Evolution of orphan genes in *Drosophila*

I n a u g u r a l – D i s s e r t a t i o n

zur

Erlangung des Doktorgrades

der Mathematisch-Naturwissenschaftlichen Fakultät

der Universität zu Köln

vorgelegt von

Tomislav Domazet-Lošo

aus Split, Kroatien

(Köln, 2003)

Berichterstatter:

Prof. Dr. Diethard Tautz

Prof. Dr. Thomas Wiehe

Tag der mündlichen Prüfung: 03. Juli 2003

Acknowledgments

I am particularly grateful to Prof. Dr. Diethard Tautz for giving me opportunity to join his group, for his patience, guidance, criticism and support. I would like to show gratitude to Prof. Dr. Thomas Wiehe, Prof. Dr. Siegfrid Roth and Dr. Angelika Stollewerk as well for accepting a membership in my thesis committee.

I thank Dr. Wim Damen, Dr. Martin Gajewski and Dr. Angelika Stollewerk for helpful suggestions and discussions during my work. With great pleasure, I thank Dr. Heidi Fusswinkel and Eva Sigmund for their kindness and help with solving numerous important administrative issues.

Many information technology problems would be very hard to solve without generous help of Alexander Pozhitkov. Our endless scientific and philosophical discussions helped me to clarify many ideas in this work. I am grateful to Hilary Dove for her overall support, especially at the beginning of my staying in Cologne. Karl Schmid generously provided me with embryo library and I am grateful for his hospitality during my visit in Jena. I appreciate discussions with Arne Nolte and his helpful experimental and theoretical hints, above all concerning sequencing. Sebastian Steinfartz organized many things in the lab professionally, and his help during the very beginning of the project is appreciated. I thank Marco Mendez-Torres for his hospitality and help during my first months in Cologne. I thank Joel Savard who was providing me with the reliable experimental protocols. I thank Michael Schoppmeier and Nikola-Michael Prpić for many useful discussions and advices. Sonja Ihle helped me with statistical problems and Christian Voolstra gave me in situ hybridisations hints. I am grateful to Susanne Krächter for helping me with sequencing. Finally, I express gratitude to all members of the Prof. Tautz lab that are not mentioned explicitly for their overall help.

Financial support from DFG (Ta99-17) is gratefully acknowledged.

In conclusion, I am particularly grateful to my parents, sister, relatives and friends and especially to Mirjana Madunić for their encouragement and support.

Abbreviations

aa – amino acid
ANOVA – analysis of variance
BLAST – basic local alignment search tool
bp – base pair
cfu – colony forming unit
DEPC – diethyl pyrocarbonate
dN – non-synonymous substitution rate
DNA – deoxyribonucleic acid
dNTP – deoxyribonucleoside triphosphate
dS – synonymous substitution rate
ENC – effective number of codons
EST – expressed sequence tag
Fop – frequency of optimal codons
GO – gene ontology database
mel – *Drosophila melanogaster*
mRNA – messanger ribonucleic acid
NCBI – national centre for biotechnology information
ORF – open reading frame
P – probability
pfu – plaque forming unit
r – Pearson's correlation coefficient
RNA – ribonucleic acid
r_s – Spearman's rank correlation coefficient
SAGE – serial analysis of gene expression
yak – *Drosophila yakuba*

Zusammenfassung

Orphan-Gene sind proteincodierende Bereiche, die kein erkennbares Homolog in entfernt verwandten Arten haben. Ein wesentlicher Anteil der bisher sequenzierten Genome besteht aus solchen Orphan-Genen, deren evolutionäre und funktionelle Bedeutung bislang nicht bekannt ist. Eine Analyse des *Drosophila melanogaster* Proteoms zeigt, dass immerhin 26 - 29% aller Proteine keine statistisch signifikanten Übereinstimmungen mit nicht aus Insekten stammenden Sequenzen haben. Entsprechend haben weder das stetige Anwachsen der Menge verfügbarer Sequenzdaten noch die Reannotation bekannter Gene den Anteil der Orphan-Gene im *Drosophila* Genom wesentlich verändert. Es konnte gezeigt werden, dass Orphan-Gene in derzeitigen genetischen Analysen deutlich unterrepräsentiert sind.

Um die evolutionären Eigenschaften von Orphan-Genen in *Drosphila* zu analysieren wurden 774 cDNA Sequenzen aus zwei *D. yakuba*-Genbibliotheken (adult und embryo) mit ihren Orthologen aus *D. melanogaster* verglichen. Eine Analyse der Substitutionsraten ergab, dass Orphan-Gene im Mittel dreimal schneller evolvieren als Nicht-Orphan-Gene, wobei die Breite der Evolutionsraten-Verteilung sich für beide Klassen ähnelt. Einzelne Orphan-Gene zeigen sehr niedrige Substitutionsraten, wie sie sonst für besonders hochkonservierte Gene typisch sind. Ein allgemeines Modell für die Evolution von Orphan Genen wurde entwickelt, dass die grossen Substitutionsratenunterschiede durch Phasen schneller und langsamer Divergenz erklährt.

Neben der Tatsache, dass Orphan-Gene unter allen untersuchten Genen unterrepräsentiert sind gibt es Hinweise darauf, dass sie generell einen weniger offensichtlichen Phänotypen haben. Eine Hypothese besagt, dass funktionell wichtige Gene einen deutlichen Phänotypen und eine verlangsamte Evolutionsrate haben. Damit übereinstimmend waren unter den untersuchten cDNA's genetisch charakterisierte Gene häufig langsam evolvierend. Interessanterweise war solch ein Zusammenhang nicht für Orphan-Gene zu beobachten. Zusätzlich spielen Orphan-Gene überproportional häufig eine Rolle für Geruchssinn, Hormonhaushalt, Puppenanheftung, Eimembranstrucktur und Wahrnehmung. Es ist anzunehmen, dass all diese Funktionen eine Bedeutung für spezifische ökologische Anpassungen haben, die sich schnell verändern und einen schwer detektierbaren mutanten Phänotypen haben.

Ein Vergleich zwischen Entwicklungsstadien zeigt, dass in der cDNA Bibliothek von Adulten doppelt so viele Orphan-Gene gefunden wurden wie in der Embryobibliothek. Eine Analyses der Gene, die Stadienspezifisch exprimiert werden, ergibt ein ähnliches Verhältnis. Zusammen mit einer bei Embryotranskripten gefundenen verringerten Evolutionsrate deutet sich deshalb eine stärkere Einschränkung für die Verwendung von Orphan-Genen in Embryos an. Die Expression von Orphan-Genen ist bei Embryos oft räumlich begrenzt, was auf eine eher lokale als ubiquitäre Verwendung hinweist. Die generellen Charackteristika von Orphan-Genen in *Drosophila* legen nahe, dass diese bei der Evolution von adaptiven Merkmalen eine Rolle spielen. Langsam evolvierende Orphan-Gene könnten von besonderem Interesse für die Bestimmung von linienspezifischen Adaptationen sein.

1. Summary

Orphan genes are protein coding regions that have no recognizable homologue in distantly related species. A substantial fraction of coding regions in any genome sequenced so far consists of such orphan genes, but their evolutionary and functional significance is not understood. A re-analysis of the *Drosophila melanogaster* proteome is presented that shows that there are still between 26 - 29% of all proteins without a significant match with non-insect sequences. Therefore, neither the growth of the database nor the re-annotations have significantly changed the proportion of orphans in the *Drosophila* genome over time. In addition, it was shown that these orphans are significantly underrepresented in the current genetic analysis.

To analyse directly the evolutionary characteristics of orphan genes in *Drosophila*, 774 sequences were compared between cDNAs retrieved from two *D. yakuba* libraries (embryo and adult) and their corresponding *D. melanogaster* orthologues. Analysis of substitution rates shows that recovered orphans evolve on average more than three times faster than non-orphan genes, although the width of the evolutionary rate distribution is similar for both classes. In particular, some orphan genes show very low substitution rates, which are comparable to otherwise highly, conserved genes. A general model for orphan gene evolution is proposed that takes these large rate differences into account and suggests that they are caused by episodic phases of fast and slow divergence.

Besides the result, that orphans are under-represented among genetically studied genes, additional findings suggest that orphan genes have less obvious phenotypes. For example, in the complete sample of the recovered cDNAs higher frequency of genetically studied genes was found among slow evolving genes, what supports the proposed hypothesis that functionally more important genes with obvious phenotypes have lower evolutionary rates. Interestingly, such relationship is lacking if only orphans are analysed. Additionally, orphans are over-represented among genes related to olfaction, hormonal activity, puparial adhesion, egg membrane structure and perception and response to abiotic stimulus. It is reasonable to expect for all of these functions to be involved in specific ecological adaptations that change easily over time, and accordingly to have mutant phenotypes which are difficult to detect.

Finally, comparison between stages shows that the cDNA library from adults yields twice as many orphan genes than the one from embryos. An analysis of only genes having stage specific expression reveals a similar figure and together with lower evolutionary rate of embryo transcripts suggests a higher constraint on use of orphan genes in embryos. Furthermore, expression of embryo orphans is more often spatially restricted compared to a random sample of genes what shows that they act in more localised rather then ubiquitous manner. Taken together, the general characteristics of orphan genes in *Drosophila* suggest that they may be involved in the evolution of adaptive traits and that slow evolving orphan genes may be particularly interesting candidate genes for identifying lineage specific adaptations.

2. Introduction

2.1 Genome sequencing projects and orphan genes

2.1.1 What are orphan genes?

A gene that has amino acid sequence similarity to other genes that belong to relatively narrow monophyletic lineages is referred to as an orphan gene. The phylogenetic group used to define orphan genes in a particular study is necessary arbitrary, often influenced by availability of the sequence data. In the most rigorous use, the term designates strictly genes specific just for one species, moreover sometimes only one strain (e.g. bacterial species), but more frequently group of closely related species is compared to the rest of the living organisms. It is reasonable to expect that genes specific to relatively closely related organisms exist. However, surprisingly, they came into focus only after first complete genomes were sequenced. Most of the genes, studied before the genome era, had sequence counterparts in distantly related organisms, scattered among more general taxonomic divisions like phyla and kingdoms. Sequence similarity between these conserved genes often implied their similar functional roles. This was the reason that genome content was envisaged in a considerably biased way. The yeast genome, as the first completely sequenced eukaryotic genome (Goffeau et al., 1996), illustrates this preconception.

2.1.2 Orphan genes and the yeast genome project

Already after the completion of the first chromosome (chromosome III) of *Saccharomyces cerevisiae* (Oliver et al., 1992) it was obvious that most of the predicted protein coding genes did not correspond to any previously encountered sequence. This finding was unexpected for an otherwise genetically extensively studied organism such as yeast. Before sequencing of the complete yeast genome started, identification of the same new gene by independent investigators had been becoming frequent; leading to the notion that the yeast genome had become over-studied. When the complete yeast genome was sequenced it was estimated,

depending on the stringency criteria applied, that 30% to 35% of 6275 predicted genes are without any match to other proteins in the gene databases or without any functional information. Inability of genetic screens to uncover substantial proportion of genes and inability of researchers to transfer functional information to these genes using sequence similarity motivated Dujon to name this unforeseen result “the mystery of orphan genes” (Dujon, 1996).

It is important to note that the term ‘orphan’ in this initial analysis of the yeast genome had a double meaning, namely coding regions without known function and coding regions without matches to other genes in the database (Dujon, 1996). However, taking into account only lack of the sequence similarity, a later study came to a similar proportion of yeast orphans (Malpertuy et al., 2000). To overcome confusion because of the initial functional connotation of the orphan definition Malpertuy and co-workres (2000) proposed the term 'maverick' gene for a gene with lack of sequence similarity to other organisms. However, the definition of orphan genes as coding regions without matches to other genes in the database is usually used (Fischer and Eisenberg, 1999; Schmid and Aquadro, 2001; Jordan et al., 2002a).

2.1.3 Orphan genes are ubiquitous in the genomes

Genome projects of the eukaryotic and prokaryotic (Fischer and Eisenberg, 1999) organisms confirmed findings in the analysis of the yeast genome. *Table 1* summarizes approximate orphan content for some completely sequenced eukaryotic genomes based on the original genome publication data. Although similarity searches in these studies were performed in a not directly comparable way, because of different databases sizes used (Spang and Vingron, 2001), differences in their content and varying significance thresholds, it can be said that almost each newly sequenced genome brought a large number of new orphan genes.

Taken together, it can be concluded that the genome sequencing projects uncovered a substantial proportion of genes without sequence similarity in other organisms that were also missed by various previous functional approaches. Although this phenomenon is not a trivial issue of the genomic and post-genomic research, a small number of studies have addressed this question, and very often just as a side topic.

Table 1. Approximate orphan gene content in some of the completely sequenced eukaryotic genomes

Organism	Year of publication	No. of genes	Orphan genes	Taxonomic group[a]	Reference
Saccharomyces cerevisiae	1996	6275	30-35 %	*Saccharomyces cerevisiae*	(Goffeau et al., 1996; Dujon, 1996)
	2000	5651	32 %	Ascomycetes	(Malpertuy et al., 2000)
Caenorhabditis elegans	1998	18600	58 %	Nematoda	(Blaxter, 1998; The C. elegans Sequencing Consortium., 1998)
Drosophila melanogaster	2000	13601	~ 30 %	*Drosophila melanogaster*	(Rubin et al., 2000; Adams et al., 2000)
	2002	13885	18.6 %	*Drosophila melanogaster*	(Zdobnov et al., 2002)
	2002	13885	34.5 %	Insecta	(Zdobnov et al., 2002)
Homo sapiens	2001	31778	25 %	*Homo sapiens*	(Lander et al., 2001)
Schizosaccharomyces pombe	2002	4824	14 %	*Schizosaccharomyces pombe*	(Wood et al., 2002)
			30 %	*Schizosaccharomyces pombe* and *Caenorhabditis elegans*	(Wood et al., 2002)
Ciona intestinalis	2002	15852	21 %	*Ciona intestinalis*	(Dehal et al., 2002)
Anopheles gambie	2002	12981	11.1 %	*Anopheles gambie*	(Zdobnov et al., 2002)
	2002	12981	29.0 %	Insecta	(Zdobnov et al., 2002)
Mus musculus	2002	22011	~ 0 %	*Mus musculus*	(Waterston et al., 2002)
	2002	22011	14 %	Mamalia	(Waterston et al., 2002)
	2002	22011	20 %	Chordata	(Waterston et al., 2002)
Arabidopsis thaliana	2000	25498	?	-	(Arabidopsis Genome Initiative, 2000)
Oryza sativa	2002	53398	~50 %	monocots	(Yu et al., 2002)

[a] Taxonomic rank used to define orphan genes. If a gene lacks sequence similarity to a sequence outside the stated taxonomic rank it is considered as an orphan gene.

2.2 Are there trivial explanations for phenomenon of orphan genes?

2.2.1 Do orphan genes code for real proteins?

The first trivial explanation, which could account for the existence of orphan genes, is that orphan genes are just over-predicted open reading frames (ORFs) that do not code for the functional proteins. Correct selecting of ORFs (which are coding for real proteins) from an 'ORFome' (total number of possible ORFs) is recognized as the main problem in defining the proteome of an organism. (Zhang, 2002; Harrison et al., 2002; Parra et al., 2003). Direct functional analysis and different types of transcriptome analysis, e.g. expressed sequenced tags (EST) projects, full length cDNA sequencing, serial analysis of gene expression (SAGE) (Velculescu et al., 1997) and microarray analysis (Shoemaker et al., 2001; Clark et al., 2002), are used to improve pure *ab initio* or homology based annotation of the genomes. However, high-throughput experimental approaches for identification of genes and their functions are still in development. As a result, reliable experimental genomic data, necessary for precise annotations and improvement of prediction tools, is still missing (Zhang, 2002).

Because of the above reasons the gene count for many completely sequenced eukaryotic genomes is still debated. Even the true size of the yeast proteome has been a point of considerable confusion, although its complete genome is available for already seven years. In the beginning, as high orphan content of the yeast genome was unexpected and confusing, several studies based on statistical properties of known genes tried to correct the gene count arguing that many of the ORFs are over-predicted (Kowalczuk et al., 1999; Mackiewicz et al., 1999; Zhang and Wang, 2000). However, when the partial genome sequences of a set of closely related Hemiascomycetou*s* yeasts became available, it was possible to support the annotation of many orphan genes based on sequence similarity. This study showed that, although the total estimated number of genes dropped by 9 % compared to the initial one (*Table 1*), the proportion of orphans, now defined as Hemiascomycetous yeast specific genes, remained the same. This result suggests that, most likely, miss-annotation is not the major determinant that can account for the existence of orphan genes, at least not in the yeast genome. However, this study does not provide evidence that the regions, having similarity to the yeast orphans, are indeed protein coding. Transcriptional analysis of these regions is indispensable to show that they

are coding for functional proteins. On the other hand, sequencing of two closely related bacterial species *Mycoplasma pneumoniae* and *Mycoplasma genitalium* brought orthologues for the most of the predicted genes (Himmelreich et al., 1997). In the same way, the recent sequencing of human (Lander et al., 2001) and mouse genome (Waterston et al., 2002), which are closely related organisms in the terms of evolutionary rates, brought support for many mammal specific orphans. However, in this case caution is necessary because of a very unreliable annotation of these genomes (Harrison et al., 2002; Xuan et al., 2003; Parra et al., 2003).

Contrary to the above findings, a direct study of four orphans from the *Drosophila melanogaster Adh* region found that their ORFs were interrupted in the closely related species *D. simulans* or *D. yakuba,* indicating that they are not real genes (Schmid and Aquadro, 2001). Taken together, it is not yet clear which proportion of orphans are functional proteins, although several studies suggest that many of them are real genes.

2.2.2 Do orphan genes reflect incompleteness of gene databases?

Based on the studies reported in the previous section it seems that reliable annotation of a genome requires sequencing of two or more closely related species and that orphan genes will have orthologues only in the closely related organisms. However, another trivial explanation for orphan genes could be that they are genes that do have homologues in other distantly related organisms but that these organisms are not yet sequenced. Indeed, complete genome sequences of many phyla are missing in the databases. On the other hand, if incompleteness of gene databases explains why most of the genes are orphans then accumulation of enough sequence information in the databases would reduce their number. However, all genome projects so far have identified a substantial fraction of open reading frames that have no similarity to the other genes in the database, demonstrating that the fraction of orphans cumulatively does not diminish (Fischer and Eisenberg, 1999; Rubin et al., 2000) (*Table 1*). Accordingly, this defies early hopes that an increasing database size would eventually reduce the number of orphan genes (Casari et al., 1996). On the other hand, there is also the possibility that the original reports about orphans are outdated and that previously classified orphans can now find matches to newly sequenced genes. Indeed, some decay in the number of bacterial orphans can be observed, but their proportion in bacterial genomes is still significantly high

(Fischer and Eisenberg, 1999). Nevertheless, rigorous tests on the current number of orphans for many sequenced genomes especially eukaryotic ones are missing.

2.2.3 Are orphan genes abundant copies of several genes?

If one takes orphan genes as reality, their abundance may alternatively be explained by a high copy number of several duplicated orphan genes. Fischer and Eisenberg (1999) tested the possibility that a high frequency of orphans in bacteria is due to the existence of paralog families of orphan genes. Nevertheless, the frequency of recovered orphan protein families was also high. Moreover, they notice that bacterial orphans are less likely to be members of paralog families compared to other proteins. This observation is unexplained and opens the question about the evolutionary dynamics of orphan genes (Fischer and Eisenberg, 1999).

2.3 Common assumptions about the evolutionary origin of orphan genes

If a substantial fraction of orphan genes code for functional proteins, then the next question is about their evolutionary origin. There are two most commonly used explanations for the lack of sequence similarity of orphan genes. The first one is that orphan genes are fast evolving genes and the second one is that they are lineage specific genes (Blaxter, 1998; Fischer and Eisenberg, 1999; Wolfe and Sharp, 1993; Malpertuy et al., 2000; Rubin et al., 2000; Schmid and Aquadro, 2001; Rubin, 2001; Dehal et al., 2002). Certainly, these two possibilities are not expected to be mutually exclusive.

Several studies indirectly approached the question of protein evolution rate of orphan genes. A lower sequence conservation between genes of unknown functions, as compared with the functionally assigned genes, has been observed for the two related bacterial species *Mycoplasma pneumoniae* and *Mycoplasma genitalium* (Himmelreich et al., 1997). As unknown function is often coupled with lack of sequence similarity to distantly related organisms this was the hint that orphan genes might have different evolutionary rates. In a more direct approach, it was shown that sequence similarity between *Kluyveromyces lactis* and *Saccharomyces cerevisiae* is lower for orphans then for non-orphans (Ozier-Kalogeropoulos et al., 1998). Similar results were obtained in the analysis of the partial genomic sequence of other closely

related yeast species (Malpertuy et al., 2000). This is also the only study, which gives a hint that at least some of the orphan genes could have reasonably low divergence rates, indicating that orphans might be lineage specific genes as well (Malpertuy et al., 2000). However, these results were based on the BLAST E-values and amino acid identities, which are rather rough measures of sequence divergence. In addition, these studies were based on the partial gene sequences derived form genomic regions, and thus they lack stronger evidence that aligned sequences are coding for real proteins.

Schmid and Tautz (1997) by genomic hybridisation studies and sequencing of orthologs from *D. melanogaster* and *D. yakuba* showed that the fraction of fast evolving genes in Drosophila is about 30%, roughly matching the percentage of orphan genes predicted in the *Drosophila* genome (Rubin et al., 2000). However, not all fast evolving genes were orphan genes. For example, a zinc-finger transcription factor and a functional homologue of a yeast chaperone gene was found in the class of fast evolving genes (Schmid et al., 1999; Wang et al., 1999). Both of these do not qualify as orphan genes as they match at least partially with known protein domains. In addition, the relationship between average rate of sequence evolution and orphan gene status could not be established unequivocally because the applied hybridisation technique lacks the sensitivity and because public databases contained in the time of that study only the yeast genome as completely sequenced eukaryotic organism.

2.4 Functional and other properties of orphan genes

2.4.1 Function of orphan genes

As mentioned in the previous part, lack of sequence conservation of orphan genes is coupled with lack of their functional assignment, not only due to the inability of researchers to infer functional information using the sequence similarity but also because phenotype information for orphan gene mutants was not obtained by the genetic studies. This was originally found in the yeast project (Oliver et al., 1992; Dujon, 1996) but was also noted in the extensive study of the *Adh* region in *Drosophila* (Ashburner et al., 1999) and the analysis of fast evolving genes (Tautz and Schmid, 1998). Comparison of genomes of bacterial strains of the same species also suggest that strain specific genes are over-represented among functionally uncharacterised genes (Jordan et al., 2002b).

Indirectly, a possible function of orphan genes can be traced trough some comparative genomics and yeast studies. For example, genomic exploration of the closely related yeast species shows that orphan genes are especially abundant among proteins involved in the extracellular secretion and in the organisation of the cell wall (Gaillardin et al., 2000). Interestingly, both of these functional classes were extensively used as taxonomic markers (Phaff, 1998). In bacterial, archaeal and eukaryotic organisms some of the proteins with narrow phyletic distribution were shown to function at the periphery of the cell. More specifically, some of them were predicted membrane proteins that may mediate the interaction of the cells with their environment (Jordan et al., 2001; Jordan et al., 2002b).

2.4.2 Distinct features of orphan genes

Several studies report some additional distinct properties of orphan genes. For example, Lipman et al. (2002) found in a comparison between two prokaryotes, yeast, *Drosophila* and humans that non-conserved genes are generally shorter than conserved ones and that their length distribution is more uniform. This could be explained if non-conserved genes are under weaker selective constraints and would thus more easily tolerate deletion mutations. The comparison between the *Drosophila* and the *Anopheles* proteome shows also that the orphans that are specific for each species have the shortest average length (Zdobnov et a. 2002).

There is also indication that orphan genes are generally lower expressed than non-orphan genes. The observation that phylogenetically conserved genes are more highly expressed tested by occurrence of ESTs was first made by Green et al. (1993) (Green et al., 1993) and was confirmed in the analysis of the *Adh* region in *Drosophila* (Ashburner et al., 1999).

2.5 Orphan genes in Drosophila melanogaster genome

The first annotation of the *D. melanogaster* genome uncovered that 28% of predicted genes has no sequence similarity to other organisms (Adams et al., 2000). However, a systematic study or orphan genes in the *Drosophila* at the genome level is still missing, although high orphan gene content was announced three years before as an important open question of fly biology (Rubin et al., 2000). Since then

only one study directly analysed the evolutionary properties of four orphan genes (Schmid and Aquadro, 2001).

2.6 Open questions

Based on the current state of the literature many of the important questions concerning orphan genes are not answered. For example, it is not clear which fraction of orphans are coding for real proteins, especially in the eukaryotic organisms. Although repeatedly noted, under-representation of orphans among studied genes was not tested on the genome level. The evolutionary origin of orphan genes is also still enigmatic. Two proposed reasons for the lack of sequence similarity of orphans, namely rapid evolution of coding sequence and/or lineage specific localization of these genes, have not yet been tested rigorously. There is definite scarcity of information concerning the function of orphan genes, although some functional roles are suggested. Moreover, protein properties, expression profiles and position in biochemical pathways are almost completely unexplored for orphan genes.

3. Aim of the study

The aim of this thesis was to study evolutionary dynamics, as well as sequence properties of the orphan genes in *Drosophila,* with view to understand their evolutionary origin and general functional patterns.

The following aspects were in special focus of this study:

- Proportion of orphan genes in the *Drosophila melanogaster* genome
- Under-representation of orphan genes in the genetic studies
- Testing of hypothesis that orphan genes are fast evolving genes
- Testing of hypothesis that functionally more important *Drosophila* genes have lower evolutionary rates
- Comparison of evolutionary rates between adult and embryo transcripts
- Expression levels of orphan genes trough ontogeny of *Drosophila* and their relation to possible genetic or developmental constraint
- Statistical analysis of functional patterns of previously characterized orphan genes
- Spatial expression of orphan genes in the *Drosophila* embryo

4. Results

4.1 Analysis of orphan genes in the D. melanogaster genome

4.1.1 Orphan gene content

As gene database content is increasing exponentially and annotation of the complete genomes is improving some change in the number of orphan genes in *Drosophila* genome can be expected. The current database was therefore re-analysed using BLASTP with the about 14,300 predicted full-length proteins of the *Drosophila melanogaster* proteome (release 2), to re-analyse whether the fraction of orphans reported previously (Rubin et al., 2000) has changed over time. As the probability of identifying a significant BLAST match depends on the size of the database (Spang and Vingron, 2001), it is not possible to use a single probability cutoff criterion for assigning orphan status. To overcome this uncertainty, a range of probability cutoffs was used. For each cutoff category, as defined through the expectation (E)-values provided by BLAST (Altschul et al., 1990; Altschul et al., 1997), the fraction of genes was determined whose matches above this cutoff occurred only in *Drosophila* or other insects.

Figure 1a shows the results for cutoff E-value classes from 10 to 10^{-100}. The number of non-matching sequences is very small at the highest E-values, but this is evidently due to many insignificant chance matches. With continuously lower E-values there is a continuous increase in the non-matching sequences and there is no obvious criterion for choosing a particular E-value as a cutoff criterion for orphan genes. Most studies prefer to take cutoff values form 10^{-3} to 10^{-5} to discriminate significant matches from 'noise' in a similar type of database search (e.g. Lipman et al., 2002), whereby the 10^{-3} cutoff value is considered as rather conservative. In this analysis for cutoff classes form 10^{-3} to 10^{-5}, the fraction of orphan genes is 26 to 29 % (marked in *Figure 1a*). When the BLAST output data were inspected manually and decision about the significance of a match was done case-by-case, most of the E-values were also fitting to the above range. Besides these arguments, additional support that the chosen cutoff values are appropriate comes from analysis of the named genes in *Drosophila* genome (see section *4.1.2*). Based on these results it

can be concluded that the fraction of the orphan genes in the *Drosophila* genome is still comparable to what has been repeatedly found in the past (Rubin et al., 2000). Therefore, neither the growth of the database nor the re-annotations have significantly changed this value over time.

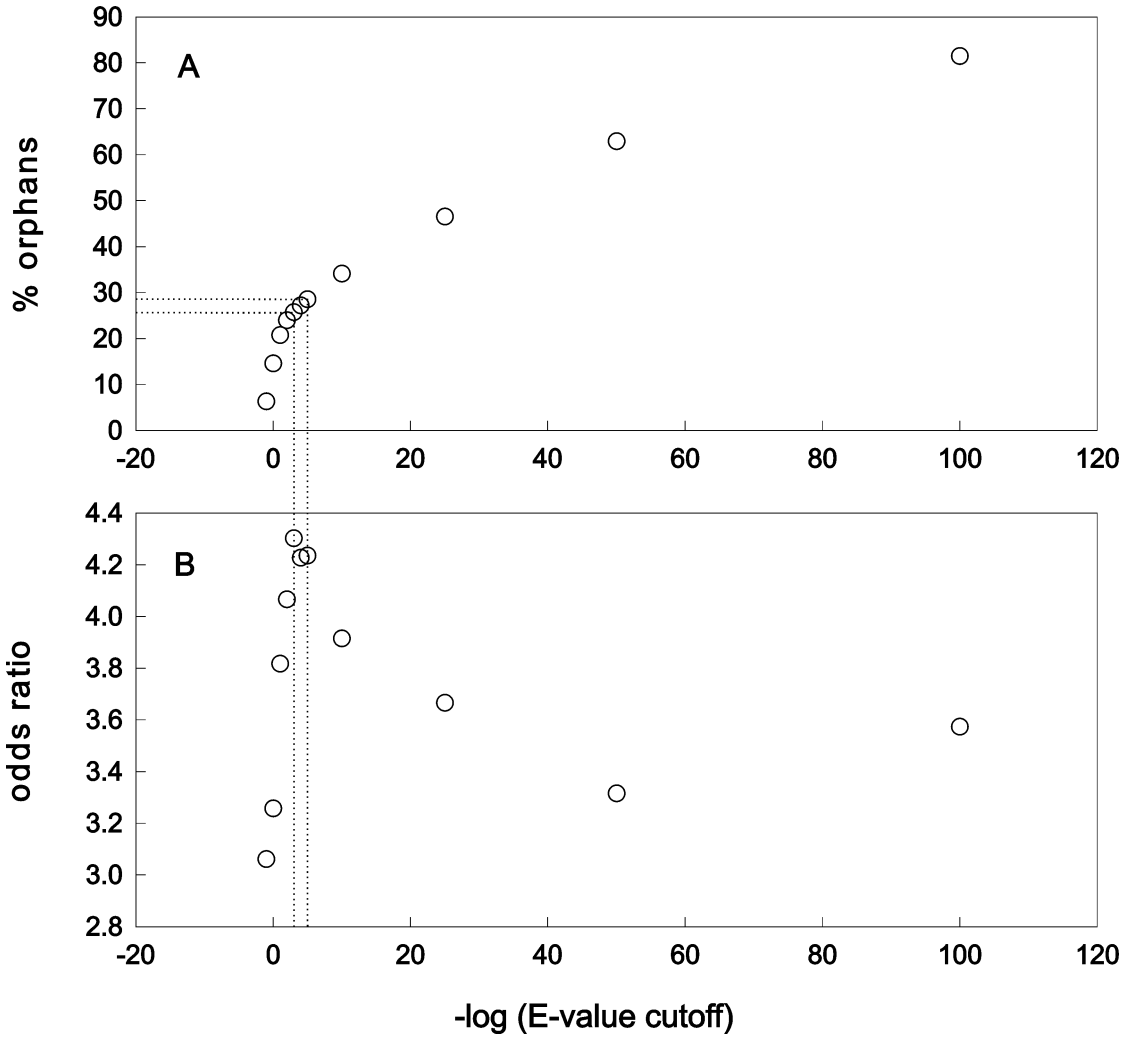

Figure 1. A: Percentage of orphans found in each cutoff category. The broken lines indicate the BLAST E-value range of 10^{-3} to 10^{-5}, for which 26 to 29% orphan genes and the highest odds ratio were found (see below) **B:** Odds-ratios for genetically studied genes in the different cutoff classes. The values indicate how much more likely one finds a genetically studied gene in the non-orphan compared to the orphan class. All values are highly significant (P = 0 Fischer's exact test).

