


1-1-2016

# Neuronal Insult Either By Exposure To Lead Or By Direct Neuronal Damage Cause Genome-Wide Changes In Dna Methylation And Histone 3 Lysine 36 Trimethylation

Arko Sen  
*Wayne State University,*

Follow this and additional works at: [http://digitalcommons.wayne.edu/oa\\_dissertations](http://digitalcommons.wayne.edu/oa_dissertations)

 Part of the [Bioinformatics Commons](#), [Genetics Commons](#), and the [Pharmacology Commons](#)

---

## Recommended Citation

Sen, Arko, "Neuronal Insult Either By Exposure To Lead Or By Direct Neuronal Damage Cause Genome-Wide Changes In Dna Methylation And Histone 3 Lysine 36 Trimethylation" (2016). *Wayne State University Dissertations*. 1663.  
[http://digitalcommons.wayne.edu/oa\\_dissertations/1663](http://digitalcommons.wayne.edu/oa_dissertations/1663)

This Open Access Dissertation is brought to you for free and open access by DigitalCommons@WayneState. It has been accepted for inclusion in Wayne State University Dissertations by an authorized administrator of DigitalCommons@WayneState.

**NEURONAL INSULT EITHER BY EXPOSURE TO LEAD OR BY DIRECT  
NEURONAL DAMAGE CAUSES GENOME-WIDE CHANGES IN DNA  
METHYLATION AND HISTONE 3 LYSINE 36 TRI-METHYLATION**

by

**ARKO SEN**

**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

\_\_\_\_\_  
\_\_\_\_\_  
\_\_\_\_\_  
\_\_\_\_\_

**© COPYRIGHT BY**

**ARKO SEN**

**2016**

**All Rights Reserved**

## DEDICATION

To my mother Urmila Sen, for her guidance, insight and encouragement.

To my father Anil Sen, for his support

To my friend Dr. S.C. Banerjee, for being the inspiration during difficult times.

To my grandmother, Gitanjali Dutta

To my cousin brother, Sayak Raza

To my aunt, Papri Raza

And the memory of grandfather, Usha Ranjan Dutta

## ACKNOWLEDGEMENTS

I would like to express my gratitude to my supervisor, Dr. Douglas Ruden. His expertise, understanding, and patience, enabled me to complete my graduate studies. I appreciate his repertoire of knowledge in classical *Drosophila* genetics, human cohort based studies and next generation sequencing, which have benefitted me greatly and allowed me to gain considerable experience in coping with the challenges of the Genome Era.

I would also like to thank my dissertation committee members Dr. P. Stemmer, Dr. T. Kocarek, Dr. H. Wu and Dr. Roger-Pique Regi, who were instrumental in the development of my projects and helped me through various obstacles I faced during my studies.

I must also acknowledge my department, Department of Pharmacology for provided me with provisions and logistical support when needed, which made my tenure as a graduate student quite smooth and hassle free.

I would also like to thank my friends in the Krawetz Lab, particularly Molly Estill, for our philosophical debates, exchanges of knowledge, skills, and venting of frustration during my graduate program, which helped enrich the experience.

I would also like to thank my family for the support they provided me through my entire life and in particular, my mother without whose love, encouragement and editing assistance, I would not have finished this thesis. I recognize that this research would not have been possible without the financial assistance of Department of Pharmacology graduate research assistantship and NIH grants awarded to the Ruden Lab and express my gratitude to those agencies.

## TABLE OF CONTENTS

Dedication	ii
Acknowledgements	iii
List of tables	vii
List of figures	ix
CHAPTER 1 DNA METHYLATION AS A PUTATIVE BIOMARKER FOR ENVIRONMENTAL EXPOSURE	1
1.1. Background	1
1.1.1.DNA methylation is critical for genomic stability and spatial temporal regulation of gene expression	1
1.1.2.DNA hydroxymethylation is stable epigenetic modification	3
1.1.3.DNA methylation and hydroxymethylation as potential early biomarkers for environmental exposures	4
1.1.4.Lead (Pb) is potent neurotoxin and mediates its harmful effect via generation of reactive oxygen species (ROS)	7
CHAPTER 2 THE HUMAN DNA METHYLOME IS RESPONSIVE TO ENVIRONMENTAL LEAD (PB) EXPOSURE	10
2.1. Early life Lead exposure causes distinct gender specific changes in the DNA methylation profile of DNA extracted from dried blood spots	10
2.1.1.Background	10

2.1.2.Methods	11
2.1.3.Results	19
2.1.4.Discussion	24
2.2. Multigenerational epigenetic inheritance in humans: DNA methylation changes associated with maternal exposure to lead can be transmitted to the grandchildren	33
2.2.1.Background	33
2.2.2.Methods	34
2.2.3.Results	40
2.2.4.Discussion	44
CHAPTER 3 LEAD (PB) EXPOSURE INDUCES CHANGES IN 5-HYDROXYMETHYLCYTOSINE CLUSTERS IN CPG ISLANDS IN HUMAN EMBRYONIC STEM CELLS AND UMBILICAL CORD BLOOD; HMEDIP-450K ARRAY	51
3.1.Background	51
3.2.Methods	52
3.3.Results	59
3.4.Discussion	68

CHAPTER 4 TRAUMATIC BRAIN INJURY CAUSES RETENTION OF LONG INTRONS  
VIA REGULATION OF LOCAL HISTONE 3 LYSINE 36 METHYLATION PROFILE IN  
THE SUB-ACUTE PHASE OF INJURY \_\_\_\_\_ 75

4.1. Background \_\_\_\_\_ 75

4.1.1. Drosophila model of traumatic brain injury \_\_\_\_\_ 75

4.1.2. Regulation of alternative splicing by epigenetic modifications \_\_\_\_\_ 78

4.2. Methods \_\_\_\_\_ 81

4.3. Results \_\_\_\_\_ 97

4.4. Discussion \_\_\_\_\_ 119

References \_\_\_\_\_ 130

Abstract \_\_\_\_\_ 140

Autobiographical Statement \_\_\_\_\_ 143



## LIST OF TABLES

Table 2.1.1: Sample BLL ( $\mu\text{g}/\text{dl}$ ) and covariates for current blood spots used for the analysis. \_28

Table 2.1.2: Single nucleotide differences in DNA methylation between females and males as estimated by fixed effect model. \_\_\_\_\_30

Table 2.1.3: Representative gene mapping to clusters which show a change in methylation of  $\geq |2\%|$  or  $|0.02|$ , at a FDR corrected p-value cutoff of 0.05. \_\_\_\_\_32

Table 2.2.1: Covariate and Blood Lead Level (BLL) information used for DNA methylation analysis for 35 samples. \_\_\_\_\_47

Table 2.2.2: Table showing a list of 6 candidate genes/CpG clusters which show Pb dependent change in DNA methylation status in CNBS exposed in-utero to high BLL (high BLL MNBS).50

Table 3.1: Differentially hydroxymethylated (5hmC) regions (DhMRs) for representative genes in hESCs exposed to either  $0.8\mu\text{M}$  Pb or  $1.5\mu\text{M}$  Pb. These clusters can be used as potential early 5hmC biomarkers of Pb exposure. \_\_\_\_\_72

Table 3.2: Differentially hydroxymethylated (5hmC) regions (DhMRs) for representative genes for male-specific, female-specific and conserved regions in umbilical cord blood DNA with blood lead levels (BLL)  $\geq 5\mu\text{g}/\text{dl}$ . These clusters can be used as potential early 5hmC biomarkers of Pb exposure. \_\_\_\_\_72

Table 3.3: Differentially methylated (5mC) regions (DMRs) for representative genes for male-specific, female-specific and conserved regions umbilical cord blood DNA with blood lead levels (BLL)  $\geq 5\mu\text{g}/\text{dl}$ . These clusters can be used as potential early 5mC biomarkers of Pb exposure. \_\_\_\_\_73

Table 4.1: ANOVA followed by TukeyHSD for difference in tag counts (read counts) of H3K36me3 ChIP-Seq peaks between 1= 1st exon, 2= 2nd exon and 3= 3rd exons of constitutive or alternative exon trios for 3rd instar larvae(GSE47248) and adult heads(GSE47280). \_\_\_\_\_125

Table 4.2: Motif enrichment analysis for 5'SS. 299 bps sequences were collected from intronic side of 5'SS for retained intron of length  $\geq 600$  (N= 421/458). \_\_\_\_\_125

Table 4.3: Motif enrichment analysis for 3'SS. 299 bps sequences were collected from intronic side of 5'SS for retained intron of length  $\geq 600$  (N= 421/458). \_\_\_\_\_126

Table 4.4: Motif enrichment analysis for 5'SS. 299 bps sequences were collected from intronic side of 5'SS for retained intron of length  $\geq 600$  (N= 345/413). \_\_\_\_\_127

Table 4.5: Motif enrichment analysis for 3'SS. 299 bps sequences were collected from intronic side of 5'SS for retained intron of length  $\geq 600$  (N=345/413). \_\_\_\_\_128

Table 4.6: Contribution of factors in regulation of intron retention. GC frequency= GC frequency, width (Kb) = intron length (Kb), logFC= RPKM(TBI)/RPKM(Control) and ACACACA frequency= frequency of ACACACA motif near the 5'SS and 3'SS allowing for 1 mismatch. \_\_\_\_\_129

## LIST OF FIGURES

Figure 1.1.4.1: Mechanism of action of Lead (Pb) exposure	7
Figure 2.1.1: Methylation profiles of DNA extracted from whole blood for male and female children are different from each other.	19
Figure 2.1.2: Lead causes common and gender-specific changes in DNA methylation.	20
Figure 2.1.3: The representative UCSC genome browser plot for candidate genes.	22
Figure 2.1.4: DNA methylations at nuclear encoded mitochondrial genes are affected by Pb.	23
Figure 2.2.1: Control analysis to account for the effect of covariates such as gender and blood cell distribution.	38
Figure 2.2.2: Maternal prenatal exposure to lead (Pb) and its effect on the child's neonatal and current blood.	39
Figure 2.2.3: UCSC genome browser pictures of differentially methylated region detected in CNBS with high BLL in MNBS.	40
Figure 2.2.4: Validation of HM450K using in-vitro studies.	43
Figure 3.1: Beta value distribution of HMeDIP-450K data	59
Figure 3.2: Characterization of CpG sites in putative 5hmC clusters.	61
Figure 3.3: Significant differentially hydroxymethylated clusters (DhMRs)	62
Figure 3.4: Validation of putative 5hmC clusters	63

Figure 3.5: Examples of CpG sites mapping to high density 5hmC region (hMRs) regions which overlapped with HMeDIP-Seq dataset (GEO accession: GSM1008199).	64
Figure 3.6: Correlations and overlap with validation sets	65
Figure 3.7: Sex-dependent difference in 5hmC and 5mC clusters.	65
Figure 3.8: Representative Pb-dependent 5hmC clusters for UCB DNA from infants prenatally exposed to Pb	66
Figure 3.9: Number of differentially hydroxymethylated (DhMRs) and methylated (DMRs) detected for UCB DNA samples in conserved, male-specific and female-specific regions.	67
Figure 4.1.1: Drosophila brain is bilaterally symmetrical (Keene and Waddell, 2007).	76
Figure 4.1.2: Schematic diagram of the MTBI (modified TBI) device. It is designed to ameliorate the effect of TBI.	77
Figure 4.1.3: Graphical representation of the process of alternative splicing process.	78
Figure 4.2.1: Exploratory analysis of RNA-sequencing profile and long term survival estimation in Drosophila model of TBI.	98
Figure 4.2.2: Characterization of alternative splicing events in Drosophila model of TBI	100
Figure 4.2.3: Sashimi plot showing intron retention event in Isocitrate dehydrogenase (IDH)(A) and stoned A/B (STNA/B) (B).	101
Figure 4.2.4: Intron retention causes degradation or nuclear retention of transcripts	103
Figure 4.2.5: Association between H3K36me3 and Alternative splicing in modEncode data.	105
Figure 4.2.6: RNAi of KDM4A causes Intron Retention (RI)	107

Figure 4.2.7: sashimi plot showing intron retention event in Isocitrate dehydrogenase (IDH) (A) and Stoned A (STNA) (B). For KDM4A-RNAi samples. _____	108
Figure 4.2.8: H3K36me3 is involved in regulating splicing changes in the 24 hours post-TBI. _____	110
Figure 4.2.9: Mutation of Smooth (SM) results in complete loss of RI post-TBI _____	112
Figure 4.2.10: Sashimi plot for RI events which are reversed in SM-mutants. _____	113
Figure 4.2.11: Enrichment of H3K36me3 centered ( $\pm 1000$ ) around CA-rich motifs. _____	114
Figure 4.2.12: Pictorial depiction of Percent Intron Retention (PIR) calculation using R/Bioconductor. _____	115
Figure 4.2.13: Percent change in RI 2 days and 7days post-Spinal Cord Injury (SCI). _____	115
Figure 4.2.14: Characterization of RI/exclusion events detected 7days post-SCI. _____	116
Figure 4.2.15: Model for intron retention in mouse SCI dataset. _____	117
Figure 4.2.16: Homology between Drosophila Smooth (SM), isoform U and human hnRNPL splicing factor. _____	118
Figure 4.2.17: Modelling RI and effect of covariate in 24 hours post-TBI samples. _____	120
Figure 4.2.18: Model for regulation of alternative splicing 24 hours post-TBI. _____	121

## **CHAPTER 1: DNA METHYLATION AS A PUTATIVE BIOMARKER FOR ENVIRONMENTAL EXPOSURE.**

### **1.1. Background**

#### **1.1.1. DNA methylation is critical for genomic stability and spatial temporal regulation of gene expression.**

In recent times, DNA methylation has emerged as one of the most widely studied epigenetic modifications. DNA methylation is the addition of a methyl (-CH<sub>3</sub>) group to 5' position of the DNA base cytosine. This reaction is mediated by a group of enzymes called DNA methyltransferases (DNMTs). The mammalian genome encodes for 3 DNMTs; DNMT1, DNMT3A, DNMT3B and one regulatory protein called DNMT3L, which lacks the catalytic domain. It is generally accepted that DNMT3A and 3B are responsible for the de-novo methylation or establishment of DNA methylation patterns prior to implantation (Okano et al., 1999; Challen et al., 2014). On the other hand, DNMT1 is responsible for maintenance of existing DNA methylation pattern in co-ordination with DNMT3A (Okano et al., 1999; Gowher et al., 2005). Studies have reported a wide range of function for DNMTs including maintenance of genomic stability, regulation of expression and DNA damage repair.

The human genome consists of about 40% of repetitive elements. Expression of these repetitive elements can lead to widespread genomic instability by promoting chromosome fragility, silencing of genes corresponding to their location or sequestering of factors essential for transcription, splicing and translation. Active LINE1 (L1) retro-transposable elements have been reported to be involved in 1 out of every 1000 spontaneous disease-producing insertions in the human genome (Hancks and Kazazian, 2016). Besides introduction of disease causing mutation expression of repeats has been reported to play

important regulatory role in several biological processes. A study by Coufal et al, 2009 suggested that expression of L1 elements might be responsible for generation of “somatic mosaicism” in the human brain and is important for differentiation of neural cells (Coufal et al., 2009; Singer et al., 2010). Regulation of expression of L1 and other repetitive elements has been reported to be inversely correlated with their methylation status. For example, mouse studies have reported that triple knockdown of DNMTs in mouse ESCs can result in loss of methylation and de-repression of in Intracisternal A particles (IAPs) (Matsui et al., 2010; Karimi et al., 2011). Alongside, suppression of retro-transposable elements DNA methylation has been associated with regulation of gene expression. The association between DNA methylation and gene expression has been reported to be dependent on the location of the methylation sites with respect to gene. Increased methylation in the promoter and 5'Untranslated region (UTR) of genes has been associated with decrease in expression of genes. This association was shown as early as 1979, when J. D. McGhee and G. D. Ginder showed that the methylation status of the promoter of the beta-globin gene inversely correlated with its expression (McGhee and Ginder, 1979). Later studies demonstrated that the methylation of promoters, recruit's methyl-binding factors such as MECP2. This results in the inhibition of binding of cis-acting transcription factors and inhibition of expression (Nan et al., 1998). This promoter associated methylation changes, usually occur in regions of high frequency of CG dinucleotides called as CpG islands (CGI). It has been estimated that approximately ~60% of human genes have CGI near promoter regions. Gene-bodies on the other hand have low density of CGIs and are extensively methylated. This has been associated in the past as markers for transcribed genes. For example, Shotgun bisulfite sequencing, which distinguishes methylated and un-methylated cytosines, suggested positive

correlation between active gene transcription and gene-body methylation (Yang et al., 2014). Other than regulation of gene and repeat expression, DNA methylation is speculated to be associated with other regulatory processes such as alternative splicing. 5mC is known to interact with CTCF (CCCTC-binding factor) and MECP2 and been reported to decrease the rate of elongation of RNA Polymerase 2 (RNA pol2) (Maunakea et al., 2013; Lev Maor et al., 2015). Decreased elongation rate of RNA pol2 has been shown to increase the efficiency of splicing specifically the processing of introns (Khodor et al., 2011). Additionally, increase in DNA methylation has been reported to be associated with increase in H3K9me3 and Heterochromatin Protein 1 (HP1). HP1 can directly interact with splicing factors such as serine/arginine-rich splicing factor SRSF1 and hnRNPs and consequently impact regulation of alternative splicing (Yearim et al., 2015).

### **1.1.2. DNA hydroxymethylation is a stable epigenetic modification.**

5mC can be further oxidized by Ten eleven translocases (TET) in a  $\alpha$ -Ketoglutarate and  $Fe^{2+}$  dependent manner to form 5-Hydroxy-Methylcytosine (5hmC). The mammalian genome encodes 3 TET proteins; TET1, 2 and 3. TET1 and 2 are mainly expressed in embryonic stem cells (ESCs) (Koh et al., 2011; Wu et al., 2011b), whereas, TET3 is highly expressed in germ line cells (Shen et al., 2014a). All 3 TET enzymes share a common dioxygenase domain and is capable of the 5mC hydroxylase activities. This systemic redundancy suggests that TET expression might be necessary for survival. Recent studies have reported that knockdown of TET in mice result in compromised development of the offspring and in majority of cases perinatal lethality. This suggests that TET is important for embryonic developments. Further in-depth investigations have revealed that perinatal lethality is mainly caused due to loss of paternal genome oxidation by TET3 expressed by the



oocytes. TET expression and consequent oxidation of 5mC is also important in somatic cell differentiation and function particularly in the brain (Sun et al., 2014; Wen et al., 2014). Studies have shown relatively high expression levels of TET3 in developing brain (Szwagierczak et al., 2010). High level of TET3 expression was correlated with relatively higher 5hmC levels in developing neurons specifically in the gene bodies of highly expressing genes (Santiago et al., 2014). These and several other studies suggest that 5hmC is a stable epigenetic modification and might have important role to play in gene regulation. Similar to 5mC, effects of change in 5hmC levels on gene expression and other associated function is dependent on the genomic location of the modification relative to the genes. Considering the 5hmC is directly downstream of 5mC, and is in essence a bulky DNA modification, it is expected to have complimentary functions. Interestingly, 5hmC has been reported to be specifically enriched near promoters containing chromatin transcription permissive; H3K4me3 and repressive; H3K27me3 marks; also known as bivalent promoters with high GC content but low CpG density (Yu et al., 2012). Such promoters are categorized to be in a “poised” state, ready to be transcribed into mRNA. 5hmC have also been reported to be enriched in intergenic cis-regulatory elements, such as active enhancers and insulator-binding sites (Stroud et al., 2011). Therefore, TET mediated oxidation of 5mC to 5hmC seems to be an important genomic modification with gene regulatory function.

### **1.1.3. DNA methylation and hydroxymethylation as potential early biomarkers for environmental exposure.**

The epigenetic profile of the embryo is dynamic and in a state of continuous flux. Post-fertilization, the paternal genome undergoes replacement of male protamine with female histones and depletion of repressive chromatin modification H3K9me2/me3 and H3K27me3

(Jenkins and Carrell, 2012). This permits the active demethylation of the paternal genome by the action of TET enzymes. The maternal genome on the other hand undergoes passive loss of methylation due to absence of DNMT1; maintenance methyl-transferase. This genome-wide loss of DNA methylation progress past morula to the blastocyst stage. During the Blastocyst stage the DNA methylation profile is re-established and participates in the process of differentiation into somatic cells.

The de-programming and reprogramming of the DNA methylation profile during development, is especially susceptible to environmental cues. Environmental exposure to toxicant during this process can cause persistent changes in the epigenomic profile and contribute to the development of diseases in adult life (DOHaD; Developmental Origins of Health and Disease hypothesis) (Wadhwa et al., 2009). In recent years' mouse models and epidemiological studies have provided compelling evidence of the association between antenatal exposure and DNA methylation. Kile et al, 2012, reported a small yet significant increase in umbilical cord blood (UCB) DNA methylation in LINE1 repetitive regions on infants exposed to Arsenic (Kile et al., 2014). Studies in similar target tissue (i.e. UCB) have also reported general hypermethylation of CPG loci within CpG islands with increasing arsenic exposure (Lu et al., 2014). Studies with other heavy metals such as Lead (Pb) and Mercury (Hg) also revealed significant associations with DNA methylation. For example, Wright and colleagues have reported inverse correlation between patella bone Pb levels in mothers and DNA methylation of LINE1 repeat elements in UCB, suggesting that methylation might serve as a marker for past Pb exposure (Wright et al., 2010). Studies of global expression patterns and their correlation with DNA methylation in mouse models of prenatal exposure to Pb have revealed a significant association between an increase in DNA

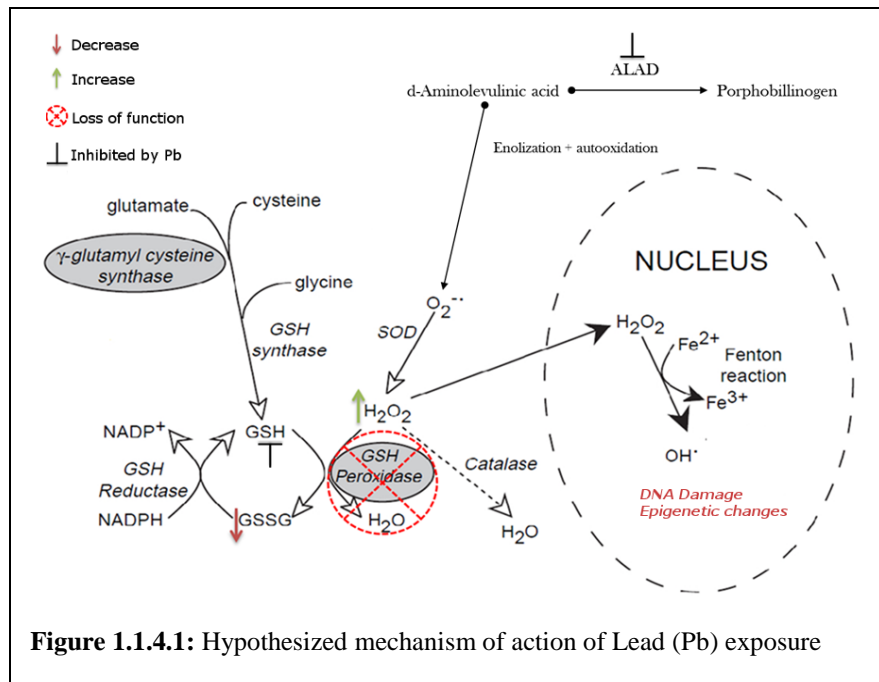
methylation and transcriptional repression of genes associated with the immune response, metal binding, metabolism and transcription (Dosunmu et al., 2012). Hanna et al, 2012 demonstrated increase in methylation of GSTM1/5 promoter on exposure to mercury (Hg) and decrease in methylation in the in COL1A2 promoter, an important component of the chorio-amniotic membrane and the uterine cervix, that correlate with high exposure to Pb in the whole blood of women undergoing in vitro fertilization (IVF) (Hanna et al., 2012). However, prenatal exposure associated changes in the epigenome can be widespread and high variable. It can be influenced by a large number of external factors including diet, xenobiotic exposures to other chemical, stress, lifestyle factor, gestation period and mother's age of conception and internal factors such as tissue type, and sex of the infant to list a few. Therefore, in view of the large background variability in DNA methylation profiles, definition of a set of genomic sub-features which can be used biomarkers of early environment exposure in the new-born is critical.

One of the major candidates for ideal environmental sensors could be Imprinted genes. Imprinting marks are established very early during development and is most likely constant in several tissues including easily accessible tissues such as peripheral blood. The usefulness of imprinted regions as environmental sensors have been demonstrated by several studies. For example, Heijmans et al. 2008 reported hypomethylation of imprinted gene IGF2 in Dutch famine survivors, who suffered from severe in-utero caloric restriction (Heijmans et al., 2008). Genomic imprinting alteration has also been noted in Imprinting Control Regions (ICR) of H19 in mouse embryos exposed to ethanol during preimplantation development (Haycock and Ramsay, 2009).

Besides DNA methylation, a few studies have indicated the possible role of DNA hydroxymethylation in mediating long-term effect of environmental exposure. Zhang et al, 2014 reported a marked increase in total 5hmC levels in the heart and spleen of rats exposed to physiologically relevant levels of Arsenic (As) through drinking water (Zhang et al., 2014). An association between arsenic metabolism and global DNA hydroxymethylation has also been reported in humans (Tellez-Plaza et al., 2014). Therefore, DNA hydroxymethylation might also be a possible candidate for early genomic biomarkers of environmental exposure. However, despite of the progress made, till date, determining a suitable and robust biomarker remain a significant challenge in epigenetic epidemiology.

#### 1.1.4. Lead(Pb) is potent neurotoxin and mediates its adverse effects on DNA methylation profile via generation of reactive oxygen species (ROS).

Pb is generally accepted as a potent neuro-toxicant. In an in-vitro model of neuronal differentiation, Senut et al, 2014, demonstrated that exposure to Pb during formation of neural



rosettes resulted in loss of expression of neuronal markers PAX6 and MS11 (Senut et al., 2014). The resulting Neuronal precursor cells (NPCs) differentiated into neurons with shorter neurites and less branching than control non-exposed neurons. However, Pb exposure

associated changes in neuronal morphology was only noted at high Pb concentrations (>1.9 $\mu$ M). This suggests that effect of Pb exposure on differentiating neurons at low exposure is much subtler and may become apparent over a long period of time. We speculate that the long term effect of Pb exposure is most likely mediated by continued generation of ROS. ROS has been known to cause extensive DNA damage and mitochondrial damage which might result in increased susceptibility to adult diseases. Pb in its divalent form is known to covalently bind to sulfhydryl groups. This allows Pb to bind to the sulfhydryl group of Glutathione (GSH) and reduce its ability to be oxidized to GSSG. Availability of oxidized glutathione is essential for the cells ability to cope with hydrogen peroxide (H<sub>2</sub>O<sub>2</sub>) generated during the cellular processes. Additionally, depletion of cellular stores of GSH causes the cell to start making glutathione from cysteine via  $\gamma$ -glutamyl cycle which is inefficient and may cause several secondary effects. Pb is also a very potent inhibitor of delta-Aminolevulinic acid dehydratase (ALAD) and Ferrochelatase. These enzymes are critical in heme synthesis. Inhibition of ALAD by Pb can potentially increase the amount of ALA (Aminolevulinic acid). Accumulation of ALA can lead to its enolization and consequently auto-oxidation leading to generation of superoxide ion.

The different hypothesized mechanisms of action of Pb are illustrated in Fig 1.1.4.1. Therefore, through inhibition of the anti-oxidant defense system and increase superoxide level Pb can cause extensive DNA damage and associated epigenetic changes. Several studies have suggested a likely association between DNA methylation and generation of ROS. ROS are well-known to cause oxidation of DNA generating 8-oxoguanine (8-OxoDG). Weitzman et al, 1994 demonstrated using synthesized oligonucleotides that replacing guanine with 8-oxoguanine resulted in hypomethylation of surrounding CpG sites by altering

the binding of the methylating enzymes (Weitzman et al., 1994). A follow-up study by Turk et al, 1995, confirmed these finding using human methyltransferases and suggested that the effect of 8-OxoDG on cytosine methylation is variable depending on the relative position of the cytosine with respect to CG dinucleotide(Turk et al., 1995). A study by O'Hagan et al, 2011 reported the recruitment of large protein complex near sites of DNA damage consisting of DNMT1, DNMT3b and SIRT1, a histone deacetylase (O'Hagan et al., 2011). Wang et al, 2008 demonstrated that SIRT1 is involved in DNA repair (Wang et al., 2008). In this study Wang and colleagues injected WT and SIRT1 null MEF cells with micro-homologous DNA damage repair reporter, pGL2 Luc vector. Then they treated the cells with EcoRI restriction enzyme which cut within the coding sequence of the luciferase gene. Following incubation, the WT cells were able to re-acquire ~70% of the original luciferase reporter level, compared to 42% for SIRT1 null cells. Another study by Fan et al, 2010, demonstrated that SIRT1 is responsible for deacetylation of Xeroderma pigmentosum group A (XPA) is a core nucleotide excision repair (NER) factor (Fan and Luo, 2010). The deacetylation of XPA was also shown improve it interaction with RPA32 and regulate the DNA repair process. Therefore, co-occurrence of DNMTs alongside a DNA repair protein SIRT1 suggests a close interplay between DNA methylation and DNA damage. In view of the evidence from these and other studies, we believe that Pb exposure alters DNA methylation by ROS mediated mechanism.

## **CHAPTER 2: THE HUMAN DNA METHYLOME IS RESPONSIVE TO ENVIRONMENTAL LEAD (PB) EXPOSURE.**

### **2.1. Early life Lead exposure causes distinct gender specific changes in the DNA methylation profile of DNA extracted from dried blood spots. (Sen et al., 2015b).**

#### **2.1.1. Background**

The human epigenome is particularly responsive to environmental cues especially during developmental years. Therefore, exposure to environmental toxin in early childhood may result in extensive changes in the epigenome which may culminate in increased susceptibility to adult diseases. Among epigenetic modifications, DNA methylation is perhaps the most widely studied and long been proposed to be excellent early biomarkers of environmental exposure. These exposure dependent DNA methylation changes have been reported to be localized in LINE1 and SINE1 repeat elements, transcription start sites, imprinted regions etc. and shown to have varied functional consequences. For example, LINE1 repeat elements were shown to be hypo-methylated in Umbilical cord blood (UCB) of mothers with high patella bone Lead (Pb) levels (Wright et al., 2010). This can result in activation of LINE1 retrotransposons and consequently increase genomic instability (Cordaux and Batzer, 2009). Occupational and environmental exposure to Arsenic has been shown linked to hypermethylation of promoter of Tumor suppressor gene p16 (Hossain et al., 2012; Lu et al., 2014). This is expected to recruit methyl-binding proteins such as MECP2 to the promoter regions and inhibit assembly of the transcription initiation complex and lead to silencing of the gene (Nan et al., 1998). Li et al, 2016, suggested using DNA methylation levels of imprinted genes as early stable biomarkers of Pb exposure (Li et al., 2016). Imprinted genes are genes which are expressed in a parent of origin specific manner. DNA methylation levels in these genes remain relatively stable from one generation to the next and

alteration due to in-utero exposure is has been associated with susceptibility to heritable disorders.

Another defining characteristic of DNA methylation is that it is dependent upon the sex of the individual. A study by Faulk et al, 2013 in mouse models reported hypermethylation of the A(vy) locus in male offspring of Viable yellow agouti (A(vy)) female mice exposed to low levels of Pb (2.1ppm) before conception through weaning(Faulk et al., 2013). This suggested that male offspring might be more sensitive than females to environmental cues. In agreement with this study other mouse models have reported similar associations. For examples males were reported to be more susceptible to hypertension in intrauterine undernutrition (do Carmo Pinho Franco et al., 2003). In human cohorts, a study by Lindburg et al, 2008 reported that female participants showed higher efficiency of arsenic metabolism compared to age-matched males providing further evidence of sex-specific responses to environmental exposures (Lindberg et al., 2008). Therefore, in view of the evidence so far we hypothesize that DNA methylation changes in response to acute Pb exposure are sex-specific.

### 2.1.2. Methods

**Recruitment/consent procedures:** Seventy-five children (3 months to 5 years of age) and their biological mothers (75) were enrolled into the study through routine visit at the WIC (Women Infant and Children). Clinics chosen for recruitment from Southwest Detroit, were selected with guidance from DHWP (Detroit Health and Wellness Department). Note: recruitment fliers were placed in each location announcing when the study personnel would be at the location to consent for the study. The biological mothers of the children were asked if they would like to enroll in the study. All of the following are exclusion criteria: 1)



Mothers born before January 1, 1987; 2) Mothers born outside of Michigan; 3) children 6 years of age and older; 4) Children who are not the biological children of the mothers; 5) biological children who were born outside of Michigan; 6) Non-English speaking individuals and those who are not fluent in English.

Approximately half (38) of the children enrolled had BLL at or above the Center for Disease and Control blood lead level of concern (equal to or greater than 5  $\mu\text{g}/\text{dL}$ ). The remainder of the children will have lower BLLs.

A finger stick blood draw was taken from the child and mother to determine BLL by placing a sample of the blood in a Lead Care Analyzer. Results were reported to the mother within 15 minutes of the blood draw. The child's BLL was reported to the Michigan Department of Community Health within a week of the testing.

The mothers of the children were asked to complete a demographic and environmental questionnaire after enrollment in the study. The questionnaire will be self-administered; staff will be present to assist any participant as needed.

**Samples and sample classification:** For the study, we selected 43 dried blood spots collected from children from Health Fairs ran in three Detroit communities, Rosa Parks, Chene, and Kettering-Butzel, because they have a high prevalence (8-11%) of high BLL in children. The study only included mothers born after January 1, 1987 in Michigan with biological children ages, 3 month to 5 years also born in Michigan. The final sample used in this study consisted of 25 male children and 18 female children. Among males 15 children had  $\text{BLL} \geq 5\mu\text{g}/\text{dl}$  and among females 11 children had  $\text{BLL} \geq 5\mu\text{g}/\text{dl}$ . The covariate and BLL information is available is Table 2.1.1.

**Lead measurements in dried blood spots:** The samples were 3 mm punch-outs of blood spots on filter paper. The samples, control filter paper punch-out blanks without blood, blanks without filter paper, and standards were all prepared in parallel at the same time in the same way, except that five times as much of the standard solutions were made to allow sufficient volume for recalibration every 20 samples. To each container was added ultra-high purity 0.25 ml of 70% HNO<sub>3</sub> (0.25 ml) and 30% H<sub>2</sub>O<sub>2</sub> (0.05 ml). These reagents were purified in an in-house sub-boiling still from reagent grade materials. After reacting overnight, samples were diluted with 10 ml of 0.5% HNO<sub>3</sub> containing 5 ppb each of Ga and Bi as internal standards. Standards (50 ml) were spiked with 0.1 ml of a solution of Fe, Mg, Zn, and Pb, prepared from 1000 ppm single element ICP-MS stock solutions (Inorganic Ventures). The final standard concentrations were Fe, 200 ppb, Mg, 20 ppb, Zn 2 ppb, Pb, 1 ppb. All containers were exposed to vacuum (~0.03 bars) for ten minutes to reduce the amount of dissolved oxygen from H<sub>2</sub>O<sub>2</sub> dissociation. Even so, some oxygen bubbles were seen during analysis in the sample line to the ICP-MS. Filter paper fibers settled to the container bottoms and did not affect the ICP-MS sample introduction system. Samples were analyzed on a PerkinElmer Elan 6100 DRC ICP-MS instrument at the Geology Department, Union College. <sup>25</sup>Mg, <sup>66</sup>Zn, and <sup>206,207,208</sup>Pb were analyzed in normal mode, and <sup>54</sup>Fe was analyzed in DRC mode using NH<sub>3</sub> reaction gas at a flow rate of 0.5 ml/minute.

**Extraction, shearing and denaturation of DNA:** DNA was isolated from dried blood spots with Qiagen EZ1 Advanced® using the DNA Investigator® reagents and protocol card. The “Stains on Fabric” preprocessing and Trace® (tip-dance) instrument protocol was used for isolation. The Quantifiler Human DNA Quantification Kit® (Applied Biosystems, Inc.) was used to determine the amount of amplifiable DNA.

Approximately 3 $\mu$ g genomic DNA was diluted in 130 $\mu$ l of buffer TE (10mM Tris (pH 8.0), and 1mM EDTA (pH 8.0)) and sheared into ~200-600 bp fragments using microcavitation (Covaris, Inc, setting: Duty Cycle = 5%, Intensity = 3, Cycles/burst = 200, Time = 75 seconds run at 6-8°C). 125 $\mu$ l of the sheared DNA samples were mixed with 330 $\mu$ l of buffer TE. The sheared DNA was denatured by boiling it in the Thermomixer at 95°C and 700rpm for ten minutes and left on ice for 10 minutes.

**HM450K bead chip array:** For this study we measured the change in DNA methylation on environmental exposure to Pb in dried blood spots using the Illumina Human Methylation 450K Bead chip array (HM450K). The HM450K assay measures DNA methylation at over ~480,000 CpG and non-CpG sites with single-base resolution (Jin and Warren, 2003; Sandoval et al., 2011). The results are represented in the form of beta ( $\beta$ ) values ranging from 0 to 1, and provide a quantitative measure of methylation for each queried CG dinucleotide methylation site (CpG site) (Sandoval et al., 2011). DNA methylation changes at CpG sites located close to each other often exhibit common behavior in response to environmental stimuli. These regions show highly correlated changes in methylation signatures and can be defined as co-regulated regions. The co-regulated regions can be assigned to specified clusters and the effect of the exposure on these clusters can be tested using the generalized estimating equation (GEE) (Sofer et al., 2013). GEE uses a weighted combination of observation to measure the effect of a covariate (in our case Pb exposure) while conserving the correlation structure of the data (Petronis and Anthony, 2003). Consequently, this approach is much less conservative and yields a greater number of differentially methylated regions compared to the traditional case-control study with a single CpG site  $\beta$  value comparison. Biochemically it makes more sense to study methylation

changes as clusters rather than single CG sites because DNA bound DNMT1 and 3a act on multiple sites in a small region when they bind to DNA (Jurkowska et al., 2011) .

Detection of methylated DNA is facilitated by two different probe types (Type 1 and Type 2 probes). The Type 1 probes or the Infinium 1 (Inf1) probes consist of the methylated bead and an un-methylated bead (Bibikova et al., 2011). If the probe for methylated DNA matches the target site, there is a single base extension which results in detection which signals into the red channel. Similarly, if an un-methylated probe binds to the DNA it signals into the green channel. The type 2 probes or the Infinium 2(Inf2) queries both methylated and unmethylated DNA on a single bead, and the ratio of incorporation of two differently-colored fluorescent nucleotides (Signals A and B) determines the methylation signal. The results are represented in the form of  $\beta$  values, specifically, the average  $\beta$  value (AVG\_Beta), representative of the average methylation level of the CpG dinucleotide, and a delta  $\beta$  value which signifies the difference in methylation levels between the control and the experimental group. The Beta or  $\beta$  for the  $i^{th}$  interrogated CpG nucleotide is:

$$\text{Beta}_i \text{ or } \beta_i = \frac{\text{Max}(y_{i,\text{methy}}, 0)}{\text{Max}(y_{i,\text{unmethy}}, 0) + \text{Max}(y_{i,\text{methy}}, 0)}$$

Where  $y_{i,\text{methy}}$  and  $y_{i,\text{unmethy}}$  are the intensities measured by the  $i^{th}$  methylated and unmethylated probes, respectively. Illumina recommends adding a constant offset  $\alpha$  (by default,  $\alpha= 100$ ) to the denominator to regularize  $\beta$  value when both methylated and unmethylated probe intensities are low. The Beta-value statistic results in a number between 0 and 1, or 0 and 100% (Du et al., 2010). The raw data was retrieved from Genome Studio methylation module version 1.8<sup>TM</sup> in the form of 2 files; a sample methylation profile and

control probe profile. Quality control, signal correction and normalization of the data was carried out using the HM450K BeadChip data processing pipeline proposed by Teschendorff et al, 2013 in R environment (R> 2.13.0) (Teschendorff et al., 2013). Several studies have indicated that the Infinium 1 and 2 probes differed in chemistry, henceforth the HM450K are two separate experiment combined as one. The Infinium 1 probes were shown to have a more stable signal and extended dynamic range compared to the Infinium 2 probes (Touleimat and Tost, 2012). Therefore, a 3-state beta mixture model is utilized to assign methylation values to specific methylation states. Then the probability of assignment to particular state is divided in quantiles and finally a methylation dependent dilation transformation is performed to preserve sample monotonicity (Teschendorff et al., 2013). Prior to analysis the beta values were corrected for Batch effect using Combat function in R and potential single nucleotide polymorphism (snp)-containing probes were removed from the analysis ( $>2.15$ ) (Johnson et al., 2007).

The HM450K array is highly reliable for locus specific methylation detection at CpG island associated methylation sites which are frequently associated with dynamic regulation during development and disease states. Bibikova et al, 2011, conclusively showed using human sample (lung tissue) that the HM450K array show 95 to 96% correlation with Whole genome bisulfite sequencing (WGBS) results(Bibikova et al., 2011). Since then several studies have explored this correlation and found this array to be highly consistent. Therefore in view of recent evidence, we believe that HM450K array can “stand on its own” as an independent system for differential methylation analysis. That said, to look at the whole genome methylation status in an un-biased way sequencing based approaches are still the

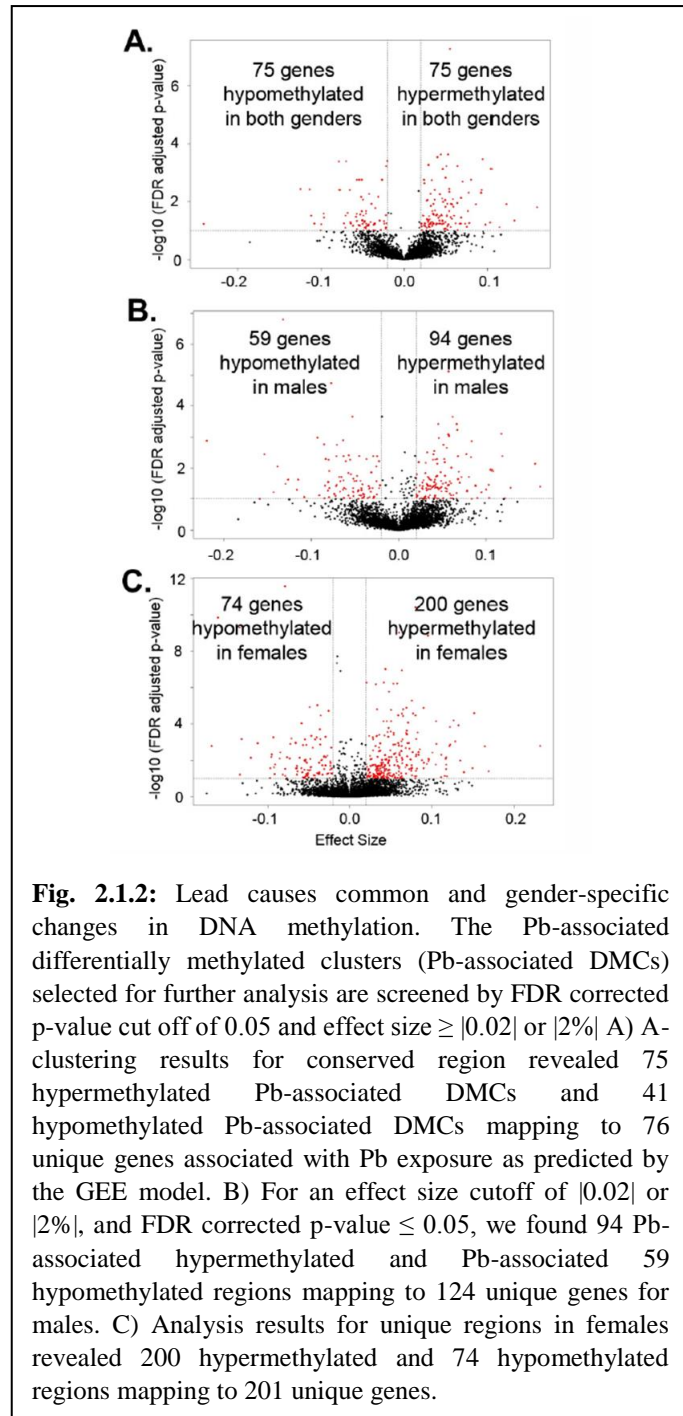
best methodology. When we have a much larger cohort, then we could use the less expensive PCR assay to validate the CpG sites that we identify in this pilot study.

