UC San Diego

UC San Diego Electronic Theses and Dissertations

Title

Integrative Analysis of 5-Hydroxymethylcytosine Signal in the Context of Gene Regulation

Permalink

https://escholarship.org/uc/item/6ck8h50w

Author

Gonzalez Avalos, Edahi

Publication Date

2022

Peer reviewed|Thesis/dissertation

eScholarship.org Powered by the California Digital Library
University of California

UNIVERSITY OF CALIFORNIA SAN DIEGO

Integrative Analysis of 5-Hydroxymethylcytosine Signal in the Context of Gene Regulation

A dissertation submitted in partial satisfaction of the
requirements for the degree Doctor of Philosophy

in

Bioinformatics and Systems Biology

by

Edahí González Avalos

Committee in charge:

Professor Anjana Rao, Chair
Professor Ferhat Ay, Co-Chair
Professor Eran Mukamel, Co-Chair
Professor Rafael Bejar
Professor Christopher Benner
Professor Olivier Harismendy

2022

Copyright

Edahí González Avalos, 2022

All rights reserved.

The Dissertation of Edahí González Avalos is approved, and it is acceptable in quality and form for publication on microfilm and electronically.

University of California San Diego

2022

DEDICATION

This dissertation is dedicated to **Daniela Samaniego Castruita**, for providing and supporting a life filled with promise and opportunity. I am incredibly blessed to be her husband.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

First, I would like to thank my advisor Anjana Rao for her support over the past seven years. Early in my career, she taught me the value of perseverance, to stand for what I think is the best desition, and that making mistake is part of the process. It was her supportiveness that brought me to be the scientist I am today.

Next, I would also like to thank my co-advisor Ferhat Ay, for he was always confident of the great work I could do, even when I doubted myself and hit low, and advocated and saw to and for me and my future. He saw a spark and nurtured to it to be lit. Always sincere and straight, with a balanced judgement he made the graduating experience a graceful one.

I would also like to thank my dissertation committee for their advices and valuable inputs in my research. Olivier Harismendy has always been an inspiration even before I came here to the US. By showing interest in my abilities, he encouraged me further to continue down the science path, to seek the flame, the opportunity. Chris Benner was key in me forging a stronger scientific personality by providing good guidance as early as in the qualifying exam. Rafael Bejar showed me the value of kind collaboration and openness by being always willing to discuss my project ideas out of the medical field, having this cross-field input made me think critically on overlooked flaws from my project. Eran Mukamel's guidance and eager enthusiasm has proved to be invaluable for the foundations of my project. He was key in questioning the machine learning approaches I planned and included in my project.

I would like to thank all past and current members of Rao Lab for their unyielding support and friendship. In particular, I would like to thank Daniela Samaniego Castruita, Vipul Shukla, Hyungeok Seo, Chan-Wang Jerry Lio, Lukas Chavez, Ageliki Tsagaratou, Atsushi Onodera, Xiaojing Yue, Hugo Sepúlveda, Romain Georges, Chao Yang, Myunggon Ko, Sara Trifari, James

Scott-Browne and Roberta Nowak. They endlessly encouraged me through the process, brainstormed strategies to help solve both my scientific and interpersonal problems.

I would also like to thank Patrick Hogan and all of his Lab's past members for always sharing with me the opportunity to work and discuss amazing and interesting science. In particular, I would like to thank Giuliana Mognol, Xiaocui He, Payal Ramchandani and Aparna Gudlur. The projects we worked together were always interesting and propelled by great discussions lead by Patrick. Patrick has a bright mind able to connect pieces of information so dexterously that it still fascinates me to this day. It was an honor working next to him and his team.

I would like to thank my colleagues and collaborators for all the advises and support. I would like to thank all of the IT core at LJI for answering all my emails and helping me trouble shoot problems on the cluster. Particularly thanks to Michael Talbott, always giving amazing feedback while fixing any issue diligently.

I thank my family. My parents, Silvia Avalos Guarneros and Rafael González Rosales, for their love and support. Without they advocating for me going out of our hometown, to go throught that crazy idea of becoming a scientist, on a city we knew no-one, none of this could have been possible. Their wisdom has helped me through tough times and reminded me to relax, to remember where we all came from, and reconnect to my roots. I would also like to thank my brothers Tonatiúh González Avalos and Rafael González Avalos, for helping me keeping it cool and unwind from time to time. Specially thanks to Tonatiúh for openly sharing his love and care. I thank my extended family: Ana María Castruita Márquez and José Alfredo Samaniego Gaxiola, without them, my Ph.D. could not have been taken into completion and become the beautiful experience it has been (really). ¡Gracias! I would like to thank my wife, Daniela Samaniego Castruita, for always being the voice of reason. She always cheered me up and helped me take

head-grounded desitions, plus celebrated my accomplishments and endured my hysterical moments. Thank you for helping me hold my act together and, first of all, to have an act what to hold. Special thanks to Baby Chami “Firstborn Silky Golden Boy”, and to Cotton Truffle “Secondborn Cotton Silver Girl” for their unconditional love. Saying “I love you” to each and both of you fills me with joy every single time. Thanks to both, my Ana Victoria and my little Amanda, my two happy girls I love to spend my free time with, and the ones I will go anywhere on earth for their sake. They are so bright and wonderful I cannot imagine my life from now on without them. PD: I love you, family.

Chapter 1, in full, is a reformatted reprint with modifications of the material as it appears as "TET enzymes augment activation-induced deaminase (AID) expression via 5-hydroxymethylcytosine modifications at the Aicda superenhancer" in *Science Immunology*, 2019 by Chan-Wang J. Lio, Vipul Shukla, Daniela Samaniego-Castruita, Edahi González-Avalos, Abhijit Chakraborty, Xiaojing Yue, David G. Schatz, Ferhat Ay, and Anjana Rao. The dissertation author was an investigator and co-author of this paper.

Chapter 2 and 3, in full, is a reformatted presentation of the material currently being prepared for submission for publication as “Linking proximal and distal 5hmC enrichment to cell-specific gene regulation with graph convolutional networks” by Edahí González-Avalos, Daniela Samaniego-Castruita, Atsushi Onodera, Xiaojing Yue, Anjana Rao, and Ferhat Ay. The dissertation author was the primary investigator and first author of this material.

VITA

2016 National Autonomous University México
Bachelor of Genomic Sciences.

2022 University of California San Diego
Doctor of Philosophy, Bioinformatics and Systems Biology

PUBLICATIONS

González-Avalos, E., Samaniego-Castruita, D., Rao, A. and Ay, F. (2022) "Linking proximal and distal 5hmC enrichment to cell-specific gene regulation with graph convolutional networks" *In Preparation 2022.*

Shukla, V.*, Samaniego-Castruita, D.*, Dong, Z., **González-Avalos, E.**, Yan, Q., Sarma, K. and Rao, A. (2021). TET deficiency perturbs mature B cell homeostasis and promotes oncogenesis associated with accumulation of G-quadruplex and R-loop structures. Nature Immunology *23*, 99–108.

Yue, X.*, Samaniego-Castruita, D.*, **González-Avalos, E.**, Li, X., Barwick, B.G.[#] and Rao, A.[#] (2021). Whole-genome analysis of TET dioxygenase function in regulatory T cells. EMBO reports *22*, e52716.

Seo, H.*, **González-Avalos, E.***, Zhang, W., Ramchandani, P., Yang, C., Lio, C.-W.J., Rao, A.[#] and Hogan, P.G.[#] (2021). BATF and IRF4 cooperate to counter exhaustion in tumor-infiltrating CAR T cells. Nature Immunology *22*, 983–995.

Onodera, A., **González-Avalos, E.**, Lio, C.-W.J., Georges, R.O., Bellacosa, A., Nakayama, T. and Rao, A. (2021). Roles of TET and TDG in DNA demethylation in proliferating and non-proliferating immune cells. Genome Biology *22*, 186.

Reilly, B.M., Luger, T., Park, S., Lio, C.-W.J., **González-Avalos, E.**, Wheeler, E.C., Lee, M., Williamson, L., Tanaka, T., Diep, D., Zhang, K., Huang, Y., Rao, A. and Bejar, R. (2020). 5-Azacytidine Transiently Restores Dysregulated Erythroid Differentiation Gene Expression in TET2-Deficient Erythroleukemia Cells. Molecular Cancer Research *19*, 451–464.

Seo, H., Chen, J., **González-Avalos, E.**, Samaniego-Castruita, D., Das, A., Wang, Y.H., López-Moyado, I.F., Georges, R.O., Zhang, W., Onodera, A., Wu, C.-J., Lu, L.-F., Hogan, P.G., Bhandoola, A. and Rao, A. (2019). TOX and TOX2 transcription factors cooperate with NR4A transcription factors to impose CD8+ T cell exhaustion. Proceedings of the National Academy of Sciences *116*, 12410–12415.

Mognol, G.P., **González-Avalos, E.**, Ghosh, S., Spreafico, R., Gudlur, A., Rao, A., Damoiseaux, R. and Hogan, P.G. (2019). Targeting the NFAT:AP-1 transcriptional complex on DNA with a small-molecule inhibitor. Proceedings of the National Academy of Sciences *116*, 9959–9968.

Lio, C.-W.J.*, Shukla, V.*, Samaniego-Castruita, D.**, **González-Avalos, E.****, Chakraborty, A., Yue, X., Schatz, D.G., Ay, F. and Rao, A. (2019). TET enzymes augment activation-induced deaminase (AID) expression via 5-hydroxymethylcytosine modifications at the Aicda superenhancer. Science Immunology *4*, eaau7523.

Tsagaratou, A., **González-Avalos, E.**, Rautio, S., Scott-Browne, J.P., Togher, S., Pastor, W.A., Rothenberg, E.V., Chavez, L., Lähdesmäki, H. and Rao, A. (2017). TET proteins regulate the lineage specification and TCR-mediated expansion of iNKT cells. Nature Immunology *18*, 45–53.

Lio, C.-W., Zhang, J., **González-Avalos, E.**, Hogan, P.G., Chang, X.[#] and Rao, A.[#] (2016). Tet2 and Tet3 cooperate with B-lineage transcription factors to regulate DNA modification and chromatin accessibility. eLife *5*, e18290.

An, J., **González-Avalos, E.**, Chawla, A., Jeong, M., López-Moyado, I.F., Li, W., Goodell, M.A., Chavez, L.[#], Ko, M.[#] and Rao, A.[#] (2015). Acute loss of TET function results in aggressive myeloid cancer in mice. Nature Communications *6*, 10071.

*These authors contributed equally to this work. [#]Co-corresponding authors.

ABSTRACT OF THE DISSERTATION

Integrative Analysis of 5-Hydroxymethylcytosine Signal in the Context of Gene Regulation

by

Edahi Gonzalez Avalos

Doctor of Philosophy in Bioinformatics and Systems Biology

University of California San Diego, 2022

Professor Anjana Rao, Chair
Professor Ferhat Ay, Co-Chair
Professor Eran Mukamel, Co-Chair

Many histone marks, obtained through chromatin immunoprecipitation (ChIP) followed by massively parallel DNA sequencing (ChIP-seq) are used as the input features of complex machine learning frameworks in the gene expression prediction task. However, a ChIP-seq assay requires access to a large number of viable cells whose nuclei are intact, a limitation if viable cells are not

available and the only source of cellular material is DNA, or if cells are subjected to processes that compromise their viability, such as formalin fixed paraffin embedding. 5-hydroxymethylcytosine (5hmC) is a stable covalent DNA modification deposited through the Ten-Eleven Translocation (TET) proteins, that is extensively associated to highly expressed genes and lineage-specific enhancers. Thus, as long as some DNA is present in a sample, 5hmC can be assessed and quantified. Through the integration of multi-omic data, we report a close correspondence between 5hmC-marked regions, chromatin accessibility and enhancer activity in B cells. We then produced generalizable machine learning methods to predict gene expression in multiple cell types using 5hmC as a standalone epigenetic feature. Finally, through the integration of 3D genomic structure data, 5hmC signal and complex machine learning frameworks, we predicted gene expression and enhancer-promoter linkages that are cell-type specific. We revealed regions that were orthogonally validated as enhancers in the literature, or had epigenetic characteristics seen in TET-responsive regulatory elements. The analyzes we conducted here highlight the potential of 5hmC signal to predict gene expression and link enhancers to their target genes, and suggest additional approaches for the study of gene regulatory networks.

Introduction

Since the discovery of Ten-Eleven Translocation (TET) proteins by our lab in 2009 (Tahiliani et al. 2009), functional studies have focused on TETs ability to facilitate DNA demethylation and modulate gene expression through oxidation of the methyl group of 5-methylcytosine (5mC). More recently, studies have highlighted the roles of TET proteins and 5hmC in heterochromatin integrity, which if compromised would be deleterious for genome stability and lead cells to oncogenic transformation. Although the mechanism by which TETs influence genome stability are still unclear, a great swath of studies has been done to explore their role in cell differentiation and the effects of their first oxidation product, 5-hydroxymethylcytosine (5hmC). Either single TET knock out (KO) or double or triple KOs (DKO and TKO respectively) generally result in impaired development or halted or incomplete differentiation. An exemplar study involved mouse models developed in our lab, bearing a TET2/3 DKO in developing T cells that resulted in the oligoclonal expansion of iNKT cells (Tsagaratou et al. 2017), followed by the aggressive development of T cell lymphomas (showing DNA damage among other hallmarks of cancer) transmissible in 100% of mice, that succumb in less than 9 weeks (Lopez-Moyado et al. 2019). Another example is the fully-penetrant B cell lymphoma that arises from a TET2/3 DKO in B cells, resulting in a fatal phenotype (most mice died within 5 months; Lio et al. 2016). In both TET2/3 T or B cell cases, individual TET2 or TET3 KO resulted in a less dramatic phenotype, proved in a different model system with a Tet2 gene germline mutation with Tet3 inducibly deleted ($Tet2^{-/-}$ $Tet3^{fl/fl}$; An et al. 2015). In this system, tamoxifen-treated mice (resulting on TET2/3 DKO) showed myeloid expansion, concomitant loss of T, B and erythroid cells, and an aggressive acute myeloid leukemia (mice succumbed within 4–5 weeks of injection). Overall, these studies shown the profound effects TET DKOs have in differentiation and cancer development.

Another characteristic of TET proteins is the genomic distribution of their 5mC-derived enzymatically oxidized products. TET deposition of 5hmC has been found to be strongly enriched in the gene bodies of the most highly expressed genes and in the most active enhancers as defined by the highest enrichment levels of both histone 3 lysine 4 monomethylation (H3K4me1) and lysine 27 acetylation (H3K27Ac) signals (Tsagaratou et al. 2014). This general genomic enrichment is found in euchromatic genomic regions as defined by Hi-C A compartment (principal component (PC) analysis of the Hi-C interaction matrix partitions the genome into A and B compartments; Lieberman-Aiden et al. 2009), and this observation is seen in haematopoietic stem/precursor cells, embryonic stem cells, pro-B cells and natural-killer-T/NKT cells, unlike 5mC that is present in both euchromatin and heterochromatin (Lopez-Moyado et al. 2019).

However, besides multiple associations of 5hmC with genomic activity and correlations of levels of gene expression, to the best of our knowledge no evaluation of 5hmC as an actual predictor of gene expression has been reported in the literature. Before exploring the use of 5hmC as a predictor of gene expression, here we detail the 5hmC genomic deposition dynamics during naïve B cell differentiation (when activated with lipopolysaccharide (LPS) and Interleukin-4 (IL-4)), that led us to observe a strong and gradual 5hmC signal enrichment on enhancers that are components of a superenhancer key for *Aicda* gene expression, required for efficient Class Switch Recombination (CSR) (Lio & Shukla et al. 2019). This observation further supported the idea of linking enhancers (as defined by 5hmC signal) with their target genes.

In this work, we studied 5hmC deposition dynamics throughout the entire B cell differentiation process (5hmC signal enrichment surveyed at 24, 48 and 72 hours after stimulation) that allowed us to pinpoint regions where 5hmC was sharply increased. These “TET-regulated” enhancers tended to bind BATF as an upstream TF required for TET recruitment to these enhancers

and subsequent 5hmC deposition. This published study is followed by a description of our unpublished studies on the analysis of how 5hmC enrichment signal predicts gene expression; first, by the sole use of 5hmC enrichment in and around the genes; and second by the integration of 3D genome structure.

In our study of the differences between activation of wildtype (WT) and TET2/3 DKO B cells, we found that CSR is affected as a result of a reduced expression of the gene *Aicda*, required for proper CSR, through a failed hydroxymethylation of two enhancers (here named TetE1 and TetE2) upstream of the *Aicda*'s TSS. We also found that BATF was required for proper 5hmC deposition at TetE2. In our study of the use of 5hmC enrichment to predict gene expression, we found that this stable DNA modification can be used to generate models that perform as well as, or even better than, state-of-the-art models using multiple histone marks. Finally, we employed Graph Convolution Networks (GCNs) where the graph's nodes are 10 kb windows and edges represent an interaction among the 10 kb windows. Using these GCNs, we integrated the 3D genome organization with 5hmC signal enrichment when solving the gene expression prediction task. Using the GNNExpainer tool in key gene-containing nodes, we gave a relative score to all their interacting windows, where the higher the ranking the more important the connection was in making the gene expression prediction.

References

An, J., González-Avalos, E., Chawla, A., Jeong, M., López-Moyado, I.F., Li, W., Goodell, M.A., Chavez, L., Ko, M. and Rao, A. (2015). Acute loss of TET function results in aggressive myeloid cancer in mice. Nature Communications *6*, 10071.

Lieberman-Aiden, E., van Berkum, N.L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., Amit, I., Lajoie, B.R., Sabo, P.J., Dorschner, M.O., Sandstrom, R., Bernstein, B., Bender, M.A., Groudine, M., Gnirke, A., Stamatoyannopoulos, J., Mirny, L.A., Lander, E.S. and Dekker, J. (2009). Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science *326*, 289–93.

Lio, C.-W.J., Shukla, V., Samaniego-Castruita, D., González-Avalos, E., Chakraborty, A., Yue, X., Schatz, D.G., Ay, F. and Rao, A. (2019). TET enzymes augment activation-induced deaminase (AID) expression via 5-hydroxymethylcytosine modifications at the Aicda superenhancer. Science Immunology *4*, eaau7523.

Lio, C.-W., Zhang, J., González-Avalos, E., Hogan, P.G., Chang, X. and Rao, A. (2016). Tet2 and Tet3 cooperate with B-lineage transcription factors to regulate DNA modification and chromatin accessibility. eLife *5*, e18290.

López-Moyado, I.F., Tsagaratou, A., Yuita, H., Seo, H., Delatte, B., Heinz, S., Benner, C. and Rao, A. (2019). Paradoxical association of TET loss of function with genome-wide DNA hypomethylation. Proceedings of the National Academy of Sciences *116*, 16933–16942.

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

Tsagaratou, A., Aijo, T., Lio, C.-W. J., Yue, X., Huang, Y., Jacobsen, S.E., Lahdesmaki, H. and Rao, A. (2014). Dissecting the dynamic changes of 5-hydroxymethylcytosine in T-cell development and differentiation. Proceedings of the National Academy of Sciences *111*, E3306–E3315.

Tsagaratou, A., González-Avalos, E., Rautio, S., Scott-Browne, J.P., Togher, S., Pastor, W.A., Rothenberg, E.V., Chavez, L., Lähdesmäki, H. and Rao, A. (2017). TET proteins regulate the lineage specification and TCR-mediated expansion of iNKT cells. Nature Immunology *18*, 45–53.

CHAPTER 1: Epigenomics analysis during murine B cell activation.

1.1 Abstract

TET enzymes are dioxygenases that promote DNA demethylation by oxidizing the methyl group of 5-methylcytosine (5mC) to 5-hydroxymethylcytosine (5hmC). Here we report a close correspondence between 5hmC-marked regions, chromatin accessibility and enhancer activity in B cells, and a strong enrichment for consensus binding motifs for basic region-leucine zipper (bZIP) transcription factors at TET-responsive genomic regions. Functionally, Tet2 and Tet3 regulate class switch recombination (CSR) in murine B cells by enhancing expression of *Aicda*, encoding the cytidine deaminase AID essential for CSR. TET enzymes deposit 5hmC, demethylate and maintain chromatin accessibility at two TET-responsive elements, TetE1 and TetE2, located within a superenhancer in the *Aicda* locus. Transcriptional profiling identified BATF as the bZIP transcription factor involved in TET-dependent Aicda expression. 5hmC is not deposited at TetE1 in activated Batf-deficient B cells, indicating that BATF recruits TET proteins to the *Aicda* enhancer. Our data emphasize the importance of TET enzymes for bolstering *AID* expression, and highlight 5hmC as an epigenetic mark that captures enhancer dynamics during cell activation.

1.2 Introduction

TET proteins (Ten-Eleven-Translocation; TET1, TET2, TET3) are Fe(II)- and α-ketoglutarate-dependent dioxygenases that catalyze the step-wise oxidation of 5-methylcytosine (5mC) to 5-hydroxymethylcytosine (5hmC), 5-formylcytosine (5fC) and 5-carboxylcytosine (5caC) (Wu et al. 2017; Tsagaratou et al. 2017). Together these oxidized methylcytosine (oxi-mC) bases are intermediates in DNA demethylation, and may also function as stable epigenetic marks. 5hmC, the most stable and abundant product of TET enzymatic activity, is highly enriched at the most active enhancers and in the gene bodies of the most highly expressed genes, and its presence at enhancers correlates with chromatin accessibility (Tsagaratou et al. 2017; Lio et al. 2016). TET proteins regulate several fundamental biological processes including lineage commitment, and play important roles in embryonic, neuronal and haematopoietic development (Lio et al. 2016).

TET proteins, particularly TET2 and TET3, have critical roles in B cell differentiation and malignancy (Tsagaratou et al. 2017). We and others have previously shown that deletion of the *Tet2* and *Tet3* with *Mb1-Cre* at early stages of mouse B cell development resulted in impaired light chain rearrangement and developmental blockade, and eventually developed an acute precursor-B-cell-derived leukemia with 100% penetrance (Scott-Browne et al. 2017; Orlanski et al. 2016). Inducible deletion of *Tet1* and *Tet2* using *Mx1-Cre* promoted the development of acute lymphoblastic leukemia derived from precursor B cells, and global loss of *Tet1* caused B cell lymphomas with an extended latency (Quivoron et al. 2011). In humans, *TET2* mutations are frequently observed in Diffuse Large B Cell Lymphoma (DLBCL), a malignancy derived from germinal center (GC) B cells (Schmitz et al. 2018; Reddy et al. 2017), suggesting that TET proteins may regulate mature B cell function. However, due to the pleiotropic functions of TET proteins,

studies of TET-mediated gene regulation are best performed in systems where TET genes are deleted acutely rather than during development.

After their development in the bone marrow, mature B cells migrate to peripheral lymphoid tissues where they encounter antigen and follicular T helper cells in GC, and participate in the generation of functional immune responses (De Silva et al. 2015). In GC, B cells diversify the variable regions of immunoglobulin (Ig) chains in a process known as somatic hypermutation (SHM) and also undergo Class Switch Recombination (CSR) to replace the constant region of immunoglobulin M (IgM) to other isotypes (IgG1, IgA, etc.). Both CSR and SHM are both orchestrated by the enzyme AID (Activation-induced cytidine deaminase, encoded by *Aicda*) (Chandra et al. 2015; Vaidynathan et al. 2015; Muramatsu et al. 2000). AID promotes CSR and SHM by generating DNA double-strand breaks at Ig switch regions and point mutations at Ig variable regions, respectively (Alt et al. 2013). Due to its high mutagenic potential (Casellas et al. 2016; Robbiani et al. 2013), AID expression is normally restricted to activated B cells and is tightly regulated.

Here we investigated the role of TET proteins during mouse B cell activation by mapping 5hmC distribution genome-wide and integrating the data with previous studies of transcriptional and epigenetic changes during B cell activation (Kieffer-Kwon et al. 2017; Kieffer-Kwon et al. 2013). We deleted the Tet2 and Tet3 genes acutely using CreERT2 to avoid secondary effects caused by prolonged TET deficiency during differentiation. We show that TET2 and TET3 regulate CSR by controlling the activation-induced up-regulation of AID mRNA and protein and that they act downstream of the basic region/leucine zipper (bZIP) transcription factor BATF (basic leucine zipper transcription factor, ATF-like), which is induced during B cell activation with more rapid kinetics than Aicda and binds concomitantly with TET proteins to two TET-responsive

elements in the Aicda locus, TetE1 and TetE2. Our study demonstrates the role of TET proteins in CSR in activated B cells and provides a detailed description of the general mechanism whereby TET proteins influence cell activation and differentiation.

1.3 Results

B cell activation promotes genome-wide deposition of 5hmC. We immunoprecipitated genomic DNA with antibodies to cytosine-5-methylenesulfonate (CMS-IP) (Huang, et al. 2012; Pastor et al. 2011) to analyze the kinetics of genome-wide 5hmC distribution in murine B cells activated with lipopolysaccharide (LPS) and interleukin-4 (IL-4), a well-characterized *in vitro* system for studying CSR (**Fig. 1.1A**). The vast majority of 5hmC-marked regions (~160,000) were shared between pre- and post-activated B cells (**Fig. 1.1B** and **Fig. S1.1A**); of the ~9,500 differentially hydroxymethylated regions (DhmRs) in 72h-activated versus naïve B cells, the majority (8,454) showed increased 5hmC ($DhmR^{72h\text{-}up}$) whereas a much smaller fraction showed a decrease ($DhmR^{down}$) (**Fig. 1.1B,C**). DhmRs were typically located more than 10 kb from the closest transcription start site (TSS) (**Fig. S1.1B**), and their 5hmC levels progressively changed with time after activation ($DhmR^{72h\text{-}up}$; **Fig. 1.1D**; $DhmR^{72h\text{-}down}$; **Fig. S1.1C**).

The oxidized methylcytosines produced by TET proteins are known intermediates in DNA demethylation (Tsagaratou et al. 2017; Pastor et al. 2013). To relate 5hmC to changes in DNA methylation, we compared 5hmC distribution in naïve and 72h-activated B cells with published whole-genome bisulfite sequencing (WGBS) data on B cells activated for 48h under similar conditions (Kieffer-Kwon et al. 2013). Although WGBS cannot distinguish 5mC and 5hmC (Huang et al. 2010), 5hmC is typically a small fraction (1-10%) of 5mC (Yue et al. 2016), thus we refer to the WGBS signal as "DNA methylation" here. Most (1097 of 1168, 94%) differentially methylated regions (DMRs) were demethylated at 48 hours and marked by 5hmC (**Fig. S1.1D,E**). Regions that showed increased 5hmC at 24, 48, and 72 hours ($DhmR^{24h\text{-}up}$, $DhmR^{48h\text{-}up}$, and $DhmR^{72h\text{-}up}$) also showed decreased DNA methylation (**Fig. 1.1E** and **Fig. S1.1F**). Overall, regions of 5hmC deposition correspond to regions of DNA hypomethylation during B cell activation, as expected

from the well-established role of 5hmC as an intermediate in DNA demethylation. Motif enrichment analysis of the 8454 $DhmR^{72h\text{-}up}$ and 1097 $DMR^{48h\text{-}down}$ regions showed that both were enriched in consensus binding sequences for transcription factors of the nuclear factor kB (NF-kB) (Rel homology domain) and bZIP families, as well as for "composite" IRF:bZIP motifs (**Fig. 1.1F** and **Fig. S1.1G**) (Li et al. 2012; Glasmacher et al. 2012; Murphy et al. 2013).

To discern the relationship between 5hmC and enhancers, naïve and activated B cell enhancers, defined by H3K4 monomethylation (H3K4me1), were stratified based on the level of H3K27 acetylation (H3K27Ac), a modification that tracks with enhancer activity (Calo et al. 2013). In both sets of enhancers, 5hmC was most highly enriched at active ($H3K4me1^{+}$ $H3K27Ac^{+}$) relative to poised ($H3K4me1^{+}$ $H3K27Ac^{-}$) enhancers (**Fig. 1.1G**). Moreover, more than 75% of previously identified superenhancers in activated B cells, defined by H3K27Ac, overlapped with at least one $DhmR^{72h\text{-}up}$ region (**Fig. 1.1H**) (Meng et al. 2014). As an example, a 3' distal element at the *Ccr4* locus showed activation-dependent gain of 5hmC and H3K27Ac, associated with concomitant loss of methylation at specific CpGs and increased mRNA expression (**Fig. S1.1I,J**). 5hmC was also associated with accessible chromatin defined by ATAC-seq (assay for transposase-accesible chromatin using sequencing) (Scott-Browne et al. 2017; Orlanski et al. 2016; Tsagaratou et al. 2017), and kinetic analysis of active enhancers, defined as differentially active between naïve and 48h-activated B cells by high accessibility and high H3K27Ac, showed that 5hmC level positively correlated with enhancer activity (**Fig. S1.1H**). Together our data show that 5hmC modification and DNA demethylation correlates with enhancer activity during B cell activation.

Comparison of WT and *Tet2/3* DKO B cells identifies TET-responsive regulatory elements. *Tet2* and *Tet3* are the two major TET proteins expressed in B cells (**Fig. 1.2A**). To evaluate the role of TET proteins in regulating B cell function, we generated mice conditionally

deficient in *Tet2* and *Tet3*, using Cre^{ERT2} and further introduced a *Rosa26-LSL-YFP* cassette to monitor Cre recombinase activity after tamoxifen treatment (LSL: *LoxP*-STOP-*LoxP* cassette in which a strong transcriptional stop is flanked by LoxP sites). Cre^{ERT2} $Tet2^{fl/fl}$ $Tet3^{fl/fl}$ *Rosa26-LSL-YFP* (DKO) and control $Tet2^{fl/fl}$ $Tet3^{fl/fl}$ *Rosa26-LSL-YFP* (WT) mice were treated for 5 days with tamoxifen, after which WT and *Tet2/3* DKO B cells were isolated and activated with LPS and IL-4 (**Fig. 1.2B**). Both *Tet2* and *Tet3* were efficiently deleted in B cells (**Fig. 1.2C**), and the YFP^+ cells showed similar frequencies of mature splenic follicular B cells (**Fig. S1.2A,B**).

Global 5hmC levels assessed by DNA dot blot were similar in WT and *Tet2/3* DKO B cells prior to activation but showed a perceptible decrease by 48h after activation, this may partially be explained due to 5hmC being passively lost as a function of cell division (Tsagaratou et al. 2017; Pastor et al. 2013) (**Fig. S1.2C**). Starting at 48h, 5hmC levels were substantially lower in *Tet2/3* DKO compared to WT B cells, indicating that Tet2 and Tet3 actively oxidize 5mC to 5hmC during B cell activation (**Fig. S1.2C**). Around 2,300 5hmC-enriched regions were significantly different between WT and DKO at 72h of activation, with substantially more regions gaining 5hmC in control compared to *Tet2/3* DKO B cells at each time point examined (**Fig. 1.2D,E**); most were located > 10kb from the nearest TSS (**Fig. S1.2D**). Of 2,139 "TET-regulated" DhmRs with higher 5hmC in WT compared to *Tet2/3* DKO B cells ("WT>DKO DhmR"), 2020 (94.4%) significantly overlapped with $DhmR^{72\text{-up}}$ regions; with decreased DNA methylation at their centers (**Fig. S1.2E**), and were enriched for RHD, bZIP and composite IRF:bZIP motifs (**Fig. 1.2F**).

Tet2 and Tet3 regulate Ig CSR. To assess the effect of *Tet2/3* deletion on the antibody response *in vivo*, we treated $Tet2^{fl/fl}$ $Tet3^{fl/fl}$ *Rosa26-LSL-YFP* Cre^{ERT2} and control $Tet2^{fl/fl}$ $Tet3^{fl/fl}$ *Rosa26-LSL-YFP* mice for 5 days with tamoxifen, followed by immunization with NP-OVA in the footpads two days later **(Fig. 1.3A)**. Acute deletion of *Tet2/3* resulted in increased numbers of total

cells and B cells in draining popliteal lymph nodes by day 7 after immunization (**Fig. S1.3A**), consistent with our previous observations that *Tet2/3* deficiency results in increased cell survival and/or proliferation (Tsagaratou et al. 2017). The overall percentage of germinal center (GC) B cells ($CD19^+$ $GL7^+$ Fas^+) was similar between WT and DKO (**Fig. 1.3B,C**), but there was a significant increase in the frequency of NP-specific B cells (**Fig. 1.3D**). There was also a moderate (~25%) decrease in GL7 MFI in *Tet2/3 DKO* B cells compared with WT control GC B cells (**Fig. S1.3B**), indicating that although TET proteins are, in general, important for cell differentiation, acute deletion of *Tet2/3* in B cells had only a limited effect on GC B cell differentiation. Because acute Tet2/3 deletion using the CreERT2 system affects multiple cell types, we did not investigate the in vivo B cell phenotype further in Tet2fl /f l Tet3f l /fl CreERT2 Rosa26 YFPLSL mice. The most notable phenotype in these mice, however, was the consistent decrease in CSR from IgM to IgG1 (**Fig. 1.3E,F**), demonstrating a role for TET proteins in regulating antibody responses in vivo, particularly the CSR.

To determine if the CSR phenotype was B-cell-intrinsic, B cells from tamoxifen-treated mice were labeled with proliferation dye (Cell-trace Violet) and activated with LPS and IL-4 for 4 days (**Fig. 1.3G**). Consistent with the CSR defect *in vivo*, we noticed a consistent decrease in IgG1 switching in *Tet2/3* DKO B cells activated *in vitro* relative to WT B cells (**Fig. 1.3H,I**). The impaired CSR in *Tet2/3* DKO is not due to Cre activity, as similar result was observed when Cre^{ERT2} $Tet2^{+/+}$ $Tet3^{+/+}$ *Rosa26-LSL-YFP* was used as control (**Fig. S1.3C**). The defect in CSR was cell-intrinsic, since it was also apparent when congenically-marked WT (CD45.1) and *Tet2/3* DKO (CD45.2) B cells were mixed and co-cultured (**Fig. S1.3D**). The difference was not due to altered proliferation, which was comparable between WT and *Tet2/3* DKO B cells (**Fig. 1.3H** and **Fig. S1.3E**). Correlating with the decrease in CSR from IgM to IgG1, the expression of circular

γ1 transcript was decreased in *Tet2/3* DKO B cells (**Fig. 1.3J**). Further, CSR to IgA was also decreased in *Tet2/3* DKO relative to WT B cells activated with anti-CD40, IL-4, IL-5, and TGFβ (**Fig. 1.3K-N**). The loss of Tet2/3 also resulted in a decrease in the differentiation of CD138+ plasma blasts/cells after in vitro activation (**Fig. S1.3F**). Reconstitution of *Tet2/3* DKO B cells with the enzymatically active catalytic domain of TET2 (Tet2CD) (Ko et al. 2013) restored CSR almost to control levels (**Fig. S1.3, H-I**), whereas an enzymatically inactive mutant of Tet2CD ($Tet2CD^{HxD}$) was ineffective (**Fig. S1.3G-I**) (Zhang et al. 2015). Together, these results indicate that *Tet2* and *Tet3* are required for optimal CSR both *in vitro* and *in vivo*.

Because the CSR defect in *Tet2/3* DKO B cells was ~50% of control, we asked whether deletion of all three TET proteins might have a more striking effect. Consistent with the very low expression of Tet1 in mature B cells (**Fig. 1.2A**), the CSR in *Tet1/2/3 TKO* was comparable to that observed in *Tet2/3* DKO mice (**Fig. S1.3J,K**). These results indicate that Tet2 and Tet3 are the major TET proteins that regulate CSR in B cells.

Tet2 and Tet3 regulate expression of the cytidine deaminase AID. CSR is a highly regulated process and involves multiple pathways, including cytokine signaling and DNA repair (Methot et al. 2017). RNA-seq analysis identified a relatively small number of genes differentially expressed between WT and *Tet2/3* DKO B cells under resting conditions and at different time points after activation (**Fig. S1.4A,B**); among these was *Aicda*, which encodes AID, the activation-induced cytidine deaminase essential for CSR. qRT-PCR analysis confirmed a >50% decrease in *Aicda* mRNA expression in *Tet2/3* DKO relative to WT B cells at each time point from 48 to 96 hours post-activation (**Fig. 1.4A-D, Fig. S1.4C,D**), a phenotype reminiscent to the dampened CSR in the case of AID haploinsufficiency (Sernandez et al. 2008; Takizawa et al. 2008) (**Fig. S1.4E-J**). Although Tet2 mRNA expression showed only minor changes in unstimulated versus

stimulated B cells (**Fig. 1.2A**), Tet2 protein expression was low in unstimulated and 24-hour-stimulated B cells, with increased expression observed at 48 hours after stimulation (**Fig. S1.4D**). The late TET2 induction parallels the late kinetics of increase in 5hmC (**Fig. 1.1C**).

To determine whether the decrease in AID expression was fully responsible for the CSR defect, we expressed WT and catalytically inactive AID ($AIDH^{56R/E58Q}$) (Papavasiliou et al. 2002) in WT and Tet2/3 DKO B cells via retroviral transduction. Retroviral expression of catalytically active AID (AIDWT), but not inactive $AIDH^{56R/E58Q}$, largely rescued the CSR defect in Tet2/3 DKO B cells (**Fig. 1.4C,D**, bottom panels). Similar to previous observations (Muramatsu et al. 2000), expression of AID in WT cells also increased the frequency of IgG1+ cells (**Fig. 1.4C,D**, top left and middle panels). Because Tet2/3 were not required for expression of the germline transcripts essential for CSR (**Fig. S1.4K**), the bulk of the CSR defect in Tet2/3 DKO B cells can be attributed to the decrease in expression of Aicda mRNA and AID protein, leading us to test the hypothesis that TET proteins control Aicda expression through distal regulatory element(s) of the Aicda gene cells (**Fig. 1.4C,D**, top left and middle panels). Despite their importance in *Aicda* expression, Tet2/3 were not required for the expression of μ and γ1 germline transcripts that are essential for CSR (**Fig. S1.4J**). These data suggest that the bulk of the CSR defect in *Tet2/3* DKO B cells can be attributed to the decrease in expression of *Aicda* mRNA and AID protein, leading us to test the hypothesis that TET proteins control *Aicda* expression through distal regulatory element(s) of the *Aicda* gene.

Genome-wide analyses identify TET-responsive regulatory elements in the Aicda locus. Multiple conserved regulatory elements influence Aicda expression (**Fig. S1.5A**), and their deletion markedly decreased Aicda expression in activated B cells (Kieffer-Kwon et al. 2013; Crouch et al. 2007; Huong Ie et al. 2013; Tran et al. 2010). Of these, the Aicda 5′ enhancer at −26

kb in the Mfap5 gene, the intergenic 5′ enhancers, and the intron 1 enhancer noticeably gain H3K27Ac and lose 5mC upon activation and have been collectively termed the Aicda "superenhancer" (**Fig. 1.5A,** middle and bottom tracks) (Kieffer-Kwon et al. 2013; Meng et al. 2014).

