A genomic approach to the study of *Tribolium castaneum* genetics, development & evolution

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

Joël Savard

aus Hauterive, Kanada

(Köln, 2004)

Berichterstatter:

Prof. Dr. Diethard Tautz

Prof. Dr. Thomas Wiehe

Tag der mündlichen Prüfung: 4. Februar 2004

Table of contents

Acknowledgments

I am particularly grateful to my supervisor Prof. Dr. Diethard Tautz for giving me the opportunity to explore with him the different avenues of genomics, development and evolution, and for guiding me through these paths. I am also grateful to him for continually providing me with incredible scientific opportunities through which I could learn, create and express myself freely. I would like to thank Prof. Dr. Thomas Wiehe, Prof. Dr. Siegfried Roth and Dr. Wim Damen for serving on my thesis committee.

The work I did during the last years was part of a larger collaborative effort. With great pleasure, I thank our German and American *Tribolium* collaborators: Prof. Dr. Martin Klingler, Dr. Susan J. Brown and Dr. Richard Beeman. I especially want to thank Prof. Dr. Martin Klingler and the members of his lab for teaching me the basics of working with *Tribolium* and for always being there to answer my questions. I am grateful to Dr. Susan J. Brown who provided me with BAC DNA, and to Dr. Jonathan Margolis and Exelixis Inc. who gave me access to their *Tribolium* EST database and provided us with the BAC library. Many thanks also to Dr. Martin Lercher for highly estimated collaboration on diverse projects, helpful discussions and critical reading of my manuscripts.

Several people from the lab also helped me at various stages of my project. Very special thanks to Vladimir Simović who worked with me on the *in situ* screen and during the BAC-ends sequencing project. All of this would have not been possible without his help. My thanks also goes to Susanne Krächter and Karin Otto for helping me during the EST and BAC-ends sequencing projects, to Dr. Martin Gajewski and Manuel Aranda for technical advices and to Alexander Pozhitkov for helping me to resolve Linux-related issues.

I would like to express my gratitude to Dr. Heidi Fußwinkel who helped me resolving complex administrative issues and to Eva Siegmund who was always there to answer my multiple questions. I am grateful to Susanne Krächter and Hilary Dove who helped me to solve several professional as well as private issues.

Some of you were always there to discuss scientific theories and to support me in my work. Thanks to Prof. Dr. Siegfied Roth, Dr. Wim Damen and Nicolas-Michael Prpić for very interesting Evo-Devo discussions. Also many thanks to Arne Nolte for sharing with me his contagious passion for biology. This was very much appreciated.

This work was supported by a grant from the HSFP (RG0303).

In conclusion, I would like to thank my parents and my family, who supported me through all the stages of my studies. I am also grateful to my friends, especially Patrycja Niewiadomski and Yannick Cornet, for their continual encouragement and support.

Abbreviations

BAC — bacterial artificial chromosome

BDGP — Berkeley *Drosophila* Genome Project

BES — BAC-end sequences

BLAST — basic local alignment search tool

bp — base pair

cDNA — complementary DNA

cM — centiMorgan

DNA — deoxyribonucleic acid

EST — expressed sequence tag

GSP — gene specific primer

HGP — Human Genome Project

kb — kilobase pair

Mb — megabase pair

mRNA — messenger RNA

My — million years

Mya — million years ago

NCBI — National Centre for Biotechnology Information

NHGRI — National Human Genome Research Institute

ORF — open reading frame

pRNAi — parental RNA interference

rDNA — ribosomal DNA

RNA — ribonucleic acid

RNAi — RNA interference

rRNA — ribosomal RNA

TGD — *Tribolium* Genome Database

UTR — untranslated region

WGS — whole-genome shotgun

Zf — zinc finger

Zusammenfassung

Während der letzten zehn Jahre ist *Tribolium castaneum* das Insekt der Wahl der vergleichenden Genetik und Entwiklungsbiologie ausserhalb der Drosophiliden geworden. Bis heute sind die meisten molekularen Studien auf die Segmentierung und die homeotischen Gene fokussiert. Um unabhängiges Wissen über die genetische Basis der Insektenentwicklung zu erlangen, wurden im Rahmen einer genomischen Studie ein EST und ein BAC Enden Sequenzierprojekt initiert.

Für das EST Projekt wurden 2.246 zufällig gewählte Klone sequenziert, aus denen 488 nicht redundante Contigs zusammengesetzt wurden. Von diesen wurden 280 Sequenzen ausgewählt, und zusammen mit 86 unabhängig klonierten mutmasslichen Transkripitonsfaktoren mittels *in situ* Hybridisierung genauer charakterisiert. Durch die Expressionsanalyse konnten mindestens 25 neue Gene isoliert werden, die wahrscheinlich in verschiedenen Aspekten der Embryonalentwicklung von *Tribolium* wie Segmentierung, Entwicklung der Extremitäten, Neurogenese, Myogenese und Musterbildung der terminalen Strukturen ein Rolle spielen. Eine vergleichende Analyse der EST Sequenzen unter evolutionären Gesichtspunkten bestätigte, dass *Tribolium* im Vergleich zu den Dipteren eine langsam evolvierende Spezies ist. Die Daten zeigten, dass Evolutionsraten aus Gen und Spezies spezifischen Raten zusammengesetzt sind, wie von der neutralen Evolutionstheorie vorhergesagt.

Bis heute deckt das BAC-Enden Sequenzierprojekt mit 8.640 Sequenzen 2,9% des *Tribolium* Genoms ab. Durch eine funktionelle Analyse eines Teils dieser BAC End Sequenzen (BES) konnten 486 mutmassliche offene Leseraster identifiziert werden. Es kann geschätzt werden, dass in den 53.000 BES die produziert werden sollen 6.900 offene Leseraster enthalten sind, und damit 18% des Genoms sequenziert werden.

Es wird gezeigt, dass die Sequenzierung zufällig ausgewählter ESTs und der BAC Enden eine leistungsfähige Methode zur Identifizierung neuer Gene ist, bei der Erstellung einer Karte des *Tribolium* Genoms hilft, und der Identifizierung von kodierenden Bereichen in genomischen Sequenzen dient.

Summary

During the last decade, *Tribolium castaneum* has become the insect of choice for comparative genetics and developmental studies outside of drosophilids. Until recently, most molecular studies have focused on the comparative analysis of early development with a focus on segmentation and homeotic genes. In order to acquire independent knowledge on the genetic basis of insect development, a genomic approach consisting of EST and BAC-ends sequencing projects has been initiated in *Tribolium*.

The EST project resulted in the production of 2,246 random sequences representing 488 non-redundant EST contigs. Of those, 280 sequences were selected, along with 86 independently cloned putative transcription factors, and further characterized by *in situ* hybridization. Expression analysis led to the identification of at least 25 novel genes putatively involved in diverse aspects of *Tribolium* embryonic development such as segmentation, appendage development, neurogenesis, myogenesis and terminal patterning. Comparative evolutionary analysis of the EST sequences verified that *Tribolium* is a slow evolving species when compared to dipterans. As predicted by the neutral theory, the data also revealed that evolutionary rates are a composite measure of both gene and species specific rates.

To date, the BAC-ends sequencing project resulted in the production of 8,640 sequences covering 2.9% of the *Tribolium* genome. A functional analysis of a subset of these BAC-end sequences (BES) allowed the identification of 486 putative ORFs. It is estimated that of the 53,000 BES to be produced, 6,900 ORFs will be found, comprising 18% of the genome.

Random sequencing of ESTs and production of BES are shown to be powerful ways to identify new genes, to help mapping the *Tribolium* genome and to identify coding regions in genomic sequences.

1. General introduction

1.1 Taxonomy

Tribolium castaneum HERBST (Coleoptera; Tenebrionidae) also known as the red flour beetle or Mehlkäfer is a common, worldwide distributed pest of cereal products. The genus *Tribolium* comprises 33 species divided in five species groups (Hinton, 1948; reviewed in Sokoloff, 1972). Only eight species are maintained as culture in laboratories. Among those, *Tribolium castaneum* and *Tribolium confusum*, which are the most widespread species, are the ones more frequently used in a scientific environment.

1.2 First steps as model organism

Although the description of *Tribolium castaneum* by Herbst goes back to 1797, its first use as a model organism had to wait until the beginning of the twentieth century. The very first experiments generally concerned the Tenebrionidae family at large rather than *Tribolium castaneum per se*. Chapman (1924a) first stressed the use of *Tribolium* in population studies while Arendsen Hein (1920; 1924a; 1924b) and Ferwerda (1928) were using *Tenebrio molitor* (mealworm) to study the heritability of characters. *T. castaneum*'s real debut on the scientific scene goes back to the 1930s when Park (1934) and Good (1936) described the general biology of *Tribolium*, and when *pearl,* the very first *T. castaneum* mutant, was described (Park, 1937).

Genetics being still at its infancy, the research involving *Tribolium* during the following two decades focused almost entirely on population ecology and pest control. Noteworthy was the work of Thomas Park, which mainly dealt with interspecies competition between *Tribolium castaneum* and *Tribolium confusum* (Park, 1948, 1954, 1957). Interestingly, what was first seen as being competition was later reinterpreted as a predator-prey interaction between the two species (Park et al., 1965; Sokoloff and Lerner, 1967). The question then raised by Alexander Sokoloff "Interactions in *Tribolium*: Competition or predator-prey?" is still open today.

A certain interest in *Tribolium* among geneticists arose only in the early 1960s with the production of the first linkage maps and the description of several mutants. Yet, on the whole, the main focus of research on *Tribolium* historically speaking has been as a model organism for population ecology and genetics, parasitology and insecticide resistance. Figure 1.1 illustrates this fact. The histogram represents the approximate number of publications describing *Tribolium* research produced every decade since 1945, both in historical and genetic fields. At first sight, one can immediately see that historical fields account for the vast majority (90%) of all the *Tribolium* related publications. Still, interest in *Tribolium* has never stopped growing, and in the last fifteen years, the number of publications in every field of biology follows an exponential growth. Here, we shall focus on the rise of *Tribolium castaneum* as model organism in genetics and development. The association of *Tribolium* with genetics has its very own history, which will be described later on.

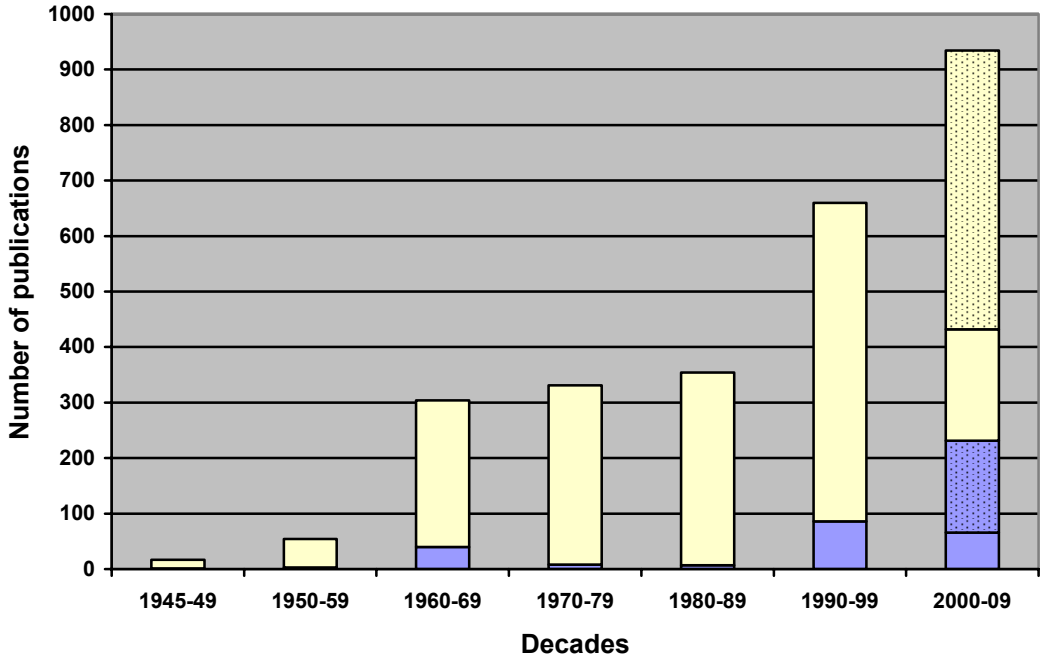

Figure 1.1 Distribution of scientific publications concerning *Tribolium* between genetics and development (blue) and other fields of biological sciences (ivory) since 1945. The shadowed areas in the 2000-09 decade are forecasts for the interval 2004-2009. Statistics were compiled by querying the ISI Web of Science database using "*Tribolium*" as query.

1.3 Why use *Tribolium* anyway?

Several factors favour the use of *Tribolium castaneum* as model organism. Generally speaking, it represents the majority. Coleoptera is the most successful group among insects and possibly among animals, if the number of species is a criterion. The coleopteran order comprises a minimum of 350,000 species. Specifically, *Tribolium* has a short life cycle that lasts about one month from zygote to reproductive adult. It can be reared in dense population, on a simple medium (flour supplemented with brewer's yeast), and in a wide range of temperature and relative humidity conditions. Females produce eggs one to two days after hatching and this for a period of four to five months. They can produce about ten to twenty eggs a day. Various stages of the life cycle can be isolated easily from the flour with sieves of different mesh. The adults can be long-lived. The average for *Tribolium castaneum* is about six months but two and a half year old males have been found (Good, 1936). Stocks require little care. At 25°C, a stock can maintain itself for four to six months before one needs to replace the medium.

Compared to *Drosophila melanogaster*, which is the major insect model organism for genetics and development, *Tribolium* has also several advantages. First, in contrast to *Drosophila*, *Tribolium* has a short germ mode of development, which is considered to be the ancestral condition among insects (Tautz et al., 1994). Second, the presence of a non-invaginated head and limb buds during embryogenesis favour the use of *Tribolium* to study head and appendage development. Third, *Tribolium* can be useful to bridge comparisons between *Drosophila* and human. It is well known that dipterans and especially *Drosophila* are fast evolving organisms (Friedrich and Tautz (1997) and Chapter 4 of this thesis), making the use of drosophilid species sometimes difficult in the context of comparative biology. And finally, every standard genetic and developmental technique developed in *Drosophila melanogaster* can also be applied in *Tribolium*.

1.4 *Tribolium* as model organism for genetics & development

As introduced earlier in this chapter, *Tribolium castaneum* has a long history as a model organism among insects alongside the well-characterized fruit fly *Drosophila melanogaster*. The situation of *Tribolium* being somewhat subordinated to *Drosophila* as model organism has influenced the field of *Tribolium* genetics since its creation.

In 1958, an informal meeting of geneticists and ecologists using *Tribolium* in their research was held at the International Genetics Congress in Montréal where it was decided to create the *Tribolium* Information Bulletin. This publication, patterned on the *Drosophila* Information Service started in 1934, was meant to store information concerning newly described mutants, linkage studies and stock lists. However, at this time not much information was available. In total, only five *T. castaneum* and two *T. confusum* mutants had been described.

During the 1960s, a first wave of genetic data was collected mainly due to the effort of Alexander Sokoloff and Peter Dawson. During an interval of a few years, more than 150 primarily spontaneous mutant phenotypes were described. In parallel, a genetic map covering the ten linkage groups of *Tribolium* was constructed. This first map, containing 35 markers, and several mutants were summarized in the first book devoted to *Tribolium* genetics: "The Genetics of *Tribolium* and Related Species" (Sokoloff, 1966). A few years later, a series of three books termed "The Biology of *Tribolium*" (Sokoloff, 1972, 1974, 1977) was written in the spirit of "The Biology of *Drosophila*" edited by Milislav Demerec in 1950. In the same way Demerec's publication was in his time a bible to any *Drosophila* researcher, the *Tribolium* counterpart was meant to be a reference work that would stimulate research in the field. However, these books represented the end of an era for *Tribolium* genetics. Perhaps because of the great enthusiasm generated by the discovery of the Hox gene cluster (Lewis, 1978) and the segmentation cascade (Nusslein-Volhard and Wieschaus, 1980) in *Drosophila melanogaster*, research in genetics and development using *Tribolium* more-or-less ceased to exist for about twenty years during the 1970s and 1980s (Figure 1.1).

It is only at the very end of the 1980s that researchers regained interest in *Tribolium*. Stimulated by the description of the antero-posterior segmentation cascade in *Drosophila*, researchers embarked on a comparative study of the cascade in both organisms to determine how conserved this patterning mechanism might be. Starting with the Hox cluster (Beeman et al., 1989; Stuart et al., 1991; Beeman et al., 1993), every level of the segmentation cascade was studied and orthologous genes such as the gap gene *hunchback* (Wolff et al., 1995), the pair-rule genes *hairy* (Sommer and Tautz, 1993) and *even-skipped* (Brown et al., 1997) as well as the segment polarity genes *engrailed* (Brown et al., 1994) and *wingless* (Nagy and Carroll, 1994) were cloned and characterized in *Tribolium*. This interest for comparative analysis between the two species extended to other areas of insect development such as head patterning (*orthodenticle* (Li et al., 1996)), dorso/ventral patterning (*twist* and *snail* (Sommer and Tautz, 1994), and *decapentaplegic* (Doctor et al., 1995; Doctor et al., 1996)), appendage development (*Distal-less* (Beermann et al., 2001) and *dachshund* (Prpic et al., 2001)) and most recently neurogenesis (*achaete-scute* (Wheeler et al., 2003)).

In parallel to the classical approach of cloning and characterizing orthologous genes, other significant initiatives were taken to facilitate the use of *Tribolium* as model organism. A new genetic map containing 131 molecular markers covering the ten linkage groups at an interval of 350 kb/cM was produced (Beeman and Brown, 1999). A mutagenic screen focusing specifically on segmentation mutants (Maderspacher et al., 1998) was initiated, bringing the number of mutant stocks available from different laboratories to several hundred. Technical innovations were also realized, making genetic manipulations possible in *Tribolium*. The universal insect transformation system (Berghammer et al., 1999) and the parental RNA interference (pRNAi) technique (Bucher et al., 2002) were first developed in *Tribolium*.

1.5 *Tribolium* in the genomic era

Although the conserved and divergent aspects of *Tribolium* and *Drosophila* development are far from being understood, almost every orthologous gene of developmental interest that could be characterized in *Tribolium* has been cloned. Yet, after more than a decade of genetic and developmental studies in *Tribolium*, only few original discoveries have been made outside of *Drosophila*. Notwithstanding the fact that comparative work done between the two species was highly significant, relatively few novel ideas concerning insect embryonic development were generated.

At the dawn of the 21st century, it became a consensus idea among *Tribolium* developmental biologists that novel genetic data had to be generated to make *Tribolium* a model organism in its own right. To circumvent the paucity of *Tribolium* specific developmental molecular markers, we decided to adopt a genomic approach to the problem and to initiate an expressed sequence tag (EST) project in *Tribolium castaneum*. Our approach can be divided into three successive steps: (1) generation and annotation of sequence data, (2) *in situ* hybridization screen and (3) RNAi functional analysis. For the purpose of my PhD work, only the first two steps are here presented. The third step is to be initiated in the near future. Sequence data that were originally meant to be generated from a single source of genetic material, ended up being obtained using three different sources (cDNA library, direct cloning and BAC library) and are therefore presented in three distinct chapters. A last chapter concerns evolutionary features of *Tribolium* as deduced from the comparative analysis of evolutionary rates between the beetle, dipterans and human.

2. The EST project

2.1 Introduction

Since DNA cycle sequencing has become a routine procedure, the generation of expressed sequence tags or ESTs is the method of choice to generate rapidly and easily a very large number of raw genetic data about the transcriptome of any organism or organ of interest. The first EST projects were initiated at the beginning of the 1990s as part of the Human Genome Project (HGP) (Adams et al., 1991). Since then, several hundreds of similar projects have been initiated, generating more than eighteen millions ESTs that cover almost every phylum of the tree of life.

Outside of drosophilids, large amounts of insect genetic information are currently publicly available only for the mosquito *Anopheles gambiae*, the honeybee *Apis mellifera* and the silkworm *Bombyx mori* whose genomes have been or are in the process of being sequenced. However, mosquitoes, honeybees and silkworms are by no means model organisms for developmental genetics, in contrast to *Tribolium castaneum*. It is therefore clear that such an EST project had to be initiated in the beetle.

By definition, an EST is a single raw sequence read of a clone randomly chosen from a cDNA library. ESTs are generally short reads (400-500 bp) described as being usually of “low quality”. The largely false idea that ESTs are of lower quality than any other sequencing reaction performed on a daily basis comes from the fact that ESTs are generated in a high throughput fashion, whereby tens of thousands of reads are produced in a very short time. Consequently, they are never visually inspected for quality and may therefore contain unnoticed sequencing errors. Nevertheless, these sequences are highly informative since they allow a quick “read” of the information contained in any cDNA library and provide a way to identify putative genes of interest. In this chapter, I present the outcome of the *Tribolium* EST project.

2.2 Materials & Methods

cDNA library

The cDNA library was readily available in the laboratory when I started the project. It was constructed by Dr. Reinhard Schröder in 1995 using the Uni-ZAP XR vector (Stratagene). The library was generated using eggs covering every stage of *Tribolium* embryonic development. cDNAs were cloned directionally into the EcoRI-XhoI site. After mass *in vivo* excision, inserts were contained in pBluescript SK II + phagemid.

DNA extraction

Clones were randomly picked and grown for 24 hours (37°C, 325 rpm) in 1.2 ml of 2xLB-Amp (50 µg/ml) [2xLB: 50 g Peptone 140, 25 g yeast extract, 25 g NaCl in 2.5 l of ddH_2O] contained in a 2.2 ml 96 deep well plate. Cells were pelleted 10 min. at 3,200 g and supernatant was discarded. The pellet was then suspended in 200 µl of Solution 1 [50 mM glucose, 25 mM Tris-Cl pH 8.0, 10 mM EDTA pH 8.0, 20 µg/ml RNase A]. Lysis was achieved by adding 200 µl of Solution 2 [0.2 N NaOH, 1% SDS]. The solution was neutralized by adding 200 µl of Solution 3 [5 M guanidine-HCl, 0.7 M KOAc pH 4.8]. Cell debris was precipitated (15 min. at 3,200 g) and 400 µl of the cleared supernatant was loaded on a 96-well Unifilter 800 GF/B plate (Whatman). These plates contain a silica membrane on which DNA binds reversibly in the presence of high concentrations of a chaotropic agent such as guanidine-HCl. The DNA was bound to the silica membrane by centrifugation (1 min. at 1,900 g) and the flow-through was discarded. Bound DNA was washed twice with 500 µl of 80% ethanol and eluted using 100 µl of ddH_2O. The resulting extract was further isopropanol precipitated or air dried and finally suspended in 25 µl of 10 mM Tris-Cl pH 8.0. Plasmid DNA preparations were stored at -20°C in 96-well plates. This method gave in average 1-2.5 µg of DNA per clone.

Sequencing reaction

Clones were sequenced using the DYEnamic ET Terminator Cycle Sequencing kit and run on a MegaBACE 1000 instrument (Amersham Biosciences). Sequencing reactions were set as follow: 3 µl DYEnamic ET reagent premix, 250 nM of primer and 2 µl of plasmid DNA (≈100-200 ng) brought to a final volume of 10 µl with

ddH_2O. Reactions were cycled (95°C, 20 sec.; 50°C, 20 sec.; 60°C, 1 min. x 40 cycles) and then purified on Sephadex G-50. Primers used are M13 reverse (GGAAACAGCTATGACCATG) and M13 forward –21 (GTAAAACGACGGCCA GT). Reactions were injected in ddH_2O at 2 kV for 45 sec. and run at 9 kV for 150 min. at 44°C.

Sequence trace analysis

Sequence traces were basecalled using Phred (Ewing and Green, 1998; Ewing et al., 1998) and assembled into contigs using Phrap. Contigs were subsequently examined by eye and non-overlapping contigs belonging to a single transcript were manually joined for the subsequent BLAST analysis.

Annotation

Preliminary functional annotation was performed on the basis of a tBLASTx analysis against the non-redundant NCBI database and a BLASTx analysis against the BDGP predicted proteins data set. Batch BLAST at NCBI was done using the BLAST client (blastcl3) program and BLASTs at BDGP were done by hand. Cut-off value was e<1e-03.

2.3 Results

A total of 2,304 randomly chosen clones were sequenced from the 5' end. The rational to generate only 5' ESTs comes from the fact that 3' ESTs are likely to be less informative since they usually contain long untranslated regions (UTRs), which do not help to assess the functional relevance of a transcript. This is less frequently the case for 5' ESTs in the context of a cDNA library where the average insert size should be around 1 kb. In this EST project, I placed a special emphasis on transcription factors since they are frequently involved in key steps of embryological development and have very specific expression patterns, which allow their use as molecular markers. From all the sequencing reactions performed, 76% were successful, resulting in the production of 1,750 5' ESTs.

Sequences were first classified in three general categories to help identify and discard contaminating sequences and to assess at the same time the quality of the cDNA library used in the project. Mitochondrial and ribosomal sequences were the first targets of such a screen. Both sequence types are present in very large copy numbers in any cell at any time and are sometimes over-represented in cDNA libraries (Bonaldo et al., 1996). Sequences of mitochondrial origin are considered real contaminants since mitochondrial transcripts do not contain the poly-A tail used to select transcripts to be included in the library. Of all the 5' EST produced, we obtained 417 mitochondrial sequences (23.8%), 411 ribosomal sequences (23.5%) and 922 other sequences of nuclear origin (52.7%). This last category is the one on which I concentrated my efforts in the subsequent steps. After assembly of the non-ribosomal nuclear sequences into contigs, we obtained 488 non-redundant single sequences. These contigs were preliminarily annotated using BLASTx results from BDGP and FlyBase annotation (when available) as guidelines. When no specific information was available from these databases, annotation was done on the basis of the NCBI BLAST result. Figure 2.1 represents the different sequence classes obtained in this way. Classes were created following the Gene Ontology classification system but in a much less exhaustive way.

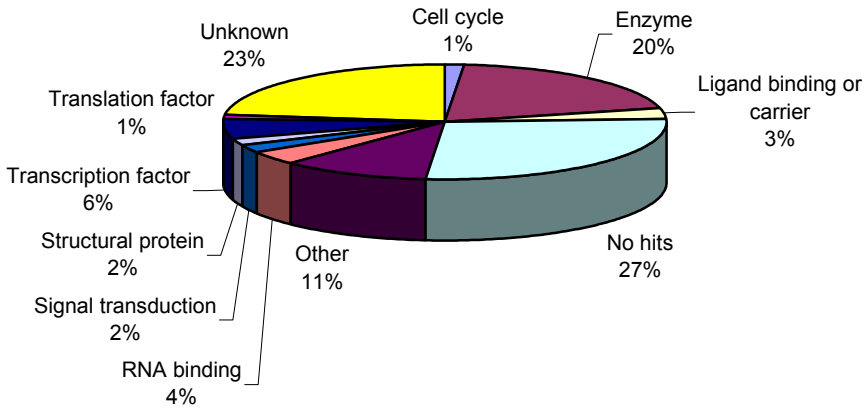

Figure 2.1 Distribution of 488 non-redundant EST contigs among sequence classes.

Although most categories are self-explanatory, others require some explanation. “No hits” comprises contigs for which no BLAST result was obtained, while “Unknown” includes contigs for which a significant BLAST hit was obtained but no functional information could be found. Good examples of genes belonging to the later category are the ones showing high similarity to predicted but unannotated genes from *Drosophila melanogaster* or *Anopheles gambiae*. *Tribolium* is, from a phylogenetic point of view, more closely related to the fruit fly and the mosquito than to any other highly represented organism in the NCBI database, therefore, the “Unknown” category is a major one. The class “Enzyme” contains every contig that putatively encode proteins for which a molecular function involving an instance of enzyme activity was found. “Other” includes contigs that could not easily fit into any category. A detailed table of EST assignation to contigs, BLAST results and preliminary annotation is available in Appendix I.

From this set of non-redundant sequences, 280 genes potentially involved in embryonic patterning were selected for further investigation by *in situ* hybridization (see Chapter 5). Included in this pool are transcription and translation factors, signal transduction proteins as well as all the contigs belonging to the “Unknown” and “No hits” classes.

In the context of a collaboration with Dr. Richard Beeman (USDA, Kansas), the 3’ ends of 496 clones were sequenced for mapping on the *Tribolium castaneum* genetic map. All the sequences produced during the EST project have been deposited in the NCBI dbEST (Genbank accession numbers CB334789-CB337245; dbEST_ID 17071305-17073761) as well as in the *Tribolium* Genome Database (TGD) and are therefore not included here.

2.4 Discussion

2.4.1 High throughput preparation of plasmid DNA and sequencing

In such a high throughput project, a key factor to success is the proper design of a production line. One must understand that protocols used in an everyday laboratory life, which are good for processing few samples in parallel, need to be modified and optimised in order to reach an output of hundreds to thousands of samples a day. Hence, much time was spent designing high throughput procedures, which were efficient but also cheap. This chapter and some of the followings contain a discussion of technical points, which might be considered trivial but are in fact the cornerstone of any high throughput approach.

As a very first step, I designed a DNA extraction protocol based on selective binding of nucleic acids on silica. This method has the advantage of being fast, highly efficient and low cost. The ability of silica to bind DNA in the presence of chaotropic agents (guanidine hydrochloride, guanidine thiocyanate, sodium iodine or the like) has been known at least since the 1960s and is still currently used in several commercially available miniprep kits. The basis of such protocols is very simple: an alkaline lysis in the most traditional way is followed not by phenol-chloroform extraction and ethanol precipitation but by binding of the DNA on a silica-based column and cleaning with an ethanol solution. The principle remains the same for processing one or 96 samples, except that single miniprep procedures rely on centrifugation to exchange liquids on the column, while all the commercial high throughput procedures rely on liquid exchanges by application of a vacuum. This step has the advantage of being totally automatable if one has a pipetting robot but has the disadvantage of being less efficient given that only one plate can be manipulated at a time. Initially, we had no pipetting robot in our facility, therefore I designed the protocol to be done using centrifugation. Such a modification allowed a single person to extract up to 768 samples (eight 96-well plates) within a normal working day. Surprisingly, the procedure turned out to be faster, cheaper and as efficient as most DNA miniprep kits available on the market.

The amount of DNA isolated by this method unfortunately turned out to be quite low. In general, 1-2.5 µg of plasmid DNA could be recovered per sample. The low yield of extraction was not due to the protocol itself but to the difficulty of obtaining saturated bacterial cultures. Since extraction was done in 96-well plates, bacterial growth was therefore also performed in a 96-well format. In these plates, culture volume is reduced (maximum 1.5 ml) and because the wells are square and not round, the oxygenation of the culture media by shaking is suboptimal. Incubating the cultures for 24 hours at 325 rpm instead of the usual 16 hours at 125 rpm had only a marginal effect in increasing the amount of recovered DNA. I tried using richer media such as TB or 2xYT to increase yields but these media decreased the overall purity of the DNA, which in turn affected negatively the quality of sequencing. The paucity of DNA material available impeded subsequent steps, making them more prone to failure. I never managed to overcome the yield problem within the time interval where ESTs were produced. In retrospect, I believe that a simple solution would be to grow bacteria in 48-well instead of 96-well plate. This modification is used to grow low copy plasmids such as BACs. In this way, one would have been able to double the volume of culture, to overcome the problem of culture oxygenation and to certainly double the yield.

