

**Bioinformatic analysis of epigenetic effects, particularly in DNA  
methylation, following different interventions**

**Sara-Jayne Thursby**

BSc (Hons) Biology

*Research conducted within*

Genomic Medicine Research Group

School of Biomedical Sciences

Faculty of Life and Health Sciences

Ulster University



*A thesis submitted for the degree of*

Doctor of Philosophy

January 2020

I confirm that the word count of this thesis is less than 100,000 words

## Acknowledgements

Thank you to my supervisor Professor Colum Walsh, for your supervision, guidance and the opportunities you have given me throughout my PhD. I would also like to express my gratitude to Dr Rachelle Irwin, for answering all my queries and supporting me throughout my research. In addition to other members of the Walsh Lab & officemates, Gareth Pollin, Catherine Scullion, Miroslava Ondicova, Catherine McBride and Zoe Angel for entertaining chats, night outs and keeping morale up throughout my PhD – it would have been so much more difficult without you guys. Thank you.

## Note to access to contents

I hereby declare that with effect from the date on which this thesis is deposited in the Research Office of Ulster University, I permit;

1. The Librarian of the University to allow the thesis to be copied in whole or in part without reference to me on the understanding that such authority applies to the provision of single copies made for study purposes or for the inclusion within the stock of another library.
2. The thesis to be made available through the Ulster Institutional Repository and/or EThOS under the terms of the Ulster eTheses Deposit Agreement which I have signed.

IT IS A CONDITION OF USE OF THIS THESIS THAT ANYONE WHO CONSULTS IT MUST RECOGNISE THAT THE COPYRIGHT RESTS WITH THE UNIVERSITY AND THEN SUBSEQUENTLY TO THE AUTHOR AND THAT NO QUOTATION FROM THE THESIS AND NO INFORMATION DERIVED FROM IT MAY BE PUBLISHED UNLESS THE SOURCE IS PROPERLY ACKNOWLEDGED.

## Abstract

Epigenetics is defined as heritable changes in gene expression without a change in the underlying DNA sequence. In this thesis I concentrate on DNA methylation and the changes that occur in response to different conditions; more particularly, I develop methods to analyse methylation data and associated transcriptional and chromatin changes and apply this to four different projects.

The first project focused on the effects of shRNA mediated DNMT1 depletion within immortalised human fibroblasts. Here we found four key classes of genes dependent on DNA methylation; protocadherins, genes involved in fat homeostasis, olfactory receptors and cancer testis antigens. In addition to an interplay with polycomb repressive complexes at certain loci. Within this project, I developed tools to examine complex loci and correlate methylation with chromatin marks.

In the second project, we sought to carry out a similar experiment, but this time investigated the effects of UHRF1 depletion within the same cell line, as UHRF1 is known to recruit DNMT1 to hemi-methylated DNA. Here we found depletion of UHRF1 caused demethylation and upregulation of endogenous retroviruses and a subsequent innate immune response. When the cells were rescued methylation did not recover but the innate immune response and expression of retroviral elements was attenuated. However, rescued cells were hypersensitive to SETDB1 and KAP1 inhibition, implicating H3K9me3 in the UHRF1-mediated repression in absence of DNA methylation. UHRF1 cell lines which were mutated to affect the H3K9me3 binding domain could not repress endogenous retroviral expression, confirming the involvement of H3K9me3 here. Here, I aided in the analysis of methylation array data for knockdown, rescue and mutant cell lines and developed a tool to

analyse repeat elements covered by the Illumina Human Methylation 450k BeadChip and MethylationEPIC arrays.

In the third project, we sought to investigate the effects of folic acid supplementation in the second and third trimester on the methylation of the offspring. Folate is a limiting factor of one carbon metabolism and as a result, DNA synthesis and DNA methylation. Following intervention, cord blood was examined using the EPIC array and we discovered a folate sensitive differentially methylated region upstream of the imprint regulator ZFP57 and verified the change in an independent cohort and within in vitro models. In this project, I helped to develop statistical models with the initial and downstream bioinformatic analysis of methylation arrays and refined a tool for the investigation of target loci from methylation array data.

In project 4, we investigated the effects of mental illness on the methylation patterns of first year university students. We observed enrichment for genes involved in the immune response and the inflammatory skin condition psoriasis, with notable hypermethylation at the late cornified envelope gene cluster involved in skin cell differentiation. Results were confirmed via wet lab approaches and validated in part in an independent cohort, adding an immune component to the aetiology of depression. In this study, I aided with the initial and downstream bioinformatic analysis of methylation arrays, including taking advantage of their ability to score copy number variation.

Finally, in project 5, I formalised the tools I had used in project 1-4 into a complete pipeline called CandiMeth (available at [www.bit.do/candimeth](http://www.bit.do/candimeth)) which can be used by people with little bioinformatics training to investigate DNA methylation at candidate genomic features. This pipeline is user-friendly, has no installation requirements and runs freely off the Galaxy framework ([www.usegalaxy.org](http://www.usegalaxy.org)) to allow users to reproducibly quantify and visualise

methylation differences among their samples and how these results correlate with different genomic features, such as repetitive elements.

Overall, in this thesis I have developed novel approaches to analysing methylation data and applied these to a range of projects, culminating in the development of a user-friendly methylation array analysis tool called CandiMeth.

## Abbreviations

ABBREVIATION	DEFINITION
<b>10FTHF</b>	10-formyltetrahydrofolate
<b>2D</b>	2-Dimensional
<b>450K ARRAY</b>	Infinium Human Methylation 450K BeadChip array
<b>5'-AZA-DC</b>	5-aza-2'-deoxycytidine
<b>5HMC</b>	5-hydroxymethylcytosine
<b>5MC</b>	5-methylcytosine
<b>5-MTHF</b>	5-methyltetrahydrofolate
<b>AFAST</b>	Aberdeen folic acid supplementation trial
<b>AML</b>	Acute myeloid leukemia
<b>ATAC-SEQ</b>	Assay for Transposase Accessible Chromatin sequencing
<b>ATP</b>	Adenosine triphosphate
<b>BED</b>	Browser Extensible Data
<b>BLAST</b>	Basic Like Alignment Search Tool
<b>BLAT</b>	Basic like alignment tool
<b>BMI</b>	Body mass index
<b>BMIQ</b>	Beta Mixture Quantile normalisation
<b>BWA</b>	Burrows Wheeler Aligner
<b>BWT</b>	Burrows Wheeler Transformation
<b>cDNA</b>	Complementary DNA
<b>CGI</b>	CpG Island
<b>CHAMP</b>	Chip analysis methylation pipeline
<b>CHIP-SEQ</b>	Chromatin Immunoprecipitation sequencing
<b>CHMM</b>	Chromatin state segmentation via Hidden Markov Model
<b>CKO</b>	Conditional knockout
<b>CNS</b>	Central nervous system
<b>CNV</b>	Copy number variation
<b>CpG</b>	Cytosine-phosphate-Guanine
<b>CRISPR</b>	Clustered Regularly Interspaced Short Palindromic Repeats
<b>CSV</b>	Comma separated variable file
<b>CTA</b>	Cancer testis antigen
<b>CTCF</b>	CCCTC-binding factor
<b>CYS</b>	Cysteine
<b>DAVID</b>	The Database for Annotation, Visualization and Integrated Discovery
<b>DDR</b>	DNA damage response
<b>DMG</b>	Dimethylglycine
<b>DMP</b>	Differentially methylated probes
<b>DMR</b>	Differentially methylated region
<b>DNA</b>	Deoxyribose nucleic acid
<b>DNMT</b>	DNA methyltransferase
<b>DNMT1</b>	DNA methyltransferase 1
<b>DNMT2</b>	DNA methyltransferase 2
<b>DNMT3A</b>	DNA methyltransferase 3A
<b>DNMT3B</b>	DNA methyltransferase 3B
<b>DNMT3L</b>	DNA methyltransferase 3L
<b>DNMTi</b>	DNA methyltransferase inhibition
<b>dsDNA</b>	Double stranded DNA

<b>dsRNA</b>	Double stranded RNA
<b>dUTP</b>	Deoxyuridine Triphosphate
<b>EBI SRA</b>	European Nucleotide Archive Sequence Read Archive
<b>EDC</b>	Epidermal differentiation complex
<b>ENCODE</b>	Encyclopedia of DNA elements
<b>EPIC ARRAY</b>	Infinium MethylationEPIC array
<b>EPIFASSTT</b>	Epigenetic effects on children's psychosocial development in a randomised trial of folic acid supplementation in second and third trimester
<b>ERV</b>	Endogenous Retrovirus
<b>ESC</b>	Embryonic stem cell
<b>EZH2</b>	Enhancer of Zeste 2
<b>FA</b>	Folic acid
<b>FANTOM5</b>	Functional ANnoTation Of the Mammalian genome
<b>FASSTT</b>	Folate acid supplementation in the second and third trimester
<b>FASTQC</b>	FASTQ quality control
<b>FBM</b>	Fat and body mass
<b>FPKM</b>	Fragments per kilobase of transcript per million mapped reads
<b>FTP</b>	File transfer protocol
<b>GB</b>	Gene body
<b>GO</b>	Gene Ontology
<b>GR</b>	Glucocorticoid receptors
<b>GSH</b>	Glutathione
<b>GWAS</b>	Genome wide association study
<b>HCP</b>	High CG content promoter
<b>HDAC</b>	Histone deacetylases
<b>HDM</b>	Histone demethylases
<b>HLA</b>	Human leukocyte antigen
<b>HMT</b>	Histone methyltransferases
<b>HOMOCYS</b>	Homocysteine
<b>HPA</b>	Hypothalamic pituitary adrenal
<b>HTERT</b>	Human telomerase reverse transcriptase
<b>HTML</b>	Hyper-text markup language
<b>IAP</b>	Intracisternal alpha particles
<b>IBMS</b>	Institute of biomedical sciences
<b>ICP</b>	Intermediate CG content promoter
<b>ICR</b>	Imprint control region
<b>IDATS</b>	Intensity Data Files
<b>IDE</b>	Interactive Development Environment
<b>IFN</b>	Interferon
<b>IPS</b>	Induced polypotent stem cells
<b>ISG</b>	Interferon stimulating genes
<b>KAT</b>	Histone acetyltransferases
<b>KD</b>	Knockdown
<b>KEGG</b>	Kyoto Encyclopedia of Genes and Genomes
<b>KO</b>	Knockout
<b>LCP</b>	Low CG content promoter
<b>LIMMA</b>	Linear models for microarray analysis
<b>LINE</b>	Long INterspersed Elements
<b>LTR</b>	Long terminal repeat



<b>MCPG</b>	methyated CpG
<b>MCPH</b>	methyated CpH
<b>MDS</b>	Myelodysplastic syndromes
<b>MDS</b>	Multi-dimensional scaling
<b>MESC</b>	Mouse ESC
<b>MET</b>	Methionine
<b>MHC</b>	Major histocompatibility cluster
<b>MNASE</b>	Micrococcal nuclease
<b>MR</b>	Mineralcorticoid receptors
<b>MRNA</b>	Messenger RNA
<b>MS</b>	Manuscript
<b>MTHFR</b>	Methylenetetrahydrofolate reductase
<b>NCBI</b>	National Center for Biotechnology Information
<b>NCRNA</b>	Non-coding RNA
<b>NGS</b>	Next generation sequencing
<b>NON-LTR</b>	Non-long terminal repeat
<b>NTD</b>	Neural tube defects
<b>OR</b>	Olfactory receptor
<b>OS</b>	Operating system
<b>PCA</b>	Principal component analysis
<b>PCDH</b>	Protocadherin
<b>PCNA</b>	Proliferating cell nuclear antigen
<b>PCR</b>	Polymerase Chain Reaction
<b>PHD</b>	Plant Homeodomain
<b>piRNA</b>	PIWI-interacting RNA
<b>PRC1</b>	Protein regulator of cytokinesis 1
<b>PRC2</b>	Protein regulator of cytokinesis 2
<b>pre-mRNA</b>	Pre-posttranscriptional modification mRNA
<b>RCSB PDB</b>	Research collaboratory for structural bioinformatics protein database
<b>RCT</b>	Randomised control trial
<b>RHS</b>	Right hand side
<b>RNA</b>	Ribonucleic acid
<b>RNAI</b>	RNA interference
<b>RNA-SEQ</b>	RNA-sequencing
<b>RPKM</b>	Reads per kilobase per million mapped reads
<b>RRBS</b>	Reduced representation bisulphite sequencing
<b>RRNA</b>	Ribosomal RNA
<b>RT-qPCR</b>	Quantitative reverse transcription polymerase chain reaction
<b>SAH</b>	S-adenosylhomocysteine
<b>SAM</b>	Sequence Alignment Map
<b>SAM</b>	S-adenosylmethionine
<b>SNP</b>	Single Nucleotide Polymorphism
<b>SRA</b>	Set and RING finger associated domain
<b>SUZ12</b>	Suppressor of zeste 12
<b>SVA</b>	Surrogate variable analysis
<b>SWAN</b>	Subset quantile Within Array Normalisation
<b>TE</b>	Transposable element
<b>TF</b>	Transcription factors
<b>THF</b>	Tetrahydrofolate

<b>TPM</b>	Transcripts per million mapped reads
<b>TRNA</b>	Transfer RNA
<b>TSS</b>	Transcription Start Site
<b>TTD</b>	Tandem tudor domain
<b>UCSC</b>	University of California Santa Cruz
<b>UHRF1</b>	Ubiquitin Like With PHD And Ring Finger Domains 1
<b>WGBS</b>	Whole Genome Bisulphite Sequencing
<b>WHO</b>	World health organization
<b>WT</b>	Wildtype
<b>β</b>	Beta Methylation

## Table of Contents

Acknowledgements.....	2
Note to access to contents.....	3
Abstract.....	4
Abbreviations.....	7
1.0 General Introduction.....	14
1.1 DNA packaging into chromatin .....	14
1.2 Histone Marks .....	16
1.3 Epigenetics and DNA Methylation .....	19
1.4 Distribution of DNA Methylation .....	19
1.4.1 Methylation at CpG Islands.....	19
1.4.2 Non-CpG Methylation .....	22
1.4.3 Gene Body Methylation .....	24
1.4.4 Methylation at Enhancers.....	28
1.4.5 Transposable Elements .....	31
1.4.6 Copy Number Variation .....	33
1.4.7 5-hydroxymethylcytosine .....	34
1.5 DNA Methylation Assessment Methods .....	37
1.5.1 Gene Specific.....	37
1.5.2 Array-Based.....	41
1.5.3 Sequence-Based.....	44
1.6 Bioinformatic Analysis.....	52
1.6.1 R .....	52
1.6.2 Array-Based Processing Methods .....	53
1.6.3 Sequence-Based Processing Methods .....	62
1.6.4 UCSC Genome Browser.....	72
1.6.5 Galaxy Bioinformatics Interface.....	73
1.6.6 Database for Annotation, Visualization and Integrated Discovery.....	73
1.7 Mechanistic Studies .....	74
1.7.1 Cellular Machinery .....	74
1.7.2 <i>DNMT1</i> .....	75
1.7.3 <i>UHRF1</i> .....	76
1.7.4 <i>Hypomorphic States</i> .....	76
1.7.5 <i>Interaction with Polycomb</i> .....	78
1.8 Epidemiology Applications.....	79

1.8.1 Dietary Intervention.....	79
1.8.2 Mental Health .....	86
1.9 Conclusion.....	90
1.10 Bibliography .....	90
1.11 Thesis Aims.....	119
2.0 PAPER-I.....	120
3.0 PAPER-II.....	146
4.0 PAPER-III.....	202
5.0 PAPER-IV.....	223
6.0 PAPER-V.....	245
7.0 General Discussion.....	297
7.1 Effects of perturbing the basic methylation machinery (Papers I and II).....	297
7.1.1 Neuroepithelial genes .....	297
7.1.2 Body mass regulation.....	302
7.1.3 The <i>UGT1A</i> detoxification gene cluster .....	303
7.1.4 Cancer-testis genes .....	304
7.1.5 Activation of innate immune genes in UHRF1-depleted differentiated human cells .....	308
7.1.6 ERV reactivation and innate immune response to Uhrf1 mutation in mouse .....	311
7.2 Methylation-deficient systems are indicative of alternative repressive mechanisms (Paper I and II) .....	313
7.2.1 UHRF1.....	313
7.2.2 DNMT1 .....	317
7.3 Differences in application of large-data analytics to human epidemiological rather than cell-line work (Paper III and VI).....	321
7.4 Effects of environment on methylation on current and future generations (Paper III and VI).....	322
7.4.1 Folic acid supplementation in the second and third trimester causes alterations in DNA methylation upstream of a key imprint regulator.....	322
7.4.2 Alterations to the DNA methylation of immune response genes in sufferers of Depression .....	328
7.5 The development of CandiMeth and possible future versions .....	333
7.6 Concluding Remarks.....	335
7.7 Bibliography .....	337
8.0 Achievements.....	354
8.1 Published Abstracts.....	354
8.2 Additional Research Training .....	354
8.3 Certificates .....	354
8.4 Conference Presentations .....	354

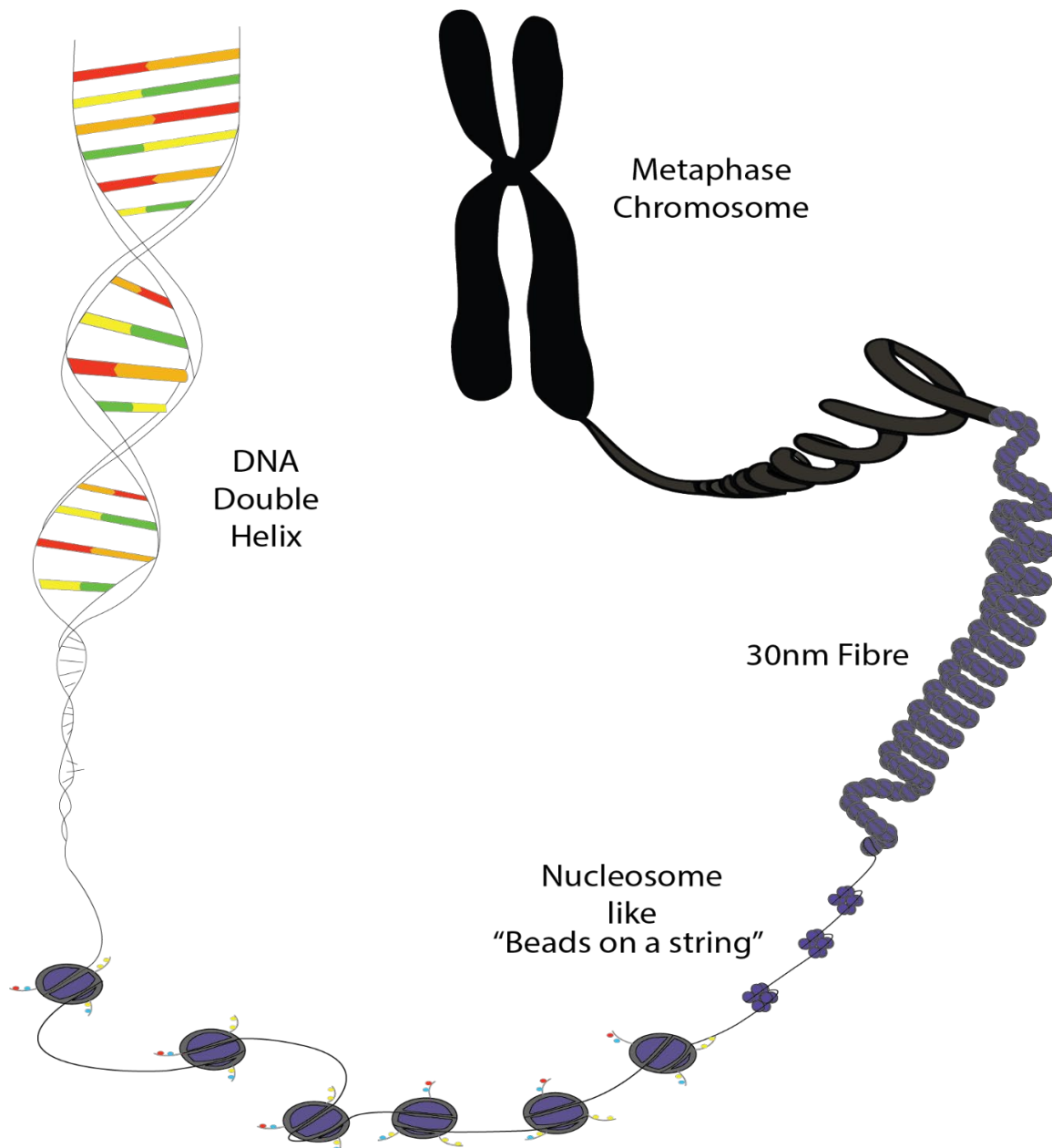
8.4.1 Oral Presentations.....	354
8.4.2 Poster Presentations.....	354
8.5 Grants.....	355
8.6 Other Achievements .....	355
8.7 Publications .....	355

## 1.0 General Introduction

In this introductory chapter, I will provide detail regarding current knowledge of DNA methylation and its effects on different regions of the genome. In addition to, the various methylation-based machinery utilised to maintain and establish these DNA methylation marks investigated within this thesis and give a brief background into histone marks as they pertain to the work presented. I will also provide background detail into the various bioinformatic tools and techniques utilised within this thesis to give insight to those not familiar with such techniques.

### 1.1 DNA packaging into chromatin

The DNA of the human genome is approximately 3 billion base pairs long, when measured end to end this equates to 2 metres in length. In order to fit this information into a nucleus with a 6µm diameter, the DNA must be heavily compacted into a structure known as chromatin (Figure 1). First, 147bp of DNA is wrapped around a four-core octamer of histone proteins, forming a nucleosome. These consist of two molecules of histone H2A, H2B, H3 and H4 each containing positively charged lysine or arginine which binds electrostatically to the negative charge of the phosphate within nucleotides (Kornberg, 1974). Each nucleosome is linked to the next via a 50bp linker DNA wrapped around a lysine rich H1 histone. This first stage of DNA compression compacts the DNA to approximately one third of its original size. The H1 linker proteins of each nucleosome interact with other nucleosomes to allow the DNA to coil into spirals of 6-8 nucleosomes, forming a structure known as a solenoid – the second stage of DNA compression. Following this, the solenoids coil further to form the transcriptionally repressive heterochromatin and subsequently, a metaphase chromosome (Fazary et al., 2017).



**Figure 1: How DNA is packaged into Chromatin.** 147bp of DNA is wrapped around a four-core octamer of histone proteins, forming a nucleosome. These consist of two molecules of histone H2A, H2B, H3 and H4 each containing positively charged lysine or arginine which binds electrostatically to the negative charge of the phosphate within nucleotides. Each nucleosome is linked to the next via a 50bp linker DNA wrapped around a lysine rich H1 histone. This first stage of DNA compression compacts the DNA to approximately one third of its original size. The H1 linker proteins of each nucleosome interact with other nucleosomes to allow the DNA to coil into spirals of 6-8 nucleosomes, forming a structure known as a solenoid – the second stage of DNA compression. Following this, the solenoids coil further to form the transcriptionally repressive heterochromatin and subsequently, a metaphase chromosome. [original figure SJ Thursby]

However, as mentioned heterochromatin is the transcriptionally repressive form of chromatin and does not allow transcriptional machinery to access the DNA of the cell and produce the proteins required. This is facilitated via euchromatin, the transcriptionally active form of chromatin, formed via the modification of histone tails which extend from the solenoid. These histone tails are targeted by chromatin re-modelling complexes, which using ATP, unwind the compressed configuration of the chromatin to enable access of the transcriptional machinery to the DNA (Tang et al., 2010).

## 1.2 Histone Marks

Histone Modifications are the regulators of chromatin configuration. They were first formally described in 1964 (Allfrey et al., 1964) and predominantly target the previously mentioned histone tails which protrude from the solenoid. These modifications occur most frequently at the N-terminal tail of the histone and act to alter the binding ability of DNA to the histone proteins and thus the chromatin structure, making the chromatin more/less accessible to transcriptional machinery (Gates et al., 2017; Rea et al., 2000). Examples of histone modifications include, methylation (generally repressive), acetylation (generally activating), phosphorylation (involved in DNA repair), ubiquitination (DNA damage signaling) and SUMOylation (generally repressive (Shiio and Eisenman, 2003)) – of which acetylation and methylation are the most well-characterised (Alaskhar et al., 2018).

To elicit a histone modification, specialized groups of enzymes are required e.g. histone acetyltransferases (KATs) and histone methyltransferases (HMTs) which add acetyl & methyl groups to the histone tails respectively and histone deacetylases (HDACs)/demethylases (HDMs) which remove these marks making histone marks essentially reversible (Wang et al., 2018). In terms of gene activity, each histone mark can also be described as activating or repressive. For example, trimethylation of histone 3 lysine 27 (H3K27me3) is associated with



transcriptional repression as it influences the binding/recruitment of certain proteins to the DNA. H3K27me3 is characteristic of the polycomb group of proteins (discussed further in 1.7.5). Whereas, methylation of histone 3 lysine 4 (H3K4me1) or acetylation of histone 3 lysine 9 (H3K9ac) has been associated with transcriptional activation (Ernst and Kellis, 2012; Gates et al., 2017). Lysine acetylation affects the overall electrical charge of the histone and as a result changes how histone interacts with DNA (Alaskhar et al., 2018). An overview of histone marks and their transcriptional associations can be found in Table 1.

In addition to epigenetic writers (HMTs) and erasers (HDMs), epigenetic reader proteins also exist. These proteins function to regulate the actions of epigenetic writers and via interaction with the histone mark, can determine its function. Examples of epigenetic readers include, methyl-binding protein MeCP2 or the histone methyltransferase SETDB1 (Alaskhar et al., 2018; Biswas and Rao, 2018).

However, the histone tails are not the only region that can be modified by such groups of enzymes, the central globular domains (nucleosomes) also host multiple histone modification sites which are involved in histone-histone and histone-DNA interactions. (Lawrence et al., 2016).

**Table 1: Histone modifications and their abbreviations and transcriptional associations (Alaskhar et al., 2018; Wang et al., 2008)**

<b>Modification</b>	<b>Abbreviation</b>	<b>Association</b>
<b>Histone 3 lysine 4 dimethylation</b>	H3K4me2	Activating
<b>Histone 3 lysine 4 monomethylation</b>	H3K4me1	Activating
<b>Histone 3 lysine 4 trimethylation</b>	H3K4me3	Activating
<b>Histone 3 lysine 9 acetylation</b>	H3K9ac	Activating
<b>Histone 3 lysine 9 dimethylation</b>	H3K9me2	Repressive
<b>Histone 3 lysine 27 acetylation</b>	H3K27ac	Activating
<b>Histone 3 lysine 27 trimethylation</b>	H3K27me3	Repressive
<b>Histone 3 lysine 36 trimethylation</b>	H3K36me3	Activating
<b>Histone 4 lysine 20 monomethylation</b>	H4K20me1	Activating

### 1.3 Epigenetics and DNA Methylation

Epigenetics can be defined as heritable, reversible changes in gene expression without changes in the underlying DNA sequence. Such modifications include histone marks and DNA methylation, both critical to regulating gene expression, imprinting, X inactivation and maintaining genomic stability. DNA methylation is the more characterised of these marks and can be defined as the addition of a methyl group to a cytosine residue within the DNA sequence. This results in different effects depending on where the alteration is located (Edwards et al., 2017; Johnson and Coghill, 1925).

### 1.4 Distribution of DNA Methylation

#### 1.4.1 Methylation at CpG Islands

In the mammalian genome, DNA methylation usually exists within CpG sites, that is a C sequentially preceded by a G with the p representing the phosphate group between these bases. The idea of CG methylation was first mentioned by Johnson and Coghill in *Tulercule bacillus* (1925). Since then, in the human genome, approximately 28 million CpG sites have been identified, distributed throughout 99% of the genome (Deaton and Bird, 2011). The CpG dinucleotide has also been found to cluster in large numbers (approximately 200bp), termed a CpG Island (CGI) making up 5% of all CpGs and 1% of the human genome. CGI are more specifically defined as: longer than 200bp, a C + G content greater than 50% and an observed/expected ratio of 0.6 or greater. These CGIs usually reside in an unmethylated state, often within promoters, but can also be found within gene bodies and intergenic sites throughout the genome – methylation at which has differing effects dependent on island location. For example, CGI methylation at transcription start sites (TSS) blocks transcription, whereas methylation within the gene body has the opposite effect and methylation at repeat regions is essential for genomic stability (Jones, 2012; Vinson and Chatterjee, 2012). These characterizations will be explained in more detail in the following paragraphs.

Methylation at CpG sites makes these cytosines more susceptible to spontaneous deamination. Spontaneous deamination of unmethylated cytosine generates uracil, which is removed by base excision repair using uracil-DNA glycosylase. However, spontaneous deamination of methylated cytosine results instead in thymine and despite the actions of thymine-DNA glycosylase and methyl-CpG-binding protein 4, this results in a C > T mutation. CGI are less susceptible to deamination due to their unmethylated state (Pfeifer, n.d.; Walsh and Xu, 2006). The CGI are thought to be kept in an unmethylated state by the binding of basal transcription factors and also specialized proteins such as CFP1, protecting the TSS from mutation. Thus, the high mutability of methyl is thought to have shaped the mammalian genome into islands and deserts with respect to CpG and explains the high coincidence of CGI and promoters (Schübeler, 2015).

Over 70% of identified CGI are in promoters, with most unmethylated and within TSS i.e. in housekeeping genes as one example (Deaton and Bird, 2011). Although, mammalian promoters fall into three categories, high CG content (HCP), intermediate CG content (ICP) or low CG content (LCP) with the CG density having varying effects on DNA methylation at that area. HCP promoters are generally unmethylated and protected from methylation by the presence of positive histone marks such as histone 3 lysine 4 methylation or via the presence of TET1 which converts 5mC to 5hmC, CFP1 is also involved in this process (Jones, 2012; Maunakea et al., 2010; Vinson and Chatterjee, 2012). The methylation state of HCPs also correlates well with gene expression. Conversely, LCP tend to be methylated in somatic cells as they are a target for de novo methylation. The expression of methylated LCP is thought to be tissue-specific and important for the silencing of germline-specific genes, imprinting and regulation of retrotransposons (Jones et al., 2015). They have also been known to be methylated during development. However, in development and in somatic

cells, the methylation status of LCPs does not always correlate with gene expression (Illingworth and Bird, 2009; Walsh and Bestor, 1999; Yoder et al., 1997).

In recent years, ICPs have also been identified as a class of CGI density promoters: although these promoters have variable methylation states dependent on gene activity their methylation status has been found to correlate well to their expression, as in HCP (Jang et al., 2017; Weber et al., 2007).

However, not all CGI reside within TSS or promoters, as previously indicated (Jones, 2012). Those outside of transcriptional units are termed 'orphan' CGIs due to the uncertainty over their significance (Deaton and Bird, 2011; Meng et al., 2015). These were discovered via CXXC Affinity Purification which isolates clusters of unmethylated cytosines. CXXC binds the protein CFP1 which recruits H3K4me3 that subsequently blocks methylation to that site. Further orphan CGI were discovered via the combination of this technique with next generation sequencing (CAP-seq) (Illingworth et al., 2010).

Orphan CGI exhibit more of a tissue-specific methylation profile than annotated promoters, ~34% of intragenic CGI are methylated in the brain, possibly to prevent spurious transcription (Illingworth et al., 2010); de novo methylation during development has also been found to affect orphan CGI more than annotated promoters, indicating that orphan CGI may be more tightly regulated than those at TSS or within promoters.

Methylated orphan CGI have been found marked with H3K4me3, a positive histone mark indicative of an active promoter. Alternative reports have revealed evidence of RNA polymerase II at these H3K4me3 sites, indicating orphan CGIs may be intragenic or alternative promoters that could give rise to novel transcripts and indicate a potential regulatory role for orphan CGI. Evidence has also arisen that certain orphan CGI may be

alternative promoters for ncRNA that could regulate gene expression – see 1.4.3 for more information on gene body methylation (Illingworth et al., 2010; Illingworth and Bird, 2009; Maunakea et al., 2010).

#### 1.4.2 Non-CpG Methylation

Although not investigated in this thesis, methylation at non-CpG sites does exist. It was originally discovered in plants in 1975, then later in bacteria, fungi and in human ESC (Cokus et al., 2008; Fuso, 2018; Jones, 2012; Lindroth et al., 2001; Rountree and Selker, 1997). Non-CpG methylation is usually denoted CpH (where H can stand for A, C or T). In prokaryotes, methylation at CpA or CpC sites has been found to aid DNA repair and protect the cell from foreign bacterial and viral genomes (Meng et al., 2015).

In Eukaryotes, a high frequency of CpH methylation has been found in mouse and human ESC and in induced pluripotent stem cells (iPS), largely CpA (Lister et al., 2009; Meng et al., 2015; Ramsahoye et al., 2000a). Approximately a quarter of all methylation found in human ESC is non-CpG derived, with the greatest enrichment in the gene body (GB)

– the function of this methylation is currently not well understood but it is thought to have a role in developmental gene regulation, such as gene repression during embryogenesis, as non-CpG methylation was lost after differentiation of ESCs but recovered after they were restored into iPS (Lister et al., 2009). This theory coincides with the work of Ramsahoye and colleagues (2000), who discovered that during the early post implantation stage of embryogenesis, de novo methylation was observed at many non-canonical sites, but this methylation was lost after development. This could be because one of the main enzymes known for maintenance methylation, DNMT1, has been shown to have no notable effect on non-CpG methylation and therefore the non-CpG methylation is not present in differentiated cells (Gowher and Jeltsch, 2001). When methylation at CpG and non-CpG sites

of DNMT1 KO mouse ESC were examined, the cells maintained their methylation at non-CpG sites but not at CpG sites. However, upon DNMT3L KO, the ESC displayed lower levels of CpA methylation, indicating DNMT3 enzymes may be more important for regulation of non-CpG methylation. The lower levels of the de-novo methylation enzymes in differentiated cells may also explain why non-CpG methylation is rare in such cell types (Jang et al., 2017).

Moreover, when the level of mCpH is reduced in ESCs the cells display a reduction in differentiation capacity. This evidence also concurs with that of Han *et al.*, (2011), who suggest that non-CpG methylation aids establishment and maintenance of cell identity.

Another theory states that at the later stages of development, non-CpG methylation may not be required due to additional mechanisms such as chromatin modification systems becoming more effective and therefore the non-CpG methylation is no longer required (Gowher and Jeltsch, 2001; Jang et al., 2017; Ma et al., 2014).

In recent years, a high frequency of non-CpG methylation has been found in human skeletal muscle, hematopoietic cells and in the brain, in neurons and glial cells – with neural cells having a higher level of non-CpG methylation than glial cells. Non-CpG methylation accounts for 53% of total 5mC within the brain, it is established and conserved throughout development and is one of the most abundant forms of neuronal methylation found within this tissue (Fuso, 2018; Lister et al., 2013; B H Ramsahoye et al., 2000). This indicates a potential regulatory role for non-CpG methylation in this tissue. Other studies have also found evidence suggesting this type of methylation is related to brain pathology and aging as the abundance of mCpH in the brain increases with age but mCpG does not (Guo et al., 2014; Lister et al., 2013; Xie et al., 2012).

### 1.4.3 Gene Body Methylation

Gene body methylation was first assessed via a genome-wide screen by Zhang (2006) & Zilberman (2007) in *Arabidopsis thaliana*. When assessed in the human genome, it became evident that over a third of DNA methylation also occupied intragenic regions i.e. the gene body (GB). Within both plants and animals, DNA methylation tends to lie within the GB but rarely at the 5' or 3' ends of genes (Flanagan and Wild, 2007). Considering that plants and animals diverged over 1.6 million years ago, yet the placement of methylation remains similar, this would suggest that GB methylation had ancestral function (Suzuki and Bird, 2008). Subsequent studies revealed GB methylation was correlated with increased transcription, as well as possibly with regulation of intragenic promoters and regulation of splicing (Yang et al., 2014) and see section 1.4.1 above.

Although the above paragraph points towards the overall effect of GB methylation, more specifically, it depends on the position of the methylation within the GB, CpG density and the histone marks present. For example, GB methylation in the first exon is tightly correlated with transcriptional silencing but this correlation does not exist for the downstream exons and introns (Brenet et al., 2011). At downstream intron and exon junctions (where nucleosome occurrence is greater), GB methylation is thought to destabilise nucleosome placement, which leads to transcriptional initiation. Therefore, GB methylation may indirectly influence splice events (Andersson et al., 2009; Luco et al., 2011; Tilgner et al., 2009). Other reports have also inferred an indirect link between intragenic DNA methylation and the regulation of splice events. Examples of this include CTCF binding, an inhibitory action also regulated by DNA methylation. CTCF binding is known to pause RNA Pol II, and since RNA Pol II has a clear role in splicing, DNA methylation may be



indirectly linked to splice events (Ehrlich et al., 2016; Lorincz et al., 2004; Maunakea et al., 2010; Shukla et al., 2011).

In terms of histone marks within the GB, methylation is inversely correlated with the positive histone mark, H3K4me3 (Barski et al., 2007) which is associated with open chromatin and particularly enriched at promoters and unmethylated CGI. In the presence of methylated H3K4me3, DNMT3A exists in an inhibited state where it cannot effectively bind to DNA to induce de-novo methylation. However, in the presence of H3K4me0, DNMT3A changes structure and binds to the nucleosome with the help of its H3K4me0 sensing accessory protein DNMT3L – a non-catalytically active form of DNMT3 (Ooi et al., 2007; Guo et al., 2015; Hashimoto et al., 2010). Also, GB methylation is positively correlated with H3K36me3 which represses aberrant transcription following RNA Pol II action (Carrozza et al., 2005; Joshi and Struhl, 2005). The PWWP domain of DNMT3A which is involved in targeting chromatin has been shown to recognise H3K36me3. This interaction has been shown to increase the activity of DNMT3A and elicit de-novo methylation at this mark (Dhayalan et al., 2010; Rondelet et al., 2016). Therefore, intragenic DNA methylation may regulate intragenic promoters by preventing their spurious transcription (Maunakea et al., 2010).

Greater than 45% of the genome consists of repetitive elements which reside primarily in CG rich intragenic regions. These intragenic repetitive elements also have intragenic promoters thus initiation of their transcription could have a negative effect on genomic stability (Yoder et al., 1997). GB methylation is thought to have evolved to protect the genome against their transcription via dense de novo methylation of their intragenic promoters (Brenet et al., 2011; Maunakea et al., 2010). This coincides with the discovery

that housekeeping genes rarely have a downstream CGI but approximately 49% of genes with lower and more heavily regulated expression patterns exhibit such downstream functional elements (Larsen et al., 1992).

In CpG poor regions, DNA methylation is inversely correlated with H3K9me3 (Schotta et al., 2004) and H4K20me3 (Li et al., 2011), repressive chromatin marks involved in compacting chromatin and repression of transcription (Hahn et al., 2011). H3K39me3 is thought to have the same function as DNA methylation in areas of low CpG density and work with H3K36me3 to suppress aberrant transcription. This was observed mechanistically in the below study.

In HCT116 DKO of DNMT1 + DNMT3B 95% of methylation was lost but a specific group of genes marked with the positive histone mark H3K36me3 retained intergenic methylation – most likely due to the actions of DNMT3A which has been linked to intergenic methylation and maintenance activity in previous work investigating the absence of DNMT1 (Taiping Chen et al., 2003; Wu et al., 2010).

In an alternative study, when HCT116 cells were treated with the demethylating agent, Aza, similar results are observed. A specific group of genes related to cellular growth and metabolic pathways rapidly remethylate, it is thought this was due to the de novo action of DNMT3B. However, upon withdrawal of the treatment, some regions displayed sustained DNA demethylation and transcriptional activation even in the presence of DNMT1 (Yang et al., 2014). This sustained demethylation was suggested to be the result of H3K27me3, the repressive histone mark most often associated with polycomb mediated repression, which has been found to invade adjacent sites in absence of DNA methylation (Reddington *et al.*,

2013). The polycomb complex has also been found to block the access of DNMT3B its target DNA (Jin et al., 2009).

As mentioned, GB methylation is highly correlated with transcriptional activation and RNA markers of transcriptional initiation have been correlated with intragenic CGI. However, only genes of intermediate level expression exhibit the highest intragenic methylation. Jjingo and colleagues (2012) proposed a theory that this occurrence could be due to the RNA Pol II levels becoming so high in highly expressed genes that it interferes with the efficiency and ability of DNMT1 to access the DNA and maintain methylation. Similar theories have been suggested since then (Jjingo et al., 2012; Lorincz et al., 2004; Rountree and Selker, 1997; Shukla et al., 2011;).

However, there are exceptions to the GB methylation and transcriptional activation correlation. According to Aran (2011) this correlation is true only for proliferating cells and cell lines. Inactive and genes in tissues with little proliferations exhibit similar levels of methylation. In their investigation, cells which had an early replication time were correlated with high levels of DNA methylation. Whereas, tissues such as the lungs, kidney and brain (low proliferative rates) fail to demonstrate a positive correlation between GB methylation and expression. Aran et al., (2011) and colleagues also suggested this was a result of low levels of DNMT3B in slowly proliferating cells, hinting that, DNMT3B may act as a transcription-coupled DNA methyltransferase in somatic tissues.

In addition to this exception, transcribed regions may contain many functional genomic features, including, promoters, enhancers and repeat elements which may require specific transcription factors in addition to intragenic DNA methylation to become upregulated (Kulis et al., 2012; Maunakea et al., 2010; Varley et al., 2013; Yang et al., 2014).

#### 1.4.4 Methylation at Enhancers

In addition to CpG islands, promoters and the gene body, there is yet another regulatory element within the human genome – enhancer regulatory sequences. These are usually intergenic regulatory sequences thought to be responsible for cellular specialization via regulation of cell- and tissue-specific expression patterns through multiple different mechanisms of action (García-González et al., 2016). Enhancers were first described in monkey tumor virus studies, when a 72bp repeating sequence was deleted in what is now known to be the SV40 enhancer (Banerji et al., 1981). This caused vastly reduced viral protein levels and reduced virus viability. In mammals, the first enhancer was found in mouse, in the immunoglobulin heavy chain gene. In this instance, gene activity was dependent on the binding of cellular specific transcription factors (TF) (Gillies et al., 1983).

The mechanism of action of enhancers is still not fully understood. Currently, there are two main theories; the binary model and the progressive model (García-González et al., 2016).

The binary model suggests that enhancers increase the proportion of molecules that activate transcription at that given locus. The progressive model suggests that enhancers increase the number of RNA molecules transcribed but not the number of molecules that initiate transcription. Overall, enhancers function as TF binding sites that bind then loop over to their target sequences approximately one kilobase away, affecting their

transcription. There are two theories regarding TF binding at enhancers, the ‘enhanceosome’ model and the bill-board model. The ‘enhanceosome’ model suggests DNA is a scaffold for TF binding complexes to form which then influences transcription. The bill-board model suggests every bound TF is independent of each other and acts a single unit.

Either way, most studies agree that the critical protein CTCF found at most enhancers helps

to form a loop with cohesin to allow enhancer/target promoter interactions (García-González et al., 2016).

Enhancers are also marked by histones which determine their activity state. Active enhancers can be characterised by the presence of H3K4me1, H3K79me3 and H3K27ac. This type of enhancer also lacks any DNA methylation. Poised enhancers remain in contact with their target DNA but lack the active H3K27ac mark and are usually marked by H3K4me1. Repressed enhancers also exhibit H3K4me1, but in addition to the repressive H3K27me3. H3K4me1 has been found to stay at active enhancers even after they disengage from their target locus, aiding in the protection of the locus from DNA methylation and maintaining the chromatin state for future enhancer use: cytosine hydroxymethylation (5hmc) has also been found to protect enhancer loci from accumulation of cytosine methylation. However, it is difficult to map enhancer/gene pairs using these histone marks and DNA methylation sites, as enhancers can be degenerate and some histone marks, such as H3K4me1 are also found at alternative regulatory sites i.e. insulators (Benetatos and Vartholomatos, 2018; Smith and Shilatifard, 2014).

It has been suggested that DNA methylation may be required for the deposition of most histone marks, with the exception of H3K4me3 which has been found independent of DNA methylation states. It is also responsible for the maintenance of repressed or poised enhancer states in a tissue specific manner (Ehrlich et al., 2016). It has been proposed that DNA methylation is regulated by the binding of DNA-directed TF that encourage DNA demethylation via the interaction with promoters/enhancers and the recruitment of TET1 for active demethylation. This demethylation is thought to be an early stage in the

activation of enhancers preceding TF binding, mediator complex conformational alterations and recruitment of RNA Pol II (Plank and Dean, 2014).

Aberrant cytosine methylation at enhancers has been correlated with many different types of cancer, imprinting disorders and chronic kidney disease development (Kerr et al., 2019; Ko et al., 2013; Qu et al., 2017; Yoon et al., 2005). As mentioned, there is a negative correlation between DNA methylation and chromatin accessibility, in addition to TF binding and DNA methylation that can activate key cancer drivers (Clermont et al., 2016). Also, as enhancers are responsible for cellular specialisation and tissue specific expression patterns, aberrant DNA methylation can cause irregularities in these instances. For example, erroneous DNA methylation at enhancers related to haematopoiesis results in failure to discrimination between foetal, adult erythropoiesis and granulopoiesis through repression of enhancers during maturation (Bell et al., 2016; Benetatos and Vartholomatos, 2018). The relationship between enhancers and DNA methylation is similarly disrupted in AML, MDS and many other types of cancer via abnormal enhancer activity - resulting in anomalous gene expression changes that can contribute to tumorigenesis (Benetatos and Vartholomatos, 2018; Clermont et al., 2016; Heyn et al., 2016).

In imprinting, for instance at the *H19/IGF2* locus, DNA methylation-regulated CTCF binds to multiple sites within the imprint control region (ICR) of this locus. These sites are typically unmethylated and vital for the inhibition of the enhancers close to this ICR. Inhibition of these enhancers leads to the expression of H19 and repression of maternal *IGF2* – irregular methylation here can also contribute to the cognate imprinting disorder Beckwith-Wiedemann syndrome (García-González et al., 2016; Hark et al., 2000; Plank and Dean,

2014). A similar situation can be observed at the *RASGRF1* imprinting locus, again regulated by the methylation sensitive CTCF (Yoon et al., 2005).

In one study into enhancer methylation dynamics and cancer plasticity, enhancer methylation was found to be indicative of patient outcome with high accuracy (Bell et al., 2016). This and similar studies have proposed using DNA methylation at enhancers as a biomarker of disease (Clermont et al., 2016; Qu et al., 2017).

#### 1.4.5 Transposable Elements

Approximately 50% of the human genome is made up of non-coding DNA termed 'selfish' or 'parasitic' DNA (Slotkin and Martienssen, 2007). Some of this consists of autonomous elements which can make copies of themselves which insert in new locations: these were discovered over 50 years ago (McClintock, 1951) and are known as transposable elements, as they can move throughout the genome. There are two main types of transposable elements, retrotransposons (class I) and DNA transposons (class II). Class I transposable elements are termed retrotransposons due to their transposition via reverse transcription, these make up the majority of retrotransposons. There are also two main sub-types of retrotransposon, those with long terminal repeats (LTRs) and those without (non-LTR). LTR-containing transposons are characterised by the presence of direct repeats at the end of the repeat element, undergo duplicative transposition and lack envelope proteins required to exit the cell. An example of an LTR retrotransposon is LTR10C which is enriched at sub-telomeric regions (Bourgeois and Boissinot, 2019; Cardoso et al., 2016; Tutton and Lieberman, 2017).

Type II transposable elements constitute 3% of TE and duplicate via a transposon-encoded protein termed a transposase (Lander et al., 2001). This protein recognises repeats which

flank this type of transposable element, excises the TE sequence from the donor sequence and into the acceptor site. The empty donor site remaining can then be filled with the same sequence via gap repair or without a replacement sequence – similar to the ‘cut and paste’ action.

The role of TE in the genome is unclear, they are suspected of having a regulatory role in gene expression (Chenais, 2015; Drongitis et al., 2019; Trizzino et al., 2018), genome evolution and X-inactivation (Cohen et al., 2007; Kapitonov and Jurka, 2005; Zhou et al., 2004). It is thought that LINE elements boost the spread of silencing away from the X-chromosome inactivation centre on the inactive X chromosome in females ensuring effective silencing (Lyon, 2006; Pinheiro and Heard, 2017). The movement of TE often has deleterious effects however, particularly in affected coding regions – leading to mutation, dysregulation and possible loss of gene activity, in addition to possible chromosome breakage, illegitimate recombination and genome rearrangement (Chenais, 2015; Kazazian et al., 1988; Lin et al., 2009).

However, eukaryotic genomes have evolved various epigenetic silencing mechanisms to inhibit the effects of TE. Such mechanisms include chromatin modifications and DNA methylation - with DNA methylation being the most effective (Chenais, 2015; Sotero-Caio et al., 2017). Inhibition of DNMT1 using 5’aza-dC in human ESC resulted in activation of LINE1 elements (Woodcock et al., 1997). In mice, intracisternal alpha particles (IAP) were activated in embryos with a hypomorphic mutation in DNMT1 (Walsh et al., 1998). DNMT3L has also been found to be required for IAP silencing in premeiotic male germ cells, suggesting methylation of TE may also be required during development to maintain silencing,



potentially explaining why de novo methylation occurs in newly integrated elements (Bourc'his and Bestor, 2004).

#### 1.4.6 Copy Number Variation

The human genome is constantly changing, this is how humans evolve and adapt to environmental changes. Although single nucleotide polymorphisms and trisomy/monosomy have been the centre of evolutionary research for many years, in the last 30 years a new intermediate variant has been identified as having a potential causative link to evolution and disease – copy number variation (CNV). Within the literature, the definition and limits of the term CNV are variable, but the most common definition refers to a DNA segment that is 1kb in length or longer which occurs at a variable copy number in comparison with a reference genome (Zarrei et al., 2015). CNVs belong to the category of structural variants – variants of the genome which alter chromosomal structure- this includes balanced changes such as inversions and translocations and unbalanced changes i.e. CNVs. CNVs can be simple in structure, such as deletions and duplication, or more complex gains/losses of a sequence at many different sites throughout the genome. They also fall into the categories of adaptive and maladaptive, for example, the CNVs in the alpha-amylase gene which enables the digestion of dietary starch is an adaptive CNV (Perry et al., 2007). Maladaptive CNVs are usually associated with disease, for example, autism (Pinto et al., 2014), schizophrenia (Girard et al., 2011) and Crohn's disease (Craddock et al., 2010).

The first casual association of a CNV with a phenotype was over 80 years ago, when it was observed that a duplication of the *Bar* gene in *Drosophila Melanogaster* was found to cause the Bar Eye phenotype (Sturtevant, 1925). Since then, CNVs arising via homologous, non-homologous and erroneous replication mechanisms have been found to be associated with disease. In a study of CNVs from the HapMap Project, for 80% of cases copy number was

correlated with gene expression, with the remaining 20% in negative correlation with gene expression. However, greater than 50% of CNV associated with gene expression were not located in coding sequences, potentially indicating the mechanisms of CNV action are diverse (Stranger et al., 2007). CNVs have been found to alter gene expression via disrupting gene interactions through position effects, deleterious genetic changes or altered gene dosage e.g. the microdeletion related to Angelman Syndrome (Gamazon and Stranger, 2015; Williams et al., 1989).

Originally, CNV analysis was done via comparative array hybridisation (Ahn et al., 2015; lafrate et al., 2004). This technique involves hybridising test DNA to reference DNA with different fluorescence labels and measuring the difference in fluorescence at different regions in the genome. Oligonucleotide arrays then became popular due to the ability to conduct comparative genomic hybridisation and use SNP-based arrays. However, these arrays can lack probes in many areas of the genome and therefore many led to the misrepresentation of structural variations (Alkan et al., 2009). The advancement of next generation sequencing (NGS) has improved this resolution issue, but now CNVs must be identified as benign or maladaptive (Sudmant et al., 2010). Lack of population-based CNV data was a limitation for resolving this issue, but now many consortiums- such as The HapMap Project (Belmont et al., 2003), Database of Genomic Variants (MacDonald et al., 2014), DECIPHER (Firth et al., 2009) and The 1000 Genome Project (Altshuler et al., 2012)- are working towards generating population-based CNV data via NGS technologies to identify further maladaptive structural variants.

#### 1.4.7 5-hydroxymethylcytosine

Although not investigated in this thesis, an alternative cytosine modification does exist, termed 5-hydroxymethylcytosine (5hmC). This modification was first discovered in T-even

bacteriophage (Wyatt and Cohen, 1952) and then at a later date in mammals (Penn et al., 1972). However, the evidence for 5hmC in mammalian DNA was not successfully reproduced for many years and led to a lack of investigation of that modification until in 2009, evidence of high levels of 5hmC was found in mESCs (Tahiliani et al., 2009) and Purkinje neurons in mouse (Kriaucionis and Heintz, 2009), with later studies elucidating its presence in other tissues (Globisch et al., 2010).

The presence of 5hmC is dependent on the presence of methylated cytosine and occurs via TET-catalysed oxidation of methylated cytosine from 5mC to 5hmC. This is also the first step in active demethylation – highlighted via the depletion of TET1 in mESCs which lead to reduced 5hmC and increased 5mC, resulting in transcriptional repression. 5hmC is then further converted into 5-formylcytosine and 5-carboxylcytosine by TETs and converted back into unmodified cytosine, most likely via base-excision repair through thymine DNA glycosylase (He et al., 2011; López et al., 2017).

5hmC may also have a role in passive demethylation. The maintenance methyltransferase exhibits a poor affinity for 5hmC and therefore during mitosis, 5hmC is not re-established on daughter strands and thus not passed on through cell division. After many rounds of cell division, this may, in theory, lead to an exponential loss in methylated cytosine and subsequently, transcriptional repression. However, as mentioned in previous sections, the effects of a loss in methylation depends on the location of the loss (Amouroux et al., 2016; Hill et al., 2014). A summary of the roles of DNA methylation can be found in Table 2.

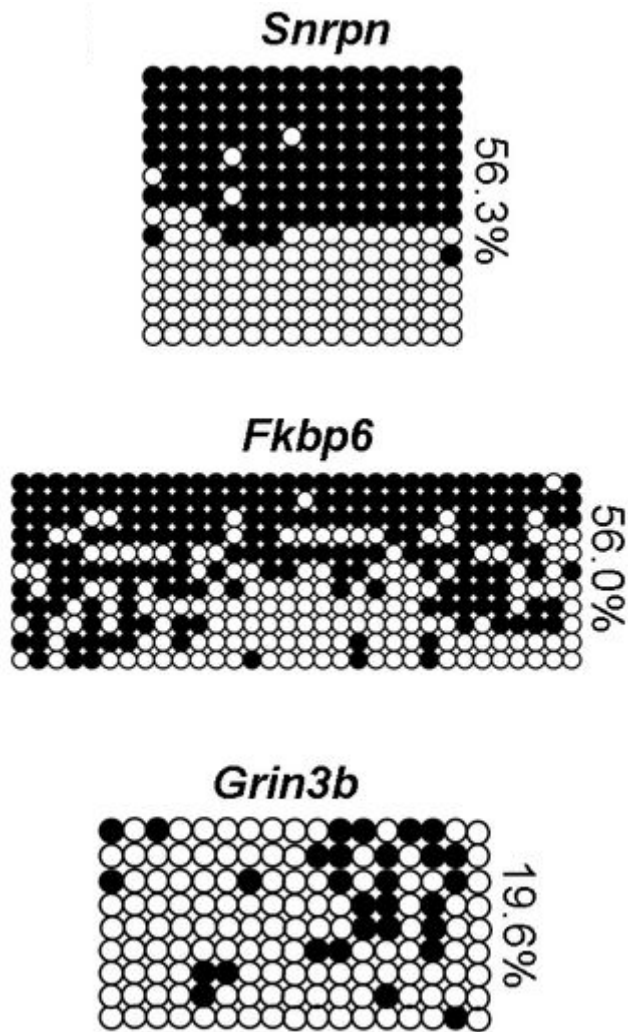
<b>Table 2: Effects of DNA methylation at different genomic locations</b>				
<b>Methylation Location</b>	<b>Genomic Location</b>	<b>Examples</b>	<b>Effects of Methylation</b>	<b>References</b>
<b>1) CpG Islands</b>	a) Promoter	TNF- $\alpha$ Promoter	Repressive	(Jang et al., 2017; Pieper et al., 2008)
	b) Intragenic	STC2	Activating	(Yang et al., 2014)
<b>2) non-CpG</b>	a) Embryogenesis	OCT4	Repressive	(Fuso, 2018; Lister et al., 2009; Bernard H. Ramsahoye et al., 2000b)
	b) Neuronal cells	RGS9	Repressive	(Fuso, 2018; Rizzardi et al., 2019)
<b>3) Gene Body</b>	a) First exon	CDKN2B	Repressive	(Brenet et al., 2011)
	b) Alternative exons	CDKN2A	Activating	(Arechederra et al., 2018)
<b>4) Enhancer</b>	a) Intragenic	TREX2	Repressive	(Weigel et al., 2019)
<b>5) Repeats</b>	a) Promoter	LINE1	Repressive	(Woodcock et al., 1997)
	b) 5'-CCGG-3' sites	IAPs	Repressive	(Walsh et al., 1998)

## 1.5 DNA Methylation Assessment Methods

### 1.5.1 Gene Specific

#### 1.5.1.1 Bisulphite Sequencing

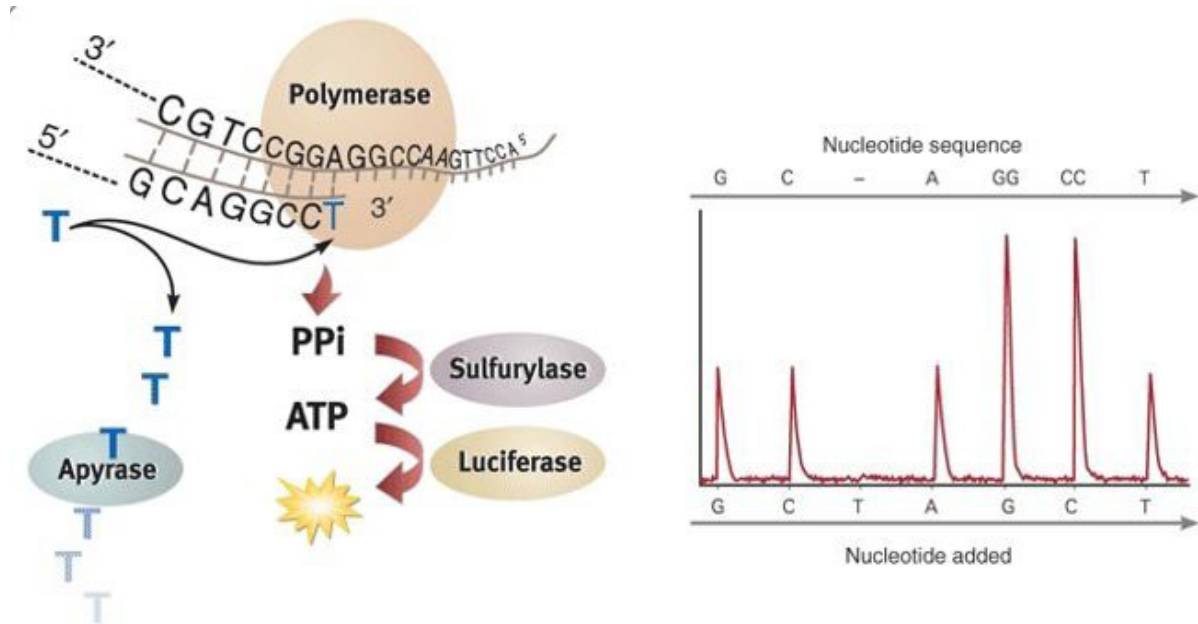
Before pyrosequencing was available, the most popular way to assess DNA methylation was bisulphite sequencing (Figure 2). This process involved using bisulphite to treat DNA, which would convert all unmethylated cytosines to thymine and leave methylated cytosines as unmodified cytosine. The DNA was then amplified via PCR and cloned into a vector. This vector was then ligated and transformed into bacteria such as *E-coli* and grown over a period of 24 hours to form individual colonies with plasmids representing individual PCR strands within them. Following this, DNA was extracted from the bacteria and sent for sequencing to see the results of the bisulphite conversion. The output sequence then showed the presence of cytosine (if protected by methylation) or thymine (if not) at each CpG site. By comparing the sequence from the PCR product with the reference genome being used, the methylation status of the sequence can be elucidated for each individual fragment of DNA. This method is useful for examining not just the methylation state of the target sequence but also for assessing the strand-specificity of the methylation, which can reveal allele-specific methylation.



**Figure 2: Example of Bisulphite sequencing.** Also known as clonal analysis. This technique can clarify the methylation state of genes across many samples. It is particularly useful for checking allele specific methylation at imprinted loci. Here, the methylation of an imprint (*Snrpn*), testis gene (*Fkbp6*) and brain gene (*Grin3b*) can be observed. The imprint has the typical all-or-nothing methylation on each allele as expected of imprints. The *Fkbp6* shows a similar spread of methylation and the methylation in the brain gene *Grin3b* is much lower in methylation across samples. Taken from (Rutledge et al., 2014).

### 1.5.1.2 Pyrosequencing

Pyrosequencing has become the standard go-to molecular biology technique for assessing site specific methylation. It uses PCR-amplified DNA, a collection of sequence-specific primers and DNA polymerase to accomplish a sequencing-by-synthesis reaction. This is facilitated via the addition of a biotin tag to either the forward or reverse primer (whichever is at the 5' end) and addition of magnetic beads to the amplified DNA. The magnetic beads bind to the biotin tag of the DNA fragments and the cartridge floods the wells with free deoxyribonucleotides. Whenever a complementary base binds to the sequencing primer a pyrophosphate is liberated. Following this, ATP sulfurylase converts the pyrophosphate into ATP in the presence of adenosine 5' phosphosulfate. The resulting ATP then catalyses the reaction of luciferin to oxyluciferin which generates a flash of visible light proportional to level of ATP. This flash is recorded by a camera and generates a pyrogram, from which the individual nucleotides within the sequence can be determined as well as the methylation of the CpG sites of interest (Figure 3).



**Figure 3: Overview of Pyrosequencing.** Magnetic beads within the pyrosequencing cartridge bind to biotin labelled DNA fragments. When the wells of the cartridge are flooded with free deoxyribonucleotides, apyrase removes the nucleotides not incorporated by RNA polymerase enzyme. When a complimentary base binds to the sequencing primer on the DNA fragment, a pyrophosphate is liberated. Then, ATP sulfurylase converts the pyrophosphate into ATP. This ATP then catalyses the reaction of luciferin to oxyluciferin and a flash of light is generated proportional to the level of ATP. The light is recorded by a camera and generated a pyrogram and any nucleotides added and the methylation of any CpG sites can be observed. Taken from (England and Pettersson, 2005)



## 1.5.2 Array-Based

### 1.5.2.1 Illumina BeadChip Arrays

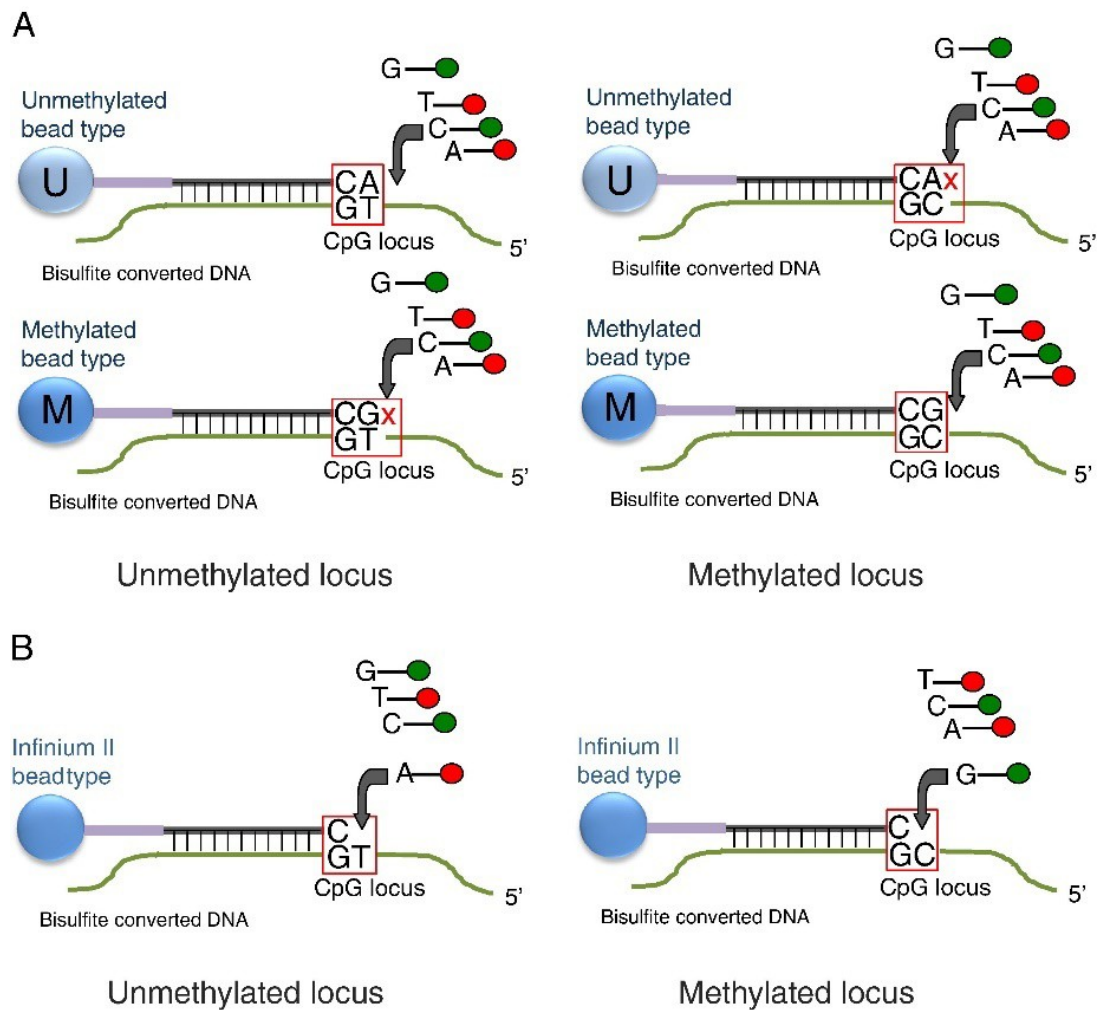
In order to provide a cost-effective method of assessing genome-wide methylation without the expense of whole genome bisulphite sequencing (WGBS), Illumina designed an oligonucleotide array which uses 850,000 probes to assess over 99% of RefSeq Genes, 95% of CpG islands and most enhancer regions discovered through the ENCODE and FANTOM5 projects (ENCODE Project Consortium, 2004; Kawai et al., 2001).

The array works via the use of two different types of probe chemistries, Infinium type I probes, and Infinium type II probes. Infinium Type I probes utilise two bead types per CpG site, one methylated and the other unmethylated, whereas Infinium Type II probes work via one bead type with a degenerate R base (see Figure 4 below). In this case, methylation of the CpG site is determined at the nucleotide level (Bibikova et al., 2011).

DNA for assessment is bisulphite converted as above, leading to the conversion of unmethylated cytosine to uracil, which is amplified as thymine following PCR amplification. Methylated cytosine will not be affected by the treatment. Following this, the fragmented DNA is hybridised to the array and, using hapten-labelled dideoxynucleotides, single base extension is conducted. After multiple rounds of immunohistochemical assays, the intensity of fluorescence is scanned using the Illumina iScan and reported in intensity data files as a beta value between 0 and 1, 1 being highly methylated and 0 representing low methylation (Morris and Beck, 2015a; Pidsley et al., 2016).

When originally designed, the array had only approximately 27,000 probes, made with Illumina's Infinium type I probe chemistry. Following this, Infinium Type II probes were invented, and the next version of the array had a mixture of two probe chemistries and

approximately 450,000 probes. Following further innovation, and re-engineering of over 2000 probe sequences, the methylation EPIC array was created with over 850,000 probes.



**Figure 4: Overview of Illumina Infinium Probe Chemistries.** The array works via the use of two different types of probe chemistries, Infinium type I probes (A), and Infinium type II probes (B). Infinium Type I probes utilise two bead types per CpG site, one methylated (M) and the other unmethylated (U). Whereas Infinium Type II probes work via one bead type with a degenerate R base. This binds to a complimentary hapten-labelled free nucleotide and the methylation of that CpG is determined at the nucleotide level following fluorescence scanning. Taken from (Bibikova et al., 2011).

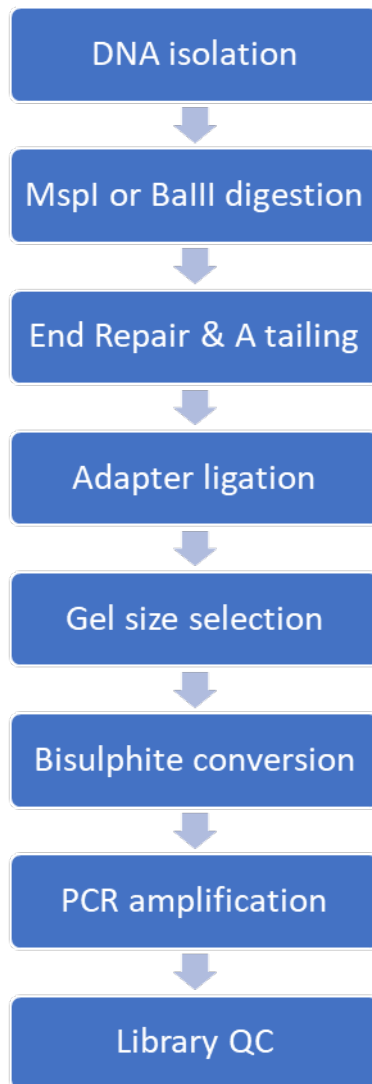
### 1.5.3 Sequence-Based

#### 1.5.3.1 Reduced Representation Bisulphite Sequencing

Reduced Representation Bisulphite Sequencing (RRBS) was an innovation of sanger sequencing that resulted in greater CpG resolution at a more cost-effective price than WGBS. RRBS is useful to assess targeted regions of the genome at high resolution. This is because high sequencing depths are not often required due to the low input yield of this technique. It covers approximately 10-15% of the CpG in the human genome (approximately 4.2 million CpG) but cannot be used to resolve 5hmC, non-CpG methylation or CpGs in regions not covered by the restriction enzyme used (Fouse et al., 2010).

After DNA extraction, a methylation-insensitive restriction enzyme is used to fragment the genome. *MspI* is usually the most common choice, which cuts at CCGG, but *BglII* can also be used. Following digestion, the 5' CG overhangs are repaired using deoxyguanosine triphosphate and deoxycytidine triphosphate nucleotides and A tails are added, as Illumina primers have a 3'-T overhang. The adapters also contain methylated cytosines so they are not converted to uracil after bisulphite conversion (Guo et al., 2015; Meissner et al., 2005).

Once the ends are repaired and adapters ligated, the fragment size is selected from an agarose gel and the DNA is bisulphite converted, modifying all unmethylated cytosines to uracil and amplifying them as thymine. The bisulphite-converted DNA is then amplified via PCR and sequenced using the Illumina platform (most commonly), see figure 5 for a schematic. The next step is to align the fragmented sequences and quantify the methylation levels on each sequence by comparison with the reference genome (Fouse et al., 2010; Guo et al., 2015; Lee et al., 2015; Meissner et al., 2005).



**Figure 5: Overview of RRBS.** Genomic DNA is fragmented with a methylation insensitive restriction enzyme such as MspI or BclI. This creates fragments of the genome with 5' overhangs which are filled and repaired using complementary nucleotides. As Illumina adapters have 3' T overhangs, poly-A tails are added to the fragmented genomic DNA to allow adapter ligation. Fragment size is calculated from running via gel electrophoresis and DNA is bisulphite converted, converting all unmethylated cytosines to uracil and amplifying them as thymine. The converted DNA fragments are then amplified using PCR and the quality of the library checked prior to sequencing, most commonly on the Illumina platform. Adapted from (Gu et al., 2011)

### 1.5.3.2 Whole Genome Bisulphite Sequencing

WGBS is the gold standard of genome-wide methylation assessment: it covers all 28 million CpG sites within the genome, in addition to any CpH, CHG or CHH sites, where H represents A, T or C. Originally, WGBS was ineffective due to the degradation of DNA resulting from adapter ligation and bisulphite conversion. Current WGBS techniques solve this issue via the use of tagging sequences on the 3' and 5' ends of the single-stranded bisulphite converted sequence. A polymerase capable of reading uracil nucleotides, plus random DNA primers are used to synthesize a complementary DNA strand with a 5' random hexamer tagging sequence. A further random hexamer sequence tag is added to the 3' end of the sequence, this allows Illumina P7 and P5 adapters to be ligated to the tagged sequence via PCR, where an index/barcode can be added between the tagging sequence and the adapter in the case of multiplexing – to distinguish sample specific sequences. The sequencing template used is the complement to the first bisulphite sequence and therefore, Read 1 from the sequencing results will be the same as the first strand of bisulphite-treated DNA (Cokus et al., 2008; Saxonov et al., 2006).

WGBS is usually sequenced using a flow cell in a HiSeq or Novaseq sequencer with 75bp paired-end reads resulting in approximately 120GB of data per human genome sample.

Methylation calling, alignment and differential calculations can then be computed using the Bismark software (Krueger and Andrews, 2011a).

### 1.5.3.3 RNA-sequencing

RNA-sequencing (RNA-seq) is the process of assessing genome wide transcript levels i.e. the transcriptome. The process is highly customisable but generally follows the same basic structure of: fragmenting RNA; constructing cDNA; synthesising the second strand of the cDNA; ligating adapters; amplifying the fragments using PCR and then sequencing the newly

created genomic library (Kukurba and Montgomery, 2015). When planning an RNA-seq experiment, many things must be considered, for example, the type of RNA to be assessed or whether the strandedness of the RNA should be kept. A total RNA extraction includes ribosomal RNA (rRNA), pre-posttranscriptional modification messenger RNA (pre-mRNA), messenger RNA (mRNA) and non-coding RNA (ncRNA) in addition to many other smaller categories of RNA. Although, other library preparation techniques are available, like polyA-enrichment. Library preparation procedures will change depending on the RNA subcategory of interest e.g. for mRNA, a poly-A enrichment protocol can be followed, as mRNA is usually characterised by a poly-A tail. Next to be considered is whether the RNA library should be stranded or unstranded (Kirby, 1956; Levin et al., 2010).

Following traditional cDNA synthesis protocols, using a reverse transcriptase and then a DNA polymerase will cause the strand information of the original RNA to be lost. To prevent this, the first strand of cDNA is synthesized using a reverse transcriptase, as normal, then chemical labels such as dUTP are added prior to synthesizing the second strand of cDNA using a DNA polymerase. Adding a dUTP will cause the second strand of cDNA to have many uracil bases which can be removed enzymatically before sequencing (Bentley et al., 2008; Borodina et al., 2011; Kukurba and Montgomery, 2015; Levin et al., 2010). Following adapter ligation, this allows the forward strand to be differentiated from the reverse strand. In addition to this, options also exist regarding the primer type for second strand cDNA synthesis. Oligo-dT primers (most common) can be used if only fully matured mRNA is required (only fully matured mRNA has a poly-A tail), or random primers can be used if all maturation states of mRNA is desired (Borodina et al., 2011; Hansen et al., 2010; Kirby, 1956).

The next step in RNA-seq library preparation is the adapter ligation. Adapters are ligated to the 5' and 3' ends of sequences, such sequences usually consist of a specific sequence that allows the fragment to attach to the flow cell and a sequencing primer for the sequencing reaction to proceed as desired. A barcode sequence may also be added, within adapter sequences if using Illumina technologies, as this will allow different samples to be run on the same flow cell lane i.e. multiplexing. This option can also save money when assessing the sequences of large-scale RNA-seq libraries (Bentley et al., 2008; Busby et al., 2013; Kukurba and Montgomery, 2015).

The final step in planning an RNA-seq experiment is choosing the appropriate sequencing depth for the experimental aims. Sequencing too shallow will result in inaccurate reads, but too much depth will result in increased variance/convoluted results. In most cases, adequate depth is defined as approximately 30-40 million reads per sample. To assess the diversity of a highly complex library, approximately 500 million reads per sample are required (Wang et al., 2011).

However, caution is advised when conducting RNA-seq as low RNA integrity will also result in inaccurate sequences. The quality of the RNA should therefore be checked at multiple stages throughout the process. Many RNA-seq library preps include 'spike-ins' - standard reference sequences set by the External RNA Controls Consortium (Jiang et al., 2011), which are included at different concentrations at different stages of the library prep procedure to assess the quality, sensitivity and coverage of a library preparation protocol (McIntyre et al., 2011; Raz et al., 2011; Volkin and Carter, 1951). Biological replicates are also favored over technical replicates in RNA-seq (unless assessment of the actual technique is desired), this provides greater re-assurance when conducting differential gene expression, due to the



inclusion of genomic variation between biological replicates (Bullard et al., 2010; Kukurba and Montgomery, 2015).

#### 1.5.3.4 Chromatin Immunoprecipitation sequencing

Chromatin Immunoprecipitation sequencing (ChIP-seq) is the addition of NGS sequencing after the immunoprecipitation of chromatin, for elucidating the binding sites of proteins of interest. The basic process is similar to that of the previous sections; after following an NGS-compatible ChIP protocol, the resulting sequences are fragmented by either sonication (non-histone protein enrichment) or MNase (histone protein enrichment). Next, the sequence ends are repaired as in previous sections and sequencing adapters/indexes are ligated to the DNA fragments. PCR-based amplification is then conducted and, if using an Illumina based sequencer, cluster generation and sequencing is then conducted (Bentley et al., 2008; Landt et al., 2012; Park, 2009).

However, like in RNA-seq in the last section, there are many considerations to be aware of when conducting a ChIP-seq experiment. First, ChIP-seq verified antibodies are essential as these have been tested to be compatible with NGS sequencing and do not have high levels of cross-reactivity with other antibodies or a high level of non-specific binding sites – which could adversely affect sequencing results. In addition to this, consideration must be given to the type of antibody used i.e. monoclonal or polyclonal, as these will bind to differing numbers of epitopes, which will again influence sequencing results. ChIP antibodies are quality tested prior to NGS using an RNAi knockdown of the protein of interest, any ChIP signal observed will be the result of non-specific antibody binding (Kidder et al., 2011; Teytelman et al., 2009).

Secondly, the cell count needed to conduct a ChIP-seq experiment needs to be clarified. There is a delicate balance between signal intensity and noise after immunoprecipitation and NGS. For abundant proteins like RNA Polymerase or histone marks, cell numbers in the range of  $1 \times 10^6$  are recommended. If the protein of interest is rare,  $10 \times 10^6$  cells are recommended (Adli et al., 2010). Biological replicates are also needed to distinguish signal enrichment from that of biological variation (Kidder et al., 2011). In addition to, using a control input chromatin sequence or a non-specific immunoglobulin antibody as a reference sequence control. This will then allow the user to conduct peak enrichment of the sequencing results in MACS2, with greater confidence in their findings. It will also allow them to assess the effects of sequence shearing, background noise and antibody cross reactivity (Feng et al., 2012; Gaspar, 2018a).

Thirdly, it is important not to over-amplify the DNA fragments, this can be checked by comparing the length of the PCR product with the original size of the adapter-ligated DNA. Overamplified DNA will exhibit a 200-300bp drift in PCR product size. PCR overamplification can be corrected computationally at the data analysis stage after sequencing, however an ideal ChIP-seq protocol would prevent an overamplification prior to sequencing by reducing the number of PCR cycles if the quantity of DNA is low (Brinkman et al., 2012; Park, 2009).

Finally, the last item to consider in a ChIP-seq experiment is the sequencing depth required. As mentioned previously, low sequencing depth may result in inaccurate sequence results. For a ChIP-seq experiment, the depth of sequencing is determined by the prevalence of binding of the protein of interest. For histone marks, deeper sequencing is required to determine the point in which sequence levels do not equate to further peak enrichment – as histone marks are rather diffuse in their binding and this can be difficult to elucidate from

background noise. The opposite is true for proteins with less dense binding sites as these are more easily clarified from background noise due to the higher sequence abundance at enriched sites (Kharchenko et al., 2008; Kidder et al., 2011).

Additionally, while single end sequencing works well for most ChIP-seq cases, if the desired protein falls within a repeating sequence region, paired end sequencing is generally preferred. This can provide deeper sequencing, improved alignment efficiency and therefore yield more representative sequencing results (Kharchenko et al., 2008; Kidder et al., 2011; Landt et al., 2012).

#### 1.5.3.5 Assay for Transposase Accessible Chromatin sequencing

Assay for Transposase Accessible Chromatin sequencing (ATAC-seq) is an NGS technique used to assess regions of open chromatin. It can also be used to map nucleosome positions and study transcription factor occupancy (in collaboration with other NGS techniques) and to identify novel enhancers/predict enhancer development (Buenrostro et al., 2015a, 2015b). The basis of ATAC-seq was originally developed for NGS library preparation but since then has been adapted into a quick and efficient technique to assess regions of open chromatin without complicated sequencing library preparations (Buenrostro et al., 2015a; Meyer and Liu, 2014).

The theory behind ATAC-seq centres around a mutant hyperactive Tn5 transposase that cuts at sites of open chromatin. ATAC-seq uses a process known as tagmentation, the simultaneous fragmenting and tagging of genomic DNA with sequencing adapters. An ATAC-seq protocol adheres to the following steps: 1) approximately 50,000 cells are harvested and lysed to obtain areas of open chromatin; 2) a Tn5 transposase then simultaneously fragments the open DNA and tags the fragments with sequencing adapters; 3) following

purification, the DNA is amplified using PCR and finally 4)sequenced using paired-end sequencing at a depth of approximately 50 million reads per sample for a human genome (Buenrostro et al., 2015a).

As with any NGS protocol, several variables must be considered. The cell numbers optimal for ATAC-seq are relatively small in comparison to other NGS library preparations (millions for ChIP-seq). The original protocol used 500 – 50,000 cells but the optimum amount depends on the aims of the experiment and the species/cell used, the general number of cells used is 25,000 to 50,000 cells. Low cell number leads to under transposition and overly high cell number leads to over transposition (Buenrostro et al., 2015a; Corces et al., 2017). The number of PCR cycles must also be controlled and equate to as few cycles as possible, to minimise PCR duplicates and over-amplification. Finally, sequencing depth must be carefully calculated. It is dependent on the size of the genome of interest, samples from the human genome require at least 50 million reads per sample. More shallow sequencing depths are adequate for a smaller genome. Paired-end sequencing is used in ATAC-seq as it provides further sequencing depth, more information on where the transposase inserted (this can be on the anti-sense strand and single stranded ATAC-seq would not detect this) and it allows PCR duplicates to be more easily identified (Buenrostro et al., 2015a, 2015b, 2013a, 2013b; Heinz et al., 2010; Kharchenko et al., 2008).

## 1.6 Bioinformatic Analysis

### 1.6.1 R

The R software platform is a command line environment commonly used for the analysis of high dimensional data, like that obtained in genome-wide association studies. After downloading R, it comes with a series of default packages for data analysis, data wrangling and graphical visualization. The user has the option to install a constantly growing number

of community-written R packages. These are sets of scripts and custom functions that can be loaded into the R environment and are usually customized to a specific purpose i.e. graphing or array analysis. R was originally built from a linux-based format and is built on the statistical language S. However, an assortment of computational languages can be used in R including, C, C++ and Fortran. Custom functions are also easily be created in R with low coding intensity and can be specific to desired data-wrangling requirements (R Core Team, 2013). To utilise the R environment a basic core graphical user interface (GUI) is provided by default with the framework, although most users choose to operate the R console via the RStudio Interactive Development Environment (IDE) (RStudio, 2015). This allows them to keep track of environmental variables, create custom scripts in the script editor, and view any graphical visualisations in the plot window. This IDE also offers debugging capabilities and frameworks for interactions with web browsers via the Shiny package (Chang, 2019).

## 1.6.2 Array-Based Processing Methods

### 1.6.2.1 Pre-processing and Normalisation

Due to normal genetic variation (polymorphisms), the different types of probe chemistries used within the Illumina BeadChip arrays, and the discovery that some Illumina probes exhibit non-specific cross reactivity in certain circumstances, it is necessary to process and normalise IDAT data prior to conducting any further analysis.

Pre-processing of array data seeks to remove defective probes that have high detection p values - such values are indicative of inadequate fluorescence scanning and therefore may provide erroneous results (Roessler et al., 2012). Those at single nucleotide polymorphism sites are removed as they could be effects of purely genomic variation and therefore their methylation values would be representative of either the intervention or reference WT population at that CpG site. Probes that have non-specifically bound to other genomic regions

may also give values unrepresentative of the methylation of their actual target loci (Chen et al., 2013).

Following pre-processing of raw IDATs, the subsequent data must be normalized to correct for the differences in the design of the two Infinium probe chemistries. The differences in the mechanisms of action of the two probe types results in two different  $\beta$  value distributions after fluorescence scanning (Dedeurwaerder et al., 2013), see figure 4 for further detail. These differences are mostly due to dye bias and differences in background or residual fluorescence when recording methylation values in both colour channels. For example, Infinium type I chemistry utilises two bead types per CpG site, one that records a methylated signal and one that records the unmethylated signal. These bead types are scanned via the same colour channel and so are not as affected as type II probes (type II probes use two different colour channels due to their degenerate bases) (Bibikova et al., 2011).

Several normalization procedures to correct for these effects have been proposed, examples of the most cited methods include; *Subset quantile Within Array Normalisation (SWAN)* and *Beta Mixture Quantile normalisation (BMIQ)*. The SWAN method (Maksimovic et al., 2012) assumes that regions with similar probe coverage will reside in similar genomic regions and therefore exhibit similar methylation profiles, which is more than often not true and therefore not an accurate representation of the data. It then divides the data on the basis of this assumption and attempts to correct the differences in results between the two probe types. However, multiple investigations have reported a reduction in data reproducibility following SWAN normalisation (Fortin et al., 2017; Wu et al., 2014).

The BMIQ method (Teschendorff et al., 2013) has no assumptions and splits the raw data into 3 types, corresponding to probes which exhibit high methylation, intermediate methylation and low methylation. It then utilises quantile normalisation to adjust the values

of the type II probes to fit the distribution of the type I profile. This method has been shown to improve data reproducibility and data quality (Dedeurwaerder et al., 2013; Fortin et al., 2017; Teschendorff et al., 2013) and is also the method used to normalise array data within this thesis. Additional normalisation strategies, including between-array normalisation are also available and have been reviewed here (Heiss and Brenner, 2015; Morris and Beck, 2015a; Triche et al., 2013)

However, it is to be noted that background correction of any residual fluorescence prior to scanning should be carried out before normalisation, as this has been found to be more effective than combination normalisation strategies (Dedeurwaerder et al., 2013, 2011).

#### 1.6.2.2 Epidemiological Based Correction Methods

In addition to the unwanted sources of variation present in cell line-based studies (see 1.7.6), clinical and human intervention-based studies present a different kind of experimental variation, due to the use of whole blood and saliva DNA collection methods. As mentioned in section 1.4, DNA methylation is highly tissue-specific. Variations in whole blood cell composition has been identified as a possible experimental confounder that should be corrected for, in order to arrive at a more accurate DNA methylation profile independent of intra-cellular immune cascades (Titus et al., 2017). While SVA should correct for both known and unknown sources of unwanted variation, cell type composition correction algorithms do exist, such as that by Houseman *et al.*, (2012), which uses regression-based models to determine the distribution of immune cells and effects on the DNA methylation results, benchmarked against validated controls consisting of known mixtures of the relevant cell types. A reference-free approach has also been developed for cases in which a validated control is not currently available (Houseman et al., 2014).

### 1.6.2.3 Cell-Line Based Correction Methods

Although sequence and array results have already gone through pre-processing to remove variation due to different types of probe chemistries, or PCR-based library bias, alternative forms of unwanted variation exist. Examples of such variation include batch effects from differing library prep personnel, differences in array readers or flow cells and different reagents utilised in prepping for the technique (Price and Robinson, 2018; Tom et al., 2017). Therefore, multiple methods have been suggested to correct for this unwanted technical variation (Alter et al., 2000; Benito et al., 2004; Leek et al., 2012). One of the most popular methods is surrogate variable analysis (SVA) (Leek and Storey, 2007). SVA can remove both known and unidentified sources of unwanted variation, in addition to working effectively even in small sample sizes.

The premise of SVA works in four steps, splitting the data into the separate sources of variation i.e. batch variables or phenotypic variables- such as age or sex. It then looks to see if those variables are exhibiting more variation in the data than they otherwise would by chance and tests to see if there is a significant association between the subset of data and that variable. It then builds a surrogate variable to model the entire dataset to determine what it would look like without that unwanted source of variation. Finally, it corrects for that variable in any later regression-based models. SVA has been shown to reduce technical variation and improve reproducibility within assessment techniques, improving the identification of anomalous differentially expressed features (Leek et al., 2012; Leek and Storey, 2007).

### 1.6.2.2 R Packages for Illumina BeadChip Array Analysis

Since the release of both the Illumina HumanMethylation450k BeadChip array and the Illumina MethylationEPIC array, many R-based packages for the analysis of methylation



array data have been released and an overview of the packages utilised within this thesis can be found below.

#### 1.6.2.2.1 Limma

Limma stands for Linear Models for MicroArrays and is mostly used to discover differentially expressed genes, their associated p-values and for correcting these p values for multiple testing. Data can be analysed within *limma* as an expression set or as a matrix of M values (rows indicating probes and columns indicating samples). M values are used here instead of  $\beta$  values as they represent a distribution more compatible with the assumptions (normality) of a *limma*-based linear model (Assenov et al., 2014). Following this, the design of the experiment is input into *limma* using a design matrix and any form of hierarchical comparison can be elicited via the model design (Smyth, 2004; Wilhelm-Benartzi et al., 2013).

*Limma* uses a moderated t-statistic to conduct differential analysis. This is similar to a t test but with an empirical Bayes-based modification – the standard deviations of the samples have been shrunk towards a common value. Prior to differential gene analysis, *limma* fits a linear model to each gene to assess the relationship between the genes and the differences in the samples. This, and the shrinkage of standard deviations allows *limma* to borrow information across genes to give the analysis greater statistical power for inference of information about each gene. From this, *limma* can then decipher if there are any differentially expressed genes (Smyth, 2004).

#### 1.6.2.2.2 Minfi

*Minfi* was one of the earliest developed R packages for the analysis of microarray data. It is based on R's version of object-oriented programming, that is S3 and S4 object classes, and allows less coding-intensive manipulation of input data. *Minfi* provides a full pipeline with all

the modules required for preliminary and downstream analysis of array data, including both differentially methylated region (DMR) finder, *bumphunter* and differentially methylated position finder, *DMP Finder*, although the latter is not recommended as a primary way of locating differential enrichment. *Minfi* also provides many forms of normalisation, both within- and between-array normalisation, the ability to compare between 450k and EPIC array platforms, surrogate variable analysis and cell type correction if using whole blood. It can also be easily integrated with *limma* for differential analysis. However, as versatile as this package is, it does not provide an option to run all analysis modules as a start-to-finish pipeline.

#### 1.6.2.2.3 RnBeads

*RnBeads* provides a complete analysis pipeline for microarray and NGS bisulphite sequencing data and is one of the most comprehensive packages to date. Whether run as one default pipeline or using each module separately, *RnBeads* provides publication-quality graphics and displays analysis results in user-friendly HyperText Mark-up Language (HTML) outputs. Within the default analysis pipeline, IDATs returned from the EPIC array can be input into R, quality control conducted to remove problematic probes or biases and normalisation performed to correct against differences between type I and type II Infinium probe chemistries. After this, an exploratory analysis of the acceptable methylation values can be carried out including comparing global DNA methylation profiles between sample groups and multiple genomic regions. Many types of dimensionality reduction like multi-dimensional scaling (MDS) and principle component analysis (PCA) can be performed using this package. *RnBeads* also displays heat maps and regional methylation profiles to identify potential experimental differences.

A differential methylation module utilising limma-based linear models is also included. This module outputs multiple types of whole genome methylation analyses, such as scatter and volcano plots, displayed in various different forms such as html format as well as the accompanying comma separated variable (csv) format. From this, tracks for UCSC genome browser can in principle be computed. However, these require an File Transfer Protocol (FTP) server to export the resultant data as a UCSC track hub, which is not a resource that the majority of biomedical science researchers have access to or experience in operating. In addition, within the default pipeline, or if desired in tailored analysis, the differential analysis module will also initiate an enrichment analysis of the differential methylation results via hypergeometric testing. This will identify gene ontology categories which demonstrate differential methylation patterns in comparison to the experimental control which can then be further investigated by both wet-lab and dry-lab processes (Assenov et al., 2014; Morris and Beck, 2015b).

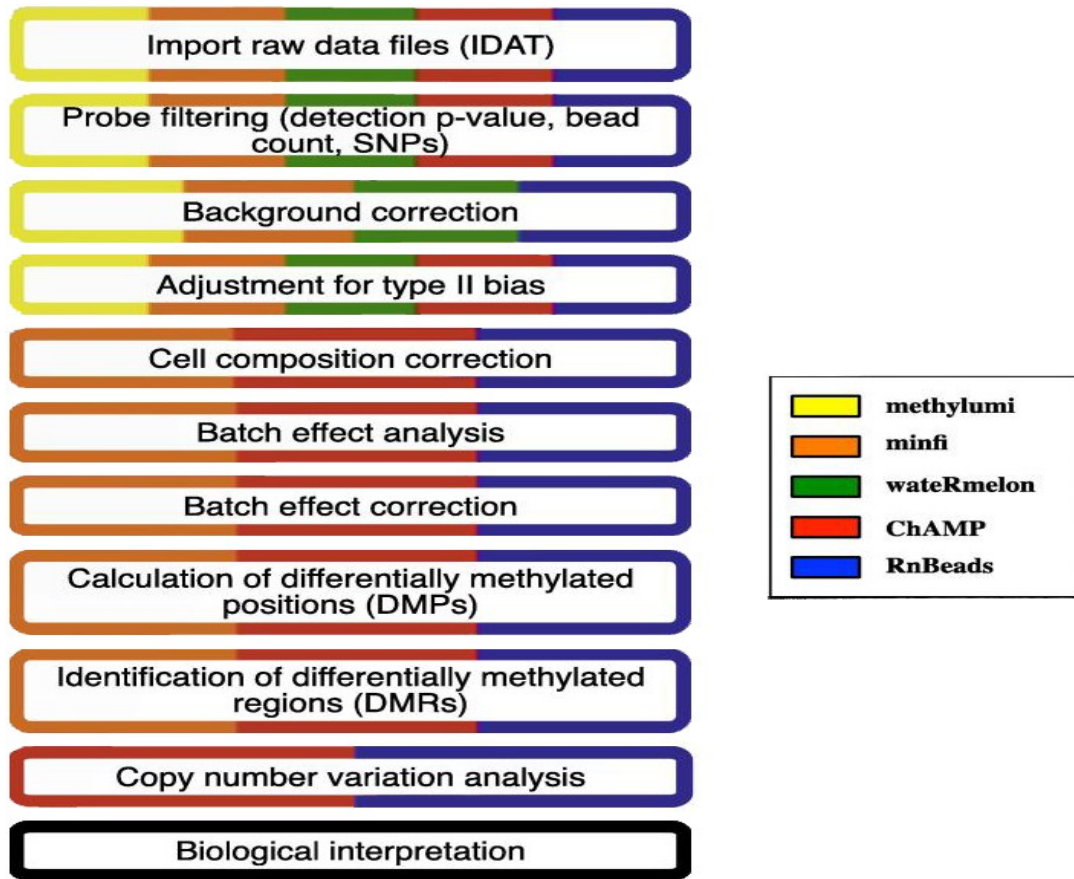
Recently, *RnBeads* has also been updated with a GUI for ease of use and many new features, including copy number variation analysis, age prediction and differential region enrichment (Müller et al., 2019).

#### 1.6.2.2.4 ChAMP

The Chip Analysis Methylation Pipeline is also one of the more popular analysis packages for microarray data and, like *RnBeads* features a complete pipeline in which all modules can be conducted via a single “run” command. It was originally created to make *minfi* and associated packages easier to use for beginner R users but has now grown into a fully comprehensive analysis pipeline (Morris et al., 2014).

*ChAMP* accepts as input to a beta matrix in addition raw IDATs, which is particularly of use if importing results from a different R package. It also offers many different types of normalisation and pre-processing options (between-array and within-array) and provides a module for investigating and correcting batch effects, as well as methods for cell type correction and CNV analysis.

Unique to *ChAMP* it offers 3 types of regional differential methylation analysis in addition to differential block finders and the option to input results into Gene Set Enrichment Analysis (GSEA). *GSEA* allows the user to assess whether their differentially methylated results are related to any biological pathway. The user can also submit their results (from *ChAMP*) for global methylation assessment, in which differentially methylated regions will be determined via *GSEA* processes. Both of these options are highly useful, as *GSEA* itself is a difficult program for new users to work. Moreover, *ChAMP* provides interactive *Shiny*-based HTML outputs and interactive *Plot-ly* based graphics – making analysis of resultant graphics easy for the end user (Morris et al., 2014; Tian et al., 2017).



**Figure 6: Overview of methylation array processing pipelines and their capabilities.** Methylation arrays need to be pre-processed and normalised before conducting differential methylation analysis. *Methylumi* (yellow) capabilities only extend as far as the pre-processing stage. Whereas, *ChAMP* (red), *RnBeads* (blue) and *minfi* (orange) provide full packages for the pre-processing and analysis of methylation array data. *RnBeads* and *ChAMP* even provide deeper copy number variation analysis for further insight into methylation array data. Adapted and updated from (Morris and Beck, 2015)

### 1.6.3 Sequence-Based Processing Methods

All NGS-based techniques produce the same specific file type, FASTQ. These are sequence-based files with quality control measurements embedded within them. FASTQ files consist of multiple entries for every read on the sequencer. Each entry consists of four lines; A sequence identifier with information about that run, the sequence recorded, a plus sign which is used as a separator and a base call quality score which can be used for quality control and later in the analysis i.e. if variant calling. If single-read sequences are read, one entry is recorded for every sample, per flow cell lane. If paired-end runs are desired, two files are created, Read 1 and Read 2 for every sample and again for every lane in the flow cell. Since FASTQ files can be large in size, they are usually compressed and output as \*.fastq.gz – most programs can also work with these files in their compressed format to save on computational resources.

As with the analysis of array-based methods, sequence-based methods also have to go through quality control and pre-processing prior to mapping and analysis. An outline of the general pipeline for NGS analysis is discussed below.

#### 1.6.3.1 FastQC

FastQC is a program developed by the Babraham Institute to provide an easy-to-use and -interpret method of quality control on sequence reads. It provides HTML and text outputs showing basic statistics, such as average sequence read length and per base sequence quality, in addition to giving an indication of whether adapters have been removed from the sequence using the over-represented sequences module. Furthermore, it provides analysis on whether the flow cell used shows any particular tile or sequence bias, as this could alter sequence read quality. From here, bad quality samples can be excluded, or the end of lesser

quality reads can be trimmed. FastQC is a Java based application and is typically only compatible with linux-based OS such as Ubuntu or MAC OS (Andrews, 2010).

#### 1.6.3.2 Adapter Trimming

Following quality control, the sequences must have the adapters, which were ligated during the library prep, removed prior to downstream analysis, as these could also alter analysis results. Illumina provides the sequences for their adapters publicly (Illumina, 2019), but some adapter trimming programs such as Trim Galore! (Krueger, 2012) can automatically detect the adapter sequence used and remove it. This results in a shortened fastq file which is then passed through FastQC again to double check for any quality issues that may have been hidden due to the adapters i.e. overrepresented sequences.

#### 1.6.3.4 Mapping

The next step in NGS analysis is to map the sequence reads to the genome of interest. This can be done using a reference genome-based approach or via a de novo genome build (higher sequence depth required). For the reference-based approach, multiple tools are available and are dependent on the sequence type i.e. for DNA-based studies such as for variant calling, WGBS etc, non-gapped aligners such as BWA (Li, 2013) or Bowtie2 (Langmead et al., 2009) can be used to map the sequence reads to the genome of interest. For RNA-based studies, a gapped aligner such as STAR (Dobin et al., 2013) or HiSAT2 (Kim et al., 2015) is required due to the lack of intronic sequences within the reads and alternative splice patterns.

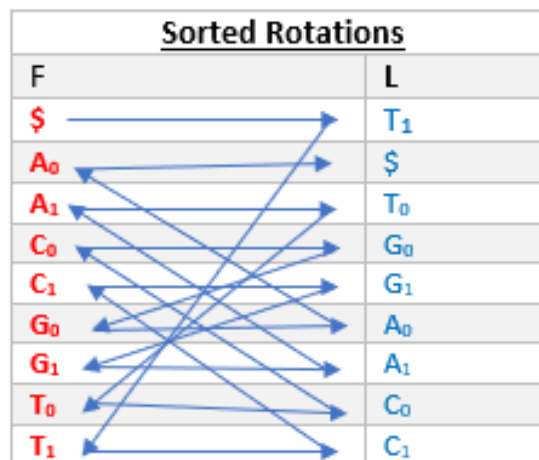
Non-gapped aligners like BWA (Li, 2013) work via the use of an index file of the genome of interest. The index file will be converted into an alternative sequence using the Burrow's Wheeler transformation (BWT). This allows frequently occurring sequence to be transformed/rotated into different character strings, with only the last column of the

transformation being saved due to the reversible nature of the BWT. This equates to a highly efficient method of storing sequence read fragments, allows quick searching through the sequences thanks to the \$ notation. This allows highly repetitive sequences to be aligned to their place in the genome with greater ease. For example, the string 'AGCTAGCT\$' would become 'GCTAGCT\$A' under the first rotation. Then, 'CTAGCT\$AG' on the second rotation, until all rotations of the sequence are calculated. Rotations are then sorted into alphabetical order and the last column of characters taken as the transformation i.e. 'TT\$GGAACC'. This allows the BWT to store 9 sequence rotations which may occur in the genome given the limited characters of A, C, G and T as one character string. The first column of the transformation is also stored to allow the transformation to be reversed.

The occurrence of each A, G, C and T is also indexed in the original sequence allowing the transformation to be reversed. These properties permit the BWT to calculate all possible sequences of a string of characters, enable it to be fast in terms of searching i.e. compare two sequences in accordance to where the dollar sign is, if the dollar sign is not in the same place in each sequence, the sequences don't match and therefore looking at the rest of the sequence does not need to occur (Figure 7). Finally, due to the index and sorting of the rotations of the original string, the BWT is reversible, allowing the original sequence fragment to be obtained and placed in the correct place in accordance to the reference genome (Li and Durbin, 2009).



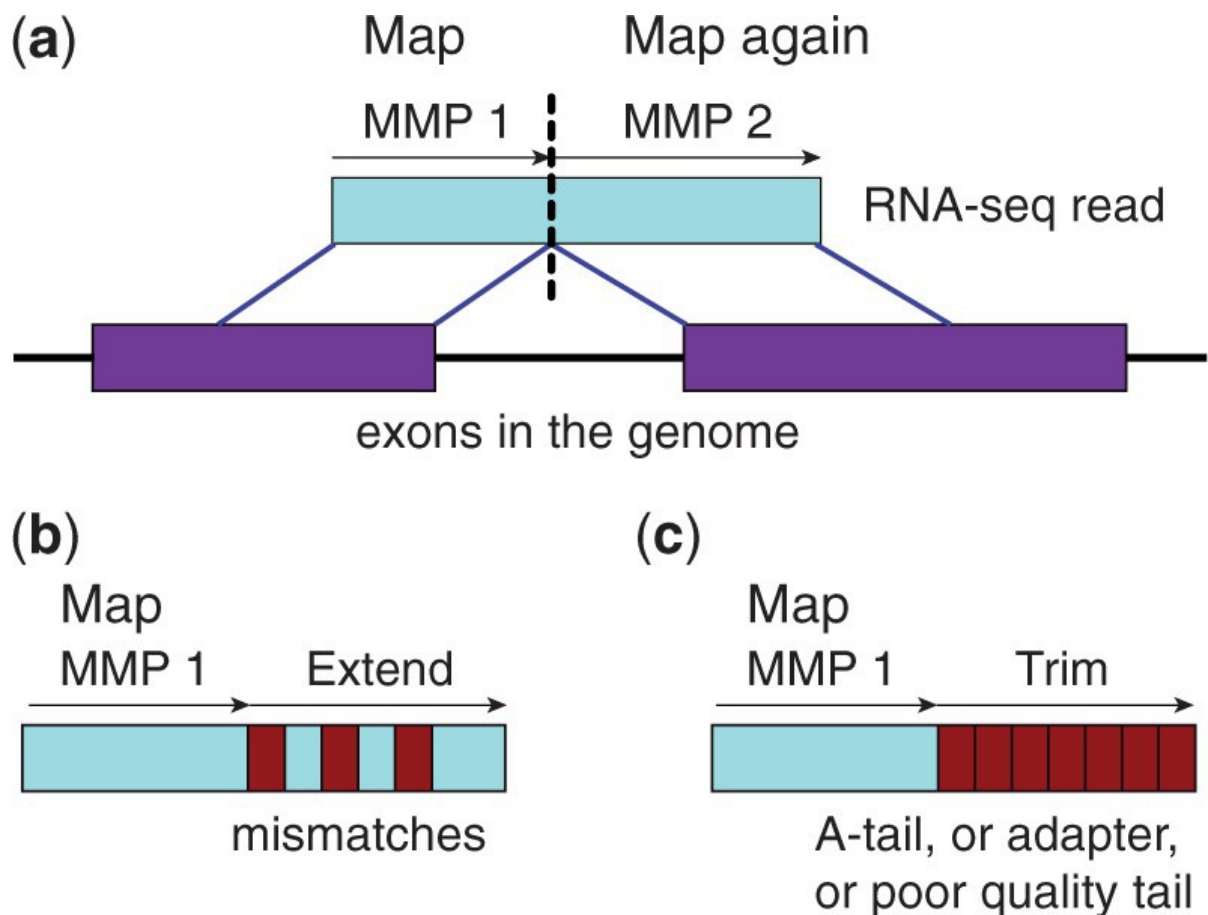
<u>Text:</u>	<u>Rotations:</u>			<u>Sorted Rotations</u>			<u>Transform:</u>
	F		L	F		L	
<b>A<sub>0</sub>G<sub>0</sub>C<sub>0</sub>T<sub>0</sub>A<sub>1</sub>G<sub>1</sub>C<sub>1</sub>T<sub>1</sub>\$</b>	<b>A<sub>0</sub></b>	GCTAGCT	<b>\$</b>	<b>\$</b>	AGCTAGC	<b>T<sub>1</sub></b>	<b>TT\$GGAACC</b>
	<b>G<sub>0</sub></b>	CTAGCT\$	<b>A<sub>0</sub></b>	<b>A<sub>0</sub></b>	GCTAGCT	<b>\$</b>	
	<b>C<sub>0</sub></b>	TAGCT\$A	<b>G<sub>0</sub></b>	<b>A<sub>1</sub></b>	GCT\$AGC	<b>T<sub>0</sub></b>	
	<b>T<sub>0</sub></b>	AGCT\$AG	<b>C<sub>0</sub></b>	<b>C<sub>0</sub></b>	TAGCT\$A	<b>G<sub>0</sub></b>	
	<b>A<sub>1</sub></b>	GCT\$AGC	<b>T<sub>0</sub></b>	<b>C<sub>1</sub></b>	T\$AGCTA	<b>G<sub>1</sub></b>	
	<b>G<sub>1</sub></b>	CT\$AGCT	<b>A<sub>1</sub></b>	<b>G<sub>0</sub></b>	CTAGCT\$	<b>A<sub>0</sub></b>	
	<b>C<sub>1</sub></b>	T\$AGCTA	<b>G<sub>1</sub></b>	<b>G<sub>1</sub></b>	CT\$AGCT	<b>A<sub>1</sub></b>	
	<b>T<sub>1</sub></b>	\$AGCTAG	<b>C<sub>1</sub></b>	<b>T<sub>0</sub></b>	AGCT\$AG	<b>C<sub>0</sub></b>	
	<b>\$</b>	AGCTAGC	<b>T<sub>1</sub></b>	<b>T<sub>1</sub></b>	\$AGCTAG	<b>C<sub>1</sub></b>	



**Figure 7: Simple representation of the Burrow's Wheeler Transformation.** Original text AGCTAGCT, \$ added to mark the end of the character string. Text is rotated one character until the \$ is reached again (Rotations column). Only the first (F) and last (L) columns are saved in the transformation process. The rotations are then sorted alphabetically via the F column, the L column is then taken as the Burrow's Wheeler Transformation of that sequence and therefore stores all possible versions of the original character string. To reverse the process the sorted rotations in the F column then follow the notations in the L column as shown by the arrows and spell out the reverse of the original text \$T<sub>1</sub>C<sub>1</sub>G<sub>1</sub>A<sub>1</sub>T<sub>0</sub>C<sub>0</sub>G<sub>0</sub>A<sub>0</sub>. Subscripts are only for illustrative purposes of the positions of the original characters. [original figure SJ Thursby]

Gapped aligners like STAR (Dobin et al., 2013) work via a different mechanism due to the presence of alternative splice sites. This results in sequences that cannot be mapped contiguously to the genome. To combat this, STAR utilises a maximal mappable prefix (MMP) approach in which it finds the longest mappable sequence from a sequence fragment and aligns that to a donor splice site (figure 8a), these sequences are called seeds. Then, MMP is repeated for all unmapped portions of the read, allowing it to identify mismatches (figure 8b) and poly-A tails (figure 8c) which are marked for extension or as anchor points respectively. STAR uses compressed suffix arrays to elicit binary searches of the genome and thus results in computational efficiency even in large genomes. Following this, all sequence seeds are stitched together within their applicable genomic region and the maximum intron frequency is determined based on the size of that genomic region. Scoring penalties are then applied for mismatches, indels, deletions and splice junction gaps and the highest scoring stitched combination is chosen as the most suitable alignment for that region (Figure 8).

WGBS and RRBS also require specialized aligners due to the methylation quantification step that must be computed prior to alignment. The sequence also must be converted back into unmodified genomic DNA to be aligned to the reference genome (Krueger and Andrews, 2011b).



**Figure 8: STAR gapped alignment Maximum Mappable Prefix (MMP) search.** STAR aligner finds a sequence which matches with the first portion of a gene, but because there is a splice junction, this read cannot map contiguously to that region. STAR aligner then utilises MMP to mark this position as a spot in which this sequence can map to (i.e. a seed, a) which is next to a donor splice site. MMP then goes on to see if there are any alternative sites this sequence could map to and scores these sites, with penalties given for mismatches (b), indels and polyA-tails (c) which can also be found using the MMP principal. Taken from (Dobin et al., 2013)

#### 1.6.3.5 Differential Analysis

After pre-processing and mapping, differential analysis of sequence-based data depends on its origin and the aims of the experiment. For whole genome sequencing, variant calling using GATK equates to differential analysis, this will allow you to see the differences between your samples in terms of SNPs, indels and deletions (McKenna et al., 2010). For RNA-seq, the gene-based frequencies should be quantified using HTseq-count (Anders et al., 2015) and differential analysis computed in DeSeq2 (Love et al., 2014) using a linear model based approach similar to the approach used in array analysis. RNA-seq data can also be normalised to provide a relative score of enrichment which can be applied across every gene examples include; fragments per kilobase of transcript per million mapped reads (FPKM), transcripts per million mapped reads (TPM), and reads per kilobase per million mapped reads (RPKM), as reviewed in (Evans et al., 2018). However, such techniques should not be used to compare across samples as they are relative to that sample and not all samples.

Differential analysis of WGBS or RRBS is computed using Metilene, this segments the genome then uses a scoring approach and a 2D Kolmogorov-Smirnov test to assess the differences between genomic regions of different samples (Jühling et al., 2016).

For ChIP-seq data, differential analysis is computed via peak calling through MACS2 (Gaspar, 2018a). Peak calling will depend on whether the enriched protein is a histone or not.

Histones, due to their diverse ranges of action, require broad peak calling to elucidate enrichment, proteins with more specific binding sites can be quantified via narrow peak calling to reveal regions of enrichment. The differences in peak calling procedures are due to variable parsing of background noise – this is harder to distinguish with histones so alternative peak calling procedures are required (Feng et al., 2012; Gaspar, 2018a; Landt et

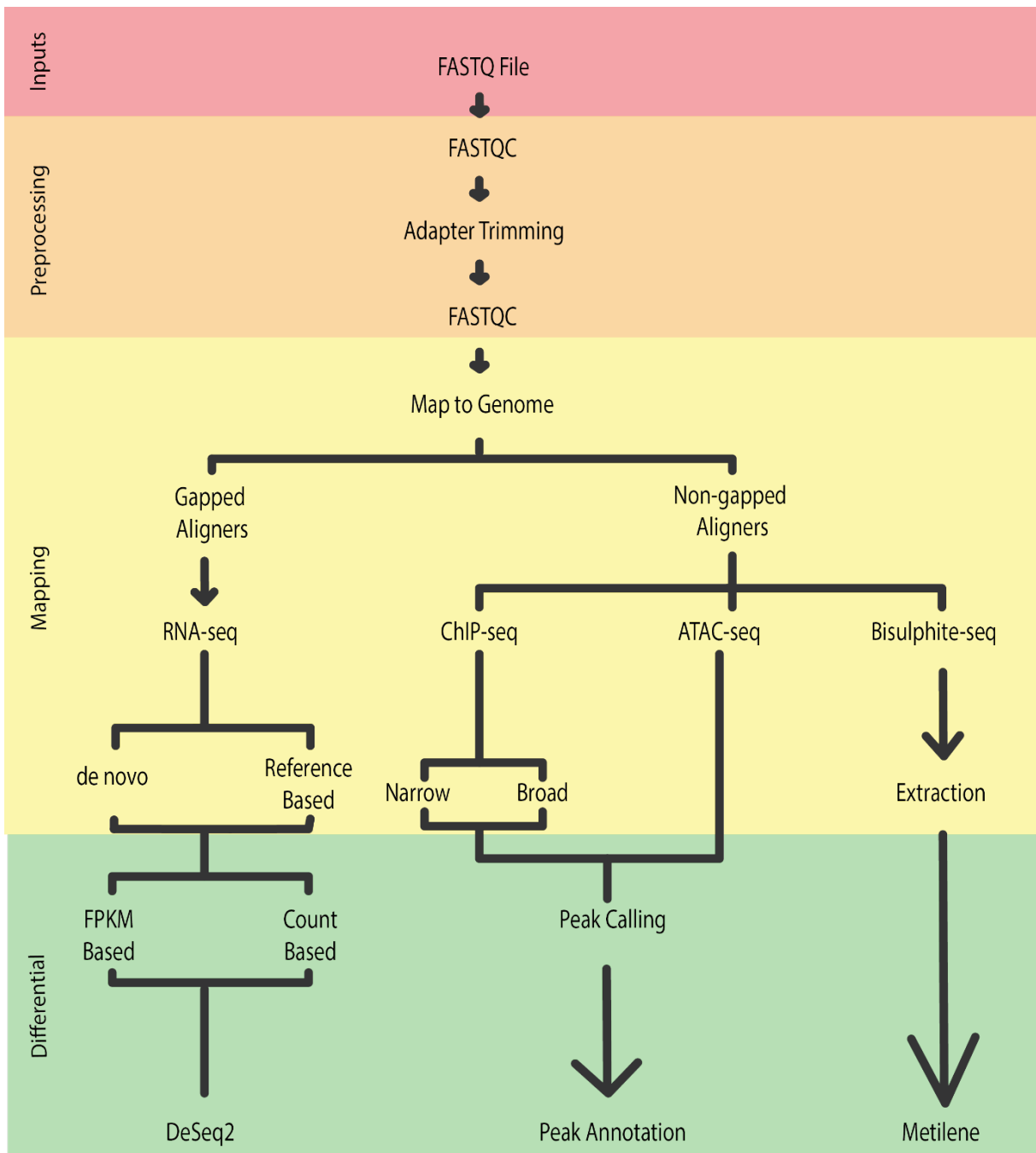
al., 2012). Further downstream processing can be computed following differential analysis, such as chromatin state segmentation algorithms, like CHMM (Ernst and Kellis, 2012). This software calculates the frequency of histone marks that occur through a region and from this determines the characteristic of that region, such as, active promoter, polycomb repressed and others (figure 9). CHIP-seq data from the ENCODE project (Ernst et al., 2011) that has been processed through this software has been in used in most investigations presented in this thesis.

Differential Analysis of ATAC-seq data also utilises peak calling but using the Genrich software. Peak calling here is needed where the Tn5 transposase cuts, and not directly on the nucleosome itself, to see the most accurate view of the chromatin accessible sites and aid in the prediction of promoter and enhancer regions (Gaspar, 2018b). An overview of common sequence-based processing techniques can be found in figure 10.

Chromatin States	State	CTCF	H3K27me3	H3K36me3	H4K20me1	H3K4me1	H3K4me2	H3K4me3	H3K27ac	H3K9ac	WCE	Candidate state annotation
	1	16	2	2	6	17	93	99	96	98	2	
2	12	2	6	9	53	94	95	14	44	1	Weak Promoter	
3	13	72	0	9	48	78	49	1	10	1	Inactive/poised Promoter	
4	11	1	15	11	96	99	75	97	86	4	Strong enhancer	
5	5	0	10	3	88	57	5	84	25	1	Strong enhancer	
6	7	1	1	3	58	75	8	6	5	1	Weak/poised enhancer	
7	2	1	2	1	56	3	0	6	2	1	Weak/poised enhancer	
8	92	2	1	3	6	3	0	0	1	1	Insulator	
9	5	0	43	43	37	11	2	9	4	1	Transcriptional transition	
10	1	0	47	3	0	0	0	0	0	1	Transcriptional elongation	
11	0	0	3	2	0	0	0	0	0	0	Weak transcribed	
12	1	27	0	2	0	0	0	0	0	0	Polycomb-repressed	
13	0	0	0	0	0	0	0	0	0	0	Heterochrom; low signal	
14	22	28	19	41	6	5	26	5	13	37	Repetitive/CNV	
15	85	85	91	88	76	77	91	73	85	78	Repetitive/CNV	

Chromatin Mark Observation Frequency (%)

**Figure 9: Frequency calculations in the CHMM algorithm.** Chromatin state segmentation algorithms such as CHMM (Ernst and Kellis, 2012) utilise ChIP-seq data to calculate the frequency of different histone marks in different genomic regions. These programs can then provide an estimate with regards to the chromatin state of that region i.e. this region correlates with active promoter histone marks (H3K4me2/me3, H3K27ac and H3K9ac). WCE indicates whole cell extract and is used as a control. Diagram adapted from (Ernst et al., 2011)



**Figure 10: Overview of sequence-based processing methods.** All sequence-based analysis methods are output in the form of a FASTQ file (inputs) these then have to be quality checked and have their sequencing adapters removed (pre-processing) prior to mapping to the genome (mapping). Different programs are required for genome alignment depending on the type of sequence data, for example, RNA-seq requires a gapped aligner due to splice junctions and the lack of introns in this sequence type. Most alternative sequences can be mapped to the genome using non-gapped aligners. After mapping to the genome, differential analysis can be carried out. Each sequence type has their own specified program for differential analysis. Programs listed are examples of popular tools used in the analysis of this sequence type, alternative programs also exist. [original figure SJ Thursby].

#### 1.6.4 UCSC Genome Browser

The UCSC Genome Browser (<https://genome.ucsc.edu.com>) was created to view sequence data in a more effective manner than BLAST or similar formats. It is a freely- available bioinformatics resource that provides an intuitive map-based feature for viewing human and other genomic builds from the results of the NCBI Reference Sequence and GenBank databases. Using UCSC Genome Browser, users can type in a gene of interest, view its structure and zoom into and out of the base sequence of that chromosome. Users can also load custom or freely-available data on to the genome browser and superimpose it (in the form of lines known as tracks) onto the genome build of interest to aid in their loci-specific investigations. Examples of such tracks include: CpG island location, SNP locations and chromatin segmentation tracks. Custom data can be uploaded on the genome browser via file upload or via bioinformatics interfaces such as Galaxy (<https://www.usegalaxy.org>). For example, utilising browser extendible data (BED) format (chromosome, start, end coordinates and a label) allows the user to generate a track to be viewed on the genome browser: visualization parameters including track name and colour can also be specified within the track header. However, sequence alignment maps, wiggle files and many other formats are also supported by the browser.

Alternative genome browsers such as Integrative Genomics Browser also exist. This browser was originally a desktop application before the developers provided a web-based platform. It is useful when viewing sequence transcripts as it displays each individual sequence fragment onto the browser from SAM files so differences can be visually assessed as well as computationally. However, it is harder to work than UCSC Genome Browser and the desktop version can be slower. For these reasons and the fact research in this thesis does not focus



on sequence heterogeneity, UCSC Genome Browser was used for all subsequent research (Robinson et al., 2011).

#### 1.6.5 Galaxy Bioinformatics Interface

Galaxy is a web-based environment which provides user-friendly computational architecture for the analysis of sequencing data without the need for programming, command line knowledge or high specification computational architecture. In addition, it allows the user to visualise their data in one of the many available graphics plug-ins or via various genome browsers, like UCSC Genome Browser. Galaxy is compatible with a host of file formats and supports the analysis of the most popular sequencing technologies. Data can also be imported into Galaxy via the UCSC table browser, EBI SRA or NCBI (Børnich et al., 2016).

Galaxy utilises a traffic light-based system indicating the status of job progress; grey – queued, yellow - in progress, green – complete. All jobs will appear in the history column of Galaxy at the right-hand side (RHS) of the browser window. One of the most useful features within Galaxy is the ability to create workflows – user friendly multi-step processes which allow the automation of repetitive tasks, like custom functions in traditional programming languages. Only output processes not hidden will show up in the Galaxy History. Workflows can also be extracted from already-conducted history jobs. The functions of Galaxy extend far beyond the above points and more can be found for the interested reader in the following references (Afgan et al., 2016; Giardine et al., 2005; Thiel, 2016).

#### 1.6.6 Database for Annotation, Visualization and Integrated Discovery

The Database for Annotation, Visualization and Integrated Discovery (DAVID) provides a free- to-use online platform for the functional analysis of genes and results from genome-wide assessment studies (<https://david.ncifcrf.gov/home.jsp>). It provides enrichment analysis across over 40 different categorised platforms including Gene Ontology (GO), KEGG

and REACTOME. The main functions of DAVID can be split into functional annotation summary, functional annotation clustering and Gene ID conversion. Functional annotation summary provides a method of converting the gene enrichment list into biological categories of interest, it also provides a modified (more conservative) Fisher's exact p value of the significance of the enrichment: as with many statistical tests, a smaller p value indicates a more significant result. A more recently-developed component is the Functional annotation clustering function, which groups the categories from the functional annotation summary and clusters them into groups of similar biological meaning (Huang et al., 2009, 2007; Huang da et al., 2009): thus, if genes are turning up under different categories such as tissue-specificity or cellular localisation it will recognise this and bring them under a single super-heading, making it easier to see overall patterns and significance. This function was particularly useful in one of the MS contained within this thesis (Paper II).

## 1.7 Mechanistic Studies

### 1.7.1 Cellular Machinery

Cellular machinery in the context of DNA methylation is complex and involves a variety of proteins, many mentioned in section 1.4. These proteins can be divided into those that read, those that write and those that erase DNA methylation.

Reader proteins such as methyl-binding domains read 5mC and have been found to regulate transcription. Writer proteins involve those that aid in the establishment of 5mC such as DNA methyltransferase (DNMT) enzymes, which add 5mC to naked cytosine at the replication fork or via *de novo* mechanisms. To date, there are 3 known catalytically active DNMTs in human: one maintenance methyltransferase, DNMT1, that prefers hemimethylated DNA (Brown and Robertson, 2007) and two *de novo* methyltransferases DNMT3A and DNMT3B, that add 5mC to unmodified naked cytosine. However the activity of

the *de novo* methyltransferases appears dependent upon the activity of their co-factor DNMT3L in many tissues (Bestor, 2000; Gowher and Jeltsch, 2018). Reports also suggest that DNMT3B may also work in conjunction with DNMT1 at the replication fork, similar to that of an auditor. DNMT2 is also known to exist but the literature indicates this is a tRNA methyltransferase (Goll et al., 2006; Lyko, 2017).

Eraser proteins aid in the removal of 5mC from methylated cytosines, such as the TET proteins which convert 5mC to 5hmC. 5hmC is then converted into 5-formylcytosine and 5-carboxylcytosine and converted back into unmodified cytosine via excision using thymine DNA glycosylase and re-synthesis using a repair polymerase (Zhang et al., 2017).

### 1.7.2 DNMT1

The maintenance methyltransferase DNMT1 was the first methyltransferase identified in mammals in 1988 due to its resemblance to its bacterial counterpart (Bestor, 1988). It was later found to be responsible for the maintenance of methylation marks preferentially at hemi-methylated DNA. During cell division, DNMT1 methylates the daughter strand using the universal carbon donor S-adenosylmethionine (Hermann et al., 2004).

In addition to maintenance methylation, DNMT1 has been associated with DNA repair pathways (Inano et al., 2000; Loughery et al., 2011). In proliferating cells, DNMT1 is ubiquitously expressed due to the role of DNMT1 in maintenance methylation. However, in post-mitotic neurons, DNMT1 is also highly expressed, but KO of DNMT1 in such cells does not affect DNA methylation (Fan et al., 2001). With this and other studies in mind, it is thought that DNMT1 may maintain methylation after DNA repair (Chuang et al., 1997; Ha et al., 2011; Mortusewicz et al., 2005).

### 1.7.3 UHRF1

In order to maintain methylation at the hemi-methylated daughter strands following cell division, DNMT1 must be guided to the newly-synthesised strand. This is partly due to the PCNA-interaction domain on the protein, but may also be facilitated via interaction with another co-factor, ubiquitin-like PHD and RING finger domain-containing protein 1 (UHRF1), also known as NP95, since Uhrf1/Np95 knockout in mouse results in a failure of DNMT1 to localise to the nucleus (Bostick et al., 2007; Sharif et al., 2007). UHRF1 is thought to guide DNMT1 to the replication fork via recognition of hemi-methylated DNA through its SRA domain (Avvakumov et al., 2008; Bostick et al., 2007; Sharif et al., 2007), this involves a novel 'flip-out' mechanism to stabilise the interaction between the DNA and UHRF1 SRA domain (Arita et al., 2008). The RING finger domain, with its E3 ubiquitin ligase activity, modifies histone 3 which is recognised by DNMT1 and aids in its recruitment to the replication foci (Berkyurek et al., 2014; Liu et al., 2013; Rothbart et al., 2012). The PHD and TTD domains then recognise unmodified arginine 2, unmodified lysine 4 and H3K9me2/3 - this aids the cementing of DNMT1 onto the correct genomic loci (Cheng et al., 2013; Foster et al., 2018; Hu et al., 2011; Rajakumara et al., 2011; C. Wang et al., 2011). Alternatively, Rothbart and colleagues (2012) proposed that H3K9me3 binding keeps DNMT1 attached to heterochromatin regions when replication was not occurring.

### 1.7.4 Hypomorphic States

DNMT1 is essential for viable embryo development. Embryos that have undergone a DNMT1 KO do not survive past embryonic day 6.5 as a result of substantial global DNA methylation loss. However, KO of DNMT1 in ESC cells does not result in lethality. DNMT1 KO ESCs show the same global loss of methylation but retain their proliferative abilities. However, when differentiation is induced these ESCs trigger the DNA damage response and

undergo apoptosis (Liao et al., 2015), as for cancer cells with an inducible KO (T. Chen et al., 2003). Tissue-specific KO of DNMT1 are consistent with a cell-autonomous cell death response: in neuroblasts for example, KO results in offspring death after birth due to respiratory difficulties, presumably due to the absence of crucial neural signals to initiate breathing (Fan et al., 2001) and in foetal pancreatic cells, there is a reduction in differentiation and an increase in p53 levels as seen in ESC and cancer cells with KO (Georgia et al., 2013). Providing evidence for the necessity of DNMT1 in genomic stability and differentiation, and a maintenance role of DNMT3b.

However, hypomorphic levels of DNMT1 do not result in lethality, although at least a 20% level of DNMT1 (truncated or not) must remain in the cells to ensure a lethal phenotype does not result (Gaudet, 2003). In adult non-cancerous cells, immortalised via the overexpression of the telomerase enzyme, a stable KD of DNMT1 in human fibroblasts (hTERT-1604) was established without resulting in lethality by the Walsh lab (K.M. O'Neill et al., 2018; Ouellette et al., 2000) – work with these fibroblasts and this model system make up the majority of the cellular work within this thesis.

Similar to DNMT1, UHRF1 KO also results in embryonic lethality but not in ESCs. Conditional KO of UHRF1 in oocytes of mice also results in lethality during the blastocyst stage of embryonic development (Maenohara et al., 2017; Sharif et al., 2016, 2007). There have been a range of studies looking at functional consequences of mutations in the gene, with quite diverse results. Mesenchymal-specific UHRF1 KO mice resulted in morphological abnormalities due to dysregulation in proliferative and differentiation abilities (Yamashita et al., 2018). In cancer cells, hypomorphic levels of UHRF1 resulted in abrupt cell cycle arrest in breast cancer cells (X. Li et al., 2011), apoptosis and increased sensitivity to DNA damage in

HCT116 cells (Arima et al., 2004; Tien et al., 2011). While these provided evidence for the importance of UHRF1 in maintenance methylation, cell cycle regulation and differentiation, the lack of identification of a single primary function for the protein was part of the motivation for the work carried out as part of Paper V in this thesis.

#### *1.7.5 Interaction with Polycomb*

First identified as regulators of the *Hox* gene cluster in *Drosophila* (Jürgens, 1985) and later in eukaryotes (Kuzmichev et al., 2002), the Polycomb repressive complex are transcription and chromatin regulatory factors composed of multi-domain binding proteins. The most well-characterised of these proteins are Polycomb Repressive Complex 1 (PRC1) and Polycomb Repressive Complex 2 (PRC2).

PRC1 is composed of a RING1 protein, one of the polycomb group ring finger 1-6 proteins, and a RANUL protein. It ubiquitinates H2A on lysine 119 to form H2AK119ub1 and can compact chromatin independent of ubiquitination via recognition of H3K27me3 and interaction with nucleosomes (Chittock et al., 2017; Ku et al., 2008).

PRC2 is composed of 3 main components, a histone methyltransferase Enhancer of Zeste 2 (EZH2), the embryonic ectoderm development (Eed) protein and suppressor of zeste 12 (Suz12). It can also bind with various additional subunits, including nucleosome remodelling factors (Nurf55), to function as one complex. PRC2 functions to deposit mono-, di- and trimethyl groups onto H3K27 via the SET domain of EZH2 (Ciferri et al., 2012; Laugesen et al., 2016; Reddington et al., 2013; Viré et al., 2006). In studies, this histone mark was diluted via cell division and recognised via PRC1, which then induced more compacted chromatin (Blackledge et al., 2014; Reddington et al., 2013). Following this, PRC2 interacted with TET1 to block cytosine methylation of the newly polycomb-repressed site. However,

PRC2 binding has been found to depend on CpG density. In CpG- rich regions, there is a negative correlation between DNA methylation and H3K27me3 observed – most likely to due 5mC inhibiting PRC2 binding. In CpG-poor regions, similar levels of 5mC and H3K27me3 have been reported (Liu et al., 2015). Transcription factors such as OCT4 and SOX2 have also been implicated in PRC2 recruitment (Holoch and Margueron, 2017). In addition to this, the histone methyltransferase subunit of PRC2, EZH2, has been found interacting with DNMT3A and DNMT3B *in vitro*. Binding of the DNMT enzymes at polycomb-repressed sites was also dependent on the presence on EZH2, resulting in EZH2 being reported to regulate DNA methylation at certain areas (Viré et al., 2006).

In the absence of 5mC, H3K27me3 has been found to invade neighbouring regions, which it cannot usually bind to (Reddington et al., 2013). This results in a dilution of binding of PRC1 & 2 at primarily polycomb-repressed sites as observed in hypomethylated mouse ESCs (Reddington et al., 2013). Erasure of core PRC2 subunits in mice results in embryonic lethality, similar to that of DNMT1 or UHRF1 (O'Carroll et al., 2001). KO of *Suz12* or *Eed*, which code for subunits of PRC2, also leads to abnormalities in the formation of hematopoietic cells, indicating a crucial role for PRC2 in development (Faust et al., 1995; Pasini et al., 2004). Furthermore EZH2, which is highly expressed in ESCs and proliferating cells, has been found to be overexpressed in certain tumour types, identifying PRC2 and EZH2 as potential oncogenic biomarkers (Kawano et al., 2016; Moritz and Trievel, 2018).

## 1.8 Epidemiology Applications

### 1.8.1 Dietary Intervention

#### 1.8.1.1 The Barker Hypothesis

Approximately 40 years ago, an investigation was published linking foetal malnutrition during gestation and coronary heart disease in later life (Barker and Osmond, 1986).