Chromatin immunoprecipitation sequencing (ChIP-seq) for TET2 showed that each of these regulatory elements was occupied by TET2 in 72-hour activated B cells (**Fig. 1.5A,** top two tracks). Among these, the −26-kb Mfap5 intronic region and the −8-kb 5′ intergenic region (here termed TetE2 and TetE1, respectively) were TET-regulated: B cell activation induced a TET-dependent increase in 5hmC (**Fig. 1.5B**), placing them in the category of WT > DKO DhmRs (**Fig. 1.2D,E**). TetE1 appears to be the prime target for TET2/3 due to its larger gain of 5hmC after activation (**Fig. 1.5B**). In contrast, regions such as the Aicda promoter were marked by 5hmC even before activation, indicating that the 5hmC at these regions was likely generated during a previous stage of B cell differentiation and then maintained until the emergence of naïve peripheral B cells.

To confirm the importance of *TetE1* in *Aicda* regulation, we deleted the enhancer using CRISPR in CH12F3 cells, a B cell line that can class-switch from IgM to IgA upon activation with anti-CD40/IL-4/TGFβ (**Fig. S1.5B,C**). We tested four clones with homozygous deletions; all showed decreased expression of *Aicda* mRNA, and in three of these, there was almost no detectable CSR (**Fig. S1.5D,E**), confirming a previous report in the context of a BAC transgene that *TetE1* was essential for *Aicda* expression (Huong Ie et al. 2013). B cell activation also induced hydroxymethylation at the IgH locus, most notably upstream of the IgG1 promoter (**Fig. S1.5F**). Given that Tet2/3 DKO B cells expressed similar levels of IgG1 germline transcripts (**Fig. S1.4K**) and that ectopic expression of AID could rescue the impairment of IgG1 CSR, the significance of TET-mediated DNA modification/demethylation at the IgH locus remains to be determined.

B cell activation induces strong H3K27Ac and DNA demethylation at TetE1 (**Fig. 1.5A**). Because bisulfite sequencing does not distinguish 5mC from 5hmC, we used oxidative bisulfite sequencing (oxBS-seq) to assess the levels of 5mC, 5hmC, and unmodified C at TetE1 in WT and Tet2/3 DKO cells [neither BS-seq nor oxBS-seq distinguish unmodified C from 5fC and 5caC, but these modified bases are ~10-fold and ~100-fold less abundant than 5hmC (Tsagaratou et al. 2017)]. CpGs in both TetE1 and the Aicda promoter displayed similar levels of 5mC and 5hmC before activation (**Fig. 1.5C** and **Fig. S1.5H**; compare 0-hour panels). At 72 hours after activation, there was a substantial decrease in 5mC in WT B cells; in contrast, both TetE1 and the Aicda promoter were methylated in Tet2/3 DKO B cells (**Fig. 1.5C,** bottom panel; compare 72-hour panels). These results indicate that TET2 and TET3 regulate Aicda expression by binding to and depositing 5hmC at TetE1 and TetE2.

TET2 and TET3 maintain chromatin accessibility at two Aicda TET-responsive elements, TetE1 and TetE2. Active regulatory regions are typically found in accessible regions of chromatin (Kundaje et al. 2015) and are marked by 5hmC (Scott-Browne et al. 2017; Tsagaratou et al. 2017). To assess the dynamics of chromatin accessibility, we performed ATAC-seq in B cells stimulated with LPS and IL-4. Activated B cells displayed progressive chromatin remodeling (**Fig. S1.6A**). Regions with increased 5hmC after activation ($DhmR^{72h\text{-}up}$) also showed increased chromatin accessibility after activation and vice versa (**Fig. S1.7A**; blue box-and-whisker plots). To understand the relationship between TET function and chromatin accessibility, we performed ATAC-seq on WT and Tet2/3 DKO B cells activated as in **Fig. 1.3G**. Of a total of ~28,000 accessible regions (**Fig. S1.6B,C**), only a minor fraction (~1.5%; 421 of 28,137) showed significant changes in accessibility between WT and Tet2/3 DKO B cells, and the differences were observed late, at 72 hours after activation (**Fig. S1.6B,D**). Of the 292 potentially TET-regulated

differentially accessible regions (DARs), defined as showing decreased accessibility in Tet2/3 DKO compared with WT B cells (WT > DKO DARs), the majority were located distal to the TSS (**Fig. S1.7B**) and a significant proportion of these (110 of 292, 37.7%) showed a TET2/3-dependent increase in 5hmC at 72 hours compared with unstimulated cells ($DhmR^{72up}$) (**Fig. S1.6C** and **Fig. S1.7C**, top). In contrast, the 129 DKO > WT DARs that were less accessible in WT compared with Tet2/3 DKO B cells and the 27,716 commonly accessible DARs were present in both TSS-proximal and TSS-distal regions and did not show significant changes in 5hmC (**Fig. S1.6C**, middle and bottom panels, and **Fig. S1.7B,C**). Analysis of DNA methylation at 48 hours after activation showed that WT > DKO DARs were further demethylated after activation, whereas DKO > WT DARs were already substantially demethylated in naïve B cells and showed no further changes after activation (**Fig. S1.7D**). Moreover, the WT > DKO DARs were enriched in bZIP and BATF:IRF motifs (**Fig. S1.7E**), similarly to those in $DhmR^{72h\text{-}up}$ (**Fig. 1.1F**) and WT > DKO DhmRs (**Fig. 1.2F**). Together, these data support our previously observed correlation of 5hmC modification with changes in chromatin accessibility.

Focusing on the *Aicda* locus, we found that activation was associated with increased accessibility at the *Aicda* enhancers *TetE1* and *TetE2* (**Fig. S1.6D**). The 5hmC modification continuously increased at these two elements until 72h, with a higher level of deposition at *TetE1* (see **Fig. 1.5B**). In contrast, the time course of increase in chromatin accessibility was quite different at the two enhancers (**Fig. S1.6D**): *TetE2* showed a rapid increase in accessibility apparent in both WT and *Tet2/3* DKO B cells at 24 h following activation, whereas the time course of increase in *TetE1* accessibility was slower, matching that of 5hmC deposition (compare **Fig. 1.5B** and **Fig. S1.6D**). Consistent with the increased accessibility, several chromatin remodelers and histone acetyl-transferases including Brg1, Chd4, p300, and to a lesser extent, Gcn5, were

recruited to *TetE1* and *TetE2* in 24h activated B cells (**Fig. S1.7F**). Interestingly, we noticed a slight decrease in chromatin accessibility at *TetE1* and *TetE2* in *Tet2/3* DKO B cells compared with WT cells at 72 hours, suggesting that TET proteins are important for maintaining the accessibility at these enhancers (**Fig. S1.6E**). Loss of TET proteins had no significant effect on chromatin accessibility at the IgH locus (**Fig. S1.5G**). Together, these data point to a consistent link between bZIP-family transcription factors, TET catalytic activity, and chromatin accessibility.

Batf acts upstream of TET at *Aicda* enhancers. Before enhancers are established during development, cell lineage specification, or activation, certain key transcription factors bind to nucleosome-associated regions and recruit chromatin remodeling complexes and histone-modifying enzymes to create active enhancers (Calo et al. 2013). To identify potential pioneer transcription factors for the Aicda locus, we took advantage of our previous motif enrichment analyses (**Fig. 1.1F**, **Fig. 1.2F** and **Fig. S1.1G**). We had observed strong enrichment for consensus binding motifs for bZIP transcription factors, at regions that progressively gained 5hmC as a function of activation (**Fig. 1.1F**, DhmRup), regions that lost DNA methylation upon activation (**Fig. S1.1G**, $DMR^{48h\text{-}down}$), regions with higher 5hmC in WT compared with Tet2/3 DKO B cells (**Fig. 1.2.F**, WT > DKO DhmRs), and regions with higher accessibility in WT versus Tet2/3 DKO B cells (**Fig. S1.7E**, DAR^{72h} WT > DKO).

On the basis of these data, we focused on bZIP transcription factors expressed in activated B cells. Consistent with previous observations (Ise et al. 2011; Betz et al. 2010), Batf mRNA and protein were induced after activation (**Fig. 1.6A** and **Fig. S1.8B**) and their expression preceded that of Aicda, as expected if BATF regulated Aicda mRNA induction (**Fig. S1.8A,B**, and **Fig. S1.4D**). In contrast, expression of Bach1 and AP-1 (Fos and Jun) family members was either low throughout (Fosl1, Fosl2, JunD, and Bach1) or moderate to high in unstimulated B cells,

potentially because these cells contained a minor population of memory B cells that express high levels of Fos and Jun (**Fig. S1.8C-E**; and Immgen.org) (Heng et al. 2008). Given the kinetics, we examined the importance of BATF in subsequent experiments.

BATF is essential for T and B cells during humoral responses (Ise et al. 2011; Betz et al. 2010), and Batf-KO B cells are defective in CSR (**Fig. S1.8F**). Genome-wide analysis of Batf binding by ChIP-seq in 72-houractivated WT and Tet2/3 DKO B cells showed very few overall differences (**Fig. S1.8G**), indicating that BATF functioned upstream or independently of TET enzymes. Nevertheless, one of two distinguishable sets of BATF ChIP-seq peaks (cluster 2 in **Fig. 1.6B**) was TETregulated, because the peaks in this cluster showed a progressive TET2/3-dependent increase in 5hmC after activation (**Fig. 1.6B,C**). In contrast, BATF peaks in cluster 1 showed no significant activationdependent increase in 5hmC (**Fig. 1.6B,** top panel). Overall, about onethird of regions with activation-induced 5hmC ($DhmR^{72\text{-}up}$) overlapped with BATF peaks (**Fig. S1.8H**), indicating that in addition to BATF, other transcription factors also have a role in facilitating TET-mediated 5hmC generation at the Aicda locus. Despite this strong functional interaction, we did not observe a substantial direct protein-protein interaction between BATF and TET2 in coimmunoprecipitation experiments in which the effects of nucleic acids were excluded (**Fig. S1.8I**), suggesting that the interaction is dependent on additional factors.

BATF bound strongly at the TetE1 and TetE2 enhancers in the Aicda locus and, to a lesser extent, to the −21-kb intergenic enhancer located between TetE1 and TetE2 (**Fig. 1.6D**, 72-hour WT and DKO; second and third tracks). Consistent with the lack of BATF expression in unstimulated cells, there was no enrichment of BATF occupancy at the Aicda enhancers at 0 hours (**Fig. 1.6D**, 0-hour WT; top track). This binding pattern resembles that of TET2 (**Fig. 1.5A**), as well as that of E2A and PU.1 (**Fig. 1.6D**) (Willis et al. 2017; Wöhner et al. 2016; Gloury et al.

2016). Moreover, BATF and JUNB associated with TETE1 and TETE2 in a human B cell lymphoblast (**Fig. S1.8L**), suggesting that the binding of BATF to these enhancers is evolutionarily conserved. To determine whether BATF acted upstream of TET, we analyzed 5hmC deposition at the TET-responsive element TetE1 in WT and Batf-deficient B cells. We found unambiguously that the absence of BATF abolished activation-induced hydroxymethylation at TetE1 (**Fig. 1.6E**). Our results are consistent with the hypothesis that BATF facilitates the recruitment of TET2 and/or TET3 to TetE1 and TetE2 and increases Aicda expression by promoting 5hmC modification and DNA demethylation at these upstream Aicda enhancers.

As mentioned above, BATF is essential for Aicda regulation. However, we cannot rule out the involvement of additional transcription factors, including other bZIP family members, in this process. Although IRF4 is required for Aicda expression and binds to TetE1 (Klein et al. 2006; Sciammas et al. 2006), 5hmC levels at TetE1 were unaffected in Irf4-deficient B cells after LPS/IL-4 stimulation (**Fig. S1.8M**). Thus, depending on cell type and conditions of stimulation, certain transcription factors preferentially function together with TET proteins, whereas others could be responsible for additional aspects of locus remodeling and gene expression.

1.4 Discussion

TET proteins (TET1, TET2 and TET3) oxidize 5mC to 5hmC, a stable epigenetic mark that is the most abundant of the three oxi-mC intermediates for DNA demethylation. Due to the pleiotropic effects of TET proteins in cells, it has been challenging to address the specific roles of TET proteins in mice with prolonged TET deficiency. Here, to circumvent this issue, we used the inducible tamoxifen-CreERT2 system to delete Tet2 and Tet3 in mature B cells, a well-established system for the molecular analysis of gene regulation during cell activation. Our data show clearly that Tet2 and Tet3 – the major TET proteins expressed in B cells – are required for efficient class switch recombination (CSR) both in vivo and in cultured cells. A primary mechanism involves TET-mediated regulation of the expression of Aicda, the essential DNA cytosine deaminase for CSR. BATF, potentially with other transcription factors, helps recruit TET proteins to two major TET-responsive regulatory elements that we have newly defined in the Aicda locus, TetE1 and TetE2. TET2 and TET3 convert 5mC to 5hmC at these regulatory elements, leading to DNA demethylation, sustaining enhancer accessibility and augmenting Aicda expression.

The biological consequences of TET loss-of-function are determined by several factors: the time course of Tet2 and Tet3 gene deletion, the stability of Tet2 and Tet3 mRNA and protein, and the rate of cell division which determines the rate of passive (i.e. replication-dependent) dilution of 5hmC. At each cell division, hemi-methylated CpGs are recognised by the maintenance UHRF1/DNMT1 DNA methyltransferase complex and converted back to symmetrically methylated CpGs, whereas hemi-hydroxymethylated CpGs are ignored and so are diluted by half (Wu et al. 2017; Tsagaratou et al. 2017). Consequently, 5hmC is present at comparable levels in quiescent (non-dividing) WT and Tet2/3 DKO B cells, thus enabling us to study the effects of acute TET deletion in activated, proliferating B cells. The progressive replication-dependent loss

of 5hmC and consequent dilution of both 5mC and 5hmC is likely to be required for optimal gene expression, explaining the long-standing observation that the induction of Aicda expression during B cell activation, and the induction of cytokine genes during Th2 differentiation, are both tightly coupled to cell division (Rush et al. 2005; Bird et al. 1998).

An optimal level of AID is crucial to maintain the necessary balance between effective antibody immune responses and unintentional C > T mutations caused by AID-mediated DNA cytidine deamination. Although Aicda haploinsufficiency results in dampened antibody responses (**Fig. S1.4, E to J**) (Sernandez et al. 2008; Takizawa et al. 2008), uncontrolled AICDA expression is associated with B cell malignancies (Compagno et al. 2017). Thus, the level and activity of AID are meticulously controlled by diverse mechanisms including a tight transcriptional regulatory program (Zan et al. 2015). At the Aicda locus, at least six regulatory elements have been identified (Fig. S1.5A); five of them, located at distances ranging from −29 to +5 kb relative to the TSS, are collectively termed the Aicda superenhancer (Kieffer-Kwon et al. 2013; Meng et al. 2014). The enhancers at −26, −21, −8, and +13 kb are all necessary for inducing Aicda expression in activated B cells, on the basis of deletion of individual enhancers in mice and the CH12 B cell line (Huong Ie et al. 2013; Tran et al. 2010). Even in naïve B cells where Aicda is not expressed, the Aicda promoter is already highly enriched in 5hmC and the −21-, intronic, and +13-kb Aicda enhancers display 5hmC and H3K27Ac (**Fig. S1.5A**). The 5hmC modification at the −26-, −21-, and +13-kb Aicda enhancers is apparent as early as the pro–B cell stage of B cell development (Scott-Browne et al. 2017), suggesting that TET-mediated 5hmC modification acts to "bookmark" regulatory elements necessary for proper gene expression in progeny cells after activation.

The vast majority of 5hmC-marked regions are present in common between naïve and activated mature B cells (this study) and between WT and TET-deficient invariant natural killer T

cells (iNKT) cells (Tsagaratou et al. 2017), supporting the hypothesis that most 5hmC-marked regions in any given cell type were laid down during previous developmental stages and thus are constitutively modified. In contrast, activationinduced 5hmC modification occurs at only a few distal elements in B cells (Fig. 1C), and 5hmC levels at these elements correlate strongly with activation-induced increases in enhancer activity defined by H3K27Ac (**Fig. 1.1G**) (Tsagaratou et al. 2014). Moreover, the majority of previously described B cell superenhancers (Kieffer-Kwon et al. 2013; Meng et al. 2014) harbor at least one activationinduced 5hmC-modified regulatory element (**Fig. 1.1H**). In the particular case of Aicda, we identified activation-induced 5hmC modification at two major TET-responsive elements, TetE1 and TetE2, both part of a superenhancer cluster located 5′ of the Aicda gene (**Fig. 1.5A**) (Kieffer-Kwon et al. 2013; Meng et al. 2014). 5hmC modification at these elements was apparent by 48 hours (**Fig. 1.5B**), preceding the marked up-regulation of Aicda mRNA at 72 hours (**Fig. S1.4C**). Tet2/3 deficiency almost eliminated the activationinduced 5hmC modification at both enhancers and resulted in diminished expression of both Aicda mRNA and AID protein (**Fig. 1.4A,B**, **Fig. 1.5B** and **Fig. S1.4C**). Thus, our in vitro data indicate that TET proteins and 5hmC are important for Aicda expression by enabling the TetE1 and TetE2 enhancers to function at full capacity. However, we cannot rule out that in addition to decreasing AID expression, TET deficiency affects CSR indirectly by impeding B cell differentiation in vivo.

Studies from our lab and others have implicated TET proteins and 5hmC in regulating chromatin accessibility. For instance, TET proteins were shown to be required for demethylation of evolutionarily conserved enhancers during zebrafish development, and morpholino-mediated knockdown of Tet1/2/3 resulted in decreased enhancer accessibility (Rush et al. 2005). In mammals, we have shown that Tet2/3-deficiency results in lower accessibility of enhancers during

T and B cell development (Scott-Browne et al. 2017; Papavasiliou et al. 2004). However, these steady state studies provide limited mechanistic insights. Here, through systematic analyses of 5hmC modification and chromatin accessibility kinetics during B cell activation, we show that 5hmC displays a time-dependent increase at regions that are differentially accessible between WT and Tet2/3 DKO B cells during; moreover, TET proteins are important for sustaining enhancer accessibility (**Fig. 1.6E; Fig. S1.6E**). We speculate that enhancer methylation limits enhancer output through recruitment of repressive complexes associated with a variety of proteins that bind methylcytosine or methylated CpGs (Bird et al. 1998), and that TET-mediated CpG hydroxymethylation and subsequent DNA demethylation are required to maintain enhancer accessibility, perhaps through recruitment of CXXC domain proteins such as Cpf1, a component of the SETD1 H3K4 methyltransferase complex (Compagno et al. 2017).

Our data strongly suggest that the bZIP transcription factor BATF is a major bZIP transcription factor responsible for TET recruitment to the Aicda locus, consistent with the identification that the bZIP motif is the singularly most enriched motif that correlated with DNA demethylation after B cell activation in human (Oakes et al. 2016). BATF is induced at the mRNA level before Aicda induction in activated B cells (**Fig. 1.6A**), and Batf deficiency in B cells is associated with a marked impairment of AID expression and CSR (Ise et al. 2011; Betz et al. 2010). Although loss of TET2 and TET3 had no significant effect on global BATF binding (**Fig. 1.6B and Fig. S1.8G**), BATF was required for 5hmC modification at TetE1 (**Fig. 1.6E**). Composite bZIP:IRF motifs and AP-1 motifs that support BATF:JUN:IRF and BATF:JUN cooperation, respectively, were enriched in our genome-wide 5hmC, ATAC, and DNA methylation datasets (**Fig. 1.1F** and **Fig. 1.2F** and **Fig. S1.1G** and **Fig. S1.7E**), consistent with previous findings that B cells lacking BATF or IRF4/IRF8 show impaired Aicda induction and CSR (Ise et al. 2011; Betz

et al. 2010; Klein et al. 2006; Sciammas et al. 2006; Lee et al. 2006). We propose that together with additional transcription factors, BATF:JunB and BATF:IRF facilitate the recruitment of TET proteins as well as chromatin remodeling complexes to diverse enhancers including the Aicda enhancers TetE1 and TetE2 in activated B cells, thereby promoting enhancer accessibility, 5hmC deposition, and DNA demethylation. We note, however, that the phenotype of Batf-KO mice and B cells is considerably more severe than that of Tet2/3 DKO mice and B cells. Thus, TET2/3 proteins likely function as one of several modulators of BATF function in CSR.

Our data emphasize the utility of 5hmC mapping for easy, one-step analysis of transcriptional and epigenetic landscapes in any cell type of interest. 5hmC is a quintessential epigenetic modification that marks the most highly enriched at active enhancers and the gene bodies of highly transcribed genes, and the relative levels of 5hmC at enhancers and gene bodies provide good estimates of enhancer function and the magnitude of transcription, respectively (Tsagaratou et al. 2014). 5hmC mapping by CMS-IP sufficed to identify all known enhancers in the Aicda locus, in a manner that was superior to both H3K27Ac and Tet2 ChIP-seq, and changes in 5hmC identified enhancers relevant to any particular process of cell activation or differentiation separately from all enhancers in the genome. Given its high chemical stability, the fact that its measurement requires only purified DNA, and the availability of methods for its sensitive and specific detection, 5hmC is an appealing epigenetic mark for studying gene regulation. Overall, 5hmC distribution contains information analogous to those from ATAC-seq and ChIP-seq for enhancer histone marks, effectively providing a transcriptional history of any given cell type written in DNA. If the genome is akin to an encyclopedia, 5hmC highlights those entries most relevant to a particular biological process.

1.5 Material and Methods

Mice. *Tet2*$^{fl/fl}$ and *Tet3*$^{fl/fl}$ mice were generated as previously described (54,55). C57BL/6J (000664), Ubc-CreERT2 (008085; described as Cre^{ERT2} herein), Rosa26-LSL-EYFP (006148), and AID-Cre (007770) were obtained form Jackson Laboratory. All mice used were 8-16 weeks in the C57BL/6 background and kept in specific-pathogen free animal facilitate at La Jolla Institute and were used according to protocols approved by the Institutional Animal Care and Use Committee. To induce Cre^{ERT2}-mediated deletion, Cre-expressing and control mice were intraperitoneally injected with 2 mg tamoxifen (Sigma) dissolved in 100 uL corn oil (Sigma) daily for 5 days.

B cell isolation and class switch recombination (CSR). B cells were isolated with EasySep Mouse B cell isolation kit (Stem Cell Technology, Canada) from splenocytes. To induce class switch recombination from IgM to IgG1, B cells ($5x10^5$-$1x10^6$ cells/ mL) were activated with 25 μg/mL LPS from *E. coli* O55:B5 (Sigma, St. Louis, MO) and 10 ng/mL rmIL-4 at 37°C 5% CO_2. For IgA switching, cells were activated with anti-CD40 (1 μg/mL, clone 1C10, Biolegend), rmIL-4 (10 ng/mL, Peprotech), rmIL-5 (10 g/mL, Peprotech), and rhTGFβ1 (1 ng/mL). Media were composed of RPMI1640 (Thermo Fisher, Waltham, MA) supplemented with 10% FBS, 1x MEM non-essential amino acids, 10 mM HEPES, 2 mM Glutamax, 1 mM sodium pyruvate, 55 μM 2-mercaptoethanol, and 50 μg/mL gentamicin (all from Thermo Fisher, Waltham, MA). To enhance *Cre*^{ERT2}-mediated deletion, cells from *Cre*^{ERT2} mice were cultured in the presence of 1 uM 4-hydroxytamoxifen (Tocris). All cytokines used above were from Peprotech (Rocky Hill, NJ).

Immunization. For 4-hydroxy-3-nitrophenylacetyl-conjugated ovalbumin (NP-OVA; Biosearch) immunization, the hapten-conjugated protein was diluted to 1 mg/mL in PBS was mixed with 1 volume of Alhydrogel (Invivogen) and injected into hind footpads (10 μg in 20 μL

per injection). Germinal center response was analyzed 7 days later and the two draining popliteal lymph nodes were pooled for analysis. Hapten-specific B cells were identified by positive staining with NP-phycoerythrin (BioSearch Technologies).

Retroviral transduction and two-step CSR. Retrovirus was produced by transfecting PlatE cells with MSCV-based retroviral vectors and pCL-Eco. Naïve B cells were stimulated with 5 μg/mL $F(ab')_2$ goat anti-mouse IgM (Jackson Immuno Research) and 10 μg/mL LPS at $1x10^6$ cells/mL for 24-48 hours. Retrovirus was added to the cells in the presence of 20mM HEPES and 0.8 μg/mL Polybrene (Millipore) and centrifuged at 3,000 rpm at 32°C for 90 mins. Cells were transferred back to 37°C 5% CO_2 incubator for another 24 hours. To induce CSR, cells were washed once with warm media and activated with LPS and IL-4 as above for 48 hours. Under this condition, CSR was inhibited and started to class switch only after LPS/IL-4 activation.

Flow cytometry. Primary cells and *in vitro* cultured cells were stained in FACS buffer (1% bovine serum albumin, 1mM EDTA, and 0.05% sodium azide in PBS) with indicated antibodies for 30 mins on ice. Cells were washed and then fixed with 1% paraformaldehyde (diluted from 4% with PBS; Affymetrix) before FACS analysis using FACS Canto II and FACS LSR II (BD Biosciences). Antibodies and dye were from BioLegend, eBioscience, and BD Pharmingen. Data were analyzed with FlowJo (FlowJo LLC, Ashland, OR).

Immunoblotting. Proteins isolated from B cells with RIPA buffer were resolved using NuPAGE 4-12% Bis-Tris gel (ThermoFisher) and transferred from gel to PVDF membrane using Wet/Tank Blotting Systems (Bio-Rad). Membrane was blocked with 5% non-fat milk (Bob's red mill) in TBSTE buffer (50 mM Tris-HCl pH 7.4, 150 mM NaCl, 0.05% Tween-20, 1 mM EDTA), incubated with indicated primary antibodies, followed by secondary antibodies conjugated with

horseradish peroxidase (HRP) and signal was detected with enhance chemiluminescence reagents and X-ray film.

Coimmunoprecipitation. Coimmunoprecipitation was performed similar to previously described (Scott-Browne et al. 2017). Briefly, in vitro activated B cells (48 hours) were washed twice with cold PBS and then resuspended in swelling buffer [5 mM tris (pH 7.5), 2 mM MgCl2, and 3 mM CaCl2] at 10 × 106 cells/ml. After 10 min on-ice incubation, cells were pelleted (400g, 5 min) and resuspended in swelling buffer with 10% glycerol. An equal volume of lysis buffer (1% NP-40 in swelling buffer with 10% glycerol) was added to the cells with constant mixing. Cells were incubated on ice for 5 min, pelleted (400g, 5 min) and resuspended in buffer C [10 mM Hepes (pH 7.9), 400 mM NaCl, and 1 mM EDTA] supplemented with benzonase (500 U/ml; Sigma) at 10 × 106 cells/ml, and incubated at 4°C for 30 min with constant mixing. Debris was removed by centrifuge at 13,000 rpm for 10 min, and supernatant (nuclear fraction) was recovered. An equal volume of 2× conversion buffer [10 mM tris-HCl (pH 7.5), 280 mM NaCl, 1 mM EDTA, 1 mM EGTA, 0.2% sodium deoxycholate, and 0.2% Triton X-100] was added to the nuclear proteins. For immunoprecipitation, 10 ug of rabbit anti-TET2 (Abcam) or control rabbit Ig (Santa Cruz Biotechnology) was added to the nuclear extract with 30 µl of protein A Dynabeads (Thermo Fisher Scientific) in the presence of benzonase (500 U/ml) and ethidium bromide (10 µg/µl), both of which inhibit the indirect binding between nuclear proteins via DNA. Reaction was carried out overnight at 4°C and washed three times with RIPA buffer without SDS [50 mM tris-HCl (pH 8.0), 150 mM NaCl, 1 mM EDTA, 0.5% sodium deoxycholate, and 1% NP-40]. Proteins were eluted from beads by heating at 70°C for 10 min with 1× LDS (lithium dodecyl sulfate) sample buffers (Thermo Fisher Scientific) with 10% 2-mercaptoethanol (Sigma). Immunoprecipitated proteins

were analyzed as described above using immunoblotting with Rabbit TrueBlot Anti-Rabbit IgG HRP (Rockland) as secondary antibody.

RNA extraction, cDNA synthesis, and quantitative RT-PCR. Total RNA was isolated with RNeasy plus kit (Qiagen, Germnay) or with Trizol (ThermoFisher, Waltham, MA) following manufactures' instruction. cDNA was synthesized using SuperScript III reverse transcriptase (ThermoFisher) and quantitative RT-PCR was performed using FastStart Universal SYBR Green Master mix (Roche, Germany) on a StepOnePlus real-time PCR system (ThermoFisher). Gene expression was normalized to *Gapdh*.

Bisulfite- (BS) and oxidative-bisulfite- (oxBS) sequencing. The BS and oxBS procedures were performed as previously described (Yue et al. 2016). Briefly, three PCR products containing C, mC, or hmC pertaining to different regions of ʎ phage genome were used as spike-ins at a ratio of 1:200 of the genomic DNA. 1.5 μg of genomic DNA mixed with spike-ins was ethanol precipitated of which 1 μg of the DNA was oxidized with potassium perruthenate ($KRuO_4$; Sigma) prior to bisulfite (BS) treatment (for oxBS) using MethylCode bisulfite conversion kit (ThermoFisher) and 0.5 μg of DNA was directly used for BS treatment. The BS and oxBS treated DNA were amplified using respective PCR primers and as well as primers specific to the spike-in PCR products with KAPA Uracil$^+$ PCR mix (Roche). The amplified products were pooled and libraries were prepared using the NEB Ultra II library preparation kit (NEB) according the manufacturer. The libraries were sequenced paired-end 250bp by 250bp using MiSeq with the MiSeq reagent kit v2 (500-cycles; Illumina).

Genome-wide 5hmC mapping by cytosine-5-methylenesulfonate immunoprecipitation (CMS-IP). Techniques for immunoprecipitation of DNA with antibodies are plagued by the fact that Igs nonspecifically immunoprecipitate DNA sequences containing CA

and other DNA repeats (Lentini et al. 2018). However, CMS is a derivative formed only after the reaction of 5hmC with sodium bisulfite, and anti-CMS antibodies recognize CMS with strong sensitivity and selectivity (Pastor et al. 2011). Lentini and colleages (Lentini et al. 2018) showed that the CMS-IP method is free of the nonspecific background of immunoprecipitation of DNA repeats, most likely because bisulfite treatment results in C > T conversions and/or single-stranded, as supposed to doublestranded, DNA was used for IP (57).

CMS-IP was performed essentially as previously described (Scott-Browne et al. 2017; Huang, et al. 2012; Pastor et al. 2011). Briefly, genomic DNA isolated from naïve and activated B cells was spiked with unmethylated lambda phage cI857 Sam7 DNA (Promega, Madison, WI, USA) and a PCR amplicon from a puromycinresistant gene at a ratio of 200:1 and 100,000:1, respectively. DNA (5 to 10 µg in 130 µl tris-EDTA buffer) was sheared with a Covaris E220 using microTUBE for 4 min. DNA was cleaned up with Ampure XP beads, processed with NEBNext End Repair and A-tail Modules (NEB, Ipswich, MA, USA), and ligated to methylated Illumina adaptors (NEB). DNA was then bisulfite-treated (MethylCode, Thermo Fisher Scientific), denatured, and immunoprecipitated with anti-CMS serum (in-house) and a mixture of protein A and G Dynabeads (Thermo Fisher Scientific). Because our goal for this study was to identify regions that undergo significant activation-induced 5hmC modification after LPS + IL-4 stimulation, we normalized the data within each sample using total sequencing read counts for each individual time point, without using 5hmC-containing oligonucleotides as spike-ins to consider the progressive dilution of 5hmC that occurs as a function of cell division (**Fig. S1.2C**). Libraries for immunoprecipitated DNA were generated by PCR with barcoded primers (NEBNext Multiplex Oligos for Illumina; NEB) for 15 cycles using KAPA HiFi HotStart Uracil+ ReadyMix

(Roche), followed by a cleanup with Ampure XP beads (Beckman Coulter), and sequenced with a HiSeq 2500 (Illumina, San Diego, CA, USA) with paired-end 50-bp reads.

Locus specific analysis of 5hmC with AbaSI-qPCR. Genomic DNA (200 ng) was treated with T4 beta-glucosyltransferase (ThermoFisher) in the presence of UDP-glucose to glycosylate 5hmC at 37°C overnight. Half of the reaction was digested with AbaSI (NEB), which is specifically active for glycosylated 5hmC, for 4 hours at 25°C followed by 15 mins at 65°C to inactivate enzymes. The uncut sample was processed as above without the addition of AbaSI. Equal amount of DNA from the above reactions was used as template for real-time PCR as described for RNA qRT-PCR using primers TetE1-CMS-qF and TetE1-CMS-qR. To monitor the degree of digestion, samples were spiked-in 1 pg control DNA with a single 5hmC-modified CpG (EpiMark 5-hmC and 5-mC Analysis Kit; NEB). The relative amount of 5hmC was calculated by the percentage of decrease in qPCR signals in digested half relative to undigested half. As control to monitor non-specific digestion (not show), a genomic region containing CpG motifs but without 5hmC modification in B cells (*Foxp3 CNS2*) was used as a control with Foxp3-CNS2-qF and Foxp3-CNS2-qR.

DNA dot blot. DNA dot blot was performed as previously described (Scott-Browne et al. 2017; Yue et al. 2016). To analyze 5hmC abundance, genomic DNA was treated with sodium bisulfite as above. DNA was diluted two-fold serially with TE buffer, denatured in 0.4 M sodium hydroxide and 10 mM EDTA at 95°C for 10 mins, and then immediately chilled on ice. Equally volume of ice-cold 2 M ammonium acetate pH 7.0 was added and incubated on ice for 10 mins. Denatured DNA were spotted on a nitrocellulose membrane using a Bio-Dot apparatus (Bio-Rad), washed with 2x SSC buffer (300 mM NaCl and 30 mM sodium citrate), and baked in a vacuum oven at 80°C for 2 hours. To detect CMS, membrane was rehydrated with TBSTE buffer and

blocked with 5% non-fat milk (Bob's red mill) in TBSTE buffer. CMS was detected with primary rabbit anti-CMS antisera (in house) following the procedures above for Immunoblotting.

Chromatin Immunoprecipitation sequencing (ChIP-seq). Chromatin immunoprecipitation was performed as described before (Scott-Browne et al. 2017). Briefly, cells were fixed with 1% formaldehyde (ThermoFisher) at room temperature for 10 mins at $1x10^6$ cell/mL in media, quenched with 125 mM glycine, washed twice with ice cold PBS. Cells were pelleted, snap-froze with liquid nitrogen, and store at -80°C until use. For Tet2-ChIP, activated cells were centrifugated at 250 x *g* for 5 mins and cell pellets were resuspended in 37°C PBS with 2mM disuccinimidyl glutarate to crosslink proteins for 30 mins at room temperature. Formaldehyde was added to a final concentration of 1% and the cells were incubated at room temperature for 10 mins with nutation. Quenching and cell storage were performed as above. To isolate nuclei for sonication, cell pellets were thawed on ice and lysed with lysis buffer (50 mM HEPES pH 7.5, 140 mM NaCl, 1mM EDTA, 10% glycerol, 0.5% NP40, 0.25% Triton-X100) for 10 mins at 4°C with rotation, washed once with washing buffer (10 mM Tris-HCl pH 8.0, 200 mM NaCl, 1 mM EDTA, 0.5 mM EGTA) and twice with shearing buffer (10 mM Tris-HCl pH 8.0, 1 mM EDTA, 0.1% SDS). Nuclei were resuspended in 1mL shearing buffer and sonicated with Covaris E220 using 1 mL milliTUBE (Covaris, Woburn, MA) for 18-20 minutes (Duty Cycle 5%, intensity 140 Watts, cycles per burst 200). After sonication, insoluble debris was removed by centrifugation at 20,000 x *g*. Buffer for chromatin was adjusted with 1 volume of 2x conversion buffer (10 mM Tris-HCl pH 7.5, 280 mM NaCl, 1 mM EDTA, 1mM EGTA, 0.2% sodium deoxycholate, 0.2% Triton-X100, 1% Halt protease inhibitors with (for H3K27Ac) or without (for BATF, Tet2) 0.1% SDS. Chromatin was pre-cleared with washed protein A dynabeads (ThermoFisher) for 2 hours, incubated with antibodies and protein A dynabeads overnight (all

procedures were at 4°C with rotation). For H3K27Ac ChIP, bead-bound chromatin was washed twice with RIPA buffer (50 mM Tris-HCl pH 8.0, 150 mM NaCl, 1 mM EDTA, 0.5% sodium deoxycholate, 1% NP-40, 0.1% SDS), once with high salt wash buffer (50 mM Tris-HCl pH 8.0, 500 mM NaCl, 1 mM EDTA, 1% NP-40, 0.1% SDS), and once with TE (10 mM Tris-HCl pH 8.0, 1 mM EDTA). For BATF ChIP, all wash buffers were as above but without SDS. For Tet2 ChIP, beads were washed three times with RIPA buffer without SDS and once with TE. Chromatin was eluted from beads with elution buffer (100 mM $NaHCO_3$, 1% SDS, 1 mg/mL RNaseA; Qiagen) twice for 30 mins each at 37°C with constant shaking. NaCl and proteinase K (Ambion) were added to the eluted chromatin at concentrations of 250 mM and 0.5 mg/mL, respectively, and de-crosslinked at 65°C overnight with constant shaking. DNA was purified with Zymo ChIP DNA Clean & Concentrator-Capped Column (Zymo Research, Irvine, CA). Library was prepared with NEB Ultra II library prep kit (NEB) following manufacture's instruction and was sequenced on an Illumina Hiseq 2500 (single-end 50 bp reads).

ATAC-seq. Procedures were as described (Scott-Browne et al. 2017). Briefly, 50,000 cells were collected by centrifugation and washed once with 50uL ice-cold PBS and centrifuged at 600 *xg* for 5 mins at 4°C. Cell pellets were resuspended in 50 µL of cold lysis buffer (10mM Tris-HCl pH 7.4, 10 mM NaCl, 3 mM $MgCl_2$, 0.1% IGEPAL CA-630) and spin down immediately at 500 *xg* for 10 mins at 4°C. Supernatant was discarded and nuclei were resuspended in 50µL transposition reaction mix (25µL 1x TD buffer fom Illuminia, 2.5µL Tn5 transposase, 22.5µL H_2O), incubated at 37°C for 30 mins, and DNA was purified with a Qiagen MinElute kit (Qiagen). Library was amplified with KAPA HiFi HotStart Real-time PCR Master Mix (Roche) using indexed primers and sequenced on an Illumina Hiseq 2500 (paired-end 50 bp reads).

