1-1-2016

Identification Of Lead-Sensitive Expression And Splicing Quantitative Trait Loci In Drosophila Melanogaster By Analysis Of Rna-Seq Data

Wen Qu
Wayne State University,

IDENTIFICATION OF LEAD-SENSITIVE EXPRESSION AND SPLICING QUANTITATIVE TRAIT LOCI IN DROSOPHILA MELANOGASTER BY ANALYSIS OF RNA-SEQ DATA

by

WEN QU

DISSERTATION

Submitted to the Graduate School

of Wayne State University,

Detroit, Michigan

in partial fulfillment of the requirements

for the degree of

DOCTOR OF PHILOSOPHY

2016

MAJOR: PHARMACOLOGY

Approved By:

_____

Advisor Date

_____

_____

_____

_____

DEDICATION

For the ancestors who paved the road ahead of the time upon whose shoulders I stand. This is also dedicated to my family and the many friends who supported me through this challenging but rewarding journey.

I would like to express my sincere gratitude to my advisor Professor Douglas Mark Ruden for his unwavering support and mentorship throughout the entire Ph.D. training.

I would like to extend my thanks to those who offered support and collegial guidance over the past five years: Professor Roger Pique-Regi, our Department Chair Professor Raymond R. Mattingly, Professor Paul Stemmer, Professor Hai-Young Wu and also our graduate officer Professor Sokol Todi and Professor Roy McCauley.

ACKNOWLEDGMENTS

Great thanks to Dr. Stuart Macdonald from the University of Kansas and Dr. Anthony Long from the University of California, Irvine for their kindness in providing the 8 founder strains of Drosophila Synthetic Population Resource (DSPR) and their recombinant inbred lines (RILs).

Special thanks to the funding authorities: National Institute of Health (5R01ES012933-10) and the Wayne State University Office of the Vice President for Research (D.M.R.).

Also, thanks to the High Performance Computing Services offered by Wayne State University for making the entire statistical analysis in a highly-efficient manner.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

GLOSSARY

(ordered alphabetically)

centiMorgans (cMs): A physical map specifies the physical position of markers on the chromosomes, in a genetic map distance is measured by the rate of cross-over events at meiosis. Two markers are d centiMorgens (cMs) apart if there is an average of d crossovers in the internvening interval in every 100 products of meiosis (Broman et al., 2003).

cis-eQTLs and trans-eQTLs: cis-eQTLs refer to genetic variants that affect a locus expression only on the same haplotype, while trans-eQTLs affect both. Therefore, cis-eQTLs tend to be "local"—close to the locus of the gene encoding the regulated transcript, while trans-eQTLs tend to be "distant"—away from the gene.

***Drosophila* Synthetic Population Resource (DSPR)**: It is comprised of a panel of approximately 1700 recombinant inbred lines (RILs) of *Drosophila melanogaster*, created by intercrossing 8 inbred founder lines (King E.G., 2012).

eQTL: expression Quantitative trait locus. QTL is a region of the genome that contributes to variation in a quantitative trait such as height, blood pressure. eQTL analysis treats the gene expression levels as quantitative traits and it searches for genomic loci that are responsible for the differential gene expression levels.

Expression heterogeneity (EH): used to describe patterns of expression variation due to unknown, unmeasured, or too complicated to measure factors (Leek and Storey, 2007). Without considering expression heterogeneity, the result will be less reliable, not only because of reduced power but also false positive signals.

LOD score: The LOD score is the logarithm of odds base 10. It is a statistical test often used for linkage and association analysis. The LOD score compares the likelihood of obtaining the test

data if the two loci are indeed associated to the likelihood of observing the same data purely by chance. Large LOD scores favor the presence of correlation.

MA plot: MA plot is a useful way to compare two groups. Each dot in the plot represents one gene. The reads on the x axis show the average expression profile throughout all the samples, while the y axis shows the log2 fold change between the lead treatment and the control.

Master-modulatory gene: We believe that the cluster of genes in each trans-eQTL hotspot is co-regulated by a gene encoded at the chromosomal locus (Ruden D.M., 2009b). The potential regulatory gene is called master-modulatory gene.

MatInspector: It is a software tool developed by Genomatix® to predict transcription factor binding sites via locating motif matches in DNA sequences (Cartharius K., 2005).

Microarray: A DNA microarray contains a specific DNA sequence (probe). It is a hybridization of DNA samples to a large collection of probes. Scientists use this technique to measure the level of gene expression or gene structure.

NMDAR: N-methyl-D-aspartic acid receptors

Pair-end reads: Short cDNA fragments can be sequenced from one or both ends (Majewski J., 2011).

QTL: a QTL represents a genomic location that is responsible for the variation in the quantitative trait of interest (eg. height, body weight).

Recombinant Inbred Lines (RILs): An organism with chromosomes that incorporate a permanent set of recombination events between chromosomes inherited from two or more inbred strains.

RNAi: RNA interference is a biological process that inhibits post transcriptional gene expression via RNA molecules such as microRNA and small interfering RNA.

RNA-seq: RNA sequencing, one of Next Generation Sequencing (NGS) applications, is sequencing mRNA present in a sample. Usually, mRNA is isolated from an organism or a tissue, converted into cDNA and cut into small fragments. Millions of those small fragments will be sequenced. Aligning these short sequences to the genome can provide information on gene expression (Majewski J., 2011).

R/qtl: is a QTL mapping software that is run in R environment. It is developed by Dr. Karl W. Broman and Dr. Saunak Sen.

sQTL: SNPs that influence the regulation of transcript isoform levels are referred to as "splicing QTL" (sQTL) (Pickrell J.K., 2010).

Trans-eQTL hotspots: This term is used to describe the chromosomal regions that influence the expression levels of multiple genes. It is a genomic locus that is associated with the regulation of a cluster of genes, regardless of their transcript locations.

CHAPTER 1 IDENTIFICATION OF EXPRESSION QTLS

Introduction

Lead Toxicity

Lead exposure has long been one of the most important topics in global public health. The major lead sources up until the 1970s when they were restricted in the United States were lead-containing paint and gasoline. The phase-out of these two sources in the US has resulted in dramatic reductions in mean blood lead level (BLL); however, lead exposure from environmental contamination remains a major world public health issue (Dietrich, 2001; White L.D., 2007). It was reported by the World Health Organization (WHO) that lead exposure is predicted to account for 143,000 deaths per year throughout the world and it is considered as one of the highest burdens in developing countries (WHOteam, 2015). Lead contamination in our city of Detroit and our neighboring city Flint have been one of the heated topics for debate in the last two years because of the Flint Lead Water Crisis, which was caused by switching Flint's water supply from Lake Huron to the more corrosive Flint River (Hanna-Attisha et al., 2015).

The long-term effects of lead poisoning on humans, especially on children, include damage to the nervous system, heart, bones, intestines, kidney and reproductive system (Jedrychowski W., 2011). In early 2012, the Centers for Disease Control (CDC) lowered the reference blood lead level for children and pregnant women from 10 μg/dl to 5 μg/dl (Bellinger, 2013). Both the WHO and CDC have emphasized that no known level of lead is considered as "safe", referring to the irreversible danger of lead exposure (Bellinger, 2013; WHOteam, 2015). On the biological and cellular level, the direct effects of lead toxicity include mitochondrial damage, oxidative stress (Adonaylo and Oteiza, 1999), disruption of calcium homeostasis (Lafond et al., 2004), alteration of neurotransmitter release, altered function of neurotransmitter and receptors (Suszkiw, 2004), and apoptosis (Oberto et al., 1996).

Lead's ability to mimic as calcium makes it able to cross the blood brain barrier (BBB) (Bradbury and Deane, 1992). The effects of lead on neurotransmission include damage of synapses, alteration of neurotransmitter receptors and causing apoptosis or necrosis in dopamine systems (Jabłońska et al., 1994). The molecular targets and genetic mechanisms of lead remain unclear, though N-methyl-D-aspartic acid receptors (NMDAR) have been believed to contribute to Pb neurotoxicity at the synapse level (Baranowska-Bosiacka I., 2012). NMDAR play a key role in synapses and also in the process of learning and memory. They were believed to become excessive stimulated by Pb toxicity and this led to excess calcium flow-in thorough NMDAR, which could lead to lethal damage to the neurons (Marchetti and Gavazzo, 2005; Baranowska-Bosiacka I., 2012).

Single Nucleotide Polymorphisms (SNPs)

The DNA is transcribed into single-stranded RNAs and then the RNAs, after splicing, are used as templates for synthesizing proteins. During this fundamental process, known as the central dogma of molecular biology, there are numerous factors influencing the protein function, such as DNA sequence variation. Single nucleotide polymorphisms, frequently called SNPs, are the most common type of genetic variation. Each SNP, by definition, represents a difference in a single DNA nucleotide. For example, most individuals might have base G at a specific genomic location, but a small population has base A instead. There are approximately 10 million SNPs in the human genome that have been characterized in whole genome sequencing projects, such as the 1000 genome project (Siva, 2008). Although most SNPs are believed to have no effects on health, some of these genetic variations cause an increased susceptibility to diseases, such as sickle-cell anemia and cystic fibrosis. SNPs may also reveal an individual's susceptibility to environmental toxic factors, an increased risk of developing certain diseases, and an atypical response to particular medical treatments or drugs (Squassina et al., 2010). SNPs can also be

used to trace evolutionary ancestries. Studies working on association of genome and disease outcomes lays the foundation for future individualized therapy, which is also called personalized medicine (Squassina et al., 2010).

Studies that correlate SNPs and diseases are called genome-wide association studies, and are also known as GWAS (pronounced “GeeWass”). However, researchers have also started to examine the correlation between SNPs and global gene expression profiles, or more precisely, steady state mRNA levels (Majewski and Pastinen, 2011).

Expression QTLs

One of the biggest challenges in biology is to understand how genetic variation alters gene expression, which is also known as genetical genomics (Mackay et al., 2009; Massouras et al., 2012; Lagarrigue et al., 2013). Genetics of gene expression has been studied in various species, such as maize (Schadt et al., 2003), yeast (Brem et al., 2002; Yvert et al., 2003; Bing N., 2005; Brem et al., 2005), roundworms (Francesconi and Lehner, 2014), flies (Hirsch et al., 2009; Massouras et al., 2012), mice (Schadt et al., 2003; Huang et al., 2009) and humans (Schadt et al., 2003; Mangravite et al., 2013; Zhang et al., 2014). Expression Quantitative Trait Loci (eQTL) analyses, which search for genomic loci that are responsible for the differential gene expression levels, has shed light on the genetic structure of transcriptional regulation. The first achievement in this field was seen in the budding yeast, where differential gene expression was shown to be segregated by parental genotypes (Brem et al., 2002).

Significant eQTLs were often categorized into two sub-groups: cis-eQTLs and trans-eQTLs. By their classical definitions, cis-eQTLs refer to genetic variants that affect a locus expression only on the same haplotype, while trans-eQTLs affect both haplotypes (Benzer, 1955; Hasin-Brumshtein et al., 2014). A haplotype is defined as a set of SNPs on one chromosome that occur together because they are tightly linked and, therefore, are from one

parent. Such information is critical for investigating the genetics of common diseases, such as those investigated in humans by the International Hapmap Project (Gibbs et al., 2003). Accordingly, cis-eQTLs tend to be “local”—near the locus of the gene encoding the regulated transcript, while trans-eQTLs tend to be “distant”—away from the locus of the regulator (Benzer, 1955; Hasin-Brumshtein et al., 2014).

During the past several years, multiple cis-eQTLs were detected in human lymphoblastoid cell lines (Pickrell et al., 2010; Lappalainen et al., 2013; Mangravite et al., 2013). Several disease—specific cis-eQTLs were also detected, one of which proved the correlation between a statin-related eQTL for the gene glycine amidinotransferase (*GATM*) and statin-induced myopathy (Mangravite et al., 2013).

In contrast to the high production of cis-eQTLs, fewer trans-eQTLs were identified, let alone disease-specific trans-eQTLs. One of the most mysterious types of eQTLs are trans-eQTL hotspots, where one single location is associated with the regulation of multiple genes, regardless of their transcript locations (Mangravite et al., 2013). The existence of trans-eQTL hotspots were previously confirmed in budding yeast in 2003, where the gene Antagonist of Mitotic Exit Network 1 (*AMN1*) was shown to trans-regulate a cluster of 12 downstream genes, irrespective of their transcript distances and located throughout the yeast genome. Trans-eQTL hotspots are usually described as being eQTL in trans-regulatory factors, such as transcription factors or signaling proteins, but these types of eQTLs have been hard to identify outside of yeast, and require further study.

Gene Expression Studies: RNA-seq and Microarrays

RNA-seq has been considered as a revolutionary tool for transcriptomics (Wang et al., 2009). This technology converts tissue RNAs to a library of DNA fragments with adaptors attached to the ends that hybridize to flow cells for next-generation DNA sequencing, such as

with the HiSeq2500 in the Wayne State University Applied Technology Genomics Core. Each fragment, up to 600 million at a time, is directly sequenced in a high-throughput manner. But before RNA-Seq, gene expression studies were mostly performed by hybridization-based microarrays. This microarray technology uses a collection of microscopic DNA spots, which contain DNA sequences that are complementary to the mRNA, to measure the expression profiles of large numbers of genes simultaneously.

Microarrays are a robust reliable method proven over decades. Furthermore, they are often more economical than Next Generation Sequencing (NGS) and have a well-established protocol for processing the data. Microarrays also have a significant advantage when working with a large number of samples. On the other hand, the advantage of RNA-seq lies in its independence to prior sequence knowledge. This enables the detection of structural variations such as alternative splicing and novel transcripts. Although both platforms include robustness and high-reproducibility, RNA-seq suffers less from numerous biases as well as background noise when measuring low abundance transcripts. For microarrays, studies have observed a sharp rise of false positives when thousands of genes were processed simultaneously (Xiao et al., 2002; Fadiel and Naftolin, 2003) and the sources of these biases are not yet well understood. For RNA-seq, the technology is inherently more sensitive in detecting low expression values since each transcript is sequenced individually.

In this chapter, we use the RNA-sequencing technology to quantify gene expression profiles and compare it with our previous microarray data (Ruden D.M., 2009b).

Previous Lab Experiments

In order to better understand how lead plays a role as a neurotoxin, our lab utilizes the *Drosophila melanogaster* model to study the effects of developmental lead exposure on steady-state mRNA levels in adult brains in order to identify lead-responsive genes. Our lab has

already shown that *Drosophila* fed with 250 µM lead acetate in standard fly food, which results in lead levels of 50-100µg/dL in tissue, results in gene expression (Ruden D.M., 2009b), synaptic (He et al., 2009), and behavioral (Hirsh H.V., 2009) changes. We have previously found that lower lead levels in the food, i.e. 50 µM lead acetate, altered the uniformity of the synaptic match between the size of the motor neuron terminal and muscle fibers at larval neuromuscular junctions (Morley E.J., 2003) and resulted in behavioral changes including courtship (Hirsch et al., 1995) and locomotor activity (Hirsch et al., 2003). In a recent study on Detroit children, our laboratory has also shown that lead exposure could have multigenerational epigenetic effects (Sen et al., 2015). We have also found that lead exposure in human embryonic stem cells can affect DNA methylation and hydroxymethylation at specific genes. However, identifying the genetic mechanism of lead induced neurotoxicity is facilitated by more detailed studies of gene regulatory networks in model organisms.

In a precursor paper, which was published in 2009, our lab performed eQTL analyses of microarray data by comparing Pb-treated whole males to the control ones. In that paper, we identified 12 genomic regions (5 in the control males and 7 in the Pb-treated males), which we called “transbands” or “trans-eQTL hotspots” because many genes were affected by a single locus and perhaps the locus contains potential lead-responsive master regulatory genes (Ruden D.M., 2009b). While it was an intriguing result, this analysis only utilized 92 genotype markers, and was performed prior to the complete genome sequencing of *Drosophila*. A further limitation of the earlier study was that each of the 12 trans-eQTL hotspots could only be restricted to a region of 5 centi-Morgans (cM), which hinders the ability to fine map the targeted genomic location, identify and verify potentially master regulatory genes.

In order to extend the earlier study, and to further validate the existence of the detected 12 trans-eQTL hotspots, our lab used another set of the *Drosophila* recombinant inbred lines

(RILs) for this study, the *Drosophila* Synthetic Population Resource (DSPR), to collect additional expression analyses. In this chapter, we used RNA-seq and focused on genomic information on 11768 genomic markers (King et al., 2012). Each sample from the DSPR was a mosaic of eight parental strains, which were from different geographic locations and should include a large collection of genetic variance. By using this information, we were able to restrict the regulatory genomic regions within 10kb. In this chapter, we present the results of these findings and provide further validation of the existence of lead-responsive trans-eQTL hotspots.

Methods

Genotype Data

The 8 founder strains of *Drosophila* Synthetic Population Resource (DSPR) and their recombinant inbred lines (RILs) were kindly provided by Dr. Stuart Macdonald from the University of Kansas and Dr. Anthony Long from the University of California, Irvine. The RILs were started with eight founder strains, A1- A8 that were of diverse geographic origins (Table S1) and may include a great deal of the genetic variation in the *Drosophila* species (King et al., 2012). Strains were first intercrossed, A1 was crossed with A2, A2 was crossed with A3, and this crossing went on until A7 was crossed with A8 (King et al., 2012). 10 F1 flies per genotype per sex were mixed altogether and continued to produce offspring (King et al., 2012). Until the 50th generation of crossing, offspring were separated and another ~25 generations of sibling inbreeding made the finished DSPR A2 subpopulation ~800 RILs contain only 1% of heterozygous founder genotype (King et al., 2012).

The DSPR constructed 96-plexed restriction-site associated DNA (RAD) libraries, which further resulted in the revelation of 10,275 SNPs (King et al., 2012). They used the hidden Markov model (HMM) to convert the SNP data to estimate the probability of the underlying founder genotype for the *Drosophila* genome (genotyping error rate: 0.5%) (King et al., 2012).

Since all RIL samples are mostly homozygous and they have in total eight parents (marked as A1-A8), there are at most eight possible genomic origins for any genomic position. The *Drosophila* genome (only chromosome X, 2, and 3; chromosome 4 was excluded) was divided into 11,769 10kb genomic segments, resulting in 11,768 markers at the junction point. The genotype dataset shows the founder name of each of the 11768 markers for all the samples.

Sample Preparation

All the fly stocks were reared at 25° C in 35 ml vial containing standard *Drosophila* 10 ml medium. To cause lead-poisoning, medium was mixed with 250 μM PbAc for lead-containing medium or 250 μM NaAc for control. This results in the *Drosophila* head containing 50-100 μg/dL lead (Ruden D.M., 2009b). Next, 79 randomly selected DSPR samples were fed, from egg to adult, either control food or lead-containing food. We did not have any technical or biological replicates in this experiment, since we prefer the maximum inclusion of RILs.

Fifty heads of adult male flies (5-10 days old) in each of the 79 strains were collected and TruSeq Cluster RNA sample prep kit from Illumina was used to prepare the samples. 1μg of RNA was used after RNA isolation. The High Sensitivity D1K ScreenTape on the Agilent TapeStation instrument and quantitative PCR on the QuantStudio 12K Flex were used to make sure the quality of library. RNA expression analyses were performed with fifty-cycle paired-end RNA-seq on the Hiseq2000™ instrument from Illumina. General read quality was verified using FastQC (Andrews, 2010). The RNA-seq raw data are available on the NCBI GEO accession: GSE83141.

Expression Profiling

Tophat2 (V2.0.8) was used to map reads against the known *Drosophila Melanogaster* (UCSC/dm3) transcriptome (Kim et al., 2013). The transcript assembly tool Cufflinks and differential expression tool Cuffdiff were utilized for gene discovery and comprehensive

expression analysis of RNA-seq data (Trapnell C., 2012). After the Cufflink pipeline, we assembled all the expression data and quantile normalized to the overall average empirical distribution across all samples first, then across all genes. Gene Ontology (http://geneontology.org/) (Ashburner et al., 2000; Consortium, 2015) was used for the GO enrichment analysis for the differentially expressed genes upon Pb exposure and GO categories of "Molecular Function" and "Biological Process" were selected.

Genome-Wide eQTL Mapping

A data analysis R package called DSPRqtlDataA (http://wfitch.bio.uci.edu/~dspr/index.html) was provided by the DSPR group (King et al., 2012). We used it to extract the genotype dataset indicating the genomic origin at 10,768 loci for each sample we used. Similar to what the DSPR group did, we performed a multiple regression—regressing gene expression profiles on the eight additive genotype probabilities with zero covariate. We also used the LOD score (Manichaikul et al., 2009) to quantify the likelihood of association between 10,768 genomic locations and 13,381 gene expression profiles among 79 paired samples (one control and one Pb-treated).

$$H0: Y = \mu + \varepsilon$$

$$H: Y = \mu + \sum Gi + \varepsilon$$

Where μ is the grand mean, *Gi* is the ith parental genotype probability.

The LOD score, which is the logarithm of odds base 10, is a statistical test commonly used for linkage and association analysis. It compares the likelihood of obtaining the test data if the two loci are indeed associated to the likelihood of observing the same data purely by chance. Positive LOD score favors the presence of correlation.

LOD score = $\log_{10}$(Likelihood of H1) –$\log_{10}$(Likelihood of H0)

After obtaining the LOD score for each genomic location and each gene expression level, we determined the significance threshold for each gene via 1000 permutations on its expression levels.

The expression levels for each gene were extracted, shuffled randomly, and the LOD score was calculated in the same way as described above but using the shuffled data. A null distribution could be formed by the one thousand LOD scores, resulted from the thousand times of shuffling.

$$p - \text{value for gene x} = \frac{\text{numbers of permutations for whose lodscore} \leq \text{observed lodscore}}{\text{total number of permutations (=1000)}}$$

After obtaining all the *p*-values, we defined eQTLs as *p*-value ≤ 0.05, including cis-eQTL as genes that have significant associations with at least one genomic location within 1Mb geographic distance and trans-eQTL as genes that have significant associations with genomic locations outside of 1Mb. For the trans-eQTL hotspot threshold, we randomly shuffled the eQTL results 10,000 times. From each of the randomization, the highest number of associated genes for an eQTL was recorded. The *p*-value was generated based on the distribution of the total 10,000 recordings. Then qvalue function in R was used to transform *p*-value into FDR.

SVA was later used to control for potential confounders (Pickrell et al., 2010) and the following model

$$H0: Y = \mu + \sum Gi + E + \varepsilon$$

$$H: Y = \mu + \sum Gi + E + \sum Gi * E + \varepsilon$$

Where E represents two conditions: control or lead-treated.

Common Motif Search by Genomatix

The list of 89 anticipated downstream genes at Chr2L: 6,250,000 was entered as the input of Gene2Promoter (Retrieval and analysis of promoters). Among 100 loci containing 201 transcripts, 100 were selected, including both experimentally verified 5' complete transcripts and

some annotated transcripts that have not yet been confirmed. Sequences of these promoters were extracted by using Genomatix optimized length (500 bp upstream of the first transcription start site (TSS) and 100 bp downstream of last TSS). After obtaining all the promoter sequences of the downstream genes, CoreSearch (Wolfertstetter et al., 1996) was used to define unknown common motifs among the sequences. Tomtom was used to search for matches with the existing pool of motif databases (Gupta et al., 2007). Interactions browser on the flybase website was used in search for protein-protein interactions (http://flybase.org/cgi-bin/get_interactions.html) (Tweedie et al., 2009).