Sequencing reactions were performed and run using Amersham Biosciences technology. The MegaBACE 1000 sequencer has the advantage of being extremely rapid. It can process 96 samples in one and a half hour. However, the technology developed by Amersham is relatively recent and not optimal. To obtain good results with this set up (600-700 bp per read) one needs to use normalized DNA samples free of EDTA or salt contamination. This was not the case here, so the average read length ended up being closer to 400-500 bp, with a high level of failure (24%). Having recently experienced ABI sequencing technology, I find this technology much simpler to handle and more robust against variations of template quality and quantity. I would therefore consider that ABI is currently more adapted than the Amersham technology for such a high throughput project.

2.4.2 EST assembly and annotation

The *Tribolium* EST project has remained at a small scale. However, it could have been otherwise. Our initial goal was the production of 10,000 5' ESTs. A full scale EST project depends of the organism and the cDNA library used but is generally in the order of 25,000 5' ESTs. Since this was a single man project and time was limited, we envisage a more restricted project. However, we quickly came across another major problem, which convinced us to stop the EST production more rapidly than initially expected. From the preliminary annotation, it turned out that 47% of all the ESTs produced were "trash" sequences of mitochondrial or ribosomal origin. In addition, assembling of all the sequences in non-redundant contigs resulted in a maximum of 586 clusters from a total of 1,750 sequences. This is equivalent to a redundancy level of 67%. Such a high value is definitely abnormal after production of so few ESTs. At this point of the project 10-20% redundancy would have been more likely. This odd result signifies that the embryonic cDNA library available for the EST project was not suited for high throughput random sequencing. Today, if another EST project would be conducted, I would advise constructing a new library and to test it beforehand. New methods of library construction allow cloning of full-length cDNAs, which was not the case in 1995 when our library was constructed. There are also ways to decrease the amount of mitochondrial sequence contamination. Normalization and subtraction of the library could also be considered. Finally, depending on the goal of the EST project, one might think about being more specific in the selection of the life stages and/or the organs to be included in the library.

At the point where raw data production came to completion, computational processing and analysis became more important. Bioinformatics and general computer knowledge thus became the second key factor to the EST project success. Considering the throughput allowed by the equipment, the point is reached very quickly where it is not possible to manage the data by hand within a respectable amount of time. Simple processes such as quality assessment, vector and low quality regions trimming as well as preliminary annotation cannot be performed on a sequence-by-sequence basis and have to be automated. To do so, the programs Phred (basecaller), Phrap (assembler) and Consed (viewer) were used (Ewing and Green, 1998; Ewing et al., 1998; Gordon et al., 1998). These programs were originally designed to assemble BAC clone

shotgun sequences generated by the HGP and are the state of the art in sequence data processing. The only difficulty with those programs is that, until recently, they could only be run under Unix-like operating systems, which are seldom used in biology laboratories. The pipeline was thus operated on a remote fashion, which had the disadvantage that one could not easily modify the assembly parameters and control the quality of the resulting assemblies. This unfortunate situation resulted in three minor problems. First, the data trimming parameters were far too stringent. This resulted in a loss of intermediate quality sequence information, which could have been valuable in the context of an EST project. Second, this hyper-stringency resulted in several unassembled contigs. As much as possible, overlapping contigs were assembled by hand but this was not always feasible. For instance, several ribosomal proteins and the 16S rRNA are represented by more than one contig (Appendix I). Third, cases of overcollapsing were not resolved. Overcollapsing occurs where sequences are wrongly included into a larger contig on the basis of short repeated sequences. Putative mitochondrial and ribosomal contigs have been found to attract unrelated sequences in seven cases, thus partially explaining the multiplication of ribosomal and rRNA contigs described above.

In the end, these difficulties of processing had no significant influence on the subsequent steps of the project. Most contigs selected for further *in situ* analysis were represented by single clones and the few that were not were checked by eye (I found only one misassembled contig in this way). The data stored in NCBI dbEST and TGD is also not negatively influenced since dbEST data is not assembled (only the vector has been trimmed) and data included in the TGD will be reprocessed in the context of the *Tribolium* genome sequencing project using proper parameters.

Functional annotation was probably the only part of the data processing that could not be easily extensively automated. Although batch BLAST can be performed at NCBI and outputs can be easily parsed, the results obtained are usually not highly informative. NCBI contains functional annotations for only few genes and frequently, because of the different ongoing EST and genome projects, these annotations are done automatically on the basis of BLAST results. The circularity of the procedure increasing chances of wrong annotations, I decided to perform functional annotation of most contigs by hand using the large amount of information available in FlyBase.

Surprisingly, 27% of the *Tribolium* ESTs did not have homologous sequences in the queried databases ("No hits" from Figure 2.1). There might be two major reasons for such a situation: (1) they are orphan genes, which are by definition not expected to be closely related to other sequences or (2) they encode UTRs, which should be poorly conserved between distantly related species. Current data does not allow discriminating between the two possibilities.

In summary, a grand total of 2,246 ESTs (1750 5' ESTs and 496 3' ESTs) have been produced and made publicly available. From these, 280 non-redundant clones were selected for further analysis by *in situ* hybridization (see Chapter 5). The importance of high quality starting materials (cDNA library and extracted DNA) and appropriate computational organization were shown to be key factors in the successful completion of such a large scale approach.

3. Cloning of Exelixis ESTs

3.1 Introduction

This chapter describes the results of a collaboration between our laboratory and Exelixis Inc. (South San Francisco, USA). Exelixis has produced in a private fashion over 8,800 ESTs derived from adult and mixed larvae cDNA libraries. In the context of this collaboration, I had the occasion to search their database for transcription factor sequences and to take them back to Köln for cloning and *in situ* analysis. Here, I present the results of the database mining and the cloning procedure.

3.2 Materials & Methods

Database mining

In June 2001, I visited Exelixis laboratories where I queried the database with two lists of subjects. The first one contained 77 Pfam IDs corresponding to all the known transcription factor and DNA binding domains, as well as few protein-protein interaction domains associated to transcription factors. The second list contained 412 gene keywords.

Annotation

Preliminary annotation was performed as described in Chapter 2. Cut-off value was e<1e-10.

mRNA isolation

Total RNA extraction from eggs, larvae and adults (1 g each) was done using a scaled up Trizol total RNA extraction procedure (GibcoBRL). Total mRNA isolation was performed using the PolyATtract mRNA Isolation System (Promega).

cDNA cloning

First-strand cDNA synthesis was done according to the SuperScript II RT protocol (Invitrogen) using 300 ng of mRNA from embryos, larvae and adults. Amplification and cloning of the target cDNA was performed using the 3'-RACE kit (Invitrogen) with the following modifications. AccuPrime SuperMix I (Invitrogen) was used instead of normal *taq* polymerase in a final reaction volume of 25 µl instead of 50 µl as suggested (half reaction). Touchdown PCR was done using the following cycling parameters: 94°C, 2 min.; [94°C, 30 sec.; 65°C to 55°C, 30 sec. with decreasing steps of 1°C; 68°C, 2 min.] x 1; [94°C, 30 sec.; 55°C, 30 sec.; 68°C, 2-4 min.] x 30. After agarose gel analysis, amplified fragments of interest were extracted and eluted in 30 µl of 10 mM Tris pH 8.0. Purified products were cloned in the TOPO-TA cloning vector (Invitrogen) as described but quarter reactions were performed. In each cloning experiment, one GSP and an anchor (GGCCACGCGTCGACTAGTAC) or two GSPs were used. A list of primers is presented in Appendix II.

DNA extraction

DNA extraction was done as presented in Chapter 2. For each gene, 4-8 clones were selected for analysis.

Sequencing reaction & sequence trace analysis

Both of these steps were performed as described in Chapter 2.

Positive clones identification

Positive clones were identified by searching the resulting contigs back against a database composed of sequences recovered from Exelixis database.

3.3 Results & Discussion

From the database mining, a total of 125 putative transcription factor sequences were obtained using the Pfam list of domain IDs and none with the list of gene keywords. This second result probably represents a flaw in the procedure since I found using the Pfam domain IDs several genes, which were part of the keywords list as well. However, time was limited and the database querying was not under my control so no further queries were possible.

Retrieved EST sequences were found to be dispersed across 26 different families of putative transcription factors or DNA binding proteins. Family assignments were done only on the basis of the Pfam domains search. The distribution of the sequences across the main transcription factor families is depicted in Figure 3.1.

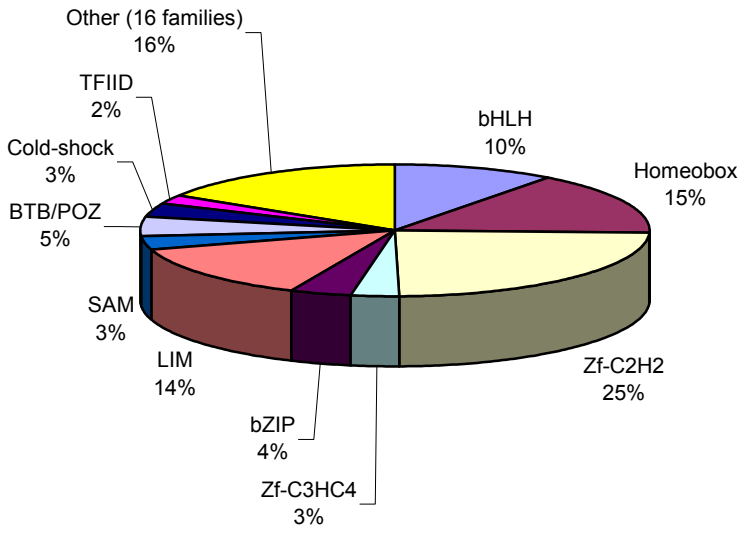

Figure 3.1 Distribution of 125 retrieved Exelixis ESTs across the Pfam transcription factor families.

Preliminary functional annotation was subsequently performed to better assess the relevance of each sequence as far as their potential role in embryonic development is concerned. BLAST results and sequences annotation can be consulted in Appendix II. From this analysis, it seems clear that a few sequences might have been misplaced in some families since eighteen sequences could not return any significant BLAST hits (e<1e-10), which is not to be expected from a gene belonging to a family characterized by a conserved domain. For convenience, the original assignment was not modified but simply labelled with a question mark. If these genes are not truly transcription factors, then they can be regarded as unknown sequences possibly distantly related to transcription factors, which makes them candidates for cloning and analysis via *in situ* hybridization.

From the pool of 125 sequences available, 99 were finally chosen for cloning. Genes already published in *Tribolium* by other workers (mainly homeobox genes) and general transcription factors were not further considered. Cloning was performed preferentially with only one gene specific primer (GSP) to decrease the cost of the operation. In addition, this allowed cloning of sequences including the 3'-UTR, which was not present in the sequences obtained from Exelixis. A mixture of embryonic, larval and adult mRNA was used as template. Although I am primarily interested in factors involved in embryonic development, Exelixis genes were retrieved from mixed larvae or adult cDNA libraries. Using only embryonic mRNA as template would have not allowed discriminating between a failed reaction and an unexpressed gene. Cloning was finally successful for 87 genes using only one GSP 59% of the time.

For each gene, up to eight clones were sequenced. Surprisingly, twelve genes gave multiple transcripts showing insertions and/or deletions ranging from 9 to 142 bp in length. I have not been able to determine if this was due to alternative splicing or to experimental artefacts and if the insertions/deletions were in-frame or not, although half of them seem to be. In addition, I found seven cases of discrepancies between Exelixis and my sequences. Here again, I was not in a position to easily discriminate between wrong assembly of Exelixis sequences, experimental artefacts or biological reality. In these cases, clones selected for *in situ* probe synthesis were the ones verifying Exelixis sequences or the most frequently obtained.

The sequences discussed here are currently not publicly available since they are protected by a confidential disclosure agreement between Exelixis and myself. However, Exelixis has agreed to release the *Tribolium* ESTs and to deposit them in the TGD where they will be publicly available.

4. Evolutionary features of *Tribolium*

4.1 Introduction

The molecular clock hypothesis states that evolutionary rates of proteins and DNA are approximately constant over time and across lineages (Zuckerkandl and Pauling, 1962, 1965; reviewed in Bromham and Penny, 2003). The existence of a molecular clock is predicted by the neutral theory, which presumes that most amino acid changes are selectively neutral (Kimura and Ohta, 1971).

However, the neutral theory also predicts rate variation between phylogenetic lineages, which may be caused by differences in mutation rate and effective population size (Ohta and Kimura, 1971; Ohta, 1987). The discovery that in mammals (Catzeflis et al., 1987; Gu and Li, 1992), rates of evolution differ among lineages has opened the door to larger scale experiments to verify to which extent rate between lineages was heterogeneous. Britten (1986) estimated the rate of evolution for different groups using DNA-DNA hybridization and gene comparisons. It was found that rates vary between groups by a factor of 5. Birds and higher primates have slow rates, while rodents, sea urchins and *Drosophila* have high rates. A study of a similar scale but restricted to vertebrates was achieved by Adachi et al. (1993) who compared the substitution rates of mitochondrial encoded proteins from eight vertebrates. They found that the rate has increased from fish to amphibians to birds to mammals by a total factor of 6.

Following the initial work by Britten (1986) indicating that *Drosophila* is a fast evolving species, several groups studied the peculiar evolutionary behaviour of this model organism. The rate of synonymous substitutions in *Drosophila* was found to be approximately 2 times higher than in rodents and 10 times higher than in primates (Moriyama, 1987). However, the rate calculated seemed to depend on the genes studied, since Sharp and Li (1989) calculated only a 3-fold difference between *Drosophila* and mammalian evolutionary rates using a different set of sequences.

Nevertheless, high rates of nuclear evolution in drosophilids seem to be a general trend. Caccone and Powell (1990), who reviewed six DNA-DNA hybridization studies of drosophilids, also determined that the nuclear genome evolves about 1 order of magnitude faster in drosophilids than in mammals. The high evolutionary rates measured in drosophilids have been suggested to be a characteristic of the whole dipteran order (Carmean et al., 1992; Carmean and Crespi, 1995; Friedrich and Tautz, 1997). When 18S and 28S rDNA are compared among insect orders, an increase of evolutionary rate is noticeable in the dipteran clade (Carmean et al., 1992; Friedrich and Tautz, 1997). This change of rate seems to have been episodic in the stem lineage of dipterans with a maximum increase of rate of about 20-fold compared to other insects.

One of the major criticisms that can be addressed to the vast majority of the studies where evolutionary rates are compared between organisms is the paucity of genetic data on which conclusions are based. To the knowledge of the author, only five studies compared more than ten nuclear genes between species, four of them being related to the controversy surrounding primate and rodent evolution. This situation has the pervasive effect that one considers rates measured mostly using a single gene as being a proper estimate of an organism trend. However, it is clear that heterogeneity in evolutionary rates is present at every level of complexity from genes to genomes to organisms. Evolution of the three genomes (nuclear, mitochondrial and chloroplastic) is uncoupled. Even within the genomes themselves, evolutionary rates are far from being homogenous. Mutation rates vary from region to region (Werman et al., 1990; Pesole et al., 1999), gene to gene (Ayala et al., 1996; Takano, 1998), and domain to domain (Miyata et al., 1980).

In order to better assess the evolutionary trend of both *Tribolium* and *Drosophila* within the insect clade and in comparison to human, the author in collaboration with Dr. Martin Lercher, took a genomic approach to the question of rate heterogeneity among lineages. Here, using 185 orthologous nuclear genes, we compare evolutionary rates between *Tribolium castaneum*, *Homo sapiens*, and two dipterans *Drosophila melanogaster* and *Anopheles gambiae*.

4.2 Materials & Methods

Sequences

EST contigs described in Chapter 2 were used as starting genetic material. From a total of 586 contigs, only 569 sequences of nuclear origin have been retained. ESTs obtained from Exelixis (Chapter 3) have not been included in this work because they constitute a biased data set composed solely of transcription factors.

Identification of orthologues

Tribolium contigs were first searched against all the *Drosophila* proteins from NCBI using BLASTx. The reading frame from the best hit was assumed to be the correct reading frame and was used to translate *Tribolium* contigs into peptides. These were then searched against all the *Homo sapiens* and *Anopheles gambiae* peptides available in NCBI using BLASTp. Sequence representativeness of queried organisms in NCBI protein database was of 15/06/2002. Reciprocal best BLAST hit between *Drosophila*, *Anopheles* and *Homo* was then used to discriminate between orthologous and paralogous sequences. If in all cases we obtained the same sequences as the best hit, we assumed that the four genes in the cluster were indeed orthologues. Cut-off value for all the BLAST searches was e<1e-10.

Alignments

Four-gene alignments were performed with ClustalW (Thompson et al., 1994). Resulting alignments were purged from gap containing positions and concatenated into a single alignment.

Estimation of evolutionary distances

Branch lengths were calculated with the maximum likelihood model by Goldman and Yang (1994) as implemented in the PAML package (Yang, 1997). We used the empirical transition matrix compiled by Jones et al. (1992) and the distribution of evolutionary rates was approximated by a discrete γ-distribution, with the shape parameter as an additional free parameter. Branch lengths were estimated without a molecular clock (i.e., different rates for every branch). When calculating rates for individual genes, we assumed a uniform rate across sites.

4.3 Results

4.3.1 Analysis of concatenated genes

From the 569 nuclear EST contigs of *Tribolium*, we were able to form 185 clusters comprising orthologous sequences from *Tribolium*, *Homo sapiens*, *Drosophila melanogaster* and *Anopheles gambiae*. Concatenation of these orthologous clusters resulted in a single alignment of 24,708 amino acids in length. Given this alignment and the topology presented in Figure 4.1, the best tree was calculated using maximum likelihood. The tree obtained without a molecular clock is (((*Dm*:0.208, *Ag*:0.190):0.099, *Tc*:0.217):0.075, *Hs*:0.438).

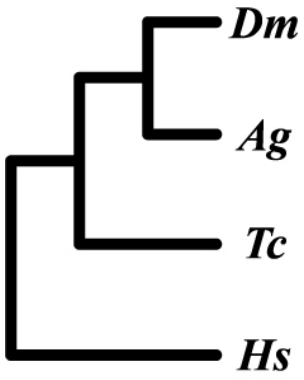

Figure 4.1 Topology used for maximum likelihood branch length estimates. Abbreviations are *Dm*: *Drosophila melanogaster*, *Ag*: *Anopheles gambiae*, *Tc*: *Tribolium castaneum*, *Hs*: *Homo sapiens*.

The divergence of dipterans (*Drosophila* and *Anopheles*) from coleopterans (*Tribolium*) was estimated at 280 Mya (Kukalova-Peck, 1991), which corresponds approximately to the primary radiation of holometabolous insect orders. The divergence between different dipterans was estimated at 210 Mya (Hennig, 1981). From the branch length measurements, we then calculated the mean rates of evolution for each branch (Table 4.1). On first sight, one can immediately see that dipterans are evolving much faster than *Tribolium*. In the time interval of 70 My between the separation of Coleoptera from the dipteran lineage and the radiation of Diptera, an episodic increase of 1.8 times the evolutionary rate of *Tribolium* was measured. Since the diversification of dipterans, the rates are still relatively faster (1.2 to 1.3 times) than in the coleopteran lineage but to a lesser extent.

Table 4.1 Evolutionary rates of each branch since the radiation of holometabolous insects

	Distance [Substitutions/site]	**Time [My]**	**Rate [10^{-3} subs/site/My]**	**Relative rate**
Dm	0.208	210	0.99	1.3
Ag	0.190	210	0.90	1.2
Tc	0.217	280	0.77	1.0
Base of Diptera	0.099	70	1.41	1.8

4.3.2 Analysis of individual genes

Figure 4.2 compares the amino acid distance of individual *Tribolium* genes, and the mean distance of *Drosophila* and *Anopheles* genes to their last common ancestor. There is a strong linear correlation between the two distances (Pearson's r^2=0.64). Assuming that a species-specific rate of evolution is equivalent to forcing the regression through the origin, the corresponding constant of proportionality is 1.34 (excluding the seven genes where either distance was larger than 1). Thus, although evolutionary rates among genes of a genome can vary extensively, these rates are correlated between species in a fashion corresponding to their relative evolutionary rates.

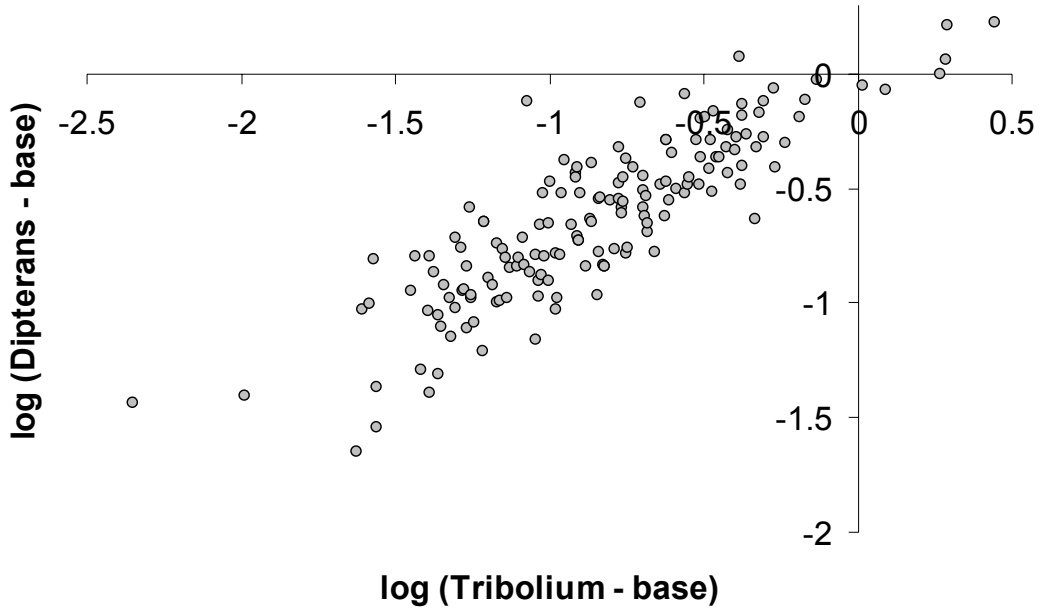

Figure 4.2 Amino acid distances of individual *Tribolium* and dipterans (*Drosophila* and *Anopheles*) genes to their last common ancestor.

4.4 Discussion

Evolutionary rates among dipteran and coleopteran lineages are shown here to be more heterogeneous than expected. Hence, using 185 nuclear transcripts we found that the Diptera lineage (here represented by *Drosophila* and *Anopheles*) has experienced an episodic increase in evolutionary rate when compared to Coleoptera (here represented by *Tribolium*). This rate subsequently dropped in the diversifying dipteran lineage. This result verifies the previous finding that *Drosophila* and probably the dipteran clade as a whole evolve faster than other insect orders by a factor of 3 to 10 (Carmean et al., 1992; Carmean and Crespi, 1995; Friedrich and Tautz, 1997).

However, the rate differences found in our genome scale analysis appear weaker than those reported earlier, with an increase of only 1.3 to 1.8-fold at varying times of the dipteran evolution. This discrepancy may be due to the fact that we calculated rates using concatenated sequences, creating a “super-gene” instead of using a single gene. Super-genes should average genome scale evolutionary rates more accurately

than a single gene and should therefore be less subject to gene-to-gene variation of rates within and between lineages. This is confirmed by the gene-specific rate comparisons in Figure 4.2, which vary widely among proteins.

This peculiar evolutionary behaviour of *Drosophila melanogaster* has strong implications considering its position of most studied model organism. When *Tribolium*, *Drosophila* and *Homo* are compared in a BLAST analysis, *Tribolium* sequences are frequently more similar to human than to their *Drosophila* counterpart. In the EST study (Chapter 2), this situation was found 33% of the time in a tBLASTx analysis (Appendix I). BLAST search of BAC-end sequences resulted in a similar finding. The best hit is a species outside of Hexapoda 24% of the time and a chordate in 18% of all cases (Chapter 6). A comparable situation was found in a BLAST analysis of honeybee ESTs (Whitfield et al., 2002). These results suggest that data from other insects might be necessary to link human genes to *Drosophila* counterparts. In this respect, *Tribolium* is likely to play a significant role as insect model organism. Indeed, it has already been the case at least once since evolutionary relationships between the *Drosophila zen* gene and human HOX3 genes have been resolved by comparison with orthologous counterparts from *Tribolium* and *Schistocerca gregaria* (Falciani et al., 1996).

The basic rate of molecular evolution on which the molecular clock is based refers to the intrinsic mutational rate supplied by the DNA replication machinery. This fundamental mutation rate, in concert with selective pressures, gives rise to variations in amino acid substitution rates between lineages but also between genes. Here, we show that although evolutionary rates among genes of a genome can vary extensively, these rates can be correlated between organisms. Thus, evolutionary rate of a gene in a given species is a good indicator of the evolutionary rate of an orthologue in another species, providing that the relative evolutionary rate between both organisms is known. This finding contradicts the work of Rodriguez-Trelles et al. (2001) who studied molecular clocks of three proteins across a range of species (30-61 per gene) and concluded that evolutionary rates are distributed erratically among genes and lineages. This discrepancy is likely to be a sample size effect. Due to the variation in relative rates, three proteins may not be sufficient to detect the linear trend shown in Figure 4.2.

In summary, we have shown that the dipterans and especially *Drosophila melanogaster* are fast evolving when compared to *Tribolium castaneum*, and that the rate of amino acid evolution of a gene in one species is a good predictor of the rate in another species. As predicted by the neutral theory, evolutionary rates are a composite measure of both gene specific and species specific rates, which together explain two thirds of the variation in evolutionary distances. These results will have important implications for the interpretation of comparative genetics and developmental experiments between *Drosophila*, *Homo* and *Tribolium*.

5. *In situ* screening of ESTs

5.1 Introduction

Within the context of a large scale EST project, functional studies are essential steps in assessing relevance of genes for developmental genetics. Preliminary functional annotation using BLAST analysis, the very first step in this direction, is useful in the way that it provides basic information about putative orthology and function of genes. However, this approach is quite limited since databases such as NCBI or FlyBase contain relevant functional annotation for a relatively limited number of genes. Hence, in the case of the *Tribolium* EST project, no functional annotation could be found 23% of the time for genes showing a significant level of similarity to our ESTs, even by querying several different databases. Moreover, a fairly large number of ESTs (27%) simply did not return significant hits to any sequences currently available in databases. Thus, a BLAST analysis can be seen in the present context as a rough but easy and rapid procedure, allowing discrimination between genes of potential interest (e.g. showing similarity to transcription factors or signalling molecules) and genes that should not be directly relevant to developmental genetics (e.g. house-keeping genes).

After attention has been narrowed from thousands to a few hundreds of sequences, the second logical step to the identification of genes involved in embryonic patterning is to analyse the expression pattern of the selected genes using *in situ* hybridization. As a common rule, it is generally assumed that a gene showing an expression pattern restricted in space and/or in time during embryonic development is likely to play a significant role in this process, thus further narrowing down the number of genes to be more deeply investigated via RNAi for example. Here, I present the results of an *in situ* hybridization screen performed on 366 genes selected from the sequence pools presented in Chapters 2 and 3.

5.2 Materials & Methods

Egg collection and fixation

Cultures of *Tribolium* adults were reared at 25°C. Eggs covering every embryological stage were collected on a weekly basis by sieving. Fixation was carried-on as described in Wigand et al. (1998).

RNA probe synthesis

DIG-labelled RNA probes were synthesized and purified according to the procedure described in the DIG RNA Labelling Mix protocol (Roche) except that half reactions were performed. Clones chosen for RNA probes synthesis are listed in Appendix III.

***In situ* hybridization**

In situ hybridization was performed as described in Lehmann and Tautz (1994).

5.3 Results & Discussion

From a possibility of 613 non-redundant EST contigs, a total 366 sequences (280 in-house ESTs and 86 Exelixis ESTs) were selected for further analysis by *in situ* hybridization. Included in this pool were all the genes possibly involved in embryonic development, including all the putative transcription factors and DNA binding proteins (26%), signalling molecules (3%) as well as all the genes of unknown function (71%). As of October 2003, the screen was completed at 78% (286 out of 366 genes screened).

From all the genes tested, an expression pattern was obtained in only 68% of all cases (194 out of 286 genes). This level is somewhat low, because most genes tested were isolated from an embryonic cDNA library and should therefore be expressed during embryogenesis. General failure of the *in situ* screen can be rejected because the gene *engrailed* used as a positive control never failed to give a signal. Failure was observed on a gene-by-gene basis. Given that one never knew what kind of signal to expect, it was sometimes difficult to differentiate between very low ubiquitous

expression and non-specific staining or to determine how long a staining reaction would need to be pursued in order to give a signal. In our case, if no specific signal could be obtained after sixteen hours of incubation, the gene was classified as showing no signal and was not further investigated.

A general distribution of the expression patterns obtained is presented in Figure 5.1 and a more detailed description of all the results is available in Appendix III. As expected, the vast majority of the patterns represent ubiquitous gene expression. Although most ubiquitously expressed genes are present at every stage of the embryonic development, few of them (12 out of 154) have an expression pattern restricted in time. For convenience, I have separated the different developmental stages of *Tribolium* in three categories: early, intermediate and late. The early stages correspond to the blastoderm and primitive pit stages before the appearance of the germ anlage (approximately 0-12 hours of development at 30°C). The intermediate stages correspond to the growing germ anlage when all the segments are added up to the point where the germ band is fully elongated (12-20 hours). And finally, the late stages include all the subsequent developmental stages from the point where the embryo begins to retract and appendages begin to appear up to the dorsal closure (20 hours to hatching). Especially frequent were the ubiquitous genes restricted to the early developmental stages (9 out of 12). These genes could be interesting candidates for early patterning of the embryo, especially for the establishment of the antero/posterior and dorso/ventral axis.

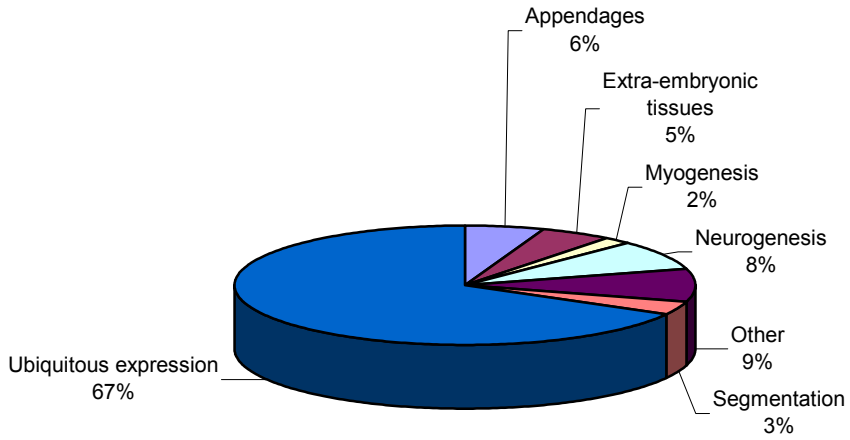

Figure 5.1 Distribution of expression patterns obtained by *in situ* hybridization.