**Statistical analysis:** For studying effects of exposure to Pb on the DNA methylation profile of UCB in male and female samples we used two independent statistical approaches. The first approach known as CpG association analysis (Barfield et al., 2012), which analyzes DNA methylation data using a fixed effect model at single CpG site level and adjacent site clustering algorithm. The second approach is called A-clustering which detected sets of correlated CpG sites and then tested the clusters for multivariate response to environmental exposure to Pb using the generalized estimation equation (GEE). The aforementioned approach is efficiently implemented using the R-package Aclust (Sofer et al., 2013).

For determining the differentially methylated clusters (DMCs) we used the recommended Aclust parameters; Spearman correlation, for calculating the distance between adjacent sites ( $\text{dist}_{(i,j)} = 1 - \text{corr}_{(i,j)}$ ), average clustering type, which require that mean distance between two sites be at least 0.25, 1000 bp distance restriction for merging of clusters, which ensures that clusters located far away from each other are not merged together based on correlation. The clustering approach is implemented with a 999 bp merge initiation step, which clusters all sites wedged between 2 high correlated sites within 999 bp of each other together, to reduce the complexity of data and the analysis time for the Aclust step. Finally, the data was analyzed using a generalized estimation equation approach and filtered for significant DMCs using FDR corrected p-value cutoff = 0.05 and exposure effect size  $\geq |0.02|$ . To determine the genomic locations of the probes belonging to individual DMCs, they were annotated using the publicly available Illumina Human Methylation 450k annotation data in R (>2.15). The target genes mapping to DMCs were individually visualized using

UCSC genomic browser. The Delta beta or the beta difference between the median of the beta values for each probe for low BLL samples and high BLL samples were mapped by the chromosomal location of the probes. If A-clustering is an effective technique for DMR identification, we hypothesized that the change in methylation status visualized using UCSC genome browser will correspond to the exposure effect (i.e. increase or decrease in methylation) predicted by GEE in the respective regions and might serve as a useful tool for visualization. Gene ontology for mapped DMCs was carried out using Hypergeometric testing implemented by the package GOstats in R (>2.15).

A-clustering takes into consideration that adjacent CpG sites are probably co-regulated by Pb exposure, therefore the differential methylation calls are made based on multiple probes rather than a single CpG site, making it considerably more



**Fig. 2.1.2:** Lead causes common and gender-specific changes in DNA methylation. The Pb-associated differentially methylated clusters (Pb-associated DMCs) selected for further analysis are screened by FDR corrected p-value cut off of 0.05 and effect size  $\geq |0.02|$  or  $|2\%|$  A) A-clustering results for conserved region revealed 75 hypermethylated Pb-associated DMCs and 41 hypomethylated Pb-associated DMCs mapping to 76 unique genes associated with Pb exposure as predicted by the GEE model. B) For an effect size cutoff of  $|0.02|$  or  $|2\%|$ , and FDR corrected p-value  $\leq 0.05$ , we found 94 Pb-associated hypermethylated and Pb-associated 59 hypomethylated regions mapping to 124 unique genes for males. C) Analysis results for unique regions in females revealed 200 hypermethylated and 74 hypomethylated regions mapping to 201 unique genes.

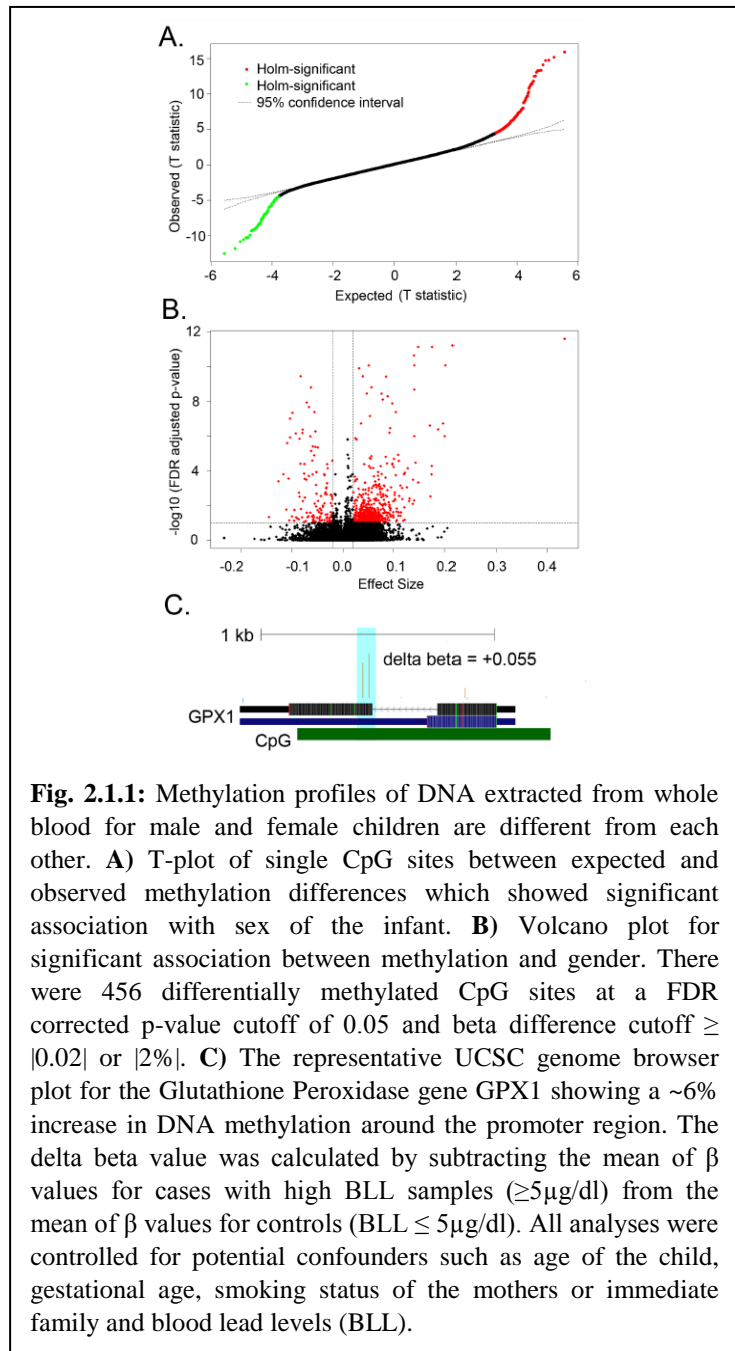
reliable compared to generalized linear model based differential methylation calling algorithm such as limma(Wessely and Emes, 2012). Moreover, the presence of multiple CpG sites with altered methylation states, in regions such as transcription start sites (TSS), is more likely to cause altered transcription factor binding and affect expression. A significant problem associated with clustering based approach is the possible introduction of gene-specific methylation bias due to varied number of probes mapping to genes. In the A-clustering based approach, the clusters of probes are based on the base-pair distance between adjacent probes, and beta value correlations across samples, but not in the context of the genomic features. Therefore, overall the number of probes in each cluster contributes very little to the differential methylation calls.

### **2.1.3. Results**

#### **Sex specific differences in DNA methylation regardless of Pb exposure**



To delineate the sex specific effects in DNA methylation we modelled the  $\beta$  values for single CpG dinucleotides as a function of sex of the individual while controlling for variations in BLL between samples and other factors such as age of the children, gestational age, age of the mother, and smoking status of the mother and immediate family (Table 2.1.1). We restricted these analyses to the autosomes because females have one inactive X chromosome with increased promoter DNA methylation (i.e., the Barr Body (Felsenfeld, 2014)). We observed 456 CpG sites which were differentially methylated between males and females at a FDR corrected P-value cutoff of 0.05 and effect size cutoff of  $|0.02|$  or  $|2\%|$  (Fig. 2.1.1 A). 365 CpG regions were hypermethylated and



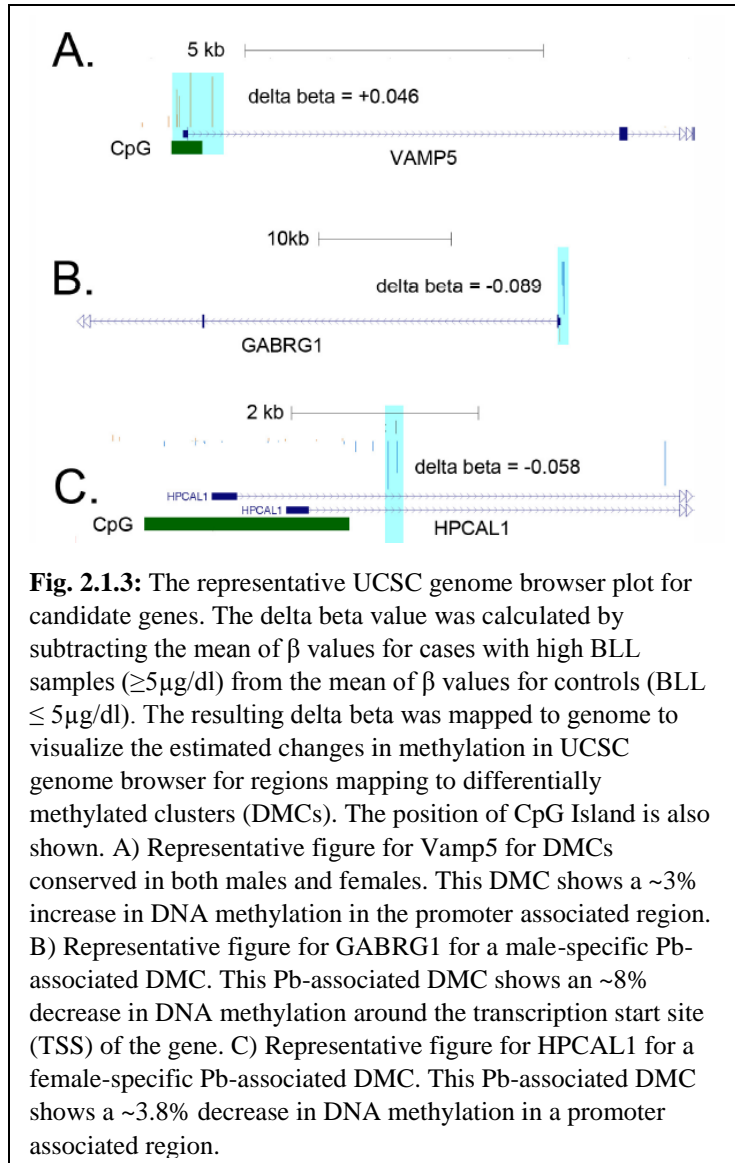
91 CpG regions were hypomethylated in females compared to males (Fig 2.1.1 B). GO mining based on overrepresentation analysis (hypergeometric testing) showed enrichment of

hypermethylated CpG regions in genes associated with neurogenesis (GO:0022008), neuronal differentiation (GO: 0030182) and oxygen and reactive oxygen species metabolic process (GO:0072593) (data not shown) such as Glutathione Peroxidase (GPX1) (Table 2.1.2 and Fig 2.1.1 C), Cytochrome P450 (CYP1A1) (Table 2.1.2), and Superoxide Dismutase 3 (SOD3) (Table 2.1.2). Hypomethylated CpG regions in females were associated with basic regulatory processes such as double-strand break repair (GO:0006302) and aerobic respiration (GO:0009060) (data not shown). We hypothesize that these differences in methylation might underlie some of the gender specific differences in the sensitivity to Pb exposure.

### **Pb exposure associated DNA methylation changes in both males and females.**

We detected sex-specific differences in DNA methylation profile. However, there might be regions of the genome which respond to Pb-exposure similarly in males and females; conserved regions. To define these regions, we aggregated the CpG sites with well correlated  $\beta$ -values and located within 1000bps into DNA methylation clusters. Then we tested the impact of high BLL ( $\geq 5\mu\text{g/dl}$ ) on these clusters using Generalized Estimating Equations (GEE) while controlling for covariates such as age of the children, gestational age, age of the mother, and smoking status of the mother and immediate family. At FDR corrected p-value cut off of 0.05 and effect size  $\geq |0.02|$  or  $|2\%|$  we found 75 hypermethylated Pb-associated DMCs and 38 hypomethylated Pb-associated DMCs mapping to 75 unique genes as predicted by the GEE model (Fig. 2.1.2 A). Gene ontology analysis of the genes mapping to Pb-associated DMCs showed overrepresentation of genes associated with differentiation of myeloid lineages (data not shown).

We interrogated the DNA methylation changes in a few representative genes by mapping the  $\Delta\beta$  values for individual CG dinucleotides to human genome (hg19) and visualized them using genome browser. One interesting gene which showed an hypermethylation in the CpG island located near the Transcription initiation sites was Vesicle-Associated Membrane Protein 5 (VAMP5) (Fig 2.1.3A and table 2.1.3). We also observed significant DNA methylation changes in genes associated with mitochondrial metabolism such as



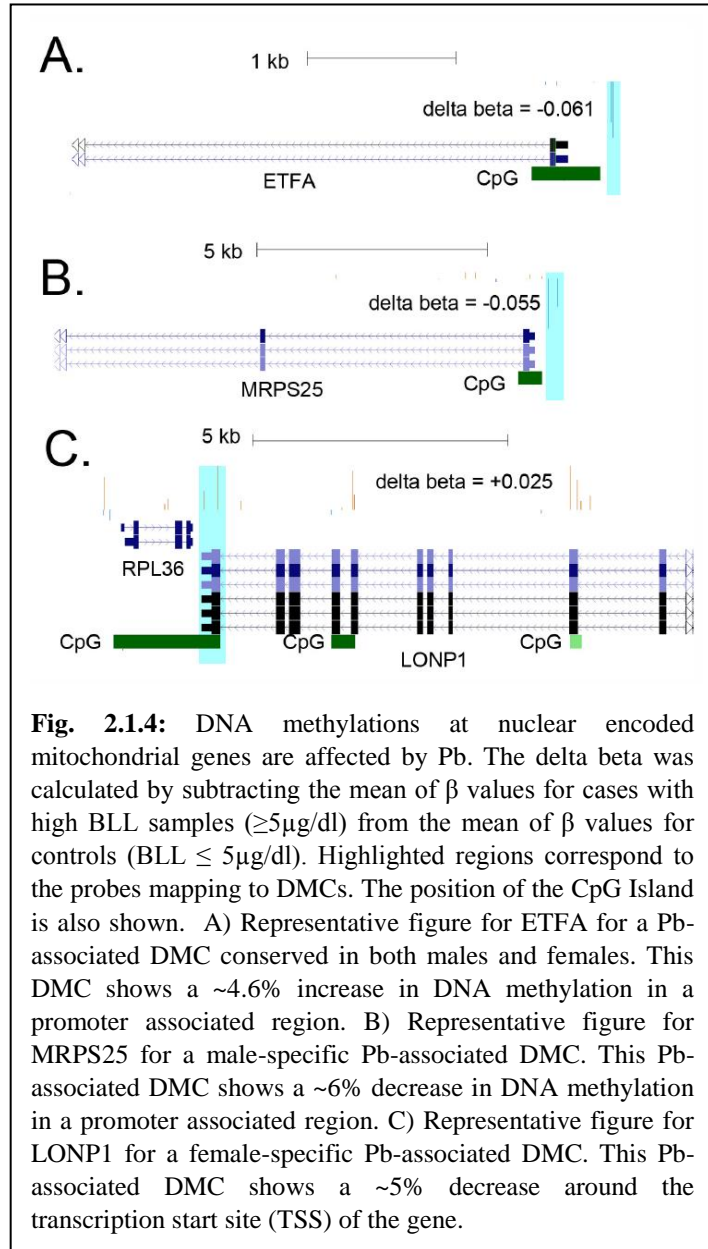
hypomethylation in the CpG Island located in the promoter region of the gene encoding the alpha subunit of Electron Transfer Flavoprotein (ETF $\alpha$ ) (table 2.1.3 and Fig 2.1.4A). These DNA methylation changes and their possible functional implications are further discussed in later sections.

## Sex-specific effects of Pb exposure on DNA methylation

The primary objective of our study was to delineate the sex specific Pb-exposure changes in DNA methylation profile. Therefore we separated the male and female samples.

We reclustered our CpG sites and reanalysed our data using GEE while controlling for potential contributing factors such as age of the children, gestational age, age of the mother, and smoking status of the mother and immediate family. For an effect size cutoff of  $|0.02|$  or  $|2\%|$ , and FDR corrected p-value  $\leq 0.05$ , we found 94 Pb-associated hypermethylated and 59 hypomethylated regions in males (Fig 2.1.2 B). Gene ontology analysis of genes mapping to male-specific Pb-associated differentially-methylated regions show an enrichment of genes associated with

leukocyte proliferation and differentiation (GO: 1902107) and calcium ion transport (GO: 0051924) (data not shown). Examples of males-specific genes that are differentially methylated by Pb are Runt-Related Transcription Factor 1 (RUNX1) (Table 3), Gamma-



Aminobutyric Acid (GABA) A Receptor, Gamma 1 (GABARG1) (Table 2.1.3 and Figure 2.1.3B) and Mitochondrial Ribosomal Protein S25 (MRPS25) (Table 2.1.3 and Figure 2.1.4B).

We also see female-specific changes in DNA methylation that are associated with Pb exposure. At p-value of  $\leq 0.05$  and exposure effect size of  $\geq |0.02|$  or  $|2\%|$  we found 200 hypermethylated and 74 hypomethylated regions in females exposed to Pb (Fig. 2.1.2C). We performed gene ontology analysis and found enrichment of hypermethylated genes for pathways such as neuron maturation (GO:0042551) and visual learning (GO:0008542) and hypomethylated genes for pathways such as regulation of type 2 immune response (GO:0002828) (data not shown). Other interesting genes which also showed a Pb-associated differential methylation included stress response genes such as Amyloid Beta (A4) Precursor Protein (APP) (Table 2.1.3) and Hypoxia inducible factor 3A (HIF3A) (Table 2.1.3), nuclear encoded mitochondrial genes such as Lon Peptidase 1, Mitochondrial (Lonp1) (Table 2.1.3 and Figure 2.1.4C) and Mitochondrial Transcription Termination Factor 4 (MTERFD2) (Table 2.1.3), transcriptional regulators like Runt associated transcription factor 3 (RUNX3) (Table 2.1.3), important regulators of signaling pathways such as Mitogen-Activated Protein Kinase 6 (MAP3K6) (Table 2.1.3) and neuronal calcium sensors such as Hippocalcin-Like 1 (HPCAL1) (Table 2.1.3 and Fig. 2.1.3C).

#### **2.1.4. Discussion**

Sex-specific response of the epigenome to exposure has been described by several studies (Pilsner et al., 2012; Faulk et al., 2013). However, few studies have described potential gene-specific biomarkers of Acute Pb exposure. In this study we attempt to answer this question using blood spots collected from 3 months to 5 years old children. Assessing

Pb-dependent DNA methylation in peripheral blood poses several challenges such as impact of dietary factors on mothers such as folate deficiency, exposure to other environmental toxicants, and cigarette smoking (Hsiung et al., 2007), variability in DNA methylation profile introduced due to shift immune cell population during early age among other. We tried to control for possible secondary effectors by including covariates such as age of the children, gestational age, age of the mother, and smoking status of the mother and immediate family during statistical model building. To control for impact of immune cells on DNA methylation in the blood we estimated blood cell type proportion using methylation data (Houseman et al., 2012). Our analysis predicted that blood cell type only contributed to  $\geq 1\%$  of the total variance in methylation profile (data not shown). In our study we observed several sexually dimorphic promoters belonging to genes associated with oxidative stress response and metabolism such as glutathione peroxidase encoding gene (GPX1) (4 to 5% higher promoter methylation in females) and Cytochrome P450 1A1 (CYP1A1) (5% higher methylation status around the TSS in females). Interestingly the CYP1A1 promoter methylation has been reported to sexually dimorphic in primary cells cultured from embryos of the Swiss Webster CWF mouse strain (Penalozza et al., 2014).

Pb is a potent inducer of oxidative stress. It has been shown to bind to sulfhydryl group of Glutathione (GSH) and reduce its ability to be oxidized GSSG. Availability of oxidized glutathione is essential for the cells ability to cope with hydrogen peroxide ( $H_2O_2$ ) generated during the cellular processes. Additionally, depletion of cellular stores of GSH causes the cell to start making glutathione from cysteine via  $\gamma$ -glutamyl cycle which is inefficient and may cause several secondary effects. Pb is also a very potent inhibitor of Delta-Aminolevulinic Acid Dehydratase (ALAD) and ferrochelatase. These enzymes are

critical in heme synthesis. Inhibition of ALAD by Pb can potentially increase the amount of ALA (Aminolevulinic Acid). Accumulation of ALA can lead to its enolization and consequently auto-oxidation leading to generation of superoxide ion. Therefore, perhaps not surprisingly we saw sex specific Pb –dependent DNA methylation changes in genes associated with ALA metabolism. In this study we report an increase in gene body DNA methylation of Lon peptidase 1 (LONP1) in females. LONP1 has been implicated in breakdown of 5-Aminolevulinic acid synthase enzyme in the mitochondrial outer membrane (Tian et al., 2011). Genebody methylation has been demonstrated to be associated with increased expression (Ball et al., 2009). We speculate that this might be an adaptive DNA methylation change which helps ameliorate the synthesis of ALA and alleviate the oxidative stress burden due to Pb exposure. Consequently, LONP1 genebody methylation on Pb-exposure might confer reduced susceptibility to oxidative stress in females.

In our previous epidemiological study using the NHANES cohort we observed a negative correlation between BLL and body mass index (BMI) (Padilla et al., 2010). Interestingly, in our study we observed sex-independent hypermethylation of the Leptin (LEP) promoter. A study by Shen et al, 2014 demonstrated that increase in DNA methylation at LEP promoters can recruit Methyl Binding Protein 2 (MBP2) leading to decreased recruitment of RNA polymerase 2 and decreased transcription in adipose tissue of obese mice (Shen et al., 2014b). Hypermethylation of LEP promoter in sample with high BLL is contradictory to the results of our epidemiological survey however does suggest an association between exposure and metabolic disorders. Consistent with the idea of Pb effecting the epigenome of metabolic regulator we found male-specific decrease in methylation in a CpG island located near the distal transcription start site of a gene encoding

an important hematopoiesis controlling transcription factor, RUNX1 and female-specific Pb-dependent hypermethylation in another member of Runt domain associated transcription factor, RUNX3. RUNX3 deficiency is shown to be associated with myeloproliferative disorder in mouse (Wang et al., 2013). Wolff *et al.* reported that the DNA methylation of RUNX3 increases with age and the process is further accelerated by smoking (Wolff et al., 2008). In combination with previous data, our data suggest that epigenetic regulation of genes associated with controlling haematopois (e.g. RUNX1 and RUNX3) can play a role in mediating the effect of Pb exposure on immune response in a sex-specific fashion.

Finally, as Pb is a potent neurotoxin we expected to see changes in methylation status of neuronal associated genes. Previous studies have reported females to be much more resistant to neurodegenerative conditions compared to males. If Pb-dependent DNA methylation changes are protective females were expected to have a higher number of significant Pb-associated DNA methylation clusters. We report hypermethylated CpG clusters several genes related to biological processes such as neuronal maturation (GO:0042551), visual learning (GO:0008542) and regulation of neurotransmitter levels (GO:0001505) in females. In contrast male showed hypomethylation of promoter region of a single Gamma-Aminobutyric Acid (GABA) A Receptor, Gamma 1 (GABRG1). Therefore, our results suggest that most acute Pb exposure dependent DNA methylation changes are protective in nature.



## Tables

**Table 2.1.1:** Sample BLL ( $\mu\text{g}/\text{dl}$ ) and covariates for current blood spots used for the analysis.

ID	Normalized concentration	Pb Gender	Age (in months)	Mother's	Gestational	Smoking
				age (in months)	age (In months)	
001-033-001-2012-B	5.098809195	Male	3	228	225	No
001-009-001-2012-B	1.609016955	Male	5	240	235	Yes
001-010-001-2012-B	4.71525216	Male	5	240	235	No
001-001-001-2012-B	0.579901708	Male	6	204	186	No
001-113-002-2012-B	4.602492159	Female	8	228	208	No
001-130-003-2012-B	4.378173778	Male	9	264	243	No
001-122-002-2011-B	1.972338713	Female	12	240	228	Yes
001-093-002-2011-B	2.204395525	Female	12	204	192	No
001-051-001-2011-B	2.27245527	Male	12	216	204	No
001-058-004-2010-B	2.772809748	Female	12	252	240	No
001-060-002-2010-B	3.035195443	Female	12	264	252	Yes
001-084-001-2011-B	5.94181757	Male	12	216	204	No
001-036-002-2011-B	8.502613523	Female	12	264	252	Yes
001-095-003-2011-B	10.20218454	Male	12	228	216	No
001-099-001-2011-B	11.04437582	Male	12	228	216	No

001-092-003-2010-B	13.0268562	Male	12	252	240	No
001-085-002-2011-B	25.59618366	Female	12	240	228	No
001-027-002-2010-B	0.19009625	Female	24	264	240	No
001-122-001-2010-B	0.856610711	Male	24	240	216	Yes
001-082-001-2009-B	0.871991444	Male	24	216	192	No
001-022-003-2010-B	1.496641473	Male	24	300	276	No
001-087-002-2010-B	2.938008435	Female	24	252	228	Yes
001-033-002-2010-B	16.97667656	Female	24	228	204	No
001-010-002-2009-B	24.20912991	Female	24	240	216	No
001-079-002-2010-B	32.80080749	Female	24	252	228	No
001-036-005-2009-B	5.809591329	Male	36	264	228	Yes
001-015-006-2009-B	6.248038356	Female	36	264	228	No
001-045-001-2008-B	6.860479927	Male	36	252	216	No
001-095-001-2009-B	7.04519292	Male	36	228	192	No
001-092-001-2009-B	8.840076423	Male	36	252	216	No
001-015-002-2010-B	9.640403264	Female	36	264	228	No
001-118-002-2009-B	11.83912714	Female	36	252	216	No
001-022-001-2008-B	2.626116004	Male	48	300	252	No
001-036-001-2007-B	4.735150984	Male	48	264	216	Yes
001-130-001-2007-B	7.626344312	Male	48	264	216	No

001-036-003-2008-B	9.288905445	Male	48	264	216	Yes
001-015-004-2008-B	9.665637279	Female	48	264	216	No
001-130-002-2007-B	10.47605773	Female	48	264	216	No
001-019-003-2007-B	17.22238377	Male	48	276	228	Yes
001-025-001-2007-B	6.436692662	Male	60	252	192	No
001-121-002-2007-B	7.206834813	Female	60	240	180	Yes
001-099-003-2007-B	7.430384158	Male	60	228	168	No
001-019-001-2006-B	9.725285685	Male	60	276	216	Yes

**Table 2.1.2:** Single nucleotide differences in DNA methylation between females and males as estimated by fixed effect model. For the analysis, the females were used as experimental group and males as the control group. All analysis was controlled for covariates such as age of the child, gestational age of the mother, age of the mother, smoking status of the household and Blood Lead Levels.

Cpg Site	Gene	Promoter associated	CpG Island	Effect size	Standard error	P.value	Holm.sig	FDR
cg22584138	SLC6A4	No	chr17:28562387-28563186	0.074	0.012	3.91E-07	FALSE	0.001
cg04555966	GCK	No	chr7:44185961-44186184	0.037	0.007	5.65E-06	FALSE	0.0089
cg09933329	GCK	No	chr7:44185961-44186184	0.056	0.012	4.9E-05	FALSE	0.041
cg03314840	PKLR	No	chr1:155264318-155265536	0.053	0.011	3.93E-05	FALSE	0.0357

cg12177922	HAX1	Yes	chr1:154244710- 154245289	0.052	0.007	1.06E- 08	TRUE	5.44E- 05
cg21763723	GPX1	Yes	chr3:49394855- 49395942	0.042	0.009	2.62E- 05	FALSE	0.027
cg25187648	GPX1	Yes	chr3:49394855- 49395942	0.054	0.007	1.38E- 08	TRUE	6.88E- 05
cg11924019	CYP1A1	No	chr15:75018186- 75019336	0.053	0.011	5.6E- 05	FALSE	0.0448
cg02891686	SOD3	No	chr4:24801109- 24801902	0.045	0.009	8.5E- 06	FALSE	0.0123
cg11304734	POLR1E	Yes	chr9:37485801- 37486099	0.045	0.009	1.47E- 05	FALSE	0.0182
cg14268223	DNAJA1	Yes	chr9:33025082- 33025797	- 0.047	0.009	2.06E- 05	FALSE	0.0237
cg13479204	HOXB3	No	chr17:46641534- 46642110	- 0.082	0.016	1.2E- 05	FALSE	0.0159
cg05323879	HOXB3	No	chr17:46641534- 46642110	0.476	-0.071	2.39E- 05	FALSE	0.0257
cg02325951	FOXP3	No	chr14:89882421- 89884278	- 0.062	0.007	9.55E- 10	TRUE	7.23E- 06
cg26355737	TFDP1	No	chr13:114292551- 114292886	- 0.034	0.006	6.91E- 07	FALSE	0.00174
cg19292611	TFDP1	NO	chr13:114292551- 114292886	- 0.023	0.004	4.24E- 06	FALSE	0.00726

**Table 2.1.3:** Representative gene mapping to clusters which show a change in methylation of  $\geq |2\%|$  or  $|0.02|$ , at a FDR corrected p-value cutoff of 0.05.

Region ID	Gene	CpG Island	Promoter associated	Effect Size	Standard error	P-value	FDR	CpGsites/ cluster
Conserved	VAMP5	chr2:85811340- 85811855	Yes	0.031	0.009	0.000839	0.0410	5
Conserved	CAPN2	chr1:223936342- 223937044	Yes	0.024	0.005	5.94E-06	0.00174	3
Conserved	PF4	chr4:74847528- 74847830	No	0.0933	0.0223	2.38E-05	0.00395	4
Conserved	LEP	chr7:127880750- 127881375	No	0.036	0.01	0.00045	0.0280	2
Conserved	ETFA	chr15:76603563- 76604026	Yes	-0.046	0.013	0.000387	0.0258	2
Male specific	RUNX1	chr21:36258952- 36259472	No	-0.084	0.025	0.00082	0.0373	5
Male specific	GABRG1	Non-CpG	No	-0.08	0.019	4.24E-05	0.00534	5
Male-specific	MRPS25	chr3:15106459- 15106971	Yes	-0.064	0.020	0.00148	0.0497	2
Female specific	APP	Non-CpG	No	0.043	0.013	0.00131	0.0366	4
Female specific	LONP1	chr19:5690127- 5692213	No	0.038	0.012	0.00143	0.0391	2
Female specific	MTERFD2	chr2:242041543- 242042026	Yes	0.021	0.003	8.73E-10	5.43E-07	7
Female specific	HPCAL1	chr2:10442308- 10444509	Yes	-0.055	0.017	0.00111	0.0323	2
Female specific	MAP3K6	chr1:27683277- 27683590	Yes	0.061	0.016	8.8E-05	0.00534	5
Female specific	HIF3A	chr19:46800053- 46800603	No	0.06	0.019	0.00150	0.0405	2
Female specific	RUNX3	chr1:25255527- 25259005	No	0.045	0.012	0.00017	0.00859	2

## **2.2. Multigenerational epigenetic inheritance in humans: DNA methylation changes associated with maternal exposure to lead can be transmitted to the grandchildren. (Sen et al., 2015c).**

### **2.2.1. Background**

In the previous section we have reported that acute exposure to heavy metal neurotoxicant Lead (Pb) can cause significant changes in DNA methylation in peripheral blood in children aged 3 months to 5 years. We also suggested that these DNA methylation changes can be used as biomarkers of early exposure to Pb. However, this study was limited to a single generation and did not address whether Pb-dependent DNA methylation changes can be epigenetically inherited and transmitted to the consequent generations.

Epigenetic inheritance of DNA methylation has been previously reported by several mouse models. For example, feeding pregnant Agouti<sup>viable yellow</sup> (A<sup>vy</sup>) mice a diet rich in methyl donors causes hyper-methylation of the Intracisternal-A-Particle (IAP) transposable element in offspring. This causes the agouti/black mottling in offspring in the direction of the pseudoagouti phenotype (Wolff et al., 1998). Exposure to fungicide Vinclozolin in F<sub>0</sub> female rats has been shown to correlate with changes in DNA methylation status of genes involved in testis formation in F<sub>3</sub> males (Skinner et al., 2013). Prenatal exposure to alcohol (PAE) has been demonstrated to cause a complex disease called fetal alcohol spectrum disorder (FASD). Adult mice born to mothers exposed to PAE have been reported to show enhanced activity of the hypothalamic-pituitary-adrenal (HPA) axis when exposed to stress (Lee et al., 2008). Studies in human cohort with 3-6 year of children diagnosed with FASD showed significant DNA methylation changes in Protocadherin genes (Laufer et al., 2015). Therefore, in view of the evidence from several studies we hypothesized that Pb exposure dependent DNA methylation changes are stable and can be inherited by later generations.

During embryonic development, the maternal and paternal genome has been shown to undergo extensive DNA demethylation. Consequently, the DNA methylation patterns are re-established during the post – blastocyst stage and continue throughout post-implantation. We hypothesized the exposure to Pb during this stage may cause stable changes in the methylome of the germ cells. As the germ cells remain relatively undifferentiated during development and are only activated during consequent pregnancy, the impact of Pb exposure on DNA methylation may also skip a generation and be transmitted from the grandmother to the grandchildren through the mother. To test this hypothesis, we selected 35 dried blood spots (DBS) collected from mother-infant pairs in Detroit. For these samples we also collected the neonatal DBS and mother neonatal DBS from the Michigan Neonatal Biobank. The expectation was that if our hypothesis is correct, then the Blood Pb levels (BLL) of the mother's neonatal DBS (representative of the grandmother's BLL during pregnancy) would be correlated with significant changes in DNA methylation in the grandchildren's neonatal DBS and the current DBS irrespective of their BLL and their mother's BLL (Figure 2A).

### 2.2.2. Methods

**Cell culture and treatment:** The human ESC line WA09 (H9) was obtained from the WiCell Research Institute (Madison, WI, USA) and maintained in a humidified incubator at 37°C with 5% CO<sub>2</sub>, as previously described (Senut et al., 2014). Briefly, undifferentiated hESCs (passages 26-39) were cultured in DMEM/F12 supplemented with knockout serum replacement, nonessential amino acids, penicillin/streptomycin, L-Glutamine, 2-mercaptoethanol, and human basic fibroblast growth factor (Life Technologies) on a feeder layer of irradiated mouse embryonic fibroblasts (GlobalStem). hESCs were passaged by mechanical dissociation every 4-6 days and their pluripotency frequently tested by

immunofluorescence staining for specific markers including Oct4 and Lin28 Stock solutions (100-fold concentrated) of Pb acetate ( $\text{Pb}(\text{C}_2\text{H}_3\text{O}_2)_2$ ) (Sigma-Aldrich) were prepared in sterile distilled water. Physiologically relevant concentrations of Pb acetate chosen on the basis of our previous work (Senut et al., 2014) were tested in this study:  $1.5\mu\text{M}$  ( $32\mu\text{g/dL}$ ). Distilled water was used as a vehicle control. Undifferentiated hESCs were acutely exposed to the different concentrations of Pb or vehicle for 24 hours, at which time the hESC colonies were dissected and their DNA was isolated.

**Samples and sample classification:** Methods were carried out in accordance with guidelines that were approved by the Wayne State University (WSU) Institutional Review Board (IRB), the Michigan Department of Community Health (MDCH) IRB, and the Michigan Neonatal Biobank (MNB) IRB. Informed consent was obtained from all subjects enrolled for the study. For the study, we selected 35 dried blood spots (DBS) collected from mother-infant pairs from Health Fairs ran in three Detroit communities, Rosa Parks, Chene, and Kettering-Butzel, because they have a high prevalence (8-11%) of high BLL in children. The study only included mothers born after January 1, 1987 in Michigan with biological children ages, 3 months to 5 years also born in Michigan. The final sample sets consisted of 25 male children and 18 female children. We also collected the neonatal DBS and mother neonatal DBS for these mother-infant pairs from the Michigan Neonatal Biobank. Blood Pb measurement from 3mm punches using Atomic absorption spectroscopy. The corresponding Pb measurement and covariate information is listed in Table 2.2.3.

**Extraction, shearing and denaturation of DNA:** DNA was isolated from dried blood spots with Qiagen EZ1 Advanced® using the DNA Investigator® reagents and protocol card. The “Stains on Fabric” preprocessing and Trace® (tip-dance) instrument



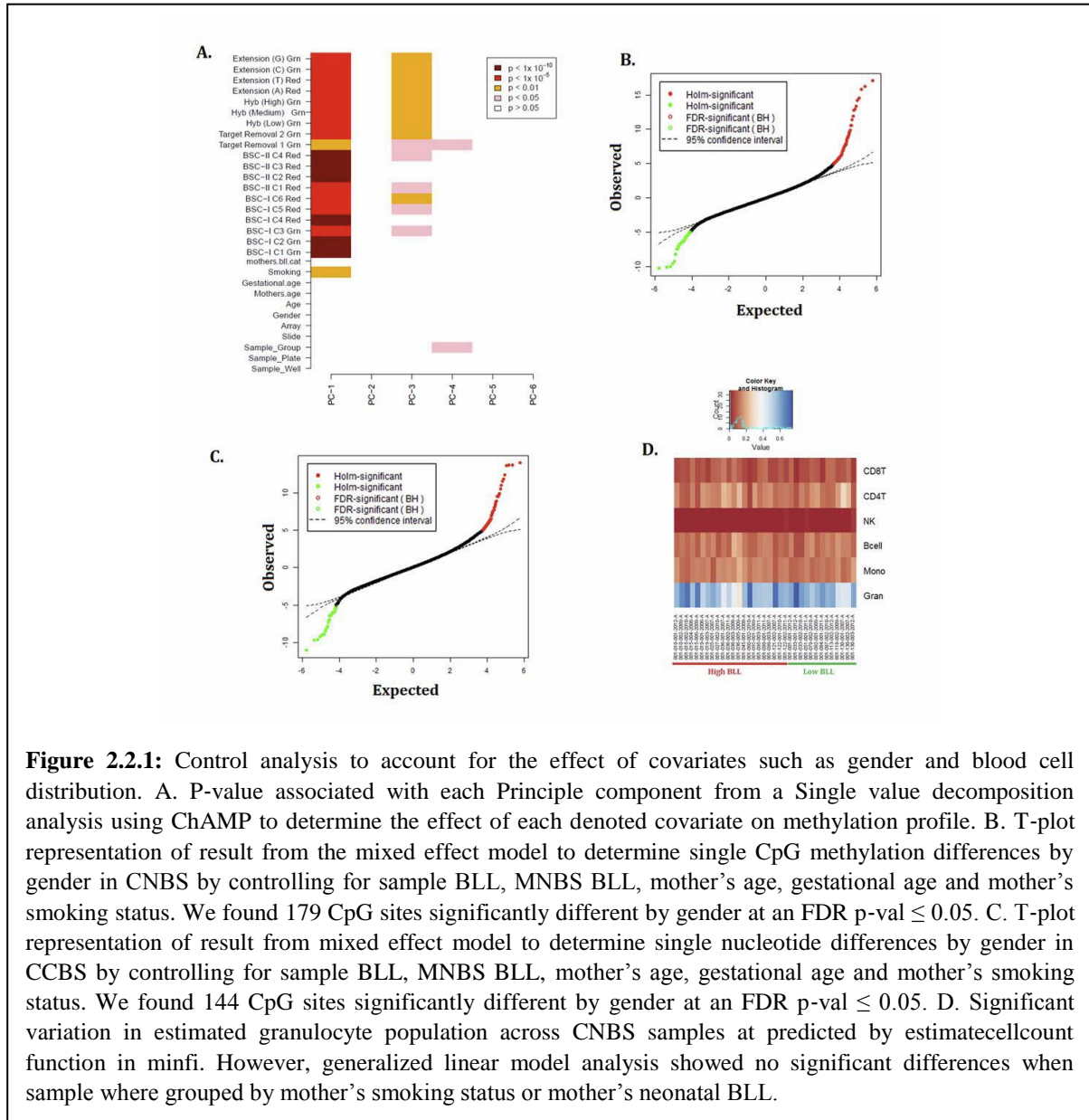
protocol was used for isolation. The Quantifiler Human DNA Quantification Kit® (Applied Biosystems, Inc.) was used to determine the amount of amplifiable DNA.

Approximately 3µg genomic DNA was diluted in 130µl of buffer TE (10mM Tris (pH 8.0), and 1mM EDTA (pH 8.0)) and sheared into ~200-600bp fragment using microcavitation (Covaris, Inc, setting: Duty Cycle = 5%, Intensity = 3, Cycles/burst = 200, Time = 75 seconds run at 6-8°C). 125µl of the sheared DNA samples were mixed with 330µl of buffer TE. The sheared DNA was denatured by boiling it in the Thermomixer at 95°C and 700rpm for ten minutes and left on ice for 10min.

**HM450K bead chip array:** For this study we measure the DNA methylation at >450,000 sites using the Human methylation 450K array. For more detail please refer to Methods 2.1.2; sub-heading HM450K bead chip array.

