Facultat de Ciències

Memòria del Treball Final de Grau

Títol del treball:

The role of Transposable Elements in regulatory networks in response to stress

Estudiant:	Quirze Rovira Castellà	Grau en Biologia
Correu electrònic:	u1912042@campus.udg.edu / quirze92@gmail.com	

Tutor:	Dr. Manuel Vera	(manuel.verar@udg.edu)
Cotutor*:	Dr. Josefa González	(josefa.gonzalez@ibe.upf-csic.es)
Empresa / institució:	Institut de Biologia Evolutiva - CSIC - UPF	

Vistiplau tutor (i cotutor*):

Nom del tutor: Manuel Vera
Nom del cotutor*: Josefa González

Empresa / institució:
Institut de Biologia Evolutiva - CSIC - UPF

Correu(s) electrònic(s): manuel.verar@udg.edu
josefa.gonzalez@ibe.upf-csic.es

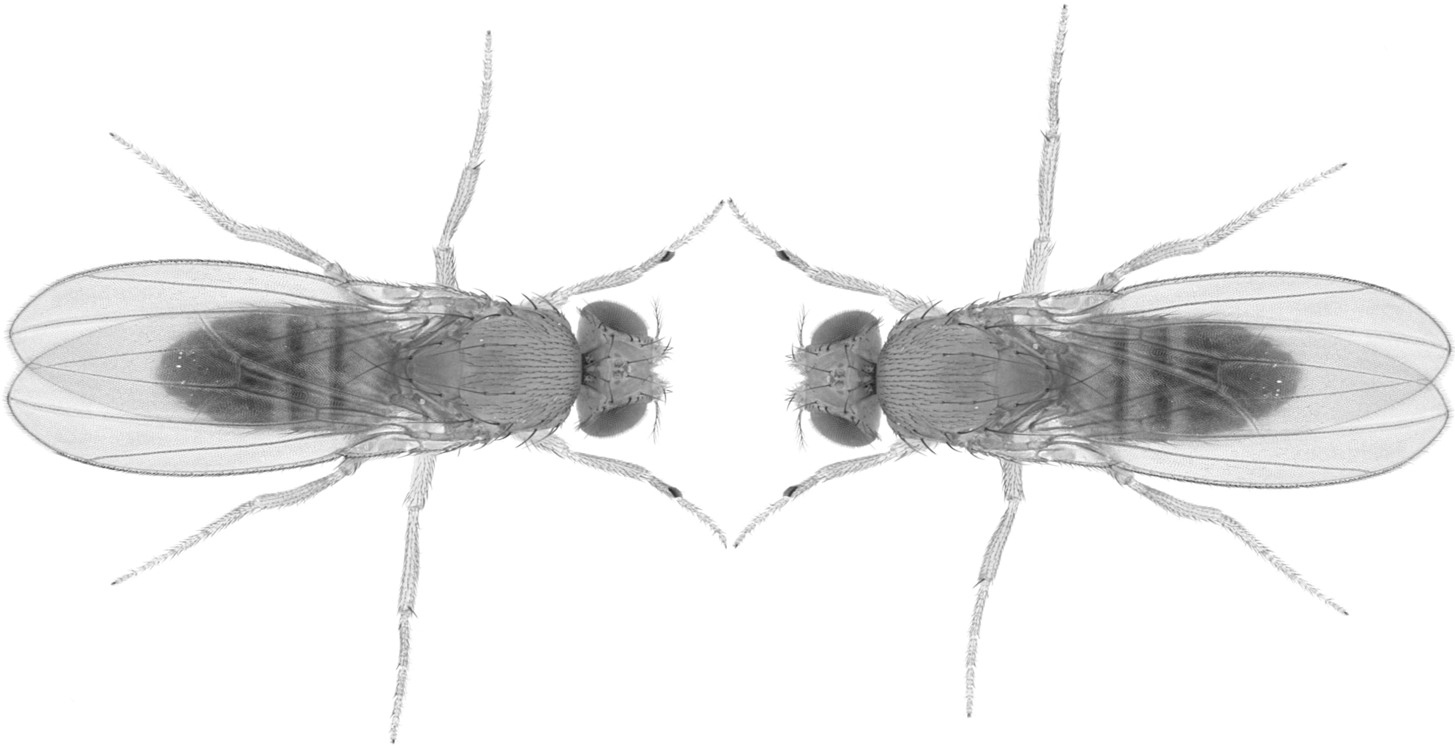
*si hi ha un cotutor assignat

Data de dipòsit de la memòria a secretaria de coordinació: 27/05/2015

Universitat de Girona

Treball de Final de Grau

The role of Transposable Elements in regulatory networks in response to stress



Quirze Rovira Castellà
Universitat de Girona
Grau en Biologia
Treball de Final de Grau
u1912042@campus.udg.edu

Tutor acadèmic: **Dr. Manuel Vera**
Correu electrònic: manuel.verar@udg.edu
Institució:
Institut de Biologia Evolutiva - CSIC - UPF
Tutora empresa: **Dr. Josefa González**
Correu electrònic: josefa.gonzalez@ibe.upf-csic.es

INDEX

ABSTRACT

The role of Transposable Elements in regulatory networks in response to stress.
Quirze Rovira

Transposable elements (TEs) are mobile DNA sequences that have the ability to jump from one location of the genome to another and create copies of them during the process. Although their activity can create mostly harmful mutation, it also represent an important source of genetic variation that have been shown to rewire and fine-tune regulatory networks. Stress response mechanisms are highly conserved across organisms and a major target of natural selection. Despite the latest advances in evolutionary biology, adaptation process remains poorly understood. Therefore, studying stress pathways provide a great opportunity to study how adaptive processes occur. In this study we aim to identify genome-wide putatively adaptive TEs that are candidates to modulate stress response by the addition of stress response elements (SREs) in the promoter of *Drosophila melanogaster* genes. We combine bioinformatics and population genetics approaches to discard SREs found by chance and select TEs having SREs that are more likely to be involved in stress response. We found 3,381 TEs out of 5,416 annotated TEs that contain at least one SRE. Because only TEs inserted in the promoter region are likely to have a biological impact in the gene modulation we reduce our dataset to 406 TEs. Moreover, we select TEs present at high frequency in the populations since they are more likely to be adaptive. Finally, our curated dataset contain 108 TEs with at least one SRE, inserted at a promoter region of a gene and at high frequency in at least one of the 8 worldwide studied populations. Out of this list we *in silico* validate 18 TEs thanks to modENCODE project and other chromatin immune-precipitation studies. Additionally, we analyzed the promoter of the 127 genes nearby the 108 TEs. The function of 22 out of this list was related to stress by genome wide association studies and gene ontology term annotations. Eighty-five TEs add preexisting or novel SREs while other 26 introduce SREs to the regions without any previous SREs. Having several SREs make the promoter region and/or the TE more likely to have an impact on the gene expression. Seventy-three TEs out of the 108 TEs contain more than one SRE and hence are more likely to be involved in stress response regulation. Our final candidate list is the starting point to functionally analyze the role of these putatively adaptive TEs in stress response.

El paper dels Elements Transposables a les xarxes reguladores de resposta a estrès. Quirze Rovira

Els Elements transposables (TEs) són seqüències de DNA mòbils capaços de saltar i crear còpies de si mateixes en el genoma. Encara que la seva activitat genera mutacions perjudicials, s'ha demostrat que també representen una font important de variació genètica amb capacitat de modular xarxes reguladores. Els mecanismes de resposta a estrès estan altament conservats entre organismes i són una diana important per a la selecció natural. Tot i els últims avenços en biologia evolutiva, els processos adaptatius segueixen sent poc coneguts i estudiar les vies d'estrès és una bona oportunitat per comprendre'ls. Aquest estudi pretén identificar els TEs putativament adaptatius del genoma, candidats a modular la resposta a estrès afegint elements de resposta a estrès (SREs) en el promotor dels gens de *Drosophila melanogaster*. Combinem eines bioinformàtiques i de genètica de poblacions per descartar SREs espuris, trobats per atzar, i seleccionar els TEs amb SREs més probables a estar involucrats en la resposta a estrès. Trobem 3,381 TEs dels 5,416 TEs anotats que contenen almenys un SRE. A causa que és més probable que els TEs inserits en el promotor tinguin un impacte en la regulació d'un gen, vam poder reduir els nostres TEs candidats a 406. Seleccionem els TEs presents en alta freqüència, en almenys una de les 8 poblacions estudiades en tot el món, ja que és més probable que siguin adaptatius. Finalment, tenim 108 TEs amb almenys un SRE inserit en el promotor d'un gen i a alta freqüència en almenys una de les poblacions estudiades. Validem *in silico* 18 TEs gràcies al projecte modENCODE i a altres estudis de immuno-precipitació de la cromatina. També analitzem el promotor dels 127 gens propers als 108 TEs. La funció de 22 d'aquests gens ha estat relacionada amb la resposta a estrès mitjançant estudis d'associació i les anotacions de Gene Ontology. Vuitanta-cinc TEs afegeixen SREs nous o preexistents, mentre que 26 introdueixen SREs en regions sense SREs previs. Tenir diversos SREs fan que la regió promotora i / o el TE tinguin major probabilitat a tenir un impacte en l'expressió gènica. Setanta-tres dels 108 TEs contenen més d'un SRE i per tant tenen major probabilitat d'estar involucrats en la regulació de la resposta a estrès. La nostra llista de candidats final és el punt de partida per analitzar funcionalment el paper d'aquests TEs putativament adaptatius en la resposta a estrès.