Other expression patterns were assigned to general categories of expression according to what could be observed. A very large number of patterns were unspecific and/or excessively dynamic. Some patterns were easily interpreted and assigned to one or several categories of expression, while others were more problematic. Many genes exhibited mixed expression patterns where, for example, a neurogenic-like expression was overlayed on a low level ubiquitous expression, or a gene with ubiquitous expression will seem to form segmental stripes or to be more strongly expressed in appendages. These cases were documented as precisely as possible and when the observations seemed uncertain, they were assigned a question mark (Appendix III). Other expression patterns were difficult to interpret considering that, contrary to *Drosophila*, no fate map and few morphological markers are available in *Tribolium*. Patterns described as being possibly specific to the terminal patterning system, tracheal system, ventral midline, presumptive mesoderm or non-neurogenic ectoderm have all been classified in the "Other" category to illustrate the uncertainty surrounding the interpretation of the expression patterns. Notwithstanding these difficulties, some highly interesting expression patterns have been obtained for genes covering every category of expression. Pictures representing the main aspect of these patterns are presented in Figure 5.2. Clearly, more work needs to be done to properly

characterize these genes. Double *in situ* using specific morphological markers and/or RNAi will be necessary to assess the biological significance of the observed patterns.

As we can see here, an *in situ* screen is a very powerful way to identify new molecular markers for developmental genetics. Nevertheless, such an approach was difficult to implement in *Tribolium*. The main difficulty was to get a sufficient amount of embryos to perform hundreds of *in situ* hybridizations. Obtaining a large number of eggs from *Tribolium* cultures is not as problematic as preparing them for *in situ* analysis. The relatively low efficiency of the egg devitellinization procedure in *Tribolium* compared to *Drosophila* is limiting in two respects. First, it is very hard to obtain enough fixed embryos to perform such a large number of *in situ* hybridization. In the present case, fixations had to be performed once a week for about half a year in order to obtain sufficient material. Second, it is virtually impossible to obtain a pool of devitellinized embryos containing an even distribution of every developmental stage. In *Tribolium*, the fixation procedure provides a large number of early and late stages but very few intermediate stages. Incidentally, the intermediate stages are usually regarded as being among the most relevant for developmental genetics since most segments are added during this period. To circumvent the problem, egg fixations restricted to a 12-20 hours interval (30°C) were performed to enrich the egg pool in intermediate stages, thus increasing the chances of obtaining few representative embryos of each developmental stage in every *in situ* analysis.

In conclusion, about 25 molecular markers covering almost every aspect of *Tribolium* embryonic development have been identified in this *in situ* hybridization screen. Most of these genes, especially the ones putatively involved in antero-posterior segmentation and appendage development, will be further investigated at least by pRNAi.

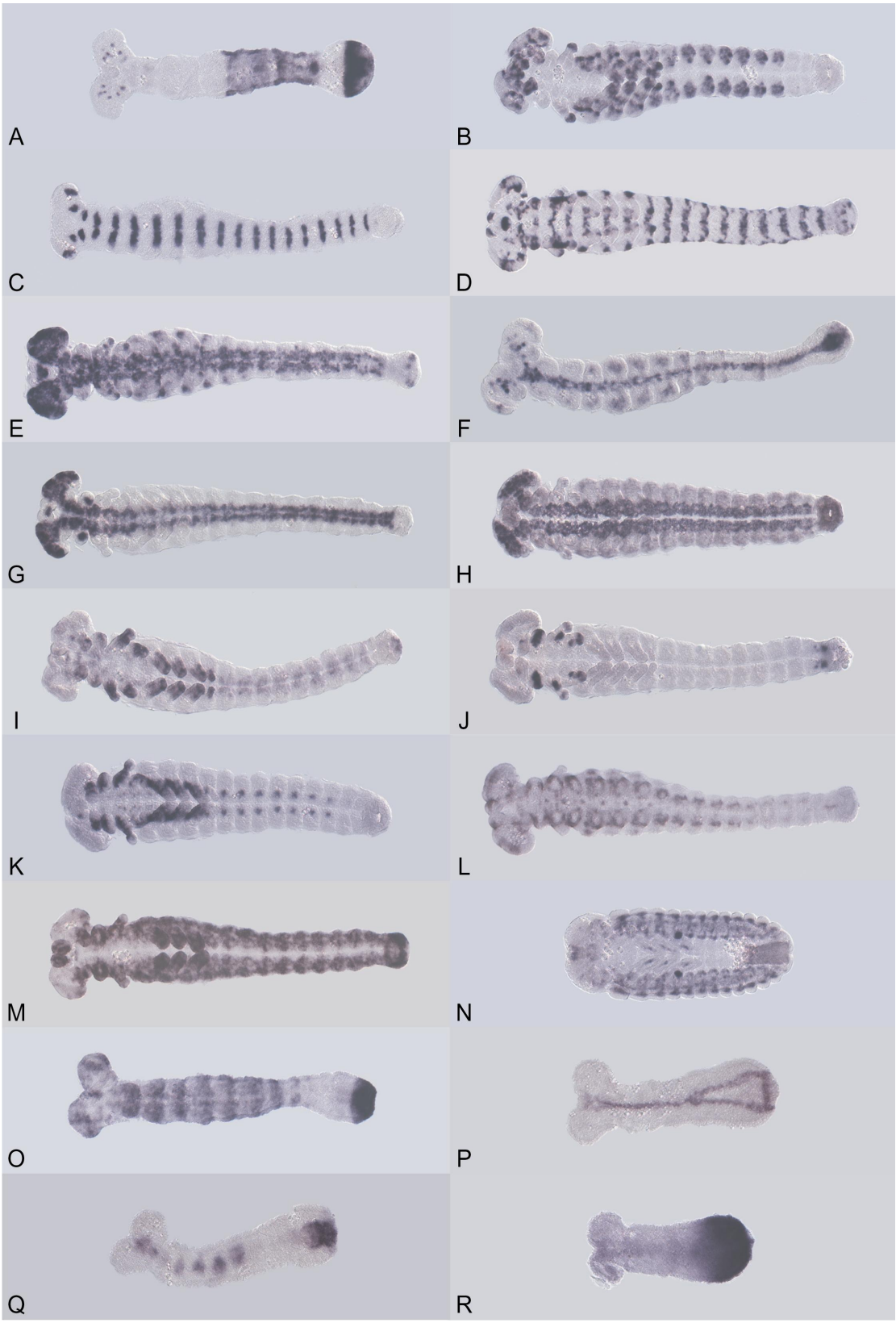

Figure 5.2 Some of the most interesting expression patterns obtained in the *in situ* hybridization screen. Magnification for all the pictures is 10X. **A-D** Putative segmentation genes. **A,B** Tc006A12 shows an expression in two phases. The first phase (**A**) is mainly characterized by a gap-like expression while the second phase (**B**) is mostly restricted to appendages. No homology to other genes was found.

C Tcex003 is a putative segment polarity gene. No homology to other genes was found. **D** Tc021D12 is a segment polarity gene homologous to *Notum* from *Drosophila*, which is part of the Wnt receptor signalling pathway. **E-H** Putative neurogenesis genes. **E** Tc007H05 is an HLH transcription factor showing homology to *E(spl) region transcript mβ* from *Drosophila*. **F** Tc025H12 shows low similarity to *modulator of the activity of Ets* from *Drosophila*, which is involved in the regulation of the EGF receptor signalling pathway. **G** Tc005B12 is an HMG-box DNA binding protein homologous to *Sox21b* from *Drosophila*. **H** Tc027H03 expression pattern is highly similar to *Tribolium achaete-scute homolog* (*Tc-ASH*), which is a proneural gene. No homology to other genes was found. **I-L** Putative appendage genes. **I** Tc001E12 shows a low level of similarity to a *Drosophila* PKC-potentiated PP1 inhibitory protein of unknown biological function. It is expressed in every appendage. **J** Tc014B08 is expressed in every head appendage and in the last abdominal segment at the presumptive position of the urogomphi. It shows low similarity to an ORF from *Anopheles*. **K** Tc025C01 is expressed in the anterior compartment of every appendage as well as in a circular area of each abdominal segment. No similarity to other genes was found. **L** Tcex019 is a homeobox gene forming rings in developing appendages. It also forms a neurogenic-like expression pattern in older embryonic stages. It is homologous to *ventral veins lacking* from *Drosophila*, which is required for the differentiation of selected neurons and glia in the central nervous system. **M-R** Other processes. **M** Tc027C03 is presumptively expressed in the non-neurogenic ectoderm. No homology to other genes was found. **N** Tc014A02 is a Zf-C2H2 transcription factor probably involved in myogenesis. No homology to other genes was found. **O** Tcex045 is Zf-C2H2 transcription factor possibly involved in the terminal patterning system. It shows high homology to *Drosophila no ocelli*, which is involved in tracheal system development. **P** Tc010C05 expression pattern possibly borders the mesoderm. No homology to other genes was found. **Q** Tc002G09 has an early segmental expression and is possibly marking the visceral mesoderm. This gene shows high similarity to a RA domain containing protein from *Drosophila*. Proteins with this domain are mostly RasGTP effectors. **R** Tcex051 is a Zf-C2H2 transcription factor expressed very strongly in the growth zone during early embryonic stages. It is expressed in the fashion of a proneural gene in late embryonic stages. It shows low similarity to genes from *Drosophila*.

6. BAC-ends sequencing project

6.1 Introduction

Since about 1990, which marks the beginning of the human sequencing project, dozens of eukaryotic genomes have been sequenced and hundreds are planned or already in progress. Most large genomes, with the exception of human and *Drosophila melanogaster*, are still at the level of draft sequences. Sequence coverage is suboptimal, gaps remain to be closed and discrepancies in the assembly of large contigs along the genome map (i.e. integrated genetic, physical and cytological maps) must be resolved. Notwithstanding the relative incompleteness of these genome projects, a tremendously large amount of information regarding gene content and genome organization can readily be retrieved from such drafts.

There exist two main strategies to sequence a genome: the hierarchical or clone-by-clone and the whole-genome shotgun (WGS) strategies (reviewed in Green, 2001). The first approach relies on the previous mapping of BAC clones to determine a minimal tiling path of clones covering the whole genome, which will then be sequenced by shotgun. The second approach bypasses the mapping step to directly shotgun the whole genome at once. Both methods have strengths and weaknesses. The clone-by-clone approach has the merit of making the assembly of contigs into an ordered genome sequence easy since every BAC clone was previously mapped. However, it also signifies that marker dense genetic and physical maps must be available at the beginning of the sequencing project. The WGS approach has the advantage of bypassing the time consuming construction of genome maps. Hence, generation of a prefinished genome sequence by WGS is much faster than by hierarchical sequencing. However, assembling tens of millions of single reads into a genome is far from being trivial and WGS usually fails to provide a properly assembled genome if no genome map is available. Even if a physical map exists, WGS will usually result in a low resolution assembly of several thousands of scaffolds (very large contigs) split by gaps of every magnitude (e.g. WGS of the

human genome (Venter et al., 2001)). In summary, WGS is a much more rapid method to generate a prefinished genome sequence, but the hierarchical approach provides a draft sequence that is easier to finish. Here, I should emphasize that no matter which approach is chosen to sequence a genome, marker dense physical and genetic maps must be available at one point during the assembly process. Currently, the trend to genome sequencing is somewhat a hybrid strategy between clone-by-clone and WGS strategies. Although there exist several variations of this hybrid strategy, one can summarize it in this way: the bulk of sequence data is first obtained by whole-genome shotgun but the final assembly and gap closing steps rely on a physical map and on sequencing of specific BAC clones.

In October 2003, *Tribolium castaneum* was granted the rank of high priority organism for genome sequencing by the National Human Genome Research Institute (NHGRI). The NHGRI was the organization responsible to coordinate the efforts of the HGP and is currently the sponsor of dozens of eukaryotic genome sequencing projects including chimpanzee, chicken, dog, cow, sea urchin, honeybee as well as eleven *Drosophila* species. Before this announcement, several initiatives had already been taken to pave the way in obtaining the full genome sequence of *Tribolium*. A high quality BAC library and an intermediate resolution genetic map (Beeman and Brown, 1999) covering every chromosome are readily available and a physical map based on BAC clone fingerprints is underway. It is in this context that we decided to establish a BAC-ends sequencing project in our laboratory. End sequences of BAC clones are highly specific markers, which can make a useful contribution to the proper completion of large scale genome projects. First, BAC-end sequences (BES) can be mapped on both genetic and physical maps. This has the advantage of increasing the density of markers on both maps in a non-random fashion. BAC libraries are meant to cover whole genomes with a high level of redundancy so BES increase density of makers more or less evenly along every chromosome. Second, they are very helpful in assembling the physical map and ordering the genomic scaffolds. Third, when a prefinished genome sequence is available, they become key players by helping to select BAC clones to close gaps. Finally, BES can provide an easy and rapid way to identify new genes.

The goal of the BAC-ends sequencing project is to produce about 53,000 BES. Producing such a large number of end sequences will allow us on the one hand to obtain high resolution physical and genetic maps and on the other hand to retrieve a large amount of sequence data. Considering a genome size of 200 Mb and a theoretical average read length of 500 bp per sequence, BES will provide a marker approximately every 3.8 kb and a sequence coverage corresponding to at least 13% of the *Tribolium* genome.

6.2 Materials & Methods

BAC library

The BAC genomic library was prepared by Exelixis during a collaboration phase with laboratories in Kansas and our laboratory. The GA-2 strain was used for this purpose. This strain was inbred by single sib pairs for twenty generations. The EcoRI library was constructed using the cloning vector pBACe3.6 (gi4878025). The theoretical coverage of the *Tribolium castaneum* genome is 20X, with an average insert size of 155 kb. The library contains 26,496 clones grided in sixty-nine 384-well plates.

DNA extraction

DNA extraction was performed by Dr. Susan J. Brown in the context of a BAC fingerprinting project. BAC clones were purified using the R.E.A.L. Prep 96 System (Quiagen). The unused part of the DNA preparations was given to us in the format of two hundred and seventy-six 96-well plates containing dried BAC DNA.

Sequencing reaction

Clones were sequenced using the ABI Prism BigDye Terminator v3.1 Cycle Sequencing kit and run on an ABI Prism 3730 DNA Analyzer (Applied Biosystems). DNA was rehydrated overnight in 18 µl of HPLC-H_2O and sequencing reaction were set up as follow: 2 µl of BigDye Terminator v3.1 Ready Reaction Mix, 4 µl of Magic-Dye (Red Rabbit), 1 µl 20 µM primer solution and 8 µl of BAC DNA for a final volume of 15 µl. Reactions were cycled (95°C, 5 min.; [95°C, 30 sec.; 45°C, 20 sec.; 60°C 4 min.] x 80 cycles) and then purified on Sephadex G-50. Primers used are Sp6 (ATTTAGGTGACACTATAGAAG) and T7 (TAATACGACTCACTATAGGG). Reactions were injected 15 sec. and run on short capillaries.

Sequence trace analysis

Basecalling and quality assessments were done using the ABI Prism DNA Sequencing Analysis Software v5.0. Raw traces were basecalled with the ABI basecaller and transferred as such to NCBI dbGSS and TGD. For statistical purpose, basecalling was redone in-house with the KB basecaller, which calculates sample quality values and allows trimming of low quality regions. The clear range of a trace was set such that every base from the ends were removed until fewer than 4 bases out of 20 had quality values less than 20. Analysis reports were produced for each sequencing run.

Annotation

Preliminary functional annotation was performed on the basis of a BLASTx analysis against the non-redundant NCBI protein database. Batch BLAST at NCBI was done using the blastcl3 program. Cut-off value was e<1e-10.

6.3 Results

As of October 2003, 8,640 BES had been generated. This number represents only about 16% of the total number of BES to be generated but it is nevertheless sufficient to gain insights concerning the quality of the data produced and to readily obtain valuable genetic information through sequence analysis. The sequencing data are continuously submitted to dbGSS and Trace Archive at NCBI as well as to TGD so these data will not be presented here.

6.3.1 Quality assessment of the BAC-end sequences

Quality was assessed using three parameters: the percentage of success, the average length of read and the average quality value. A successful read was defined as a sequence of at least 200 bp in length with an average quality value of at least 20. The results of the BES quality analysis are summarized in Table 6.1.

Table 6.1 BES quality analysis	
# of sequence	8,640
% of success	85%
bp sequenced	5.9 Mb
% of the genome	2.9%
Average length of read	810 bp
Average quality value	38

The quality analysis is a combined measure of both T7 and Sp6 sequencing reactions. It is noteworthy that even if the average length of read and quality value is the same for both primers, a small discrepancy exists between the percentages of success of both reactions. T7 reactions are successful 90% of the time against 80% for Sp6. Several Sp6 derived primers with higher GC contents and several modifications to the sequencing protocol were made in an attempt to increase the percentage of success of Sp6 reactions without any positive outcome. Recently, the problem was resolved by decreasing the annealing temperature from 50°C to 45°C. The average success rate, which is currently of 85%, is expected to increase to 90-95%. We can estimate that an amount of sequence data equivalent to at least 18% of the total genome will be produced in the *Tribolium* BAC-ends sequencing project.

6.3.2 Preliminary sequence analysis

A preliminary sequence analysis was conducted over the first 3,648 BES by means of a BLASTx analysis in the same way as it was performed for the ESTs (Chapter 2). Since in this case sequences are not of transcriptional but of genomic origin, the procedure was restricted to the identification of putative ORFs. From the BLASTx analysis against the non-redundant NCBI protein database, 622 significant hits were obtained (e-value < 1e-10). From those, 134 vector hits (21.5%) detected by a search against the pBACe3.6 vector sequence and twenty putative contaminating sequences (3%) were discarded. Putative contaminants are referred as sequences of microbial origin, which are detected at a very high frequency in the BLAST analysis. In the present case, only a hypothetical protein from *Plasmodium yoelii yoelii* detected twenty times was rejected.

We considered that the remaining 468 BES contained putative ORFs. This signifies that on average a putative gene is found in 13% of the single BES reads. Distribution of the top BLAST hit for these 468 ORFs across kingdoms and metazoan phyla is presented in Figure 6.1. To better assess the value of the BLAST results obtained, e-value ranges were split in 1e-10>e>1e-25 and e<1e-25 intervals. The first interval indicates putative ORFs and the second putative orthology between the query and the subject of the BLAST analysis. 1e-25 is an arbitrary cut-off for putative orthology and has not been empirically determined. All the hits are distributed more-or-less equally between both intervals (48.5% and 51.5% respectively)

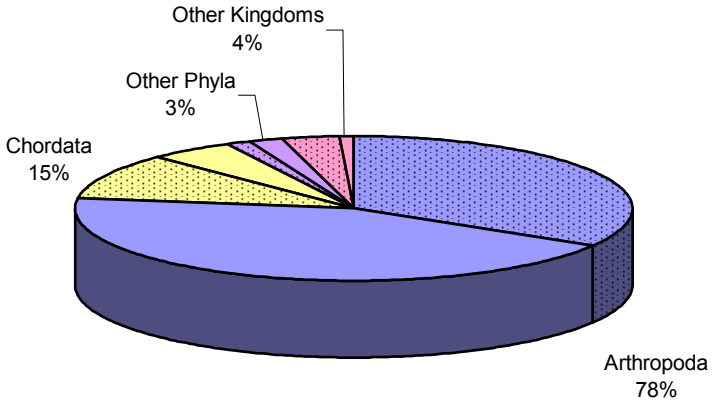

Figure 6.1 Distribution of putative ORFs among kingdoms and metazoan phyla. The shadowed slice of each group corresponds to the 1e-10>e>1e-25 interval and the clear one to the e<1e-25 interval.

Not surprisingly, the vast majority of the BES show high similarity to arthropod (78%) or chordate (18%) sequences while other phyla or kingdoms are seldom represented (7% altogether). This result was expected since almost all of the animal sequences present in NCBI databases come from these two phyla. Further subdivision of the Arthropoda slice into classes and orders (Figure 6.2) help to exemplify the bias in top hits obtained. In this case, Diptera and Lepidoptera that are the only hexapod orders for which large data sets are currently available, monopolize most of the top hits (85% for the Diptera alone).

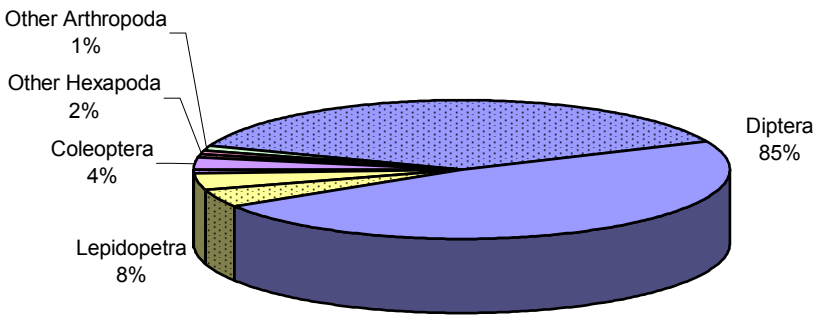

Figure 6.2 Distribution of putative ORFs among Arthropoda classes and orders. Slice colour scheme is as in Figure 6.1.

However, the trend is rapidly changing. Not so long ago *Drosophila melanogaster* was virtually the only heavily sequenced insect and thus the one usually retrieved by BLAST analysis. A closer look at the distribution of hits in Figure 6.2 reveals that the dipteran slice surface begins to shrink at the expense of other orders for which sequence data is rapidly growing. Most notable are the cases of lepidopterans represented by the silkworm *Bombyx mori* and coleopterans represented by *Tribolium* itself. The honeybee sequence data are not yet available in protein databases and is, except in one case, not present in the BLAST analysis. This situation might change in the near future.

Distribution of hits within the Diptera also deserves a closer look (Figure 6.3). Since the release of the data from the *Anopheles gambiae* genome project, top hits distribution has shifted from a majority of *Drosophila* to a majority of *Anopheles* sequences. One can also note that most *Anopheles* hits are in the 1e-25 interval (60%) while only 48% of the *Drosophila* hits can pretend to the same confidence interval. Taken together, these results would suggest that proteins in *Drosophila melanogaster* are more divergent than in *Anopheles gambiae* when compared to *Tribolium*.

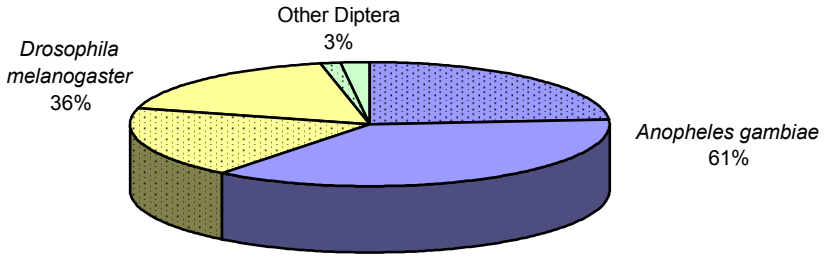

Figure 6.3 Distribution of putative ORFs among Diptera species. Slice colour scheme is as in Figure 6.1.

Classification of BAC-ends ORFs into functional classes on the basis of the BLAST analysis is in progress and is therefore not explicitly presented here. However, preliminary results suggest that putative orthologues of *Drosophila* transcription and translation factors, signalling molecules as well as genes involved in dorso/ventral axis determination, neurogenesis, wing, eye, antenna and sex morphogenesis have been identified. Moreover, genes or genomic regions such as *hedgehog* and the Hox cluster, which are well known in *Tribolium,* have also been found.

6.4 Discussion

The *Tribolium* BAC-ends sequencing project is by itself slightly different from the other projects presented in this thesis since the work is still in progress and will hopefully be completed in the perspective of the *Tribolium* genome project. This situation implicates that the whole analysis is preliminary and somewhat incomplete. Along with a discussion of the results already obtained, I shall therefore take this opportunity to suggest improvements to the data analysis procedure.

The data quality obtained is quite surprising. 85% sequencing success for an average read length of 810 bp is extremely high when compared to a large BAC-ends sequencing project such as the one for human where in average 65% success rate and 460 bp read length was obtained over more than 300,000 BES (Zhao et al., 2000). Our rate of success is probably accurate, because we find for the most part that either these sequencing reactions provide long, high quality reads, or they fail entirely. However, average read length could be subject to interpretation. Our data have been analysed with the software provided by ABI while the human BAC-ends sequencing project used Phred. I believe that average read length in our case is overestimated, because quality value assessment by Phred appears to be more stringent than by the ABI basecaller. It would be more advantageous to use the Phred basecaller to assess quality of the BES produced because this basecaller will be the one used for processing of all the traces in prevision of the genome sequence assembly.

The BLAST analysis resulted in the identification of 468 putative ORFs in 13% of the BES. From this number, we can readily estimate that about 6,900 ORFs will be identified during the course of the *Tribolium* BAC-ends sequencing project. This estimate includes a certain level of redundancy, which has not been estimated. Nevertheless, the high rate at which we find genes suggests that the *Tribolium* genome is compact, possibly containing a low level of repetitive sequences. If it is the case, this will ease the sequencing and assembly of the genome, because repetitive elements are difficult to sequence and problematic to assemble, especially in the context of a whole-genome shotgun approach. From the genes found in the BES, some are already known and mapped in *Tribolium*. The Hox cluster and *hedgeho*g are the only examples for the moment but this result is encouraging. Identification of previously mapped genes, ESTs and other sequence-tagged sites in the BES facilitates mapping of the corresponding BAC clones on the genetic map, which will aid in the integration of physical and genetic maps. In this respect, it is worth mentioning that a subset of the BES produced in Köln are mapped by Dr. Richard Beeman in Kansas, thus increasing the marker density of the *Tribolium* linkage map.

Currently the BES annotation is done only by BLASTx of the NCBI non-redundant database. This approach is simple but minimally informative. Most sequences in the NCBI database are very badly annotated if they are annotated at all.

A BLAST-only approach requires a time consuming ORF-by-ORF analysis to obtain an accurate preliminary functional annotation on the basis of which genes will be selected for cloning and further studies via *in situ* hybridization and pRNAi. One of the obvious modifications of the approach is to search all the NCBI databases along with the very well annotated data from *Drosophila* and the EST data from *Tribolium*. This last step is important since it would potentially allow linking mapped ESTs to specific BAC clones. This large scale BLAST analysis would require database integration, which could potentially be performed at TGD. In addition, BES should be searched for conserved protein domains using Pfam or Interpro databases. Obviously, if a BES contains a protein motif, it will return a highly significant BLAST hit but it is always much easier to understand the biological significance of "Homeobox" or "Zf-C2H2" than "ENSANGP00000013231" or "CG5669" (unannotated proteins from *Anopheles* and *Drosophila* respectively).

In summary, the BAC-ends sequencing project initiated in *Tribolium* will provide 53,000 markers covering every chromosome at an interval of 3.8 kb, resulting in the identification of about 6,900 putative genes and in the production of sequence data equivalent to 18% of the genome. These BES will play key roles in the assembling and finishing of the complete genome sequence of *Tribolium castaneum*.

7. Conclusion

The aim of this study was to identify new genes involved in *Tribolium castaneum* embryonic development. This goal has been achieved by the production and functional analysis of 10,973 ESTs, cDNA clones and BES. Sequence analysis is far from being completed, but we already know that at least 25 genes of high interest, covering almost every aspect of *Tribolium* embryonic development, have been identified and are now being further characterized. This is a relatively small outcome, but more functional analyses are to be performed and many more genes will certainly be discovered and characterized in the near future. Sequence data produced also helped to clarify the evolutionary trend of *Tribolium* and to better illustrate the concept of the molecular clock. Finally, in the context of the *Tribolium* genome project, the work presented here provides key elements for the production of a genome map and contributes significantly to the *Tribolium* Genome Database by providing most of the content.

During the course of this project, I have been able at several occasions to realize, mostly in a frustrating way, the key role that a bioinformatic infrastructure plays in the successful completion of such a project. From the point where a sequence trace is produced to the moment where few genes are chosen for functional characterization, data analysis is done *in silico*. Consequently, without proper data treatment, analysis and storage facilities it is virtually impossible to get all the substance out of the data. Although I have been able to overcome most of the problems associated to data analysis, much work remains to be done in order to ease and standardize in-house data treatment and to make the TGD fully operational.

In retrospective, such a high throughput approach might seem an odd enterprise for a single person especially within a PhD degree. Hence, the data presented in this thesis could have potentially been produced within one or two weeks by any genome sequencing facility. However, such a homemade approach is worthwhile in two respects. First, it can be adapted to any organism of interest with a relatively low amount of effort, thus easily providing large amounts of raw data for genes discovery

and beyond. Second, it can provide the initial information necessary for the selection of an organism for whole genome sequencing. Availability of a large amount of sequence data along with libraries, maps, and databases are all elements demonstrating the interest of the scientific community for an organism.

All of these considerations taken together, I consider this study a good starting point to a genomic approach for the study of *Tribolium* genetics, development and evolution. The work presented here will certainly help to finally make of *Tribolium castaneum* a model organism of its own.