**Statistical analysis:** For studying the gender specific effects we CpG association analysis (Barfield et al., 2012), which analyzes DNA methylation data using a mixed effect model at single CpG site level. For determining the Pb-dependent changes in DNA methylation we used adjacent site clustering algorithm, A-clustering to detect sets of correlated CpG sites and then tested the clusters for multivariate response to environmental exposure to Pb using the generalized estimation (GEE) equation approach. The aforementioned approach is efficiently implemented using the R-package Aclust (Sofer et al., 2013). For determining the differentially methylated clusters we used the recommended Aclust parameters; Spearman correlation, for calculating the distance between adjacent sites ( $\text{dist}_{(i,j)} = 1 - \text{corr}_{(i,j)}$ ), average clustering type, which require that mean distance between two sites be at least 0.25, 1000 bp distance restriction for merging of clusters, which ensures that clusters located far away from each other are not merged together based on correlation. The

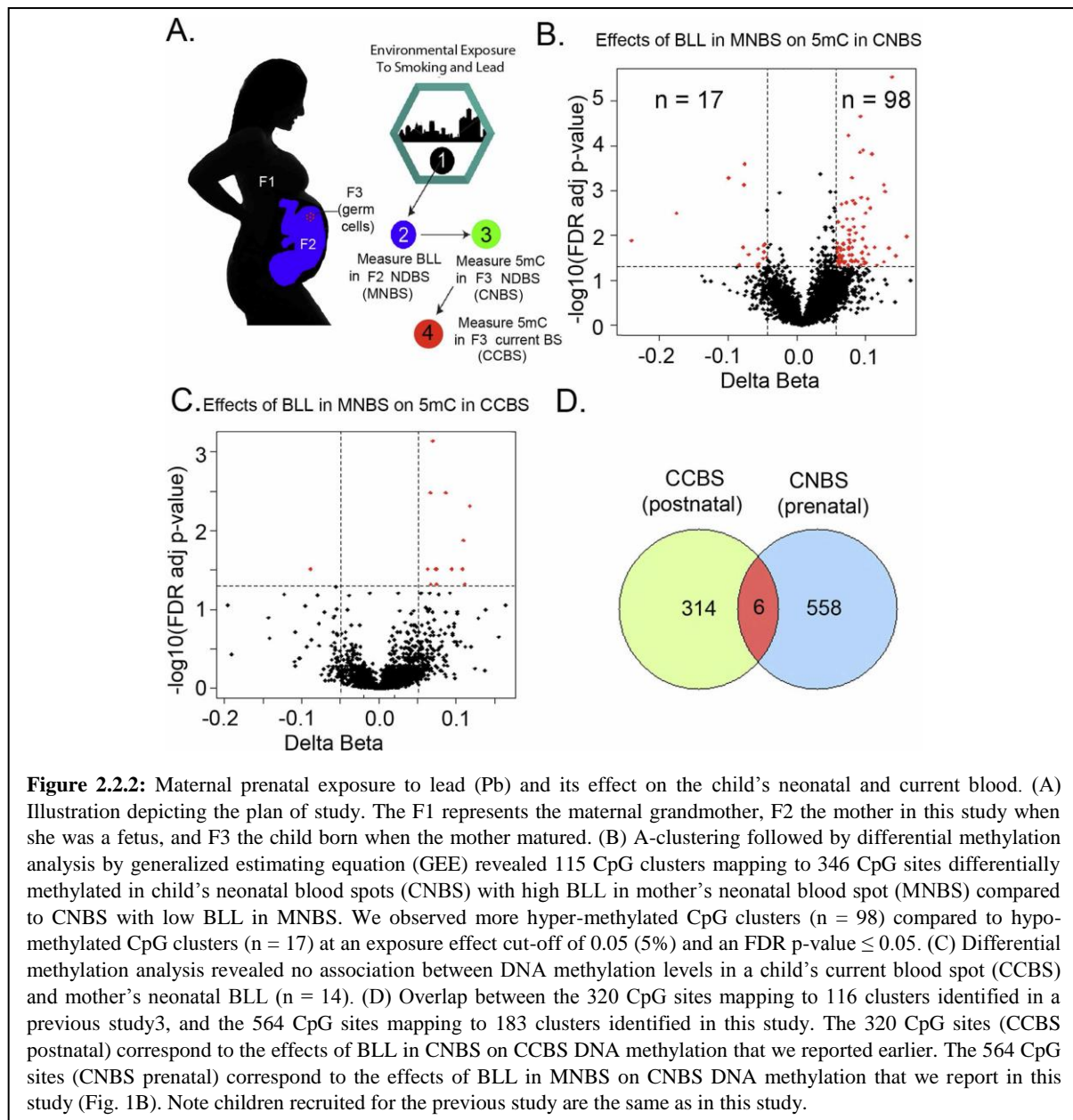
clustering approach is implemented with a 999bp merge initiation step, which clusters all sites wedged between 2 high correlated sites within 999 bps of each other together, to reduce the complexity of data and the analysis time for the Aclust step. Finally, the data was analyzed using a generalized estimation equation approach and filtered for significant DMCs using FDR corrected p-value cutoff = 0.05 and exposure effect size  $\geq |0.02|$ . To determine the genomic locations of the probes belonging to individual DMCs, they were annotated using the publicly available Illumina Human Methylation 450k annotation data in R (>2.15). The target genes mapping to DMCs were individually visualized using UCSC genomic browser. The Delta beta or the beta difference between the median of the beta values for each probe for low BLL samples and high BLL samples were mapped by the chromosomal location of the probes. If A-clustering is an effective technique for differentially methylated identification, we hypothesized that the change in methylation status visualized using UCSC genome browser and Integrative genome viewer (IGV) will correspond to the exposure effect (i.e. increase or decrease in methylation) predicted by GEE in the respective regions and might serve as a useful tool for visualization. Single value decomposition (SVD) was implemented using the package ChAMP and estimation of cell distribution was implemented using the package minfi. All statistical analysis was implemented in R. The enhancer regions were downloaded as .bed files from Andersson et al, 2014 (Andersson et al., 2014), and the distance from the middle of the robust enhancer sites to the middle of the cluster was calculated using R package ChIPpeakAnno. Then the clusters were subsetted by Maximum distance < 301 and exposure effect size  $\geq |0.05|$ .



**Figure 2.2.1:** Control analysis to account for the effect of covariates such as gender and blood cell distribution. A. P-value associated with each Principle component from a Single value decomposition analysis using ChAMP to determine the effect of each denoted covariate on methylation profile. B. T-plot representation of result from the mixed effect model to determine single CpG methylation differences by gender in CNBS by controlling for sample BLL, MNBS BLL, mother’s age, gestational age and mother’s smoking status. We found 179 CpG sites significantly different by gender at an FDR p-val  $\leq 0.05$ . C. T-plot representation of result from mixed effect model to determine single nucleotide differences by gender in CCBS by controlling for sample BLL, MNBS BLL, mother’s age, gestational age and mother’s smoking status. We found 144 CpG sites significantly different by gender at an FDR p-val  $\leq 0.05$ . D. Significant variation in estimated granulocyte population across CNBS samples at predicted by estimatecellcount function in minfi. However, generalized linear model analysis showed no significant differences when sample where grouped by mother’s smoking status or mother’s neonatal BLL.

Analysis of BS-Seq data: The bisulfite sequencing data was aligned to the BS genome using bismark (Krueger and Andrews, 2011). The \*.sam files from bismark was used to call

5mC in CPG context for non-treated and control samples using R package *methylKit*(Akaline *et al.*, 2012). The methylation calls were further filtered by coverage (min = 10) and possible PCR duplicates were removed by discarding the bases with coverage more than 99.9th



percentile of coverage distribution. Then the number of C's and T's were calculated for the given HM450K co-regulated clusters. Finally, percent (%) methylation for calculated for the clusters to determine the sample

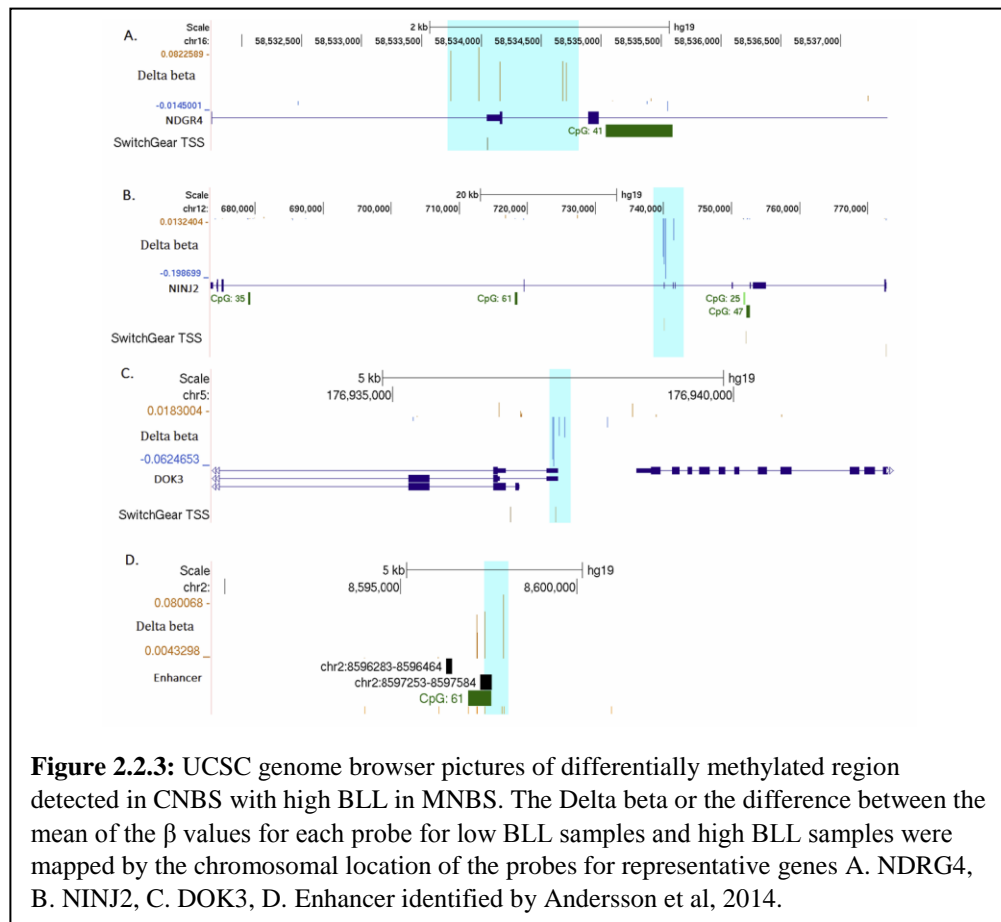
### 2.2.3. Results

#### Measurement of Lead (Pb) concentration in dried blood spots (DBS)

Blood lead levels (BLL) was measured in 3mm punches from mother's and children's neonatal and current dried blood spots by Atomic absorption mass spectrometry (Table 2.2.1). For future reference, we will refer to the child's neonatal blood spots as CNBS, the mother's neonatal blood spots as MNBS, and the child's current blood spots as CCBS (Fig. 2.2.2A).

#### Contribution of covariates to detectable changes in DNA methylation in dried blood spot (DBS).

DNA methylation measured in peripheral blood can be influenced by several factors. Variation in DNA methylation measurement related to



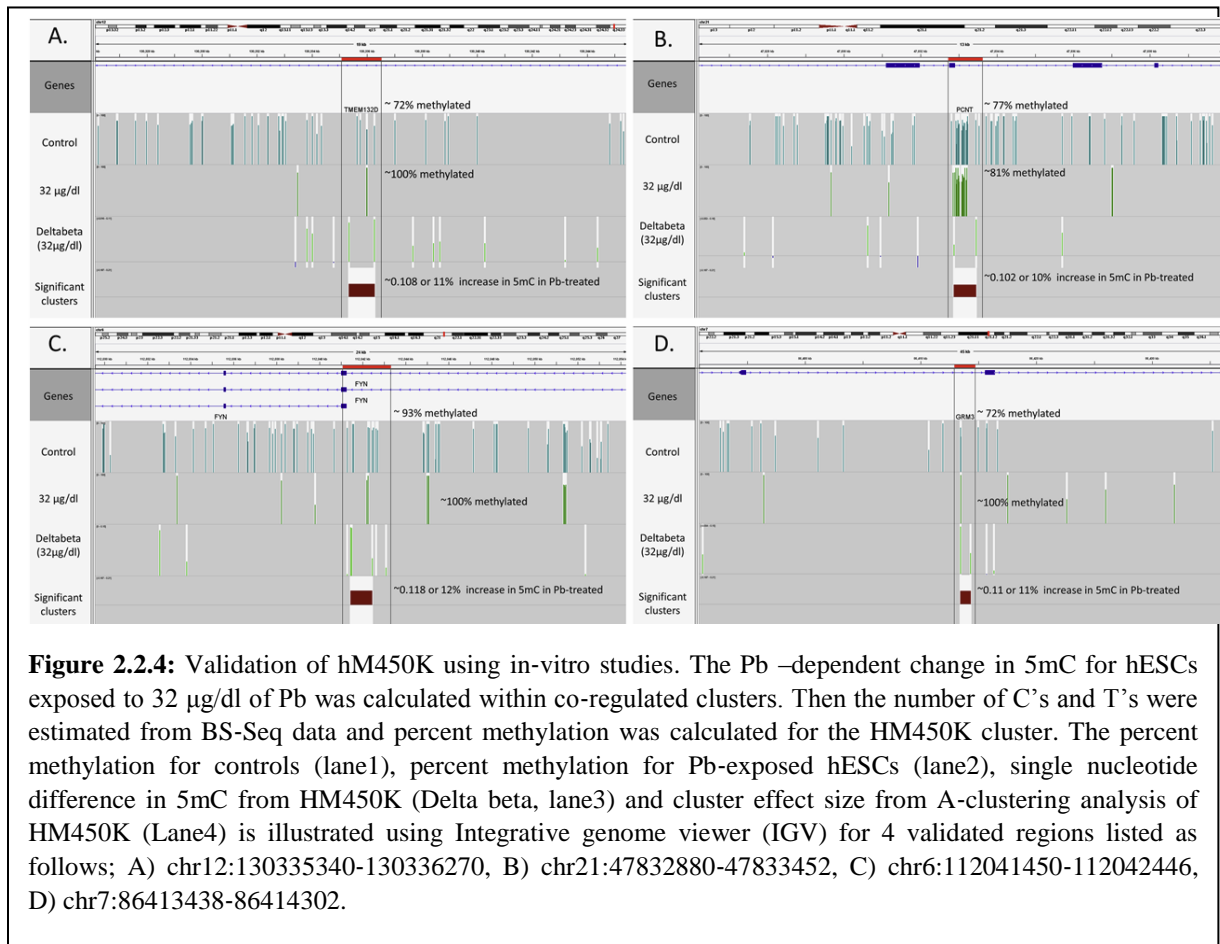
secondary factors needs to be controlled and accounted for before differential testing. We attempted to understand the influence of some of these covariates using Single value decomposition (SVD). Our analysis demonstrated that among all the covariates tested (except experimental variations) smoking was by far the strongest predictor of changes in DNA methylation in peripheral blood (Fig 2.2.1A) in CCBS and CNBS. Interestingly the sample BLL also seemed to contribute the significant variations in methylation profile. Our previous studies have demonstrated that the impact on DNA methylation is sex-specific is small but presents. As SVD was unable to tease out the impact of sex on the DNA methylation profile of peripheral blood we used mixed effect model to estimate single CpG differences. In CNBS, we found 179 autosomal CpG sites with sex-dependent DNA methylation effects at an FDR corrected P-value  $\leq 0.05$ . For CCBS we found relatively similar amounts of differentially methylated CpG sites (N=144) suggesting that DNA methylation changes are probably not established early in development. Finally using a statistical framework proposed by Housemann et al, 2012 we looked at the relationship between blood-cell-type proportion and by mother's smoking status or mother's neonatal BLL in CNBS samples. We chose to study the CNBS samples because the impact of smoking or neonatal high BLL in mother's will perhaps be more pronounce and detectable in a relatively under developed immune system. We did observe large variability in granulocyte population across CNBS samples however; environmental exposure did not impact granulocyte distribution. Therefore, using these control and quality control analysis we were able to define the covariates to use for differential testing.

**Grandmothers' BLL during gestation correlates with gene specific changes in grandchild's DNA methylation levels.**

We hypothesized that DNA methylation changes caused in germ cells due to Pb exposure in grandmothers during gestation will be preserved in the CNBS and CCBS of the grandchild (Fig 2.2.2A). To test this hypothesis, we modelled DNA methylation in CNBS/CCBS as function of BLL in MNBS, surrogates for grandmother's BLL during gestation. The BLL in MNBS was converted to discrete measurements high BLL ( $\geq 5\mu\text{g/dl}$ ) or low BLL ( $\leq 5\mu\text{g/dl}$ ) for ease of calculation. We used sample BLLs, age of the mother, gender of the child, smoking status and gestational age of the mother as covariates for the model. At a significance level (FDR corrected p-value  $\leq 0.05$ ), we found 183 CpG clusters which were differentially methylated in CNBS with high BLLs in MNBS compared to low BLL in MNBS. We observed more hyper-methylated CpG clusters (n = 151) compared to hypo-methylated CpG clusters (n= 32) at an exposure effect cut-off of 0.02 (i.e., a 2% change in DNA methylation). Further increasing the cut-off for differential methylation calls to 0.05 or 5% resulted in significant reduction in identified 5mC clusters (n=115). However, the trend in DNA methylation patterns was preserved, with a greater number of hyper-methylated regions (n =98) compared to hypo-methylated regions (n=17) (Fig. 2.2.2B). Interestingly, similar analysis in CCBS reported limited DNA methylation changes; 14 CpG clusters mapping to 37 CpG sites at exposure effect size cut-off of 0.05/5% or 0.02/2% (Fig 2.2.2C). Some striking examples of these changes are shown in Fig 2.2.3 and are discussed in details in later sections.

## Validation of HM450K array using in-vitro Pb-exposure model.

HM450K array can detect a ~20% changes in DNA methylation status with a confidence of >95%. We modelled our differentially methylated regions as clusters of co-regulated CpG sites located within 1kb of each other (A-clustering). Consequently, this



allowed us to call subtle differences in DNA methylation profile ( $\pm 10\% \geq x \leq \pm 20\%$ ). To validate that A-clustering can detect Pb dependent changes as low as  $\pm 10\%$ , we treated H9 embryonic stem cells (hESCs) with 32µg/dl or 1.5 µM Pb for 24 hours. Then we extracted the DNA and we measured the DNA methylation with HM450K array. 279 clusters significant at a FDR corrected p-value cut-off of 0.05 and exposure effect size cutoff of  $\pm 10\%$  was selected for validation using Whole Genome Bisulfite Sequencing (WGBS). We



mapped the sequences in context of CpG and calculated the % methylation within the HM450K cluster windows (n=279) for Pb-treated samples and non-treated controls. We were able to get significant coverage (min=10 reads/CpG) for 10 out of 279 CpG clusters. 4 out of 10 clusters showed positive association between exposure effect size from HM450K and percent difference in methylation from WBS (Fig 2.2.4).

#### **2.2.4. Discussion**

In this study we present in-direct evidence that Pb-exposure during gestation in grandmothers is correlated with DNA methylation changes in the peripheral neonatal blood in their grandchildren. However, these changes are “normalized” during the course of early development and cannot be detected in child’s current blood. This suggests that DNA methylation is relatively stable during prenatal stages however most likely undergo replication dependent dilution during early post-natal years. We observed limited overlap between the DNA methylation changes in CpG clusters detected on acute exposure to Pb in the CCBS (Sen et al., 2015b) and CpG clusters detected in CNBS in this study. Therefore, this demonstrates that epigenetic effect of Pb exposure is dependent on the stage of development during which the exposure occurs and the duration of the exposure. In an earlier study from our lab, we reported that acute and chronic exposures to Pb during differentiation into neurons have significantly divergent effects on the morphology and DNA methylation status of the developing neurons compared to direct Pb-exposure in differentiated neurons (Senut et al., 2014). Our results in an in-vitro model further emphasizes the impact of exposure time and developmental stage on DNA methylation status.

In this study we detected Pb-dependent changes within ~1kb regions containing co-regulated CpG sites (See methods). This allowed us to detect methylation difference at a cut-

off of  $\pm 10\%$ . To further justify using a lower methylation difference cut-off we validated few of the CpG clusters detected using HM450K using WGBS in a stem cell model of Pb-exposure (Senut et al., 2014). Based on this limited, very expensive and data-intensive WGBS study, we demonstrate that the lower DNA methylation difference detection limit with the HM450K assay can be achieved if multiple CpG sites are modelled into CpG regions using A-clustering.

We observed several interesting DNA methylation clusters in CNBS which seemed to be associated with BLL of grandmother's during gestation. We reported a 14% increase in DNA methylation in cluster of 5 CpG sites near the TSS of Brain Development-Related Molecule 1 or N-Myc Downstream-Regulated Gene (NDRG4) for CNBS with high BLL in MNBS (Fig 2.2.3A). We speculate that this may lead to their down-regulation during development and contribute to cognitive deficits reported in Pb-exposure studies in other human cohorts. In support of our hypothesis studies in mouse models demonstrated that ablation of NDRG4 lead to reduced levels of Brain Derived Neurotrophic Factor (BDNF). This was consequently shown to cause impaired spatial learning and memory (Yamamoto et al., 2011). We also report ~ 24% decrease in the promoter of Nerve Injury Induced Protein 2 (NINJ2) in a small cluster of 2 CpG sites (Fig 2.2.3B). We propose that down-regulation of promoter methylation of NINJ2 might lead to upregulation of its expression. This is consistent with the observation that NINJ2 has been shown to upregulated in Schwann cells surrounding the distal segment of injured sensory neurons and has been shown to promote extensive neurite growth during neurodevelopment (Araki and Milbrandt, 2000) . In a human cohort NINJ2 has been shown to undergo hypermethylation in peripheral blood of patients with Borderline personality disorders (BPD) (Teschler et al., 2013). Therefore, in view of the

evidence from our study, we speculate that NINJ2 might be one of the primary responders to in-utero exposure to Pb and might be involved in neuro-protection. NINJ2 and NDRG4 are the strongest candidates for potential transgenerational biomarkers of Pb exposure for studies in larger cohorts and will be the subject of future studies. We also observed differentially methylated CpG clusters in genes associated with the immune system such as TRPV2 (Zhang et al., 2012), and shared promoter of Docking Protein 3 (DOK3) and DEAD (Asp-Glu-Ala-Asp) Box Polypeptide 41(DDX41) (Peng et al., 2013) (Table 2.2.2). This suggested that in-utero Pb exposure might have widespread impact on the immune system (Table 2.2.2) (Youssef et al., 1996; Dyatlov and Lawrence, 2002).

In conclusion, our pilot study provides indirect evidence that Pb-exposure in women during childbirth can affect the locus-specific DNA methylation status of her grandchildren. However, the altered DNA methylation profiles of the grandchildren's blood are apparently "normalized" during post-natal development. Also, fetal germ-line exposure to Pb apparently has different epigenetic consequences than acute childhood exposure. It remains to be determined whether Pb-exposure dependent epigenetic changes are observed in larger and more diverse cohorts, and whether they affect neurodevelopment or other phenotypes associated with high BLL.

**Table 2.2.1:** Covariate and Blood Lead Level (BLL) information used for DNA methylation analysis for 35 samples. Sample names ending with A (Child's Neonatal blood spots), and with B (Child's current blood spots).

Sample	sample bll	mother's bll	Gender	Age (month)	Mother's age (month)	Gestational age (month)	Smoking
001-027-002-2010-B	0.19009625	6.148736497	F	24	264	240	No
001-001-001-2012-B	0.579901708	2.745941529	M	6	204	186	No
001-122-001-2010-B	0.856610711	7.29056368	M	24	240	216	Yes
001-082-001-2009-B	0.871991444	1.378498216	M	24	216	192	No
001-122-002-2011-B	1.972338713	7.29056368	F	12	240	228	Yes
001-051-001-2011-B	2.27245527	0.28218839	M	12	216	204	No
001-087-002-2010-B	2.938008435	4.252003701	F	24	252	228	Yes
001-060-002-2010-B	3.035195443	36.17606133	F	12	264	252	Yes
001-130-003-2012-B	4.378173778	3.376888045	M	9	264	243	No
001-113-002-2012-B	4.602492159	3.011403372	F	8	228	208	No
001-010-001-2012-B	4.71525216	56.64450079	M	5	240	235	No
001-036-001-2007-B	4.735150984	11.26302255	M	48	264	216	Yes
001-033-001-2012-B	5.098809195	2.244241237	M	3	228	225	No
001-036-005-2009-B	5.809591329	11.26302255	M	36	264	228	Yes
001-084-001-2011-B	5.94181757	1.768640127	M	12	216	204	No
001-015-006-2009-B	6.248038356	6.413477367	F	36	264	228	No
001-025-001-2007-B	6.436692662	5.171338965	M	60	252	192	No
001-045-001-2008-B	6.860479927	5.987767511	M	36	252	216	No
001-095-001-2009-B	7.04519292	5.636077432	M	36	228	192	No
001-121-002-2007-B	7.206834813	6.614628519	F	60	240	180	Yes
001-099-003-2007-B	7.430384158	7.507768472	M	60	228	168	No
001-130-001-2007-B	7.626344312	3.376888045	M	48	264	216	No

001-036-002-2011-B	8.502613523	11.26302255	F	12	264	252	Yes
001-036-003-2008-B	9.288905445	11.26302255	M	48	264	216	Yes
001-015-002-2010-B	9.640403264	6.413477367	F	36	264	228	No
001-015-004-2008-B	9.665637279	6.413477367	F	48	264	216	No
001-019-001-2006-B	9.725285685	6.999675563	M	60	276	216	Yes
001-095-003-2011-B	10.20218454	5.636077432	M	12	228	216	No
001-130-002-2007-B	10.47605773	3.376888045	F	48	264	216	No
001-099-001-2011-B	11.04437582	7.507768472	M	12	228	216	No
001-118-002-2009-B	11.83912714	2.722341717	F	36	252	216	No
001-033-002-2010-B	16.97667656	2.244241237	F	24	228	204	No
001-019-003-2007-B	17.22238377	6.999675563	M	48	276	228	Yes
001-010-002-2009-B	24.20912991	56.64450079	F	24	240	216	No
001-079-002-2010-B	32.80080749	2.671777556	F	24	252	228	No
001-051-001-2011-A	1.241657755	0.28218839	M	NA	216	204	No
001-082-001-2009-A	-0.559041588	1.378498216	M	NA	216	192	No
001-084-001-2011-A	1.627906418	1.768640127	M	NA	216	204	No
001-033-001-2012-A	2.675718869	2.244241237	M	NA	228	225	No
001-033-002-2010-A	4.48833828	2.244241237	F	NA	228	204	No
001-079-002-2010-A	4.078826258	2.671777556	F	NA	252	228	No
001-118-002-2009-A	1.368164285	2.722341717	F	NA	252	216	No
001-001-001-2012-A	0.132033982	2.745941529	M	NA	204	186	No
001-113-002-2012-A	0.526838178	3.011403372	F	NA	228	208	No
001-130-003-2012-A	1.559798609	3.376888045	M	NA	264	243	No
001-130-001-2007-A	1.67457733	3.376888045	M	NA	264	216	No
001-130-002-2007-A	4.367599524	3.376888045	F	NA	264	216	No
001-087-002-2010-A	1.054781846	4.252003701	F	NA	252	228	Yes
001-025-001-2007-A	8.759664027	5.171338965	M	NA	252	192	No

001-095-003-2011-A	-0.002739693	5.636077432	M	NA	228	216	No
001-095-001-2009-A	1.616803451	5.636077432	M	NA	228	192	No
001-045-001-2008-A	-3.885461602	5.987767511	M	NA	252	216	No
001-027-002-2010-A	0.497807044	6.148736497	F	NA	264	240	No
001-015-002-2010-A	0.667523822	6.413477367	F	NA	264	228	No
001-015-006-2009-A	1.383208565	6.413477367	F	NA	264	228	No
001-015-004-2008-A	1.728121508	6.413477367	F	NA	264	216	No
001-121-002-2007-A	-4.3374629	6.614628519	F	NA	240	180	Yes
001-019-003-2007-A	1.497650833	6.999675563	M	NA	276	228	Yes
001-019-001-2006-A	3.819612838	6.999675563	M	NA	276	216	Yes
001-122-001-2010-A	-7.576020476	7.29056368	M	NA	240	216	Yes
001-122-002-2011-A	-0.034270196	7.29056368	F	NA	240	228	Yes
001-099-003-2007-A	2.376179089	7.507768472	M	NA	228	168	No
001-099-001-2011-A	9.40974033	7.507768472	M	NA	228	216	No
001-036-003-2008-A	-0.214657358	11.26302255	M	NA	264	216	Yes
001-036-005-2009-A	0.233450692	11.26302255	M	NA	264	228	Yes
001-036-002-2011-A	1.617716682	11.26302255	F	NA	264	252	Yes
001-036-001-2007-A	2.896384326	11.26302255	M	NA	264	216	Yes
001-060-002-2010-A	2.494466541	36.17606133	F	NA	264	252	Yes
001-010-002-2009-A	0.587111426	56.64450079	F	NA	240	216	No
001-010-001-2012-A	1.04007402	56.64450079	M	NA	240	235	No

**Table 2.2.2:** Table showing a list of 6 genes/CpG clusters which show Pb dependent change in DNA methylation status in CNBS exposed in-utero to high BLL (high BLL MNBS). CpG sites, chromosome and location of CpG island, Location relative to the Promoter of the gene (i.e. CpG sites located at the promoter (Yes) or transcription start site (TSS)), Effect size, the average change in DNA methylation at the CpG island (e.g., 0.05 is a 5% increase in average DNA methylation), CpG sites/cluster (the number of CpG sites with significant changes in the cluster) is illustrated in the following table.

Gene	CpG island	Promoter	Effect size	Standard error	P-value	FDR	CpG sites/cluster
NDRG4	chr16:58535040-58535596	No (TSS)	0.14	0.04	0.000729	0.028	5
NINJ2	non CpG	Yes	-0.24	0.06	0.000168	0.013	2
TRPV2	non-CpG	Yes	-0.11	0.02	1.29E-06	0.0005	2
DOK3	non-CpG	No (TSS)	-0.08	0.02	4.46E-07	0.00030	4
APOA5	chr11:116661034-116661410	No	0.05	0.01	0.00077	0.029	3
Enhancer	chr2:8596907-8597573	NA	0.090329	0.028656	0.00162	0.0427	2

## **CHAPTER 3: LEAD (PB) EXPOSURE INDUCES CHANGES IN 5-HYDROXYMETHYLCYTOSINE CLUSTERS IN CPG ISLANDS IN HUMAN EMBRYONIC STEM CELLS AND UMBILICAL CORD BLOOD; HMEDIP-450K ARRAY. (SEN ET AL., 2015A)**

### **3.1. Background**

In the previous sections we have demonstrated that DNA methylation can be used as important biomarkers for acute and chronic exposure to Pb. Furthermore, we have shown the DNA methylation changes due to Pb exposure can be transmitted from grandmother to grandchildren most likely via the mother's germ cells. Even though DNA methylation is a stable epigenetic modification it can be removed from the genome by either replication dependent passive dilution or active demethylation. Active demethylation require oxidation of 5 methylcytosine (5mC) to 5-hydroxymethylcytosine (5hmC) by a family of dioxygenases called ten-eleven-translocation enzymes (TET). 5hmC has been specifically been shown to be enriched in the mature brain. This suggests that they play a crucial role in brain development (Kriaucionis and Heintz, 2009; Tahiliani et al., 2009). Furthermore, 5hmC has also been shown to be localized in promoter regions of actively transcribed genes suggesting they might have a regulatory role to play in transcription (Stroud et al., 2011; Szulwach et al., 2011). Studies exploring the relationship between 5hmC and heavy metal exposure have been limited. Zhang et al., 2014 reported a marked increase in total 5hmC levels in the heart and spleen of rats exposed to physiologically relevant levels of Arsenic (As) through drinking water (Zhang et al., 2014). Tellez-Plaza et a, 2014 reported significant correlation between Arsenic metabolism (Urinary Arsenic concentration) and %5hmC in the peripheral blood of 48 volunteers (Tellez-Plaza et al., 2014). These studies presented compelling evidence suggesting that 5hmC might be better biomarker for Pb exposure.



Detection of 5hmC usually requires enrichment of 5hmC containing DNA using antibody based enrichments. The isolated DNA fragments are then individually tested using PCR based reaction or Next Generation Sequencing (NGS). While both the methods are effective, PCR based methods are time consuming and NGS is expensive. Therefore, the main objective for this study was to develop a cost-effective method to enrich and quantitatively measure 5hmC in human cohort. For enrichment of high density regions of 5hmC modification, we generated randomly-sheared DNA fragments and immunoprecipitated (IP) them with 5hmC specific antibodies. Due to the limited sensitivity of the antibodies, only fragments with large number of 5hmC sites were expected to be precipitated. Then we used the high-throughput Infinium™ Human Methylation 450K (HM450K) array (Illumina, Inc.) to determine the 5hmC profile of the IP'ed fragments. We named this novel modification of the HM450K array the HMeDIP-450K array.

### 3.2. Methods

**hESC culture and Pb exposure:** The human ESC line WA09 (H9) (Thomson *et al.* 1998) was obtained from the WiCell Research Institute (Madison, WI, USA) and maintained in a humidified incubator at 37°C with 5% CO<sub>2</sub>, as previously described (Senut *et al.*, 2014). Briefly, undifferentiated hESCs (passages 26-39) were cultured in DMEM/F12 supplemented with knockout serum replacement, nonessential amino acids, penicillin/streptomycin, L-Glutamine, 2-mercaptoethanol, and human basic fibroblast growth factor (Life Technologies) on a feeder layer of irradiated mouse embryonic fibroblasts (Globalstem). hESCs were passaged by mechanical dissociation every 4-6 days and their pluripotency frequently tested by immunofluorescence staining for specific markers including Oct4 and Lin28 Stock solutions (100-fold concentrated) of Pb acetate (Pb(C<sub>2</sub>H<sub>3</sub>O<sub>2</sub>)<sub>2</sub>) (Sigma-Aldrich) were

prepared in sterile distilled water. Two physiologically relevant concentrations of Pb acetate chosen on the basis of our previous work (Senut et al., 2014) were tested in this study: 0.8 $\mu$ M (16  $\mu$ g/dL) and 1.5 $\mu$ M (32  $\mu$ g/dL). Distilled water was used as a vehicle control. Undifferentiated hESCs were acutely exposed to the different concentrations of Pb or vehicle for 24 hours, at which time the hESC colonies were dissected and DNA was isolated (Sen et al., 2015a).

**Extracting genomic DNA from hESCs:** DNA was isolated from ~1 million cells with Qiagen EZ1 Advanced® using the DNA Investigator® reagents and protocol. DNA concentrations were quantified by UV spectrophotometry using the DropSense96® Microplate Spectrophotometer (Trinean), and the purity was assessed based on the A260/A280 and A260/A230 ratios (Sen et al., 2015a).

**Samples and sample classification:** The Early Life Exposure in Mexico to Environmental Toxicants (ELEMENT) cohort consists of over 4000 mother-infant pairs, belonging to a low income population recruited from 1994 onwards from three hospitals in Mexico City (Mexican Social Security Institute, Manuel Gea Gonzalez Hospital, and National Institute of Perinatology) (Pilsner et al., 2009; Sen et al., 2015a). We used umbilical cord blood (UCB) from the fourth wave of the ELEMENT study (Each of the four waves had approximately 1000 children participants). The blood lead concentrations of umbilical cord blood were determined using atomic absorption spectroscopy (Model 3000; PerkinElmer, Chelmsford, MA, USA) at the metals laboratory of the American British Cowdray Hospital in Mexico City (Pilsner et al., 2009; Sen et al., 2015a). Out of the initial pool of samples, 412 samples were randomly selected for DNA extraction. DNA extraction was carried out at the Harvard-Partners Center for Genetics and Genomics. Unfortunately, RNA was not extracted

so it is not possible to follow up these findings with gene expression analyses. For our study, we randomly selected UCB from 24 males and 24 female children from the 1<sup>st</sup> and 4<sup>th</sup> quartiles of Pb levels. The UCB for male children had a minimum blood lead level (BLL) of 0.59  $\mu\text{g}/\text{dl}$  and highest BLL of 7.21 $\mu\text{g}/\text{dl}$ . The UCB for female children had a minimum BLL of 0.68 $\mu\text{g}/\text{dl}$  and highest BLL of 10.53  $\mu\text{g}/\text{dl}$ . Five male children and seven female children had BLL in UCB  $\geq 5\mu\text{g}/\text{dl}$  and classified as high BLL group. Details on the 48 sample mother-infant cohort samples can be found in previous studies (Sen et al., 2015a).

**Shearing and denaturation of DNA:** A similar protocol was used for shearing and denaturation of DNA for hESCs and UCB. Approximately 3 $\mu\text{g}$  genomic DNA was diluted in 130 $\mu\text{l}$  of buffer TE (10mM Tris (pH 8.0), and 1mM EDTA (pH 8.0)) and sheared into ~200-600bp fragment using micro cavitation (Covaris, Inc, setting: Duty Cycle = 5%, Intensity = 3, Cycles/burst = 200, Time = 75 seconds run at 6-8°C). 125 $\mu\text{l}$  of the sheared DNA samples were mixed with 330 $\mu\text{l}$  of buffer TE. The sheared DNA was denatured by boiling it in the Thermomixer at 95°C and 700rpm for ten minutes and left on ice for 10min.

**Immunoprecipitation and extraction of DNA:** A similar protocol was used for DNA extracted from hESCs and UCB. 51  $\mu\text{l}$  of 10x immunoprecipitation buffers and 3 $\mu\text{l}$  of 5mC or 5hmC specific antibodies at a concentration of 1 $\mu\text{g}/\mu\text{l}$  was added to the denatured DNA sample and incubated for  $\geq 2$ hours at 4°C with tipping. Then, 25 $\mu\text{l}$  Dynabeads™ protein G magnetic beads was added to the samples and incubated overnight at 4°C with tipping. The residual beads were collected with a magnetic rack and washed 3 times with 700 $\mu\text{l}$  1xIP buffer. The bead was then re-suspended in 20mg/ml Proteinase K solution and incubated for 3hours at 50°C and 800rpm. The residual beads were pelleted using a magnetic rack. The supernatant was collected for phenol/chloroform extraction. For

extraction, the supernatant was treated with 250µl phenol: chloroform: isoamyl alcohol at a ratio of 25:24:1, vortexed and then spun down for 2 minutes at 14,000rpm. Then, the supernatant was treated with 250µl chloroform, vortexed and spun down for 2 minutes. Finally, DNA was precipitated by adding 20µl 5M NaCl, 1µl glycogen and 500µl 100% ethanol and incubated at -20°C for 30 minutes. The samples were then pelleted at 4°C and 14,000rpm for 20 minutes and allowed to dry at room temperature for 10 minutes. The dried out samples are re-suspended in 25-50µl water and readied for the HM450K bead chip array for methylation analysis.

**HM450K bead chip array:** For this study, we coupled Immunoprecipitation with 5hmC antibodies with the Human methylation 450K array. For more details, refer to the 2.1.2. Methods; sub-heading HM450K bead chip array.

**DNA digestion with PvuRst1I and DNA sequencing:** PvuRts1I (Pvu) restriction enzyme can directly cleave hydroxymethylated DNA 12-14 bps away from the 5hmC site (Szwagierczak et al., 2011). We developed a new technique that we call Pvu-seq which allows direct detection of 5hmC without chemical modification of 5hmC or bisulfite conversion. A paper describing this technique was published in *BMC Genomics* (Cingolani et al., 2013). Briefly, the whole genomic DNA extracted from control and Pb-treated hESCs, was digested with PvuRts1I and sequenced using 50 bps paired end sequencing reads in Illumina™ HiSeq 2500.

**Statistical analysis for HMeDIP- 450K data:** For studying effects of exposure to Pb on the 5hmC profile for in-vitro Pb-hESCs and UCB DNA, we used the Adjacent site clustering algorithm (A-clustering) proposed by Sofer et al, 2013(Sofer et al., 2013; Sen et al., 2015a). We used A-clustering to detect sets of correlated CpG sites and then tested the

clusters for multivariate response to environmental exposure to Pb using the generalized estimation equation approach. The aforementioned approach is efficiently implemented using the R-package Aclust (Sofer et al., 2013). For determining the differentially hydroxymethylated regions (DhMRs) we used modified Aclust parameters; Spearman correlation, for calculating the distance between adjacent sites ( $\text{dist}(i,j) = 1 - \text{corr}(i,j)$ ), average clustering type, which require that mean distance between two sites be at least 0.25, 300 bps distance restriction for merging of clusters, which ensures that clusters located far away from each other are not merged together based on correlation. The clustering approach is implemented with a 299 bps merge initiation step, which clusters all sites wedged between 2 high correlated sites within 299 bps of each other together, to reduce the complexity of data and the analysis time for the A-clustering step. The minimum fragment length pulled down by 5hmC antibody was 300bps, therefore to increase the probability of detecting high density 5hmC clusters only the distance restriction for clustering was restricted to 300 bps. Finally, the data was analyzed using a generalized estimation equation approach and filtered for significant DhMRs using FDR corrected p-value cutoff = 0.05, exposure effect size  $\geq |0.02|$  and number of CpG sites per cluster  $\geq 5$ . These regions are can be defined as Pb-dependent high density 5hmC clusters. To determine the genomic locations of the probes belonging to individual DhMCs, they were annotated using the publicly available Illumina Human Methylation 450k annotation data in R (>2.15). The target genes mapping to DhMRs were individually visualized using UCSC genomic browser(Sen et al., 2015a). The Delta beta or the beta difference between the median of the beta values for each probe for low BLL samples and high BLL samples were mapped by the chromosomal location of the probes.

### **Statistical analysis of the effect of sex in HMeDIP-450K array for UCB DNA:**

The single nucleotide differences 5hmC and association with the sex of the infant was determined using a mixed effect model which was implemented using the package CpGassoc in R(R>2.15)(Barfield et al., 2012). All analysis was conducted controlling for covariates such as BLL, socioeconomic status, gestational age and smoking status of the mothers and birth weight of the infant. After determination of single nucleotide differences the sample was either separated into males and females and analyzed separately using the A-clustering approach described previously to determine the male-specific and female-specific 5hmC changes in co-regulated hMRs. Alternatively, the sex of the infant was used as a covariate which enabled us to determine the Pb-dependent changes in hMRs conserved between male and female infant. All analysis was conducted while controlling for socioeconomic status, gestational age and smoking status of the mothers and birth weight of the infant.

**Cell type estimation:** Cell type estimation was done using the estimateCellCount function in Minfi (R>2.15)(Aryee et al., 2014). The raw dataset directly read from the .idat files were used to estimate cell count based on the HM450K data generated for flow sorted blood cells from 6 adult males available in Bioconductor.

**Statistical analysis of HM450K for UCB DNA:** All CpG sites mapping to putative 5hmC clusters are removed from the beta matrix. The single nucleotide differences 5hmC and association with the sex of the infant was determined using a mixed effect model which was implemented using the package CpGassoc in R(R>2.15) [26]. All analyses were conducted controlling for covariates such as BLL, socioeconomic status, gestational age and smoking status of the mothers and birth weight of the infant. For determining the differentially methylated regions (DMRs) we used the modified Aclust parameters;

Spearman correlation, for calculating the distance between adjacent sites ( $\text{dist}(i,j) = 1 - \text{corr}(i,j)$ ), average clustering type, which require that mean distance between two sites be at least 0.25, 300 bps distance restriction for merging of clusters. The clustering approach is implemented with a 299 bps merge initiation step. Finally, the data was analyzed using a generalized estimation equation approach and filtered for significant DhMCs using FDR corrected p-value cutoff = 0.05 and exposure effect size  $\geq |0.02|$ . To determine the genomic locations of the probes belonging to individual DMCs, they were annotated using the publicly available Illumina™ Human Methylation 450k annotation data in R (>2.15).