El papel de los Elementos Transponibles en las redes reguladoras de respuesta a estrés. Quirze Rovira

Los Elementos Transponibles (TEs) son secuencias de DNA móviles capaces de saltar y crear copias de sí mismas en el genoma. Aunque su actividad genera mutaciones perjudicial, se ha demostrado que también representan una fuente importante de variación genética con capacidad de modular redes reguladoras. Los mecanismos de respuesta a estrés están altamente conservados entre organismos y son una diana importante para la selección natural. A pesar de los últimos avances en biología evolutiva, los procesos adaptativos siguen siendo poco conocidos y estudiar las vías de estrés es una buena oportunidad para comprenderlos. Este estudio pretende identificar los TEs putativamente adaptativos del genoma, candidatos a modular la respuesta a estrés añadiendo elementos de respuesta a estrés (SREs) en el promotor de los genes de *Drosophila melanogaster*. Combinamos herramientas bioinformáticas y de genética de poblaciones para descartar SREs espurios, encontrados por azar, y seleccionar los TEs con SREs más probables a estar involucrados en la respuesta al estrés. Encontramos 3381 TEs de los 5416 TEs anotados que contienen al menos un SRE. Debido a que es más probable que los TEs insertados en el promotor tengan un impacto en la regulación de un gen, pudimos reducir nuestros TEs candidatos a 406. Seleccionamos los TEs presentes en alta frecuencia, en al menos una de las 8 poblaciones estudiadas en todo el mundo, ya que es más probable que sean adaptativos. Tenemos 108 TEs con al menos un SRE insertado en el promotor de un gen y a alta frecuencia en al menos una de las poblaciones estudiadas. Validamos *in silico* 18 TEs gracias al proyecto modENCODE y a otros estudios de inmuno-precipitación de la cromatina. También analizamos el promotor de los 127 genes cercanos a los 108 TEs. La función de 22 de estos genes ha sido relacionada con la respuesta a estrés mediante estudios de asociación y las anotaciones de Gene Ontology. Ochenta y cinco TEs añaden SREs nuevos o preexistentes, mientras que 26 introducen SREs en regiones sin SREs previos. Tener varios SREs hacen que la región promotora y/o el TE tengan mayor probabilidad a tener un impacto en la expresión génica. Setenta y tres de los 108 TEs contienen más de un SRE y por tanto tienen mayor probabilidad de estar involucrados en la regulación de la respuesta a estrés. Nuestra lista de candidatos final es el punto de partida para analizar funcionalmente el papel de estos TEs putativamente adaptativos en la respuesta a estrés.

INTRODUCTION

Evolutionary biology aims at comprehending the mechanisms that produced the diversity of life and the origin of new species (Futuyma 2013). The availability of new techniques and the latest technology advances such as whole genome sequencing has revolutionized our capacity to study the genetic basis of differences among and within species (Nadeau and Jiggins 2010). Despite recent progress, the mechanisms underlying adaptation are still poorly understood.

Gonzalez Lab aims at understanding how organisms adapt to their environment. Adaptation is the basis for highly important processes such as the survival of organisms to changes in the environment, host-pathogen interactions and resistance to pesticides, antibiotics and cancer chemotherapies (Daub et al. 2013, Tiffin and Ross-Ibarra 2014). Therefore, a better understanding of the adaptation process can provide useful knowledge relevant to other fields beyond evolutionary biology.

Transposable elements (TE) are an exceptional tool to study adaption. Their ability to replicate and insert in the genome produce mostly harmful mutations, for example by disrupting a coding region of a housekeeping gene or by increasing the chance of ectopic recombination leading to duplications or deletions (Kidwell and Lisch 2001). However, a proportion of the mutations induced by TEs are expected to be adaptive (Feschotte 2008, Rebollo et al. 2012). Our goal is to find those TE that have an adaptive role using *Drosophila melanogaster* as a model organism.

D. melanogaster is an exceptional model organism to study adaptation. It is a relative young species, in the evolutionary timescale, which has its origin in Africa and recently has spread all over the world (David and Capy 1988). This particular case, similar with human evolutionary history, is highly interesting to study how organisms have adapted to different environments. We expect that adaptive mutations that help this species to adapt to the new environments are present high frequency (Sundaram et al. 2014).

There are already well-known cases of adaptive TEs identified in the Gonzalez Lab: one adaptive TE was found to mediate resistance to oxidative stress (Guio et al. 2014),

another TE mediates resistance to xenobiotics (Mateo et al. 2014), and a third TE is associated with shorter developmental time (Ullastres et al. 2015).
Besides these three examples, Daborn et al. (2002) and Aminetzach et al. (2005) reported adaptive TEs involved in insecticide resistance also in *D. melanogaster*.

Interestingly, four of the five previous described adaptive TEs in *D. melanogaster* are involved in stress response. Stress is a primary aspect of life and a major aspect of natural selection. Stress-defense systems are found in any cell of the organisms and it is widespread form prokaryotes to eukaryotes. This wide range indicates that defense mechanisms were present in a very early time at the evolutionary tree (Korsloot et al. 2004). Surprisingly similitude at the biochemical level through different species and different environments are found for stress response mechanisms. Thus, it is not rare to found shared routes through different stress response pathways.
This study aims at identifying a catalogue of putatively adaptive TEs that add transcription factor binding sites (TFBS) related to stress response, called stress response elements (SREs), in the upstream regions of *D. melanogaster* genes. We have focused on five different stress-responses that are highly conserved from *Drosophila* to mammals (Korsloot et al. 2004): heat shock stress, hypoxia, oxidative stress, xenobiotic stress and heavy metal stress.
Heat shock is studied by exposing flies to a supraoptimal temperature for a period of time. It induces the action of heat shock proteins, which involve four aspects: (i) stabilizing essential structural proteins, (ii) assisting transfer of proteins across membranes, (iii) supporting refolding of proteins denaturated by stress, and (iv) supporting the degradation of aberrant proteins. Heat shock proteins are involved not only in heat shock stress but also in general stress response (Feder and Hofmann 1999). Some proteins of this family are highly inducible in stress conditions. Heat shock elements (HSE) are binding sites were transcription factor (TF) related in heat shock (HSF) recognize and promotes gene expression of nearby genes (Santoro 2000, Enoki and Sakurai 2011).
Oxidative stress involves many phenomena that are unavoidable consequences of aerobic metabolism. Reactive oxygen species (ROS) are free radical molecules generated by cellular processes that can damage many macromolecules in the cell such as proteins, DNA, and lipids. Protection against ROS is evolutionarily necessary and antioxidants systems have been develop to deal with this issue (Sykiotis and Bohmann

2010, Taguchi et al. 2011). The activity of antioxidant related genes are modulated by TF activity (Nguyen et al. 2003). Those genes present antioxidant response elements (ARE) sites were specific antioxidant TF can bind.

Xenobiotic stress is involved in the degradation of many foreign molecules and includes resistance to insecticide. It is closely related to antioxidant routes since many xenobiotic molecules create ROS in their metabolic pathway (Vorrink and Domann 2014). Similar to other stress-response pathways, xenobiotic related genes appear to be regulated by xenobiotic response factor (XRF), which binds short DNA sequences called xenobiotic response elements (XRE) (Abel and Haarmann-Stemmann 2010).

Metallothionein protein is the principal component of metal stress response system. This protein has high affinity for free metal ions (Zn^{2+}, Cd^{2+} and Cu^{+}) reducing their concentrations. This protein contains binding sites for metal transcription factor (MTF), which is called metal response elements (MRE), in the promoter region but ARE are also present (Egli et al. 2003, Balamurugan et al. 2004, Bahadorani et al. 2010).

Hypoxia condition is also a widely conserved stress response. Hypoxia-inducible factor (HIF) mediates hypoxia response by binding to hypoxia response elements (HRE). This suggests again a conserved evolution for stress response (Zhou and Haddad 2013).

In order to identify TEs that add SREs to the promoter regions of *D. melanogaster* genes, we performed bioinformatics prediction of SREs for all 5,416 TEs annotated in the genome of this species. Furthermore, we used population genetics data and *ad hoc* scripts filters to select the TEs that are more likely to be functional. The result of this study is a dataset of 108 TEs putative to be involved in stress response.

OBJECTIVES

The goal of this study is to identify a catalogue of putatively adaptive TEs that add SREs in the promoter regions of *D. melanogaster* genes. Also, we aim to determine which of the selected TEs are more likely to be adaptive in order to experimentally validate them in the future.

MATERIAL AND METHODS

Databases

We have used two transcription factor databases, Transfac and Jaspar. Transfac Public database contains experimentally validated positional weight matrices (PWM) for many transcription factor proteins in eukaryote organisms (Matys et al. 2006). Jaspar core database contains a curated, non-redundant set of PWM from published articles, that is, this database is constantly updated. By non-redundant, Jaspar means that there are not many models for the same factor. Thus, this database gives the "best" model for each transcription factor (Mathelier et al. 2014).

Selection of Stress Response Elements (SREs)

Stress response elements are evolutionary conserved motifs due to their biological function. For this reason, it is a great model to study how adaptation works across organisms.

We have selected 5 different SREs previously described in the Introduction section: Antioxidant response element (ARE), hypoxia response element (HRE), heat shock response element (HSE), xenobiotic response element (XRE) and metal response element (MRE).

