomes (VICOSO and CHARLESWORTH 2006).

One pattern that has emerged from the genomic analysis of *Drosophila melanogaster* is that there is a significant excess of gene duplications in which a new autosomal gene has arisen from an X-linked parental gene through retrotransposition (BETRÁN *et al.* 2002). Most of these new autosomal genes appear to be functional and are expressed in testis (BETRÁN *et al.* 2002). Several of these genes that have been studied in detail show evidence of adaptive evolution and/or functional diversification (BETRÁN *et al.* 2002; BETRÁN and LONG 2003; Betrán *et al.* 2006; KALAMEGHAM *et al.* 2006). Another pattern that has emerged from functional genomic studies is that genes with male-enriched expression are underrepresented on the X chromosome (PARISI *et al.* 2003; RANZ *et al.* 2003). For example, about 19% of all *D. melanogaster* genes reside on the X chromosome, but only 11% of the genes with a twofold or greater male bias in expression are X-linked (HAMBUCH and PARSCH 2005). Furthermore, the male-biased genes that are X-linked tend to show less sex bias in their expression than those that are autosomal (CONNALLON and KNOWLES 2005).

A number of hypotheses have been put forth to explain the above observations (ROGERS *et al.* 2003; SCHLÖTTERER 2003; OLIVER and PARISI 2004). To explain the large excess of retrotransposed genes that have "escaped" the X chromosome, BETRÁN *et al.* (2002) proposed the X inactivation hypothesis, which posits that genes with a beneficial effect late in spermatogenesis are selectively favored to be autosomally located. Otherwise, their expression would be prevented by male germline X inactivation, which is supposed to occur early in spermatogenesis at a time when autosomal genes are still actively transcribed. Early X inactivation could also explain the paucity of genes with male-biased expression on the X chromosome: if X-linked genes cannot be expressed in the later stages of spermatogenesis, then one would expect to see fewer X-linked genes with enriched expression in adult males. In particular, this should be true for genes expressed in the male germline and those encoding sperm proteins, which has been observed (PARISI *et al.* 2003; DORUS *et al.* 2006).

Male germline X inactivation, however, cannot completely explain the observations. For instance, male-biased genes that are expressed only in somatic cells, where X inactivation does not occur, are also significantly underrepresented on the X chromosome (PARISI *et al.* 2003; SWANSON *et al.* 2003). An alternative explanation that accommodates this observation invokes sexual antagonism, that is, evolutionary conflict between males and females. The fixation probability of an X-linked, sexually-antagonistic mutation is expected to differ from that of an autosomal one, with the direction of this difference depending on the dominance coefficient (RICE 1984, CHARLESWORTH *et al.* 1987). If the antagonistic effects are (at least partly) dominant, then female-beneficial/male-harmful mutations will accumulate on the X chromosome, while male-beneficial/female-harmful mutations will be removed from the X. This is because the X chromosome spends two-thirds of its evolutionary history in females and, thus, is more often under selection in the background of this sex. Since genes with sex-biased expression may be prime targets for sexually antagonistic mutations, the above scenario could lead to an excess of female-biased genes and a paucity of male-biased genes on the X (RANZ *et al.* 2003), resulting in "feminization" or "demasculinization" of this chromosome (PARISI *et al.* 2003).

A hypothesis that combines the concepts of sexual antagonism and X inactivation was proposed by WU and XU (2003). This hypothesis, termed SAXI (sexually antagonistic X inactivation), suggests that natural selection has favored the movement of sexually antagonistic X-linked genes whose expression is beneficial to males to the autosomes, leaving those beneficial to females on the X. Over evolutionary time, the accumulation of female-beneficial/male-harmful genes on the X leads to selection for X inactivation in the male germline, particularly during the later stages of spermatogenesis where the effects of sexual antagonism are expected to be greatest (WU and XU 2003). The hypotheses of BETRÁN *et al.* (2002) and WU and XU (2003) assume that the X chromosome becomes inactive before the autosomes during spermatogenesis. This phenomenon has been established in mammals and nematodes (RICHLER *et al.* 1992; KELLY *et al.* 2002; FONG *et al.* 2002). However, the evidence for male germline X inactivation in *Drosophila* has been equivocal. LIFSCHYTZ and LINDSLEY (1972) cited cytological observations and genetic experiments to argue that X inactivation during spermatogenesis was common to most animal species with heterogametic males, including *D. melanogaster*. However, similar evidence was used to argue against X inactivation in *Drosophila* (MCKEE and HANDEL 1993). A later study of the expression of sperm-specific proteins in transgenic *Drosophila* provided experimental support for X inactivation (HOYLE *et al.* 1995). Here the authors used a testis-specific promoter to drive the expression of altered forms of β-tubulins in the male germline and noted that X-linked inserts of the constructs showed reduced expression relative to autosomal inserts. Although this result was consistent with X inactivation, there were some limitations. For instance, the sample sizes were small for each of the expression constructs, with only one or two X-linked inserts per construct. Furthermore, the expression level of the genes was only roughly estimated from protein abundance on electrophoresis gels.

A more recent experimental study failed to find support for male germline X inactivation in *Drosophila* (RASTELLI and KURODA 1998). These authors examined the expression and intracellular location of the MLE protein (encoded by *maleless*), as well as the acetylation pattern of histone H4, in male germline cells. Although MLE is known to be involved in X chromosome hypertranscription in somatic cells, presumably through the recruitment of histone acetylation factors (GU *et al.* 1998; SMITH *et al.* 2000), it does not associate specifically with the X chromosome in male germ cells. Furthermore, H4 acetylation at lysine 16, which is thought to be a reliable marker for active transcription, was observed equally on the X chromosome and the autosomes. Thus, there was no evidence for dosage compensation or X inactivation in the male germline. However, it is not necessary that these two processes occur through the same mechanism, or that they rely on the same proteins required for somatic cell dosage compensation. Indeed, a microarray analysis of germline gene expression indicated that dosage compensation does occur in the male germline (GUPTA *et al.* 2006). Because these microarray experiments used reproductive tissues that contained somatic cells and germline cells from all stages of gametogenesis, they could not directly address the issue of early X inactivation. However, the fact that most X-linked genes showed similar levels of expression in both male and female reproductive tissues suggests that, if X chromosome inactivation does occur in the male germline, it does not have a large effect on the global pattern of sex-biased gene expression.

In this study, we perform a more rigorous experimental test for X inactivation in the male germline. Using a transgenic construct in which the expression of a reporter gene is driven by the promoter of the autosomal, testis-specific *ocnus* (*ocn*) gene, we show that autosomal inserts are expressed specifically in males and in testis. X-linked inserts, in contrast, show greatly reduced levels of expression. These results hold for a large sample of independent insertions and for two different transformation vectors and, thus, provide strong support for inactivation of the X chromosome during *Drosophila* spermatogenesis.

## 1.1 Materials and Methods

### 1.1.1 Transformation vector construction

Two different expression vectors that combined the *ocn* promoter of *D. melanogaster* with the *lacZ* coding region of *E. coli* were generated using standard techniques (Sambrook *et al.* 1989). For the first, we PCR-amplified a 150-bp fragment of *D. melanogaster* genomic DNA that spanned bases 25,863,383 - 25,863,532 of chromosome 3R in genome release 5.1 (http://www.flybase.org). The amplified region includes 80 bases of 5' flanking sequence and 70 bases of 5' UTR of the *ocn* gene (*CG7929*), corresponding to bases –165 to –16 relative to the A in the ATG start codon. We chose to end the promoter fragment at –16 because the preceeding sequence presented a good target for PCR-primer design; we know of no functional reason to include or exclude the final 15 bp before the start codon. The PCR product was cloned directly into the pCR2.1-TOPO vector (Invitrogen, Carlsbad, CA). The identity and orientation of the cloned fragment were confirmed by restriction analysis. A 3.5-kb *Not*I fragment containing the complete *E. coli lacZ* open reading frame was excised from the plasmid pCMV-SPORT-βgal (Invitrogen) and inserted into the *Not*I site of the above plasmid, just downstream of the *ocn* promoter and in the same orientation. A 3.6-kb fragment containing the *ocn* promoter and the *lacZ* coding region was then excised as an *Spe*I/*Xba*I fragment and cloned into the *Spe*I site of the p*P[wFl]* transformation vector. This vector is based on the *P* transposable element and contains the *D. melanogaster white* (*w*) gene as a selectable marker (Siegal and Hartl 1996). The final construct was designated p*P[wFl-ocn-lacZ]* (Figure 1.1A).

The second expression vector contained the *ocn* promoter described above as well as the *ocn* 3' UTR sequence (Figure 1.2). The *ocn* promoter was excised from the pCR2.1-TOPO vector as a *Bam*HI/*Xba*I fragment and inserted into the *Bam*HI/*Xba*I sites of the plasmid pUC18 (Invitrogen). The *ocn* 3' UTR sequence was PCR-amplified from genomic DNA corresponding to bases 25,862,721 – 25,862,830 of chromosome arm 3R (bases –16 to +93 relative to the T in the TGA stop codon) and cloned into the pCR2.1-TOPO vector. After confirming the identity and orientation of the cloned fragment by restriction analysis, a *Hind*III fragment (where one *Hind*III site was internal to the 3' UTR fragment, occurring 30 bp from the 5' end) was extracted and inserted into the *Hind*III site of the pUC18 plasmid containing the *ocn* promoter, such that the promoter and 3' UTR were in the same orientation. An *Spe*I fragment containing both the promoter and the 3' UTR was then excised and cloned into the *Xba*I site of the YES vector  (PATTON *et al.* 1992). This vector is also based on the *P*

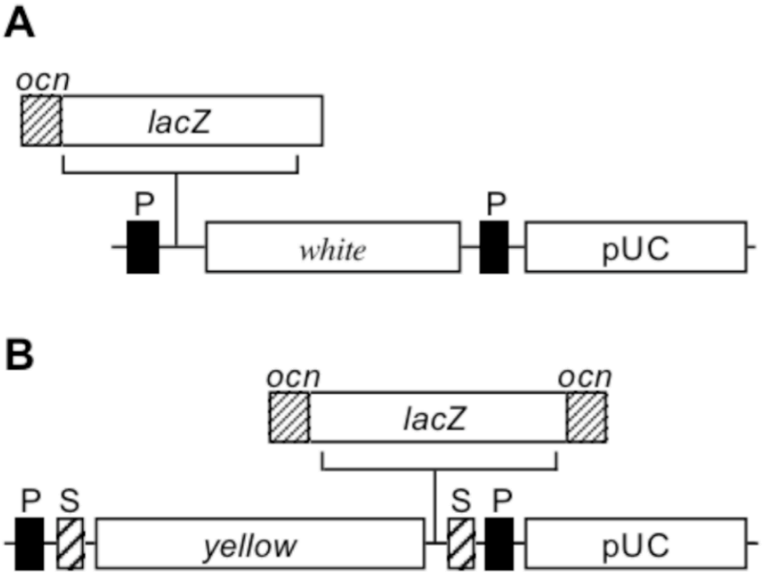
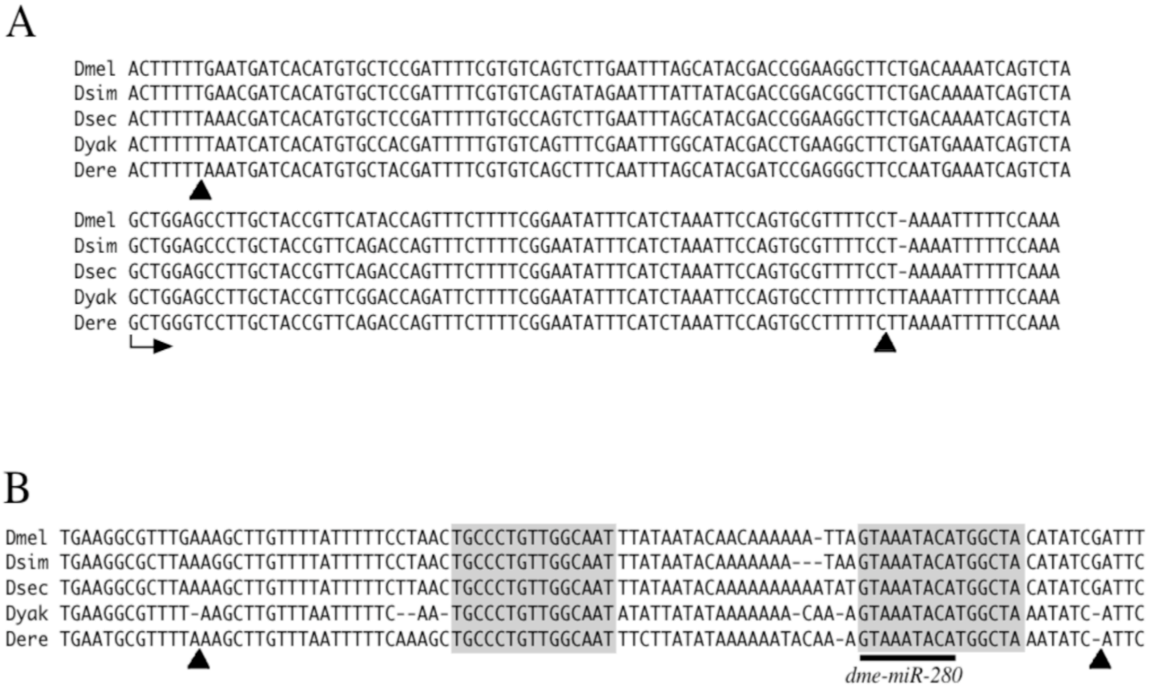


**Figure 1.1 Schematic diagram of the *ocn-lacZ* expresssion construct.** – (A) The *ocn* promoter fused to the *lacZ* open reading frame was inserted into the p*P[wFl]* transformation vector, which contains the white gene as a selectable marker. The boundaries of the DNA inserted into the *Drosophila* genome are indicated by "P". The portion of the plasmid used for replication in *E. coli* is labeled "pUC". (B) The p*P[YEStes-lacZ]* vector. The *ocn* promoter and 3′ UTR were fused to respective ends of the *lacZ* open reading frame and inserted into the YES transformation vector. Binding sites for the Suppressor of Hairy-wing protein, which functions as a chromosomal insulator, are labeled "S". (Figure designed by W. HENSE.)

transposable element and contains the *yellow* (*y*) gene of *D. melanogaster* as a selectable marker. Additionally, it contains binding sites for the *suppressor of Hairy-wing* protein that flank the inserted DNA and serve to insulate it from position effects caused by random insertion of the vector into the genome (PATTON *et al.* 1992). The resulting transformation vector was designated as YEStes (YES vector for testes specific expression) and contains the *ocn* promoter and 3' UTR separated by unique *Xba*I and *Not*I restriction sites. To complete the expression construct, a 3.5-kb *Not*I fragment of the plasmid pCMV-SPORT-βgal containing the complete *lacZ* open reading frame was cloned into the *Not*I site of the YEStes vector in the appropriate orientation. This final construct was designated p*P[YEStes-lacZ]* (Figure 1.1B).



```
A

Dmel ACTTTTTGAATGATCACATGTGCTCCGATTTTCGTGTCAGTCTTGAATTTAGCATACGACCGGAAGGCTTCTGACAAAATCAGTCTA
Dsim ACTTTTTGAACGATCACATGTGCTCCGATTTTCGTGTCAGTATAGAATTTATTATACGACCGGACGGCTTCTGACAAAATCAGTCTA
Dsec ACTTTTTAAACGATCACATGTGCTCCGATTTTTGTGCCAGTCTTGAATTTAGCATACGACCGGAAGGCTTCTGACAAAATCAGTCTA
Dyak ACTTTTTTAATCATCACATGTGCCACGATTTTTGTGTCAGTTTCGAATTTGGCATACGACCTGAAGGCTTCTGATGAAATCAGTCTA
Dere ACTTTTTAAATGATCACATGTGCTACGATTTTCGTGTCAGCTTTCAATTTAGCATACGATCCGAGGGCTTCCAATGAAATCAGTCTA

Dmel GCTGGAGCCTTGCTACCGTTCATACCAGTTTCTTTTCGGAATATTTCATCTAAATTCCAGTGCGTTTTCCT-AAAATTTTTCCAAA
Dsim GCTGGAGCCCTGCTACCGTTCAGACCAGTTTCTTTTCGGAATATTTCATCTAAATTCCAGTGCGTTTTCCT-AAAATTTTTCCAAA
Dsec GCTGGAGCCTTGCTACCGTTCAGACCAGTTTCTTTTCGGAATATTTCATCTAAATTCCAGTGCGTTTTCCT-AAAAATTTTTCAAA
Dyak GCTGGAGCCTTGCTACCGTTCGGACCAGATTCTTTTCGGAATATTTCATCTAAATTCCAGTGCCTTTTTCTTAAAATTTTTCCAAA
Dere GCTGGGTCCTTGCTACCGTTCAGACCAGTTTCTTTTCGGAATATTTCATCTAAATTCCAGTGCCTTTTTCTTAAAATTTTTCCAAA

B

Dmel TGAAGGCGTTTGAAAGCTTGTTTTATTTTTCCTAACTGCCCTGTTGGCAATTTATAATACAACAAAAAA-TTAGTAAATACATGGCTACATATCGATTT
Dsim TGAAGGCGCTTAAAGGCTTGTTTTATTTTTCCTAACTGCCCTGTTGGCAATTTATAATACAAAAAAA---TAAGTAAATACATGGCTACATATCGATTC
Dsec TGAAGGCGCTTAAAAGCTTGTTTTATTTTTCTTAACTGCCCTGTTGGCAATTTATAATACAAAAAAAAAATATGTAAATACATGGCTACATATCGATTC
Dyak TGAAGGCGTTTT-AAGCTTGTTTAATTTTTC--AA-TGCCCTGTTGGCAATATATTATATAAAAAAA-CAA-AGTAAATACATGGCTAAATATC-ATTC
Dere TGAATGCGTTTTAAAGCTTGTTTAATTTTTCAAAGCTGCCCTGTTGGCAATTTCTTATATAAAAAATACAA-AGTAAATACATGGCTAAATATC-ATTC
```

**Figure 1.2 Sequence alignment of the *ocn* promoter and 3′ UTR.** – (A) Alignment of the *ocn* 5′ flanking and 5' UTR sequences of *D. melanogaster*, *D. simulans*, *D. sechellia*, *D. yakuba*, and *D. erecta*. The arrowheads indicate the boundaries of the *ocn* promoter sequence included in our expression constructs. The transcriptional start site is indicated by an arrow. (B) Alignment of the *ocn* 3' UTR sequences of *D. melanogaster*, *D. simulans*, *D. sechellia*, *D. yakuba*, and *D. erecta*. The two conserved regions are shaded. The arrowheads indicate the boundaries of the 3′ UTR sequence included in our expression construct. (Preliminary data provided by J. PARSCH.)

### 1.1.2 Germline transformation

Plasmid DNA of the above expression constructs was purified using the QIAprep Spin kit (QIAGEN, Hilden, Germany) and used for microinjection of early stage embryos of the *y w*; Δ2-3, *Sb*/TM6 strain of *D. melanogaster* following standard procedures (SPRADLING and RUBIN 1982; RUBIN and SPRADLING 1982). Because it carries both the *y* and *w* mutations, this strain could be used for both transformation vectors. The Δ2-3 *P* element on the third chromosome served as source of transposase (ROBERTSON *et al.* 1988). Following transformation, all lines were crossed to a *y w* stock to remove the transposase source.

In cases where the transgene insertion was linked to the Δ2-3 source of transposase, the inserts were immediately re-mobilized by crossing transformed males to *y w* females and selecting offspring carrying the transgene, but not the Δ2-3 element. These flies were then mated to *y w* flies of the opposite sex to establish stable transgenic lines.

X-linked insertions were identified by crossing transformed males to *y w* females and following inheritance of the phenotypic marker ($y^+$ or $w^+$): crosses in which all daughters, but no sons, showed the marker phenotype revealed X linkage. Some X-linked insertions were mobilized to the autosomes by the following procedure. Transformed females were mated to *y w*; Δ2-3, *Sb*/TM6 males and the male offspring carrying both the transgene and the Δ2-3 source of transposase were mated to *y w* females. From this cross, we selected male offspring carrying the transgene (which could not be on the X chromosome inherited from the mother). These males were mated to *y w* females to establish stable transformed lines with new autosomal insertions of the transgene.

To map the intrachromosomal location of the transgene insertions, the genomic sequence flanking the *P*-element vector was determined by sequencing the products of inverse PCR (BELLEN *et al.* 2004). Briefly, genomic DNA was extracted from insertion-bearing flies and digested with either *Hpa*II or *Hin*P1I. The digestion products were self-ligated and used as a template for PCR with primer pairs Pry1 (CCTTA GCATG TCCGT GGGGT TTGAA T) / Pry2 (CTTGC CGACG GGACC ACCTT ATGTT ATT) and Plac1 (CACCC AAGGC TCTGC TCCCA CAAT) / Plac4 (ACTGT GCGTT AGGTC CTGTT CATT GTT) to determine 3’ or 5’ flanking sequences, respectively. PCR products were sequenced with BigDye v1.1 chemistry on a 3730 automated sequencer (Applied biosystems, Foster City, CA) using the PCR primers as sequencing primers. In all cases, the chromosomal locations assigned by inverse PCR were consistent with those determined by genetic crosses.

### 1.1.3 β-galactosidase assays

To determine *in vivo* expression levels of our transgenic constructs, we measured the level of β-galactosidase activity in transformed flies. For all autosomal insert lines, transformed males were mated to *y w* females and offspring heterozygous for the transgene insertion were used for assays. For transformants with X-linked inserts, females were mated to *y w* males and offspring heterozygous (female) or hemizygous (male) for the transgene insertion were used for assays. In all cases, the offspring were collected shortly after eclosion and separated by sex until they were assayed at age 5-7 days. All flies were raised on cornmeal-molasses medium at 25 °C.

For assays of β-galactosidase activity, five adult flies of the same sex were homogenized in 150 µl of a buffer containing 0.1M tris-HCl, 1mM EDTA, and 7mM 2-mercaptoethanol at pH 7.5. After incubation on ice for 15 min, the homogenates were centrifuged at 12,000 g for 15 min at 4 °C and the supernatant containing soluble proteins was retained. For each assay, 50 µl of this supernatant were combined with 50 µl of assay buffer [200 mM sodium phosphate (pH 7.3), 2 mM $MgCl_2$, 100 mM 2-mecaptoethanol] containing 1.33 mg/ml *o*-nitrophenyl-β-D-galactopyranoside. β-galactosidase activity was measured by following the change in absorbance at 420 nm over 30 min at 25 °C. β-galactosidase activity units were quantified as the change in absorbance per minute multiplied by 1000 (mOD/min). For all transformed lines, we performed at least two technical and two biological replicates (always in equal numbers), where the former used the same soluble protein extraction and the latter used extractions from independent cohorts of flies. The activity of each line was calculated as the mean over all replicates, with the variance and standard error calculated among replicates. For comparisons between chromosomes or vectors, we averaged over the means of the individual lines and used the among-line variation to calculate variance, standard error, and CV. This approach is conservative, as the among-line differences (position effects) tended to be the largest source of variation. Statistical tests for differences between groups were performed using non-parametric methods, such as the Mann-Whitney *U* test, that do not rely on estimates of variance. For our purposes this approach is conservative.

For lines that showed β-galactosidase activity in adult males, we also performed assays on gonadectomized males. This was done following the above protocol, after removal of the testes by manual dissection. For visualizing β-galactosidase activity in whole tissues, we incubated dissected testes in the above assay buffer containing 1 mg/ml ferric ammonium citrate and 1.8 mg/ml of S-GAL sodium salt (Sigma-Aldrich, Munich, Germany) for 6 hours at 37 °C.

### 1.1.4 Quantitative reverse-transcription PCR (qRT-PCR)

To measure expression at the level of transcription (mRNA abundance), we performed qRT-PCR using a TaqMan assay (Applied Biosystems, Foster City, CA) designed specifically to our transgene (*i.e.*, spanning the junction between the *ocn* 5' UTR and the *lacZ* coding region). For this, 1 μg of DNase I-treated total RNA isolated from heterozygous (autosomal insertions) or hemizygous (X insertions) males was reverse transcribed using Superscript II reverse transcriptase and random hexamer primers (Invitrogen) according to the manufacturer's protocol. A 1:10 dilution of the resulting cDNA was used as template for PCR on a 7500 Fast Real-Time PCR System (Applied Biosystems). The average threshold cycle value (Ct) was calculated from two technical replicates per sample. Expression of the transgene was standardized relative to the ribosomal protein gene *RpL32* (*CG7939*, TaqMan probe ID Dm02151827). Relative expression values were determined by the ΔΔCt method according the formula $2^{-(\Delta Ct_{x} - \Delta Ct_{min})}$, where $\Delta Ct_x = Ct_{transgene} - Ct_{RpL32}$ for a given transformed line *x*, and $\Delta Ct_{min}$ represents the corresponding value of the line displaying the lowest level of transgene relative to *RpL32* expression. Statistical analyses were performed as described above for β-galactosidase activity.

## 1.2 RESULTS

### 1.2.1 Identification and functional analysis of the *ocn* promoter

The *ocn* gene is expressed specifically in testis and encodes a protein abundant in mature sperm (DORUS *et al.* 2006; PARSCH *et al.* 2001). It is part of a cluster of three tandemly duplicated genes on chromosome arm 3R that are present in all species of the *D. melanogaster* species subgroup and shares greatest homology to the neighboring *janusB* (*janB*) gene, which is also expressed in testis. Although *ocn* lies only 250 bp distal to *janB*, it produces a unique transcript that does not overlap with that of *janB* (PARSCH *et al.* 2001). The first half of the *janB*-*ocn* intergenic region is highly diverged among species of the *D. melanogaster* subgroup and cannot be aligned unambiguously. However, the portion just upstream of the *ocn* start codon is well conserved, suggesting that it has regulatory function (Figure 1.2). We refer to this region as the *ocn* promoter. To test its ability to drive tissue-specific gene expression, we fused it to the open reading frame of the *Escherichia coli lacZ* gene, which encodes the enzyme β-galactosidase (Figure 1.1A). Transgenic flies with autosomal insertions of *P[wFl-ocn-lacZ]* showed reporter gene expression specifically in testis, as expected (Figure 1.3).

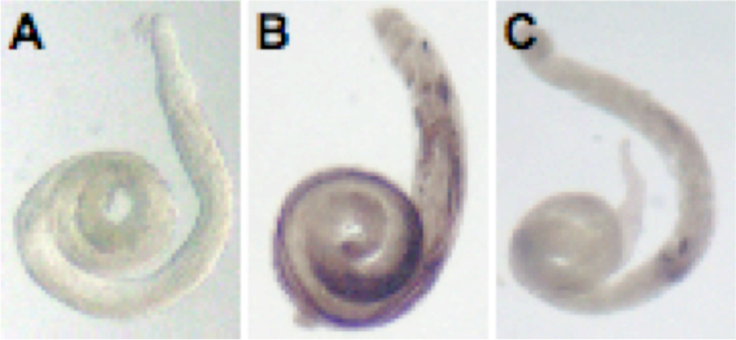


**Figure 1.3 Reporter gene expression in testes.** – Testes were dissected and incubated with S-GAL, which forms a black precipitate in the presence of β-galactosidase. Shown are testes from *y w* males (negative control) (A), *y w* males with an autosomal insertion of *P[wFl-ocn-lacZ]* (B), and *y w* males with an X-linked insertion of *P[wFl-ocn-lacZ]* (C). Staining was performed in parallel for the same length of time. The strongest signal is in the proximal region of the autosomal-insert testis. Note that weak staining is visible in the proximal region of the X-insert testis. (Testis dissection and staining performed by W. HENSE; pictures taken by W. HENSE together with J. PARSCH.)

### 1.2.2 Comparison of autosomal and X-linked insertions

Overall, we obtained 15 independent autosomal insertions of *P[wFl-ocn-lacZ]*. The mean β-galactosidase activity in adult males was 8.67 units, while that in adult females was 0.34 units. The difference between the sexes was highly significant (Mann-Whitney *U* test, $P < 0.001$). The mean β-galactosidase activity of gonadectomized males was 0.24 units, which was significantly less than whole males (Mann-Whitney *U* test, $P < 0.01$).

If the X chromosome is inactivated before the autosomes during spermatogenesis, then one would expect transgenic lines with X-linked insertions of *P[wFl-ocn-lacZ]* to show lower levels of reporter gene expression than those with autosomal insertions. This is indeed what we observe. In total, we obtained 10 independent X-linked insertions of *P[wFl-ocn-lacZ]*. All of these lines showed reduced β-galactosidase activity in adult males relative to the autosomal-insertion lines (Figures 1.3 and 1.4). On average, the activity difference between autosomal and X-linked insertions was 7-fold (8.67 *versus* 1.19 units), and the difference between the two groups was highly significant (Mann-Whitney *U* test, $P < 0.001$). Although β-galactosidase activity was very low for the X-linked insertions, it was significantly greater than zero. Assuming a normal distribution of activity among the X-insertion lines, the 95% confidence interval was 0.82–1.56 units. Five of the autosomal insertion lines (the last five in Figure 1.4) were obtained through the re-mobilization of X-linked inserts (see MATERIALS AND METHODS), demonstrating that the reduction in expression was not caused by undesired sequence changes in the *ocn* promoter or *lacZ* coding sequence, but instead was a direct result of X-linkage.