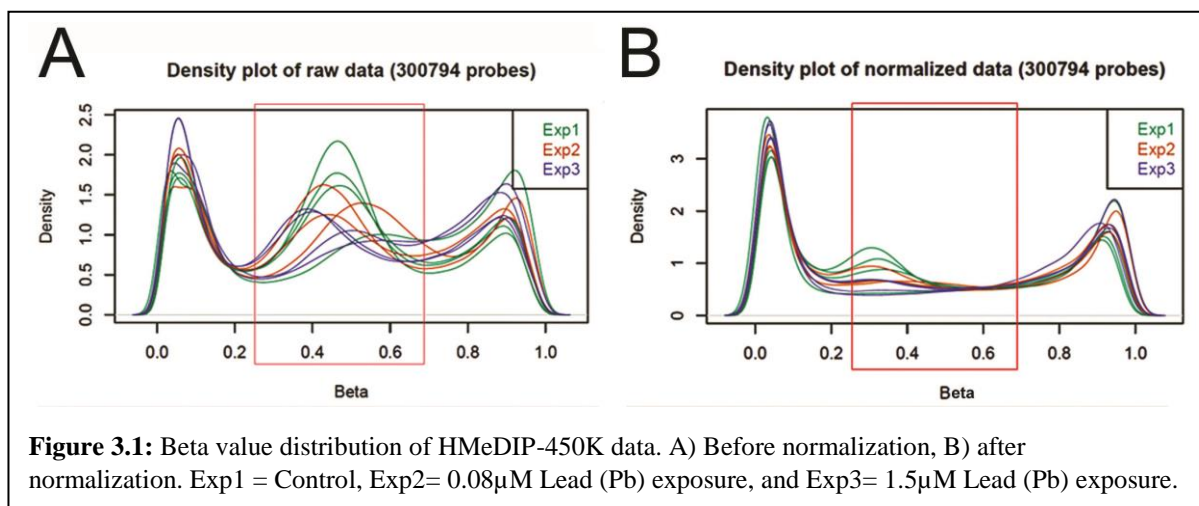
**Statistical analysis for Pvu-Seq data:** Pvu-Seq was used to confirm the presence of 5hmC in the predicted Pb-dependent high density 5hmC clusters (Cingolani et al., 2013). Using the approach implemented in the *chipseq* package in R/Bioconductor, peaks significantly above the noise distribution was estimated and visualized in using Integrative genome viewer (IGV)(Sen et al., 2015a). Briefly, the peak calling for pvu-seq is based on the presumption that the number of times a genomic position is sequenced (coverage where  $\text{coverage} = \text{Read Length (nt)} * \text{Total Reads Number} * / * \text{Genome Length (nt)}$ ) follows a *Poisson distribution*. Therefore, from this information we can estimate a coverage cut-off for which the FDR corrected p-value  $\leq 0.05$ . For our study the estimated coverage cutoff was 4.5. PvuRts1I (Pvu) restriction enzyme can directly cleave hydroxymethylated DNA 12-14 bps away from the 5hmC site(Szwagierczak et al., 2011) sites. Therefore, we extended our peak region to  $\pm 20$  on either side. Then using GenomicRanges package in R we determined HMeDIP-450K 5hmC regions were overlapped with the predicted Pvu-Seq peak to confirm the presence of 5hmC. Validation was only done for the Pb non-treated hESCs. For visualization of the data, we converted the sorted Pvu-Seq, .bam files into .bed file using

*bedtools* and calculated the median count in 25bps windows along the genome and plotted them as histograms using Integrative genome viewer (IGV). The location of the CpG sites/probes mapping to high density 5hmC regions (hMRs) where obtained from the Illumina Human Methylation 450k annotation data in R and plotted as a separate track in IGV. The overlapped regions were queried on IGV to generate region-specific 5hmC overlap. HMeDIP-Seq data was downloaded from GEO (GSM1008199) and overlapped HMeDIP-450K using Genomic Ranges package in R.

**Gene ontology analysis:** Gene ontology analysis was done using GOstats in R/Bioconductor. Briefly, for a gene ontology (GO) class a hypergeometric probability is calculated, denoting whether the number of gene belong to the GO term is larger than expected (Falcon and Gentleman, 2007). We use this method to look at the suggestive association of our gene list with gene – ontological categories. As the gene list is short, this assists in manual curation of the dataset rather than providing the exact biological targets.

### 3.3. Results

#### Normalization and analysis of HMeDIP-450K array on the stem-cell model of environmental exposure.





The  $\beta$ -values from HM450K array have a bimodal distribution with the right peak representing un-methylated probes and left peak representing methylated probes (Fig 3.1A). The HMeDIP-450K array has a prototypical distribution characterized by a central peak (Fig 3.1A). We speculate that this central peak is most likely due to accumulation of HM450K probes which did not hybridize to the IP'ed DNA. Normalization and background correction using well established HM450K normalization pipelines (see methods) effectively gets rid of these peaks (Fig 3.1B).

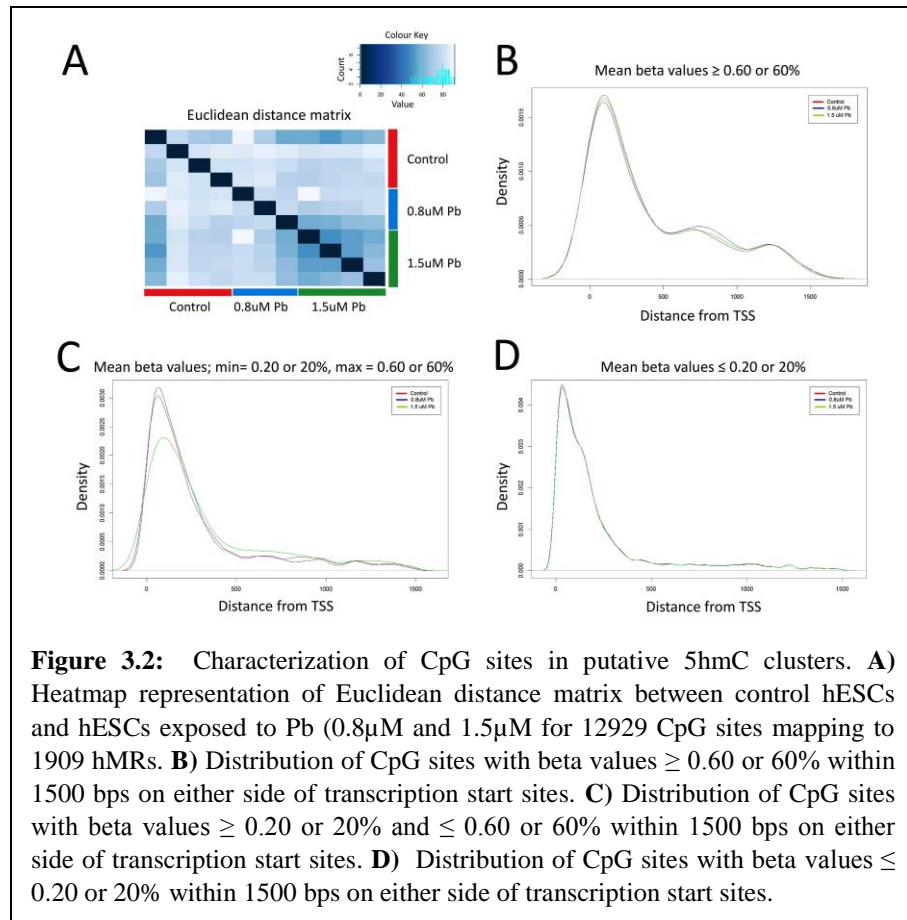
5hmC antibodies are most likely to bind to regions of high 5hmC density. Based on this assumption we pre-clustered our CpG sites into co-regulated regions (see methods) and then tested the impact of exposure on these co-regulated clusters. One caveat of this analysis is that some of these clusters will contain 5mC sites which lie in close proximity to high density 5hmC clusters. We tested our modified array on a stem-cell model of environmental exposure (see methods)(Senut et al., 2014). At a correlation distance (d) cut-off of 0.25, maximum cluster size  $\leq 300$  bps and minimum number of mapped CpG probes  $\geq 5$  we detected 1909 putative high density CpG clusters consisting of 12929 unique CpG sites. We defined these clusters/regions as co-regulated high density 5hmC regions (hMRs).

5hmC has been shown to localize around the transcription start sites of active transcribing genes (Stroud et al., 2011; Szulwach et al., 2011). We found that 41% of the CpG sites belonging to hMRs mapped near TSS ( $\pm 1500$ bps). Irrespective of  $\beta$  values, the density distribution of these hydroxymethylated CpG sites gradually decreased when moving away from the TSS (Fig. 3.2B-D).

Next, we investigated the impact of Pb-exposure on DNA hydroxymethylation of the pre-defined hMRs using GEE (see methods). We found that exposure to  $0.8 \mu\text{M}$  Pb caused a  $\geq +2\%$  increase in 5hmCs in 36 DhMRs and a  $\leq -2\%$  decrease in 70 DhMRs at a FDR

corrected significance cut-off of 0.05 (Fig. 3.3A). Similarly, exposure to  $1.5 \mu\text{M}$  Pb concentrations caused a  $\geq 2\%$  increase in 5hmC in 38 DhMRs and a  $\leq 2\%$  decrease in 162 DhMRs at a FDR corrected significance cut-off

of 0.05 (Fig. 3.3B). The DhMRs for hESCs exposed to  $0.8 \mu\text{M}$  Pb were associated with genes implicated in processing of mRNA such as RNA phosphodiester bond hydrolysis,

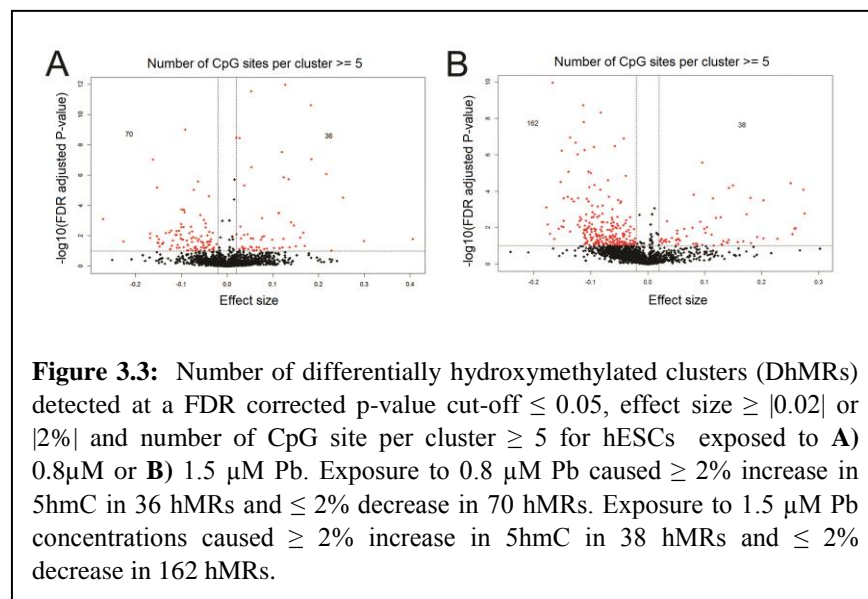


endonucleolytic (GO: 0090502, p-value= 0.000396921) and pyrimidine deoxyribonucleotide metabolic process (GO: 0009219, p-value=0.002472255) (data not shown). For hESCs exposed to 1.5 $\mu$ M Pb, gene ontology categories were associated with cellular proliferation associated processes such as cell division (GO: 0051301, p-value= 1.16E-05), mitosis (GO:0007067, p-value= 0.00013) (data not shown).

### Validation of HMeDIP-450K array using PVU-Sequencing.

We wanted to validate the DhMRs detected by using a bisulfite independent

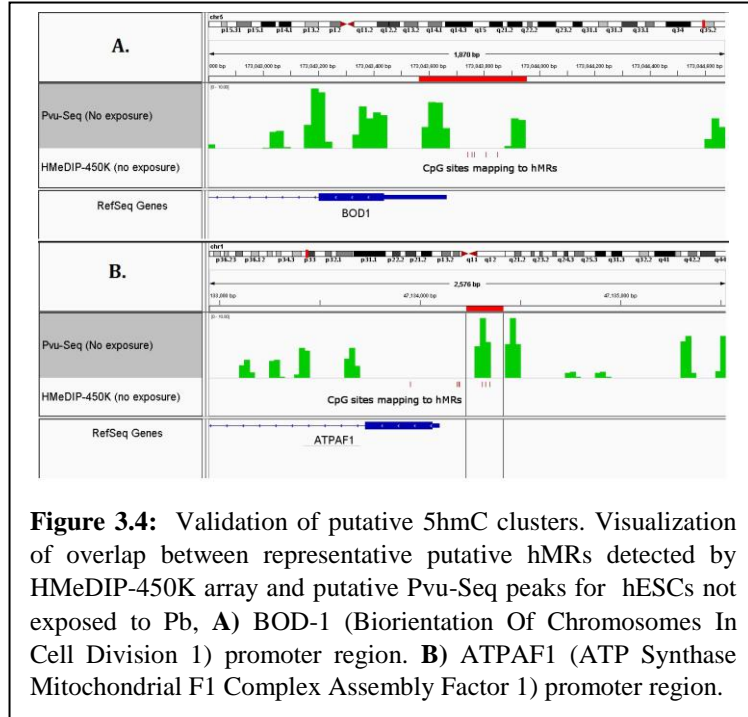
technique called PVU-Seq (see methods). The PvuRts1i enzyme cuts 12-15 nucleotides away from the 5hmC sites (Szwagierczak et al., 2011). Therefore, regions which show



significantly higher coverage compared to the surrounding noise levels (peaks) are indicative of the presence of 5hmC within  $\pm 12-15$  nucleotides from the genomic location of the peaks (Cingolani et al., 2013). We found that Pvu-Seq can only validate the presence or absence of 5hmC site (s) and cannot be used quantitatively at the sequencing depth that we used in this paper (Sen et al., 2015a). We mapped the significant PVU-Seq peaks for DNA isolated from control un-treated H9 stem cells and filtered them by FDR corrected p-value  $\leq 0.05$ . Finally, we overlapped the significant peaks with the 12929 probe positions for 1909 hMRs. We observed that 5661/12929 (44%) CpG probes contained significant PVU-Seq

peaks (see methods for peak calling method) located within  $\pm 20$ bps around the CpG probe position. Examples of an overlapped region are indicated (Fig 3.4A and B).

We further validated our findings from PVU-Seq using meta-analysis of HMeDIP-Seq data from H9-hESCs (GEO accession, GSM1008199) (Gao et al., 2013). This study was focused on determining the 5hmC distribution on pluripotent hESCs. Fortuitously the cell line used, the culture protocol and the IP protocol for this study were almost identical to

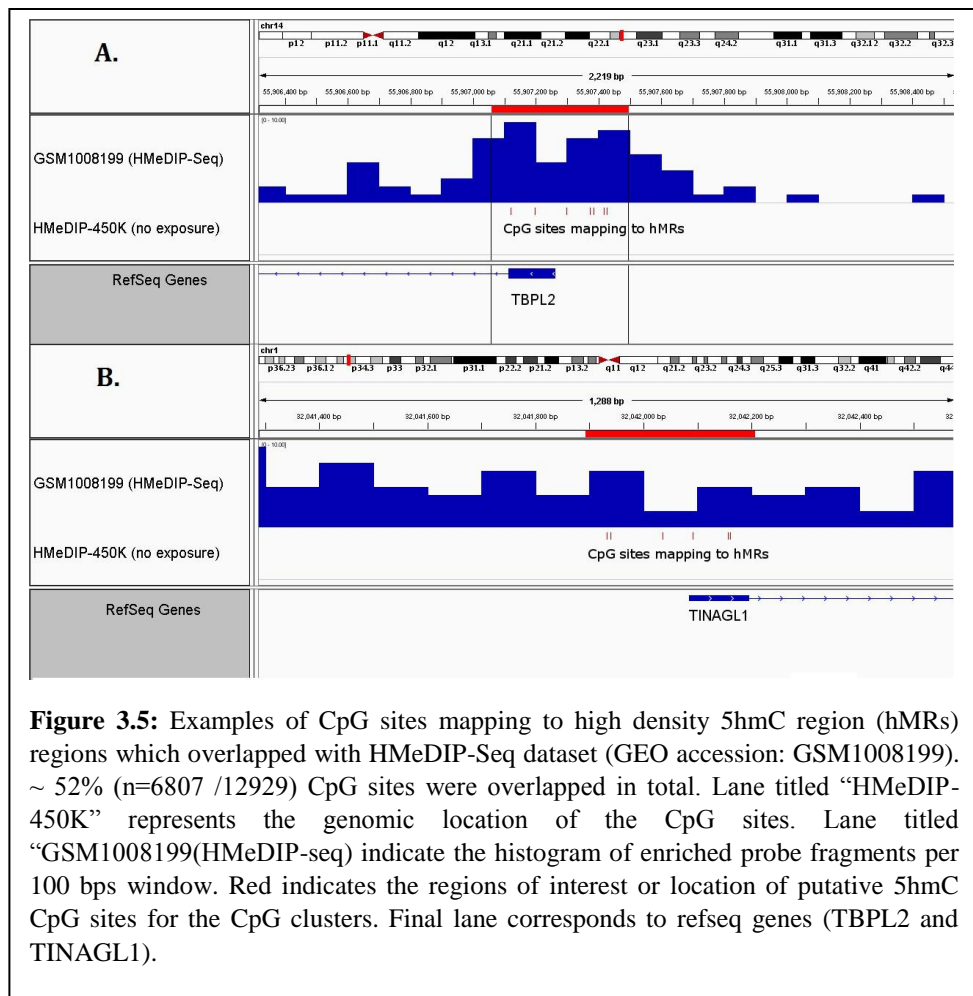


our non-treated hESCs. 6807/12929 (52%) CpG probes mapping to our target hMRs contained at-least 1 significant HMeDIP-Seq peak located  $\pm 20$ bps around the probe sites. Furthermore, we report a positive correlation between ( $r^2=0.36$  or 36%) overlapping HMeDIP-450K sites and number of enriched fragments from the HMeDIP-Seq dataset (Fig 3.6A). Examples of overlapping region are illustrated in Fig 3.5A and B. Finally, 3160/12929 CpG sites were detected in both HMeDIP-Seq data from Gao et al, 2013 and our PVU-Seq data (Fig 3.6B). Our validation experiments clearly indicate that HMeDIP-HM450K is able to interrogate well represented 5hmC regions in the genome.

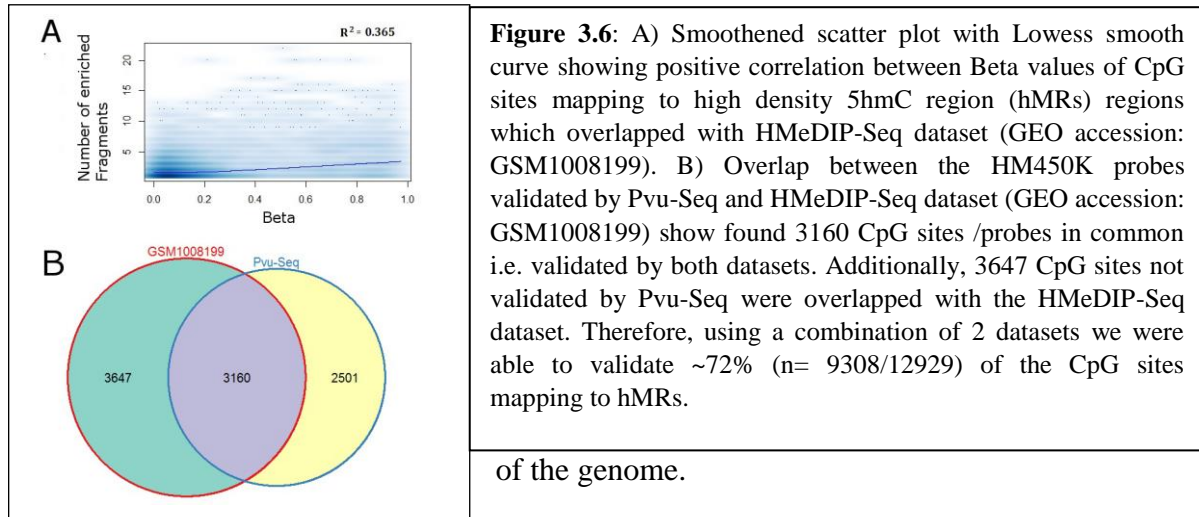
## Pb-exposure associated changes in co-responding high density 5hmC clusters in umbilical cord blood samples.

In the previous section, we demonstrated that our modified version of the HM450K array is capable of detection and quantification of densely populated 5hmC regions. Therefore, we used the HMeDIP-450K array to look for Pb-dependent changes in 5hmC in UCB DNA isolated from the ELEMENT cohort (see methods). In our earlier studies we have reported Pb-exposure causes sex-specific changes in DNA methylation in peripheral blood of children aged 3months to 5 years (Sen et al., 2015b). Therefore, we hypothesized that the 5hmC profile of the genome might also be sex-specific and contribute to sex-specific susceptibility

to adult diseases caused due to prenatal Pb exposure. We modelled the association between Pb-exposure and 5hmC levels at single CpG-sites using a mixed-effect

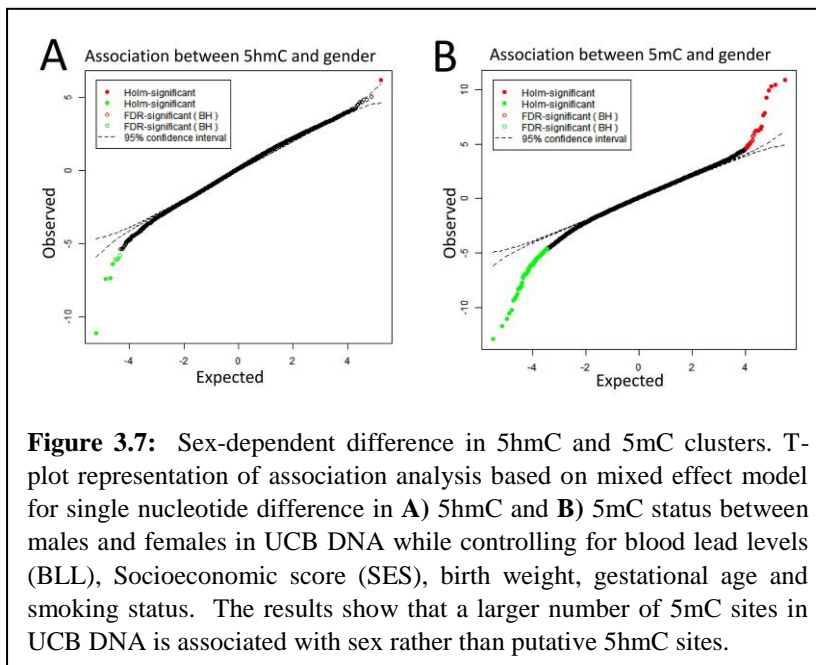


model. Our analysis revealed only 9 CpG probes which were significantly different between males and females (Fig 3.7A), suggesting that sex has a minimal effect on the 5hmC profile



of the genome.

Nevertheless, we performed differential hydroxymethylation analysis using sex as a covariate in our model (see methods). These regions we define as sex-independent/conserved regions. At a FDR corrected p-value  $\leq 0.05$ , we found 4 DhMRs which showed at-least  $\geq +2\%$  Pb-associated increase in hydroxymethylation and 12 DhMRs which showed  $\leq -2\%$  Pb-associated decrease in 5hmC in conserved DhMRs. Some important examples of conserved



regions are shown in Table 3.2. For example, we observed  $13 \pm 3\%$  Pb-exposure dependent decrease in a DhMR of 5 probes mapping to CpG island (chr1:110230238-110230614) near the TSS

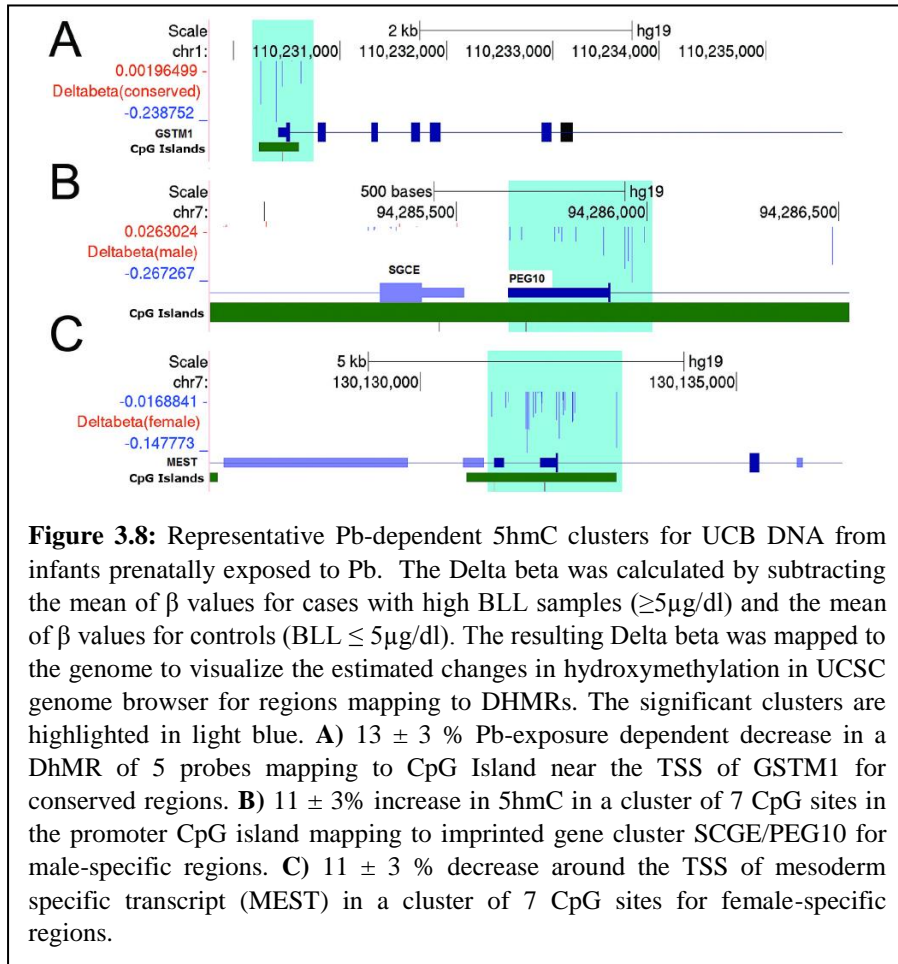
of the Glutathione-S-transferase subunit mu 1 gene *GSTM1* (Fig. 3.8A).

Though the impact of Pb-exposure of 5hmC was limited, we believed that it was still worthwhile to investigate the male and female samples separately. At an FDR corrected p-

value  $\leq 0.05$  and  $\Delta\beta$  or  $\beta$  difference  $\geq \pm 2\%$  we found 8 DhMRs which showed an increase and 14 DhMRs which showed a decrease in

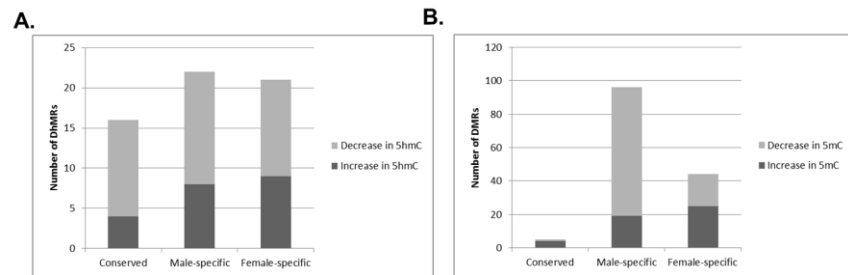
hydroxymethylation in UCB DNA for Pb-exposed male children (Fig. 3.9A). We defined these as male specific DhMRs. For

example, male specific DhMRs mapped to genes such as shared promoter CpG island (chr7: 94284858-94286527) paternally expressed imprinted gene *PEG10* and Sarcoglycan (*SGCA*) ( $11 \pm 3\%$  decrease) (Fig. 3.8B). Similarly, we detected 9 DhMRs which showed an increase and 12 DhMRs which showed a decrease only in female; female specific DhMRs. For examples, Female specific DhMRs mapped to genes such as CpG Island (chr7-130130739-130133111) of paternally imprinted mesoderm specific transcript (*MEST*) ( $11 \pm 3\%$  decrease) (Fig. 3.8C).



## Lead exposure associated changes in co-regulated high density 5mC clusters in umbilical cord blood samples.

In the earlier section using the HMeDIP-450K array we demonstrated that Pb-exposure has limited sex-specific impact on DNA hydroxymethylation profile of UCB DNA. Therefore, we wanted to explore if limited sex-specificity was also shared characteristic of the DNA methylation profile



**Figure 3.9:** Number of differentially hydroxymethylated (DhMRs) and methylated (DMRs) detected for UCB DNA samples in conserved, male-specific and female-specific regions. A) At a FDR corrected p-value  $\leq 0.05$ , we found 4 DhMRs which showed  $\geq 2\%$  Pb-associated increase in hydroxymethylation and 12 DhMRs which showed  $\leq 2\%$  Pb-associated decrease in hydroxymethylation in conserved DhMRs, 8 DhMRs which showed an Pb-dependent increase in hydroxymethylation and 14 DhMRs showed Pb-dependent decrease in hydroxymethylation in UCB DNA for male infants, 9 DhMRs which showed a Pb-dependent increase in 5hmC and 12 DhMRs which showed a Pb-dependent decrease in 5hmC for female infants. B) For conserved regions at an exposure effect size cut-off  $\geq |0.02|$  or  $|2\%$  and FDR corrected p-value  $\leq 0.05$ , we found 4 Pb-associated DMRs which were hypermethylated and 1 DMR which was hypomethylated. Out of 44 DMRs consisting of 122 CpG sites in females, 25 DMRs were hypermethylated, while 19 DMRs were hypomethylated in UCB of female infants. For males, using the same criteria for filtering we found 96 DMRs mapping to 390 probes or CpG sites. 19 DMRs were hypermethylated and 77 DMRs were hypomethylated in UCB for male infants.

of the UCB DNA. We used the HM450K array to determine the association between 5mC profile of UCB DNA and sex of the children. We filtered the HM450K dataset and excluded all probes mapping to hMRs from the Hm450K probe list (see methods). Then we used a mixed effect model (see methods) to determine the CpG sites which were differentially methylated in UCB DNA depending on the sex of the infants. Our analysis revealed that 300 CpG sites were significantly different between males and females (Fig. 3.7B), suggesting that, contrary to our findings for 5hmC, sex has a large effect on the 5mC profile of UCB DNA. To further understand this association at a regional level we modelled CpG sites as co-



regulated 5mC clusters using sex as a covariate. For conserved regions at an exposure effect size cut-off  $\geq \pm 2\%$  and FDR corrected p-value  $\leq 0.05$ , we found 4 Pb-associated differentially methylated regions (DMRs). This result further demonstrated that 5mC profile of the genome is highly sex-specific.

Therefore, we separated out the male and female samples and re-ran our regional differential methylation analysis. At an FDR corrected p-value  $\leq 0.05$  and  $\Delta\beta$  or  $\beta$  difference  $\geq \pm 2\%$  we identified 44 DMRs consisting of 122 CpG sites in females. Out of these 44 DMRs, 25 were hypermethylated, and 19 DMRs were hypomethylated in females (Fig 3.9B). For males, using the same criteria for filtering, we found 96 DMRs mapping to 390 CpG sites. Of these 96 DMRs, 19 were hypermethylated and 77 DMRs were hypomethylated in males (Fig. 3.9B). Our finding suggests that the 5mC profile for UCB DNA from male infants is more sensitive to Pb exposure compared to females. Furthermore, male-specific DMRs were enriched for functional categories such as telencephalon development (GO: 0021537, p-value= 4.24783E-05) and glial cell differentiation (GO: 0010001, p-value= 0.000151812) (data not shown). These genes were not affected by Pb-exposure in females.

### 3.4. Discussion

In this study we present an efficient and inexpensive method to study the 5hmC profile of the DNA called as the HMeDIP-450K array. This method is only restricted to regions of high 5hmC density which can be efficiently IP'ed using commercially available 5hmC antibody. Using this method, we have demonstrated that 5hmC profile of UCB DNA shows limited sex specific changes in response to prenatal Pb exposure. In contrast, 5mC was highly sex-dependent. Furthermore, 5mC in UCB DNA from male infants was much more susceptible prenatal exposure to Pb compared to the females. Sex-specific susceptibility to

exposure has been reported by several studies in animal models and human cohorts. For example, low BLL has been shown to have much more profound impact on cognitive performance of boys compared to girls (Koller et al., 2004). This is consistent with the observation that childhood Pb exposure is correlated with much reduced gray matter volume in males compared to females (Cecil et al., 2008). Some of these sex-biased susceptibility has been attributed to down-regulation of expression of genes required for learning and behavior like Arc, and transcription factors such as Stat1 (Schneider et al., 2012). Epigenetically, studies by Pilsner and colleagues have demonstrated that prenatal exposure to Pb is associated with global changes in 5mC dependent on sex of the individual. Furthermore, over 580 autosomal CpG sites have been shown to have sex-specific differences in DNA from buccal swabs from healthy adult subject. Therefore, our study provides further additional evidence 5mC profile of the genome is indeed largely dependent on sex and may underlie increased susceptibility of males to environmental insults.

Interestingly, though we observed limited impact of prenatal Pb-exposure on DNA methylation profile of UCB DNA, we observed significant 5hmC decrease near transcription start sites (TSS) of GSTM1 and GSTM5. GST or Glutathione-S-Transferases are a group of phase 2 metabolic enzymes responsible for conjugation of reduced form of glutathione to xenobiotic substrates. Genetically, GSTM1 class of enzymes has been shown to highly polymorphic and has been implicated in development of bladder cancer (Engel et al., 2002). Interestingly, polymorphisms in the GSTM1 gene have also been associated with lead-induced inflammatory response in human cohorts (Sirivarasai et al., 2015). This suggests that GSTM1 is an important regulator of stress response. Therefore, it's perhaps not surprising to observe 5hmC changes near the TSS of GSTM1 and M5. We speculate that Pb-dependent

decrease in 5hmC might be responsible for activating transcription (Wu et al., 2011a) of GST genes and may be protective in nature.

Additionally, we did observe a limited number of 5hmC clusters which showed sex-specific response Pb-exposure. Interestingly these clusters were restricted to genes associated with paternal and maternal imprinting. Imprinted genes are expressed in a parent-of-origin specific fashion. DNA methylation of imprinted genes has long been hypothesized as biomarker of exposure. In this study we report, Male-specific DhMRs mapped to promoter region of a paternally expressed imprinted gene locus consisting of 2 genes; paternally expressed 10 (PEG10) and sarcolycaan epsilon gene (SGCE). Yamaguchi et al, 2013 demonstrated that imprinted SGCE/PEG10 locus was completely methylated in TET1 paternal KO embryos which lead to early embryonic lethality in mouse model (Yamaguchi et al., 2013). We predict that a decrease in 5hmC levels in this region will overtime lead to the accumulation of methylation marks and consequently increase susceptibility to developmental disorder. Similarly, in females, we saw a decrease in 5hmC in the CpG islands located in the TSS of the Mesoderm specific transcript (MEST), also known as PEG1. Significant decrease in DNA methylation of PEG1 locus was observed in mothers diagnosed with diabetes mellitus compared to healthy controls (El Hajj et al., 2013). This result suggests aberrant changes in methylation in PEG1 locus might be associated with metabolic disorders.

To determine the ability of the HMeDIP-450K array to detect 5hmC we tried to validate some of the differential methylated sites detected in un-treated H9 Human embryonic stem cells (hESCs) using 2 independent methods. In the first method, we used the PvuRts1I (Pvu) restriction enzyme to digest the DNA from H9 hESCs and sequenced the

products. This method is significantly different from the HM450K array as it does not rely on sodium bisulfite treatment of DNA. Sodium bisulfite treatment is a critical step in preparation of samples for HM450K array as it is required for the cytosine(C) to thymine(T) (C>T) conversion of un-methylated C. This facilitates the differentiation of methylated-C from the unmethylated-C which allows detection of 5mC and 5hmC. PVU-Seq is at best qualitative in nature and can be utilized to detect the presence or absence of 5hmC.

We performed secondary validation using a HMeDIP-Seq meta-dataset generated from H9 hESCs. The method of preparation of DNA prior to detection using the HMeDIP-450K array and HMeDIP-Seq is exactly the same. However, HMeDIP-450K array interrogates pre-selected CpG sites while the HMeDIP-Seq is unbiased and whole genome. Furthermore, HMeDIP-Seq is limited by the sequencing depth i.e. increasing the sequencing depth will lead to detection of larger number of 5hmC sites. Therefore, for this set we expected higher number of HMeDIP-450K probes to be located near HMeDIP-Seq peak regions.

Given the difference in method of preparation of DNA, detection capabilities and limitation of the corresponding techniques, we expected to see low (~25%) to moderate (~60%) overlap between the HMeDIP-450K CpG probes and the PVU-Seq and HMeDIP-Seq peaks. Correspondingly we observed only 44% of the HMeDIP-450K was located near PVU-peaks and 52% located near HMeDIP-Seq peaks. Therefore, in this study we decided to annotate all the HMeDIP-450K probes validated by either the PVU-Seq or the HMeDIP-Seq dataset as our high confidence 5hmC probe sets (N= 9308/12929).

In conclusion, in this study we present a novel modification of the HM450K array which is capable of detecting subtle changes in densely populated 5hmC clusters. We show

that 5hmC modification are less impacted by sex of the individual compared to 5mC and might be better suited as early biomarkers of Pb exposure especially in Umbilical cord blood.

Table

**Table 3.1:** Differentially hydroxymethylated (5hmC) regions (DhMRs) for representative genes in hESCs exposed to either 0.8 $\mu$ M Pb or 1.5 $\mu$ M Pb. These clusters can be used as potential early 5hmC biomarkers of Pb exposure.

Pb exposure ( $\mu$ M)	Gene	CpG Island	Promoter associated	Effect size	Standard error	P-value	FDR	Number of sites per cluster
0.8	ATP5G1	chr17:46969695-46970125	yes	-0.078	0.023	0.00063	0.01504	7
0.8	ATPAF1	chr1:47133674-47134395	yes	0.028	0.004	1.50E-11	3.59E-09	7
0.8	ETHE1	chr19:44031127-44031504	yes	-0.063	0.011	2.21E-08	2.63E-06	6
0.8	HSD17B4	chr5:118788125-118788428	yes	-0.074	0.021	0.00045	0.01186	6
0.8	AGO2	chr8:141646075-141646514	yes	-0.073	0.014	9.40E-08	9.27E-06	5
1.5	CCND1	chr11:69451136-69458596	yes	-0.034	0.010	0.00091	0.01333	11
1.5	NUF2	chr1:163291479-163292021	yes	-0.148	0.043	0.00053	0.00925	5
1.5	RAD51	chr15:40986871-40987772	yes	-0.074	0.025	0.00306	0.03123	6
1.5	GOLPH3	chr5:32173598-32174837	yes	-0.066	0.022	0.00335	0.03324	7
1.5	SOD1	chr21:33031734-33032657	yes	-0.064	0.021	0.00290	0.03022	6

**Table 3.2:** Differentially hydroxymethylated (5hmC) regions (DhMRs) for representative genes for male-specific, female-specific and conserved regions in umbilical cord blood DNA with blood lead levels (BLL)  $\geq 5\mu$ g/dl. These clusters can be used as potential early 5hmC biomarkers of Pb exposure.

Group	Gene	CpG island	around TSS	Imprinted gene	Effect size	Standard error	P-value	FDR	Number of sites per cluster
Conserved	GSTM1	chr1:11023023-8-110230614	yes	no	-0.139	0.035	6.1E-05	0.000777	5
Conserved	GSTM5	Non-CpG	yes	no	-0.212	0.033	9.82E-11	3.8E-08	8
Conserved	H19	chr11:2019565-2019863	yes	maternally expressed	-0.106	0.025	1.59E-05	0.000307	5
Conserved	DNAH6	chr2:84743463-84743685	yes	no	0.044	0.008	4.57E-08	4.82E-06	6
Male-specific	GOLPH3	chr5:32173598-32174837	yes (promoter)	no	0.038	0.013	0.002885	0.018132	7
Male-specific	SGCA/PEG10	chr7:94284858-94286527	yes (promoter)	paternally expressed	-0.113	0.032	0.000364	0.005345	9
Female-specific	MEST	chr7:13013073-9-130133111	yes	paternally expressed	-0.112	0.031	0.000291	0.002521	7
Female-specific	DDAH2	chr6:31695894-31698245	yes (promoter)	no	0.034	0.011	0.001535	0.008076	5

**Table 3.3:** Differentially methylated (5mC) regions (DMRs) for representative genes for male-specific, female-specific and conserved regions umbilical cord blood DNA with blood lead levels (BLL)  $\geq 5\mu\text{g/dl}$ . These clusters can be used as potential early 5mC biomarkers of Pb exposure.