RNA-sequencing with Smart-seq. Smart-seq was performed as described previously (Tsagaratou et al. 2017; Picelli et al. 2014). Briefly, total RNA was isolated from naïve and activated B cells with Trizol (ThermoFisher) and the integrity of the RNA was accessed with TapeStation RNA Analysis ScreenTape or Bioanalyzer RNA pico kit (Agilent). 10ng of RNA was reverse transcribed using oligo-dT_{30} VN primer in the presence of Template Switching Oligo (TSO) with SuperScript II reverse transcriptase. cDNA was pre-amplified with IS PCR primers and PCR products were cleaned up with Ampure XP beads. One ng of PCR product was used to generate library using NexteraXT library prep kit (Illumina) and tagmentated DNA was amplified for a 12 cycles PCR and purified with AmpureXP beads. Libraries were sequenced on an Illumina Hiseq 2500 with single-end 50 bp reads.

Statistical analyses. Statistical analyses and bar plots were performed and plotted with Prism 7 or R (v3.3.3). The bar graph and dot plots shown indicate mean and SE. Most experiments were analyzed using two-tailed unpaired t test or Wilcoxon rank-sum test, as indicated in the figure legends unless otherwise stated. The reference genome used was mm10. Heatmaps and profile plots were generated using DeepTools (Ramírez et al. 2016).

CMSIP analysis. Paired-end reads (50bp) were mapped to the mouse genome mm10 GRCm38 (Dec. 2011) from UCSC, using BSMAP (V.2.74) (-v 4 -R -n 1 -w 2 -r 0 -q 20 -R -p 8) (Xi et al. 2009). Reads that mapped to the spike-in control (Lambda and Puro) were filtered out from the Sam file using awk. Tag directories were created with the remaining reads using makeTagDirectory from HOMER (Heinz & Benner et al. 2010) (-genome mm10 -tbp 1 –checkGC). Peaks were called with findPeaks from HOMER (-style histone -o auto -i). Peaks from all samples were merged with mergePeaks from HOMER into a master table. Quantile

normalization was applied to all raw counts files and differentially enriched 5hmC regions were identified with edgeR (Robinson et al. 2010); a p adjusted value of ≤ 0.05 was used as a cutoff.

H3K27Ac ChIP analysis. Single end reads (50bp) were mapped to the mouse genome mm10 GRCm38 (Dec. 2011) from UCSC with Bowtie (V.1.1.2). Reads were sorted and PCR duplicates were removed using SortSam and MarkDuplicates, respectively from Picard Tools (V.2.7.1). Tag directories were created with makeTagDirectory (-genome mm10 –checkGC) from HOMER, and peaks were called with findPeaks (-region). Peaks from all samples were merged with mergePeaks from HOMER into a master table. Quantile normalization was applied to all raw counts files and differentially enriched 5hmC regions were identified with edgeR (Robinson et al. 2010); a p adjusted value of ≤ 0.05 was used as a cutoff.

BATF ChIP analysis. Single end reads (50bp) were mapped to the mouse genome mm10 GRCm38 (Dec. 2011) from UCSC with Bowtie (v1.1.2). Reads were sorted and PCR duplicates were removed using SortSam and MarkDuplicates, respectively from Picard Tools (V.2.7.1). Tag directories were created with makeTagDirectory (-genome mm10 –checkGC) from HOMER, and peaks were called with findPeaks (-style factor -o auto).

Definition of preferentially active enhancers. Preferentially active enhancers (**Fig. S1.1H**) were defined as distal H3K27Ac regions (> 1kb from TSS) that had ATAC-seq and H3K27Ac peaks overlapping in at least 50% of either peak/region; overlapping was calculated with intersectBed -f 0.5 -f 0.5 –e (Bedtools v2.26.0). The differential enrichment in an enhancer was called if it contains a differentially enriched H3K27Ac region as well as at least one differentially accessible region.

ATAC-seq analysis. Paired-end reads (100 bp) were mapped to the mouse genome mm10 GRCm38 (Dec. 2011) from UCSC using Bowtie 1.0.0 ("-p 8 -m 1 --best --strata -X 2000 -S --fr --chunkmbs 1024."). Reads that failed this alignment step were filtered for Illumina adapters and low quality using "Trim Galore!" ("--paired --nextera --length 37 -- stringency 3 --three_prime_clip_R1 1 --three_prime_clip_R2 1") and re-mapped using the same parameters. Both mapping results were merged and processed together to remove duplicates using picard-tools-1.94 MarkDuplicates. Mitochondrial and Chromosome Y reads were excluded. Subnucleosomal fragments were obtained with SAMtools and awk to identify DNA fragments that were less than or equal to 100 nt in length. These fragments were used to call peaks using HOMER (v4.9.1) findPeaks function for each replicate ("-size 500 -region -center -P .1 -LP .1 -poisson .1 -style dnase") and all the peak sets were merged to generate a global set. Peaks overlapping with ENCODE blacklisted regions (ENCODE Project Consortium 2012) were removed. From each sample, Tn5 insertion sites were obtained by isolation of the initial 9bp of mapped reads (Buenrostro et al. 2013) which were used to compute the number of transposase insertions per peak using MEDIPS (Chavez et al. 2010). Raw reads from all samples were quantile-normalized prior to differential coverage analysis using edgeR without TMM (Trimmed mean of M-values) normalization. Only regions with more than 32 normalized reads across the samples per comparison. Differentially accessible regions were defined by an adjusted p value (FDR) lower than 0.05 and a log2 fold enrichment higher equal than 1.

OxBS analysis. OxBS-seq reads were mapped to both the mouse genome mm10 GRCm38 (Dec. 2011) from UCSC and the phage Lambda genome (GenBank: J02459.1) using bsmap-2.90 (" -v 15 -w 3 -p 4 -S 1921 -q 20 -A AGATCGGAAGAGC -r 0 -R -V 2 "). The mapping results were separated into reads belonging to the mm10 genome and each of the three loci from lambda

used for oxidation and conversion efficiency calculation. Methylation calls from lambda- and mm10-derived reads were obtained using bsmap-2.90 function methratio.py (" -u -p -g -i "correct" -x CG,CHG,CHH "). Conversion efficiencies as well as posterior probabilities of methylation, hydroxymethylation and unmodified cytosine were calculated by luxGLM v.0.666 (prior probabilities used for for C, hmC and mC "998,1,1", "6,2,72" and "1,998,1" respectively) (Äijö et al. 2016). Following genomic positions from lambda used for oxidation and BS treatment efficiencies: chrLambda:22893-23053 C; chrLambda:23765-23925 hmC; chrLambda:47335-47495 mC.

WGBS analysis. WGBS reads were mapped to both the mouse genome mm10 GRCm38 (Dec. 2011) from UCSC. Bisulfite conversion efficiency was estimated based on cytosine methylation in non-CpG context. For all the samples the bisulfite conversion efficiency was higher than 0.9996. Duplicated reads caused by PCR amplification were removed by BSeQC (v1.2.0) applying a p value cutoff Poisson distribution test in removing duplicate reads (1e5) (Lin et al. 2013). Consequently, a maximum of three stacked reads at the same genomic location were allowed and kept for further analysis. In addition, BSeQC was employed for removing DNA methylation artifacts introduced by end repair during adaptor ligation. Overlapping segments of two mates of a pair were reduced to only one copy to avoid considering the same region twice during the subsequent DNA methylation quantification. To estimate CpG DNA methylation at both DNA strands, methratio.py script was executed from BSMAP (v2.90) (-u -r -z -g - i "correct" -x CG). To identify differentially methylated cytosines and regions (DMCs and DMRs), a naïve B cells dataset and was compared to a activated B cell replicate using RADmeth methpipe-3.4.2 (adjust -bins 1:100:1 ; merge -p 0.05) (Song et al. 2013).

RNA-seq analysis. RNA-seq samples at four different time-points collected from WT and DKO conditions were first mapped to the mouse genome mm10/GRCm38 using both Hisat2 (Kim et al. 2015) ("--no-mixed --no-discordant --add-chrname –dta") and Tophat2 (Kim et al. 2013) ("--no-novel-juncs") alignment programs separately. Aligned bam files obtained from both the programs were further used to generate the Hisat2- and Tophat2-specific counts using HTseqcount program (Anders et al. 2015) (default parameters). Hisat2- and Tophat2-specific count files at each time point for WT and DKO conditions were then used to identify the differentially expressed genes (FDR < 0.05) between matching time points using edgeR program (Robinson et al. 2010). Potential batch effects were removed using svaseq program (Leek 2014). Finally, the common differentially expressed genes obtained from both Hisat2- and Tophat2-specific list were used to perform the downstream analysis.

Genome-browser track generation for ChIP-seq. ChIP-seq results from TET2, Ig control, E2A, PU.1, p300, and GCN5 were processed as follow to generate the genome browser tracks. Fastq files were mapped to mm10 reference genome with Bowtie 2 (v2.1.0) with "- very-sensitive". The mapped SAM files were converted to BAM using Samtools (v1.7) view –h –F 4, and duplicates were removed using Picard (v2.7.1). BigWig files were made by first generating a BedGraph files from the filtered Bam files using Bedtools (v2.26.0) genomecov followed by bedGraphToBigWig (v4) with read counts normalized to 10,000,000 reads.

Miscellaneous analyses of regions. For distance between regions to the closest TSS was analyzed with HOMER software with "annotatePeaks.pl –annStats". Overlap between regions was analyzed by "bedtools intersect" with no requirement for the degree of overlapping. The degree of significance for overlap between superenhancers and test regions was estimated by Fisher exact test using "bedtools fisher".

Time-series analysis. For a unified analysis of the RNA-seq time-course data (0hr to 96hr) from WT samples, TC-seq package (Wu & Gu 2018) was used on the combined RNA-seq read counts, obtained after applying Tophat2 (Kim et al. 2013) and Hisat2 (Kim et al. 2015) alignment programs (Described in the previous RNA-seq analysis part). TC-seq utilizes GLM method implemented in edgeR package (Robinson et al. 2010) to detect the differential events in gene expression. Differential analysis was performed between "0hr" to the rest of the time points, and the significant differential events were extracted whenever a log2FC > 2 or < -2 and FDR < 0.05 criteria was satisfied. To detect the temporal pattern of the differential gene expressions (RPKM values), a soft clustering algorithm implemented in TCseq program was applied ("algo = 'cm', k = 6, standardize = TRUE"). Finally, the differential genes were assigned to a cluster (C1- C6) representing a specific temporal pattern of expression, if the membership probability of the genes to a cluster is 0.5 or more.

Published datasets. Naïve H3K4me1 (0h): SRR1535686, SRR1535685. Activated H3K4me1 (48h): SRR1014530. SRR1087900. Naive WGBS (0h): SRR1003257. Activated WGBS (48h): SRR1020523. Naive PU1 (0h): SRR2976278. Activated PU1 (48h): SRR1014532. Naïve DSG control (0h): SRR3158132. Activated E2A DSG anti-CD40/IL-4 (48h): SRR3158146. Naïve Brg1 (0h): SRR3619348. Naïve Chd4 (0h): SRR3619349. Naïve Gcn5 (0h): SRR3619350. Naïve p300 (0h): SRR3619356. Activated Brg1 (24h): SRR3619334. Activated Chd4 (24h): SRR3619335. Activated Gcn5 (24h): SRR3619336. Activated p300 (24h): SRR3619342.

1.6 Figures

Figure 1.1 Dynamic changes in 5hmC during B cell activation.
(A) Flow-chart of experiments. B cells were activated with LPS+IL-4 for the indicated times prior to genome-wide analyses. **(B)** Of a total of 159,305 5hmC-enriched regions in B cells activated for 72h, while most regions (grey, 94.1%) display similar 5hmC levels, 8,454 (blue, 5.3%) show increased 5hmC and 928 (red, 0.6%) show decreased 5hmC relative to naïve B cells. Note that 193 regions represented only in naïve B cells are not shown. **(C)** Number of differential hydroxymethylated region (DhmR) showing increased and decreased 5hmC at respective time points after activation, of a total number of ~160,000 5hmC-marked regions present in naïve and activated B cells (**Fig. S1.1A**). **(D)** Heatmaps showing the kinetics of 5hmC in the 8,454 regions with increased 5hmC at 72h compared to naïve B cells (*left panels*), but no increase in the same number of 5hmC-marked regions common to naïve and 72h-activated B cells (*middle panels).* *Right panels*, only a small number of random genomic regions are marked with 5hmC. For a similar analysis of the 1,121 regions that lose 5hmC after 72h of B cell activation, see **Fig. S1.1C**. 5hmC enrichment is shown as normalized reads per 100 bp bin. **(E)** The 85 and 1,953 regions with increased 5hmC in 24h- and 48h- activated B cells relative to naïve B cells show decreased "methylation" (bisulfite-resistant cytosine, 5mC+5hmC) at their centers 48h after activation. Average methylation was calculated for each 200 bp bin across 6kb. **(F)** Significant enrichment for consensus RHD (NFκB), IRF:bZIP, and bZIP transcription factor binding motifs in 8,454 regions $DhmR^{72h\text{-}up}$ showing increased 5hmC in 72h-activated relative to naïve B cells. Common 5hmC-enriched regions were used as background for analysis. Y-axis indicates the fold enrichment versus background, circle size indicates the percentage of regions containing the respective motif, and the color indicates the significance (Log_{10} *p* vaule). **(G)** 5hmC is enriched at active ($H3K4me1^{+}$ $H3K27Ac^{hi}$) relative to poised ($H3K4me1^{+}$ $H3K27Ac^{lo}$) enhancers in both activated and naïve B cells. Y and X axes indicate the levels (log_2) of H3K4me1 and H3K27Ac relative to input, respectively. **(H)** A substantial fraction of super-enhancers (76.7%, 352 of 459) identified by high H3K27Ac enrichment overlap with $DhmR^{72h\text{-}up}$ at which 5hmC is increased in activated (72h) relative to naïve B cells. Fisher exact test was used to analyze the significance. ***, $p<0.01$ ($p=8.9203\text{x}10^{-266}$). n.s., not significant. **(I)** Genome browser view of the *Ccr4* locus (mm10; chr9:114,484,000-114,501,000) as an example of a genomic region marked by increased 5hmC, increased H3K27Ac and decreased CpG methylation in activated compared to naïve B cells. Red track indicates CpGs that were included for analysis based on coverage. **(J)** Kinetics of increase in *Ccr4* mRNA expression (by RNA-seq) in activated B cells. See also **Fig. S1.1**.

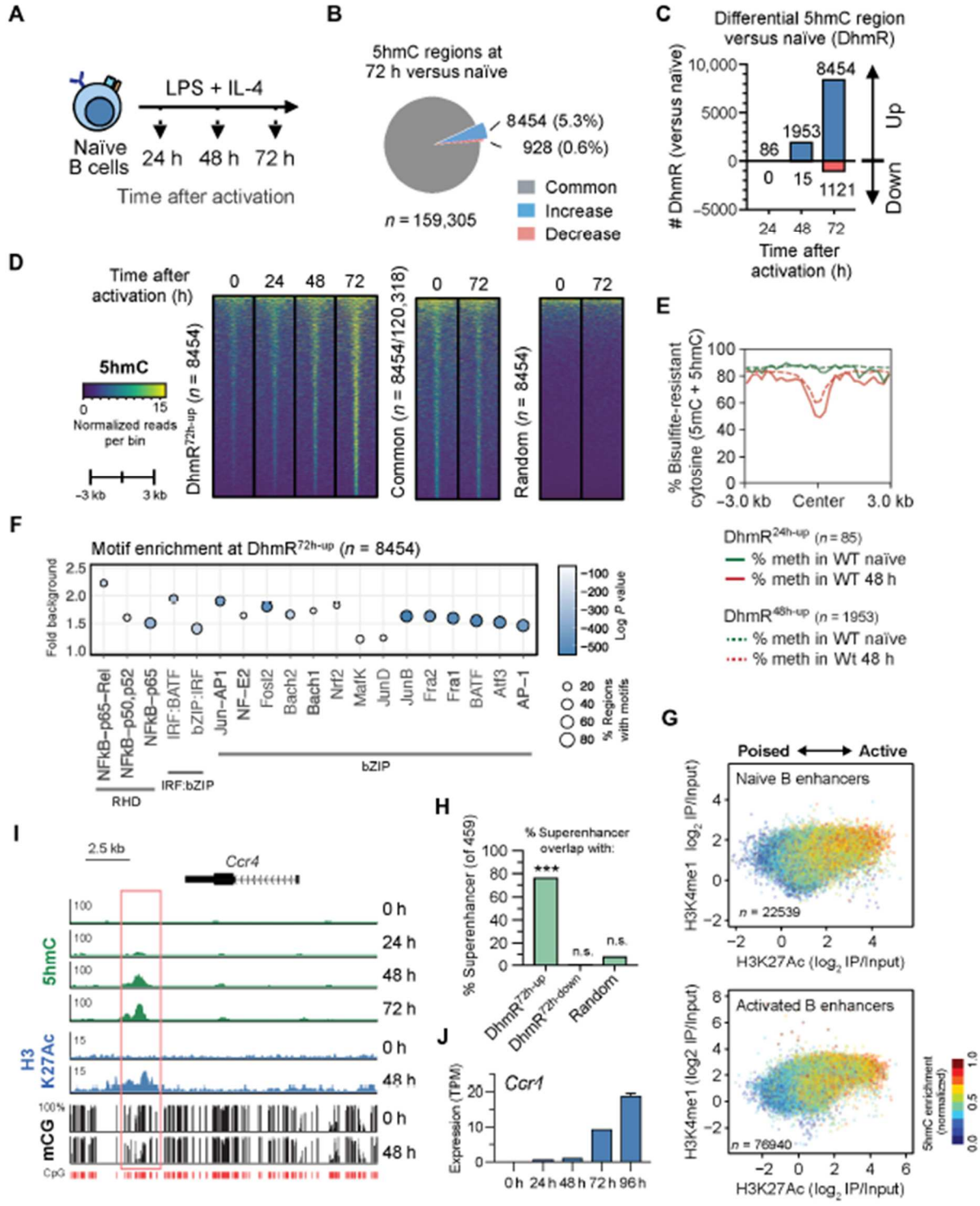

A
LPS + IL-4
Naïve B cells
24 h
48 h
72 h
Time after activation
B
5hmC regions at 72 h versus naïve
8454 (5.3%)
928 (0.6%)
n = 159,305
Common
Increase
Decrease
C
Differential 5hmC region versus naïve (DhmR)
DhmR (versus naïve)
10,000
5000
0
-5000
8454
1953
86
0
15
1121
Up
Down
24
48
72
Time after activation (h)
D
Time after activation (h)
0
24
48
72
5hmC
0
15
Normalized reads per bin
-3 kb
3 kb
DhmR72h-up (n = 8454)
Common (n = 8454/120,318)
Random (n = 8454)
E
% Bisulfite-resistant cytosine (5mC + 5hmC)
-3.0 kb
Center
3.0 kb
DhmR24h-up (n = 85)
% meth in WT naïve
% meth in WT 48 h
DhmR48h-up (n = 1953)
% meth in WT naïve
% meth in Wt 48 h
F
Motif enrichment at DhmR72h-up (n = 8454)
Fold background
Log P value
% Regions with motifs
NFkB–p65–Rel
NFkB–p50,p52
NFkB–p65
IRF:BATF
bZIP:IRF
Jun–AP1
NF–E2
Fosl2
Bach2
Bach1
Nrf2
MafK
JunD
JunB
Fra2
Fra1
BATF
Atf3
AP–1
RHD
IRF:bZIP
bZIP
G
Poised
Active
Naive B enhancers
n = 22539
H3K4me1 log2 IP/Input
H3K27Ac (log2 IP/Input)
Activated B enhancers
n = 76940
5hmC enrichment (normalized)
H
% Superenhancer overlap with:
% Superenhancer (of 459)
***
n.s.
n.s.
DhmR72h-up
DhmR72h-down
Random
I
2.5 kb
Ccr4
5hmC
H3 K27Ac
mCG
CpG
0 h
24 h
48 h
72 h
J
Ccr1
Expression (TPM)
0 h 24 h 48 h 72 h 96 h

Figure 1.2. Comparison of 5hmC modification in WT and *Tet2/3* DKO B cells.
(A) Mean mRNA expression levels for TET family members (from RNA-seq) in WT naïve and activated B cells. TPM, transcript per million. **(B)** Description of mice and flow chart of experiment. **(C)** *Tet2* and *Tet3* are efficiently deleted. *Tet2* and *Tet3* expression in B cells from tamoxifen-treated WT control and *Tet2/3* DKO mice (described in **Fig. 1.2D**) were analyzed by qRT-PCR. Data were normalized to *Gapdh* within sample and subsequently to the value from WT. Representative of two independent experiments with three technical replicates is shown. ***, $p<0.01$. **(D)** Number of regions differentially marked with 5hmC (DhmR) between WT and *Tet2/3* DKO B cells as a function of time after activation. **(E)** Heatmaps showing the kinetics of 5hmC enrichment signals from WT (*left panels*) and *Tet2/*3-DKO *(right panels)* at the differentially hydroxymethylated regions ($DhmR^{72h}$) between WT and DKO (72h, **D**). Regions with decreased 5hmC in DKO are shown on top (WT>DKO, n=2,139) and those with increased 5hmC on bottom (DKO>WT, n=184). 5hmC enrichment is shown in normalized reads per 100 bp bin. **(F)** Strong enrichment for consensus RHD (NFκB), IRF:bZIP (IRF:BATF) and bZIP transcription factor binding motifs in the "TET-dependent" regions with decreased 5hmC in 72h-activated *Tet2/3* DKO relative to WT B cells ($DhmR^{72h\text{-}WT>DKO}$, n=2,139). Common 5hmC-enriched regions were used as background for analysis. Y-axis indicates the fold enrichment versus background, circle size indicates the percentage of regions containing the respective motif, and the color indicates the significance (Log_{10} *p* vaule). See also **Fig. S1.2**.

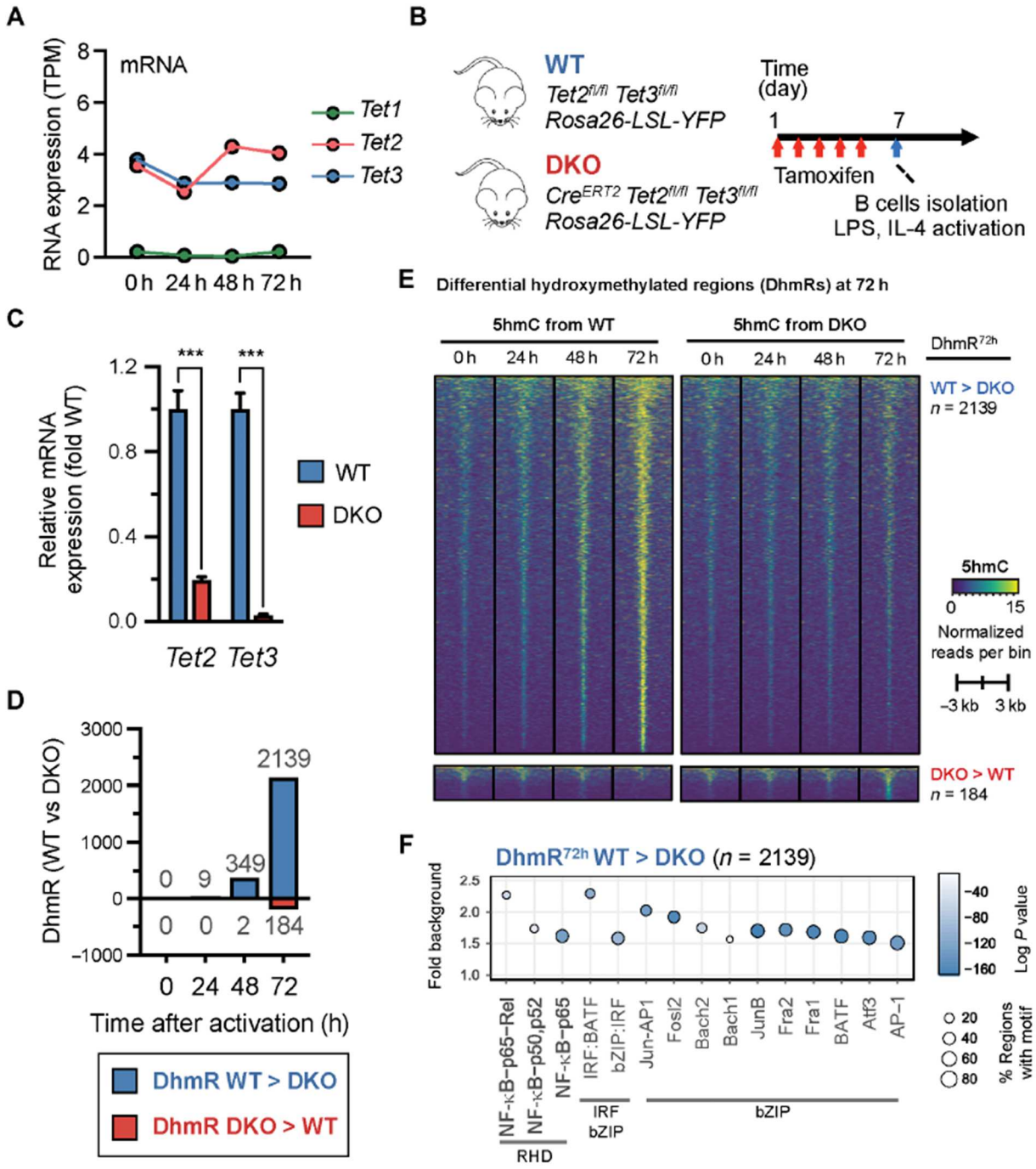
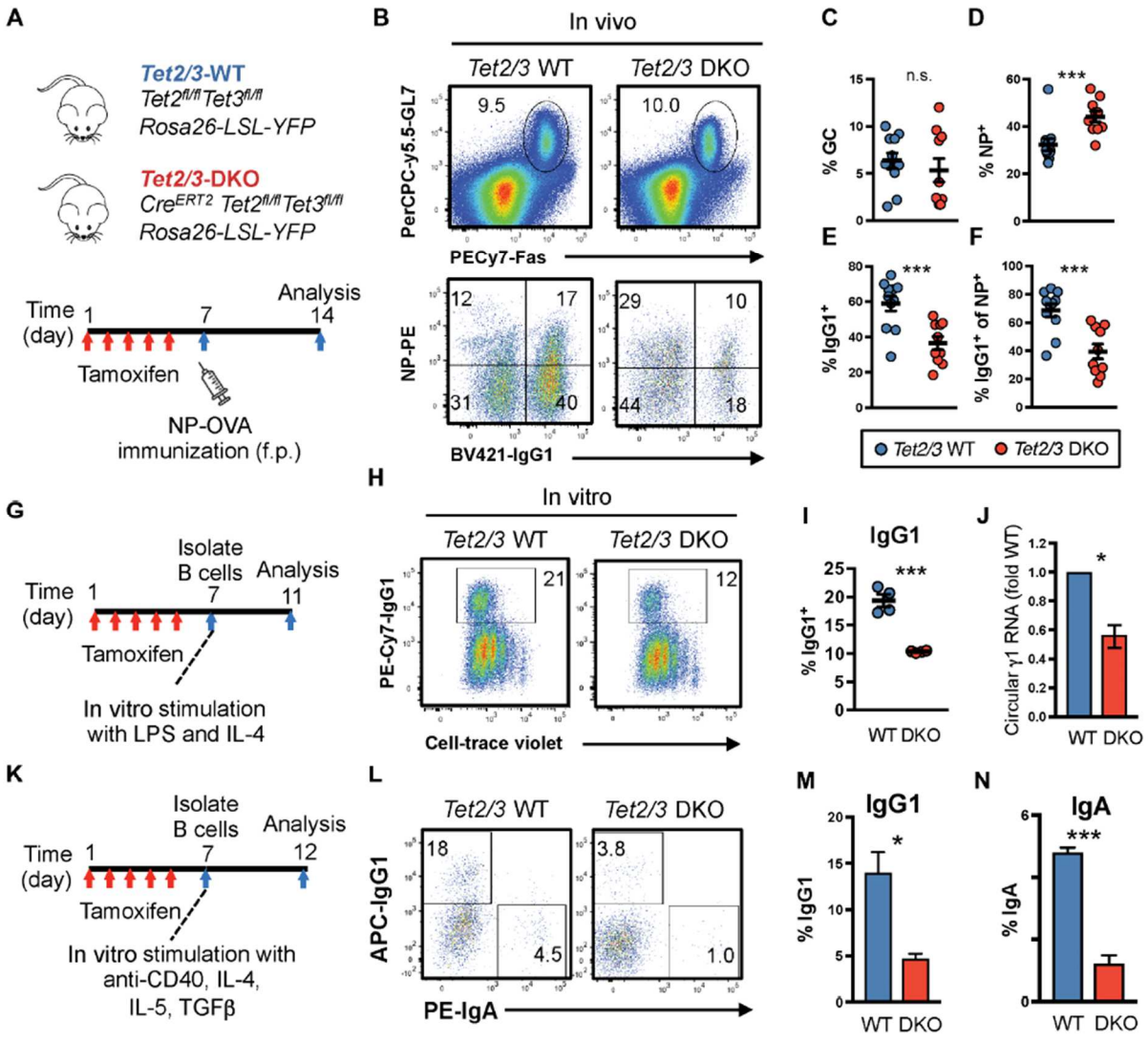

A
mRNA
RNA expression (TPM)
Tet1
Tet2
Tet3
0 h 24 h 48 h 72 h
B
WT
Tet2fl/fl Tet3fl/fl
Rosa26-LSL-YFP
DKO
CreERT2 Tet2fl/fl Tet3fl/fl
Rosa26-LSL-YFP
Time (day)
1
7
Tamoxifen
B cells isolation
LPS, IL-4 activation
C
Relative mRNA expression (fold WT)
***
***
WT
DKO
Tet2 Tet3
D
DhmR (WT vs DKO)
0 9 349 2139
0 0 2 184
0 24 48 72
Time after activation (h)
DhmR WT > DKO
DhmR DKO > WT
E
Differential hydroxymethylated regions (DhmRs) at 72 h
5hmC from WT
5hmC from DKO
0 h 24 h 48 h 72 h
DhmR72h
WT > DKO
n = 2139
5hmC
Normalized reads per bin
−3 kb 3 kb
DKO > WT
n = 184
F
DhmR72h WT > DKO (n = 2139)
Fold background
Log P value
% Regions with motif
NF-κB-p65-Rel
NF-κB-p50,p52
NF-κB-p65
IRF:BATF
bZIP:IRF
Jun-AP1
Fosl2
Bach2
Bach1
JunB
Fra2
Fra1
BATF
Atf3
AP-1
RHD
IRF bZIP
bZIP

Figure 1.3. TET proteins facilitate class switch recombination (CSR) *in vitro* and *in vivo.*
(A) Flow chart of experiment to assess CSR *in vivo*. *f.p.,* foot pad. **(B)** *Upper panels*, flow cytometry plots showing equivalent frequencies of $CD19^+GL7^+Fas^+$ germinal center B (GCB) cells at the draining popliteal lymph nodes from WT and *Tet2/3* DKO mice after treated with tamoxifen and immunized with NP-OVA as indicated in **(A)**. *Lower panels*, flow cytometry plots showing decreased frequencies of IgG1-switched cells among GCB cells in *Tet2/3* DKO (YFP^+ GCB-gated) compared to WT mice (GCB-gated). **(C-F)** Quantifications of experiments shown in **(B)**. Data shown are aggregated results from two independent experiments. Means and standard errors are shown. WT, n=11; DKO, n=12. **(G)** Flow chart of experiment to assess CSR (IgG1 switching) *in vitro*. Cells were labeled with Cell-Trace violet and activated for 4 days with LPS (25 ug/mL) and rmIL-4 (10 ng/mL). **(H-I)** Flow cytometry plots (**H**) and quantification of experiments (**i**) show decreased frequencies of IgG1-switched B cells in *Tet2/3* DKO (n=4) compared to WT (n=4) mice. Data were representative from at least three independent experiments. **(J)** Circular gamma 1 transcript, generated after successful IgG1 switching, was quantified by qRT-PCR and normalized to *Gapdh* and then to the level of WT. Representative of two independent experiment is shown with three technical replicates. **(K)** Flow chart of experiment to assess CSR (IgG1- and IgA-switching) in cell culture. Cells were activated for 5 days with anti-CD40 (1 ug/mL), rmIL-4 (10 ng/mL), rmIL-5 (10 ng/mL), and rhTGFbeta (1 ng/mL). **(L-M)** Flow cytometry plots **(L)** and quantification of experiments **(M, N)** showing decreased frequencies of IgG1- **(M)** and IgA-switched cells **(N)** in *Tet2/3* DKO compared to WT cells. Data shown are representative from three independent experiments with three technical replicates. Statistical significance was calculated using unpaired two-tailed *t*-test. n.s., not significant. ***, $p<0.01$. *, $p<0.05$. See also **Fig. S1.3**.


A
Tet2/3-WT
Tet2fl/flTet3fl/fl
Rosa26-LSL-YFP
Tet2/3-DKO
CreERT2 Tet2fl/flTet3fl/fl
Rosa26-LSL-YFP
Time (day)
1
7
14
Analysis
Tamoxifen
NP-OVA immunization (f.p.)
B
In vivo
Tet2/3 WT
Tet2/3 DKO
PerCPC-y5.5-GL7
PECy7-Fas
NP-PE
BV421-IgG1
9.5
10.0
12
17
31
40
29
10
44
18
C
% GC
n.s.
D
% NP+
***
E
% IgG1+
***
F
% IgG1+ of NP+
***
Tet2/3 WT
Tet2/3 DKO
G
Isolate B cells
Analysis
Time (day)
1
7
11
Tamoxifen
In vitro stimulation with LPS and IL-4
H
In vitro
Tet2/3 WT
Tet2/3 DKO
PE-Cy7-IgG1
Cell-trace violet
21
12
I
IgG1
% IgG1+
***
WT DKO
J
Circular γ1 RNA (fold WT)
*
WT DKO
K
Isolate B cells
Analysis
Time (day)
1
7
12
Tamoxifen
In vitro stimulation with anti-CD40, IL-4, IL-5, TGFβ
L
Tet2/3 WT
Tet2/3 DKO
APC-IgG1
PE-IgA
18
4.5
3.8
1.0
M
IgG1
% IgG1
*
WT DKO
N
IgA
% IgA
***
WT DKO

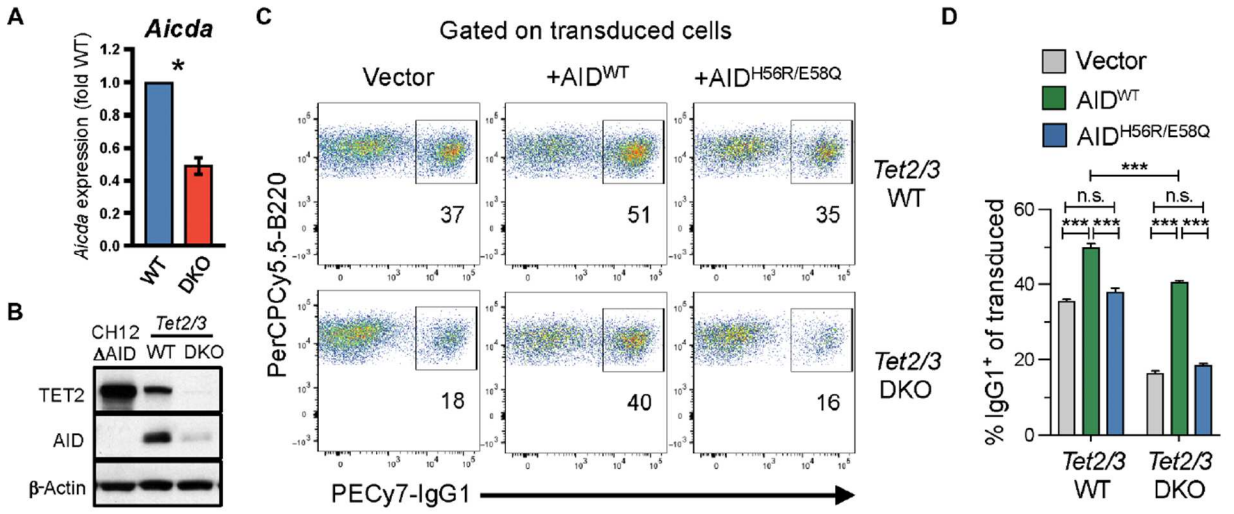
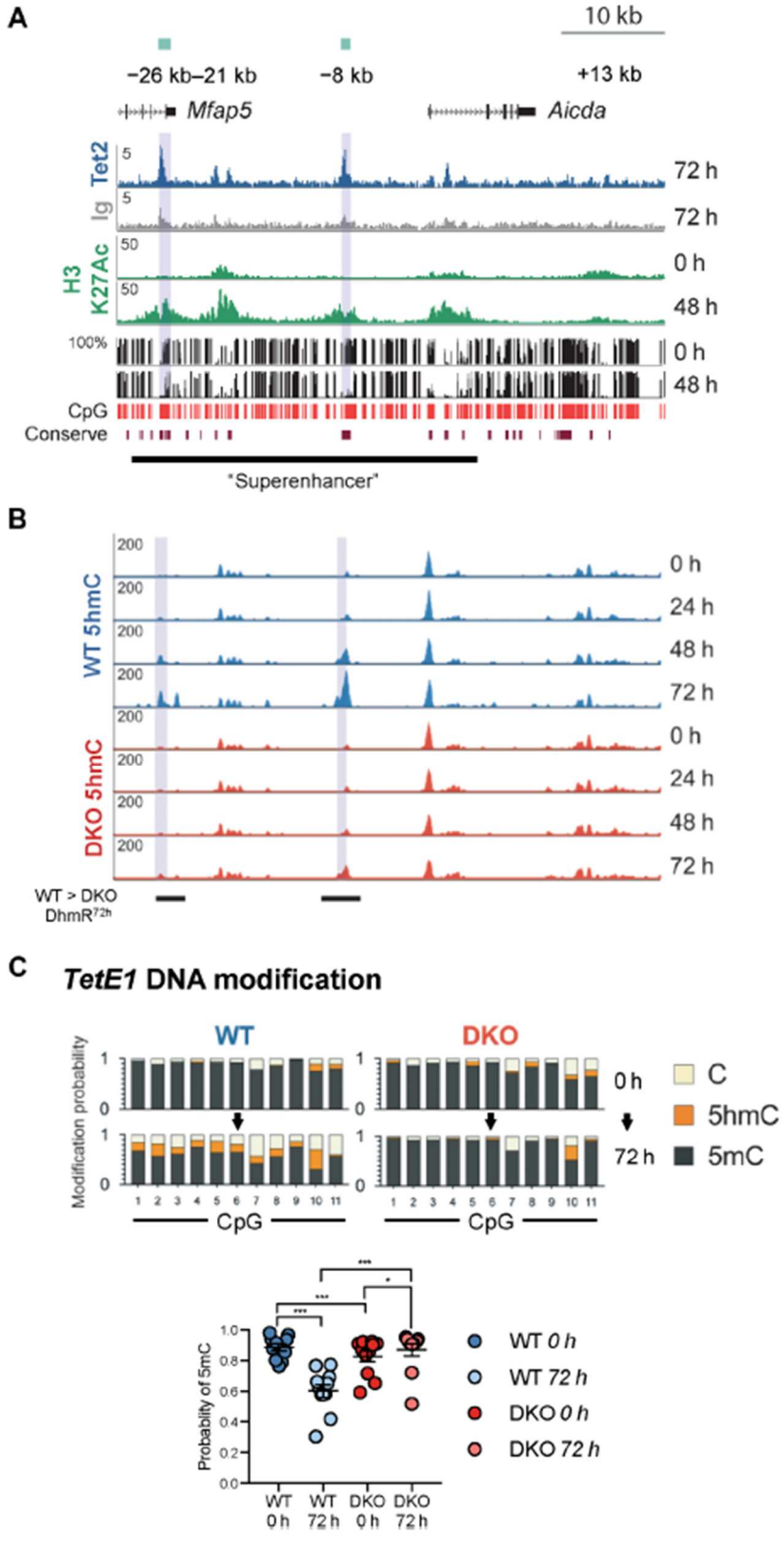


Figure 1.4. Tet2/3 facilitate CSR by regulating expression of the cytidine deaminase AID. **(A)** qRT-PCR analysis *Aicda* mRNA expression in WT and *Tet2/3*-DKO B cells activated 4 days with LPS and IL4. *Aicda* expression was normalized to *Gapdh* and then to the level in WT. Result shows ~50% decrease of *Aicda* mRNA expression in *Tet2/3* DKO relative to WT B cells. Data shown are representative of two independent experiments with three technical replicates. *, $p<0.05$. **(B)** Immunoblotting of whole cell lysates showed a substantial decrease of AID and Tet2 protein expression in *Tet2/3*-DKO relative to WT B cells activated for 4 days. Left lane contains lysate from the AID-KO CH12 B cells as a control for the specificity of anti-AID antibody. Beta-Actin was used as loading control. Data shown are representative of two independent experiments. See also **Fig. S1.4D. (C,D)** WT and *Tet2/3*-DKO B cells were retrovirally transduced with empty vector expressing Thy1.1 (*left panels*), wild-type AID (AID^{WT}, *middle panels*), or catalytically inactive AID (AID^{MUT}, *right panels*). Cells were gated on live transduced B cells ($CD19^+$ $Thy1.1^+$). Representative flow cytometry plots **(C)** and quantification **(D)** are shown. Data shown are representative of three independent experiments. n.s., not significant. ***, $p<0.01$. See also **Fig. S1.4**.