When available, we used the PWM described in *Drosophila* (only for ARE and HSF). Otherwise we used vertebrate or human PWM. In total, we worked with 13 PWM (Table 1).

Table 1. Selected PWM for each SRE.

	Drosophila melanogaster		**Homo sapiens**		**Others**	
	Jaspar	**Transfac**	**Jaspar**	**Transfac**	**Jaspar**	**Transfac**
ARE	**MA0530.1**		MA0150.1		MA0150.2	
HRE			**MA0259.1**	**M00236** **M00539**	MA0259.1 (ver)	
HSE		**M00028** **M00163** **M00164** **M00165** **M00166**	MA0486.1	M00146 M00147	MA0319.1 (S.cer) PF0068.1 (Ver) PF0141.1 (Ver)	
XRE				**M00235** **M00237**	MA0006.1 (MM)	AHR_01 (MM)
MRE					**PB0044.1 (MM)** **PB0148.1 (MM)**	

Dataset

We analyzed the 5,416 TEs annotated in the release 6.02 of the *D. melanogaster* genome available at http://flybase.org.

From those 5,416 TEs our dataset is reduced to 1,636 TEs for which the frequency can be estimated. These 1,636 TEs do not contain: TEs from the INE-1 family, TEs inserted inside other TEs, and TEs inserted inside parts of segmental duplications (Fiston-Lavier et al. 2015). We obtained TE population frequency in 8 worldwide populations of *D. melanogaster*: one African population from Rwanda (DPGP2, Pool et al. 2012); two European populations sequenced in our lab from Italy and Sweden; and 5 populations from North America: North Carolina (DGRP, Huang et al. 2014), Florida, Georgia, South Carolina and Maine (Bergland et al. 2014).

SREs in TEs

In order to informatically predict the binding position for each TF, we used the TFBSTools software (Tan 2015). TFBSTools is an R package for the analysis and manipulation of transcription factor binding sites and transcription factor profile matrices. This program is specially designed for Jaspar matrices, however, we run it with the 13 selected matrices by manually introducing their parameters. The program output calculates the score parameter for each predicted hit by a pairwise alignment of each PWM. We used the relative score value parameter (between 0 and 1) with respect of the score range of the associated PWM. This parameter is being used to qualify predicted hits. TFBSTools default settings have a score threshold of 0.75.

Redundant hits

For some specific matrices, TFBSTools finds 2 hits for the same TF in the exactly same position but on the opposite strand. We have cleaned the output of the program by selecting the duplicated hits with the highest score value and eliminating the other hits.

Overlapping SREs

Our data showed that some SREs have overlapping coordinates. Biologically it is not possible that more than one TF binds to the same position. Thus, we create a script that joins overlapping hits into a single combine hit. We considered overlapping hits those that overlap at least 1bp.

Additionally, a single copy of Heat shock transcription factor is not able to bind to the DNA: the active binding form is a homotrimer of the protein. Four of the five selected PWM for HSF contain three repeated single HSE binding sites, while the other one (M00028) is a single repeat. However, Enoki and Sakurai (2011) have proved that two single HSE repeats separated by a 5bp gap can be recognized by HSF. For this reason, we created a script that removes the single hits of M00028, and it joins the hits separated by a 5bp gap into a single HSE hit.

Filters

Several filters were applied to the data with the purpose of selecting only those TEs more likely to be adding functional SREs.

Gene distance filter

Only TEs in the promoter region (1kb upstream) of a gene are likely to have a biological impact on that gene (Hoskins et al. 2011). Thus, we selected TEs that are located inside the 1kb upstream region of a gene or that overlap with this region.

Frequency filter

The majority of TE-induced mutations are deleterious and thus they are present at low population frequencies (Kidwell and Lisch 2001). However, adaptive TEs are expected to be present at high frequencies in the populations (Feschotte 2008, Rebollo et al. 2012). We thus focused on TEs that are present in at least 15% of the strains analyzed. Frequencies were obtained from 8 worldwide populations (see Dataset) (Fiston-Lavier et al. 2015). A TE was considered to be present at high frequency if its presence was high at least one of the populations. INE TE family was excluded from the study due to the particular dynamics of this family: all INE TEs are old insertions that are fixed in the populations (Yang et al. 2006).

SREs in promoter regions

For those TEs that add SREs to the promoter region of nearby genes and are present at 15% or higher frequency in populations, we look for SREs in the promoter region of the nearby gene. We have created a script that subtracts the 1,000bp upstream region for each one of these genes. For each promoter region sequence we look for SREs with TFBSTools software, we eliminate the repeated hits and we run the overlapping SREs script. We also create a script that removes the SREs hits coming from the TE inserted in the promoter region. Thus, we obtained the number of SREs added by the TE and the number of SREs present in the upstream region of the gene.

Experimental evidence for SREs predictions

We look for experimental evidence for SREs predicted hits by looking to modEncode ChIP-seq data. modENCODE only has for *cnc* and *h* the only transcription factors related in any of our studied stress cases. We also check other chromatin immune-precipitation (ChIP) publications for HSF and CBP proteins (Guertin and Lis 2010, Holmqvist and Mannervik 2013).

Genome Wide Association Study (GWAS) evidence for a role of the selected genes in stress response

To find additional evidence for the role of the genes nearby selected TEs in stress response, we looked for GWAS publications in which candidate genes for the five stress-response studied in this work were identified. We checked four GWAS studies related to oxidative stress (Zou et al. 2000, Girardot et al. 2004, Landis et al. 2004, Weber et al. 2012), two GWAS for heat stress response (Sorensen et al. 2005), one GWAS study for hypoxia response (Azad et al. 2009).

Gene Ontology (GO) biological process

Gene Ontology project provides controlled vocabularies of defined terms representing gene product properties. GO terms are divided in three main categories: cellular component, molecular function and biological process. We used the Batch downloaded tool from Flybase to retrieve the GO annotations for biological process of selected genes. Stress response related GO terms were selected.

RESULTS

We have used TFBSTools R package to look for SREs in the 5,416 TEs annotated in the *D. melanogaster* genome. We focus on five SREs that are highly conserved across organisms: Antioxidant response element (ARE), hypoxia response element (HRE), heat shock response element (HSE), xenobiotic response element (XRE) and metal response element (MRE). We have filtered the output of this tool to end up with a list of putatively adaptive TEs that add SREs to gene promoter regions. In the sections below, we described the results of each one of these steps.

Repeated hits filter

We have filtered the TFBSTools output in order to remove the redundant hit errors (see Material and Methods). The total number of hits predicted by TFBSTools for each PWM and the number of repeated hits for each PWM are shown on Table 2.

Table 2. Number of repeated hits for each PWM

	ARE	HRE			HSE		
	MA0530.1	**MA0259.1**	**M00236**	**M00539**	**M00028**	**M00163**	**M00164**
Total hits	300,686	157,589	55,034	13,648	596,004	29,926	25,325
Repeated hits	1,102	17,271	7,786	6,823	0	0	742
% of repeated hits	0.37	10.96	14.15	49.99	0	0	2.93

	HSE		XRE		MRE	
	M00165	**M00166**	**M00235**	**M00237**	**PB0044.1**	**PB0148.1**
Total hits	25,345	22,329	108,720	10,631	22,723	161,094
Repeated hits	0	295	208	0	32	416
% of repeated hits	0	1.32	0.19	0	0.14	0.26

MA0259.1, M00236 and M00539 had a repeat rate significantly higher than all the other SREs, suggesting HRE can be bound on both strands.

We tested that all repeated hits were eliminated by running the same script on the clean dataset. As expected, no duplicated hits were found.

Prediction of SREs in TEs, data filtering, and threshold selection

Five thousand one hundred sixty-two out of 5,416 TEs contain at least one predicted SREs by the default parameters of TFBSTools. Only TEs inserted in the promoter region of a gene are likely to have a biological impact on that gene (Hoskins et al. 2011). Thus, we selected from the TEs that do contain SREs the TEs that are located in gene promoter regions. We filtered out the TEs that were present at low frequency in populations (<15%) because we expect adaptive TEs to be present at high frequencies in the populations (Sundaram et al. 2014). We include a TE in our study if it is present at high frequencies in any of the 8 worldwide populations studied. Finally, we discarded elements from the INE family because the dynamic of this family is different from that of all the other families in the genome: all INE elements are old and fixed in populations (see Materials and Methods).