Group	Gene	CpG Island	around TSS	Effect size	Standard error	P-value	FDR	Number of sites per cluster
Conserved	PIK3R1	chr5:67584213-67584451	yes (promoter)	0.030	0.006	1.67E-06	0.00912	2

Male-specific	GLI2	chr2:121624827-121625209	no	-0.040	0.010	4.57E-05	0.00536	2
Male-specific	FGF20	chr8:16859044-16859452	yes	-0.036	0.008	1.99E-06	0.000774	3
Male-specific	SLITRK5	chr13:88329394-88329885	yes	0.042	0.012	0.000772	0.040357	3
Male-specific	TOP1MT	chr8:144416712-144417054	yes(promoter)	-0.024	0.007	0.000831	0.042691	2
Female-specific	MBP	chr18:74843360-74845426	yes(promoter)	-0.059	0.015	0.000121	0.032948	2
Female-specific	SLC2A1	chr1:43423467-43424768	yes	0.036	0.010	0.000237	0.039979	3
Female-specific	GJB3	chr1:35246876-35247599	yes	0.043	0.010	5.85E-06	0.007541	2

## **CHAPTER 4: TRAUMATIC BRAIN INJURY CAUSES RETENTION OF LONG INTRONS VIA REGULATION OF LOCAL HISTONE 3 LYSINE 36 METHYLATION PROFILE IN THE SUB-ACUTE PHASE OF INJURY.**

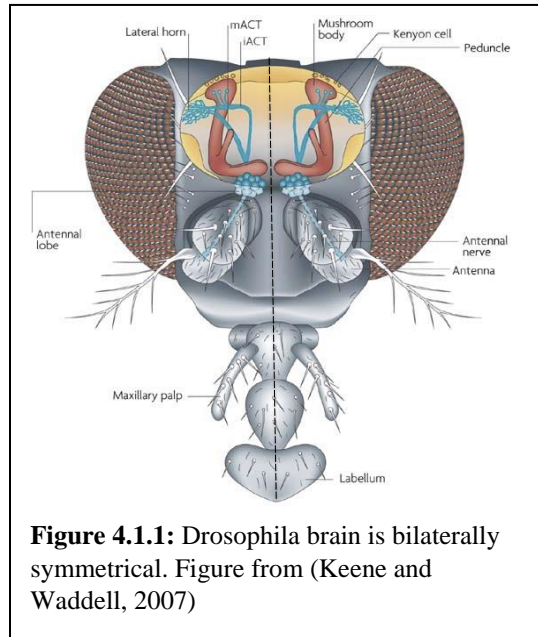
### **4.1. Background**

#### **4.1.1. Drosophila model of traumatic brain injury**

Traumatic brain injury (TBI) is one of the leading causes of death in the world. TBI is typically characterized by primary damage due to physical trauma leading to neuronal cell death followed by secondary long-term damage mediated by a cascade of complex events. The primary damage encompasses physical damage such as contusion, damage to blood vessels (haemorrhage) and axonal shearing (Langlois et al., 2006; Maas et al., 2008). Profile for secondary damage can vary from tissue atrophy (Bramlett and Dietrich, 2007), glutamate exotoxicity, inflammation, mitochondrial damage and ROS generation leading to DNA damage among others (Povlishock and Christman, 1995). Till date, mouse is the most common organism used to study TBI. Mouse models of TBI are mainly divided in 3 types; fluid percussion injury (FPI) (Dixon et al., 1987), controlled cortical impact (CCI) injury (Lighthall, 1988; Dixon et al., 1991), weight drop–impact acceleration injury (Marmarou et al., 1994), and blast injury (Cernak et al., 1996). Briefly the FPI utilizes a pendulum striking the piston of a reservoir of fluid to generate a fluid pressure pulse to the intact dura through a craniotomy. This model can closely simulate a Closed Head Trauma (CHI), i.e. trauma without skull fracture. It has been generally observed to cause focal lesions and secondary neuronal damage in thalamus and hippocampus. The second model, CCI, utilizes a rigid surface (e.g. Piston) to cause injury to exposed, intact dura. CCI generally causes more diffuse injury, causing Blood Brain Barrier dysfunction, deformation of the cortex and long term progressive hypoxia in the brain. The weight drop-impact acceleration model, uses a



guided free falling weight to impact an exposed (Feeney's weight drop model) or protected (Marmarou's weight drop model) dura. This usually causes cortical contusion progressing from a white matter hemorrhage and necrosis within 24 hours of the injury. If using protection (metal helmet on the skull midline between lambda and bregma) bilateral damage of vasculature alongside the midline in most commonly noted. The CCI usually produce neuronal damage most commonly observed car accidents. Finally, the last model i.e. Blast injury model is meant to simulate the effect of blast

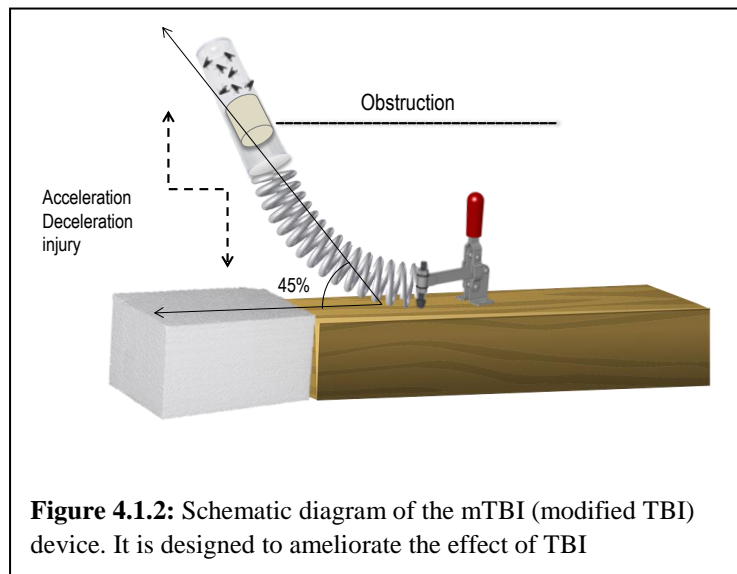


waves on the CNS. This type of injury is commonly seen in military personnel subjected to explosive detonations. They commonly cause axonal injury in peripheral nervous system and diffuse systemic injury such as myelinated axonopathy, microvasculopathy, chronic neuroinflammation and neurodegeneration. Despite the wide range of TBI models available in mice, these experiments can be expensive. Moreover phenotypic and molecular outcomes can be high variable due to complex structure and wiring of the mammalian brain.

Interestingly, the macroscopic structure of the Drosophila brain is very similar in many respects to the mammalian brain (Fig 4.1.1). For example, Drosophila brain is bilaterally symmetrical and divided into 3 regions; protocerebrum, deutocerebrum and tritocerebrum homologous to forebrain, midbrain and hindbrain in human. Moreover, it covered by a cuticle, separated from the brain by a fluid haemolymph making the structure very similar to human cranium. In lieu of the structural similarities between human and

Drosophila brain, the Wassarman lab developed an inexpensive and reproducible model to simulate TBI in *Drosophila melanogaster* (Fruit Fly) using an in-house device known as the high impact trauma (HIT) device (Katzenberger et al., 2013). Using the HIT device, they demonstrated that it is possible to reproduce defining characteristics typical to closed head trauma (CHI) in humans such as temporary incapacitation, ataxia, and neurodegeneration in *Drosophila* (Fig 4.1.2). For example, the *Drosophila* showed a reduction in lifespan, dependent upon the number of strikes, similar to reports from epidemiological studies in humans. The authors also demonstrated an increase in neurodegeneration (i.e. formation

of vacuolar structures) dependent upon the age and the number of primary TBI incidences, similar to that observed in mouse models. The HIT device can simulate a TBI caused by rapid acceleration-deceleration coupled by physical



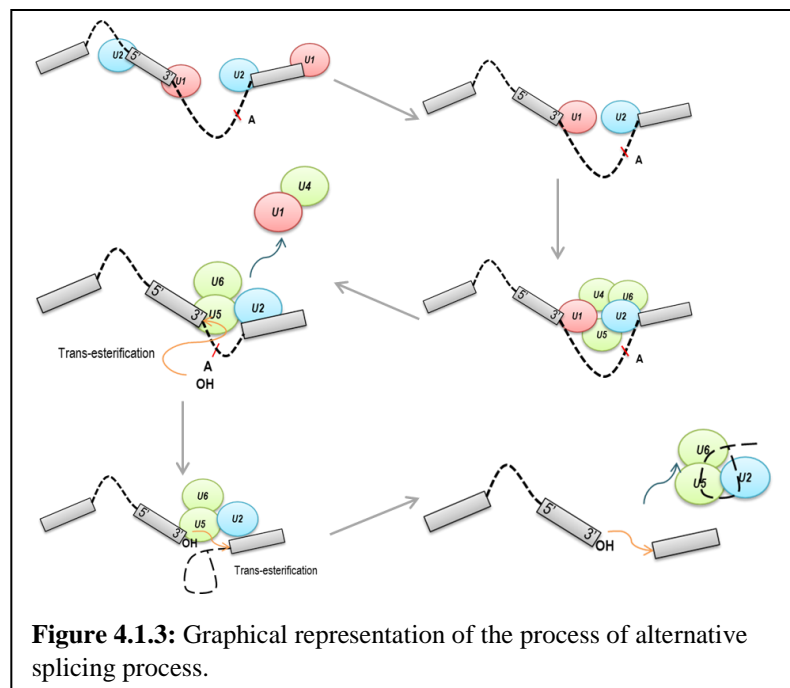
trauma and might be reminiscent of a car crash. We modified the HIT device to ameliorate the effect of the TBI, by erecting an obstructive canopy at an approximate angle of  $45^\circ$  (Fig 4.1.2). This helped us mitigate the variability in draw between biological replicates. Using this model, we investigated the effect of TBI on the transcriptome at 2 time points; early period/acute phase (4 hours), and late period/sub-acute phase (24 hours) in 0 to 5 days old flies. We hypothesized that the diversity of the transcriptome and expression profile will be different between the early and late period. Secondly, we speculated that the changes in the

transcriptome between early and late period will be correlated with local epigenetic features such as chromatin accessibility (e.g. heterochromatin vs. euchromatin).

#### 4.1.2. Regulation of alternative splicing by epigenetic modifications in neuronal damage.

Alternative splicing is tightly regulated process which allows for a single gene to encode for a large number of proteins. Briefly, this process begins by demarcation of the exon boundaries by U1 and U2 small nuclear RNAs and protein complex. These small nuclear RNA recognize the 5' Splice Sites and 3' Splice Sites by directly base pairing with unprocessed mRNA and form an exon definition complex. The exon definition complex

undergoes structural re-arrangement into an intron spanning complex. This is followed by the recruitment of the tri-snRNP U4-U6-U5 complex which allows the transition of the intron spanning complex into a catalytically active state. The main purpose of this reaction



is to facilitate the spatial proximity between the 5'SS, branch site and 3'SS. This is followed by a transesterification reaction which results joining of the 5'SS with the branch point (usually Adenine) and formation of a *Lariat intermediate*. This is followed by a second transesterification reaction that joins together the adjacent exons and release the *Lariat* as a by-product of the splicing process (Fig 4.1.3).

The process of splicing is further regulated by interaction of the splicing machinery with other accessory proteins which bind to exonic and intronic sequences. For example, Serine/Arginine rich splicing factors (SR) proteins have been shown to bind to exonic sequences and allow differential processing of exons. One of the best studied examples is the sex-dependent exon inclusion/exclusion of *doublesex* (dsx) gene in *Drosophila* (Steinmann-Zwicky, 1994). Exon 4 of dsx contains a 13nt repeat sequence (or dsxRE) which function as a splicing enhancer (Chandler et al., 2003). This enhancer sequence is recognized by SR protein RBP1 and splicing regulator TRA2. However, binding of RBP1 and TRA2 to dsxRE is dependent on presence of TRA. TRA is only expressed in female therefore exon 4 is included in females and excluded in males. Functions of SR proteins are antagonized by Heterogeneous nuclear ribonucleoproteins (hnRNP) (Han et al., 2010). As the name implies, hnRNP group of nuclear proteins can participate in regulation of diverse alternative splicing events. For example, hnRNP-A1 has been hypothesized to antagonize the function of SR protein Splicing Factor 2 (SF2) (Mayeda et al., 1993), hnRNP-L has been shown to bind to intronic sequences and prevent inclusion of cassette exon (Hung et al., 2008), hnRNP-H regulates switching between long and short isoform of A-raf protein (Rauch et al., 2010). Besides regulation of alternative splicing hnRNPs proteins are also responsible for nuclear-cytoplasmic transport of mRNAs, mRNA stability and transcriptional regulation. hnRNP-A2 has been shown to provide the cytoplasmic localization signal of myelin basic protein and control its localization in oligodendrocytes (Muller et al., 2013), hnRNP-K has been shown to regulate expression of c-myc by associating directly with the elongating polymerase (Michelotti et al., 1996) . Due to complex nature of the interaction between splicing factors, much of the alternative splicing regulatory process is yet to be well-understood.

Originally, it was believed that alternative splicing occurs post-transcriptionally. However early evidence from studies conducted by Kim et al,1997 and Hirose et al, 1999 demonstrated the direct interaction between splicing factors and the nascent RNA template (Kim et al., 1997; Hirose et al., 1999), suggesting the role of Transcriptional machinery is regulation of alternate splicing events. These observations were further corroborated by Rosbash lab, who demonstrated that *Drosophila* S2 cells with a Slow Polymerase mutant (RPII<sub>215C4</sub>) mutant processed intron with a greater efficiency compared to WT flies (Khodor et al., 2011). Therefore, in view of the current evidence, we can conclude that splicing and transcription are not mutually independent processes.

Interactions between the splicing machinery and the elongating polymerase also suggest that histone modifications might a regulatory role of alternative splicing. In particular, H3K36me3 has been hypothesized to be associated with alternative splicing in several studies. Yoh et al, 2008 demonstrated that SETD2 is recruited to the C-terminal domain of elongating Polymerase by SPT6 and is associated with its locus specific methyl transferase activity (Yoh et al., 2008). Studies by the Ahringer lab demonstrated that alternative exons are marked by lower levels of H3k36me3 modification compared to flanking constitutive exons suggesting a probable association between this histone mark and Alternative Splicing (Kolasinska-Zwierz et al., 2009). H3k36me3 has also been shown to directly interact with the splicing machinery. A study by Luco et al, 2010 reported that H3K36me3 can indirectly recruit Poly-pyrimidine binding protein (PTB) through MORF-related gene 15 (MRG15) protein and cause exclusion of exons (Luco et al., 2010; Pradeepa et al., 2012; Zhou et al., 2014). Interestingly, H3K36me3 can also indirectly interact with SR proteins such as SRSF1 through accessory proteins PSIP1 and cause exon inclusion in genes,

such as exon 5 of DIAP2 and exon7 of VCAN (Luco et al., 2010; Pradeepa et al., 2012; Zhou et al., 2014). However, studies looking at genome-wide association between H3K36me3 and alternative splicing are limited.

#### 4.2. Methods

**Sample for survival estimation and exploratory RNA sequencing:** Sample consisted of 0 to 5days old  $w^{1118}$  flies. These flies were collected in sturdy plastic vials and subjected to Traumatic brain injury using the High Impact trauma (HIT) device. For the survival assay, the male and female flies were separated and 25 flies per vial were used for each condition (control, 1 strike, 2 strike). The flies were subjected to TBI, at a spring deflection of approx.  $45^\circ$  to ameliorate the impact of trauma. The 2<sup>nd</sup> strike was performed after a recovery time of 5mins. These flies were maintained till ~98% of the flies in the control vial were deceased. Survival estimations and standard error calculation was done using R/Bioconductor. For RNA sequencing analysis heads from 50 males and 50 female flies for control, 1 strike and 2 strikes were pulled off manually one at a time and immediately transferred to RNA-later (500 $\mu$ l). Fly head collection was done at 2 time-points 4 hours and 24 hours. The male and female were used as biological replicates. More explanation of study design for RNA-Sequencing studies is provided in later sections.

**Bulk preparation of fly heads:** 0-5days old  $w^{1118}$  flies were transferred to 50 ml tubes on ice. Approximately 3 ml of flies was collected per tube. These flies were snap frozen by dipping them in liquid nitrogen and vortexed vigorously to detach the head (for 3mins; 15 secs per turn). The vial was dipped in liquid nitrogen between consecutive turns to prevent them from thawing. The vortexed flies were passed through a sieve 720 $\mu$ m pore size. This allowed the separation of the bodies from the detached heads. The detached heads were

collected on another sieve 410µm pore size. This sieve allowed the separation of heads from other detached body parts. The fly heads were transferred to 2ml Eppendorf tubes and stored in -80°C this further use.

**Sample for Sub-cellular fractionation:** 0-5days old  $w^{1118}$  flies were collected in bulk and separated into 2 batches. Batch 1 was used as control and prepared for subcellular fractionation. The fly heads were collected in RNAlater (500µl) to prevent RNA degradation. Batch 2 was transferred to plastic vial and subjected to 1 strike and collected after 24 hours in RNA-later (500µl). Samples were stored in -20°C prior to fractionation. Approximately, 500µg of heads were collected per condition in 2ml Eppendorf tubes.

**Sample for ChIP-sequencing:** 0-5days old  $w^{1118}$  flies were collected in bulk and separated into 2 batches. Batch 1 was used as control and prepared for tissue homogenization and DNA isolation. The fly heads were collected in 1XPBS (phosphate buffer saline) containing Protease halt cocktail (10µl per ml of PBS). Batch 2 was transferred to plastic vial and subjected to 1 strike and collected after 24 hours in 1XPBS (+halt cocktail). Approximately, 500µg of heads were collected per condition in 2ml Eppendorf tubes.

**Sample for RNA-sequencing of 3<sup>rd</sup> instar larvae and SM-mutant flies:** Most of the gene knockdowns were performed by crossing da-Gal4 flies, which have ubiquitous expression, to UAS-RNAi flies using the following stocks from the Bloomington, Indiana stock center:  $y[1] \text{ sc}[*] v[1]; Py[+t7.7] v[+t1.8]=\text{TRiP.HMS01304 attP2}$  (expresses dsRNA for RNAi of Kdm4A (FBgn0033233) under UAS control);  $y[1] v[1]; Py[+t7.7] v[+t1.8]=\text{TRiP.HMS02273 attP40}$  (expresses dsRNA for RNAi of Idh (FBgn0001248) under UAS control); P{GAL4-da.G32}, which expresses Gal4 in all cells;  $y[1] v[1]; Py[+t7.7] v[+t1.8]=\text{TRiP.JF01320 attP2}$  (Expresses dsRNA for RNAi of Kdm2 (FBgn0037659) under

UAS control); y[1] sc[\*] v[1]; Py[+t7.7] v[+t1.8]=TRiP.HMS00488 attP2 (Expresses dsRNA for RNAi of CG7200 (FBgn0032671) under UAS control); y[1] sc[\*] v[1]; Py[+t7.7] v[+t1.8]=TRiP.HMS00775 attP2 (Expresses dsRNA for RNAi of CG8165 (FBgn0037703) under UAS control); and y[1] sc[\*] v[1]; Py[+t7.7] v[+t1.8]=TRiP.HMS00575 attP2 (Expresses dsRNA for RNAi of Utx (FBgn0260749) under UAS control). The smooth-RNAi lines were not viable when crossed to da-Gal4, so instead we used a hypomorphic allele over a deficiency in a hemizygous combination; cn[1] P{PZ} sm[05338]/CyO; ry[506] (aka. sm [4]) and w[1118]; Df(2R)BSC820, P+PBac w[+mC]=XP3.WH3 BSC820/SM6a..

**Tissue homogenization:** For tissue homogenization 4 stainless steel 3.2mm beads (Qiagen cat no. 69990) was added to 2ml Eppendorf tubes. The samples were subjected to homogenization using TissuelyserLT (Qiagen) using the following setting; 50 oscillations (1/s) for 2min. The homogenized samples were centrifuged using a low-speed bench top centrifuge and supernatant containing cells were collected for downstream processing.

**Subcellular fractionation:** The RNAlater has a specific gravity greater than cells therefore it is difficult to recover the complete homogenate for prepared samples. To optimize the recovery, the RNAlater was diluted using 1XPBS (1:2). Subcellular fractionation was carried out using RNA Subcellular Isolation Kit (Active motif; 25501). The nuclear fraction was further treated with DNase to degrade contaminating DNA using DNase treatment kit (Active motif, 25503). The isolated RNA was measured using NanoDrop™ 1000 Spectrophotometer.

**RNA isolation:** RNA was isolated from heads or 3<sup>rd</sup> instar larvae using the Qiagen EZ1 RNA tissue minikit (Cat No./ID: 959034) and the isolated RNA was measured using NanoDrop™ 1000 Spectrophotometer.



**Chromatin immunoprecipitation:** The homogenized tissue collected was fixed by incubating with 16% Formaldehyde (final concentration 1%) for 10 min at RT. Then 10XGlycine (final concentration 1X) was added and samples were incubated at RT for 5 mins, to quench the activity of Formaldehyde. After incubation the samples were centrifuged at 3000g for 5min and supernatant was removed to obtain cell pellets. The formaldehyde fixed cell pellets were washed 2 times with PBS (1X) + HC (3000g for 5min) and then reconstituted in Membrane extraction buffer + HC (Pierce Chromatin Prep module, 26158). Cell pellets were broken up by pipetting with a p200 and vortexing for 15 secs and incubated on ice for 15 min. The lysed cells were recovered by centrifugation (9000g for 3min at 4°C) and reconstituted in 200 or 300µl of Nuclease free water (depending upon the size of the pellet, decided arbitrarily). Sonication was carried out using Covaris S2 Series 2 Focused Ultra-sonicator at intensity=5, duty cycle: 10%, 200cycle/burst, for 5-8 mins (depending upon size of the pellet and presence of debris). The sonicated nuclei were recovered using centrifugation (10000g for 5min) and reconstituted in Nuclear extraction buffer + HC (Pierce Chromatin Prep module, 26158). The nuclei were incubated on ice for 15 min which 15sec of vortexing every 5 min. The nuclei were centrifuged at 9000g for 5 min and supernatant containing cleaved chromatin was collect for downstream analyses. The DNA concentrations were measured using Qbit. Chromatin bound DNA was collected from mass preps (N=4) of control and TBI heads (24 hours post-TBI) and pooled and redistributed for immunoprecipitation using Anti-Histone H3 (tri methyl K36) antibody - ChIP Grade (ab9050). We used total of 0.5µg/reaction of DNA for control and TBI heads. The efficiency of the sonication and quality of IP was assessed using 2200 TapeStation System. The IP'ed

DNA was sequenced using 100bps paired-end reads in technical replicates. Quality control metric from sequencing runs were estimated using HOMER (<http://homer.salk.edu/homer>).

**RNA sequencing and ChIP sequencing:** Libraries were prepared using the Illumina™ TruSeq Stranded Total RNA. All sequencing was done using 100 bps paired-ends on Hi-Seq2500.

### **Statistical analysis**

**RNA sequencing dataset used for meta-analysis:** For the first study i.e. RNA sequencing of spinal cord tissue from mouse subjected to spinal contusion 2days or 7days post-injury, the data was directly downloaded from Gene Expression Omnibus (GEO accession number: GSE45376). For the second study i.e. RNA sequencing of hippocampal and cortical tissue from mice subjected to MTBI, the raw .fastq files were obtained from the authors.

For the first mouse dataset the authors performed contusive spinal cord injury in 24 female C57BI/6J mice and collected the RNA from spinal tissue 2days (Acute) and 7days (Sub-acute) post-SCI. The RNA was sequenced using 100bps paired end reads. This model was very similar to our experimental design. The second dataset consisted of sham and TBI cortical and hippocampal samples collected from Male C57BL/6J mice in biological triplicates. The RNA isolated from these samples was sequenced using 50 bps paired-end sequencing.

**RNA-Sequencing alignments:** Prior to alignment the sequencing adaptors were trimmed using FASTQ Toolkit v1.0 in Illumina Basespace. FastQC was used to generate quality metrics for assessment of .fastq files. RNA sequencing alignment was done using TopHat v2.1.0 or HISAT 0.1.6-beta using *default* alignment parameters. After alignment the

mapped reads was further filtered by mappability score ( $\text{MAPQ} \geq 10$ ). The quality controlled .bam files were sorted by genomic position of the reads using *samtools-0.1.19*. PCR duplicate reads were processed and removed using *rmdup* function (options -S) in *samtools*. The sorted duplicate removed .bam files were further assessed and visualized using Integrative Genome Viewer (<https://www.broadinstitute.org/igv/v1.4>). The Mouse and Drosophila dataset were both processed using this pipeline. For Drosophila Melanogaster UCSC genomic build *dm3* was used for read alignments. For Mus Musculus UCSC genomic build *mm10* was used for read alignment.

**ChIP-Sequencing alignments:** All ChIP sequencing alignments were done using *Bowtie2* using the *dm3* build of the Drosophila genome as reference. After alignment the mapped reads was further filtered by mappability score ( $\text{MAPQ} \geq 10$ ). The quality controlled .bam files were sorted by genomic position of the reads using *samtools*. PCR duplicate reads were processed and removed using *rmdup* function (options -S) in *samtools*. The sorted duplicate removed .bam files were further assessed and visualized using Integrative Genome Viewer (<https://www.broadinstitute.org/igv/v1.4>). The tag density and Quality metrics such as GC content and genomic nucleotide frequencies relative (-50, 50) to the 5' end of ChIP-fragment were calculated using HOMER (<http://homer.salk.edu/homer>). (Suppl. figure 8)

**Gene expression analysis:** Reads from the processed .bam files were overlapped with ENSEMBL exon annotation extracted from UCSC in R/Bioconductor (Packages; *GenomicAlignments*, *GenomicFeatures*). The *dm3* genomic-build was used for Drosophila datasets and *mm10* genomic-build was used for the Mouse dataset. The read count per exon was computed using the preset “Union” mode. (Refer to *summarizeoverlaps*). The exon counts were next combined to give the gene read counts. Differential gene expression

analysis from computed read counts was carried out using DESEQ2 in R/Bioconductor. The read count and DESEQ2 parameters used were identical between the Drosophila and mouse dataset.

For Drosophila model of TBI RNA were prepared in bulk for 50 males and 50 females for 3 separate conditions; control, 1 strike and 2 strikes at 2 time points; 4 hours and 24 hours. This meant that for the exploratory analysis we had a total of 12 samples. The negative binomial distribution (see DESEQ2 vignette) (Love et al., 2014) was fitted using treatment conditions; control, 1 strike and 2 strikes as classifiers and sex of the flies and time of collection as categorical covariates (read distribution  $\sim$  conditions + sex + time of collection). This enabled us to control for any potential sources of gene expression variations which may arise due to the sex of the flies. Additionally, this allowed us to use the male and the female  $w^{1118}$  flies were used as biological replicates. Male and female  $w^{1118}$  non-TBI flies collected at 4 hours and 24 hours were used as controls for all differential expression estimations. The differentially expressed genes were further filtered using FDR and logFC cut-off which are further discussed in the results section.

**Exon usage analysis:** Number of aligned reads was counted within disjoint exonic bins rather than exons using the “Union” mode of read-counting in summarizeoverlaps. Then DEXSEQ was used to determine relative exon-usage while controlling for overall gene expression. Relative exon usage can be defined as follows;

Exon usage= # transcripts from the gene that contain target exon/# all transcripts from the gene

The read counting and DEXSEQ parameters used were identical between the Drosophila and mouse dataset.

For *Drosophila* TBI the negative binomial distribution for modelling read counts (see DEXSEQ vignette) was fitted using treatment conditions; control, 1 strike and 2 strikes as classifiers and sex of the flies and time of collection as categorical covariates. For each target gene DEXSEQ models read distribution as a function of  $\sim$ conditions + exon + sex:exon+ time:exon + conditions:exon, where “:” indicate existence of possible interaction/correlations between covariates. Then this complete model is compared against a *null* model;  $\sim$ conditions + exon + sex:exon+ time:exon, to determine the effect of treatment conditions of exon usage. For the mouse SCI dataset DEXSEQ analysis was performed using  $\sim$  conditions + exon + conditions:exon as the complete model and  $\sim$ conditions + exons as the *partial/null* model. Further details on DEXSEQ can be found in the package vignette and the paper (Anders et al., 2012).

**Mixture of Isoforms (MISO) analysis:** Exon usage exclusively relies on reads mapping to exons. However, much more accurate estimation of alternative splicing changes can be inferred from intronic and junction reads. For this purpose, we used Mixture of isoform (MISO) to estimate the Percent Spliced in (PSI) value for each annotated splicing feature. Annotations for splicing features were provided by modENCODE consortium and classified into 5 representative classes; Alternative 3’SS, Alternative 5’SS, Skipped exons, mutually exclusive exons, and Retained introns. The difference in PSI value ( $\Delta$ PSI) between control and treatment (TBI) samples were estimated using Bayesian factor analysis. The comparisons were further filtered using a  $\Delta$ PSI cut-off of 0.05 or 5%, Bayesian factor  $\geq 10$  and number of exclusion and inclusion reads  $\geq 10$  (see *MISO documentation*). The significant events common between males and females were selected and correlated to give the final list of sex-independent splicing changes. As majority of the significant events were retained

introns downstream annotation and visualization was done only with retained intron genes. The retained introns were overlapped with maximum overhang length of 10 bps with introns annotation obtained from the ENSEMBL genes UCSC dm3 build of the genome and annotated with respective gene and transcript annotation. Visualization of MISO results was done using 2 independent approaches. In the first approach, the  $\log_{10}(\text{RPKM})$  (reads per kilobase per million) for respective splicing features were plotted using sashimi plot (<https://miso.readthedocs.io/en/fastmiso/>). In the second approach, the normalized coverage was for exonic and intronic regions was calculated using HOMER (<http://homer.salk.edu/homer>).

**Characterization of introns:** Introns and their respective lengths in base-pairs were obtained for ENSEMBL genes from UCSC dm3 build of the Drosophila genome. The lengths were log transformed ( $\log_2$ ) and their density distribution was determined. Then the density distribution was modelled as a mixture of N=2 normal distribution using *Gaussian mixture model*. This allowed us to determine the natural cut-off for long introns. The distributions inferred from the model were plotted using R/Bioconductor. The GC content was determined directly from the fasta sequence corresponding to the dm3 build of the Drosophila genome and plotted using *ggplot2* in R/Bioconductor. Maximum entropy scores (MaxEntScores), was used to discriminate between weak and strong splice sites flanking the retained introns. Briefly, short sequence motif around 5'SS and 3'SS for retained introns was collected depending upon the strand for the gene. MaxEntScores was calculated using the *score5* and *score3* functions in R/Bioconductor (Package *spliceSites*) and was confirmed using the online version ([http://genes.mit.edu/burgelab/maxent/Xmaxentscan\\_scoreseq.html](http://genes.mit.edu/burgelab/maxent/Xmaxentscan_scoreseq.html)). Intron characterization for retained introns in mouse dataset was carried out using exactly the

same pipeline. We used mouse ENSEMBL genome UCSC build mm10 for the mouse datasets.

**Calculating Transcripts per million (TPM):** For determining relationship between transcript abundance and RI we calculated the Transcript Per Million (TPM) for all transcripts containing retained introns for all samples.

$$\text{TPM} = (\text{Rg} * 10^6) / (\text{Tn} * \text{Lg})$$

Where Rg is total read counts for the transcript, Tn is the sum of all length normalized total transcript read counts, and length is the length of the transcript in bps. Following TPM calculations, we tested for significant differences in TPM between representative sample groups using *Fisher Exact Test* (Package; *edgeR*) and filtered the transcripts by FDR corrected p-value of 0.05 and  $\Delta\text{TPM} \geq \pm 20$ . The  $\Delta\text{TPM}$  were plotted against  $-\log_{10}(\text{FDR})$  in R/Bioconductor. The same method for TPM calculation was used for *Drosophila* and mouse datasets.

**Calculating Splicing rate (SR):** Co-transcriptional splicing rate or Splicing rate (SR) was calculated following the same principle proposed by Wickramasinghe et al, 2015 (Wickramasinghe et al., 2015). The splicing rate is defined as the normalized read counts of the first intron versus the last intron of each transcript. Calculations in equation form are as follows;

$$f_i = C_i / L_i \text{ where } C_i \text{ is the read count for last introns and } L_i \text{ is the length of last intron}$$

$$f_k = C_k / L_k \text{ where } C_k \text{ is the read count for first introns and } L_k \text{ is the length of first intron}$$

$$f_{i-1} = C_{i-1} / L_{i-1} \text{ where } C_{i-1} \text{ is the read count for last but one exon and } L_{i-1} \text{ is the length of last but one exon}$$

$$f_{k+1} = C_{k+1} / L_{k+1} \text{ where } C_{k+1} \text{ is the read count for } 2^{\text{nd}} \text{ exon and } L_{k+1} \text{ is the length of } 2^{\text{nd}} \text{ exon}$$

$F_i = f_i / (f_{i-1} \times N)$  where  $N =$  library size of the sample

$F_k = f_k / (f_{k+1} \times N)$  where  $N =$  library size of the sample

$SR = F_k / F_i$  where  $SR =$  splicing rate

Transcripts selected for SR calculation was selected based on following criteria; belongs to a single transcript gene, number of exon for the transcript/gene  $\geq 4$ .

**Meta-analysis of ChIP sequencing datasets:** Exon-trio analysis (ETA) for ChIP sequencing data was adapted from Kolasinska-Zwierz, et al, 2009(Kolasinska-Zwierz et al., 2009). Briefly, we used ENSEMBL gene model annotation from BioMart to create random sets of Constitutive and Alternative exon trios. The constitutive trio (CE trios) consisted of 3 adjacent CE. The alternative trio (AE trios) consisted of 1 AE flanked by 2 CE in a constitutive-alternative-constitutive configuration. We selected an average of unique 1-3 trios per candidate gene. Then we counted the number of reads mapping to exon belonging to pre-defined trios using *MEDIPS* package from R/Bioconductor. We compared each ChIP set to their inputs within 200bps sliding windows and selected exons which showed at least +2fold enrichment of ChIP peaks over input at an FDR corrected p-value  $\leq 0.05$ . For the list of enriched exons, we filtered them such that all 3 members of the trio are present and significantly enriched. We labelled these member as 1= exon1, 2=exon2, 3=exon3 based on their genomic organization and position. Finally, we compared the tag counts (read counts) for each member of the constitutive or alternative trio to each other using ANOVA followed by Tukey-HSD test in R/Bioconductor.

The relative enrichment of H3K36me3 peaks in exons and introns where done using the *MEDIPS* package from R/Bioconductor. Briefly, the differential tag densities within 200 bp regions were estimated using MEDIPS. Regions which showed a  $\geq + 2$ -fold enrichment at



an FDR corrected p-value  $\leq 0.05$  over input, was considered as significant peaks. The target peaks were overlapped with regions of interest (ROIs). Criteria for selecting ROIs are as follows; introns and exons containing significant peaks were  $\geq 200$  bps in length, all introns and exons of the transcripts had significant peaks in them (N=3752). Reads Per Kilobase per Million (RPKM) for respective ROIs were plotted with *ggplo2* in R/Bioconductor.

**Analysis of ChIP-sequencing data:** We used 0.5 $\mu$ g DNA per ChIP reaction. This is a relatively small amount compared to recommended 50 to 100 $\mu$ g. Therefore, we expected relatively less enrichment of H3k36me3 peaks with large background signal. We tried multiple peak calling strategies including the more commonly used peak caller such as MACS1.4.2. We found that combination of MEDIPS in R/Bioconductor and HOMER was more suited for low enrichment high background data. We sequenced our sample using 100bps paired-end reads for more accurate estimation of fragment length and used technical replicates to control for sequencing biases.

We estimated 2 important quality metrics GC% bias and Genomic Nucleotide Frequency relative to read positions using HOMER. These are illustrated in Supplemental Figure 9. These quality metrics were well within expected values. Following the quality control for our H3K36me3 ChIP-seq datasets we determined significant peaks within 200bps regions using MEDIPS. We implemented 2 additional controls in peak calls; firstly, replaced all reads which map to exactly the same start and end positions by only one representative read. Secondly, we specified that the ChIP-seq data used for peak calling was paired end, for more accurate estimation of fragment length. Significant peak calls were made against input controls for TBI and Control samples and filtered by an FDR corrected P-value  $\leq 0.1$  or 10%. Additionally, significant peaks were called for H3K36me3 modENCODE data using an FDR

corrected P-value  $\leq 0.05$  and  $\log_{2}FC \geq +2$ fold. The significant peaks called for  $w^{1118}$  control, TBI, and H3k36me3 modENCODE samples were aggregated and averaged across all differential RIs detected in TBI samples. The RIs with significant peaks were overlapped between the  $w^{1118}$  control, TBI, and H3k36me3 modENCODE samples using *Vennable* in R/Bioconductor. For the common RIs the RPKM (Reads per Kilo-base per Million) were compared between  $w^{1118}$  control and TBI using Welch 2sample t-test.

For determining the H3K36me3 levels around the CA-rich motifs we reanalyzed our H3k36me3 data for  $w^{1118}$  control and TBI using HOMER. We called peaks against their respective input control using the following parameters; “*-fdr 0.1 -P 0.1 -F 0 -L 0 -LP 0.1 -style factor*”. The peaks detected were overlapped with MEDIPS peaks detected with intronic region for  $w^{1118}$  control and TBI (only peaks with minimum overlap 10 bases were kept). We re-centered these peaks around CA-rich motifs (ACACACA allowing 1 mismatch) and estimated the read coverage  $\pm 1000$  bps around the CA-rich regions. The results were plotted using *ggplot2* in R/Bioconductor.

**Splicing factor discovery:** The median intron size of retained introns was  $\sim 3000$ bps (min=115, max= 36560). Therefore, only retained introns with size  $\geq 600$  bps (421/458) was considered for motif analysis. For potential splicing factor binding motifs, 300bp regions were collected from the intronic side of the 5’SS and 3’SS. Sequences for these target regions were obtained and Motif enrichment analysis was carried out using DREME (Discriminative Regular Expression Motif Elicitation). DREME uses reshuffled input sequences as control to calculate motif enrichment. There are wide varieties of motif enrichment analysis software available publicly, which use varied mathematical models for enrichment test. Therefore we wanted to confirm our finding in DREME using different analysis strategy; MotifRG

(R/Bioconductor). We searched for enrichment of motifs against the randomly selected control sets of ~300bps windows from the dm3 build of the Drosophila genome. The enrichment analysis was run with 5 bootstrapping test to estimate score variance and filtered by bootstrap P-value  $\leq 0.1$  (MotifRG default parameters). Only the Top 10 results from MotifRG were considered.

**Calculating RI in mammalian datasets:** For mouse SCI and MTBI, the Percent RI (PIR) calculations were adapted from Braunschweig et al, 2015.

PIR= (Number of inclusion reads X 100)/ (Number of exclusion reads + Number of inclusion reads)

For calculating the *number of inclusion reads*, the first step was to define a list of target regions for read counting. We selected all annotated introns per transcripts from ENSMEBL (UCSC mm10 build) as our target regions. For this pre-defined set of targets, we counted the reads using summarizeoverlaps function in “Union” mode. *Inclusion reads* is comprised of 2 types of reads; mid-intron reads i.e. reads which lie within the intron and exon-intron junction reads which span the edges of the intron. “Union” mode of read counting accounts for the position of the read with respect to the target regions and any overlapping feature. Therefore, it is efficiently able to assign mid-intron and exon-intron junction reads to the respective targets. A better illustration of “Union” mode of counting reads can be found in Figure 4.2.12.

Exclusion reads on the other hand are equivalent to junction reads i.e. reads which maps across exon-exon junctions. Using the utilities provided in the spliceSites package in R/Bioconductor we mapped the gap-sites corresponding to the junction reads for each

sample. These gap-sites are representative of introns. For each gap-sites aka introns we counted the total number of junction read alignment and calculate RPGM values.

RPGM= (Number of aligns per splice site X  $10^6$ )/Number of gapped aligns per probe.

Then we filtered our gap-sites such that they have an  $RPGM \geq 5$  and are well correlated (~95% CI) across replicated RNA-Sequencing samples. The final list consisted of a set of novel and pre-annotated introns/gap-sites. We further filtered our curated intron sets and only considered gap-sites (introns) which overlapped with the targets used for counting the *inclusion reads* (maximum overhang length  $\leq 3$ bps), i.e. pre-annotated introns. The number of reads corresponding to these final set of gap sites are defined as *exclusion reads*.

Finally, for each intron/sample we calculated the PIR values as described before, and compared the PIR values between respective groups of samples using *Fisher Exact Test*. The results were further filtered using  $\Delta PIR \geq \pm 10\%$  and FDR corrected p-value  $\leq 0.05$ . For visualization of intron exclusion/retention the Normalized coverage was estimated using HOMER and plotted using Genome Browser.

**Modelling of intron retention in Drosophila and SCI datasets:** For modelling intron retention, we selected all the intron retention and exclusion events irrespective of the  $\Delta PSI$  and Bayes factor values and converted them in categorical variables ( $\Delta PSI > 0 \sim TRUE$ ,  $\Delta PSI < 0 \sim FALSE$ ). For determining the frequency of ACACACA motif we divided the introns in equal halves and counted the number of ACACACA motif allowing 1 mismatch at any base position. The frequency was calculated as follows.

ACACACA frequency= ACACACA count \*2/intron length

Then we modelled categorical  $\Delta$ PSI as a function of GC frequency, intron length (width), Log-fold change in mean RPKM of H3k36me3 peaks mapping to RI in 24 hours post-TBI samples compared to controls and frequency of ACACACA motif (1 mismatch) near the 3'SS or 5'SS.

$Cat(\Delta$ PSI)  $\sim$  GC frequency + width(Kb)\*logFC\*ACACACA frequency

To determine the best model, we used *step* function to select a formula-based model by AIC. Following the selection of the best model the coefficients were extracted and plots were made using ggplot2.

For modelling intron retention and exclusion in mouse SCI datasets we selected all RI/exclusion events (N=30460) irrespective of their FDR corrected p-value. For these events we divided the introns in equal halves and counted the number of ACACACA motif allowing 1 mismatch at any base position. Only sequences near 3'SS was chosen for further study as motif analysis placed the ACACACA motif near the 3'SS in mouse model of SCI. The frequency was calculated as follows.

ACACACA frequency= ACACACA count \*2/intron length

Then we modelled  $\Delta$ PIR (continuous) as a linear function of GC frequency, intron length (width) and frequency of ACACACA motif (1 mismatch) near the 3'SS. We selected the best model based on AIC using the *step* function in R/Bioconductor. The best model was as follows:

$\Delta$ PIR  $\sim$  GC frequency + width(Kb)\* ACACACA frequency

**Gene ontology analysis:** Gene ontology (GO) analysis was performed using DAVID bioinformatics resources (<https://david-d.ncifcrf.gov>). The background gene set for GO analysis were selected as follows;

1. gene expression; 5923 genes which were expressed in both controls and 24 hours post-TBI samples were used as background,
2. exon-usage; 9056 genes containing well represented exons (counts/exon  $\geq 2$ ) was used as background,
3. RI, *Drosophila*, dm3; 4209 genes containing RI events detected both in males and females were used as controls

PIR 7 days post-SCI, mm10; 6378 genes which were used for differential PIR test were used as background. If no, GO categories were found to be enriched than the all ENSEMBL genes (dm3 or mm10) were used as background. The Fold enrichment (over background) was plotted against the  $-\log_{10}(\text{FDR})$  for visualization of significant GO categories in ggplot2. The GO categories with  $\text{FDR} \leq 0.1$  was labelled in red.

