Dynamic regulation of DNA methylation in human T-cell biology

Antonio Lentini
Dynamic regulation of DNA methylation
in human T-cell biology

Antonio Lentini

Division of Children’s and Women’s Health
Department of Clinical and Experimental Medicine
Faculty of Medicine and Health Sciences
Linköping University, Linköping, Sweden
Linköping 2019
Supervisor
Docent Colm E. Nestor
Department of Clinical and Experimental Medicine (IKE), Linköping University, Sweden

Co-supervisors
Professor Mikael Benson
Department of Clinical and Experimental Medicine (IKE), Linköping University, Sweden
Principal Research Engineer Huan Zhang
Department of Clinical and Experimental Medicine (IKE), Linköping University, Sweden
Senior Lecturer Mika Gustafsson
Department of Physics, Chemistry and Biology (IFM), Linköping University, Sweden

Faculty opponent
Senior Lecturer Ian R. Adams
MRC Institute of Genetics & Molecular Medicine, University of Edinburgh, UK

Examination board
Associate Professor Joëlle Rüegg
Institute for Environmental Medicine (IMM), Karolinska Institute, Stockholm,
Associate Professor Carlos Guerrero Bosagna
Department of Physics, Chemistry and Biology (IFM), Linköping University, Sweden
Professor Marie Larsson
Department of Clinical and Experimental Medicine (IKE), Linköping University, Sweden
Table of contents

Dynamisk reglering av DNA-metylering vid T-cellsdifferentiering ......................................... 1
Abstract ...................................................................................................................................... 3
List of included papers ............................................................................................................... 5
List of other papers not included ............................................................................................... 6
Frequently used abbreviations ................................................................................................... 7
Introduction ................................................................................................................................ 9
DNA methylation ................................................................................................................... 9
  DNA demethylation ................................................................................................................ 10
  DNA methylation and chromatin ........................................................................................... 12
  DNA methylation in development ....................................................................................... 13
  DNA methylation in disease ............................................................................................... 13
T-cell biology ....................................................................................................................... 14
  T-cell development ........................................................................................................... 14
  T-cell differentiation ......................................................................................................... 15
  T-cell memory formation ................................................................................................. 17
Epigenetic regulation of T-cell differentiation ................................................................ 18
Measuring DNA methylation in mammals ................................................................................. 18
  Base-resolution techniques .............................................................................................. 18
  Enrichment-based techniques .......................................................................................... 20
  Bioinformatics analysis of genome-wide DNA methylation data ................................... 21
Materials and methods ............................................................................................................. 23
  Ethics statement ................................................................................................................ 23
  Material from healthy individuals (Paper I) ..................................................................... 23
  Material from murine cell lines (Paper II) ....................................................................... 23
  Cell isolation ........................................................................................................................ 23
    Isolation of peripheral blood mononuclear cells (PBMCs) (Paper I) .............................. 23
    Magnetic-activated cell sorting (MACS) (Paper I) ....................................................... 23
    Fluorescence-activated cell sorting (FACS) (Paper I) ................................................... 23
  Cell lines .............................................................................................................................. 24
    Jurkat (Paper I) ............................................................................................................ 24
    Mouse embryonic stem cells (Paper II) ...................................................................... 24
  Cell culture manipulation ..................................................................................................... 24
    In vitro CD4+ T-cell polarisation (Paper I) ................................................................. 24
    Cell cycle and division detection (Paper I) .................................................................. 24
    Gene expression analysis ............................................................................................... 24
    Microarray (Paper I) ..................................................................................................... 24
    Quantitative PCR (qPCR) (Paper I) ............................................................................. 25
  Global analysis of DNA methylation ................................................................................... 25
    Immuno-dot blot (Paper I & II) .................................................................................. 25
    Enzyme-linked immunosorbent assay (ELISA) (Paper II) .......................................... 25
    Mass spectrometry (Paper II) ...................................................................................... 25
  Genome-wide analysis of DNA methylation ....................................................................... 26
Dynamisk reglering av DNA-metylering vid T-cellsdifferentiering


DNA-metylering kan fästas direkt på DNA för att kontrollera hur gener uttrycks men upprätthålls även under längre tid vilket skapar ett epigenetiskt minne av vilka gener som kan uttryckas i varje celltyp. Etableringen av DNA-metyleringsmönster är en viktig del av differentieringen av T-celler men hur dessa mönster regleras över tid är fortfarande okänt. Metylering fästs oftast direkt på cytosinbaser i DNA för att bilda 5-metylcytosin (5mC). Borttagning av DNA-metylering kräver däremot flera enzymatiska reaktioner som oftast startas via omvandling från 5mC till 5-hydroximetylcytosin (5hmC). Tillsammans skapar detta ett mycket komplext system av reglering då olika DNA-modifikationer ofta är kopplade till olika funktioner. Till exempel är 5mC vanligtvis associerad med lågt uttryckta gener men 5hmC ofta markerar högt uttryckta gener. Denna avhandling syftar till att undersöka hur DNA-metylering regleras dynamiskt vid T-cellsdifferentiering.

För detta ändamål användes storskaliga profileringstekniker och kombinerade genuttryck med mätningar av 5mC och 5hmC över hela genomet. Denna data användes för att konstruera en modell av hur epigenetisk reglering sker över tid under differentieringsprocessen. Detta visade att tidig T-cellsaktivering åtföljdes av omfattande berikning av 5hmC vilket senare resulterade i de-metylering vid celldelning och differentiering. Globala förändringar av DNA-metylering via 5mC var inte bara ett tecken på var de-metylering skulle ske under T-cellsdifferentiering utan återfanns och kvarstod även i långlivade minnes-T-celler. Resultaten tyder på att grundandet av epigenetiska landskap i T-celler initieras tidigt under aktiveringsfasen, innan celler har etablerat ett stabilt cell-öde. Landskapen kan sedan bibehållas under cellens livstid som ett epigenetiskt minne. Regionerna som genomgick omvandling av DNA-metylering var även berikade för genetiska varianter kopplade till autoimmuna sjukdomar. Många av dessa varianter kunde direkt påverka proteinbindning vilket potentiellt skulle kunna påverka de genregulatoriska nätverken. Detta illustrerar det komplexa samspelet mellan genetiska och epigenetiska funktioner samt hur dessa mönster kan påverkas vid sjukdom.

Sammantaget visar denna avhandling hur integrerad användning av högupplöst epigenomisk data kan användas för att studera komplexa biologiska system över tid samt hur tekniker för att mäta epigenetiska signaler systematiskt kan karakteriseras för att identifiera och korrigera metodfel för att förbättra genomiska mättekniker.
Abstract

T helper cells play a central role in orchestrating immune responses in humans. Upon encountering a foreign antigen, T helper cells are activated followed by a differentiation process where the cells are specialised to help combating the infection. Dysregulation of T helper cell activation, differentiation and function has been implicated in numerous diseases, including autoimmunity and cancer. Whereas gene-regulatory networks help drive T-cell differentiation, acquisition of stable cell states require heritable epigenetic signals, such as DNA methylation. Indeed, the establishment of DNA methylation patterns is a key part of appropriate T-cell differentiation but how this is regulated over time remains unknown. Methylation can be directly attached to cytosine residues in DNA to form 5-methylcytosine (5mC) but the removal of DNA methylation requires multiple enzymatic reactions, commonly initiated by the conversion into 5-hydroxymethylcytosine (5hmC), thus creating a highly complex regulatory system. This thesis aimed to investigate how DNA methylation is dynamically regulated during T-cell differentiation.

To this end, we employed large-scale profiling techniques combining gene expression as well as genome-wide 5mC and 5hmC measurements to construct a time-series model of epigenetic regulation of differentiation. This revealed that early T-cell activation was accompanied by extensive genome-wide deposition of 5hmC which resulted in demethylation upon proliferation. Early DNA methylation remodelling through 5hmC was not only indicative of demethylation events during T-cell differentiation but also marked changes persisting long-term in memory T-cell subsets. These results suggest that priming of epigenetic landscapes in T-cells is initiated during early activation events, preceding any establishment of a stable lineage, which are then maintained throughout the cells lifespan. The regions undergoing remodelling were also highly enriched for genetic variants in autoimmune diseases which we show to be functional through disruption of protein binding. These variants could potentially disrupt gene-regulatory networks and the establishment of epigenetic priming, highlighting the complex interplay between genetic and epigenetic layers. In the course of this work, we discovered that a commonly used technique to study genome-wide DNA modifications, DNA immunoprecipitation (DIP)-seq, had a false discovery rate between 50-99% depending on the modification and cell type being assayed. This represented inherent technical errors related to the use of antibodies resulting in off-target binding of repetitive sequences lacking any DNA modifications. These sequences are common in mammalian genomes making robust detection of rare DNA modifications very difficult due to the high background signals. However, off-target binding could easily be controlled for using a non-specific antibody control which greatly improved data quality and biological insight of the data. Although future studies are advised to use alternative methods where available, error correction is an acceptable alternative which will help fuel new discoveries through the removal of extensive background signals.

Taken together, this thesis shows how integrative use of high-resolution epigenomic data can be used to study complex biological systems over time as well as how these techniques can be systematically characterised to identify and correct errors resulting in improved detection.
List of included papers

**Paper I:**


*Joint first author

**Paper II:**


**Erratum:**

In **Paper I**, author contributions of ANTONIO LENTINI was incorrectly credited at the time of printing this thesis. Correct contributions are the following: A.L. performed experiments, data analysis and wrote the paper.
List of other papers not included

Nikolaos Doumpas, Franziska Lampart, Mark D Robinson, ANTONIO LENTINI, Colm E Nestor, Claudio Cantù, Konrad Basler. TCF/LEF dependent and independent transcriptional regulation of Wnt/β-catenin target genes. 2018 The EMBO Journal, e98873

Richard R Meehan, John P Thomson, ANTONIO LENTINI, Colm E Nestor, Sari Pennings. DNA methylation as a genomic marker of exposure to chemical and environmental agents. 2018 Current opinion in chemical biology 45, 48-56

Lina Mattson, ANTONIO LENTINI, Danuta R Gawel, Tejaswi VS Badam, Mikael Benson, Torbjorn Ledin, Colm E Nestor, Mika Gustafsson, Jordi Serra-Musach, Janne Bjorkander, Zou Xiang, Huan Zhang. Potential Involvement of Type I Interferon Signaling in Immunotherapy in Seasonal Allergic Rhinitis. 2016 Journal of immunology research


Colm E Nestor, Fredrik Barrenäs, Hui Wang, ANTONIO LENTINI, Huan Zhang, Sören Bruhn, Rebecka Jörnsten, Michael A Langston, Gary Rogers, Mika Gustafsson, Mikael Benson. DNA methylation changes separate allergic patients from healthy controls and may reflect altered CD4+ T-cell population structure. 2014 PLoS genetics 10(1), e1004059

Huan Zhang, Colm E Nestor, Shuli Zhao, ANTONIO LENTINI, Barbara Bohle, Mikael Benson, Hui Wang. Profiling of human CD4+ T-cell subsets identifies the TH2-specific noncoding RNA GATA3-AS1. 2013, Journal of Allergy and Clinical Immunology 132(4), 1005-1008
### Frequently used abbreviations

In alphabetical order:

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<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>5mC</td>
<td>5-methylcytosine</td>
</tr>
<tr>
<td>5modC</td>
<td>5mC/5hmC/5fC/5caC</td>
</tr>
<tr>
<td>ChIP</td>
<td>Chromatin immunoprecipitation</td>
</tr>
<tr>
<td>CMS</td>
<td>Cytosine-5-methylenesulfonate</td>
</tr>
<tr>
<td>DIP</td>
<td>DNA immunoprecipitation</td>
</tr>
<tr>
<td>DNMT</td>
<td>DNA methyltransferase</td>
</tr>
<tr>
<td>FDR</td>
<td>False discovery rate</td>
</tr>
<tr>
<td>HSC</td>
<td>Haematopoietic stem cell</td>
</tr>
<tr>
<td>LD</td>
<td>Linkage disequilibrium</td>
</tr>
<tr>
<td>MBD</td>
<td>Methyl binding domain</td>
</tr>
<tr>
<td>mESC</td>
<td>Mouse embryonic stem cell</td>
</tr>
<tr>
<td>NT</td>
<td>Naive T-cell</td>
</tr>
<tr>
<td>PBMC</td>
<td>Peripheral blood mononuclear cell</td>
</tr>
<tr>
<td>Seal</td>
<td>Selective chemical labelling</td>
</tr>
<tr>
<td>SNV</td>
<td>Single nucleotide variant</td>
</tr>
<tr>
<td>STR</td>
<td>Short tandem repeat</td>
</tr>
<tr>
<td>TCM</td>
<td>Central memory T-cell</td>
</tr>
<tr>
<td>TCR</td>
<td>T-cell receptor</td>
</tr>
<tr>
<td>TDG</td>
<td>Thymine DNA Glycosylase</td>
</tr>
<tr>
<td>T_EM</td>
<td>Effector memory T-cell</td>
</tr>
<tr>
<td>TF</td>
<td>Transcription factor</td>
</tr>
<tr>
<td>T_H</td>
<td>T helper cell</td>
</tr>
<tr>
<td>TKO</td>
<td>Triple knockout</td>
</tr>
<tr>
<td>WGBS</td>
<td>Whole-genome bisulfite sequencing</td>
</tr>
</tbody>
</table>
Introduction

The genome encodes the functional programs of every cell in the human body. Thus, flexibility in how the genome sequence is interpreted is required to give rise to the multitude of different cell-types and transcriptional programs in humans. Epigenetics (from Greek ‘epi-’ meaning “on, upon, above”) facilitates this flexibility by chemically modifying DNA and/or altering its structure and organisation to affect changes in gene expression without changing the underlying DNA sequence. Since epigenetic mechanisms are tightly coupled with cellular development, it is not surprising that disruption of epigenetic pathways is commonly observed in differentiation disorders such as cancer. Whereas the exact definition of ‘epigenetics’ remains diffuse and highly debated, this thesis focuses on epigenetic regulation of gene expression by covalent modification of DNA.