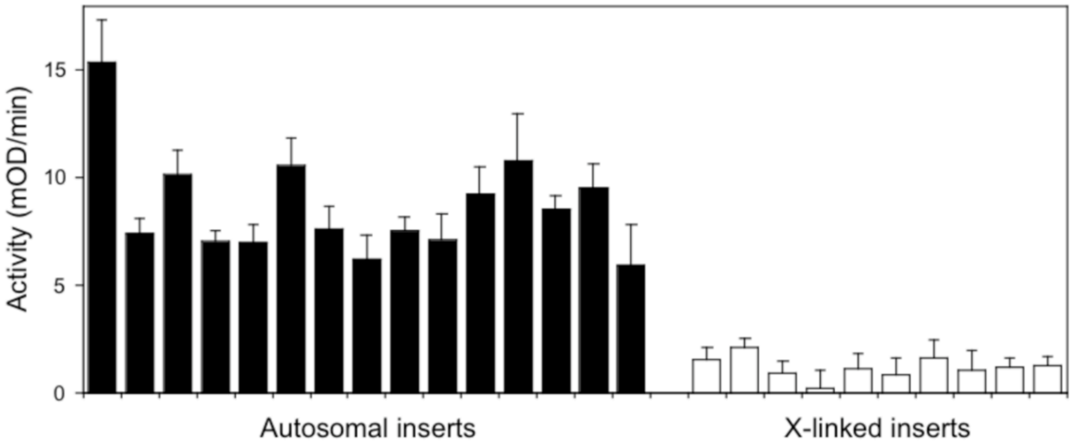


**Figure 1.4 Average β-galactosidase activity of adult male flies with autosomal (solid bars) or X-linked (open bars) insertions of *P[wFl-ocn-lacZ]*.** – Each bar represents a different transformed line with a unique, independent transgene insertion. Error bars indicate the standard error of the mean, calculated from the variance among all replicate measurements within each independent insertion line. (Assays performed and figure designed by W. HENSE.)

Because the assays of β-galactosidase activity measure expression at the level of protein abundance, it is possible that they do not reflect underlying levels of transcription. To test this, we performed quantitative reverse-transcription PCR (qRT-PCR) to estimate the relative transcript abundance of a subset of eight transformed lines, including four with autosomal and four with X-linked inserts. The autosomal inserts had significantly higher transgene expression at the level of mRNA (Mann-Whitney *U* test, $p = 0.02$), with the relative expression difference being 5-fold (Figure 1.5B), which corresponds well to the observed difference in β-galactosidase activity and suggests that the enzymatic assays provide a reliable estimate of expression.

### 1.2.3 Effect of chromosomal insulator sequences

To test if the reduced expression of the X-linked *ocn-lacZ* transgenes could be attributed to the presence of localized transcriptional repressors bound to the X chromosome, we performed additional experiments using the *P[YEStes-lacZ]* transformation vector (Figure 1.1B), which contains binding sites for the *suppressor of Hairy-wing* protein. These binding sites flank the inserted transgene and serve to insulate it from the effects of external transcriptional regulators (PATTON *et al.* 1992). We obtained 12 independent autosomal insertions of *P[YEStes-lacZ]* and these lines showed male- and testis-specific expression of the *lacZ* reporter gene. The mean β-galactosidase activity in adult males was 1.84 units, which was significantly greater than that of adult females (mean = 0.42; Mann-Whitney *U* test, $P < 0.001$) or gonadectomized males (mean = 0.22; Mann-Whitney *U* test, $P < 0.001$).

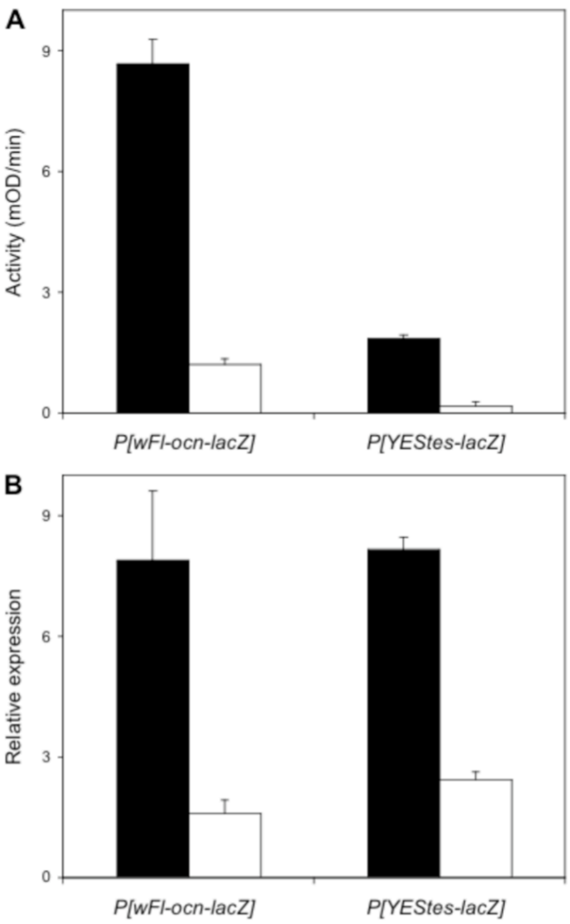


**Figure 1.5 Expression levels of autosomal (solid bars) and X-linked (open bars) insertions of the two *ocn-lacZ* transformation vectors shown in Figure 1.1 (designed by W. HENSE)**

(A) Average β-galactosidase activity of adult males. (Assays performed by W. HENSE.)

(B) Relative expression measured by qRT-PCR. Transcript abundance was standardized to that of the ribosomal protein gene *RpL32* and is given in arbitrary units. Error bars indicate the standard error of the mean, calculated from the variance among the means of the independent insertion lines. (Done in collaboration with J. BAINES.)

We also obtained 10 independent insertions of *P[YEStes-lacZ]* on the X chromosome. Adult males of these lines had a mean β-galactosidase activity of 0.17 units, which differed significantly from the autosomal-insert lines (Mann-Whitney *U* test, $P < 0.001$), but did not differ significantly from zero (95% confidence interval = -0.09–0.43). The reduction in reporter β-galactosidase activity caused by X linkage was >10-fold (Figure 1.5A). We also assayed expression at the level of transcript abundance by performing qRT-PCR on a subset of eight transformed lines (four with autosomal and four with X-linked inserts). Again, the X chromosome insertion lines showed significantly less transgene expression than the autosomal insertion lines (Mann-Whitney *U* test*,* $p = 0.02$). The reduction in reporter gene expression measured by qRT-PCR was 3.4-fold (Figure 1.5B). Thus, the presence of the chromosomal insulator sequences did not alleviate transcriptional repression of the X-linked transgenes.

For adult males with autosomal insertions, the coefficient of variation (CV) for β-galactosidase activity was lower among the *P[YEStes-lacZ]* transformed lines (CV = 0.16) than among the *P[wFl-ocn-lacZ]* transformed lines (CV = 0.28). A more pronounced difference was seen at the level of mRNA abundance, where the CVs for *P[YEStes-lacZ]* and *P[wFl-ocn-lacZ]* transformants were 0.07 and 0.44, respectively. This suggests that the insulator sequences successfully reduced position effect variation caused by the chromosomal context of the insertion. The *P[YEStes-lacZ]* transformants, however, showed significantly less β-galactosidase activity than the *P[wFl-ocn-lacZ]* transformants (Mann-Whitney *U* test, $P < 0.001$; Figure 1.5A). Interestingly, this difference was not detectable at the level of mRNA abundance (Figure 1.5B), which suggests additional, post-transcriptional regulation of the *P[YEStes-lacZ]* transgenes.

## 1.3 Discussion

Although a number of hypotheses regarding genome and sex chromosome evolution assume that the *Drosophila* X chromosome becomes transcriptionally inactive before the autosomes during spermatogenesis, little direct evidence for this has been reported. Our experimental results indicate that X chromosome inactivation does occur in *Drosophila* and that it can have a considerable effect on gene expression in the male germline. In total, we examined 27 autosomal and 20 X-linked insertions of a testis-specific reporter gene in two different transformation vectors. In all cases, transformed lines with autosomal insertions showed significantly greater transgene expression than their X-linked counterparts, with the differences in expression ranging from 3.4- to 10-fold. The consistency of these results across a large number of independent insertions suggests that this transcriptional inactivity is a global property of the X chromosome. The fact that we observe the same pattern when using a vector that insulates the transgene from external transcriptional regulators further suggests that inactivation of the X chromosome in the male germline occurs through a major structural change, rather than by the binding of localized transcriptional repressors.

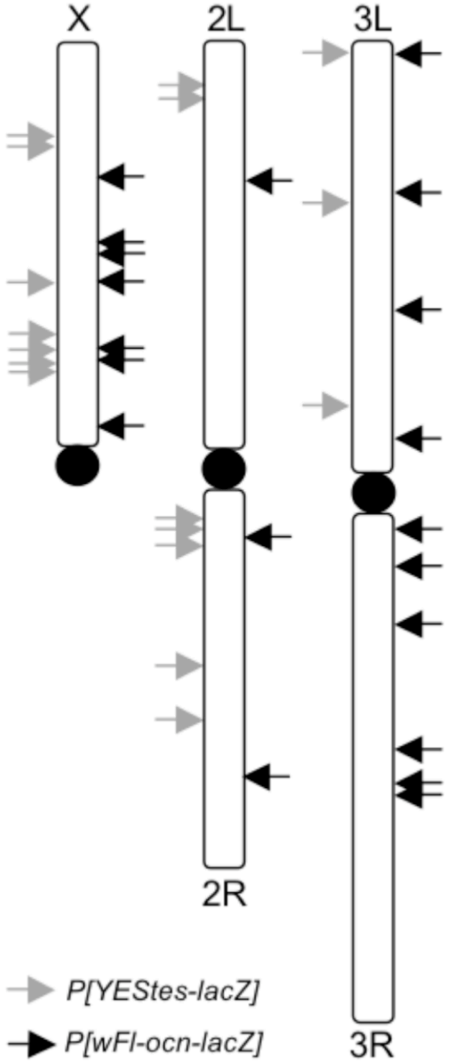


**Figure 1.6 Chromosomal location of the transgene insertions.**
Arrows indicate the insertion sites of *P[wFl-ocn-lacZ]* (black) and *P[YEStes-lacZ]* (gray) transgenes as determined by inverse PCR. Nine additional inserts could be assigned only to the X chromosome or autosomes by genetic crosses and are not shown. (Done in collaboration with J. BAINES.)

Could our results be explained by something other than male germline X inactivation? One possibility is that there is an insertional bias of our transgenes that differs between the X chromosome and the autosomes. For example, X-linked inserts could preferentially target inactive or heterochromatic regions. To investigate this, we used inverse PCR to map the insertion sites (Figure 1.6). We find that the insertions span the euchromatic regions of the X and autosomes, with many being in or near genes (Table 1.1). Thus, our mapping results run counter to the expectations of insertional bias as a cause of the observed differences in

## Table 1.1

**Chromosomal locations of transgene insertions (compiled by J. BAINES)**

| Line | Chrom | Cytological<br>Band | Coordinate (v5.1) | Location | Comment | Proximal gene | Distal gene |
|---|---|---|---|---|---|---|---|
| wol12X | X | 7B1 | 7231447* | intergenic | | *CG18155* | *CG1435* |
| wol21X | X | 10E3 | 11699401 | *CG4147* | in exon | | |
| wol13X | X | 11E3 | 13101216 | *CG1903* | in intron | | |
| wol23X | X | 19F1 | 20994197 | intergenic | | *CG15445* | *CG34120* |
| wol24X | X | 10E3 | 11687344* | *CG15224* | in intron | | |
| wol20X | X | 15A7 | 16677891* | intergenic | | *CG9623* | *CG4742* |
| wol19X | X | 16A1 | 17197389* | *CG5445* | in exon | | |
| wol25X | X | n.m. | | | | | |
| wol5X | X | n.m. | | | | | |
| wol4 | 2L | 27F4 | 7423613 | intergenic | | *CG5261* | *CG5229* |
| wol7 | 2R | 42C6 | 2603250 | *CG3409* | in exon | | |
| wol9 | 2R | 56E1 | 15518667* | *CG9218* | in exon | | |
| wol11 | 3L | 61C9 | 749342 | intergenic | | *CG13897* | *CG1007* |
| wol6 | 3L | 66C12 | 8414592 | intergenic | | *CG32354* | *CG7037* |
| wol18 | 3L | 70F4 | 14751002* | *CG33261* | in exon | | |
| wol16 | 3L | 79A2 | 21872663 | intergenic | | *CG14563* | *CG7437* |
| wol2 | 3R | 82E4 | 790802* | heterochrom | | | |
| wol1 | 3R | 84B1 | 2799036* | intergenic | | *CG41463* | *CG41464* |
| wol14 | 3R | 85F10 | 5920571* | intergenic | | *CG5361* | *CG6203* |
| wol3 | 3R | 89E11 | 12882012 | CG5201 | in intron | | |
| wol15 | 3R | 91D4 | 14743978 | CG17836 | in exon | | |
| wol17 | 3R | 91F4 | 14983880* | CG11779 | in intron | | |
| wol10 | 2 | n.m. | | | | | |
| wol8 | 2 | n.m. | | | | | |
| ylz22X | X | 4F9 | 5312216 | *CG3249* | in intron | | |
| ylz9X | X | 5A12 | 5574020 | intergenic | | *CG3171* | *CG15779* |
| ylz20X | X | 11E1 | 13022326 | *CG32368* | in exon | | |
| ylz15X | X | 14A1 | 15834425 | *CG9126* | in intron | | |
| ylz19X | X | 16A1 | 17195978 | *CG8649* | in exon | | |
| ylz18X | X | 16A1 | 17196628 | *CG5445* | in exon | | |
| ylz17X | X | 16B10 | 17552922 | *CG5870* | in intron | | |
| ylz16X | X | n.m. | | | | | |
| ylz21X | X | n.m. | | | | | |
| ylz23X | X | n.m. | | | | | |
| ylz6 | 2L | 23A3 | 2753160 | *CG9894* | in intron | | |
| ylz4 | 2L | 24C2 | 3730445 | intergenic | | *CG2822* | *CG10019* |
| ylz11 | 2R | 41F9 | 1642051 | *CG12792* | in exon | | |
| ylz10 | 2R | 42C3 | 2549792 | *CG15845* | in exon | | |
| ylz5 | 2R | 43E16 | 3670803 | *CG1555* | in exon | | |
| ylz7 | 2R | 50A13 | 9389601 | *CG6033* | in exon | | |
| ylz3 | 2R | 53F8 | 12984754 | *CG8938* | in intron | | |
| ylz12 | 3L | 61C9 | 749342 | intergenic | | *CG13897* | *CG1007* |
| ylz13 | 3L | 66D8 | 8609567 | *CG6282* | in exon | | |
| ylz8 | 3L | 76C5 | 19784609 | *CG8742* | in exon | | |
| ylz1 | 2 | n.m. | | | | | |
| ylz2 | 3 | n.m. | | | | | |

| | |
|---|---|
| n.m. | = not mapped by inverse PCR |
| * | = approximate location, precise insertion site not obtained |
| wol | = *P[wFl-ocn-lacZ]* |
| ylz | = *P[YEStes-lacZ]* |

expression. Another possibility is that insertion of the transgenes onto the X chromosome may cause rearrangements or other disruptions to the gene or promoter that prevent proper expression. However, by re-mobilizing multiple, independent X inserts to new autosomal locations, we have shown that their expression can be restored. Thus, the X-linked insertions must have been intact. Finally, a lack of proper dosage compensation of transgenes inserted onto the X chromosome could possibly lead to reduced expression. We consider this unlikely for two reasons. First, X chromosome dosage compensation has been shown to occur on a global level in the *Drosophila* germline (GUPTA *et al.* 2006). Second, the expression assays for the autosomal-insert lines were performed on flies heterozygous for the insertion. Thus, even if dosage compensation did not occur, we would expect to observe equal expression of X-linked and autosomal transgenes. Any degree of dosage compensation would result in higher activity in the X-insertion lines, which makes our test conservative.

The use of the *ocn* promoter may make our experimental system especially sensitive to the effects of male germline X inactivation for two reasons. First, the promoter fragment used here is rather short (150 bp) and, thus, may be abnormally influenced by differences in chromatin environment between the autosomes and the X chromosome. It should be noted, however, that other known testis-specific promoters are also relatively short, in the range of 76-390 bp (MICHIELS *et al.* 1989; YANICOSTAS and LEPESANT 1990; NURMINSKY *et al.* 1998). Second, *ocn* is likely to be expressed relatively late in spermatogenesis, where the effects of X inactivation should be pronounced. The *ocn* gene was originally identified as one encoding a protein abundant in the testes of mature males, but absent from those of immature males (PARSCH *et al.* 2001). Our observation that β-galactosidase activity imparted by the *ocn-lacZ* transgenes is greatest in proximal regions of the testis (Figure 1.3) also supports its relatively late expression. Furthermore, levels of β-galactosidase activity, as well as transgene transcript abundance as measured by qRT-PCR, are at least 50-fold lower in the third larval instar stage, where spermatogenesis is not yet complete, than in adult males (not shown). Thus, it may be that a large proportion of *ocn* expression occurs after the X chromosome is inactivated. Indeed, if X-linked genes expressed early in spermatogenesis are hypertranscribed through a dosage compensation mechanism (GUPTA *et al.* 2006), the effects of later X inactivation may be masked. Finally, we wish to point out that, although testis-expressed genes are underrepresented on the X chromosome, they are not absent. Thus, many X-linked genes involved in spermatogenesis must be expressed at levels sufficient for proper function. This may be a result of their (hyper)transcription early in spermatogenesis. Recently, it has been noted that a region of the X chromosome is enriched for newly-evolved, testis-expressed genes (LEVINE *et al.* 2006; BEGUN *et al.* 2007; CHEN *et al.* 2007), which suggests that this region may escape germline X inactivation. One of our transgene inserts falls within ~500 kb of this interval, but does not differ in expression from other X-linked insertions. A higher density of X-linked transgene insertions may reveal specific regions that escape inactivation.

Overall, *P[YEStes-lacZ]* transformants had much lower β-galactosidase activity than *P[wFl-ocn-lacZ]* transformants (Figure 1.5A). This difference was not observable at the level of mRNA (Figure 1.5B), suggesting additional regulation at the level of translation. There are two major differences between the vectors that could account for this. The first is the *suppressor of Hairy-wing* chromosomal insulator sequences in *P[YEStes-lacZ]* (Figure 1.1). However, it seems unlikely that these insulator sequences, which lie far outside of the transcriptional unit, would be involved in posttranscriptional regulation. Furthermore, putting the transgenes into a genetic background homozygous for a mutant *suppressor of Hairy-wing* allele had no effect on levels of β-galactosidase activity (Figure 1.7). The second difference is that *P[YEStes-lacZ]* contains the *ocn* 3' untranslated region (UTR) (Figure 1.1). Although functional information for this 3' UTR is lacking, the presence of two conserved sequence blocks suggests that it may play a role in the regulation of expression (Figure 1.2).

Our finding that a testis-specific gene is not properly expressed when located on the X chromosome provides compelling experimental evidence for male germline X inactivation in *Drosophila*, something that was first proposed over thirty years ago (LIFSCHYTZ and LINDSLEY 1972). It is also consistent with a selective explanation for the overabundance of retrotransposed genes that have moved from the X to the autosomes (BETRÁN *et al.* 2002). If such genes have a beneficial effect when expressed in testis (especially in later stages of spermatogenesis), then selection would favor the maintenance of autosomal copies. The acquisition of expression late in spermatogenesis may even predispose a gene to adaptive evolution, as testis-expressed genes appear to be targets of positive selection more often than genes of other expression classes (PRÖSCHEL *et al.* 2006). Our results also have relevance to the SAXI hypothesis (WU and XU 2003), which proposes that sexual antagonism leads to the selective relocation of male-beneficial genes expressed late in spermatogenesis to the autosomes. After all such genes have been relocated, selection could favor global inactivation of the X chromosome during spermatogenesis to prevent the expression of female-beneficial genes that have a harmful effect when expressed in males. Alternately, the X may be inactive at this stage simply because it no longer contains genes with the proper regulatory sequences required for male germline expression. Our results are consistent with the former scenario, as the *ocn* promoter, which drives testis-specific expression on autosomes, does not function properly when relocated to the X chromosome.

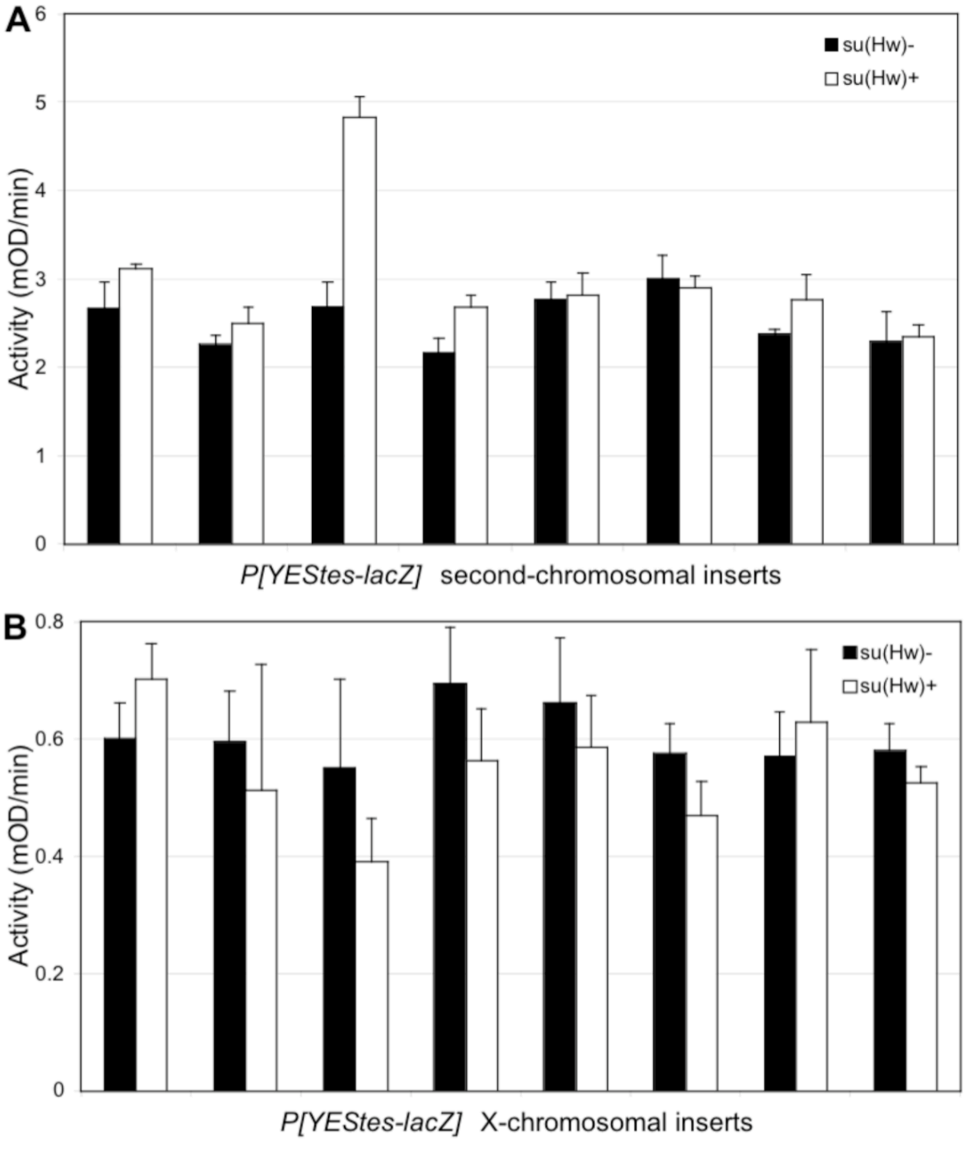


**Figure 1.7 Effect of *suppressor of Hairy-wing* genetic background on transgene expression.** The β-galactosidase activity imparted by the transgenes was measured in a background where the third chromosome was homozygous for either the mutant *su(Hw)*$^{8}$ allele (solid bars) or the wild-type allele (open bars). (A) Activity comparison of eight heterozygous second-chromosomal insertions (done by W. HENSE). (B) Activity comparison of eight hemizygous X-chromosomal insertions (done by J. BAINES). In both cases, the genetic background had no significant effect on activity (two-tailed Wilcoxon signed ranks test, $p > 0.10$). Error bars indicate the standard error of the mean, calculated from the variance among all replicate measurements within each independent insertion line.

# Chapter 2

# The contribution of *cis*-regulatory polymorphism to intraspecific expression variation of the *Drosophila melanogaster CG13360* gene

EVOLUTION, the process that shaped all existing life on Earth, requires heritable molecular variation that gives rise to phenotypic diversity upon which natural selection can act. Typically, the first step in transforming the molecular variation encoded in the genotype into an organism's phenotype is the temporally and spatially regulated production of an mRNA transcript of a gene to initiate its expression. The ultimate source of variation is mutation, which is essentially stochastic by nature. There are two fundamentally distinct types of mutations that may influence the phenotype: structural mutations, which cause an amino acid change in a protein, and regulatory mutations, which modify the amount of protein produced. Therefore, it is possible that a particular phenotypic change can be achieved by different modes. For instance, an increased rate of an enzymatic reaction can be caused by either a more effective enzyme (due to a structural change near the catalytic active site) or by the presence of a higher amount of enzyme (due to a regulatory change). KING and WILSON (1975) were among the first to propose that many phenotypic changes between organisms are not caused by structural mutations in the coding sequence of genes, but instead by regulatory changes, which may also play a substantial role in adaptation and speciation (see also BRITTEN and DAVIDSON 1969). Their hypothesis came mainly from the observation that humans and chimpanzees are around 99% identical at the DNA level, whereas their phenotype (including morphology and cognitive capability) differs considerably.

Because DNA sequencing techniques had been developed more rapidly than methods to quantify gene expression, research of the last decades mainly focused on sequence variation among and between species and its interplay with evolutionary forces such as natural selection, sexual selection, recombination, demography, and on the background level of DNA sequence variation caused by mutation and random genetic drift. This finally culminated in a wealth of DNA polymorphism and divergence data that could be used to test the neutral theory of molecular evolution (KIMURA 1968, 1983; for tests see *e.g.* TAJIMA 1989; HUDSON *et al.* 1987; MCDONALD and KREITMAN 1991; reviews of NIELSEN 2005; NIELSEN *et al.* 2007). It only recently became technically possible to include gene expression data into evolutionary analysis. Using microarrays, the extent of variation in transcript abundance has been surveyed in several taxa including *Drosophila*, fish, yeast, and humans (CAVALIERI *et al.* 2000; TOWNSEND *et al.* 2003; JIN *et al.* 2001; OLEKSIAK *et al.* 2002; LO *et al.* 2003; ENARD *et al.* 2002; MEIKLEJOHN *et al.* 2003; HUTTER *et al.* 2008). Further microarray studies identified expression differences between the two sexes (PARISI *et al.* 2003; RANZ *et al.* 2003). All of these studies revealed extensive variability at the level of mRNA abundance, raising questions about the forces governing and maintaining such variation. Despite a broad capacity for rapid gene expression evolution (RIFKIN *et al.* 2005), there is evidence for pervasive stabilizing selection for an optimal transcript level of most genes (LEMOS *et al.* 2005). Moreover, there are expression data that support a neutral theory of gene expression evolution, as interspecific expression often appears to diverge in a clock-like fashion (KHAITOVICH *et al.* 2006; WITTKOPP *et al.* 2008).

Once the extent of gene expression variation became clear, the immediate question was which specific molecular mechanisms are responsible for it and how this variation is reflected at the DNA level, as some part of the expression differences between individuals of a species was shown to be heritable (reviewed by WRAY *et al.* 2003). Among the first explanatory principles to account for differences in gene expression were *cis*-regulatory sequences nearby the regulated gene and *trans* factors that bind *cis*-regulatory DNA to influence transcription initiation, but are, however, encoded elsewhere in the genome. Recent studies in *Drosophila* addressing the relative contribution of *cis* and *trans* factors to expression divergence found that *cis*-regulatory sequences can act allele-specifically and independently of *trans* factors on gene expression, whereas *trans* factors that affect expression are always accompanied by changes in *cis* (WITTKOPP *et al.* 2004). These studies examined allele-specific gene expression by pyro-sequencing of cDNA in a hybrid genetic background representing all *trans* factors, thus they were not able to determine specific polymorphisms responsible for differential expression, neither in *cis* nor in *trans*. Furthermore, on the same chromosomal or genomic scale ANDOLFATTO (2005) found evidence for adaptive evolution of non-coding DNA in *Drosophila*, intergenic DNA that is thought to harbor a lot of functionally significant regulatory sequences. He estimated that 40 to 70% of non-coding DNA is evolutionary constrained relative to synonymous sites, and that up to 60% of nucleotide divergence in these regions was driven to fixation by positive selection. Thus, adaptive changes in non-coding DNA may be more prevalent than those in proteins. Moreover, ANDOLFATTO (2005) extended the statistical test of MCDONALD and KREITMAN (1991) to the analysis of non-coding DNA, which has also been done with several other such tests (reviewed by HAHN 2007).