Results

Differential Expression Caused by Chronic Lead Poisoning

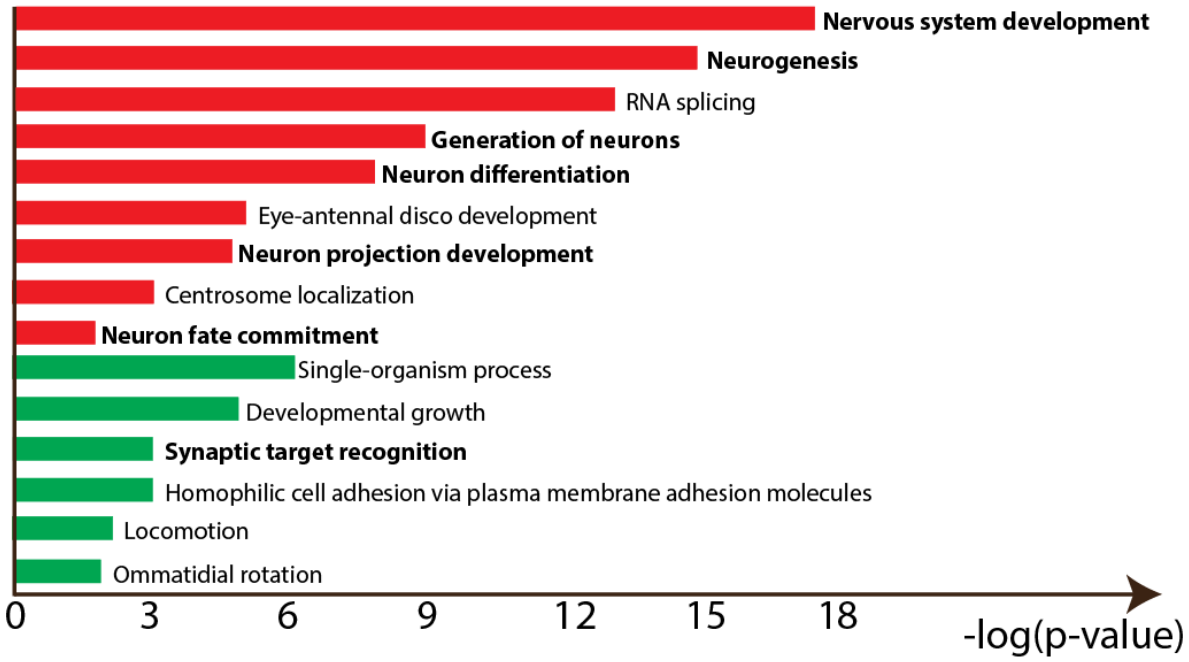
In order to further understand the trans-eQTL hotspots detected in our 2009 microarray paper (Ruden et al., 2009), we collected RNA-seq data on 79 recombinant inbred lines (RILs) selected from The *Drosophila* Synthetic Population Resource (DSPR) (King et al., 2012). The DSPR was composed of a panel of ~1600 *Drosophila* lines (King et al., 2012). The lines were initiated with eight parental strains A1-A8 that are from different geographic origins and should include a good mix of genetic variation in the *Drosophila* species which were intercrossed for 50 generations and then inbred for another 25 (King et al., 2012). We randomly selected 79 lines from the synthetic population and offspring were fed, from egg to adult, either control food (containing 250 μM NaAc) or Pb-treated food (containing 250μM PbAc). 50 heads of adult male flies (5-10 days old) in each strain were collected and RNA expression analyses were performed (See Methods). As a result, we had 79 control and 79 Pb-treated RNA-seq samples which we could analyze for differentially expressed genes.

Dramatic effects were seen on gene expression profiles after lead poisoning: 2698 among the 13381 expressed genes, including 68 exhibiting over 50% of change in expression

levels. (20%, false discovery rate (FDR) < 0.0001, 0.214 ± 0.223 mean absolute log2 change ± s.d.) (Fig.1). Among the responders, 2038 genes were upregulated after lead treatment, among which nervous system development and neurogenesis were the topmost enriched gene ontology (GO) categories (Fig.2). On the other hand, among the 660 genes downregulated upon Pb exposure, developmental growth and synaptic target recognition were among the most enriched GO categories (Fig.2). These results were consistent with our expectation, since only *Drosophila* heads were collected on sample preparation and the neurotransmitters at the synaptic levels has long been considered as the main targets for lead neurotoxicity (Baranowska-Bosiacka I., 2012). Genes that are metal responders, like Metallothionein B, C, D and E, and neuro-related genes like Nacalpha, dhd, and RpS5b were among the strongest responders. N-Methyl-D-Aspartate 1 (NMDA1) and its Receptors (NMDAR1 and NMDAR2), previously identified as Pb target at the synapse level (Marchetti and Gavazzo, 2005; Baranowska-Bosiacka I., 2012), were also among the differentially expressed genes (NMDA1: logFC=7.809, FDR=0.014; NMDAR1: logFC=1.004, FDR=0.005; NMDAR2: logFC=-1.150, FDR=0.004).

Fig.1. Lead (Pb) Treatment Altered the Gene Expression Levels among *Drosophila Melanogaster* Male Head Samples. MA plots for change in gene expression (n=2698) comparing Pb-treated (n=79) with control-treated samples (n=79). M= $\log_2$(P/C), A= ($\log_2$(C)+$\log_2$(P))/2, where P: Pb-treated FPKM values; C: control FPKM values. Red dots: genes expression profiles were not significantly changed; Cyan dots: genes expressions were significantly changed (0.214 ± 0.223 mean log2 fold changes s.d, FDR<0.0001).

Fig.2. Gene Ontology Enrichment Analysis of Lead Treatment in the *Drosophila Melanogaster* Male Head Samples. Gene Ontology (http://geneontology.org/) was used to detect over represented GO categories in RNA-seq data (FDR <0.0001). Y-axis shows the minus logarithm of each significant GO ID's p-value (after Bonferroni correction for multiple testing). Significant GO IDs among upregulated genes after Pb exposure were colored in red and GO ID among downregulated genes in green. GO IDs related with synapses and neuronal functions were highlighted in **bold**.



Identification of Cis- and Trans- eQTLs

After identifying genes that were affected by Pb treatment, we worked on identifying expression quantitative trait loci (eQTLs)— the genomic region with genetic variants that affect gene expression levels. In most eQTL studies (Ruden et al., 2009; Mangravite et al., 2013), SNPs were used to represent the genotype. However, in our study, each sample was a mosaic of the eight parental lines (A1-A8) (details in Method) and we used directly the information provided by the DSPR— the genetic contribution by the parental genotypes, which means the parental line a certain genomic region of the offspring was inherited from. With this type of genotype information, the eQTL was defined as a genomic location where gene expressions were associated with differential parental contributions.

The readily available DSPR R package was designed for single gene eQTL search (http://wfitch.bio.uci.edu/~dspr/Tools/Tutorial/index.html); therefore, we re-structured it to allow multiple gene eQTL searches (see Methods). Using the newly modified R program, we computed the LOD score to quantify the likelihood of association between the genomic locations and the gene expressions, and 1000 permutations were run to estimate the threshold of statistical significance (see Methods). In total, 1,536 cis-eQTLs (FDR ≤ 10%) and 952 trans-eQTLs were identified (1000 permutation threshold at 0.05). Among the genes with cis-eQTLs, 774 genes were shared among control and lead-treated, along with 547 control-specific and 215 lead-specific (Fig.3A).

One example of the control-specific cis-eQTL was shown in Fig.4A. In this example, left two panels showed all the lodscores for the gene CG2807 at each of the 11768 evenly divided genomic locations for both control and Pb-treated status. The high peak in the control panel indicated strong association with the corresponding genomic location on the x-axis but this signal disappeared after lead treatment (Fig.4A, second to the left panel). We also noticed that

the strongest peak overlapped with the transcript location (green dashed line); this indicated that the gene CG2807 is not only a control-specific eQTL but also a cis-eQTL.

In order to further explore the parental contribution of the genomic location at the highest peak in control, we sub-grouped the gene expression levels according to their parental genotypes at this associated location (Chr2L: 1,770,000) and used a boxplot to show their expression levels (Fig. 4B, right two panels). From the figure, samples originally from A2, A3 and A4 have significantly higher expression levels than samples from A5, A6 and A7 in control, while this difference was greatly reduced in Pb-treated samples. This allelic heterogeneity was also widely observed in DSPR female head eQTL study (King et al., 2014).

In addition to the control-specific cis-eQTLs, there is an example of Pb-specific cis-eQTL in Fig.4C. On the other hand, among the 952 genes with trans-eQTLs, 50 genes were shared among control and lead-treated, along with 645 control-specific and 257 Pb-treated (Fig.3B, one examples of control-specific trans-eQTL in Fig.4D and another Pb-specific trans-eQTL in Fig.4E).

Fig.3. Venn Diagrams Demonstrating Overlaps between Control-specific eQTLs and Pb-specific eQTLs. (A) gene numbers for cis-eQTLs. (B) gene numbers for trans-eQTLs.

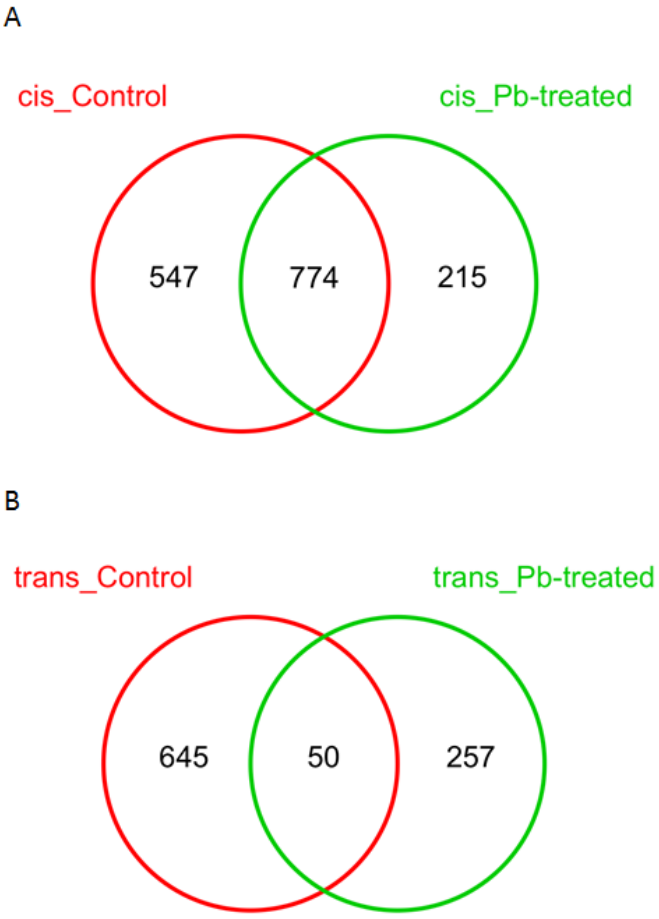
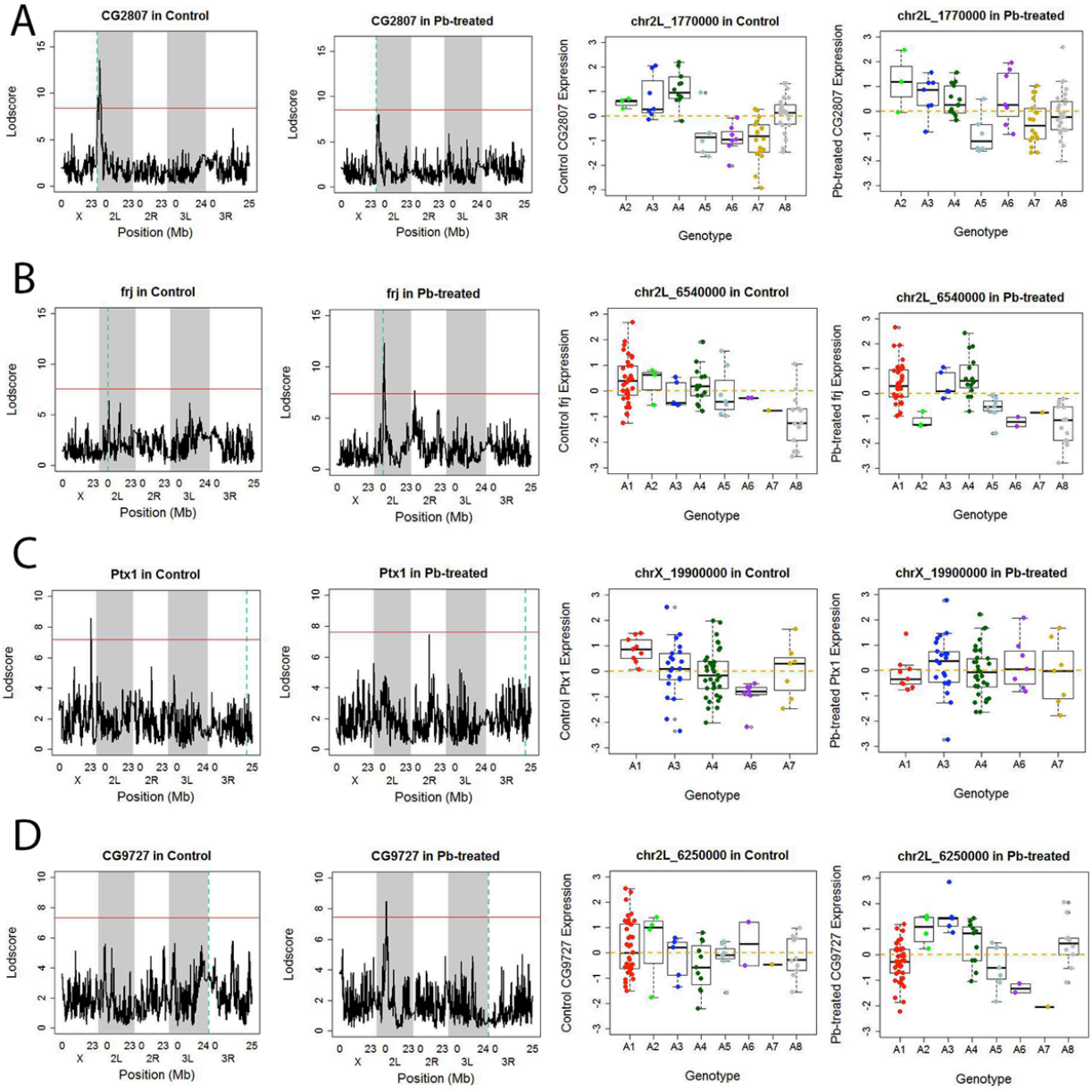


Fig.4. Examples of cis- and trans- eQTLs. (A) shows one example of control-specific cis-eQTL. In the left two panels, the x-axis represents the *Drosophila* genomic locations and y-axis represents the lodscore of the gene. The red horizontal line indicates the threshold for *p*-value to be 0.05 after 1000 permutation test. The green dash vertical line indicates the location of the gene. If it overlaps with the peak, which suggests strong correlation between the gene and the corresponding location, it is referred as a cis-eQTL, meaning the regulator is near the downstream gene. Since this phenomena only occurred in control data but not in Pb-treated one, this genomic location Chr2L: 1,770,000 is a control-specific cis-eQTL for gene CG2807. In the right two panels, association of the Chr2L: 1,770,000 location, which has the highest lodscore in control samples, with quantile normalized CG2807 expression levels following control (*p*-value< 0.001) and Pb-treated (not significant). Samples originally from A2, A3 and A4 parental lines exhibited higher expression levels, while samples from A5, A6 and A7 parental lines showed lower expression levels in control. After lead was introduced, this phenomenon disappeared. (B) is one example of Pb-specific cis-eQTL. (C) is one example of control-specific trans-eQTL. (D) is one example of Pb- specific trans-eQTL.



After searching for all possible associations among 13,381 gene expression profiles against 11,768 genomic locations, we visualized the entire significant associations with an eQTL map (Fig.5A for control panel and Fig.5B for Pb-treated panel). Each of the colored dots represents one significant correlation between the genetic location displayed on the x-axis and the gene on the y-axis (significance at 0.05 for 1000 permutation). There was a prominent diagonal band in both control and lead-treated map. It showed that transcript locations of these genes were similar to their eQTL locations, thus the cluster of genes belong to cis-eQTLs. On the other hand, there were also some distinguished vertical bands, indicating any one of these genomic regions was associated with a list of genes across the entire chromosome. These genomic loci with a high density of eQTLs are usually called trans-eQTL hotspots (Joo et al., 2014; King et al., 2014) or trans-eQTL bands (Rockman M.V., 2006). In total, we got 6 control and 7 Pb-treated trans-eQTL hotspots (Fig.6, Table 1). Among them, 4 were Pb-sensitive hotspots: 3 Pb-specific and 1 control-specific (Table 1, highlighted in red).

Fig.5. eQTL Map. All significant associations were shown in an eQTL map with eQTL locations (genomic loci) on x-axis and transcript locations (gene loci) on y-axis. (A) Associations for control samples only. Each of the green dots indicates a significant association between the corresponding eQTL location and the gene at the transcript location. (B) eQTL Interactive Map for Pb-treated samples only. Each of the red dots indicates a significant signal. (C) eQTL Interactive Map combining both control and Pb-treated samples. Shared significant signals were marked as brown, with Pb-specific signals as red and control-specific ones as green.

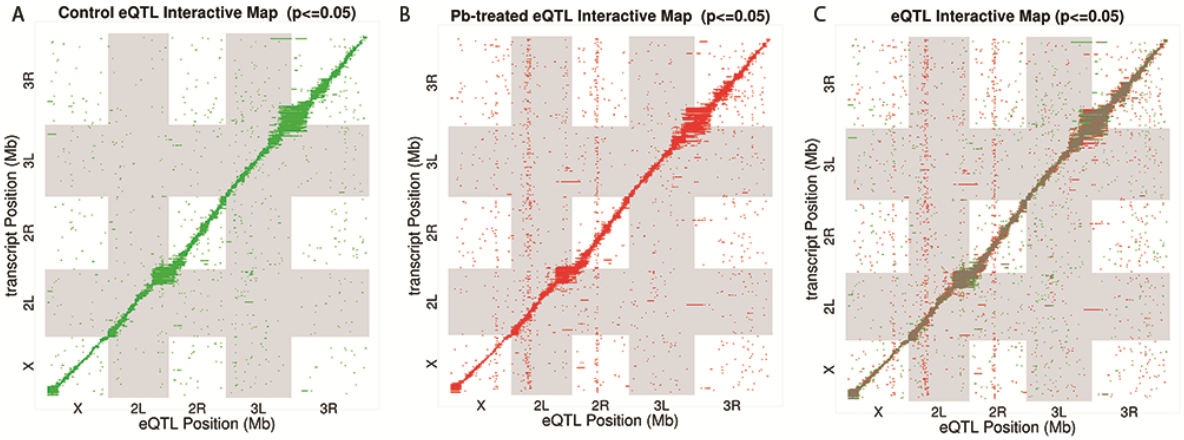


Fig.6. The Distribution of Trans-eQTL Hotspots among the *Drosophila* Genome. 6 for control (green, marked above the genomic axis) and 7 for Pb-treated (red, marked under the genomic axis) trans-eQTL hotspots were detected in total. Chromosome 4 and heterochromatic chromosomes were excluded due to lack of genomic information from the DSPR group. None of the trans-eQTL hotspots were detected in Chromosome X. Numbers on top/ bottom of each trans-eQTL hotspot represented the number of associated genes at the peak locus.

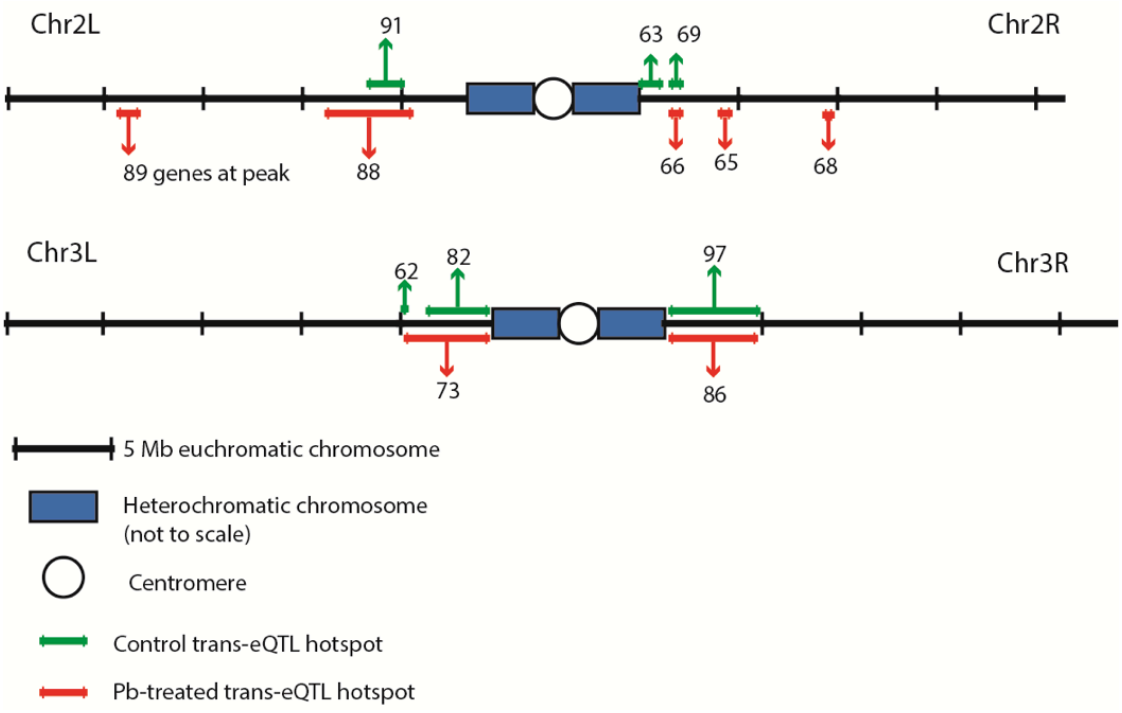


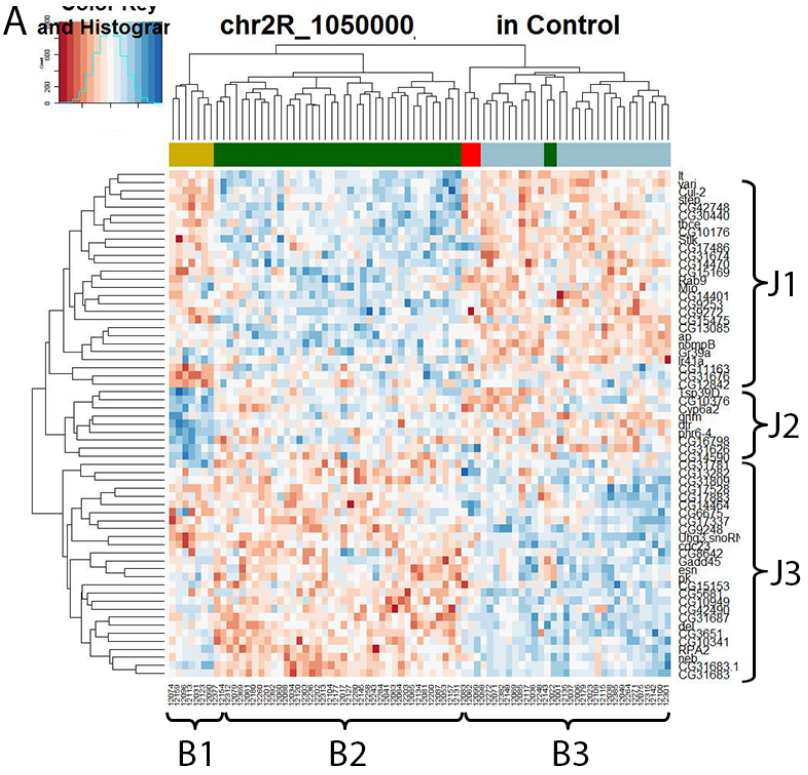
Table 1. Detailed Information about the Pb-responsive Trans-eQTL hotpots.

status	chr	start	end	Length (Mb)	peak location	#genes @ peak	*p*-value
control	chr2L	18510000	20590000	2.08	20330000	91	<0.0001
control	chr2R	100000	1090000	0.99	1050000;1080000	63	0.001
control	chr2R	1370000	1820000	0.45	1600000	69	0.0001
control	chr3L	20170000	20200000	0.03	20170000-20200000	62	0.001
control	chr3L	20780000	24360000	3.58	22680000;22690000……	82	<0.0001
control	chr3R	90000	5780000	4.88	5090000	97	<0.0001
Pb-treated	chr2L	6130000	7060000	0.93	6250000	89	<0.0001
Pb-treated	chr2L	16790000	21240000	4.45	20290000	88	<0.0001
Pb-treated	chr2R	1590000	1990000	0.4	1640000;1650000	66	0.0001
Pb-treated	chr2R	4540000	5110000	0.57	4570000	65	0.0001
Pb-treated	chr2R	9290000	9400000	0.11	9290000;9320000	68	0.0001
Pb-treated	chr3L	20160000	24360000	3.20	22360000; 22370000; 22380000	73	<0.0001
Pb-treated	chr3R	90000	5590000	4.69	590000; 600000; 610000	86	<0.0001

Genetic Dissection of the Trans-eQTL Hotspots

To further explore the mechanism of the trans-eQTL hotspot, we first looked at the stable trans-eQTL hotspots, meaning signals that were present in both control and Pb-treated (one example in Fig.7). A heatmap was used to show the regulations of the associated genes in the presence or absence of chronic lead exposure. To do this, expression profiles of all the associated genes were extracted into a subset and the hierarchical clustering analysis (Eisen et al., 1998) was used to display the expression patterns (Fig.7). In Fig.7, all associated genes were arranged based on the similarity of their expression pattern making genes (right list) divided into three groups (J1, J2 and J3) and samples (bottom list) into three groups (B1, B2 and B3). Interestingly, the segregation of samples according to the expression pattern actually overlapped with the genetic contribution of the parental genotypes (the color-coded bar above the heatmap): samples originally from A4 (dark green) showed lower J3 expression pattern and higher J1+J2 expression pattern, while samples from A5 (light blue) had the exact opposite pattern.