Figure 1.5. Tet2 and Tet3 control *Aicda* expression via TET-responsive elements *TetE1* and *TetE2*.

Diagram shows two conserved TET-responsive elements *TetE1* and *TetE2* at the 5' of the *Aicda* gene (labeled with green rectangles and grey shades). **(A)** *Top two tracks*. ChIP-seq analysis showed that Tet2 (*blue track*) specifically bound to multiple elements in the *Aicda* locus (mm10; chr6:122,523,500-122,576,500) after activation when compared to Ig control (*grey track*). *Middle tracks* (green). Increased H3K27 acetylation at the upstream and intronic regulatory elements of *Aicda* after activation. *Bottom tracks*. Activation induced DNA demethylation at *TetE1* and *TetE2*. Whole genome bisulfite sequencing (WGBS) showing DNA methylation (5mC+5hmC) in naïve and 48h-activated B cells (*mCG, black tracks*). CpGs included in the analysis are indicated by red lines (*red track*). Bottom track indicates the conserved DNA elements among placental animals ("Conserve"). Previously identified super-enhancer is indicated. For Tet2 and Ig, scales indicate per 10 million reads; for H3K27Ac, quantile-normalized reads; for bisulfite sequencing, percentage of bisulfite-resistant cytosine. **(B)** Activation induced Tet2/3-dependent 5hmC deposition at *Aicda* distal elements. WT and *Tet2/3*-DKO B cells were activated as in **Fig. 1.3G** with LPS and IL-4 as a function of time. DNA was purified and 5hmC enrichment was detected by CMS-IP (see *Materials and Methods*). Significant differential 5hmC-enriched regions between WT and DKO after 72h-activation were indicated at the bottom (WT>DKO DhmR^{72h}). Scales indicate quantile-normalized reads. **(C)** Tet2/3 deposit 5hmC and demethylate *Aicda* TET-responsive element *TetE1* and promoter. CpG modifications (5hmC, 5mC, and C) at *TetE1* (*top panels*) and *promoter* (*bottom panels*) were analyzed by oxidative bisulfite sequencing (oxBS-seq; *Materials and Methods*) using DNA isolated from WT and *Tet2/3*-DKO B cells before and after activation. Although 5hmC and 5mC can be distinguished by oxBS-seq, unmodified C and minuscule amount of fC and caC were recognized as "C", all of which are sensitive to deamination by bisulfite treatment. See also **Fig. S1.5**.


A
10 kb
−26 kb–21 kb
−8 kb
+13 kb
Mfap5
Aicda
Tet2
Ig
H3K27Ac
72 h
72 h
0 h
48 h
0 h
48 h
100%
CpG
Conserve
"Superenhancer"
B
WT 5hmC
DKO 5hmC
0 h
24 h
48 h
72 h
0 h
24 h
48 h
72 h
WT > DKO
DhmR72h
C
TetE1 DNA modification
WT
DKO
Modification probability
0 h
72 h
C
5hmC
5mC
CpG
CpG
Probability of 5mC
WT 0 h
WT 72 h
DKO 0 h
DKO 72 h

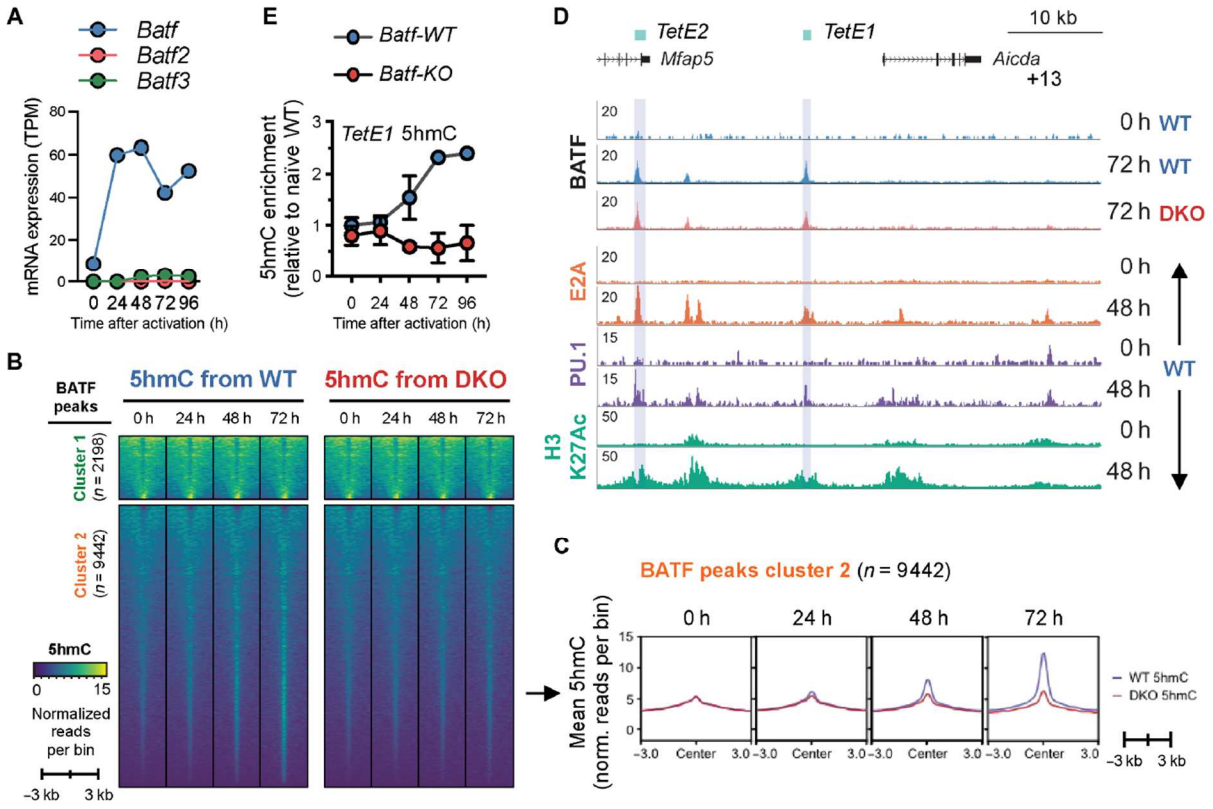


Figure 1.6. BATF facilitates TET protein-mediated hydroxymethylated at *TetE1*.
(A) Mean mRNA expression level of *Batf* family members (*Batf1-3*) in B cells activated with LPS and IL-4 as a function of time. Data shown are from RNA-seq with two independent replicates. TPM, transcript per million. **(B,C) BATF binding correlates with 5hmC-enrichment.** WT BATF peaks (n=11,640) were divided into two clusters based on the pattern of 5hmC distribution. **(B)** Cluster 1 (n=2,198; *top panels*) showed a broad 5hmC distribution, with the 5hmC level remained unchanged after activation and in the absence of Tet2/3 (*upper panels, compare "5hmC from WT" to "5hmC from* DKO*"*). In contrast, a substantial portion of regions in cluster 2 (n=9,422) showed a progressive Tet-dependent 5hmC modification after activation (*lower panels*) and is further illustrated in **(C)** as line plots. Data shown are mean enrichment per 100 bp bin. **(D)** Recruitment of BATF and other transcription factors to *Aicda* enhancers. *Upper three tracks*, genome browser view of BATF-binding in unstimulated and 72h-activated WT (*blue*) and *Tet2/3*-DKO B cells (*red*) at the *Aicda* locus. Note that the major BATF-binding sites are at *TetE1* and *TetE2,* and the loss of *Tet2/3* has no significant effect on BATF recruitment (compare WT and DKO; also see **Fig. S1.8G**; two independent experiments). Activation also induced E2A and PU.1 binding to *Aicda* enhancers (orange and purple tracks). Coordinate for locus is chr6:122,523,500-122,576,500 (mm10). See also **Fig. S1.8**. **(E)** BATF is required for 5hmC modification at TetE1. Batf-WT and Batf-KO B cells were activated with LPS and IL-4 for 4 days. 5hmC modification at TetE1 was quantified using AbaSI-qPCR.

1.7 Supplemental Figures

Figure S1.1. TET-mediated DNA hydroxymethylation correlates with demethylation and enhancer activity.
(A) Similar total numbers of 5hmC-enriched regions between naïve and activated B cells. **(B)** Box-and-whisker plot showing that differentially hydroxymethylated regions (DhmRs) in activated vs naïve B cells (see **Fig. 1.1C**) are located on average more than 10 kb from the closest transcription start site (TSS). **(C)** Heatmaps showing the kinetics of 5hmC modification at the 1,121 regions with decreased 5hmC in 72h-activated vs naïve B cells ($DhmR^{72h\text{-}down}$; see **Fig. 1.1C**). 5hmC enrichment is shown as normalized reads per 100 bp bin. **(D)** *Left*, the vast majority of differentially methylated regions (DMRs) with altered WGBS signal (5mC+5hmC) in naïve vs 48h-activated B cells show decreased DNA methylation. *Right*, plot of average DNA methylation (bisulfite-resistant cytosine 5mC+5hmC) at the $DMR^{48h\text{-}down}$ (n=1,097) in 48h-activated vs naïve B cells. Average methylation is measured per 200 bp bin. **(E)** Heatmaps showing the kinetics of change (increase) in 5hmC at the 1,097 $DMR^{48h\text{-}down}$ with decreased methylation 48h post-activation. 5hmC enrichment is shown in normalized reads per 100 bp bin. **(F)** Heatmap showing decreased DNA methylation at the 8454 $DhmR^{72h\text{-}up}$ regions. **(G)** Strong enrichment for consensus binding motifs of NFκB, IRF:bZIP, bZIP, and other transcription factors in the 1,097 regions that became demethylated after 48h of activation ($DMR^{48h\text{-}down}$). Random genomic regions were used as background for motif analysis. Y-axis indicates the fold enrichment versus background, circle size indicates the percentage of regions containing the respective motif, and the color indicates the significance (Log_{10} *p* vaule). **(H)** 5hmC levels track with enhancer activity. *Left*, MA plot showing differentially active enhancers between naïve and 48h-activated B cells were classified based on the significant difference in H3K27Ac and accessibility (ATAC-seq) into enhancers preferentially active in naïve B cells (green, "Naïve>48h") and enhancers preferentially active in 48 h-activated B cells (orange, "48>naïve"). The remaining enhancers not meeting the above criteria were classified as common active enhancers (grey, "Common"). *Right*, mean level of 5hmC per bin (50 bp) at the +/- 10kb interval to the center was plotted for each type of active enhancer. Note that the 5hmC levels from the "Common" enhancers (*middle)* are also plotted as dotted lines for naïve-B-active (*top*) and activated-B-active enhancers (*bottom*) as reference.

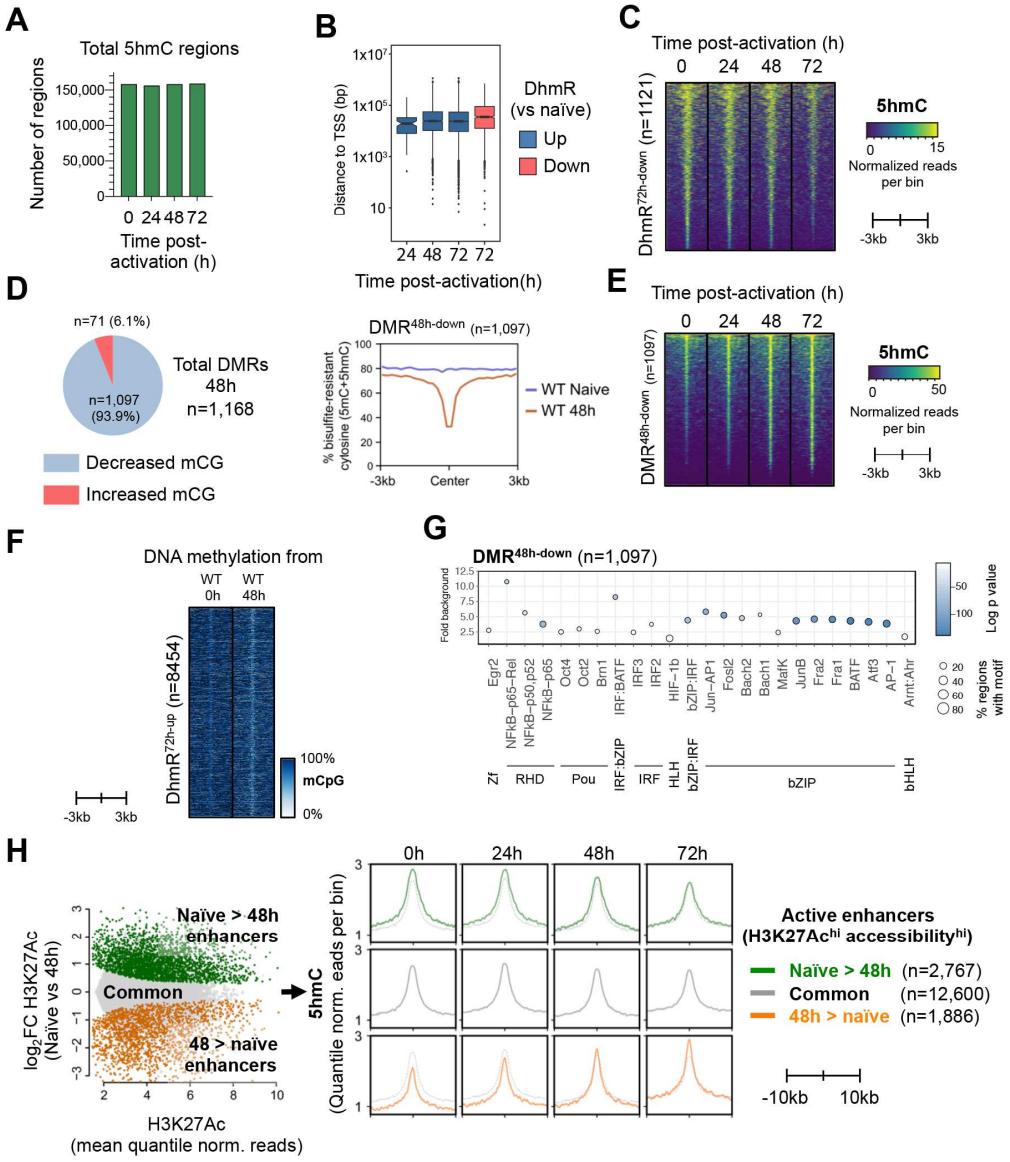

A
Total 5hmC regions
Number of regions
150,000
100,000
50,000
0
0 24 48 72
Time post-activation (h)
B
Distance to TSS (bp)
1x10^7
1x10^5
1x10^3
10
24 48 72 72
DhmR (vs naïve)
Up
Down
Time post-activation(h)
C
Time post-activation (h)
0 24 48 72
DhmR72h-down (n=1121)
5hmC
0 15
Normalized reads per bin
-3kb 3kb
D
n=71 (6.1%)
n=1,097 (93.9%)
Total DMRs 48h n=1,168
Decreased mCG
Increased mCG
DMR48h-down (n=1,097)
% bisulfite-resistant cytosine (5mC+5hmC)
100
80
60
40
20
0
-3kb Center 3kb
WT Naive
WT 48h
E
Time post-activation (h)
0 24 48 72
DMR48h-down (n=1097)
5hmC
0 50
Normalized reads per bin
-3kb 3kb
F
DNA methylation from
WT 0h
WT 48h
DhmR72h-up (n=8454)
100%
mCpG
0%
-3kb 3kb
G
DMR48h-down (n=1,097)
Fold background
12.5
10.0
7.5
5.0
2.5
Log p value
-50
-100
Egr2
NFkB–p65–Rel
NFkB–p50,p52
NFkB–p65
Oct4
Oct2
Brn1
IRF:BATF
IRF3
IRF2
HIF–1b
bZIP:IRF
Jun–AP1
Fosl2
Bach2
Bach1
MafK
JunB
Fra2
Fra1
BATF
Atf3
AP–1
Arnt:Ahr
% regions with motif
20
40
60
80
Zf
RHD
Pou
IRF:bZIP
IRF
HLH
bZIP:IRF
bZIP
bHLH
H
log2FC H3K27Ac (Naïve vs 48h)
Naïve > 48h enhancers
Common
48 > naïve enhancers
H3K27Ac (mean quantile norm. reads)
5hmC
0h
24h
48h
72h
(Quantile norm. eads per bin)
Active enhancers (H3K27Achi accessibilityhi)
Naïve > 48h (n=2,767)
Common (n=12,600)
48h > naïve (n=1,886)
-10kb 10kb

Figure S1.2. Phenotypic features of WT and *Tet2/3* DKO B cells.
(A) Comparable splenic mature B cell populations in *Tet2/3*-conditionally deleted mice. WT (*Tet2/3-flox Rosa26-LSL-YFP*) and DKO (*Cre^{ERT2} Tet2/3-flox Rosa26-LSL-YFP*) mice were treated as in **Fig. 1.2B** and the phenotype of splenic B cells were analyzed on day 7 after the initial tamoxifen injection. Plots were first gated on live single cells based on FSC/SSC (*first panel*) and total (WT) or YFP$^+$ (DKO) CD19$^+$ B cells were subsequently gated (*second panel*), followed by analysis of mature and immature B cells (*third panel*); and follicular (FO) and marginal zone (MZ) B cells (fourth panel). **(B)** Similar percentages of YFP$^+$ cells in total (CD19$^+$; middle panel) and mature (CD19$^+$ AA4.1^{lo}) B cells. **(C)** Total 5hmC levels in WT and *Tet2/3* DKO B cells assessed by cytosine 5-methylenesulphonate (CMS) dot blot (see *Materials and Methods*). Note that 5hmC levels decrease in *Tet2/3* DKO B cells only after several rounds of cell division (>48h). **(D)** Histograms showing the distance from the TSS to the TET-regulated DhmR regions differentially marked with 5hmC in 72h-activated *Tet2/3* DKO relative to WT B cells (see **Fig. 1.2D, E**). The 2,139 and 184 DhmRs with decreased (*left*, DhmR^{72h} WT>DKO) and increased (*right*, DhmR^{72h} DKO>WT) 5hmC after activated for 72h 72h-activated *Tet2/3* DKO relative to WT B cells are located on average more than 10 kb from the closest TSS. **(E) TET-mediated 5hmC modifications mark DNA demethylation.** *Left panels*, heatmaps show DNA methylation status in naïve and 48h-activated WT B cells (WGBS, 5mC+5hmC) at the 2,139 and 184 DhmR regions with decreased (*top*, WT>DKO) and increased (*bottom*, DKO>WT) 5hmC in 72h-activated *Tet2/3* DKO vs WT B cells. *Right panels*, plots of the average decrease in bisulfite-resistant modifications (5mC+5hmC) per bin (200 bp) at these regions. The majority of the 2,139 WT>DKO DhmRs at 72h show decreased DNA methylation in activated WT B cells (*top*); the 184 DKO>WT DhmR at 72h are fully methylated and unchanged in *Tet2/3* DKO vs WT (*bottom*).

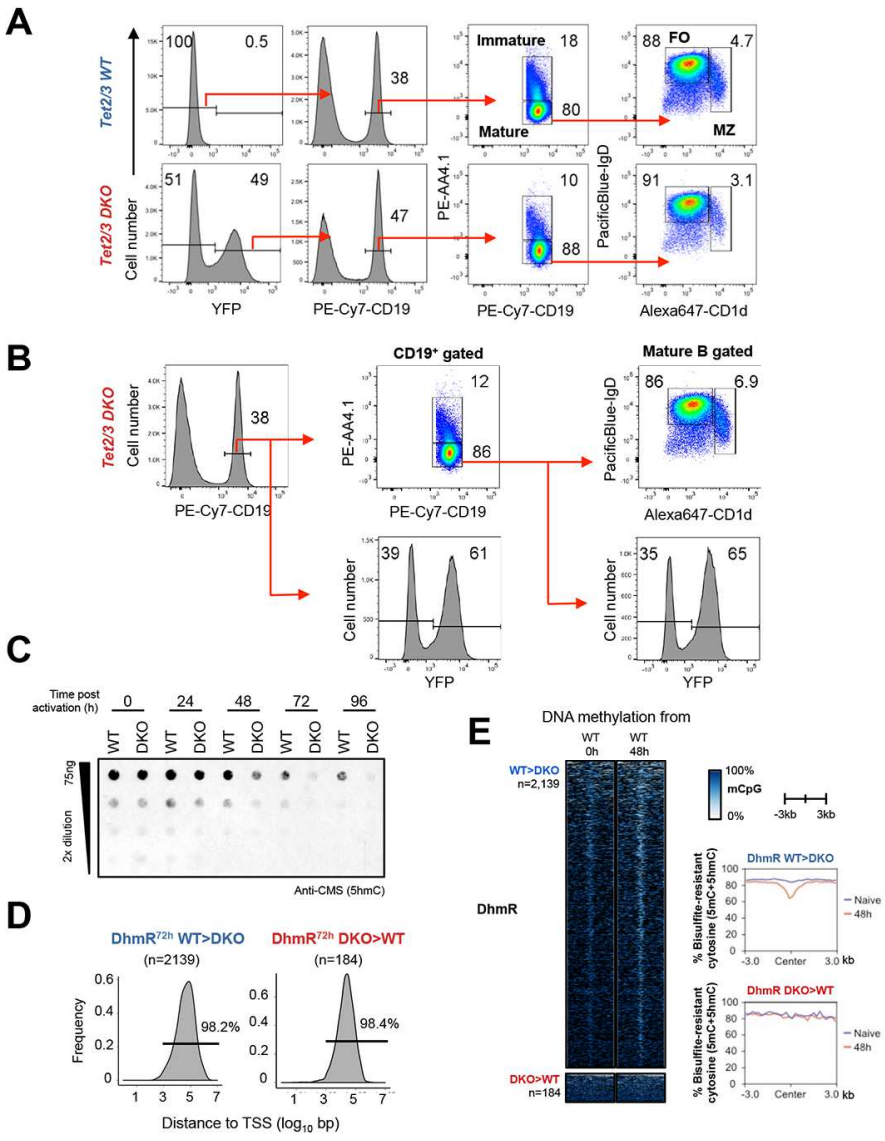

A
Tet2/3 WT
Tet2/3 DKO
Cell number
YFP
PE-Cy7-CD19
PE-AA4.1
PacificBlue-IgD
Alexa647-CD1d
Immature
Mature
FO
MZ
B
CD19+ gated
Mature B gated
C
Time post activation (h)
0
24
48
72
96
WT
DKO
75ng
2x dilution
Anti-CMS (5hmC)
D
DhmR72h WT>DKO
(n=2139)
DhmR72h DKO>WT
(n=184)
Frequency
98.2%
98.4%
Distance to TSS (log10 bp)
E
DNA methylation from
WT 0h
WT 48h
WT>DKO n=2,139
DhmR
DKO>WT n=184
100% mCpG
0%
-3kb
3kb
DhmR WT>DKO
DhmR DKO>WT
% Bisulfite-resistant cytosine (5mC+5hmC)
Naive
48h
Center
3.0 kb

Figure S1.3. TET family proteins are important for B-cell-intrinsic CSR.

(A) Total cell number in draining lymph nodes (left), percentage of $CD19^+$ cells (middle), and number of $CD19^+$ B cells (right) from **Fig. 1.3A**. *, *p<0.05*. **(B)** Level of GL7 (geometric mean of fluorescence intensity, MFI) on WT and DKO GC B cells. *, *p<0.05*. **(C)** CSR defect is not caused by Cre activity. Cre^{ERT2} Rosa26-LSL-YFP (*Cre^{ERT2} WT*) and *Cre^{ERT2} $Tet2^{fl/fl}$ $Tet3^{fl/fl}$ Rosa26-LSL-YFP* (*Cre^{ERT2}* DKO) mice were injected with tamoxifen as in **Fig. 1.2.B**. Isolated B cells were activated with LPS and IL-4 in the presence of 4-hydroxytamoxifen for 4 days and $\%IgG1^+$ cells were analyzed (gated on live $CD19^+$ YFP^+). One representative of two experiments is shown. n=2 for each genotype. *, p<0.05. **(D-E) (D)** CSR defect is cell-intrinsic. *(Left) $Tet2^{+/+}$ $Tet3^{+/+}$* WT CD45.1 and *Tet2/3*-DKO CD45.2 mice were treated as in **Fig. 1.3A** and isolated B cells were labeled with Cell-Trace violet, 1:1 mixed, and activated with LPS and IL-4 for 4 days. *(Right)* Cells were gated based on CD45.1 and CD45.2 and the percentages of IgG1-switched cells in WT and DKO are shown. Cells from the same well are connected with lines. **(E)** Co-cultured WT and *Tet2/3*-DKO B cells showed similar proliferation profiles. Data shown are representative of two independent experiments with four technical replicates for each genotype. **(F)** Percentage of $CD138^+$ cells in the non-switched ($IgG1^-$ IgA^-) population from **Fig. 1.3K-3N**. **(G-I)(G)** Flow chart of experiment to assess the importance of TET catalytic activity in CSR. **(H)** Flow cytometry plots and **(I)** quantification of WT and *Tet2/3* DKO B cells transduced with empty vector (*left panels*), TET2 wild-type catalytic domain (Tet2CD, *middle panels*), and Tet2 HxD mutant catalytic domain ($Tet2CD^{HxD}$, *right panels*) shows that TET catalytic activity can partly rescue the CSR to IgG1. Data shown are representative of two independent experiments with two technical replicates. n.s., not significant. ***, *p<0.01*. *, *p<0.05*. **(J-K**) Deletion of all three TET proteins (*Tet1/2/3 TKO*) results in a similar decrease in CSR as deletion of Tet2 and Tet3 (*Tet2*/3-DKO). **(J)** *Tet1/2/3-flox Cre^{ERT2} Rosa26-LSL-YFP* (TKO) and control *Tet1/2/3-flox Rosa26-LSL-YFP* (WT) mice were treated with tamoxifen and immunized with NP-OVA as in **Fig. 1.3A** and GC response and CSR were analyzed on day 7. Flow cytometry plots showed the percentage of GCB ($CD38^{lo}GL7^{hi}$) in WT and *Tet1/2/3-TKO* lymph node cells gated on total (WT) and YFP^+ (TKO) live $CD19^+$ B cells (*left panels)*. Antigen-specific (NP-PE) and class-switched cells (IgG1) were analyzed among GC B cells. **(K)** Quantification of the experiments showed in **(I)**. Data shown are aggregated results from two independent experiments. WT, n=7; TKO, n=7. Statistical significance was calculated using unpaired two-tailed *t*-test. n.s., not significant. ***, *p<0.01*. *, *p < 0.05*.

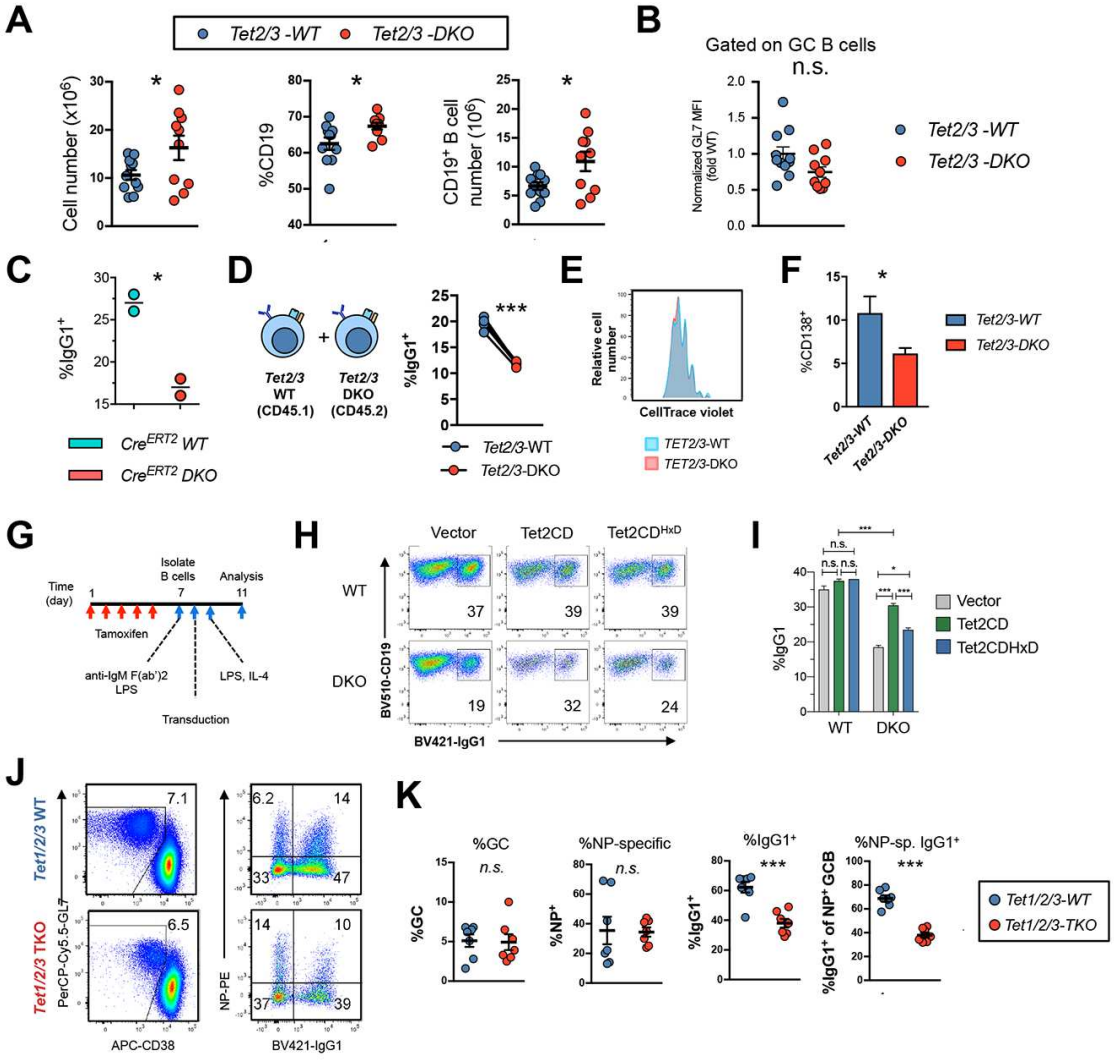

A
Tet2/3 -WT
Tet2/3 -DKO
Cell number (x10⁶)
%CD19
CD19⁺ B cell number (10⁶)
B
Gated on GC B cells
n.s.
Normalized GL7 MFI (fold WT)
Tet2/3 -WT
Tet2/3 -DKO
C
%IgG1⁺
CreERT2 WT
CreERT2 DKO
D
Tet2/3 WT (CD45.1)
Tet2/3 DKO (CD45.2)
%IgG1⁺
Tet2/3-WT
Tet2/3-DKO
E
Relative cell number
CellTrace violet
TET2/3-WT
TET2/3-DKO
F
%CD138⁺
Tet2/3-WT
Tet2/3-DKO
G
Time (day)
Isolate B cells
Analysis
Tamoxifen
anti-IgM F(ab')2
LPS
LPS, IL-4
Transduction
H
Vector
Tet2CD
Tet2CDHxD
WT
DKO
BV510-CD19
BV421-IgG1
I
%IgG1
Vector
Tet2CD
Tet2CDHxD
WT
DKO
J
Tet1/2/3 WT
Tet1/2/3 TKO
PerCP-Cy5.5-GL7
APC-CD38
NP-PE
BV421-IgG1
K
%GC
%NP-specific
%IgG1⁺
%NP-sp. IgG1⁺
%NP⁺
%IgG1⁺ of NP⁺ GCB
Tet1/2/3-WT
Tet1/2/3-TKO

Figure S1.4. Decreased *Aicda* expression in *Tet2/3*-DKO B cells.
WT and *Tet2/3*-DKO B cells were activated as in **Fig. 1.3G** and the transcriptomes were analyzed by RNA-seq (see *Materials and Methods* for details). **(A)** Number of differentially expressed genes between WT and *Tet2/3*-DKO B cells as a function of time after activation. Relatively few genes show alterations in their expression. **(B)** List of all differentially expressed genes between WT and *Tet2/3*-DKO B cells. *Aicda* (indicated by red arrows) was one of the genes expressed at significantly lower levels in DKO B cells at all time points analyzed. Color scale indicates Log_2 fold change between WT and DKO. **(C-D)** TET2 and TET3 are required for potent *Aicda* expression. *Aicda* mRNA **(C)** and protein **(D)** expression were analyzed by qRT-PCR and western blot as a function of time after activation. Results show increased AID expression with time after activation of WT B cells, and a consistent decrease in *Tet2/3* DKO relative to WT B cells. **(E-J)** Haploinsufficiency of *Aicda* results in decreased CSR. Mice with the indicated genotypes were immunized with 10 μg of NP-OVA mixed with Alum via footpad injection, and the draining lymph nodes were analyzed by flow cytometry at day 7 post-immunization. Heterozygous *Aicda-Cre* mice were used to model *Aicda* haploinsufficiency as the knocked-in Cre recombinase disrupted the endogenous *Aicda* expression. Representative flow cytometric analysis of **(E)** germinal center B cells (GCB; $CD38^{lo}$ $GL7^{hi}$) and **(F)** CSR to IgG1. **(G-J)** Statistical analyses of the populations (means and standard errors) are shown (n=4 each). Data are representative of two independent experiments. Unpaired two-tailed *t*-test was used to calculate statistical significance and the *p* values are indicated. **(K)** TET2 and TET3 are not required for expression of germline transcripts. WT and DKO B cells were activated for 4 days, and *m* and *g*1 germline transcripts were analyzed by qRT-PCR. Data were normalized to *Gapdh* and to WT level as in **Fig. 1.4A**. n.s., not significant.

