Genome-wide Distribution and Regulation of DNA Methylation and Hydroxymethylation in Prostate Cancer

by

Thuy Linh Ho

A thesis submitted in conformity with the requirements for the degree of Master of Science
Department of Laboratory Medicine and Pathobiology
University of Toronto

© Copyright by Thuy Linh Ho 2014
Genome-wide Distribution and Regulation of DNA Methylation and Hydroxymethylation in Prostate Cancer

Thuy Linh Ho
Master of Science
Department of Laboratory Medicine and Pathobiology
University of Toronto
2014

Abstract

Prostate cancer (PCa) is the most common malignancy in men. Epigenetic alterations, including DNA methylation, significantly influence disease pathogenesis. Aberrant methylation of the cytosine base (5mC) transcriptionally regulates several genes associated with PCa development. Recent discovery that 5mC marks are converted to DNA hydroxymethylation (5hmC) marks by ten-eleven translocation (TET) proteins has opened avenues for epigenome regulation.

The intent of my MSc research was to define the genome-wide DNA methylation and hydroxymethylation patterns in prostate cell lines, benign and cancerous, using next-generation sequencing technology and bioinformatics analysis. I defined regions of differential methylation between normal and cancer, and suggest novel biomarkers for the disease. Additionally, this is the first study to discover the landscape of DNA hydroxymethylation in PCa, which suggests enrichment in specific genomic regions. Furthermore, I revealed pathways potentially regulated by 5mC and 5hmC marks to expand our knowledge of aberrant epigenetic processes in PCa pathogenesis.
Acknowledgments

I would like to thank my supervisor, Dr. Bharati Bapat for her continuous support and critical advice during my Master's research project. Bharati is a leader in the field of epigenetic biomarkers and I admire her dedication and passion for the discoveries and ideas she is a part of.

I thank members of the Bapat lab for creating a positive research environment. Ken: thank you for mentoring me when I first started in the lab. You were a patient teacher and I learned strong work ethic from you. Ekaterina, Andrea, Thomas, Nicole, Julia, and Fang: you all were always available to provide advice and feedback when my experiments were going awry. Thank you for helping me prepare for presentations and ultimately my thesis defense.

I would like to acknowledge Dr. Gary Bader and Ruth Isserlin for their guidance and advice in pathways-based analysis performed during this project.

I thank the members of my MSc advisory committee: Dr. Susan Done, Dr. Neil Fleshner, and Dr. Michael Taylor. I appreciate all of your ideas and suggestions towards improving my project and for being supportive of the work I have presented to you. A special thank you to Dr. George Charames for chairing my defense.

I thank Eugene for being my best friend; you believed in me and supported me through everything. Thank you to CLO for designing beautiful diagrams for my thesis and being an amazing friend.

During my Master's project I was supported by: The Bill and Kathleen Troost Innovation Grant of the Canadian Cancer Society (Grant #701456), University of Toronto Fellowship, Lorne F. Lambier, Q.C. Scholarship, OSOTF Studentship at Mount Sinai Hospital, and Paul Starita Graduate Student Fellowship.
Contributions

All work in Chapter 2 and data analysis was performed by Linh T. Ho, with the exception that KK, a research assistant, completed the first replicate of MBD-seq samples. RI, a research associate from Dr. Gary Bader’s lab, provided advice on pathway-based analyses and developed a GREAT plugin for the Cytoscape program.

The work in Chapter 3 and data analysis was primarily contributed by Linh T. Ho. JG, a laboratory technician, performed the dot blot analysis on DNA samples extracted by LTH and assisted with RT-qPCR validation experiments. AS, a PhD student from Dr. Philip Marsden’s lab, performed the liquid chromatography-tandem mass spectrometry analysis on DNA samples extracted by LTH. NWA, a post-doctoral fellow, assisted with the design of RT-qPCR validation studies. RI, a research associate from Dr. Gary Bader’s lab, provided advice on pathway-based analyses.

The work in the Appendix and data analysis was primarily contributed by Linh T. Ho. JG, a laboratory technician, generated stable TET2 knockdown clones in RWPE-1 cells, completed western blot analysis for TET2 protein expression, and assisted with RT-qPCR experiments for TET2 mRNA expression. KK, a research assistant and EOM, a PhD student, extracted RNA from prostate cancer cell lines.

# Table of Contents

ABSTRACT ................................................................................................................................. ii
ACKNOWLEDGEMENTS ........................................................................................................ iii
CONTRIBUTIONS.................................................................................................................... iv
TABLE OF CONTENTS .............................................................................................................. v
LIST OF TABLES ..................................................................................................................... viii
LIST OF FIGURES .................................................................................................................... ix
LIST OF APPENDICES ............................................................................................................ x
LIST OF ABBREVIATIONS ......................................................................................................... xi

## CHAPTER 1. GENERAL INTRODUCTION

1.1 EPIGENETIC ALTERATIONS ................................................................. 1
   1.1.1 DNA Methylation ................................................................. 1
   1.1.2 Histone Modifications and Nucleosome Remodeling .......... 2
   1.1.3 MicroRNAs .......................................................................... 4

1.2 DNA METHYLATION PATTERNS AND EFFECTS ON GENE EXPRESSION .... 6
   1.2.1 CpG Island Methylation ..................................................... 6
   1.2.2 Global Hypomethylation in Cancer ................................ 8
   1.2.3 Locus-specific Hypermethylation in Cancer .................. 9

1.3 DNA HYDROXYMETHYLATION PATTERNS ................................. 14
   1.3.1 DNA Hydroxymethylation and DNA Demethylation Pathways 14
   1.3.2 TET Proteins .................................................................... 18

1.4 CURRENT TECHNOLOGY IN THE DETECTION OF DNA METHYLATION AND DNA HYDROXYMETHYLATION ................................................................. 19
   1.4.1 Genome-wide DNA Methylation Detection Techniques 19
   1.4.2 Global and Locus-specific Detection Techniques of DNA Hydroxymethylation ......................................................... 21

1.5 PROSTATE CANCER ......................................................................... 24
   1.5.1 Incidence, Survival, and Risk Factors ............................. 24
   1.5.2 Screening, Prognosis, and Treatment ............................ 24
   1.5.3 Genetic Alterations ............................................................ 26
   1.5.4 Epigenetic Alterations ...................................................... 28

1.6 HYPOTHESIS AND OBJECTIVES .................................................. 31
   1.6.1 Hypothesis .......
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.2.2</td>
<td>DNA Methylation Enrichment: MBD2 and MeDIP-capture</td>
<td>33</td>
</tr>
<tr>
<td>2.2.3</td>
<td>Real Time Quantitative PCR (RT-qPCR) following DNA Methylation Enrichment: MBD2- and MeDIP-qPCR</td>
<td>33</td>
</tr>
<tr>
<td>2.2.4</td>
<td>Next-Generation Sequencing of MBD2-enriched DNA: MBD-seq</td>
<td>34</td>
</tr>
<tr>
<td>2.2.5</td>
<td>Bioinformatic Analysis of MBD-seq</td>
<td>34</td>
</tr>
<tr>
<td>2.2.6</td>
<td>Pathway-based Analysis of MBD-seq</td>
<td>34</td>
</tr>
<tr>
<td>2.3</td>
<td>RESULTS</td>
<td>36</td>
</tr>
<tr>
<td>2.3.1</td>
<td>Optimization of Methylation Enrichment Techniques, MBD2- and MeDIP-capture</td>
<td>36</td>
</tr>
<tr>
<td>2.3.2</td>
<td>Relative Absolute and Differential Analysis of MBD-seq in Prostate Cell lines, Benign and Cancerous</td>
<td>38</td>
</tr>
<tr>
<td>2.3.3</td>
<td>Validation of MBD-seq Results using Methylation Microarrays</td>
<td>41</td>
</tr>
<tr>
<td>2.3.4</td>
<td>Pathways Associated and Potentially Regulated by Differentially Methylated Regions (DMRs) in Prostate Cell lines</td>
<td>43</td>
</tr>
<tr>
<td>2.4</td>
<td>DISCUSSION</td>
<td>48</td>
</tr>
</tbody>
</table>

**CHAPTER 3. INVESTIGATING THE HYDROXYMETHYLOMEN IN PROSTATE CELL LINES**

<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.1</td>
<td>INTRODUCTION</td>
<td>54</td>
</tr>
<tr>
<td>3.2</td>
<td>MATERIALS AND METHODS</td>
<td>55</td>
</tr>
<tr>
<td>3.2.1</td>
<td>Cell Culture and DNA Extraction</td>
<td>55</td>
</tr>
<tr>
<td>3.2.2</td>
<td>DNA Hydroxymethylation Enrichment: hMeSeal</td>
<td>55</td>
</tr>
<tr>
<td>3.2.3</td>
<td>Next-Generation Sequencing of Glucosylated 5hmC-enriched DNA: hMeSeal-seq</td>
<td>55</td>
</tr>
<tr>
<td>3.2.4</td>
<td>Bioinformatic Analysis of hMeSeal-seq</td>
<td>55</td>
</tr>
<tr>
<td>3.2.5</td>
<td>Global DNA Hydroxymethylation Detection using Dot Blot Assay</td>
<td>56</td>
</tr>
<tr>
<td>3.2.6</td>
<td>Locus-specific DNA Hydroxymethylation Detection using hMeSeal-qPCR</td>
<td>57</td>
</tr>
<tr>
<td>3.2.7</td>
<td>Pathway-based Analysis of hMeSeal-seq</td>
<td>57</td>
</tr>
<tr>
<td>3.3</td>
<td>RESULTS</td>
<td>59</td>
</tr>
<tr>
<td>3.3.1</td>
<td>Relative Absolute Analysis of hMeSeal-seq in Prostate Cell lines, Benign and Cancerous</td>
<td>59</td>
</tr>
<tr>
<td>3.3.2</td>
<td>Differential Analysis of hMeSeal-seq in Prostate Cell lines, Benign and Cancerous</td>
<td>64</td>
</tr>
<tr>
<td>3.3.3</td>
<td>Validation of hMeSeal-seq Results using Dot Blot Assay</td>
<td>66</td>
</tr>
<tr>
<td>3.3.4</td>
<td>Validation of hMeSeal-seq Results using hMeSeal-qPCR</td>
<td>68</td>
</tr>
<tr>
<td>3.3.5</td>
<td>Pathways Associated and Potentially Regulated by Relative Absolute Regions of Hydroxymethylation in Prostate Cell lines</td>
<td>70</td>
</tr>
<tr>
<td>3.4</td>
<td>DISCUSSION</td>
<td>74</td>
</tr>
</tbody>
</table>

**CHAPTER 4. CONCLUSIONS AND FUTURE DIRECTIONS**

<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.1</td>
<td>CONCLUSIONS</td>
<td>79</td>
</tr>
<tr>
<td>4.2</td>
<td>FUTURE DIRECTIONS</td>
<td>81</td>
</tr>
<tr>
<td>APPENDIX</td>
<td>.........................................................................................</td>
<td>84</td>
</tr>
</tbody>
</table>
S1.1 SUPPLEMENTARY MATERIAL AND METHODS

S1.1.1 Endogenous TET2-knockdown in Cell lines 84
S1.1.2 Real Time Quantitative PCR (RT-qPCR) 84
S1.1.3 Protein Extraction and Western Blot Analysis 85

S1.2 SUPPLEMENTARY FIGURES 86

REFERENCES ....................................................................................................................90
List of Tables

Table 1. Representative list of significant differentially methylated regions (DMRs) in 22Rv1 cells, validated on an Agilent Human CpG Island methylation microarray. 42

Table 2. Representative lists of significant DMRs in RWPE-1 and 22Rv1 cells, validated on an Illumina Infinium HumanMethylation450 BeadChips methylation microarray. 42

Table 3. Representative list of candidate DNA methylation biomarkers, identified through pathway-based analysis using GREAT and Post-analysis gene sets 46

Table 4. Pilot analysis of differentially hydroxymethylated regions (DHMRs) identified in RWPE-1 and 22Rv1 cells, validated using hMeSeal-qPCR. 69
## List of Figures

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Figure 1</td>
<td>Epigenetic model.</td>
<td>5</td>
</tr>
<tr>
<td>Figure 2</td>
<td>DNA methylation patterns in normal and cancer cells.</td>
<td>13</td>
</tr>
<tr>
<td>Figure 3</td>
<td>DNA demethylation pathways.</td>
<td>17</td>
</tr>
<tr>
<td>Figure 4</td>
<td>MBD2- and MeDIP-qPCR in 22Rv1 cells using regions of known methylation status from an Agilent Human CpG Island methylation array.</td>
<td>37</td>
</tr>
<tr>
<td>Figure 5</td>
<td>MBD-seq methodology pipeline.</td>
<td>39</td>
</tr>
<tr>
<td>Figure 6</td>
<td>Relative absolute methylated regions and significant differentially methylated regions (DMRs) identified in RWPE-1 and 22Rv1 cells using MBD-seq.</td>
<td>40</td>
</tr>
<tr>
<td>Figure 7</td>
<td>Pathways associated with differentially methylated regions (DMRs) in 22Rv1 in promoter regions overlapping CpG islands (CGIs), identified by MBD-seq.</td>
<td>45</td>
</tr>
<tr>
<td>Figure 8</td>
<td>Pathways associated with differentially methylated regions (DMRs) in RWPE-1 and 22Rv1 cells, overlapping non-CpG islands (non-CGIs), identified by MBD-seq.</td>
<td>47</td>
</tr>
<tr>
<td>Figure 9</td>
<td>hMeSeal-seq methodology pipeline.</td>
<td>61</td>
</tr>
<tr>
<td>Figure 10</td>
<td>Diagnostic enrichment plots of hMeSeal-seq reads and MACS analysis of RWPE-1 and 22Rv1 cells.</td>
<td>62</td>
</tr>
<tr>
<td>Figure 11</td>
<td>Relative absolute regions of hydroxymethylation identified in RWPE-1 and 22Rv1 cells using hMeSeal-seq.</td>
<td>63</td>
</tr>
<tr>
<td>Figure 12</td>
<td>Preliminary differentially hydroxymethylated regions (DHMRs) described in RWPE-1 and 22Rv1 cells using hMeSeal-seq.</td>
<td>65</td>
</tr>
<tr>
<td>Figure 13</td>
<td>Dot blot analysis of global levels of hydroxymethylation in RWPE-1 and 22Rv1 cells.</td>
<td>67</td>
</tr>
<tr>
<td>Figure 14</td>
<td>Preliminary validation of differentially hydroxymethylated regions (DHMRs) identified in RWPE-1 and 22Rv1 cells, using hMeSeal-qPCR.</td>
<td>69</td>
</tr>
</tbody>
</table>
Figure 15. Pathways associated with relative absolute regions of hydroxymethylation in RWPE-1 and 22Rv1 cells identified by hMeSeal-seq.

Figure 16. Pathways associated with relative absolute regions of hydroxymethylation in RWPE-1 cells identified by hMeSeal-seq, and further correlated with RWPE-1 DNase-seq dataset (ENCODEx).
List of Appendices

Supplemental Figure 1. Enrichment map of genes significantly hypermethylated in RWPE-1 and 22Rv1 cells, identified by MBD-seq. 86

Supplemental Figure 2. Diagnostic enrichment plot of hMeSeal-seq samples from RWPE-1 and 22Rv1. 87

Supplemental Figure 3. TET messenger RNA expression profiles across prostate cell lines. 88