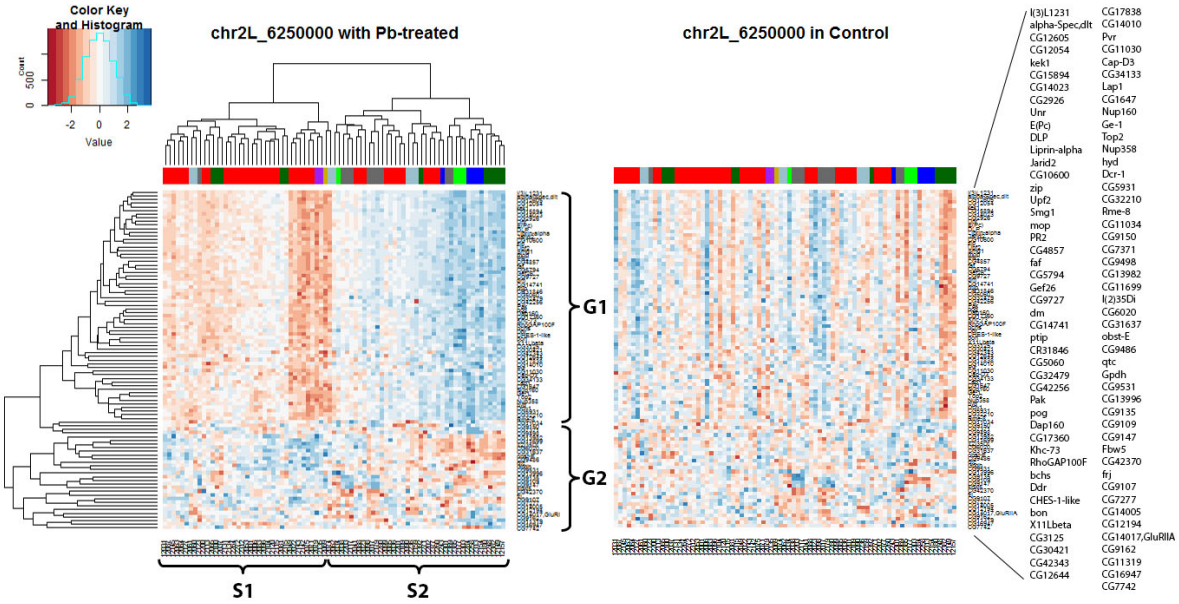
Fig.7. Stable Trans-eQTL Hotspot at Chr2R: 1,050,000 (p-value < 0.05 at 1000 permutation threshold). Hierarchical clustering analysis was done according to the expression profiles of the Chr2R: 1,050,000 associated genes (p-value < 0.05). On the heatmap plotted by using the control expression data, the bottom list indicates all the sample names and the right list indicates all the associated genes. Color-coded bar above the heatmap and below the dendrogram indicates the original parent of each sample listed at the bottom at this specific location. Color legend in the color-coded bar: red: A1, green: A2, blue: A3, dark green: A4, light blue: A5, purple: A6, gold: A7, darkgray: A8.



Not only did we find this correlation between expression traits and parental contribution at the stable trans-eQTL hotspots, but also in lead-responsive ones. Here, as an example for Pb-sensitive trans-eQTL hotspots, we used the one that located at Chr2L: 6,250,000 that contains and contained 89 associated genes. The hierarchical clustering analysis was also used to present expression data graphically (Fig.8) and it showed that all the hotspot-associated genes could be divided into two groups (G1, G2) and all the samples could be divided into two groups (S1, S2) according to the gene expression profiles. It appeared that genes from G1 exhibited lower expression levels in sample group S1 but higher in S2, while genes belong to the G2 presented the opposite phenomena. With the help of the color coded bar on top of the heatmap, a clear segregation was shown among samples based on their original parents: the expression pattern of samples from A2 (green) and A3 (blue) was in contrast with that of samples from A6 (purple) and A7 (gold).

However, not all parents show differential influences on downstream genes, such as A1 (red) and A4 (dark green). This suggested that different strains of *Drosophila* species might respond differently to Pb exposure and this might be reflected by regulation of some key eQTL loci and their downstream gene expression levels. Compared with the Pb-specific trans-eQTL hotspot that contained 89 associated genes, only 28 associated genes were observed at the same genomic locus in control status. If we kept the order of gene list and sample list in the Pb-treated heatmap (Fig.8 left panel) but replaced with corresponding control data, we would find an entire disruption of the expression pattern present upon lead exposure (Fig.8 right panel). This confirmed that this hotspot at Chr2L: 6,250,000 locus is a lead-responsive trans-eQTL hotspot.

Fig.8. Trans-eQTL hotspot at Chr2L: 6,250,000 (*p*-value < 0.05). Hierarchical clustering analysis was done according to the expression profiles of the Chr2L: 6,250,000 associated genes (*p*-value < 0.05). On the left heatmap plotted by using the Pb-specific trans-eQTL, the bottom list indicates all the sample names and the right list indicates all the associated genes. The heatmap on the right was created by maintaining the order of the sample names and associated gene names in the Pb-treated plot on the left but replacing with control expression data. The expression patterns formed in Pb-treated data were totally disrupted after replacing with the control data, suggesting this trans-eQTL hotspot could only be observed in expression levels after lead exposure. Color-coded bar above the heatmap and below the dendrogram indicates the original parent of each sample listed at the bottom at this specific location. Color legend in the color-coded bar: red: A1, green: A2, blue: A3, dark green: A4, light blue: A5, purple: A6, gold: A7, darkgray: A8.



In order to take a deeper look at the genes associated with the Chr2L: 6,250,000 genomic location upon lead exposure, we searched for their GO enrichment categories (Attrill et al., 2015). Those 89 associated genes could actually be categorized into 5 groups: neuro-related, metal-related, response to stimuli and immune system, other metabolic processes, and unknown function (Table 2A). We noticed that genes in G1 were mainly related to neuronal function (18 out of 61, 30%), while genes in G2 were mostly metabolic processes (18 out of 28, 64%) (see details in Table 2A). We also recognized that genes in G1 (52 out of 61, 85%) were Pb-specific trans-eQTLs at Chr2L: 6,250,000 (Table 2B, examples in Fig.4D & Fig.9A, B), while genes in G2 (22 out of 28, 78%) were more likely in close proximity of the eQTL locus and were stable cis-eQTLs (Table 2B, examples in Fig.9C, D). Among the rest of the signals, a few were Pb-specific cis-eQTLs (Table 2B, one example from G1 in Fig.9E and one example from G2 in Fig.4B).

Table 2. (A) GO Function Categories for the Associated Genes at Chr: 6,250,000. (B) eQTL Types for Genes at G1 and G2.

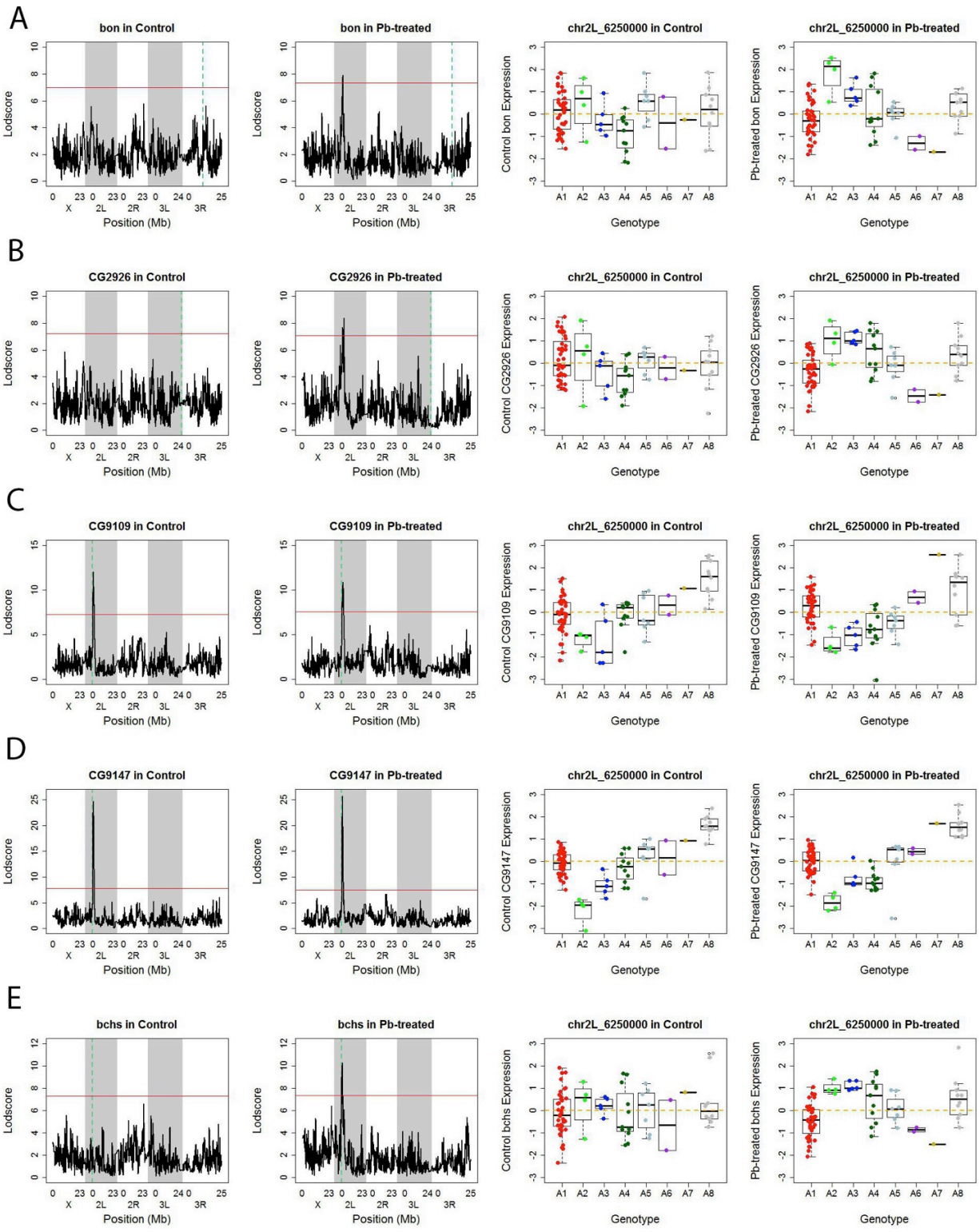
(A)

Group	Neuron-related	Metal ion binding	Response to stimuli, immune system, cell death, DNA damage	Other metabolic processes	Behavior (mating, etc.)	Unknown function	total
G1	18 (30%)	7 (11%)	12 (20%)	13 (21%)	1 (1%)	10 (16%)	61
G2	1 (3%)	1 (3%)	2 (7%)	18 (64%)	1 (3%)	5 (18%)	28
total	19 (21%)	8 (9%)	14 (16%)	31 (35%)	2 (2%)	15 (17%)	89

(B)

Group	Both control & Pb-treated Cis-eQTL at Chr2L:6,250,000	Pb-specific Trans-eQTL at Chr2L:6,250,000	Pb-specific cis-eQTL at Chr2L:6,250,000	others	total
G1	3 (5%)	52 (85%)	2 (3%)	4 (6%)	61
G2	22 (78%)	4 (14%)	2 (7%)	0	28
total	25 (28%)	56 (63%)	4 (4%)	4 (4%)	89

Fig.9. Examples of Trans-eQTL Associated Genes. (A) shows one example of lead-specific trans-eQTL in G1 family. In the left two panels, the x-axis represents the *Drosophila* genomic locations and y-axis represents the lodscore of the gene. The red horizontal line indicates the threshold for *p*-value to be 0.05 after 1000 permutation test. The green dash vertical line indicates the location of the gene. (B) is another example of lead-specific trans-eQTLs in G1 family. (C) is one example of stable cis-eQTLs. (D) is another example of stable cis-eQTLs. (E) is an example of the Pb-specific cis-eQTLs in G1 family.



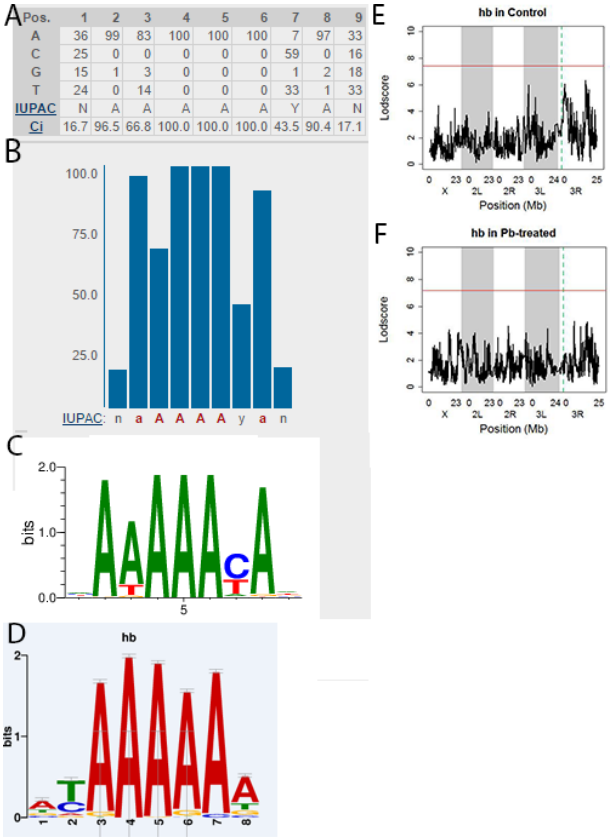
It has long been proposed that a transcription factor is a natural candidate for being the regulator of the trans-eQTL hotspots (Yvert et al., 2003). It has been hypothesized that the eQTL location may have influence over the affinity of a certain linked transcription factor and the transcription factor has multiple associations with downstream genes. This hypothesis serves as a perfect candidate explanation for trans-eQTL hotspots. However, it has been controversial ever since and not many studies have discussed about it. Yvert *et al.* (Yvert et al., 2003) mentioned that few *trans* variations have strong correlations with known or predicted transcription factors in their yeast research.

In our case, we used the CoreSearch in German Genomatix software, a tool that could define unknown common motifs from a set of unaligned DNA sequences (Wolfertstetter et al., 1996), to search for common nucleotide motifs of the downstream genes associated with the same trans-eQTL hotspot. 100 promoter sequences of all the 89 downstream genes at Chr2L: 6,250,000 were extracted by Gene2Promoter, a tool in Genomatix to provide promoter sequences of all genes annotated from the genomes (http://www.genomatix.com). As a result, AAAAAYA (Y: C or T) was the common motif generated after searching among the 100 promoter sequences (Fig.10). We used Tomtom, a software for quantifying similarity between query motif and motifs from the exisiting databases to see whether this identified motif would match with any of the previously discovered ones (Gupta et al., 2007).

It turned out that hunchback (hb) has a shared motif with the AAAAAYA (*p*-value= 8.11e-04, Fig.10D). hb, as a protein-coding gene, locates at Chr3R and its main function involves generations of neurons and neuroblast fate determination (Isshiki et al., 2001; Tran et al., 2010; Attrill et al., 2015). hb, as a transcription factor, has been shown to be necessary for regulation of the first-born glial cell fates, leading a sequence of transcription factors at the cell fate specification stage (Isshiki et al., 2001). Interestingly, hb was not detected to be an eQTL by

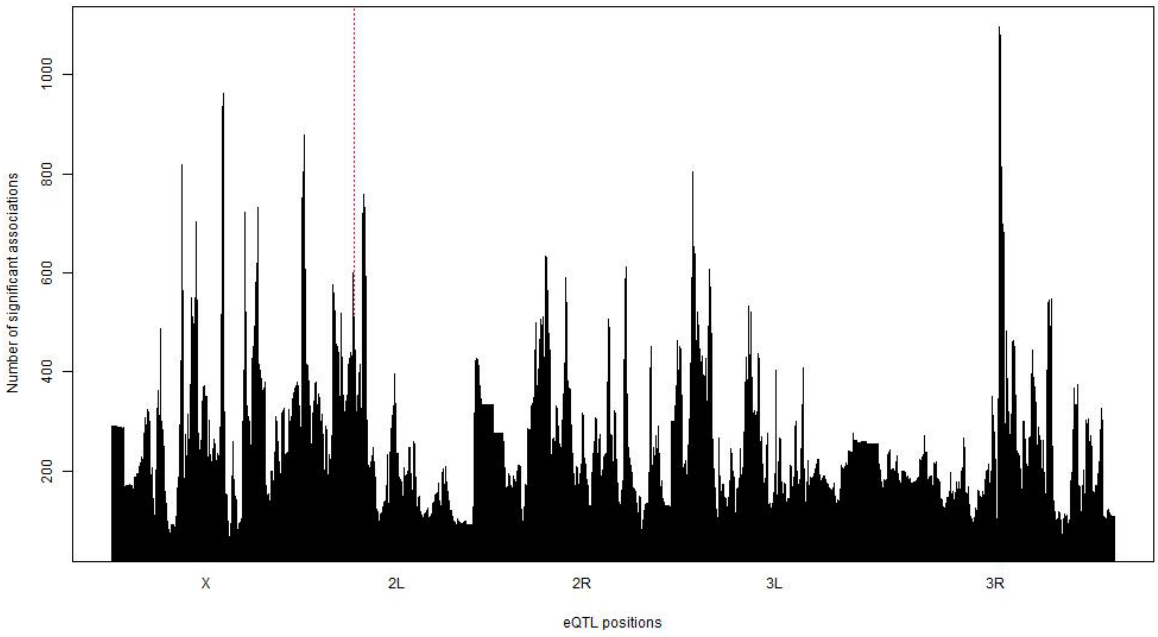
itself (Fig.10E, F). There were also no known protein-protein interactions between hb and any of the associated genes at the trans-eQTL hotspot.

Fig.10. Common Motif Shared by Associated Genes with the Trans-eQTL Hotspot. (A) The distribution and frequencies of each basepairs. (B) The nucleotide distribution matrix shows the nucleotide frequencies observed in aligned binding sites of the corresponding transcription factor. Basepairs in red indicate high information content, which means the matrix exhibits a high conservation (> 60%) at this position. Genomatix made the basepairs in capital letters denote the core sequence used by MatInspector. The core sequence of a matrix is defined as the (usually 4) highest conserved, consecutive positions of the matrix. (C) Common motif of the associated genes at the trans-eQTL hotspot. (D) The common motif detected in (C) resembles hb with *p*-value to be 8.11e-04. (E) - (F): The lodscore plot of hunchback.



Our next consideration was to verify the existence of the trans-eQTL hotspot at Chr2L: 6,250,000. For eQTL analysis, one of the major concern is the expression heterogeneity (EH) (Pickrell et al., 2010; Joo et al., 2014). We used Surrogate Variable Analysis (SVA) to test whether the trans-eQTL hotspot would still be considered as significant after controlling EH (Pickrell et al., 2010). As a result, the trans-eQTL hotspot at Chr2L: 6,250,000 locus was still among the strong peaks after the SVA processing (Fig.11). This indicated that the Pb-sensitive trans-eQTL hotspot could be considered as a true positive result.

Fig.11. Trans-eQTL Hotspots Targeted after the SVA. Numbers of significant associated genes identified after SVA process (*p*-value ≤ 0.05). The lead-responsive trans-eQTL hotspot at Chr2L: 6,250,000 (red dashed line) was still one of the strong peaks after SVA processing.



After controling for the EH, another way to validate the trans-eQTL hotspots is by using another set of Pb-treated expression data and see if similar expression pattern existed as well. Our lab had another set of lead-treated *Drosophila* microarray data back in 2009 (Ruden et al., 2009). There are some differences between the microarray data and current RNA-seq data. For example, the microarray dataset was a two-way eQTL analysis, meaning the samples were originally from two parents (comparison in experimental design in Table 3). Also, in the previous study we analyzed RNA isolated from whole adult males, whereas, this study analyzed RNA isolated from adult male heads.

We applied our current eQTL-detection method to the microarray expression dataset and as a result, we found that marker 27B, which is the closest to Chr2L: 6,250,000, showed no significant trans-eQTL signals (data not shown). However, when we extracted all expression levels of the available microarray probes for the 89 genes identified by the current RNA-seq data and ran for the hierarchical clustering heatmap at 27B, we found similar expression segregation patterns as previously. In the left panel (Pb-treated) of Fig.12, genes could be roughly divided into three groups: g1, g2 and g3 according to the similarity of the expression pattern. We noticed that most genes from g1 (10 out of 12, 83.3%) and g3 (29 out of 34, 85.3%) belong to RNA-seq G1 group (Fig.8), while most genes from g2 (20 out of 29, 70.0%) were the same as G2. The right panel of Fig.12 was created by maintaining the order of the samples at the bottom and associated genes on the right but replacing Pb-treated expression data with control ones.