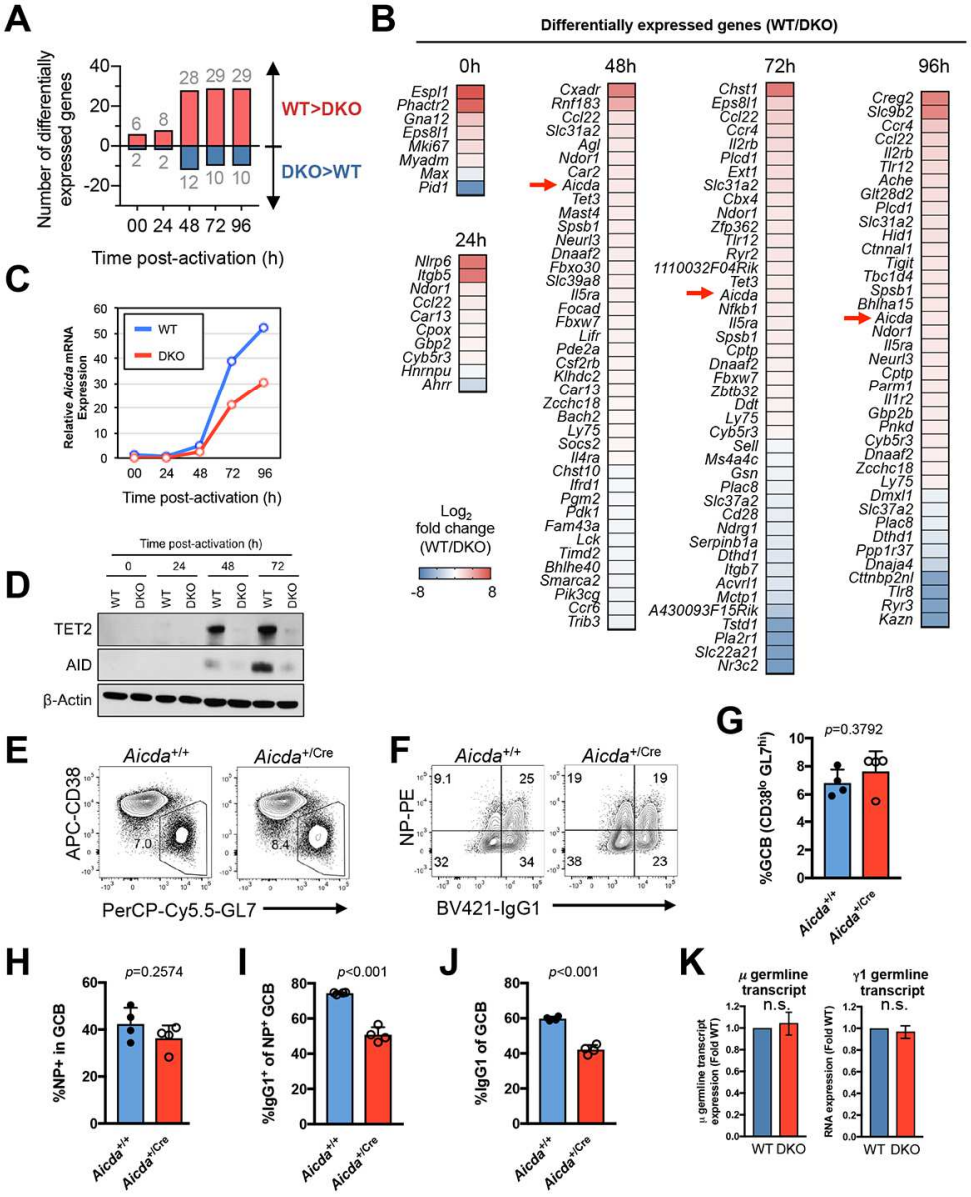

A
Number of differentially expressed genes
40
20
0
-20
6
8
28
29
29
2
2
12
10
10
WT>DKO
DKO>WT
00 24 48 72 96
Time post-activation (h)
B
Differentially expressed genes (WT/DKO)
0h
Espl1
Phactr2
Gna12
Eps8l1
Mki67
Myadm
Max
Pid1
24h
Nlrp6
Itgb5
Ndor1
Ccl22
Car13
Cpox
Gbp2
Cyb5r3
Hnrnpu
Ahrr
48h
Cxadr
Rnf183
Ccl22
Slc31a2
Agl
Ndor1
Car2
Aicda
Tet3
Mast4
Spsb1
Neurl3
Dnaaf2
Fbxo30
Slc39a8
Il5ra
Focad
Fbxw7
Lifr
Pde2a
Csf2rb
Klhdc2
Car13
Zcchc18
Bach2
Ly75
Socs2
Il4ra
Chst10
Ifrd1
Pgm2
Pdk1
Fam43a
Lck
Timd2
Bhlhe40
Smarca2
Pik3cg
Ccr6
Trib3
72h
Chst1
Eps8l1
Ccl22
Ccr4
Il2rb
Plcd1
Ext1
Slc31a2
Cbx4
Ndor1
Zfp362
Tlr12
Ryr2
1110032F04Rik
Tet3
Aicda
Nfkb1
Il5ra
Spsb1
Cptp
Dnaaf2
Fbxw7
Zbtb32
Ddt
Ly75
Cyb5r3
Sell
Ms4a4c
Gsn
Plac8
Slc37a2
Cd28
Ndrg1
Serpinb1a
Dthd1
Itgb7
Acvrl1
Mctp1
A430093F15Rik
Tstd1
Pla2r1
Slc22a21
Nr3c2
96h
Creg2
Slc9b2
Ccr4
Ccl22
Il2rb
Tlr12
Ache
Glt28d2
Plcd1
Slc31a2
Hid1
Ctnnal1
Tigit
Tbc1d4
Spsb1
Bhlha15
Aicda
Ndor1
Il5ra
Neurl3
Cptp
Parm1
Il1r2
Gbp2b
Pnkd
Cyb5r3
Dnaaf2
Zcchc18
Ly75
Dmxl1
Slc37a2
Plac8
Dthd1
Ppp1r37
Dnaja4
Cttnbp2nl
Tlr8
Ryr3
Kazn
Log2 fold change (WT/DKO)
-8
8
C
Relative Aicda mRNA Expression
WT
DKO
Time post-activation (h)
D
Time post-activation (h)
0
24
48
72
WT
DKO
TET2
AID
β-Actin
E
Aicda+/+
Aicda+/Cre
APC-CD38
7.0
8.4
PerCP-Cy5.5-GL7
F
Aicda+/+
Aicda+/Cre
NP-PE
9.1
25
32
34
19
19
38
23
BV421-IgG1
G
p=0.3792
%GCB (CD38lo GL7hi)
H
p=0.2574
%NP+ in GCB
I
p<0.001
%IgG1+ of NP+ GCB
J
p<0.001
%IgG1 of GCB
K
μ germline transcript
n.s.
γ1 germline transcript
n.s.
μ germline transcript expression (Fold WT)
RNA expression (Fold WT)
WT DKO

Figure S1.5. The TET-responsive element *TetE1* regulates CSR and *Aicda* mRNA expression in the CH12 B cell.

(A) Diagram depicts the relative position of TET-responsive elements *TetE1* and *TetE2* to previously identified *Aicda* distal and intronic enhancers. "Region" IV, II, III are from Tran et al. 2010; "CNS" V-X are from Crouch et al. 2007; "Enhancer" E1-E5 from Kieffer-Kwon et al. 2013. Note that the promoter-proximal element is not depicted. Coordinates for the shown locus are chr6:122,523,500-122,576,500 (mm10). **(B-E**) *TetE1* is important for regulating *Aicda* expression and CSR. **(B)** Scheme for *TetE1* deletion in CH12 cells with CRISPR. **(C)** Four clones were identified with homozygous deletion of *TetE1* as examined by PCR followed by gel electrophoresis. A clone with heterozygous deletion (Het) and a WT control are shown as controls. **(D-E)** WT and TetE1-deletion clones were stimulated with CIT (anti-CD40, IL-4, TGFbeta) for two days. **(D)** *Aicda* mRNA expression and **(E)** CSR to IgA were analyzed by qRT-PCR and flow cytometry, respectively. Results show that deletion of *TetE1* decreased *Aicda* mRNA expression and abrogated CSR. **(F)** 5hmC modification at the IgH locus. Genome browser view of the IgH locus (chr12:113,211,000-113,445,000; mm10) showing H3K27Ac in unstimulated and 48h activated B cells (*top two panels*), followed by DNA methylation in unstimulated and 48h-activated B cells (*black histograms*), CpG covered in analysis (*red histograms*), and 5hmC modifications in WT and DKO as indicated. Regions with increased 5hmC modification after activation ($DhmR^{72h\text{-}up}$) are indicated by horizontal bars. **(G)** Chromatin accessibility at the IgH locus. Genome browser view of ATAC-seq data from activated WT and DKO B cells. There is no statistically significant difference in chromatin accessibility between the two genotypes. **(H)** DNA modification at the *Aicda* promoter. *Top*, CpG modifications (5hmC, 5mC, and C) *Aicda* promoter were analyzed by oxBS-seq as in **Fig. 1.5C**. *Bottom*, the overall CpG methylation probability was quantified. Methylation at the *Aicda* promoter at 72 h was significantly increased in DKO compared to WT. *, $p < 0.05$. n.s., not significant.

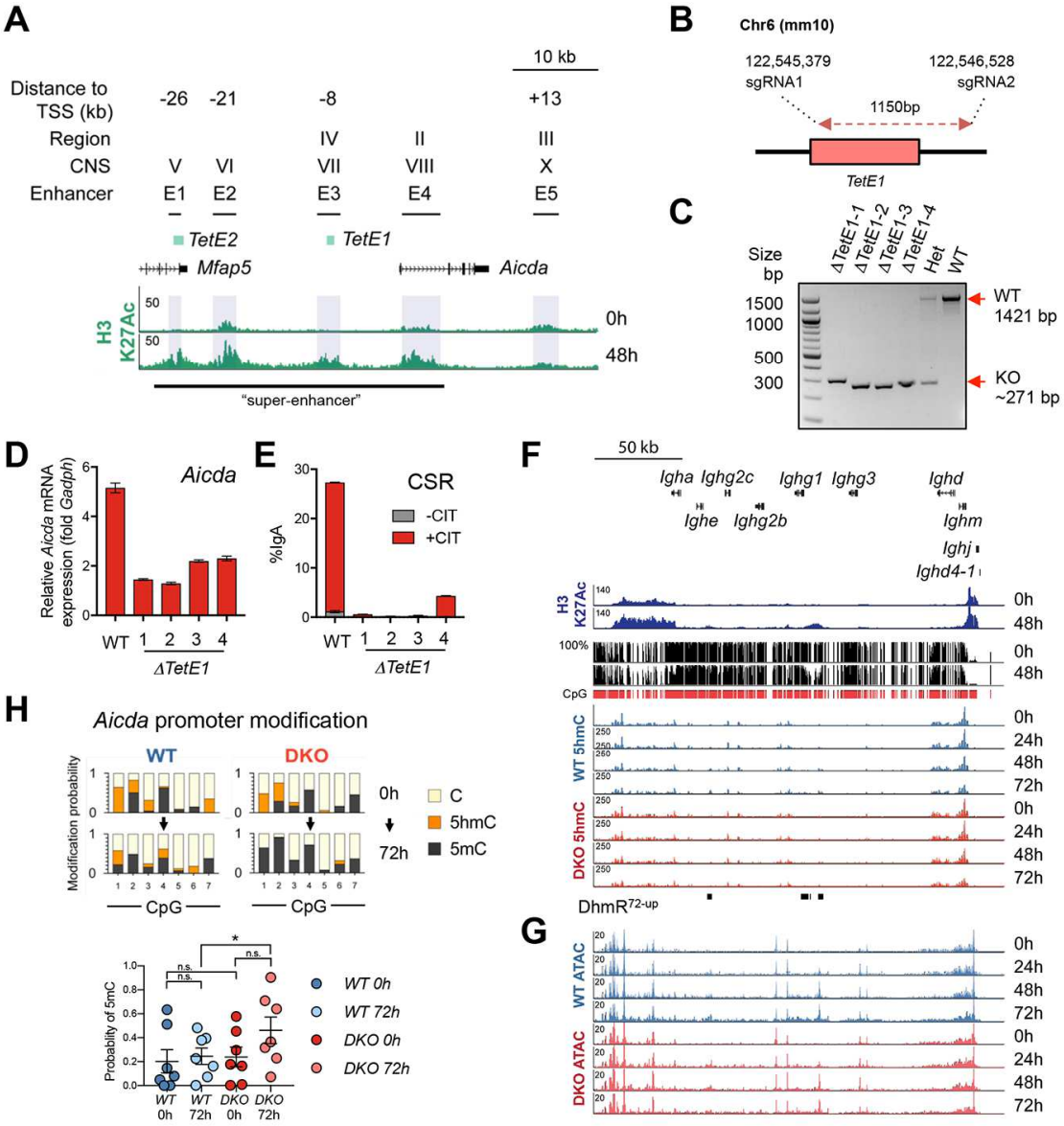

A
10 kb
Distance to TSS (kb)
-26
-21
-8
+13
Region
IV
II
III
CNS
V
VI
VII
VIII
X
Enhancer
E1
E2
E3
E4
E5
TetE2
TetE1
Mfap5
Aicda
H3K27Ac
50
50
0h
48h
"super-enhancer"
B
Chr6 (mm10)
122,545,379
sgRNA1
122,546,528
sgRNA2
1150bp
TetE1
C
Size bp
ΔTetE1-1
ΔTetE1-2
ΔTetE1-3
ΔTetE1-4
Het
WT
1500
1000
500
300
WT
1421 bp
KO
~271 bp
D
Aicda
Relative Aicda mRNA expression (fold Gadph)
WT
1
2
3
4
ΔTetE1
E
CSR
%IgA
-CIT
+CIT
WT
1
2
3
4
ΔTetE1
F
50 kb
Igha
Ighg2c
Ighg1
Ighg3
Ighd
Ighe
Ighg2b
Ighm
Ighj
Ighd4-1
H3K27Ac
0h
48h
100%
0h
48h
CpG
WT 5hmC
0h
24h
48h
72h
DKO 5hmC
0h
24h
48h
72h
DhmR72-up
G
WT ATAC
0h
24h
48h
72h
DKO ATAC
0h
24h
48h
72h
H
Aicda promoter modification
WT
DKO
Modification probability
0h
72h
C
5hmC
5mC
CpG
CpG
Probability of 5mC
n.s.
n.s.
*
n.s.
WT 0h
WT 72h
DKO 0h
DKO 72h
WT 0h
WT 72h
DKO 0h
DKO 72h

Figure S1.6. Tet proteins sustain enhancer accessibility.
(A) B cell activation induces global changes in chromatin accessibility. WT B cells were activated with LPS and IL-4 and chromatin accessibility was profiled by ATAC-seq at different times. Numbers indicate differentially accessible regions (DARs) between activated (act.) and naïve B cells with FDR < 0.05 and fold change above $\log_2(1.5)$ or below $\log_2(0.67)$. **(B)** Loss of TET proteins results in decreased chromatin accessibility at later time points. Numbers of DARs between WT and *Tet2/3*-DKO B cells activated for different times are shown. The difference between WT and DKO B cells was minimal at time points earlier than 72h. DARs were selected based on FDR < 0.05 and fold change above $\log_2(1.5)$ or below $\log_2(0.67)$. **(C)** Tet2/3-dependent accessible regions are hydroxymethylated. Heatmaps show the kinetics of 5hmC modification at differentially accessible regions (DARs) between WT and *Tet2/3*-DKO B cells. Regions that are more accessible in WT (WT>DKO, n=292), less accessible in WT (DKO>WT, n=129), and commonly accessible (n=27,716) are shown in the top, middle, and bottom panels, respectively. WT>DKO DARs show progressive 5hmC enrichment only in WT (top left panels) but not in DKO (top right panels) B cells, demonstrating that 5hmC modification at these regions is *Tet2/3*-dependent. The DKO>WT DARs (n=129) and common regions (n=27,716) show no apparent difference between naïve and activated B cells and between 5hmC from WT and DKO B cells. 5hmC enrichment is shown as normalized reads per 100 bp bin. **(D)** TET2 and TET3 maintain chromatin accessibility at the *Aicda* Tet-responsive elements *TetE1* and *TetE2*. Genome browser view of ATAC-seq data showing the accessibility profile at *Aicda* locus in WT (blue, top 4 tracks) and DKO (red, bottom 4 tracks) B cells. Note that *TetE1* and *TetE2* were among the DAR at 72 h (DAR^{72h} WT>DKO) as indicated at the bottom. Coordinates for the *Aicda* locus are chr6:122,523,500-122,576,500 (mm10). **(E)** Plot of mean chromatin accessibility at the DARs between WT and *Tet2/3*-DKO B cells after 72h of activation (as in **(C)** top and middle panels). Top panels, WT>DKO DARs (n=292); bottom panels, DKO>WT DARs (n=129). Y-axes indicate the mean ATAC signals (normalized ATAC-seq reads per 100 bp bin) from WT (blue line) and DKO (red line) B cells activated as indicated. The difference between WT and DKO is apparent at 72h. See also **Fig. S1.7**.

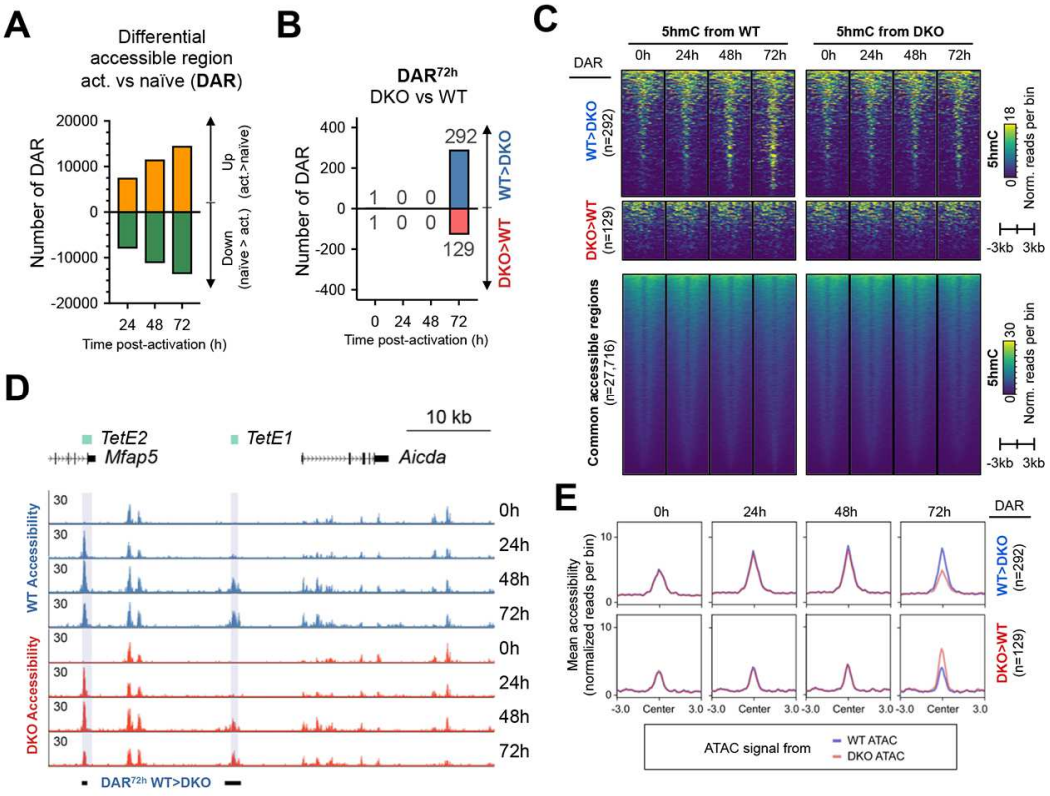

A
Differential accessible region act. vs naïve (DAR)
Number of DAR
Time post-activation (h)
Up (act.>naive)
Down (naive > act.)
B
DAR72h DKO vs WT
Number of DAR
Time post-activation (h)
WT>DKO
DKO>WT
C
5hmC from WT
5hmC from DKO
DAR
WT>DKO (n=292)
DKO>WT (n=129)
Common accessible regions (n=27,716)
5hmC
Norm. reads per bin
D
10 kb
TetE2
TetE1
Mfap5
Aicda
WT Accessibility
DKO Accessibility
DAR72h WT>DKO
E
Mean accessibility (normalized reads per bin)
Center
ATAC signal from
WT ATAC
DKO ATAC

Figure S1.7. Analysis of TET-dependent accessible regions.
(A) Correlation between 5hmC and accessibility. Comparison of mean chromatin accessibility (ATAC-seq) between WT and DKO B cells is shown for activation-induced 5hmC-enriched regions ($DhmR^{72h\text{-}up}$, n=8454, **Fig. 1.1C**). Regions with increased 5hmC in WT cells after activation also show increased accessibility (blue, ATAC-WT). DKO cells show decreased accessibility, as shown for the *Aicda* locus in **Fig. S1.6D**. Statistical significance between WT and DKO at each time point was calculated by Kolmogorov-Smirnov test with Bonferroni correction using the family-wise error rate. n.s., not significant. ***, p adj. < 0.01. The exact adjusted *p* values are 0.06, 6.01e-05, 1.79e-07, 3.611e-11 for 0h, 24h, 48h, 72h, respectively. **(B)** TET facilitates Increased accessibility at distal elements. Histograms showing the distance of DARs (WT>DKO and DKO>WT) and commonly accessible regions (Common) from the closest TSS. Majority of the Tet-facilitated accessible regions (WT>DKO, n=292) are distal elements (>1000bp; 92.8%). **(C)** TET2/3-dependent accessible regions are hydroxymethylated. Line plots showing the kinetics of mean 5hmC modification at differentially accessible regions (DARs) between WT and *Tet2/3* DKO for the data depicted in **Fig. S1.6C**. Regions that are more accessible in WT (WT>DKO, n=292), less accessible in WT (DKO>WT, n=129), and commonly accessible (n=27,716) are shown in the top, middle, and bottom panels, respectively. 5hmC enrichment is shown as normalized reads per 100 bp bin. Note the Tet-dependent 5hmC modification at these DARs (compare 72h panels). **(D)** TET-facilitated accessible regions are further demethylated after activation. Left, heatmaps showing the DNA modification status (5mC+5hmC) in naïve and 48h-stimulated WT B cells at WT>DKO DARs (i.e. Tet-facilitated accessible regions; n=292) and DKO>WT DARs (n=129). *Right*, plots summarizing the data in the heatmaps; the y-axis indicates the level of bisulfite-resistant cytosine (5mC+5hmC). In WT B cells, regions that lose accessibility in *Tet2/3* DKO B cells relative to WT (WT>DKO) also show a decrease in modification (mostly 5mC) after activation. **(E)** Enrichment for consensus IRF:bZIP (IRF:BATF) and bZIP transcription factor binding motifs in the 292 Tet-facilitated accessible regions, which show increased accessibility in WT relative to *Tet2/3* DKO B cells at 72h. No significant motif enrichment was detected at DKO>WT DARs (n=129). Commonly accessible regions were used as background for the analysis. Y-axis indicates the fold enrichment versus background, circle size indicates the percentage of regions containing the respective motif, and the color indicates the significance (Log_{10} *p* value). **(F)** B cell activation induces recruitment of chromatin regulators to *Aicda* distal elements. Genome browser view of ChIP-seq data before and after B cell activation showing inducible binding of the chromatin remodelling complex components Brg1 and Chd4, and the histone acetyltransferases p300 and Gcn5, to the TET-responsive *Aicda* elements *TetE1* and *TetE2* in naïve and activated WT B cells. Scale indicates reads per 10 million. Coordinates for locus are chr6:122,523,500-122,576,500 (mm10).

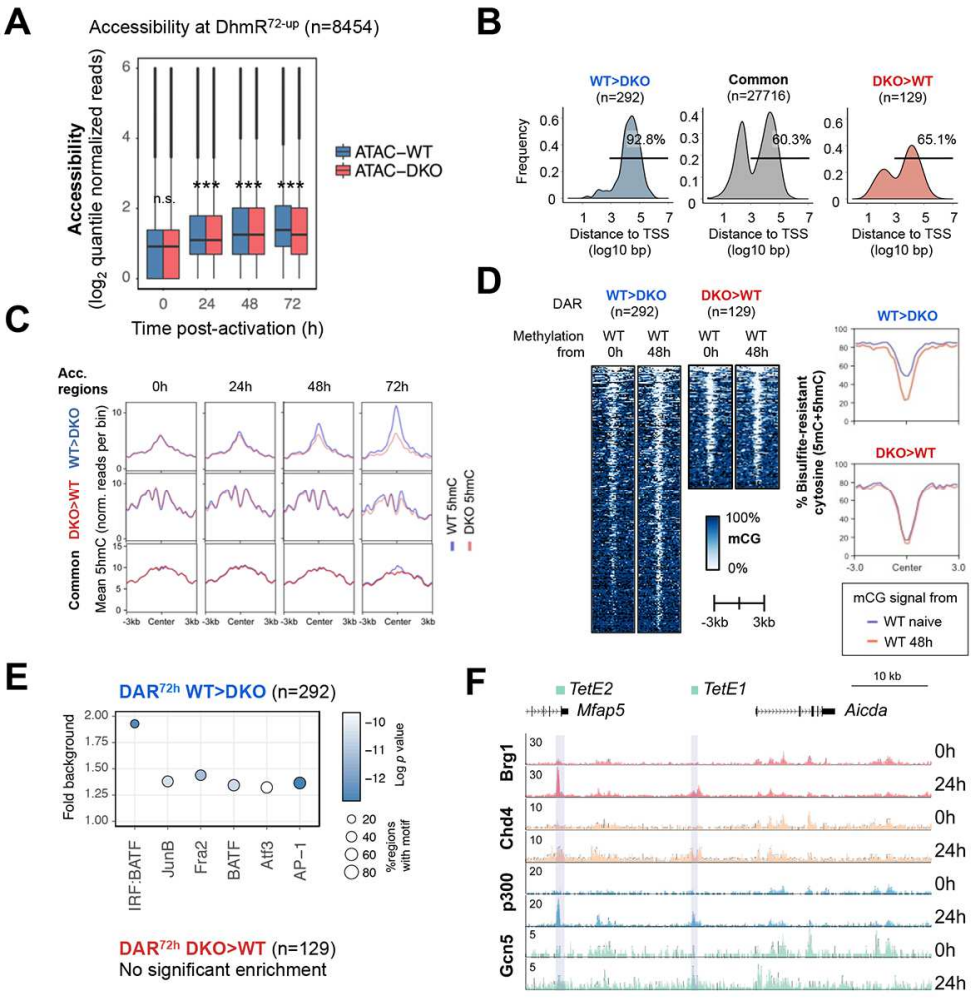
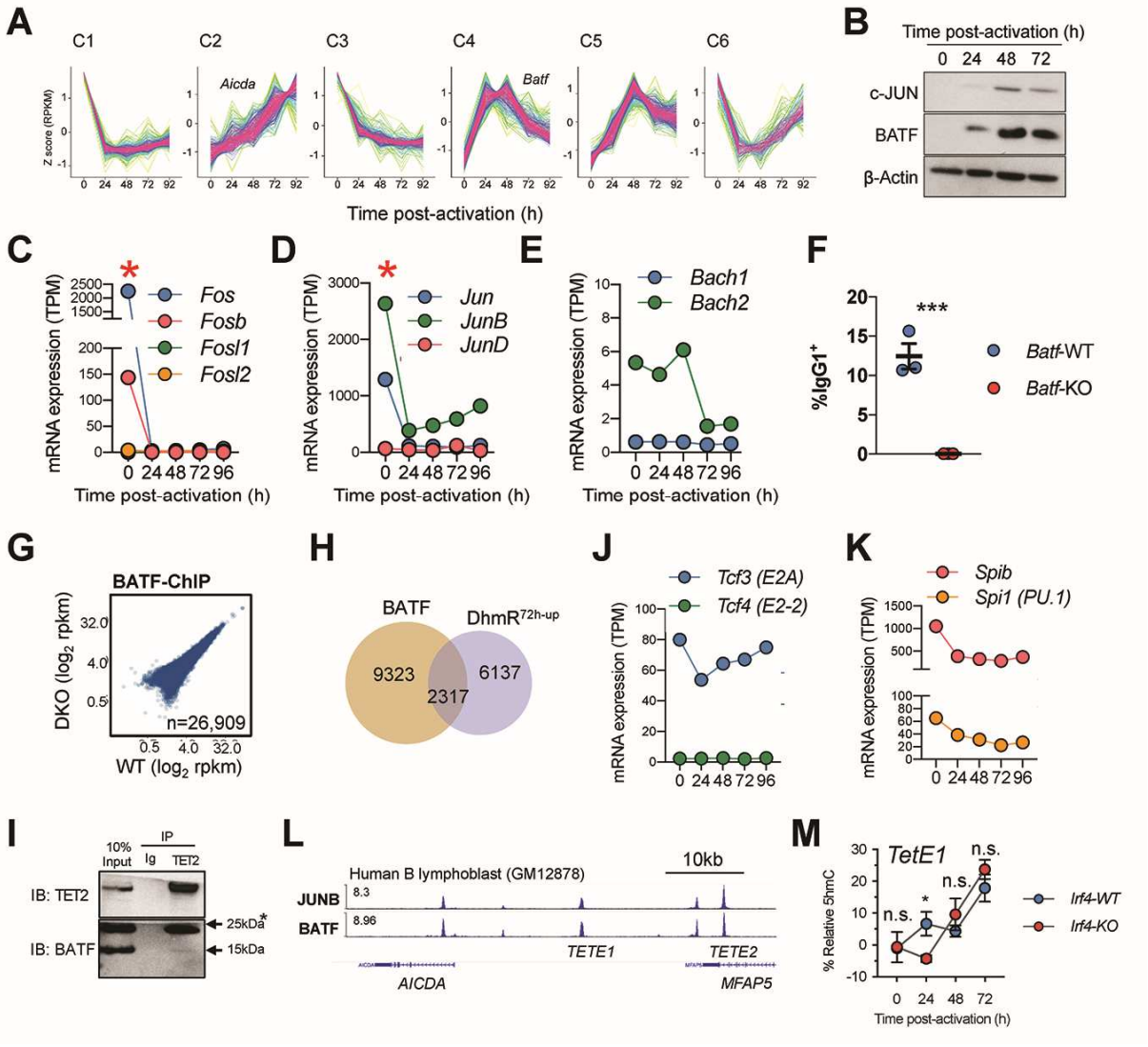

A
Accessibility at DhmR72-up (n=8454)
Accessibility
(log2 quantile normalized reads)
ATAC–WT
ATAC–DKO
n.s.
***
Time post-activation (h)
B
WT>DKO
(n=292)
Common
(n=27716)
DKO>WT
(n=129)
92.8%
60.3%
65.1%
Frequency
Distance to TSS
(log10 bp)
C
Acc. regions
0h
24h
48h
72h
WT>DKO
DKO>WT
Common
Mean 5hmC (norm. reads per bin)
WT 5hmC
DKO 5hmC
D
DAR
WT>DKO
(n=292)
DKO>WT
(n=129)
Methylation from
WT 0h
WT 48h
100%
mCG
0%
% Bisulfite-resistant cytosine (5mC+5hmC)
mCG signal from
WT naive
WT 48h
E
DAR72h WT>DKO (n=292)
Fold background
Log p value
%regions with motif
IRF:BATF
JunB
Fra2
BATF
Atf3
AP-1
DAR72h DKO>WT (n=129)
No significant enrichment
F
TetE2
TetE1
10 kb
Mfap5
Aicda
Brg1
Chd4
p300
Gcn5
0h
24h

Figure S1.8. AP-1 proteins in activated B cells.
(A) Analysis of temporal gene expression modules (TC-seq). Gene expression was analyzed in WT cells activated for various times and genes were clustered based on their temporal expression patterns. Six clusters (C1-C6), or expression patterns, were identified. *Aicda* and *Batf* are found in C2 and C4, respectively. Y-axis indicates the Z-score calculated using RPKM. X-axis indicates the time post-activation. For details, see *Materials and Methods*. **(B-E)** Expression of AP-1 proteins. **(B)** Protein expression of c-JUN and BATF in unstimulated and activated B cells was analyzed by western blot. Mean mRNA expression of selected AP-1 proteins in naïve and activated B cells including *Fos* family **(C)**, *Jun* family **(D)**, and *Bach* family **(E)**. Note that the high basal expression of *Fos*, *FosB*, *Jun* and *JunB* prior to stimulation might reflect the presence of a minor population of contaminating by memory or other B cells (red asterisks)". TPM, transcript per million. **(F)** BATF is required for CSR. B cells were isolated from WT and *Batf-KO* and activated with LPS and IL-4 for 3 days. Class switch recombination to IgG1 was analyzed by flow cytometry (gated on live $CD19^{+}$). Data shown are representative of three independent experiments (n=3 each). Means and standard errors are shown. Statistical significance was calculated using unpaired two-tailed *t*-test. ***, $p<0.01$. **(G)** Tet proteins are not necessary for genome-wide BATF binding. Plots show highly similar distribution of BATF in WT and DKO 72h-activated B cells as analyzed by ChIP-seq with two independent replicates. Shown are the comparison of the BATF enrichment in WT and DKO B cells at the 26,909 regions integrated from the joined peaks from two replicates each of WT and DKO. Axes depict the $\log_2$ rpkm (read per kilobase per million) using quantile-normalized reads for each region analyzed. No region was significantly different between WT and DKO using an adjusted *p* value of 0.05. **(H)** Overlap between BATF binding sites and regions with activation-induced 5hmC modification. Venn diagram showing the number of overlapping regions between BATF peaks and $DhmR^{72h\text{-}up}$. **(I)** The interaction between BATF and TET2 was analyzed by co-immunoprecipitation using nuclear extracts from 48h-activated B cells. The pulldown was carried out in the presence of benzonase and ethidium bromide to minimize non-specific interactions via nucleic acids. The 25kDa band is non-specific as it also appears in *Batf*-KO (not shown). **(J-K)** Expression of E-box and Ets family proteins. Mean mRNA expression for E-box and Ets family proteins, from RNA-seq experiments with two independent replicates. TPM, transcripts per million. **(L)** JUNB and BATF bind to *AICDA* enhancers in human B cells. JUNB and BATF binding in human B cell lymphoblast GM12878 at the *AICDA* locus are shown (Hg38; chr12:8,598,356-8,655,770). The approximate locations for *TETE1* and *TETE2* are indicated based on sequence conservation. Data were originally from ENCODE project, processed by CistromeDB, and were viewed using WashU Epigenome Browser. **(M)** *Irf4*-deficiency has no significant effect on *TetE1* hydroxymethylation. B cells from *Cd19-Cre* (WT) and *Cd19-Cre Irf4-flox* (KO) was analyzed as in **Fig. 1.6E**. Two biological replicates with 3 technical replicates each. *, *$p<0.05$; n.s., not significant.*


A
C1
C2
C3
C4
C5
C6
Aicda
Batf
Z score (RPKM)
Time post-activation (h)
B
Time post-activation (h)
0 24 48 72
c-JUN
BATF
β-Actin
C
mRNA expression (TPM)
Fos
Fosb
Fosl1
Fosl2
Time post-activation (h)
D
Jun
JunB
JunD
E
Bach1
Bach2
F
%IgG1+
***
Batf-WT
Batf-KO
G
BATF-ChIP
DKO (log2 rpkm)
WT (log2 rpkm)
n=26,909
H
BATF
DhmR72h-up
9323
2317
6137
J
Tcf3 (E2A)
Tcf4 (E2-2)
K
Spib
Spi1 (PU.1)
I
10% Input
IP
Ig
TET2
IB: TET2
IB: BATF
25kDa
15kDa
L
Human B lymphoblast (GM12878)
10kb
JUNB
BATF
TETE1
TETE2
AICDA
MFAP5
M
TetE1
n.s.
% Relative 5hmC
Irf4-WT
Irf4-KO

1.8 Author Contributions

C.-W.J.L. and V.S. conceptualized experiments, acquired and analyzed the data, and performed statistical analyses for in vitro and in vivo experiments. D.S.-C. and E.G.-A. performed the majority of the bioinformatics and related statistical analyses (ATAC-seq, CMS-IP, enhancer analysis, WGBS, oxBS-seq, and ChIP-seq) and proofread the manuscript. A.C. performed the expression clustering analysis (TC-seq) and advised on RNA-seq analysis. C.-W.J.L. performed initial bioinformatics analyses, motif analysis, and data visualization. X.Y. provided key reagents and assistance for oxBS-seq experiments. F.A. supervised the bioinformatics analysis and reviewed the manuscript. D.G.S. provided advice and key reagents, interpreted data, and reviewed the manuscript. A.R. supervised and interpreted the data. C.-W.J.L., V.S., and A.R. wrote the manuscript.

1.9 Acknowledgements

We would like to thank you Dr. Uttiya Basu for providing the AID antibody; Dr. Kefei Yu providing the Aicda-KO CH12F3 cells; Dr. Paolo Casali and Dr. Hong Zan for discussion; Laura Hempleman for assisting animal experiments; Cheryl Kim, Lara Nosworthy, Denise Hinz, and Robin Simmons (LJI Flow Cytometry Core) for help with cell sorting; Jeremy Day and Nick Wlodychak (LJI Functional Genomics Center) for assistance with next generation sequencing. C.-W.J.L. was supported by a Cancer Research Institute Irvington Postdoctoral Fellowship. V.S. is supported by Leukemia and Lymphoma Society Postdoctoral Fellowship. A.R. is supported by the National Institutes of Health (NIH) grants R35 CA210043 and R01 AI109842. D.G.S. supported by NIH Grant R01 AI127642. F.A. and A.C. have been partially supported by Institute Leadership Funds from La Jolla Institute for Allergy and Immunology and by the NIH Grant R01 MH111267. Funding for Illumina HiSeq 2500 and BD FACSAria II is supported by NIH (NIH S10OD016262, NIH S10RR027366).

Chapter 1, in full, is a reprint with modifications as it appears in "TET enzymes augment activation-induced deaminase (AID) expression via 5-hydroxymethylcytosine modifications at the Aicda superenhancer", *Science Immunology* (2019), published on April 26th. DOI: 10.1126/sciimmunol.aau7523. The dissertation author was an investigator and co-author of this paper. Other authors include Chan-Wang J. Lio, Vipul Shukla, Daniela Samaniego-Castruita, Abhijit Chakraborty, Xiaojing Yue, David G. Schatz, Ferhat Ay, Anjana Rao.

References

Äijö, T., Yue, X., Rao, A. and Lähdesmäki, H. (2016). LuxGLM: a probabilistic covariate model for quantification of DNA methylation modifications with complex experimental designs. Bioinformatics *32*, i511–i519.

Alt, F.W., Zhang, Y., Meng, F.L., Guo, C. and Schwer, B., 2013. Mechanisms of programmed DNA lesions and genomic instability in the immune system. Cell *152*, 417-429.

Anders, S., Pyl, P.T. and Huber, W. (2014). HTSeq--a Python framework to work with high-throughput sequencing data. Bioinformatics *31*, 166–169.

Betz, B.C., Jordan-Williams, K.L., Wang, C., Kang, S.G., Liao, J., Logan, M.R., Kim, C.H. and Taparowsky, E.J., 2010. Batf coordinates multiple aspects of B and T cell function required for normal antibody responses. Journal of Experimental Medicine *207*, 933-942.

Bird, J.J., Brown, D.R., Mullen, A.C., Moskowitz, N.H., Mahowald, M.A., Sider, J.R., Gajewski, T.F., Wang, C.R. and Reiner, S.L., 1998. Helper T cell differentiation is controlled by the cell cycle. Immunity *9*, 229-237.

Buenrostro, J.D., Giresi, P.G., Zaba, L.C., Chang, H.Y. and Greenleaf, W.J. (2013). Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. Nature Methods *10*, 1213–1218.

Calo, E. and Wysocka, J., 2013. Modification of enhancer chromatin: what, how, and why?. Molecular cell *49*, 825-837.

Casellas, R., Basu, U., Yewdell, W.T., Chaudhuri, J., Robbiani, D.F. and Di Noia, J.M., 2016. Mutations, kataegis and translocations in B cells: understanding AID promiscuous activity. Nature Reviews Immunology *16*, 164-176.

Chandra, V., Bortnick, A. and Murre, C., (2015). AID targeting: old mysteries and new challenges. Trends in immunology *36*, 527-535.

Chavez, L., Jozefczuk, J., Grimm, C., Dietrich, J., Timmermann, B., Lehrach, H., Herwig, R. and Adjaye, J. (2010). Computational analysis of genome-wide DNA methylation during the differentiation of human embryonic stem cells along the endodermal lineage. Genome Research *20*, 1441–1450.