Supplemental Figure 4. Messenger RNA and protein profiles of stable TET2 KD RWPE-1 cells. 89
## List of Abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Full Form</th>
</tr>
</thead>
<tbody>
<tr>
<td>2-HG</td>
<td>2-Hydroxyglutarate</td>
</tr>
<tr>
<td>2-OG</td>
<td>2-Oxyglutarate</td>
</tr>
<tr>
<td>5caC</td>
<td>5-Carboxylcytosine</td>
</tr>
<tr>
<td>5fC</td>
<td>5-Formylcytosine</td>
</tr>
<tr>
<td>5hmC</td>
<td>5-Hydroxymethylcytosine</td>
</tr>
<tr>
<td>5hmU</td>
<td>5-Hydroxymethyluracil</td>
</tr>
<tr>
<td>5mC</td>
<td>5-Methylcytosine</td>
</tr>
<tr>
<td>ADT</td>
<td>Androgen deprivation therapy</td>
</tr>
<tr>
<td>AID</td>
<td>Activation induced deaminase</td>
</tr>
<tr>
<td>AGO2</td>
<td>Argonaute 2</td>
</tr>
<tr>
<td>Alu</td>
<td>Arthrobacter luteus</td>
</tr>
<tr>
<td>APC</td>
<td>Adenomatous polyposis coli</td>
</tr>
<tr>
<td>APOBEC</td>
<td>Apolipoprotein B editing complex protein</td>
</tr>
<tr>
<td>AR</td>
<td>Androgen receptor</td>
</tr>
<tr>
<td>BER</td>
<td>Base excision repair</td>
</tr>
<tr>
<td>B-GT</td>
<td>T-phage β-glucosyltransferase</td>
</tr>
<tr>
<td>BMP</td>
<td>Bone morphogenetic protein</td>
</tr>
<tr>
<td>bp</td>
<td>Base pair</td>
</tr>
<tr>
<td>BPE</td>
<td>Bovine pituitary extract</td>
</tr>
<tr>
<td>BPH</td>
<td>Benign prostatic hyperplasia</td>
</tr>
<tr>
<td>BRCA1</td>
<td>Breast cancer 1</td>
</tr>
<tr>
<td>BS-seq</td>
<td>Bisulfite sequencing</td>
</tr>
<tr>
<td>CDKN2A</td>
<td>Cyclin-dependent kinase inhibitor 2A</td>
</tr>
<tr>
<td>CDKN2B</td>
<td>Cyclin-dependent kinase inhibitor 2B</td>
</tr>
<tr>
<td>cDNA</td>
<td>Complementary DNA</td>
</tr>
<tr>
<td>CGI</td>
<td>CpG island</td>
</tr>
<tr>
<td>CHD</td>
<td>Chromodomain helicase DNA-binding</td>
</tr>
<tr>
<td>ChIP-seq</td>
<td>Chromatin immunoprecipitation sequencing</td>
</tr>
<tr>
<td>Abbreviation</td>
<td>Full Form</td>
</tr>
<tr>
<td>--------------</td>
<td>-----------</td>
</tr>
<tr>
<td>Chr</td>
<td>Chromosome</td>
</tr>
<tr>
<td>CIMP</td>
<td>CpG island methylator phenotype</td>
</tr>
<tr>
<td>c-myc</td>
<td>V-Myc avian myelocytomatosis viral oncogene homolog</td>
</tr>
<tr>
<td>CpG</td>
<td>Cytosine-phosphate-guanosine dinucleotide</td>
</tr>
<tr>
<td>CRC</td>
<td>Colorectal cancer</td>
</tr>
<tr>
<td>CRPC</td>
<td>Castration-resistant prostate cancer</td>
</tr>
<tr>
<td>Ct</td>
<td>Cycle threshold</td>
</tr>
<tr>
<td>CXCL1</td>
<td>Chemokine (C-X-C motif) ligand 1</td>
</tr>
<tr>
<td>DAC</td>
<td>5-Aza-2'-deoxycytidine</td>
</tr>
<tr>
<td>DHMR</td>
<td>Differentially hydroxymethylated region</td>
</tr>
<tr>
<td>DHT</td>
<td>5α-Dihydrotestosterone</td>
</tr>
<tr>
<td>DMR</td>
<td>Differentially methylated region</td>
</tr>
<tr>
<td>DNA</td>
<td>Deoxyribonucleic acid</td>
</tr>
<tr>
<td>DNase I HS</td>
<td>DNase I hypersensitive sites</td>
</tr>
<tr>
<td>DNase-seq</td>
<td>DNase I hypersensitive sites sequencing</td>
</tr>
<tr>
<td>DNMT</td>
<td>DNA methyltransferase</td>
</tr>
<tr>
<td>DNMT1</td>
<td>DNA methyltransferase 1</td>
</tr>
<tr>
<td>DNMT3A</td>
<td>DNA methyltransferase 3A</td>
</tr>
<tr>
<td>DNMT3B</td>
<td>DNA methyltransferase 3B</td>
</tr>
<tr>
<td>dsDNA</td>
<td>Double-stranded DNA</td>
</tr>
<tr>
<td>ECL</td>
<td>Enhanced chemiluminescence</td>
</tr>
<tr>
<td>EGF</td>
<td>Epidermal growth factor</td>
</tr>
<tr>
<td>ELISA</td>
<td>Enzyme-linked immunosorbent assay</td>
</tr>
<tr>
<td>ENCODE</td>
<td>The Encyclopedia of DNA Elements</td>
</tr>
<tr>
<td>ESCs</td>
<td>Embryonic stem cells</td>
</tr>
<tr>
<td>ERG</td>
<td>v-ets erythroblastosis virus E26 homolog</td>
</tr>
<tr>
<td>ETS</td>
<td>v-ets erythroblastosis virus E26 oncogene</td>
</tr>
<tr>
<td>EZH2</td>
<td>Enhancer of zeste homolog 2</td>
</tr>
<tr>
<td>FBP2</td>
<td>Fructose-1,6-bisphosphatase-2</td>
</tr>
<tr>
<td>FBS</td>
<td>Fetal bovine serum</td>
</tr>
<tr>
<td>FOXA1</td>
<td>Forkhead box protein A1</td>
</tr>
<tr>
<td>gDNA</td>
<td>Genomic DNA</td>
</tr>
<tr>
<td>Acronym</td>
<td>Definition</td>
</tr>
<tr>
<td>-----------</td>
<td>-------------------------------------------------</td>
</tr>
<tr>
<td>GLIB</td>
<td>Glucosylation, periodate oxidation, biotinylation</td>
</tr>
<tr>
<td>glucMS-qPCR</td>
<td>Glucosylation-coupled methylation-sensitive quantitative PCR</td>
</tr>
<tr>
<td>GREAT</td>
<td>Genomic Regions Enrichment of Annotations Tool</td>
</tr>
<tr>
<td>GS</td>
<td>Gleason score</td>
</tr>
<tr>
<td>GSEA</td>
<td>Gene Set Enrichment Analysis</td>
</tr>
<tr>
<td>GSTP1</td>
<td>Glutathione S-transferase pi 1</td>
</tr>
<tr>
<td>H2A</td>
<td>Histone 2A</td>
</tr>
<tr>
<td>H2A.Z</td>
<td>Variant of histone H2A</td>
</tr>
<tr>
<td>H2B</td>
<td>Histone 2B</td>
</tr>
<tr>
<td>H3</td>
<td>Histone 3</td>
</tr>
<tr>
<td>H3K4me2</td>
<td>Histone H3 lysine 4 dimethylation</td>
</tr>
<tr>
<td>H3K4me3</td>
<td>Histone H3 lysine 4 trimethylation</td>
</tr>
<tr>
<td>H3K9me3</td>
<td>Histone H3 lysine 9 trimethylation</td>
</tr>
<tr>
<td>H3K18ac</td>
<td>Histone H3 lysine 19 acetylation</td>
</tr>
<tr>
<td>H3K27me3</td>
<td>Histone H3 lysine 27 trimethylation</td>
</tr>
<tr>
<td>H4</td>
<td>Histone 4</td>
</tr>
<tr>
<td>H4K16ac</td>
<td>Histone H4 lysine 16 acetylation</td>
</tr>
<tr>
<td>H4K20me3</td>
<td>Histone H4 lysine 20 trimethylation</td>
</tr>
<tr>
<td>HAT</td>
<td>Histone acetyltransferase</td>
</tr>
<tr>
<td>HDAC</td>
<td>Histone deacetylase</td>
</tr>
<tr>
<td>HDAC1</td>
<td>Histone deacetylase 1</td>
</tr>
<tr>
<td>HDM</td>
<td>Histone demethylase</td>
</tr>
<tr>
<td>HELLs</td>
<td>Helicase, lymphoid-specific</td>
</tr>
<tr>
<td>HGNC</td>
<td>HUGO Gene Nomenclature Committee</td>
</tr>
<tr>
<td>HGPIN</td>
<td>High-grade prostatic intraepithelial neoplasia</td>
</tr>
<tr>
<td>hMeDIP-seq</td>
<td>Hydroxymethylated DNA immunoprecipitation combined with NGS</td>
</tr>
<tr>
<td>hMeSeal</td>
<td>Hydroxymethyl selective chemical labeling</td>
</tr>
<tr>
<td>hMeSeal-qPCR</td>
<td>Hydroxymethyl selective chemical labeling combined with RT-qPCR</td>
</tr>
<tr>
<td>hMeSeal-seq</td>
<td>Hydroxymethyl selective chemical labeling combined with NGS</td>
</tr>
<tr>
<td>HMT</td>
<td>Histone methyltransferase</td>
</tr>
<tr>
<td>HOXA9</td>
<td>Homeobox A9</td>
</tr>
<tr>
<td>HOXD3</td>
<td>Homeobox D3</td>
</tr>
<tr>
<td>Abbreviation</td>
<td>Full Form</td>
</tr>
<tr>
<td>--------------</td>
<td>-----------</td>
</tr>
<tr>
<td>HOXD8</td>
<td>Homeobox D8</td>
</tr>
<tr>
<td>HPLC-MS</td>
<td>High-performance liquid chromatography–mass spectrometry</td>
</tr>
<tr>
<td>HRP</td>
<td>Horseradish peroxidase</td>
</tr>
<tr>
<td>IDH1</td>
<td>Isocitrate dehydrogenase 1</td>
</tr>
<tr>
<td>IDH2</td>
<td>Isocitrate dehydrogenase 2</td>
</tr>
<tr>
<td>IGF1</td>
<td>Insulin-like growth factor 1</td>
</tr>
<tr>
<td>IHC</td>
<td>Immunohistochemical</td>
</tr>
<tr>
<td>JBP-1</td>
<td>J-binding protein 1</td>
</tr>
<tr>
<td>kb</td>
<td>Kilobase</td>
</tr>
<tr>
<td>KD</td>
<td>Knockdown</td>
</tr>
<tr>
<td>KEGG</td>
<td>Kyoto Encyclopedia of Genes and Genomes</td>
</tr>
<tr>
<td>KLK3</td>
<td>Kallikrein-related peptidase 3</td>
</tr>
<tr>
<td>KRuO4</td>
<td>Potassium perruthenate</td>
</tr>
<tr>
<td>K-SFM</td>
<td>Keratinocyte serum free medium</td>
</tr>
<tr>
<td>LC-MS/MS</td>
<td>Liquid chromatography–tandem mass spectrometry</td>
</tr>
<tr>
<td>LHRH</td>
<td>Luteinizing hormone-releasing hormone</td>
</tr>
<tr>
<td>LINE-1</td>
<td>Long interspersed nuclear element-1</td>
</tr>
<tr>
<td>LSD1</td>
<td>Lysine-specific demethylase 1</td>
</tr>
<tr>
<td>MACS</td>
<td>Model-based analysis for ChIP-seq</td>
</tr>
<tr>
<td>MAPK</td>
<td>Mitogen-activated protein kinases</td>
</tr>
<tr>
<td>MBD</td>
<td>Methyl-CpG binding domain protein</td>
</tr>
<tr>
<td>MBD2</td>
<td>Methyl-CpG binding domain protein 2</td>
</tr>
<tr>
<td>MBD3</td>
<td>Methyl-CpG binding domain protein 3</td>
</tr>
<tr>
<td>MECP2</td>
<td>Methyl-CpG-binding protein 2</td>
</tr>
<tr>
<td>MeDIP</td>
<td>Methylated DNA immunoprecipitation</td>
</tr>
<tr>
<td>MeDIP-seq</td>
<td>Methylated DNA immunoprecipitation combined with NGS</td>
</tr>
<tr>
<td>miRNA</td>
<td>Micro ribonucleic acid</td>
</tr>
<tr>
<td>MLH1</td>
<td>MutL homolog 1</td>
</tr>
<tr>
<td>MRE-seq</td>
<td>Methylation-sensitive restriction enzyme sequencing</td>
</tr>
<tr>
<td>mRNA</td>
<td>Messenger ribonucleic acid</td>
</tr>
<tr>
<td>MSI</td>
<td>Microsatellite instability</td>
</tr>
<tr>
<td>MSigDB</td>
<td>The Molecular Signatures Database</td>
</tr>
<tr>
<td>Abbreviation</td>
<td>Full Form</td>
</tr>
<tr>
<td>--------------</td>
<td>-----------</td>
</tr>
<tr>
<td>ncRNA</td>
<td>Non-coding ribonucleic acid</td>
</tr>
<tr>
<td>NDR</td>
<td>Nucleosome-depleted region</td>
</tr>
<tr>
<td>NGC</td>
<td>Non-genotoxic carcinogen</td>
</tr>
<tr>
<td>NGS</td>
<td>Next-generation sequencing</td>
</tr>
<tr>
<td>NNX3.1</td>
<td>NK3 homeobox 1</td>
</tr>
<tr>
<td>NOMe-seq</td>
<td>Nucleosome occupancy and methylome sequencing</td>
</tr>
<tr>
<td>Non-CGI</td>
<td>Non-CpG island</td>
</tr>
<tr>
<td>NSAA</td>
<td>Non-steroidal anti-androgens</td>
</tr>
<tr>
<td>NURD</td>
<td>Nucleosome remodeling deacetylase</td>
</tr>
<tr>
<td>oxBS-seq</td>
<td>Oxidative bisulfite sequencing</td>
</tr>
<tr>
<td>p38</td>
<td>Mitogen-activated protein kinase 14</td>
</tr>
<tr>
<td>p53</td>
<td>Tumor protein 53</td>
</tr>
<tr>
<td>Pc</td>
<td>Polycomb</td>
</tr>
<tr>
<td>PCa</td>
<td>Prostate cancer</td>
</tr>
<tr>
<td>PCA3</td>
<td>Prostate cancer antigen 3</td>
</tr>
<tr>
<td>PcG</td>
<td>Drosophila Polycomb group</td>
</tr>
<tr>
<td>PIN</td>
<td>Prostatic intraepithelial neoplasia</td>
</tr>
<tr>
<td>PRC1</td>
<td>Polycomb repressive complex 1</td>
</tr>
<tr>
<td>PRC2</td>
<td>Polycomb repressive complex 2</td>
</tr>
<tr>
<td>PSA</td>
<td>Prostate-specific antigen</td>
</tr>
<tr>
<td>PTEN</td>
<td>Phosphatase and tensin homolog</td>
</tr>
<tr>
<td>RASSF1A</td>
<td>Tumor suppressor protein RDA32</td>
</tr>
<tr>
<td>RB1</td>
<td>Retinoblastoma 1</td>
</tr>
<tr>
<td>RISC</td>
<td>RNA-induced silencing complex</td>
</tr>
<tr>
<td>RNA</td>
<td>Ribonucleic acid</td>
</tr>
<tr>
<td>RRBS</td>
<td>Reduced representation bisulfite sequencing</td>
</tr>
<tr>
<td>rRNA</td>
<td>Ribosomal ribonucleic acid</td>
</tr>
<tr>
<td>RT-qPCR</td>
<td>Real-time quantitative polymerase chain reaction</td>
</tr>
<tr>
<td>SEM</td>
<td>Standard error of the mean</td>
</tr>
<tr>
<td>shRNA</td>
<td>Short hairpin RNA</td>
</tr>
<tr>
<td>SMUG</td>
<td>Single-strand-selective monofunctional uracil-DNA glycosylase</td>
</tr>
<tr>
<td>SNP</td>
<td>Single-nucleotide polymorphism</td>
</tr>
<tr>
<td>Abbreviation</td>
<td>Full Form</td>
</tr>
<tr>
<td>--------------</td>
<td>-----------</td>
</tr>
<tr>
<td>SSC</td>
<td>Saline-sodium citrate</td>
</tr>
<tr>
<td>ssDNA</td>
<td>Single-stranded DNA</td>
</tr>
<tr>
<td>TAB-seq</td>
<td>TET-assisted bisulfite sequencing</td>
</tr>
<tr>
<td>TBP</td>
<td>TATA box binding protein</td>
</tr>
<tr>
<td>TBS-T</td>
<td>Tris-buffered saline and tween 20</td>
</tr>
<tr>
<td>TCGA</td>
<td>The Cancer Genome Atlas</td>
</tr>
<tr>
<td>TDG</td>
<td>Thymine-DNA glycosylase</td>
</tr>
<tr>
<td>TET</td>
<td>Ten-eleven translocation protein</td>
</tr>
<tr>
<td>TET1</td>
<td>Ten-eleven translocation protein 1</td>
</tr>
<tr>
<td>TET2</td>
<td>Ten-eleven translocation protein 2</td>
</tr>
<tr>
<td>TET3</td>
<td>Ten-eleven translocation protein 3</td>
</tr>
<tr>
<td>TGFβ2</td>
<td>Transforming growth factor beta 2</td>
</tr>
<tr>
<td>TIMPs</td>
<td>Tissue inhibitors of metalloproteases</td>
</tr>
<tr>
<td>TMPRSS2</td>
<td>Transmembrane protease, serine 2</td>
</tr>
<tr>
<td>TPM4</td>
<td>Tropomyosin 4</td>
</tr>
<tr>
<td>TRUS</td>
<td>Transrectal ultrasonography</td>
</tr>
<tr>
<td>TrxG</td>
<td>Drosophila Trithorax group</td>
</tr>
<tr>
<td>TSG</td>
<td>Tumor suppressor gene</td>
</tr>
<tr>
<td>TSS</td>
<td>Transcription start site</td>
</tr>
<tr>
<td>UHRF1</td>
<td>Ubiquitin-like with PHD and ring finger domains 1</td>
</tr>
<tr>
<td>UNG</td>
<td>Uracil-DNA glycosylase</td>
</tr>
<tr>
<td>WDR76</td>
<td>WD repeat domain 76</td>
</tr>
</tbody>
</table>
Chapter 1
General Introduction

1.1 Epigenetic Alterations

Epigenetic events are heritable and reversible DNA modifications that do not alter the underlying DNA sequence [1, 2]. These modifications broadly include DNA methylation, histone modifications, nucleosome remodeling, and non-coding RNAs (ncRNAs) (Fig. 1). Epigenetic patterns are assembled by “writers” and "erasers", and these enzymes initiate the addition or removal of a modification to the DNA or histones. Epigenetic patterns are recognized by enzymes termed “readers”, which interpret these patterns and catalyze the change in chromatin organization. Lastly, "editors" are enzymes that modulate and monitor epigenetic marks. Moreover, epigenetic modifications are capable of regulating gene expression in a tissue and developmental-specific manner, whereby these silencing mechanisms can become dysregulated in diseased states [3]. Strikingly, these writers, readers, and editors of epigenetic marks are frequently mutated in cancers, alluding to the interaction between genetic and epigenetic reprogramming to ultimately orchestrate gene expression in development and disease.

1.1.1 DNA Methylation

The most widely studied epigenetic modification is DNA methylation (5mC), whereby a methyl group (–CH3) is added to the 5’-carbon of the cytosine base [4, 5]. The methylation reaction is catalyzed by DNA methyltransferases (DNMTs) onto CpG dinucleotides, that is, a cytosine directly followed by a guanine base [6]. DNMT3A and DNMT3B enzymes establish de novo methylation during early development [7], while DNMT1 maintains DNA methylation during mitosis through faithful propagation of symmetrically methylated CpGs, by associating with DNA replication machinery [8, 9]. Nevertheless, all three DNMTs are required for methylation maintenance and embryonic development [10, 11].

Currently, 5mC marks in animals have mainly been studied in the context of CpG methylation, where the role of non-CpG methylation is unknown. The majority of the genome is CpG-sparse and observed to be
methylated. When methylated CpGs occur within endogenous repeats and transposable elements, they are able to repress transcription [12-14]. However, 10% of these dinucleotides exist in regions of dense CpG content, referred to as CpG islands (CGIs). These CGIs are generally unmethylated in normal conditions and are commonly observed at transcription start sites (TSSs) of genes involved in housekeeping and development [15]. When these genes are actively expressed, the promoter region is generally associated with a nucleosome-depleted region (NDR), which is surrounded by histone variant H2A.Z and marked with trimethylation of histone H3 at lysine 4 (H3K4me3) [16]. On the other hand, the repression of these genes can occur by polycomb proteins or several alternate mechanisms [17]. Studies have largely characterized the distribution and function of CpG island methylation, however improved technologies have begun to elucidate the role of methylation outside of CpG islands (non-CGI).

1.1.2 Histone Modifications and Nucleosome Remodeling

Histone modifications are another category of epigenetic marks involved in regulating transcription. The basic unit of chromatin is a nucleosome, which is an octamer consisting of four pairs of core proteins, termed histones. These pairs are composed of identical copies of histones H2A, H2B, H3, and H4 and each nucleosome is further wrapped by a 146 bp section of DNA [18]. The N-terminal tails of histone polypeptides protrude from the nucleosome and are susceptible to over 100 different covalent modifications. These histone modifications are deposited post-transcriptionally and include variations of methylation, acetylation, phosphorylation, and ubiquitination [19]. Histone modifications are involved in transcriptional regulation through either direct organization of chromatin structure into an open or closed state, or by recruiting other factors to the chromatin [19]. Current knowledge is weak concerning the specific function of each modification, however, technological advances have allowed the characterization of several histone marks. For example, the modification of acetylation is described as a transcriptionally active mark when found on histone H3 and H4, whereas the methylation modification is associated both with activation and repression of transcription [19]. That is, methylation of histone H3 at lysine 4 (H3K4) and lysine 36 (H3K36) are active marks, while methylation of H3 lysine 9 (H3K9), lysine 27 (H3K27), and H4 lysine 20 (H4K20) are repressive marks. The writers involved in regulating acetylation and methylation modifications are histone acetylases (HATs) and histone methyltransferases (HMTs), respectively. In contrast, the erasers, which remove these marks, are termed histone deacetylases (HDACs)
and histone demethyltransferases (HDMs) [20]. Importantly, emerging evidence supports that alterations in histone modifications or their writers and erasers, are implicated in carcinogenesis [21].

_Drosophila_ Polycomb group (PcG) and Trithorax group (TrxG) families are repressive and activating complexes, respectively, that are antagonistic in regulating transcription [22, 23]. PcG and TrxG complex target genes are involved in development, stem cell differentiation, and carcinogenesis [24]. Importantly, these complexes can write and read histone modifications. PcG complexes are responsible for writing and reading repressive marks. For example, enhancer of zeste homolog 2 (EZH2) is a member of the polycomb repressive complex 2, or PRC2 (a PcG complex), which catalyzes the trimethylation of histone H3, lysine 27 (H3K27me3). Further Polycomb (Pc), which is a member of the polycomb repressive complex 1, or PRC1 (a PcG complex), binds H3K27me3 contributing to gene silencing by interfering with transcription initiation [25]. Additionally, many PcG and TrxG proteins are able to erase and write histone marks and these proteins can further be placed within the same complex to facilitate rapid switch in active or repressive histone modifications.

Nucleosomes are organized into regularly spaced sets, and connected by linker regions of DNA [26]. Most genomic DNA is occupied by nucleosomes and they are positioned in favorable sites throughout the genome. This positioning occurs in both genic and intergenic regions, in a cell-type specific manner. Commonly, nucleosomes are depleted in enhancer regions, promoters, and transcription termination sites [25-27]. Nucleosome positioning is a dynamic process, and describes the position of a base pair and whether it is placed at the same position on a nucleosome in a fraction of cells. On the other hand, nucleosome occupancy characterizes whether a base pair contains any nucleosomes, regardless of the position. Thus, positioning is a more specific description of how the nucleosome and a base pair interact, while occupancy gives a general pattern of how often the base pair interacts with a nucleosome. Both positioning and occupancy are imperative to proper chromatin organization and resulting biological processes [27, 28]. A nucleosome will restrict the access of DNA-binding proteins to DNA and thus are master regulators of transcription and gene regulation. Meanwhile, an open chromatin state is usually marked with DNase I hypersensitivity sites, which are central markers for cis-regulatory regions such as enhancers, insulators, promoters, and transcription factor binding sites [29]. Interestingly, specific transcription factors termed as “pioneer factors”, can target nucleosomes and recruit additional proteins to open the chromatin. For example, this process provides RNA polymerase II with the ability to access
DNA and initiate transcription [30]. Further, chromatin-remodeling complexes are responsible for the sliding of nucleosomes, thus regulating transcription, DNA replication, and DNA repair [31]. These complexes belong to subgroups termed: SWI/SNF, chromodomain and helicase-like domain (CHD), ISWI, and INO90.

1.1.3 MicroRNAs

MicroRNAs (miRNAs) are endogenous, small, non-coding RNAs that are single stranded and approximately 20-23 nucleotides long. These miRNAs are capable of regulating transcription by either degrading the messenger RNA (mRNA) itself or by inhibiting translation [32, 33]. These processes are initiated when a mature miRNA binds to the 3' untranslated region of a target mRNA with either complete or partial complementarity. As a result, complete complementarity leads to degradation of the target mRNA, as the mature miRNA is incorporated into the RNA-induced silencing complex (RISC). That is, protein argonaute 2 (AGO2), a component of the RISC complex, cleaves the mRNA and renders it non-functional. On the other hand, partial complementarity represses translation, as the mRNA cannot be translated while bound by the miRNA [32, 34]. A single miRNA may have hundreds of target genes, while a single gene can be targeted by many miRNAs. Moreover, miRNAs have roles in development, differentiation, cell cycle regulation, apoptosis, and tumorigenesis [34, 35].
Figure 1. Epigenetic model. An overview of epigenetic mechanisms including DNA methylation, histone modifications, nucleosome remodeling, and RNA-based mechanisms. Diagram by Ho LT and Lo C.
1.2 DNA Methylation Patterns and Effects on Gene Expression

1.2.1 CpG Island Methylation

CpG islands (CGIs) are characterized by an elevated G+C base content, are approximately 1 kilobase (kb) in length, and are generally unmethylated [36-38]. CGIs are associated with approximately 70% of promoter regions in humans, underlining their role in transcription [39]. Interestingly, about 50% of CGIs do not overlap TSSs and are referred to as “orphan” CGIs, as their function is undefined [40]. It has been suggested that these orphan CGIs are observed within intragenic alternative promoters and intergenic regions, and may be sites of transcriptional initiation at imprinted genes and occur in a tissue-specific manner [40-42]. Furthermore, CGI methylation within a gene body has been recognized to correlate with an actively transcribed gene and does not block transcriptional elongation. Evidence suggests that genic methylation may regulate differential expression through alternative promoters or splicing, as 5mC alters the stability of RNA polymerase II elongation [43, 44]. These studies are particularly interesting, as genic CGIs have been reported to interact with modulators that are associated with closed chromatin when observed at TSSs. Modulators were found to be methyl-CpG-binding protein 2 (MECP2) and repressive histone marks, such as trimethylated histone H3, lysine 9 (H3K9me3). These findings allude to the “methylation paradox”, where CGI methylation within promoters represses transcription, while genic CGI methylation leads to active gene expression. Accordingly, evidence suggests that CGI methylation is functional during transcriptional initiation and does not contribute to silencing machinery during transcriptional elongation.

A low percentage of CGI promoters are selected for methylation during normal development and consequently become stably silenced [45-47]. The first mechanism of silencing is proposed as direct blocking of transcription factor binding by 5mC marks [48], while the second mechanism involves 5mC recognition by methyl-binding domain proteins (MBDs) and subsequent recruitment of histone deacetylases (HDACs) to instruct a repressed chromatin state [49, 50]. Most notably, this pattern is found in X-inactivation and genomic imprinting and in these cases, methylation appeared to be a secondary event that “locked” previously silenced genes [51, 52]. It has been observed that the initial silencing
mechanism may involve the presence of a nucleosome, leading to recruitment of DNMT3A, and resulting in de novo methylation [53]. Importantly, another study found that de novo methylation is not deposited in NDRs flanked by active histone marks [7]. Dysregulation of CGI promoter methylation has been widely reported in cancer, providing an avenue for a majority of DNA methylation studies investigating diseased states [54]. Interestingly, many of these promoters have shown previous marking by polycomb repressive complexes (PcG) during development [55, 56]. Indeed, evidence suggests that CGI methylation in promoter regions is not a transcriptional initiating event, but rather promotes a closed chromatin state in tandem with chromatin remodeling complexes, though the mechanism is not well understood.