Similar to the results in the RNA-seq data, this heatmap produced by the microarray data showed that the expression patterns formed in Pb-treated status were disrupted in control, suggesting the genes forming expression patterns could only be observed in expression levels after lead exposure, but not in control. This also indicated that the segregated expression

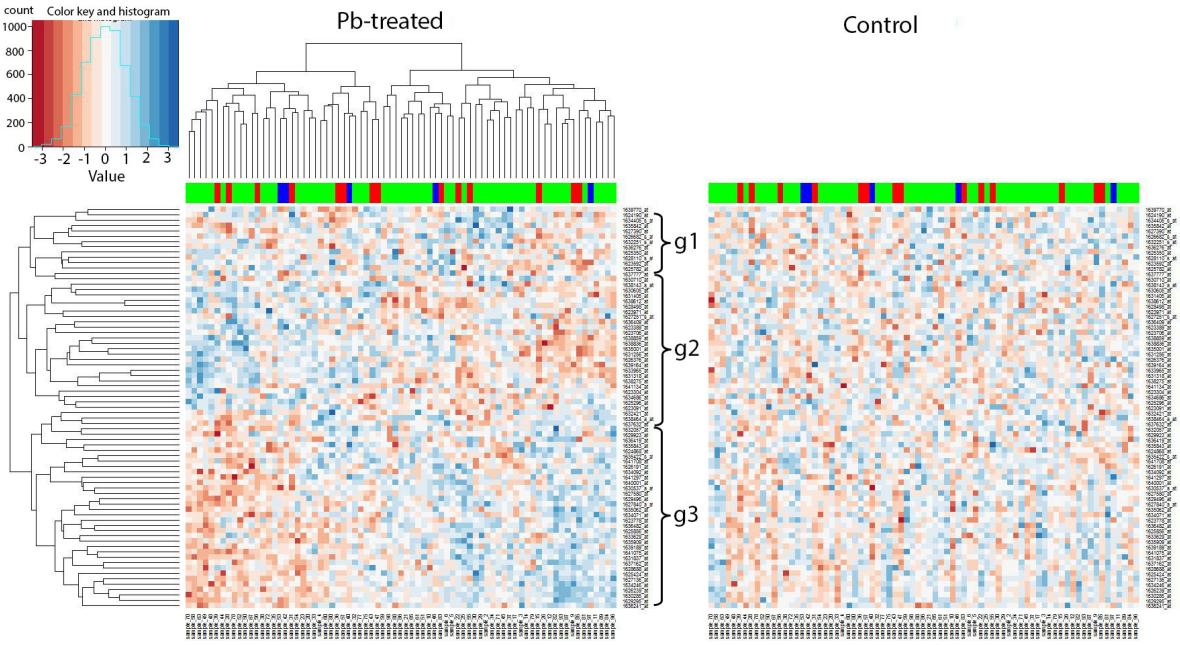
patterns at the eQTL locus were found in both RNA-seq data and in microarray data. However, the color-coded bar above the microarray heatmap, which indicated the original parent of each sample at this 27B location, showed no significant difference based on parental choice. This showed that the two parental lines—Oregon R and Russian 2B (no overlaps with the eight parental lines used in RNA-seq), have no differential influence over associated gene expression profiles at this 27B (Chr2L: 6,250,000) locus, explaining why this location was not detected as a trans-eQTL hotspot in the microarray experiment in the first place. This also demonstrates that the eight-way analysis which includes more genetic variations is more robust and should include more eQTLs than the two-way analysis.

Table 3. Experimental Design and Result Comparison between the Microarray in 2009 and RNA-seq in 2012.

	Microarray in 2009			RNA-seq in 2012		
Geno Types	two-way			eight-way		
Genomic information	SNPs			Genomic origins		
Numbers of Genomic locations	92 (markers)			11768		
Numbers of Samples in each condition	75			79		
Numbers of Genes detected	~14000 (18,952 probesets)			13381		
Condition	mixing 250μM lead acetate in the fly food as lead exposure					
Sample collected	Whole male *Drosophila*			Male *Drosophila* head		
Technique used	Microarray			RNA sequencing		
The criteria for significant eQTLs	The 1000 permutation LOD scores have a P-value of less than 0.0001			The 1000 permutation LOD scores have a P-value of less than 0.05		
Definition of cis-eQTL	Significant eQTLs within a 5cM sliding window			Significant eQTLs within 1Mb		
Definition of trans-eQTL	Significant eQTLs outside the 5cM sliding window			Significant eQTLs outside the 1Mb		
Definition of trans-eQTL hotspots	96 probesets in a 5cM window			50 genes in a 1Mb window		
Numbers of cis-eQTLs detected* Control-only|overlap|Pb-only	405	440	544	547	774	215
Numbers of trans-eQTLs detected* Control-only|overlap|Pb-only	948	357	1191	645	50	257

*Please note that the numbers of cis- and trans- eQTL detected in microarray assay or in RNA-seq assay are not comparable due to differential genomic information.

Fig.12. Trans-eQTL hotspot at 27E (nearest to Chr2L: 6,250,000). Expression levels of candidate downstream genes detected with RNA-seq were extracted from the microarray data and hierarchical clustering analysis was performed accordingly. In the left panel (Pb-treated), genes could be roughly divided into three groups: g1, g2 and g3. Most genes from g1 and g3 belong to G1 from the RNA-seq data, while most genes from g2 were the same as G2. The right heatmap was created by keeping the order of the sample names and associated gene names in the Pb-treated plot on the right but replacing with control expression data. The expression patterns formed in Pb-treated data disappeared in control data, suggesting the genes forming expression patterns could only be observed in expression levels after lead exposure. Color-coded bar above the heatmap and below the dendrogram indicates the original parent of each sample listed at the bottom at this specific location. Color legend in the Color-coded bar above the heatmap: red: Oregon R (ORE), green: Russian 2B (2B), blue: heterozygous. No segregation based on the parental origin was seen, suggesting these two parental lines do not differ in expression levels and this also explains why this has not been detected as a trans-eQTL hotspot in the microarray data.



Further Analyses on the Microarray Data

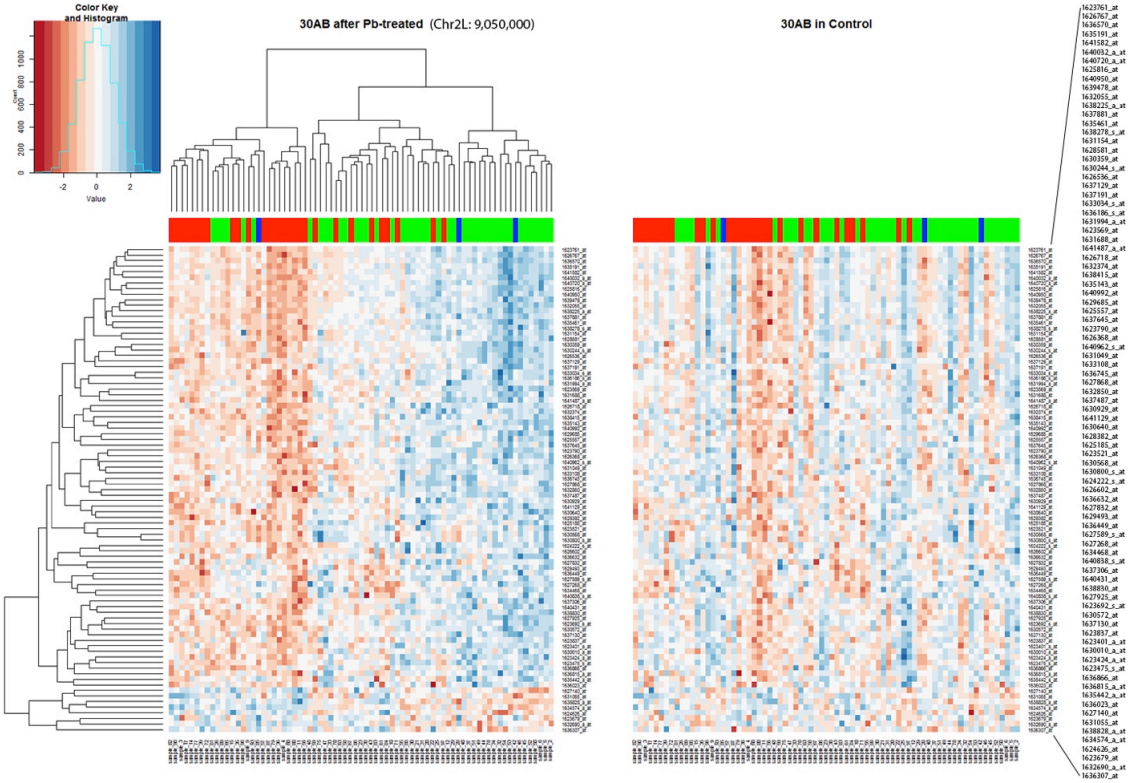
In 2009, we used R/qtl (Broman et al., 2003) to estimate the presence of eQTLs on the set of lead-treated *Drosophila* microarray (Ruden et al., 2009). As a result, our lab found 5 control trans-eQTL hotspots and 7 lead-treated ones (Experimental design and Result Comparison shown in Table 3) (Ruden et al., 2009). Due to the limited amount of the genotype information, which contained only 92 markers throughout the entire genome, our lab was, at the time, unable to shorten the trans-eQTL hotspots within 5cM. In order to further explore the mysterious trans-eQTL hotspots, our lab performed this RNA-seq analysis on Pb-treated *Drosophila* male heads.

After identifying eQTLs by using the method provided by the DSPR group, we applied the same one on the previous microarray data. As a result, we found 7 overlapping results with that detected in the 2009 paper (Table 4). Among the overlaps, one lead-responsive trans-eQTL hotspots was at the cytological location of 30AB and was within 1.7 Mb from the trans-eQTL hotspot we found at Chr2L: 6,250,000 in the RNA-seq data. In order to see whether these two trans-eQTL hotspots are somewhat related, we used a venngraph to compare their associated genes. Although both Chr2L: 6,250,000 in RNA-seq and 30AB in microarray showed nice expression segregation between samples originally from their parents (Fig.8 for Chr2L: 6,250,000 RNA-seq data & Fig.13 for 30AB microarray data), we found no overlaps between the two lists of associated genes. We concluded that the two trans-eQTL hotspots searched in the different expression data sets should actually be considered as two different ones.

Table 4. Seven out of the Twelve Trans-eQTL Hotspots were Reproduced by our Current Method to Target Trans-eQTLs.

	Control		**Pb-treated**	
	Results in 2009	Results using current method	Results in 2009	Results using current method
3E			✓	✓
27B	✓	✓		
30AB			✓	✓
50DF	✓			
57F			✓	
63A			✓	
65A			✓	✓
70C	✓			
72A	✓			
73D	✓	✓	✓	✓
77E		✓	✓	✓

Fig.13. Significant Trans-eQTL Hotspot Detected after Reanalyzing Microarray Data.



Preliminary Deficiency Validation Test

After the detection of trans-eQTL hotspot, we used w^{1118} flies that cause deficiency in the proposed trans-eQTL area and see if we could detect the differential expression levels of these potential downstream genes. We tried with three genomic regions (Bloomington Deficiency Kits stk# 24124, 9605, 8835, 8674, and 9341), most of which are shorter than 1Mb long (Cook et al., 2012; Roote and Russell, 2012). RNA-seq was then performed after lead exposure to test the potential downstream effects. Unfortunately, none of the deficiencies showed any significant influence on the potential downstream genes. In the future analyses, we would use the highly efficient genome modification method—CRISPER-CAS9 (Clustered regularly interspaced short palindromic repeats) technique to pinpoint the trans-eQTL hotspots in the eight parental strains instead of using the wildtype (Yu et al., 2013).

Discussion & Conclusion

Here we investigated gene expression in *Drosophila* heads from 79 different 8-way RILs to identify lead-responsive cis- and trans- eQTLs. We also went one step further to provide the further evidence for the existence of the controversial lead-responsive trans-eQTL hotspots. With the help of the clustering analyses, we confirmed that the expression traits of the progeny could be sub-grouped based on the genetic contributions of the parents.

There are several advantages of eQTL analyses using RNA-seq compared with using microarrays. First, RNA-seq avoids the possibility of false positive reads due to the limitation of the microarray technology (Xiao et al., 2002; Fadiel and Naftolin, 2003). Second, the abundant genotype information, which includes 11768 underlying parental haplotype structures, makes it more likely to pinpoint the eQTL loci, while the previous microarray eQTL analysis only contain 92 genomic markers, each of which was at least 5cM wide (Ruden et al., 2009). Third, this time

we have more parental lines involved (eight-way versus two-way), which should include more genetic variations that are present in *Drosophila* species.

Another criterion worth mentioning is the sample size. Due to the financial concerns, we could only afford to analyze 79 RNA-seq samples with no replicates. We might have identified more trans-eQTL hotspots if we included more samples.

The DSPR group mapped genome-wide expression variation in 2014. They generated an eQTL interactive map and found two trans-eQTL hotspots. However, they did not have an exposure model, and the trans-eQTL did not overlap with the hotspots identified in our study. This is not surprising since their experiments included more genetic differences: heterozygotes from parental population groups A and B (both A1, A2 and B1, B2) (King et al., 2014), while we only considered a subset of homozygotes in one parental subgroup, which is A2. Furthermore, they worked with heterozygotes due to inbreeding depression (King et al., 2014) and we did not have that problem when processing the fly lines. Therefore, our genetic information contains only A1-A8 and each of our samples was a homozygous mosaic of the eight parental lines, while samples used in the DSPR paper were originally from 16 founders A1-A8, B1-B8 (line A8 and B8 are actually the same and therefore were also referred to as AB8) and were heterozygous. In their paper, there was another finding that most of their eQTLs were multi-allelic (King et al., 2014) and same phenomena have been observed in our study.

In contrast to the DSPR group, we included developmental lead poisoning as another perturbation and searched for lead-responsive eQTLs. We have successfully identified lead-responsive trans-eQTL hotspots. We found that some new trans-eQTL hotspots were formed in response to lead poisoning and some existing trans-eQTL hotspots disappeared after lead treatment. The clustering analysis has shown the samples from different parental genetic origins responded differently in downstream gene expression profiles before or after lead exposure.

Previous papers have hypothesized that gene expression profiling patterns associated with trans-eQTL hotspots reflect biological pathways (Wu C., 2008); however, we ended up with no enriched pathways among the associated genes. Our next step will be to identify and knock down genes responsible for the trans-eQTL hotspots and to determine if the expression levels of the proposed downstream genes are influenced. We also plan to include longevity and behavioral test to determine if the differential expression changes in different parental strains could provide a protective mechanism to respond to lead poisoning.

In conclusion, RNA-seq technology is a powerful tool in obtaining genome-wide expression profiles and identifying cis-and trans-eQTLs. The hierarchical clustering analyses display the expression patterns of the eQTL-associated genes and show that they segregate by genotype. We have successfully made progress in understanding how trans-eQTL hotspots alter the susceptibility to lead exposure, opening up a gate towards the mechanisms of trans-eQTL hotspots, as well as the neurotoxicity of lead.

CHAPTER 2: SPLICING QTLS

Introduction

Alternative Splicing

After the discovery of splicing in the Adenovirus *hexon* gene in 1977 (Sambrook, 1977), Walter Gilbert proposed in 1978 that different combinations of exons and introns, namely "alternative splicing" (AS), could produce different mRNA isoforms of a gene (Gilbert, 1978; Modrek and Lee, 2002). The disparity between the expected 150,000 or more human genes and the surprising actual report of under 32,000 later suggested an underestimated role for alternative splicing in the production of an increased variety of mRNAs and proteins (Pennisi, 2000; Venter et al., 2001). It has been estimated that AS is a crucial form of gene regulation affecting about 60-90% of human genes (Modrek and Lee, 2002) and over 40% of *Drosophila* genes (Stolc et al., 2004). Mutations that affect mRNA splicing and AS were also considered to be highly linked with disease occurrences (Singh and Cooper, 2012). It has also been estimated that 15% of human disease mutations lie within splicing sites and 22% of disease-related SNPs may affect splicing (Krawczak et al., 2007; Lim et al., 2011).

The *Drosophila Dscam* gene exemplifies one of the most extreme examples of alternative splicing. *Dscam* (Down Syndrome Cell Adhesion Molecule) is a cell surface protein which gives rise to over 30,000 potential alternatively spliced isoforms in the *Drosophila* nervous system (Schmucker and Flanagan, 2004). The human homologue *DSCAM* was first discovered as a candidate disease gene for the central and peripheral nervous system defects associated with Down syndrome (Yamakawa et al., 1998). The *Drosophila Dscam* was later found to have the extreme structural diversity and is essential for neural circuit assembly (Schmucker et al., 2000; Hattori et al., 2007). Its diversity allowed each neuron to have a unique pattern on its cell membrane, which made self-recognition possible (Lawrence Zipursky and Grueber, 2013;

Armitage et al., 2015). It has also been shown that *Dscam* regulated interactions between neurons through isoform-specific homophilic binding or repulsion (Wojtowicz et al., 2004; Tadros et al., 2016). Its role in the insect cellular immune system has also been suggested since 2005 (Watson et al., 2005). Even after years of study, many questions remain unanswered, such as how *Dscam* mRNA isoforms are selectively expressed and how homophilic interactions are translated into binding or repulsing responses during neurogenesis (Schmucker and Flanagan, 2004).

Splicing Quantitative Trait Locus (sQTLs)

A quantitative trait locus (QTL) is a sequence of DNA (the locus) that is associated with variation in a phenotype (the quantitative trait) (Miles and Wayne, 2008). Splicing QTLs (sQTLs) distinguish relative splicing isoform abundance. Significant sQTLs could be categorized into two groups: cis- and trans-sQTLs. In the previous chapter, we described the identification of cis- and trans-eQTLs in *Drosophila* (Ruden D.M., 2009a). In this chapter, we focus on cis- and trans-sQTLs. By definition, cis-sQTLs refer to genetic variants that affect the splicing event of a locus only on the same haplotype, while trans-sQTLs affect both haplotypes (Benzer, 1955; Hasin-Brumshtein et al., 2014). Therefore, cis-sQTLs tend to be “local”, near the locus of the gene encoding the regulated transcript, while trans-sQTLs tend to be “distant”, away from the regulator (Benzer, 1955; Hasin-Brumshtein et al., 2014). For the purposes of this paper, an sQTL is defined as genetic variants that are associated with changes in the splicing ratios of transcripts (Monlong et al., 2014).

Our previous expression QTL (eQTL) study was a gene expression analysis on a lead-treated *Drosophila* model (Ruden D.M., 2009b). In our original study, we identified both lead-responsive cis-eQTLs and trans-eQTLs using Affymetrix *Drosophila* gene expression microarrays (Ruden D.M., 2009b). We also identified a QTL linked with developmental

behavioral effects of lead exposure (Hirsch et al., 2009). In this current study, in order to identify sQTLs, which cannot be identified with standard gene-expression microarrays, we used the RNA-seq analysis. This original eQTL study was on a recombinant inbred line (RIL) set derived from two parental lines. In this current study, we used another set of RILs provided by the *Drosophila* Synthetic Population Resource (DSPR) (King et al., 2012). This time, we used the same RNA-seq data to explore Pb-responsive sQTLs. We found hundreds of positive candidate sQTLs, among which *Dscam1* was one of the most significant sQTLs both on the exon level and on the transcript level.

Methods

Genotyping and Sample Collection

The genotyping and sample collection protocol were the same as the methods section shown in Chapter 1. The RNA-seq data are publicly available on the NCBI GEO accession: GSE83141.

Expression Quantification

RNA-seq reads were mapped to the UCSC/dm3 *D.melanogaster* references genome (track: Flybase Genes) using TopHat (Karolchik et al., 2004; Kim et al., 2013).

We first used the coverageBED function in BEDTools to quantify raw counts of the exons and transcripts (Quinlan and Hall, 2010). In our second trial, htseq was used (Anders et al., 2014).

ANOVA Test

All analyses were performed in R. After calculating division of exon reads to its corresponding transcript reads, quantile normalized, and confounding factors were removed by PCA (n.pc=3), we used the following model to target sQTLs:

$$H0: Y = \mu + \varepsilon$$

$$H: Y = \mu + \sum Gi * E + \varepsilon$$

Where μ is the grand mean, *Gi* is the ith parental genotype probability, E represents two environmental conditions: control or lead-treated.

For transcript expression levels, transcript reads were quantile normalized, and confounding factors were removed by PCA (n.pc=4). Expression data for genes that have more than one isoform were subgrouped (n=3975).

$$H0: Y = \mu + \sum Gi * Iso + \sum Gi * E + \varepsilon$$

$$H: Y = \mu + \sum Gi * E * Iso + \varepsilon$$

Where μ is the grand mean, Iso is isoform type, *Gi* is the ith parental genotype probability, E represents two environmental conditions: control or lead-treated.

Definition of the Significant sQTLs

After obtaining all the p-values indicating the likelihood of association between each genomic location and each transcript, qvalue function in R (library: qvalue) was used to transform the *p*-values into FDRs and we defined *p*-values $\leqslant$ 0.0001, corresponding FDRs $\leqslant$ 0.39, as significant signals. We next randomly shuffled the entire set of sQTL results 10,000 times. From each of the randomization, the highest number of associated genes detected for a significant sQTL was recorded. The *p*-value for the trans-eQTL hotspot at Chr3L:18,810,000 locus was generated based on the distribution of the total 10,000 recordings.

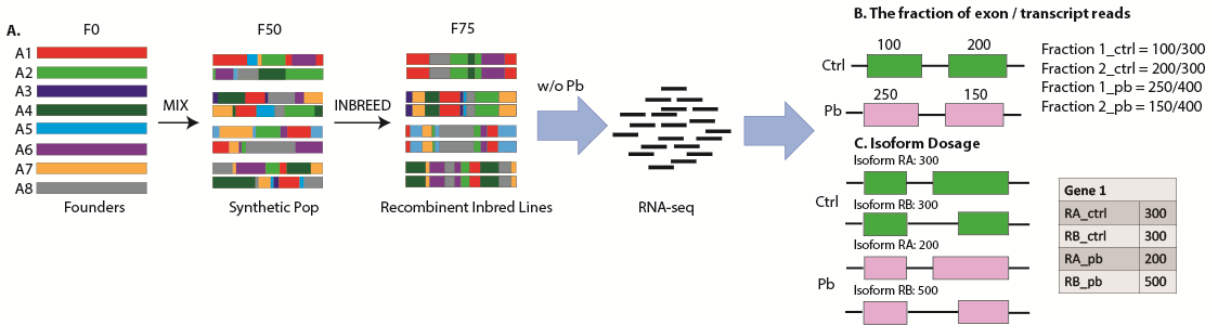
GO Enrichment Analysis

Upload/Convert ID tool from Flybase.org was used to convert the annotation symbols into official symbols (Tweedie et al., 2009). GOseq (Young et al., 2010; Young et al., 2012), which is an R package for conducting GO analysis for RNA-seq data, was used for the GO enrichment analysis for the differentially expressed genes upon Pb exposure and GO categories of "Molecular Function" and "Biological Process" were selected.

Results

After identifying eQTLs affected by Pb treatment, we tried to identify sQTLs – the genomic regions where genetic variants affect splicing events. In most sQTL studies (Lappalainen et al., 2013; Battle et al., 2014; Kurmangaliyev et al., 2015; Ongen and Dermitzakis, 2015), SNPs were used to represent the genotype information. However, in our study, each fly line was a mosaic of the eight parental lines (A1-A8) (Fig.14A) and the genetic contribution by the parental genotypes, meaning the parental line a certain genomic region of the offspring was inherited from. With this type of genotype information, the sQTL was defined as a genomic locus where gene expressions were associated with differential parental contribution.

Fig.14. Workflow in Search for sQTLs. A) The design of the recombinant inbred lines. Strains were initiated with eight founder strains A1-A8 with a diverse geographic origin. In the first generation, lines were intercrossed with each other and 10 F1 flies per genotype per sex were mixed together to establish the next generation. This mix went on until the 50th generation when flies were separated and Inbreeding continued for another 25 generations, leading to a total of ~1600 completed recombinant inbred lines. After samples were treated with or without Pb treatment, RNA-seq analysis was performed and two methods were used to target sQTLs: B) the fraction of exon reads to transcript reads and C) Isoform dosage.