Following this, many additional investigations published similar results, indicating a link between the foetal environment and later life disease states (Brown et al., 1995; Centers for Disease Control, 1992; Czeizel and Dudás, 1993; Department of Health, 1992; Jacob et al., 1998; MRC Vitamin Study Research Group, 1991; Sohn et al., 2003; Stanner and Yudkin, 2001). This link and subsequent studies became known as the foetal origins of adult disease (FOAD) hypothesis or the Barker hypothesis and suggested that multiple chronic illnesses, including Diabetes Mellitus (Liu et al., 2018), may be the result of suboptimal foetal environments, e.g. periods of severely restricted calorie intake, which result in developmental plasticity to aid and promote survival (Barker, 2004; Barker and Osmond, 1986).

The Barker hypothesis states that maternal nutrient alterations will elicit alterations in the epigenome of the child (Barker, 2004). Recent research has also highlighted that this is particularly true in relation to folic acid supplementation during pregnancy (Irwin et al., 2019, 2016; McGarel et al., 2017; McNulty et al., 2011).

Studies of the Dutch Famine Winter (1944-1945) revealed that calorie- and nutrient-deficient status for mothers during late gestation resulted in low birth weight children. These children exhibited indications of an altered methylome and it was suspected that this predisposed them to coronary artery disease and insulin resistance later in their lives (Schulz, 2010; Stein et al., 2004).

However mothers of the Leningrad siege, who had undergone similar calorie and nutrient restrictions for almost twice as long as that of the Dutch Famine Winter, did not show any indications of low birth weight or potentially disadvantageous changes in the methylome (Stanner and Yudkin, 2001; Tobi et al., 2015, 2009). The results of these two studies are

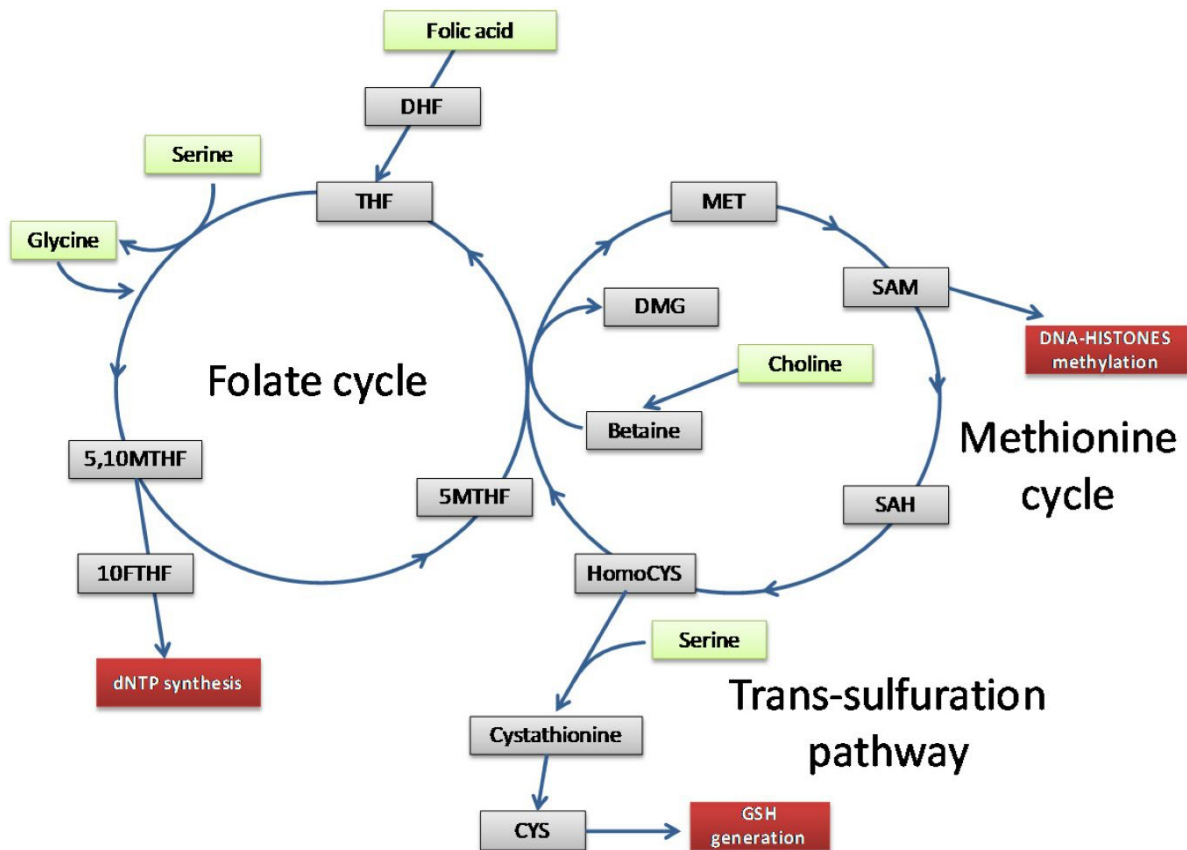


significantly different, but they and similar studies that have been influenced by the Barker Hypothesis serve as an indication of the current interest in the effects of maternal dietary variation on offspring birth weight and later life disease states, both in first and subsequent generations (Brown et al., 1995; Lumey et al., 2012; Tobi et al., 2015, 2009).

#### 1.8.1.2 One Carbon Metabolism

One possible mechanism by which dietary interventions might cause changes in the methylome are through alterations in one carbon metabolism. For the establishment and maintenance of DNA methylation, a methyl group is required to enable DNMTs to methylate an unmodified cytosine base. The universal carbon donor S-adenosylmethionine (SAM) provides the carbon for this modification and hails from the metabolism of folate and other micronutrients including methionine, vitamin B12 and choline. However, the human body cannot produce these substances *de novo*, therefore they are taken from the diet. Since DNA methylation and DNA synthesis are highly active processes during gestation this results in folic acid becoming a limiting factor during this period (Farias et al., 2015).

In one carbon metabolism (figure 11), the addition of a carbon molecule to dietary folate from serine or glycine results in the formation of tetrahydrofolate (THF). Further addition of a methyl group to THF results in the formation of 5-methyltetrahydrofolate (5-MTHF). Following a vitamin B12-catalysed reaction with homocysteine, the 5-MTHF forms methionine, the majority of which is converted to SAM, the universal carbon donor (Clare et al., 2019; Irwin et al., 2016).



**Figure 11: Folic acid and one carbon metabolism.** Folic acid is converted to dihydrofolate (DHF) then tetrahydrofolate (THF) and following the conversion of Serine to Glycine, THF is converted to 5,10-methyltetrahydrofolate (5,10MTHF). A portion of this 5,10MTHF is converted to 10-formyltetrahydrofolate (10FTHF) to be used in the formation of deoxyribonucleotide triphosphate (dNTP). The remainder is then converted into 5-methyltetrahydrofolate (5MTHF) and following a conversion of Betaine to dimethylglycine (DMG) the 5MTHF is converted to methionine (MET) and then S-adenosylmethionone (SAM) i.e. the universal carbon donor. SAM can then be utilised in DNA methylation establishment or converted to S-adenosylhomocysteine (SAH) and then homocysteine (HomoCYS) to be used in the Trans-sulfuration pathway. Here, HOMO CYS is converted into Cystathionine, via the addition of Serine, and then to cysteine (CYS) to be used in glutathione (GSH) generation. Image taken from (Rizzo et al., 2018).

### 1.8.1.3 Suboptimal Folate Levels & DNA Methylation-associated Disease

#### 1.8.1.3.1 The Role of Folate in NTD Prevention

During the first trimester, DNA synthesis, cell division and growth are the fundamentals of this developmental period (Yiu and Li, 2015). Therefore it is acknowledged that sufficient *in utero* folate is essential to the correct closure of the neural tube and brain development (Czeizel and Dudás, 1993). Incorrect closure of the neural tube can result in spina bifida or anencephaly (van Gool et al., 2018).

During the second and third trimester DNA methylation is being established in parallel to neurological development. Although, under current guidelines folic acid supplementation is not recommended during the second and third trimester. This is a period of high carbon donor requirement. This could affect the establishment of epigenetic marks or lead to restrictions in neurological development (Irwin et al., 2016; McGarel et al., 2017; McNulty et al., 2011; Pentieva et al., 2012).

#### 1.8.1.3.2 Maternal Folate Supplementation & DNA Methylation in Offspring

While the Dutch famine Winter and the Leningrad siege do provide epidemiological evidence for the Barker hypothesis, recent studies have been more controlled/regimented with investigations into the supply of carbon donor rich foods or DNA methylation of the offspring designed to attempt to elaborate on the mechanistic side of this hypothesis.

In the variable yellow agouti mouse ( $A^{vy}/a$ ) model, supplementation of micronutrients related to one carbon metabolism during pregnancy led to alterations expression of the  $A^{vy}$  gene in the offspring, causing alterations in body mass and a change in the coat colour of the mice from yellow to brown (Wolff et al., 1998). Upon further investigation, brown mice demonstrated reduced levels of obesity, healthier blood pressure, and lower risk of insulin resistance and tumour development (Wolff et al., 1987; Yen et al., 1994). These changes

were directly related to changes in the methylation of a newly-arisen regulatory region of the gene, which turned out to be a de novo insertion of an endogenous IAP retrovirus. Read-through transcription from the IAP LTR was driving ectopic transcription of the Agouti gene, leading to obesity and yellow colour: silencing of the IAP was associated with methylation of the LTR and concurrent reversion of the Agouti gene to its normal mode of regulation and transcription. While a fascinating case study, ERV-driven endogenous genes are very rare and such metastable epialleles may represent an evolutionary oddity. However, it did support the theory that the mother's one carbon nutritional status can affect the epigenome of the offspring.

In a study of pregnant mothers and offspring in rural Gambia, the season of conception was shown to alter DNA methylation at 9 endogenous genes which showed signs of being metastatic epialleles. DNA methylation increased for those born within the rainy season, a time of nutritional hardship affecting one carbon levels, in comparison to those born outside of that season (Waterland et al., 2010). A similar randomised control trial of the seasonal effects of conception in rural Gambia also noted sex-specific effect of one carbon donor supplementation. Decreases in methylation at the imprint-related *IGF2* locus was found in females and the same for *GTL2* in males (Cooper et al., 2014). An alternative study into folic acid supplementation during the periconception growth period demonstrated epigenetic plasticity in the imprinted gene *IGF2* within the offspring, dependent on whether the mother had had supplementary folic acid (Steeegers-Theunissen et al., 2009).

In the Aberdeen Folic Acid Supplementation Trial, blood samples had initially been collected at childbirth (as detailed in, Charles et al., 2005), then analysed for DNA methylation and saliva obtained as a follow-up 47 years after the intervention. Results from the follow-up samples found a

dose-responsive reduction in the DNA methylation of *PDGFRA* (one singular Illumina array probe only), a gene related to the occurrence of NTDs. In addition to this, 46 regions were highlighted as differentially methylated between placebo and treatment, including members of the HLA cluster and regulators of embryonic development (*PAX8*) (Cheung et al., 2003; Richmond et al., 2018).

As these studies had a number of limitations, an in-house randomised control trial addressing some of these concerns (Pentieva et al., 2012) was conducted to assess the effect of folate acid supplementation in the second and third trimester (FASSTT) of pregnancy. In particular, the study was a Randomized Controlled Trial and as such, was designed to directly test the effects of presence or absence of the nutrient, unlike observational studies such as those in the Gambia, and to do so in shorter time periods than in the Aberdeen study. Preliminary results were indicative of a reduction in plasma homocysteine, a hormone related to premature delivery, pre-eclampsia and low birth weight (Wang et al., 2015). When this study was accessed in a follow-up trial, the offspring of placebo group mothers demonstrated restricted neurodevelopment and indications of an altered methylome and transcriptome (Caffrey et al., 2018). It was suggested this was related to nutritional status during pregnancy and the subsequent effects on early life development (Irwin et al., 2016). Following assessment at an early post-natal stage, the offspring of treatment mothers demonstrated improved cognitive development (Pentieva et al., 2012), with similar affects observed again at age six (McGarel et al., 2017). A further follow-up investigation of the DNA methylation of the children at birth was carried out and is the subject of Paper III in this thesis.

### 1.8.1.3.3 In-Utero Exposure to Cigarette Smoke

In utero exposure to cigarette smoke has been associated with miscarriage, low birth weight, and developmental difficulties like congenital heart defects (Alberg et al., 2014; Alverson et al., 2011; Blohm et al., 2008). Effects of smoking have also been seen in later generations and reports indicate this effect may have an epigenetic component (Magnus et al., 2015; Rehan et al., 2013; Spindel and McEvoy, 2016). A study of cigarette smoke exposure in pregnant mice noted global DNA methylation alterations in addition to an upregulation in the expression of inflammatory cytokines like *ERK1* in the offspring (Chen et al., 2018). Moreover, in human over 6000 CpG sites were identified in a meta-analysis as being differentially methylated in the offspring of smoking mothers. Alterations at some of these sites, such as *BMP4* (lung development) or *PRDM8* (neurological development), were also observed in follow-up studies of the offspring many years later (Joubert et al., 2016). In a longitudinal study of prenatal exposure to cigarette smoke, *MYO1G* and *CNTNAP2* were found to be differentially methylated at birth, during childhood, and 17 years later in adolescence, adding further to the evidence for the strong effects of in utero exposure to cigarette smoke (Lee et al., 2015; Richmond et al., 2015).

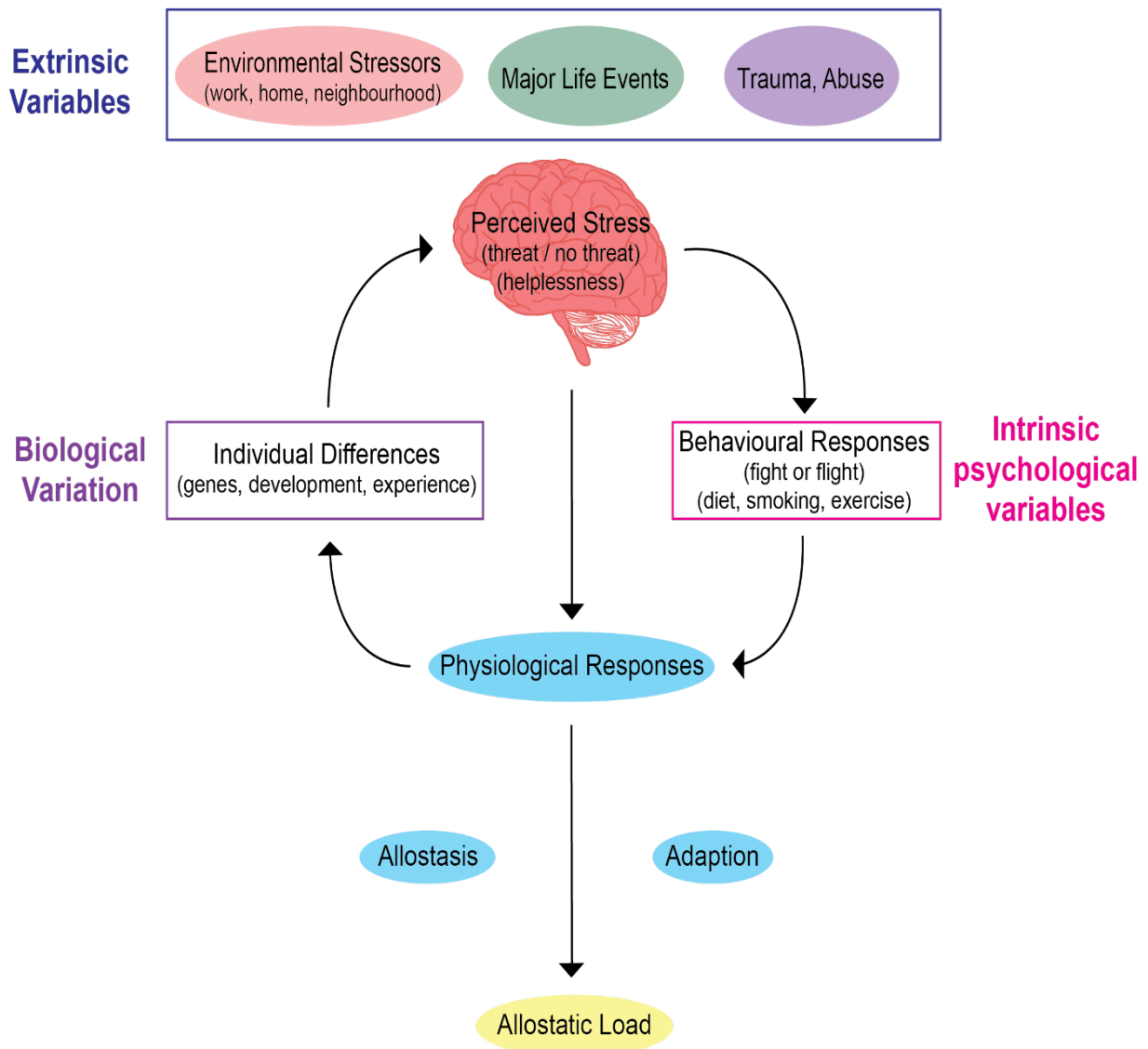
## 1.8.2 Mental Health

### 1.8.2.1 Intrinsic, Extrinsic Variables & Allostatic Load

It has been noted that genetic factors influence gene expression but extrinsic variables such as major life events, trauma or abuse may also lead to alterations in transcription, potentially due to changes in methylation. Such changes could lead for example to the development of a mental illness like depression (Boyle et al., 2005; Labonté et al., 2013). This effect is known as allostatic load (figure 12) and is defined as chronic or long-term exposure to stress, which leads to dysregulation of stress response systems i.e. the hypothalamic-pituitary adrenal axis

(HPA) and eventual disadvantageous effects on the brain and body (Juruena et al., 2018; Oberlander et al., 2008; Todkar et al., 2016).

## Allostatic load model of stress



**Figure 12: Allostatic load model of stress.** Long term exposure to traumatic events or high stress situations (Extrinsic variables) can lead to dysregulation of behavioural responses (Intrinsic psychological variables). These changes are dependent on individual genetic differences (biological variation) but long-term exposure may lead to changes in physiological responses such as the stress response. These changes are termed allostatic load and can have disadvantageous effects on the body. Adapted from (McEwen et al., 2015)



### 1.8.2.2 Glucocorticoids and the HPA Axis

Glucocorticoids such as Cortisol are released following HPA activation and seek to adapt physiological systems to potentially hazardous external stimulus, while also inhibiting further HPA axis stimulation via negative feedback. To achieve negative feedback, cortisol binds to two types of receptors; mineralcorticoid receptors (MR) and glucocorticoid receptors (GR), with greater affinity to the former (Burford et al., 2017; Labonte et al., 2012). MR are involved in regulating the normal concentrations of cortisol found within blood. However when cortisol concentrations are high as a result of repeated HPA axis activation, cortisol binds to GR to discontinue its production and cease the stress response (Casavant et al., 2019).

It is the determinants of HPA axis stimulation which are thought to cause dysregulation e.g. traumatic childhood events, genetic profile, or current stress levels (Burford et al., 2017). Continuous exposure to cortisol can be extremely hazardous to the body as it alters glucose, fat and protein metabolism, in addition to altering immune sensitivity and blood pressure. Some studies also suggest (Ising et al., 2008) that such extreme exposure to cortisol may alter the epigenetic profile of certain genes including *FKBP5* – which regulates the affinity to which cortisol can bind to GR and extinguish the stress response. Upregulation of this gene may lead to alterations in behaviour, which correlate with symptoms of anxiety, depression and many other mental illnesses (Mulder et al., 2017; Paquette et al., 2014).

### 1.8.2.3 HPA Overstimulation and Chronic Stress

Overstimulation of the HPA axis has been linked to suicide in previous studies (Labonté et al., 2012; Labonté et al., 2013). In a post-mortem investigation of the prefrontal cortex of 53 major depression-diagnosed suicide completers, considerable hypermethylation was observed at some CpG sites in comparison to non-psychiatric controls (Haghighi et al., 2014). Aberrant hippocampal DNA methylation has also been observed in suicide completers that

were victims of childhood abuse, in addition to decreased levels of hippocampal glucocorticoid receptor activity, which as mentioned regulates HPA axis activation (Boyle et al., 2005).

## 1.9 Conclusion

DNA methylation plays a major role in regulating gene expression in development and disease, accomplishing this goal in collaboration with histone modifications and the polycomb complex, as discussed above. In this thesis, I describe cellular, enzymatic and human interventions affecting DNA methylation, and try to establish the effects of these alterations on genome-wide methylation, developing new tools for analysis in the process.

This thesis is composed of a number of papers and manuscripts which I was an important contributor to: I will preface each with a brief statement of my role and in the General Discussion expand a little on this and explain how the transition from one paper to the next occurred.

## 1.10 Bibliography

- Adli, M., Zhu, J., Bernstein, B.E., 2010. Genome-wide chromatin maps derived from limited numbers of hematopoietic progenitors. *Nat. Methods* 7, 615–618.
- Afgan, E., Baker, D., van den Beek, M., Blankenberg, D., Bouvier, D., Čech, M., Chilton, J., Clements, D., Coraor, N., Eberhard, C., Grüning, B., Guerler, A., Hillman-Jackson, J., Von Kuster, G., Rasche, E., Soranzo, N., Turaga, N., Taylor, J., Nekrutenko, A., Goecks, J., B., G., D., B., J., G., T., O., K., W., M., R., C.A., M., F., C., M., G., J., S., J., I., D., B., P.A., K., M.L., S., D., B., E., A., L.D., P., J., H., J., G., 2016. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2016 update. *Nucleic Acids Res.* 44, W3–W10.
- Ahn, J.W., Coldwell, M., Bint, S., Ogilvie, C.M., 2015. Array comparative genomic hybridization (Array CGH) for detection of genomic copy number variants. *J. Vis. Exp.*
- Alaskhar Alhamwe, B., Khalaila, R., Wolf, J., Bülow, V., Harb, H., Alhamdan, F., Hii, C.S., Prescott, S.L., Ferrante, A., Renz, H., Garn, H., Potaczek, D.P., 2018. Histone modifications and their role in epigenetics of atopy and allergic diseases. *Allergy, Asthma Clin. Immunol.*
- Alberg, A.J., Shopland, D.R., Cummings, K.M., 2014. The 2014 Surgeon General’s Report: Commemorating the 50th Anniversary of the 1964 Report of the Advisory Committee to the US Surgeon General and Updating the Evidence on the Health Consequences of Cigarette Smoking. *Am. J. Epidemiol.*
- Alkan, C., Kidd, J.M., Marques-Bonet, T., Aksay, G., Antonacci, F., Hormozdiari, F., Kitzman, J.O.,

Baker, C., Malig, M., Mutlu, O., Sahinalp, S.C., Gibbs, R.A., Eichler, E.E., 2009. Personalized copy number and segmental duplication maps using next-generation sequencing. *Nat. Genet.* 41, 1061–7.

Allfrey, V.G., Faulkner, R., Mirsky, A.E., 1964. Acetylation and methylation of histones and their possible role In *The. Proc. Natl. Acad. Sci. United States* 51, 786–794.

Alter, O., Brown, P.O., Botstein, D., 2000. Singular value decomposition for genome-Wide expression data processing and modeling. *Proc. Natl. Acad. Sci. U. S. A.* 97, 10101–10106.

Altshuler, D.M., Durbin, R.M., Abecasis, G.R., Bentley, D.R., Chakravarti, A., Clark, A.G., Donnelly, P., Eichler, E.E., Flicek, P., Gabriel, S.B., Gibbs, R.A., Green, E.D., Hurles, M.E., Knoppers, B.M., Korbel, J.O., Lander, E.S., Lee, C., Lehrach, H., Mardis, E.R., Marth, G.T., McVean, G.A., Nickerson, D.A., Schmidt, J.P., Sherry, S.T., Wang, J., Wilson, R.K., Dinh, H., Kovar, C., Lee, S., Lewis, L., Muzny, D., Reid, J., Wang, M., Fang, X., Guo, X., Jian, M., Jiang, H., Jin, X., Li, G., Li, J., Li, Y., Li, Z., Liu, X., Lu, Y., Ma, X., Su, Z., Tai, S., Tang, M., Wang, B., Wang, G., Wu, H., Wu, R., Yin, Y., Zhang, W., Zhao, J., Zhao, M., Zheng, X., Zhou, Y., Gupta, N., Clarke, L., Leinonen, R., Smith, R.E., Zheng-Bradley, X., Grocock, R., Humphray, S., James, T., Kingsbury, Z., Sudbrak, R., Albrecht, M.W., Amstislavskiy, V.S., Borodina, T.A., Lienhard, M., Mertes, F., Sultan, M., Timmermann, B., Yaspo, M.L., Fulton, L., Fulton, R., Weinstock, G.M., Balasubramaniam, S., Burton, J., Danecek, P., Keane, T.M., Kolb-Kokocinski, A., McCarthy, S., Stalker, J., Quail, M., Davies, C.J., Gollub, J., Webster, T., Wong, B., Zhan, Y., Auton, A., Yu, F., Bainbridge, M., Challis, D., Evani, U.S., Lu, J., Nagaswamy, U., Sabo, A., Wang, Y., Yu, J., Coin, L.J.M., Fang, L., Li, Q., Li, Z., Lin, H., Liu, B., Luo, R., Qin, N., Shao, H., Wang, B., Xie, Y., Ye, C., Yu, C., Zhang, F., Zheng, H., Zhu, H., Garrison, E.P., Kural, D., Lee, W.P., Fung Leong, W., Ward, A.N., Wu, J., Zhang, M., Griffin, L., Hsieh, C.H., Mills, R.E., Shi, X., Von Grotthuss, M., Zhang, C., Daly, M.J., Depristo, M.A., Banks, E., Bhatia, G., Carneiro, M.O., Del Angel, G., Genovese, G., Handsaker, R.E., Hartl, C., McCarroll, S.A., Nemes, J.C., Poplin, R.E., Schaffner, S.F., Shakir, K., Yoon, S.C., Lihm, J., Makarov, V., Jin, H., Kim, W., Cheol Kim, K., Rausch, T., Beal, K., Cunningham, F., Herrero, J., McLaren, W.M., Ritchie, G.R.S., Gottipati, S., Keinan, A., Rodriguez-Flores, J.L., Sabeti, P.C., Grossman, S.R., Tabrizi, S., Tariyal, R., Cooper, D.N., Ball, E. V., Stenson, P.D., Barnes, B., Bauer, M., Keira Cheetham, R., Cox, T., Eberle, M., Kahn, S., Murray, L., Peden, J., Shaw, R., Ye, K., Batzer, M.A., Konkel, M.K., Walker, J.A., MacArthur, D.G., Lek, M., Herwig, R., Shriver, M.D., Bustamante, C.D., Byrnes, J.K., De La Vega, F.M., Gravel, S., Kenny, E.E., Kidd, J.M., Maples, B.K., Moreno-Estrada, A., Zakharia, F., Halperin, E., Baran, Y., Craig, D.W., Christoforides, A., Homer, N., Izatt, T., Kurdoglu, A.A., Sinari, S.A., Squire, K., Xiao, C., Sebat, J., Bafna, V., Ye, K., Burchard, E.G., Hernandez, R.D., Gignoux, C.R., Haussler, D., Katzman, S.J., James Kent, W., Howie, B., Ruiz-Linares, A., Dermitzakis, E.T., Lappalainen, T., Devine, S.E., Liu, X., Maroo, A., Tallon, L.J., Rosenfeld, J.A., Michelson, L.P., Min Kang, H., Anderson, P., Angius, A., Bigham, A., Blackwell, T., Busonero, F., Cucca, F., Fuchsberger, C., Jones, C., Jun, G., Li, Y., Lyons, R., Maschio, A., Porcu, E., Reinier, F., Sanna, S., Schlessinger, D., Sidore, C., Tan, A., Kate Trost, M., Awadalla, P., Hodgkinson, A., Lunter, G., Marchini, J.L., Myers, S., Churchhouse, C., Delaneau, O., Gupta-Hinch, A., Iqbal, Z., Mathieson, I., Rimmer, A., Xifara, D.K., Oleksyk, T.K., Fu, Y., Liu, X., Xiong, M., Jorde, L., Witherspoon, D., Xing, J., Browning, B.L., Alkan, C., Hajirasouliha, I., Hormozdiari, F., Ko, A., Sudmant, P.H., Chen, K., Chinwalla, A., Ding, L., Dooling, D., Koboldt, D.C., McLellan, M.D., Wallis, J.W., Wendl, M.C., Zhang, Q., Tyler-Smith, C., Albers, C.A., Ayub, Q., Chen, Y., Coffey, A.J., Colonna, V., Huang, N., Jostins, L., Li, H., Scally, A., Walter, K., Xue, Y., Zhang, Y., Gerstein, M.B., Abyzov, A., Balasubramanian, S., Chen, J., Clarke, D., Fu, Y., Habegger, L., Harmanci, A.O., Jin, M., Khurana, E., Jasmine Mu, X., Sis, C., Degenhardt, J., Stütz, A.M., Keira Cheetham, R., Church, D., Michaelson, J.J., Blackburne, B., Lindsay, S.J., Ning, Z., Frankish, A., Harrow, J., Mu, X.J., Fowler, G., Hale, W., Kalra, D., Barker, J., Kelman, G., Kulesha, E., Radhakrishnan, R., Roa, A., Smirnov, D., Streeter, I., Toneva, I., Vaughan, B., Ananiev, V., Belaia, Z., Beloslyudtsev, D., Bouk, N., Chen, C., Cohen, R., Cook, C., Garner, J., Hefferon, T., Kimelman,

- M., Liu, C., Lopez, J., Meric, P., O'Sullivan, C., Ostapchuk, Y., Phan, L., Ponomarov, S., Schneider, V., Shekhtman, E., Sirotkin, K., Slotta, D., Zhang, H., Barnes, K.C., Beiswanger, C., Cai, H., Cao, H., Gharani, N., Henn, B., Jones, D., Kaye, J.S., Kent, A., Kerasidou, A., Mathias, R., Ossorio, P.N., Parker, M., Reich, D., Rotimi, C.N., Royal, C.D., Sandoval, K., Su, Y., Tian, Z., Tishkoff, S., Toji, L.H., Via, M., Wang, Y., Yang, H., Yang, L., Zhu, J., Bodmer, W., Bedoya, G., Ming, C.Z., Yang, G., Jia You, C., Peltonen, L., Garcia-Montero, A., Orfao, A., Dutil, J., Martinez-Cruzado, J.C., Brooks, L.D., Felsenfeld, A.L., McEwen, J.E., Clemm, N.C., Duncanson, A., Dunn, M., Guyer, M.S., Peterson, J.L., Lacroute, P., 2012. An integrated map of genetic variation from 1,092 human genomes. *Nature* 491, 56–65.
- Alverson, C.J., Strickland, M.J., Gilboa, S.M., Correa, A., 2011. Maternal Smoking and Congenital Heart Defects in the Baltimore-Washington Infant Study. *Pediatrics* 127, e647–e653.
- Amouroux, R., Nashun, B., Shirane, K., Nakagawa, S., Hill, P.W.S., D'Souza, Z., Nakayama, M., Matsuda, M., Turp, A., Ndjetehe, E., Encheva, V., Kudo, N.R., Koseki, H., Sasaki, H., Hajkova, P., 2016. De novo DNA methylation drives 5hmC accumulation in mouse zygotes. *Nat. Cell Biol.* 18, 225–233.
- Anders, S., Pyl, P.T., Huber, W., 2015. HTSeq-A Python framework to work with high-throughput sequencing data. *Bioinformatics* 31, 166–169.
- Andersson, R., Enroth, S., Rada-Iglesias, A., Wadelius, C., Komorowski, J., 2009. Nucleosomes are well positioned in exons and carry characteristic histone modifications. *Genome Res.* 19, 1732–41.
- Andrews, S., 2010. FastQC: A quality control tool for high throughput sequence data.
- Aran, D., Toperoff, G., Rosenberg, M., Hellman, A., 2011. Replication timing-related and gene body-specific methylation of active human genes. *Hum. Mol. Genet.* 20, 670–680.
- Arechederra, M., Daian, F., Yim, A., Bazai, S.K., Richelme, S., Dono, R., Saurin, A.J., Habermann, B.H., Maina, F., 2018. Hypermethylation of gene body CpG islands predicts high dosage of functional oncogenes in liver cancer. *Nat. Commun.* 9.
- Arima, Y., Hirota, T., Bronner, C., Mousli, M., Fujiwara, T., Niwa, S., Ishikawa, H., Saya, H., 2004. Down-regulation of nuclear protein ICBP90 by p53/p21Cip1/WAF1-dependent DNA-damage checkpoint signals contributes to cell cycle arrest at G1/S transition. *Genes Cells* 9, 131–42.
- Arita, K., Ariyoshi, M., Tochio, H., Nakamura, Y., Shirakawa, M., 2008. Recognition of hemimethylated DNA by the SRA protein UHRF1 by a base-flipping mechanism. *Nature* 455, 818–21.
- Assenov, Y., Müller, F., Lutsik, P., Walter, J., Lengauer, T., Bock, C., 2014. Comprehensive analysis of DNA methylation data with RnBeads. *Nat. Methods* 11, 1138–40.
- Avvakumov, G. V., Walker, J.R., Xue, S., Li, Y., Duan, S., Bronner, C., Arrowsmith, C.H., Dhe-Paganon, S., 2008. Structural basis for recognition of hemi-methylated DNA by the SRA domain of human UHRF1. *Nature* 455, 822–5.
- Banerji, J., Rusconi, S., Schaffner, W., 1981. Expression of a  $\beta$ -globin gene is enhanced by remote SV40 DNA sequences. *Cell* 27, 299–308.
- Barker, D.J.P., 2004. The developmental origins of chronic adult disease. In: *Acta Paediatrica, International Journal of Paediatrics, Supplement.* pp. 26–33.
- Barker, D.J.P., 2004. Developmental origins of adult health and disease. *J. Epidemiol. Community Health* 58, 114–5.
- Barker, D.J.P., Osmond, C., 1986. Infant mortality, childhood nutrition, and ischaemic heart disease in England and Wales. *Lancet* 1, 943–6.

- Barski, A., Cuddapah, S., Cui, K., Roh, T.-Y., Schones, D.E., Wang, Z., Wei, G., Chepelev, I., Zhao, K., 2007. High-Resolution Profiling of Histone Methylations in the Human Genome. *Cell* 129, 823–837.
- Bell, R.E., Golan, T., Sheinboim, D., Malcov, H., Amar, D., Salamon, A., Liron, T., Gelfman, S., Gabet, Y., Shamir, R., Levy, C., 2016. Enhancer methylation dynamics contribute to cancer plasticity and patient mortality. *Genome Res.* 26, 601–611.
- Belmont, J.W., Hardenbol, P., Willis, T.D., Yu, F., Yang, H., Ch'Ang, L.Y., Huang, W., Liu, B., Shen, Y., Tam, P.K.H., Tsui, L.C., Waye, M.M.Y., Wong, J.T.F., Zeng, C., Zhang, Q., Chee, M.S., Galver, L.M., Kruglyak, S., Murray, S.S., Oliphant, A.R., Montpetit, A., Chagnon, F., Ferretti, V., Leboeuf, M., Phillips, M.S., Verner, A., Duan, S., Lind, D.L., Miller, R.D., Rice, J., Saccone, N.L., Taillon-Miller, P., Xiao, M., Sekine, A., Sorimachi, K., Tanaka, Y., Tsunoda, T., Yoshino, E., Bentley, D.R., Hunt, S., Powell, D., Zhang, H., Matsuda, I., Fukushima, Y., Macer, D.R., Suda, E., Rotimi, C., Adebamowo, C.A., Aniagwu, T., Marshall, P.A., Matthew, O., Nkwodimmah, C., Royal, C.D.M., Leppert, M.F., Dixon, M., Cunningham, F., Kanani, A., Thorisson, G.A., Chen, P.E., Cutler, D.J., Kashuk, C.S., Donnelly, P., Marchini, J., McVean, G.A.T., Myers, S.R., Cardon, L.R., Morris, A., Weir, B.S., Mullikin, J.C., Feolo, M., Daly, M.J., Qiu, R., Kent, A., Dunston, G.M., Kato, K., Niikawa, N., Watkin, J., Gibbs, R.A., Sodergren, E., Weinstock, G.M., Wilson, R.K., Fulton, L.L., Rogers, J., Birren, B.W., Han, H., Wang, H., Godbout, M., Wallenburg, J.C., L'Archevêque, P., Bellemare, G., Todani, K., Fujita, T., Tanaka, S., Holden, A.L., Collins, F.S., Brooks, L.D., McEwen, J.E., Guyer, M.S., Jordan, E., Peterson, J.L., Spiegel, J., Sung, L.M., Zacharia, L.F., Kennedy, K., Dunn, M.G., Seabrook, R., Shillito, M., Skene, B., Stewart, J.G., Valle, D.L., Clayton, E.W., Jorde, L.B., Chakravarti, A., Cho, M.K., Duster, T., Foster, M.W., Jasperse, M., Knoppers, B.M., Kwok, P.Y., Licinio, J., Long, J.C., Ossorio, P., Wang, V.O., Rotimi, C.N., Spallone, P., Terry, S.F., Lander, E.S., Lai, E.H., Nickerson, D.A., Abecasis, G.R., Altshuler, D., Boehnke, M., Deloukas, P., Douglas, J.A., Gabriel, S.B., Hudson, R.R., Hudson, T.J., Kruglyak, L., Nakamura, Y., Nussbaum, R.L., Schaffner, S.F., Sherry, S.T., Stein, L.D., Tanaka, T., 2003. The international HapMap project. *Nature* 426, 789–796.
- Benetatos, L., Vartholomatos, G., 2018. Enhancer DNA methylation in acute myeloid leukemia and myelodysplastic syndromes. *Cell. Mol. Life Sci.*
- Benito, M., Parker, J., Du, Q., Wu, J., Xiang, D., Perou, C.M., Marron, J.S., 2004. Adjustment of systematic microarray data biases. *Bioinformatics* 20, 105–114.
- Bentley, D.R., Balasubramanian, S., Swerdlow, H.P., Smith, G.P., Milton, J., Brown, C.G., Hall, K.P., Evers, D.J., Barnes, C.L., Bignell, H.R., Boutell, J.M., Bryant, J., Carter, R.J., Keira Cheetham, R., Cox, A.J., Ellis, D.J., Flatbush, M.R., Gormley, N.A., Humphray, S.J., Irving, L.J., Karbelashvili, M.S., Kirk, S.M., Li, H., Liu, X., Maisinger, K.S., Murray, L.J., Obradovic, B., Ost, T., Parkinson, M.L., Pratt, M.R., Rasolonjatovo, I.M.J., Reed, M.T., Rigatti, R., Rodighiero, C., Ross, M.T., Sabot, A., Sankar, S. V., Scally, A., Schroth, G.P., Smith, M.E., Smith, V.P., Spiridou, A., Torrance, P.E., Tzonev, S.S., Vermaas, E.H., Walter, K., Wu, X., Zhang, L., Alam, M.D., Anastasi, C., Aniebo, I.C., Bailey, D.M.D., Bancarz, I.R., Banerjee, S., Barbour, S.G., Baybayan, P.A., Benoit, V.A., Benson, K.F., Bevis, C., Black, P.J., Boodhun, A., Brennan, J.S., Bridgham, J.A., Brown, R.C., Brown, A.A., Buermann, D.H., Bundu, A.A., Burrows, J.C., Carter, N.P., Castillo, N., Catenazzi, M.C.E., Chang, S., Neil Cooley, R., Crake, N.R., Dada, O.O., Diakoumakos, K.D., Dominguez-Fernandez, B., Earnshaw, D.J., Egbujor, U.C., Elmore, D.W., Etchin, S.S., Ewan, M.R., Fedurco, M., Fraser, L.J., Fuentes Fajardo, K. V., Scott Furey, W., George, D., Gietzen, K.J., Goddard, C.P., Golda, G.S., Granieri, P.A., Green, D.E., Gustafson, D.L., Hansen, N.F., Harnish, K., Haudenschild, C.D., Heyer, N.I., Hims, M.M., Ho, J.T., Horgan, A.M., Hoschler, K., Hurwitz, S., Ivanov, D. V., Johnson, M.Q., James, T., Huw Jones, T.A., Kang, G.D., Kerelska, T.H., Kersey, A.D., Khrebtukova, I., Kindwall, A.P., Kingsbury, Z., Kokko-Gonzales, P.I., Kumar, A., Laurent, M.A., Lawley, C.T., Lee, S.E., Lee,

- X., Liao, A.K., Loch, J.A., Lok, M., Luo, S., Mammen, R.M., Martin, J.W., McCauley, P.G., McNitt, P., Mehta, P., Moon, K.W., Mullens, J.W., Newington, T., Ning, Z., Ling Ng, B., Novo, S.M., O'Neill, M.J., Osborne, M.A., Osnowski, A., Ostadan, O., Paraschos, L.L., Pickering, L., Pike, A.C., Pike, A.C., Chris Pinkard, D., Pliskin, D.P., Podhasky, J., Quijano, V.J., Raczy, C., Rae, V.H., Rawlings, S.R., Chiva Rodriguez, A., Roe, P.M., Rogers, J., Rogert Bacigalupo, M.C., Romanov, N., Romieu, A., Roth, R.K., Rourke, N.J., Ruediger, S.T., Rusman, E., Sanches-Kuiper, R.M., Schenker, M.R., Seoane, J.M., Shaw, R.J., Shiver, M.K., Short, S.W., Sizto, N.L., Sluis, J.P., Smith, M.A., Ernest Sohna Sohna, J., Spence, E.J., Stevens, K., Sutton, N., Szajkowski, L., Tregidgo, C.L., Turcatti, G., Vandevondele, S., Verhovsky, Y., Virk, S.M., Wakelin, S., Walcott, G.C., Wang, J., Worsley, G.J., Yan, J., Yau, L., Zuerlein, M., Rogers, J., Mullikin, J.C., Hurles, M.E., McCooke, N.J., West, J.S., Oaks, F.L., Lundberg, P.L., Klenerman, D., Durbin, R., Smith, A.J., 2008. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* 456, 53–59.
- Berkyurek, A.C., Suetake, I., Arita, K., Takeshita, K., Nakagawa, A., Shirakawa, M., Tajima, S., 2014. The DNA methyltransferase Dnmt1 directly interacts with the SET and RING finger-associated (SRA) domain of the multifunctional protein Uhrf1 to facilitate accession of the catalytic center to hemi-methylated DNA. *J. Biol. Chem.* 289, 379–86.
- Bestor, T.H., 1988. Cloning of a mammalian DNA methyltransferase. *Gene* 74, 9–12.
- Bestor, T.H., 2000. The DNA methyltransferases of mammals. *Hum. Mol. Genet.* 9, 2395–2402.
- Bibikova, M., Barnes, B., Tsan, C., Ho, V., Klotzle, B., Le, J.M., Delano, D., Zhang, L., Schroth, G.P., Gunderson, K.L., Fan, J.-B., Shen, R., 2011. High density DNA methylation array with single CpG site resolution. *Genomics* 98, 288–95.
- Biswas, S., Rao, C.M., 2018. Epigenetic tools (The Writers, The Readers and The Erasers) and their implications in cancer therapy. *Eur. J. Pharmacol.*
- Blackledge, N.P., Farcas, A.M., Kondo, T., King, H.W., McGouran, J.F., Hanssen, L.L.P., Ito, S., Cooper, S., Kondo, K., Koseki, Y., Ishikura, T., Long, H.K., Sheahan, T.W., Brockdorff, N., Kessler, B.M., Koseki, H., Klose, R.J., 2014. Variant PRC1 complex-dependent H2A ubiquitylation drives PRC2 recruitment and polycomb domain formation. *Cell* 157, 1445–1459.
- Blohm, F., Fridén, B., Milsom, I., 2008. A prospective longitudinal population-based study of clinical miscarriage in an urban Swedish population. *BJOG An Int. J. Obstet. Gynaecol.* 115, 176–182.
- Børnich, C., Grytten, I., Hovig, E., Paulsen, J., Čech, M., Sandve, G.K., 2016. Galaxy Portal: interacting with the galaxy platform through mobile devices. *Bioinformatics* 32, 1743–5.
- Borodina, T., Adjaye, J., Sultan, M., 2011. A strand-specific library preparation protocol for RNA sequencing. In: *Methods in Enzymology*. Academic Press Inc., pp. 79–98.
- Bostick, M., Kim, J.K., Estève, P.-O., Clark, A., Pradhan, S., Jacobsen, S.E., 2007. UHRF1 plays a role in maintaining DNA methylation in mammalian cells. *Science* 317, 1760–4.
- Bourc'his, D., Bestor, T.H., 2004. Meiotic catastrophe and retrotransposon reactivation in male germ cells lacking Dnmt3L. *Nature* 431, 96–9.
- Bourgeois, Y., Boissinot, S., 2019. On the Population Dynamics of Junk: A Review on the Population Genomics of Transposable Elements. *Genes (Basel)*. 10.
- Boyle, M.P., Brewer, J.A., Funatsu, M., Wozniak, D.F., Tsien, J.Z., Izumi, Y., Muglia, L.J., 2005. Acquired deficit of forebrain glucocorticoid receptor produces depression-like changes in adrenal axis regulation and behavior. *Proc. Natl. Acad. Sci. U. S. A.* 102, 473–8.
- Brenet, F., Moh, M., Funk, P., Feierstein, E., Viale, A.J., Socci, N.D., Scandura, J.M., 2011. DNA

methylation of the first exon is tightly linked to transcriptional silencing. *PLoS One* 6.

- Brinkman, A.B., Gu, H., Bartels, S.J.J., Zhang, Y., Matarese, F., Simmer, F., Marks, H., Bock, C., Gnirke, A., Meissner, A., Stunnenberg, H.G., 2012. Sequential ChIP-bisulfite sequencing enables direct genome-scale investigation of chromatin and DNA methylation cross-talk. *Genome Res.* 22, 1128–38.
- Brown, A.S., Susser, E.S., Lin, S.P., Neugebauer, R., Gorman, J.M., 1995. Increased risk of affective disorders in males after second trimester prenatal exposure to the Dutch hunger winter of 1944-45. *Br. J. Psychiatry* 166, 601–6.
- Brown, K.D., Robertson, K.D., 2007. DNMT1 knockout delivers a strong blow to genome stability and cell viability. *Nat. Genet.* 39, 289–290.
- Buenrostro, J.D., Giresi, P.G., Zaba, L.C., Chang, H.Y., Greenleaf, W.J., 2013a. Transposition of native chromatin for multimodal regulatory analysis and personal epigenomics. *Nat. Methods* 10, 1213.
- Buenrostro, J.D., Giresi, P.G., Zaba, L.C., Chang, H.Y., Greenleaf, W.J., 2013b. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat. Methods* 10, 1213–1218.
- Buenrostro, J.D., Wu, B., Chang, H.Y., Greenleaf, W.J., 2015a. ATAC-seq: A Method for Assaying Chromatin Accessibility Genome-Wide. *Curr. Protoc. Mol. Biol.* 109, 21.29.1-9.
- Buenrostro, J.D., Wu, B., Litzenburger, U.M., Ruff, D., Gonzales, M.L., Snyder, M.P., Chang, H.Y., Greenleaf, W.J., 2015b. Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature* 523, 486–90.
- Bullard, J.H., Purdom, E., Hansen, K.D., Dudoit, S., 2010. Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. *BMC Bioinformatics* 11.
- Burford, N.G., Webster, N.A., Cruz-Topete, D., 2017. Hypothalamic-Pituitary-Adrenal Axis Modulation of Glucocorticoids in the Cardiovascular System. *Int. J. Mol. Sci.* 18.
- Busby, M.A., Stewart, C., Miller, C.A., Grzeda, K.R., Marth, G.T., 2013. Scotty: A web tool for designing RNA-Seq experiments to measure differential gene expression. *Bioinformatics* 29, 656–657.
- Caffrey, A., Irwin, R.E., McNulty, H., Strain, J.J., Lees-Murdock, D.J., McNulty, B.A., Ward, M., Walsh, C.P., Pentieva, K., 2018. Gene-specific DNA methylation in newborns in response to folic acid supplementation during the second and third trimesters of pregnancy: epigenetic analysis from a randomized controlled trial. *Am. J. Clin. Nutr.* 107, 566–575.
- Cardoso, A.R., Oliveira, M., Amorim, A., Azevedo, L., 2016. Major influence of repetitive elements on disease-associated copy number variants (CNVs). *Hum. Genomics*.
- Carrozza, M.J., Li, B., Florens, L., Suganuma, T., Swanson, S.K., Lee, K.K., Shia, W.-J., Anderson, S., Yates, J., Washburn, M.P., Workman, J.L., 2005. Histone H3 Methylation by Set2 Directs Deacetylation of Coding Regions by Rpd3S to Suppress Spurious Intragenic Transcription. *Cell* 123, 581–592.
- Casavant, S.G., Cong, X., Fitch, R.H., Moore, J., Rosenkrantz, T., Starkweather, A., 2019. Allostatic Load and Biomarkers of Stress in the Preterm Infant: An Integrative Review. *Biol. Res. Nurs.* 21, 210–223.
- Centers for Disease Control, 1992. Recommendations for the use of folic acid to reduce the number of cases of spina bifida and other neural tube defects. *Morb. Mortal. Wkly. report.* 41, 1–8.

- Charles, D.H.M., Ness, A.R., Campbell, D., Smith, G.D., Whitley, E., Hall, M.H., 2005. Folic acid supplements in pregnancy and birth outcome: Re-analysis of a large randomised controlled trial and update of Cochrane review. *Paediatr. Perinat. Epidemiol.*
- Chen, H., Li, G., Chan, Y.L., Chapman, D.G., Sukjamnong, S., Nguyen, T., Annissa, T., McGrath, K.C., Sharma, P., Oliver, B.G., 2018. Maternal E-cigarette exposure in mice alters DNA methylation and lung cytokine expression in offspring. *Am. J. Respir. Cell Mol. Biol.*
- Chen, T., Ueda, Y., Dodge, J.E., Wang, Z., Li, E., 2003. Establishment and maintenance of genomic methylation patterns in mouse embryonic stem cells by Dnmt3a and Dnmt3b. *Mol. Cell. Biol.* 23, 5594–605.
- Chen, T., Ueda, Y., Dodge, J.E., Wang, Z., Li, E., 2003. Establishment and Maintenance of Genomic Methylation Patterns in Mouse Embryonic Stem Cells by Dnmt3a and Dnmt3b. *Mol. Cell. Biol.* 23, 5594–5605.
- Chen, Y.A., Lemire, M., Choufani, S., Butcher, D.T., Grafodatskaya, D., Zanke, B.W., Gallinger, S., Hudson, T.J., Weksberg, R., 2013. Discovery of cross-reactive probes and polymorphic CpGs in the Illumina Infinium HumanMethylation450 microarray. *Epigenetics* 8, 203–209.
- Chenais, B., 2015. Transposable elements in cancer and other human diseases. *Curr. Cancer Drug Targets* 15, 227–42.
- Cheng, J., Yang, Y., Fang, J., Xiao, J., Zhu, T., Chen, F., Wang, P., Li, Z., Yang, H., Xu, Y., 2013. Structural insight into coordinated recognition of trimethylated histone H3 lysine 9 (H3K9me3) by the plant homeodomain (PHD) and tandem tudor domain (TTD) of UHRF1 (ubiquitin-like, containing PHD and RING finger domains, 1) protein. *J. Biol. Chem.* 288, 1329–39.
- Cheung, L., Messina, M., Gill, A., Clarkson, A., Learoyd, D., Delbridge, L., Wentworth, J., Philips, J., Clifton-Bligh, R., Robinson, B.G., 2003. Detection of the PAX8-PPAR gamma fusion oncogene in both follicular thyroid carcinomas and adenomas. *J. Clin. Endocrinol. Metab.* 88, 354–7.
- Chiappinelli, K.B., Strissel, P.L., Desrichard, A., Li, H., Henke, C., Akman, B., Hein, A., Rote, N.S., Cope, L.M., Snyder, A., Makarov, V., Buhu, S., Slamon, D.J., Wolchok, J.D., Pardoll, D.M., Beckmann, M.W., Zahnow, C.A., Mergoub, T., Chan, T.A., Baylin, S.B., Strick, R., 2015. Inhibiting DNA Methylation Causes an Interferon Response in Cancer via dsRNA Including Endogenous Retroviruses. *Cell* 162, 974–986.
- Chittock, E.C., Latwiel, S., Miller, T.C.R., Müller, C.W., 2017. Molecular architecture of polycomb repressive complexes. *Biochem. Soc. Trans.*
- Chuang, L.S., Ian, H.I., Koh, T.W., Ng, H.H., Xu, G., Li, B.F., 1997. Human DNA-(cytosine-5) methyltransferase-PCNA complex as a target for p21WAF1. *Science* 277, 1996–2000.
- Ciferri, C., Lander, G.C., Maiolica, A., Herzog, F., Aebersold, R., Nogales, E., 2012. Molecular architecture of human polycomb repressive complex 2. *Elife* 1.
- Clare, C.E., Brassington, A.H., Kwong, W.Y., Sinclair, K.D., 2019. One-Carbon Metabolism: Linking Nutritional Biochemistry to Epigenetic Programming of Long-Term Development. *Annu. Rev. Anim. Biosci.* 7, 263–287.
- Clermont, P.-L., Parolia, A., Liu, H.H., Helgason, C.D., 2016. DNA methylation at enhancer regions: Novel avenues for epigenetic biomarker development. *Front. Biosci. (Landmark Ed.)* 21, 430–46.)
- Cohen, D.E., Davidow, L.S., Erwin, J.A., Xu, N., Warshawsky, D., Lee, J.T., 2007. The DXPas34 repeat regulates random and imprinted X inactivation. *Dev. Cell* 12, 57–71.
- Cokus, S.J., Feng, S., Zhang, X., Chen, Z., Merriman, B., Haudenschild, C.D., Pradhan, S., Nelson, S.F.,



- Pellegrini, M., Jacobsen, S.E., 2008. Shotgun bisulphite sequencing of the Arabidopsis genome reveals DNA methylation patterning. *Nature* 452, 215–219.
- Cooper, W.N., Khulan, B., Owens, S., Elks, C.E., Seidel, V., Prentice, A.M., Belteki, G., Ong, K.K., Affara, N.A., Constancia, M., Dunger, D.B., 2014. DNA methylation profiling at imprinted loci after periconceptional micronutrient supplementation in humans: Results of a pilot randomized controlled trials. *World Rev. Nutr. Diet.*
- Corces, M.R., Trevino, A.E., Hamilton, E.G., Greenside, P.G., Sinnott-Armstrong, N.A., Vesuna, S., Satpathy, A.T., Rubin, A.J., Montine, K.S., Wu, B., Kathiria, A., Cho, S.W., Mumbach, M.R., Carter, A.C., Kasowski, M., Orloff, L.A., Risca, V.I., Kundaje, A., Khavari, P.A., Montine, T.J., Greenleaf, W.J., Chang, H.Y., 2017. An improved ATAC-seq protocol reduces background and enables interrogation of frozen tissues. *Nat. Methods* 14, 959–962.
- Craddock, N., Hurler, M.E., Cardin, N., Pearson, R.D., Plagnol, V., Robson, S., Vukcevic, D., Barnes, C., Conrad, D.F., Giannoulatou, E., Holmes, C., Marchini, J.L., Stirrups, K., Tobin, M.D., Wain, L. V., Yau, C., Aerts, J., Ahmad, T., Andrews, T.D., Arbury, H., Attwood, A., Auton, A., Ball, S.G., Balmforth, A.J., Barrett, J.C., Barroso, I., Barton, A., Bennett, A.J., Bhaskar, S., Blaszczyk, K., Bowes, J., Brand, O.J., Braund, P.S., Bredin, F., Breen, G., Brown, M.J., Bruce, I.N., Bull, J., Burren, O.S., Burton, J., Byrnes, J., Caesar, S., Clee, C.M., Coffey, A.J., Connell, J.M.C., Cooper, J.D., Dominiczak, A.F., Downes, K., Drummond, H.E., Dudakia, D., Dunham, A., Ebbs, B., Eccles, D., Edkins, S., Edwards, C., Elliot, A., Emery, P., Evans, D.M., Evans, G., Eyre, S., Farmer, A., Ferrier, I.N., Feuk, L., Fitzgerald, T., Flynn, E., Forbes, A., Forty, L., Franklyn, J.A., Freathy, R.M., Gibbs, P., Gilbert, P., Gokumen, O., Gordon-Smith, K., Gray, E., Green, E., Groves, C.J., Grozeva, D., Gwilliam, R., Hall, A., Hammond, N., Hardy, M., Harrison, P., Hassanali, N., Hebaishi, H., Hines, S., Hinks, A., Hitman, G.A., Hocking, L., Howard, E., Howard, P., Howson, J.M.M., Hughes, D., Hunt, S., Isaacs, J.D., Jain, M., Jewell, D.P., Johnson, T., Jolley, J.D., Jones, I.R., Jones, L.A., Kirov, G., Langford, C.F., Lango-Allen, H., Lathrop, G.M., Lee, J., Lee, K.L., Lees, C., Lewis, K., Lindgren, C.M., Maisuria-Armer, M., Maller, J., Mansfield, J., Martin, P., Massey, D.C.O., McArdle, W.L., McGuffin, P., McLay, K.E., Mentzer, A., Mimmack, M.L., Morgan, A.E., Morris, A.P., Mowat, C., Myers, S., Newman, W., Nimmo, E.R., O'Donovan, M.C., Onipinla, A., Onyiah, I., Ovington, N.R., Owen, M.J., Palin, K., Parnell, K., Pernet, D., Perry, J.R.B., Phillips, A., Pinto, D., Prescott, N.J., Prokopenko, I., Quail, M.A., Rafelt, S., Rayner, N.W., Redon, R., Reid, D.M., Renwick, A., Ring, S.M., Robertson, N., Russell, E., Clair, D.S., Sambrook, J.G., Sanderson, J.D., Schuilenburg, H., Scott, C.E., Scott, R., Seal, S., Shaw-Hawkins, S., Shields, B.M., Simmonds, M.J., Smyth, D.J., Somaskantharajah, E., Spanova, K., Steer, S., Stephens, J., Stevens, H.E., Stone, M.A., Su, Z., Symmons, D.P.M., Thompson, J.R., Thomson, W., Travers, M.E., Turnbull, C., Valsesia, A., Walker, M., Walker, N.M., Wallace, C., Warren-Perry, M., Watkins, N.A., Webster, J., Weedon, M.N., Wilson, A.G., Woodburn, M., Wordsworth, B.P., Young, A.H., Zeggini, E., Carter, N.P., Frayling, T.M., Lee, C., McVean, G., Munroe, P.B., Palotie, A., Sawcer, S.J., Scherer, S.W., Strachan, D.P., Tyler-Smith, C., Brown, M.A., Burton, P.R., Caulfield, M.J., Compston, A., Farrall, M., Gough, S.C.L., Hall, A.S., Hattersley, A.T., Hill, A.V.S., Mathew, C.G., Pembrey, M., Satsangi, J., Stratton, M.R., Worthington, J., Deloukas, P., Duncanson, A., Kwiatkowski, D.P., McCarthy, M.I., Ouwehand, W.H., Parkes, M., Rahman, N., Todd, J.A., Samani, N.J., Donnelly, P., 2010. Genome-wide association study of CNVs in 16,000 cases of eight common diseases and 3,000 shared controls. *Nature* 464, 713–720.
- Czeizel, A.E., Dudás, I., 1993. Prevention of the first occurrence of neural-tube defects by periconceptional vitamin supplementation. *Obstet. Gynecol. Surv.* 48, 395–397.
- Deaton, A.M., Bird, A., 2011. CpG islands and the regulation of transcription. *Genes Dev.* 25, 1010–22.
- Dedeurwaerder, S., Defrance, M., Bizet, M., Calonne, E., Bontempi, G., Fuks, F., 2013. A

- comprehensive overview of Infinium Human Methylation450 data processing. *Brief. Bioinform.* 15, 929–941.
- Dedeurwaerder, S., Defrance, M., Calonne, E., Denis, H., Sotiriou, C., Fuks, F., 2011. Evaluation of the Infinium Methylation 450K technology. *Epigenomics* 3, 771–784.
- Department of Health, 1992. Folic acid and the prevention of neural tube defects. Heywood, United Kingdom.
- Dhayalan, A., Rajavelu, A., Rathert, P., Tamas, R., Jurkowska, R.Z., Ragozin, S., Jeltsch, A., 2010. The Dnmt3a PWWP domain reads histone 3 lysine 36 trimethylation and guides DNA methylation. *J. Biol. Chem.* 285, 26114–26120.
- Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., Gingeras, T.R., 2013. STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics* 29, 15–21.
- Drongitis, D., Aniello, F., Fucci, L., Donizetti, A., 2019. Roles of Transposable Elements in the Different Layers of Gene Expression Regulation. *Int. J. Mol. Sci.*
- Edwards, J.R., Yarychivska, O., Boulard, M., Bestor, T.H., 2017. DNA methylation and DNA methyltransferases. *Epigenetics Chromatin* 10, 23.
- Ehrlich, K.C., Paterson, H.L., Lacey, M., Ehrlich, M., 2016. Focus: Epigenetics: DNA Hypomethylation in Intragenic and Intergenic Enhancer Chromatin of Muscle-Specific Genes Usually Correlates with their Expression. *Yale J. Biol. Med.* 89, 441.
- ENCODE Project Consortium, 2004. The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science* (80-. ). 306, 636–640.
- England, R., Pettersson, M., 2005. Pyro Q-CpG<sup>TM</sup>: Quantitative analysis of methylation in multiple CpG sites by Pyrosequencing<sup>®</sup>. *Nat. Methods* 2.
- Ernst, J., Kellis, M., 2012. ChromHMM: automating chromatin-state discovery and characterization. *Nat. Methods* 9, 215–6.
- Ernst, J., Kheradpour, P., Mikkelson, T.S., Shores, N., Ward, L.D., Epstein, C.B., Zhang, X., Wang, L., Issner, R., Coyne, M., Ku, M., Durham, T., Kellis, M., Bernstein, B.E., 2011. Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature* 473, 43–49.
- Evans, C., Hardin, J., Stoebel, D.M., 2018. Selecting between-sample RNA-Seq normalization methods from the perspective of their assumptions. *Brief. Bioinform.* 19, 776–792.
- Fan, G., Beard, C., Chen, R.Z., Csankovszki, G., Sun, Y., Siniiaia, M., Biniszkiwicz, D., Bates, B., Lee, P.P., Kühn, R., Trumpp, A., Poon, C.-S., Wilson, C.B., Jaenisch, R., 2001. DNA Hypomethylation Perturbs the Function and Survival of CNS Neurons in Postnatal Animals. *J. Neurosci.* 21, 788–797.
- Farias, N., Ho, N., Butler, S., Delaney, L., Morrison, J., Shahrzad, S., Coomber, B.L., 2015. The effects of folic acid on global DNA methylation and colonosphere formation in colon cancer cell lines. *J. Nutr. Biochem.* 26, 818–826.
- Faust, C., Schumacher, A., Holdener, B., Magnuson, T., 1995. The eed mutation disrupts anterior mesoderm production in mice. *Development* 121, 273–85.
- Fazary, A.E., Ju, Y.H., Abd-Rabboh, H.S.M., 2017. How does chromatin package DNA within nucleus and regulate gene expression? *Int. J. Biol. Macromol.*
- Feng, J., Liu, T., Qin, B., Zhang, Y., Liu, X.S., 2012. Identifying ChIP-seq enrichment using MACS. *Nat.*

Protoc. 7, 1728–1740.

- Firth, H. V., Richards, S.M., Bevan, A.P., Clayton, S., Corpas, M., Rajan, D., Vooren, S. Van, Moreau, Y., Pettett, R.M., Carter, N.P., 2009. DECIPHER: Database of Chromosomal Imbalance and Phenotype in Humans Using Ensembl Resources. *Am. J. Hum. Genet.* 84, 524–533.
- Flanagan, J.M., Wild, L., 2007. An epigenetic role for noncoding RNAs and intragenic DNA methylation. In: *Genome Biology*.
- Fortin, J.P., Triche, T.J., Hansen, K.D., 2017. Preprocessing, normalization and integration of the Illumina HumanMethylationEPIC array with minfi. *Bioinformatics* 33, 558–560.
- Foster, B.M., Stolz, P., Mulholland, C.B., Montoya, A., Kramer, H., Bultmann, S., Bartke, T., 2018. Critical Role of the UBL Domain in Stimulating the E3 Ubiquitin Ligase Activity of UHRF1 toward Chromatin. *Mol. Cell* 72, 739–752.e9.
- Fouse, S.D., Nagarajan, R.O., Costello, J.F., 2010. Genome-scale DNA methylation analysis. *Epigenomics* 2, 105–17.
- Fuso, A., 2018. Non-CpG Methylation Revised. *Epigenomes* 2, 22.
- Gamazon, E.R., Stranger, B.E., 2015. The impact of human copy number variation on gene expression. *Brief. Funct. Genomics* 14, 352–7.
- García-González, E., Escamilla-Del-Arenal, M., Arzate-Mejía, R., Recillas-Targa, F., 2016. Chromatin remodeling effects on enhancer activity. *Cell. Mol. Life Sci.*
- Gaspar, J.M., 2018a. Improved peak-calling with MACS2. *bioRxiv* 496521.
- Gaspar, J.M., 2018b. Genrich.
- Gates, L.A., Shi, J., Rohira, A.D., Feng, Q., Zhu, B., Bedford, M.T., Sagum, C.A., Jung, S.Y., Qin, J., Tsai, M.-J., Tsai, S.Y., Li, W., Foulds, C.E., O'Malley, B.W., 2017. Acetylation on histone H3 lysine 9 mediates a switch from transcription initiation to elongation. *J. Biol. Chem.* 292, 14456–14472.
- Gaudet, F., 2003. Induction of Tumors in Mice by Genomic Hypomethylation. *Science (80-. )*. 300, 489–492.
- Georgia, S., Kanji, M., Bhushan, A., 2013. DNMT1 represses p53 to maintain progenitor cell survival during pancreatic organogenesis. *Genes Dev.* 27, 372–7.
- Giardine, B., Riemer, C., Hardison, R.C., Burhans, R., Elnitski, L., Shah, P., Zhang, Y., Blankenberg, D., Albert, I., Taylor, J., Miller, W., Kent, W.J., Nekrutenko, A., 2005. Galaxy: a platform for interactive large-scale genome analysis. *Genome Res.* 15, 1451–5.
- Gillies, S.D., Morrison, S.L., Oi, V.T., Tonegawa, S., 1983. A tissue-specific transcription enhancer element is located in the major intron of a rearranged immunoglobulin heavy chain gene. *Cell* 33, 717–28.
- Girard, S.L., Gauthier, J., Noreau, A., Xiong, L., Zhou, S., Jouan, L., Dionne-Laporte, A., Spiegelman, D., Henrion, E., Diallo, O., Thibodeau, P., Bachand, I., Bao, J.Y.J., Tong, A.H.Y., Lin, C.-H., Millet, B., Jaafari, N., Joober, R., Dion, P.A., Lok, S., Krebs, M.-O., Rouleau, G.A., 2011. Increased exonic de novo mutation rate in individuals with schizophrenia. *Nat. Genet.* 43, 860–3.
- Globisch, D., Münzel, M., Müller, M., Michalakakis, S., Wagner, M., Koch, S., Brückl, T., Biel, M., Carell, T., 2010. Tissue distribution of 5-hydroxymethylcytosine and search for active demethylation intermediates. *PLoS One* 5, e15367.
- Goll, M.G., Kirpekar, F., Maggert, K.A., Yoder, J.A., Hsieh, C.-L., Zhang, X., Golic, K.G., Jacobsen, S.E.,

- Bestor, T.H., 2006. Methylation of tRNA<sup>Asp</sup> by the DNA methyltransferase homolog Dnmt2. *Science* 311, 395–8.
- Gowher, H., Jeltsch, A., 2001. Enzymatic properties of recombinant Dnmt3a DNA methyltransferase from mouse: the enzyme modifies DNA in a non-processive manner and also methylates non-CpA sites. *J. Mol. Biol.* 309, 1201–1208.
- Gowher, H., Jeltsch, A., 2018. Mammalian DNA methyltransferases: New discoveries and open questions. *Biochem. Soc. Trans.*
- Gu, H., Smith, Z.D., Bock, C., Boyle, P., Gnirke, A., Meissner, A., 2011. Preparation of reduced representation bisulfite sequencing libraries for genome-scale DNA methylation profiling. *Nat. Protoc.* 6, 468–481.
- Guo, H., Zhu, P., Guo, F., Li, X., Wu, X., Fan, X., Wen, L., Tang, F., 2015. Profiling DNA methylome landscapes of mammalian cells with single-cell reduced-representation bisulfite sequencing. *Nat. Protoc.* 10, 645–59.
- Guo, J.U., Su, Y., Shin, J.H., Shin, J., Li, H., Xie, B., Zhong, C., Hu, S., Le, T., Fan, G., Zhu, H., Chang, Q., Gao, Y., Ming, G.L., Song, H., 2014. Distribution, recognition and regulation of non-CpG methylation in the adult mammalian brain. *Nat. Neurosci.* 17, 215–222.
- Guo, X., Wang, L., Li, J., Ding, Z., Xiao, J., Yin, X., He, S., Shi, P., Dong, L., Li, G., Tian, C., Wang, J., Cong, Y., Xu, Y., 2015. Structural insight into autoinhibition and histone H3-induced activation of DNMT3A. *Nature* 517, 640–644.
- Ha, K., Lee, G.E., Pali, S.S., Brown, K.D., Takeda, Y., Liu, K., Bhalla, K.N., Robertson, K.D., 2011. Rapid and transient recruitment of DNMT1 to DNA double-strand breaks is mediated by its interaction with multiple components of the DNA damage response machinery. *Hum. Mol. Genet.* 20, 126–40.
- Haghighi, F., Xin, Y., Chanrion, B., O'Donnell, A.H., Ge, Y., Dwork, A.J., Arango, V., Mann, J.J., 2014. Increased DNA methylation in the suicide brain. *Dialogues Clin. Neurosci.* 16, 430–8.
- Hahn, M.A., Wu, X., Li, A.X., Hahn, T., Pfeifer, G.P., 2011. Relationship between gene body DNA methylation and intragenic H3K9ME3 and H3K36ME3 chromatin marks. *PLoS One* 6.
- Han, H., Cortez, C.C., Yang, X., Nichols, P.W., Jones, P.A., Liang, G., 2011. DNA methylation directly silences genes with non-CpG island promoters and establishes a nucleosome occupied promoter. *Hum. Mol. Genet.* 20, 4299–4310.
- Hansen, K.D., Brenner, S.E., Dudoit, S., 2010. Biases in Illumina transcriptome sequencing caused by random hexamer priming. *Nucleic Acids Res.* 38, e131.
- Hark, A.T., Schoenherr, C.J., Katz, D.J., Ingram, R.S., Levorse, J.M., Tilghman, S.M., 2000. CTCF mediates methylation-sensitive enhancer-blocking activity at the H19/Igf2 locus. *Nature* 405, 486–489.
- Hashimoto, H., Vertino, P.M., Cheng, X., 2010. Molecular coupling of DNA methylation and histone methylation. *Epigenomics* 2, 657–669.
- He, Y.F., Li, B.Z., Li, Z., Liu, P., Wang, Y., Tang, Q., Ding, J., Jia, Y., Chen, Z., Li, N., Sun, Y., Li, X., Dai, Q., Song, C.X., Zhang, K., He, C., Xu, G.L., 2011. Tet-mediated formation of 5-carboxylcytosine and its excision by TDG in mammalian DNA. *Science* (80-. ). 333, 1303–1307.
- Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y.C., Laslo, P., Cheng, J.X., Murre, C., Singh, H., Glass, C.K., 2010. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol. Cell* 38, 576–89.

- Heiss, J.A., Brenner, H., 2015. Between-array normalization for 450K data. *Front. Genet.* 5.
- Hermann, A., Gowher, H., Jeltsch, A., 2004. Biochemistry and biology of mammalian DNA methyltransferases. *Cell. Mol. Life Sci.*
- Heyn, H., Vidal, E., Ferreira, H.J., Vizoso, M., Sayols, S., Gomez, A., Moran, S., Boque-Sastre, R., Guil, S., Martinez-Cardus, A., Lin, C.Y., Royo, R., Sanchez-Mut, J. V., Martinez, R., Gut, M., Torrents, D., Orozco, M., Gut, I., Young, R.A., Esteller, M., 2016. Epigenomic analysis detects aberrant super-enhancer DNA methylation in human cancer. *Genome Biol.* 17.
- Hill, P.W.S., Amouroux, R., Hajkova, P., 2014. DNA demethylation, Tet proteins and 5-hydroxymethylcytosine in epigenetic reprogramming: An emerging complex story. *Genomics.*
- Holoch, D., Margueron, R., 2017. Mechanisms Regulating PRC2 Recruitment and Enzymatic Activity. *Trends Biochem. Sci.* 42, 531–542.
- Houseman, E.A., Accomando, W.P., Koestler, D.C., Christensen, B.C., Marsit, C.J., Nelson, H.H., Wiencke, J.K., Kelsey, K.T., 2012. DNA methylation arrays as surrogate measures of cell mixture distribution. *BMC Bioinformatics* 13.
- Houseman, E.A., Molitor, J., Marsit, C.J., 2014. Reference-free cell mixture adjustments in analysis of DNA methylation data. *Bioinformatics* 30, 1431–9.
- Hu, L., Li, Z., Wang, P., Lin, Y., Xu, Y., 2011. Crystal structure of PHD domain of UHRF1 and insights into recognition of unmodified histone H3 arginine residue 2. *Cell Res.* 21, 1374–8.
- Huang, D.W., Sherman, B.T., Lempicki, R.A., 2009. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res.* 37, 1–13.
- Huang, D.W., Sherman, B.T., Tan, Q., Kir, J., Liu, D., Bryant, D., Guo, Y., Stephens, R., Baseler, M.W., Lane, H.C., Lempicki, R.A., 2007. DAVID Bioinformatics Resources: expanded annotation database and novel algorithms to better extract biology from large gene lists. *Nucleic Acids Res.* 35, W169-75.
- Huang da, W., Sherman, B., Lempicki, R., 2009. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.* 4, 44–57.
- Iafate, A.J., Feuk, L., Rivera, M.N., Listewnik, M.L., Donahoe, P.K., Qi, Y., Scherer, S.W., Lee, C., 2004. Detection of large-scale variation in the human genome. *Nat. Genet.* 36, 949–951.
- Illingworth, R.S., Bird, A.P., 2009. CpG islands - 'A rough guide.' *FEBS Lett.* 583, 1713–1720.
- Illingworth, R.S., Gruenewald-Schneider, U., Webb, S., Kerr, A.R.W., James, K.D., Turner, D.J., Smith, C., Harrison, D.J., Andrews, R., Bird, A.P., 2010. Orphan CpG islands identify numerous conserved promoters in the mammalian genome. *PLoS Genet.* 6, e1001134.
- Illingworth, R.S., Gruenewald-Schneider, U., Webb, S., Kerr, A.R.W., James, K.D., Turner, D.J., Smith, C., Harrison, D.J., Andrews, R., Bird, A.P., 2010. Orphan CpG Islands Identify Numerous Conserved Promoters in the Mammalian Genome. *PLoS Genet.* 6, e1001134.
- Illumina, 2019. Illumina Adapter Sequences.
- Inano, K., Suetake, I., Ueda, T., Miyake, Y., Nakamura, M., Okada, M., Tajima, S., 2000. Maintenance-type DNA methyltransferase is highly expressed in post-mitotic neurons and localized in the cytoplasmic compartment. *J. Biochem.* 128, 315–21.
- Irwin, R.E., Pentieva, K., Cassidy, T., Lees-Murdock, D.J., McLaughlin, M., Prasad, G., McNulty, H., Walsh, C.P., 2016. The interplay between DNA methylation, folate and neurocognitive

development. *Epigenomics* 8, 863–79.

- Irwin, R.E., Thursby, S.-J., Ondičová, M., Pentieva, K., McNulty, H., Richmond, R.C., Caffrey, A., Lees-Murdock, D.J., McLaughlin, M., Cassidy, T., Suderman, M., Relton, C.L., Walsh, C.P., 2019. A randomized controlled trial of folic acid intervention in pregnancy highlights a putative methylation-regulated control element at ZFP57. *Clin. Epigenetics* 11, 31.
- Ising, M., Depping, A.-M., Siebertz, A., Lucae, S., Unschuld, P.G., Kloiber, S., Horstmann, S., Uhr, M., Müller-Myhsok, B., Holsboer, F., 2008. Polymorphisms in the FKBP5 gene region modulate recovery from psychosocial stress in healthy controls. *Eur. J. Neurosci.* 28, 389–398.
- Jacob, R.A., Gretz, D.M., Taylor, P.C., James, S.J., Pogribny, I.P., Miller, B.J., Henning, S.M., Swendseid, M.E., 1998. Moderate folate depletion increases plasma homocysteine and decreases lymphocyte DNA methylation in postmenopausal women. *J. Nutr.* 128, 1204–12.
- Jang, H.S., Shin, W.J., Lee, J.E., Do, J.T., 2017. CpG and non-CpG methylation in epigenetic gene regulation and brain function. *Genes (Basel)*.
- Jjingo, D., Conley, A.B., Yi, S. V., Lunyak, V. V., Jordan, I.K., 2012. On the presence and role of human gene-body DNA methylation. *Oncotarget* 3, 462–474.
- Jiang, L., Schlesinger, F., Davis, C.A., Zhang, Y., Li, R., Salit, M., Gingeras, T.R., Oliver, B., 2011. Synthetic spike-in standards for RNA-seq experiments. *Genome Res.* 21, 1543–51.
- Jin, B., Yao, B., Li, J.-L., Fields, C.R., Delmas, A.L., Liu, C., Robertson, K.D., 2009. DNMT1 and DNMT3B Modulate Distinct Polycomb-Mediated Histone Modifications in Colon Cancer. *Cancer Res.* 69, 7412–7421.
- Johnson, T.B., Coghill, R.D., 1925. Researches on pyrimidines. C111. The discovery of 5-methylcytosine in tuberculinic acid, the nucleic acid of the tubercle bacillus. *J. Am. Chem. Soc.* 47, 2838–2844.
- Jones, M.J., Goodman, S.J., Kobor, M.S., 2015. DNA methylation and healthy human aging. *Aging Cell* 14, 924–32.
- Jones, P.A., 2012. Functions of DNA methylation: islands, start sites, gene bodies and beyond. *Nat. Rev. Genet.* 13, 484–492.
- Joshi, A.A., Struhl, K., 2005. Eaf3 Chromodomain Interaction with Methylated H3-K36 Links Histone Deacetylation to Pol II Elongation. *Mol. Cell* 20, 971–978.
- Joubert, B.R., Felix, J.F., Yousefi, P., Bakulski, K.M., Just, A.C., Breton, C., Reese, S.E., Markunas, C.A., Richmond, R.C., Xu, C.J., Küpers, L.K., Oh, S.S., Hoyo, C., Gruziova, O., Söderhäll, C., Salas, L.A., Baiz, N., Zhang, H., Lepeule, J., Ruiz, C., Ligthart, S., Wang, T., Taylor, J.A., Duijts, L., Sharp, G.C., Jankipersadsing, S.A., Nilsen, R.M., Vaez, A., Fallin, M.D., Hu, D., Litonjua, A.A., Fuemmeler, B.F., Huen, K., Kere, J., Kull, I., Munthe-Kaas, M.C., Gehring, U., Bustamante, M., Saurel-Coubizolles, M.J., Quraishi, B.M., Ren, J., Tost, J., Gonzalez, J.R., Peters, M.J., Håberg, S.E., Xu, Z., Van Meurs, J.B., Gaunt, T.R., Kerkhof, M., Corpeleijn, E., Feinberg, A.P., Eng, C., Baccarelli, A.A., Benjamin Neelon, S.E., Bradman, A., Merid, S.K., Bergström, A., Herceg, Z., Hernandez-Vargas, H., Brunekreef, B., Pinart, M., Heude, B., Ewart, S., Yao, J., Lemonnier, N., Franco, O.H., Wu, M.C., Hofman, A., McArdle, W., Van Der Vlies, P., Falahi, F., Gillman, M.W., Barcellos, L.F., Kumar, A., Wickman, M., Guerra, S., Charles, M.A., Holloway, J., Auffray, C., Tiemeier, H.W., Smith, G.D., Postma, D., Hivert, M.F., Eskenazi, B., Vrijheid, M., Arshad, H., Antó, J.M., Dehghan, A., Karmaus, W., Annesi-Maesano, I., Sunyer, J., Ghantous, A., Pershagen, G., Holland, N., Murphy, S.K., Demeo, D.L., Burchard, E.G., Ladd-Acosta, C., Snieder, H., Nystad, W., Koppelman, G.H., Relton, C.L., Jaddoe, V.W.V., Wilcox, A., Melén, E., London, S.J., 2016. DNA Methylation in Newborns and Maternal Smoking in Pregnancy: Genome-wide Consortium Meta-analysis. *Am. J. Hum. Genet.* 98, 680–696.

- Jühling, F., Kretzmer, H., Bernhart, S.H., Otto, C., Stadler, P.F., Hoffmann, S., 2016. Metilene: Fast and sensitive calling of differentially methylated regions from bisulfite sequencing data. *Genome Res.* 26, 256–262.
- Jürgens, G., 1985. A group of genes controlling the spatial expression of the bithorax complex in *Drosophila*. *Nature* 316, 153–155.
- Juruena, M.F., Bocharova, M., Agustini, B., Young, A.H., 2018. Atypical depression and non-atypical depression: Is HPA axis function a biomarker? A systematic review. *J. Affect. Disord.* 233, 45–67.
- Kapitonov, V. V., Jurka, J., 2005. RAG1 core and V(D)J recombination signal sequences were derived from Transib transposons. *PLoS Biol.* 3, e181.
- Kawai, J., Shinagawa, A., Shibata, K., Yoshino, M., Itoh, M., Ishii, Y., Arakawa, T., Hara, A., Fukunishi, Y., Konno, H., Adachi, J., Fukuda, S., Aizawa, K., Izawa, M., Nishi, K., Kiyosawa, H., Kondo, S., Yamanaka, I., Saito, T., Okazaki, Y., Gojobori, T., Bono, H., Kasukawa, T., Saito, R., Kadota, K., Matsuda, H., Ashburner, M., Batalov, S., Casavant, T., Fleischmann, W., Gaasterland, T., Gissi, C., King, B., Kochiwa, H., Kuehl, P., Lewis, S., Matsuo, Y., Nikaido, I., Pesole, G., Quackenbush, J., Schriml, L.M., Staubli, F., Suzuki, R., Tomita, M., Wagner, L., Washio, T., Sakai, K., Okido, T., Furuno, M., Aono, H., Baldarelli, R., Barsh, G., Blake, J., Boffelli, D., Bojunga, N., Carninci, P., de Bonaldo, M.F., Brownstein, M.J., Bult, C., Fletcher, C., Fujita, M., Gariboldi, M., Gustincich, S., Hill, D., Hofmann, M., Hume, D.A., Kamiya, M., Lee, N.H., Lyons, P., Marchionni, L., Mashima, J., Mazzarelli, J., Mombaerts, P., Nordone, P., Ring, B., Ringwald, M., Rodriguez, I., Sakamoto, N., Sasaki, H., Sato, K., Schönbach, C., Seya, T., Shibata, Y., Storch, K.-F., Suzuki, H., Toyo-oka, K., Wang, K.H., Weitz, C., Whittaker, C., Wilming, L., Wynshaw-Boris, A., Yoshida, K., Hasegawa, Y., Kawaji, H., Kohtsuki, S., Hayashizaki, Y., Burdett, T., Dylag, M., Emam, I., Ikeo, K., Ishi-i, J., Tamura, T., Gojobori, T., Tateno, Y., Su, A., Wiltshire, T., Batalov, S., Lapp, H., Ching, K., Block, D., Kapushesky, M., Emam, I., Holloway, E., Kurnosov, P., Zorin, A., Malone, J., Wu, C., Orozco, C., Boyer, J., Leglise, M., Goodale, J., Batalov, S., Shiraki, T., Kondo, S., Katayama, S., Waki, K., Kasukawa, T., Kawaji, H., Qi, Y., Liu, Y., Rong, W., Kanamori-Katayama, M., Itoh, M., Kawaji, H., Lassmann, T., Katayama, S., Kojima, M., Forrest, A., Kawaji, H., Rehli, M., Baillie, J.K., Hoon, M., Haberle, V., Carninci, P., Sandelin, A., Lenhard, B., Katayama, S., Shimokawa, K., Ponjavic, J., Kawaji, H., Kasukawa, T., Fukuda, S., Katayama, S., Kai, C., Kawai, J., Suzuki, H., Forrest, A., Nimwegen, E., Daub, C., Balwierz, P., Irvine, K., Kawaji, H., Severin, J., Lizio, M., Waterhouse, A., Katayama, S., Irvine, K., Povey, S., Lovering, R., Bruford, E., Wright, M., Lush, M., Wain, H., Ashburner, M., Ball, C., Blake, J., Botstein, D., Butler, H., Cherry, J., Mungall, C., Torniai, C., Gkoutos, G., Lewis, S., Haendel, M., Smith, B., Ashburner, M., Rosse, C., Bard, J., Bug, W., Ceusters, W., Mitsuhashi, N., Fujieda, K., Tamura, T., Kawamoto, S., Takagi, T., Okubo, K., Rosse, C., Mejino, J., Itoh, M., Kojima, M., Nagao-Sato, S., Saijo, E., Lassmann, T., Kanamori-Katayama, M., Anders, S., Huber, W., Robinson, M., McCarthy, D., Smyth, G., Rayner, T., Rocca-Serra, P., Spellman, P., Causton, H., Farne, A., Holloway, E., Sansone, S., Rocca-Serra, P., Field, D., Maguire, E., Taylor, C., Hofmann, O., Meyer, L., Zweig, A., Hinrichs, A., Karolchik, D., Kuhn, R., Wong, M., Bernstein, B., Stamatoyannopoulos, J., Costello, J., Ren, B., Milosavljevic, A., Meissner, A., Mons, B., Haagen, H., Chichester, C., Hoen, P., Dunnen, J., Ommen, G., Severin, J., Lizio, M., Harshbarger, J., Kawaji, H., Daub, C., Hayashizaki, Y., Andersson, R., Gebhard, C., Miguel-Escalada, I., Hoof, I., Bornholdt, J., Boyd, M., Freeman, T., Goldovsky, L., Brosch, M., Dongen, S., Maziere, P., Grocock, R., Whetzel, P., Noy, N., Shah, N., Alexander, P., Nyulas, C., Tudorache, T., Dongen, S., Abreu-Goodger, C., Bizer, C., Heath, T., Berners-Lee, T., Patrinos, G., Cooper, D., Mulligen, E., Gkantouna, V., Tzimas, G., Tatum, Z., Fujibuchi, W., Kiseleva, L., Taniguchi, T., Harada, H., Horton, P., Yamashita, R., Sugano, S., Suzuki, Y., Nakai, K., Kawaji, H., Lizio, M., Itoh, M., Kanamori-Katayama, M., Kaiho, A., Nishiyori-Sueki, H., Takahashi, H., Lassmann, T., Murata, M., Carninci, P., Kawaji, H., Hayashizaki, Y., Daub, C., Lassmann, T.,

- Hayashizaki, Y., Daub, C., Djebali, S., Davis, C., Merkel, A., Dobin, A., Lassmann, T., Mortazavi, A., Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Quinlan, A., Hall, I., Enright, A., Dongen, S., Ouzounis, C., Beissbarth, T., Speed, T., Subramanian, A., Tamayo, P., Mootha, V., Mukherjee, S., Ebert, B., Gillette, M., Bryne, J., Valen, E., Tang, M., Marstrand, T., Winther, O., Piedade, I., Cock, P., Antao, T., Chang, J., Chapman, B., Cox, C., Dalke, A., Tatum, Z., Roos, M., Gibson, A., Taschner, P., Thompson, M., Schultes, E., Suzuki, T., Nakano-Ikegaya, M., Yabukami-Okuda, H., Hoon, M., Severin, J., Saga-Hatano, S., 2001. Gateways to the FANTOM5 promoter level mammalian expression atlas. *Nature* 409, 685–690.
- Kawano, S., Grassian, A.R., Tsuda, M., Knutson, S.K., Warholic, N.M., Kuznetsov, G., Xu, S., Xiao, Y., Pollock, R.M., Smith, J.S., Kuntz, K.K., Ribich, S., Minoshima, Y., Matsui, J., Copeland, R.A., Tanaka, S., Keilhack, H., 2016. Preclinical evidence of anti-tumor activity induced by EZH2 inhibition in human models of synovial sarcoma. *PLoS One* 11.
- Kazazian, H.H., Wong, C., Youssoufian, H., Scott, A.F., Phillips, D.G., Antonarakis, S.E., 1988. Haemophilia A resulting from de novo insertion of L1 sequences represents a novel mechanism for mutation in man. *Nature* 332, 164–6.
- Kerr, K., McAnaney, H., Flanagan, C., Maxwell, A.P., McKnight, A.J., 2019. Differential methylation as a diagnostic biomarker of rare renal diseases: a systematic review. *BMC Nephrol.* 20, 320.
- Kharchenko, P. V., Tolstorukov, M.Y., Park, P.J., 2008. Design and analysis of ChIP-seq experiments for DNA-binding proteins. *Nat. Biotechnol.* 26, 1351–1359.
- Kidder, B.L., Hu, G., Zhao, K., 2011. ChIP-Seq: Technical considerations for obtaining high-quality data. *Nat. Immunol.*
- Kim, D., Langmead, B., Salzberg, S.L., 2015. HISAT: A fast spliced aligner with low memory requirements. *Nat. Methods* 12, 357–360.
- KIRBY, K.S., 1956. A new method for the isolation of ribonucleic acids from mammalian tissues. *Biochem. J.* 64, 405–408.
- Ko, Y.A., Mohtat, D., Suzuki, M., Park, A.S.D., Izquierdo, M.C., Han, S.Y., Kang, H.M., Si, H., Hostetter, T., Pullman, J.M., Fazzari, M., Verma, A., Zheng, D., Grealley, J.M., Susztak, K., 2013. Cytosine methylation changes in enhancer regions of core pro-fibrotic genes characterize kidney fibrosis development. *Genome Biol.* 14.
- Kornberg, R.D., 1974. Chromatin structure: A repeating unit of histones and DNA. *Science* (80- ). 184, 868–871.
- Kriaucionis, S., Heintz, N., 2009. The nuclear DNA base 5-hydroxymethylcytosine is present in Purkinje neurons and the brain. *Science* 324, 929–30.
- Krueger, F., 2012. Trim Galore!
- Krueger, F., Andrews, S.R., 2011a. Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. *Bioinformatics* 27, 1571–2.
- Krueger, F., Andrews, S.R., 2011b. Bismark: A flexible aligner and methylation caller for Bisulfite-Seq applications. *Bioinformatics* 27, 1571–1572.
- Ku, M., Koche, R.P., Rheinbay, E., Mendenhall, E.M., Endoh, M., Mikkelsen, T.S., Presser, A., Nusbaum, C., Xie, X., Chi, A.S., Adli, M., Kasif, S., Ptaszek, L.M., Cowan, C.A., Lander, E.S., Koseki, H., Bernstein, B.E., 2008. Genomewide analysis of PRC1 and PRC2 occupancy identifies two classes of bivalent domains. *PLoS Genet.* 4, e1000242.
- Kukurba, K.R., Montgomery, S.B., 2015. Topic Introduction RNA Sequencing and Analysis.



- Kulis, M., Heath, S., Bibikova, M., Queirós, A.C., Navarro, A., Clot, G., Martínez-Trillos, A., Castellano, G., Brun-Heath, I., Pinyol, M., Barberán-Soler, S., Papasaikas, P., Jares, P., Beà, S., Rico, D., Ecker, S., Rubio, M., Royo, R., Ho, V., Klotzle, B., Hernández, L., Conde, L., López-Guerra, M., Colomer, D., Villamor, N., Aymerich, M., Rozman, M., Bayes, M., Gut, M., Gelpí, J.L., Orozco, M., Fan, J.-B., Quesada, V., Puente, X.S., Pisano, D.G., Valencia, A., López-Guillermo, A., Gut, I., López-Otín, C., Campo, E., Martín-Subero, J.I., 2012. Epigenomic analysis detects widespread gene-body DNA hypomethylation in chronic lymphocytic leukemia. *Nat. Genet.* 44, 1236–1242.
- Kuzmichev, A., Nishioka, K., Erdjument-Bromage, H., Tempst, P., Reinberg, D., 2002. Histone methyltransferase activity associated with a human multiprotein complex containing the enhancer of zeste protein. *Genes Dev.* 16, 2893–2905.
- Labonté, B., Suderman, M., Maussion, G., Lopez, J.P., Navarro-Sánchez, L., Yerko, V., Mechawar, N., Szyf, M., Meaney, M.J., Turecki, G., 2013. Genome-Wide Methylation Changes in the Brains of Suicide Completers. *Am. J. Psychiatry* 170, 511–520.
- Labonte, B., Yerko, V., Gross, J., Mechawar, N., Meaney, M.J., Szyf, M., Turecki, G., 2012. Differential Glucocorticoid Receptor Exon 1B, 1C, and 1H Expression and Methylation in Suicide Completers with a History of Childhood Abuse. *Biol. Psychiatry* 72, 41–48.
- Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., Fitzhugh, W., Funke, R., Gage, D., Harris, K., Heaford, A., Howland, J., Kann, L., Lehoczky, J., Levine, R., McEwan, P., McKernan, K., Meldrim, J., Mesirov, J.P., Miranda, C., Morris, W., Naylor, J., Raymond, C., Rosetti, M., Santos, R., Sheridan, A., Sougnez, C., Stange-Thomann, N., Stojanovic, N., Subramanian, A., Wyman, D., Rogers, J., Sulston, J., Ainscough, R., Beck, S., Bentley, D., Burton, J., Clee, C., Carter, N., Coulson, A., Deadman, R., Deloukas, P., Dunham, A., Dunham, I., Durbin, R., French, L., Grafham, D., Gregory, S., Hubbard, T., Humphray, S., Hunt, A., Jones, M., Lloyd, C., McMurray, A., Matthews, L., Mercer, S., Milne, S., Mullikin, J.C., Mungall, A., Plumb, R., Ross, M., Shownkeen, R., Sims, S., Waterston, R.H., Wilson, R.K., Hillier, L.W., McPherson, J.D., Marra, M.A., Mardis, E.R., Fulton, L.A., Chinwalla, A.T., Pepin, K.H., Gish, W.R., Chissoe, S.L., Wendl, M.C., Delehaunty, K.D., Miner, T.L., Delehaunty, A., Kramer, J.B., Cook, L.L., Fulton, R.S., Johnson, D.L., Minx, P.J., Clifton, S.W., Hawkins, T., Branscomb, E., Predki, P., Richardson, P., Wenning, S., Slezak, T., Doggett, N., Cheng, J.F., Olsen, A., Lucas, S., Elkin, C., Uberbacher, E., Frazier, M., Gibbs, R.A., Muzny, D.M., Scherer, S.E., Bouck, J.B., Sodergren, E.J., Worley, K.C., Rives, C.M., Gorrell, J.H., Metzker, M.L., Naylor, S.L., Kucherlapati, R.S., Nelson, D.L., Weinstock, G.M., Sakaki, Y., Fujiyama, A., Hattori, M., Yada, T., Toyoda, A., Itoh, T., Kawagoe, C., Watanabe, H., Totoki, Y., Taylor, T., Weissenbach, J., Heilig, R., Saurin, W., Artiguenave, F., Brottier, P., Bruls, T., Pelletier, E., Robert, C., Wincker, P., Rosenthal, A., Platzer, M., Nyakatura, G., Taudien, S., Rump, A., Smith, D.R., Doucette-Stamm, L., Rubenfield, M., Weinstock, K., Hong, M.L., Dubois, J., Yang, H., Yu, J., Wang, J., Huang, G., Gu, J., Hood, L., Rowen, L., Madan, A., Qin, S., Davis, R.W., Federspiel, N.A., Abola, A.P., Proctor, M.J., Roe, B.A., Chen, F., Pan, H., Ramser, J., Lehrach, H., Reinhardt, R., McCombie, W.R., De La Bastide, M., Dedhia, N., Blöcker, H., Hornischer, K., Nordsiek, G., Agarwala, R., Aravind, L., Bailey, J.A., Bateman, A., Batzoglou, S., Birney, E., Bork, P., Brown, D.G., Burge, C.B., Cerutti, L., Chen, H.C., Church, D., Clamp, M., Copley, R.R., Doerks, T., Eddy, S.R., Eichler, E.E., Furey, T.S., Galagan, J., Gilbert, J.G.R., Harmon, C., Hayashizaki, Y., Haussler, D., Hermjakob, H., Hokamp, K., Jang, W., Johnson, L.S., Jones, T.A., Kasif, S., Kasprzyk, A., Kennedy, S., Kent, W.J., Kitts, P., Koonin, E. V., Korf, I., Kulp, D., Lancet, D., Lowe, T.M., McLysaght, A., Mikkelsen, T., Moran, J. V., Mulder, N., Pollara, V.J., Ponting, C.P., Schuler, G., Schultz, J., Slater, G., Smit, A.F.A., Stupka, E., Szustakowki, J., Thierry-Mieg, D., Thierry-Mieg, J., Wagner, L., Wallis, J., Wheeler, R., Williams, A., Wolf, Y.I., Wolfe, K.H., Yang, S.P., Yeh, R.F., Collins, F., Guyer, M.S., Peterson, J., Felsenfeld, A., Wetterstrand, K.A., Myers, R.M., Schmutz, J., Dickson, M., Grimwood, J., Cox, D.R., Olson, M. V., Kaul, R., Raymond, C., Shimizu, N.,

- Kawasaki, K., Minoshima, S., Evans, G.A., Athanasiou, M., Schultz, R., Patrinos, A., Morgan, M.J., 2001. Initial sequencing and analysis of the human genome. *Nature* 409, 860–921.
- Landt, S.G., Marinov, G.K., Kundaje, A., Kheradpour, P., Pauli, F., Batzoglou, S., Bernstein, B.E., Bickel, P., Brown, J.B., Cayting, P., Chen, Y., DeSalvo, G., Epstein, C., Fisher-Aylor, K.I., Euskirchen, G., Gerstein, M., Gertz, J., Hartemink, A.J., Hoffman, M.M., Iyer, V.R., Jung, Y.L., Karmakar, S., Kellis, M., Kharchenko, P. V, Li, Q., Liu, T., Liu, X.S., Ma, L., Milosavljevic, A., Myers, R.M., Park, P.J., Pazin, M.J., Perry, M.D., Raha, D., Reddy, T.E., Rozowsky, J., Shores, N., Sidow, A., Slattery, M., Stamatoyannopoulos, J.A., Tolstorukov, M.Y., White, K.P., Xi, S., Farnham, P.J., Lieb, J.D., Wold, B.J., Snyder, M., 2012. ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome Res.* 22, 1813–31.
- Langmead, B., Trapnell, C., Pop, M., Salzberg, S.L., 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* 10, R25.
- Larsen, F., Gundersen, G., Lopez, R., Prydz, H., 1992. CpG islands as gene markers in the human genome. *Genomics* 13, 1095–1107.
- Laugesen, A., Højfeldt, J.W., Helin, K., 2016. Role of the polycomb repressive complex 2 (PRC2) in transcriptional regulation and cancer. *Cold Spring Harb. Perspect. Med.* 6.
- Lawrence, M., Daujat, S., Schneider, R., 2016. Lateral Thinking: How Histone Modifications Regulate Gene Expression. *Trends Genet.*
- Lee, E.-J., Rath, P., Liu, J., Ryu, D., Pei, L., Noonepalle, S.K., Shull, A.Y., Feng, Q., Litofsky, N.S., Miller, D.C., Anthony, D.C., Kirk, M.D., Lattera, J., Deng, L., Xin, H.-B., Wang, X., Choi, J.-H., Shi, H., 2015. Identification of Global DNA Methylation Signatures in Glioblastoma-Derived Cancer Stem Cells. *J. Genet. Genomics* 42, 355–71.
- Lee, K.W.K., Richmond, R., Hu, P., French, L., Shin, J., Bourdon, C., Reischl, E., Waldenberger, M., Zeilinger, S., Gaunt, T., McArdle, W., Ring, S., Woodward, G., Bouchard, L., Gaudet, D., Smith, G.D., Relton, C., Paus, T., Pausova, Z., 2015. Prenatal exposure to maternal cigarette smoking and DNA methylation: Epigenome-wide association in a discovery sample of adolescents and replication in an independent cohort at birth through 17 years of age. *Environ. Health Perspect.* 123, 193–199.
- Leek, J.T., Johnson, W.E., Parker, H.S., Jaffe, A.E., Storey, J.D., 2012. The sva package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics* 28, 882–3.
- Leek, J.T., Johnson, W.E., Parker, H.S., Jaffe, A.E., Storey, J.D., 2012. The SVA package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics* 28, 882–883.
- Leek, J.T., Storey, J.D., 2007. Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS Genet.* 3, 1724–1735.
- Levin, J.Z., Yassour, M., Adiconis, X., Nusbaum, C., Thompson, D.A., Friedman, N., Gnirke, A., Regev, A., 2010. Comprehensive comparative analysis of strand-specific RNA sequencing methods. *Nat. Methods* 7, 709–715.
- Li, H., 2013. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM.
- Li, H., Durbin, R., 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25, 1754–1760.
- Li, X., Meng, Q., Rosen, E.M., Fan, S., 2011. UHRF1 confers radioresistance to human breast cancer

- cells. *Int. J. Radiat. Biol.* 87, 263–73.
- Li, Z., Nie, F., Wang, S., Li, L., 2011. Histone H4 Lys 20 monomethylation by histone methylase SET8 mediates Wnt target gene activation. *Proc. Natl. Acad. Sci.* 108, 3116–3123.
- Liao, J., Karnik, R., Gu, H., Ziller, M.J., Clement, K., Tsankov, A.M., Akopian, V., Gifford, C.A., Donaghey, J., Galonska, C., Pop, R., Reyon, D., Tsai, S.Q., Mallard, W., Joung, J.K., Rinn, J.L., Gnirke, A., Meissner, A., 2015. Targeted disruption of DNMT1, DNMT3A and DNMT3B in human embryonic stem cells. *Nat. Genet.* 47, 469–478.
- Lin, C., Yang, L., Tanasa, B., Hutt, K., Ju, B. gun, Ohgi, K., Zhang, J., Rose, D.W., Fu, X.D., Glass, C.K., Rosenfeld, M.G., 2009. Nuclear Receptor-Induced Chromosomal Proximity and DNA Breaks Underlie Specific Translocations in Cancer. *Cell* 139, 1069–1083.
- Lindroth, A.M., Cao, X., Jackson, J.P., Zilberman, D., McCallum, C.M., Henikoff, S., Jacobsen, S.E., 2001. Requirement of CHROMOMETHYLASE3 for Maintenance of CpXpG Methylation. *Science* (80-. ). 292, 2077–2080.
- Lister, R., Mukamel, E.A., Nery, J.R., Urich, M., Puddifoot, C.A., Johnson, N.D., Lucero, J., Huang, Y., Dwork, A.J., Schultz, M.D., Yu, M., Tonti-Filippini, J., Heyn, H., Hu, S., Wu, J.C., Rao, A., Esteller, M., He, C., Haghghi, F.G., Sejnowski, T.J., Behrens, M.M., Ecker, J.R., 2013. Global Epigenomic Reconfiguration During Mammalian Brain Development. *Science* (80-. ). 341, 1237905–1237905.
- Lister, R., Pelizzola, M., Downen, R.H., Hawkins, R.D., Hon, G., Tonti-Filippini, J., Nery, J.R., Lee, L., Ye, Z., Ngo, Q.-M., Edsall, L., Antosiewicz-Bourget, J., Stewart, R., Ruotti, V., Millar, A.H., Thomson, J.A., Ren, B., Ecker, J.R., 2009. Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature* 462, 315.
- Liu, L., Wang, W., Sun, J., Pang, Z., 2018. Association of famine exposure during early life with the risk of type 2 diabetes in adulthood: a meta-analysis. *Eur. J. Nutr.* 57, 741–749.
- Liu, X., Gao, Q., Li, P., Zhao, Q., Zhang, J., Li, J., Koseki, H., Wong, J., 2013. UHRF1 targets DNMT1 for DNA methylation through cooperative binding of hemi-methylated DNA and methylated H3K9. *Nat. Commun.* 4.
- Liu, X., Yang, J., Wu, N., Song, R., Zhu, H., 2015. Evolution and Coevolution of PRC2 Genes in Vertebrates and Mammals. In: *Advances in Protein Chemistry and Structural Biology*. Academic Press Inc., pp. 125–148.
- López, V., Fernández, A.F., Fraga, M.F., 2017. The role of 5-hydroxymethylcytosine in development, aging and age-related diseases. *Ageing Res. Rev.* 37, 28–38.
- Lorincz, M.C., Dickerson, D.R., Schmitt, M., Groudine, M., 2004. Intragenic DNA methylation alters chromatin structure and elongation efficiency in mammalian cells. *Nat. Struct. Mol. Biol.* 11, 1068–1075.
- Loughery, J.E.P., Dunne, P.D., O’Neill, K.M., Meehan, R.R., McDaid, J.R., Walsh, C.P., 2011. DNMT1 deficiency triggers mismatch repair defects in human cells through depletion of repair protein levels in a process involving the DNA damage response. *Hum. Mol. Genet.* 20, 3241–3255.
- Love, M.I., Huber, W., Anders, S., 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* 15.
- Luco, R.F., Allo, M., Schor, I.E., Kornblihtt, A.R., Misteli, T., 2011. Epigenetics in alternative pre-mRNA splicing. *Cell* 144, 16–26.
- Lumey, L., Terry, M.B., Delgado-Cruzata, L., Liao, Y., Wang, Q., Susser, E., McKeague, I., Santella,

- R.M., 2012. Adult global DNA methylation in relation to pre-natal nutrition. *Int. J. Epidemiol.* 41, 116–123.
- Lyko, F., 2017. The DNA methyltransferase family: a versatile toolkit for epigenetic regulation. *Nat. Rev. Genet.* 19, 81–92.
- Lyon, M.F., 2006. Do LINEs have a role in X-chromosome inactivation? *J. Biomed. Biotechnol.*
- Ma, H., Morey, R., O’Neil, R.C., He, Y., Daughtry, B., Schultz, M.D., Hariharan, M., Nery, J.R., Castanon, R., Sabatini, K., Thiagarajan, R.D., Tachibana, M., Kang, E., Tippner-Hedges, R., Ahmed, R., Gutierrez, N.M., Van Dyken, C., Polat, A., Sugawara, A., Sparman, M., Gokhale, S., Amato, P., Wolf, D.P., Ecker, J.R., Laurent, L.C., Mitalipov, S., 2014. Abnormalities in human pluripotent cells due to reprogramming mechanisms. *Nature* 511, 177–83.
- MacDonald, J.R., Ziman, R., Yuen, R.K.C., Feuk, L., Scherer, S.W., 2014. The Database of Genomic Variants: A curated collection of structural variation in the human genome. *Nucleic Acids Res.* 42.
- Maenohara, S., Unoki, M., Toh, H., Ohishi, H., Sharif, J., Koseki, H., Sasaki, H., 2017. Role of UHRF1 in de novo DNA methylation in oocytes and maintenance methylation in preimplantation embryos. *PLoS Genet.* 13.
- Magnus, M.C., Håberg, S.E., Karlstad, Ø., Nafstad, P., London, S.J., Nystad, W., 2015. Grandmother’s smoking when pregnant with the mother and asthma in the grandchild: the Norwegian Mother and Child Cohort Study. *Thorax* 70, 237–43.
- Maksimovic, J., Gordon, L., Oshlack, A., 2012. SWAN: Subset-quantile within array normalization for illumina infinium HumanMethylation450 BeadChips. *Genome biology*, 13(6), R44. *Genome Biol.* 13, R44.
- Maunakea, A.K., Nagarajan, R.P., Bilenky, M., Ballinger, T.J., Dsouza, C., Fouse, S.D., Johnson, B.E., Hong, C., Nielsen, C., Zhao, Y., Turecki, G., Delaney, A., Varhol, R., Thiessen, N., Shchors, K., Heine, V.M., Rowitch, D.H., Xing, X., Fiore, C., Schillebeeckx, M., Jones, S.J.M., Haussler, D., Marra, M.A., Hirst, M., Wang, T., Costello, J.F., 2010. Conserved role of intragenic DNA methylation in regulating alternative promoters. *Nature* 466, 253–257.
- McCLINTOCK, B., 1951. Chromosome organization and genic expression. *Cold Spring Harb. Symp. Quant. Biol.* 16, 13–47.
- McEwen, B.S., Gray, J.D., Nasca, C., 2015. Redefining neuroendocrinology: Stress, sex and cognitive and emotional regulation. *J. Endocrinol.*
- McGarel, C., McNulty, H., Strain, J.J., Cassidy, T., Mcloughlin, M., McNulty, B., Rollins, M., Marshall, B., Ward, M., Molloy, A.M., Pentieva, K., 2017. Effect of folic acid supplementation during pregnancy on cognitive development of the child at 6 years: preliminary results from the FASSTT Offspring Trial. *Br J Nutr Am J Clin Nutr JAMA* 103, 445–452.
- McIntyre, L.M., Lopiano, K.K., Morse, A.M., Amin, V., Oberg, A.L., Young, L.J., Nuzhdin, S. V., 2011. RNA-seq: technical variability and sampling. *BMC Genomics* 12, 293.
- McKenna, Aaron, Matthew Hanna, Eric Banks, Andrey Sivachenko, Kristian Cibulskis, Andrew Kernytsky, Kiran Garimella, David Altshuler, Stacey Gabriel, M.D., DePristo, and M.A., McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., DePristo, M.A., 2010. The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* 20, 254–260.

- McNulty, B., Pentieva, K., Marshall, B., Ward, M., Molloy, A.M., Scott, J.M., McNulty, H., 2011. Women's compliance with current folic acid recommendations and achievement of optimal vitamin status for preventing neural tube defects. *Hum. Reprod.* 26, 1530–1536.
- Meissner, A., Gnirke, A., Bell, G.W., Ramsahoye, B., Lander, E.S., Jaenisch, R., 2005. Reduced representation bisulfite sequencing for comparative high-resolution DNA methylation analysis. *Nucleic Acids Res.* 33, 5868–5877.
- Meng, H., Cao, Y., Qin, J., Song, X., Zhang, Q., Shi, Y., Cao, L., 2015. DNA methylation, its mediators and genome integrity. *Int. J. Biol. Sci.* 11, 604–17.
- Meyer, C.A., Liu, X.S., 2014. Identifying and mitigating bias in next-generation sequencing methods for chromatin biology. *Nat. Rev. Genet.*
- Moritz, L.E., Trievel, R.C., 2018. Structure, mechanism, and regulation of polycomb-repressive complex 2. *J. Biol. Chem.*
- Morris, T.J., Beck, S., 2015a. Analysis pipelines and packages for Infinium HumanMethylation450 BeadChip (450k) data. *Methods* 72, 3–8.
- Morris, T.J., Beck, S., 2015b. Analysis pipelines and packages for Infinium HumanMethylation450 BeadChip (450k) data. *Methods* 72, 3–8.
- Morris, T.J., Butcher, L.M., Feber, A., Teschendorff, A.E., Chakravarthy, A.R., Wojdacz, T.K., Beck, S., 2014. ChAMP: 450k Chip Analysis Methylation Pipeline. *Bioinformatics* 30, 428–430.
- Mortusewicz, O., Schermelleh, L., Walter, J., Cardoso, M.C., Leonhardt, H., 2005. Recruitment of DNA methyltransferase I to DNA repair sites. *Proc. Natl. Acad. Sci. U. S. A.* 102, 8905–9.
- MRC Vitamin Study Research Group, 1991. Prevention of neural tube defects: results of the Medical Research Council Vitamin Study. *MRC Vitamin Study Research Group. Lancet* 338, 131–137.
- Mulder, R.H., Rijlaarsdam, J., Luijk, M.P.C.M., Verhulst, F.C., Felix, J.F., Tiemeier, H., Bakermans-Kranenburg, M.J., Van Ijzendoorn, M.H., 2017. Methylation matters: FK506 binding protein 51 (FKBP5) methylation moderates the associations of FKBP5 genotype and resistant attachment with stress regulation. *Dev. Psychopathol.* 29, 491–503.
- Müller, F., Scherer, M., Assenov, Y., Lutsik, P., Walter, J., Lengauer, T., Bock, C., 2019. RnBeads 2.0: comprehensive analysis of DNA methylation data. *Genome Biol.* 20, 55.
- Murphy, T.M., Crawford, B., Dempster, E.L., Hannon, E., Burrage, J., Turecki, G., Kaminsky, Z., Mill, J., 2017. Methyloomic profiling of cortex samples from completed suicide cases implicates a role for PSORS1C3 in major depression and suicide. *Transl. Psychiatry* 7.
- O'Carroll, D., Erhardt, S., Pagani, M., Barton, S.C., Surani, M.A., Jenuwein, T., 2001. The polycomb-group gene *Ezh2* is required for early mouse development. *Mol. Cell. Biol.* 21, 4330–6.
- O'Neill, K.M., Irwin, R.E., Mackin, S.-J., Thursby, S.-J., Thakur, A., Bertens, C., Masala, L., Loughery, J.E.P., McArt, D.G., Walsh, C.P., 2018. Depletion of DNMT1 in differentiated human cells highlights key classes of sensitive genes and an interplay with polycomb repression. *Epigenetics Chromatin* 11, 12.
- O'Neill, K.M., Irwin, R.E., Mackin, S.-J., Thursby, S.-J., Thakur, A., Bertens, C., Masala, L., Loughery, J.E.P., McArt, D.G., Walsh, C.P., 2018. Depletion of DNMT1 in differentiated human cells highlights key classes of sensitive genes and an interplay with polycomb repression. *Epigenetics and Chromatin* 11.
- Oberlander, T.F., Weinberg, J., Papsdorf, M., Grunau, R., Misri, S., Devlin, A.M., 2008. Prenatal

- exposure to maternal depression, neonatal methylation of human glucocorticoid receptor gene (NR3C1) and infant cortisol stress responses. *Epigenetics* 3, 97–106.
- Ooi, S.K.T., Qiu, C., Bernstein, E., Li, K., Jia, D., Yang, Z., Erdjument-Bromage, H., Tempst, P., Lin, S.P., Allis, C.D., Cheng, X., Bestor, T.H., 2007. DNMT3L connects unmethylated lysine 4 of histone H3 to de novo methylation of DNA. *Nature* 448, 714–717.
- Ouellette, M.M., McDaniel, L.D., Wright, W.E., Shay, J.W., Schultz, R.A., 2000. The establishment of telomerase-immortalized cell lines representing human chromosome instability syndromes. *Hum. Mol. Genet.* 9, 403–11.
- Paquette, A.G., Lester, B.M., Koestler, D.C., Lesueur, C., Armstrong, D.A., Marsit, C.J., 2014. Placental FKBP5 Genetic and Epigenetic Variation Is Associated with Infant Neurobehavioral Outcomes in the RICHS Cohort. *PLoS One* 9, e104913.
- Park, P.J., 2009. ChIP-seq: Advantages and challenges of a maturing technology. *Nat. Rev. Genet.*
- Pasini, D., Bracken, A.P., Jensen, M.R., Lazzerini Denchi, E., Helin, K., 2004. Suz12 is essential for mouse development and for EZH2 histone methyltransferase activity. *EMBO J.* 23, 4061–71.
- Penn, N.W., Suwalski, R., O’Riley, C., Bojanowski, K., Yura, R., 1972. The presence of 5-hydroxymethylcytosine in animal deoxyribonucleic acid. *Biochem. J.* 126, 781–790.
- Pentieva, K., McGarel, C., McNulty, B., Ward, M., Elliott, N., Strain, J.J., Rollins, M.D., McNulty, H., Czeizel, A.E., Dudás, I., Julvez, J., Fortuny, J., Mendez, M., Torrent, M., Ribas-Fitó, N., Sunyer, J., 2012. Effect of folic acid supplementation during pregnancy on growth and cognitive development of the offspring: a pilot follow-up investigation of children of FASSTT study participants. *Proc. Nutr. Soc.* 71, E139.
- Perry, G.H., Dominy, N.J., Claw, K.G., Lee, A.S., Fiegler, H., Redon, R., Werner, J., Villanea, F.A., Mountain, J.L., Misra, R., Carter, N.P., Lee, C., Stone, A.C., 2007. Diet and the evolution of human amylase gene copy number variation. *Nat. Genet.* 39, 1256–1260.
- Pfeifer, G.P., n.d. Mutagenesis at Methylated CpG Sequences. In: *DNA Methylation: Basic Mechanisms*. Springer-Verlag, Berlin/Heidelberg, pp. 259–281.
- Pidsley, R., Zotenko, E., Peters, T.J., Lawrence, M.G., Risbridger, G.P., Molloy, P., Van Dijk, S., Muhlhauser, B., Stirzaker, C., Clark, S.J., 2016. Critical evaluation of the Illumina MethylationEPIC BeadChip microarray for whole-genome DNA methylation profiling. *Genome Biol.* 17, 208.
- Pieper, H.C., Evert, B.O., Kaut, O., Riederer, P.F., Waha, A., Wüllner, U., 2008. Different methylation of the TNF-alpha promoter in cortex and substantia nigra: Implications for selective neuronal vulnerability. *Neurobiol. Dis.* 32, 521–527.
- Pinheiro, I., Heard, E., 2017. X chromosome inactivation: New players in the initiation of gene silencing. *F1000Research*.
- Pinto, D., Delaby, E., Merico, D., Barbosa, M., Merikangas, A., Klei, L., Thiruvahindrapuram, B., Xu, X., Ziman, R., Wang, Z., Vorstman, J.A.S., Thompson, A., Regan, R., Pilorge, M., Pellecchia, G., Pagnamenta, A.T., Oliveira, B., Marshall, C.R., Magalhaes, T.R., Lowe, J.K., Howe, J.L., Griswold, A.J., Gilbert, J., Duketis, E., Dombroski, B.A., De Jonge, M. V, Cuccaro, M., Crawford, E.L., Correia, C.T., Conroy, J., Conceição, I.C., Chiochetti, A.G., Casey, J.P., Cai, G., Cabrol, C., Bolshakova, N., Bacchelli, E., Anney, R., Gallinger, S., Cotterchio, M., Casey, G., Zwaigenbaum, L., Wittemeyer, K., Wing, K., Wallace, S., van Engeland, H., Tryfon, A., Thomson, S., Soorya, L., Rogé, B., Roberts, W., Poustka, F., Mouga, S., Minshew, N., McInnes, L.A., McGrew, S.G., Lord, C., Leboyer, M., Le Couteur, A.S., Kolevzon, A., Jiménez González, P., Jacob, S., Holt, R., Guter,

- S., Green, J., Green, A., Gillberg, C., Fernandez, B.A., Duque, F., Delorme, R., Dawson, G., Chaste, P., Café, C., Brennan, S., Bourgeron, T., Bolton, P.F., Bölte, S., Bernier, R., Baird, G., Bailey, A.J., Anagnostou, E., Almeida, J., Wijsman, E.M., Vieland, V.J., Vicente, A.M., Schellenberg, G.D., Pericak-Vance, M., Paterson, A.D., Parr, J.R., Oliveira, G., Nurnberger, J.I., Monaco, A.P., Maestrini, E., Klauck, S.M., Hakonarson, H., Haines, J.L., Geschwind, D.H., Freitag, C.M., Folstein, S.E., Ennis, S., Coon, H., Battaglia, A., Szatmari, P., Sutcliffe, J.S., Hallmayer, J., Gill, M., Cook, E.H., Buxbaum, J.D., Devlin, B., Gallagher, L., Betancur, C., Scherer, S.W., 2014. Convergence of genes and cellular pathways dysregulated in autism spectrum disorders. *Am. J. Hum. Genet.* 94, 677–94.
- Plank, J.L., Dean, A., 2014. Enhancer function: Mechanistic and genome-wide insights come together. *Mol. Cell.*
- Price, E.M., Robinson, W.P., 2018. Adjusting for batch effects in DNA methylation microarray data, a lesson learned. *Front. Genet.* 9.
- Qu, Y., Siggens, L., Cordeddu, L., Gaidzik, V.I., Karlsson, K., Bullinger, L., Döhner, K., Ekwall, K., Lehmann, S., Lennartsson, A., 2017. Cancer-specific changes in DNA methylation reveal aberrant silencing and activation of enhancers in leukemia. *Blood* 129, e13–e25.
- R-Core-Team, 2013. R: a language and environment for statistical computing.
- Rajakumara, E., Wang, Z., Ma, H., Hu, L., Chen, H., Lin, Y., Guo, R., Wu, F., Li, H., Lan, F., Shi, Y.G., Xu, Y., Patel, D.J., Shi, Y., 2011. PHD Finger Recognition of Unmodified Histone H3R2 Links UHRF1 to Regulation of Euchromatic Gene Expression. *Mol. Cell* 43, 275–284.
- Ramsahoye, B.H., Biniszkiwicz, D., Lyko, F., Clark, V., Bird, A.P., Jaenisch, R., 2000a. Non-CpG methylation is prevalent in embryonic stem cells and may be mediated by DNA methyltransferase 3a. *Proc. Natl. Acad. Sci. U. S. A.* 97, 5237–5242.
- Ramsahoye, B.H., Biniszkiwicz, D., Lyko, F., Clark, V., Bird, A.P., Jaenisch, R., 2000. Non-CpG methylation is prevalent in embryonic stem cells and may be mediated by DNA methyltransferase 3a. *Proc. Natl. Acad. Sci. U. S. A.* 97, 5237–42.
- Ramsahoye, B.H., Biniszkiwicz, D., Lyko, F., Clark, V., Bird, A.P., Jaenisch, R., 2000b. Non-CpG methylation is prevalent in embryonic stem cells and may be mediated by DNA methyltransferase 3a. *Proc. Natl. Acad. Sci. U. S. A.* 97, 5237–5242.
- Raz, T., Kapranov, P., Lipson, D., Letovsky, S., Milos, P.M., Thompson, J.F., 2011. Protocol dependence of sequencing-based gene expression measurements. *PLoS One* 6, e19287.
- Rea, S., Eisenhaber, F., O’Carroll, D., Strahl, B.D., Sun, Z.W., Schmid, M., Opravil, S., Mechtler, K., Ponting, C.P., Allis, C.D., Jenuwein, T., 2000. Regulation of chromatin structure by site-specific histone H3 methyltransferases. *Nature* 406, 593–599.
- Reddington, J.P., Perricone, S.M., Nestor, C.E., Reichmann, J., Youngson, N.A., Suzuki, M., Reinhardt, D., Dunican, D.S., Prendergast, J.G., Mjoseng, H., Ramsahoye, B.H., Whitelaw, E., Grealley, J.M., Adams, I.R., Bickmore, W.A., Meehan, R.R., 2013. Redistribution of H3K27me3 upon DNA hypomethylation results in de-repression of Polycomb target genes. *Genome Biol.* 14, R25.
- Reddington, J.P., Perricone, S.M., Nestor, C.E., Reichmann, J., Youngson, N.A., Suzuki, M., Reinhardt, D., Dunican, D.S., Prendergast, J.G., Mjoseng, H., Ramsahoye, B.H., Whitelaw, E., Grealley, J.M., Adams, I.R., Bickmore, W.A., Meehan, R.R., 2013. Redistribution of H3K27me3 upon DNA hypomethylation results in de-repression of Polycomb target genes. *Genome Biol.* 14.
- Rehan, V.K., Liu, J., Sakurai, R., Torday, J.S., 2013. Perinatal nicotine-induced transgenerational asthma. *Am. J. Physiol. Lung Cell. Mol. Physiol.* 305, L501-7.

- Richmond, R.C., Sharp, G.C., Herbert, G., Atkinson, C., Taylor, C., Bhattacharya, S., Campbell, D., Hall, M., Kazmi, N., Gaunt, T., McArdle, W., Ring, S., Davey Smith, G., Ness, A., Relton, C.L., 2018. The long-term impact of folic acid in pregnancy on offspring DNA methylation: follow-up of the Aberdeen Folic Acid Supplementation Trial (AFAST). *Int. J. Epidemiol.*
- Richmond, R.C., Simpkin, A.J., Woodward, G., Gaunt, T.R., Lyttleton, O., McArdle, W.L., Ring, S.M., Smith, A.D.A.C., Timpson, N.J., Tilling, K., Smith, G.D., Relton, C.L., 2015. Prenatal exposure to maternal smoking and offspring DNA methylation across the lifecourse: Findings from the Avon Longitudinal Study of Parents and Children (ALSPAC). *Hum. Mol. Genet.* 24, 2201–2217.
- Rizzardi, L.F., Hickey, P.F., Rodriguez DiBlasi, V., Tryggvadóttir, R., Callahan, C.M., Idrizi, A., Hansen, K.D., Feinberg, A.P., 2019. Neuronal brain-region-specific DNA methylation and chromatin accessibility are associated with neuropsychiatric trait heritability. *Nat. Neurosci.* 22, 307–316.
- Rizzo, A., Napoli, A., Roggiani, F., Tomassetti, A., Bagnoli, M., Mezzanzanica, D., 2018. One-carbon metabolism: Biological players in epithelial ovarian cancer. *Int. J. Mol. Sci.*
- Robinson, J.T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E.S., Getz, G., Mesirov, J.P., 2011. Integrative genomics viewer. *Nat. Biotechnol.* 29, 24–6.
- Roessler, J., Ammerpohl, O., Gutwein, J., Hasemeier, B., Anwar, S.L., Kreipe, H., Lehmann, U., 2012. Quantitative cross-validation and content analysis of the 450k DNA methylation array from Illumina, Inc. *BMC Res. Notes* 5.
- Rondelet, G., Dal Maso, T., Willems, L., Wouters, J., 2016. Structural basis for recognition of histone H3K36me3 nucleosome by human de novo DNA methyltransferases 3A and 3B. *J. Struct. Biol.* 194, 357–367.
- Rothbart, S.B., Krajewski, K., Nady, N., Tempel, W., Xue, S., Badeaux, A.I., Barsyte-Lovejoy, D., Martinez, J.Y., Bedford, M.T., Fuchs, S.M., Arrowsmith, C.H., Strahl, B.D., 2012. Association of UHRF1 with methylated H3K9 directs the maintenance of DNA methylation. *Nat. Struct. Mol. Biol.* 19, 1155–1160.
- Rountree, M.R., Selker, E.U., 1997. DNA methylation inhibits elongation but not initiation of transcription in *Neurospora crassa*. *Genes Dev.* 11, 2383–2395.
- RStudio, R.T., 2015. RStudio: integrated development for R.
- Rutledge, C.E., Thakur, A., O'Neill, K.M., Irwin, R.E., Sato, S., Hata, K., Walsh, C.P., 2014. Ontogeny, conservation and functional significance of maternally inherited DNA methylation at two classes of non-imprinted genes. *Development* 141, 1313–23.
- Saxonov, S., Berg, P., Brutlag, D.L., 2006. A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters. *Proc. Natl. Acad. Sci. U. S. A.* 103, 1412–1417.
- Schotta, G., Lachner, M., Sarma, K., Ebert, A., Sengupta, R., Reuter, G., Reinberg, D., Jenuwein, T., 2004. A silencing pathway to induce H3-K9 and H4-K20 trimethylation at constitutive heterochromatin. *Genes Dev.* 18, 1251–62.
- Schübeler, D., 2015. Function and information content of DNA methylation. *Nature* 517, 321–326.
- Schulz, L.C., 2010. The Dutch hunger winter and the developmental origins of health and disease. *Proc. Natl. Acad. Sci. U. S. A.*
- Sharif, J., Endo, T.A., Nakayama, M., Karimi, M.M., Shimada, M., Katsuyama, K., Goyal, P., Brind'Amour, J., Sun, M.A., Sun, Z., Ishikura, T., Mizutani-Koseki, Y., Ohara, O., Shinkai, Y., Nakanishi, M., Xie, H., Lorincz, M.C., Koseki, H., 2016. Activation of Endogenous Retroviruses in



- Dnmt1<sup>-/-</sup> ESCs Involves Disruption of SETDB1-Mediated Repression by NP95 Binding to Hemimethylated DNA. *Cell Stem Cell* 19, 81–94.
- Sharif, J., Muto, M., Takebayashi, S.I., Suetake, I., Iwamatsu, A., Endo, T.A., Shinga, J., Mizutani-Koseki, Y., Toyoda, T., Okamura, K., Tajima, S., Mitsuya, K., Okano, M., Koseki, H., 2007. The SRA protein Np95 mediates epigenetic inheritance by recruiting Dnmt1 to methylated DNA. *Nature* 450, 908–912.
- Shiio, Y., Eisenman, R.N., 2003. Histone sumoylation is associated with transcriptional repression. *Proc. Natl. Acad. Sci. U. S. A.* 100, 13225–13230.
- Shukla, S., Kavak, E., Gregory, M., Imashimizu, M., Shutinoski, B., Kashlev, M., Oberdoerffer, P., Sandberg, R., Oberdoerffer, S., 2011. CTCF-promoted RNA polymerase II pausing links DNA methylation to splicing. *Nature* 479, 74–9.
- Slotkin, R.K., Martienssen, R., 2007. Transposable elements and the epigenetic regulation of the genome. *Nat. Rev. Genet.*
- Smith, E., Shilatifard, A., 2014. Enhancer biology and enhanceropathies. *Nat. Struct. Mol. Biol.*
- Smyth, G.K., 2004. Linear Models and Empirical Bayes Methods for Assessing Differential Expression in Microarray Experiments. *Stat. Appl. Genet. Mol. Biol.* 3, 1–25.
- Sohn, K.-J., Stempak, J.M., Reid, S., Shirwadkar, S., Mason, J.B., Kim, Y.-I., 2003. The effect of dietary folate on genomic and p53-specific DNA methylation in rat colon. *Carcinogenesis* 24, 81–90.
- Sotero-Caio, C.G., Platt, R.N., Suh, A., Ray, D.A., 2017. Evolution and diversity of transposable elements in vertebrate genomes. *Genome Biol. Evol.* 9, 161–177.
- Spindel, E.R., McEvoy, C.T., 2016. The role of nicotine in the effects of maternal smoking during pregnancy on lung development and childhood respiratory disease: Implications for dangers of e-cigarettes. *Am. J. Respir. Crit. Care Med.*
- Stanner, S.A., Yudkin, J.S., 2001. Fetal Programming and the Leningrad Siege Study. *Twin Res.* 4, 287–292.
- Steegers-Theunissen, R.P., Obermann-Borst, S.A., Kremer, D., Lindemans, J., Siebel, C., Steegers, E.A., Slagboom, P.E., Heijmans, B.T., 2009. Periconceptional Maternal Folic Acid Use of 400 µg per Day Is Related to Increased Methylation of the IGF2 Gene in the Very Young Child. *PLoS One* 4, e7845.
- Stein, A.D., Zybert, P.A., van de Bor, M., Lumey, L.H., 2004. Intrauterine famine exposure and body proportions at birth: The Dutch Hunger Winter. *Int. J. Epidemiol.* 33, 831–836.
- Stranger, B.E., Forrest, M.S., Dunning, M., Ingle, C.E., Beazley, C., Thorne, N., Redon, R., Bird, C.P., de Grassi, A., Lee, C., Tyler-Smith, C., Carter, N., Scherer, S.W., Tavaré, S., Deloukas, P., Hurler, M.E., Dermitzakis, E.T., 2007. Relative impact of nucleotide and copy number variation on gene expression phenotypes. *Supporting Online Material. Science* (80- ). 315, 848–53.
- Sturtevant, A.H., 1925. The Effects of Unequal Crossing over at the Bar Locus in *Drosophila*. *Genetics* 10, 117–47.
- Sudmant, P.H., Kitzman, J.O., Antonacci, F., Alkan, C., Malig, M., Tsalenko, A., Sampas, N., Bruhn, L., Shendure, J., Eichler, E.E., Altshuler, D.L., Durbin, R.M., Abecasis, G.R., Bentley, D.R., Chakravarti, A., Clark, A.G., Collins, F.S., De La Vega, F.M., Donnelly, P., Egholm, M., Flicek, P., Gabriel, S.B., Gibbs, R.A., Knoppers, B.M., Lander, E.S., Lehrach, H., Mardis, E.R., McVean, G.A., Nickerson, D.A., Peltonen, L., Schafer, A.J., Sherry, S.T., Wang, J., Wilson, R.K., Deiros, D., Metzker, M., Muzny, D., Reid, J., Wheeler, D., Li, J., Jian, M., Li, G., Li, R., Liang, H., Tian, G.,

Wang, B., Wang, W., Yang, H., Zhang, X., Zheng, H., Ambrogio, L., Bloom, T., Cibulskis, K., Fennell, T.J., Jaffe, D.B., Shefler, E., Sougnez, C.L., Gormley, N., Humphray, S., Kingsbury, Z., Koko-Gonzales, P., Stone, J., McKernan, K.J., Costa, G.L., Ichikawa, J.K., Lee, C.C., Sudbrak, R., Borodina, T.A., Dahl, A., Davydov, A.N., Marquardt, P., Mertes, F., Nietfeld, W., Rosenstiel, P., Schreiber, S., Soldatov, A. V., Timmermann, B., Tolzmann, M., Affourtit, J., Ashworth, D., Attiya, S., Bachorski, M., Buglione, E., Burke, A., Caprio, A., Celone, C., Clark, S., Conners, D., Desany, B., Gu, L., Guccione, L., Kao, K., Keibel, A., Knowlton, J., Labrecque, M., McDade, L., Mealmaker, C., Minderman, M., Nawrocki, A., Niazi, F., Pareja, K., Ramenani, R., Riches, D., Song, W., Turcotte, C., Wang, S., Dooling, D., Fulton, L., Fulton, R., Weinstock, G., Burton, J., Carter, D.M., Churcher, C., Coffey, A., Cox, A., Palotie, A., Quail, M., Skelly, T., Stalker, J., Swerdlow, H.P., Turner, D., De Witte, A., Giles, S., Bainbridge, M., Challis, D., Sabo, A., Yu, F., Yu, J., Fang, X., Guo, X., Li, Y., Luo, R., Tai, S., Wu, H., Zheng, X., Zhou, Y., Marth, G.T., Garrison, E.P., Huang, W., Indap, A., Kural, D., Lee, W.P., Leong, W.F., Quinlan, A.R., Stewart, C., Stromberg, M.P., Ward, A.N., Wu, J., Lee, C., Mills, R.E., Shi, X., Daly, M.J., DePristo, M.A., Ball, A.D., Banks, E., Browning, B.L., Garimella, K. V., Grossman, S.R., Handsaker, R.E., Hanna, M., Hartl, C., Kernytsky, A.M., Korn, J.M., Li, H., Maguire, J.R., McCarroll, S.A., McKenna, A., Nemesh, J.C., Philippakis, A.A., Poplin, R.E., Price, A., Rivas, M.A., Sabeti, P.C., Schaffner, S.F., Shlyakhter, I.A., Cooper, D.N., Ball, E. V., Mort, M., Phillips, A.D., Stenson, P.D., Sebat, J., Makarov, V., Ye, K., Yoon, S.C., Bustamante, C.D., Boyko, A., Degenhardt, J., Gravel, S., Gutenkunst, R.N., Kaganovich, M., Keinan, A., Lacroute, P., Ma, X., Reynolds, A., Clarke, L., Cunningham, F., Herrero, J., Keenen, S., Kulesha, E., Leinonen, R., McLaren, W.M., Radhakrishnan, R., Smith, R.E., Zalunin, V., Zheng-Bradley, X., Korbil, J.O., Stütz, A.M., Bauer, M., Cheetham, R.K., Cox, T., Eberle, M., James, T., Kahn, S., Murray, L., Fu, Y., Hyland, F.C., Manning, J.M., McLaughlin, S.F., Peckham, H.E., Sakarya, O., Sun, Y.A., Tsung, E.F., Batzer, M.A., Konkel, M.K., Walker, J.A., Albrecht, M.W., Amstislavskiy, V.S., Herwig, R., Parkhomchuk, D. V., Agarwala, R., Khouri, H.M., Morgulis, A.O., Paschall, J.E., Phan, L.D., Rotmistrovsky, K.E., Sanders, R.D., Shumway, M.F., Xiao, C., Auton, A., Iqbal, Z., Lunter, G., Marchini, J.L., Moutsianas, L., Myers, S., Tumian, A., Knight, J., Winer, R., Craig, D.W., Beckstrom-Sternberg, S.M., Christoforides, A., Kurdoglu, A.A., Pearson, J. V., Sinari, S.A., Tembe, W.D., Haussler, D., Hinrichs, A.S., Katzman, S.J., Kern, A., Kuhn, R.M., Przeworski, M., Hernandez, R.D., Howie, B., Kelley, J.L., Melton, S.C., Anderson, P., Blackwell, T., Chen, W., Cookson, W.O., Ding, J., Kang, H.M., Lathrop, M., Liang, L., Moffatt, M.F., Scheet, P., Sidore, C., Snyder, M., Zhan, X., Zöllner, S., Awadalla, P., Casals, F., Idaghdour, Y., Keebler, J., Stone, E.A., Zilversmit, M., Jorde, L., Xing, J., Aksay, G., Hajirasouliha, I., Hormozdiari, F., Kidd, J.M., Sahinalp, S.C., Chen, K., Chinwalla, A., Ding, L., Koboldt, D.C., McLellan, M.D., Wallis, J.W., Wendl, M.C., Zhang, Q., Albers, C.A., Ayub, Q., Balasubramaniam, S., Barrett, J.C., Chen, Y., Conrad, D.F., Danecek, P., Dermitzakis, E.T., Hu, M., Huang, N., Hurles, M.E., Jin, H., Jostins, L., Keane, T.M., Le, S.Q., Lindsay, S., Long, Q., MacArthur, D.G., Montgomery, S.B., Parts, L., Tyler-Smith, C., Walter, K., Zhang, Y., Gerstein, M.B., Abyzov, A., Balasubramaniam, S., Bjornson, R., Du, J., Grubert, F., Habegger, L., Haraksingh, R., Jee, J., Khurana, E., Lam, H.Y., Leng, J., Mu, X.J., Urban, A.E., Zhang, Z., Coafra, C., Dinh, H., Kovar, C., Lee, S., Nazareth, L., Wilkinson, J., Scott, C., Gharani, N., Kaye, J.S., Kent, A., Li, T., McGuire, A.L., Ossorio, P.N., Rotimi, C.N., Su, Y., Toji, L.H., Brooks, L.D., Felsenfeld, A.L., McEwen, J.E., Abdallah, A., Juenger, C.R., Clemm, N.C., Duncanson, A., Green, E.D., Guyer, M.S., Peterson, J.L., 2010. Diversity of human copy number variation and multicopy genes. *Science* (80- ). 330, 641–646.