Several studies have reported that CGI promoter methylation does not correlate well with gene silencing [57-59]. One proposed mechanism for this outcome is that genic CGIs overlap the sites regulating transcriptional initiation of non-coding RNAs (ncRNA) on the anti-sense strand, which in turn regulate the expression of the transcript from the sense strand. Consequently, CGI methylation silences the ncRNA, while the sense transcript becomes derepressed [60]. Although CGI methylation does not completely correlate with expression profiles, it was discovered that regions flanking CGIs are able to distinguish between cell types and remarkably, demonstrate improved correlation with gene expression [61]. The flanking regions are termed CpG shores, which are characterized as 500bp-2kb regions surrounding CGIs and consist of lower CpG density than CGIs.

CpG island methylator phenotype (CIMP) describes a subset of tumors with multiple gene promoters that have acquired CpG island methylation [62]. CIMP was first proposed as a methylation signature in colorectal cancer (CRC) [63], although the term has since been applied to a range of different cancer types, for example, breast and bladder cancers [64, 65]. Even in the context of CRC, CIMP status does not have a standard definition, which poses an obstacle when determining its incidence. That is, it is not clear what distinguishes a CIMP-positive from a CIMP-negative tumor [66]. Currently, the molecular basis for CIMP is not well characterized, perhaps due to the non-standardized definition. However, studies observe that CIMP is influenced by environmental and lifestyle factors [67]. Interestingly, a major link to the molecular basis of CIMP was found through the identification of IDH1/2 and TET2 mutations in CIMP-positive leukemia and glioma tumors, lending to a possible cause of CGI hypermethylation [68, 69]. Interestingly, CIMP status has been associated with specific molecular and clinicopathological features [64], highlighting its ability to stratify patients for diagnostic and therapeutic purposes. Accordingly, CIMP has shown
potential in improving prognosis prediction and response to treatment [70]. Ultimately, it remains to be shown whether CIMP can be applied to all cancer types or if it is necessary to generate a specific CIMP definition unique to each cancer.

1.2.2 Global Hypomethylation in Cancer

A hallmark of human cancers is a global decrease in DNA methylation, or hypomethylation, compared to normal tissues [71, 72] (Fig. 2). Although, the pattern has been long been observed in various cancers [73], the degree of loss is dependent on tumor stage, grade, and individual sample [74, 75]. Additionally, depending on the tumor type, DNA hypomethylation can increase with tumor progression [73, 76]. Remarkably, hypomethylation has been reported in normal tissues adjacent to neoplastic cells suggesting that this is an early event in carcinogenesis [77, 78]. Accordingly, hypomethylation has potential as a prognostic and survival marker in numerous cancers including gastric, ovarian, and microsatellite instability (MSI)-high colorectal cancer tumors [79-81].

Hypomethylation is commonly observed in repeat elements, and these are the major contributors to loss of methylation in cancer. The regions typically affected are tandem and interspersed repeat elements [82-84]. That is, hypomethylation of long interspersed nuclear element-1 (LINE-1) and Arthrobacter luteus (Alu) interspersed repeat elements are often observed and have been described in a range of cancers [85-87]. Whereas hypomethylation at tandem repeats is seen in centromeric and juxtacentromeric satellite DNA in breast and ovarian cancers [88, 89]. Occasionally, interspersed repeat elements can contain promoter regions, which in turn, may regulate expression of corresponding genes [90, 91]. Notably, cancer hypomethylation is a result of DNA demethylation and has not previously been established in a cancer stem cell model [92].

The introduction of genome-wide methylation studies has allowed the identification of hypomethylation within gene bodies and intergenic regions surrounding genes associated with metastasis. Additionally, differential signatures of hypomethylated regions have been reported both across normal tissues types, and when comparing cancer and matched normal tissue [93, 94]. These differential sites occur in both shorter regions and longer blocks throughout the genome [12, 95]. In general, hypomethylated regions
correlate with increased gene expression, for example, imprinted genes impacted in cancer [96].
Currently, it is debatable whether oncogenes are up-regulated through genic hypomethylation, as studies
reporting this correlation did not investigate methylation at promoter regions. Further, many oncogenes
found to be hypomethylated did not contain CGIs in their promoters [97, 98]. Additionally, it has been
reported that promoter regions contain a low proportion of both tissue-specific and cancer-associated
differential hypomethylated regions. It has been shown that enhancer regions, which incur a loss of
methylation, overlap transcription regulatory factor binding sites associated with differentiation [99]. The
subsequent binding proteins are termed “pioneer factors”, and are involved in exposing an enhancer
previously repressed by closed chromatin, eventually recruiting effector transcription factors. Pioneer
factors have been implicated in carcinogenesis, for example FOXA1 is associated with prostate cancer
[100, 101].

In addition, there are broad regions of hypomethylation associated with an increase in copy number,
allowing increased gene expression [102]. It has been suggested that these broad regions reflect higher
order chromatin organization that is dependent on several factors such as: the frequency and spacing of
the broad regions, the CpG content, nucleosome density, and groups of co-regulated genes. This leads to
the cancer phenomenon where longer regions of acquired hypermethylation are located next to a region
that has undergone hypomethylation [103]. These two regions are divided by a strong demarcation and
illustrate the pattern of DNA demethylation. Interestingly, cancer-associated DNA demethylation can
appear to overshoot the target region, resulting in a “passenger event” of hypomethylation [104]. The
cause of “passenger events” could be due to relaxed specificity in the DNA demethylation pathway,
however the mechanism is not currently well characterized. Finally, it has been well established that DNA
hypomethylation in cancer leads to chromosomal instability and tumor development, as proven by
murine models expressing hypomorphic DNMT1 [105, 106].

1.2.3 Locus-specific Hypermethylation in Cancer

An important driver in carcinogenesis is aberrant gain in DNA methylation, in a locus-specific manner.
More specifically, cancer-associated hypermethylation is most commonly characterized in CGI promoter
regions of tumor suppressor genes (TSGs), leading to their repression [107] (Fig. 2). The first report of CGI
promoter hypermethylation was in *retinoblastoma 1 (RB1)* and this was observed to inactivate the gene [108]. Since then, numerous tumor suppressor genes have been discovered to incur hypermethylation in the promoter CGIs, such as *cyclin-dependent kinase inhibitor 2A (CDKN2A)*, *mutL homolog 1 (MLH1)*, and *breast cancer 1 (BRCA1)* [109-111]. Evidence supporting the ability of methylated CGI promoters to silence TSGs is substantial. One major support is the consequential activation of repressed genes following 5-aza-2'-deoxycytidine (DAC) treatment, a demethylating agent [112]. Although promoter CGI methylation is likely a secondary event in the gene repression model, its removal is capable of re-establishing gene expression. Additionally, there are reports of other genes, which are inactivated by CGI promoter hypermethylation whose functions involve DNA repair, cell-cycle regulation, and apoptosis [113, 114].

Recent genome-wide analysis has revealed that a majority of hypermethylated CGI promoters in cancer were previously methylated in their non-neoplastic state. Further, these genes exhibited a low expressing or inactive state prior to cancer [115, 116]. This leads to the hypothesis that these are “passenger genes” in that they are not involved in the initiation of carcinogenesis. Meanwhile, the subsets of genes that acquire aberrant methylation in cancer and were previously expressed in normal tissue, are “driver genes” which directly contribute to disease progression [104]. Indeed, hypermethylation could inactivate genes whose repression allows cancer cells to survive [117]. Notably, these findings find a similarity between CGI methylation during cancer and normal development, that is, hypermethylation occurs after the initial event of gene repression.

On the other hand, genes marked and repressed by polycomb repressive complexes (PRCs) in embryonic stem cells (ESCs), have a tendency to become hypermethylated in cancer [118]. As PRCs target lineage-specific genes in ESCs, it is hypothesized that in cancer, PRC target genes contribute to a stem-cell-like state where differentiation is inhibited [119]. Consequently, hypermethylation of promoter CGIs would correlate with increased cancer aggressiveness. This pattern may occur through PRC recruitment of DNMTs to the promoter CGI. In fact, it has been reported that DNMT3A and 3B can bind EZH2, a component of the PRC2 complex [120]. However, this mechanism has been shown to be tissue-specific [121]. In addition, H3K27me3, a repressive histone mark associated with PRC2, and DNA methylation do not usually co-localize at promoter regions [122].
A final model is of particular interest, as it supports the association of CIMP-positive tumors with improved disease prognosis, as reported in breast cancer [123]. This hypothesis suggests that genes are silenced in an effort to inhibit cancer progression. Evidence suggests that metastatic cells require activation of a set of genes, rather than repression [124]. Moreover, this model lends to the notion that genes previously repressed and hypermethylated in normal tissues, are already acting to protect from cancer development.

The mechanism behind CGI hypermethylation is not well characterized and several models are being investigated. One model, (previously mentioned above) involves recruitment of DNMT1 to the CGI promoter by PRC2 complex. Another hypothesis suggests that instead of being repressed, these targeted CGI promoters are poised, using histone marks for either gene activation or repression. Further, the resulting transcriptional pattern is determined during differentiation [125]. On the other hand, histone marks may protect promoter CGIs from DNA methylation; for example, H3K4me3 and H2A.Z are inversely correlated with 5mC marks in normal cells [126, 127]. It has been shown that DNMTs are unable to bind histones bearing H3K4me3 [128], while the mechanism behind H2A.Z-mediated methylation protection is unknown.

Promoter CGI hypermethylation may occur due to defects in the DNA demethylation pathway (described in Section 1.3.1). Various cancers have reported a reduction in expression of TET proteins, which are the enzymes responsible for initiating this pathway [129, 130]. Thus, DNA methylation marks could be aberrantly maintained in disease state, however it is not well understood why disrupted TET function would target specific genes.

Another hypothesis supports that sequence-specific CGI methylation is determined by transcription factors which have shown binding affinity for DNMT3A and 3B. However, motifs found to be associated with hypermethylated promoters have not been well reproduced [131]. Another layer of protection suggests that actively transcribed genes are resistant to methylation, although the relationship may be correlative rather than causative. That is, increased transcription levels might only be associated will lower occurrence of CGI methylation, and not confer a protection against it.
Clearly, the function and mechanism of CGI promoter methylation is a complex phenomenon. Remarkably, although the biological basis of cancer-associated hypermethylation is unclear, these aberrations are currently a widely promising option for epigenetic biomarkers and therapeutic targets as their diagnostic and prognostic potential has been described in a variety of cancers.
Figure 2. DNA methylation patterns in normal and cancer cells. Global loss of DNA methylation marks (left side) and gene specific CpG island hypermethylation (right side) generally lead to an increase in gene expression and a gene silencing, respectively. These aberrant methylation events are hallmarks of cancer cells. Diagram by Olkhov-Mitsel E.
1.3 DNA Hydroxymethylation Patterns

1.3.1 DNA Hydroxymethylation and DNA Demethylation Pathways

As early as 1952, it was reported that 5mC is converted to 5-hydroxymethylcytosine (5hmC) marks in T-bacteriophage and mammalian tissues [132, 133]. However, only recently, it was re-discovered that 5mC marks are oxidized into 5hmC by ten-eleven translocation (TET) proteins in mammalian DNA [134]. This is a remarkable finding as it suggests an initial step in the process of DNA demethylation, whereby 5mC returns to an unmodified cytosine base. In the passive demethylation model, 5hmC is a transient intermediate, as it cannot be read by DNMT1 leading to loss of 5mC marks after several DNA replications [135]. In contrast, the active demethylation model suggests that 5hmC is a stable mark that can be deaminated into 5-hydroxymethyluracil (5hmU) and removed by thymine DNA glycosylase (TDG) [136]. On the other hand, 5hmC may be further oxidized by TET enzymes into 5-formylcytosine (5fC) and then 5-carboxylcytosine (5caC), to eventually be removed by TDG-mediated base excision repair (BER) pathways [130] (Fig. 3). It has since been suggested that active DNA demethylation likely occurs through the latter pathway, as the deamination enzymes activation induced deaminase (AID) and apolipoprotein B editing complex proteins (APOBEC), show higher affinity for unmodified cytosines and further, their over expression does not increase the levels of 5fC and 5caC [137].

It was initially proposed that 5mC is a stable epigenetic mark, however its dynamic nature was revealed with reports of global demethylation stages during embryonic development [138] and global hypomethylation in human cancers [139]. Mechanisms have been postulated regarding both passive and active DNA demethylation pathways, although the passive pathway is more widely accepted [140, 141]. Support for the active pathway is being generated as overexpression of 5hmC glycosylases [142] and TET genes [143] lead to abundance in oxidated derivatives of 5mC (5hmC, 5fC, 5caC) and further demethylation. Overall, there is no consensus on the mechanism of this pathway and it remains to be shown how DNA demethylation occurs in vivo [141].

Currently, the biological role of 5hmC is undefined. Studies have shown 5hmC enrichment in gene bodies of highly active genes, TSS of inactive genes [144], and other cis-regulatory regions, such as enhancers and transcription factor binding sites [145, 146]. Coupled with observations that 5hmC is not recognized
by methyl-CpG binding proteins (MBDs) [147] suggests that 5hmC regulates transcription either by blocking MBD proteins or by modifying chromatin accessibility. In addition, 5hmC has distinct readers, or binding partners, from 5mC such as UHRF1 [148], MBD3/NURD complex [149], and several other proteins involved in transcriptional regulation and DNA repair [150]. Interestingly, 5hmC marks are preferentially bound by WD repeat domain 76 (WDR76), which interacts with a DNA helicase named helicase, lymphoid-specific (HELS) [151]. The HELLS protein has been shown to regulate DNA methylation [152], thus suggesting the ability of 5hmC to contribute to gene transcription and moreover, lending to its role as a stable epigenetic mark.

The global distribution of 5hmC is highly varied across normal tissues, unlike 5mC marks, which are generally stable between tissues. The highest content of 5hmC is found in brain and embryonic stem cells (ESCs), and the lowest levels are found in glandular tissues and blood [153]. Additionally, it has been shown that 5hmC levels are reduced when cells are introduced to cell culture [153]. These findings suggest that 5hmC is part of the passive DNA demethylation pathway, in that cells with low proliferative rates, such as brain have high 5hmC content, while global levels of 5hmC begin to reduce in cells with significant proliferation, such as blood.

Interestingly, it has been reported that high levels of 5hmC are important for undifferentiated ESCs [154] and that several solid tumors, such as breast, colon and prostate, show depletion of global 5hmC profile compared to normal tissue [129, 155, 156]. These findings suggest that 5hmC negatively regulates tumorigenesis. Further, as 5hmC can block DNMT1 from depositing 5mC marks [135], this regulation may be severely suppressed in cancer, allowing dysregulated methylation signatures and therefore abnormal gene expression. Additionally, as 5hmC marks are greatly reduced in differentiated cells, it could potentially develop into a cancer therapeutic. Another study showed that non-genotoxic carcinogen (NGC) exposure, which is used to initiate a mouse model for liver tumors, led to activation of genes marked with 5hmC at promoter-proximal regions [157]. This suggests that 5hmC contributes to transcriptional regulation and further indicates that 5hmC profiling may be used to monitor responses to drugs or other treatments.

Recently, the 5mC oxidative derivatives 5fC and 5caC were mapped in mouse ESCs. They showed extremely low enrichment, likely due to their transient nature [158, 159]. Through knockdown studies of
TDG, it was found that 5fC and 5caC marks are enriched in gene bodies, proximal-promoter regions, and enhancers. Moreover, this knockdown model supports an active DNA demethylation model.
Figure 3. DNA demethylation pathways. Proposed (A) passive and (B) active models for DNA demethylation. Deamination pathway is faded as evidence is weak for this mechanism [137]. C: unmodified cytosine, 5mC: 5-methylcytosine, 5hmC: 5-hydroxymethylcytosine, 5hmU: 5-hydroxymethyluracil, 5caC: 5-carboxycytosine, UHRF1: ubiquitin-like, containing PHD and RING finger domains 1, DNMT1: DNA methyltransferase 1, TET: Ten-eleven translocation proteins, AID: activation induced deaminase, APOBEC: Apolipoprotein B Editing Complex protein, TDG: thymine-DNA glycosylase, BER: base-excision repair, SMUG: single-strand-selective monofunctional uracil-DNA glycosylase, UNG: uracil-DNA glycosylase. Diagram by Ho LT and Lo C.
1.3.2 TET Proteins

TET proteins are a family of mammalian dioxygenase enzymes, consisting of TET1, TET2, and TET3 [160, 161], which have common catalytic domain structures dependent on 2-oxoglutarate and iron (II) suggesting they have similar enzymatic activity [162]. However, each TET consists of unique features [163, 164] indicating distinct binding-affinities and functions that are currently being explored.

TET enzymes are able to sequentially oxidize 5hmC into additional oxidative forms of 5mC. These products, known as 5fC and 5caC, are distributed in far less quantities than 5hmC and their functions are still unknown [130, 158, 159]. Evidently, revealing the catalytic roles of TET enzymes will contribute to understanding the process of DNA demethylation.

The mechanisms involved in targeting TET proteins to specific genomic loci have not been completely described. However, it is apparent that each protein has specific recruitment complexes; shows involvement on active, repressive, and bivalent regulatory regions; and is dependent on developmental stage or tissue type [149, 165-167]. In addition, the co-localization of TET proteins with 5mC and 5hmC marks, at specific loci, remains to be fully revealed.

Expression levels of each TET protein are distinct across human tissues. For instance, TET1 expression is highest in ESCs and may be essential in self-renewal [160], TET2 is the most highly expressed TET protein in many normal tissues [168], and TET3 shows ubiquitous expression in normal tissues [169]. In addition, it has been reported that TET proteins are depleted in human solid tumors, including prostate, compared to normal tissue [129], although the underlying mechanism is undefined. Interestingly, it has been shown that TET2 is frequently mutated in myeloid malignancies [170], is implicated in melanoma progression [171], and may be a tumor suppressor gene [172].

It has been suggested that the inhibition of TET co-factors may contribute to the significant loss of 5hmC marks in cancer. TET enzymes are dependent on 2-oxoglutarate (2-OG), which is generated from the oxidative carboxylation of isocitrate. Further, this reaction is catalyzed by isocitrate dehydrogenases
Studies have revealed that both IDH1 and IDH2 genes incur gain-of-function mutations in various cancers [174, 175]. These mutated enzymes catalyze the production of 2-hydroxyglutarate (2-HG), which can antagonize 2-OG-dependent reactions [176]. Ultimately, mutations in IDH genes can inhibit TET enzyme function and consequently, reduce the conversion of 5mC to 5hmC marks. Studies are ongoing to reveal the role of the Krebs cycle in epigenetic regulation.

1.4 Current Technology in the Detection of DNA Methylation and DNA Hydroxymethylation

1.4.1 Genome-wide DNA Methylation Detection Techniques

Before the introduction of next-generation sequencing (NGS) in 2005, genome-wide methylation studies were performed using microarray technology [177]. These methods were largely based on sodium bisulfite modification of DNA [178]. This technique converts all unmodified cytosines to uracil bases and subsequent Sanger sequencing or microarray analysis displays the uracil bases as thymines. Meanwhile, 5mC marks are read as cytosines. A major pitfall of bisulfite modification is the inability to discriminate between 5mC and 5hmC marks, in that both marks are resistant to modification and are eventually read as cytosines post-modification and analysis [179]. Nonetheless, bisulfite sequencing is the “gold standard” for DNA methylation detection as it is easily detectable by sequencing and remains the only tool to generate a single-nucleotide resolution of the methylome.

Without this disadvantage, microarrays have their own inherent limitations. Microarrays consist of probes targeting specific regions in the genome. In the context of methylation, such arrays were designed to map CpG island methylation and further, probes are synthesized based on previous knowledge about the regions being targeted. Thus, only a small portion of the genome can be analyzed [177, 180]. The introduction of high density microarrays, such as the Illumina Infinium 450K BeadChIP [181], provided greater coverage of CpGs, however the power of next-generation greatly exceeds this technology.

There are approximately 28 million CpGs in the genome and NGS technology enables close to full coverage of these sites. For example, one study performed shotgun bisulfite sequencing of stem cells and fibroblasts and was able to detect 94% of all cytosine bases [182]. The methods developed to detect
methylation using NGS are affinity/enrichment-based using protein domains or antibody detection, methylation-sensitive restriction enzymes, and bisulfite modification.

Affinity/enrichment-based techniques include those utilizing proteins that recognize 5mC marks and antibodies generated to bind methyl groups. One example of the former subgroup is termed MBD-seq [183]. This technique captures methylated DNA that has been randomly fragmented using the methyl-CpG binding domain of MBD proteins. A major benefit of MBD enrichment is that the genomic DNA captured is double-stranded, and thus the integrity of the DNA is maintained. In addition, MBD2 is specific for 5mC marks and can discriminate from other modified cytosines. However, MBD-seq requires greater amounts of input DNA, and pooling of samples may be required [184]. In terms of antibody-capture, the technique is termed MeDIP-seq or methyl-DNA immunoprecipitation combined with NGS [185], and its strength is the capacity to detect regions of low-density CpGs and thus, is not biased towards CpG islands. In addition, MeDIP requires low amounts of starting DNA [186]. However, antibodies are known for non-specific binding as the anti-5mC antibody can bind other methylated bases. Lastly, MeDIP requires denaturation of genomic DNA to single strands, which may detect methylation that is allele-specific. This has posed a problem for NGS technology as adaptors would not ligate to single stranded DNA, however, this issue has been circumvented and a pipeline was developed to ligate adaptors prior to denaturation [186].

The enrichment efficiencies of MBD-seq and MeDIP-seq have been compared in multiple studies, as these two techniques both result in moderately high resolution of the genome, do not modify any nucleotides, and are affordable alternatives to BS-seq [187-190]. Most studies observed that although MBD-seq generated a greater number of sequencing reads and higher relative enrichment, it was biased towards regions of higher CpG density. Meanwhile, MeDIP-seq was biased towards regions of lower CpG density. Interestingly, one study found these techniques to be 99% concordant using binary methylation calls, though another study suggests that the techniques did not always agree when discovering differentially methylated regions. Ultimately, both methods enrich for regions of methylated DNA, and thus bioinformatic analysis reveals relative enrichment for methylation rather than absolute. While these methods cannot substitute for the high resolution mapping from BS-seq, they were found to be effective in determining general enrichment patterns throughout the genome.
Methylation-sensitive restriction enzyme-based methods, termed MRE-seq [191], involve the use of restriction enzymes, such as methylation-sensitive enzyme HpaII and methylation-insensitive enzyme MspI. This technique detects a higher resolution of the methylome compared to affinity-based methods [192]. A limitation of this type of detection is that only regions of the genome that contain the specific recognition sequences or cut sites can be interrogated. In addition, the amount of enrichment can depend on the efficiency of the digestion. One group combined MeDIP- and MRE-seq in order to utilize the benefits of both techniques, and were able to define high-resolution regions of methylation that were allele-specific [188].