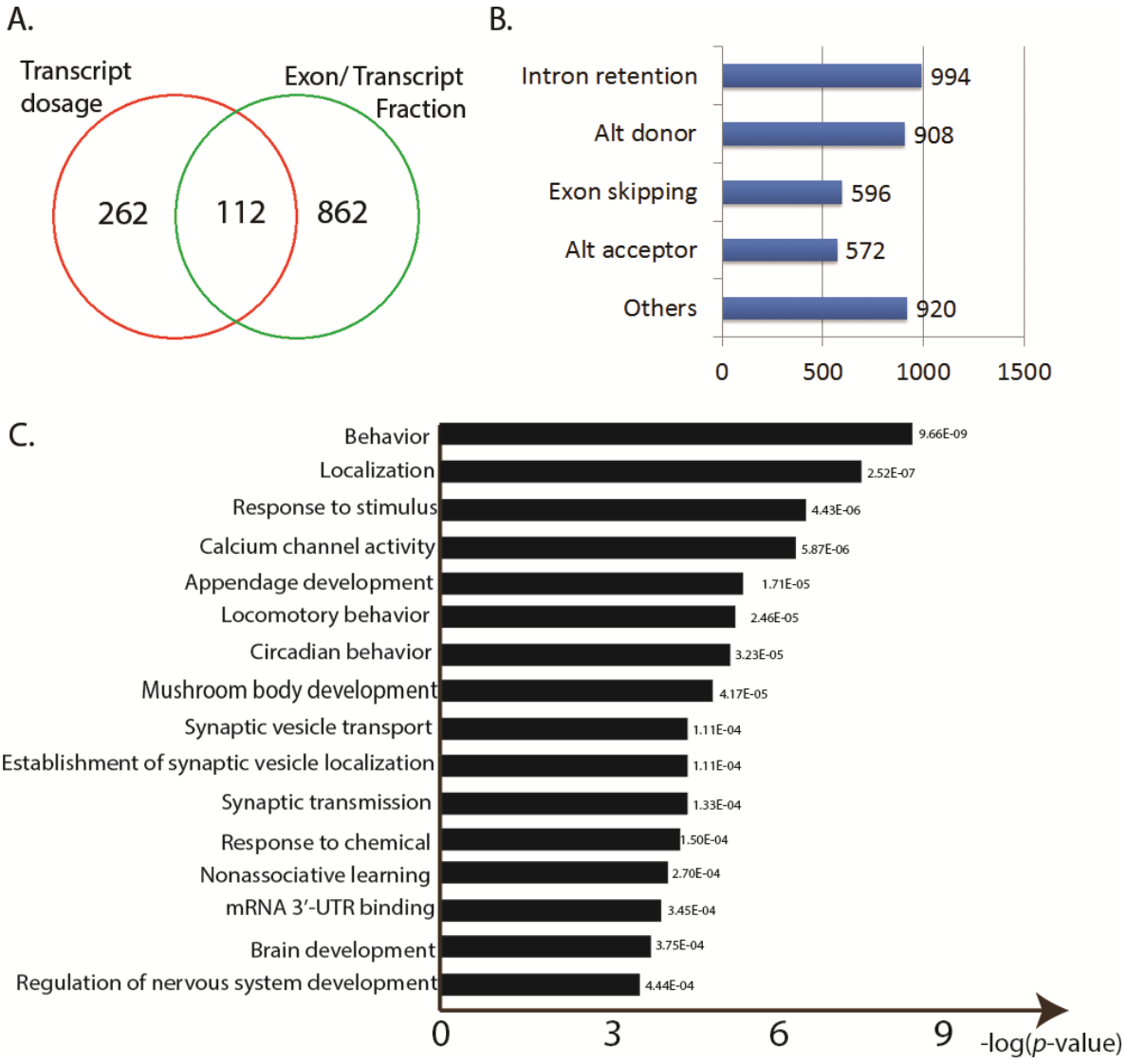


Reads were mapped by Tophat2 (Trapnell C., 2012; Kim et al., 2013) and quantified to exons and transcripts by Htseq (Anders et al., 2014). We used two ways to detect sQTLs: 1) use the fraction of reads in a transcript that falls in a given exon as the quantitative trait and run for fraction QTLs; 2) directly target the differential transcript dosage in the same gene (Fig.14C, D, see Methods). Exon/transcript fraction considers changes within each exon, while the second method captures events where both exon reads and transcript reads change in the same direction, probably missed by the first method.

Here, we used ANOVA analysis to detect Pb-responsive sQTLs (see Methods) (Hoaglin and Welsch, 1978). In total, we obtained 974 Pb-responsive sQTLs by calculating exon/transcript fraction and 374 by isoform dosage, with 112 shared ones (*p*-value <0.0001, FDR <0.39) (Fig.15A). We then used the Alternative splicing transcriptional landscape visualization tool (ASTALAVISTA) (Foissac and Sammeth, 2007) to determine the types of events represented by the entire set of sQTLs (Fig.15B). The four main AS types were intron retention (n=994), Alternative donor splicing (n=908), exon skipping (n=596) and alternative acceptor splicing (n=572) (Fig.15C).

The identified sQTLs were also run through a GO enrichment analysis. The top enriched categories were “behavior” (*p*-value = 9.66E-09) and “response to stimulus” (*p*-value = 4.43E-06) and “calcium channel activity” (*p*-value = 5.87E-06) (Fig.15C). Neural developmental related GO categories were also among the most significant: “mushroom body development” (*p*-value = 4.17E-05), “synaptic vesicle transport” (*p*-value = 4.17E-05), “non-associative learning” (*p*-value =2.70E-04), “brain development” (*p*-value =3.75E-04), and “regulation of nervous system development” (*p*-value =4.44E-04) (Fig.15C). Other over-represented GO categories include “locomotory behavior” (*p*-value = 2.46E-05), “response to chemical” (*p*-value = 1.50E-04), and “mRNA 3’-UTR binding” (*p*-value =3.45E-04) (Fig.15C).

Fig.15. Properties of the Pb-responsive sQTLs. A) Venn graph showing the overlapping sQTLs targeted between two methods. B) Numbers of different AS events found among the identified sQTLs. C) GO enrichment analysis for the sQTLs.



One of the most significant sQTLs is *Dscam1* linked with Chr3L:2,790,000 (*q*-value<1.11E-08). Transcript expressions of samples originally from the A3 parent were altered significantly, while others were not (Fig.16). In samples that were inherited from A3 parent at Chr3L:2,790,000 locus, RT, RU and RW isoforms were upregulated after developmental lead exposure, RV was downregulated, while RAE was remained steadily. This suggested that A3 strain responded to lead poisoning differently from the rest of the parents by altered usage of the various isoforms. In order to explore deeper into the exon usage in *Dscam1*, we used the Integrative Genomics Viewer (IGV) (Thorvaldsdóttir et al., 2013), which is a popular visualization tool for integrated genomic data. We noticed that reads for exon 7, 8, 10 and 11 were increased after lead treatment in A3 samples, while in other samples the expression change was in the reversed direction (Fig.17). However, not all exons were affected in the same way. For exon 18 and 19, read counts in all samples were upregulated after lead exposure (Fig.17). It is possible that differential exon usage resulting from lead neurotoxicity is a part of the compensatory pathway after lead poisoning, but future research is needed to tackle this problem.

Fig.16. Differential *Dscam1* Isoform Expression Upon Lead Exposure among Samples Originally from Chr3L: 2,790,000. Green boxes represent control samples. Red boxes represent Pb-treated samples.

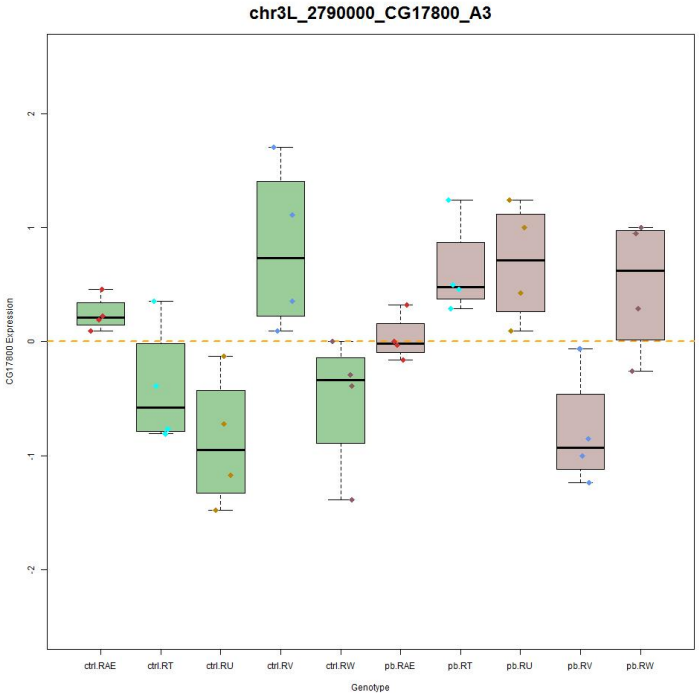
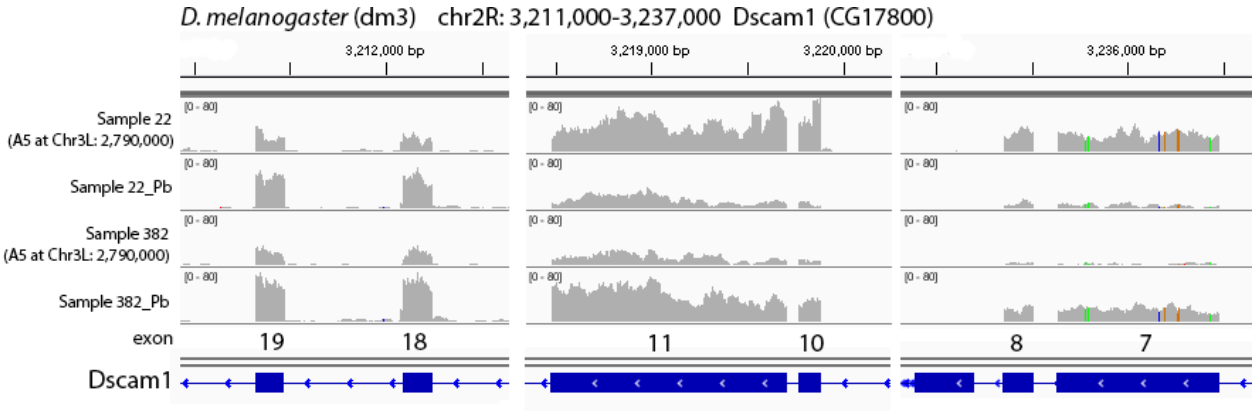
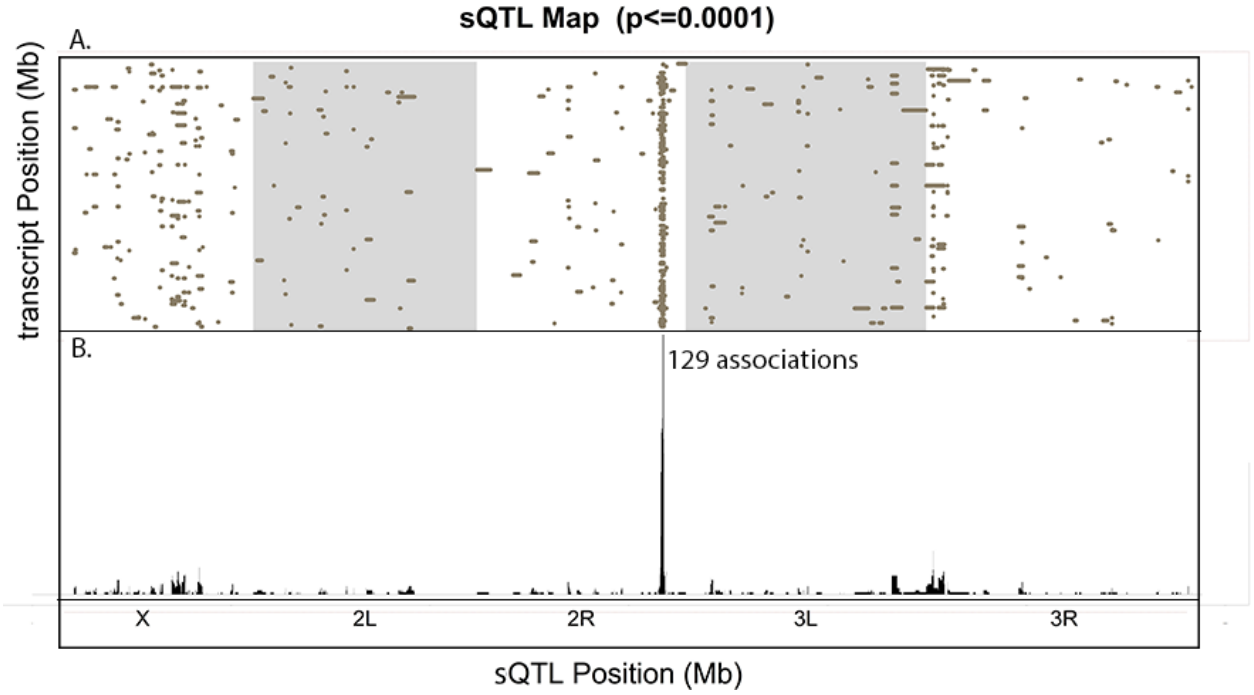


Fig.17. Visualization of Different Exon Usage by RNA-seq w/o Lead Treatment. Sample 22, one of the examples that were originally from A5 strain at Chr3L: 2,790,000, has reduced expression of exon 7, 8, 10 and 11 with lead treatment, while sample 382, one of the examples that were originally from A3 at the same locus, has increased expression of the same exons. However, not all exons share the same feature. For exon 18 and 19, read counts were all increased after lead exposure.



We then visualized the distribution of all the significant associations with a comprehensive sQTL map (Fig.18). Each dot represents one significant association between the genetic locus shown on the x-axis and the transcript on the y-axis (FDR <0.39). Among the scattered dots, there was one prominent vertical band consisting of a high density of dots, demonstrating that the genomic locus was associated with a cluster of transcripts regardless of their loci. This is similar to what has been observed in many eQTL studies, where genomic loci linked with abnormally high numbers of eQTLs were called trans-eQTL hotspots (Joo et al., 2014; King et al., 2014) or trans-eQTL bands (Rockman M.V., 2006). More interestingly, this trans-sQTL hotspot at Chr3L:18,810,000 locus is Pb-responsive (*p*-value < 1E-10).

Fig.18. The Comprehensive sQTL Map. A) All significant signals were shown and each dot represents one association between the eQTL location on the x-axis and transcript location on the y-axis. B) The sum of all the associated dots for each genomic location.



Here, a heatmap by hierarchical clustering analysis was used to visualize the expression of the downstream genes w/o lead poisoning at this trans-sQTL hotspot (Fig.19) (Eisen et al., 1998). In Fig.19, the entire set of correlated transcripts in each of the 79 samples was clustered based on the similarity of their expression profiles. Interestingly, there was a clear segregation when the control data was used (Fig.19A), while this expression pattern disappeared by replacing with the Pb-treated data (Fig.19B), where both column and row have exactly the same order as the control heatmap. Additionally, the color-coded bar on top of each heatmap represents each sample's (sample name at the bottom) original parent at this Chr3L: 18,810,000 locus.

Samples originally from grey parent (A8) showed distinct expression patterns from green (A2) parents in a majority of the associated transcripts. This suggested that in normal condition, A8 have a different expression patterns compared to A2; however, this difference was suppressed or disrupted after lead poisoning. We also noticed that "cation channel activity" was the topmost GO category (*p*-value= 8.73E-06) for the list of 129 genes. And there is only one protein-coding gene bypassing this Chr3L: 18,810,000 locus—CG14073. This is also among the 129 assoicated sQTLs: Fig.20 showed differential expression of isoform RB after lead exposure for samples originally from A2. Currently, there is no known molecular function for this gene but experimental evidence has shown that it is involved in the wing disc dorsal/ventral pattern formation (Bejarano et al., 2008). When we expanded the search to the genomic region of Chr3L:18,800,000~ 18,820,000, the region was estimated to have at least 50 donor sites and 92 acceptor sites by the Berkeley *Drosophila* Genome Project (Reese et al., 1997).

Fig.19. The Heatmap of Trans-sQTL Hotspot. A) All associated signals represented in control expression data; B) Heatmap generated by maintaining both column and row name order in control but replacing with Pb-treated data. The list of ordered isoforms on the right side was in Table 5. Color legend: red: A1 parent, green: A2, blue: A3, dark green: A4, light blue: A5, purple: A6, gold: A7, darkgray: A8.

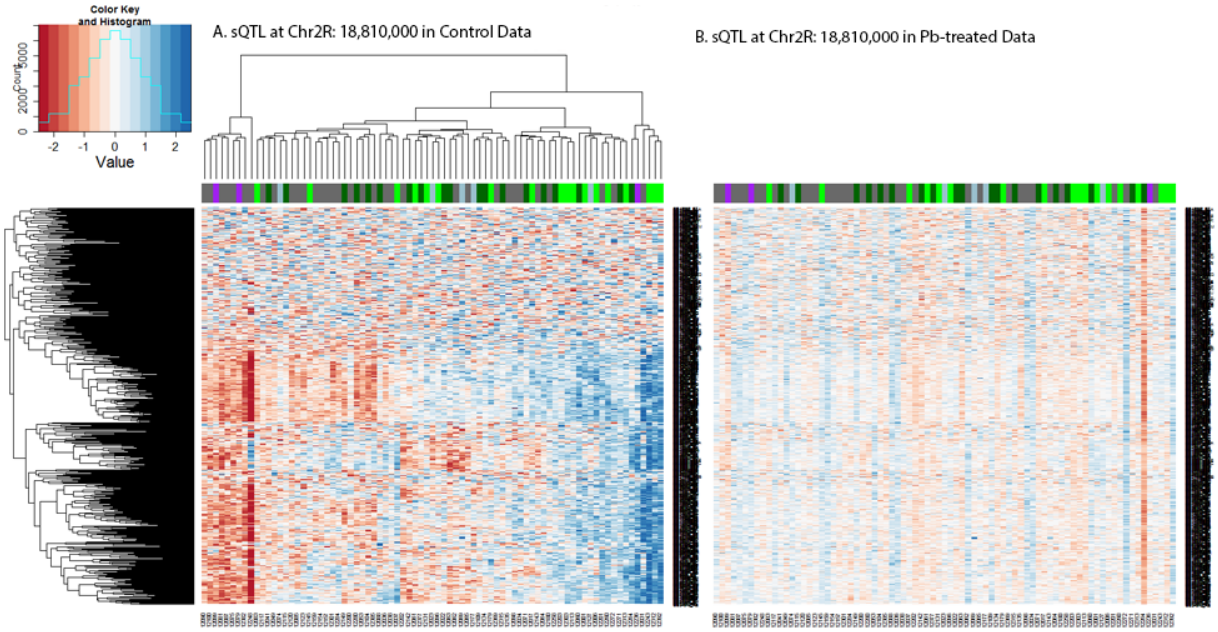
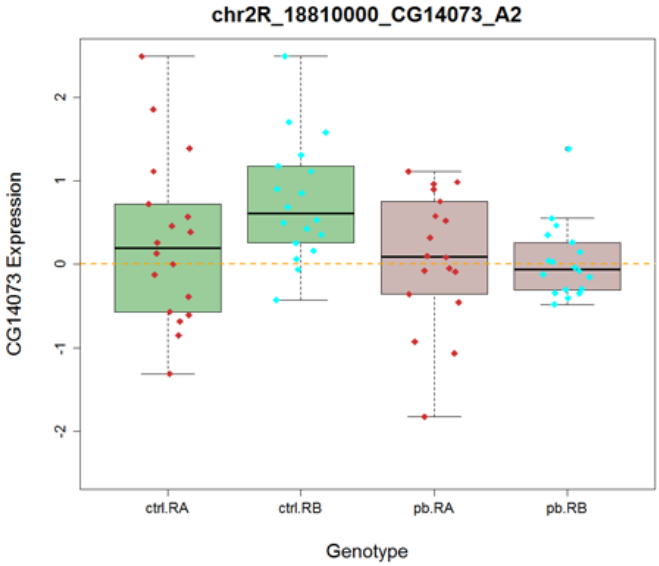


Fig.20. Slight Decreased Expression of CG14073-RB after Lead Exposure for Samples Originally from A2.



Discussions & Conclusions

In this paper, we used two methods to detect Pb-responsive sQTLs. The first one, which is the fraction of exon reads to the transcript reads, searches splicing events on exon levels and all types of genetic–related AS events that cause change either in exon or in transcript after lead treatment will be selected. The other method compares the transcript counts among isoforms. It is on the transcript levels and will select those have differential isoform dosage after lead exposure.

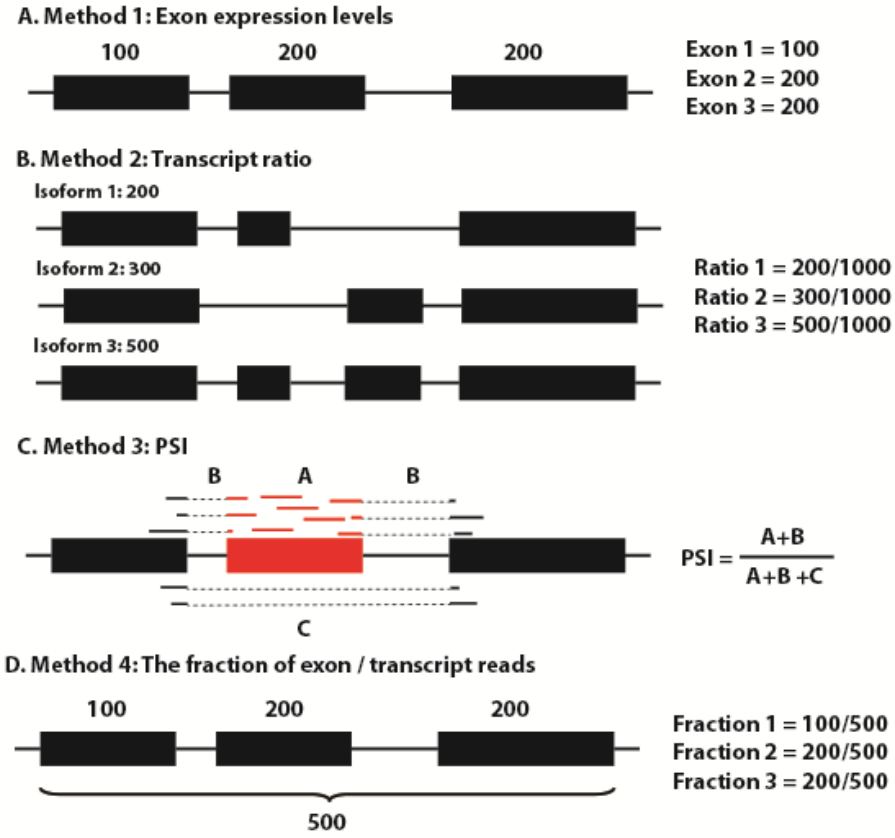
The combination of two methods resulted in 1236 significant Pb-responsive sQTLs.

Generally, to target sQTLs, there are five major approaches (Fig.21): 1) simply use exon expression profiles as the quantitative trait and this could also be referred to as exon QTLs (Montgomery et al., 2010; Lappalainen et al., 2013; Gymrek, 2014); 2) the proportion of each transcript quantification of the sum of all transcripts per gene (Lappalainen et al., 2013; Battle et al., 2014; Gymrek, 2014); 3) Percent spliced in (PSI) (Lappalainen et al., 2013; Zhao et al., 2013; Gymrek, 2014; Kurmangaliyev et al., 2015); 4) use the fraction of reads in a gene that falls in a given exon as the phenotype, as used in this paper (Pickrell J.K., 2010; Gymrek, 2014); 5) Multivariate approaches, such as sQTLseekeR (Gymrek, 2014; Monlong et al., 2014).

The sQTLseekeR is a multivariate model called for each gene consisting of the relative abundance of each transcript (Monlong et al., 2014). It calculated the variability of splicing ratios of a gene across samples by using a MANOVA-like distance-based approach and then compared the variability of the splicing ratios within genotypes with the variability among genotypes. We have run our data through the sQTLseekeR pipeline. However, no significant association was returned. One of the potential explanation for this result is that the sQTLseekeR was originally designed to incorporate the genotypes as SNP information (0 for ref/ref, 1 for ref/mutated, 2 for mutated/mutated), but our genotype data, which represent the original parents

of evenly distributed genomic locations (A1-A8, representing 8 parents), pose potential challenges to process the data (Monlong et al., 2014).