A more recent study of WANG *et al.* (2008), however, analyzed chromosome-substitution lines of two behavioral races of *D. melanogaster* and found that as little as 3% of differentially expressed genes are purely *cis*-regulated, mainly because around 80% of expression differences are controlled by at least two chromosomes. The fraction of *cis*-regulated genes rises to about 14% if additional *trans* regulation (either additive or epistatic) is included (and even to 32% if strongly differentiated genes are considered). This suggests that, at least in intraspecific comparisons, both *trans* effects and *cis*-by-*trans* effects play a major role in gene regulation. These results contrast to those of WITTKOPP *et al.* (2004), which focused on interspecific expression differences. This suggests that there might be a significant difference of *cis* and *trans* effects in their contribution to inter- and intraspecific expression variation (WITTKOPP *et al.* 2008). By investigating dominance relationships and their effect on differential gene expression of *D. melanogaster* populations, LEMOS *et al.* (2008) could confirm the latter. They also used chromosome-substitution lines to measure gene expression in homozygous and heterozygous flies. When an expression difference between two homozygous is not found in a heterozygote, it is assumed that of one of the alleles is recessive. More than 70% of differentially expressed genes surveyed by LEMOS *et al.* (2008) showed this feature. In addition, expression variation due to *trans* factors exhibits greater deviations from additivity because of dominant/recessive alleles, whereas it is this greater additivity present in the *cis*-regulation of genes that allows them to be a better subject for natural selection. If natural selection later drives speciation, these findings would altogether confirm an important role of *cis* sequences in interspecific expression divergence, while intraspecific expression evolution would be mainly left to the realm of *trans* effects.

In addition to experiments that assess the relative importance of *cis* and *trans* regulation, it is nowadays possible to determine the selective constraints operating on non-coding DNA and even to detect evidence of past positive selection in non-coding DNA. However, neither of the above approaches allows for an in-depth investigation of *cis*-regulatory DNA elements. This requires additional experimental work. Such experimental studies have revealed that regulatory sequences are usually short in length, degenerate in their sequence, and variably located relative to the transcription start site (WRAY *et al.* 2003), which makes them extremely difficult to track down. Furthermore, recent studies have found that there might be sequence conservation despite functional divergence and, vice versa, functional convergence when sequence similarity is absent (WITTKOPP 2006; HARE *et al.* 2008). Despite these rather disappointing results, scientists in the field of "evo-devo" gathered some inspiring examples of morphological and/or physiological evolution, many of which turned out to be driven by *cis*-regulatory DNA changes (see *e.g.* for *Drosophila* GOMPEL *et al.* 2005; PRUD'HOMME *et al.* 2006; for stickleback fish SHAPIRO *et al.* 2004; for humans HAMBLIN and DI RIENZO 2000), thereby opening again the polarizing question of whether *cis*-regulatory and coding sequence mutations make qualitatively different contributions to phenotypic evolution (*e.g.* STERN 2000; CARROLL 2005; HOEKSTRA and COYNE 2007).

In this study we employed transgenics in *D. melanogaster* to examine the effect of two alternative versions of a putative *cis*-regulatory promoter in the expression of a reporter gene. To do this, we selected a gene of *D. melanogaster* (*CG13360*) whose expression level showed a bimodal distribution among eight laboratory strains (MEIKLEJOHN *et al.* 2003). We fused the *CG13360* upstream region from two of these strains, which differed in expression as well as DNA sequence at multiple sites, to the *E. coli* gene that encodes β-galactosidase and compared their *in vivo* enzymatic activities in transgenic flies. For this we utilized the "waffle" transformation vector of SIEGAL and HARTL (1996) to overcome position-effect variation, which notoriously limits the sensitivity of transgenic experiments. Since the genetic background including all relevant *trans* factors was identical in all flies we used, we were able to estimate the relative contribution of the promoter region to observed intraspecific expression differences. Overall, we observed only a minor difference in gene expression caused by the different promoter sequences, and only in particular genetic backgrounds. This suggests that there is a slight *cis*-by-*trans* effect influencing *CG13360* expression variation, but most of the intraspecific variation is not caused by *cis*-regulatory variation in the proximal promoter.

## 2.1 MATERIALS AND METHODS

### 2.1.1 Fly strains

We used eight strains of *D. melanogaster*, which we grouped into two populations, an African population from Zimbabwe consisting of *Zim53*, *Zim(S)2*, *Zim29*, and *Zim30* (highly inbred), and a non-African, "cosmopolitan" population from various locations in the USA (two laboratory strains: *Canton-S* (*Can-S*) and *Oregon-R* (*Ore-R*); one isofemale line from St. Louis (*StL*)) and one strain from Japan (*Hikone-R* (*Hik-R*)). The gene expression of gene *CG13360* in these eight strains as measured by the study of MEIKLEJOHN *et al.* (2003) was the starting point for our study (Figure 2.1). All flies were raised on cornmeal-molasses medium at 25 °C.

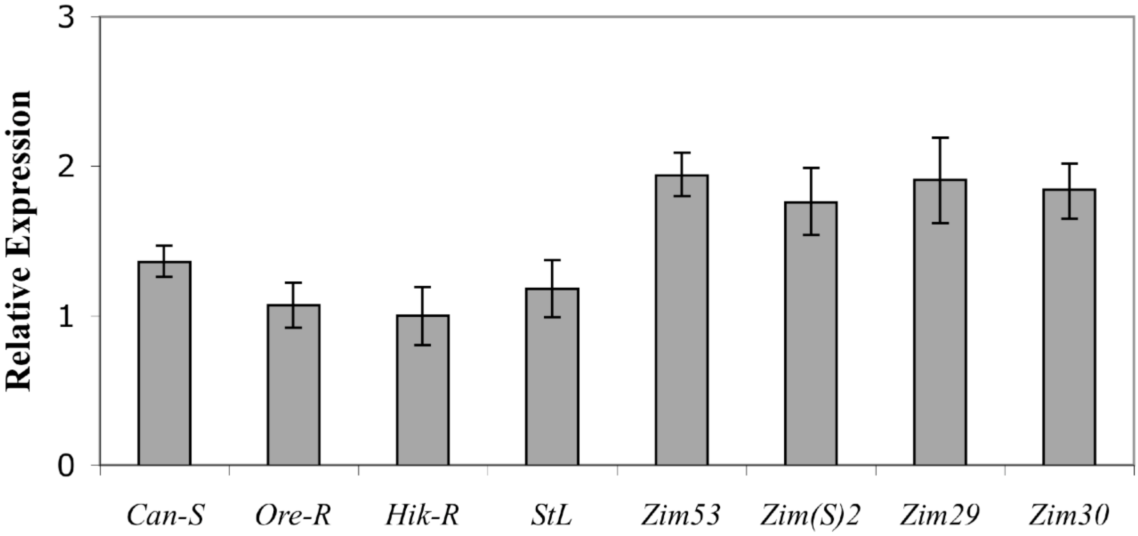


**Figure 2.1 Microarray expression results results of gene *CG13360*.** – A previous microarray survey of *Drosophila melanogaster* measured gene expression in eight strains from Zimbabwe, Africa, the U.S., and Japan. The four strains from Zimbabwe showed significantly higher expression than each of the remaining "cosmopolitan" strains. Error bars show 95% CI; *Can-S*: *Canton-S*, *Ore-R*: *Oregon-R*, *Hik-R*: *Hikone-R*, *StL*: *St. Louis*, *Zim*: *Zimbabwe* (Data taken from MEIKLEJOHN *et al.* 2003; figure designed by W. HENSE.)

### 2.1.2 Sequencing of gene *CG13360*

The upstream promoter region, the 5' UTR, and a large part of the coding sequence of gene *CG13360* was PCR-amplified in all eight strains using the primer pair 5' – CTTGG CCATG ACGCA ATG – 3' / 5' – AATGC GAGGG AAACG AAA – 3' (forward/reverse primer), while for sequencing the upstream promoter the latter primer was exchanged with the following reverse primer: 5' – CGGCG GTTTC TTCGA CTG – 3'. Sequencing was performed on PCR products using ExoSAP-IT® (USB, Cleveland) and applying BigDye v1.1 chemistry on a 3730 automated sequencer (Applied Biosystems, Foster City, USA).

### 2.1.3 Transformation vector construction

The amplified promoter region spans almost all of the 5'UTR and around 1.2 kb of the region directly upstream of the gene (*CG13360* spans the genomic region from 681,882 to 684,122 (in reverse orientation) of the X chromosome in genome release 5.10 including the 5' UTR; the amplified promoter region covers 1272 bp from 684,055 to 685,327. Thus, it contains also 67 of 75 bp of 5' UTR sequence (the A of the start codon ATG is at position 684,047). The remaining distance to the next upstream gene (*CG16989* starting at 687,201) is 1874 bp. Each of the promoters of *Zim53* and *Hikone-R* was PCR-amplified (forward primer sequence: 5' – GCCTA TATGC GCCTC AAGAC CC – 3'; reverse primer sequence: 5' – GCTGT CCTTT CTGGC TGCG – 3') and cloned separately into the plasmid vector pCR2.1-TOPO (Invitrogen, Carlsbad, CA) using standard techniques (SAMBROOK *et al.* 1989). After verifying the correct orientation by restriction analysis, a 3.4-kb *Not*I fragment of the plasmid vector pCMV-SPORT-ßgal containing the entire coding sequence of the *E. coli lacZ* gene (encoding β-galactosidase) was cloned into the *Not*I site of both of the TOPO vectors with the putative promoters (located downstream of the promoter insert). The correct orientation of the *Not*I insert was again confirmed by restriction analysis. The last step before cloning the two promoter-*lacZ* constructs into the *P*-element vector p*P[wFl]* (see below) was the introduction of an *Xho*I linker (New England Biolabs, Ipswich, MA) into the *Spe*I site of one of the TOPO vectors containing the promoter. For this we used the plasmid with the *Hikone-R* promoter.

The transformation vector p*P[wFl]* (SIEGAL and HARTL 1996) has two important features that make it ideal for our purposes. First, the functional part of it is flanked by *P*-element terminal repeat sequences that enable it to be stably integrated into the genome of *D. melanogaster* by transposition and germline transformation (see below). Second, this functional part has two cloning sites for the integration of genetic elements that one is interested in comparing with each other, *i.e.*, in our case the two promoter-*lacZ* constructs (transgene coplacement, see below).

To finish our transformation vector, the promoter-*lacZ* construct of *Hikone-R* was cloned as an *Xho*I fragment into the *Xho*I site of p*P[wFl]* (cloning site 2) resulting in a plasmid designated p*P[wFl-HikproB]*. After confirmation of the appropriate orientation (reverse) by restriction analysis, the second promoter-*lacZ* construct (that of *Zim53*) was inserted into cloning site 1 of p*P[wFl-HikproB]* (as a *Bam*HI/*Xba*I fragment into the *Bam*HI/*Spe*I site). The resulting plasmid vector (named p*P[wFl-Zim53proB-HikproB]*, Figure 2.2) was used for germline transformation.

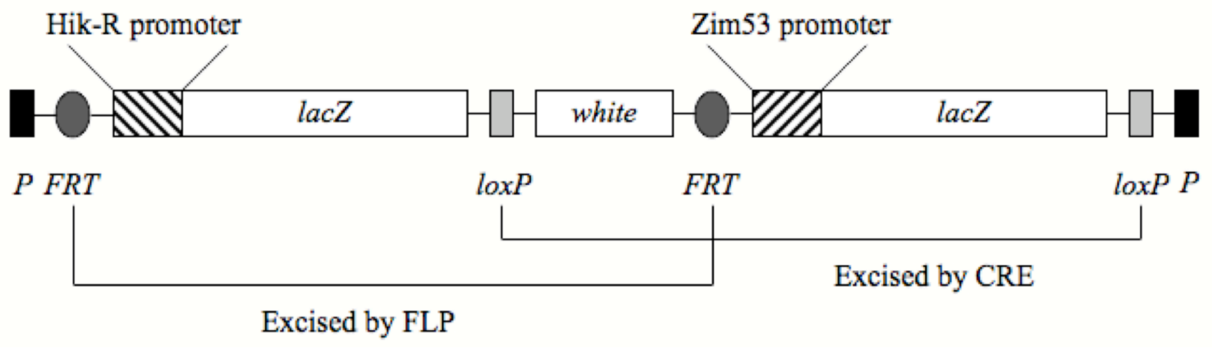


**Figure 2.2 The vector p*P[wFl-Zim53proB-HikproB]* used for transgene coplacement.** – Two promoter-*lacZ* constructs were inserted into the cloning sites of the transformation vector p*P[wFl]*, both flanking the selectable marker gene *white*. After successful transformation of the double construct two site-specific recombinases, FLP and CRE, excise the Hik-R promoter-*lacZ* and the *Zim53* promoter-*lacZ* construct, respectively, along with the *white* gene by utilizing their target sequences, *FRT* and *loxP*, respectively. (Figure designed by W. HENSE.)

### 2.1.4 Germline transformation by transgene coplacement

A solution of the above-described vector in water (concentration around 200 ng/µl) was used for *P*-element mediated germline transformation (RUBIN and SPRADLING 1982, SPRADLING and RUBIN 1982). To inject the plasmid vector, freshly laid eggs of our injection stock *y w*; Δ2-3, *Sb*/TM6, which carries a stable source of transposase, Δ2-3, marked with *Stubble* (*Sb*) (ROBERTSON *et al.* 1988), were collected from molasses plates (in time intervals of around 20 minutes), quickly dechorionated and desiccated for 2-4 minutes. Afterwards, the plasmid construct was injected into the posterior end of the embryo using a FemtoJet® microinjector and a TransferMan® NK micromanipulator (both from Eppendorf, Hamburg, Germany). Fly embryos showing any sign of unequal distribution of cytoplasm were regarded as too far developed and hence discarded. Injected eggs/embryos were kept at appropriate humidity for up to 48 hours and monitored for surviving larvae. These were put on standard food vials and, if surviving to adult flies, mated with flies of the *y w* strain of the opposite sex.

### 2.1.5 Fly care and maintenance

Since our transformation vector carries a phenotypic marker (the *mini-white* ($w^+$) gene of *D. melanogaster*, which is located between the two cloning sites (see below)), transformant flies could easily be identified by their red eye color and were mated to *y w* flies to remove the source of transposase to establish stable transformant lines. If the Δ2-3 element (marked with *Sb* bristles) did not segregate from red eye color ($w^+$), this was indicative of a transgenic insert on the *Sb* chromosome. These strains could not be maintained as stable lines and were used for immediate remobilization crosses with *y w* flies. Offspring from these crosses that had red eyes and wild-type bristles represented mobilized new transgenic lines. With this method the number of lines each representing a different chromosomal location was increased to a total of around 50.

In the next step we attempted to make all of the fly strains homozygous with regard to the transgenic insert. For this we utilized a *D. melanogaster* stock with multiple phenotypic markers (*y w*; *CyO*/*Sco*; *Ubx*/*Sb*) and a series of genetic crosses. The crossing scheme made it also possible to determine the chromosome (X, second or third) each construct was inserted in.

Since the following partial removal of inserts requires them to be located on the third chromosome and in homozygous state, we continued our analysis only with the ten lines that met these criteria.

### 2.1.6 Excision of either of the promoter-*lacZ* constructs

As already mentioned above, the specific structure of the transformation vector p*P[wFl]* allows for the comparison of two genetic elements. This is done in two steps: first, by the joint integration of the elements into a single, but random genomic location (described above) and, second, by the subsequent precise removal of the one or the other element.

For the latter, p*P[wFl]* additionally provides two systems for site-specific recombination, the Cre/*loxP* system of the bacteriophage P1, and the FLP/*FRT* system of *Saccharomyces cerevisiae*, where CRE and FLP proteins represent the site-specific recombinases and *loxP* and *FRT* the respective recognition sites. One pair of recognition sequences flanks the first cloning site and the $w^+$ gene (*loxP*), whereas the second pair does so with the region encompassing the $w^+$ gene to cloning site 2 (*FRT*). This construction allows for the precise removal of either of the two genetic elements (promoter-*lacZ* constructs) by crossing in appropriate fly strains carrying one of the two recombinases.

To do this, we first crossed virgins of each of our ten transgenic lines to males of a fly stock that harbours genes on balanced third chromosomes that encode the two recombinases CRE and FLP (*y w*; MKRS, *FLP*/TM6B, *cre*). As the transgenic flies were homozygous and had two copies of the promoter-*lacZ* constructs, each offspring inherited from the maternal side one of them and one of the recombinases on the homologous, paternally inherited third chromosome. After that we separated flies (females as virgins) carrying different recombinases (the MKRS, FLP chromosome is marked with *Sb*, whereas the TM6B, *cre* chromosome is marked $w^+$), crossed them with each other and treated them henceforth in independent, parallel crossing schemes. In one of them flies were heat-shocked as first instar larvae (for 1 hour at 38 °C) thereby activating FLP. In the second series of crosses they were grown at 25 °C allowing CRE to be active. A last cross removed the still present recombinase leaving the partially excised insert homozygous. These lines could then be maintained stably for generations.

For the β-galactosidase enzymatic assays we used heterozygous flies and for each transgenic line (for which now two sub-strains exist) mated one homozygous male to 3-4 *y w* females. For the analysis of the influence of the *Zim53* genetic background on *lacZ* expression we first reversed the cross and mated *Zim53* males to homozygous transgenic females. Here, female offspring of the cross were assayed. Secondly, to assay the *Zim53* X chromosome's influence also in male flies, we crossed male transgenic flies to females that came from a cross of *Zim53* males and X-balanced females (FM7j). Finally, to investigate the impact of an African X chromosome on reporter gene activity in general, we took transgenic males and used them for crosses with females of a *D. melanogaster* strain carrying only the X chromosome of *Zim157* and chromosomes 2 and 3 from a North-American stock.

### 2.1.7 Enzymatic assays

Adult flies from these crosses were collected within a time interval of three days (starting on the first day of eclosion) and stored alive to a final age of 6-8 or 3-5 days (two different age classes). To determine *in vivo* levels of β-galactosidase activity in our transgenic flies, we homogenized six male or female flies in 150 µl of a buffer containing 0.1 M tris-HCl, 1 mM EDTA, and 7 mM 2-mercaptoethanol at pH 7.5. The homogenates were then incubated on ice for 15 min and afterwards centrifuged at 12,000 *g* for 15 min at 4 °C. The supernatant with all soluble proteins of the flies was used immediately for the assays. Each of the homogenates was used twice, thus providing two technical replicates for each set of flies. In addition we obtained enough flies from each cross to perform assays on two biological replicates, *i.e.* different sets of six male flies. As a consequence, we could average β-galactosidase activity over a total of four replicates for each of the ten lines of transgenic insertions.

For each assay, we took 50 µl of the protein homogenate and added 50 µl of assay buffer consisting of 200 mM sodium phosphate (pH 7.3), 2 mM $MgCl_2$, 100 mM 2-

mercaptoethanol with *o*-nitrophenyl-β-D-galactopyranoside as substrate (in a concentration of 1.33 mg/ml). In a plate-reading spectrophotometer, we were able to follow the change in absorbance, which was caused by reporter gene activity, at a wavelength of 420 nm during 46 min with reads every 2 minutes.

For the statistical analysis of our expression results, we took advantage of the special design of the transformation vector. The previous problem of position-effect variation, *i.e.* the dependence of expression levels of transgenic constructs on chromosomal location, is overcome by transgene coplacement (SIEGAL and HARTL 1996, 1998; PARSCH 2004). After removal of one or the other promoter-*lacZ* variant, we end up with a pair of two sub-strains of each independent transgenic line with each sub-strain carrying only one variant allele of the promoter-*lacZ* construct, but at exactly the same position where its partner sub-strain carries the alternate allele. In statistical terms, this allows for a paired *t*-test, which greatly increases statistical power when activity of the paired transgenes is correlated. By this method the number of independent transgene inserts needed to detect significant differences is greatly reduced and it is possible to detect more subtle differences in transgene expression than with non-paired methods.

## 2.2 RESULTS

In this study we compared the ability of putative promoter sequences to drive allele-specific expression of a reporter gene. These promoter sequences came from the *D. melanogaster* gene *CG13360* that was previously shown to be differentially expressed between eight highly inbred lab strains of *D. melanogaster*. To do this, we cloned a large upstream region of two of these strains with an approximate length of 1.2 kb into the plasmid vector p*P[wFl]*, fused the *lacZ* reporter gene just downstream of them, and performed the method of transgene coplacement to directly compare the two reporter gene activities at ten different locations on the third chromosome.

| | | 1132 | 1092 | 1067 | 1060 | 1045 | 1009 | 814 | 813 | 812 | 811 | 808 | 807 | 806 | 767 | 758 | 709 | 576 | 543 | 542 | 541 | 540 | 539 | 538 | 537 | 536 | 480 | 456 | 455 | 454 | 453 | 452 | 451 | 450 | 449 | 448 | 447 | 403 | 347 | 246 | 166 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| non-African | *Can-S* | A | G | G | A | T | C | - | - | - | - | A | G | A | C | C | A | C | T | C | A | G | A | A | T | T | G | C | - | - | - | - | - | - | - | - | - | C | C | A | C |
| | *Ore-R* | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . |
| | *Hik-R* | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . |
| | *StL* | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . |
| African | *Zim53* | . | . | . | G | . | . | T | A | C | A | C | A | G | G | G | C | T | . | . | . | . | . | . | . | A | A | . | . | . | . | . | . | . | . | . | . | T | T | G | T |
| | *Zim(S)2* | T | T | A | G | A | . | T | A | C | A | C | A | G | G | G | C | T | . | . | . | . | . | . | . | A | A | . | . | . | . | . | . | . | . | . | . | . | . | . | . |
| | *Zim29* | . | . | . | . | . | T | T | A | C | A | C | A | G | G | G | C | . | . | . | . | . | . | . | . | . | A | T | C | A | G | C | C | C | A | C | A | . | T | . | T |
| | *Zim30* | . | . | . | . | . | . | T | A | C | A | C | A | G | G | G | C | T | - | - | - | - | - | - | - | - | A | . | . | . | . | . | . | . | . | . | . | . | T | . | . |

**Figure 2.3 Sequence alignment of the upstream region of gene *CG13360*.** – Shown are the segregating sites of all 8 strains used in this chapter (consisting of two populations: African and non-African) in the region that was cloned into the "waffle" vector (which was done only with the alleles of *Hik-R* and *Zim53*). Numbers are counted down and are given relative to the A in the start codon ATG. Shaded are the fixed indel polymorphism (at positions -814 to -811) and the 6 fixed SNPs. (Sequencing performed and figure designed by W. HENSE.)

### 2.2.1 Selection and sequencing of gene *CG13360*

We chose to investigate the putative promoter region of the X-linked gene *CG13360* of *D. melanogaster* because in a microarray survey, it showed significant expression differences among eight strains from locations in Zimbabwe (Africa), Japan, and the United States (MEIKLEJOHN *et al.* 2003). These expression differences were grouped: the African population (*Zim(S)2*, *Zim29*, *Zim30*, and *Zim53*) had an average relative expression of 1.86, whereas the corresponding value of the remaining strains (the non-African "cosmopolitan" population) was only 1.15 (Figure 2.1). We called this a "fixed expression difference" between the two populations. Sequencing of around 1.2 kb of the upstream part of the gene also revealed a number of fixed DNA polymorphisms between the two populations; among them one indel polymorphism and six SNPs in a ~100-bp window and another SNP around 230 bp further downstream, which all are located in the upstream region and not in the 5' UTR. The polymorphic differences are at positions -480, -709, -758, -767, -806 to -808 (all SNPs), and -811 to -814 (indel) relative to the A in the ATG start codon (Figure 2.3). Moreover there are two more fixed differences in the first intron of *CG13360*, which were not included in subsequent transgene constructs.

We selected the two strains that showed the largest difference in relative expression of *CG13360*, *i.e. Hikone-R* (relative expression 1.00) and *Zim53* (1.94), and designed PCR primers to amplify the putative promoter of this gene in order to functionally analyze the significance of DNA polymorphisms in driving allelic gene expression.

### 2.2.2 Expression of transgenic inserts

First, we found that our *lacZ* reporter gene shows activity in all of our transgenic flies. In total, we generated 10 pairs of fly strains each of which allows for one comparison of homologous putative promoter sequences at a distinct genomic location. Due to experimental restrictions imposed by transgene coplacement we only used transgenic inserts on the third chromosome for our analysis. These already show considerable variation in transgene activity: the lowest value being 1.345 mOD/min and the highest 10.621 mOD/min, the overall average is 4.362 mOD/min (standard deviation: 2.407 mOD/min; coefficient of variation (CV): 55.18%; Figure 2.4A). These absorbance values came from measurements of adult male flies that were 6 to 8 days old. Since gene expression in general is also known to be age-dependent we repeated the enzymatic assays with flies of a second age class, this time flies aged 3 to 5 days. Here, we obtained very similar results suggesting that reporter gene

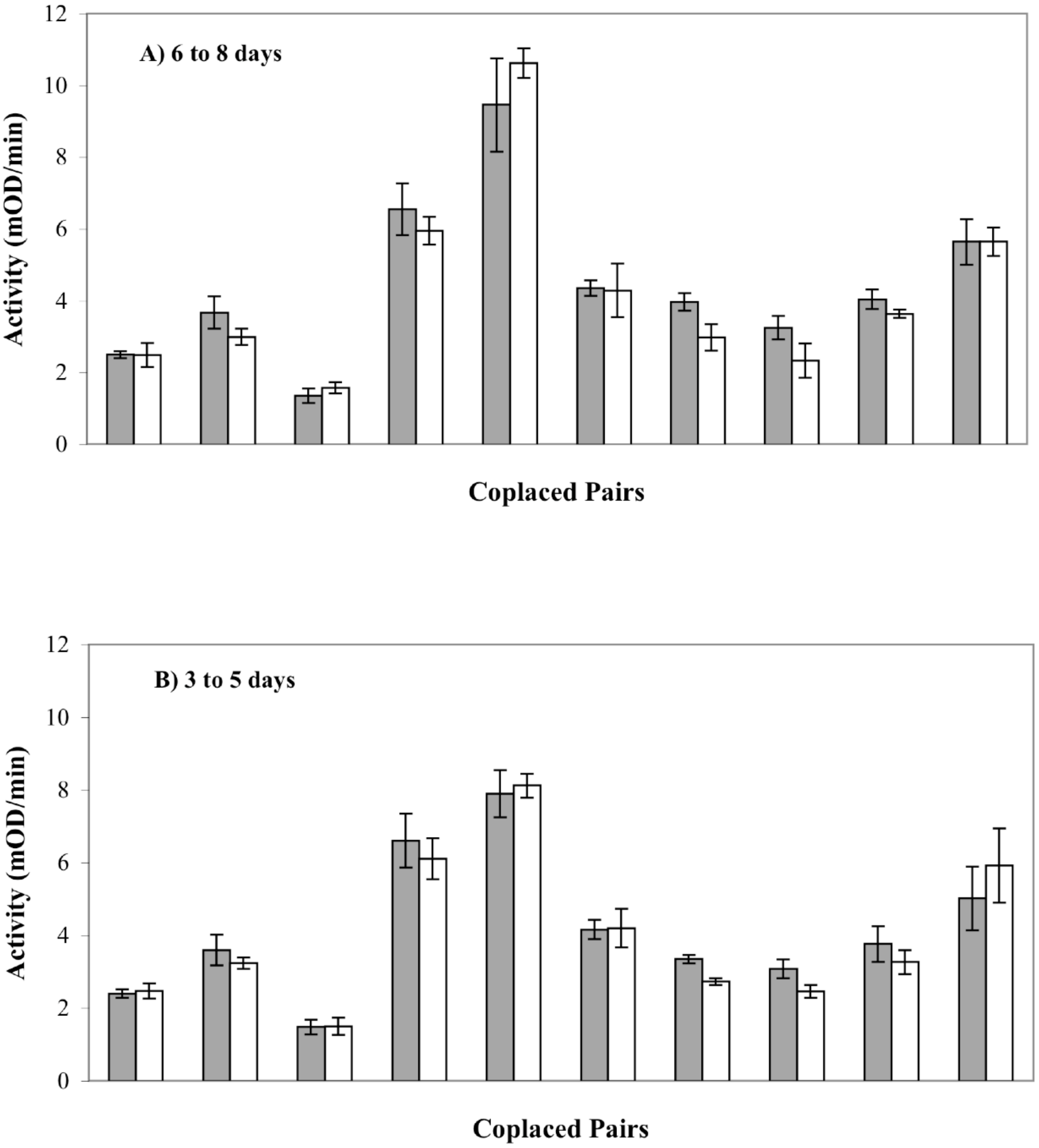


**Figure 2.4 β-galactosidase activity in male *D. melanogaster*.** – Enzymatic activity of *lacZ* driven by the *Zim53* promoter (filled bars) and the *Hik-R* promoter (open bars) in transgenic flies that were (A) 6 to 8, and (B) 3 to 5 days old. Each pair shows the measurement of one of 10 unique, independent transgene insertions on the 3rd chromosome. Error bars indicate ± 1 standard deviation from the mean. (Assays performed and figure designed by W. HENSE.)

expression is rather constant within the age span of 3 to 8 days. The mean absorbance over all 20 measurements (10 for each promoter construct) was this time 4.068 mOD/min with a standard deviation of 1.949 units (CV = 47.91%). We also tested for an age effect in expression and found an only marginally significant difference in one set of transgenes, *i.e.* the transgenes with the putative promoter of *Zim53* fused to *lacZ* (paired *t*-test, two-tailed

$p = 0.058$), whereas the second set of promoter-*lacZ* constructs (with the promoter of *Hik-R*) displayed a virtually indistinguishable expression of β-galactosidase between the two tested age classes (paired *t*-test, two-tailed $p = 0.365$). Thus, the age of the flies used for enzyme assays has almost no influence on transgene expression, since even in the case of the *Zim53* promoter the average fold difference in expression of older to younger flies was only 1.064, that means a 6% increase with time.