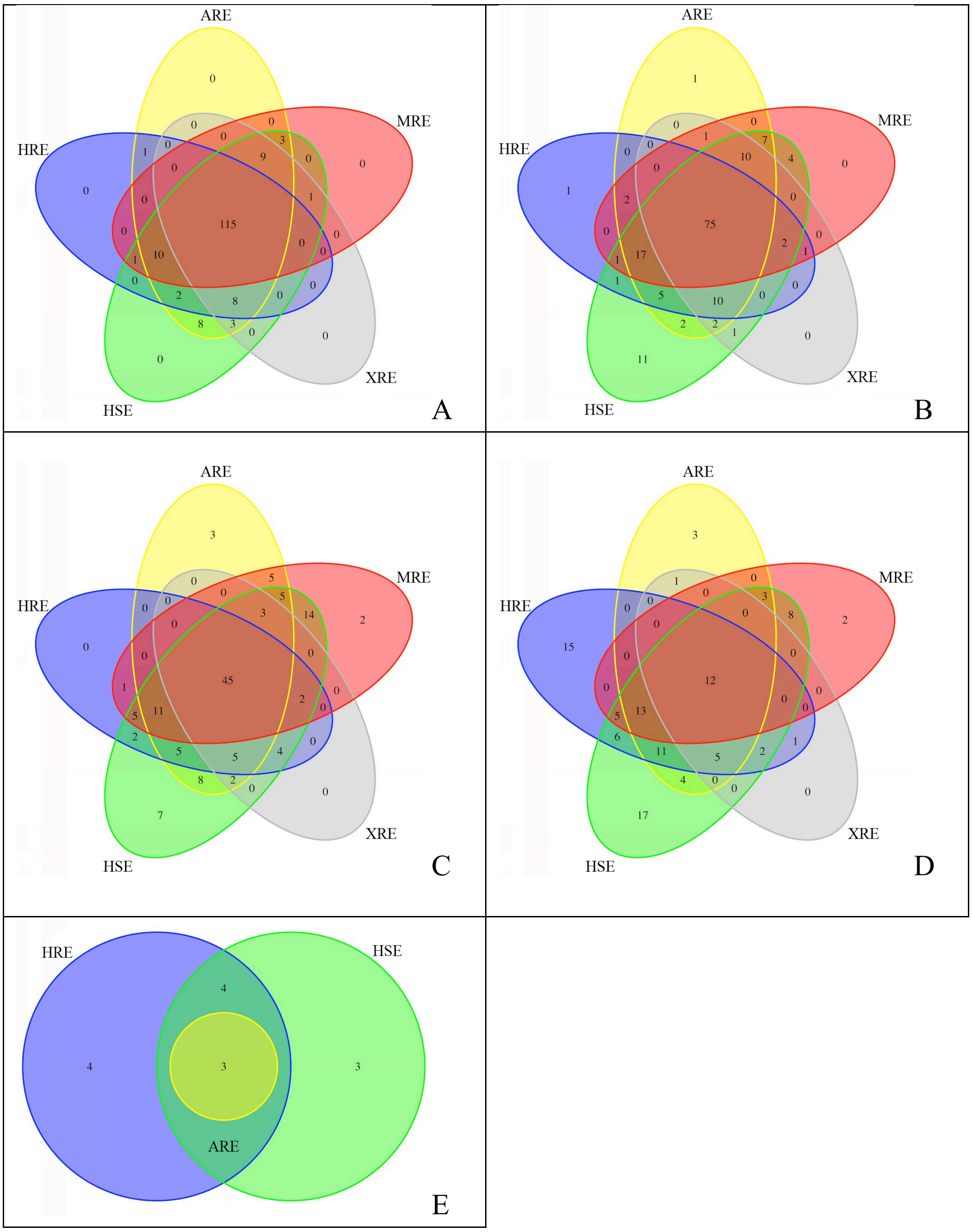
Once all the filters were applied the number of TEs was reduced from 2,131 to 14 and from 5,162 to 161 depending on the score (Table 3). As expected, the number of TEs that contain SREs decreased as the score value becomes higher.

Table 3. Number of TEs that accomplish the filtering criteria in each step of the process for various score values thresholds. (#: Number).

score	#TE with SREs	#TE close tan 1kb 5'	INE-1	#TE in high frequency
0.75	5,162	695	383	161
0.80	5,000	667	362	154
0.85	4,421	565	290	129
0.90	3,381	406	172	108
0.95	2,131	237	4	14

We plotted the number of TEs adding each one of the five SREs studied depending on the score used (Figure 1). For the 0.75 threshold (Figure 1A), we found that 71% of the TEs contain all five SREs. This percentage is smaller for the 0.80 (Figure 1B) and 0.85 (Figure 1C) score threshold (48.7% and 34.9% respectively) being the 0.90 threshold the one that shows a more drastic reduction: only 11% of the TEs add all SREs (Figure 1D). Because it is not likely that most of the TEs contain all five SREs, these results suggest that 0.90 is a good threshold to identify functional SREs.

Figure 1. Venn Diagrams representing number of TEs that contain each one of the five SREs for 0.75 (A), 0.80 (B), 0.85 (C), 0.90 (D) and 0.95 (D) score thresholds.



Bioinformatic prediction of SREs in promoter regions of genes close to selected TEs

We look for evidence of SREs in the promoter regions of genes located nearby our selected TEs. In some cases a TE can be in the promoter region of more than one gene. This case makes possible for a TE to affect the expression of more than one gene. Thus, we took into account all the genes that are downstream of a selected TE (Table 4).

Table 4. Number of putatively adaptive TEs and number of the corresponding nearby genes. (#: Number).

score	# TEs	# genes nearby TEs
0.75	161	192
0.80	154	184
0.85	129	156
0.90	108	127
0.95	14	15

We checked whether (i) the TE introduces new SREs in a promoter region where there were no SREs; (ii) the TE adds the same type of SREs that were already present in the promoter regions; and (iii) the TE adds different SREs in a promoter region that do not contain those particular SREs.

For score values less than 0.90, there are no TEs that add different SREs in a promoter region that do not contain those SREs (Table 5). Most of the TEs add the same SRE that was already present in the promoter region. Although with score 0.90, TEs adding the same SRE are still the more abundant, we also find TEs that add SREs to regions that did not have SREs or had different SREs. Thus we decided to focus on those TEs with a 0.90 threshold.

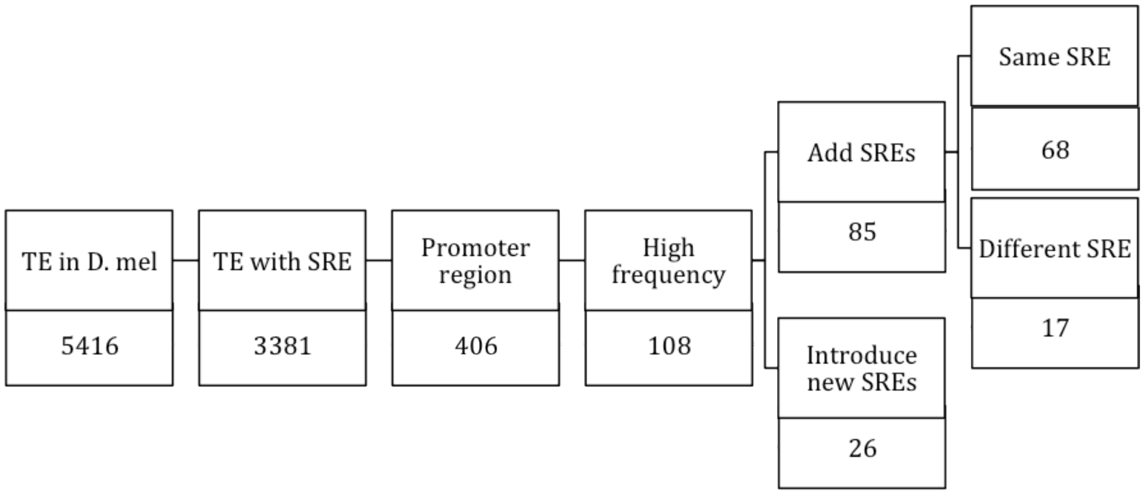
Table 5. Number of TEs for the three described scenarios: TEs that introduce new SREs in a promoter region where there were no SREs (**#TE introduce SRE**), TEs that add the same type of SREs that were already present in the promoter regions (**#TE add same SREs**) and TEs that add different SREs in a promoter region that do not contain those SREs (**#TE add different SREs)**.

score	# TEs introduce SREs	# TEs add same SREs	# TEs add different SREs
0.75	1	160	0
0.80	3	151	0
0.85	7	122	0
0.90	26	68	17
0.95	8	5	1

Note that the total number of TEs is higher that the total number of selected TEs because some TEs are close to several genes. (#: Number).

Overall, we end up with a dataset of 108 TEs that are likely to add SREs to the promoter region of their nearby genes (Figure 2). These 108 TEs included TEs that introduce SREs to the promoter regions of genes lacking SREs (26 TEs), and 85 TEs that add SREs to the promoter regions of genes that already had at least one SRE. Note that in the same region one TE can introduce new SREs to the promoter region of a gene and the same TE can add SREs to another gene. This happens in three cases in our dataset.

Figure 2. Summary of the filtering process used in our study including the numbers of TEs selected in each step for our working threshold of 0.90.



In silico **validation of SREs in our selected genes**

ChIP technique is used to study the interaction between proteins and DNA, and find specific regions were proteins are associated (Landt et al. 2012). This technique is combined with massive sequencing (ChIP-Seq) in order to perform a global map of protein binding sites. We thus looked for ChIP-Seq studies for our selected TFs. The modENCODE project validates experimentally the localization of TFBS for 38 functionally diverse TFs using ChIP experiments (Negre et al. 2011). Two of those TFs, *cnc* and *hairy*, are related to oxidative stress response and hypoxia response respectively. We check for evidence for *cnc* and *hairy* in 127 genes nearby our selected 108 TEs. Note that we can not validate SREs located inside a TE because modENCODE excludes repetitive sequences from their analyses.

Only one predicted ARE was validated by *cnc* ChIP-Seq data, specifically the same SRE studied by Guio et al. (2014). On the other hand, five *hairy* binding sites match with predicted sites in the promoter regions of our gene list, suggesting a hypoxia response function for those genes.

Besides the two TFs studied in the modENCODE project, ChIP-Seq data is also available for HSF. Using this data, we validated two HSE in promoter regions of the selected gene list (Guertin and Lis 2010).