Compagno, M., Wang, Q., Pighi, C., Cheong, T.C., Meng, F.L., Poggio, T., Yeap, L.S., Karaca, E., Blasco, R.B., Langellotto, F. and Ambrogio, C., (2017). Phosphatidylinositol 3-kinase δ blockade increases genomic instability in B cells. Nature *542*, 489-493.

Crouch, E.E., Li, Z., Takizawa, M., Fichtner-Feigl, S., Gourzi, P., Montaño, C., Feigenbaum, L., Wilson, P., Janz, S., Papavasiliou, F.N. and Casellas, R., (2007). Regulation of AID expression in the immune response. Journal of Experimental Medicine *204*, 1145-1156.

De Silva, N.S. and Klein, U., (2015). Dynamics of B cells in germinal centres. Nature reviews immunology *15*, 137-148.

Glasmacher, E., Agrawal, S., Chang, A.B., Murphy, T.L., Zeng, W., Vander Lugt, B., Khan, A.A., Ciofani, M., Spooner, C.J., Rutz, S. and Hackney, J., (2012). A genomic regulatory element that directs assembly and function of immune-specific AP-1–IRF complexes. Science *338*, 975-980.

Gloury, R., Zotos, D., Zuidscherwoude, M., Masson, F., Liao, Y., Hasbold, J., Corcoran, L.M., Hodgkin, P.D., Belz, G.T., Shi, W. and Nutt, S.L., (2016). Dynamic changes in Id3 and E-protein activity orchestrate germinal center and plasma cell development. Journal of Experimental Medicine *213*, 1095-1111.

Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y.C., Laslo, P., Cheng, J.X., Murre, C., Singh, H. and Glass, C.K. (2010). Simple Combinations of Lineage-Determining Transcription Factors Prime cis-Regulatory Elements Required for Macrophage and B Cell Identities. Molecular Cell *38*, 576–589.

Heng, T.S., Painter, M.W., Elpek, K., Lukacs-Kornek, V., Mauermann, N., Turley, S.J., Koller, D., Kim, F.S., Wagers, A.J., Asinovski, N. and Davis, S., (2008). The Immunological Genome Project: networks of gene expression in immune cells. Nature immunology *9*, 1091-1094.

Huang, Y., Pastor, W.A., Shen, Y., Tahiliani, M., Liu, D.R. and Rao, A., (2010). The behaviour of 5-hydroxymethylcytosine in bisulfite sequencing. PloS one *5*, e8888.

Huang, Y., Pastor, W.A., Zepeda-Martínez, J.A. and Rao, A., (2012). The anti-CMS technique for genome-wide mapping of 5-hydroxymethylcytosine. Nature protocols *7*, 1897-1908.

Huong, L.T., Kobayashi, M., Nakata, M., Shioi, G., Miyachi, H., Honjo, T. and Nagaoka, H., (2013). In vivo analysis of Aicda gene regulation: a critical balance between upstream enhancers and intronic silencers governs appropriate expression. PLoS One *8*, e61433.

Ise, W., Kohyama, M., Schraml, B.U., Zhang, T., Schwer, B., Basu, U., Alt, F.W., Tang, J., Oltz, E.M., Murphy, T.L. and Murphy, K.M., (2011). The transcription factor BATF controls the global regulators of class-switch recombination in both B cells and T cells. Nature immunology *12*, 536-543.

Kieffer-Kwon, K.R., Nimura, K., Rao, S.S., Xu, J., Jung, S., Pekowska, A., Dose, M., Stevens, E., Mathe, E., Dong, P. and Huang, S.C., (2017). Myc regulates chromatin decompaction and nuclear architecture during B cell activation. Molecular cell *67*, 566-578.

Kieffer-Kwon, K.R., Tang, Z., Mathe, E., Qian, J., Sung, M.H., Li, G., Resch, W., Baek, S., Pruett, N., Grøntved, L. and Vian, L., (2013). Interactome maps of mouse gene regulatory domains reveal basic principles of transcriptional regulation. Cell *155*, 1507-1520.

Kim, D., Langmead, B. and Salzberg, S.L. (2015). HISAT: a fast spliced aligner with low memory requirements. Nature Methods *12*, 357–360.

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

Klein, U., Casola, S., Cattoretti, G., Shen, Q., Lia, M., Mo, T., Ludwig, T., Rajewsky, K. and Dalla-Favera, R., (2006). Transcription factor IRF4 controls plasma cell differentiation and class-switch recombination. Nature immunology *7*, 773-782.

Ko, M., An, J., Bandukwala, H.S., Chavez, L., Äijö, T., Pastor, W.A., Segal, M.F., Li, H., Koh, K.P., Lähdesmäki, H. and Hogan, P.G., (2013). Modulation of TET2 expression and 5-methylcytosine oxidation by the CXXC domain protein IDAX. Nature *497*, 122-126.

Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., Heravi-Moussavi, A., Kheradpour, P., Zhang, Z., Wang, J., Ziller, M.J. and Amin, V., (2015). Integrative analysis of 111 reference human epigenomes. Nature *518*, 317-330.

Lee, C.H., Melchers, M., Wang, H., Torrey, T.A., Slota, R., Qi, C.F., Kim, J.Y., Lugar, P., Kong, H.J., Farrington, L. and van der Zouwen, B., (2006). Regulation of the germinal center gene program by interferon (IFN) regulatory factor 8/IFN consensus sequence-binding protein. The Journal of experimental medicine *203*, 63-72.

Leek, J.T. (2014). svaseq: removing batch effects and other unwanted noise from sequencing data. Nucleic Acids Research *42*, e161–e161.

Lentini, A., Lagerwall, C., Vikingsson, S., Mjoseng, H.K., Douvlataniotis, K., Vogt, H., Green, H., Meehan, R.R., Benson, M. and Nestor, C.E., (2018). A reassessment of DNA-immunoprecipitation-based genomic profiling. Nature methods *15*, 499-504.

Li, P., Spolski, R., Liao, W., Wang, L., Murphy, T.L., Murphy, K.M. and Leonard, W.J., (2012). BATF–JUN is critical for IRF4-mediated transcription in T cells. Nature *490*, 543-546.

Lin, X., Sun, D., Rodriguez, B., Zhao, Q., Sun, H., Zhang, Y. and Li, W. (2013). BSeQC: quality control of bisulfite sequencing experiments. Bioinformatics *29*, 3227–3229.

Lio, C.W., Zhang, J., González-Avalos, E., Hogan, P.G., Chang, X. and Rao, A., (2016). Tet2 and Tet3 cooperate with B-lineage transcription factors to regulate DNA modification and chromatin accessibility. Elife *5*, e18290.

Meng, F.L., Du, Z., Federation, A., Hu, J., Wang, Q., Kieffer-Kwon, K.R., Meyers, R.M., Amor, C., Wasserman, C.R., Neuberg, D. and Casellas, R., (2014). Convergent transcription at intragenic super-enhancers targets AID-initiated genomic instability. Cell *159*, 1538-1548.

Methot, S.P. and Di Noia, J.M., (2017). Molecular mechanisms of somatic hypermutation and class switch recombination. Advances in immunology *133*, 37-87.

Muramatsu, M., Kinoshita, K., Fagarasan, S., Yamada, S., Shinkai, Y. and Honjo, T., (2000). Class switch recombination and hypermutation require activation-induced cytidine deaminase (AID), a potential RNA editing enzyme. Cell *102*, 553-563.

Murphy, T.L., Tussiwand, R. and Murphy, K.M., (2013). Specificity through cooperation: BATF–IRF interactions control immune-regulatory networks. Nature reviews immunology *13*, 499-509.

Oakes, C.C., Seifert, M., Assenov, Y., Gu, L., Przekopowitz, M., Ruppert, A.S., Wang, Q., Imbusch, C.D., Serva, A., Koser, S.D. and Brocks, D., (2016). DNA methylation dynamics during B cell maturation underlie a continuum of disease phenotypes in chronic lymphocytic leukemia. Nature genetics *48*, 253-264.

Orlanski, S., Labi, V., Reizel, Y., Spiro, A., Lichtenstein, M., Levin-Klein, R., Koralov, S.B., Skversky, Y., Rajewsky, K., Cedar, H. and Bergman, Y., (2016). Tissue-specific DNA demethylation is required for proper B-cell differentiation and function. Proceedings of the National Academy of Sciences *113*, 5018-5023.

Papavasiliou, F.N. and Schatz, D.G., (2002). The activation-induced deaminase functions in a postcleavage step of the somatic hypermutation process. The Journal of experimental medicine *195*, 1193-1198.

Pastor, W.A., Aravind, L. and Rao, A., (2013). TETonic shift: biological roles of TET proteins in DNA demethylation and transcription. Nature reviews Molecular cell biology *14*, 341-356.

Pastor, W.A., Pape, U.J., Huang, Y., Henderson, H.R., Lister, R., Ko, M., McLoughlin, E.M., Brudno, Y., Mahapatra, S., Kapranov, P. and Tahiliani, M., (2011). Genome-wide mapping of 5-hydroxymethylcytosine in embryonic stem cells. Nature *473*, 394-397.

Picelli, S., Faridani, O.R., Björklund, Å.K., Winberg, G., Sagasser, S. and Sandberg, R., (2014). Full-length RNA-seq from single cells using Smart-seq2. Nature protocols *9*, 171-181.

Quivoron, C., Couronné, L., Della Valle, V., Lopez, C.K., Plo, I., Wagner-Ballon, O., Do Cruzeiro, M., Delhommeau, F., Arnulf, B., Stern, M.H. and Godley, L., (2011). TET2 inactivation results in pleiotropic hematopoietic abnormalities in mouse and is a recurrent event during human lymphomagenesis. Cancer cell *20*, 25-38.

Ramírez, F., Ryan, D.P., Grüning, B., Bhardwaj, V., Kilpert, F., Richter, A.S., Heyne, S., Dündar, F. and Manke, T., (2016). deepTools2: a next generation web server for deep-sequencing data analysis. Nucleic acids research *44*, W160-W165.

Reddy, A., Zhang, J., Davis, N.S., Moffitt, A.B., Love, C.L., Waldrop, A., Leppa, S., Pasanen, A., Meriranta, L., Karjalainen-Lindsberg, M.L. and Nørgaard, P., (2017). Genetic and functional drivers of diffuse large B cell lymphoma. Cell *171*, 481-494.

Robbiani, D.F. and Nussenzweig, M.C., (2013). Chromosome translocation, B cell lymphoma, and activation-induced cytidine deaminase. Annual Review of Pathology: Mechanisms of Disease *8*, 79-103.

Robinson, M.D., McCarthy, D.J. and Smyth, G.K. (2009). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics *26*, 139–140.

Rush, J.S., Liu, M., Odegard, V.H., Unniraman, S. and Schatz, D.G., (2005). Expression of activation-induced cytidine deaminase is regulated by cell division, providing a mechanistic basis for division-linked class switch recombination. Proceedings of the National Academy of Sciences *102*, 13242-13247.

Schmitz, R., Wright, G.W., Huang, D.W., Johnson, C.A., Phelan, J.D., Wang, J.Q., Roulland, S., Kasbekar, M., Young, R.M., Shaffer, A.L. and Hodson, D.J., (2018). Genetics and pathogenesis of diffuse large B-cell lymphoma. New England Journal of Medicine *378*, 1396-1407.

Sciammas, R., Shaffer, A.L., Schatz, J.H., Zhao, H., Staudt, L.M. and Singh, H., (2006). Graded expression of interferon regulatory factor-4 coordinates isotype switching with plasma cell differentiation. Immunity *25*, 225-236.

Scott-Browne, J.P., Lio, C.W.J. and Rao, A., (2017). TET proteins in natural and induced differentiation. Current opinion in genetics & development *46*, 202-208.

Sernández, I.V., De Yébenes, V.G., Dorsett, Y. and Ramiro, A.R., (2008). Haploinsufficiency of activation-induced deaminase for antibody diversification and chromosome translocations both in vitro and in vivo. PloS one *3*, e3927.

Song, Q., Decato, B., Hong, E.E., Zhou, M., Fang, F., Qu, J., Garvin, T., Kessler, M., Zhou, J. and Smith, A.D. (2013). A Reference Methylome Database and Analysis Pipeline to Facilitate Integrative and Comparative Epigenomics. PLoS ONE *8*, e81148.

Takizawa, M., Tolarová, H., Li, Z., Dubois, W., Lim, S., Callen, E., Franco, S., Mosaico, M., Feigenbaum, L., Alt, F.W. and Nussenzweig, A., (2008). AID expression levels determine the extent of cMyc oncogenic translocations and the incidence of B cell tumor development. Journal of Experimental Medicine *205*, 1949-1957.

The ENCODE Project Consortium, An integrated encyclopedia of DNA elements in the human genome. Nature 489, 57–74 (2012).

Tran, T.H., Nakata, M., Suzuki, K., Begum, N.A., Shinkura, R., Fagarasan, S., Honjo, T. and Nagaoka, H., (2010). B cell–specific and stimulation-responsive enhancers derepress Aicda by overcoming the effects of silencers. Nature immunology *11*, 148-154.

Tsagaratou, A., Äijö, T., Lio, C.W.J., Yue, X., Huang, Y., Jacobsen, S.E., Lähdesmäki, H. and Rao, A., (2014). Dissecting the dynamic changes of 5-hydroxymethylcytosine in T-cell development and differentiation. Proceedings of the National Academy of Sciences *111*, E3306-E3315.

Tsagaratou, A., González-Avalos, E., Rautio, S., Scott-Browne, J.P., Togher, S., Pastor, W.A., Rothenberg, E.V., Chavez, L., Lähdesmäki, H. and Rao, A., (2017). TET proteins regulate the lineage specification and TCR-mediated expansion of i NKT cells. Nature immunology *18*, 45-53.

Tsagaratou, A., Lio, C.W.J., Yue, X. and Rao, A., (2017). TET methylcytosine oxidases in T cell and B cell development and function. Frontiers in immunology *8*, 220.

Vaidyanathan, B. and Chaudhuri, J., (2015). Epigenetic codes programing class switch recombination. Frontiers in immunology *6*, 405.

Willis, S.N., Tellier, J., Liao, Y., Trezise, S., Light, A., O'Donnell, K., Garrett-Sinha, L.A., Shi, W., Tarlinton, D.M. and Nutt, S.L., (2017). Environmental sensing by mature B cells is controlled by the transcription factors PU. 1 and SpiB. Nature communications *8*, 1-14.

Wöhner, M., Tagoh, H., Bilic, I., Jaritz, M., Poliakova, D.K., Fischer, M. and Busslinger, M., (2016). Molecular functions of the transcription factors E2A and E2-2 in controlling germinal center B cell and plasma cell development. Journal of Experimental Medicine *213*, 1201-1221.

Wu, M., and Gu, L., TCseq: Time course sequencing data analysis. R package version 1.4.0., (2018).

Wu, X. and Zhang, Y., (2017). TET-mediated active DNA demethylation: mechanism, function and beyond. Nature Reviews Genetics *18*, 517-534.

Xi, Y. and Li, W. (2009). BSMAP: whole genome bisulfite sequence MAPping program. BMC Bioinformatics *10*, 232.

Yue, X., Trifari, S., Äijö, T., Tsagaratou, A., Pastor, W.A., Zepeda-Martínez, J.A., Lio, C.W.J., Li, X., Huang, Y., Vijayanand, P. and Lähdesmäki, H., (2016). Control of Foxp3 stability through modulation of TET activity. Journal of Experimental Medicine *213*, 377-397.

Zan, H. and Casali, P., (2015). Epigenetics of peripheral B-cell differentiation and the antibody response. Frontiers in immunology *6*, 631.

Zhang, Q., Zhao, K., Shen, Q., Han, Y., Gu, Y., Li, X., Zhao, D., Liu, Y., Wang, C., Zhang, X. and Su, X., (2015). Tet2 is required to resolve inflammation by recruiting Hdac2 to specifically repress IL-6. Nature *525*, 389-393.

CHAPTER 2: Prediction of gene expression through the use of 5hmC immunoprecipitation enrichment profiles.

2.1 Abstract

5-hydroxymethylcytosine (5hmC) signal enrichment across the gene body positively correlates with the levels of gene expression, but the power of 5hmC enrichment to predict the expression status of a gene remains unexplored. By integrating machine learning techniques and 5hmC inmunoprecipitation data, we show that 5hmC signal across the promoter and gene body can be used to classify genes with respect to their expression status. We generated models with high predictive ability, as measured by the area under the receiver operator curves (AUROC or AUC) and F1 scores. We show that our predictive models are generalizable across different immune cells, in other words, can be trained in one cell type and used to predict gene expression in another cell type, suggesting their broad applicability. Our models showed a median AUC score of 0.87 across all samples when trainining was done per sample and 0.86 when the model was trained using all samples. We also showed a median F1 score of 0.81 and 0.8 in the same context as above. Finally, we found that models trained in immune cells and tested in embryonic stem cells have reduced predictive power (0.74 median AUC score), possibly due to the unique interplay of Tet1 protein and Polycomb repressive complex 2 (PRC2) in stem cells.

2.2 Introduction

5-methylcytosine is a covalent DNA modification catalyzed *de novo* by DNA Methyltransferase 3A (DNMT3A) and 3B (DNMT3B) proteins, and maintained during DNA replication by the DNMT1/UHRF1 protein complex (Moore et al. 2012). 5-hydroxymethylcytosine (5hmC) is a DNA epigenetic mark that is a product of 5-methylcytosine oxidation mediated through the family of Ten-Eleven Translocation (TET) proteins (Tahiliani et al. 2009; Kriaucionis et al. 2009; Pastor et al. 2013). The mammalian TET family is comprised of three enzymes, Tet1, Tet2, and Tet3. TET enzymes are dioxygenases that convert 5mC to 5hmC, 5-formylcytosine (5fC), and 5-carboxylcytosine (5caC) (Ito et al. 2011; He et al. 2011; An et al. 2017). These three oxidised methylcytosine derivatives are essential intermediates in all known mechanisms of DNA demethylation (Tsagaratou et al. 2017; Lio et al. 2020). Our lab and others have developed immunoprecipitation or capture assays to survey 5hmC signal genome-wide, such as GLIB-seq (Pastor et al. 2012), CMS-IP (Ko & Huang et al. 2010; Huang et al. 2012), hMe-Seal (Song et al. 2017), nano-hmC-Seal (Han et al. 2016; Gabrieli et al. 2018), and hMEDIP (Song et al. 2011; Taiwo et al. 2012). Independent of the method used to capture 5hmC, this epigenetic mark has been consistently associated with active genomic regions or "epigenetically dynamic loci" (Szulwach et al. 2011). 5hmC is particularly enriched in cell-specific active enhancers (Szulwach et al. 2011; Tsagaratou et al. 2014), accessible genomic regions (Lio et al. 2016; Lio & Shukla et al. 2019), as well as in euchromatin and transcribed regions (López-Moyado et al. 2019; Nestor et al. 2011). Additionally, we have previously shown that highly expressed genes in murine B cells, in double positive, CD4 and CD8 single positive thymocytes, as well as naïve T cells, Th1 and Th2 cells, have high 5hmC levels across the gene body and Transcription Termination Sites (TTS) when compared to their Transcriptional Start Sites (TSS) (Tsagaratou et al. 2014; Tsagaratou et al.

2017; Szulwach et al. 2011). This enrichment pattern has been also observed in multiple other cell-types, such as in neurons (Stoyanova et al. 2021), cardiomyocytes (Greco et al. 2016), colon epithelia (Uribe-Lewis et al. 2020), liver (Ivanov et al. 2013), myeloid and megakaryocytic erythroid progenitors (Tekpli et al. 2016), and more (Han et al. 2016; Alberge et al. 2020). To the best of our knowledge, no further research has been done to predict gene expression patterns using 5hmC levels across the genome.

However, there have been several previous attempts to predict gene expression: from the mere use of the DNA sequence (Beer et al. 2004; Zrimec et al. 2020; Agarwal & Shendure. 2020), methylation information (Li et al. 2015), accessibility signals (Natarajan et al. 2012), landmark genes (Li et al. 2019), and by integration of multiple histone marks (Singh et al. 2016; Singh et al. 2017). More recent methods also integrated 3D chromatin structure information improving accuracy of gene expression prediction (Zeng et al. 2019; Avsec et al. 2021). Most of these studies made use of powerful Machine Learning (ML) techniques including Deep Learning.

Machine learning algorithms detect patterns in data (Eraslan et. al. 2019) either with labeled training data (supervised learning, e.g. classification of digits from hand writing, or facial expressions), or without (unsupervised learning, e.g., clustering of patients, segmentation of the genome). A subdiscipline of machine learning is known as deep learning (LeCun et. al. 2015), or deep neural networks (DNN). The reason why these networks can discover complex features in the datasets, *e.g.* by using a set of labeled output in supervised training, is because they can integrate increasingly complex representations of the datasets and its interactions. An example of these networks is the fully connected deep neural network (FCDNN), composed of simple mathematical units, known as neurons, grouped by interconnected layers, carrying sequentially more complex operations as more layers are added. Singh and collaborators (2016), used five

histone H3 marks (H3K4me1, H3K4me3, H3K9me3, H3K27me3 & H3K36me3) to train a deep neural network in a binary classification task to predict gene expression (labels High and Low) of genes in 56 different cell-types using the REMC database, with an average diagnostic ability per network of 0.8. However, the results they reported varied depending on the cell analyzed, and the generalizability of the generated models (cross-cell-type predictions) was not tested.

To perform ChIP-seq assays (Johnson et al. 2007) for histone marks, we require access to a large number of viable cells whose nuclei are intact. This could be a limitation if viable cells are not available and the only source of cellular material is DNA, or if cells are subjected to processes that compromise their viability, such as formalin-fixed paraffin embedded (FFPE) preserved samples. Since 5hmC is a stable, covalent DNA modification and would enable the study of samples where viable cells were inaccessible, we asked if it was possible to use 5hmC as an alternative method to predict gene expression. In this project, we explored the use of machine learning algorithms to predict binary classification of gene expression in multiple mouse cell types by using only the magnitude and distribution of the 5hmC signal.

2.3 Results

Datasets. Supervised machine learning models require the existence of both inputs into the proposed model (here termed "features"), and the output corresponding to the value of the predicted target (termed "labels"). We downloaded RNA-seq datasets for gene expression profiling and 5hmC-immunoprecipitation sequencing datasets (using multiple techniques) for 153 samples representing 40 different cell types from the published literature. **Table S2.1** and **Table S2.2** contain all the GEO information and relevant information of the cell-types, for which we acquired 5hmC enrichment and input (control) profiles. **Table S2.3** summarizes all the GEO information and relevant information of the cell-types, for which we acquired RNA-seq gene expression profiles. Finally, **Table S2.4** shows the triad of 5hmC enrichment, its corresponding 5hmC input, and the matched gene expression profile for each cell type and replicate (if available) used for the analyzes in this study.

For each sample, 5hmC enrichment and the 5hmC input signal were processed together to produce the inputs into our proposed model, whereas RNA-seq profiling was processed to obtain a label for gene expression as either "High" or "Low". For each cell type, we normalized gene expression to RPKM and labeled a gene as "High" if its signal was above the median gene expression for that sample, otherwise the gene was labeled as "Low" (**Fig. 2.1A**). The combined set of 5hmC enrichment and corresponding gene expression values were divided into three datasets: training set comprising ~19000 genes (85% of the total), validation set comprising ~1300 of the genes (7.5% of the total) and the test set (also ~1300 genes, 7.5% of the total) used after the final network parameters were inferred, as a way to evaluate the model on unseen data. We used a total of 230 features per gene: 100 variable-sized bins (to account for varying gene lengths) spanning the entire length of the gene body defined as the base pairs (bp) between the TSS and the TTS, 15

100-bp bins spanning the upstream region of the TSS, 15 100-bp bins spanning the downstream region of the TTS (**Fig. 2.1B**). Separately, we used 100 bins of size 100-bp covering the +/-5kb of the TSS to get a detailed representation of 5hmC signal at the promoter (**Fig. 2.1C**). We used the UCSC genome annotation database for the Dec. 2011 (GRCm38/mm10) assembly of the mouse genome, and excluded genes with size smaller than 1000 bp. The 5hmC and input enrichment per bin were calculated and normalized by bin and library size to adjust for the different size of each gene, as well as for differences in sequencing depth. For the regions with higher signal in the control experiments compared to the capture assay, we set the enrichment value to 0.

Baseline Methods. For our first approach to evaluate the performance of 5hmC as a predictor of gene expression in a binary manner, we trained three models for each individual sample: logistic regression (LRg), support vector machines (SVM) and random forest models (RFo), powerful, well-established, off-the-shelf machine learning methods (Fernández-Delgado et al. 2014; van Os et al. 2018). To evaluate the performance of the trained models on each cell type analyzed in an unbiased manner, we calculated the area under the curve (AUC) scores from the Receiver Operating Characteristic (ROC). The range of values the AUC score can take is between 0 and 1, values closer to 1 corresponding to more successful predictions (AUC for random guessing in a binary classification setting with balanced classes would be 0.5). Under default parameters, we found that 5hmC signals displayed a promising predictive power with these three conventional machine learning methods (median AUC values 0.83, 0.78 and 0.76 for LRg, RFo, and SVM, respectively, **Fig. 2.1A** and **Table 2.1**), and that the predictive power varied from cell type to cell type. T cells (naïve CD4 and CD8 T cells, iNKT cells and CD8 Single Positive thymocytes) yielded the highest prediction accuracies, with an AUC above 0.9 for these cell-specific models. On the contrary, networks trained on mESC, cardiomyocytes and mouse embryonic fibroblasts (MEFs)

performed poorly (**Table S2.5**). Unexpectedly, logistic regression outperformed the two other more sophisticated, ensemble-based, learning algorithms (Fernández-Delgado et al. 2014; Kirasich et al. 2018); a possible explanation is that logistic regression performs better when there is a balance between the number of explanatory and noise variable (as long as explanatory variables dominate), whereas random forest has a higher true/false positive as the number of explanatory variables increases in a dataset (Kirasich et al. 2018). The relatively high amount of input features we are implementing in our framework (230 bins) may be a possible source of explanation why Logistic Regression outperformed both ensemble-based learning algorithms.

Deep Neural Networks. Deep fully connected networks are powerful approximations empirically capable of learning complex functions (LeCun et al. 2015). We used fully connected (FC) deep neural networks (DNN) (FCDNN), in an attempt to study if these highly adaptable networks could enhance the ability of 5hmC to predict gene expression. After some hyperparameter tuning (**Table 2.2** presents the minimum, mean and maximum AUC scores obtained in each of the 153 samples when testing the model results for tuning different parameter combinations of layers and neurons per layer), we trained our FCDNNs using the following configuration that also controls ovefitting: hidden layers (hl = {3}), neurons per layer (n = {200,100,50}), learning rate (lr={0.0001}), Dropout Chance (Pd = {0.85}), Decay rate (Dr={0.975}) Decay schedule (Ds = {20}), L2 beta regularization weight (L2 = {0.01}) and minibatch_size (Mbsize = {128}). We found that our FCDNN models outperformed all SVM and RFo models. Additionally, FCDNN outperformed LRg in 148/153 samples tested, thus achieving a higher predictive ability in most conditions (**Fig. 2.2A**, **Fig. S2.1A**, **Fig. S2.1B**, and **Table S2.5**).

The samples we analyzed included mouse embryonic stem cells (in the "ESC" category we are including other types of pluripotent cells, such as iPSC) as well as more differentiated cells.

5hmC is mostly deposited by Tet1 in ESC, unlike differentiated cells where Tet2 and Tet3 are the main mechanism (Neri et al. 2013). Moreover, a major difference between ESCs and other cell types is the correlation between 5hmC and H3K27me3 that is unique to ESCs (Hagihara et al. 2021). Tet1 is a key element in the deposition of both 5hmC and the histone marks H3K27me3 and H2AK119Ub at facultative and pericentromeric heterochromatin; these histone marks are both associated with repression of nearby genes [need a reference]. These observations, and the ESC-specific functional interplay between Tet1 and the PRC2 complex (Cartron et al. 2013) suggest that the roles 5hmC plays in mESC vary from those observed in differentiated cells.

To explore how generalizable the FCDNN models utilizing 5hmC could be in gene expression prediction, we first trained a network using all of the different training and validation datasets to obtain a "Whole" model (whole-dataset model). This model obtained an AUC score of 0.82 and a F1 score of 0.75 (the harmonic mean of the precision and recall, where the highest value of 1 indicates perfect precision and recall). We then tested the trained model's predictive ability in each of the samples' unseen test dataset (**Table S2.5** and **S2.6** for AUC and F1 scores respectively). Each sample had their AUC score and we sorted these values to study if the Whole-model had a bias towards any particular kind of cell type. We found that the ESC samples were highly enriched for having lower AUC values, suggesting that the Whole model's AUC was separating differentiated from the ESCs samples (**Table S2.5**, see the "Whole" column under the "Multi-sample models" column group). We sorted in increasing order all of our sample's AUC scores (calculated using the "whole" model's weights), and using this order we plotted the cumulative frequency of finding a mESC sample (**Table S2.5**, sample rows labeled as ESC in the "Subgroup" column), calculated as the ratio of mESC samples seen from the 34 total ESCs samples (**Fig. 2.1B**, Y-axis) for any given position in the sorted scores (**Fig. 2.1B**, X-axis; 153 total positions, equal to

all our samples). Midway through the sorted scores (position 77 in the X-axis, yellow-filled romboid over black line for the Whole model), 82% of the ESC were being covered, meaning that most of the the ESC were below the median across the distribution of observed AUC scores. To further study the observed separation of predictive ability for ESC vs immune-related cell-types, like T-cells (**Table S2.5**, T-cell samples with an AUC score higher than 0.9 are bolded under column titled "Whole"), as well as to explore deeper the generalizability of the 5hmC models, we generated a subset compossed of immune cell samples ("Immuno" set), and another subset composed of all the ESC samples ("ESC" set) (see column "Subgroup" for sample membership in both **Table S2.5** and **Table S2.6** tables). We then trained two new models named "Immuno" and "ESC", and obtained AUC scores of 0.89 and 0.83, and F1 scores of 0.83 and 0.74, respectively. **Table 2.3** and **Table 2.4** show the summary statistics of AUC score distribution of each group's unseen test data. The relatively lower scores achieved by the ESC model could potentially be explained by the bivalent role 5hmC plays in ESC as discussed above. Since the model relies on 5hmC enrichment to predict gene expression, having 5hmC also deposited in repressed genes and inactive regions may hamper the dissection of what genes are actually being expressed (Wu & D'Alessio et al. 2011).

The test sets' genes are comprised of the same randomly chosen genes across all the samples in a group, thus our models never see the 5hmC signal from those genes in any cell type in either the training or the validation datasets. In order to better assess the generalizability of our predictions to completely unseen cell types, we repeated our training by withholding a number of samples from the training, and used them as test sets in the final AUC calculation (**Table S2.7)**. The AUC scores we obtain from this setting were similar to witholding randomly chosen gene sets

across all cell types suggesting that our predictive models generalize well to samples not seen by the model (**Table S2.7**, see "Withheld" row results compared to "All Samples" row results).

Since the new two models were specialized for either ESC or Immunological cell-types, we suspected that the Immune model would perform worse with ESC-like samples compared to either the ESC or the Whole model. First, we tested the prediction ability of these two trained models in each of the samples' unseen test dataset. From the distribution of AUC scores across all samples, we obtained an AUC mean of 0.81 (**Table 2.3**) for the Immuno and ESC models (F1 mean scores of 0.72 each, **Table 2.4**), a result that was similar to each model's validation dataset. Then, we repeated the analysis of the relative prediction rankings (AUCs ranked low to high) for ESC samples with respect to all other cell types using the three trained networks (Whole, Immuno, ESC) (**Fig. 2.1B)**. All 34 ESC samples were among the worst performing (bottom 50%) for the Immuno model, unlike the ESC model, where the AUC scores of ESC samples were more evenly distributed across the whole ranking. These results are in line with observations that the role of the 5hmC mark may be different between differentiated and undifferentiated cell-types (Wu & D'Alessio et al. 2011).

Although the specialized models excel in the predictions associated to the cell-types they were trained on (**Fig. 2.2C**), when comparing the summary statistics on the distribution of each model's AUC or F1 scores across all samples (**Table 2.3** and **Table 2.4** respectively; highest value per column across the multi-sample models are in bold), none of the two specialized models outperformed the whole model's mean or median values (**Fig. S2.2**; see **Table S2.5** and **Table S2.6** for all the AUC and F1 scores respectively). The whole model used not only the Immuno and ESC subsets, but also the remainder of the samples, and this exposure may explain why it learned features generalizable to diverse cell-types.

When assigning a gene an expression label (such as "ON" and "OFF") the median of a cell's gene expression is the usual threshold to have a balanced label group (splitted the data in two equally sized groups). However, this threshold does not bear much biological significance; variation in replicates from the same cell-type may move a gene's expression above or below the expression threshold. We explored the accuracy of our specialized models per gene expression quartiles and found that the best performing quartiles were the 1st and the 4th quartiles (bottom 25% and top 25% expressed genes), whereas the 2nd and 3rd quartiles performed poorly, particularly true for the ESC model (**Fig S2.3**). This means that our model accurately captures the lowest and highest expressed gene categories whereas those with intermediate expression values, close to the median threshold, are more challenging to predict.

Finally, we wanted to explore what were the most important features in performing the gene expression prediction task. To this end, we implemented DeepLift (Shrikumar et al. 2019), a tool that gives a significance score to each of the features of a DNN relative to the state of the network after a "reference" signal (e.g. any gene's 5hmC signal distribution) is processed by the network. If we activated the network (the state of a network when all features are processed and a prediction is made) by the use of features associated to a gene labeled as "High", and decoded it by the features of another "High" gene, the significance scores assigned to each feature may be very low, since the network activation state may not change (**Fig. 2.3A**). On the contrary, if we activate the network by the use of features associated to a "High" gene and and decode it by the use of a "Low" gene, or *vice-versa* (e.g., the use of opposites), this generates an artificial contrast between the state and the reference set of signals that almost all features become "important" (**Fig. 2.3B**). In order to obtain a distribution of relative significances per feature, we fed DeepLift the networks activated by neutral signal. This neutral signal was generated using randomly sampled

genes (equal number of High and Low genes) and averaging their signal for each of the 230 bins. For each label ("High" and "Low"), we ran this process for the Whole network, and found that the features representing the TSS, and those surrounding the promoter, are the most relevant features for the gene expression prediction task using 5hmC (**Fig. 2.3C**). This is consistent with previous studies finding that the signals slightly downstream of TSSs are the most informative (Cheng et al. 2011) and that epigenetic features in or near the promoter region were the most informative in the gene expression prediction task (Dong et al. 2012; Singh et al. 2016; Singh et al. 2017). Then, we ran this process for both the Immuno and the ESC models (**Fig. 2.4**) and, although in both models the bins representing up to 600 bp downstream the TSS are the most relevant (**Fig. 2.4A**), the ranking of their significance score follows opposite directions in each model. For the ESC model the most significant bin is located 600 bp downstream the TSS, followed by the bins upstream, whereas the bin representing the TSS is the most significant for the Immuno model, followed by the bins downstream (**Fig. 2.4B,** green circles).

Overall, our results show that 5hmC signal enrichment on its own can be used to effectively predict gene expression, and that the model trained in all the available datasets is sufficiently generalizable to cell types that are not represented in the training.

2.4 Discussion

5hmC signal enrichment has been associated with positive gene expression. Here we explore this association further by successfully employing FCDNN that models signal from 5hmC enrichement methods to predict gene expression. When we calculated the AUC in models trained and tested on the same cell type, we obtained a median AUC of 0.87 (across 153 samples from 40 different cell types each with 1 to 10 replicates, 2 in most cases). We extended these models to a diverse range of cell types and also produced a generalizable model "Whole" with a median AUC of 0.86 and median F1 score of 0.80 that can be used in cells from different context. This is comparable to other state-of-the-art models that use the genomic distribution of histone marks, and complex network architectures such as kernels and convolutions in DeepChrome (Singh et al. 2016), and a hierarchy of multiple Long Short-Term Memory modules with recurrent and memory cells in AttentiveChrome (Singh et al. 2017) with F1 scores of 0.69 and 0.62 respectively. Both methods had an AUC score of 0.80, however, they trained a specific model per each cell-type and have not considered the generalizability of their predictions to unseen cell types. To our knowledge, our group is the first to implement a deep learning framework to predict gene expression from 5hmC signal alone.

In all samples, our Deep Neural Network framework always outperformed SVM and Random Forest models, as well as Logistic Regression with few exceptions (see **Table S2.5**, all six sample exceptions are bold in the "LRg" column under "Models trained in one sample" column group). Although exceptions occurred with some mESC samples having an exceptionally good AUC using Deep Neural Networks, as high as 0.9 (see the bold numbers in column "FCDNN", under "Models trained in one sample", from **Table S2.5**), from the compendium of all cell-types, mESC-related samples were generally low. Our findings may be an additional piece of evidence

supporting the bivalent roles that 5hmC plays in ESC , where Tet1 has a dual function of promoting transcription of pluripotency factors and participating in the repression of Polycomb-targeted developmental regulators (Wu & D'Alessio et al. 2011). We note that, regardless of these contrasting 5hmC roles between differentiated and ESC datasets, when we analyzed the highest relevant features in either ESCs or differentiated cells, we found that the bins corresponding to the TSS-proximal region were the most significant regions for both. Whether these Deep Neural Networks could be further tuned through the network structure and configuration (e.g. convolutional networks, total layer, kernels, neurons, etc.), optimizers, regulators and more hyperparameters to enhance the ESC model prediction is an open question.

Other groups have been successful in the gene expression prediction task by integrating a wide range of histone marks (Sekhon et al. 2018; Zeng et al. 2019), using convolutions such as DeepChrome (Singh et al. 2016), or more complex DNN frameworks such as AttentiveChrome, with a median AUC score of 0.80 (Singh et al. 2017). For these methods the data required to generate the input features is dependent on sequencing assays that rely on intact cells to perform protein immunoprecipitation (IP) of histones and other proteins from chromatin (this process involves crosslinking of intact nuclei). Compared to proteins and most cellular organelles, DNA is more stable and can last longer out of the laboratory environment. 5hmC is a stable and covalent mark, therefore as long as some DNA is present in a sample, 5hmC can be assessed and quantified, as recently shown for cell-free (circulating or plasma) DNA from normal and cancer patients (Song et al. 2016). Moreover, 5hmC immunoprecipitation techniques that allow the use of minimal DNA starting material without compromising quality data, such as nano-hMe-Seal (Han et al. 2016; Gabrieli et al. 2018), can be used to get a sense of gene expression profiles of samples where DNA is scarce and RNA is degraded.