Reduced representation bisulfite sequencing (RRBS) is a technique that combines restriction enzyme digestion with bisulfite sequencing [193]. This allows for detection of CpG-dense regions and subsequent single-nucleotide resolution of these regions. Thus, higher resolution of the genome is attained with a reduced cost, although the genome-wide coverage is lower than enrichment techniques [184]. In addition, RRBS can detect methylation in a small sample, which makes this technique appealing for tumor tissue analyses. Lastly, as RRBS is based on bisulfite modification, it is not able to distinguish between 5mC and 5hmC marks.

### 1.4.2 Global and Locus-specific Detection Techniques of DNA Hydroxymethylation

Since its re-discovery in 2009, technology has been rapidly developing in order to detect DNA hydroxymethylation marks throughout the genome. The first studies developed detected 5hmC marks at a global level. These techniques include antibody detection, high-performance liquid chromatography combined with mass spectrometry (HPLC-MS), and immunohistochemical (IHC) analyses. To better interrogate 5hmC marks, genome-wide techniques have been developed which generally modify the 5hmC base prior to capture followed by next-generation sequencing. These locus-specific detection methods can be categorized into three subgroups: affinity/enrichment-based, restriction enzyme digestion-based, or single-nucleotide resolution methods.
Antibody detection can be performed through two different methods, both using a primary antibody to specifically recognize the hydroxymethylated group on the cytosine base. The first technique is by dot blotting [194], which subjects total genomic DNA to the primary antibody and fixes the complex to a membrane, similar to a southern blot. Dot blotting is a simple, inexpensive technique, however it is only semi-quantitative. Alternatively, an enzyme-linked immunosorbent assay (ELISA) subjects single stranded DNA to a well coated with primary antibody [195]. Antibody-based detection can further be combined with next-generation technology in a technique termed hMeDIP-seq [195], where single stranded DNA is first enriched using primary antibody and the subsequent enriched fraction is sequenced. Major disadvantages of antibody detection are non-specific antibody binding and a bias towards high-density 5hmC regions.

High-performance liquid chromatography combined with mass spectrometry (HPLC-MS) is currently the most accurate technique to detect global levels of 5hmC marks [195]. DNA is first hydrolyzed into nucleosides and passed through a liquid chromatography column to elute differently weighted bases and further are detected using absorption qualities. Thus, MS is able to differentiate between 5hmC, 5mC, and unmodified cytosine. Although HPLC-MS is sensitive and quantitative, it is an expensive technique that requires extensive training and expertise. Still, HPLC remains the “gold standard” for global detection and quantification of 5hmC.

Immunohistochemical (IHC) analysis became an attractive technique as it is capable of staining single cells and can visualize 5hmC patterns amongst a tissue cell population [134]. IHC uses an anti-5hmC antibody and could ultimately transition into a diagnostic technique. For example, IHC staining could differentiate between normal and neoplastic cells from a biopsy sample or archival tissue. Limitations of IHC are the semi-quantitative nature and the need for standardized preparation, in that this technique is very sensitive.

Most affinity-based methods for mapping the hydroxymethylome involve the protection of 5hmC prior to capture. One method, termed hydroxymethyl selective chemical labeling (hMeSeal) [196], uses the T-phage β-glucosyltransferase (B-GT) enzyme to glucosylate 5hmC marks, which is then conjugated to an azide group. The azide is recognized by biotin and subsequently captured with a streptavidin bead. Alternative techniques using the B-GT enzyme are: glucosylation, periodate oxidation, biotinylation or GLIB [197],
which involves oxidizing 5hmC post-glucosylation to yield aldehydes; and J-binding protein 1(JBP-1)-capture [198], which utilizes the JBP-1 dioxygenase protein to bind 5hmC marks post-glucosylation. These affinity-based methods can further be assessed for genome-wide mapping of 5hmC marks using next-generation sequencing or through locus-specific detection using RT-qPCR. Advantages of affinity-based methods are specificity and efficient enrichment of DNA fragments. However, quantification is relative and the techniques require a large amount of starting DNA.

Restriction enzyme-based technique can be performed by a method termed glucosylation-coupled methylation-sensitive quantitative PCR (glucMS-qPCR) [199]. The genomic DNA is first subjected to glucosylation from B-GT enzyme. By modifying the 5hmC mark, hydroxymethylated fragments of DNA are protected from restriction enzyme digestion, for example, using MspI, which is a DNA methylation insensitive enzyme. A second reaction is performed using HpaII, which is a 5mC sensitive enzyme that cannot distinguish between 5mC and 5hmC marks. The difference in DNA content, detected between the two reactions, is considered an absolute quantification of 5hmC at specific loci. While this method allows absolute quantification, it is biased towards 5hmC occurring at CCGG cut sites, recognized by the enzymes.

In order to detect 5hmC marks at a single-nucleotide resolution, variations of bisulfite sequencing (BS-seq) were developed. One technique is termed oxidative bisulfite sequencing (oxBS-seq) [200] and uses KRuO4 to oxidize 5hmC marks to 5fC, while all other cytosine bases (unmodified cytosine, 5mC, 5fC, 5caC) are resistant to oxidation. Subsequently, the reaction is subjected to bisulfite modification, which converts all bases, except 5mC, to uracils. Post-sequencing, all modified and unmodified cytosine bases are read as thymines, while 5mC is the only base to be read as a cytosine. In order to determine the distribution of 5hmC marks, a second reaction of classical BS-seq must be performed, in which the resulting sequence allows both 5mC and 5hmC to be read as cytosines. The sequences from oxBS-seq and BS-seq are subtracted in order to identify the location of 5hmC marks. Recently, another novel technique was developed and termed TET-assisted bisulfite sequencing (TAB-seq) [201], which exploits the ability of TET proteins to oxidize all modified cytosines to an unmodified state. Prior to this reaction, 5hmC marks are glucosylated and protected from TET-mediated oxidation. That is, following bisulfite treatment and sequencing, only 5hmC marks are read as cytosines. Although these are attractive techniques to detect absolute levels at loci enriched for 5hmC, they are still laborious, expensive, and require high sequencing.
depth. Additionally, oxBS-seq can lead to extensive DNA degradation and requires a large amount of DNA. Thus, TAB-seq is a milder approach, however, TET enzymes are very sensitive and require caution to maintain stability.

Clearly, diverse 5hmC detection methods exist and the optimal technique depends on the plan of study. Both global and genome-wide detection methods are available, yet the low enrichment of 5hmC marks in many tissue types may become the limiting factor, rather than the technique itself.

1.5 Prostate Cancer

1.5.1 Incidence, Survival, and Risk Factors

Prostate cancer (PCa) is the most commonly diagnosed malignancy in Canadian men, and the third leading cause of death due to cancer [202]. It is estimated that 23,600 men will be diagnosed with the disease in 2014, while the expected number of deaths is 4,000. However, rates of mortality have been declining and this is likely due to improved treatment options for patients and earlier detection of the disease. It is commonly stated that most men will die with prostate cancer rather than from the disease, and this is exemplified by the diagnostic age to average around 60 years, while cases of death generally occur over the age of 80 [202]. Prostate cancer is a slow growing tumor and the progression of the disease can last 10 to 15 years [203]. Importantly, the survival rate of patients diagnosed at localized stage prostate cancer is approximately 95% [202], while this number drastically drops to less than 30% when considering patients diagnosed at advanced stage cancer. This lends to the fact that treatment options for the former are curative, while treatments of the latter are palliative (discussed in Section 1.5.2). The greatest risk factor for developing PCa is age, with an incidence rate of 80% by the age of 80 [204]. Other common risk factors are family history, African-American ethnicity, and high fat diet [205].

1.5.2 Screening, Prognosis, and Treatment

Prostate cancer is categorized as an adenocarcinoma and generally manifests from high-grade prostatic intraepithelial neoplasia (HGPIN), as HGPIN patients have a 50% risk of developing PCa in a 10-year period
In the early 1990’s, a rise in the incidence of prostate cancer occurred largely due to the introduction of serum-based prostate-specific antigen (PSA) testing [208]. This test provides a simple indicator for the disease, in that patients with PCa often exhibit elevated serum levels of PSA. PSA is thought to be a biologically relevant marker as it is defined as a glycoprotein, encoded by the kallikrein-3 (KLK3) gene, which is normally secreted by the prostate gland in low amounts [209]. PSA testing was initially accepted as an accurate early detection biomarker, however many disadvantages have been associated with PSA. For instance, there is no distinct PSA value that determines whether a patient is prostate cancer positive or not. That is, a high PSA value does not always indicate disease and there are patients with confirmed disease that show low PSA levels [210]. Additionally, PSA is secreted in non-neoplastic states such as benign prostatic hyperplasia (BPH) and prostatitis [211, 212]. Consequently, PSA monitoring has led to over-diagnosis and over-treatment of the disease. Moreover, PSA is not able to differentiate a slow-growing or indolent cancer from a tumor that will become aggressive. This poses as a significant pitfall as most patients diagnosed with PCa are older and treatment might be unnecessary.

To combat the issues of over-diagnosis and consequential over-treatment of PCa patients, active surveillance and watchful waiting programs have been introduced to diagnosed patients who require different risk management regimes. Active surveillance [213] involves close monitoring of the patient in order to detect any unfavorable changes that may occur and perhaps indicate aggressive disease. The goal of this program is to detect changes early and treat a patient in a curative manner. In contrast, watchful waiting [213] is a program established for the subset of patients whom are diagnosed at an older age and may be dealing with other health complications. These men will not be treated for PCa until symptoms arise, which could be at an advanced stage of the disease. Thus, treatment goals are palliative.

Additionally, the use of prostate cancer antigen 3 (PCA3) testing has been utilized as another potential biomarker for the disease [214]. It is a urine-based test to detect PCA3, a non-coding RNA, which is secreted by prostate cancer cells. PCA3 is shown to be over-expressed in prostate cancer and studies have shown that the use of both PSA and PCA3 testing can accurately predict a positive biopsy. PCA3 testing is not yet performed in most Canadian clinics, however it is commercially available.

Current diagnostic tools for the disease involve the Gleason grading system, digital rectal exam, and needle biopsies. The Gleason grading system [215, 216] is utilized by a pathologist, which allows
classification of tumor tissue based on its architectural pattern. Using a point scale, a tumor is assessed for the two most common grades or patterns. Subsequently, the sum of these grades is termed the Gleason Score (GS) of the tumor where, in general, a GS of <6 is low-grade disease, GS7 is intermediate, and GS>8 is high grade. GS is currently the best prognostic indicator of the disease. However, there are issues in the reproducibility of scoring as these vary between pathologists [217]. The digital rectal exam [218] is an internal exam of the rectum performed by finger insertion and palpitation of the prostate gland for abnormalities. Although the exam itself is painful and prone to false positives, it remains a reliable diagnostic tool prior to biopsy, especially for cases that secrete low levels of PSA. Finally, transrectal ultrasonography (TRUS) guided needle biopsy is a method to assess tumor size and obtain patient tissue samples, where up to 12 cores of tumor can be extracted. This is a painful procedure that can lead to infection, however a positive biopsy is required for a full diagnosis of the disease [219, 220].

When patients are diagnosed at the local disease stage, cancer is confined to the prostate gland. Many cases can be cured using radical prostatectomy, which is a safe and widely used surgical procedure to remove a tumor [221]. Alternatively, radiotherapy can be performed [222]. If cases progress to an advanced or a metastatic stage, the main therapy option is hormonal or androgen deprivation therapy (ADT), which targets androgen production, as these are required for prostate growth and subsequent disease progression [223]. This therapy can be achieved through the use of luteinizing hormone-releasing hormone (LHRH) agonists [224]. Additionally hormone therapy can block the androgen receptor (AR) using androgen antagonists or non-steroidal anti-androgens (NSAA) [225, 226]. Most patients respond to treatment and tumor size will reduce for a period of time, while the tumor is androgen-dependent. However, prostate cancers can become unresponsive to treatment, and resume growth in an androgen-independent manner. This stage of the disease is termed castration-resistant prostate cancer (CRPC) [227, 228]. Treatment options include additional hormonal therapy or chemotherapy, although the intent is palliative [229, 230].

1.5.3 Genetic Alterations

Genomic alterations have shown to be important contributors to prostate cancer development. These can include somatic mutations, gene deletions or amplifications, and chromosomal rearrangements.
Frequently reported genetic alterations are the chromosomal rearrangements between PSA-regulated gene *transmembrane protease, serine 2* (*TMPRSS2*) and *v-ets erythroblastosis virus E26* oncogenes (*ETS*) family of transcription factor genes [231]. *TMPRSS2* is an androgen-responsive gene that is highly expressed in both normal and cancerous prostate epithelium [232]. ETS members interact with other transcription factors to regulate proliferation, differentiation, angiogenesis, and apoptosis [233]. The most common genetic rearrangement observed in prostate cancer involves the fusion between *TMPRSS2* and the *ERG* gene, a member of the ETS family, which is reported in over 50% of cases [234]. These genes are found on chromosome 21q22 and this particular fusion (in total there are seventeen reported between *TMPRSS2* and *ERG*) is the only one to encode a protein, though its function is unknown. However, as *TMPRSS2* is regulated by androgens, which are up-regulated in PCa, *ERG* oncogene is consequently overexpressed. *TMPRSS2-ERG* poses as a potential biomarker for the disease as it occurs in early PCa development and is not reported in BPH or other benign prostate diseases. Notably, *TMPRSS2-ERG* is detected in 20% of PIN lesions and evidence suggests that the fusion does not participate in driving carcinogenesis, in that other mutations are required. Interestingly, in some *TMPRSS2-ERG*-positive tumors, rearrangements occur within *AR* and *ERG* DNA-binding sites, suggesting that the translocation is linked to transcriptional activities of these two genes [235].

Androgen signaling is a major pathway involved in normal development and PCa carcinogenesis [236, 237], whereby many genetic alterations have been associated with members of this axis. The action of androgens occurs through the binding of androgen receptor (AR), which belongs to a superfamily of nuclear receptors. When AR is unbound it is found in the cytosol, in a heterocomplex consisting of heat shock proteins. Once the androgen ligand has bound, AR dissociates from the complex, dimerizes, and translocates to the nucleus [238]. Here it functions as a transcription factor, interacting at AR-responsive elements throughout the genome. This interaction recruits RNA polymerase II and numerous co-factors, ultimately activating transcription of target genes [239]. Various genetic alterations involved in AR signaling are associated with prostate cancer progression. Those include amplification of the *AR* locus leading to its over-expression, gain-of-function mutations, and splice variants of *AR* gene [239]. These alterations are of particular concern, as major treatment options for advanced disease and CRPC target androgens and the androgen receptor. Thus, a poor response to treatment could stem from a genetic mutation. For instance, the amplification of *AR* gene is associated with hormone therapy failure after the tumor has undergone androgen ablation therapy [240]. These patients could benefit from androgen
blockage as the AR in these tumor cells becomes highly sensitive. In addition, AR somatic mutations have been reported, including missense mutation T877A, which allows ligands other than 5α-dihydrotestosterone (DHT) to activate AR [241]. Patients with this mutation could halt any androgen antagonist therapies, as this treatment could lead to “withdrawal syndrome”.

Other genetic alterations frequently observed in prostate cancer involve the following tumor suppressor genes: *tumor protein 53 (p53)* missense mutations [242], *phosphatase and tension homolog (PTEN)* and *NK3 homeobox 1 (NKX3.1)* deletions [243, 244], and *v-myc myelocytomatosis viral oncogene homolog (c-myc)* amplification [245].

### 1.5.4 Epigenetic Alterations

Increasing evidence suggests that the development of prostate cancer involves interplay between both genetic and epigenetic factors. It has been shown that epigenetic alterations, or epi-mutations, are more frequent than genetic mutations in the context of this disease [246]. The most thoroughly characterized epigenetic mark in prostate cancer is DNA methylation. Furthermore, these modifications have been suggested as promising biomarkers for the disease, for diagnostic and prognostic purposes [5]. A majority of the aberrant methylation signatures detected in prostate cancer occur in CGIs of promoter regions, that is, these regions are hypermethylated compared to normal tissues [247]. Over 90 genes have been identified as hypermethylated in prostate cancer, including *adenomatous polyposis coli (APC)*, *tumor suppressor protein RDA32 (RASSF1A)*, and *homeobox D3 (HOXD3)* [248, 249], and have been shown to be involved in cell cycle control, hormone response, DNA repair, tumor invasion, and apoptosis. The leading example of CGI promoter hypermethylation occurs on *glutathione S-transferase pi 1 (GSTP1)* gene [250], which leads to its inactivation and has been described in over 35 studies [251]. *GSTP1* is involved in cell detoxification and has been shown to play a protective role against DNA damage [252]. As a promising biomarker for the disease, *GSTP1* has shown a higher specificity rate than PSA (over 90%) and is able to differentiate between PCa and BPH or PIN lesions [253, 254]. Further, the level of *GSTP1* hypermethylation has positively correlated with aggressiveness of disease and predicts recurrence post-treatment [255]. On the other hand, global hypomethylation, associated with advanced metastatic PCa, is reported to promote chromosomal instability [256, 257]. In addition, approximately 50% of PCa cases have shown
hypomethylation of DNA repeat elements, such as LINE-1 [258]. Currently, there is controversy surrounding the initial aberrant methylation signature in PCa. It has been suggested that CGI promoter hypermethylation is an early event, as normal adjacent tissues already displayed this pattern [256]. However, it was previously reported that hypomethylation is the first alteration as exemplified in a PIN mouse model [259].

Histone modifications and chromatin remodeling have shown to be imperative in regulating gene expression in prostate cancer. Few studies have investigated global alterations of histone modifications in PCa. The first study of global histone marks used IHC analysis, and found that presence of H3K18ac and H3K4me2 were indicative of recurrence in low-grade prostate tumors [260]. Another study found that an increase in global H3K27me3 was detected in metastatic tumors compared to normal tissues [261]. Meanwhile, various studies have shown that a global loss of H4K16ac and H4K20me3 marks in prostate tumors can predict recurrence [260, 262]. Still, it remains to be shown whether histone modifications can be used for early detection of the disease. Further, the writers and erasers involved in regulating histone modifications are often disrupted. These enzymes include HMTs, HDMs, and HDACs. Enhancer of zeste homolog 2 (EZH2) is a HMT that is well studied in prostate cancer, and was shown to be overexpressed in the disease [263]. EZH2 is a subunit of the PRC2 complex, which catalyzes H3K27me3, a repressive mark [264]. EZH2 has shown direct interaction with DNMTs, and has been reported to co-localize at hypermethylated promoter regions that are inactivated [120]. Further, prostate cancer cells with increased EZH2 expression have shown high rates of proliferation and advanced tumor aggressiveness [265]. In addition, HDACs such as histone deacetylase 1 (HDAC1) are overexpressed in prostate cancer and have shown further dysregulation in CRPC [266]. As another example, lysine-specific demethylase 1 (LSD1) is involved in repressing transcription and its overexpression is associated with hormone-refractory PCa and increased risk of recurrence [267]. Although these aberrations are not describing epigenetic modifications themselves, these studies highlight that epigenetics are indeed important in the development of prostate cancer, in that the enzymes regulating these patterns are altered.

Over 50 miRNAs have been identified as disrupted in prostate cancer [268]. Evidence has shown that differential miRNA expression can distinguish cancer from normal tissues, alluding to their potential as diagnostic and prognostic biomarkers. To date, many miRNAs have not been validated for their contribution to the disease, however several have shown influence on epigenetic reprogramming, cell
cycle, invasion, and interestingly, to maintain androgen-independent growth [269]. Several different mechanisms and players have shown to be implicated in the miRNA and AR-signaling axis. In this relationship, miRNAs such as miR-221/222 have been shown to regulate androgen-independent growth through PSA up-regulation, whereas androgens have alternatively been reported to repress the miRNAs themselves [270]. Additionally, miR-146a is a proposed tumor suppressor gene, as its inactivation is associated with hormone-refractory PCa [271].

Although increasing evidence supports the use of epigenetic biomarkers for the prognosis of PCa, few are transitioning into clinical trials perhaps due to lack of individual biomarker strength and biological significance [272]. Evidently, further studies are required to characterize global and gene-specific epigenetic profiles in PCa to better understand their specific contribution in prostate cancer development.
1.6 Hypothesis and Objectives

1.6.1 Hypothesis

I hypothesize that genome-wide studies of DNA methylation and hydroxymethylation modifications in primary prostate cell lines will demonstrate that prostate cancer undergoes a loss of 5hmC and accumulation of 5mC at specific genomic regions. Moreover, this study may improve our understanding of the biology of prostate cancer and DNA methylation-based prostate cancer biomarkers.

1.6.2 Aims of Study

I will investigate epigenetic regulation and DNA demethylation in prostate cancer through two aims.

1) Assessment of global 5mC profiles of benign and cancerous prostate cell lines and associated pathways implicated in prostate cancer progression.

2) Establishment of global 5hmC profiles of benign and cancerous prostate cell lines and novel pathways implicated in normal prostate and prostate cancer progression.
Chapter 2
Investigating the Methylome of Prostate Cell lines

2.1 Introduction

As of Sept 2014, a total of seven genome-wide methylation studies using next-generation sequencing technology have been reported in prostate cancer [122, 273-278]. In contrast, over 30 studies have been completed using methylation microarray analyses. Of the NGS studies reported, five were performed using prostate cancer cells and three studies utilized tumor tissue, exemplifying that although these techniques are novel, they have been successfully applied to patient samples. Interestingly, one study used a novel technique termed nucleosome occupancy and methylome sequencing (NOMe-seq), which simultaneously investigated nucleosome occupancy and DNA methylation marks. They discovered that methylation marks were disrupted both at promoter regions and distal regulatory regions of prostate cancer cells compared to normal [278]. Additionally, most studies used techniques based on bisulfite modification, thus the distribution of the hydroxymethylome is included in these analyses [122, 273, 274, 276, 277]. Further, a majority of studies compared cancer to normal cells and found major differences in a CpG island context, generally in promoter regions, suggesting either a biological significance or a bias towards high CpG-density detection. Lastly, one study observed methylation changes between TMPRSS2-ERG-positive and negative tumors, demonstrating a link between genomic and epigenetic changes in prostate cancer [275].