Finally, we also look for ChIP-Seq studies for CBP/p300 protein, which is a general co-activator in metazoan: binding of this protein to the DNA is directly linked to other TF binding (Holmqvist and Mannervik 2013). We found additional ten genes in our list that have evidence of CBP/p300 binding on their promoter region.

GWAS and GO evidences for stress related biological function of selected genes

To check whether any of the 127 genes located downstream of the 108 selected TEs have been previously identified as candidate stress-response genes, we look for GWAS for stress response genes. In total we obtained GWAS evidences for 12 selected genes: 10 genes were previously identified as candidates for oxidative stress response (Zou et al. 2000, Girardot et al. 2004, Landis et al. 2004, Weber et al. 2012), 2 genes showed response in hypoxia conditions (Azad et al. 2009) and 2 genes showed response to heat shock conditions (Girardot et al. 2004, Sorensen et al. 2005). Note that the two heat shock responsive genes identified were also found in the oxidative stress response GWAS (Girardot et al. 2004).

Additionally, we have checked if any of our selected genes had a GO functional annotation involved in stress response using the Batch download tool from Flybase. We found 14 genes that had a stress related GO functional annotation in our selected gene list (Table 6).

Table 6. Putatively adaptive TEs with SREs identified in this work. An underbar between SREs represents combine regions of SREs. Table is separated by: TEs that add the same type of SREs that were already present in the promoter regions (A), TEs that add different SREs in a promoter region that do not contain those (B) and TEs that introduce new SREs in a promoter region where there were no SREs (C). ChIP-Seq validation (Guertin and Lis 2010, Negre et al. 2011, Holmqvist and Mannervik 2013), GWAS (Zou et al. 2000, Girardot et al. 2004, Landis et al. 2004, Sorensen et al. 2005, Azad et al. 2009, Weber et al. 2012).

Flybase TE identified	SREs inside TE	Gene	SREs in gene promoter	ChIP-Seq validation [1] | GWAS candidate [2] | GO stress response terms [3]
A - TEs adds same SREs in gene promoter region				
FBti0064246, *FBti0018868*, *FBti0062946*[1]	ARE	*CR44391*, *TM4SF*, *karr*[1]	ARE, HRE, HSE, MRE_MRE | ARE, HSE_HSE | HSE_HSE, ARE_HSE	CBP/p300
FBti0061018	ARE_HSE	*Noa36*	HRE, HSE_HSE	
FBti0019530	ARE, ARE_HSE	*CG3021*	ARE, HRE, HSE, HRE_HSE, HSE_MRE, HRE_HSE_XRE	
FBti0018999[3]	ARE ,HRE, ARE_HSE, HRE_XRE	*Gr43a*[3]	HRE, HSE, HRE_HSE	sensory perception of taste | cellular response to fructose stimulus | regulation of feeding behavior | adult feeding behavior | regulation of appetite | sensory perception of sweet taste
FBti0062853	ARE, HRE, HSE_HSE	*CR42546*	HRE, ARE_ARE, HSE_HSE	
FBti0018962	ARE, HRE, HSE_HSE, ARE_HSE	*CG3649*	HSE_HSE	
FBti0019167	ARE, HRE, HSE_HSE, ARE_HSE, HRE_HSE, HSE_MRE	*CG42809*	HRE, HSE_HSE, HRE_XRE	
FBti0020147	ARE, HRE, HSE_HSE, HRE_HSE	*Met75Cb*	HRE_HRE, ARE_HRE	
FBti0061736, *FBti0019309*[3]	ARE, HRE, HSE_HSE, HRE_XRE, HRE_MRE	*CR45923*, *Hpr1*[3]	HRE, HRE_HSE | ARE, HRE, XRE, HSE_MRE	mRNA export from nucleus | mRNA export from nucleus in response to heat stress | signal transduction
FBti0020010[1][3], *FBti0020111*[3]	ARE, HRE, HSE, HSE_HSE	*dpr20*[1][3], *CG13473*[3]	HRE, HSE, XRE, HSE_HSE | HRE_HRE, HSE_HSE	modENCODE *h* | sensory perception of chemical stimulus | cell redox homeostasis | glycerol ether metabolic process
FBti0019620	ARE, HRE, HSE, MRE, HRE_HRE, HSE_HSE, HRE_XRE, HSE_MRE	*CR45604*	HRE, HSE_MRE	
FBti0019068	ARE, HRE, HSE, MRE, HSE_HSE	*sl*	HSE_HSE	
FBti0019354	ARE, HRE, HSE, MRE, HSE_HSE, MRE_MRE, ARE_HSE, HSE_MRE	*CG14692*	ARE, HSE	
FBti0019210	ARE, HRE, HSE, XRE, MRE, HSE_HSE, ARE_HSE, HRE_HSE	*spir*	ARE, MRE, HSE_HSE	
FBti0019381[1]	ARE, HRE, MRE, HRE_HRE, HSE_HSE	*CG42788*[1]	ARE, HRE, HSE_HSE	CBP/p300
FBti0019359[1][2][3]	ARE, HRE, MRE, HSE_HSE	*Hsp70Aa*[2][3], *Hsp70Ab*[1][2][3]	ARE, HSE_HSE | ARE, HSE_HSE	HSF | hypoxia | heat shock-mediated polytene chromosome puffing | response to heat | response to hypoxia
FBti0019758, *FBti0019975*, *FBti0019750*	ARE, HRE, MRE, HSE_HSE, ARE_HSE, HRE_XRE, HSE_MRE	*CG14946*, *CG1850*, *CG15399*	HRE, MRE, HSE_HSE, HSE_MRE | HRE, HSE | HSE, MRE, HSE_HSE	
FBti0020015[2][3],	ARE, HRE, XRE, MRE, HSE_HSE,	*LysP*[2][3], *LysS*[2][3],	HRE, ARE_ARE | ARE | HRE, HSE	oxidative stress | antimicrobial humoral

FBti0019086[3]	MRE_MRE, HRE_XRE	*Obp18a*[3]		response | sensory perception of chemical stimulus
FBti0019782	ARE, HSE_HSE	*CG13970*	MRE, HSE_HSE	
FBti0020018	ARE, HSE_HSE, HRE_XRE	*stet*	HSE	
FBti0019012	ARE, HSE, HSE_HSE	*fab1*, *CG33981*	HSE | HSE	
FBti0018940, *FBti0063191*[2], *FBti0061734*, *FBti0062910*[1], *FBti0060088*	HRE	*tRNA:E4:56Fa*, *tRNA:E4:56Fb*, *tRNA:CR30220*, *Est-6*[2], *COX4L*, *Df31*[1], *CG32500*	ARE, HSE_HSE, HRE_XRE | ARE, HRE_XRE | ARE, HSE_HSE, HRE_XRE | HRE, XRE | HRE, HRE_HRE, HSE_HSE, HSE_MRE | HRE | HRE	CBP/p300 | oxidative stress | heat shock stress
FBti0060255	HRE_HRE	*CG11523*, *Csp*	HRE, HRE_XRE | HRE	
FBti0061318	HRE_HSE_XRE	*CG14613*	HRE_HRE, HSE_HSE, MRE_MRE, HSE_MRE, HRE_HSE_XRE	
FBti0062912	HRE, ARE_HSE, HRE_XRE	*CR43144*	HRE, HSE_HSE	
FBti0063779	HRE, HRE_HSE	*CG1239*	HRE_HRE, HSE_HSE	
FBti0063688[3]	HRE, HSE_HSE	*Lcp65Ag1*[3]	HSE, MRE, HSE_HSE	larval chitin-based cuticle development | chitin-based cuticle development
FBti0019411	HRE, HSE_MRE	*MFS9*	HRE, HSE_HSE	
FBti0019169	HRE, HSE, HRE_HRE, HSE_HSE, ARE_HSE, HSE_MRE	*CG33640*	HSE_HSE	
FBti0019480	HRE, HSE, HSE_HSE	*yellow-h*	HRE, HSE	
FBti0062442	HRE, MRE, HRE_HRE, HSE_HSE, HRE_HSE	*stnB*, *stnA*, *CR45009*	ARE, HSE_HSE, MRE_MRE, HRE_HSE | ARE, HSE_HSE, MRE_MRE, HRE_HSE | HRE	
FBti0019297	HRE, XRE, ARE_HSE	*CG14659*	HSE_HSE, ARE_HSE, HSE_MRE	
FBti0060187, *FBti0020057*[2], *FBti0020128*, *FBti0019453*[1]	HSE	*CG7370*, *CG8628*[2], *Dbp73D*, *Cnx99A*[1]	HRE, HSE | HSE | HRE_HRE, ARE_HSE | HRE, HSE_HSE	modENCODE *h* | oxidative stress
FBti0062779[1][2], *FBti0061582*, *FBti0063188*, *FBti0063589*[1][3], *FBti0019124*, *FBti0063758*[3], *FBti0060621*, *FBti0061129*[1], *FBti0061485*[1]	HSE_HSE	*La*[1][2], *RtGEF*, *CG6340*, *CR43281*, *Tak1*[1][3], *mRpL28*, *Trpm*[3], *TpnC41C*, *be*[1], *CG34200*[1]	ARE, HRE, HSE_MRE | ARE, HSE_MRE | HRE, HSE_HSE | HRE, HSE_HSE, HRE_HSE_XRE | HRE, HSE | HRE, HSE_HSE, HRE_HSE, HRE_XRE | HSE_HSE | HRE, MRE, HSE_HSE | HSE_HSE | HSE, MRE	modENCODE *h* | CBP/p300 | HSF | oxidative stress | defense response | defense response to virus | defense response to Gram-negative bacterium | cellular response to lipopolysaccharide | positive regulation of biosynthetic process of antibacterial peptides active against Gram-negative bacteria | zinc ion homeostasis | divalent metal ion transport
FBti0062874	HSE_HSE, HRE_HSE_MRE	*CG34283*	HSE_HSE	
FBti0019721, *FBti0019594*	HSE_HSE, HSE_MRE	*CG14613*, *CG14438*	HRE_HRE, HSE_HSE, MRE_MRE,	