DNA methylation

Shortly after DNA was identified as the genetic material in 1944, 5-methylcytosine (5mC) was discovered as a minor constituent of mammalian DNA [1]. DNA methyltransferases (DNMTs) can covalently attach a methyl-group to Cytosine, preferentially in the context of a CpG dinucleotide. As 5mC is highly mutagenic, CpGs are relatively underrepresented in mammalian genomes. Moreover, CpG distribution is highly non-random, typified by large regions of very low CpG density punctuated by regions of very high CpG density. Unlike the majority of CpGs, cytosines in these high density islands are typically not methylated and situated in the promoters of genes, predominantly highly expressed house-keeping genes and developmentally important genes [2, 3]. In contrast, tissue-specific genes typically have low CpG density promoters that are heavily methylated in non-target tissues [4]. DNA methylation plays an important role in development through its central role in processes including X-inactivation, genomic imprinting and silencing of transposable elements, where loss of DNMTs leads to developmental defects and genomic instability in mammals [5]. Deposition of 5mC is typically associated with transcriptional silencing mediated by two known mechanisms [6]; firstly, 5mC can block DNA binding factors and DNA polymerase elongation via steric hindrance, and secondly, 5mC binding proteins can recruit co-repressor molecules which silence transcription. More recently, functional proteomics screens have suggested that several homeodomain-containing transcription factors (TFs) may also bind methylated DNA [7, 8], but these findings require further replication and validation before their biological significance can start to be dissected.

The de novo methyltransferases DNMT3A and DNMT3B typically attach methyl-groups to previously unmodified CpGs whereas the maintenance methyltransferase, DNMT1, acts on CpG sites methylated on a single strand of a double-stranded DNA molecule (hemi-methylated DNA), attaching a methyl group to the newly synthesised strand during DNA replication [6] (Figure 1A). This system of establishment and maintenance of 5mC patterns allows cells to create cell states that are faithfully propagated to the daughter cells [4]. Whereas DNA methylation helps maintain stable cell identities, it is not the main driver of lineage-commitment [9]. While DNMT1 contains RFTS and CXXC domains which target replication
foci and unmethylated DNA, respectively, DNMT3A/B lack strong DNA binding domains but have PWWP and ADD domains facilitating chromatin interactions [10]. To achieve highly specific methylation patterns, DNMTs must therefore be recruited by other proteins to mediate their actions [10]. For instance, UHRF1 is a well-known interaction partner of DNMT1 that contains an SRA domain with a strong preference for hemi-methylated DNA directing DNMT1 to its substrate [10]. DNMT3A/B complexes on the other hand, associate non-specifically with histone H3 molecules unmethylated at lysine 4 (H3K4). Certain regulatory elements, such as enhancers and high CpG density promoters, are bound by TFs blocking DNA access and/or facilitating H3K4 methylation, therefore shielding these regions from DNMT3A/B binding [10]. Furthermore, DNMT3A/B can be recruited by- and form complexes with TFs which provide sequence specificity [11]. A notable property of certain developmentally important TFs is the preference to bind methylated DNA motifs [11]. These TFs are often key lineage-defining factors that recruit other protein complexes which help establish a permissive chromatin landscape through opening of heterochromatin and removal of DNA methylation to drive gene expression [11].

**Figure 1. The regulation of DNA methylation patterns in mammals.** (A) DNA methylation patterns are established and maintained by DNA methyltransferase (DNMT) enzymes. (B) Ten-Eleven Translocation (TET) family of enzymes mediate the sequential conversion of 5-methylcytosine (5mC) to 5-hydroxymethylcytosine (5hmC), 5-formylcytosine (5fC) and 5-carboxylcytosine (5caC) which may result in DNA demethylation.

**DNA demethylation**

While several different DNA demethylation pathways have been hypothesised in mammals, identifying a mammalian DNA demethylase has proven controversial [12]. Certain species, such as Trypanosoma and Leishmania, contain base J (β-D-glucosyl hydroxymethyluracil) which is produced through hydroxylation by JBP proteins followed by glycosylation of the exocyclic methyl group of thymine [13]. In 2009, computational prediction of human orthologues of JBP identified the TET family of enzymes as being capable of hydroxylation of 5mC into 5-hydroxymethylcytosine (5hmC) [14, 15]. Simultaneously, 5hmC was found to be present at high levels in human neurons, suggesting a potential regulatory function [16]. TET enzymes were also found to further catalyse stepwise oxidation of 5hmC into 5-formylcytosine (5fC) and 5-carboxylcytosine (5caC) [17], both of which could be efficiently excised by Thymine DNA Glycosylase (TDG) and DNA repair pathways to yield unmodified cytosine [18, 19] (Figure 1B). In contrast to 5mC, there is no known maintenance pathway for 5hmC.
and its derivatives which are gradually lost over cell divisions [20]. This ‘passive’ loss of modifications through dilution is likely the major DNA demethylation pathway throughout development as lineage specification is typically accompanied by cellular replication [20].

![Figure 2. DNA methylation and chromatin cross-talk.](image)

(Top) DNA is wrapped around histone octamers to form a nucleosome. Together with histone 1 (not shown), the nucleosome can be coiled into chromatin fibres. The N-terminal tails of the histones can carry a wide range of post-translational modifications which have been associated with specific transcriptional states. (Bottom) Multiple layers of epigenetic information form distinct genomic states in a combinatorial manner with permissive, repressive or bivalent contexts. K, lysine; R, arginine; Ac, acetylation; hMe, hydroxymethylation; Me, methylation; Ub, ubiquitination.

Although both TET1/3 contain an N-terminal CXXC domain, TET2 completely lacks DNA binding capacity, likely originating from a chromosomal inversion event [14]. Unlike other CXXC domains, TET1/3 recognise both unmethylated and methylated cytosines allowing a
greater flexibility in sequence selectivity [21]. The oxidised derivatives 5hmC/5fC/5caC are typically found at lineage-specific enhancers and active- or poised promoters of low to intermediate CpG density [21]. Similar to 5mC binders, certain proteins have been observed to have specific non-overlapping preference for each of the different oxidised derivatives of 5mC [22, 23]. Notably, 5fC and 5caC have more binding partners than 5mC or 5hmC, including TFs, DNA repair proteins and chromatin regulators [22, 23]. However, as 5fC and 5caC are slowly generated [17] and rapidly excised [18, 19], only very low levels are present in mammalian genomes (~65–4,000 fold lower than 5hmC [17]) and their biological relevance as epigenetic modifications remains largely unknown.

DNA methylation and chromatin

While simple organisms such as bacteria can use DNA methylation as their main mode of epigenetic signalling [24], large mammalian genomes require multiple layers of regulation to allow for complex gene expression profiles in a wide variety of cell types [25]. Human DNA is wrapped around histone proteins, forming a nucleosome, which can further be coiled into chromatin, allowing regulation not only on the levels of DNA but also through higher-order structures and post-translational modifications of histones [26]. Although each epigenetic layer contains certain information, these layers are typically interlinked, creating a complex mode of regulation (Figure 2). Open euchromatin is generally marked by histone acetylation and H3K4 methylation whereas silent heterochromatin typically contains hypo-acetylated H3/H4, H3K27me3 and H3K9me3 [27]. While acetylation of histone tails alters the relative net charge against DNA, relaxing chromatin compaction and enabling transcription [28], histone methylation can be associated with either activation or repression depending on position and state [29]. Genome-wide mapping of histone modifications has enabled high-resolution determination of chromatin architecture where functional roles of histone modifications are typically inferred from correlation with transcription [30, 31]. Although chromatin landscapes and DNA methylation provide seemingly overlapping information, a large portion of genes are directly regulated by 5mC in the absence of histone modifications [32]. Conversely, histone modifications can provide a more nuanced landscape at actively transcribed genes [33].

Heterochromatin and 5mC are tightly linked and DNMTs have been shown to directly interact with H3K9 methyltransferases to mediate silent chromatin [34]. However, as genome-wide 5hmC profiles were generated, it became clear that it provided a landscape largely distinct from 5mC as it was found in active regions of euchromatin, such as in enhancers and highly transcribed genes [35]. Although active promoters are marked by H3K4 methylation and lack cytosine modifications (hereafter referred to as ‘5modC’) at the transcription start site (TSS), they are commonly bound by TET1 indicating actions independent of catalytic activity [36]. Indeed, TET1 has been found to mediate dual roles in regulation by recruiting Polycomb repressive complexes to silence gene expression at a subset of genes [36, 37]. Furthermore, both TET1 and TET2 have been shown to interact with DNMT3A in both competitive and cooperative manners to dynamically regulate gene expression through modulation of both methylation- and chromatin landscapes [38, 39]. Another unexpected interaction is the recruitment of OGT, which mediates histone O-GlcNAcylation, and SET1/COMPASS
mediating H3K4 methylation by TET proteins to active promoters, positively regulating gene expression [40]. While 5mC is typically added late during pericentric heterochromatin formation [41], accumulating evidence suggests that 5hmC remodelling precedes chromatin opening [42] (unpublished observations) although histone acetylation and transcription likely occurs before complete conversion back to unmodified cytosine [43]. Even though the exact temporal relationships between regulatory layers remains largely unexplored, the last decade has revealed the vast complexity and cross-talk within the epigenetic machinery.

**DNA methylation in development**

During embryonic development, mammalian genomes undergo global erasure of 5mC which is then re-established by de novo methylation as cells differentiate into more restricted lineages [44]. The global loss of 5mC in the paternal genome during the first cell replication coincides with a rapid conversion to 5hmC mediated by TET3 [45-47] where deletion of Tet3 leads to reduced foetal survival [46]. Notably, whereas individual or combined deletion of Tet1/2 leads to developmental defects [48], impaired embryogenesis is only observed in Tet1/2/3 triple knockouts (TKOs) [49] indicating important and non-overlapping roles of TETs during development. A second wave of global 5mC erasure occurs in primordial germ cells (PGCs) to provide a largely blank slate for germ cell development [50]. Although the exact mechanisms of how 5mC is re-established remains unclear, the importance of DNA methylation during early development is highlighted by the embryonic lethal phenotype of DNMT1 loss [51]. However, not all genomic regions are controlled equally and several imprinted loci escape demethylation [52]. A seemingly paradoxical finding is that genome-wide 5mC erasure is the most extreme at LINE1 elements [53] as activation of retrotransposons by 5mC loss leads to genomic instability and cell death in the male germ cells [54]. While this window of demethylation enables LINE1 mobilisation, it also activates ‘defence’ genes that are solely repressed by DNA methylation [55]. This includes the PIWI pathway protein Mili and Tex19.1 which can post-transcriptionally target LINE1 elements for silencing [56, 57].