2.2 Material and Methods

2.2.1 Cell Culture and DNA Extraction

Normal human prostate epithelial cell line, RWPE-1, was obtained from the American Type Culture Collection (ATCC). Human prostate cancer cell line, 22Rv1, was provided by Dr. E. Diamandis (Mount Sinai Hospital). RWPE-1 cells were cultured with Keratinocyte Serum Free Medium (K-SFM) (Invitrogen) supplemented with 0.05 mg/ml bovine pituitary extract (BPE) and 5 ng/ml human recombinant epidermal
growth factor (EGF). 22Rv1 cells were cultured with RPMI 1640 (Mount Sinai Hospital) with 10% fetal bovine serum (FBS). All cells were cultured as a monolayer and maintained in a humidified incubator at 37°C with 5% CO₂. Genomic DNA was extracted from cells after trypsinization, using the QIAamp DNA Mini Kit (Qiagen) following the protocol provided.

### 2.2.2 DNA Methylation Enrichment: MBD2 and MeDIP-capture

Genomic DNA extracted from RWPE-1 and 22Rv1 cells was sonicated into approximately 100-300 bp fragments using a Vibracell Disrupter (SONICS). Sheared genomic DNA (4 ug) was incubated with methyl-CpG binding domain 2 (MBD2) protein and magnetic streptavidin beads provided by the MethylMiner Kit (Invitrogen). Bound DNA fragments were eluted with the highest concentration of NaCl buffer (2000 mM). Bound, unbound, and input DNA fragments were precipitated using MinElute PCR Purification Kit (Qiagen) with the final elution in UltraPure Distilled Water (Invitrogen). Bound and input DNA was quantified using Qubit 2.0 Fluorometer (Invitrogen). Optimization was performed for the MeDIP Kit (Diagenode) by denaturing sheared genomic DNA (2 ug) from 22Rv1 cells and immunoprecipitated overnight at 4°C, according to the protocol provided. However, NGS was not performed on MeDIP Kit elutions.

### 2.2.3 Real Time Quantitative PCR (RT-qPCR) following DNA Methylation Enrichment: MBD2- and MeDIP-qPCR

DNA methylation levels were analyzed by RT-qPCR using the 7500 Real Time PCR Instrument (Applied Biosystems). Methylation levels were normalized to 5% of input genomic DNA. The PCR assays comprised of 10 ul of iTaq Universal SYBR Green Supermix (Bio-Rad), 50 ng of methylated or input DNA, and 0.25 uM of each primer with a total volume of 20 ul. All PCR assays included a non-template control, using UltraPure Distilled Water (Invitrogen) as the template. The PCR conditions were: 10 min at 95°C, 40 cycles of denaturation for 15 sec at 95°C, and annealing for 1 min at 60°C. The sequences of primers amplifying CDKN2A are 5’-TGG GTT TGT AGA AGC AGG CAT-3’ (forward), 5’-AGC CAG CTT GCG ATA ACC AAA G-3’ (reverse); TGFβ2 are 5’-CCA TCT ACA ACA GCA CCA GGG ACT-3’ (forward), 5’-GTA GTA CTC TTC GTC
GCT CCT CTC-3' (reverse); and HOXD8 are 5'-AAC TTG CGT TCG TCT GCC CT-3' (forward), 5'-ACA GAA ACG TTC TGA GGC GGG AAA-3' (reverse).

2.2.4 Next-Generation Sequencing of MBD2-enriched DNA: MBD-seq

Bound, MBD2-enriched DNA (20 ng) and input samples (non-enriched) from 22Rv1 and RWPE-1 cells were submitted in triplicate for library preparation (NEBNext® ChIP-Seq Library Prep Reagent Set for Illumina) and high-throughput sequencing using the HiSeq 2000 and 2500 (Illumina) at The Centre for Applied Genomics (The Hospital for Sick Children). Each library generated approximately 50 million paired-end reads with 5X coverage of CpGs.

2.2.5 Bioinformatic Analysis of MBD-seq

Sequenced reads were mapped to the reference human genome (GRCh37, hg19) using Bowtie (v0.12.7) [279]. Significantly enriched regions/peaks of methylation were determined using model-based analysis of ChIP-seq (MACS) algorithm [280], by comparing bound, enriched samples to input, non-enriched samples. Annotation was performed using ChIPpeakAnno [281] and Annovar [282] programs to determine specific genomic features: promoter regions (2.5 kb upstream and 500 bp downstream from the nearest TSS), genic (further defined as 500 bp downstream from the nearest TSS), exons, introns, and intergenic (defined as greater than 100 kb up/downstream from the nearest TSS). UCSC CpG island definitions were used to define CpG islands. MACS data was further analyzed using DiffBind algorithm [283] to determine consensus peaks common between biological replicates, and these were subsequently defined as differentially methylated regions (DMRs) between RWPE-1 and 22Rv1 cells.

2.2.6 Pathway-based Analysis of MBD-seq

Genomic region lists were generated to represent differentially methylated regions (DMRs) identified between RWPE-1 and 22Rv1 cells. Functional and pathway enrichment analysis were performed on genomic region lists using the Genomic Regions Enrichment of Annotations Tool (GREAT) [284]. GREAT
analyses were performed using the binomial test over genomic regions to assess for significance. This option is recommended by GREAT when inputting large region lists. Subsequently, GREAT results were organized into functional enrichment maps using the visualizing software, Cytoscape [285] or organized into bar graphs using the recommended statistic of \(-\log_{10}(\text{binomial raw p-value})\) for ranking. Ontologies shown were most significantly enriched and the following ontology categories were omitted from analysis, as the resulting outputs were not biologically relevant: Phenotype Data and Human Disease, Gene Expression, Regulatory Motifs, Gene Families (except for HGNC Gene Families). Additionally, enriched genomic region lists were subjected to post-analysis with The Molecular Signatures Database (MSigDB) curated gene sets [286] using the post-analysis plugin within Cytoscape. The gene set termed “Genes reported to be methylated in PCa”, was generated from the systematic review by Day and Bianco-Motto (2013).
2.3 Results

2.3.1 Optimization of Methylation Enrichment Techniques, MBD2- and MeDIP-capture

To determine the optimal DNA methylation (5mC) enrichment technique, I compared two different strategies: methyl-CpG binding domain (MBD)-capture and immunoprecipitation-based methods. I validated the 5mC enrichment efficiencies, using enriched genomic DNA from 22Rv1 cells and analyzed genes of known methylation status: cyclin-dependent kinase inhibitor 2A (CDKN2A), transforming growth factor, beta 2 (TGFβ2), and homeobox D8 (HOXD8) (Fig. 4). I found that the DNA methylation levels, determined on an Agilent Human CpG Island methylation microarray previously performed in the lab on 22Rv1 cells [287], was more accurately represented using the MBD-capture method compared to the immunoprecipitation method. For instance, the MBD-capture was more sensitive compared to MeDIP as genes that were observed as methylated on the array, such as CDKN2A and TGFβ2, were detected in the methylated fraction of the MethylMiner kit, and in contrast, there was no observation in the methylated fraction of the MeDIP kit. In terms of specificity, MBD-capture was able to discriminate between methylated and unmethylated DNA fragments, as CDKN2A and TGFβ2 were exclusively observed in the methylated fraction, however MeDIP kit detected these genes in the unmethylated fraction. Further, MBD-capture produced higher enrichment yields compared to MeDIP, as exemplified by an unmethylated gene from the array, HOXD8, which was detected in the unmethylated fraction of both the MethylMiner kit and MeDIP kit, although the enrichment was greater in the former. Thus, I continued performing 5mC enrichment using the MBD-capture strategy.
Figure 4. MBD2- and MeDIP-qPCR in 22Rv1 cells using regions of known methylation status from an Agilent Human CpG Island methylation array. (a-b) CDKN2A and TFGB2 were previously shown as methylated, while (c) HOXD8 was unmethylated in 22Rv1 cells. Both the MethylMiner kit and MeDIP kit produced a methylated and unmethylated fraction, and these fractions were subjected to RT-qPCR. $2^{-\Delta Ct}$ method normalized to 5% of input DNA, where error bars indicate standard deviation.
2.3.2 Relative Absolute and Differential Analysis of MBD-seq in Prostate Cell lines, Benign and Cancerous

When MBD-capture is combined with NGS, this technique is referred to as MBD-seq (Fig. 5). In order to determine genome-wide regions of relative absolute methylation using MBD-seq, peak calling was performed using MACS analysis [280] on triplicates of RWPE-1 and 22Rv1, where each sample was compared to an input sample (Fig. 6a-b). Input samples were non-enriched, sheared genomic DNA derived from RWPE-1 and 22Rv1 cells, respectively. This analysis found that 22Rv1 cells generally incurred less global 5mC marks compared to RWPE-1 cells (Rep 2, 3), however Replicate 1 demonstrated the opposite result.

To define methylated regions that were differential between RWPE-1 and 22Rv1 cell lines, DiffBind analysis of the MACS datasets was performed. Consensus peaks, or regions common between MACS peaks triplicates, were generated and analyzed for differential enrichment between RWPE-1 and 22Rv1 cells. These regions were entitled differentially methylated regions (DMRs) and represent hypermethylated regions in one cell type versus the other. Overall, a total of 55 631 significant DMRs (p-value <0.04, FDR <0.15, fold-change >2) were identified in RWPE-1 cells and 52 852 DMRs in 22Rv1 (Fig. 6c-d), which corresponds to 10 956 and 13 369 unique RefSeq genes, respectively. The total number of DMRs overlapping CpG islands (CGIs) was greater in 22Rv1 cells (6 249 DMRs, 12%) compared to RWPE-1 (1 363 DMRs, 2.4%) (Fig. 6d). Stratifying CGIs by genomic features, I found that 22Rv1 cells displayed a greater percentage of DMRs in all regions, compared to RWPE-1 (Fig. 6f), where the most striking increase was observed within promoter regions (93.6% in 22Rv1, 14.9% in RWPE-1). When all DMRs were stratified by genomic location: promoter, genic, or intergenic; 22Rv1 cells revealed increased regions of DMRs in all features, with the exception of intergenic regions, when compared to RWPE-1 (Fig. 6e). Further, intronic regions of hypermethylation showed a greater distinction between the two cell lines, compared to exonic regions.
Figure 5. MBD-seq methodology pipeline. (Step 1-5b) Cell line DNA was enriched for 5mC marks using the MBD-capture and MeDIP-capture, however only the (Step 6) former was submitted for next-generation sequencing (NGS) using the Illumina HiSeq 2000/2500. (Step 7-8) Bioinformatics analysis was performed using MACS and Diffbind software, while (Step 9) pathway-based analysis was performed using GREAT program. Diagram by Ho LT and Lo C.
Figure 6. Relative absolute methylated regions and significant differentially methylated regions (DMRs) identified in RWPE-1 and 22Rv1 cells using MBD-seq. (a-b) Total number of absolute regions/peaks of methylation, called by MACS in three biological replicates. (c) Volcano plot representing threshold of significance determined using p-value (<0.04) and FDR (<0.15) cutoff, where each point represents a DMR. (d) Total significantly hypermethylated regions identified and stratified by CpG island overlap. (e) DMRs were categorized into genomic features and (f) further stratified by CpG island overlap.
2.3.3 Validation of MBD-seq Results using Methylation Microarrays

To confirm that the MBD-seq technique successfully detected CpG methylation, I used two independent methylation microarrays previously performed in our lab. Firstly, I compared significant DMRs in 22Rv1 between MBD-seq and an Agilent Human CpG Island methylation microarray [287]. In this microarray, untreated 22Rv1 cells were compared to those treated with 5-aza-2’-deoxycytidine (DAC), a demethylating agent. I assessed regions/probes within the hypomethylated fraction, which demonstrated reduced methylation signal (p<0.05) post-DAC treatment, as these regions were suggested as methylated in the untreated cells. I took a candidate region approach, and analyzed the top 65 ranking probes that were significantly hypomethylated. Following comparison with all significant DMRs discovered through MBD-seq, I identified common regions of methylation (Table 1).

Additionally, I analyzed data from an Illumina Infinium HumanMethylation450 BeadChips methylation microarray that sampled both untreated RWPE-1 and 22Rv1 cells (unpublished data). I used raw beta values to determine regions/probes with high methylation signal, that is, values ranging from 0.9 to 1 [288, 289]. Regions assessed were not normalized or tested for statistical significance between RWPE-1 and 22Rv1 cells, thus were potentially methylated in both cell lines. A candidate region approach was taken, where probes were ranked based on beta value and 12 probes were chosen for comparison with MBD-seq, based on the beta value observed for the other cell line. That is, a high scoring probe for untreated RWPE-1 was chosen (beta value >0.9), if it was detected as a low scoring probe (beta value <0.4) in untreated 22Rv1 cells. Through this analysis, I was able to identify several significant DMRs common to the microarray dataset (Table 2).
Table 1. Representative list of significant DMRs in 22Rv1 cells, discovered using MBD-seq, which were validated on an Agilent Human CpG Island methylation microarray.

<table>
<thead>
<tr>
<th>Chr</th>
<th>Start</th>
<th>End</th>
<th>Gene name</th>
<th>Gene feature</th>
<th>CpG island overlap</th>
<th>P-value</th>
<th>FDR</th>
<th>Abs Fold-change</th>
<th>Agilent Probe ID</th>
</tr>
</thead>
<tbody>
<tr>
<td>chr3</td>
<td>147138566</td>
<td>147137756</td>
<td>ZIC1</td>
<td>Intergenic</td>
<td>ID_5022</td>
<td>7.55E-10</td>
<td>9.24E-07</td>
<td>10.02263732</td>
<td>A_17_P02656394</td>
</tr>
<tr>
<td>chr10</td>
<td>21823888</td>
<td>21824615</td>
<td>MLLT10</td>
<td>Promoter</td>
<td>ID_13913</td>
<td>0.000205642</td>
<td>0.004029016</td>
<td>4.85706821</td>
<td>A_17_P07119347</td>
</tr>
<tr>
<td>chr5</td>
<td>172655222</td>
<td>172662046</td>
<td>NKX2-5</td>
<td>Gene body</td>
<td>ID_7372</td>
<td>0.01718216</td>
<td>0.086122538</td>
<td>2.677702748</td>
<td>A_17_P04471754</td>
</tr>
<tr>
<td>chr20</td>
<td>25370276</td>
<td>25371050</td>
<td>ABHD12</td>
<td>Promoter</td>
<td>ID_26044</td>
<td>1.00E-07</td>
<td>2.40E-05</td>
<td>8.148078208</td>
<td>A_17_P11128252</td>
</tr>
</tbody>
</table>

DMRs: differentially methylated regions, chr: chromosome, FDR: false discovery rate, Abs: absolute.

Table 2. Representative lists of significant DMRs in RWPE-1 (blue) and 22Rv1 (red) cells, discovered using MBD-seq, which were validated on an Illumina Infinium HumanMethylation450 BeadChips methylation microarray.

<table>
<thead>
<tr>
<th>Chr</th>
<th>Start</th>
<th>End</th>
<th>Gene name</th>
<th>Gene feature</th>
<th>CpG island overlap</th>
<th>P-value</th>
<th>FDR</th>
<th>Abs Fold-change</th>
<th>Illumina Probe ID</th>
</tr>
</thead>
<tbody>
<tr>
<td>chr11</td>
<td>120040298</td>
<td>120042080</td>
<td>TRIM29</td>
<td>Intergenic</td>
<td>No</td>
<td>0.042415693</td>
<td>0.16375125</td>
<td>2.750738972</td>
<td>cg1605559</td>
</tr>
<tr>
<td>chr12</td>
<td>131158518</td>
<td>131160135</td>
<td>RIMBP2</td>
<td>Intergenic</td>
<td>No</td>
<td>0.020855049</td>
<td>0.098786739</td>
<td>3.130368298</td>
<td>cg20532558</td>
</tr>
<tr>
<td>chr3</td>
<td>197104424</td>
<td>197106454</td>
<td>DLG1</td>
<td>Intergenic</td>
<td>No</td>
<td>0.0156628</td>
<td>0.085021962</td>
<td>2.665314699</td>
<td>cg19873496</td>
</tr>
<tr>
<td>chr8</td>
<td>6637445</td>
<td>6638738</td>
<td>XKR5</td>
<td>Intergenic</td>
<td>No</td>
<td>0.009213247</td>
<td>0.056322313</td>
<td>3.381177788</td>
<td>cg24422826</td>
</tr>
<tr>
<td>chr11</td>
<td>122888559</td>
<td>122890557</td>
<td>BSX</td>
<td>Intergenic</td>
<td>No</td>
<td>0.000866519</td>
<td>0.010893845</td>
<td>4.229924017</td>
<td>cg10744926</td>
</tr>
<tr>
<td>chr1</td>
<td>181381906</td>
<td>181383003</td>
<td>CACNA1E</td>
<td>Intergenic</td>
<td>ID_1975</td>
<td>0.000242913</td>
<td>0.004513826</td>
<td>5.15793315</td>
<td>cg20741134</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Chr</th>
<th>Start</th>
<th>End</th>
<th>Gene name</th>
<th>Gene feature</th>
<th>CpG island overlap</th>
<th>P-value</th>
<th>FDR</th>
<th>Abs Fold-change</th>
<th>Illumina Probe ID</th>
</tr>
</thead>
<tbody>
<tr>
<td>chr17</td>
<td>46641060</td>
<td>46642303</td>
<td>HOXB3</td>
<td>Gene body</td>
<td>ID_22076</td>
<td>0.011177408</td>
<td>0.064170297</td>
<td>3.077192445</td>
<td>cg13479204</td>
</tr>
<tr>
<td>chr1</td>
<td>67772797</td>
<td>67774278</td>
<td>IL12RB2</td>
<td>Promoter</td>
<td>ID_1219</td>
<td>0.003534966</td>
<td>0.029204968</td>
<td>3.368932901</td>
<td>cg20630547</td>
</tr>
<tr>
<td>chr11</td>
<td>503112</td>
<td>505251</td>
<td>RNH1</td>
<td>Gene body</td>
<td>ID_14919</td>
<td>0.001489308</td>
<td>0.015956196</td>
<td>3.676519721</td>
<td>cg23687468</td>
</tr>
<tr>
<td>chr7</td>
<td>50341702</td>
<td>50344760</td>
<td>IKZF1</td>
<td>Promoter</td>
<td>ID_9465</td>
<td>0.000632527</td>
<td>0.008748138</td>
<td>4.160073383</td>
<td>cg16499856</td>
</tr>
<tr>
<td>chr12</td>
<td>47472367</td>
<td>47472934</td>
<td>AMIGO2</td>
<td>Gene body</td>
<td>No</td>
<td>8.28E-05</td>
<td>0.002218315</td>
<td>5.493274524</td>
<td>cg18593793</td>
</tr>
<tr>
<td>chr3</td>
<td>44002345</td>
<td>44004036</td>
<td>TMEM42</td>
<td>Promoter</td>
<td>ID_4416</td>
<td>5.94E-05</td>
<td>0.001783035</td>
<td>5.010557733</td>
<td>cg25190974</td>
</tr>
</tbody>
</table>

DMRs: differentially methylated regions, chr: chromosome, FDR: false discovery rate, Abs: absolute.
2.3.4 Pathways Associated and Potentially Regulated by Differentially Methylated Regions (DMRs) in Prostate Cell lines

To investigate pathways potentially regulated by DMRs in RWPE-1 and 22Rv1 cells, I used pathway-based analysis on regions categorized by genomic location as follows:

**Promoter CpG island (CGI) DMRs:**
It has been shown that CpG island hypermethylation in promoter regions is a common event in cancer [247-249], and these regions are well documented for their contribution to transcriptional regulation [107]. Thus, I investigated the biological role of genes mapping to this feature in 22Rv1 cells. Pathway analyses showed that the top pathways to be potentially regulated by DMRs were RNA polymerase I transcription, glycolysis, telomere maintenance, and gene silencing (binomial p<2.57E-03) (Fig. 7a). Further, I observed that a subset of these genes, such as tropomyosin 4 (TPM4), homeobox A9 (HOXA9), and chemokine (C-X-C motif) ligand 1 (CXCL1), were previously shown to be methylated in prostate cancer and were known KEGG cancer genes (Fig. 7b, Table 3). From this analysis, I identified functionally relevant promoters, which were significantly and differentially methylated in 22Rv1 compared to RWPE-1 cells.

**Genic non-CGI DMRs:**
When examining pathways involved in hypermethylated genic regions, I found that significant non-CGI DMRs in RWPE-1 were enriched for DNA repair, RNA polymerase II transcription, and hormone biosynthesis, among others (binomial p<4.99E-12) (Fig. 8a). I performed an identical analysis for significant non-CGI DMRs in 22Rv1 cells and found that a variety of processes previously implicated in prostate cancer were enriched in these regions, including GTPase mediated signaling, IGF1 pathway, EGF receptor signaling pathway, and androgen signaling (binomial p<4.88E-09) (Fig. 8b). Finally, a common pathway significantly enriched for non-CGI genic hypermethylation in both RWPE-1 and 22Rv1 cells, was observed to be cellular organization.

**Intergenic non-CGI DMRs:**
The role of methylation in intergenic regions is not entirely understood, however these are common sites for hypomethylation in cancer and evidence shows that this event is likely associated with genomic instability [290, 291]. I discovered that significant non-CGI intergenic DMRs in RWPE-1 had pathway over-
representation for epithelium development, heparan sulfate biosynthesis, RNA polymerase II transcription, and PRC2 complex and subunit targets (binomial p<3.86E-49) (Fig. 8c). In a parallel analysis, non-CGI DMRs in intergenic regions of 22Rv1 cells were found to be enriched for translational regulation, regulation of BMP signaling, and semaphorin activity (binomial p<1.46E-06) (Fig. 8d). Importantly, although pathways enrichment for this analysis of 22Rv1 was statistically significant, the level of significance was reduced compared to those pathways enriched in RWPE-1.
Figure 7. Pathways associated with differentially methylated regions (DMRs) in 22Rv1 in promoter regions overlapping CpG islands (CGIs), identified by MBD-seq. (a) Over-represented pathways for DMRs overlapping promoter regions determined by GREAT analysis and (b) organized into a functional enrichment map using Cytoscape program. GREAT gene sets were not annotated in the enrichment map. Post-analysis was performed on the enrichment map using gene sets of interest (green and purple triangles) to discover candidate 5mC biomarkers (pink lines).
Table 3. Representative list of candidate DNA methylation biomarkers, derived from significant DMRs detected by MBD-seq of 22Rv1 cells, and identified through pathway-based analysis using GREAT tool and the gene set KEGG: Pathways in cancer.