4.1.2 Genetically studied orphan and non-orphan genes

In *Drosophila*, one can take the fact that a gene has been named as an approximate indicator that it has been genetically studied, i.e. that a described mutant exists for it. Therefore, the relative proportion of genetically studied genes was analysed in all cutoff categories. There are currently 3,633 named genes in *Drosophila*, which correspond to about 26 % of the known ORFs. Differences in the number of named genes were compared between the orphan and non-orphan sample for each cutoff category (*Table 2*) and corresponding odds-ratios were calculated (*Figure 1b*). The results show that named genes, independent of the chosen E-value cutoff, are more likely to occur in the non-orphan class. The odds ratio of finding named genes among non-orphans compared to orphans is the highest for 10^{-3} to 10^{-5} cutoff class, supporting the notion that orphan genes are less likely to be recovered in the current genetic screens.

Interestingly, the odds ratio analysis has a peak at the same cutoff range as the one chosen to re-estimate the proportion of orphan genes in the *Drosophila* genome (see section *4.1.1*). This supports independently the correctness of the chosen cutoff range, because such a peak can be expected only for a non-orphan sample with the lowest level of incorrectly assigned genes. This reasoning is based on the assumption of general over-representation of named genes in the non-orphan sample. Accordingly, more loose or stringent cutoff values then the optimal one would change the odd ratio by introducing false positives or excluding false negatives (*Figure 1*).

Additionally, it is interesting to note that even if E-value cutoff of 10^{-50} is used as threshold for a significant match, 58% of the named *Drosophila* genes are still among non-orphans, although the non-orphan class contains for this threshold only 37 % of all genes. This is indication, if BLAST E-value is taken as a rough measure of sequence conservation, that genetic studies have focused on phylogenetically strongly conserved genes, whereas lineage specific and phylogenetically broadly distributed weekly conserved genes were more likely to be overlooked.

Table 2. Number of named genes in the orphan and non-orphan sample (complete *Drosophila* genome)

Genes (E-value cutoff = 10)		
	Not named	**Named**
Non-orphan	9805 (73.5 %)	3538 (26.5 %)
Orphan	806 (89.5 %)	95 (10.5 %)
Genes (E-value cutoff = 1)		
	Not named	**Named**
Non-orphan	8755 (72.0 %)	3411 (28.0 %)
Orphan	1856 (89.3 %)	222 (10.7 %)
Genes (E-value cutoff = e-1)		
	Not named	**Named**
Non-orphan	7949 (70.4 %)	3340 (29.6 %)
Orphan	2662 (90.1 %)	293 (9.9 %)
Genes (E-value cutoff = e-2)		
	Not named	**Named**
Non-orphan	7531 (69.5 %)	3301 (30.5 %)
Orphan	3080 (90.3 %)	332 (9.7 %)
Genes (E-value cutoff = e-3)		
	Not named	**Named**
Non-orphan	7296 (68.9 %)	3286 (31.1 %)
Orphan	3315 (90.5 %)	347 (9.5 %)
Genes (E-value cutoff = e-4)		
	Not named	**Named**
Non-orphan	7117 (68.6 %)	3255 (31.4 %)
Orphan	3494 (90.2 %)	378 (9.8 %)
Genes (E-value cutoff = e-5)		
	Not named	**Named**
Non-orphan	6949 (68.3 %)	3231 (31.7 %)
Orphan	3662 (90.1 %)	402 (9.9 %)
Genes (E-value cutoff = e-10)		
	Not named	**Named**
Non-orphan	6292 (67.1 %)	3091 (32.9 %)
Orphan	4319 (88.9 %)	542 (11.1 %)
Genes (E-value cutoff = e-25)		
	Not named	**Named**
Non-orphan	4866 (63.9 %)	2748 (36.1 %)
Orphan	5745 (86.7 %)	885 (13.3 %)
Genes (E-value cutoff = e-50)		
	Not named	**Named**
Non-orphan	3157 (59.8 %)	2122 (40.2 %)
Orphan	7454 (83.1 %)	1511 (16.9 %)
Genes (E-value cutoff = e-100)		
	Not named	**Named**
Non-orphan	1377 (52.2 %)	1263 (47.8 %)
Orphan	9234 (79.6 %)	2370 (20.4 %)

Differences were significant in all comparisons ($P = 0$, two-sided Fisher's exact test).

4.2 Comparative analysis of expressed genes in D. yakuba

4.2.1 Orphan gene content in the sample of expressed genes

Analysis of expressed genes allows avoiding mistakes due to wrong annotations. To study directly the evolutionary characteristics of orphan genes, cDNA libraries were prepared from *D. yakuba* embryos and adults and clones were picked randomly from these. The clones were initially 5'-sequenced to check for redundant clones and the non-redundant clones were then fully sequenced to high quality. Comparisons with the *D. melanogaster* genome sequence allowed to unequivocally identify the corresponding *D. melanogaster* orthologue in all cases. The full *D. melanogaster* gene sequence was then taken to determine whether it is an orphan applying the rather conservative cutoff criterion of $E > 10^{-4}$.

Approximately 400 non-redundant cDNAs were obtained from each of the two libraries (371 from the adult and 403 from the embryo library). Among these, 81 genes were found in both libraries and just one of them was orphan. The embryo library contains 42 and the adult library 81 orphan genes. To be certain that only true orphans were included, clones in which a weak match with an InterPro domain was present were removed, although significance of these weak matches may be questionable. This curation yielded 34 orphan genes for the embryo library (8.4 %) and 73 (19.7 %) for the adult library, which is highly significant difference ($P < 0.001$). This difference is analysed in more detail in sections *4.3.2* and *4.4*. On the other hand, the percentages are lower than one would have expected from the whole genome scan (27.1 % in the 10^{-4} class). This could either suggest that many of the genomic orphans are indeed due to wrong annotations (Schmid and Aquadro, 2001), or that orphans are generally lower expressed than non-orphan genes, with a corresponding under-representation in cDNA libraries. That less conserved genes may be generally lower expressed has also been noted before (see Introduction, *2.4.2*).

4.2.2 Genetically studied genes in the sample of expressed genes

Named genes are strongly under-represented among identified orphans. The odds ratio analysis shows that in the embryo library it is almost eight times and in the

adult library it is almost three times less likely to find a named gene in the orphan class than in the non-orphan class (*Table 3*). Still, 4 orphan genes in the embryo library and 15 in the adult library are previously named genes, but it is interesting to look at the nature of the named genes in the orphan class (*Table 4*). In the adult library genes with available functional information are involved in immune response, behaviour, oxygen deprivation or regulation of circadian rhythm and flight. All these functions can be expected to be important in a specific ecological context. Interestingly, for several of these mutants are not known, i.e. they were named because of other reasons.

Table 3. Number of named genes in the orphan and non-orphan sample (genes recovered in this study)

Genes (embryo)		
	Not named	**Named**
Orphan	30 (88.2 %)	4 (11.8 %)
Non-orphan	181 (49.1 %)	188 (50.9 %)
$P = 7.1 \times 10^{-6}$		
Genes (adult)		
	Not named	**Named**
Orphan	58 (79.5 %)	15 (20.5 %)
Non-orphan	169 (56.9 %)	128 (43.1%)

$P = 0.0004$, two-sided Fisher's exact test

Table 4. Previously named orphan genes that were identified among *D. yakuba* cDNA sequences

Name	Function	Mutants
Adult library		
ACP53EA	Accessory gland-specific peptide 53Ea	6 alleles known
AttA	Attacin-A, a gram-negative antibacterial peptide	none
AttD	Attacin-D, a putative antibacterial peptide	none
Cp16	Chorion protein 16 - structural protein of the chorion	none
Dpt	Diptericin, a gram-negative antibacterial peptide	none
DptB	Diptericin B, a putative antibacterial peptide	none
fau	An anoxia-regulated novel gene	none
fln	Required for thick filament in flight muscle	viable, but flightless
fok	Associated with kinesin-like molecule	none
l(2)k09913	Unknown function	recessive lethal
Mst89B	Testis specific expression, function unknown	none
Noe	Nervous system expression, function unknown	none
Os9	Olfactory system expression, function unknown	none
to	Circadian rythm regulated gene	rythm defective
yellow-c	Possibly involved in cuticle development	none
Embryo library		
GATAd	Non-specific RNA polymerase II transcription factor	none
mael	Involved in oocyte nucleus migration	recessive lethal
Tom	Interacts genetically with Su(H)	recessive lethal
Df31	Component of the chromatin	recessive lethal

4.2.3 Sequence properties of the expressed orphan genes

The identified orphan genes differ also in several other respects from non-orphan genes. They are on average more than 100 amino acids shorter, have lower GC content, lower codon usage bias and fewer exons. All of these differences are statistically significant (*Table 5*). Likewise, the number of paralogs is lower in the orphan sample. If two samples are compared, not taking into account the number of paralogs per gene (*Table 6*), the difference is significant but not large. Interestingly, when the number of paralogs for each gene is included, non-orphan genes have on average more than four times more paralogs ($N_{ORPHAN} = 2.7 \pm 0.6$; $N_{NON\text{-}ORPHAN} = 12.3 \pm 1.3$; $P = 0.006$, Mann-Whitney U test).

Table 5. Statistical comparisons between orphan and non-orphan cDNAs.

	Orphans	**Non-orphans**		
No.	106	586		
	Mean ± 1SE	Mean ± 1SE	*t test*	*P*
aa length	224 ± 13	356 ± 14	-4.994	7.5 x 10-7
GC	0.541 ± 0.0050	0.553 ± 0.0020	-2.231	0.026
GC3	0.638 ± 0.0122	0.688 ± 0.0049	-3.950	8.6 x 10-5
ENC	47.7 ± 0.79	44.22 ± 0.35	3.872	1.2 x 10-4
Fop	0.527 ± 0.0120	0.591 ± 0.0054	-4.726	2.8 x 10-6
Exon number	2.5 ± 0.16	3.5 ± 0.09	-5.545	1.2 x 10-7

Mean and standard errors of the mean are given. Significance of differences were tested using Student's t. Values are derived from the full length *D. melanogaster* homologues of the D. yakuba cDNAs. GC is general GC content, GC3 is GC content at third codon positions. ENC (effective number of codons) and Fop (frequency of optimal codons) are measures of codon usage bias.

Table 6. Genes with paralogues in the orphan and non-orphan sample

Genes	**Paralogues**	
	0	≥1
Orphan	62 (57.9%)	45 (42.1%)
Non-orphan	313 (46.9%)	354 (53.1%)

Difference is significant (*P* = 0.032, 2-sided, Fischer's exact test). Numbers in parenthesis represent percent of genes in the respective class. The analyzed sample consists of *D. melanogaster* genes, which are homologues to the non-redundant cDNAs recovered from D. yakuba. Each gene was compared by BLASTP against the complete *D. melanogaster* coding sequence (FlyBase Release 2). If a gene had at least one BLASTP hit with an E-value $<10^{-10}$ it was considered to have a paralogue in the *D. melanogaster* genome.

4.3 Substitution rates of the expressed genes in D. yakuba

4.3.1 Substitution rates of orphan and non-orphan genes

Substitution rates at coding (*dN*) and non-coding (*dS*) positions were determined for embryo (381) and adult (356) *D. yakuba* cDNAs aligned to the corresponding *D. melanogaster* genes. In this data set, 71 cDNAs were present in both libraries. Removing the respective shorter cDNA from these duplicate pairs yielded a non-redundant set of 659 cDNAs. None of the genes has a *dN*/*dS* ratio larger than one, which would be indicative of fast evolution due to positive selection. For 18 non-redundant genes (2 orphans and 16 non-orphans) it was not possible to reject the hypothesis that their rate is significantly different from one (*Figure 2*). However, many of these genes showed only a small total number of substitutions (Appendix *Table 20*).

Table 7 summarizes the rate comparisons. As a class, orphan genes have a more than three times higher non-synonymous substitution rate compared to non-orphan genes ($dN_{\text{ORPHAN}} = 0.062$ versus $dN_{\text{NON-ORPHAN}} = 0.020$). When the adult and embryo transcripts are compared separately, orphan genes from the embryo library are evolving more than four times faster compared to non-orphans, while adult orphan genes almost three times faster (*Table 7*). A similar trend but with a lower proportion is seen for the synonymous substitution rates ($dS_{\text{ORPHAN}} = 0.335$ versus $dS_{\text{NON-ORPHAN}} = 0.277$) in the complete sample, and when embryo and adult transcripts are considered separately (*Table 7*).

Several studies reported positive correlation between *dN* and *dS* in different organisms including *Drosophila* (Duret and Mouchiroud, 2000; Comeron and Kreitman, 1998; Dunn et al., 2001). In this study, significant correlation between *dN* and *dS* is also detected for the complete sample ($r_{\text{ALL GENES}} = 0.443$, $P = 3.5 \times 10^{-22}$), and in both subclasses ($r_{\text{ORPHAN}} = 0.487$, $P = 2.4 \times 10^{-7}$; $r_{\text{NON-ORPHAN}} = 0.408$, $P = 5.2 \times 10^{-22}$). Therefore, this correlation may at least partially account for the increased *dS* rates of orphans. In mammals neighbouring effects like double mutation at adjacent sites were proposed to explain this correlation (Duret and Mouchiroud, 2000). In *Drosophila* it is suggested that relaxed constraint exists on both kinds of substitutions in a particular codon (Comeron and Kreitman, 1998).

Although *dN* and *dS* are correlated, the *dN*/*dS* ratio of orphan genes is on average 2.5 times higher than of non-orphan genes (*Table 7*), indicating that orphan proteins are less constrained by purifying selection. Taken together these results rule out the null-hypothesis that orphan and non-orphan genes have equal rates of evolution. Although orphan genes evolve on average significantly faster than non-orphan genes, there is nonetheless a broad distribution of different rates for both classes of genes (*Figure 3* and *Figure 4*). Intriguingly, sequences with very low divergence rates (*dN* < 0.0032, *dN*/*dS* < 0.02) were found in the orphan gene class, which is in the range of highly conserved non-orphan genes. Thus, orphan genes are not necessarily all fast evolving genes.

Table 7. Substitution rate comparisons between orphan and non-orphan cDNAs

cDNA	Variable	Orphans	Non-orphans	Ratio	*t* test	*P* value
All	*dS*	0.335 ± 0.0130 (n = 100)	0.277 ± 0.0060 (n = 559)	1.2	3.814	1.5×10^{-4}
	dN	0.062 ± 0.0077 (n = 100)	0.020 ± 0.0014 (n = 559)	3.1	7.562	8.5×10^{-12}
	dN/*dS*	0.171 ± 0.0157 (n = 100)	0.068 ± 0.0043 (n = 559)	2.5	7.928	7.8×10^{-13}
Embryo	*dS*	0.323 ± 0.0240 (n =31)	0.265 ± 0.0078 (n =350)	1.2	2.098	0.037
	dN	0.069 ± 0.0189 (n = 31)	0.016 ± 0.0013 (n = 350)	4.3	3.388	5.1×10^{-4}
	dN/*dS*	0.182 ± 0.0345 (n = 31)	0.060 ± 0.0052 (n = 350)	3.0	4.257	1.7×10^{-4}
Adult	*dS*	0.344 ± 0.0157 (n= 70)	0.266 ± 0.0079 (n = 286)	1.3	4.382	1.5×10^{-5}
	dN	0.063 ± 0.0082 (n = 70)	0.022 ± 0.0022 (n = 286)	2.9	6.753	1.3×10^{-9}
	dN/*dS*	0.172 ± 0.0177 (n = 70)	0.073 ± 0.0086 (n = 286)	2.4	7.104	6.7×10^{-12}

Mean and standard errors of the mean are given. Significance of differences was tested using Student's *t*.

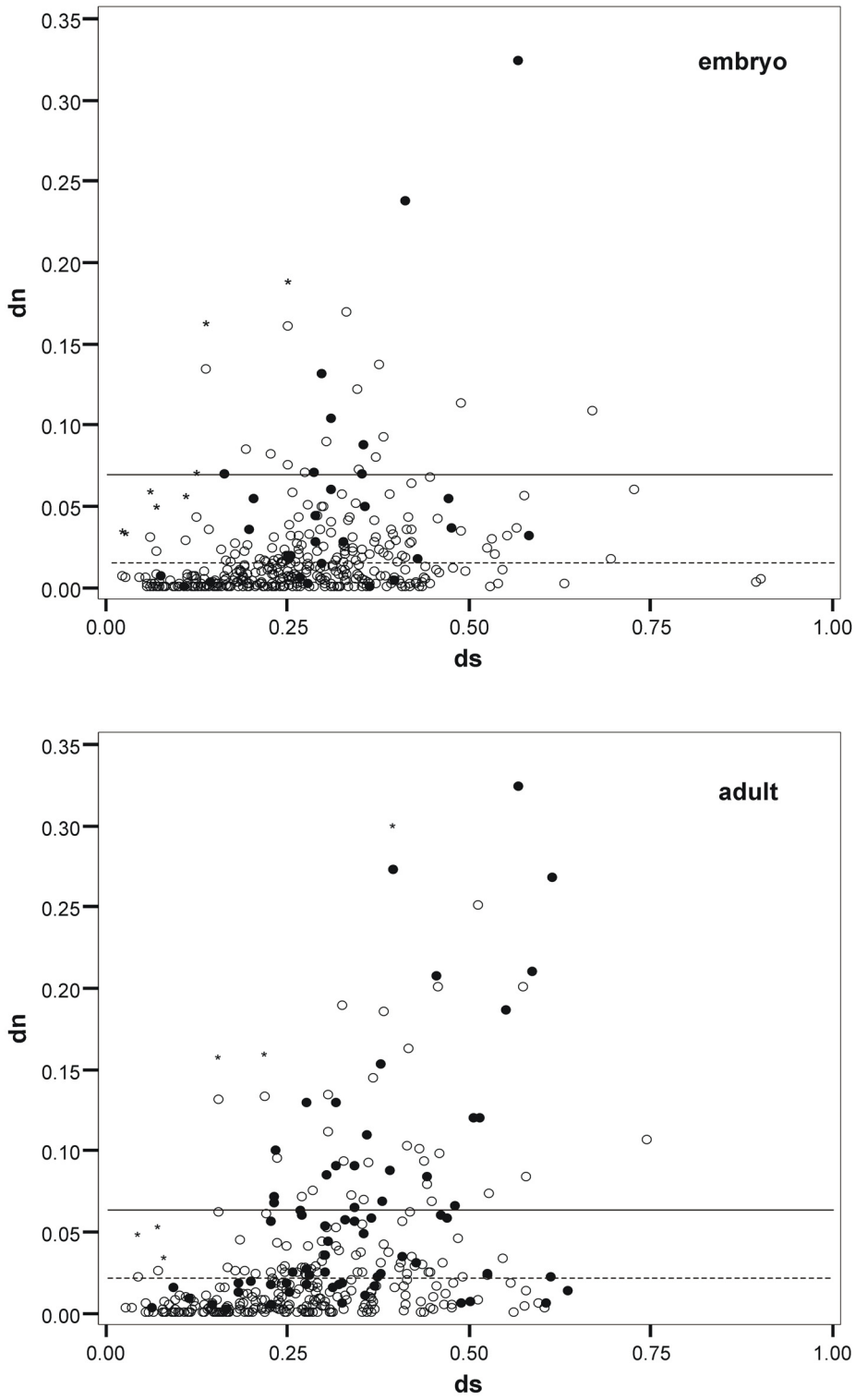

Figure 2. Scatter plot of the nucleotide substitution rates at synonymous (*dS*) and non-synonymous (*dN*) sites for the embryo (above) and the adult library (below). Orphan genes are represented as filled circles and non-orphan genes as open circles. The mean of the *dN´*s for the orphan genes is marked as solid line and for non-orphan genes as dashed line. Genes for which the null hypothesis that *dS* and *dN* are equal can not be rejected are marked with a star.

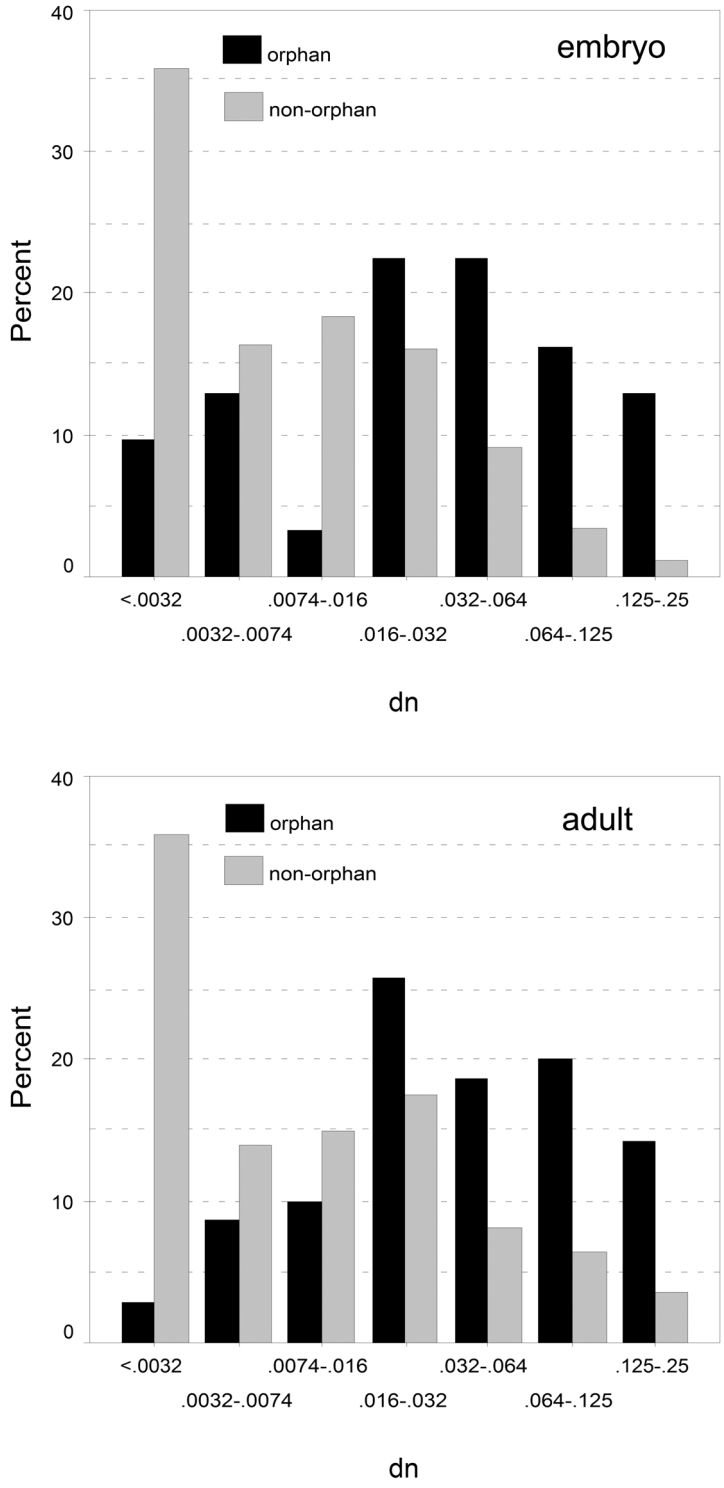

Figure 3. Discrete distribution of non-synonymous substitutions (*dN*) for the embryo (above) and the adult (below) library. The percentages of genes falling into the respective *dN* value classes are represented by black (orphans) and gray (non-orphans) columns. Note the logarithmic scale for representing the *dN* value classes.

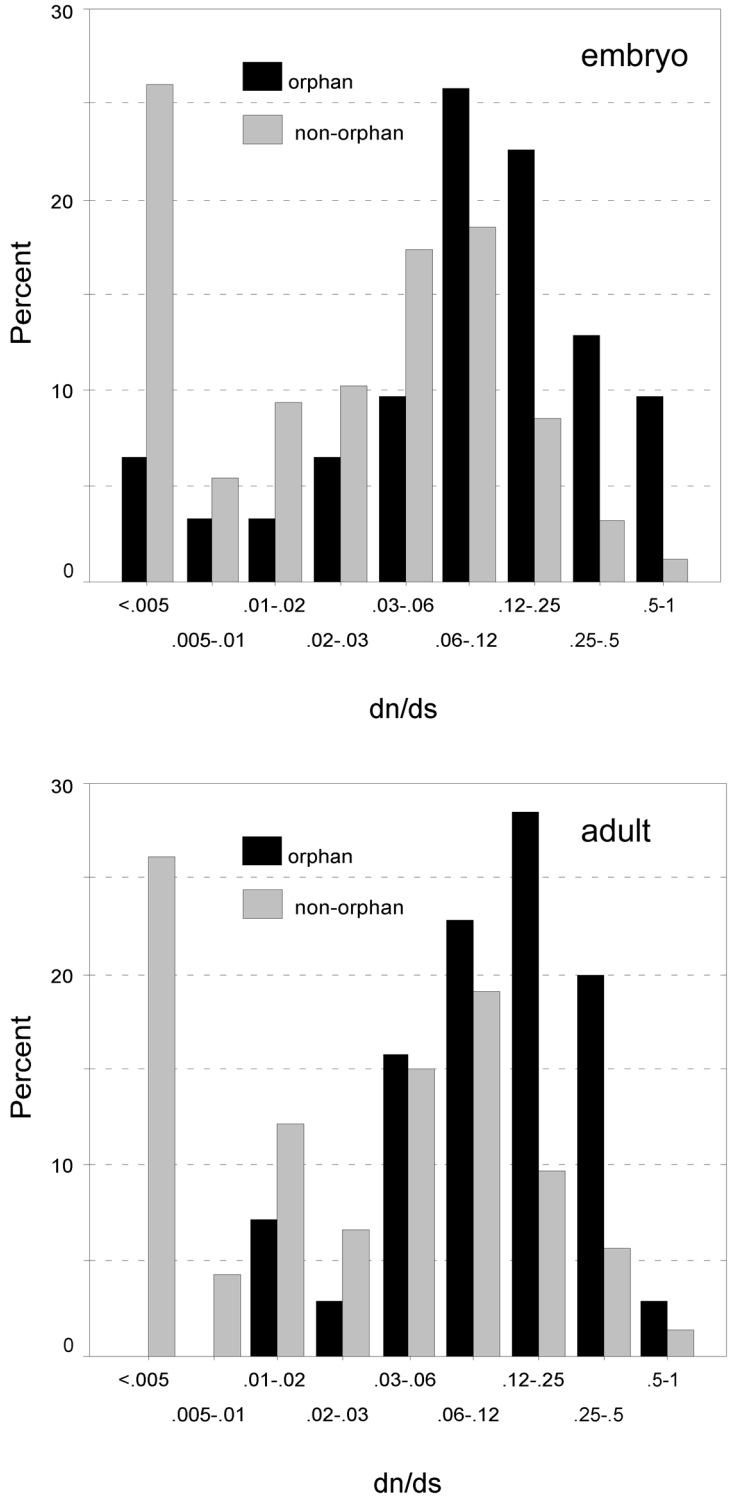

Figure 4. Discrete distribution of *dN/dS* ratios for the embryo (above) and the adult (below) library. The percentages of genes falling into the respective *dN*/*dS* value classes are represented by black (orphans) and gray (non-orphans) columns. Note the logarithmic scale for representing the *dN*/*dS* ratio classes.

4.3.2 Substitution rates of embryo and adult transcripts

The proportion of orphan genes is higher among adult transcripts (see section *3.2.1*) and therefore it is interesting to analyse how this difference may influence substitution rates between adults and embryos. There are fewer highly conserved orphan genes in the adult library than in the embryo library (*Figure 4*), but the average non-synonymous substitution rate, synonymous substitution rate and *dN/dS* ratio are nonetheless not significantly different for the orphan genes in both libraries (*Table 8*). The same is true for the non-orphan genes (*Table 8*). Thus, the fact that the average *dN* and *dN*/*dS* ratios are higher among the cDNAs recovered from the adult library (dN_{ADULT} = 0.030 versus dN_{EMBRYO} = 0.020; dN/dS_{ADULT} = 0.093 versus dN/dS_{EMBRYO} = 0.070) is apparently solely due to the fact that there are more orphan genes among adult transcripts.

Table 8. Substitution rate comparisons between cDNAs from the adult and embryo library

cDNA	Variable	Adult	Embryo	Ratio	*t* test	*P* value
All	*dS*	0.281 ± 0.0072 (n = 356)	0.270 ± 0.0075 (n = 381)	1.0	-1.061	0.289
	dN	0.030 ± 0.0026 (n = 356)	0.020 ± 0.0021 (n = 381)	1.5	3.321	0.001
	dN/dS	0.093 ± 0.0068 (n = 356)	0.070 ± 0.0058 (n = 381)	1.3	2.770	0.006
Orphan	*dS*	0.344 ± 0.0157 (n= 70)	0.323 ± 0.0240 (n =31)	1.1	0.741	0.460
	dN	0.063 ± 0.0082 (n = 70)	0.069 ± 0.0189 (n = 31)	0.9	- 0.175	0.861
	dN/dS	0.172 ± 0.0177 (n = 70)	0.182 ± 0.0345 (n = 31)	0.9	- 0.071	0.943
Non-orphan	*dS*	0.266 ± 0.0079 (n = 286)	0.265 ± 0.0078 (n =350)	1.0	0.037	0.971
	dN	0.022 ± 0.0022 (n = 286)	0.016 ± 0.0013 (n = 350)	1.4	1.883	0.060
	dN/dS	0.073 ± 0.0086 (n = 286)	0.060 ± 0.0052 (n = 350)	1.2	1.403	0.161

Mean and standard errors of the mean are given. Significance of differences was tested using Student's *t*. The 81 clones that were found in both libraries were excluded from the comparisons.

4.3.3 Substitution rates and genetically studied genes

If one assumes that slow evolving genes have important and more general functions, than the probability of recovery of these genes by classical functional genetic methods would be higher than for the fast evolving genes. This would hold under the assumption that the dispensability of genes is correlated with the rate of protein evolution. A recent study showed that this is indeed the case in bacteria (Jordan et al., 2002b; Jordan et al., 2002a), but for eukaryotic organisms the situation is not completely clear (Hurst and Smith, 1999; Hirsh and Fraser, 2001; Jordan et al., 2002a). As was mentioned before, in *Drosophila* one can take the fact that a gene has been named as an approximate indicator that an observable phenotype exists for it (section *4.1.2*). Thus, evolutionary rates calculated for the genes recovered in this study give an opportunity to test hypothesis that a clear phenotype is correlated with evolutionary rate. This analysis can be done for all genes or just specifically for orphans and non-orphans.