**DNA methylation in disease**

Only a handful diseases are known to be caused by defects in DNA methylation, notable examples being RETT-, ICF- and fragile-X syndromes [58]. However, global 5mC changes are commonly observed in autoimmune diseases [59] and all types of cancer [60] representing a promising biomarker for disease detection [61]. Indeed, DNMTs and TETs are commonly mutated in cancer, most frequently in haematopoietic malignancies [10, 62]. Although cancer cells exhibit a general phenotype with global loss of methylation combined with promoter hypermethylation [63], the latter typically occurs at genes already silent in the cell of origin [64]. While cancer epigenomes have been intensively studied for over 10 years, the function of DNA methylation changes in cancer still remains largely unknown [60]. DNMT3A and TET2 in particular are frequent hotspots for mutations in aging healthy individuals and are highly associated with clonal haematopoiesis [65]. Notably, up to 80% of peripheral T-cell lymphomas of CD4+ lineages harbour TET2 mutations indicating its importance in T-cell development [66].
T-cell biology

The human immune system can be broadly divided into innate- and adaptive immunity. The function of the innate immune system is to produce a rapid general response as it is largely encoded within the functions of the mediating cells. Adaptive immunity, on the other hand, produces highly specialised responses depending on encountered antigens. Lymphocytes make up the majority of the adaptive immune system in humans. These cells are broadly divided into B- and T-cells based on their organ of origin, Bone marrow (historically Bursa of Fabricius in birds) or Thymus, respectively. While B-cells are critical for humoral immunity through production of antibodies, T-cells can be further subdivided into CD8⁺ or CD4⁺ cells based on their surface molecules where the former plays a central role in cell-mediated immunity by directly killing infected cells. Meanwhile, CD4⁺ T-cells, commonly referred to as ‘T helper cells’, mediate their action through orchestrating immune responses through production of signalling molecules, cytokines. The produced cytokines can help maturation and class-switching of B-cells as well as assist activation of CD8⁺ T-cells and other immune cells.

Figure 3. Thymocyte development. Thymus-seeding progenitors (TSPs) enter the thymus from the bone marrow. The immature early thymic progenitors (ETP) express CD34 and have lymphomyeloid potential. Restriction of these progenitors to the T-cell lineage is marked by expression of CD1a in pre-T-cells. These early, double negative cells do not express CD4 or CD8, nor a T-cell receptor. As T-cell maturation progress, expression of CD34 is lost and cells gain double positive CD4 and CD8 expression (DP). Surface CD3 expression coincides with the expression of a functional T-cell receptor. Finally, those cells surviving positive and negative selection lose expression of either CD4 or CD8 as mature single positive T-cells (SP).

T-cell development

The thymus is a unique organ in that it provides a training ground specifically required for T-cells. After haematopoietic stem cell (HSC)-derived lymphocyte progenitors enter the thymus, they undergo a series of differentiation steps typically defined by the surface expression of CD4 and CD8 (Figure 3). Once the CD4⁺CD8⁻ double-negative (DN) CD34⁺ early thymus progenitors (ETP) commit to the T-cell lineage, the differentiation capacity of non-T-cells is lost and CD1a expression is gained [67]. From this step onwards, IL-7 signalling
is an important driver of human T-cell development where blockage of IL-7R largely prevents further differentiation of committed CD34+CD1a+ thymocytes [68]. As the cell travels along the thymic cortex, it starts rearranging the T-cell receptor (TCR) in a stepwise-manner (TCRδ > γ > β > α) and cells that pass signalling through the pre-TCR, β-selection, will initiate expression of CD4 and CD8, becoming double-positive (DP) thymocytes. When the DP cells complete TCRα rearrangement, the TCRαβ complex becomes expressed on the surface and they down-regulate IL-7Rα, becoming unresponsive to IL-7 signalling. The DP thymocytes are then positively selected for a productive interaction between the TCRαβ and self MHC complexes (MHC class I or II; known as HLA –A, –B and –C, or –DR, –DQ and –DP in humans, respectively) expressed on the thymic epithelium followed by negative selection of self-reactive cells. Finally, the cells will mature into CD4 or CD8 single-positive (SP) cells and lose expression of the opposite co-receptor based on their MHC specificity, MHC-II-restricted thymocytes will develop into the CD4+ lineage and MHC-I-restricted thymocytes CD8+ lineage [69].

**T-cell differentiation**

After development in the thymus, T-cells enter the blood circulation in a dormant ‘naïve’ state until they encounter their cognate pathogenic signal. Cells in vertebrate species constantly present endogenous peptides through MHC-I complexes on the cell surface which are surveyed by CD8+ T-cells, potentially triggering elimination. Certain cell types however, so-called antigen-presenting cells (APCs), such as dendritic cells, express MHC-II molecules and are able to present and process exogenous peptides to mediate CD4+ T-cell responses [70].

The TCR complex is made up of the TCRαβ chains, responsible for antigen recognition, as well as CD3 which is responsible for signal transduction [71]. Interestingly, there is an obligate requirement for co-expression of both molecules in human T-cells as CD3 mutants also lose TCR surface expression [72]. Early studies revealed that stimulation with anti-CD3 could activate T-cells to produce IL-2 but this signalling alone resulted in a non-responsive state (anergy) without co-stimulation, for instance via CD28. Together, CD3 and CD28 stimulation promotes robust T-cell proliferation, survival, metabolism and cytokine production [71]. Within 48 hours of activation, T-cells double in size as they enter the cell cycle in preparation for cell division. These morphological changes are coupled to major chromatin remodelling events leading to waves of gene activation [73].

T-cell activation upon encountering a foreign antigen is followed by differentiation into specialised effector subsets which help combat the foreign pathogen. While CD4+ T helper (TH) subsets are increasingly recognised as highly plastic, three main lineages of effector subsets are commonly defined based on their signature cytokine production; IFN-γ for TH1, IL-4 for TH2 and IL-17 for TH17. Each lineage is driven by subset-specific, self-reinforcing cytokine responses and TFs which help establish a stable lineage cell fate while repressing transcriptional networks of other lineages (*Figure 4*) [74].
Figure 4. T-helper cell differentiation. Upon being presented with an antigen, Naïve T-cells (NT) differentiate and engage specific transcriptional programs depending on the cytokine microenvironment. Multiple functional subsets have been described based on cytokine production driven by lineage-specific transcription factors. Cytokine-producing effector cells help direct immune responses and clearance of pathogens but have also been implicated in immune-related diseases, such as autoimmunity and allergy.

T_h1 differentiation is initiated mainly in the presence of IL-12 but also IFN-γ which activates STAT1/4 causing induction of TBX21 (also known as T-bet) leading to production of IFN-γ which helps combat parasites and intracellular pathogens [75]. While IFN-γ drives early TBX21 expression through STAT1, IL-12 signalling through STAT4 maintains TBX21 expression levels after termination of TCR signalling [76]. Although STAT4 can directly bind and promote local epigenetic changes at Ifng [77], inputs from both STAT4 and TBX21 are required for stable Ifng expression [78]. TBX21 further promotes T_h1 cell expansion through upregulation of IL-12Rβ2, reinforcing responses to IL-12 [75]. Furthermore, TBX21 plays a key role in balancing lineage fates as overexpression in T_h2 cells leads to production of IFN-γ and suppression IL-4 while mice lacking TBX21 show severe defects in T_h1 differentiation with subsequent skew towards T_h2 cytokine production [75]. However, IFN-γ may be redundantly regulated by both TBX21 and T-box family member EOMES (Eomesodermin) as CD4+ T-cells lacking the former still produce IFN-γ whereas production is abolished in double knockouts, suggesting that the main role of TBX21 may be to suppress the T_h2 master regulator GATA3 [75].

Differentiation of T_h2 is initiated in the presence of IL-2 and IL-4 which activates STAT5/6 and upregulates GATA3 expression leading to production of IL-4, -5 and -13 facilitating B-cell class switching to IgE and recruitment of eosinophils and basophils to help defence against helminths [79]. T_h2 responses also help protect against anaphylaxis from toxins such as snake-
and bee venoms [80]. Although low levels of STAT5 are enough to promote proliferation and survival, strong STAT5 signals are required for T\(_h2\) differentiation where it may maintain accessibility of the \(Il4\) locus [75]. While constitutively activated STAT5 can induce IL-4 production without regulating GATA3, this induction fails in the absence of GATA3 [74]. To this end, STAT6 activation through IL-4 signalling is both necessary and sufficient to drive GATA3 expression [74]. While GATA3 is critical for T-cell development, conditional deletion of Gata3 in peripheral CD4\(^+\) T-cells only impairs T\(_h2\) differentiation [75]. Not only does GATA3 collaborate with STAT5 to induce IL-4 production, it is alone indispensable for IL-5 and IL-13 production in T\(_h2\) cells [81]. Furthermore, GATA3 also selectively stimulates commitment and proliferation of T\(_h2\) cells [81] while suppressing T\(_h1\) responses through downregulation of Stat4 [82].

While T-cell differentiation has been extensively studied in vitro, this model relies on forcing the cells towards specific lineages in addition to inhibiting other fates. It is now generally acknowledged that T-cell subsets display a high degree of functional plasticity [83]. Furthermore, single human T-cell clones give rise to a multitude of different subsets and lineage fates regardless of pathogenic stimuli and polarised responses are likely driven by selective expansion rather than specific priming [84]. This allows for a highly flexible immune system where multiple functional subsets are generated but only the appropriate effector cells are preferentially expanded, depending on the pathogen.

**T-cell memory formation**

After the primary immune response is resolved, the majority of effector cells undergo apoptosis, with the exception of a small surviving population of antigen-specific memory cells that mediate a rapid response upon secondary stimuli [85]. Although the signals underlying activation and polarisation of CD4\(^+\) T-cells are well-studied, the transition from effector to memory cell remains poorly understood [85]. In contrast to the tightly regulated activation process, few signals are needed for transition into memory. In fact, the most well-known starting point for memory transition is the removal of antigen and inflammatory cytokines, suggesting that memory responses are primed during early stages of activation or differentiation [85]. Indeed, optimal memory generation can be affected by key activation signals such as IL-2, CD28 and TCR signalling [85].

Compared to naïve T-cells (NT), the memory compartment produces rapid and robust recall responses [85]. The memory T-cell pool is likely highly complex but a few broadly defined functional subsets have been described [86]. Similarly to NT, central memory T-cells (T\(_{CM}\)) produce mainly IL-2 upon stimulation but are more readily activated and rapidly differentiate into effector cells [87]. T\(_{CM}\) also express CCR7, allowing homing to secondary lymphoid organs [87]. In contrast, effector memory T-cells (T\(_{EM}\)) retain the capacity to produce effector-associated cytokines and produce immediate responses when re-stimulated [87]. T\(_{EM}\) may more closely resemble effector T\(_h\) cells by expression of subset-specific chemokine receptors that facilitate tissue homing [87]. Whereas both T\(_{CM}\) and T\(_{EM}\) cells rapidly respond to stimuli, T\(_{EM}\)
cells have a decreased capacity for proliferation as well as shortened telomeres indicative of their more terminally differentiated state [87].

Epigenetic regulation of T-cell differentiation

While naïve and memory cells harbour very different functions, the vast majority of genes are expressed at similar levels during resting states, suggesting that acquisition of a stable differentiated T-cell lineage is largely maintained by epigenetic mechanisms [88]. Evidence of the importance of DNA methylation in T-cell biology can be seen during early development where HSCs lacking DNMT1 fail to suppress myelo-erythroid lineages and do not produce lymphoid cells [89]. Furthermore, conditional deletion of Dnmt1 in CD4+ T-cells impairs memory differentiation while effector cytokine production is increased upon activation of naïve T-cells [90]. Notably, loss of DNMT1 does not affect Tbx21 or Gata3 expression, suggesting that deregulated cytokine production is not due to impaired lineage choice [91]. Similarly, while HSCs lacking DNMT3A/B have defects in differentiation of all major lineages [92], T-cells conditionally lacking DNMT3A show reduced plasticity and fail to selectively silence IFN-γ and IL-4 upon activation [93]. Furthermore, a similar phenotype to Dnmt1 and Dnmt3a deletion, with increased IFN-γ and IL-4 production, is also observed in CD4+ T-cells lacking the methyl-DNA binding protein MBD2 [93]. Similarly to loss of DNMT function, loss of Tet2 result in altered differentiation of both myeloid and lymphoid precursors [94] but loss of cytokine expression in activated T-cells [95]. Thus, the regulation of DNA methylation is likely a key process in the appropriate differentiation of mammalian T-cell lineages.