			HSE_MRE, HRE_HSE_XRE | MRE_MRE	
FBti0019116, *FBti0019123*[2], *FBti0019379*, *FBti0019627*[1], *FBti0060388*	HSE_MRE	*Sfp23F*, *CR44791*, *Jon25Bii*[2], *CG6974*, *CR44266*[1], *CG31731*	HRE, HSE, HSE_HSE | HSE, HSE_HSE | ARE, HRE, MRE, ARE_ARE, HRE_HSE | ARE, HRE, HSE_HSE, ARE_HSE | HRE, HSE_HSE | ARE, HRE, MRE, HSE_HSE	CBP/p300 | oxidative stress
FBti0063555	HSE, HSE_HSE	*CG31612*	ARE, XRE, MRE, HSE_HSE	
FBti0060649	HSE, HSE_HSE, HRE_XRE	*CG14452*	HRE, ARE_HSE	
FBti0061084[1][2]	HSE, MRE	*Rab18*[1][2]	ARE, HRE, HSE, HSE_HSE	CBP/p300 | oxidative stress
FBti0059791	MRE, MRE_MRE, ARE_HSE	*tRNA:K2:42Ac*	HRE, HSE	
B - TEs adds different SREs in gene promoter region				
FBti0018880[1]	ARE, HSE, HSE_HSE, ARE_HSE_MRE	*Jheh2*[1]	HRE	modENCODE *cnc*
FBti0061428	ARE, MRE, HSE_HSE	*CG31809*	HRE	
FBti0060121[3]	ARE, XRE	*Ir40a*[3]	HSE_HSE	detection of chemical stimulus involved in sensory perception | detection of chemical stimulus involved in sensory perception of smell | olfactory behavior
FBti0018940, *FBti0019755*[2], *FBti0019841*, *FBti0060103*, *FBti0061511*, *FBti0062979*, *FBti0063300*, *FBti0063994*	HRE	*tRNA:CR30449*, *tRNA:CR30454*, *LManV*[2], *Lamp1*, *CG32507*, *az2*, *CG34332*, *Snm1*, *CG32797*	ARE, HSE_HSE | ARE, HSE_HSE | HSE | ARE, HSE_HSE | HSE_HSE, HSE_MRE | HSE_HSE | ARE | HSE_HSE | MRE, HSE_HSE	oxidative stress | heat shock stress
FBti0060734	HRE_XRE	*CG14451*	ARE, HSE_HSE	
FBti0064133	HSE_HSE	*CR32511*	HRE, XRE	
FBti0019721	HSE_HSE, HSE_MRE	*CG14618*	HRE	
FBti0061084	HSE, MRE	*RpL35*	ARE, HRE	
FBti0018876, *FBti0019133*[2]	MRE_MRE	*Epac*, *RapGAP1*[2]	HRE, ARE_ARE, HRE_HSE, HSE_XRE | ARE, HRE, HSE, HRE_HSE	oxidative stress
C - TEs introduce SREs in gene promoter region				
FBti0018961	ARE, HRE, HRE_HRE, HSE_HSE, ARE_HSE	*CG42336*		
FBti0019827	ARE, HRE, HSE_HSE	*del*		
FBti0020117	ARE, HRE, HSE_HSE, ARE_HSE, HSE_MRE	*Best4*		

FBti0019622	ARE, HRE, HSE, HSE_HSE, ARE_HSE	*CG11106*	
FBti0019747	ARE, HRE, HSE, HSE_HSE, HRE_HSE	*CG12446*	
FBti0019436	ARE, HRE, HSE, MRE, HSE_HSE	*CR45226*, *CR46083*	
FBti0020026[3], *FBti0019276*, *FBti0020152*[1]	ARE, HRE, HSE, MRE, HSE_HSE, ARE_HSE	*Jafrac2*[3], *CR45128*, *term*[1], *CR45389*	CBP/p300 | response to oxidative stress | defense response to bacterium | cellular response to cold
FBti0020127	ARE, HRE, MRE, ARE_ARE, HSE_HSE, ARE_HSE	*CR45179*	
FBti0019165[1]	ARE, HRE, MRE, HSE_HSE	*snRNA:U5:34A*[1]	CBP/p300
FBti0019353[1], *FBti0019754*	ARE, HRE, MRE, HSE_HSE, ARE_HSE, HRE_XRE, HSE_MRE	*CG31406*[1], *CR43855*	CBP/p300
FBti0019082[1]	ARE,HSE,HRE_HRE,HSE_HSE	*CG18259*[1], *CR6900*	modENCODE *h*
FBti0019333	ARE, MRE, HSE_HSE, ARE_HSE, HRE_XRE, HSE_MRE	*CG7900*	
FBti0060155	HRE	*CG3635*	
FBti0062284	HRE_HRE	*CG8915*	
FBti0019217[1]	HRE, ARE_HSE, HRE_HSE	*Oseg5*[1]	CBP/p300
FBti0060834	HRE, HSE	*CG33494*	
FBti0020162[3], *FBti0019197*[2]	HRE, HSE_HSE	*brv1*[3], *CG15152*[2], *CG44476*	oxidative stress | cellular response to cold | detection of temperature stimulus involved in thermoception
FBti0019993	HRE, HSE_MRE	*CR44459*	
FBti0018951	HRE, HSE, MRE, HSE_HSE, HRE_HSE	*CR45140*	
FBti0019090[3]	HRE, XRE, HSE_HSE, ARE_HSE	*Cdc42*[3], *CG14229*	melanotic encapsulation of foreign target | sensory perception of touch
FBti0061551, *FBti0019062*	HSE	*CaMKII*, *Lsd-2*	

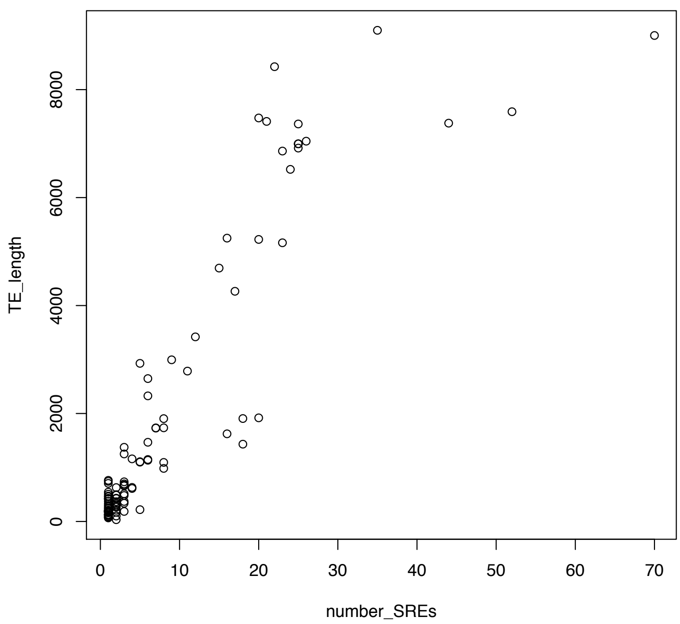
TEs that add only one SRE are not likely to have an impact on the gene expression of nearby genes. Otherwise the addition of several SREs allows the union of many TFs that increase significantly the gene expression (Xie et al. 2010). We look for the number of SREs en each TE for each three scenarios: A, B and C (Table 7). Thirty-five out of 108 TEs contains unique SRE hits. Thus, the other 73 TEs are more likely to affect the gene expression of nearby genes.

Table 7. Number of SREs in the three different scenarios: TEs that add the same type of SREs that were already present in the promoter regions (**#TEs add same SREs)**, TEs that add different SREs in a promoter region that do not contain those (**#TEs add different SREs**) and TEs that introduce new SREs in a promoter region where there were no SREs (**#TEs introduce SREs)**. (#: Number)

	1 SRE	2 - 4 SREs	5 - 9 SREs	10-20 SREs	21-30 SREs	>30 SREs
#TEs add same SREs	8	7	2	0	0	0
#TEs add different SREs	23	20	12	3	7	3
#TEs introduce SREs	4	6	3	8	4	1
Total:	35	33	17	11	11	4

Finally, we also tested whether there is a correlation between the number of SREs and the length of TE (Figure 3). We found that there is a strong correlation between number of SREs added by a TE and the length of that TE. We performed a nonparametric test because our data is not normally distributed (Spearman correlation test value is 0.867 with a p-value less than $2.2e^{-16}$).

Table 3. Distribution between number of SREs and TEs length.



DISCUSSION

We have used bioinformatics tools to predict five different SREs. Because SREs are short sequences that could be find by chance in many genomic regions, we have applied some filters to distinguish spurious from actual signals.