Suzuki, M.M., Bird, A., 2008. DNA methylation landscapes: Provocative insights from epigenomics. *Nat. Rev. Genet.*

Tahiliani, M., Koh, K.P., Shen, Y., Pastor, W.A., Bandukwala, H., Brudno, Y., Agarwal, S., Iyer, L.M., Liu, D.R., Aravind, L., Rao, A., 2009. Conversion of 5-methylcytosine to 5-hydroxymethylcytosine in mammalian DNA by MLL partner TET1. *Science* (80- ). 324, 930–935.

- Tang, L., Nogales, E., Ciferri, C., 2010. Structure and function of SWI/SNF chromatin remodeling complexes and mechanistic implications for transcription. *Prog. Biophys. Mol. Biol.* 102, 122–8.
- Teschendorff, A.E., Marabita, F., Lechner, M., Bartlett, T., Tegner, J., Gomez-Cabrero, D., 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.
- Teytelman, L., Özyaydin, B., Zill, O., Lefrançois, P., Snyder, M., Rine, J., Eisen, M.B., 2009. Impact of chromatin structures on DNA processing for genomic analyses. *PLoS One* 4.
- Thiel, W.H., 2016. Galaxy Workflows for Web-based Bioinformatics Analysis of Aptamer High-throughput Sequencing Data. *Mol. Ther. Nucleic Acids* 5, e345.
- Tian, Y., Morris, T.J., Webster, A.P., Yang, Z., Beck, S., Feber, A., Teschendorff, A.E., 2017. ChAMP: Updated methylation analysis pipeline for Illumina BeadChips. *Bioinformatics* 33, 3982–3984.
- Tien, A.L., Senbanerjee, S., Kulkarni, A., Mudbhary, R., Goudreau, B., Ganesan, S., Sadler, K.C., Ukomadu, C., 2011. UHRF1 depletion causes a G<sub>2</sub>/M arrest, activation of DNA damage response and apoptosis. *Biochem. J.* 435, 175–185.
- Tilgner, H., Nikolaou, C., Althammer, S., Sammeth, M., Beato, M., Valcárcel, J., Guigó, R., 2009. Nucleosome positioning as a determinant of exon recognition. *Nat. Struct. Mol. Biol.* 16, 996–1001.
- Titus, A.J., Gallimore, R.M., Salas, L.A., Christensen, B.C., 2017. Cell-type deconvolution from DNA methylation: A review of recent applications. *Hum. Mol. Genet.*
- Tobi, E.W., Lumey, L.H., Talens, R.P., Kremer, D., Putter, H., Stein, A.D., Slagboom, P.E., Heijmans, B.T., 2009. DNA methylation differences after exposure to prenatal famine are common and timing- and sex-specific. *Hum. Mol. Genet.* 18, 4046–4053.
- Tobi, E.W., Sliker, R.C., Stein, A.D., Suchiman, H.E.D., Slagboom, P.E., van Zwet, E.W., Heijmans, B.T., Lumey, L., 2015. Early gestation as the critical time-window for changes in the prenatal environment to affect the adult human blood methylome. *Int. J. Epidemiol.* 44, 1211–1223.
- Todkar, A., Granholm, L., Aljumah, M., Nilsson, K.W., Comasco, E., Nylander, I., 2016. HPA Axis Gene Expression and DNA Methylation Profiles in Rats Exposed to Early Life Stress, Adult Voluntary Ethanol Drinking and Single Housing. *Front. Mol. Neurosci.* 8, 90.
- Tom, J.A., Reeder, J., Forrest, W.F., Graham, R.R., Hunkapiller, J., Behrens, T.W., Bhangale, T.R., 2017. Identifying and mitigating batch effects in whole genome sequencing data. *BMC Bioinformatics* 18.
- Triche, T.J., Weisenberger, D.J., Van Den Berg, D., Laird, P.W., Siegmund, K.D., 2013. Low-level processing of Illumina Infinium DNA Methylation BeadArrays. *Nucleic Acids Res.* 41.
- Trizzino, M., Kapusta, A., Brown, C.D., 2018. Transposable elements generate regulatory novelty in a tissue-specific fashion. *BMC Genomics* 19.
- Tutton, S., Lieberman, P.M., 2017. A role for p53 in telomere protection. *Mol. Cell. Oncol.*
- van Gool, J.D., Hirche, H., Lax, H., De Schaepdrijver, L., 2018. Folic acid and primary prevention of neural tube defects: A review. *Reprod. Toxicol.*
- Varley, K.E., Gertz, J., Bowling, K.M., Parker, S.L., Reddy, T.E., Pauli-Behn, F., Cross, M.K., Williams, B.A., Stamatoyannopoulos, J.A., Crawford, G.E., Absher, D.M., Wold, B.J., Myers, R.M., 2013. Dynamic DNA methylation across diverse human cell lines and tissues. *Genome Res.* 23, 555–567.

- Vinson, C., Chatterjee, R., 2012. CG methylation. *Epigenomics*.
- Viré, E., Brenner, C., Deplus, R., Blanchon, L., Fraga, M., Didelot, C., Morey, L., Van Eynde, A., Bernard, D., Vanderwinden, J.M., Bollen, M., Esteller, M., Di Croce, L., De Launoit, Y., Fuks, F., 2006. The Polycomb group protein EZH2 directly controls DNA methylation. *Nature* 439, 871–874.
- Volkin, E., Carter, C.E., 1951. The Preparation and Properties of Mammalian Ribonucleic Acids. *J. Am. Chem. Soc.* 73, 1516–1519.
- Walsh, C.P., Bestor, T.H., 1999. Cytosine methylation and mammalian development. *Genes Dev.* 13, 26–34.
- Walsh, C.P., Chaillet, J.R., Bestor, T.H., 1998. Transcription of IAP endogenous retroviruses is constrained by cytosine methylation. *Nat. Genet.* 20, 116–117.
- Walsh, C.P., Xu, G.L., 2006. Cytosine methylation and DNA repair. In: *Current Topics in Microbiology and Immunology*. Springer Verlag, pp. 283–315.
- Wang, C., Shen, J., Yang, Z., Chen, P., Zhao, B., Hu, W., Lan, W., Tong, X., Wu, H., Li, G., Cao, C., 2011. Structural basis for site-specific reading of unmodified R2 of histone H3 tail by UHRF1 PHD finger. *Cell Res.*
- Wang, Y., Ghaffari, N., Johnson, C.D., Braga-Neto, U.M., Wang, H., Chen, R., Zhou, H., 2011. Evaluation of the coverage and depth of transcriptome by RNA-Seq in chickens. *BMC Bioinformatics* 12 Suppl 10, S5.
- Wang, Y., Yuan, Q., Xie, L., 2018. Histone Modifications in Aging: The Underlying Mechanisms and Implications. *Curr. Stem Cell Res. Ther.* 13, 125–135.
- Wang, Y., Zhao, N., Qiu, J., He, X., Zhou, M., Cui, H., Lv, L., Lin, X., Zhang, C., Zhang, H., Xu, R., Zhu, D., Dang, Y., Han, X., Zhang, H., Bai, H., Chen, Y., Tang, Z., Lin, R., Yao, T., Su, J., Xu, X., Liu, X., Wang, W., Ma, B., Liu, S., Qiu, W., Huang, H., Liang, J., Wang, S., Ehrenkranz, R.A., Kim, C., Liu, Q., Zhang, Y., 2015. Folic acid supplementation and dietary folate intake, and risk of preeclampsia. *Eur. J. Clin. Nutr.* 69, 1145–1150.
- Wang, Z., Zang, C., Rosenfeld, J.A., Schones, D.E., Barski, A., Cuddapah, S., Cui, K., Roh, T.Y., Peng, W., Zhang, M.Q., Zhao, K., 2008. Combinatorial patterns of histone acetylations and methylations in the human genome. *Nat. Genet.* 40, 897–903.
- Waterland, R.A., Kellermayer, R., Laritsky, E., Rayco-Solon, P., Harris, R.A., Travisano, M., Zhang, W., Torskaya, M.S., Zhang, J., Shen, L., Manary, M.J., Prentice, A.M., 2010. Season of conception in rural gambia affects DNA methylation at putative human metastable epialleles. *PLoS Genet.* 6, 1–10.
- Weber, M., Hellmann, I., Stadler, M.B., Ramos, L., Pääbo, S., Rebhan, M., Schübeler, D., 2007. Distribution, silencing potential and evolutionary impact of promoter DNA methylation in the human genome. *Nat. Genet.* 39, 457–66.
- Weigel, C., Chaisaingmongkol, J., Assenov, Y., Kuhmann, C., Winkler, V., Santi, I., Bogatyrova, O., Kaucher, S., Bermejo, J.L., Leung, S.Y., Chan, T.L., Lasitschka, F., Bohrer, M.H., Marx, A., Haußen, R.H., Von Herold-Mende, C., Dyckhoff, G., Boukamp, P., Delank, K.W., Hörmann, K., Lippert, B.M., Baier, G., Dietz, A., Oakes, C.C., Plass, C., Becher, H., Schmezer, P., Ramroth, H., Popanda, O., 2019. DNA methylation at an enhancer of the three prime repair exonuclease 2 gene (TREX2) is linked to gene expression and survival in laryngeal cancer. *Clin. Epigenetics* 11.
- Wilhelm-Benartzi, C.S., Koestler, D.C., Karagas, M.R., Flanagan, J.M., Christensen, B.C., Kelsey, K.T.,

- Marsit, C.J., Houseman, E.A., Brown, R., 2013. Review of processing and analysis methods for DNA methylation array data. *Br. J. Cancer*.
- Williams, C.A., Gray, B.A., Hendrickson, J.E., Stone, J.W., Cantu, E.S., 1989. Incidence of 15q deletions in the angelman syndrome: A survey of twelve affected persons. *Am. J. Med. Genet.* 32, 339–345.
- Winston Chang, Joe Cheng, JJ Allaire, Y.X. and J.M., 2019. shiny: Web Application Framework for R.
- Wolff, G.L., Kodell, R.L., Moore, S.R., Cooney, C.A., 1998. Maternal epigenetics and methyl supplements affect agouti gene expression in Avy/a mice. *FASEB J.* 12, 949–57.
- Wolff, G.L., Roberts, D.W., Morrissey, R.L., Greenman, D.L., Allen, R.R., Campbell, W.L., Bergman, H., Nesnow, S., Frith, C.H., 1987. Tumorigenic responses to lindane in mice: potentiation by a dominant mutation. *Carcinogenesis* 8, 1889–97.
- Woodcock, D.M., Lawler, C.B., Linsenmeyer, M.E., Doherty, J.P., Warren, W.D., 1997. Asymmetric methylation in the hypermethylated CpG promoter region of the human L1 retrotransposon. *J. Biol. Chem.* 272, 7810–7816.
- Wu, H., Coskun, V., Tao, J., Xie, W., Ge, W., Yoshikawa, K., Li, E., Zhang, Y., Sun, Y.E., 2010. Dnmt3a-dependent nonpromoter DNA methylation facilitates transcription of neurogenic genes. *Science* (80-. ). 329, 444–447.
- Wu, M.C., Joubert, B.R., Kuan, P.F., Håberg, S.E., Nystad, W., Peddada, S.D., London, S.J., 2014. A systematic assessment of normalization approaches for the Infinium 450K methylation platform. *Epigenetics* 9.
- WYATT, G.R., COHEN, S.S., 1952. A new pyrimidine base from bacteriophage nucleic acids. *Nature* 170, 1072–3.
- Xie, W., Barr, C.L., Kim, A., Yue, F., Lee, A.Y., Eubanks, J., Dempster, E.L., Ren, B., 2012. Base-Resolution Analyses of Sequence and Parent-of-Origin Dependent DNA Methylation in the Mouse Genome. *Cell* 148, 816–831.
- Yamashita, M., Inoue, K., Saeki, N., Ideta-Otsuka, M., Yanagihara, Y., Sawada, Y., Sakakibara, I., Lee, J., Ichikawa, K., Kamei, Y., Iimura, T., Igarashi, K., Takada, Y., Imai, Y., 2018. Uhrf1 is indispensable for normal limb growth by regulating chondrocyte differentiation through specific gene expression. *Dev.* 145.
- Yang, X., Han, H., DeCarvalho, D.D., Lay, F.D., Jones, P.A., Liang, G., 2014. Gene body methylation can alter gene expression and is a therapeutic target in cancer. *Cancer Cell* 26, 577–590.
- Yen, T.T., Gill, A.M., Frigeri, L.G., Barsh, G.S., Wolff, G.L., 1994. Obesity, diabetes, and neoplasia in yellow A(vy)/- mice: ectopic expression of the agouti gene. *FASEB J.* 8, 479–88.
- Yiu, T.T., Li, W., 2015. Pediatric cancer epigenome and the influence of folate. *Epigenomics*.
- Yoder, J.A., Walsh, C.P., Bestor, T.H., 1997. Cytosine methylation and the ecology of intragenomic parasites. *Trends Genet.*
- Yoon, B., Herman, H., Hu, B., Park, Y.J., Lindroth, A., Bell, A., West, A.G., Chang, Y., Stablewski, A., Piel, J.C., Loukinov, D.I., Lobanenko, V. V., Soloway, P.D., 2005. Rasgrf1 Imprinting Is Regulated by a CTCF-Dependent Methylation-Sensitive Enhancer Blocker. *Mol. Cell. Biol.* 25, 11184–11190.
- Zarrei, M., MacDonald, J.R., Merico, D., Scherer, S.W., 2015. A copy number variation map of the human genome. *Nat. Rev. Genet.*

- Zhang, Y.W., Wang, Z., Xie, W., Cai, Y., Xia, L., Easwaran, H., Luo, J., Yen, R.-W.C., Li, Y., Baylin, S.B., 2017. Acetylation Enhances TET2 Function in Protecting against Abnormal DNA Methylation during Oxidative Stress. *Mol. Cell* 65, 323–335.
- Zhou, L., Mitra, R., Atkinson, P.W., Hickman, A.B., Dyda, F., Craig, N.L., 2004. Transposition of hAT elements links transposable elements and V(D)J recombination. *Nature* 432, 995–1001.

## 1.11 Thesis Aims

The overall aim of this thesis was to identify the effects of alterations to DNA methylation via cell line and human based interventions, in addition to the development of tools and pipelines for the processing and analysis of large amounts of data.

To achieve this aim, the objectives of this thesis were as follows:

- 1) Use the R platform for Statistical Computing and Galaxy Bioinformatics Interface to conduct analysis and quality control of microarray outputs
- 2) Use the R platform for Statistical Computing and Galaxy Bioinformatics Interface to develop more efficient processes for gene target analysis
- 3) To examine potential mechanisms in an attempt to explain the differences in methylation observed within array results e.g. to align array data with ENCODE chromatin configuration or tracks available within UCSC
- 4) To analyze microarray outputs within human studies as well as in cell lines and to adjust analysis protocols due to greater variability within human based studies
- 5) To improve the downstream analysis of bioinformatic data for those not experienced in the processing of high-dimensional data

## 2.0 PAPER-I

### **Depletion of DNMT1 in differentiated human cells highlights key classes of sensitive gene and an interplay with polycomb repression**

Karla M. O'Neill, Rachelle E. Irwin, Sarah-Jayne Mackin, Sara-Jayne Thursby, Avinash Thakur, Ciske Bertens, Laura Masala, Jayne E.P. Loughery, Darragh G. McArt, Colum P. Walsh

The main aims of this paper were to:

- Develop a non-cancerous differentiated cell line with hypomorphic levels of DNMT1
- To investigate the genome-wide effects of depletion of DNMT1
- To investigate the transcriptional response of DNMT1 depletion and its correlation with DNA methylation

### **CONTRIBUTION**

To this paper, I developed an initial simple Galaxy workflow into the CandiMeth prototype to allow easier quantification of candidate features and applied it to derive box-and-whisker and other quantitative outputs. I then further developed this to allow us to analyse overlap of hypomethylated probes with ENCODE chromatin state segmentation data to discover the correlation between hypomethylated probes and polycomb-repressed/ heterochromatin/low signal marks. I also conducted a similar overlap with ENCODE chromatin state segmentation data with hypermethylated probes such as that found at the UGT1A cluster and found a correlation between hypermethylated probes and weak/poised promoters. I contributed a number of the final illustrations and commented on the MS.



RESEARCH

Open Access



# Depletion of DNMT1 in differentiated human cells highlights key classes of sensitive genes and an interplay with polycomb repression

Karla M. O'Neill<sup>1,5†</sup>, Rachele E. Irwin<sup>1†</sup>, Sarah-Jayne Mackin<sup>1</sup>, Sara-Jayne Thursby<sup>1</sup>, Avinash Thakur<sup>1,6</sup>, Ciske Bertens<sup>1,2</sup>, Laura Masala<sup>1,3</sup>, Jayne E. P. Loughery<sup>1</sup>, Darragh G. McArt<sup>4</sup> and Colum P. Walsh<sup>1\*</sup>

## Abstract

**Background:** DNA methylation plays a vital role in the cell, but loss-of-function mutations of the maintenance methyltransferase *DNMT1* in normal human cells are lethal, precluding target identification, and existing hypomorphic lines are tumour cells. We generated instead a hypomorphic series in normal hTERT-immortalised fibroblasts using stably integrated short hairpin RNA.

**Results:** Approximately two-thirds of sites showed demethylation as expected, with one-third showing hypermethylation, and targets were shared between the three independently derived lines. Enrichment analysis indicated significant losses at promoters and gene bodies with four gene classes most affected: (1) protocadherins, which are key to neural cell identity; (2) genes involved in fat homeostasis/body mass determination; (3) olfactory receptors and (4) cancer/testis antigen (CTA) genes. Overall effects on transcription were relatively small in these fibroblasts, but CTA genes showed robust derepression. Comparison with siRNA-treated cells indicated that shRNA lines show substantial remethylation over time. Regions showing persistent hypomethylation in the shRNA lines were associated with polycomb repression and were derepressed on addition of an EZH2 inhibitor. Persistent hypermethylation in shRNA lines was, in contrast, associated with poised promoters.

**Conclusions:** We have assessed for the first time the effects of chronic depletion of DNMT1 in an untransformed, differentiated human cell type. Our results suggest polycomb marking blocks remethylation and indicate the sensitivity of key neural, adipose and cancer-associated genes to loss of maintenance methylation activity.

**Keywords:** DNMT1, EZH2, Protocadherin, Body mass, Cancer/testis antigen

## Background

DNA methylation is an important mechanism for epigenetic regulation of genes in both mouse and human [1]. It occurs mainly at the CpG dinucleotide, and methylation at this symmetrical site is efficiently maintained during replication by the action of the DNA methyltransferase

1 (DNMT1) enzyme [2]. Methylation is known to play an important role in regulating imprinted loci [3], genes on the inactive X chromosome [4] and germline-specific genes [5] in mouse.

Where methylation occurs at the promoter of a gene, it is strongly associated with the silencing of transcription, particularly if there is a high density of CpGs, a so-called CpG island (CGI). However, studies have shown that most CGI are intrinsically protected from methylation [6, 7] and only a small number shows dynamic changes during development, mostly in the three classes mentioned above [5, 8], though there may be others which have not

\*Correspondence: cp.walsh@ulster.ac.uk

†Karla M. O'Neill and Rachele E. Irwin contributed equally to this work

<sup>1</sup>Genomic Medicine Research Group, Centre for Molecular Biosciences, School of Biomedical Sciences, Ulster University, Cromore Road, Coleraine BT52 1SA, UK

Full list of author information is available at the end of the article

yet been clearly defined. As you move outward from an island, the shores and shelves show higher levels of methylation and greater dynamic response [9], though here the link to changes in gene activity is less clear [10]. Methylation is also associated with larger regions of inert chromatin, such as the inactive X, pericentromeric repeats and regions rich in transposable elements [1], generally consistent with a repressive role. Recent genome-wide surveys have also indicated that high levels of methylation are found in the bodies of active genes, where they may facilitate transcription [11, 12]. In keeping with this, we and others recently showed that artificially decreasing intragenic methylation levels reduced steady-state transcript levels, consistent with a positive role for methylation in the gene body [11–13].

Another major system for epigenetic repression is via histone modification, particularly by the polycomb group of proteins, with EZH2 being one of the main enzymes involved [14]. A number of studies suggest an interplay between polycomb- and DNMT-mediated repression, with a generally negative correlation between DNA methylation and the H3K27me3 mark deposited by EZH2 [15, 16]. Supporting this, a loss of DNA methylation caused a reshaping of the histone landscape and derepression of some polycomb targets in mouse ES cells [17], suggesting that DNA methylation helps to determine where polycomb marks are deposited.

While DNMT1 is the main maintenance methyltransferase, there also appears to be an important role for the de novo enzymes DNMT3A and DNMT3B in complementing that activity at some loci [18, 19]. In order to clarify which genes are most sensitive to DNMT1 loss in human, a number of studies have been carried out using mutations within the gene to assess the effects of loss of methylation [19–22]. While this has been a fruitful approach in mouse embryonic stem (ES) cells, where null mutants are tolerated, differentiation of the mouse cells leads to cell death [20, 22, 23], whereas DNMT1 disruption in human ES cells is not tolerated even in undifferentiated cells [24]. Genetic ablation in adult differentiated cells also leads to cell death within a few cell cycles, before passive demethylation of the genome can occur [23, 25]. One of the best-studied systems in humans consists of HCT116 colon cancer cells carrying a hypomorphic allele in the DNMT1 gene together with a DNMT3B knockout (HCT116 DKO cells) [26–28]. Blattler et al. [29] found that there was widespread and relatively uniform demethylation across the genome in the DKO cells, with small effects at CGI (most of which are normally unmethylated anyway) and relatively few genes showing derepression. There was no enrichment by gene ontology (GO) analysis, but some effect at enhancers: however, this is complicated by the presence of the DNMT3B

knockout alleles. Acute depletion of DNMT1 using an siRNA-mediated approach in embryonal carcinoma cells also found regions of low CpG density (open sea, shelf) to be the most affected by loss of methylation [70]. Among the small number of dysregulated genes, there was some enrichment for cell morphogenesis and phosphorylation pathways.

Neither of these cancer cell lines, however, are a good model for the normal differentiated cell as they are transformed, aneuploid, hypermethylated, and contain a number of different mutations in key regulatory genes. Additionally, acute depletion of DNMT1 results in cell cycle delay, triggering of the DNA damage response and increased rates of cell death [24, 25, 30], making it difficult to separate acute and chronic effects.

To circumvent some of the difficulties outlined above, we generated a series of isogenic human cell lines derived from the hTERT-immortalised normal fibroblast line hTERT1604 as previously described [30]. These are normosomic and non-transformed, and by using a stably incorporated plasmid with an shRNA targeting *DNMT1* we were able to isolate a number of clonally derived lines to allow identification of any cell line-specific effects. While these showed initially the range of shared features indicative of a global response to the loss of this critical regulator, including cell cycle delay, demethylation of imprinted genes and others, they could be cultured for longer under selection [30], allowing identification of loci with particular sensitivity for decreased maintenance methyltransferase activity. Here we set out to completely characterise the methylation changes seen in the cell lines using the Illumina Infinium HumanMethylation450 BeadChip (450k) array platform [31] and subsequent analysis using the RnBeads pipeline [32]. These approaches were chosen due to their high reproducibility and low inter-operator variability, ensuring the reliable and sensitive detection of alterations in methylation. A sample of the observations was then further verified using locus-specific assays. In addition and for the same reasons, we used the HT-12 Expression v4 BeadChip array, to assay changes in transcription in our cell lines.

## Methods

### Cell culture

The parental or wild-type (WT) adherent hTERT1604 lung fibroblast cell line [33] was cultured in 4.5 g/l glucose DMEM (ThermoFisher, Loughborough, UK) supplemented with 10% FBS and 2× NEAA (Gibco/ThermoFisher). Generation of the hTERT1604 cell lines stably depleted of DNMT1 using a pSilencer construct (ThermoFisher) has been previously described [30]. Knockdown (KD) cells were maintained as for WT, but medium was supplemented with 150 µg/ml hygromycin

B (Invitrogen/ThermoFisher, Paisley, UK), which was removed at least 48 h before any experimental procedure. Treatment of cells with siRNA for 24 h was as previously described [34]: for the pulse-chase experiment cells were afterwards allowed to recover in normal media and passaged as required for up to 36 days. The siRNA (Dharmacon ON-TARGETplus SMARTpool) for *DNMT1* and *DNMT3B*, as well as scrambled control, was obtained from Invitrogen/ThermoFisher. HCT116 and double knockout (DKO) cells [27] were cultured in 1 g/l glucose DMEM (Gibco) supplemented with 10% FBS and 1× NEAA (Gibco). DZNeP (Sigma-Aldrich, Dorset, UK) was used at a final concentration of 1 μM.

### DNA extraction and bisulphite conversion

Genomic DNA was harvested from cells in log phase of growth. Samples were incubated overnight at 55 °C in lysis buffer [50 mM Tris pH 8, 0.1 M EDTA (both Sigma-Aldrich), 0.5% SDS, 0.2 mg/ml proteinase K (Roche, West Sussex, UK)], with rotation, and DNA was subsequently isolated using the standard phenol/chloroform/isoamyl alcohol (25:24:1 pH8, Sigma-Aldrich) extraction method. DNA quality was verified using gel electrophoresis and UV absorbance measurements at 260/280 and 260/230 nm using a Nanodrop UV spectrophotometer (Labtech International, Ringmer, UK). Bisulphite conversion of 500 ng of DNA was carried out using the EpiTect bisulphite kit (Qiagen, Crawley, UK) according to the manufacturer's instructions.

### Hybridisation to 450K array and bioinformatic analyses

Three samples from each cell line were used to prepare DNA, with at least one biological repeat in each set. DNA was assessed for purity and integrity as above prior to quantification using the Quant-iT PicoGreen dsDNA assay kit (Thermo Fisher Scientific) as per manufacturer's instructions. In total, 500 ng of high-quality bisulphite-converted (Zymo Research) DNA was checked for purity and fragmentation on a bioanalyser and then loaded on the Infinium Human Methylation 450 BeadChip [31] and imaged using an Illumina iScan (Cambridge Genomic Services). Output files in IDAT format were processed using the RnBeads [32] methylation analysis package (v1.0.0) which carries out all the analysis from import to differential methylation within the R platform (3.2.0). Briefly, quality control used the built-in probes on the array and included filtering out of probes containing SNPs, and checking for hybridisation performance. Normalisation was then carried out using the SWAN method in minfi [35] after background subtraction with methylumi.noob. The exploratory analysis module was used to generate probe density distributions and scatter graphs. The differential methylation analyses was based

on a combined ranking score, which combined absolute effect size, relative effect sizes and p-values from statistical modelling into one score where rank is computed as the most conservative value among mean difference in means, mean in quotients and combined *p* value across sites in the region: the enrichment analysis used the combined rank among the 1000 best-ranking regions and a hypergeometric test to identify GO terms in the AmiGO 2 database [36]. Pairwise comparison of triplicate samples from each cell line against WT hTERT was also made to determine change in beta value and associated combined p-value, adjusted for multiple comparison using false discovery rate (FDR). Some tailored analyses were also carried out using custom scripts in R. Additional GO studies were performed using DAVID (v6.7) [37].

We used the GALAXY platform [38] to map sites showing highly reproducible changes (FDR < 0.05) against the locations of RefSeq genes or ChromHMM regions on the UCSC genome browser [39] for each cell line. GO category genes which showed changes in methylation at multiple sites in more than one KD cell line were scored as true hits (Yes in the FDR column), while GO categories with few or no sites reproducibly altered across replicates (FDR > 0.05) or where methylation changes were small (< 0.1 β), inconsistent in direction, or not found in more than one KD cell line, were scored as false positives. Absolute β levels were used to measure median methylation across genes of interest using custom workflows in GALAXY, with further statistical analyses in Statistical Package for the Social Sciences software (SPSS) version 22.0 (SPSS UK Ltd).

### Locus-specific methylation analysis

Amplification was carried out using the PyroMark PCR kit (Qiagen) with 2 μl bisulphite-converted DNA, 12.5 μl MasterMix, 2.5 μl CoralLoad Concentrate, 1.25 μl each primer (10 μM) and 5.5 μl nuclease-free H<sub>2</sub>O using the following conditions: 15 min at 95 °C followed by 45 cycles of 94 °C for 30 s, 56 °C for 30 s, 72 °C for 30 s and a final elongation step of 72 °C for 10 min. Pyrosequencing was carried out on the PyroMark Q24 System, according to the manufacturer's instructions (Qiagen). Most assays were designed in-house using the PyroMark Assay Design software 2.0 (*LEP*, *MAGEA12*, *OR10J5*, *OR51E2*, *OR2AG1*, *PCDHA2*, *PCDHC4*, *UGT1A1*, *UGT1A4*) prior to synthesis (Metabion, Germany): see Additional file 1: Table S1 for details: *DAZL*, *SYCP3*, *D4Z4* and *NBL2* were as described [34, 40]. In some cases, pre-designed pyrosequencing primers were obtained from Qiagen (*GABRQ* PM00133483, *GHSR* PM00014350, *SNRPN* PM00168252). Clonal analysis was carried out as previously described [30].

### Hybridisation to HT-12 microarray and bioinformatic analyses

Total RNA was extracted using the RNeasy minikit (Qiagen) as per manufacturer's instructions, including a DNase step. RNA integrity was verified via gel electrophoresis, and quality and quantity were verified using a SpectroStar (BMG Labtech, Aylesbury, UK) and a bioanalyser (Agilent Technologies, Cheadle, UK). Two hundred nanograms of total RNA underwent linear amplification using the Illumina TotalPrep RNA Amplification Kit (Life Technologies/ThermoFisher, Paisley, UK) following the manufacturer's instructions. Microarray experiments were performed at Cambridge Genomic Services, University of Cambridge, using the HumanHT-12 v4 Expression BeadChip (Illumina, Chesterford, UK). After scanning the data were loaded in GenomeStudio (Illumina) and then processed in R (version 3.2.2). The data were filtered to remove any non-expressed probes using the detection p-value from Illumina, transformed using the variance stabilization transformation (VST) from lumi and normalised using the quantile method. Comparisons were made using the limma package with results corrected for multiple testing using false discovery rate (FDR) testing.

### RNA and protein analysis

Transcriptional assays at individual loci using RT- and RT-qPCR were carried out essentially as in [34]: primer sequences are listed in Additional file 1: Table S1. Protein was extracted from cells growing in log phase using protein extraction buffer (50 mM Tris-HCl, 150 mM NaCl, 1% Triton-X, 10% glycerol, 5 mM EDTA; all Sigma-Aldrich) and 0.5 µl protease inhibitor mix (Sigma-Aldrich). For Western blotting, 30 µg protein was denatured in the presence of 5 µl 4× LDS sample buffer (Invitrogen) and 2 µl 10× reducing agent (Invitrogen) in a total volume of 20 µl nuclease-free water (Qiagen) via incubation at 70 °C. Proteins were separated by SDS-PAGE and then electroblotted onto a nitrocellulose

membrane (Invitrogen) and blocked in 5% non-fat milk for 1 h at room temperature (RT). Membranes were incubated with anti-DNMT1 (a kind gift from Guoliang Xu) and anti-β-actin (Abcam ab8226) overnight at 4 °C, followed by HRP-conjugated secondary antibody incubation at RT using ECL (Invitrogen).

### Statistical analysis

Statistical analysis was performed by the RnBeads package, or separately in Excel (Microsoft Office Professional Plus 2013), Prism (Graphpad) or SPSS (v22.0). Experiments were carried out in triplicate and included at least one biological replicate. PCR results were analysed using Student's paired *t*-test. Pyrosequencing results were analysed by ANOVA within representative runs and using Student's *t*-test on the average of multiple runs. Error bars on all graphs show standard error of the mean (SEM) or in the case of HT12 array data, 95% confidence interval (CI), unless otherwise stated. Asterisks are used to represent probability scores as follows: \**p* < 0.05; \*\**p* < 0.01; \*\*\**p* < 0.005 or n.s. not significant.

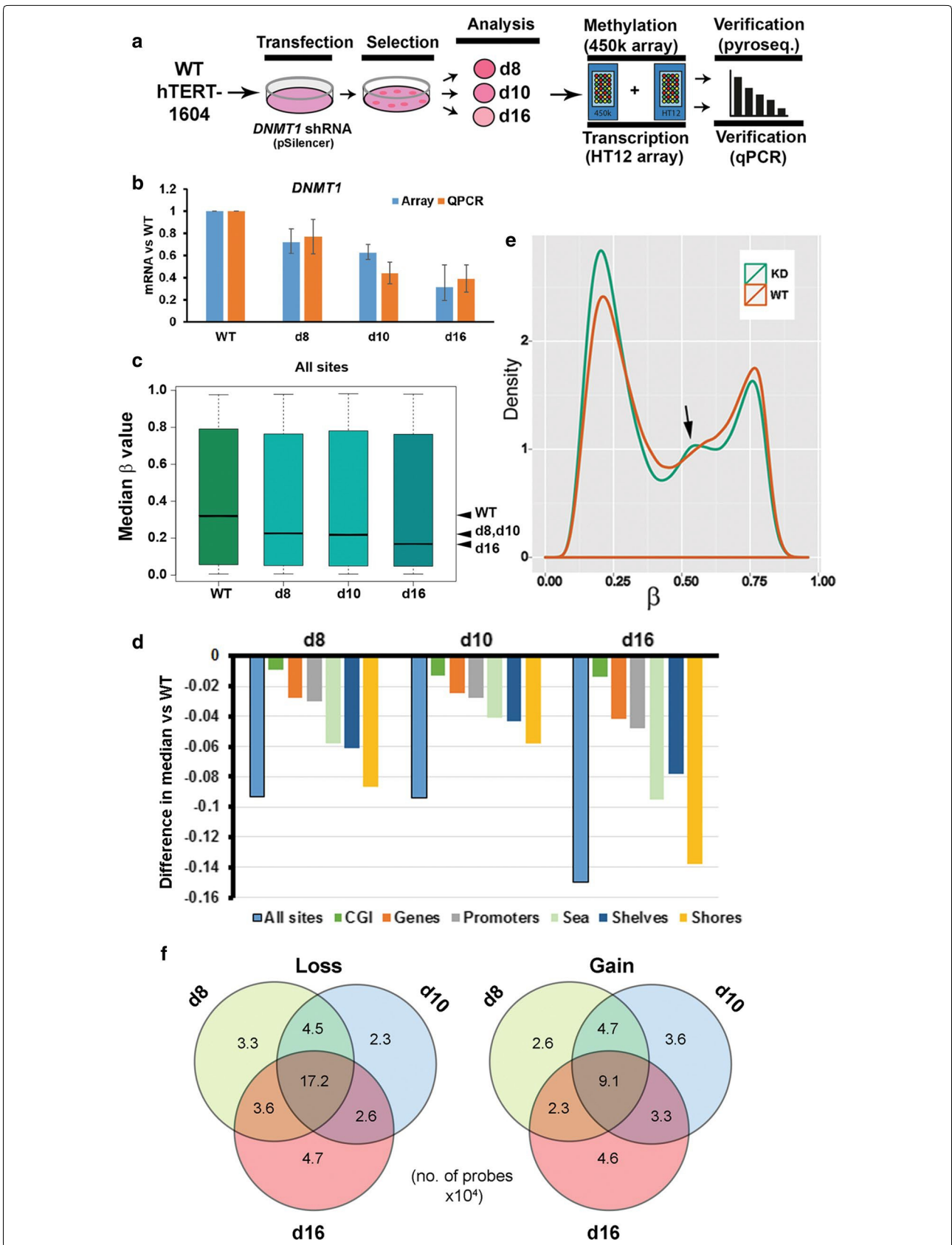
### Results

#### Generation of isogenic hTERT1604 fibroblast cell lines

Isogenic lines carrying an shRNA construct targeting *DNMT1* were generated by transfecting the hTERT-immortalised human lung fibroblast cell line hTERT-1604 with pSilencer plasmid containing an shRNA (Fig. 1a). The generation and initial characterisation of these isogenic cell lines have been previously described [30]. Here we took two sublines typical of the intermediate levels of knockdown (KD) seen (d8 and d10) as well as one line (d16) with relatively low levels of mRNA, with good agreement between reverse transcription quantitative PCR (qPCR) and array results (Fig. 1b; all *p* < 0.05 except d8 array). We also confirmed knockdown at the protein level using Western blotting, with HCT116 cells mutated in *DNMT1* and *DNMT3B* [27] as controls (Additional file 3: Fig. S2A).

(See figure on next page.)

**Fig. 1** Cell line generation and overall changes seen in methylation levels. **a** Experimental approach: WT hTERT1604 fibroblasts were transfected with shRNA-containing plasmid and grown in selective medium; colonies of resistant cells were expanded, and three (d8, d10, d16) showing reduced *DNMT1* levels were then analysed using genome-wide methylation and transcription arrays on the Illumina platform. **b** Levels of *DNMT1* mRNA in cell lines from array and qPCR: error bars represent 95% confidence intervals around median, and standard error of the mean (SEM), respectively. All three knockdown (KD) lines were significantly depleted at *p* < 0.05 for both assays (except d8 array). **c** Overall methylation levels in WT and KD cells as measured by 450K: a  $\beta$  value of 1 equates to 100% methylation. Median values are indicated by the line, and whiskers represent interquartile range. The positions of the medians are also indicated at right (arrowheads). **d** The difference in median  $\beta$  value between each KD cell line and WT is shown first for all sites assayed (see above) and then for each type of genomic element. CGI, CpG island; shore, region adjacent to CGI; shelf, adjacent to shore; sea, all other. See also Additional file 3: Fig. S2B. **e** Probe density distributions; in KD there is a decrease in the number of fully methylated sites ( $\beta$  closer to 1) and an increase in the number of unmethylated sites ( $\beta$  closer to 0), as well as in probes showing intermediate levels of methylation (arrow). **f** Numbers of sites ( $\times 10^4$ ) showing significant changes in methylation (FDR < 0.05) compared to WT: the set of common sites is largest in each case, with close to twice as many sites commonly losing methylation in comparison with those gaining



### Characterisation of overall changes in absolute methylation levels in depleted lines

Using the 450K array [31] and processing in RnBeads [32] to assess methylation levels across the genome (Fig. 1c), there was still a wide range of methylation values (given for the array as a value  $\beta$  ranging from 0 to 1) in KD lines as compared to WT, but the median values were decreased as expected in all three with d8 being comparable to d10, while d16 was lower (arrowheads at right). Principle components analysis and examination of the sites showing greatest differences in methylation between the stable lines confirmed that d8 and d10 were most similar (Additional file 2: Fig. S1). Probes on the array were annotated by location relative to genomic features, and while all regions showed a decrease in methylation, the difference in median values was smallest for CGI, which were unmethylated anyway in parental cells ( $\beta < 0.1$  in WT), while the separation in medians was greatest at shelves and shores, where methylation levels were higher (Additional file 3: Fig. S2B). This can most clearly be seen by plotting the difference in medians (Fig. 1d). Both WT and the KD cell lines showed the typical bimodal probe density distribution pattern reported in most cell types [31] (Fig. 1e). Overall, there was an increase in the numbers of less methylated probes ( $\beta < 0.25$ ) in the KD cell lines and a decrease in the numbers of highly methylated probes ( $\beta > 0.65$ ). For individual regions CGI again showed the smallest change, while gene bodies (genes) appeared most altered (Additional file 3: Fig. S2C).

To determine whether methylation was lost stochastically in each KD cell line given the variation seen (Additional file 2: Fig. S1), or was more targeted, we determined the degree to which affected sites were shared between the three cell lines (Fig. 1f). The largest set of sites losing methylation ( $17.2 \times 10^4$ ) was that shared between all three KD lines, supporting a non-random loss. A spike in numbers of probes showing intermediate levels of methylation ( $\beta \sim 0.50$ ) in KD cell lines in the density profile plot (Fig. 1e, arrow) had indicated that a possible gain in methylation might also be occurring at some sites. Analysis showed that a substantial number ( $9.1 \times 10^4$ ) of sites gaining methylation are shared between all three KD lines, indicating reproducible gains in methylation at particular CpGs.

### Overall pattern of sites showing significant differential methylation on DNMT1 depletion

We compared WT cells to all three KD lines using the RnBeads package in R and combined rank scoring (see methods). This confirmed that d16 has the greatest number of demethylated sites using a false discovery rate (FDR) cut-off of  $p < 0.05$ , but at  $p < 0.001$  all three lines have comparable numbers of hypo- and hypermethylated

sites (Additional file 4: Fig. S3A), with more sites losing than gaining. An analysis of the 1000 best-ranking sites highlights sites common to all three KD lines (Additional file 4: Fig. S3B), confirming that there are large numbers of sites which respond in the same way in each KD, with an excess of probes showing loss over gain.

We then looked to see whether shared probes were enriched in any particular gene region. As we were interested in changes which might cause altered transcription, we focussed on CGI, promoters and gene bodies (hereafter genes) rather than shores, shelves or open sea, where correlations with transcriptional output are harder to assess. Using a hypergeometric test in RnBeads, both promoters and genes, but not CGI, showed significant enrichment in demethylated probes for particular gene ontology (GO) terms. Table 1 indicates the top 3 ontology classes under biological process (BP) and molecular function (MF). For loss of methylation, examining common genes and processes suggested that three classes of genes were common to the enriched GO terms, which we grouped as follows: (1) genes involved in neuroepithelial differentiation; (2) genes involved in fat homeostasis/body mass (FBM); and (3) olfactory receptor genes (groups 1–3 in Table 1), all of which will be dealt with below. The only orphan GO term whose members had multiple high-confidence demethylated sites was GO:0007506 gonadal mesoderm formation, which largely consists of members of the *TSPY* gene family on the Y chromosome. For gain of methylation, the same was true in that a relatively small number of histone modifier genes (group 4), represented under several GO terms, were responsible for many of the hits. In addition, the GO terms for glucuronosyltransferase activity (GO:0015020) and for regulation of megakaryocyte differentiation were also represented (Table 1). These were then curated by looking for sites showing reproducible changes (FDR  $< 0.05$ ) in all KD lines (described more fully in “Methods” section), which indicated strong support [Yes (Y) in confirm column, Table 1] for all GO categories showing loss, but only in two showing gain (GO:0015020 and GO:0004984). We then set about verifying these targets.

### Loss of methylation at the protocadherin gamma gene cluster particularly affects the A and B class variable genes

A main contributor to the enrichment of neuroepithelial genes are the protocadherin genes. Protocadherin  $\alpha$ ,  $\beta$  and  $\gamma$  (*PCDHA*, *PCDHB* and *PCDHG*) genes are located in three linked clusters on chromosome 5 and give rise to neural cell–cell adhesion proteins, with significant loss of methylation across the whole region in all three cell lines (Additional file 4: Fig. S3C). The  $\alpha$  and  $\gamma$  proteins have a variable extra-cellular recognition domain, either A, B

**Table 1 Gene ontology analysis for differentially methylated sites**

Type	GO FID	P	OR	Ex	Obs	Total	GO Term	Grp	confirm
<i>Loss</i>									
<i>Promoter</i>									
BP	0098609	0.0011	3.0454	4.2148	12	189	Cell-cell adhesion	1	Y
	0007156	0.0011	3.4722	3.0998	10	139	Homophilic cell adhesion via plasma membrane	1	Y
	0010982	0.0015	88.2036	0.0669	2	3	Regulation of high-density lipoprotein particle clearance	2	Y
MF	0004888	0.0001	1.9709	24.2681	44	1055	Transmembrane signaling receptor activity	3	Y
	0005509	0.0001	2.2488	14.3768	30	625	Calcium ion binding	1	Y
	0004871	0.0003	1.7441	33.6302	54	1462	Signal transducer activity	3	Y
<i>Gene</i>									
BP	0007506	0	130.3775	0.1339	5	7	<i>Gonadal mesoderm development</i>		Y
	0032375	0.0001	25.9783	0.2295	4	12	Negative regulation of cholesterol transport	2	Y
	0045409	0.0001	77.705	0.0956	3	5	Negative regulation of interleukin-6 biosynthetic process	2	Y
MF	0008083	0.0009	3.5742	3.0015	10	158	Growth factor activity	3	Y
	0004984	0.0014	2.5939	6.136	15	323	Olfactory receptor activity	3	Y
	0038023	0.0014	1.7776	22.9102	38	1206	Signalling receptor activity	3	Y
<i>Gain</i>									
<i>Promoter</i>									
BP	0035574	0	443.1106	0.4729	14	15	Histone H4-K20 demethylation	4	N
	0045653	0	147.6833	0.5359	14	17	Negative regulation of megakaryocyte differentiation		N
	0016577	0	26.4022	1.0404	15	33	Histone demethylation	4	N
MF	0035575	0	452.3692	0.4637	14	15	Histone demethylase activity (H4-K20 specific)	4	N
	0032451	0	21.0879	1.1747	15	38	Demethylase activity	4	N
	0015020	0	10.1109	0.8965	7	29	<i>Glucuronosyltransferase activity</i>		Y
<i>Gene</i>									
BP	0035574	0	280.0725	0.4039	14	16	Histone H4-K20 demethylation	4	N
	0045653	0	140.0181	0.4544	14	18	Negative regulation of megakaryocyte differentiation		N
	0006335	0	31.0869	0.8078	14	32	DNA replication-dependent nucleosome assembly	4	N
MF	0035575	0	287.2955	0.3942	14	16	Histone demethylase activity (H4-K20 specific)	4	N
	0032451	0	24.654	0.9856	15	40	Demethylase activity	4	N
	0004984	0	4.4768	7.9586	31	323	Olfactory receptor activity	3	Y

BP biological process, MF molecular function, GO FID gene ontology family identification code, P probability value, OR odds ratio, Ex expected number of hits, Obs observed number, Total total number of genes in that family, Grp-see below; confirm Y/N, confirmation given by FDR tracks Yes/No

Groups (Grp): 1 = neuroepithelium; 2 = Fat homeostasis/body mass (FBM); 3 = olfactory receptor; 4 = histone modifier

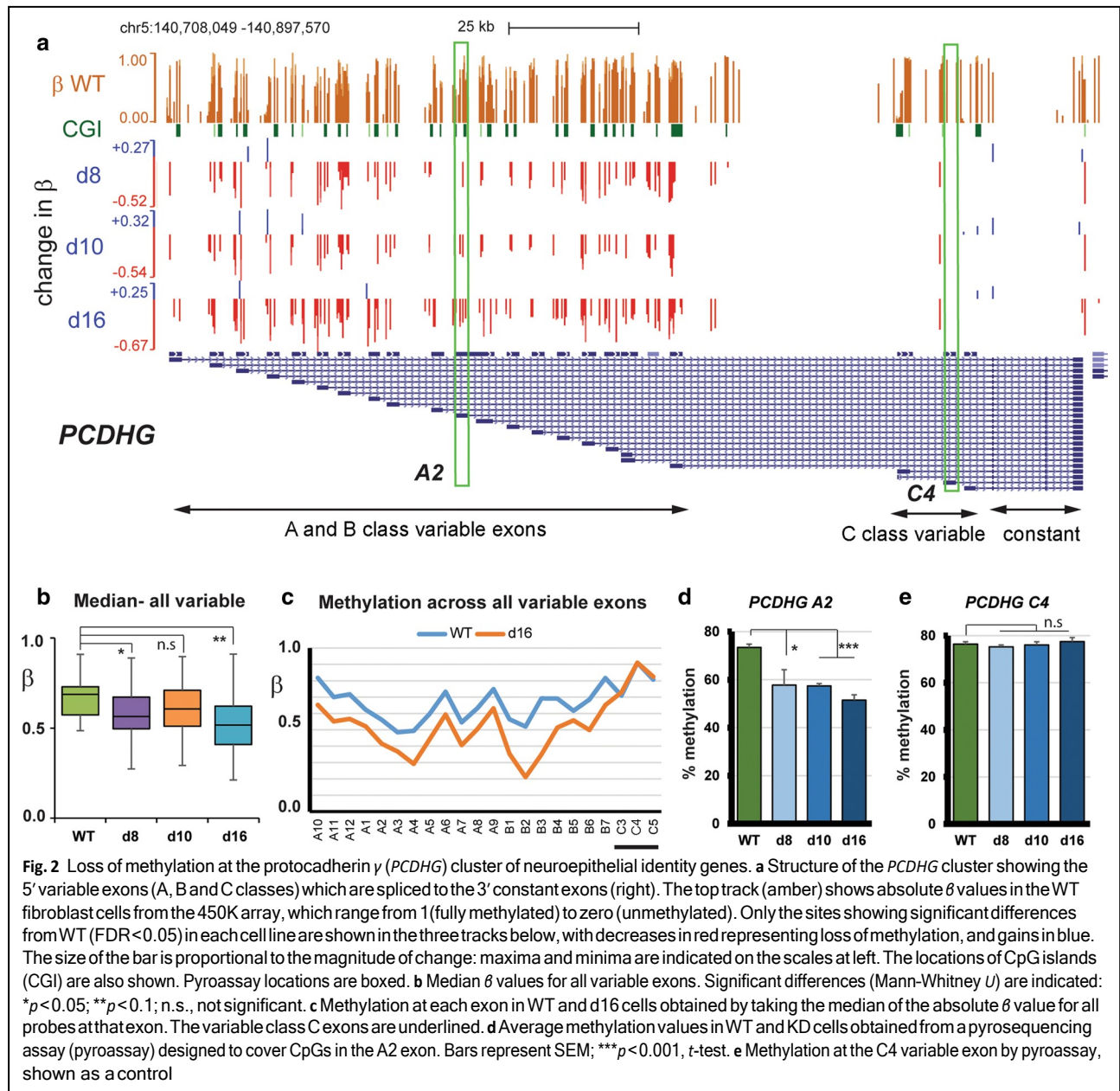
or C-type, attached to a constant transmembrane and intracellular domain. This is achieved at the gene level by alternative 5' exons encoding the variable region being spliced to the constant region exons. Figure 2a shows the tracks containing sites with significant (FDR < 0.05) methylation differences between KD and WT cells for the *PCDHG* cluster. These reveal loss of methylation (in red in Fig. 2a) at most A and B class variable exons in all three KD cell lines, but not at the C class variable or the constant exons. Array probes were present in this region, and examination of the absolute rather than relative methylation (amber, top track in Fig. 2a) confirmed high levels of methylation in WT, where median  $\beta$  values were high for all variable exons (Fig. 2b). Methylation decreased in all three KD lines, with d10 showing the least effect

(Fig. 2b). Methylation was substantially altered at all A and B class variable exons, but not at the C class (Fig. 2c). We could experimentally verify the loss of methylation at A2 (Fig. 2d), and no change at C4 (Fig. 2e), using pyrosequencing assays (pyroassay).

Some demethylation of other neuroepithelial genes in this GO category was also seen from the array, such as *S100P*, *ROBO1* and *PAX6*, with significant ( $p < 0.05$ ) demethylation of *S100P* in two-thirds of KD cell lines confirmed by pyrosequencing (not shown).

#### Loss of methylation at other targets including fat homeostasis/body mass (FBM) genes

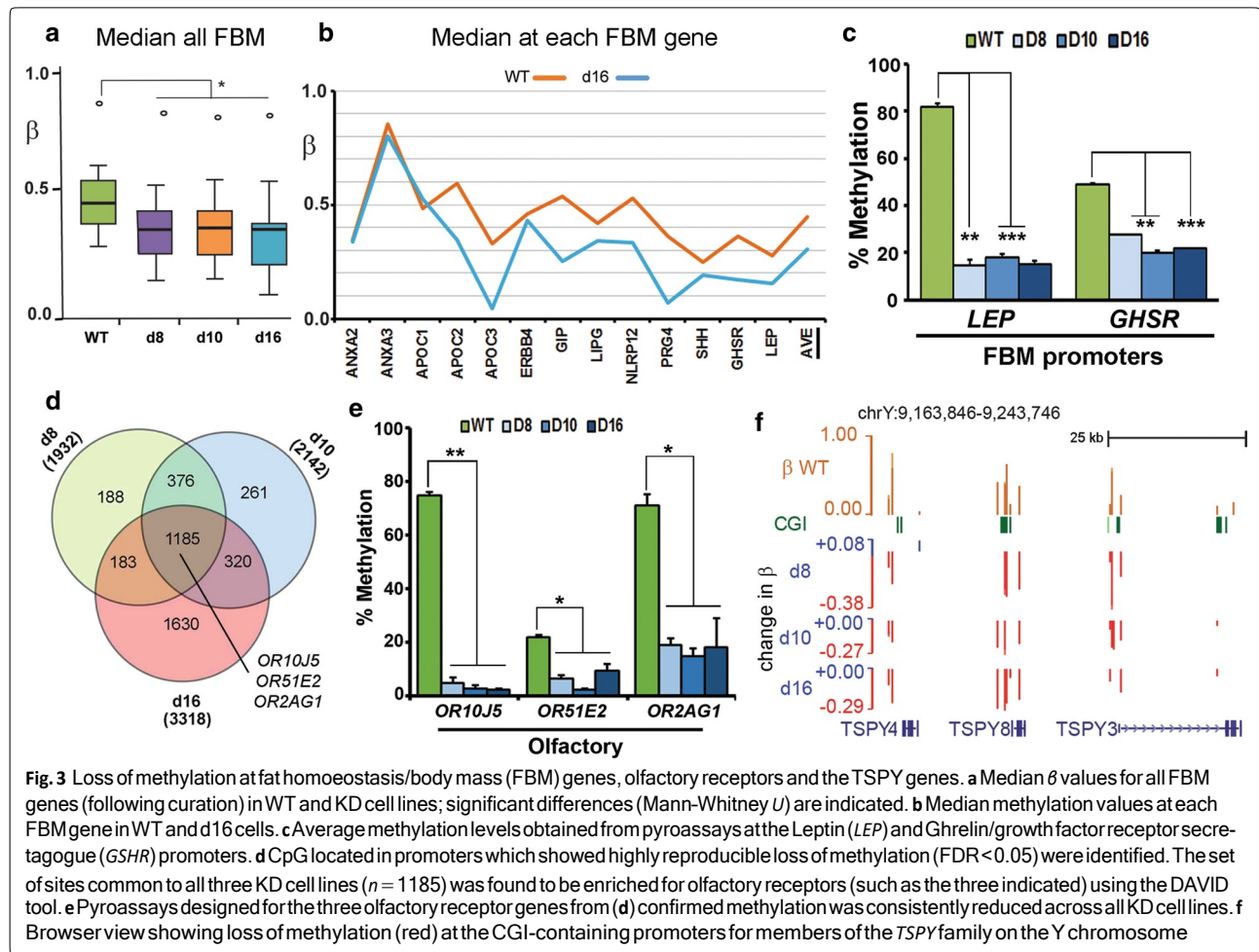
Another class of genes showing enrichment all appear to be involved in some aspect of triglyceride processing,



energy homeostasis and body weight regulation (Table 1), including leptin (*LEP*), ghrelin/growth hormone secretagogue receptor (*GHSR*) and genes encoding the very low density lipoproteins *APOC1*, *APOC2* and *APOC3*. Median levels of methylation in the gene bodies were approximately 45% in WT ( $\beta = 0.45$ ) and showed significant ( $p < 0.05$ ) decreases in the KD lines (Fig. 3a). Most individual genes also showed substantial loss, with the exception of the *ANXA* genes (Fig. 3b). Loss of methylation at the *LEP* and *GHSR* promoters was confirmed using pyroassay (Fig. 3c).

Olfactory receptor (OR) genes appeared in a number of GO categories as having lost methylation, though some gains in the gene body were also indicated (Table 1). ORs encode G protein-coupled receptor proteins and are members of a large gene family, many of which are grouped into major clusters, particularly on chromosome 11 [41]. To buffer against stochastic effects due to the large gene family involved, we carried out a second analysis starting instead with sites in promoters showing reliable methylation loss compared to WT (FDR<0.05) in the triplicates of each KD line



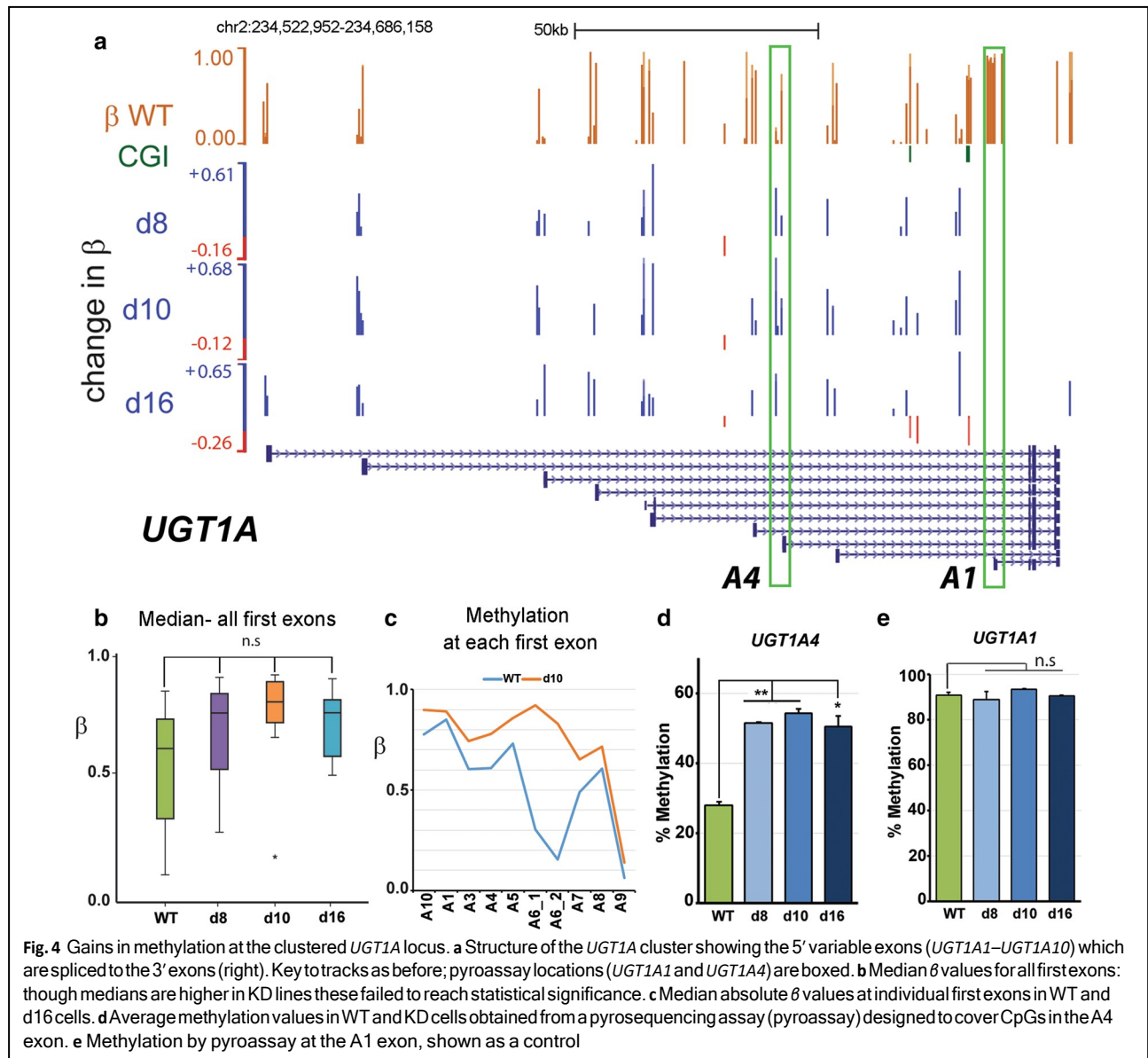


and then overlapping these (Fig. 3d) to see which sites were common to all three KD cell lines (Additional file 5: Table S2). Ontology analysis of these common sites using DAVID independently highlighted signaling receptor genes and more particularly olfactory receptors ( $n = 21$ ). This group of OR genes also showed significant demethylation compared to WT (Kruskal–Wallis,  $p < 0.05$ ) across the genes when median methylation at all available probes was analysed (Additional file 4: Fig. S3D). We chose three of these genes—*OR10J5*, *OR51E2* and *OR2AG1*—located on different chromosomes and could verify loss of methylation in all KD lines (Fig. 3e).

The final GO category of genes (GOFMID:0007506) showing loss of methylation (Table 1) consists largely of the *TSPY* gene family (*TSPY1-4*, 8 and 10) located on the Y chromosome and thought to be implicated in both normal gonadal development and in gonadoblastoma [42]. These also showed clear evidence of demethylation (Fig. 3f).

### Gains in methylation affect the *UGT1A* locus

As indicated above, with respect to gains in methylation only two of the GO classes identified in the genome-wide screen (Table 1) contained multiple sites showing significant gains in methylation (FDR < 0.05, > 0.1 gain in  $\beta$ ). One of these was the olfactory genes, discussed above: the other GO term GO:0015020 was largely comprised of members of the *UGT1A* family. This gene family has a similar structure to the *PCDHG* cluster, where unique alternate 5' exons splice to common 3' exons, but in this case codes for a series of nine UDP-glucuronosyltransferase enzymes (UGTs). Substantial gains in methylation can be seen at the upstream promoters controlling the 5' exons (Fig. 4a), most of which lack CGI. Median methylation levels also showed clear increases overall in the KD lines (Fig. 4b), though these did not reach significance. Most individual exons also showed a sharp increase (Fig. 4c), with A1 being a clear exception in all lines. We confirmed a significant gain in methylation in each cell line at A4 (Fig. 4d) but no alteration at A1 (Fig. 4e). In contrast to the clear gains in

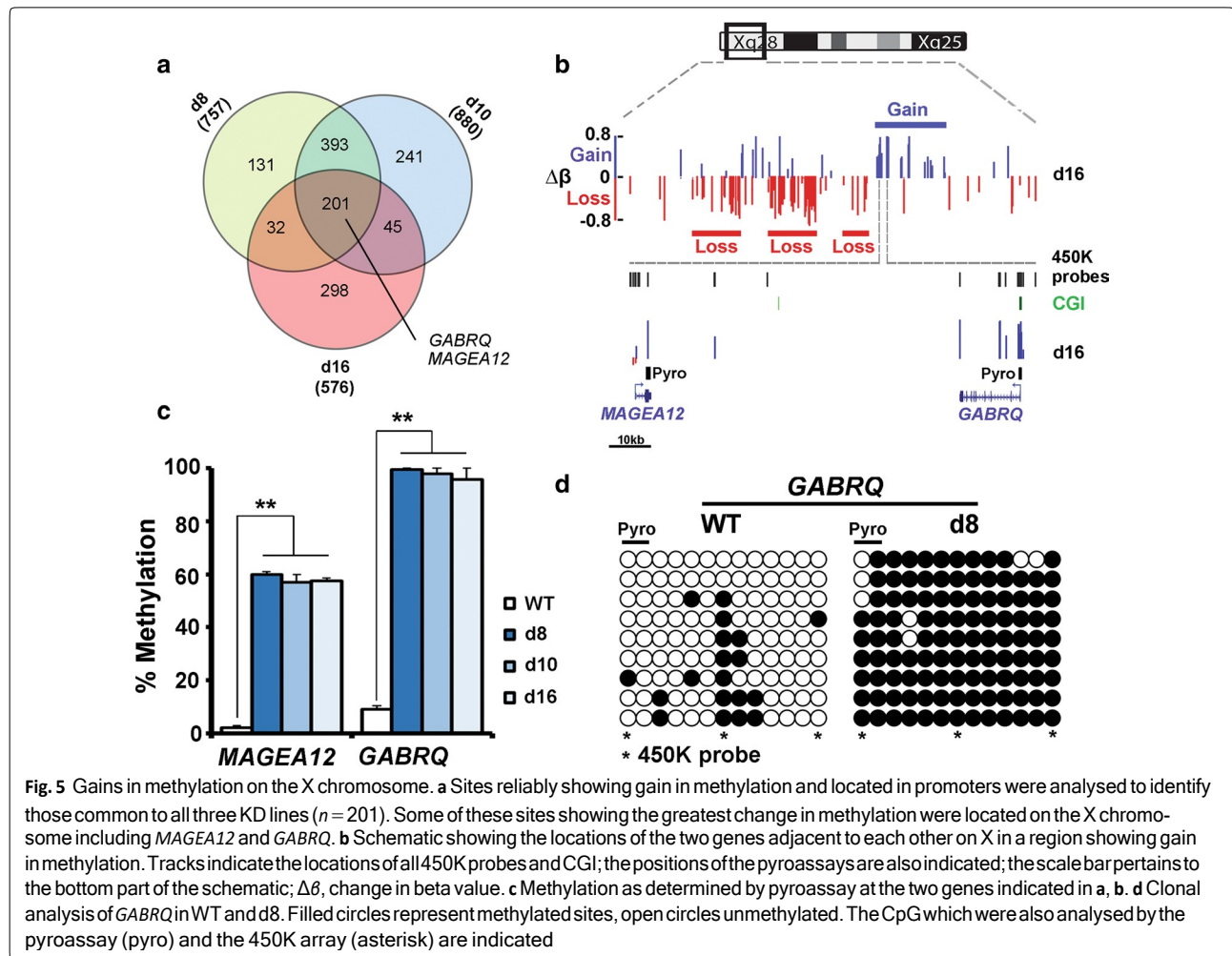


all three lines for *UGT1A*, the histone modifier group also identified as gaining methylation (Table 1, group 4) contained few FDR-supported sites and these often did not overlap between cell lines, with median  $\beta$  levels also not differing significantly (Additional file 4: Fig. S3E).

**A cluster of loci showing gain of methylation on the X chromosome**

Given that there were considerable numbers of probes showing gain in methylation, but few of the GO classes from the RnBeads analysis contained testable targets by our criteria, we tried an alternative analysis as for the OR above. Sites associated with promoters and which

showed reliable (FDR<0.05) gains were identified in each KD line, and then the lists of cognate genes were compared to find those which were common to all three cell lines (Fig. 5a). Examination of the 201 promoters from this analysis (Additional file 5: Table S2) failed to show any significantly enriched terms in DAVID. However, several of the genes showing the greatest gain in methylation were located on the X chromosome, including *GABRQ* and members of the *MAGE* family of cancer/testis antigens such as *MAGEA12*. Mapping of FDR sites to the X chromosome showed that adjacent domains could vary in methylation level by more than 80% in either direction (Fig. 5b). Pyroassays for *GABRQ* and the neighbouring



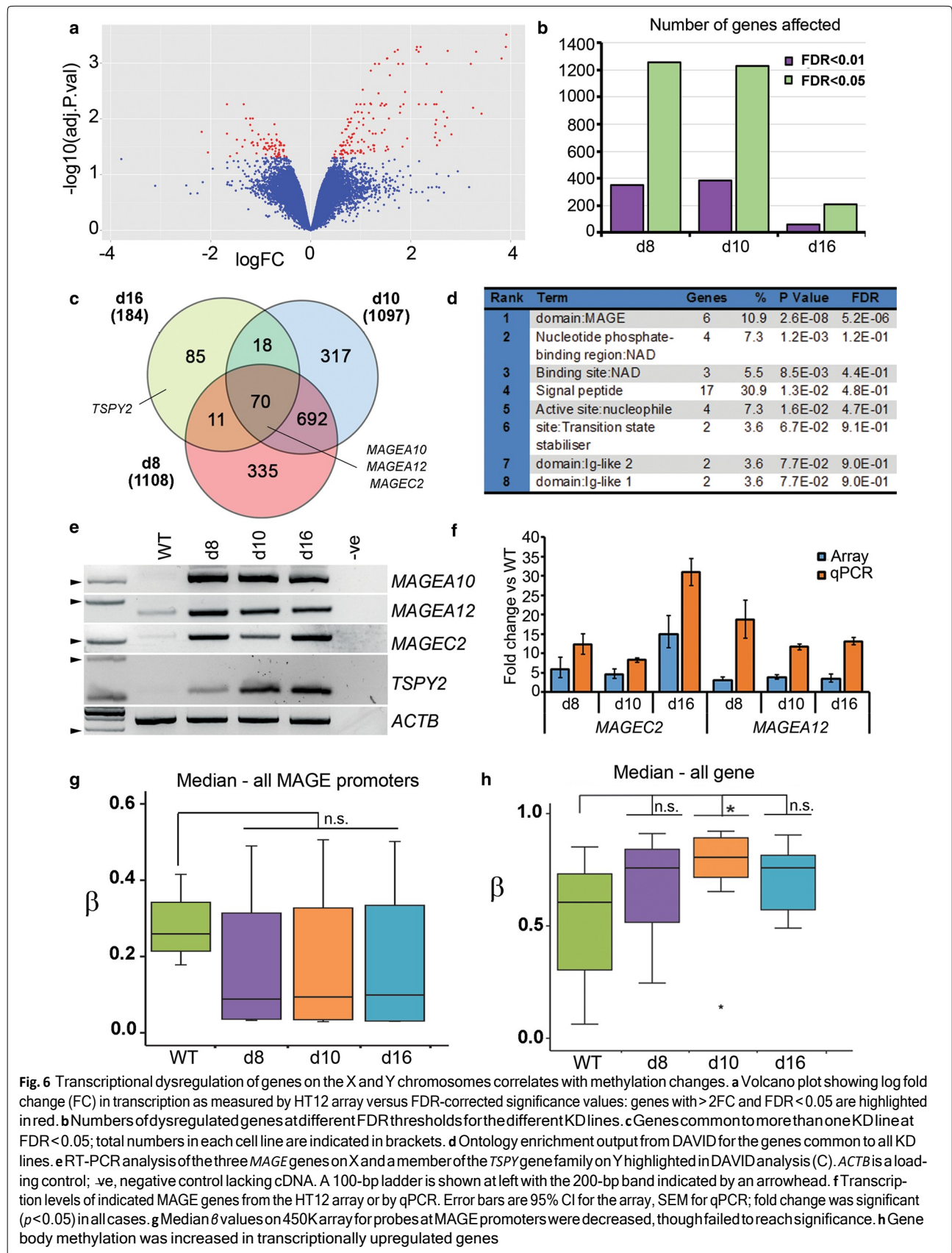
**Fig. 5** Gains in methylation on the X chromosome. **a** Sites reliably showing gain in methylation and located in promoters were analysed to identify those common to all three KD lines ( $n = 201$ ). Some of these sites showing the greatest change in methylation were located on the X chromosome including *MAGEA12* and *GABRQ*. **b** Schematic showing the locations of the two genes adjacent to each other on X in a region showing gain in methylation. Tracks indicate the locations of all 450K probes and CGI; the positions of the pyroassays are also indicated; the scale bar pertains to the bottom part of the schematic;  $\Delta\beta$ , change in beta value. **c** Methylation as determined by pyroassay at the two genes indicated in **a**, **b**. **d** Clonal analysis of *GABRQ* in WT and d8. Filled circles represent methylated sites, open circles unmethylated. The CpG which were also analysed by the pyroassay (pyro) and the 450K array (asterisk) are indicated

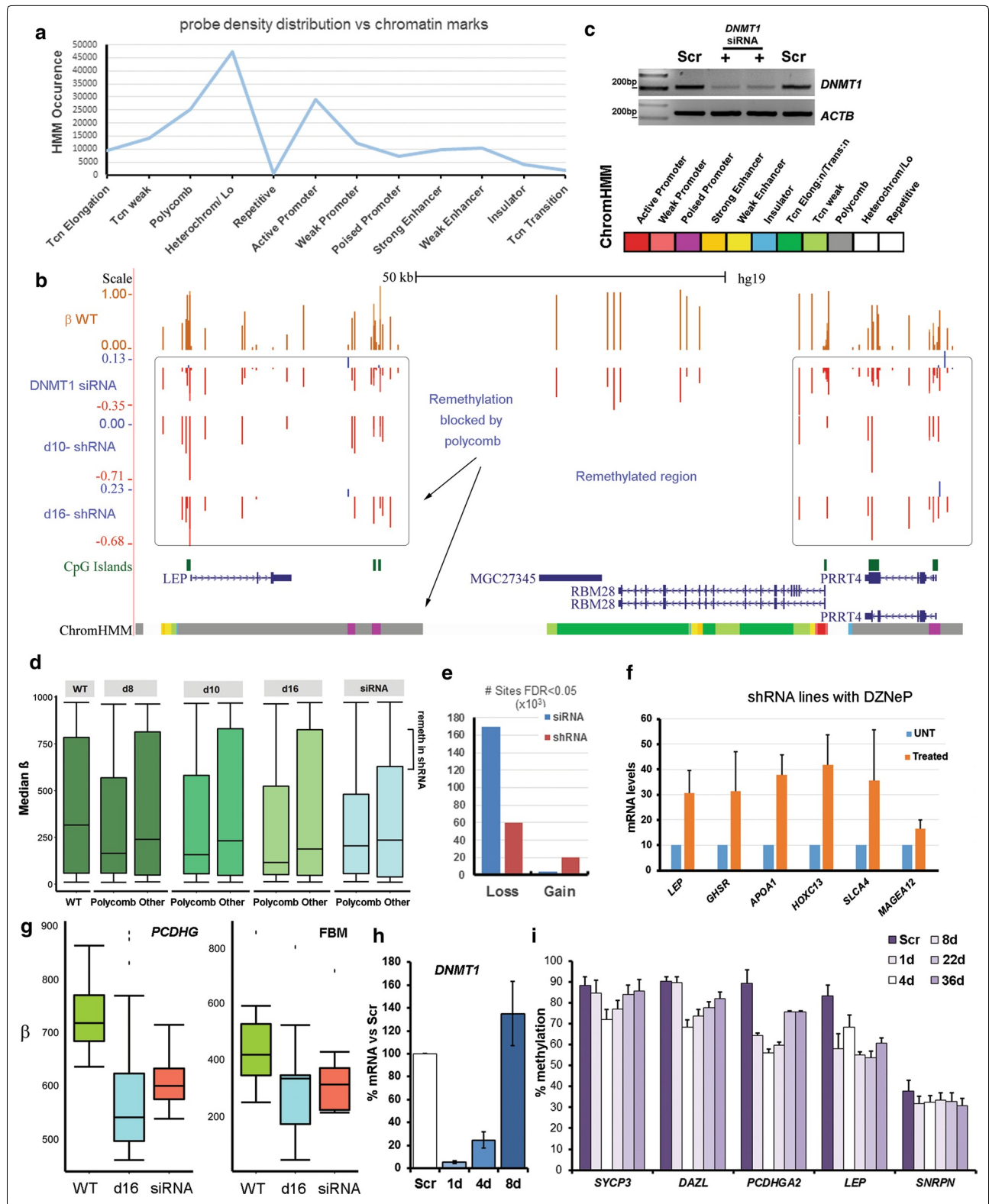
*MAGEA12* gene confirmed significant gains in methylation at the *GABRQ* promoter and in the *MAGEA12* gene body (Fig. 5c). Clonal analysis for *GABRQ* indicated a uniform increase in methylation (78 vs. 16%) across all adjacent CpG at this locus (Fig. 5d). Both direction and degree of change in methylation were highly correlated between pyrosequencing and the 450K array across all sites which were covered by both types of assay ( $r = 0.916$  for loss of methylation  $r = 0.818$  for gain in methylation).

**Transcriptional changes are enriched at cancer/testis antigen genes on X and Y**

To see whether methylation changes were accompanied by large-scale changes in transcription, we carried out a genome-wide screen using the HT12 array which assays most RefSeq genes. Figure 7a shows the distribution of changes comparing d8 and WT: genes which showed > 2 fold change (FC) and with scores of  $p < 0.05$  are highlighted, with the greater spread to the right indicating a greater tendency to derepression. Relatively small

numbers of genes were affected (Fig. 6b), particularly at higher stringency ( $FDR < 0.01$ ), and d16 showed fewest dysregulated genes. To determine common targets, we looked for shared genes (Fig. 6c). DAVID analysis on the genes common to all three ( $n = 70$ ; Additional file 6: Table S3) indicated significant enrichment for genes coding for MAGE domains (Fig. 6d). MAGE genes on the X chromosome were previously identified as showing large changes in methylation (Fig. 5): also appearing here was a *TSPY* family member (Table 1, Fig. 3f). Upregulation of members of these gene classes could be verified by RT-PCR (Fig. 6e) and showed similar direction of change to the array, and greater magnitude, by RT-qPCR (Fig. 6f). Consistent with the transcriptional upregulation, median methylation levels at the promoters of these genes were lower than WT (Fig. 6g). Interestingly, there was an overall increase in intragenic (as opposed to promoter) methylation in the larger group of transcriptionally dysregulated genes common to d8 and d10 ( $n = 764$ , see Fig. 6h and Additional file 6: Table S3), which may





(See figure on previous page.)

**Fig. 7** Methylation loss is concentrated at regions normally repressed by polycomb. **a** Distribution of probes showing significant loss per chromatin state—numbers of probes are shown at left, chromatin states below: tcn, transcription; heterochrom/Lo, heterochromatin or low signal; repetitive, repeat DNA. **b** Region around the *LEP* gene: tracks as before, with the addition of data from cells treated with siRNA for 72 h (top). A track showing ChromHMM chromatin states from NHLF foetal lung fibroblasts is shown at bottom: grey, polycomb-repressed; green, transcriptionally active (full colour key at top right). **c** DNMT1 mRNA levels by qPCR following treatment with siRNA (+) for 72 h compared with scrambled control (Scr). *ACTB* is shown as a control; ladder as above. **d** Median  $\beta$  values for all regions (WT) compared to medians for polycomb-repressed regions (Polycomb), or all other regions (Other) in the cell lines indicated at top; remeth, remethylated. **e** Numbers of probes showing loss and gain in methylation in hTERT cells following treatment with siRNA for 72 h compared with the shRNA lines (averaged); #, number. **f** mRNA levels for the indicated genes in shRNA lines treated with the EzH2 inhibitor DZNep; UNT, untreated; bars represent SEM, experiment carried out in duplicate. **g** Median  $\beta$  values for all variable exons at the *PCDHG* locus (left) and for fat/body mass genes (FBM, right): compared 16 shRNA lines with cells treated with siRNA. **h** DNMT1 mRNA levels in WT cells exposed to siRNA for 48 h, then allowed to recover in normal medium; comparisons were made to a scrambled siRNA negative control (Scr). **i** Methylation levels by pyroassay at the loci indicated during the transient KD and recovery shown in (h); timepoints are in days. All loci showed significant loss of methylation: *LEP* and *SNRPN* showed no significant gain versus lowest methylation level, while *PCDHGA2* showed no significant gain between d22 and d36

reflect increasing gene body methylation accompanying transcription.

### Regions hypomethylated in shRNA lines correlate with polycomb repression

To investigate why losses in methylation occurred at the same positions in all KD lines, we used ENCODE data to look at chromosomal distribution, replication timing and chromatin features which might be important, since the DNMTs have no DNA sequence specificity themselves. Of these, the chromatin marks were most informative, in particular the ChromHMM dataset on lung fibroblasts which partitioned the genome into different types of chromatin based on a set of distinguishing histone marks and other features [43]. This indicated that probes significantly losing methylation in our shRNA lines are most densely distributed across regions which are normally polycomb-repressed or are heterochromatic/low-signal regions in lung fibroblasts (Fig. 7a). Specifically, many regions show a striking correlation between polycomb marking and methylation loss, such as the *LEP* and neighbouring *PRRT4* genes (Fig. 7b): in contrast, the intervening *MGC27345* and *RBM28* genes at that locus, which are highly methylated in WT cells (top track), show little or no loss of methylation and have chromatin marks associated with transcription.

These data suggested that polycomb-repressed regions might be more susceptible to demethylation than others. To test whether these regions lost methylation more readily than others, we treated hTERT1604 with siRNA for 72 h, which led to acute depletion of the *DNMT1* mRNA (Fig. 7c). We found, however, that there was little difference between polycomb-repressed and other regions in terms of demethylation in the siRNA-treated lines (Fig. 7d), in contrast to the shRNA lines where losses were concentrated at the former (Fig. 7d). This could also be seen at the *LEP* locus, where *MGC27345* and *RBM28* now showed loss of methylation following

siRNA treatment (Fig. 7b, siRNA track). Also of note, almost no probes showed gains in methylation relative to WT in the siRNA cells (Fig. 7e), indicating that this effect is associated exclusively with chronic treatment. These results suggested that gains of methylation had occurred only in shRNA lines and had effectively restored methylation to near WT levels at most regions outside of those marked as polycomb-repressed.

Since transcriptional analysis did not highlight dysregulation of polycomb regions in shRNA cells (Fig. 6d), we tested to see whether polycomb-mediated repression was being maintained there in the absence of DNA methylation. To do this, we treated with DZNep, an inhibitor of EZH2, and confirmed the upregulation of a positive control gene *SLCA4* (Fig. 7f) as previously reported [44]. Likewise, *HOXC13*—a known polycomb target—showed derepression (Fig. 7f). The FBM genes marked by polycomb including *LEP* showed reactivation to a comparable degree to *SLCA4*, whereas the *MAGEA12* gene which is in a heterochromatic region not marked by polycomb showed little effect (Fig. 7f).

To further investigate the difference between acute and chronic DNMT1 depletion in these cells, we first examined the effects of acute depletion by siRNA on the loci identified in the stable lines: this confirmed that loci such as the clustered protocadherins and the fat/body mass genes also lose methylation on short-term depletion by siRNA (Fig. 7g). Following treatment, cells were then allowed to recover in the absence of siRNA for an extended period (36 days). DNMT1 levels returned to normal rapidly (Fig. 7h). Examination of the methylation response at various gene classes was very instructive. Germline genes (*SYCP3*, *DAZL*), which are known to become de novo methylated to high levels during somatic differentiation [5, 34], showed initial loss versus a scrambled control (Scr), followed by remethylation over time to near WT levels (Fig. 7i), confirming that the hTERT cells possess sufficient de novo activity to remethylate

the genome, as already suggested (Fig. 7b–e). Imprinted genes are normally unable to regain methylation somatically [45], and we could confirm that the *SNRPN* imprint control region failed to remethylate (Fig. 7i). The polycomb-marked genes *LEP* and *PCDHGA2* were also refractory to de novo methylation, either showing no gain (*LEP*) or reaching a plateau at an intermediate level of recovery only (*PCDHGA2*) (Fig. 7i).

### Gain in methylation is associated with poised promoters in shRNA lines

Having established that loss of methylation in shRNA lines is linked to polycomb repression, we wished to determine what features are associated with gains in methylation in these chronically depleted cell lines. As indicated, gains were not seen genome-wide following acute depletion using siRNA (Fig. 7e) and specific loci such as *UGT1A* showed instead loss of methylation on acute treatment (Fig. 8a, siRNA track), suggesting that hypermethylation is associated with longer-term culture of the shRNA-containing cell lines. To investigate what features might be associated with such loci, we looked to see which chromatin states in shRNA lines showed the highest median  $\beta$  for probes which gained methylation and the largest difference in methylation (Fig. 8b). This identified weak and poised promoter categories, and comparing shRNA lines to WT (Fig. 8c), the median values were more different for poised than for weak promoters (0.4 vs. 0.2, Cohen's *D* test). These results suggested that poised promoters attract de novo methylation particularly strongly. Consistent with this, hypermethylation in the shRNA lines is centred around the *UGT1A* promoters and not the common 3' exons (Fig. 8a). A heterochromatic location may contribute to over-methylation, since genes in adjacent active chromatin show restoration of normal methylation (Fig. 8a, compare siRNA to d10, d16 for *DGKD*), but not hypermethylation. While *UGT1A* transcription levels were very low compared to expressing cells by RT-qPCR (not shown), available HT12 array data showed a consistent decrease in transcription in all three shRNA lines (Fig. 8d, left), correlated with gains in methylation at the cognate promoters (Fig. 8d, right).

Further analysis confirmed that while gains in methylation were seen across all the *UGT1A* exons in all shRNA lines (Fig. 8e), all of these exons showed a loss of methylation following acute depletion with siRNA. We took advantage of our transient depletion and recovery experiment (Fig. 7h, i) to examine levels of methylation at *UGT1A4* using pyrosequencing: this showed that while the region indeed loses methylation on acute depletion, it undergoes steady de novo methylation following recovery and at day 36 was the only gene examined whose

methylation exceeded that seen in the scrambled control (32.4 vs. 31.3%), suggesting that these genes are indeed susceptible to hypermethylation.

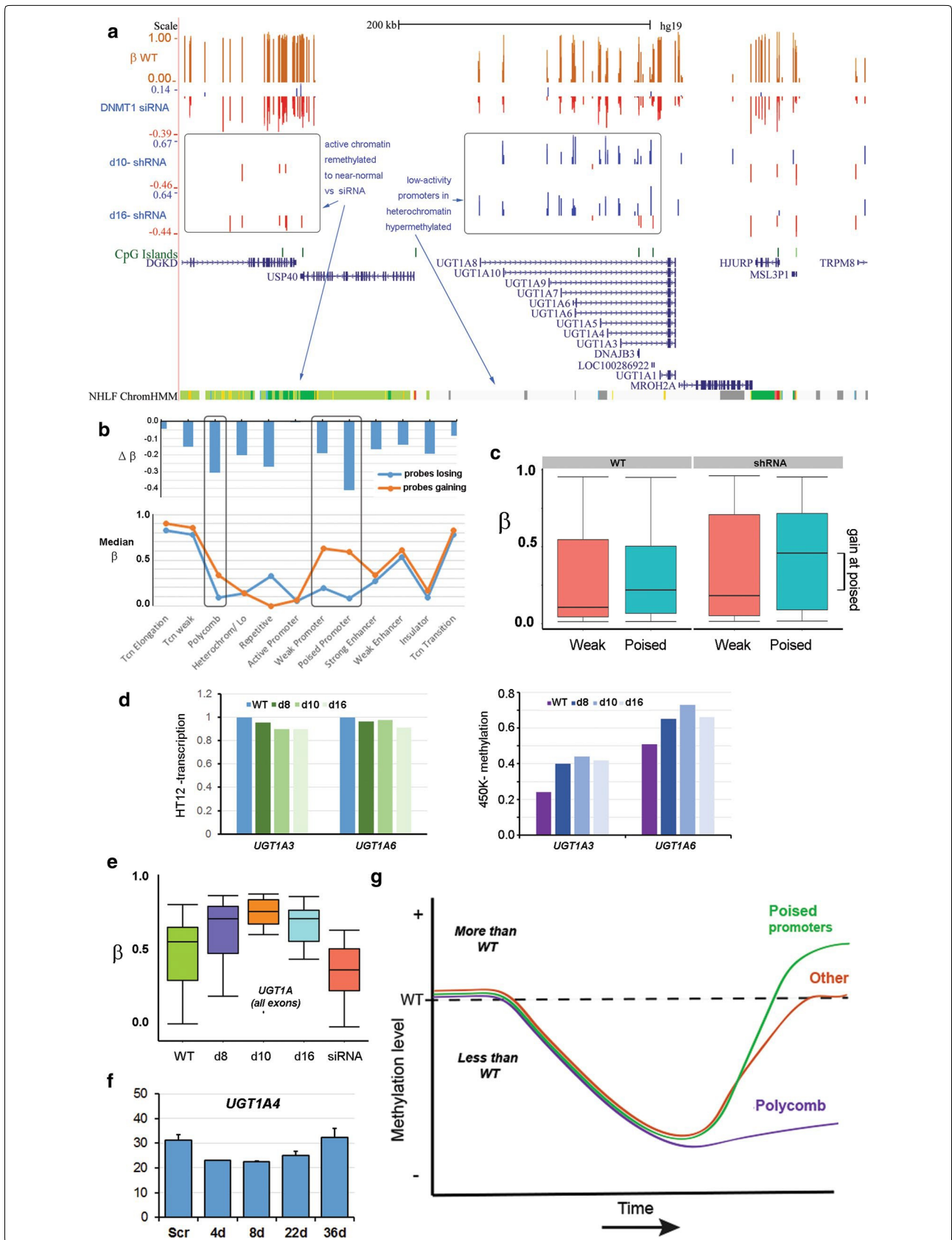
One possible reason for the gains in methylation seen in the shRNA lines could be over-expression of a de novo enzyme. Previous reports have indicated that between them, DNMT3B and DNMT1 account for the majority of methylation in cultured adult human cells and that there may be a role for DNMT3B in maintenance as well as de novo methylation [27]. We saw little change in *DNMT3B* levels in the *DNMT1* KD lines from the HT12 transcriptional array (Additional file 7: Fig. S4A) or RT-PCR (not shown), indicating that gains in methylation are not due to *DNMT3B* over-expression. To investigate a possible role in maintenance methylation, we carried out a transient siRNA treatment and could achieve robust knockdown of *DNMT3B* in the cells (Additional file 7: Fig. S4B). While some germline genes showed little effect, loci previously shown to require DNMT3B including *D4Z4* and *NBL2* did show loss of methylation (Additional file 7: Fig. S4C), confirming that we had achieved a functional depletion. Examination of the loci identified in our *DNMT1* shRNA clones showed that these loci also showed loss of methylation in *DNMT3B* KD cells (Additional file 7: Fig. S4C), suggesting that loci which remain hypomethylated in the shRNA clones also require input from DNMT3B to retain WT methylation levels.

## Discussion

### Summary and model

We and others have previously shown that acute depletion of DNMT1 using siRNA triggered the DNA damage response and cell cycle perturbations in human cell lines, making it difficult to identify genes which are directly controlled by methylation. Here we used isogenic shRNA-containing derivatives of a normosomic lung fibroblast cell line to look at the effects of chronic depletion of the protein. We characterised the alterations in methylation and transcription using microarrays in three different cell lines, processing them using a highly reproducible pipeline, and verified changes using locus-specific pyrosequencing or RT-qPCR assays. Additionally, we compared the effects on methylation of this chronic depletion to the effects of acute depletion using siRNA, as well as investigating possible contributions by DNMT3B. Finally, we investigated the correlations between chromatin state and DNA methylation and showed a role for polycomb-mediated repression at some of the loci.

Our results show that while both siRNA and shRNA-treated cells lose methylation overall as would be expected, only the latter show gains in methylation, most likely reflecting selection against the deleterious effects





(See figure on previous page.)

**Fig. 8** Methylation gain is concentrated at poised promoters. **a** *UGT1A* locus showing siRNA treatment data (top), shRNA lines (middle) and chromatin states (bottom); grey, heterochromatin/low signal; green, transcriptionally active (for full key see previous fig). **b** Median  $\beta$  levels for probes gaining and losing in shRNA lines (bottom) and median changes in methylation ( $\Delta\beta$ ) versus WT for different chromatin states. **c** Boxplots of methylation values for probes falling within weak and poised promoter chromatin regions in WT or shRNA lines (averaged). **d** Transcription at the *UGT1A3* and *UGT1A6* genes decreases (relative to WT, set to 1) in all three shRNA lines as methylation ( $\beta$  value) increases, as indicated by HT12 and 450K arrays, respectively. **e** Median methylation ( $\beta$ ) across all *UGT1A* exons decreases in siRNA-treated cells, but shows gains in all shRNA lines. **f** Methylation at *UGT1A2* during the transient KD and recovery experiment shown in Fig. 7h, i; differences are significant between control (Scr) and d4, but not Scr versus d36. **g** Model for methylation changes which occurred over time following chronic (shRNA) depletion of *DNMT1*: while polycomb-marked regions (purple) resisted remethylation, most regions ("other", red) regained normal or near-normal levels, while poised promoters (green) tended to become hypermethylated

of hypomethylation during clonal expansion and culture. Figure 8e shows what we propose to have occurred: shRNA treatment gave initial widespread demethylation in all three clonal lines, since each line shows the presence of some highly demethylated sites distributed across the genome, but methylation seems to have recovered at most CpGs (Fig. 8e red line). Comparison to normal chromatin patterns in human lung fibroblasts indicated that remaining hypomethylation in the expanded cells was concentrated at regions normally marked for repression by polycomb (Fig. 8e purple line), while the smaller number of regions becoming hypermethylated relative to the parental cell line are associated with poised promoters (green line). TET expression was not detected, and the cells had little or no 5-hydroxymethylation (5hmC; data not shown), in keeping with other reports [46], suggesting that the hypermethylation does not represent 5hmC. Likewise, no over-expression of DNMT3B was detected.

In terms of what type of gene was particularly affected by chronic DNMT1 KD, the enrichment analyses and laboratory verification consistently pointed at the same small group of gene categories, namely (1) neuroepithelial genes, and in particular the protocadherins; (2) fat homeostasis/body mass genes; (3) olfactory receptors; (4) the cancer/testis antigens; and (5) the *UGT1A* complex.

### Protocadherins are major targets of DNA methylation in human cells

Emerging evidence suggests that the clustered protocadherin genes may be central to specifying individual neural cell identity [47, 48] and they have been shown to become heavily methylated during embryonic development in mouse [49], suggesting that stable repression of non-transcribing copies is a programmed event during development. Recent work has shown that DNMT3B is important for de novo methylation at these loci and suggested that dysregulated expression may contribute to the phenotype in immunodeficiency, chromosome abnormalities and facial anomalies (ICF) syndrome [50], where

*DNMT3B* is frequently mutated [51], and we found that depletion of DNMT3B was accompanied by loss of methylation at *PCDHGA2*. The *PCDHA* and *PCDHB* loci are heterochromatic and show persistent loss of methylation, as does the 5' end of the *PCDHG* locus which is polycomb-repressed, but not the 3' end which shows little loss of methylation and has instead chromatin marks associated with weak transcription (Additional file 3: Fig. S2B). Meehan and co-workers recently showed that long-term loss of DNA methylation in mouse *Dnmt1*  $-/-$  ESC cells led to spreading of polycomb marks (in particular H3K27me3): their analyses singled out the *Pcdh* genes, which were heavily methylated in WT but not mutant ESC, as also shown by others [52]. Reddington et al. [17] also showed an increase in H3K27me3. A similar sequence of events in our human cells would cause an increase in H3K27me3 on *PCDH* genes and potentially help block remethylation. The sensitivity of the protocadherin cluster to methylation changes may explain why these genes are frequently identified in screens for differentially methylated loci in cancer [53]. The lack of derepression in our stable fibroblast cells is unsurprising here since expression of these genes is restricted to neurons [54]: they are also, with the exception of part of the *PCDHG* complex, heterochromatic rather than polycomb-repressed and may as such be harder to reactivate.

### Fat/body mass genes can be repressed by DNA methylation and polycomb

Currently, there is much interest in the possibility that altered diet, folate status or exposure to environmental toxins may lead to stable changes in the human methylome which particularly affect metabolic processes, as this offers an attractive mechanism by which it may be possible to partly explain the foetal origins of adult disease [55, 56]. Enrichment analysis in our cells identified the FBM genes involved in the common processes of lipid storage and body mass homeostasis, including *LEP*, *GHSR* and the *APOC* cluster. These loci are readily demethylated on acute DNMT1 depletion and remain demethylated in chronically depleted cells where many other loci have

recovered methylation. These loci are heavily marked by polycomb in normal fibroblasts, rather than being heterochromatic, which can potentially explain both their resistance to remethylation and their lack of transcriptional depression in the stable lines. In keeping with this, inhibition of the polycomb repressor EZH2 which generates H3K27me3 marks could reactivate these genes, as well as the canonical polycomb targets the *HOX* genes. These results suggest that in cells which have both DNA methylation and polycomb-mediated repression, both layers of repression must be removed to achieve gene activation. Interestingly a recent report by Hajkova and colleagues showed that reprogramming of germ cells in mouse also required both removal of DNA methylation and alteration of polycomb marks [57].

### Olfactory genes are methylated and largely inert

Olfactory receptors are also involved in specification of neural cell identity, where individual receptors are expressed in only a small group of cells in the olfactory epithelium [58]. They are largely monoallelically expressed, and methylation has been implicated as playing a role in their control [59, 60]. The OR gene family is the largest in the genome, with approx. 380 active members, many organised into “gene factories” where they are flanked by many more pseudogenes and repeats, such as the large cluster on chr11 [41]. These regions are often transcriptionally inert and heterochromatic, which together with the requirement for tissue-specific factors may explain their lack of derepression.

### Cancer/testis antigen genes are particular targets for demethylation and activation

The *TSPY* and *MAGE* genes fall into a functionally defined group known as the cancer/testis antigen (CTA) genes ([61, 62]; <http://www.cta.lncc.br/>) which are expressed during testis development normally, but which are aberrantly expressed in some tumours, such as melanoma and gonadoblastoma (e.g. *TSPY2*). This latter property makes them of particular interest for cancer immunotherapy, and monoclonal antibodies against some CTA members have already gained clinical approval [63]. CTA genes have been shown previously to lose methylation and become derepressed in several cancer cell types after treatment with the methyltransferase inhibitor 5'aza-2-deoxycytidine (Aza) [64–66] and in the HCT116 DNMT1 mutant line [66, 67] using locus-specific approaches. Our study (1) shows in an

unbiased genomic screen that CTA genes are the genes most affected by loss of maintenance activity, (2) shows this for the first time in a normal, differentiated cell line and (3) highlights the subset of CTA genes which are particularly dependent on maintenance activity to keep them repressed. It is noteworthy that the majority of these genes are on the X chromosome, which shows major fluxes in methylation in our stable lines. The genes are largely associated with heterochromatin, rather than polycomb repression, and do not respond to EZH2 inhibition, but rather directly to loss of methylation, which may reflect some difference in heterochromatin marking on the X. Strategies to demethylate and turn on these genes in tumour cells (e.g. with Aza) to facilitate cancer vaccine development may be worthwhile to pursue, given that these genes are the most responsive to loss of methylation in our cell lines.

### UGT1A genes and other poised promoters are susceptible to hypermethylation

From the enrichment analysis, the *UGT1A* gene cluster was highlighted in terms of genes gaining methylation. These genes are known to be highly expressed in skin fibroblasts postnatally, and to be repressed in non-expressing tissues by methylation [68, 69]. The WT cells already had substantial levels of methylation but the increased methylation in the stable cell lines led to small but consistent decreases in transcription on the HT12 array, though levels were so low these could not be confirmed by Taqman qPCR (data not shown). It may be that the particular marks associated with a recent inactivation of the *UGT1A* cluster in the fibroblasts during adaptation to cell culture led to an increased de novo activity here, and in our transient KD experiment we saw the greatest gains in methylation at *UGT1A4*. Consistent with this, hypermethylation relative to the WT cells was associated with weak and poised promoters genome-wide, and the latter showed the greatest tendency to gain methylation above normal WT levels in the shRNA-containing lines.

### Lack of transcriptional changes in part due to polycomb

It is notable that while there was widespread changes in methylation in the KD cell lines, this was not accompanied by large-scale transcriptional derepression, with only a few hundred genes showing dysregulation, and the fold change in transcription being small. Of the four gene classes identified as most affected in terms of methylation, only one—that containing the *TSPY* and *MAGE*

genes—showed robust transcriptional derepression. A lack of global changes in transcription, also reported by others [29, 70], is likely due to in part to the absence of transcription factors in fibroblasts needed to transcribe neural or adipocyte genes at high levels. However, many of the regions showing most persistent hypomethylation are polycomb-marked and this is likely to be sufficient in itself, as it is for example in *Drosophila*, to maintain repression of these genes. However, we could show that in the presence of an EZH2 inhibitor, polycomb-marked loci which lacked DNA methylation, such as those involved in fat homeostasis/body mass regulation, became upregulated, along with canonical polycomb targets such as the HOX genes. Our results therefore indicate both that the polycomb system is sufficient in itself to repress and also that polycomb-repressed regions appear to be refractive to remethylation, which may be due to the action of FBXL10 [71]. It has previously been proposed that the two systems work in parallel, with their own sets of targets and a degree of mutual exclusivity [15–17]: our results would support such a conclusion.

### Comparison to other recent work

Two recent studies have also examined the effects of DNMT1 mutation on DNA methylation and gene transcription in human, albeit in cancer cells [29, 70]. Acute depletion of DNMT1 using an siRNA-mediated approach found, as we did, regions of low CpG density (open sea, etc.) to be most affected, but differed in finding more evidence for cell morphogenesis and phosphorylation pathways being affected [70]. This might reflect differences between acute and chronic depletion and the high levels of cell death during acute depletion. Blattler and colleagues [29] also found that relatively few genes were dysregulated in *DNMT1/3B* double KO HCT116 cells, but some cancer/testis genes (the related GAGE genes) were upregulated, along with Krüppel-associated box genes, while chaperonins figured prominently among down-regulated genes. The latter two gene classes may therefore be more dependent on DNMT3B, or the combination of DNMT1 and 3B, for their maintenance; alternatively the differences may be due to the experiment being carried out in colon cancer cells rather than, as here, in non-transformed fibroblasts.

### Conclusions

In conclusion, our study sheds new light on the loci which are most sensitive to sustained loss of maintenance activity in humans and shows an interplay between polycomb and DNA methylation-mediated repression in these differentiated cells.

## Additional files

**Additional file 1: Table S1.** Details of the primers used in this study.

**Additional file 2: Figure S1.** Variation between shRNA clonal lines. (A) Relative similarities between cell lines based on principal component analysis (PCA) of the 450K data; three independent cultures of each line were analysed. Note the clustering of lines d8R and d10R. The fraction of total variance explained by each component is indicated in brackets. (B) The 1000 sites most variably methylated between cell lines were used for hierarchical clustering. The location of sites with respect to CpG island is indicated at left. Beta values are depicted as shades from red (low) to blue (high).

**Additional file 3: Figure S2.** Changes in methylation levels by genomic element. (A) Protein levels in knockdown lines by western blotting. As a control HCT116 colon cancer cells which are WT or have a homozygous mutation in *DNMT1* (KO) are shown: the DNMT1-specific top band is indicated by the arrowhead at right. (B) Median levels of methylation are shown for each genomic element (listed at top). The positions of medians are also indicated at right (arrowheads). The differences between WT and KD medians were used to plot Fig. 1d. (C) Density distribution of methylation at the three main elements involved in gene regulation, shown by cell line. Demethylation seems most marked at gene bodies (Genes), indicated by increased density of probes at low methylation ( $\beta$ ) values.

**Additional file 4: Figure S3.** Further analysis of enriched genes. (A) Total numbers of sites showing significant changes in methylation at different false discovery rates (FDR). Some sites showing gain were found in each KD cell line alongside the more numerous sites showing loss. (B) Differential methylation between WT and all KD lines using the 1000 best-ranking sites as identified by RnBeads (red). The majority of high-scoring sites common to all three lines lost methylation, but approx. one-third showed gain. (C) Methylation changes at neural identity genes on chromosome 5. Protocadherins in the  $\alpha$  and  $\gamma$  families (*PCDHA* and *PCDHG* genes) have a clustered arrangement, while genes for the  $\beta$  family members are arranged individually. Tracks are as in Fig. 3. The position of the C class variable exons in the *PCDHA* and *PCDHG* clusters are also shown: gain in methylation relative to the siRNA-treated cells can be seen in the boxed regions, which includes the *PCDHG* constant exons, corresponding to transcriptionally active chromatin (green). (D) Median  $\beta$  values for gene bodies for olfactory receptors identified by DAVID: differences were significant by Mann-Whitney U (MWU). (E) Median  $\beta$  values for the promoters of genes in the histone modifier group identified by enrichment analysis in Table 1. No significant differences between WT and KD were found by MWU.

**Additional file 5: Table S2.** Details of the hypomethylated and hypermethylated genes from Figs. 3d and 5a, respectively.

**Additional file 6: Table S3.** Details of the genes showing transcriptional changes in KD cell lines from Fig. 6c.

**Additional file 7: Figure S4.** Role of DNMT3B in hTERT1604. (A) DNMT3B mRNA levels from the HT12 transcription array (3 probes) did not differ substantially in *DNMT1* shRNA cell lines from WT cells. (B) Successful depletion of *DNMT3B* mRNA using siRNA for 48hr, versus a scrambled control (Scr). (C) Methylation levels by pyroassay at the indicated loci: KD, knockdown. Methylation levels at 72hr were similar (not shown).

### Authors' contributions

KON and REI supervised and carried out the majority of wet laboratory work and assembled figures; SJM and SJT carried out the majority of the bioinformatics analyses; AT carried out the DNMT3B work; CB and LM contributed results on specific loci; JL carried out initial KD experiments; DGM supervised and carried out bioinformatics analyses; CPW designed the experiments, carried out bioinformatics analyses, interpreted results and wrote the MS. All authors read and approved the final manuscript.

## Author details

<sup>1</sup> Genomic Medicine Research Group, Centre for Molecular Biosciences, School of Biomedical Sciences, Ulster University, Cromore Road, Coleraine BT52 1SA, UK. <sup>2</sup> AcademieLife Science, Engineering & Design, Saxion University, M.H. Tromplaan 28, 7500 Enschede, Netherlands. <sup>3</sup> Department of Obstetrics and Gynecology, University of Sassari, Via Vienne 2, 7100 Sassari, Italy. <sup>4</sup> Centre for Cancer Research and Cell Biology, Queen's University Belfast, Belfast BT9 7AE, UK. <sup>5</sup> Present Address: The Wellcome-Wolfson Institute for Experimental Medicine, Queen's University Belfast, Belfast BT9 7AE, UK. <sup>6</sup> Present Address: Terry Fox Laboratory, BC Cancer Research Centre, 675 West 10th Avenue, Room 13-112, Vancouver, BC V5Z 1L3, Canada.

## Acknowledgements

We thank Julien Bauer, Daniel Harkin and Steven McLoughlin for help with bioinformatics analysis, Andrew Irwin, Lee McCahon, Anna McLaughlin and Bob Goodman for technical help, Paul Thompson for advice and other members of the laboratory for critical comments. We are obliged to Bert Vogelstein for the HCT116 WT and DKO cells.

## Competing interests

The authors declare that they have no competing interests.

## Availability of data and materials

Data from the 450K and HT-12 arrays have been deposited with the Gene Expression Omnibus database at the National Centre for Biotechnology Information, USA, under the Series number GSE90012. Supplementary Figures and Tables are available in the online version. Cell lines or other materials are available from the corresponding author on request.

## Ethics approval and consent to participate

Not applicable.

## Funding

Work in the Walsh laboratory was funded by grants from the Medical Research Council (MR/J00773/1) and the ESRC/BBSRC (ES/N000323/1).

## Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Received: 4 January 2018 Accepted: 21 March 2018

Published online: 29 March 2018

## References

- Smith ZD, Meissner A. DNA methylation: roles in mammalian development. *Nat Rev.* 2013;14:204-20.
- Edwards JR, Yarychivska O, Boulard M, Bestor TH. DNA methylation and DNA methyltransferases. *Epigenetics Chromatin.* 2017;10:23. <https://doi.org/10.1186/s13072-017-0130-8>
- Li E, Beard C, Jaenisch R. Role for DNA methylation in genomic imprinting. *Nature.* 1993;366:362-5.
- Beard C, Li E, Jaenisch R. Loss of methylation activates Xist in somatic but not in embryonic cells. *Genes Dev.* 1995;9:2325-34.
- Weber M, Hellmann I, Stadler MB, Ramos L, Paabo S, Rebhan M, et al. Distribution, silencing potential and evolutionary impact of promoter DNA methylation in the human genome. *Nat Genet.* 2007;39:457-66.
- Krebs AR, Dessus-Babus S, Burger L, Schübeler D. High-throughput engineering of a mammalian genome reveals building principles of methylation states at CG rich regions. *Elife [Internet].* 2014;3:e04094.
- Wachter E, Quante T, Merusi C, Arczewska A, Stewart F, Webb S, et al. Synthetic CpG islands reveal DNA sequence determinants of chromatin structure. *Elife [Internet].* 2014;3:1-16.
- Walsh CP, Bestor TH. Cytosine methylation and mammalian development. *Genes Dev [Internet].* 1999;13:26-34.
- Irizarry RA, Ladd-Acosta C, Wen B, Wu Z, Montano C, Onyango P, et al. The human colon cancer methylome shows similar hypo- and hypermethylation at conserved tissue-specific CpG island shores. *Nat Genet [Internet].* 2009;41:178-86.
- Jones PA. Functions of DNA methylation: islands, start sites, gene bodies and beyond. *Nat Rev Genet [Internet].* 2012;13:484-92.
- Neri F, Krepelova A, Incarnato D, Maldotti M, Parlato C, Galvagni F, et al. Dnmt3L antagonizes DNA methylation at bivalent promoters and favors DNA methylation at gene bodies in ESCs. *Cell.* 2013;155:121-34.
- Wu H, Coskun V, Tao J, Xie W, Ge W, Yoshikawa K, et al. Dnmt3a-dependent nonpromoter DNA methylation facilitates transcription of neurogenic genes. *Science.* 2010;329:444-8.
- Irwin RE, Thakur A, O'Neill KM, Walsh CP. 5-Hydroxymethylation marks a class of neuronal gene regulated by intragenic methylcytosine levels. *Genomics [Internet].* 2014;104:383-92.
- Margueron R, Reinberg D. The Polycomb complex PRC2 and its mark in life. *Nature [Internet].* 2011;469:343-9.
- Lynch MD, Smith AJH, De Gobbi M, Flenley M, Hughes JR, Vernimmen D, et al. An interspecies analysis reveals a key role for unmethylated CpG dinucleotides in vertebrate Polycomb complex recruitment. *EMBO J [Internet].* 2012;31:317-29.
- Weinhofer I, Hehenberger E, Roszak P, Hennig L, Köhler C. H3K27me3 profiling of the endosperm implies exclusion of polycomb group protein targeting by DNA methylation. *PLoS Genet.* 2010;6:1-14.
- Reddington JP, Perricone SM, Nestor CE, Reichmann J, Youngson NA, Suzuki M, et al. Redistribution of H3K27me3 upon DNA hypomethylation results in de-repression of Polycomb target genes. *Genome Biol.* 2013;14:R25.
- Goll MG, Bestor TH. Eukaryotic cytosine methyltransferases. *Annu Rev Biochem [Internet].* 2005;74:481-514.
- Lei H, Oh SP, Okano M, Juttermann R, Goss KA, Jaenisch R, et al. De novo DNA cytosine methyltransferase activities in mouse embryonic stem cells. *Development.* 1996;122:3195-205.
- Li E, Bestor TH, Jaenisch R. Targeted mutation of the DNA methyltransferase gene results in embryonic lethality. *Cell.* 1992;69:915-26.
- Okano M, Xie S, Li E. Cloning and characterization of a family of novel mammalian DNA (cytosine-5) methyltransferases. *Nat Genet.* 1998;19:219-20.
- Okano M, Bell DW, Haber DA, Li E. DNA methyltransferases Dnmt3a and Dnmt3b are essential for de novo methylation and mammalian development. *Cell.* 1999;99:247-57.
- Jackson-Grusby L, Beard C, Possemato R, Tudor M, Fambrough D, Csankovszki G, et al. Loss of genomic methylation causes p53-dependent apoptosis and epigenetic deregulation. *Nat Genet.* 2001;27:31-9.
- Liao J, Karnik R, Gu H, Ziller MJ, Clement K, Tsankov AM, et al. Targeted disruption of DNMT1, DNMT3A and DNMT3B in human embryonic stem cells. *Nat Genet.* 2015;47:469-78.
- Chen T, Hevi S, Gay F, Tsujimoto N, He T, Zhang B, et al. Complete inactivation of DNMT1 leads to mitotic catastrophe in human cancer cells. *Nat Genet.* 2007;39:391-6.
- Rhee I, Jair KW, Yen RW, Lengauer C, Herman JG, Kinzler KW, et al. CpG methylation is maintained in human cancer cells lacking DNMT1. *Nature.* 2000;404:1003-7.
- Rhee I, Bachman KE, Park BH, Jair KW, Yen RW, Schuebel KE, et al. DNMT1 and DNMT3b cooperate to silence genes in human cancer cells. *Nature.* 2002;416:552-6.
- Egger G, Jeong S, Escobar SG, Cortez CC, Li TW, Saito Y, et al. Identification of DNMT1 (DNA methyltransferase 1) hypomorphs in somatic knockouts suggests an essential role for DNMT1 in cell survival. *Proc Natl Acad Sci USA.* 2006;103:14080-5.
- Blattler A, Yao L, Witt H, Guo Y, Nicolet CM, Berman BP, et al. Global loss of DNA methylation uncovers intronic enhancers in genes showing expression changes. *Genome Biol [Internet].* 2014;15:469.
- Loughery JE, Dunne PD, O'Neill KM, Meehan RR, McDaid JR, Walsh CP. DNMT1 deficiency triggers mismatch repair defects in human cells through depletion of repair protein levels in a process involving the DNA damage response. *Hum Mol Genet.* 2011;20:3241-55.
- Bibikova M, Barnes B, Tsan C, Ho V, Klotzle B, Le JM, et al. High density DNA methylation array with single CpG site resolution. *Genomics.* 2011;98:288-95.
- Assenov Y, Muller F, Lutsik P, Walter J, Lengauer T, Bock C. Comprehensive analysis of DNA methylation data with RnBeads. *Nat Methods.* 2014;11:1138-40.

33. Ouellette MM, McDaniel LD, Wright WE, Shay JW, Schultz RA. The establishment of telomerase-immortalized cell lines representing human chromosome instability syndromes. *Hum Mol Genet.* 2000;9:403-11.
34. Rutledge CE, Thakur A, O'Neill KM, Irwin RE, Sato S, Hata K, et al. Ontogeny, conservation and functional significance of maternally inherited DNA methylation at two classes of non-imprinted genes. *Development.* 2014;141:1313-23.
35. Aryee MJ, Jaffe AE, Corrada-Bravo H, Ladd-Acosta C, Feinberg AP, Hansen KD, et al. Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. *Bioinformatics.* 2014;30:1363-9.
36. Carbon S, Ireland A, Mungall CJ, Shu S, Marshall B, Lewis S, et al. AmiGO: online access to ontology and annotation data. *Bioinformatics.* 2009;25:288-9.
37. Huang DW, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc.* 2009;4:44-57.
38. Giardine B, Riemer C, Hardison RC, Burhans R, Eltnitski L, Shah P, et al. Galaxy: a platform for interactive large-scale genome analysis. *Genome Res.* 2005;15:1451-5.
39. Karolchik D, Baertsch R, Diekhans M, Furey TS, Hinrichs A, Lu YT, et al. The UCSC genome browser database. *Nucleic Acids Res.* 2003;31:51-4.
40. Kondo T, Bobek MP, Kuick R, Lamb B, Zhu X, Narayan A, et al. Whole-genome methylation scan in ICF syndrome: hypomethylation of non-satellite DNA repeats D4Z4 and NBL2. *Hum Mol Genet.* 2000;9:597-604.
41. Glusman G, Yanai I, Rubin I, Lancet D. The complete human olfactory subgenome. *Genome Res.* 2001;11:685-702.
42. Lau YFC, Li Y, Kido T. Gonadoblastoma locus and the TSPY gene on the human y chromosome. *Birth Defects Res Part C Embryo Today Rev.* 2009;87:114-22.
43. Ernst J, Kheradpour P, Mikkelsen TS, Shores N, Ward LD, Epstein CB, et al. Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature [Internet].* 2011;473:43-9.
44. Fujiwara T, Saitoh H, Inoue A, Kobayashi M, Okitsu Y, Katsuo Y, et al. 3-Deazaneplanocin A (DZNep), an inhibitor of S-adenosylmethionine-dependent methyltransferase, promotes erythroid differentiation. *J Biol Chem.* 2014;289:8121-34.
45. Tucker KL, Beard C, Dausmann J, Jackson-Grusby L, Laird PW, Lei H, et al. Germ-line passage is required for establishment of methylation and expression patterns of imprinted but not of nonimprinted genes. *Genes Dev.* 1996;10:1008-20.
46. Nestor CE, Ottaviano R, Reddington J, Sproul D, Reinhardt D, Dunican D, et al. Tissue type is a major modifier of the 5-hydroxymethylcytosine content of human genes. *Genome Res.* 2012;22:467-77.
47. Yagi T. Molecular codes for neuronal individuality and cell assembly in the brain. *Front Mol Neurosci [Internet].* 2012;5:45.
48. Rubinstein R, Thu CA, Goodman KM, Wolcott HN, Bahna F, Manneppalli S, et al. Molecular logic of neuronal self-recognition through protocadherin domain interactions. *Cell.* 2015;163:629-42.
49. Borgel J, Guibert S, Li Y, Chiba H, Schübeler D, Sasaki H, et al. Targets and dynamics of promoter DNA methylation during early mouse development. *Nat Genet [Internet].* 2010;42:1093-100.
50. Toyoda S, Kawaguchi M, Kobayashi T, Tarusawa E, Toyama T, Okano M, et al. Developmental epigenetic modification regulates stochastic expression of clustered Protocadherin genes, generating single neuron diversity. *Neuron.* 2014;82:94-108.
51. Xu GL, Bestor TH, Bourc'his D, Hsieh CL, Tommerup N, Bugge M, et al. Chromosome instability and immunodeficiency syndrome caused by mutations in a DNA methyltransferase gene. *Nature.* 1999;402:187-91.
52. Otani J, Kimura H, Sharif J, Endo TA, Mishima Y, Kawakami T, et al. Cell cycle-dependent turnover of 5-hydroxymethyl cytosine in mouse embryonic stem cells. *PLoS ONE.* 2013;8:e82961.
53. van Roy F. Beyond E-cadherin: roles of other cadherin superfamily members in cancer. *Nat Rev Cancer [Internet].* 2014;14:121-34.
54. Chen WV, Maniatis T. Clustered protocadherins. *Development [Internet].* 2013;140:3297-302.
55. Barker DJP. The developmental origins of chronic adult disease. *Acta Paediatr Suppl.* 2004;93:26-33.
56. Irwin RE, Pentieva K, Cassidy T, Lees-Murdock DJ, McLaughlin M, Prasad G, et al. The interplay between DNA methylation, folate and neurocognitive development. *Epigenomics [Internet].* 2016;8:863.
57. Hill PWS, Leitch HG, Requena CE, Sun Z, Amouroux R, Roman-Trufero M, et al. Epigenetic reprogramming enables the transition from primordial germ cell to gonocyte. *Nature [Internet].* 2018;555:392-6.
58. McClintock TS. Achieving singularity in mammalian odorant receptor gene choice. *Chem Senses.* 2010;35:447-57.
59. MacDonald JL, Gin CSY, Roskams AJ. Stage-specific induction of DNA methyltransferases in olfactory receptor neuron development. *Dev Biol.* 2005;288:461-73.
60. Colquitt BM, Markenscoff-Papadimitriou E, Duffie R, Lomvardas S. Dnmt3a regulates global gene expression in olfactory sensory neurons and enables odorant-induced transcription. *Neuron.* 2014;83:823-38.
61. Simpson AJG, Caballero OL, Jungbluth A, Chen Y-T, Old LJ. Cancer/testis antigens, gametogenesis and cancer. *Nat Rev Cancer.* 2005;5:615-25.
62. Almeida LG, Sakabe NJ, de Oliveira AR, Silva MCC, Mundstein AS, Cohen T, et al. CTdatabase: a knowledge-base of high-throughput and curated data on cancer-testis antigens. *Nucleic Acids Res.* 2009;37:D816.
63. Gjerstorff MF, Andersen MH, Ditzel HJ. Oncogenic cancer/testis antigens: prime candidates for immunotherapy. *Oncotarget.* 2015;6:15772-87.
64. Samlowski WE, Leachman SA, Wade M, Cassidy P, Porter-Gill P, Busby L, et al. Evaluation of a 7-day continuous intravenous infusion of decitabine: inhibition of promoter-specific and global genomic DNA methylation. *J Clin Oncol.* 2005;23:3897-905.
65. Karpf AR. A potential role for epigenetic modulatory drugs in the enhancement of cancer/germ-line antigen vaccine efficacy. *Epigenetics.* 2006;1:116-20.
66. Koslowski M, Bell C, Seitz G, Lehr HA, Roemer K, Müntefering H, et al. Frequent nonrandom activation of germ-line genes in human cancer. *Cancer Res.* 2004;64:5988-93.
67. James SR, Link PA, Karpf AR. Epigenetic regulation of X-linked cancer/germline antigen genes by DNMT1 and DNMT3b. *Oncogene.* 2006;25:6975-85.
68. Belanger AS, Tojcic J, Harvey M, Guillemette C. Regulation of UGT1A1 and HNF1 transcription factor gene expression by DNA methylation in colon cancer cells. *BMC Mol Biol.* 2010;11:9.
69. Sumida K, Kawana M, Kouno E, Itoh T, Takano S, Narawa T, et al. Importance of UDP-glucuronosyltransferase 1A1 expression in skin and its induction by UVB in neonatal hyperbilirubinemia. *Mol Pharmacol.* 2013;84:679-86.
70. Tiedemann RL, Putiri EL, Lee J-H, Hlady RA, Kashiwagi K, Ordog T, et al. Acute depletion redefines the division of labor among DNA methyltransferases in methylating the human genome. *Cell Rep [Internet].* 2014;9:1554-66.
71. Boulard M, Edwards JR, Bestor TH (2015) FBXL10 protects polycomb-targeted genes from hypermethylation. *Nat Genet* 47(5):479-85.