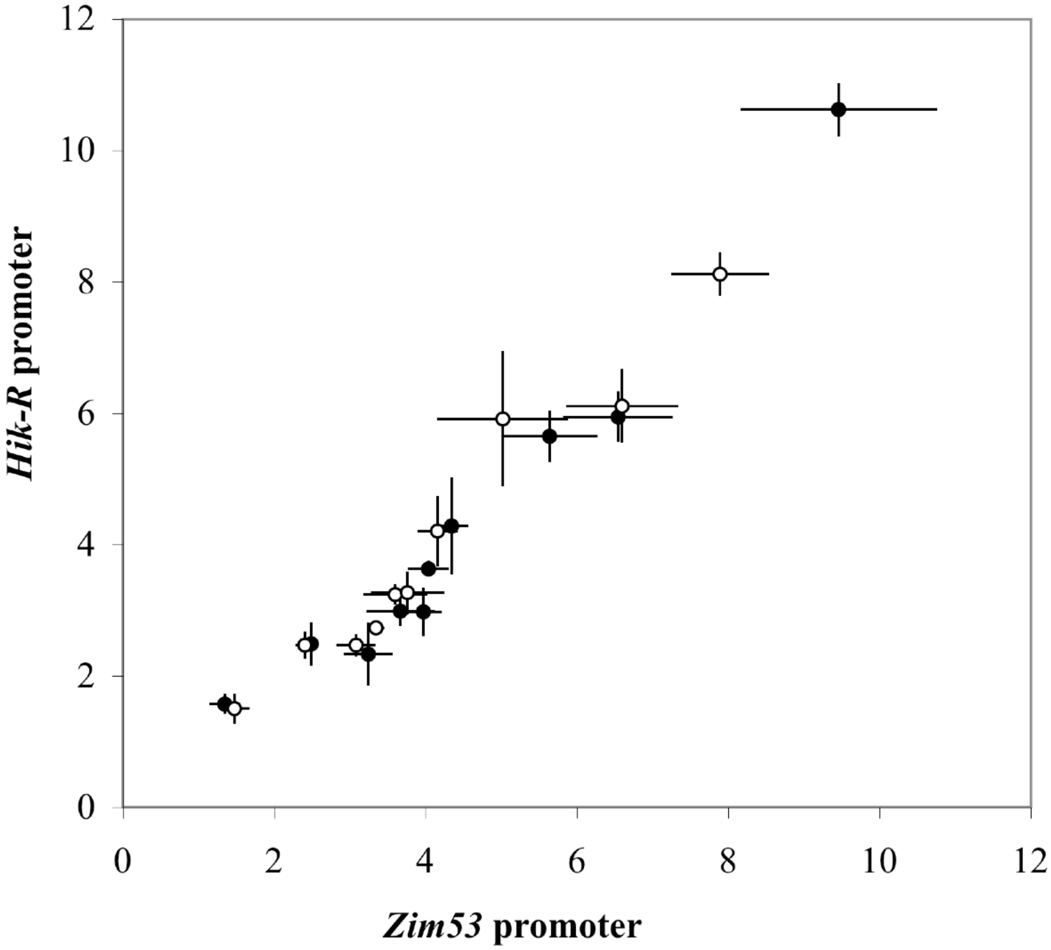


**Figure 2.5 Correlation of alternative promoter activities.** – There is a positive correlation between the activities of the two promoter-*lacZ* constructs, both in the young (3 to 5 days old; open circles) and the old male *D. melanogaster* flies (6 to 8 days; filled circles), thereby demonstrating that transgene coplacment is an effective means to reduce position-effect variation. Pearson's *R* values are 0.974 ($p < 0.0001$) and 0.976 ($p < 0.0001$), respectively. Each circle corresponds to an insertion at a unique, independent location on the 3rd chromosome. Error bars indicate ± 1 standard deviation from the mean. (Figure designed by W. HENSE.)

In all comparisons considered so far the expression of pairs of transgenic insertions was highly correlated (Figure 2.5). For example, the correlation coefficient *R* of coplaced inserts of 6-to-8-day-old flies is 0.976 ($p < 0.0001$), the same value for the younger is flies is 0.974 ($p < 0.0001$), thus supporting the advantage of transgene coplacement as a means to overcome the problem of position-effect variation by directly comparing two transgenic variants at exactly the same position in the genome.

### 2.2.3 Variable reporter gene expression driven by different promoters

As a next step in our analysis we compared the effect of the two different promoter sequences in driving the expression of *lacZ*. The difference in allelic gene expression of the gene we derived the promoters from, *i.e. CG13360*, as measured by a microarray assay of adult male flies was 1.94-fold (with the higher expression in the strain *Zim53*). However, when we assayed male flies in a genetic background coming originally from the injection fly stock (*y w*; Δ2-3, *Sb*/TM6) and from a double-recessive marker strain (*y w*), we observe no expression difference between the *Zim53-lacZ* and the *Hik-R-lacZ* constructs. Neither the male flies of age 6 to 8 days (Figure 2.4A) nor the younger ones (3 to 5 days; Figure 2.4B) showed a consistent and significant pattern of expression difference. If the cloned upstream sequence of *Zim53* carrying several mutations compared to the version of *Hik-R* were responsible for the 1.94-fold difference in gene expression, then we would expect that the corresponding reporter gene construct with *lacZ* exhibits at least a fraction of this difference. But not only are the observed differences at each chromosomal position very small, there fails to be a consistent signal of higher expression in one of the two alternative insertions compared to the other. Of all ten cases we obtained, at five genomic locations the expression of *lacZ* driven by the *Zim53* promoter is higher than the one driven by the *Hik-R* promoter, in the remaining cases the situation is reversed, an altogether rather random distribution. The statistical tests we applied are also – as expected – non-significant. A two-tailed paired *t*-test for the 6-to-8-day-old and the 3-to-5-day-old flies resulted in *p* values of 0.291 and 0.406, respectively. This result favors the scenario in which the putative promoters do not influence the allelic expression. This view is further supported by similar results observed in the two age classes we assayed.

Because the above assays were performed only in male transgenic flies heterozygous for the transgene we next considered the possibility that an expression difference is revealed when assaying female flies. Although the original microarray results came from hybridizations of male cDNA and the gene *CG13360* is not known to be expressed sex-specifically, we cannot *a priori* exclude this possibility. Since the two assayed age classes were not significantly different we restricted this analysis of females flies to the class of younger flies. When doing this, however, we could not find any difference to the results

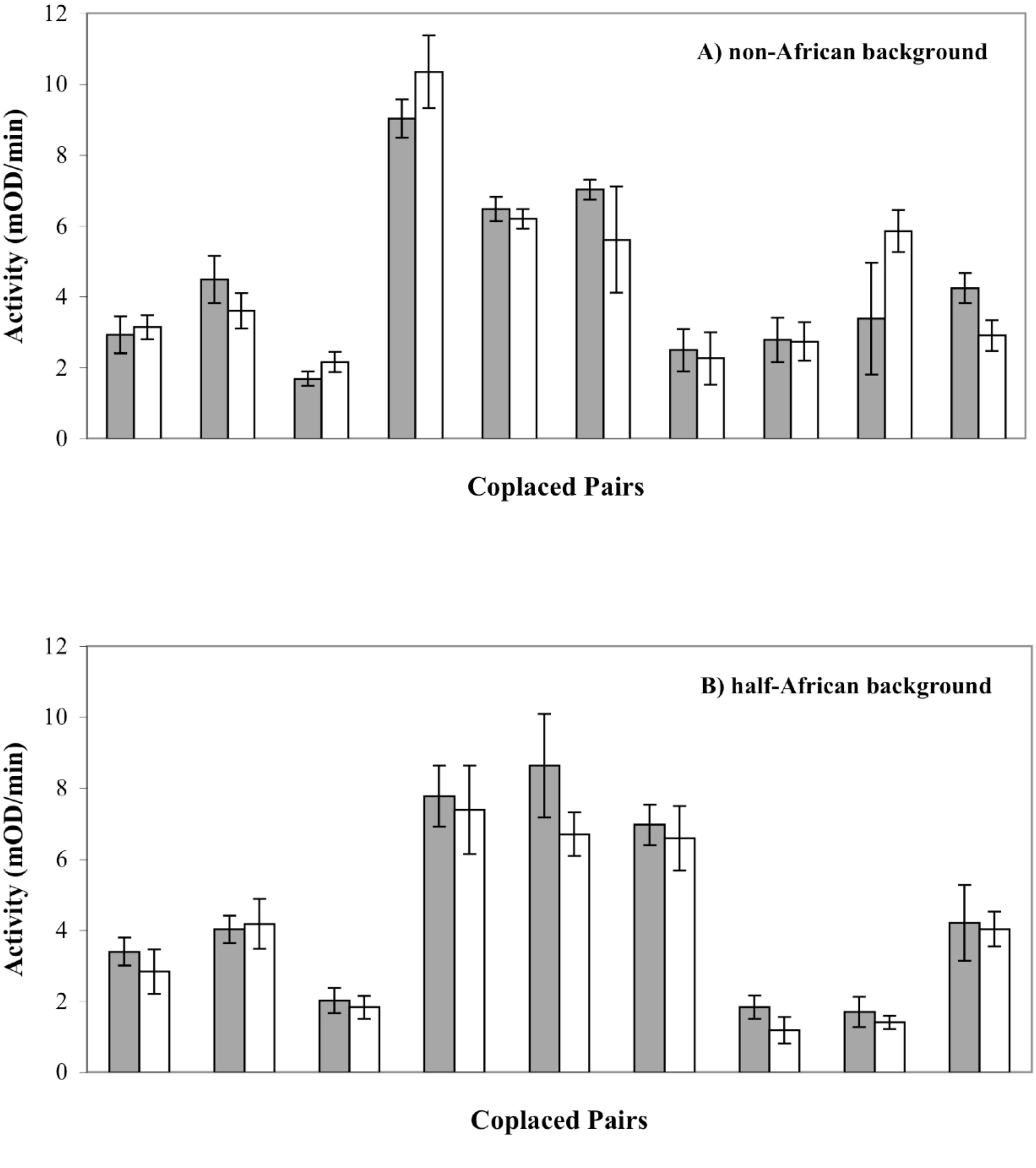


**Figure 2.6 β-galactosidase activity in female *D. melanogaster*.** – Transgenic *lacZ* expression in 3-to-5-day-old female flies driven by the *Zim53* (filled bars) and the *Hik-R* (open bars) promoter, (A) in a completely non-African chromosomal background, and (B) in a genetic background containing a full haploid African chromosome set, including the X chromosome, from *Zim53*. In (B) a consistent (with one exception) and significant pattern of higher activity in the *Zim53* promoter-*lacZ* construct emerges (paired *t*-test, two-tailed $p = 0.037$). Error bars indicate ± 1 standard deviation from the mean. (Assays performed and figure designed by W. HENSE.)

obtained in males (Figure 2.6A). This time the promoter driving higher reporter gene expression was the one from *Zim53* in 6 instances. The appropriate statistical test gave a non-significant result again (paired *t*-test, two-tailed $p = 0.942$) with an additionally weaker correlation coefficient of paired inserts of only 0.885 ($p = 0.0006$). Finally, we also compared male to female expression in the above data sets so far. There seems to be no sex-bias in reporter gene activity, thus confirming a previous expression measurement of gene *CG13360* (RANZ *et al.* 2003). When data sets were separated, both the *t*-test for the *Zim53* and the *Hik-R* promoter constructs were not significant (both two-tailed *p* values 0.48) in a comparison of male to female expression. If the data are pooled, the correlation drops down to a value of $R = 0.702$ ($p = 0.0005$; Figure 2.7), due to a weaker correlation of the *Hik-R* half of the data (separate $R = 0.620$, $p = 0.055$; that of *Zim53* half: $R = 0.800$, $p = 0.005$).

Summarizing, the above indicate that there is promoter activity present in the two variants of upstream regions of the gene *CG13360*, but that this activity does not differ between variants. Neither transgenic construct shows consistently higher expression than the other, and the result holds for two age classes of male flies and at least one age class of female flies.

Gene expression is known to be governed not only by *cis*-regulatory regions such as upstream promoters and enhancers or regulatory DNA within introns, it can also be influenced post-transcriptionally by UTRs or by codon usage bias. When considering transcription initiation, chromatin structure and condensation must be modified and loosened to allow the transcription machinery consisting of several proteins (the *trans* factors) to access the appropriate regulatory DNA regions. In this sense, protein-DNA and also protein-protein interactions are assumed to play a crucial role in transcription and gene expression. Therefore polymorphisms in these *cis* and *trans* factors are believed to be responsible for both intra- and interspecific gene expression differences. Applying this to our present study where it was so far impossible to recreate the expression differences observed in the gene *CG13360* in its natural genetic environment when transferred into a controlled experimental setting, we next investigated the role of *trans* factors on a chromosomal scale. The *trans* factors are generally encoded elsewhere in the genome compared to the affected gene making it extremely difficult to identify and trace those factors. In order to see whether the genetic background as a whole is influential in our case, we placed our transgenic constructs into a more natural genetic environment and crossed male flies carrying the transgene with females of the *Zim53* strain. Thus, at least a haploid genome with all the *trans* factors from the natural fly stock in which the expression difference was actually observed would be present.

When doing this, we observed an interesting effect. Indeed, the presence of a haploid African chromosome set from *Zim53* leads to the emergence of a more consistent expression pattern that the other data set failed to reveal (Figure 2.6B). After excluding one of the strains where the chromosomal location of the insert might potentially have an additional counter-effect on expression, in 8 of the 9 remaining positions on the third chromosome the *Zim53* promoter drives a higher enzyme production than the alternative promoter version (*Hik-R*).

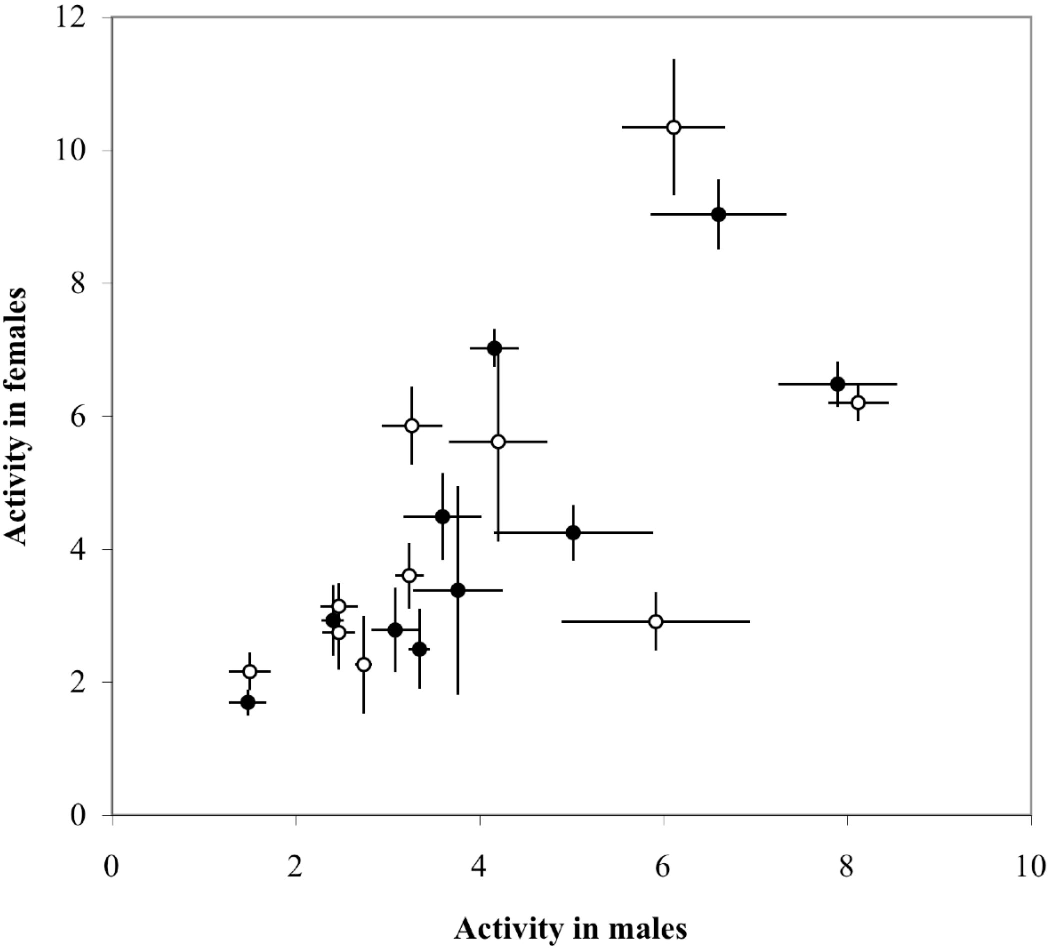


**Figure 2.7 Expression correlation between sexes in transgenic *D. melanogaster*.** – β-galactosidase activity in 3-to-5-day-old female and male flies carrying either the *Zim53* (filled circles) or *Hik-R* (open circles) promoter-*lacZ* constructs shows a correlation that is weaker than the one between different transgenic constructs (Pearson's $R = 0.702$, $p = 0.0005$). Error bars indicate ± 1 standard deviation from the mean. (Figure designed by W. HENSE.)

The expression ratios detected range from 1.044 to 1.549 with an average of 1.163, *i.e.* there is a 16% increase in enzyme production from the *Zim53* promoter version. Although this value is much smaller than the original 1.94-fold difference measured for *CG13360* by microarrays, it proves to be marginally significant (paired *t*-test, two-tailed $p = 0.037$).

Another caveat for this result is that we obtained it only in female flies. Opposite to our usual procedure we performed assays on the other sex, as we wanted the African X chromosome be present. First we planned to cross male transgenic flies to females of *Zim53* to do the enzyme assays on the sons. That way we could have tested for a whole haploid African genome, since

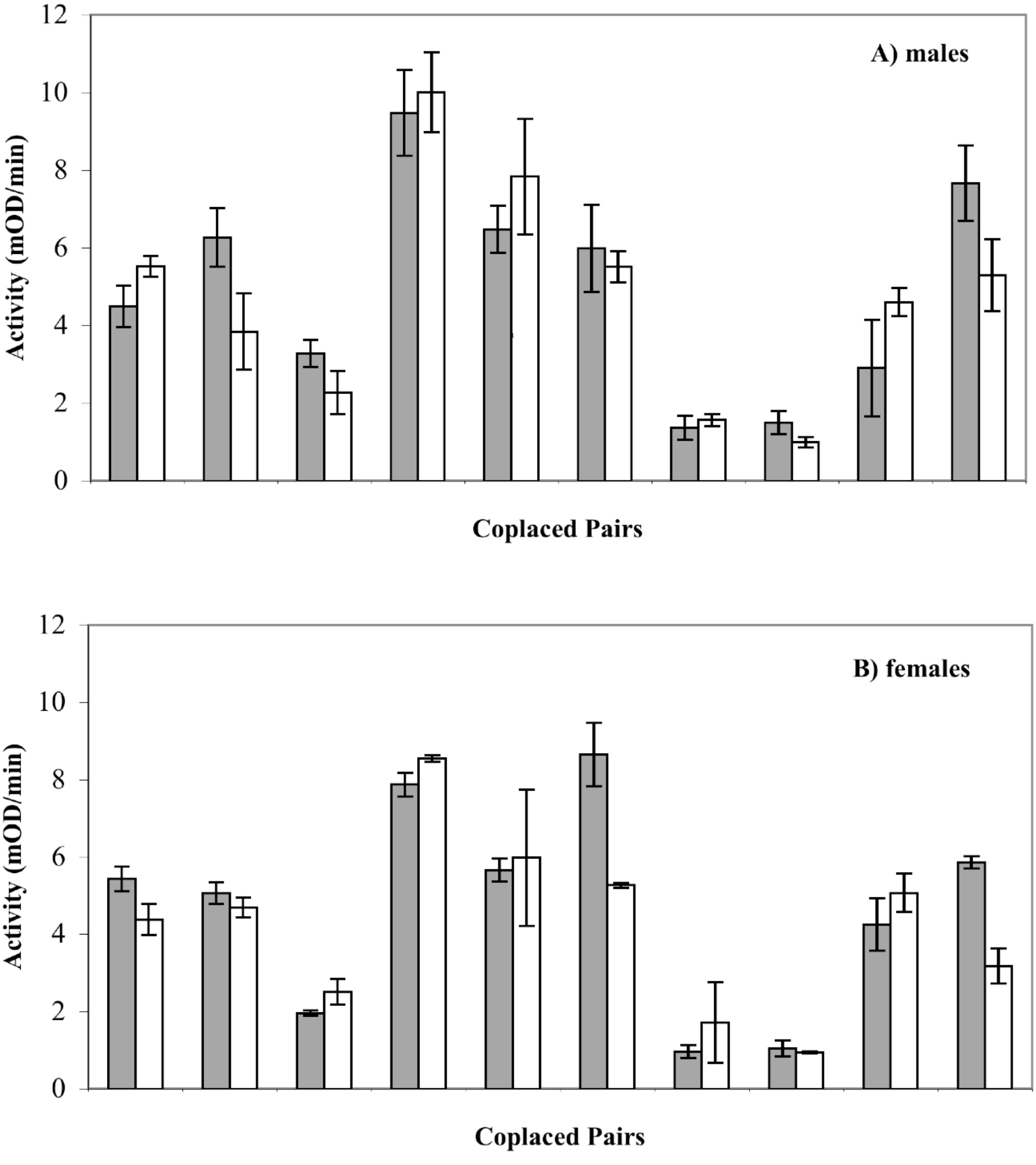


**Figure 2.8 β-galactosidase activity in *D. melanogaster* carrying a *Zim157* X chromosome.** – Expression of the *lacZ* transgene in (A) male, and (B) female transgenic flies driven by the *Zim53* (filled bars) and the *Hik-R* (open bars) promoters in the presence of a single X chromosome from the strain *Zim157*. Both patterns were non-significant (paired *t*-test, two-tailed *p* = 0.677 and 0.367, respectively). Error bars indicate ± 1 standard deviation from the mean. (Assays performed and figure designed by W. HENSE.)

sons of this cross inherited the X chromosome from the African mother. Due to a marked mate choice preference exhibited even in the absence of preferred males and sole presence of unpreferred ones, the African females were almost completely successful in repelling the latter, the transgenic male flies. Thus, we were forced to take female offspring from the reverse cross in order to have an African X chromosome. We nevertheless also performed enzyme assays on males coming from this reverse cross and found a negative result (paired *t*-test for difference between *Zim53* and *Hik-R* promoters: two-tailed $p$ = 0.58). These males were only lacking the African X chromosome, otherwise they had the identical genotype as the females where the effect was measured. Since female flies with the non-African background also showed no significant difference in reporter gene expression (Figure 2.6), it is like that the effect requires the African X chromosome.

To measure the African X chromosome's regulatory impact in male transgenic flies, we first introduced the X of another African *D. melanogaster* strain from Zimbabwe (*Zim157*) into the transgenic flies. The strain we used contains only the X chromosome from Africa, whereas chromosomes 2 and 3 were derived from a North-American *D. melanogaster* line. Here again, although we focused on an X chromosome that is assumed to be quite close to the one from *Zim53* in terms of evolutionary distance, we could not detect any differences in promoter activity on *lacZ* expression, neither in heterozygous males (paired *t*-test, two-tailed $p$ = 0.677) nor females (two-tailed $p$ = 0.367) (Figure 2.8), demonstrating that there is variation affecting expression even among Zimbabwe strains. We also were able to bring *Zim53*'s X chromosome into male transgenic flies by utilizing a fly stock with a balanced X chromosome (FM7j). Female offspring from a cross of male *Zim53* and female FM7j flies carrying only a haploid chromosome set of *Zim53* showed a much weaker mating preference than pure *Zim53* females. Thus we were able to mate these females to male transgenic flies to obtain male offspring with the African X chromosome. As in the case of females carrying the African X chromosome we obtained a clearer expression difference of *lacZ* driven by the two promoter variants. After excluding the last of our 10 transgenic lines again, in seven of nine transgenic strains β-galactosidase activity was higher when driven by the *Zim53* promoter. One of the two outliers shows almost exactly the same activity, whereas in the remaining one the difference is slightly higher (Figure 2.9). Nevertheless the statistical test proves to be significant (paired *t*-test, two-tailed $p$ = 0.044). The correlation between coplaced inserts is again very strong (Pearson's $R$ = 0.954, two-tailed $p$ < 0.0001). Furthermore the average fold-difference in the activity ratio of *Zim53* to *Hik-R* is 1.268, and higher than the respective value in females of 1.16 (see above). Thus, in males, there is a 26% increase in expression driven by the *Zim53* promoter relative to the *Hik-R* promoter. Since the original microarray was performed with male *D. melanogaster* and measured a fold-difference of 1.94, one would expect the difference to be higher in males than in females. However, a difference of 1.26-fold is still much lower than 1.94-fold, suggesting that other factors contribute to expression differences detected in the microarray experiment (see DISCUSSION).

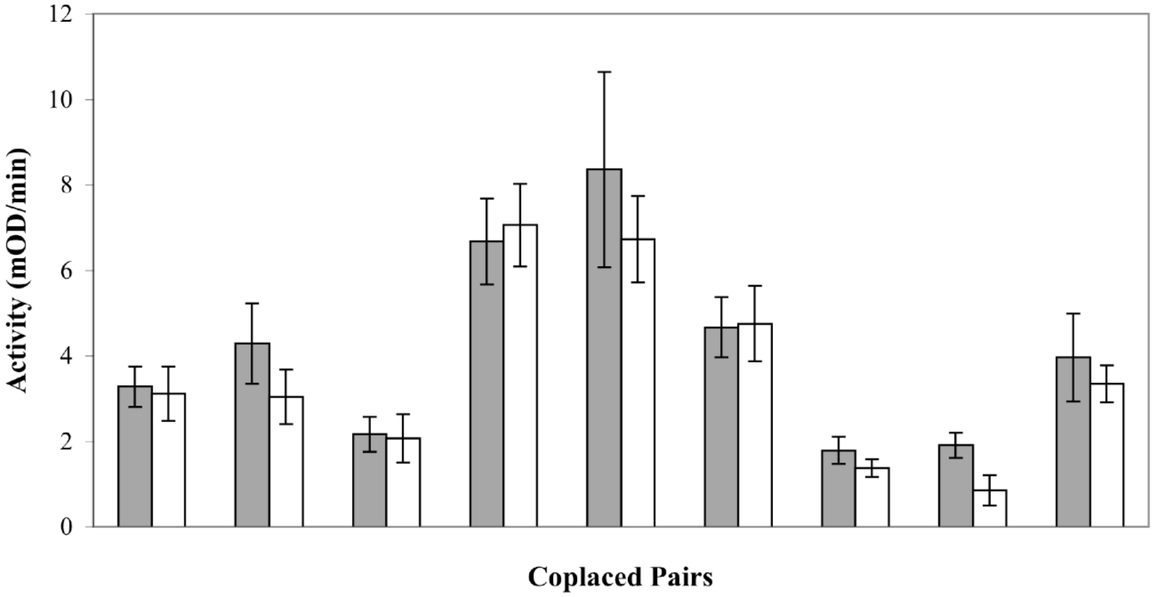


**Figure 2.9 β-galactosidase activity in male *D. melanogaster* with an African X chromosome from *Zim53*.** – Expression of the *lacZ* gene in transgenic flies driven by the *Zim53* (filled bars) and the *Hik-R* (open bars) promoter shows a weak, but nevertheless significant pattern (as in the case of females; Figure 2.6B). A two-tailed *t*-test results in a *p* value of 0.044. Error bars indicate ± 1 standard deviation from the mean. (Assays performed and figure designed by W. HENSE.)