Measuring DNA methylation in mammals

While knowledge of the absolute amounts of DNA modifications may yield some biological insight, the majority of the biological information is encoded in the spatial location of the modifications along the genome. Determination the genome-wide profiles of DNA modifications is typically accomplished through specific labelling of the target modification coupled with high-throughput sequencing. Detection methods for genome-wide mapping of DNA modifications can be broadly divided into two categories; base-resolution-sequencing and enrichment-based-sequencing.

Base-resolution techniques

Spontaneous hydrolytic deamination is one of the most common sources of endogenous DNA damage preferentially targeted to cytosine and 5mC, leading to conversion of uracil and thymine, respectively [96]. Exploiting this reaction, chemical deamination with sodium bisulfite instead preferentially converts cytosine while the conversion rate of 5mC remains very slow, allowing selective identification of 5mC under ideal conditions (uracil is read as thymine while 5mC is read as cytosine during PCR) [97]. When applied as whole-genome bisulfite sequencing (WGBS), over 90% of all CpG sites can be investigated at single-base resolution [98]. Due to the genome-wide quantitative readout of WGBS, it is generally considered the gold standard method of profiling 5mC.
As CpG dinucleotides are underrepresented in the genome, a large portion of WGBS data is uninformative in regards to methylation status. Alternatively, converted fragments can be hybridised to beadchips containing around 480,000 CpG probes [99] allowing quantitative base-resolution measurement at specific CpGs for a fraction of the cost of WGBS. However, while covering majority of promoters, the arrays are underrepresented in certain regulatory elements, such as enhancers. Strangely, this problem persists even in newer iterations of these arrays although they cover nearly twice as many CpGs [100]. Another problem with methylation arrays are that a subset of probes may cross-hybridise or map incorrectly to their target sequence [101]. Moreover, as these arrays contain two different probe types they may introduce type-specific dye biases which are typically corrected for computationally [102].

Conversely, reduced representation bisulfite sequencing (RRBS) offers a more targeted view of the genome (~0.5%) through the use of BglII digestion followed by size selection of fragments [103]. This approach results in higher coverage of enhancers but is generally more expensive than arrays [104]. Furthermore, regions covered by RRBS are generally biased towards high CpG-density regions and results can be inconsistent due to the usage of restriction enzymes [104].

However, as DNA base complexity is effectively reduced upon bisulfite conversion, balanced artificial libraries need to be included to avoid sequencing errors [98]. Furthermore, bisulfite conversion needs to be highly efficient to both distinguish 5mC from unconverted cytosine while avoiding deamination of 5mC to thymine [105]. Although bisulfite sequencing is regarded as the gold standard method for quantitative detection of 5mC, it cannot distinguish between 5mC and 5hmC [106-108]. For this reason, numerous different methods have been developed for profiling of each individual DNA modification [21]. These techniques typically rely on enzymatic- or chemical conversion of modified bases coupled with bisulfite conversion. However, bisulfite treatment is known to degrade DNA, especially long fragments, therefore recent efforts have focused on the development of bisulfite-free methods for base-resolution detection of DNA modifications [109, 110]. Furthermore, as bisulfite conversion largely reduces the complexity of the DNA, especially for rare modifications, care has to be taken to minimize PCR amplification biases [111].

Another base resolution approach to mapping DNA modifications is single molecule real time (SMRT) sequencing. SMRT-seq utilizes zero-mode waveguides to measure incorporation of labelled nucleotides by a DNA polymerase in real-time [112, 113]. As base modifications alter polymerase kinetics, DNA modifications can be directly inferred from the time lag between nucleotide incorporations [114]. Although SMRT-Seq offers potential measurement of all DNA modifications simultaneously without prior amplification, high error rates (13-15%) and costs are currently limiting whole-genome implementation to small genomes [114]. Furthermore, as each polymerase is holding approximately 12 nucleotides simultaneously, local sequence context is a major determinant of signal and modifications can give rise to a false signal by affecting the incorporation of surrounding nucleotides [115]. Compared to bisulfite sequencing, which estimates modification rate directly from the sequence (~5–10x
coverage required), modification calling in SMRT-seq is based on statistical tests and extremely deep coverage (~150x) is required to identify low-abundance modifications [116].

**Enrichment-based techniques**

Although bisulfite sequencing and SMRT-seq provide information about the methylation status of individual CpGs, high sequencing coverage is required for robust methylation calls making these techniques prohibitively expensive. Furthermore, as CpGs are underrepresented in mammalian genomes, a high number of sequencing reads are likely to be uninformative regarding 5mC content. For these reasons, methods that specifically enrich for modified regions in DNA are often employed.

One of the earliest methods used the methyl-CpG-binding domain (MBD) from MeCP2 to affinity-purify CpGs which are then eluted using a salt gradient to yield fractions with different amounts of 5mC [117]. While MBD has the highest affinity for high-density CpG regions, low- and medium density regions could be recovered at lower NaCl concentrations, although this would require several sequencing runs for a single sample [118]. Similarly, DNA immunoprecipitation (DIP) utilises a modification-specific antibody to capture modified DNA fragments [119]. While MBD enrichment requires symmetrically methylated CpGs in double-stranded DNA [120], DIP works on single-stranded DNA and can therefore be used to profile hemi-methylated sites [121]. In profiling 5mC at CpG sites, DIP-seq is highly concordant with MBD-seq [122] albeit with a preference for lower CpG densities [123]. As new antibodies were developed, DIP-seq enabled detection of all currently known 5modC in mammals [124]. While 5mC DIP profiles broadly match those obtained with bisulfite sequencing at highly methylated CpGs [122], several studies suggested that antibodies used in DIP also enriched for unmodified DNA fragments [125-127].

Soon after the discovery that TET enzymes mediated 5hmC conversion, selective chemical labelling (Seal) approaches were developed in the absence of a commercially available antibody to produce genome-wide 5hmC profiles [128, 129]. These techniques utilize β-glucosyltransferase (β-GT) to transfer a glucose moiety from uridine diphosphate glucose (UDP-Glu) to the hydroxyl group of 5hmC followed by attachment of a biotin tag which can be enriched using streptavidin [128, 129]. The technique can further be extended to profile 5fC through blocking of 5hmC with unmodified glucose followed by selective reduction of 5fC to 5hmC using NaBH₄ and biotin-labelling [130]. While antibodies can show density-dependence, Seal techniques provide unbiased genome-wide profiles with less background [126, 129], albeit with slightly lower DNA recovery [126].

A key aspect of all enrichment-based techniques is that only relative abundances are measured and a control sample must be included to account for experimental confounders and uneven sequencing coverage [131]. Both non-specific IgG and non-enriched INPUT DNA have been suggested as appropriate controls for antibody-based methods [131] but INPUT controls have prevailed as the most commonly used in DIP-seq protocols [132, 133]. Even though the NIH Roadmap Epigenomics project produced both IgG and INPUT controls for
5mC DIP-seq [122], the consortium never published a comparison between the two controls and the data remained publicly unavailable. It is likely that the choice of INPUT controls were due to both practical and theoretical reasons; firstly, INPUT DNA was readily available and did not require an extra experiment, secondly, IgG controls were thought to only enrich small amounts of DNA which would not produce an even genomic background required for early peak finding algorithms [131]. While these algorithms would move on to model local backgrounds to avoid local biases and division-by-zero problems [134], INPUT samples remained preferred. For Seal-seq, both INPUT and non-biotin labelled DNA produced similar profiles with low- to no enrichment [128]. Although streptavidin has been reported to bind DNA 8-oxoguanine [135], the binding affinity may not be strong enough to persist through washing.

While cost-effective, enrichment-based techniques do not provide a quantitative estimate of DNA modifications at any given location and resolution is limited to fragment size which is typically around 200 bp.

**Bioinformatics analysis of genome-wide DNA methylation data**

Computational analysis of epigenomic data generated by both base-resolution and enrichment–based profiling of DNA modifications poses several unique challenges. As millions of short sequence reads are generated, they must first be aligned to the genome to provide a functional location. Simple algorithms, such as BLAST, optimise alignment based on local sequence similarity [136]. While effective for a low number of short sequences, this becomes too resource-intensive for millions of reads from highly complex mammalian genomes. Modern alignment tools generally convert genome sequences into more efficiently indexed data structures against which reads are queried to identify the best matching location [137]. As bisulfite conversion changes the DNA sequence, it will no longer exactly match the reference. Furthermore, due to asymmetrical modifications, the reference must not only include both original strands but also their converted form (C>T and G>A for forward and complementary strands, respectively) making WGBS alignment highly complex as this data must be combined to identify 5mC sequence substitutions [138]. Conversely, enrichment-based sequencing produces highly localised reads forming enriched ‘peaks’. Peak-calling algorithms vary based on assumptions of underlying data and statistical frameworks but generally have some key steps in common [139]. As only a fraction of each fragment is typically sequenced, reads must first be extended to represent the original fragment. If both ends of the fragment are sequenced as paired-ends, the insert length can be directly inferred while for single-end sequencing, the insert length can be estimated based on the average distance between read pile-ups on opposite strands (i.e. each end of the fragment) [140]. Next, enriched peaks are identified relative to the local background which can be either a control sample, modelled statistical noise or both [134]. Inclusion of a control sample is considered critical as it can reduce false positive regions by removing experimental and/or technical artefacts [141] and generally lowers false discovery rate (FDR) [142]. Finally, enriched peaks are filtered based on some user-defined criteria, typically based on fold enrichment or statistical significance.
While peak-calling algorithms work well for highly enriched regions, identification of diffuse and weakly enriched regions remains challenging.

Although sequencing technology has provided an unprecedented resolution in which to view the genome, it is not without flaws. Only a set portion of the known genome is uniquely mappable at a given read length and reads often map to multiple locations. Mappability is especially problematic for short reads and bisulfite converted DNA where up to 16% of all CpGs may not be uniquely mapped even with 100 bp reads [144]. Furthermore, around 50% of the human genome consists of repetitive sequences that can exist as multiple copies dispersed throughout the genome. The repetitiveness of the human genome is a feature that has been known for decades [145] and sequence alignment tools generally deal with ambiguously mapping reads by random assignment, best match, reporting all hits or completely ignoring the read [146]. While mapping algorithms have yet to fully address multi-mapping, several issues related to mismapping can be resolved by increasing read length but not all [144]. For fragments spanning non-unique regions, it may be enough with one end to map uniquely to fully place the whole fragment, highlighting the strength of paired-end sequencing in these cases [146]. Furthermore, the emergence of long read sequencing holds great promise in resolving repetitive regions and filling in gaps in the human genome [147, 148].
Materials and methods

Ethics statement

Material from healthy individuals (Paper I)

Studies were approved by Linköping University ethics board, Diary Number: 2013/408-32.

Material from murine cell lines (Paper II)

DNA was obtained from established mESC cell line J1, no ethical permission was required.

Cell isolation

Isolation of peripheral blood mononuclear cells (PBMCs) (Paper I)

Buffy coats were obtained via the Linköping University Hospital blood transfusion clinic from healthy blood donors after written consent. This blood fraction is largely depleted of red blood cells and plasma and PBMCs were further enriched by density centrifugation.

Magnetic-activated cell sorting (MACS) (Paper I)

To isolate either total CD4⁺ or NT cells, magnetic separation was performed using positive or negative magnetic selection, respectively. Briefly, surface antigens are labelled with antibodies conjugated with superparamagnetic nanoparticles which are captured in columns under a magnetic field. As monocytes express CD4 at low levels [149], there is a risk of contamination using positive selection unless additional sorting steps are included. However in our hands, monocyte contamination was typically < 1%, as determined by cell size and complexity as well as CD4 signal strength, and MACS has routinely resulted in cell purities ranging from 95 to 99%.