Although we carefully normalized the 5hmC enrichment datasets such that the signals across each feature were comparable, each 5hmC signal capture assay may add uncontrolled variation that, if corrected, could increase the models' performances. For future work, we would like to study the technical variations from each capture assay across samples, unfortunately a limiting factor that is faced on this analysis is the lack of individual samples, or cell-types, covered by more than one technique. Even with these adversities, our models achieved high AUC scores similar or higher than those using protein-based assays such as histone/ chromatin immunoprecipitation and/or multiple enrichment datasets. Finally, for most of the samples, the best performing network was the one trained specifically on the same sample as expected. However, the overall low prediction performance in some samples (see bottom tail in **Fig. S2.1B**) warrants further studies of the possible sources of unexplored variation, to explain why 5hmC profiles for these few samples (including macrophages, embryonic bodies, bergman glia, Th1 and CD4 SP T cells) were not predictive of gene expression.

In summary, we demonstrated that 5hmC signal can be used to train machine learning methods of varying complexity in the binary classification task of gene expression prediction. These kinds of processes are achieved to great extent thanks to deep learning's ability to juggle gigantic amounts of existing data, as well as to automatically filter in/out relevant features and integrate complex interactions, providing us with important tools.

2.5 Materials and Methods

RNA-seq analysis. All RNA-seq datasets were processed using STAR. We downloaded the raw reads and mapped them to the UCSC genome annotation database for the Dec. 2011 (GRCm38/mm10) assembly of the mouse genome. Counts per gene were obtained using FeatureCounts. Similar results were obtained when using STAR's count algorithm. For the generation of the output labels, we RPKM-normalized the RNA signal expression and took the median gene expression as the threshold to divide and label genes as "High" and "Low" (above or below threshold, respectively).

5hmC-seq and Input analysis. All enrichment datasets were processed in the same pipeline. We downloaded the raw reads and mapped them to the mm10 genome reference assembly using Bsmap for CMS-IP and bwa for the other tools. Unmapped reads were remapped after using trimgalore from trimming and added to the mapping results after both files were sorted with samtools. PCR duplicates were estimated and removed using Picard tools' MarkDup. Reads mapping in the blacklisted regions were removed before further analysis. We generated HOMER's TagDirectories followed by HOMER's makeMultiWig tracks for visualization in the genome browser. We took the promoters, TSS and TTS from the UCSC genome annotation database for the Dec. 2011 (GRCm38/mm10) assembly of the mouse genome. For each gene longer than 1000 bases, we extended the promoter both upstream and dowmstream by 5,000bp, and divided these 10,000bp stretches into 100 equally-sized bins (100bp per bin). We also took 1,500bp regions both upstream of the TSS and downstream of the TTS, resulting in 15 equally sized 100bp-bins for each per gene. We split the gene body (from TSS to TTS) into 100 equally-sized bins (per gene only, unlike the other bins). We used these bins to obtain the raw 5hmC signal from the mapping results

and proceeded to RPKM-normalization based on the sequencing depth per sample and the size of each bin.

Machine Learning Baseline methods. All Three methods implemented as baseline, Logistic Regression, Random Forest, and Support Vector Machines, were run with default parameters in R (version 3.3.3), from packages "tibble", "randomForest" and "e1071" respectively, using all the 230 bins as the explanatory variable and the gene expression as the target. The Train, Validation and Test datasets per sample were split into 85/7.5/7.5% form the total. For the AUC scores, we used the library pROC's roc function.

Deep Neural Networks. We developed our DNN frameworks in Python's TensorFlow and translated them into Keras for their analysis with DeepChrome. We always used CPU, since the complexity of the networks did not require us to use GPU or intense memory requirements. After the hyperparameters survey, we trained our FCDNNs using the following hyperparameters: hidden layers (hl = {3}), neurons per layer (n = {200,100,50}, respectively), learning rate (lr={0.0001}), Dropout Chance (Pd = {0.85}), Decay rate (Dr={0.975}) Decay schedule (Ds = {20}), L2 beta regularization weight (L2 = {0.01}) and minibatch_size (Mbsize = {128}). For the models consisting more than one replicate/sample in training, such as the general model "Whole", and the other models "Immuno" or "ESC", we increased the number of Epochs to 60. The Train, Validation and Test datasets per sample were split into 85/7.5/7.5% form the total.

We used DeepLift with target layer index ({-2}), this computes explanations with respect to the logits, appropriate for our network architecture. The score layer index we used was ({0}) which correspond to the scores for the input layer.

2.6 Figures

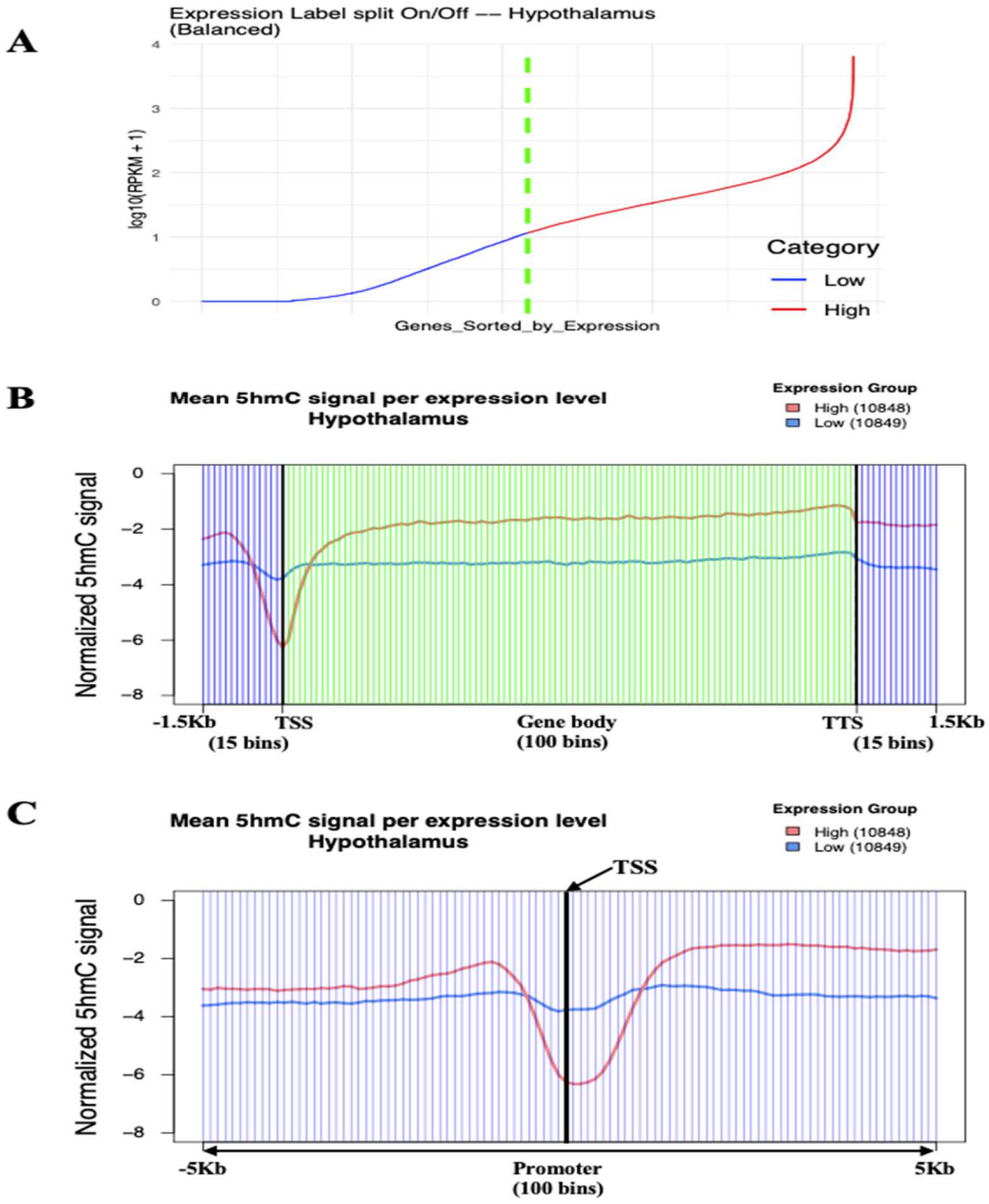


Figure 2.1. Sample normalization and Input Generation.
(A) Gene expression obtained from RNA-seq of hypothalamus (whole brain region, Lin et al. 2017) were RPKM-normalized and from the resulting expression, we divided (dashed diagonal green line) the genes into two equal groups as Low and High expression (blue and red lines respectively). **(B-C)**. Average 5hmC signal per expression group through the gene body (**B**) and promoter regions (**C**). Vertical lines in **(B)** and **(C)** represent single divisions (bins), blue lines (Promoter, TSS and TTS) represent bins with fixed size and green dashed lines (Gene Body) represent bins with variable size. Labelling and bin results from a biological replicate of Hypothalamus cell-type are shown in this figure as representation.

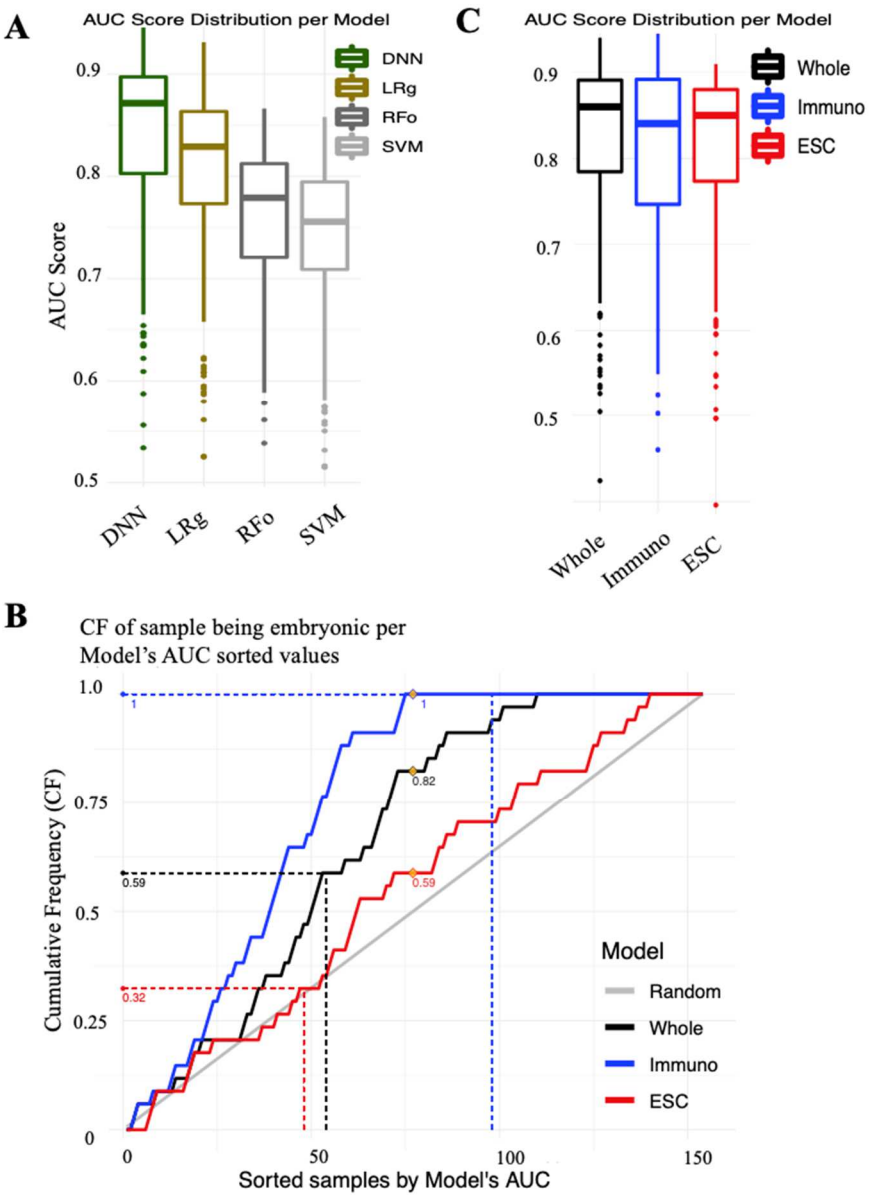


Figure 2.2. Evaluation of different methods to predict gene expression from 5hmC signal. **(A)** AUC score distribution for baseline machine learning models: logistic regression (LRg), golden, random forest (FRo), gray, and SVM (light gray) next to the scores from our fully connected de neural network (DNN), green, for each sample. Our framework outperforms all of the baselines. **(B)** Cumulative frequency of a sample being an ESC. The range in the X-axis the total samples we used in this study (153 total). Each model trained in multiple samples (Whole, Immuno and ESC; black, blue and red solid lines, respectively) was used to process each sample and calculate its' AUC score, sorted increasingly. Each model's validation AUC score is indicated by the dashed lines. Whole model had 82% of ESC samples below the median AUC score (X-axis position 77) **(C)** AUC score distribution for specialized models Immuno and ESC (blue and red respectively) next to the scores from the model trained in all the available datasets "Whole", which outperforms the specialized models when predicting the expression labels of each sample.

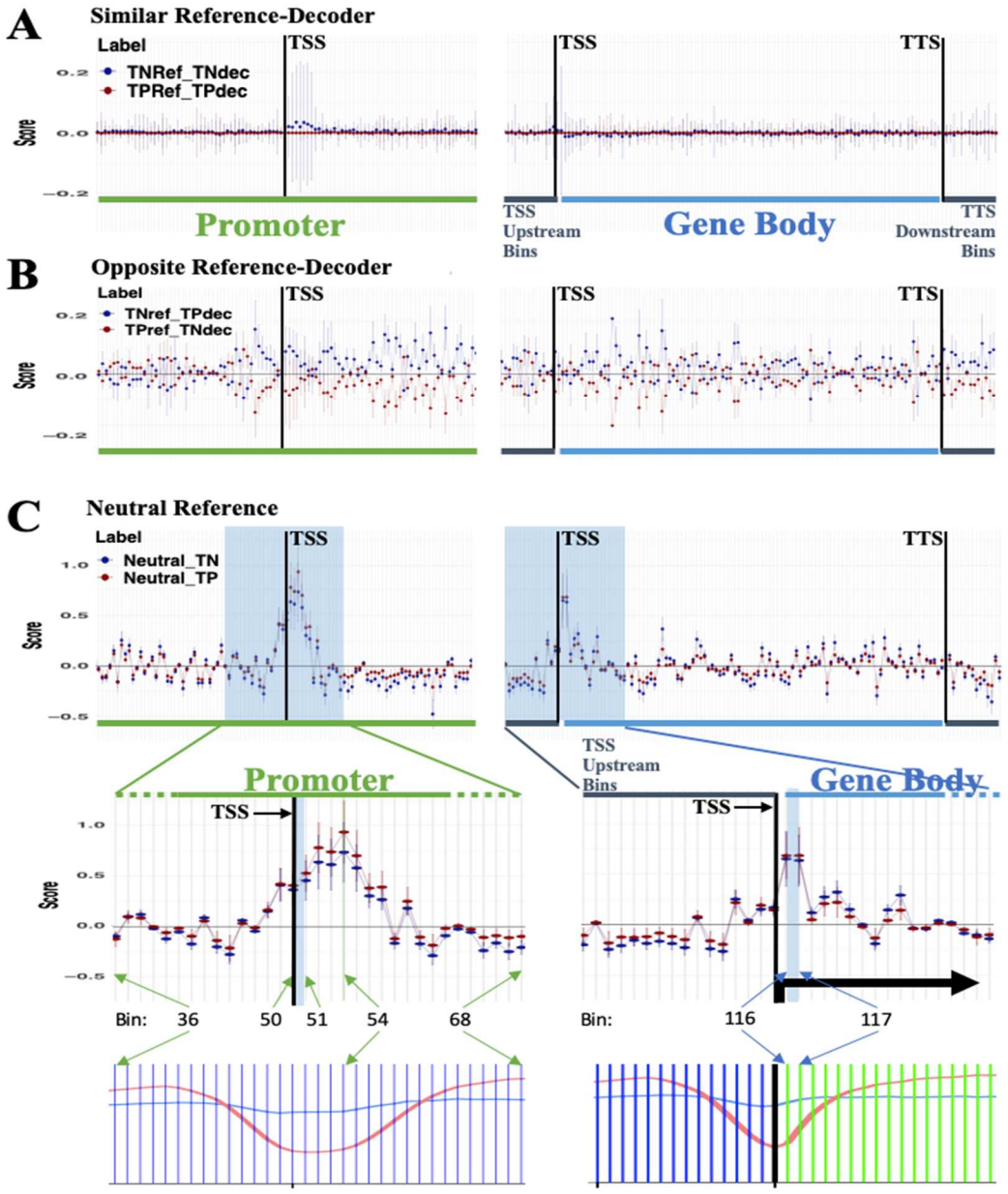


Figure 2.3. DeepLift Significance scores.
(**A.-C.**) Distribution of DeepLift significance scores throughout the 230 bins using different combination of genes (input signal) for network activation and decoding. (**A**) When using true positives (TP, red), or true negatives (TN, blue), for both activation and decoding, no enrichment for any specific position is found in either Immuno (top) or Embryo (bottom) models. (**B**) Providing a decoding signal to contrast with the activation signal (e.g. TP as activation and TN as decoding signal), the enrichment pattern of significance scores is noisy. (**C**) When we use a neutral signal (e.g. all 230 bins have random signal, or all zeroes) we obtained clear significance (compare Y axis across panels) near the 10 bins surrounding the TSS (see **Fig. 2.1C**; peak at bin 54, corresponding to 400bp downstream TSS) and 6 bins downstream of the start of the gene body (see **Fig. 2.1B**; peak at bin 116 and 117).

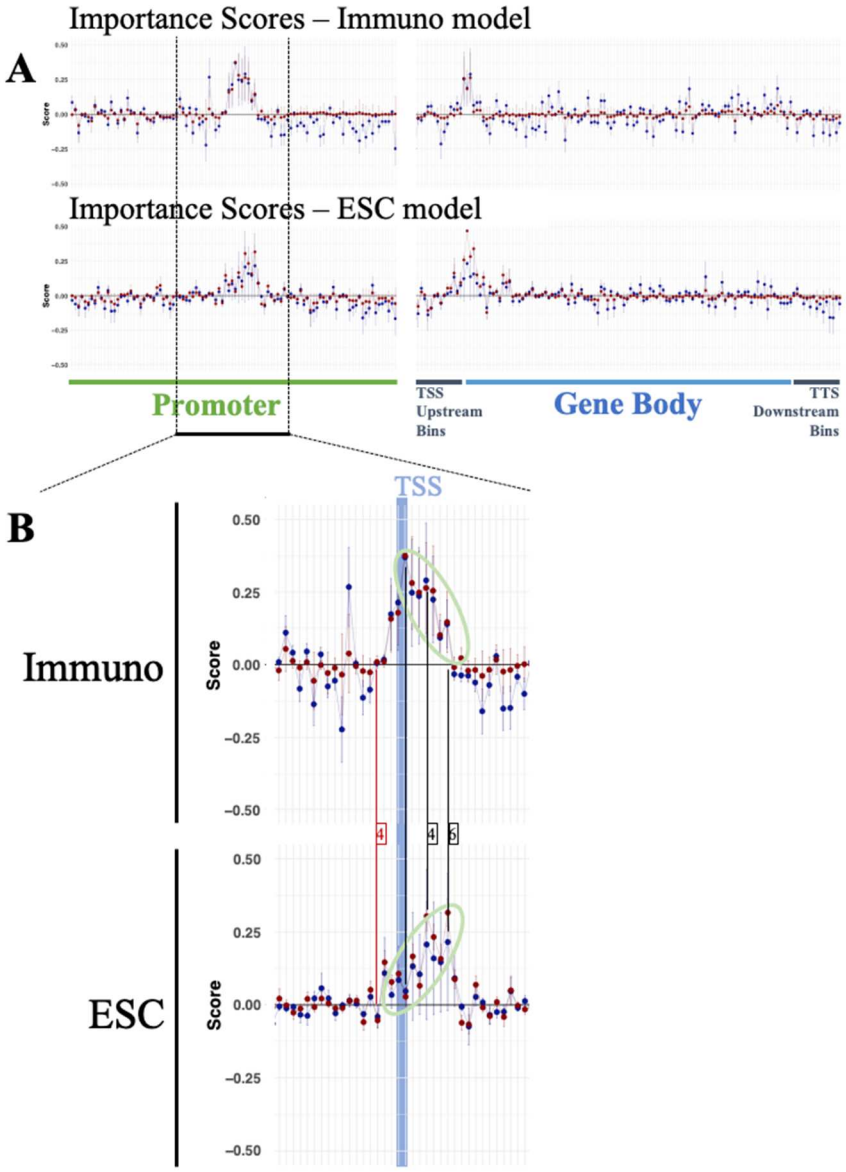


Figure 2.4. DeepLift Significance scores.
(**A**) Distribution of DeepLift significance scores throughout the 230 bins using a neutral signal (e.g. all 230 bins have random signal, or all zeroes) for network activation and true positives (TP, red) or true negatives (TN, blue) genes for decoding. (**B**) Zoom-in of the TSS for the Immuno and the ESC models. The most significant bins for the Immuno model are found directly over the TSS and the following most significant bins are located downstream this position up to the 6th -to- 7th bin. For the ESC model, the 6th bin has the most significance and the following most significant bins are located upstream, towards the TSS, whereas the bins downstream (7th bin onward) had a sharp loss of significance.

2.7 Tables

Table 2.1. All samples's AUC score distribution for each traditional machine learning tool on the gene expression prediction task.

Model	Min	1stQ.	Median	Mean	3rdQ.	Max
Logistic Regression (LRg)	0.5249	0.7733	0.8290	0.8046	0.8634	0.9311
Random Forest (RFo)	0.5384	0.7209	0.7791	0.7594	0.8124	0.8663
Support Vector Machines (SVM)	0.5148	0.7091	0.7557	0.7392	0.7946	0.8581

Table 2.2. Hyperparameter tuning of total connected layers and neurons per layer. Shown are the *Min*, *Mean* and *Max* values across all AUC scores per sample per configuration.

Neurons per layer	Min	Mean	Max
200,100,50	0.5142	0.8129	**0.9179**
100,100,100	0.5001	**0.8135**	0.9135
50,50,50	0.4811	0.8077	0.9133
200,200	0.5210	0.8123	0.9109
100,100	**0.5225**	0.8113	0.9135
50,50	0.4939	0.8055	0.9113

Table 2.3. Summary statistics of the AUC scores per DNN model processing each samples' unseen test datasets "Final results".

Model	Min	1stQ.	Median	Mean	3rdQ.	Max
Whole	0.4242	**0.7846**	**0.8601**	**0.8186**	0.8912	0.9405
Immuno	**0.5025**	0.8719	0.892	0.8741	**0.9294**	**0.9494**
ESC	0.5483	0.7821	0.8266	0.7987	0.869	0.8964
Sample-specific	0.5330	0.8020	0.8710	0.8384	0.8970	0.9450

Table 2.4. Summary statistics of the F1 scores per DNN model processing each samples' unseen test datasets "Final Results".

Model	Min	1stQ.	Median	Mean	3rdQ.	Max
Whole	**0.1342**	**0.7008**	**0.7951**	**0.7354**	0.8249	0.8835
Immuno	0.4043	0.8064	0.8265	0.7931	**0.8647**	**0.8878**
ESC	0.5737	0.6854	0.7411	0.7373	0.8074	0.8340
Sample-specific	0.5280	0.7370	0.8090	0.7788	0.8310	0.8920

2.8 Supplemental Data, Tables and Figures

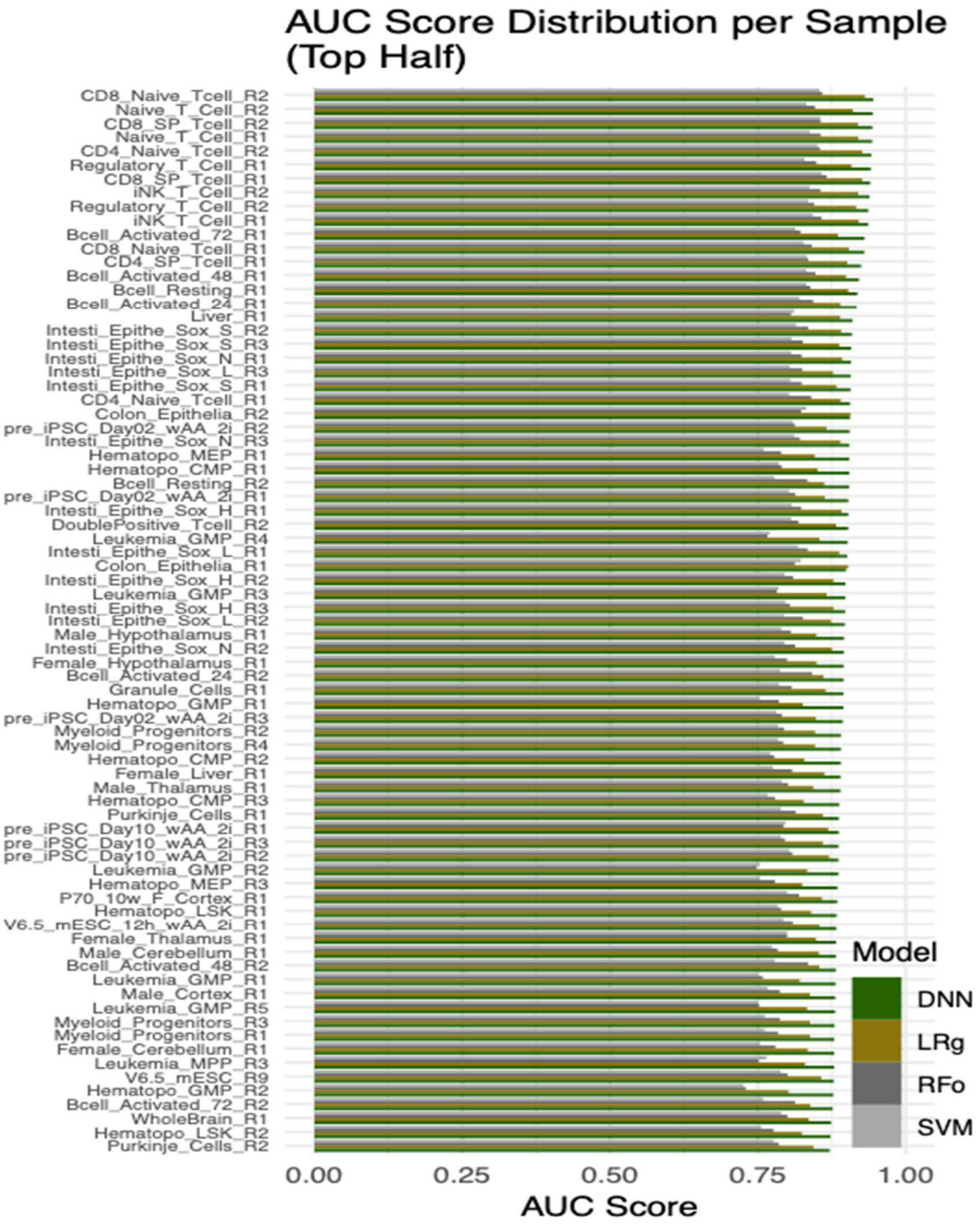


Figure S2.1A. AUC Scores per sample for each machine learning method (TOP).
The AUC scores were sorted with respect to the scores obtained in the DNN model. There is a total of 153 samples and the sorted dataset was split on two to visualize each sample ID, here we show the top 76 samples.

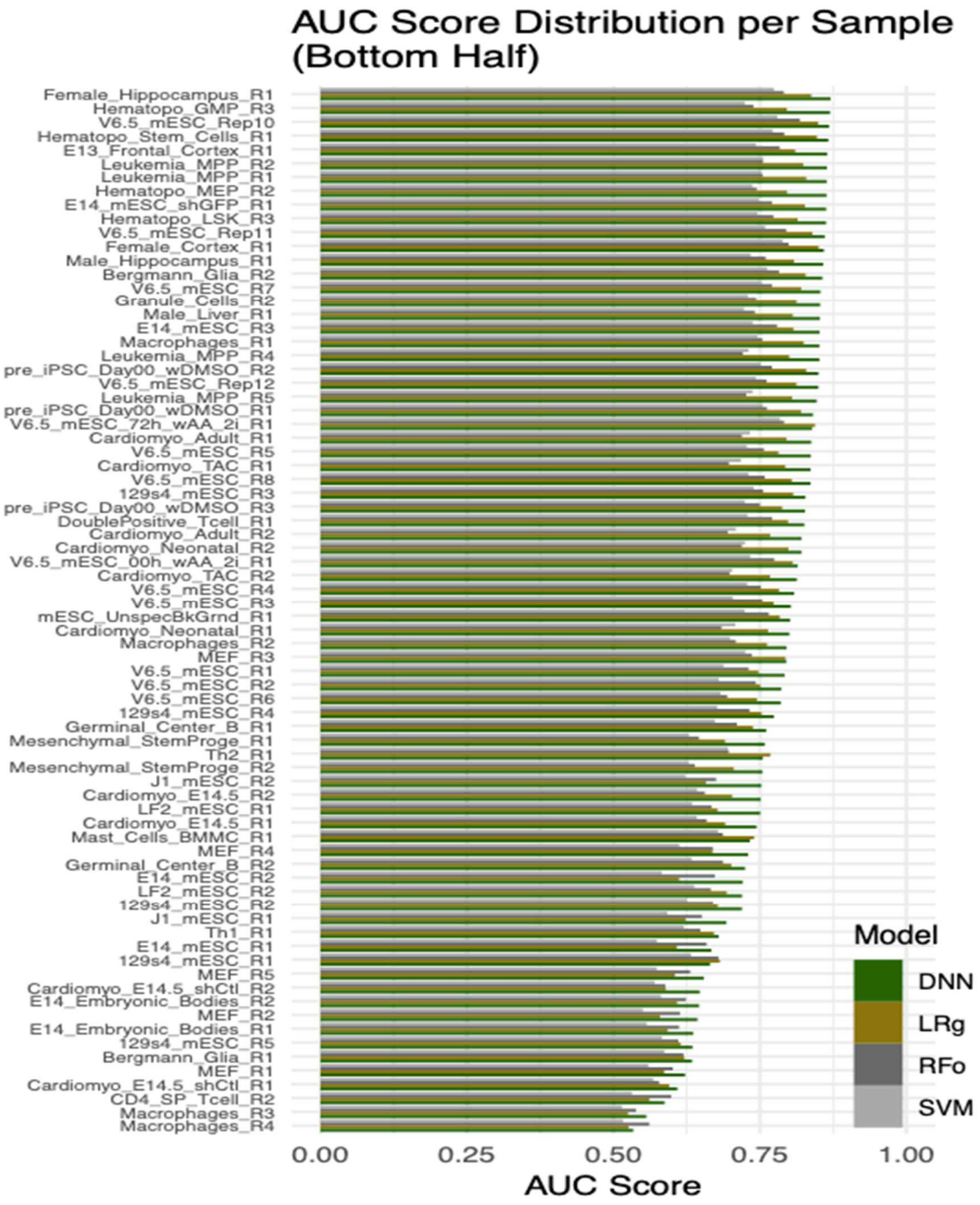


Figure S2.1B. AUC Scores per sample for each machine learning method (BOTTOM). The AUC scores were sorted with respect to the scores obtained in the DNN model. There is a total of 153 samples and the sorted dataset was split on two to visualize each sample ID, here we show the bottom 77 samples.

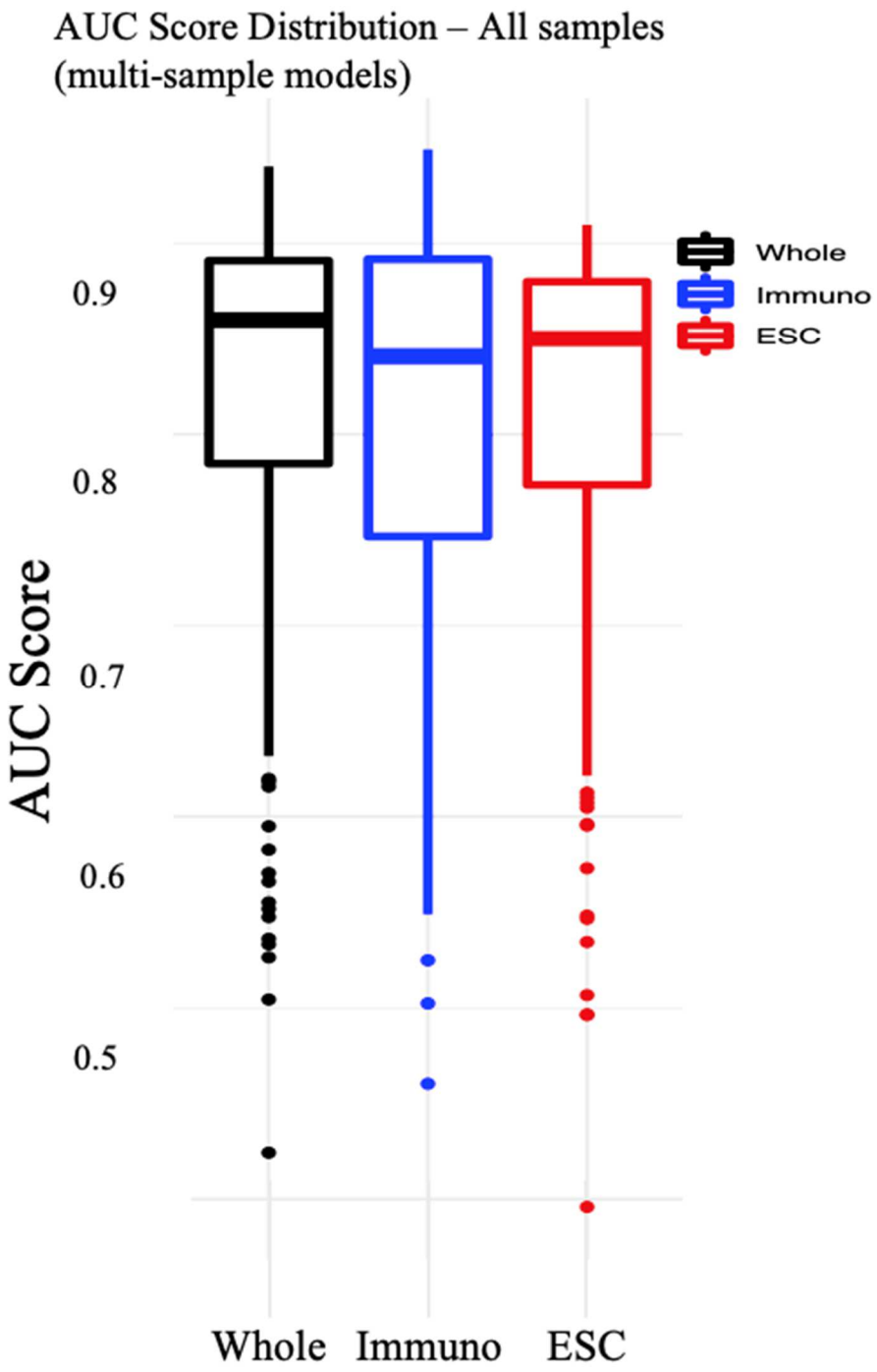


Figure S2.2. AUC Score distribution per specialized model processing the entire dataset. The boxplots represent the AUC score discribution calculated in the test set of all the datasets when using each of the specialized models "Whole", "Immuno" and "ESC". The "Whole" model has the best performance overall likely due to being trained in a higher spectrum of cell types.

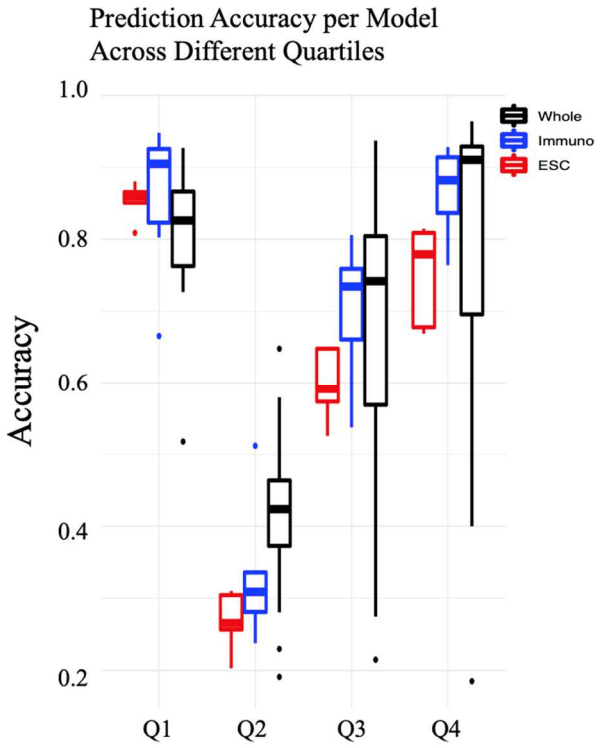


Figure S2.3. Accuracy distribution of the predicted labels per specialized model processing the tests datasets.
Quartiles were defined using the RPKM-normalized gene expression. Q1 represent the bottom 25% of genes based on expression mostly complised of genes with no expression, whereas Q4 represent the 25% top most expressed genes. The ESC model had low accuracy in both Q2 and Q3, where Q2 had the lowest accuracy of all.

Table S2.1. Publication and GEO/project information of 5hmC Enricment downloaded data.