<table>
<thead>
<tr>
<th>Chr</th>
<th>Start</th>
<th>End</th>
<th>Gene name</th>
<th>P-value</th>
<th>FDR</th>
<th>Abs Fold-change</th>
</tr>
</thead>
<tbody>
<tr>
<td>chr14</td>
<td>54421058</td>
<td>54424162</td>
<td>BMP4</td>
<td>1.38127E-11</td>
<td>1.69795E-07</td>
<td>11.7435394</td>
</tr>
<tr>
<td>chr13</td>
<td>37005379</td>
<td>37007418</td>
<td>CCNA1**</td>
<td>1.95065E-10</td>
<td>4.65966E-07</td>
<td>10.7376342</td>
</tr>
<tr>
<td>chr19</td>
<td>16178102</td>
<td>16179679</td>
<td>TPM4*</td>
<td>2.50925E-10</td>
<td>5.19502E-07</td>
<td>10.1785378</td>
</tr>
<tr>
<td>chr15</td>
<td>99192832</td>
<td>99194842</td>
<td>IGF1R</td>
<td>2.73603E-10</td>
<td>5.37715E-07</td>
<td>10.4247708</td>
</tr>
<tr>
<td>chr7</td>
<td>27205602</td>
<td>27207163</td>
<td>HOXA9*</td>
<td>5.02307E-10</td>
<td>7.28797E-07</td>
<td>10.002373</td>
</tr>
<tr>
<td>chr4</td>
<td>74734545</td>
<td>74735788</td>
<td>CXCL1*</td>
<td>1.92565E-09</td>
<td>1.595E-06</td>
<td>9.7660195</td>
</tr>
<tr>
<td>chr7</td>
<td>116963723</td>
<td>116964941</td>
<td>WNT2</td>
<td>2.83537E-09</td>
<td>2.08136E-06</td>
<td>9.39767681</td>
</tr>
<tr>
<td>chr7</td>
<td>128827889</td>
<td>128828973</td>
<td>SMO</td>
<td>3.43699E-09</td>
<td>2.32923E-06</td>
<td>9.52152665</td>
</tr>
</tbody>
</table>

DMRs: differentially methylated regions, chr: chromosome, FDR: false discovery rate, Abs: absolute.
*Genes reported to be methylated in prostate cancer [292], **Methylation status identified in a separate study [293].
Figure 8. Pathways associated with differentially methylated regions (DMRs) in RWPE-1 and 22Rv1 cells, overlapping non-CpG islands (non-CGIs), identified by MBD-seq. Over-represented pathways for DMRs overlapping (a-b) genic/gene body regions and (c-d) intergenic regions, determined by GREAT analysis.
2.4 Discussion

In this chapter, I defined the genome-wide methylation patterns in prostate cell lines, RWPE-1 (benign) and 22Rv1 (cancerous), using an innovative technique combining 5mC-enrichment with next-generation sequencing (NGS) technology, termed MBD-seq. Further, I discovered pathways implicated in DNA methylation-mediated regulation. Differential analyses revealed that the genomic distribution of hypermethylated regions were similar between cell lines, however when stratified by occurrence within a CpG island (CGI), promoter regions showed the highest contribution of CGI methylation (93.6%) in cancer cells. Genic/gene body and intergenic region analysis demonstrated that in both normal and cancer cells, methylation generally occurs in a non-CGI manner. Through pathway-based analyses, I observed that CGI-hypermethylated promoters in 22Rv1 share common biological functions and these may lead to the discovery of relevant DNA methylation biomarkers distinguishing between normal and prostate cancer. Additionally, genic and intergenic methylation were found to harbor over-represented pathways that were altered between normal and prostate cancer cells.

I have chosen to utilize these prostate cell lines to model the disease, as they are representative of normal and cancer conditions. That is, RWPE-1 is an immortalized epithelial cell line, which contains basal keratins, is non-tumorigenic in mice, non-invasive, expresses p53 and Rb, and is androgen responsive [294-296]. In terms of a prostate cancer model, I used 22Rv1 cell line. This is a xenografted epithelial cell line, derived from the primary tumor of a patient diagnosed with Gleason score (GS) 9 prostate cancer. The cell line is characterized as cancerous prostate epithelium, as it contains luminal keratins, is tumorigenic in mice, motile and invasive, and is androgen receptor (AR) positive and androgen responsive [297-299]. Moreover, both cell lines are well documented in the literature and are commonly used as prostate cancer models.

A 5mC-capture based method was used to enrich for methylated fragments throughout the genome. These fragments were bound at methylated CpGs by the methyl-CpG binding domain of MBD2 protein. This method of capture was selected instead of the 5mC-immunoprecipitation method, termed MeDIP, as pilot studies found MBD2-capture performed more optimally. That is, CDKN2A, TGFβ2, and HOXD8 were genes of known methylation status, and these profiles were effectively detected using the MBD2-capture
method (Fig. 4). Importantly, I found that MeDIP method was unable to correctly identify methylated genes, and this could be due to antibody preference for regions of lower CpG density [186, 187]. That is, CpG island methylation could be difficult to detect using this method, as CGIs are CpG-dense regions. These findings corroborate with other studies that have shown several advantages of MBD-capture compared to MeDIP, such as ease of using double-stranded DNA, higher specificity in CpG recognition, maintenance of DNA integrity, and reduced bias for genic locations [183, 188, 190, 300, 301]. MBD-enrichment techniques have been successfully applied to prostate cancer cells [93], providing evidence supporting this detection method, especially in the field of -omics analyses.

MBD-seq data was first analyzed using MACS analysis [280], which compared each enriched sample to a corresponding input, or non-enriched sample. I have chosen to use MACS algorithm for MBD-seq absolute analysis as it has been applied to recent high-throughput methylation studies, is reputable in background normalization, and it outperformed alternative algorithms designed for methylation sequencing [187, 302, 303]. From the MACS analysis, I observed that 22Rv1 cells generally displayed a lower number of methylated regions compared to RWPE-1 (Fig. 6a-b), albeit at a conservative level. This is an expected result, as a hallmark of various cancers is a global loss of methylation compared to normal state [256, 257] and another prostate cancer study confirmed that the aberrant hypomethylation pattern was moderate [155].

Next, I sought to determine differential methylation profiles in normal compared to cancer cells. This is a pertinent analysis, as mounting evidence supports the potential of DNA methylation biomarkers in improving the diagnostic and prognostic evaluation of prostate cancer [248, 249, 304]. In that sense, the methylation differences identified between RWPE-1 and 22Rv1 cells would be a useful resource to validate known biomarkers and determine novel candidates. Additionally, the roles of 5mC marks in normal and cancer methylation have been well characterized in promoter CGIs, while the function of genic and intergenic methylation is still under investigation. Thus, I aimed to address this gap in knowledge.

The differential analysis was performed using DiffBind algorithm, which is commonly used to process ChIP-seq peak calling programs, such as MACS [283, 305], and identified differentially methylated regions (DMRs) between RWPE-1 and 22Rv1 sequencing reads. In this pipeline, DMRs represent regions showing greater methylation signal, or hypermethylation, in one cell line (fold change >1) compared to the other
cell line dataset. Alternatively, DMRs also represent a region of hypomethylation in the compared cell line. That is, a DMR identified in RWPE-1 cells is not only considered to be hypermethylated in these cells, but is consequently hypomethylated in 22Rv1. Further, these DMRs include regions that are exclusively methylated in one cell line and not the other. For instance, a DMR in 22Rv1 cells may represent a region that was not methylated in RWPE-1. Thus, although the range in methylation patterns is thoroughly evaluated using this differential analysis, DMRs of interest must be interrogated at the MACS analysis or absolute level to ensure a properly defined methylation status.

In my resulting differential methylation analysis of prostate cancer and normal prostate epithelium cell lines, I found that cancer cells generally incurred a lower number of global 5mC marks compared to normal, which is consistent with previous genome-wide methylation patterns in the disease [155]. Investigation of the distribution of DNA methylation across CGIs revealed that hypermethylation of these regions concentrated around promoter regions (Fig. 6f), and these may lead to down-regulation of gene expression. This was an expected result based on the gene-specific gain in methylation signature that is commonly described in prostate cancer [247-249]. Alternatively, non-CGI DMRs most commonly occurred within intronic regions (Fig. 6f). These results could be due to the unproportional representation of regions within gene bodies, where introns make up roughly 24% of the genome compared to exon regions at 1.1% [306]. Evidence has shown the role of non-CGI methylation in genic regions to be implicated in increased gene expression [307, 308], alternate promoters [41], and splicing [309]. On the other hand, CGI gene body methylation in exonic regions has been hypothesized to be involved in regulating the transcription of non-coding RNA [310]. Lastly, intergenic DMRs occurring outside of CGIs were found in moderate numbers. Intergenic methylation has been associated with genomic stability [290, 291] and repression of enhancer activity [311], however the role within transcription factor binding sites is controversial, not always resulting in reduced activity [312, 313]. These results emphasize that CGI methylation is indeed abundant in cancer cells, particularly in promoter and exonic regions, whereas a majority of 5mC marks are occurring outside of CGIs, among intronic and intergenic regions (Fig. 6e-f).

To determine the functional role of methylation patterns identified in specific genomic locations, I performed pathways-based analysis using the GREAT tool [284]. GREAT was designed to analyze ChIP-seq datasets, map cis-regulatory regions to their putative target genes, and subsequently annotate the biological processes associated with these genes. Advantages associated with the binomial region-based
analysis utilized by GREAT, are the consideration of multiple sample regions targeting a single gene, a group of proximal or distal regions targeting a common set of genes, and ultimately predicting functions regulated by sample regions rather than limiting to gene expression correlations or protein-protein interactions [286, 314]. Additionally, GREAT has been successfully implemented when analyzing localized epigenomic modifications, such as DNA methylation and hydroxymethylation studies [145]. Using alternate gene-based pathways tools, such as Gene Set Enrichment Analysis (GSEA) [286], I was unable to include numerous DMRs per gene as this tool requires the dataset to be systematically filtered, allowing only one DMR to be assigned per gene. Further, a single gene cannot be used in both the normal and cancer phenotype, which leads to a bias in the resulting pathways enriched. In the MBD-seq datasets, GSEA analysis revealed greater enrichment in 22Rv1 DMRs (10 901 genes, 1 370 gene sets enriched at pval<0.05, FDR<0.25) compared to RWPE-1 (7 763 genes, 97 gene sets enriched at pval<0.05, FDR<0.25), demonstrating that more genes were assigned to the cancer phenotype (Supplemental Fig. 1). When using gene-based algorithms, these biases will exist as GSEA, along with many commonly used pathways tools; have been developed to analyze gene expression datasets [286, 314].

GREAT analysis of biologically significant DMRs occurring within CGI promoters of prostate cancer cells were enriched for RNA polymerase I transcription and gene silencing (Fig. 7a). As most DNA methylation studies have focused on cancer-associated hypermethylation in promoter CGIs, it has been shown that the functional significance of these aberrant developments indeed represses gene expression [2, 315], and these findings largely support our pathways analyses. Interestingly, roles in RNA polymerase I transcription, glycolysis, and telomere maintenance were discovered (Fig. 7a). Although RNA polymerase I has not shown direct regulation by methylation, ribosomal biogenesis has been implicated in prostate cancer, suggesting that silencing of enzymes involved could contribute to the disease. For example, *SnorRNA 50*, which mediates methylation of 28S rRNA, has been implicated as a potential tumor suppressor in PCa [316]. Additionally, glycolysis is generally up-regulated in prostate cancer [317], and one study shows that gastric cancer incurs promoter methylation of *fructose-1,6-bisphosphatase-2* (*FBP2*), an enzyme involved in regulating glucose metabolism. Aberrant methylation led to silencing of *FBP2* and consequently, when re-expressed, tumor formation was abrogated [318]. Lastly, telomere shortening has been associated with prostate carcinogenesis and the methylation and inactivation of players regulating this phenomenon could allow disease progression [319]. Ultimately, these mechanisms require further investigation to demonstrate the specific regulatory role of DNA methylation. Following, I used MSigDB
gene sets of interest and found that genes previously reported as methylated in prostate cancer, such as tropomyosin 4 (TPM4), homeobox A9 (HOXA9), and chemokine (C-X-C motif) ligand 1 (CXCL1) (Fig 7b, Table 3) could be discovered with MBD-seq methodology. This innovative analysis was valuable in filtering a large dataset for biologically relevant candidates, rather than top-scoring DMRs, which are frequently the basis for epigenetic biomarkers discovered in the cancer field.

Additionally, GREAT analysis revealed that significant DMRs within non-CGI genic regions were enriched for RNA polymerase II transcription, and hormone biosynthesis in RWPE-1 cells (Fig. 8a). As mentioned previously, studies have shown that gene body methylation plays a role in promoting gene expression, and regulating alternate promoters and splicing events [41, 307-309]. My analyses are supportive of these studies, and introduce novel implicated pathways such as hormone biosynthesis (Fig. 8a). It is likely that this is a relevant biological process, as hormone regulation is up-regulated in PCa [320] and these genic DMRs are, in fact, hypomethylated in 22Rv1 cells. This differential 5mC pattern may contribute to the proper maintenance of hormone biosynthesis in that regulators involved demonstrate a loss of methylation, leading to increased gene expression. On the other hand, the genic DMRs identified in 22Rv1 analysis, were enriched for various pathways previously associated with prostate cancer such as IGF1 pathway [320, 321], EGF receptor signaling pathway [322], and androgen signaling [323]. This suggests that genic hypermethylation in cancer may contribute to the dysregulation of these pathways, and more specifically, may promote up-regulation of these classic pathways by promoting transcriptional elongation of oncogenes involved.

Lastly, non-CGI methylation among intergenic regions were of interest, as these vast regulatory gene features have largely remained unstudied, in the context of their regulation through DNA methylation. It is generally understood that intergenic regions undergo hypomethylation during the development of cancer, leading to chromosomal instability [105, 106]. I found that non-CGI DMRs within intergenic regions of RWPE-1 were enriched for epithelium development, heparan sulfate biosynthesis, RNA polymerase II transcription, and PRC2 complex and subunit targets (Fig. 8c). Although GREAT analysis did not find pathways specifically annotated for chromosomal instability, polycomb repressive complex 2 (PRC2) is known to inform a closed chromatin state, which contributes to proper chromatin organization [324]. Interestingly, it has been shown that PRC2 can target DNA repeat elements in mouse ESCs. Thus, the loss of this pattern could contribute to chromosomal instability [325]. In terms of heparin sulfate
function, reports have shown its association with proliferation and invasion in prostate cells [326]. In addition, regulation of RNA polymerase II opens another avenue to regulate gene expression through methylation, which could perhaps be working in tandem with promoter CGI methylation or against it. That is, methylated intergenic regions in normal condition may maintain and promote RNA polymerase II transcription, while hypomethylated intergenic regions in cancer may contribute to the up or down-regulation of gene expression.
Chapter 3
Investigating the Hydroxymethylome of Prostate Cell lines

3.1 Introduction

Limited studies have investigated DNA hydroxymethylation patterns in prostate tissues, both normal and cancer. Utilizing immunohistochemical analysis, three studies have shown a global reduction in 5hmC marks in tumors compared to normal tissue [129, 155, 156]. Interestingly, the first study observed that normal prostate tissue demonstrated stronger staining for 5hmC marks in luminal compared to basal cells. That is, luminal cells are the more terminally differentiated cell type and this suggests that 5hmC patterns are differentiation-dependent. Further, in prostate tumors, the study found that 5hmC marks were not associated with clinicopathological features, where even low-grade tumors were significantly depleted in global 5hmC levels. These reports provide evidence that the loss of 5hmC is potentially an early event in PCa carcinogenesis [155]. In further support, another study found that 5hmC staining was mutually exclusive to Ki67, which is a marker of proliferation. Indeed, it was demonstrated that tumor tissue had stronger Ki67 staining compared to normal, and the inverse pattern was observed for 5hmC marks [156]. Additionally, normal prostate epithelium has shown high correlation between abundant global 5hmC and H3K27me3 levels. Both marks were markedly reduced in adjacent tumor cells, however, the direct link of these marks was not investigated [327]. One study showed that TET1 could inhibit cancer growth and metastasis in prostate cancer cells, LNCaP and 22Rv1, through tissue inhibitors of metalloproteases (TIMPs). This study stained for TET1 protein and found that 30% of tumors exhibited reduced levels [328]. Recently, it was shown that TET2 commonly harbors single-nucleotide polymorphisms (SNPs) in a subset of metastatic PCa tumors, however the contributing mechanism requires further analysis [329]. Nonetheless, the distribution of 5hmC marks in primary PCa and their role in the development of the disease is unknown.
3.2 Materials and Methods

3.2.1 Cell Culture and DNA Extraction

Previously described in Section 2.2.1.

3.2.2 DNA Hydroxymethylation Enrichment: hMeSeal

Genomic DNA extracted from RWPE-1 and 22Rv1 cells was sonicated into approximately 100-300 bp fragments using a Q125 sonicator (Qsonica). Sheared genomic DNA (10 ug) was specifically glucosylated at 5hmC by T4-phage β-glucosyltransferase, and incubated with biotin and streptavidin beads provided by the Hydroxymethyl Collector Kit (Active Motif). This process of enrichment is termed hydroxymethyl selective chemical labeling, or hMeSeal. The final elution of bound, unbound, and input DNA fragments were precipitated according to the protocol provided with the final elution in UltraPure Distilled Water (Invitrogen). Bound and input DNA was quantified using Qubit 2.0 Fluorometer (Invitrogen).

3.2.3 Next-Generation Sequencing of Glucosylated 5hmC-enriched DNA: hMeSeal-seq

Each replicate of 5hmC-enriched sample was prepared by pooling up to four hMeSeal reactions. Following, bound/glucosylated 5hmC-enriched DNA (15 ng), and input samples (non-enriched) from RWPE-1 and 22Rv1 cells were submitted in triplicate for library preparation (NEBNext® ChIP-Seq Library Prep Reagent Set for Illumina) and high-throughput sequencing using the HiSeq 2500 (Illumina) at The Centre for Applied Genomics (The Hospital for Sick Children). Each library generated approximately 50 million paired-end reads, with 10X coverage of CpGs.

3.2.4 Bioinformatic Analysis of hMeSeal-seq

Sequenced reads were mapped to the reference human genome (GRCh37, hg19) using Bowtie (v0.12.7) [279]. Repitools package [330] was used as an enrichment diagnostic screen of sequenced samples, enriched and input. Significantly enriched regions/peaks of hydroxymethylation were determined using
model-based analysis of ChIP-seq (MACS) algorithm [280], by comparing bound, enriched samples to input, non-enriched samples. Annotation of relative absolute hydroxymethylated regions was performed using Annovar program [282], where promoter regions were not defined and thus, are included in gene body regions. Remaining genomic features: exon, intron, and intergenic; were defined by the program. Annotation of differentially hydroxymethylated regions (DHMRs) was performed using ChIPpeakAnno [281] and Annovar [282] programs to determine specific genomic features: genic, exons, introns, and intergenic (previously described in Section 2.2.5). UCSC CpG island definitions were used to define CpG islands. MACS data was further analyzed using DiffBind algorithm [283] to determine consensus peaks based on one biological replicate per cell line, and these were subsequently defined as differentially hydroxymethylated regions (DHMRs) between RWPE-1 and 22Rv1 cells. RWPE-1 DNase-seq data from The Encyclopedia of DNA Elements (ENCODE) project (GEO accession: GSM1008595) was correlated to hMeSeal-seq datasets at the absolute and differential level, in both RWPE-1 and 22Rv1 cells.

### 3.2.5 Global DNA Hydroxymethylation Detection using Dot Blot Assay

Human 5mC and 5hmC genomic DNA standards (Zymo) and mouse tissue-derived genomic DNA of known 5mC and 5hmC percentage (Zymo) were used to optimize the dot blotting protocol. DNA was denatured in 0.4 M NaOH/10 mM EDTA for 10 min at 95 degrees Celsius and spotted onto a positively charged nylon membrane (Roche) using a Dot Blot apparatus (Durocher lab, Mount Sinai Hospital), and cross-linked by alkaline fixation. Membrane was washed with 2X saline-sodium citrate (SSC) buffer, air dried for 1 hour, and blocked with 5% non-fat milk in tris-buffered saline and tween 20 (TBS-T) for 1 hour, at room temperature. Following, the membrane was incubated with anti-5hmC antibody (Active Motif), overnight at 4 degrees Celsius. After washing, membrane was incubated with horseradish peroxidase (HRP)-conjugated anti-rabbit IgG secondary antibody (Life Technologies) for 1 hour. Membrane was treated with enhanced chemiluminescence (ECL) for visualization. Relative quantification was performed using 0.04% methylene blue in 0.3 M sodium acetate, pH 5.2 and ImageJ software.
3.2.6 Locus-specific DNA Hydroxymethylation Detection using hMeSeal-qPCR

Following hMeSeal technique (described in Section 3.2.2), DNA hydroxymethylation levels were analyzed by RT-qPCR using the 7500 Real Time PCR Instrument (Applied Biosystems). Hydroxymethylation levels were calculated using $2^{-\Delta\Delta Ct}$ method, relative to 1% of input genomic DNA, and normalized to 0.04% of input genomic DNA. A negative control reaction was performed according to the Hydroxymethyl Collector protocol (Active Motif), where UDP-azide-glucose was excluded from the glucosylation reaction. In addition, 2.5 ug of sheared genomic DNA was subjected to each glucosylation reaction, for both sample and negative control reactions.

The PCR assays comprised of 5 ul of PerfeCTa® SYBR® Green FastMix®, Low ROX (Quanta Biosciences), 3 ul of bound DNA elution (hydroxymethylated or negative control), and 5 uM of each primer, with a total volume of 10 ul. The 1% and 0.04% of input PCR assays were completed using the identical conditions, and input DNA stocks were prepared at 8.33 ng/ul and 3.33 ng/ul, respectively. All PCR assays included a non-template control, using UltraPure Distilled Water (Invitrogen) as the template. The PCR conditions were: 30 sec at 95°C, 40 cycles of denaturation for 5 sec at 95°C, and annealing for 30 sec at 60°C. Primers were designed using the specifications recommended by Hydroxymethyl Collector protocol (Active Motif) and targeted differentially hydroxymethylated regions (DHMRs) identified by hMeSeal-seq. The sequences of primers amplifying CDKN2B are 5'-ACA TCG GCG ATC TAG GTT CC -3' (forward), 5'-GAT CCC AAC GGA GTC AAC C -3' (reverse); and AR are 5'-CCC CTG ACT CAG CAA CAT TC -3' (forward), 5'-GCC TTT CAA TCC ACA TTA GGG -3' (reverse).