Fluorescence-activated cell sorting (FACS) (Paper I)

Flow cytometry is based on light scattering and fluorescence emission of single cells contained in droplets generated by hydrodynamic focusing [150]. While simultaneously being hit with a light source for detection, a charging electrode can apply a charge to the droplets enabling electrostatic cell sorting through the downstream use of charged deflection plates [150]. Primary T-cell memory subsets were sorted from total CD4⁺ T-cell samples on the basis of surface expression of the chemokine receptor CCR7 [86]. Purity of sorted NT (CD4⁺CD45RO⁻CCR7⁺), TCM (CD4⁺CD45RO⁺CCR7⁻) and TEM (CD4⁺CD45RO⁻CCR7⁻) was determined by reanalysis of sorted samples and resulted in an average purity of 97%. CD45RA expression of sorted NT cells was also verified through reanalysis.
Cell lines

*Jurkat (Paper I)*

The human CD4⁺ T-cell leukaemia line Jurkat was developed in the 1980s and has historically been used to study TCR signalling as activation triggers IL-2 production similarly to primary human cells [151].

*Mouse embryonic stem cells (Paper II)*

DNA from wild-type or *Dnmt* TKO J1 mESCs [152] was provided by R.R. Meehan (University of Edinburgh).

Cell culture manipulation

*In vitro CD4⁺ T-cell polarisation (Paper I)*

Primary NT cells were activated using plate-bound anti-CD3 and soluble anti-CD28 in the presence of IL-2, IL-12 and anti-IL-4 or IL-2, IL-4, anti-IFN-γ and anti-IL-12 for differentiation of T₈₁ or T₈₂ lineages, respectively [153]. Although the engagement of lineage-specific transcription factors will inhibit the opposing lineage, this is further enhanced by the inclusion of neutralising antibodies. Stimuli was replenished after 3 days in culture.

*Cell cycle and division detection (Paper I)*

Proliferation of cells was tracked using the fluorescent dye carboxyfluorescein diacetate succinimidyl ester (CFSE). This dye is highly membrane-permeable due to the two acetate side chains which are cleaved by esterases once inside the cell. The succinimidyl moiety reacts with amine groups to form a highly stable amide bond and while labelled short-lived molecules are typically depleted within the first 24 h, long-lived molecules, such as DNA, can be labelled for up to 6 months. The intensity of CFSE is progressively halved during cell division enabling detection of up to 8 cell divisions [154].

Conversely, the ethidium bromide analogue propidium iodide (PI) binds nucleic acids and enables quantification of DNA content per cell allowing estimation of cell cycle phase [155]. As PI is not membrane-permeable, live cells have to be permeabilised by fixation to allow staining. Furthermore, fixed cells have to be RNase treated as PI also binds RNA and care has to be taken to avoid cell aggregates as doublets could be misinterpreted as one cell containing twice the DNA amount. Single cells can easily be distinguished by flow cytometry based on signal pulse area compared to signal pulse width.

Gene expression analysis

*Microarray (Paper I)*

RNA was spiked with synthetic control transcripts, converted to cDNA through first-strand synthesis then *in vitro* transcribed, amplified and labelled with Cy3-CTP. cRNA was purified and quantified then fragmented and hybridised to Agilent SurePrint G3 Human Gene
Expression 8x60K v2 microarrays and scanned on an in-house microarray scanner. Raw intensities were extracted and data was background corrected and quantile normalised using the R package LIMMA [156]. Control- and unannotated probes were removed and undetected probes (background + 10%) were excluded.

These microarrays cover the majority of known genes as well as around half of known transcript isoforms. When these experiments were carried out in 2014, microarrays were still commonly used to determine gene expression and sample processing, biases and normalisation procedures were relative well-known and standardised [156]. Contemporary analyses also revealed that although RNA sequencing (RNA-seq) was more sensitive to smaller changes in expression, the vast majority of differential expression was comparable between the two methods [157] while microarrays were considerably less expensive [158]. Since then, the cost of RNA-seq has massively decreased while both quantification methods and statistical frameworks for differential expression have become more sophisticated and sensitive [159, 160]. Although the dynamic range of microarray is constrained compared to other techniques, data typically show a good correlation with both qPCR [161] and RNA-seq [162].

Quantitative PCR (qPCR) (Paper I)

RNA was converted to cDNA through reverse transcription and quantified using specific TaqMan probes in a real-time cycler. Relative expression was calculated using the comparative ΔΔCt method [163] using GUSB as a control gene which we have previously found to be stable in T-cells through microarray analysis [164].

Global analysis of DNA methylation

Immuno-dot blot (Paper I & II)

DNA was heat denatured and dotted onto a positively charged nylon membrane under vacuum. Membranes were briefly washed and air-dried, UV cross-linked then blocked using a casein buffer. Incubation with primary 5modC antibodies was followed by three TBS-T washes and incubation with an HRP-conjugated secondary antibody. Finally, membranes were washed twice in TBS-T then once in TBS and visualised by UV light using chemiluminescence.

Enzyme-linked immunosorbent assay (ELISA) (Paper II)

Reaction plates were coated with DNA fragments overnight at 37°C then blocked and washed followed by incubation with primary 5modC antibodies. Plates were then washed and incubated with an HRP-conjugated secondary antibody. Finally, plates were washed and visualised using a microplate reader.

Mass spectrometry (Paper II)

DNA was heat denatured and cleaved into nucleoside 5’ monophosphates using a nuclease followed by precipitation of proteins and a buffer exchange. Samples were separated and analysed by ultra performance liquid chromatography (UPLC) coupled with a triple quadrapole
mass spectrometer where individual bases were detected in three time windows corresponding to C and 5hmC, 5mC and 5caC or 5fC using multiple reaction monitoring (MRM).

Genome-wide analysis of DNA methylation

Infinium 450k methylation arrays (Paper I)

Profiling of DNA methylation on Infinium 450k methylation arrays was performed at Barts and the London Genome Centre, Queen Mary University of London. Raw data files were pre-processed using subset-quantile within array normalisation (SWAN) as implemented in the R package minfi [165] to control for probe-type dye biases. Differentially methylated probes were defined by $P < 0.05$ and an absolute 5mC change of 10, 20 or 30% compared to NT for 1 day, 5 day and memory cells, respectively.

DNA immunoprecipitation (DIP) sequencing (Paper I)

DNA was fragmented by sonication to a peak size around 200 bp (range 50 – 500 bp) then end-repaired and dA-tailed. Illumina sequencing adapters were ligated and samples were heat denatured where 10% INPUT DNA was set aside. Next, adapter-ligated fragments were subjected to immunoprecipitation overnight at 4°C using a 5hmC-specific antibody. Antibody-bound fragments were captured using magnetic Protein G beads then washed three times and digested using proteinase K. IP and INPUT samples were amplified by PCR for 13 and 9 cycles, respectively, to avoid over-amplification of INPUT samples. Samples were separated on a 2% low melting point agarose gel and 100 – 400 bp fragments were selected and purified. Libraries were paired-end sequenced on a HiSeq 2000 platform at Beijing Genome Institute (BGI). Data was aligned to the human hg19 reference genome using bowtie (-v 1 --best) [166] and enriched peaks were identified using MACS [142] using corresponding controls. Differential enrichment was performed using the MACS sub-module macs2diff.

This DIP-seq protocol retains strand information due to the directionality of forked adapters, allowing analysis of asymmetric modifications [121], although this information remains largely underutilised [167]. While base-resolution bisulfite sequencing remains the gold standard for 5mC profiling, no such standard has yet been established for 5hmC. Techniques have been developed for detection of 5hmC at base-resolution but typically suffer from high coverage requirements and compound experimental confounders due to multiple conversion steps which were not computationally addressed until recently [168]. Enrichment-based alternatives to DIP-seq were anti-cytosine-5-methylenesulfonate (CMS) and 5hmC-Seal but the required CMS antibody was not commercially available and Seal lacked a standardised protocol, unlike DIP-seq [133].

Locus-specific analysis of DNA methylation

DIP-qPCR (Paper II)

DNA was fragmented by sonication to a peak size around 400 (range 100 – 1,000 bp) then heat denatured. 10% INPUT was set aside and immunoprecipitation was performed as for DIP-
seq. qPCR was performed using specific primers and SYBR green reagents. Primer efficiencies were determined by 5-fold dilution series and calculated based on the linear regression in log10 scale against cycle threshold (Ct). Samples were quantified based on primer regression statistics then normalised to INPUT:

\[
\text{Efficiency} = -1 + 10^{(-1/slope)} \\
\text{Quantity} = 10^{(Ct-\text{intercept})/slope}
\]

Primers utilised had an average efficiency of 111% (± 17%).

*Methylation-sensitive restriction digestion (Paper I)*

DNA was glycosylated using T4-βGT then digested using MspI or HpaII and analysed using qPCR. MspI does not cleave glycosylated 5hmC while HpaII only cleaves unmodified DNA allowing for identification of 5hmC and 5mC+5hmC, respectively. An uncut control is included to identify total DNA amount and the full reaction is run in duplicate without T4-βGT. In this control reaction, MspI cleaves 5mC, 5hmC and unmodified recognition sites. Approximate inner cytosine modification level can then be estimated from Ct values at as:

\[
\begin{align*}
C^{5\text{hmCGG}}\% &= (\text{MspIT4-βGT}\times(\text{DNA}_{\text{control}}/\text{DNA}_{\text{T4-βGT}}) - \text{MspI}_{\text{control}})/\text{DNA}_{\text{control}} \\
C^{5\text{mCGG}}\% &= (\text{HpaII}_{\text{control}} - \text{MspIT4-βGT}\times(\text{DNA}_{\text{control}}/\text{DNA}_{\text{T4-βGT}}))/\text{DNA}_{\text{control}} \\
C^{\text{CGG}}\% &= (\text{DNA}_{\text{control}} - \text{HpaII}_{\text{control}})/\text{DNA}_{\text{control}}
\end{align*}
\]

As this technique does not rely on antibodies or bisulfite conversion, it offers independent single CpG validation for 5mC and 5hmC detection techniques but is limited to CpGs in certain recognition sites (C^CGG). Primers utilised had an average efficiency of 101% (± 27%).

*Analysis of single nucleotide variant (SNV) interactions*

*Mapping of variants in high linkage disequilibrium (LD) (Paper I)*

Genome-wide association study (GWAS) single nucleotide variants (SNVs) (n = 73,196) were obtained from GWASDB2 (accessed October 2015) [169] and variants in high linkage disequilibrium (LD) (\(r^2 > 0.8\)) were extracted from 1000 Genomes Phase 3 v5 LD blocks [170] using the SNIPA-tool [171], resulting in 600,320 variants in total.

All variants were then overlapped with regions gaining 5hmC by day 1 of *in vitro* polarization and compared with variant frequencies in randomly selected genomic regions of equal size to the 5hmC set, excluding all centromeric and telomeric DNA, which were bootstrapped 10,000 times.
Electrophoretic mobility shift assays (EMSA) (Paper I)

Cy5-labelled 45 bp oligonucleotides corresponding to major and minor alleles were synthesised. Nuclear protein extracts were prepared through mechanical homogenisation of cells in a hypotonic buffer followed by addition of NP-40 to an end-concentration of 0.5%, vortexed and centrifuged. Supernatants were discarded and nuclei were washed then resuspended and incubated in a hypertonic buffer for 20 min at 4°C with rotation. Samples were centrifuged and supernatants containing nuclear proteins were collected and stored at -80°C. Nuclear proteins were blocked using poly(dI-dC) for 10 min on ice then Cy5-labelled oligonucleotides were added for an additional 20 min. DNA-protein complexes were separated on a 5.3% polyacrylamide gel and visualised on a phosphoimager. Binding for each allele pair was optimised individually and replicated 5-11 times. Bands were assessed manually and ImageJ was used to quantify and validate diffuse differences.

Reanalysis of epigenomic data

Chromatin immunoprecipitation (ChIP) sequencing (Paper I & II)

For paper I, pre-processed histone ChIP-seq data was obtained for T-cells (NT [172], T_h1 and T_h2 [173]).

For paper II, raw histone ChIP-seq data for mESCs was obtained [174] and mapped to the mouse mm9 genome reference using bowtie2 (-N 1 –L 30) [175].

Long-range chromatin interactions (Paper I)

Pre-processed promoter capture Hi-C (cHi-C) data was obtained for Jurkat cells [176]. Data was restricted to interactions on the same chromosome and within 300 kb of SNVs of interest. Genomic regions were partitioned into 5 kb bins to estimate interaction densities and central bins were excluded to avoid over-saturation. Sashimi plots of interactions were visualised using the R package Sushi [177].

DIP-seq (Paper II)

Data was obtained for mouse [36, 121, 124, 125, 130, 178] and mapped using bowtie2 (-N 1 –L 30) [175] to the mouse mm9 genome reference. Peak-calling was performed using MACS2 (--bw=200 –p 1e-5) [142].