Curation of TFBSTools output

We showed that HRE has a repeated hit rate significantly higher than others SREs (Table 2). Looking at the core sequence for each HIF PWMs revealed that they are palindromic sequences. Consequently, TFBSTools create a hit in both strands of the sequence. M00259, which has the highest repeated rate, has a perfect palindromic core sequence while the others differ in some nucleotides:

M00259 PWM core sequence: NNNNN(A/G)TCACGTGA(T/C)NNNNN
M00259 reverse sequence: NNNNN(T/C)AGTGCACT(A/G)NNNNN

M00236 PWM core sequence: NN(T/C/A)(T/C/A)NNGTGCACNNNNNN
M00236 reverse sequence: NN(A/G/T)(A/G/T)NNCACGTGNNNNNN

MA0259.1 PWM core sequence: NNACGTGN
MA0259.1 reverse sequence: NNTGCACN

Although TFBSTools program considers 0.75 score value enough to recognize a SREs position, our data shows different patterns when the score threshold was changed. We do not expect to find all five SREs in the majority of TEs as we observed with 0.75, 0.80 and 0.85 tresholds. We discarded 0.95 threshold because it appears to be too strict as no prediction sites for XRE and MRE are identified with this threshold. Thus, we decide to be conservative and take 0.90 score value as a threshold so we get a more likely pattern (Figure 1).

Moreover, for thresholds less than 0.90 there is no case of TE inserting different SREs in a promoter region that do not contain those SREs. As the threshold becomes stricter the possibility to find a SREs by chance decreases. Consequently, the chance that the same SREs are found in both TE and promoter region also decreases. Thus, the change in the pattern observed at 0.90 threshold supports this score value as a good threshold (Table 5).

Functional annotation and population genetics

One of the applied main filters is the distance to the nearby genes because 1kb upstream regions are likely to have a biological impact on that gene (Hoskins et al. 2011). Consequently, TEs that add SREs out of those regions were discarded from the analysis. Recently, it has been shown that first introns have also a role in regulation (Yepiskoposyan et al. 2006). These regions could also be included to detect additional putatively adaptive TEs. (see section *Future perspective*).

Taking advantage of the next-generation sequencing (NGS) techniques estimating accurate TE frequencies in the population is now possible. We took information of 8 worldwide *Drosophila* populations to discard TEs that are unlikely to be involved in adaptive processes. Most of the TEs are assumed to be deleterious and are observed at low frequencies (Barron et al. 2014). Then, the neutral TEs could increased in frequency and eventually get fixed. Some TE insertions can increase the fitness of the individual that carries them and play a role in genome adaptation (Barron et al. 2014). Hence, we considered that high frequent TEs are more likely to be adaptive than TEs at low frequencies. Thus, we discard those TEs that are at low frequency in the population. We additionally decide to discard INE-1 TEs family for its particular dynamics. Previous analyses of this family indicate that a burst of transposition occurred in the ancestor of *D. melanogaster* and it have remained inactive (Yang et al. 2006). As the split from the closest species, *D. simulans*, is estimated to happened before the world colonization of *D. melanogaster* (Tamura et al. 2004, Duchen et al. 2013), this family is probably not relevant in recent adaptive processes.

We have run TFBSTools program for all 5,416 annotated TEs in *D. melanogaster* genome. Unfortunately, we only have accurate TE frequency estimates for 1,636 TEs. Among these TEs, as a result of our study, we obtained a dataset of 108 TEs that are likely to be adaptive by the addition of SREs to the promoter region of *D. melanogaster* genes.

***In silico* validation of the identified SREs**

Based on the results of previous ChIP-Seq studies, we validated a total of 17 SREs in the promoter regions of our selected genes (Guertin and Lis 2010, Holmqvist and Mannervik 2013). SREs present inside TEs can not be validated because repetitive sequence like TEs are excluded in ChIP-Seq experiments due to mapping difficulties of the reads. We could solve this issue by finding unique SNPs for each TE. Then we could map the raw reads data of ChIP-Seq studies with our sequences using BLAST tool. In this work, we did not have enough time to apply this method to find evidence of SREs inside our selected TEs, but it should be done in the future.

Nine out of 17 selected genes contain a single CBP/p300 binding. CBP/p300 is a general co-activator: it binds to other TFs and increase the expression of their target genes (Holmqvist and Mannervik 2013). We can consider that those 9 genes have a TF binding their promoter region, but it can be stress related or not. Still, we considered informative to know the binding of CBP/p300 co-activator because it is suggesting modulation of those genes by TFs.

ChIP studies provide evidence for *in vivo* binding between proteins and DNA (Landt et al. 2012). Union of TFs to promoter regions strongly suggests a modulation of those genes. Although, gene expression modulation is highly complex and more evidence are needed to confirm real function of the identified SREs.

Additional evidence for the role of our selected TEs in stress-response

We used GWAS and GO functional annotation to look for additional evidence of our 108 candidate TEs in stress-response.

GWAS studies provide evidence for gene expression variation for a specific stress condition but they can not be linked with TF activity. We checked 6 GWAS and confirm 12 genes of our list to be related to oxidative, heat shock or hypoxia stresses. Five out of 10 genes related to oxidative stress are present in both Girardot et al. (2004) and Landis et al. (2004) studies. Moreover in four of them an ARE and/or HSE are found in either the TE and/or the promoter region. HSF function is related not only with heat shock stress but also with general stress response. The other gene contains a HRE position in both the TE insertion and the promoter region. This result strongly suggests a biological function involved in oxidative stress response of these genes.

Hsp70Aa and *HSP70Ab* (Heat shock proteins) genes appear in Azad et al. (2009) hypoxia GWAS. This result suggests the wide function of heat shock proteins in other stress response as has been previously suggested (Santoro 2000). Unfortunately no GWAS were found for xenobiotic and metal stress in *D. melanogaster*.

There was no GO biological function annotated for 40 out of 127 selected genes. Eleven out of 87 genes have a stress related biological function GO term annotated. 6 of them match with other ChIP and/or GWAS evidences.

In summary, we found ChIP, GWAS, or GO evidence for stress related biological function for a total of 41 genes located nearby 35 TEs. Thus, these 35 TEs are more likely to be adding functional SREs.

Furthermore, because unique SREs are unlikely to have an impact on the gene expression of nearby genes, we expect that TEs containing binding regions of multiple TFs are more likely to affect the expression of neighbouring genes (Xie et al. 2010). Seventy-three out of 108 putatively adaptive TEs add more than one SRE and are thus more likely to be adding functional SREs (Table 7). Thus, 20 out of the 73 TEs, are more likely to be adding functional SREs because they add more than one SREs and are nearby genes involved in stress response.

Future perspective

Currently, we are experimentally validating 4 of our putatively adaptive TEs related to oxidative stress or heat shock stress (*FBti0061428*, *FBti0060121*, *FBti0019721* and *FBti0061084*). We select those TEs for their number of SREs related to their TE length, that is, they add several SREs and are short. We also take into account the SREs present in the promoter region of nearby genes and the gene function. *FBti0061428* and *FBti0060121* add 2 and 3 SREs hits respectively in a promoter region that already contain 1 SRE. *FBti0061428* and *FBti0060121* are inserted in the promoter region of *Ir40a* and *CG31809* genes respectively. These two genes have been related to insecticide resistance, which strongly indicate a function in stress response (Kalajdzic et al. 2012, Kain et al. 2013, Silbering et al. 2011, Croset et al. 2010). The other two selected genes (*RpL35* and *CG14618*) do not have a clear stress related function, but they have a similar scenario as the other two selected TEs.

Further bioinformatic analysis might be done to identify more TEs that add SREs in *D. melanogaster*. For example, there are evidences that confirm that first introns contain regulatory sequences such as TFBS (Yepiskoposyan et al. 2006). Therefore, we are currently running TFBSTools software to find SREs in TEs inserted in the first introns.

CONCLUSION

Our results show that the addition of SREs by TEs in the promoter region of *D. melanogaster* genes is not an uncommon event. We identified 108 TEs that are likely to be introducing functional SREs. Out of these 108 TEs, 73 add more than one SRE and are therefore more likely to be functional. Further experimental validation is needed in order to prove the functionality of these TEs.

REFERENCES

Abel J, Haarmann-Stemmann T. 2010. An introduction to the molecular basics of aryl hydrocarbon receptor biology. Biol Chem 391:1235-1248.

Aminetzach YT, Macpherson JM, Petrov DA. 2005. Pesticide resistance via transposition-mediated adaptive gene truncation in Drosophila. Science 309:764-767.

Azad P, Zhou D, Russo E, Haddad GG. 2009. Distinct mechanisms underlying tolerance to intermittent and constant hypoxia in Drosophila melanogaster. PLoS One 4:e5371.

Bahadorani S, Mukai S, Egli D, Hilliker AJ. 2010. Overexpression of metal-responsive transcription factor (MTF-1) in Drosophila melanogaster ameliorates life-span reductions associated with oxidative stress and metal toxicity. Neurobiol Aging 31:1215-1226.

Balamurugan K, Egli D, Selvaraj A, Zhang B, Georgiev O, Schaffner W. 2004. Metal-responsive transcription factor (MTF-1) and heavy metal stress response in Drosophila and mammalian cells: a functional comparison. Biol Chem 385:597-603.

Barron MG, Fiston-Lavier AS, Petrov DA, Gonzalez J. 2014. Population genomics of transposable elements in Drosophila. Annu Rev Genet 48:561-581.

Bergland AO, Behrman EL, O'Brien KR, Schmidt PS, Petrov DA. 2014. Genomic evidence of rapid and stable adaptive oscillations over seasonal time scales in Drosophila. PLoS Genet 10:e1004775.