Submit your next manuscript to BioMed Central and we will help you at every step:

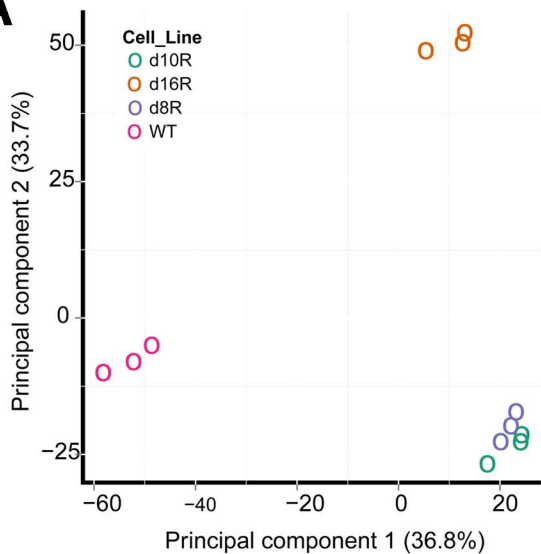
- We accept pre-submission inquiries
- Our selector tool helps you to find the most relevant journal
- We provide round the clock customer support
- Convenient online submission
- Thorough peer review
- Inclusion in PubMed and all major indexing services
- Maximum visibility for your research

Submit your manuscript at  
[www.biomedcentral.com/submit](http://www.biomedcentral.com/submit)

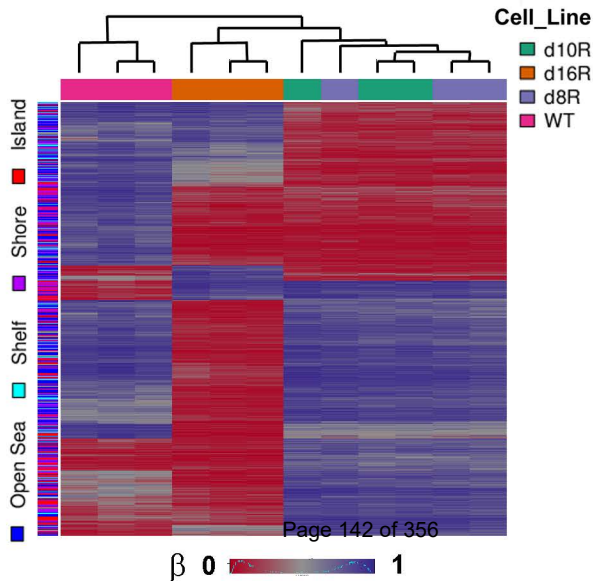


# Supp.Fig.1

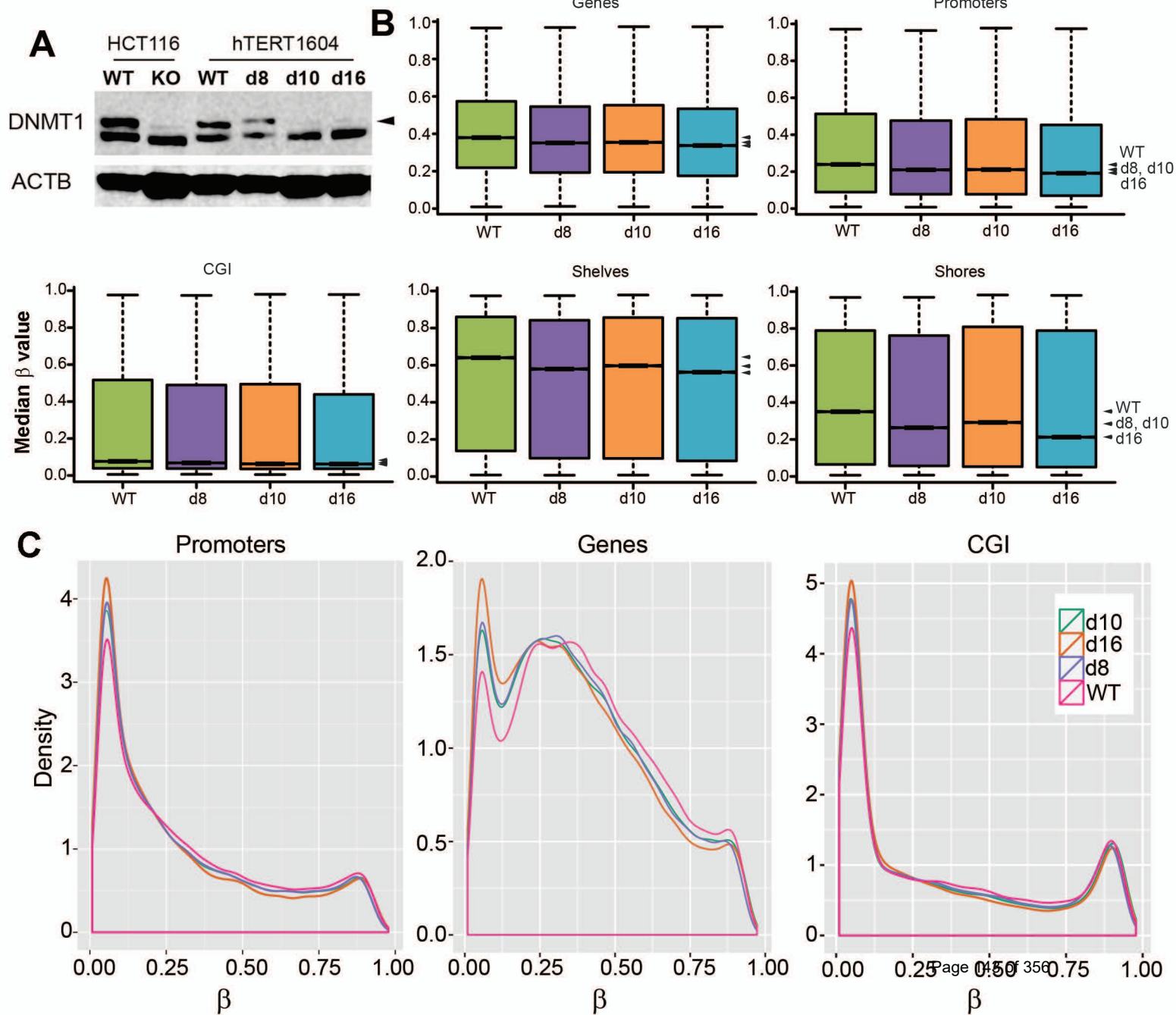
## A



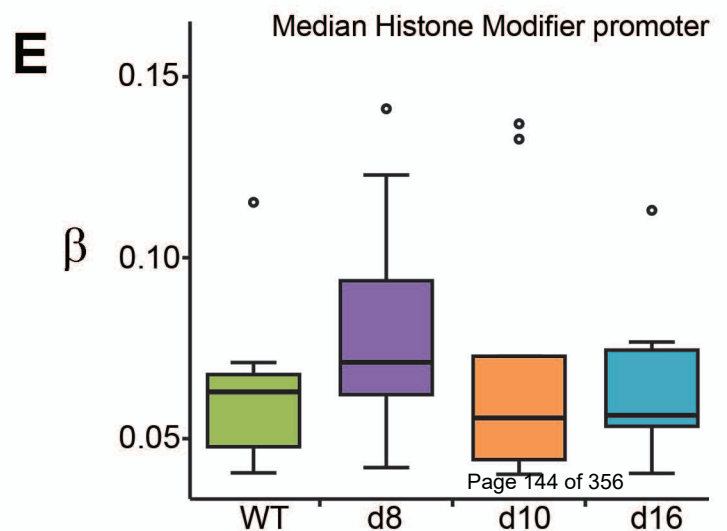
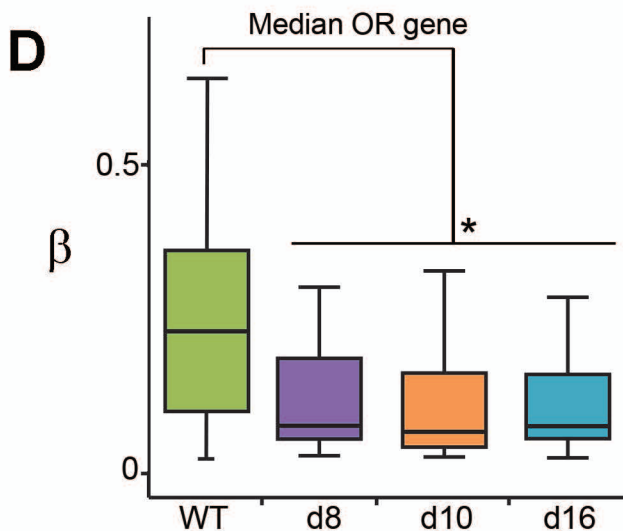
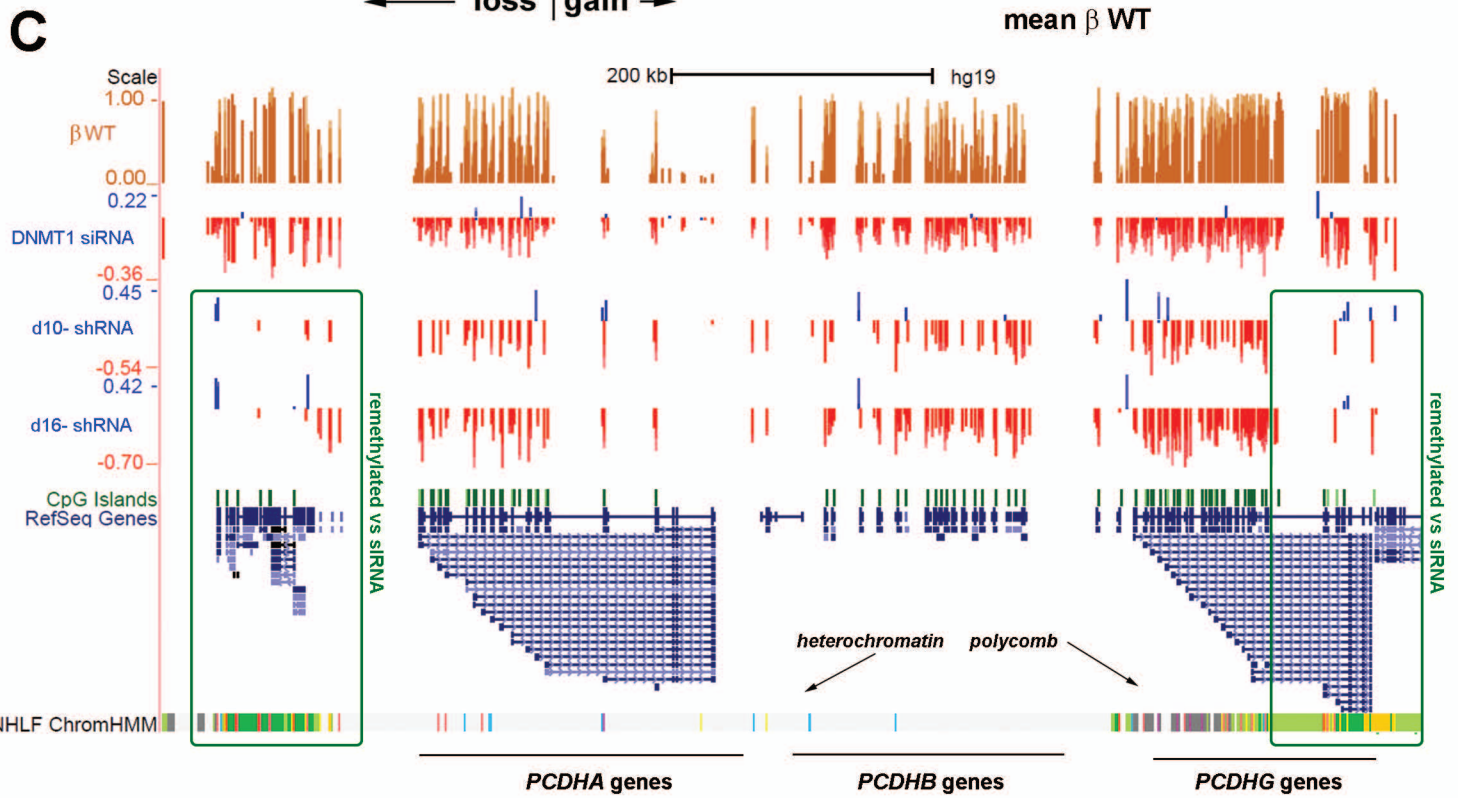
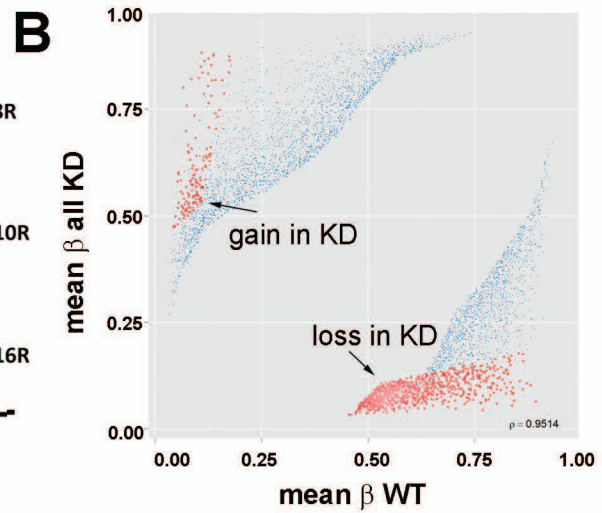
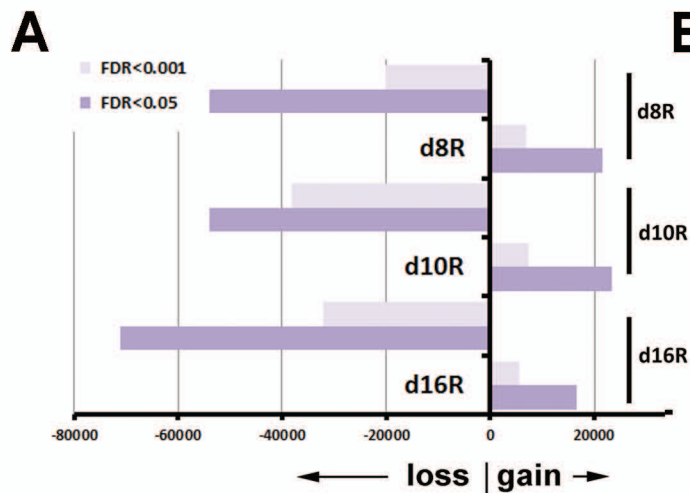
## B



# Supp.Fig.2



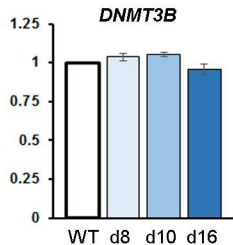
# Supp.Fig.3



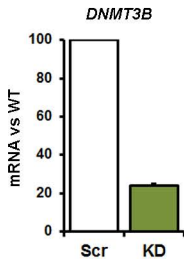


# Supp. Fig.4

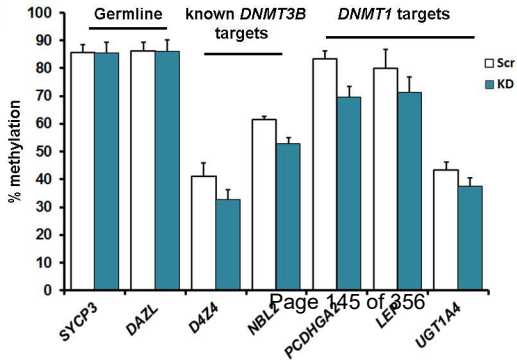
## A



## B



## C



## 3.0 PAPER-II

### **A novel and conserved mechanism for silencing transposable elements**

Rachelle E. Irwin, Catherine Scullion, Sara-Jayne Thursby, Meiling Sun, Avinash Thakur, Scott Rothbart, Gou-Liang Xu, Colum P. Walsh

The main aims of this paper were to:

- Establish a non-cancerous cell line depleted in UHRF1
- Investigate the genome wide effects of UHRF1 depletion on DNA methylation
- To compare the effect on DNA methylation of UHRF1 depletion to that of DNMT1 depletion

### **CONTRIBUTION**

For this paper, I did many of the overall comparisons of methylation between cell lines in R, as well as further developing visual tools such as a graph theme for the MS figures. I contributed substantially to the analysis of transcription data from the HT-12 array, both in developing the viral response gene expression signature of the cell lines and in overall gene enrichment analysis. In order to extract latent data on repeats from the 450K array, I developed new tools in CandiMeth. I further helped process array results from mutant UHRF1 rescues and made subsequent absolute beta and delta beta tracks for viewing on UCSC genome browser. I also reanalysed RNA-seq data of a UHRF1 KD in HCT116 cells from (Chiappinelli et al., 2015), helped visualize RT-qPCR data of repetitive elements in CRISPR mouse UHRF1 alterations via plotting the data as a heatmap and generated the molecular model for the PHD mutations in the human protein.

## **A novel and conserved mechanism for silencing transposable elements**

Irwin RE<sup>1</sup>, Scullion C<sup>1,2</sup>, Thursby SJ<sup>1</sup>, Sun M<sup>2</sup>, Thakur A<sup>1,4</sup>, Rothbart S<sup>3</sup>, Xu G-L<sup>2</sup> and Walsh CP<sup>1,\*</sup>

<sup>1</sup>Genomic Medicine Research Group, Biomedical Sciences, Ulster University, Coleraine, BT52 1SA, UK, <sup>2</sup>Shanghai Institutes of Biological Sciences, 320 Yue Yang Road, Shanghai 200031, China and <sup>3</sup>Van Andel Research Institute, 333 Bostwick Ave, Grand Rapids, MI 49503, USA.

<sup>4</sup>Current Address: Terry Fox Laboratory & Dept. of Medical Genetics, University of British Columbia, Vancouver V6T 1Z4, Canada.

\*Corresponding author: Tel +44(0)28 7012 4484; Email: [cp.walsh@ulster.ac.uk](mailto:cp.walsh@ulster.ac.uk)

## Abstract

While epigenetic mechanisms are known to be important for suppression of Class I transposable elements (TE), relatively little is still understood about the proteins which regulate these mechanisms, and cellular responses to their absence. The UHRF1 protein can interact with both DNA methylation and repressive chromatin marks and has been linked to a range of possible functions. To determine its primary role in adult tissues we first established stable knockdowns in normal human lung fibroblasts. While these showed the expected genome-wide loss of DNA methylation, transcriptional changes were instead dominated by a single response, namely activation of innate immune signalling, consistent with de-repression of TEs. We confirmed using mechanistic approaches that 1) TEs were demethylated and transcriptionally activated, producing double-stranded RNA; 2) activation of interferons and interferon-stimulated genes was crucial to the cellular response and 3) that this pathway was conserved in a number of other adult cell types. Restoring UHRF1 in either transient- or stable knockdown systems could abrogate TE reactivation and interferon response. Interestingly, UHRF1 could impose TE repression in the absence of DNA methylation, but not if the protein contained point mutations affecting H3K9me3 binding. To look at conservation of this pathway we introduced similar point mutations in the mouse *Uhrf1*: homozygous mutants died by mid-gestation with severe developmental delay and failed both to establish DNA methylation and to fully maintain suppression of TEs post-implantation. Our results therefore point to a conserved role for UHRF1 as a key regulator of retrotransposon suppression in differentiated tissues even in the absence of DNA methylation.

## Introduction

DNA methylation is known to play an important role in mice in maintaining suppression at many genes which are transcriptionally inactivated during development and differentiation (Smith and Meissner, 2013), such as those on the inactive X chromosome (Beard et al., 1995), silent alleles of imprinted genes (Li et al., 1993), inactive olfactory receptor genes (McClintock, 2010), some protocadherins (Kawaguchi et al., 2008), and certain germline genes (Weber et al., 2007). These roles have largely been established by introducing mutations or deletions in the genes encoding the DNA methyltransferases either in the whole embryo, or in specific tissues such as the brain or germ line. DNA methylation has also been known for some time to be important for suppression of endogenous retroviruses (ERV) in mice, as hypomorphic mutations in the maintenance methyltransferase DNMT1 result in widespread derepression of Intracisternal A Particles (*IAP*), a young and mobile class of ERV specific to rodents (Walsh et al., 1998).

Less is known about the transcriptional response to loss of DNA methylation in human, where developmental models are lacking. Studies there have been hampered by a strong cell-autonomous DNA damage response which occurs even in undifferentiated cells lacking DNMT1, and acute loss of the enzyme results in cell death within a few cell generations through triggering a DNA damage response (Chen et al., 2007; Liao et al., 2015; Loughery et al., 2011). To circumvent this, we recently generated a hypomorphic series in human by selecting for integration of an shRNA in a normosomic, untransformed normal lung fibroblast cell line hTERT-1604 (O'Neill et al., 2018). Here we found that chronic depletion of DNMT1 resulted in loss of methylation at some targets known from mice, such as protocadherins and olfactory genes, but also at some gene classes specific to human such as the cancer/testis antigen (CTA) genes. In fact, the transcriptional response in these cells was dominated by up-

regulation of the CTA genes, the bulk of which are clustered on chromosomes X and Y (Almeida et al., 2009; Simpson et al., 2005). ]

Early work with pan-DNMT inhibitors (DNMTi) such as 5-azacytidine (5-AZA) or 5-aza-2'-deoxycytidine (5-AZA-CdR) also showed that CTA genes were one of the major direct targets of methylation-mediated repression in adult human tissues (James et al., 2006; Samlowski et al., 2005). These genes are normally expressed to varying levels in testis, but repressed elsewhere in the body in a methylation-dependent manner. Tumour cells often show spontaneous genome-wide hypomethylation and derepression of CTA genes, with presentation of fragments of these proteins as neo-antigens on the cell surface (Karpf, 2006). Recent studies have shown that, as well as acting on CTA, DNMTi treatment led to demethylation and transcriptional up-regulation of endogenous retroviruses (ERV) (Chiappinelli et al., 2015; Roulois et al., 2015). ERV are a type of Class I transposable element (TE) which transpose using a copy-and-paste mechanism going through an RNA intermediate, whereas Class II elements use cut-and-paste. The presence of double-stranded RNA (dsRNA) from ERV in the cytoplasm was recognised by the dsRNA sensors DDX58 (RIG1) and MDA5 (IFIH1), which triggered IRF7 signalling through the mitochondrial protein MAVS. IRF7 translocated to the nucleus and up-regulated interferons (IFN) and interferon-stimulated genes (ISG) which include dsRNA sensors and other upstream components in a feedback loop, triggering an innate immune response including presentation of antigens at the surface and cell-cell signalling (Chiappinelli et al., 2015; Roulois et al., 2015). The extent to which these effects are due to loss of DNA methylation only, or to secondary effects of the inhibitors is currently unclear, since viral mimicry has not been fully characterised in cells carrying DNMT1 mutations (Cai et al., 2017; Chiappinelli et al., 2015) and 5-AZA-CdR is known to affect levels of the histone methyltransferase G9a (Wozniak et al., 2007), while Aza is mainly incorporated in RNA not DNA (Stresemann et al., 2006).

While the effects of DNMTi in humans and studies using mouse mutants have implicated DNA methylation in TE suppression, other mechanisms are also at work to ensure transcriptional suppression and avoid genomic disruption during periods of DNA hypomethylation in germ and stem cells, principally H3K9 trimethylation (Hajkova et al., 2002; Hill et al., 2018; Lees-Murdock et al., 2003). Consistent with this, loss of H3K9me3 leads to up-regulation of TEs and TE-neighbouring genes in mouse stem cells (Karimi et al., 2011). Recent work in human leukaemia has also shown that SETDB1, a H3K9 methyltransferase, was required for repression of both long terminal repeat (LTR)-containing TEs such as ERV, and non-LTR TEs such as the long interspersed nuclear elements (LINEs) (Cuellar et al., 2017). However H3K9me3 levels decrease in differentiated human cells, where DNA methylation is thought to take over as the primary suppressive mechanism (Kassiotis and Stoye, 2016; Mikkelsen et al., 2007).

Mutations in the Ubiquitin-like with PHD and ring finger domains 1 (*Uhrf1*) gene (aka *Np95*) were initially characterised as phenocopying loss of DNMT1 in mouse and resulted in widespread hypomethylation of the genome and dysregulation of imprinted genes, as well as TE such as *IAP* and *LINE-1* (Bostick et al., 2007; Sharif et al., 2007). In cells lacking UHRF1 the DNMT1 protein did not localise correctly to the nucleus and the paired tandem tudor (TTD)- plant homeodomain (PHD) region of UHRF1 has been proposed to allow interaction with chromatin even during mitosis by binding to histone 3 lysine 9 trimethylation (H3K9me3) (Rothbart et al., 2012, 2013). Reports regarding the role of UHRF1 and more specifically H3K9me binding in DNA methylation have varied. Mutations in the TTD-PHD region that affect H3K9me3 binding by UHRF1 have been shown in human to decrease DNA methylation at ribosomal DNA repeats in HeLa cells (Rothbart et al., 2012), but effects at single-copy genes and other regions of the human genome are unknown. In mouse, mutations in the same region gave only a 10% decrease in DNA methylation, which was genome-wide and not just restricted

to repeats (Zhao et al., 2016). Other studies in mouse suggested that UHRF1 mutations caused widespread loss of methylation, but that it only played a minor role in TE suppression (Sharif et al., 2016). In contrast, mutations in the zebrafish homologue were reported to result in ERV derepression in the developing embryo and activation of the innate immune system (Chernyavskaya et al., 2017) as for DNMT1 in human, but through double-stranded DNA rather than dsRNA signalling. Conversely, a recent report by the same group indicated that UHRF1 KO in mouse liver was not sufficient in itself to de-repress TE (Wang et al., 2019).

There is therefore a lack of clarity regarding the role of UHRF1, what the cellular response to loss of this important epigenetic regulator would be, what genes would be most affected, and what the dependence, if any, of DNA methylation on the TTD-PHD domain would be. As complete ablation of UHRF1 caused cell death in mouse ES cells once differentiated (Bostick et al., 2007; Sharif et al., 2007), as well as in differentiated human cells (REI, MS, GLX, CPW data not shown), we used the same approach we recently took with DNMT1 (O'Neill et al., 2018) and generated a hypomorphic series using shRNA in a hTERT-immortalised normal fibroblast cell line as before. An unbiased genome-wide screen showed widespread loss of methylation across most regions, but the major transcriptional response was consistent with viral mimicry, including upregulation of innate immune and CTA genes. This appeared to be triggered by demethylation of TE and the appearance of dsRNA in the cytoplasm. Rescuing the cells with intact UHRF1 could restore TE repression and switch off the viral response. Interestingly, this occurred even in the absence of DNA methylation. Blocking H3K9me3-mediated silencing via knockdown of KAP1, SETDB1 or mutation of the binding pocket on UHRF1 prevented ERV suppression, suggesting this is upstream of DNA methylation. Consistent with this, the same binding pocket mutations cause loss of methylation at ERV in mice post-implantation, with concomitant derepression of ERV and innate immune genes.



## Results

### **Widespread DNA demethylation in cells depleted of UHRF1 is accompanied by a specific innate immune response**

We used our previously-described approach to generate human differentiated cells lacking UHRF1 (Fig.1A). Briefly, normal human fibroblasts which have been immortalised using hTERT (hTERT-1604) were transfected with a construct containing shRNA and individual integrants selected for. Two rounds of experiments were carried out using shRNA targeting the main body (prefix U e.g. U5) or 3'UTR (prefix UH e.g. UH4, UH5) of the gene, results were indistinguishable: those for the index line UH4 are shown here as an example, results from other clones were similar. Initial screening was using reverse transcription-polymerase chain reaction (RT-PCR): cells showing depletion were further expanded and UHRF1 mRNA levels checked by quantitative RT-PCR (RT-qPCR; Fig.1B, Fig.S1A) as well as checking protein levels by western blotting (Fig.1C, Fig.S1B). Lines showing depletion were further analysed using HT12 arrays for transcription, which verified low *UHRF1* levels (Fig.1C), as well as 450K array for DNA methylation. Median methylation levels, expressed as a  $\beta$  value between 1 (fully methylated) and 0 (no methylation) were lower than WT in UH4 (Fig. 1D), and were comparable to the levels seen in our most severe hypomorph for *DNMT1* (d16, Fig.1D), generated using shRNA in a similar manner and previously described in detail (O'Neill et al., 2018). Notably, multiple *UHRF1* hypomorphs with accompanying lower methylation were more readily isolated (Fig.S1C) than *DNMT1* hypomorphs, suggesting a more severe effect of the latter on cell viability.

Examination of the transcription profile using the HT12 array indicated that more probes showed significant differences between UH4 and WT than for d16 versus WT (>8000 UH4 vs <500 d16 using a false discovery rate (FDR) of 0.05, Fig.1E). DNA demethylation was

widespread across the genome in UH4, with over half of all promoters (n=18,826), yellow circle Fig.1F) showing significant ( $>0.1$ ) decreases in  $\beta$  value. When these were compared with genes showing up-regulation from the set of dysregulated transcripts the majority of genes (82.5%) showed demethylation but no derepression, which may reflect either no effect, of the absence of cell type-specific transcription factors required to activate them. A relatively small percentage (10.7%) were both demethylated and upregulated, consistent with a direct role for DNA methylation in their suppression in this cell type (Fig.1F). This included several previously-characterised gene categories known to be regulated, at least in part, by DNA methylation, such as Cancer/Testis antigen (CTA) genes and olfactory receptors (OR), as we described previously in DNMT1 hypomorphs like d16 (O'Neill et al., 2018). Interestingly a third group of genes (6.8%) showed no demethylation but were nevertheless up-regulated (Fig.1F): this suggests an indirect response of genes in this category to loss of DNA methylation.

To investigate transcriptional response more closely, we then carried out gene ontology (GO) analysis of all up-regulated transcripts (not just those which are demethylated) from UH4 cells using the DAVID clustering tool (Huang et al., 2009). Top hits in this analysis (Fig.1G) included several sub-classes of CTA genes (GAGE SPANX, MAGE). The other enriched gene categories included Type I interferon (IFN) signalling, antiviral response and MHC antigen presentation (Fig.1G). In fact, all 10 of the top 10 categories are related to the so-called “viral mimicry” state previously noted in cells treated with DNA methyltransferase inhibitors (Chiappinelli et al., 2015; Roulois et al., 2015).

### **Innate immune signalling is crucial to the cellular response following loss of UHRF1**

The viral mimicry state induced by methyltransferase inhibitors was a response to the presence of dsRNA in the cell, which is detected by specific sensors in the cytoplasm (Fig.2A). These

signal through the MAVS complex in the mitochondrion, releasing transcription factors (TFs) which turn on both interferons (IFN) and interferon-stimulated genes (ISG) in the nucleus (Jensen and Thomsen, 2012). Some of these latter are themselves components of the pathway such as the sensors DDX58 and OAS1 and the TF STAT1, leading to positive feedback-mediated amplification (broken arrows, Fig.2A). Consistent with this, a profile consisting of genes detected as up-regulated in our GO analysis, combined with previously reported viral defence genes, showed a clear up-regulation in UH4, but not d16 cells, compared to WT (Fig.2B). Profiling of the transcriptional response from the HT12 array (Fig.2C) showed activation of components from several parts of the pathway shown in A. Changes in transcription level were most marked for ISG which are at the bottom of the cascade (Fig.2C, right-hand side), including genes with anti-viral and cell death effects, whereas transcriptional changes were least marked or absent for TFs and sensors (Fig.2C, left -hand side), as previously reported for this innate immune pathway (Cuellar et al., 2017). Notably, three of the genes unique to our profile and not previously reported are linked to T-cell signalling (Fig.2C, RHS). We verified sample genes from various parts of the pathway using RT-qPCR (Fig.2D), with results consistent in direction, though not always in magnitude, between the array and the RT-qPCR. Notably, there was no evidence for up-regulation of components of the dsDNA response pathway from our array analysis, consistent with findings in cells exposed to DNA methyltransferase inhibitors. There was also a poor correlation between methylation and transcription for the IFN and ISG genes (not shown), as reported previously for DNMT inhibitor treatment (Roulois et al., 2015), confirming the response is indirect.

In order to investigate the dependence of cellular response on the activation of this innate immune pathway, we tested our model mechanistically. Inhibition of MAVS with siRNA in UH4 caused significant down-regulation of downstream ISG such as *IFI27* (Fig.2E). This included *OAS2* (Fig.2A), which although it is activated by dsRNA and therefore a sensor

(Donovan et al., 2015), is also an ISG and up-regulated transcriptionally by anti-viral signalling (Fig.2C,D) in a feedback loop. Our GO analysis of transcriptional response in UH4 highlighted enrichment for genes involved in type I IFN signalling (Fig.1D), which included *IRF9* and *STAT1* (Fig.2C). Type I IFN binding at the cell surface can activate JAK kinases, which phosphorylate STAT1 and STAT2, causing them to dimerise (Ivashkiv and Donlin, 2014): IRF9 can then associate with these dimers, forming a complex termed the ISGF3 transcription factor, which enters the nucleus and upregulates IFNs and ISGs (Fig.2A). To test if this was happening, we treated cells for 4-7d with Ruxolitinib (RUX), a small-molecule inhibitor of JAK kinases and indeed found a significant down-regulation of target ISGs (Fig.2F).

Our analysis so far suggested that components of the innate immune response were upregulated by depletion of *UHRF1* in the UH4 cells, including type I interferons (Fig.1G) as well as other cell surface and secreted signalling factors such as CCL5 and LY6E (Fig.2C). To test for cell-cell signalling, we transferred media from tissue plates containing UH4 cells to plates with WT cells (Fig.2G): this resulted in up-regulation of ISG including *OAS2* and *IFI27*. All of the results above are consistent with an up-regulation of the dsRNA sensing pathway in the cells, presumably in response to the presence of dsRNA in the cytoplasm of UH4 cells (Fig.2A). Treatment of WT cells with polyI:C, a form of dsRNA, but not with dsDNA, caused up-regulation of the same genes as seen in the UH4 line, confirming that the transcriptional response is consistent with exposure to dsRNA (Fig.2H).

### **The presence of dsRNA correlated with transcriptional derepression and loss of DNA methylation at transposable elements (TE)**

Type I interferon response can be triggered in cells when dsRNA is detected in the cytoplasm: this normally only occurs on infection of cells with viruses which produce dsRNA during their replication cycle, but can also occur if endogenous retroviruses and other Class I TE are

derepressed (Chiappinelli et al., 2015; Cuellar et al., 2017; Roulois et al., 2015). Staining of cells with the J2 monoclonal antibody is a sensitive and specific test for the presence of dsRNA (Weber et al., 2006) and gave a clear positive response in UH4, but not WT cells (Fig.3A). To test for derepression of TEs we used RT-qPCR (Fig.3B) for family members previously shown to be most active in response to epigenetic inhibitors (Cai et al., 2017; Cuellar et al., 2017), as transposable elements are not covered on the HT12 array. This indicated that members of several HERV families were transcriptionally up-regulated, including elements of the HERV-F (HERV-FC2), HERV-H (HERV-H) and HERV-W (HERV-W1) families (Fig.3C). As the fold change was small for a number of the HERVs, but J2 staining was much stronger than seen using polyI:C, suggesting the presence of large amounts of dsRNA, we considered that other Class I TE besides the HERV group might also be up-regulated in UH4. We therefore examined LINE-1 elements, a non-LTR TE which can stimulate an IFN response and which are present at much higher copy number than HERV in the genome (Cuellar et al., 2017). RT-qPCR was again consistent with up-regulation of some of these elements (*L1-PBA*, *L1PI*) in UH4 compared to WT controls (Fig.3B, D). Although small in magnitude (~2-fold), the absolute amount of dsRNA generated would be larger due to the greater copy number of the elements involved.

Having established that specific elements were activated in UH4 cells, we examined control regions in these genes (Fig.3B), where methylation has been shown to act in a repressive capacity. Using pyrosequencing assays (pyroassays) covering multiple CG dinucleotides, we found consistent and significant demethylation of the TE showing derepression, including *HERV-FC2*, *HERV-H* and *LINE-1* (Fig.3E). Examination of individual CG in these regions confirmed significant demethylation across the entire region assayed (Fig.3F). While the 450K array was not designed to assay repetitive elements, a substantial number of probes overlap with regions labelled as TE on the *RepeatMasker* track in UCSC.

Using in-house scripts and a GALAXY workflow we assayed methylation across all TEs in the genome, which showed a significant decrease in median methylation ( $p < 2.2 \times 10^{-16}$ , Kruskal-Wallis test) and greater variability in UH4 cells (Fig.3G). For regions with sufficient probe coverage (Fig.3H), we also found evidence of substantial demethylation of several individual TE families (*HERV-FC2*, *HERV-H*, *LINE-1*, *HERV-3*), but not for all elements (*HERV-K22*).

While the analyses above have concentrated on the UH4 cell line, we confirmed demethylation and up-regulation for TE and ISG for a number of other independently-derived clones from the two rounds of transfection (Fig.S1 C-E).

### **A conserved interferon response follows TE demethylation in multiple cell types**

Our results so far strongly supported a role for UHRF1 in methylation and repression of TE in the hTERT-1604 normal fibroblast line and showed that stable depletion resulted in a robust innate immune response targeted against the dsRNA. We then wished to examine the timing of these events, both in the non-transformed hTERT1604 and in different transformed cells to determine whether loss of UHRF1 triggers the same transcriptional response. Further, we wished to determine whether cells could recover from the loss of the protein and re-establish repression. To this end we carried out a transient or “hit-and-run” experiment (Fig.4A) where we exposed cells to siRNA against *UHRF1* for 48hrs, then switched the cells to normal medium without siRNA and allowed them to recover for up to three weeks. RT-qPCR showed that *UHRF1* levels were effectively depleted to ~25% by 96hrs, after which point they steadily recovered, reaching and even slightly exceeding levels seen in scrambled controls (SCR) by 14 days (14D-Fig. 4B). Consistent with observations in our stable knockdown clones, *HERV-H* mRNA levels were increased versus scrambled controls, starting already at 96hrs, and climbed steadily until 14D, at which point they started to decrease and were back at levels seen in SCR control by 21D (Fig.4B). The ISG gene *IFI27* showed comparable dynamics, increasing from

7D and then decreasing to normal levels or below by 21D. Examination of methylation levels at TE by pyroassay showed loss of methylation at the promoter regions already at 3D (Fig.4C). Interestingly, methylation showed only a modest gain (difference vs 3D not significant by T-test) during the recovery period and remained significantly lower than WT out to 21D, beyond the period during which transcription of the TE and ISG had already normalised (Fig.4C). This was true for both average methylation and levels at individual sites across the promoters (Fig.4D, differences between 7D and 21D not significant except CG2,  $p < 0.05$ ).

We then sought to determine if similar transcriptional responses would be seen in tumour cells. To this end, we performed an identical transient KD and recovery experiment in SKMEL melanoma cell lines, which have a more epithelial character. While transient KD was less efficient in these cells, *UHRF1* levels were depleted to ~50% by 7D, then rapidly recovered to levels seen in scrambled controls by 14D (Fig.4E). This was accompanied by activation of TE and ISG, peaking between 14D-21D, after which point transcription started to decrease again for the ERV, while the ISG was still on but more variable (Fig.4E).

Additionally, we reanalysed a publicly-available dataset (Cai et al., 2017) where *UHRF1* was depleted in HCT116 colon cancer cells using adenovirus-mediated transfection of shRNA and where a limited analysis of ISG by RT-qPCR had been reported. Analysing instead the whole RNA-seq dataset using GO analysis, we found that the top enriched gene class was indeed Type I interferon response, with CTA activation accounting for two more of the top 5 categories (Fig.4F). An additional category among the top 7 was piRNA/meiotic silencing (Fig.4E). Taken together with the results above, this confirmed that *UHRF1* depletion led to reproducible TE demethylation and derepression in multiple cell types, evoking a strong innate immune response.

## **Rescuing stable KD clones with UHRF1 can restore TE repression without re-establishing normal DNA methylation levels**

The transient experiments above suggested, importantly, that TE repression could be re-established without full remethylation (Fig.4C,D). In order to confirm this in a more stable system, we undertook to rescue *UHRF1* expression in UH4 cells by transfecting them with full-length cDNA lacking the 3'UTR which is targeted by the shRNA in UH4 (Fig.5A). Western blotting confirmed the presence of the full-length, FLAG-tagged protein in rescues, termed WT10 (Fig.5A). The WT10 cells showed clear restoration of repression (Fig.5B) at HERVs (*HERV-FC2*) and LINE-1 elements (*L1PBA*). Reinforcing this, normalisation of ISG levels was also seen in WT10 cells by RT-qPCR (Fig.5C). Analysis of overall transcription by HT12 array confirmed widespread shut-down of the innate immune response, with genes from most components of the pathway returning to normal or near-normal levels (Fig.5D, black columns), with the exception of a few genes (*GTSF1*, *BST2*). In contrast, examination of the methylation levels using 450k arrays showed that, despite the presence of WT UHRF1 protein, median methylation levels in WT10 were indistinguishable from UH4 (Fig.5E). There was no increase in methylation ( $\beta$ ) in WT10 vs. UH4 over HERV elements, as confirmed by both array and pyroassay analysis (Fig.5F,G). The same was true of LINE-1 TEs, where methylation at individual sites across the promoter also showed no significant change (Fig.5F,G). These results, taken together with the transient experiments in Figure 4 above, indicate that UHRF1 can restore TE repression even when DNA methylation levels cannot be fully re-established.

## **Hypomethylated cell lines rescued using mutated proteins implicate the Histone 3 tail binding domain of UHRF1 in TE repression**

Since wild type unmutated UHRF1 protein was able to restore TE repression in WT10 cells despite DNA methylation remaining low on these elements, we reasoned that there might



remain another epigenetic mark on the retrotransposons which could be recognised by the protein (Fig.6A). The repressive chromatin mark H3K9me3 is also associated with TE and can be read by UHRF1 through its paired tandem-tudor domain/plant homeodomain (TTD-PHD) (Rothbart et al., 2012, 2013). Western blotting with an antibody to H3K9me3 indicated that levels of this modification were not substantially affected in UHRF1 KD cells (UH4) versus WT (Fig.6B), suggesting that the cells retain this mark. Levels were also not markedly increased in the UH4 cells rescued with full-length intact UHRF1 (WT10), indicating that the reestablishment of repression seen there was not a result of greatly increased H3K9me3 levels (Fig.6B). As a control, even transient depletion of the SETDB1 enzyme responsible for trimethylating H3K9 using siRNA in WT cells was sufficient to give marked loss of H3K9me3 (Fig.6B). These results suggest that TE repression tracks more with UHRF1 levels than H3K9me3, but that H3K9me3 is retained at TE in the absence of DNA methylation and so could potentially act as a cue for repression when UHRF1 was restored.

The UHRF1 protein has been previously shown, through both crystallographic and binding studies, to engage the histone 3 tail through its TTD-PHD region (Fig.6C) with key residues including D334/E335 (PHD) which holds the tail in place, and Y188 (TTD) which interacts with H3K9me3 (Rothbart et al., 2012, 2013). We used the same constructs as before to rescue UH4 cells and isolate clones expressing FLAG-tagged UHRF1 proteins containing these mutations in either the TTD (TTD9) or PHD (PHD1, PHD4, PHD10) domains (Fig.6D). As expected, these expressed the rescued protein to readily-detectable levels, with the variation normally seen with clones (Fig.6D). Unlike cells rescued with intact protein however (WT10, WT18), the cell lines containing mutated UHRF1 showed poor and variable repression of TE (Fig.6F). Furthermore, cells with the point mutations were positive for dsRNA in the cytoplasm using J2 staining (Fig.6G). In this respect they resembled cells with no UHRF1 (UH4), whereas cells rescued with WT protein (WT18) showed little or no staining (Fig.6G). In keeping with

the failure to repress ERV, there remained a robust ISG response (Fig.6H) in the UH4 cell lines rescued with the point-mutated UHRF1 (PHD1, PHD4, PHD10, TTD9), but not when the same UH4 cells were rescued with intact protein (WT10, WT18).

### **Mutations in the PHD domain of mouse UHRF1 cause hypomethylation and transcriptional derepression of TE in developing embryos**

While our results so far implied that UHRF1 can potentially bind the H3K9me3 mark on TE leading to repression of transcription from these elements, we did not see marked de novo DNA methylation in either the stable (Fig.6) or transient (Fig.5) experiments in human. Since we have previously shown de novo methylation activity in these cells is sufficient to restore methylation to WT at some genes (O'Neill et al., 2018), we considered that these adult cells may instead lack other factors required for de novo DNA methylation which are only found earlier in development. To examine the dependence of de novo methylation and TE repression on an intact H3K9me3 binding domain in UHRF1, we generated mouse embryos containing mutations in the PHD domain matching those used in human (Fig.7A). To do so, we crossed C57BL/6 and DBA/2 mice to generate 1-cell embryos, which we then injected with a single-guide RNA targeting the region around the DE amino acids in the PHD domain, together with an oligo containing the desired replacement nucleotides as well as an mRNA for the CAS9 enzyme. The first round of injections (n=306) and embryo transfers (n=11) resulted in no pups, suggesting that mutations were leading to embryonic lethality (Fig.7A).

Consistent with this, homozygous mutant embryos (-/-) harvested at embryonic days 8.75 (e8.75) from round 2 of injections showed developmental delay and hypomethylation of *IAP* compared to WT (Fig.7B, top). Both the retardation and the hypomethylation were more severe by e9.5 compared to WT (+/+) or heterozygous (+/-) embryos (Fig.7B, lower panels). A further round of injection gave one heterozygous founder animal which survived (#13); this

animal was then back-crossed for one generation before intercrossing the heterozygous offspring to generate litters containing all three genotypes (Fig.7A). Examination of DNA methylation at *IAP* LTR showed highly significant decreases in homozygous embryos (HOM) compared to WT or heterozygous (HET) littermates (Fig.7C). This decreased methylation was seen across the whole promoter region assayed (Fig.7D). Highly significant decreases in DNA methylation compared to WT was also seen both in 5'UTR regions of both *IAP* and *LINE-1* elements assayed (Fig.7E).

We also examined transcription of TE in the homozygous (HOM) embryos. These showed significant derepression of *IAP* and *musD* ERV as assayed by RT-qPCR (Fig. 7F), although increases were very variable across individual embryos. Analysis of both ERV and ISG transcription results from RT-qPCR confirmed that retroviral elements belonging to several classes, as well as interferon alpha and a number of ISG, were all generally more active in HOM mutants than in HET or WT (Fig.7G). *Uhrfl* mRNA levels were even across all embryos assayed, consistent with a point-mutated transcript (Fig.7G).

## **Discussion**

We showed here that depletion of UHRF1 protein in differentiated human cells, either transiently or using stable models, causes loss of DNA methylation, up-regulation of TEs and an innate immune response. This was linked to the presence of dsRNA in the cytoplasm, likely originating from derepressed TE, since in rescued cells where the TE have been silenced the dsRNA disappeared. Notably this rescue effect can occur without reintroducing DNA methylation, suggesting a separate mechanism for TE repression independent of methylated cytosine, but still dependent on UHRF1. Mutation in the PHD/TTD domain strongly implicated H3K9me3 as the signal which allowed UHRF1 to bind to TE and repress them in demethylated cells. Consistent with this, mutating the H3K9me3 binding pocket of UHRF1 in mouse

prevented the protein from recruiting DNA methylation to mouse TEs post-implantation, concomitant with embryonic lethality, TE up-regulation and an innate immune response.

### **UHRF1 plays a conserved role in TE suppression**

The data we present here therefore strongly supports an important role for UHRF1 in suppressing TEs which is conserved across species. We showed this here in four different systems using a variety of approaches: 1) stable knockdown in normal human lung fibroblasts 2) transient knockdown in human skin cells 3) bioinformatic analysis of published data on colon cancer cells and 4) mutations in the endogenous gene in mouse embryos. In 3/4 of these cases, we found depletion or mutation of UHRF1 gave up-regulation of TEs and in all four, that it induced an innate immune response targeted against dsRNA. Demethylation was seen at most HERV classes examined in our human cell lines, as well as at the more numerous LINE-1 elements in the genome, and we could detect transcriptional activation of several young HERVs and LINE-1 subtypes which have been reported to be recently active and can be derepressed in response to DNMTi treatment (Cai et al., 2017; Chiappinelli et al., 2015; Roulois et al., 2015) or loss of H3K9me3 (Cuellar et al., 2017). In mouse, demethylation was seen at IAP and LINE-1 elements, and derepression detected for the young TE IAPez-GAG and musD, previously seen to be reactivated in response to DNA demethylation (Bourc'his and Bestor, 2004; Hata et al., 2006; Sharif et al., 2016; Walsh et al., 1998).

The demethylation and activation of TEs on loss of UHRF1 was consistent with previous reports from zebrafish (Chernyavskaya et al., 2017) and mouse (Sharif et al., 2007) embryos, the former also reporting innate immune activation. TE activation and demethylation have also previously been reported in mouse embryonic (Sharif et al., 2007) and neural (Ramesh et al., 2016) stem cells. There have also been some studies which have only detected low-levels of activation of TEs on loss of UHRF1 (Sharif et al., 2016; Wang et al., 2019),

which may be in part due to epigenetic compensation, whereby spreading of repressive histone marks from neighbouring repressed regions can compensate for loss of DNA methylation (Reddington et al., 2013; Wang et al., 2019). However given the derepression of TEs and strong immune response seen by ourselves and others in whole embryos and differentiated cells from three different species, the bulk of the evidence clearly points to a conserved role for the protein in maintaining suppression. UHRF1 is likely to be most important in differentiated tissues, since *Uhrfl*<sup>-/-</sup> ESC are viable until differentiated in vitro whereupon they die (Bostick et al., 2007; Sharif et al., 2007), and our mouse mutant showed lethality at around mid-gestation, consistent with a post-implantation defect. Likewise, mutations in adult stem cell populations in mouse were not lethal until the cells began to differentiate (Ramesh et al., 2016; Wang et al., 2019). It is notable that, despite the many roles attributed to UHRF1, the transcriptional response in our fibroblast cells was dominated by the innate immune activation triggered by the dsRNA, indicating that this is the main cellular response to loss of the protein. However complete loss of function of the gene was lethal in differentiated human cell lines as well as mouse (CPW, REI, data not shown), so we cannot exclude other roles for the protein below the viability threshold. Responses in cell lines (SKMEL, HCT116) with a more epithelial character also showed a strong but less dominant innate immune response, suggesting that responses may show some variation by cell type, which would be consistent with differences in innate immune signalling abilities (Barlow et al., 1984; Kassiotis and Stoye, 2016). However, in all cases TE reactivation and some degree of interferon signalling was seen.

### **A novel function for UHRF1 in TE repression in the absence of DNA methylation**

While some previous work had therefore indicated a role for UHRF1 in TE suppression, this had been tightly coupled with its role in assisting the DNA methyltransferases to localise to the nucleus (Bostick et al., 2007; Sharif et al., 2007, 2016). In contrast, we show here that repression can occur in the absence of DNA methylation. This was shown in i) transient

experiments, where endogenous UHRF1 levels were allowed to recover to normal, and ii) in stable KD experiments, where multiple cell lines were derived from UH4 by rescuing with a wild-type version of the protein (WT10, WT18). In both cases, suppression of both LTR- and non-LTR TEs was seen, as well as switch-off of the innate immune response to viral infection, with the disappearance of dsRNA from the cytoplasm in the case of the stable lines. However neither the transient or stable cell lines showed any significant restoration of DNA methylation at TEs, as assessed using both arrays and pyrosequencing. These results strongly suggest that the presence of UHRF1 alone is sufficient to restore repression, at least in these fibroblast cell lines. These results have three important implications: 1) that DNA methylation in itself is not necessary to repress TEs, at least to a level low enough not to trigger the innate immune response, in these cells; 2) that the UHRF1 protein can mediate repression of the retrotransposons through a mechanism independent of DNA methylation and 3) that there must remain some epigenetic information associated with the TE that allowed UHRF1 to recognise and repress them once protein levels were restored.

It is known from many different studies that TEs can be transcriptionally suppressed by a number of methods, including H3K4 deacetylation, H3K27me<sub>3</sub> and- most commonly seen- H3K9me<sub>3</sub> (Kassiotis and Stoye, 2016). While it may seem initially surprising that repression of TE was seen in cells where UHRF1 levels were restored to normal but DNA methylation remained low, the addition of DNA methyl groups to the TE DNA is thought to be a relatively late stage in repression (Rowe et al., 2013). Studies in mouse ESC lacking all DNA methylation showed that TE derepression was modest (<5-fold) (Matsui et al., 2010), but in differentiating mouse embryos or adult cells IAP derepression was orders of magnitude higher (Matsui et al., 2010; Walsh et al., 1998). In contrast, loss of SETDB1 in ESC gave robust derepression of TE (Matsui et al., 2010). There is also considerable overlap in TE which are labelled with both H3K9me<sub>3</sub> and DNA methylation (Karimi et al., 2011), suggesting that this dual marking may

represent a defence-in-depth against inadvertent activation of these deleterious elements. Consistent with this, H3K9me3 levels remain high on TE in germ cells during the crucial window during development when DNA methylation is reprogrammed to allow resetting of genomic imprints and the histone mark is required at that time for suppression (Hill et al., 2018).

### **H3K9me3 binding was required for UHRF1-dependent TE suppression**

Given a) the tight linkage between H3K9me3, DNA methylation and TE suppression; b) that UHRF1 had been shown to bind H3K9me3, and c) that it could re-establish TE suppression in UH4 cells where there was little DNA methylation, we speculated that the protein might still be able to recognise and bind H3K9me3 marks to facilitate repression. This hypothesis was supported by three lines of evidence here. Firstly, the UH4 cells which lacked DNA methylation and showed TE and innate immune activation nevertheless retained H3K9me3 at levels similar to those in the parental cell line. Secondly, point mutations in the H3K9me3 recognition component of UHRF1 prevented the protein from repressing TE in multiple independent clones with at least two different mutations. Finally, introduction of the TTD mutation into the mouse homologue prevented the accumulation of DNA methylation at TE post-implantation, resulting in TE derepression, innate immune activation and embryonic lethality.

Since loss of UHRF1 is concomitant with loss of DNA methylation in the cells, derepression could be due to either alone, or a combination of both. It is well-established that mutations in mouse DNA methyltransferases can cause derepression of TEs (Bourc'his and Bestor, 2004; Hata et al., 2006; Walsh et al., 1998), while treatment with DNMTi have the same effect in mouse and human (Chiappinelli et al., 2015; Roulois et al., 2015). Interestingly however mutations in human DNMTs have given more equivocal results, with only a partial

signal of TE reactivation reported by Chiappinelli and colleagues (Chiappinelli et al., 2015). We also saw no evidence of an innate immune signal in our isogenic DNMT1-depleted cells (this study and (O'Neill et al., 2018)), or in the cells when treated with 5-AZA-CdR (Mackin et al., 2018), despite hypomethylation of many TE (not shown). However there are limitations to our previous experiments, since a substantial degree of remethylation had occurred in the DNMT1 KD lines (which may be necessary for the survival and outgrowth of clones (Chen et al., 2007; Liao et al., 2015; Loughery et al., 2011)) and this may have targeted young TE which could induce an innate immune response. Likewise, 5-AZA-CdR treatment may not have been sufficiently prolonged or at too high a dose to detect the relatively slow onset of innate immune activation (Chiappinelli et al., 2015; Roulois et al., 2015). Nevertheless, the absence of any clear TE reactivation or innate immune response in those experiments may indicate that loss of UHRF1 is less tolerated than loss of DNMT1 due to its repressive effects on TE which are independent of DNA methylation. Given current interest among cancer biologists in the use of hypomethylating agents to boost tumor cell response to immunotherapy (Jones et al., 2019), further work to tease apart the relative roles of DNMT1 and UHRF1 in TE suppression and innate immune activation among different human cell types would seem warranted.

The UHRF1 protein has several functional domains, including the SRA domain, which binds to hemi-methylated DNA (Arita et al., 2008; Hashimoto et al., 2008) and the RING and UBL domains, which catalyse the transfer of ubiquitin to histone 3 (DaRosa et al., 2018; Foster et al., 2018). We showed here that there was a failure to repress TEs and turn off the innate immune reaction in our cell lines when the PHD or TTD domains were mutated, highlighting the essential role of these regions that cannot be compensated for by other domains. From crystallographic (Xie et al., 2012) and binding studies (Rothbart et al., 2012) it has been shown that the TTD mutation used decreased the protein's ability to interact with H3K9me3, while the PHD mutations interfered with the protein's ability to hold the H3 tail in position and



abrogated binding completely (Rothbart et al., 2013). In our mouse model containing the same mutation in the PHD domain, DNA methylation levels did not increase after implantation at TEs as it normally does (Jahner and Jaenisch, 1985; Okano et al., 1999). This may indicate a failure to efficiently recruit the protein to these elements following the wave of de novo methylation which occurs mid-gestation. In terms of phenotype and timing, the PHD mutation resembles the *Dnmt1<sup>N/N</sup>* mutation in mice (Li et al., 1992) rather than the more severe *Dnmt1<sup>S/S</sup>* or *Dnmt1<sup>C/C</sup>* (Lei et al., 1996) or *Uhrfl<sup>-/-</sup>* mutants (Bostick et al., 2007; Sharif et al., 2007), all of which died at earlier stages, suggesting we have generated a hypomorphic mutation which decreases rather than abrogates function. TE derepression and innate immune signalling were variable among mouse embryos homozygous for our EA=>DD mutation in *Uhrfl*. This may reflect segregation of background alleles in the cross, or a degree of stochasticity in TE activation (Kazachenka et al., 2018) and/or immune response (Barlow et al., 1984). The roughly inverse relationship between embryos positive for TE transcripts and those with innate immune activation may also suggest that embryos showing strong immune signalling are successfully clearing cells showing TE upregulation and vice versa. Loss of DNA methylation in mouse embryos is also known to have multiple effects on imprinting (Li et al., 1993), X-inactivation (Beard et al., 1995) and repression of germline genes (Borgel et al., 2010), which are likely to contribute to variability. Notwithstanding the inter-embryo differences in transcription, a highly consistent and reproducible decrease in DNA methylation at TE was seen in all homozygous mutant embryos tested, firmly establishing the requirement for H3 tail binding by UHRF1 for successful recruitment of DNA methylation to these selfish DNA elements during development.

## **Conclusions**

We have confirmed here using a variety of approaches in both human cells and in mouse embryos that UHRF1 is required to suppress TE expression, consistent with some earlier

reports in mouse and zebrafish. Additionally, we have shown that suppression of TE can be achieved by UHRF1 in the absence of DNA methylation, uncovering a novel mechanism for suppression of these elements. This pathway appears to rely on H3K9me3 binding by UHRF1 as mutation of the cognate binding domain on the protein prevents TE suppression in mouse and human cells. Further work is required to determine exactly how UHRF1 can repress these selfish DNA elements and what other components of the cellular machinery are required.

## **Materials and Methods**

### **Cell culture and transfections**

The wild-type (hTERT1604) lung fibroblast cell line (Ouellette et al., 2000) and derivatives were cultured in 4.5g/l glucose DMEM with 10% FBS and 2× NEAA (all Thermo-Fisher Scientific, Loughborough, UK). SK-MEL-28 cells (kind gift of Paul Thompson) were cultured in 4.5g/l glucose DMEM supplemented with 10% FBS. The hTERT1604 cell lines stably depleted of DNMT1 have been previously described (Loughery et al., 2011; O'Neill et al., 2018). Stable depletion of UHRF1 in hTERT1604 (U5/U10/UH4 lines) for this study used pGIPz Lentiviral shRNAmirs (Horizon/Dharmacon), see Table S5 for sequences, used according to the manufacturer's instructions. Briefly, overlapping primers incorporating siRNA sequences to target UHRF1 were made and ligated into pGIPz. The vector was linearized using *XhoI* and *MluI*, then 1µg transfected into WT cells using Lipofectamine 2000 (Thermo-Fisher Scientific) prior to selection in puromycin (Sigma-Aldrich, Dorset, UK) to isolate single colonies, which were then expanded; selection was removed 24 hours (24hrs) prior to any experimental analysis. Rescue cell lines (UH+R10/18, PHD1/4/10, TTD9) were generated by transfecting UH4 cells with pCMV plasmids containing full length UHRF1 cDNA which was either intact (WT) or contained functional mutations in either the PHD or

TTD domains as previously described (Rothbart et al., 2013); individual colonies were selected in G418 (Sigma-Aldrich) and expanded as above.

For transient knock-down experiments,  $1 \times 10^6$  cells/well were seeded in 6-well plates prior to reverse transfection basically as before (O'Neill et al., 2018) using 100nM ON-TARGETplus SMARTpool siRNA (Table S6) or scrambled control (all ThermoFisher Scientific). Post-transfection, cells were cultured in complete medium to allow recovery, with extraction of RNA and DNA up to 28 days after addition of siRNA. For drug treatment (see Table S7) Ruxolitinib (Absource, München, Germany) was dissolved in DMSO and added to culture media at a final concentration of  $2 \mu\text{M}$ ; negative controls contained just DMSO. For analysis of dsRNA and dsDNA sensing pathways, cells were treated at a final concentration of  $10 \mu\text{g/ml}$  Poly(I:C) or sonicated salmon sperm DNA (Agilent, Stockport, UK) for 72hrs, with fresh media and drug every 24hrs; the nucleic acids were dissolved in sterile phosphate-buffered saline (PBS), heated at  $50^\circ\text{C}$  and cooled on ice to achieve re-annealing into double strands prior to treatment. For the media transfer test, UH4 cells were seeded and grown for 72hrs, then media transferred onto the wild type hTERT1604 cells, which were grown for another 72hrs.

### **Immunohistochemical staining**

Cells were seeded onto glass slides pre-sterilized with 100% ethanol and UV light and allowed to attach overnight. Cells were then fixed in 4% paraformaldehyde in PBS, 10mins before quenching with 0.1M glycine (Sigma-Aldrich), then permeabilized with 0.1% Triton X-100, 15mins and preblocked with 2% BSA (Sigma-Aldrich) for 1hr, room temperature (RT). Slides were incubated with J2 primary antibody (Scicons, Szirák, Hungary, see Table S4 for antibodies) at 1:200 in 2% BSA overnight at  $4^\circ\text{C}$ . The next day, slides were washed and incubated with anti-mouse IgG AlexaFluor 546 (Invitrogen, Paisley, Scotland) antibody at 1:1000, 1hr, RT before washing and adding DAPI mounting media (Santa Cruz Biotechnology,

Heidelberg, Germany). Fluorescent images were taken with a Nikon Eclipse E400 phase contrast microscope and processed using Adobe Photoshop (Maidenhead, UK).

### **DNA analyses**

Genomic DNA was extracted from cells growing in log phase, with each cell line done in triplicate, including one biological replicate. DNA preparation, bisulfite conversion and array hybridization was essentially as previously described (Mackin et al., 2018; O'Neill et al., 2018). Briefly, DNA was isolated using the QIAmp DNA Blood Mini Kit (Qiagen, Crawley, UK), assessed for integrity and quality using a range of measures including agarose gel electrophoresis, UV absorbance and Quant-iT PicoGreen dsDNA assay (Thermo Fisher Scientific). Purified DNA was sent to Cambridge Biological Services where bisulfite conversion was performed using the EZ DNA Methylation kit (Zymo Research, California, USA) and samples were loaded onto the Infinium HumanMethylation450 BeadChip (Bibikova et al., 2011) and imaged using the Illumina iScan.

For pyrosequencing, DNA (500ug) was bisulfite-converted in-house as above, then PCR-amplified using the PyroMark PCR kit using Qiagen's pyrosequencing primer assays or those designed in-house (see Table S1) via the PyroMark Assay Design Software 2.0 (Qiagen). Reaction conditions were as follows: 95°C, 15mins; followed by 45 cycles of 94°C, 30secs; 56°C, 30secs and 72°C, 30secs; final elongation 72°C, 10mins, with products verified on agarose gels prior to pyrosequencing using the PyroMark Q24 (Qiagen).

### **RNA analyses**

RNA was extracted from cells growing in log phase using the RNEasy Mini Kit (Qiagen, Crawley, UK), including a DNase step, according to manufacturer's instructions. Complementary DNA (cDNA) was reverse transcribed in a reaction containing 250-500ng total RNA, 0.5uM dNTPs, 0.25ug random primers (Roche, UK), 1x reverse transcriptase buffer

and 200U RevertAid reverse transcriptase in a total volume of 20ul. Reaction conditions were as follows: 25°C, 10 minutes (mins); 42°C, 60mins; 70°C, 10mins. cDNA was stored at -80°C until use. Each RT-PCR reaction contained 1ul cDNA from the above reaction, 1x buffer, 0.4mM dNTPs, 1uM primers (Table S2), MgCl<sub>2</sub> concentration specific to the primers and 0.01 U Taq polymerase. Reaction condition were as follows: 94°C, 3mins; followed by cycles of 94°C, 30 seconds (secs); gene-specific annealing temperature for 1min; 72°C, 1min; with final elongation at 72°C, 5mins. RT-qPCRs were performed using 1× LightCycler 480 SYBR Green I Master (Roche), 0.5 μM primers (Table S3) and 1μl cDNA. Reactions were run on the LightCycler 480 II (Roche), with an initial incubation step of 95°C, 10mins; followed by 50 cycles of 95°C, 10secs; 60°C, 10secs and 72°C, 10 secs. Expression was normalised to *HPRT*, and relative expression was determined using the  $\Delta\Delta C_T$  method.

Array work was carried out essentially as previously described (Mackin et al., 2018; O'Neill et al., 2018): briefly, total RNA was extracted from each cell line growing in log phase in triplicate, including at least one biological replicate, and was assessed for integrity and quantity using a SpectroStar (BMG Labtech, Aylesbury, UK) and bioanalyser (Agilent Technologies, Cheshire, UK) prior to sending to Cambridge Analytical Services for linear amplification using the Illumina TotalPrep RNA Amplification Kit (Life Technologies/ThermoFisher, Paisley, UK) followed by hybridization to the HumanHT-12 v4 Expression BeadChip.

### **Bioinformatics and statistical analysis**

Output files in IDAT format were processed and bioinformatic analysis was carried out using the RnBeads (Assenov et al., 2014) methylation analysis package (v1.0.0) as previously described (Mackin et al., 2018). In order to map CpG sites showing highly reproducible changes (FDR < 0.05) against the locations of RefSeq genes on the UCSC genome browser (Karolchik et al., 2003) for each cell line, we employed a bespoke workflow termed CandiMeth

(Thursby, Irwin, Walsh, submitted) on the Galaxy platform (Giardine et al., 2005). Absolute  $\beta$  levels were used to measure median methylation across genes of interest using CandiMeth, with further statistical analyses in Statistical Package for the Social Sciences software (SPSS) version 22.0 (SPSS UK Ltd).

Agilent arrays GSE93142 and GSE93135 from SuperSeries GSE93136 from (Cai et al., 2017) were processed using the R package GEOquery (2.46.15), annotation package hgug4112a.db (3.8) and annotation table for Agilent-014850 Whole Human Genome Microarray 4x44K G4112F (Probe Name version) from GEO to obtain log<sub>2</sub> normalized fold changes (FC) per probe. Gene Ontology analysis through DAVID (Huang et al., 2009) was then computed using the top 500 genes with greater than 1.5 FC.

Statistical analysis for pyrosequencing and RT-qPCR data using Student's paired t-test employed Microsoft Excel (Microsoft Office Professional Plus 2016). Experiments were carried out at least in triplicate and included at least one biological replicate in all cases except Supp. Fig.1, one biological repeat only. Error bars on all graphs represent standard error of the mean (SEM), or in the case of HT12 array data, 95% confidence interval (CI), unless otherwise stated. Asterisks are used to represent probability scores as follows: \* $p < 0.05$ ; \*\* $p < 0.01$ ; \*\*\* $p < 0.001$  or n.s. not significant.

### **Protein analysis**

Cells growing in log phase were harvested for protein extraction using the protein extraction buffer (50 mM Tris-HCl, 150 mM NaCl, 1% Triton-X, 10% glycerol, 5 mM EDTA; all Sigma-Aldrich) and 0.5  $\mu$ l protease inhibitor mix (Sigma-Aldrich). Western blotting was carried out essentially as before (O'Neill et al., 2018): in brief, 30 $\mu$ g protein was denatured at 70°C in the presence of 5 $\mu$ l 4 $\times$  LDS sample buffer and 2 $\mu$ l 10 $\times$  reducing agent (Invitrogen) in a total volume of 20 $\mu$ l nuclease-free water (Qiagen). Proteins were separated by SDS-PAGE and

electroblotted onto a nitrocellulose membrane (Thermo-Fisher Scientific), then blocked in 5% non-fat milk for either 1h at RT or overnight at 4°C. Membranes were incubated with the primary antibody overnight (Table S4) overnight at 4 C, followed by HRP-conjugated secondary antibody incubation at RT using ECL (Thermo-Fisher Scientific).

### **Generation of the UHRF1 PHD D334/E335AA mutant mice**

*B6D2F1* female mice (5 weeks old) were super-ovulated by intraperitoneal injection of pregnant mare's serum gonadotropin (PMSG, 5 IU). Mice were injected with human chorionic gonadotropin (hCG, 5 IU) 48h later and mated with *B6D2F1* males overnight. Zygotes were collected from the oviducts of female mice at embryonic day (E) 0.5. The cumulus cells were removed by incubation in 1% hyaluronidase/M2 medium before washing with fresh M2 medium and recovery for 6h at 37°C in a 5% CO<sub>2</sub> incubator. SpCas9 mRNA (100ng/μl), *Uhrf1*-crRNA (50ng/μl) and a 112-bp single-stranded oligodeoxynucleotide (ssODN, 10ng/μl) which was flanked by homologous arms corresponding to exon 7 of *Uhrf1* (see Table S8 for sequences) were mixed immediately before microinjection using a FemtoJet microinjector, set to  $P_c = 10-15$  hPa, and  $P_i = 40-50$  hPa. Successfully microinjected zygotes were incubated in KSOM at 37 °C in a 5% CO<sub>2</sub> incubator for 72h until they reached the blastocyst stage and transferred into the uteri of pseudopregnant *B6D2F1* females. To investigate CRISPR/Cas9-mediated mutation in the *Uhrf1* gene, genomic DNA was prepared from 3wk-old mouse tails. The genomic regions flanking the gRNA target were amplified by PCR using specific primers (Supplementary Table X). The PCR amplicons were ligated into the pClone007 vector and sequenced.

## References

- Almeida, L.G., Sakabe, N.J., de Oliveira, A.R., Silva, M.C.C., Mundstein, A.S., Cohen, T., Chen, Y.T., Chua, R., Gurung, S., Gnjatic, S., et al. (2009). CTdatabase: A knowledge-base of high-throughput and curated data on cancer-testis antigens. *Nucleic Acids Res.* *37*.
- Arita, K., Ariyoshi, M., Tochio, H., Nakamura, Y., and Shirakawa, M. (2008). Recognition of hemi-methylated DNA by the SRA protein UHRF1 by a base-flipping mechanism. *Nature*.
- Assenov, Y., Muller, F., Lutsik, P., Walter, J., Lengauer, T., and Bock, C. (2014). Comprehensive analysis of DNA methylation data with RnBeads. *Nat. Methods* *11*, 1138–1140.
- Barlow, D.P., Randle, B.J., and Burke, D.C. (1984). Interferon synthesis in the early post-implantation mouse embryo. *Differentiation* *27*, 229–235.
- Beard, C., Li, E., and Jaenisch, R. (1995). Loss of methylation activates Xist in somatic but not in embryonic cells. *Genes Dev* *9*, 2325–2334.
- Bibikova, M., Barnes, B., Tsan, C., Ho, V., Klotzle, B., Le, J.M., Delano, D., Zhang, L., Schroth, G.P., Gunderson, K.L., et al. (2011). High density DNA methylation array with single CpG site resolution. *Genomics* *98*, 288–295.
- Borgel, J., Guibert, S., Li, Y., Chiba, H., Schübeler, D., Sasaki, H., Forné, T., Weber, M., Schubeler, D., Sasaki, H., et al. (2010). Targets and dynamics of promoter DNA methylation during early mouse development. *Nat. Genet.* *42*, 1093–1100.
- Bostick, M., Kim, J.K., Esteve, P.-O., Clark, A., Pradhan, S., and Jacobsen, S.E. (2007). UHRF1 Plays a Role in Maintaining DNA Methylation in Mammalian Cells. *Science* (80-. ). *317*, 1760–1764.
- Bourc'his, D., and Bestor, T.H. (2004). Meiotic catastrophe and retrotransposon reactivation in male germ cells lacking Dnmt3L. *Nature* *431*, 96–99.
- Cai, Y., Tsai, H.C., Yen, R.W.C., Zhang, Y.W., Kong, X., Wang, W., Xia, L., and Baylin, S.B. (2017). Critical threshold levels of DNA methyltransferase 1 are required to maintain DNA methylation across the genome in human cancer cells. *Genome Res.*
- Chen, T., Hevi, S., Gay, F., Tsujimoto, N., He, T., Zhang, B., Ueda, Y., and Li, E. (2007). Complete inactivation of DNMT1 leads to mitotic catastrophe in human cancer cells. *Nat. Genet.* *39*, 391–396.
- Chernyavskaya, Y., Mudbhary, R., Zhang, C., Tokarz, D., Jacob, V., Gopinath, S., Sun, X., Wang, S., Magnani, E., Madakashira, B.P., et al. (2017). Loss of DNA methylation in zebrafish embryos activates retrotransposons to trigger antiviral signaling. *Development* *144*, 2925–2939.
- Chiappinelli, K.B., Strissel, P.L., Desrichard, A., Li, H., Henke, C., Akman, B., Hein, A., Rote, N.S., Cope, L.M., Snyder, A., et al. (2015). Inhibiting DNA Methylation Causes an Interferon Response in Cancer via dsRNA Including Endogenous Retroviruses. *Cell* *162*, 974–986.
- Cuellar, L., Herzner, A.M., Zhang, X., Goyal, Y., Watanabe, C., Friedman, B.A., Janakiraman, V., Durinck, S., Stinson, J., Arnott, D., et al. (2017). Silencing of retrotransposons by SET DB1 inhibits the interferon response in acute myeloid leukemia. *J.*



Cell Biol. 216, 3535–3549.

DaRosa, P.A., Harrison, J.S., Zelter, A., Davis, T.N., Brzovic, P., Kuhlman, B., and Klevit, R.E. (2018). A Bifunctional Role for the UHRF1 UBL Domain in the Control of Hemi-methylated DNA-Dependent Histone Ubiquitylation. *Mol. Cell*.

Donovan, J., Whitney, G., Rath, S., and Korennykh, A. (2015). Structural mechanism of sensing long dsRNA via a noncatalytic domain in human oligoadenylate synthetase 3. *Proc. Natl. Acad. Sci.*

Foster, B.M., Stolz, P., Mulholland, C.B., Montoya, A., Kramer, H., Bultmann, S., and Bartke, T. (2018). Critical Role of the UBL Domain in Stimulating the E3 Ubiquitin Ligase Activity of UHRF1 toward Chromatin. *Mol. Cell*.

Giardine, B., Riemer, C., Hardison, R.C., Burhans, R., Elnitski, L., Shah, P., Zhang, Y., Blankenberg, D., Albert, I., Taylor, J., et al. (2005). Galaxy: a platform for interactive large-scale genome analysis. *Genome Res.* 15, 1451–1455.

Hajkova, P., Erhardt, S., Lane, N., Haaf, T., El-Maarri, O., Reik, W., Walter, J., and Surani, M. (2002). Epigenetic reprogramming in mouse primordial germ cells. *Mech Dev* 117, 15.

Hashimoto, H., Horton, J.R., Zhang, X., Bostick, M., Jacobsen, S.E., and Cheng, X. (2008). The SRA domain of UHRF1 flips 5-methylcytosine out of the DNA helix. *Nature*.

Hata, K., Kusumi, M., Yokomine, T., Li, E., and Sasaki, H. (2006). Meiotic and epigenetic aberrations in Dnmt3L-deficient male germ cells. *Mol. Reprod. Dev.* 73, 116–122.

Hill, P.W.S., Leitch, H.G., Requena, C.E., Sun, Z., Amouroux, R., Roman-Trufero, M., Borkowska, M., Terragni, J., Vaisvila, R., Linnett, S., et al. (2018). Epigenetic reprogramming enables the transition from primordial germ cell to gonocyte. *Nature* 555, 392–396.

Huang, W., Sherman, B.T., and Lempicki, R.A. (2009). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.* 4, 44–57.

Ivashkiv, L.B., and Donlin, L.T. (2014). Regulation of type I interferon responses. *Nat. Rev. Immunol.* 14, 36–49.

Jahner, D., and Jaenisch, R. (1985). Retrovirus-induced de novo methylation of flanking host sequences correlates with gene inactivity. *Nature* 315, 594–597.

James, S.R., Link, P.A., and Karpf, A.R. (2006). Epigenetic regulation of X-linked cancer/germline antigen genes by DNMT1 and DNMT3b. *Oncogene* 25, 6975–6985.

Jensen, S., and Thomsen, A.R. (2012). Sensing of RNA Viruses: a Review of Innate Immune Receptors Involved in Recognizing RNA Virus Invasion. *J. Virol.* 86, 2900–2910.

Jones, P.A., Ohtani, H., Chakravarthy, A., and De Carvalho, D.D. (2019). Epigenetic therapy in immune-oncology. *Nat. Rev. Cancer* 19, 151–161.

Karimi, M.M., Goyal, P., Maksakova, I.A., Bilenky, M., Leung, D., Tang, J.X., Shinkai, Y., Mager, D.L., Jones, S., Hirst, M., et al. (2011). DNA methylation and SETDB1/H3K9me3 regulate predominantly distinct sets of genes, retroelements, and chimeric transcripts in mESCs. *Cell Stem Cell* 8, 676–687.

Karolchik, D., Baertsch, R., Diekhans, M., Furey, T.S., Hinrichs, A., Lu, Y.T., Roskin, K.M., Schwartz, M., Sugnet, C.W., Thomas, D.J., et al. (2003). The UCSC Genome Browser

Database. *Nucleic Acids Res* 31, 51-4.

Karpf, A.R. (2006). A potential role for epigenetic modulatory drugs in the enhancement of cancer/germ-line antigen vaccine efficacy. *Epigenetics* 1, 116–120.

Kassiotis, G., and Stoye, J.P. (2016). Immune responses to endogenous retroelements: Taking the bad with the good. *Nat. Rev. Immunol.* 16, 207–219.

Kawaguchi, M., Toyama, T., Kaneko, R., Hirayama, T., Kawamura, Y., and Yagi, T. (2008). Relationship between DNA methylation states and transcription of individual isoforms encoded by the protocadherin-?? gene cluster. *J. Biol. Chem.* 283, 12064–12075.

Kazachenka, A., Bertozzi, T.M., Sjoberg-Herrera, M.K., Walker, N., Gardner, J., Gunning, R., Pahita, E., Adams, S., Adams, D., and Ferguson-Smith, A.C. (2018). Identification, Characterization, and Heritability of Murine Metastable Epialleles: Implications for Non-genetic Inheritance. *Cell* 175, 1259-1271.e13.

Lees-Murdock, D.J., Felici, M. De, Walsh, C.P., De Felici, M., and Walsh, C.P. (2003). Methylation dynamics of repetitive DNA elements in the mouse germ cell lineage. *Genomics* 82, 230–237.

Lei, H., Oh, S.P., Okano, M., Juttermann, R., Goss, K.A., Jaenisch, R., and Li, E. (1996). De novo DNA cytosine methyltransferase activities in mouse embryonic stem cells. *Development* 122, 3195-205.

Li, E., Bestor, T.H., and Jaenisch, R. (1992). Targeted mutation of the DNA methyltransferase gene results in embryonic lethality. *Cell* 69, 915–926.

Li, E., Beard, C., and Jaenisch, R. (1993). Role for DNA methylation in genomic imprinting. *Nature* 366, 362–365.

Liao, J., Karnik, R., Gu, H., Ziller, M.J., Clement, K., Tsankov, A.M., Akopian, V., Gifford, C.A., Donaghey, J., Galonska, C., et al. (2015). Targeted disruption of DNMT1, DNMT3A and DNMT3B in human embryonic stem cells. *Nat. Genet.* 47, 469–478.

Loughery, J.E.E., Dunne, P.D.D., O’Neill, K.M.M., Meehan, R.R.R., McDaid, J.R.R., and Walsh, C.P.P. (2011). DNMT1 deficiency triggers mismatch repair defects in human cells through depletion of repair protein levels in a process involving the DNA damage response. *Hum. Mol. Genet.* 20, 3241–3255.

Mackin, S.-J., O’Neill, K.M., and Walsh, C.P. (2018). Comparison of DNMT1 inhibitors by methylome profiling identifies unique signature of 5-aza-2’deoxycytidine. *Epigenomics in press*.

Matsui, T., Leung, D., Miyashita, H., Maksakova, I.A., Miyachi, H., Kimura, H., Tachibana, M., Lorincz, M.C., and Shinkai, Y. (2010). Proviral silencing in embryonic stem cells requires the histone methyltransferase ESET. *Nature* 464, 927–931.

McClintock, T.S. (2010). Achieving singularity in mammalian odorant receptor gene choice. *Chem. Senses* 35, 447–457.

Mikkelsen, T.S., Ku, M., Jaffe, D.B., Issac, B., Lieberman, E., Giannoukos, G., Alvarez, P., Brockman, W., Kim, T.K., Koche, R.P., et al. (2007). Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature* 448, 553–560.

O’Neill, K.M., Irwin, R.E., Mackin, S.-J., Thursby, S.-J., Thakur, A., Bertens, C., Masala, L.,

- Loughery, J.E.P., McArt, D.G., and Walsh, C.P. (2018). Depletion of DNMT1 in differentiated human cells highlights key classes of sensitive genes and an interplay with polycomb repression. *Epigenetics and Chromatin* *11*.
- Okano, M., Bell, D.W., Haber, D.A., and Li, E. (1999). DNA methyltransferases Dnmt3a and Dnmt3b are essential for de novo methylation and mammalian development. *Cell* *99*, 247-57.
- Ouellette, M.M., McDaniel, L.D., Wright, W.E., Shay, J.W., and Schultz, R.A. (2000). The establishment of telomerase-immortalized cell lines representing human chromosome instability syndromes. *Hum. Mol. Genet.* *9*, 403–411.
- Ramesh, V., Bayam, E., Cernilogar, F.M., Bonapace, I.M., Schulze, M., Riemenschneider, M.J., Schotta, G., and Götz, M. (2016). Loss of Uhrf1 in neural stem cells leads to activation of retroviral elements and delayed neurodegeneration. *Genes Dev.* *30*, 2199–2212.
- Reddington, J.P., Perricone, S.M., Nestor, C.E., Reichmann, J., Youngson, N.A., Suzuki, M., Reinhardt, D., Dunican, D.S., Prendergast, J.G., Mjoseng, H., et al. (2013). Redistribution of H3K27me3 upon DNA hypomethylation results in de-repression of Polycomb target genes. *Genome Biol.* *14*, R25.
- Rothbart, S.B., Krajewski, K., Nady, N., Tempel, W., Xue, S., Badeaux, A.I., Barsyte-Lovejoy, D., Martinez, J.Y., Bedford, M.T., Fuchs, S.M., et al. (2012). Association of UHRF1 with methylated H3K9 directs the maintenance of DNA methylation. *Nat. Struct. Mol. Biol.* *19*, 1155–1160.
- Rothbart, S.B., Dickson, B.M., Ong, M.S., Krajewski, K., Houliston, S., Kireev, D.B., Arrowsmith, C.H., and Strahl, B.D. (2013). Multivalent histone engagement by the linked tandem tudor and PHD domains of UHRF1 is required for the epigenetic inheritance of DNA methylation. *Genes Dev.*
- Roulois, D., Loo Yau, H., Singhania, R., Wang, Y., Danesh, A., Shen, S.Y., Han, H., Liang, G., Jones, P.A., Pugh, T.J., et al. (2015). DNA-Demethylating Agents Target Colorectal Cancer Cells by Inducing Viral Mimicry by Endogenous Transcripts. *Cell* *162*, 961–973.
- Rowe, H.M., Friedli, M., Offner, S., Verp, S., Mesnard, D., Marquis, J., Aktas, T., and Trono, D. (2013). De novo DNA methylation of endogenous retroviruses is shaped by KRAB-ZFPs/KAP1 and ESET. *Development* *140*, 519–529.
- Samlowski, W.E., Leachman, S.A., Wade, M., Cassidy, P., Porter-Gill, P., Busby, L., Wheeler, R., Boucher, K., Fitzpatrick, F., Jones, D.A., et al. (2005). Evaluation of a 7-day continuous intravenous infusion of decitabine: inhibition of promoter-specific and global genomic DNA methylation. *J. Clin. Oncol.* *23*, 3897–3905.
- Sharif, J., Muto, M., Takebayashi, S., Suetake, I., Iwamatsu, A., Endo, T.A., Shinga, J., Mizutani-Koseki, Y., Toyoda, T., Okamura, K., et al. (2007). The SRA protein Np95 mediates epigenetic inheritance by recruiting Dnmt1 to methylated DNA. *Nature* *450*, 908–912.
- Sharif, J., Endo, T.A., Nakayama, M., Karimi, M.M., Shimada, M., Katsuyama, K., Goyal, P., Brind'Amour, J., Sun, M.-A., Sun, Z., et al. (2016). Activation of Endogenous Retroviruses in Dnmt1<sup>-/-</sup> ESCs Involves Disruption of SETDB1-Mediated Repression by NP95 Binding to Hemimethylated DNA. *Cell Stem Cell* *19*, 81–94.
- Simpson, A.J.G., Caballero, O.L., Jungbluth, A., Chen, Y.-T., and Old, L.J. (2005). Cancer/testis antigens, gametogenesis and cancer. *Nat. Rev. Cancer* *5*, 615–625.

- Smith, Z.D., and Meissner, A. (2013). DNA methylation: roles in mammalian development. *Nat. Rev.* *14*, 204–220.
- Stresemann, C., Brueckner, B., Musch, T., Stopper, H., and Lyko, F. (2006). Functional diversity of DNA methyltransferase inhibitors in human cancer cell lines. *Cancer Res.*
- Walsh, C.P.P., Chaillet, J.R.R., and Bestor, T.H.H. (1998). Transcription of IAP endogenous retroviruses is constrained by cytosine methylation. *Nat Genet* *20*, 116-7.
- Wang, S., Zhang, C., Hasson, D., Desai, A., SenBanerjee, S., Magnani, E., Ukomadu, C., Lujambio, A., Bernstein, E., and Sadler, K.C. (2019). Epigenetic Compensation Promotes Liver Regeneration. *Dev. Cell.*
- Weber, F., Wagner, V., Rasmussen, S.B., Hartmann, R., and Paludan, S.R. (2006). Double-Stranded RNA Is Produced by Positive-Strand RNA Viruses and DNA Viruses but Not in Detectable Amounts by Negative-Strand RNA Viruses. *J. Virol.* *80*, 5059–5064.
- Weber, M., Hellmann, I., Stadler, M.B., Ramos, L., Paabo, S., Rebhan, M., and Schubeler, D. (2007). Distribution, silencing potential and evolutionary impact of promoter DNA methylation in the human genome. *Nat. Genet.* *39*, 457–466.
- Wozniak, R.J., Klimecki, W.T., Lau, S.S., Feinstein, Y., and Futscher, B.W. (2007). 5-Aza-2'-deoxycytidine-mediated reductions in G9A histone methyltransferase and histone H3 K9 di-methylation levels are linked to tumor suppressor gene reactivation. *Oncogene* *26*, 77–90.
- Xie, S., Jakoncic, J., and Qian, C. (2012). UHRF1 double Tudor domain and the adjacent PHD finger Act together to recognize K9me3-containing histone H3 tail. *J. Mol. Biol.* *415*, 318–328.
- Zhao, Q., Zhang, J., Chen, R., Wang, L., Li, B., Cheng, H., Duan, X., Zhu, H., Wei, W., Li, J., et al. (2016). Dissecting the precise role of H3K9 methylation in crosstalk with DNA maintenance methylation in mammals. *Nat. Commun.*

## Figure Legends

### Figure 1. Loss of DNA methylation in UHRF1-deficient human fibroblast cells indirectly triggered a viral defence response

A) Schematic overview: normal fibroblasts (hTERT1604) were stably transfected with shRNA, then single cells expanded and evaluated for genome-wide transcription (HT12) and methylation (450K) using microarrays: hits were verified using pyrosequencing and RT-qPCR.

(B) Decreased *UHRF1* mRNA levels in the index UH4 line from HT12 array and RT-qPCR; error bars are 95% confidence interval and standard error of the mean, respectively, \*\*\*  $p < 0.001$ .

(C) Western blot showing levels of UHRF1 protein and GAPDH as a loading control

(D) Boxplot showing median and inter-quartile range of DNA methylation ( $\beta$ ) values across all probes from the 450K array in parental (WT) and UHRF1-depleted (UH4) cells; difference between samples significant at  $p < 2.2 \times 10^{-16}$  (Kruskal-Wallis test). The same cell type depleted of DNMT1 (d16) from our previously published study are shown for comparison.

(E) Number of probes on transcription (HT12) array showing greater than 2-fold changes at the indicated false discovery rate (FDR) values, comparing UHRF1- (UH4) and DNMT1-depleted (d16) cells

(F) Comparison of genes showing  $>10\%$  ( $0.1 \beta$ ) demethylation and those showing transcription changes  $>2$ -fold. Most upregulated genes are not demethylated, with exceptions being the cancer/testis antigens (CTA), some histones (HIST) and olfactory receptors (OR).

(G) Gene Ontology (GO) analysis on genes showing the most transcriptional derepression regardless of methylation. The top 10 categories from DAVID functional annotation clustering with enrichment scores are shown and are all related to an innate immune viral defense response. The x-axis represents group enrichment score, the geometric mean (in  $-\log$  scale) of the p-values of the individual subcategories. IFN, interferon; MHC, major histocompatibility complex; different subclasses of CTA are shown in brackets.

**Figure 2. Interferons and interferon-stimulated genes (ISG) involved in dsRNA detection were crucial to the cellular response to loss of UHRF1**

(A) Model for possible pathway triggering ISG response in UH4 based on GO analysis and literature. Signalling from the dsRNA sensors would converge on the MAVS complex if dsRNA was detected, leading to release of transcription factors (TFs) which trigger upregulation of both interferons (IFN) and ISG. Many ISG are also part of the pathway, leading to positive feedback (dashed arrows). Signalling to other cells can also occur (dashed red arrows). Inhibition of the pathway using siRNA against MAVS and the STAT inhibitor RUX are indicated. (B) Average fold change (FC) versus WT cells (set to 1) for viral defence genes from HT12 transcription array. (C) Many components of the signalling pathway are upregulated on the transcriptional array: actual FC for *IFI27* was 118. (D) Verification of selected array targets from different parts of the pathway using RT-qPCR. (E) An siRNA was used to knock down (KD) MAVS for the indicated period before assaying the named genes using RT-qPCR. (F) UH4 cells were treated with the JAK/STAT inhibitor RUX for the indicated time before carrying out RT-qPCR on the named targets. (G) Schematic (left) of experiment where media which had been exposed to UH4 cells was transferred to WT cells (WT+UH4 media), before assaying transcription by RT-qPCR (right). (H) Exposure of WT cells to dsRNA (poly I:C), but not dsDNA, results in up-regulation of the same ISG as seen in UH4, measured here by RT-qPCR. Error bars in all experiments represent standard error of the mean (SEM); \* $p < 0.05$ ; \*\* $p < 0.01$ , Student's T-test.

**Figure 3. Transposable elements were demethylated and transcriptionally activated by depletion of UHRF1**

(A) WT and UH4 cells were stained with J2 monoclonal antibody (red), used for detection of viral dsRNA; nuclei were counterstained with DAPI (blue). (B) Locations of primers used to

assay methylation (pyro) at the promoters and transcription (qPCR) relative to the transposable elements (TE) indicated. LTR, long terminal repeat; UTR, untranslated region; ORF, open reading frame. (C) RT-qPCR for the indicated *HERV* elements showing fold-change over WT. Asterisks as in Fig.1; n.s., not significant. (D) RT-qPCR for the *LINE-1* retrotransposon elements indicated. (E) Percentage DNA methylation (% meth) at the promoters of the indicated retrotransposons was determined using pyroassay. (F) Methylation across individual CG dinucleotides in the pyroassays indicated in E. \*\*\* $p < 0.001$ , Student's T-test. (G) Methylation values ( $\beta$ ) at all probes overlapping *HERV* elements from the 450K array, difference significant at  $p < 2.2 \times 10^{-16}$  (Kruskal-Wallis test). (H) Methylation at the indicated retroviral elements using probes from the 450K array; *LINE-1* difference significant at  $p < 2.2 \times 10^{-16}$  (Kruskal-Wallis test), others n.s. but with lower probe numbers. Error bars indicate SEM (C-F) or 95% confidence interval (G, H).

**Figure 4. Demethylation of TEs precedes reactivation and an interferon response in multiple cell types depleted of *UHRF1***

(A) A “hit-and-run” strategy was employed to establish timing of events: indicated cell types were exposed to small interfering RNA (siRNA) targeting *UHRF1*, or a scrambled control (SCR), for 48hrs, then fresh medium without siRNA added and cells allow to recover before sampling. (B) RT-qPCR showing initial loss of *UHRF1* is followed by recovery to above initial levels by 14 days (14D). Levels of transcript for a representative ERV (*HERV-H*) and ISG (*IFI27*) are shown. FC, fold-change; error bars and statistics as above. (C) Average methylation levels at representative TE as determined by pyroassay; error bars are SD. Methylation recovery by 21days, when transcription is already repressed, is still not significant. (D) Differences at the most highly-methylated individual CG sites for the *LINE-1* assay compared to WT, error bars are SD; 7d vs 21d n.s. except CG2,  $p < 0.05$  (E) RT-qPCR analysis of SKMEL melanoma cells treated as in B. (F) GO analysis of genes showing transcriptional upregulation

in HCT116 colon cancer cells following 90% KD of *UHRF1* (Cai et al, 2018). Enrichment scores etc as above. Cells were analysed 7days after adenoviral delivery of shRNA; raw data were obtained from GEO (GSE93136).

**Figure 5. Rescuing cells with UHRF1 could abrogate TE reactivation and interferon response without restoring DNA methylation**

(A) Schematic (top) showing rescue strategy: a plasmid containing a selectable marker and a full-length *UHRF1* cDNA lacking the 3'UTR targeted by the shRNA was transfected into UH4 cells and resistant colonies expanded. Western blot (bottom) detected the presence of the full-length FLAG-tagged UHRF1 in index daughter cell line WT10 (B) RT-qPCR of the indicated retrotransposons showing repression in the WT10 cell line derived from UH4 by introducing full-length cDNA; UH4 vs the original hTERT1604 (WT) is shown for comparison; error bars are SEM. (C) RT-qPCR showing repression of the ISG IFI27 in WT10 cells; error bars are not visible for WT10. (D) HT12 array results for WT10 confirm most ISG involved in the response pathway are down-regulated again with 1-2 exceptions (e.g. *BST2*, *GTSF1*). (E) 450K analysis indicated in contrast that genome-wide methylation ( $\beta$ ) levels were largely unchanged from the parental UH4 cells in the WT10 derivatives (F) Methylation across all *HERV* and all *LINE-1* elements assessed by 450K (G) Confirmation of array results using pyrosequencing; there was no significant gain in methylation in WT10 vs UH4 cells; error bars are SD.

**Figure 6. Knockdown cells cannot be rescued with UHRF1 proteins containing mutations known to affect H3K9me3 binding**

(A) Model showing possible differences between original WT cells and WT10 rescues, which may retain another chromatin mark in the absence of DNA methylation; TE are known to have repressive H3K9me3 chromatin marks added by the SETDB1 enzyme. (B) Western blot showing levels of H3K9me3 are indeed largely unchanged in hTERT1604 (WT), the UHRF1



KD (UH4) and its rescue line (WT10); actin was used as a loading control (ACTB). Levels in HeLa and SETDB1 knockdown (KD) cells are shown as positive and negative controls, respectively. (C) Model of the paired PHD-TTD domain of UHRF1 interacting with H3K9me<sub>3</sub>, showing the location of point mutations in the PHD (D334, E335) and TTD (Y188) domains previously shown to affect H3K9me<sub>3</sub> binding. (D) Schematic showing approach; UH4 cells were transfected with cDNA as before, but containing the point mutations, and colonies expanded. (E) Example western blot of rescued lines testing for FLAG-tagged proteins. (F) RT-qPCR for individual retrotransposons in the various rescued lines indicated; though variable, repression was generally seen in cells rescued with intact wild-type UHRF1 (WT10, WT18) but not in those containing point mutations in the PHD-TTD region (PHD1, PHD4, PHD10, TTD9); error bars represent SEM. (G) Rescuing UH4 cells with UHRF1 protein caused a shut-down of dsRNA production (WT18) as detected by J2 antibody (red), but not if the protein contained point mutations affecting H3K9me<sub>3</sub> binding (PHD1); nuclei were counterstained with DAPI (blue). (H) ISG response to dsRNA is still seen in UH4 cells rescued with point-mutated UHRF1 (PHD1, PHD4, PHD10, TTD9), but not in cells rescued with intact protein (WT10, WT18).

**Figure 7. A conserved role for the PHD domain of UHRF1 in TE methylation and repression in mouse**

(A) Strategy used to generate mutant mice containing point mutations in the PHD domain. Inter-strain hybrid zygotes were injected with: (i) a donor molecule for Homologous Recombination (HR) containing the mutations to match the D334A/E335A in human; (ii) a single-guide RNA for the targeted region in *Uhrfl* and (iii) an mRNA for Cas9. Round 1 (Rd.1) of injections gave no mice, so pregnant females from Rd.2 were sacrificed at e8.75 and e9.5 to examine embryos (see B). Further injections (Rd.3) resulted in a single heterozygous (HET) pup (#13), who was back-crossed once, then offspring inter-crossed (F1) to generate

homozygous (HOM) as well as heterozygous embryos. (B) Embryos from Rd.2 injections: at e8.75 homozygous mutant (-/-) mutants showed developmental delay (left); methylation differences by pyroassay at the 5'UTR of *IAP* elements did not reach significance (right) compared to WT (+/+). By e9.5 delay was more marked (left) and methylation differences significant (right). Heterozygotes were indistinguishable morphologically from WT at both e8.75 and e9.5 (e.g #1). (C) Overall methylation (% meth) at the *IAP* LTR is lower, as assessed by pyroassay, in homozygous mutant embryos derived by breeding as shown above from the HET founder in Rd.3; error bars represent SEM, n=number of embryos. (D) Methylation was significantly decreased across each CG site in the *IAP* LTR. (E) Methylation differences at *LINE-1* elements and in *IAP* 5'UTR were also significantly decreased in mutants. (F) RT-qPCR of individual ERVs, though variable, showed significant derepression overall in homozygous mutant (HOM) embryos compared to WT or HET embryos. (G) Heat map showing differences in expression of the transposable elements (TE), ISG and IFN (indicated at top) across embryos of the genotypes shown at left. Individual embryos and genes are indicated at right and bottom, respectively. *Uhrfl* mRNA levels are also indicated as a control.

## Supplementary Materials

### Figure S1. Independent UHRF1 KD lines show similar phenotypes

(A) RT-qPCR of *UHRF1* mRNA levels in U5 and U10 cell lines independently derived from hTERT1604 by integration of shRNA; error bars are SEM. (B) Western blot showing levels of UHRF1 protein in U5 and U10; the index cell line UH4 is shown for comparison; GAPDH is a loading control. (C) DNA methylation ( $\beta$ ) levels from 450K array analysis of the three *UHRF1* KD lines compared to WT; the *DNMT1* KD line d16 is shown for reference. (D) RT-qPCR showing derepression of the indicated ERV in all three lines; due to high variability not all differences reached significance. (E) Transcription of the indicated ISG genes was significant in all three lines.

**Table S1: Pyrosequencing primers**

<b>Human pyrosequencing primers</b>			
<b>Gene</b>	<b>Primer</b>	<b>Modification</b>	<b>Oligo sequence (5'-3')</b>
<i>HERV-FC2</i>	FW		TTTGGTTTTTTTTGTTAGGATTAGTGA
	RV	biotin	AAATCTCCTCCCCAATTTTAACACCA
	SEQ		TTTTTGTTAGGATTAGTGAATT
<i>HERV-H</i>	FW		AGGGTTTGTGTGAGTAATAAAGTT
	RV	biotin	ACTCCTACCCCCCAAAAACAAACT
	SEQ		AGTTTTTAATTATTTGGGTGT
<i>LINE-1</i>	FW		GGGAGGAGTTAAGATGGT
	RV	biotin	ATAAACCCCATACCTCAA
	SEQ		GGGAGGAGTTGGATGGT
<b>Mouse pyrosequencing primers</b>			
<b>Gene</b>	<b>Primer</b>	<b>Modification</b>	<b>Oligo sequence (5'-3')</b>
<i>Iap 5-UTR</i>	FW		GGGTTGTAGTTAATTAGGGAGTGATA
	RV	biotin	ACAATTAAATCCTTCTTAACAATCTACTT
	SEQ		ATTTTGGTTTGTGTGT
<i>Iap LTR</i>	FW	biotin	GGTTTTGGAATGAGGGATTTT
	RV		CTCTACTCCATATACTCTACCTTC
	SEQ		ATACTCTACCTTCCCC
<i>Line-1</i>	FW		GTAGAAGTATAGAGGGGTTGAGGTA
	RV	biotin	ACAATTCCCAAATAATACAAACTCT
	SEQ		AGTATTTTGTGTGGGT

**Table S2: RT-PCR primers**

<b>Gene</b>	<b>Primer</b>	<b>Oligo sequence (5'-3')</b>
<i>ACT-B</i>	FW	GGACTTCGAGCAAGAGATGG
	RV	AGCACTGTGTTGGCGTACAG
<i>UHRF1</i>	FW	TGAGGACATGTGGGATGAGA
	RV	GTCCTGGAGTTCATCTGGA

**Table S3: RT-qPCR primers**

<b>Human RT-qPCR primers</b>		
<b>Gene</b>	<b>Primer</b>	<b>Oligo sequence (5'-3')</b>
<i>IFI27</i>	FW	TCTGTCCACCCTCTGCTTCT
	RV	GGCATGGTTCTCTTCTCTGC
<i>OAS2</i>	FW	AGGAAGGTAGCGCATCTTGA
	RV	CACCTCTGTCCGACTTGGTT
<i>STAT1</i>	FW	CCCACTCTGATCAACTTTTGC
	RV	GGCCTGTTGAAGATGCTTGT
<i>ISG15</i>	FW	ACCTACGAGGTACGGCTGAC
	RV	GGTGGAGGCCCTTAGCTC
<i>HERV-FC2</i>	FW	TTTCCACCGCTGGTAATAG
	RV	AGGCTAAGGATTCGGCTGAG
<i>HERV-H</i>	FW	TTCACTCCATCCTTGGCTAT
	RV	CGTCGAGTATCTACGAGCAAT
<i>HERV-W1</i>	FW	AAGAATCCCTAAGCCTAGCTGG
	RV	GCCTAATTAGCATTTTAGTGAGCTC
<i>LIPBA</i>	FW	CTTTGCAGACACTCCCCAGT
	RV	GGTCTAGCCACCCAGCAG
<i>LIP1</i>	FW	TGCCCTAAAAGAGCTCCTGA
	RV	TGTTTTTGCAGTGGCTGGTA
<i>UHRF1</i>	FW	TGAGGACATGTGGGATGAGA
	RV	GTCCCTGGAGTTCATCTGGA
<i>HPRT</i>	FW	AGCCCTGGCGTCGTGATTAGT
	RV	CCCGTTGAGCACACAGAGGCCTA
<b>Mouse RT-qPCR primers</b>		
<i>Gapdh</i>	FW	CAACTACATGGTCTACATGTTC
	RV	CTCGCTCCTGGAAGATG
<i>IAPez-gag</i>	FW	CACGCTCCGGTAGAATACTTACAAAT
	RV	CCTGTCTAACTGCACCAAGGTAAAAT
<i>MusD</i>	FW	GAATATGTCTAATACGCTAGCCTTTCC
	RV	GTAATGTCTGCCCTAGTATCTTGTT
<i>Ifn-alpha</i>	FW	CTGCTGGCTGTGAGGACATA
	RV	AGGAAGAGAGGGCTCTCCAG

<b><i>ERV-K10C gag</i></b>	FW	ATGTGAGCTAGCTGTAAAGAAGGAC
	RV	CTCTCTGTTTCTGACATACTTTCCTGT
<b><i>ERV-K10C</i></b>	FW	CCAAATAGCCCTACCATATGTCAG
	RV	GTATACTTTCTTCTTCAGGTCCAC
<b><i>ERV-L gag</i></b>	FW	TTCTTCTAGACCTGTAACCAGACTCA
	RV	TCCTTAGTAGTGTAGCGAATTCCTC
<b><i>Line1-ORF2</i></b>	FW	GACATAGACTAACAACTGGCTACACAAAC
	RV	GGTAGTGTCTATCTTTTTTCTCTGAGATGAG
<b><i>Ifi27</i></b>	FW	TAACTGGTCCTCATGGCGTT
	RV	CCCCTTCGAACCAGCTAGAA
<b><i>Uhrf1</i></b>	FW	AAAACGCCCTGAGTTTTTCGC
	RV	CCGATGTACTCTCTCACGGC
<b><i>Isg15</i></b>	FW	AGCAATGGCCTGGGACCTAA
	RV	AGACCCAGACTGGAAAGGGT
<b><i>Ifit2</i></b>	FW	CTGGGGAAACTATGCTTGGGT
	RV	ACTCTCTCGTTTTGGTTCTTGG
<b><i>Irf7</i></b>	FW	CCCATCTTCGACTTCAGCAC
	RV	TGTAGTGTGGTGACCCTGC

**Table S4. Antibodies**

Target	Supplier	Cat. number	Raised in	Clonality	Dilution	Size (kDa)	Application
<b>Primary antibodies</b>							
<i>UHRF1</i>	SC	373750	Mouse	Mono	1:100	90	WB
<i>GAPDH</i>	CST	14C10	Rabbit	Mono	1:10000	36	WB
<i>FLAG</i>	SA	F1804	Mouse	Mono	1:1000	120	WB
<i>dsRNA</i>	SCICONS	10010200	Mouse	Mono	1:200	X	IF
<b>Secondary antibodies</b>							
<b>AR-IgG</b>	SC	sc-2004	Goat	Poly	1:10000	X	WB
<b>AM-IgG</b>	SA	A9044	Rabbit	Poly	1:5000	X	WB
<b>AM-IgG</b>	Invitrogen	A10036	Donkey	Mono	1:1000	X	IF

SC (Santa Cruz Biotechnology); CST (Cell Signalling Technologies); SA (Sigma-Aldrich); Mono (monoclonal); Poly (polyclonal); kDa (Kilodaltons).

**Table S5. shRNA**

Target	Supplier	Cat. number	Mature antisense sequence	Cell lines
<i>UHRF1</i>	Dharmacon	V3LHS_353720	TGACATTGCGCACCACCCT	U prefix (U5, U10)
<i>UHRF1</i>	Dharmacon	V3LHS_413692	AACGTTATATCTTTCTTGG	UH prefix (UH4, UH5)

**Table S6. siRNA – ON-TARGETplus Human - SMARTpool**

Target	Supplier	Cat. number
<i>KAP1</i>	Horizon Discovery	L-005046-00-0005
<i>SETDB1</i>	Horizon Discovery	L-020070-00-0005
<i>UHRF1</i>	Horizon Discovery	L-006977-00-0005
<i>MAVS</i>	Horizon Discovery	L-024237-00-0005



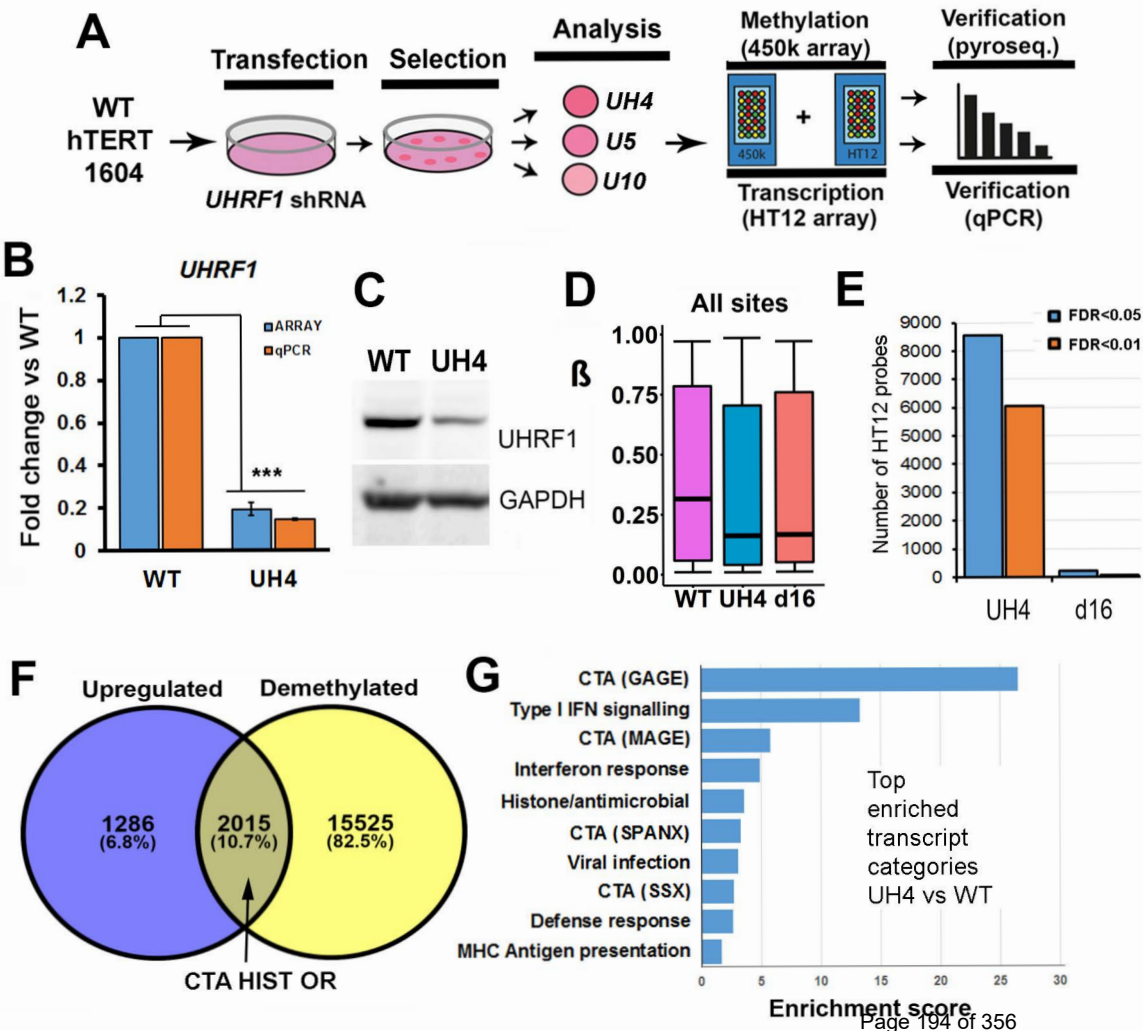
**Table S7. Pharmacological Inhibitors**

Target	Supplier	Cat. number
<b>Ruxolitinib (INCB018424)</b>	Selleckchem	S1378
<i>5-Aza-2'-deoxycytidine</i>	Sigma	A3656

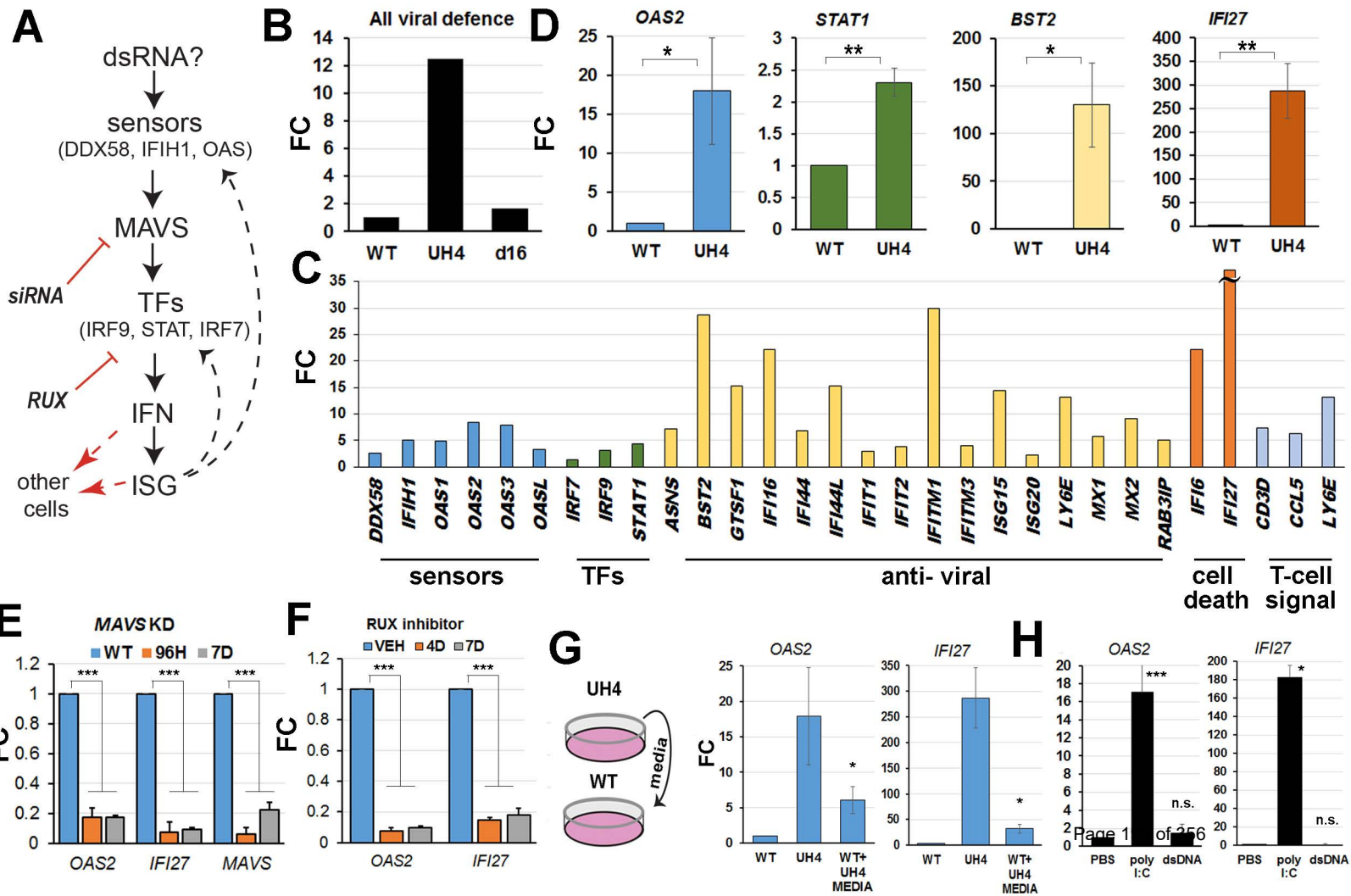
**Table S8. Mouse CRISPR guide/oligo sequences**

Target	Oligo sequence (5'-3') FW	Oligo sequence (5'-3') RV
<b>sgRNA sequence</b>		
<i>UHRF1</i> <i>exon 3</i>	GTTGTGTGATGAGTGTGACA	AAAAAAGCACCGACTCGGTG
<b>PHD mutant HDR oligo sequence</b>		
<i>UHRF1</i> <i>exon 7</i>	GTGCCTGCCATGTGTGTGGTGGGCGCGAGGCTCCTGAGAAACAGCTGTTGT GTGCTGCGTGCGATATGGCCTTCCACCTGTACTGCCTGAAGCCACCGCTCA CCTCTGTCCC	

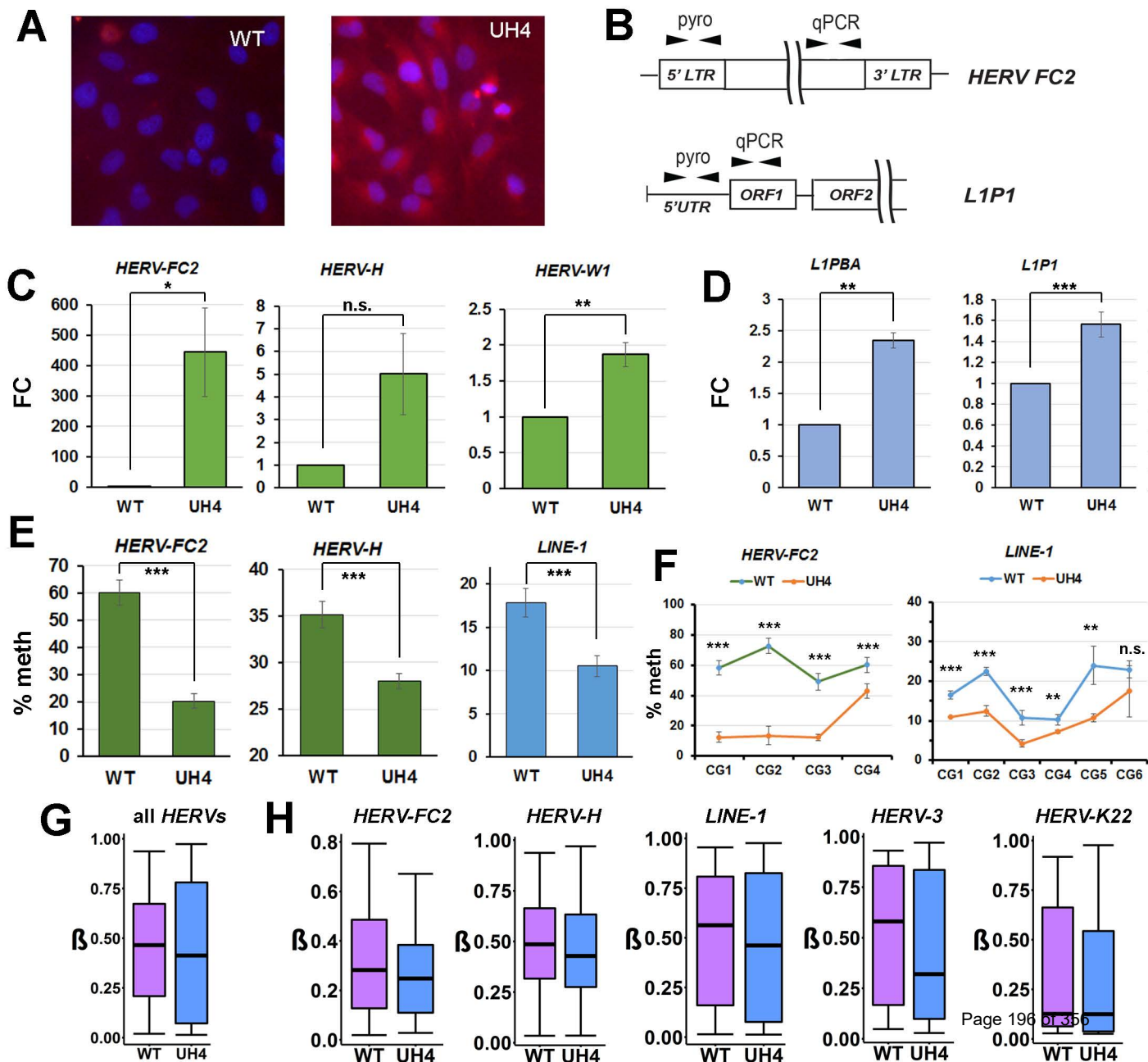
# Figure 1



# Figure 2

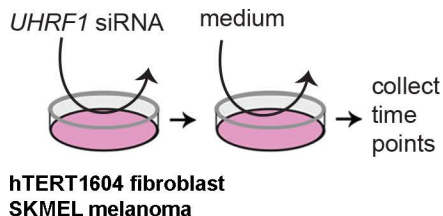


# Figure 3



# Figure 4

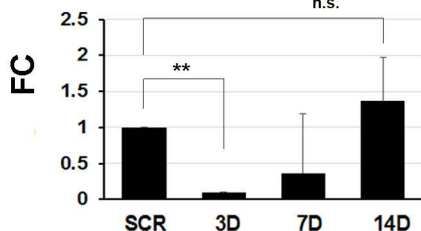
**A**



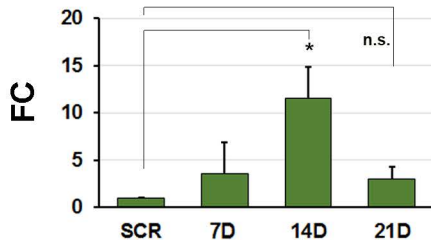
**B**

**hTERT1604 fibroblasts**

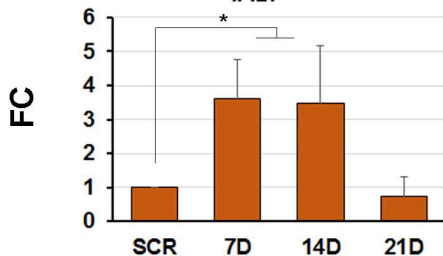
*UHRF1*



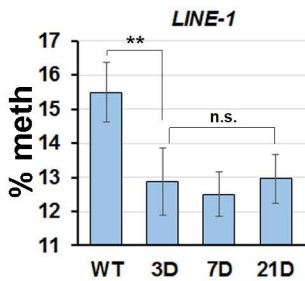
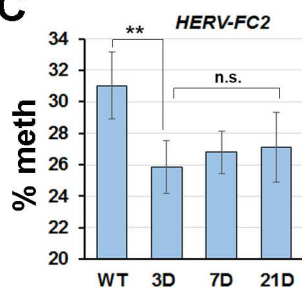
*HERV-H*



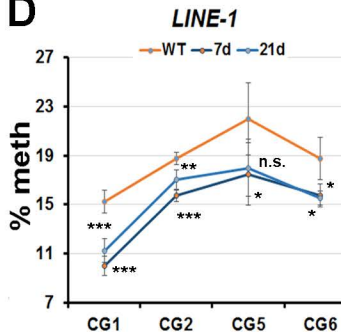
*IFI27*



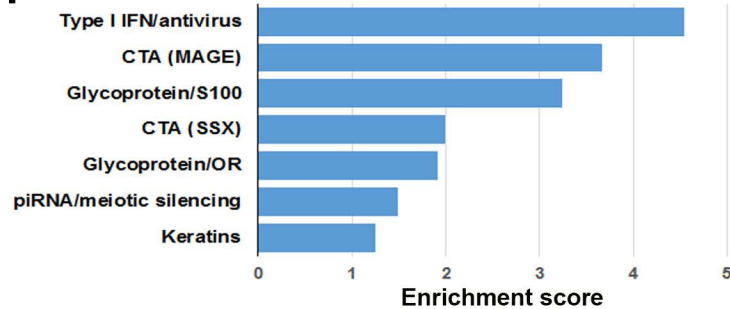
**C**



**D**



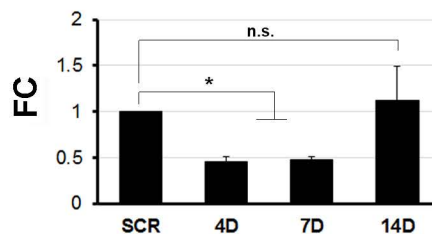
**HCT116 colon cancer**



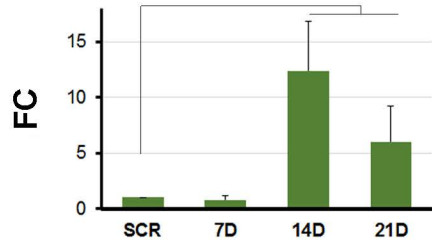
**SKMEL melanoma**

**E**

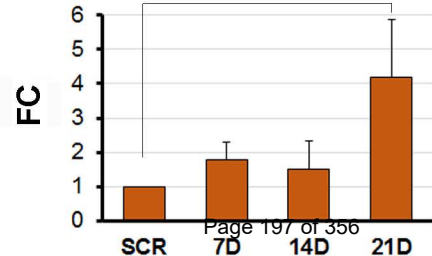
*UHRF1*



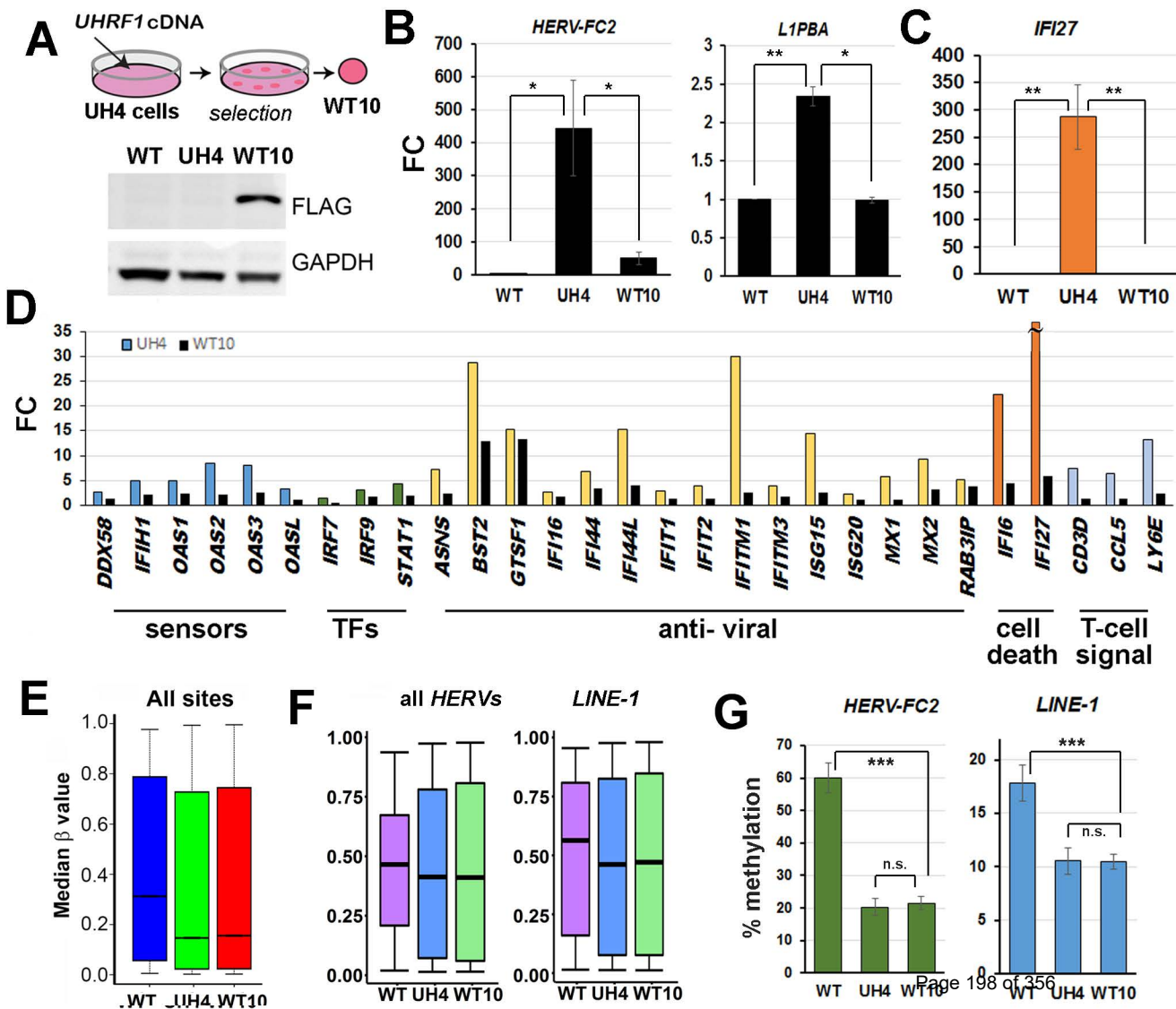
*HERV-H*



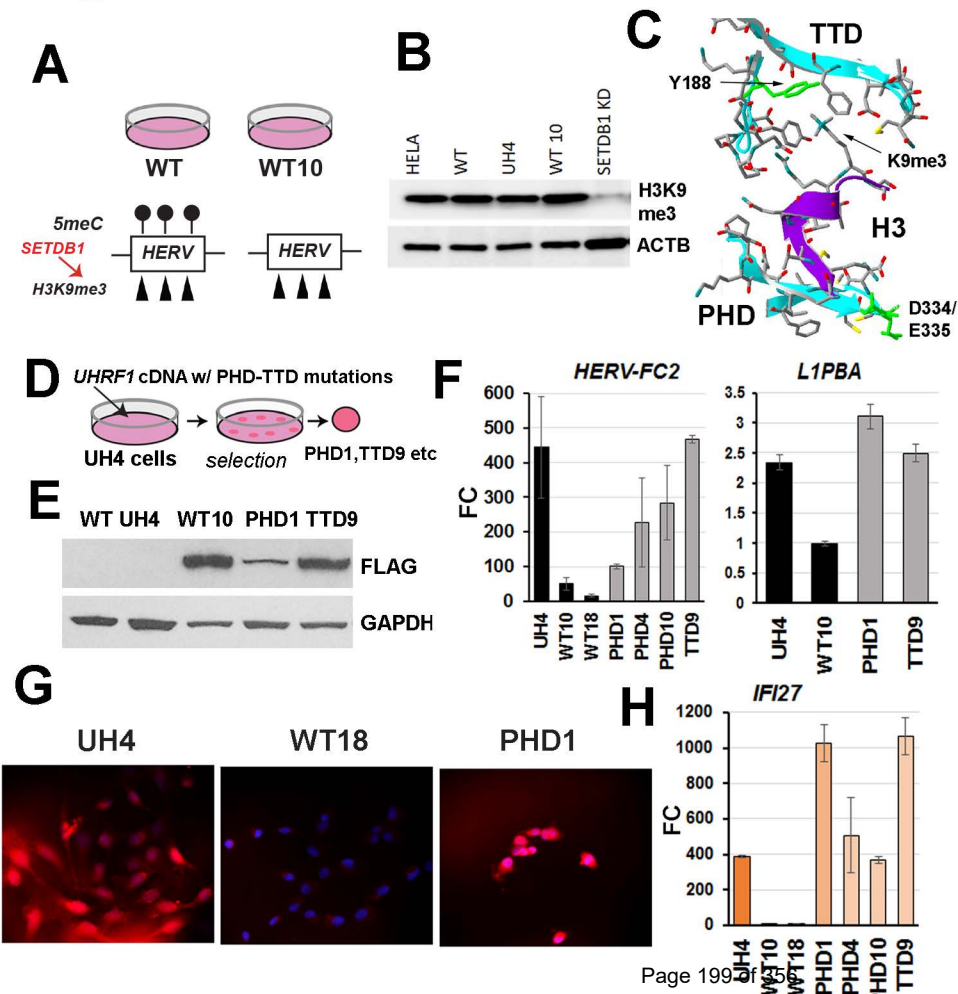
*IFI27*



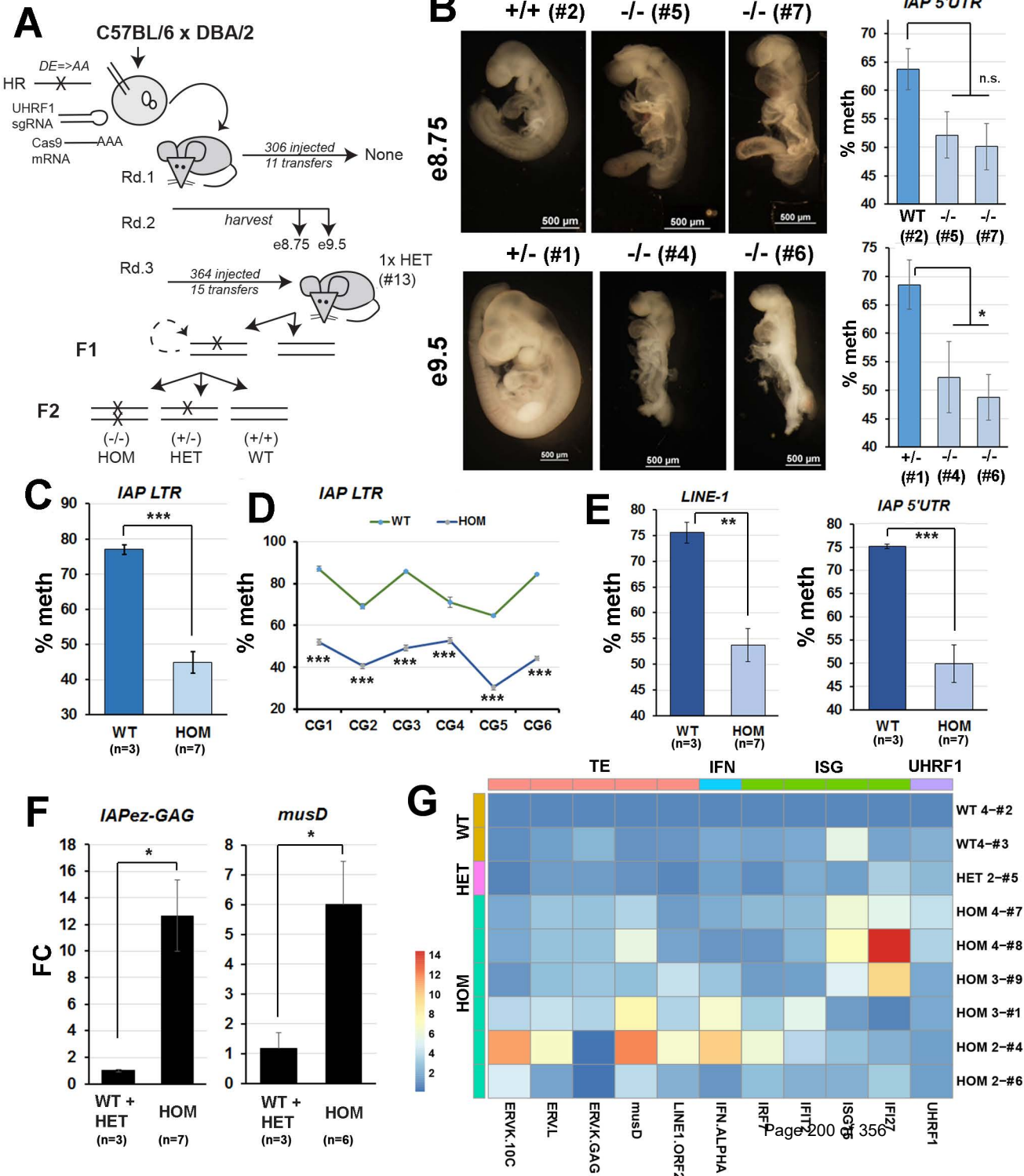
# Figure 5



# Figure 6

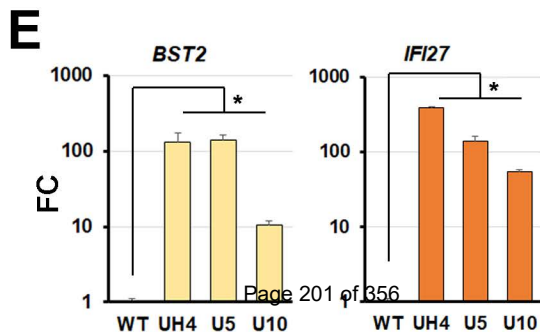
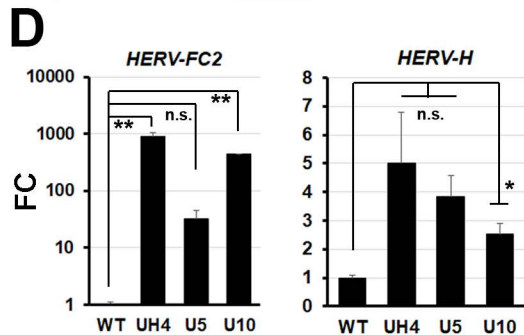
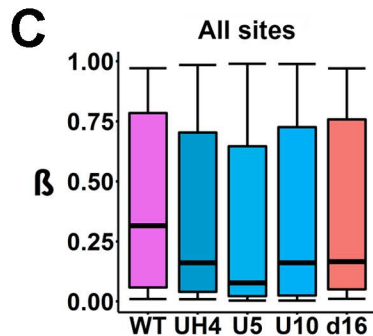
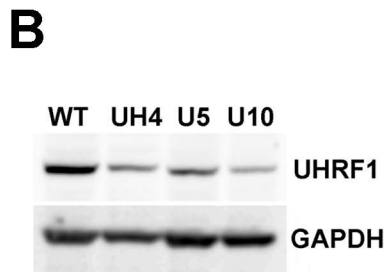
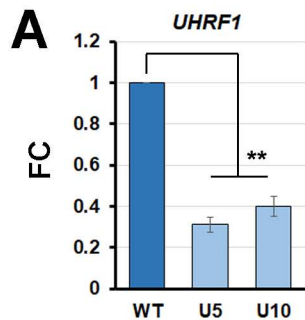


# Figure 7





# Suppl. Fig.1



## 4.0 PAPER-III

### **A Randomised controlled trial of folic acid intervention in pregnancy highlights a putative methylation-regulated control element at ZFP57**

Rachelle E. Irwin, Sara-Jayne Thursby, Miroslava Ondicova, Kristina Pentieva, Helene McNulty, Rebecca Richmond, Aoife Caffrey, Diane J. Lees-Murdock, Marian McLaughlin, Tony Cassidy, Matthew Suderman, Caroline L. Relton and Colum P. Walsh

The aims of this paper were to:

- Assess the effects on the DNA methylation of the offspring of mothers of the FASSTT trial
- Elucidate whether there are any gene class specific effects from this RCT
- Investigate whether the results obtained align with those of observational studies into folic acid supplementation throughout all of gestation

### **CONTRIBUTION**


For this paper, I independently conducted EPIC array analysis of the data in RnBeads and Limma, conducted tissue type correction and SVA analysis and carried these alterations through to differential analysis. I made absolute beta and delta beta tracks for UCSC genome browser from the results of the array. I computed the Manhattan plot of the suspected fDMR upstream of ZFP57 and assessed the methylation of selected imprints in placebo and treatment groups using the updated Galaxy pipeline, now being called CandiMeth. I also generated a QQ plot with the EPIC array data from this study to test for population stratification effects and compared and computed comparison statistics of our results to that of the AFAST trial in Aberdeen.

RESEARCH

Open Access



# A randomized controlled trial of folic acid intervention in pregnancy highlights a putative methylation-regulated control element at *ZFP57*

Rachelle E. Irwin<sup>1</sup>, Sara-Jayne Thursby<sup>1</sup>, Miroslava Ondiřová<sup>1</sup>, Kristina Pentieva<sup>2</sup>, Helene McNulty<sup>2</sup>, Rebecca C. Richmond<sup>4</sup>, Aoife Caffrey<sup>2</sup>, Diane J. Lees-Murdock<sup>1</sup>, Marian McLaughlin<sup>3</sup>, Tony Cassidy<sup>3</sup>, Matthew Suderman<sup>4</sup>, Caroline L. Relton<sup>4</sup> and Colum P. Walsh<sup>1\*</sup> 

## Abstract

**Background:** Maternal blood folate concentrations during pregnancy have been previously linked with DNA methylation patterns, but this has been done predominantly through observational studies. We showed recently in an epigenetic analysis of the first randomized controlled trial (RCT) of folic acid supplementation specifically in the second and third trimesters (the EpiFASST trial) that methylation at some imprinted genes was altered in cord blood samples in response to treatment. Here, we report on epigenome-wide screening using the Illumina EPIC array (~ 850,000 sites) in these same samples ( $n = 86$ ).

**Results:** The top-ranked differentially methylated promoter region (DMR) showed a gain in methylation with folic acid (FA) and was located upstream of the imprint regulator *ZFP57*. Differences in methylation in cord blood between placebo and folic acid treatment groups at this DMR were verified using pyrosequencing. The DMR also gains methylation in maternal blood in response to FA supplementation. We also found evidence of differential methylation at this region in an independent RCT cohort, the AFAST trial. By altering methylation at this region in two model systems in vitro, we further demonstrated that it was associated with *ZFP57* transcription levels.