Whole-genome bisulfite sequencing (WGBS) (Paper II)

Data was obtained for E14 mESCs [179] and aligned using Bismark (-N 1) [180] to mm9. Methylation levels in CpG, CHG and CHH contexts (where H = A, T or C) were extracted with a minimum coverage requirement of 5.
Seal and anti-CMS techniques (Paper II)

Data was obtained for 5hmC- and 5fC-Seal for mESCs [130] and mapped to the mouse mm9 genome reference using bowtie2 (-N 1 –L 30) [175]. Anti-CMS data was obtained for mESCs [129] and aligned using Bismark (-N 1) [180] to mm9.

Other statistical considerations

Paper I

Proportional differences were calculated using Fisher’s exact test. Differences in relative values was calculated using T-tests with FDR correction if multiple testing. Motif enrichment in peaks was performed using Hypergeometric tests as implemented in homer2 [181]. Gene ontology (GO) term overrepresentation test was performed using modified Fisher’s exact test with FDR correction as implemented in DAVID [182].

Paper II

Motif enrichment was performed using Hypergeometric tests as implemented in homer2 [181]. GO term overrepresentation test was performed using Fisher’s exact test with Bonferroni correction as implemented in PANTHER [183]. Differences in relative enrichment was calculated using T-tests with FDR correction if multiple testing. To control for uneven groups, bootstrap resampling was performed to estimate empirical distributions where 10,000 permutations were performed. Specific considerations regarding Paper II are outlined and discussed in detail in the results and discussion section.
Results and Discussion

Paper I

A fundamental concept in cellular differentiation is that cell states are established and/or maintained by epigenetic mechanisms [184]. These processes are typically studied in stem cells and whereas the developmental stages in these systems are generally well-known, access to these primary cell types is very limited in humans. As the immune system needs to react to a myriad of different pathogens, many lineages in peripheral blood are not terminally differentiated. A unique property of CD4+ T-cells is their ability to differentiate into multiple different stable subsets with distinct effector functions. T-cell differentiation is associated with widespread epigenetic re-programming [185]. Moreover, active DNA demethylation has previously been suggested to occur in activated T-cells [186] and we hypothesised that this may be driven by 5hmC remodelling.

![Figure 5. T-cell differentiation as a tool to study DNA demethylation. (A) Cell division measured by carboxyfluorescein succinimidyl ester (CFSE) dilution. (B) DNA replication measured by intracellular propidium iodide (PI) DNA staining. (C) Global DNA 5hmC content measured by immuno-dot blot. Data shown as mean ± s.d. for n = 5 biological replicates. *P < 0.05 against NT, T-test. (D) Expression of lineage-specific signature genes during in vitro T-cell differentiation determined by qPCR relative to GUSB. Data shown as mean ± s.d. for n = 3 biological replicates.](image)

Validation of T-cell differentiation as a model to study DNA demethylation pathways

Using cells purified from healthy blood donors, we employed a time-series design of in vitro differentiation (polarisation) of naïve T-cells (NT) towards T_{h1} or T_{h2} lineages. Time points were chosen as 0 h- (NT), 24 h- and 5 day stimulation to encompass resting-, activated-, and differentiated states, respectively. This design would also allow inference of how early locus-specific changes affects outcomes at day 5. However as differentiation is accompanied by
massive cell expansion, distinguishing between active- and passive DNA demethylation events may be difficult. Indeed, by day 5 over 99% of all cells had divided at least once with the majority of cells having divided up to 5 times (Figure 5A). Conversely, activated cells did not enter the cell cycle until day two (Figure 5B), making the 24 h time point an ideal window in which to study DNA methylation dynamics in the absence of replication. While global 5hmC content was largely depleted by day 5, levels were retained after 24 h activation (Figure 5C). Induction of gene expression was already observed at 24 h for key lineage-defining genes (Figure 5D) as well as around 4,000 other genes (not shown) and we hypothesised that 5hmC remodelling at this time point, if any, would be highly gene-specific. To answer this, we generated genome-wide 5mC and 5hmC profiles for each time point.

**Genome-wide 5hmC remodelling initiates demethylation during T-cell differentiation**

Similar to our 5hmC measurements using immuno-dot blots (Figure 5C), we observed a global loss of 5hmC over gene bodies by day 5 as well as genome-wide loss of 5mC (Figure 6A). Conversely, a major increase in 5hmC-enriched regions, primarily at introns, was observed by 24 h of activation (Figure 6B). The distribution of 5hmC across these regions was highly similar that observed for T-cell specific enhancers (H3K4me1) [173] (5hmC = 53%, 9%, 32% and enhancers = 57%, 3%, 40% for introns, exons and intergenic regions, respectively) suggesting that 5hmC remodelling occurs primarily at enhancers. Indeed, 63-67% of regions gaining 5hmC by day 1 overlapped T-cell H3K4me1 peaks [173] (not shown) and marked key T-cell genes (Figure 6C). The rapid increase in 5hmC during activation followed by a subsequent loss of 5mC during cell division suggested a mode of demethylation through 5hmC conversion followed by passive, replication-dependent loss. An alternative explanation is due to loss of DNA methylation maintenance but DNMT1 and UHRF1 remained highly expressed throughout the differentiation series (Figure 6D and not shown). To infer demethylation dynamics directly from our data, we compared 5mC changes over time within regions gaining 5hmC by day 1 and found that these regions were highly predictive of 5mC loss (Figure 6E). As 5hmC regions may be biased by genomic context, i.e. enhancers, we also stratified genome-wide 5mC changes by day 5 and measured relative 5hmC enrichment at these sites, revealing again that sites losing 5mC by day 5 were enriched for 5hmC by day 1 (Figure 6F). Collectively, these results suggested a mechanism by which T-cell enhancers are primed with 5hmC followed by passive DNA demethylation to yield a permissive transcriptional landscape upon terminal differentiation. Such a dynamic system has previously been described for imprint erasure in PCGs [187] but, to our knowledge, never in the human immune system. Conversely, active DNA demethylation through 5hmC has previously been described during human monocyte differentiation [188]. However as monocytes differentiate post-mitotically, the DNA demethylation process is likely exclusively reliant on active removal in contrast to our observations in T-cells. As DIP-seq only provides relative values, it may be biased in a system with massive cell division, such as T-cell differentiation, as modification amounts are rapidly lost globally. Indeed, spike-ins have been suggested to normalise for varying global modification levels [189] but this approach has not yet been widely adopted. However, as our main focus was on 5hmC changes in the absence of replication, the majority of these issues...
could be mitigated. Furthermore, as global 5hmC was measured using dot blots (Figure 5C) we were aware of major dilution effects and avoided over-normalisation of data.

Figure 6. Genome-wide DNA demethylation mediated by 5hmC. (A) Gene body 5hmC levels measured by DIP-seq. (B) Genomic distribution of 5hmC enriched peaks. ***P < 0.001, Fisher’s exact test. (C) Gene ontology enrichment for genes gaining 5hmC peaks at day 1. (D) Expression of DNMT genes during T-cell differentiation measured by qPCR relative to GUSB. Data shown as mean ± s.d. for n = 3 biological replicates. (E) Genome-wide 5mC change in regions gaining 5hmC or random background probes. (F) Genome-wide 5hmC enrichment at 5mC sites. ***P < 0.001 against all other groups, one-way ANOVA, Tukey’s test.

5hmC priming establishes epigenetic memory in T-cells

We have previously shown that T-cells passively lose 5hmC under prolonged cell culture, likely due to the lack of the TET co-factor vitamin C [190]. Furthermore, as in vitro differentiation may not mirror physiological conditions, we wanted to validate our findings in vivo. Primary Th1- and Th2- specific memory cells can be isolated from blood [191] but the surface markers for these subsets are often ambiguous [192] and whether the Th1/Th2 paradigm holds true in vivo for humans is debatable. Instead, we isolated two functionally well-defined T-cell memory subsets (TCM and TEM) [86]. These memory subsets exhibited a high degree of DNA methylation remodelling, both global 5hmC content (Figure 7A) and genome-wide 5mC (Figure 7B) which clearly separated each T-cell subset (Figure 7C), indicating the
important role of DNA methylation in establishing stable epigenetic memory of lineage [4]. As early activation signals such as IL-2, CD28- and TCR engagement have been shown to greatly affect CD4+ memory formation even though proliferative responses remain normal [193-195], we hypothesised that early 5hmC priming during activation may help establish stable memory 5mC states. First we compared in vitro profiles with memory cells and found that sites losing 5mC by day 5 of differentiation showed a high overlap with demethylated sites in memory cells (odds ratio TCM = 151 and TEM = 244, P < 2.2^{-16}, Fisher’s exact test) (Figure 7D), suggesting that our in vitro model of DNA methylation remodelling faithfully mirrors processes in vivo. Next, we investigated sites gaining 5hmC during activation and found these regions to be preferentially demethylated in memory cells (Figure 7E), suggesting that 5hmC priming can establish long-term epigenetic programs. Importantly, 5mC loss in 5hmC regions was not due to passive loss through cell division as randomly sampled CpG sites did not show the same degree of demethylation (Figure 7E). As bisulfite conversion cannot distinguish 5mC and 5hmC [106-108], we verified that sites were generally not marked by 5hmC and lost passively by cell division using methylation-sensitive restriction digestion (Figure 7F).

Figure 7. Early DNA methylation remodelling persists in T-cell memory. (A) Global DNA 5hmC content measured by immuno-dot blot. Data shown as mean ± s.d. for n = 3 biological replicates. *P < 0.05, **P < 0.01, T-test. (B) Genome-wide 5mC levels measured using 450k arrays. CpGs with methylation levels > 80% were considered as 5mC. Data represent median and first and third quartiles with whiskers extending 1.5 * IQR for n = 7-9 biological replicates. (C) Principle component analysis (PCA) for n = 7-9 biological replicates. (D) Venn diagram of sites losing 5mC compared to NT (E) Genome-wide 5mC changes in regions gaining 5hmC or random background probes. (F) Absolute modification levels measured using methylation-sensitive restriction digestion. Together, these results show that early 5hmC remodelling permeates into long-term established T-cell memory programs. CD4+ memory has long been thought to be established during early activation [85] and 5mC maintenance has been shown to be critical for T-cell memory formation [90]. Our findings that regions primed by 5hmC during early activation...
remain demethylated in long-lived memory cells is quite surprising. The global 5mC levels in both T_{CM} and T_{EM} cells suggest that they have undergone several rounds of cell division, a conclusion that has since been supported by an independent study [196], and memory formation is typically achieved once activating signals are removed [85]. One possible explanation is that DNA demethylation, together with chromatin remodelling, helps establish a permissive landscape without initiating transcription, similar to observations in monocyte differentiation [197], enabling establishment of a memory phenotype upon later signals. Alternatively, as 5hmC is typically deposited in a strand-specific manner [198] and lack any maintenance enzyme, early DNA methylation remodelling may contribute to asymmetric cell fates seen during CD4^+ T-cell differentiation [199]. Speculatively, this model would explain how both effector and memory lineages can be derived from a single T-cell clone within the first cell divisions [200] but remains to be fully experimentally addressed.