8. Literature

Adachi J, Cao Y, Hasegawa M (1993) Tempo and mode of mitochondrial DNA evolution in vertebrates at the amino acid sequence level: rapid evolution in warm-blooded vertebrates. J Mol Evol 36:270-281

Adams MD, Kelley JM, Gocayne JD, Dubnick M, Polymeropoulos MH, Xiao H, Merril CR, Wu A, Olde B, Moreno RF, Kerlavage AR, McCombie WR, Venter JC (1991) Complementary DNA sequencing: expressed sequence tags and human genome project. Science 252:1651-1656

Arendsen Hein SA (1920) Studies on variation in the mealworm *Tenebrio molitor* L. I. Biological and genetical notes on *Tenebrio molitor* L. J Genet 10:227-263

Arendsen Hein SA (1924a) Selektionsversuche mit Prothorax- und Elytravariationen bei *Tenebrio molitor*. Entomol Mitt 153:243-275

Arendsen Hein SA (1924b) Studies on variation in the mealworm *Tenebrio molitor* L. II. Variation in tarsi and antennae. J Genet 14:1-38

Ayala FJ, Barrio E, Kwiatowski J (1996) Molecular clock or erratic evolution? A tale of two genes. Proc Natl Acad Sci USA 93:11729-11734

Beeman RW, Brown SJ (1999) RAPD-based genetic linkage maps of *Tribolium castaneum*. Genetics 153:333-338

Beeman RW, Stuart JJ, Brown SJ, Denell RE (1993) Structure and function of the homeotic gene complex (HOM-C) in the beetle *Tribolium castaneum*. Bioessays 15:439-444

Beeman RW, Stuart JJ, Haas MS, Denell RE (1989) Genetic analysis of the homeotic gene complex (HOM-C) in the beetle *Tribolium castaneum*. Dev Biol 133:196-209

Beermann A, Jay DG, Beeman RW, Hulskamp M, Tautz D, Jurgens G (2001) The *Short antennae* gene of *Tribolium* is required for limb development and encodes the orthologue of the *Drosophila* Distal-less protein. Development 128:287-297

Berghammer AJ, Klingler M, Wimmer EA (1999) A universal marker for transgenic insects. Nature 402:370-371

Bonaldo MF, Lennon G, Soares MB (1996) Normalization and subtraction: two approaches to facilitate gene discovery. Genome Res 6:791-806

Britten RJ (1986) Rates of DNA sequence evolution differ between taxonomic groups. Science 231:1393-1398

Bromham L, Penny D (2003) The modern molecular clock. Nat Rev Genet 4:216-224

Brown SJ, Parrish JK, Beeman RW, Denell RE (1997) Molecular characterization and embryonic expression of the *even-skipped* ortholog of *Tribolium castaneum*. Mech Dev 61:165-173

Brown SJ, Patel NH, Denell RE (1994) Embryonic expression of the single *Tribolium engrailed* homolog. Dev Genet 15:7-18

Bucher G, Scholten J, Klingler M (2002) Parental RNAi in *Tribolium* (Coleoptera). Curr Biol 12:R85-R86

Caccone A, Powell JR (1990) Extreme rates and heterogeneity in insect DNA evolution. J Mol Evol 30:273-280

Carmean D, Crespi BJ (1995) Do long branches attract flies? Nature 373:666

Carmean D, Kimsey LS, Berbee ML (1992) 18S rDNA sequences and the holometabolous insects. Mol Phylogenet Evol 1:270-278

Catzeflis FM, Sheldon FH, Ahlquist JE, Sibley CG (1987) DNA-DNA hybridization evidence of the rapid rate of muroid rodent DNA evolution. Mol Biol Evol 4:242-253

Doctor J, Bennett R, Sanchez-Salazar J, Pletcher M, Denell R (1995) Sequence and expression of a TGF-beta / *decapentaplegic*-like gene in the red flour beetle, *Tribolium castaneum*. Dev Biol 170:743

Doctor J, Sanchez-Salazar J, Pletcher M, Bennett R, Denell R (1996) The *decapentaplegic* gene of the red flour beetle, *Tribolium castaneum*, is similar to *Drosophila dpp* in sequence, structure, and expression. Dev Biol 175:380

Ewing B, Green P (1998) Base-calling of automated sequencer traces using phred. II. Error probabilities. Genome Res 8:186-194

Ewing B, Hillier L, Wendl MC, Green P (1998) Base-calling of automated sequencer traces using phred. I. Accuracy assessment. Genome Res 8:175-185

Falciani F, Hausdorf B, Schroder R, Akam M, Tautz D, Denell R, Brown S (1996) Class 3 Hox genes in insects and the origin of *zen*. Proc Natl Acad Sci USA 93:8479-8484

Ferwerda FP (1928) Genetische Studien am Mehlkäfer. Genetica 11:1-111

Friedrich M, Tautz D (1997) An episodic change of rDNA nucleotide substitution rate has occurred during the emergence of the insect order Diptera. Mol Biol Evol 14:644-653

Goldman N, Yang ZH (1994) Codon-based model of nucleotide substitution for protein-coding DNA sequences. Mol Biol Evol 11:725-736

Good NE (1936) The flour beetles of the genus *Tribolium*. USDA Tech Bull 498:1-58

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

Green ED (2001) Strategies for the systematic sequencing of complex genomes. Nat Rev Genet 2:573-583

Gu X, Li WH (1992) Higher rates of amino acid substitution in rodents than in humans. Mol Phylogenet Evol 1:211-214

Hennig W (1981) Insect Phylogeny, New York

Hinton HE (1948) A synopsis of the genus *Tribolium* Macleay with some remarks on the evolution of its species-group (Coleoptera, Tenebrionidae). Bull Entomol Res 39:13-55

Jones DT, Taylor WR, Thornton JM (1992) The rapid generation of mutation data matrices from protein sequences. Comput Appl Biosci 8:275-282

Kimura M, Ohta T (1971) On the rate of molecular evolution. J Mol Evol 1:1-17

Kukalova-Peck J (1991) Fossil history and evolution of hexapod structures. In: Nauman ID (ed) The Insects of Australia. Melbourne University Press, Melbourne, Vol. Vol. 1 p 141-179

Lehmann R, Tautz D (1994) I*n situ* hybridization to RNA. In: Methods in Cell Biology, Vol. 44 p 575-598

Lewis EB (1978) A gene complex controlling segmentation in *Drosophila*. Nature 276:565-570

Li YB, Brown SJ, Hausdorf B, Tautz D, Denell RE, Finkelstein R (1996) Two *orthodenticle*-related genes in the short-germ beetle *Tribolium castaneum*. Dev Genes Evol 206:35-45

Maderspacher F, Bucher G, Klingler M (1998) Pair-rule and gap gene mutants in the flour beetle *Tribolium castaneum*. Dev Genes Evol 208:558-568

Miyata T, Yasunaga T, Nishida T (1980) Nucleotide sequence divergence and functional constraint in mRNA evolution. Proc Natl Acad Sci USA 77:7328-7332

Moriyama EN (1987) Higher rates of nucleotide substitution in *Drosophila* than in mammals. Jpn J Genetics 62:139-147

Nagy LM, Carroll S (1994) Conservation of *wingless* patterning functions in the short-germ embryos of *Tribolium castaneum*. Nature 367:460-463

Nusslein-Volhard C, Wieschaus E (1980) Mutations affecting segment number and polarity in *Drosophila*. Nature 287:795-801

Ohta T (1987) Very slightly deleterious mutations and the molecular clock. J Mol Evol 26:1-6

Ohta T, Kimura M (1971) On the constancy of the evolutionary rate of cistrons. J Mol Evol 1:18-25

Park T (1934) Observations on the general biology of the flour beetle *Tribolium confusum*. Quart Rev Biol 9:36-54

Park T (1937) The inheritance of the mutation "pearl" in the flour beetle *Tribolium castaneum* Herbst. Am Nat 71:143-157

Park T (1948) Experimental studies of inter-species competition. 1. Competition between populations of the flour beetles *Tribolium confusum* Duval and *Tribolium castaneum* Herbst. Ecol Monogr 18:265-307

Park T (1954) Experimental studies of inter-species competition. 2. Temperature, humidity, and competition in 2 species of *Tribolium*. Physiol Zool 27:177-238

Park T (1957) Experimental studies of inter-species competition. 3. Relation of initial species proportion to competitive outcome in populations of *Tribolium*. Physiol Zool 30:22-40

Park T, Mertz DB, Grodzinski W, Prus T (1965) Cannibalistic predation in populations of flour beetles. Physiol Zool 38:289-321

Pesole G, Gissi C, De Chirico A, Saccone C (1999) Nucleotide substitution rate of mammalian mitochondrial genomes. J Mol Evol 48:427-34

Prpic NM, Wigand B, Damen WG, Klingler M (2001) Expression of *dachshund* in wild-type and *Distal-less* mutant *Tribolium* corroborates serial homologies in insect appendages. Dev Genes Evol 211:467-477

Rodriguez-Trelles F, Tarrio R, Ayala FJ (2001) Erratic overdispersion of three molecular clocks: GPDH, SOD, and XDH. Proc Natl Acad Sci USA 98:11405-11410

Sokoloff A (1966) The Genetics of *Tribolium* and Related Species. Academic Press, New York, 212 p.

Sokoloff A (1972) The Biology of *Tribolium* With Special Emphasis on Genetic Aspects 1. Oxford University Press, London, 300 p.

Sokoloff A (1974) The Biology of *Tribolium* With Special Emphasis on Genetic Aspects 2. Oxford University Press, London, 610 p.

Sokoloff A (1977) The Biology of *Tribolium* With Special Emphasis on Genetic Aspects 3. Oxford University Press, London, 612 p.

Sokoloff A, Lerner IM (1967) Laboratory ecology and mutual predation of *Tribolium* species. Am Nat 101:261-276

Sommer RJ, Tautz D (1993) Involvement of an orthologue of the *Drosophila* pair-rule gene *Hairy* in segment formation of the short germ-band embryo of *Tribolium* (Coleoptera). Nature 361:448-450

Sommer RJ, Tautz D (1994) Expression patterns of *twist* and *snail* in *Tribolium* (Coleoptera) suggest a homologous formation of mesoderm in long and short germ band insects. Dev Genet 15:32-37

Stuart JJ, Brown SJ, Beeman RW, Denell RE (1991) A deficiency of the homeotic complex of the beetle *Tribolium*. Nature 350:72-74

Takano TS (1998) Rate variation of DNA sequence evolution in the *Drosophila* lineages. Genetics 149:959-970

Tautz D, Friedrich M, Schroder R (1994) Insect embryogenesis - what is ancestral and what is derived? Development Supplement:193-199

Thompson JD, Higgins DG, Gibson TJ (1994) Clustal W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res 22:4673-4680

Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA, Gocayne JD, Amanatides P, Ballew RM, Huson DH, Wortman JR, Zhang Q, Kodira CD, Zheng XQH, Chen L, Skupski M, Subramanian G, Thomas PD, Zhang JH, Miklos GLG, Nelson C, Broder S, Clark AG, Nadeau C, McKusick VA, Zinder N, Levine AJ, Roberts RJ, Simon M, Slayman C, Hunkapiller M, Bolanos R, Delcher A, Dew I, Fasulo D, Flanigan M, Florea L, Halpern A, Hannenhalli S, Kravitz S, Levy S, Mobarry C, Reinert K, Remington K, Abu-Threideh J, Beasley E, Biddick K, Bonazzi V, Brandon R, Cargill M, Chandramouliswaran I, Charlab R, Chaturvedi K, Deng ZM, Di Francesco V, Dunn P, Eilbeck K, Evangelista C, Gabrielian AE, Gan W, Ge WM, Gong FC, Gu ZP, Guan P, Heiman TJ, Higgins ME, Ji RR, Ke ZX, Ketchum KA, Lai ZW, Lei YD, Li ZY, Li JY, Liang Y, Lin XY, Lu F, Merkulov GV, Milshina N, Moore HM, Naik AK, Narayan VA, Neelam B, Nusskern D, Rusch DB, Salzberg S, Shao W, Shue BX, Sun JT, Wang ZY, Wang AH, Wang X, Wang J, Wei MH, Wides R, Xiao CL, Yan CH, et al. (2001) The sequence of the human genome. Science 291:1304-1351

Werman SD, Davidson EH, Britten RJ (1990) Rapid evolution in a fraction of the *Drosophila* nuclear genome. J Mol Evol 30:281-289

Wheeler SR, Carrico ML, Wilson BA, Brown SJ, Skeath JB (2003) The expression and function of the *achaete-scute* genes in *Tribolium castaneum* reveals conservation and variation in neural pattern formation and cell fate specification. Development 130:4373-4381

Whitfield CW, Band MR, Bonaldo MF, Kumar CG, Liu L, Pardinas JR, Robertson HM, Soares MB, Robinson GE (2002) Annotated expressed sequence tags and cDNA microarrays for studies of brain and behavior in the honey bee. Genome Res 12:555-566

Wigand B, Bucher G, Klingler M (1998) A simple whole mount technique for looking at *Tribolium* embryos. Tribolium Information Bulletin 38:281-283

Wolff C, Sommer R, Schroder R, Glaser G, Tautz D (1995) Conserved and divergent expression aspects of the *Drosophila* segmentation gene *hunchback* in the short germ band embryo of the flour beetle *Tribolium*. Development 121:4227-4236

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

Zhao S, Malek J, Mahairas G, Fu L, Nierman W, Venter JC, Adams MD (2000) Human BAC ends quality assessment and sequence analyses. Genomics 63:321-332

Zuckerkandl E, Pauling L (1962) Molecular disease, evolution, and genic heterogeneity. In: Kasha M, Pullman B (eds) Horizons in Biochemistry. Academic Press, New York, Vol. 189-225

Zuckerkandl E, Pauling L (1965) Evolutionary divergence and convergence in proteins. In: Bryson V, Vogel HJ (eds) Evolving Genes and Proteins. Academic Press, New York, Vol. 97-166

9. Appendix

Appendix I – Summary of the EST project

TcEST Contigs	NCBI tBLASTx (*Drosophila*)	NCBI tBLASTx (Human)	BDGP BLASTx (e<1e-03)	Putative function	General function
Tc012D01R, Tc012D01F	3e-36	3e-34	CG13090	Molybdopterin cofactor sulfurase	Enzyme
Tc025C03R	8e-55	2e-48	Cyp1	Cyclophilin	Enzyme
Tc001F10R, Tc001F10F	2e-68	2e-65	BG, DS00004.11	Signal peptidase	Enzyme
Tc001H11R, Tc013A05R	1e-47	3e-31	ial	Protein kinase	Enzyme
Tc002G02R, Tc002G02F	3e-17	1.1	CG14321	Unknown	Unknown
Tc002G12R, Tc002G12F	2e-16	5e-16	EST LD34570 from BDGP BLASTn	Ribosome associated membrane protein (from NCBI tBLASTx result)	Unknown
Tc004B04R, Tc004B04F	1e-133	1e-126	Fib	Small nuclear ribonucleoprotein	RNA binding
Tc004E02F, Tc004E02R	1e-13	6e-05	CG11455	Unknown	Unknown
Tc004E12R, Tc004E12F	1e-27	1.9	BAC clones BACR09N11 and BACR40A15 from NCBI tBLASTx	Unknown	Unknown
Tc020G10R, Tc004F04R	3e-90	1e-89	Rack1	Signal transducer	Signal transduction
Tc004G02R, Tc004G02F	5e-72	8e-66	viaf1	Unknown	Unknown
Tc005C05R, Tc020F10R	8e-12	1.5	CG8668	UDP-galactose beta-N-acetylglucosamine beta-1,3-galactosyltransferase	Enzyme
Tc005C07R, Tc005C07F	1e-76	9e-66	TfIIEbeta	General transcription factor	Transcription factor
Tc005G07F, Tc005G07R	2e-43	5e-35	Srp19	RNA binding, signal recognition particule	RNA binding
Tc005G11R, Tc005G11F	7e-44	7e-39	lwr	Protein nucleus import, possible gap gene	Ligand binding or carrier
Tc005G12R, Tc005G12F	1e-106	1e-96	Ef2b	Translation elongation factor	Translation factor

TcEST Contigs	NCBI tBLASTx (*Drosophila*)	NCBI tBLASTx (Human)	BDGP BLASTx (e<1e-03)	Putative function	General function
Tc006A09R, Tc023F08R	3e-20	2.3	CG2004	Unknown	Unknown
Tc006B10R, Tc006B10F	7e-51	9e-34	Gst2	Glutathione transferase	Enzyme
Tc006D05R, Tc006D05F	7e-28	1e-07	KLP54D	Kinesin motor	Other
Tc006E11R, Tc006E11F	3e-69	8e-54	CG11999	Unknown	Unknown
Tc006F08R, Tc023B08R, Tc023B08F	2e-63	7e-51	eIF-3p66	Translation initiation factor	Translation factor
Tc007C06R, Tc007C06F	3e-26	6e-06	CG14331	Unknown	Unknown
Tc007D01F, Tc007D01R	1e-57	4e-50	awd	Nucleoside-diphosphate kinase	Enzyme
Tc008B11R, Tc008B11F	2e-21	2e-13	CG14672	Unknown	Unknown
Tc008C06R, Tc008C06F	1e-149	1e-108	CG1972	Unknown	Unknown
Tc012B02R, Tc008C08R	9e-58	4e-50	CG6543	Short-chain enoyl-CoA hydratase	Enzyme
Tc008D05F, Tc008D05R	5e-24	1e-21	BAC clone BACR21H10 from BDGP BLASTn	Unknown	Unknown
Tc008F12R, Tc008F12F	5e-27	8e-19	GalNAC-T1	Polypeptide N-acetylgalactosaminyltransferase	Enzyme
Tc008G09R, Tc008G09F	4e-43	5e-30	ARP-like	Unknown	Unknown
Tc008H12F, Tc008H12R	3e-17	1e-14	CG6513	Sulfonylurea receptor ligand, signal transduction	Signal transduction
Tc009D08F, Tc009D08R	1e-42	1e-36	CG5454	RNA binding, mRNA splicing	RNA binding
Tc009F01F, Tc009F01R	6e-59	9e-45	CG8745	Ornithine-oxo-acid aminotransferase	Enzyme
Tc012E11R, Tc012E11F	2e-31	4e-08	BAC clone BACR46G10, similarity to Hm GL004 in NCBI tBLASTx	Unknown	Unknown
Tc014C04R, Tc014C04F	1e-94	5e-83	DHPR	Dihydropteridine reductase	Enzyme
Tc015C05R, Tc015C05F	2e-24	3e-16	BAC clone BACR32M04	Unknown	Unknown
Tc015F02R, Tc015F02F	1e-102	3e-90	Cyc C	Cell cycle regulator, cyclin	Cell cycle regulator
Tc015F08F, Tc015F08R	6e-06	1.1	CG13043	Unknown	Unknown
Tc016E01R, Tc016E01F	8e-61	2e-41	desat1	Stearoyl-CoA desaturase	Enzyme
Tc016G01R, Tc016G01F	7e-56	5e-24	Tsp66E	Tumor suppressor	Other
Tc016G03R, Tc016G03F	2e-34	9e-21	Cyp18a1	Cytochrome P450, steroid biosynthesis	Other
Tc016H02R, Tc016H02F	7e-29	1.1	PebIII	Ligand binding or carrier	Ligand binding or carrier
Tc016H10R, Tc016H10F	4e-34	4.4e-01	CG14430	Unknown	Unknown

TcEST Contigs	NCBI tBLASTx (*Drosophila*)	NCBI tBLASTx (Human)	BDGP BLASTx (e<1e-03)	Putative function	General function
Tc020A11R, Tc020A11F	1e-40	2e-39	CG17401	LIM transcription factor	Transcription factor
Tc020C08R, Tc020C08F	2e-55	2e-55	dod	Peptidylprolylisomerase	Enzyme
Tc020F02R, Tc020F02F	2e-18	3.1	CG12926	Tocopherol binding, lignand binding or carrier	Ligand binding or carrier
Tc020G04R, Tc020G04F	2e-22	6.8e-02	rad201	DNA repair protein	Other
Tc020H03R, Tc020H03F	2e-16	7e-03	l(3)3670	Unknown	Unknown
Tc020H06R, Tc020H06F	2e-42	1e-30	kappaB-Ras	RAS small monomeric GTPase	Enzyme
Tc020H10R, Tc020H10F	5e-19	2e-11	Tsp42Ee	Cell adhesion molecule	Other
Tc021A12F, Tc021A12R	6e-54	1e-52	CG6724	Signal transduction	Signal transduction
Tc021G09R, Tc021G09F	2e-21	2e-21	CG11500	Unknown	Unknown
Tc021G11R, Tc021G11F	2e-53	3e-24	Mlc2	Ligand binding or carrier, muscle motor	Ligand binding or carrier
Tc022C08R, Tc022C08F	1e-141	1e-133	Arp66B	Structural protein of cytoskeleton	Structural protein
Tc022F05R, Tc022F05F	1e-38	3e-36	CG18591	Unknown	Unknown
Tc023C04F, Tc023C04R	3e-49	3e-15	genomic scaffold	Unknown	Unknown
Tc023C11R, Tc023C11F	6e-17	1e-16	CG4452	Unknown	Unknown
Tc023F10R, Tc023F10F	2e-28	3e-14	CG3817	Unknown	Unknown
Tc024E11F, Tc024E11R	7e-24	6e-15	CG9723	Unknown	Unknown
Tc025C04R, Tc025C04F	1e-44	2e-41	CG11490	GTPase activator	Other
Tc025G04F, Tc025G04R	2e-16	1e-14	CG9455	serpin	Other
Tc026A08R, Tc026A08F	7e-31	1e-12	CG15747	Unknown	Unknown
Tc026C02R, Tc026C02F	4e-31	1.7	S	Unknown	Unknown
Tc026D12R, Tc026D12F	1e-24	1.9e-01	CG8927	Unknown	Unknown
Tc026E07F, Tc026E07R	2e-53	3e-08	wupA	actin binding	Other
Tc001A08R, Tc020H01F, Tc008C04R, Tc020H01R, Tc001A08F	0	0	Ef1alpha100E	Translation elongation , actin binding	Translation factor
Tc001A09R, Tc006C02R, Tc001A09F	3e-94	7e-89	Cyp1	Cyclophilin	Enzyme
Tc003B11R, Tc011A11R, Tc003B11F	3e-09	5.5e-01	CG7215	Ubiquitin-like	Other
Tc025B10R, Tc003E01R, Tc025B10F	1e-45	1e-41	Eb1	Microtubule binding, cytoskeletal structural protein	Structural protein

TcEST Contigs	NCBI tBLASTx (*Drosophila*)	NCBI tBLASTx (Human)	BDGP BLASTx (e<1e-03)	Putative function	General function
Tc027H04R, Tc004C01R, Tc004C01F	2e-49	1e-48	CG13926	Unknown	Unknown
Tc004D02R, Tc027G06R, Tc004E01R	5e-09	1.7e-01	CG6469	Structural protein of larval cuticle	Structural protein
Tc013G07R, Tc004F06R, Tc010B01R	8e-51	3e-10	CG4692	Hydrogen-translocating F-type ATPase	Enzyme
Tc004F07R, Tc004F07F, Tc020F06R	1e-89	4e-76	Prosbeta5	Multicatalytic endopeptidase	Enzyme
Tc009C09F, Tc005D12R, Tc009C09R	3e-36	2e-18	CG1458	Unknown	Unknown
Tc005F12R, Tc014B10F, Tc014B10R	9e-12	1.8	Ocho	Unknown	Unknown
Tc024D12R, Tc007A12R, Tc025C02R	8e-04	4.6e-01	CG12991	Unknown	Unknown
Tc007C02R, Tc014D03R, Tc014D03F	2e-37	2e-06	Fer2LCH	Ferrous iron binding, lignand binding or carrier	Ligand binding or carrier
Tc007E02R, Tc025F12R, Tc013H12R	3e-57	3e-56	FK506-bp2	FK506 binding, peptidylprolyl isomerase	Enzyme
Tc007E12R, Tc025B03R, Tc007E12F	2e-59	2e-53	Thioloase	Acetyl-CoA C-acyltransferase	Enzyme
Tc008B07R, Tc021B09R, Tc026C04R	1e-15	1e-04	CG10374	Chaperone	Other
Tc009G10F, Tc025E08R, Tc009G10R	3e-26	1e-09	CG3420	Unknown	Unknown
Tc009H09R, Tc015D07R, Tc009H09F	3e-27	7e-20	CG10874	Unknown	Unknown
Tc014E01R, Tc014E01F, Tc010E03R	1e-68	6e-63	Px6005	peroxidase	Enzyme
Tc011A04R, Tc013E04R, Tc013E04F	2e-51	2.1	BcDNA, GH02976	Structural protein of peritrophic membrane	Structural protein
Tc011A12R, Tc027H10F, Tc027H10R	9e-36	7e-27	Gst2	Glutathione transferase	Enzyme
Tc025H12F, Tc025H09R, Tc025H12R	2e-15	9e-08	aop	Ets-domain transcription factor	Transcription factor
Tc003G05R, Tc001D04F, Tc001D04R, Tc022G10R	4e-16	2e-06	CG6961	RNA binding	RNA binding
Tc014E02R, Tc026G03R, Tc022E11R, Tc001G05R	7e-62	2e-59	His2Av	Histone, chromatin assembly/disassembly	Other
Tc021C06R, Tc005B09R, Tc001H04R, Tc001H04F	5e-48	4e-34	Rbp1-like	RNA binding	RNA binding
Tc003C07R, Tc025C12R, Tc004A04R, Tc004A04F	2e-15	6.7e-01	CG10112	Unknown	Unknown
Tc007H05R, Tc010E05R, Tc005H05R, Tc007H05F	2e-29	2e-18	HLHmbeta	HLH transcription factor	Transcription factor
Tc011B02R, Tc024F11R, Tc013F02R, Tc022B08R, Tc011B02F	1e-98	3e-98	alphaTub84B	Structural protein of cytoskeleton	Structural protein
Tc023E06R, Tc015E03R, Tc020C12R, Tc020C12F	3e-26	6e-25	CG6783	Ligand binding or carrier	Ligand binding or carrier
Tc003D10R, Tc023F06R, Tc022B12R, Tc022B12F	6e-59	7e-55	bic	Bicaudal-like transcription factor	Transcription factor

TcEST Contigs	NCBI tBLASTx (*Drosophila*)	NCBI tBLASTx (Human)	BDGP BLASTx (e<1e-03)	Putative function	General function
Tc026F10R, Tc013A11R, Tc024H03R, Tc003H05R, Tc024H03F	1e-137	1e-128	ran	RAN small monomeric GTPase	Enzyme
Tc025H01R, Tc010C07F, Tc025B07R, Tc027A02R, Tc010C07R	1e-58	5e-55	BEST, CK01296	Lignand binding or carrier	Ligand binding or carrier
Tc012B12R, Tc001A10R, Tc006C04R, Tc002E03R, Tc020F07R, Tc001A10F	8e-91	3e-26	CG4800	Unknown	Unknown
Tc010C04R, Tc008F02R, Tc001H08R, Tc001B01R, Tc001H08F, Tc001B01F	2e-27	2e-26	l(2)efl	chaperone	Other
Tc024G08R, Tc015H05R, Tc024G02R, Tc002H08R, Tc026G09R, Tc002H08F	2e-35	4e-23	BcDNA, GM12291	Unknown	Unknown
Tc024B07R, Tc015C07R, Tc003C04R, Tc014A07R, Tc013D06R, Tc013D06F	1e-134	1e-118	Mus209	DNA repair protein	Other
Tc005B11R, Tc023C09R, Tc007A09R, Tc004G06R, Tc015A04R, Tc010H10R	6e-69	1.2e-01	chic	Actin binding	Other
Tc007A10R, Tc003B08R, Tc008D08R, Tc021G12R, Tc020B03R, Tc008D08F	4e-43	1e-29	smt3	Protein tagging, protein-nucleus import	Other
Tc013E03R, Tc011E03R, Tc008F03R, Tc008E02R, Tc009G04R, Tc016G08R, Tc008F03F	1e-86	2e-69	eIF-5A	translation factor	Translation factor
Tc012C08R, Tc022C05R, Tc022H07R, Tc002G03R, Tc025C10F, Tc012C08F, Tc021H03R, Tc025C10R	1e-36	2e-16	CG15715	Zf-C2H2 protein	Transcription factor
Tc021H10R, Tc015E10R, Tc021H10F, Tc008A09R, Tc009B06R, Tc010B12R, Tc027D12R	6e-16	1.9e-01	Nlp	DNA binding	Other
Tc023E12R, Tc009H07R, Tc026F07R, Tc001D11R, Tc001C11R, Tc001C11F, Tc026F07F, Tc025D06R	2e-11	4e-11	guf	Ornithine decarboxylase inhibitor	Other
Tc001F03F, Tc022D04R, Tc001F03R, Tc001H03R, Tc011E12R, Tc011D01R, Tc027A05R, Tc020H08R, Tc005H12R	3e-34	1e-32	CG12262	Acyl-CoA deshydrogenase	Enzyme
Tc011B11R, Tc012B06R, Tc005F02R, Tc002H06R, Tc006H09R, Tc006C10R, Tc020F09R, Tc002A06R, Tc006H09F	1e-102	1e-96	eff	ubiquitin conjugating enzyme	Enzyme
Tc010A01R, Tc014H07R, Tc014E04R, Tc007H06R, Tc016D10R, Tc011A01R, Tc006D02R, Tc020F05R, Tc008H07R, Tc023H01R, Tc022E06R, Tc006D02F	3e-19	6	CG10407	Unknown	Unknown
Tc013F09F, Tc005E06F, Tc004G03F, Tc011A08F, Tc015C04R, Tc013F09R, Tc004G03R, Tc005E06R, Tc007F12R, Tc015F10R, Tc011A08R, Tc024B04R	8e-07	1.8e-02	CG4784	Cuticle protein	Structural protein
Tc014C10R, Tc004G01R, Tc007F02R, Tc010C09R, Tc009C12R, Tc005G01R, Tc005H08R, Tc005F07R, Tc003C01R, Tc007B12R, Tc010F07R, Tc007C11R, Tc004A06R	2e-86	6e-09	tsr	Actin binding	Other

TcEST Contigs	NCBI tBLASTx (*Drosophila*)	NCBI tBLASTx (Human)	BDGP BLASTx (e<1e-03)	Putative function	General function
Tc020E11R, Tc015A01R, Tc021D06R, Tc010A07R, Tc007A01R, Tc023B01R, Tc020G08R, Tc002G01R, Tc002F03R, Tc021H06R, Tc012F10R, Tc006E07R, Tc024F12R	3e-84	1e-83	His3.3A	Histone	Other
Tc015G03R, Tc025F02R, Tc027B06R, Tc002C06R, Tc003D07R, Tc005F10R, Tc027A01R, Tc009A05R, Tc020F01R, Tc002G08R, Tc005D08R, Tc005H10R, Tc005D10R, Tc024B05R, Tc008D06R, Tc008B06R, Tc014D11R	7e-09	1.2e-01	CG6469	Cuticle protein	Structural protein
Tc007F10R	1e-28	2e-17	CG1268	Hydrogen-translocating V-type ATPase	Enzyme
Tc009D07R	3e-29	2e-27	CG6668	Unknown	Unknown
Tc008A03R, Tc008A03F	8e-68	2e-67	Hsp83	Chaperone	Other
Tc016B10F, Tc016B10R	8e-33	5.7e-02	CG9175	DNA binding, WD40 motif	Transcription factor (possible)
Tc024G10R, Tc024G10F	1e-13	8e-11	CG9914	Unknown	Unknown
Tc025D04R	2e-08	3e-02	Hsc70Cb	Chaperone	Other
Tc007F09F, Tc007F09R	1e-15	1e-08	cos	Motor protein/microtubule binding	Other
Tc021E09R, Tc011G03R, Tc021E09F	5e-19	1e-14	SF2	Ore-mRNA splicing factor, RNA binding	RNA binding
Tc006A05R	9e-06	1.7e-01	CG5100	Unknown	Unknown
Tc008F01R, Tc008F01F	2e-76	7e-74	Rpn11	Endopeptidase	Enzyme
Tc020E06R	9e-26	3e-23	CG12141	Lysine-tRNA ligase	Enzyme
Tc020A08F, Tc020A08R	4e-34	7e-32	Pk91C	Protein serine/threonine kinase/protein kinase	Enzyme
Tc020E10R	2e-37	5e-05	lola	BTB/POZ, Zf-C2H2 transcription factor	Transcription factor
Tc013C01R, Tc013C01F	2e-39	4e-16	CG4914	Endopeptidase	Enzyme
Tc005B01R	8e-26	8e-10	CG4923	Unknown	Unknown
Tc001B09F, Tc001B09R	8e-40	5e-39	CG8258	Chaperonin ATPase	Enzyme
Tc022C10R, Tc022C10F	1e-60	8e-60	CG5651	Ribonuclease inhibitor, enzyme inhibitor	Other
Tc005H04R	2e-20	2e-18	pcan	Heparin sulfate proteoglycan, cell adhesion molecule	Structural protein
Tc001B11F, Tc001B11R	8e-43	3e-34	CG4908	Chaperonin ATPase	Enzyme
Tc021G03R	1e-11	3e-04	gol	Zf-C3HC4 transcription factor	Transcription factor