Fig.21. Major Methods for Detecting sQTLs.



Most of the sQTL studies were performed in human cell lines. One study by Kurmangaliyev et. al., which has claimed to be the first sQTL study in *Drosophila*, searched for genotype-specific alternative splicing donor/ acceptor sites by using 81 *Drosophila* hybrid strains generated by crossing natural populations to a single inbred reference line (Kurmangaliyev et al., 2015). They found 59 AS donor/ acceptor events by performing 120,240 association tests (Kurmangaliyev et al., 2015). In our study, we detected 1236 Pb-responsive AS events by running >1,255,422,008 association tests (106681 exon/transcript reads *11768 genomic loci) and our detection should not only include alternative donor/ acceptor splicing but also other types of AS events.

The identification of *Dscam1* as one of the most significant sQTLs helps to further understand the isoform usage and changes after Pb exposure. Schumucher et al. have shown that the overexpression of one *Dscam* isoform resulted in strong dominant phenotypes in mushroom body neurons (Schmucker et al., 2000). In 2004, Schumucher and Flanagan suggested either that different neurons express different *Dscam* isoforms or that isoforms need to be present at a precise concentration or a certain development time period (Schmucker and Flanagan, 2004).

The diversity of *Dscam* isoforms has been shown to allow neurons having differential patterns on its cell membrane and interacting through isoform-specific hemophilic binding (Wojtowicz et al., 2004; Lawrence Zipursky and Grueber, 2013; Armitage et al., 2015; Tadros et al., 2016). In our analysis, we found expression alterations in *Dscam1* both on the exon level and on the transcript level. However, we have few ideas on how to interpret: why such changes occurred after lead exposure and how this could contribute to the neural developmental damage. Future studies might consider combining sQTL analysis with other molecular and cellular experiments in order to better understand the lead neurotoxicology.

Our study is the first to link sQTL analysis with an environmental toxin in *Drosophila*. However, there are limitations in our study: 1) the RNA-seq data were prepared as 50 bp paired-end. However, 100 bp paired-end reads were considered to enhance splicing junction detection significantly (Chhangawala et al., 2015). 2) Both methods in this paper rely on known transcript annotation and transcript level quantifications.

In conclusion, we have shown that sQTL analysis is a useful way in understanding alternative splicing mechanisms and the neuro-toxicology of environmental toxin. We discovered widespread genetic variation affecting the splicing events. Our characterization of causal regulatory variation sheds light on the mechanisms of neurotoxicity of lead, and allows us to infer putative causal variants for hundreds of environmental toxic-associated loci.

Table 5. The List of Ordered Isoform after the Hierarchical Clustering Analysis

[1] "CG12052_RZ" "CG17800_RAA" "CG32464_RJ" "CG32464_RB" "CG17800_RS" "CG31716_RE" "CG31284_RA"
[8] "CG1070_RB" "CG32491_RL" "CG4527_RC" "CG16765_RB" "CG17800_RBB" "CG10693_RB" "CG31716_RC"
[15] "CG1725_RJ" "CG17800_RAI" "CG42260_RB" "CG17800_RAK" "CG34373_RH" "CG33555_RD" "CG34416_RK"
[22] "CG17800_RU" "CG5020_RJ" "CG10693_RH" "CG6282_RA" "CG17689_RB" "CG42275_RD" "CG17800_RAM"
[29] "CG42281_RE" "CG10693_RA" "CG33232_RA" "CG17800_RA" "CG1725_RC" "CG12052_RU" "CG31689_RA"
[36] "CG32490_RC" "CG17800_RB" "CG4894_RA" "CG1725_RD" "CG5640_RC" "CG4821_RB" "CG32538_RA"
[43] "CG12052_RC" "CG12052_RN" "CG10693_RC" "CG16765_RD" "CG32491_RG" "CG42275_RE" "CG9660_RD"
[50] "CG14619_RB" "CG5020_RH" "CG9660_RC" "CG17838_RD" "CG17838_RA" "CG33555_RH" "CG32464_RR"
[57] "CG33555_RF" "CG31349_RC" "CG10706_RD" "CG10693_RG" "CG32158_RE" "CG17800_RP" "CG17800_RD"
[64] "CG12052_RG" "CG17800_RK" "CG32490_RO" "CG9059_RA" "CG32498_RG" "CG17800_RT" "CG1693_RB"
[71] "CG17800_RAY" "CG17800_RAC" "CG10693_RE" "CG12052_RB" "CG17800_RAO" "CG33555_RB" "CG5020_RC"
[78] "CG4527_RB" "CG17800_RX" "CG16765_RJ" "CG10693_RQ" "CG34412_RI" "CG17800_RC" "CG17800_RAS"
[85] "CG32158_RB" "CG4527_RE" "CG31349_RB" "CG32538_RC" "CG34416_RF" "CG17800_RE" "CG7029_RC"
[92] "CG34412_RF" "CG10693_RO" "CG17800_RQ" "CG12052_RT" "CG1725_RI" "CG17800_RL" "CG42275_RG"
[99] "CG34365_RF" "CG4894_RD" "CG17800_RAQ" "CG32498_RD" "CG17838_RE" "CG32464_RH" "CG1228_RD"
[106] "CG1725_RB" "CG6671_RA" "CG42403_RB" "CG8566_RC" "CG33183_RB" "CG7893_RA" "CG12052_RY"
[113] "CG17800_RAE" "CG17800_RBE" "CG32498_RM" "CG10618_RF" "CG32490_RH" "CG32490_RI" "CG12052_RE"
[120] "CG17800_RAP" "CG10693_RK" "CG32529_RC" "CG10693_RN" "CG17800_RAN" "CG17800_RV" "CG32498_RK"
[127] "CG17800_RZ" "CG32498_RI" "CG32464_RU" "CG17800_RAV" "CG32158_RF" "CG7125_RE" "CG16765_RK"
[134] "CG42275_RF" "CG17838_RH" "CG1725_RE" "CG33989_RE" "CG17800_RAZ" "CG3136_RC" "CG17800_RAW"
[141] "CG32490_RN" "CG10706_RC" "CG7145_RD" "CG32498_RO" "CG17800_RN" "CG17800_RAB" "CG17800_RBH"
[148] "CG17800_RAF" "CG17800_RAD" "CG17800_RAX" "CG32490_RL" "CG33232_RC" "CG31536_RC" "CG15072_RC"

[155] "CG8174_RB" "CG5659_RC" "CG6998_RD" "CG5685_RB" "CG2822_RC" "CG33232_RD" "CG32464_RG"
[162] "CG8669_RB" "CG42275_RI" "CG31349_RI" "CG32490_RJ" "CG6703_RD" "CG8385_RA" "CG2822_RB"
[169] "CG17800_RM" "CG17800_RG" "CG16765_RC" "CG32464_RN" "CG1725_RH" "CG11680_RA" "CG32498_RJ"
[176] "CG18250_RB" "CG17838_RB" "CG17800_RBA" "CG17800_RR" "CG8007_RA" "CG17800_RO" "CG32491_RQ"
[183] "CG17800_RY" "CG34416_RL" "CG34416_RG" "CG33555_RG" "CG16765_RG" "CG17800_RAR" "CG32498_RB"
[190] "CG10693_RJ" "CG17800_RAG" "CG17800_RAU" "CG17800_RAL" "CG8174_RC" "CG17800_RAJ" "CG17800_RI"
[197] "CG2225_RF" "CG42274_RF" "CG10693_RP" "CG32158_RC" "CG34365_RE" "CG10706_RF" "CG6703_RA"
[204] "CG32498_RL" "CG32555_RB" "CG7029_RB" "CG10377_RB" "CG17800_RF" "CG10693_RI" "CG34416_RJ"
[211] "CG17800_RW" "CG17800_RH" "CG17800_RAT" "CG17800_RAH" "CG17800_RBD" "CG17800_RBC" "CG2225_RB"
[218] "CG2225_RC" "CG32423_RD" "CG12052_RM" "CG9821_RB" "CG32498_RE" "CG42252_RB" "CG17907_RB"
[225] "CG32464_RS" "CG10693_RL" "CG10693_RD" "CG6827_RB" "CG6827_RA" "CG13521_RA" "CG9059_RD"
[232] "CG9059_RC" "CG12052_RR" "CG2225_RE" "CG33989_RD" "CG34412_RC" "CG9660_RE" "CG34412_RB"
[239] "CG32464_RQ" "CG34344_RC" "CG34341_RC" "CG4527_RD" "CG12690_RA" "CG13521_RB" "CG34412_RG"
[246] "CG5055_RA" "CG32555_RC" "CG5020_RA" "CG15427_RE" "CG5060_RA" "CG1063_RA" "CG1063_RB"
[253] "CG15427_RC" "CG17090_RB" "CG9674_RD" "CG8639_RB" "CG8639_RC" "CG42403_RC" "CG4467_RA"
[260] "CG4467_RB" "CG8566_RB" "CG34416_RI" "CG34344_RA" "CG9239_RB" "CG9674_RA" "CG4821_RA"
[267] "CG6282_RB" "CG8566_RE" "CG8566_RD" "CG34365_RD" "CG33957_RB" "CG6671_RC" "CG5685_RA"
[274] "CG1070_RA" "CG5627_RB" "CG5627_RA" "CG33183_RA" "CG32464_RK" "CG42403_RG" "CG33143_RC"
[281] "CG6703_RB" "CG3954_RA" "CG32158_RG" "CG15427_RA" "CG15028_RB" "CG33183_RC" "CG32529_RA"
[288] "CG42260_RA" "CG18250_RC" "CG9660_RA" "CG5060_RB" "CG42281_RF" "CG17838_RG" "CG33989_RC"
[295] "CG10693_RF" "CG32464_RD" "CG42260_RC" "CG32529_RD" "CG32809_RB" "CG17912_RB" "CG10693_RM"
[302] "CG18250_RA" "CG4894_RB" "CG4059_RA" "CG42403_RF" "CG33989_RF" "CG1725_RL" "CG11206_RD"
[309] "CG5020_RB" "CG4527_RA" "CG4894_RC" "CG10706_RH" "CG6671_RB" "CG10706_RE" "CG17838_RC"

[316] "CG4059_RB" "CG17838_RF" "CG12052_RF" "CG9239_RA" "CG16765_RH" "CG16765_RF" "CG34362_RA"
[323] "CG34362_RB" "CG34341_RB" "CG32632_RC" "CG32538_RB" "CG32490_RR" "CG32464_RT" "CG32464_RP"
[330] "CG32498_RA" "CG3136_RA" "CG6998_RB" "CG7125_RD" "CG7125_RC" "CG15009_RC" "CG15009_RB"
[337] "CG32464_RM" "CG33957_RC" "CG9660_RF" "CG8385_RH" "CG10618_RD" "CG10618_RE" "CG10618_RB"
[344] "CG33555_RE" "CG5020_RL" "CG31689_RC" "CG32464_RC" "CG7145_RB" "CG8385_RI" "CG8385_RF"
[351] "CG8983_RB" "CG7145_RA" "CG32158_RD" "CG10618_RC" "CG8669_RD" "CG33232_RB" "CG33080_RB"
[358] "CG33183_RD" "CG32491_RAA" "CG32491_RM" "CG34416_RE" "CG5020_RK" "CG7971_RA" "CG7971_RD"
[365] "CG34416_RN" "CG17912_RA" "CG12052_RL" "CG12052_RA" "CG32491_RT" "CG12052_RQ" "CG12052_RO"
[372] "CG12052_RX" "CG12052_RW" "CG32491_RC" "CG7893_RB" "CG1228_RB" "CG5020_RI" "CG34373_RF"
[379] "CG42274_RC" "CG32491_RH" "CG10077_RA" "CG7971_RC" "CG32491_RN" "CG32491_RP" "CG32491_RV"
[386] "CG32491_RAB" "CG32491_RE" "CG32491_RR" "CG32491_RD" "CG32491_RAC" "CG32491_RF" "CG32491_RK"
[393] "CG32491_RX" "CG32491_RY" "CG32491_RW" "CG32491_RZ" "CG32491_RO" "CG32491_RS" "CG32491_RJ"
[400] "CG32491_RI" "CG32491_RB" "CG12052_RD" "CG31716_RD" "CG33275_RC" "CG32490_RA" "CG34373_RD"
[407] "CG32498_RC" "CG4821_RC" "CG42274_RD" "CG17090_RA" "CG17800_RBF" "CG17800_RBG" "CG31716_RG"
[414] "CG17800_RJ" "CG31716_RB" "CG34416_RM" "CG4357_RB" "CG34416_RH" "CG10706_RG" "CG8547_RB"
[421] "CG31689_RB" "CG42275_RB" "CG32491_RU" "CG32491_RA" "CG5685_RC" "CG6282_RC" "CG34392_RD"
[428] "CG33275_RB" "CG32688_RA" "CG42492_RC" "CG42275_RC" "CG34344_RB" "CG11711_RB" "CG11711_RA"
[435] "CG7125_RB" "CG1725_RG" "CG42281_RG" "CG42281_RH" "CG6998_RA" "CG12052_RH" "CG6998_RC"
[442] "CG14619_RE" "CG5659_RA" "CG7125_RA" "CG3954_RC" "CG4070_RB" "CG5659_RB" "CG32632_RB"
[449] "CG6703_RE" "CG34412_RE" "CG1725_RK" "CG12690_RB" "CG42492_RB" "CG42238_RB" "CG1725_RA"
[456] "CG32158_RA" "CG15072_RA" "CG9674_RC" "CG9674_RB" "CG1070_RD" "CG32555_RA" "CG32498_RN"
[463] "CG33275_RA" "CG3954_RB" "CG34412_RH" "CG7029_RA" "CG33555_RC" "CG13316_RA" "CG14619_RA"
[470] "CG12052_RV" "CG12052_RK" "CG3136_RB" "CG2225_RA" "CG17907_RA" "CG10077_RB" "CG17342_RB"

[477] "CG17342_RA" "CG10706_RA" "CG33097_RB" "CG11206_RB" "CG12052_RJ" "CG13316_RB" "CG11727_RB"
[484] "CG3920_RB" "CG7971_RB" "CG12054_RA" "CG14619_RC" "CG34373_RG" "CG12052_RS" "CG17077_RD"
[491] "CG3920_RA" "CG42281_RI" "CG31716_RA" "CG6703_RC" "CG2822_RA" "CG14619_RD" "CG7029_RD"
[498] "CG6919_RA" "CG42275_RH" "CG15028_RC" "CG32498_RF" "CG10377_RC" "CG11172_RB" "CG8174_RA"
[505] "CG11172_RA" "CG32423_RA" "CG32423_RC" "CG32423_RB" "CG9821_RA" "CG13316_RC" "CG10377_RA"
[512] "CG4070_RA" "CG34365_RC" "CG42281_RD" "CG42274_RB" "CG31349_RF" "CG31349_RG" "CG31349_RH"
[519] "CG31349_RE" "CG5640_RB" "CG8260_RA" "CG15072_RB" "CG8260_RB" "CG31284_RB" "CG4357_RA"
[526] "CG32464_RO" "CG32490_RG" "CG32490_RP" "CG32490_RM" "CG6923_RB" "CG5055_RB" "CG31729_RA"
[533] "CG31729_RB" "CG11680_RC" "CG11680_RB" "CG16971_RB" "CG16971_RD" "CG16971_RC" "CG33080_RA"
[540] "CG11711_RD" "CG1725_RF" "CG6919_RB" "CG8007_RB" "CG34392_RC" "CG7893_RC" "CG11206_RC"
[547] "CG33143_RB" "CG32688_RB" "CG32464_RF" "CG8385_RB" "CG31120_RA" "CG31120_RB" "CG31536_RE"
[554] "CG6016_RB" "CG8983_RA" "CG15009_RA" "CG16747_RC" "CG16747_RA" "CG32464_RA" "CG16747_RB"
[561] "CG8385_RE" "CG8385_RC" "CG8669_RA" "CG6923_RA" "CG1228_RC" "CG31689_RD" "CG32464_RI"
[568] "CG4821_RD" "CG13784_RB" "CG13784_RC" "CG1228_RA" "CG16765_RI" "CG8547_RA" "CG32245_RC"
[575] "CG6016_RA" "CG32490_RQ" "CG11711_RC" "CG32245_RB" "CG32490_RE" "CG32245_RA" "CG5640_RA"
[582] "CG42238_RA" "CG31284_RC" "CG32490_RK" "CG17077_RB" "CG15427_RD" "CG1070_RF" "CG17689_RA"
[589] "CG9660_RB" "CG11727_RA" "CG42492_RA" "CG11206_RA" "CG42252_RC" "CG12054_RB" "CG1693_RA"
[596] "CG32809_RD" "CG1070_RC" "CG12052_RI" "CG12052_RP" "CG1070_RE" "CG7971_RE" "CG42252_RD"
[603] "CG31349_RA" "CG42281_RJ" "CG17077_RC" "CG4527_RF" "CG33097_RA" "CG34373_RE"

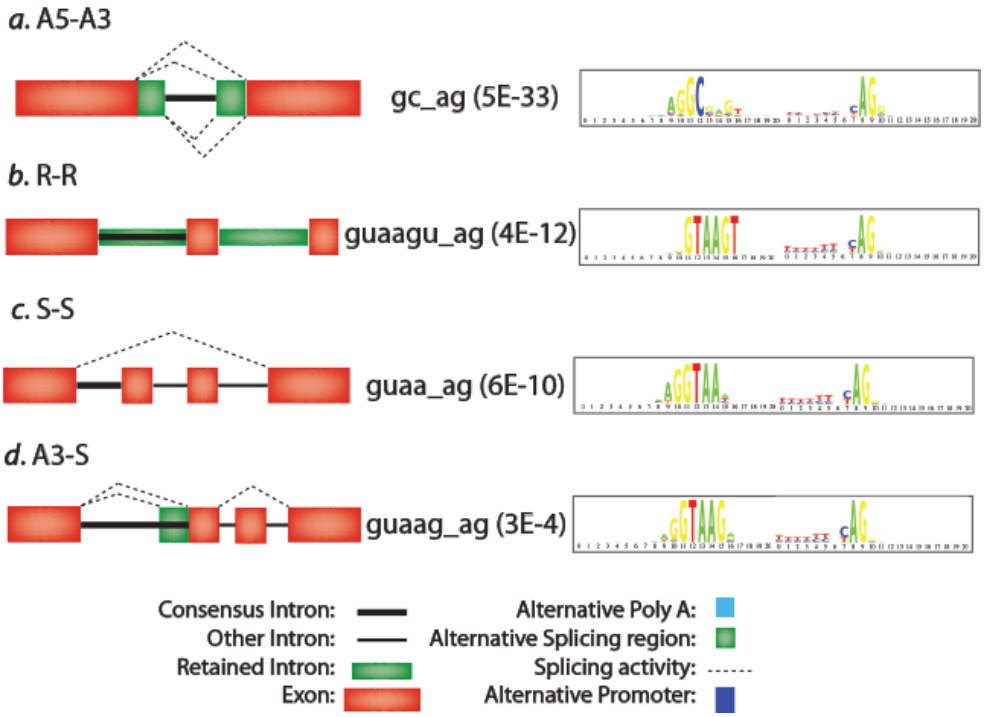
CHAPTER 3. CONSENSUS SEQUENCE IN ALTERNATIVE SPLICING

The almost invariant consensus sequence for mRNA splicing in animals and plants is gu_ag, where gu is the splice donor sequence and ag is the splice acceptor sequence. A longer splice donor consensus sequence in most mammals is guragu, where r is either g or a (Mount, 1982; Black, 2003). The splice acceptor consensus sequence is preceded by a branch point sequence, which contains an adenine, which is ligated to the 5' splice site ribonucleotide to form the intron lariat, and a polypyrimidine tract (c or u), which is between the branch point and the splice acceptor sequence. While the short gu_ag consensus sequence of introns is clearly not sufficient to differentiate amongst the multitude of alternative splicing events, surprisingly little is known about what other sequence information is required to regulate alternative RNA splicing (Ladd and Cooper, 2002; Barash et al., 2010; Witten and Ule, 2011).

Alternative RNA splicing occurs in almost all human genes and vastly increases the number of proteins and transcripts that an organism can produce (Pan et al., 2008; Wang et al., 2008). Exons that are involved in all types of RNA splicing can be classified into five major categories: 1) exons containing alternative 5' splice sites (A5), 2) exons containing alternative 3' splice sites (A3), 3) retained or invariant exons (R), 4) skipped exons (S), 5) mutually exclusive exons (ME) (Ast, 2004; Sugnet et al., 2004). In addition to these five types of exons, there are also exons that contain an alternative promoter (APr) and exons that contain an alternative poly A (APA) site. Since an intron can be flanked by two exons, APr can only be at the 5' end, and APA can only be at the 3' end, there are in total 36 possible pair-wise categories that are distinguished by the combinations of the above 7 AS types. Here, we present all AS types in the form Xa-Xb, where X is one of the seven types of exon, and the Xa exon precedes the Xb exon in the same gene. For example, the class A5-A3 is an intron that is flanked by an upstream exon with an alternative 5' splice site and a downstream exon with an alternative 3' splice site.

In Fig.22a, all 4 types of splicing, indicated with dashed lines, would generate introns in the A5-A3 class. Fig.22b-d show R-R, S-S, and A3-S classes of introns and the consensus sequences that are most significantly enriched for these classes of introns. Notice that there are only 36 possibly combinations for the 7 types of exons rather than 49 (i.e., $7^2 = 49$) because alternative poly A (APA) is never first (Xa) and alternative promoter (APr) is never second (Xb) in the Xa-Xb nomenclature system.

Fig.22. Intron Motifs that Are Over-represented in Four Representative Alternative Splicing Classes. Left, diagram of alternative splicing classes A5-A3, R-R, S-S, and A3-S. Middle, consensus sequence with most significant p-values for enrichment in this intron class (in parenthesis). Right, "logo plot" of consensus sequences. The larger the letter, the more frequent the nucleotide. a, A3-A5 (Alternative 5' splice site followed by an alternative 3' splice site). b, R-R (Retained exon followed by another retained exon). c, S-S (Skipped exon followed by another skipped exon). d, A3-S (Alternative 3' splice site followed by a skipped exon).



Here we present bioinformatics evidence to support our hypothesis of the 36 different AS types with unique consensus sequences. To test this hypothesis, we determined whether the 36 AS types are enriched for a particular paired consensus sequence(s) that is derived from both ends of the intron and flanking exon regions. We first generated a table of paired splice donor and acceptor consensus sequences, from the most common to the least common. For statistical reasons, we selected an arbitrary cutoff of each paired consensus sequence being represented by at least 100 introns. Using a modification of our program SnpEff (Cingolani et al., 2012), which classifies sequences in any sequenced genome, we analyzed the genomes of 50 different plant and animal species.

The total number of different types of paired consensus sequences ranged from one in baker's yeast, Saccharomyces cerevisiae, guaugu_ag, which has only 282 introns, all of which are always flanked by invariant exons (R-R), to 95 different consensus sequences in the marmoset, which has 184,882 introns. The average number of introns in the 50 species that we analyzed was 116,288 with a standard deviation of 45,266. Almost half of the animals' genomes we analyzed have between 40 and 50 different types of paired consensus sequences with at least 100 introns in each type. The 42 different paired consensus sequences in humans, which are in at least 100 introns, in rank order from most common to least common (Table 6) were analyzed individually to determine whether they are enriched or depleted for any of 36 AS types.

Table 6. The Top 42 Ranked Intron Consensus Sequences in Humans. Rank, the most to the least common consensus sequence. Donor-Acceptor (5S-3S), the intron sequences of the donor and acceptor sequences. Count, the count number of introns that have the indicated consensus sequence (N>100). 1-42, the total number of introns in rank 1-42 is 213,949, which represents 99.44% of the total number of introns in humans.