The analysis of non-synonymous substitution rates and of *dN/dS* ratio for the complete sample shows that there is a significant difference in the proportion of named genes (genetically studied genes) between the slow and fast evolving group irrespective of the threshold used (*Table 9* and *Table 10*). The same holds when non-orphan genes are considered separately (*Table 11* and *Table 12*). On the other hand, the pattern is opposite for the orphan genes, namely the proportion of named genes is not significantly different for slow and fast evolving orphan genes for all thresholds. Similarly, there is significant rank correlation between naming and evolutionary rate for the complete (*dN*: r_s = -0.293, $P << 0.001$; *dN/dS*: r_s = -0.248, $P << 0.001$) and the non-orphan sample (*dN*: r_s = -0.255, $P << 0.001$; *dN/dS*: r_s = -0.206, $P << 0.001$), but not for the orphan genes (*dS*: r_s = -0.055, P = 0.56; *dN/dS*: r_s = -0.039, P = 0.7). Taking into account that orphans are also under-represented among genetically studied genes, this suggests that most of them have less obvious phenotypes, even if some of them have rather low evolutionary rates.

Table 9. Number and proportion of named genes for different levels of non-synonymous substitution rate (*dN*) in the complete sample

	dN	**Not named**	**Named**
dN ($P = 1.7 \times 10^{-12}$)	≤0.007	141 (42.6 %)	190 (57.4 %)
	>0.007	278 (68.5 %)	128 (31.5 %)
	dN	**Not named**	**Named**
dN ($P = 6.6 \times 10^{-15}$)	≤0.01	165 (43.2 %)	217 (56.8 %)
	>0.01	254 (71.5 %)	101 (28.5 %)
	dN	**Not named**	**Named**
dN ($P = 1.2 \times 10^{-8}$)	≤0.03	295 (51.4 %)	279 (48.6 %)
	>0.03	124 (76.1 %)	39 (23.9 %)

Differences were tested using two-sided Fisher's exact test.

Table 10. Number and proportion of named genes for different levels of selective constraint (*dN/dS*) in the complete sample

	dN/dS	**Not named**	**Named**
dN/dS ($P = 1.2 \times 10^{-8}$)	≤0.03	151 (45.3 %)	182 (54.7 %)
	>0.03	268 (66.3 %)	136 (33.7 %)
	dN/dS	**Not named**	**Named**
dN/dS ($P = 9.5 \times 10^{-10}$)	≤0.06	216 (48.0 %)	234 (52.0 %)
	>0.06	203 (70.7 %)	84 (29.3 %)
	dN/dS	**Not named**	**Named**
dN/dS ($P = 6.8 \times 10^{-6}$)	≤0.1	295 (52.3 %)	269 (47.7 %)
	>0.1	124 (71.7 %)	49 (28.3 %)

Differences were tested using two-sided Fisher's exact test.

Table 11. Number and proportion of named genes for different levels of selective constraint (*dN/dS*) in the orphan and non-orphan sample

Genes	*dN/dS*	**Not named**	**Named**
Orphan (P = 0.241)	≤0.03	9 (69.2 %)	4 (30.8 %)
	>0.03	74 (84.1 %)	14 (15.9 %)
Non-orphan (P = 1.7 x 10^{-7})	≤0.03	142 (44.4 %)	178 (55.6 %)
	>0.03	194 (61.4 %)	122 (38.6 %)

Genes	*dN/dS*	**Not named**	**Named**
Orphan (P = 0.242)	≤0.06	20 (74.1 %)	7 (25.9 %)
	>0.06	63 (85.1 %)	11 (14.9 %)
Non-orphan (P = 3.6 x 10^{-6})	≤0.06	142 (46.3 %)	178 (53.7 %)
	>0.06	194 (65.7 %)	122 (34.3 %)

Genes	*dN/dS*	**Not named**	**Named**
Orphan (P = 1)	≤0.1	39 (81.3 %)	9 (18.8 %)
	>0.1	44 (83.0 %)	9 (17.0 %)
Non-orphan (P = 0.001)	≤0.1	256 (49.6 %)	260 (50.4 %)
	>0.1	80 (66.7 %)	40 (33.3 %)

Differences were tested using two-sided Fisher's exact test.

Table 12. Number and proportion of named genes for different levels of non-synonymous substitution rate (*dN*) in the orphan and non-orphan sample

Genes	*dN*	**Not named**	**Named**
Orphan (P = 0.124)	≤0.007	9 (64.3 %)	5 (35.7 %)
	>0.007	74 (85.1 %)	13 (14.9 %)
Non-orphan ($P = 2.4 \times 10^{-8}$)	≤0.007	132 (41.6 %)	185 (58.4 %)
	>0.007	204 (63.9 %)	115 (36.1 %)
Genes	*dN*	**Not named**	**Named**
Orphan (P = 0.155)	≤0.01	11 (68.8 %)	5 (31.3 %)
	>0.01	72 (84.7 %)	13 (15.3 %)
Non-orphan ($P = 6.6 \times 10^{-10}$)	≤0.01	154 (42.1 %)	212 (57.9 %)
	>0.01	182 (67.4 %)	88 (32.6 %)
Genes	*dN*	**Not named**	**Named**
Orphan (P = 1)	≤0.03	38 (82.6 %)	8 (17.4 %)
	>0.03	45 (81.8 %)	10 (18.2 %)
Non-orphan ($P = 2.7 \times 10^{-6}$)	≤0.1	257 (48.7 %)	271 (51.3 %)
	>0.1	79 (73.1 %)	29 (26.9 %)

Differences were tested using two-sided Fisher's exact test.

4.4 Genes with stage specific and non-restricted expression

4.4.1 Stage specific genes

To further examine the so far observed pattern of different evolutionary rates between genes expressed in embryos and adults (sections *4.2.1* and *4.3.2*), the data set from this study was compared against *D. melanogaster* EST information from public databases. As numerous *D. melanogaster* ESTs retrieved from adult and embryo cDNA libraries are available, it was possible to define adult and embryo specific EST sets among the genes studied here. The genes recovered from the *D. yakuba* adult library and their *D. melanogaster* orthologues were considered adult specific when no TBLASTN match among *D. melanogaster* embryo ESTs was found. In a similar way embryo specific genes were chosen, dividing the original *D. yakuba* non-redundant data set (n = 692) into the three classes: genes expressed only in the embryo (n = 59), genes expressed only in the adult (n = 117), and non-restricted genes that are expressed in both stages (n = 516).

4.4.2 Substitution rates of stage specific and non-restricted genes

The three expression classes (embryo, non-restricted and adult) show significant differences in non-synonymous substitution rates by one-way ANOVA ($F_{(2, 656)} = 49.180$, $P = 1.7 \times 10^{-13}$). The comparison shows that non-restricted genes have the lowest average substitution rate, followed by genes expressed only in the embryo and genes expressed only in the adult stage (*Figure 5*). All of these differences are significant in the *post hoc* pair wise comparisons at the 0.01 level (*Table 13*).

To distinguish specific differences between orphan and non-orphan genes these groups were analysed separately. When only orphan genes are considered (*Figure 5*) expression status has, as before, a significant effect on the non-synonymous substitution rates ($F_{(2, 97)} = 4.393$, $P = 0.015$). However, the average *dN* rates of orphans have a different pattern compared to the complete sample. *dN* is increasing from embryo, over non-restricted genes up to the adult class (*Figure 5*). Still, in the pair wise comparisons the only significant difference in the average *dN* rates is between embryo and adult class. It is interesting to note that the magnitude

of this difference (three times) is higher compared to the one in the complete data set (1.9 times) (*Table 13*). The separate analysis of non-orphan genes gives a pattern similar to the complete sample analysis ($F(2, 556) = 27.240$, $P = 5.2 \times 10^{-12}$), except that *dN* rate in the adult class, although higher, is not significantly different from the embryo class (*Figure 5* and *Table 13*).

The expression class has also a significant effect on the *dN*/*dS* ratio ($F(2, 656) = 35.573$, $P = 1.8 \times 10^{-13}$) (*Figure 6* and *Table 13*). As for the analysis of non-synonymous rates of the complete sample, it is clear that adult specific genes have the highest *dN*/*dS* ratio compared to embryo specific and non-restricted genes. When only orphans were considered, adult specific orphan genes have a higher *dN*/*dS* rate compared to embryo specific orphans, but the difference is not any more significant, probably due to the correlation between *dN* and *dS* (see section *4.3.1*). The non-orphan sample reveals higher *dN*/*dS* rates of stage specific genes compared to non-restricted genes, however no significant difference between embryo and adult class can be detected.

Taken together, these results show that average substitution rates are the highest for genes specifically expressed in adults compared to the embryo specific and non-restricted genes. Orphan genes are the major cause of this difference, as can be seen by the separate analysis of orphans and non-orphans. On the other hand, non-restricted genes have on average the lowest substitution rates, whereby non-restricted non-orphan genes contribute the most to this low average rate.

These results support the previous analysis (see section *4.3.2*), which suggested that the protein sequences of the embryo transcripts are evolving slower compared to the adult transcripts. The above analysis shows that the difference is even more pronounced when only genes having a stage specific expression are considered. For example, the adult specific transcripts have on average a 1.9 times higher non-synonymous substitution rate compared to embryo specific transcripts *(Table 13)*, while the previous analysis, where all transcripts found in the two libraries were taken into account, showed only a 1.5 times higher rate *(Table 8)*.

4.4.3 Protein length of stage specific and non-restricted genes

As *Figure 7* and *Table 13* show, the expression class has a significant effect on average protein length (one-way ANOVA, F (2, 689) = 14.229, $P = 8.8 \times 10^{-7}$). Non-restricted proteins have the longest protein sequence followed by embryo and adult specific proteins, but in pair wise comparisons, the only significant difference detected is the one between non-restricted and adult genes (*Table 13*). The separate analysis of orphan genes does not show significant influence of the expression class (F (2, 103) = 2.499, $P = 0.087$), while for non-orphan genes the pattern is the same as for the complete sample (F (2, 583) = 5.129, $P = 0.006$) (*Figure 7* and *Table 13*). However, the differences found in this analysis are less pronounced compared to obviously shorter average protein length in adults if all transcripts independent of stage specific expression are considered ($Laa_{ADULT} = 250 \pm 9.9$; $Laa_{EMBRYO} = 397 \pm 18.8$; t test = 7.792; $P = 2.2 \times 10^{-14}$).

4.4.4 Different expression of orphan genes in embryos and adults

The proportion of recovered orphan genes among adult transcripts is more than two times higher than among embryo transcripts (see section 4.2.1). Therefore, it is interesting to further analyse the use of orphan genes in embryos and adults when only genes with stage specific expression are taken into account. Interestingly, expression of stage specific genes between libraries is biased by itself (14.6 % embryo versus 30.8 % adult specific genes in corresponding libraries; $P = 7.1 \times 10^{-8}$, two-sided Fisher's exact test). Among these stage specific genes, 19 genes (25.4 %) in the embryo and 49 (43%) in the adult class were orphans ($P = 0.031$) (*Table 14)*. The lowest number of orphans was found among non-restricted genes 43 (7.2%) (*Table 14)*. Altogether, these results show that orphans and specifically expressed genes are used more often in the adult stage.

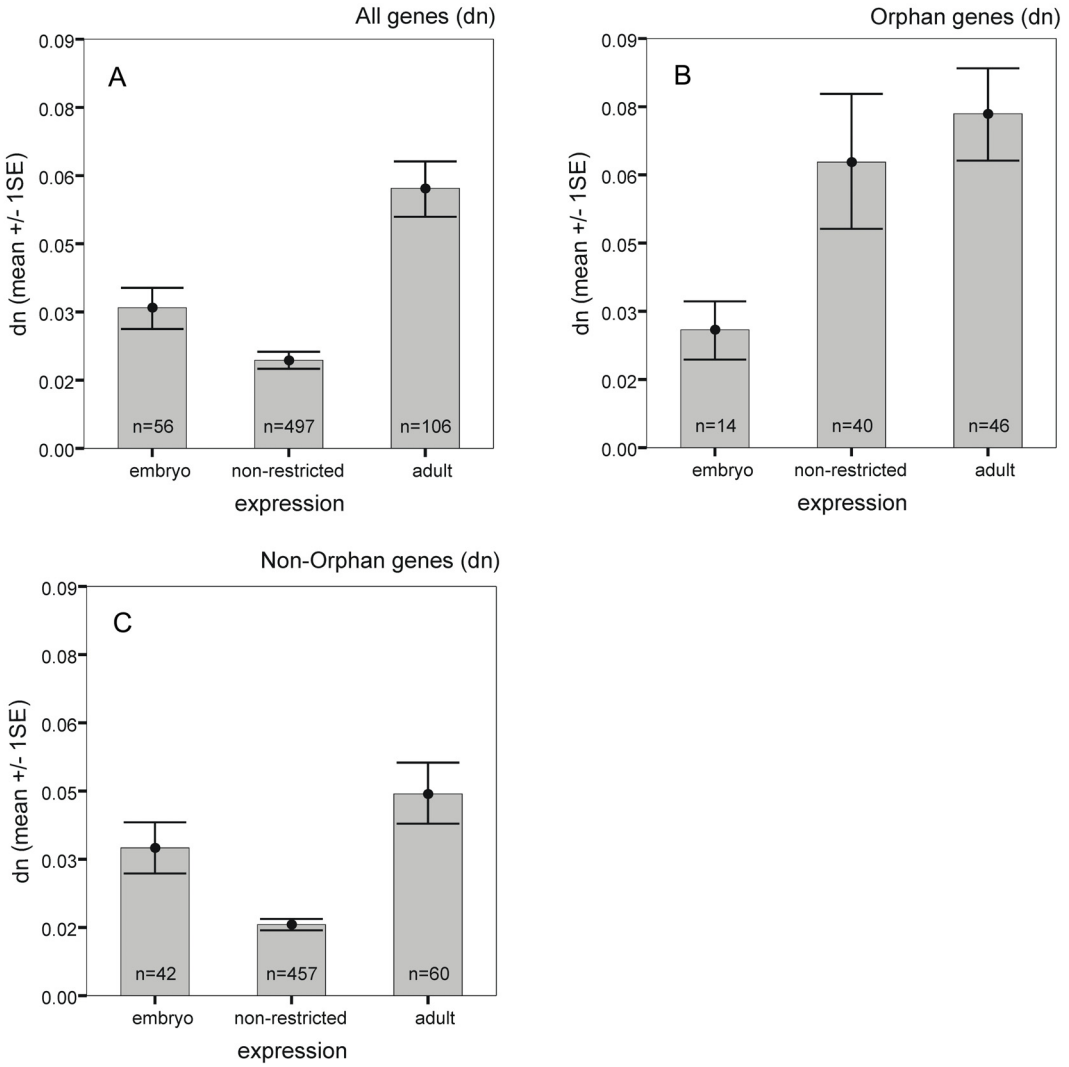

Figure 5. Mean of non-synonymous substitution rates (*dN*). Embryo, non-restricted and adult expression classes were analyzed. Error bars show one standard error of the mean. Differences between means were tested using the transformed *dN* data by one-way ANOVA (see Material and Methods). Hochberg's GT2 *post hoc* test was used in pair wise comparisons (**A**) Complete data set analysis. Expression class has a significant effect on *dN* ($F(2, 656) = 49.180$, $P = 1.7 \times 10^{-13}$) and accounts for 13.1% of the *dN* variance. In all pair wise comparisons difference between expression classes is significant at the 0.01 level (**B**) Orphan gene analysis. Expression class has a significant effect on *dN* ($F(2, 97) = 4.393$, $P = 0.015$) and accounts for 8.3% of the *dN* variance. Single significant difference in pair wise comparisons is between embryo and adult class ($P = 0.017$). (**C**) Non-orphan gene analysis. Expression class has a significant effect on *dN* ($F(2, 556) = 27.240$, $P = 5.2 \times 10^{-12}$) and accounts for 8.9 % of the *dN* variance. There are two significant differences in pair wise comparisons: between the embryo and non-restricted class ($P = 1.6 \times 10^{-4}$) and the adult and non-restricted class ($P = 4.3 \times 10^{-10}$).

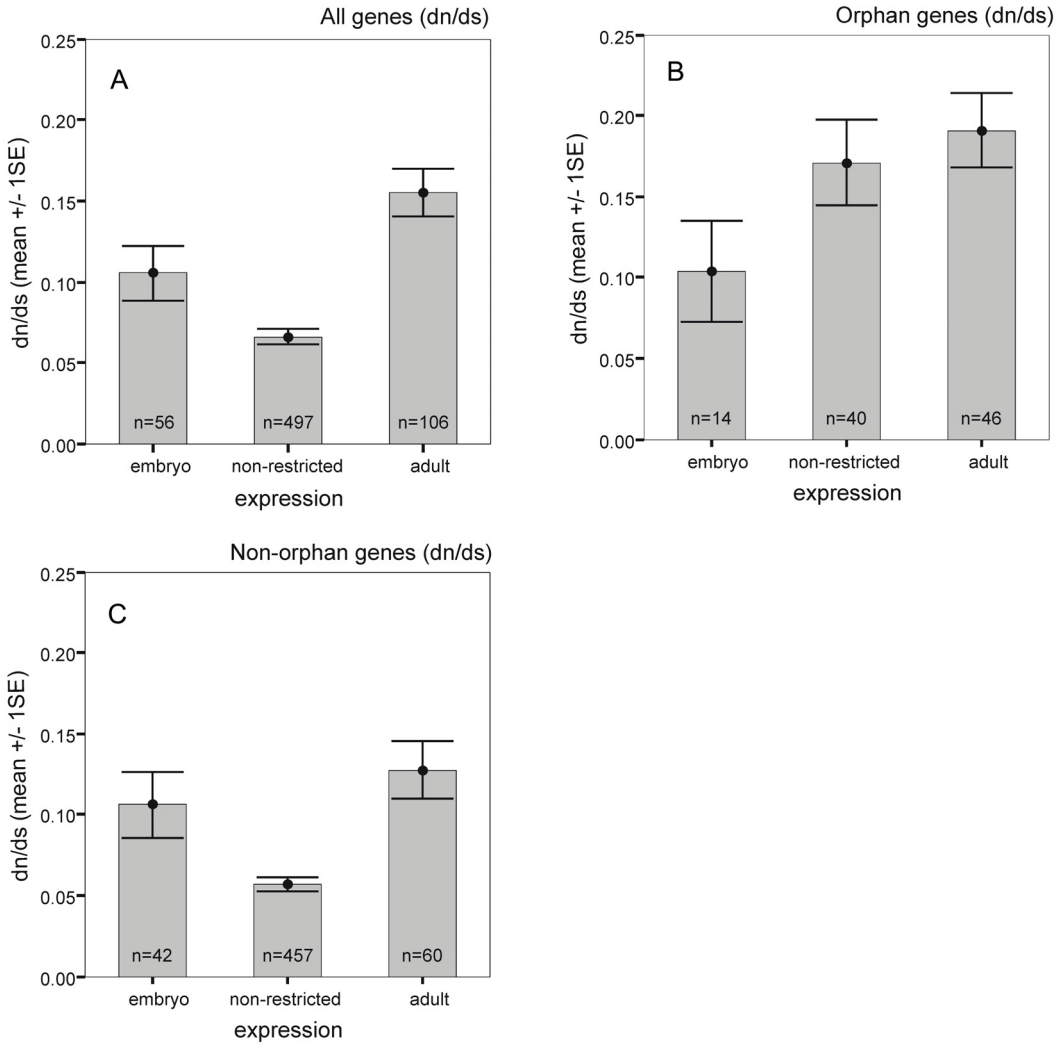

Figure 6. Mean of ratio of non-synonymous and synonymous substitution rates (*dN/dS*). Embryo, non-restricted and adult expression were analyzed. Error bars show one standard error of mean. Differences between means were tested on the transformed *dN/dS* data by one-way ANOVA (see Material and Methods). Hochberg's GT2 *post hoc* test was used in pair wise comparisons. (**A**) Complete data set analysis. Expression class has a significant effect on *dN/dS* ($F(2, 656) = 35.573$, $P = 1.8 \times 10^{-13}$) and accounts for 9.8% of the *dN/dS* variance. In all pair wise comparisons difference between expression classes is significant at the 0.05 level. (**B**) Orphan gene analysis. Expression class has no significant effect on *dN/dS* ($F(2, 97) = 2.896$, $P = 0.060$), nevertheless pattern is similar to *dN* differences for orphan genes (previous figure) (**C**) Non-orphan gene analysis. Expression class has a significant effect on *dN/dS* ($F(2, 556) = 18.113$, $P = 2.4 \times 10^{-8}$) and accounts for 6.1 % of the *dN/dS* variance. There are two significant differences in pair wise comparisons: between the embryo and non-restricted class ($P = 0.002$) and the adult and non-restricted class ($P = 6.0 \times 10^{-7}$).

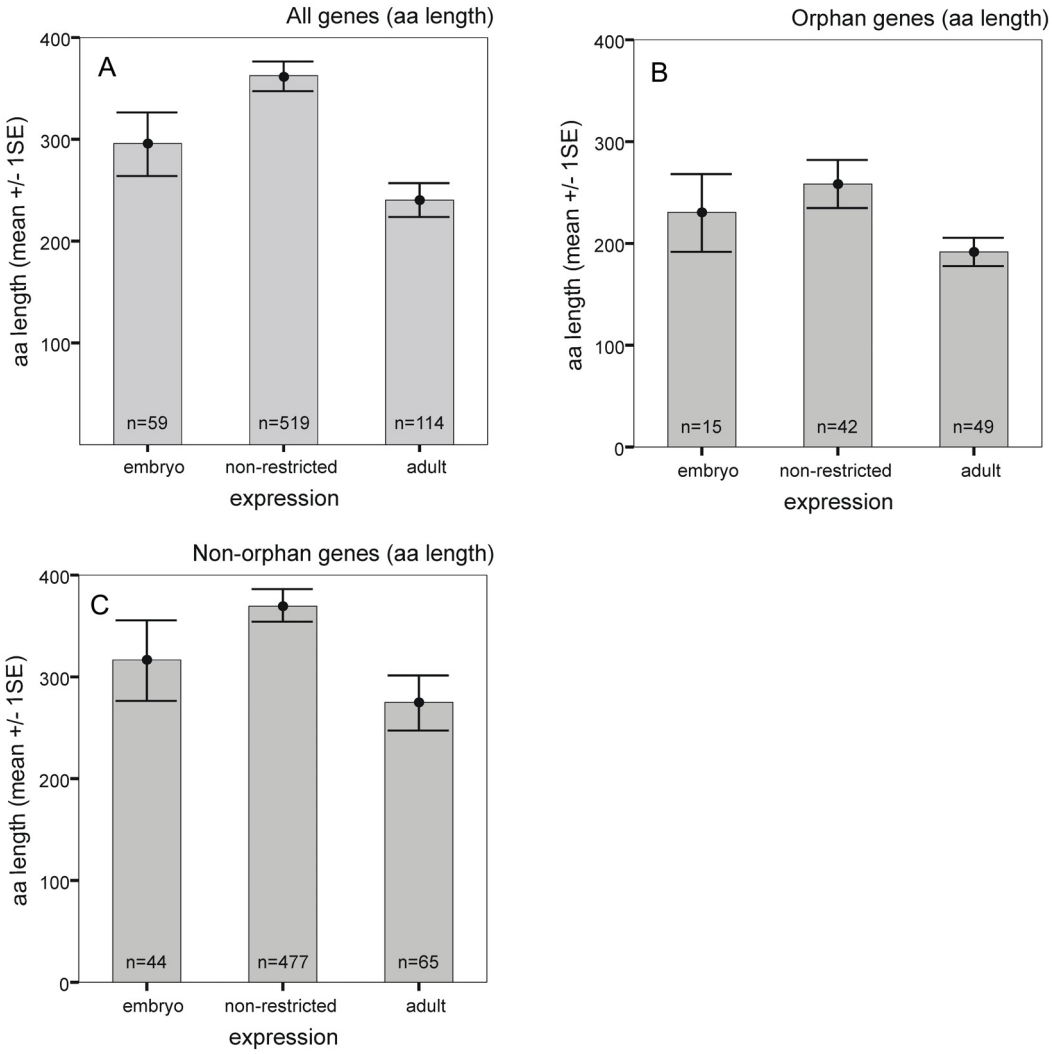

Figure 7. Mean of protein amino acid length. Embryo, non-restricted and adult expression classes were considered. Error bars show one standard error of mean. Differences between means were tested by one-way ANOVA (see Material and Methods). Hochberg's GT2 *post hoc* test was used in pair wise comparisons. (**A**) Complete data set analysis. Expression class has a significant effect on protein length (F (2, 689) = 14.229, $P = 8.8 \times 10^{-7}$) and accounts for 4% of the protein length variance. There is a single significant difference in pair wise comparisons between non-restricted and adult class ($P = 5.1 \times 10^{-7}$). (**B**) Orphan gene analysis. Expression class has no significant effect on protein length (F (2, 103) = 2.499, P = 0.087). (**C**) Non-orphan gene analysis. Expression class has a significant effect on protein length (F (2, 583) = 5.129, P = 0.006) and accounts for 1.7 % of the protein length variance. There is a single significant difference in pair wise comparisons between non-restricted and adult class (P = 0.005).

Table 13. One-way ANOVA *post hoc* pair wise comparison between means of the three expression classes.

Variable	Average values			Pair wise comparison		
	Exp. class	N	Mean ± 1SE	Exp.class	*P* value	Ratio
dN						
	Embryo	56	0.031 ± 0.0045	- Non-res.	0.005	1.6 (e/nr)
Complete sample	Non-res.	494	0.019 ± 0.0016	- Adult	0	3.1 (a/nr)
	Adult	109	0.058 ± 0.0064	- Embryo	0.001	1.9 (a/e)
	Embryo	14	0.026 ± 0.0063	- Non-res.	NS	
Orphan genes	Non-res.	39	0.056 ± 0.0136	- Adult	NS	
	Adult	47	0.079 ± 0.0113	- Embryo	0.017	3.0 (a/e)
	Embryo	42	0.033 ± 0.0057	- Non-res.	1.6×10^{-4}	2.1 (e/nr)
Non-orphan genes	Non-res.	455	0.016 ± 0.0012	- Adult	4.3×10^{-4}	2.7 (a/nr)
	Adult	62	0.043 ± 0.0067	- Embryo	NS	
dS						
	Embryo	56	0.326 ± 0.0202	- Non-res.	0.01	1.1 (a/e)
Complete sample	Non-res.	494	0.269 ± 0.0061	- Adult	8.4×10^{-7}	1.3 (a/nr)
	Adult	109	0.344 ± 0.0130	- Embryo	NS	
	Embryo	14	0.275 ± 0.0222	- Non-res.	NS	
Orphan genes	Non-res.	39	0.315 ± 0.0219	- Adult	NS	
	Adult	47	0.370 ± 0.0188	- Embryo	0.05	1.3 (a/e)
	Embryo	42	0.343 ± 0.0256	- Non-res.	0.002	1.3 (e/nr)
Non-orphan genes	Non-res.	455	0.265 ± 0.0064	- Adult	0.004	1.2 (a/nr)
	Adult	62	0.325 ± 0.0176	- Embryo	NS	
dN/dS						
	Embryo	56	0.106 ± 0.0170	- Non-res.	0.008	1.6 (e/nr)
Complete sample	Non-res.	494	0.065 ± 0.0045	- Adult	0	2.4 (a/nr)
	Adult	109	0.156 ± 0.0147	- Embryo	0.022	1.5 (a/e)
	Embryo	14	0.104 ± 0.0316	- Non-res.	NS	
Orphan genes	Non-res.	39	0.161 ± 0.0254	- Adult	NS	
	Adult	47	0.199 ± 0.0257	- Embryo	NS	
	Embryo	42	0.106 ± 0.0202	- Non-res.	0.002	1.9 (e/nr)
Non-orphan genes	Non-res.	455	0.057 ± 0.0042	- Adult	6.0×10^{-7}	2.2 (a/nr)
	Adult	62	0.124 ± 0.0177	- Embryo	NS	1.2 (a/e)
Protein length (aa)						
	Embryo	59	295 ± 31	- Non-res.	NS	
Complete sample	Non-res.	516	363 ± 15	- Adult	5.1×10^{-7}	1.5 (nr/a)
	Adult	117	238 ± 17	- Embryo	NS	
	Embryo	15	231 ± 38	- Non-res.	NS	
Orphan genes	Non-res.	41	262 ± 24	- Adult	NS	
	Adult	50	192 ± 14	- Embryo	NS	
	Embryo	44	317 ± 40	- Non-res.	NS	
Non-orphan genes	Non-res.	475	371 ± 16	- Adult	0.005	1.4 (nr/a)
	Adult	67	272 ± 26	- Embryo	NS	

Hochberg's GT2 *post hoc* test was used in pair wise comparisons. Note that the sample size is smaller in the analysis of *dN*, *dS* and *dN/dS* compared to the analysis of protein length. The reason is that substitution rates were not calculated for the genes having very short cDNA lengths. Letters in parenthesis designate expression classes used to calculate the ratio (a – adult, e - embryo, nr – non-restricted).

Table 14. Differences in the number of orphan and non-orphan genes between three expression classes (pair wise comparisons)

Expression	**Genes**	
	Orphan	Non-orphan
Embryo	15 (25.4 %)	44 (74.6 %)
Adult	49 (43.0 %)	65 (57.0 %)
$P = 0.031$		

Expression	**Genes**	
	Orphan	Non-orphan
Embryo	15 (25.4 %)	44 (74.6 %)
Non-modulated	43 (7.2 %)	557 (92.8 %)
$P = 5.1 \times 10^{-5}$		

Expression	**Genes**	
	Orphan	Non-orphan
Non-modulated	43 (7.2 %)	557 (92.8 %)
Adult	49 (43.0 %)	65 (57.0 %)
$P = 1.1 \times 10^{-19}$		

Differences were tested using two-sided Fisher's exact test.

4.5 Functional patterns of previously characterised orphan genes

To trace potential functional roles of orphan genes, it is possible to group previously studied orphan genes using their molecular function, biological process or cellular localization through the Gene Ontology (GO) database assignment (Ashburner et al., 2000). The controlled vocabulary of the Gene Ontology database allows statistical analysis of such data sets (Castillo-Davis and Hartl, 2003). With a view to find common functional patterns, the orphan genes obtained in the whole genome scan (section *4.1.1*) using BLAST E-value cutoff of 10^{-3} were tested for over-representation of particular GO terms compared to the complete genome of *D. melanogaster.* The statistical comparison was done using hypergeometric distribution implemented in GeneMerge (see materials and methods, section *6.4*). Even though only a small proportion of genes in the orphan sample has functional information (4.7% in the biological process and 6.8% in the molecular function section) some conclusion about functions and processes where orphans are prevalent can be made.

Table 15 and *Table 16* summarise the results. Previously characterised orphan genes are obviously over-represented among genes involved in olfaction, hormonal activity, puparial adhesion and egg membrane organization, all functions which one would expect to be important for specific ecological adaptations. It is also easy to notice, especially in biological process analysis, that orphan genes are over-represented in the pathways involved in communication of the organism with the environment (*Table 15*).