Taken together, our results suggest a functional regulatory role of polymorphisms within a 1.2-kb region upstream of *CG13360*, but it is only observable in a background containing the *Zim53* X chromosome. Such a regulatory role was not observable when flies were analyzed that had an entirely non-African chromosomal background, such as the flies from the derived *D. melanogaster* strains we used for transgenesis. In this genetic background, both male and female flies showed no significant expression differences. Moreover, in the experiments with males we tested two different age classes both of which were negative. The situation, however, changes with the presence of the *Zim53* X chromosome. Summing up, we can conclude that one or more *trans* factors residing on this X chromosome interact with the cloned promoter region on the third chromosome, where all of our analyzed inserts lie, since all the flies had an otherwise identical genetic background. The exceptions were the second-to-last series of flies with the *Zim157* X chromosome and the flies taken for the final enzymatic assays, *i.e.* male *D. melanogaster* with the appropriate African X chromosome. Here, due to the crossing scheme we employed the assayed flies had one second and third chromosome (where the transgene was inserted) from the non-African fly stocks, whereas the other of these two chromosomes came either from *Zim53* or FM7j, the latter again expected to better resemble the non-African stock.

## 2.3 Discussion

The evolution of gene expression is currently a topic of hot debate. After detection of substantial variation in gene expression not only between, but also within species, as revealed by recent microarray studies in diverse taxa, evolutionary biologists started to focus on evolutionary principles and molecular factors governing it. Since expression levels of genes are to some degree heritable, genetic elements must be at least partially responsible for the variance in expression. Among these are *cis*-regulatory factors like promoters and enhancers (but also silencers) that lie within close proximity of the gene in question, and *trans* factors that normally are transcription factors encoded elsewhere in the genome.

In this study, we therefore investigated the functional role of a putative *cis*-regulatory region in gene expression by fusing it to the *lacZ* reporter gene and analyzing its enzyme activity in transgenic *D. melanogaster*. Although we found an effect of this *cis*-regulatory promoter region on reporter gene expression in an appropriate genetic background, it was very weak and statistically only marginally significant. According to our experimental results, the genetic background seems to require the presence of the X chromosome of the *D. melanogaster* strain *Zim53* to exhibit at least this weak effect. The strain *Zim53* is the one in which the gene we derived one of the alleles of putative promoter from (*CG13360*) showed a 1.94-fold higher expression than in the strain *Hik-R* (MEIKLEJOHN *et al.* 2003). Indeed, the corresponding promoter-*lacZ* construct also showed the higher reporter gene expression. It was, however, not possible to obtain an equally large difference, since the fold-difference of *lacZ* expression was on average only 1.26 and 1.16 in males and females, respectively, suggesting that there must exist more regulatory factors than the potential ones in the cloned promoter region. Among them there could be regulatory sites in the two introns (a large first one of 568 nucleotides including two more fixed differences and a small one of only 64 nucleotides) or in the 3'UTR of *CG13360*, acting during transcription initiation or post-transcriptionally (by microRNAs and their respective binding sites). These categories of *cis* sequences are known to harbor regulatory functions but were missing in our promoter-*lacZ* constructs. Furthermore, different codon usage in the two species *D. melanogaster* and *E. coli* (where the reporter gene *lacZ* comes from) could also account for the residual expression difference between our experimental and the natural conditions in which expression was measured. Surely, the two genes *CG13360* and *lacZ* come from very diverse species, but as long as the ability of a putative regulatory region to control gene expression is considered, it should not matter which gene is located downstream and of which origin it is. However, synonymous codon usage can have an influence on gene expression, especially when a bacterial gene like *lacZ* whose codons are optimized for the *E. coli* tRNA pool is transferred into a eukaryotic organism like *Drosophila* that provides a distinct pool of tRNAs. Interestingly, PARSCH (2004) reported similar quantitative results in a study that investigated the functional role of an 8-bp sequence in the 3'UTR of the *Adh* gene of *D. melanogaster* in expression. Here, using the same technical approach of transgene coplacement, the original 2-fold expression difference in the natural genetic setting was also reduced to an average 1.16 in the experimental setting, a heterologous reporter gene construct consisting of the human cytomegalovirus (CMV) promoter, *lacZ* and either of the two variants of the 3'UTR. Together these examples show that is sometimes not possible to restore a naturally occurring expression difference under experimental conditions by solely focusing on one known regulation mechanism.

In addition to the above we cannot exclude that there are more *cis* factors further apart (more upstream in 5' direction) that influence expression of *CG13360* and hence *lacZ*. Our putative promoter was limited in length due to use of restriction enzymes we applied to fuse the promoter to the reporter gene. The entire intergenic region between *CG13360* and the neighboring upstream gene (*CG16989*) including the 5' UTR of *CG13360* is 3154 bp in length of which 1272 bp were covered in our putative promoter by PCR amplification, thus leaving 1874 bp of the total intergenic region outside our construct. This reasoning assumes that regulatory *cis* sequences can only occur in the intergenic region to the next gene, which is a conservative assumption. Recent research found that additional *cis*-regulatory sequences such as enhancers and silencers could often be found further away, sometimes several kilobases not only upstream, but also downstream of the target gene, making it difficult in general to identify and localize such regions. Another intricate problem is that many transcription factor binding sites are so short in length (around 6 to 30 bp) that even with perfect sequence conservation they would be difficult to find by computational approaches simply because the statistical signal of such short sequences would be too low or the false positive rate too high.

The knowledge of all *cis*-regulatory elements of a particular gene would enable one to assess the relative importance of *cis* versus *trans* factors in gene expression. However, as binding site predictions mainly come from bioinformatics approaches, they are in need of experimental verification which is generally much more time-consuming. An experimental analysis of a whole set of *cis* factors (perhaps in different combinations) would require new experimental techniques with a good deal of improvement in the accuracy of measurement, as the total variance in gene expression is (despite the experimental noise that is inherent in current methodologies) the sum of many factors which interact with each other resulting in a delicate dynamic and flexible balance (steady state). With the method employed in this study we were able to focus only on one aspect of expression regulation and end up with a marginally significant result, although transgene coplacement already eliminates position-effect variation. Depending on the size of the effect (here: a difference in gene expression) in the natural genetic environment one would like to investigate and taking into account that this size is likely to decrease in the experiment (due to the other, neglected factors' influence), the experimenter is well-advised to carefully choose candidate genes and employ and develop further state-of-the-art methods. In our case it might have been better to select a gene with a larger expression difference. On the other hand, the method of coplacement was very effective. By this means, the difference in expression necessary to yield a statistically significant result is greatly reduced. If we had had to use a normal instead of a paired *t*-test for statistical analysis, we would have obtained a non-significant result (two-tailed $p = 0.687$) and the necessary difference would have been much larger (absolute: 2.53; relative: a 63% increase compared to the average expression of the *Hik-R* or a 29% decrease compared the average *Zim53* expression). In constrast, an average difference as small as 0.45 or 11% already yields a significant result when using coplacement. When regarding the correlation coefficients of coplaced insertions, which range from 0.828 to 0.977, the method of transgene coplacement proves to be a valuable means to avoid position-effect variation and the need for a much larger number of transgene locations in the genome. But the main advantage of this method is the ability to detect more subtle expression differences of transgenes. However, even this precise method is set a limit when expression differences of the order of much less than around 10% are to be revealed (also depending on absolute expression values).

The initiation of transcription is mediated by an interaction of proteins and DNA and also protein-protein interactions. Together, these proteins form a complex that opens the chromatin structure to enable RNA polymerase to access the beginning of a gene's coding sequence. Taking this into account, even the complete knowledge of *cis*-acting DNA factors would not suffice to understand the evolution of gene expression. It is only the advanced DNA sequencing technique that makes evolutionary biologists dealing with these questions focus on *cis*-regulatory sequences. But mutations both in *cis* and in *trans* factors contribute to the evolution of gene expression, and it seems likely that neither is more prevalent than the other. Furthermore, the molecular interaction opens up the possibility of compensatory mutations and co-evolution, as in the evolution of mRNA secondary structure (CHEN *et al.* 1999; LANDRY *et al.* 2005). Indeed, what is a *cis* sequence worth without an appropriate binding partner? Or, vice versa: Can a transcription factor be effective without a docking site on the DNA? In CHAPTER 1 of this thesis a short upstream DNA fragment of the *ocnus* gene was shown to impart testis-specific expression of a transgene. This example, however, also demonstrates that the presence of a suitable transcription factor is crucial for expression since the same fragment also exists in females upstream of the *ocnus* gene without expression, and existed in transgenic females, but without reporter gene activity.

What can be further done with the results of this study to explore the functional significance of our cloned promoter sequence? First, applying the above to our example of gene *CG13360* we could easily test whether the X chromosome with its *trans* factors alone is sufficient to drive allele-specific expression and whether the polymorphisms found in the cloned upstream region are neutral to expression. To do this, we could hybridize the two *D. melanogaster* strains, *Zim53* and *Hik-R*, thereby exchanging the X chromosomes of both, and see if expression differs. In males, this approach would be straightforward, whereas in females, the $F_1$ generation would carry one X chromosome of each strain. Another generation would be required to get the *Zim53* X chromosome homozygous in the *Hik-R* background and vice versa. These experiments could be performed either with flies carrying the additional transgenic 3rd chromosome with subsequent enzyme assays (as done before), or alternatively, without transgenic background by measuring *CG13360* mRNA with qRT-PCR methods.

Secondly, although minor effects from chromosomes 2 and 3 cannot be excluded, future experiments could include the generation of recombinant inbred lines (RILs) to produce a mosaic of the *Zim53* and some other, *e.g.* European, X chromosome. By this way, it would be possible to narrow down the region in which the *trans* factor(s) can be found. The problem, however, that would arise is the starting effect size of around 16 to 26% expression difference at most which is likely to decrease as there might be more than one *trans* factor on the X chromosome contributing to this 16 to 26%. Thus, the sensitivity of the method applied here could quickly reach its limits.

Thirdly, to accelerate and support the above we can make use of the strain *Zim157*. In the last set of experiments of this study, we analyzed its X chromosome's influence on transgene expression. As we could not find any significant effect, the factors responsible for expression differences are to be found in genetic loci where this strain differs from *Zim53*. As both of these naturally occurring fly strains, which are highly inbred, are derived from an African population in Zimbabwe, there is reason to believe that their genomes are not excessively divergent (in comparison to laboratory strains such as *Hik-R*). However, since *D. melanogaster* has originated from sub-Saharan Africa, the Zimbabwean strains represent the putative ancestral population with plenty of evolutionary time to diverge or form structured sub-populations. A survey of single nucleotide polymorphisms on the X chromosome done in a European and an African population (including *Zim157*, but not *Zim53*) of *D. melanogaster* could help facilitate the search for functional differences. On the other hand, the number of these functional differences is expected to be small, so the search for them could become overly difficult and time-consuming. A restriction to known transcription factors could help here, while it remains unclear whether to look for a structural change in this factor (a non-synonymous mutation) or for another regulatory mutation.

Finally, efforts could be made to bring the 3$^{rd}$ chromosome (and also the 2$^{nd}$ chromosomes that were so far not used for our analyses) with the promoter-*lacZ* insertion into an otherwise complete *Zim53* background. That way the double set of potential transcription factors on the second chromosome (or on the third) and X chromosome (only in females) would be present and testable as to whether this is to influence transgene expression. This could the be done in flies that are either homozygous or heterozygous for the insert, thereby also checking if a certain level of transcription factors has different influence on one or two copies of the transgene.

While transgenic approaches like ours are suited to investigate functional properties of *cis*-regulatory DNA, they merely provide evidence in a proof-of-principle way. In this context, however, it was good to select the gene *CG13360* about which only scarce information exists. Thus, new insights about gene expression regulation have been gained in an unbiased manner so that they should in principle be applicable to a majority of genes. Nevertheless, should the present approach be extended to a larger number of candidate genes with differences in expression, it would be advisable to also include genes of which a better *a* *priori* knowledge exists. This could facilitate follow-up experimental work, once the starting experiments have revealed some new and interesting finding.

An extension to more candidate genes would not only improve the generalization of results, but also enable a comprehensive analysis of polymorphisms in regulatory DNA. Apart from the problem of where to locate binding sites for gene regulating proteins, a statistical analysis of non-coding DNA (*i.e.* intergenic regions and introns) that would only have to be sequenced will be capable of assessing the genetic variability in these regions. Questions to be addressed are: How variable is non-coding DNA within and among populations and species? Are the levels of variability different from those at synonymous sites? Are there signatures of selection in these regions? If so, what type of selection was acting? To answer these questions, current statistical methods might not be sensitive enough and must be improved in resolution to reliably locate a target of selection. This could be complicated by the fact that transcription factor binding sites are not necessarily well-conserved in sequence and arrangement. HARE *et al.* (2008) sequenced the complete *even-skipped* (*eve*) locus with the entire surrounding region of well-known transcription factor binding sites in a number of *Drosophila* species and six species of scavenger flies (Sepsidae), which share basic patterns of developmental gene expression. The overall sequence similarity was low, especially between the two groups of distantly related species, fruit and scavenger flies, which did not come as a surprise. However, when the regulatory DNA of scavenger flies was introduced into the *D. melanogaster* genome, it produced nearly identical expression patterns of the *eve* gene during embryonic development. The fine-scale analysis revealed substantial re-arrangement of transcription factor binding sites with an element of conservation: specific pairs of sites were either overlapping or adjacent to each other. The authors conclude that large-scale arrangement of binding sites can alter as long as specific requirements about fine-scale arrangement are met. This example demonstrates that known transcription factor binding sites are so small (6 – 30 bp) that already rearranging them can decrease sequence similarity to such an extent that despite their presence motif-finding methods based on sequence conservation would mostly fail to detect them. Thus, unless an *a priori* knowledge about binding sites is available finding new examples of them by sequence conservation remains a difficult task. Surely, this issue also depends on the genetic divergence of the species under consideration.

Another question about gene expression regulation and its association with *cis*-regulatory polymorphisms is as to whether the type of mutation makes a difference. Is it only a point mutation (generating a SNP) that can significantly alter gene expression, or does it have to be more of them? What role do insertions and deletions play? Are they perhaps more effective in modifying expression? In a study of DABORN *et al.* (2002) the insertion of an *Accord* transposable element into the 5' region of the gene *Cyp6g1* (a cytochrome p450 gene) was shown to impart overtranscription of the gene, thereby conferring DDT resistance, and that this overexpression is necessary and sufficient for resistance. Or is it really the relative arrangement of several binding sites to each other? Is it then the spacing of the binding sites what matters more? Moreover it is also about existing and known binding sites on the one hand, and their capacity to maintain proper function in presence of mutations on the other. To address this question it is necessary to know more about the physical process of protein-DNA binding and its relationship to the production of different amounts of transcript. In addition, chromatin status and epigenetics can also contribute to gene expression. This would clearly lead away from evolutionary biology, but it could likewise clearly help elucidate evolutionary questions about gene expression regulation by finding all possible mechanisms of expression regulation and their degree of heritability. Thus, evolutionary biologists and population geneticists can develop an idea of where and how to look for the outcome of selection.

An interesting finding of this study was that the chromosomal context in which the transgenic insertions had been embedded had a major effect on transgene expression. This was already observable with a restriction to fly strains with 3rd chromosome inserts. To generalize this also to insertions on the 2nd chromosome and partially the X chromosome (which due to its role as sex chromosome behaves differently in terms of gene content, molecular evolution and expression patterns) it would be recommendable to map the chromosomal position of our inserts. If it turned out that there is no insertional bias among the 3rd chromsome inserts, then there would be reason to believe that also inserts on the 2nd chromosome would show such a degree in expression variation caused by chromosomal environment, thus resulting in a general feature of (at least) autosomes. If chromosomal context in otherwise genetically identical flies makes such a difference, genes should change their position in the genome in order escape the local expression regime, either to increase or decrease their transcript abundance. Indeed, genes would actually have a way to do so, by means of duplication by retrotransposition. Here, an mRNA can be reverse transcribed into cDNA and afterwards integrated randomly into the genome. The gene normally loses its intronic sequences and becomes a pseudogene due to a lack of proper *cis*-regulatory DNA in the genetic surroundings. But if by chance this new environment provides the latter, it can also become an active duplicated gene open the processes of neo- and subfunctionalisation. It has been shown that in *Drosophila* there is an excess of autosomal duplicated genes that were derived from X-chromosomal parental genes by retrotransposition. As many of these genes were expressed specifically in testis leading to a hypothesis regarding X chromosome evolution, this example nevertheless shows the possibility of gene relocation for better expression. Although those testis-expressed genes are thought to have been relocated in order to escape X inactivation, there might also be reasons for autosomal genes to look for an optimal chromosomal position for receiving rough-scale expression levels, while the process of fine-scale tuning of expression could be governed by *cis*-regulatory polymorphisms. An alternative view would be that because all of our transgenic inserts had an otherwise identical genetic background, the chromosomal environment of each insert provides important additional *cis*-regulatory binding sites that account for the observed significant reporter gene expression differences. So it might be worth searching for these binding sites in the chromosomal neighborhood, once the position of each of our inserts in the genome has been determined. It is not unlikely that it is generally a combination of local regulating sequences, both silencers and enhancers, that together establish a spatial and temporal gene expression pattern required for proper gene performance, and that the number of such regulating sequences increases with the functional role of the gene in the organism. The more often a gene is used and the more functions it is required for, the more complex the regulating system will be, with more binding sites for transcription factors to allow for a precisely adjustable expression. Not surprisingly, many of the genes with a very large *cis* region are developmental genes that orchestrate this highly elaborate process.

# Chapter 3

## Experimental increase of codon bias in the *Drosophila Adh* gene has no effect on ADH protein expression

DESPITE the redundancy of the genetic code, synonymous codons are not used with equal frequency – a phenomenon known as codon bias (IKEMURA 1981). Codon bias is apparent in the genomes of a wide array of organisms including eubacteria, archaea, and both unicellular and multicellular eukaryotes; it is essentially a universal property of genomes. The two main hypotheses that have been proposed to account for synonymous codon bias are 1) mutational bias (including biased gene conversion), and 2) natural selection for translational accuracy and/or efficiency (reviewed by AKASHI 2001; DURET 2002).

In *Drosophila*, several lines of evidence suggest that codon bias results from natural selection for translational accuracy and/or efficiency. The lack of a significant association between intronic and synonymous site base composition indicates that mutational bias cannot account for codon bias (VICARIO *et al.* 2007). Optimal codons, those synonymous codons whose usage shows a statistically significant increase in frequency with increasing gene expression (DURET and MOUCHIROUD 1999), tend to match the most abundant species of isoaccepting tRNA (MORIYAMA and POWELL 1997). Codon bias is most extreme in highly expressed genes (SHARP and LI 1986; DURET and MOUCHIROUD 1999) and is significantly higher in the functionally constrained codons of proteins (AKASHI 1994). These observations support the hypothesis that codon bias results from natural selection for translational accuracy and efficiency (BULMER 1991), referred to herein as the translational selection hypothesis.

Although there is a substantial body of indirect evidence for translational selection driving synonymous codon usage in *Drosophila*, direct experimental evidence for the translational selection hypothesis is comparatively sparse. Experimental reduction of codon bias in the leucine codons of the alcohol dehydrogenase (*Adh*) gene, the most highly biased codon family in one of the most highly expressed genes in the *Drosophila melanogaster* genome, resulted in a significant reduction in ADH protein expression (CARLINI and STEPHAN 2003) and rendered flies less tolerant to ecologically relevant levels of environmental ethanol

**Table 3.1**

**Codon usage bias in the leucine codons of *D. melanogaster***
**(compiled by N. ANDERSON and D. CARLINI)**

| Codon | Genome-wide<br>Usage[a] (%) | Genome-wide<br>RSCU[b] | Weakly expressed<br>RSCU[c] | Highly expressed<br>RSCU[d] | ΔRSCU[e] |
|---|---|---|---|---|---|
| TTA | 4.30 | 0.26 | 0.38 | 0.21 | -0.17 |
| TTG | 17.70 | 1.06 | 1.18 | 1.08 | -0.10 |
| CTA | 9.20 | 0.55 | 0.57 | 0.44 | -0.13 |
| CTC | 15.40 | 0.92 | 0.93 | 0.90 | -0.03 |
| CTG | 43.50 | 2.61 | 2.33 | 2.81 | 0.48 |
| CTT | 9.80 | 0.59 | 0.61 | 0.56 | -0.05 |

[a] Genome-wide codon usage (n = 13,464 genes) from HAMBUCH and PARSCH (2005)
[b] RSCU = relative synonymous codon usage (SHARP *et al.* 1986)
[c] Genes in the lowest 5% of expression determined by microarray hybridization (GIBSON *et al.* 2004)
[d] Genes in the highest 5% of expression determined by microarray hybridization (GIBSON *et al.* 2004)
[e] The difference in RSCU between highly expressed and weakly expressed genes. Codons for which ΔRSCU > 0 are defined as optimal codons (DURET and MOUCHIROUD 1999)

(CARLINI 2004). However, to date no studies have been conducted to examine the functional effects of experimentally increased codon bias in *Drosophila*. This is a significant consideration, because the levels of codon bias observed in the most highly expressed genes rarely approach the theoretical maximum. At present, it is unclear whether this reflects the shape of the fitness curve for codon bias (*i.e.*, diminishing returns due to tRNA saturation), interference from adaptive amino acid substitutions within the same gene (Betancourt and PRESGRAVES 2002; COMERON and KREITMAN 2002; HAMBUCH and PARSCH 2005), or some trade-off between translational selection and other factors which influence synonymous codon usage such as mRNA stability (CARLINI *et al.* 2001; CHAMARY and HURST 2005a), exonic splice enhancers (WILLIE and MAJEWSKI 2004; CHAMARY and HURST 2005b; PARMLEY and HURST 2007), and/or transcription driven mutagenesis (HOEDE *et al*. 2006).

In this study we build on previous work (CARLINI and STEPHAN 2003), again focusing on leucine codons in the *Adh* gene because of the high levels of codon bias observed in *Drosophila* leucine codons (Table 3.1). Overall, *Adh* has a frequency of optimal codon usage (Fop; IKEMURA 1981) of 75%, and 20 of its 27 (74%) leucine codons are the optimal CTG. To investigate the effect of increasing codon bias on ADH protein expression, we performed site-directed mutagenesis to replace the seven suboptimal leucine codons with the optimal CTG codon. The *in vivo* ADH activity imparted by the mutant allele was compared to that of the wild-type allele in stable transformed lines that otherwise lacked a functional *Adh* gene. Using standard transformation methods, we were unable to detect a difference in ADH expression between wild-type and mutant transformants. However, the use of a more sensitive transformation method that eliminates genomic position-effect variation on transgene expression (SIEGAL and HARTL 1996, 1998; PARSCH 2004) revealed a marginally significant decrease in ADH expression in transformants with the mutant *Adh* allele. These results suggest that there are diminishing returns of increased codon bias with respect to translational efficiency and/or that additional selective constraints limit optimal codon use in the *Adh* gene.

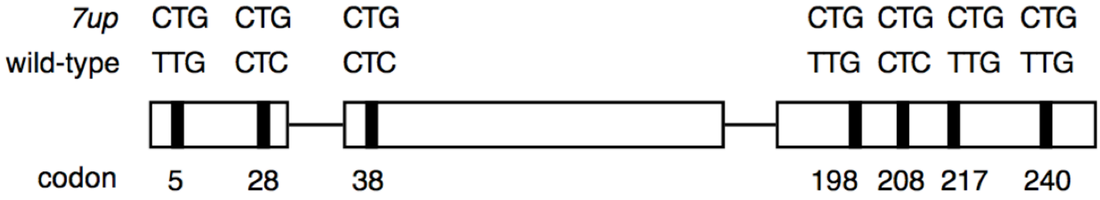
## 3.1 MATERIALS AND METHODS

### 3.1.1 Site-directed mutagenesis

The *Adh Wa-F* allele (KREITMAN 1983) was used as the wild-type allele for all experiments. For mutagenesis, an 8.6-kb *Bgl*II fragment containing the complete *Adh* transcriptional unit and ~5.5 kb of upstream flanking region was excised from the plasmid pΔWaf2a (CHOUDHARY and LAURIE 1991) and inserted into the *Bam*HI site of the vector pBluescript SK+ (Stratagene, La Jolla, CA). To facilitate subsequent cloning steps, an *Xho*I linker sequence (New England Biolabs, Ipswich, MA) was then inserted into the *Spe*I site. The resulting plasmid, designated as pBSX2a, served as the template for site-directed mutagenesis.

The wild-type *Adh* allele contains 27 leucine codons, seven of which are not the optimal CTG (Figure 3.1). All seven of these codons (either TTG or CTC) were changed to CTG using the QuikChange XL site-directed mutagenesis kit (Stratagene, La Jolla, CA). The primer pairs used for mutagenesis were as follows (given are the forward primers in 5'– 3' orientation, the reverse primers are complementary; the mutated nucleotide is underlined): Leu5$_{T\rightarrow C}$ (CC ATG TCG TTT ACT $\underline{C}$TG ACC AAC AAG AAC GTG ATT TTC GTG GCC G), Leu28$_{C\rightarrow G}$ (C ACC AGC AAG GAG CTG CT$\underline{G}$ AAG CGC GAT CTG AAG GTA AC), Leu38$_{C\rightarrow G}$ (G AAC CTG GTG ATC CT$\underline{G}$ GAC CGC ATT GAG AAC CCG GC), Leu198$_{T\rightarrow C}$ (G CAC ACG TTC AAC TCC TGG $\underline{C}$TG GAT GTT GAG CCT CAG G), Leu208$_{C\rightarrow G}$ (GTT GCC GAG AAG CTG CT$\underline{G}$ GCT CAT CCC ACC CAG C), Leu217$_{T\rightarrow C}$ (ACC CAG CCC TCG $\underline{C}$TG GCC TGC GCC GAG AAC), and Leu240$_{T\rightarrow C}$ (CC ATC TGG AAA CTG GAC $\underline{C}$TG GGC ACC CTG GAG GC). The resulting *Adh* allele with seven suboptimal codons replaced by optimal codons was designated as *7up* and its sequence was confirmed using a

MegaBACE automated sequencer and the DYEnamic ET terminator cycle sequencing kit (Amersham Biosciences, Buckinghamshire, UK).



*7up* CTG CTG CTG CTG CTG CTG CTG
wild-type TTG CTC CTC TTG CTC TTG TTG
codon 5 28 38 198 208 217 240

**Figure 3.1 Comparison of wild-type and *7up* alleles of *Adh*. –** Seven suboptimal leucine TTG or CTC codons in the wild-type allele were replaced with optimal CTG codons to construct the *Adh 7up* allele. Boxes indicate exons, horizontal lines indicate introns. The locations of the suboptimal codons within the coding sequence are indicated by black rectangles and the amino acid positions are given below. (Mutagenesis and transformation vector construction performed by S. HUTTER.)

### 3.1.2 Transformation vector construction

For standard *P*-element mediated germline transformation we used the YES transformation vector, a *P*-element vector containing the *D. melanogaster yellow* (*y*) gene as a selectable marker (PATTON *et al.* 1992). The YES vector was used in previous experiments involving the reduction of codon bias in the *Drosophila Adh* gene (CARLINI and STEPHAN 2003) and its use in the present study thus provides a means of directly comparing the effects of increasing codon bias with previous results. To introduce the *7up* allele of *Adh* into the YES vector, an 8.6-kb *Cla*I fragment containing *7up* was excised from the plasmid pBSX2a (described above) and ligated into the *Cla*I site of the YES vector. The sequence of the *7up* allele in the YES vector was confirmed by DNA sequencing using a LI-COR 4300 automated sequencer and the SequiTherm EXCEL II DNA cycle sequencing kit cycle (Epicentre Biotechnologies, Madison, WI). The final transformation vector was designated as p*P[YES-7up]*.

For transgene coplacement, an 8.6-kb *Bgl*II fragment containing the wild-type *Adh* gene was excised from the plasmid pΔWaf2a and cloned into the *Bam*HI site of the vector p*P[wFl]* (SIEGAL and HARTL 1996). This vector contains two cloning sites for inserts that are to be compared, each flanked by target sequences for a different site-specific recombinase, FLP or Cre. The cloning and recombination sites also flank the *mini-white* (*w*) gene of *D. melanogaster*, which serves as a selectable eye-color marker. The *Bam*HI site is located upstream of the *w* gene and is referred to as cloning site 1. The *7up* allele of *Adh* was excised from the pBSX2a mutagenesis vector as an 8.6-kb *Xho*I fragment and inserted into the *Xho*I site of p*P[wFl]* (cloning site 2 located downstream of the *w* gene), which already contained the wild-type allele at cloning site 1. This final vector was designated as p*P[wFl-2a-7up]*. In this construct, the two alleles of *Adh* are arranged in a head-to-head orientation, meaning that they are transcribed from opposite strands of the DNA. This is not expected to affect their relative expression in a systematic way, as long as pairs of coplaced alleles are compared (see below).