Rank	5S-3S	Count	Rank	5S-3S	Count	Rank	5S-3S	Count
0	ALL	215155	15	gucag_ag	1650	30	gc_ag	230
1	gugagu_ag	30585	16	gugugu_ag	1421	31	gucgg_cag	202
2	guaag_ag	29538	17	guuagu_ag	1415	32	guca_ag	198
3	guaagu_ag	28972	18	guuugu_ag	1113	33	guccg_ag	162
4	gugag_ag	26627	19	gcaagu_ag	967	34	gcagg_ag	153
5	guaa_ag	22188	20	guuggu_ag	929	35	guucgu_ag	147
6	gua_ag	20040	21	gugggu_ag	918	36	guauccuu_ag	144
7	guagg_ag	12474	22	gugug_ag	912	37	auauccuu_ac	124
8	guaugu_ag	6312	23	guggg_ag	719	38	gua_ugguuucag	118
9	guaug_ag	5552	24	gucugu_ag	450	39	guaag_uguucag	117
10	guggg_cag	5168	25	gucug_ag	371	40	gu_ugguuuuag	113
11	gu_ag	4901	26	gugcgu_ag	311	41	gcaug_ag	112
12	guga_ag	3332	27	gcaag_ag	301	42	gu_uuugagacag	109
13	gucagu_ag	2439	28	gugcg_ag	255	Total (1-42)		213,943 (99.44%)
14	gugcg_cag	1904	29	guauccuuu_ag	250			

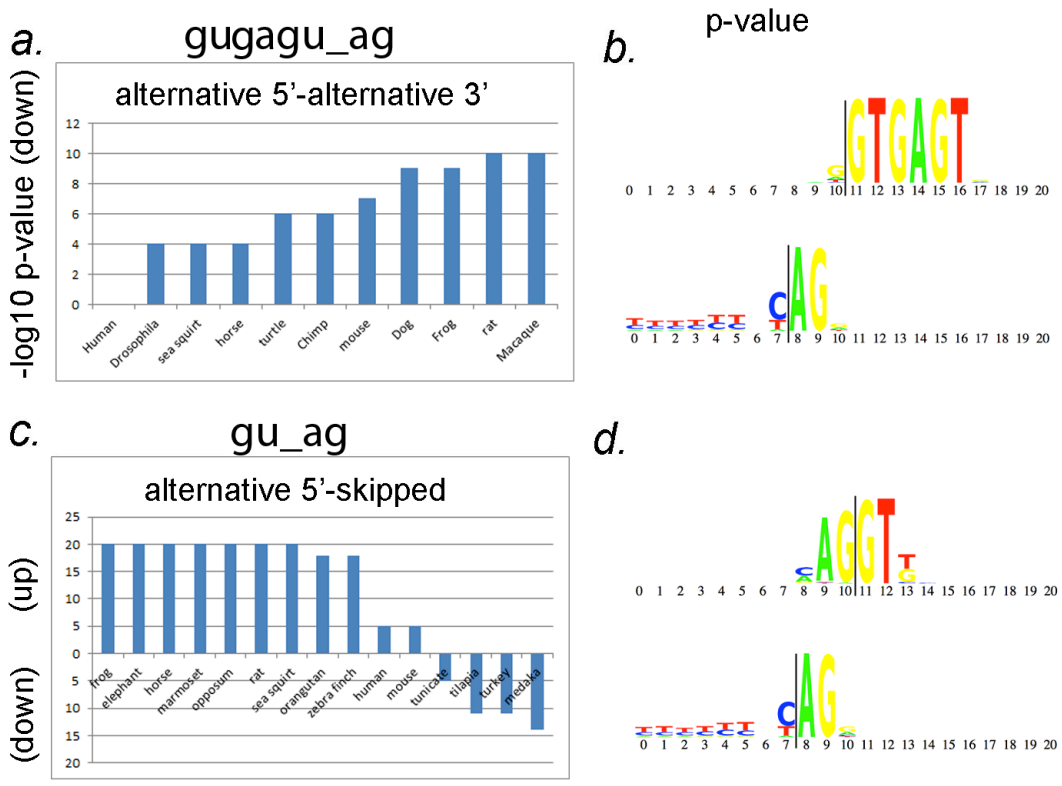
In our analyses of the 5 most studied species, the most frequent intron class in most of the species is R-R, (e.g., 79% for H.sapiens, 29% for M. musculis, 62% for D. melanogaster, 80% for C. elegans and 91% for A. thaliana) which means a invariant exon is followed by another invariant exon (Fig.22b). The second most common intron class, in most of the 50 species analyzed, is S-S (e.g., 5% for H. sapiens, 22% for M. musculus, 4% for D. melanogaster, and 3% for C. elegans), which means that two consecutive 7 exons are skipped, either together or individually, in mature RNA (Fig.22c).

Other studies have also suggested that exon skipping is the most frequently occurring alternative splicing event. For example, it was found that over one third of exons can be skipped (~38%) (Ast, 2004; Sugnet et al., 2004) and “pathological” exon skipping is commonly seen in diseases with multiple disrupted alternative splicing events, especially in cancer (Watson and Watson, 2010). We then compared alternative splicing consensus sequences among the 50 species to determine whether they are evolutionarily conserved and whether they are enriched in the same classes of AS types. The top two consensus sequences that are shared by the greatest number of species are gugagu_ag and gu_ag (Fig.23). The motif gugagu_ag is enriched for the intron class A5-A3 in 10 of the 50 species (Fig.23a), and the motif gu_ag is enriched in the intron class A5-S in 11 of the 50 species and is depleted in the intron class A5-S in 4 of the 50 species (Fig.23b).

Humans and mouse share 80% of all alternative RNA splicing motifs. When we looked for a possible reason why D. melanogaster, C. elegans and A. thaliana share only small portion of significant motifs with human (14%, 26% and 29%), we found that, although the canonical sequence gu_ag is the most highly conserved (98%), the third base after the splicing donor “gu” varies. The base adenine was hardly ever observed in the third position of the intron donor sequence in D. melanogaster or C. elegans (less than 1%), while adenine is the most common

nucleotide in the third position in the splice donor, i.e., gua_ag, for both human (58.4%) and mouse (58.3%).

Fig.23. Conserved Splicing Motifs in 50 Species. a, The alternative 5' splice site – alternative 3' splice site (A5-A3) class is under represented for the consensus sequence gugagu_ag in 10 of the 50 species analyzed. b, Donor and acceptor motif structure for gugagu_ag class. The splice donor (gu) starts at 11 and the splice acceptor (ag) ends at position 9 (vertical lines). c, The alternative 5' splice site – skipped exon (A5-S) class is enriched for the consensus sequence gu_ag in 11 of the 50 species and depleted in 4 of the 50 species analyzed. d, Donor and acceptor motif structure for gu_ag class.



Moreover, there are many practical uses for understanding the differential dosage of the AS types. For example, many diseases, including cancer, have mutations that cause changes in alternative RNA splicing that contribute to pathogenesis (Watson and Watson, 2010). It was estimated that at least 15% to 50% of mutations that cause human diseases affect splice-site selection (Wang and Cooper, 2007; Singh and Cooper, 2012). Here, we show how differential dosage of the AS types helps to interpret human genetic diseases that are caused by mutations near splice donor and acceptor sites (Singh and Cooper, 2012). Using the databases of disease-causing mutations at spliced 3' and 5' splice sites, dbass5 and dbass (http://www.dbass.org.uk/dbass5/viewlist.aspx) (Singh and Cooper, 2012), we analyzed all intron mutations at intron positions +3, +4, +5 and +6 (the first intron nucleotide at the splice donor is +1) and successfully correlated the alternative RNA splicing code to 96 different mutations in 56 genes.

One of the examples showed that Menkes disease (MD), which has several alleles in the ATP7A gene that are associated with alternative splicing defects, is a lethal disorder of copper metabolism that lead to severe neurological degeneration (Møller et al., 2000). Occipital horn syndrome (OHS) is a milder allelic form that is caused by partial loss of function of the ATP7A gene (Møller et al., 2000). Both MD and OHS are caused by mutations in the intronic sequences of the ATP7A gene, which encodes an ATPase that is responsible for copper efflux from cells (Fig.24b) (Nissim-Rafinia and Kerem, 2002). In the ATP7A gene, two splice-site mutations (IVS6+1G>A, IVS6+5G>A) for MD and one (IVS6+6T>A) for OHS were identified in a previous study (Fig.24) (Møller et al., 2000).

The main biological effect of the mutation in the first position of the splice donor site of intron 6 (gu to au) is cryptic downstream splice donor usage followed by exon 7 skipping (Fig.24c) (Møller et al., 2000). Exon skipping and cryptic splice site activation are typical results

of mutations in any of the four core consensus bases, gu_ag, and can be explained without our hypothesis. However, why the ATP7A mutation in position 5 of intron 6 (IVS6+5G>A) has such a severe effect on alternative splicing was previously not understood since this is outside of the canonical gu_ag consensus sequence (Fig.24d) (Møller et al., 2000). Now, we can better explain the alternative splicing phenotypes caused by the mutations the 5th position of the 5' splice site of intron 6 of ATP7A.

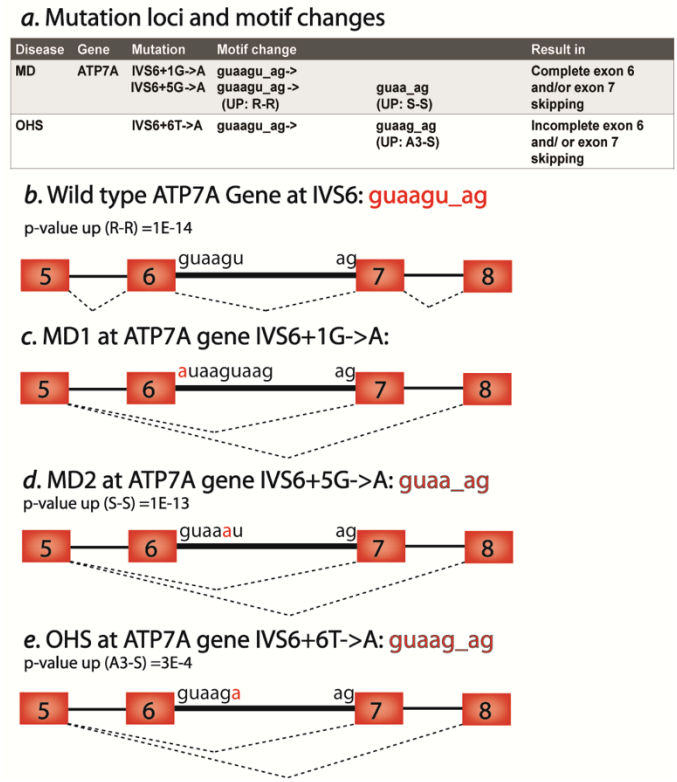
The wild-type sequence guaagu_ag corresponds to a paired consensus sequence that is overrepresented for R-R, which means that there is little or no alternative splicing in the wild-type ATP7A gene for this intron (Fig.22a). However, the 5th position mutation (Fig.24d) corresponds to the guaa_ag paired consensus sequence that is over-represented for the intron class S-S (Fig.22c). Therefore, the alternative RNA splicing code helps explain why two adjacent exons, exons 6 and 7, are skipped as the result of the mutation in the 5th position (Fig.24d). A similar argument can be made for the milder ATP7A mutation in OHS, (IVS6+6T>A), which leads to a motif change to guaag_ag, which is an overrepresented motif for the intron class A3-S, and leads to incomplete exon 6 and/or exon 7 skipping and cryptic splice site usage 50 nucleotides downstream of the normal 5' splice site in intron 6, at a second guaag_ag sequence (Fig.24e).

In the OHS allele, exon 6 becomes an A3 exon because the 5' splice site of exon 5 can join with the normal 3' splice site or exon 6 or the alternative 3' splice sites of exon 7 or exon 8 (Fig.24e). We note that the above analysis for ATP7A intron 6 is an over simplification of what is required to predict the effect of an intron mutation because multiple consensus sequences are often enriched or depleted in several of the 36 types on introns. For example, the wild type ATP7A intron 6 consensus sequence, guaagu_ag, corresponds to a consensus sequence that is enriched for R-R, R-S, and S-APA (Fig.24a). Therefore, in order to predict the outcome of a

mutation in a consensus sequence, one must determine which intron classes are uniquely enriched when a mutation is present that was not enriched in the wild-type sequence. The sixth position mutation in OHS has the intron sequence guaag_ag which is enriched in A3-S and R-R. This might explain why both A3-S and R-R splicing events are induced by the OHS mutation (Fig.24e).

Similarly, the fifth position mutation in MD2 has 10 the sequence guaa_ag, which is only enriched in the intron type S-S. This might explain why S-S splicing events are induced by the MD2 mutation (Fig.24d). Similar to the MD disease, the alternative RNA splicing code might also be used to explain +3 to +6 intron mutations in neurofibromatosis type 1 (NF1), one of the most prevalent inherited disorders in human (Hastings and Krainer, 2001), beta thalassemia (HBB) (Felber et al., 1982), and many other human diseases (data not shown).

Fig.24. ATP7A Mutations in Menke's Disease. a. Summary of the mutation loci and motif changes. b. Wild type ATP7A intron 6 has the sequence guaagu_ag. The p-value (up) for this sequence in the Retained – Retained (R-R) class of introns is 1E-14. c. The MD1 mutation (IVS6+1G>A) in ATP7A causes complete exon 6 and/or exon 7 skipping and cryptic splice site usage at the 5th position in the intron at the sequence guaag_ag. d. The MD2 mutation (IVS6+5G>A) in ATP7A causes complete exon 6 and/or exon 7 skipping and has the sequence guaa_ag. The p-value (up) for this sequence in the skipped – skipped (S-S) class of introns is 1E-13. e. The OHS mutation (IVS6+6T>A) in ATP7A is a weaker allele that causes incomplete exon 6 and/or exon 7 skipping and cryptic splice site usage at a second guaag_ag motif 50 base pairs downstream of the splice donor site. The p-value (up) for guaag_ag in the Alternative 3' splice site – skipped class of introns is 3E-4.



In addition to the canonical splicing pathway, which uses the gu_ag consensus sequence, there are also non-canonical (a.k.a., minor) splicing pathways that sometimes do not use the gu_ag consensus (Padgett, 2012). The canonical splicing pathway generally uses the U1 and U2 small RNAs in their splicing mechanism, always at gu_ag introns, while the non-canonical pathway uses U11 and U12 small RNAs, at both gu_ag and au_ac introns. The U12-like introns also have several conserved nucleotides that flank the splice donor and splice acceptor sequences (Padgett, 2012). When we searched for U12-like consensus sequences in the lists of intron consensus sequences, we found that human and mouse share the top three U12-like sequence matches: (1) guauccuuu_ag (Rank 29, Table 6), (2) auauccuu_ac (Rank 37, Table 6) and (3) guauccuu_ag (Rank 36, Table 6). The U12- like motif guauccuuu_ag is also the best match with the U12-like splicing pathway in A. thaliana. Curiously, both D. melanogaster and C.elegans have the weakest matches to the U12-like splicing sequence, gugggu_cag and guucguuuuu_uuucag, respectively, even though they are presumably evolutionarily closer to humans than plants.

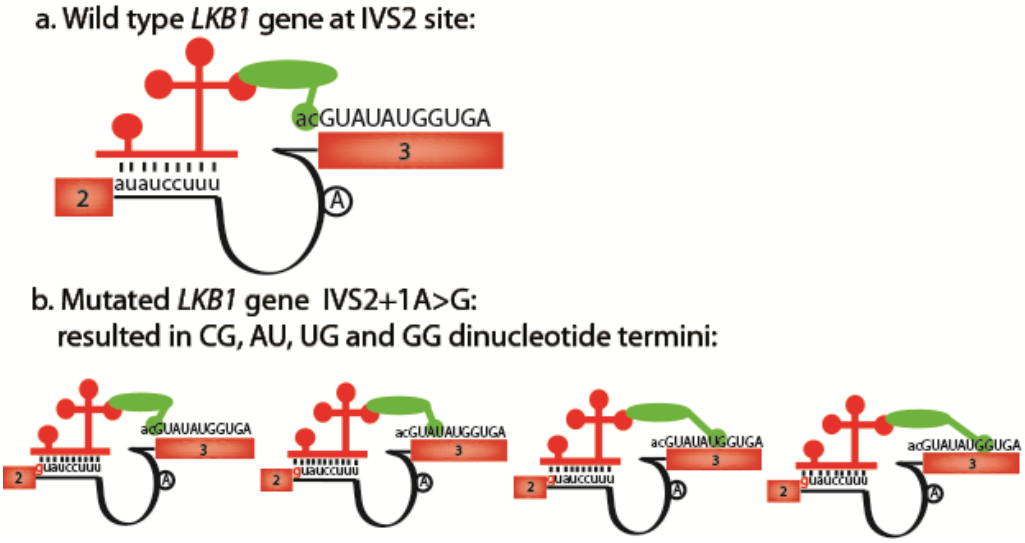
As we showed with mutations that affect the major splicing machinery, mutations that affect the minor splicing machinery can also be better interpreted with the paired consensus sequence motifs that we identified. One example involves a tumor suppressor gene, LKB1, whose splice acceptor mutation in the second intron is thought to cause Peutz-Jeghers syndrome (PJS) (Hastings et al., 2005). This mutation changes the splice junction sequence from auauccuu_ac to guauccuu_ac, and causes aberrant splicing, even though the mutation is changing a non-canonical ‘au’ splice donor to a canonical ‘gu’ splice donor (Fig.25a).

Perusing the alternative RNA splicing code, we noticed that the wild-type LKB1, auauccuu_ac, is present, but the sequence found in PJS, guauccuu_ac, is not present on the paired RNA splicing consensus sequence table in humans (Table 6). Therefore, even though

the consensus sequence table indicates that the splice donor sequence guauccuu is a good minor splice donor sequence, the paired-sequence analyses indicate that the 'gu' core splice donor sequence must be paired with another canonical splice acceptor sequence, 'ag', even in U12-type introns.

In other words, our analyses suggest that there are at least two distinct classes of U12-type introns in humans; one with the core sequence gu_ag and the other with au_ac, and the machinery that recognizes the two ends of the introns in the U12-type splicesosomes cannot be swapped. This hypothesis might also help explain the unusual splicing reactions at the 3' splice site to be multiple cryptic dinucleotide termini (such as cg, au, ug and gg) observed from different patients since no "ag" is present in vicinity of the splice acceptor site (Fig.25b) (Hastings et al., 2005).

Fig.25. The Alternative mRNA Splicing Code Predicts the Effects of a U12-type Intron Mutation, IVS2+1A>G, in the LKB1 gene. a, Schematic of human LKB1 wild-type gene sequence, has a "auauccuu_ac" U12-like intron consensus sequence. Exon numbers are shown in boxes, sequences belong to exons are uppercase; lines represent introns, sequences belonging to introns are lowercase. The 5' and 3' splice site recognition machinery of the U12 splicesosome complex are shown schematically. The lariat site is shown as an 'A' in a black circle. b, The mutation in Peutz-Jegher's syndrome (IVS2+1A>G) changes the U12-like consensus sequence from auauccuu_ac to guauccuu_ag. However, since no "ag" is detected at the 3' splice site, cg, au, ug and gg become alternative dinucleotide termini.



In summary, the alternative RNA splicing code can also be used to better understand how human germline disease mutations can affect alternative RNA splicing and lead to disease etiology. However, future biochemical experiments are needed to test the hypothesis that the many classes of paired alternative RNA splicing events in humans with paired consensus sequences have unique macromolecular complexes that regulate RNA maturation. Future bioinformatics analyses are needed to predict how a particular splice site mutation in any of the first or last few nucleotides in an intron precisely affects alternative splicing. The alternative splicing code should help inform both of these endeavors.

REFERENCES

Adonaylo V and Oteiza PI (1999) Lead intoxication: antioxidant defenses and oxidative damage in rat brain. *Toxicology* **135**:77-85.

Anders S, Pyl PT and Huber W (2014) HTSeq—a Python framework to work with high-throughput sequencing data. *Bioinformatics*:btu638.

Andrews S (2010) FastQC: A quality control tool for high throughput sequence data. *Reference Source*.

Armitage SA, Peuß R and Kurtz J (2015) Dscam and pancrustacean immune memory–a review of the evidence. *Developmental & Comparative Immunology* **48**:315-323.

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM and Sherlock G (2000) Gene Ontology: tool for the unification of biology. *Nat Genet* **25**:25-29.

Ast G (2004) How did alternative splicing evolve? *Nature Reviews Genetics* **5**:773-782.

Attrill H, Falls K, Goodman JL, Millburn GH, Antonazzo G, Rey AJ, Marygold SJ and Consortium F (2015) FlyBase: establishing a Gene Group resource for Drosophila melanogaster. *Nucleic acids research*:gkv1046.

Baranowska-Bosiacka I. ea (2012) Neurotoxicity of lead. Hypothetical molecular mechanisms of synaptic function disorders. *Neurologia i Neurochirurgia Polska* **46**:569-578.

Barash Y, Calarco JA, Gao W, Pan Q, Wang X, Shai O, Blencowe BJ and Frey BJ (2010) Deciphering the splicing code. *Nature* **465**:53-59.

Battle A, Mostafavi S, Zhu X, Potash JB, Weissman MM, McCormick C, Haudenschild CD, Beckman KB, Shi J and Mei R (2014) Characterizing the genetic basis of transcriptome diversity through RNA-sequencing of 922 individuals. *Genome research* **24**:14-24.

Bejarano F, Luque CM, Herranz H, Sorrosal G, Rafel N, Pham TT and Milán M (2008) A gain-of-function suppressor screen for genes involved in dorsal–ventral boundary formation in the Drosophila wing. *Genetics* **178**:307-323.

Bellinger DC (2013) Prenatal exposures to environmental chemicals and children's neurodevelopment: an update. *Saf Health Wrk* **4**:1-11.

Benzer S (1955) Fine structure of a genetic region in bacteriophage. *Proceedings of the National Academy of Sciences of the United States of America* **41**:344.

Bing N. ea (2005) Genetical genomics analysis of a yeast segregant population for transcription network inference. *Genet Society of American*:533-542.

Black DL (2003) Mechanisms of alternative pre-messenger RNA splicing. *Annual review of biochemistry* **72**:291-336.

Bradbury M and Deane R (1992) Permeability of the blood-brain barrier to lead. *Neurotoxicology* **14**:131-136.

Brem RB, Storey JD, Whittle J and Kruglyak L (2005) Genetic interactions between polymorphisms that affect gene expression in yeast. *Nature* **436**:701-703.

Brem RB, Yvert G, Clinton R and Kruglyak L (2002) Genetic dissection of transcriptional regulation in budding yeast. *Science* **296**:752-755.

Broman KW, Wu H, Sen Ś and Churchill GA (2003) R/qtl: QTL mapping in experimental crosses. *Bioinformatics* **19**:889-890.

Cartharius K. ea (2005) MatInspector and beyond: promoter analysis based on transcription factor binding sites. *Bioinformatics* **21**:2933-2942.

Chhangawala S, Rudy G, Mason CE and Rosenfeld JA (2015) The impact of read length on quantification of differentially expressed genes and splice junction detection. *Genome biology* **16**:1-10.

Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, Land SJ, Lu X and Ruden DM (2012) A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly* **6**:80-92.

Consortium TGO (2015) Gene Ontology Consortium: going forward. *Nucleic Acids Research* **43**:D1049-D1056.

Cook RK, Christensen SJ, Deal JA, Coburn RA, Deal ME, Gresens JM, Kaufman TC and Cook KR (2012) The generation of chromosomal deletions to provide extensive coverage and subdivision of the Drosophila melanogaster genome. *Genome Biol* **13**:R21.