Table 15. Rank scores for over-representation of Biological Process terms in the orphan gene sample compared to the complete *D. melanogaster* genome

GO Biol. Process term	Genome frac.	Orphan frac.	Raw e-score	e-score	Description
GO:0007608	0.0045	0.0152	8.21E-20	1.96E-17	Olfaction
GO:0007606	0.0062	0.0171	1.43E-15	3.41E-13	Chemosensory perception
GO:0009593	0.0062	0.0171	1.43E-15	3.41E-13	Perception of chemical substance
GO:0007600	0.0074	0.0175	7.23E-12	1.72E-09	Sensory perception
GO:0009582	0.0083	0.0178	8.32E-10	1.98E-07	Perception of abiotic stimulus
GO:0007594	0.0007	0.0029	2.70E-06	0.0006	Puparial adhesion
GO:0009628	0.0109	0.0178	4.28E-05	0.0102	Response to abiotic stimulus
GO:0009581	0.0114	0.0181	7.99E-05	0.0190	Perception of external stimulus
GO:0007304	0.0011	0.0032	0.0002	0.0579	Eggshell formation
GO:0007591	0.0010	0.0029	0.0007	0.1742	Molting cycle (sensu Insecta)
GO:0007582	0.0015	0.0036	0.0016	0.3900	Physiological processes
GO:0007305	0.0003	0.0013	0.0034	0.7994	Vitelline membrane formation
GO:0007306	0.0006	0.0019	0.0034	0.8176	Insect chorion formation

Biological Process (BP) terms are from the Gene Ontology database. Only terms with raw e-scores below 0.05 are shown. Scores are based on hypergeometric distribution. Raw e-scores were calculated with Bonferroni correction excluding singleton terms, while e-scores were calculated with Bonferroni correction for all terms. Genome fraction represents the proportion of the genes in the complete *D. melanogaster* genome (12843 genes) having a corresponding BP term assignment. The orphan fraction represents the proportion of the orphan genes in the orphan sample (3039 genes) having a corresponding BP Function term assignment. There are 257 BP terms among orphan genes and 146 orphan genes have BP information.

Table 16. Rank scores for over-representation of Molecular Function terms in the orphan gene sample compared to the complete *D. melanogaster* genome

GO Mol. Func. term	Genome frac.	Orphan frac.	Raw e-score	e-score	Description
GO:0004984	0.0040	0.0152	4.60E-24	1.19E-21	Olfactory receptor activity
GO:0008141	0.0007	0.0029	2.70E-06	0.0007	Puparial glue (sensu Diptera)
GO:0001584	0.0115	0.0175	0.0005	0.1193	Rhodopsin-like receptor activity
GO:0005179	0.0023	0.0052	0.0005	0.1334	Hormone activity
GO:0005180	0.0023	0.0052	0.0005	0.1334	Peptide hormone
GO:0008316	0.0003	0.0013	0.0034	0.8666	Structural constituent of vitelline membrane (sensu Insecta)
GO:0005213	0.0006	0.0019	0.0034	0.8863	Structural constituent of chorion (sensu Insecta)
GO:0004930	0.0136	0.0184	0.0065	1	G-protein coupled receptor activity
GO:0008613	0.0002	0.0010	0.0140	1	Diuretic hormone activity
GO:0005549	0.0018	0.0032	0.0319	1	Odorant binding activity
GO:0005184	0.0013	0.0026	0.0326	1	Neuropeptide hormone activity

Molecular Function terms are from the Gene Ontology database. Only terms with raw e-scores below 0.05 are shown. Scores are based on hypergeometric distribution. Raw e-scores were calculated with Bonferroni correction excluding singleton terms, while e-scores were calculated with Bonferroni correction for all terms. The genome fraction column represents the proportion of the genes in the complete *D. melanogaster* genome (12843 genes) that have a corresponding Molecular Function term assignment. The orphan fraction column represents the proportion of the genes in the orphan sample (3039 genes) that have a corresponding Molecular Function term assignment. Altogether there are 269 Molecular Function terms among orphan genes and 213 orphan genes have Molecular Function information.

4.6 Spatially restricted expression of orphan genes in Drosophila embryo

It was shown in mammals that genes with localised and tissue specific expression have increased evolutionary rates (Duret and Mouchiroud, 2000). As orphans have increased evolutionary rates also, it was appealing to test if their expression is localised. Expression patterns of all orphan genes recovered from the embryo library were analysed by whole mount in situ hybridisation. Expression was classified as specific if any kind of spatially restricted expression was observed. The general information about expression patterns is summarized in *Table 18.* A random sample of expression patterns from the same cDNA library obtained previously (Schmid, 1996) was statistically compared to the sample of embryo orphans *Table 17*. The result shows that expression of embryo orphans is more often spatially restricted compared to the random sample of genes suggesting that they act more often in a localised rather than ubiquitous manner.

Table 17. Comparison of expression patterns between random sample and orphan genes from *Drosophila yakuba*

Expression	**Random sample**	**Orphans**
Spatially restricted	29	22
Homogenous	76	12
Total	105	34

G = 14.33 (Williams's correction), $P < 0.001$

Table 18. Expression and substitution rates of embryo orphans

Appendix ID	Name (*D. melanogaster* orthologue)	Expression	*dN/dS*	*dN*	*dS*
4	CG18111	specific	0.0543	0.0316	0.5823
17	CG13741	specific	0.5769	0.2377	0.4121
26	mael, CG11254	specific	0.3362	0.1041	0.3095
32	CG3227	specific	0.1533	0.0441	0.2879
46	CG13512	unspecific	0.4441	0.1314	0.2958
62	CG11051	specific	0.7453	0.4926	0.661
66	GATAd ,CG5034	unspecific			
81	CG4440	specific	0.0782	0.0195	0.2499
93	CG7543	specific			
97	CG13011	specifc	0.0089	0.0025	0.2775
99	CG15188	unspecific	0.001	0.0004	0.3626
110	CG10978	unspecific	0.0119	0.0047	0.3961
137	CG12487	specific	0.2499	0.0712	0.2851
139	CG15189	specific	0.1782	0.0352	0.1975
141	CG14112	specific	0.0486	0.0144	0.2959
159	Df31 ,anon1A4, l(2)k05815	specific	0.2705	0.0551	0.2036
216	CG6583	unspecific	0.0245	0.0066	0.2678
232	CG13878	specific			
233	CG11100	specific	0.0687	0.0172	0.2509
281	CG13339	specific	0.1408	0.0501	0.356
293	CG9795	unspecific	0.2483	0.0877	0.3532
302	Tom, anon-fast-evolving-1F6	specific	0.001	0.0001	0.1069
307	CG18145	unspecific	0.0847	0.0276	0.3263
308	CG14639	unspecific	0.077	0.0366	0.4761
313	CG2046	unspecific	0.1963	0.0608	0.31
324	BG:DS08249.4	specific	0.4322	0.0701	0.1622
327	CG10799	unspecific	0.574	0.3247	0.5657
337	EG:25E8.4	specific	0.0763	0.0193	0.2531
350	CG9188	unspecific	0.1985	0.0697	0.3511
370	CG13043		0.0259	0.0037	0.1443
378	CG6803	unspecific	0.0908	0.0068	0.0746
389	CG18178	specific	0.1155	0.0543	0.4706
391	CG1157	specific	0.0415	0.0178	0.4288
403	CG14915	specific	0.0963	0.0278	0.2884

5. Discussion

5.1 Evolutionary scenarios for the origin of orphan genes

5.1.1 Orphan genes are a reality

The definition of orphan genes is necessarily vague. It depends on the statistics of the probability cutoff calculation, the size of the database and the species representation in the database. An E-value of > 10^{-4} and an extra screening step against the InterPro domain database have been chosen to define the set of orphan genes among the *D. yakuba* cDNA sequences. These criteria are conservative although I would expect the results not to be very different if more relaxed criteria such as E-values > 10^{-6} (Lipman et al., 2002) would be used. Another question concerns the species representation that one should use for the exclusion criterion. Insects were taken as a group within which a match was allowed. This is rather arbitrary and is more dictated by the fact that there are only few EST or genomic sequences available from the nearest evolutionary relative of insects, the crustaceans (Friedrich and Tautz, 1995). The full genome sequence from another Dipteran insect, *Anopheles*, has recently become available (Holt et al., 2002). The *Anopheles* genome has been specifically searched with all orphan genes defined in this study and 56% of them had no corresponding match in *Anopheles*. Zdobnov et al. (2002) find that 18.6% of the *Drosophila* genes and 11.1% of the *Anopheles* genes are orphans that are only found in the respective species in a pairwise comparison, which roughly matches the figure in this study.

The main reason why insects as a whole were chosen as an exclusion criterion in database search was to make the results in this survey comparable to previous studies. Therefore it can be concluded that although the number of sequences in the databases have increased with exponential rates, it seems that the percentage of coding regions that show no similarity to previously sequenced genes is not getting smaller. It is therefore clear that orphan genes are a reality that needs to be explained.

5.1.2 Evolutionary scenarios

There are three possible reasons why a gene can be an orphan gene.

(i) The gene has newly evolved in a particular evolutionary lineage, either through a recombination of exons from other genes, or by a recruitment of a randomly occurring open reading frame. In the former case, it should show at least domain similarity to other genes and would therefore not be an orphan. The latter case would lead directly to an orphan, as a random ORF would not be expected to show similarity to known genes. On the other hand, random ORFs are unlikely to code for a useful protein domain. In fact, it seems likely that today's existing protein domains have evolved very early on from short peptides, under conditions which are not any more prevalent in today's organisms (Lupas et al., 2001).

(ii) The gene was an ancestrally shared gene, but was lost in most evolutionary lineages, giving the appearance of a lineage specific orphan gene. This explanation may well apply to some orphans. The different evolutionary lineages are currently not well represented in the database. A *Drosophila* gene that has no homologue in yeast, plants, nematodes and vertebrates may still be present for example in platyhelminths, annelids or cnidarians, in which case one would not call it an orphan. On the other hand, given the large number of orphans in any of the well analysed lineages, it seems almost impossible to picture an ancestor, which would have had all these genes.

(iii) The gene evolves so fast that a similarity cannot be traced after a certain evolutionary distance. That such fast evolving genes exist in *Drosophila* has been shown previously (Schmid and Tautz, 1997). They diverge with rates between 0.3 - 1% per million year, implying that it would not even be possible to trace them among all Diptera. On the other hand, the data presented here show that many orphan genes do not evolve fast, at least not in the *D. melanogaster* - *D. yakuba* comparison that has been chosen. In fact, some of them evolve so slow that they should be present in all organisms, if they would always have had this slow divergence rate.

5.2 A model for orphan evolution

5.2.1 The model

The considerations above show that a more complex scenario is required to explain the existence of orphan genes and their evolutionary patterns. A scheme is proposed that tries to integrate the general knowledge on the evolution of genes, as well as the new data that are presented here. The scheme starts with the assumption that a new gene is initially created through a duplication of an existing gene (*Figure 8*). Such a duplicated gene can either be lost, or can be recruited into an accessory or redundant function (Krakauer and Nowak, 1999; Lynch and Conery, 2000). Because of the relaxed selective constraint, it will go through a phase of fast evolution (Lynch and Conery, 2000), during which it may loose most or all of the sequence similarity to the "parent" gene. However, at a certain point during evolution, it might become integrated into a new pathway, because evolutionary novelties have arisen in the respective lineage. During the time of integration into the new pathway, one can expect that the gene goes first through a phase of fast adaptive evolution, which would make it even more different from its "parent" gene. But once it has reached a new optimal state, it will be under strong purifying selection, implying slow evolution from this point onwards (*Figure 8*).

5.2.2 Implications of the model

This scenario has several important implications, both for the evolutionary history, as well as for the possible function of orphan genes. Because an initial gene duplication is assumed that leads eventually to an orphan, more refined structure based methods for the analysis of protein similarities (Koretke et al., 2002) may eventually help to identify the gene from which the orphan was derived. In terms of function, this scenario suggests that orphans have only accessory functions during the phase where they evolve fast, and are involved in important, but lineage specific functions when they evolve slowly. This would explain why they are under-represented in genetic screens, because such functions are usually not assessed in genetical screens. If the presented scenario is right, it points immediately to a class of genes that should be particularly interesting for studying the genetics of evolutionary divergence, namely the very slow evolving orphan genes. They can be

viewed as signatures of genetic pathways that have been newly acquired in a particular lineage and that are of special importance for the respective lineage.

One of the previously annotated orphan genes that have been recovered among *D. yakuba* cDNAs, the *flightin* gene, is indeed an excellent candidate for a lineage specific adaptation. It has a *dN*/*dS* ratio of 0.015 and is thus among the group of highly conserved orphan genes. Its function was thoroughly studied in *Drosophila* (Vigoreaux et al., 1993; Vigoreaux et al., 1998; Reedy et al., 2000). Mutations have no effect on viability or fecundity, but have a specific effect on the ultrastructure and function of the flight muscle. It appears that the gene is specifically required to increase the frequency at which the maximum power of the flight muscle is delivered to the wing. This could be seen as a rather specific adaptation for Dipterans. Slow evolving orphan genes should therefore deserve special attention in the future, both with respect to their evolutionary divergence patterns as well as their genetic functions.

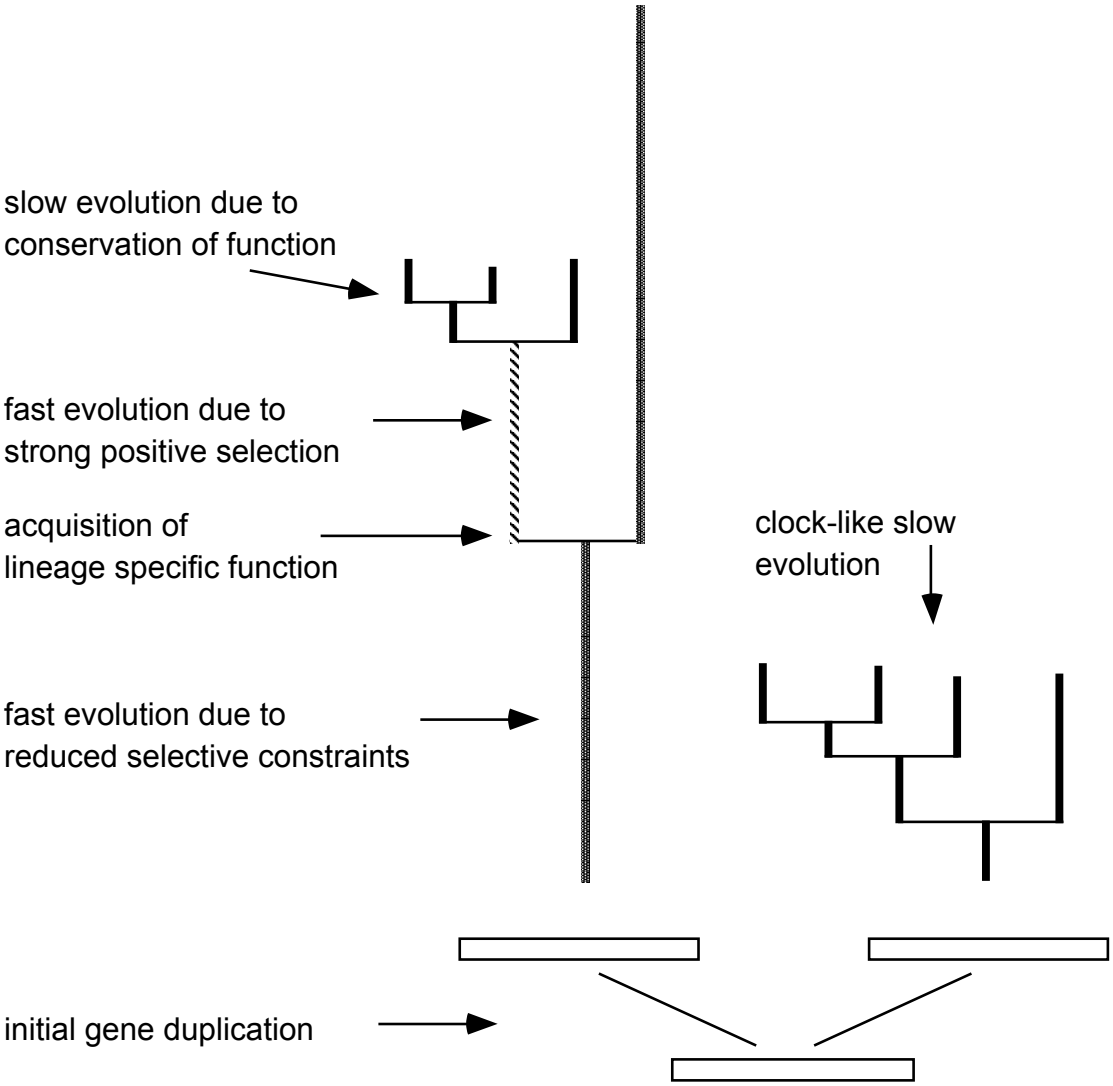

Figure 8. Model for the evolution of orphan genes. The model assumes an initial gene duplication, after which selective constraints in one of the duplicated genes become relaxed. This leads to a fast evolutionary divergence (left), indicated by a long branch in the topology. After a lineage splitting event, the gene may become integrated into a new central function in one lineage, but not in the other, where it continues to evolve fast because of reduced constraints. The new function in the first lineage implies that the gene would go through a phase of adaptive evolution, which would also result in a long branch, depending on how many aminoacid changes occurred during the phase of adaptation. But once an adaptive peak is reached, further evolution is slowed down and the branches become short. At this time, the gene may have lost all sequence similarity to its parent gene, but not necessarily its structural similarity. The parent gene (right topology) would undergo the same lineage splitting events, but would continue to have short branches in all lineages, because it has retained its original function. This model suggests the existence of three types of divergence modes: (1) fast divergence of genes which may, or may not yet have lost their sequence similarity to their parent gene, (2) fast divergence due to positive selection and (3) slow evolving orphan genes. Note that the model would apply in a similar way, if the initial gene would not have been created through a pure gene duplication, but through recruitment and recombination of exons from other genes or even after a gene has lost its original function in the context of a speciation event.

5.3 Differences between adults and embryos

5.3.1 Overall difference

Almost double as many orphan genes were found among the cDNAs of the adult library than in the embryo library. Because libraries used in this study were not normalized, this could have two reasons. Either orphan genes have a higher expression level in adults, which would result in a higher probability of recovery, or there are indeed more orphan genes specifically expressed in adults. Even without differentiating between these possibilities this means that orphan gene products are more abundant in adults. Because orphans evolve faster on average, this has the consequence that the average evolutionary rate of adult cDNAs is higher.

This finding can be compared with a previous study, which used crosshybridisation between RNAs of various *Drosophila* species (Powell et al., 1993). In this study it was found that RNAs from adults appeared to evolve faster when closely related species are compared, but not in the comparisons among more distantly related species. It was originally assumed that this could be due to differences in rates of silent substitutions in genes expressed at different stages of development as well as lineage specific shifts in codon usage (Powell et al., 1993). However, it would now seem possible that the differences in orphan gene expression could explain these results as well. Assuming that there is a higher expression of orphan genes in adults, and taking into account that these evolve faster, one would see more divergence among adult transcripts than among embryonic ones in comparisons among closely related species. On the other hand, the fast evolution of many orphan genes would lead to a complete loss of crosshybridisation between more distantly related species (Schmid and Tautz, 1997) and the signal that is measured by crosshybridisation would be mainly due to the more conserved non-orphan genes. Accordingly, a differential signal between embryonic and adult transcripts would vanish. Thus, we conclude that the crosshybridisation studies by Powell et al. (1993) are fully in line with our findings.

5.3.2 Stage specific genes

The analysis of the ESTs recovered in this study does not allow to differentiate between genes that are specifically expressed in embryos or adults. In fact, it can be expected that a substantial number of the genes that have been recovered are expressed at both stages. However, this means that the difference between embryos and adults should even be more pronounced when genes are compared that are specifically expressed in each stage. Indeed analysis of only the stage specific genes, which were selected based on abundance of their ESTs in public databases, confirms this reasoning.

A similar specific analysis was done by Castillo-Davis and Hartl (2002) for two nematode species. They selected the early and late expressed genes on the basis of quantitative expression data from microarray experiments in *Caenorhabditis elegans* and compared then the substitution rates with respect to the orthologue sequences retrieved from the *C. briggsae* genome project. In contrast to this work, they do not find any differences in non-synonymous substitution rates between early and late expressed genes. This result could have different reasons. *C. elegans* and *C. briggsae* are molecularly more divergent than *D. melanogaster* and *D. yakuba*, as can be inferred from the average synonymous substitution rates (average *dS C. elegans*/*C. briggsae* > 1; average *dS D.melanogster*/*D.yakuba* < 0.3). Thus, there might have been a bias against fast evolving genes in this study, because it focussed on a subset of unequivocally alignable orthologous genes.

But there might also be a biological reason for this difference between the two studies. The post-embryonic stages in nematodes are less divergent than in flies. The adult fly uses a completely different habitat than the embryos and larvae and it is likely to be subject to many different differential adaptations. If orphan genes are more often involved in such adaptations and if these evolve generally faster, one could expect a more pronounced difference in evolutionary rates between early and late stages in flies than in nematodes.

5.3.3 Developmental constraint

The higher proportion of orphan genes among adult RNAs can also be seen in the context of possible developmental constraints. Embryos go through a stage during early development which looks morphologically very similar even among very distantly related animal taxa and which has been called the phylotypic stage (Sander, 1983). It was proposed that the phylotypic stage represents a point in development where structural and network constraints place limits on morphological variability (Raff, 1996). Given that all developmental processes ultimately depend on the activity of a specific set of genes, some level of constraint on the variability of proteins expressed during the embryonic and phylotypic stage may be expected. If such a constraint exists, its signature may therefore be present in the coding and/or regulatory sequences. In the nematode study (Castillo-Davis and Hartl, 2002), the analysis of evolutionary rates did not confirm this expectation, although this may be partly due to a sampling bias (discussed above). However, the study did find differences with respect to the number of paralogous genes expressed in the different stages, which do suggest a stronger constraint on genes involved in embryonic development.

This study uncovered clear differences in evolutionary rates caused by a differential representation of orphan genes between the stages and in the number of stage specific orphan genes, but not with respect to the number of paralogous genes (not shown**)**. Intriguingly though, another clear difference between adult and embryonic transcripts was found that points also to a constraint. In this study, the proteins expressed in embryos are on average 150 amino acids longer than those expressed in the adult. This exceeds the difference that could be expected from the larger number of short orphan genes in adults *(Table 5)*. A possible explanation would be that proteins expressed in the embryo are involved in more protein-protein interactions, possibly to safeguard the developmental pathways. The analysis of yeast genes shows that proteins that are involved in more protein-protein interactions also tend to evolve more slowly (Fraser et al., 2002).

5.4 Proteins under adaptive pressure

Swanson et al. (2001) compared the sequences of ESTs from the male accessory gland of *Drosophila simulans* to their orthologues in its close relative *Drosophila melanogaster*. Among these, they found also many fast evolving genes and even several with an excess of non-synonymous versus synonymous substitutions. This demonstrates that genes, which can be expected to be under continuous pressure of new adaptations, such as accessory gland-specific seminal fluid proteins, are indeed subject to fast evolutionary divergence at the molecular level. This is also confirmed by the comparative systematic analysis of immunity-related genes between *Anopheles* and *Drosophila*, which show a marked deficit of orthologues and excessive gene expansions (Christophides et al., 2002). The overrepresentation of certain functions among orphans in *Drosophila* that were found in this study (see section 4.5) suggests also that these might play a role in specific ecological adaptations that change easily over time.

5.5 Conclusion

The role of orphan genes in the evolutionary process remains enigmatic. From the evidence discussed in this thesis, it would seem most likely that they are often involved in specific ecological adaptations. They might thus be the raw material for micro-evolutionary divergence, while macro-evolutionary differences are more likely to be caused by changes in regulatory interactions of highly conserved developmental genes (Carroll, 2001).

6. Materials and Methods

General molecular biology methods were performed, if not otherwise stated, as described in Sambrook et al. (1989). The following fly stocks were used in this study: *Drosophila yakuba* (wild type obtained from Prof. Dr. Michael Ashburner laboratory) and *Drosophila melanogaster* (Oregon R).

6.1 Database search

6.1.1 D. melanogaster proteome analysis

The *Drosophila melanogaster* proteome (release 2) comprising 14334 proteins was downloaded from Flybase. After removal of 38 5´-truncated proteins a BLASTP search was carried out against the non-redundant GenBank peptide database using the NCBI network BLAST client (blastcl3) and the following parameters: BLOSUM62 matrix, SEG filtering on and expectation cutoff of 10. After parsing the BLAST output using MuSeqBox (Xing and Brendel, 2001) installed locally, the resulting $2.1x10^6$ query/hit pairs were sorted into a Microsoft Access database. For each cutoff, the number of genes without match outside insects (orphans) and with match outside insects (non-orphans) was determined. The insect assignment was done according to the NCBI taxonomy rank classes. In addition, for each cutoff category the number of named genes was determined. For all genes retrieved from *D. yakuba* the full-length orthologue from *D. melanogaster* was used to search for protein domains via InterProScan v2.2 (Zdobnov and Apweiler, 2001) installed locally.

6.1.2 D. melanogaster EST database search

D. melanogaster EST data were downloaded from Flybase and NCBI EST database. As *D. melanogaster* ESTs are recovered form cDNA libraries constructed from different tissues and stages the data set was divided into embryo (99 617 ESTs) and adult sample (113 484 ESTs). The majority of these ESTs were derived from

normalized cDNA libraries, and thus the proportions of transcripts in this data set do not represent real expression levels. Nevertheless, the large number of the sequenced transcripts permits some conclusions about differences in the expression between stages, especially if data are analysed just by considering presence or absence of a particular transcript in a given library. *D. melanogaster* orthologues of cDNAs recovered from *D. yakuba* were compared against the set of adult and embryo ESTs using TBLASTN. In this analysis, a match having E-value less then 0.001 was considered significant.

6.2 cDNA libraries and sequencing

cDNA libraries were constructed from *D. yakuba* embryonic (0-14 hours) and adult (varying posteclosion times) stages using the Uni-ZAP XR Library Construction Kit (Stratagene) according to the instructions of the supplier.

6.2.1 D. yakuba 0-14 h embryo library

The *Drosophila yakuba* 0-14 h embryo library was constructed previously (Schmid, 1996). In this work, an aliquot of the primary embryo library containing 1.3 x 10^5 pfu was amplified once, yielding 3.24 x 10^{11} pfu. An aliquot ($1.3x10^7$ pfu) of the amplified library was mass excised to give clones in the pBluescript SK- plasmid vector (2.8 x 10^6 cfu), which were used for sequencing.

6.2.2 D. yakuba adult library

Total RNA was extracted from 1g of fresh material using a modified guanidine isothiocyanate procedure (Stratagene) as follows. Homogenisation of tissue and subsequent adding of sodium acetate was done according to the protocol of the manufacturer. After this steps one volume chloroform extraction was included. Chloroform and water phase were separated by centrifugation for 10 min on 6000xg at 4 °C. This step was added to improve separation of phenol and water phase in the subsequent step of the original protocol. Total RNA was dissolved in 2 ml of DEPC-treated water.

mRNA was isolated using the Poly(A) Quick mRNA Isolation Kit (Stratagene) according to the instructions of the supplier. cDNA was obtained from 3.3 μg D. yakuba mRNA. cDNA size fractions grater than 500 bp were selected for cloning. Cloning was done in 1 μg of Lambda ZAP II XR vector. 1μl of ligation reaction was packaged using Gigapack III Gold Packaging Extract. The primary library (4 x 10^6 pfu) was amplified yielding 9.24 x 10^{12} pfu. An aliquot (4 x 10^8 pfu) of the amplified library was mass excised with ExAssist helper phage (Stratagene) to give clones in pBluescript SK- plasmid vector (6 x 10^9 cfu).

6.2.3 Preparation of plasmid DNA and sequencing

Randomly picked colonies were grown in 1.2 ml 2xLB media in 96-deep-well blocks for 30 hours on 37 °C. Plasmids were isolated applying an alkaline lyses - diatomaceous earth miniprep protocol optimized for 96 well plates as follows. Cells were harvested by centrifugation at 3220 x g for 10 min. After removal of media, cells were resuspended in 200 μl of resuspension buffer (50 mM glucose, 25 mM Tris-HCl pH8.0, 10 mM EDTA pH 8.0). After cell lysis (200 μl of 0.2 M NaOH, 1% SDS) samples were neutralized by 200 μl of neutralizing buffer (3.6 M GHCl, 1.2 M K acetate pH 5.5) and centrifuged for 15 min at 3220 x g. The supernatant (500μl) of each was transferred to a new 96-deep-well block and mixed with 200 μl of diatomaceous earth suspension (16.8 g diatomaceous earth, 5 ml 1M Tris-HCl pH 8.0, 6 M Guanidine hydrochloride filled up to 100 ml). Samples were transferred in a 96-well filter plate (Whatman GF/B) and centrifuged for 5 min at 2500xg. Two washing steps with 500 μl washing buffer (20 mM Tris-HCl pH 8.0, 2 mM EDTA pH 8.0, 0.2 M NaCl, 50% ethanol) and one with 250 μl 80% ethanol were performed by centrifugation for 10 min at 2500 x g. Plasmid DNA bound to diatomaceous earth was eluted with 100 μl of 10 mM Tris pH 8.0 preheated to ~65 °C. After ~15 min of incubation on room temperature, plates were centrifuged at 2500xg for 10 min. Plasmids were sodium acetate – isopropanol precipitated and washed twice with 70 % ethanol. Samples were dissolved in 25 μl of 5 mM Tris pH 8.0. Integrity of plasmids and concentration were checked by agarose gel electrophoresis.

The clone inserts were fully sequenced directly from plasmids or from PCR products after amplification with standard T3/T7 and internal primers. The cDNA insert was cycle sequenced in 10 μl reaction volume using ~200 ng of plasmid

template, 2 or 4 μl of ET-Terminator mix (Amersham) and 5 pmol of primer. Cycle sequencing was done in 40 cycles [20s 95C°, 15s 50°C, 1 minute 60°C]. Sephadex G-50 columns were used for clean up of reactions samples which were then sealed and stored on –20 °C prior to sequencing injection. Sequencing reactions were run on a MegaBACE 1000 capillary sequencer (Amersham – Molecular Dynamics). For injection as well as for the run varying voltage and time were applied (from 40kVs up to 200kVs for injection, 9 kV-120min or 4 kV-400min for the run). To decrease injection failures, increase read length and improve sequencing quality several runs were performed per plate with different injection and run conditions.

6.3 Basecalling and contig assembly

Raw sequence data were basecalled applying the MegaBACE Sequence Analysis Software Version 2.1 (Cimarron 2.19.5 Slim Phredify basecaller). For each library all electropherogams were separately basecalled again using PHRED and assembly was done trough PHRAP (Ewing et al., 1998; Ewing and Green, 1998). Contigs and basecalling was inspected using CONSED (Gordon et al., 1998; Gordon et al., 2001). *D. yakuba* cDNA contigs and *D. melanogaster* ortholog CDS detected by BLAST were trimmed and adjusted in the same reading frame using BioEdit Version 5.0.9. Protein sequences were aligned in the frame using ClustalW (Thompson et al., 1994). Comparison of sequenced clones with *D. melanogaster* orthologues showed that the *D. yakuba* sequences were on average 62% fulllength for the embryo library and 77% for the adult library.

6.4 Evolutionary rates, sequence analysis and statistics

Nonsynonymous (*dN*) and synonymous (*dS*) rates were estimated by the maximum likelihood method implemented in PAML v3.1 package using the F3x4 codon frequency model (Yang, 1997). The null hypothesis that *dN* and *dS* are equal was tested comparing -2[log(L_0) log(L_1)] with the χ^2 distribution with 1 degree of freedom, where L_1 is log likelihood when *dN* and *dS* were estimated as two free parameters and L_0 is log likelihood having *dN* equal to *dS*. Codon usage bias measured as effective number of codons (ENC) or frequency of optimal codons (Fop), GC3 and GC content and amino acid length were calculated for *D. melanogaster* - *D. yakuba* orphan ortholog pairs using CodonW.