### 3.1.3 Germline transformation

Germline transformation using the p*P[YES-7up]* vector was performed by microinjection of *y w*; *Adh*$^{fn6}$; Δ2-3, *Sb*/TM6 embryos. *Adh*$^{fn6}$ is a null allele (splicing defect) that produces no detectable ADH protein (BENYAJATI *et al.* 1982). The Δ2-3 *P* insertion on the third chromosome served as the source of transposase (ROBERTSON *et al.* 1988). Following injection, surviving adults were crossed to *y w*; *Adh*$^{fn6}$ flies and transformant offspring were identified by their wild-type body color. Additional lines with inserts at unique chromosomal locations were generated through mobilization crosses as follows. Transformants carrying insertions on the X chromosome were crossed to the *y w*; *Adh*$^{fn6}$; Δ2-3, *Sb*/TM6 stock and transformants carrying insertions linked to the *Sb* marker (*i.e.*, those with insertions linked to the source of transposase) were crossed to the *y w*; *Adh*$^{fn6}$ stock. Mobilized insertions were identified as *y*$^{+}$ offspring where the *y*$^{+}$ marker was not segregating with the same chromosome as the parental insert. When necessary, further crosses to the *y w*; *Adh*$^{fn6}$ stock were performed to remove the Δ2-3 source of transposase and establish stable transformed lines. Southern blots were performed to confirm that transformant lines contained a single insertion of the transgene (one line was found to contain a double insert and was not used in subsequent analyses). Only autosomal-insertion lines were used for subsequent analysis. For comparison, previously-described transformants carrying the wild-type *Adh* allele were used (PARSCH *et al.* 1999, 2000; CARLINI and STEPHAN 2003).

Germline transformation using the p*P[wFl-2a-7up]* vector was performed by microinjection of *y w*; Δ2-3, *Sb*/TM6 embryos. This strain carries the endogenous *Adh* gene that was later removed through crossing (see below). Successfully transformed flies showing red eye color were crossed to *y w* flies to remove the source of transposase (if still present) and establish stable transformed lines. In cases where the transgene inserted onto the third chromosome carrying the transposase gene, the insert was immediately re-mobilized by crossing with *y w* flies and selecting for offspring with red eyes and lacking the *Sb* marker (indicating the absence of the chromosome carrying the transposase gene). Transformed lines were then crossed to a strain with multiple phenotypic markers (*y w*; *CyO/Sco*; *Ubx/Sb*) to determine which chromosome contained the insertion and to establish homozygous lines for each independent insertion. Only lines with insertions on the third chromosome were used for subsequent transgene coplacement.

### 3.1.4 Transgene coplacement

Following the protocol of SIEGAL and HARTL (1996), females of the transformed fly strains homozygous for p*P[wFl-2a-7up]* insertions were mated with males from a stock carrying both the *FLP* and *cre* recombinase genes (*y w*; *MKRS*, *FLP*/*cre*, TM6B), thereby producing offspring with one of the recombinase genes on one third chromosome and the transgenic insert on the other. These two types of flies were separated and treated independently. In the first treatment, *cre* expression was induced by rearing the flies at 25 °C to excise the wild-type *Adh* allele along with the *w* gene. In the second treatment, *FLP* expression was induced by heat shock at 38 °C for 1 hour during the first larval instar stage, which resulted in the removal of the *7up Adh* allele together with the *w* marker gene. In both cases, successfully excised alleles generated flies with white eyes. Additional crosses to the above marker strains were performed to remove the recombinase genes and establish lines homozygous for their respective *Adh* inserts. This resulted in matched pairs of fly strains with homozygous third chromosome insertions of either the wild-type or the *7up* allele of *Adh*.

Since the original injection stock and all other flies used in the above crossing scheme carried the endogenous *Adh* gene on chromosome 2, we performed additional crosses to remove the endogenous gene so that the two introduced *Adh* alleles could be tested in an otherwise *Adh*-null background. This was done with crosses to a stock of *y w*; *Adh$^{fn6}$* flies and the above mentioned strain *y w*; *CyO/Sco*; *Ubx/Sb*.

### 3.1.5 ADH activity assays

Males of all transformed lines were crossed to *y w*; *Adh$^{fn6}$* females to produce offspring heterozygous for their respective *Adh* insertion in an otherwise *Adh*-null genetic background. These offspring (males aged 6–8 days) were used for ADH assays following standard protocols (MARONI 1978) using isopropanol as the substrate. ADH activity units were defined as µmol of NAD$^+$ reduced per minute per mg of total protein (multiplied by 100). For the *P[YES-7up]* transformants, the total protein concentration of the crude extracts was determined using the *RC DC* Protein Assay kit (Bio-Rad Laboratories, Hercules, CA). For the

*P[wFl-2a-7up]* transformants, total protein concentration was estimated by the method of LOWRY *et al.* (1951).

The above crosses were repeated in two separate blocks, and from each cross two independent cohorts of five flies each were used for ADH assays. This resulted in a total of four ADH activity measurements for each transformed line. For the *P[YES-7up]* transformants, a one-way nested ANOVA was used to test the null hypothesis of no difference in ADH activity between genotypes. For the *P[wFl-2a-7up]* transformants, the coplaced pairs of alleles at each genomic location were used for a paired *t*-test for ADH activity differences between wild-type and *7up* lines.

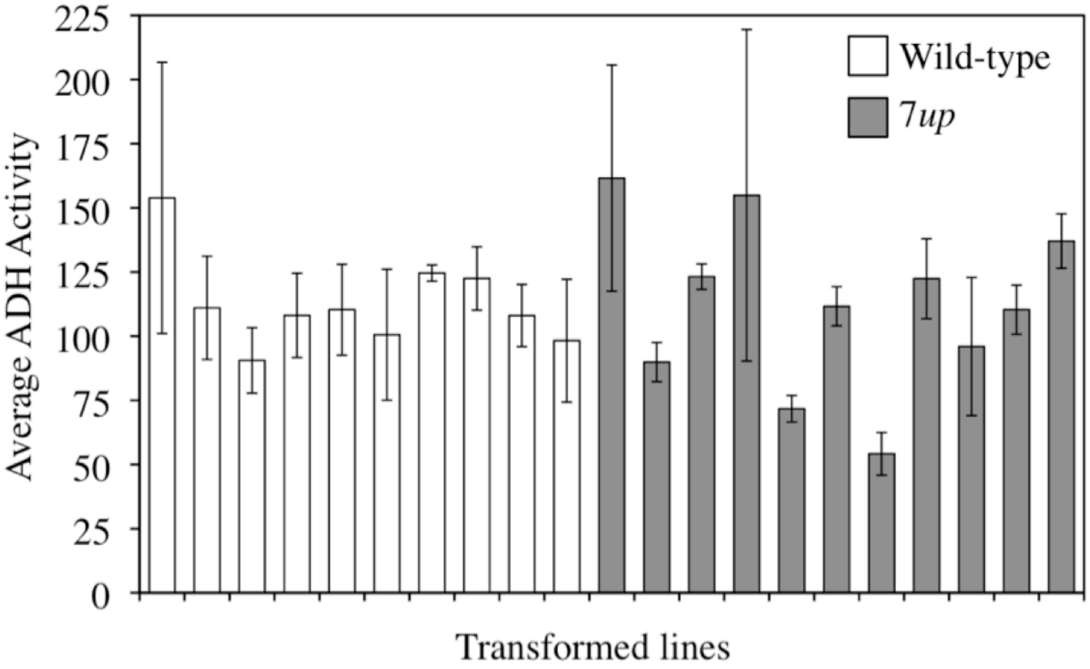


**Figure 3.2 Comparison of enzymatic activity: YES vector.** – ADH activity of wild-type (open bars) and *7up* (filled bars) lines obtained from standard *P*-element transformation. ADH activity did not differ between the two genotypes (Nested ANOVA, $p = 0.953$), although there was significant position-effect variation among lines within genotypes (Nested ANOVA, $p = 0.043$). Error bars indicate ± 1 standard deviation from the mean. (Assays performed and figure designed by N. ANDERSON and D. CARLINI.)

## 3.2 RESULTS

We used site-directed mutagenesis to create an allele of *Adh* in which the seven suboptimal leucine codons present in the wild-type sequence were replaced by the optimal codon, CTG (Figure 3.1). This mutant allele was designated as *7up* and was compared to the wild-type allele in transformed lines of *D. melanogaster* that otherwise lacked a functional *Adh* gene. Because the amino acid sequences encoded by the wild-type and *7up* alleles were identical, any differences in ADH activity could be attributed to differences in ADH protein production.

Using standard *P*-element transformation (YES vector), we compared the ADH activity of 10 independent transformed lines with the wild-type *Adh* allele and 11 independent transformed lines with the *7up* allele (Figure 3.2). We observed no difference in ADH activity between the two genotypes (Nested ANOVA, $p$ = 0.953). The average ADH activity of the wild-type lines (112.79 ± 16.83) was virtually identical to that of the *7up* lines (112.09 ± 31.44). Due to the random insertion location of the *Adh* transgenes in the *Drosophila* genome using the YES vector, substantial position-effect variation was observed among lines within genotypes (Nested ANOVA, $p$ = 0.043).

To avoid the problem of position effect variation and increase our power to detect a difference between the wild-type and *7up* alleles, we repeated the above experiment using the method of transgene coplacement (using the "waffle" vector; SIEGAL and HARTL 1996). This method allows us to introduce both alleles into the same chromosomal location and then remove one or the other allele through site-specific recombination. As a result, the two alleles can be compared in an otherwise identical genomic context. In total, we obtained 9 pairs of transformed lines with coplaced wild-type and *7up* alleles on the third chromosome. Overall, the wild-type transformants had slightly higher ADH activity (on average 5% higher than *7up* transformants), which was marginally significant (paired *t*-test, two-tailed $p$ = 0.058; Figure 3.3).

There was a highly significant positive correlation between the ADH activities of transformants with coplaced alleles (Pearson's $R$ = 0.95, $p$ < 0.001; Figure 3.4), which leads to this method having much greater sensitivity than the standard approach. Given our observed variance, the smallest difference between wild-type and *7up* ADH activity that could be detected as significant by a paired *t*-test (two-tailed $p \leq 0.05$) is 5.4%. Our observed difference (5.0%) lies just within this limit. In contrast, if transformants with coplaced alleles are not paired (as would be the case using the standard approach), the smallest difference that could be detected as significant by an unpaired *t*-test (two-tailed $p \leq 0.05$) is 20.1%.

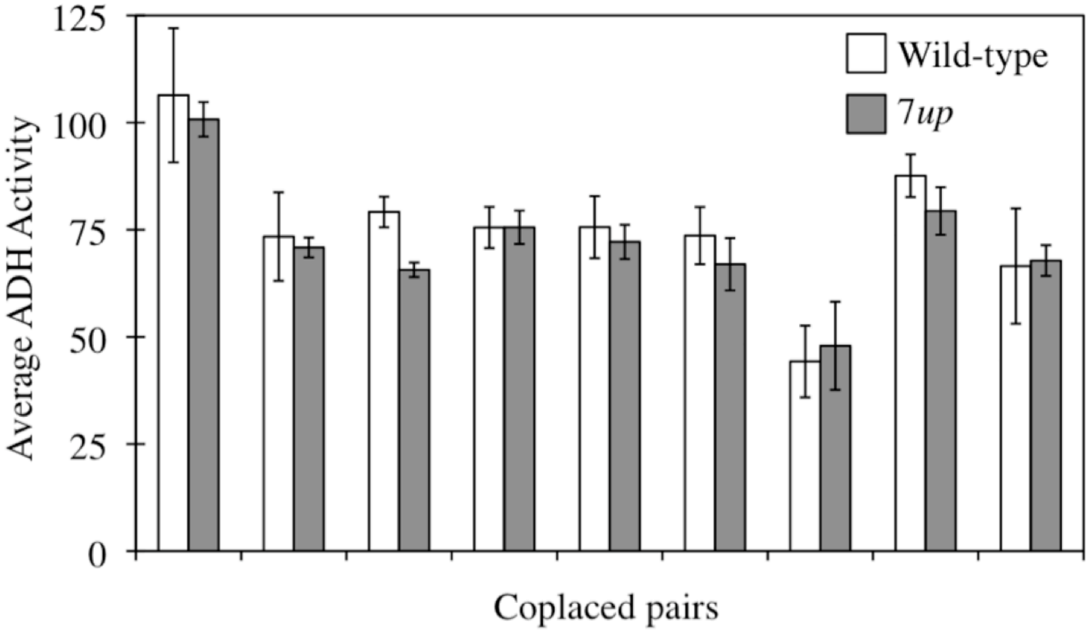


**Figure 3.3 Comparison of enzymatic activity: "waffle" vector.** – Pairwise comparisons of ADH activity between wild-type (open bars) and *7up* (filled bars) transformants generated by transgene coplacement. The overall mean ratio of wild-type to *7up* ADH activity is 1.050 (paired *t*-test, two-tailed $p = 0.058$). Error bars indicate ± 1 standard deviation from the mean. (Assays performed and figure designed by W. HENSE.)

## 3.3 DISCUSSION

Previous work has shown that experimentally decreasing codon bias in the *D. melanogaster Adh* gene leads to a reduction in ADH protein expression (CARLINI and STEPHAN 2003). For example, the introduction of six suboptimal leucine codons reduced ADH expression by 19% – a result consistent with the translational selection hypothesis. In the current study, we have further tested this hypothesis by performing the reverse experiment: codon bias was increased by replacing seven suboptimal leucine codons present in the wild-type *Adh* gene with the optimal leucine codon, CTG. However, the introduction of these optimal codons did not lead to an increase in ADH expression. Our transgene coplacement experiment even suggests that it may have decreased ADH expression. In the following, we consider several possible explanations for this result.

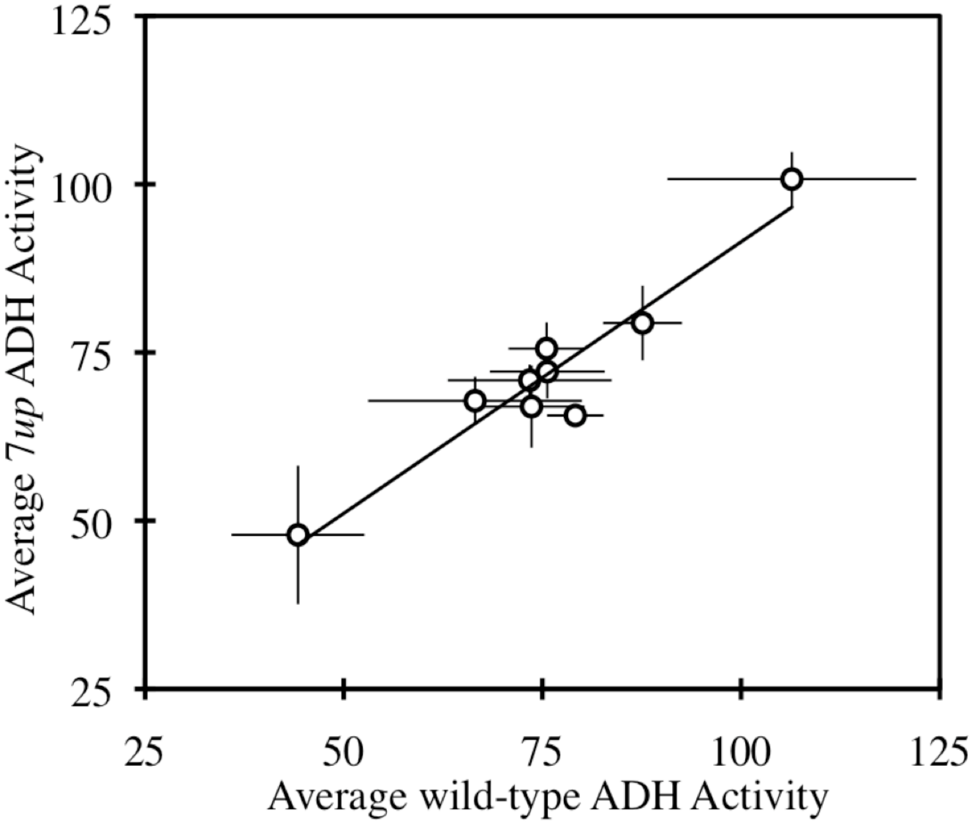


**Figure 3.4 Correlation of transgenic inserts.** – Pairwise comparisons of ADH activity between wild-type (open bars) and *7up* (filled bars) transformants generated by transgene coplacement. The overall mean ratio of wild-type to *7up* ADH activity is 1.050 (paired *t*-test, two-tailed *p* = 0.058). Error bars indicate ± 1 standard deviation from the mean. (Figure designed by W. HENSE.)

An important consideration is that the previous experiments decreased codon bias by replacing leucine codons of the wild-type *Adh* gene with the rarely-used leucine codon CTA – a codon that does not occur naturally in the *Adh* gene. It may be that the introduction of CTA codons has a much stronger effect on protein expression than the introduction of CTG codons. In the wild-type *Adh* sequence, only seven of the 27 leucine codons are not the optimal CTG. This limited our experimental options, as only CTC or TTG codons could be altered. Although both of these codons are used less frequently in highly-expressed genes than CTG, they are not as strongly avoided as CTA (Table 3.1) (CHEN *et al.* 1999). For a limited set of genes, the CTC codon even demonstrated a significant increase in its use as codon bias within a gene increases, causing it to be defined as one of two “preferred” leucine codons (AKASHI 1994, 1995), although our genome-wide analysis of codon usage and gene expression presented in Table 3.1 indicates that CTC codons are used less frequently in highly expressed genes. Nevertheless, there is likely to be an asymmetry in the effects of introducing CTA *versus* CTG codons. Furthermore, it is possible that there are diminishing returns to increasing codon bias with respect to translational efficiency. Because the wild-type *Adh* gene already shows very strong bias in synonymous codon usage, increasing this bias further may have little or no impact on ADH expression. Perhaps this is because the tRNA pool is already saturated, and increasing leucine codon bias has negligible effects because it is the charged $tRNA^{Leu}$ that is limiting translation.

Another possibility is that the optimal codon substitutions improved the efficiency and/or accuracy of translation, but these beneficial effects were obviated by deleterious effects on mRNA stability and/or splicing. Such a scenario could account for the slight reduction of ADH activity observed in *7up* transformants. It is possible to examine the potential effects on mRNA stability by comparing the folding free energies of the most stable global mRNA secondary structures of the wild-type and *7up* transcripts. Since the optimal substitutions involved four T→C changes and three C→G changes, it follows that the *7up* version of *Adh* would have a more stable global secondary structure, with T→C changes increasing the opportunity for more stable G-C pairings and C→G changes allowing for additional G-U pairings at the RNA level. Using the program mfold (ZUKER 2003), we found the most stable secondary structure of the adult primary transcript of the wild-type *Adh* allele to have a folding free energy of –588.5 kcal/mole. The folding free energy of the most stable structure for the *7up* adult primary transcript was –594.9 kcal/mole, a difference of –6.4 kcal/mole. The average folding free energy of the 10 most stable structures for the *7up* adult primary transcripts was significantly less than that of the 10 most stable wild-type adult primary transcripts (*7up*: –589.87 kcal/mole, wild-type: –584.00 kcal/mole; two-tailed *t*-test: $p < 0.001$). Similar results were obtained when considering only the coding regions of the wild-type and *7up* sequences (two-tailed *t*-test: $p < 0.01$). Although statistically significant, it is unlikely that these differences in mRNA stability are large enough to have an effect on translation. Two recent genome-wide studies have shown that global mRNA stablility (measured as folding free-energy of full length mRNA) is not correlated with gene expression in *Drosophila* (STENØIEN and STEPHAN 2005; ECK and STEPHAN 2008).

We also compared the local structures within each of the most stable global structures of the wild-type and *7up* mRNAs and found no evidence for biologically significant differences among local structures. For example, the 13 most stable helices among those in the *7up* and wild-type mRNAs were identical and ranged from –60.1 (24 bp) to –13.3

kcal/mole (7 bp). The remaining helices ranged from –13.3 (7 bp) to –1.3 kcal/mole (2 bp), and no differences greater than 0.9 kcal/mole were observed in a ranked list of helices. None of these minor differences in local structures appears to be sufficient to differentially inhibit the helicase activity of the ribosome, which has been experimentally demonstrated to be capable of melting a highly stable 27 bp helix (–52.1 kcal/mole) without dissociation from the mRNA (TAKYAR *et al*. 2005). Furthermore, the *7up* mutations did not alter any of the putative secondary structural elements identified by previous covariation analysis of multiple *Drosophila* species or by previous experimental manipulation (KIRBY *et al*. 1995; CARLINI *et al.* 2001; PARSCH *et al*. 1997; BAINES *et al*. 2004). We also determined a consensus mRNA secondary structure for *Adh* coding sequences of 12 *Drosophila* species from the recent 12 genomes project using RNAalifold (GRUBER *et al.* 2008). RNAalifold determines a consensus structure of a set of aligned sequences by averaging free energy contributions over all sequences while also scoring covariations to account for compensatory mutations. None of the seven nucleotides we altered were within stem regions of the consensus structure, lending support to the conclusion that the 7*up* mutations did not significantly alter secondary structure. However, we were unable to obtain a reliable alignment of *Adh* pre-mRNA sequences due to substantial variation in the non-coding nucleotides among the 12 sequences, so there remains a possibility that the mutated nucleotides pair with non-coding portions of the pre-mRNA, although the analyses on the *D. melanogaster* wild-type and *7up* pre-mRNA sequences described above do not indicate that this is the case.

The *7up* synonymous substitutions could also reduce ADH expression by altering one or more exonic splicing enhancer (ESE). These *cis*-acting motifs tend to occur near exon-intron boundaries and are enriched in As and diminished in Cs, precisely opposite the pattern observed for optimal codons in *Drosophila* (VICARIO *et al*. 2007). A recent genome-wide survey of *D. melanogaster* exons provided evidence of a trade-off between the use of translationally optimal codons and the regulation of splicing (WARNECKE and HURST 2007). Because three of the *7up* mutations (codons 28, 38, and 208) involved C→G changes, it can be reasoned that they favored the creation of ESEs. However, the other four mutations (codons 5, 198, 217, and 240) involved T→C changes, presumably resulting in the disruption of ESEs. We used two software applications to determine if the *7up* mutations altered splicing motifs in *Adh*. To date most work on the identification of ESE motifs has focused on mammals but many of the SR proteins, which recognize and bind the mRNA at ESEs, are strongly conserved within the metazoa so that many of the ESE motifs identified in mammals are therefore likely to be functional in *Drosophila*. ESEfinder3.0 (CARTEGNI *et al*. 2003) was used to locate these comparatively well-characterized ESEs. Recently a set of ESEs has been identified in *Drosophila* using both the RESCUE-ESE approach of FAIRBROTHER *et al.* (2002) that was successfully used to identify human ESEs as well as ELPH, a general purpose Gibbs sampler for finding sequence motifs (PERTEA *et al.* 2007). The SEE ESE software application (http://www.cbcb.umd.edu/software/SeeEse/index.html) was used to determine whether any putative *Drosophila* ESEs were disrupted or created by the *7up* mutations. The results from these analyses indicate that differences between wild-type and *7up* in ESE content were minimal and that, overall, the *7up* mutations led to a roughly twofold increase in the number of ESEs in the *Adh* gene (Table 3.2). Thus, if anything, these differences are biased in favor of increased splicing efficiency of the *7up* allele, which cannot account for the observed reduction in ADH protein expression.

## Table 3.2

**Total number of exonic splicing enhancers (ESEs) in the wild-type and 7up *Adh* coding sequences as predicted by two methods (compiled by N. ANDERSON and D. CARLINI.)**

| | ESEfinder 3.0 | | SEE ESE | |
|---|---|---|---|---|
| Leucine codon | Wild-type | *7up* | Wild-type | *7up* |
| 5 | 0 | 1 | 0 | 0 |
| 28 | 0 | 2 | 0 | 0 |
| 38 | 1 | 1 | 0 | 0 |
| 198 | 0 | 0 | 1 | 0 |
| 208 | 2 | 2 | 1 | 3 |
| 217 | 1 | 2 | 0 | 0 |
| 240 | 0 | 0 | 0 | 0 |
| Total | 4 | 8 | 2 | 3 |

A final possibility is that the suboptimal leucine codons present in the wild-type *Adh* gene play a functional role in translational pausing, which has been implicated as a requirement for proper protein folding (BUCHAN and STANSFIELD 2007). If so, we would expect that the degree of functional constraint at these codons would be comparable to that at optimal leucine codons. We evaluated this by comparing levels of overall sequence divergence (including synonymous and nonsynonymous substitutions) at homologous positions in the *Adh* genes of 12 *Drosophila* species (*DROSOPHILA* 12 GENOMES CONSORTIUM 2007). In pairwise comparisons among the 12 *Adh* homologs, the average nucleotide sequence divergence for the entire coding sequence was 16.76%, whereas that at the seven suboptimal leucine codons was 25.76% (Table 3.3). For comparison, we calculated the average sequence divergence at the twenty preferred CTG leucine codon positions and obtained a value of 19.67%. Because these comparisons involved a relatively wide range of divergence times, we also calculated sequence divergence for two more restricted subsets of taxa, i) in the subgenus *Sophophora*, and ii) in the *melanogaster* subgroup for *Adh* as a whole, for the seven suboptimal, and for the 20 optimal leucine positions and found that the pattern was more extreme in the more restrictive taxonomic groups (Table 3.3). Thus, if anything, there appears to be less functional constraint at the seven suboptimal leucine codon positions, consistent with AKASHI (1995) and inconsistent with the idea that these suboptimal codons are adaptively positioned to ensure proper co-translational folding of the nascent polypeptide.

Although we cannot rule out minor effects, our analyses suggest that the *7up* mutations were unlikely to significantly alter the *Adh* mRNA secondary structure, splicing code, or translational pausing. If so, it implies that native levels of optimal codon usage in the leucine codons of the *Adh* gene cannot be altered to improve translational efficiency and/or accuracy. This may be due to a saturation of the available tRNA$^{Leu}$ pool. Future experiments that alter codon bias and tRNA expression individually and in combination could test this hypothesis, as well as shed light on the coevolutionary dynamics that led to the emergence of codon bias as a ubiquitous feature of genomes.

## Table 3.3

**Average uncorrected pairwise sequence divergences (%) for the entire coding region, at the 20 optimal leucine codons, and at the seven suboptimal leucine codons of the *Adh* gene (compiled by N. ANDERSON and D. CARLINI.)**

| Region compared | *melanogaster* subgroup[a] | Subgenus *Sophophora*[b] | 12 *Drosophila* species[c] |
|---|---|---|---|
| Entire *Adh* coding region | 3.61 | 11.70 | 16.76 |
| 20 optimal CTG codons | 0.67 | 11.25 | 19.67 |
| 7 suboptimal codons | 6.67 | 24.34 | 25.76 |

[a] *D. melanogaster*, *D. simulans*, *D. sechellia*, *D. yakuba*, and *D. erecta*

[b] *melanogaster* subgroup + *D. ananassae*, *D. pseudoobscura*, *D. persimilis*, and *D. willistoni*

[c] *Sophophora* subgenus + *D. mojavensis*, *D. virilis*, and *D. grimshawi*

# Concluding Discussion

THE regulation of genetic activity or gene expression has become a tremendously important topic in current biology since it is key to our understanding of many quantitative aspects of life in general. As shown by research of the last decades and even centuries, living organisms appear as genetic entities whose genetical inventory – to resume the INTRODUCTION of this thesis – has first to become unleashed by physical and later on also chemical stimuli in order to build up the phenotypically visible body of animals, plants, fungi and, although to a much smaller degree, also bacteria. The form that arises from this process of development is not restricted to the emergent macromolecular continuum, but goes down even to the discontinuous level of single molecules like proteins, and has a defined structure with a likewise demand on space. On all of these levels, morphology or form is always accompanied by and deeply connected to its function (physiology), and the ever tantalizingly puzzling question has been as to how this form emerges, whereas it is the function that determines the selective or fitness value of a given form. Through this interaction species are able to change which process is termed *descent with modification* or simply *evolution*. And it is this interaction that gave rise to the new field of *evolutionary developmental genetics* or *evo-devo*. Although the scope of this field was at the center of evolutionary thought it became possible only in recent years to address old questions with newly developed experimental techniques and methods as well as new data that was brought to light by new high-throughput methods in the fields of genomics, transcriptomics, and proteomics. In this thesis, transgenics, one of these experimental methods, were applied to address three questions regarding the regulation of gene expression and its implications for evolutionary biology.