3.2.7 Pathway-based Analysis of hMeSeal-seq

Genomic region lists were generated to represent relative absolute hydroxymethylated regions identified in RWPE-1 and 22Rv1 cells. Functional and pathway enrichment analysis were performed on Genomic region lists using the Genomic Regions Enrichment of Annotations Tool (GREAT) [284]. GREAT analyses were performed using the binominal test over genomic regions to assess for significance. Subsequently, GREAT results were organized into bar graphs using the recommended statistic of $-\log_{10}$ (binomial raw p-value) for ranking. Ontologies shown were most significantly enriched and the following ontology
categories were omitted from analysis, as the resulting outputs were not biologically relevant: Phenotype Data and Human Disease, Gene Expression, Regulatory Motifs, Gene Families (except for HGNC Gene Families).
3.3 Results

3.3.1 Relative Absolute Analysis of hMeSeal-seq in Prostate Cell lines, Benign and Cancerous

The genome-wide detection technique I used to discover hydroxymethylation marks is termed hMeSeal-seq, as it combines 5hmC-enrichment, by hydroxymethyl selective chemical labeling (hMeSeal), with NGS technology (Fig. 9). Due to the novelty of this detection technique, we first used Repitools package to perform a diagnostic test on the enrichment of hydroxymethylated DNA in enriched samples compared to input, or non-enriched sample. The shape of the input sample curve is representative of a random and sporadic sample, equally enriched across the genome, where any shift from this curve is determined as enrichment at specific genomic loci [330]. I observed that enrichment levels varied between RWPE-1 triplicates, as shown by the extent of curve shift between input and replicate samples in Figure 10a. In addition, 22Rv1 results indicated two replicates (Rep 1, 3) were not enriched compared to input, as these curves coincided with the input sample curve in Figure 10b. Once I performed MACS analysis [280], the results confirmed that replicates were behaving differently, as varying numbers of absolute regions of hydroxymethylation were called (Fig. 10d). Ultimately, these analyses prompted the continuation of further comparisons solely using one replicate (Rep 2) from RWPE-1 and 22Rv1 cells, as these demonstrated the highest levels of enrichment (Fig. 10a-d, shifted yellow lines).

To investigate the genome-wide pattern of relative absolute hydroxymethylation, I used the MACS analysis of Replicate 2 from RWPE-1 and 22Rv1 cells and observed a dramatic loss in global hydroxymethylation marks in 22Rv1 cells compared to RWPE-1 (Fig. 11a). That is, a total of 34,654 hydroxymethylated regions were called in RWPE-1 cells and 16,427 regions in 22Rv1. Further, there was a low number of hydroxymethylated regions that overlapped CpG islands in both RWPE-1 and 22Rv1 cells (1,745 and 295 regions, respectively), however there was a higher proportion found in RWPE-1. Additionally, I correlated these absolute regions of hydroxymethylation with an ENCODE DNase-seq dataset, which mapped the DNase I hypersensitivity sites (HS) in RWPE-1 cells (described in Section 3.2.4). This dataset is of interest, as DNase I HS commonly mark open chromatin [29]. The correlative analysis revealed that overall, a greater number of hydroxymethylated regions in RWPE-1 were overlapping DNase I HS compared to 22Rv1 (Fig. 11a, 5.0% in RWPE-1, 1.8% in 22Rv1). Moreover, although the number and
proportion of CGI and DNase I HS overlap with hydroxymethylated marks was quite similar, I found that the correlation became stronger for DNase I HS, when these features were expanded. That is, I considered the overlap of 5hmC marks within a 5kb distance of CGIs and DNase I HS, and this identified a greater number of hydroxymethylated regions overlapping DNase I HS compared to CGIs in both cell lines (Fig. 11b, 32.5% in RWPE-1, 20.9% in 22Rv1).

I investigated the distribution of these absolute hydroxymethylated regions, in RWPE-1 and 22Rv1 cells, according to genomic location: genic/gene body and intergenic. Further, genic regions were subdivided into exonic and intronic regions. From this analysis, I observed that both cell lines showed a similar distribution pattern, in that intronic (38.6% in RWPE-1, 34.3% in 22Rv1) and intergenic regions (50.8% in RWPE-1, 60.7% in 22Rv1) had high enrichment for 5hmC marks (Fig. 11c). Interestingly, the global loss of hydroxymethylation in 22Rv1 cells was evenly distributed across the genome as I found that RWPE-1 had a greater number of hydroxymethylated regions across all genomic features (Fig. 11c). Additionally, when 5hmC marks were stratified by overlap with RWPE-1 DNase I HS (ENCODE), the highest percentage of hydroxymethylation within a genomic feature was identified in exons and the lowest percentage in intergenic regions, in both RWPE-1 and 22Rv1 cells (Fig. 11d).
Figure 9. hMeSeal-seq methodology pipeline. (Step 1-5b) Cell line DNA was enriched for 5hmC marks using hydroxymethyl selective chemical labeling (hMeSeal) and (Step 6) was submitted for next-generation sequencing (NGS) using the Illumina HiSeq 2500. (Step 7-8) Bioinformatics analysis was performed using MACS and DiffBind software, while (Step 9) pathway-based analysis was performed using GREAT program. Diagram by Ho LT and Lo C.
Figure 10. Diagnostic enrichment plots of hMeSeal-seq reads and MACS analysis of RWPE-1 and 22Rv1 cells. (a-b) Diagnostic plots comparing the curve of input (non-enriched) sample (black line) to biological replicates 1-3 (green, yellow, red line, respectively). (c-d) Total number of relative absolute regions/peaks of hydroxymethylation, called by MACS in three biological replicates.
Figure 11. Relative absolute regions of hydroxymethylation identified in RWPE-1 and 22Rv1 cells using hMeSeal-seq. (a) Total absolute regions of hydroxymethylation, stratified by CpG island (CGI) and RWPE-1 DNase-seq (ENCODE) overlap, where (b) CGIs and DNase I hypersensitivity sites (HS) were expanded 5kb up- and downstream. (c) Regions were categorized into genomic features and (d) further stratified by RWPE-1 DNase I HS overlap.
3.3.2 Differential Analysis of hMeSeal-seq in Prostate Cell lines, Benign and Cancerous

Using the MACS analyses of one hMeSeal-seq replicate from RWPE-1 and 22Rv1 cells compared to respective input samples, I performed a pilot differential analysis between these two datasets using DiffBind to determine differentially hydroxymethylated regions (DHMRs). Overall, 20,490 DHMRs were identified for RWPE-1 cells and 14,362 DHMRs for 22Rv1 (Fig. 12a), showing a decrease in hyper-hydroxymethylated regions in 22Rv1 compared to RWPE-1 cells. Further, I observed a lower proportion of 22Rv1 DHMRs occurring within CpG islands (CGIs) and RWPE-1 DNase I hypersensitivity sites (HS) compared to RWPE-1 cells (Fig. 12a), paralleling the pattern shown in absolute regions of hydroxymethylation (Fig. 11a). Subsequently, DHMRs were categorized into genomic features, where promoter regions were categorized separately from genic regions. I found promoter regions had relatively low levels of hyper-hydroxymethylation, in both RWPE-1 and 22Rv1 cells (Fig. 12b). Further, RWPE-1 cells showed that genic regions were most abundant in DHMRs, followed by intergenic regions. The opposite pattern was observed in 22Rv1 cells, in which most DHMRs were found in intergenic regions (Fig. 12b). Additionally, genomic locations were stratified by RWPE-1 DNase I HS overlap, revealing that promoters containing DHMRs, in both RWPE-1 and 22Rv1 cells, had the highest correlation compared to all other genomic features (Fig. 12c).
Figure 12. Preliminary differentially hydroxymethylated regions (DHMRs) described in RWPE-1 and 22Rv1 cells using hMeSeal-seq. (a) Total number of DHMRs, stratified by CpG island (CGI) and RWPE-1 DNase-seq (ENCODE) overlap. (b) DHMRs were categorized into genomic features and (c) further stratified by RWPE-1 DNase I HS overlap.
3.3.3 Validation of hMeSeal-seq Results using Dot Blot Assay

To verify the number of relative absolute regions of hydroxymethylation detected by hMeSeal-seq technique, I used the dot blotting assay to analyze global levels of 5hmC. I first optimized, the dot blotting technique using positive controls: human genomic DNA oligos, in which 100% of cytosine bases were hydroxymethylated; and mouse brain genomic DNA, in which 0.548% of cytosine bases were hydroxymethylated (Methods described in Section 3.2.5).

Dot blotting analysis of genomic DNA, extracted from two biological replicates of RWPE-1 and 22Rv1 cells, revealed both cell lines contained 5hmC marks, as expected from hMeSeal-Seq analysis (Fig. 13a-b). Additionally I found that global 5hmC levels in 22Rv1 were reduced compared to RWPE-1 cells (Fig. 13b; average 5hmC intensity at 200 ng dilution; RWPE-1 100% and 22Rv1 17.5% ± 0.07). The methylene blue staining was used as a loading control for total genomic DNA. These preliminary results validated the decrease in number of MACS peaks called from hMeSeal-seq analyses.
Figure 13. Dot blot analysis of global levels of hydroxymethylation in RWPE-1 and 22Rv1 cells. (a) Representative micrographs of dot blotting assays of RWPE-1 and 22Rv1 genomic DNA, where 500ng was serially diluted and incubated with anti-5hmC antibody. Methylene blue staining was used as total genomic DNA loading control. (b) Global detection and quantification of 5hmC marks at 200ng dilution, relative to 5hmC intensity in RWPE-1 cells are shown as ± SEM from 2 biological replicates.
3.3.4 Validation of hMeSeal-seq Results using hMeSeal-qPCR

In order to validate preliminary differentially hydroxymethylated regions (DHMRs) identified by hMeSeal-seq technique, I used the locus-specific detection method of hMeSeal-qPCR. That is, genomic DNA extracted from RWPE-1 and 22Rv1 cells was sonicated and subjected to 5hmC-enrichment using the Hydroxymethyl Collector kit, and these enriched fractions were analyzed using RT-qPCR.

First, I identified biologically relevant DHMRs to investigate. I used MSigDB gene sets (KEGG Pathways in Cancer, KEGG Prostate Cancer) to generate a list of genes previously implicated in cancer and these were compared with a gene list compiled of nearest genes mapping to DHMRs discovered in RWPE-1 and 22Rv1 cells. I found common genes between these two lists and filtered the identified DHMRs further, by genomic location. That is, I continued analysis on genes with DHMRs occurring within genic regions. Subsequently, I designed primers to amplify various regions along these candidate DHMRs, as the specific cytosine base bearing a hydroxymethylation mark cannot be detected using hMeSeal-seq technology. These primers were tested on a hMeSeal sample-enriched fraction and a negative control-enriched fraction, using RWPE-1 and 22Rv1 genomic DNA in both reactions. In addition, primers were subjected to two input samples of RWPE-1 and 22Rv1 genomic DNA: one containing 1% (25ng) of the DNA subjected to a hMeSeal reaction; and the second containing 0.04% (1ng). These input samples were used in order to analyze RT-qPCR results using the delta-delta Ct method. Moreover, the 1% (25 ng) input sample was used as a mock reference gene, while the 0.04% (1ng) input sample was used for fold-change normalization.

In preliminary studies, I identified DMHRs in the gene body of cyclin-dependent kinase inhibitor 2B (CDKN2B) and androgen receptor (AR) to be validation candidates in RWPE-1 and 22Rv1 cells, respectively (Table 4). In one biological replicate, I observed a positive fold-change of CDKN2B in RWPE-1 sample-enriched fraction when compared to an RWPE-1 negative control-enriched fraction and a 22Rv1 sample-enriched fraction (RWPE-1 FC=3.7, 22Rv1 FC=0.3, data not shown). Similarly, I found a positive fold-change of AR in 22Rv1 sample-enriched fraction (22Rv1 FC=5.6, RWPE-1 FC=0.2, data not shown). However, when this experiment was repeated in two additional biological replicates, the fold-change enrichment was not significant for either CDKN2B or AR (Fig. 14a-b).
Table 4. Pilot analysis of DHMRs identified in RWPE-1 (blue) and 22Rv1 (red) cells, discovered using hMeSeal-seq, which were validated using hMeSeal-qPCR.

<table>
<thead>
<tr>
<th>Chr</th>
<th>Start</th>
<th>End</th>
<th>Gene name</th>
<th>Gene feature</th>
<th>CpG island overlap</th>
<th>Abs Fold-change</th>
</tr>
</thead>
<tbody>
<tr>
<td>chr9</td>
<td>22008729</td>
<td>22009011</td>
<td>CDKN2B</td>
<td>Gene body</td>
<td>ID_11555</td>
<td>4.87036472</td>
</tr>
<tr>
<td>chrX</td>
<td>66915387</td>
<td>66915893</td>
<td>AR</td>
<td>Gene body</td>
<td>No</td>
<td>11.62570884</td>
</tr>
</tbody>
</table>

DHMR: differentially hydroxymethylated region, chr: chromosome, Abs: absolute.

Figure 14. Preliminary validation of differentially hydroxymethylated regions (DHMRs) identified in RWPE-1 and 22Rv1 cells, using hMeSeal-qPCR. (a) CDKN2B was found to be a DHMR in hMeSeal-seq analysis of RWPE-1 cells and (b) AR was found to be a DHMR in 22Rv1 cells. No significant difference between cell lines was observed for both genes. $2^{-\Delta\Delta Ct}$ method normalized to 0.04% of input DNA, where error bars indicate SEM across 3 biological replicates.
3.3.5 Pathways Associated and Potentially Regulated by Relative Absolute Regions of Hydroxymethylation in Prostate Cell lines

To investigate pathways potentially regulated by relative absolute regions of hydroxymethylation in RWPE-1 and 22Rv1 cells, I used GREAT database for pathway-based analysis on regions categorized by genomic location, including exon, intron, or intergenic regions. The intergenic sub-group was further stratified according to RWPE-1 DNase I HS (ENCODE) overlap.

The biological function of hydroxymethylation in both normal and cancer state, remains to be elucidated. Firstly, genomic regions of particular interest are gene bodies, as they are abundant in 5hmC marks and have been implicated in gene expression regulation [144]. Thus, I sought to determine whether prostate cells incur similar regulation by hydroxymethylation. I performed this analysis by separating genic 5hmC marks, in RWPE-1 and 22Rv1 cells, into those occurring within exonic and intronic regions, to examine any similarities or differences in pathways analysis results. I observed that absolute regions of hydroxymethylation within exon regions of RWPE-1 cells were enriched for translational regulation, DNA repair, and RNA polymerase II transcription (binomial p<1.19E-07) (Fig. 15a), while hydroxymethylated exonic regions in 22Rv1 resulted in pathway enrichment for mismatch repair, histone transcription, genes up-regulated in metastatic prostate tumors, and carnitine transport (binomial p<1.05E-08) (Fig. 15b).

Additionally, when absolute regions of hydroxymethylation were stratified by overlap with intronic regions, I found that RWPE-1 cells showed pathway over-representation of GTPase activity and RNA polymerase II transcription (binomial p<4.43E-11) (Fig. 15c), while 22Rv1 were enriched for a variety of biological functions such as mismatch repair, ATP metabolism, immune response, and several prostate tumor gene expression studies as annotated by MSigDB databases (binomial p<1.81E-02) (Fig. 15d).

Pathways enriched in intronic regions of both RWPE-1 and 22Rv1 cells were cellular organization and MAPK binding (Fig. 15c-d).

Secondly, the presence of 5hmC within intergenic regions has been shown to mark cis-regulatory regions and euchromatin [145, 150, 331]. To further address these patterns in prostate cells, I correlated intergenic occurring hydroxymethylation marks with RWPE-1 DNase I hypersensitivity sites (HS) (ENCODE), which are known to mark open chromatin. Pathways analysis of absolute hydroxymethylated
regions overlapping intergenic regions revealed that RWPE-1 cells were enriched for housekeeping genes expressed in normal tissues, apoptosis regulation, p38 signaling, BMP signaling, Notch signaling, and RNA polymerase II transcription, among others (binomial p<1.76E-02) (Fig. 16).
Figure 15. Pathways associated with relative absolute regions of hydroxymethylation in RWPE-1 and 22Rv1 cells identified by hMeSeal-seq. Over-represented pathways for absolute regions of hydroxymethylation overlapping (a-b) exonic regions and (c-d) intronic regions, determined by GREAT analysis.
Figure 16. Pathways associated with relative absolute regions of hydroxymethylation in RWPE-1 cells identified by hMeSeal-seq, and further correlated with RWPE-1 DNase-seq dataset (ENCOD). Over-represented pathways for absolute regions of hydroxymethylation overlapping intergenic regions occurring within RWPE-1 DNase I hypersensitivity sites (HS), determined by GREAT analysis.
3.4 Discussion

In this chapter, I discovered genome-wide hydroxymethylation patterns in prostate cell lines, RWPE-1 (benign) and 22Rv1 (cancerous), using a novel technique in sequencing \(5\text{hmC}\) marks. This affinity-based method couples hydroxymethyl selective chemical labeling (hMeSeal) with next-generation sequencing (NGS) and is termed hMeSeal-seq. Relative absolute analyses revealed that, although a global loss of hydroxymethylated marks occurred in cancer compared to normal cells, the genomic distribution of these regions were similar across both cell lines. However, when stratified by occurrence within RWPE-1 DNase I hypersensitivity sites (HS), exonic regions demonstrated the highest contribution of correlative hydroxymethylation marks. Intronic and intergenic region analysis revealed that in both normal and cancer cells, \(5\text{hmC}\) marks were abundant in these regions. Further, I investigated pathways potentially implicated in DNA hydroxymethylation-mediated regulation. Through pathway-based analyses, I observed that when genomic regions were further subdivided, exons and introns enriched for distinct biological processes both within a cell line and across normal and cancer cells. Additionally, when RWPE-1 intergenic hydroxymethylation was correlated with RWPE-1 DNase I HS, over-represented pathways suggest specific regulation of prostate cancer-associated signaling pathways.

As hMeSeal-seq is a recently developed methodology for 5hmC-sequencing [196], I experienced several difficulties while optimizing the protocol. Previous studies have shown that 5hmC levels are low in prostate tissue and I have found these levels to be increasingly lower in cell culture, which is a recognized phenomenon [153]. For example, we subjected genomic DNA, from RWPE-1 and 22Rv1 cells, to liquid chromatography-tandem mass spectrometry (LC-MS/MS) (Marsden Lab, St. Michael’s Hospital) to evaluate global levels of 5hmC. We observed that LC-MS/MS method, although highly sensitive [332], was unable to detect such a low abundance of hydroxymethylation suggesting that 5hmC is found in less than 0.01% of the genome. In order to enrich for 5hmC marks, I used the Hydroxymethyl Collector Kit, as it was shown to be the most efficient method for capturing this mark at a locus-specific and genome-wide level, compared to other techniques available [333]. Indeed, I was able to enrich for 5hmC and the DNA yield from the kit was approximately 0.03% of input DNA. Although the percentage of enrichment was reliable, many reactions were pooled to obtain sufficient DNA required in library preparation. During the first
submission for sequencing, two replicates of 22Rv1 failed library preparation, despite passing the prior quality control step. The cause of failure is still unknown, as all samples were treated identically. As the length of enriched regions reduces, along with the total amount of enriched DNA, the success of library preparation can become more sporadic and unpredictable [334]. Meanwhile, I repeated the Hydroxymethyl Collector Kit and re-submitted these samples for sequencing. However, the re-submitted samples of 22Rv1 (Rep 1,3) were not enriched for 5hmC when compared to input samples (Fig. 10a-b) and this was validated by a lesser number of peaks called by MACS (Fig. 10c-d). The reason for non-enrichment was traced to a change in protocol by the company and incorrectly labeled T4-phage β-glucosyltransferase. Due to these circumstances, I removed two replicates of 22Rv1 from the hMeSeal-seq analyses. Remaining, are three replicates of RWPE-1, one replicate of 22Rv1, and input samples.

In order to determine relative absolute regions of hydroxymethylation, I performed analysis on MACS datasets of Replicate 2 from both RWPE-1 and 22Rv1 cells. This was an appropriate selection of replicates, as Replicate 2 was observed to be the most robustly enriched compared to other replicates of RWPE-1 (Fig. 10a), and further diagnostic analysis revealed that Replicate 2 of RWPE-1 was more highly enriched compared to that of 22Rv1 (Supplemental Fig. 2). Subsequently, I sought to correlate this discovery dataset of absolute hydroxymethylated regions with other known cis-regulatory marks in RWPE-1 or 22Rv1. Through literature and GEO datasets search, I uncovered a lack of diversity in the epigenetic profiling completed for prostate cell lines, in that DNA methylation studies were currently the most common. However, the ENCODE project had collected data from one DNase-seq study, which profiled genome-wide DNase I HS in RWPE-1 cells (GEO accession: GSM1008595). It has been shown that DNase I HS are indicators of euchromatin and associate with several gene features, such as promoters, to enhance gene expression [29]. As 5hmC marks have also been implicated in expression up-regulation [144], I thought it fitting to correlate the datasets. From this analysis, I observed that absolute regions of hydroxymethylation identified in RWPE-1 showed increased correlation with RWPE-1 DNase I HS compared to the 22Rv1 dataset (Fig. 11a). Moreover, I instilled further confidence that this correlation was not due to chance, by increasing the boundaries of both CGIs and DNase I HS, and revealed that the latter resulted in a more robust correlation compared to CGIs (Fig. 11b). This suggested that DNase I HS were more functionally relevant in terms of DNA hydroxymethylation genomic distribution, which has been demonstrated in fetal and adult liver cells [335]. Further, human ESC CGIs have portrayed a depletion of 5hmC mark occurrence [201, 331], although the biological significance was not known. Moreover, the low
abundance of 5hmC marks observed in CGIs of 22Rv1 cells (Fig. 11a-b), suggested that although 5hmC is an oxidative derivative of 5mC, its functional role is unlike that of DNA methylation. And if it does indeed, regulate a similar function, it is through an alternative mechanism that is not highly dependent on CGIs.