Statistical calculations were done by *SPSS for Windows Release 10.0.7.* Variables used in the statistical analysis, which were not normally distributed, were transformed using different power and log transformations (*Table 19*). Kolmogorov-Smirnov test of goodness-of-fit to the normal distribution were performed and the transformation, which gave the lowest Z, was used in further analysis, although qualitatively the same results were obtained without transformation in all tests. Means are reported with ± one standard error of the mean. Correlations were tested by Pearson's correlation coefficient (r) and for non-normally distributed variables by Spearman's rank correlation coefficient (r_s).

Over-representation of particular Gene Ontology GO term (Ashburner et al., 2000) in the orphan sample compared to the complete genome of *D. melanogaster* was tested using hypergeometric distribution implemented in GeneMerge software (Castillo-Davis and Hartl, 2003). GeneMerge algorithm gives two score values. Raw e-score is calculated without Bonferroni correction for singletons (terms which are present just once in a sample and thus can not be over-represented) while e-score takes into account this correction.

Table 19. Transformation of variables used in the statistical analysis

Variable	Transformation
dN	$(dN)^{1/2}+ (dN+1)^{1/2}$
dS	-
dN/dS	$(dN/dS)^{1/2}+ (dN/dS +1)^{1/2}$
ENC	$(ENC+0.5)^{0.6}$
Fop	-
GC	$(GC)^2$
GC3	$(GC3)^2$
N of exons	$(N+3/8)^{1/2}$
N of paralogues	-
Protein length (Laa)	log (Laa)

6.5 Expression analysis

Expression analysis of embryos was done by whole-mount in situ hybridisation (Tautz and Pfeifle, 1989; Lehmann and Tautz, 1994) using a RNA probe from *D. yakuba*. A gene was considered to be expressed specifically if any kind of spatially restricted expression pattern was detected.

7. Literature

Adams, M. D., Celniker, S. E., Holt, R. A., Evans, C. A., Gocayne, J. D., Amanatides, P. G., et al. (2000). The genome sequence of *Drosophila melanogaster*. Science *287*, 2185-95

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

Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res *25*, 3389-402.

Arabidopsis Genome Initiative (2000). Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. Nature *408*, 796-815.

Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., et al. (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium . Nat Genet *25*, 25-9.

Ashburner, M., Misra, S., Roote, J., Lewis, S. E., Blazej, R., Davis, T., Doyle, C., Galle, R., et al. (1999). An exploration of the sequence of a 2.9-Mb region of the genome of *Drosophila melanogaster*: the Adh region. Genetics *153*, 179-219

Blaxter, M. (1998). *Caenorhabditis elegans* is a nematode. Science *282*, 2041-6.

Carroll, S., Wetherbee, S., and Grenier, J. (2001). From DNA to Diversity (Blackwell Science).

Casari, G., De Daruvar, A., Sander, C., and Schneider, R. (1996). Bioinformatics and the discovery of gene function. Trends Genet *12*, 244-5.

Castillo-Davis, C.I. and Hartl, D.L. (2002). Genome Evolution and Developmental Constraint in *Caenorhabditis elegans*. Mol Biol Evol *19*, 728-35.

Castillo-Davis, C.I. and Hartl, D.L. (2003). GeneMerge-post-genomic analysis, data mining, and hypothesis testing. Bioinformatics *19*, 891-2.

Christophides, G.K., Zdobnov, E., Barillas-Mury, C., Birney, E., Blandin, S., Blass, C., Brey, P.T., et al. (2002). Immunity-related genes and gene families in *Anopheles gambiae*. Science *298*, 159-65.

Clark, T.A., Sugnet, C.W., and Ares, M. Jr (2002). Genomewide analysis of mRNA processing in yeast using splicing-specific microarrays. Science *296*, 907-10.

Comeron, J. M. and Kreitman, M. (1998). The correlation between synonymous and non-synonymous substitutions in *Drosophila*: mutation, selection or relaxed constraints? Genetics *150*, 767-75.

Dehal, P., Satou, Y., Campbell, R.K., Chapman, J., Degnan, B., De Tomaso, A., Davidson, B., et al. (2002). The draft genome of *Ciona intestinalis*: insights into chordate and vertebrate origins. Science *298*, 2157-67.

Dujon, B. (1996). The yeast genome project: what did we learn? Trends in Genetics *12*, 263-270

Dunn, K. A., Bielawski, J. P., and Yang, Z. (2001). Substitution rates in *Drosophila* nuclear genes: implications for translational selection. Genetics *157*, 295-305.

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

Ewing, B. and Green, P. (1998). Base-calling of automated sequencer traces using phred. II. Error probabilities. Genome Res *8*, 186-94.

Ewing, B., Hillier, L., Wendl, M. C., and Green, P. (1998). Base-calling of automated sequencer traces using phred. I. Accuracy assessment. Genome Res *8*, 175-85.

Fischer, D. and Eisenberg, D. (1999). Finding families for genomic ORFans. Bioinformatics *15*, 759-62

Fraser, H.B., Hirsh, A.E., Steinmetz, L.M., Scharfe, C., and Feldman, M.W. (2002). Evolutionary rate in the protein interaction network. Science *296*, 750-2.

Friedrich, M. and Tautz, D. (1995). Ribosomal DNA phylogeny of the major extant arthropod classes and the evolution of myriapods. Nature *376*, 165-7.

Gaillardin, C., Duchateau-Nguyen, G., Tekaia, F., Llorente, B., Casaregola, S., Toffano-Nioche, C., et al. (2000). Genomic exploration of the hemiascomycetous yeasts: 21. Comparative functional classification of genes. FEBS Lett *487*, 134-49.

Goffeau, A., Barrell, B. G., Bussey, H., Davis, R. W., Dujon, B., Feldmann, H., Galibert, F., et al. (1996). Life with 6000 genes. Science *274*, 546, 563-7

Gordon, D., Abajian, C., and Green, P. (1998). Consed: a graphical tool for sequence finishing. Genome Res *8*, 195-202.

Gordon, D., Desmarais, C., and Green, P. (2001). Automated finishing with autofinish. Genome Res *11*, 614-25.

Green, P., Lipman, D., Hillier, L., Waterston, R., States, D., and Claverie, J. M. (1993). Ancient conserved regions in new gene sequences and the protein databases. Science *259*, 1711-6.

Harrison, P.M., Kumar, A., Lang, N., Snyder, M., and Gerstein, M. (2002). A question of size: the eukaryotic proteome and the problems in defining it. Nucleic Acids Res *30*, 1083-90.

Himmelreich, R., Plagens, H., Hilbert, H., Reiner, B., and Herrmann, R. (1997).

Comparative analysis of the genomes of the bacteria Mycoplasma pneumoniae and Mycoplasma genitalium. Nucleic Acids Res *25*, 701-12

Hirsh, A. E. and Fraser, H. B. (2001). Protein dispensability and rate of evolution. Nature *411*, 1046-9.

Holt, R.A., Subramanian, G.M., Halpern, A., Sutton, G.G., Charlab, R., Nusskern, D.R., et al. (2002). The genome sequence of the malaria mosquito *Anopheles gambiae*. Science *298*, 129-49.

Hurst, L. D. and Smith, N. G. (1999). Do essential genes evolve slowly? Curr Biol *9*, 747-50.

Jordan, I.K., Kondrashov, F.A., Rogozin, I.B., Tatusov, R.L., Wolf, Y.I., and Koonin, E.V. (2001). Constant relative rate of protein evolution and detection of functional diversification among bacterial, archaeal and eukaryotic proteins. Genome Biol *2*, R53

Jordan, I.K., Rogozin, I.B., Wolf, Y.I., and Koonin, E.V. (2002a). Essential genes are more evolutionarily conserved than are nonessential genes in bacteria. Genome Res *12*, 962-8.

Jordan, I.K., Rogozin, I.B., Wolf, Y.I., and Koonin, E.V. (2002b). Microevolutionary genomics of bacteria. Theor Popul Biol *61*, 435-47.

Koretke, K.K., Russell, R.B., and Lupas, A.N. (2002). Fold recognition without folds. Protein Sci *11*, 1575-9.

Kowalczuk, M., Mackiewicz, P., Gierlik, A., Dudek, M. R., and Cebrat, S. (1999). Total number of coding open reading frames in the yeast genome. Yeast *15*, 1031-4.

Krakauer, D.C. and Nowak, M.A. (1999). Evolutionary preservation of redundant duplicated genes. Semin Cell Dev Biol *10*, 555-9.

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., et al. (2001). Initial sequencing and analysis of the human genome. Nature *409*, 860-921

Lehmann, R. and Tautz, D. (1994). In situ hybridization to RNA. Methods Cell Biol *44*, 575-98.

Lipman, David J., Souvorov, Alexander, Koonin, Eugene V., Panchenko, Anna R., and Tatusova, Tatiana A. (2002). The relationship of protein conservation and sequence length. BMC Evolutionary Biology *2*,

Lupas, A.N., Ponting, C.P., and Russell, R.B. (2001). On the evolution of protein folds: are similar motifs in different protein folds the result of convergence, insertion, or relics of an ancient peptide world? J Struct Biol *134*, 191-203.

Lynch, M. and Conery, J.S. (2000). The evolutionary fate and consequences of duplicate genes. Science *290*, 1151-5.

Mackiewicz, P., Kowalczuk, M., Gierlik, A., Dudek, M. R., and Cebrat, S. (1999). Origin and properties of non-coding ORFs in the yeast genome. Nucleic Acids

Res *27*, 3503-9.

Malpertuy, A., Tekaia, F., Casaregola, S., Aigle, M., Artiguenave, F., Blandin, G., Bolotin-Fukuhara, M., et al. (2000). Genomic exploration of the hemiascomycetous yeasts: 19. Ascomycetes- specific genes. FEBS Lett *487*, 113-21.

Oliver, S. G., van der Aart, Q. J., Agostoni-Carbone, M. L., Aigle, M., Alberghina, L., Alexandraki, D., et al. (1992). The complete DNA sequence of yeast chromosome III. Nature *357*, 38-46.

Ozier-Kalogeropoulos, O., Malpertuy, A., Boyer, J., Tekaia, F., and Dujon, B. (1998). Random exploration of the *Kluyveromyces lactis* genome and comparison with that of *Saccharomyces cerevisiae*. Nucleic Acids Res *26*, 5511-24

Parra, G., Agarwal, P., Abril, J.F., Wiehe, T., Fickett, J.W., and Guigo, R. (2003). Comparative gene prediction in human and mouse. Genome Res *13*, 108-17.

Phaff, H.J. (1998). Chemotaxonomy based on the polysaccharide composition of the cell walls and capsules. In The Yeast, a Taxonomic Study. C.P. Kurtzman and J.W. Fell, eds. (Amsterdam: Elsevier).

Powell, J. R., Caccone, A., Gleason, J. M., and Nigro, L. (1993). Rates of DNA evolution in *Drosophila* depend on function and developmental stage of expression. Genetics *133*, 291-8.

Raff, R.A. (1996). The shape of life. (The University of Chicago Press, Chicago, Ill.)

Reedy, M. C., Bullard, B., and Vigoreaux, J. O. (2000). Flightin is essential for thick filament assembly and sarcomere stability in *Drosophila* flight muscles. J Cell Biol *151*, 1483-500.

Rubin, G. M. (2001). The draft sequences. Comparing species. Nature *409*, 820-1

Rubin, G. M., Yandell, M. D., Wortman, J. R., Gabor Miklos, G. L., Nelson, C. R., Hariharan, I. K., et al. (2000). Comparative genomics of the eukaryotes. Science *287*, 2204-15

Sander, K. (1983). The evolution of patterning mechanisms: gleanings from insect embryogenesis and spermatogenesis. In Development and evolution: the sixth symposium of the British Society for Developmental Biology. B.C. Goodwin, N. Holder, and C.C. Wylie, eds. (Cambridge University Press), pp. 137-159.

Schmid, K.J. and Aquadro, C.F. (2001). The evolutionary analysis of "orphans" from the *Drosophila* genome identifies rapidly diverging and incorrectly annotated genes. Genetics *159*, 589-98.

Schmid, K. J., Nigro, L., Aquadro, C. F., and Tautz, D. (1999). Large number of replacement polymorphisms in rapidly evolving genes of *Drosophila*. Implications for genome-wide surveys of DNA polymorphism. Genetics *153*, 1717-29.

Schmid, K. J. and Tautz, D. (1997). A screen for fast evolving genes from

Drosophila. Proc Natl Acad Sci U S A *94*, 9746-50.

Schmid, K. Isolierung und Characterisierung von schnell evolvirenden Genen aus *Drosophila melanogaster*. (1996). München, Fakultät für Biologie, Ludwig-Maxililians-Universität München.

Shoemaker, D.D., Schadt, E.E., Armour, C.D., He, Y.D., Garrett-Engele, P., McDonagh, P.D., et al. (2001). Experimental annotation of the human genome using microarray technology. Nature *409*, 922-7.

Spang, R. and Vingron, M. (2001). Limits of homology detection by pairwise sequence comparison. Bioinformatics *17*, 338-42.

Swanson, W.J., Yang, Z., Wolfner, M.F., and Aquadro, C.F. (2001). Positive Darwinian selection drives the evolution of several female reproductive proteins in mammals. Proc Natl Acad Sci U S A *98*, 2509-14.

Tautz, D. and Pfeifle, C. (1989). A non-radioactive in situ hybridization method for the localization of specific RNAs in *Drosophila* embryos reveals translational control of the segmentation gene hunchback. Chromosoma *98*, 81-5.

Tautz, D. and Schmid, K. J. (1998). From genes to individuals: developmental genes and the generation of the phenotype. Philos Trans R Soc Lond B Biol Sci *353*, 231-40.

The C. elegans Sequencing Consortium. (1998). Genome sequence of the nematode C. elegans: a platform for investigating biology. Science *282*, 2012-8

Thompson, J. D., Higgins, D. G., and Gibson, T. J. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res *22*, 4673-80.

Velculescu, V.E., Zhang, L., Zhou, W., Vogelstein, J., Basrai, M.A., Bassett, D.E. Jr, Hieter, P., Vogelstein, B., and Kinzler, K.W. (1997). Characterization of the yeast transcriptome. Cell *88*, 243-51.

Vigoreaux, J. O., Hernandez, C., Moore, J., Ayer, G., and Maughan, D. (1998). A genetic deficiency that spans the flightin gene of *Drosophila melanogaster* affects the ultrastructure and function of the flight muscles. J Exp Biol *201* , 2033-44.

Vigoreaux, J. O., Saide, J. D., Valgeirsdottir, K., and Pardue, M. L. (1993). Flightin, a novel myofibrillar protein of *Drosophila* stretch-activated muscles. J Cell Biol *121*, 587-98.

Wang, Z. G., Schmid, K. J., and Ackerman, S. H. (1999). The *Drosophila* gene 2A5 complements the defect in mitochondrial F1-ATPase assembly in yeast lacking the molecular chaperone Atp11p. FEBS Lett *452*, 305-8

Waterston, R.H., Lindblad-Toh, K., Birney, E., Rogers, J., Abril, J.F., Agarwal, P., Agarwala, R., et al. (2002). Initial sequencing and comparative analysis of the mouse genome. Nature *420*, 520-62.

Wolfe, K.H. and Sharp, P.M. (1993). Mammalian gene evolution: nucleotide sequence divergence between mouse and rat. J Mol Evol *37*, 441-56.

Wood, V., Gwilliam, R., Rajandream, M.A., Lyne, M., Lyne, R., Stewart, A., Sgouros, J., et al. (2002). The genome sequence of *Schizosaccharomyces pombe*. Nature *415*, 871-80.

Xing, L. and Brendel, V. (2001). Multi-query sequence BLAST output examination with MuSeqBox. Bioinformatics *17*, 744-5.

Xuan, Z., Wang, J., and Zhang, M.Q. (2003). Computational comparison of two mouse draft genomes and the human golden path. Genome Biol *4*, R1

Yang, Z. (1997). PAML: a program package for phylogenetic analysis by maximum likelihood. Comput Appl Biosci *13*, 555-6.

Yu, J., Hu, S., Wang, J., Wong, G.K., Li, S., Liu, B., Deng, Y., Dai, L., Zhou, Y., Zhang, X., et al. (2002). A draft sequence of the rice genome (*Oryza sativa L. ssp. indica*). Science *296*, 79-92.

Zdobnov, E. M. and Apweiler, R. (2001). InterProScan - an integration platform for the signature-recognition methods in InterPro. Bioinformatics *17*, 847-8.

Zdobnov, E.M., von Mering, C., Letunic, I., Torrents, D., Suyama, M., Copley, R.R., Christophides, G.K., et al. (2002). Comparative genome and proteome analysis of *Anopheles gambiae* and *Drosophila melanogaster*. Science *298*, 149-59.

Zhang, C. T. and Wang, J. (2000). Recognition of protein coding genes in the yeast genome at better than 95% accuracy based on the Z curve. Nucleic Acids Res *28*, 2804-14.

Zhang, M.Q. (2002). Computational prediction of eukaryotic protein-coding genes. Nat Rev Genet *3*, 698-709.

8. Appendix

8.1 Appendix A – Overview of D. yakuba cDNA clones

No.	Library	Gene name	Duplicates	Dup. excl.	Name	Orphan	*Anopheles*	Expression	Paralogues	Exons	*dN*	*dS*	*dN/dS*	H_0: *dN/dS*=1
1	Adult	troponin I, wupA	No	-	Real name	No	Yes	Non-modulated	0	6	0.0035	0.0272	0.1301	
2	Adult	guf	No	-	Real name	No	Yes	Non-modulated	0	7	0.0111	0.1004	0.1102	
3	Adult	CG18180	No	-	Identifier	No	Yes	Non-modulated	123	1	0.0308	0.4039	0.0762	
4	Adult	Ser4	No	-	Real name	No	Yes	Non-modulated	136	1	0.0282	0.1659	0.1698	
5	Adult	epsilonTry1	No	-	Real name	No	Yes	Non-modulated	176	1	0.0174	0.2518	0.0689	
6	Adult	CG7768	No	-	Identifier	No	Yes	Non-modulated	15	3	0.0032	0.4760	0.0067	
7	Adult	CG7170	No	-	Identifier	No	Yes	Non-modulated	111	1	0.0205	0.4136	0.0497	
8	Adult	CG2229	No	-	Identifier	No	Yes	Non-modulated	149	1	0.0161	0.3958	0.0406	
9	Adult	CG6467	No	-	Identifier	No	Yes	Non-modulated	165	2	0.1035	0.4141	0.2499	
10	Adult	CG9344	No	-	Identifier	No	Yes	Non-modulated	1	2	0.0003	0.2847	0.0010	
11	Adult	RpS3	Yes	-	Real name	No	Yes	Non-modulated	0	1	0.0059	0.1449	0.0410	
12	Adult	CG5770	No	-	Identifier	Yes	No	Adult	7	3	0.0588	0.3640	0.1614	
13	Adult	RpS12	Yes	Yes	Real name	No	Yes	Non-modulated	0	3	0.0001	0.0983	0.0010	
14	Adult	CG18001	Yes	-	Identifier	No	Yes	Non-modulated	0	1	0.0004	0.3571	0.0010	
15	Adult	CG5390	No	-	Identifier	No	Yes	Non-modulated	157	4	0.2004	0.5724	0.3501	
16	Adult	oho23B	Yes	-	Real name	No	Yes	Non-modulated	0	4	0.0002	0.1527	0.0010	
17	Adult	CG10423	Yes	-	Identifier	No	Yes	Non-modulated	0	3	0.0001	0.0922	0.0010	
18	Adult	CG2099	Yes	Yes	Identifier	No	Yes	Non-modulated	0	4	0.0249	0.2908	0.0858	
19	Adult	yip7	No	-	Real name	No	Yes	Non-modulated	159	2	0.0727	0.3367	0.2159	
20	Adult	sqh	Yes	-	Real name	No	Yes	Non-modulated	15	3	0.0026	0.2378	0.0110	
21	Adult	Pglym	No	-	Real name	No	Yes	Non-modulated	2	1	0.0072	0.2888	0.0248	
22	Adult	CG14500	No	-	Identifier	No	Yes	Non-modulated	1	1	0.2005	0.4551	0.4406	
23	Adult	MtnA	No	-	Real name	No	No	Adult	0	2	-	-	-	
24	Adult	CG1124	No	-	Identifier	Yes	Yes	Non-modulated	13	4	0.0174	0.3195	0.0543	
25	Adult	Mlc2	Yes	Yes	Real name	No	Yes	Non-modulated	7	3	0.0000	0.0287	0.0010	
26	Adult	Arr1	No	-	Real name	No	Yes	Adult	3	3	0.0090	0.1167	0.0775	
27	Adult	Uev1A, CG10640	No	-	Real name	No	Yes	Non-modulated	0	4	0.0001	0.0912	0.0010	
28	Adult	Transaldolase, CG2827	No	-	Real name	No	Yes	Non-modulated	0	3	0.0002	0.2477	0.0010	

No.	Library	Gene name	Duplicates	Dup. excl.	Name	Orphan	*Anopheles*	Expression	Paralogues	Exons	*dN*	*dS*	*dN/dS*	H_0: *dN/dS*=1
29	Adult	Vha16	No	-	Real name	No	Yes	Non-modulated	5	4	0.0000	0.0477	0.0010	
30	Adult	CG7118	No	-	Identifier	No	Yes	Non-modulated	113	1	0.0087	0.2932	0.0297	
31	Adult	CG18081	No	-	Identifier	No	Yes	Non-modulated	1	2	0.0053	0.4120	0.0128	
32	Adult	mtacp1	No	-	Real name	No	Yes	Non-modulated	0	4	0.0221	0.0439	0.5032	H0 not rejected
33	Adult	Ser99Dc	No	-	Real name	No	Yes	Non-modulated	135	2	0.0190	0.2902	0.0655	
34	Adult	RpS20	Yes	-	Real name	No	Yes	Non-modulated	0	4	0.0032	0.2128	0.0151	
35	Adult	CG4692	No	-	Identifier	No	Yes	Non-modulated	0	3	0.0002	0.1892	0.0010	
36	Adult	CG6666	No	-	Identifier	No	Yes	Non-modulated	1	2	0.0029	0.2708	0.0106	
37	Adult	CG8462	No	-	Identifier	No	Yes	Non-modulated	2	2	0.0755	0.2841	0.2658	
38	Adult	CG3203	Yes	Yes	Identifier	No	Yes	Non-modulated	0	3	0.0001	0.0765	0.0010	
39	Adult	CG10320	No	-	Identifier	No	Yes	Adult	0	2	0.0082	0.0956	0.0854	
40	Adult	CG12848	No	-	Identifier	No	Yes	Adult	0	1	0.0086	0.5103	0.0168	
41	Adult	CG9091	Yes	-	Identifier	No	Yes	Non-modulated	1	3	0.0001	0.1476	0.0010	
42	Adult	CG14933	No	-	Identifier	No	No	adult	0	1	0.0622	0.4172	0.1490	
43	Adult	CG11151	No	-	Identifier	No	Yes	Non-modulated	0	2	0.0035	0.1638	0.0211	
44	Adult	CG5329	No	-	Identifier	Yes	No	Adult	0	2	0.1198	0.5047	0.2373	
45	Adult	RpS9	Yes	Yes	Real name	No	Yes	Non-modulated	0	2	0.0001	0.0948	0.0010	
46	Adult	Ag5r	No	-	Real name	No	Yes	Non-modulated	16	2	0.0361	0.2987	0.1207	
47	Adult	RpL7	No	-	Real name	No	Yes	Non-modulated	1	3	0.0021	0.2026	0.0105	
48	Adult	RpS14a	Yes	-	Real name	No	Yes	Non-modulated	1	3	0.0001	0.1499	0.0010	
49	Adult	Cyp1	Yes	-	Real name	No	Yes	Non-modulated	16	2	0.0026	0.2192	0.0118	
50	Adult	Sec61beta	No	-	Real name	No	Yes	Non-modulated	0	3	0.0002	0.1505	0.0010	
51	Adult	rab1	Yes	-	Real name	No	Yes	Non-modulated	52	5	0.0002	0.1601	0.0010	
52	Adult	CG4759	No	-	Identifier	No	Yes	Non-modulated	0	2	0.0001	0.1168	0.0010	
53	Adult	CG3609	No	-	Identifier	No	Yes	Non-modulated	2	4	0.0122	0.1882	0.0647	
54	Adult	RpL19	Yes	-	Real name	No	Yes	Non-modulated	0	3	0.0021	0.2689	0.0079	
55	Adult	CG7283	Yes	-	Identifier	No	Yes	Non-modulated	1	3	0.0019	0.1616	0.0116	
56	Adult	CG5885, BEST:CK01296	No	-	Identifier	No	Yes	Non-modulated	0	2	0.0003	0.2795	0.0010	
57	Adult	CG10992	No	-	Identifier	No	Yes	Non-modulated	6	4	0.0209	0.3374	0.0620	
58	Adult	CG6910	No	-	Identifier	No	Yes	Adult	0	4	0.0056	0.3569	0.0156	
59	Adult	CG16743	No	-	Identifier	No	No	Adult	3	1	0.1860	0.3817	0.4873	
60	Adult	CG9336	No	-	Identifier	Yes	Yes	Non-modulated	2	3	0.0230	0.2787	0.0824	

No.	Library	Gene name	Duplicates	Dup. excl.	Name	Orphan	*Anopheles*	Expression	Paralogues	Exons	*dN*	*dS*	*dN/dS*	H_0: *dN/dS*=1
61	Adult	14-3-3zeta	Yes	-	Real name	No	Yes	Non-modulated	1	6	0.0000	0.0162	0.0010	
62	Adult	Nap1	No	-	Real name	No	Yes	Non-modulated	3	6	0.0285	0.2734	0.1042	
63	Adult	RpS25	Yes	Yes	Real name	No	Yes	Non-modulated	0	2	0.0001	0.0591	0.0010	
64	Adult	CG17244	No	-	Identifier	Yes	No	Adult	0	1	0.1203	0.5124	0.2348	
65	Adult	CG15043	No	-	Identifier	Yes	Yes	Non-modulated	1	2	0.0906	0.3157	0.2870	
66	Adult	Nacalpha	Yes	-	Real name	No	Yes	Non-modulated	1	2	0.0065	0.2382	0.0271	
67	Adult	CG14235	Yes	-	Identifier	No	Yes	Non-modulated	0	2	0.0002	0.1578	0.0010	
68	Adult	CG13315	No	-	Identifier	Yes	No	Non-modulated	0	1	0.0127	0.1817	0.0698	
69	Adult	Cys	Yes	-	Real name	No	Yes	Non-modulated	2	2	0.0838	0.5773	0.1451	
70	Adult	RfeSP, CG7361	No	-	Real name	No	Yes	Non-modulated	0	2	0.0081	0.2139	0.0376	
71	Adult	RpL31 ,CG1821	Yes	Yes	Real name	No	Yes	Non-modulated	0	2	0.0002	0.1956	0.0010	
72	Adult	Fer1HCH	Yes	-	Real name	No	Yes	Non-modulated	1	3	0.0121	0.1936	0.0624	
73	Adult	l(2)06225 , CG6105	No	-	Real name	No	Yes	Non-modulated	1	4	0.0002	0.2014	0.0010	
74	Adult	CG11015	Yes	-	Identifier	No	Yes	Non-modulated	1	3	0.0075	0.1196	0.0625	
75	Adult	CG5453	No	-	Identifier	Yes	No	Non-modulated	0	4	0.0677	0.2313	0.2926	
76	Adult	CG13585	No	-	Identifier	No	Yes	Adult	0	3	0.0453	0.1848	0.2449	
77	Adult	RpL13	Yes	Yes	Real name	No	Yes	Non-modulated	0	3	0.0002	0.1547	0.0010	
78	Adult	RpS27A	Yes	-	Real name	No	Yes	Non-modulated	9	2	0.0001	0.0639	0.0010	
79	Adult	RpS6	Yes	-	Real name	No	Yes	Non-modulated	1	2	0.0025	0.1688	0.0151	
80	Adult	bic	Yes	-	Real name	No	Yes	Non-modulated	2	2	0.0078	0.2179	0.0359	
81	Adult	yip6	Yes	Yes	Real name	No	Yes	Non-modulated	0	4	0.0022	0.3254	0.0067	
82	Adult	CG7584	No	-	Identifier	Yes	Yes	Non-modulated	1	2	0.0238	0.5235	0.0454	
83	Adult	Vha13	No	-	Real name	No	Yes	Non-modulated	0	3	0.0033	0.2048	0.0163	
84	Adult	CG6503	No	-	Identifier	Yes	Yes	Adult	0	1	0.0187	0.3235	0.0577	
85	Adult	CG4800	Yes	Yes	Identifier	No	Yes	Non-modulated	0	1	0.0105	0.4502	0.0233	
86	Adult	CG1883	Yes	Yes	Identifier	No	Yes	Non-modulated	0	4	0.0021	0.0947	0.0227	
87	Adult	sp2	No	-	Real name	No	Yes	Adult	24	4	0.1345	0.3048	0.4413	
88	Adult	Ef1beta	No	-	Real name	No	Yes	Non-modulated	2	2	0.0023	0.2396	0.0098	
89	Adult	CG8869	No	-	Identifier	No	Yes	Non-modulated	133	1	0.0279	0.4038	0.0692	
90	Adult	CG8857	Yes	-	Identifier	No	Yes	Non-modulated	0	5	0.0001	0.0819	0.0010	
91	Adult	CG17280	No	-	Identifier	No	Yes	Non-modulated	0	2	0.0146	0.2480	0.0588	
92	Adult	CG7808	No	-	Identifier	No	Yes	Non-modulated	0	4	0.0059	0.1847	0.0320	