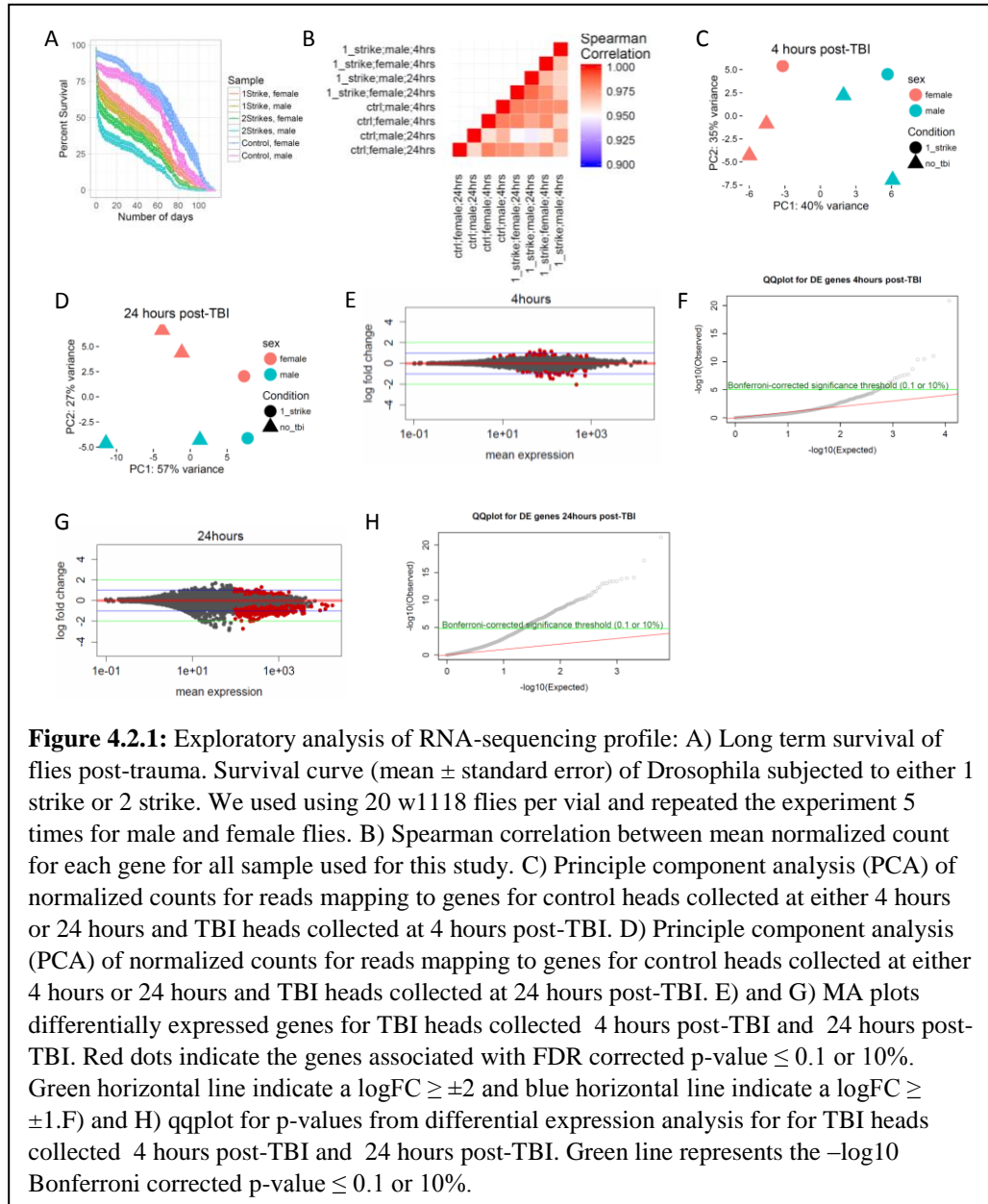
### 4.3. Results

#### Characterization of the transcriptome of *Drosophila* heads post-TBI.

We simulated closed head TBI in *Drosophila* using a similar experimental set-up, as described by the Wassarman lab (Katzenberger et al. 2013). Briefly, 0-5 days-old w1118 flies were collected in sturdy vials and TBI was inflicted using an in-house HIT (high impact trauma) device. Instead using a 90° spring deflection, we used a 45° spring deflection to attenuate the impact. Estimation of long term survival reported decrease in long term survival of TBI-flies. Furthermore, we also observed decrease in survival with increase in the number of strikes (see methods) (Fig 4.2.1A).

We collected heads at 2 time-points; 4 hours' post-trauma and 24 hours' post-trauma and characterized the transcriptomic profile using RNA sequencing. We observed that the expression profile of the heads post-trauma was very similar to control heads (Fig 4.2.1B).

Principal component analysis (PCA) of the expression profile revealed that majority of the sample-specific variability (approx. 40%) at 4hours post-



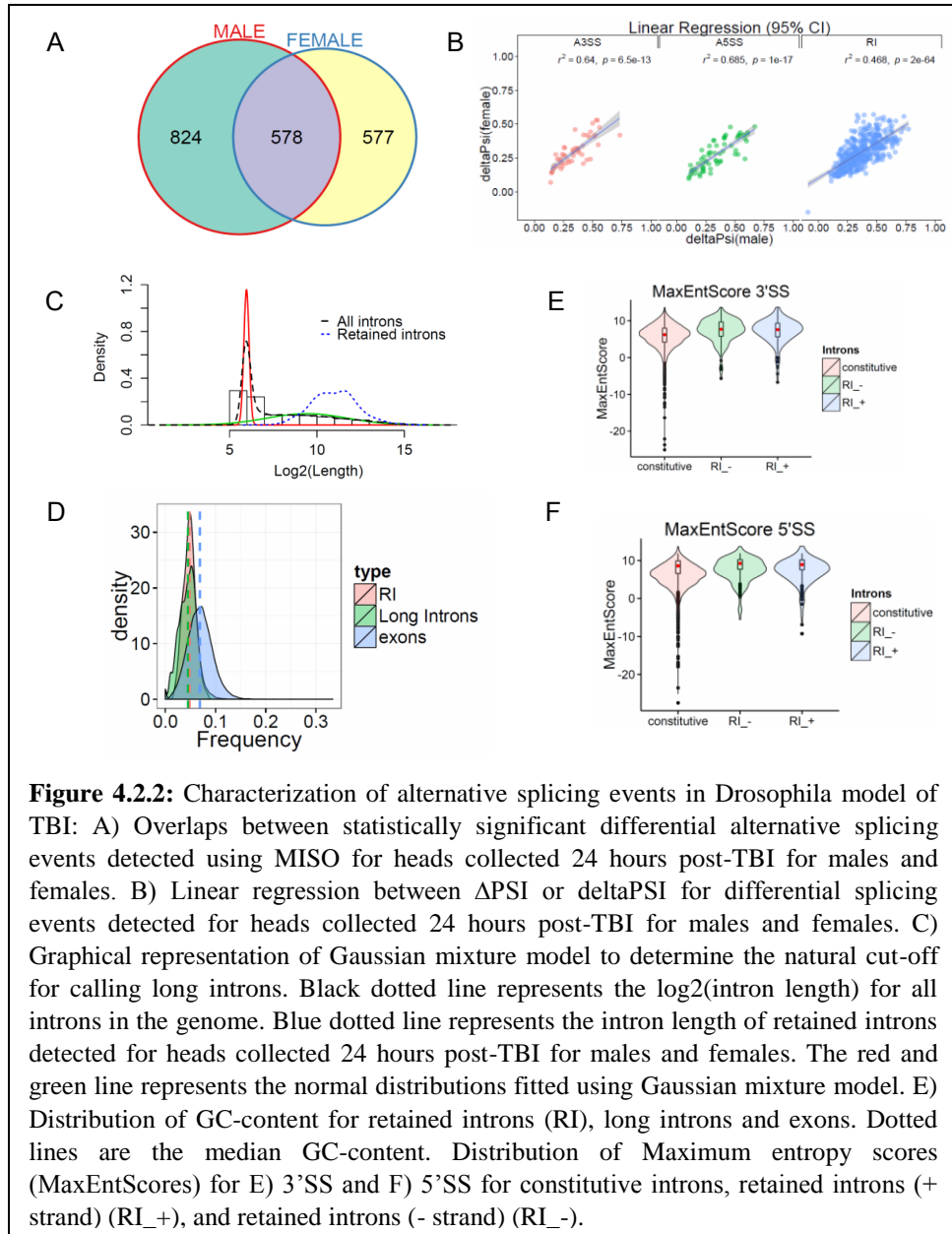
trauma was due to the sex of the flies (Fig 4.2.1C). Interestingly, 24 hours post-TBI, the expression profile showed observation separation dependent on the TBI (Fig 4.2.1D). The number of differentially expressed (DE) genes at an FDR corrected p-value cut-off of

0.1/10% was much greater at 24 hours (N= 1698) (Fig 4.2.1E) compared to 4 hours (N=145) (Fig 4.2.1G). Further filtering the DE genes by  $\log_{2}FC \geq \pm 1$  fold, restricted the list to 222 DE genes in heads collected 24 hours after TBI and 17 DE genes in heads collected 4hrs after TBI. Our result suggests that attenuated traumatic brain injury causes only mild change in steady state mRNA expression of genes and most of these changes are induced at 24 hours post-TBI. Therefore, change in expression of genes does not explain the decrease in fitness of the TBI flies.

RNA sequencing data can be used to interrogate the alternative splicing. Using exon centric analysis (see methods) we estimated the changes in alternative splicing 4hours and 24hours post-trauma separately for male and female flies. Then we only considered the sex-independent changes i.e. common between males and females for further study.



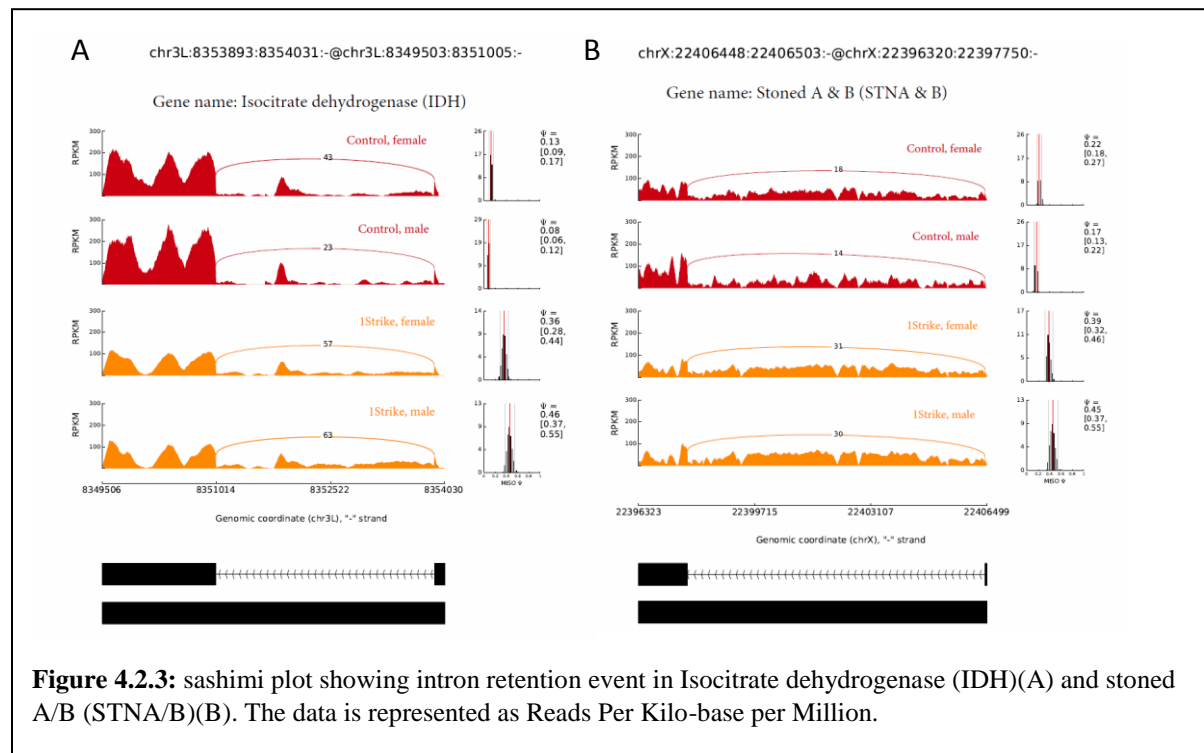
At 4 hours for a  $\Delta$ PSI (difference in Percent Spliced In)  $\geq \pm 0.05$  or  $\pm 5\%$  and Bayes Factor  $\geq 10$ , we observed very few changes. Interestingly, at 24 hours we observed a widespread induction of sex-independent splicing



changes (N=578) (Fig 4.2.2A). Furthermore, annotation of splicing events demonstrated that approx. 79% (N=458) of these changes were classified as intron retention (RI) (Fig 4.2.2B). These RI was enriched for genes involved in processes such as GO:0048489~synaptic vesicle transport, GO:0005811~lipid particles and GO:0005829~cytosol (data not shown). Examples of intron retention events in long intron of Isocitrate dehydrogenase (IDH) and synaptic

transport associated protein Stoned A/B (STNA/B) are illustrated in figure 4.2.3A and 4.2.3B.

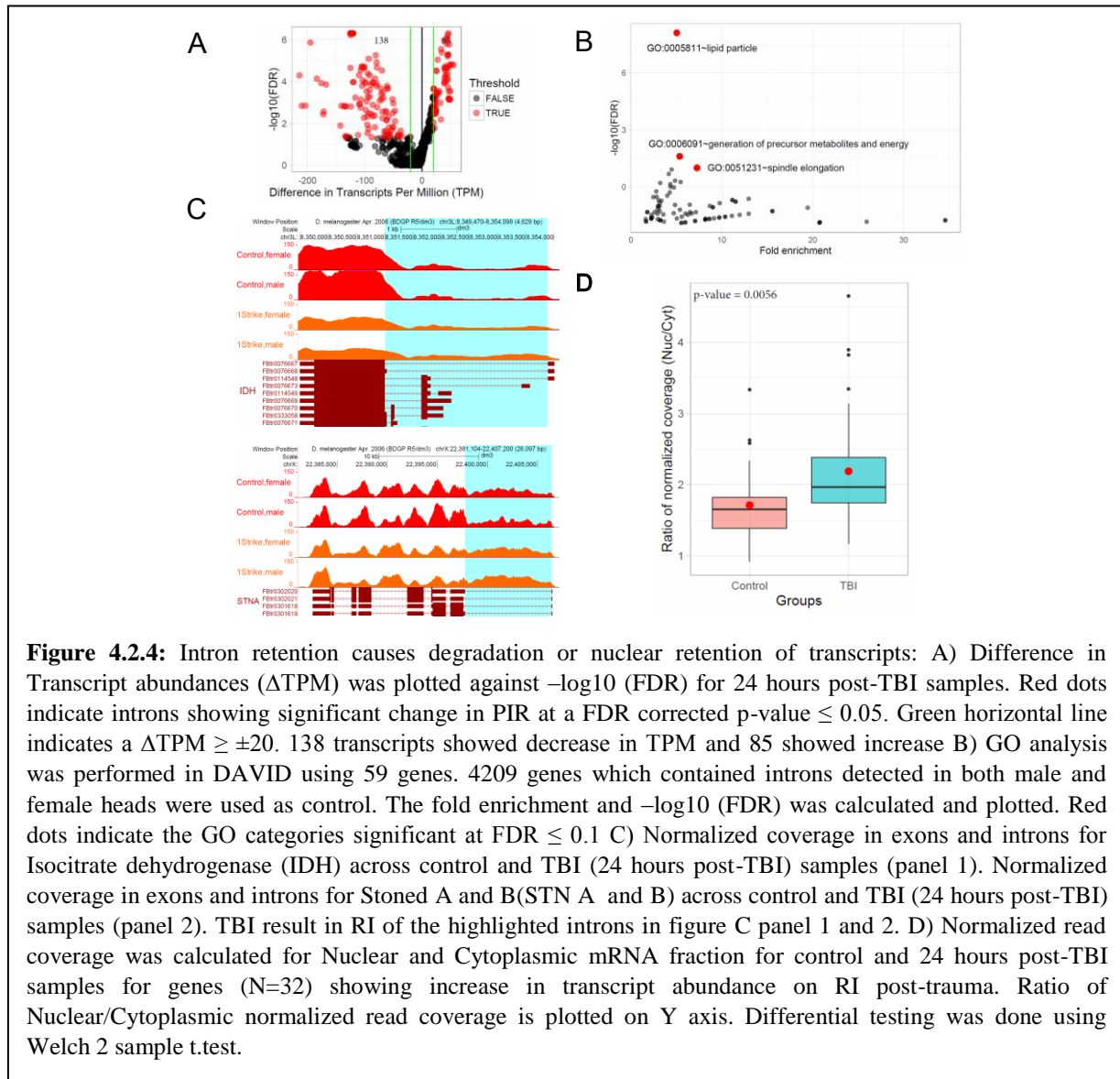
### Characterization of intron retention events 24 hours' post-trauma in *Drosophila* heads.



RI has been shown to be associated with specific defining characteristics of introns including GC content, intron length and flanking splice site strength. To estimate the intron length cut-off for classifying intron we modelled the log-transformed length of all introns in the *Drosophila* genome (dm3) as a mixture of 2 normal distributions (Gaussian mixture model). Using this model, we estimated the natural cut-off for calling long introns to be 81bps. We observed that all log<sub>2</sub>-transformed length of RIs was >81bps i.e. long introns (Fig 4.2.2C). The GC frequency of the RIs was very similar to other long introns in the dm3 genome and slightly lower than exons (Fig 4.2.2D). Studies have demonstrated that it is possible to differentiate between true and decoy splice sites using Maximum entropy scores

(MaxEntScore). Using the model proposed by Yeo et al, 2004 (Yeo and Burge, 2004) we calculated the MaxEntScore of 5'SS and 3'SS short sequence motifs for constitutive introns and retained introns. We defined constitutively spliced introns (constitutive) as introns which are efficiently processed and is not covered by even a single read in any sample. The maximum entropy scores at the 5'SS and 3'SS for retained introns were not significantly different from constitutive introns (Fig 4.2.2E and 4.2.2F). This suggested that the RI events are not caused due to activation of decoy splice sites during TBI (Fig 4.2.2E and 4.2.2F).

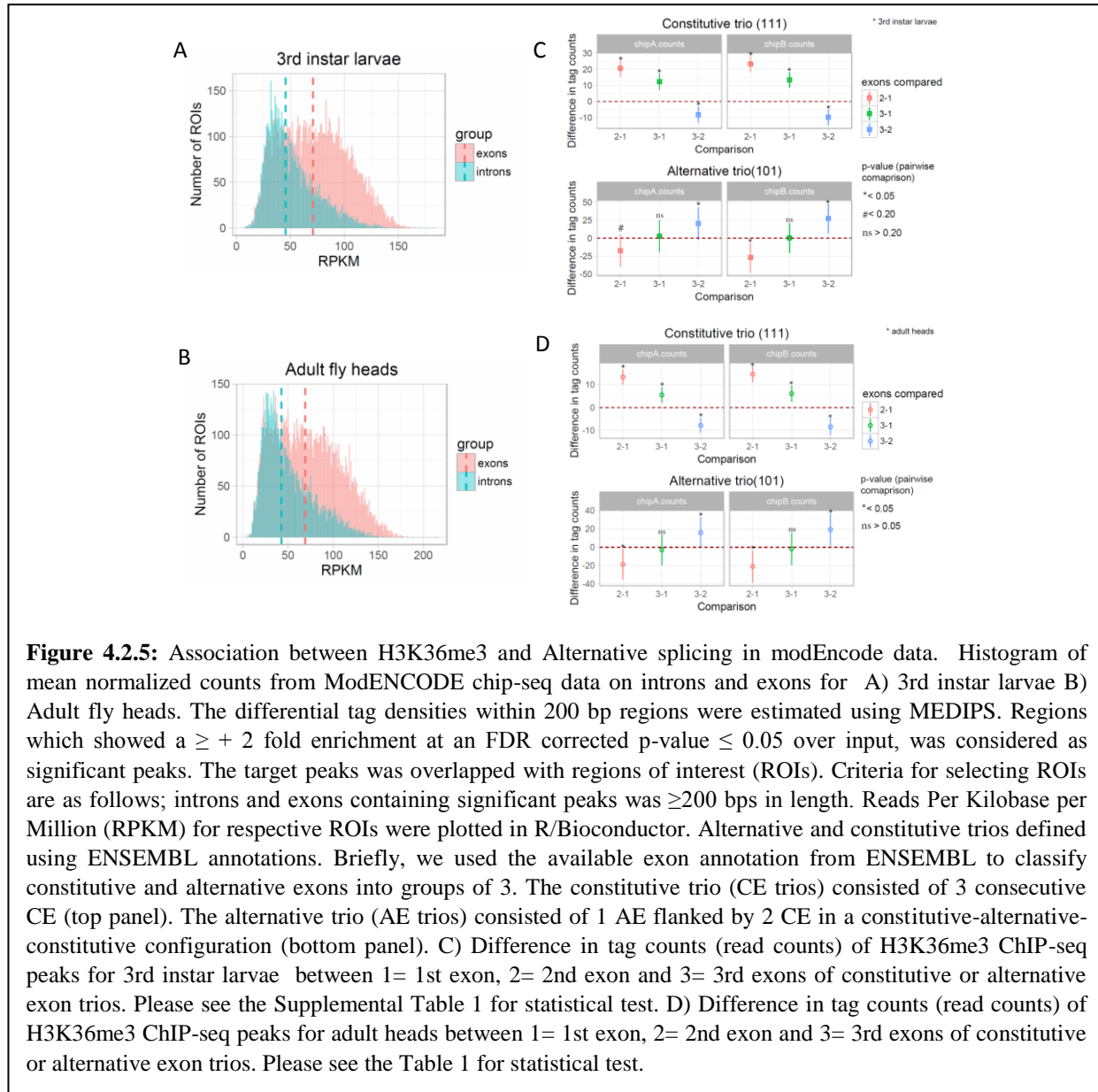
## Consequence of intron retention.



Inefficient processing of long introns has been demonstrated to introduce Pre-Mature Termination codons (PTC) and target the transcript to Nonsense Mediated Decay (NMD) (Chang et al., 2007; Isken and Maquat, 2007). Therefore, we hypothesized that transcripts showing RI will most likely show an appreciable decrease in transcript abundance or transcripts per million (TPM) (see methods). At an FDR corrected p-value  $\leq 0.05$  and  $\Delta$ TPM

$\geq \pm 20$  we observed 138 transcripts which showed a decrease and 85 transcripts which showed an increase in transcript abundance (Fig 4.2.4A). The 138 transcripts which showed reduced transcript abundance on RI, belonged to 59 genes. GO enrichment analysis (see methods) revealed significant enrichment of genes belonging to GO:0005811~lipid particles, GO: 0006091~generation of precursor metabolites and energy and GO:0051231~spindle elongation (Fig 4.2.4B). Examples are illustrated in figure 4.2.4C. The 85 transcripts which showed increased expression on RI showed significantly higher coverage in the nuclear fraction compared to cytoplasm fraction in TBI samples compared to WT controls. This suggested that these transcripts were getting accumulated inside the nucleus post-TBI (Fig 4.2.4D).

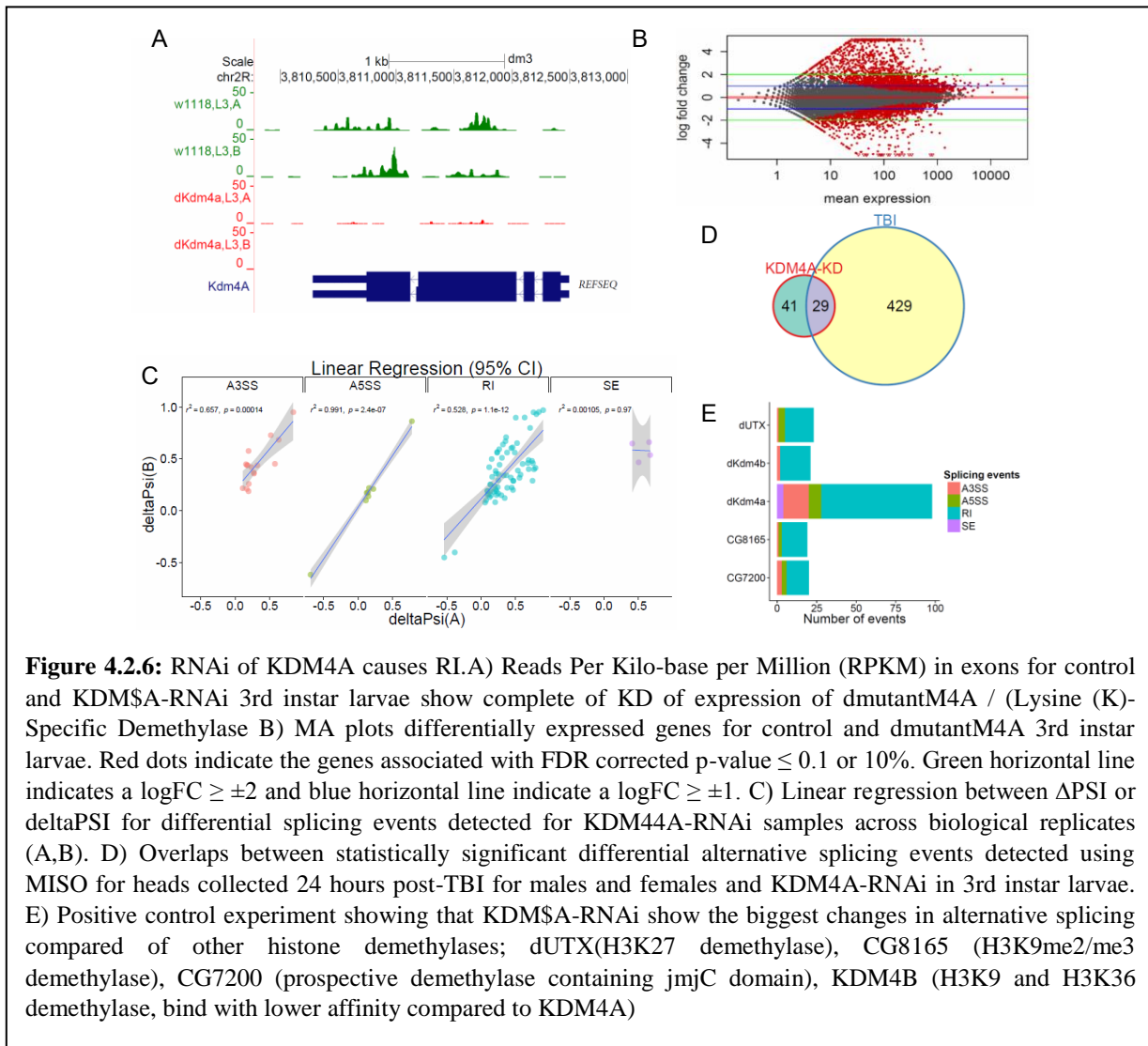
## H3K36me3 may play a regulatory role in splicing



A study by Ahringer's lab had previously reported association between histone modifications specifically H3k36me3 and splicing in model systems such as *C.elegans* and mice (Kolasinska-Zwierz et al., 2009). We wanted to determine if this association is conserved in *Drosophila*. To understand this association, we obtained ChIP-seq data for H3k36me3 in 3<sup>rd</sup> instar larvae (GSE47248) and fly-heads (GSE47280) from modEncode

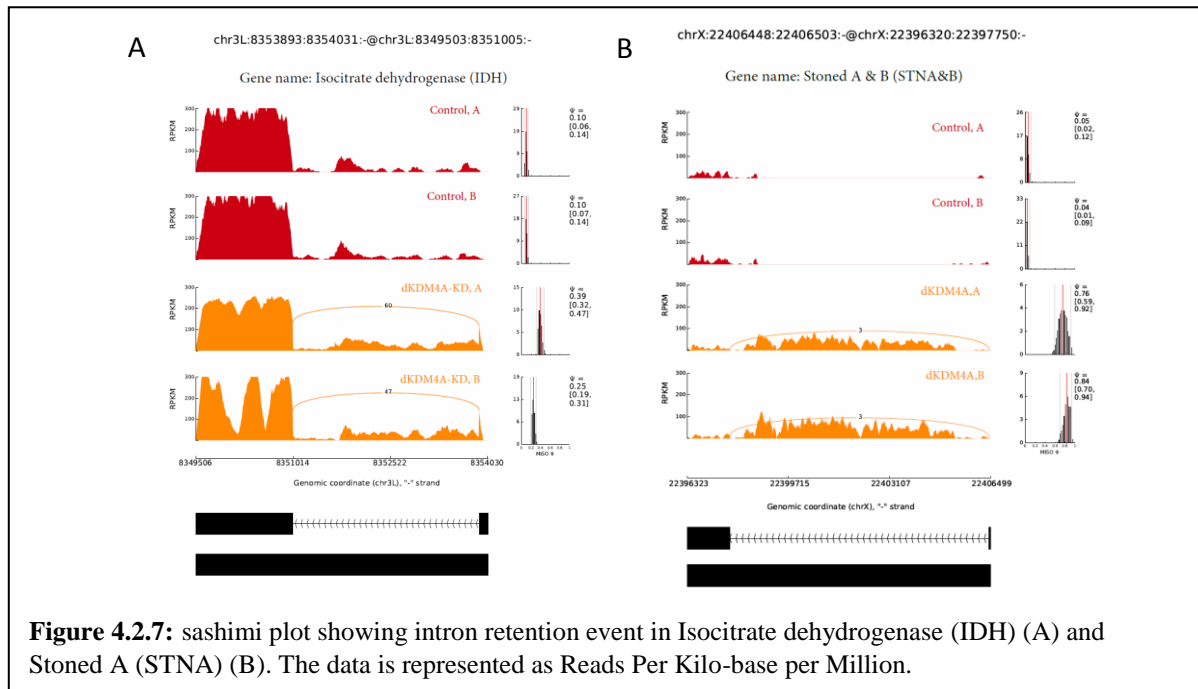
consortium. For these we called significant peaks within 200 bps target regions against the input controls at an FDR corrected p-value  $\leq 0.05$  and  $\log_2FC \geq 2$  (see methods). For the significant peaks we estimated the Reads Per Kilobase per Million (RPKM) within exonic and intronic regions. Density distribution of RPKM reported lower levels of H3k36me3 in introns compared to exons in both 3<sup>rd</sup> instar larvae and adult heads (Fig 4.2.5A and B). To further understand the association between alternative splicing and H3k36me3 we used the available exon annotation from ENSEMBL to classify constitutive and alternative exons into groups of 3 or trios. The constitutive trio (CE trios) consisted of 3 adjacent CE. The alternative trio (AE trios) consisted of 1 AE flanked by 2 CE in a constitutive-alternative-constitutive configuration (see methods). We observed that the H3K36me3 was depleted from the AE relative to the flanking CE in both 3<sup>rd</sup> instar larvae and adult heads (Fig 4.2.5C and D). One-sided ANOVA reported significant difference between the alternative and constitutive exons for the alternative trio (Table 4.2.1) for one of the replicates (P-adjusted  $\leq 0.1$ ) in 3<sup>rd</sup> instar larvae but the pattern was well conserved from the 3<sup>rd</sup> instar larvae to adult heads.

## RNAi of H3k36me3 demethylase causes intron retention in IDH.



H3K36me3 is removed from the genome by Jumanji-C (JmJc) domain containing histone demethylase specifically; KDM4A (Lin et al., 2012; Crona et al., 2013). Therefore, we hypothesized that RNAi of KDM4A will cause significant changes in RI. Homozygous loss of function mutations of KDM4A in *Drosophila* has been shown to cause developmental arrest (Tsurumi et al., 2013). Therefore, we attempted to conditionally knockdown KDM4A in adult heads. We were unable to get an appreciable and stable knockdown of the gene. KDM4A has been shown to be well expressed in 3<sup>rd</sup> instar larvae (Lorbeck et al., 2010). As



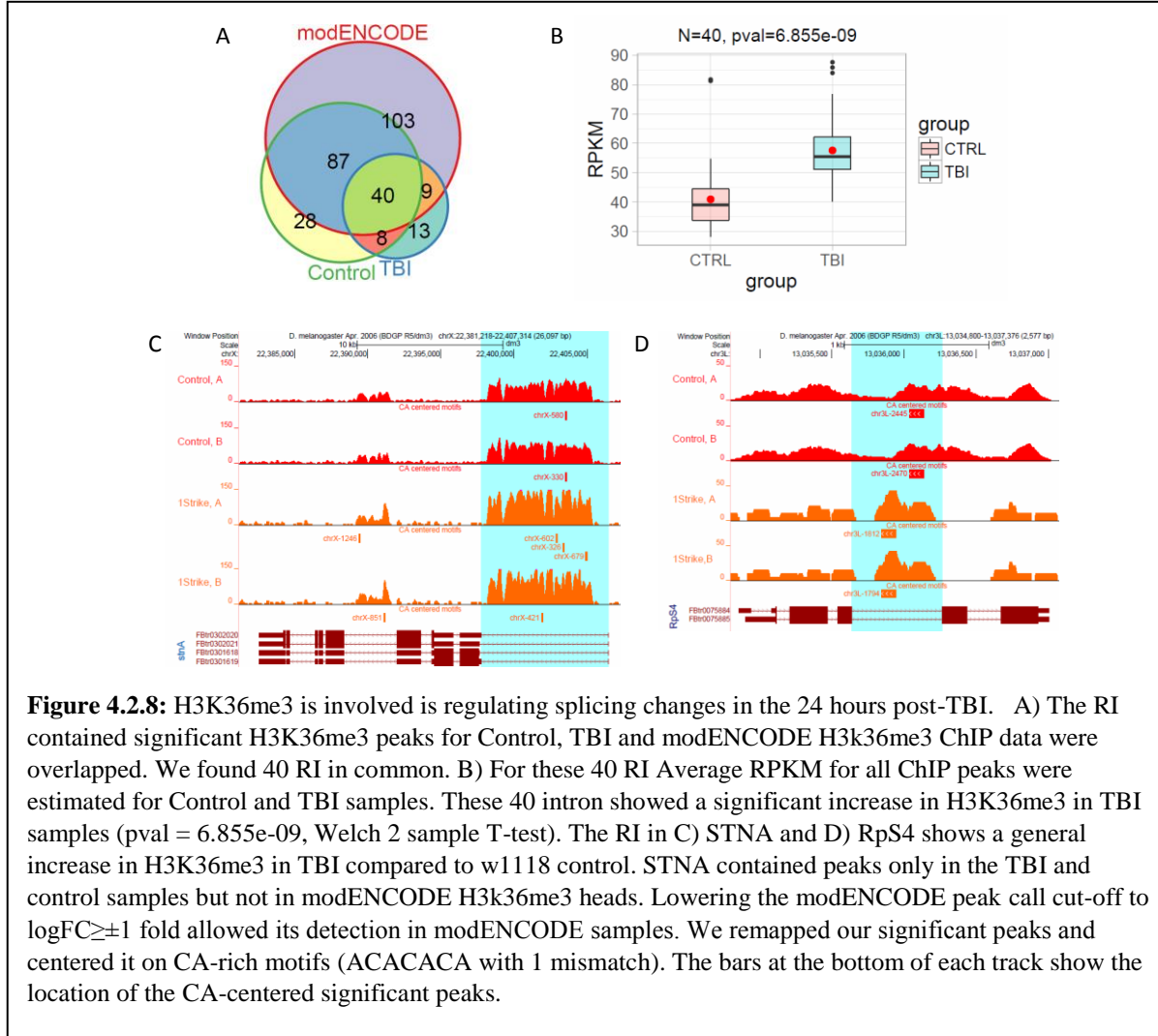


this was proof of concept experiment and we were trying to define the impact KDM4A on splicing regulation irrespective of physical trauma we decided to continue our study with 3<sup>rd</sup> instar larvae. We knocked down KDM4A in 3<sup>rd</sup> instar larvae using the UAS/GAL4 RNAi system (see methods) (Fig. 4.2.6A) and performed RNA-sequencing using 100bps paired-end reads. KDM4A activity has been reported to be associated with transcriptional repression in mammalian cells (Zhang et al., 2005; Huang and Dixit, 2011). Consistent with this observation, differential expression analysis comparing dKDM4A-RNAi to controls showed much larger number of genes which show increase in expression at an FDR corrected p-value  $\leq 0.1/10\%$  and log-fold change  $\geq \pm 2$ fold on KDM4A knockdown, as illustrated in Figure 4.2.6B. Therefore, in conclusion the KDM4A-RNAi in 3<sup>rd</sup> instar larvae were efficient and cause expected gene expression changes.

We compared the alternative splicing profile of KDM4A-RNAi 3<sup>rd</sup> instar larvae to wild-type w<sup>1118</sup> 3<sup>rd</sup> instar larvae. At a  $\Delta$ PSI  $\geq |0.05|$  or  $|5\%|$  and Bayes Factor  $\geq 10$ , we found

98 differential splicing changes. 70/98 events were RI and showed a significant ~52% correlation between biological replicates (Fig 4.2.6C). Notably, 29/70 RI events detected in KDM4A-RNAi were also detected 24 hours post trauma (Fig 4.2.6D). These common events included metabolic regulators such as IDH and PYK, and synaptic transport proteins such as STNA/B (Fig 4.2.7A and B). As additional controls we also performed KD of other histone demethylases such as H3K27me3 histone demethylase, UTX (Copur and Muller, 2013), H3K9me3 demethylase dKDM3A (CG8165) (Herz et al., 2014), KDM4B and JmjC-domain containing uncharacterized protein CG7200 in *Drosophila* pupae/ 3<sup>rd</sup> instar larvae. These histone demethylases showed limited number of splicing changes suggesting that KDM4A is a major regulator of splicing events (Fig 4.2.6E). Our RNAi study and the previous meta-analysis of H3K36me3 in 3<sup>rd</sup> instar larvae and adult heads demonstrate an association between H3k36me3 and splicing.

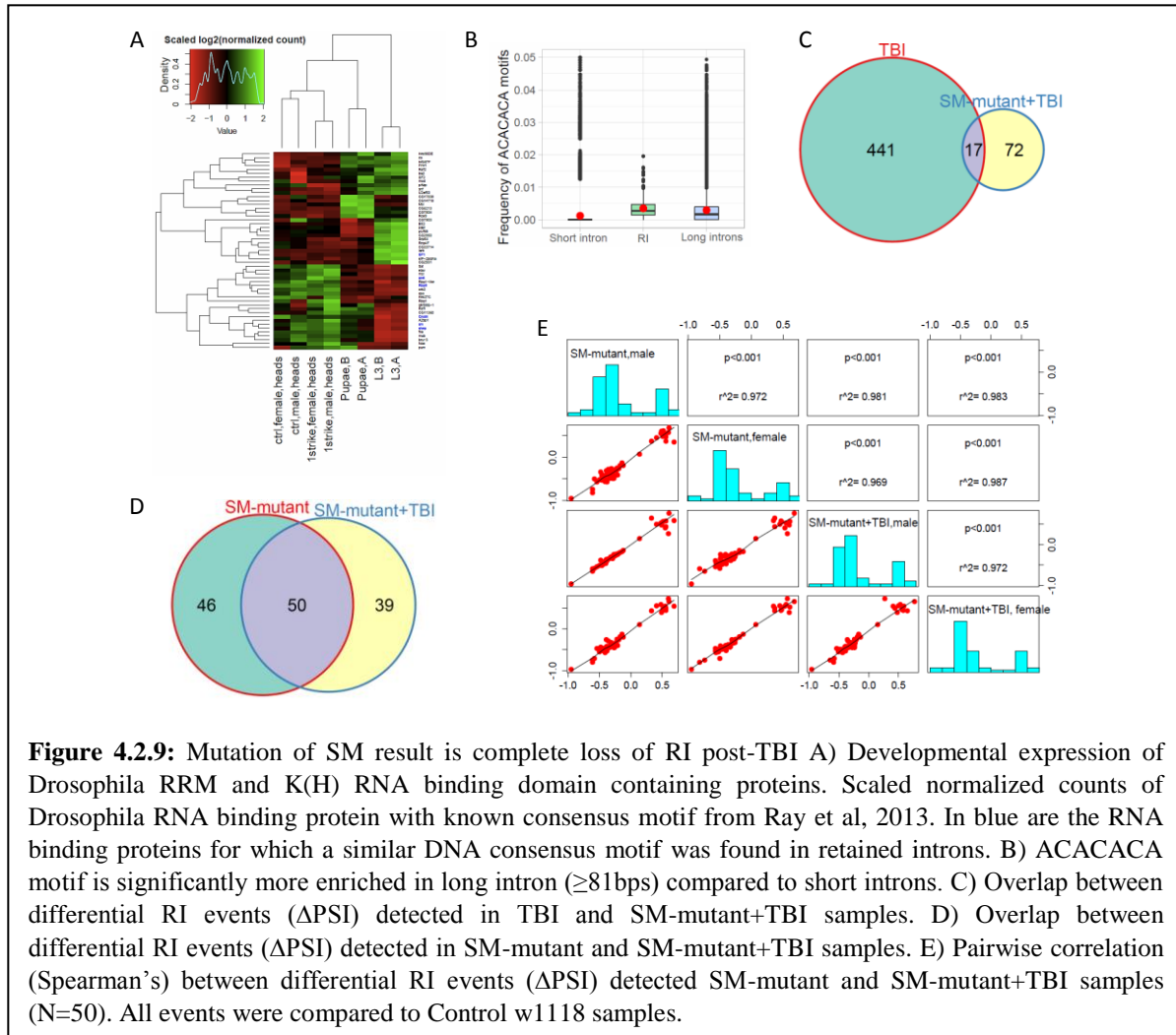
## H3K36me3 is involved in regulating splicing changes in the 24 hours post-TBI.



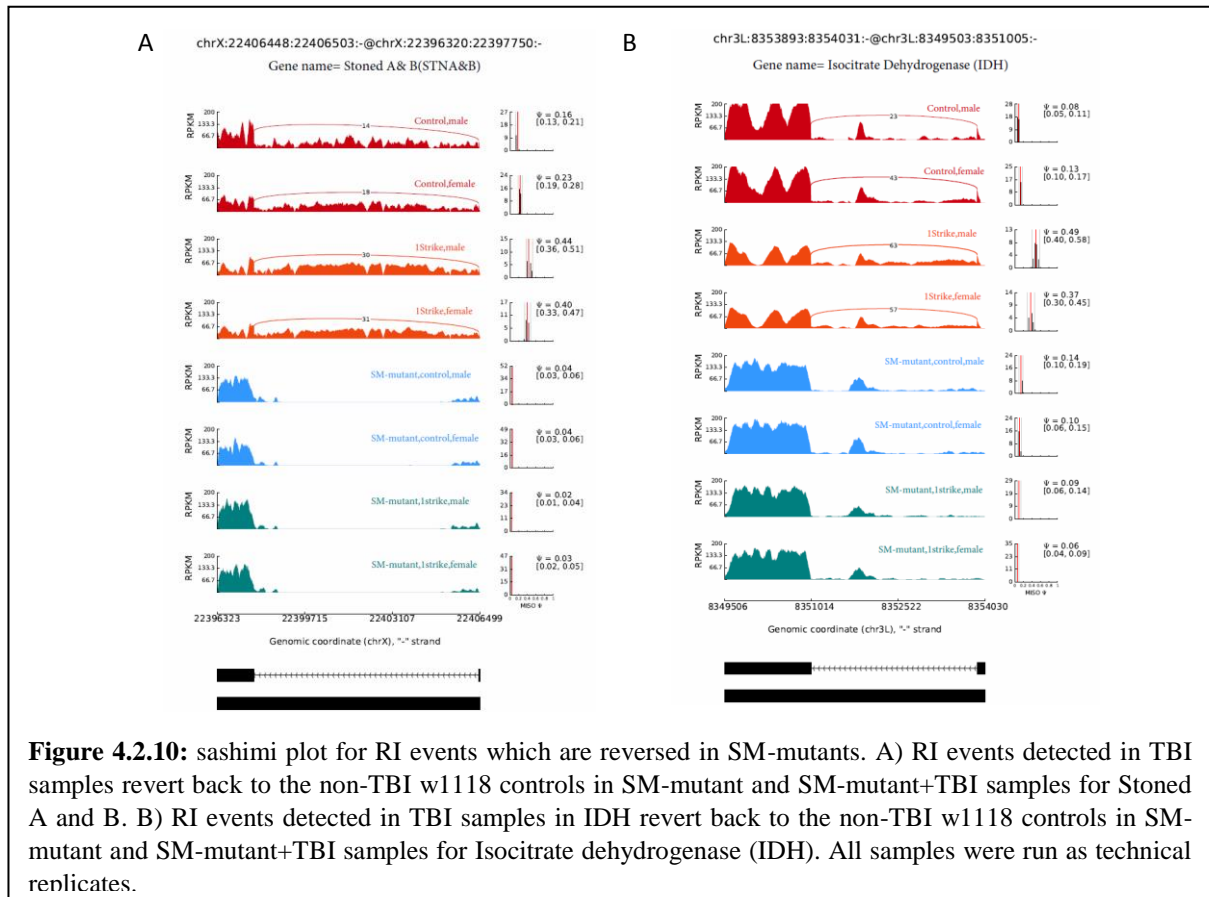
In view of the evidence from our KDM4A-RNAi studies in 3<sup>rd</sup> instar larvae we hypothesized that TBI results in increase in H3K36me3 within introns. To confirm this hypothesis, we performed ChIP for H3K36me3 in w<sup>1118</sup> control and TBI heads 24 hours' post-trauma and sequenced them using 100bps paired end reads. We called peaks against respective input control using 200 bps sliding windows. These peaks were further filtered by an FDR corrected  $p\text{-value} \leq 0.1$  or 10%. For the significantly enriched peaks we estimated the mean RPKM within RI. Only RIs with minimum of 8 reads were considered as well-

represented in our dataset. We detected 48/458 RIs which contained significant peaks in the w<sup>1118</sup> control and 24 hours post-TBI heads. To provide an independent validation of our H3k36me3 ChIP, we also calculated the mean RPKM of significant peaks for modENCODE H3k36me3 heads (GSE47280) within RIs (see methods). 40/458 RIs were represented in w1118 control, 24 hours post-TBI heads and modENCODE H3k36me3 ChIP-seq datasets (Fig 4.2.8A). For 40 RIs, the mean RPKM of H3K36me3 peaks for 24 hours post-TBI heads were significantly higher compared to the w1118 heads (Fig 4.2.8B). Normalized coverage plots for H3K36me3 for the long intron of STNA/B and RpS4 are illustrated in Fig 4.2.8C and 4.2.8D. In STNA we observed higher level of H3K36me3 in introns compared to exons. This is in agreement between basal levels of intron retention in the STNA long intron. On TBI, the H3k36me3 is significantly increased, which corresponds to increased inclusion of the long intron (Fig 4.2.8C).