TcEST Contigs	NCBI tBLASTx (*Drosophila*)	NCBI tBLASTx (Human)	BDGP BLASTx (e<1e-03)	Putative function	General function
Tc020E02R	4e-39	7e-30	CG3808	RNA binding	RNA binding
Tc003D11R, Tc003D11F	5e-11	3e-06	CDC45L	DNA replication factor	Other
Tc020B06R, Tc020B06F	1e-19	3e-15	Cp1	Cathepsin L	Enzyme
Tc006H12R, Tc006H12F	1e-46	2e-39	M(2)21AB	Methionine adenosyltransferase	Enzyme
Tc007H07R, Tc007H07F	3e-25	7e-14	CG12026	Unknown	Unknown
Tc014B04F, Tc014B04R	2e-84	8e-83	Cdk7	Cyclin-dependent protein kinase/general RNA polymerase II transcription factor (TFIIH)/protein serine/threonine kinase/protein kinase	Transcription factor
Tc025E11F, Tc025E11R	1e-09	8e-07	fzy	Cell cycle regulator, degradation of cyclin	Cell cycle
Tc014G01R, Tc014G01F	1e-35	2e-25	Snr1	Transcription factor	Transcription factor
Tc008C02R	2e-06	8.3e-02	unknown gene	Unknown	Unknown
Tc008D01R, Tc008D01F	2e-25	2e-20	CG5116	Unknown	Unknown
Tc012D05F, Tc012D05R	6e-53	1e-43	CG12262	Acyl-CoA dehydrogenase	Enzyme
Tc001H01R	7e-07	1e-05	nop5	tRNA processing	Other
Tc004G12R, Tc004G12F	7e-24	9e-20	l(2)35Aa	Polypeptide N-acetylgalactosaminyltransferase	Enzyme
Tc021D12R, Tc021D12F	3e-37	1	CG13076	Unknown	Unknown
Tc027G01R	1e-17	7e-04	CG4552	Unknown	Unknown
Tc007H03R	1e-04	6.3e-01	CG16980	Unknown	Unknown
Tc008G05F, Tc008G05R	3e-35	4e-29	SMC1	Motor, chromatid cohesion	Other
Tc001B02F, Tc001B02R	5e-37	5e-32	CG11980	Unknown	Unknown
Tc026F06R, Tc021B06F, Tc004D06R, Tc021B06R, Tc026F06F	1e-72	2e-58	CG8243	Enzyme activator	Other
Tc005D01R, Tc005A02R, Tc021A06R, Tc021A06F	6e-31	2e-20	Pros54	Endopeptidase	Enzyme
Tc014D08F, Tc014D08R	4e-29	2e-28	CG17540	Pre-mRNA splicing factor, RNA binding	RNA binding
Tc006F12F, Tc006F12R	4e-15	2e-10	Cyp4g1	Cytochrome P450	Other
Tc016G11R, Tc016F09R, Tc022A04R, Tc021E03R, Tc022A02R, Tc022B07R, Tc021E03F, Tc022F04F, Tc022F04R	3e-92	2e-83	Gapdh2	Glyceraldehyde 3-phosphate dehydrogenase, glycolysis	Enzyme
Tc022D12F, Tc022D12R	4e-21	2e-16	tam	DNA-directed DNA polymerase	Enzyme
Tc026F01R, Tc024H01F, Tc024H01R	4e-21	2e-16	CG6391	Diphosphoinositol polyphosphate phosphohydrolase	Enzyme

TcEST Contigs	NCBI tBLASTx (*Drosophila*)	NCBI tBLASTx (Human)	BDGP BLASTx (e<1e-03)	Putative function	General function
Tc001G04R, Tc001G04F, Tc007G08F, Tc007G08R	2e-86	3e-74	U2af50	Pre-mRNA splicing factor, RNA binding	RNA binding
Tc015C03F, Tc015C03R, Tc005B07F, Tc005B07R	2e-78	2e-63	RpII15	DNA-directed RNA polymerase II	Enzyme
Tc010A08F, Tc010A08R	8e-44	3e-15	Keren	Signal transducer	Signal transduction
Tc010H01F, Tc010H01R	4e-35	3e-08	CG14946	Oxidoreductase	Enzyme
Tc010B05F, Tc010B05R	9e-59	1e-44	CG4603	Zf-C2H2 transcription factor	Transcription factor
Tc010F06F, Tc010F06R	6e-05	8e-05	CG14549	Unknown	Unknown
Tc008F05F, Tc008F05R	1e-47	3e-39	CG12162	Unknown	Unknown
Tc013C03F, Tc013C03R	3e-39	3e-32	CG7842	S-malonyltransferase	Enzyme
Tc013H09F, Tc013H09R	2e-18	5e-16	CG10440	Unknown	Unknown
Tc009H03F, Tc009H03R	1e-25	4e-13	PitsIre	Protein kinase	Enzyme
Tc016A05F, Tc016A05R	3e-71	2e-62	CG5602	DNA ligase, DNA repair protein	Enzyme
Tc016F04F, Tc016F04R	1e-32	3e-26	CG15626	Apoptosis inhibitor	Other
Tc016B01F, Tc016B01R	2e-32	3e-27	CG7683	Cyclin, cell cycle regulator	Cell cycle regulator
Tc003D12F, Tc003D12R	2e-51	3e-51	CG7955	Transporter	Ligand binding or carrier
Tc023C02F, Tc023C02R	4e-05	7.1e-01	CG1561	Unknown	Unknown
Tc002F02F, Tc002F02R	6e-04	1.5e-02	CG10309	Zf-C2H2 transcription factor	Transcription factor
Tc027G12F, Tc027G12R	8e-12	3e-11	Rad51	DNA repair protein	Other
Tc012A09R, Tc012A09F	3e-61	1e-54	nop5	tRNA processing	Other
Tc008H02F, Tc008H02R	1e-39	3e-37	Noa36	Unknown	Unknown
Tc023E09F, Tc023E09R	1e-34	1.3e-01	CG11757	Unknown	Unknown
Tc014G06R, Tc014G06F	3e-39	7e-27	Gbeta5	Heterotrimeric G-protein GTPase, signal transduction	Signal transduction
Tc012E10F, Tc012E10R	6e-34	5e-32	CG8045	Unknown	Unknown
Tc007E06F, Tc007E06R	1e-27	2e-08	CG2079	Ligand binding or carrier	Ligand binding or carrier
Tc012C10R, Tc001D06R, Tc023E11R, Tc007G03R, Tc001D06F	1e-95	1e-88	14-3-3zeta	Diacylglycerol-activated/phospholipid dependent protein kinase C inhibitor	Other
Tc007F05R, Tc022E09R, Tc021D08R, Tc026E03R, , Tc021D08F, Tc014F10R	3e-93	6e-05	CG4115	Ligand binding or carrier	Ligand binding or carrier

TcEST Contigs	NCBI tBLASTx (*Drosophila*)	NCBI tBLASTx (Human)	BDGP BLASTx (e<1e-03)	Putative function	General function
Tc027F11R, Tc026H08F, Tc026H08R	4e-31	1e-27	CG10077	RNA binding, RNA helicase	Enzyme
Tc001C09R, Tc001C09F, Tc005D11F, Tc007A07R, Tc005D11R	1e-90	3e-63	CG5677	Signal peptidase	Enzyme
Tc016C09R, Tc024G04R, Tc022A10R, Tc015A02F, Tc015A02R	3e-46	3e-45	CG3403	Unknown	Unknown
Tc023G02F, Tc023G02R	2e-46	2e-40	CG14616	Unknown	Unknown
Tc020D10F, Tc020D10R	1e-60	1e-57	CG10932	Acetyl-CoA C-acetyltransferase	Enzyme
Tc002G06R, Tc004H04R, Tc007B03R, Tc011B04R, Tc016D05R, Tc004H04F	2e-80	2e-54	TBPH	RNA binding	RNA binding
Tc012F03R, Tc012F03F	2e-14	2e-11	CG10321	Zf-C2H2 transcription factor	Transcription factor
Tc014G04F, Tc014G04R	8e-39	9e-24	CG3304	Unknown	Unknown
Tc026B12F, Tc015B11R, Tc026B12R	4e-38	6e-30	AG01	Translation initiation factor	Translation factor
Tc025G06F, Tc016D02F, Tc025G06R, Tc016D02R	7e-07	3.1	CG15497	Unknown	Unknown
Tc026E01R, Tc021G06R, Tc014A11R, Tc014E07R, Tc008C10R, Tc024C10R, Tc027F01R, Tc009C06R, Tc020D02R, Tc020D02F	9e-14	7.8	CG1240	Unknown	Unknown
Tc002G09F, Tc002G09R	7e-31	1.1	CG8965	Unknown	Unknown
Tc025E03R, Tc014A02R, Tc014A02F	4e-04	2.8e-01	CG17181	Zf-C2H2 transcription factor	Transcription factor
Tc011E10F, Tc011E10R	3e-13	8e-04	CG2159	Diacylglycerol kinase	Enzyme
Tc013B11R, Tc022F02R, Tc022F02F	3e-40	2e-33	eas	Ethanolamine kinase, choline kinase	Enzyme
Tc012H01F, Tc012H01R	1e-31	6e-22	Atpalpha	Sodium/potassium-exchanging ATPase	Enzyme
Tc024D02F, Tc010E11R, Tc024D02R	1e-27	3e-26	CG7085	Unknown	Unknown
Tc021C08F, Tc021C08R	4e-28	4e-22	tws	Protein phosphatase	Enzyme
Tc010B11F, Tc010B11R	2e-24	9e-23	B52	Pre-mRNA splicing factor, RNA binding	RNA binding
Tc009C05F, Tc009C05R	2e-14	2e-07	Hsp67Bb	Chaperone	Other
Tc012H02F, Tc012H02R	1e-56	2e-30	Hrb27C	RNA binding, ribonucleoprotein	RNA binding
Tc008F08F, Tc008F08R	7e-69	8e-53	Gprk2	Protein serine/threonine kinase, G-protein coupled receptor kinase	Enzyme
Tc009A08F, Tc009A08R	9e-38	3e-09	mod(mdg4)	BTB/POZ transcription factor	Transcription factor
Tc014C06F, Tc014C06R	9e-27	3.3e-01	CG8756	Chitin binding	Other

TcEST Contigs	NCBI tBLASTx (*Drosophila*)	NCBI tBLASTx (Human)	BDGP BLASTx (e<1e-03)	Putative function	General function
Tc024E01R, Tc008B09R, Tc008B09F	3e-06	6.2e-02	Dsp1	Transcription co-repressor, DNA binding, HMG-box	Transcription factor
Tc026C11F, Tc026C11R	1e-04	3.6	CG11030	Unknown	Unknown
Tc008D07F, Tc008D07R	2e-25	4e-25	CG10336	Unknown	Unknown
Tc014A05F, Tc014A05R	7e-03	2.7e-01	CG5620	DNA binding	Transcription factor (possible)
Tc011F04R, Tc011F11R, Tc006F06R, Tc025D12R, Tc005E10R, Tc011E09R, Tc024E08F, Tc025D12F	1e-53	9e-49	BcDNA, LD08534	DNA repair protein	Other
Tc010H07F, Tc010H07R	2e-21	3e-18	DNApol-alpha50	DNA replication factor	Other
Tc016C05F, Tc016C05R	2e-20	1e-07	CG8964	Unknown	Unknown
Tc026H02F, Tc026H02R	4e-04	1.1e-02	kto	Transcription co-activator, receptor	Transcription factor
Tc021E12F, Tc021E12R, Tc024C06R, Tc021C09R	2e-66	7e-65	skpA	Cell cycle regulator	Cell cycle regulator
Tc002D10F, Tc002D10R	4e-50	5e-39	Vha16	Hydrogen-transporting ATP synthase	Enzyme
Tc005D02R, Tc005D02F	7e-52	6e-51	pum	RNA binding	RNA binding
Tc011H01F, Tc011H01R	1e-18	2e-13	sno	Unknown	Unknown
Tc025B02F, Tc009E07R, Tc025B02R, Tc007B05R, Tc007C03R	2e-42	1e-22	sqd	RNA binding, ribonucleoprotein	RNA binding
Tc021E04F, Tc021E04R	7e-66	1.4e-02	fax	Unknown	Unknown
Tc009A01F, Tc009A01R	4e-11	1e-05	CG9947	Unknown	Unknown
Tc013E01F, Tc013E01R	5e-38	4e-31	Nurf-38	Inorganic diphosphatase	Enzyme
Tc001C02R, Tc001C02F	6e-28	2e-26	NADH dehydrogenase-ubiquinone	Electron transfer	Enzyme
Tc001A03F, Tc001A03R	1e-69	1e-64	ND75 NADH dehydrogenase-ubiquinone	Enzyme	Enzyme
Tc023H03F, Tc023H03R, Tc013D11R, Tc005F04R	1e-02	2.1	CG11339	Actin binding	Other
Tc004B11R, Tc004F05R, Tc012A10R, Tc026G12R, Tc012A10F	3e-05	7.2e-01	CG7709	Unknown	Unknown
Tc009C10R, Tc020A05R	1.1e-01	9.8e-01	CG1150	Unknown	Unknown
Tc014A10F, Tc014A10R	1e-04	3.8e-02	EG, BACR42I17.2	Unknown	Unknown

TcEST Contigs	NCBI tBLASTx (*Drosophila*)	NCBI tBLASTx (Human)	BDGP BLASTx (e<1e-03)	Putative function	General function
Tc001B08F, Tc013C11R, Tc010E07R, Tc010E01R, Tc025B06R, Tc026B03R, Tc004C09R, Tc005G05R, Tc011B06R, Tc011H02R, Tc011A05R, Tc007D02R, Tc001B08R, Tc001A01R, Tc008H01F, Tc009G09R, Tc020E12R, Tc001A02R, Tc001G11R, Tc010B06R, Tc006D11R, Tc026E02R, Tc007F04R, Tc021A07R, Tc012D06R, Tc021E08R, Tc007A08R, Tc016G02R, Tc021H05R, Tc023F01R, Tc004F10R, Tc027B04R, Tc020B01R, Tc003E05R, Tc009B02R, Tc004B08R, Tc012D03R, Tc022H01R, Tc011E08R, Tc007B08R, Tc004G09R, Tc022A03R, Tc023D03R, Tc022F10R, Tc001A01F, Tc001A02F, Tc008H01R, Tc001G11F	3e-25	6.2	a10	Odorant binding, pheromone binding	Ligand binding or carrier
Tc015H11R, Tc023E03R, Tc002F11R, Tc027A09R, Tc004H09R	7e-45	5e-39	CoVa	Cytochrome-c oxidase subunit Va	Enzyme
Tc008A12F, Tc008A12R	3e-10	8e-10	Rep3	Unknown	Unknown
Tc020G09R	1e-12	2.8e-01	CG8444	ATP-dependant DNA helicase	Enzyme
Tc015G11F, Tc015G11R	6e-11	2e-15	CG5288	Galactokinase	Enzyme
Tc001B03F, Tc001B03R	4e-04	6e-07	CG4678	Carboxypeptidase	Enzyme
Tc001C05R, Tc001C05F	4e-04	6e-07	CG8980	Protein phosphatase inhibitor	Other
Tc001C06R, Tc001C06F	1e-39	9e-42	CG1349	RNA-binding protein regulatory subunit, ThiJ motif	RNA binding
Tc002A08R, Tc010F02R	3e-04	2e-08	CG13585	Unknown	Unknown
Tc003G03R, Tc003G03F	5e-54	4e-63	Rpb4	DNA-directed RNA polymerase II	Enzyme
Tc004A11R, Tc004A11F	3e-42	3e-68	CG1696	Unknown	Unknown
Tc005A12R, Tc005A12F	1e-17	4e-24	CG10951	Serine/threonine kinase	Enzyme
Tc005C12R, Tc005C12F	3e-76	3e-78	CG9667	Unknown	Unknown
Tc007B10R, Tc007B10F	3e-70	5e-80	CG3527	Unknown	Unknown
Tc007E07R, Tc007E07F	2e-37	1e-111	BcDNA, GH12558	Long/short chain 3-hydroxyacyl-CoA dehydrogenase	Enzyme
Tc007H09R, Tc007H09F	2e-81	9e-83	CG11876	Pyruvate dehydrogenase	Enzyme
Tc013A03R, Tc008G01R	4e-37	2e-38	CG8674	Unknown	Unknown
Tc008G10R, Tc008G10F	1e-36	1e-39	BG, DS00929.2	Cytoskeletal structural protein	Structural protein
Tc009E09R, Tc009E09F	1e-108	1e-121	Prosalpha6	Multicatalytic endopeptidase	Enzyme

TcEST Contigs	NCBI tBLASTx (*Drosophila*)	NCBI tBLASTx (Human)	BDGP BLASTx (e<1e-03)	Putative function	General function
Tc010H06R, Tc010H06F	5e-22	7e-23	CG8786	Zf-C3HC4	Transcription factor
Tc012B04R, Tc012B04F	5e-16	1e-36	CG10326	Unknown	Unknown
Tc016C02F, Tc016C02R	3e-50	2e-54	Su(var)3-9	Translation initiation factor	Translation factor
Tc021G07F, Tc021G07R	8e-68	2e-87	l(2)05070	Multicatalytic endopeptidase	Enzyme
Tc023H09R, Tc023H09F	3e-43	3e-46	CG6523	Thioredoxin	Enzyme
Tc024A10R, Tc024A10F	3e-41	3e-73	CG6835	Glutathione synthetase	Enzyme
Tc024G06R, Tc024G06F	1e-23	1e-23	CkIIbeta	Protein/casein kinase	Enzyme
Tc026B05R, Tc026B05F	5e-60	9e-82	CG10306	Unknown	Unknown
Tc026D04R, Tc026D04F	3e-40	4e-42	CG9027	Superoxide dismutase	Enzyme
Tc026H09R, Tc026H09F	1e-15	7e-16	CG11825	Unknown	Unknown
Tc027D10R, Tc027D10F	6e-67	4e-69	CG9773	Unknown	Unknown
Tc003C05F, Tc003D05R, Tc003C05R	8e-64	1e-75	CG9548	Unknown	Unknown
Tc013F10R, Tc013F10F, Tc004A07R	1e-59	6e-62	CG10682	Ubiquitin conjugating enzyme	Enzyme
Tc020D05R, Tc011E05R, Tc006G04R	4e-36	3e-36	R	RAS small monomeric GTPase	Enzyme
Tc007C05R, Tc007C05F, Tc026B01R	2e-52	1e-54	CG10590	Unknown	Unknown
Tc008E01R, Tc008C05R, Tc008E01F	4e-54	3e-60	CG7823	RHO GDP-dissociation inhibitor	Other
Tc008E06R, Tc025A10R, Tc008E06F	1e-109	1e-115	Arf79F	GTP binding, ARF small monomeric GTPase	Enzyme
Tc026F11R, Tc015A05R, Tc015A05F	1e-120	1e-132	Pros29	Multicatalytic endopeptidase	Enzyme
Tc022G01R, Tc022G01F, Tc015C06R	5e-07	5e-08	SmB	Small nuclear ribonucleoprotein	RNA binding
Tc020H12R, Tc020H12F, Tc021F04R	2e-44	7e-60	CG6766	Unknown	Unknown
Tc023H11R, Tc023E05R, Tc023H11F	1e-109	1e-120	l(2)03709	Unknown	Unknown
Tc025G07F, Tc008H11R, Tc025G07R	2e-08	1e-21	CG6272	bZIP transcription factor	Transcription factor
Tc002C07R, Tc002G07R, Tc014A04R, Tc002G07F	3e-56	1e-57	His2Av	Chromatin assembly/disassembly	Other
Tc001H06F, Tc024G03R, Tc023C03R, Tc010C01R, Tc022B02R, Tc016G09R, Tc006F02R, Tc012B08R, Tc001H06R, Tc014H02R, Tc015B02R	2e-47	6e-49	His4r	Chromatin assembly/disassembly	Other
Tc004B01R	4e-42	7e-44	CG14435	Unknown	Unknown
Tc011H08R	8e-23	6e-39	CG3004	Heterotrimaric G-protein	Signal transduction

TcEST Contigs	NCBI tBLASTx (*Drosophila*)	NCBI tBLASTx (Human)	BDGP BLASTx (e<1e-03)	Putative function	General function
Tc015A10R	3e-20	2e-24	CG5853	ATP-binding cassette (ABC) transporter	Ligand binding or carrier
Tc005B12R	2e-41	5e-43	D	HMG-box DNA binding protein	Transcription factor
Tc004E05R, Tc009A03R, Tc014C03R, Tc001E10R, Tc001C04R, Tc001E10F, Tc001C04F	1e-111	1e-112	betaTub56D	Structural protein of cytoskeleton	Structural protein
Tc027H02F, Tc014C11F, Tc027H02R, Tc013D04R, Tc014C11R	4e-50	4e-61	Arc-p34	Actin binding	Other
Tc026G06F, Tc026G06R	1e-27	4e-37	CG3957	Signal transduction	Signal transduction
Tc020G11F, Tc020G11R	4e-32	1e-59	CG1532	Glyoxalase	Enzyme
Tc025D09F, Tc025D09R	2e-09	5e-11	CG6345	Unknown	Unknown
Tc016G12F, Tc016G12R	3e-92	3e-92	CG11700	Contain ubiquitin domain	Unknown
Tc008G02F, Tc008G02R	2e-55	3e-58	Rab5	GTP binding, RAB small monomeric GTPase	Enzyme
Tc012G11F, Tc023B12F, Tc012G11R, Tc023B12R	1e-55	8e-57	CG5525	Chaperonin ATPase	Enzyme
Tc023A02F, Tc025B04F, Tc025B04R, Tc023A02R	4e-09	7e-12	Cg11228	Receptor signaling protein serine/threonine kinase	Enzyme
Tc006E10F, Tc006E10R	4e-20	3e-31	CG6364	Uridine kinase	Enzyme
Tc024G01R, Tc020D04R, Tc024D08R, Tc006D12R, Tc004B03R, Tc005B04R, Tc026E06R, Tc008F04R, Tc004B03F	1e-111	1e-119	sesB	Carrier, ATP/ADP antiporter	Ligand binding or carrier
Tc009C11R, Tc014E03R, Tc014E03F	2e-20	9e-25	CG7718	CDP-diacylglycerol-glycerol-3-phosphate 3-phosphatidyltransferase	Enzyme
Tc005C06F, Tc005C06R	2e-09	4e-10	CG2982	Unknown	Unknown
Tc023A12F, Tc023A12R, Tc002H03R	4e-45	4e-48	CG9149	Acetyl-CoA C-acetyltransferase	Enzyme
Tc014B02F, Tc014B02R	8e-20	3e-48	CG6479	Unknown	Unknown
Tc003D03F, Tc003D03R	6e-07	1e-08	BcDNA, LD21129	Unknown	Unknown
Tc013G08F, Tc013G08R	9e-02	5e-06	CycA	Cell cycle regulator	Cell cycle regulator
Tc027A12F, Tc027A12R	2e-42	2e-48	wds	Signal transduction	Signal transduction
Tc022G03F, Tc022G03R	2e-42	2e-48	CG6750	Unknown	Unknown
Tc015D01F, Tc015D01R	5e-29	6e-33	CG7041	Chromatin binding	Other
Tc008A11R, Tc008A11F	2e-42	2e-53	Rnrs	Ribonucleoside-diphosphate reductase	Enzyme
Tc016C01F, Tc016C01R	6e-33	1e-42	CG10627	Phosphoacetylglucosamine mutase	Enzyme

TcEST Contigs	NCBI tBLASTx (*Drosophila*)	NCBI tBLASTx (Human)	BDGP BLASTx (e<1e-03)	Putative function	General function
Tc007H02F, Tc002G10F, Tc007H02R, Tc002G10R	5e-21	2e-30	mts	Protein phosphatase	Enzyme
Tc010H11F, Tc006B02F, Tc006B02R, Tc010H11R	9e-06	6e-08	l(2)08492	Unknown	Unknown
Tc016C03F, Tc016C03R	5e-34	1e-35	CG1513	Oxysterol binding	Other
Tc024C05F, Tc024C05R	8e-28	4e-31	CG1696	Unknown	Unknown
Tc004H02F, Tc004H02R	9e-23	7e-23	Tsc1	Cell cycle regulator	Cell cycle regulator
Tc008G08F, Tc008G08R	5e-17	5e-18	Akt1	Protein serine/threonine kinase/protein kinase	Enzyme
Tc020B11R, Tc020B11F	3e-14	1e-21	E2f	E2f/TDP transcription factor	Transcription factor
Tc011B03F, Tc011B03R	4e-14	8e-15	AP-1gamma	Transporter, vesicle coating	Ligand binding or carrier
Tc001E06F, Tc026H03R, Tc001E06R	4e-56	1e-67	Uev1A	Ubiquitin conjugating enzyme	Enzyme
Tc027D06R, Tc008H03R, Tc025B01R, Tc025B01F	9e-14	1e-16	e(y)2	Transcription factor	Transcription factor
Tc001E11F, Tc001E11R	4e-21	2e-22	CG3861	Citrate synthase	Enzyme
Tc008D03F, Tc008D03R	1e-30	6e-31	CG3876	Unknown	Unknown
Tc008G03R	1.2	3.4e-02	chd1	Helicase	Enzyme
Tc016B02F, Tc016B02R	7e-06	1e-08	CG7816	Unknown	Unknown
Tc027B08F, Tc027B08R	1e-09	5e-11	CG13191	Unknown	Unknown
Tc027H06R, Tc006A08R	1e-03	3e-04	mof	Histone acetyltransferase	Enzyme
Tc009A04F, Tc009A04R	1e-10	3e-24	CG16745	Contains Zf-traf domain	Signal transduction
Tc006B11F, Tc006B11R	8e-06	3e-09	CG6313	Unknown	Unknown
Tc005H02F, Tc005H02R	6e-05	3e-05	CG7331	Unknown	Unknown
Tc003F11F, Tc003F11R	2e-03	1e-05 (*Mus musculus*)	CG5834	Unknown	Unknown
Tc006G02F, Tc010D02R, Tc013B03R, Tc006G02R	1e-103	1e-114	CG11935	Hydrolase	Enzyme
Tc015G04R, Tc008H05R, Tc008H05F	1e-36	6e-49	CG12000	Multicatalytic endopeptidase	Enzyme
Tc009D09R, Tc020D12R, Tc009D09F	1e-93	3e-94	Cam	Calcium binding	Other
Tc015B06R, Tc015B06F, Tc006B03R, Tc010D07R	1e-19	1e-20	CG11822	Nicotinic acetylcholine-activated cation selective channel	Other
Tc001A05F	4e-48	1e-48	AP-47	Vesicle coating	Other
Tc001C01F, Tc001C01R	2e-28	3e-45	CG10576	Methionyl aminopeptidase	Enzyme

TcEST Contigs	NCBI tBLASTx (*Drosophila*)	NCBI tBLASTx (Human)	BDGP BLASTx (e<1e-03)	Putative function	General function
Tc013A04F, Tc013A04R	7e-08	2e-10	CG17598	Protein serine/threonine phosphatase	Enzyme
Tc023E02F, Tc023E02R	4e-04	4e-06	CG8149	Unknown	Unknown
Tc015F11F, Tc015F11R	5e-34	3e-39	CG7662	Contain PDZ domain for signaling molecule	Signal transduction
Tc004A08R, Tc004A08F	1e-32	1e-44	XM_007381.3 [RTN-1] (*Homo sapiens*)	Neuroendocrine secretion or cell trafficking	Other
Tc008B02R, Tc008B02F	7e-05	1e-10	Unactive progesterone receptor (*Homo sapiens*)	Unknown	Unknown
Tc010F10R, Tc011A03R	1e-06	2e-34	ORF (*Homo sapiens*)	Membrane protein, Unknown	Unknown
Tc012E06F, Tc012E06R	7e-07	2e-13	U1 snRNA from tBLASTx	RNA splicing	RNA binding
Tc021H01R, Tc021H01F	4e-38	4e-55	Micorsomal signal peptidase 25 KDA subunit (SPC25) (*Homo sapiens*)	Signal peptidase	Enzyme
Tc012D08R, Tc014D04R, Tc020D08R, Tc011G01R, Tc027D09R, Tc009D01R, Tc004C10R, Tc007H12R, Tc013E06R, Tc021C01R, Tc005G10R, Tc002A11R, Tc021F02R, Tc023D10R, Tc009E10R	1e-02	4e-33	c6.1A (*Homo sapiens*)	Unknown	Unknown
Tc022D08F, Tc022D08R	2e-03	2e-28	Epsilon-COP (*Homo sapiens*)	Vesicle coating	Other
Tc006C07F, Tc006C07R	5e-10	2e-13	Putative G-protein coupled receptor (*Homo sapiens*)	Unknown	Unknown
Tc026A09F, Tc026A09R	5e-07	2e-15	BEST, LD29743	3-hydroxyisobutyrate dehydrogenase	Enzyme
Tc014B06F, Tc014B06R	4.10E+00	8e-04 (*Mus musculus*)	BAB27018 (*Mus musculus*)	Unknown	Unknown
Tc001E03F, Tc001E03R	2e-03	1e-06 (*Schizosacchromyces pombe*)	cki2 (Schizosacchromyces pombe)	Casein kinase	Enzyme
Tc026B10F, Tc026B10R	1e-08	3e-09 (*Xenopus Leavis*)	U1 snRNA (*Xenopus leavis*)	RNA splicing	RNA binding
Tc009B05R, Tc008H06R	1.6	3e-21	KDEL (*Homo sapiens*)	Endoplasmic reticulum protein, Unknown	Unknown
Tc011C02R, Tc011A06R, Tc004D08R, Tc002F07R, Tc004D08F	1.6	9e-66 (*Derobrachus geminatus*)	Apolipophorin-III (*Derobrachus geminatus*)	Apolipoprotein, ligand binding or carier	Ligand binding or carrier
Tc020B04F, Tc020B04R	5.5	1e-03	Cyclin A1 (*Homo sapiens*)	Cell cycle protein	Cell cycle regulator
Tc007F06F, Tc007F06R	3.1	9e-38	TOLLIP protein (*Homo sapiens*)	Signal transduction	Signal transduction
Tc024C03F, Tc024C03R	2.5	7e-04	Ancient ubiquitous 46 kDa protein AUP1 (*Homo sapiens*)	Ubiquitous, unknown	Unknown