No.	Library	Gene name	Duplicates	Dup. excl.	Name	Orphan	*Anopheles*	Expression	Paralogues	Exons	*dN*	*dS*	*dN/dS*	H_0: *dN/dS*=1
93	Adult	CG8495	No	-	Identifier	No	Yes	Non-modulated	0	3	0.0002	0.1614	0.0010	
94	Adult	CG12775	Yes	-	Identifier	No	Yes	Non-modulated	0	3	0.0029	0.2427	0.0120	
95	Adult	CG8332	Yes	Yes	Identifier	No	Yes	Non-modulated	0	4	0.0002	0.1967	0.0010	
96	Adult	RpL15	Yes	Yes	Real name	No	Yes	Non-modulated	0	2	0.0003	0.3088	0.0010	
97	Adult	CG11501	No	-	Identifier	Yes	No	Adult	0	1	-	-	-	
98	Adult	Ics	No	-	Real name	No	No	Adult	33	1	0.1073	0.7423	0.1446	
99	Adult	CG4046	Yes	Yes	Identifier	No	Yes	Non-modulated	0	5	0.0000	0.0436	0.0010	
100	Adult	RpS19	Yes	Yes	Real name	No	Yes	Non-modulated	1	3	0.0001	0.1126	0.0010	
101	Adult	RpL14	Yes	-	Real name	No	Yes	Non-modulated	0	4	0.0024	0.0777	0.0314	
102	Adult	RpP1	Yes	-	Real name	No	Yes	Non-modulated	1	2	0.0259	0.1779	0.1457	
103	Adult	RpL32	Yes	Yes	Real name	No	Yes	Non-modulated	0	2	0.0001	0.1065	0.0010	
104	Adult	sta	Yes	-	Real name	No	Yes	Non-modulated	0	3	0.0002	0.1954	0.0010	
105	Adult	eIF-5A	Yes	-	Real name	No	Yes	Non-modulated	0	4	0.0266	0.0702	0.3788	H0 not rejected
106	Adult	Scp1	No	-	Real name	No	Yes	Adult	0	4	0.0047	0.1218	0.0388	
107	Adult	CG1475	Yes	-	Identifier	No	Yes	Non-modulated	0	3	0.0002	0.1675	0.0010	
108	Adult	alphaTry	No	-	Real name	No	Yes	Non-modulated	183	1	0.0251	0.3447	0.0729	
109	Adult	RpS18	Yes	Yes	Real name	No	Yes	Non-modulated	0	3	0.0002	0.1566	0.0010	
110	Adult	RpS3A	Yes	Yes	Real name	No	Yes	Non-modulated	0	2	0.0003	0.2628	0.0010	
111	Adult	CG1678	No	-	Identifier	No	No	Adult	0	1	0.1333	0.2185	0.6100	H0 not rejected
112	Adult	Qm	No	-	Real name	No	Yes	Non-modulated	1	5	0.0003	0.2801	0.0010	
113	Adult	CG4716	No	-	Identifier	Yes	No	Adult	0	1	0.1536	0.3765	0.4081	
114	Adult	CG13324	No	-	Identifier	Yes	Yes	Adult	1	1	0.0141	0.6342	0.0222	
115	Adult	CG9762	Yes	-	Identifier	No	Yes	Non-modulated	0	4	0.0259	0.1905	0.1361	
116	Adult	CG16978	No	-	Identifier	Yes	No	Adult	0	1	0.1298	0.2759	0.2169	
117	Adult	crl	No	-	Real name	No	Yes	Non-modulated	21	4	0.0083	0.1523	0.0543	
118	Adult	RpL27a	Yes	Yes	Real name	No	Yes	Non-modulated	0	6	0.0001	0.1474	0.0010	
119	Adult	CG6398	Yes	-	Identifier	No	Yes	Non-modulated	0	5	0.0002	0.1691	0.0010	
120	Adult	BG:DS06874.1	No	-	Identifier	Yes	No	Adult	6	1	0.0489	0.3545	0.1381	
121	Adult	CG2177	No	-	Identifier	No	Yes	Non-modulated	0	3	0.0256	0.2628	0.0973	
122	Adult	CG7380	No	-	Identifier	No	Yes	Non-modulated	0	2	0.0138	0.5779	0.0238	
123	Adult	CG5497	No	-	Identifier	No	Yes	Adult	0	3	0.0288	0.2457	0.1173	

No.	Library	Gene name	Duplicates	Dup. excl.	Name	Orphan	*Anopheles*	Expression	Paralogues	Exons	*dN*	*dS*	*dN/dS*	H_0: *dN/dS*=1
124	Adult	CG3751	Yes	Yes	Identifier	No	Yes	Non-modulated	0	2	0.0002	0.1638	0.0010	
125	Adult	CG1468	No	-	Identifier	No	No	Adult	3	2	-	-	-	
126	Adult	ctp	No	-	Real name	No	Yes	Non-modulated	2	5	0.0001	0.1137	0.0010	
127	Adult	SmB	No	-	Real name	No	Yes	Non-modulated	0	2	0.0047	0.1838	0.0257	
128	Adult	CG7025	No	-	Identifier	No	Yes	Non-modulated	19	5	0.0142	0.3368	0.0422	
129	Adult	CG17571	No	-	Identifier	No	Yes	Non-modulated	182	1	0.0322	0.2961	0.1086	
130	Adult	ATPsyn-gamma	Yes	-	Real name	No	Yes	Non-modulated	0	2	0.0022	0.1175	0.0190	
131	Adult	CG6770	No	-	Identifier	No	Yes	Non-modulated	0	1	0.0004	0.4324	0.0010	
132	Adult	NEW001	No	-	Identifier	Yes	Yes	Adult	3	2	0.0629	0.2667	0.2357	
133	Adult	CG9288	No	-	Identifier	No	Yes	Non-modulated	0	2	0.0703	0.3548	0.1982	
134	Adult	CG17022	No	-	Identifier	No	No	Adult	4	1	0.2510	0.5107	0.4914	
135	Adult	CG12643	No	-	Identifier	Yes	No	Non-modulated	0	1	0.0177	0.2269	0.0780	
136	Adult	Gip	No	-	Real name	No	Yes	Non-modulated	0	1	0.0074	0.4336	0.0171	
137	Adult	Syb	Yes	Yes	Real name	No	Yes	Non-modulated	2	4	-	-	-	
138	Adult	RpS13	Yes	-	Real name	No	Yes	Non-modulated	0	3	0.0055	0.1348	0.0408	
139	Adult	CG13335	No	-	Identifier	Yes	No	Non-modulated	0	1	0.0127	0.1817	0.0698	
140	Adult	Rpn7	No	-	Real name	No	Yes	Non-modulated	1	1	0.0049	0.4747	0.0104	
141	Adult	CG6084	No	-	Identifier	No	Yes	Non-modulated	6	4	0.0048	0.1678	0.0289	
142	Adult	ox	No	-	Real name	No	Yes	Adult	0	2	0.0075	0.0798	0.0942	H0 not rejected
143	Adult	CG3344	No	-	Identifier	No	Yes	Non-modulated	2	4	0.0249	0.2449	0.1018	
144	Adult	Pdh	No	-	Real name	No	Yes	Adult	20	3	0.0082	0.3305	0.0249	
145	Adult	CG11876	No	-	Identifier	No	Yes	Non-modulated	1	6	0.0021	0.1847	0.0116	
146	Adult	Rpn11	No	-	Real name	No	Yes	Non-modulated	1	4	0.0001	0.1338	0.0010	
147	Adult	CG15111_new_annotation	No	-	Identifier	No	Yes	Non-modulated	0	3	-	-	-	
148	Adult	CG3893	No	-	Identifier	No	Yes	Non-modulated	0	2	-	-	-	
149	Adult	CG5134	No	-	Identifier	No	No	Non-modulated	0	3	0.0031	0.3535	0.0088	
150	Adult	DptB, CG10794	No	-	Real name	Yes	Yes	Adult	1	2	0.0250	0.3007	0.0831	
151	Adult	RfaBp	No	-	Real name	No	Yes	Non-modulated	1	8	-	-	-	
152	Adult	fln	No	-	Real name	Yes	Yes	Adult	0	4	0.0026	0.1637	0.0158	
153	Adult	MtnB	No	-	Real name	No	Yes	Adult	1	2	-	-	-	
154	Adult	EG:23E12.3	No	-	Identifier	No	Yes	Adult	2	10	0.0356	0.3436	0.1037	
155	Adult	RpL30, CG10652	Yes	Yes	Real name	No	Yes	Non-modulated	0	2	0.0037	0.2332	0.0157	

No.	Library	Gene name	Duplicates	Dup. excl.	Name	Orphan	*Anopheles*	Expression	Paralogues	Exons	*dN*	*dS*	*dN/dS*	H_0: *dN/dS*=1
156	Adult	Tsp42Ee	Yes	-	Real name	No	Yes	Non-modulated	21	5	0.1312	0.1545	0.8493	H0 not rejected
157	Adult	CG4338	Yes	-	Identifier	No	Yes	Non-modulated	0	1	0.0108	0.3292	0.0329	
158	Adult	CG17575	No	-	Identifier	No	Yes	Adult	15	3	0.0187	0.2738	0.0683	
159	Adult	CG15198	No	-	Identifier	Yes	No	Adult	0	1	0.0848	0.3026	0.2803	
160	Adult	PebIII	No	-	Real name	No	Yes	Adult	2	2	0.0248	0.4436	0.0559	
161	Adult	CG3446	No	-	Identifier	No	Yes	Non-modulated	0	3	0.0027	0.2527	0.0108	
162	Adult	CG18284	No	-	Identifier	No	Yes	Adult	21	5	0.0609	0.2202	0.2766	
163	Adult	CG3760 anon2C9	No	-	Identifier	No	Yes	Non-modulated	1	4	0.0155	0.1572	0.0986	
164	Adult	CG12373	No	-	Identifier	No	Yes	Non-modulated	0	1	0.0349	0.4195	0.0831	
165	Adult	FK506-bp2	Yes	Yes	Real name	No	Yes	Non-modulated	4	3	0.0050	0.3651	0.0138	
166	Adult	CG13356	No	-	Identifier	No	Yes	Non-modulated	26	2	0.0381	0.3197	0.1192	
167	Adult	Prosbeta5	Yes	-	Real name	No	Yes	Non-modulated	4	2	0.0233	0.2743	0.0849	
168	Adult	CG10465	No	-	Identifier	No	Yes	Non-modulated	4	2	0.0092	0.2937	0.0314	
169	Adult	CG18594	No	-	Identifier	No	Yes	Adult	7	1	0.0044	0.1222	0.0358	
170	Adult	CG11892	No	-	Identifier	No	Yes	Non-modulated	41	2	0.0427	0.3810	0.1122	
171	Adult	CG7224	No	-	Identifier	No	Yes	Adult	1	2	0.0183	0.4779	0.0382	
172	Adult	CG17202	No	-	Identifier	No	Yes	Adult	0	1	0.0737	0.5258	0.1402	
173	Adult	CG8009	No	-	Identifier	No	Yes	Non-modulated	1	4	0.0107	0.1380	0.0774	
174	Adult	CG2108	No	-	Identifier	No	Yes	Non-modulated	51	3	0.0127	0.2053	0.0617	
175	Adult	CG4666	No	-	Identifier	No	Yes	Non-modulated	1	2	0.0171	0.4531	0.0378	
176	Adult	Damm	No	-	Real name	No	Yes	Adult	2	5	0.1444	0.3665	0.3940	
177	Adult	mago	Yes	-	Real name	No	Yes	Non-modulated	0	2	0.0003	0.2770	0.0010	
178	Adult	CG17239	No	-	Identifier	No	Yes	Non-modulated	174	1	0.1893	0.3248	0.5829	
179	Adult	CG7033	No	-	Identifier	No	Yes	Non-modulated	7	3	0.0211	0.2290	0.0923	
180	Adult	BG:DS00941.14	No	-	Identifier	Yes	No	Adult	4	2	0.0192	0.1994	0.0962	
181	Adult	CG17327	No	-	Identifier	No	Yes	Non-modulated	0	1	0.0006	0.5606	0.0010	
182	Adult	CG9894	Yes	Yes	Identifier	No	No	Non-modulated	1	4	0.0001	0.0549	0.0010	
183	Adult	Mlf	Yes	-	Real name	No	Yes	Non-modulated	0	4	0.0099	0.2406	0.0413	
184	Adult	CG14558	No	-	Identifier	No	Yes	Non-modulated	7	5	0.0087	0.3646	0.0238	
185	Adult	CG15304	No	-	Identifier	No	Yes	Non-modulated	0	2	0.0000	0.0447	0.0010	H0 not rejected
186	Adult	CG14105	No	-	Identifier	No	Yes	Adult	0	3	0.0308	0.4619	0.0667	
187	Adult	CG15027	No	-	Identifier	No	Yes	Adult	0	2	0.0171	0.3710	0.0460	

No.	Library	Gene name	Duplicates	Dup. excl.	Name	Orphan	*Anopheles*	Expression	Paralogues	Exons	*dN*	*dS*	*dN/dS*	H_0: *dN/dS*=1
188	Adult	CG16826	No	-	Identifier	Yes	No	Adult	1	1	0.1297	0.3165	0.4097	
189	Adult	CG7770	No	-	Identifier	No	Yes	Non-modulated	0	3	0.0439	0.2962	0.1482	
190	Adult	Arc-p34	No	-	Real name	No	Yes	Non-modulated	0	7	0.0023	0.2130	0.0109	
191	Adult	eEF1delta	Yes	-	Real name	No	Yes	Non-modulated	1	3	0.0417	0.3151	0.1324	
192	Adult	CG7778	No	-	Identifier	Yes	No	Adult	6	3	0.0534	0.3010	0.1772	
193	Adult	CG6746	Yes	-	Identifier	No	Yes	Non-modulated	1	1	0.0041	0.3007	0.0136	
194	Adult	EG:34F3.5	No	-	Identifier	Yes	Yes	Adult	39	3	0.0065	0.6044	0.0107	
195	Adult	Rpn9	No	-	Real name	No	Yes	Non-modulated	0	5	0.0043	0.2978	0.0143	
196	Adult	CG2998	Yes	Yes	Identifier	No	Yes	Non-modulated	1	2	0.0001	0.0982	0.0010	
197	Adult	l(2)04154	Yes	-	Real name	No	Yes	Non-modulated	29	11	0.0623	0.1534	0.4060	
198	Adult	CG3040	No	-	Identifier	No	Yes	Adult	5	1	0.0171	0.2832	0.0603	
199	Adult	CG12374	No	-	Identifier	No	Yes	Adult	0	5	0.0074	0.2057	0.0359	
200	Adult	CG12859	No	-	Identifier	Yes	Yes	Non-modulated	0	2	0.0109	0.3554	0.0306	
201	Adult	CG11069	No	-	Identifier	No	Yes	Non-modulated	17	12	0.0003	0.2743	0.0010	
202	Adult	CG11455	No	-	Identifier	No	Yes	Non-modulated	0	2	0.0003	0.2837	0.0010	
203	Adult	CG9897	No	-	Identifier	No	Yes	Adult	70	2	0.0790	0.4412	0.1790	
204	Adult	to	No	-	Real name	Yes	Yes	Adult	14	3	0.0220	0.6117	0.0360	
205	Adult	primo-1	No	-	Real name	No	Yes	Adult	1	7	0.0003	0.2911	0.0010	
206	Adult	CG13603	No	-	Identifier	Yes	Yes	Non-modulated	0	4	0.0129	0.2512	0.0514	
207	Adult	wal	No	-	Real name	No	Yes	Non-modulated	0	4	0.0040	0.2270	0.0175	
208	Adult	CG14645	No	-	Identifier	Yes	Yes	Adult	3	1	0.0251	0.2553	0.0982	
209	Adult	Kisir	No	-	Real name	No	Yes	Non-modulated	0	1	0.0003	0.3065	0.0010	
210	Adult	CG6921	No	-	Identifier	No	Yes	Non-modulated	17	4	0.0122	0.4661	0.0263	
211	Adult	ProsMA5	No	-	Real name	No	Yes	Non-modulated	12	3	0.0048	0.2970	0.0160	
212	Adult	CG4108	No	-	Identifier	No	Yes	Non-modulated	2	2	0.0004	0.4378	0.0010	
213	Adult	Alas	No	-	Real name	No	Yes	Non-modulated	3	2	0.0336	0.5462	0.0615	
214	Adult	CG10219	No	-	Identifier	No	Yes	Non-modulated	0	4	0.0182	0.2417	0.0754	
215	Adult	CG6364	No	-	Identifier	No	Yes	Non-modulated	3	5	0.0003	0.3112	0.0010	
216	Adult	CG13155	No	-	Identifier	Yes	No	Adult	0	2	0.0881	0.3909	0.2254	
217	Adult	CG10570	No	-	Identifier	Yes	No	Adult	0	2	0.0056	0.1449	0.0389	
218	Adult	CG9354	No	-	Identifier	No	Yes	Non-modulated	1	2	0.0024	0.1448	0.0167	
219	Adult	Transferrin	No	-	Real name	No	Yes	Non-modulated	2	5	0.0188	0.2332	0.0804	

No.	Library	Gene name	Duplicates	Dup. excl.	Name	Orphan	*Anopheles*	Expression	Paralogues	Exons	*dN*	*dS*	*dN/dS*	H_0: *dN/dS*=1
220	Adult	CG17508	No	-	Identifier	Yes	Yes	Non-modulated	0	2	0.0714	0.2306	0.3095	
221	Adult	CG9306	No	-	Identifier	No	Yes	Non-modulated	0	3	0.0126	0.2870	0.0439	
222	Adult	ATPsyn-beta	Yes	-	Real name	No	Yes	Non-modulated	9	3	0.0023	0.3074	0.0074	
223	Adult	CG5548	No	-	Identifier	No	Yes	Non-modulated	0	2	0.0064	0.1890	0.0337	
224	Adult	CG10527	No	-	Identifier	No	Yes	Non-modulated	3	3	0.0027	0.3610	0.0075	
225	Adult	CG5445	No	-	Identifier	No	No	Non-modulated	0	4	0.0268	0.2414	0.1112	
226	Adult	CG5547	No	-	Identifier	No	Yes	Non-modulated	4	3	0.0036	0.1407	0.0253	
227	Adult	me31B	No	-	Real name	No	Yes	Non-modulated	32	5	0.0033	0.0342	0.0961	
228	Adult	CG3321	No	-	Identifier	No	Yes	Non-modulated	0	2	0.0082	0.2850	0.0288	
229	Adult	CG10911	No	-	Identifier	Yes	No	Non-modulated	33	2	0.0838	0.4412	0.1899	
230	Adult	regucalcin	No	-	Real name	No	Yes	Non-modulated	1	3	0.0309	0.3765	0.0820	
231	Adult	yps	No	-	Real name	No	Yes	Non-modulated	0	4	-	-	-	
232	Adult	CG8093	No	-	Identifier	No	Yes	Adult	21	2	0.0000	0.0292	0.0010	
233	Adult	CG17472	No	-	Identifier	Yes	No	Adult	0	2	0.1863	0.5486	0.3395	
234	Adult	CG17737	No	-	Identifier	No	Yes	Non-modulated	0	3	0.0375	0.3503	0.1071	
235	Adult	CG10252	No	-	Identifier	No	Yes	Adult	1	4	0.0019	0.4091	0.0047	
236	Adult	CG9978	No	-	Identifier	No	Yes	Adult	13	3	0.0002	0.1530	0.0010	
237	Adult	fbl	No	-	Real name	No	Yes	Non-modulated	0	7	0.0051	0.2527	0.0201	
238	Adult	BG:DS00941.15, CG7968	No	-	Identifier	Yes	No	Adult	3	2	0.0175	0.2742	0.0637	
239	Adult	CG7542	No	-	Identifier	No	Yes	Adult	172	1	0.0531	0.3035	0.1750	
240	Adult	CG9259	No	-	Identifier	No	Yes	Adult	31	1	0.0281	0.2684	0.1048	
241	Adult	CG6922	No	-	Identifier	No	Yes	Non-modulated	1	4	0.0254	0.4457	0.0570	
242	Adult	CG7181	No	-	Identifier	Yes	No	Adult	0	2	0.0185	0.1822	0.1017	
243	Adult	CG18778	No	-	Identifier	No	No	Adult	0	2	-	-	-	
244	Adult	Cp16	No	-	Real name	Yes	No	Adult	0	2	0.0565	0.2266	0.2493	
245	-	-	-	-	-	-	-	-	-	-	-	-	-	
246	Adult	CG5399	No	-	Identifier	Yes	No	Non-modulated	0	3	0.0661	0.4799	0.1377	
247	Adult	CG1324	No	-	Identifier	No	No	Adult	0	2	0.0930	0.3594	0.2589	
248	Adult	RpL40	Yes	-	Real name	No	Yes	Non-modulated	8	2	0.0001	0.1469	0.0010	
249	Adult	Scsalpha	No	-	Real name	No	Yes	Non-modulated	2	4	0.0023	0.1436	0.0161	
250	Adult	CG9568	No	-	Identifier	Yes	No	Adult	1	3	0.0569	0.3413	0.1668	
251	Adult	CG13328	No	-	Identifier	Yes	Yes	Adult	0	3	-	-	-	

No.	Library	Gene name	Duplicates	Dup. excl.	Name	Orphan	*Anopheles*	Expression	Paralogues	Exons	*dN*	*dS*	*dN/dS*	H_0: *dN/dS*=1
252	Adult	CG8769	No	-	Identifier	Yes	Yes	Non-modulated	0	5	0.0607	0.2694	0.2253	
253	Adult	CG14061	No	-	Identifier	Yes	Yes	Adult	0	1	0.1093	0.3579	0.3054	
254	Adult	CG5317	No	-	Identifier	No	Yes	Non-modulated	1	4	0.0158	0.1635	0.0964	
255	Adult	CG12408	No	-	Identifier	No	Yes	Adult	19	1	0.0148	0.2646	0.0561	
256	Adult	Tsp42Ed	No	-	Real name	No	Yes	Non-modulated	19	5	0.0105	0.3169	0.0330	
257	Adult	mbf1	No	-	Real name	No	Yes	Non-modulated	0	4	0.0030	0.2238	0.0132	
258	Adult	CG7787	Yes	Yes	Identifier	No	Yes	Non-modulated	0	2	0.0130	0.3072	0.0424	
259	Adult	CG1532	No	-	Identifier	No	Yes	Non-modulated	0	4	0.0072	0.3658	0.0197	
260	Adult	smp-30	No	-	Real name	No	Yes	Non-modulated	3	2	0.0409	0.2764	0.1480	
261	Adult	Hsp67Bb, CG4456	No	-	Real name	No	Yes	Non-modulated	2	3	0.0239	0.2732	0.0873	
262	Adult	CG2852	Yes	-	Identifier	No	Yes	Non-modulated	14	3	0.0086	0.2375	0.0361	
263	Adult	fan , CG7919	No	-	Real name	No	Yes	Adult	1	3	0.1008	0.4298	0.2344	
264	Adult	FKBP59	No	-	Real name	No	Yes	Non-modulated	6	8	0.0076	0.3016	0.0252	
265	Adult	CG8586	No	-	Identifier	No	Yes	Adult	146	3	0.0988	0.4586	0.2155	
266	Adult	AttD, CG7629	No	-	Real name	Yes	No	Adult	3	2	0.0356	0.3004	0.1186	
267	Adult	CG9066	No	-	Identifier	No	Yes	Non-modulated	1	4	0.0275	0.2820	0.0976	
268	Adult	ran	No	-	Real name	No	Yes	Non-modulated	46	2	0.0004	0.3514	0.0010	
269	Adult	AttA	No	-	Real name	Yes	No	Adult	3	2	0.0310	0.4254	0.0729	
270	Adult	tsr	Yes	Yes	Real name	No	Yes	Non-modulated	1	4	0.0002	0.1786	0.0010	
271	Adult	RpL18A	Yes	Yes	Real name	No	Yes	Non-modulated	0	2	0.0051	0.0871	0.0581	
272	Adult	Dpt	No	-	Real name	Yes	No	Adult	1	1	0.0654	0.3422	0.1912	
273	Adult	Elongin-C	Yes	-	Real name	No	Yes	Non-modulated	0	2	0.0000	0.0000	0.0010	H0 not rejected
274	Adult	CG18067	No	-	Identifier	No	No	Adult	0	2	0.1631	0.4149	0.3932	
275	Adult	CG2543	No	-	Identifier	No	Yes	Adult	1	5	0.0939	0.4377	0.2147	
276	Adult	CG3683	No	-	Identifier	No	Yes	Non-modulated	0	4	0.0166	0.2409	0.0690	
277	Adult	CG11752	No	-	Identifier	Yes	Yes	Non-modulated	0	1	0.0272	0.2751	0.0989	
278	Adult	CG5582	No	-	Identifier	No	Yes	Non-modulated	0	5	0.0155	0.2794	0.0555	
279	Adult	CG11024	No	-	Identifier	No	Yes	Non-modulated	1	3	0.0162	0.2902	0.0559	
280	Adult	Fad	No	-	Real name	No	Yes	Non-modulated	6	1	0.0254	0.3519	0.0721	
281	Adult	CG3199	No	-	Identifier	Yes	No	Adult	1	2	0.0576	0.3290	0.1749	
282	Adult	eIF5, CG9177	No	-	Real name	No	Yes	Non-modulated	0	7	0.0002	0.2180	0.0010	
283	Adult	CG3214	No	-	Identifier	No	Yes	Adult	0	4	0.0140	0.3646	0.0383	

No.	Library	Gene name	Duplicates	Dup. excl.	Name	Orphan	*Anopheles*	Expression	Paralogues	Exons	*dN*	*dS*	*dN/dS*	H_0: *dN/dS*=1
284	Adult	CG3305	Yes	-	Identifier	No	Yes	Non-modulated	0	4	0.0957	0.2356	0.4060	
285	Adult	LysS	No	-	Real name	No	Yes	Adult	10	1	0.0061	0.5932	0.0104	
286	Adult	CG9471	No	-	Identifier	No	Yes	Adult	0	2	0.0130	0.3651	0.0356	
287	Adult	Trip1	No	-	Real name	No	Yes	Non-modulated	7	3	0.0025	0.1642	0.0152	
288	Adult	CG17494	No	-	Identifier	No	Yes	Non-modulated	1	4	0.0934	0.3269	0.2856	
289	Adult	CG10005	No	-	Identifier	No	Yes	Non-modulated	0	3	0.0003	0.2617	0.0010	
290	Adult	thioredoxin	No	-	Real name	No	Yes	Non-modulated	6	4	0.0254	0.2266	0.1122	
291	Adult	CG5703	No	-	Identifier	No	Yes	Non-modulated	1	3	0.0003	0.2741	0.0010	
292	Adult	Hs2st	No	-	Real name	No	Yes	Non-modulated	1	5	0.0050	0.2340	0.0213	
293	Adult	chp	No	-	Real name	No	Yes	Adult	50	15	0.0002	0.2454	0.0010	
294	Adult	CG1844	No	-	Identifier	No	No	Non-modulated	1	3	-	-	-	
295	Adult	CG10472	No	-	Identifier	No	Yes	Non-modulated	171	3	0.0238	0.4144	0.0574	
296	Adult	fok	No	-	Real name	Yes	Yes	Non-modulated	0	3	0.0060	0.4879	0.0123	
297	Adult	CG14482	Yes	-	Identifier	No	No	Non-modulated	0	2	0.0001	0.1013	0.0010	
298	Adult	CG14022	No	-	Identifier	No	Yes	Adult	2	3	0.0001	0.1354	0.0010	
299	Adult	CG13095	No	-	Identifier	No	Yes	Non-modulated	12	1	0.0231	0.5249	0.0440	
300	Adult	ocn	No	-	Real name	No	Yes	Adult	3	3	0.0295	0.4348	0.0679	
301	Adult	cta	Yes	-	Real name	No	Yes	Non-modulated	7	5	0.0123	0.2367	0.0519	
302	Adult	porin	No	-	Real name	No	Yes	Non-modulated	3	4	0.0000	0.0398	0.0010	
303	Adult	CG8417	No	-	Identifier	No	Yes	Adult	0	3	0.0028	0.4495	0.0063	
304	Adult	CG13618	No	-	Identifier	Yes	Yes	Non-modulated	16	3	0.0089	0.1147	0.0778	
305	Adult	CG5844	No	-	Identifier	No	Yes	Non-modulated	7	4	0.0022	0.3686	0.0060	
306	Adult	fau, anoxia	No	-	Real name	Yes	Yes	Non-modulated	0	8	0.0032	0.0616	0.0512	
307	Adult	CG1534	No	-	Identifier	No	Yes	Non-modulated	1	3	0.0001	0.0899	0.0010	
308	Adult	CG8229	No	-	Identifier	Yes	No	Non-modulated	0	3	0.0162	0.0928	0.1744	
309	Adult	BG:DS04095.3 , CG4959	No	-	Identifier	Yes	No	Adult	0	2	0.0170	0.3686	0.0460	
310	Adult	Spat	No	-	Real name	No	Yes	Non-modulated	0	3	0.0045	0.5743	0.0078	
311	Adult	CG10799	Yes	-	Identifier	Yes	No	Non-modulated	0	1	0.3247	0.5657	0.5740	
312	Adult	Cam	No	-	Real name	No	Yes	Non-modulated	27	3	0.0001	0.0790	0.0010	
313	Adult	CG5902	No	-	Identifier	No	Yes	Non-modulated	0	3	0.0056	0.2836	0.0196	
314	Adult	CG13309	No	-	Identifier	Yes	No	Adult	6	1	0.0905	0.3412	0.2651	
315	Adult	CG11342	No	-	Identifier	No	Yes	Adult	0	1	0.0568	0.4070	0.1395	

No.	Library	Gene name	Duplicates	Dup. excl.	Name	Orphan	*Anopheles*	Expression	Paralogues	Exons	*dN*	*dS*	*dN/dS*	H_0: *dN/dS*=1
316	Adult	CG10589	No	-	Identifier	Yes	No	Adult	0	1	0.0241	0.3783	0.0638	
317	Adult	Cyp4d2	No	-	Real name	No	Yes	Adult	83	5	0.0225	0.4900	0.0459	
318	Adult	yellow-c	No	-	Real name	Yes	Yes	Non-modulated	12	3	0.0062	0.3252	0.0189	
319	Adult	CG8043	No	-	Identifier	No	Yes	Non-modulated	0	2	0.0249	0.2457	0.1012	
320	Adult	CG17337	No	-	Identifier	No	Yes	Non-modulated	0	5	0.0243	0.2043	0.1190	
321	Adult	CSN5	No	-	Real name	No	Yes	Non-modulated	1	4	0.0004	0.3649	0.0010	
322	Adult	CG9920	No	-	Identifier	No	Yes	Adult	1	2	0.0045	0.2752	0.0165	
323	Adult	CG12203	No	-	Identifier	No	Yes	Adult	0	3	0.0103	0.3624	0.0283	
324	Adult	CG11852	No	-	Identifier	Yes	Yes	Adult	16	4	0.0074	0.5012	0.0148	
325	Adult	kappaTry	No	-	Real name	No	Yes	Adult	157	1	0.0410	0.2475	0.1657	
326	Adult	CG5945	No	-	Identifier	Yes	Yes	Adult	8	4	0.0439	0.3045	0.1443	
327	Adult	CG13442	No	-	Identifier	No	Yes	Adult	1	6	0.0724	0.3364	0.2152	
328	Adult	EG:22E5.9	No	-	Identifier	No	Yes	Adult	0	4	0.0163	0.4085	0.0400	
329	Adult	CG18543	No	-	Identifier	No	No	Non-modulated	0	1	0.0546	0.3520	0.1551	
330	Adult	Os9,CG10658	No	-	Real name	Yes	No	Non-modulated	0	2	0.0690	0.3794	0.1817	
331	Adult	CG17333	No	-	Identifier	No	Yes	Adult	0	3	0.0377	0.3886	0.0971	
332	Adult	CG9645	No	-	Identifier	No	Yes	Adult	106	5	0.0717	0.2695	0.2659	
333	Adult	BM-40/SPARC	No	-	Real name	No	Yes	Non-modulated	0	3	0.0188	0.5566	0.0337	
334	Adult	His2A	No	-	Real name	No	Yes	Non-modulated	0	4	0.0001	0.1144	0.0010	
335	Adult	CG12347	No	-	Identifier	Yes	No	Adult	19	1	0.1004	0.2330	0.4310	
336	Adult	RpL23a	Yes	-	Real name	No	Yes	Non-modulated	1	3	0.0026	0.1385	0.0189	
337	Adult	CG8701	No	-	Identifier	Yes	No	Adult	7	1	0.0581	0.4685	0.1241	
338	Adult	CG2254	No	-	Identifier	No	Yes	Non-modulated	18	4	0.0461	0.4839	0.0953	
339	Adult	CG12253, BEST:LD08487	No	-	Identifier	Yes	No	Non-modulated	0	3	0.0222	0.3733	0.0596	
340	Adult	CG2471	No	-	Identifier	No	Yes	Adult	2	4	0.0110	0.4113	0.0267	
341	Adult	CG17672	No	-	Identifier	No	Yes	Non-modulated	0	5	0.0100	0.2013	0.0495	
342	Adult	RpL46	No	-	Real name	No	Yes	Non-modulated	0	3	0.0000	0.0209	0.0010	H0 not rejected
343	Adult	CG4370	No	-	Identifier	No	Yes	Non-modulated	2	5	0.0035	0.4459	0.0078	
344	Adult	CG9332	No	-	Identifier	No	Yes	Non-modulated	5	9	0.0433	0.2357	0.1836	
345	Adult	BcDNA:GH06048	No	-	Identifier	Yes	No	Adult	10	2	0.2684	0.6137	0.4373	
346	Adult	CG14619	No	-	Identifier	No	Yes	Adult	22	12	0.0040	0.2493	0.0160	
347	Adult	CG7217	No	-	Identifier	No	Yes	Non-modulated	0	3	0.0053	0.4650	0.0115	