Gene expression regulation can take place in many ways and at least on two levels, the level of individual genes or groups of related genes (as in the case of bacterial operons) and on a larger scale which involves entire chromosomes. Molecular biological questions regarding this are the nature of the mechanism of gene regulation, evolutionary questions, on the other hand, deal with the heritable part of this process, the variability in the molecules involved, and finally the modes of selection responsible for maintaining or purging molecular variants with advantageous or deleterious forms of gene regulation, respectively. In this thesis, I tried to shed light on a mechanism by which the X chromosome of *Drosophila melanogaster* becomes transcriptionally inactivated in the male germline, during spermatogenesis. Genome sequencing studies in the fruit fly in the recent past have revealed interesting properties of genomes concerning the distribution of genes. One of these findings has been that there is an excess of retrotransposed genes where a formerly X-linked copy had been relocated to an autosome, and that in many cases those genes stay active and are not transformed into a pseudogene. Furthermore, most of these transposed genes are transcribed in the testis, *i.e.* the male germline (BETRÁN *et al.* 2002). A second observation that came from genomic studies was that genes whose expression is enriched in or even entirely restricted to male individuals are underrepresented on the X chromosome (PARISI *et al.* 2003; RANZ *et al.* 2003). Moreover, the degree of male bias in expression is negatively correlated with the probability of a gene's being located on the X chromosome (CONNALLON and KNOWLES 2005). Among the first hypotheses to explain the observed was the one about an X chromosome inactivation that only occurs in the male germline. If this happens male-biased genes should be selectively favored to be autosomally located, especially if they are expressed late during spermatogenesis and in testis, to avoid the inactivation of the X chromosome which is thought to happen early in the process of spermatogenesis, when the autosomal genes are still actively transcribed (BETRÁN *et al.* 2002). An important point to make here is that this type of inactivation is restricted to the male germline, in contrast to and not to be confused with the well-known X inactivation that occurs in female mammalian somatic cells as a means to enable dosage compensation of X-linked genes (LYON 1961), since male mammals represent the heterogametic sex in this taxonomic group, which is the case in *Drosophila*, too.

Because the X chromosome inactivation hypothesis is not able to explain all observations of genomic studies concerning the distribution of male-biased or male-enriched genes, *e.g.* the lack of male-biased genes on the X chromosome even if they are entirely expressed in somatic cells and not in the germline (PARISI *et al.* 2003; SWANSON *et al.* 2003), additional attempts to explain the patterns were made. The phenomenon of sexual antagonism is described as the presence of mutually deleterious effects that sex-biased genes can have on the respective other sex (RICE 1984, CHARLESWORTH *et al.* 1987). For instance, a gene that is active and biased in expression in females of an organism may have detrimental effects on males, and *vice versa*. Given this and the fact that the X chromosome as genetic entity spends more time in females than in males, the hypothesis of sexual antagonism predicts an accumulation of female-beneficial/male-detrimental genes on the X and, at the same time, a removal of male-beneficial/female-detrimental genes from the X. The same would be true for mutations in these kinds of genes. Over evolutionary time this would altogether lead to a “feminization” or “demasculinization” of the X chromosome (PARISI *et al.* 2003). Finally, there is a combination of the two hypotheses above, the SAXI hypothesis (sexually antagonistic X inactivation; WU and XU 2003). It claims that due to the effects of sexual antagonism the X chromosome will eventually be inactivated when the feminization of the X has reached a level at which it is no longer tolerable in males, and there especially in the germline where the antagonistic effects are expected to be most pronounced.

Since conclusive evidence for X inactivation had been lacking, an experimental test for such an inactivation was put forth in this thesis. A testis-specific reporter gene construct was used for transgenesis of *D. melanogaster* flies, and it was shown that insertions of this construct on the X chromosome were expressed at very low levels, close to zero, while insertions on the two autosomes, the second and third chromosome, show substantial activity in expression. The fold-difference in expression was up to 10-fold. I tested this construct by using two different transgenesis vectors and by measuring the reporter gene expression by assaying both reporter enzyme activity and levels of transcript abundance (by quantitative RT-PCR). The insertion sites and chromosomal locations of the almost 50 different transgene insertions on the X and the autosomes were furthermore mapped by inverse PCR. It turned out that the possibility of an insertional bias could be excluded since the insertions covered the euchromatic portions of all chromosomes. Thus, X chromosome inactivation indeed seems to take place in the male germline (reporter gene activity in tissues other than testis was negligible). What would be worth investigating in the future is whether the entire X chromosome gets inactivated or if specific regions or parts of it can escape it. An important observation regarding this is that the X chromosome is not completely devoid of male-biased or testis-expressed genes. Thus, these genes may be either transcribed or hyper-transcribed early in spermatogenesis and/or fall into regions of the X chromosome that are omitted from inactivation. A specific region of the X even shows an accumulation of newly-evolved, testis-expressed genes (LEVINE *et al.* 2006; BEGUN *et al.* 2007; CHEN *et al.* 2007), and it would be tempting to try and get some insertions near or even into this region. The expression of the reporter gene in such a case is expected to be restored. Furthermore it would be interesting to determine the size of such an escaping region and, in the case that it is considerably small, to identify more of those regions. Finally, given the SAXI hypothesis and the distribution of sex-biased genes in *D. melanogaster*, a possible consequence of the rationale behind it would be that the Y chromosome, which exists only in male flies, should then be enriched with male-biased and testis-expressed genes, in addition to a removal of female-biased and ovary-expressed genes from this chromosome (“defeminization/masculinization”). However, the Y chromosome of the fruit fly is almost completely heterochromatic and thus harbors only a small number of functional genes. The process of Y chromosome degeneration, the molecular causes, and the selection pressure supporting this process thus must have been much stronger than selection for the transfer of male-biased and testis-expressed genes to the Y chromosome as an ideal chromosomal location for them. And it is interesting to speculate about these selection pressures.

Thus, in the first chapter of this thesis I demonstrated that X chromosome inactivation during spermatogenesis in *D. melanogaster* is very likely to occur. However, about the evolutionary reasons of this process, *i.e.* the fitness advantage of an inactivated X chromosome in male fruit flies can be only speculation (well-reasoned speculation nevertheless) since many evolutionary processes are unique events in evolutionary and hence historical time, which can hardly be repeated (because the circumstances and conditions at that particular time were themselves unique) and might be irreversible. Moreover, evolutionary processes cannot be directly observed, but instead only be reconstructed with a more or less high certainty by looking on the present-day outcome and carefully interpreting observable facts on the basis of knowledge about evolutionary processes gained so far. Taking this into account, the SAXI hypothesis nevertheless appears as a plausible and attractive explanation as it requires only sexual antagonism (a known and well-supported fact) to explain the inactivation of the X chromosome, something that also occurs in other taxa and in somatic cells (LYON 1961).

Whereas the topic of the first chapter was a type of regulation that involves an entire chromosome, I switched to the level of individual genes and their regulation in chapters 2 and 3. DNA and histone modifications such as methylation and acetylation can prepare chromatin in a way that active transcription of single genes and groups of genes become possible (WANG *et al.* 2004). The main effects of these modifications are the loosening of the dense DNA packaging that is organized in chromatin. That way the transcription machinery consisting of several protein and enzyme complexes is capable of binding to specific DNA sequence motifs to initiate transcription (Interestingly, DNA nucleotide sequences not only possess information in form of their coding potential for polypeptides and proteins, but also structural information as binding partners for protein molecules and hence perhaps a second, structural code). Again the evolutionary question arose what the consequences of variability in the molecules involved are. Among these molecules are protein factors that bind to the DNA (mostly termed transcription factors) and the portion of the DNA close to a gene that is actually bound by the protein factors. The former are called the *trans* factors (since they are themselves encoded elsewhere in the genome in *trans*), the former the *cis* factors (because they lie normally rather close to the transcribed gene in question, on the same continuous DNA strand in *cis*) (WITTKOPP *et al.* 2004). Focusing on the *cis*-regulatory part of transcription initiation and controlling for the genetic background consisting of all relevant *trans* factors, I functionally investigated the putative involvement of DNA sequence variation in *cis*-regulatory DNA in producing different levels of reporter gene transcripts. Here, I started with a survey of gene transcription levels of almost the whole genome of *D. melanogaster* in eight fly strains (MEIKLEJOHN *et al.* 2003) and selected one gene with a pattern of a fixed expression difference between the African half of the eight strains and the remaining half consisting of four non-African fly strains. After sequencing a large upstream region of this gene in all eight strains and revealing a number of fixed sequence polymorphisms that corresponded to the expression differences I tried to functionally evaluate these sequence polymorphisms with regard to their possible role in driving the respective fixed expression difference. The approach that was performed with a bacterial reporter gene construct yielded a remarkable result since an expression difference in reporter gene activity was restored only in the presence of an appropriate African genetic background with all of the *trans* factors. Especially the African X chromosome was required for this expression difference to appear. Thus, it is obviously an interplay of *cis* and *trans* factors that contributes to gene expression, at least in the case of the one gene I analyzed. To explore the conditions of gene expression on a broader scale of many genes or many functional categories of genes one could try and identify *trans*-acting factors on a genomic scale as well as the corresponding DNA sequence motifs. To do this, computational approaches seem appropriate on a rough scale, which should then be analyzed in more detail by experiments. What could turn out as a most valuable tool is a high-throughput method to identify binding motifs and their binding proteins. Recently, ChIP-on-chip technology was developed to achieve exactly this goal (APARICIO *et al.* 2004). Briefly, a protein of interest (like a transcription factor) is allowed to bind all its target DNA sequences *in vivo*. After lysing the cells and shearing the DNA fragments of naked DNA and some with the bound protein are separated by immunoprecipitation (chromatin immunoprecipitation, ChIP) with an antibody specific to the protein. The DNA fragments can then be purified from the protein and labeled with fluorochromes to be poured over the surface of a microarray that carries single stranded DNA fragments as probes (the "chip" part of the name). That way the DNA fragments of interest, which were formerly bound by the protein of interest, can be identified by determining their sequence. Certainly, also this new method will have its technical limitations concerning its accuracy, but it is equally certain that scientific advancement can only be made when it is accompanied by technological progress. For many scientific topics the latter is crucial since the final questions have already been asked and formulated, and they are only awaiting means and methods to address and tackle them. It must also be pointed out that the above-mentioned progress is not restricted to advances in wet-lab technology, but also involves new mathematical and statistical approaches, since many high-throughput methods produce vast amounts of raw data that must be processed in order to gain new insights ("knowledge discovery and data mining"). Regarding the question posed in CHAPTER 2 it is interesting to ask whether *cis*-regulatory polymorphisms alone can make the difference in expression, or whether it is the presence or absence of binding motifs that can better explain expression differences. In the former case it is desirable to decipher a putative *cis*-regulatory code and also ask what the physical basis of its role in transcription initiation and maintenance is. Here, collaboration with physicists could help pursue this plan. In the latter case, however, the role of mobile genetic elements, namely transposable elements, could reinforce research of gene expression regulation (FESCHOTTE and PRITHAM 2007) since they seem to be ideal candidates for providing *cis*-regulatory sequences for a majority of genes due to their ability to relocate themselves within the genome of an organism.

One intricate problem concerning gene expression regulation is the large number of molecular factors that can contribute to it. In the second chapter there was the question of the role of *cis*-regulatory elements in initiating gene transcription. The *trans* factors, which also contribute to gene expression, were controlled for by focusing on a certain genetic background. Once transcription is started and a gene transcript is produced the number of factors even increases making it more difficult to assess the role of a single of these factors in gene regulation. The exchange of synonymous codons is thought to be governed by weak positive selection for translational accuracy and efficiency (AKASHI 2001; DURET 2002). Among the number of gene regulatory processes beyond transcription are some (apart from translation) that can possibly influence the amount of protein produced from the transcript. For instance, exonic splicing enhancers can influence the production of a mature mRNA transcript (CHAMARY and HURST 2005b; PARMLEY and HURST 2007), and mRNA secondary structures can be thermodynamically more or less stable which in turn affects its melting at the ribosomes (CARLINI *et al.* 2001; CHAMARY and HURST 2005a), thereby causing a translational delay. Finally translational pausing that enables the proper three-dimensional folding of the nascent protein presents another possibility to influence gene expression post-transcriptionally (BUCHAN and STANSFIELD 2007). In CHAPTER 3 I reported on an experimental exchange of seven leucine codons in a well-known gene of *D. melanogaster*. This exchange made the gene consist of entirely optimal leucine codons so that the expectation regarding gene expression was an increase in enzymatic activity. However, even after applying two different experimental methods (with the second one being much more sensitive) the expected increase was not observable. Instead the enzymatic activity even decreased relative to the wild-type form of the gene. Although the mRNA secondary structures of the wild-type and the mutated form of the gene were different in their folding free energy, this difference can be regarded as too small to have an effect on translation. An analysis of exonic splicing enhancers showed that the number of such enhancers differs between the two alleles of the gene, but in the wrong direction, *i.e.* the mutated form exhibiting lower enzymatic activity had a larger number of enhancers. The most likely explanation for this is that the increase in the number of splicing enhancers was simply too low to show an effect. Furthermore, computational approaches to determine the number of such enhancers might still have a considerable false-positive rate and hence must be used with caution. Finally, also the possibility of translational pausing caused by the presence of suboptimal leucine codons in the wild-type allele could be rejected since a comparison of nucleotide divergence of the wild-type form of the gene in twelve *Drosophila* genomes revealed much less functional constraint in the seven suboptimal codons than in the optimal ones and the entire coding region. Thus, the most likely explanation for the observed are diminishing returns to the increase of codon bias caused by a saturation of the tRNA$^{Leu}$ pool. This means that the leucine codon composition was already sufficiently optimized, mainly with regard to this tRNA pool. Further experiments concerning codon usage bias should therefore attempt to alter codon bias and tRNA abundance (via its expression) separately and in combination to investigate this new hypothesis originating from the present results.

Concluding, this dissertation strove to shed light on gene expression regulation and its evolutionary implications. The regulation of gene expression, from transcription to translation with additional higher-level layers, appears to be an extremely complex phenomenon in biology. Step by step new modes of regulation are discovered without having completely understood the hitherto known. This thesis focused on an example of higher-level regulation with the discovery that the X chromosome of *D. melanogaster* is inactivated during early spermatogenesis, *i.e.* in the male germline. The evolutionary hypothesis explaining this differs from the well-known X chromosome inactivation that occurs in female mammalian somatic cells. In CHAPTER 2 the role of *cis*-regulatory polymorphisms in intraspecific expression variation is analyzed, whereas in the last chapter codon bias and the question whether there is weak selection for the accuracy and efficiency of translation by optimizing codon usage is the evolutionary topic. Especially in the last two chapters it was also shown that the effects on gene expression were very small when considering only one aspect. Together with various other mechanisms of gene regulation the ones highlighted here enable the organism to react to a change in environmental conditions and to fine-tune gene expression to the actual requirements. From an evolutionary standpoint it is crucial to know what amount of this expression variation is encoded in the genetic material, the DNA. This is because, roughly speaking, only the genetic material is heritable and hence offers the potential to adapt over evolutionary time-scales. Furthermore, organisms are not only well-adapted in terms of the structure of their constituting molecules, but moreover also in the quantity of the latter. As there are so many ways to regulate genes with a large potential for buffering, it raises the question as to whether traces of adaptation in quantity can be detected in the genetic material at all. This, however, would be uttermost important in order to evaluate the evolutionary dynamics of change, with change not being restricted to the quality of molecules, one of the main goals of molecular evolutionary biology (population genetics). To do this, experimental and statistical/mathematical methods must be further improved in their accuracy and sensitivity.

Finally, I wish to point out that many of the insights gained in this dissertation are extendable to much larger groups of organisms than fruit flies. *Drosophila melanogaster* was the model organism of choice during this thesis since it has been introduced into genetical analysis some 100 years ago and is therefore very well known (MORGAN 1910). It was continuously developed further, also into a molecular genetic tool with the opportunity for genetic transformation (RUBIN and SPRADLING 1982; SPRADLING and RUBIN 1982). This was extensively used during the course of this thesis. The overall topic of gene expression regulation is very general so that the findings should in principal be applicable to at least eukaryotic organisms. X chromosome inactivation due to sexual antagonism is a possibility open to all organisms with a chromosomal mode of sex determination. In the second of such determination systems (the ZW system) it would be predicted that it is the female sex (which is heterogametic) where the corresponding Z chromosome could be inactivated. The initiation of transcription with its involvement of *cis* and *trans* factors is even a feature present in all organisms. The same is true of codon usage bias, which appears as a ubiquitous property of all genomes. Thus, a certain generality of the results presented here is warranted.

At last, to resume the thoughts at the beginning of this discussion, I want to emphasize that, although not particularly placed in evolutionary developmental genectics, the work presented here and the basic method used throughout this thesis, germline transformation of *Drosophila* flies, is also applicable in the field of "evo-devo" where old problems at least seem now accessible to investigation, *e.g.* the significance of *cis*-regulation in phenotypic evolution (in both inter- and intraspecific comparisons), by focusing evolutionary analysis on developmentally important genes. With more and more whole genome sequences available there is now the complete foundation given to tackle the conundrum of the relationship of form and function together with their connection to the genetic inventory of organisms:

> *"Auf diesem Gebiet liegt bereits eine Reihe höchst interessanter und viel versprechender Arbeiten vor, [...] und es spricht vieles dafür, dass die lebendige Gestalt einst über diesen entwicklungsbiologischen Weg ursächlich verstanden werden kann. [...] Nicht zuletzt sind es ja diese komplexen Formen der Gestalt und ihrer Metamorphosen, die nicht nur zur Erkenntnis der Evolution geführt haben, sondern uns auch das Grundproblem des Lebendigen vor Augen führen: auf welchen Prinzipien nämlich seine* Information *beruht. Somit ist die Gestaltforschung gleichsam der Anfang und das Ziel aller biologischen Erkenntnis."*
>
> ROBERT KASPAR, in: BECKER *et al.* 1994, vol. 4, p. 54

# Summary

THE results presented in this dissertation contribute to our understanding of gene expression regulation from an evolutionary point of view. Using a well-established model organism, the fruit fly *Drosophila melanogaster*, not only as an observational, but also as a manipulative genetic tool, I investigate three separate aspects of the process by which the information that is stored in the DNA of organisms is "unleashed" or transformed into biological meaning, which ultimately is form and function.

In CHAPTER 1, I demonstrate that X chromosome inactivation (and hence gene regulation on a chromosomal scale) takes place in the male germline of *D. melanogaster*. In contrast to X inactivation in female mammals, which occurs in somatic cells as a mechanism of dosage compensation, this type of inactivation is restricted to spermatogenesis and assumed to have been established during genome evolution as a way to avoid deleterious effects associated with sexual antagonism. By *P*-element mediated germline transformation, nearly 50 independent insertions of a testis-specific reporter gene construct were obtained and their respective reporter gene activities were assayed by measuring enzymatic activity and by qRT-PCR. Autosomal insertions of this construct showed the expected high levels of male- and testis-specific expression. In contrast, insertions on the X chromosome showed little or no transgene expression. Since the X-chromosomal insertions covered the euchromatic portions of the chromosome (as determined by inverse PCR), an insertional bias for the lack of expression on the X could be excluded. The effect appears to be a global property of the X chromosome. Only the testis-specificity of the transgenic construct is required for this effect to appear, which supports a selective hypothesis for X inactivation and may explain several observations regarding the distribution of male- and testis-expressed genes in the *Drosophila* genome.

In CHAPTER 2, I examine putative *cis*-regulatory sequences and their ability to drive allele-specific gene expression. After microarray studies revealed extensive variability in the primary trait of gene expression among diverse taxa, a current question evolutionary biologists have to face is what the underlying genetic source for this variability is. Apart from epigenetic mechanisms, there is a dispute as to whether regulatory sequences nearby the expressed gene (*cis* factors) and factors encoded elsewhere in the genome (*trans* factors) contribute in a qualitatively and quantitatively different way to gene expression variation. To investigate this, I selected a gene from *D. melanogaster* that was previously shown to exhibit consistent expression differences between African and non-African ("cosmopolitan") strains and cloned the respective upstream flanking regions into a reporter gene construct to compare directly their effects on gene expression (after successfully integrating them into the fruit fly genome). The observed effect was small, but significant, and appeared only in transgenic flies in the presence of an X chromosome from the original African fly strain. These results suggest that, in addition to upstream *cis*-regulatory elements, *trans*-acting factors (especially on the X chromosome) contribute to the observed expression difference between strains.

Finally, in CHAPTER 3 I investigate the phenomenon of codon usage bias through its relationship to gene expression. Due to the redundancy of the genetic code, many of the proteinogenic amino acids are encoded by more than one codon. Thus it is possible to change synonymous codons in the coding sequence of a gene without altering the amino acid sequence of the encoded polypeptide. Whether or not this has any consequence for the amount of protein produced (translational efficiency) is the topic of this chapter. I directly compared the enzymatic activity imparted by two alleles of the *D. melanogaster* alcohol dehydrogenase gene (*Adh*) that differed in seven leucine codons. There was almost no difference in the ADH enzymatic activity imparted by the two alleles, even though one allele consisted of entirely optimal leucine codons and the other contained seven suboptimal leucine codons. Since the latter allele was the wild-type form of *Adh*, these results suggest that the *Adh* gene is already sufficiently optimized in its leucine codon composition (and perhaps also in its general codon composition). Attempts to increase the number of optimal leucine codons may even have a negative effect in terms of enzyme production, possibly due to a saturation of the tRNA pool and/or the consequences of altered mRNA secondary structures.

# Literature cited

ADAMS, M.D., S.E. CELNIKER, R.A. HOLT, C.A. EVANS, J.D. GOCAYNE, *et al.*, 2000 The genome sequence of *Drosophila melanogaster*. Science **287:** 2185-2195.

AKASHI, H., 1994 Synonymous codon usage in *Drosophila melanogaster*: natural selection and translational accuracy. Genetics **136:** 927-935.

AKASHI, H., 1995 Inferring weak selection from patterns of polymorphism and divergence at "silent" sites in *Drosophila* DNA. Genetics **139:** 1067-1076.

AKASHI, H., 2001 Gene expression and molecular evolution. Curr. Opin. Genet. Dev. **11:** 660-666.

ANDOLFATTO, P., 2005 Adaptive evolution of non-coding DNA in *Drosophila*. Nature **437:** 1149-1152.

APARICIO, O., J.V. GEISBERG, and K. STRUHL, 2004 Chromatin immunoprecipitation for determining the association of proteins with specific genomic sequences *in vivo*. Curr. Protoc. Cell Biol. **Chapter 17:** Unit 17.7.

BACHTROG, D., 2005 Sex chromosome evolution: Molecular aspects of Y-chromosome degeneration in *Drosophila*. Genome Res. **15:** 1393-1401.

BAINES, J.F., J. PARSCH, and W. STEPHAN, 2004 Pleiotropic effect of disrupting a conserved sequence involved in a long-range compensatory interaction in the *Drosophila Adh* gene. Genetics **166:** 237-242.

BECKER, U., S. GANTER, and C. JUST (eds.), 1994 Herder-Lexikon der Biologie. Heidelberg; Berlin; Oxford: Spektrum, Akademischer Verlag.

BEGUN, D.J., H.A. LINDFORS, A.D. KERN, and C.D. JONES, 2007 Evidence for *de novo* evolution of testis-expressed genes in the *Drosophila yakuba*/*Drosophila erecta* clade. Genetics **176:** 1131-1137.

BELLEN, H.J., R.W. LEVIS, G. LIAO, Y. HE, J.W. CARLSON, *et al.*, 2004 The BDGP gene disruption project: Single transposon insertions associated with 40% of *Drosophila* genes. Genetics **167:** 761-781.

BENYAJATI, C., A.R. PLACE, N. WANG, E. PENTZ, and W. SOFER, 1982 Deletions at intervening sequence splice sites in the alcohol dehydrogenase gene of *Drosophila*. Nucleic Acids Res. **10:** 7261-7272.

BETANCOURT, A.J., and D.C. PRESGRAVES, 2002 Linkage limits the power of natural selection in *Drosophila*. Proc. Natl. Acad. Sci. U.S.A. **99:** 13616-13620.

BETRÁN, E., and M. LONG, 2003 *Dntf-2r*, a young *Drosophila* retroposed gene with specific male expression under positive Darwinian selection. Genetics **164:** 977-988.

BETRÁN, E., K. THORNTON, and M. LONG, 2002 Retroposed new genes out of the X in *Drosophila*. Genome Res. **12:** 1854-1859.

BETRÁN, E., Y. BAI, and M. MOTIWALE, 2006 Fast protein evolution and germ line expression of a *Drosophila* parental gene and its young retroposed paralog. Mol. Biol. Evol. **23:** 2191-2202.

BRITTEN, R.J., and E.H. DAVIDSON, 1969 Gene regulation for higher cells: a theory. Science **165:** 349-357.

BUCHAN, J.R., and I. STANSFIELD, 2007 Halting a cellular production line: responses to ribosomal pausing during translation. Biol. Cell **99:** 475-487.

BULMER, M., 1987 Coevolution of codon usage and transfer RNA abundance. Nature **325:** 728-730.

BULMER, M., 1991 The selection-mutation-drift theory of synonymous codon usage. Genetics **129:** 897-907.

CARLINI, D.B., 2004 Experimental reduction of codon bias in the *Drosophila alcohol dehydrogenase* gene results in decreased ethanol tolerance of adult flies. J. Evol. Biol. **17:** 779-785.

CARLINI, D.B., and W. STEPHAN, 2003 *In vivo* introduction of unpreferred synonymous codons into the *Drosophila Adh* gene results in reduced levels of ADH protein. Genetics **163:** 239-243.

CARLINI, D.B., Y. CHEN, and W. STEPHAN, 2001 The relationship between third-codon position nucleotide content, codon bias, mRNA secondary structure and gene expression in the drosophilid alcohol dehydrogenase genes *Adh* and *Adhr*. Genetics **159:** 623-633.

CARROLL, S.B., 2005 Evolution at two levels: on genes and form. PLoS Biol. **3:** e245.

CARTEGNI, L., J. WANG, Z. ZHU, M.Q. ZHANG, and A.R. KRAINER, 2003 ESEfinder: a web resource to identify exonic splicing enhancers. Nucleic Acids Res. **31:** 3568-3571.

CAVALIERI, D., J.P. TOWNSEND, and D.L. HARTL, 2000 Manifold anomalies in gene expression in a vineyard isolate of *Saccharomyces cerevisiae* as revealed by microarray analysis. Proc. Natl. Acad. Sci. U.S.A. **97:** 12369-12374.

CHAMARY, J.V., and L.D. HURST, 2005a Evidence for selection on synonymous mutations affecting stability of mRNA secondary structure in mammals. Genome Biol. **6:** R75.

CHAMARY, J.V., and L.D. HURST, 2005b Biased codon usage near intron-exon junctions: selection on splicing enhancers, splice-site recognition or something else? Trends Genet. **21:** 256-259.

CHARLESWORTH, B., 1996 The evolution of chromosomal sex determination and dosage compensation. Curr. Biol. **6:** 149-162.

CHARLESWORTH, B., J.A COYNE, and N.H. BARTON, 1987 The relative rates of evolution of sex chromosomes and autosomes. Am Nat **130:** 113-146.

CHEN, K., and N. RAJEWSKY, 2007 The evolution of gene regulation by transcription factors and microRNAs. Nat. Rev. Genet. **8:** 93-103.

CHEN, S.T., H.C. CHENG, D.A. BARBASH, and H.P. YANG, 2007 Evolution of *hydra*, a recently evolved testis-expressed gene with nine alternative first exons in *Drosophila melanogaster*. PLoS Genet. **3:** e107.

CHEN, Y., D.B. CARLINI, J.F. BAINES, J. PARSCH, J.M. BRAVERMAN, S. TANDA, and W. STEPHAN, 1999 RNA secondary structure and compenatory evolution. Genes Genet. Syst. **74:** 271-286.

CHOUDHARY, M., and C.C. LAURIE, 1991 Use of *in vitro* mutagenesis to analyze the molecular basis of the difference in *Adh* expression associated with the allozyme polymorphism in *Drosophila melanogaster*. Genetics **129:** 481-488.

COMERON, J.M., and M. KREITMAN, 2002 Population, evolutionary, and genomic consequences of interference selection. Genetics **161:** 389-410.

CONNALLON, T., and L.L. KNOWLES, 2005 Intergenomic conflict revealed by patterns of sex-biased gene expression. Trends Genet. **21:** 495-499.

CREMER, T., M. CREMER, S. DIETZEL, S. MÜLLER, I. SOLOVEI, and S. FAKAN, 2006 Chromosome territories – a functional nuclear landscape. Curr. Opin. Cell Biol. **18:** 307-316.