**Conclusions:** These results strengthen the link between folic acid supplementation during later pregnancy and epigenetic changes and identify a novel mechanism for regulation of *ZFP57*. This trial was registered 15 May 2013 at [www.isrctn.com](http://www.isrctn.com) as ISRCTN19917787.

**Keywords:** Folic acid, DNA methylation, Cord blood, Offspring, Imprinting, *ZFP57*

## Background

Folate is an essential B vitamin required for viable embryonic and fetal development and as an important dietary constituent throughout life, fundamental in cellular biosynthesis and DNA methylation pathways [1, 2]. Folic acid (FA) is the oxidized, and more stable, synthetic form of folate which is exclusively found in supplements and fortified foods [3]. Well-established evidence from randomized

controlled trials [4, 5] has led to recommendations, in place globally, that women should consume 400  $\mu\text{g}/\text{d}$  FA from prior to conception until the end of the first trimester in order to protect against neural tube defects (NTDs) [6, 7]. Despite the identification of a relationship between maternal folate status and NTDs as early as 40 years ago, information on the mechanism behind the benefit of FA supplementation with respect to NTDs remains to be fully elucidated (reviewed in [8]), as does the relationship of FA, NTDs, and DNA methylation [9]. There is however little dispute in regards to the protective effect of folic acid supplementation before and in early pregnancy, which was proven in clinical trials to reduce NTDs by approximately

\* Correspondence: [cp.walsh@ulster.ac.uk](mailto:cp.walsh@ulster.ac.uk)

<sup>1</sup>Genomic Medicine Research Group, School of Biomedical Sciences, Ulster University, Coleraine BT52 1SA, UK

Full list of author information is available at the end of the article



© The Author(s). 2019 Open Access This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated.

70% [10]. Furthermore, there remains a lack of evidence as to whether it is beneficial to mother and/or child to continue this supplementation throughout the entire pregnancy [11, 12]. FA supplementation during pregnancy has been associated with health benefits such as reduced risk of low birth weight [13], language delay [14], autism [15] and reduced risk of psychosis and other pediatric problems [16, 17]. In addition, observational studies have indicated that FA supplement use by mothers during pregnancy is associated with better cognitive health and brain development in the child [14, 18, 19], possibly related to the fact that there is a brain growth spurt at the end of the second trimester [20, 21]. However, there may also be potential adverse effects from excess folate in later pregnancy, an aspect which would also benefit from further exploration [12].

At a molecular level, there is some evidence in human that epigenetic changes could be the mechanism underpinning some of the effects of folate, both in the first trimester [8] in the prevention of NTDs, and also in the second and third trimester, as reviewed elsewhere [2]. Folate is essential for the production of S-adenosylmethionine (SAM), which provides the methyl group to the DNA methyltransferases (DNMTs), which carry out DNA methylation. DNA methylation is an essential means of maintaining transcriptional silencing at many different classes of genes when it occurs at promoter and enhancer elements, including endogenous retroviruses, genes on the inactive X, and imprinted genes [22] but can also facilitate transcription when occurring in the gene body [23–25]. DNA methylation is vital for embryonic survival and development, as mice carrying mutations in the DNA methyltransferases die in utero or shortly after birth [26, 27]. Some DNA methylation marks are inherited from the parents in the form of differential methylation on the paternal or maternal copy. This includes both the canonical imprinted loci, as well as some germline and neuronal genes [25, 28, 29], all of which methylation plays a direct role in controlling transcription. Both animal and human studies have indicated that the fetal epigenome is vulnerable to environmental exposures, such as methyl group availability from the maternal diet [30–36].

Imprinted genes are a paradigm for the transmission of epigenetic information across generations. Methylation differences between the paternal and maternal copies of imprinted genes are established in the germ cells and are known to be important for transcriptional regulation. Accordingly, inappropriate loss or gain of methylation at imprint control regions (ICR) is a diagnostic feature for several human disorders. These regions are protected from the wave of demethylation which occurs prior to implantation by several factors, such as PGC7/STELLA [37] and ZFP57, a Krueppel-associated box (KRAB) domain zinc finger protein [38, 39]. Several studies to date have centered on analyzing the effects of nutrition in particular

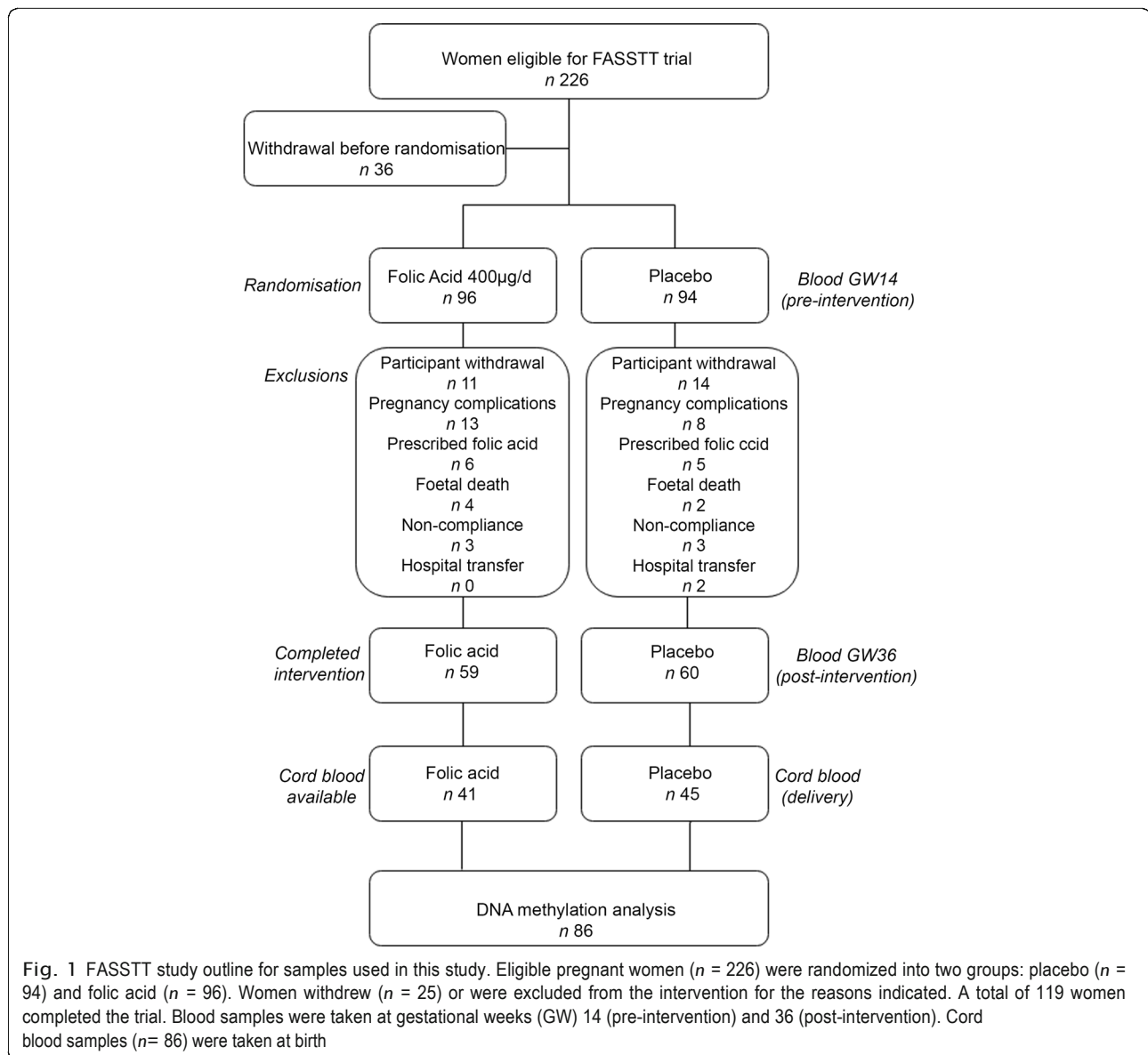
on imprinted genes [31–33, 40] and have shown that not only can altered diet result in an altered epigenotype, but it can also affect phenotype and predisposition to childhood and adulthood disease [41].

We have previously reported data from a randomized controlled trial of Folic Acid Supplementation in the Second and Third Trimester (The FASSTT Trial; ISRCTN19917787) where we found supplementation led to significant protection against folate depletion in mothers and offspring [42] and more recently that this led to differences in DNA methylation at some imprinted loci by using a candidate gene approach [43]. Here, we used the Infinium Methylation EPIC Beadchip Array to profile genome-wide DNA methylation levels in cord blood in an unbiased screen for regions susceptible to DNA methylation changes in response to altered FA levels. We report here that the top candidate region affected is a differentially methylated region (DMR) upstream of the gene encoding ZFP57. We verified our finding using pyrosequencing in cord blood and also show that the region responds to FA supplementation in maternal blood. Additionally, we confirm that altering methylation results in changes in ZFP57 transcription.

## Results

### Maternal FA supplementation significantly improves folate status in mother and baby

For the current analysis, the same 86 cord blood samples from the FASSTT trial (outlined in Fig. 1) which had been analyzed previously for candidate gene methylation [43] were used: a summary of the most pertinent characteristics are given in Table 1 for convenience. At baseline (gestational week 14 (GW14)), there were no detectable differences between the treatment and placebo groups in maternal characteristics, dietary folate intakes, serum or red blood cell (RBC) folate concentrations, or in *MTHFR* status, as expected following randomization. There were also no significant differences in neonatal characteristics such as weight, length, and head circumference (Table 1). However, as a result of treatment with FA during trimesters 2 and 3, maternal serum and RBC folate became significantly different between placebo and treated group, as previously reported from this trial. The normal decline in maternal folate biomarkers previously reported from observational studies during pregnancy is mirrored in the placebo group where serum folate decreased from 48.8 to 23.6 nmol/L between GW14 and GW36 (Table 1). FA supplementation served to protect the mothers in the treatment group, where folate concentrations remained stable over the course of pregnancy (i.e., serum folate 45.8 nmol/L at GW14 and 46.5 nmol/L at GW36). Cord serum and RBC folate concentrations were also



**Fig. 1** FASSTT study outline for samples used in this study. Eligible pregnant women ( $n = 226$ ) were randomized into two groups: placebo ( $n = 94$ ) and folic acid ( $n = 96$ ). Women withdrew ( $n = 25$ ) or were excluded from the intervention for the reasons indicated. A total of 119 women completed the trial. Blood samples were taken at gestational weeks (GW) 14 (pre-intervention) and 36 (post-intervention). Cord blood samples ( $n = 86$ ) were taken at birth

significantly higher in infants of the mothers supplemented with FA compared with those from the placebo mothers (Table 1). RBC folate concentrations in mothers and offspring were strongly correlated ( $r = 0.619$ ;  $p < 0.001$ , Additional file 1: Figure S1).

#### Widespread alterations to DNA methylation levels in cord blood in response to late gestation maternal FA supplementation

DNA was purified from cord blood and quantified prior to bisulfite conversion and hybridization to the Infinium Methylation EPIC Beadchip Array, which covers more than 850,000 CpG sites distributed across the genome. Methylation values are expressed as a decimal value  $\beta$  between 0.0 (no methylation) and 1.0

(fully methylated). Data were analyzed and visualized using the RnBeads package in RStudio (see methods section). As a control, a quantile-quantile (QQ) plot of observed versus expected chi-squared values was generated and showed no evidence of population substructure effects (Additional file 2: Figure S2). Figure 2a is a scatterplot showing mean  $\beta$  value for each CpG site analyzed in treated versus placebo samples. Overall, methylation at individual CpG remains closely correlated ( $\rho = 0.998$ ) between the two groups as expected, with most sites falling along the diagonal. Sites which differed in methylation between placebo and treatment groups were automatically ranked by RnBeads, which uses a combination of the change in mean methylation, the quotient of mean methylation

**Table 1** General characteristics of participants from the EpiFASSTT trial

Characteristic	Placebo (n=45)		Folic acid (n=41)		P value
	N = 45		N = 41		
	Mean	SD	Mean	SD	
<b>Maternal characteristics (GW14)</b>					
Age (years)	28.9	3.5	29.4	3.9	0.513
BMI (kg/m <sup>2</sup> )	25.2	3.9	24.9	4.6	0.768
Smoker <i>n</i> (%)	8 (18)		6 (15)		0.693
Alcohol <i>n</i> (%)	3 (7)		1 (2)		0.618
Parity ( <i>n</i> )	1 (1.1)		1 (1.0)		0.915
MTHFR 677TT genotype <i>n</i> (%)	5 (11)		2 (5)		0.291
<b>Dietary intakes</b>					
Energy (MJ/d)	8.170	1.717	7.732	1.595	0.280
Dietary folate equivalents (μg/d)	364	172	387	152	0.582
Vitamin B12 (μg/d)	4.1	1.9	3.9	1.8	0.791
<b>Neonatal characteristics</b>					
Gestational age (weeks)	40.1	1.3	40.0	1.1	0.540
Sex, male <i>n</i> (%)	22 (49)		22 (54)		0.659
Birth weight (g)	3610	475	3557	465	0.601
Birth length (cm)	51.5	2.6	51.1	2.2	0.499
Head circumference (cm)	34.9	1.2	34.8	1.4	0.907
Apgar score at 5 min	8.4	0.4	9.0	0.3	0.220
Caesarian <i>n</i> (%)	11 (24)		10 (24)		0.995
<b>B-vitamin biomarkers</b>					
<b>Maternal pre-intervention (GW14)</b>					
Serum folate (nmol/L)	48.8	19.8	45.8	19.5	0.469
RBC folate (nmol/L)	1185	765	1181	649	0.978
Serum B12 (pmol/L)	224	79	217	79	0.601
<b>Maternal post-intervention (GW36)</b>					
Serum folate (nmol/L)	23.6	17.9	46.5	24.8	< 0.001*
RBC folate (nmol/L)	991	404	1556	658	< 0.001*
Serum B12 (pmol/L)	168	51	157	60	0.229
<b>Cord blood</b>					
Serum folate (nmol/L)	68.3	24.8	91.7	36.7	0.004*
RBC folate (nmol/L)	1518	597	1877	701	0.024*
Serum B12 (pmol/L)	276	155	251	107	0.776

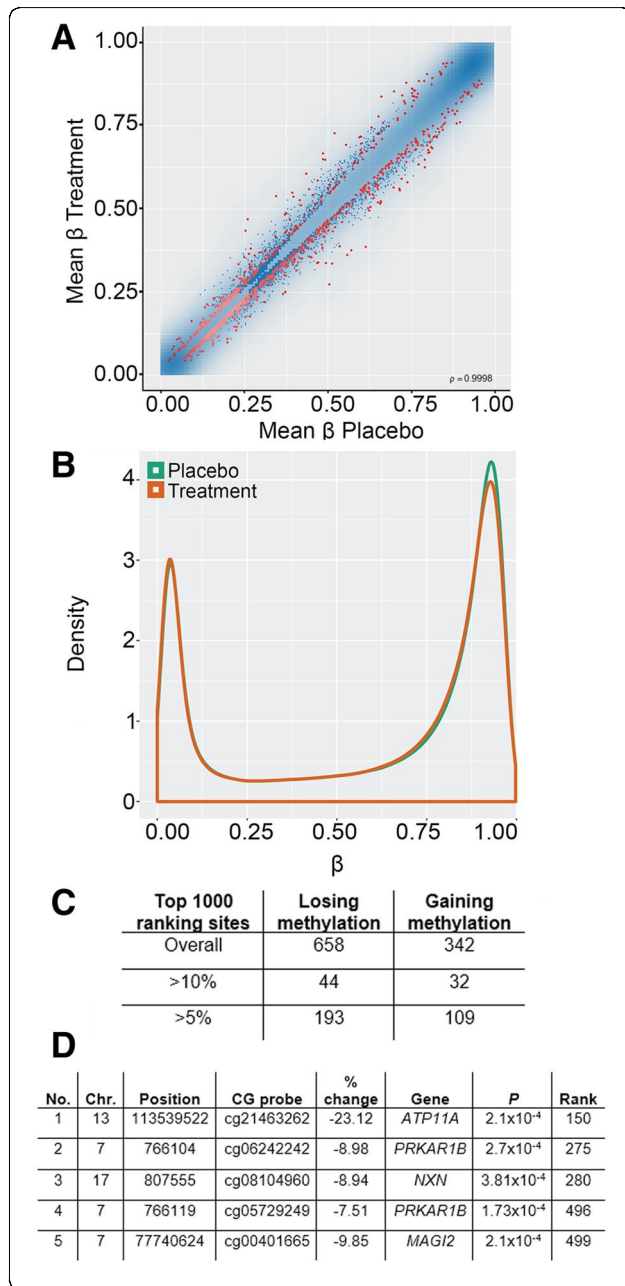
Statistical comparisons by independent *t* test (continuous variables) or  $\chi^2$  test (categorical variables)

GW gestational week, BMI body mass index, RBC red blood cell

\**p* < 0.05

and the combined *p* value, and the 1000 top-ranking sites are highlighted in red in Fig. 2a. This metric was developed to take into account not only *p* value but the magnitude of the change in methylation and in our experience is a more reliable indicator of biologically meaningful differences than *p* value alone. Sites falling along either side of the diagonal, representing gains and losses in methylation after treatment, can both be seen, with a tendency to greater numbers of sites losing. Consistent with this, a

methylation density distribution plot shows that after treatment there was a clear decrease in the numbers of sites in the top quartile for methylation ( $\beta = 0.75-1.00$ ; Fig. 2b). Taking the top 1000 ranking sites overall, approximately 2/3 (*n* = 658) lost and 1/3 (*n* = 342) gained methylation (Fig. 2c). However, the magnitude of these changes was generally modest, with only 302 (193 + 109) losing or gaining more than 5% methylation, the minimum change which we could potentially



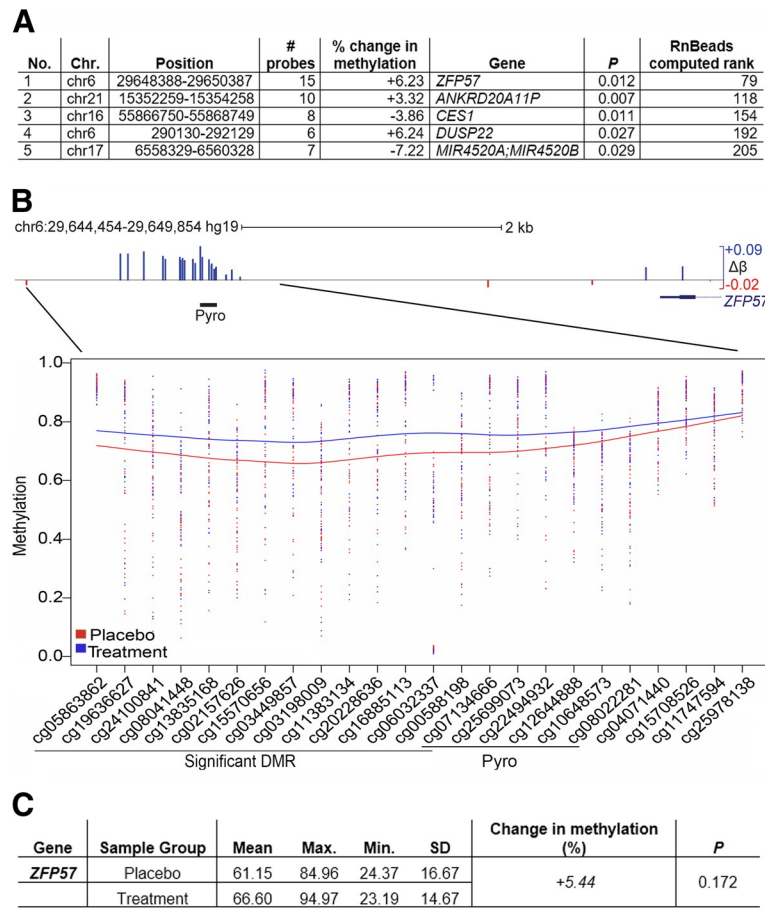
**Fig. 2** Widespread alterations to DNA methylation levels in cord blood in response to late gestation maternal folic acid supplementation. **a** Scatterplot comparing mean methylation levels ( $\beta$  values 1 = 100%; 0 = 0% methylation) at individual probes in placebo and treated groups. The 1000 top-ranking sites between groups are highlighted in red;  $\rho$  = correlation value. **b** Probe methylation density plot comparing the distributions of methylation values per sample group. In the treatment group, there is a decrease in the number of fully methylated sites ( $\beta > 0.75$ ). **c** Split in top 1000 ranking sites losing or gaining methylation overall. Also shown are numbers of sites showing changes greater than 5% or 10%. **d** Top 5 differentially methylated sites overall, sorted by combined rank, the value being computed as the maximum (i.e., worst) value among the mean quotient log, mean difference in methylation and  $p$  value ( $P$ ). No., number; Chr, chromosome; Position, coordinates in *hg19* human genome release; CG probe, identity number of the CpG probe on the EPIC array; % change, difference in mean  $\beta$  value expressed as %; Gene, nearest gene; P, probability (uncorrected); Rank, RnBeads computed ranking value (lowest being best)

verify using pyrosequencing, and only 76 sites losing or gaining more than 10% (Fig. 2c).

We examined the top-ranking sites as identified by RnBeads (Fig. 2d): of these, the CpG site in the *ATP11A* gene contained a single nucleotide polymorphism (SNP) missed by the quality control routines; the same was true of the CpG at the *MAGI2* gene. The presence of the SNPs at these CpGs leads to the erroneous appearance of a change in methylation, so these were discounted. Two of the other top-ranked sites were at the *PRKAR1B* locus, which encodes a regulatory subunit of cyclic AMP-dependent protein kinase A, and one was at *NXN*, a member of the thioredoxin superfamily; however, all three were listed as located in the respective gene body and so are less likely to contribute to transcriptional control. Nevertheless, to verify these, we used a second method utilizing commercial pyrosequencing methylation assays (pyroassays) designed to query the same CpGs. These reported smaller average differences in methylation between treated and placebo groups than seen with the array of 6.6% for cg08104960 at *NXN*, and 4.2% (cg06242242) and 2.2% for (cg05729249) for the sites at *PRKAR1B*: only the site at *NXN* was significant ( $p = 0.002$ ,  $t$  test).

#### Identification and verification of a differentially methylated region upstream of *ZFP57*

Given that single sites are more susceptible to confounders such as the presence of SNPs and show only moderate accuracy on verification, and to maximize our chances of finding biologically significant changes, we also looked for genomic intervals showing coherent alterations in methylation across multiple neighboring sites [44], rather than isolated CpGs. Figure 3a lists the top 5 differentially methylated regions (DMR) found at promoters, ordered by RnBeads ranking which is here computed by combining measures at adjacent sites



**Fig. 3** Top ranking promoter regions included imprint regulator gene *ZFP57*. **a** Top 5 differentially methylated regions (DMR) at promoters, sorted by combined RnBeads rank (smallest to largest) as for Fig. 2d above, except combining values across all the CpG sites in the DMR as detailed in the “Methods” section. Abbreviations as above except # probes, number of probes on EPIC array included in DMR. **b** Top: genome browser tracks showing the region around the DMR upstream of *ZFP57*, genomic coordinates in *hg19* human genome release, and scale as shown. EPIC array probes showing differential methylation (blue, gain; red, loss) are indicated, with size indicating the magnitude of change. The start of the *ZFP57* gene and the position of the pyrosequencing assay (Pyro) are also shown.  $\Delta\beta$ , mean difference in  $\beta$  value between placebo and FA-treated groups; maximum gain and loss also shown (+ 0.09  $\beta$  = 9% methylation). Bottom: Loess plot of  $\beta$  values across the region, with CpG identification numbers from array below; those forming the DMR defined by RnBeads are indicated, as well as sites analyzed by pyroassay. Each dot represents  $\beta$  value in an individual sample, with lines representing smoothed averages; color code is indicated at left. **c** Results of pyroassay covering the six sites indicated in **b**. Sample groups: cord blood DNA from placebo ( $n = 45$ ) and FA-treated ( $n = 41$ ). Mean, average of the individual means in that group; Max., largest of the mean methylation values in that group; Min, lowest mean in group; SD, standard deviation for the means; Change, difference in % methylation seen between groups; P, probability (Student’s  $t$  test)

using a linear hierarchical model as described in the “Methods” section: uncorrected  $p$  value and % change in methylation are also shown for comparison. For the top 5 regions, *ZFP57* was of particular interest and is dealt with below. Two others (*CES1*, a liver carboxylesterase, and *ANKRD20A11P*, a pseudogene) showed less than 5% change in methylation and so could not be verified: *DUSP22* which has a larger change is also a pseudogene. The last DMR is located at a microRNA cluster *MIR4520A/B* and loses approximately 7.22% overall in the treatment group, averaged over a number of well-spaced CpG. Due to pyrosequencing assay design constraints, we could only cover one site (cg08750459) from the array at

this locus but that site showed reasonable concordance (loss of 12.24% ( $p = 0.008$ ) in array and 9.45% ( $p = 0.006$ ) by pyroassay). The function of these microRNAs remains obscure however.

Of more interest in the context of this cohort was the highest ranking promoter DMR identified using RnBeads [45], which was located on chromosome 6, the closest gene being the known regulator of genomic imprinting *ZFP57*. The identified DMR consisted of 15 CpG sites and mapped approximately 3 kb upstream of the first exon of the gene, a region containing additional adjacent sites also gaining methylation. Figure 3b shows a genomic map of the first exon of *ZFP57* and



the upstream region, overlaid with a track showing the locations of EPIC probes and whether they gained or lost methylation. Also shown is a graph of averaged methylation values at the numbered CpG probes from the array in placebo and treatment groups, showing a clear difference in methylation extending beyond the DMR. To confirm these results using a second method, we designed a pyrosequencing methylation assay (pyroassay) to cover some of these CpG sites, as shown in Fig. 3b. Due to the CpG density of this region, thus difficulty in pyrosequencing primer design, our pyroassay is not directly overlapping all CpGs identified by RnBeads as the DMR but is inside the area showing methylation differences. We then carried out PCR and pyrosequencing for all the samples. The overall gain in methylation at the CpGs covered by the pyroassay ( $n = 6$ ) was very similar in magnitude and direction to that seen over the neighboring CpG by the array (+ 5.44% vs + 6.23%, respectively—Fig. 3a, c).

Demethylation of the upstream region was accompanied by increased *ZFP57* transcription

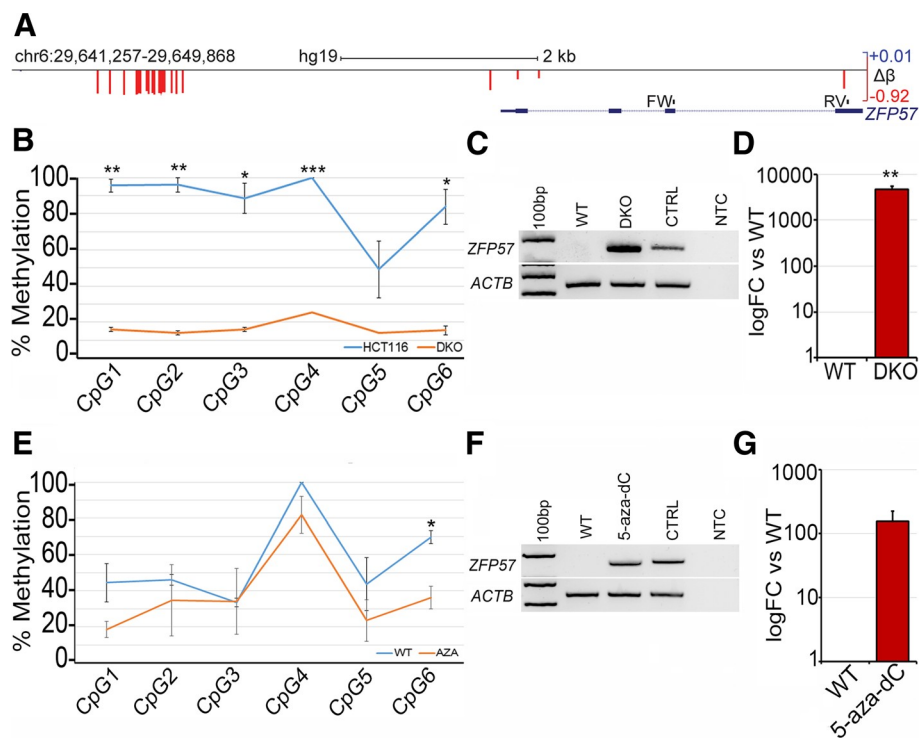
Having established that methylation differences at the upstream DMR are evident between FA-supplemented and placebo-treated controls, we wished to test mechanistically if such differences could impact on transcription from the downstream gene. To do this, we first used a well-established model, the paired colorectal cancer lines HCT116 and its derivative HCT116 DKO (double knockout), which carries mutations in two of the methyltransferase genes *DNMT1* and *DNMT3B* and is known to be hypomethylated at many loci [46]. Methylation array data available in-house showed differential methylation between the parental or wild type HCT116 (WT) and paired DKO cells at the same region upstream of *ZFP57* found in the FASSTT cohort, indicated by red colored bars whose height is proportional to the loss of methylation (Fig. 4a); this indicates that *DNMT1* and *DNMT3B* are required for methylation at this locus. We confirmed these results using our pyroassay, which showed > 80% methylation in WT HCT116 cells and a drop to < 20% in DKO cells ( $p = < 0.001$ ) (Fig. 4b).

To determine if methylation at this upstream region can regulate transcription at the *ZFP57* gene 3 kb downstream, we designed primers to cover part of the transcript as shown in Fig. 5a (FW/RV) and carried out reverse transcription on mRNA from the cells followed by polymerase chain reaction (RT-PCR). While minimal transcript could be detected in the HCT116 WT cells, which are heavily methylated, signal was readily apparent in the demethylated DKO cells (Fig. 4c). We confirmed this expression pattern quantitatively using RT-qPCR (Fig. 4d). While these results show that the gene can be de-repressed in response to loss of methylation, it is

normally not expressed in colon cells, from which HCT116 were derived, so we used the neuroblastoma cell line SH-SY5Y to test the effect of methylation changes on transcription in a neural cell type. *ZFP57* is normally transcribed in neural tissue as well as early embryo [47], but shows some methylation in the SH-SY5Y cells, which may be due to differences among neural cell types, or reflect accumulation of methylation during culture; however, these cells are likelier than HCT116 to contain neural-specific transcription factors. Here, we used a second method to perturb methylation, namely treatment with the DNA methyltransferase inhibitor 5'-aza-2'-deoxycytidine (5-aza-dC). Exposure of the cells to this small molecule inhibitor caused loss of methylation at the upstream region (Fig. 4e). RT-PCR confirmed that *ZFP57* was de-repressed upon treatment with 5-aza-dC (Fig. 4f). Quantification of mRNA levels with RT-qPCR again indicated a substantial increase in transcription from the gene in response to loss of methylation (Fig. 4g).

Greater variability at imprinted DMR in folate-treated samples

These results suggest that the increased methylation seen at the *ZFP57* upstream region will lead to decreased transcription. Since *ZFP57* plays a role in maintaining methylation specifically at imprinted genes, we examined methylation levels at these regions using data from the EPIC array. We used germline differentially methylated regions as defined by [48] and assessed average methylation across all probes which fell within these intervals. We excluded DMR which were flagged as acquiring methylation differences somatically and also germline DMR where methylation as assessed by the array fell outside the 35–65% methylation range defined as normal in that study. This left 15 imprinted germline DMR for which the median methylation level fell within the normal range in the placebo group (Additional file 3: Figure S3A). Comparing the samples from the folate supplemented group, only the maternally imprinted neuronatin gene (*NNAT*) showed a small but significant loss of methylation in the treatment group ( $p = 0.022$ , Mann-Whitney *U* test (MWU)) but there was no significant difference between placebo and treatment for any other DMR. However, it was notable that 11/15 DMR showed a significantly greater variability in methylation in treated participants ( $p = < 0.001$ , chi-squared test), which can be seen from the greater interquartile range (IQR—see Additional file 3: Figure S3A). Along with this greater variability in the treatment group, the median methylation levels trended lower than the placebo group for almost all imprinted genes (Additional file 3: Figure S3A). We repeated this analysis using



**Fig. 4** *ZFP57* upstream region is a methylation-dependent regulator of transcription at this locus. **a** Schematic as in Fig. 3 above but showing difference in methylation ( $\Delta\beta$ ) between HCT116 WT cells vs HCT116 DKO cells. The intron/exon structure and positions of the forward (FW) and reverse (RV) primers for RT-(q)PCR on the *ZFP57* gene are also shown. **b** Methylation levels at individual CpG covered by the pyrosequencing assay in WT (HCT116) and knockout (DKO) cells. Values are shown as mean  $\pm$  SD for each site: \* $p < 0.05$ ; \*\* $p < 0.01$ ; \*\*\* $p < 0.001$ . **c** RT-PCR showing upregulation using the primers indicated in **a**, key as above. CTRL, positive control (human reference total RNA); NTC, negative control (no template control); 100bp, size standards ladder; *ACTB*,  $\beta$ -actin loading control. **d** Confirmation of upregulation by RT-qPCR using the same primers, values normalized to *HPRT*; FC, fold change. **e** Methylation levels using pyroassay as in **B** but in 5-aza-dC treated SH-SY5Y cells (5-aza-dC), as compared to untreated (UT). **f** RT-PCR for 5-aza-dC treated cells from **e**. **g** RT-qPCR confirmation of *ZFP57* upregulation in 5-aza-dC-treated SH-SY5Y cells

imprinted DMR as defined by Court et al. [49], which defines slightly larger DMR based on an analysis of Illumina 450 K data. After applying similar criteria as above, this left 14 DMR suitable for comparison. Using these genomic intervals, again, only *NNAT* showed a significantly different level of methylation in treated samples ( $p = 0.022$ , MWU; Additional file 3: Figure S3B), although *PLAG1* was also close to significant ( $p = 0.072$ , MWU). Again, the IQR for the imprints showed greater variability in the treated than placebo groups ( $p < 0.001$ , chi-squared test) and medians tended to be lower in the FA-treated group (Additional file 3: Figure S3B).

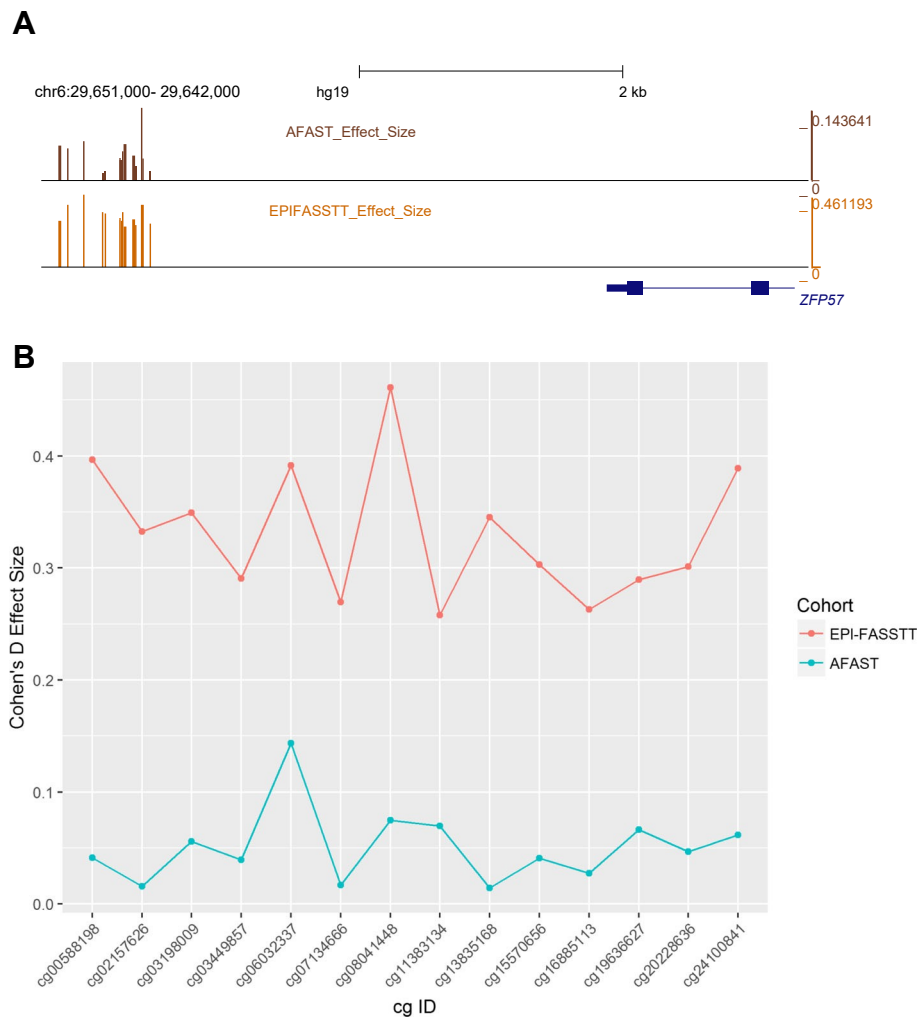
**Increased *ZFP57* methylation in response to FA in maternal blood samples**

In order to investigate the effects of FA in maternal tissue, and to elucidate if this differentially methylated region upstream of *ZFP57* was directly responsive, we carried out pyrosequencing on matched maternal buffy coat samples at GW14 ( $n = 24$ ) and GW36 ( $n = 24$ ) (i.e.,

comparing the same mother’s blood sample taken before and after intervention). Pyrosequencing analysis confirmed that FA-supplemented mothers show a 5.51% increase in DNA methylation levels at this DMR after late gestation supplementation ( $p = 0.609$ ), in contrast to non-supplemented mothers, whose methylation levels decreased 1.51% at GW36 ( $p = 0.826$ ) (Table 2).

**Effect of FA at the *ZFP57* DMR in a second cohort**

In order to test the generality of the effect of folic acid intervention on this genomic region, we examined data from a second randomized-controlled trial. The Aberdeen Folic Acid Supplementation Trial (AFAST) was an RCT using two doses of folic acid (0.2 and 5 mg/day vs placebo) during pregnancy, with intervention starting at antenatal booking at  $< 30$  weeks gestational age [50]. The study was conducted in the late 1960s, and recently, Richmond and colleagues [35] followed up on the offspring born to the mothers who had participated in the trial, mean present age of 47 years. Saliva samples were collected from those who could be identified and



**Fig. 5** Comparison of AFAST and EpiFASSTT data for the DMR. **a** Effect size (Cohen's *d*) at each CpG in the *ZFP57* DMR was calculated by comparing high dose and placebo from the AFAST study and plotted against the locus (top track). A similar analysis was done for the EpiFASSTT data (bottom track). Maxima are indicated at right, scale bar and location at top; note: no other CpG outside the DMR are shown in this analysis. **b** The two sets of values from **a** are plotted on the same scale to give an indication of comparability

consented, with subsequent 450 k array analysis conducted using modeling approaches to correct for hidden variables such as cell counts [35]. Examination of the CpG in the *ZFP57* DMR which we had identified in the EpiFASSTT cohort showed the same trends in the AFAST high-folate cohort versus placebo, with change in a positive direction across the whole region (Fig. 5a), although effect size was lower at each site in the AFAST study (Fig. 5b).

### Discussion

We have previously reported DNA methylation differences at imprinted loci using cord blood from the EpiFASSTT trial of folic acid (FA) supplementation in later pregnancy by using a candidate gene approach. Here, we used the same samples to carry out an unbiased

genome-wide screen for methylation differences using the EPIC array. The top hit was a differentially methylated region upstream of the imprint controller *ZFP57*, and we separately verified methylation differences by pyroassay. This region responded to FA supplementation in maternal blood as well as in cord blood and showed differences between FA-treated and untreated in an independent cohort [50]. Altered methylation at *ZFP57* was associated with increased variation in methylation at imprinted loci in cord blood. We also showed using two separate cell line models that altering methylation at the *ZFP57* upstream region can affect transcription, indicating a potential feedback mechanism may be operating here. We were also able to identify and verify methylation changes at a number of other individual CpG sites including some in the gene bodies of the *NXN* and *PRKAR1B* genes and at the start

**Table 2** ZFP57 methylation in maternal blood pre- and post-intervention. DNA methylation levels of ZFP57 DMR in maternal blood samples at GW14 and GW36.

Sample group	Gestational week (GW)	Mean methylation (%)	Standard deviation (SD)	Change in methylation (%)	<i>p</i> values
Treatment ( <i>n</i> = 24)	GW14	57.47	15.37	+ 5.51	0.609
	GW36	62.98	14.94		
Placebo ( <i>n</i> = 24)	GW14	64.36	6.58	-1.51	0.826
	GW36	62.85	7.13		

GW gestational week, SD standard deviation

of the *MIR4520A/B* gene, but these were less likely to have functional consequences. It is notable also that we found more decreases in methylation genome-wide than increases, which may seem counter-intuitive; however, we and others have reported similar response to FA previously [43, 51]. It has been suggested that FA may cause feedback inhibition by altering the SAM to SAH ratio and therefore the intracellular methylation potential [52].

Uncovering a DMR at a region controlling *ZFP57* transcription as the top hit in an unbiased screen was particularly striking in the EpiFASSTT randomized controlled trial, where we have already shown, using a candidate gene approach, that methylation levels were perturbed at some imprinted loci. The primary importance of *ZFP57*, as described in the literature from mechanistic work, is in maintaining imprinting, and it is currently the only protein known to be dedicated solely or largely to this epigenetic process [53]. *ZFP57* was discovered as a maternal-zygotic effect gene which was required in mice for establishing methylation at some imprints in the oocyte, and for maintaining all imprints, both maternal and paternal, in the preimplantation embryo [38]. It does this by binding to a conserved hexamer consensus sequence (5'-TGCme5CGC-3) found at all imprinting control regions (ICRs) [54, 55], recognizing the methylated CpG in this motif, as shown in a crystallographic study [56]. Deletion of mouse *Zfp57* causes a loss of methylation from the modified parental allele by mid-gestation, with subsequent dysregulation of transcription at imprinted loci and embryonic lethality [55]. Importantly, mutations in the human homolog *ZFP57* are also associated with hypomethylation of multiple imprinted loci, indicating a conserved role in human for this gene in maintaining imprints [39].

Although this is the first report, to our knowledge, from a randomized controlled trial of FA intervention which implicates methylation changes at *ZFP57*, it was previously reported from a small observational study (*n* = 23) that maternal folate concentrations in the third trimester were associated with changes at a DMR at the same genomic location [51] when cord blood DNA methylation levels at birth were profiled. While that study reported a loss rather than gain of methylation, it was not an RCT but an

observational study, and so could not test the effects of folate supplementation directly in a controlled fashion: there were many other differences in study design, numbers of participants, and analysis methods. It should also be noted that the high folate group in that study had levels of serum folate almost twice those seen in our treated samples (74.59 +/- 6.1 nmol/L Amarasekera et al. vs 46.5 +/- 19.5 nmol/L GW36 treated group in this study), highlighting that we are protecting normal folate levels rather than elevating them. Although the largest-to-date observational study, comprising a meta-analysis of the MoBa (*n* = 1275) and Generation R (*n* = 713) cohorts, did not identify this region as a top hit, they could confirm that five CpG sites within this 923 bp region were significantly altered, though not the direction of change [57]. These two papers reporting changes from different observational studies nevertheless lend considerable support to this being a true folate-sensitive DMR. We could also verify using a separate biological assay the magnitude and direction of change in methylation, a gain of 5.44% in the treatment group, at the DMR in cord blood by using pyrosequencing (*p* = 0.172). Furthermore, by comparing the mother's pre- and post-intervention, we could show that this region also gained methylation in the treated mothers, but lost methylation in the placebo group, providing a further degree of validation.

To extend our findings, we also used data from one of the few other RCTs testing the role of folic acid during pregnancy, the AFAST study [50]. We found a small effect (Cohen's *d* < 0.2) at all the CpG across the *ZFP57* DMR, whereas there was a medium effect (Cohen's *d* < 0.5) seen at the same region in the EpiFASSTT study. The effect in AFAST was only seen with the high dose of FA (5 mg/day) vs placebo, rather than the lower dose (200 µg/day) which was closer to that used in EpiFASSTT (400 µg/day), and the effect size was smaller than that seen in EpiFASSTT. There may be a number of reasons why effect size was smaller in AFAST: (1) the time between exposure and measurement is much greater, with median age 47 years in AFAST, vs newborns in EpiFASSTT; (2) the AFAST participants used were recruited significantly later than other groups (20.2 weeks for high dose vs 16.3 for low dose), meaning that there was less time spent exposed to the additional FA while in the womb; (3) the AFAST

DNA samples were derived from saliva, while the Epi-FASSTT DNA samples are from cord blood; and (4) the final numbers for the AFAST comparisons were very low (5 mg/day  $n = 23$ ; placebo  $n = 43$ ). Notwithstanding these limitations, the AFAST study showed a similar effect in terms of direction and magnitude at the same region upstream of *ZFP57*, providing further evidence that this is a bona fide FA sensor.

Given the role of *ZFP57* in imprint maintenance, we also took advantage of the array to examine imprinted genes in our samples. Of these, only the maternal imprint *NNAT* (neuronatin) showed a small but significant loss of methylation in the treatment group, consistent with other evidence [58]. *NNAT* is highly expressed in the brain and placental tissue and functions during brain development to regulate ion channels and maintain hindbrain and pituitary segment identity [59]. *ZFP57* is essential for the maintenance of this imprint [38]. Induction of increasing mRNA levels of *NNAT* commences at midgestation in association with neurogenesis and peaks upon neuroepithelial proliferation and neuroblast formation [60], which would coincide with when folate concentrations increased in the treated group. Although we previously reported significant differences overall at *IGF2*, and at some CpG for *GRB10*, in our candidate gene approach using these samples [43], that was based on pyroassays which covered smaller regions of the imprinted DMR, whereas the probes from the array are more dispersed and cover a larger area. It was also notable that, while there was little change at other imprinted DMR as assessed by the array, there did appear to be an increase in the variability of methylation at these regions, an effect which was small but statistically significant and consistent with findings from a mouse model where FA supplementation increased variance in methylation levels across generations [61]. Given that *ZFP57* has a role in maintaining imprints, increased methylation at the upstream controller as seen in our FA-treated samples should lead to decreased transcription of *ZFP57*, which could potentially lead to reduced ability to maintain imprints and increased variability in methylation at the ICR. These possibilities can be further explored using our in vitro cell models.

It remains to be established from mechanistic studies in mouse whether *ZFP57* plays any role in maintaining methylation in vivo in the post-implantation embryo. It is also possible that methylation of the DMR in human blood may not reflect the methylation levels seen at earlier stages, or in tissues which normally express the gene, which includes oocytes and some neural cells. It may be that methylation levels at the *ZFP57* DMR reported here reflect changes which have occurred in the cord and maternal bloods independently of what is occurring in the germline, and this would need to be assessed. It is also quite likely, given that imprints are thought to be

established much earlier during development, that it would not be until the next generation that effects at imprinted germline DMRs could be seen. In this context, several studies have pointed to transgenerational rather than intergenerational effects at imprinted loci [62, 63]. It should also be noted that methylation levels varied substantially across the *ZFP57* DMR and between individuals (max = 94.97, min = 20.95), unlike the imprinted DMR which vary much less and may be buffered against methylation changes by multiple mechanisms.

In addition to its well-established role in imprinting, *ZFP57* has also been proposed to act as a transcriptional repressor in Schwann cells, which comprise the principal glia of the peripheral nervous system [47]. Recent work from our group has indicated children born from mothers supplemented with FA in late gestation have psychosocial developmental benefits, scoring significantly higher for emotional intelligence and resilience in comparison with children not exposed to FA supplementation in later pregnancy [64]. Further work needs to be carried out to check if there are any other novel targets of *ZFP57* which may be affected in later childhood and adulthood.

We sought to clarify whether an increase in methylation at the *ZFP57* DMR as seen in this RCT would have a substantial effect on the production of the protein. In order to explore whether changes in methylation can alter transcription, we utilized cell lines where the only variable was the presence or absence of DNA methylation. Our results from these two systems (HCT116 cells with methyltransferase deficiency and SH-SY5Y cells treated with an inhibitor) showed that altering methylation alone can cause changes in transcription at the *ZFP57* locus and that this is linked to changes in methylation at the DMR. Our results therefore support the hypothesis that the DMR represents an upstream control element for the gene, which we have shown from the RCT is sensitive to methyl donor status in the diet. Little is currently known about the factors controlling *ZFP57* transcription. Interestingly, the region containing the DMR does not appear to be conserved in mice and so may represent a human-specific element. However, it has features characteristic of a control element, as from examining publicly available datasets on the UCSC genome browser, there are DNase I hypersensitive sites present here and data suggesting transcription factors may bind. We are currently exploring these aspects of the work further.

## Conclusions

Despite the limitations discussed above, we have nevertheless shown conclusively that a region upstream of the imprint controller *ZFP57* shows changes in methylation in mothers in response to intervention during later

pregnancy with FA, a methyl donor, and that this effect is also evident in the cord blood in their offspring. Our findings are borne out by other observational studies as well as an independent RCT [50]. We have also clearly demonstrated that altering methylation is sufficient in itself to cause changes in transcription of the gene. These results have implications for the control of imprinting by environmental inputs and uncover a novel transcriptional control element which may be involved in this process.

## Methods

### Study design and sample collection

Samples were acquired from the FASSTT (folic acid supplementation in the second and third trimester) study cohort, a previously conducted double-blinded, randomized controlled trial in Northern Ireland described in full previously [42, 43]. To summarize in brief, women with singleton pregnancies were recruited at approximately 14 weeks of gestation from antenatal clinics at the Causeway Hospital, Coleraine ( $n = 226$ ; Fig. 1). Women were excluded from participation if they were taking medication known to interfere with B-vitamin metabolism or if they had any vascular, renal, hepatic, or gastrointestinal disease, epilepsy, or had a previous NTD-affected pregnancy. Prior to randomization,  $n = 36$  women withdrew from the study. The remaining eligible participants at the end of their first trimester were randomized into two groups; one group received 400  $\mu\text{g}/\text{d}$  folic acid ( $n = 96$ ) and the other a placebo in pill form ( $n = 94$ ) until the end of their pregnancy. Randomization was done on a double-blind basis. Maternal non-fasting blood samples were taken at gestational week 14 (GW14), prior to intervention commencement, and at GW36, towards the end of the intervention. The study was completed by 119 women, as 71 participants were excluded during the study (see Fig. 1). A total of  $n = 37$  women were excluded from the folic acid group for the following reasons: participant withdrawal  $n = 11$ , pregnancy complications  $n = 13$ , prescribed folic acid  $n = 6$ , fetal death  $n = 6$ , non-compliance  $n = 6$ . A total of  $n = 34$  women were excluded from the placebo group for the following reasons: participant withdrawal  $n = 14$ , pregnancy complications  $n = 8$ , prescribed folic acid  $n = 5$ , fetal death  $n = 2$ , non-compliance  $n = 3$ , hospital transfer  $n = 2$ . Umbilical cord blood samples were collected after the expulsion of the placenta at delivery, along with birth weight, length, head circumference, mode of delivery, and Apgar score.

### Blood sample processing and B-vitamin biomarker determination

Blood samples were collected in EDTA-lined tubes, kept refrigerated, and processed within 4 h (excepting cord

blood, processed within 24 h). Blood samples were analyzed for serum and red blood cell folate and vitamin B12 via microbiological assay as previously described [65, 66]. The buffy coat was used for methylenetetrahydrofolate reductase (*MTHFR*) 677C > T genotyping as described [67]. Quality control was affirmed by repeated analysis of stored batches of pooled samples. Intra- and inter-assay CVs were  $\leq 8.2\%$  for serum and RBC folate and  $\leq 10.4\%$  for serum vitamin B12.

### Maternal dietary analysis

Dietary data was collected using a 4d food diary in combination with a food-frequency questionnaire during the second trimester of pregnancy, with particular emphasis on a B-vitamin-fortified food intake. Dietary analysis was carried out using WISP version 3.0 (Tinuviel Software, UK) modified to segregate naturally occurring folate in foods versus folic acid fortification of foods; these were combined to enable calculation of dietary folate equivalents.

### Cell culture

HCT116 and double knockout (DKO) cells [46] were cultured in 1 g/L glucose DMEM supplemented with 10% FBS and  $1\times$  NEAA (Thermo Scientific, Loughborough, UK). SH-SY5Y cells were cultured in DMEM/F12 medium supplemented with 10% FBS (Thermo Scientific). For treatment with 5'-aza-2-deoxycytidine (5-aza-dC) (Sigma-Aldrich, Dorset, UK), SH-SY5Y cells were seeded onto a 90-mm plate in complete medium, and the following day medium was replaced and supplemented with 5-aza-dC at a final concentration of 1  $\mu\text{M}$ , which was renewed at 24-h intervals up to 72 h. Cells were then harvested for DNA and RNA extraction.

### Transcriptional analysis

RNA was extracted using the RNeasy Mini kit (Qiagen, Crawley, UK) according to manufacturer's instructions. Complementary DNA (cDNA) was synthesized and RT-qPCR/RT-PCR were carried out as previously [29]. Primer sequences are listed in Additional file 4: Table S1. Human reference total RNA was used as a positive control for expression (Clontech, UK).

### DNA extraction, bisulfite conversion, and Infinium MethylationEPIC Beadchip Array

Genomic DNA was extracted from cultured cells as previously described [25] and from cord blood using the QiAMP DNA Blood Mini kit (Qiagen), according to manufacturer's instructions. Purity and integrity of DNA were assessed by agarose gel electrophoresis and using the Nanodrop 2000 spectrophotometer (Labtech International, Ringmer, UK). DNA quantification was determined using Quant-IT PicoGreen dsDNA Assay Kit

(Invitrogen, Paisley, UK). The DNA at a concentration of 50 ng/μl was sent to Cambridge Genomic Services (Cambridge, UK), who bisulfite converted the DNA in-house using the EZ DNA Methylation Kit (Zymo Research, California, USA) prior to hybridization to the Infinium Human Methylation EPIC BeadChip Array and scanning with the Illumina iScan according to manufacturer's instructions (Illumina, Chesterford, UK).

#### Bioinformatic analysis

*GenomeStudio* (Illumina v3.2) was used for initial data processing. Subsequently, *idat* files were imported into the *RnBeads* package (version 1.6.1) [45] in the freely available statistical software platform R (version 3.1.3) using the R Studio interface (Version 0.99.903). Samples were quality control checked including removal of probes with missing values, containing SNPs, or of poor quality using the *greedyCut* algorithm, then sex chromosomes were removed from the analysis. Background correction was carried out using *methylumi.noob* and the methylation values of the remainder probes were normalized using *bmiq* [68]. Initial data exploration in *RnBeads* used principal components analysis (PCA) to explore potential correlations between the groups and known confounders such as BMI, smoking, and gender. In addition, in order to account for any hidden confounding variables in the dataset, surrogate variable analysis was carried out using the *sva* package with the Buja and Eyboglu algorithm from (1992) [69] Briefly, potential surrogate variables such as age, sample plate, Sentrix ID, and Sentrix Position were tested for association with the target variable sample group using PCA and any surrogate variable with a high correlation to sample group was adjusted for and incorporated into the making of the *limma* based linear model. The methylation intensities for each probe, each representing a CpG site, were represented as  $\beta$  values (ranging from 0, unmethylated, to 1, fully methylated), and these were plotted against genomic loci (based on *hg19*-Human Genome Build 19) using *GALAXY* software (<https://usegalaxy.org/>) [70] in order to visualize changes in DNA methylation on the University of California at Santa Cruz genome browser (<https://genome.ucsc.edu/>) as described previously [71].

#### Bisulfite pyrosequencing

Primers spanning the probes of interest from the array were designed using the PyroMark Assay Design Software 2.0 and bisulfite-treated DNA PCR-amplified using the PyroMark PCR kit prior to analysis on a PyroMark Q24 according to manufacturer's instruction (Qiagen). The primer sequences are summarized in Additional file 4: Table S1. Amplification was carried out as follows: 95 °C for 15 min, followed by 45 cycles of 95 °C for 30 s, 56 °C for 30 s, and 72 °C for 30 s, with a final

elongation step at 72 °C for 10 min. Products were verified via gel electrophoresis prior to pyrosequencing analysis.

#### Statistical analysis

Statistical analysis was performed using the Statistical Package for the Social Sciences software (SPSS) (Version 22.0; SPSS UK Ltd., Chertsey, UK). The results are expressed as mean  $\pm$  SD, except where otherwise stated. For normalization purposes, variables were log transformed before analysis, as appropriate. Differences between treatment groups for participant characteristics were assessed using an independent *t* test for continuous variables or chi-square for categorical variables. Pyrosequencing data and RT-qPCR data were analyzed using Student's *t* test to identify statistical differences between intervention groups. A *p* value < 0.05 was considered significant. Differential methylation analysis was conducted in *RnBeads* (see above) on a site and region level. The normalized  $\beta$  values were converted into *M* values ( $M = \log_2(\beta/(1-\beta))$ ) and differential methylation between samples (placebo vs. treatment) was estimated with hierarchical linear models using *limma*. Ranking was automatically carried out in *RnBeads* and was based on the combination of the average difference in means across all sites in the promoter regions of the sample groups, the mean of quotients in mean methylation, and the combined *p* value, which was calculated from all site *p* values in the region using a generalization of Fisher's method [72]. The smaller the combined rank for a region, the more evidence for differential methylation it exhibits.

#### Additional files

**Additional file 1:** Figure S1. Correlation between folate levels in cord blood and mother. Scatterplot shows log-converted red blood cell folate (RCF) levels in nanomoles per liter (nmol/l) at gestational week 36 (GW36) for mothers (post-intervention) and matched cord blood. The line of best fit shows significant correlation between mothers and offspring ( $r = 0.619$ ;  $p < 0.001$ ). (PDF 460 kb)

**Additional file 2:** Figure S2. QQ plot shows no evidence of population substructure effects. The observed Chi-squared ( $\chi^2$ ) values (open circles), plotted as  $-\log_{10}$  of the *p* value for both sample groups, fit tightly to the expected  $\chi^2$  values (red line), indicating little evidence of association due to population substructure effects and that the top hits which deviate from the line (right-hand side) are likely to represent true differences due to loci with large effects. (PDF 332 kb)

**Additional file 3:** Figure S3. Median methylation levels at imprint control regions. Methylation levels at imprint control regions (ICR) were assessed by matching EPIC array probes to the imprint germline DMR intervals defined by [48] (A) or [49] (B) then taking the average (median) across each. The identities of each ICR and number of probes are indicated below. Boxes show the median and interquartile range for the individual averages from each group (Placebo  $n = 45$ , Treated  $n = 41$ ), whiskers represent the range of values, dots indicate outliers. (PDF 1518 kb)

**Additional file 4:** Table S1. Pyrosequencing and transcriptional primer sets used in this study. Pyroassay primers are given as bisulfite converted

sequence. The same primers were used for both RT-PCR and RT-qPCR. (DOCX 15 kb)

#### Abbreviations

5-aza-dC: 5'aza-2'deoxyctidine; AFAST: Aberdeen Folic Acid Supplementation Trial; BBSRC: Biotechnology and Biological Sciences Research Council; BMI: Body mass index; DKO: Double knockout; DMR: Differentially methylated region; DNMT: DNA methyltransferases; ESRC: Economic and Social Research Council; FA: Folic acid; FASSTT: Folic acid supplementation in second and third trimester; GW: Gestational week; ICR: Imprint control region; IQR: Interquartile range; KRAB: Krueppel-associated box; MRC: Medical Research Council; MTHFR: Methylene tetrahydrofolate reductase; MWU: Mann-Whitney *U* test; NNAT: Neuronatin; NTD: Neural tube defects; ORECNI: Office for Research and Ethics Committees Northern Ireland; Pyroassay: Pyrosequencing methylation assay; QQ: Quantile-quantile; RCT: Randomized controlled trial; RT-PCR: Reverse transcription-polymerase chain reaction; SAM: S-Adenosylmethionine; SPSS: Statistical Package for the Social Sciences; WT: Wild type; ZFP57: Zinc finger protein 57

#### Acknowledgements

The authors are grateful to the other members of the Walsh and Relton labs who provided valuable feedback on the work.

#### Funding

Work was supported by grants jointly funded by the Economic and Social Research Council (ESRC) and Biotechnology and Biological Sciences Research Council (BBSRC), grant refs: ES/N000323/1 (CPW) and ES/N000498/1 (CLR). RCR, MS and CLR work in a unit supported by the Medical Research Council (MC\_UU\_12013/1, MC\_UU\_12013/2 and MC\_UU\_12013/8).

#### Availability of data and materials

The datasets used and analyzed during the current study are available where appropriate from the corresponding author on reasonable request and subject to governance regulations at Ulster (EpiFASSTT): for data from the AFAST study contact C. Relton.

#### Authors' contributions

CPW, KP, and HM designed and planned the work. REI and MO carried out the lab work. REI and SJT performed the bioinformatics analysis for EpiFASSTT. RR carried out the analysis for AFAST. AC and DLM helped with EpiFASSTT samples and statistics. MM and TC advised on biopsychosocial correlations. MS advised on bioinformatics approaches for both cohorts. CLR advised on overall approaches and coordinated the AFAST comparison. REI and CPW wrote the paper. All authors commented on the final manuscript. All authors read and approved the final manuscript.

#### Ethics approval and consent to participate

The Office for Research and Ethics Committees Northern Ireland (ORECNI) granted ethical approval (reference 05/ Q2008/21) and each participant gave written informed consent upon recruitment.

#### Consent for publication

Not applicable

#### Competing interests

The authors declare that they have no competing interests.

## Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

#### Author details

<sup>1</sup>Genomic Medicine Research Group, School of Biomedical Sciences, Ulster University, Coleraine BT52 1SA, UK. <sup>2</sup>Nutrition Innovation Centre for Food and Health, School of Biomedical Sciences, Ulster University, Coleraine, UK. <sup>3</sup>Psychology Institute, Ulster University, Coleraine, UK. <sup>4</sup>MRC Integrative Epidemiology Unit, Bristol Medical School, University of Bristol, Bristol, UK.

Received: 3 November 2018 Accepted: 21 January 2019

Published online: 18 February 2019

#### References

- Crider KS, Yang TP, Berry RJ, Bailey LB. Folate and DNA methylation: a review of molecular mechanisms and the evidence for folate's role. *Am Soc Nutr.* 2012;3:21–38.
- Irwin RE, Pentieva K, Cassidy T, Lees-Murdock DJ, McLaughlin M, Prasad G, et al. The interplay between DNA methylation, folate and neurocognitive development. *Epigenomics.* 2016;8(6):863–79.
- Bailey LB, Stover PJ, McNulty H, Fenech MF, Gregory JF, Mills JL, et al. Biomarkers of nutrition for development—folate review. *J Nutr.* 2015;145(7):1636S–80S.
- Vitamin MRC. Study Research Group. Prevention of neural tube defects: results of the Medical Research Council Vitamin Study. MRC Vitamin Study Research Group. *Lancet.* 1991;338:131–7.
- Czeizel AE, Dudás I. Prevention of the first occurrence of neural-tube defects by periconceptional vitamin supplementation. *Obstet Gynecol Surv.* 1993;48:395–7.
- Centers for Disease Control. Recommendations for the use of folic acid to reduce the number of cases of spina bifida and other neural tube defects. *Morb Mortal Wkly Report.* 1992;41:1–8.
- Department of Health. Folic acid and the prevention of neural tube defects. Report from an expert advisory group. Health Publications Unit, Heywood, Lancashire, 1992.
- Greene NDE, Stanier P, Moore GE. The emerging role of epigenetic mechanisms in the etiology of neural tube defects. *Epigenetics.* 2011;6(7):875–83.
- Roctus A, Jansen K, Geet C, Freson K. Nutri-epigenomic studies related to neural tube defects: does folate affect neural tube closure via changes in DNA methylation? *Mini-Reviews Med Chem.* 2015;15(13):1095–102.
- Blom HJ, Shaw GM, Den Heijer M, Finnell RH. Neural tube defects and folate: case far from closed. *Nat Rev Neurosci.* 2006;7(9):724–31.
- Mills JL, Molloy AM, Reynolds EH. Do the benefits of folic acid fortification outweigh the risk of masking vitamin B12 deficiency? *BMJ.* 2018;360:k724.
- Schrott R, Murphy SK. Folic acid throughout pregnancy: too much? *Am J Clin Nutr.* 2018;107:497–8.
- Hodgetts V, Morris R, Francis A, Gardosi J, Ismail K. Effectiveness of folic acid supplementation in pregnancy on reducing the risk of small-for-gestational age neonates: a population study, systematic review and meta-analysis. *BJOG.* 2015;122:478–90.
- Roth C, Magnus P, Schjølberg S, Stoltenberg C, Surén P, McKeague IW, et al. Folic acid supplements in pregnancy and severe language delay in children. *JAMA.* 2011;306:1566–73.
- Wang M, Li K, Zhao D, Li L. The association between maternal use of folic acid supplements during pregnancy and risk of autism spectrum disorders in children: a meta-analysis. *Mol Autism.* 2017;8:51.
- Eryilmaz H, Dowling KF, Huntington FC, Rodriguez-Thompson A, Soare TW, Beard LM, et al. Association of prenatal exposure to population-wide folic acid fortification with altered cerebral cortex maturation in youths. *JAMA Psychiatry.* 2018;75(9):918–28.
- Barua S, Kuizon S, Junaid MA. Folic acid supplementation in pregnancy and implications in health and disease. *J Biomed Sci.* 2014;21:77.
- Julvez J, Fortuny J, Mendez M, Torrent M, Ribas-Fitó N, Sunyer J. Maternal use of folic acid supplements during pregnancy and four-year-old neurodevelopment in a population-based birth cohort. *Paediatr Perinat Epidemiol.* 2009;23:199–206.
- Villamor E, Rifas-Shiman SL, Gillman MW, Oken E. Maternal intake of methyl-donor nutrients and child cognition at 3 years of age. *Paediatr Perinat Epidemiol.* 2012;26:328–35.
- de Graaf-Peters VB, Hadders-Algra M. Ontogeny of the human central nervous system: what is happening when? *Early Hum Dev.* 2006;82:257–66.
- Nyaradi A, Li J, Hickling S, Foster J, Oddy WH. The role of nutrition in children's neurocognitive development, from pregnancy through childhood. *Front Hum Neurosci.* 2013;7:97.
- Smith ZD, Meissner A. DNA methylation: roles in mammalian development. *Nat Rev.* 2013;14:204–20.
- Wu H, Coskun V, Tao J, Xie W, Ge W, Yoshikawa K, et al. Dnmt3a-dependent nonpromoter DNA methylation facilitates transcription of neurogenic genes. *Science (80-).* 2010;329:444–7.



24. Neri F, Krepelova A, Incamato D, Maldotti M, Parlato C, Galvagni F, et al. XNmt3L antagonizes DNA methylation at bivalent promoters and favors DNA methylation at gene bodies in ESCs. *Cell*. 2013;155(1):121–34.
25. Irwin RE, Thakur A, O' Neill KM, Walsh CP. 5-Hydroxymethylation marks a class of neuronal gene regulated by intragenic methylcytosine levels. *Genomics*. 2014;104:383–92.
26. Li E, Bestor TH, Jaenisch R. Targeted mutation of the DNA methyltransferase gene results in embryonic lethality. *Cell*. 1992;69:915–26.
27. Okano M, Bell DW, Haber DA, Li E. DNA methyltransferases Dnmt3a and Dnmt3b are essential for de novo methylation and mammalian development. *Cell*. 1999;99:247–57.
28. Borgel J, Guibert S, Li Y, Chiba H, Schübeler D, Sasaki H, et al. Targets and dynamics of promoter DNA methylation during early mouse development. *Nat Genet*. 2010;42:1093–100.
29. Rutledge CE, Thakur A, O' Neill KM, Irwin RE, Sato S, Hata K, et al. Ontogeny, conservation and functional significance of maternally inherited DNA methylation at two classes of non-imprinted genes. *Development*. 2014;141:1313–23.
30. Dolinoy DC. The agouti mouse model: an epigenetic biosensor for nutritional and environmental alterations on the fetal epigenome. *Nutr Rev*. 2008;66:S7–11.
31. Steegers-Theunissen RP, Obermann-Borst SA, Kremer D, Lindemans J, Siebel C, Steegers EA, et al. Periconceptional maternal folic acid use of 400 µg per day is related to increased methylation of the IGF2 gene in the very young child. *PLoS One*. 2009;4(11):e7845.
32. Heijmans BT, Tobi EW, Stein AD, Putter H, Blauw GJ, Susser ES, et al. Persistent epigenetic differences associated with prenatal exposure to famine in humans. *Proc Natl Acad Sci*. 2008;105:17046–9.
33. Haggarty P, Hoad G, Campbell DM, Horgan GW, Piyathilake C, McNeill G. Folate in pregnancy and imprinted gene and repeat element methylation in the offspring. *Am J Clin Nutr*. 2013;97:94–9.
34. Dominguez-Salas P, Moore SE, Baker MS, Bergen AW, Cox SE, Dyer RA, et al. Maternal nutrition at conception modulates DNA methylation of human metastable epialleles. *Nat Commun*. 2014;5:3746.
35. Richmond RC, Sharp GC, Herbert G, Atkinson C, Taylor C, Bhattacharya S, et al. The long-term impact of folic acid in pregnancy on offspring DNA methylation: follow-up of the Aberdeen Folic Acid Supplementation Trial (AFAST). *Int J Epidemiol*. 2018;47(3):928–37.
36. Pauwels S, Ghosh M, Duca RC, Bekaert B, Freson K, Huybrechts I, et al. Maternal intake of methyl-group donors affects DNA methylation of metabolic genes in infants. *Clin Epigenetics*. 2017;9:16.
37. Nakamura T, Arai Y, Umehara H, Masuhara M, Kimura T, Taniguchi H, et al. PGC7/Stella protects against DNA demethylation in early embryogenesis. *Nat Cell Biol*. 2007;9:64–71.
38. Li X, Ito M, Zhou F, Youngson N, Zuo X, Leder P, et al. A maternal-zygotic effect gene, Zfp57, maintains both maternal and paternal imprints. *Dev Cell*. 2008;15:547–57.
39. Mackay DJG, Callaway JLA, Marks SM, White HE, Acerini CL, Boonen SE, et al. Hypomethylation of multiple imprinted loci in individuals with transient neonatal diabetes is associated with mutations in ZFP57. *Nat Genet*. 2008;40:949–51.
40. Cooper WN, Khulan B, Owens S, Elks CE, Seidel V, Prentice AM, et al. DNA methylation profiling at imprinted loci after periconceptional micronutrient supplementation in humans: results of a pilot randomized controlled trials. *World Rev Nutr Diet*. 2014;26(5):1782–90.
41. Barker DJP. The developmental origins of chronic adult disease. *Acta Paediatr Int J Paediatr Suppl*. 2004;93(446):26–33.
42. McNulty B, McNulty H, Marshall B, Ward M, Molloy AM, Scott JM, et al. Impact of continuing folic acid after the first trimester of pregnancy: findings of a randomized trial of folic acid supplementation in the second and third trimesters. *Am J Clin Nutr*. 2013;98:92–8.
43. Caffrey A, Irwin RE, McNulty H, Strain JJ, Lees-Murdock DJ, McNulty BA, et al. Gene-specific DNA methylation in newborns in response to folic acid supplementation during the second and third trimesters of pregnancy: epigenetic analysis from a randomized controlled trial. *Am J Clin Nutr*. 2018;107:566–75.
44. Teschendorff AE, Relton CL. Statistical and integrative system-level analysis of DNA methylation data. *Nat Rev Genet*. 2018;19(3):129–47.
45. Assenov Y, Müller F, Lutsik P, Walter J, Lengauer T, Bock C. Comprehensive analysis of DNA methylation data with RnBeads. *Nat Methods*. 2014;11:1138–40.
46. Rhee I, Bachman KE, Park BH, Jair KW, Yen RWC, Schuebel KE, et al. DNMT1 and DNMT3b cooperate to silence genes in human cancer cells. *Nature*. 2002;416:552–6.
47. Alonso MBD, Zoidl G, Taveggia C, Bosse F, Zoidl C, Rahman M, et al. Identification and characterization of ZFP-57, a novel zinc finger transcription factor in the mammalian peripheral nervous system. *J Biol Chem*. 2004;279:25653–64.
48. Woodfine K, Huddleston JE, Murrell A. Quantitative analysis of DNA methylation at all human imprinted regions reveals preservation of epigenetic stability in adult somatic tissue. *Epigenetics and Chromatin*. 2011;4(1):1.
49. Court F, Tayama C, Romanelli V, Martin-Trujillo A, Iglesias-Platas I, Okamura K, et al. Genome-wide parent-of-origin DNA methylation analysis reveals the intricacies of human imprinting and suggests a germline methylation-independent mechanism of establishment. *Genome Res*. 2014;24:554–69.
50. Charles DHM, Ness AR, Campbell D, Smith GD, Whitley E, Hall MH. Folic acid supplements in pregnancy and birth outcome: re-analysis of a large randomised controlled trial and update of Cochrane review. *Paediatr Perinat Epidemiol*. 2005;19:2.
51. Amarasekera M, Martino D, Ashley S, Harb H, Kesper D, Strickland D, et al. Genome-wide DNA methylation profiling identifies a folate-sensitive region of differential methylation upstream of ZFP57-imprinting regulator in humans. *FASEB J*. 2014;28:4068–76.
52. Christensen KE, Mikael LG, Leung KY, Lévesque N, Deng L, Wu Q, et al. High folic acid consumption leads to pseudo-MTHFR deficiency, altered lipid metabolism, and liver injury in mice. *Am J Clin Nutr*. 2015;101:646–58.
53. Mackin SJ, Thakur A, Walsh CP. Imprint stability and plasticity during development. *Reproduction*. 2018;156:43–55.
54. Strogantsev R, Krueger F, Yamazawa K, Shi H, Gould P, Goldman-Roberts M, et al. Allele-specific binding of ZFP57 in the epigenetic regulation of imprinted and non-imprinted monoallelic expression. *Genome Biol*. 2015;16:112.
55. Quenneville S, Verde G, Corsinotti A, Kapopoulou A, Jakobsson J, Offner S, et al. In embryonic stem cells, ZFP57/KAP1 recognize a methylated hexanucleotide to affect chromatin and DNA methylation of imprinting control regions. *Mol Cell*. 2011;44:361–72.
56. Liu Y, Toh H, Sasaki H, Zhang X, Cheng X. An atomic model of Zfp57 recognition of CpG methylation within a specific DNA sequence. *Genes Dev*. 2012;26(21):2374–9.
57. Joubert BR, Den Dekker HT, Felix JF, Bohlin J, Ligthart S, Beckett E, et al. Maternal plasma folate impacts differential DNA methylation in an epigenome-wide meta-analysis of newborns. *Nat Commun*. 2016;7:10577.
58. Hoyo C, Daltveit AK, Iversen E, Benjamin-Neelon SE, Fuemmeler B, Schildkraut J, et al. Erythrocyte folate concentrations, CpG methylation at genomically imprinted domains, and birth weight in a multiethnic newborn cohort. *Epigenetics*. 2014;9:1120–30.
59. Dou D, Joseph R. Cloning of human neuronatin gene and its localization to chromosome-20q11.2-12: the deduced protein is a novel "proteolipid.". *Brain Res*. 1996;723:8–22.
60. Kikyo N, Williamson CM, John RM, Barton SC, Beechey CV, Ball ST, et al. Genetic and functional analysis of neuronatin in mice with maternal or paternal duplication of distal Chr 2. *Dev Biol*. 1997;190:66–77.
61. Li CCY, Cropley JE, Cowley MJ, Preiss T, Martin DIK, Suter CM. A sustained dietary change increases epigenetic variation in isogenic mice. *PLoS Genet*. 2011;7(4):e1001380.
62. Serpeloni F, Radtke K, de Assis SG, Henning F, Nätt D, Elbert T. Grandmaternal stress during pregnancy and DNA methylation of the third generation: an epigenome-wide association study. *Transl Psychiatry*. 2017;7:e1202.
63. Bygren LO, Tinghög P, Carstensen J, Edvinsson S, Kaati G, Pembrey ME, et al. Change in paternal grandmother's early food supply influenced cardiovascular mortality of the female grandchildren. *BMC Genet*. 2014;15:12.
64. Henry LA, Cassidy T, McLaughlin M, Pentieva K, McNulty H, Walsh CP, et al. Folic acid supplementation throughout pregnancy: psychological developmental benefits for children. *Acta Paediatr Int J Paediatr*. 2018;107:1370–1378.
65. Molloy AM, Scott JM. Microbiological assay for serum, plasma, and red cell folate using cryopreserved, microtiter plate method. *Methods Enzymol*. 1997;281:43–53.
66. Kelleher BP, Broin SD. Microbiological assay for vitamin B12 performed in 96-well microtitre plates. *J Clin Pathol*. 1991;44:592–5.

67. Frosst P, Blom HJ, Milos R, Goyette P, Sheppard CA, Matthews RG, Boers GJ, den Heijer M, Kluijtmans LA, van den Heuvel LP. *Nat Genet.* 1995;10(1):111-3.
68. Teschendorff AE, Marabita F, Lechner M, Bartlett T, Tegner J, Gomez-Cabrero D, et al. A beta-mixture quantile normalization method for correcting probe design bias in Illumina Infinium 450 k DNA methylation data. *Bioinformatics.* 2013;29:189–96.
69. Leek JT, Johnson WE, Parker HS, Jaffe AE, Storey JD. The SVA package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics.* 2012;28:882–3.
70. Giardine B, Riemer C, Hardison RC, Burhans R, Elnitski L, Shah P, et al. Galaxy: a platform for interactive large-scale genome analysis. *Genome Res.* 2005;15:1451–5.
71. Mackin SJ, O'Neill KM, Walsh CP. Comparison of DNMT1 inhibitors by methylome profiling identifies unique signature of 5-aza-2'deoxyctidine. *Epigenomics.* 2018;10(8):1085–101.
72. Makambi KH. Weighted inverse chi-square method for correlated significance tests. *J Appl Stat.* 2003;30:225–34.

Ready to submit your research? Choose BMC and benefit from:

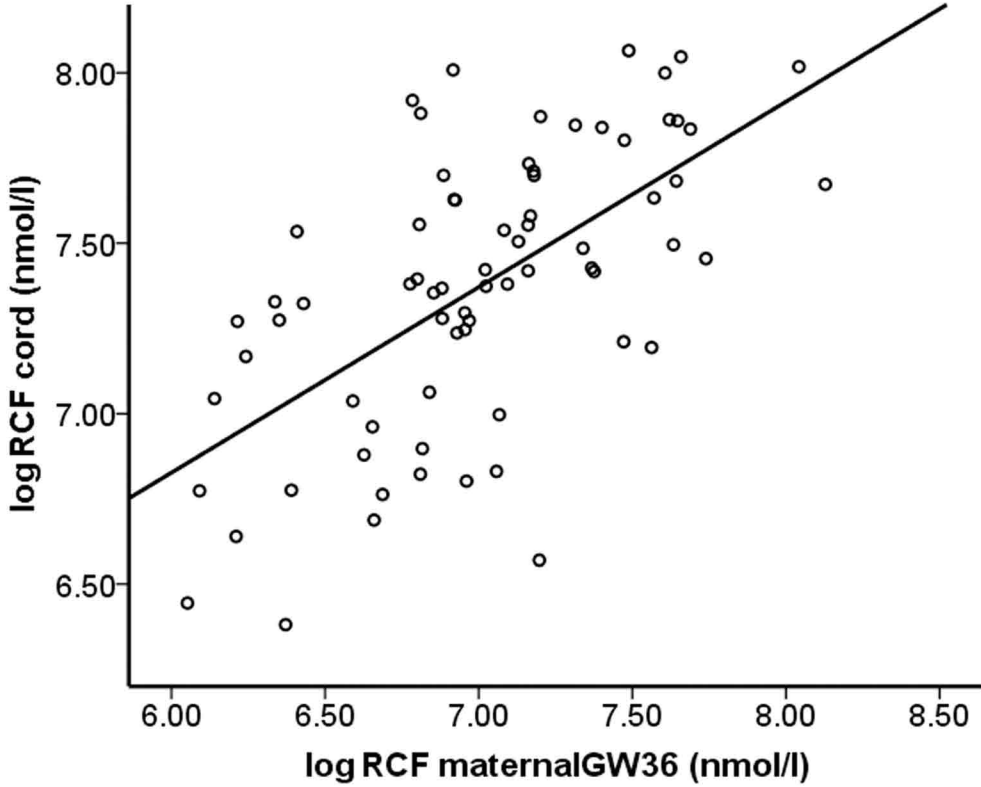
- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

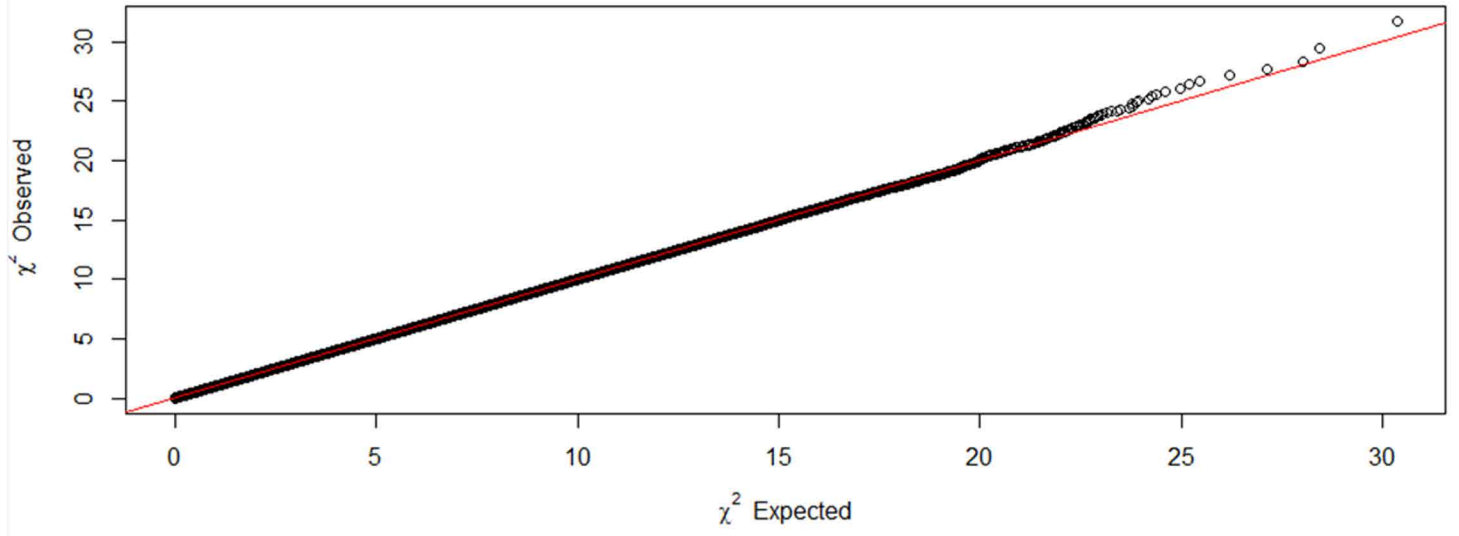
Learn more [biomedcentral.com/submissions](https://www.biomedcentral.com/submissions)



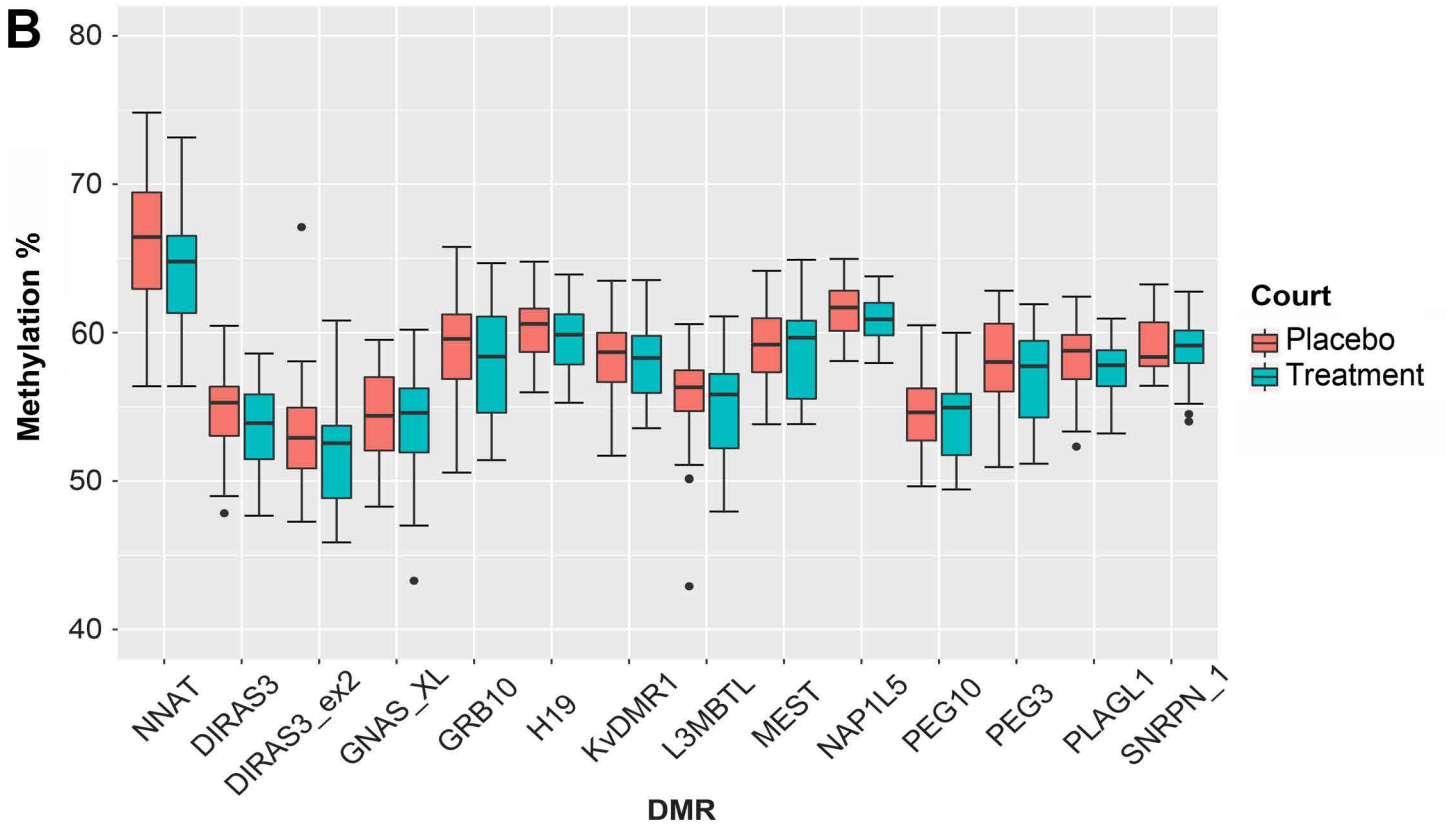
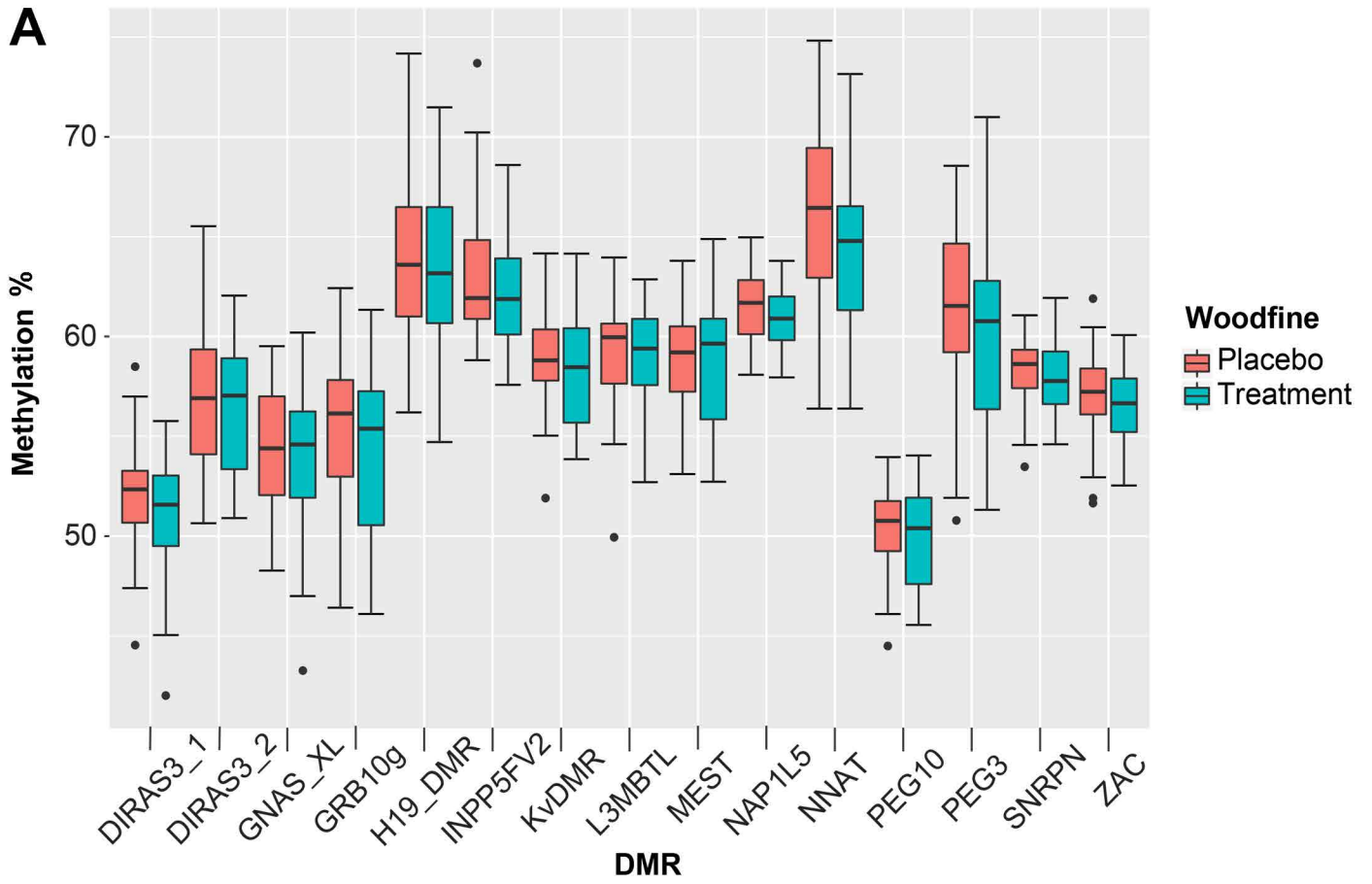
# Suppl.Fig.1



# Suppl. Fig.2



# Suppl.Fig.3



## Supplementary Table 1

Primer sequences used for pyrosequencing and transcriptional analysis

Application	Gene	Primer	Sequence 5'-3'
Pyrosequencing	<i>ZFP57</i>	FW	GGGATTTTTTTTAGTTATTGTTTTGTAT
		RV – 5'Btn	ACTAACAAACCCCTACTTTACCAAAC
		Seq	ATTGTTTTGTATTTATTTATTAGA
	<i>NXN</i>	FW – 5'Btn	TAGTAAAGTTTGGGGAAGG
		RV	ACACCATAAACTAAAACCAATCTAT
		Seq	CCATAAACTAAAACCAATCTATC
	<i>PRKAR1B</i>	FW	TTAGGGGTAGGTTTAGGTTTATAGT
		RV – 5'Btn	CCAACCTACCTACTAAACCTTATC
		Seq	GGTAGGTTTAGGTTTATAGTT
	<i>MIR4520A/B</i>	FW	GTTTAAATTTTTTTTTTGATTTGGATAGAAA
		RV – 5'Btn	AAAACATACCCTCAATTCCAAAAAAT C
		Seq	TTTTTTTTTGATTTGGATAGAAAATA
RT-qPCR/RT-PCR	<i>ZFP57</i>	FW	CCCAAACACAGAAGGCCTTT
		RV	GGTCCTGTCCATAGTCCCAG
	<i>ACTB</i>	FW	GGACTTCGAGCAAGAGATGG
		RV	AGCACTGTGTTGGCGTACAG
	<i>HPRT</i>	FW	AGCCCTGGCGTCGTGATTAGT
		RV	CCCGTTGAGCACACAGAGGCCTA

Based on Human Genome Build 19; all primers listed 5' to 3'.

Abbreviations: Btn, Biotinylated

## 5.0 PAPER-IV

### **Methylome profiling of young adults with depression supports a link with immune response and psoriasis**

Coral R. Lapsley, Rachelle Irwin, Margaret McLafferty, Sara-Jayne Thursby, Siobhan M. O'Neill, Anthony J. Bjourson, Colum P. Walsh, Elaine K. Murray

The main aims of this paper were to:

- Assess the genome wide effects of depression and self-harm or suicide attempt on the DNA methylation profiles of registering first year university students
- Investigate whether there are any gene classes preferentially effected by this disorder
- Compare the results obtained to that of similar studies into depression and DNA methylation

### **CONTRIBUTION**

For this paper, I independently conducted EPIC array analysis of the data in *RnBeads*, *Limma* and *ChAMP*. I created absolute beta and delta tracks of the EPIC array results of this study on UCSC genome browser. Further, I conducted the candidate gene analysis of the LCE cluster using the improved CandiMeth workflow. I also computed the CNV analysis for each subject per chromosome to check for deletions or variation in copy number. Here too, I computed a QQ plot for the study from the results of the EPIC array to check for population stratification effects. Finally, I compared the array results of this study to that of (Murphy et al., 2017) using effect size and Cohen's D test.

RESEARCH

Open Access

# Methylome profiling of young adults with depression supports a link with immune response and psoriasis



Coral R. Lapsley<sup>1†</sup>, Rachele Irwin<sup>2†</sup>, Margaret McLafferty<sup>1,3</sup>, Sara Jayne Thursby<sup>2</sup>, Siobhan M. O'Neill<sup>3</sup>, Anthony J. Bjourson<sup>1</sup>, Colum P. Walsh<sup>2</sup> and Elaine K. Murray<sup>1\*</sup> 

## Abstract

**Background:** Currently the leading cause of global disability, clinical depression is a heterogeneous condition characterised by low mood, anhedonia and cognitive impairments. Its growing incidence among young people, often co-occurring with self-harm, is of particular concern. We recently reported very high rates of depression among first year university students in Northern Ireland, with over 25% meeting the clinical criteria, based on DSM IV. However, the causes of depression in such groups remain unclear, and diagnosis is hampered by a lack of biological markers. The aim of this exploratory study was to examine DNA methylation patterns in saliva samples from individuals with a history of depression and matched healthy controls.

**Results:** From our student subjects who showed evidence of a total lifetime major depressive event (MDE,  $n = 186$ ) we identified a small but distinct subgroup ( $n = 30$ ) with higher risk scores on the basis of co-occurrence of self-harm and attempted suicide. Factors conferring elevated risk included being female or non-heterosexual, and intrinsic factors such as emotional suppression and impulsiveness. Saliva samples were collected and a closely matched set of high-risk cases ( $n = 16$ ) and healthy controls ( $n = 16$ ) similar in age, gender and smoking status were compared. These showed substantial differences in DNA methylation marks across the genome, specifically in the late cornified envelope (LCE) gene cluster. Gene ontology analysis showed highly significant enrichment for immune response, and in particular genes associated with the inflammatory skin condition psoriasis, which we confirmed using a second bioinformatics approach. We then verified methylation gains at the LCE gene cluster at the epidermal differentiation complex and at *MIR4520A/B* in our cases in the laboratory, using pyrosequencing. Additionally, we found loss of methylation at the *PSORSC13* locus on chromosome 6 by array and pyrosequencing, validating recent findings in brain tissue from people who had died by suicide. Finally, we could show that similar changes in immune gene methylation preceded the onset of depression in an independent cohort of adolescent females.

(Continued on next page)

\* Correspondence: [e.murray@ulster.ac.uk](mailto:e.murray@ulster.ac.uk)

†Coral R. Lapsley and Rachele Irwin contributed equally to this work.

<sup>1</sup>Northern Ireland Centre for Stratified Medicine, School of Biomedical Sciences, Ulster University, C-TRIC, Altnagelvin Hospital, Derry/Londonderry, UK

Full list of author information is available at the end of the article



© The Author(s). 2020 **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated in a credit line to the data.



(Continued from previous page)

**Conclusions:** Our data suggests an immune component to the aetiology of depression in at least a small subgroup of cases, consistent with the accumulating evidence supporting a relationship between inflammation and depression. Additionally, DNA methylation changes at key loci, detected in saliva, may represent a valuable tool for identifying at-risk subjects.

**Keywords:** Depression, Suicide, DNA methylation, Inflammation, Epidermal differentiation complex, Psoriasis

## Background

Depression is a highly prevalent, complex mental health disorder characterised by a range of debilitating symptoms. It affects over 300 million people globally [1] and is responsible for more years lost to disability (YLD) than any other condition, with a total of 76.4 million YLD [2]. Mental health problems, including depression, often emerge before age 18 with the period from 18 to 25 having been highlighted as a susceptible time in a person's life [3]. In particular, high prevalence rates of mental health problems and suicidality have been found among university students [4, 5]. Northern Ireland (NI) has one of the highest incidences of mental illness in Western Europe [6] and the highest rate of suicide in the UK, a rate which continues to increase [7]. The trans-generational impact of the years of conflict in NI have been mooted as one potential contributor to this [8]. We recently reported on prevalence rates of mental health disorders, self-harm and suicidality in a large cohort ( $n = 739$ ) of first year NI university students [9] and found that, consistent with the other recent studies [10–13], rates were high, with more than 50% of new undergraduate students reporting any lifetime mental disorder. Rates of depression and suicidal ideation were particularly high (24.2% and 31.0% respectively). Consistent with other studies suggesting that self-harm is the strongest predictor of suicidal behaviour [14–17], 122/155 individuals who self-harmed (78.7%) reported suicidal ideation in our cohort [7]. These results highlighted the high incidence of co-occurring depression, self-harm and suicide amongst young people entering university in our study population.

The aetiology of depression is very complex, but epidemiological studies indicate that genetic and environmental interactions are both implicated in disease pathology [18–20]. There is a genetic component to the aetiology of psychiatric disorders, including depression, which has been demonstrated in twin and family studies [21] indicating up to 40% heritability. The most recent meta-analysis of genome-wide association studies (GWAS), including 246,363 cases of depression and 561,190 controls, identified 102 independent regions reaching genome-wide significance associated with depression, including genes and pathways involved in synaptic structure and neurotransmission [22]. In terms of

environmental causes, severe childhood adversity and trauma including both verbal and physical abuse, neglect and parental mental disorders are also major contributing factors to the development of mood disorders and suicidal behaviours [23–25]. Other childhood adversities of varying severity, e.g. parental loss, bullying and socio-economic status are all associated with increased incidence of depression in later life [24, 26].

There are several well-discussed theories of depression including the monoamine theory [27], HPA axis dysregulation in response to stress [28], and in particular the emerging role of inflammation [29, 30]. Epidemiological research indicates that up to 70% of individuals with autoimmune and inflammatory diseases, such as rheumatoid arthritis and heart disease, experience depression [31]. Chronic stress, a major risk factor for depression, can activate inflammatory response in both the periphery and CNS through the hypothalamic pituitary adrenal (HPA) axis [32]. Impaired negative feedback in the HPA axis resulting in high levels of cortisol lead to the production of pro-inflammatory cytokines, chemokines and acute phase proteins from macrophages through the activation of NF- $\kappa$ B [33, 34]. Peripheral inflammatory signals are detected by microglia in the brain, which then initiate their own inflammatory cascade through the activation of CNS cytokines, reactive oxygen species (ROS) and reactive nitrogen species (RNS) [35], ultimately leading to alterations in serotonin signalling and changes in mood.

Peripheral levels of pro-inflammatory cytokines IL-6 and TNF- $\alpha$  are elevated in depression patients who were SSRI resistant compared to patients with depression, but in remission whose cytokine levels were similar to matched healthy controls [36, 37]. In addition, IL-12 and IL-4 were found to decrease in patients receiving a course of sertraline treatment [38]. C-reactive protein (CRP) is elevated in peripheral blood from depressed patients and significantly decreased from baseline following successful treatment with the SSRI, sertraline [37], further support for the link between inflammation and depression and the potential use of immune markers to stratify patients.

Our understanding of these mechanisms on a molecular level however remains poor. The genes implicated in each of these are different. For example, polymorphisms

in components of the serotonin system such as the serotonin-transporter-linked *5-HTTLPR* involved with monoamine levels affect predisposition to anxiety and depression [39, 40]. In contrast, polymorphisms in stress-related genes such as 5-HT transporter and *CRF* may instead modify susceptibility to depression according to HPA axis models. More recently, there has been great interest in possible epigenetic rather than genetic changes to components in either the HPA axis, such as glucocorticoid receptor (GR), or the inflammation pathway.

Epigenetic modifications such as DNA methylation, in contrast to DNA polymorphisms, can be influenced by environmental factors and provides a potential mechanism through which life events such as childhood trauma and stress, major risk factors of depression, can lead to the biological, and ultimately behavioural, changes associated with depression [41]. Epigenetic mechanisms could therefore be a key mediator of the interplay between biological vulnerability and life events leading to the behavioural changes seen in depression. In this context, there has been much recent interest in methylation changes at the glucocorticoid receptor (GR) and at genes related to corticotropin releasing hormone action, all with potential roles of an HPA axis model [42–44]. In contrast, a recent study from Murphy and colleagues using the Illumina 450K Beadarray chip uncovered instead significant association between self-reported depression and methylation changes at genes related to immune function in peripheral blood samples, particularly the *LTB4R* and *TRIM39-RPP21* loci, [45]. An earlier study by the same team found significant methylation changes at the *PSORC13* gene, involved in the inflammatory skin condition psoriasis, in completed suicide cases [46]. Given the differing targets implicated in these studies, further work in additional cohorts could help to clarify the main pathways showing epigenetic changes and afford greater insight into potential mechanisms involved.

As indicated above, we reported high rates of depression as well as co-occurring self-harm and suicide risk in a cohort of university entrants [7, 9]. In this study, we wished to (1) investigate the potential external and internal drivers of depression with and without experience of self-harm and a suicide attempt in this cohort; (2) conduct an initial genome-wide screen for DNA methylation differences in a subset of cases with highest levels of risk; (3) verify methylation changes at top-ranking loci using a second method and (4) compare our findings to other recent work in the area.

We were able to confirm that a set of shared risk factors greatly increased the chances of co-occurring depression, self-harm and suicidal ideation. In an initial comparison of saliva samples from students displaying all three conditions and a closely-matched set of

controls, we identified significant enrichment for immune response genes among those showing differential methylation. Closer examination highlighted genes involved in psoriasis, including several novel targets (*LCE* and *MIR4520A/B*). All regions could be verified by pyrosequencing. With due consideration of the limitations of the study, these findings nevertheless suggest a significant link between psoriasis and depression, self-harm and suicidal risk that can be detected in peripheral tissues.

## Results

### Risk factors in the student population

In order to determine associations between socio-demographic variables and depression, suicidality and self-harm, logistic regression analysis was undertaken (Table 1). Of the total  $N=739$  students who completed the survey, 24.2% ( $n = 186$ ) showed evidence of a lifetime Major Depressive Event (MDE). We attempted to identify discrete groups within these MDE sufferers using stratification on the basis of risk factors. Several demographic risk factors were significantly correlated with depression, self-harm or suicide attempt, or a combination of these three. In particular, females were more likely to develop depression with comorbid suicide attempt and self-harm ( $OR = 3.082, p < 0.05$ ), in comparison with males. Older students ( $>21$  years old) were nearly twice as likely to have depression ( $OR = 1.921, p < .05$ ), compared to students under the age of 21. In contrast to heterosexual students, those who stated they were non-heterosexual were nearly four times more likely to have experienced depression with comorbid suicide attempt and self-harm ( $OR = 3.384, p < 0.05$ ). Interestingly, none of the extrinsic factors examined including finances, bullying or maltreatment, were significantly correlated with depression, self-harm and/or suicidality. However, in relation to intrinsic factors such as emotional regulation, students who indicated suppression were more likely to have depression with co-occurring suicide attempt and self-harm ( $OR = 1.128, p < 0.01$ ), compared to those who reported reappraisal, which was a protective factor ( $OR = 0.924, p < 0.01$ ).

All values represent odds ratio; *SH* self-harm; significant results in bold; \* $p < 0.05$ , \*\* $p < 0.01$

### Selection of cases and controls

The logistic regression results suggested a particularly high-risk subpopulation within our study, namely students reporting depression, self-harm and suicidal ideation, which might display epigenetic differences from healthy controls. Of the 739 fully completed Student Wellbeing survey responses a total of only 30 participants reported depression, self-harm and a suicide attempt. As age, gender and smoking status are known

**Table 1** Logistic regression analyses of correlates of depression

Demographics	Depression only (no SH/attempt)	Dep and self-harm and attempt	Dep and self-harm with no attempt	Dep and attempt with no self-harm
<i>N</i> = 739	( <i>n</i> = 92)	( <i>n</i> = 30)	( <i>n</i> = 51)	( <i>n</i> = 13)
Demographic risk factors				
Gender				
Female	1.127	<b>3.082*</b>	1.635	1.188
Male	1.0	1.0	1.0	1.0
Age				
21 and over	<b>1.921*</b>	1.161	0.796	3.996
Under 21	1.0	1.0	1.0	1.0
Sexuality				
Non-heterosexual	1.164	<b>3.384*</b>	1.076	1.872
Heterosexual	1.0	1.0	1.0	1.0
Extrinsic risk factors				
Finances				
Enough	0.638	0.939	1.014	1.363
Comfortable	0.816	0.540	0.618	0.000
Well to do	0.308	1.049	1.024	4.438
Poor	1.0	1.0	1.0	1.0
Bullying				
Physical bullying	1.023	1.172	0.825	1.454
Verbal bullying	1.134	1.169	1.319	1.037
Ignoring bullying	1.100	1.033	1.409	1.000
Cyber bullying	0.811	0.961	0.983	1.253
Maltreatment				
Physical Abuse	0.642	0.879	0.883	0.394
Emotional Abuse	1.334	1.540	1.195	1.588
Intrinsic risk factors				
Impulsivity	1.111	<b>1.659**</b>	1.129	0.892
Emotion regulation				
Reappraisal	1.010	<b>0.924**</b>	0.973	0.986
Suppression	<b>1.074**</b>	<b>1.128**</b>	1.016	1.043

confounders in DNA methylation analyses [47, 48], we chose 16 cases for which we could identify closely matched controls based on these criteria, where controls had no life-time history of mental health problems (Table 2). The average age in both groups was 23 years ( $\pm 5.4$ ), and they contained equal numbers of males (4) and females (12) each, as well as an identical spread of smoking status (Table 2). All cases had also experienced MDE in the 12 months prior to the survey. The average age of onset of depression was 14, average age of suicide attempt and onset of self-harm was 16 years. On average, cases had experienced a MDE for at least two weeks for an average of 7 years since onset. Saliva samples from these participants were collected and DNA isolated from the samples using standard protocols as described

under methods below. Following quality control checks on the DNA, it was then subjected to genome-wide methylation analysis using the Infinium Methylation EPIC 850K Beadchip array.

#### Gains in methylation at immune response genes

Principal component analysis of the methylation status of all 32 samples using the RnBeads analysis package [49] in RStudio demonstrated separation by gender which confirmed established sex differences in methylation (Fig. 1a), but not by smoking status or age (not shown), which indicated that these latter are not major confounding factors in this study. As a control, a quantile-quantile plot was carried out, which showed no evidence of stratification effects among the samples

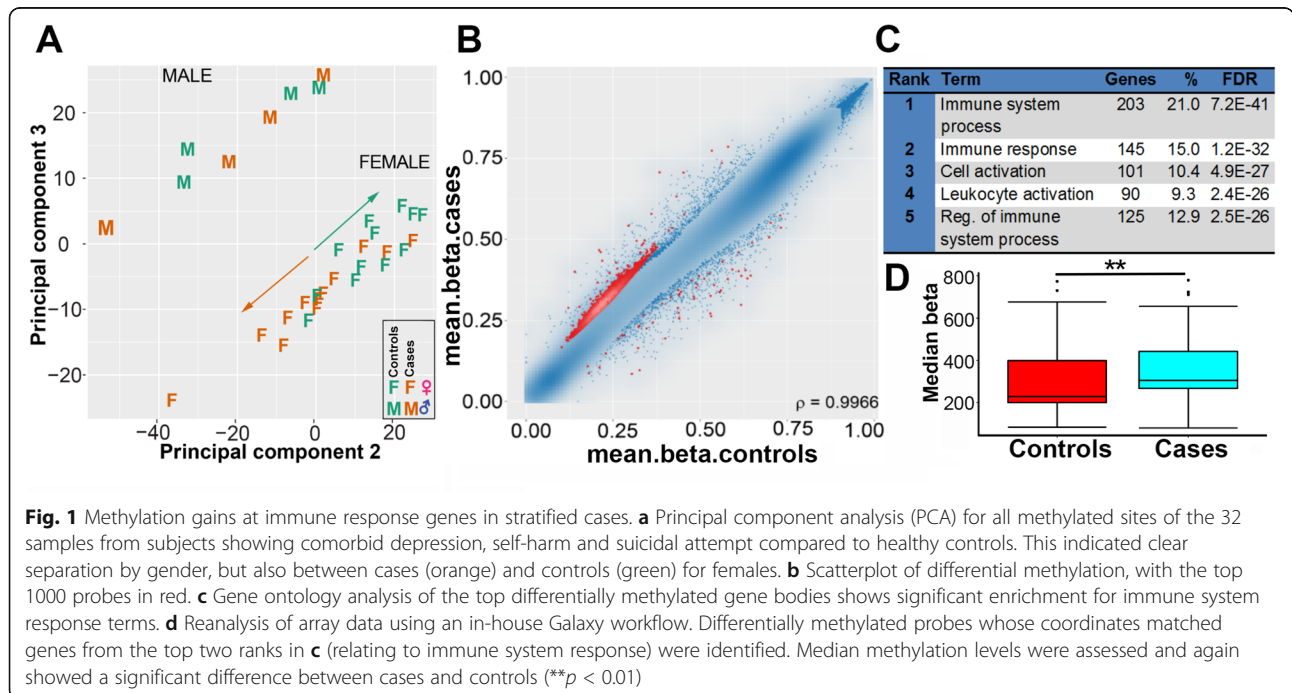
**Table 2** Characteristics of samples analysed by EPIC array

Demographics	Controls (n = 16)	Cases (n = 16)
Age, mean (range ± SD)	23 (18–32 ± 5.4)	23 (18–32 ± 5.0)
Gender		
Male (%)	4 (25)	4 (25)
Female (%)	12 (75)	12 (75)
Smoking status		
Past (%)	1 (6.2)	1 (6.2)
Daily (%)	6 (37.6)	6 (37.6)
Occasional (%)	1 (6.2)	1 (6.2)
Never (%)	8 (50)	8 (50)
Physical health		
Infectious (%)	0 (0)	0 (0)
Blood or immune (%)	0 (0)	1 (6.2)
Endocrine (%)	0 (0)	0 (0)
Eye or ear (%)	1 (6.2)	0 (0)
Neurological (%)	0 (0)	0 (0)
Heart or circulatory (%)	0 (0)	0 (0)
Respiratory (%)	0 (0)	2 (12.5)
Digestive (%)	0 (0)	2 (12.5)
Skin (%)	0 (0)	5 (31.3)
Musculoskeletal (%)	0 (0)	2 (12.5)

(Suppl. Fig.1). Among the female participants, there was also reasonable separation between cases and controls (Fig. 1a), with female cases clustered in the negative quartiles of the PCA and female controls towards the positive quartiles.

Given that females are at higher risk of co-occurring depression, self-harm and suicide attempt (Table 1) and the separation of cases and controls among females in the PCA (Fig. 1a), we concentrated further bioinformatics analysis on female samples. RnBeads used a combination of the difference in mean methylation (beta value), the quotient of mean methylation and the *p* value to rank the sites showing differential methylation, which we and others have found to be a more reliable indicator of biologically significant differences than *p* value alone, which often highlighted sites showing very small differences in methylation unlikely to be of functional significance. A scatterplot of the top 1000 ranked CpG sites with differential methylation between cases and controls in females displayed predominantly gains of methylation in the cases sample group (Fig. 1b).

In order to determine common features between the top ranking differentially methylated sites, gene ontology (GO) analysis for genes gaining methylation was carried out using DAVID software [50] which indicated strong enrichment scores for immune response terms (Fig. 1c). Common GO term for both promoters and genes included immune system process (GO:0002376), immune response (GO:0006955), cell activation (GO:0001775) and regulation of immune system process (GO:0002682), with very low predicted false discovery rates



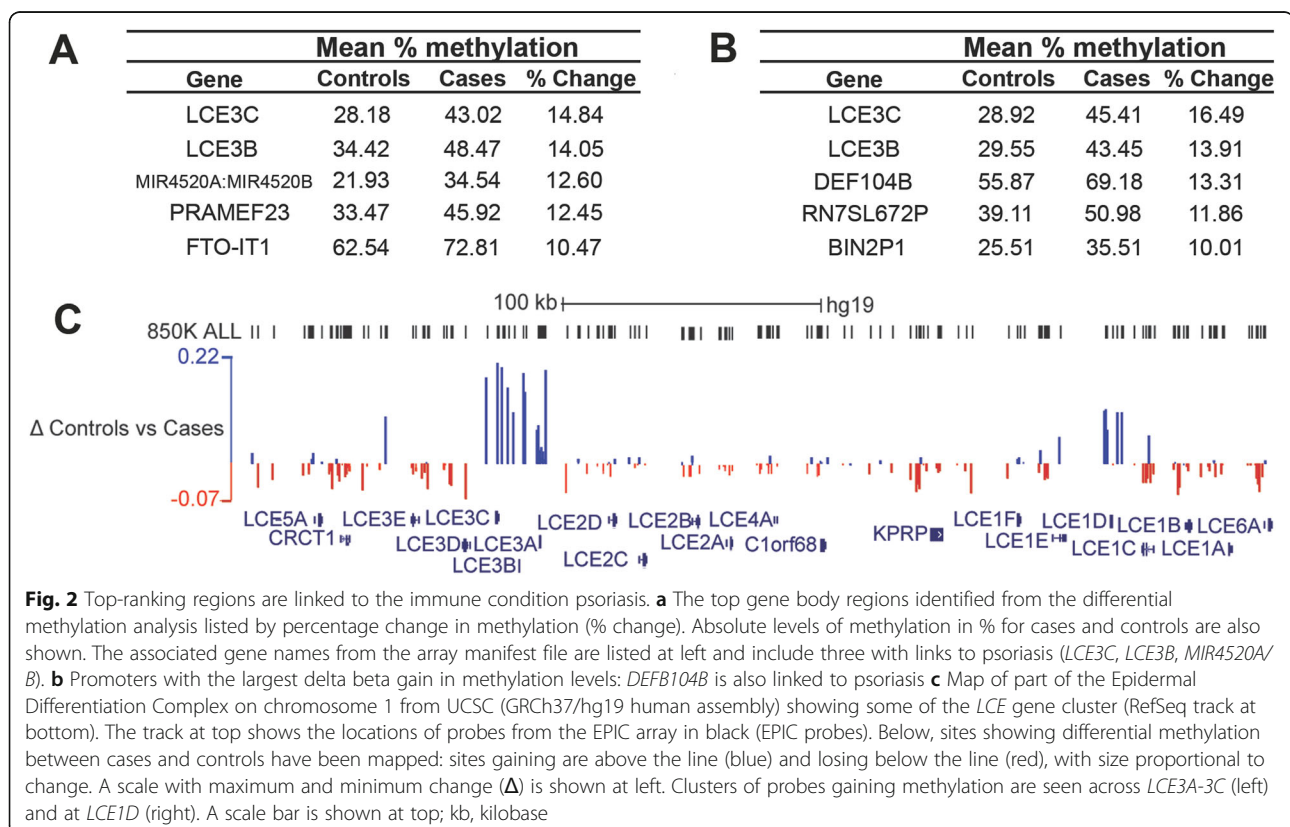
(FDR). Promoters and genes which showed loss in methylation were instead enriched for GO categories related to epidermal and keratin genes, but with higher FDR rates indicating lower likelihood of being true hits, most likely due to the smaller number of genes showing loss (data not shown).

To verify gains in methylation using a different bioinformatic approach, gene names common across the GO categories related to immune response were compiled (Table S1). We used an in-house developed workflow in Galaxy termed *CandiMeth* (Thursby and Walsh, in prep) to map differentially methylated probes to the human genome map, as previously described [51], found those which fell within promoters (defined as starting 500 bp 5' of the first exon) of these immune genes, and extracted the mean methylation values in cases and controls. Comparison of the average methylation across these genes between the two groups confirmed significant ( $p < 0.01$ ) gains in mean methylation levels in the cases relative to the controls (Fig. 1d).

#### Top-ranking regions include several loci linked with psoriasis and skin conditions

To examine more closely the immune response targets identified by the genome-wide scan, and to identify

those where methylation differences could be verified in the laboratory, we ranked the top hits showing gains in methylation by the magnitude of the difference in methylation ( $\Delta\beta$ ), both at gene bodies (Fig. 2a) and promoters (Fig. 2b). The Late Cornified Envelope-3C (LCE3C) and -3B (LCE3B) loci featured at the top of both lists and showed substantial gains in methylation ( $> 10\%$ ) in cases versus controls (Fig. 2a, b). These genes are part of a family which encode components of the stratum corneum of the skin and are thought to play a role in skin differentiation. The *LCE3* genes in particular have been linked to the development of psoriasis, a chronic inflammatory skin disease characterised by hyperproliferation of the epidermis and changes to keratinocyte differentiation [52]. Many of the *LCE* family members are clustered together on chromosome 1q21.3 in a region known as the epidermal differentiation complex (EDC) which contains multiple other genes expressed in the upper layers of the skin. A small deletion encompassing *LCE3B* and part of *LCE3C* (*LCE3C\_LCE3B-del*) is found in a substantial fraction of psoriasis sufferers [52–54]. As hemizyosity would affect methylation ratios, we checked for copy number variation (CNV) at this locus in our participants using the *R* package *DNAcopy*. No evidence for a CNV on chromosome

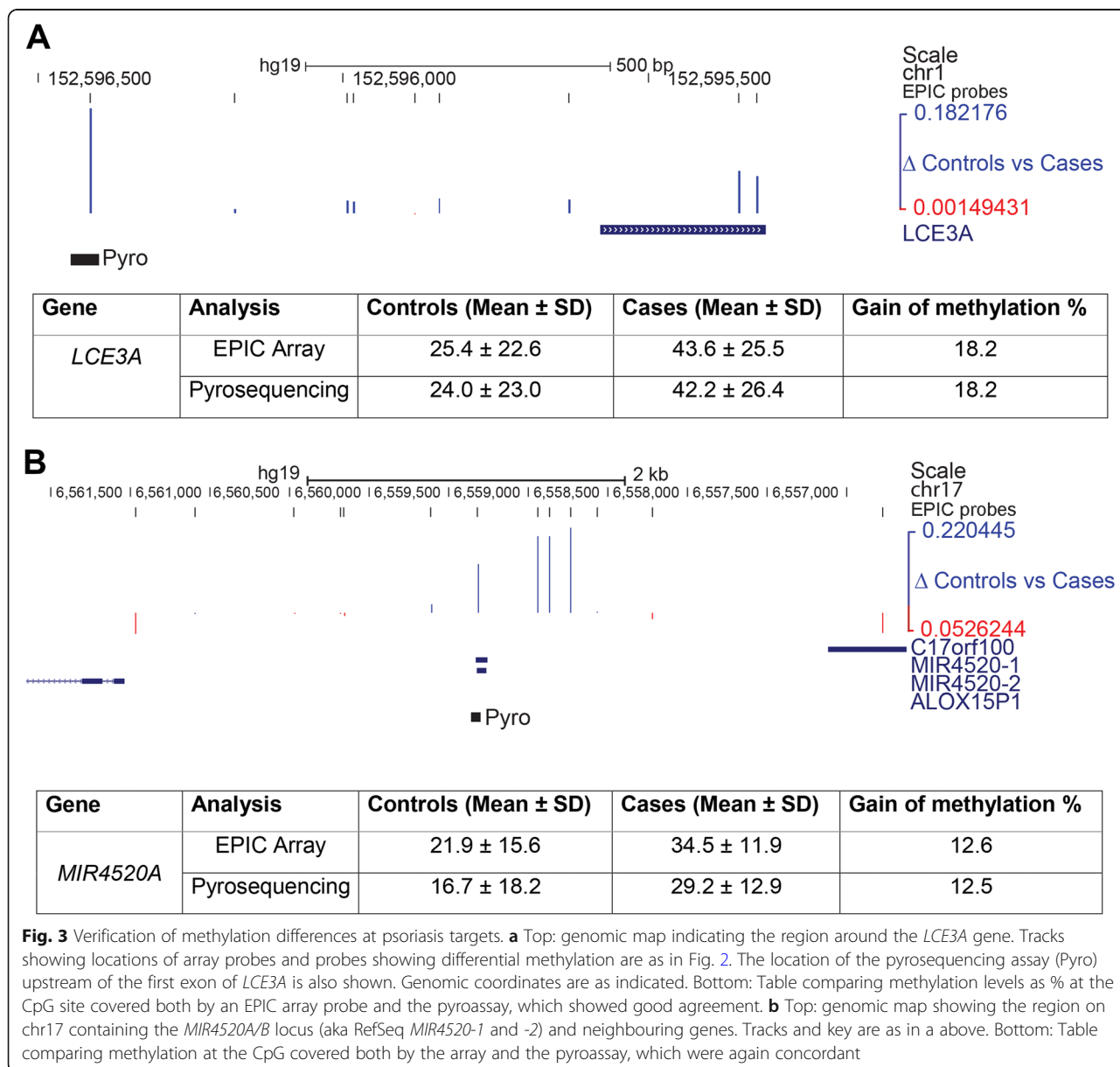


1q21 was found using this method, despite successfully detecting an isolated CNV in one patient on chromosome 5 (Fig.S2).

We mapped the differentially methylated sites identified in the screen to the human genome (hg19) and found clusters of sites comprising a differentially methylated region (DMR) not only at the *LCE3A-3B* locus, but also further along in the cluster at the *LCE1D-1C* locus (Fig. 2c), further supporting the association of depression with methylation changes at the *LCE* genes in this region. In order to confirm methylation differences at promoter regions, we used a commercially-available pyrosequencing assay to assess a site ~ 700 bp upstream of the *LCE3A* transcriptional start site (Fig. 3a). This site showed a gain of methylation of 18.2% in

cases compared to controls (42.2 vs 24.0), identical to that seen using the array (18.2% gain: 43.6% vs 25.4%).

Interestingly another locus the microRNA cluster *MIR4520A/B* on chromosome 17 (Fig. 2a), which was identified as a top hit by the genome-wide assay, has also been linked to psoriasis [55]. We also used a pyrosequencing assay (Fig. 3b) to assess the methylation difference at this region. While absolute levels of methylation at this site were lower by pyroassay than seen using the array (controls 16.7% pyro vs 21.95 array; cases 29.2% vs 34.5%), the approximate level of methylation is similar, and the gains in methylation seen between controls and cases was almost identical (12.5% pyroassay vs 12.6% array) (Fig. 3b).



**Validation of the *PSORS1C3* DMR**

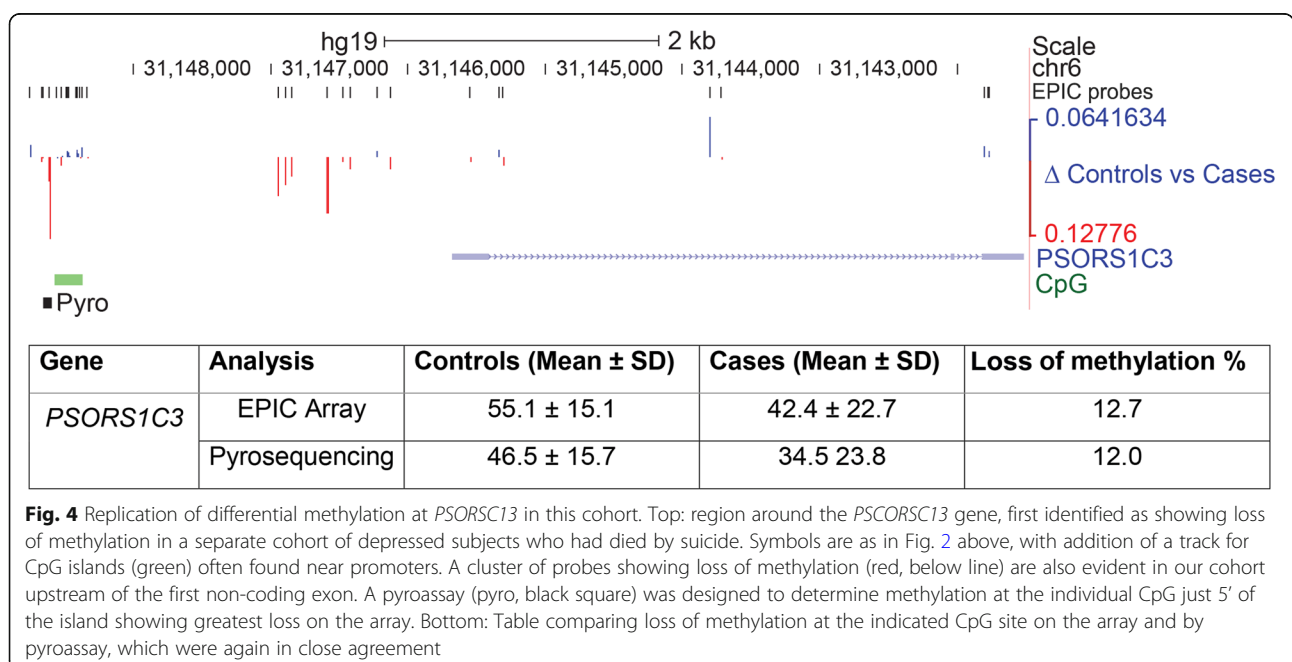
Recently an independent epigenetic screen by Murphy and colleagues (2017) also identified a link between psoriasis and depression. On comparing methylation in brain regions BA11 and BA25 of depression-suicide cases and normal controls, they identified Psoriasis Susceptibility 1 Candidate 3 (*PSORS1C3*) as one of the top hits [46]. A region comprising 12 CpG sites across the *PSORS1C3* locus showed loss of methylation (rather than gain) in depression-suicide cases compared to controls in that study. While this locus was not identified as a major target in our screen, there were a number of differences in sample size, tissue type and clinical diagnostics used which could account for this. An examination of the differentially methylated sites in our study confirmed however that there were a number of CpG showing loss of methylation immediately upstream of the *PSORS1C3* gene in an area likely to contain the promoter (Fig. 4a), with a maximum loss seen of 12.7%. To verify that there were differences in methylation between our cases and controls at this region, we designed a pyroassay for this site, which confirmed a loss of methylation. The results again showed good concordance both in direction (loss) and magnitude (12.0% vs 12.7%) between the pyroassay and the array (Fig. 4b).

**Alterations to immune genes precede development of depression in an independent cohort**

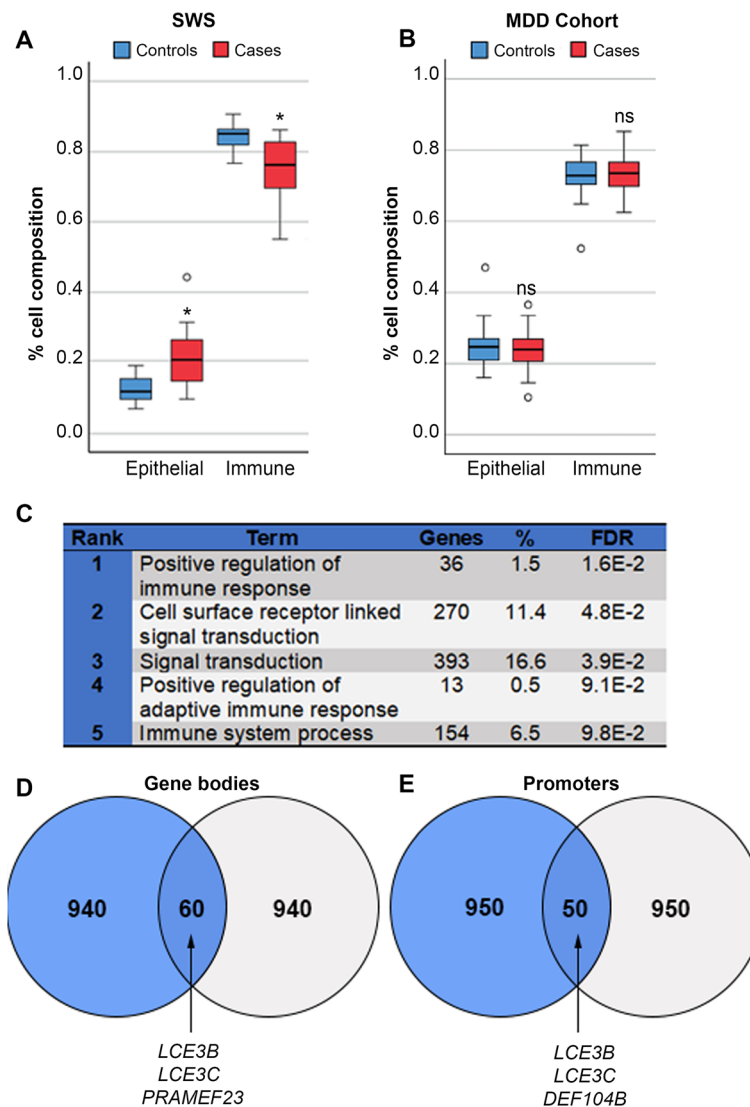
One possible complication with regards to working with saliva is that the samples may differ significantly in the ratios of different cell types present. While surrogate variable analysis as performed can compensate for such

hidden variables, an estimate of cell counts in the samples would be valuable. However, these could not be done on saliva, so we sought instead to estimate cell ratios using methods based on the array data alone, a so-called reference-free method. These methods are based on the observation that some methylation sites on the array show characteristic levels of methylation in specific tissues, independently of effects at other sites [56]. We employed the recently-developed EpiDISH algorithm [57] which is more suitable for saliva than the original methods developed for blood. As can be seen in Fig. 5a, EpiDISH indicated that the saliva samples from the Student Wellbeing Study cases had significantly different proportions of epithelial ( $p = 0.011$ , Kruskal-Wallis  $H = 6.16$ ) and immune cells ( $p = 0.013$ , Kruskal-Wallis  $H = 6.453$ ) from the controls. This suggested that some of the immune gene signature seen in our cohort may be due to differences in immune status at the time of sample collection in the cases versus controls. While this finding is valuable in itself as a biomarker, most of the changes seen in overall methylation and in specific classes of genes will be independent of the small number of sites used to identify tissue type. We wished therefore to further investigate if some of these methylation changes in immune genes may precede the overt changes in cell numbers and be linked to earlier stages in the development of depression.

A recently published study examined methylation patterns in female children, born to mothers with major depressive disorder (MDD), who were at increased risk of developing depression [58]. Saliva was collected from these girls ~ 13 years and analysed with the Illumina



**Fig. 4** Replication of differential methylation at *PSORS1C3* in this cohort. Top: region around the *PSORS1C3* gene, first identified as showing loss of methylation in a separate cohort of depressed subjects who had died by suicide. Symbols are as in Fig. 2 above, with addition of a track for CpG islands (green) often found near promoters. A cluster of probes showing loss of methylation (red, below line) are also evident in our cohort upstream of the first non-coding exon. A pyroassay (pyro, black square) was designed to determine methylation at the individual CpG just 5' of the island showing greatest loss on the array. Bottom: Table comparing loss of methylation at the indicated CpG site on the array and by pyrosequencing, which were again in close agreement



**Fig. 5** Alterations to immune genes precede development of depression in an independent cohort. **a** EpiDISH cell-type fraction estimation for epithelial and immune cells for the Student Wellbeing Study (SWS); differences were significant by Kruskal-Wallis test ( $*p < 0.05$ ). **b** EpiDISH analysis of saliva samples taken from at-risk adolescent girls prior to development of major depressive disorder (MDD; 58); ns, not significant, KW test. **c** Gene ontology analysis of the top 3000 ranked differentially methylated gene bodies from the MDD study shows significant enrichment for immune system response terms. **d** Top 1000 gene bodies gaining methylation in the SWS and the MDD cohorts showing overlap; top hits already identified in the SWS cohort are indicated. **e** Top 1000 promoter regions indicating shared targets between the two cohorts as in **d**

EPIC chip, then the children followed longitudinally, with many developing MDD. Due to the excellent match with phenotype, gender, tissue and method of assessment this is an ideal cohort for examining changes which may occur prior to the onset of depression. We obtained raw data from the authors and analysed them using the same pipeline described above. Looking at the children who went on to develop MDD, despite the fact that they show no sign of differences in cell counts between cases and controls (Fig. 5b), we found that many of the same top GO categories involved immune system response (Fig. 5c). Likewise, there was overlap between

many of the top-ranked gene bodies (Fig. 5d) and promoters (Fig. 5e) showing gain in methylation between the two studies, with many of the best hits (*LCE3B*, *LCE3C*, *PRAMEF23*, *DEF104B*) being common to both, supporting our theory that immune gene alterations may prefigure MDD.