No.	Library	Gene name	Duplicates	Dup. excl.	Name	Orphan	*Anopheles*	Expression	Paralogues	Exons	*dN*	*dS*	*dN/dS*	H_0: *dN/dS*=1
348	Adult	CG17377	No	-	Identifier	Yes	Yes	Adult	0	5	-	-	-	
349	Adult	CG16756	No	-	Identifier	No	Yes	Adult	2	2	0.0523	0.3159	0.1656	
350	Adult	CG6543	No	-	Identifier	No	Yes	Non-modulated	10	3	0.0084	0.2788	0.0302	
351	Adult	CG4413	No	-	Identifier	No	Yes	Non-modulated	138	4	0.0290	0.4237	0.0685	
352	Adult	Pka-C3	No	-	Real name	No	No	Adult	1	8	-	-	-	
353	Adult	CG10837	No	-	Identifier	No	Yes	Non-modulated	0	4	0.0686	0.4470	0.1534	
354	Adult	Noe	No	-	Real name	Yes	No	Adult	0	1	0.2734	0.3952	0.6919	H0 not rejected
355	Adult	CG8415	Yes	-	Identifier	No	Yes	Non-modulated	0	5	0.0001	0.0655	0.0010	
356	Adult	Mp20	No	-	Real name	No	Yes	Non-modulated	2	3	0.0098	0.1093	0.0899	
357	Adult	CG12292	No	-	Identifier	No	Yes	Adult	0	2	0.0192	0.2314	0.0828	
358	Adult	CG15697	Yes	-	Identifier	No	Yes	Non-modulated	0	3	0.0064	0.0551	0.1165	
359	Adult	CG10424	No	-	Identifier	No	Yes	Non-modulated	0	2	0.0291	0.2889	0.1007	
360	Adult	l(2)k09913	No	-	Real name	Yes	Yes	Non-modulated	0	5	0.0155	0.3109	0.0498	
361	Adult	CG8661	No	-	Identifier	Yes	No	Adult	2	2	0.0346	0.4067	0.0850	
362	Adult	ACP53EA, CG8622	No	-	Real name	Yes	No	Adult	1	2	0.1298	0.2759	0.4703	
363	Adult	CG8309	Yes	Yes	Identifier	No	Yes	Non-modulated	0	3	0.0001	0.1482	0.0010	
364	Adult	UbcD10	Yes	Yes	Real name	No	Yes	Non-modulated	22	1	0.0122	0.2607	0.0470	
365	Adult	CG6537	No	-	Identifier	No	Yes	Non-modulated	0	3	0.0037	0.6036	0.0061	
366	Adult	CG12699	No	-	Identifier	Yes	No	Adult	1	2	0.2108	0.5849	0.3604	
367	Adult	CG8588	No	-	Identifier	No	No	Non-modulated	1	6	0.1116	0.3057	0.3650	
368	Adult	CG13061	No	-	Identifier	Yes	No	Adult	0	2	0.0184	0.2484	0.0741	
369	Adult	Mst89B, CG6864	No	-	Real name	Yes	No	Adult	0	2	0.2076	0.4532	0.4581	
370	Adult	CG6115	No	-	Identifier	Yes	Yes	Adult	0	2	0.0049	0.2268	0.0217	
371	Adult	CG9129	No	-	Identifier	Yes	No	Adult	1	1	0.0599	0.4595	0.1304	
1	Embryo	RpS27A	Yes	Yes	Real name	No	Yes	Non-modulated	9	2	0.0001	0.0562	0.0010	
2	Embryo	CG4111	No	-	Identifier	No	Yes	Non-modulated	0	3	0.0100	0.3681	0.0272	
3	Embryo	RpS3	Yes	Yes	Real name	No	Yes	Non-modulated	0	2	0.0076	0.1184	0.0642	
4	Embryo	CG18111	No	-	Identifier	Yes	No	Non-modulated	3	2	0.0316	0.5823	0.0543	
5	Embryo	RpL1	No	-	Real name	No	Yes	Non-modulated	0	4	0.0017	0.1401	0.0120	
6	Embryo	CG7424	No	-	Identifier	No	Yes	Non-modulated	0	3	0.0000	0.0359	0.0010	
7	Embryo	RpL31	Yes	-	Real name	No	Yes	Non-modulated	0	2	0.0002	0.1956	0.0010	
8	Embryo	Act57B	No	-	Real name	No	Yes	Non-modulated	13	2	0.0318	0.5507	0.0577	

No.	Library	Gene name	Duplicates	Dup. excl.	Name	Orphan	*Anopheles*	Expression	Paralogues	Exons	*dN*	*dS*	*dN/dS*	H_0: *dN/dS*=1
9	Embryo	CG9184	No	-	Identifier	No	No	Embryo	1	2	0.0360	0.2946	0.1221	
10	Embryo	Prosbeta5	Yes	Yes	Real name	No	Yes	Non-modulated	4	2	0.0051	0.4399	0.0117	
11	Embryo	Hsc70-4	No	-	Real name	No	Yes	Non-modulated	10	2	0.0267	0.1775	0.1506	
12	Embryo	RpS12	Yes	-	Real name	No	Yes	Non-modulated	0	3	0.0001	0.0983	0.0010	
13	Embryo	RpS25	Yes	-	Real name	No	Yes	Non-modulated	0	2	0.0001	0.0738	0.0010	
14	Embryo	RpP0	No	-	Real name	No	Yes	Non-modulated	0	3	0.0017	0.2334	0.0072	
15	Embryo	Su(var)205	No	-	Real name	No	Yes	Non-modulated	5	5	0.0236	0.2641	0.0894	
16	Embryo	TpnC47D	No	-	Real name	No	Yes	Embryo	19	4	0.0027	0.1216	0.0225	
17	Embryo	CG13741	No	-	Identifier	Yes	No	Non-modulated	0	3	0.2377	0.4121	0.5769	
18	Embryo	Tsp66E	No	-	Real name	No	Yes	Non-modulated	12	7	0.0061	0.0259	0.2344	H0 not rejected
19	Embryo	14-3-3zeta	Yes	Yes	Real name	No	Yes	Non-modulated	1	6	0.0000	0.0343	0.0010	
20	Embryo	Ip259	No	-	Real name	No	Yes	Non-modulated	0	2	0.0002	0.2224	0.0010	
21	Embryo	CG6877	No	-	Identifier	No	Yes	Non-modulated	0	4	0.0069	0.2532	0.0271	
22	Embryo	CG10850	No	-	Identifier	No	Yes	Non-modulated	0	3	0.0346	0.4870	0.0710	
23	Embryo	CG4076	No	-	Identifier	No	Yes	Embryo	0	3	0.0637	0.4204	0.1515	
24	Embryo	CG18347	No	-	Identifier	No	Yes	Non-modulated	39	6	0.0028	0.2108	0.0134	
25	Embryo	stathmin	No	-	Real name	No	Yes	Non-modulated	7	8	0.0037	0.1788	0.0205	
26	Embryo	mael, CG11254	No	-	Real name	Yes	Yes	Non-modulated	0	5	0.1041	0.3095	0.3362	
27	Embryo	Updo	No	-	Real name	No	Yes	Non-modulated	184	3	0.0063	0.4361	0.0145	
28	Embryo	CG7834	No	-	Identifier	No	Yes	Non-modulated	0	2	0.0020	0.3932	0.0051	
29	Embryo	BcDNA:GM12291	No	-	Identifier	No	Yes	Non-modulated	0	2	0.0025	0.2285	0.0111	
30	Embryo	Jafrac1	No	-	Real name	No	Yes	Non-modulated	6	2	0.0020	0.4445	0.0046	
31	Embryo	CG3203	Yes	-	Identifier	No	Yes	Non-modulated	0	3	0.0044	0.0884	0.0502	
32	Embryo	CG3227	No	-	Identifier	Yes	Yes	Non-modulated	2	2	0.0441	0.2879	0.1533	
33	Embryo	CG7375	No	-	Identifier	No	Yes	Non-modulated	21	4	0.0070	0.1218	0.0578	
34	Embryo	asf1	No	-	Real name	No	Yes	Non-modulated	0	1	0.0155	0.3997	0.0387	
35	Embryo	Traf1	No	-	Real name	No	Yes	Non-modulated	2	2	0.0043	0.2527	0.0170	
36	Embryo	CG2950	No	-	Identifier	No	Yes	Non-modulated	0	9	0.0332	0.2801	0.1186	
37	Embryo	Scamp	No	-	Real name	No	Yes	Non-modulated	0	5	0.0059	0.3974	0.0148	
38	Embryo	CG2249	No	-	Identifier	No	Yes	Non-modulated	0	3	0.0004	0.3848	0.0010	
39	Embryo	Hel25E	No	-	Real name	No	Yes	Non-modulated	32	8	0.0016	0.2069	0.0076	
40	Embryo	Syx4 , EG:95B7.1	No	-	Real name	No	Yes	Embryo	1	6	0.0160	0.1717	0.0933	

No.	Library	Gene name	Duplicates	Dup. excl.	Name	Orphan	*Anopheles*	Expression	Paralogues	Exons	*dN*	*dS*	*dN/dS*	H_0: *dN/dS*=1
41	Embryo	Hn	No	-	Real name	No	Yes	Non-modulated	2	5	0.0039	0.3951	0.0099	
42	Embryo	CG8298	No	-	Identifier	No	Yes	Non-modulated	5	4	0.0411	0.3683	0.1116	
43	Embryo	CG18661	No	-	Identifier	No	Yes	Embryo	0	1	0.0108	0.5450	0.0199	
44	Embryo	CG14683	No	-	Identifier	No	Yes	Non-modulated	0	2	0.0053	0.2955	0.0179	
45	Embryo	tacc	No	-	Real name	No	Yes	Non-modulated	0	11	0.0896	0.3039	0.2947	
46	Embryo	CG13512	No	-	Identifier	Yes	No	Non-modulated	0	1	0.1314	0.2958	0.4441	
47	Embryo	CG17347	No	-	Identifier	No	Yes	Embryo	0	2	0.0273	0.4181	0.0652	
48	Embryo	m6	No	-	Real name	No	Yes	Embryo	0	5	-	-	-	
49	Embryo	CG12391	No	-	Identifier	No	Yes	Non-modulated	49	2	0.0851	0.1921	0.4430	
50	Embryo	dup	No	-	Real name	No	Yes	Non-modulated	0	4	0.0495	0.2983	0.1659	
51	Embryo	CG4094	No	-	Identifier	No	Yes	Non-modulated	3	5	0.0005	0.5289	0.0010	
52	Embryo	BcDNA:LD29885	No	-	Identifier	No	Yes	Non-modulated	0	1	0.0003	0.3211	0.0010	
53	Embryo	tep2	No	-	Real name	No	Yes	Non-modulated	4	3	0.0356	0.3923	0.0909	
54	Embryo	ATPsyn-d	No	-	Real name	No	Yes	Non-modulated	1	2	0.0044	0.4035	0.0109	
55	Embryo	CG8707	No	-	Identifier	No	Yes	Non-modulated	0	4	0.0057	0.2381	0.0237	
56	Embryo	CG10228	No	-	Identifier	No	Yes	Non-modulated	8	12	0.0004	0.4080	0.0010	
57	Embryo	CG6171	No	-	Identifier	No	Yes	Embryo	0	2	0.1610	0.2493	0.6456	H0 not rejected
58	Embryo	Mage, CG10059	No	-	Real name	No	Yes	Embryo	0	1	0.0759	0.2496	0.3040	
59	Embryo	BcDNA:LD23830	No	-	Identifier	No	Yes	Non-modulated	0	3	0.0105	0.2144	0.0490	
60	Embryo	Hsc70-3, CG4147	No	-	Real name	No	Yes	Non-modulated	10	2	0.0077	0.1125	0.0684	
61	Embryo	Tbp-1	No	-	Real name	No	Yes	Non-modulated	28	3	0.0002	0.2451	0.0010	
62	Embryo	CG11051	No	-	Identifier	Yes	No	Non-modulated	0	3	0.4926	0.6610	0.7453	H0 not rejected
63	Embryo	RpL30, CG10652	Yes	-	Real name	No	Yes	Non-modulated	0	2	0.0000	0.0167	0.0010	H0 not rejected
64	Embryo	CkIIalpha	No	-	Real name	No	Yes	Non-modulated	103	7	-	-	-	
65	Embryo	yip2	No	-	Real name	No	Yes	Non-modulated	3	4	0.0068	0.2622	0.0258	
66	Embryo	GATAd ,CG5034	No	-	Real name	Yes	Yes	Non-modulated	0	6	-	-	-	
67	Embryo	CG10585	No	-	Identifier	No	Yes	Non-modulated	1	3	0.0002	0.2400	0.0010	
68	Embryo	trn	No	-	Real name	No	Yes	Non-modulated	58	2	0.0089	0.1794	0.0496	
69	Embryo	CG10068	No	-	Identifier	No	Yes	Non-modulated	0	2	0.0263	0.2697	0.0977	
70	Embryo	CG4338	Yes	Yes	Identifier	No	Yes	Non-modulated	0	1	0.0111	0.3239	0.0342	
71	Embryo	hts	No	-	Real name	No	Yes	Non-modulated	0	3	-	-	-	
72	Embryo	UbcD10	Yes	-	Real name	No	Yes	Non-modulated	22	1	0.0099	0.2492	0.0396	

No.	Library	Gene name	Duplicates	Dup. excl.	Name	Orphan	*Anopheles*	Expression	Paralogues	Exons	*dN*	*dS*	*dN/dS*	H_0: *dN/dS*=1
73	Embryo	CG9796	No	-	Identifier	No	Yes	Non-modulated	3	4	0.0132	0.2882	0.0456	
74	Embryo	CG12141	No	-	Identifier	No	Yes	Non-modulated	2	5	-	-	-	
75	Embryo	fry	No	-	Real name	No	Yes	Non-modulated	3	8	0.0046	0.1877	0.0247	
76	Embryo	CG9915	No	-	Identifier	No	Yes	Non-modulated	1	7	0.0309	0.2314	0.1334	
77	Embryo	Gapdh1	No	-	Real name	No	Yes	Non-modulated	2	2	0.0072	0.3847	0.0186	
78	Embryo	CG7718	No	-	Identifier	No	Yes	Non-modulated	0	2	0.0142	0.3110	0.0456	
79	Embryo	Pros54	No	-	Real name	No	Yes	Non-modulated	0	3	-	-	-	
80	Embryo	CG7461	No	-	Identifier	No	Yes	Non-modulated	8	1	0.0040	0.3283	0.0121	
81	Embryo	CG4440	No	-	Identifier	Yes	No	Embryo	0	2	0.0195	0.2499	0.0782	
82	Embryo	sds22	No	-	Real name	No	Yes	Non-modulated	17	2	0.0265	0.3909	0.0679	
83	Embryo	rgr	No	-	Real name	No	Yes	Non-modulated	173	1	0.0028	0.6294	0.0044	
84	Embryo	CG14206	No	-	Identifier	No	Yes	Non-modulated	1	1	0.0063	0.1115	0.0566	
85	Embryo	fur2	No	-	Real name	No	Yes	Embryo	10	16	0.0023	0.4171	0.0055	
86	Embryo	Arc32, CG13867	No	-	Real name	No	Yes	Non-modulated	0	2	0.0002	0.2156	0.0010	
87	Embryo	CG10674	No	-	Identifier	No	Yes	Embryo	0	2	0.0079	0.2587	0.0305	
88	Embryo	CG9772	No	-	Identifier	No	Yes	Embryo	2	5	0.0825	0.2257	0.3658	
89	Embryo	RpL9	No	-	Real name	No	Yes	Non-modulated	0	3	0.0043	0.1934	0.0220	
90	Embryo	CG15877	No	-	Identifier	No	Yes	Embryo	0	2	0.0586	0.2558	0.2291	
91	Embryo	Gtp-bp	No	-	Real name	No	Yes	Non-modulated	1	4	0.0004	0.4301	0.0010	
92	Embryo	CG10306	No	-	Identifier	No	Yes	Non-modulated	0	2	0.0086	0.2305	0.0372	
93	Embryo	CG7543	No	-	Identifier	Yes	No	Embryo	0	8	-	-	-	
94	Embryo	tpi	No	-	Real name	No	Yes	Non-modulated	0	2	0.0046	0.2947	0.0155	
95	Embryo	fne	No	-	Real name	No	Yes	Non-modulated	24	4	0.0001	0.0682	0.0010	
96	Embryo	CG4699	No	-	Identifier	No	Yes	Non-modulated	0	9	0.0311	0.3703	0.0840	
97	Embryo	CG13011	No	-	Identifier	Yes	Yes	Embryo	0	3	0.0025	0.2775	0.0089	
98	Embryo	CG17523	No	-	Identifier	No	Yes	Non-modulated	26	1	0.0434	0.4160	0.1043	
99	Embryo	CG15188	No	-	Identifier	Yes	Yes	Embryo	2	2	0.0004	0.3626	0.0010	
100	Embryo	BG:DS05899.3	No	-	Identifier	No	Yes	Non-modulated	6	1	0.0354	0.4128	0.0857	
101	Embryo	CG10419	No	-	Identifier	No	Yes	Non-modulated	0	2	0.0439	0.2898	0.1513	
102	Embryo	grh	No	-	Real name	No	Yes	Non-modulated	22	5	0.0001	0.1116	0.0010	
103	Embryo	l(2)04154	Yes	Yes	Identifier	No	Yes	Non-modulated	29	11	0.0027	0.1965	0.0139	
104	Embryo	Pfk	No	-	Real name	No	Yes	Non-modulated	0	2	0.0003	0.3014	0.0010	

No.	Library	Gene name	Duplicates	Dup. excl.	Name	Orphan	*Anopheles*	Expression	Paralogues	Exons	*dN*	*dS*	*dN/dS*	H_0: *dN/dS*=1
105	Embryo	CG14103	No	-	Identifier	No	Yes	Non-modulated	2	4	0.0275	0.2625	0.1047	
106	Embryo	CG5375	No	-	Identifier	No	Yes	Non-modulated	0	5	0.0426	0.4562	0.0933	
107	Embryo	CaMKI	No	-	Real name	No	Yes	Non-modulated	149	6	0.0078	0.1947	0.0402	
108	Embryo	LamC	No	-	Real name	No	Yes	Non-modulated	10	4	0.0002	0.2170	0.0010	
109	Embryo	CG14482	Yes	Yes	Identifier	No	No	Non-modulated	0	2	0.0001	0.1013	0.0010	
110	Embryo	CG10978	No	-	Identifier	Yes	Yes	Non-modulated	0	3	0.0047	0.3961	0.0119	
111	Embryo	Tm1	No	-	Real name	No	Yes	Non-modulated	44	10	0.0000	0.0112	0.0010	H0 not rejected
112	Embryo	CG15347	No	-	Identifier	No	Yes	Non-modulated	3	2	0.0345	1.2559	0.0274	
113	Embryo	Sep-02	No	-	Real name	No	Yes	Non-modulated	6	2	0.0025	0.5396	0.0046	
114	Embryo	CG9091	Yes	Yes	Identifier	No	Yes	Non-modulated	1	3	0.0001	0.1476	0.0010	
115	Embryo	RpS19	Yes	-	Real name	No	Yes	Non-modulated	1	3	0.0027	0.1113	0.0240	
116	Embryo	Tsp42Ee	Yes	Yes	Real name	No	Yes	Non-modulated	21	5	0.1346	0.1380	0.9757	H0 not rejected
117	Embryo	RpP1	Yes	Yes	Real name	No	Yes	Non-modulated	1	2	0.0261	0.1954	0.1335	
118	Embryo	RpL29	No	-	Real name	No	Yes	Non-modulated	0	3	0.0164	0.3429	0.0478	
119	Embryo	Cys	Yes	Yes	Real name	No	Yes	Non-modulated	2	2	0.0678	0.4458	0.1522	
120	Embryo	RACK1	No	-	Real name	No	Yes	Non-modulated	49	3	0.0114	0.2749	0.0415	
121	Embryo	CG1943	No	-	Identifier	No	Yes	Non-modulated	0	2	0.0076	0.1873	0.0404	
122	Embryo	RpS18	Yes	-	Real name	No	Yes	Non-modulated	0	3	0.0002	0.1693	0.0010	
123	Embryo	mira	No	-	Real name	No	Yes	Embryo	31	1	0.0131	0.3124	0.0419	
124	Embryo	RpS20	Yes	Yes	Real name	No	Yes	Non-modulated	0	4	0.0032	0.2128	0.0151	
125	Embryo	Nacalpha	Yes	Yes	Real name	No	Yes	Non-modulated	1	2	0.0242	0.5237	0.0462	
126	Embryo	CG10686	No	-	Identifier	No	Yes	Non-modulated	25	6	0.0277	0.3682	0.0751	
127	Embryo	LanB2	No	-	Real name	No	Yes	Non-modulated	32	9	0.0187	0.2778	0.0675	
128	Embryo	Int6	No	-	Real name	No	Yes	Non-modulated	0	3	0.0004	0.4179	0.0010	
129	Embryo	BEST:GH02921	No	-	Identifier	No	Yes	Non-modulated	171	4	0.0578	0.3892	0.1486	
130	Embryo	CG7006	No	-	Identifier	No	Yes	Embryo	0	2	0.0022	0.3627	0.0061	
131	Embryo	Prosalpha6	No	-	Real name	No	Yes	Non-modulated	10	5	0.0120	0.2949	0.0405	
132	Embryo	CG7003	No	-	Identifier	No	Yes	Non-modulated	0	5	-	-	-	
133	Embryo	Mhc	No	-	Real name	No	Yes	Non-modulated	132	15	0.0001	0.1350	0.0010	
134	Embryo	Ef1gamma	No	-	Real name	No	Yes	Non-modulated	0	3	0.0061	0.2562	0.0238	
135	Embryo	betaTub56D	No	-	Real name	No	Yes	Non-modulated	12	2	0.0001	0.1202	0.0010	
136	Embryo	ade5	No	-	Real name	No	Yes	Non-modulated	1	5	0.0089	0.4585	0.0195	

No.	Library	Gene name	Duplicates	Dup. excl.	Name	Orphan	*Anopheles*	Expression	Paralogues	Exons	*dN*	*dS*	*dN/dS*	H_0: *dN/dS*=1
137	Embryo	CG12487	No	-	Identifier	Yes	No	Embryo	1	1	0.0712	0.2851	0.2499	
138	Embryo	CG1475	Yes	Yes	Identifier	No	Yes	Non-modulated	0	3	0.0002	0.1876	0.0010	
139	Embryo	CG15189	No	-	Identifier	Yes	Yes	Embryo	2	4	0.0352	0.1975	0.1782	
140	Embryo	Sod2	No	-	Real name	No	Yes	Non-modulated	0	2	0.0092	0.2377	0.0389	
141	Embryo	CG14112	No	-	Identifier	Yes	No	Embryo	1	1	0.0144	0.2959	0.0486	
142	Embryo	PyK	No	-	Real name	No	Yes	Non-modulated	4	4	0.0024	0.1314	0.0179	
143	Embryo	CG4334	No	-	Identifier	No	Yes	Embryo	3	5	0.0039	0.1744	0.0223	
144	Embryo	CG11367	No	-	Identifier	No	Yes	Embryo	1	4	0.0152	0.2778	0.0549	
145	Embryo	CG9926	No	-	Identifier	No	No	Embryo	0	1	0.1377	0.3743	0.3679	
146	Embryo	CG17768	No	-	Identifier	No	Yes	Embryo	2	4	0.0060	0.2509	0.0240	
147	Embryo	CG2944	No	-	Identifier	No	Yes	Non-modulated	2	3	0.0044	0.1856	0.0238	
148	Embryo	eEF1delta	Yes	Yes	Real name	No	Yes	Non-modulated	1	3	0.0429	0.3341	0.1285	
149	Embryo	mus209	No	-	Real name	No	Yes	Embryo	1	2	0.0003	0.3316	0.0010	
150	Embryo	Lam	No	-	Real name	No	Yes	Non-modulated	15	5	-	-	-	
151	Embryo	FK506-bp2	Yes	-	Real name	No	Yes	Non-modulated	4	3	0.0124	0.2351	0.0528	
152	Embryo	CG16974	No	-	Identifier	No	Yes	Non-modulated	56	4	0.0175	0.2106	0.0832	
153	Embryo	oho23B	Yes	Yes	Real name	No	Yes	Non-modulated	0	4	0.0002	0.1886	0.0010	
154	Embryo	Mlc2	Yes	-	Real name	No	Yes	Non-modulated	7	3	0.0000	0.0251	0.0010	
155	Embryo	eIF-5A	Yes	Yes	Real name	No	Yes	Non-modulated	0	4	0.0222	0.0680	0.3268	H0 not rejected
156	Embryo	HmgZ	No	-	Real name	No	Yes	Non-modulated	4	3	0.0001	0.0807	0.0010	
157	Embryo	CG3183	No	-	Identifier	No	Yes	Non-modulated	0	1	0.0137	0.3480	0.0393	
158	Embryo	RpS17	No	-	Real name	No	Yes	Non-modulated	0	3	0.0031	0.0725	0.0422	
159	Embryo	Df31 ,anon1A4, l(2)k05815	No	-	Real name	Yes	No	Non-modulated	0	3	0.0551	0.2036	0.2705	
160	Embryo	CG7283	Yes	Yes	Identifier	No	Yes	Non-modulated	1	3	0.0021	0.1941	0.0106	
161	Embryo	His3.3A	No	-	Real name	No	Yes	Non-modulated	1	2	0.0002	0.1824	0.0010	
162	Embryo	CG12740	No	-	Identifier	No	Yes	Non-modulated	0	4	0.0058	0.1121	0.0519	
163	Embryo	CG6846	No	-	Identifier	No	Yes	Non-modulated	0	2	0.0002	0.2396	0.0010	
164	Embryo	CG18624	No	-	Identifier	No	Yes	Non-modulated	0	2	0.0166	0.3157	0.0525	
165	Embryo	RpL40	Yes	Yes	Real name	No	Yes	Non-modulated	8	2	0.0001	0.1187	0.0010	
166	Embryo	RpL7A	No	-	Real name	No	Yes	Non-modulated	0	5	0.0002	0.2441	0.0010	
167	Embryo	RpL36	No	-	Real name	No	Yes	Non-modulated	0	3	0.0034	0.3612	0.0094	
168	Embryo	RpS6	Yes	Yes	Real name	No	Yes	Non-modulated	1	2	0.0170	0.2303	0.0736	

No.	Library	Gene name	Duplicates	Dup. excl.	Name	Orphan	*Anopheles*	Expression	Paralogues	Exons	*dN*	*dS*	*dN/dS*	H_0: *dN/dS*=1
169	Embryo	RpL19	Yes	Yes	Real name	No	Yes	Non-modulated	0	3	0.0038	0.1820	0.0211	
170	Embryo	eIF-5C, CG2922	No	-	Real name	No	Yes	Non-modulated	0	9	0.0022	0.0672	0.0326	
171	Embryo	RpS14a	Yes	Yes	Real name	No	Yes	Non-modulated	1	3	0.0001	0.1411	0.0010	
172	Embryo	CG8857	Yes	Yes	Identifier	No	Yes	Non-modulated	0	5	0.0001	0.0653	0.0010	
173	Embryo	CG3751	Yes	-	Identifier	No	Yes	Non-modulated	0	2	0.0001	0.1379	0.0010	
174	Embryo	alphaTub84B	No	-	Real name	No	Yes	Non-modulated	12	2	0.0002	0.1582	0.0010	
175	Embryo	CG11522	No	-	Identifier	No	Yes	Non-modulated	0	3	0.0103	0.1842	0.0561	
176	Embryo	NHP2	No	-	Real name	No	Yes	Non-modulated	1	2	0.0081	0.2280	0.0355	
177	Embryo	dhd	No	-	Real name	No	Yes	Embryo	5	1	0.0000	0.0143	0.0010	H0 not rejected
178	Embryo	RpP2	No	-	Real name	No	Yes	Non-modulated	1	2	0.0076	0.1627	0.0469	
179	Embryo	CG2099	Yes	-	Identifier	No	Yes	Non-modulated	0	4	0.0221	0.2901	0.0761	
180	Embryo	Ef2b	No	-	Real name	No	Yes	Non-modulated	5	6	0.0055	0.3427	0.0161	
181	Embryo	RpL27a	Yes	-	Real name	No	Yes	Non-modulated	0	6	0.0001	0.1304	0.0010	
182	Embryo	qm	No	-	Real name	No	Yes	Non-modulated	0	7	0.0020	0.2999	0.0067	
183	Embryo	CG15697	Yes	Yes	Identifier	No	Yes	Non-modulated	0	3	0.0064	0.0551	0.1165	
184	Embryo	HmgD	No	-	Real name	No	Yes	Non-modulated	4	2	0.0037	0.1478	0.0253	
185	Embryo	RpS9	Yes	-	Real name	No	Yes	Non-modulated	0	2	0.0001	0.0765	0.0010	
186	Embryo	CG4800	Yes	-	Identifier	No	Yes	Non-modulated	0	1	0.0111	0.4122	0.0270	
187	Embryo	CG3278	No	-	Identifier	No	Yes	Embryo	0	4	0.0404	0.3103	0.1301	
188	Embryo	CG4046	Yes	-	Identifier	No	Yes	Non-modulated	0	5	0.0000	0.0436	0.0010	
189	Embryo	RpS15A, CG2033	No	-	Real name	No	Yes	Non-modulated	1	2	0.0004	0.3521	0.0010	
190	Embryo	CG18001	Yes	Yes	Identifier	No	Yes	Non-modulated	0	1	0.0004	0.3571	0.0010	
191	Embryo	RpL15	Yes	-	Real name	No	Yes	Non-modulated	0	2	0.0003	0.3088	0.0010	
192	Embryo	RpL18A	Yes	-	Real name	No	Yes	Non-modulated	0	2	0.0001	0.1263	0.0010	
193	Embryo	CG12775	Yes	Yes	Identifier	No	Yes	Non-modulated	0	3	0.0033	0.2662	0.0126	
194	Embryo	RpS4	No	-	Real name	No	Yes	Non-modulated	0	6	0.0028	0.0618	0.0449	
195	Embryo	RpS3A	Yes	-	Real name	No	Yes	Non-modulated	0	2	0.0017	0.2522	0.0068	
196	Embryo	yip6	Yes	-	Real name	No	Yes	Non-modulated	0	4	0.0003	0.3171	0.0010	
197	Embryo	sta	Yes	Yes	Real name	No	Yes	Non-modulated	0	3	0.0003	0.2633	0.0010	
198	Embryo	CG3195	No	-	Identifier	No	Yes	Non-modulated	0	4	0.0001	0.0600	0.0010	
199	Embryo	tsr	Yes	-	Real name	No	Yes	Non-modulated	1	4	0.0002	0.1670	0.0010	
200	Embryo	sop	No	-	Real name	No	Yes	Non-modulated	0	2	0.0081	0.1113	0.0728	