## Discovery of splicing factor binding motifs.



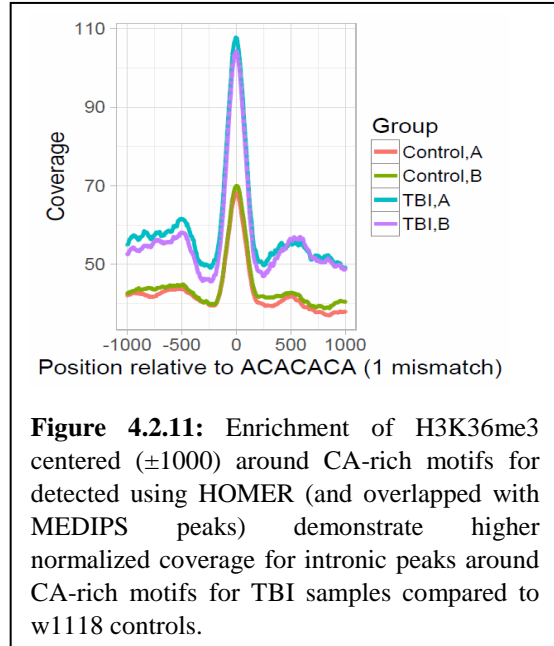
We searched for splicing factor binding motifs 300bps near either the 3' or 5'SS for RI  $\geq 600$ bps in length (421/458) (see methods). We found enrichment of CA rich motif (ACACACA) motifs near both the 3'SS and 5'SS (Table 4.2.1 and 4.2.2). The CA rich motif has been shown to bind to Smooth (SM) splicing factor (Ray et al., 2013). SM share approx. 42% identity with hnRNPL class of splicing regulator in humans (data not shown) and has been reported to be involved in regulation of splicing. We also found other motif for RNA binding proteins such as SHEP (AT-rich motif), ARET (GT-rich motif), RBP9 (T rich motif)



near the 5'SS (Table 4.2.2) and SHEP (AT-rich motif), SF1 near the 3'SS. These splicing factors are highly expression in adult heads and did not show any change in expression in 24 hours' post-trauma (Fig 4.2.9A). As the CA-rich binding motif is found near both the 5'SS and 3'SS of the retained introns with DREME we considered SM to be our strongest candidate. Furthermore, the frequency of ACACACA motif allowing 1 mismatch was significantly higher in long intron ( $\geq 81$  bps) compared to short introns (Fig 4.2.9B). Presence of CA-rich motifs in long introns might be defining characteristic of *Drosophila* long introns allowing its recognition and efficient processing.

To further understand the role of SM in regulation of splicing especially RI, we obtained a SM-mutant line ( $sm^4$ ). The  $sm^4$  mutant is semi-lethal over a deficiency (Df) (Karpen and Spradling, 1992). The resultant  $sm^4$ /Df flies survive to adulthood but are short

lived with a median age of about ~30 days (Layalle et al., 2005). The short lived phenotype of  $sm^4/Df$  mutants is mainly due to reduced arborization in the chemosensory neurons which culminates into a feeding defect (Layalle et al., 2005). After confirming the phenotypic characteristic of the  $sm^4$  mutant (data not shown) we subjected them to a single strike and collected the heads 24 hours after the TBI. We



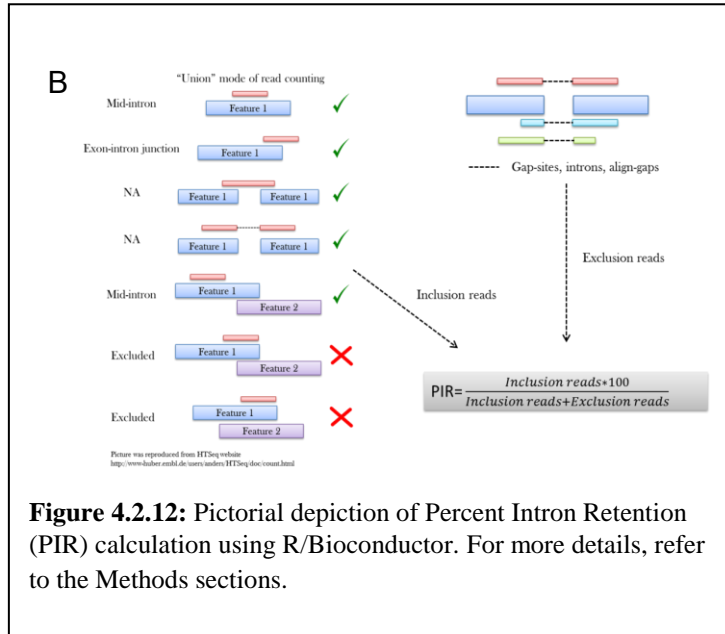
estimated the  $\Delta$ PSI using the MISO compared to the  $w^{1118}$  controls separately for males and females. We only considered the events which were overlapped between males and females for further analysis. We observed that  $sm^4/Df$  mutation rendered the flies insensitive to TBI i.e. RI events observed 24 hours post-TBI were not detected in  $sm^4/Df$  mutants subjected to TBI (Fig 4.2.9C). To further confirm this association, we performed RNA-sequencing of  $sm^4/Df$  mutant flies without trauma and found that the RI detected were well overlapped (Fig 4.2.9D) and well-correlated with  $sm^4/Df$  +TBI (Fig 4.2.9E). Examples of RI in  $sm^4/Df$  and  $sm^4/Df$  +TBI are illustrated in Figure 4.2.10 A and B.

### H3k36me3 pattern around SM-binding ISS.

To understand the relationship between SM and H3K36me3 we centered our H3k36me3 intronic peaks on the CA-rich motif (ACACACA) (see methods). We observed that the heads collected 24hours post-TBI showed higher levels of H3K36me3 compared to the controls (Fig 4.2.11). The ACACACA motif positions (100bps centered regions) surrounded by significant H3K36me3 peaks ( $\pm 1000$  bps) is indicated in figure 8C and D.

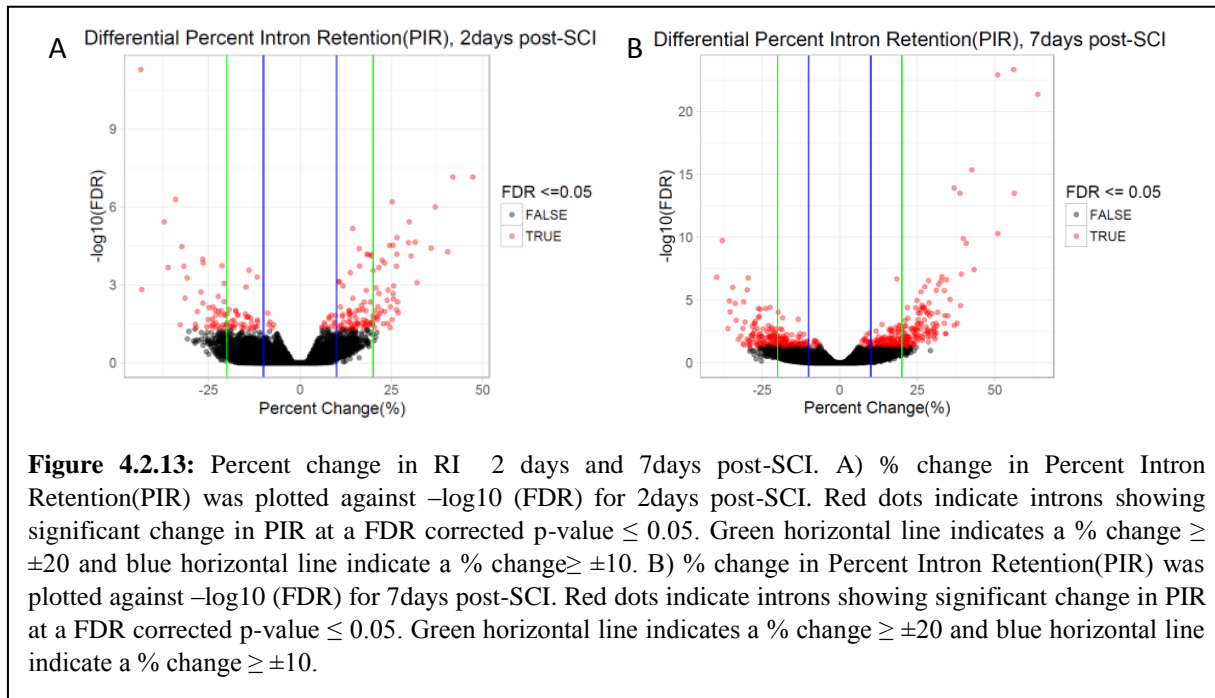
Paired with the observation from RNA-Sequencing sm<sup>4</sup>/DF mutants (Fig 4.2.9) we speculate that increased H3K36me3 level recruit SM (hnRNPL) splicing factors to introns and regulate RI.

### Intron retention in mouse model of spinal cord injury and traumatic brain injury.



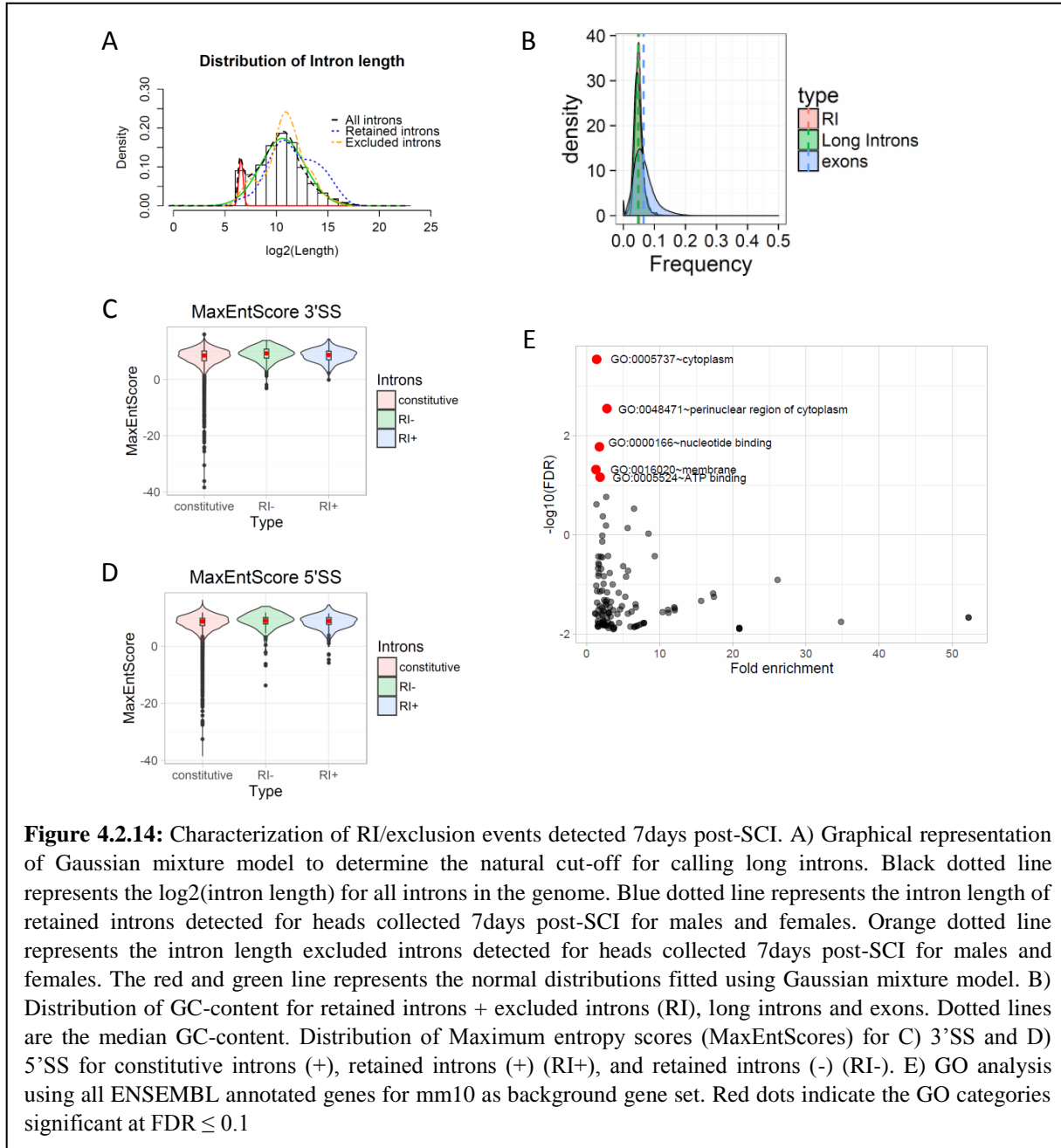
**Figure 4.2.12:** Pictorial depiction of Percent Intron Retention (PIR) calculation using R/Bioconductor. For more details, refer to the Methods sections.

In this study using a modified version of *Drosophila* model of TBI we report widespread RI 24 hours post-TBI. To verify if the RI plays an important role in transcriptomic stability post-trauma in mammals we performed meta-analysis of 2 datasets. The first dataset consisted of 24 female C57BI/6J mice subjected to contusive spinal cord





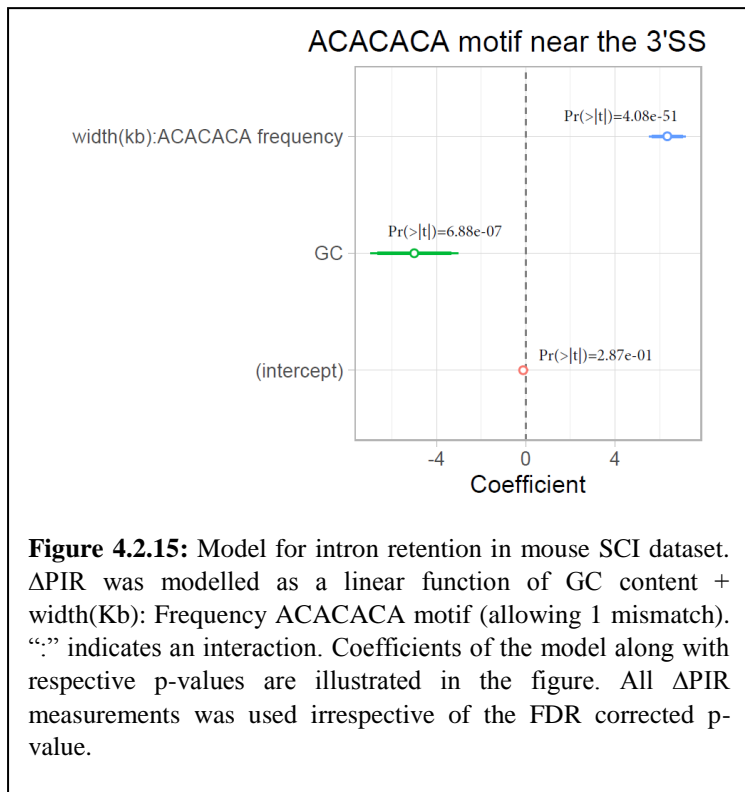
injury (Chen et al., 2013). Post-injury the RNA was extracted from tissues at 2days and 7days and sequenced using 100 bps paired-end libraries. This experimental design was quite similar to our model. Contrary to the consensus in the field, Braunschweig et al, 2014, reported widespread RI in mammalian systems. The authors demonstrated the RI functions as



a quality control signal to remove incorrectly spliced transcripts either by nuclear retention

or/and by Nonsense mediated decay (Braunschweig et al., 2014). For purposes of detecting RI in mammalian systems we adapted the approach used by this study. Briefly, we used control and SCI RNA-sequencing data to detect alignment gaps/ gap-sites. These gap-sites are representative of exon-exon junction. For quantification of the number of reads aligned to the gap sites, we calculated the Reads Per Million Gapped (RPGM) for all detected gap-sites per sample. Then we selected the

gap sites for which the RPGM were well correlated (~95% CI) between biological replicates and had a  $RPGM \geq 5$  (data not shown). The resulting output consisted of a list of putative introns (pre-annotated + novel introns) and the number of reads mapping to each respective gap-sites. These reads were classified as “Intron Exclusion Reads”.



Among the putative introns, only pre-annotated introns were considered for analysis. The final list included approx. 50,000 introns. For these introns, we counted the number of reads which fell within the introns and spanning the intron-exon boundary. These we defined as “Intron Inclusion Reads”. Then we calculated Percent RI (PIR) as a ratio of inclusion reads and summation of inclusion and exclusion reads. Finally, to test for significant differences between conditions we performed a Fisher Exact test (Fig 4.2.12).

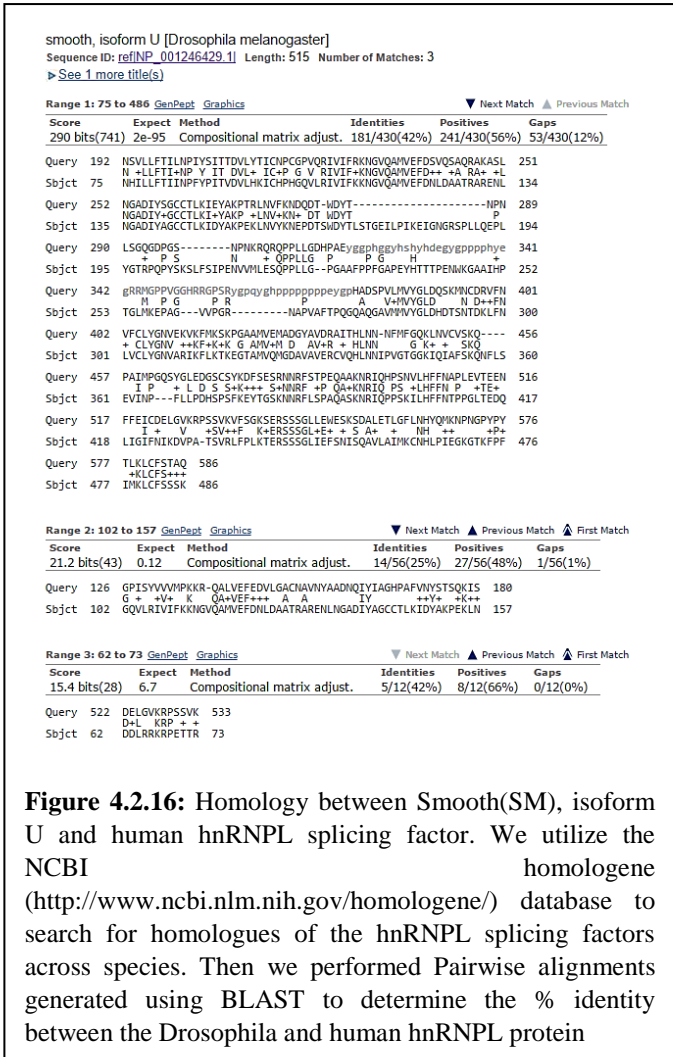
At an FDR corrected p-value  $\leq 0.05$  and  $\Delta\text{PIR} \geq \pm 10\%$  we observed 189 at 2days and 413 differential events at 7days (Fig. 4.2.13A and B). We characterized the intron length, GC-content and Splice site strength for 413 RI/exclusion events detected 7days post-SCI. Using the strategy described earlier, we calculated the natural cut-off for long introns to be approximately  $>128$  bps. The

RI/excluded introns detected to be differentially spliced 7days post-SCI were  $>128\text{bps}$  in length (Fig 4.2.14A).

The average GC content of spliced introns was not different from the average GC-content of all long introns in the mouse genome and slightly lower than that of exons (Fig 4.2.14B).

Similarly, the splice site strength represented as maximum entropy scores (MaxEntScore) for 5'SS and 3'SS for spliced introns was not statistically different from constitutive introns (Fig 4.2.14C and D). These

observations indicate similar characteristics of RI/excluded introns post-SCI in mouse and post-TBI in Drosophila. However, only a limited number of genes showing RI/exclusion showed significant changes in transcript abundance (at an FDR corrected p-value cut-off of  $0.05$  and  $\Delta\text{TPM} \geq \pm 20$ ) (data not shown). GO analysis using all ENSEMBL genes as



**Figure 4.2.16:** Homology between Smooth(SM), isoform U and human hnRNPL splicing factor. We utilize the NCBI [homologene](http://www.ncbi.nlm.nih.gov/homologene/) (<http://www.ncbi.nlm.nih.gov/homologene/>) database to search for homologues of the hnRNPL splicing factors across species. Then we performed Pairwise alignments generated using BLAST to determine the % identity between the Drosophila and human hnRNPL protein

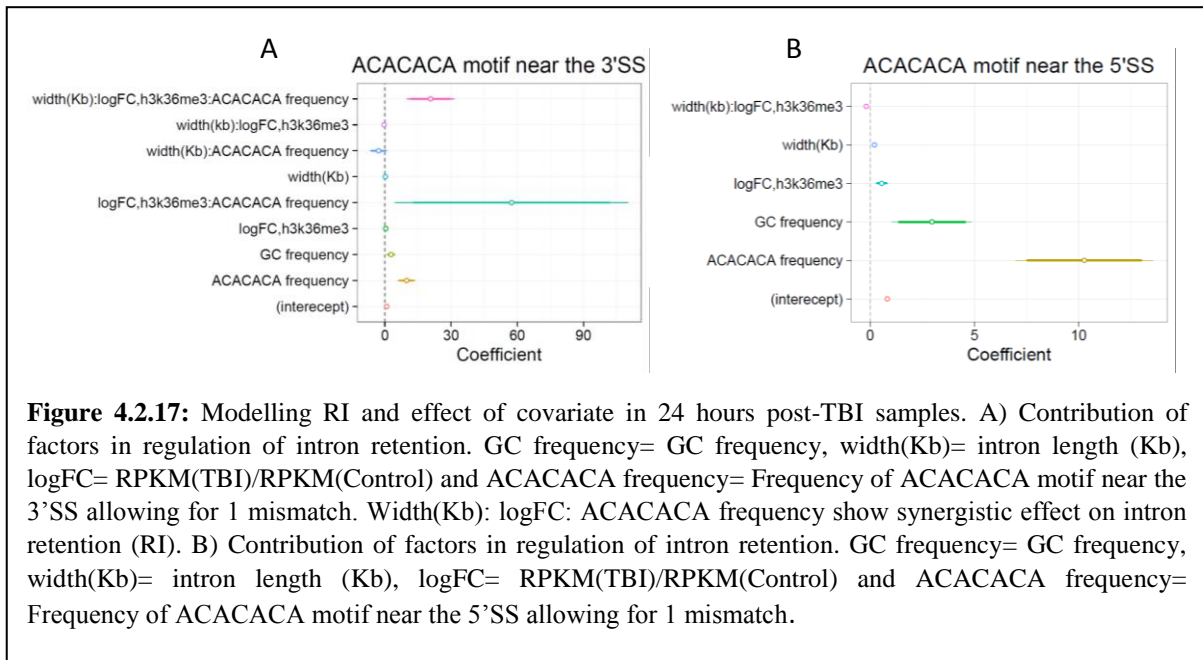
background revealed enrichment ( $FDR \leq 0.1$ ) of processes such as GO:0048471~perinuclear region of cytoplasm, and GO:0000166~nucleotide binding and GO:0005524~ATP binding (Fig 14E). This included lysosomal genes implicated in neurodegeneration such as Cathepsin D (CTSD) (Fig 4.2.15 top panel) (Cataldo et al. 1995; Koike et al. 2000), and Vacuole protein sorting 18 (VPS18) (Peng et al. 2012a; Peng et al. 2012b) (Fig 4.2.15 bottom panel). Motif analysis using DREME using the same parameter as *Drosophila* dataset revealed enrichment of CA-rich motif near the 3'SS of the differential RI/exclusion events suggesting that the functional association between hnRNPL binding and intron processing might be conserved across species (Table 4.2.4 and 4.2.5). To further understand this relationship between CA-rich motif and RI/exclusion events in mouse model of SCI, we modelled  $\Delta$ PIR as a linear function of GC content, ACACACA frequency with one mismatch near the 3'SS and width of the introns (Kb) (see methods). Our analysis revealed that demonstrated that high frequency of ACACACA motif near 3'SS of long intron is associated with RI not exclusion (Fig 4.2.15).

#### 4.4. Discussion

The role of intron retention in regulating transcript abundance under neurodegenerative disease conditions remain poorly understood. In this study using a modified version of *Drosophila* model of TBI, we provide preliminary evidence that, retention of long introns (>81bps) functionally tunes the transcriptome in response to neuronal damage. We observed increased intron retention coupled with decreased transcript abundance in genes associated with the citric acid cycle such as Isocitrate dehydrogenase (IDH), Enolase (ENO), Pyruvate Kinase (PYK) and Aconitase (ACON) 24 hours' post-trauma. ACON catalyzes the conversion of cis-Aconitate to D-isocitrate (Beinert et al., 1996)

and IDH catalyzes the conversion of D-isocitrate to  $\alpha$ -Ketoglurate (Gabriel et al., 1986; Yen and Schenkein, 2012). Therefore, we speculate that downregulation of expression of TCA cycle genes are a consequence of oxidative stress caused by TBI and will result in downregulation of mitochondrial metabolism.

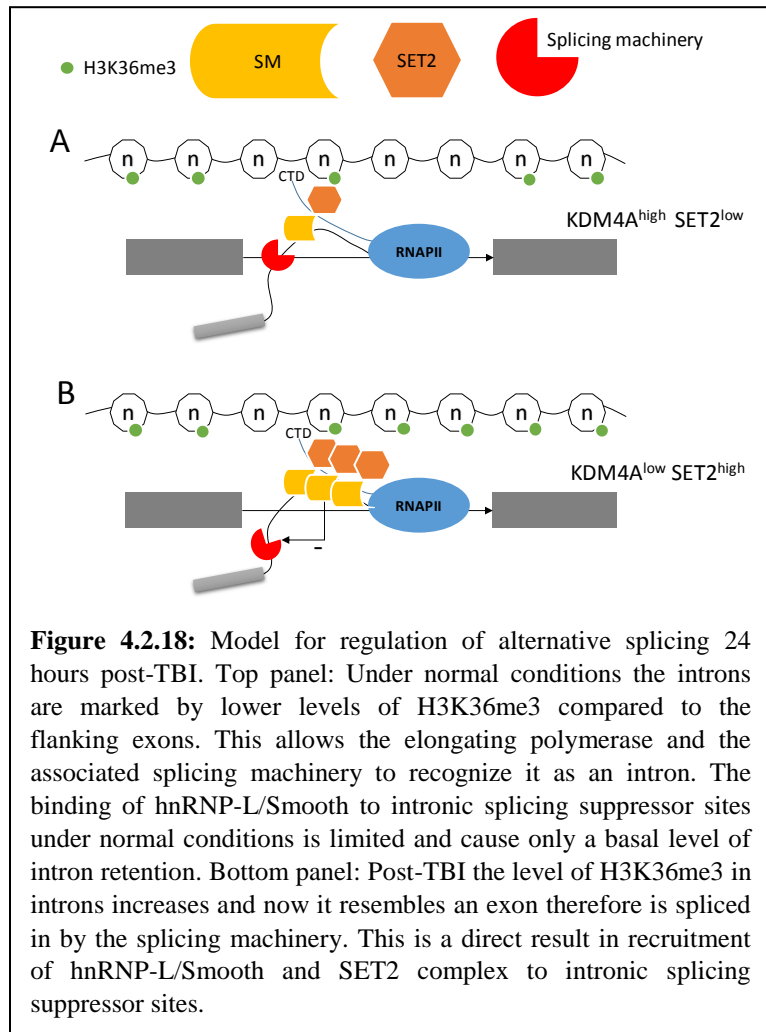
Downregulation of mitochondrial metabolism is neuroprotective. Restoration of mitochondrial function using creatinine in Down syndrome patient-derived astrocytic cells,



resulted in increased ROS production and consequently loss of cell viability (Helguera et al., 2013). Our data indicates that TCA cycle genes such as IDH, ACON, PYK and ENO might be critical regulators of the neuroprotection. Dysregulation of expression of genes such as IDH and ACON can significantly impact the production of a critical metabolite;  $\alpha$ -Ketoglutarate ( $\alpha$ -KG).  $\alpha$ -Ketoglutarate is an important co-factor in regulating all dioxygenase reactions in the cell including H3k36me3 demethylation (Klose and Zhang, 2007). In a recent study by Carry et al, 2015, direct perturbation of  $\alpha$ -KG/succinate levels was shown to be sufficient to cause changes in H3K27me3 and DNA hydroxymethylation in mouse

embryonic stem cells suggesting  $\alpha$ -KG is an important metabolic effector of epigenetic modification (Carey et al., 2015). We hypothesized, that decrease in total  $\alpha$ -KG reserve due to downregulation of expression of TCA cell cycle might lead to decreased activity of KDM4A; H3K36me3 demethylase. Increased levels of H3K36me3 might interact with splicing suppressor to their Intronic splicing suppressor sites (ISS) and result in intron retention. The potential link between metabolism and splicing especially RI is strongly suggestive.

We observed enrichment of CA-rich motifs near the 5'SS and 3'SS of RI detected 24 hours' post-trauma. CA-rich motif has been shown to bind to hnRNPL and LL class of splicing factor (Hung et al., 2008; Ray et al., 2013).



Functionally the binding of these splicing factors has been implicated in the regulation of a wide range of alternative splicing events such as exon skipping, and suppression of variable exons. Interestingly, RNAi mediated KD of hnRNPL /LL has been shown to cause intron retention in CD55 and STRA6 genes in HeLa cell lines (Hung et al., 2008). This suggested

that binding to hnRNPL /LL might be required for efficient processing of introns. *Drosophila* genome encode for a homologue of the hnRNPL protein, known as Smooth(SM). SM contains 3 ranges corresponding to RRM (RNA binding) domains which share approximately 42%, 25% and 42% identity with human hnRNPL protein (Fig 4.2.16). To investigate the relationship between SM protein and intron retention in our TBI model, we used a SM-mutant ( $sm^4$ ) line (Karpen and Spradling, 1992; Layalle et al., 2005). We used the  $sm^4/DF$  mutant instead of RNAi lines because these flies have a well characterized and easily observable phenotype. The SM-mutants have an average life-span of ~30days due to defective feeding behavior. This phenotype has been attributed to the lack of chemosensory axon arborization in the leg neuromere (Layalle et al., 2005). Subjecting the  $sm^4/DF$  mutants to TBI resulted in loss of RI events observed 24 hours' post-trauma in  $w^{1118}$  flies. In some cases, for example in STNA/B 1<sup>st</sup> intron, even the basal level of RI was reversed. This result provided strong evidence that SM binding to Intronic splicing suppressor (ISS) sites is necessary for RI 24 hours post-TBI. This result was contrary to our expectations. However, it is worth noting that Splicing Factors have multiple binding sites along the genes. Their impact on RNA splicing may vary depending upon their binding to Intronic or exonic sites. In this study, we hypothesize that SM (hnRNPL) is most likely being recruited to ISS which result in RI. As the  $sm^4/DF$  is short lived it also suggests that TBI-associated increase in RI might be protective in nature.

Alternative splicing in eukaryotic organism has been shown to occur co-transcriptionally. Therefore, the splicing machinery is in close proximity with surrounding histone modification which may interact directly or indirectly with splicing factors and regulate splicing. One of the modifications that has been implicated in regulation of splicing

is H3K36me3. As mentioned before, H3K36me3 has been shown to be depleted from alternative exons compared to flanking constitutive exons in *C.Elegans* and Mice. This seemingly functional relationship seems to well conserved in *Drosophila*. Furthermore, consistent with our hypothesis we observed increase in H3K36me3 modification within RI in 24 hours post-TBI heads. This result was further substantiated by our KDM4A-RNAi studies in 3rd instar larvae. We were able to recreate some of the major RI found 24 hours post-TBI in KDM4A-RNAi 3rd instar larvae including genes associated with mitochondrial metabolism (IDH and PYK) and neuronal transport protein STNA/B. Re-mapping all intronic reads 1000bps around the SM motif (ACACACA) also showed an enrichment of H3k36me3 in 24 hours post-TBI heads compared to controls within introns. Therefore, multiple lines of evidence suggest possible relationship between splicing and H3k36me3. We tried to model all intron retention ( $\Delta\text{PSI}>0$ ) and exclusion events ( $\Delta\text{PSI}<0$ ) in our TBI samples as a linear function of log fold-change in H3k36me3 peaks in TBI compared to control samples, GC frequency, intron length and frequency of ACACACA motif (1 mismatch) near the 3'SS or 5'SS (see methods). At a p-value cut-off  $\leq 0.001$  we found that the interaction between the intron length, log fold-change in H3k36me3 peaks and frequency of ACACACA motif near 3'SS was predictive of intron retention ( $\Delta\text{PSI}>0$ ) and showed synergistic affects (Table 4.2.6, Fig 4.2.17A and B).

A study by Yuan et al, 2009 reported hnRNPL to be an integral part of the lysine trimethylase 3A (KTM3A) complex (Yuan et al. 2009). KMT3A (also known as HYPB or hSET2) is a histone methyltransferase specifically shown to increase H3K36me3 levels in mammals and *Drosophila* (Bell et al. 2007; Edmunds et al. 2008). The authors demonstrated that RNA interference against KMT3A or hnRNPL down-regulates exclusively the















H3K36me3 mark in HeLa cells. This suggests that hnRNPL might be required for hSET2 methyltransferase activity. Therefore, based on the results from this study and our results we propose that 24 hours' post-trauma, there is an increase in recruitment of SM (hnRNPL) binding proteins to the 3'SS of long introns. This leads to recruitment of SET2 HMT to the ACACACA sites and increases the levels of H3K36me3 within these introns (Fig 18). Simultaneously, an increase in mitochondrial stress post-TBI reduces the availability of  $\alpha$ -KG, and, thereby, causes a decrease in KDM4A histone demethylase activity. Therefore, the disruption of balance between the SETD2 histone methyltransferase and KDM4A histone demethylase activities contribute to an increase in H3K36me3 within long introns.

## Tables

**Table 4.1:** ANOVA followed by TukeyHSD for difference in tag counts (read counts) of H3K36me3 ChIP-Seq peaks between 1= 1st exon, 2= 2nd exon and 3= 3rd exons of constitutive or alternative exon trios for 3rd instar larvae(GSE47248) and adult heads(GSE47280).








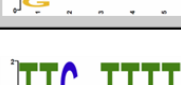

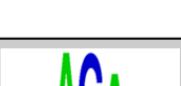
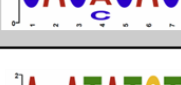
diff	lwr	upr	p.adj	Comparison	type	group
13.37804	10.06118	16.6949	0	2-1	CT	adult heads, A
5.506706	2.189848	8.823565	0.000295	3-1	CT	adult heads, A
-7.87134	-11.1882	-4.55448	8.09E-08	3-2	CT	adult heads, A
14.6052	11.06998	18.14042	0	2-1	CT	adult heads, B
6.123988	2.588766	9.65921	0.000146	3-1	CT	adult heads, B
-8.48122	-12.0164	-4.94599	5.69E-08	3-2	CT	adult heads, B
-18.7569	-35.3136	-2.20013	0.021761	2-1	AT	adult heads, A
-2.88908	-19.4458	13.66766	0.911434	3-1	AT	adult heads, A
15.86778	-0.68896	32.42452	0.063595	3-2	AT	adult heads, A
-21.0012	-38.3085	-3.69388	0.012544	2-1	AT	adult heads, B
-1.7153	-19.0226	15.592	0.970517	3-1	AT	adult heads, B
19.28587	1.978582	36.59317	0.024585	3-2	AT	adult heads, B
20.6737	15.60103	25.74637	0	2-1	CT	Larvae, A
12.37894	7.306264	17.45161	3.25E-08	3-1	CT	Larvae, A
-8.29477	-13.3674	-3.22209	0.000375	3-2	CT	Larvae, A
23.28809	18.5096	28.06659	0	2-1	CT	Larvae,B
13.39256	8.614069	18.17106	0	3-1	CT	Larvae,B
-9.89553	-14.674	-5.11704	3.67E-06	3-2	CT	Larvae,B
-17.427	-39.9339	5.079979	0.163815	2-1	AT	Larvae,A
2.851521	-19.6554	25.35846	0.952205	3-1	AT	Larvae,A
20.27848	-2.22846	42.78542	0.087159	3-2	AT	Larvae,A
-26.9479	-47.575	-6.32084	0.006393	2-1	AT	Larvae,B
0.275923	-20.3512	20.90302	0.999454	3-1	AT	Larvae,B
27.22387	6.596766	47.85097	0.005779	3-2	AT	Larvae,B

**Table 4.2:** Motif enrichment analysis for 5'SS. 299 bps sequences were collected from intronic side of 5'SS for retained intron of length  $\geq 600$  (N= 421/458). Motif enrichment analysis was run using reshuffled sequences as background (Parameters; . Strand Handling=Only the given strand is processed, E-value Threshold=0.05, Max Motif Count=10).

IUPAC	P-Value	E-Value	Motif (5'SS)	RNA Motif (Ray et al, 2013)	Gene
GTRAGT	1.4e-039	9.1e-035		Not found	Not found
TAYATAY	3.5e-014	2.3e-009			SHEP
TGTGTGKG	3.3e-013	2.2e-008			ARET/TDP4 3
CRCACAY	2.4e-011	1.6e-006			SM
BCGAAA	1.2e-009	7.8e-005		Not found	Not found
MTCGATT	1.6e-008	1.0e-003		Not found	Not found
GGAAAA	2.5e-007	1.6e-002		Not found	Not found
TTTRTTT	4.9e-007	3.2e-002			RBP9




\*\*Parameters used in DREME; Strand Handling=Only the given strand is processed, E-value Threshold=0.05, Max Motif Count= 10

**Table 4.3:** Motif enrichment analysis for 3'SS. 299 bps sequences were collected from intronic side of 5'SS for retained intron of length  $\geq 600$  (N= 421/458). Motif enrichment analysis was run using reshuffled sequences as background (Parameters; . Strand Handling=Only the given strand is processed, E-value Threshold=0.05, Max Motif Count=10).

IUPAC	P-Value	E-Value	Motif (5'SS)	RNA Motif (Ray et al, 2013)	Gene
TTBCAG	3.2e-018	2.1e-013		Not found	Not found
ATAYATWT	1.6e-012	1.1e-007			SHEP
ACTAATT	4.7e-008	3.0e-003			SF1
AHACAAA	5.3e-008	3.4e-003		Not found	Not found
RCAGA	1.6e-007	1.0e-002		Not found	Not found
TTCKTTTT	1.8e-007	1.1e-002		Not found	Not found
CACMCAC	4.9e-007	3.1e-002			SM
AYATATGT	6.0e-007	3.8e-002		Not found	Not found






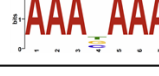

\*\*Parameter; Strand Handling= Only the given strand is processed, E-value Threshold=0.05, Max Motif Count=10.

**Table 4.4:** Motif enrichment analysis for 5'SS. 299 bps sequences were collected from intronic side of 5'SS for retained intron of length  $\geq 600$  (N= 345/413). Motif enrichment analysis was run using reshuffled sequences as background (Parameters; . Strand Handling=Only the given strand is processed, E-value Threshold=0.05, Max Motif Count=10).

IUPAC (5'SS)	P-value	E-value	Motif	RNA motif (Ray et al, 2013)	Gene
GTRAG	2.6e-011	1.6e-006		Not found	Not found
TTTTAAW	2.9e-007	1.8e-002		Not found	Not found
TATTTTMT	7.2e-007	4.4e-002		Not found	Not found

\*\*Parameter; Strand Handling= Only the given strand is processed, E-value Threshold=0.05, Max Motif Count=10.

**Table 4.5:** Motif enrichment analysis for 3'SS. 299 bps sequences were collected from intronic side of 5'SS for retained intron of length  $\geq 600$  (N=345/413). Motif enrichment analysis was run using reshuffled sequences as background (Parameters; . Strand Handling=Only the given strand is processed, E-value Threshold=0.05, Max Motif Count=10).

IUPAC (3'SS)	P-value	E-value	Motif	RNA motif (Ray et al, 2013)	Gene
TTTTYTTY	1.1e-009	6.7e-005		Not found	Not found
RGGAAR	5.8e-009	3.4e-004		Not found	Not found
CACASAS	2.9e-009	1.7e-004		Not found	Not found
YTSTGTGT	7.0e-009	4.1e-004		Not found	Not found
GTGCTGKG	1.6e-008	9.5e-004		Not found	Not found
AAABAAA	4.5e-008	2.6e-003			KHDRBS1

\*\*Parameter; Strand Handling= Only the given strand is processed, E-value Threshold=0.05, Max Motif Count=10.

**Table 4.6:** Contribution of factors in regulation of intron retention. GC frequency= GC frequency, width(Kb)= intron length (Kb), logFC= RPKM(TBI)/RPKM(Control) and ACACACA frequency= Frequency of ACACACA motif near the 5'SS and 3'SS allowing for 1 mismatch.