## Discussion

Our previous work identified a high rate of depression among University students in Northern Ireland [9]. Here, we further analysed this cohort and found that while there were several extrinsic or intrinsic risk factors



which were significantly correlated with depression on its own, or depression and any one other feature, depressed students with both self-harm and suicide attempt formed a distinct sub-group with higher risk scores. Significant risk factors for this group included being female or non-heterosexual, and having higher impulsivity and emotional suppression with poorer re-appraisal ability. Hypothesizing that this much smaller sub-group might also show distinct epigenetic marks, we looked for differential methylation patterns in saliva samples from this group. Female cases separated from controls and showed overall tendency to gain in methylation. Methylation differences were significantly enriched in immune-related genes, with a number of top hits, including the *LCE3* genes and *MIR4520A/B*, being linked to the inflammatory skin condition psoriasis. We confirmed methylation differences at several loci by pyrosequencing. Additionally, the psoriasis gene *PSORSIC3*, recently identified as showing altered methylation in post-mortem brain from suicide completers, was also differentially methylated in our saliva samples. Finally, we saw alterations in methylation at some of the same immune-related genes in an independent cohort of teenage girls prior to onset of depression, suggesting these changes are occurring early in the etiology of the disease.

One of our major findings was the clear differential methylation between cases and controls we identified in immune-related genes. Depression has been previously linked with several chronic inflammatory conditions including diabetes, rheumatoid arthritis and cardiovascular disease [31, 59]. Inflammatory cytokines, IL-6, TNF- $\alpha$  and IFN- $\gamma$  are consistently upregulated in individuals with depression [32, 60] and many antidepressant medications have anti-inflammatory effects [61]. Furthermore, recent methylome analysis of whole blood also reported that depression-related methylation differences were enriched in pathways related to immune function [45], consistent with what we have identified for the first time in saliva. DMRs in immune response genes and their link to immune dysregulation warrant further investigation as potential biomarkers for depression.

Psoriasis is one of the most common inflammatory skin conditions, and affects up to 125 million people worldwide [62]. While being noncontagious and nonlethal, it nevertheless can be painful and disfiguring and can lead to severe disruptions in everyday social interactions and personal relationships. Psoriasis tends to develop between the ages of 15 and 25 and can lead to an impairment of social development due to attendant feelings of self-consciousness and embarrassment [63]. The average age of onset for psoriasis therefore shows notable overlap with that for depression, commonly occurring around 15 years old [3]. In a large UK population-based study individuals with severe psoriasis were also

reported to have an increased susceptibility to depression, anxiety and suicidality [64]. In that study, patients with psoriasis had a 39% increased chance of depression, which increased to 72% for severe psoriasis. Significantly, randomized controlled trials have shown that control of psoriasis symptoms can lead to improvements in psychological outcomes [65, 66]. Paediatric patients in particular have been shown to be at increased risk of developing depression [67]. Individuals with psoriasis display fatigue and sleep deprivation, which has been linked to the concomitant pruritis (itching) and pain and is linked to depression and obstructive sleep apnea in this group [68]. Insomnia has been recently highlighted as a particular risk factor for self-harm and suicide in university students [7, 69].

As indicated earlier, genes in the *LCE* cluster, and particularly *LCE3* homologues, have been strongly linked with psoriasis. The genes lie in the Epidermal Differentiation Complex (EDC) on chromosome 1, and genome-wide association studies (GWAS) have identified a major psoriasis susceptibility locus (*PSORS4*) in this region [70, 71]. A separate GWAS in the Chinese Han population identified two SNPs in the *LCE3A* gene, and three in *LCE3D* as particularly associated with psoriasis. The EDC also contains other skin genes such as *Filaggrin family member 2 (FLG2)* and *Cornulin*: interestingly experimental disruption of the skin barrier resulted in down-regulation of *LCE5A*, *LCE2B* and *FLG2* but upregulation of *LCE3A*, *Involucrin* and *Hornerin* [72]. *LCE3* genes show marked difference in expression between psoriatic lesions and normal skin, but not between prelesional skin and control [72–74], consistent with roles in skin repair rather than development per se. A small deletion encompassing parts of *LCE3C* and *LCE3B* (*LCE3C\_LCE3B-del*) has also been identified as a risk factor for psoriasis in a number of populations [52]. While we tested for CNV affecting this region in our cohort, we found no evidence for any deletions. The *LCE* cluster has also been reported to interact epistatically with the *PSORS1* locus at the HLA-complex on Chromosome 1 [53, 75].

Furthering the link with psoriasis, other top hits from our screen are also connected with this condition. The *MIR4520A/B* locus, which produces two microRNA *mir4520A* and *-B*, was also a top-ranking region from our screen. While little is currently known about this microRNA, next-generation sequencing of small RNAs from normal versus psoriatic skin highlighted *mir4520A* as one of the most abundant novel miRNA expressed in psoriatic skin [55] which was significantly downregulated in psoriatic lesions. Although it has not been firmly established whether transcription of this miR is epigenetically controlled, increased DNA methylation at this region may indicate downregulation of this gene in our

cohort. Another top hit was *DEFB104B*, which is a beta-defensin and part of a family of antimicrobial and cytotoxic peptides made by neutrophils. Defensins are expressed during inflammatory conditions, including psoriasis [76]. A total of 4 CpG sites were hypermethylated at the promoter regions of *DEFB104B* in our cases, suggesting perhaps a form of silencing or reduced mRNA expression and thus reduction/suppression of the innate immune response in depressive cases.

As detailed above, a number of loci involved in psoriasis were identified among our cases with depression and co-occurring self-harm and suicide attempt. Recently, Mill and colleagues (2017) compared methylation in two cortical brain regions from depressed suicide completers and non-psychiatric sudden-death controls and also identified a psoriasis susceptibility locus *PSORSIC3* as the main target affected [46]. They found a loss of methylation at a DMR upstream of the gene, which they verified by pyrosequencing and could replicate in a second set of suicide samples. While *PSORSCIC3* was not identified as a high-ranking DMR in our screen, there were a number of important differences between our studies: (1) we were using saliva samples not brain, and tissue type is known to have a major effect on methylation patterns [77] as seen even between brain regions [46]; (2) our samples were from subjects reporting suicide attempt at most, not completion and (3) there were a number of technical differences in diagnosis, processing and analysis, including the use of different chips (450K vs 850K EPIC). Nevertheless, on examination of the *PSORSIC3* locus in our cohort we found a DMR at this region which also showed loss of methylation in our samples (in contrast to the gains seen at other loci). We could also verify this using pyrosequencing, with good agreement in direction and magnitude of difference. These results are important as they (1) confirm methylation changes at this locus in another cohort displaying depression; (2) further link psoriasis, depression and suicidal thoughts and behaviour; (3) indicate that changes seen in the brain may also be mirrored in peripheral tissues such as saliva (4) suggest the change precedes death by suicide and therefore may have utility as a predictive tool.

The exact nature of the overlap between depression and psoriasis warrants further investigation. Traditionally, depression was thought to be a secondary consequence of living with a chronic physical condition such as psoriasis. However, the accumulating evidence that depression itself has an inflammatory component suggests that there may be common aetiology which can lead to mental health disorders, physical health problems, or both in a given individual. In our current sample, none of the depressed group reported psoriasis, indicating that while there is overlap in risk on a molecular level this does not necessarily manifest as co-occurrence of the two conditions.

While replication of the *PSORSC13* finding in saliva is encouraging, there has been debate over whether findings in peripheral tissue in general will parallel those in the organ most likely to be primarily involved, in this case the brain. A recent study evaluated DNA methylation patterns in the blood and saliva using the 450K BeadChip to assess the correlation of the two sample sources with secondary data from brain tissue. Although concordance was poor overall, methylation patterns in saliva were more similar to the brain methylome than blood [78]. The development of biomarkers that can be used to improve the diagnosis of depression, or those predictive of response to treatment, requires them to be easily accessible for sampling, so identification of reliable markers of depression in the periphery have more clinical utility than those in the brain. Saliva is a very promising potential biomarker discovery tissue due to the non-invasive sampling method. A concern is whether cell composition differences between cases and controls might be a confounding effect. The collection method utilised here involved the lysis of the cells so the specific cell types present could not be directly assessed. EpiDISH analysis indicated that the saliva samples from the cases had different proportions of both cell types from the controls, and that these were significant. Thus, it is likely that the methylation profile in part may reflect a difference in cell count in the cases vs controls. From the point of view of a biomarker, this is still a valuable finding as it can help to identify people with depression based on a heightened altered cell profile. On another level, we must consider that the surrogate variable analysis (SVA) and correction applied to our cohort will have accounted in part for this, suggesting that the immune gene hits highlighted in the analysis are genuine targets: these two levels of information are analogous to the cell count estimators such as EpiDISH, which can determine cell types independently of the top hits in the differential methylation analysis. Furthermore, immune genes were the top GO categories, and LCE and other immune genes the top hits, in the independent MDD cohort where there was no evidence of cell count skewing and prior to MDD onset, strongly supporting an underlying immune system link with depression.

The current exploratory study was carried out on a relatively small subgroup from the larger available student cohort as an initial investigation into the viability of using DNA methylation in saliva samples as potential biomarkers of depression. However, we have taken a stratified approach here in sub-classifying the cases of depression and, using logistic regression identified the small group of students who represent the most severe cases of depression with self-harm and suicide attempt. Our sample screened by array here represented more than half (16/30) of that high-risk subgroup: by limiting

the samples chosen we could more closely match these cases to controls with no history of mental disorders in terms of known confounders in methylation analysis, namely age, gender and smoking. The stratification approach may explain why clear differences were observed between the groups despite the small sample size overall. Significant hits were determined using a combined rank approach across adjacent sites, which took into account not only  $p$  value, but also magnitude and quotient for the changes in methylation, and is considered a more reliable indicator of biologically meaningful differences than  $p$  value alone [79]. Methyomic profiling of additional samples across a broad spectrum of individuals with depression are necessary to determine whether these changes are representative of depressive cohorts generally and to assess their utility as biomarkers.

## Conclusions

While this study is exploratory in nature, and has a number of caveats as indicated above, it nevertheless shows a novel linkage between epigenetic changes detected in saliva, and a particular category of depression with self-harm and suicidal attempt. Furthermore, our study clearly implicates changes at genes involved in the chronic inflammatory condition psoriasis, supporting emerging evidence from a number of epidemiological studies. Future work will include the analysis of a larger cohort if possible, as well as investigating specific intrinsic influences such as childhood adversity, and other clinical/phenotypic information. Additionally, it will be valuable to explore the potential mechanistic role of methylation in controlling transcription from these loci. Further analyses will also determine whether these markers have clinical utility in identifying or sub-classifying depression, or in predicting therapeutic response.

## Methods

### Ethics

Ethical approval was obtained from Ulster University Research Ethics Committee (REC/15/0004).

### Design

The Ulster University Student Wellbeing Study (UUSWS) has been described in detail elsewhere [7, 9] and was conducted as part of the WHO World Mental Health International College Student Project (WMH-ICS). The UUSWS study is being conducted as part of the WHO World Mental Health International College Student Project (WMH-ICS). An observational, longitudinal cohort study design is used for all studies. Prospective studies, such as this, can be very beneficial in that recall issues are minimised, sequences or patterns of events can be established and causal relationships may be inferred.

## Recruitment

All students commencing Ulster University in September 2015 were emailed a participant information sheet. First year students were recruited during registration where they gave written consent, provided a saliva sample and were given a unique, anonymous number to complete an online mental health survey clinically validated against the Diagnostic and Statistical Manual of Mental Disorders (DSM-IV) [80].

## Survey responses

The survey instrument was adapted from the WMH Composite International Diagnostic Interview (CIDI), version 3.0 [81], designed to be validated against the criteria of ICD-10 and DSM-IV disorders. Although these measures are self-report, good concordance has been found between the CIDI and clinical assessments [82]. Participants completed a section on emotional problems including depression, bi-polar disorder, anxiety, panic attacks or panic disorder and other serious emotional problems. Suicidal behaviour and non-suicidal self-injury (NSSI) questions were included from the Self-Injurious Thoughts and Behaviours Interview [83]. Impulsivity was measured by asking the participants if they often act without thinking, a Likert scale ranging from 1 'strongly agree' to 6 'strongly disagree' from the Student Experience and Student Expectations questionnaire [84]. Bullying was measured by asking participants how often they experienced the following: (1) you were bullied at school physically (i.e. repeatedly punched, shoved or physically hurt)? (2) You were bullied at school verbally (i.e. teased, called names). (3) You were bullied at school by someone who purposefully ignored you, excluded you, or spread rumours about you behind your back? You were bullied over the internet (e.g. Facebook, Twitter) or by text messaging? The questionnaire used a Likert scale ranging from 1 'very often' to 5 'never'. These questions were adapted from The Bully Survey [85]. Maltreatment was measured by asking participants how often they experienced the following using a Likert scale ranging from 1 'very often' to 5 'never': (1) physical abuse—you were physically abused at home; (2) emotional abuse—you were emotionally abused at home. The questions were adapted from the Adverse Childhood Experiences Scale [86]. Emotion regulation was measured using the Emotion Regulation Questionnaire, which consists of two dimensions of emotion regulation, reappraisal (six questions) and suppression (four questions), related to how well they control or manage their emotions. The instrument utilises a 7-point Likert scale. High scores for reappraisal are optimal while low scores for suppression indicate better emotion regulation strategies [87]. Logistic regression analysis was used to explore relationships between socio-demographic risk factors for individuals with depression, and comorbid suicidality and/or self-harm.

### Case selection

Cases ( $n = 16$ ) were selected from students who met the criteria for life-time (LT) major depressive episode, and who also reported suicide attempt and self-harm. Life-time depression is determined based on the response to seven questions (Likert scale) corresponding to DSM-IV criteria for depression. To calculate LT depression the first 6 symptoms/questions were recoded to 4 = “all or most of the time and 0 = none of the time, and summed. If at least 1 of the first 4 symptoms was “all or most of the time” and the sum of all six symptoms was at least 15 then participants met the criteria for depression. Suicidality, including thoughts, plans and attempts and self-harm, was assessed using items from the Self-Injurious Thoughts and Behaviour Interview [83]. If a participant responded yes to either of two questions asking about thoughts of hurting or killing themselves, or responded yes to direct questions on plan or attempt, they met the criteria for suicidal behaviour. Self-harm was assessed by asking the participant the following question: did you ever do something to hurt yourself on purpose, without wanting to die? (e.g. cutting yourself, hitting yourself, or burning yourself)? If they responded yes, they met the criteria for self-harm and were asked some further questions with regards to the number of times and what age this began. Healthy controls ( $n = 16$ ) were participants who reported no mental health problems, and strictly matched by age, gender and smoking status.

### Sample collection

Saliva samples were collected using Oragene OG-500 kits (DNA Genotek, Ontario Canada), enabling the self-collection and stabilisation of DNA at room temperature as per manufacturer guidelines.

### DNA extraction, bisulphite conversion and EPIC Beadchip Array

Saliva samples were incubated for 2 h at 56 °C, and DNA isolation carried out with PrepIT.L2P (DNA Genotek Inc., Canada) as per the manufacturer’s instructions. The purity and integrity of the genomic DNA preparations were assessed by agarose gel electrophoresis, and the quantity of DNA was determined using Quant-IT PicoGreen dsDNA Assay Kit (Invitrogen, Paisley, UK). In preparation for DNA methylation analysis, 500 ng of DNA was bisulphite converted using the EZ DNA Methylation Kit (Zymo Research, CA, USA) according to manufacturer’s instructions. Genome-wide DNA methylation profiles were generated using the Infinium Methylation EPIC Beadchip Array, and the Beadchip images captured using an Illumina iScan (Cambridge Genomic Services, Cambridge, UK) for matched cases ( $n = 16$ ) and controls ( $n = 16$ ).

### Bioinformatic analysis

Data was analysed using the *RnBeads* package (version 1.6.1) [49] on the freely available statistical software platform *R* (version 3.1.3). All samples passed quality control and were subjected to pre-processing, which involves filtering of probes and normalisation. Probes removed included those with a missing value (NA), probes at SNP-enriched sites, and bad quality probes determined by *greedy* algorithm. Background correction was carried out using *methylnumi.noob v2.32.0* [88] and the methylation values of the remainder probes were normalized using *bmiq* [89]. Copy number variation (CNV) was assessed using the *DNACopy* package v1.60.0 [90]. In order to account for any hidden confounding variables in the dataset, surrogate variable analysis was carried out using the *limma* method [91]. The methylation intensities for each probe, representing a CpG site, were represented as  $\beta$  values (ranging from 0, unmethylated, to 1, fully methylated) and these were plotted against genomic loci (based on Human Genome Build 19) using an in-house developed workflow in GALAXY v19.01 (<https://usegalaxy.org/>) [92] called *CandiMeth* (Thursby and Walsh, in prep) in order to visualise changes in DNA methylation in UCSC (<https://genome.ucsc.edu/>) and quantify differences across specific genomic intervals. Subsequent gene ontology (GO) analyses were performed using DAVID v6.7 (<https://david.ncifcrf.gov/>) [50]. Cell type composition estimation was performed in RStudio using EpiDISH v2.2.2 [57].

### Pyrosequencing

Bisulphite pyrosequencing was carried out in order to verify changes in methylation at loci of interest from the Infinium MethylationEPIC Beadchip Array. Primers spanning the probes of interest from the array were designed using the PyroMark Assay Design Software 2.0 (Qiagen, Manchester UK). Bisulfite-treated DNA was PCR-amplified using the PyroMark PCR kit (Qiagen, Hilden, Germany) according to manufacturer’s instruction. The primer sequences and PCR conditions are summarized in Supplementary Table 1. Amplification was carried out as follows: 95 °C for 15 min, followed by 45 cycles of 95 °C for 30 s, 56 °C for 30 s, and 72 °C for 30 s, with a final elongation step at 72 °C for 10 min. Pyrosequencing was performed as per manufacturer’s instructions on the PyroMark Q24 system (Qiagen, Hilden, Germany), and methylation levels were analysed using PyroMark Q24 1.010 software (Qiagen, Hilden, Germany).

### Statistical analysis

Differential methylation analysis was conducted on site and region level for healthy controls and cases samples. The normalized  $\beta$  values of the Infinium MethylationEPIC Beadchip Array data were converted into  $M$  values

( $M = \log_2(\beta/(1-\beta))$ ) and differential methylation between samples (cases vs. healthy controls) was estimated with hierarchical linear models using limma. On the region level (i.e. genes, promoters, CpG islands), differential methylation was computed based on the average difference in means across all sites in a specified region of the sample groups and the mean of quotients in mean methylation as well as a combined p-value, which was calculated from all site p-values in the region using a generalization of Fisher's method [93]. In addition, each region was assigned a rank based on each of these criteria. The smaller the combined rank for a region, the more evidence for differential methylation it exhibits.

The top 1000 ranking genes of each region was input into DAVID. We used DAVID software to determine significance of each gene ontology category, calculated using a modified Fisher's exact test (EASE score) which was FDR-corrected. Pyrosequencing data were analysed using Student's *t* test to identify statistical differences between cases and healthy controls. EpiDISH data was analysed using Kruskal-Wallis test to identify statistical differences between cell type composition for cases and controls in each cohort. A *p* value < 0.05 was considered significant.

### Supplementary information

**Supplementary information** accompanies this paper at <https://doi.org/10.1186/s13148-020-00877-7>.

**Additional file 1: Table S1.** Pyrosequencing primers.

**Additional file 2: Table S2.** Immune-related genes analysed for gains in methylation at promoter.

**Additional file 3: Figure S1.** Absence of population substructure effects. A quantile–quantile (QQ) plot showing observed vs. expected  $-\log_{10}(p)$  values for association at all CpG sites. The x-axis shows the expected  $-\log_{10}(p)$  value, the y-axis the observed  $-\log_{10}(p)$  value; the red line indicates the expected distributions under the null hypothesis and the black dots were the observed values. A close match at lower significance values indicated no systematic inflation of *P* was seen due to unaccounted-for stratification effects.

**Additional file 4: Figure S2.** Absence of deletions or duplications at top differentially methylated loci. EPIC array probe data was analysed using the *DNACopy* package in R to look for variations indicating copy number variation (CNV): an example output plot from subject 225 (Healthy Control) is shown. Probe index number is shown along the x-axis, while gain/loss in copy number, expressed as the log<sub>10</sub> ratio, is shown on the Y-axis; dots coming away from the line indicate probes showing gains or losses of signal consistent with regional duplications/deletions. No significant CNVs were detected in the Epidermal Differentiation Complex (EDC) region on chromosome 1q21, but the approach successfully detected a CNV on chromosome 5 in one participant (arrow at right) not overlapping any of the differentially methylated regions, shown here as a positive control for sensitivity. *DNACopy* plots were carried out for all samples and failed to detect copy number variation (CNV) at other top hits.

### Abbreviations

MDE: Major Depressive Episode; LCE: Late cornified envelope; MIR4520A/B: MicroRNA 4520 A/B; PSORSC13: Psoriasis Susceptibility 1 Candidate 3; YLD: Years lost to disability; 5-HTTLPR: Serotonin-transporter-linked receptor; 5-HT: 5-HT transporter; CRF: Corticotropin-releasing factor; GR: Glucocorticoid

receptor; LTB4R: Leukotriene B4 Receptor; TRIM39-RPP21: Tripartite motif-containing 39 - Ribonuclease P Protein Subunit P21; SH: Self-harm; CpG: Cytosine guanine dinucleotide; GO: Gene Ontology; EDC: Epidermal differentiation complex; CNV: Copy Number Variation; DMR: Differentially methylated region; IL6: Interleukin 6; TNF- $\alpha$ : Tumour Necrosis Factor alpha; IFN- $\gamma$ : Interferon Gamma; PSORS4: Psoriasis susceptibility 4; FLG2: Filaggrin family member 2; HLA-complex: The human leukocyte antigen complex; DEFB104B: Defensin Beta 104B; CID: Composite International Diagnostic Interview; NSSI: Non-suicidal self-injury; LT: Lifetime

### Acknowledgements

We thank the Northern Ireland Centre for Stratified Medicine and School of Psychology staff and postgraduate students for assistance with the recruitment during registration week. We also thank the Students Union, Student Support and Student Administration Services at Ulster University, and Inspire Students, for their assistance in this study. We would also like to thank Dr. Randy Auerbach and his team at Harvard University for their assistance in data management and analysis. We would like to thank Dr Kathryn Humphreys, Sarah Moore and Prof Michael Kobor for kindly providing raw data files from their recently published study as an independent validation cohort [60]. In particular, we would also like to thank all the participants who took part in this study.

### Authors' contributions

C.L, M.McL, S.ON, A.JB and E.M were involved in the design and participant recruitment of the Ulster University Student Wellbeing Study. M.McL examined the initial survey information and carried out logistic regression analysis. C.L selected the samples for DNA analysis and C.L and R.I prepared the DNA samples for the Illumina Infinium Methylation EPIC array. R.I and S.J.T analysed the array data and performed the gene ontology analysis, supervised by CPW. Validation pyrosequencing was carried out by C.L, with R.I's support with primer design. C.P.W. and E.M supervised the methylation analysis overall and C.L., R.I., C.P.W. and E.M. wrote the manuscript. All authors read and approved the final version.

### Funding

Work was supported by a programme grant jointly from the European Union Regional Development Fund (ERDF) EU Sustainable Competitiveness Programme for N. Ireland/NI Public Health Agency (HSC R&D) and Ulster University (A.JB, EM) as well as PhD awards from the Department of Employment & Learning (CL, MMcL, SJT), an Economic and Social Research Council/Biotechnology and Biological Sciences Research Council (ESC/BSRC) joint grant ES/N000323/1 (CPW) and an NI Clinical Research Facility grant (EM,CPW).

### Availability of data and materials

The datasets used and/or analysed during the current study are available from the corresponding author on reasonable request.

### Ethics approval and consent to participate

Ethical approval was obtained from Ulster University Research Ethics Committee (REC/15/0004), and participants provided written consent.

### Consent for publication

No individual participant data is presented in this paper, therefore consent for publication is not applicable.

### Competing interests

Not applicable

### Author details

<sup>1</sup>Northern Ireland Centre for Stratified Medicine, School of Biomedical Sciences, Ulster University, C-TRIC, Altnagelvin Hospital, Derry/Londonderry, UK. <sup>2</sup>Genomics Medicine Research Group, School of Biomedical Sciences, Ulster University, Coleraine Campus, Coleraine, UK. <sup>3</sup>School of Psychology, Ulster University, Coleraine Campus, Coleraine, UK.

Received: 30 September 2019 Accepted: 28 May 2020

Published online: 15 June 2020

## References

- World Health Organisation. Depression and other common mental disorders: global health estimates. Geneva; 2017. Report No.: CC BY-NC-SA 3.0 IGO.
- Smith K. Mental health: a world of depression. *Nature*. 2014;515(7526):181.
- Kessler RC, Amminger GP, Aguilar-Gaxiola S, Alonso J, Lee S, Ustun TB. Age of onset of mental disorders: a review of recent literature. *Curr Opin Psychiatry*. 2007;20(4):359–64.
- Mortier P, Cuijpers P, Kiekens G, Auerbach RP, Demyttenaere K, Green JG, et al. The prevalence of suicidal thoughts and behaviours among college students: a meta-analysis. *Psychol Med*. 2018;48(4):554–65.
- Mortier P, Demyttenaere K, Auerbach RP, Cuijpers P, Green JG, Kiekens G, et al. First onset of suicidal thoughts and behaviours in college. *J Affect Disord*. 2017;207:291–9.
- Bunting BP, Murphy SD, O'Neill SM, Ferry FR. Lifetime prevalence of mental health disorders and delay in treatment following initial onset: evidence from the Northern Ireland study of health and stress. *Psychol Med*. 2012;42(8):1727–39.
- O'Neill S, McLafferty M, Ennis E, Lapsley C, Bjourson T, Armour C, et al. Socio-demographic, mental health and childhood adversity risk factors for self-harm and suicidal behaviour in college students in Northern Ireland. *J Affect Dis*. 2018;239:58–65.
- O'Connor RC, O'Neill SM. Mental health and suicide risk in Northern Ireland: a legacy of the troubles? *Lancet Psychiatry*. 2015;2(7):582–4.
- McLafferty M, Lapsley CR, Ennis E, Armour C, Murphy S, Bunting BP, et al. Mental health, behavioural problems and treatment seeking among students commencing university in Northern Ireland. *PLoS One*. 2017;12(12):e0188785.
- Eisenberg D, Gollust SE, Golberstein E, Hefner JL. Prevalence and correlates of depression, anxiety, and suicidality among university students. *Am J Orthop*. 2007;77(4):534–42.
- Bewick B, Koutsopoulou G, Miles J, Slaa E, Barkham M. Changes in undergraduate students' psychological well-being as they progress through university. *Stud High Educ*. 2010;35(6):633–45.
- American College Health Association. American college health association-National College Health Assessment II: fall 2015 reference group undergraduates executive summary. MD: Hanover; 2016.
- Castillo LG, Schwartz SJ. Introduction to the special issue on college student mental health. *J Clin Psychol*. 2013;69(4):291–7.
- Gunnell D, Hawton K, Ho D, Evans J, O'Connor S, Potokar J, et al. Hospital admissions for self harm after discharge from psychiatric inpatient care: cohort study. *BMJ*. 2008;337:a2278.
- Guan K, Fox KR, Prinstein MJ. Nonsuicidal self-injury as a time-invariant predictor of adolescent suicide ideation and attempts in a diverse community sample. *J Consult Clin Psychol*. 2012;80(5):842–9.
- Scott LN, Pilkonis PA, Hipwell AE, Keenan K, Stepp SD. Non-suicidal self-injury and suicidal ideation as predictors of suicide attempts in adolescent girls: a multi-wave prospective study. *Compr Psychiatry*. 2015;58:1–10.
- Ribeiro JD, Franklin JC, Fox KR, Bentley KH, Kleiman EM, Chang BP, et al. Self-injurious thoughts and behaviors as risk factors for future suicide ideation, attempts, and death: a meta-analysis of longitudinal studies. *Psychol Med*. 2015;46(2):225–36.
- Ptak C, Petronis A. Epigenetic approaches to psychiatric disorders. *Dialogues Clin Neurosci*. 2010;12(1):25–35.
- Bagot RC, Labonté B, Peña CJ, Nestler EJ. Epigenetic signaling in psychiatric disorders: stress and depression. *Dialogues Clin Neurosci*. 2014;16(3):281–95.
- Klengel T, Binder EB. Epigenetics of stress-related psychiatric disorders and gene x environment interactions. *Neuron*. 2015;86(6):1343–57.
- Sullivan PF, Neale MC, Kendler KS. Genetic epidemiology of major depression: review and meta-analysis. *Am J Psychiatry*. 2000;157(10):1552–62.
- Howard DM, Adams MJ, Clarke T-K, Hafferty JD, Gibson J, Shirali M, et al. Genome-wide meta-analysis of depression identifies 102 independent variants and highlights the importance of the prefrontal brain regions. *Nat Neurosci*. 2019;22(3):343–52.
- Dube SR, Anda RF, Felitti VJ, Chapman DP, Williamson DF, Giles WH. Childhood abuse, household dysfunction, and the risk of attempted suicide throughout the life span: findings from the adverse childhood experiences study. *Jama*. 2001;286(24):3089–96.
- Kessler RC, Davis CG, Kendler KS. Childhood adversity and adult psychiatric disorder in the US National Comorbidity Survey. *Psychol Med*. 1997;27(5):1101–19.
- McLafferty M, Armour C, McKenna A, O'Neill S, Murphy S, Bunting B. Childhood adversity profiles and adult psychopathology in a representative Northern Ireland study. *Journal of Anxiety Disorders*. 2015;35:42–8.
- Kessler RC, McLaughlin KA, Green JG, Gruber MJ, Sampson NA, Zaslavsky AM, et al. Childhood adversities and adult psychopathology in the WHO world mental health surveys. *Br J Psychiatry*. 2010;197(5):378–85.
- Iversen L. The Monoamine Hypothesis of Depression. *Biology of Depression: Wiley-VCH Verlag GmbH*; 2008. p. 71–86.
- Pariante CM, Lightman SL. The HPA axis in major depression: classical theories and new developments. *Trends Neurosci*. 2008;31(9):464–8.
- Miller AH, Maletic V, Raison CL. Inflammation and its discontents: the role of cytokines in the pathophysiology of major depression. *Biol Psychiatry*. 2009;65(9):732–41.
- Slavich GM, Irwin MR. From stress to inflammation and major depressive disorder: a social signal transduction theory of depression. *Psychol Bull*. 2014;140(3):774–815.
- Almond M. Depression and Inflammation: Examining the link: Inflammatory conditions may precipitate or perpetuate depression, but the precise relationship is unclear 2013;12:25-6-32.
- Raison CL, Capuron L, Miller AH. Cytokines sing the blues: inflammation and the pathogenesis of depression. *Trends Immunol*. 2006;27(1):24–31.
- Pace TWW, Hu F, Miller AH. Cytokine-effects on glucocorticoid receptor function: relevance to glucocorticoid resistance and the pathophysiology and treatment of major depression. *Brain Behav Immun*. 2007;21(1):9–19.
- Zunszain PA, Anacker C, Cattaneo A, Carvalho LA, Pariante CM. Glucocorticoids, cytokines and brain abnormalities in depression. *Prog Neuro-Psychopharmacol Biol Psychiatry*. 2011;35(3):722–9.
- Yirmiya R, Rimmerman N, Reshef R. Depression as a microglial disease. *Trends Neurosci*. 2015;38(10):637–58.
- Eller T, Vasar V, Shlik J, Maron E. Effects of bupropion augmentation on pro-inflammatory cytokines in escitalopram-resistant patients with major depressive disorder. *Journal of psychopharmacology (Oxford, England)*. 2009;23(7):854–8.
- O'Brien SM, Scott LV, Dinan TG. Antidepressant therapy and C-reactive protein levels. *Br J Psychiatry*. 2006;188:449–52.
- Sutcgil L, Oktenli C, Musabak U, Bozkurt A, Cansever A, Uzun O, et al. Pro- and anti-inflammatory cytokine balance in major depression: effect of sertraline therapy. *Clin Dev Immunol*. 2007;2007:76396.
- Collier DA, Stober G, Li T, Heils A, Catalano M, Di Bella D, et al. A novel functional polymorphism within the promoter of the serotonin transporter gene: possible role in susceptibility to affective disorders. *Mol Psychiatry*. 1996;1(6):453–60.
- Lesch K-P, Bengel D, Heils A, Sabol SZ, Greenberg BD, Petri S, et al. Association of anxiety-related traits with a polymorphism in the serotonin transporter gene regulatory region. *Science (New York, NY)*. 1996;274(5292):1527–31.
- Mitchell C, Schnepfer LM, Notterman DA. DNA methylation, early life environment, and health outcomes. *Pediatr Res*. 2016;79(1-2):212–9.
- Weaver IC, Meaney MJ, Szyf M. Maternal care effects on the hippocampal transcriptome and anxiety-mediated behaviors in the offspring that are reversible in adulthood. *Proc Natl Acad Sci U S A*. 2006;103(9):3480–5.
- Korosi A, Baram TZ. The central corticotropin releasing factor system during development and adulthood. *Eur J Pharmacol*. 2008;583(2-3):204–14.
- Jokinen J, Boström AE, Dadfar A, Ciuculete DM, Chazittzoffis A, Åsberg M, et al. Epigenetic changes in the CRH gene are related to severity of suicide attempt and a general psychiatric risk score in adolescents. *EBioMedicine*. 2017;27:123–33.
- Crawford B, Craig Z, Mansell G, White I, Smith A, Spaul S, et al. DNA methylation and inflammation marker profiles associated with a history of depression. *Hum Mol Genet*. 2018;27(16):2840–50.
- Murphy TM, Crawford B, Dempster EL, Hannon E, Burrage J, Turecki G, et al. Methyloomic profiling of cortex samples from completed suicide cases implicates a role for PSORS1C3 in major depression and suicide. *Transl Psychiatry*. 2017;7:e989.
- Zellinger S, Kuhnel B, Klopp N, Baurecht H, Kleinschmidt A, Gieger C, et al. Tobacco smoking leads to extensive genome-wide changes in DNA methylation. *PLoS One*. 2013;8(5):e63812.







48. Horvath S, Zhang Y, Langfelder P, Kahn RS, Boks MP, van Eijk K, et al. Aging effects on DNA methylation modules in human brain and blood tissue. *Genome Biol.* 2012;13(10):R97.
49. Assenov Y, Muller F, Lutsik P, Walter J, Lengauer T, Bock C. Comprehensive analysis of DNA methylation data with RnBeads. *Nat Methods.* 2014;11(11):1138–40.
50. Huang da W, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc.* 2009;4(1):44–57.
51. O'Neill KM, Irwin RE, Mackin SJ, Thursby SJ, Thakur A, Bertens C, et al. Depletion of DNMT1 in differentiated human cells highlights key classes of sensitive genes and an interplay with polycomb repression. *Epigenetics Chromatin.* 2018;11(1):12.
52. Shen C, Gao J, Yin X, Sheng Y, Sun L, Cui Y, et al. Association of the Late Cornified Envelope-3 genes with psoriasis and psoriatic arthritis: a systematic review. *Journal of Genetics and Genomics.* 2015;42(2):49–56.
53. de Cid R, Riveira-Munoz E, Zeeuwen PL, Robarge J, Liao W, Dannhauser EN, et al. Deletion of the late cornified envelope LCE3B and LCE3C genes as a susceptibility factor for psoriasis. *Nat Genet.* 2009;41(2):211–5.
54. Coto E, Santos-Juanes J, Coto-Segura P, Diaz M, Soto J, Queiro R, et al. Mutation analysis of the LCE3B/LCE3C genes in psoriasis. *BMC medical genetics.* 2010;11:45.
55. Joyce CE, Zhou X, Xia J, Ryan C, Thrash B, Menter A, et al. Deep sequencing of small RNAs from human skin reveals major alterations in the psoriasis miRNAome. *Hum Mol Genet.* 2011;20(20):4025–40.
56. Houseman EA, Accomando WP, Koestler DC, Christensen BC, Marsit CJ, Nelson HH, et al. DNA methylation arrays as surrogate measures of cell mixture distribution. *BMC Bioinformatics.* 2012;13(1):86.
57. Teschendorff AE, Breeze CE, Zheng SC, Beck S. A comparison of reference-based algorithms for correcting cell-type heterogeneity in epigenome-wide association studies. *BMC Bioinformatics.* 2017;18(1):105.
58. Humphreys KL, Moore SR, Davis EG, MacIsaac JL, Lin DTS, Kobor MS, et al. DNA methylation of HPA-axis genes and the onset of major depressive disorder in adolescent girls: a prospective analysis. *Transl Psychiatry.* 2019;9(1):245.
59. Gibson DS, Drain S, Kelly C, McGilligan V, McClean P, Atkinson SD, et al. Coincidence versus consequence: opportunities in multi-morbidity research and inflammation as a pervasive feature. *Expert Review of Precision Medicine and Drug Development.* 2017;2(3):147–56.
60. Dowlati Y, Herrmann N, Swardfager W, Liu H, Sham L, Reim EK, et al. A meta-analysis of cytokines in major depression. *Biol Psychiatry.* 2010;67(5):446–57.
61. Powell TR, Smith RG, Hackinger S, Schalkwyk LC, Uher R, McGuffin P, et al. DNA methylation in interleukin-11 predicts clinical response to antidepressants in GENDEP. *Transl Psychiatry.* 2013;3(9):e300.
62. Griffiths CEM, van der Walt JM, Ashcroft DM, Flohr C, Naldi L, Nijsten T, et al. The global state of psoriasis disease epidemiology: a workshop report. *Br J Dermatol.* 2017;177(1):e4–7.
63. de Jager MEA, de Jong EMGJ, van de Kerkhof PCM, Evers AWM, Seyger MMB. An inpatient comparison of quality of life in psoriasis in childhood and adulthood. *J Eur Acad Dermatol Venereol.* 2011;25(7):828–31.
64. Kurd SK, Troxel AB, Crits-Christoph P, Gelfand JM. The risk of depression, anxiety and suicidality in patients with psoriasis: a population-based cohort study. *Arch Dermatol.* 2010;146(8):891–5.
65. Tyring S, Gottlieb A, Papp K, Gordon K, Leonardi C, Wang A, et al. Etanercept and clinical outcomes, fatigue, and depression in psoriasis: double-blind placebo-controlled randomised phase III trial. *Lancet.* 2006;367(9504):29–35.
66. Langley RG, Feldman SR, Han C, Schenkel B, Szapary P, Hsu MC, et al. Ustekinumab significantly improves symptoms of anxiety, depression, and skin-related quality of life in patients with moderate-to-severe psoriasis: results from a randomized, double-blind, placebo-controlled phase III trial. *J Am Acad Dermatol.* 2010;63(3):457–65.
67. Kimball AB, Wu EQ, Guerin A, Yu AP, Tsaneva M, Gupta SR, et al. Risks of developing psychiatric disorders in pediatric patients with psoriasis. *J Am Acad Dermatol.* 2012;67(4):651–7.e1–2.
68. Gupta MA, Simpson FC, Gupta AK. Psoriasis and sleep disorders: a systematic review. *Sleep Med Rev.* 2016;29:63–75.
69. Russell K, Allan S, Beattie L, Bohan J, MacMahon K, Rasmussen S. Sleep problem, suicide and self-harm in university students: a systematic review. *Sleep Med Rev.* 2019;44:58–69.
70. Duffin KC, Chandran V, Gladman DD, Krueger GG, Elder JT, Rahman P. Genetics of psoriasis and psoriatic arthritis: update and future direction. *J Rheumatol.* 2008;35(7):1449–53.
71. Chandran V. The genetics of psoriasis and psoriatic arthritis. *Clin Rev Allergy Immunol.* 2013;44(2):149–56.
72. de Koning HD, van den Bogaard EH, Bergboer JGM, Kamsteeg M, van Vlijmen-Willems IMJJ, Hitomi K, et al. Expression profile of cornified envelope structural proteins and keratinocyte differentiation-regulating proteins during skin barrier repair. *Br J Dermatol.* 2012;166(6):1245–54.
73. Jackson B, Tilli CMLJ, Hardman MJ, Avilion AA, MacLeod MC, Ashcroft GS, et al. Late cornified envelope family in differentiating epithelia—response to calcium and ultraviolet irradiation. *J Invest Dermatol.* 2005;124(5):1062–70.
74. Bergboer JGM, Zeeuwen PLJM, Schalkwijk J. Genetics of psoriasis: evidence for epistatic interaction between skin barrier abnormalities and immune deviation. *J Invest Dermatol.* 2012;132(10):2320–31.
75. Hüffmeier U, Bergboer JGM, Becker T, Armour JA, Traupe H, Estivill X, et al. Replication of LCE3C–LCE3B CNV as a risk factor for psoriasis and analysis of interaction with other genetic risk factors. *J Invest Dermatol.* 2010;130(4):979–84.
76. Harder J, Bartels J, Christophers E, Schroder JM. Isolation and characterization of human beta-defensin-3, a novel human inducible peptide antibiotic. *J Biol Chem.* 2001;276(8):5707–13.
77. Novak P, Stampfer MR, Munoz-Rodriguez JL, Garbe JC, Ehrlich M, Futscher BW, et al. Cell-type specific DNA methylation patterns define human breast cellular identity. *PLoS One.* 2012;7(12):e52299.
78. Smith AK, Kilaru V, Klengel T, Mercer KB, Bradley B, Conneely KN, et al. DNA extracted from saliva for methylation studies of psychiatric traits: evidence tissue specificity and relatedness to brain. *Am J Med Genet B Neuropsychiatr Genet.* 2015;168B(1):36–44.
79. Teschendorff AE, Relton CL. Statistical and integrative system-level analysis of DNA methylation data. *Nat Rev Genet.* 2018;19(3):129–47.
80. Bell CC. DSM-IV: diagnostic and statistical manual of mental disorders. *JAMA.* 1994;272(10):828–9.
81. Kessler RC, Ustun TB. The WHO world mental health surveys: global perspectives on the epidemiology of mental disorders. New York; 2008.
82. Haro JM, Arbabzadeh-Bouchez S, Brugha TS, de Girolamo G, Guyer ME, Jin R, et al. Concordance of the composite international diagnostic interview version 3.0 (CIDI 3.0) with standardized clinical assessments in the WHO world mental health surveys. *Int J Methods Psychiatr Res.* 2006;15(4):167–80.
83. Neck MK, Holmberg EB, Photos VI, Michel BD. Self-injurious thoughts and behaviors interview: development, reliability, and validity in an adolescent sample. *Psychol Assess.* 2007;19(3):309–17.
84. University of Northumbria. Round A. A survey of student attitudes, experiences and expectations on selected vocational courses at the University of Northumbria. 2005.
85. Swearer SM, Cary PT. Perceptions and attitudes toward bullying in middle school youth. *J Appl Sch Psychol.* 2003;19(2):63–79.
86. Felitti VJ, Anda RF, Nordenberg D, Williamson DF, Spitz AM, Edwards V, et al. Relationship of childhood abuse and household dysfunction to many of the leading causes of death in adults. The adverse childhood experiences (ACE) study. *Am J Prev Med.* 1998;14(4):245–58.
87. Gross JJ, John OP. Individual differences in two emotion regulation processes: implications for affect, relationships, and well-being. *J Pers Soc Psychol.* 2003;85(2):348–62.
88. methylumi: Handle Illumina methylation data. R package version 2.32.0. [Internet]. 2019.
89. Teschendorff AE, Marabita F, Lechner M, Bartlett T, Tegner J, Gomez-Cabrero D. A beta-mixture quantile normalization method for correcting probe design bias in Illumina Infinium 450 k DNA methylation data. *Bioinformatics.* 2013;29.
90. Seshan V, Olshen A. DNACopy: DNA copy number data analysis. R package version 1.60.0. 2019.
91. Leek JT, Johnson WE, Parker HS, Jaffe AE, Storey JD. The sva package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics.* 2012;28(6):882–3.
92. Giardine B, Riemer C, Hardison RC, Burhans R, Eltnitski L, Shah P, et al. Galaxy: a platform for interactive large-scale genome analysis. *Genome Res.* 2005;15(10):1451–5.
93. Makambi K. Weighted inverse chi-square method for correlated significance tests. *J Appl Stat.* 2003;30(2):225–34.

## Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

## Supplementary Materials

**Table S1: Pyrosequencing primers**

Gene	Primers
	F1 GGA ATG ATA AAA GGG AAG TAG GAA
<i>LCE3A</i>	R1  ACC CTA ACT TCA AAA CAT ATA AAC TAA
	S1  AAG GGA AGT AGG AAA TT
<i>MIR4520A</i>	F1 GTT TAA ATT TTT TTT TGA TTT GGA TAG AAA
	R1  AAA ACA TAC CCT CAA TTC CAA AAA AAT C
	S1  TTT TTT TTG ATT TGG ATA GAA AAT A
<i>PSORS1C3</i>	F1 GGA GGT TTT TAT TGG TT GGA GTT GT
	R1  AAA TCA CCC CTC CCA CTA CTA A
	S1  ATT GGA TTG GAG TTG TT



**Table S2: Immune-related genes analysed for gains in methylation at promoter**

---

<b>Gene names [UCSC]</b>
CCL22 CCL23 CCL3 CCL4L1 CCL4L2 CCR1 CCR7 CXCL10 CD101 CD177 CD22 CD244 CD248 CD28 CD300LB CD300LD CD300LF CD300A CD33 CD34 CD40 CD48 CD5L CD5 CD6 CD74 CD79B CD84 CD93 NLRC3 NLRC4 NLRP12 NLRP3 DEFA1 DEFA1B DEFA3 DEFA4 IGHV4-39 IGHV7-81 IGKV2-30 IGKV2D-30 IGLC2 IGLL1 IGLV2-11 IGLV2-8 LILRA2 LILRA3 LILRB1 LILRB2 LILRB4 LST1

---

Promoter regions were defined as the interval from -500bp to +1bp from the transcriptional start site: methylation data for these regions in controls vs cases are presented in Fig.1D.

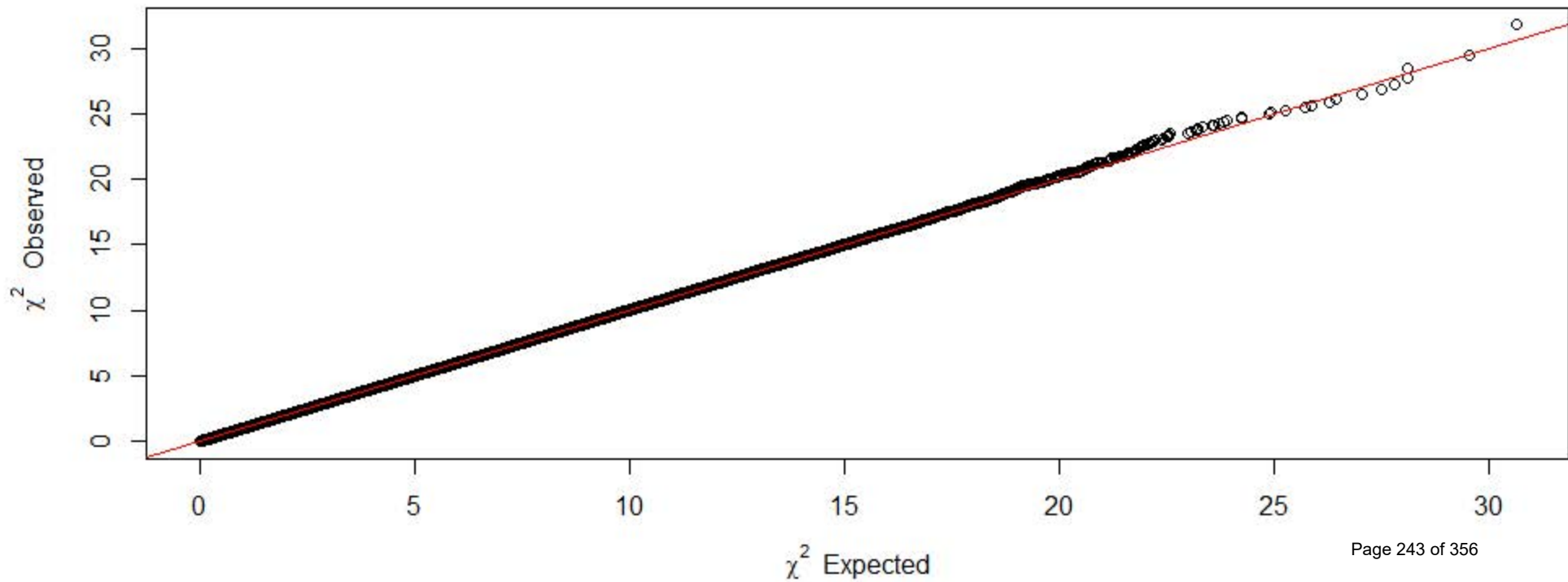
### **Fig. S1 Absence of population substructure effects**

A quantile–quantile (QQ) plot showing observed vs. expected  $-\log_{10}(p \text{ values})$  for association at all CpG sites. The x-axis shows the expected  $-\log_{10}(p \text{ value})$ , the y-axis the observed  $-\log_{10}(p \text{ value})$ : the red line indicates the expected distributions under the null hypothesis and the black dots were the observed values. A close match at lower significance values indicated no systematic inflation of P was seen due to unaccounted-for stratification effects.

### **Fig. S2 Absence of deletions or duplications at top differentially methylated loci**

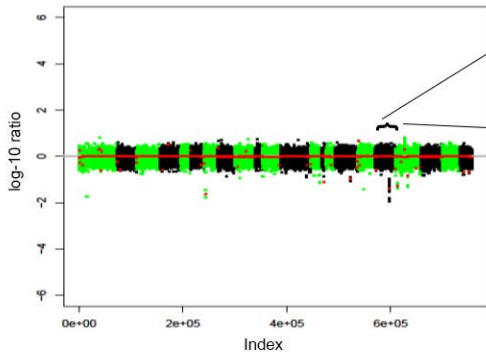
EPIC array probe data was analysed using the *DNACopy* package in R to look for variations indicating copy number variation (CNV): an example output plot from subject 225 (Healthy Control) is shown. Probe index number is shown along the x-axis, while gain/loss in copy number, expressed as the log-10 ratio, is shown on the Y-axis; dots coming away from the line indicate probes showing gains or losses of signal consistent with regional duplications/deletions. No significant CNVs were detected in the Epidermal Differentiation Complex (EDC) region on chromosome 1q21, but the approach successfully detected a CNV on chromosome 5 in one participant (arrow at right) not overlapping any of the differentially methylated regions, shown here as a positive control for sensitivity. *DNACopy* plots were carried out for all samples and failed to detect copy number variation (CNV) at other top hits.

# Figure S1

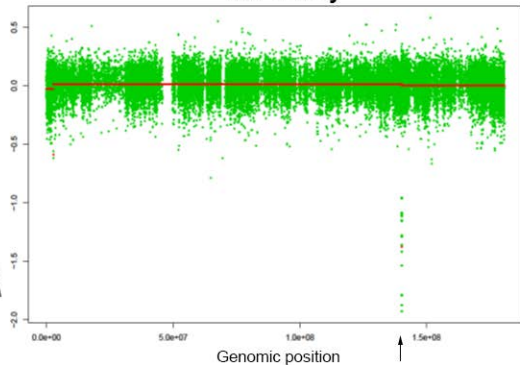


# Figure S2

## Whole genome



## chr 5 only



## 6.0 PAPER-V

### **CandiMeth: Powerful yet simple visualization and quantification of DNA methylation at candidate genes**

Sara-Jayne Thursby, Darin K. Lobo, Kristina Pentieva, Shu-Dong Zhang, Rachelle E. Irwin,  
Colum P. Walsh

The main aims of this paper were to:



- Provide an online user-friendly method of feature specific methylation analysis from the results on epigenome-wide methylation arrays
- Permit feature specific analysis of the outputs of multiple R-based methylation array processing pipelines

#### **CONTRIBUTION**

This paper represents the culmination of the development of the Galaxy workflow which I had been modifying through Papers I-IV from the original simple one generated by CPW. As well as massive redesign, I had to re-code much of it using SED commands, with the help of DKL. I conducted the majority of testing/improving of the workflow and created and populated the associated GitHub repository, as well as all the test and example data histories within Galaxy. I drafted the manuscript, carried out all suggested edits and made the majority of the figures.

## TECHNICAL NOTE

# CandiMeth: Powerful yet simple visualization and quantification of DNA methylation at candidate genes

Sara-Jayne Thursby <sup>1</sup>, Darin K. Lobo<sup>1,2</sup>, Kristina Pentieva <sup>3</sup>,  
Shu-Dong Zhang<sup>4</sup>, Rachele E. Irwin <sup>1</sup> and Colum P. Walsh <sup>1,\*</sup>

<sup>1</sup>Genomic Medicine Research Group, School of Biomedical Sciences, Ulster University, 1 Cromore Road, Coleraine, BT52 1SA, UK; <sup>2</sup>Present address: Republic Polytechnic, 9 Woodlands Avenue 9, Singapore 738964; <sup>3</sup>Nutrition Innovation Centre for Food & Health (NICHE), School of Biomedical Sciences, Ulster University, 1 Cromore Road, Coleraine, BT52 1SA, UK and <sup>4</sup>Stratified Medicine Research Groups, School of Biomedical Sciences, Ulster University, 1 Cromore Road, Coleraine, BT52 1SA, UK

\*Correspondence address. Colum P. Walsh, Genomic Medicine Research Group, School of Biomedical Sciences, Ulster University UK. Tel: +44 28 7012 4484; E-mail: [cp.walsh@ulster.ac.uk](mailto:cp.walsh@ulster.ac.uk)  <http://orcid.org/0000-0001-9921-7506>

## Abstract

**Background:** DNA methylation microarrays are widely used in clinical epigenetics and are often processed using R packages such as ChAMP or RnBeads by trained bioinformaticians. However, looking at specific genes requires bespoke coding for which wet-lab biologists or clinicians are not trained. This leads to high demands on bioinformaticians, who may lack insight into the specific biological problem. To bridge this gap, we developed a tool for mapping and quantification of methylation differences at candidate genomic features of interest, without using coding. **Findings:** We generated the workflow "CandiMeth" (Candidate Methylation) in the web-based environment Galaxy. CandiMeth takes as input any table listing differences in methylation generated by either ChAMP or RnBeads and maps these to the human genome. A simple interface then allows the user to query the data using lists of gene names. CandiMeth generates (i) tracks in the popular UCSC Genome Browser with an intuitive visual indicator of where differences in methylation occur between samples or groups of samples and (ii) tables containing quantitative data on the candidate regions, allowing interpretation of significance. In addition to genes and promoters, CandiMeth can analyse methylation differences at long and short interspersed nuclear elements. Cross-comparison to other open-resource genomic data at UCSC facilitates interpretation of the biological significance of the data and the design of wet-lab assays to further explore methylation changes and their consequences for the candidate genes. **Conclusions:** CandiMeth (RRID:SCR.017974; Biotools: CandiMeth) allows rapid, quantitative analysis of methylation at user-specified features without the need for coding and is freely available at <https://github.com/sjthursby/CandiMeth>.

**Keywords:** galaxy; methylation; workflow; DNA methylation; arrays; epigenetics; EWAS

## Introduction

Epigenetics can be defined as stable, and most often heritable, changes to the chromatin that do not alter the DNA sequence itself but still affect gene expression and/or are required to maintain genomic stability [1]. These modifications consist of

reversible marks such as cytosine DNA methylation or histone modifications, each critical to gene expression regulation, imprinting, X-inactivation, and many other processes from mammalian gestation to later life [1].

Cytosine DNA methylation is the most common and thoroughly investigated of these epigenetic alterations. It is charac-

Received: 6 December 2019; Revised: 12 April 2020; Accepted: 26 May 2020

© The Author(s) 2020. Published by Oxford University Press. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

terized by the addition of a methyl group to a cytosine residue, many of which are located within so-called CpG islands (CGI) close to gene promoters [2]. High levels of DNA methylation at promoters aid in the stable long-term repression of the cognizant genes, such as can be seen on the inactive X chromosome in mammals [3]. Methylation at control elements such as insulators or enhancers can also help regulate regional gene expression, with multiple examples being seen among imprinted genes [4] or gene clusters such as the protocadherins [5]. High levels of methylation are seen on selfish DNA elements such as endogenous retroviruses, where they play an important role in their suppression [6] as well as at inert regions of the genome such as pericentromeric repeats [7]. More recently, methylation through the body of the gene has been recognized as contributing to maintaining gene transcription levels at highly expressed genes [8, 9]. As well as showing such developmental programming, DNA methylation is susceptible to environmental influence, with inputs such as diet [10, 11] and exposure to pollutants such as cigarette smoke [12] having clear and reproducible effects on methylation levels, sparking great interest in analysis at a population level, particularly in humans [13].

Advances in sequencing technology have allowed us to quantify and analyse methylation via whole-genome bisulphite sequencing at ~28 million CpG resolution [14]. While this technique remains the gold standard for whole-genome methylation assessment, it can be very expensive, and when there are hundreds of samples to be tested and analysed prohibitively so; quantifying small differences reproducibly between multiple samples is also challenging. An alternative technology known as a microarray, which predates the era of whole-genome bisulphite sequencing, is often a popular solution for such cases, where a lower CpG resolution is satisfactory but where greater intersample reproducibility is required [15]. A popular choice here is the Illumina Infinium Methylation BeadChip array [15], which currently covers 850,000 CpG sites across the human genome, including 99% of RefSeq genes and large numbers of enhancers and other features. This can help elucidate the effects of an intervention across hundreds of samples in a cost-effective manner. There are many packages across multiple computational languages to analyse the outputs from these arrays such as RnBeads [16] or ChAMP [17], but these pipelines operate in the statistical programming environment R and require some coding. Additionally, the output file formats can be overwhelming and difficult to investigate further without experience in data analytics and bioinformatics. This situation is exacerbated by the typically higher number of samples in epidemiological or intervention studies where such arrays are commonly used.

To help solve this predicament, we developed a Galaxy workflow known as CandiMeth, which takes the main output from such methylation analysis pipelines and pairs this with a list of features that the user may wish to investigate. The workflow first generates tracks showing both absolute methylation levels in samples and differences in methylation between samples. These can be viewed via the University of California Santa Cruz (UCSC) genome browser and overlaid with other available tracks such as CpG island, enhancers, chromatin immunoprecipitation (ChIP) data, and so forth to allow data exploration and more intuitive analysis. This also facilitates the design of assays to cover specific CGs using BLAT. The workflow can then help confirm any patterns observed by quantifying data across the identified regions or features, e.g., methylation differences at specific sets of genes between cases and controls. It also has a bespoke analysis allowing estimation of methylation differences

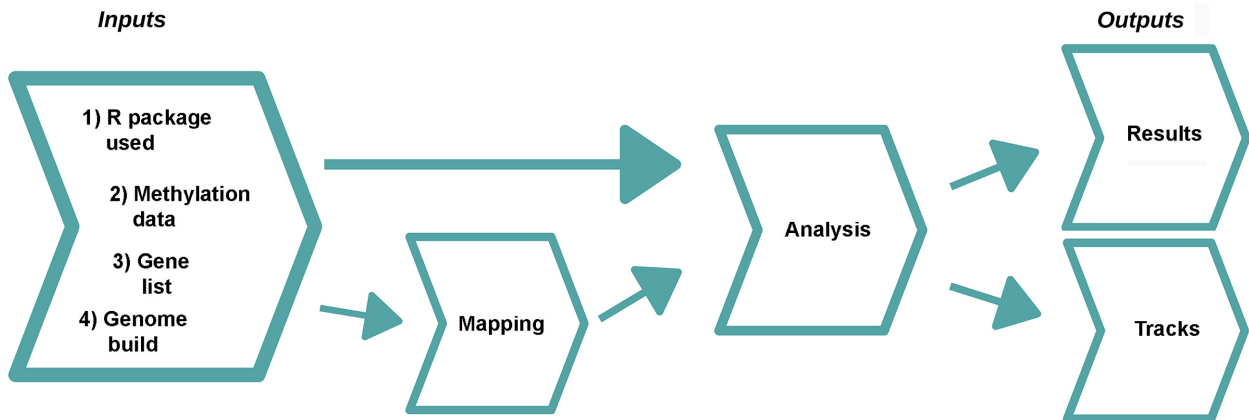
at repetitive sequences by leveraging the RepeatMasker tracks at UCSC. The workflow removes the need for further analysis in R and increases reproducibility by using an automated process, but in a more user-friendly manner.

## Methods

CandiMeth (CandiMeth, [RRID:SCR.017974](#)) (Biotoools: CandiMeth) is designed to work downstream of DNA methylation analysis pipelines in R. It was developed initially using RnBeads as reference but has been subsequently successfully run with ChAMP and other packages (see below). ChAMP (ChAMP, [RRID:SCR.012891](#)) [18] and RnBeads (RnBeads, [RRID:SCR.010958](#)) [15, 18] are end-to-end pipelines in R that can take raw data files such as IDATs and bam files from microarray readers or sequencers and process these to allow data exploration, visualization, and comparison. For array data, which is the main area where CandiMeth addresses an unmet need, IDAT files containing raw values for the red and green channels for each of ~850,000 probes are exported from the microarray reader. RnBeads/ChAMP can perform quality control, remove probes with low signal or overlapping with single-nucleotide polymorphisms (SNPs), and provide a cleaned dataset giving absolute levels of methylation as  $\beta$  or M values. The packages can also facilitate exploratory visualization through principal component analysis or similar and allow grouping of data prior to looking for differential methylation. Probes showing substantial differences in methylation ( $\Delta\beta$ ) can be identified and then ranked on the basis of a variety of parameters, including probability of occurrence (P-value),  $\Delta\beta$ , false discovery rate (FDR), or a combination of several of these. The packages can look for enriched gene sets using gene ontologies/GSEA [19] and visualize differences for annotated categories of array probe such as promoter and gene body.

While packages for array analysis provide genome-level data such as whether promoters in general are losing or gaining methylation, querying specific gene sets that might give more biological insight cannot be easily done in this or other R packages with similar functionality without extracting the processed dataset and writing bespoke code. Visualization of the data against the genome map is also of great attraction for the wet-lab biologist but is also not easily done within these packages. While RnBeads can map methylation values to the genome as customized tracks, this can only be carried out if a local instance is installed on the user's server, which requires substantial investment for set-up and maintenance. ChAMP does not currently provide tracks at all, to our knowledge. Typically, many biologists have specific genes that are of interest to them, or they may want to examine the area in which top sites are located and determine whether adjacent probes are also losing or gaining methylation. A ready way of assessing the degree to which methylation is changing across a particular region and the exact location of the probes also greatly facilitates the design of gene-specific assays such as primer sets for pyrosequencing or clonal analysis. It is also generally of interest to try and leverage the enormous pool of publicly available data accessible through UCSC Genome Browser tracks to explore possible novel correlations between methylation changes in a particular dataset and other genome characteristics such as replication timing, histone modifications, or similar.

We therefore wished to develop a user-friendly non-computationally intensive method of candidate feature investigation that avoided the command line but was more powerful than browser-only interfaces. To this end we chose the



**Figure 1:** Overview of CandiMeth workflow. When the CandiMeth workflow is started the user needs to specify as Inputs (left): (i) the type of R package used; (ii) the methylation data, normally in the form of a differential methylation table generated by the package; (iii) a list of genes to be analysed; and (iv) a human genome build to match the data to. The methylation data are then mapped (centre left) to the genome and sites overlapping features of interest analysed (centre right). The data are then output quantitatively as Results and visually as Tracks (right).

Galaxy (Galaxy, [RRID:SCR.006281](https://doi.org/10.26434/chemrxiv-2019-006281)) platform [20], which is a free open-source environment for user-friendly and reproducible bioinformatics [21]. It provides a variety of data manipulation and analysis tools via a web interface with no prior installation or dependency packages required, with results stored within the Galaxy infrastructure and every action producing a new history entry so the original data are never compromised via destructive edits. Galaxy also allows users to aggregate analysis steps into repeatable pipelines called workflows, which can be easily shared, along with the histories, via URL or username. These can allow biologists with little bioinformatics experience to conduct complex analyses on their own data within a system that has a low maintenance requirement and with little worry over data storage or data corruption. Moreover, workflows can be published to a repository such as GitHub ([RRID:SCR.002630](https://doi.org/10.26434/chemrxiv-2019-002630)) or MyExperiment ([RRID:SCR.001795](https://doi.org/10.26434/chemrxiv-2019-001795)) [22] or within a scientific journal—further encouraging open data science and reproducibility. Galaxy also provides many plugins such as interactive visualization software to view results, the option to export results to genome browsers, and the option to configure tools, or indeed an entire Galaxy instance, to the desired end-user needs.

### Overview of workflow

The main process undertaken by CandiMeth is to take as input the methylation data from an R pipeline such as RnBeads or ChAMP and (i) visualize the data as tracks in the UCSC Genome Browser and (ii) analyse the methylation differences relative to genomic features specified by the user. The workflow comprises 3 main steps: Inputs, Feature Mapping, and Analysis (Fig. 1). There are also 4 items required at input stage: the user must (i) indicate the R package used with the keywords “RnBeads,” “ChAMP,” or “Custom,” then supply (ii) the methylation data, (iii) a list of the genes of interest, and (iv) specify the human genome build to be used, e.g., hg19. The basic workflow for CandiMeth is that the genes of interest are mapped to the reference genome and then cross-referenced with the input methylation data to get feature-specific statistics. The workflow can currently look at either the promoters (–500 to +1 bp relative to transcription start site; suffix “\_P” on results) or gene bodies (the transcription unit; “\_GB”), or both parts of the gene together (“\_all”). We have found this to be a particularly useful split because the current consensus is that promoters and gene bodies can show opposite

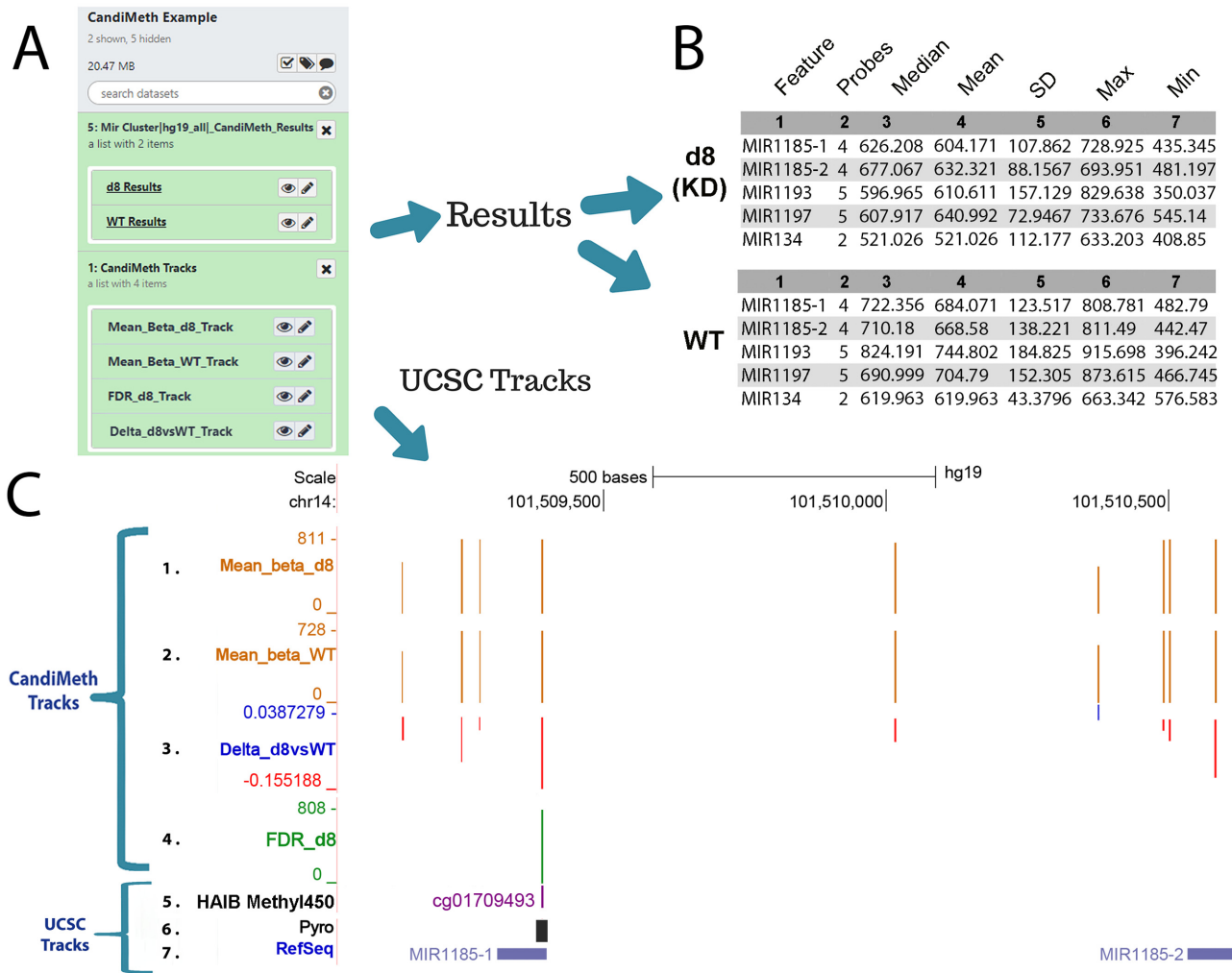
methylation patterns, with methylation at the promoter largely associated with repression, whereas gene body methylation instead is a feature of transcribed genes. Outputs are then grouped in the history into 2 types, Results or Tracks (Fig. 1). The methylation data from the R packages are output as a standard differential methylation table as routinely generated, and either a single table comparing 2 groups, or several tables can be processed at once as inputs, e.g., comparing different experimental conditions with the control. Each comparison will result in a separate table and tracks, grouped together and given a condition-specific identifier to avoid confusion. The CandiMeth workflow, together with the example datasets used and a step-by-step tutorial, are available on GitHub [23]. CandiMeth is optimized to work on the latest version of Galaxy (19.0) through the Galaxy website [20, 24], thus making it platform-independent. For users who have their own instance of Galaxy, the workflow can be downloaded and imported via a link on the GitHub page, where a .yaml file is also available.

### Example outputs

To illustrate the type of analysis that can be done, Fig. 2 shows outputs from 1 of the example dataset runs. Here we used as input 1 of our previously published differential methylation tables generated by RnBeads (NCBI Gene Expression Omnibus [GEO] identifier GSE90012; the table is also given as Suppl. Table 1) [25]. The experiment compared wild-type hTERT1604 human fibroblast cells (WT) and a clonal derivative with a stable knock-down (KD) of the maintenance DNA methyltransferase DNMT1 (d8 KD), which gave large alterations in DNA methylation levels, very suitable for the purpose of illustration here. The second item needed for CandiMeth, namely, features of interest, was in this case a set of microRNA (MIR) genes not analysed in the original article, which was input here simply as a list of names (given in Supp. File 2). CandiMeth first mapped the MIR locations to the human genome (in this case hg19), then analysed the co-occurrence of probes at these locations. The results appeared in Galaxy as 2 grouped sets of datasets (Fig. 2A): “Mir Cluster | hg19 all | CandiMeth Results” and “CandiMeth Tracks.”

The Results set contained an output table for each condition, namely, KD (d8) and WT cells (Fig. 2B, first 5 rows of each shown). Each table consisted of 7 numbered columns. It should be noted here that methylation values from the array are expressed as a





**Figure 2:** Case Study 1: analysing new genes in a published dataset. Our previously published dataset GSE90012 using RnBeads to compare methylation levels in cells deficient in DNA methyltransferase 1 (d8 KD) with wild type (WT) was reanalysed for methylation levels at microRNA (MIR) using CandiMeth. (A) The workflow generated 2 grouped sets of outputs (white boxes on left) on completion, “Mir Cluster | hg.19.all | CandiMeth Results” containing links to the tabular quantitative data and “CandiMeth Tracks” with links to the tracks on UCSC (B) CandiMeth Results box expanded: a separate dataset for each cell line is generated showing the list of candidates, probe coverage, median, and a variety of other statistics for each gene analysed (top 5 rows only shown). (C) CandiMeth Tracks: UCSC Genome Browser view, accessible via the eye symbols on the Galaxy history shown in (A): (From the top down) Scale bar, size of region in kilobases of DNA; chr1, chromosome number and exact coordinates from the hg19 genome build. (1–4) CandiMeth tracks: (1) Mean.beta.d8, absolute methylation track reflecting array output going from 1, no methylation, to 1,000, fully methylated, e.g., 811 = 81.1%, maximum and minimum indicated at left; (2) Mean.beta.WT, absolute methylation in WT; (3) delta.d8vsWT, a differential methylation track showing proportional change going from  $-1.0$  (100% loss, red) to  $+1.0$  (100% gain, blue), e.g.,  $-0.155$  = loss of 15.5% compared to WT; (4) FDR.D8, a significance score track showing only those sites whose differential methylation meets the cut-off criterion of a 0.05 false discovery rate. (5–7) Examples of some of the tracks available through the UCSC Genome Browser, which can be aligned and directly compared to CandiMeth tracks: (5) HAIB MethyI450, data on comparative methylation from ENCODE projects; (6) Pyro, the BLAT tool in UCSC, which can be used to find primers for pyrosequencing to cover 1 or multiple CG; (7) RefSeq track, showing the location of the top 2 MIR from (B).

number from 1 (no methylation) to 1,000 (fully or 100% methylated) to facilitate visualization. The numbered columns correspond to (1) Feature, the candidate region of interest, in this case each of the MIR in the initial list; (2) Probes, the number of array probes that are found in the specified feature; (3) Median, the methylation value that is the median of all probes mapping to that feature, e.g., 626.208/1,000 is the median of all probes at MIR1185-1, or 62.6% methylated; (4) Mean, the mean methylation value across all probes; (5) SD, the standard deviation; (6) Max, the maximum probe value seen in the feature; and (7) Min, the minimum probe value (Fig. 2B). It can be seen that methylation values are much lower in the DNA methyltransferase-depleted cells (d8) for each miR compared to the parental or WT cells, e.g., MIR1185-1 62.6% median methylation in d8 vs 72.2% in

KD. It can be seen that, while usually in reasonable agreement, in some cases the median and mean vary substantially, and having data on the numbers of probes can be useful for deciding confidence in the results and on any threshold to be applied.

In the Tracks folder CandiMeth also generated 4 tracks on the UCSC Genome Browser (Fig. 2C, 1–4), which can be visualized by clicking on the eye icon on the Galaxy datasets under CandiMeth Tracks in Fig. 2A (clicking on each track overlaid it on the previous one to generate the cumulative view shown).

Tracks 1 and 2 are absolute methylation (raw  $\beta$ ) tracks, denoted as “Mean.beta” in CandiMeth outputs. These show the methylation per probe for all probes in the differential methylation table that passed quality control and other screening steps in RnBeads, and not just the feature-specific (here MIR) probes,

as we have found that the genomic methylation context is very valuable to consider when looking at features. In other words, even if Promoters is selected at input, the tracks will show all probes, including those in the gene body and other regions. Track 1 is the DNMT1-depleted cell line (“Mean.beta.d8”) data, and Track 2 is from the WT cells (“Mean.beta.WT”).

Track 3 is the  $\Delta\beta$  track (“Delta.d8vsWT”) showing the difference between methyltransferase-deficient and WT cells. These are BedGraph files like Tracks 1 and 2, but because methylation can be higher or lower in 1 sample versus another, the visualization is different from the absolute methylation tracks. Instead, gains in methylation in the experimental condition are shown as blue columns above the zero (no change) line, and losses are shown as red columns below the line, with a change of +1 being 100% increase and  $-1$  being  $-100\%$ , i.e., an array probe going from 100% methylated to 0% methylated. The Delta track also allows the user to see how many array probes in a region are showing large differences in methylation and whether a differentially methylated region (DMR) identified by RnBeads extends farther than originally estimated [26]. Note that this track shows all differences in methylation, however small: the FDR-corrected probes are shown in the next track.

Last, an FDR-corrected track (“FDR.D8,” Track 4) was also produced: this only showed information for those probes where the R package has assessed the FDR to be  $<0.05$  because this is a statistical cut-off implemented by many array users. This is an excellent method for visualizing only CG that have high-confidence differences in methylation between samples. Here, only a single probe passed the FDR threshold and is shown: the absolute methylation level at the probe is given because P-values would not scale correctly.

One of the most powerful features of using this approach is that data can easily and more intuitively be compared to other UCSC tracks (Fig. 2C, 5–7). The specific CpG site can be identified in UCSC, e.g., by right-clicking on the column on the track, or by typing the CG identity into the UCSC browser search window, which will then pull out a track with the site highlighted, in this case the ENCODE project’s HAIB Methyl450 (Fig. 2C, Track 5). A particularly useful tool in this context is UCSC’s BLAT, which can be used to help ensure that primers designed to verify methylation differences at specific regions of interest by pyrosequencing or similar do indeed overlap the crucial sites (Fig. 2C Track 6, Pyro), in this case the FDR-significant site. Off-the-rack assays for each CG on the EPIC array can also now be purchased commercially. Other UCSC tracks shown in Fig. 2C include the RefSeq track (Track 7), invaluable for identifying well-curated genes rather than predicted or rare products. These tracks were all overlaid on the CandiMeth tracks, allowing the user to see whether methylation changes were located in or near any of these features. These are examples only; any track available through UCSC or that can be called through Galaxy can potentially be aligned with the CandiMeth tracks.

### Data preparation and inputs

A complete User Guide document with step-by-step tutorials is available [23]; here we describe more general features of the workflow. As indicated, CandiMeth runs in the Galaxy environment: users must first create an account and copy the CandiMeth test history and workflows to their account, as explained in the Guide. Once these simple steps have been carried out the first time, they do not need to be repeated. When CandiMeth is being run, the initial window will look as shown in Fig. 3: the workflow occupies the central window, while the example data

and datasets required for the workflow are in the History window at right; the left window Tools will not be used. Upon initialization, the workflow window will look as shown, with 1 Yes/No choice and 4 fields (numbered 1–4) to fill in. We recommend saving the outputs of CandiMeth to a new history when initiating the pipeline. This will (i) make it possible to continue working on other tasks while CandiMeth is running in the background—the workflow can take a while to run depending on server usage and (ii) segregate the current job from the reference datasets in the CandiMeth initial history, which avoids cluttering the initial history or causing problems if a particular run fails and generates incompletely processed datasets. The 4 fields are the 4 forms of inputs required, as indicated in the example above and dealt with below.

#### Input Type 1: R package used

CandiMeth works downstream of R-based packages that are designed to process epigenome-wide datasets. The 2 most popular packages (by Bioconductor download) ChAMP and RnBeads both automatically generate tabular data outputs that are suitable as input for CandiMeth without further processing, but the tables are in slightly different formats. Therefore, CandiMeth users should select either “RnBeads” or “ChAMP” when asked which R package was used. CandiMeth also supports other packages via a “Custom” keyword.

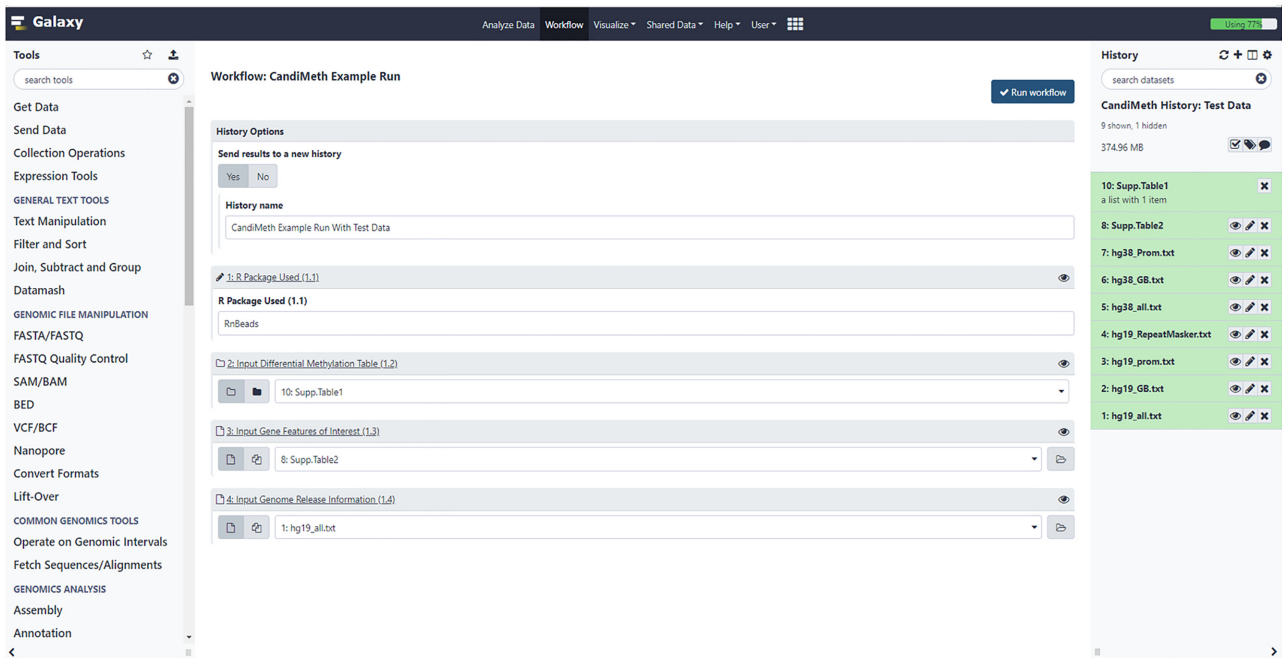
#### Input Type 2: Differential methylation table

The user needs to identify the location of, or upload directly, a copy of the output table from the R package containing the differential methylation data. For RnBeads this can be found via the html interface by opening “differential\_methylation.html” and choosing the desired comparison table. Once uploaded, the differential methylation table must be converted to a dataset collection through a 1-step operation (see the User Guide [23, 27]), which allows all the data from the table to be processed at once. An example table is available in dataset collection format in the CandiMeth default history; the raw table itself is also available as Supp. Table 1. In addition, an example ChAMP output is also available as Suppl. Table 5 in the CandiMeth History.

If the custom option is chosen at Input 1 above, the user can input a data frame of any origin as long as it follows the default CandiMeth format, namely: Chromosome; Start; cgid; mean.X; mean.Y; the difference between the 2 groups; and the FDR-corrected P-value (where X and Y equal the names of the experimental and control groups, respectively). Data frames can also be rearranged in Galaxy using the text manipulation tools “cut” and “join” within the Galaxy tool panel to produce an acceptable input table. We hope to extend the number of preformatted options beyond RnBeads and ChAMP to reduce the need for custom inputs in future.

#### Input Type 3: Gene features of interest

Here the user can choose which features they want to investigate. This can be done in a customized fashion, but commonly biologists initially want to see how much methylation is present across well-defined genomic features such as genes. This can easily be done in CandiMeth by following the commands `>Get Data >upload File >paste/Fetch` and then typing the official gene names, 1 per line, into the window that opens there (see Step-by-Step Guide). Alternatively, they can be uploaded as a list in a tab-delimited file format at this step. To facilitate initial trials, the MIR gene names used above have been preloaded into the default CandiMeth history for use and are also supplied as Supp.



**Figure 3:** User Interface for the workflow. Screenshot of the workflow start window (*middle pane*) that appears on right-clicking >CandiMeth>Run. The right-hand side shows the CandiMeth starting history, where preloaded data used with the workflow can be found, together with any user-uploaded datasets. Galaxy tools (*left*) are not used. For the workflow the user chooses whether to save results to a new history (recommended), then specifies (1) which R package was used to pre-process the data, e.g., RnBeads; (2) the dataset collection table of pre-processed data—available sets will appear in the drop-down menu; (3) a list of the genes/other features of interest to analyse; and (4) the reference genome to be mapped to, e.g., hg19. Once all 4 have been decided, the user clicks on the blue “Run workflow” button at top right to initiate a run.

Table 2. The features associated with the gene names are then mapped to the genome using the genomic data discussed next.

#### Input Type 4: Genome information

An important part of the CandiMeth workflow is the parsed human genome information used to assign array probes to various genomic features. Example human genome build information used for the mapping part of the CandiMeth pipeline can be found within the CandiMeth history (right-hand pane in Fig. 3). The data provided here cover 2 genome assemblies, hg19/hg38, and will aid the mapping of candidate features to promoters, whole gene body region, or both (hg19\_all option) as defined by RefSeq [28].

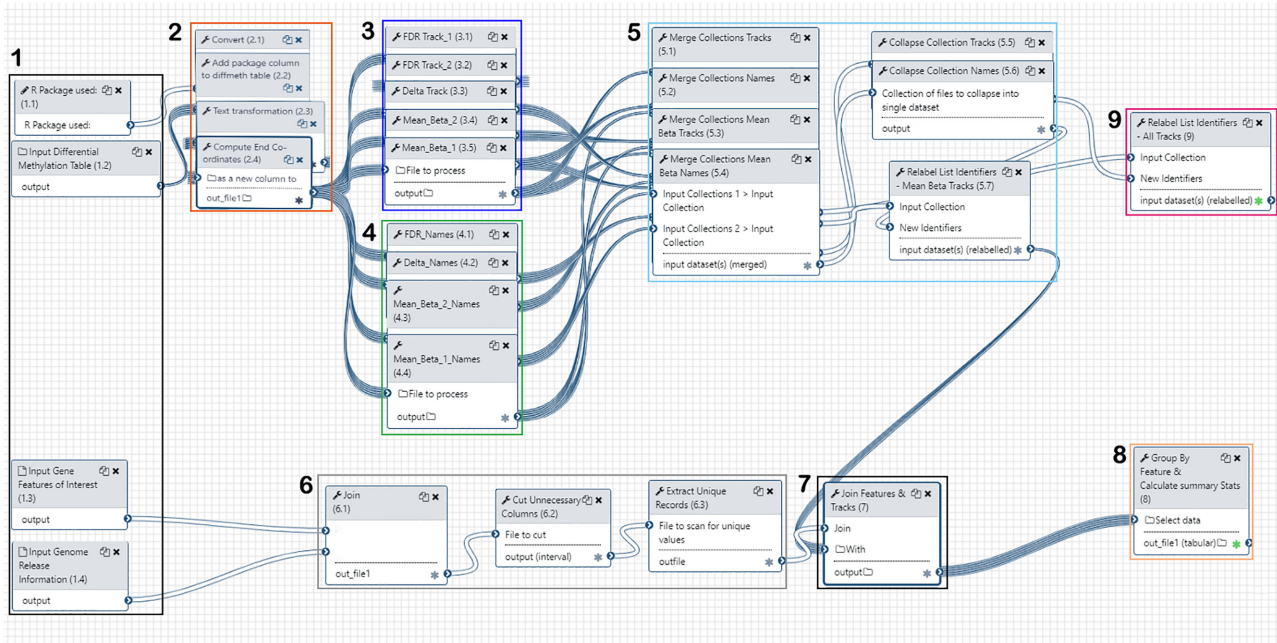
Using CandiMeth, users can query RefSeq-defined genes or repeats to obtain the same types of information as can be obtained by analysis in an R package. One advantage here however is that the simultaneous visualization allows the user to inspect the match between probe location and gene structure for candidate regions of interest: e.g., the initial screen may indicate changes in promoter methylation from the manifest-defined promoter, when inspection shows that all of the probes lie in the first exon of a single-exon gene and therefore are in fact gene body, the discrepancy being due to the definition of promoter in the manifest. CandiMeth allows the user to refine or alter the promoter definition to exclude bases downstream of the transcriptional start site, for example, and re-evaluate. An approximation of promoter areas of these RefSeq genes was generated for the example data analysis and was defined as the region from 500 bp upstream to the first base ( $-500 \rightarrow +1$  bp) and is available in the CandiMeth history [29] mentioned above. Similarly, probes were also parsed into gene body and repeat categories for CandiMeth to facilitate user analysis of effects over

these types of genomic intervals for their candidate genes of interest.

#### Processing steps

Fig. 4 shows a workflow editor view of CandiMeth: different sections have been numbered for ease of reference here.

1. Inputs: Inputs are indicated at left; R package used to generate the table (1.1), differential methylation table (1.2), features of interest (1.3), and parsed genome information specific to that type of interval, e.g., promoters (1.4). Once the 4 input types have been decided (see aforementioned examples) the workflow proceeds as follows.
2. Generation of a standardized data frame between RnBeads and ChAMP: First the CSV file output from the R package is processed by converting the delimiters used into tabs (2.1), then the keyword identifier for that package (either “RnBeads,” “ChAMP,” or “Custom”) added to the differential methylation table (2.2) to form an extra column. A table is then output showing the chr, start, cgid, mean methylation between control and experimental groups, the difference between these experimental groups, and FDR-corrected  $P$ -value (2.3). Subsequently, the end coordinates for each cg site are calculated and added to this table (2.4), so the data can be configured to run on UCSC Genome Browser at a later stage in the workflow.
- 3,4. Track generation and naming: Differential Table inputs from RnBeads (1.2) are converted into a variety of tracks compatible with UCSC Genome Browser. These include 2 absolute methylation tracks (3.4, 3.5) in this case, 1 FDR track showing only FDR significant sites (3.1), and 1  $\Delta\beta$  track (3.3) showing the difference in  $\beta$ -value between the 2 absolute methylation



**Figure 4:** Galaxy Workflow Editor View of CandiMeth. Detailed view of the workflow using the editing tool in Galaxy. Steps in the workflow have been grouped for clarity. (1) Inputs: here the user indicates which R package was used to analyse their array using the keywords “RnBeads,” “ChAMP,” or “Custom” (1.1), identifies the differential methylation table resulting from this R package (1.2) and the genomic features that they wish to analyse (1.3), and specifies the desired genome build (1.4). (2) Standardizing the input data: Using the R package information in 1.1 and the differential methylation table in 1.2, CandiMeth generates a table showing the chromosome location, start, end, mean methylation in the control and experimental groups, the difference between these groups, and the FDR-adjusted P-value. (3) Track generation: maps the data on absolute as well as differential methylation from the table to the genome build. (4) Track naming: generates unambiguous labels for each type of track. (5) Merging of tracks and names: this ensures logical labelling and grouping of tracks. (6) Feature mapping: this maps the specific features to the same genome build. (7) Compilation of feature methylation: this parses the data in the tracks to only examine the features of interest. (8) Output Tables: these contain summary statistics on the features of interest and are 1 major output. (9) Output Tracks: the user can also see the mapping on which the summary statistics are based, which allows them to see areas adjacent to the features of interest, and overlay other UCSC tracks, as well as use tools such as BLAT.

tracks. Track and results names (4.1–4.4) are also generated from the differential table inputs: this is an important step because both absolute methylation data for individual samples and a number of types of comparison data must be separated and given logical and intuitive names to allow easy identification among the multiple output datasets. The workflow uses a number of pre-existing tools available in Galaxy to carry out these steps (Table 1).

- Merging of tracks and names: Following track creation (3), the resultant tracks and their names are merged into separate dataset collections (5.1–5.4) and then collapsed into singular dataset collections (5.5, 5.6), one for all comparative tracks (5.5), one for all comparative track names, and one for all absolute methylation (mean  $\beta$ ) tracks (5.4) with their associated names (5.7). The mean  $\beta$  tracks will be used for feature investigation later in the workflow. The results here are compilations containing information on methylation at each probe across the genome in each sample, or the differences in methylation at specific probes between pairs of samples.
- Feature mapping: Features of interest (1.3) input by the user such as a particular set of genes are joined (6.1) to the specified genome release information (1.4) using the Paste tool. The gene features of interest are overlapped with the genome release information to obtain the desired genome intervals using AWK (6.2). Any repeated columns or rows that are no longer required are discarded and unique records extracted (6.3). The output here is a set of genomic coordinates matching only the specific features of interest, e.g., a specific set of genes.

- Compilation of methylation data for features: The dataset collection containing now correctly named absolute methylation tracks (5.7) is now joined with the mapped features of interest (7). This allows the generation of feature-specific statistics.
- Outputs: Feature-specific statistics such as mean methylation over all probes in each feature, median, maximum, etc. (see below), are tabulated and form 1 major output (8). The comparative tracks (generated in 3) are also given unambiguous final names, collated, and output as a dataset collection called “CandiMeth Tracks” (9, with green stars marking final output states).

## Output files

The CandiMeth workflow produced as indicated above under Example outputs 2 main types of output files:

### Tables

Results tables all follow the same layout: feature name, probe coverage, median methylation, mean methylation, standard deviation, maximum, and minimum. A partial example of a tabular output for the set of miRs used in the example above is shown in Fig. 2B (first 5 lines) and given in full in Suppl. Table 3. Methylation values for the features can then be plotted within Galaxy via their integrated visualization software or the Table can be exported and downloaded then plotted within the user’s preferred visualization software such as Prism or Excel as desired.

**Table 1:** List of Galaxy tools used

Tool name	Tool ID	Version	CandiMeth step	Reference
Convert delimiters to TAB	convert_characters	1.0.0	2.1	[30]
Add column to an existing dataset	add_value	1.0.0	2.2	[31]
Text transformation with SED	tp_sed_tool	1.1.1	2.3	[32]
Compute an expression on every row	column_maker	1.2.0	2.4	[33]
Merge collections into single list of datasets	__MERGE_COLLECTIONS__	1.0.0	5	[34]
Relabel list identifiers from contents of a file	__RELABEL_FROM_FILE__	1.0.0	5.7/9	[35]
Collapse Collection into single dataset in order of collection	collapse_dataset	4.1.0	5.5/5.6	[36]
Paste 2 files side by side	Paste1	1.0.0	6.1	[37]
Text reformatting using AWK	tp_awk_tool	1.1.1	6.2	[38]
Unique occurrences of each record	tp_sorted_uniq	1.1.0	6.3	[39]
Join the intervals of 2 datasets side by side	tp_easyjoin_tool	1.0.0	7	[40]
Group data by a column and perform aggregate operations on other columns	Grouping1	2.1.4	8	[41]

### Tracks

CandiMeth produced 4 different tracks from the differential methylation table input in the first step, of 3 different kinds (absolute methylation, relative differences in methylation [ $\Delta\beta$ ], and FDR-significant methylation difference), as shown in the example above for a cell line system.

### Findings

The utility of the CandiMeth workflow may be best illustrated by some case studies.

#### Case Study 1: Application to array results from model systems

One straightforward use of CandiMeth that has found common use in our laboratory and among collaborators is to test a specific gene set, as illustrated by the *MIR* example above (Fig. 2). To do this, the user only has to specify a list of the names of the genes they are interested in, together with the genome release, then upload a table containing differential methylation data. This can either be one generated by the bioinformatics team in-house; one that was supplied, typically when array services are brought in; or one that was generated from publicly available array data such as our dataset GSE90012 described previously [42] and used above.

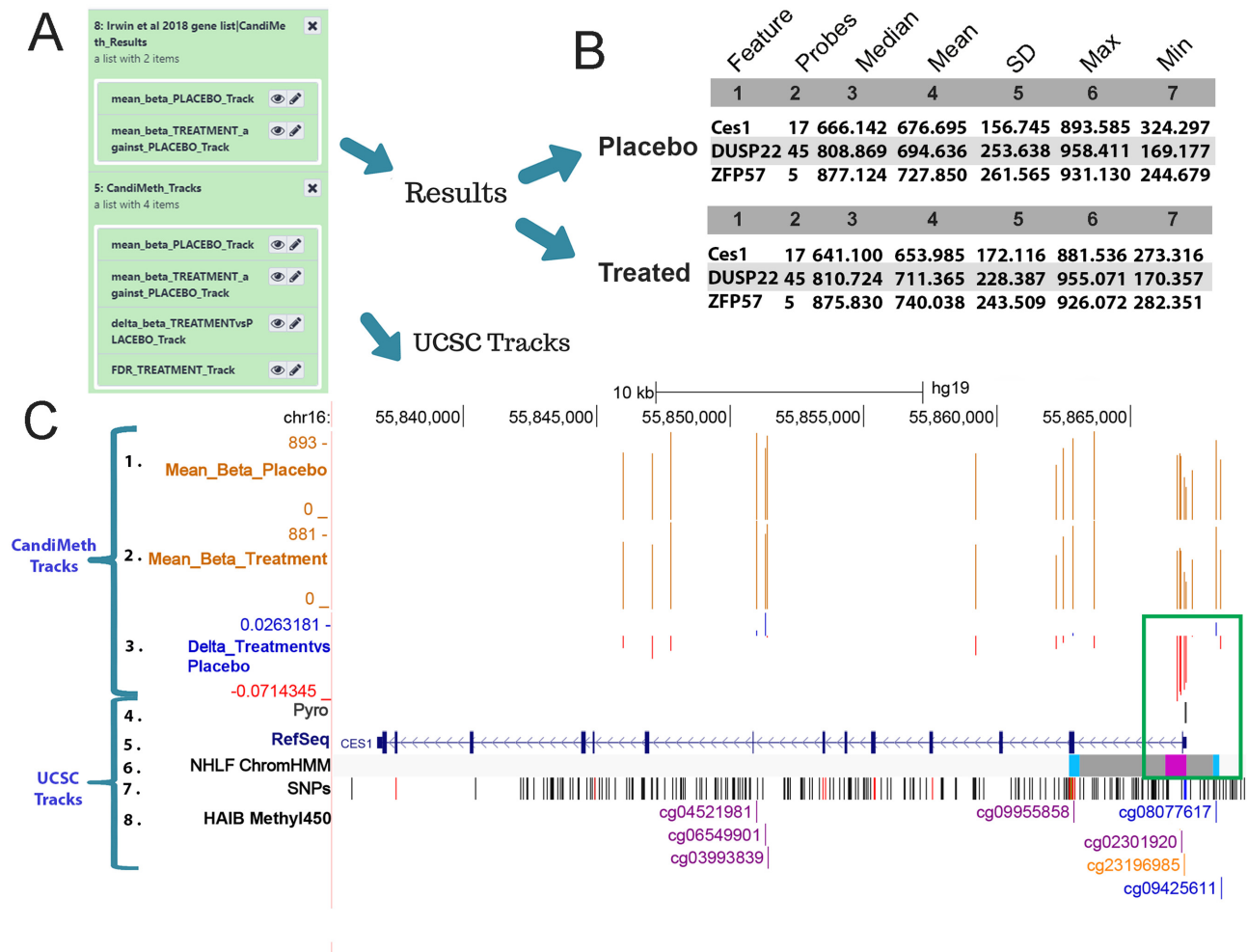
#### Case Study 2: Application to EWAS study outputs

A major application of methylation array technology is in epigenome-wide association studies (EWAS). CandiMeth can provide a very useful tool for quickly examining in detail and quantifying methylation differences around candidate regions identified either by the R-based packages or from the literature. Fig. 5 shows the application of this approach to an EWAS that we have recently published containing data from 86 participants divided into 45 receiving placebo and 41 receiving folic acid supplementation during trimesters 2 and 3 of pregnancy to assess the potential positive effects of prolonging this vitamin supplementation beyond the currently recommended preconception and first trimester periods [26]. Output differential methylation tables from RnBeads were used as input for CandiMeth, together with the names of the top candidate promoters reported earlier. This produced a collection of outputs (Fig. 5A) including a set of tabular Results for the 2 groups Placebo and

Treatment, as well as a set of Tracks. The latter included absolute mean  $\beta$ ,  $\Delta\beta$ , and an FDR track, although the latter returned the message “#No FDR significant sites” (not shown), often the case for EWAS if the sample set was small or the perturbation mild. Clicking through to the tabular results (Fig. 5B) showed tables indicating the number of probes present at each promoter and mean methylation, revealing, e.g., that median methylation at the *CES1* promoter is 2.5% lower in folic acid-treated participants than placebo ( $666.142 - 641.100 = 25.042/1,000 = 0.025$ , or 2.5%).

Examination of the CandiMeth Tracks (Fig. 5C) was however also informative here. This BedGraph track type is set by default to scale to the maximum loss and gain on visualization, so that when the UCSC browser is opened on a genomic region of interest, not only are the maximum loss and gain shown, but the graph is scaled to these, meaning that even when small differences in methylation occur, as typically seen in epidemiological studies, the areas of the genome with the greatest changes can be easily identified at a glance. In-house testing has found  $\Delta\beta$  tracks to be particularly useful because it can easily be seen whether a feature contains any probes with methylation differences between samples big enough to assess by other means—e.g., pyrosequencing can accurately assess differences in methylation  $>5\%$ . It can be easily seen from the  $\Delta\beta$  (Track 3) that the biggest loss of methylation was 7% ( $-0.071$ ). The clustering of sites losing methylation at the promoter is also striking (boxed in green) compared to the rest of the gene, suggestive of a step-change in methylation at this important regulatory element rather than a point source. The seamless integration of BLAT [43] meant that designing primers to verify methylation changes could be done very intuitively and the area covered by the assay mapped against the methylation data to confirm that the assay could confirm methylation levels at the exact same location (Fig. 5C Track 4 “Pyro”).

It was also seen from the absolute methylation levels in the samples (Tracks 1, 2, values for promoters given in Fig. 5B) that loss of methylation at the *CES1* promoter occurred against a background of high methylation at this region, which suggested that this control element is normally methylated and silenced, a type that often responds to even small losses of methylation. Additional data to corroborate this could be obtained by examining chromatin state data available through the ChomHMM track in UCSC (Fig. 5C, Track 6), which showed that the promoter falls into the “poised promoter” category (colour-coded pink) and is regulated in part by polycomb-group proteins (grey shading). A



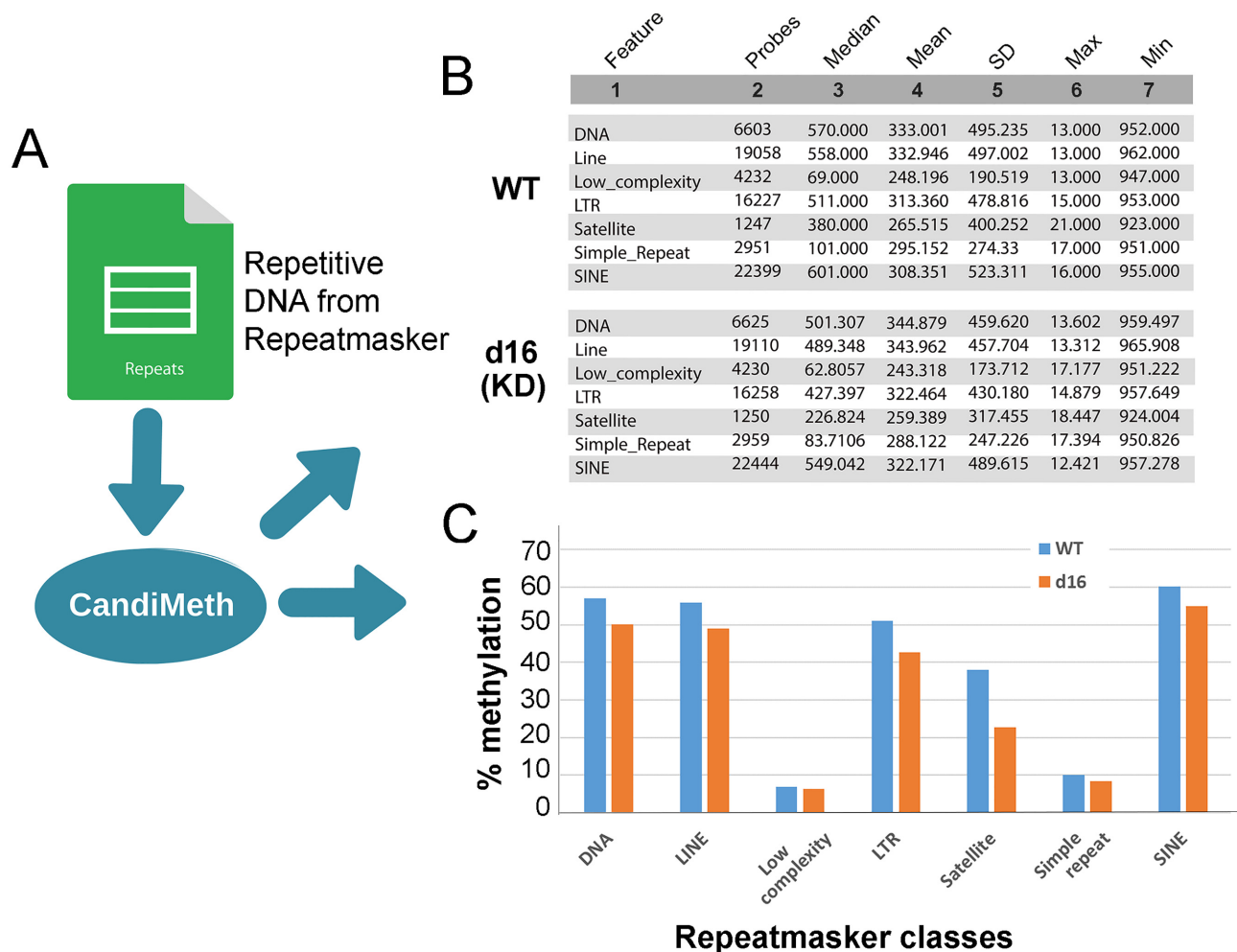
**Figure 5:** Case Study 2: Using CandiMeth to mine EWAS data. Example data from an EWAS dataset comparing 45 placebo and 41 treated samples from a randomized controlled trial of a folic acid intervention during the second and third trimester of pregnancy. (A) Output Results and Tracks from the workflow when the RnBeads differential methylation table and a list of the top-ranked differentially methylated promoters were used as inputs. (B) Summary statistics generated by the workflow indicates the number of probes and methylation values (from 1 to 1,000) for the top promoters. (C) Tracks view for the CES1 locus showing the absolute levels of methylation (Tracks 1, 2) as well as the most differentially methylated probes (Track 3) located at the promoter (boxed in green). Comparison to ChromHMM data in UCSC (Track 6) shows this to be a poised promoter (pink). Identification of individual CG (numbered in Track 8) facilitated the design of a pyrosequencing assay (Track 4) covering the CG to be validated in the laboratory.

low likelihood of SNPs at the pyroassay region could be confirmed by examination of the Common SNPs dataset (Fig. 5C, Track 7) and individual CpGs labelled by searching using the UCSC query window, and their status in other public datasets highlighted if desired (Fig. 5C, Track 8). Thus CandiMeth allowed quick examination of candidate regions, quantification of differences specifically at these, the assessment of sites that could be verified in the laboratory, exclusion of confounding SNPs, and eased assay design and gave additional valuable insights through mining of UCSC datasets using only a few simple inputs and no coding.

### Case Study 3: Analysis of methylation at genomic repeats such as LINE1

Many studies looking for epigenetic changes also try to assess DNA methylation outside of the coding regions. One common approach is to assess methylation at a highly repetitive interspersed repeat such as LINE1, which is found scattered throughout the genome at ~500,000 copies, so in theory sam-

pling methylation across many locations. This normally has to be done using a separate wet-lab assay such as pyrosequencing because the 450 K and EPIC arrays are designed to cover genes and their associated control elements, not repetitive DNA. However, as has been noted elsewhere [29, 42], a substantial number of probes on the arrays, particularly the EPIC, nevertheless fall within repeats such as LINES and SINES. Taking advantage of this, we parsed data from the RepeatMasker track on UCSC to allow mapping and quantification of methylation at the major repeat classes using array data (Fig. 6A). By simply listing the categories of repeat given by RepeatMasker (as in Suppl. Table 4), it is possible to obtain summary statistics indicating the numbers of probes overlapping the respective elements, together with median methylation, and so forth, from any differential methylation table, in this case from our experiment comparing WT and DNMT1-deficient cell lines (Fig. 6B). It can be seen from the tables that very substantial numbers of probes on the EPIC map to the various repeat classes, with ~20,000 probes in LINE elements spread across the genome, and equal numbers in SINE elements, with satellite repeats near centromeres showing the



**Figure 6:** Case Study 3: Analysis of LINEs and SINEs. Use of CandiMeth to give an overview of methylation at repetitive elements. (A) Data on repeat location and type from the RepeatMasker track on UCSC has been parsed and made available through the workflow: users can therefore simply type in the name(s) of a class of repeats as a query. (B) Example outputs showing probe coverage of repeats on the EPIC array and methylation statistics for each repeat class in a DNMT1 knockdown cell line (d16 KD) versus WT. (C) Tables from B were exported, median methylation levels converted to percent, and then graphed to highlight differences between the 2 cell lines: decreases in methylation are seen at some (LINE, SINE, satellite) but not other repeats (low complexity, simple).

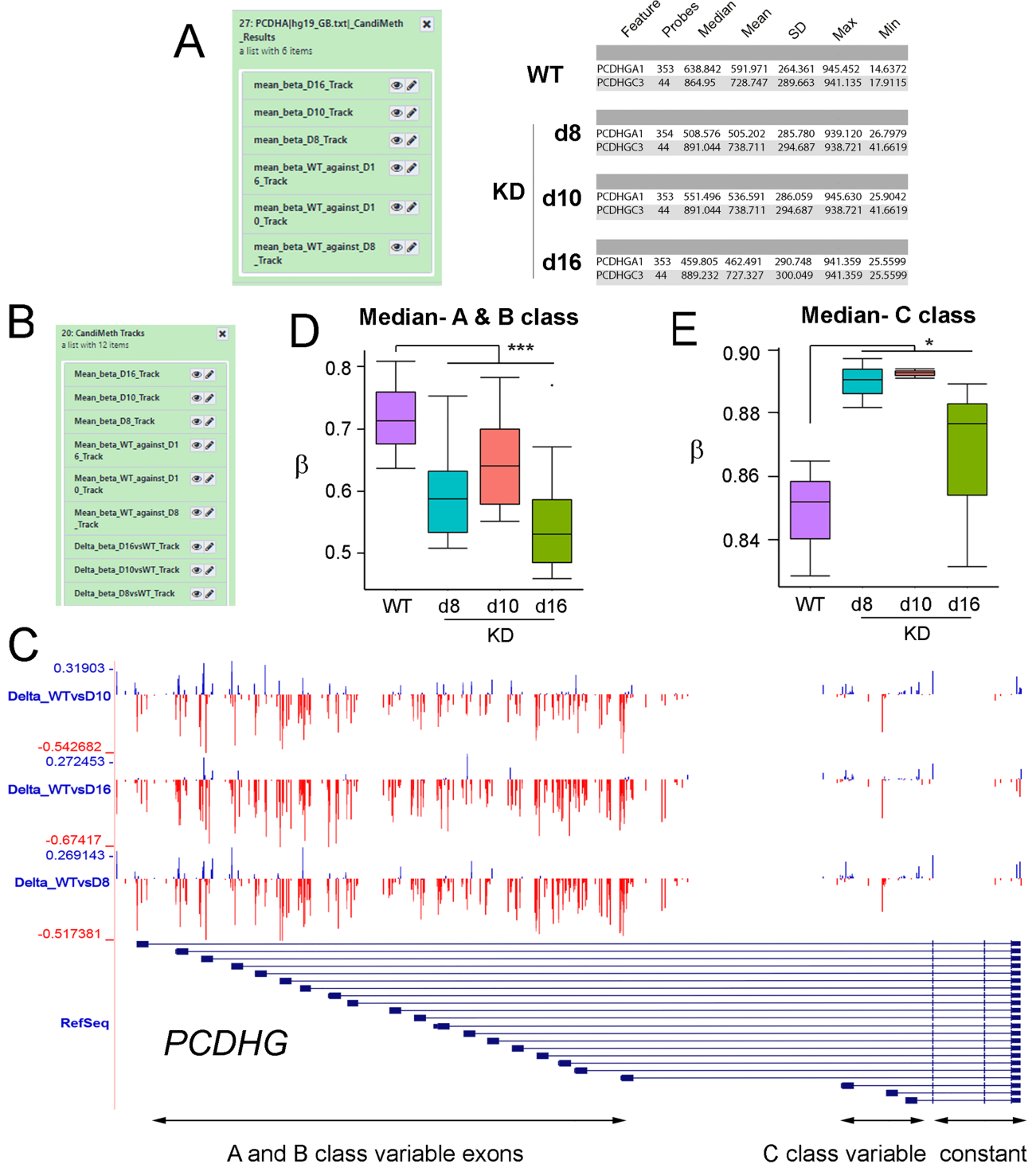
lowest coverage, at  $\sim 1,000$ . The summary data were exported to Excel and graphed to highlight where the greatest differences lay (Fig. 6C), which showed that satellite sequences appear to be most demethylated on average, with notable decreases at LINE and long terminal repeat (LTR)-containing elements too, which would include endogenous retroviruses for example, whereas low-complexity and simple repeats show almost no changes, despite good probe coverage (Fig. 6B). Thus CandiMeth allowed straightforward assessment of repeat methylation across the genome without the need for wet-lab analysis and gave novel insights into the differential effects of DNMT1 loss on individual repetitive DNA classes.

#### Case Study 4: Analysis of methylation changes seen at a large complex gene locus in multiple samples using parallel processing in CandiMeth

A powerful feature of CandiMeth is the ability to process data from multiple differential methylation analyses at once. To illustrate this, we took 3 sets of comparisons between the independently derived DNMT1 KD cell lines described earlier (d8, d10, and d16), each of which had been compared to the parental WT

cell line, and processed them simultaneously. In our earlier publication [25] we had found differences between the variable A and B classes and the variable C class of exons at the important neurodevelopmental gene cluster *Protocadherin  $\beta$*  (*PCDHB*), with the A and B classes showing severe loss of methylation but no change at the C class. This highlighted differences between these classes, which indicate (i) a hyper-dependence on DNMT1 for maintenance of methylation levels and (ii) a potential difference in methylation dependence that may track with allele usage because the A and B classes show monoallelic expression but not the C class. Here, we wished to examine the neighbouring *PCDHG* locus, which has a similar structure, and see whether the same effect could be seen there.

We therefore generated a candidate region list containing the names of the  $\gamma$ -cluster genes and input this as our candidate feature list input to CandiMeth, together with the 3 differential methylation tables from RnBeads (d8 vs WT, d10 vs WT, d16 vs WT). All 3 sets are processed at once (Fig. 7A, left) and give as outputs data on absolute methylation levels in each KD line as well as the WT parental line (which will not vary), from which summary tables were derived specific to the *PCDHG* exons; example data for 1 A and 1 C exon in each cell



**Figure 7: Case Study 4: Parsing data from a complex gene locus using parallel processing.** Analysis of methylation at variable exons in the large (~200 kb) clustered *Protocadherin  $\gamma$*  (*PCDHG*) locus on human chromosome 5. (A) Results: changes at the variable exons across 3 independent cell lines deficient in DNMT1 (d8, d10, d16) were scored using the RnBeads tables comparing each to WT as input, together with a list of variable exon names. Example quantitative outputs are shown at right for WT and knockdown (KD) cells. (B) Tracks: part of the output set of tracks is shown, which included mean  $\beta$ , differential methylation, and FDR significant sites (not shown) for all cell lines, generated simultaneously in 1 run. (C) The UCSC browser view available by following the links in (B). The region covering the A and B class variable exons appears to show more loss of methylation while the C class appear to show predominantly gains; however, this is not exclusive and many probes lie between exons. (D, E) Data on probes that lie solely in exons and not introns, obtained through Results (B), were exported and grouped as indicated. The numbers were then converted back into  $\beta$ -values and graphed. This confirmed that methylation was lost on average at the A and B class exons, while the C class predominantly gained methylation. \*\*\* $P < 0.001$ , \* $P < 0.05$  by Kruskal-Wallis test.



line only are shown (Fig. 7A, right). Interestingly, the summary statistics indicated that, while levels of methylation appeared to be decreased across A and B class variable exons at this locus too (e.g., *PCDHGA1* 63.8% median methylation in WT vs 50.9%, 55.1%, and 46% in d8, d10, and d16, respectively), median methylation at C class variable exons appeared to be increasing rather than remaining constant (e.g., *PCDHGC3* 86.5% in WT vs 89.1%, 89.1%, and 88.9% in d8, d10, and d16).

CandiMeth additionally generated Tracks outputs including the full range of tracks for each input table (absolute methylation in WT and each KD,  $\Delta\beta$  and FDR for each vs WT). In Fig. 7C we show the differential methylation ( $\Delta\beta$ ) tracks, from which it appeared that methylation was largely lost across the region of the gene containing the A and B class variable exons (Fig. 7C, region boxed in red), although some gains (blue peaks) could be seen particularly in the d10 track. Additionally, given the size of the region (~200 kb) it cannot be assessed whether many of the probes lie in the introns rather than the exons themselves. For the C class exons (Fig. 7C, blue box at right) most changes appeared to be gains (blue) although peak sizes were smaller and interspersed with some individual large losses in red. To resolve the exact nature of the changes seen, the tabular data (Fig. 7A) were exported and median values across all A and B exons vs WT generated, converted back to  $\beta$ -value to allow direct comparison to previous results [26], and plotted (Fig. 7D). This clearly showed a general loss of methylation at A and B class exons in all 3 cell lines ( $P < 0.001$  vs WT by Kruskal-Wallis test), although the effect was least marked in the d10 cell line. When values were averaged in a similar fashion across the C class variable exons, however (Fig. 7E), we saw a clear gain of methylation in all 3 cell lines ( $P < 0.05$ , Kruskal-Wallis test). The reason this effect was not noted before is likely to be because our previous examination of the C class exons at *PCDHG* used the FDR-significant probes only, and as can be seen the magnitude of the gains at the C class exons is much smaller than the losses at the A and B classes (compare scales in Fig. 7D and E).

The analysis thus confirmed and extended observations from our previous study that the A and B class variable exons at the clustered protocadherin loci are hypersensitive to loss of DNMT1 across multiple independently derived cell lines, suggesting a strong dependence on this enzyme for maintenance of epigenetic state at this important neurodevelopmental locus. Furthermore, we have uncovered new evidence for differences between the A and B exons and the C exons, which may reflect divergent transcriptional control, or an increased transcription across the C class exons in response to loss of DNMT1, in line with observations that intragenic DNA methylation is associated with transcription at active loci [9, 43]. In terms of CandiMeth functionality, the study highlights the ability of the workflow to process multiple comparisons in parallel and the value of being able to directly compare the visual outputs and the quantitative data where complex genetic loci are being examined, giving insights into the underlying biology.

## Conclusions and Future Directions

CandiMeth provides a user-friendly non-computationally intensive method of candidate feature investigation. With a minimum of training and no coding, users of CandiMeth can set up and run quite advanced exploratory and confirmatory analyses and use the rich set of existing data in UCSC to formulate and test hypotheses regarding the methylation changes that they are seeing.

In future versions, we hope to add support for further methylation processing pipelines and continue to grow the CandiMeth history with additional genomic data such as DNA hypersensitivity sites. In addition to the current pipeline, we also wish to make CandiMeth more intuitive via the creation of a Galaxy tool that would allow the pipeline to be extended to whole-genome bisulphite sequencing or RNA-sequencing data and would also allow further analysis options for those with a private instance of Galaxy.

## Availability of Supporting Source Code and Requirements

Project Name: CandiMeth  
 Project home page: <https://github.com/sjthursby/CandiMeth>  
 Operating system: [www.usegalaxy.org](http://www.usegalaxy.org)  
 License: GNU GPL

## Availability of Supporting Data and Materials

All supporting data and materials are available in the GigaScience GigaDB database [30].

## Supplementary Materials

**Supplementary File 1:** CandiMeth User Guide. A complete Guide to setting up and using CandiMeth, including some background on Galaxy and UCSC browser, how to import the workflow and example files, tutorials on the use of the example data, and further guidance and instruction.

**Supplementary Table 1:** Example Differential Methylation Table generated by RnBeads from GSE90012 for input to CandiMeth. Table comparing wild-type hTERT1604 human fibroblasts (WT) and a clonally derived daughter cell line with depleted levels of DNA methyltransferase 1 (d8) from GEO database entry GSE90012, used as Input 2 to CandiMeth in Case Study 1 (Fig. 2).

**Supplementary Table 2:** MIR gene list used to query data from GSE90012. List of human microRNA genes (MIR) used as Input 3 to CandiMeth in Case Study 1.

**Supplementary Table 3:** Methylation summary for MIR genes derived by CandiMeth. Full table of Results for MIR methylation in GSE90012 WT vs DNMT1-depleted (d8) cells given as output from CandiMeth (Fig. 2B).

**Supplementary Table 4:** Classes of repetitive DNA sequence that can be analysed. List of repetitive DNA classes as given by RepeatMasker and that can be used as Input 3 by CandiMeth to query datasets, as in Case Study 3 (Fig. 6).

**Supplementary Table 5:** Example Differential Methylation Table from ChAMP. Example data in ChAMP format for use in tutorial as Input 2 in CandiMeth.

## Abbreviations

BLAT: BLAST-Like Alignment Tool; bp: base pairs; CGI: CpG island; ChIP: chromatin immunoprecipitation; CSV: comma-separated values; DMR: differentially methylated region; EWAS: epigenome-wide association studies; FDR: false discovery rate; GEO: Gene Expression Omnibus; GSEA: Gene Set Enrichment Analysis; HGNC: Human Genome Nomenclature Committee; kb: kilobase pairs; KD: knock-down; LINE: long interspersed nuclear element; LTR: long terminal repeat; MIR: microRNA; NCBI: National Center for Biotechnology Information; RefSeq: Reference Sequence; SINE: short interspersed nuclear element; SNPs:

single-nucleotide polymorphisms; UCSC: University of California Santa Cruz; WT: wild type.

## Competing Interests

The authors declare that they have no competing interests.

## Funding

Work was funded in part by grants from the Medical Research Council (MR/J007773/1), the EpiFASST grant from the ESRC/BBSRC(ES/N000323/1), and the HDHL EpiBrain award from the BBSRC (BB/S020330/1).

## Authors' Contributions

S.J.T. generated and tested the workflow and accompanying datasets and drafted the manuscript; D.K.L. helped with SED and workflow debugging; R.E.I. provided supervision and feedback on early versions; K.P. and S.D.Z. provided guidance and comments; C.P.W. designed the study, generated early versions of the workflow, and co-wrote the manuscript; all authors commented on and approved the final version.

## Acknowledgements

We are grateful to members of the C.P.W. lab for beta-testing of the workflow and accompanying histories in Galaxy, and especially to Sarah Cairns for help with the Guide that accompanies the workflow on GitHub.

## References

- Schübeler D. Function and information content of DNA methylation. *Nature* 2015;**517**:321–6.
- Deaton AM, Bird A. CpG islands and the regulation of transcription. *Genes Dev* 2011;**25**:1010–22.
- Lyko F. The DNA methyltransferase family: a versatile toolkit for epigenetic regulation. *Nat Rev Genet* 2017;**19**:81–92.
- Bartolomei MS, Ferguson-Smith AC. Mammalian genomic imprinting. *Cold Spring Harb Perspect Biol* 2011;**3**:a002592.
- Mountoufaris G, Canzio D, Nwakeze CL, et al. Writing, reading, and translating the clustered protocadherin cell surface recognition code for neural circuit assembly. *Annu Rev Cell Dev Biol* 2018;**34**:471–93.
- Walsh CP, Chaillet JR, Bestor TH. Transcription of IAP endogenous retroviruses is constrained by cytosine methylation. *Nat Genet* 1998;**20**:116–7.
- Xu GL, Bestor TH, Bourc'his D, et al. Chromosome instability and immunodeficiency syndrome caused by mutations in a DNA methyltransferase gene. *Nature* 1999;**402**:187–91.
- Irwin RE, Thakur A, O' Neill KM, et al. 5-Hydroxymethylation marks a class of neuronal gene regulated by intragenic methylcytosine levels. *Genomics* 2014;**104**(5):383–92.
- Wu H, Coskun V, Tao J, et al. Dnmt3a-dependent nonpromoter DNA methylation facilitates transcription of neurogenic genes. *Science* 2010;**329**:444–8.
- Sapienza C, Issa J-P. Diet, nutrition, and cancer epigenetics. *Annu Rev Nutr* 2016;**36**:665–81.
- Stevens AJ, Rucklidge JJ, Kennedy MA. Epigenetics, nutrition and mental health. Is there a relationship? *Nutr Neurosci* 2018;**21**:602–13.
- Vaz M, Hwang SY, Kagiampakis I, et al. Chronic cigarette smoke-induced epigenomic changes precede sensitization of bronchial epithelial cells to single-step transformation by KRAS mutations. *Cancer Cell* 2017;**32**:360–76.e6.
- Abdul QA, Yu BP, Chung HY, et al. Epigenetic modifications of gene expression by lifestyle and environment. *Arch Pharm Res* 2017;**40**:1219–37.
- Suzuki M, Liao W, Wos F, et al. Whole-genome bisulfite sequencing with improved accuracy and cost. *Genome Res* 2018;**28**:1364–71.
- Bibikova M, Barnes B, Tsan C, et al. High density DNA methylation array with single CpG site resolution. *Genomics* 2011;**98**:288–95.
- Assenov Y, Müller F, Lutsik P, et al. Comprehensive analysis of DNA methylation data with RnBeads. *Nat Methods* 2014;**11**:1138–40.
- Morris TJ, Butcher LM, Feber A, et al. ChAMP: 450k Chip analysis methylation pipeline. *Bioinformatics* 2014;**30**:428–30.
- Tian Y, Morris TJ, Webster AP, et al. ChAMP: Updated methylation analysis pipeline for Illumina BeadChips. *Bioinformatics* 2017;**33**:3982–4.
- Subramanian A, Tamayo P, Mootha VK, et al. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A* 2005;**102**:15545–50.
- Galaxy Homepage. [www.usegalaxy.org](http://www.usegalaxy.org). accessed 1 May 2020.
- Giardine B, Riemer C, Hardison RC, et al. Galaxy: A platform for interactive large-scale genome analysis. *Genome Res* 2005;**15**:1451–5.
- Goble CA, De Roure DC. myExperiment: Social networking for workflow-using e-scientists. In: WORKS '07: Proceedings of the 2nd workshop on Workflows in support of large-scale science. New York, NY: ACM; 2007:1–2.
- CandiMeth Github. <https://github.com/sjthursby/CandiMeth>. Accessed 10th October 2019.
- Afgan E, Baker D, Batut B, et al. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update. *Nucleic Acids Res* 2018;**46**(W1):W537–44.
- O'Neill KM, Irwin RE, Mackin S-J, et al. Depletion of DNMT1 in differentiated human cells highlights key classes of sensitive genes and an interplay with polycomb repression. *Epigenetics Chromatin* 2018;**11**(1):12.
- Irwin RE, Thursby S-J, Ondičová M, et al. A randomized controlled trial of folic acid intervention in pregnancy highlights a putative methylation-regulated control element at ZFP57. *Clin Epigenetics* 2019;**11**:31.
- Batut B, Hiltmann S, Bagnacani A, et al. Collections: Using dataset collection. [galaxyproject.github.io](http://galaxyproject.github.io). 2017.
- O'Leary NA, Wright MW, Brister JR, et al. Reference sequence (RefSeq) database at NCBI: Current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res* 2016;**44**:D733–45.
- Thursby S-J. CandiMeth History. 2019. <https://github.com/sjthursby/CandiMeth>. Accessed 2 February 2020.
- Thursby S, Lobo DK, Pentieva K, et al. Supporting data for "CandiMeth: Powerful yet simple visualization and quantification of DNA methylation at candidate genes." *Gigascience Database* 2020. <http://doi.org/10.5524/100753>.
- Galaxy Team. Convert delimiters to TAB. [https://toolshed.g2.bx.psu.edu/view/devteam/convert\\_characters/8a53d7f02ce4](https://toolshed.g2.bx.psu.edu/view/devteam/convert_characters/8a53d7f02ce4). Accessed 2 December 2019.
- Von Kuster G, Cock P, Soranzo N, et al. Add Column. 2018. [https://toolshed.g2.bx.psu.edu/view/devteam/add\\_value/745871c0b055](https://toolshed.g2.bx.psu.edu/view/devteam/add_value/745871c0b055). Accessed 7 August 2019.

33. Gruening B. Text transformation with SED. 2016. <https://github.com/bgruening/galaxytools/tree/master/tools/sed>. Accessed 6 August 2019.
34. Van Den Beek M. Compute expression on every row. 2017. [https://github.com/galaxyproject/tools-iuc/blob/1bf170897075f6ca390128cd5666afcd944f3aa/tools/column\\_maker/column\\_maker.xml](https://github.com/galaxyproject/tools-iuc/blob/1bf170897075f6ca390128cd5666afcd944f3aa/tools/column_maker/column_maker.xml). Accessed 7 August 2019.
35. Mabon P, Chilton J, Van Den Beek M. Merge collections. 2018. [https://github.com/galaxyproject/galaxy/blob/ed6f5d45a9ecbf0cf8a85341e0213b82a9132a46/lib/galaxy/tools/merge\\_collection.xml](https://github.com/galaxyproject/galaxy/blob/ed6f5d45a9ecbf0cf8a85341e0213b82a9132a46/lib/galaxy/tools/merge_collection.xml). Accessed 2 February 2020.
36. Van Den Beek M, Chilton J. Relabel list identifiers from contents of a file. 2018. [https://github.com/galaxyproject/galaxy/blob/ed6f5d45a9ecbf0cf8a85341e0213b82a9132a46/lib/galaxy/tools/relabel\\_from\\_file.xml](https://github.com/galaxyproject/galaxy/blob/ed6f5d45a9ecbf0cf8a85341e0213b82a9132a46/lib/galaxy/tools/relabel_from_file.xml). Accessed 6 August 2019.
37. Mabon P. Collapse collection. 2017. [https://github.com/pha-c-nml/galaxy\\_tools/blob/master/tools/collapse\\_collection/merge.xml](https://github.com/pha-c-nml/galaxy_tools/blob/master/tools/collapse_collection/merge.xml). Accessed 7 February 2020.
38. Galaxy Team. Paste two files side by side. 2017. <https://github.com/galaxyproject/galaxy/blob/ed6f5d45a9ecbf0cf8a85341e0213b82a9132a46/tools/filters/pasteWrapper.xml>. Accessed 2 February 2020.
39. Gruening B, Gamaleldin H, Soranzo N, et al. Text Processing Awk. 2018. [https://github.com/bgruening/galaxytools/blob/5c6486dead878a8c9521e1d6d50b3a537a2ec2b0/tools/text\\_processing/text\\_processing/awk.xml](https://github.com/bgruening/galaxytools/blob/5c6486dead878a8c9521e1d6d50b3a537a2ec2b0/tools/text_processing/text_processing/awk.xml). Accessed 6 August 2019.
40. Gruening B, Gamaleldin H. Unique occurrences of each record. 2015. [https://github.com/bgruening/galaxytools/blob/5c6486dead878a8c9521e1d6d50b3a537a2ec2b0/tools/text\\_processing/text\\_processing/unordered\\_uniq.xml](https://github.com/bgruening/galaxytools/blob/5c6486dead878a8c9521e1d6d50b3a537a2ec2b0/tools/text_processing/text_processing/unordered_uniq.xml). Accessed 6 August 2020.
41. Baker D, Matthias B, Coraor N, et al. Group data by a column and perform aggregate operations on other columns. 2011. <https://github.com/galaxyproject/galaxy/blob/d06718ef24202d2fcf3d6e8bfaa9a862044f90f/tools/stats/grouping.xml>. Accessed 12 February 2020.
42. O'Neill KM, Irwin RE, Mackin S-J, et al. Depletion of DNMT1 in differentiated human cells highlights key classes of sensitive genes and an interplay with polycomb repression. *Epigenetics Chromatin* 2018; 11:12.
43. Kent WJ. BLAT—the BLAST-Like Alignment Tool. *Genome Res* 2002;12:656–64.

# CandiMeth User Guide

## Contents

1. Overview: where to start .....	2
2. Getting set up.....	3
3. Using CandiMeth with the sample data provided .....	6
3.1 Tutorial with the example data provided .....	6
3.1.1 Inputs and starting the run .....	6
3.1.2 Overview of outputs from the microRNA analysis .....	8
3.1.3 Working with the output Results tables .....	9
3.1.4 Graphing data directly in Galaxy .....	10
3.1.5 Exporting data from Galaxy to Excel .....	11
3.1.6 Working with output Tracks.....	14
3.1.7 Exporting browser views as graphics files .....	20
3.1.8 Quantifying methylation in different parts of the gene .....	22
3.2 Looking at a new set of genes in the current methylation array dataset.....	23
3.3 Looking at repetitive DNA elements such as LINES and ERV .....	25
3.4 Using ChAMP-generated methylation data .....	27
4. Uploading and working with your own differential methylation data .....	28
4.1 Locating data files in RnBeads .....	28
4.2 Locating data files in ChAMP .....	30
4.3 Uploading your data files to Galaxy .....	31
Appendix 1. Primer for those unfamiliar with the UCSC genome browser .....	33
Appendix 2. Quick guide to the Galaxy web environment .....	36
Appendix 3. Importing CandiMeth to a custom Galaxy instance .....	37

## 1. Overview: where to start

CandiMeth is aimed primarily at wet-lab biologists working in the general area of biomedical sciences who have access to DNA methylation data and want to look at methylation levels at specific candidate genes or regions, but are not adept at coding or programming.

Most commonly, the end-user of CandiMeth would have a report generated in html by one of the two main DNA methylation array processing pipelines RnBeads or ChAMP, which will give overviews of the data such as PCA analysis, graphs etc and then links to the processed data in the form of differential methylation tables showing beta values from the arrays, differences in methylation between samples and p values.