TcEST Contigs	NCBI tBLASTx (*Drosophila*)	NCBI tBLASTx (Human)	BDGP BLASTx (e<1e-03)	Putative function	General function
Tc014B11F, Tc014B11R	3.2	5e-14	Hypothetical protein FLJ10407 (*Homo sapiens*)	Unknown	Unknown
Tc007B09F, Tc007B09R	2.2	4e-05	Hypothetical protein SBBI48 (*Homo sapiens*)	Unknown	Unknown
Tc003G09F, Tc003G09R	5.6e-02	5e-05	Hypothetical protein KIAA0627 (*Homo sapiens*)	Cytoplasmic linker protein	Other
Tc002C11F, Tc002C11R	9.2e-01	8e-03	Estrogen receptor binding site associated antigen 9 (*Homo sapiens*)	Cytoplasmic molecule, involved in apoptosis	Other
Tc007A02F, Tc011D09F, Tc007A02R, Tc011D09R	3.5	3e-62	KIAA1576 (*Homo sapiens*)	Unknown	Unknown
Tc027B02R	1e-09	2e-13	CYMBP (*Homo sapiens*)	c-myc binding protein	Transcription factor
Tc023B03R	2e-03	2e-14	CG5316	Zf-C2H2 transcription factor	Transcription factor
Tc012H03R	n/a	n/a	Zen (*Tribolium castaneum*)	Homeobox transcription factor	Transcription factor
Tc001D07F, Tc001D07R	n/a	n/a	engrailed (*Tribolium castaneum*)	Homeobox transcription factor	Transcription factor
Tc015E05F, Tc015E05R	n/a	n/a	Cytochrome P450 monooxygenase (*Tribolium castaneum*)	Electron transfer	Enzyme
Tc012G05F, Tc012G05R	n/a	n/a	Unknown (*Tribolium castaneum*)	Unknown	Unknown
Tc001F02R, Tc001F02F	n/a	n/a	Unknown (*Tribolium castaneum*)	Unknown	Unknown
Tc015B05R, Tc015D04R, Tc012E08R, Tc003H03R, Tc025A02R, Tc015G01R, Tc023D05R, Tc002E01R, Tc021D03R, Tc022F12R, Tc013H01R, Tc002A04R, Tc002D02R, Tc016A09R, Tc012F07R, Tc025B11R, Tc022H08R, Tc007B01R, Tc016G05R, Tc022F09R, Tc014E12R, Tc012A03R, Tc012A06R, Tc002G05R, Tc015H01R, Tc027E12R, Tc021A03R, Tc026B04R, Tc022G07R, Tc011B09R, Tc008G04R, Tc005B10R, Tc027G03R, Tc027E08R, Tc024H06R, Tc015D08R, Tc022C03R	n/a	n/a	Rp S2	Ribosomal protein	Ribosomal protein
Tc015E08R, Tc008F09R, Tc014G05R, Tc026G02R, Tc015C09R, Tc009G06R, Tc008D12R, Tc026C10R, Tc007D09R, Tc004B02R, Tc026B08R, Tc016G04R	n/a	n/a	Rp S2	Ribosomal protein	Ribosomal protein
Tc010B03R, Tc015D05R, Tc016F07R, Tc012A12R, Tc012F08R, Tc027C04R, Tc003B05R	n/a	n/a	Rp S3	Ribosomal protein	Ribosomal protein

TcEST Contigs	NCBI tBLASTx (*Drosophila*)	NCBI tBLASTx (Human)	BDGP BLASTx (e<1e-03)	Putative function	General function
Tc013G12R, Tc003H12R, Tc027E10R, Tc001A11R	n/a	n/a	Rp S4	Ribosomal protein	Ribosomal protein
Tc003F03R, Tc027G02R, Tc021A05R, Tc004D12R, Tc010F03R, Tc013C10R, Tc001C12R, Tc026F02F, Tc001C12F, Tc026F02R	n/a	n/a	Rp S5	Ribosomal protein	Ribosomal protein
Tc012C02R, Tc001H05R, Tc001H05F, Tc010A11R, Tc010F04R, Tc011B05R, Tc009C01R, Tc006B12R	n/a	n/a	Rp S5	Ribosomal protein	Ribosomal protein
Tc026G05R, Tc025A05R, Tc013F12R, Tc007E01R, Tc005C04R, Tc006F03R, Tc012F05R	n/a	n/a	Rp S6	Ribosomal protein	Ribosomal protein
Tc001B06F, Tc013D02R, Tc006B06R, Tc014A06R, Tc004D04R, Tc005E09R, Tc001B06R	n/a	n/a	Rp S7	Ribosomal protein	Ribosomal protein
Tc009E02R, Tc009G12R, Tc003D04R, Tc011H10R, Tc007E11R, Tc002C08R	n/a	n/a	Rp S8	Ribosomal protein	Ribosomal protein
Tc009C02R, Tc013H03R, Tc023F03R, Tc010D03R, Tc010C02R, Tc014H01R, Tc014H11R, Tc020A07R, Tc008C03R, Tc020E07R	n/a	n/a	Rp S9	Ribosomal protein	Ribosomal protein
Tc004C02R, Tc010D06R, Tc025G12R	n/a	n/a	Rp S11	Ribosomal protein	Ribosomal protein
Tc001F08R, Tc015D06R, Tc022H09R, Tc023H04R, Tc012B09R, Tc001F08F	n/a	n/a	Rp S11	Ribosomal protein	Ribosomal protein
Tc001D01F, Tc024D10R, Tc021C04R, Tc015A09R, Tc001D01R	n/a	n/a	Rp S12	Ribosomal protein	Ribosomal protein
Tc003G02R, Tc011C11R, Tc004C06R	n/a	n/a	Rp S13	Ribosomal protein	Ribosomal protein
Tc010H09R, Tc012B05R, Tc027G11R, Tc015H02R, Tc008C07R, Tc012B03R, Tc009G11R, Tc007D12R, Tc013G10R, Tc020G02R, Tc010F08R, Tc022H11R, Tc020A09R, Tc008C01R, Tc015A03R, Tc012C12R, Tc009G11F, Tc007E05F, Tc003A12F, Tc010G10R, Tc027F06R, Tc007E05R, Tc003A12R, Tc007D03R	n/a	n/a	Rp S14	Ribosomal protein	Ribosomal protein
Tc025C06R, Tc004F11R	n/a	n/a	Rp S15	Ribosomal protein	Ribosomal protein
Tc013D12R, Tc002G11R, Tc009C04R, Tc023E08R, Tc010A05R, Tc023C01R, Tc020B08R, Tc004C04R	n/a	n/a	Rp S16	Ribosomal protein	Ribosomal protein
Tc007A04R, Tc026E12R, Tc025E01R, Tc024B03R, Tc005A08R	n/a	n/a	Rp S17	Ribosomal protein	Ribosomal protein
Tc016D06R, Tc025A03R, Tc002A03R, Tc022F06F, Tc016B11R, Tc003H09R, Tc021G01R, Tc022F06R	n/a	n/a	Rp S19	Ribosomal protein	Ribosomal protein
Tc023H12R, Tc021F03R	n/a	n/a	Rp S20	Ribosomal protein	Ribosomal protein

TcEST Contigs	NCBI tBLASTx (*Drosophila*)	NCBI tBLASTx (Human)	BDGP BLASTx (e<1e-03)	Putative function	General function
Tc027G08R, Tc005E08R	n/a	n/a	Rp S21	Ribosomal protein	Ribosomal protein
Tc026E09R, Tc010G08R, Tc009G02R	n/a	n/a	Rp S22	Ribosomal protein	Ribosomal protein
Tc012E01R	n/a	n/a	Rp S23	Ribosomal protein	Ribosomal protein
Tc013F06R, Tc004F12R, Tc026E05R, Tc023F02R	n/a	n/a	Rp S24	Ribosomal protein	Ribosomal protein
Tc001D10F, Tc001D10R, Tc010G12R, Tc009C03R	n/a	n/a	Rp S25	Ribosomal protein	Ribosomal protein
Tc012E12R, Tc007F03R, Tc024F03R, Tc013B04R, Tc009C08R, Tc012A08R, Tc023A05R, Tc003A06R, Tc003B06R, Tc025G05R, Tc015A11R	n/a	n/a	Rp S25	Ribosomal protein	Ribosomal protein
Tc015B01R, Tc011B12R, Tc021G10R, Tc010F01R	n/a	n/a	Rp S26	Ribosomal protein	Ribosomal protein
Tc024F01R, Tc003F08R, Tc004C05R	n/a	n/a	Rp S26	Ribosomal protein	Ribosomal protein
Tc026G11R, Tc009F12R, Tc016E11R	n/a	n/a	Rp S27	Ribosomal protein	Ribosomal protein
Tc014D09R, Tc027G10R, Tc012F09R, Tc004H07R	n/a	n/a	Rp S27A	Ribosomal protein	Ribosomal protein
Tc010G07R, Tc001D05F	n/a	n/a	Rp S28	Ribosomal protein	Ribosomal protein
Tc016F03R	n/a	n/a	Rp S29	Ribosomal protein	Ribosomal protein
Tc006E12R, Tc009A07R, Tc008A06R	n/a	n/a	Rp S29	Ribosomal protein	Ribosomal protein
Tc002F05R, Tc004A03R, Tc021H09R, Tc009A02R	n/a	n/a	Rp L3	Ribosomal protein	Ribosomal protein
Tc005F01R	n/a	n/a	Rp L4	Ribosomal protein	Ribosomal protein
Tc024D03R, Tc024D03F	n/a	n/a	Rp L6e	Ribosomal protein	Ribosomal protein
Tc022F03R, Tc003G12R, Tc015G07R, Tc014H12R, Tc020C06R, Tc012G12R	n/a	n/a	Rp L7	Ribosomal protein	Ribosomal protein
Tc027D03R	n/a	n/a	Rp L7A	Ribosomal protein	Ribosomal protein
Tc027B12R, Tc005B02R, Tc021F08R, Tc001H02R, Tc016A10R, Tc001H02F	n/a	n/a	Rp L8	Ribosomal protein	Ribosomal protein

TcEST Contigs	NCBI tBLASTx (*Drosophila*)	NCBI tBLASTx (Human)	BDGP BLASTx (e<1e-03)	Putative function	General function
Tc023G09R, Tc020C11R, Tc009C07R, Tc026E10R, Tc022D11R, Tc026F04R, Tc016F01R, Tc011D11R, Tc014G07R, Tc009B07R, Tc027D08R, Tc006B09R, Tc010B04R	n/a	n/a	Rp L9	Ribosomal protein	Ribosomal protein
Tc023B10R, Tc027E02R, Tc016E06R, Tc016E02R, Tc020E08R, Tc006H07R, Tc027H01R, Tc020A04R, Tc012D02R, Tc013H04R	n/a	n/a	Rp L10	Ribosomal protein	Ribosomal protein
Tc002A07R, Tc024G09R	n/a	n/a	Rp L10	Ribosomal protein	Ribosomal protein
Tc011H07R, Tc024H10R, Tc016H04R, Tc005F08R	n/a	n/a	Rp L10A	Ribosomal protein	Ribosomal protein
Tc001E07R, Tc001E07F	n/a	n/a	Rp L11	Ribosomal protein	Ribosomal protein
Tc025B12R, Tc025B12F	n/a	n/a	Rp L12	Ribosomal protein	Ribosomal protein
Tc027C11R, Tc013B10R, Tc001G10R, Tc013A07R, Tc014D02R, Tc001G10F	n/a	n/a	Rp L12	Ribosomal protein	Ribosomal protein
Tc010F11R, Tc003C06R, Tc025A09R	n/a	n/a	Rp L13	Ribosomal protein	Ribosomal protein
Tc020A02R, Tc020A02F	n/a	n/a	Rp L13	Ribosomal protein	Ribosomal protein
Tc027G05R, Tc025F03R, Tc013C08R, Tc009H01R, Tc010E12R	n/a	n/a	Rp L13A	Ribosomal protein	Ribosomal protein
Tc009D05R	n/a	n/a	Rp L14	Ribosomal protein	Ribosomal protein
Tc016D11R, Tc021C12R, Tc023G10R, Tc027B07R, Tc015F09R, Tc016D03R, Tc008A07R, Tc008B12R, Tc020A06R, Tc024H02R, Tc027F02R, Tc014F04R, Tc012B11R, Tc014F03R	n/a	n/a	Rp L14	Ribosomal protein	Ribosomal protein
Tc003A09R, Tc011E04R, Tc027A06R, Tc027E06R, Tc004E03R, Tc016F12R	n/a	n/a	Rp L15	Ribosomal protein	Ribosomal protein
Tc004E11R, Tc014C05R, Tc026A02R, Tc008E07R, Tc025D07R, Tc008E04R, Tc009F11R	n/a	n/a	Rp L17	Ribosomal protein	Ribosomal protein
Tc021B03F, Tc021B03R	n/a	n/a	Rp L17	Ribosomal protein	Ribosomal protein
Tc014E05R, Tc004A12R, Tc015E02R, Tc020A01R	n/a	n/a	Rp L18A	Ribosomal protein	Ribosomal protein
Tc022E07R, Tc022C11R, Tc024G11R, Tc012G08R, Tc012F11R, Tc003D06R	n/a	n/a	Rp L19	Ribosomal protein	Ribosomal protein

TcEST Contigs	NCBI tBLASTx (*Drosophila*)	NCBI tBLASTx (Human)	BDGP BLASTx (e<1e-03)	Putative function	General function
Tc001B10F, Tc012A07R, Tc023B05R, Tc013F11R, Tc012G06R, Tc008B10R, Tc027A10R, Tc011F06R, Tc012C11R, Tc001B10R, Tc002E07R, Tc012C11F	n/a	n/a	Rp L21	Ribosomal protein	Ribosomal protein
Tc021F12R, Tc007C09R, Tc022B11R, Tc002B08R	n/a	n/a	Rp L22	Ribosomal protein	Ribosomal protein
Tc015F03R	n/a	n/a	Rp L23	Ribosomal protein	Ribosomal protein
Tc011A09R, Tc016E10R, Tc015E04R, Tc010A04R, Tc008E12R	n/a	n/a	Rp L23A	Ribosomal protein	Ribosomal protein
Tc024C04R, Tc024C04F	n/a	n/a	Rp L24	Ribosomal protein	Ribosomal protein
Tc004H11R, Tc024A04R, Tc021G08R	n/a	n/a	Rp L24	Ribosomal protein	Ribosomal protein
Tc001D02F, Tc016D08R, Tc015H04R, Tc001D02R, Tc005D09R	n/a	n/a	Rp L27	Ribosomal protein	Ribosomal protein
Tc020G07R, Tc011F10R, Tc011H09R, Tc027D04R, Tc001C03F, Tc024C07R, Tc001C03R, Tc020G07F, Tc008B03R	n/a	n/a	Rp L27A	Ribosomal protein	Ribosomal protein
Tc004F03R	n/a	n/a	Rp L27A	Ribosomal protein	Ribosomal protein
Tc020F04R, Tc023B11R, Tc003F02R	n/a	n/a	Rp L28	Ribosomal protein	Ribosomal protein
Tc015D12R, Tc027B01R, Tc022E12R	n/a	n/a	Rp L30	Ribosomal protein	Ribosomal protein
Tc013B01R, Tc009A12R, Tc013D01R, Tc016H09R, Tc020C01R, Tc023E01R	n/a	n/a	Rp L32	Ribosomal protein	Ribosomal protein
Tc023G08R, Tc022D05R, Tc008H09R, Tc023A11R	n/a	n/a	Rp L34	Ribosomal protein	Ribosomal protein
Tc012D12R, Tc016H08R, Tc023F04R, Tc024F08R	n/a	n/a	Rp L35	Ribosomal protein	Ribosomal protein
Tc021B02R, Tc003H01R	n/a	n/a	Rp L35A	Ribosomal protein	Ribosomal protein
Tc001G01F, Tc022H10R, Tc011C12R, Tc001G01R	n/a	n/a	Rp L36	Ribosomal protein	Ribosomal protein
Tc001B04F, Tc001B04R	n/a	n/a	Rp L37A	Ribosomal protein	Ribosomal protein
Tc005F05R	n/a	n/a	Rp L38	Ribosomal protein	Ribosomal protein
Tc025H04R, Tc005E04R, Tc012E02R	n/a	n/a	Rp L38	Ribosomal protein	Ribosomal protein

TcEST Contigs	NCBI tBLASTx (*Drosophila*)	NCBI tBLASTx (Human)	BDGP BLASTx (e<1e-03)	Putative function	General function
Tc027H11R, Tc015E07R	n/a	n/a	Rp L40	Ribosomal protein	Ribosomal protein
Tc013H10R, Tc023A03R, Tc008B04R, Tc012C05R, Tc014C08R, Tc016H07R, Tc010G03R, Tc022E02R, Tc023G06R	n/a	n/a	Rp L44	Ribosomal protein	Ribosomal protein
Tc014H05R, Tc023C05R, Tc021C03R	n/a	n/a	Rp P0	Ribosomal protein	Ribosomal protein
Tc016A04R, Tc009F02R, Tc010D08R, Tc008F07R, Tc010A09R, Tc015F06R, Tc011F08R	n/a	n/a	Rp P2	Ribosomal protein	Ribosomal protein
Tc015D11R, Tc026H10R, Tc003A01R, Tc022D02R, Tc004C08R	n/a	n/a	Rp P2	Ribosomal protein	Ribosomal protein
Tc001F07R, Tc024C11R, Tc014F12R, Tc024E07R, Tc009F04R, Tc009A09R, Tc011F07R, Tc021E06R, Tc022F01R, Tc001F07F	n/a	n/a	Rp SA	Ribosomal protein	Ribosomal protein
Tc013D10R, Tc007G09R, Tc004E09R, Tc016A08R, Tc004H10R, Tc007C12R, Tc021D09R, Tc013H02R, Tc006F05R, Tc005G03R, Tc013F04R, Tc003D01R, Tc004G10R, Tc004E06R, Tc015H07R, Tc005D03R, Tc016B06R, Tc001E04R, Tc004D10R, Tc027B11R, Tc010D12R, Tc011G12R, Tc007H08R, Tc010E08R, Tc001E04F, Tc001H12R, Tc026F08R, Tc001H12F, Tc021H07R	n/a	n/a	Cytochrome oxidase subunit I	Mitochondrial sequence	Mitochondrial sequence
Tc007H01R, Tc007D04R, Tc011B01R, Tc021D11R, Tc009F06R, Tc025E07R, Tc014A08R, Tc020D03R, Tc026D07R, Tc005A11R, Tc002H12R, Tc008F10R, Tc006D04R, Tc008G07R, Tc015D02R, Tc007G12R, Tc016A11R, Tc016A12R, Tc027G07R, Tc023D07R, Tc026H06R, Tc024A12R, Tc024E02R, Tc011D07R, Tc005C02R, Tc004F01R, Tc006F11R, Tc009G03R, Tc026C08R, Tc005H11R, Tc012D10R, Tc016F08R, Tc004B05R, Tc024D07R, Tc003A07R, Tc013D09R, Tc013A06R, Tc005G02R, Tc003E06R, Tc004G04R, Tc022C04R, Tc001E01F, Tc010D01R, Tc007A06R, Tc025F04R, Tc015G12R, Tc001E01R, Tc016B07R, Tc001B05F, Tc003F12F, Tc016F08F, Tc021F11R, Tc001B05R, Tc023G04R, Tc010A06R, Tc003F12R, Tc016E05R, Tc026G01R, Tc002D06R, Tc027D02R	n/a	n/a	Cytochrome oxidase subunit II	Mitochondrial sequence	Mitochondrial sequence
Tc007G02R, Tc013H11R, Tc006D07R, Tc003H08R, Tc009F07R, Tc013C12R, Tc013C02R, Tc014G03R, Tc008D04R, Tc027H08R, Tc003A02R, Tc009B09R, Tc010G04R, Tc001G12R, Tc012H04R, Tc001G12F	n/a	n/a	Cytochrome oxidase subunit III	Mitochondrial sequence	Mitochondrial sequence

TcEST Contigs	NCBI tBLASTx (*Drosophila*)	NCBI tBLASTx (Human)	BDGP BLASTx (e<1e-03)	Putative function	General function
Tc024B08R, Tc020B05R, Tc021C05R, Tc014H08R, Tc009E06R, Tc015H08R, Tc014B05R, Tc022B06R, Tc020C03R, Tc021B05R, Tc013A08R, Tc010B10R, Tc008D09R, Tc022E01R, Tc013F07R, Tc020H11R, Tc025C09R, Tc016D04R, Tc026F09R, Tc007F07R, Tc025H07R, Tc005C08R, Tc016G07R, Tc026H04R, Tc007A03R, Tc021H11R, Tc021B07R, Tc011H04R, Tc024C08R, Tc011D10R, Tc003G06R, Tc001G08R, Tc016H05R, Tc015C11R, Tc004E07R, Tc026G04R, Tc001G08F, Tc014A09F, Tc021C05F, Tc005C08F, Tc007B06R, Tc026E04R, Tc005A06R, Tc011H06R, Tc022A12R, Tc021B08R, Tc013B02R, Tc022B03R, Tc013G02R, Tc021B11R, Tc010B10F, Tc014A09R	n/a	n/a	Cytochrome b	Mitochondrial sequence	Mitochondrial sequence
Tc010E10R, Tc026A12F	n/a	n/a	NADH dehydrogenase subunit 5	Mitochondrial sequence	Mitochondrial sequence
Tc021F10F, Tc021F10R	n/a	n/a	ATP synthase	Mitochondrial sequence	Mitochondrial sequence
Tc007D05R, Tc025G11R, Tc022B10R, Tc004E10R, Tc012H06R, Tc014B09R, Tc026C09R, Tc014G12R, Tc023D04R, Tc007G10R, Tc002E05R, Tc005F03R, Tc002H09R, Tc008G12R, Tc003C11R, Tc025A07R, Tc025E12R, Tc012G07R, Tc014B12R, Tc025E05R, Tc005C10R, Tc015A12R, Tc022H06R, Tc024E04R, Tc027A11R, Tc025H08R, Tc024C02R, Tc006C08R, Tc026A11R, Tc020B10R, Tc021H12R, Tc009G08R, Tc009B10R, Tc005A03R, Tc024C09R, Tc015E01R, Tc026D01R, Tc008A10R	n/a	n/a	16S rRNA	Mitochondrial sequence	Mitochondrial sequence

TcEST Contigs	NCBI tBLASTx (*Drosophila*)	NCBI tBLASTx (Human)	BDGP BLASTx (e<1e-03)	Putative function	General function
Tc009E08R, Tc023G11R, Tc022B09R, Tc023A07R, Tc026A07R, Tc024B11R, Tc025F06R, Tc004D05R, Tc008F06R, Tc025D10R, Tc013C06R, Tc014G09R, Tc015E06R, Tc015F01R, Tc023F11R, Tc026D08R, Tc023C10R, Tc006D06R, Tc013E07R, Tc013C07R, Tc016B08R, Tc014E10R, Tc002F10R, Tc002B05R, Tc025D08R, Tc020C10R, Tc010A02R, Tc013D07R, Tc013E02R, Tc002F01R, Tc024H11R, Tc008A01R, Tc011C08R, Tc024A06R, Tc015H03R, Tc001G02R, Tc016D01R, Tc007D06R, Tc006G12R, Tc010H03R, Tc004D03R, Tc022H02R, Tc007B02R, Tc008E05R, Tc009F08R, Tc005E05R, Tc025H05R, Tc011E06R, Tc004D07R, Tc004C07R, Tc024E03R, Tc020B12R, Tc011G02R, Tc023G03R, Tc020C04R, Tc015F01F, Tc012C04R, Tc004H03R, Tc021C07R, Tc004A05R, Tc006G03R, Tc004H05R, Tc009E11R, Tc012A04R, Tc005D06R, Tc006B08R, Tc014A03R, Tc003G04R, Tc004C03R, Tc009D02R, Tc004D11R, Tc007C10R, Tc021F07R, Tc020D11R, Tc021A11R, Tc006H08R, Tc020F08R, Tc022B01R, Tc008H04R, Tc003C02R, Tc015G10R, Tc009B11R, Tc004A10R, Tc013H06R, Tc014C01R, Tc013A10R, Tc001G06F, Tc004B10R, Tc004F08R, Tc025F11R, Tc004H12R, Tc010C12R, Tc008A08R, Tc006D08R, Tc009D11R, Tc011C01R, Tc001G03R, Tc020G01R, Tc006C05R, Tc009D04R, Tc004G08R, Tc022G02R, Tc006H01R, Tc010H12R, Tc004A09R, Tc003B07R, Tc006A10R, Tc013F01R, Tc006D03R, Tc025B05R, Tc008B08R, Tc012H08R, Tc022C02R, Tc025H11R, Tc014F02R, Tc012G01R, Tc013C04R, Tc002F09R, Tc014F06R, Tc024D09R, Tc003A05R, Tc022A09R, Tc008B05R, Tc024E09R, Tc013H08R, Tc001F09R, Tc014H04R, Tc003B04R, Tc006C12R, Tc001E05R, Tc015H10R, Tc025H03R, Tc010G06R, Tc001E08R, Tc020B02R, Tc023E04R, Tc007C07R, Tc015D03R, Tc014E11F, Tc027F12R, Tc013E08R, Tc009A10R, Tc001G03F, Tc020E09R, Tc004E04R, Tc025G10R, Tc026D06R, Tc010D04R, Tc012A02R, Tc015B10R, Tc001E08F, Tc027F10R, Tc015F04R, Tc020F11R, Tc020F12R, Tc001E09R, Tc022H04R, Tc011F02R, Tc001G06R, Tc006G04F, Tc009G07R, Tc010A03R, Tc015H09F, Tc015F05R, Tc003B01R, Tc011F12R, Tc001F09F, Tc024G07R, Tc013C05R, Tc020D09R, Tc007G05R, Tc020C09R, Tc025H06R, Tc010E06R, Tc026D11R, Tc001E05F, Tc027F08R, Tc003H02R, Tc001E09F, Tc025E04R, Tc021B10R, Tc009H04R, Tc026A04R, Tc013C09R, Tc021B01R, Tc006H04R, Tc004C12R, Tc016B05R, Tc006B07R, Tc024D05R, Tc023G11F, Tc015H09R, Tc014E11R, Tc003G04F	n/a	n/a	16S rRNA	Mitochondrial sequence	Mitochondrial sequence

TcEST Contigs	NCBI tBLASTx (*Drosophila*)	NCBI tBLASTx (Human)	BDGP BLASTx (e<1e-03)	Putative function	General function
Tc020H04F, Tc020H04R, Tc020D06R, Tc025E06R	n/a	n/a	16S rRNA	Mitochondrial sequence	Mitochondrial sequence
Tc026B06R, Tc014E09R, Tc002F06R, Tc005G04R, Tc025D11R, Tc013H07R, Tc024H07R, Tc011A02R, Tc026A03R, Tc013E05R, Tc010C11R, Tc013B09R, Tc002D09R, Tc010F12R, Tc025C11R, Tc010H08R, Tc007H04R, Tc010A10R, Tc025B09R, Tc001F12R, Tc001A12R, Tc001G09R, Tc001F12F, Tc001D08R, Tc010E09R, Tc005C01R, Tc001D08F, Tc001A12F, Tc009G01R	n/a	n/a	Mitochondrial sequence	Mitochondrial sequence	Mitochondrial sequence
Tc024E06R, Tc025F05R, Tc012G04R	n/a	n/a	Mitochondrial sequence	Mitochondrial sequence	Mitochondrial sequence
Tc014D01R, Tc007F08R, Tc005C11R, Tc023B06R	n/a	n/a	Mitochondrial sequence	Mitochondrial sequence	Mitochondrial sequence
Tc025H10R, Tc014D07R, Tc011F03R, Tc021D02R, Tc023C08R, Tc026F03R, Tc020D01R, Tc005H01R	n/a	n/a	Mitochondrial sequence	Mitochondrial sequence	Mitochondrial sequence
Tc001F05R, Tc001D12R, Tc004H01R, Tc001D12F, Tc001F05F	n/a	n/a	Mitochondrial sequence	Mitochondrial sequence	Mitochondrial sequence
Tc010H02R, Tc002F04R, Tc003E02R	n/a	n/a	Mitochondrial sequence	Mitochondrial sequence	Mitochondrial sequence
Tc022D01R	n/a	n/a	Mitochondrial sequence	Mitochondrial sequence	Mitochondrial sequence
Tc025F09R, Tc023D09R	n/a	n/a	Mitochondrial sequence	Mitochondrial sequence	Mitochondrial sequence
Tc020A03F, Tc020A03R	No hits	No hits	n/a	n/a	n/a
Tc009H06R	No hits	No hits	n/a	n/a	n/a
Tc001B07F, Tc001B07R	No hits	No hits	n/a	n/a	n/a
Tc001B12F, Tc001B12R	No hits	No hits	n/a	n/a	n/a
Tc001C08R, Tc001C08F	No hits	No hits	n/a	n/a	n/a
Tc001C10R, Tc001C10F	No hits	No hits	n/a	n/a	n/a
Tc001E12R, Tc001E12F	No hits	No hits	n/a	n/a	n/a
Tc001F04F, Tc001F04R	No hits	No hits	n/a	n/a	n/a
Tc001F06F, Tc001F06R	No hits	No hits	n/a	n/a	n/a
Tc003B02F, Tc003B02R	No hits	No hits	n/a	n/a	n/a
Tc003H04R, Tc003H04F	No hits	No hits	n/a	n/a	n/a
Tc004B09F, Tc004B09R	No hits	No hits	n/a	n/a	n/a

TcEST Contigs	NCBI tBLASTx (*Drosophila*)	NCBI tBLASTx (Human)	BDGP BLASTx (e<1e-03)	Putative function	General function
Tc005H07F, Tc005H07R	No hits	No hits	n/a	n/a	n/a
Tc006A01F, Tc006A01R	No hits	No hits	n/a	n/a	n/a
Tc006A07R, Tc006A07F	No hits	No hits	n/a	n/a	n/a
Tc006A11R, Tc006A11F	No hits	No hits	n/a	n/a	n/a
Tc006A12F, Tc006A12R	No hits	No hits	n/a	n/a	n/a
Tc006E08R, Tc006E08F	No hits	No hits	n/a	n/a	n/a
Tc021E10R, Tc006E09R	No hits	No hits	n/a	n/a	n/a
Tc008A05F, Tc008A05R	No hits	No hits	n/a	n/a	n/a
Tc008E08R, Tc008E08F	No hits	No hits	n/a	n/a	n/a
Tc008E09R, Tc008E09F	No hits	No hits	n/a	n/a	n/a
Tc008E11R, Tc010D09R	No hits	No hits	n/a	n/a	n/a
Tc009B03F, Tc009B03R	No hits	No hits	n/a	n/a	n/a
Tc009H12R, Tc009H12F	No hits	No hits	n/a	n/a	n/a
Tc010C05F, Tc010C05R	No hits	No hits	n/a	n/a	n/a
Tc010C10F, Tc010C10R	No hits	No hits	n/a	n/a	n/a
Tc010G05F, Tc010G05R	No hits	No hits	n/a	n/a	n/a
Tc011C04F, Tc011C04R	No hits	No hits	n/a	n/a	n/a
Tc011E02R, Tc011E02F	No hits	No hits	n/a	n/a	n/a
Tc011G04R, Tc011G04F	No hits	No hits	n/a	n/a	n/a
Tc012D11F, Tc012D11R	No hits	No hits	n/a	n/a	n/a
Tc014B08R, Tc014B08F	No hits	No hits	n/a	n/a	n/a
Tc014C02F, Tc014C02R	No hits	No hits	n/a	n/a	n/a
Tc014C09R, Tc014C09F	No hits	No hits	n/a	n/a	n/a
Tc014F01F, Tc014F01R	No hits	No hits	n/a	n/a	n/a
Tc015B03F, Tc015B03R	No hits	No hits	n/a	n/a	n/a
Tc016G10F, Tc016G10R	No hits	No hits	n/a	n/a	n/a
Tc020B07F, Tc020B07R	No hits	No hits	n/a	n/a	n/a
Tc021E05R, Tc021E05F	No hits	No hits	n/a	n/a	n/a
Tc025E09R, Tc025E09F	No hits	No hits	n/a	n/a	n/a