5hmC priming may be disrupted in disease

While TETs have been reported to act at distinct loci [201], they also exhibit overlapping binding sites [202]. To provide site-specificity, TETs are likely recruited by TFs. Indeed, key lineage factors have previously been shown to regulate locus-specific 5hmC levels in T-cells [95]. Furthermore, regions gaining 5hmC during T-cell activation were highly enriched for binding sites and partners of AP-1 (Figure 8A) which plays key roles in T-cell activation and function [202]. As TF binding can be directly influenced by the underlying sequence [203], it is tempting to speculate that disruption of TF binding by mutations may impact TET and/or DNMT recruitment and DNA methylation remodelling. While these complex interactions remains largely unexplored, the majority of genetic variants in multiple paediatric autoimmune diseases are enriched near genes involved in T-cell activation and cytokine signalling [204] undergoing 5hmC remodelling (not shown). Indeed, regions gaining 5hmC by day 1 were highly enriched for genetic variants associated with multiple autoimmune diseases (Figure 8B). As genetic material is inherited as non-random blocks in humans [205], we not only included reported single nucleotide variants (SNVs) but also those in high linkage disequilibrium (LD) (see methods) as reported SNVs are often not functional [206]. Unfortunately, the vast majority of SNVs are found in non-coding regions and are rarely experimentally validated in regards to function [207]. To this end, we manually selected SNVs overlapping regions gaining 5hmC at day 1 of T-cell differentiation associated with immune-related diseases, T-cell related genes and/or were predicted to break TF binding motifs [208]. To determine whether these candidate SNVs could affect TF binding, labelled DNA oligonucleotides were synthesised to contain either the major or minor allele then let to complex with T-cell specific nuclear proteins from either cell lines (Jurkat) or primary CD4^+ T-cells. Notably, 16 out of 22 tested variants (73%) produced allele-specific protein binding (Figure 8C and not shown) suggesting that TF binding in 5hmC remodelling regions is a frequent genetic target in disease. While this approach allowed inference of binding in 45 bp regions, the average 5hmC peak was 300 bp. Surprisingly, SNVs were not normally distributed along peaks (Shapiro-Wilk normality test P = 2x10^{-15}) but instead occurred in clusters of < 75 bp (not shown), suggesting that genetic variation is perhaps non-randomly targeted to binding sites. While providing cell-specific binding data, this approach is limited by lack of any
information regarding the identity of bound proteins and likewise, transiently bound proteins can be missed [209]. Alternatively, DNA can be labelled with biotin to enrich and profile bound proteins and while this approach was considered, we did not perform it. Furthermore, as some SNVs altered CpG sites, pre-treatment with methylating enzymes such as M.SssI could have been used to investigate the effect of DNA methylation on protein binding. Nevertheless, our results provided valuable functional information of how immune-related SNVs may regulate TF binding and potentially locus-specific 5hmC levels. Furthermore, experimental validation of SNV function remains critical as sophisticated computational functional prediction [208] only produced moderate agreement (AUROC = 0.68, not shown).

Figure 8. DNA methylation remodelling in disease. (A) Transcription factor binding motifs in regions gaining 5hmC by day 1, calculated using hypergeometric test. (B) Enrichment of disease-associated SNVs in regions gaining 5hmC by day 1, calculated using bootstrap resampling n = 10,000. (C) Allele-specific binding of T-cell proteins to autoimmune SNVs, red text indicates minor allele. Data representative of n = 5 technical replicates (D) Chromatin interactions of rs7203150 with nearby gene promoters, cHi-C, capture Hi-C. (E) T-cell co-expression matrix of genes in chromatin interaction regions, *P < 0.05 Spearman rank correlation.

While binding results provided information regarding TF binding, the majority of regulatory regions act through long-range interactions [210]. To link SNV-5hmC regions to genes, we incorporated T-cell-specific (Jurkat) long-range interaction maps (capture Hi-C) between autoimmune SNVs and gene promoters [176]. While genome-wide Hi-C interactions are typically limited to large regions due to high coverage requirements, capture Hi-C sacrifices genome-scale detection for pre-defined regions at a higher resolution, allowing detection of all promoter interactions within 500 kb [176]. This allowed not only association of regions to genes but also suggested how SNVs may pleiotropically regulate multiple genes through
analysis of co-expression of interacting regions (Figure 8D,E). Taken together, our results show that genetic variation in autoimmune diseases can disrupt cell-specific TF binding which could potentially lead to dysregulated 5hmC priming during T-cell activation with pleiotropic effects on expression of surrounding genes. Indeed, chromatin interaction maps for activated human CD4+ T-cells have since been produced [211] revealing that autoimmune SNVs can directly influence expression of IL2RA, for which the promoter is marked by 5hmC and demethylation upon activation (Figure 7F) as well as validating the IL24-IL10 interaction in Jurkat cells that we show produce allele-specific TF binding in CD4+ T-cells (not shown).

Summary

Despite reports of locus-specific active DNA demethylation in CD4+ T-cells [186, 212], the underlying mechanisms have largely gone undetermined. The only model currently supported is active demethylation mediated by 5hmC [213]. By profiling both 5mC and 5hmC at multiple time points, we were able to identify regions increasing in 5hmC during T-cell activation followed by loss upon cell division. These regions not only marked genes important for T-cell activation and function but were also selectively demethylated in primary memory subsets, suggesting that 5hmC may also prime long-term memory stability. Indeed, genome-wide 5mC landscapes were distinct in T-cell memory populations whereas resting expression profiles were highly similar, indicating that memory is maintained at the epigenetic level. Furthermore, regions primed with 5hmC during early activation were highly enriched for autoimmune SNVs, many of which are able to affect T-cell specific TF binding, thus potentially affecting TET recruitment. We further show that these regions may exhibit pleiotropic gene regulatory effects through chromatin interactions and co-expression. Collectively, these results highlight the importance of dynamic DNA methylation remodelling in T-cell differentiation and points towards novel disease mechanisms in autoimmune disease.

Paper II

While our results from T-cells suggested a mainly passive mechanism of DNA demethylation, we observed loss of 5hmC at a rate higher than passive dilution in regions undergoing demethylation (Figure 6F) suggesting contribution of active mechanisms, at least in part. However, to more accurately characterise active DNA demethylation in T-cells, both 5fC and 5caC needs to be profiled [214]. As both 5fC and 5caC antibodies were available [124], mapping these modifications using DIP-seq would produce highly complementary data to our 5hmC profiles in T-cell differentiation. However, before starting the experiments, we sought to reanalyse previously published 5fC and 5caC DIP-seq data [124] to gain insight into genome-wide active DNA demethylation. Surprisingly, upon visual examination 5fC and 5caC profiles were highly similar to 5mC and 5hmC. Although all 5modC share the same substrate origin, the relative levels vary by orders of magnitude [17] and 5fC and 5caC were biologically unlikely to co-exist with more abundant modifications at such high rates. These observations were partly answered with the inclusion of an IgG control which revealed near identical enrichment profiles whereas INPUT controls produced no enrichment (Figure 9A). Historically, INPUT controls have been the preferred control as IgG was assumed to produce
non-specific signals [131]. However, the simple observation that IgG enrichment was near identical to 5modC suggested that these assumptions were false in DIP-seq but also raised new questions regarding the nature of this off-target binding.

![Figure 9. Characterisation of off-target binding in DIP-seq.](image)

**Figure 9. Characterisation of off-target binding in DIP-seq.** (A) Signal tracks for cytosine modifications in mouse embryonic stem cells (mESCs). WGBS, whole-genome bisulfite sequencing; STRs, short tandem repeats. (B) Consensus motif enrichment for IgG reads compared to INPUT reads. Representative of n = 3 biologically independent samples. SAT, satellite. (C) Signal tracks in wild-type (WT) or Dnmt triple-knockout (TKO) mESCs. (D) Enrichment plots over IgG peaks for WT, Dnmt or Tet TKO mESCs. Data shown as mean (left) or mean ± 95% confidence interval (c.i.; right) for n = 1 or 3 biologically independent samples, respectively. (E) Enrichment in Dnmt TKO mESCs over selected regions using DIP-qPCR and DIP-seq. Data shown as mean ±s.d. for n = 3 biologically independent samples (DIP-qPCR). (F) Enrichment in mESCs with different 5hmC profiling techniques over 5hmC or IgG peaks.

**Identification and characterisation of off-target binding in DIP-seq**

Enrichment profiles suggested that the IgG signal was at least in part localised to short tandem repeats (STRs) which was previously reported as an attribute of the 5hmC antibody [125, 126]. Indeed, IgG reads were highly enriched for STR sequences but also other DNA repeats, such as satellites (Figure 9B). To rule out any cross-reactivity between modifications for commercially available antibodies, immuno-dot blots and ELISAs were performed which
confirmed antibody specificities (not shown). This was further confirmed genome-wide by analysis of DIP-seq profiles from Dnmt or Tet triple-knockout (TKO) mESCs (Figure 9C, D) [36], suggesting inherent modification-independent binding. As off-target binding was high in repetitive regions which are known to cause mapping problems in short read sequencing [146], we validated IgG binding in Dnmt TKO mESCs using DIP-qPCR (Figure 9E), confirming that off-target binding was not due to mismapping of sequence reads. To test whether off-target binding was method specific, we compared mESC 5hmC data generated using DIP-seq [36], Seal [128] and anti-CMS [129] techniques. Interestingly, only DIP-seq produced off-target binding (Figure 9F) even though anti-CMS is also antibody-based, suggesting that pre-treatment of sodium bisulfite to convert 5hmC to CMS was enough to disrupt sequence-dependent off-target binding. Indeed, anti-CMS has been suggested to produce lower background levels compared to DIP-seq [129] which is likely due to lower off-target binding. While IgG-enriched regions showed a higher CA/TG and lower CG frequency in DIP-seq (not shown), PCR biases could be excluded through comparative analysis of DIP-seq and Seal data which showed near identical patterns (not shown). Collectively, these results show that IgG antibodies cause sequence-dependent off-target binding of repetitive DNA in DIP-seq. Enrichment-based sequencing is well-known to produce false positive signals in high copy-number regions, such as centromeric- and telomeric repeats [215, 216], whereas IgG DIP off-target signals are mainly in intronic repeats, such as STRs, suggesting a different mechanism. Indeed, artificially high read counts in high copy-number regions are likely derived through sample amplification as they are also observed in INPUT samples whereas exaggerated enrichment of certain sequences, such as STRs, is only observed with the use of an antibody, suggesting direct binding of the STRs to the antibody. In support of this, MBD-seq has been performed using an MBD protein fused either with GST tag [118] or the Fc domain of a human IgG1 antibody [217] where only the Fc-fusion shows enrichment of STRs [125, 218]. This suggests that off-target binding to STRs is primarily mediated by the Fc domain.

Off-target binding can efficiently be corrected using IgG controls

Having identified IgG off-target binding, we wanted to understand the degree to which this error has affected DIP-seq data, as it has rarely been controlled for. Pairwise comparison of genome-wide enrichment revealed that IgG enrichment could explain a substantial proportion of 5modC data ($r^2$ $5mC = 0.42$, $5hmC = 0.35$, $5fC = 0.55$, $5caC = 0.55$, INPUT = 0.05) (Figure 10A), suggesting that a non-trivial portion of DIP-seq data is due to off-target binding. This genome-wide correlation between IgG and 5modC in DIP-seq has previously been remarked upon but never explored [125]. To better quantify the contribution of off-target binding we developed a method for determining the empirical false discovery rate (FDR) from 5modC DIP-seq data using either IgG or INPUT controls (see methods). This revealed that for off-target binding (INPUT controls) resulted in FDRs of around 50% for 5mC and 5hmC while rates close to 99% were seen for 5fC and 5caC (Figure 10B). Conversely, FDR was tightly controlled for all 5modC by IgG controls at around 5% (Figure 10B), which is close to the estimated genome-wide mapping error rate [175]. Not only were the error rates very high but also showed an abundance-dependence as the lowest abundance modifications showed the highest error rates (Figure 10B). As repetitive DNA is constant between samples, the off-target
signal is likely to decrease with increased on-target binding. Indeed, data from mESCs where 5fC and 5caC levels were greatly increased through Tdg knockdown [124] showed a decrease in FDR of up to 50% when using INPUT controls, whereas the FDR remained largely unchanged using IgG (Figure 10C). Not only did IgG controls maintain FDR around ~5% but also greatly improved the overlap of 5hmC DIP-seq with anti-CMS and Seal, compared to INPUT, suggesting that IgG controls effectively remove background while maintaining the true signal (Figure 10D). This was further evident with the increased 5hmC signal over TET1-bound regions when using IgG controls (Figure 10E).

Figure 10. Data correction using IgG controls. (A) Correlation matrix of genome-wide DIP-seq peaks per Mb calculated as pairwise Pearson correlation coefficient squared ($r^2$). (B) Estimated false discovery rate (FDR) using IgG or INPUT controls in mESCs. Data shown as mean ± s.d. for 5caC $n = 2$, 5fC $n = 2$, 5hmC $n = 7$ and 5mC $n = 6$ biologically independent samples. (C) Estimated FDR in control (shCtrl) or Tdg knockdown (shTDG) mESCs. Data shown as mean of $n = 2$ biologically independent samples. (D) Venn diagram of 5hmC peak overlap in mESCs with different techniques and controls. (E) GO term enrichment for top genes ($n = 500$) enriched for 5hmC using different techniques and controls. P-values calculated using PANTHER biological processes. (F) TET1 binding over 5hmC peaks using IgG or INPUT controls. Data shown as relative to IgG ChIP-seq.