If you have tables of differential methylation values from RnBeads or ChAMP and want to look at specific genes then CandiMeth is a suitable tool.

CandiMeth will allow you to visualise the methylation at your genes of interest, as well as quantify it. Although there are various options to allow you to visualise the methylation, we recommend the UCSC genome browser, as this has a wide range of extra data on the genes which can enrich and inform your analysis.

- For those not familiar with the UCSC browser, see the Primer (Appendix 1)

CandiMeth basically takes the differential methylation data output by the RnBeads or ChAMP pipelines and uses the Galaxy web-based bioinformatics platform to map the data to the human genome and to assess and quantify regions of interest.

- For those unfamiliar with the Galaxy environment, see the Quick guide (Appendix 2)

CandiMeth users do not need to be very adept with either UCSC browser or Galaxy for CandiMeth to work, as the interface is very simple and the outputs in terms of visualisation and quantification very straightforward. If you have used the UCSC browser before, and have looked over the Galaxy webpage, then you should be able to try CandiMeth with little prior preparation.


- For those comfortable with the UCSC browser and who understand the basic screen layout in Galaxy, you need to get set up with the CandiMeth workflow and example data (Section 2)
- After that, we recommend using CandiMeth the first time with the sample data provided (Section 3)
- Once CandiMeth is working on your computer, you can try uploading and working with your own data (Section 4)
- Some additional help is available to enable conversion of ChAMP outputs to Excel files for import into CandiMeth (Section 4.2)
- If you run Galaxy on a local server or have a customised version, we provide a guide to importing CandiMeth below (Appendix 3)

## 2. Getting set up

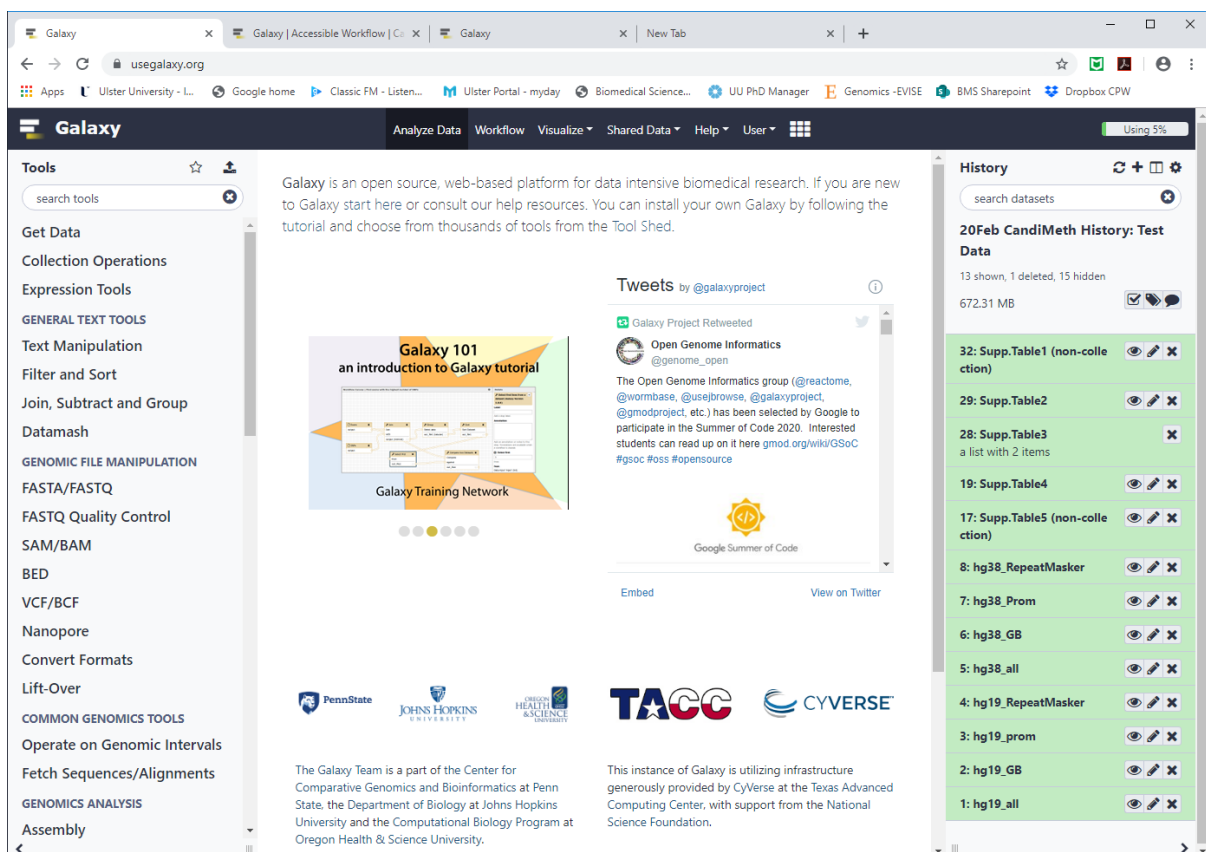
[Note: CandiMeth, Galaxy and UCSC work best in Chrome or Firefox rather than Explorer. If Chrome is not your default browser, cut and paste the web-addresses in brackets into your browser window instead.]

- 1) CandiMeth runs in the Galaxy on-line environment: you will therefore need a working Galaxy account, which can be created for free [here](http://www.usegalaxy.org) (<http://www.usegalaxy.org>) - those not familiar with Galaxy are referred to the brief introduction below

[Tip! It may take a couple of seconds for the hyperlink to open.]

- 2) Once you have an account, click [here](http://bit.do/candimeth-history) (<http://bit.do/candimeth-history>) and click on the  button at the top RHS of the page – this will create a History in your own Galaxy account containing the reference genome information and some example data to be used in the CandiMeth workflow

It should look something like this:



The screenshot shows the Galaxy web interface. On the left is a 'Tools' sidebar with categories like 'Get Data', 'Expression Tools', and 'Genomic File Manipulation'. The main content area features a 'Galaxy 101' tutorial, a tweet from the Open Genome Informatics group, and logos for PennState, Johns Hopkins, and TACC. On the right is a 'History' sidebar showing a collection of datasets, including 'Supp.Table1' through 'Supp.Table5' and 'hg38\_RepeatMasker' through 'hg19\_all'.

- 3) The input Differential Methylation Table has to be converted from a table into the form of a Dataset Collection: this is in case there are multiple differential methylation tables to be assessed, then CandiMeth can assess them all simultaneously and present them in the typical Results and Tracks outputs, as opposed to multiple outputs that might make your history very crowded or initiate multiple histories that may become confusing due to their number.


There are two example files which can be converted and used, one which is an *RnBeads* output (Suppl.Table1) and one a *ChAMP* output (Suppl.Table5).

- a) Click on the already checked box at the top of the History panel (mouse over shows “Operations on multiple datasets”): this will cause checkboxes to appear beside all of your datasets as well as some choices to appear at top
- b) Check the box beside the Differential Methylation Table dataset e.g. Suppl.Table1
- c) Under the pulldown menu beside “For all selected” choose “Build Dataset List”
- d) In the window that appears, you can give the collection a new name e.g. “All Probes Set1” and click “Create”
- e) A new entry will appear in the RHS with the new name and “a list with 1 (or more) items”- this is the Dataset Collection and is now ready to be processed by CandiMeth
- f) -Click on the Check box at top right again as in (a) to go back to normal list view in RHS window

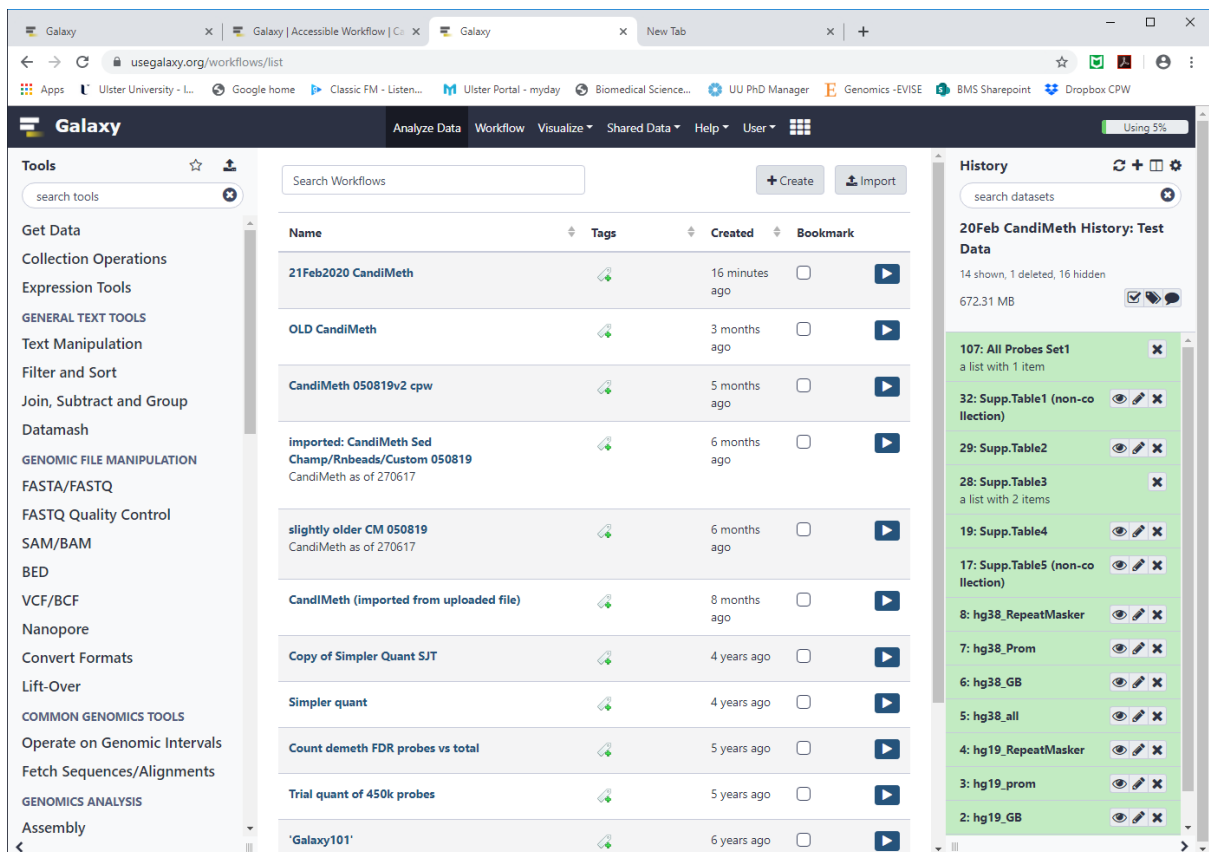
[Note: In the examples here and below, names which the user can choose themselves are in green]

It should look something like this at RHS, with the new Dataset Collection (“All Probes Set1”) at top



- 4) Following this, click [here](http://bit.do/candimeth) (<http://bit.do/candimeth>) and then click the  button at the top RHS of the page – this will import the CandiMeth workflow [CandiMeth] to your Galaxy account so you can use it. Click on “Start using this workflow” in the window that appears to bring you to the Galaxy webpage again

This should look something like this:



The screenshot shows the Galaxy web interface. The central window displays a table of workflows with the following data:

Name	Tags	Created	Bookmark
21Feb2020 CandiMeth		16 minutes ago	<input type="checkbox"/>
OLD CandiMeth		3 months ago	<input type="checkbox"/>
CandiMeth 050819v2 cpw		5 months ago	<input type="checkbox"/>
imported: CandiMeth Sed Champ/Rnbeads/Custom 050819 CandiMeth as of 270617		6 months ago	<input type="checkbox"/>
slightly older CM 050819 CandiMeth as of 270617		6 months ago	<input type="checkbox"/>
CandiMeth (imported from uploaded file)		8 months ago	<input type="checkbox"/>
Copy of Simpler Quant SJT		4 years ago	<input type="checkbox"/>
Simpler quant		4 years ago	<input type="checkbox"/>
Count demeth FDR probes vs total		5 years ago	<input type="checkbox"/>
Trial quant of 450k probes		5 years ago	<input type="checkbox"/>
'Galaxy101'		6 years ago	<input type="checkbox"/>

The right-hand side shows the 'History' panel for '20Feb CandiMeth History: Test Data'. It lists several datasets, including '107: All Probes Set1', '32: Supp.Table1 (non-co llection)', '29: Supp.Table2', '28: Supp.Table3', '19: Supp.Table4', '17: Supp.Table5 (non-co llection)', '8: hg38\_RepeatMasker', '7: hg38\_Prom', '6: hg38\_GB', '5: hg38\_all', '4: hg19\_RepeatMasker', '3: hg19\_prom', and '2: hg19\_GB'.

-the central window shows all workflows available to you: CandiMeth should be at the top if it was the last one you imported (the example above shows others the author was using too, will be absent). The RHS window should still be your Test data history.

If you navigate away from this view for whatever reason, you can find it again by going to the top of the Galaxy homepage and clicking on the "workflows" option

You should now be ready for your first test run with the sample data provided.




## 3. Using CandiMeth with the sample data provided

### 3.1 Tutorial with the example data provided

The first example takes data from an experiment where two cell lines were compared: one was normal or wild type (WT), the other had lower levels of the DNMT1 methyltransferase enzyme which methylates DNA (d8). DNA from the two types of cell was isolated and run on the 450K Illumina array to determine methylation levels. This data was processed using the RnBeads pipeline, which logged methylation at each position in each cell line, as well as comparing the levels at each position to see if they were significantly different between the cell lines. The RnBeads analysis was output as a table (Supp.Table1). Here we want to see if there are any differences in methylation at certain microRNA genes between the two cell lines: the list of genes is given in Supp.Table2. To do this we choose a version of the human genome map to work with (the version called hg19) and ask to look at all probes associated with the microRNA (hg19\_all).

#### 3.1.1 Inputs and starting the run

To start, click on the CandiMeth workflow in your Galaxy page (see step 2.4 above) and on the pull-down menu at RHS marked  choose > Run

In the History Options at the top of CandiMeth, “Send results to a new history” click “Yes” (light grey) and give the new history a name of your choosing e.g. “[Date/run identifier] CandiMeth My Test Data”

[NOTE: we advise you to give each new history a unique identifier to avoid confusion]

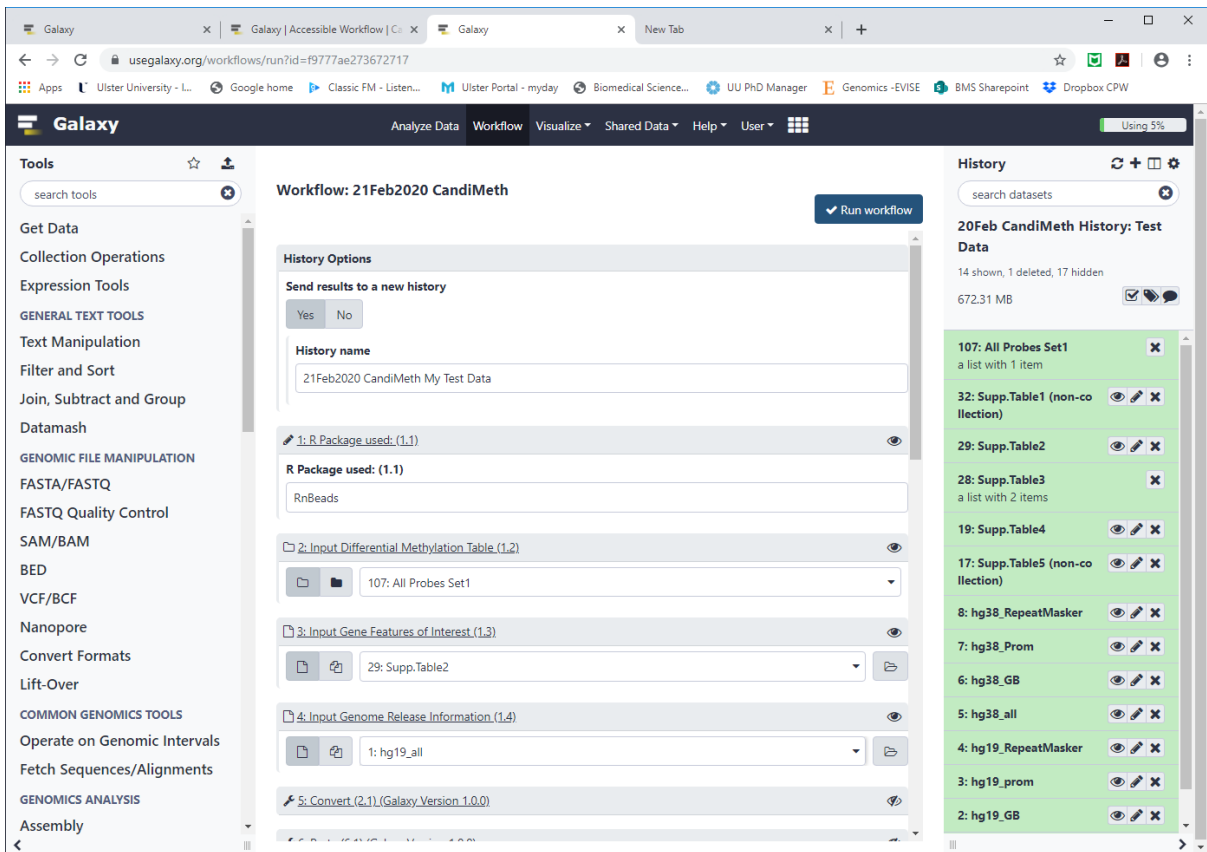
1. Under 1: R Package Used: (1.1) enter ‘RnBeads’
2. For 2: Input Differential Methylation Table (1.2) choose “All Probes Set1”

[Note: This was the example name used in step 2.3 above, alter as required]

3. At 3: Input Gene Features of Interest (1.3) choose “Supp.Table2”
4. For 4: Input Genome Release Information (1.4) choose “hg19\_all”

You can now click the blue ‘Run workflow’ button at top right

The Workflow start window with all the above options chosen should look like the screenshot on the next page:

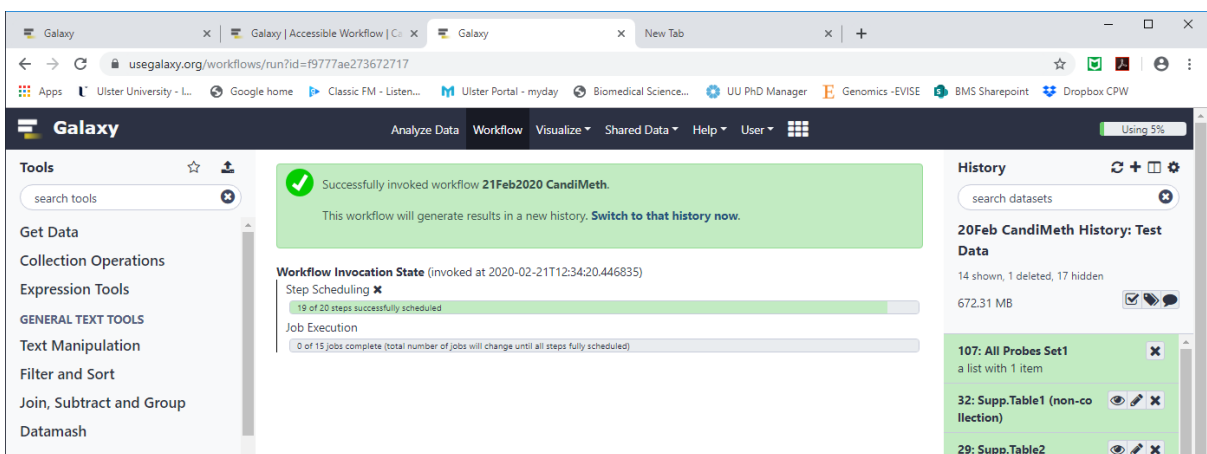



If all goes well, you should see a large green tick in the main (middle) window of Galaxy and the following text:

“Successfully invoked workflow [CandiMeth](#)

This workflow will generate results in a new history. [Switch to that history now.](#)”

The window should also show two status bars, “Step scheduling” and “Job Execution” which will update you on the progress of the jobs.



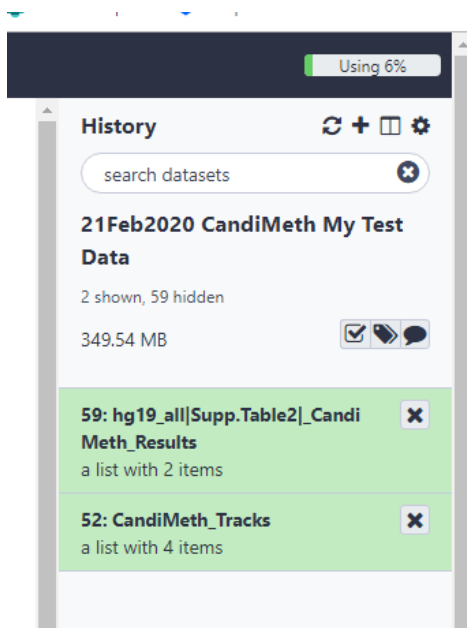
A typical test run with the data above may take ~15 mins to complete. Once both bars have gone green the new  history “[Date/run identifier] CandiMeth My Test Data” is ready.

You can navigate to the new history at any point by following the link “Switch to that history now” at any point, or by navigating between histories using the “Switch to” function at top of the History pane on the RHS.

### 3.1.2 Overview of outputs from the microRNA analysis

CandiMeth produces two types of outputs; tabular Results and genome browser Tracks.

The outputs will look something like this for the example data above :



The two green boxes represent the Results and Tracks output collections respectively. The general format of these is as follows, with details changing one each run. The blue text in each is a link to a more detailed list-

Results:

Number: [genome release\\_probeset|input gene list|\\_CandiMeth\\_Results](#)  
a list with x items

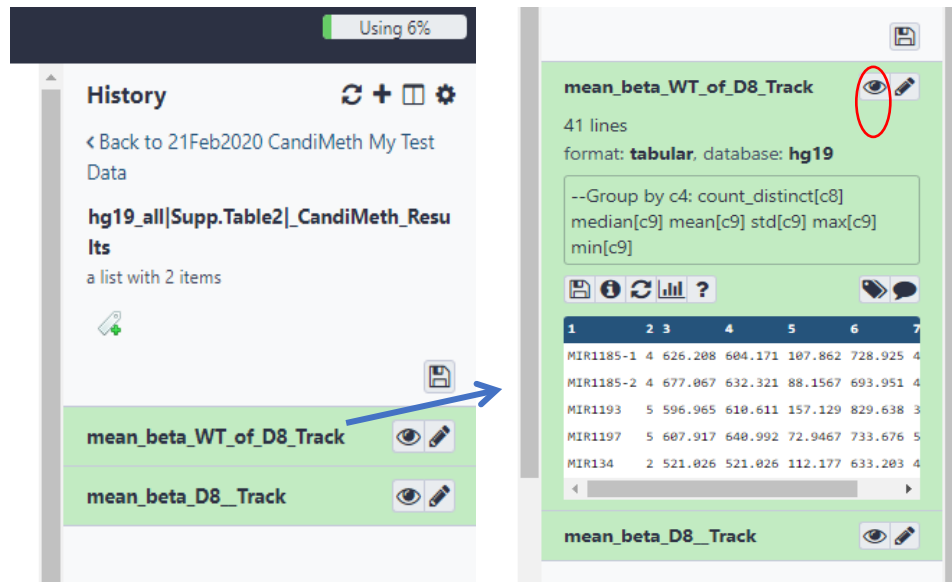
Tracks:

Number: [CandiMeth\\_Tracks](#)  
A list with x items

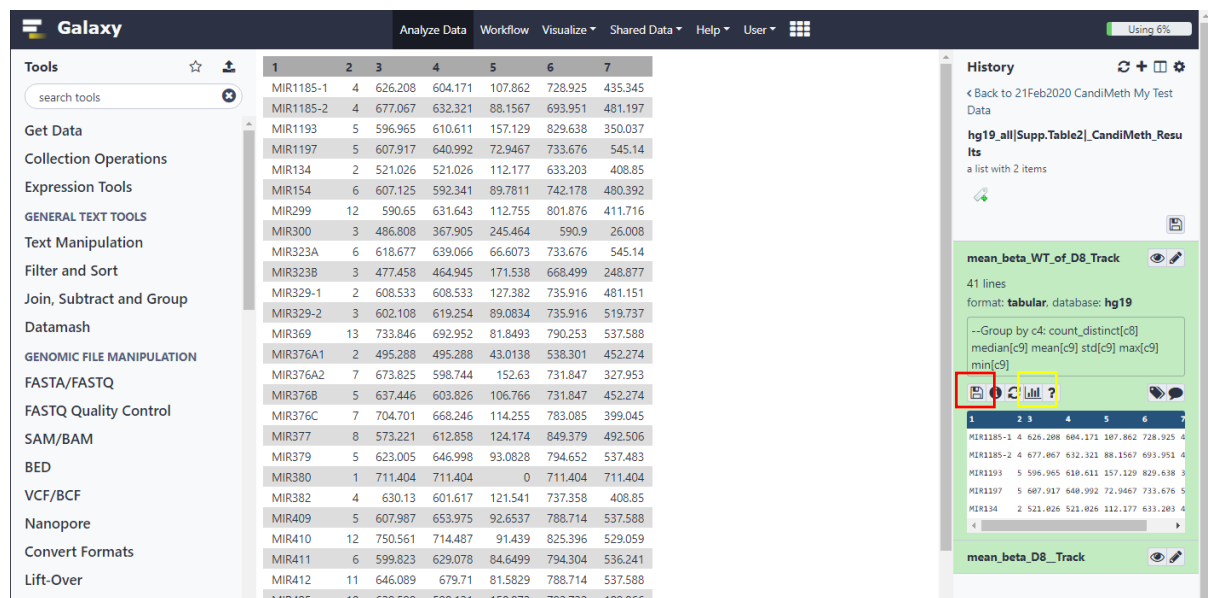
For the Results, the [genome release\\_probeset](#) and [gene list](#) are variables which were decided at the start of the run (see 3.1.1 above), so the choices are recorded in the outputs for clarity. Probeset refers to whether methylation across the promoter only, the gene body only, or both (all) is to be analysed. We will now look at the Results tables (3.1.3) and Tracks (3.1.6).

### 3.1.3 Working with the output Results tables

- To access the tabular results, click on the link saying *CandiMeth\_Results* which will open a new window at RHS. For the example data here, this will show the two items in the list (see screenshot at left below). One is a table of methylation values across microRNA genes in the WT cells [mean\\_beta\\_WT\\_of\\_D8\\_Track](#) and the second the methylation values in the cells with the lower DNMT1 enzyme levels [mean\\_beta\\_D8\\_Track](#).




- Clicking on the [mean\\_beta\\_WT\\_of\\_D8\\_Track](#) will show a preview (screenshot at right above) of the first 5 lines of the data table, as well as the header and other information.
- To see a full table, click on the eye symbol (circled above) and the full table of data will appear in the central Galaxy window as shown below




There are no headers in Galaxy but the key to the columns is as follows:-

- 1) Name of gene, 2) Number of array probes, 3) Median methylation, 4) Standard deviation, 5) Mean methylation, 6) Maximum value and 7) Minimum value

You can work with the results directly in Galaxy, using the Galaxy graphing and stats tool, or you can save  (red box in screenshot above) the data from Galaxy onto your computer to allow you to work with it in other graphing and statistics programs such as Excel and SPSS, see section 3.1.5 below for this latter option.

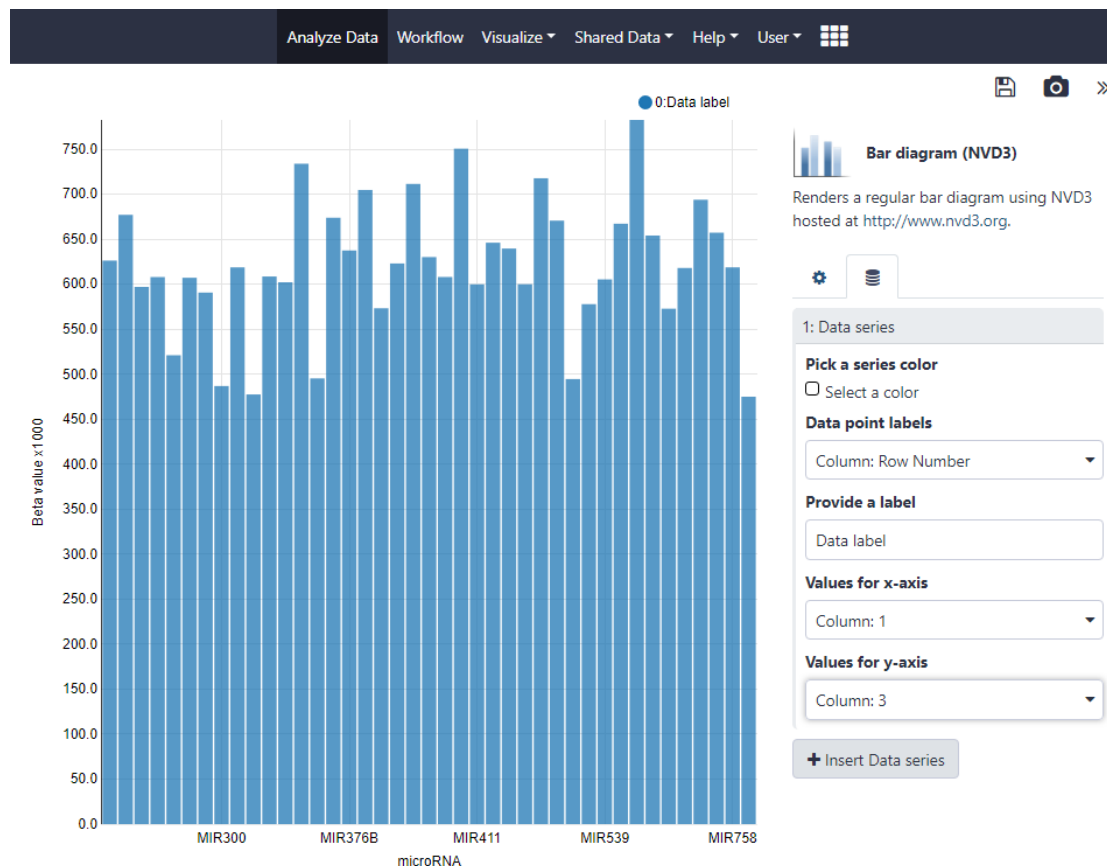
### 3.1.4 Graphing data directly in Galaxy

This option is available to the user for each table: click on the  button in the RHS window (boxed in yellow above). As an example we will use the *Bar diagram (NVD3)* option in the central Galaxy window.


Below we reproduce the first part of the table for reference:-

1	2	3	4	5	6	7
MIR1185-1	4	626.208	604.171	107.862	728.925	435.345
MIR1185-2	4	677.067	632.321	88.1567	693.951	481.197
MIR1193	5	596.965	610.611	157.129	829.638	350.037
MIR1197	5	607.917	640.992	72.9467	733.676	545.14
MIR134	2	521.026	521.026	112.177	633.203	408.85
MIR154	6	607.125	592.341	89.7811	742.178	480.392
MIR299	12	590.65	631.643	112.755	801.876	411.716
MIR300	3	486.808	367.905	245.464	590.9	26.008

The key to the column headers is as above (3.1.3 step 3): Name, number of probes, median etc  
 On the Bar diagram tool window in the centre window we generated the following chart showing median methylation at the microRNA by choosing column 3 (median) in the Data series window, and adding labels etc to the graph. This is one example, there are many other options to explore.



### 3.1.5 Exporting data from Galaxy to Excel

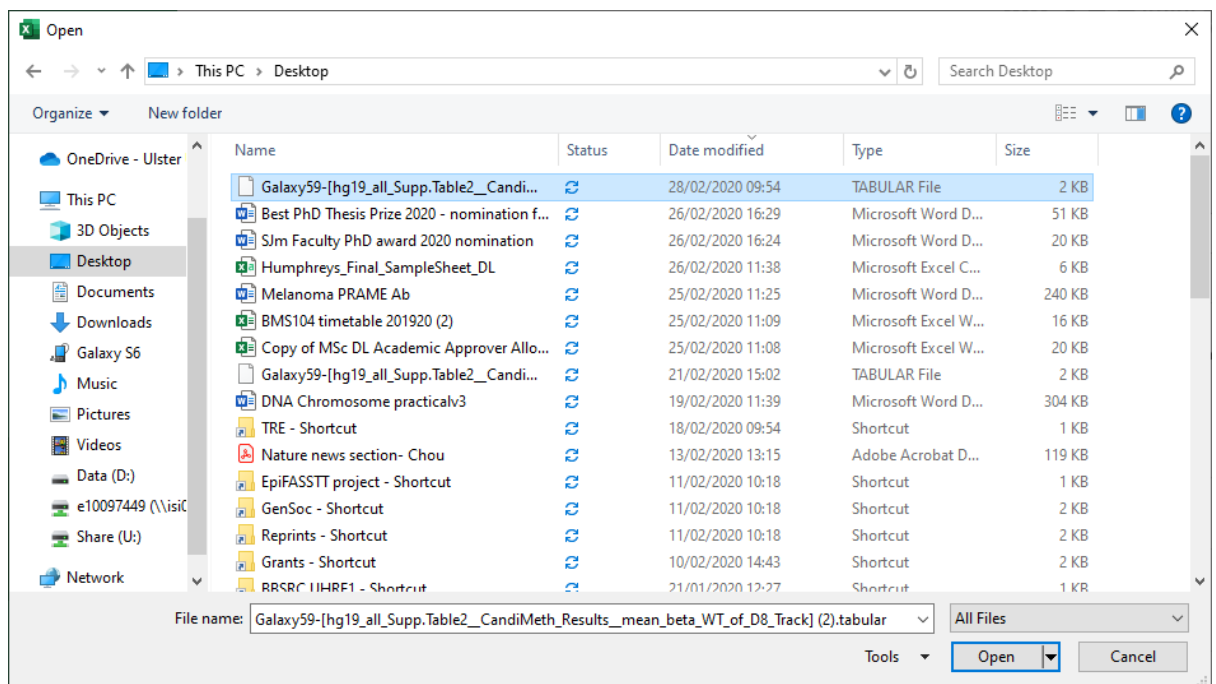
Tables of generated results can be downloaded and imported into programs such as Excel and Notebook by selecting the  button just above the individual output in the RHS window (see p9): this will download the data in a generic tabular file format. The name of the file will be similar to that seen before (see 3.1.2 above), but will have some additional information, an example is-

GalaxyNumber: genome\_release\_probeset|input gene list|\_CandiMeth\_Results\_mean\_beta.tabular

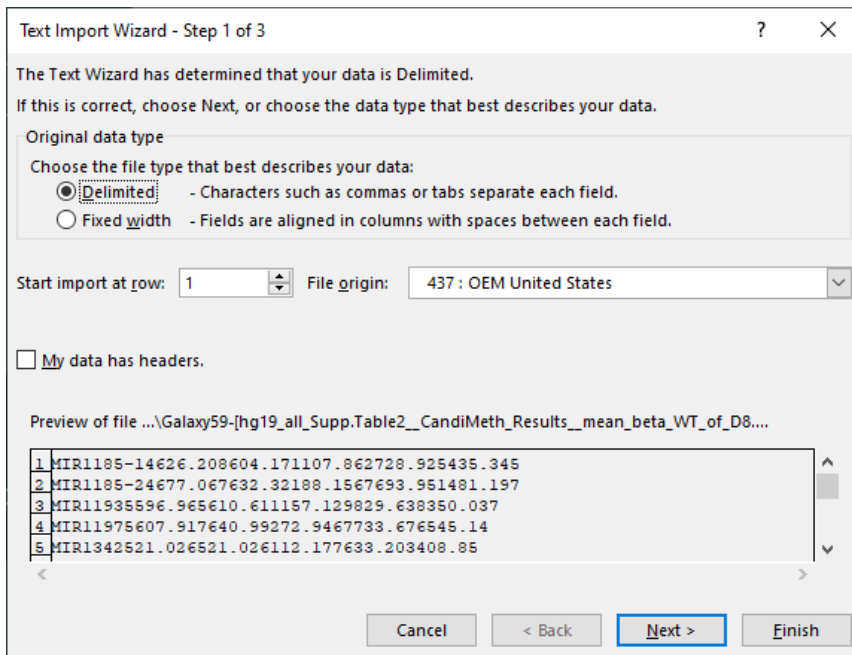
This file will have a `.tabular` suffix which allows it to be imported into a number of different programs such as SPSS or Excel. You should first save or move it to a specific folder on your computer.

To import a `.tabular` file into Excel:

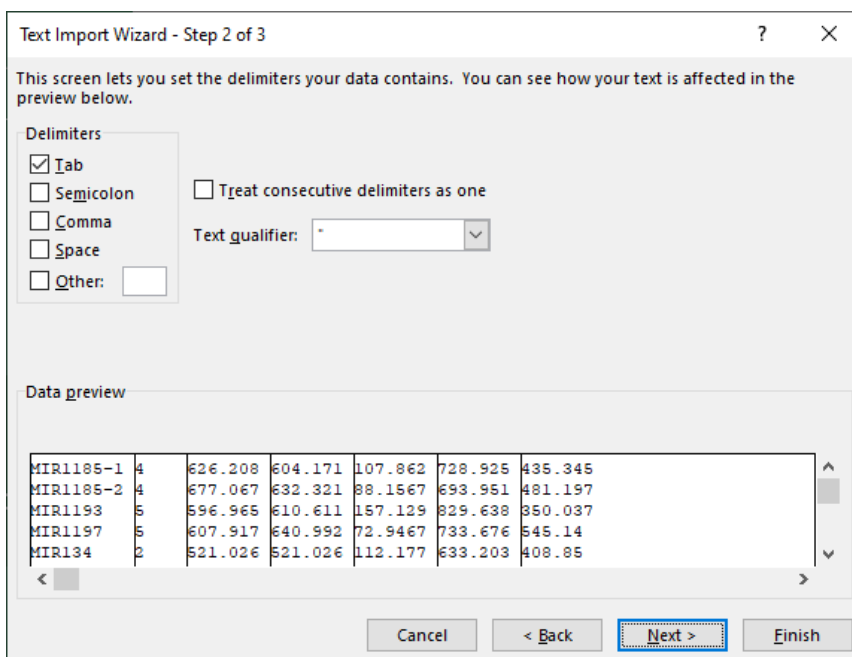
1. Open Excel and using the Open command, locate the folder containing the `.tabular` file [Note: the file may not be visible unless you choose “all files” in the pull-down menu to the right of the *File name* window, since it is not a standard Excel suffix (.xls) ]



2. Select the Galaxy file you want to import into Excel as above, then click Open. This will cause the Text Import Wizard window to automatically open



- Go with the default option *Delimited*, click next and on the next window the default *Tab*,



- In the third and final window choose *General* (for formatting of columns): the imported file should then automatically open in Excel and look like the window below-

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	MIR1185-1	4	626.208	604.171	107.862	728.925	435.345						
2	MIR1185-2	4	677.067	632.321	88.1567	693.951	481.197						
3	MIR1193	5	596.965	610.611	157.129	829.638	350.037						
4	MIR1197	5	607.917	640.992	72.9467	733.676	545.14						
5	MIR134	2	521.026	521.026	112.177	633.203	408.85						
6	MIR154	6	607.125	592.341	89.7811	742.178	480.392						
7	MIR299	12	590.65	631.643	112.755	801.876	411.716						
8	MIR300	3	486.808	367.905	245.464	590.9	26.008						
9	MIR323A	6	618.677	639.066	66.6073	733.676	545.14						

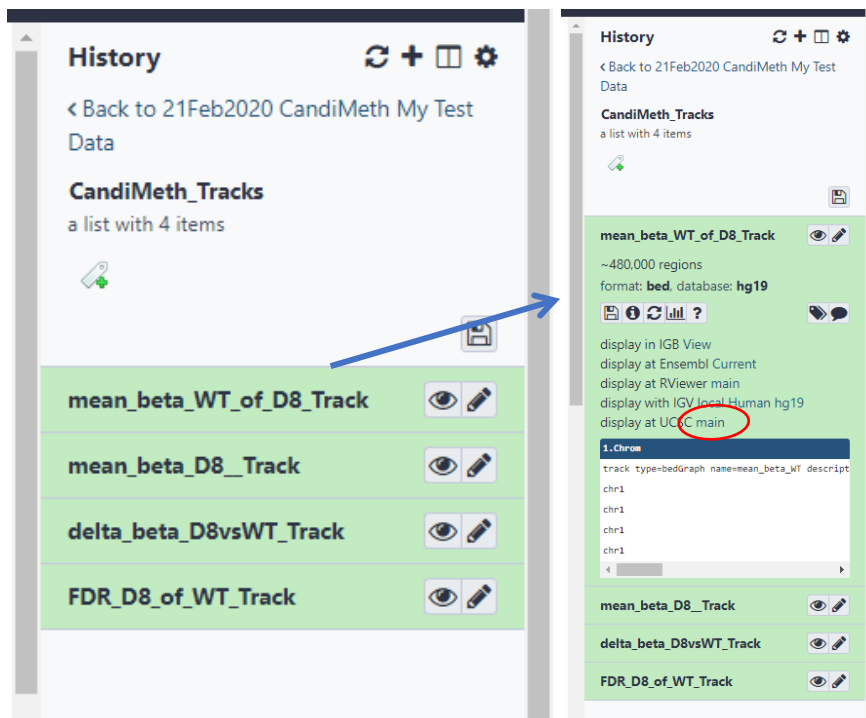
5. This should exactly match your data output from Galaxy. There are no headers in Galaxy so you should add the row of output labels at top yourself:-  
column A) Name of gene, B) Number of array probes, C) Median methylation, D) Standard deviation, E) Mean methylation, F) Maximum value and G) Minimum value.
6. It would be sensible to give the Excel table a simpler name reflecting the data type eg “MicroRNA methylation in WT cells”, but each file will have the unique identifiers automatically embedded in the long file name by default
7. In a similar fashion the data on methylation in the D8 cells can also be imported into an Excel file. Data from these files can then be cut and pasted into one file to allow direct graphing and statistical comparisons in Excel as indicated in the CandiMeth paper and bibliography therein

To navigate back to the window showing both Tracks and Results, just click on the link at the top of the RHS screen which should say “<Back to....My Test Data” or similar



### 3.1.6 Working with output Tracks

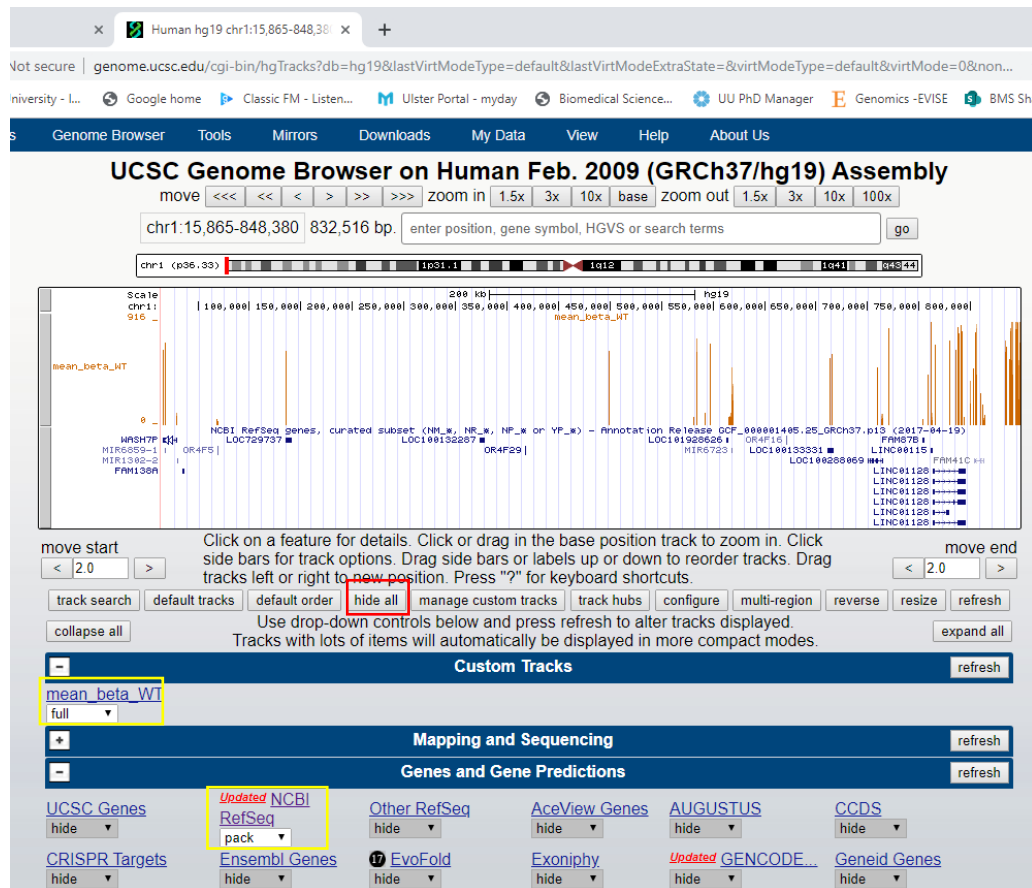
To access the Tracks generated as output, from the results History generated from your run, click on the link saying “CandiMeth\_Tracks (a list with x items)”, which will open a new window at RHS. For the example data here, this will show the four items in the list below (see screenshot at left). These are the two tracks showing absolute methylation (beta value) in 1)the WT cells used as a control [mean\\_beta\\_WT\\_of\\_D8\\_Track](#) and 2)the D8 cells which have decreased DNMT1 levels [mean\\_beta\\_D8\\_Track](#). There are also two tracks showing comparisons between the WT and D8: these are 3)the track showing difference in methylation (delta beta or  $\Delta\beta$ ) between WT and D8 called [delta\\_beta\\_D8vsWT\\_Track](#) and 4)a track showing only those probes where the difference in methylation is significant at a false discovery rate of 0.05, called [FDR\\_D8\\_of\\_WT\\_Track](#).



Clicking on the name of the Track at left e.g. [mean\\_beta\\_WT\\_of\\_D8\\_Track](#) will show the preview window (see RHS above). This should say under the title “~480,000 regions” for the 450K array and “format:bed, database:hg19” indicating that a type of track called a BED file has been generated, using the hg19 edition of the human genome map. The first five lines of the track data will also be shown, but this is a long table with 480,000 rows! To visualise the data, we instead:-

1. Click on one of the “display” options in the preview window: CandiMeth is optimised for use with the UCSC browser, so click on “display at UCSC [main](#)” by following the hyperlink in blue.
2. There will be a small delay, then a new tab will open in the browser, taking you to the familiar UCSC genome browser page, with the data from your first track displayed at the top (see next page for screenshot).
3. The default tracks on UCSC include roughly one from every major group (blue header), and at writing were [UCSC\\_Genes](#), [NCBI\\_RefSeq](#), [Publications](#), [GTEx gene](#), [ENCODE regulation](#), [Conservation](#) and [dbSNP\\_153](#). Your track will appear under the [Custom Tracks](#) header at top as [mean\\_beta\\_WT](#). As this makes the window quite complex and busy, you

may wish to simplify the view by clicking on the “hide all” button (boxed in red on screenshot below), then add back just the locations of genes using [NCBI\\_RefSeq](#) >pack and [mean\\_beta\\_WT](#)>full (both boxed in yellow below). This should give you a simpler view which is easier to work with, similar to that below when using the test data:

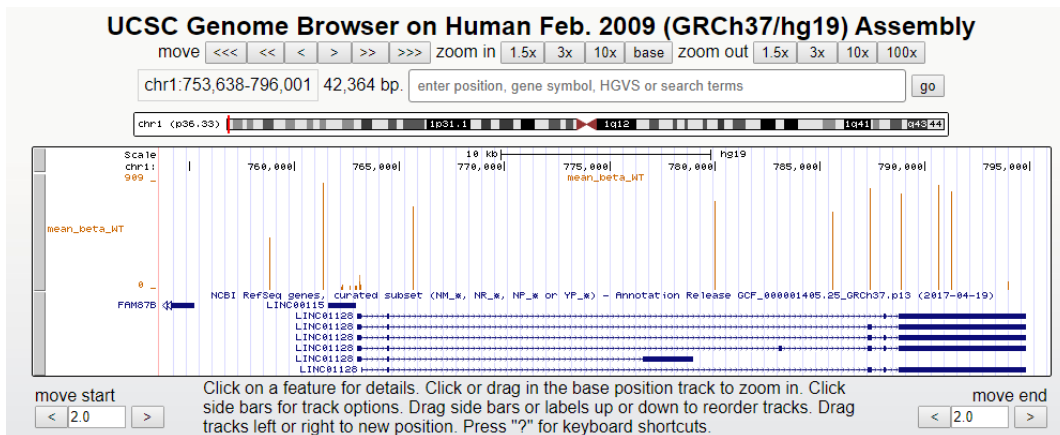


4. You now have a display showing the start of the genome map (chromosome 1p) with genes displayed in blue on the [NCBI\\_RefSeq](#) track at bottom, with methylation levels in the WT cells shown as peaks in the [mean\\_beta\\_WT](#) track along the top in brown.

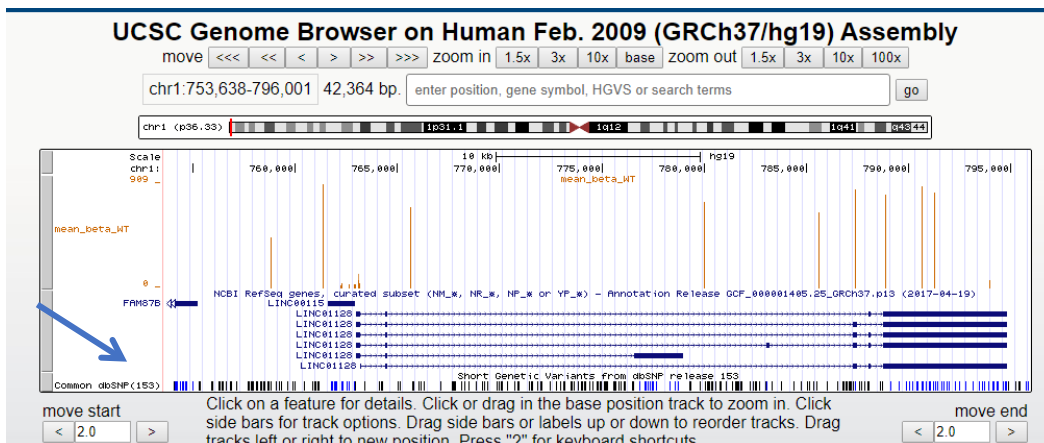
The height of each peak in the [mean\\_beta\\_WT](#) track corresponds to the methylation level at that position, with the minimum and maximum values seen in this window displayed at left as beta values x 1000 (0 and 916, equivalent to 0% and 91.6% methylated respectively).

5. This is a fully zoomable map as usual for UCSC: to illustrate, if you draw a box around the gene just visible at right above, *LINC01128*, this will magnify the view of that gene, showing the locations and extent of methylation at each probe across the gene (below)

[Note: if a window opens for “Drag-and-select” simply tick “don’t show this again” and “zoom in”]



- To overlay any other data on this map, simply choose from the UCSC pull-down menus: e.g. to show the locations of common SNPS, choosing [dbSNP\\_153](#) dense under the *Variation* header further down the UCSC main controls will overlay a track with this information underneath the other tracks (arrow below)

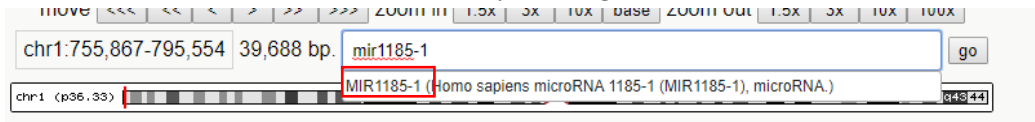


- The screen currently only shows the data from one of your four tracks generated using the example data: to bring up the next track, follow steps 1-2 above for [mean\\_beta\\_D8\\_Track](#): this will open a new tab in your browser showing the new data AND the track you already generated
- Do the same (steps 1-2) for the remaining two tracks [delta\\_beta\\_D8vsWT\\_Track](#) and [FDR\\_D8\\_of\\_WT\\_Track](#); the last window you open will now show ALL FOUR tracks (see screenshot below) and other tabs can be closed

[Note: leave the Galaxy tab open in the background, to allow access to Results etc]

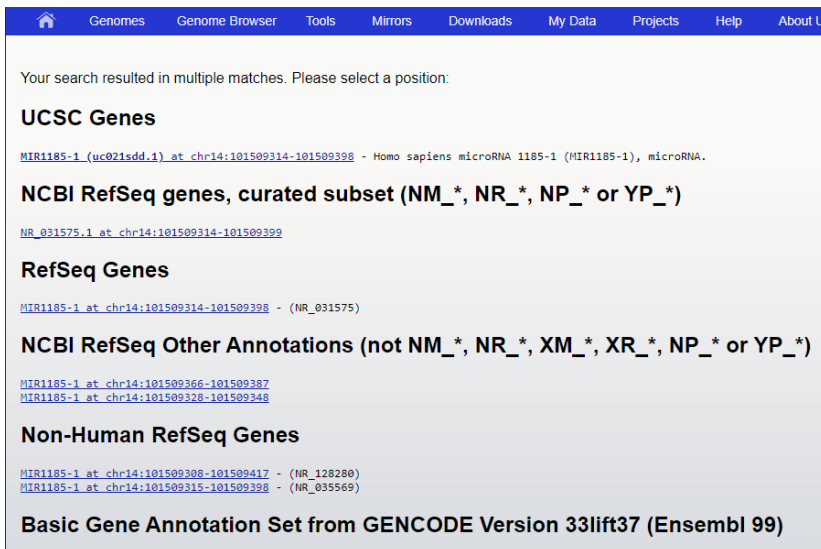


12. Note that individual tracks can be toggled on or off with the buttons under Custom tracks, then hitting “refresh”: this can be particularly useful to just look at differences in methylation (delta beta)
13. While the Tracks open by default at the start of the genome map (chr1p), you can look at any gene in the human genome by typing its name into the search box at top of the screen (yellow box at top of last screenshot)
14. For this example, type “MIR1185-1” into the box: as you type, the name should appear under the box- click on the name to take you straight there

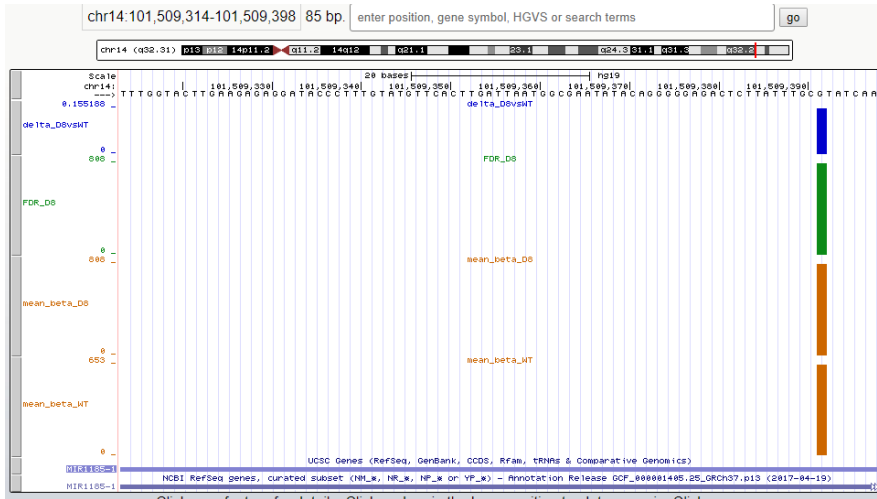


[Note: this is a short-cut to the location of this MIR gene as decided by UCSC]

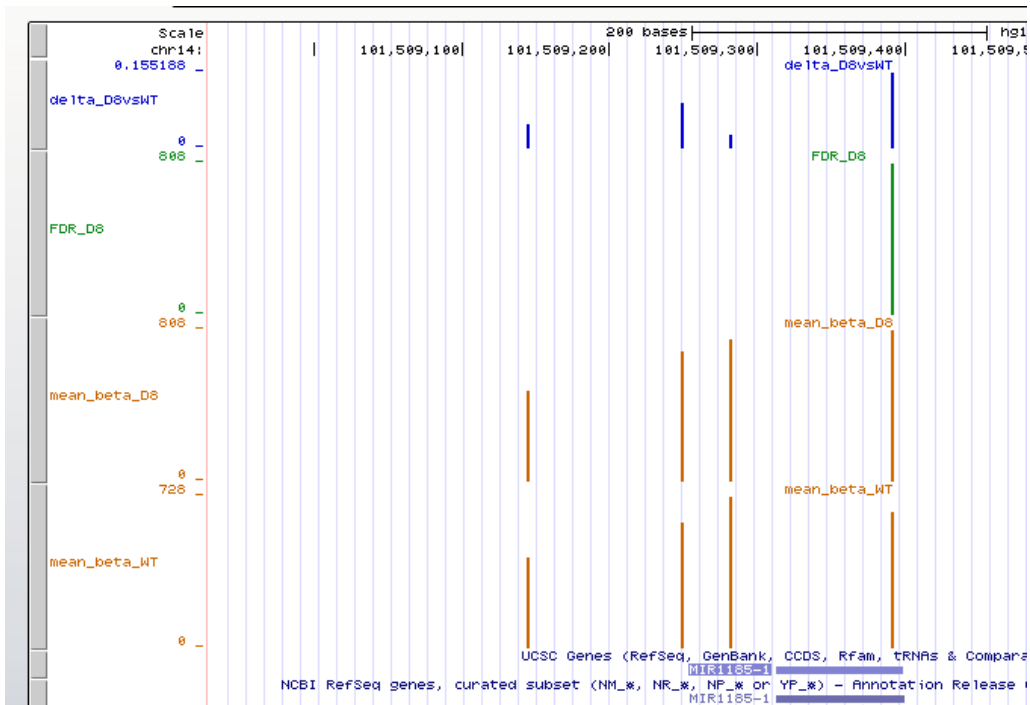
15. Alternatively, if you type MIR1185-1 and hit “Go”, the next screen will give you a set of alternatives, based on who has mapped the gene



- you can click on any of the options to bring you to the map location indicated: hitting UCSC will bring you to the same location as in step 15 above. This screen can be useful when there is some dispute over map location of genomic features.
16. Following either steps 15 or 16 above will bring you to a zoomed-in map of MIR1185-1: this shows ONLY the body of this small gene, so information from only one array probe is visible at right at large thick bars



17. To get a better view of the promoter and surroundings, use the 10X zoom-out buttons



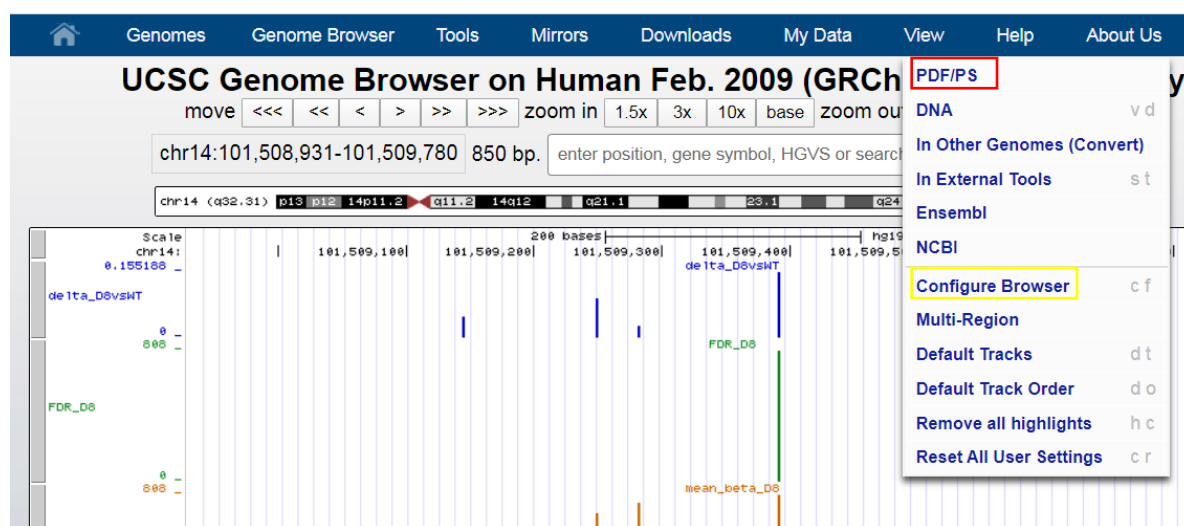
18. This shows a view of the four probes associated with this MIR: the mean and median methylation of these four probes were captured in the Results tables under 3.1.3 above

19. From the Tracks here it can be seen that the methylation is much lower at the probe furthest away from the gene (to left above), while the only probe showing significant differences in methylation between WT cells and those with low levels of DNMT1 (D8) is the probe at right (green above)

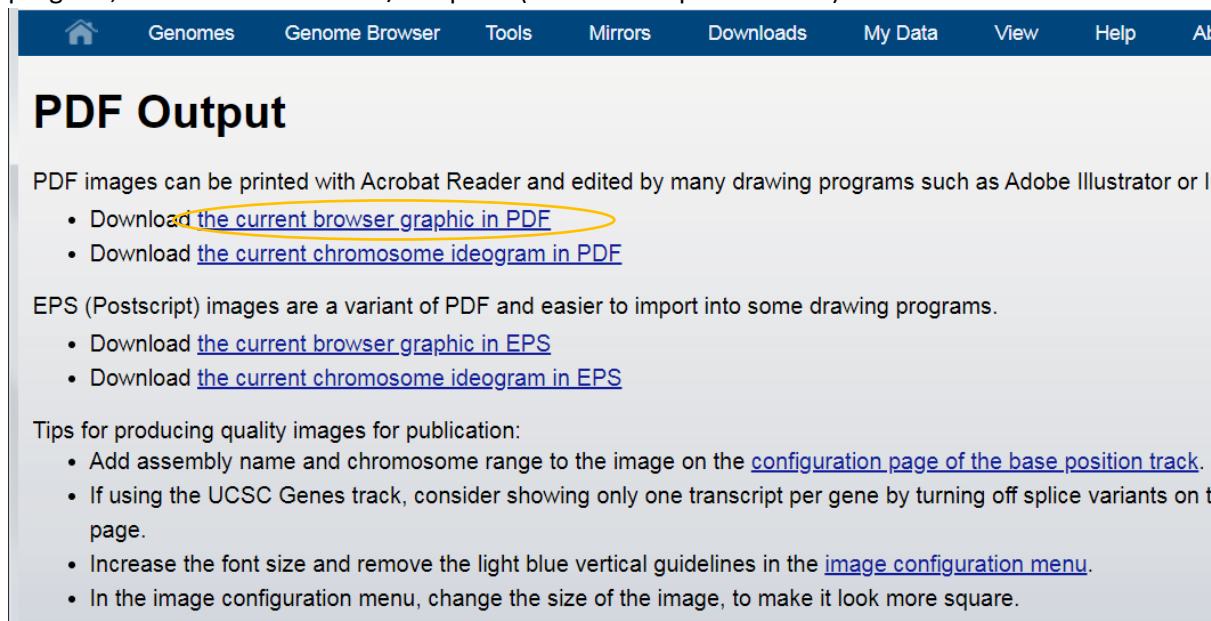
### 3.1.7 Exporting browser views as graphics files

It is often the case that the user wants to show a particular UCSC genome browser view of the data. You may also wish to modify the view slightly by, for example, removing the grid lines (a common request from journals). These facilities are provided by UCSC and can be accessed as follows:-

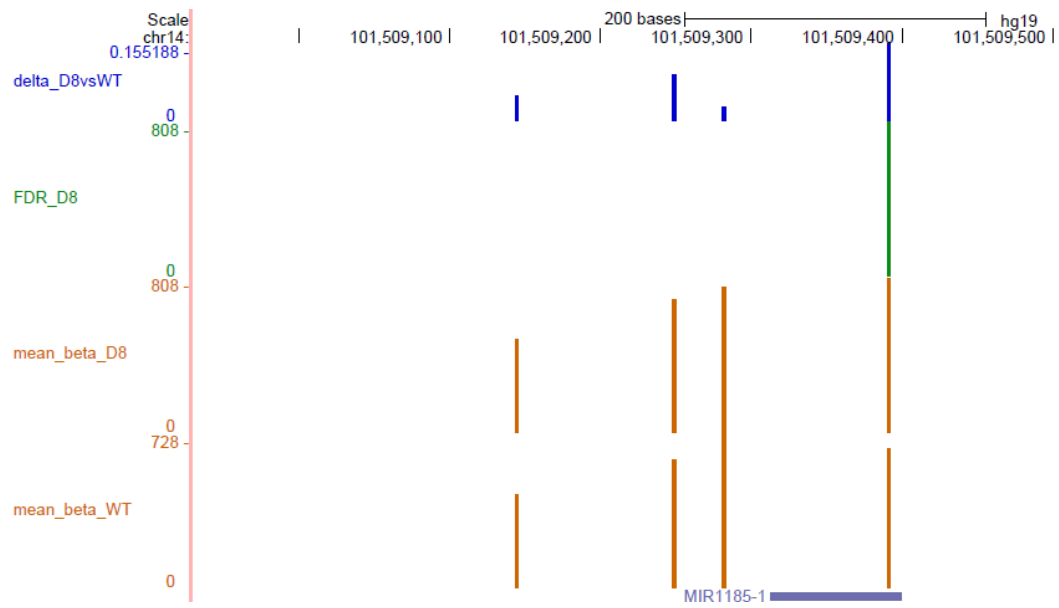
1. On the blue top ribbon in the UCSC browser window, click on View button, which will bring up a number of options in a pull-down menu



2. To first tidy the image, you can click on >Configure Browser (yellow box above)
3. Uncheck the boxes for “Show light blue vertical guidelines” and “Display description above each track” and click the gray “submit” box at top left
4. Your browser will return to the image you were viewing, which should now have no gridlines or labels in the middle of the screen
5. To export this view in a format you can include in documents, or further adjust in another program, click on the View>PDF/PS option (red box in top screenshot)



6. A new screen will appear (below) with a number of options: to save as a PDF file click on the top option (orange box above) [Note the tips for publication-quality images here]



7. This should open a new screen showing the PDF version of the genome browser view (above), which can be downloaded and inserted in documents
8. As well as PDFs, postscript (PS) file format is also supported: most graphics software programs can import files in one or the other format for further adjustment if needed e.g. Adobe Illustrator or Photoshop




### 3.1.8 Quantifying methylation in different parts of the gene

In the example above, we looked at methylation across the whole gene locus (promoter and gene body) for the microRNA genes in our list. CandiMeth is however designed to look at different parts of the gene, as these can often behave differently.

#### 3.1.8.1 Promoter methylation only

To ONLY look at the methylation in the promoter regions of the microRNA genes, the same settings as in 3.1.1 above can be used EXCEPT that under Input 4 choose “hg19\_prom”

[Note: if you have already run CandiMeth, remember to switch back to the original history containing the test data “[Date/run identifier] CandiMeth My Test Data” History by using the “switch history” tool  in the top RHS of the screen: choose the history needed by clicking the grey “Switch to” button at top left of the History, then >Analyse data on the top bar of Galaxy. You should return to the standard Galaxy view but with the desired History in the RHS window]

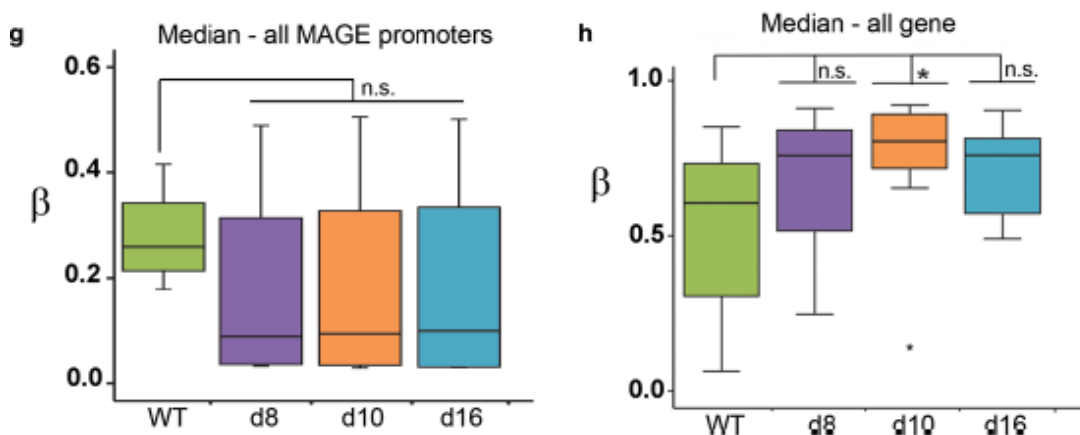
The output tables in the Results folder will now only average the methylation across the probes found in the MIR promoters (defined as -500bp to +1bp from the gene start).

#### 3.1.8.2 Gene body methylation only

To ONLY look at the methylation in the gene bodies of the microRNA genes, the same settings as in 3.1.1 above can be used EXCEPT that under Input 4 choose “hg19\_GB”

The output tables in the Results folder will now only average the methylation across the probes found in the MIR gene bodies (defined as +1bp from the gene start, through all of the exons and introns, to the transcriptional end site (TES)).

That these two parts of the genes can vary significantly, or even show opposite effects, is well-documented in the literature and can be illustrated by the graphics below:

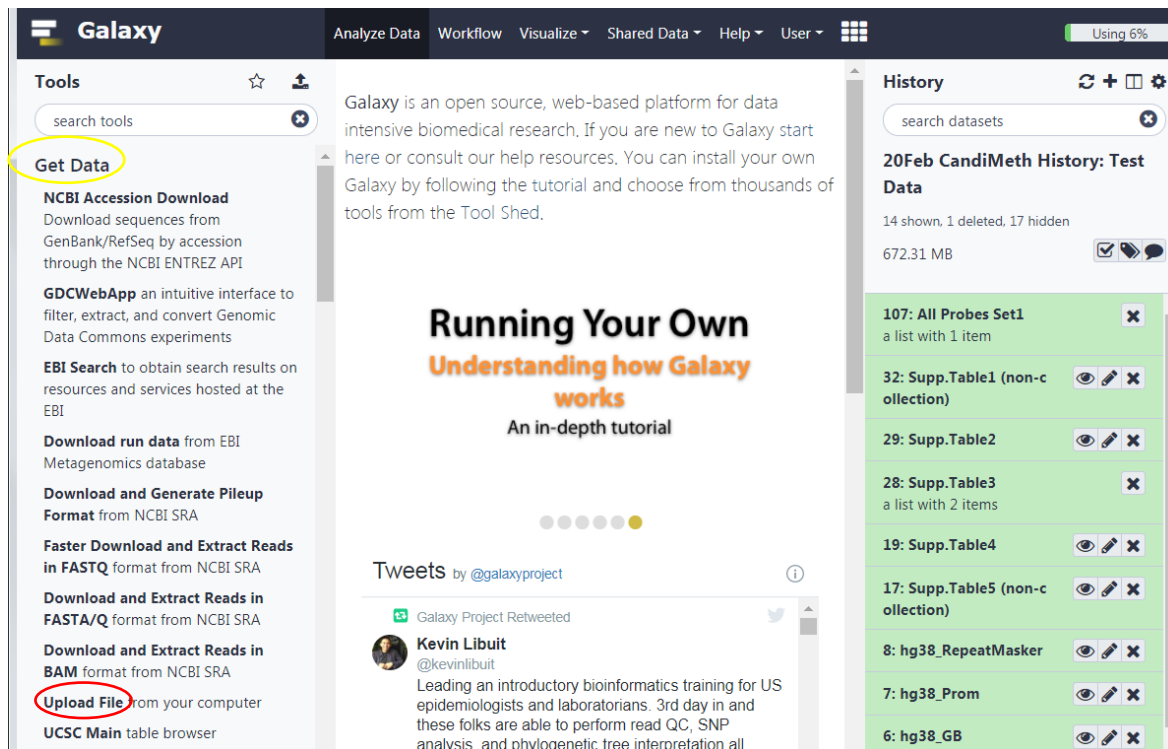


This shows that median methylation ( $\beta$  value) of the promoters of MAGE genes decreased in D8 and other DNMT1- depleted cell lines relative to WT (left), while methylation at the gene bodies went up (right). [graphic generated in SPSS after CandiMeth analysis, see O’Neill et al E&C 2018]

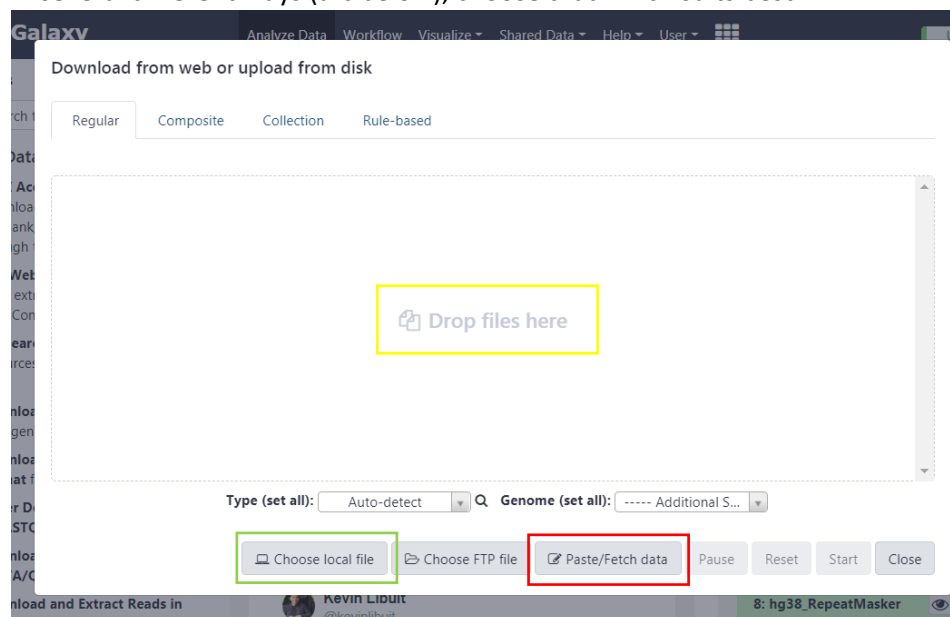
### 3.2 Looking at a new set of genes in the current methylation array dataset

In the example above, the microRNA (MIR) gene list (Supp.Table 2) was used to query the methylation array data from the comparison of DNMT1-depleted and WT cells (Supp.Table1, converted into a collection All Probes Set 1). Once array data has been uploaded and converted however it is perfectly possible to look at any other gene or genes you are interested in.

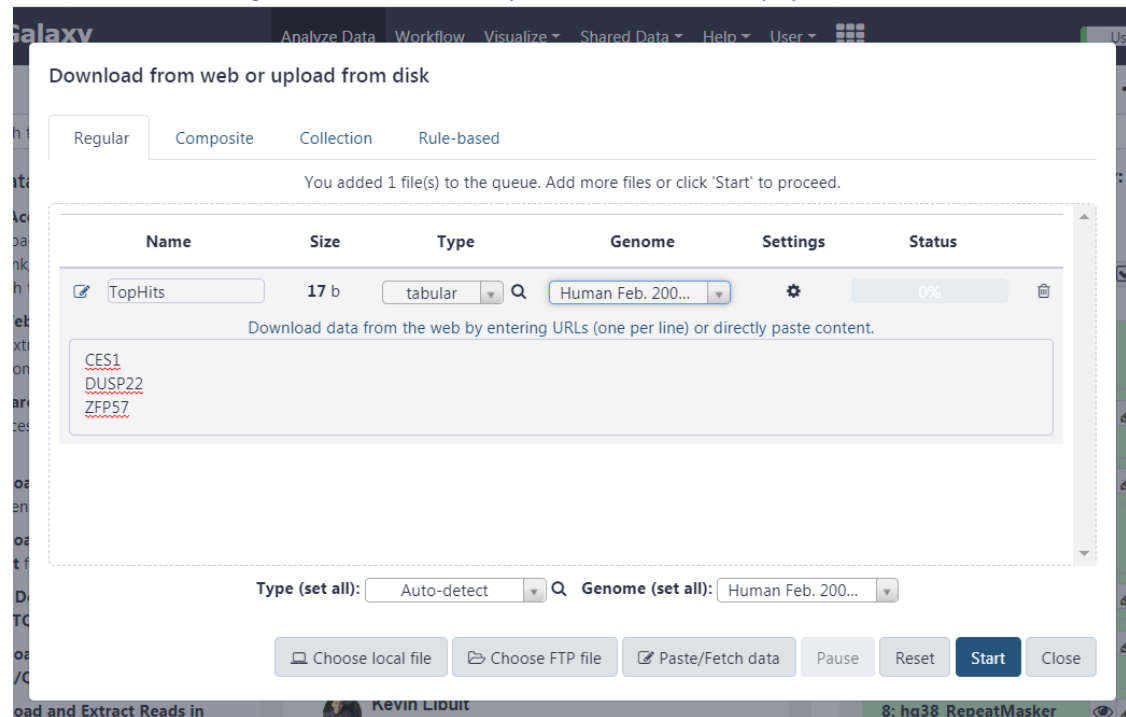
- 1) Navigate back to the History containing the Test data, which includes All Probes Set1 (the data you will query) by using the navigation symbol as before (see note in 3.1.8.1 above)
- 2) Now on the left hand side (LHS) Tools window choose >Get data (yellow oval on screenshot) then >Upload file from your computer (red oval)



- 3) This will open a new window where you can put in the names of the genes you are interested in in several different ways (a-c below), choose that which suits best:



- a) Click on **>Paste/Fetch data** (red square in screenshot above), then just type or paste in the names of the genes you wish to investigate onto separate lines as in the example below (here the three genes from Case Study 2, main CandiMeth paper)





-Along the top of the window, give the new list a **Name** e.g. **"TopHits"**, choose **>tabular** under **Type**, and under **Genome** choose **>Human Feb.2009 (GRCh37/h19) (hg19)** [this will appear as an option if you start to type *hg19*]

-Click **"Start"**: the file should upload to Galaxy and appear as a separate dataset on the RHS with the name you gave it, in this case **"TopHits"**: you can close the Upload window and go to step 4

- b) If you have a longer or more complex list, this can be written in a word-processing program such as Word and saved as a *text only* or *Plain text* file (\*.txt), before uploading directly using the uploaded directly from a .txt file format (e.g. [Supp Table 2](#)) by following steps 1-3 above i.e. **>Upload file >Choose local file**, the format should be Tabular
- c) You can also simply drag and drop a text-only file created as in (b) into the window shown at the start of step 3 above
- 4) This new list can then be used to query the array data by Running CandiMeth and choosing **"TopHits"** as Input 3 (See 3.1.1 step 3) instead of the MIR gene list in Supp.Table2
- 5) The Results folder will now contain Tables showing methylation levels for the new list of genes

### 3.3 Looking at repetitive DNA elements such as LINES and ERV

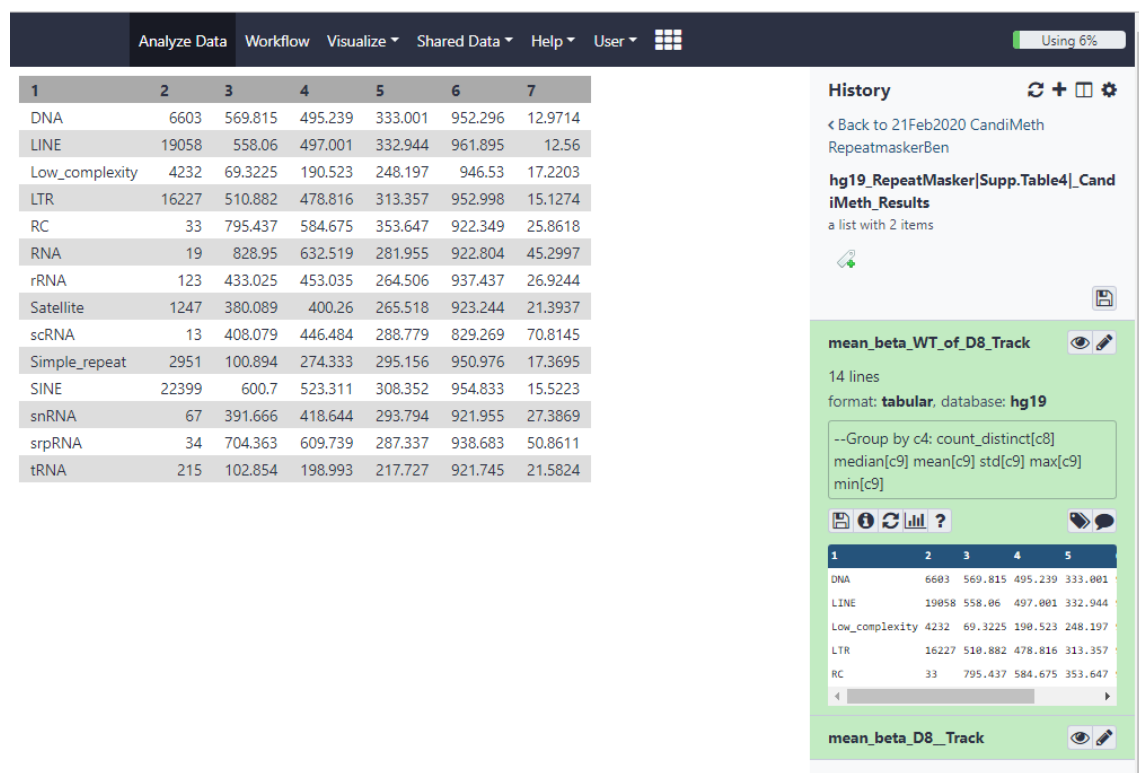
Many epigenome-wide studies are interested in assessing methylation at repetitive DNA elements instead of endogenous genes, often using a wet-lab analysis technique such as pyrosequencing to assay LINE-1 methylation for example. A substantial number of probes on the 450K and EPIC arrays fall within repetitive DNA elements however, allowing analysis of methylation across these elements. As an example, we can look at methylation across repetitive elements in the WT and d8 dataset. To do so we:

1. Switch to the original history containing the test data “[Date/run identifier] CandiMeth My Test Data” using the “switch history” tool  in the top RHS of the screen [see note in 3.1.8.1 above]
2. Under >Workflow on the top black Galaxy header, click on the CandiMeth workflow and on the pull-down menu at RHS marked  choose > Run
3. Choose to Send the results to a new history e.g. “[Date/run identifier] CandiMeth Repeats”
4. Under 1: R Package Used: (1.1) enter ‘RnBeads’
5. For 2: Input Differential Methylation Table (1.2) choose “All Probes Set1”
6. At 3: Input Gene Features of Interest (1.3) choose “Supp.Table4”

[Note: this contains the names of the different types of repetitive DNA as identified by the *RepeatMasker* program (see below)]

7. For 4: Input Genome Release Information (1.4) choose “hg19\_RepeatMasker”

You can now click the blue ‘Run workflow’ button at top right. Results will appear in the new History and should resemble the table below for the WT cells :



The screenshot shows the Galaxy web interface. On the left, a table displays the results of the workflow for various repetitive DNA elements. On the right, the History panel shows the workflow steps and a preview of the output table.

1	2	3	4	5	6	7
DNA	6603	569.815	495.239	333.001	952.296	12.9714
LINE	19058	558.06	497.001	332.944	961.895	12.56
Low_complexity	4232	69.3225	190.523	248.197	946.53	17.2203
LTR	16227	510.882	478.816	313.357	952.998	15.1274
RC	33	795.437	584.675	353.647	922.349	25.8618
RNA	19	828.95	632.519	281.955	922.804	45.2997
rRNA	123	433.025	453.035	264.506	937.437	26.9244
Satellite	1247	380.089	400.26	265.518	923.244	21.3937
scRNA	13	408.079	446.484	288.779	829.269	70.8145
Simple_repeat	2951	100.894	274.333	295.156	950.976	17.3695
SINE	22399	600.7	523.311	308.352	954.833	15.5223
snRNA	67	391.666	418.644	293.794	921.955	27.3869
srpRNA	34	704.363	609.739	287.337	938.683	50.8611
tRNA	215	102.854	198.993	217.727	921.745	21.5824

The History panel shows the workflow steps and a preview of the output table:

```

--Group by c4: count_distinct[c8]
median[c9] mean[c9] std[c9] max[c9]
min[c9]

```

1	2	3	4	5
DNA	6603	569.815	495.239	333.001
LINE	19058	558.06	497.001	332.944
Low_complexity	4232	69.3225	190.523	248.197
LTR	16227	510.882	478.816	313.357
RC	33	795.437	584.675	353.647

In the output Table, the different types of repetitive element as identified by the *RepeatMasker* algorithm [[www.repeatmasker.org](http://www.repeatmasker.org)] are indicated, together with the number of probes etc in the same output format as before ie

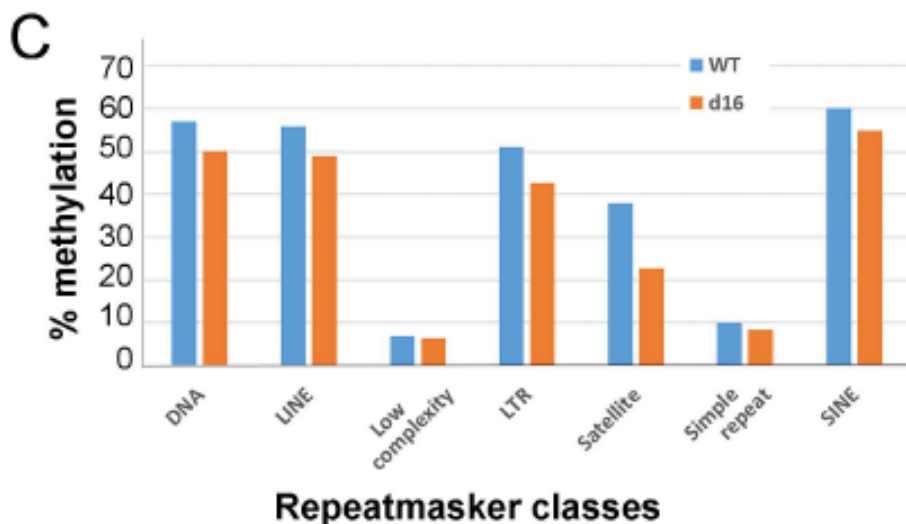
- 1) Name of gene, 2) Number of array probes, 3) Median methylation, 4) Standard deviation, 5) Mean methylation, 6) Maximum value and 7) Minimum value

The Names of the elements can be found on the RepeatMasker track in UCSC. As can be seen from the example Table above, a number of repeats are covered by less than 1000 probes, which may be less reliable. Names of the classes with >1000 probes on the array includes:-

- DNA DNA repeat elements
- LINE Long Interspersed Nuclear Elements such as LINE1
- Low\_complexity Low complexity repeats which do not fall into other categories
- LTR LTR-containing elements such as endogenous retroviruses (ERV)
- Satellite Satellite repeats, found near the centromeres
- Simple\_repeat Largely microsatellites, which are interspersed
- SINE Short Nuclear Interspersed Elements such as Alu elements

Methylation varies greatly across these elements as can be seen from the minimum and maximum values, but comparisons of median methylation can nevertheless be valuable.

In Case Study 3 in the main CandiMeth paper for example, methylation in DNMT1-depleted cells (d16 in this case) can be seen to affect satellite repeats, but have little effect on microsatellites (Simple\_repeat), many of which would lack any CG.



### 3.4 Using ChAMP-generated methylation data

While the above examples all use the data processed by the RnBeads package in R, CandiMeth can also work with data which has instead been processed using the ChAMP package. An example dataset has been provided in the Test History, Suppl. Table 5. This has been uploaded as an Excel output (.csv), so it needs first to be converted into a dataset collection (See also Section 2, Step 3).

1. Convert the ChAMP csv file to a dataset collection
  - a. Click on “Operations on multiple datasets” at top RHS
  - b. Check the box beside Suppl. Table 5
  - c. Under “For all selected” choose “Build Dataset List”
  - d. Give the collection a new name e.g. “All probes ChAMP1”
  - e. Once the new dataset collection appears, click on the “Operations” box again to return to normal view
  
2. Choosing inputs and starting the run
  - a. Click on Workflows on the top ribbon and choose CandiMeth and click on the arrow
  - b. Choose the option to send the results to a new History e.g. “[date] ChAMP test1”
  - c. Under 1: R Package Used: (1.1) enter ‘ChAMP’
  - d. For 2: Input Differential Methylation Table (1.2) choose “All Probes ChAMP1”  
  
[Note: This was the example name used in step 1(d) above, alter as required]
  - e. At 3: Input Gene Features of Interest (1.3) choose “Suppl. Table2”
  - f. For 4: Input Genome Release Information (1.4) choose “hg19\_all”
  
3. Once the workflow has finished, similar tables of Results and Tracks should appear as before (see sections 3.1.1-3.1.8) for these sample microRNA data, and all the same types of operations (looking at promoters vs gene bodies, repeat analysis, new gene queries etc) can be carried out

[Note: If your outputs from ChAMP are not normally being produced as the .csv files needed for Step 1 above, please show whoever is running the ChAMP pipeline Appendix 3 below, which contains the few lines of coding needed to do this]

## 4. Uploading and working with your own differential methylation data

For this you need at least one file containing information on methylation differences between two samples produced from either RnBeads or ChAMP.

### 4.1 Locating data files in RnBeads

1. If you received your data back as a completed [Report](#) folder with an [index.html](#) page then click on that, which should bring up a list of all reports, including differential methylation:









## RnBeads Analysis

### Table of Contents

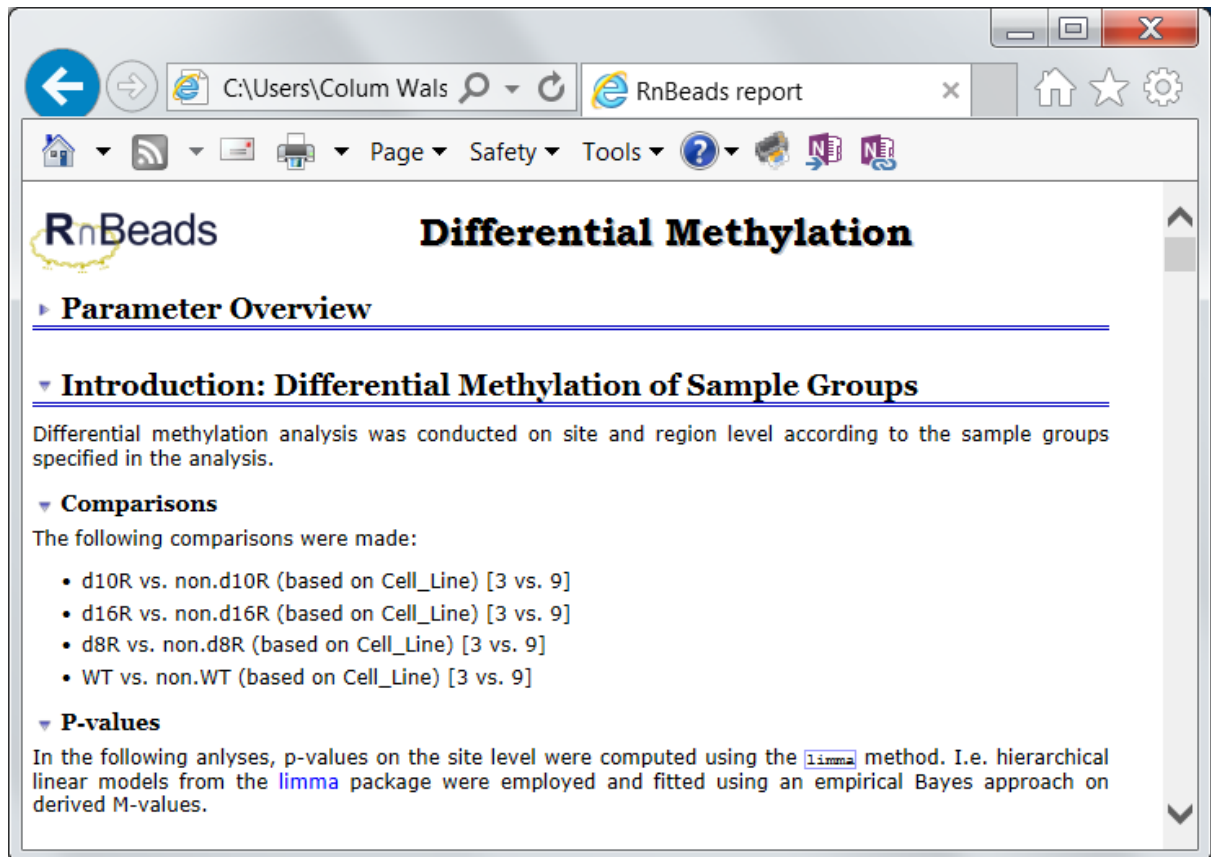
The following listing contains links to all reports generated or scheduled by RnBeads. A short description of each report is also provided.

The log file [analysis.log](#) presents a detailed account of all performed activities.

	<b>Data Import</b> This report describes the loading of the data into RnBeads.
	<b>Quality Control</b> This report performs assay quality validation.
	<b>Preprocessing</b> This report presents the filtering and normalization steps applied to the dataset.
	<b>Tracks and Tables</b> This report provides contains information on exported data, generated genome browser tracks and sample summary tables.
	<b>Exploratory Analysis</b> This reports describes sample subgroups, methylation profiles and associations with sample annotations.
	<b>Differential Methylation</b> This report identifies differentially methylated sites and regions between sample groups.

[Note: if Differential Methylation is absent then this type of analysis has not yet been done]

2. Click on the Differential Methylation report and look to see what comparisons have been done, for example:



3. Each comparison listed will generate a differential methylation table labelled cmp1, cmp2 etc
4. Scroll down the page to find links to the actual differential methylation files, under the heading Differential Methylation Tables (boxed in green below):

- min.covg.g1,min.covg.g2: minimum coverage of groups 1 and 2 respectively
- max.covg.g1,max.covg.g2: maximum coverage of groups 1 and 2 respectively
- covg.thresh.nsamples.g1,covg.thresh.nsamples.g2: number of samples in group 1 and 2 respectively exceeding the coverage threshold (5) for this site.

The tables for the individual comparisons can be found here:

- [d10R vs. non.d10R \(based on Cell\\_Line\)](#)
- [d16R vs. non.d16R \(based on Cell\\_Line\)](#)
- [d8R vs. non.d8R \(based on Cell\\_Line\)](#)
- [WT vs. non.WT \(based on Cell\\_Line\)](#)

### ▼ Region Level

Differential methylation on the region level was computed based on a variety of metrics. Of particular

5. These hyperlinks take you the file itself- the address can be seen by holding the mouse over the link. These files are usually located in the “differential\_methylation\_data” folder of your Results folder and is named something similar to “diffMethTable\_site\_cmp1.csv” for comparison 1 etc

[NB: the file must have data on all sites i.e. contain “\_site\_” to work for all types of analysis]

6. You can Open or Save As to copy the file to a new location for uploading, or use the original file in the differential\_methylation\_data folder

[Tip: these are usually large files and may take some time to download/upload]



## 4.2 Locating data files in ChAMP

If your ChAMP related output has not been produced as a .csv file outside of R, please see the below instructions on how to write the differential methylation table to a .csv file. This code will need to be implemented in R while using the ChAMP package, so it may be appropriate to pass these on to whoever is providing bioinformatics support for the project.

-For just one comparison:

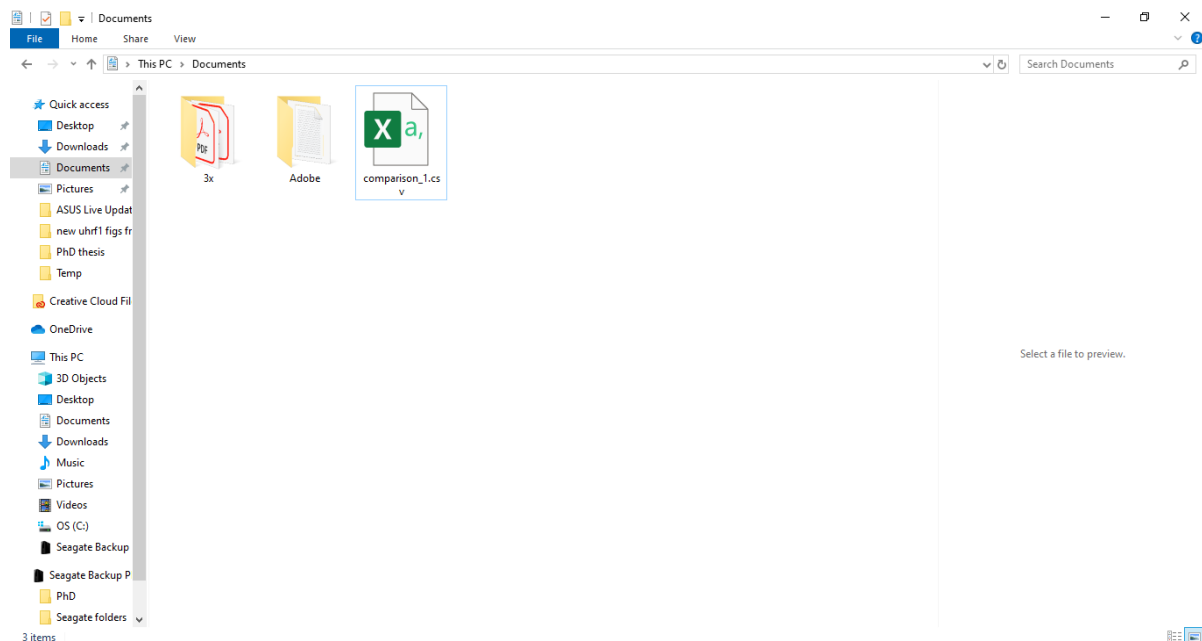
```
write.csv(myDMP[[x]], file = "comparison1.csv", quote = FALSE)
```

(where x is the element number of the file comparison you wish to write to the .csv file and myDMP is the resulting object of running champ.DMP() as within the ChAMP [vignette](https://www.bioconductor.org/packages/3.7/bioc/vignettes/ChAMP/inst/doc/ChAMP.html#section-differential-methylation-probes) (<https://www.bioconductor.org/packages/3.7/bioc/vignettes/ChAMP/inst/doc/ChAMP.html#section-differential-methylation-probes>))

-For the output of multiple comparisons:

```
compnames <- -names(myDMP)
for(i in 1:length(compnames)){write.csv(myDMP[[i]], file
= paste(compnames[i], ".csv", sep=""), quote=FALSE)}
```

This will create all probes differential methylation tables within your documents folder as below:



When opened, this file will look similar to the image below:

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	
1		logFC	AveExpr	t	P.Value	adj.P.Val	B	WT_AVG	EXP_AVG	deltaBeta	CHR	MAPINFO	Strand	Type	gene	feature	cgi	feat.cgi	UCSC_CpCDS	DHS	Enhancer	
2	cg0559916	0.129839	0.360831	6.217882	1.94E-06	0.083747	4.90793	0.295911	0.425751	0.129839	5	1.25E+08	R	II		IGR	opensea	IGR-opensea	NA	NA	TRU	
3	cg1667253	0.117385	0.282915	6.042221	2.98E-06	0.083747	4.479385	0.224223	0.341608	0.117385	1	6510652	F	II	ESPN	Body	shore	Body-shor	chr1:65111	NA	NA	TRU
4	cg0201998	0.113914	0.357914	5.890808	4.34E-06	0.083747	4.10767	0.300957	0.414871	0.113914	3	50311211	R	II	SEMA3B	Body	island	Body-islar	chr3:50311	NA	NA	TRU
5	cg0793368	0.10605	0.488299	5.822059	5.14E-06	0.083747	3.93823	0.435274	0.541324	0.10605	19	3025347	R	II	TLE2	Body	shelf	Body-shel	chr19:3025	NA	NA	TRU
6	cg1249788	0.059135	0.657643	5.707706	6.84E-06	0.083747	3.655548	0.628075	0.68721	0.059135	10	1.33E+08	F	II	TCERGIL	Body	shore	Body-shoi	chr10:133	NA	NA	TRU
7	cg0199196	0.047756	0.734435	5.634966	8.20E-06	0.083747	3.475213	0.710557	0.758314	0.047756	2	1.14E+08	R	II	IL1RN	S'UTR	opensea	S'UTR-opensea	NA	NA	TRU	
8	cg0223684	0.113669	0.389223	5.439063	1.34E-05	0.083747	2.987765	0.332389	0.446057	0.113669	16	7364246	R	II	A2BP1	Body	opensea	Body-opensea	NA	NA	TRU	
9	cg1197281	0.062665	0.556712	5.355805	1.65E-05	0.083747	2.779911	0.52538	0.588045	0.062665	15	76484379	F	I	C15orf27	Body	island	Body-islar	chr15:764	NA	NA	TRU
10	cg2261385	0.138493	0.22454	5.345275	1.70E-05	0.083747	2.753597	0.153207	0.291701	0.138493	3	1.95E+08	F	I		IGR	opensea	IGR-opensea	TRUE	NA	TRU	
11	cg0005577	-0.05119	0.830404	-5.31804	1.82E-05	0.083747	2.685502	0.855997	0.804811	-0.05119	5	1.45E+08	F	II		IGR	opensea	IGR-opensea	NA	NA	TRU	
12	cg2097973	0.114813	0.367945	5.309181	1.86E-05	0.083747	2.663361	0.310538	0.425352	0.114813	1	2.45E+08	R	II		IGR	opensea	IGR-opensea	NA	NA	TRU	
13	cg0072904	0.044176	0.09889	5.281203	2.00E-05	0.083747	2.593372	0.076802	0.120978	0.044176	2	1.72E+08	F	II	GAD1	Body	island	Body-islar	chr2:1716	NA	TRU	
14	cg2014276	0.110788	0.308423	5.247293	2.18E-05	0.083747	2.508496	0.253029	0.363817	0.110788	19	55416928	F	II	NCR1	TSS1500	opensea	TSS1500-opensea	NA	TRU		
15	cg2200664	0.080171	0.290543	5.232348	2.26E-05	0.083747	2.471076	0.250458	0.330629	0.080171	4	1.12E+08	F	I		IGR	shore	IGR-shore	chr4:1115	NA	TRU	
16	cg2026230	0.107102	0.292745	5.180412	2.58E-05	0.083747	2.340965	0.239194	0.346296	0.107102	7	1.56E+08	F	II	SHH	Body	shore	Body-shoi	chr7:1556	NA	TRU	
17	cg2450619	0.129859	0.291302	5.174767	2.62E-05	0.083747	2.326815	0.226372	0.356232	0.129859	7	1.58E+08	F	II	PTPRN2	Body	shelf	Body-shel	chr7:1581	NA	TRU	
18	cg0893630	0.133711	0.29709	5.172257	2.63E-05	0.083747	2.320526	0.230234	0.363946	0.133711	19	6481945	R	II	DENND1C	TSS200	opensea	TSS200-opensea	NA	TRU		
19	cg0391713	0.117834	0.512032	5.168749	2.66E-05	0.083747	2.311731	0.453116	0.570949	0.117834	17	70120182	F	II	SOX9	Body	island	Body-islar	chr17:701	NA	TRU	
20	cg0017811	0.095648	0.388105	5.163693	2.69E-05	0.083747	2.299058	0.340281	0.435929	0.095648	8	1.05E+08	F	II	RIMS2	Body	shore	Body-shoi	chr8:1052	NA	TRU	
21	cg0763885	0.095333	0.353629	5.163307	2.69E-05	0.083747	2.298091	0.305962	0.401295	0.095333	12	99137308	R	II	ANKS1B	Body	shelf	Body-shel	chr12:991	NA	TRU	
22	cg2193254	0.144786	0.330371	5.148693	2.79E-05	0.083747	2.261453	0.257978	0.402764	0.144786	7	4850345	F	II	RADIL	Body	shore	Body-shoi	chr7:4848	NA	TRU	
23	cg0261313	0.070646	0.991532	5.146462	2.81E-05	0.083747	2.255872	0.261304	0.81125	0.070646	10	1.02E+08	F	II		IGR	shore	IGR-shoi	chr10:103	NA	TRU	

This file is in comma separated variable (.csv) format and can now be uploaded to Galaxy to be used in CandiMeth, as detailed in the next section and in section 3.2 above.

### 4.3 Uploading your data files to Galaxy

This is essentially as in section 3.2 above, where you uploaded a new list of gene names. In brief,

1. Navigate back to the History containing the Test data
2. Using the Tools window (LHS) choose >Get Data > Upload File > Choose Local File and locate the methylation data you wish to upload (e.g. diffMethTable\_site\_cmp1.csv). Alternatively, you can drag and drop it in.
3. Set "Type (set all)" to whatever kind of file you are uploading, for example, RnBeads based outputs are usually comma separated variable files (.csv)

[Tip: The default "Auto-detect" setting works well for most file formats]

4. Set "Genome (set all)" to either Human Feb. 2009 (GRCh37/hg19) or Human Dec. 2013 (GRCh38/hg38) depending on what genome your array was mapped to.



[Tip: You can type in Human here to bring up all the human genomes and save time]

The window should look something like this:-

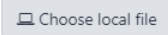
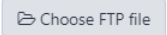





## Download from web or upload from disk

Regular Composite Collection Rule-based

You added 1 file(s) to the queue. Add more files or click 'Start' to proceed.

Name	Size	Type	Genome	Settings	Status
 Dataset2 BEN diffMet	147.9 MB	csv	Human Feb. 200...		0%

Type (set all): csv Genome (set all): Human Feb. 200...

5. Click "Start": the file should upload to Galaxy and the upload window will go green and a tick appear under "Status"- this can now be closed
6. In the main window, the new data will appear as a dataset on the RHS: this goes from grey to orange then finally green if all is well
7. NB: the input Differential Methylation Table still has to be converted from a table into the form of a Dataset Collection as before (see Section 2 above)
8. Give the collection a more specific name e.g. "Dataset2 [your name]" and click "Create"
9. A new entry should appear in the RHS with the new name and "a list with 1 (or more) items"-this is the Dataset Collection and is now ready to be processed by CandiMeth

## Appendix 1. Primer for those unfamiliar with the UCSC genome browser

As a result of major efforts from many scientists around the world, the complete genetic code present in our DNA (our genome) has been characterised and mapped. As we have ~4 billion individual “bits” or DNA bases, split into 23 chromosomes, one of the main problems that arose was how to find the information for any specific gene, and relate it to its surrounding genes and other information we might have such as whether the gene was associated with specific diseases. In answer to these problems, a new way of looking at genetic information called a genome browser was devised, which works much like a web browser, but specific to our genes and information associated with them. One of the most popular such browsers is the one devised by the University of California at Santa Cruz (UCSC), called for short the UCSC genome browser. The UCSC browser can be used to look at information from other species too, such as mouse and many others: a link to human genome browser version can be found here: <https://genome-euro.ucsc.edu/cgi-bin/hgGateway?redirect=manual&source=genome.ucsc.edu> and should look like:

Request:	Genome Browser Response:
chr7	Displays all of chromosome 7
chrUn_gl000212	Displays all of the unplaced contig gl000212
20p13	Displays region for band p13 on chr 20
chr3:1-1000000	Displays first million bases of chr 3, counting from p-arm telomere
chr3:1000000+2000	Displays a region of chr3 that spans 2000 bases, starting with position 1000000

A simple way to think of the UCSC genome browser is to think of it as basically a zoomable map, not unlike Google maps. It shows us a representation of chromosomes in our cells, with the positions of the different genes indicated on each. We can take a virtual tour of our genome by scanning around, or zoom in to look at specific regions in greater detail. Most usefully, we can search for a genetic “address” and the browser will locate it for us and take us to a close-up view of the gene and its surroundings.

To get started, write in the name of a gene in the “Position/Search term” box (outlined in red above). Like many apps, the browser tries to guess your destination, so start by writing the name for the human hemoglobin gene “Hbb” in the box: the “best guess” will appear under the box, choose that using the mouse so that it is highlighted and appears in the box (right) and click the big blue “Go” button.

Position/Search Term

hbb

HBB (Homo sapiens hemoglobin, beta (HBB), mRNA.)

HBBP1 (Homo sapiens hemoglobin, beta pseudogene 1 (HBBP1), non-coding RNA.)

This will open a new view in the browser which may be quite crowded with information, like a busy map. This is because, like the maps we use to navigate in everyday life where we can find the nearest coffee shop or petrol station, new interesting information and locations are constantly being added to the genome map, a bit like a news feed, and this can make it quite busy.

To get a simple map with just the Hemoglobin gene, do the following:

- Click on >Hide all (red box below)
- Under UCSC genes, choose “dense” (yellow box below)

The view should now look something like this:

The screenshot shows the UCSC Genome Browser interface for the Hemoglobin gene on human chromosome 11. The top navigation bar includes links for Genomes, Genome Browser, Tools, Mirrors, Downloads, My Data, View, Help, and About Us. The main title is "UCSC Genome Browser on Human Feb. 2009 (GRCh37/hg19) Assembly". Below the title is a search bar and a zoom control. The chromosome ideogram shows the location of the gene on chromosome 11. The scale bar indicates the genomic coordinates. The track controls include "move start", "move end", "track search", "default tracks", "default order", "hide all", "add custom tracks", "track hubs", "configure", "multi-region", "reverse", "resize", "refresh", "collapse all", and "expand all". The "Mapping and Sequencing" section is expanded, showing a grid of tracks with dropdown menus for each. The "UCSC Genes" track is highlighted with a yellow box, and the "hide all" button is highlighted with a red box.

- Here we see the Hemoglobin gene (labelled UCSC gene) outlined on the map as black boxes joined by thin lines: these are the exons (boxes) and introns (lines)
- Above there is a little picture or ideogram of human chromosome 11, showing where the gene is located (red line at left)

There is a wealth of other data available for each gene, each available through the pull-down menus shown. Each generates a new line or “track” on the map below the gene, showing the new information in parallel. To get a flavour of this, the user can try “default tracks” (green box above) which will show the same gene, but with information below it from every major group (blue header):

- UCSC\_Genes The default map, best current estimate of the start and end of the gene
- NCBI\_RefSeq Gene start and end as defined conservatively, based on curated data
- Publications Scientific papers with links to this gene
- GTEx gene Data on where the gene is thought to be expressed
- ENCODE Clues as to how the gene may be regulated based on epigenetic info
- Conservation Showing which regions of the gene may be conserved in vertebrates
- dbSNP\_153 Known variations in the DNA sequence at this gene

These lines of information are fully clickable in the top window: right-clicking or double-clicking can bring you to pages with further information on each track.

We can also add our own data to the genome maps using our own tracks, which is what we do using CandiMeth (see section 3.1.6 f in main Guide). More details on how to export data, views of the maps and more can also be found there.

Like political maps, the human genome map is constantly being updated with the latest borders and newest information. These updates are known as genome builds or genome assemblies, an example is the GRCh37/hg19 (Genome Release Consortium human build 37) genome build used in section 3.1.

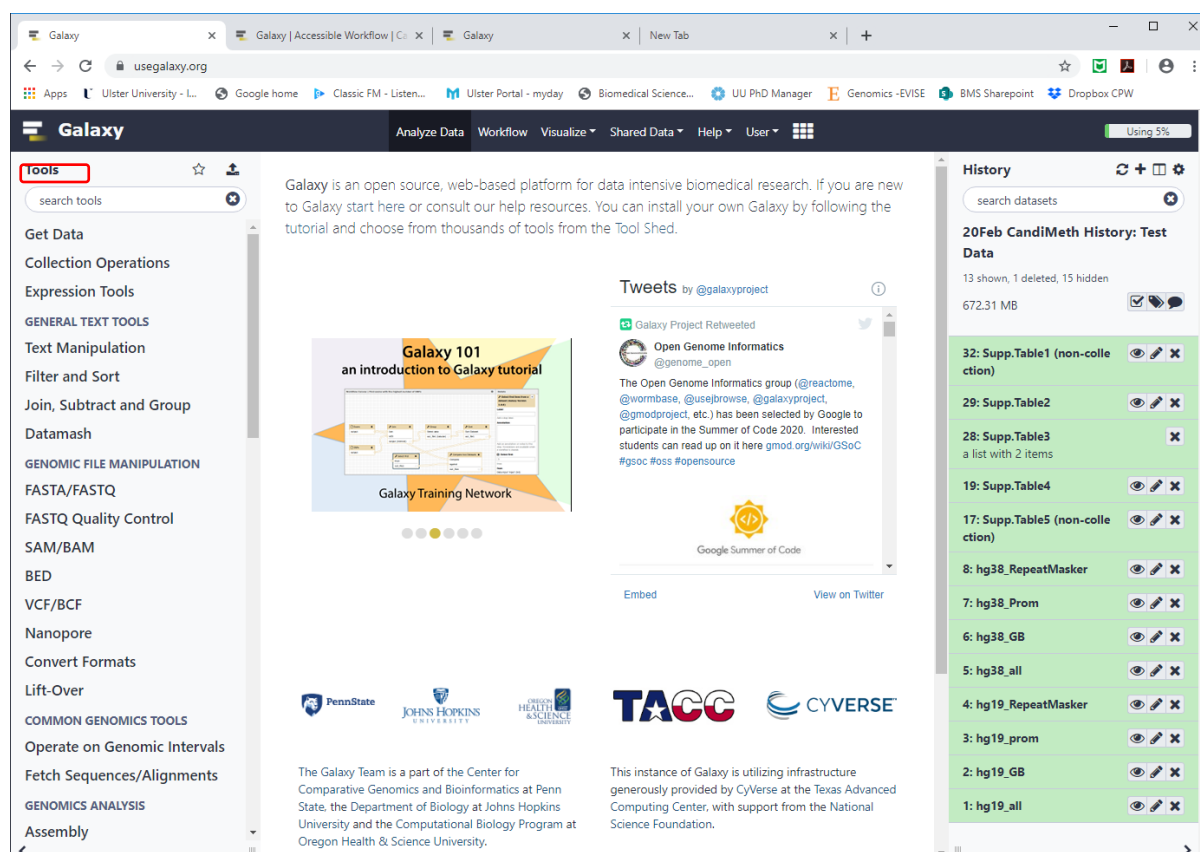
Further documentation on the UCSC genome browser can be found here:



<https://genome.ucsc.edu/goldenPath/help/hgTracksHelp.html>

-Happy browsing!

## Appendix 2. Quick guide to the Galaxy web environment

Galaxy is a free online environment for user friendly data science. It can be located here, <https://usegalaxy.org/> and requires users to create an account and log-in to utilise the service (registration for this service can be found here, [https://usegalaxy.org/root/login?is\\_logout\\_redirect=true](https://usegalaxy.org/root/login?is_logout_redirect=true)). Once logged in the home page should look something similar to the below image:

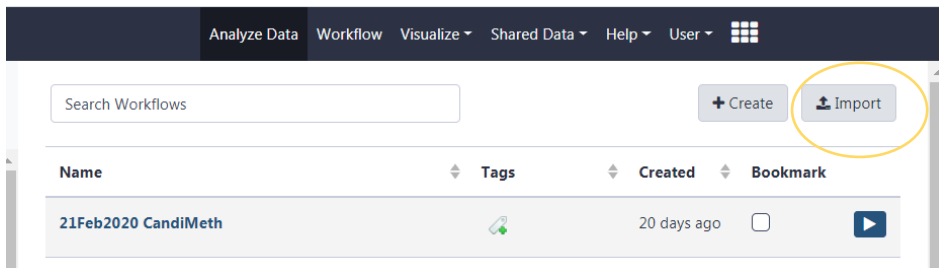


Here, tools can be found in the tools panel on the left-hand side and any process that the user has executed can be found in the history column in the right-hand side. Every process a user executes creates a new item in the history column. Multiple histories can be created via the gear  icon and multiple histories can be viewed using the book-like  icon. Data can be uploaded to the Galaxy interface, as detailed in section 4.3 and using the “Get data” function (red box above) the user can load data from publicly available external sources such as UCSC genome browser or NCBI.

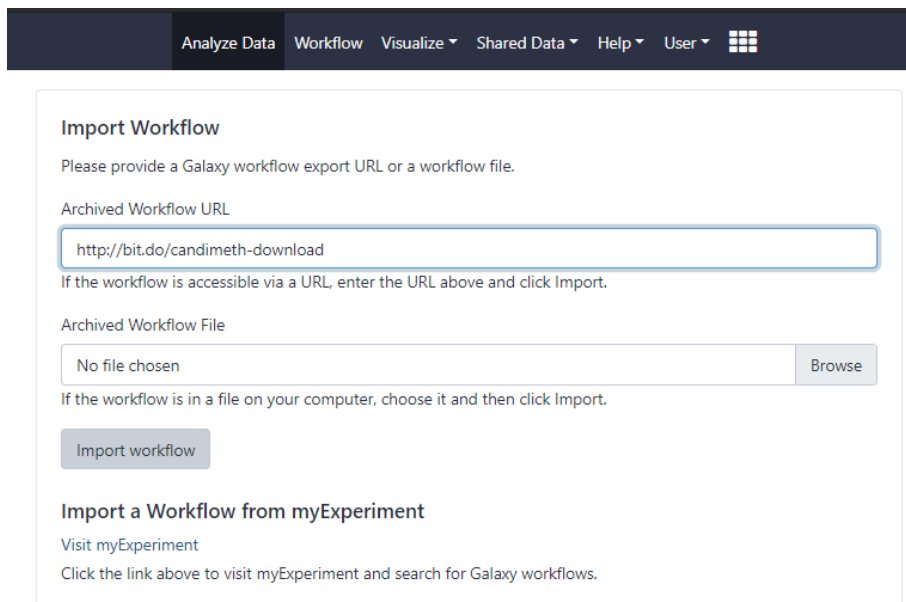
Further detail regarding the Galaxy interface can be found here (<https://galaxyproject.org/tutorials/g101/>) or in the Galaxy Training Network (<https://training.galaxyproject.org/>).

## Appendix 3. Importing CandiMeth to a custom Galaxy instance

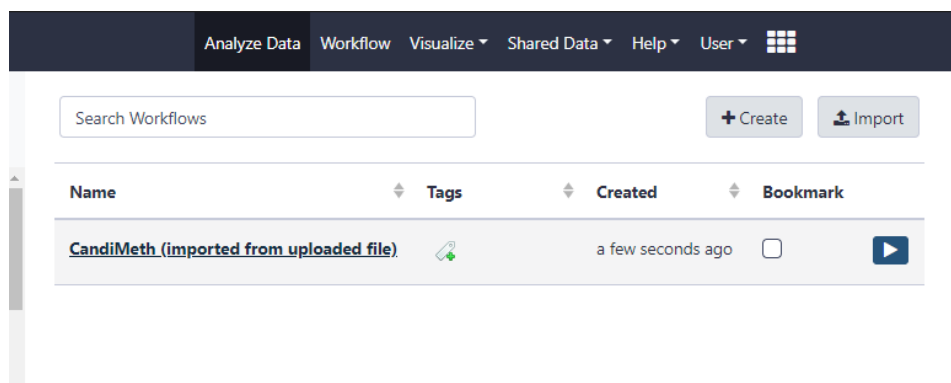
1. To download and upload CandiMeth to an alternative Galaxy instance, navigate to the workflow tab using the “workflow” tab at the top of the Galaxy window.
2. Click on the import button (circled in yellow here) at the top right of the window



3. Paste this link <http://bit.do/candimeth-download> into the “Archived Workflow URL” section of the import workflow screen and click on import workflow, as below.



4. CandiMeth should then show up in the list of workflows available to you, as below.





## 7.0 General Discussion

In this thesis, differences in DNA methylation between sample sets have been studied, with some of these sample sets arising from cell line work and some from groups of human participants in various studies. The common denominator here is that the studies have all involved arrays and other high-dimensional genomic datasets which have required extensive bioinformatic handling, which has been a major component of my work. This has led to several publications and the development of a new tool for analysis of such data. In this final chapter, I wish to relate the results of the presented work to that of recently published research to enable conclusions, limitations and the direction of further work to be established.

### 7.1 Effects of perturbing the basic methylation machinery (Papers I and II)

Depletion of DNMT1 highlighted several classes of genes regulated by DNA methylation including Protocadherins (PCDH), Fat/Body Mass (FBM) genes and olfactory receptor (OR) genes.

#### 7.1.1 Neuroepithelial genes

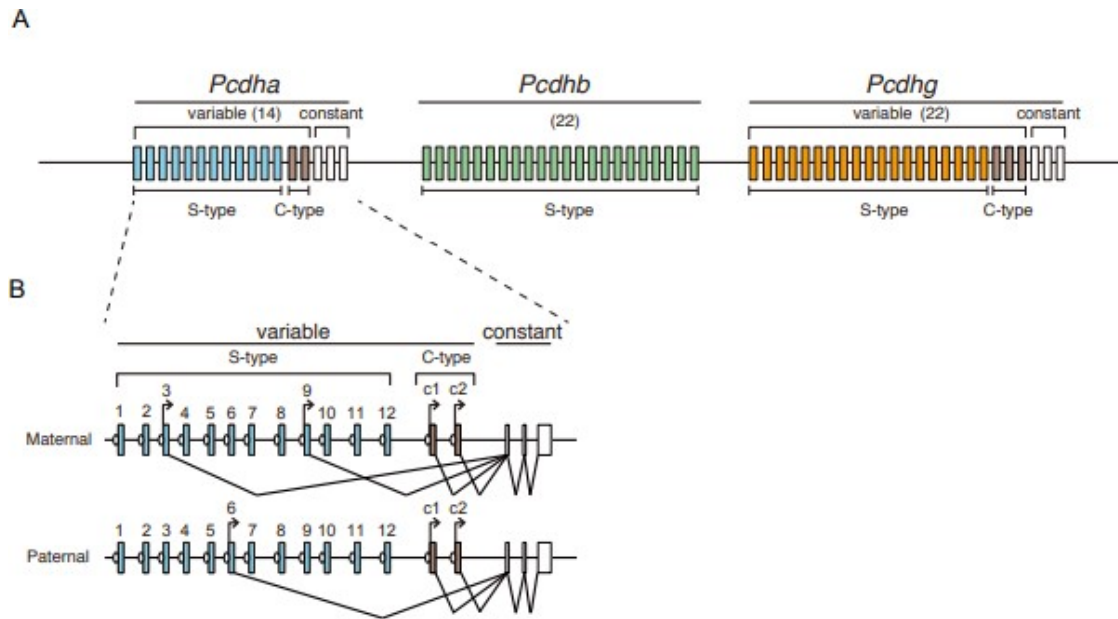
The OR and PCDH genes have a number of striking features in common: 1)they are usually expressed in the nervous system; 2)they form neural cell-cell adhesion proteins; 3)they are essential to neural cell identity (El Hajj et al., 2017) (Monahan and Lomvardas, 2015); 4)many show monoallelic expression and 5)they are known to be regulated, at least in part, by epigenetic means (El Hajj et al., 2017). It was unclear to us initially which of these features, singly or in combination, explained why these gene classes were enriched in the DNMT1 KD cells, and why the same gene classes were not highlighted as clearly in the GO analysis of methylation changes in UHRF1 KD cells. To gain a better understanding of why

this might be, we had to look a little more closely at the structure, function and expression patterns of these genes.

PCDH and OR genes both encode types of extracellular receptor, important for different types of extra-cellular binding. In the case of the OR genes, they form one of the largest gene families in the human genome and have diversified to bind different odorant molecules: binding triggers a G-protein coupled signaling cascade in the cell, triggering a neural response. Odorant cells are located in the naso-epithelial membranes and individual OR receptors are only expressed by a small number of cells in a cluster. Since only one OR out of the hundreds available is expressed, the others become inactivated and heterochromatinised, eventually accumulating DNA methylation as well. In fact, only a single allele of each OR is usually expressed, in a system analogous to antibody variable chain expression. OR gene expression is thus tightly linked to cellular identity for the expressing cells (Antunes and Simoes de Souza, 2016; Maßberg and Hatt, 2018; Monahan and Lomvardas, 2015).

The PCDH gene expression is also immensely variable between neural cells, but they solve the problem of generating diversity through another mechanism. Rather than having a large battery of separate genes from which one is chosen, two of the major sets of PCDH genes,  $\alpha$  and  $\beta$ , are arranged in clusters with alternative splicing generating different functional isoforms using combinatorial coding, in a similar fashion to the immunoglobulin cluster (Figure 1) (Collins and Watson, 2018). Here again the genes code for extracellular receptors, but rather than detecting odorant molecules these bind to PCDH proteins on neighbouring cells and repel receptors on their own cell surface. The expression of different domains on the mature proteins allows neighbouring cells to distinguish self from non-self. Again, only

one combination of exons is used in each cell, often monoallelically, with inactive exons being heterochromatinised and eventually methylated as well (Canzio and Maniatis, 2019; Hirayama and Yagi, 2017; Mountoufaris et al., 2018; Peek et al., 2017).



**Figure 1: Structure and expression of PCDH gene cluster.** Structure of the PCDHA, PCDHB and PCDHG clusters as found in mouse. They are located on a 900kb stretch of DNA in chromosome 18 and consist of variable (shown in brackets) and constant exons as indicated (A). The PCDHA cluster is shown to demonstrate the clusters monoallelic expression. Every variable exon exhibits its own promoter (circle in front of the exon) found upstream of the coding region. These promoters are spliced and form the constant exons. Stochastic (S) isoforms are expressed by each allele while constitutive (C) isoforms can be found expressed via both alleles. Adapted from (Hirayama and Yagi, 2017).

We thus considered that it was most likely that the enrichment for OR and PCDH genes in the DNMT1 KD might reflect the fact that inactive members of both clusters become heavily methylated, particularly in cell lines. Since they are highly methylated, but not derepressed when methylation is removed in fibroblasts due to the absence of tissue-specific transcription factors (no up-regulation on RNA arrays), they would be more likely to show hypomethylation in the fibroblasts since there will be no selection against this. In the case of the OR genes this would be compounded by the fact that there is a large family of separate genes (Monahan and Lomvardas, 2015), artificially inflating the GO scores.

However, this could not be the whole answer for the PCDH genes, since there is a relatively small number and several other factors suggested that the PCDH gene sensitivity was important: 1)PCDH gene enrichment was not seen in many of our studies, unlike OR genes which were enriched in a number of experiments (see paper II Fig 1, (Mackin et al., 2018)) and we thus discounted as being of special significance in paper I 2)the enrichment reflected in part the very high levels of DNA methylation seen at this cluster, including at CGI, which are rarely methylated except when important for regulation (Deaton and Bird, 2011); 3)the cluster is also regulated by CTCF, which in some cases is CG-sensitive (Golan-Mashiach et al., 2012; Mountoufaris et al., 2018), further suggesting an important role and 4)we did not see a general enrichment for all other loci known to be heterochromatinised and inactivated in our DNMT1 lines (such as CTA genes, immunoglobulin genes, germline genes or imprinted genes). On this latter point, this is exactly what we did see in UHRF1-depleted cells, where there was so much general demethylation of all heterochromatinised genes that no clear pattern of enrichment was apparent on looking at the methylation arrays. We concluded that PCDH genes are particularly sensitive to loss of DNMT1 compared to many other

epigenetically regulated genes, and that it is particularly important to maintain methylation on these genes in adult.

Consistent with this picture, there is dedicated de novo activity from DNMT3B which can restore any methylation loss at PCDH clusters, as we showed using DNMT3B KD in the hTERT cells (Suppl. Figs, Paper I). Consistent with this picture, Toyoda and colleagues showed in an elegant paper which was published while our manuscript was under development, that DNMT3B is particularly important for establishing methylation at inactive PCDH variants (Toyoda et al., 2014). Additionally, Rajarajan and colleagues also recently published work showing that SNPs which affect DNA methylation at PCDH cluster genes are associated with mental health phenotypes (Rajarajan et al., 2018).

Interestingly PCDH clusters have also been shown to be involved in cell proliferation/death pathways and can alter the WNT pathway signaling (O’Leary et al., 2011). For example, overexpression of PCDHGA1 has been found to upregulate the WNT pathway (Mah and Weiner, 2017). Moreover, PCDHGC3 can downregulate WNT signaling (Mah et al., 2016), which correlates with the discovery of PCDH cluster dysregulation in various forms of cancer, like hypermethylation of PCDH in Wilm’s Tumor (Dallosso et al., 2009). A role in triggering cell death would be consistent with neural arborisation and pruning during CNS development, as discussed above.

#### 7.1.2 Body mass regulation

In addition to PCDH, loss of methylation was also observed in genes related to fat and body mass homeostasis from enrichment analyses. The most common theme to these genes is some form of triglyceride processing, however, many candidate genes also share an immune link; ANXA2 (Liu and Hajjar, 2016; Y. Liu et al., 2015), disease association (GHSR; (Z.

Liu et al., 2015)) and even a link to cancer (ANXA2; (Kpetemey et al., 2015), ERBB2; (Kushwaha et al., 2019), GHSR; (Jandaghi et al., 2015; Lin and Hsiao, 2017), SHH; (Samkari et al., 2015)), as well as a link to embryogenesis (NLRP2; (Peng et al., 2017) SHH; (Lopez-Rios, 2016)). These associations and the aberrant methylation found via this intervention suggest a link to the foetal origins of adult disease hypothesis, as discussed in section 1.8.1.1.

However, little difference in the methylation of ANXA2 and APOC1 following DNMT1 depletion was observed. The lack of effect here could be because these genes are not solely regulated by DNA methylation, as there is a small gain in the APOC1 gene, or that specific transcription factors are required to derepress these genes. Moreover, these genes do not seem to be marked by polycomb or poised promoter chromatin states but the transcription status of these genes would need to be assessed to further investigate this, in addition to the binding sites of H3K27me3/H3K4me3. However, the probe resolution at the gene body of APOC1 is quite low, this could affect methylation results at this gene and explain the small change.

### 7.1.3 The *UGT1A* detoxification gene cluster

In addition to losses in methylation upon DNMT1 KD, there were unexpected gains in methylation, particularly at the UGT1A locus and a variety of genes known as Cancer Testis Antigens (CTA) on the X chromosome. UGT1A genes can be found on chromosome 2q37 and are responsible for glucuronidation, inactivating their targets so they are excreted. Targets include steroid hormones, bilirubin, carcinogens, bile acids and therapeutic drugs (Hu et al., 2016). It is thought that DNA methylation within the UGT1A locus acts as a guide to facilitate correct alternative splicing patterns (Habano et al., 2015). It has also been noted that UGT1A1 and UGT1A10 are regulated by DNA methylation but not much has been established regarding the rest of this gene family (Oda et al., 2014; Yasar et al., 2013). Upon

siRNA KD of DNMT3B, the methylation of DNMT1 targets PCDHGA2, LEP and UGT1A4 lost methylation indicating that, although DNMT3B did not appear to be overexpressed in these cells, it still provided maintenance methylation at these targets (Paper I Supplementary figure 4B). This could explain the hypermethylation observed at the UGT1A locus, in addition to the location of the UGT1A promoter locations within heterochromatic regions which may be more susceptible to methylation (adjacent active non-heterochromatic regions show a restoration of normal methylation). However, literature has stated that stable KDs in long term culture are affected by a cumulative gain in methylation (Bork et al., 2010; Ehrlich et al., 1982; Gordon et al., 2014; Landan et al., 2012). This is more likely the reason behind the unexpected gains in methylation at this locus as hypermethylation was not observed within an siRNA KD of DNMT1. Although, many of the gains observed in shRNA KD of DNMT1 do overlap with the poised promoter category of chromatin state segmentation from the ENCODE project (Ernst et al., 2011a).

#### 7.1.4 Cancer-testis genes

CTA genes are a large family of genes usually expressed in the testis and placenta but have also been found to have abnormal expression in some cancers (De Plaen et al., 1994; Salmaninejad et al., 2016), such as gonadoblastoma (Kido and Lau, 2014) and melanoma (Hagiwara et al., 2016). Evidence also suggests that CTA are a form of expressed pseudogenes. CTA based pseudogenes have been isolated in both human and other mammalian genomes at MAGEA, MAGEB, SSX and CT47 gene loci (De Plaen et al., 1994; Gordeeva, 2018; Güre et al., 2002; Zhao et al., 2012). According to our KD of DNMT1, and many other investigations using Aza (Mackin et al., 2018; Salmaninejad et al., 2016; Wrangle et al., 2013), the CTA genes are directly regulated by methylation. Upon treatment with Aza, Weber and colleagues (1994) discovered CTA genes, such as MAGE-A1, lose methylation and



become transcriptionally active in melanoma (Weber et al., 1994) and later studies demonstrated their regulation by methylation in many other cancer types including non-small cell lung cancers - identifying CTA genes as a potential target for intervention during cancer treatment (Juergens et al., 2011; Wrangle et al., 2013). Since then, many CTA-based immunotherapies have been tested as potential cancer immunotherapies (NCT00960752, NCT00960752, NCT00960752, <https://clinicaltrials.gov/>) and some are being combined with Aza to increase the efficacy of the CTA treatment (Adair and Hogan, 2009).

CTA genes were also demethylated and subsequently upregulated in our UHRF1 KD.

Although, not identified in the GO analysis, CTA genes did show transcriptional upregulation in our transcriptional array and demethylation of some promoters with hypermethylation of gene bodies. These alterations were later verified using a candidate gene assessment as available in CandiMeth (Paper II, Paper V) and certain targets were selected for validation via pyrosequencing.