TcEST Contigs	NCBI tBLASTx (*Drosophila*)	NCBI tBLASTx (Human)	BDGP BLASTx (e<1e-03)	Putative function	General function
Tc025E10R, Tc025E10F	No hits	No hits	n/a	n/a	n/a
Tc027H03F, Tc027H03R	No hits	No hits	n/a	n/a	n/a
Tc024E10R, Tc001H10R, Tc001H10F	No hits	No hits	n/a	n/a	n/a
Tc007B04R, Tc005A07R, Tc007B04F	No hits	No hits	n/a	n/a	n/a
Tc015E12R, Tc016G06R, Tc005A09R	No hits	No hits	n/a	n/a	n/a
Tc005B06F, Tc005B06R, Tc005D07R	No hits	No hits	n/a	n/a	n/a
Tc021D04F, Tc021D04R, Tc005E03R	No hits	No hits	n/a	n/a	n/a
Tc011G11R, Tc007F01R, Tc007F01F	No hits	No hits	n/a	n/a	n/a
Tc023E07R, Tc011A07R, Tc011A07F	No hits	No hits	n/a	n/a	n/a
Tc012G10R, Tc011G07R, Tc012G10F	No hits	No hits	n/a	n/a	n/a
Tc026D03F, Tc025H02R, Tc026D03R	No hits	No hits	n/a	n/a	n/a
Tc003G01R, Tc002D12R, Tc010C06R, Tc002D12F	No hits	No hits	n/a	n/a	n/a
Tc012G09F, Tc007G06R, Tc012G09R, Tc005E02R	No hits	No hits	n/a	n/a	n/a
Tc006G09F, Tc016H12R, Tc012D04R, Tc006G09R	No hits	No hits	n/a	n/a	n/a
Tc025C05R, Tc011H03R, Tc009E05R, Tc009E05F	No hits	No hits	n/a	n/a	n/a
Tc025D03R, Tc020C02R, Tc024F07R, Tc020C02F	No hits	No hits	n/a	n/a	n/a
Tc005B05F, Tc027H12R, Tc021H08R, Tc005B05R, Tc021H08F	No hits	No hits	n/a	n/a	n/a
Tc009H10R, Tc021H04R, Tc005H03R, Tc015A08R, Tc012H11R, Tc001F11R	No hits	No hits	n/a	n/a	n/a
Tc011D06R, Tc021D05R, Tc005A05R, Tc011D06F, Tc004D09R, Tc005A01R	No hits	No hits	n/a	n/a	n/a
Tc009A11R, Tc016F10R, Tc013G03R, Tc005B03R, Tc013G03F, Tc012A11R, Tc023G05R	No hits	No hits	n/a	n/a	n/a
Tc020A12R, Tc001F01F, Tc027E05R, Tc005A04R, Tc003E03R, Tc004G05R, Tc025A08R, Tc007E10R, Tc007B11R, Tc020G05R, Tc010D05R, Tc008H10R, Tc020H07R, Tc002D08R, Tc001F01R, Tc006B05R, Tc008C09R, Tc023A04R, Tc009E04R	No hits	No hits	n/a	n/a	n/a
Tc024B02R	No hits	No hits	n/a	n/a	n/a
Tc026C03R	No hits	No hits	n/a	n/a	n/a
Tc015C01R	No hits	No hits	n/a	n/a	n/a
Tc005C03R	No hits	No hits	n/a	n/a	n/a

TcEST Contigs	NCBI tBLASTx (*Drosophila*)	NCBI tBLASTx (Human)	BDGP BLASTx (e<1e-03)	Putative function	General function
Tc007E03R	No hits	No hits	n/a	n/a	n/a
Tc015F12R	No hits	No hits	n/a	n/a	n/a
Tc012E04R	No hits	No hits	n/a	n/a	n/a
Tc027C03R	No hits	No hits	n/a	n/a	n/a
Tc006H10R	No hits	No hits	n/a	n/a	n/a
Tc015D10R	No hits	No hits	n/a	n/a	n/a
Tc020G03R	No hits	No hits	n/a	n/a	n/a
Tc012C07R	No hits	No hits	n/a	n/a	n/a
Tc009H11R	No hits	No hits	n/a	n/a	n/a
Tc015C08R	No hits	No hits	n/a	n/a	n/a
Tc003D02R	No hits	No hits	n/a	n/a	n/a
Tc025D01R	No hits	No hits	n/a	n/a	n/a
Tc005E12R	No hits	No hits	n/a	n/a	n/a
Tc001A04R	No hits	No hits	n/a	n/a	n/a
Tc026G10R	No hits	No hits	n/a	n/a	n/a
Tc020C05R	No hits	No hits	n/a	n/a	n/a
Tc007G04R	No hits	No hits	n/a	n/a	n/a
Tc026D10R	No hits	No hits	n/a	n/a	n/a
Tc012F01R	No hits	No hits	n/a	n/a	n/a
Tc014G02R	No hits	No hits	n/a	n/a	n/a
Tc004H08F, Tc004A01R, Tc004H08R	No hits	No hits	n/a	n/a	n/a
Tc025G03R, Tc025G03F	No hits	No hits	n/a	n/a	n/a
Tc026A10F, Tc026A10R	No hits	No hits	n/a	n/a	n/a
Tc001E02R, Tc014F07R, Tc027B10R, Tc001D03R, Tc001E02F, Tc001D03F	No hits	No hits	n/a	n/a	n/a
Tc007C04R, Tc023A01F, Tc023A01R	No hits	No hits	n/a	n/a	n/a
Tc012G02F, Tc012G02R	No hits	No hits	n/a	n/a	n/a
Tc025B08F, Tc025B08R	No hits	No hits	n/a	n/a	n/a
Tc021C11F, Tc021C11R	No hits	No hits	n/a	n/a	n/a

TcEST Contigs	NCBI tBLASTx (*Drosophila*)	NCBI tBLASTx (Human)	BDGP BLASTx (e<1e-03)	Putative function	General function
Tc013G01F, Tc013G01R	No hits	No hits	n/a	n/a	n/a
Tc013A09F, Tc013A09R	No hits	No hits	n/a	n/a	n/a
Tc010A12F, Tc010A12R	No hits	No hits	n/a	n/a	n/a
Tc025A01F, Tc025A01R	No hits	No hits	n/a	n/a	n/a
Tc021E07F, Tc021E07R	No hits	No hits	n/a	n/a	n/a
Tc003F01F, Tc003F01R	No hits	No hits	n/a	n/a	n/a
Tc022G12F, Tc022G12R	No hits	No hits	n/a	n/a	n/a
Tc023C07F, Tc024H12R, Tc023C07R, Tc027H05R	No hits	No hits	n/a	n/a	n/a
Tc015F07F, Tc025G08R, Tc015F07R	No hits	No hits	n/a	n/a	n/a
Tc014G11F, Tc014G11R	No hits	No hits	n/a	n/a	n/a
Tc023D06F, Tc023D06R	No hits	No hits	n/a	n/a	n/a
Tc010C03F, Tc010C03R	No hits	No hits	n/a	n/a	n/a
Tc016A02F, Tc016A02R	No hits	No hits	n/a	n/a	n/a
Tc009B12F, Tc009B12R	No hits	No hits	n/a	n/a	n/a
Tc023D01F, Tc023D01R	No hits	No hits	n/a	n/a	n/a
Tc013G04F, Tc013G04R	No hits	No hits	n/a	n/a	n/a
Tc004A02F, Tc004A02R	No hits	No hits	n/a	n/a	n/a
Tc025F10F, Tc025F10R	No hits	No hits	n/a	n/a	n/a
Tc023H02F, Tc023H02R	No hits	No hits	n/a	n/a	n/a
Tc004E08F, Tc004E08R	No hits	No hits	n/a	n/a	n/a
Tc016F11F, Tc016F11R	No hits	No hits	n/a	n/a	n/a
Tc007E09F, Tc007E09R	No hits	No hits	n/a	n/a	n/a
Tc025C01F, Tc025C01R	No hits	No hits	n/a	n/a	n/a
Tc009E01F, Tc009E01R	No hits	No hits	n/a	n/a	n/a
Tc007E04F, Tc007E04R	No hits	No hits	n/a	n/a	n/a
Tc003H06F, Tc003H06R	No hits	No hits	n/a	n/a	n/a
Tc006C06F, Tc023B09R, Tc010D11R, Tc006C06R, Tc010C08R, Tc012E05R	No hits	No hits	n/a	n/a	n/a
Tc013B06F, Tc013B06R	No hits	No hits	n/a	n/a	n/a

TcEST Contigs	NCBI tBLASTx (*Drosophila*)	NCBI tBLASTx (Human)	BDGP BLASTx (e<1e-03)	Putative function	General function
Tc014H09F, Tc014H09R	No hits	No hits	n/a	n/a	n/a
Tc009A06F, Tc009A06R	No hits	No hits	n/a	n/a	n/a
Tc006E05R, Tc006E05F	No hits	No hits	n/a	n/a	n/a
Tc012C03R, Tc012C03F	No hits	No hits	n/a	n/a	n/a
Tc012F04R, Tc012F04F	No hits	No hits	n/a	n/a	n/a
Tc004B12F, Tc004B12R	No hits	No hits	n/a	n/a	n/a
Tc005G08F, Tc005G08R	No hits	No hits	n/a	n/a	n/a
Tc020G06F, Tc020G06R	No hits	No hits	n/a	n/a	n/a
Tc025F01F, Tc025F01R	No hits	No hits	n/a	n/a	n/a
Tc011H11R, Tc013B12R, Tc003C08R, Tc010E04R, Tc011C03R, Tc011G05R, Tc008E03R, Tc006D10R, Tc002A10R, Tc005H09R, Tc006D01R, Tc013F08R, Tc003D08R, Tc012C01R, Tc020E01R, Tc020A10R, Tc008B01R, Tc006F07R, Tc015C02R	No hits	No hits	n/a	n/a	n/a
Tc021C10R	No hits	No hits	n/a	n/a	n/a

Appendix II – Summary of Exelixis EST data

Exelixis ID	Internal ID	Domain	NCBI tBLASTx (species, e value)	BDGP BLASTx (gene, e value)	Selected for cloning?	5' GSP from Primer3	3' GSP from Primer3	Cloned?
ex|977283	Tcex001	bHLH?	No hits	No hits	Yes	CGCCGGGAGATTCGCCGATTGAAGTC	None	Yes
ex|1005841	Tcex002	bHLH?	No hits	No hits	Yes	GGCTAGGCGAAAAAGCTGTA	None	Yes
ex|1006503	Tcex003	bHLH?	No hits	No hits	Yes	AAGAGAAGGAGTCCGACCGGATTATG	None	Yes
ex|1006255	Tcex004	bHLH	Drosophila 2e-35	tango (tgo) 1.7e-29	Yes	CCGTTGTGTTGTAAACAGTCTAGT	GCCTTCATGTGGGCTACG	Yes
ex|1006866	Tcex005	bHLH	Drosophila 3e-37	tango (tgo) 2.6e-43	Yes	AGTTGTCGACGCAAGAACCT	CGCACTGTTTGATCCTGACA	Yes
ex|1006089	Tcex006	bHLH	Drosophila 2e-18	CG11867 6.8e-18	Yes	ATGCCATTCTGTCCATTTCC	TCGCCCAACTTCTTCAACTT	Not completed
ex|1024069	Tcex007	bHLH	Drosophila 1e-64	Clock (Clk) 1.2e-54	Yes	AAACGGAAATCTCGAAACTTGA	CTCCCCCGAGTAGCCACTAT	Yes
ex|1005789	Tcex008	bHLH	Drosophila 4e-26	extra macrochaetae (emc) 3.3e-18	Yes	GGGTTCGTTCGCTCGTAGT	None	Yes
ex|1023810	Tcex009	bHLH	Tribolium twist 5e-18	n/a	No	n/a	n/a	n/a
ex|1090954	Tcex010	bHLH	Drosophila 2e-54	bigmax 9.9e-46	Yes	CACACCCAAGCTGAACAGAA	None	Yes
ex|1023959	Tcex011	bHLH	Rat 6e-29, Drosophila 1e-23	Helix loop helix protein 106 (HLH106) 4.8e-24	Yes	AGGATTCCTCAAGGGACTGG	None	Yes
ex|1024213	Tcex012	bHLH	Tribolium Atonal homolog 1 (Tath1) 2e-25	n/a	No	n/a	n/a	n/a
ex|1023816	Tcex013	bHLH	Tribolium Atonal homolog 2 (Tath2) 9e-22	n/a	No	n/a	n/a	n/a
ex|1004883	Tcex014	Homeobox	Drosophila 3e-47	achintya 2.3e-39	Yes	AAACGTGCTTTGGTGTTGCT	None	Yes
ex|1024207	Tcex015	Homeobox	Tribolium even-skipped e-142	n/a	No	n/a	n/a	n/a
ex|2057780	Tcex016	Homeobox	Tribolium cephalothorax e-163	n/a	No	n/a	n/a	n/a
ex|1090806	Tcex017	Homeobox/ bHLH	Drosophila 6e-89	extradenticle (exd) 2.9e-71	Yes	AAATCTACCACCAAGAACTGGAAAAG	None	Yes
ex|1023812	Tcex018	Homeobox	Tribolium fushi-tarazu 0.0	n/a	No	n/a	n/a	n/a
ex|977293	Tcex019	Homeobox	Artemia 5e-64, Drosophila 9e-64	ventral veins lacking (vvl) 9.1e-52	Yes	TTCGCCAAACAGTTCAAACA	None	Yes
ex|1023977	Tcex020	Homeobox	Tribolium orthodenticle-1 0.0	n/a	No	n/a	n/a	n/a

Exelixis ID	Internal ID	Domain	NCBI tBLASTx (species, e value)	BDGP BLASTx (gene, e value)	Selected for cloning?	5' GSP from Primer3	3' GSP from Primer3	Cloned?
ex|1023969	Tcex021	Homeobox	Tribolium orthodenticle-2 0.0	n/a	No	n/a	n/a	n/a
ex|1023818	Tcex022	Homeobox	Tribolium Abdominal-A e-122	n/a	No	n/a	n/a	n/a
ex|2057777	Tcex023	Homeobox	Tribolium Abdominal-B e-143	n/a	No	n/a	n/a	n/a
ex|1023869	Tcex024	Homeobox	Tribolium caudal A e-154	n/a	No	n/a	n/a	n/a
ex|1023867	Tcex025	Homeobox	Tribolium caudal B e-179	n/a	No	n/a	n/a	n/a
ex|1024209	Tcex026	Homeobox	Tribolium ultrathorax 7e-42	n/a	No	n/a	n/a	n/a
ex|978215	Tcex027	Homeobox?	No hits	No hits	Yes	ACGAAACGCGTTCGGCTTAG	None	Yes
ex|3260632	Tcex028	Homeobox	Tribolium Deformed1 2e-90	n/a	No	n/a	n/a	n/a
ex|1023975	Tcex029	Homeobox	Tribolium Deformed 0.0	n/a	No	n/a	n/a	n/a
ex|1319541	Tcex030	Homeobox	Tribolium Maxillopedia 0.0	n/a	No	n/a	n/a	n/a
ex|1023865	Tcex031	Homeobox	Tribolium Zerknuellt e-173	n/a	No	n/a	n/a	n/a
ex|3260626	Tcex032	Homeobox	Tribolium prothoraxless e-180	n/a	No	n/a	n/a	n/a
ex|1007536	Tcex033	Zf-C2H2	Human 1e-44, Drosophila 4e-37	CG14435 9.7e-32	Yes	AATAATAAGCGCCGGTAAAGC	None	Yes
ex|977989	Tcex034	Zf-C2H2	Drosophila 4e-23	tramtrack (ttk) 1.1e-18	Yes	CATTTTGCCGGGTCTAGAAA	CCAGAGGACGTTCAAAAACAA	Not completed
ex|1004285	Tcex035	Zf-C2H2	Echinoderm 5e-17, Drosophila 3e-16	Adult enhancer factor 1 (Aef1) 4.0e-17	Yes	GTCGAAAACGAGCAACAACC	None	Yes
ex|1024005	Tcex036	Zf-C2H2	Human 4e-14, Drosophila 5e-11	CG9932 1.4e-17	Yes	ATGCGAGCACTGTCCCTTTA	None	Yes
ex|1090556	Tcex037	Zf-C2H2	Human 1e-17, Drosophila 3e-14	CG4360 9.4e-15	Yes	TGTTCATCCAGAGGCTGAGA	TCATCGCAGATGAGACAAACA	No
ex|1007080	Tcex038	Zf-C2H2	Human 4e-44, Drosophila 2e-37	CG5669 8.9e-25	Yes	GGATGGATGCCACAAGATTT	None	Yes
ex|1009838	Tcex039	Zf-C2H2	Human 3e-36, Drosophila 8e-34	zinc-finger-motif-protein 8.0e-25	Yes	CAACAAAGTCATCCGAGACG	GTTTCGCCCAGAAAAACAAA	Yes
ex|1090926	Tcex040	Zf-C2H2	Drosophila 7e-15	CG3407 2.6e-13	Yes	GTCCGGTGAGTTTTGTGGAA	GCCGAATCCACACTCCTAAT	Yes
ex|1023873	Tcex041	Zf-C2H2	Tribolium Krueppel 6e-51	n/a	No	n/a	n/a	n/a

Exelixis ID	Internal ID	Domain	NCBI tBLASTx (species, e value)	BDGP BLASTx (gene, e value)	Selected for cloning?	5' GSP from Primer3	3' GSP from Primer3	Cloned?
ex|1024093	Tcex042	Zf-C2H2	Tribolium Hunchback 0.0	n/a	No	n/a	n/a	n/a
ex|1008714	Tcex043	Zf-C2H2	Mouse 4e-60, Drosophila 3e-51	males absent on the first (mof) 1.0e-37	Yes	TTCCGAAACAATGCGTAAAA	None	Yes
ex|977617	Tcex044	Zf-C2H2	Chicken 1e-70, Drosophila 2e-70	CG3850 1.1e-64	Yes	AGGAGCGTCATCCAACACAC	None	Yes
ex|1007906	Tcex045	Zf-C2H2	Drosophila 4e-43	no ocelli (noc) 1.5e-51	Yes	ATACACCTTTGGGGGCGTA	None	Yes
ex|1090940	Tcex046	Zf-C2H2	Drosophila 8e-23	CG9638 2.1e-29	Yes	GCATCTCGTGCCTGTGTTT	None	Yes
ex|1024179	Tcex047	Zf-C2H2?	No hits	No hits	Yes	TTGTCGGCACGAAACAATTA	TGCGAGAAGTTCTTCGTGAG	Yes
ex|1006651	Tcex048	Zf-C2H2	Mouse 5e-50, Drosophila 1e-45	Zn72D 1.5e-37	Yes	CATATTCGTGGAGCCAAACA	TTCACGTAAAGCAGCACGTC	Yes
ex|1009462	Tcex049	Zf-C2H2	Drosophila 1e-20	longitudinals lacking (lola) 3.1e-9	Yes	CGTTTTGAATGCCTTTGGTT	CTTCACCGAAACTATGTATGTGC	Yes
ex|1005462	Tcex050	Zf-C2H2?	No hits	No hits	Yes	GCAGCGCTACAACTGCGAGTTTTGCG	None	Yes
ex|1005221	Tcex051	Zf-C2H2	Drosophila 5e-17	CG12701 1.3e-15	Yes	AACGGCCCCTTACTGCTACT	None	Yes
ex|1008716	Tcex052	Zf-C2H2?	No hits	No hits	Yes	GCGTGGGTTTTGCTGATATT	GCGGTACTGGAAGACTTTGG	No
ex|1007825	Tcex053	Zf-C2H2	Human 4e-11, Drosophila 5e-11	CG14962 3.8e-16	Yes	GTCCGCAAAAACTGCCATAA	TTTCACACTTGGAGGCAGTG	Yes
ex|1008016	Tcex054	Zf-C2H2	Drosophila 2e-29	EP2237 1.3e-29	Yes	TGGAACCCCAAATGAACATC	None	Yes
ex|1008562	Tcex055	Zf-C2H2?	No hits	No hits	Yes	GTTTAGGTTTTAGCGAGAACATG	None	Yes
ex|1006069	Tcex056	Zf-C2H2	Human 3e-36, no Drosophila	CG17395 4.3e-29	Yes	CTGAGCCAGCAAAACCTGAT	None	Yes
ex|1006735	Tcex057	Zf-C2H2	Drosophila 1e-33	CG8209 1.6e-34	Yes	CGTCCCCAAAATTAACCACA	None	Yes
ex|1024211	Tcex058	Zf-C2H2	Tribolium Snail 1e-28	n/a	No	n/a	n/a	n/a
ex|1006321	Tcex059	Zf-C2H2	Human 2e-14, Drosophila 2e-9	CG5204 7.7e-14	Yes	CAAGCAATACCAGTGCGAGA	None	Yes
ex|1008738	Tcex060	Zf-C2H2?	No hits	No hits	Yes	CCGTGATCTCGTTCAACACTGGCAAG	None	Yes
ex|978241	Tcex061	Zf-C2H2?	No hits	tramtrack (ttk) 1.3e-10	Yes	AACATGCCACATCTGCAAAG	None	Yes
ex|1004148	Tcex062	Zf-C2H2	Drosophila 9e-14	CG8159 2.3e-20	Yes	GCCTGTCATCTGTCATTTTTCTC	None	Yes
ex|1091285	Tcex063	Zf-C3HC4	Human 4e-64 Drosophila 1e-50	CG5382 7.3e-50	Yes	TGGCCCCCATGTATTTTATT	None	Yes
ex|1090898	Tcex064	Zf-C3HC4	Drosophila 2e-30	neuralized (neur) 2.1e-25	Yes	TCGGAAGGAGTGGTGTGTCT	GTGAACCCGAACCGAATAAC	Not completed
ex|1023985	Tcex065	Zf-C3HC4	Drosophila 3e-57	Plenty of SH3s (POSH) 3.9e-48	Yes	GTCTCGGATTATCGCCCATC	CTTGTTTTCCGCCACATTCT	Yes

Exelixis ID	Internal ID	Domain	NCBI tBLASTx (species, e value)	BDGP BLASTx (gene, e value)	Selected for cloning?	5' GSP from Primer3	3' GSP from Primer3	Cloned?
ex|1006407	Tcex066	Zf-C3HC4	Human 6e-47, Drosophila 1e-38	CG11982 3.2e-11	Yes	GCCGAATGTGTGTGTGAATG	CCACGGCATATAGGACAGGT	No
ex|1008488	Tcex067	bZIP?	No hits	No hits	Yes	CGATTTTCGTGTGGTACCTG	CGCTATGTCGCTAATGCTGT	Yes
ex|1005105	Tcex068	bZIP	Human 1e-18, Drosophila 7e-15	Cyclic-AMP response element binding protein A (CrebA) 1.2e-14	Yes	AGTTGGTTGAGGTGGAAACG	CCCGTCAACGTACTCCTTTT	Yes
ex|978067	Tcex069	bZIP?	No hits	No hits	Yes	CGGTGACGACGAATACGTAAC	TCACTGTCGGCTTCTAAGGAA	Yes
ex|1091291	Tcex070	bZIP	Drosophila 5e-21	Jun-related antigen (Jra) 5.0e-19	Yes	CGCAGCCTCTAAATGTCGTT	None	Yes
ex|1004723	Tcex071	bZIP	Drosophila 3e-24	X box binding protein-1 (Xbp1) 2.5e-24	Yes	ACGACGTCAACGAAGCAAG	None	Yes
ex|1003440	Tcex072	Myb-like	Drosophila 7e-49	CG3168 4.8e-40	Yes	AGGAGGAACAGCAGGAGAAA	None	Yes
ex|1006671	Tcex073	Myb-like	Human 3e-61, Drosophila 4e-54	metastasis-associated-1-like-protein 2.6e-39	Yes	AGGGCGTCTGAGGCTAATCT	GTCGGTGTGACATTGCTGAC	Yes
ex|3260624	Tcex074	Forkhead	Tribolium Forkhead 0.0	n/a	No	n/a	n/a	n/a
ex|1004623	Tcex075	Forkhead	Human 3e-54, Drosophila 2e-47	forkhead domain 59A (fd59A) 8.0e-44	Yes	TTGCGGTGGAGTTTAGTGTG	GGTCCAGCGTCCAGTAGTTG	Not completed
ex|1007492	Tcex076	Cold-shock	Drosophila 6e-58	CG7015 1.2e-54	Yes	CGAGCCCAACAGTCTTGAAC	None	Yes
ex|1006609	Tcex077	Cold-shock	Drosophila 5e-53	ypsilon schachtel (yps) 4.9e-44	Yes	GAAAGGCAGGAAGGAGGAAG	TTTGACTCTCACGTGGTGGA	Yes
ex|1007066	Tcex078	Cold-shock	Drosophila 5e-80	CG7015 7.3e-65	Yes	AGCGCCGTTACAAAAATAGC	CCGGAAGGTAGACGGGTTAT	Yes
ex|1006615	Tcex079	Cold-shock	Drosophila 2e-47	ypsilon shachtel (yps) 2.2e-39	Yes	GGAAGGAAGAACATCGTGGA	TTTGACTCTCACGTGGTGGA	No
ex|1005379	Tcex080	TFIID?	No hits	No hits	No	n/a	n/a	n/a
ex|1008462	Tcex081	TFIID	Drosophila 2e-64	TBP-associated factor 80kD (Taf80) 2.4e-53	No	n/a	n/a	n/a
ex|1003516	Tcex082	TFIID	Drosophila 9e-52	TATA box binding protein-related factor 2 (Trf2) 7.4e-46	No	n/a	n/a	n/a
ex|977375	Tcex083	Chromo Domain	Human 1e-35, Drosophila 5e-21	CG7041 3.6e-25	Yes	ATAACGACCTCGGGACACCT	TTGGGAACCTGAGGAGAATTT	Not completed
ex|1007146	Tcex084	LIM	Drosophila e-103	Paxillin (Pax) 2.3e-85	Yes	CCCAAGCAGAACTTGGACTC	None	Yes

Exelixis ID	Internal ID	Domain	NCBI tBLASTx (species, e value)	BDGP BLASTx (gene, e value)	Selected for cloning?	5' GSP from Primer3	3' GSP from Primer3	Cloned?
ex|1090142	Tcex085	LIM	Mouse 9e-42, Drosophila 7e-21	CG4656 5.7e-17	Yes	GTTCGGAACTCTTCCACCAT	None	Yes
ex|1320121	Tcex086	LIM	Manduca 6e-30, Drosophila 6e-21	CG17401 4.0e-19	Yes	GCTAGTTCAACTTCGCGTTTG	None	Yes
ex|1006239	Tcex087	LIM	Drosophila 2e-42	tailup (tup) 1.5e-40	Yes	TAAGTTCGTTCGTCGGCTGT	TCGCTTGCAGTAGGTTTTCC	Not completed
ex|977203	Tcex088	LIM	Drosophila 4e-51	Muscle LIM protein at 60A (Mlp60A) 1.6e-40	Yes	GGTCCGTAAGGACACCATTG	None	Yes
ex|1005229	Tcex089	LIM	Drosophila 1e-15	No hits	Yes	CGGTATCCCGATTGAAGAGA	None	Yes
ex|1005227	Tcex090	LIM	Drosophila 3e-16	CG17467 2.5e-17	Yes	CGTGAGAAAGACGCAGAAGA	None	Yes
ex|1009388	Tcex091	LIM	Drosophila 3e-34	No hits	Yes	GTTGCCATAAACCGATTTCC	None	Yes
ex|1006281	Tcex092	LIM	Drosophila e-138	CG11916 1.0e-59	Yes	GCCCGTCTTCAAAGGATAGTC	AACGCGTGGGGAATTTCT	Yes
ex|1006445	Tcex093	LIM?	No hits	No hits	Yes	AAGGCATCTGTTCGACTGCT	None	Yes
ex|1009520	Tcex094	LIM	Drosophila 1e-77	Muscle LIM protein at 84B (Mlp84B) 2.5e-65	Yes	GGAACAATGTGGCACAAGAA	None	Yes
ex|1320191	Tcex095	LIM	Drosophila 6e-39	CG8242 1.1e-34	Yes	GGAAAACTCTTTGGCAACTCG	None	Yes
ex|1004701	Tcex096	LIM	Drosophila 5e-95	CG11063 7.0e-93	Yes	TACGCGGCAAAGCTTTCTAT	None	Yes
ex|1008722	Tcex097	LIM	Manduca 3e-33, Drosophila 2e-24	CG17401 7.5e-25	Yes	GTGGTGCGAATTTTCCCTTA	None	Yes
ex|1008371	Tcex098	LIM	Drosophila 3e-44	CG9489 1.7e-39	Yes	CGCGCAAAGTCGGTACAT	ATCTCCTTCCCCGTGTAGGT	Yes
ex|1319659	Tcex099	LIM	Drosophila 2e-66	Muscle LIM protein at 84B (Mlp84B) 2.4e-51	Yes	ACACCATTGAGGACCTCTCC	TGGCCGTAACCATATCCTTT	Yes
ex|1006597	Tcex100	LIM	Drosophila 7e-19	CG12969 1.7e-27	Yes	CGCACAGACCGAAGTTTTGT	None	Yes
ex|1024043	Tcex101	HMG?	No hits	No hits	Yes	CCGCAACTACCGCAGATTTA	GGGTAATACGCCAGATGATTG	Yes
ex|1023804	Tcex102	HMG	Drosophila 3e-24	Sox box protein 15 (Sox15) 1.2e-20	Yes	GTCCGCGTTTCAAAGACAG	CCACTTCTTGCCCAACATTT	No
ex|1010024	Tcex103	SAM	Drosophila 2e-62	CG11199 1.1e-57	Yes	AGACCTCAGGCCCAACAAC	AGGAATGCGTATCGAAGCTG	Yes
ex|1008920	Tcex104	SAM	Drosophila 2e-33	CG7915 4.6e-37	Yes	AAAAGCCAAGCATTTGTTGC	CGACGGACTAGCGGTCTTAT	Yes
ex|1008572	Tcex105	SAM	Human 2e-64, Drosophila 3e-54	CG11867 1.1e-42	Yes	AGCCGACTAATGCGCTAAAT	TTTCGGCGCTATGGTTTTTA	Yes
ex|1005929	Tcex106	SAM	Human 9e-83, Drosophila 3e-82	CG4719 1.6e-67	Yes	CGAAATTGGAGGGTGATGAT	None	Yes
ex|1090324	Tcex107	PHD	Drosophila 3e-27	bip2 7e-23	Yes	GGGAACGAGGTGTGGATTT	None	Yes