Parameters	Estimate	Std. Error	z value	Pr(> z )	ACACACA loc
(Intercept)	0.828783	0.046809	17.70578	3.78E-70	3'SS
GC frequency	2.804937	0.975779	2.874563	0.004046	3'SS
Width(Kb)	0.245127	0.020074	12.21138	2.70E-34	3'SS
logFC	0.40754	0.164387	2.479145	0.01317	3'SS
ACACACA frequency	9.868323	1.998088	4.938884	7.86E-07	3'SS
Width(Kb): logFC	-0.34452	0.04238	-8.12926	4.32E-16	3'SS
Width (Kb): ACACACA frequency	-2.87142	1.960563	-1.46459	0.143033	3'SS
logFC: ACACACA frequency	57.4091	26.96022	2.1294	0.033221	3'SS
Width (Kb): logFC: ACACACA frequency	20.7102	5.408959	3.828871	0.000129	3'SS
(Intercept)	0.828428	0.046727	17.72924	2.49E-70	5'SS
GC	2.967163	0.975796	3.040762	0.00236	5'SS
Width(Kb)	0.21057	0.014196	14.8332	8.94E-50	5'SS
logFC	0.558975	0.150455	3.715239	0.000203	5'SS
CA_freq_ss5	10.27271	1.675534	6.131004	8.73E-10	5'SS
Width(Kb): logFC	-0.18084	0.025237	-7.1657	7.74E-13	5'SS

## REFERENCES

- Akalin A, Kormaksson M, Li S, Garrett-Bakelman FE, Figueroa ME, Melnick A and Mason CE (2012) methylKit: a comprehensive R package for the analysis of genome-wide DNA methylation profiles. *Genome biology* **13**:R87.
- Anders S, Reyes A and Huber W (2012) Detecting differential usage of exons from RNA-seq data. *Genome research* **22**:2008-2017.
- Andersson R, Gebhard C, Miguel-Escalada I, Hoof I, Bornholdt J, Boyd M, Chen Y, Zhao X, Schmidl C, Suzuki T, Ntini E, Arner E, Valen E, Li K, Schwarzfischer L, Glatz D, Raithel J, Lilje B, Rapin N, Bagger FO, Jorgensen M, Andersen PR, Bertin N, Rackham O, Burroughs AM, Baillie JK, Ishizu Y, Shimizu Y, Furuhashi E, Maeda S, Negishi Y, Mungall CJ, Meehan TF, Lassmann T, Itoh M, Kawaji H, Kondo N, Kawai J, Lennartsson A, Daub CO, Heutink P, Hume DA, Jensen TH, Suzuki H, Hayashizaki Y, Muller F, Consortium F, Forrest AR, Carninci P, Rehli M and Sandelin A (2014) An atlas of active enhancers across human cell types and tissues. *Nature* **507**:455-461.
- Araki T and Milbrandt J (2000) Ninjurin2, a novel homophilic adhesion molecule, is expressed in mature sensory and enteric neurons and promotes neurite outgrowth. *The Journal of neuroscience : the official journal of the Society for Neuroscience* **20**:187-195.
- Aryee MJ, Jaffe AE, Corrada-Bravo H, Ladd-Acosta C, Feinberg AP, Hansen KD and Irizarry RA (2014) Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. *Bioinformatics* **30**:1363-1369.
- Ball MP, Li JB, Gao Y, Lee JH, LeProust EM, Park IH, Xie B, Daley GQ and Church GM (2009) Targeted and genome-scale strategies reveal gene-body methylation signatures in human cells. *Nature biotechnology* **27**:361-368.
- Barfield RT, Kilaru V, Smith AK and Conneely KN (2012) CpGassoc: an R function for analysis of DNA methylation microarray data. *Bioinformatics* **28**:1280-1281.
- Beinert H, Kennedy MC and Stout CD (1996) Aconitase as IronminisignSulfur Protein, Enzyme, and Iron-Regulatory Protein. *Chemical reviews* **96**:2335-2374.
- Bibikova M, Barnes B, Tsan C, Ho V, Klotzle B, Le JM, Delano D, Zhang L, Schroth GP, Gunderson KL, Fan JB and Shen R (2011) High density DNA methylation array with single CpG site resolution. *Genomics* **98**:288-295.
- Bramlett HM and Dietrich WD (2007) Progressive damage after brain and spinal cord injury: pathomechanisms and treatment strategies. *Progress in brain research* **161**:125-141.
- Braunschweig U, Barbosa-Morais NL, Pan Q, Nachman EN, Alipanahi B, Gonatopoulos-Pournatzis T, Frey B, Irimia M and Blencowe BJ (2014) Widespread intron retention in mammals functionally tunes transcriptomes. *Genome research* **24**:1774-1786.
- Carey BW, Finley LW, Cross JR, Allis CD and Thompson CB (2015) Intracellular alpha-ketoglutarate maintains the pluripotency of embryonic stem cells. *Nature* **518**:413-416.
- Cecil KM, Brubaker CJ, Adler CM, Dietrich KN, Altabe M, Egelhoff JC, Wessel S, Elangovan I, Hornung R, Jarvis K and Lanphear BP (2008) Decreased brain volume in adults with childhood lead exposure. *PLoS medicine* **5**:e112.
- Cernak I, Savic J, Malicevic Z, Zunic G, Radosevic P, Ivanovic I and Davidovic L (1996) Involvement of the central nervous system in the general response to pulmonary blast injury. *The Journal of trauma* **40**:S100-104.

- Challen GA, Sun D, Mayle A, Jeong M, Luo M, Rodriguez B, Mallaney C, Celik H, Yang L, Xia Z, Cullen S, Berg J, Zheng Y, Darlington GJ, Li W and Goodell MA (2014) Dnmt3a and Dnmt3b have overlapping and distinct functions in hematopoietic stem cells. *Cell stem cell* **15**:350-364.
- Chandler DS, Qi J and Mattox W (2003) Direct repression of splicing by transformer-2. *Molecular and cellular biology* **23**:5174-5185.
- Chang YF, Imam JS and Wilkinson MF (2007) The nonsense-mediated decay RNA surveillance pathway. *Annual review of biochemistry* **76**:51-74.
- Chen K, Deng S, Lu H, Zheng Y, Yang G, Kim D, Cao Q and Wu JQ (2013) RNA-seq characterization of spinal cord injury transcriptome in acute/subacute phases: a resource for understanding the pathology at the systems level. *PLoS one* **8**:e72567.
- Cingolani P, Cao X, Khetani RS, Chen CC, Coon M, Sammak A, Bollig-Fischer A, Land S, Huang Y, Hudson ME, Garfinkel MD, Zhong S, Robinson GE and Ruden DM (2013) Intronic non-CG DNA hydroxymethylation and alternative mRNA splicing in honey bees. *BMC genomics* **14**:666.
- Copur O and Muller J (2013) The histone H3-K27 demethylase Utx regulates HOX gene expression in Drosophila in a temporally restricted manner. *Development* **140**:3478-3485.
- Cordaux R and Batzer MA (2009) The impact of retrotransposons on human genome evolution. *Nature reviews Genetics* **10**:691-703.
- Coufal NG, Garcia-Perez JL, Peng GE, Yeo GW, Mu Y, Lovci MT, Morell M, O'Shea KS, Moran JV and Gage FH (2009) L1 retrotransposition in human neural progenitor cells. *Nature* **460**:1127-1131.
- Crona F, Dahlberg O, Lundberg LE, Larsson J and Mannervik M (2013) Gene regulation by the lysine demethylase KDM4A in Drosophila. *Developmental biology* **373**:453-463.
- Dixon CE, Clifton GL, Lighthall JW, Yaghamai AA and Hayes RL (1991) A controlled cortical impact model of traumatic brain injury in the rat. *Journal of neuroscience methods* **39**:253-262.
- Dixon CE, Lyeth BG, Povlishock JT, Findling RL, Hamm RJ, Marmarou A, Young HF and Hayes RL (1987) A fluid percussion model of experimental brain injury in the rat. *Journal of neurosurgery* **67**:110-119.
- do Carmo Pinho Franco M, Nigro D, Fortes ZB, Tostes RC, Carvalho MH, Lucas SR, Gomes GN, Coimbra TM and Gil FZ (2003) Intrauterine undernutrition--renal and vascular origin of hypertension. *Cardiovascular research* **60**:228-234.
- Dosunmu R, Alashwal H and Zawia NH (2012) Genome-wide expression and methylation profiling in the aged rodent brain due to early-life Pb exposure and its relevance to aging. *Mechanisms of ageing and development* **133**:435-443.
- Du P, Zhang X, Huang CC, Jafari N, Kibbe WA, Hou L and Lin SM (2010) Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis. *BMC Bioinformatics* **11**:587.
- Dyatlov VA and Lawrence DA (2002) Neonatal lead exposure potentiates sickness behavior induced by Listeria monocytogenes infection of mice. *Brain, behavior, and immunity* **16**:477-492.
- El Hajj N, Pliushch G, Schneider E, Dittrich M, Muller T, Korenkov M, Aretz M, Zechner U, Lehnen H and Haaf T (2013) Metabolic programming of MEST DNA methylation by intrauterine exposure to gestational diabetes mellitus. *Diabetes* **62**:1320-1328.
- Engel LS, Taioli E, Pfeiffer R, Garcia-Closas M, Marcus PM, Lan Q, Boffetta P, Vineis P, Autrup H, Bell DA, Branch RA, Brockmoller J, Daly AK, Heckbert SR, Kalina I, Kang D, Katoh T, Lafuente A, Lin HJ, Romkes M, Taylor JA and Rothman N (2002) Pooled analysis and meta-analysis of



- glutathione S-transferase M1 and bladder cancer: a HuGE review. *American journal of epidemiology* **156**:95-109.
- Falcon S and Gentleman R (2007) Using GOstats to test gene lists for GO term association. *Bioinformatics* **23**:257-258.
- Fan W and Luo J (2010) SIRT1 regulates UV-induced DNA repair through deacetylating XPA. *Molecular cell* **39**:247-258.
- Faulk C, Barks A, Liu K, Goodrich JM and Dolinoy DC (2013) Early-life lead exposure results in dose- and sex-specific effects on weight and epigenetic gene regulation in weanling mice. *Epigenomics* **5**:487-500.
- Gabriel JL, Zervos PR and Plaut GW (1986) Activity of purified NAD-specific isocitrate dehydrogenase at modulator and substrate concentrations approximating conditions in mitochondria. *Metabolism: clinical and experimental* **35**:661-667.
- Gao F, Xia Y, Wang J, Luo H, Gao Z, Han X, Zhang J, Huang X, Yao Y, Lu H, Yi N, Zhou B, Lin Z, Wen B, Zhang X, Yang H and Wang J (2013) Integrated detection of both 5-mC and 5-hmC by high-throughput tag sequencing technology highlights methylation reprogramming of bivalent genes during cellular differentiation. *Epigenetics : official journal of the DNA Methylation Society* **8**:421-430.
- Gowher H, Stockdale CJ, Goyal R, Ferreira H, Owen-Hughes T and Jeltsch A (2005) De novo methylation of nucleosomal DNA by the mammalian Dnmt1 and Dnmt3A DNA methyltransferases. *Biochemistry* **44**:9899-9904.
- Han SP, Tang YH and Smith R (2010) Functional diversity of the hnRNPs: past, present and perspectives. *The Biochemical journal* **430**:379-392.
- Hancks DC and Kazazian HH, Jr. (2016) Roles for retrotransposon insertions in human disease. *Mobile DNA* **7**:9.
- Hanna CW, Bloom MS, Robinson WP, Kim D, Parsons PJ, vom Saal FS, Taylor JA, Steuerwald AJ and Fujimoto VY (2012) DNA methylation changes in whole blood is associated with exposure to the environmental contaminants, mercury, lead, cadmium and bisphenol A, in women undergoing ovarian stimulation for IVF. *Human reproduction* **27**:1401-1410.
- Haycock PC and Ramsay M (2009) Exposure of mouse embryos to ethanol during preimplantation development: effect on DNA methylation in the h19 imprinting control region. *Biology of reproduction* **81**:618-627.
- Heijmans BT, Tobi EW, Stein AD, Putter H, Blauw GJ, Susser ES, Slagboom PE and Lumey LH (2008) Persistent epigenetic differences associated with prenatal exposure to famine in humans. *Proceedings of the National Academy of Sciences of the United States of America* **105**:17046-17049.
- Helguera P, Seiglie J, Rodriguez J, Hanna M, Helguera G and Busciglio J (2013) Adaptive downregulation of mitochondrial function in down syndrome. *Cell metabolism* **17**:132-140.
- Herz HM, Morgan M, Gao X, Jackson J, Rickels R, Swanson SK, Florens L, Washburn MP, Eissenberg JC and Shilatifard A (2014) Histone H3 lysine-to-methionine mutants as a paradigm to study chromatin signaling. *Science* **345**:1065-1070.
- Hirose Y, Tacke R and Manley JL (1999) Phosphorylated RNA polymerase II stimulates pre-mRNA splicing. *Genes & development* **13**:1234-1239.
- Hossain MB, Vahter M, Concha G and Broberg K (2012) Environmental arsenic exposure and DNA methylation of the tumor suppressor gene p16 and the DNA repair gene MLH1: effect of arsenic metabolism and genotype. *Metallomics : integrated biometal science* **4**:1167-1175.

- Houseman EA, Accomando WP, Koestler DC, Christensen BC, Marsit CJ, Nelson HH, Wiencke JK and Kelsey KT (2012) DNA methylation arrays as surrogate measures of cell mixture distribution. *BMC bioinformatics* **13**:86.
- Hsiung DT, Marsit CJ, Houseman EA, Eddy K, Furniss CS, McClean MD and Kelsey KT (2007) Global DNA methylation level in whole blood as a biomarker in head and neck squamous cell carcinoma. *Cancer epidemiology, biomarkers & prevention : a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology* **16**:108-114.
- Huang X and Dixit VM (2011) Cross talk between ubiquitination and demethylation. *Molecular and cellular biology* **31**:3682-3683.
- Hung LH, Heiner M, Hui J, Schreiner S, Benes V and Bindereif A (2008) Diverse roles of hnRNP L in mammalian mRNA processing: a combined microarray and RNAi analysis. *Rna* **14**:284-296.
- Isken O and Maquat LE (2007) Quality control of eukaryotic mRNA: safeguarding cells from abnormal mRNA function. *Genes & development* **21**:1833-1856.
- Jenkins TG and Carrell DT (2012) Dynamic alterations in the paternal epigenetic landscape following fertilization. *Frontiers in genetics* **3**:143.
- Jin P and Warren ST (2003) New insights into fragile X syndrome: from molecules to neurobehaviors. *Trends in biochemical sciences* **28**:152-158.
- Johnson WE, Li C and Rabinovic A (2007) Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* **8**:118-127.
- Jurkowska RZ, Jurkowski TP and Jeltsch A (2011) Structure and function of mammalian DNA methyltransferases. *Chembiochem : a European journal of chemical biology* **12**:206-222.
- Karimi MM, Goyal P, Maksakova IA, Bilenky M, Leung D, Tang JX, Shinkai Y, Mager DL, Jones S, Hirst M and Lorincz MC (2011) DNA methylation and SETDB1/H3K9me3 regulate predominantly distinct sets of genes, retroelements, and chimeric transcripts in mESCs. *Cell stem cell* **8**:676-687.
- Karpen GH and Spradling AC (1992) Analysis of subtelomeric heterochromatin in the Drosophila minichromosome Dp1187 by single P element insertional mutagenesis. *Genetics* **132**:737-753.
- Katzenberger RJ, Loewen CA, Wassarman DR, Petersen AJ, Ganetzky B and Wassarman DA (2013) A Drosophila model of closed head traumatic brain injury. *Proceedings of the National Academy of Sciences of the United States of America* **110**:E4152-4159.
- Keene AC and Waddell S (2007) Drosophila olfactory memory: single genes to complex neural circuits. *Nature reviews Neuroscience* **8**:341-354.
- Khodor YL, Rodriguez J, Abruzzi KC, Tang CH, Marr MT, 2nd and Rosbash M (2011) Nascent-seq indicates widespread cotranscriptional pre-mRNA splicing in Drosophila. *Genes & development* **25**:2502-2512.
- Kile ML, Houseman EA, Baccarelli AA, Quamruzzaman Q, Rahman M, Mostofa G, Cardenas A, Wright RO and Christiani DC (2014) Effect of prenatal arsenic exposure on DNA methylation and leukocyte subpopulations in cord blood. *Epigenetics : official journal of the DNA Methylation Society* **9**:774-782.
- Kim E, Du L, Bregman DB and Warren SL (1997) Splicing factors associate with hyperphosphorylated RNA polymerase II in the absence of pre-mRNA. *The Journal of cell biology* **136**:19-28.
- Klose RJ and Zhang Y (2007) Regulation of histone methylation by demethylination and demethylation. *Nature reviews Molecular cell biology* **8**:307-318.

- Koh KP, Yabuuchi A, Rao S, Huang Y, Cunniff K, Nardone J, Laiho A, Tahiliani M, Sommer CA, Mostoslavsky G, Lahesmaa R, Orkin SH, Rodig SJ, Daley GQ and Rao A (2011) Tet1 and Tet2 regulate 5-hydroxymethylcytosine production and cell lineage specification in mouse embryonic stem cells. *Cell stem cell* **8**:200-213.
- Kolasinska-Zwiercz P, Down T, Latorre I, Liu T, Liu XS and Ahringer J (2009) Differential chromatin marking of introns and expressed exons by H3K36me3. *Nature genetics* **41**:376-381.
- Koller K, Brown T, Spurgeon A and Levy L (2004) Recent developments in low-level lead exposure and intellectual impairment in children. *Environmental health perspectives* **112**:987-994.
- Kriaucionis S and Heintz N (2009) The nuclear DNA base 5-hydroxymethylcytosine is present in Purkinje neurons and the brain. *Science* **324**:929-930.
- Krueger F and Andrews SR (2011) Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. *Bioinformatics* **27**:1571-1572.
- Langlois JA, Rutland-Brown W and Wald MM (2006) The epidemiology and impact of traumatic brain injury: a brief overview. *The Journal of head trauma rehabilitation* **21**:375-378.
- Laufer BI, Kapalanga J, Castellani CA, Diehl EJ, Yan L and Singh SM (2015) Associative DNA methylation changes in children with prenatal alcohol exposure. *Epigenomics* **7**:1259-1274.
- Layalle S, Coessens E, Ghysen A and Dambly-Chaudiere C (2005) Smooth, a hnRNP encoding gene, controls axonal navigation in Drosophila. *Genes to cells : devoted to molecular & cellular mechanisms* **10**:119-125.
- Lee S, Choi I, Kang S and Rivier C (2008) Role of various neurotransmitters in mediating the long-term endocrine consequences of prenatal alcohol exposure. *Annals of the New York Academy of Sciences* **1144**:176-188.
- Lev Maor G, Yearim A and Ast G (2015) The alternative role of DNA methylation in splicing regulation. *Trends in genetics : TIG* **31**:274-280.
- Li Y, Xie C, Murphy SK, Skaar D, Nye M, Vidal AC, Cecil KM, Dietrich KN, Puga A, Jirtle RL and Hoyo C (2016) Lead Exposure during Early Human Development and DNA Methylation of Imprinted Gene Regulatory Elements in Adulthood. *Environmental health perspectives* **124**:666-673.
- Lighthall JW (1988) Controlled cortical impact: a new experimental brain injury model. *Journal of neurotrauma* **5**:1-15.
- Lin CH, Paulson A, Abmayr SM and Workman JL (2012) HP1a targets the Drosophila KDM4A demethylase to a subset of heterochromatic genes to regulate H3K36me3 levels. *PLoS one* **7**:e39758.
- Lindberg AL, Ekstrom EC, Nermell B, Rahman M, Lonnerdal B, Persson LA and Vahter M (2008) Gender and age differences in the metabolism of inorganic arsenic in a highly exposed population in Bangladesh. *Environmental research* **106**:110-120.
- Lorbeck MT, Singh N, Zervos A, Dhatta M, Lapchenko M, Yang C and Elefant F (2010) The histone demethylase Dmel\Kdm4A controls genes required for life span and male-specific sex determination in Drosophila. *Gene* **450**:8-17.
- Love MI, Huber W and Anders S (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome biology* **15**:550.
- Lu G, Xu H, Chang D, Wu Z, Yao X, Zhang S, Li Z, Bai J, Cai Q and Zhang W (2014) Arsenic exposure is associated with DNA hypermethylation of the tumor suppressor gene p16. *Journal of occupational medicine and toxicology* **9**:42.
- Luco RF, Pan Q, Tominaga K, Blencowe BJ, Pereira-Smith OM and Misteli T (2010) Regulation of alternative splicing by histone modifications. *Science* **327**:996-1000.

- Maas AI, Stocchetti N and Bullock R (2008) Moderate and severe traumatic brain injury in adults. *The Lancet Neurology* **7**:728-741.
- Marmarou A, Foda MA, van den Brink W, Campbell J, Kita H and Demetriadou K (1994) A new model of diffuse brain injury in rats. Part I: Pathophysiology and biomechanics. *Journal of neurosurgery* **80**:291-300.
- Matsui T, Leung D, Miyashita H, Maksakova IA, Miyachi H, Kimura H, Tachibana M, Lorincz MC and Shinkai Y (2010) Proviral silencing in embryonic stem cells requires the histone methyltransferase ESET. *Nature* **464**:927-931.
- Maunakea AK, Chepelev I, Cui K and Zhao K (2013) Intragenic DNA methylation modulates alternative splicing by recruiting MeCP2 to promote exon recognition. *Cell research* **23**:1256-1269.
- Mayeda A, Helfman DM and Krainer AR (1993) Modulation of exon skipping and inclusion by heterogeneous nuclear ribonucleoprotein A1 and pre-mRNA splicing factor SF2/ASF. *Molecular and cellular biology* **13**:2993-3001.
- McGhee JD and Ginder GD (1979) Specific DNA methylation sites in the vicinity of the chicken beta-globin genes. *Nature* **280**:419-420.
- Michelotti EF, Michelotti GA, Aronsohn AI and Levens D (1996) Heterogeneous nuclear ribonucleoprotein K is a transcription factor. *Molecular and cellular biology* **16**:2350-2360.
- Muller C, Bauer NM, Schafer I and White R (2013) Making myelin basic protein -from mRNA transport to localized translation. *Frontiers in cellular neuroscience* **7**:169.
- Nan X, Ng HH, Johnson CA, Laherty CD, Turner BM, Eisenman RN and Bird A (1998) Transcriptional repression by the methyl-CpG-binding protein MeCP2 involves a histone deacetylase complex. *Nature* **393**:386-389.
- O'Hagan HM, Wang W, Sen S, Destefano Shields C, Lee SS, Zhang YW, Clements EG, Cai Y, Van Neste L, Easwaran H, Casero RA, Sears CL and Baylin SB (2011) Oxidative damage targets complexes containing DNA methyltransferases, SIRT1, and polycomb members to promoter CpG Islands. *Cancer cell* **20**:606-619.
- Okano M, Bell DW, Haber DA and Li E (1999) DNA methyltransferases Dnmt3a and Dnmt3b are essential for de novo methylation and mammalian development. *Cell* **99**:247-257.
- Padilla MA, Elobeid M, Ruden DM and Allison DB (2010) An examination of the association of selected toxic metals with total and central obesity indices: NHANES 99-02. *International journal of environmental research and public health* **7**:3332-3347.
- Penalosa CG, Estevez B, Han DM, Norouzi M, Lockshin RA and Zakeri Z (2014) Sex-dependent regulation of cytochrome P450 family members Cyp1a1, Cyp2e1, and Cyp7b1 by methylation of DNA. *FASEB journal : official publication of the Federation of American Societies for Experimental Biology* **28**:966-977.
- Petronis KR and Anthony JC (2003) Social epidemiology, intra-neighbourhood correlation, and generalised estimating equations. *Journal of epidemiology and community health* **57**:914; author reply 914.
- Pilsner JR, Hall MN, Liu X, Ilievski V, Slavkovich V, Levy D, Factor-Litvak P, Yunus M, Rahman M, Graziano JH and Gamble MV (2012) Influence of prenatal arsenic exposure and newborn sex on global methylation of cord blood DNA. *PLoS ONE* **7**:e37147.
- Pilsner JR, Hu H, Ettinger A, Sanchez BN, Wright RO, Cantonwine D, Lazarus A, Lamadrid-Figueroa H, Mercado-Garcia A, Tellez-Rojo MM and Hernandez-Avila M (2009) Influence of prenatal lead exposure on genomic methylation of cord blood DNA. *Environmental health perspectives* **117**:1466-1471.

- Povlishock JT and Christman CW (1995) The pathobiology of traumatically induced axonal injury in animals and humans: a review of current thoughts. *Journal of neurotrauma* **12**:555-564.
- Pradeepa MM, Sutherland HG, Ule J, Grimes GR and Bickmore WA (2012) Psip1/Ledgf p52 binds methylated histone H3K36 and splicing factors and contributes to the regulation of alternative splicing. *PLoS genetics* **8**:e1002717.
- Rauch J, O'Neill E, Mack B, Matthias C, Munz M, Kolch W and Gires O (2010) Heterogeneous nuclear ribonucleoprotein H blocks MST2-mediated apoptosis in cancer cells by regulating A-Raf transcription. *Cancer research* **70**:1679-1688.
- Ray D, Kazan H, Cook KB, Weirauch MT, Najafabadi HS, Li X, Gueroussov S, Albu M, Zheng H, Yang A, Na H, Irimia M, Matzat LH, Dale RK, Smith SA, Yarosh CA, Kelly SM, Nabet B, Mecnas D, Li W, Laishram RS, Qiao M, Lipshitz HD, Piano F, Corbett AH, Carstens RP, Frey BJ, Anderson RA, Lynch KW, Penalva LO, Lei EP, Fraser AG, Blencowe BJ, Morris QD and Hughes TR (2013) A compendium of RNA-binding motifs for decoding gene regulation. *Nature* **499**:172-177.
- Sandoval J, Heyn H, Moran S, Serra-Musach J, Pujana MA, Bibikova M and Esteller M (2011) Validation of a DNA methylation microarray for 450,000 CpG sites in the human genome. *Epigenetics : official journal of the DNA Methylation Society* **6**:692-702.
- Santiago M, Antunes C, Guedes M, Sousa N and Marques CJ (2014) TET enzymes and DNA hydroxymethylation in neural development and function - how critical are they? *Genomics* **104**:334-340.
- Schneider JS, Anderson DW, Talsania K, Mettil W and Vadigepalli R (2012) Effects of developmental lead exposure on the hippocampal transcriptome: influences of sex, developmental period, and lead exposure level. *Toxicological sciences : an official journal of the Society of Toxicology* **129**:108-125.
- Sen A, Cingolani P, Senut MC, Land S, Mercado-Garcia A, Tellez-Rojo MM, Baccarelli AA, Wright RO and Ruden DM (2015a) Lead exposure induces changes in 5-hydroxymethylcytosine clusters in CpG islands in human embryonic stem cells and umbilical cord blood. *Epigenetics : official journal of the DNA Methylation Society*:0.
- Sen A, Heredia N, Senut MC, Hess M, Land S, Qu W, Hollacher K, Dereski MO and Ruden DM (2015b) Early life lead exposure causes gender-specific changes in the DNA methylation profile of DNA extracted from dried blood spots. *Epigenomics* **7**:379-393.
- Sen A, Heredia N, Senut MC, Land S, Hollacher K, Lu X, Dereski MO and Ruden DM (2015c) Multigenerational epigenetic inheritance in humans: DNA methylation changes associated with maternal exposure to lead can be transmitted to the grandchildren. *Sci Rep* **5**:14466.
- Senut MC, Sen A, Cingolani P, Shaik A, Land SJ and Ruden DM (2014) Lead exposure disrupts global DNA methylation in human embryonic stem cells and alters their neuronal differentiation. *Toxicological sciences : an official journal of the Society of Toxicology* **139**:142-161.
- Shen L, Inoue A, He J, Liu Y, Lu F and Zhang Y (2014a) Tet3 and DNA replication mediate demethylation of both the maternal and paternal genomes in mouse zygotes. *Cell stem cell* **15**:459-470.
- Shen W, Wang C, Xia L, Fan C, Dong H, Deckelbaum RJ and Qi K (2014b) Epigenetic modification of the leptin promoter in diet-induced obese mice and the effects of N-3 polyunsaturated fatty acids. *Scientific reports* **4**:5282.
- Singer T, McConnell MJ, Marchetto MC, Coufal NG and Gage FH (2010) LINE-1 retrotransposons: mediators of somatic variation in neuronal genomes? *Trends in neurosciences* **33**:345-354.
- Sirivarasai J, Kaojarern S, Chanprasertyothin S, Panpunuan P, Petchpoung K, Tatsaneeyapant A, Yoovathaworn K, Sura T, Kaojarern S and Sritara P (2015) Environmental lead exposure,

- catalase gene, and markers of antioxidant and oxidative stress relation to hypertension: an analysis based on the EGAT study. *BioMed research international* **2015**:856319.
- Skinner MK, Haque CG, Nilsson E, Bhandari R and McCarrey JR (2013) Environmentally induced transgenerational epigenetic reprogramming of primordial germ cells and the subsequent germ line. *PLoS one* **8**:e66318.
- Sofer T, Schifano ED, Hoppin JA, Hou L and Baccarelli AA (2013) A-clustering: a novel method for the detection of co-regulated methylation regions, and regions associated with exposure. *Bioinformatics* **29**:2884-2891.
- Steinmann-Zwicky M (1994) Sex determination of the Drosophila germ line: tra and dsx control somatic inductive signals. *Development* **120**:707-716.
- Stroud H, Feng S, Morey Kinney S, Pradhan S and Jacobsen SE (2011) 5-Hydroxymethylcytosine is associated with enhancers and gene bodies in human embryonic stem cells. *Genome biology* **12**:R54.
- Sun W, Zang L, Shu Q and Li X (2014) From development to diseases: the role of 5hmC in brain. *Genomics* **104**:347-351.
- Szulwach KE, Li X, Li Y, Song CX, Han JW, Kim S, Namburi S, Hermetz K, Kim JJ, Rudd MK, Yoon YS, Ren B, He C and Jin P (2011) Integrating 5-hydroxymethylcytosine into the epigenomic landscape of human embryonic stem cells. *PLoS genetics* **7**:e1002154.
- Szwagierczak A, Brachmann A, Schmidt CS, Bultmann S, Leonhardt H and Spada F (2011) Characterization of PvuRts1I endonuclease as a tool to investigate genomic 5-hydroxymethylcytosine. *Nucleic acids research* **39**:5149-5156.
- Szwagierczak A, Bultmann S, Schmidt CS, Spada F and Leonhardt H (2010) Sensitive enzymatic quantification of 5-hydroxymethylcytosine in genomic DNA. *Nucleic acids research* **38**:e181.
- Tahiliani M, Koh KP, Shen Y, Pastor WA, Bandukwala H, Brudno Y, Agarwal S, Iyer LM, Liu DR, Aravind L and Rao A (2009) Conversion of 5-methylcytosine to 5-hydroxymethylcytosine in mammalian DNA by MLL partner TET1. *Science* **324**:930-935.
- Tellez-Plaza M, Tang WY, Shang Y, Umans JG, Francesconi KA, Goessler W, Ledesma M, Leon M, Laclaustra M, Pollak J, Guallar E, Cole SA, Fallin MD and Navas-Acien A (2014) Association of global DNA methylation and global DNA hydroxymethylation with metals and other exposures in human blood DNA samples. *Environmental health perspectives* **122**:946-954.
- Teschendorff AE, Marabita F, Lechner M, Bartlett T, Tegner J, Gomez-Cabrero D and Beck S (2013) A beta-mixture quantile normalization method for correcting probe design bias in Illumina Infinium 450 k DNA methylation data. *Bioinformatics* **29**:189-196.
- Teschler S, Bartkuhn M, Kunzel N, Schmidt C, Kiehl S, Dammann G and Dammann R (2013) Aberrant methylation of gene associated CpG sites occurs in borderline personality disorder. *PLoS one* **8**:e84180.
- Tian Q, Li T, Hou W, Zheng J, Schrum LW and Bonkovsky HL (2011) Lon peptidase 1 (LONP1)-dependent breakdown of mitochondrial 5-aminolevulinic acid synthase protein by heme in human liver cells. *The Journal of biological chemistry* **286**:26424-26430.
- Touleimat N and Tost J (2012) Complete pipeline for Infinium((R)) Human Methylation 450K BeadChip data processing using subset quantile normalization for accurate DNA methylation estimation. *Epigenomics* **4**:325-341.
- Tsurumi A, Dutta P, Shang R, Yan SJ and Li WX (2013) Drosophila Kdm4 demethylases in histone H3 lysine 9 demethylation and ecdysteroid signaling. *Sci Rep* **3**:2894.

- Turk PW, Laayoun A, Smith SS and Weitzman SA (1995) DNA adduct 8-hydroxyl-2'-deoxyguanosine (8-hydroxyguanine) affects function of human DNA methyltransferase. *Carcinogenesis* **16**:1253-1255.
- Wadhwa PD, Buss C, Entringer S and Swanson JM (2009) Developmental origins of health and disease: brief history of the approach and current focus on epigenetic mechanisms. *Seminars in reproductive medicine* **27**:358-368.
- Wang CQ, Motoda L, Satake M, Ito Y, Taniuchi I, Tergaonkar V and Osato M (2013) Runx3 deficiency results in myeloproliferative disorder in aged mice. *Blood* **122**:562-566.
- Wang RH, Sengupta K, Li C, Kim HS, Cao L, Xiao C, Kim S, Xu X, Zheng Y, Chilton B, Jia R, Zheng ZM, Appella E, Wang XW, Ried T and Deng CX (2008) Impaired DNA damage response, genome instability, and tumorigenesis in SIRT1 mutant mice. *Cancer cell* **14**:312-323.
- Weitzman SA, Turk PW, Milkowski DH and Kozlowski K (1994) Free radical adducts induce alterations in DNA cytosine methylation. *Proceedings of the National Academy of Sciences of the United States of America* **91**:1261-1264.
- Wen L, Li X, Yan L, Tan Y, Li R, Zhao Y, Wang Y, Xie J, Zhang Y, Song C, Yu M, Liu X, Zhu P, Li X, Hou Y, Guo H, Wu X, He C, Li R, Tang F and Qiao J (2014) Whole-genome analysis of 5-hydroxymethylcytosine and 5-methylcytosine at base resolution in the human brain. *Genome biology* **15**:R49.
- Wessely F and Emes RD (2012) Identification of DNA methylation biomarkers from Infinium arrays. *Frontiers in genetics* **3**:161.
- Wickramasinghe VO, Gonzalez-Porta M, Perera D, Bartolozzi AR, Sibley CR, Hallegger M, Ule J, Marioni JC and Venkitaraman AR (2015) Regulation of constitutive and alternative mRNA splicing across the human transcriptome by PRPF8 is determined by 5' splice site strength. *Genome biology* **16**:201.
- Wolff EM, Liang G, Cortez CC, Tsai YC, Castelao JE, Cortessis VK, Tsao-Wei DD, Groshen S and Jones PA (2008) RUNX3 methylation reveals that bladder tumors are older in patients with a history of smoking. *Cancer research* **68**:6208-6214.
- Wolff GL, Kodell RL, Moore SR and Cooney CA (1998) Maternal epigenetics and methyl supplements affect agouti gene expression in Avy/a mice. *FASEB journal : official publication of the Federation of American Societies for Experimental Biology* **12**:949-957.
- Wright RO, Schwartz J, Wright RJ, Bollati V, Tarantini L, Park SK, Hu H, Sparrow D, Vokonas P and Baccarelli A (2010) Biomarkers of lead exposure and DNA methylation within retrotransposons. *Environmental health perspectives* **118**:790-795.
- Wu H, D'Alessio AC, Ito S, Wang Z, Cui K, Zhao K, Sun YE and Zhang Y (2011a) Genome-wide analysis of 5-hydroxymethylcytosine distribution reveals its dual function in transcriptional regulation in mouse embryonic stem cells. *Genes & development* **25**:679-684.
- Wu H, D'Alessio AC, Ito S, Xia K, Wang Z, Cui K, Zhao K, Sun YE and Zhang Y (2011b) Dual functions of Tet1 in transcriptional regulation in mouse embryonic stem cells. *Nature* **473**:389-393.
- Yamaguchi S, Shen L, Liu Y, Sandler D and Zhang Y (2013) Role of Tet1 in erasure of genomic imprinting. *Nature* **504**:460-464.
- Yamamoto H, Kokame K, Okuda T, Nakajo Y, Yanamoto H and Miyata T (2011) NDRG4 protein-deficient mice exhibit spatial learning deficits and vulnerabilities to cerebral ischemia. *J Biol Chem* **286**:26158-26165.
- Yang X, Han H, De Carvalho DD, Lay FD, Jones PA and Liang G (2014) Gene body methylation can alter gene expression and is a therapeutic target in cancer. *Cancer cell* **26**:577-590.

- Yearim A, Gelfman S, Shayevitch R, Melcer S, Glaich O, Mallm JP, Nissim-Rafinia M, Cohen AH, Rippe K, Meshorer E and Ast G (2015) HP1 is involved in regulating the global impact of DNA methylation on alternative splicing. *Cell reports* **10**:1122-1134.
- Yen KE and Schenkein DP (2012) Cancer-associated isocitrate dehydrogenase mutations. *The oncologist* **17**:5-8.
- Yeo G and Burge CB (2004) Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. *Journal of computational biology : a journal of computational molecular cell biology* **11**:377-394.
- Yoh SM, Lucas JS and Jones KA (2008) The Iws1:Spt6:CTD complex controls cotranscriptional mRNA biosynthesis and HYPB/Setd2-mediated histone H3K36 methylation. *Genes & development* **22**:3422-3434.
- Youssef SA, El-Sanousi AA, Afifi NA and El Brawy AM (1996) Effect of subclinical lead toxicity on the immune response of chickens to Newcastle disease virus vaccine. *Research in veterinary science* **60**:13-16.
- Yu M, Hon GC, Szulwach KE, Song CX, Zhang L, Kim A, Li X, Dai Q, Shen Y, Park B, Min JH, Jin P, Ren B and He C (2012) Base-resolution analysis of 5-hydroxymethylcytosine in the mammalian genome. *Cell* **149**:1368-1380.
- Zhang D, Spielmann A, Wang L, Ding G, Huang F, Gu Q and Schwarz W (2012) Mast-cell degranulation induced by physical stimuli involves the activation of transient-receptor-potential channel TRPV2. *Physiological research / Academia Scientiarum Bohemoslovaca* **61**:113-124.
- Zhang D, Yoon HG and Wong J (2005) JMJD2A is a novel N-CoR-interacting protein and is involved in repression of the human transcription factor achaete scute-like homologue 2 (ASCL2/Hash2). *Molecular and cellular biology* **25**:6404-6414.
- Zhang J, Mu X, Xu W, Martin FL, Alamdar A, Liu L, Tian M, Huang Q and Shen H (2014) Exposure to arsenic via drinking water induces 5-hydroxymethylcytosine alteration in rat. *The Science of the total environment* **497-498C**:618-625.
- Zhou HL, Luo G, Wise JA and Lou H (2014) Regulation of alternative splicing by local histone modifications: potential roles for RNA-guided mechanisms. *Nucleic acids research* **42**:701-713.



**ABSTRACT****NEURONAL INSULT EITHER BY EXPOSURE TO LEAD OR BY DIRECT NEURONAL DAMAGE CAUSE GENOME-WIDE CHANGES IN DNA METHYLATION AND HISTONE 3 LYSINE 36 TRI-METHYLATION.**

by

**ARKO SEN****December 2016****Advisor:** Dr. Douglas Ruden**Major:** Pharmacology**Degree:** Doctor of Philosophy

Prenatal and postnatal exposure to pervasive neuro-toxicants such as Lead (Pb) has been reported to causes extensive and diverse changes in the epigenetic profile. Among epigenetic modification, DNA methylation (5mC) is perhaps the most widely studied and has been proposed to be potential early biomarkers for Pb toxicity. Several studies have demonstrated the association between Pb-exposure and 5mC. However most of these studies are restricted to looking at a specific set of target genes or repetitive elements. Therefore, one of the main objectives of our study was to use an unbiased genome-wide approach to look at Pb-exposure associated changes in 5mC. To this end, we used the Human methylation 450K (HM450K) high density array to quantitatively measure the Pb-associated 5mC changes. The sample for this study consisted of DNA extracted from neonatal and current blood spots from a mother-infant cohort in Detroit, USA and Umbilical cord blood DNA from a mother-infant cohort from Mexico City, Mexico. We observed that Pb-exposure associated 5mC changes in whole blood and UCB are sex-specific. Furthermore, some of these 5mC changes are heritable and can be transmitted from the grandmother to the grandchildren. To further our

understanding of the relationship between Pb-exposure and 5mC, we wanted to look at the impact of Pb-exposure on DNA demethylation, specifically the dynamic changes in 5-hydroxymethylcytosine (5hmC) profile. To study these changes in the 5hmC profile, we used a novel modification of the HM450K, which we named HMeDIP-450K array. Using the HMeDIP-450K array we demonstrated that 5hmC showed a much larger number of sex-independent changes. Interestingly, a vast majority Pb-dependent 5mC and 5hmC clusters mapped to either gene implicated in neurodegeneration and regulation of mitochondrial processes such as NINJ2, VAMP5, GSTM1, GSTM5 etc. 5mC and 5hmC are potent regulators of gene expression and their dysregulation can cause widespread changes in the transcriptome and may contribute to neurodegenerative phenotype. Besides 5mC and 5hmC, transcriptomic changes can also be regulated by dynamic changes in histone methylation profile and alternative splicing. To study these changes, especially in context of neurodegeneration we used a *Drosophila* model of traumatic brain injury (TBI). Using a modified version of this model, we subjected *w<sup>1118</sup>* fruit flies to mild closed head trauma. To determine the transcriptomic changes which contribute to survival post TBI, we collected fly heads from the survivors at 2 time points; 4 hours and 24 hours' post-trauma. Mild TBI using our modified TBI protocol had limited impact on the expression profile of genes but showed large perturbations in alternative splicing (AS) regulation 24 hours' post-trauma. Classification of these AS changes showed selective retention of long introns (>81bps). Some of these genes also showed a significant reduction in transcript abundance and were specifically involved in mitochondrial metabolism. The retained introns were enriched for CA-rich motifs known to bind to Smooth (SM), an hnRNPL class of splicing factor. Mutating SM (*sm<sup>4</sup>/DF*) resulted in reversal of intron retention observed 24 hours' post-

trauma. This observation suggested that SM is critical regulator of Intron retention in fly heads. Interestingly, ChIP-sequencing for H3K36me3 revealed increased levels in retained introns post-trauma. Additionally, higher H3k36me3 was also observed around intronic SM-binding motifs post-trauma which suggested that increased level of H3k36me3 might be recruiting SM to their Intronic Splicing Suppressor sites and cause RI in the *Drosophila* model of TBI. Together our studies in human cohort and *Drosophila* sheds some light on the complex multi-layered mechanism regulating gene-expression especially under neurotoxic and neurodegenerative conditions.

## **AUTOBIOGRAPHICAL STATEMENT**

Complex interaction between existing environmental condition and inherent biological processes especially during development underlie the susceptibility of the individual to diseases in later life. Understanding this interaction especially in context of neuronal development has been the central theme and focus of my studies. During my tenure as a Graduate student at Wayne State University I have attempted to address this problem using diverse experimental set-ups which has allowed me to develop versatility, adaptability and an ability to apply the knowledge acquired from previous experience to broader biological questions. Besides empirical knowledge my tenure has also afforded me the opportunity for personal improvement. The countless number of seminars and meetings I have attended and presented has helped me to develop my ability to convey my thought in a succinct and comprehensive format and helped me improve upon my public speaking and teaching skills. Collaboration and continuous interaction with my peers in the scientific community at Wayne State University has helped me develop a spirit of teamwork and cooperation which has been instrumental in completion of my project.

In future, I will continue to work in the field of pharmacology and toxicology as a Post-Graduate researcher and one day, hope to become a member of the academic community. Scientific research has its origins in a very fundamental human character - curiosity. I wish to be one to add fuel to this flame called curiosity. My long term goals include contributing to the development of new concepts in toxicogenomic and pharmacology and imparting my knowledge and experience to the next generation of researchers and scientist.