No.	Library	Gene name	Duplicates	Dup. excl.	Name	Orphan	*Anopheles*	Expression	Paralogues	Exons	*dN*	*dS*	*dN/dS*	H_0: *dN/dS*=1
201	Embryo	CG10423	Yes	Yes	Identifier	No	Yes	Non-modulated	0	3	0.0001	0.0922	0.0010	
202	Embryo	CG7808	No	-	Identifier	No	Yes	Non-modulated	0	4	0.0046	0.2004	0.0228	
203	Embryo	RpS26	No	-	Real name	No	Yes	Non-modulated	0	2	0.0004	0.4348	0.0010	
204	Embryo	RpL13	Yes	-	Real name	No	Yes	Non-modulated	0	3	0.0022	0.1515	0.0143	
205	Embryo	CG8415	Yes	Yes	Identifier	No	Yes	Non-modulated	0	5	0.0001	0.0597	0.0010	
206	Embryo	CHRAC-14	No	-	Real name	No	Yes	Embryo	0	1	0.0165	0.2173	0.0758	
207	Embryo	cdc2c	No	-	Real name	No	Yes	Non-modulated	165	6	0.0004	0.4116	0.0010	
208	Embryo	CG7163	No	-	Identifier	No	Yes	Non-modulated	2	3	0.1224	0.3453	0.3546	
209	Embryo	CG8444	No	-	Identifier	No	Yes	Non-modulated	0	3	0.0003	0.3201	0.0010	
210	Embryo	CG1866	No	-	Identifier	No	Yes	Non-modulated	13	3	0.0578	0.3241	0.1783	
211	Embryo	CG15141	No	-	Identifier	No	Yes	Non-modulated	0	2	-	-	-	
212	Embryo	CG10702	No	-	Identifier	No	Yes	Non-modulated	3	6	0.0437	0.2641	0.1654	
213	Embryo	CG8309	Yes	-	Identifier	No	Yes	Non-modulated	0	3	0.0019	0.1656	0.0116	
214	Embryo	CG4394	No	-	Identifier	No	Yes	Embryo	1	5	0.0038	0.4262	0.0089	
215	Embryo	SsRbeta	No	-	Real name	No	Yes	Non-modulated	0	4	0.0139	0.2098	0.0665	
216	Embryo	CG6583	No	-	Identifier	Yes	Yes	Embryo	0	3	0.0066	0.2678	0.0245	
217	Embryo	GM130	No	-	Real name	No	Yes	Non-modulated	52	5	0.0334	0.3084	0.1083	
218	Embryo	Dph5	No	-	Real name	No	Yes	Non-modulated	0	7	0.0002	0.1540	0.0010	
219	Embryo	Pros35	No	-	Real name	No	Yes	Non-modulated	11	3	0.0251	0.3280	0.0766	
220	Embryo	BcDNA:LD08534	No	-	Identifier	No	Yes	Non-modulated	0	1	0.0025	0.3545	0.0072	
221	Embryo	EG:34F3.8	No	-	Identifier	No	Yes	Non-modulated	1	4	0.0046	0.3569	0.0128	
222	Embryo	CG4949	No	-	Identifier	No	No	Embryo	0	2	0.0132	0.1900	0.0696	
223	Embryo	Rapgap1	No	-	Real name	No	Yes	Non-modulated	3	4	0.0028	0.1953	0.0143	
224	Embryo	sqh	Yes	Yes	Real name	No	Yes	Non-modulated	15	3	0.0002	0.2292	0.0010	
225	Embryo	wdn	No	-	Real name	No	Yes	Non-modulated	0	3	0.0087	0.2047	0.0426	
226	Embryo	maf-S	No	-	Real name	No	Yes	Non-modulated	1	2	0.0003	0.3406	0.0010	
227	Embryo	EG:22E5.4	No	-	Identifier	No	Yes	Embryo	0	4	0.0284	0.4195	0.0677	
228	Embryo	Nxt1	No	-	Real name	No	Yes	Embryo	0	3	0.0604	0.7265	0.0831	
229	Embryo	CG10731	No	-	Identifier	No	Yes	Non-modulated	0	1	0.0219	0.3287	0.0666	
230	Embryo	CG14043	No	-	Identifier	No	Yes	Non-modulated	0	1	0.0192	0.4079	0.0471	
231	Embryo	Aats-val	No	-	Real name	No	Yes	Non-modulated	6	5	0.0064	0.1885	0.0340	
232	Embryo	CG13878	No	-	Identifier	Yes	No	Non-modulated	0	3	-	-	-	

No.	Library	Gene name	Duplicates	Dup. excl.	Name	Orphan	*Anopheles*	Expression	Paralogues	Exons	*dN*	*dS*	*dN/dS*	H_0: *dN/dS*=1
233	Embryo	CG11100	No	-	Identifier	Yes	Yes	Embryo	1	3	0.0172	0.2509	0.0687	
234	Embryo	CG5064	No	-	Identifier	No	Yes	Non-modulated	0	4	0.0266	0.3256	0.0817	
235	Embryo	CG1240	No	-	Identifier	No	Yes	Non-modulated	0	2	0.0030	0.4342	0.0070	
236	Embryo	CG5879	No	-	Identifier	No	Yes	Non-modulated	47	5	0.0002	0.1941	0.0010	
237	Embryo	CG6488	No	-	Identifier	No	Yes	Non-modulated	0	5	0.0080	0.3463	0.0231	
238	Embryo	C3G	No	-	Real name	No	Yes	Non-modulated	7	13	0.0002	0.1885	0.0010	
239	Embryo	CG13770	No	-	Identifier	No	Yes	Non-modulated	22	6	0.1135	0.4876	0.2328	
240	Embryo	CG8031	No	-	Identifier	No	Yes	Non-modulated	0	6	0.0139	0.1893	0.0735	
241	Embryo	CG11357	No	-	Identifier	No	Yes	Embryo	4	4	0.0433	0.1245	0.3476	H0 not rejected
242	Embryo	CG18609	No	-	Identifier	No	Yes	Non-modulated	15	3	0.0365	0.5638	0.0647	
243	Embryo	Jhl-26 ,CG3767	No	-	Real name	No	Yes	Non-modulated	18	3	0.0802	0.3700	0.2168	
244	Embryo	CG15012	No	-	Identifier	No	Yes	Non-modulated	0	4	0.0181	0.2374	0.0764	
245	Embryo	cta	Yes	Yes	Real name	No	Yes	Non-modulated	0	5	0.0168	0.2473	0.0681	
246	Embryo	xl6	No	-	Real name	No	Yes	Non-modulated	10	2	-	-	-	
247	Embryo	CG11583	No	-	Identifier	No	Yes	Non-modulated	0	3	0.0047	0.3299	0.0144	
248	Embryo	CG12792	No	-	Identifier	No	Yes	Embryo	13	1	0.0326	0.3945	0.0827	
249	Embryo	CG8326	No	-	Identifier	No	Yes	Embryo	0	2	0.0232	0.2455	0.0945	
250	Embryo	CG9894	Yes	-	Identifier	No	No	Non-modulated	1	4	0.0000	0.0317	0.0010	
251	Embryo	CG12384	No	-	Identifier	No	Yes	Non-modulated	0	4	0.0111	0.1666	0.0665	
252	Embryo	CG12054	No	-	Identifier	No	Yes	Non-modulated	28	8	0.0170	0.1641	0.1033	
253	Embryo	Bap60	No	-	Real name	No	Yes	Non-modulated	0	3	0.0072	0.3260	0.0220	
254	Embryo	CG4645	No	-	Identifier	No	Yes	Non-modulated	0	3	0.0240	0.3384	0.0709	
255	Embryo	mxc	No	-	Real name	No	Yes	Embryo	1	2	0.0498	0.2967	0.1680	
256	Embryo	CG8830	No	-	Identifier	No	Yes	Non-modulated	9	4	0.0389	0.2510	0.1551	
257	Embryo	SrpRbeta	No	-	Real name	No	Yes	Non-modulated	0	5	0.0158	0.2415	0.0655	
258	Embryo	CG6410	No	-	Identifier	No	Yes	Non-modulated	0	3	0.0112	0.4535	0.0246	
259	Embryo	CG8332	Yes	-	Identifier	No	Yes	Non-modulated	0	4	0.0002	0.1721	0.0010	
260	Embryo	CG7777	No	-	Identifier	No	Yes	Non-modulated	5	6	0.0254	0.2988	0.0851	
261	Embryo	Mlp60A	No	-	Real name	No	Yes	Non-modulated	4	9	0.0000	0.0161	0.0010	H0 not rejected
262	Embryo	CG3983	No	-	Identifier	No	Yes	Non-modulated	4	5	0.0002	0.2325	0.0010	
263	Embryo	CG17524	No	-	Identifier	No	Yes	Non-modulated	27	1	0.0354	0.4188	0.0845	
264	Embryo	CG10855	No	-	Identifier	No	Yes	Non-modulated	0	3	-	-	-	

No.	Library	Gene name	Duplicates	Dup. excl.	Name	Orphan	*Anopheles*	Expression	Paralogues	Exons	*dN*	*dS*	*dN/dS*	H_0: *dN/dS*=1
265	Embryo	CG2943	No	-	Identifier	No	Yes	Non-modulated	0	13	0.0070	0.2489	0.0283	
266	Embryo	CG9849	No	-	Identifier	No	Yes	Non-modulated	0	2	0.0157	0.3271	0.0479	
267	Embryo	ci	No	-	Real name	No	Yes	Non-modulated	105	6	0.0192	0.2470	0.0775	
268	Embryo	Hmu	No	-	Real name	No	Yes	Non-modulated	38	4	-	-	-	
269	Embryo	CG9238	No	-	Identifier	No	Yes	Non-modulated	1	5	-	-	-	
270	Embryo	CG2852	Yes	Yes	Identifier	No	Yes	Non-modulated	14	3	0.0003	0.2560	0.0010	
271	Embryo	RpS13	Yes	Yes	Real name	No	Yes	Non-modulated	0	3	0.0055	0.1242	0.0443	
272	Embryo	eIF-1A	No	-	Real name	No	Yes	Non-modulated	0	3	0.0001	0.1225	0.0010	
273	Embryo	mod(mdg4)58.8	No	-	Real name	No	Yes	Non-modulated	0	3	0.0177	0.6934	0.0256	
274	Embryo	alpha-Adaptin	No	-	Real name	No	Yes	Non-modulated	4	9	-	-	-	
275	Embryo	lap2	No	-	Real name	No	Yes	Non-modulated	3	3	0.0105	0.4936	0.0213	
276	Embryo	CG1598	No	-	Identifier	No	Yes	Non-modulated	0	4	0.0029	0.3700	0.0077	
277	Embryo	RpL14	Yes	Yes	Real name	No	Yes	Non-modulated	0	4	0.0025	0.0563	0.0446	
278	Embryo	CG3305	Yes	Yes	Identifier	No	Yes	Non-modulated	0	4	-	-	-	
279	Embryo	CG3939, EG:140G11.5	No	-	Identifier	No	Yes	Non-modulated	1	2	0.0412	0.3318	0.1242	
280	Embryo	RN-tre	No	-	Real name	No	Yes	Non-modulated	0	5	0.0056	0.1344	0.0419	
281	Embryo	CG13339	No	-	Identifier	Yes	Yes	Embryo	0	2	0.0501	0.3560	0.1408	
282	Embryo	BcDNA:GH02435	No	-	Identifier	No	Yes	Non-modulated	21	2	0.0071	0.2891	0.0244	
283	Embryo	eIF-4a	No	-	Real name	No	Yes	Non-modulated	33	5	0.0062	0.1782	0.0348	
284	Embryo	CG7706	No	-	Identifier	No	Yes	Embryo	0	3	0.0506	0.2777	0.1822	
285	Embryo	CG1696	No	-	Identifier	No	Yes	Non-modulated	3	4	0.0002	0.2044	0.0010	
286	Embryo	RpL32	Yes	-	Real name	No	Yes	Non-modulated	0	2	0.0001	0.1222	0.0010	
287	Embryo	CG32775-PA	No	-	Identifier	No	Yes	Non-modulated	4	2	0.0092	0.2742	0.0335	
288	Embryo	S6kII	No	-	Real name	No	Yes	Non-modulated	170	2	0.0102	0.3146	0.0326	
289	Embryo	skpA	No	-	Real name	No	Yes	Non-modulated	6	2	0.0050	0.8999	0.0056	
290	Embryo	ken	No	-	Real name	No	Yes	Non-modulated	32	3	0.0060	0.3026	0.0197	
291	Embryo	His1	No	-	Real name	No	Yes	Non-modulated	0	1	-	-	-	
292	Embryo	RpL17A	No	-	Real name	No	Yes	Non-modulated	0	3	0.0001	0.0823	0.0010	
293	Embryo	CG9795	No	-	Identifier	Yes	Yes	Non-modulated	0	10	0.0877	0.3532	0.2483	
294	Embryo	CG1316	No	-	Identifier	No	Yes	Non-modulated	3	8	0.0053	0.3524	0.0151	
295	Embryo	SamDC	No	-	Real name	No	Yes	Non-modulated	0	7	0.0192	0.3881	0.0496	
296	Embryo	CG12338	No	-	Identifier	No	Yes	Non-modulated	1	5	0.0298	0.5304	0.0562	

No.	Library	Gene name	Duplicates	Dup. excl.	Name	Orphan	*Anopheles*	Expression	Paralogues	Exons	*dN*	*dS*	*dN/dS*	H_0: *dN/dS*=1
297	Embryo	CG17493	No	-	Identifier	No	Yes	Non-modulated	22	2	0.0185	0.3175	0.0583	
298	Embryo	CG9342	No	-	Identifier	No	Yes	Non-modulated	0	5	0.0161	0.2442	0.0661	
299	Embryo	eff	No	-	Real name	No	Yes	Non-modulated	24	5	0.0000	0.0201	0.0010	
300	Embryo	CG10217	No	-	Identifier	No	Yes	Non-modulated	1	4	0.0046	0.3013	0.0152	
301	Embryo	CG2915	No	-	Identifier	No	Yes	Non-modulated	18	3	0.0202	0.3288	0.0615	
302	Embryo	Tom, anon-fast-evolving-1F6	No	-	Real name	Yes	Yes	Embryo	2	1	0.0001	0.1069	0.0010	
303	Embryo	CG5220	No	-	Identifier	No	Yes	Non-modulated	3	3	0.0049	0.2687	0.0181	
304	Embryo	CG11665	No	-	Identifier	No	Yes	Non-modulated	6	1	0.0040	0.1499	0.0265	
305	Embryo	Cap, CG9748	No	-	Real name	No	Yes	Non-modulated	61	4	0.0002	0.2396	0.0010	
306	Embryo	Elongin-C	Yes	Yes	Real name	No	Yes	Non-modulated	0	2	0.0000	0.0275	0.0010	H0 not rejected
307	Embryo	CG18145	No	-	Identifier	Yes	Yes	Embryo	0	4	0.0276	0.3263	0.0847	
308	Embryo	CG14639	No	-	Identifier	Yes	Yes	Non-modulated	18	2	0.0366	0.4761	0.0770	
309	Embryo	CG13089	No	-	Identifier	No	Yes	Non-modulated	0	4	0.0320	0.2641	0.1210	
310	Embryo	BcDNA:GH04753	No	-	Real name	No	Yes	Non-modulated	27	3	0.0050	0.1642	0.0303	
311	Embryo	rab1	Yes	Yes	Real name	No	Yes	Non-modulated	52	5	0.0040	0.1633	0.0242	
312	Embryo	Auxilin, CG1107	No	-	Identifier	No	Yes	Non-modulated	51	11	0.0147	0.2284	0.0644	
313	Embryo	CG2046	No	-	Identifier	Yes	Yes	Non-modulated	0	3	0.0608	0.3100	0.1963	
314	Embryo	fh	No	-	Real name	No	Yes	Non-modulated	0	2	0.0707	0.2727	0.2591	
315	Embryo	CG10038	No	-	Identifier	No	Yes	Embryo	0	5	0.0235	0.3523	0.0668	
316	Embryo	CG1471	No	-	Identifier	No	Yes	Non-modulated	0	6	0.0080	0.3209	0.0248	
317	Embryo	Cyp1	Yes	Yes	Real name	No	Yes	Non-modulated	16	2	0.0029	0.2136	0.0136	
318	Embryo	bic	Yes	Yes	Real name	No	Yes	Non-modulated	2	2	0.0287	0.1103	0.2606	H0 not rejected
319	Embryo	CG9358	No	-	Identifier	No	Yes	Embryo	2	1	0.1091	0.6691	0.1631	
320	Embryo	CG1883	Yes	-	Identifier	No	Yes	Non-modulated	0	4	0.0021	0.1005	0.0208	
321	Embryo	sesB	No	-	Real name	No	Yes	Non-modulated	27	4	0.0062	0.0453	0.1359	
322	Embryo	CG1746	No	-	Identifier	No	Yes	Non-modulated	0	3	0.0069	0.0226	0.3043	H0 not rejected
323	Embryo	mago	Yes	Yes	Real name	No	Yes	Non-modulated	0	2	0.0003	0.2663	0.0010	
324	Embryo	BG:DS08249.4	No	-	Identifier	Yes	No	Embryo	0	3	0.0701	0.1622	0.4322	
325	Embryo	Aats-thr, CG5353	No	-	Identifier	No	Yes	Non-modulated	1	4	0.0108	0.2224	0.0485	
326	Embryo	CG11738	No	-	Identifier	No	Yes	Embryo	0	1	0.0035	0.8946	0.0040	
327	Embryo	CG10799	Yes	Yes	Identifier	Yes	No	Non-modulated	0	1	0.3247	0.5657	0.5740	
328	Embryo	CG6724	No	-	Identifier	No	Yes	Non-modulated	27	4	0.0068	0.3764	0.0182	

No.	Library	Gene name	Duplicates	Dup. excl.	Name	Orphan	*Anopheles*	Expression	Paralogues	Exons	*dN*	*dS*	*dN/dS*	H_0: *dN/dS*=1
329	Embryo	CG7048	No	-	Identifier	No	Yes	Non-modulated	0	3	0.0321	0.2548	0.1260	
330	Embryo	ATPsyn-gamma	Yes	Yes	Real name	No	Yes	Non-modulated	0	2	0.0001	0.0925	0.0010	
331	Embryo	Pk17E	No	-	Real name	No	Yes	Non-modulated	125	9	0.0132	0.2640	0.0498	
332	Embryo	icln	No	-	Real name	No	Yes	Embryo	0	2	0.0183	0.3524	0.0518	
333	Embryo	CG6736	Yes	Yes	Identifier	No	No	Non-modulated	0	2	0.0098	0.1350	0.0729	
334	Embryo	CG8580	No	-	Identifier	No	Yes	Non-modulated	0	4	0.0111	0.3279	0.0340	
335	Embryo	CG13626	No	-	Identifier	No	Yes	Non-modulated	0	2	0.0067	0.2950	0.0227	
336	Embryo	CG9324	No	-	Identifier	No	Yes	Non-modulated	0	3	0.0241	0.2276	0.1061	
337	Embryo	EG:25E8.4	No	-	Identifier	Yes	Yes	Non-modulated	0	2	0.0193	0.2531	0.0763	
338	Embryo	aly	No	-	Real name	No	Yes	Non-modulated	1	2	0.0204	0.1766	0.1157	
339	Embryo	Pp2A-29B	No	-	Real name	No	Yes	Non-modulated	0	7	0.0028	0.1543	0.0183	
340	Embryo	Iola	No	-	Real name	No	Yes	Non-modulated	0	1	0.0205	0.2636	0.0776	
341	Embryo	CG10960	No	-	Identifier	No	Yes	Non-modulated	26	4	0.0516	0.3433	0.1503	
342	Embryo	CG8097	No	-	Identifier	No	Yes	Non-modulated	0	3	0.1691	0.3310	0.5108	
343	Embryo	CG15481	No	-	Identifier	No	Yes	Non-modulated	1	1	-	-	-	
344	Embryo	mod(mdg4)55.3	No	-	Real name	No	Yes	Non-modulated	0	5	0.0213	0.3812	0.0559	
345	Embryo	Mlf	Yes	Yes	Real name	No	Yes	Non-modulated	0	4	-	-	-	
346	Embryo	CG6478	No	-	Identifier	No	No	Embryo	26	2	0.0567	0.5754	0.0985	
347	Embryo	RpL23a	Yes	Yes	Real name	No	Yes	Non-modulated	1	3	0.0313	0.0596	0.5247	H0 not rejected
348	Embryo	BcDNA:LD21969	No	-	Identifier	No	No	Non-modulated	1	3	0.0723	0.3482	0.2077	
349	Embryo	CG12750	No	-	Identifier	No	Yes	Non-modulated	3	4	0.0929	0.3806	0.2441	
350	Embryo	CG9188	No	-	Identifier	Yes	No	Non-modulated	0	3	0.0697	0.3511	0.1985	
351	Embryo	Fkbp13	No	-	Real name	No	Yes	Non-modulated	4	5	0.0075	0.3478	0.0215	
352	Embryo	kin17	No	-	Real name	No	Yes	Non-modulated	0	2	0.0022	0.4055	0.0055	
353	Embryo	GSTD1	No	-	Real name	No	Yes	Non-modulated	26	2	0.0155	0.2204	0.0704	
354	Embryo	CG13068	No	-	Identifier	No	No	Embryo	9	2	0.0209	0.3620	0.0577	
355	Embryo	CG17471	No	-	Identifier	No	Yes	Non-modulated	3	9	0.0030	0.2416	0.0124	
356	Embryo	CG16868	No	-	Identifier	No	Yes	Non-modulated	0	6	0.0113	0.1840	0.0612	
357	Embryo	CG6249	No	-	Identifier	No	Yes	Non-modulated	0	2	0.0228	0.3380	0.0674	
358	Embryo	Arf102F	No	-	Real name	No	Yes	Non-modulated	15	4	0.0078	0.3312	0.0237	
359	Embryo	Hsp67Bc	No	-	Real name	No	Yes	Embryo	9	1	0.0109	0.2281	0.0476	
360	Embryo	CG5171	No	-	Identifier	No	Yes	Non-modulated	2	4	0.0290	0.3985	0.0729	

No.	Library	Gene name	Duplicates	Dup. excl.	Name	Orphan	*Anopheles*	Expression	Paralogues	Exons	*dN*	*dS*	*dN/dS*	H_0: *dN/dS*=1
361	Embryo	CG13298	No	-	Identifier	No	Yes	Non-modulated	0	3	0.0003	0.2564	0.0010	
362	Embryo	CG7956	No	-	Identifier	No	Yes	Non-modulated	4	9	0.0021	0.2619	0.0082	
363	Embryo	CG8029	No	-	Identifier	No	Yes	Non-modulated	1	4	0.0306	0.3381	0.0905	
364	Embryo	CG9762	Yes	Yes	Identifier	No	Yes	Non-modulated	0	4	0.0225	0.1892	0.1187	
365	Embryo	Pka-C1	No	-	Real name	No	Yes	Non-modulated	156	3	0.0001	0.1378	0.0010	
366	Embryo	dom	No	-	Real name	No	Yes	Non-modulated	55	15	0.0129	0.2790	0.0461	
367	Embryo	CG14235	Yes	Yes	Identifier	No	Yes	Non-modulated	0	2	0.0001	0.0718	0.0010	
368	Embryo	smg	No	-	Real name	No	Yes	Non-modulated	0	5	0.0352	0.1412	0.2496	
369	Embryo	Tm2	No	-	Real name	No	Yes	Non-modulated	18	3	0.0001	0.0936	0.0010	
370	Embryo	CG13043	No	-	Identifier	Yes	Yes	Non-modulated	8	1	0.0037	0.1443	0.0259	
371	Embryo	Fib	No	-	Real name	No	Yes	Non-modulated	76	3	0.0002	0.2309	0.0010	
372	Embryo	Syb	Yes	-	Real name	No	Yes	Non-modulated	2	4	0.0003	0.2802	0.0010	
373	Embryo	CG10084	No	-	Identifier	No	Yes	Non-modulated	0	6	0.0146	0.2531	0.0576	
374	Embryo	awd	No	-	Real name	No	Yes	Non-modulated	0	2	0.0001	0.1490	0.0010	
375	Embryo	CG11015	Yes	Yes	Identifier	No	Yes	Non-modulated	1	3	0.0075	0.1196	0.0625	
376	Embryo	CG6398	Yes	Yes	Identifier	No	Yes	Non-modulated	0	5	0.0002	0.1528	0.0010	
377	Embryo	ifc	No	-	Real name	No	Yes	Embryo	0	2	0.0003	0.3141	0.0010	
378	Embryo	CG6803	No	-	Identifier	Yes	Yes	Non-modulated	0	9	0.0068	0.0746	0.0908	
379	Embryo	CG12163	No	-	Identifier	No	Yes	Non-modulated	11	2	0.0237	0.3761	0.0631	
380	Embryo	methyl-CpG-binding-domain-like	No	-	Real name	No	Yes	Non-modulated	0	3	0.0066	0.2425	0.0274	
381	Embryo	CG3907	No	-	Identifier	No	Yes	Non-modulated	0	7	0.0229	0.1579	0.1453	
382	Embryo	CG4882	No	-	Identifier	No	Yes	Non-modulated	0	3	-	-	-	
383	Embryo	AP-2sigma	No	-	Real name	No	Yes	Non-modulated	2	4	0.0003	0.2868	0.0010	
384	Embryo	Aats-asp	No	-	Real name	No	Yes	Non-modulated	4	5	0.0133	0.3013	0.0441	
385	Embryo	CG8583	No	-	Identifier	No	Yes	Non-modulated	0	4	0.0146	0.2414	0.0603	
386	Embryo	emb	No	-	Real name	No	Yes	Non-modulated	0	7	0.0052	0.2599	0.0201	
387	Embryo	Pabp2	No	-	Real name	No	Yes	Non-modulated	0	6	0.0183	0.4109	0.0444	
388	Embryo	CG8498	No	-	Identifier	No	Yes	Embryo	7	3	0.0138	0.3327	0.0415	
389	Embryo	CG18178	No	-	Identifier	Yes	No	Non-modulated	0	2	0.0543	0.4706	0.1155	
390	Embryo	CG10473	No	-	Identifier	No	Yes	Non-modulated	12	6	0.0101	0.1856	0.0545	
391	Embryo	CG1157	No	-	Identifier	Yes	Yes	Embryo	1	3	0.0178	0.4288	0.0415	

No.	Library	Gene name	Duplicates	Dup. excl.	Name	Orphan	*Anopheles*	Expression	Paralogues	Exons	*dN*	*dS*	*dN/dS*	H_0: *dN/dS*=1
392	Embryo	CG2998	Yes	-	Identifier	No	Yes	Non-modulated	1	2	0.0001	0.0982	0.0010	
393	Embryo	CG9410	No	-	Identifier	No	Yes	Non-modulated	0	3	0.0157	0.2602	0.0604	
394	Embryo	Mlc1	No	-	Real name	No	Yes	Non-modulated	1	5	0.0078	0.0681	0.1149	
395	Embryo	Fer1HCH	Yes	Yes	Real name	No	Yes	Non-modulated	1	3	0.0123	0.2252	0.0545	
396	Embryo	CG17385	No	-	Identifier	No	Yes	Embryo	196	3	0.0022	0.4254	0.0052	
397	Embryo	CG7787	Yes	-	Identifier	No	Yes	Non-modulated	0	2	0.0130	0.4389	0.0297	
398	Embryo	CG12091	No	-	Identifier	No	Yes	Non-modulated	2	1	0.0020	0.2711	0.0074	
399	Embryo	CG1939	No	-	Identifier	No	Yes	Non-modulated	1	3	0.0120	0.4776	0.0252	
400	Embryo	CG15022	No	-	Identifier	No	No	Embryo	27	2	-	-	-	
401	Embryo	CG6182	No	-	Identifier	No	Yes	Embryo	0	1	0.0208	0.5337	0.0390	
402	Embryo	ATPsyn-beta	Yes	Yes	Real name	No	Yes	Non-modulated	9	3	0.0023	0.3089	0.0076	
403	Embryo	CG14915	No	-	Identifier	Yes	No	Embryo	0	1	0.0278	0.2884	0.0963	

Table 20. Overview of *D. yakuba* cDNA clones. Library: cDNA library from which a gene was recovered. Gene Name: Fly Base gene name. Duplicates: 81 genes are present in the embryo and adult library. Dup. excl.: One gene form a duplicate pair had been excluded when complete sample was analyzed. Name: If a gene was studied previously real name is assigned otherwise annotation identifier is given. Orphan: If a gene has no BLAST match outside insects (E-value cutoff 0.0001) and lacks protein domains it is considered to be an orphan. Anopheles: Sequence similarity of a given gene in the Anopheles genome (TBLASTN, E-value cutoff 0.001). Expression: Stage specific expression for a given gene. Paralogs: Number of paralogs for a given gene (BLASTP E-value cutoff 1x 10^{-10}). Exons: Number of exons of *D. melanogaster* orthologue *dN*: non-synonymous substitution rate calculated by the maximum likelihood method (see material and methods). *dS*: synonymous substitution rate calculated by the maximum likelihood method (see material and methods). *dN/dS* – ratio of *dN* and *dS*. H_0*: dN/dS=1:* Genes for which null-hypothesis that *dN* and *dS* are equal was not possible to reject.

Erklärung

Ich versichere, daß ich die von mir vorgelegte Dissertation selbständig angefertigt, die benutzten Quellen und Hilfsmittel vollständig angegeben und die Stellen der Arbeit - einschließlich Tabellen, Karten und Abbildungen -, die anderen Werken im Wortlaut oder dem Sinn nach entnommen sind, in jedem Einzelfall als Entlehnung kenntlich gemacht habe; daß diese Dissertation noch keiner anderen Fakultät oder Universität zur Prüfung vorgelegen hat; daß sie - abgesehen von unten angegebenen Teilpublikationen - noch nicht veröffentlicht worden ist sowie, daß ich eine solche Veröffentlichung vor Abschluß des Promotionsverfahrens nicht vornehmen werde. Die Bestimmungen dieser Promotionsordnung sind mir bekannt. Die von mir vorgelegte Dissertation ist von Prof. Dr. Diethard Tautz betreut worden.

Teilpublikationen:

Tomislav Domazet-Lošo and Diethard Tautz (2003) An evolutionary analysis of orphan genes in *Drosophila.* Genome Research *(in review)*

Gabriel Marais, Tomislav Domazet-Lošo, Brian Charlesworth and Diethard Tautz. Expression level, recombination and the evolutionary rates in *Drosophila. (in preparation)*

Datum: Unterschrift:

Lebenslauf

Tomislav Domazet-Lošo
Rennebergstr. 1
D-50931 Köln

Name	Tomislav Domazet-Lošo
Geburtsdatum/-ort	31.01.1974 in Split, Kroatien
Staatsangehörigkeit	kroatisch
Familienstand	ledig
Eltern	Davor und Jasminka Domazet-Lošo

1980 - 1988	Grundschule in Split, Kroatien
1988 - 1992	Mathematisch-Naturwissenschaftliches Gymnasium in Split, Kroatien
19. Juni 1992	Erfolgreicher Abschluß der Abiturprüfung
1992 - 1997	Studium der Biologie an der Universität Zagreb, Kroatien
30. September 1997	Diplom in Biologie, Universität Zagreb, Kroatien
1997 - 1998	9 monatige Leistung des Grundwehrdienstes (Kroatien)
1998 - 2000	Wissenschaftlicher Mitarbeiter am Institut “Ruđer Bošković”, Abteilung für Molekular Genetik, Zagreb, Kroatien
2000 - 2003	Doktorarbeit am Lehrstuhl für Evolutionsgenetik bei Prof. Tautz an der Universität zu Köln. Titel: “Evolution of orphan genes in *Drosophila*”
2003	Voraussichtlicher Abschluß der Promotion