Taken together, these results revealed that an IgG control is critical to fully control for off-target binding in DIP-seq. As off-target signals display an inverse relationship with
modification abundance, it is likely to vary between different cell types and conditions where IgG controls will be particularly important. The exact sensitivity of DIP-seq remains to be fully elucidated as off-target binding is affected by experimental conditions. However, based on modification abundances alone [219, 220], the limit of detection using DIP-seq can be loosely estimated to >10^6 modifications per base in a mammalian genome (~one-tenth of 5hmC in most healthy tissues). As IgG off-target binding is likely mediated by the Fc portion (see above), it will not compete with on-target Fab binding. If antibody binding is fully saturated on both Fab and Fc domains, FDR would never go below an average of around 50% which we indeed observe in real data. However, off-target binding is likely weaker than on-target as repeated IP enriches for modified fragments [221], suggesting that off-target binding may also be reduced through optimised experimental procedures.

**Correction of off-target binding improves data quality and epigenetic insight**

Due to the extensive off-target binding, we wondered how these errors have affected data interpretation. Functional roles of modifications were unlikely to change as these are typically not inferred from profiling but from the mediating enzymes and/or more targeted experiments. Instead, we focused on the localisation and cross-talk of the different modifications. Early studies disagreed on the exact location and overlap between 5hmC and 5mC [36, 121, 129, 178, 222]. We hypothesised that these discrepancies were due to methodological differences as some studies used either DIP-seq with IgG controls, alternative techniques or DNA microarrays [36, 129, 222], whereas others used DIP-seq with INPUT controls [121, 178]. Indeed, controlling for off-target binding greatly reduced the overlap between 5mC and 5mC modifications (Figure 11A), highlighting a more distinct landscape and suggesting less redundant functions of each DNA modification. Unsurprisingly, IgG correction altered the association of repetitive elements with 5modC enriched regions, especially for low complexity regions and AG-repeats (Figure 11B). Whereas 5mC, 5hmC and 5caC enrichment decreased in STRs and CA-repeats, 5fC enrichment remained with up to 60% of peaks located in STRs (Figure 11B), suggesting a likely function of 5fC in STRs. Indeed, TDG was recently shown to possess glycosylase activity towards 5fC in both CpG and CpA dinucleotides and accumulation of 5fC and 5caC in CA-repeat containing genes was associated with silencing upon Tdg knockdown [223]. Unlike 5caC, 5fC seems to be selectively maintained in CA-repeats as it is typically rapidly excised by TDG [19]. This could possibly be mediated by enzyme kinetics as TDG activity towards 5fC is lower in CpA compared to CpG dinucleotides [223]. Furthermore, the enrichment of 5fC in CA-repeats explain the concomitant presence of the substrates 5mC and 5hmC albeit at greatly exaggerated levels without controlling for off-target binding [223]. Next, we wanted to determine how off-target binding in DIP-seq have affected associations with chromatin landscapes. Chromatin immunoprecipitation (ChIP)-seq has well-described biases towards euchromatin [224] and often produce false positive enrichment in active promoters [225]. Indeed, ENCODE histone ChIP-seq data [174] did not show enrichment in IgG DIP-seq regions (Figure 11C), suggesting that IgG off-target binding observed in DIP only occurs with free DNA. Whereas INPUT controls resulted in modification-specific, yet diffuse, chromatin associations for both 5mC and 5hmC DIP-seq peaks, IgG correction produced highly distinct profiles matching known functions (Figure
Enrichment in 5mC regions was improved for heterochromatin marks H3K27me3 and H3K9me3 as well as gene-body H3K36me3 which together with 5mC is involved in mRNA splicing [226] while association with the actively transcribed promoter mark H3K4me3 remained low (Figure 11D). Conversely, 5hmC showed the greatest improvement for permissive modifications, such as enhancer H3K4me1 and active gene promoter H3K9ac, whereas enrichment for heterochromatin H3K27me3 and H3K9me3 showed only small changes (Figure 11E). Together, these results demonstrate the profound effects correction of off-target binding can have on DIP-seq data interpretation. This was mainly due to extensive removal of data noise and would thus allow identification of more subtle enrichment. Indeed, IgG correction had a clear impact on signal-to-noise ratios estimated from wild-type and Dnmt TKO profiles (median S/N IgG = 5.84, INPUT = 0.84, random = 0.70, not shown).

Figure 11. IgG correction alters data interpretation. (A) Venn diagram (top) and paired line plot (bottom) of 5mC and 5hmC overlap using IgG or INPUT controls. Data shown as mean and individual data points of $n = 6$ biologically independent samples (bottom). P-values calculated using two-tailed paired T-test. (B) Fraction of enriched 5modC peaks identified using IgG or INPUT overlapping repetitive elements (top) or dinucleotide repeats (bottom) for 5caC $n = 2$, 5fC $n = 2$, 5hmC $n = 7$ and 5mC $n = 6$ biologically independent samples. Data shown as mean ± s.d. P-values calculated using two-tailed T-test. STR, short tandem repeat; LC, low complexity. (C) Relative enrichment of ENCODE mESC histone ChIP-seq data for $n = 26$ biologically independent samples in IgG DIP-seq peaks or random regions of same size and chromosome. Data shown as median and first and third quartiles with whiskers extending 1.5 * IQR. P-values calculated using two-tailed T-test. (D,E) Enrichment of ENCODE mESC histone ChIP-seq data for 5mC- (D) or 5hmC peaks (E) using IgG or INPUT controls. Data presented as mean (IgG) and bootstrapped mean (INPUT) of H3K27ac $n = 2$, H3K36me3 $n = 4$, H3K4me1 $n = 6$, H3K4me3 $n = 4$, H3K9ac $n = 2$, H3K27me3 $n = 2$, H39me3 $n = 4$ biologically independent samples, #P < 1e-5, bootstrap resampling (n = 100,000).
Summary

While errors related to off-target binding in DIP-seq have remained largely unknown, symptoms have been observed and described in the literature [125-127]. Through comparative analysis of multiple modifications and techniques, we were able to identify the cause of major off-target binding events in DIP-seq as binding of low complexity sequences, such as repeats, directly to IgG antibodies. We reveal that these errors can be corrected for using IgG control samples resulting in increased data quality and biological insight through removal of noise. Although IgG controls have long been suggested as an appropriate control [131], it is rarely used. This is likely based on the incorrect assumption that IgG will not retrieve enough material for sequencing. Going forward, inclusion of IgG controls will enable normalisation for the effects of off-target binding in DIP, allowing faithful characterisation of functional methylome landscapes even at lower modification abundances.
Concluding remarks

Epigenetic remodelling plays a central role in cell differentiation and development. Yet, the exact mechanisms of how epigenetic states are established and dysregulated in disease remain largely unknown. This thesis aimed to resolve a piece of this puzzle using T-cell differentiation which is a well-defined system for differentiation of somatic cells.

We reveal the complex interplay between DNA modifications in T-cells and how early remodelling events can result in long-term establishment of cellular identity. The regions undergoing DNA methylation remodelling were enriched for genetic variants associated with autoimmunity suggesting that establishment of the epigenetic landscape is a frequent target for mutations in disease. The majority of the variants tested had the ability to affect protein binding, illustrating how the underlying DNA sequence can affect multiple layers of epigenetic regulation in a cell-type specific manner. A key aspect of these findings is the dynamic change observed over time for DNA methylation. T-cell differentiation is commonly studied only at end-points where a stable lineage has been established but our results clearly show the power of time-series approaches for studying epigenetic regulation. While we have shed light on the DNA methylation dynamics in mature T-cells, it remains to fully separate active- from passive demethylation as well as determining the individual contribution of each DNMT and TET enzyme in establishing stable lineage-commitment.

As part of our investigations into active DNA demethylation we discovered that DIP-seq, which is a commonly used technique to study DNA modifications, produced a high degree of false positive results. We traced these errors to off-target binding of IgG antibodies utilised in DIP and reveal that off-target binding has greatly impacted data quality and that DIP may be impracticable for rare modifications. Although we show that these errors can be minimised by the use of IgG controls, future profiling of rare modifications using DIP-seq will require sequencing to near saturation to mitigate the extensive off-target binding. The use of antibodies has had a crucial yet controversial role in biological research but the further development and wider implementation of sensitive non-antibody based techniques will play a crucial role in epigenomic research in the years to come. Until then, our findings outline how and why appropriate controls can be used to minimise errors related to antibodies in DIP-seq.

In summary, this work provides a step towards understanding how dynamic regulation of DNA modifications helps establish lineage-commitment in health and disease as well as contributing to significant improvements in genome-wide methods used to study these modifications.
Acknowledgements

To Colm Nestor, my supervisor and mentor, who has always provided support and given me the freedom to explore and evolve. Everything I am in science I owe to you, thank you for continually believing in me and telling me to stop working and go home every now and then. Hopefully the work of your future students will be able to provide you with one more pool.

My co-supervisors Mikael Benson, who never misses an opportunity to teach me ‘wisdoms’ of life and always pushes me to become a better leader. One day I will provide the perfect elevator pitch you have always requested. Huan Zhang for all the support and always caring about my well-being. 心地感謝您. Mika Gustafsson for always peaking my curiosity by never providing a simple answer to any of my questions.

To the research group, past and present. Cathrine Lagerwall for always providing positive energy and possibly being the only person with organisational skills. Olof Rundquist for all the fun social events and always knowing the chemical reactions we others have no idea about. Maike Bensberg for instantly absorbing knowledge and always asking more questions, keep questioning everything and never doubt yourself. Karolos Douvataniotis for always making work a fun place to be in, one day I hope to finally understand your handwriting. My old student Björn Gylemo, who is both older and stronger than me, keep working hard but do not forget to find the fun and joy in research.

All co-workers who have provided an enjoyable environment for ‘scientific’ discussion. Special thanks to Danuta Gawel for all the interesting discussions related to bioinformatics, statistics and biology. Annelie Lindström for all the fun kayak trips and other events as well as all the help related to the PhD program. Björn Ingelsson for all the help with proteins and always sharing fun personal stories. As well as all the other people, past and present, who have made Cell biology floor 13 a fun place to work in. Extra thanks to the Flow Cytometry Unit and Molecular Biology Unit for excellent training and support.

To past and present collaborators in Helmholtz Zentrum München, MIT, MRC Edinburgh and Gothenburg University for making my visits and stays fun and enjoyable.

Finally, to all good friends and family for always being there when most needed, through good times and bad. Thank you for keeping my head above the water (and always hydrated).
References

32. Fouse, S.D., et al., Promoter CpG methylation contributes to ES cell gene regulation in parallel with 
33. Wang, Z., et al., Combinatorial patterns of histone acetylations and methylations in the human genome. 
34. Fuks, F., et al., The DNA methyltransferases associate with HP1 and the SUV39H1 histone 
35. Stroud, H., et al., 5-Hydroxymethylcytosine is associated with enhancers and gene bodies in human 
36. Williams, K., et al., TET1 and hydroxymethylcytosine in transcription and DNA methylation fidelity. 
37. Wu, H., et al., Dual functions of Tet1 in transcriptional regulation in mouse embryonic stem cells. Nature, 
38. Zhang, X., et al., DNMT3A and TET2 compete and cooperate to repress lineage-specific transcription 
39. Gu, T., et al., DNMT3A and TET1 cooperate to regulate promoter epigenetic landscapes in mouse 
40. Deplus, R., et al., TET2 and TET3 regulate GlcNAcylation and H3K4 methylation through OGT and 
41. Lehner, B., et al., Swi39h-Mediated Histone H3 Lysine 9 Methylation Directs DNA Methylation to 
42. Sardina, J.L., et al., Transcription Factors Drive Tet2-Mediated Enhancer Demethylation to Reprogram 
43. Iqbal, K., et al., Reprogramming of the paternal genome upon fertilization involves genome-wide 
44. Di Giacomo, M., et al., Multiple epigenetic mechanisms and the piRNA pathway enforce LINE1 silencing 


53


Papers

The papers associated with this thesis have been removed for copyright reasons. For more details about these see:

http://urn.kb.se/resolve?urn=urn:nbn:se:liu:diva-155116
Dynamic regulation of DNA methylation in human T-cell biology

Antonio Lentini

Faculty of Medicine and Health Sciences
Linköping University Medical Dissertation No. 1671, 2019
Department of Clinical and Experimental Medicine
Linköping University
SE-581 83 Linköping, Sweden
www.liu.se