DABORN, P.J., J.L. YEN, M.R. BOGWITZ, G. LE GOFF, E. FEIL, S. JEFFERS, N. TIJET, *et al.*, 2002 A single p450 allele associated with insecticide resistance in *Drosophila*. Science **297:** 2253-2256.

DARWIN, C., 1859 The origin of species by means of natural selection, or the preservation of favoured races in the struggle for life. Murray, London. In: DARWIN, C., 1985 The origin of species, Penguin classics, London.

DORUS, S., S.A. BUSBY, U. GERIKE, J. SHABANOWITZ, D.F. HUNT, and T.L. KARR, 2006 Genomic and functional evolution of the *Drosophila melanogaster* sperm proteome. Nat. Genet. **38:** 1440-1445.

*DROSOPHILA* 12 GENOMES CONSORTIUM, 2007 Evolution of genes and genomes on the *Drosophila* phylogeny. Nature **450:** 203-218.

DURET, L., 2002 Evolution of synonymous codon usage in metazoans. Curr. Opin. Genet. Dev. **12:** 640-649.

DURET, L., and D. MOUCHIROUD, 1999 Expression pattern and, surprisingly, gene length shape codon usage in *Caenorhabditis*, *Drosophila*, and *Arabidopsis*. Proc. Natl. Acad. Sci. U.S.A. **96:** 4482-4487.

ECK, S., and W. STEPHAN, 2008 Determining the relationship of gene expression and global mRNA stability in *Drosophila melanogaster* and *Escherichia coli* using linear models. Gene **424:** 102-107.

ENARD, W., P. KHAITOVICH, J. KLOSE, S. ZÖLLNER, F. HEISSIG, *et al.*, 2002 Intra- and interspecific variation in primate gene expression patterns. Science **296:** 340-343.

ENGELS, W.R., 1992 The origin of *P* elements in *Drosophila melanogaster*. Bioessays **14:** 681-686.

FAIRBROTHER, W.G., R.F. YEH, P.A. SHARP, and C.B. BURGE, 2002 Predictive identification of exonic splicing enhancers in human genes. Science **297:** 1007-1013.

FEDOROVA, E., and D. ZINK, 2008 Nuclear architecture and gene regulation. Biochim. Biophys. Acta **1783:** 2174-2184.

FESCHOTTE, C., 2008 Transposable elements and the evolution of regulatory networks. Nat. Rev. Genet. **9:** 397-405.

FESCHOTTE, C., and E.J. PRITHAM, 2007 DNA transposons and the evolution of eukaryotic genomes. Annu. Rev. Genet. **41:** 331-368.

FILIPOWISZ, W., S.N. BHATTACHARYYA, and N. SONNENBERG, 2008 Mechanisms of post-transcriptional regulation by microRNAs: are the answers in sight? Nat. Rev. Genet. **9:** 102-114.

FONG, Y., L. BENDER, W. WANG, and S. STROME, 2002 Regulation of the different chromatin states of autosomes and X chromosomes in the germ line of *C. elegans*. Science **296:** 2235-2238.

GANGULY, R., K.D. SWANSON, K. RAY, and R. KRISHNAN, 1992 A *Bam*HI repeat element is predominantly associated with the degenerating neo-Y chromosome of *Drosophila miranda* but absent in the *Drosophila melanogaster* genome. Proc. Natl. Acad. Sci. U.S.A. **89:** 1340-1344.

GOMPEL, N., B. PRUD'HOMME, P.J. WITTKOPP, V.A. KASSNER, and S.B. CARROLL, 2005 Chance caught on the wing: *cis*-regulatory evolution and the origin of pigment patterns in *Drosophila*. Nature **433:** 481-487.

GRUBER, A.R., R. LORENZ, S.H. BERNHART, R. NEUBÖCK, and I.L. HOFACKER, 2008 The Vienna RNA websuite. Nucleic Acids Res. **36:** W70-74.

GU, W., P. SZAUTER, and J.C. LUCCHESI, 1998 Targeting of MOF, a putative histone acetyl transferase, to the X chromosome of *Drosophila melanogaster*. Dev. Genet. **22:** 56-64.

GUPTA, V., M. PARISI, D. STURGILL, R. NUTTALL, M. DOCTOLERO, *et al.*, 2006 Global analysis of X-chromosome dosage compensation. J. Biol. **5:** 3.

HAHN, M.W., 2007 Detecting natural selection on *cis*-regulatory DNA. Genetica **129:** 7-18.

HAMBLIN, M.T., A. DI RIENZO, 2000 Detection of the signature of natural selection in humans: evidence from the *Duffy* blood group locus. Am. J. Hum. Genet. **66:** 1669-1679.

HAMBUCH, T.M., and J. PARSCH, 2005 Patterns of synonymous codon usage in *Drosophila melanogaster* genes with sex-biased expression. Genetics **170:** 1691-1700.

HARE, E.E., B.K. PETERSON, V.N. IYER, R. MEIER, and M.B. EISEN, 2008 Sepsid *even-skipped* enhancers are functionally conserved in *Drosophila* despite lack of sequence conservation. PLoS Genet. **4:** e1000106.

HOEDE, C., E. DENAMUR, and O. TENAILLON, 2006 Selection acts on DNA secondary structures to decrease transcriptional mutagenesis. PLoS Genet. **2:** 1697-1701.

HOEKSTRA, H.E., and J.A. COYNE, 2007 The locus of evolution: evo devo and the genetics of adaptation. Evolution **61:** 995-1016.

HOYLE, H.D., J.A. HUTCHENS, F.R. TURNER, and E.C. RAFF, 1995 Regulation of beta-tubulin function and expression in *Drosophila* spermatogenesis. Dev. Genet. **16:** 148-170.

HUDSON, R.R., M. KREITMAN, and M. AGUADÉ, 1987 A test of neutral molecular evolution based on nucleotide data. Genetics **116:** 153-159.

HUTTER, S., S.S. SAMINADIN-PETER, W. STEPHAN, and J. PARSCH, 2008 Gene expression variation in African and European populations of *Drosophila melanogaster*. Genome Biol. **9:** R12.

IKEMURA, T., 1981 Correlation between the abundance of *Escherichia coli* transfer RNAs and the occurrence of the respective codons in its protein genes: a proposal for a synonymous codon choice that is optimal for the *E. coli* translational system. J. Mol. Biol. **151:** 389-409.

JACOB, F., and J. MONOD, 1961 Genetic regulatory mechanisms in the synthesis of proteins. J. Mol. Biol. **3:** 318-356.

JIN, W., R.M. RILEY, R.D. WOLFINGER, K.P. WHITE, G. PASSADOR-GURGEL, and G. GIBSON, 2001 The contributions of sex, genotype and age to transcriptional variance in *Drosophila melanogaster*. Nat. Genet. **29:** 389-395.

KALAMEGHAM R., D. STURGILL, E. SIEGFRIED, and B. OLIVER, 2006 *Drosophila mojoless*, a retroposed gsk-3, has functionally diverged to acquire an essential role in male fertility. Mol. Biol. Evol. **24:** 732-742.

KELLY, W.G., C.E. SCHANER, A.F. DERNBURG, M.H. LEE, S.K. KIM, *et al.*, 2002 X-chromosome silencing in the germline of *C. elegans*. Development **129:** 479-492.

KHAITOVICH, P., W. ENARD, M. LACHMANN, and S. PÄÄBO, 2006 Evolution of primate gene expression. Nat. Rev. Genet. **7:** 693-702.

KIDWELL, M.G., J.F. KIDWELL, and J.A. SVED, 1977 Hybrid dysgenesis in *Drosophila melanogaster*: A sydrome of aberrant traits including mutation, sterility and male recombination. Genetics **86:** 813-833.

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, Cambridge.

KING, M.C., and A.C. WILSON, 1975 Evolution at two levels in humans and chimpanzees. Science **188:** 107-116.

KIRBY, D.A., S.V. MUSE, and W. STEPHAN, 1995 Maintenance of pre-mRNA secondary structure by epistatic selection. Proc. Natl. Acad. Sci. U.S.A. **92:** 9047-9051.

KREITMAN, M., 1983 Nucleotide polymorphism at the alcohol dehydrogenase locus of *Drosophila melanogaster*. Nature **304:** 412-417.

LANCTÔT, C., T. CHEUTIN, M. CREMER, G. CAVALLI, and T. CREMER, 2007 Dynamic genome architecture in the nuclear space: regulation of gene expression in three dimensions. Nat. Rev. Genet. **8:** 104-115.

LANDRY, C.R., P.J. WITTKOPP, C.H. TAUBES, J.M. RANZ, A.G. CLARK, and D.L. HARTL, 2005 Compensatory *cis-trans* evolution and the dysregulation of gene expression in interspecific hybrids of *Drosophila*. Genetics **171:** 1813-1822.

LEMOS, B., C.D. MEIKLEJOHN, M. CACERES, and D.L. HARTL , 2005 Rates of divergence in gene expression profiles of primates, mice, and flies: stabilizing selection and variability among functional categories. Evolution **59:** 136-137.

LEMOS, B., L.O. ARARIPE, P. FONTANILLAS, and D.L. HARTL, 2008 Dominance and the evolutionary accumulation of *cis*- and *trans*-effects on gene expression. Proc. Natl. Acad. Sci. U.S.A. **105:** 14471-14476.

LEONHARDT, H., and M.C. CARDOSO, 2000 DNA methylation, nuclear structure, gene expression and cancer. J. Cell. Biochem. Suppl. **35:** 78-83.

LEVINE, M.T., C.D. JONES, A.D. KERN, H.A. LINDFORS, and D.J. BEGUN, 2006 Novel genes derived from noncoding DNA in *Drosophila melanogaster* are frequently X-linked and exhibit testis-biased expression. Proc. Natl. Acad. Sci. U.S.A. **103:** 9935-9939.

LEWIS, E.B., 1992 The 1991 Albert Lasker Medical Awards. Clusters of master control genes regulate the development of higher organisms. JAMA **267:** 1524-1531.

LIFSCHYTZ, E., and D.L. LINDSLEY, 1972 The role of X-chromosome inactivation during spermatogenesis. Proc. Natl. Acad. Sci. U.S.A. **69:** 182-186.

LO, H.S., Z. WANG, Y. HU, H.H. YANG, S. GERE, K.H. BUETOW, and M.P. LEE, 2003 Allelic variation in gene expression is common in the human genome. Genome Res. **13:** 1855-1862.

LOWRY, O.N., N.J. ROSENBROUGH, L.A. FARR, and R.J. RANDALL, 1951 Protein measurement with the Folin phenol reagent. J. Biol. Chem. **193:** 265-275.

LYON, M.F., 1961 Gene interaction in the X-chromosome of the mouse (*Mus musculus* L.). Nature **190:** 372-373.

MARONI, G., 1978 Genetic control of alcohol dehydrogenase levels in *Drosophila*. Biochem. Genet. **16:** 509-523.

MAYNARD SMITH, J., and E. SZATHMARY, 1995 The major transitions in evolution. Oxford, Oxford University Press, 360 p.

MAYNARD SMITH, J., and J. HAIGH, 1974 The hitch-hiking effect of a favourable gene. Genet. Res. **23:** 23-35.

MCDONALD, J.H., and M. KREITMAN, 1991 Adaptive protein evolution at the *Adh* locus in *Drosophila*. Nature **351:** 652-654.

MCKEE, B.D., and M.A. HANDEL 1993, Sex chromosomes, recombination, and chromatin conformation. Chromosoma **102:** 71-80.

MEIKLEJOHN, C.D., J. PARSCH, J.M. RANZ, and D.L. HARTL, 2003 Rapid evolution of male-biased gene expression in *Drosophila*. Proc. Natl. Acad. Sci. U.S.A. **100:** 9894-9899.

MICHIELS, F., A. GASCH, B. KALTSCHMIDT, and R. RENKAWITZ-POHL, 1989 A 14 bp promoter element directs the testis specificity of the *Drosophila* beta 2 tubulin gene. EMBO J. **8:** 1559-1565.

MORGAN, T.H., 1910 Sex limited inheritance in *Drosophila*. Science **32:** 120-122.

MORIYAMA, E.N., and J.R. POWELL, 1997 Codon usage bias and tRNA abundance in *Drosophila*. J. Mol. Evol. **45:** 514-523.

NIELSEN, R., 2005 Molecular signatures of natural selection. Annu. Rev. Genet. **39:** 197-218.

NIELSEN, R., I. HELLMANN, M. HUBISZ, C. BUSTAMANTE, and A.G. CLARK, 2007 Recent and ongoing selection in the human genome. Nat. Rev. Genet. **8:** 857-868.

NURMINSKY, D.I., M.V. NURMINSKAYA, D. DE AGUIAR, and D.L. HARTL, 1998 Selective sweep of a newly evolved sperm-specific gene in *Drosophila*. Nature **396:** 572-575.

NÜSSLEIN-VOLHARD, C., and E. WIESCHAUS, 1980 Mutations affecting segment number and polarity in *Drosophila*. Nature **287:** 795-801.

OLEKSIAK, M.F., G.A. CHURCHILL, and D.L. CRAWFORD, 2002 Variation in gene expression within and among natural populations. Nat. Genet. **32:** 261-266.

OLIVER, B., and M. PARISI, 2004 Battle of the Xs. Bioessays **26:** 543-548.

PARISI, M., R. NUTTALL, D. NAIMAN, G. BOUFFARD, J. MALLEY, J. ANDREWS, S. EASTMAN, and B. OLIVER, 2003 Paucity of genes on the *Drosophila* X chromosome with male-biased expression. Science **299:** 697-700.

PARMLEY, J.L., and L.D. HURST, 2007 Exonic splicing regulatory elements skew synonymous codon usage near intron-exon boundaries in mammals. Mol. Biol. Evol. **24:** 1600-1603.

PARSCH, J., C.D. MEIKLEJOHN, E. HAUSCHTECK-JUNGEN, P. HUNZIKER, 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.

PARSCH, J., 2004 Functional analysis of *Drosophila melanogaster* gene regulatory sequences by transgene coplacement. Genetics **168:** 559-561.

PARSCH, J., J.M. BRAVERMAN, and W. STEPHAN, 2000 Comparative sequence analysis and patterns of covariation in RNA secondary structures. Genetics **154:** 909-921.

PARSCH, J., S. TANDA, and W. STEPHAN, 1997 Site-directed mutations reveal long-range compensatory interactions in the *Adh* gene of *Drosophila melanogaster*. Proc. Natl. Acad. Sci. U.S.A. **94:** 928-933.

PARSCH, J., W. STEPHAN, and S. TANDA, 1999 A highly conserved sequence in the 3'-untranslated region of the *Drosophila Adh* gene plays a functional role in *Adh* expression. Genetics **151:** 667-674.

PATTON, J.S., X.V. GOMES, and P.K. GEYER, 1992 Position-independent germline transformation in *Drosophila* using a cuticle pigmentation gene as a selectable marker. Nucleic Acids Res. **20:** 5859-5860.

PERTEA, M., S.M. MOUNT, and S.L. SALZBERG, 2007 A computational survey of candidate exonic splicing enhancer motifs in the model plant *Arabidopsis thaliana*. BMC Bioinformatics **8:** 159.

PIRROTTA, V., 1988 Vectors for *P*-mediated transformation in *Drosophila*. Biotechnology **10:** 437-456.

PRÖSCHEL, M., Z. ZHANG, and J. PARSCH, 2006 Widespread adaptive evolution of *Drosophila* genes with sex-biased expression. Genetics **174:** 893-900.

PRUD'HOMME, B., N. GOMPEL, A. ROKAS, V.A. KASSNER, T.M. WILLIAMS, S.D. YEH, J.R. TRUE, and S.B. CARROLL, 2006 Repeated morphological evolution through *cis*-regulatory changes in a pleiotropic gene. Nature **440:** 1050-1053.

RANZ, J.M., C.I. CASTILLO-DAVIS, C.D. MEIKLEJOHN, and D.L. HARTL, 2003 Sex-dependent gene expression and evolution of the *Drosophila* transcriptome. Science **300:** 1742-1745.

RASTELLI, L., and M.I. KURODA, 1998 An analysis of maleless and histone H4 acetylation in *Drosophila melanogaster* spermatogenesis. Mech. Dev. **71:** 107-117.

RICE, W.R., 1984 Sex chromosomes and the evolution of sexual dimorphism. Evolution **38:** 735-742.

RICE, W.R., 1996 Evolution of the Y sex chromosome in animals. Bioscience **46:** 331-343.

RICHLER, C., H. SOREQ, and J. WAHRMAN, 1992 X inactivation in mammalian testis is correlated with inactive X-specific transcription. Nat. Genet. **2:** 192-195.

RIFKIN, S.A., D. HOULE, J. KIM, and K.P. WHITE, 2005 A mutation accumulation assay reveals a broad capacity for rapid evolution of gene expression. Nature **438:** 220-223.

ROBERTSON, H.M., C.R. PRESTON, R.W. PHILLIS, D.M. JOHNSON-SCHLITZ, W.K. BENZ, and W.R. ENGELS, 1988 A stable genomic source of *P* element transposase in *Drosophila melanogaster*. Genetics **118:** 461-470.

ROGERS, D.W., M. CARR, and A. POMIANKOWSKI, 2003 Male genes: X-pelled or X-cluded? Bioessays **25:** 739-741.

RUBIN, G.M., and A.C. SPRADLING, 1982 Genetic transformation of *Drosophila* with transposable element vectors. Science **218:** 348-353.

SAMBROOK, J., E.F. FRITSCH, and T. MANIATIS, 1989 Molecular cloning: A laboratory manual. Ed. 2 Cold Spring Harbor (NY): Cold Spring Harbor Laboratory Press. 1659 p.

SCHERMELLEH, L., F. SPADA, and H. LEONHARDT, 2008 Visualization and measurement of DNA methyltransferase activity in living cells. Curr. Protoc. Cell Biol. **Chapter 22:** Unit 22.12.

SCHLÖTTERER, C., 2003 Where do male genes live? Science **299:** 670-671.

SHAPIRO, M.D., M.E. MARKS, C.L. PEICHEL, B.K. BLACKMAN, K.S. NERENG, B. JÓNSSON, D. SCHLUTER, and D.M. KINGSLEY, 2004 Genetic and developmental basis of evolutionary pelvic reduction in threespine sticklebacks. Nature **428:** 717-723.

SHARP, P.M., and W. H. LI, 1986 An evolutionary perspective on synonymous codon usage in unicellular organisms. J. Mol. Evol. **24:** 28-38.

SHARP, P.M., T.M. TUOHY, and K.R. MOSURSKI, 1986 Codon usage in yeast: cluster analysis clearly differentiates highly and lowly expressed genes. Nucleic Acids Res. **14:** 5125-5143.

SIEGAL, M.L., and D.L. HARTL, 1996 Transgene coplacement and high efficiency site-specific recombination with the Cre/*loxP* system in *Drosophila*. Genetics **144:** 715-726.

SIEGAL, M.L., and D.L. HARTL, 1998 An experimental test for lineage-specific position effects on alcohol dehydrogenase (*Adh*) genes in *Drosophila*. Proc. Natl. Acad. Sci. U.S.A. **95:** 15513-15518.

SMITH, E.R., A. PANNUTI, W. GU, A. STEURNAGEL, R.G. COOK, *et al.*, 2000 The *Drosophila* MSL complex acetylates histone H4 at lysine 16, a chromatin modification linked to dosage compensation. Mol. Cell. Biol. **20:** 312-318.

SPADA, F., U. ROTHBAUER, K. ZOLGHADR, L. SCHERMELLEH, and H. LEONHARDT, 2006 Regulation of DNA methyltransferase 1. Adv. Enzyme Regul. **46:** 224-234.

SPRADLING, A.C., and G.M. RUBIN, 1982 Transposition of cloned P elements into *Drosophila* germ line chromosomes. Science **218:** 341-347.

ST. JOHNSTON, D., and C. NÜSSLEIN-VOLHARD, 1992 The origin of pattern and polarity in the *Drosophila* embryo. Cell **68:** 201-219.

STEINEMANN, M., and S. STEINEMANN, 2000 Common mechanisms of Y chromosome evolution. Genetica **109:** 105-111.

STEINEMANN, S., and M. STEINEMANN, 2001 Biased distribution of repetitive elements: A landmark for neo-Y chromosome evolution in *Drosophila miranda*. Cytogenet. Cell Genet. **93:** 228-233.

STENØIEN, H.K., and W. STEPHAN, 2005 Global mRNA stability is not associated with levels of gene expression in *Drosophila melanogaster* but shows a negative correlation with codon bias. J. Mol. Evol. **61:** 306-314.

STERN, D.L., 2000 Evolutionary developmental biology and the problem of variation. Evolution **54:** 1079-1091.

SWANSON, W.J., A.G. CLARK, H.M. WALDRIP-DAIL, M.F. WOLFNER, and C.F. AQUADRO, 2001 Evolutionary EST analysis identifies rapidly evolving male reproductive proteins in *Drosophila*. Proc. Natl. Acad. Sci. U.S.A. **98:** 7375-7379.

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

TAKYAR, S., R.P. HICKERSON, and H.F. NOLLER, 2005 mRNA helicase activity of the ribosome. Cell **20:** 49-58.

TOWNSEND, J.P., D. CAVALIERI, and D.L. HARTL, 2003 Population genetic variation in genome-wide gene expression. Mol. Biol. Evol. **20:** 955-963.

VENTER, J.C., M.D. ADAMS, E.W. MYERS, P.W. LI, R.J. MURAL, *et al.*, 2001 The sequence of the human genome. Science **291:** 1304-1351.

VICARIO, S., E. MORIYAMA, and J. POWELL, 2007 Codon usage in twelve species of *Drosophila*. BMC Evol. Biol. **7:** 226.

VICOSO, B., and B. CHARLESWORTH, 2006 Evolution on the X chromosome: Unusual patterns and processes. Nat. Rev. Genet. **7:** 645-653.

WADA, A., and A. SUYAMA, 1986 Local stability of DNA and RNA secondary structure and its relation to biological functions. Prog. Biophys. Mol. Biol. **47:** 113-157.

WANG, H.Y., F. YONGGUI, M.S. MCPEEK, X. LU, S. NUZHDIN, A. XU, J. LU, M.L. WU, and C.I. WU, 2008 Complex genetic interactions underlying expression differences between *Drosophila* races: analysis of chromosome substitutions. Proc. Natl. Acad. Sci. U.S.A. **105:** 6362-6367.

WANG, Y., W. FISCHLE, W. CHEUNG, S. JACOBS, S. KHORASANIZADEH, and C.D. ALLIS, 2004 Beyond the double helix: writing and reading the histone code. Novartis Found. Symp. **259:** 3-17.

WANG, Z., and C.B. BURGE, 2008 Splicing regulation: from a parts list of regulatory elements to an integrated splicing code. RNA **14:** 802-814.

WARNECKE, T., and L.D. HURST, 2007 Evidence for a trade-off between translational efficiency and splicing regulation in determining synonymous codon usage in *Drosophila melanogaster*. Mol. Biol. Evol. **24:** 2755-2762.

WILLIE, E., and J. MAJEWSKI, 2004 Evidence for codon bias selection at the pre-mRNA level in eukaryotes. Trends Genet. **20:** 534-538.

WITTKOPP, P.J., 2006 Evolution of *cis*-regulatory sequence and function in Diptera. Heredity **97:** 139-147.

WITTKOPP, P.J., B.K. HAERUM, and A.G. CLARK, 2004 Evolutionary changes in *cis* and *trans* gene regulation. Nature **430:** 85-88.

WITTKOPP, P.J., B.K. HAERUM, and A.G. CLARK, 2008 Regulatory changes underlying expression differences within and between *Drosophila* species. Nat. Genet. **40:** 346-350.

WRAY, G.A., M.W. HAHN, E. ABOUHEIF, J.P. BALHOFF, M. PIZER, M.V. ROCKMAN, and L.A. ROMANO, 2003 The evolution of transcriptional regulation in eukaryotes. Mol. Biol. Evol. **20:** 1377-1419.

WU, C.I., and E.Y. XU, 2003 Sexual antagonism and X inactivation – The SAXI hypothesis. Trends Genet. **19:** 243-247.

YANICOSTAS, C., and J.A. LEPESANT, 1990 Transcriptional and translational *cis*-regulatory sequences of the spermatocyte-specific *Drosophila janusB* gene are located in the 3′ exonic region of the overlapping *janusA* gene. Mol. Gen. Genet. **224:** 450-458.

ZUKER, M., 2003 Mfold web server for nucleic acid folding and hybridization prediction. Nucleic Acids Res. **31:** 3406-3415.

# Curriculum Vitae

| | |
|---|---|
| **Name:** | Winfried Karl Hense |
| **Date and Place of Birth:** | 17th May 1975, Bobingen |
| **Nationality:** | German |
| **Marital Status:** | single |

**Education:**

| | |
|---|---|
| 2004 – 2008 | PhD student, University of Munich (LMU), Germany |
| 1998 – 2004 | Diploma in Biology, University of Munich (LMU), Germany |
| 1996 – 1998 | Studies of Physics, University of Augsburg, Germany |
| 1986 – 1995 | Abitur, Gymnasium Königsbrunn, Germany |

**Fellowships:**

| | |
|---|---|
| Jan – Jun 2006 | Marie Curie Fellowship, European Commission: Bioinformatics training at the University of Manchester, UK |

# Publications

**HENSE W.**, J.F. BAINES, and J. PARSCH, (2007) X chromosome inactivation during *Drosophila* spermatogenesis. PLoS Biol 5: e273.

**HENSE W.**, N. ANDERSON, S. HUTTER, W. STEPHAN, J. PARSCH, and D.B. CARLINI, Experimental increase of codon bias in the *Drosophila Adh* gene has no effect on ADH protein expression. Genetics (*in review*)

# Conference Contributions

**HENSE, W.**, and J. PARSCH (August 2005) Molecular evolution of the *Drosophila janus* and *ocnus* genes. Poster presentation at the 10th Congress of the *European Society of Evolutionary Biology* in Krakow, Poland

**HENSE, W.**, J.F. BAINES, and J. PARSCH (March 2007) An experimental test of the X-inactivation hypothesis. Poster presentation at the 48th *Annual Drosophila Research Conference* in Philadelphia, USA

**HENSE, W.**, J.F. BAINES, and J. PARSCH (August and September 2007) X chromosome inactivation during *Drosophila* spermatogenesis. Poster presentation at the 11th Congress of the *European Society of Evolutionary Biology* in Uppsala, Sweden, and oral presentation at the 20th *European Drosophila Research Conference* in Vienna, Austria

CARLINI, D.B., N. ANDERSON, **W. HENSE**, S. HUTTER, and J. PARSCH (June 2008) Experimental increase of codon bias in the *Drosophila Adh* gene results in no effect on ADH protein expression. Poster presentation at the Annual Meeting of the *Society for Molecular Biology and Evolution* in Barcelona, Spain

# Acknowledgments

First, I would like to thank Professor John Parsch for giving me the opportunity to conduct my PhD research in evolutionary genetics, especially in the field of *Drosophila* genetics; it was very inspiring to also work on whole and living organisms, and not only parts of them. His wealth of scientific ideas and his professionalism made this thesis possible, his calmness and patience were very helpful in the final phase of writing the dissertation. His generosity was not only noticeable on our annual excursion to the *Oktoberfest*, but also allowed me to travel to several conferences during the course of the last years; one of the benefits of working in scientific research.

I would also like to thank Professor Wolfgang Stephan for taking over the *Zweitgutachen*.

I owe many thanks to Dr. Casey Bergman for letting me come to Manchester and learn bioinformatics and its applications to evolutionary genetics. In this regard I am also very grateful to Sonia, Nora, Francesco, but also Sara, Milena, and José for having a relly great Mancunian time.

Hedwig Gebhart shared her rich lab experience with me and introduced me to practical work. Thank you also for constantly providing a fully equipped lab in spite of a lack of time.

Pleuni Pennings encouraged and supported me in the phase of writing the thesis: thanks a lot!

I am thankful to Nicolas Svetec for sharing his great knowledge regarding fruit flies and for providing some of his fly strains.

Finally I would like to say thanks to everyone in the Munich Evolutionary Biology lab. For having kind and pleasant chats, many thanks to Katrin Kümpfbeck, Yvonne Cämmerer, Anne Wilken (additionally for trying to win a fortune of money with me at a quiz show), Traudl Feldmaier-Fuchs (additionally for conversations on literature), Anica Vrljic, and Kawsar Bhuiyan; for generating a cheerful work climate, thanks to Zhi Zhang, Matthias Pröschel, John Baines, Sarah Peter, Xiao Liu, Sonja Grath, Lena Müller, and Miriam Linnenbrink; for a lot of support and help in the lab I am also grateful to Sergej Nowoshilow and Theofanis Karaletsos. Eventually I am tremendously grateful to Simone Lange, Anja

Hörger, Iris Fischer, Hilde Lainer, and Claus Kemkemer for raising the social and working atmosphere by many orders of magnitude and for providing a sanctuary down the floor where to feel comfortable (almost elysian).

Last, but not least, I must thank my parents for constant and patient support during the last years…