The resulting absolute hydroxymethylated regions identified in RWPE-1 and 22Rv1 cells revealed a global loss in hydroxymethylation in 22Rv1, supporting immunohistochemical analyses reporting dramatic loss of 5hmC marks in prostate tumors compared to normal [129, 155, 156]. In terms of locus-specific patterns of hydroxymethylation, evidence has shown that 5hmC is enriched in highly expressed gene bodies in human neural tissues [196, 336], and at enhancer regions, repetitive elements, and other cis-regulatory elements [145, 150, 331]. In my study, I observed a high occurrence of hydroxymethylation marks within introns and intergenic regions (Fig. 11c), which supported previously described patterns. However, it has been shown that genic regions are more highly enriched for 5hmC marks compared to intergenic regions, and this discordance may be explained by the elevated percentage of intergenic regions, which envelop a majority of the genome. That is, the number of hydroxymethylated regions identified for genic and intergenic regions were comparable, suggesting that in proportion to total genomic distance, genic regions of prostate cells are indeed enriched for 5hmC. In an analogous comparison, I found greater absolute numbers of 5hmC within intronic regions compared to exonic regions, however, introns are more commonly found amongst the genome. Similarly, this suggested that exon enrichments were under-represented and indeed, could be abundant in hydroxymethylation. Interestingly, when 5hmC marks were stratified by DNase I HS overlap, exonic regions were determined to incur the highest percentage of correlation (Fig. 11d), adding more evidence to the importance of exonic hydroxymethylation. It is important to note that the absolute analysis did not include annotation of promoter regions and thus, promoters are included in genic and exonic region analyses. This implies that the percentage correlation in exonic regions could be due, in part, to the presence of promoter regions.

Subsequently, I performed a preliminary differential analysis using DiffBind algorithm [283] and identified potential differentially hydroxymethylated regions (DHMRs). DHMRs represent regions of higher hydroxymethylation signal in one cell line compared to the other, and thus can be described as hyper-hydroxymethylated. On the other hand, hyper-hydroxymethylated regions in one cell line may be defined as hypo-hydroxymethylated in the opposite cell line, allowing for a detailed analysis of normal
and cancer gain or loss of 5hmC. In this analysis, subtle distribution differences were observed in comparison to the absolute analysis, however overall, intron and intergenic regions revealed the highest abundance of DHMRs. Moreover, RWPE-1 showed a higher number of DHMRs in intron regions than intergenic, while 22Rv1 cells showed the opposite trend (Fig. 12b). These results may suggest an important role for intronic 5hmC marks in normal cells, while cancer cells lose these intronic marks and gain intergenic 5hmC. Studies have shown low abundance of hydroxymethylation in intron regions in mouse brain and liver tissue [333], however a cancer-specific gain of intergenic hydroxymethylation has not been reported. Lastly, DHMRs were annotated for promoter regions and these features demonstrated the highest level of correlation with RWPE-1 DNase I HS, in both cell lines. These results may define a relevant relationship between accessible chromatin and DNA hydroxymethylation, as bivalent promoters in mouse ESCs have shown abundance in 5hmC marks [144, 165, 197].

A limitation of hMeSeal-seq technique is bias generated from library preparation and NGS procedure. One such bias I experienced was a high rate of duplication in the number of reads aligned during MACS analysis. This finding is generally viewed as a concern towards false positives as the peaks being called during MACS were not broad, but rather small areas of enrichment. High rates of duplication could be due to over-amplification during library preparation, a phenomenon that becomes difficult to manage when starting with low DNA quantities and low diversity in region enrichment [337]. Optimizing this technique would establish an effective quality control step prior to NGS, where enriched samples could be interrogated for “positive control” hydroxymethylated genes in prostate cancer. This is an imperative goal as no single gene has been reported to harbor 5hmC marks in prostate, normal or cancer.

To validate the discovery set from hMeSeal-seq methodology, dot blotting analysis was used to investigate findings from the absolute analysis. The dot blotting technique was developed as a simple and sensitive technique to detect global levels of DNA hydroxymethylation using an anti-5hmC antibody [194]. According to pilot analysis, the dot blot assay was able to reproduce the pattern of global reduction in 5hmC marks observed in 22Rv1 compared to RWPE-1 cells (Fig. 13a-b). In addition, hMeSeal-qPCR technique, which used identical 5hmC-enrichment conditions as hMeSeal-seq followed by RT-qPCR, was used to analyze specific loci of interest from the pilot differential analysis. This method allowed for analysis within originally enriched samples and removed any biases generated from library preparation and NGS procedure. In my trial analysis, I found that CDKN2B and AR were potential candidate genes
(Table 4), however these were not significant across three biological samples (Fig. 14a-b). These results could be due to the lack of power in differential analyses as only one replicate from each cell line were used, or could suggest that enrichment between different biological replicates varies. Causes of enrichment variability could be differences in cell culture passage number, duration of culture, and cell cycle stage.

Few studies have investigated the functional role of DNA hydroxymethylation and the biological processes regulated by these marks. These marks have been strongly implicated in regulating gene transcription [27, 28, 78-80]. Relative absolute analysis of hydroxymethylation marks clearly highlighted the importance of genomic location. Accordingly, I approached pathway-based analyses by investigating exon and intron regions independently. The results revealed that hydroxymethylation within exons of normal cells enriched for translational regulation and RNA polymerase II transcription, whereas cancer cell enrichment was preferential towards mismatch repair and histone transcription, and carnitine transport (Fig. 15a-b), which is involved in cell death regulation in the context of prostate cancer [338]. These findings are interesting as cancer-associated hyper-hydroxymethylation may be up-regulating key pathways that allow the cancer cell to survive. Additionally, intronic regions in normal cells showed higher enrichment for RNA polymerase II transcription, while cancer cells did not enrich strongly for a single pathway. Rather, intron hydroxymethylation in cancer cells were associated with a variety of functions and most notably, were enriched for prostate tumor gene expression studies (Fig. 15c-d). These intronic analyses suggested that the involvement of 5hmC-regulated gene expression occurs within these regions rather than exonic regions. Interestingly, exonic regions were most highly enriched for translational regulation, which has not been reported as a function of 5hmC marks. This suggests a level of functional or protein regulation and may co-exist alongside transcriptional regulation roles of hydroxymethylation. Finally, cis-regulatory regions marked with 5hmC have been associated with transcription factor binding sites and enhancer functions in human ESCs [145, 146]. Through pathways-based analysis, I examined intergenic regions overlapping RWPE-1 DNase I HS in normal cells, and found over-representation of several pathways involved in prostate cancer progression such as p38 signaling [339], BMP signaling [340], and Notch signaling [341], in addition to RNA polymerase II transcription and apoptosis regulation (Fig. 16). This suggested that in normal cells, intergenic patterns of DNA hydroxymethylation are pertinent in the proper maintenance of these pathways, and that loss of these localized 5hmC marks in cancer may contribute to the development of the disease.
Chapter 4
Conclusions and Future Directions

4.1 Conclusions

The goal of this thesis project was to characterize the genome-wide distribution of epigenetic modifications contributing to the progression of prostate cancer, namely DNA methylation and hydroxymethylation. In Chapter 2, I defined differentially methylated regions (DMRs), which distinguished between RWPE-1 (benign) and 22Rv1 (cancerous) cells, using a powerful next-generation sequencing approach termed MBD-seq. In Chapter 3, I discovered relative absolute regions of hydroxymethylation in both normal and prostate cancer cells with a recently developed and innovative strategy termed hMeSeal-seq. These epigenetic analyses were paired with subsequent pathways-based analysis to determine the biological processes regulated by genomic feature-specific 5mC and 5hmC mark patterns. This study highlights the ability of high-throughput sequencing data and intensive bioinformatics analyses, to characterize epigenomic signatures, which can differentiate between normal and cancer cells. I found that my MBD-seq analysis was able to validate previously performed genome-wide methylation studies in prostate cancer, however, this is the first study to investigate locus-specific hydroxymethylation patterns in prostate cells and thus requires further replication and validation studies.

MBD-seq technique has been shown to provide accurate detection of DMRs and in-depth sequencing coverage at feasible costs [300]. Using this technique I found that significant differentially methylated regions (DMRs) occurring within promoter regions robustly distinguished between normal and cancer cells when stratified by CpG island overlap. Further, integrative pathways analysis was able to identify novel and previously implicated candidate biomarkers in prostate cancer. Interestingly, non-CpG island methylation marks were abundant in genic and intergenic regions and their functional significance was highlighted when pathways analyses revealed that normal and cancer DMR patterns enriched for discrete biological processes. For example, in genic regions, these processes included mRNA transcription and hormone biosynthesis in normal cells, while cancer cells were enriched for known PCa pathways such as
IGF1 and EGF receptor signaling. Indeed, the identified pathways warrant further validation to interpret their potential 5mC-mediated regulation and consequential involvement in disease development.

Recently, hMeSeal-seq was developed to detect 5hmC marks by specifically protecting the modified base through a glucosylation reaction [196]. In comparison with other capture-based methods, hMeSeal-seq performs with the least amount of bias and highest level of enrichment [333]. In performing this technique, I observed a global loss in hydroxymethylation in cancer cells compared to normal, and discovered a high number of 5hmC marks within intronic and intergenic regions. When these regions were subjected to pathway-based analyses, normal and cancer cells demonstrated distinct over-representation of regulated processes, suggesting functionally relevant hydroxymethylation signatures, which are condition-specific. Interestingly, when intergenic 5hmC marks in normal cells were correlated to DNase I hypersensitivity sites (HS), multiple pathways associated with prostate cancer progression were enriched, such as BMP and Notch signaling. These results suggest a cis-regulatory relationship between 5hmC and DNase I HS marks, which may be required to maintain normal prostate cell function. Accordingly, the mechanism and consequence of this regulatory role requires further investigation.

The hMeSeal-seq results are quite preliminary, in that I faced many obstacles in optimizing the technique. Prostate cells have been shown to contain a low abundance in 5hmC marks compared to other cell types such as ESCs and neural tissues, and cell culture conditions are known to further depress these marks [153]. I was unable to detect global 5hmC levels using LC-MS/MS technique, however I was able to perform preliminary validation by dot blotting assay using anti-5hmC antibody. In addition, pilot studies used hMeSeal-qPCR to assess specific DHMR loci and revealed lack of significance across biological replicates. This could be due to enrichment variability or sub-optimal number of replicates subjected to differential analysis. Finally, low DNA yields from hMeSeal technique and consequently low 5hmC enrichment levels were prone to NGS biases, such as high duplication rates. These biases may be overcome by optimizing library preparation conditions.

In summary, this thesis illustrates the power of next-generation sequencing technology and the advent of -omic based epigenetic studies to combat the complexity of a highly regulatory genome. Our MBD-seq analyses were able to depict an in-depth methylation landscape, which was capable of identifying additional 5mC biomarker candidates within CpG island promoters through pathway-based analyses. The
hMeSeal-seq discovery dataset defines expected enrichment of 5hmC in cis-regulatory regions and paves the way for future studies to continue strengthening the trends observed. The promise of biomarker potential for 5hmC is evident, as dramatic loss during cancer development was validated and this poses opportunity for therapeutic avenues, where a 5hmC signature could be re-established from the aberrant cancer condition. Moreover, the biological role of 5hmC is an important aspect to uncover, as its relationship with DNA methylation remains to be resolved in prostate cancer, and could be pivotal in understanding the pathogenesis of this epigenetic disease.

### 4.2 Future Directions

There is growing interest to integrate next-generation sequencing technology into the clinical setting, giving perspective to the applicability of these types of big data studies. The initial matter is the ability to reproduce MBD-seq and hMeSeal-seq techniques in biological samples. In this thesis, my disease model of study were cell lines which provide several benefits when optimizing a novel technique such as unlimited DNA resources, controlled cell culture conditions, and reproducible results. Based on promising findings, these techniques should be further applied to patient samples. That is, biological samples, such as tissue, serum, and urine are appealing sources of DNA. Several groups have reported success in applying MBD-seq to patient tissue derived from schizophrenia cases and head and neck tumors [342, 343]. Whole blood samples were studied in aging patients [344], however urine samples have yet to been subjected to this methodology. In terms of 5hmC investigation, hMeSeal-seq has been applied to tissue derived from liver samples [335], and has not yet been examined in serum and urine samples. The ability to use fixed-paraffin archival samples for 5hmC-enrichment and NGS has yet to be explored.

A major limitation of 5mC and 5hmC enrichment techniques is the microgram requirement of sample DNA, which is a large amount of tissue when using archived samples. Alternative techniques detecting genome-wide methylation marks, such as MeDIP-seq and RRBS-seq, have been used by groups who reported the use of one microgram of tissue [300]. The major alternative techniques for detecting 5hmC marks are hMeDIP-seq and single-nucleotide detection using TAB-seq or oxBS-seq. However, the latter two techniques are costly, labor intensive, and limited to few distribution companies. Ultimately, 5hmC
detection techniques are still being optimized and developed, for example, by utilizing microarray analyses [345].

In the hMeSeal-seq study, future studies would include completing additional replicates for 22Rv1 (cancerous) cells in order to generate statistically significant differentially hydroxymethylated regions (DHMRs) between normal and cancer cells. However, prior to this, further optimization of library preparation conditions is necessary to reduce the rate of duplication. In the event that this cannot be achieved, an advanced bioinformatics normalization program may be required and developed for this type of genome-wide sequencing read pattern. In terms of differential analysis, there is room to improve the pipeline. Novel bioinformatics packages are constantly being released, for example, PePr program was recently designed to determine differential epigenetic patterns [346]. It is understood that bioinformatics analyses of epigenomic data is currently being optimized. Numerous algorithms are available and used by different groups, and thus, no defined pipeline exists yet for genome-wide DNA methylation or hydroxymethylation analyses.

Validation studies for MBD-seq could continue by comparing findings to other publicly available datasets for RWPE-1 and 22Rv1 cells, as well as patient tumor samples. For example, the The Cancer Genome Atlas (TCGA) recently completed methylome analysis of prostate tumor samples, and this is a valuable resource for 5mC biomarker validation. In terms of hMeSeal-seq validation, primers can be redesigned to target MACS relative absolute regions of hydroxymethylation in RWPE-1 and 22Rv1, rather than DHMRs. There is a normalization bias within DiffBind analysis that performs optimally with MACS replicates, of which we did not have for the hMeSeal-seq study. In addition, a second locus-specific validation can be performed using TAB-seq technique, where the region of interest is sequenced at a single-base pair resolution, rather than the entire genome. This would provide ultimate confidence that hMeSeal-seq is capable of identifying regions enriched for 5hmC.

As 5hmC is an oxidative derivative of 5mC, the ultimate correlation would compare MBD-seq and hMeSeal-seq datasets. This would provide a detailed epigenetic landscape for prostate cells, both normal and cancer, and would elucidate genomic features prone to DNA demethylation in cancer. Additionally, the role of TET protein regulation of these marks is a critical area of interest. It has been shown that TET proteins are reduced in solid tumors, including prostate [129] [328]. However, the TET expression profile
in prostate cell lines has not been investigated. Thus, I performed a pilot study and observed a reduction in TET2 and TET3 messenger RNA expression in prostate cancer cell lines: 22Rv1, LNCaP, DU-145, and PC3, compared to normal, RWPE-1 cells (Supplemental Fig. 3b-c). As TET2 was highly expressed in normal cells, I hypothesized that it may be responsible for maintenance of 5hmC abundance in normal and resulting hypo-hydroxymethylation may occur when TET2 is lost in cancer. Following, I performed a stable knockdown (KD) of TET2 in RWPE-1 cells. These TET2 KD RWPE-1 clones demonstrated significant knockdown efficiency of up to 73% (pval<0.01) compared to a scrambled control clone (Supplemental Fig. 4). The knockdown efficiency reflects mRNA expression of TET2, transcript variant 1, which is translated to the most common isoform of the protein in normal tissue [168]. Future studies will use MBD-seq and hMeSeal-seq techniques to describe the resulting change in the epigenome in comparison to the parental dataset, and investigate genomic features or genes epigenetically regulated by TET2.

Pathway-based analyses of both MBD-seq and hMeSeal-seq regions identified enrichment of RNA polymerase II transcription in genic and intergenic regions. Although DNA methylation has been defined for its role in gene repression when occurring in CpG island promoters, the inverse relationship is shown when 5mC occurs in gene bodies [43, 44]. Additionally, DNA hydroxymethylation has been found to mark genic regions of highly expressed genes [144] and we have shown that they also correlate with DNase I HS, alluding to 5hmC’s role in mRNA transcription. Accordingly, we are currently performing RNA-seq analysis of normal and cancer cells, to allow correlation of epigenetic and gene expression patterns. This analysis will help characterize the pathways defined from my epigenetic studies, and further focus on genes that are not only biologically relevant, but also transcriptionally regulated by 5mC and 5hmC marks.
Appendix

S1.1 Supplementary Material and Methods

S1.1.1 Endogenous TET2-knockdown in Cell lines

Four shRNA plasmids targeting both transcript variants (NM_017628.4, NM_001127208.2) of human TET2 and one control non-silencing shRNA were prepared from glycerol stocks (Sigma-Aldrich) according to the protocol provided. Plasmids of shRNA (1 ug) and 3 ul of X-tremeGENE 9 DNA Transfection Reagent (Roche) were diluted in 100 ul Opti-MEM I media (Gibco). This transfection mixture was incubated with plated RWPE-1 cells (4.0 x10^5 cells) according to the protocol provided. Transfection mixture was incubated for 24 hr to obtain stable knockdown. Stable colonies were trypsinized and re-plated, then further serially diluted with K-SFM media supplemented with puromycin (0.2 ug/ml), until colonies propagated.

S1.1.2 Real Time Quantitative PCR (RT-qPCR)

Total RNA of TET2 knockdown (KD) RWPE-1 cell lines was extracted using a phenol-chloroform extraction protocol. Purity and concentration were determined by the NanoDrop 1000 Spectrophotometer (Thermo Scientific). RNA (1 ug) was reverse transcribed with iScript cDNA Synthesis Kit (Bio-Rad).

TET expression levels were analyzed by RT-qPCR using the 7500 Real Time PCR Instrument (Applied Biosystems). TET expression levels were normalized to TBP and/or β-actin expression. The PCR assays comprised of 10 ul of iTaq Universal SYBR Green Supermix (Bio-Rad), 50 ng of methylated or input DNA, and 0.25 uM of each primer with a total volume of 20 ul. All PCR assays included a non-template control, using UltraPure Distilled Water (Invitrogen) as the template. The PCR conditions were: 10 min at 95°C, 40 cycles of denaturation for 15 sec at 95°C, and annealing for 1 min at 60°C. The same procedure was followed to analyze cDNA of human prostate cancer cell lines 22Rv1, LNCaP, DU-145, and PC3 (provided by KK and EOM).
The sequences of primers amplifying *TET1* are 5'-TCT GTT GTT GTG CCT CTG GA-3' (forward), 5'-GCC TTT AAA ACT TTG GGC TTC-3' (reverse); *TET2* are 5'-AAA GAT GAA GGT CCT TTT TAT ACC C-3' (forward), 5'-TTT ACC CTG TCC AAA CCT T-3' (reverse); *TET3* are 5'-CCA TTG CAA AGT GGG TGA-3' (forward), 5'-CGC ACC AGG CAG AGT AGC-3' (reverse); *TBP* are 5'-TGT ATC CAC AGT GAA TCT TGG TTG-3' (forward), 5'-GTT TCG TGG CTC TCT TAT CCT C-3' (reverse); and *β-actin* are 5'-ATC ATG TTT GAG ACC TTC AA-3' (forward), 5'-CAT CTC TTG CTC GAA GTC CA-3' (reverse).

### S1.1.3 Protein Extraction and Western Blot analysis

Total protein lysate of RWPE-1 and *TET2* KD RWPE-1 cells was extracted with one tablet of cOmplete Mini, EDTA-free (Roche) diluted in 100 ul of RIPA buffer (Invitrogen), centrifuged at 10,000 g for 10 min at 4°C, and supernatant was collected. The expression levels of proteins were examined by sodium dodecyl sulphate/ polyacrylamide gel electrophoresis, followed by western blot analysis with anti-*TET2* antibody (R1086-vp, Abiocode) and anti-*β*-actin antibody (Abcam).
Supplemental Figure 1. Enrichment map of genes significantly hypermethylated in RWPE-1 and 22Rv1 cells, identified by MBD-seq. Enrichment map (EM) of genes significantly and differentially methylated in all gene features following MBD-seq of RWPE-1 and 22Rv1 cells, generated by GSEA database and Cytoscape program (enrichment p-val<0.005, FDR<0.1, similarity cutoff>0.5). Nodes (blue and red circles) represent a subset of enriched genes in a similar functional pathway or gene set. Light blue circles were manually drawn and annotated to highlight common biological processes.
Supplemental Figure 2. Diagnostic enrichment plot of hMeSeal-seq samples from RWPE-1 and 22Rv1. Enrichment diagnostic graphs comparing the curve of input (non-enriched) sample (green, blue lines) to biological replicate 2 (black, red lines) in RWPE-1 and 22Rv1 cells, respectively.
Supplemental Figure 3. TET messenger RNA expression profiles across prostate cell lines. Expression profiles in benign (RWPE-1) and cancerous (22Rv1, LNCaP, DU-145, PC3) cells, where (a) TET1, (b) TET2, and (c) TET3 mRNA levels were quantified using RT-qPCR, $2^{-\Delta\Delta Ct}$ method normalized to β-actin cDNA. Error bars indicate standard deviation.
Supplemental Figure 4. Messenger RNA and protein profiles of stable TET2 KD RWPE-1 cells. (a) Stable TET2 KD RWPE-1 cells using lipofection of shRNA analyzed by RT-qPCR, 2−ΔΔCt method normalized to TBP cDNA. KD indicates knockdown and error bars indicate standard deviation. P-values measured by Student’s t-test, *p-value<0.05, **p-value<0.01. (b) Western blot analysis of TET2 KD RWPE-1 lysates, using primary anti-TET2 antibody, where anti-β-actin antibody was used as loading control.
References


190. Nair, S.S., et al., *Comparison of methyl-DNA immunoprecipitation (MeDIP) and methyl-CpG binding domain (MBD) protein capture for genome-wide DNA methylation analysis reveal CpG sequence coverage bias.* Epigenetics, 2011. 6(1): p. 34-44.