Dietrich KN, et al (2001) Early exposure to lead and juvenile delinquency. *Neurotoxicol Teratol* **23**:511-518.

Eisen MB, Spellman PT, Brown PO and Botstein D (1998) Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences* **95**:14863-14868.

Fadiel A and Naftolin F (2003) Microarray applications and challenges: a vast array of possibilities. *Int Arch Biosci* **1**:111-1121.

Felber BK, Orkin SH and Hamer DH (1982) Abnormal RNA splicing causes one form of α thalassemia. *Cell* **29**:895-902.

Foissac S and Sammeth M (2007) ASTALAVISTA: dynamic and flexible analysis of alternative splicing events in custom gene datasets. *Nucleic acids research* **35**:W297-W299.

Francesconi M and Lehner B (2014) The effects of genetic variation on gene expression dynamics during development. *Nature* **505**:208-211.

Gibbs RA, Belmont JW, Hardenbol P, Willis TD, Yu F, Yang H, Ch'ang L-Y, Huang W, Liu B and Shen Y (2003) The international HapMap project. *Nature* **426**:789-796.

Gilbert W (1978) Why genes in pieces? *Nature* **271**:501.

Gupta S, Stamatoyannopoulos JA, Bailey TL and Noble WS (2007) Quantifying similarity between motifs. *Genome biology* **8**:R24.

Gymrek M (2014) The complicated world of splice QTLs. *Blog melissagymrekcom*.

Hanna-Attisha M, LaChance J, Sadler RC and Champney Schnepp A (2015) Elevated Blood Lead Levels in Children Associated With the Flint Drinking Water Crisis: A Spatial Analysis of Risk and Public Health Response. *Am J Public Health*:e1-e8.

Hasin-Brumshtein Y, Hormozdiari F, Martin L, Van Nas A, Eskin E, Lusis AJ and Drake TA (2014) Allele-specific expression and eQTL analysis in mouse adipose tissue. *BMC genomics* **15**:1.

Hastings ML and Krainer AR (2001) Pre-mRNA splicing in the new millennium. *Current opinion in cell biology* **13**:302-309.

Hastings ML, Resta N, Traum D, Stella A, Guanti G and Krainer AR (2005) An LKB1 AT-AC intron mutation causes Peutz-Jeghers syndrome via splicing at noncanonical cryptic splice sites. *Nature structural & molecular biology* **12**:54-59.

Hattori D, Demir E, Kim HW, Viragh E, Zipursky SL and Dickson BJ (2007) Dscam diversity is essential for neuronal wiring and self-recognition. *Nature* **449**:223-227.

He T, Hirsch H, Ruden D and Lnenicka G (2009) Chronic lead exposure alters presynaptic calcium regulation and synaptic facilitation in Drosophila larvae. *Neurotoxicology* **30**:777-784.

Hirsch H, Barth M, Luo S, Sambaziotis H, Huber M, Possidente D, Ghiradella H and Tompkins L (1995) Early visual experience affects mate choice of Drosophila melanogaster. *Animal behaviour* **50**:1211-1217.

Hirsch HV, Mercer J, Sambaziotis H, Huber M, Stark DT, Torno-Morley T, Hollocher K, Ghiradella H and Ruden DM (2003) Behavioral effects of chronic exposure to low levels of lead in Drosophila melanogaster. *Neurotoxicology* **24**:435-442.

Hirsch HV, Possidente D, Averill S, Despain TP, Buytkins J, Thomas V, Goebel WP, Shipp-Hilts A, Wilson D and Hollocher K (2009) Variations at a quantitative trait locus (QTL) affect development of behavior in lead-exposed Drosophila melanogaster. *Neurotoxicology* **30**:305-311.

Hirsh H.V. ea (2009) Variations at a Quantitative Trait Locus (QTL) affect development of behavior in lead-exposed Drosophila melanogaster. *NeuroToxicol* **30**:305-311.

Hoaglin DC and Welsch RE (1978) The hat matrix in regression and ANOVA. *The American Statistician* **32**:17-22.

Huang G-J, Shifman S, Valdar W, Johannesson M, Yalcin B, Taylor MS, Taylor JM, Mott R and Flint J (2009) High resolution mapping of expression QTLs in heterogeneous stock mice in multiple tissues. *Genome research* **19**:1133-1140.

Isshiki T, Pearson B, Holbrook S and Doe CQ (2001) Drosophila neuroblasts sequentially express transcription factors which specify the temporal identity of their neuronal progeny. *Cell* **106**:511-521.

Jabłońska L, Walski M and Rafałowska U (1994) Lead as an inductor of some morphological and functional changes in synaptosomes from rat brain. *Cellular and molecular neurobiology* **14**:701-709.

Jedrychowski W. ea (2011) Intrauterine exposure to lead may enhance sensitization to common inhalant allergens in early childhood: A prospective prebirth cohort study. *Environ Res* **111**:119-124.

Joo JWJ, Sul JH, Han B, Ye C and Eskin E (2014) Effectively identifying regulatory hotspots while capturing expression heterogeneity in gene expression studies. *Genome biology* **15**:r61.

Karolchik D, Hinrichs AS, Furey TS, Roskin KM, Sugnet CW, Haussler D and Kent WJ (2004) The UCSC Table Browser data retrieval tool. *Nucleic acids research* **32**:D493-D496.

Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R and Salzberg SL (2013) TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol* **14**:R36.

King E.G. ea (2012) Genetic dissection of a model complex trait using the Drosophila Synthetic Population Resource. *Genome Res* **22**:1558-1566.

King EG, Merkes CM, McNeil CL, Hoofer SR, Sen S, Broman KW, Long AD and Macdonald SJ (2012) Genetic dissection of a model complex trait using the Drosophila Synthetic Population Resource. *Genome research* **22**:1558-1566.

King EG, Sanderson BJ, McNeil CL, Long AD and Macdonald SJ (2014) Genetic dissection of the Drosophila melanogaster female head transcriptome reveals widespread allelic heterogeneity. *PLoS Genet* **10**:e1004322.

Krawczak M, Thomas NS, Hundrieser B, Mort M, Wittig M, Hampe J and Cooper DN (2007) Single base‐pair substitutions in exon–intron junctions of human genes: nature, distribution, and consequences for mRNA splicing. *Human mutation* **28**:150-158.

Kurmangaliyev YZ, Favorov AV, Osman NM, Lehmann K-V, Campo D, Salomon MP, Tower J, Gelfand MS and Nuzhdin SV (2015) Natural variation of gene models in Drosophila melanogaster. *BMC genomics* **16**:1.

Ladd AN and Cooper TA (2002) Finding signals that regulate alternative splicing in the post-genomic era. *Genome Biol* **3**:1-16.

Lafond J, Hamel A, Takser L, Vaillancourt C and Mergler D (2004) Low environmental contamination by lead in pregnant women: effect on calcium transfer in human placental syncytiotrophoblasts. *Journal of Toxicology and Environmental Health, Part A* **67**:1069-1079.

Lagarrigue S, Martin L, Hormozdiari F, Roux P-F, Pan C, Van Nas A, Demeure O, Cantor R, Ghazalpour A and Eskin E (2013) Analysis of allele-specific expression in mouse liver by RNA-Seq: a comparison with Cis-eQTL identified using genetic linkage. *Genetics* **195**:1157-1166.

Lappalainen T, Sammeth M, Friedländer MR, AC't Hoen P, Monlong J, Rivas MA, Gonzàlez-Porta M, Kurbatova N, Griebel T and Ferreira PG (2013) Transcriptome and genome sequencing uncovers functional variation in humans. *Nature* **501**:506-511.

Lawrence Zipursky S and Grueber WB (2013) The molecular basis of self-avoidance. *Annual review of neuroscience* **36**:547-568.

Leek JT and Storey JD (2007) Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS Genet* **3**:1724-1735.

Lim KH, Ferraris L, Filloux ME, Raphael BJ and Fairbrother WG (2011) Using positional distribution to identify splicing elements and predict pre-mRNA processing defects in human genes. *Proceedings of the National Academy of Sciences* **108**:11093-11098.

Mackay TF, Stone EA and Ayroles JF (2009) The genetics of quantitative traits: challenges and prospects. *Nature Reviews Genetics* **10**:565-577.

Majewski J and Pastinen T (2011) The study of eQTL variations by RNA-seq: from SNPs to phenotypes. *Trends in Genetics* **27**:72-79.

Majewski J. ea (2011) The study of eQTL variations by RNA-seq: from SNPs to phenotypes. *Cell* **27**:72-79.

Mangravite LM, Engelhardt BE, Medina MW, Smith JD, Brown CD, Chasman DI, Mecham BH, Howie B, Shim H and Naidoo D (2013) A statin-dependent QTL for GATM expression is associated with statin-induced myopathy. *Nature* **502**:377-380.

Manichaikul A, Moon JY, Sen Ś, Yandell BS and Broman KW (2009) A model selection approach for the identification of quantitative trait loci in experimental crosses, allowing epistasis. *Genetics* **181**:1077-1086.

Marchetti C and Gavazzo P (2005) NMDA receptors as targets of heavy metal interaction and toxicity. *Neurotoxicity research* **8**:245-258.

Massouras A, Waszak SM, Albarca-Aguilera M, Hens K, Holcombe W, Ayroles JF, Dermitzakis ET, Stone EA, Jensen JD and Mackay TF (2012) Genomic variation and its impact on gene expression in Drosophila melanogaster. *PLoS Genet* **8**:e1003055.

Miles C and Wayne M (2008) Quantitative trait locus (QTL) analysis. *Nature Education* **1**:208.

Modrek B and Lee C (2002) A genomic view of alternative splicing. *Nature genetics* **30**:13-19.

Møller LB, Tümer Z, Lund C, Petersen C, Cole T, Hanusch R, Seidel J, Jensen LR and Horn N (2000) Similar splice-site mutations of the ATP7A gene lead to different phenotypes: classical Menkes disease or occipital horn syndrome. *The American Journal of Human Genetics* **66**:1211-1220.

Monlong J, Calvo M, Ferreira PG and Guigó R (2014) Identification of genetic variants associated with alternative splicing using sQTLseekeR. *Nature communications* **5**.

Montgomery SB, Sammeth M, Gutierrez-Arcelus M, Lach RP, Ingle C, Nisbett J, Guigo R and Dermitzakis ET (2010) Transcriptome genetics using second generation sequencing in a Caucasian population. *Nature* **464**:773-777.

Morley E.J. ea (2003) Effects of chronic lead exposure on the neuromuscular junction in Drosophila larvae. *NeuroToxicol* **24**:35-41.

Mount SM (1982) A catalogue of splice junction sequences. *Nucleic acids research* **10**:459-472.

Nissim-Rafinia M and Kerem B (2002) Splicing regulation as a potential genetic modifier. *TRENDS in Genetics* **18**:123-127.

Oberto A, Marks N, Evans HL and Guidotti A (1996) Lead (Pb+ 2) promotes apoptosis in newborn rat cerebellar neurons: pathological implications. *Journal of Pharmacology and Experimental Therapeutics* **279**:435-442.

Ongen H and Dermitzakis ET (2015) Alternative Splicing QTLs in European and African Populations. *The American Journal of Human Genetics* **97**:567-575.

Padgett RA (2012) New connections between splicing and human disease. *Trends in Genetics* **28**:147-154.

Pan Q, Shai O, Lee LJ, Frey BJ and Blencowe BJ (2008) Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nature genetics* **40**:1413-1415.

Pennisi E (2000) And the gene number is...? *Science* **288**:1146-1147.

Pickrell J.K. ea (2010) Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature* **464**:768-772.

Pickrell JK, Marioni JC, Pai AA, Degner JF, Engelhardt BE, Nkadori E, Veyrieras J-B, Stephens M, Gilad Y and Pritchard JK (2010) Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature* **464**:768-772.

Quinlan AR and Hall IM (2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**:841-842.

Reese MG, Eeckman FH, Kulp D and Haussler D (1997) Improved splice site detection in Genie. *Journal of computational biology* **4**:311-323.

Rockman M.V. ea (2006) Genetics of global gene expression. *Nat Genet* **7**:862-872.

Roote J and Russell S (2012) Toward a complete Drosophila deficiency kit. *Genome Biol* **13**:149.

Ruden D.M. ea (2009a) Gene-Environment Interactions: Drosophila as a model for Toxicogenomics of Lead. *Encyclopedia of Env Health*.

Ruden D.M. ea (2009b) Genetical toxicogenomics in Drosophila identifies master-modulatory loci that are regulated by developmental exposure to lead. *NeuroToxicol* **30**:898-914.

Ruden DM, Chen L, Possidente D, Possidente B, Rasouli P, Wang L, Lu X, Garfinkel MD, Hirsch HV and Page GP (2009) Genetical toxicogenomics in Drosophila identifies master-modulatory loci that are regulated by developmental exposure to lead. *Neurotoxicology* **30**:898-914.

Sambrook J (1977) Adenovirus amazes at Cold Spring Harbor. *Nature* **268**:101.

Schadt EE, Monks SA, Drake TA, Lusis AJ, Che N, Colinayo V, Ruff TG, Milligan SB, Lamb JR and Cavet G (2003) Genetics of gene expression surveyed in maize, mouse and man. *Nature* **422**:297-302.

Schmucker D, Clemens JC, Shu H, Worby CA, Xiao J, Muda M, Dixon JE and Zipursky SL (2000) Drosophila Dscam is an axon guidance receptor exhibiting extraordinary molecular diversity. *Cell* **101**:671-684.

Schmucker D and Flanagan JG (2004) Generation of recognition diversity in the nervous system. *Neuron* **44**:219-222.

Sen A, Heredia N, Senut MC, Land S, Hollocher K, Lu X, Dereski MO and Ruden DM (2015) Multigenerational epigenetic inheritance in humans: DNA methylation changes associated with maternal exposure to lead can be transmitted to the grandchildren. *Sci Rep* **5**:14466.

Singh RK and Cooper TA (2012) Pre-mRNA splicing in disease and therapeutics. *Trends in molecular medicine* **18**:472-482.

Siva N (2008) 1000 Genomes project. *Nature biotechnology* **26**:256-256.

Squassina A, Manchia M, Manolopoulos VG, Artac M, Lappa-Manakou C, Karkabouna S, Mitropoulos K, Del Zompo M and Patrinos GP (2010) Realities and expectations of pharmacogenomics and personalized medicine: impact of translating genetic knowledge into clinical practice. *Pharmacogenomics* **11**:1149-1167.

Stolc V, Gauhar Z, Mason C, Halasz G, van Batenburg MF, Rifkin SA, Hua S, Herreman T, Tongprasit W and Barbano PE (2004) A gene expression map for the euchromatic genome of Drosophila melanogaster. *Science* **306**:655-660.

Sugnet CW, Kent WJ, Ares M and Haussler D (2004) Transcriptome and genome conservation of alternative splicing events in humans and mice, in *Pacific Symposium on Biocomputing* pp 66-77.

Suszkiw JB (2004) Presynaptic disruption of transmitter release by lead. *Neurotoxicology* **25**:599-604.

Tadros W, Xu S, Akin O, Caroline HY, Shin GJ-e, Millard SS and Zipursky SL (2016) Dscam Proteins Direct Dendritic Targeting through Adhesion. *Neuron* **89**:480-493.

Thorvaldsdóttir H, Robinson JT and Mesirov JP (2013) Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Briefings in bioinformatics* **14**:178-192.

Tran KD, Miller MR and Doe CQ (2010) Recombineering Hunchback identifies two conserved domains required to maintain neuroblast competence and specify early-born neuronal identity. *Development* **137**:1421-1430.

Trapnell C. ea (2012) Differential gene and transcript expression analysis of RNA-seq experiments with Tophat and Cufflinks. *Nat Protocols* **7**:562-578.

Tweedie S, Ashburner M, Falls K, Leyland P, McQuilton P, Marygold S, Millburn G, Osumi-Sutherland D, Schroeder A and Seal R (2009) FlyBase: enhancing Drosophila gene ontology annotations. *Nucleic acids research* **37**:D555-D559.

Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA and Holt RA (2001) The sequence of the human genome. *science* **291**:1304-1351.

Wang ET, Sandberg R, Luo S, Khrebtukova I, Zhang L, Mayr C, Kingsmore SF, Schroth GP and Burge CB (2008) Alternative isoform regulation in human tissue transcriptomes. *Nature* **456**:470-476.

Wang G-S and Cooper TA (2007) Splicing in disease: disruption of the splicing code and the decoding machinery. *Nature Reviews Genetics* **8**:749-761.

Wang Z, Gerstein M and Snyder M (2009) RNA-Seq: a revolutionary tool for transcriptomics. *Nature reviews genetics* **10**:57-63.

Watson FL, Püttmann-Holgado R, Thomas F, Lamar DL, Hughes M, Kondo M, Rebel VI and Schmucker D (2005) Extensive diversity of Ig-superfamily proteins in the immune system of insects. *Science* **309**:1874-1878.

Watson PM and Watson DK (2010) Alternative splicing in prostate and breast cancer. *Open cancer journal* **3**:62-76.

White L.D. ea (2007) New and evolving concepts in the neurotoxicology of lead. *Toxicol Appl Pharmacol* **225**:1-27.

WHOteam (2015) WHO | Lead poisoning and health, in, World Health Organization.

Witten JT and Ule J (2011) Understanding splicing regulation through RNA splicing maps. *Trends in genetics* **27**:89-97.

Wojtowicz WM, Flanagan JJ, Millard SS, Zipursky SL and Clemens JC (2004) Alternative splicing of Drosophila Dscam generates axon guidance receptors that exhibit isoform-specific homophilic binding. *Cell* **118**:619-633.

Wolfertstetter F, Frech K, Herrmann G and Werner T (1996) Identification of functional elements in unaligned nucleic acid sequences by a novel tuple search algorithm. *Computer applications in the biosciences: CABIOS* **12**:71-80.

Wu C. ea (2008) Gene set enrichment in eQTL data identifies novel annotations and pathway regulators. *Plos Genet* **4**.

Xiao Y, Segal MR, Rabert D, Ahn AH, Anand P, Sangameswaran L, Hu D and Hunt CA (2002) Assessment of differential gene expression in human peripheral nerve injury. *BMC genomics* **3**:28.

Yamakawa K, Huo Y-K, Haendel MA, Hubert R, Chen X-N, Lyons GE and Korenberg JR (1998) DSCAM: a novel member of the immunoglobulin superfamily maps in a Down syndrome region and is involved in the development of the nervous system. *Human Molecular Genetics* **7**:227-237.

Young MD, Wakefield MJ, Smyth GK and Oshlack A (2010) Method Gene ontology analysis for RNA-seq: accounting for selection bias. *Genome Biol* **11**:R14.

Young MD, Wakefield MJ, Smyth GK and Oshlack A (2012) goseq: Gene Ontology testing for RNA-seq datasets.

Yu Z, Ren M, Wang Z, Zhang B, Rong YS, Jiao R and Gao G (2013) Highly efficient genome modifications mediated by CRISPR/Cas9 in Drosophila. *Genetics* **195**:289-291.

Yvert G, Brem RB, Whittle J, Akey JM, Foss E, Smith EN, Mackelprang R and Kruglyak L (2003) Trans-acting regulatory variation in Saccharomyces cerevisiae and the role of transcription factors. *Nature genetics* **35**:57-64.

Zhang X, Gierman HJ, Levy D, Plump A, Dobrin R, Goring HH, Curran JE, Johnson MP, Blangero J and Kim SK (2014) Synthesis of 53 tissue and cell line expression QTL datasets reveals master eQTLs. *BMC genomics* **15**:1.

Zhao K, Lu Z-x, Park JW, Zhou Q and Xing Y (2013) GLiMMPS: obust statistical model for regulatory.

ABSTRACT

IDENTIFICATION OF LEAD-SENSITIVE EXPRESSION AND SPLICING QUANTITATIVE TRAIT LOCI IN *DROSOPHILA MELANOGASTER* BY ANALYSIS OF RNA-SEQ DATA

by

WEN QU

December 2016

Advisor: Dr. Douglas M. Ruden

Major: Pharmacology

Degree: Doctor of Philosophy

Lead exposure has long been one of the most important topics in global public health since it is a potent developmental neurotoxin. Here, we conducted an expression QTL (eQTLs) analysis, which is genome-wide association analysis of genetic variants with differential gene expression, in the male heads of 79 *Drosophila melanogaster* recombinant inbred lines originally from eight parental strains in the presence or absence of developmental exposure to 250 μM lead acetate. The aim was to study the effects of lead exposure on gene expression and identify the lead-responsive genes. After detecting 1,536 cis-eQTLs and 952 trans-eQTLs (1000 permutation threshold at 0.05), we focused our analysis on lead-sensitive "trans-eQTL hotspots," defined as genomic regions that are associated with a cluster of genes in a lead-dependent manner. We noticed that the genes associated with one of the 13 detected trans-eQTL hotspots, Chr2L: 6,250,000 could be roughly divided into two groups based on their differential expression profile patterns and different categories of function. We visualized the expression of all the associated genes in the trans-eQTL hotspot with hierarchical clustering. Besides the overall expression profile patterns, the heat maps displayed the segregation of

differential parental genetic contributions. This suggested that trans-regulatory regions with different genetic contributions from the parental lines have significantly different expression changes after lead exposure. We believe that the lead-responsive trans-eQTL hotspots generated in this study could improve our understanding of genetic dissection of transcript abundance and provide insights into the mechanisms of how environmental toxins affect transcriptional pathways.

In a follow-up study, we also found lead-responsive sQTLs. The identification of lead-responsive sQTLs provides further evidence that different parental genomic contribution can cause significantly differential isoform usage after developmental lead exposure. Great achievements have been made in understanding how trans-sQTL hotspots alter the susceptibility to lead exposure, opening up a gate towards the mechanisms of trans-sQTL hotspots, as well as the neurotoxicity of lead.

Chapter 1 is currently under minor revision in Neurotoxicology. Chapter 2 will be submitted to Neurotoxicology after the first submission is accepted. Chapter 3 will probably be submitted to Frontiers in Genetics.

AUTOBIOGRAPHICAL STATEMENT

Education

2011/08- 2016/09..... Ph.D. Wayne State University, U.S.A

Majoring in Pharmacology (GPA: 3.88)

2007/09- 2011/06..... B. Sc. China Pharmaceutical University, P.R. China

Majoring in Pharmacy (GPA: 3.58)

Publications

Qu W, Pique-Regi R, Ruden D. Lead Modulates trans- and cis-eQTLs in Drosophila melanogaster Heads. Minor revision in Neurotoxicology.

Qu W, Cingolani P, Zeeberg B, Ruden D. A Novel Alternative RNA Splicing Code that is Conserved in Animals and Plants. Submitted to Frontiers in Genetics.

Qu W, Pique-Regi R, Ruden D. Identification of Pb-responsive splicing QTLs in Drosophila Heads. Ready for submission.

Ruden D, Cingolani P, Sen A, Qu W, Wang L, Senut M, Garfinkel M, Sollars V, Lu X. Epigenetics as an answer to Darwin's "special difficulty," Part 2: Natural selection of metastable epialleles in honeybee castes. Front Genet. 2015 Feb 24;6:60. doi: 10.3389/fgene.2015.00060. eCollection 2015.

Sen A, Heredia N, Senut M, Hess M, Land S, Qu W, Hollacher K, Dereski M, Ruden D. Early life lead exposure causes gender specific changes in the DNA methylation profile of DNA extracted from dried blood spots. Epigenomics. 2015;7(3):379-93. doi: 10.2217/epi.15.2.

Wu M, Shen Q, Yang Y, Zhang S, Qu W, Chen J, Sun H, Chen S. Disruption of YPS1 and PEP4 Genes Reduces Proteolytic Degradation of Secreted HAS/PTH in Pichia pastoris GS115. J Ind Microbiol Biotechnol. 2013; 40(6) 589-599.