Croset V, Rytz R, Cummins SF, Budd A, Brawand D, Kaessmann H, Gibson TJ, Benton R. 2010. Ancient protostome origin of chemosensory ionotropic glutamate receptors and the evolution of insect taste and olfaction. PLoS Genet 6:e1001064.

Daborn PJ, Yen JL, Bogwitz MR, Le Goff G, Feil Eet al.,. 2002. A single p450 allele associated with insecticide resistance in Drosophila. Science 297:2253-2256.

Daub JT, Hofer T, Cutivet E, Dupanloup I, Quintana-Murci L, Robinson-Rechavi M, Excoffier L. 2013. Evidence for polygenic adaptation to pathogens in the human genome. Mol Biol Evol 30:1544-1558.

David JR, Capy P. 1988. Genetic variation of Drosophila melanogaster natural populations. Trends Genet 4:106-111.

Duchen P, Zivkovic D, Hutter S, Stephan W, Laurent S. 2013. Demographic inference reveals African and European admixture in the North American Drosophila melanogaster population. Genetics 193:291-301.

Egli D, Selvaraj A, Yepiskoposyan H, Zhang B, Hafen E, Georgiev O, Schaffner W. 2003. Knockout of 'metal-responsive transcription factor' MTF-1 in Drosophila by homologous recombination reveals its central role in heavy metal homeostasis. EMBO J 22:100-108.

Enoki Y, Sakurai H. 2011. Diversity in DNA recognition by heat shock transcription factors (HSFs) from model organisms. FEBS Lett 585:1293-1298.

Feder ME, Hofmann GE. 1999. Heat-shock proteins, molecular chaperones, and the stress response: evolutionary and ecological physiology. Annu Rev Physiol 61:243-282.

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

Fiston-Lavier AS, Barron MG, Petrov DA, Gonzalez J. 2015. T-lex2: genotyping, frequency estimation and re-annotation of transposable elements using single or pooled next-generation sequencing data. Nucleic Acids Res 43:e22.

Futuyma DJ. 2013. Evolution: Sinauer Associates, Incorporated.

Girardot F, Monnier V, Tricoire H. 2004. Genome wide analysis of common and specific stress responses in adult drosophila melanogaster. BMC Genomics 5:74.

Guertin MJ, Lis JT. 2010. Chromatin landscape dictates HSF binding to target DNA elements. PLoS Genet 6:e1001114.

Guio L, Barron MG, Gonzalez J. 2014. The transposable element Bari-Jheh mediates oxidative stress response in Drosophila. Mol Ecol 23:2020-2030.

Holmqvist PH, Mannervik M. 2013. Genomic occupancy of the transcriptional co-activators p300 and CBP. Transcription 4:18-23.

Hoskins RA, Landolin JM, Brown JB, Sandler JE, Takahashi Het al.,. 2011. Genome-wide analysis of promoter architecture in Drosophila melanogaster. Genome Res 21:182-192.

Huang W, Massouras A, Inoue Y, Peiffer J, Ramia Met al.,. 2014. Natural variation in genome architecture among 205 Drosophila melanogaster Genetic Reference Panel lines. Genome Res 24:1193-1208.

Kain P, Boyle SM, Tharadra SK, Guda T, Pham C, Dahanukar A, Ray A. 2013. Odour receptors and neurons for DEET and new insect repellents. Nature 502:507-512.

Kalajdzic P, Oehler S, Reczko M, Pavlidi N, Vontas J, Hatzigeorgiou AG, Savakis C. 2012. Use of mutagenesis, genetic mapping and next generation transcriptomics to investigate insecticide resistance mechanisms. PLoS One 7:e40296.

Kidwell MG, Lisch DR. 2001. Perspective: transposable elements, parasitic DNA, and genome evolution. Evolution 55:1-24.

Korsloot A, van Gestel CAM, van Straalen NM. 2004. Environmental Stress and Cellular Response in Arthropods: Taylor & Francis.

Landis GN, Abdueva D, Skvortsov D, Yang J, Rabin BE, Carrick J, Tavare S, Tower J. 2004. Similar gene expression patterns characterize aging and oxidative stress in Drosophila melanogaster. Proc Natl Acad Sci U S A 101:7663-7668.

Landt SG, Marinov GK, Kundaje A, Kheradpour P, Pauli Fet al.,. 2012. ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. Genome Res 22:1813-1831.

Mateo L, Ullastres A, Gonzalez J. 2014. A transposable element insertion confers xenobiotic resistance in Drosophila. PLoS Genet 10:e1004560.

Mathelier A, Zhao X, Zhang AW, Parcy F, Worsley-Hunt Ret al.,. 2014. JASPAR 2014: an extensively expanded and updated open-access database of transcription factor binding profiles. Nucleic Acids Res 42:D142-147.

Matys V, Kel-Margoulis OV, Fricke E, Liebich I, Land Set al.,. 2006. TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. Nucleic Acids Res 34:D108-110.

Nadeau NJ, Jiggins CD. 2010. A golden age for evolutionary genetics? Genomic studies of adaptation in natural populations. Trends Genet 26:484-492.

Negre N, Brown CD, Ma L, Bristow CA, Miller SWet al.,. 2011. A cis-regulatory map of the Drosophila genome. Nature 471:527-531.

Nguyen T, Sherratt PJ, Pickett CB. 2003. Regulatory mechanisms controlling gene expression mediated by the antioxidant response element. Annu Rev Pharmacol Toxicol 43:233-260.

Pool JE, Corbett-Detig RB, Sugino RP, Stevens KA, Cardeno CMet al.,. 2012. Population Genomics of sub-saharan Drosophila melanogaster: African diversity and non-African admixture. PLoS Genet 8:e1003080.

Rebollo R, Romanish MT, Mager DL. 2012. Transposable elements: an abundant and natural source of regulatory sequences for host genes. Annu Rev Genet 46:21-42.

Santoro MG. 2000. Heat shock factors and the control of the stress response. Biochem Pharmacol 59:55-63.

Silbering AF, Rytz R, Grosjean Y, Abuin L, Ramdya P, Jefferis GS, Benton R. 2011. Complementary function and integrated wiring of the evolutionarily distinct Drosophila olfactory subsystems. J Neurosci 31:13357-13375.

Sorensen JG, Nielsen MM, Kruhoffer M, Justesen J, Loeschcke V. 2005. Full genome gene expression analysis of the heat stress response in Drosophila melanogaster. Cell Stress Chaperones 10:312-328.

Sundaram V, Cheng Y, Ma Z, Li D, Xing X, Edge P, Snyder MP, Wang T. 2014. Widespread contribution of transposable elements to the innovation of gene regulatory networks. Genome Res 24:1963-1976.

Sykiotis GP, Bohmann D. 2010. Stress-activated cap'n'collar transcription factors in aging and human disease. Sci Signal 3:re3.

Taguchi K, Motohashi H, Yamamoto M. 2011. Molecular mechanisms of the Keap1-Nrf2 pathway in stress response and cancer evolution. Genes Cells 16:123-140.

Tamura K, Subramanian S, Kumar S. 2004. Temporal patterns of fruit fly (Drosophila) evolution revealed by mutation clocks. Mol Biol Evol 21:36-44.

Tan G. 2015. TFBSTools: Software package for transcription factor binding site (TFBS) analysis. R package version 1.6.0, http://jaspar.genereg.net/. Version 1.6.0.

Tiffin P, Ross-Ibarra J. 2014. Advances and limits of using population genetics to understand local adaptation. Trends Ecol Evol 29:673-680.

Ullastres A, Petit N, Gonzalez J. 2015. Exploring the Phenotypic Space and the Evolutionary History of a Natural Mutation in Drosophila melanogaster. Mol Biol Evol.

Vorrink SU, Domann FE. 2014. Regulatory crosstalk and interference between the xenobiotic and hypoxia sensing pathways at the AhR-ARNT-HIF1alpha signaling node. Chem Biol Interact 218:82-88.

Weber AL, Khan GF, Magwire MM, Tabor CL, Mackay TF, Anholt RR. 2012. Genome-wide association analysis of oxidative stress resistance in Drosophila melanogaster. PLoS One 7:e34745.

Xie D, Chen CC, Ptaszek LM, Xiao S, Cao X, Fang F, Ng HH, Lewin HA, Cowan C, Zhong S. 2010. Rewirable gene regulatory networks in the preimplantation embryonic development of three mammalian species. Genome Res 20:804-815.

Yang HP, Hung TL, You TL, Yang TH. 2006. Genomewide comparative analysis of the highly abundant transposable element DINE-1 suggests a recent transpositional burst in Drosophila yakuba. Genetics 173:189-196.

Yepiskoposyan H, Egli D, Fergestad T, Selvaraj A, Treiber C, Multhaup G, Georgiev O, Schaffner W. 2006. Transcriptome response to heavy metal stress in Drosophila reveals a new zinc transporter that confers resistance to zinc. Nucleic Acids Res 34:4866-4877.

Zhou D, Haddad GG. 2013. Genetic analysis of hypoxia tolerance and susceptibility in Drosophila and humans. Annu Rev Genomics Hum Genet 14:25-43.

Zou S, Meadows S, Sharp L, Jan LY, Jan YN. 2000. Genome-wide study of aging and oxidative stress response in Drosophila melanogaster. Proc Natl Acad Sci U S A 97:13726-13731.