Exelixis ID	Internal ID	Domain	NCBI tBLASTx (species, e value)	BDGP BLASTx (gene, e value)	Selected for cloning?	5' GSP from Primer3	3' GSP from Primer3	Cloned?
ex|1008532	Tcex108	Zf-CCCH	Drosophila 3e-50	CG10084 3.9e-53	Yes	CCTGGCACACCTACACAAGA	CGCTGTTTTTCCGCTTTTAT	Yes
ex|1007566	Tcex109	BTB/POZ?	No hits	No hits	Yes	AGTGGATATTGCCGTCTGTTG	None	Yes
ex|1006457	Tcex110	BTB/POZ	Drosophila 5e-38	modifier of mdg4 (mod(mdg4)) 8.4e-32	Yes	GGTGGCTGAGGTGTAACAAG	CGCTTTTCAACTCCTCTGGT	Yes
ex|1023859	Tcex111	BTB/POZ	Drosophila 4e-38	modifier of mdg4 (mod(mdg4)) 8.4e-33	Yes	AAACACTTACTTTCGCCCACA	None	Yes
ex|1006287	Tcex112	BTB/POZ	Drosophila 1e-66	longitudinals lacking (lola) 4.2e-59	Yes	CCAACAGTTTTGTCTGAGATGG	None	Yes
ex|1007002	Tcex113	BTB/POZ	Drosophila 2e-56	fruitless (fru) 9.3e-54	Yes	AGAACTAGTACCTCACCGACGTG	None	Yes
ex|1005404	Tcex114	BTB/POZ	Mouse 2e-42, Drosophila 4e-28	diablo (dbo) 3.6e-31	Yes	CACTTCCCAAGCCAGTGAAC	None	Yes
ex|1009508	Tcex115	SAND	Drosophila 1e-47	Deformed epidermal autoregulatory factor-1 (Deaf1) 1.9e-31	Yes	CTCATCACCAGCACCTTCAA	None	Yes
ex|1008876	Tcex116	SNF2	Drosophila 2e-94	brahma (brm) 3.1e-77	Yes	TCGAGGAGAAGTATTGTGACCA	AATCCGCTCTTCAACCGAAT	Yes
ex|1006077	Tcex117	DM?	No hits	No hits	Yes	TGAAGAGCCTATGCATCATTT	GTACGCGCCTGTGGTGTA	Yes
ex|1006257	Tcex118	Enhancer of zeste-like	Drosophila 1e-45	Enhancer of zeste (E(z)) 1.3e-37	Yes	AATGCACCCCAAATATCGAC	None	Yes
ex|1023686	Tcex119	Enhancer of zeste-like	Drosophila 5e-40	Enhancer of zeste (E(z)) 3.7e-38	Yes	ATGGTGCCTGTGCTTAGAGG	CCGTCCTGGTCCAGTAAATC	Yes
ex|1009828	Tcex120	Heat-shock?	No hits	No hits	Yes	GCGTCCGGAAAAAGTAACAT	GGCATTAGCATCAACAAATGG	No
ex|1007400	Tcex121	Zf-AN1	Mouse 4e-50, Drosophila 1e-35	CG12795 7.1e-36	Yes	TCTAGTGAACCCTCGCACCT	None	Yes
ex|977475	Tcex122	Bromo Domain	Drosophila 2e-21	brahma (brm) 7.6e-18	Yes	TAAGATTTTGGGCCGAATTG	None	Yes
ex|1320067	Tcex123	bicaudal-like	Drosophila 6e-56	bicaudal (bic) 1.1e-46	Yes	TGCAAAAGACTGTGCAGTGG	None	Yes
ex|1009914	Tcex124	TEA	Human 1e-46, Drosophila 8e-36	scalloped (sd) 3.8e-46	Yes	CCAAGCCAAATTGAAGGTACA	GCCCAGAATTTGACAAGGAA	Yes
ex|1024099	Tcex125	TFIIS	Drosophila 2e-53	RNA polymerase II elongation factor (TfIIS) 3.2e-49	No	n/a	n/a	n/a

Appendix III – Summary of the EST *in situ* screen

Clone	General function	In situ pattern description	Category
Tc001A04	No hits	n/a	n/a
Tc001A10	Unknown	Ubiquitous - all stages	Ubiquitous
Tc001B02	Unknown	Ubiquitous - all stages	Ubiquitous
Tc001B07	No hits	No signal	No signal
Tc001B12	No hits	Ubiquitous - all stages	Ubiquitous
Tc001C08	No hits	Ubiquitous - all stages (low level), early segmental stripes in elongating germ band, higher expression in developing appendages and possible neurogenesis pattern in very old embryos.	Ubiquitous, segmentation, appendages, neurogenesis
Tc001C10	No hits	n/a	n/a
Tc001E02	No hits	Ubiquitous all stages but expression excluded from ventral midline at late stages	Ubiquitous, other
Tc001E12	No hits	Ubiquitous - early stages, thoracic expression in elongating germ band, appendages	Ubiquitous, appendages
Tc001F01	No hits	No signal	No signal
Tc001F02	Unknown	No signal	No signal
Tc001F04	No hits	Ventral midline	Other
Tc001F06	No hits	n/a	n/a
Tc001H10	No hits	Ubiquitous - early stages	Ubiquitous
Tc002A08	Unknown	No signal	No signal
Tc002D12	No hits	Serosa (?), yolk cells, ubiquitous - late stages	Extra-embryonic tissues, ubiquitous
Tc002F02	Zf-C2H2 transcription factor	No signal	No signal
Tc002G02	Unknown	n/a (failure)	n/a
Tc002G09	Unknown	Tail bud, possible segmental stripes, appendages, lateral segmental spots (tracheal openings?)	Segmentation, appendages, other
Tc002H08	Unknown	Ubiquitous - all stages	Ubiquitous
Tc003B02	No hits	n/a (failure)	n/a
Tc003C05	Unknown	n/a	n/a
Tc003D02	No hits	n/a	n/a
Tc003D03	Unknown	Dot-like expression	Other
Tc003F01	No hits	Ubiquitous - all stages	Ubiquitous

Clone	General function	In situ pattern description	Category
Tc003F11	Unknown	n/a (failure)	n/a
Tc003H04	No hits	No signal	No signal
Tc003H06	No hits	No signal	No signal
Tc004A02	No hits	Ubiquitous - all stages (?), stronger in appendages	Ubiquitous, appendages
Tc004A04	Unknown	Serosa	Extra-embryonic tissues
Tc004A11	Unknown	Ubiquitous - all stages	Ubiquitous
Tc004B01	Unknown	Ubiquitous - all stages	Ubiquitous
Tc004B09	No hits	n/a (failure)	n/a
Tc004B12	No hits	Ubiquitous - all stages	Ubiquitous
Tc004C01	Unknown	Ubiquitous - all stages	Ubiquitous
Tc004E02	Unknown	Ubiquitous - all stages	Ubiquitous
Tc004E08	No hits	Ubiquitous - all stages	Ubiquitous
Tc004E12	Unknown	No signal	No signal
Tc004F04	Signal transduction	Ubiquitous - all stages	Ubiquitous
Tc004G02	Unknown	Neurogenesis (?), background intense	Neurogenesis
Tc004H08	No hits	Ubiquitous - all stages	Ubiquitous
Tc005B01	Unknown	No signal	No signal
Tc005B05	No hits	Ubiquitous - all stages	Ubiquitous
Tc005B06	No hits	Ubiquitous - all stages	Ubiquitous
Tc005B12	HMG-box transcription factor	Neurogenesis	Neurogenesis
Tc005C03	No hits	n/a	n/a
Tc005C06	Unknown	Neurogenesis (?), background intense	Neurogenesis
Tc005C12	Unknown	Ubiquitous - all stages	Ubiquitous
Tc005E12	No hits	n/a	n/a
Tc005G08	No hits	Ubiquitous - all stages	Ubiquitous
Tc005G11	Ligand binding or carrier	Ubiquitous - all stages	Ubiquitous
Tc005H02	Unknown	Ubiquitous - all stages	Ubiquitous
Tc005H07	No hits	Ubiquitous - all stages	Ubiquitous
Tc006A01	No hits	No signal	No signal
Tc006A05	Unknown	Ubiquitous - all stages	Ubiquitous

Clone	General function	In situ pattern description	Category
Tc006A07	No hits	Ubiquitous - all stages	Ubiquitous
Tc006A11	No hits	Gap-like expression domain in elongating germ band, myogenesis	Segmentation, myogenesis
Tc006A12	No hits	Gap-like expression domain in elongating germ band, head lobes, ring in developing appendages, lateral segmental spots	Segmentation, neurogenesis, appendages, other
Tc006B02	Unknown	Ubiquitous - all stages	Ubiquitous
Tc006B11	Unknown	Ubiquitous - early stages	Ubiquitous
Tc006C06	No hits	Ubiquitous - all stages	Ubiquitous
Tc006C07	Unknown	Ubiquitous - all stages	Ubiquitous
Tc006D02	Unknown	Ubiquitous - all stages	Ubiquitous
Tc006E05	No hits	Ubiquitous - all stages	Ubiquitous
Tc006E08	No hits	Serosa	Extra-embryonic tissues
Tc006E09	No hits	n/a	n/a
Tc006E11	Unknown	No signal	No signal
Tc006G09	No hits	No signal	No signal
Tc006H10	No hits	n/a	n/a
Tc007A12	Unknown	No signal	No signal
Tc007B04	No hits	Ubiquitous - all stages	Ubiquitous
Tc007B09	Unknown	Yolk cells, ubiquitous - all stages, ring of expression in legs	Extra-embryonic tissues, ubiquitous, appendages
Tc007B10	Unknown	Ubiquitous - all stages	Ubiquitous
Tc007C05	Unknown	No signal	No signal
Tc007C06	Unknown	No signal	No signal
Tc007E03	No hits	No signal	No signal
Tc007E04	No hits	Ubiquitous - all stages	Ubiquitous
Tc007E09	No hits	Serosa, amnion	Extra-embryonic tissues
Tc007F01	No hits	Ubiquitous - all stages	Ubiquitous
Tc007F06	Signal transduction	No signal	No signal
Tc007G04	No hits	n/a	n/a
Tc007H03	Unknown	Ubiquitous - all stages	Ubiquitous
Tc007H05	HLH transcription factor	Neurogenesis, rings in legs	Neurogenesis, appendages

Clone	General function	In situ pattern description	Category
Tc007H07	Unknown	Ubiquitous - all stages, neurogenesis (?)	Ubiquitous, neurogenesis
Tc008A05	No hits	n/a	n/a
Tc008A12	Unknown	No signal	No signal
Tc008B01	No hits	No signal	No signal
Tc008B02	Unknown	Ubiquitous - all stages	Ubiquitous
Tc008B09	HMG-box transcription factor	Ubiquitous - all stages	Ubiquitous
Tc008B11	Unknown	No signal	No signal
Tc008C02	Unknown	Serosa	Extra-embryonic tissues
Tc008C06	Unknown	No signal	No signal
Tc008D01	Unknown	No signal	No signal
Tc008D03	Unknown	Ubiquitous - all stages	Ubiquitous
Tc008D05	Unknown	Ubiquitous - all stages	Ubiquitous
Tc008D07	Unknown	No signal	No signal
Tc008E08	No hits	No signal	No signal
Tc008E09	No hits	No signal	No signal
Tc008E11	No hits	No signal	No signal
Tc008F05	Unknown	Ubiquitous - all stages, neurogenesis	Ubiquitous, neurogenesis
Tc008G09	Unknown	No signal	No signal
Tc008H02	Unknown	No signal	No signal
Tc008H12	Signal transduction	No signal	No signal
Tc009A01	Unknown	No signal	No signal
Tc009A04	Signal transduction	Yolk cells, ventral midline, spots in appendages	Extra-embryonic tissues, other
Tc009A06	No hits	n/a (failure)	n/a
Tc009A08	BTB/POZ transcription factor	Ubiquitous - all stages	Ubiquitous
Tc009B03	No hits	No signal	No signal
Tc009B12	No hits	n/a (failure)	n/a
Tc009C09	Unknown	n/a (failure)	n/a
Tc009C10	Unknown	n/a (failure)	n/a
Tc009D07	Unknown	n/a (failure)	n/a

Clone	General function	In situ pattern description	Category
Tc009E01	No hits	Ubiquitous - all stages	Ubiquitous
Tc009E05	No hits	Early segmental pattern in elongating germ band, neurogenesis	Segmentation, neurogenesis
Tc009G10	Unknown	No signal	No signal
Tc009H06	No hits	n/a (failure)	n/a
Tc009H09	Unknown	No signal	No signal
Tc009H11	No hits	n/a	n/a
Tc009H12	No hits	Ubiquitous - all stages	Ubiquitous
Tc010A08	Signal transduction	Ubiquitous - all stages	Ubiquitous
Tc010A12	No hits	Ubiquitous - all stages	Ubiquitous
Tc010B05	Zf-C2H2 transcription factor	Ubiquitous - all stages	Ubiquitous
Tc010C03	No hits	Ubiquitous - all stages	Ubiquitous
Tc010C05	No hits	Ventral midline	Other
Tc010C10	No hits	Ubiquitous - all stages	Ubiquitous
Tc010F06	Unknown	Ubiquitous - early stages, neurogenesis	Ubiquitous, neurogenesis
Tc010G05	No hits	Ubiquitous - late stages	Ubiquitous
Tc010H06	Zf-C3HC4 transcription factor	No signal	No signal
Tc011A07	No hits	Ubiquitous - late stages	Ubiquitous
Tc011C04	No hits	Ubiquitous - late stages	Ubiquitous
Tc011D06	No hits	Ubiquitous - all stages	Ubiquitous
Tc011D09	Unknown	Ubiquitous - all stages. Expression is granulated. Anything to do with cell proliferation?	Ubiquitous, other
Tc011E02	No hits	Ubiquitous - late stages	Ubiquitous
Tc011G04	No hits	No signal	No signal
Tc011H01	Unknown	Ubiquitous - all stages	Ubiquitous
Tc011H08	Signal transduction	Ubiquitous - all stages	Ubiquitous
Tc012A10	Unknown	Serosa	Extra-embryonic tissues
Tc012B04	Unknown	Ubiquitous - all stages	Ubiquitous
Tc012C03	No hits	Ubiquitous - all stages	Ubiquitous
Tc012C07	No hits	Ubiquitous - all stages	Ubiquitous
Tc012D11	No hits	No signal	No signal
Tc012E04	No hits	Ubiquitous - all stages	Ubiquitous

Clone	General function	In situ pattern description	Category
Tc012E11	Unknown	Ubiquitous - all stages	Ubiquitous
Tc012F01	No hits	Ubiquitous - all stages	Ubiquitous
Tc012F03	Zf-C2H2 transcription factor	No signal	No signal
Tc012F04	No hits	No signal	No signal
Tc012G02	No hits	No signal	No signal
Tc012G05	Unknown	No signal	No signal
Tc012G09	No hits	Ubiquitous - all stages	Ubiquitous
Tc012G10	No hits	Ubiquitous - all stages	Ubiquitous
Tc013A03	Unknown	Ubiquitous - all stages	Ubiquitous
Tc013A09	No hits	Lateral segmental spots (tracheal openings?)	Other
Tc013B06	No hits	Ubiquitous - all stages	Ubiquitous
Tc013G01	No hits	Ubiquitous - all stages	Ubiquitous
Tc013G03	No hits	Ubiquitous - all stages	Ubiquitous
Tc013G04	No hits	Ubiquitous - all stages	Ubiquitous
Tc013H09	Unknown	Ubiquitous - all stages (?), rings in legs	Ubiquitous, appendages
Tc014A02	Zf-C2H2 transcription factor	Myogenesis	Myogenesis
Tc014A05	Transcription factor (possible)	No signal	No signal
Tc014A10	Unknown	Ubiquitous - all stages	Ubiquitous
Tc014B02	Unknown	No signal	No signal
Tc014B06	Unknown	No signal	No signal
Tc014B08	No hits	Head appendages, last abdominal segment	Appendages
Tc014B10	Unknown	Ubiquitous - all stages	Ubiquitous
Tc014B11	Unknown	Ubiquitous - all stages	Ubiquitous
Tc014C02	No hits	Ubiquitous - all stages	Ubiquitous
Tc014C09	No hits	Ubiquitous - all stages	Ubiquitous
Tc014F01	No hits	Ubiquitous - all stages	Ubiquitous
Tc014G01	Transcription factor	Midline expression in early germ band, in fully extended germ band expression ubiquitous excluding the ventral midline. Weak neurogenesis-like pattern.	Neurogenesis, other
Tc014G02	No hits	Neurogenesis	Neurogenesis
Tc014G04	Unknown	Ubiquitous - all stages	Ubiquitous

Clone	General function	In situ pattern description	Category
Tc014G06	Signal transduction	Ubiquitous - all stages	Ubiquitous
Tc014G11	No hits	Ubiquitous - all stages	Ubiquitous
Tc014H09	No hits	No signal	No signal
Tc015A02	Unknown	Ubiquitous - all stages	Ubiquitous
Tc015B03	No hits	Ubiquitous - all stages	Ubiquitous
Tc015C01	No hits	Ubiquitous - all stages	Ubiquitous
Tc015C05	Unknown	Ubiquitous - all stages	Ubiquitous
Tc015C08	No hits	Ubiquitous - all stages	Ubiquitous
Tc015D10	No hits	Ubiquitous - all stages, lateral segmental spots (tracheal openings?)	Ubiquitous, other
Tc015E12	No hits	Ubiquitous - all stages	Ubiquitous
Tc015F07	No hits	Ubiquitous - all stages	Ubiquitous
Tc015F08	Unknown	Ubiquitous - all stages	Ubiquitous
Tc015F11	Signal transduction	Ubiquitous - all stages	Ubiquitous
Tc015F12	No hits	Ubiquitous - all stages	Ubiquitous
Tc016A02	No hits	n/a	n/a
Tc016B02	Unknown	Ubiquitous - all stages	Ubiquitous
Tc016B10	Transcription factor (possible)	Ubiquitous - all stages	Ubiquitous
Tc016C02	Translation factor	Ubiquitous - all stages	Ubiquitous
Tc016C05	Unknown	n/a	n/a
Tc016C05	Unknown	Ubiquitous - all stages, expression stronger in appendages	Ubiquitous, appendages
Tc016D02	Unknown	Ubiquitous - all stages	Ubiquitous
Tc016F11	No hits	Ubiquitous - all stages	Ubiquitous
Tc016G10	No hits	Ubiquitous - all stages	Ubiquitous
Tc016G12	Unknown	Ubiquitous - all stages	Ubiquitous
Tc016H10	Unknown	Ubiquitous - all stages	Ubiquitous
Tc020A03	No hits	Ubiquitous - all stages	Ubiquitous
Tc020A11	LIM transcription factor	Ubiquitous - all stages	Ubiquitous
Tc020B07	No hits	Amnion	Extra-embryonic tissues
Tc020B11	E2f/TDP transcription factor	Pattern?	Other
Tc020C02	No hits	Ubiquitous - early and intermediate stages	Ubiquitous

Clone	General function	In situ pattern description	Category
Tc020C05	No hits	No signal	No signal
Tc020D02	Unknown	Ubiquitous - all stages	Ubiquitous
Tc020D08	Unknown	Pattern?	Other
Tc020E10	BTB/POZ, Zf-C2H2 transcription factor	Ubiquitous - all stages	Ubiquitous
Tc020G03	No hits	Ubiquitous - all stages	Ubiquitous
Tc020G06	No hits	No signal	No signal
Tc020H03	Unknown	Ubiquitous - all stages	Ubiquitous
Tc020H12	Unknown	Ubiquitous - all stages	Ubiquitous
Tc021A12	Signal transduction	No signal	No signal
Tc021C10	No hits	No signal	No signal
Tc021C11	No hits	Ubiquitous - all stages	Ubiquitous
Tc021D04	No hits	No signal	No signal
Tc021D12	Unknown	Anterior marker in early stages (blastoderm), segment polarity-like expression	Segmentation, other
Tc021E04	Unknown	Ubiquitous - all stages, stronger expression in the ventral midline	Ubiquitous, other
Tc021E05	No hits	Ubiquitous - all stages	Ubiquitous
Tc021E07	No hits	Granulated (spotted) expression in late embryos in abdomen, inside appendages and in tracheal openings (?)	Other
Tc021G03	Zf-C3HC4 transcription factor	Ventral midline	Other
Tc021G09	Unknown	Ubiquitous - all stages	Ubiquitous
Tc021H04	No hits	n/a	n/a
Tc022B12	Bicaudal-like transcription factor	Ubiquitous - all stages, stronger expression in muscles	Ubiquitous, myogenesis
Tc022F05	Unknown	Ubiquitous - all stages	Ubiquitous
Tc022G03	Unknown	No signal	No signal
Tc022G12	No hits	No signal	No signal
Tc023A01	No hits	Ubiquitous - all stages	Ubiquitous
Tc023B03	Zf-C2H2 transcription factor	No signal	No signal
Tc023C02	Unknown	Serosa, ubiquitous - all stages	Extra-embryonic tissues, ubiquitous
Tc023C04	Unknown	Ubiquitous - all stages	Ubiquitous
Tc023C07	No hits	Ubiquitous - all stages	Ubiquitous
Tc023C11	Unknown	Ubiquitous - all stages	Ubiquitous

Clone	General function	In situ pattern description	Category
Tc023D01	No hits	Ubiquitous - all stages	Ubiquitous
Tc023D06	No hits	No signal	No signal
Tc023E02	Unknown	Ubiquitous - all stages	Ubiquitous
Tc023E09	Unknown	No signal	No signal
Tc023F08	Unknown	No signal	No signal
Tc023F10	Unknown	Ubiquitous - all stages	Ubiquitous
Tc023G02	Unknown	No signal	No signal
Tc023H02	No hits	Ubiquitous - all stages	Ubiquitous
Tc023H11	Unknown	Ubiquitous - all stages	Ubiquitous
Tc024B02	No hits	No signal	No signal
Tc024C03	Unknown	No signal	No signal
Tc024C05	Unknown	No signal	No signal
Tc024D02	Unknown	No signal	No signal
Tc024E11	Unknown	No signal	No signal
Tc024G10	Unknown	Ubiquitous - all stages	Ubiquitous
Tc025A01	No hits	Ubiquitous - all stages	Ubiquitous
Tc025B01	Transcription factor	Ubiquitous - all stages	Ubiquitous
Tc025B08	No hits	Ubiquitous - all stages	Ubiquitous
Tc025C01	No hits	Anterior part of the appendages	Appendages
Tc025C10	Zf-C2H2 transcription factor	Ubiquitous - all stages	Ubiquitous
Tc025D01	No hits	n/a	n/a
Tc025D09	Unknown	No signal	No signal
Tc025E09	No hits	Ubiquitous - all stages	Ubiquitous
Tc025E10	No hits	No signal	No signal
Tc025F01	No hits	Ubiquitous - all stages	Ubiquitous
Tc025F10	No hits	No signal	No signal
Tc025G03	No hits	Ubiquitous - all stages	Ubiquitous
Tc025G07	bZIP transcription factor	No signal	No signal
Tc025H12	Ets-domain transcription factor	Neurogenesis	Neurogenesis
Tc026A08	Unknown	No signal	No signal

Clone	General function	In situ pattern description	Category
Tc026A10	No hits	Ubiquitous - all stages	Ubiquitous
Tc026B05	Unknown	Ubiquitous - all stages	Ubiquitous
Tc026B12	Translation factor	Ubiquitous - all stages	Ubiquitous
Tc026C02	Unknown	n/a	n/a
Tc026C03	No hits	n/a	n/a
Tc026C11	Unknown	Ubiquitous - all stages, granulated in late embryos, neurogenesis-like pattern in late embryos	Ubiquitous, neurogenesis, other
Tc026D03	No hits	No signal	No signal
Tc026D10	No hits	n/a	n/a
Tc026D12	Unknown	Serosa	Extra-embryonic tissues
Tc026G06	Signal transduction	n/a	n/a
Tc026G10	No hits	n/a	n/a
Tc026H02	Transcription factor	Ubiquitous - all stages	Ubiquitous
Tc026H09	Unknown	No signal	No signal
Tc027A12	Signal transduction	Ubiquitous - all stages	Ubiquitous
Tc027B02	Transcription factor	No signal	No signal
Tc027B08	Unknown	No signal	No signal
Tc027C03	No hits	Ubiquitous - early stages, ventral midline in elongating germ band then only "dorsal ectoderm" stained in late embryos, rings in legs	Ubiquitous, appendages, other
Tc027D10	Unknown	No signal	No signal
Tc027G01	Unknown	No signal	No signal
Tc027H03	No hits	Neurogenesis	Neurogenesis
Tcex001	No hits	No signal	No signal
Tcex002	No hits	Ubiquitous - all stages	Ubiquitous
Tcex003	No hits	Segment polarity pattern	Segmentation
Tcex004	bHLH transcription factor	n/a	n/a
Tcex005	bHLH transcription factor	n/a	n/a
Tcex007	bHLH transcription factor	n/a	n/a
Tcex008	bHLH transcription factor	Ubiquitous - all stages	Ubiquitous
Tcex010	bHLH transcription factor	No signal	No signal
Tcex011	bHLH transcription factor	No signal	No signal

Clone	General function	In situ pattern description	Category
Tcex014	Homeobox transcription factor	Ubiquitous - all stages	Ubiquitous
Tcex019	Homeobox transcription factor	Basal ring in appendages, neurogenesis, lateral segmental spots, rings in legs	Appendages, neurogenesis, other
Tcex027	No hits	No signal	No signal
Tcex033	Zf-C2H2 transcription factor	Ubiquitous - early stages	Ubiquitous
Tcex035	Zf-C2H2 transcription factor	n/a	n/a
Tcex036	Zf-C2H2 transcription factor	n/a	n/a
Tcex038	Zf-C2H2 transcription factor	Neurogenesis	Neurogenesis
Tcex039	Zf-C2H2 transcription factor	n/a	n/a
Tcex040	Zf-C2H2 transcription factor	n/a	n/a
Tcex043	Zf-C2H2 transcription factor	No signal	No signal
Tcex044	Zf-C2H2 transcription factor	n/a	n/a
Tcex045	Zf-C2H2 transcription factor	Head lobes, tail bud, weak segmental stripes	Segmentation, other
Tcex046	Zf-C2H2 transcription factor	n/a	n/a
Tcex047	No hits	n/a	n/a
Tcex048	Zf-C2H2 transcription factor	n/a	n/a
Tcex049	No hits	n/a	n/a
Tcex050	No hits	No signal	No signal
Tcex051	Zf-C2H2 transcription factor	Ubiquitous - early stages, tail bud, neurogenesis	Ubiquitous, neurogenesis, other
Tcex053	Zf-C2H2 transcription factor	n/a	n/a
Tcex054	Zf-C2H2 transcription factor	No signal	No signal
Tcex055	No hits	n/a	n/a
Tcex056	Zf-C2H2 transcription factor	Ubiquitous - all stages	Ubiquitous
Tcex057	Zf-C2H2 transcription factor	n/a	n/a
Tcex059	Zf-C2H2 transcription factor	Ubiquitous - early stages	Ubiquitous
Tcex060	No hits	No signal	No signal
Tcex061	Zf-C2H2 transcription factor	No signal	No signal
Tcex062	Zf-C2H2 transcription factor	n/a	n/a
Tcex063	Zf-C3HC4 transcription factor	No signal	No signal
Tcex065	Zf-C3HC4 transcription factor	n/a	n/a

Clone	General function	In situ pattern description	Category
Tcex067	No hits	n/a	n/a
Tcex068	bZIP transcription factor	n/a	n/a
Tcex069	No hits	n/a	n/a
Tcex070	bZIP transcription factor	n/a	n/a
Tcex071	bZIP transcription factor	n/a	n/a
Tcex072	Myb-like protein	Ubiquitous - all stages	Ubiquitous
Tcex073	Myb-like protein	n/a	n/a
Tcex076	Cold-shock transcription factor	n/a	n/a
Tcex077	Cold-shock transcription factor	n/a	n/a
Tcex078	Cold-shock transcription factor	n/a	n/a
Tcex084	LIM transcription factor	n/a	n/a
Tcex085	LIM transcription factor	No signal	No signal
Tcex086	LIM transcription factor	n/a	n/a
Tcex088	LIM transcription factor	Myogenesis	Myogenesis
Tcex089	No hits	n/a	n/a
Tcex090	LIM transcription factor	No signal	No signal
Tcex091	No hits	No signal	No signal
Tcex092	LIM transcription factor	n/a	n/a
Tcex093	No hits	n/a	n/a
Tcex094	LIM transcription factor	Myogenesis	Myogenesis
Tcex095	LIM transcription factor	n/a	n/a
Tcex096	LIM transcription factor	No signal	No signal
Tcex097	LIM transcription factor	n/a	n/a
Tcex098	LIM transcription factor	n/a	n/a
Tcex099	LIM transcription factor	n/a	n/a
Tcex100	LIM transcription factor	n/a	n/a
Tcex101	No hits	n/a	n/a
Tcex103	SAM transcription factor	n/a	n/a
Tcex104	SAM transcription factor	n/a	n/a
Tcex105	SAM transcription factor	n/a	n/a

Clone	General function	In situ pattern description	Category
Tcex106	SAM transcription factor	n/a	n/a
Tcex107	PHD transcription factor	n/a	n/a
Tcex108	Zf-CCCH transcritpion factor	n/a	n/a
Tcex109	No hits	No signal	No signal
Tcex110	BTB/POZ transcription factor	n/a	n/a
Tcex111	BTB/POZ transcription factor	No signal	No signal
Tcex112	BTB/POZ transcription factor	n/a	n/a
Tcex113	BTB/POZ transcription factor	No signal	No signal
Tcex114	BTB/POZ transcription factor	No signal	No signal
Tcex115	SAND transcription factor	Ubiquitous - all stages	Ubiquitous
Tcex116	SNF2 transcription factor	n/a	n/a
Tcex117	No hits	n/a	n/a
Tcex118	Enhancer of zeste-like transcription factor	Ubiquitous - all stages (?)	Ubiquitous
Tcex119	Enhancer of zeste-like transcription factor	n/a	n/a
Tcex121	Zf-AN1 transcription factor (possible)	No signal	No signal
Tcex122	Bromo-domain transcription factor	Ubiquitous - all stages	Ubiquitous
Tcex123	Bicaudal-like transcription factor	No signal	No signal
Tcex124	TEA transcription factor	n/a	n/a

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."

Köln, den 21.11.2003 Joël Savard

Teilpublikationen:

Keine

Lebenslauf

Name	Joël Savard
Geburtsdatum/ort	09.02.1973 in Hauterive, Kanada
Nationalität	kanadisch
Familienstand	ledig

Schulbildung:

1979-1985	Grundschule Laflamme, Saint-Jean-sur-Richelieu, Kanada
1985-1990	Gymnasium Marcel-Landry, Saint-Jean-sur-Richelieu, Kanada
1990-1992	Collège St-Jean, Saint-Jean-sur-Richelieu, Kanada
	Abschluss Juni 1992: Collège Abschluss der Wissenschaft

Studium:

1992-1996	Sherbrooke Universität, Sherbrooke, Kanada
	Abschluss 26. April 1996: Bachelor of Science (B.Sc.) in Biologie und Biotechnologie
1997-2000	McGill Universität, Montréal, Kanada
	Abschluss 8. Juni 2000: Master of Science (M.Sc.) in Biologie
seit 2000	Universität zu Köln, Köln, Deutschland Doktorarbeit am Institut für Genetik in der Abteilung für Evolutionsgenetik unter der Leitung von Prof. Dr. Diethard Tautz