The differences observed between these two KD could be because of the more widespread regulatory effects of UHRF1, including cell cycle regulation, regulation of the DNA damage response (DDR) in addition to regulation of DNMT1 (Li et al., 2018; Xie and Qian, 2018).

Alternatively, and perhaps more likely since DNMT1 is also implicated in all these functions, is that the cells do not tolerate the loss of the DNMT1 protein as well. Many reports suggest a KD of DNMT1 causes replicative stress, resulting in proliferation being halted at the G0/G1 phase and triggers DDR pathways which ultimately results in apoptosis (Brown and Robertson, 2007; Liao et al., 2015; Loughery et al., 2011; Milutinovic et al., 2003; Sharif et al., 2016). Fitting with this idea is the hypothesis that, upon loss of DNMT1 the DDR response facilitates the removal of cells with low levels of DNMT1 (and thus lack

methylation) and exhibits a strong preference for those which remain methylated, masking any sign of immune upregulation (Loughery et al., 2011). An alternative investigation also reported global hypomethylation in a DNMT1 KD once the DDR pathway had been blocked (Unterberger et al., 2006). Other data from the Walsh lab have also shown that DNMT1 depletion, but not reduction in UHRF1, triggers PARylation of proteins, a prelude to one form of DDR-mediated cell death (Scullion and Walsh, unpublished data). It is not entirely clear why UHRF1 does not elicit the same response, but it may be due to a less immediate role in base modification. Overall differences in these two KDs can be observed in Table 1.

DNMT1 KD			UHRF1 KD				
Group	Methylation		Transcription	Group	Methylation		Transcription
	Promoter	GB			Promoter	GB	
OR	Red	Red	-	OR	Red	Red	-
CTA	Red	Light Green	Dark Green	CTA	Red	Light Green	Dark Green
PCDH	Red	Red	-	ERV	Red	Red	Light Green
FBM	Red	-	-	IFN	-	-	Light Green
UGT1A	Light Green	Light Green		ISG	-	-	Red

**Table 1: Overall effects of depletion of DNMT1 and UHRF1 in adult immortalised fibroblasts.** Upon knockdown of DNMT1 hypomethylation was observed throughout PCDH, OR, CTA and FBM gene clusters. However, hypermethylation was observed at the UGT1A cluster. Upon UHRF1 knockdown, similar hypomethylation could be observed at CTA loci. In addition to, hypomethylation at ERV and VDG with increases in transcription of these related genes. In addition to increases in transcription at IFN and ISG genes. Abbreviations: Protocadherin (PCDH), Olfactory (OR), Cancer Testis Antigen (CTA), Fat and Body Mass genes (FBM), endogenous retroviral elements (ERV), Interferon genes (IFN), Interferon stimulated genes (ISG) ‘-’ indicates no applicable data available. Red indicates demethylation (downregulation in the context of transcription). Light green indicates hypermethylation (upregulation in the context of transcription). Dark Green indicates vast upregulation in the context of transcription.

#### 7.1.5 Activation of innate immune genes in UHRF1-depleted differentiated human cells

However, the CTA genes were not the only gene categories transcriptionally upregulated in the UHRF1 KD. Upon GO of the genes which were upregulated, all ten hits were involved in the innate immune response including; interferon genes (IFN), interferon stimulated genes (ISG) and components of the major histocompatibility complex (MHC) in addition to the CTA genes. While there is a genome-wide demethylation in UHRF1 KD cells, demethylation at the IFN and ISG was not significantly correlated with their expression (data not shown), suggesting an indirect up-regulation here, mirroring the findings of Chiapinelli and colleagues (2015), termed this state 'viral mimicry' as it mimics the response to invading viruses by the innate immune system when they replicate within the cell. They discovered this via Aza-mediated DNA methylation inhibition in a variety of cancer cell lines, including activation of  $INF\beta$  through the JAK/STAT pathway (Chiappinelli et al., 2015), as in our KD of UHRF1.

Roulois and Colleagues also uncovered an innate immune response in colorectal cancer cell lines via Aza-mediated DNMTi but after clustering, split their results into four groups depending on whether the genes displayed early or late responses to Aza and their correlation between methylation and expression. The groups observed here are very similar to what is occurring in our UHRF1 KD, only a low percentage of genes in our study showed demethylation and upregulation and the majority of genes assessed were demethylated but not upregulated – indicating regulation independent of methylation. Also, Roulois and colleagues mention that even though they observed interferon (INF) upregulation and demethylation there was no additional INF protein present within the cell (Roulois et al., 2015), this may be an point to assess in our UHRF1 KD cell lines as it indicates that an alternative mechanism is abrogating the formation of INF proteins even after transcriptional

upregulation. A similar clustering effect was also found in an separate similar study (Wrangle et al., 2013).

Interestingly, there were minor differences in the pathways Chiappinelli and colleagues uncovered compared to ours: for example, they found one of their biggest changes in response to DNMT inhibition (DNMTi) was from the transcription factor IRF7 (Chiappinelli et al., 2015), but in our work in adult fibroblasts (Paper II) we found more upregulation in IRF9. In addition to this, Aza has been found to affect G9A/GLP histone methyltransferases which again could be influencing results (Ferry et al., 2017). However, Chiappinelli and colleagues did compare their results to a HCT116 DKO deficient in DNMT1 and DNMT3B and found relatively similar results in comparison to their Aza mediated DNMTi, indicating the viral response may be a definitive consequence of DNMTi, independent of the mechanism of depletion.

Consistent with what we saw, Chiappinelli et al. found that IRF7 was hypermethylated at the promoter in only 1 of their 23 epithelial ovarian cancer cell lines, and that there was a poor correlation between IFN/ISG gene hypomethylation and their expression, and concluded that an alternative mechanism may cause the viral defence response when IRF7 is not silenced. This matches well with a review of the viral immune response from Strick and similar colleagues which shows that  $INF\alpha$  and  $INF\beta$  activation following DNA methylation intervention can be mediated by IRF7 or IRF3 upregulation (Strick et al., 2016). It is also suspected that IRF9 is involved further down the viral response pathway and works with STAT proteins to initiate interferon stimulated gene (ISG) transcription (Chiappinelli et al., 2015). While IRF7 upregulation is observed in our UHRF1 KDs, it may be that IRF9 is more important in our normal fibroblast cell lines as opposed to cancerous epithelial cell lines.

There is also a greater ISG response found using alternative inhibitors of UHRF1 (Cuellar et al., 2017; Liu et al., 2016), in the latter of which upregulation of T-cell signaling genes was shown (Cuellar et al., 2017), which is also found in our UHRF1 KDs.

UHRF1 KD in HCT116 was previously assessed for viral defence gene (VDG) upregulation by Cai et al (2017), but primarily to attempt to find a potentially viable method of genome-wide DNA demethylation. Interestingly, they found similar levels of IRF7 and IRF9 upregulation upon a 60% UHRF1 KD and greater upregulation after combination treatment with Dacogen (Aza) (Cai et al., 2017b). In addition, similar results were found by Wrangle (2013) upon Aza-mediated DNMTi as in our hTERT UHRF1 KDs (Wrangle et al., 2013), in addition to high IRF7 and IRF9 upregulation in Li (2014) (Li et al., 2014). Conversely, when Cai and colleagues assessed the same VDG panel in HCT116, shRNA-mediated in DNMT1 hypomorphic HCT116 cell lines gave high VDG upregulation, but not in those cells in the absence of shRNA (Cai et al., 2017b). This indicates there is a threshold for VDG upregulation, at least in cancerous epithelial cell lines (Cai et al., 2017b; Chiappinelli et al., 2015; Roulois et al., 2015). Thus, our DNMT1 KD cell lines may not have low enough levels of DNMT1 to elicit an innate immune response.

We assessed the transcription of the same VDG list in the UHRF1 WT and UHRF1 KD and found a large increase in the transcription of these VDG between WT and KD. This encouraged us to develop a larger VDG panel and a gene panel for interferon and interferon stimulated genes to assess the extent of immune activation in our UHRF1 KD cells.

Although, it is of note that our UHRF1 and DNMT1 KDs were composed of immortalised fibroblasts which may have different proliferative and DNA methylation profiles than the HCT116 cells utilised in the Cai, Chiappinelli and Roulois investigations (Cai et al., 2017a;

Chiappinelli et al., 2015; Roulois et al., 2015). HCT116 cells are also a cancerous cell line, these types of cell lines are known to have abnormal methylation patterns in comparison to non-cancerous cell lines (Kamińska et al., 2019).

#### 7.1.6 ERV reactivation and innate immune response to Uhrf1 mutation in mouse

In contrast to our results in human cells, where UHRF1 but not DNMT1 depletion resulted in ERV reactivation, in mouse Sharif (2016) reported that DNMT1 but not UHRF1 depletion resulted in mouse ERV derepression (Sharif et al., 2016). Upon multiple conditional Kos (cKO) of DNMT1, UHRF1 and a DNMT1 and UHRF1 DKO, IAPez upregulation was greater in an ESC DNMT1 cKO as opposed to the UHRF1 ESC cKO or the double cKO of both enzymes together (Sharif et al., 2016). However, the level of IAP upregulation found in the DNMT1 ESC cKO was not as high as that found in the original KO by Walsh Chaillet and Bestor in 1998 (Walsh et al., 1998).

In addition, Hutnick (2010) conducted a similar KO of DNMT1 in mouse ESCs and did not observe ERV upregulation until induction of differentiation (Hutnick et al., 2010). Also, in our UHRF1 CRISPR mediated KO of UHRF1 in mouse embryos, we can replicate a similar immune response as within the UHRF1 KD in hTERT and SKMEL melanoma cells. In addition, our WT and CRISPR mediated UHRF1 heterozygote mouse (#1) does look morphologically similar to the WT and Cre-mediated mouse UHRF1 conditional knockout (UHRF1cKO) in Sharif et al 2016 (Sharif et al., 2016). In addition to, the differences in response between mouse (as in Sharif 2016) and human (our UHRF1 KD), Sharif and colleagues also use Cre-mediated conditional KOs, which could be considered a transient KD system. Therefore, their results could indicate the effects of acute depletion as opposed to long term shRNA mediated KDs.

When we conducted a CRISPR KO of UHRF1 in mice, we found a very reproducible demethylation of ERV elements but the reaction at IFN and ISG related genes was a lot more variable in our mice, in comparison to that of Sharif and colleagues 2016. However, there is a lack of methylation and transcriptional arrays that cover ERV elements at a high resolution. This led to the 450k array being repurposed to assess if we could obtain reliable coverage of ERV methylation and transcriptional response of ERV elements being measured by RT-qPCR. In order to visualise the results of the variable transcriptional response, I plotted the results as a heatmap to instead show overall patterns in transcriptional change as opposed to multiple graphical visualisations. Further discussion on this mouse work and the ERV response can be found in the next section.

UHRF1 mutants in Zebrafish also display an innate immune response, in addition to mutant DNMT1 Zebrafish embryos, but not to as high a degree (Chernyavskaya et al., 2017). Of note however, was the observation of cytosolic dsDNA as well as dsRNA in the UHRF1 mutant (Chernyavskaya et al., 2017). This was investigated but not observed in our UHRF1 KD (Paper II). They also observed MAVS upregulation in their UHRF1 (IRF7 and STAT1 upregulated in this mutant too) and DNMT1 mutant Zebrafish embryos (Chernyavskaya et al., 2017). Therefore, since they did not test for any specific member of the dsDNA pathway like MYD88 or TLR9 (Strick et al., 2016) but did observe MAVS upregulation (a key component in dsRNA signaling (Strick et al., 2016)) - the upregulation of STING could be due to the DDR response, since as cells become apoptotic they release dsDNA into the cytoplasm and are removed from the embryo. This theory also matches well with the upregulation of macrophages within mutant Zebrafish embryos (Chernyavskaya et al., 2017). Despite some inconsistencies between the two papers from the Sadler lab then, the



upregulation of IRF9 in siRNA-mediated UHRF1 KD of human hepatoma cells (Chernyavskaya et al., 2017) nevertheless matches well with that observed in our UHRF1 KD (Paper II).

Of additional note was that the gene ontology results implicated the PIWI-interacting RNA (piRNA) pathway in cellular response to UHRF1 KD (Paper II). From examination of the benchmark set of viral response genes from Cai et al, it was also clear that at least 1-2 of these were also involved in this pathway. piRNA has been found to interact with UHRF1 to repress transposable elements in male germ cells (Dong et al., 2019): disruption of piRNA in mice led to upregulation of transposable elements and male sterility due to malformations in testes development. In this conditional KO of the fourth exon of UHRF1, decreases in H3K9me2/H3K9me3 and increases in H3K4me3 were found in cKO spermatocytes. UHRF1 was also found to interact with the arginine methyltransferase PRMT5, which is known to suppress transposable elements and interact with piRNA proteins. Upon cKO, mRNA of PRMT5 was also decreased and it is thought that PRMT5 could affect UHRF1 localisation in the nucleus of spermatocytes. Upregulation of LINE1 and TE was again found in UHRF1 cKO in mouse testes and spermatocytes (Dong et al., 2019).

## 7.2 Methylation-deficient systems are indicative of alternative repressive mechanisms (Paper I and II)

### 7.2.1 UHRF1

Upon rescue with full length UHRF1 cDNA, we found that methylation genome-wide, including at ERVs, does not recover. However, ERV expression and the innate immune response previously observed appears to be attenuated independent of methylation. This indicated a possible alternative mechanism of repression was facilitated upon rescue of UHRF1. However, DNA methylation is not the only known method of ERV repression, microRNAs (Schorn et al., 2017), H3K9me3 including SETDB1/KAP1 based mechanisms

(Rowe et al., 2013), H3K27me3 (Li et al., 2017; Wang et al., 2019) and piRNA have all been found to repress ERVs in the absence of DNA methylation (Dong et al., 2019). In ESC, the primary method of ERV repression is via H3K9me3, with secondary assistance from DNA methylation (Bulut-Karslioglu et al., 2014; Karimi et al., 2011; Li et al., 2018; Matsui et al., 2010), but DNA methylation appears to be the primary mechanism in differentiated cells (Kassiotis and Stoye, 2016; Mikkelsen et al., 2007; Rollins et al., 2006; Smith et al., 2012).

Wang and colleagues knocked out UHRF1 in the livers of mice and found transposable elements to be subsequently enriched for both H3K27me3 and H3K9me3. However, H3K27me3 enrichment was only redistributed to hypomethylated TEs, such as the 5' region of SINEs and LTRs with high CpG density, and not those that had retained residual methylation or had low CpG density. Interestingly, H3K27me3 was not found at IAPs (Wang et al., 2019). This matches well with the work of Walter and colleagues: here, embryonic stem cells were depleted of methylation via TET-mediated active demethylation to inhibit the action of DNMT3A and DNMT3B (Walter et al., 2016). This work suggested that there were 3 categories of transposable elements; category 1 showed co-occupation of the TEs with H3K9me3 and H3K27me3 (found in LINEs); category 2 showed H3K9me3 marking only (found in IAPEz elements) and finally category 3, which showed a transit from H3K9me3 to H3K27me3 on MERVL elements in response to DNA methylation depletion (Walter et al., 2016). The redistribution of H3K9me3 after loss of methylation at IAPEz matches what we suspect occurs with our UHRF1 KD. However, we did not see a large change in H3K9me3 when tested by western blot (Paper II). H3K9me3 is restricted to the 5' UTR of LINEs but occurs evenly across ERV elements and could even be found in neighbouring genomic regions, whereas H3K27me3 could only be observed within the sequences of TEs (Bulut-Karslioglu et al., 2014; Walter et al., 2016).

Walter and colleagues also created a haploinsufficient SETDB1 /KAP1 mutants and found that category B elements like IAPEz elements showed high upregulation and failure to repress IAPEz or ERV elements (Walter et al., 2016), which is similar to what occurs in our UHRF1 SETDB1 and KAP1 inhibitions (Paper II). In other studies KAP1 KD in ESCs led to a marked increase in retroviral expression and a decrease in SETDB1 binding, with a subsequent depletion of H3K9me3 marks at these TEs (Matsui et al., 2010). This adds to the evidence that, in our UHRF1 KD, in the absence of DNA methylation, H3K9me3 could be the alternative mechanism repressing ERVs and IAP elements through primarily SETDB1/KAP1 mediated mechanisms as opposed to Suv39h based mechanisms (Bulut-Karslioglu et al., 2014; Matsui et al., 2010; Rowe et al., 2013). The afore-mentioned papers do point towards a complex system of H3K9me3, H3K27me3 and DNA methylation-mediated TE repression. However, this requires further investigation in our UHRF1 KD to clarify the effects in adult human cells. This could be accomplished via ChIP-seq enriching for H3K9me, H3K27me3, SETDB1 and KAP1, in addition to, assessments of DNA methylation, such as RRBS before and after KD of UHRF1 in these fibroblasts. This would give greater coverage in repetitive element loci. It may also be of interest to conduct RNA-seq to assess the transcriptional response of ERVs and innate immune genes following KD of UHRF1 as identified in the Chiappinelli study of 2015 that was discussed in the previous section (Chiappinelli et al., 2015). Obtaining genome wide assessments of methylation, transcription and certain histone marks and protein binding sites may also identify any alternative changes that been overlooked due to the concentration of RT-qPCR and western blotting on genes and histones related to the innate immune response.

To investigate the possibility that H3K9me3 marks in DNA methylation-deficient cells may act as a signal for UHRF1 to recognise ERVs, in-house generated UHRF1 mutants with

alterations to vital residues within the TTD/PHD domains that bind H3K9me3 were created. There was a failure to repress ERVs and a continued upregulation of the innate immune response observed. This was also observed *in vivo* in mice containing the same mutations, which died mid-gestation, suggesting the TTD-PHD domain in UHRF1 is required to bind to H3K9me3 on ERV, and as previous literature states (Rothbart et al., 2013, 2012), may be conserved through different species (Paper II). In addition, mice with a mutation of the H3K9me3 binding pocket seemed to show an increasing differential in methylation from WT as development proceeded (Paper II), suggesting a failure to gain methylation at TEs post implantation. This would be consistent with observations in ESC that mutants only became lethal on differentiation, so DNA methylation may be a long-term method of repression. This was also shown in our DNMT1 hypomorphs (Paper I). I also modelled the altered amino acid mutation in our mutant rescues using Swiss DeepView software (Guex et al., 1999). This software allows the user to create a 3D model of protein interactions via the amino acid sequence that can be downloaded from the RCSB PDB database (Berman, 2000). This allowed us to see the alterations that occurred following Alanine alterations and how these changes might affect the H3K9me3 binding pocket of the TTD-PHD domain.

Furthermore, SETDB1 null mouse embryonic fibroblasts, which were not deficient in DNA methylation, did not show IAP, MusD or LINE1 derepression (Matsui et al., 2010). In HeLa cells, Tie and colleagues (2018) discovered overexpression of ERVs and ZNF genes following KAP1 KO (these were ERVs that were also thought to be bound by KAP1 according to ENCODE data), with additional upregulation of ZNF genes. However, depletion of KAP1 in PBMC or primary adult cells did not result in activation of ISG, but these cells have intact methylation (Tie et al., 2018). Therefore, it may be possible that, in the absence of DNA methylation, these cells (Paper II, UHRF1 KD hTERT-1604 cells) revert to primarily histone-

mediated repression, similar to when DNA methylation is reprogrammed during embryonic development (Zeng and Chen, 2019). Again, further investigation is enquired to confirm this, including H3K9me3, H3K27me3, KAP1 and SETB1 ChIP-seq in WT, KD, rescue and mutants of our UHRF1 cell lines. As mentioned previously, methylation and transcriptional assessment via sequence-based techniques such as RRBS, to give greater coverage of the DNA methylation changes occurring in repeat elements and paired end total RNA-seq to enrich for repeat elements and account for ERV transcriptional bidirectionality.

Clinical potential of the use of UHRF1 could be similar to that of the demethylating agent Aza, in which demethylation encourages a viral response similar to that of when the body is attacked by a virus. This upregulation of the immune system may help to fight a cancerous phenotype possibly with greater efficacy in conjunction with existing therapeutics such as histone deacetylase inhibitors (Eckschlager et al., 2017).

## 7.2.2 DNMT1

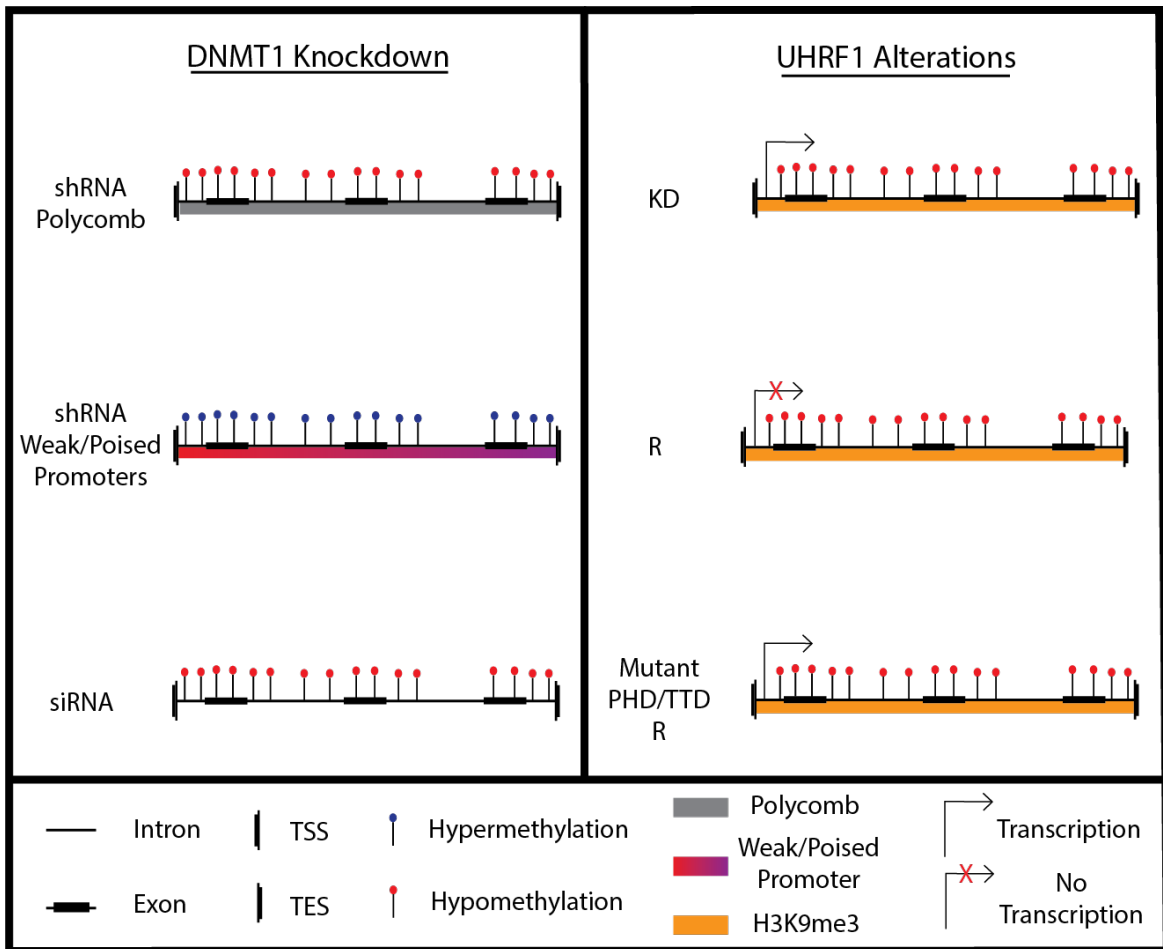
For DNMT1-depleted stable cell lines, the gains in methylation in particular in the cell lines were very puzzling when we began our analysis. It was to be expected that we should see losses in methylation, and indeed these occurred at the same positions in all the cell lines with overlap of the probes losing methylation in different shRNA lines, which was reassuring in that it suggested the effects were not stochastic but instead certain regions were more sensitive to loss. The gains were also overlapping between lines, and were harder to explain: did lowering DNMT1 levels somehow cause a loss of fidelity? Or was there an up-regulation of de novo activity, which was somehow found in new regions? In order to try and determine if there was any correlation between losses and gains, and any other feature of the genome, we mapped the 450K probes distribution using CandiMeth and compared gains/losses with a number of other tracks, including replication timing, CGI distribution and several ENCODE tracks (ENCODE Project Consortium, 2004). Of all the tracks examined, the

ChomHMM tracks (Ernst et al., 2011b) seemed to show some degree of correlation, with hypomethylated probes distributed across polycomb repressed and heterochromatic/low signal regions according to ENCODE chromatin state segmentation data (Ernst et al., 2011b). In order to quantitate this, chromatin state segmentation data was downloaded from UCSC (Paper I) – polycomb repressed targets were reactivated in comparison to the positive control, indicating they were subject to polycomb repression. Similarly, when Hill and colleagues (2018) conducted a triple knockout of all DNMTs in mouse germ cells they discovered inhibition of both polycomb repressive complexes and DNA methylation was required to mimic gonadal epigenetic reprogramming (Hill et al., 2018), this again matches with the effects of EZH2 inhibition in our DNMT1 KD cells (Paper I).

Brikman and colleagues (2012) showed a redistribution of H3K27me3 upon triple knockout of all DNMTs in mouse embryonic stem cells via ChIP-seq. They also elucidated that DNA methylation and H3K27me3 followed a negative correlation within CpG rich regions, but in CpG poor regions, similar levels of DNA methylation and H3K27me3 could co-exist (Brinkman et al., 2012). This could be due to the affinity of PRC2 for CpG-dense regions lacking DNA methylation (Laugesen et al., 2016), but proving this would require further investigation. It may also be of interest to conduct a similar ChIP-seq experiment to assess definitive evidence of redistribution of H3K27me3 between WT and DNMT1 KD in adult cells, in addition to repeating the shRNA KD with tighter time intervals sampled to see if the gains in methylation are due to long term culture or experimental intervention. When compared to shRNA-mediated KD of DNMT1, siRNA-mediated KD of DNMT1 displayed little difference between polycomb and non-polycomb regions in terms of demethylation (Paper I). After the siRNA KD was allowed to recover for 36 days, the polycomb-repression marked genes failed to remethylate, similar to what is observed in the shRNA KD at these polycomb regions (Paper I).

This failure to remethylate matches with current literature stating that after H3K27me3 deposition, PRC2 interacts with TET1 to inhibit methylation of the newly -repressed DNA (Neri et al., 2013). This inhibition of DNA methylation by TET1 and PRC2 may be why little transcriptional response is shown after the drop in methylation of these polycomb-repression marked genes. However, it could also be due to (at least in part) a lack of the correct transcription factors available to transcribe these previously methylation-repressed genes. Alternative studies have also pointed towards the possibility that these two processes (DNA methylation and polycomb repression) work in parallel (Li et al., 2018; Schlesinger et al., 2007; Viré et al., 2006; Widschwendter et al., 2007).

Interestingly, no gains in DNA methylation were observed in the siRNA KD: loci such as UGT1A showed loss of methylation upon siRNA treatment. Transcriptional levels of the de-novo methyltransferase DNMT3B were also assessed but did not differ from WT (Paper I). This would indicate that the gains observed in the shRNA KD were likely due to long-term culture. Also of interest, knockdown of DNMT3B resulted in loss of methylation at germline and PCDH genes, known targets of DNMT3B and DNMT1, indicating that in absence of DNMT1, DNMT3B may have a maintenance methylation role, as previously suggested within the literature (Chen et al., 2003; Elliott et al., 2016; Jones and Liang, 2009; Liang et al., 2002). However, gains in methylation did overlap strongly with weak and poised promoter chromatin segmentation categories, with slight accompanying changes in transcription. Looking forward, ChIP-seq enrichment for H3K4me2, H3K4me3 and H3K9ac would help to assess whether the gains in methylation observed in the shRNA KD were from the accumulation of de-novo methylation in long term culture or the redistribution of histone marks. An overview of the indications of alternative repressive mechanisms can be observed in figure 2.



**Figure 2: Indications of alternative repressive mechanisms in methylation deficient systems.**

Upon shRNA DNMT1 KD sites that were hypomethylated correlated most with those of polycomb repression (grey bar) from ENCODE chromatin state segmentation data (Ernst et al., 2011b). Whereas sites hypermethylated in shRNA DNMT1 KD correlated most with weak/poised promoters (red/purple bar). However, when a siRNA DNMT1 KD was completed, hypomethylated sites showed no preference to polycomb above any other chromatin state suggesting the correlation to polycomb repressed and weak/poised promoters was due to the long-term nature of the shRNA DNMT1 KD. In UHRF1 shRNA KD genome wide hypermethylation was observed with an increase in ERV transcription and upregulation of the innate immune response. This upregulation of ERV elements and the innate immune response was attenuated upon rescue of the KD with a full functioning copy of UHRF1 but genome wide hypomethylation was still present. Upon rescue of the UHRF1 KD cells with UHRF1 with a mutated PHD/TTD domain, similar results were observed as in the KD, with upregulation of ERV elements and the innate immune response. This indicates that H3K9me3 (orange bar) may be important in silencing ERVs in the absence of DNA methylation.



### 7.3 Differences in application of large-data analytics to human epidemiological rather than cell-line work (Paper III and VI)

Up to this point I have been discussing the application of bioinformatics approaches to cell line work, where effect sizes are large and technical repeats are tightly grouped. The cost-effectiveness and reproducibility of the array meant however that collaborating labs were very interested in applying this approach to epidemiological studies. We found that these come with a different set of parameters and limitations, effect sizes can be small and subjects often do not tightly group together because of differences in genetics and lifestyle choices i.e. age, BMI, alcohol intake, smoking, sex, ethnicity and any medication that they may be on. These differences in lifestyle choices and ethnicity can cause differences in the methylation results recorded. Therefore, these factors have to be normalised for, similar to normalising the type I and type II probes mentioned in the introduction (see section 1.6.2.1). There can also be differing levels of immune cells and other tissues in whole blood samples, these differences also must be checked and if they classify as potentially confounding for the DNA methylation results obtained, they need to be corrected for.

In order to check if any of these differences exhibit an effect on results, the data must first be explored and visualised. Principle Component Analysis (PCA) and Multi-dimensional Scaling (MDS) can provide a measure of dimensionality reduction to large scale epidemiological studies. Through PCA and MDS all samples can be plotted and will cluster via similar variances. If particular samples cluster independent of the intervention, this should be investigated further as it is evidence of a confounding factor.

The affected samples can either be removed i.e. like the 7 removed from the EPIFASSTT study, due to skewing the age distribution of participants, or they can be corrected for prior to differentiation analysis using a process known as surrogate variable analysis (SVA). SVA seeks to reduce the effect of confounders on the data collected and can also identify and

correct for hidden confounders in the data that may be from unaccounted for variation in results, such as batch effects in EPIC array results.

Alternative investigations, such as cell type correction should also be investigated if using whole blood as these samples can be composed of multiple different tissue types and therefore have the potential to effect results. Facilities for this assessment can be found in *Minfi* as well as *RnBeads* and *ChAMP* but *SVA* should normalise for any such hidden variable if present. I conducted cell type investigations via the *Minfi* package for the data presented in paper III to check for any skew in the distribution of tissue types from whole blood.

Fortunately, there was no skew in the distribution of cell types in our whole blood.

*RnBeads*, *ChAMP* and alternative packages also provide correction for copy number variation. Due to the literature surrounding SNPs in the LCE3B/3C genes in paper VI. I utilised the *ChAMP* and *DNAcopy* packages (Seshan VE, 2018) to investigate the possibility of a copy number variation (CNV) in the cohort utilised. After profiling each chromosome of each subject in paper VI no evidence of a CNV was obtained.

## 7.4 Effects of environment on methylation on current and future generations (Paper III and VI)

### 7.4.1 Folic acid supplementation in the second and third trimester causes alterations in DNA methylation upstream of a key imprint regulator

Approximately 40 years ago, an investigation was published detailing a link between foetal nutrition and coronary heart disease in the later life of the offspring. Many other similar investigations have thus been published detailing a similar link and resulting in the formation of the foetal origins of adult disease hypothesis (further detail in section 1.8.1.1). Folic acid is one of the limiting factors in one carbon metabolism as mentioned in section 1.8.1.2 and therefore DNA synthesis and DNA methylation. Currently, folic acid

supplementation is only recommended during the first trimester of pregnancy, when DNA synthesis and cell division is at its highest rate and to prevent neural tube closure defects. However, during the second and third trimesters the levels of DNA methylation and epigenetic reprogramming are at their highest, processes that depend on one carbon metabolism and as a result folic acid. Therefore, we wished to investigate the effects of folic acid supplementation in the second and third trimester on the resulting offspring (Paper III).

We found that maternal folic acid supplementation improves the folic acid status of the mother and offspring. Given this increased level of FA and the role it plays in methyl donor supply, it was hypothesized that methylation in the infant should increase, but this was not what we or others have observed (Amarasekera et al., 2014a; Caffrey et al., 2018). However, there are many other limiting factors to one carbon metabolism, including homocysteine concentration (Clare et al., 2019). The level of homocysteine has been observed as lower in mothers with the MTHFR mutation MTHFRC677T (Crider et al., 2011). In addition, mothers with this mutation have been associated with low folate status as this mutation results in the MTHFR enzyme having lower catalytic activity and subsequently lower production of methionine, a precursor to SAM – the universal one carbon donor which is essential for DNA methylation (Bagley and Selhub, 1998; Liu and Ward, 2010). Reports have also indicated high levels of SAM may modulate MTHFR and result in surplus folate being converted into thymidylate or purines (Crider et al., 2011). Reports indicating that excess folic acid may trigger a negative feedback mechanism via alterations to the SAM:SAH ratio (Christensen et al., 2015) would therefore fit with the lower genome-wide methylation observed in the treatment group (Paper III).

Additionally, we observed a region upstream of ZFP57 as the top hit in our DMR screen which appears to control ZFP57 expression (Paper III). This is of great interest as ZFP57 is an imprint regulator (Li et al., 2008) and a candidate gene analysis in the same cohort had already revealed altered methylation at certain imprinted loci (Caffrey et al., 2018). We could also see differences in methylation variability at imprints between treatment and placebo groups in the EPIFASSTT study (Paper III).

ZFP57 was originally discovered by Li and colleagues in 2008 as regulator of genomic imprinting in mice: upon insertion of a null allele of a functioning copy of ZFP57 resulted in full embryonic lethality respectively in mid gestation (Li et al., 2008). When homozygous embryos were compared to their heterozygous litter mates, DNA methylation was found to be severely affected at multiple imprints, including NNAT (the imprint most affected from the EPIFASSTT screen), which does demonstrate interaction with ZPF57 (Anvar et al., 2016)). Li and colleagues (2008) also found that loss of zygotic ZFP57 can impede the maintenance of DNA methylation even in the presence of maternal ZFP57. Conversely, in absence of maternal ZFP57, zygotic ZFP57 can rescue imprint maintenance. However, the methylation profile of the paternal chromosome cannot be maintained in the absence of both the maternal and zygotic forms of ZFP57, leading them to conclude that ZFP57 was a maternal-zygotic effect gene (Li et al., 2008).

ZFP57 protects genomic imprints from epigenetic reprogramming via forming a complex with KAP1 (Messerschmidt, 2012; Quenneville et al., 2011). This complex can then recognise its methylated consensus sequence [TG]GCCGC, which occurs at a high frequency at imprint control regions (Anvar et al., 2016; Liu et al., 2012; Quenneville et al., 2011). The KAP1 component of the ZFP57-KAP1 complex recruits SETDB1, the histone methyltransferase

mentioned in the previous section, in addition to the DNMT enzymes to maintain methylation at the imprint control regions of these imprints and also deposit H3K9me3 (Zuo et al., 2012). Mutations of ZFP57 have been associated with transient neonatal diabetes mellitus type 1, with marked hypomethylation found at imprinted regions (Baglivo et al., 2013; Mackay et al., 2008; Touati et al., 2019).

The FASSTT Trial (McNulty et al., 2013) was one of the first randomised control trials (RCT) of folic acid supplementation in the second and third trimester and the sequential EPIFASSTT trial poses the first evidence from an RCT that folic acid supplementation effects the methylation of the imprint regulator ZFP57. Before this RCT, an observational trial by Amarasekera 2014 had found lower methylation at this region upstream of ZFP57 (Amarasekera et al., 2014b). However, there were many differences between the EPIFASSTT RCT (Paper III) and this observational study (Amarasekera et al., 2014b) including, a smaller sample size (n=23), recruitment from an allergy clinic where 73% of subjects in the high folate test group had a history of allergenic disease, use 2 purified immune cell types (as opposed to using whole blood as in EPIFASSTT) and use of a 450k array as opposed to the improved EPIC array as within EPIFASSTT. Amarasekera and colleagues' independent and study subjects were also from the same source-cohort (Amarasekera et al., 2014b) which questions the verification of their results in their independent cohort.

To verify our methylation results from the EPIFASSTT RCT, I independently conducted the array analysis in *RnBeads*, analysed results and constructed a hierarchical linear model in *limma* to verify differentially methylated probes such as the region upstream of ZFP57. I also conducted analysis with and without SVA and tissue type correction to assess the differences in DMRs and to ensure array processing was reliable.

In Joubert et al 2016, five CpGs from our proposed folate sensitive DMR upstream of ZFP57 did exhibit differential methylation in one of the largest folate supplementation observational trials in the MoBa and Generation R cohorts, but the direction of change was not clarified (Joubert et al., 2016). Nevertheless, I was able to replicate our findings in an independent RCT, the AFAST study (Charles et al., 2005), but to a lesser effect size. This could be due to the greater age of the participants used in this study or the use of saliva instead of cord blood as a measure of DNA methylation.

Many other studies have noted a change in DNA methylation in response to folic acid supplementation or depletion. In the agouti mouse model, Waterland and colleagues (2010) discovered a change in the coat colour of mice offspring whose mothers had been supplemented with methyl-donor nutrients like folic acid (Waterland et al., 2010). Offspring of female sheep fed a methyl-donor constrained diet had a higher body mass, blood pressure, alterations to the immune system response and demonstrated signs of insulin resistance in comparison to controls (Sinclair et al., 2007). Haggarty (2013) discovered higher methylation at IGF2 and lower methylation at PEG3 following supplement use after week 12 of gestation (Haggarty et al., 2013). Similar changes in the methylation of IGF2 were also found in Steegers-Theunissen et al (2009). Following folic acid supplementation, children had 4.5% higher methylation at IGF2 in comparison to those without supplementation. Alterations to imprint methylation were also found in a multi-ethnic observational trial (Hoyo et al., 2014), including lower DNA methylation at NNAT as observed in the EPI-FASSTT trial (Paper III). IGF2 amongst others was also altered in the original candidate gene approach of the FASSTT trail (Caffrey et al., 2018). Interestingly, ZFP57 has been found to regulate the methylation of the above-mentioned imprinting genes (Anvar et al., 2016; Takahashi et al., 2019).

Moreover, the effects of folic acid supplementation extend beyond imprints. In a study of rural Gambian women, variation in seasonal nutrient intake surrounding the period of conception caused alterations in the DNA methylation of 13 plasma biomarkers (Dominguez-Salas et al., 2014). Sex specific changes in the DNA methylation of the HSD11B2 gene, which controls foetal exposure to glucocorticoids and has been linked to low birth weight and were observed upon folate acid supplementation in rats (Zhao et al., 2014). Folic acid supplementation caused decreased methylation at this gene in only males but caused increased birth weight in females. In addition, folic acid supplementation has been implicated in lowering the risk of stroke and other cardiovascular diseases (Li et al., 2016).

Conversely, in a folate-replete population, maternal intake of folic acid and other methyl-donors periconceptionally or during the second trimester was not associated with alterations to LINE1 methylation (Boeke et al., 2012). This may indicate again a potential feedback mechanism, directing one carbon metabolites to alternative tasks such as purine or pyrimidine synthesis. Nevertheless, further studies would be needed to clarify this.

ZFP57 has also been implicated as a transcriptional repressor in Schwann cells of the peripheral nervous system (Alonso et al., 2004). Interestingly, children of the mothers in the folic acid supplementation group within the EPI-FASSTT study, have been found to have increased emotional intelligence, resilience and psychosocial benefits (Henry et al., 2018). In addition to improved cognitive performance (McNulty et al., 2019). Further work should be done to investigate the methylation of neural genes, via CandiMeth and a network style approach like Ingenuity Pathway Analysis within the treatment cohort to assess further alterations from supplementation of this methyl-donor. It may also be of interest to look at

the methylation of these neural genes and ZFP57 in the publicly available methylation data of NTD investigations or those with anencephaly.

Of note, there appears to be an association between the stage of gestation and the effects of exposure to folic acid supplementation/restriction of methyl-donors, intergenerationally and transgenerationally. Heijmans and colleagues (2008) discovered that those exposed to famine in the early stages of gestation exhibited hypomethylation at the IGF2 gene decades after exposure. However, their same-sex siblings exposed in late gestation did not show any alterations in DNA methylation at the same region – indicative of intergenerational effects of maternal nutrition during gestation (Heijmans et al., 2008). Further studies on the agouti mouse model mentioned earlier, demonstrated that the DNA methylation changes that occurred from folic acid supplementation in the previous generation remained in offspring not exposed to any supplements (Cropley et al., 2006) – example of intergenerational effects of maternal nutrition during gestation. Additionally, follow-up work to the AFAST study found alterations to DNA methylation 47 years after the folic acid supplementation trial had taken place (Richmond et al., 2018), an example of transgenerational effects of maternal nutrition during gestation.

#### 7.4.2 Alterations to the DNA methylation of immune response genes in sufferers of Depression

The other major epidemiological study I was involved with was regarding potential epigenetic markers of mental health. During a WHO World Mental Health International College Student project a high prevalence of depression was discovered in university students within the United Kingdom (Auerbach et al., 2018). Within Northern Ireland in particular, the prevalence was especially high (O'Neill et al., 2018). The NI data was derived from a student cohort at Ulster: further analysis of this cohort revealed a sub-set of these students, those



with depression, self-harm and a suicide attempt, had significantly higher shared risk factors such as being female and non-heterosexual (Paper IV). Given the higher risk score of this sub-group and the links between depression and immune dysfunction (Robson et al., 2017), it was hypothesized that this group may show epigenetic differences in comparison with healthy controls. Results showed epigenetic dysregulation at multiple immune-related genes and a link to psoriasis – an immune condition which has also been linked with depression (Pariser et al., 2016). To obtain these results I conducted our array analysis independently in different package such as ChAMP and RnBeads. GO of array results indicated upregulation of the immune system response in top ranking promoters. From this GO result, I used CandiMeth to assess the methylation of these immune system response genes and found a significant difference in the methylation of these genes between cases and controls. We also had difficulty with tissue type correction as no reference set existed for saliva samples (the tissue type taken during this investigation). However, after conducting SVA, exploring the data via PCA and checking recent literature we came to the conclusion that SVA should remove variation from differences in tissue type composition (Teschendorff and Relton, 2018).

In a global burden of disease study in 2016, it was revealed that over 168 million people suffer from depression (Vos et al., 2017). Depression is a complex disorder characterised by multiple physical and neurological symptoms including; fatigue, social withdrawal, anhedonia, low mood and insomnia (Rahim and Rashid, 2017). As mentioned, depression has been linked to immune dysregulation (Robson et al., 2017). This theory was proposed over 20 years ago (Maes et al., 1991) and multiple studies have supported this theory since then (Ménard et al., 2016; Miller and Raison, 2016; Won and Kim, 2016). Depression has also been linked with many other chronic immune inflammatory disorders including,

cardiovascular disease (Halaris, 2017), diabetes (Anderson et al., 2001), asthma (Zielinski et al., 2000) and psoriasis (Patel et al., 2017), the latter being the most commonly linked to depression (Patel et al., 2017).

Interestingly, we found alterations to the methylation of the Late Cornified Envelope (LCE) cluster within our study which have also been linked to psoriasis (Hüffmeier et al., 2010).

This cluster is located on chromosome 1 and is part of the Epidermal Differentiation Complex (EDC), a 2 Mb region responsible for growth, repair and terminal differentiation of keratinocytes which will ultimately aid the cornified envelope of the skin (Abhishek and Krishnan, 2016). It is under epigenetic regulation by histone modification (Luis et al., 2011) and regulation by microRNA (Kretz et al., 2013). Dysregulation of this complex has been associated with psoriasis in addition to other dermatological disorders (Abhishek and Krishnan, 2016). The LCE cluster is divided into 6 groups and deletion of the LCE3B and LCE3C genes was found to be significantly associated with psoriasis, in a genome wide CNV analysis of 2831 European and American subjects (De Cid et al., 2009). An alternative GWAS study noted the same association between LCE3B and LCE3C deletion and psoriasis.

Further work into the expression of the LCE cluster in normal versus psoriatic skin discovered that LCE3B and LCE3C exhibited negligible expression in normal skin but were significantly upregulated in psoriatic skin (Bergboer et al., 2011), this was also noted in an alternative study (Guttman-Yassky et al., 2009). Additionally, expression of the group 1, 2, 5 and 6 LCE group genes were significantly downregulated in psoriatic skin compared to normal controls (Bergboer et al., 2011). A meta-analysis of psoriatic patients also revealed that 90% of sufferers exhibited heterozygous or homozygous deletions of the LCE3B and LCE3C genes (Riveira-Munoz et al., 2011). Following these investigations, it was suspected

that LCE3 genes encoded skin barrier repair proteins and absence of these genes lead to discrepancies in skin barrier function and susceptibility to psoriasis (Bergboer et al., 2011; Riveira-Munoz et al., 2011). An assessment of CNVs was conducted in our cohort but no evidence of a deletion at this region resulted (Paper IV). This was conducted as mentioned in section 7.3.

Moreover, a psoriasis susceptibility loci PSORS4 has also been found in the EDC on chromosome 1q21, close to the LCE genes, in Chinese (Chen et al., 2009), European (Liu et al., 2008) and Italian cohorts (Capon et al., 1999). In addition, interactions have been found between the LCE cluster on chromosome 1q21 and PSORS1 on chromosome 6q21 (Capon et al., 1999; De Cid et al., 2009; Hüffmeier et al., 2010). A further meta-analysis of multiple cohorts found that this latter interaction was only present in Dutch and American cohorts but not in Italian, Japanese or Mongolian cohorts (Riveira-Munoz et al., 2011). This meta-analysis also revealed that the LCE3C-LCE3B and PSORS1 interaction was dependent on the presence of the HLA-cw06 allele (Riveira-Munoz et al., 2011). This allele was a part of the PSORS1 susceptibility loci on chromosome 6 and involved in the major histocompatibility complex (Sagoo et al., 2004), a key part of immune system responses (Wieczorek et al., 2017).

Alternative top hits from this study include the MIR4520A/B locus which encodes two microRNA with the same names as their respective genes (Timis and Orasan, 2018). Not much is known about the regulation of this locus, but it has been identified as a top hit in a psoriasis study of small RNAs (Joyce et al., 2011). As mentioned previously, miRNAs may regulate the EDC (Kretz et al., 2013; Timis and Orasan, 2018) and that MIR4520A/B have also been identified as top hits in a psoriasis study (Joyce et al., 2011), it may be of interest to

examine the function of MIR4520A/B and their regulation to see whether they have a regulatory role in the EDC or if they are/contribute to a psoriasis susceptibility loci. To facilitate this, current publications should be checked to examine the known effectors of MIR4520A and MIR4520B. Following this, epigenome wide and genome wide association studies should be analysed for the expression/methylation of MIR4520A/B upon different interventions, before interventions to investigate the link between MIR4520A/B and any psoriasis susceptibility loci can begin.

DEFB104B was also identified as a top hit in our study. It encodes a  $\beta$ -defensin protein, a category of antimicrobial peptides involved in immune system responses (Premratanachai et al., 2004). Certain  $\beta$ -defensin genes have also been identified as tumour suppressor genes and their epigenetic dysregulation is associated with many chronic diseases including cancer (Xu et al., 2016), depression (Liu et al., 2018) and psoriasis (Stuart et al., 2012). CNV of  $\beta$ -defensins on chromosome 8 (location of DEFB104B) has also been identified as a risk factor in psoriasis (Hollox et al., 2008; Stuart et al., 2012).

A direct split between the male and female subjects on the basis of methylation was also discovered in our study. This could be because of the different levels of HPA-axis activation initiated in different sexes following stressful events (Muscatell et al., 2015) and why the female sex has been identified as a potential risk factor for developing depression. A recent review of the literature into sex-specific differences in the development of depression suggested that women have a more pro-inflammatory response to stressors with decreased sensitivity to glucocorticoids. This upregulation of glucocorticoids and inflammatory cytokines can then cause alterations to neurotransmission and synaptic plasticity (Bekhbat and Neigh, 2018).

Although, the current study does possess limitations, such as the small sample size. It

nevertheless acts as an informative pilot into the link between psoriasis and depression in those of university age, confirms saliva as an appropriate tissue for DNA methylation assessment in mental health conditions. It would be useful to obtain, in future, the genotype of the subjects in relation to the LCE3C/LCE3B-del and possibly other genotypes such as CNV of  $\beta$ -defensin which has also been identified as a risk-loci in psoriasis (Hollox et al., 2008; Stuart et al., 2012). In addition to expanding this pilot study into the DNA methylation of depressed vs healthy students. Since it is suspected that the LCE cluster is under regulation by DNA methylation, yet little evidence exists to support this. A greater sample size would also provide this investigation with greater statistical power and as a result increase the probability that this study is representative of DNA methylation in depressed students compared to healthy controls. Transcriptional data, such as RNA-seq on the samples in this study may also aid the investigation into whether the LCE cluster is regulated by DNA methylation.

### 7.5 The development of CandiMeth and possible future versions

The cost effectiveness of Illumina methylation arrays has inspired multiple epidemiological studies to investigate their results at a molecular scale. Examples include the data presented in paper IV. However, after preliminary analysis of array data and subsequent GO analysis, it can be difficult for those without much bioinformatics experience to utilise their array results for further investigation. Platforms such as Galaxy ([www.usegalaxy.org](http://www.usegalaxy.org)) have aided in this cause but matters that would be simple to bioinformaticians such as candidate gene investigation remain difficult to those not used with the command line or high dimensional data structures, such as those output from methylation array analysis with RnBeads or many other packages. To this end we wished to plug this gap in current software provisions via creating an online workflow in the Galaxy interface.

This workflow would allow users to easily investigate candidate genes and features from the results of multiple epigenome wide association studies (Paper V). The workflow, later called CandiMeth, allows users to overlap their methylation array results with popular UCSC genome browser tracks like RepeatMasker and hopefully in the future chromatin state segmentation tracks or CpG islands. In addition to, providing approximations for the promoter and gene bodies of all Reference Sequence defined genes in the hg19 and hg38 human genome assemblies.

CandiMeth was modelled around RnBeads primarily as it was the R package used within the Walsh lab to analyse methylation arrays, but is now compatible with ChAMP outputs as well as the option to upload custom outputs as long as they match the data format described in the CandiMeth paper. CandiMeth was primarily targeted at pipeline-based R-packages as these usually will produce outputs such as differential methylation tables without bespoke coding and can therefore be input with ease into Galaxy and therefore CandiMeth. This allowed the wet-lab biologists of the Walsh lab to investigate for example, all 4 results files for the DNMT1 KD (Paper I) or UHRF1 KD and R (Paper II) via creating a one column text file in the Galaxy interface of the features they wanted to investigate and selecting this file from the drop-down menus of the CandiMeth workflow. CandiMeth has no installation process required to make this process as simple as possible. Furthermore, CandiMeth also provides users with UCSC compatible tracks which when viewed within UCSC genome browser show the absolute methylation of WT and experimental results (also known as absolute beta tracks), the difference between WT and experimental results (delta beta tracks) and tracks which show the absolute methylation of only those probes that satisfy an FDR criteria of  $q < 0.05$ . This allows users to view the results of their intervention at a gene/cluster of interest to see if the difference at that site is worthy of further investigation and provides seamless

integration with BLAT for primer design, in the case of loci of interest. A comprehensive guide and GitHub repository were also created with step-by-step instructions on how to use CandiMeth via Galaxy with test data and using the user's own data, in addition to how to integrate those results with the RepeatMasker track. CandiMeth and the accompanying guide therefore providing a simple and easy to use process which improves result reproducibility and gives those with little bioinformatics training more control over their large-scale data.

Future amendments to CandiMeth include, integration with further R-packages (like Minfi), UCSC tracks i.e. chromatin state segmentation or CpG island tracks and potentially expanding CandiMeth to include different types of input files, such as RNA-seq. Candimeth could also be improved to facilitate downstream analysis from new analysis suites which take machine learning or multi-omics approaches to differential analysis, such pipelines include Bigmelon (Gorrie-Stone et al., 2019) or PyMethylProcess (Levy et al., 2019) amongst others (Heiss and Just, 2019; Prelot et al., 2018; Song et al., 2019; Wang et al., 2019).

## 7.6 Concluding Remarks

Through bioinformatics approaches I have investigated in this thesis the effects of DNA methylation depletion in normal adult cells with implications for potential repressive mechanisms upon depletion of DNA methylation (polycomb repression), added to the knowledge base regarding a maintenance methyltransferase function for DNMT3B and provided inferences on a potential epi-therapeutic pathway in adult cells following depletion of UHRF1. Using human intervention studies, I have determined effects on the offspring of folic acid supplementation in the second and third trimester, including a key effect on an imprint regulator and were able to replicate the effect in cell-line based models – adding to the growing evidence of a beneficial effect of folic acid supplementation during

late gestation and rationale for investigation into the effects later in the life of the child and on the mother. I have also added to the evidence of an immune component to the aetiology of depression and provided an informative basis for further study to elucidate potential genetic or epigenetic predispositions to the mentally ill state of the cases in comparison to the control subjects. I have also provided a user-friendly web-based workflow known as CandiMeth to allow those with little bioinformatics knowledge to investigate features of interest from the results of epigenome wide methylation arrays.



## 7.7 Bibliography

- Abhishek, S., Krishnan, S.P., 2016. Epidermal differentiation complex: A review on its epigenetic regulation and potential drug targets. *Cell J.*
- Adair, S.J., Hogan, K.T., 2009. Treatment of ovarian cancer cell lines with 5-aza-2'-deoxycytidine upregulates the expression of cancer-testis antigens and class I major histocompatibility complex-encoded molecules. *Cancer Immunol. Immunother.* 58, 589–601.
- Alonso, M.B.D., Zoidl, G., Taveggia, C., Bosse, F., Zoidl, C., Rahman, M., Parmantier, E., Dean, C.H., Harris, B.S., Wrabetz, L., Müller, H.W., Jessen, K.R., Mirsky, R., 2004. Identification and characterization of ZFP-57, a novel zinc finger transcription factor in the mammalian peripheral nervous system. *J. Biol. Chem.* 279, 25653–25664.
- Amarasekera, M., Martino, D., Ashley, S., Harb, H., Kesper, D., Strickland, D., Saffery, R., Prescott, S.L., 2014a. Genome-wide DNA methylation profiling identifies a folate-sensitive region of differential methylation upstream of ZFP57-imprinting regulator in humans. *FASEB J.* 28, 4068–4076.
- Amarasekera, M., Martino, D., Ashley, S., Harb, H., Kesper, D., Strickland, D., Saffery, R., Prescott, S.L., 2014b. Genome-wide DNA methylation profiling identifies a folate-sensitive region of differential methylation upstream of ZFP57-imprinting regulator in humans. *FASEB J.* 28, 4068–76.
- Anderson, R.J., Freedland, K.E., Clouse, R.E., Lustman, P.J., 2001. The prevalence of comorbid depression in adults with diabetes: A meta-analysis. *Diabetes Care* 24, 1069–1078.
- Antunes, G., Simoes de Souza, F.M., 2016. Olfactory receptor signaling. *Methods Cell Biol.* 132, 127–145.
- Anvar, Z., Cammisa, M., Riso, V., Baglivo, I., Kukreja, H., Sparago, A., Girardot, M., Lad, S., Feis, I. De, Cerrato, F., Angelini, C., Feil, R., Pedone, P. V., Grimaldi, G., Riccio, A., 2016. ZFP57 recognizes multiple and closely spaced sequence motif variants to maintain repressive epigenetic marks in mouse embryonic stem cells. *Nucleic Acids Res.* 44, 1118–1132.
- Auerbach, R.P., Mortier, P., Bruffaerts, R., Alonso, J., Benjet, C., Cuijpers, P., Demyttenaere, K., Ebert, D.D., Green, J.G., Hasking, P., Murray, E., Nock, M.K., Pinder-Amaker, S., Sampson, N.A., Stein, D.J., Vilagut, G., Zaslavsky, A.M., Kessler, R.C., 2018. WHO world mental health surveys international college student project: Prevalence and distribution of mental disorders. *J. Abnorm. Psychol.* 127, 623–638.
- Bagley, P.J., Selhub, J., 1998. A common mutation in the methylenetetrahydrofolate reductase gene is associated with an accumulation of formylated tetrahydrofolates in red blood cells. *Proc. Natl. Acad. Sci. U. S. A.* 95, 13217–20.
- Baglivo, I., Esposito, S., De C, L., Sparago, A., Anvar, Z., Riso, V., Cammisa, M., Fattorusso, R., Grimaldi, G., Riccio, A., Pedone, P. V., 2013. Genetic and epigenetic mutations affect the DNA binding capability of human ZFP57 in transient neonatal diabetes type 1. *FEBS Lett.* 587, 1474–1481.

- Bekhbat, M., Neigh, G.N., 2018. Sex differences in the neuro-immune consequences of stress: Focus on depression and anxiety. *Brain. Behav. Immun.*
- Bergboer, J.G.M., Tjabringa, G.S., Kamsteeg, M., Van Vlijmen-Willems, I.M.J.J., Rodijk-Olthuis, D., Jansen, P.A.M., Thuret, J.Y., Narita, M., Ishida-Yamamoto, A., Zeeuwen, P.L.J.M., Schalkwijk, J., 2011. Psoriasis risk genes of the late cornified envelope-3 group are distinctly expressed compared with genes of other LCE groups. *Am. J. Pathol.* 178, 1470–1477.
- Berman, H.M., 2000. The Protein Data Bank / Biopython. *Presentation* 28, 235–242.
- Boeke, C.E., Baccarelli, A., Kleinman, K.P., Burris, H.H., Litonjua, A.A., Rifas-Shiman, S.L., Tarantini, L., Gillman, M.W., 2012. Gestational intake of methyl donors and global LINE-1 DNA methylation in maternal and cord blood: Prospective results from a folate-replete population. *Epigenetics* 7, 253–260.
- Bork, S., Pfister, S., Witt, H., Horn, P., Korn, B., Ho, A.D., Wagner, W., 2010. DNA methylation pattern changes upon long-term culture and aging of human mesenchymal stromal cells. *Aging Cell* 9, 54–63.
- Brinkman, A.B., Gu, H., Bartels, S.J.J., Zhang, Y., Matarese, F., Simmer, F., Marks, H., Bock, C., Gnirke, A., Meissner, A., Stunnenberg, H.G., 2012. Sequential ChIP-bisulfite sequencing enables direct genome-scale investigation of chromatin and DNA methylation cross-talk. *Genome Res.* 22, 1128–38.
- Brown, K.D., Robertson, K.D., 2007. DNMT1 knockout delivers a strong blow to genome stability and cell viability. *Nat. Genet.* 39, 289–290.
- Bulut-Karslioglu, A., DeLaRosa-Velázquez, I.A., Ramirez, F., Barenboim, M., Onishi-Seebacher, M., Arand, J., Galán, C., Winter, G.E., Engist, B., Gerle, B., O'Sullivan, R.J., Martens, J.H.A., Walter, J., Manke, T., Lachner, M., Jenuwein, T., 2014. Suv39h-Dependent H3K9me3 Marks Intact Retrotransposons and Silences LINE Elements in Mouse Embryonic Stem Cells. *Mol. Cell* 55, 277–290.
- Caffrey, A., Irwin, R.E., McNulty, H., Strain, J.J., Lees-Murdock, D.J., McNulty, B.A., Ward, M., Walsh, C.P., Pentieva, K., 2018. Gene-specific DNA methylation in newborns in response to folic acid supplementation during the second and third trimesters of pregnancy: Epigenetic analysis from a randomized controlled trial. *Am. J. Clin. Nutr.* 107, 566–575.
- Cai, Y., Tsai, H.-C., Yen, R.-W.C., Zhang, Y.W., Kong, X., Wang, W., Xia, L., Baylin, S.B., 2017a. Critical threshold levels of DNA methyltransferase 1 are required to maintain DNA methylation across the genome in human cancer cells. *Genome Res.* 27, 533–544.
- Cai, Y., Tsai, H.C., Yen, R.W.C., Zhang, Y.W., Kong, X., Wang, W., Xia, L., Baylin, S.B., 2017b. Critical threshold levels of DNA methyltransferase 1 are required to maintain DNA methylation across the genome in human cancer cells. *Genome Res.* 27, 533–544.
- Canzio, D., Maniatis, T., 2019. The generation of a protocadherin cell-surface recognition code for neural circuit assembly. *Curr. Opin. Neurobiol.*
- Capon, F., Novelli, G., Semprini, S., Clementi, M., Nudo, M., Vultaggio, P., Mazzanti, C., Gobello, T., Botta, A., Fabrizi, G., Dallapiccola, B., 1999. Searching for psoriasis

- susceptibility genes in Italy: Genome scan and evidence for a new locus on chromosome 1. *J. Invest. Dermatol.* 112, 32–35.
- Charles, D.H.M., Ness, A.R., Campbell, D., Smith, G.D., Whitley, E., Hall, M.H., 2005. Folic acid supplements in pregnancy and birth outcome: Re-analysis of a large randomised controlled trial and update of Cochrane review. *Paediatr. Perinat. Epidemiol.*
- Chen, H., Toh, T.K.L., Szeverenyi, I., Ong, R.T.H., Theng, C.T.S., McLean, W.H.I., Seielstad, M., Lane, E.B., 2009. Association of skin barrier genes within the PSORS4 locus is enriched in Singaporean Chinese with early-onset psoriasis. *J. Invest. Dermatol.* 129, 606–614.
- Chen, T., Ueda, Y., Dodge, J.E., Wang, Z., Li, E., 2003. Establishment and maintenance of genomic methylation patterns in mouse embryonic stem cells by Dnmt3a and Dnmt3b. *Mol. Cell. Biol.* 23, 5594–605.
- Chernyavskaya, Y., Mudbhary, R., Zhang, C., Tokarz, D., Jacob, V., Gopinath, S., Sun, X., Wang, S., Magnani, E., Madakashira, B.P., Yoder, J.A., Hoshida, Y., Sadler, K.C., 2017. Loss of dna methylation in zebrafish embryos activates retrotransposons to trigger antiviral signaling. *Dev.* 144, 2925–2939.
- Chiappinelli, K.B., Strissel, P.L., Desrichard, A., Li, H., Henke, C., Akman, B., Hein, A., Rote, N.S., Cope, L.M., Snyder, A., Makarov, V., Buhu, S., Slamon, D.J., Wolchok, J.D., Pardoll, D.M., Beckmann, M.W., Zahnow, C.A., Mergoub, T., Chan, T.A., Baylin, S.B., Strick, R., 2015. Inhibiting DNA Methylation Causes an Interferon Response in Cancer via dsRNA Including Endogenous Retroviruses. *Cell* 162, 974–986.
- Christensen, K.E., Mikkelsen, L.G., Leung, K.Y., Lévesque, N., Deng, L., Wu, Q., Malysheva, O. V., Best, A., Caudill, M.A., Greene, N.D.E., Rozen, R., 2015. High folic acid consumption leads to pseudo-MTHFR deficiency, altered lipid metabolism, and liver injury in mice. *Am. J. Clin. Nutr.* 101, 646–658.
- Clare, C.E., Brassington, A.H., Kwong, W.Y., Sinclair, K.D., 2019. One-Carbon Metabolism: Linking Nutritional Biochemistry to Epigenetic Programming of Long-Term Development. *Annu. Rev. Anim. Biosci.* 7, 263–287.
- Collins, A.M., Watson, C.T., 2018. Immunoglobulin light chain gene rearrangements, receptor editing and the development of a self-tolerant antibody repertoire. *Front. Immunol.*
- Crider, K.S., Zhu, J.-H., Hao, L., Yang, Q.-H., Yang, T.P., Gindler, J., Maneval, D.R., Quinlivan, E.P., Li, Z., Bailey, L.B., Berry, R.J., 2011. MTHFR 677C>T genotype is associated with folate and homocysteine concentrations in a large, population-based, double-blind trial of folic acid supplementation. *Am. J. Clin. Nutr.* 93, 1365–1372.
- Cropley, J.E., Suter, C.M., Beckman, K.B., Martin, D.I.K., 2006. Germ-line epigenetic modification of the murine Avy allele by nutritional supplementation. *Proc. Natl. Acad. Sci. U. S. A.* 103, 17308–17312.
- Cuellar, L., Herzner, A.M., Zhang, X., Goyal, Y., Watanabe, C., Friedman, B.A., Janakiraman, V., Durinck, S., Stinson, J., Arnott, D., Cheung, T.K., Chaudhuri, S., Modrusan, Z., Doerr, J.M., Classon, M., Haley, B., 2017. Silencing of retrotransposons by SET DB1 inhibits the interferon response in acute myeloid leukemia. *J. Cell Biol.* 216, 3535–3549.

- Dallosso, A.R., Hancock, A.L., Szemes, M., Moorwood, K., Chilukamarri, L., Tsai, H.-H., Sarkar, A., Barasch, J., Vuononvirta, R., Jones, C., Pritchard-Jones, K., Royer-Pokora, B., Lee, S.B., Owen, C., Malik, S., Feng, Y., Frank, M., Ward, A., Brown, K.W., Malik, K., 2009. Frequent Long-Range Epigenetic Silencing of Protocadherin Gene Clusters on Chromosome 5q31 in Wilms' Tumor. *PLoS Genet.* 5.
- De Cid, R., Riveira-Munoz, E., Zeeuwen, P.L.J.M., Robarge, J., Liao, W., Dannhauser, E.N., Giardina, E., Stuart, P.E., Nair, R., Helms, C., Escaramís, G., Ballana, E., Martín-Ezquerria, G., Heijer, M. Den, Kamsteeg, M., Joosten, I., Eichler, E.E., Lázaro, C., Pujol, R.M., Armengol, L., Abecasis, G., Elder, J.T., Novelli, G., Armour, J.A.L., Kwok, P.Y., Bowcock, A., Schalkwijk, J., Estivill, X., 2009. Deletion of the late cornified envelope LCE3B and LCE3C genes as a susceptibility factor for psoriasis. *Nat. Genet.* 41, 211–215.
- De Plaen, E., Traversari, C., Gaforio, J.J., Szikora, J.P., De Smet, C., Brasseur, F., van der Bruggen, P., Lethé, B., Lurquin, C., Chomez, P., De Backer, O., Boon, T., Arden, K., Cavenee, W., Brasseur, R., 1994. Structure, chromosomal localization, and expression of 12 genes of the MAGE family. *Immunogenetics* 40, 360–369.
- Deaton, A.M., Bird, A., 2011. CpG islands and the regulation of transcription. *Genes Dev.* 25, 1010–22.
- Dominguez-Salas, P., Moore, S.E., Baker, M.S., Bergen, A.W., Cox, S.E., Dyer, R.A., Fulford, A.J., Guan, Y., Laritsky, E., Silver, M.J., Swan, G.E., Zeisel, S.H., Innis, S.M., Waterland, R.A., Prentice, A.M., Hennig, B.J., 2014. Maternal nutrition at conception modulates DNA methylation of human metastable epialleles. *Nat. Commun.* 5.
- Dong, J., Wang, X., Cao, C., Wen, Y., Sakashita, A., Chen, S., Zhang, J., Zhang, Y., Zhou, L., Luo, M., Liu, M., Liao, A., Namekawa, S.H., Yuan, S., 2019. UHRF1 suppresses retrotransposons and cooperates with PRMT5 and PIWI proteins in male germ cells. *Nat. Commun.* 10.
- Eckschlager, T., Plch, J., Stiborova, M., Hrabeta, J., 2017. Histone deacetylase inhibitors as anticancer drugs. *Int. J. Mol. Sci.*
- Ehrlich, M., Gama-Sosa, M.A., Huang, L.H., Midgett, R.M., Kuo, K.C., Mccune, R.A., Gehrke, C., 1982. Amount and distribution of 5-methylcytosine in human DNA from different types of tissues or cells. *Nucleic Acids Res.* 10, 2709–2721.
- El Hajj, N., Dittrich, M., Haaf, T., 2017. Epigenetic dysregulation of protocadherins in human disease. *Semin. Cell Dev. Biol.*
- Elliott, E.N., Sheaffer, K.L., Kaestner, K.H., 2016. The 'de novo' DNA methyltransferase Dnmt3b compensates the Dnmt1-deficient intestinal epithelium. *Elife* 5.
- ENCODE Project Consortium, 2004. The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science* (80-81). 306, 636–640.
- Ernst, J., Kheradpour, P., Mikkelsen, T.S., Shores, N., Ward, L.D., Epstein, C.B., Zhang, X., Wang, L., Issner, R., Coyne, M., Ku, M., Durham, T., Kellis, M., Bernstein, B.E., 2011a. Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature* 473, 43–49.
- Ernst, J., Kheradpour, P., Mikkelsen, T.S., Shores, N., Ward, L.D., Epstein, C.B., Zhang, X., Wang, L., Issner, R., Coyne, M., Ku, M., Durham, T., Kellis, M., Bernstein, B.E., 2011b.

- Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature* 473, 43–49.
- Ferry, L., Fournier, A., Tsusaka, T., Adelmant, G., Shimazu, T., Matano, S., Kirsh, O., Amouroux, R., Dohmae, N., Suzuki, T., Fillion, G.J., Deng, W., de Dieuleveult, M., Fritsch, L., Kudithipudi, S., Jeltsch, A., Leonhardt, H., Hajkova, P., Marto, J.A., Arita, K., Shinkai, Y., Defossez, P.A., 2017. Methylation of DNA Ligase 1 by G9a/GLP Recruits UHRF1 to Replicating DNA and Regulates DNA Methylation. *Mol. Cell* 67, 550–565.e5.
- Golan-Mashiach, M., Grunspan, M., Emmanuel, R., Gibbs-Bar, L., Dikstein, R., Shapiro, E., 2012. Identification of CTCF as a master regulator of the clustered protocadherin genes. *Nucleic Acids Res.* 40, 3378–3391.
- Gordeeva, O., 2018. Cancer-testis antigens: Unique cancer stem cell biomarkers and targets for cancer therapy. *Semin. Cancer Biol.*
- Gordon, K., Clouaire, T., Bao, X.X., Kemp, S.E., Xenophontos, M., De Las Heras, J.I., Stancheva, I., 2014. Immortality, but not oncogenic transformation, of primary human cells leads to epigenetic reprogramming of DNA methylation and gene expression. *Nucleic Acids Res.* 42, 3529–3541.
- Gorrie-Stone, T.J., Smart, M.C., Saffari, A., Malki, K., Hannon, E., Burrage, J., Mill, J., Kumari, M., Schalkwyk, L.C., 2019. Bigmelon: tools for analysing large DNA methylation datasets. *Bioinformatics* 6, 981–986.
- Guex, N., Diemand, A., Peitsch, M.C., 1999. Protein modelling for all. *Trends Biochem. Sci.* 24, 364–367.
- Güre, A.O., Wei, I.J., Old, L.J., Chen, Y.T., 2002. The SSX gene family: Characterization of 9 complete genes. *Int. J. Cancer* 101, 448–453.
- Guttman-Yassky, E., Suárez-Fariñas, M., Chiricozzi, A., Nograles, K.E., Shemer, A., Fuentes-Duculan, J., Cardinale, I., Lin, P., Bergman, R., Bowcock, A.M., Krueger, J.G., 2009. Broad defects in epidermal cornification in atopic dermatitis identified through genomic analysis. *J. Allergy Clin. Immunol.* 124.
- Habano, W., Kawamura, K., Iizuka, N., Terashima, J., Sugai, T., Ozawa, S., 2015. Analysis of DNA methylation landscape reveals the roles of DNA methylation in the regulation of drug metabolizing enzymes. *Clin. Epigenetics* 7, 105.
- Haggarty, P., Hoad, G., Campbell, D.M., Horgan, G.W., Piyathilake, C., McNeill, G., 2013. Folate in pregnancy and imprinted gene and repeat element methylation in the offspring. *Am. J. Clin. Nutr.* 97, 94–99.
- Hagiwara, Y., Sieverling, L., Hanif, F., Anton, J., Dickinson, E.R., Bui, T.T.T., Andreeva, A., Barran, P.E., Cota, E., Nikolova, P. V., 2016. Consequences of point mutations in melanoma-associated antigen 4 (MAGE-A4) protein: Insights from structural and biophysical studies. *Sci. Rep.* 6.
- Halaris, A., 2017. Inflammation-associated co-morbidity between depression and cardiovascular disease. In: *Current Topics in Behavioral Neurosciences*. Springer Verlag.
- Heijmans, B.T., Tobi, E.W., Stein, A.D., Putter, H., Blauw, G.J., Susser, E.S., Slagboom, P.E., Lumey, L.H., 2008. Persistent epigenetic differences associated with prenatal exposure to famine in humans. *Proc. Natl. Acad. Sci. U. S. A.* 105, 17046–17049.

- Heiss, J.A., Just, A.C., 2019. Improved filtering of DNA methylation microarray data by detection p values and its impact on downstream analyses. *Clin. Epigenetics* 11, 15.
- Henry, L.A., Cassidy, T., McLaughlin, M., Pentieva, K., McNulty, H., Walsh, C.P., Lees-Murdock, D., 2018. Folic Acid Supplementation throughout pregnancy: psychological Hill, P.W.S., Leitch, H.G., Requena, C.E., Sun, Z., Amouroux, R., Roman-Trufero, M., Borkowska, M., Terragni, J., Vaisvila, R., Linnett, S., Bagci, H., Dharmalingham, G., Haberle, V., Lenhard, B., Zheng, Y., Pradhan, S., Hajkova, P., 2018. Epigenetic reprogramming enables the transition from primordial germ cell to gonocyte. *Nature* 555, 392–396.
- Hirayama, T., Yagi, T., 2017. Regulation of clustered protocadherin genes in individual neurons. *Semin. Cell Dev. Biol.*
- Hollox, E.J., Huffmeier, U., Zeeuwen, P.L.J.M., Palla, R., Lascorz, J., Rodijk-Olthuis, D., Van De Kerkhof, P.C.M., Traupe, H., De Jongh, G., Heijer, M. Den, Reis, A., Armour, J.A.L., Schalkwijk, J., 2008. Psoriasis is associated with increased  $\beta$ -defensin genomic copy number. *Nat. Genet.* 40, 23–25.
- Hoyo, C., Daltveit, A.K., Iversen, E., Benjamin-Neelon, S.E., Fuemmeler, B., Schildkraut, J., Murtha, A.P., Overcash, F., Vidal, A.C., Wang, F., Huang, Z., Kurtzberg, J., Seewaldt, V., Forman, M., Jirtle, R.L., Murphy, S.K., 2014. Erythrocyte folate concentrations, CpG methylation at genomically imprinted domains, and birth weight in a multiethnic newborn cohort. *Epigenetics* 9, 1120–1130.
- Hu, D.G., Mackenzie, P.I., McKinnon, R.A., Meech, R., 2016. Genetic polymorphisms of human UDP-glucuronosyltransferase (UGT) genes and cancer risk. *Drug Metab. Rev.*
- Hüffmeier, U., Bergboer, J.G.M., Becker, T., Armour, J.A., Traupe, H., Estivill, X., Riveira-Munoz, E., Mössner, R., Reich, K., Kurrat, W., Wienker, T.F., Schalkwijk, J., Zeeuwen, P.L.J.M., Reis, A., 2010. Replication of LCE3C-LCE3B CNV as a risk factor for psoriasis and analysis of interaction with other genetic risk factors. *J. Invest. Dermatol.* 130, 979–984.
- Hutnick, L.K., Huang, X., Loo, T.C., Ma, Z., Fan, G., 2010. Repression of retrotransposal elements in mouse embryonic stem cells is primarily mediated by a DNA methylation-independent mechanism. *J. Biol. Chem.* 285, 21082–21091.
- Jandaghi, P., Hoheisel, J., Riazalhosseini, Y., 2015. GHSR hypermethylation: A promising pan-cancer marker. *Cell Cycle* 14, 689–690.
- Jones, P.A., Liang, G., 2009. Rethinking how DNA methylation patterns are maintained. *Nat. Rev. Genet.* 10, 805–811.
- Joubert, B.R., den Dekker, H.T., Felix, J.F., Bohlin, J., Ligthart, S., Beckett, E., Tiemeier, H., van Meurs, J.B., Uitterlinden, A.G., Hofman, A., Håberg, S.E., Reese, S.E., Peters, M.J., Kulle Andreassen, B., Steegers, E.A.P., Nilsen, R.M., Vollset, S.E., Midttun, Ø., Ueland, P.M., Franco, O.H., Dehghan, A., de Jongste, J.C., Wu, M.C., Wang, T., Peddada, S.D., Jaddoe, V.W. V., Nystad, W., Duijts, L., London, S.J., 2016. Maternal plasma folate impacts differential DNA methylation in an epigenome-wide meta-analysis of newborns. *Nat. Commun.* 7, 10577.
- Joyce, C.E., Zhou, X., Xia, J., Ryan, C., Thrash, B., Menter, A., Zhang, W., Bowcock, A.M., 2011. Deep sequencing of small RNAs from human skin reveals major alterations in the psoriasis miRNAome. *Hum. Mol. Genet.* 20, 4025–4040.

- Juergens, R.A., Wrangle, J., Vendetti, F.P., Murphy, S.C., Zhao, M., Coleman, B., Sebree, R., Rodgers, K., Hooker, C.M., Franco, N., Lee, B., Tsai, S., Delgado, I.E., Rudek, M.A., Belinsky, S.A., Herman, J.G., Baylin, S.B., Brock, M. V, Rudin, C.M., 2011. Combination epigenetic therapy has efficacy in patients with refractory advanced non-small cell lung cancer. *Cancer Discov.* 1, 598–607.
- Kamińska, K., Białkowska, A., Kowalewski, J., Huang, S., Lewandowska, M.A., 2019. Differential gene methylation patterns in cancerous and non-cancerous cells. *Oncol. Rep.* 42, 43–54.
- Karimi, M.M., Goyal, P., Maksakova, I.A., Bilenky, M., Leung, D., Tang, J.X., Shinkai, Y., Mager, D.L., Jones, S., Hirst, M., Lorincz, M.C., 2011. DNA methylation and SETDB1/H3K9me3 regulate predominantly distinct sets of genes, retroelements, and chimeric transcripts in mescs. *Cell Stem Cell* 8, 676–687.
- Kassiotis, G., Stoye, J.P., 2016. Immune responses to endogenous retroelements: Taking the bad with the good. *Nat. Rev. Immunol.*
- Kido, T., Lau, Y.F.C., 2014. The Y-located gonadoblastoma gene TSPY amplifies its own expression through a positive feedback loop in prostate cancer cells. *Biochem. Biophys. Res. Commun.* 446, 206–211.
- Kpetemey, M., Dasgupta, S., Rajendiran, S., Das, S., Gibbs, L.D., Shetty, P., Gryczynski, Z., Vishwanatha, J.K., 2015. MIEN1, a novel interactor of Annexin A2, promotes tumor cell migration by enhancing AnxA2 cell surface expression. *Mol. Cancer* 14.
- Kretz, M., Siprashvili, Z., Chu, C., Webster, D.E., Zehnder, A., Qu, K., Lee, C.S., Flockhart, R.J., Groff, A.F., Chow, J., Johnston, D., Kim, G.E., Spitale, R.C., Flynn, R.A., Zheng, G.X.Y., Aiyer, S., Raj, A., Rinn, J.L., Chang, H.Y., Khavari, P.A., 2013. Control of somatic tissue differentiation by the long non-coding RNA TINCR. *Nature* 493, 231–235.
- Kushwaha, P.P., Gupta, S., Singh, A.K., Kumar, S., 2019. Emerging Role of Migration and Invasion Enhancer 1 (MIEN1) in Cancer Progression and Metastasis. *Front. Oncol.*
- Landan, G., Cohen, N.M., Mukamel, Z., Bar, A., Molchadsky, A., Brosh, R., Horn-Saban, S., Zalcenstein, D.A., Goldfinger, N., Zundelovich, A., Gal-Yam, E.N., Rotter, V., Tanay, A., 2012. Epigenetic polymorphism and the stochastic formation of differentially methylated regions in normal and cancerous tissues. *Nat. Genet.* 44, 1207–1214.
- Laugesen, A., Højfeldt, J.W., Helin, K., 2016. Role of the polycomb repressive complex 2 (PRC2) in transcriptional regulation and cancer. *Cold Spring Harb. Perspect. Med.* 6.
- Levy, J.J., Titus, A.J., Salas, L.A., Christensen, B.C., 2019. PyMethylProcess - highly parallelized preprocessing for DNA methylation array data. *bioRxiv* 604496.
- Li, H., Chiappinelli, K.B., Guzzetta, A.A., Easwaran, H., Yen, R.-W.C., Vata-palli, R., Topper, M.J., Luo, J., Connolly, R.M., Azad, N.S., Stearns, V., Pardoll, D.M., Davidson, N., Jones, P.A., Slamon, D.J., Baylin, S.B., Zahnow, C.A., Ahuja, N., 2014. Immune regulation by low doses of the DNA methyltransferase inhibitor 5-azacitidine in common human epithelial cancers. *Oncotarget* 5, 587–98.
- Li, P., Wang, L., Bennett, B.D., Wang, J., Li, J., Qin, Y., Takaku, M., Wade, P.A., Wong, J., Hu, G., 2017. Rif1 promotes a repressive chromatin state to safeguard against endogenous retrovirus activation. *Nucleic Acids Res.* 45, 12723–12738.
- Li, T., Wang, L., Du, Y., Xie, S., Yang, X., Lian, F., Zhou, Z., Qian, C., 2018. Structural and

- mechanistic insights into UHRF1-mediated DNMT1 activation in the maintenance DNA methylation. *Nucleic Acids Res.* 46, 3218–3231.
- Li, X., Ito, M., Zhou, F., Youngson, N., Zuo, X., Leder, P., Ferguson-Smith, A.C., 2008. A Maternal-Zygotic Effect Gene, *Zfp57*, Maintains Both Maternal and Paternal Imprints. *Dev. Cell* 15, 547–557.
- Li, Y., Huang, T., Zheng, Y., Muka, T., Troup, J., Hu, F.B., 2016. Folic Acid Supplementation and the Risk of Cardiovascular Diseases: A Meta-Analysis of Randomized Controlled Trials. *J. Am. Heart Assoc.* 5.
- Liang, G., Chan, M.F., Tomigahara, Y., Tsai, Y.C., Gonzales, F.A., Li, E., Laird, P.W., Jones, P.A., 2002. Cooperativity between DNA methyltransferases in the maintenance methylation of repetitive elements. *Mol. Cell. Biol.* 22, 480–91.
- Liao, J., Karnik, R., Gu, H., Ziller, M.J., Clement, K., Tsankov, A.M., Akopian, V., Gifford, C.A., Donaghey, J., Galonska, C., Pop, R., Reyon, D., Tsai, S.Q., Mallard, W., Joung, J.K., Rinn, J.L., Gnirke, A., Meissner, A., 2015. Targeted disruption of DNMT1, DNMT3A and DNMT3B in human embryonic stem cells. *Nat. Genet.* 47, 469–478.
- Lin, T.C., Hsiao, M., 2017. Ghrelin and cancer progression. *Biochim. Biophys. Acta - Rev. Cancer.*
- Liu, D., Ray, B., Neavin, D.R., Zhang, J., Athreya, A.P., Biernacka, J.M., Bobo, W. V., Hall-Flavin, D.K., Skime, M.K., Zhu, H., Jenkins, G.D., Batzler, A., Kalari, K.R., Boakye-Agyeman, F., Matson, W.R., Bhasin, S.S., Mushiroda, T., Nakamura, Y., Kubo, M., Iyer, R.K., Wang, L., Frye, M.A., Kaddurah-Daouk, R., Weinshilboum, R.M., 2018. Beta-defensin 1, aryl hydrocarbon receptor and plasma kynurenine in major depressive disorder: Metabolomics-informed genomics. *Transl. Psychiatry* 8.
- Liu, J. (Jenny), Ward, R.L., 2010. Folate and One-Carbon Metabolism and Its Impact on Aberrant DNA Methylation in Cancer. In: *Advances in Genetics*. pp. 79–121.
- Liu, M., Ohtani, H., Zhou, W., Ørskov, A.D., Charlet, J., Zhang, Y.W., Shen, H., Baylin, S.B., Liang, G., Grønbaek, K., Jones, P.A., 2016. Vitamin C increases viral mimicry induced by 5-aza-2'-deoxycytidine. *Proc. Natl. Acad. Sci. U. S. A.* 113, 10238–10244.
- Liu, W., Hajjar, K.A., 2016. The annexin A2 system and angiogenesis. *Biol. Chem.*
- Liu, Y., Helms, C., Liao, W., Zaba, L.C., Duan, S., Gardner, J., Wise, C., Miner, A., Malloy, M.J., Pullinger, C.R., Kane, J.P., Saccone, S., Worthington, J., Bruce, I., Kwok, P.Y., Menter, A., Krueger, J., Barton, A., Saccone, N.L., Bowcock, A.M., 2008. A genome-wide association study of psoriasis and psoriatic arthritis identifies new disease loci. *PLoS Genet.* 4.
- Liu, Y., Myrvang, H.K., Dekker, L. V., 2015. Annexin A2 complexes with S100 proteins: Structure, function and pharmacological manipulation. *Br. J. Pharmacol.*
- Liu, Y., Toh, H., Sasaki, H., Zhang, X., Cheng, X., 2012. An atomic model of *Zfp57* recognition of CpG methylation within a specific DNA sequence. *Genes Dev.*
- Liu, Z., Wang, W., Li, Q., Tang, M., Li, J., Wu, W., Wan, Y., Wang, Z., Bao, S., Fei, J., 2015. Growth hormone secretagogue receptor is important in the development of experimental colitis. *Cell Biosci.* 5, 12.



- Lopez-Rios, J., 2016. The many lives of SHH in limb development and evolution. *Semin. Cell Dev. Biol.*
- Loughery, J.E.P., Dunne, P.D., O'Neill, K.M., Meehan, R.R., McDaid, J.R., Walsh, C.P., 2011. DNMT1 deficiency triggers mismatch repair defects in human cells through depletion of repair protein levels in a process involving the DNA damage response. *Hum. Mol. Genet.* 20, 3241–3255.
- Luis, N.M., Morey, L., Mejetta, S., Pascual, G., Janich, P., Kuebler, B., Roma, G., Nascimento, E., Frye, M., Di Croce, L., Benitah, S.A., 2011. Regulation of human epidermal stem cell proliferation and senescence requires polycomb-dependent and independent functions of cbx4. *Cell Stem Cell* 9, 233–246.
- Mackay, D.J.G., Callaway, J.L.A., Marks, S.M., White, H.E., Acerini, C.L., Boonen, S.E., Dayanikli, P., Firth, H. V., Goodship, J.A., Haemers, A.P., Hahnemann, J.M.D., Kordonouri, O., Masoud, A.F., Oestergaard, E., Storr, J., Ellard, S., Hattersley, A.T., Robinson, D.O., Temple, I.K., 2008. Hypomethylation of multiple imprinted loci in individuals with transient neonatal diabetes is associated with mutations in ZFP57. *Nat. Genet.* 40, 949–951.
- Mackin, S.-J., O'Neill, K.M., Walsh, C.P., 2018. Comparison of DNMT1 inhibitors by methylome profiling identifies unique signature of 5-aza-2'deoxyctidine. *Epigenomics* 10, 1085–1101.
- Maes, M., Bosmans, E., Suy, E., Vandervorst, C., DeJonckheere, C., Raus, J., 1991. Depression-related disturbances in mitogen-induced lymphocyte responses and interleukin-1 $\beta$  and soluble interleukin-2 receptor production. *Acta Psychiatr. Scand.* 84, 379–386.
- Mah, K.M., Houston, D.W., Weiner, J.A., 2016. The  $\gamma$  3-Protocadherin-C3 isoform inhibits canonical Wnt signaling by binding to and stabilizing Axin1 at the membrane. *Sci. Rep.* 6.
- Mah, K.M., Weiner, J.A., 2017. Regulation of Wnt signaling by protocadherins. *Semin. Cell Dev. Biol.*
- Maßberg, D., Hatt, H., 2018. Human olfactory receptors: Novel cellular functions outside of the nose. *Physiol. Rev.*
- Matsui, T., Leung, D., Miyashita, H., Maksakova, I.A., Miyachi, H., Kimura, H., Tachibana, M., Lorincz, M.C., Shinkai, Y., 2010. Proviral silencing in embryonic stem cells requires the histone methyltransferase ESET. *Nature* 464, 927–931.
- McNulty, B., McNulty, H., Marshall, B., Ward, M., Molloy, A.M., Scott, J.M., Dornan, J., Pentieva, K., 2013. Impact of continuing folic acid after the first trimester of pregnancy: Findings of a randomized trial of folic acid supplementation in the second and third trimesters. *Am. J. Clin. Nutr.* 98, 92–98.
- McNulty, H., Rollins, M., Cassidy, T., Caffrey, A., Marshall, B., Dornan, J., McLaughlin, M., McNulty, B.A., Ward, M., Strain, J.J., Molloy, A.M., Lees-Murdock, D.J., Walsh, C.P., Pentieva, K., 2019. Effect of continued folic acid supplementation beyond the first trimester of pregnancy on cognitive performance in the child: a follow-up study from a

- randomized controlled trial (FASSTT Offspring Trial). *BMC Med.* 17, 196.
- Ménard, C., Hodes, G.E., Russo, S.J., 2016. Pathogenesis of depression: Insights from human and rodent studies. *Neuroscience*.
- Messerschmidt, D.M., 2012. Should I stay or should I go, protection and maintenance of DNA methylation at imprinted genes. *Epigenetics* 7, 969–975.
- Mikkelsen, T.S., Ku, M., Jaffe, D.B., Issac, B., Lieberman, E., Giannoukos, G., Alvarez, P., Brockman, W., Kim, T.K., Koche, R.P., Lee, W., Mendenhall, E., O'Donovan, A., Presser, A., Russ, C., Xie, X., Meissner, A., Wernig, M., Jaenisch, R., Nusbaum, C., Lander, E.S., Bernstein, B.E., 2007. Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature* 448, 553–560.
- Miller, A.H., Raison, C.L., 2016. The role of inflammation in depression: From evolutionary imperative to modern treatment target. *Nat. Rev. Immunol.*
- Milutinovic, S., Zhuang, Q., Niveleau, A., Szyf, M., 2003. Knockdown of DNA methyltransferase 1 triggers an intra-S-phase arrest of DNA replication and induction of stress response genes. *J. Biol. Chem.* 278, 14985–14995.
- Monahan, K., Lomvardas, S., 2015. Monoallelic Expression of Olfactory Receptors. *Annu. Rev. Cell Dev. Biol.* 31, 721–740.
- Mountoufaris, G., Canzio, D., Nwakeze, C.L., Chen, W. V., Maniatis, T., 2018. Writing, Reading, and Translating the Clustered Protocadherin Cell Surface Recognition Code for Neural Circuit Assembly. *Annu. Rev. Cell Dev. Biol.* 34, 471–493.
- Murphy, T.M., Crawford, B., Dempster, E.L., Hannon, E., Burrage, J., Turecki, G., Kaminsky, Z., Mill, J., 2017. Methylomic profiling of cortex samples from completed suicide cases implicates a role for PSORS1C3 in major depression and suicide. *Transl. Psychiatry* 7.
- Muscattell, K.A., Dedovic, K., Slavich, G.M., Jarcho, M.R., Breen, E.C., Bower, J.E., Irwin, M.R., Eisenberger, N.I., 2015. Greater amygdala activity and dorsomedial prefrontal-amygdala coupling are associated with enhanced inflammatory responses to stress. *Brain. Behav. Immun.* 43, 46–53.
- Neri, F., Incarnato, D., Krepelova, A., Rapelli, S., Pagnani, A., Zecchina, R., Parlato, C., Oliviero, S., 2013. Genome-wide analysis identifies a functional association of Tet1 and Polycomb repressive complex 2 in mouse embryonic stem cells. *Genome Biol.* 14, R91.
- O'Leary, R., Reilly, J.E., Hanson, H.H., Kang, S., Lou, N., Phillips, G.R., 2011. A variable cytoplasmic domain segment is necessary for  $\gamma$ -protocadherin trafficking and tubulation in the endosome/lysosome pathway. *Mol. Biol. Cell* 22, 4362–4372.
- O'Neill, S., McLafferty, M., Ennis, E., Lapsley, C., Bjourson, T., Armour, C., Murphy, S., Bunting, B., Murray, E., 2018. Socio-demographic, mental health and childhood adversity risk factors for self-harm and suicidal behaviour in College students in Northern Ireland. *J. Affect. Disord.* 239, 58–65.
- Oda, S., Fukami, T., Yokoi, T., Nakajima, M., 2014. Epigenetic regulation of the tissue-specific expression of human UDP-glucuronosyltransferase (UGT) 1A10. *Biochem. Pharmacol.* 87, 660–667.

- Pariser, D., Schenkel, B., Carter, C., Farahi, K., Brown, T.M., Ellis, C.N., 2016. A multicenter, non-interventional study to evaluate patient-reported experiences of living with psoriasis. *J. Dermatolog. Treat.* 27, 19–26.
- Patel, N., Nadkarni, A., Cardwell, L.A., Vera, N., Frey, C., Patel, N., Feldman, S.R., 2017. Psoriasis, Depression, and Inflammatory Overlap: A Review. *Am. J. Clin. Dermatol.* 18, 613–620.
- Peek, S.L., Mah, K.M., Weiner, J.A., 2017. Regulation of neural circuit formation by protocadherins. *Cell. Mol. Life Sci.*
- Peng, H., Liu, H., Liu, F., Gao, Y., Chen, J., Huo, J., Han, J., Xiao, T., Zhang, W., 2017. NLRP2 and FAF1 deficiency blocks early embryogenesis in the mouse. *Reproduction* 154, 245–251.
- Prelot, L., Draisma, H., Anasanti, M.D., Balkhiyarova, Z., Wielscher, M., Yengo, L., Sebert, S., Ala-Korpela, M., Froguel, P., Jarvelin, M.-R., Kaakinen, M., Prokopenko, I., 2018. Machine Learning in Multi-Omics Data to Assess Longitudinal Predictors of Glycaemic Trait Levels. *bioRxiv* 358390.
- Premratanachai, P., Joly, S., Johnson, G.K., McCray, P.B., Jia, H.P., Guthmiller, J.M., 2004. Expression and regulation of novel human  $\beta$ -defensins in gingival keratinocytes. *Oral Microbiol. Immunol.* 19, 111–117.
- Quenneville, S., Verde, G., Corsinotti, A., Kapopoulou, A., Jakobsson, J., Offner, S., Baglivo, I., Pedone, P. V., Grimaldi, G., Riccio, A., Trono, D., 2011. In embryonic stem cells, ZFP57/KAP1 recognize a methylated hexanucleotide to affect chromatin and DNA methylation of imprinting control regions. *Mol. Cell* 44, 361–372.
- Rahim, T., Rashid, R., 2017. Comparison of depression symptoms between primary depression and secondary-to-schizophrenia depression. *Int. J. Psychiatry Clin. Pract.* 21, 314–317.
- Rajarajan, P., Borrman, T., Liao, W., Schrode, N., Flaherty, E., Casiño, C., Powell, S., Yashaswini, C., LaMarca, E.A., Kassim, B., Javidfar, B., Espeso-Gil, S., Li, A., Won, H., Geschwind, D.H., Ho, S.M., MacDonald, M., Hoffman, G.E., Roussos, P., Zhang, B., Hahn, C.G., Weng, Z., Brennand, K.J., Akbarian, S., 2018. Neuron-specific signatures in the chromosomal connectome associated with schizophrenia risk. *Science* (80- ). 362.
- Reddington, J.P., Perricone, S.M., Nestor, C.E., Reichmann, J., Youngson, N.A., Suzuki, M., Reinhardt, D., Dunican, D.S., Prendergast, J.G., Mjoseng, H., Ramsahoye, B.H., Whitelaw, E., Grealley, J.M., Adams, I.R., Bickmore, W.A., Meehan, R.R., 2013. Redistribution of H3K27me3 upon DNA hypomethylation results in de-repression of Polycomb target genes. *Genome Biol.* 14.
- Richmond, R.C., Sharp, G.C., Herbert, G., Atkinson, C., Taylor, C., Bhattacharya, S., Campbell, D., Hall, M., Kazmi, N., Gaunt, T., McArdle, W., Ring, S., Davey Smith, G., Ness, A., Relton, C.L., 2018. The long-term impact of folic acid in pregnancy on offspring DNA methylation: follow-up of the Aberdeen Folic Acid Supplementation Trial (AFASST). *Int. J. Epidemiol.*
- Riveira-Munoz, E., He, S.M., Escaramís, G., Stuart, P.E., Hüffmeier, U., Lee, C., Kirby, B., Oka, A., Giardina, E., Liao, W., Bergboer, J., Kainu, K., De Cid, R., Munkhbat, B., Zeeuwen, P.L.J.M., Armour, J.A.L., Poon, A., Mabuchi, T., Ozawa, A., Zawirska, A., Burden, A.D., Barker, J.N., Capon, F., Traupe, H., Sun, L.D., Cui, Y., Yin, X.Y., Chen, G., Lim, H.W., Nair,

- R.P., Voorhees, J.J., Tejasvi, T., Pujol, R., Munkhtuvshin, N., Fischer, J., Kere, J., Schalkwijk, J., Bowcock, A., Kwok, P.Y., Novelli, G., Inoko, H., Ryan, A.W., Trembath, R.C., Reis, A., Zhang, X.J., Elder, J.T., Estivill, X., 2011. Meta-analysis confirms the LCE3C-LCE3B deletion as a risk factor for psoriasis in several ethnic groups and finds interaction with HLA-Cw6. *J. Invest. Dermatol.* 131, 1105–1109.
- Robson, M.J., Quinlan, M.A., Blakely, R.D., 2017. Immune System Activation and Depression: Roles of Serotonin in the Central Nervous System and Periphery. *ACS Chem. Neurosci.*
- Rollins, R.A., Haghghi, F., Edwards, J.R., Das, R., Zhang, M.Q., Ju, J., Bestor, T.H., 2006. Large-scale structure of genomic methylation patterns. *Genome Res.* 16, 157–163.
- Rothbart, S.B., Dickson, B.M., Ong, M.S., Krajewski, K., Houliston, S., Kireev, D.B., Arrowsmith, C.H., Strahl, B.D., 2013. Multivalent histone engagement by the linked tandem tudor and PHD domains of UHRF1 is required for the epigenetic inheritance of DNA methylation. *Genes Dev.* 27, 1288–1298.
- Rothbart, S.B., Krajewski, K., Nady, N., Tempel, W., Xue, S., Badeaux, A.I., Barsyte-Lovejoy, D., Martinez, J.Y., Bedford, M.T., Fuchs, S.M., Arrowsmith, C.H., Strahl, B.D., 2012. Association of UHRF1 with methylated H3K9 directs the maintenance of DNA methylation. *Nat. Struct. Mol. Biol.* 19, 1155–1160.
- Roulois, D., Loo Yau, H., Singhanian, R., Wang, Y., Danesh, A., Shen, S.Y., Han, H., Liang, G., Jones, P.A., Pugh, T.J., O'Brien, C., De Carvalho, D.D., 2015. DNA-Demethylating Agents Target Colorectal Cancer Cells by Inducing Viral Mimicry by Endogenous Transcripts. *Cell* 162, 961–973.
- Rowe, H.M., Friedli, M., Offner, S., Verp, S., Mesnard, D., Marquis, J., Aktas, T., Trono, D., 2013. De novo DNA methylation of endogenous retroviruses is shaped by KRAB-ZFPs/KAP1 and ESET. *Dev.* 140, 519–529.
- Sagoo, G.S., Tazi-Ahnini, R., Barker, J.W.N., Elder, J.T., Nair, R.P., Samuelsson, L., Traupe, H., Trembath, R.C., Robinson, D.A., Iles, M.M., 2004. Meta-analysis of genome-wide studies of psoriasis susceptibility reveals linkage to chromosomes 6p21 and 4q28-q31 in Caucasian and Chinese Hans population. *J. Invest. Dermatol.* 122, 1401–1405.
- Salmaninejad, A., Zamani, M.R., Pourvahedi, M., Golchehre, Z., Hosseini Bereshneh, A., Rezaei, N., 2016. Cancer/Testis Antigens: Expression, Regulation, Tumor Invasion, and Use in Immunotherapy of Cancers. *Immunol. Invest.*
- Samkari, A., White, J., Packer, R., 2015. SHH inhibitors for the treatment of medulloblastoma. *Expert Rev. Neurother.*
- Schlesinger, Y., Straussman, R., Keshet, I., Farkash, S., Hecht, M., Zimmerman, J., Eden, E., Yakhini, Z., Ben-Shushan, E., Reubinoff, B.E., Bergman, Y., Simon, I., Cedar, H., 2007. Polycomb-mediated methylation on Lys27 of histone H3 pre-marks genes for de novo methylation in cancer. *Nat. Genet.* 39, 232–236.
- Schorn, A.J., Gutbrod, M.J., LeBlanc, C., Martienssen, R., 2017. LTR-Retrotransposon Control by tRNA-Derived Small RNAs. *Cell* 170, 61–71.e11.
- Seshan VE, O.A., 2018. DNACopy: DNA copy number data analysis.
- Sharif, J., Endo, T.A., Nakayama, M., Karimi, M.M., Shimada, M., Katsuyama, K., Goyal, P., Brind'Amour, J., Sun, M.A., Sun, Z., Ishikura, T., Mizutani-Koseki, Y., Ohara, O., Shinkai,

- Y., Nakanishi, M., Xie, H., Lorincz, M.C., Koseki, H., 2016. Activation of Endogenous Retroviruses in Dnmt1<sup>-/-</sup> ESCs Involves Disruption of SETDB1-Mediated Repression by NP95 Binding to Hemimethylated DNA. *Cell Stem Cell* 19, 81–94.
- Sinclair, K.D., Allegrucci, C., Singh, R., Gardner, D.S., Sebastian, S., Bispham, J., Thurston, A., Huntley, J.F., Rees, W.D., Maloney, C.A., Lea, R.G., Craigon, J., McEvoy, T.G., Young, L.E., 2007. DNA methylation, insulin resistance, and blood pressure in offspring determined by maternal periconceptional B vitamin and methionine status. *Proc. Natl. Acad. Sci. U. S. A.* 104, 19351–19356.
- Smith, A.K., Kilaru, V., Klengel, T., Mercer, K.B., Bradley, B., Conneely, K.N., Ressler, K.J., Binder, E.B., 2015. DNA extracted from saliva for methylation studies of psychiatric traits: Evidence tissue specificity and relatedness to brain. *Am. J. Med. Genet. Part B Neuropsychiatr. Genet.* 168, 36–44.
- Smith, Z.D., Chan, M.M., Mikkelsen, T.S., Gu, H., Gnirke, A., Regev, A., Meissner, A., 2012. A unique regulatory phase of DNA methylation in the early mammalian embryo. *Nature* 484, 339–344.
- Song, X., Ji, J., Gleason, K.J., Yang, F., Martignetti, J.A., Chen, L.S., Wang, P., 2019. Insights into impact of DNA copy number alteration and methylation on the proteogenomic landscape of human ovarian cancer via a multi-omics integrative analysis. *Mol. Cell. Proteomics* 18, S52–S65.
- Strick, R., Strissel, P.L., Baylin, S.B., Chiappinelli, K.B., 2016. Unraveling the molecular pathways of DNA-methylation inhibitors: human endogenous retroviruses induce the innate immune response in tumors.
- Stuart, P.E., Hüffmeier, U., Nair, R.P., Palla, R., Tejasvi, T., Schalkwijk, J., Elder, J.T., Reis, A., Armour, J.A.L., 2012. Association of  $\beta$ -defensin copy number and psoriasis in three cohorts of European origin. *J. Invest. Dermatol.* 132, 2407–2413.
- Takahashi, N., Coluccio, A., Thorball, C.W., Planet, E., Shi, H., Offner, S., Turelli, P., Imbeault, M., Ferguson-Smith, A.C., Trono, D., 2019. ZNF445 is a primary regulator of genomic imprinting. *Genes Dev.* 33, 49–54.
- Teschendorff, A.E., Relton, C.L., 2018. Statistical and integrative system-level analysis of DNA methylation data. *Nat. Rev. Genet.*
- Tie, C.H., Fernandes, L., Conde, L., Robbez-Masson, L., Sumner, R.P., Peacock, T., Rodriguez-Plata, M.T., Mickute, G., Gifford, R., Towers, G.J., Herrero, J., Rowe, H.M., 2018. KAP 1 regulates endogenous retroviruses in adult human cells and contributes to innate immune control. *EMBO Rep.* 19.
- Timis, T.L., Orasan, R.I., 2018. Understanding psoriasis: Role of miRNAs (review). *Biomed. Reports.*
- Touati, A., Errea-Dorransoro, J., Nouri, S., Halleb, Y., Pereda, A., Mahdhaoui, N., Ghith, A., Saad, A., Perez de Nanclares, G., H'mida ben brahim, D., 2019. Transient neonatal diabetes mellitus and hypomethylation at additional imprinted loci: novel ZFP57 mutation and review on the literature. *Acta Diabetol.* 56, 301–307.
- Toyoda, S., Kawaguchi, M., Kobayashi, T., Tarusawa, E., Toyama, T., Okano, M., Oda, M., Nakauchi, H., Yoshimura, Y., Sanbo, M., Hirabayashi, M., Hirayama, T., Hirabayashi, T., Yagi, T., 2014. Developmental epigenetic modification regulates stochastic expression

of clustered Protocadherin genes, generating single neuron diversity. *Neuron* 82, 94–108.

Unterberger, A., Andrews, S.D., Weaver, I.C.G., Szyf, M., 2006. DNA Methyltransferase 1 Knockdown Activates a Replication Stress Checkpoint. *Mol. Cell. Biol.* 26, 7575–7586.

Viré, E., Brenner, C., Deplus, R., Blanchon, L., Fraga, M., Didelot, C., Morey, L., Van Eynde, A., Bernard, D., Vanderwinden, J.M., Bollen, M., Esteller, M., Di Croce, L., De Launoit, Y., Fuks, F., 2006. The Polycomb group protein EZH2 directly controls DNA methylation. *Nature* 439, 871–874.

Vos, T., Abajobir, A.A., Abbafati, C., Abbas, K.M., Abate, K.H., Abd-Allah, F., Abdulle, A.M., Abebo, T.A., Abera, S.F., Aboyans, V., Abu-Raddad, L.J., Ackerman, I.N., Adamu, A.A., Adetokunboh, O., Afarideh, M., Afshin, A., Agarwal, S.K., Aggarwal, R., Agrawal, A., Agrawal, S., Ahmad Kiadaliri, A., Ahmadi, H., Ahmed, M.B., Aichour, A.N., Aichour, I., Aichour, M.T.E., Aiyar, S., Akinyemi, R.O., Akseer, N., Al Lami, F.H., Alahdab, F., Al-Aly, Z., Alam, K., Alam, N., Alam, T., Alasfoor, D., Alene, K.A., Ali, R., Alizadeh-Navaei, R., Alkerwi, A., Alla, F., Allebeck, P., Allen, C., Al-Maskari, F., Al-Raddadi, R., Alsharif, U., Alsowaidi, S., Altirkawi, K.A., Amare, A.T., Amini, E., Ammar, W., Amoako, Y.A., Andersen, H.H., Antonio, C.A.T., Anwari, P., Ärnlöv, J., Artaman, A., Aryal, K.K., Asayesh, H., Asgedom, S.W., Assadi, R., Atey, T.M., Atnafu, N.T., Atre, S.R., Avila-Burgos, L., Avokpaho, E.F.G.A., Awasthi, A., Ayala Quintanilla, B.P., Ba Saleem, H.O., Bacha, U., Badawi, A., Balakrishnan, K., Banerjee, A., Bannick, M.S., Barac, A., Barber, R.M., Barker-Collo, S.L., Bärnighausen, T., Barquera, S., Barregard, L., Barrero, L.H., Basu, S., Battista, B., Battle, K.E., Baune, B.T., Bazargan-Hejazi, S., Beardsley, J., Bedi, N., Beghi, E., Béjot, Y., Bekele, B.B., Bell, M.L., Bennett, D.A., Bensenor, I.M., Benson, J., Berhane, A., Berhe, D.F., Bernabé, E., Betsu, B.D., Beuran, M., Beyene, A.S., Bhala, N., Bhansali, A., Bhatt, S., Bhutta, Z.A., Biadgilign, S., Bienhoff, K., Bikbov, B., Birungi, C., Biryukov, S., Bisanzio, D., Bizuayehu, H.M., Boneya, D.J., Boufous, S., Bourne, R.R.A., Brazinova, A., Brugha, T.S., Buchbinder, R., Bulto, L.N.B., Bumgarner, B.R., Butt, Z.A., Cahuana-Hurtado, L., Cameron, E., Car, M., Carabin, H., Carapetis, J.R., Cárdenas, R., Carpenter, D.O., Carrero, J.J., Carter, A., Carvalho, F., Casey, D.C., Caso, V., Castañeda-Orjuela, C.A., Castle, C.D., Catalá-López, F., Chang, H.Y., Chang, J.C., Charlson, F.J., Chen, H., Chibalabala, M., Chibueze, C.E., Chisumpa, V.H., Chitheer, A.A., Christopher, D.J., Ciobanu, L.G., Cirillo, M., Colombara, D., Cooper, C., Cortesi, P.A., Criqui, M.H., Crump, J.A., Dadi, A.F., Dalal, K., Dandona, L., Dandona, R., Das Neves, J., Davitoiu, D. V., De Courten, B., De Leo, D., Degenhardt, L., Deiparine, S., Dellavalle, R.P., Deribe, K., Des Jarlais, D.C., Dey, S., Dharmaratne, S.D., Dhillon, P.K., Dicker, D., Ding, E.L., Djalalinia, S., Do, H.P., Dorsey, E.R., Dos Santos, K.P.B., Douwes-Schultz, D., Doyle, K.E., Driscoll, T.R., Dubey, M., Duncan, B.B., El-Khatib, Z.Z., Ellerstrand, J., Enayati, A., Endries, A.Y., Ermakov, S.P., Erskine, H.E., Eshrati, B., Eskandarieh, S., Esteghamati, A., Estep, K., Fanuel, F.B.B., Farinha, C.S.E.S., Faro, A., Farzadfar, F., Fazeli, M.S., Feigin, V.L., Fereshtehnejad, S.M., Fernandes, J.C., Ferrari, A.J., Feyissa, T.R., Filip, I., Fischer, F., Fitzmaurice, C., Flaxman, A.D., Flor, L.S., Foigt, N., Foreman, K.J., Franklin, R.C., Fullman, N., Fürst, T., Furtado, J.M., Futran, N.D., Gakidou, E., Ganji, M., Garcia-Basteiro, A.L., Gebre, T., Gebrehiwot, T.T., Geleto, A., Gemechu, B.L., Gesesew, H.A., Gething, P.W., Ghajar, A., Gibney, K.B., Gill, P.S., Gillum, R.F., Ginawi, I.A.M., Giref, A.Z., Gishu, M.D., Giussani, G., Godwin, W.W., Gold, A.L., Goldberg, E.M., Gona, P.N., Goodridge, A., Gopalani, S.V., Goto, A., Goulart, A.C., Griswold, M., Gughani, H.C., Gupta, R., Gupta, R., Gupta, T., Gupta, V., Hafezi-Nejad, N., Hailu, A.D., Hailu, G.B., Hamadeh, R.R., Hamidi, S., Handal, A.J., Hankey, G.J., Hao, Y., Harb, H.L., Hareri, H.A., Haro, J.M., Harvey, J.,

Hassanvand, M.S., Havmoeller, R., Hawley, C., Hay, R.J., Hay, S.I., Henry, N.J., Heredia-Pi, I.B., Heydarpour, P., Hoek, H.W., Hoffman, H.J., Horita, N., Hosgood, H.D., Hostiuc, S., Hotez, P.J., Hoy, D.G., Htet, A.S., Hu, G., Huang, H., Huynh, C., Iburg, K.M., Igumbor, E.U., Ikeda, C., Irvine, C.M.S., Jacobsen, K.H., Jahanmehr, N., Jakovljevic, M.B., Jassal, S.K., Javanbakht, M., Jayaraman, S.P., Jeemon, P., Jensen, P.N., Jha, V., Jiang, G., John, D., Johnson, C.O., Johnson, S.C., Jonas, J.B., Jürisson, M., Kabir, Z., Kadel, R., Kahsay, A., Kamal, R., Kan, H., Karam, N.E., Karch, A., Karema, C.K., Kasaeian, A., Kassa, G.M., Kassaw, N.A., Kassebaum, N.J., Kastor, A., Katikireddi, S.V., Kaul, A., Kawakami, N., Keiyoro, P.N., Kengne, A.P., Keren, A., Khader, Y.S., Khalil, I.A., Khan, E.A., Khang, Y.H., Khosravi, A., Khubchandani, J., Kieling, C., Kim, D., Kim, P., Kim, Y.J., Kimokoti, R.W., Kinfu, Y., Kisa, A., Kissimova-Skarbek, K.A., Kivimaki, M., Knudsen, A.K., Kokubo, Y., Kolte, D., Kopec, J.A., Kosen, S., Koul, P.A., Koyanagi, A., Kravchenko, M., Krishnaswami, S., Krohn, K.J., Kuate Defo, B., Kucuk Bicer, B., Kumar, G.A., Kumar, P., Kumar, S., Kyu, H.H., Lal, D.K., Laloo, R., Lambert, N., Lan, Q., Larsson, A., Lavados, P.M., Leasher, J.L., Lee, J.T., Lee, P.H., Leigh, J., Leshargie, C.T., Leung, J., Leung, R., Levi, M., Li, Y., Li, Y., Li Kappe, D., Liang, X., Liben, M.L., Lim, S.S., Linn, S., Liu, A., Liu, P.Y., Liu, S., Liu, Y., Lodha, R., Logroscino, G., London, S.J., Looker, K.J., Lopez, A.D., Lorkowski, S., Lotufo, P.A., Low, N., Lozano, R., Lucas, T.C.D., Macarayan, E.R.K., Magdy Abd El Razek, H., Magdy Abd El Razek, M., Mahdavi, M., Majdan, M., Majdzadeh, R., Majeed, A., Malekzadeh, R., Malhotra, R., Malta, D.C., Mamun, A.A., Manguerra, H., Manhertz, T., Mantilla, A., Mantovani, L.G., Mapoma, C.C., Marczak, L.B., Martinez-Raga, J., Martins-Melo, F.R., Martopullo, I., März, W., Mathur, M.R., Mazidi, M., McAlinden, C., McGaughey, M., McGrath, J.J., McKee, M., McNellan, C., Mehata, S., Mehndiratta, M.M., Mekonnen, T.C., Memiah, P., Memish, Z.A., Mendoza, W., Mengistie, M.A., Mengistu, D.T., Mensah, G.A., Meretoja, A., Meretoja, T.J., Mezgebe, H.B., Micha, R., Milllear, A., Miller, T.R., Mills, E.J., Mirarefin, M., Mirrakhimov, E.M., Misganaw, A., Mishra, S.R., Mitchell, P.B., Mohammad, K.A., Mohammadi, A., Mohammed, K.E., Mohammed, S., Mohanty, S.K., Mokdad, A.H., Mollenkopf, S.K., Monasta, L., Hernandez, J.M., Montico, M., Moradi-Lakeh, M., Moraga, P., Mori, R., Morozoff, C., Morrison, S.D., Moses, M., Mountjoy-Venning, C., Mruts, K.B., Mueller, U.O., Muller, K., Murdoch, M.E., Murthy, G.V.S., Musa, K.I., Nachega, J.B., Nagel, G., Naghavi, M., Naheed, A., Naidoo, K.S., Naldi, L., Nangia, V., Natarajan, G., Negasa, D.E., Negoi, I., Negoi, R.I., Newton, C.R., Ngunjiri, J.W., Nguyen, C.T., Nguyen, G., Nguyen, M., Nguyen, Q. Le, Nguyen, T.H., Nichols, E., Ningrum, D.N.A., Nolte, S., Nong, V.M., Norrving, B., Noubiap, J.J.N., O'Donnell, M.J., Ogbo, F.A., Oh, I.H., Okoro, A., Oladimeji, O., Olagunju, A.T., Olagunju, T.O., Olsen, H.E., Olusanya, B.O., Olusanya, J.O., Ong, K., Opio, J.N., Oren, E., Ortiz, A., Osgood-Zimmerman, A., Osman, M., Owolabi, M.O., Pa, M., Pacella, R.E., Pana, A., Panda, B.K., Papachristou, C., Park, E.K., Parry, C.D., Parsaeian, M., Patten, S.B., Patton, G.C., Paulson, K., Pearce, N., Pereira, D.M., Perico, N., Pesudovs, K., Peterson, C.B., Petzold, M., Phillips, M.R., Pigott, D.M., Pillay, J.D., Pinho, C., Plass, D., Pletcher, M.A., Popova, S., Poulton, R.G., Pourmalek, F., Prabhakaran, D., Prasad, N., Prasad, N.M., Purcell, C., Qorbani, M., Quansah, R., Rabiee, R.H.S., Radfar, A., Rafay, A., Rahimi, K., Rahimi-Movaghar, A., Rahimi-Movaghar, V., Rahman, M., Rahman, M.H.U., Rai, R.K., Rajsic, S., Ram, U., Ranabhat, C.L., Rankin, Z., Rao, P.V., Rao, P.C., Rawaf, S., Ray, S.E., Reiner, R.C., Reinig, N., Reitsma, M.B., Remuzzi, G., Renzaho, A.M.N., Resnikoff, S., Rezaei, S., Ribeiro, A.L., Ronfani, L., Roshandel, G., Roth, G.A., Roy, A., Rubagotti, E., Ruhago, G.M., Saadat, S., Sadat, N., Safdarian, M., Safi, S., Safiri, S., Sagar, R., Sahathevan, R., Salama,

J., Salomon, J.A., Salvi, S.S., Samy, A.M., Sanabria, J.R., Santomauro, D., Santos, I.S., Santos, J.V., Santric Milicevic, M.M., Sartorius, B., Satpathy, M., Sawhney, M., Saxena, S., Schmidt, M.I., Schneider, I.J.C., Schöttker, B., Schwebel, D.C., Schwendicke, F., Seedat, S., Sepanlou, S.G., Servan-Mori, E.E., Setegn, T., Shackelford, K.A., Shaheen, A., Shaikh, M.A., Shamsipour, M., Shariful Islam, S.M., Sharma, J., Sharma, R., She, J., Shi, P., Shields, C., Shigematsu, M., Shinohara, Y., Shiri, R., Shirkoobi, R., Shirude, S., Shishani, K., Shrimel, M.G., Sibai, A.M., Sigfusdottir, I.D., Silva, D.A.S., Silva, J.P., Silveira, D.G.A., Singh, J.A., Singh, N.P., Sinha, D.N., Skiadaresi, E., Skirbekk, V., Slepak, E.L., Sligar, A., Smith, D.L., Smith, M., Sobaih, B.H.A., Sobngwi, E., Sorensen, R.J.D., Sousa, T.C.M., Sposato, L.A., Sreeramareddy, C.T., Srinivasan, V., Stanaway, J.D., Stathopoulou, V., Steel, N., Stein, D.J., Stein, M.B., Steiner, C., Steiner, T.J., Steinke, S., Stokes, M.A., Stovner, L.J., Strub, B., Subart, M., Sufiyan, M.B., Suliankatchi Abdulkader, R., Sunguya, B.F., Sur, P.J., Swaminathan, S., Sykes, B.L., Sylte, D.O., Tabarés-Seisdedos, R., Taffere, G.R., Takala, J.S., Tandon, N., Tavakkoli, M., Taveira, N., Taylor, H.R., Tehrani-Banihashemi, A., Tekelab, T., Temam Shifa, G., Terkawi, A.S., Tesfaye, D.J., Tesso, B., Thamsuwan, O., Thomas, K.E., Thrift, A.G., Tiruye, T.Y., Tobe-Gai, R., Tollanes, M.C., Tonelli, M., Topor-Madry, R., Tortajada, M., Touvier, M., Tran, B.X., Tripathi, S., Troeger, C., Truelsen, T., Tsoi, D., Tuem, K.B., Tuzcu, E.M., Tyrovolas, S., Ukwaja, K.N., Undurraga, E.A., Uneke, C.J., Updike, R., Uthman, O.A., Uzochukwu, B.S.C., Van Boven, J.F.M., Varughese, S., Vasankari, T., Venkatesh, S., Venketasubramanian, N., Vidavalur, R., Violante, F.S., Vladimirov, S.K., Vlassov, V.V., Vollset, S.E., Wadilo, F., Wakayo, T., Wang, Y.P., Weaver, M., Weichenthal, S., Weiderpass, E., Weintraub, R.G., Werdecker, A., Westerman, R., Whiteford, H.A., Wijeratne, T., Wiysonge, C.S., Wolfe, C.D.A., Woodbrook, R., Woolf, A.D., Workicho, A., Wulf Hanson, S., Xavier, D., Xu, G., Yadgir, S., Yaghoubi, M., Yakob, B., Yan, L.L., Yano, Y., Ye, P., Yimam, H.H., Yip, P., Yonemoto, N., Yoon, S.J., Yotebieng, M., Younis, M.Z., Zaidi, Z., Zaki, M.E.S., Zegeye, E.A., Zenebe, Z.M., Zhang, X., Zhou, M., Zipkin, B., Zodpey, S., Zuhlke, L.J., Murray, C.J.L., 2017. Global, regional, and national incidence, prevalence, and years lived with disability for 328 diseases and injuries for 195 countries, 1990-2016: A systematic analysis for the Global Burden of Disease Study 2016. *Lancet* 390, 1211–1259.

Walsh, C.P., Chaillet, J.R., Bestor, T.H., 1998. Transcription of IAP endogenous retroviruses is constrained by cytosine methylation. *Nat. Genet.* 20, 116–117.

Walter, M., Teissandier, A., Pérez-Palacios, R., Bourc'his, D., 2016. An epigenetic switch ensures transposon repression upon dynamic loss of DNA methylation in embryonic stem cells. *Elife* 5.

Wang, S., Zhang, C., Hasson, D., Desai, A., SenBanerjee, S., Magnani, E., Ukomadu, C., Lujambio, A., Bernstein, E., Sadler, K.C., 2019. Epigenetic Compensation Promotes Liver Regeneration. *Dev. Cell* 50, 43–56.e6.

Wang, X., Branciamore, S., Gogoshin, G., Ding, S., Rodin, A.S., 2019. New analysis framework incorporating mixed mutual information and scalable Bayesian networks for multimodal high dimensional genomic and epigenomic cancer data. *bioRxiv* 812446.

Waterland, R.A., Kellermayer, R., Laritsky, E., Rayco-Solon, P., Harris, R.A., Travisano, M., Zhang, W., Torskaya, M.S., Zhang, J., Shen, L., Manary, M.J., Prentice, A.M., 2010. Season of conception in rural gambia affects DNA methylation at putative human metastable epialleles. *PLoS Genet.* 6, 1–10.

Weber, J., Salgaller, M., Samid, D., Johnson, B., Herlyn, M., Lassam, N., Treisman, J., Rosenberg, S.A., 1994. Expression of the MAGE-1 Tumor Antigen Is Up-Regulated by



- the Demethylating Agent 5-Aza-2'-Deoxycytidine. *Cancer Res.* 54, 1766–1771.
- Widschwendter, M., Fiegl, H., Egle, D., Mueller-Holzner, E., Spizzo, G., Marth, C., Weisenberger, D.J., Campan, M., Young, J., Jacobs, I., Laird, P.W., 2007. Epigenetic stem cell signature in cancer. *Nat. Genet.* 39, 157–158.
- Wieczorek, M., Abualrous, E.T., Sticht, J., Álvaro-Benito, M., Stolzenberg, S., Noé, F., Freund, C., 2017. Major histocompatibility complex (MHC) class I and MHC class II proteins: Conformational plasticity in antigen presentation. *Front. Immunol.*
- Won, E., Kim, Y.-K., 2016. Stress, the Autonomic Nervous System, and the Immune-kynurenine Pathway in the Etiology of Depression. *Curr. Neuropharmacol.* 14, 665–673.
- Wrangle, J., Wang, W., Koch, A., Easwaran, H., Mohammad, H.P., Vendetti, F., Vancrackinge, W., Demeyer, T., Du, Z., Parsana, P., Rodgers, K., Yen, R.-W., Zahnow, C.A., Taube, J.M., Brahmer, J.R., Tykodi, S.S., Easton, K., Carvajal, R.D., Jones, P.A., Laird, P.W., Weisenberger, D.J., Tsai, S., Juergens, R.A., Topalian, S.L., Rudin, C.M., Brock, M. V, Pardoll, D., Baylin, S.B., 2013. Alterations of immune response of Non-Small Cell Lung Cancer with Azacytidine. *Oncotarget* 4, 2067–79.
- Xie, S., Qian, C., 2018. The growing complexity of UHRF1-mediated maintenance DNA methylation. *Genes (Basel)*.
- Xu, D., Zhang, B., Liao, C., Zhang, W., Wang, W., Chang, Y., Shao, Y., 2016. Human beta-defensin 3 contributes to the carcinogenesis of cervical cancer via activation of NF- $\kappa$ B signaling. *Oncotarget* 7, 75902–75913.
- Yasar, U., Greenblatt, D.J., Guillemette, C., Court, M.H., 2013. Evidence for regulation of UDP-glucuronosyltransferase (UGT) 1A1 protein expression and activity via DNA methylation in healthy human livers. *J. Pharm. Pharmacol.* 65, 874–883.
- Zeng, Y., Chen, T., 2019. DNA methylation reprogramming during mammalian development. *Genes (Basel)*.
- Zhao, Q., Caballero, O.L., Simpson, A.J.G., Strausberg, R.L., 2012. Differential Evolution of MAGE Genes Based on Expression Pattern and Selection Pressure. *PLoS One* 7, e48240.
- Zhao, Y., Gong, X., Chen, L., Li, L., Liang, Y., Chen, S., Zhang, Y., 2014. Site-specific methylation of placental HSD11B2 gene promoter is related to intrauterine growth restriction. *Eur. J. Hum. Genet.* 22, 734–740.
- Zielinski, T.A., Sherwood Brown, E., Nejtek, V.A., Khan, D.A., Moore, J.J., John Rush, A., 2000. Depression in asthma: Prevalence and clinical implications. *Prim. Care Companion J. Clin. Psychiatry* 2, 153–158.
- Zuo, X., Sheng, J., Lau, H.T., McDonald, C.M., Andrade, M., Cullen, D.E., Bell, F.T., Iacovino, M., Kyba, M., Xu, G., Li, X., 2012. Zinc finger protein ZFP57 requires its co-factor to recruit DNA methyltransferases and maintains DNA methylation imprint in embryonic stem cells via its transcriptional repression domain. *J. Biol. Chem.* 287, 2107–2118.

## 8.0 Achievements

### 8.1 Published Abstracts

- Depletion Of Dnmt1 In Differentiated Human Cells Highlights Key Classes Of Dependent Genes K M. O’Neill, R Irwin , SJ Mackin, A Thakur, SJ Thursby, C Bertens, L Masala, J Loughery, D McArt and CP. Walsh - 20th Meeting of the Irish Society of Human Genetics: Friday 15th September 2017 Croke Park, Dublin, The Ulster Medical Journal, 87, 1, 2018
- Candidate gene methylation quantification (CandiMeth) within the Galaxy Bioinformatics Interface. Sara-Jayne Thursby, Sarah-Jayne Mackin et al., F1000Research, 7, 7 2018

### 8.2 Additional Research Training

- International Leadership & Management Award completed to level 5

### 8.3 Certificates

- Royal Statistical Society Advanced R Workshop (2017)
- Ulster University R Workshop (2017)

### 8.4 Conference Presentations

#### 8.4.1 Oral Presentations

- Oral Presentation: Epigenetics in Health and Disease Workshop (Ulster University Coleraine, Northern Ireland. April 10<sup>th</sup> 2019 – abstract not published)
- Oral Presentation: Galaxy Community and Open Source Bioinformatics Conference (Portland, Oregon. 24<sup>th</sup> June – 2<sup>nd</sup> July 2018; abstract and slides - <https://f1000research.com/slides/7-1060>)
- Oral Presentation: IBMS All Ireland Postgraduate Conference (Sligo, Ireland. 20<sup>th</sup> – 21<sup>st</sup> June 2017 – abstract not published)

#### 8.4.2 Poster Presentations

- Poster Presentation: Galaxy Community and Open Source Bioinformatics Conference (Portland, Oregon. 24<sup>th</sup> June – 2<sup>nd</sup> July 2018, abstract and poster [https://static.sched.com/hosted\\_files/gccbosc2018/b7/f1000research-210815.pdf](https://static.sched.com/hosted_files/gccbosc2018/b7/f1000research-210815.pdf))

- Poster Presentation: European Alliance for Personalised Medicine (Belfast, Northern Ireland, 27-30<sup>th</sup> November 2017 – abstract not published)
- Poster Presentation: Irish Society for Human Genetics (Dublin, Ireland. 15<sup>th</sup> September 2017 – <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5849961/>)

## 8.5 Grants

- Galaxy Community Fellowship Award (£1600)
- Genetics Society Training Grant (£1000)
- Cancer Translational Research Group (£1000)
- Doctoral Researcher Initiative (£500)
- ‘WhyR’ Conference Grant (£450)
- Santander Initiative (£250)
- IBMS All Ireland Postgraduate Conference Bursary (£50)

## 8.6 Other Achievements

- Treasurer and Head of Public Relations for the Biomedical Sciences Postgraduate Society
- STEM Ambassador for Northern Ireland
- Conducted SEO and Traffic analysis for The Genetics Society of the UK
- Aided in the re-design of The Genetics Society UK website ([www.genetics.org.uk](http://www.genetics.org.uk)) for their 100-year anniversary

## 8.7 Publications

- O’Neill, K.M., Irwin, R.E., Mackin, S.-J., Thursby, S.-J., Thakur, A., Bertens, C., Masala, L., Loughery, J.E.P., McArt, D.G., Walsh, C.P., 2018. Depletion of DNMT1 in differentiated human cells highlights key classes of sensitive genes and an interplay with polycomb repression. *Epigenetics and Chromatin* 11.

I conducted a significant portion of the bioinformatic analysis such as the overlap between hypomethylated and hypermethylated probes with that of ENCODE chromatin state segmentation tracks and started to develop a CandiMeth prototype.

- Irwin, R.E., Thursby, S.-J., Ondičová, M., Pentieva, K., McNulty, H., Richmond, R.C., Caffrey, A., Lees-Murdock, D.J., McLaughlin, M., Cassidy, T., Suderman, M., Relton, C.L., Walsh, C.P., 2019. A randomized controlled trial of folic acid intervention in pregnancy highlights a putative methylation-regulated control element at ZFP57. *Clin. Epigenetics* 11, 31.

I conducted the majority of the EPIC array analysis for this RTC in RnBeads and Limma, in addition to SVA and cell-type correction. I made all UCSC genome browser tracks for this study and generated a QQ plot to check for population stratification effects. I also compared the results of this trial to an independent cohort and made improvements to the CandiMeth prototype.

- Amenyah, S.D., Hughes, C.F., Ward, M., Rosborough, S., Deane, J., Thursby, S.-J., Walsh, C.P., Kok, D.E., Strain, J.J., McNulty, H., Lees-Murdock, D.J., 2020. Influence of nutrients involved in one-carbon metabolism on DNA methylation in adults—a systematic review and meta-analysis. *Nutr. Rev.*

I conducted array analysis and GO for targets of cardiovascular disease that resulted from the Kok, Dieuwertje 2015 study of nutrient supplementation in the elderly.

- Thursby, S.J., Mackin, S.J., Irwin, R.E., Walsh, C.P., CandiMeth: Candidate gene methylation quantification within the Galaxy Bioinformatics Interface (2020) (In Review, *GigaScience*).

I designed this workflow and co-wrote the manuscript as detailed in <http://bit.do/candimeth>

- Irwin, R.E., Lapsley, C., Thursby, S.J., Walsh, C.P., Murray, E. AESUS study regarding methylation changes in depression and suicide. (2020) (In Review, *Clinical Epigenetics*).

I conducted EPIC array analysis in multiple different programs, created UCSC genome tracks, conducted CNV analysis for each participant and plotted a QQ plot to assess for population bias.

- Irwin, R.E., Scullion, C., Thursby, S.J., Sun, M., Thakur, A., Rothbart, S., Xu, G-L., Walsh, C.P. UHRF1 is Required to Suppress Viral Mimicry Through Both DNA Methylation-Dependent and Independent Mechanisms (2020) (Assembled Manuscript).

I conducted EPIC array analysis for the majority of the cell lines and mapped repeat regions as defined by RepeatMasker to probes within the EPIC array in order to assess the changes in methylation at repetitive elements. I also analysed RNA-seq data from an independent study to compare to our UHRF1 KD to a similar UHRF1 KD to investigate the differences.