GEO/Project	Sequencing Technique	PubmedID	Citation
PRJEB2462	hMeDIPSeq	21460836	Ficz et al. 2011
GSE28682	CMSSeq	21552279	Pastor et al. 2011
GSE28682	GLIBSeq	21552279	Pastor et al. 2011
GSE44566	GLIBSeq	23987249	Neri et al. 2013
GSE49191	CMSSeq	24270360	Jeong et al. 2013
GSE50201	CMSSeq	24474761	Huang et al. 2014
GSE46202	hMeSealSeq	24672749	Zhu et al. 2014
GSE47894	hMeSealSeq	24757056	Leung et al. 2014
GSE59121	CMSIP	25071199	Tsagaratou et al. 2014
GSE59213	CMSIP	25071199	Tsagaratou et al. 2014
GSE59718	hMeSealSeq	26239807	Delatte et al. 2015
GSE73611	hMeSealSeq	26586431	Zhao et al. 2015
GSE77845	GLIBSeq	27160912	Montagner et al. 2016
GSE77967	hMeSeal	27477909	Han et al. 2016
GSE66847	hMeDIPSeq	27489048	Greco et al. 2016
GSE66834	CMSIP	27869820	Tsagaratou et al. 2017
E-MTAB-5167	hMeDIPSeq	28125731	Lin et al. 2017
GSE74390	hMeSealSeq	28440315	Pan et al. 2017
GSE98964	hMeDIPSeq	28813659	Pan et al. 2017
GSE42880	hMeDIPSeq	28847947	Mellén et al. 2017
GSE81222	hMeDIPSeq	29020633	Kweon et al. 2017
GSE100073	hMeSealSeq	29290626	Li et al. 2018
GSE95720	hMeDIPSeq	29482634	Coluccio et al. 2018
GSE104828	hMeSealSeq	29908294	Chu et al. 2018
GSE100957	CMSSeq	30001199	Gu et al. 2018
GSE117919	hMeDIP	30220521	Sardina et al. 2018
GSE111700	hMeDIPSeq	30274972	Dominguez et al. 2018
GSE109540	hMeSealSeq	30325306	Hrit et al. 2018
GSE119500	hMeSealSeq	30325306	Hrit et al. 2018
GSE47966	CMSIP	23828890	Lister et al. 2013
GSE116208	CMSIP	31028100	Lio et al. 2019
GSE119077	hMeSeal	31583744	Tran et al. 2019
GSE131442	hMeSealSeq	31862843	Raab et al. 2019
GSE113694	CMSIP	34288360	Yue et al. 2021

Table S2.2. Publication and GEO/project information of 5hmC Input downloaded data.

GEO/Project	Sequencing Technique	PubmedID	Citation
GSE28682	CMSSeq	21552279	Pastor et al. 2011
GSE28682	GLIBSeq	21552279	Pastor et al. 2011
GSE49191	CMSSeq	24270360	Jeong et al. 2013
GSE50201	CMSSeq	24474761	Huang et al. 2014
GSE59121	CMSIP	25071199	Tsagaratou et al. 2014
GSE59213	CMSIP	25071199	Tsagaratou et al. 2014
GSE77967	hMeSeal	27477909	Han et al. 2016
GSE66847	hMeDIPSeq	27489048	Greco et al. 2016
GSE66834	CMSIP	27869820	Tsagaratou et al. 2017
E-MTAB-5167	hMeDIPSeq	28125731	Lin et al. 2017
GSE42880	hMeDIPSeq	28847947	Mellén et al. 2017
GSE81222	hMeDIPSeq	29020633	Kweon et al. 2017
GSE95720	hMeDIPSeq	29482634	Coluccio et al. 2018
GSE100957	CMSSeq	30001199	Gu et al. 2018
GSE111700	hMeDIPSeq	30274972	Dominguez et al. 2018
GSE47966	CMSIP	23828890	Lister et al. 2013
GSE116208	CMSIP	31028100	Lio et al. 2019
GSE119077	hMeSeal	31583744	Tran et al. 2019
GSE131442	hMeSealSeq	31862843	Raab et al. 2019
GSE113694	CMSIP	34288360	Yue et al. 2021

Table S2.3. Publication and GEO/project information of gene expression profiling downloaded data.

GEO/Project	Sequencing Technique	PubmedID	Citation
PRJNA30467	RNASeq	18516045	Mortazavi et al. 2018
PRJEB2462	RNASeq	21460836	Ficz et al. 2011
GSE20898	RNASeq	21867929	Wei et al. 2011
GSE30578	RNASeq	22379031	Kirigin et al. 2012
GSE31234	RNASeq	22500808	Zhang et al. 2012
GSE31235	RNASeq	22500808	Zhang et al. 2012
GSE44566	RNASeq	23987249	Neri et al. 2013
GSE49191	RNASeq	24270360	Jeong et al. 2013
GSE50201	RNASeq	24474761	Huang et al. 2014
GSE47894	RNASeq	24757056	Leung et al. 2014
GSE59121	RNASeq	25071199	Tsagaratou et al. 2014
GSE59213	RNASeq	25071199	Tsagaratou et al. 2014
GSE60101	RNASeq	25103404	Lara-Astiaso et al. 2014
GSE73611	RNASeq	26586431	Zhao et al. 2015
GSE72628	RNASeq	26607761	An et al. 2015
GSE77845	RNASeq	27160912	Montagner et al. 2016
GSE77967	RNASeq	27477909	Han et al. 2016
GSE66847	RNASeq	27489048	Greco et al. 2016
GSE66834	RNASeq	27869820	Tsagaratou et al. 2017
GSE71513	RNASeq	27941798	Mathur et al. 2016
E-MTAB-5167	RNASeq	28125731	Lin et al. 2017
GSE98964	RNASeq	28813659	Pan et al. 2017
GSE42880	RNASeq	28847947	Mellén et al. 2017
GSE81222	RNASeq	29020633	Kweon et al. 2017
GSE100073	RNASeq	29290626	Li et al. 2018
GSE95720	RNASeq	29482634	Coluccio et al. 2018
GSE104828	RNASeq	29908294	Chu et al. 2018
GSE100957	RNASeq	30001199	Gu et al. 2018
GSE115714	RNASeq	30212902	Lloret-Llinares et al. 2018
GSE111700	RNASeq	30274972	Dominguez et al. 2018
GSE109540	RNASeq	30325306	Hrit et al. 2018
GSE119500	RNASeq	30325306	Hrit et al. 2018
GSE47966	RNASeq	23828890	Lister et al. 2013
GSE116208	RNASeq	31028100	Lio et al. 2019
GSE130898	RNASeq	31371437	Welte et al. 2019
GSE127933	RNASeq	31519808	Vanheer et al. 2019
GSE119077	RNASeq	31583744	Tran et al. 2019
GSE131442	RNASeq	31862843	Raab et al. 2019
GSE113694	RNASeq	34288360	Yue et al. 2021
GSE59213	RNASeq	25071199	Tsagaratou et al. 2014
GSE60101	RNASeq	25103404	Lara-Astiaso et al. 2014

Table S2.4. Associated PMID study of 5hmC, input and expression profile per sample.

		Pubmed ID Project for		
SampleName	**Replicates**	**5hmC**	**Input**	**Expression**
129s4_mESC	1-2	30001199	30001199	30001199
129s4_mESC	3-4	29482634	29482634	29482634
129s4_mESC	5	29020633	29020633	29020633
Bcell_Activated_24	1-2	31028100	31028100	31028100
Bcell_Activated_48	1-2	31028100	31028100	31028100
Bcell_Activated_72	1-2	31028100	31028100	31028100
Bcell_Resting	1-2	31028100	31028100	31028100
Bergmann_Glia	1-2	28847947	28847947	28847947
CD4_Naive_Tcell	1-2	25071199	25071199	25071199
CD4_SinglePositive_Tcell	1-2	25071199	25071199	22379031
CD8_Naive_Tcell	1-2	25071199	25071199	25071199
CD8_SinglePositive_Tcell	1-2	25071199	25071199	27869820
Cardiomyocites_Adult	1-2	27489048	27489048	27489048
Cardiomyocites_E14.5	1-2	27489048	27489048	27489048
Cardiomyocites_E14.5_shControl	1-2	27489048	27489048	27489048
Cardiomyocites_Neonatal	1-2	27489048	27489048	27489048
Cardiomyocites_TAC	1-2	27489048	27489048	27489048
Colon_Epithelia	1-2	26239807	31862843	27941798
DoublePositive_Tcell	1-2	25071199	25071199	22500808
E13_Frontal_Cortex	1	23828890	23828890	23828890
E14_Embryonic_Bodies	1-2	21460836	29020633	30212902
E14_mESC	1-3	21460836	29020633	30212902
E14_mESC_shGFP	1	23987249	27477909	23987249
Female_Cerebellum	1	28125731	28125731	28125731
Female_Cortex	1	28125731	28125731	28125731
Female_Hippocampus	1	28125731	28125731	28125731
Female_Hypothalamus	1	28125731	28125731	28125731
Female_Liver	1	28125731	28125731	28125731
Female_Thalamus	1	28125731	28125731	28125731
Germinal_Center_B	1-2	30274972	30274972	30274972
Granule_Cells	1-2	28847947	28847947	28847947
Common_Myeloyd_Prog	1-3	27477909	27477909	25103404
Granulocyte_Monocyte_Prog	1-3	27477909	27477909	25103404
Hematopo_LSK	1-3	27477909	27477909	26607761
Megakaryo_Erythr_Prog	1-3	27477909	27477909	25103404
Hematopoietic_Stem_Cells	1	24270360	24270360	24270360
Intestinal_Epithelial_Sox_High	1-3	31862843	31862843	31862843
Intestinal_Epithelial_Sox_Low	1-3	31862843	31862843	31862843
Intestinal_Epithelial_Sox_Negative	1-3	31862843	31862843	31862843
Intestinal_Epithelial_Sox_Sublow	1-3	31862843	31862843	31862843
J1_mESC	1-2	21460836	29020633	21460836
LF2_mESC	1-2	30325306	29020633	30325306
Leukemia_GMP	1-5	27477909	27477909	27477909
Leukemia_Multipotent_Prog	1-5	27477909	27477909	27477909
Liver	1	23987249	28125731	28125731
MEF	1-2	30220521	27477909	31519808
MEF	3-4	23987249	27489048	31519808
MEF	5	21460836	27489048	31519808
Macrophages	1-2	29908294	24270360	29908294
Macrophages	3-4	28813659	24270360	28813659
Male_Cerebellum	1	28125731	28125731	28125731
Male_Cortex	1	28125731	28125731	28125731

Table S2.4. Associated PMID study of 5hmC, input and expression profile per sample (continued).

		Pubmed ID Project for		
SampleName	**Replicates**	**5hmC**	**Input**	**Expression**
Male_Hippocampus	1	28125731	28125731	28125731
Male_Hypothalamus	1	28125731	28125731	28125731
Male_Liver	1	28125731	28125731	28125731
Male_Thalamus	1	28125731	28125731	28125731
Mast_Cells_BMMC	1	27160912	27477909	27160912
Mesenchymal_Stem-Progenitors	1-2	29290626	27489048	29290626
Myeloid_Progenitors	1-2	26586431	27477909	26586431
Myeloid_Progenitors	3-4	28440315	27477909	26586431
Naive_T_Cell	1-2	34288360	34288360	34288360
P70_10weeks_Frontal_Cortex	1	23828890	23828890	23828890
Purkinje_Cells	1-2	28847947	28847947	28847947
Regulatory_T_Cell	1-2	34288360	34288360	34288360
Th1	1-2	25071199	25071199	21867929
V6.5_mESC_00h_wAA_2i	1	31583744	31583744	31583744
V6.5_mESC_12h_wAA_2i	1	31583744	31583744	31583744
V6.5_mESC_72h_wAA_2i	1	31583744	31583744	31583744
V6.5_mESC	1-3	24474761	24474761	24474761
V6.5_mESC	4-6	21552279	21552279	24474761
V6.5_mESC	7-12	27477909	27477909	31371437
WholeBrain	1	23987249	23828890	ENCODE
iNK_T_Cell	1-2	27869820	27869820	27869820
mESC_UnspecifiedBackground	1	24757056	24474761	24757056
pre_iPSC_Day00_wDMSO	1-3	31583744	31583744	31583744
pre_iPSC_Day02_wAA_2i	1-3	31583744	31583744	31583744
pre_iPSC_Day10_wAA_2i	1-3	31583744	31583744	31583744

Table S2.5. Calculated AUC scores per Model (as referred in the manuscript) for each sample's test dataset.

Samples	Models Trained in one sample				Multi-sample Models			
	LRg	RFo	SVM	FCDNN	Whole	Immuno	ESC	Subgroup
129s4_mESC_Rep1	**0.681**	0.679	0.632	0.665	0.645	0.653	0.639	ESC
129s4_mESC_Rep2	0.678	0.669	0.625	0.719	0.690	0.694	0.693	ESC
129s4_mESC_Rep3	0.806	0.755	0.739	0.827	0.783	0.679	0.819	ESC
129s4_mESC_Rep4	0.752	0.732	0.677	0.774	0.760	0.746	0.760	ESC
129s4_mESC_Rep5	0.614	0.610	0.582	0.635	0.644	0.636	0.650	ESC
Bcell_Activated_24_Rep1	0.889	0.843	0.820	0.917	0.891	0.915	0.841	Immuno
Bcell_Activated_24_Rep2	0.860	0.841	0.787	0.895	0.831	0.876	0.758	Immuno
Bcell_Activated_48_Rep1	0.899	0.847	0.831	0.921	0.901	0.922	0.861	Immuno
Bcell_Activated_48_Rep2	0.854	0.835	0.778	0.882	0.793	0.847	0.720	Immuno
Bcell_Activated_72_Rep1	0.885	0.822	0.813	0.930	0.919	0.932	0.884	Immuno
Bcell_Activated_72_Rep2	0.838	0.812	0.759	0.876	0.786	0.843	0.709	Immuno
Bcell_Resting_Rep1	0.903	0.838	0.832	0.918	0.898	0.919	0.859	Immuno
Bcell_Resting_Rep2	0.862	0.833	0.777	0.904	0.841	0.881	0.759	Immuno
Bergmann_Glia_Rep1	0.620	0.619	0.587	0.634	0.424	0.460	0.396	-
Bergmann_Glia_Rep2	0.828	0.782	0.762	0.856	0.857	0.829	0.856	-
CD4_Naive_Tcell_Rep1	0.890	0.840	0.802	0.906	0.837	0.885	0.773	Immuno
CD4_Naive_Tcell_Rep2	0.926	0.855	0.851	0.942	**0.929**	0.941	0.902	Immuno
CD4_SP_Tcell_Rep1	0.901	0.835	0.832	0.924	0.893	0.924	0.843	Immuno
CD4_SP_Tcell_Rep2	0.561	0.598	0.531	0.587	0.504	0.502	0.496	Immuno
CD8_Naive_Tcell_Rep1	0.903	0.841	0.827	0.930	0.897	0.931	0.848	Immuno
CD8_Naive_Tcell_Rep2	0.931	0.858	0.854	0.945	**0.936**	0.946	0.909	Immuno
CD8_SP_Tcell_Rep1	0.926	0.866	0.858	0.940	**0.932**	0.942	0.901	Immuno
CD8_SP_Tcell_Rep2	0.919	0.856	0.855	0.943	**0.940**	0.948	0.908	Immuno
Cardiomyo_Adult_Rep1	0.795	0.719	0.732	0.837	0.829	0.840	0.811	-
Cardiomyo_Adult_Rep2	0.767	0.695	0.709	0.821	0.822	0.813	0.812	-
Cardiomyo_E14.5_Rep1	0.690	0.659	0.642	0.744	0.753	0.737	0.740	-
Cardiomyo_E14.5_Rep2	0.702	0.656	0.642	0.751	0.757	0.721	0.743	-
Cardiomyo_E14.5_shCtl_Rep1	0.595	0.577	0.568	0.609	0.619	0.636	0.604	-
Cardiomyo_E14.5_shCtl_Rep2	0.589	0.588	0.570	0.647	0.615	0.646	0.596	-
Cardiomyo_Neonatal_Rep1	0.764	0.684	0.708	0.800	0.804	0.796	0.784	-
Cardiomyo_Neonatal_Rep2	0.798	0.718	0.724	0.821	0.826	0.804	0.812	-
Cardiomyo_TAC_Rep1	0.793	0.697	0.717	0.836	0.827	0.830	0.814	-
Cardiomyo_TAC_Rep2	0.767	0.698	0.702	0.813	0.799	0.805	0.785	-
Colon_Epithelia_Rep1	**0.902**	0.812	0.822	0.899	0.900	0.902	0.880	-
Colon_Epithelia_Rep2	**0.906**	0.823	0.831	0.905	0.916	0.911	0.899	-
DoublePositive_Tcell_Rep1	0.797	0.770	0.728	0.825	0.754	0.808	0.710	Immuno
DoublePositive_Tcell_Rep2	0.882	0.818	0.806	0.903	0.866	0.900	0.820	Immuno
E13_Frontal_Cortex_Rep1	0.810	0.783	0.743	0.865	0.830	0.811	0.801	-
E14_Embryonic_Bodies_Rep1	0.592	0.612	0.556	0.636	0.526	0.525	0.573	ESC
E14_Embryonic_Bodies_Rep2	0.608	0.623	0.581	0.646	0.533	0.549	0.548	ESC
E14_mESC_Rep1	0.608	0.658	0.574	0.667	0.566	0.567	0.595	ESC
E14_mESC_Rep2	0.612	0.673	0.582	0.720	0.618	0.628	0.634	ESC
E14_mESC_Rep3	0.807	0.779	0.737	0.851	0.844	0.794	0.857	ESC
E14_mESC_shGFP_Rep1	0.826	0.770	0.747	0.863	0.849	0.781	0.871	ESC
Female_Cerebellum_Rep1	0.834	0.779	0.753	0.879	0.878	0.868	0.862	-
Female_Cortex_Rep1	0.850	0.799	0.788	0.859	0.860	0.820	0.869	-
Female_Hippocampus_Rep1	0.837	0.790	0.774	0.870	0.880	0.846	0.879	-
Female_Hypothalamus_Rep1	0.849	0.799	0.778	0.895	0.895	0.872	0.888	-
Female_Liver_Rep1	0.862	0.808	0.775	0.890	0.866	0.878	0.846	-
Female_Thalamus_Rep1	0.848	0.799	0.799	0.882	0.882	0.857	0.881	-
Germinal_Center_B_Rep1	0.738	0.710	0.672	0.760	0.750	0.770	0.699	Immuno
Germinal_Center_B_Rep2	0.701	0.686	0.633	0.724	0.704	0.728	0.654	Immuno
Granule_Cells_Rep1	0.864	0.807	0.786	0.895	0.894	0.885	0.874	-

Table S2.5. Calculated AUC scores per Model (as referred in the manuscript) for each sample's test dataset (continued).

Samples	Models Trained in one sample				Multi-sample Models			
	LRg	RFo	SVM	FCDNN	Whole	Immuno	ESC	Subgroup
Granule_Cells_Rep2	0.812	0.743	0.729	0.853	0.853	0.847	0.835	-
Hematopo_CMP_Rep1	0.851	0.790	0.785	0.904	0.902	0.906	0.874	Immuno
Hematopo_CMP_Rep2	0.828	0.777	0.771	0.890	0.888	0.891	0.857	Immuno
Hematopo_CMP_Rep3	0.827	0.779	0.766	0.887	0.884	0.884	0.859	Immuno
Hematopo_GMP_Rep1	0.826	0.785	0.753	0.894	0.898	0.897	0.880	Immuno
Hematopo_GMP_Rep2	0.801	0.729	0.725	0.877	0.867	0.880	0.836	Immuno
Hematopo_GMP_Rep3	0.796	0.738	0.724	0.869	0.870	0.876	0.848	Immuno
Hematopo_LSK_Rep1	0.839	0.789	0.783	0.883	0.876	0.876	0.847	Immuno
Hematopo_LSK_Rep2	0.824	0.776	0.755	0.872	0.869	0.872	0.835	Immuno
Hematopo_LSK_Rep3	0.814	0.773	0.745	0.863	0.861	0.864	0.832	Immuno
Hematopo_MEP_Rep1	0.846	0.789	0.760	0.904	0.905	0.910	0.880	Immuno
Hematopo_MEP_Rep2	0.796	0.744	0.737	0.863	0.869	0.871	0.852	Immuno
Hematopo_MEP_Rep3	0.824	0.778	0.753	0.884	0.884	0.892	0.860	Immuno
Hematopo_Stem_Cells_Rep1	0.847	0.791	0.772	0.866	0.863	0.840	0.850	Immuno
Intesti_Epithe_Sox_H_Rep1	0.891	0.823	0.807	0.903	0.903	0.892	0.891	-
Intesti_Epithe_Sox_H_Rep2	0.877	0.809	0.795	0.898	0.900	0.887	0.886	-
Intesti_Epithe_Sox_H_Rep3	0.878	0.804	0.796	0.897	0.900	0.881	0.893	-
Intesti_Epithe_Sox_L_Rep1	0.887	0.834	0.818	0.901	0.905	0.894	0.890	-
Intesti_Epithe_Sox_L_Rep2	0.873	0.826	0.794	0.897	0.900	0.887	0.886	-
Intesti_Epithe_Sox_L_Rep3	0.877	0.825	0.804	0.907	0.911	0.900	0.896	-
Intesti_Epithe_Sox_N_Rep1	0.892	0.824	0.806	0.907	0.908	0.899	0.894	-
Intesti_Epithe_Sox_N_Rep2	0.875	0.812	0.795	0.895	0.899	0.892	0.883	-
Intesti_Epithe_Sox_N_Rep3	0.889	0.821	0.811	0.905	0.908	0.902	0.895	-
Intesti_Epithe_Sox_S_Rep1	0.883	0.824	0.805	0.906	0.912	0.901	0.901	-
Intesti_Epithe_Sox_S_Rep2	0.890	0.835	0.814	0.909	0.912	0.899	0.900	-
Intesti_Epithe_Sox_S_Rep3	0.887	0.826	0.807	0.908	0.914	0.902	0.903	-
J1_mESC_Rep1	0.622	0.650	0.591	0.692	0.582	0.589	0.612	-
J1_mESC_Rep2	0.657	0.675	0.622	0.752	0.667	0.674	0.676	-
LF2_mESC_Rep1	0.677	0.667	0.634	0.750	0.570	0.577	0.609	-
LF2_mESC_Rep2	0.693	0.666	0.638	0.719	0.594	0.606	0.624	-
Leukemia_GMP_Rep1	0.820	0.758	0.751	0.881	0.886	0.889	0.860	-
Leukemia_GMP_Rep2	0.833	0.747	0.752	0.886	0.891	0.894	0.874	-
Leukemia_GMP_Rep3	0.866	0.782	0.785	0.897	0.892	0.904	0.866	-
Leukemia_GMP_Rep4	0.854	0.766	0.770	0.902	0.900	0.906	0.879	-
Leukemia_GMP_Rep5	0.833	0.752	0.750	0.880	0.882	0.893	0.857	-
Leukemia_MPP_Rep1	0.828	0.754	0.752	0.864	0.871	0.870	0.852	-
Leukemia_MPP_Rep2	0.824	0.755	0.755	0.864	0.871	0.874	0.847	-
Leukemia_MPP_Rep3	0.829	0.751	0.764	0.879	0.884	0.885	0.853	-
Leukemia_MPP_Rep4	0.800	0.720	0.730	0.851	0.858	0.852	0.841	-
Leukemia_MPP_Rep5	0.804	0.726	0.737	0.847	0.848	0.856	0.821	-
Liver_Rep1	0.889	0.806	0.811	0.910	0.910	0.895	0.908	-
MEF_Rep1	0.586	0.601	0.559	0.622	0.536	0.562	0.496	-
MEF_Rep2	0.579	0.613	0.550	0.643	0.551	0.576	0.506	-
MEF_Rep3	0.793	0.736	0.725	0.795	0.790	0.774	0.768	-
MEF_Rep4	0.669	0.669	0.612	0.730	0.701	0.705	0.685	-
MEF_Rep5	0.604	0.630	0.574	0.654	0.641	0.645	0.621	-
Macrophages_Rep1	0.824	0.754	0.746	0.851	0.837	0.840	0.820	-
Macrophages_Rep2	0.762	0.708	0.699	0.795	0.784	0.782	0.778	-
Macrophages_Rep3	0.524	0.538	0.514	0.556	0.547	0.557	0.534	-
Macrophages_Rep4	0.525	0.561	0.516	0.533	0.555	0.555	0.546	-
Male_Cerebellum_Rep1	0.853	0.783	0.774	0.882	0.884	0.873	0.865	-
Male_Cortex_Rep1	0.838	0.786	0.766	0.881	0.874	0.832	0.882	-
Male_Hippocampus_Rep1	0.808	0.759	0.734	0.858	0.859	0.829	0.857	-

Table S2.5. Calculated AUC scores per Model (as referred in the manuscript) for each sample's test dataset (continued).

Samples	Models Trained in one sample				Multi-sample Models			
	LRg	RFo	SVM	FCDNN	Whole	Immuno	ESC	Subgroup
Male_Hypothalamus_Rep1	0.849	0.805	0.79	0.896	0.890	0.864	0.888	-
Male_Liver_Rep1	0.806	0.741	0.722	0.852	0.778	0.807	0.759	-
Male_Thalamus_Rep1	0.843	0.801	0.790	0.890	0.887	0.850	0.887	-
Mast_Cells_BMMC_Rep1	**0.740**	0.686	0.678	0.732	0.731	0.729	0.703	Immuno
Mesenchymal_StemProge_Rep1	0.690	0.646	0.628	0.758	0.757	0.769	0.721	-
Mesenchymal_StemProge_Rep2	0.705	0.638	0.628	0.754	0.751	0.769	0.726	-
Myeloid_Progenitors_Rep1	0.838	0.784	0.761	0.879	0.885	0.885	0.868	Immuno
Myeloid_Progenitors_Rep2	0.846	0.794	0.784	0.890	0.891	0.893	0.876	Immuno
Myeloid_Progenitors_Rep3	0.838	0.787	0.761	0.879	0.885	0.885	0.868	Immuno
Myeloid_Progenitors_Rep4	0.846	0.792	0.784	0.890	0.891	0.893	0.876	Immuno
Naive_T_Cell_Rep1	0.919	0.856	0.837	0.943	**0.939**	0.948	0.908	Immuno
Naive_T_Cell_Rep2	0.910	0.846	0.832	0.944	**0.940**	0.949	0.909	Immuno
P70_10w_F_Cortex_Rep1	0.858	0.819	0.799	0.884	0.877	0.832	0.882	-
Purkinje_Cells_Rep1	0.859	0.814	0.789	0.887	0.886	0.857	0.877	-
Purkinje_Cells_Rep2	0.844	0.785	0.777	0.871	0.872	0.845	0.862	-
Regulatory_T_Cell_Rep1	0.908	0.849	0.828	0.941	**0.934**	0.945	0.900	Immuno
Regulatory_T_Cell_Rep2	0.916	0.845	0.835	0.937	**0.924**	0.935	0.887	Immuno
Th1_Rep1	0.671	0.648	0.620	0.679	0.631	0.675	0.607	Immuno
Th2_Rep1	**0.767**	0.696	0.695	0.755	0.694	0.745	0.665	Immuno
V6.5_mESC_00h_wAA_2i_Rep1	0.805	0.774	0.733	0.814	0.801	0.732	0.824	ESC
V6.5_mESC_12h_wAA_2i_Rep1	0.853	0.809	0.793	0.883	0.851	0.760	0.874	ESC
V6.5_mESC_72h_wAA_2i_Rep1	**0.844**	0.791	0.784	0.838	0.775	0.653	0.817	ESC
V6.5_mESC_Rep1	0.747	0.730	0.688	0.792	0.765	0.740	0.783	ESC
V6.5_mESC_Rep10	0.849	0.818	0.779	0.868	0.837	0.751	0.866	ESC
V6.5_mESC_Rep11	0.839	0.794	0.759	0.860	0.830	0.731	0.859	ESC
V6.5_mESC_Rep12	0.812	0.761	0.743	0.849	0.808	0.710	0.842	ESC
V6.5_mESC_Rep2	0.750	0.742	0.679	0.786	0.763	0.745	0.780	ESC
V6.5_mESC_Rep3	0.773	0.754	0.703	0.802	0.792	0.755	0.806	ESC
V6.5_mESC_Rep4	0.782	0.751	0.728	0.808	0.795	0.747	0.813	ESC
V6.5_mESC_Rep5	0.781	0.756	0.728	0.836	0.812	0.774	0.826	ESC
V6.5_mESC_Rep6	0.744	0.694	0.682	0.785	0.773	0.727	0.788	ESC
V6.5_mESC_Rep7	0.820	0.770	0.753	0.853	0.809	0.707	0.844	ESC
V6.5_mESC_Rep8	0.804	0.758	0.730	0.836	0.797	0.698	0.826	ESC
V6.5_mESC_Rep9	0.857	0.800	0.788	0.878	0.848	0.760	0.877	ESC
WholeBrain_Rep1	0.835	0.799	0.789	0.873	0.843	0.772	0.859	-
iNK_T_Cell_Rep1	0.920	0.857	0.843	0.936	**0.933**	0.943	0.900	Immuno
iNK_T_Cell_Rep2	0.919	0.855	0.837	0.938	**0.932**	0.942	0.898	Immuno
mESC_UnspecBkGrnd_Rep1	0.783	0.765	0.724	0.801	0.786	0.708	0.807	-
pre_iPSC_Day00_wDMSO_Rep1	0.820	0.762	0.754	0.840	0.844	0.792	0.856	ESC
pre_iPSC_Day00_wDMSO_Rep2	0.829	0.770	0.751	0.850	0.850	0.804	0.857	ESC
pre_iPSC_Day00_wDMSO_Rep3	0.787	0.750	0.724	0.827	0.819	0.774	0.822	ESC
pre_iPSC_Day02_wAA_2i_Rep1	0.863	0.813	0.803	**0.903**	0.883	0.834	0.894	ESC
pre_iPSC_Day02_wAA_2i_Rep2	0.866	0.812	0.809	**0.905**	0.886	0.840	0.896	ESC
pre_iPSC_Day02_wAA_2i_Rep3	0.848	0.790	0.781	0.893	0.881	0.835	0.891	ESC
pre_iPSC_Day10_wAA_2i_Rep1	0.869	0.793	0.796	0.887	0.867	0.795	0.886	ESC
pre_iPSC_Day10_wAA_2i_Rep2	0.870	0.808	0.803	0.886	0.868	0.792	0.887	ESC
pre_iPSC_Day10_wAA_2i_Rep3	0.859	0.795	0.789	0.886	0.864	0.780	0.886	ESC
Mixed Models' Test dataset					0.8200	0.8750	0.8069	-

Table S2.6. Calculated F1 scores per Model (as referred in the manuscript) for each sample's test dataset.

Samples	FCDNN	Whole	Immuno	ESC	Subgroup
129s4_mESC_Rep1	0.630	0.634	0.636	0.628	ESC
129s4_mESC_Rep2	0.695	0.647	0.656	0.644	ESC
129s4_mESC_Rep3	0.766	0.756	0.637	0.774	ESC
129s4_mESC_Rep4	0.719	0.699	0.675	0.712	ESC
129s4_mESC_Rep5	0.609	0.572	0.444	0.592	ESC
Bcell_Activated_24_Rep1	0.852	0.826	0.850	0.737	Immuno
Bcell_Activated_24_Rep2	0.833	0.761	0.820	0.635	Immuno
Bcell_Activated_48_Rep1	0.859	0.825	0.854	0.764	Immuno
Bcell_Activated_48_Rep2	0.810	0.729	0.795	0.615	Immuno
Bcell_Activated_72_Rep1	0.864	0.849	0.867	0.786	Immuno
Bcell_Activated_72_Rep2	0.814	0.726	0.794	0.598	Immuno
Bcell_Resting_Rep1	0.849	0.812	0.845	0.748	Immuno
Bcell_Resting_Rep2	0.837	0.781	0.821	0.649	Immuno
Bergmann_Glia_Rep1	0.533	0.153	0.093	0.201	-
Bergmann_Glia_Rep2	0.789	0.805	0.779	0.806	-
CD4_Naive_Tcell_Rep1	0.844	0.775	0.828	0.688	Immuno
CD4_Naive_Tcell_Rep2	0.892	0.862	0.880	0.812	Immuno
CD4_SP_Tcell_Rep1	0.864	0.815	0.857	0.735	Immuno
CD4_SP_Tcell_Rep2	0.630	0.486	0.444	0.504	Immuno
CD8_Naive_Tcell_Rep1	0.871	0.830	0.868	0.769	Immuno
CD8_Naive_Tcell_Rep2	0.875	0.865	0.879	0.813	Immuno
CD8_SP_Tcell_Rep1	0.873	0.856	0.868	0.782	Immuno
CD8_SP_Tcell_Rep2	0.876	0.862	0.873	0.795	Immuno
Cardiomyo_Adult_Rep1	0.759	0.764	0.775	0.745	-
Cardiomyo_Adult_Rep2	0.747	0.755	0.744	0.744	-
Cardiomyo_E14.5_Rep1	0.665	0.657	0.609	0.646	-
Cardiomyo_E14.5_Rep2	0.683	0.672	0.596	0.652	-
Cardiomyo_E14.5_shCtl_Rep1	0.592	0.494	0.474	0.509	-
Cardiomyo_E14.5_shCtl_Rep2	0.594	0.457	0.492	0.458	-
Cardiomyo_Neonatal_Rep1	0.730	0.724	0.713	0.714	-
Cardiomyo_Neonatal_Rep2	0.751	0.749	0.716	0.739	-
Cardiomyo_TAC_Rep1	0.768	0.749	0.735	0.724	-
Cardiomyo_TAC_Rep2	0.741	0.730	0.726	0.727	-
Colon_Epithelia_Rep1	0.817	0.824	0.808	0.791	-
Colon_Epithelia_Rep2	0.829	0.853	0.835	0.831	-
DoublePositive_Tcell_Rep1	0.771	0.633	0.718	0.557	Immuno
DoublePositive_Tcell_Rep2	0.836	0.756	0.806	0.677	Immuno
E13_Frontal_Cortex_Rep1	0.807	0.761	0.731	0.703	-
E14_Embryonic_Bodies_Rep1	0.636	0.543	0.637	0.634	ESC
E14_Embryonic_Bodies_Rep2	0.657	0.473	0.601	0.573	ESC
E14_mESC_Rep1	0.664	0.587	0.659	0.662	ESC
E14_mESC_Rep2	0.684	0.560	0.626	0.626	ESC
E14_mESC_Rep3	0.790	0.790	0.748	0.790	ESC
E14_mESC_shGFP_Rep1	0.804	0.788	0.708	0.807	ESC
Female_Cerebellum_Rep1	0.809	0.815	0.813	0.814	-
Female_Cortex_Rep1	0.826	0.810	0.786	0.820	-
Female_Hippocampus_Rep1	0.820	0.811	0.790	0.818	-
Female_Hypothalamus_Rep1	0.848	0.824	0.805	0.827	-
Female_Liver_Rep1	0.822	0.810	0.821	0.800	-
Female_Thalamus_Rep1	0.838	0.819	0.806	0.831	-
Germinal_Center_B_Rep1	0.687	0.558	0.506	0.474	Immuno
Germinal_Center_B_Rep2	0.663	0.470	0.404	0.422	Immuno
Granule_Cells_Rep1	0.830	0.824	0.824	0.815	-

Table S2.6. Calculated F1 scores per Model (as referred in the manuscript) for each sample's test dataset (continued).

Samples	FCDNN	Whole	Immuno	ESC	Subgroup
Granule_Cells_Rep2	0.787	0.800	0.779	0.786	-
Hematopo_CMP_Rep1	0.832	0.837	0.840	0.789	Immuno
Hematopo_CMP_Rep2	0.826	0.833	0.830	0.772	Immuno
Hematopo_CMP_Rep3	0.826	0.810	0.814	0.759	Immuno
Hematopo_GMP_Rep1	0.826	0.823	0.819	0.781	Immuno
Hematopo_GMP_Rep2	0.813	0.801	0.817	0.773	Immuno
Hematopo_GMP_Rep3	0.804	0.810	0.818	0.780	Immuno
Hematopo_LSK_Rep1	0.820	0.815	0.815	0.774	Immuno
Hematopo_LSK_Rep2	0.813	0.808	0.813	0.757	Immuno
Hematopo_LSK_Rep3	0.801	0.802	0.807	0.764	Immuno
Hematopo_MEP_Rep1	0.835	0.831	0.835	0.795	Immuno
Hematopo_MEP_Rep2	0.790	0.747	0.771	0.703	Immuno
Hematopo_MEP_Rep3	0.820	0.808	0.828	0.753	Immuno
Hematopo_Stem_Cells_Rep1	0.810	0.773	0.726	0.719	Immuno
Intesti_Epithe_Sox_H_Rep1	0.835	0.827	0.834	0.818	-
Intesti_Epithe_Sox_H_Rep2	0.828	0.826	0.827	0.817	-
Intesti_Epithe_Sox_H_Rep3	0.833	0.832	0.817	0.822	-
Intesti_Epithe_Sox_L_Rep1	0.828	0.832	0.835	0.818	-
Intesti_Epithe_Sox_L_Rep2	0.825	0.839	0.828	0.821	-
Intesti_Epithe_Sox_L_Rep3	0.843	0.848	0.843	0.827	-
Intesti_Epithe_Sox_N_Rep1	0.838	0.846	0.840	0.830	-
Intesti_Epithe_Sox_N_Rep2	0.831	0.831	0.833	0.814	-
Intesti_Epithe_Sox_N_Rep3	0.833	0.841	0.837	0.829	-
Intesti_Epithe_Sox_S_Rep1	0.833	0.849	0.836	0.829	-
Intesti_Epithe_Sox_S_Rep2	0.842	0.845	0.832	0.817	-
Intesti_Epithe_Sox_S_Rep3	0.839	0.849	0.839	0.834	-
J1_mESC_Rep1	0.695	0.578	0.659	0.668	-
J1_mESC_Rep2	0.702	0.586	0.657	0.623	-
LF2_mESC_Rep1	0.733	0.357	0.466	0.442	-
LF2_mESC_Rep2	0.695	0.370	0.518	0.447	-
Leukemia_GMP_Rep1	0.812	0.813	0.819	0.775	-
Leukemia_GMP_Rep2	0.828	0.828	0.828	0.787	-
Leukemia_GMP_Rep3	0.829	0.826	0.841	0.792	-
Leukemia_GMP_Rep4	0.839	0.835	0.847	0.810	-
Leukemia_GMP_Rep5	0.806	0.810	0.823	0.789	-
Leukemia_MPP_Rep1	0.800	0.795	0.800	0.755	-
Leukemia_MPP_Rep2	0.805	0.791	0.812	0.742	-
Leukemia_MPP_Rep3	0.812	0.798	0.818	0.757	-
Leukemia_MPP_Rep4	0.783	0.789	0.787	0.760	-
Leukemia_MPP_Rep5	0.775	0.770	0.789	0.727	-
Liver_Rep1	0.837	0.831	0.771	0.814	-
MEF_Rep1	0.617	0.550	0.650	0.487	-
MEF_Rep2	0.593	0.578	0.655	0.486	-
MEF_Rep3	0.746	0.714	0.689	0.695	-
MEF_Rep4	0.657	0.503	0.504	0.480	-
MEF_Rep5	0.611	0.643	0.682	0.624	-
Macrophages_Rep1	0.784	0.434	0.257	0.396	-
Macrophages_Rep2	0.707	0.134	0.044	0.121	-
Macrophages_Rep3	0.596	0.252	0.174	0.262	-
Macrophages_Rep4	0.528	0.283	0.168	0.304	-
Male_Cerebellum_Rep1	0.828	0.824	0.810	0.820	-
Male_Cortex_Rep1	0.827	0.809	0.789	0.813	-
Male_Hippocampus_Rep1	0.806	0.805	0.787	0.812	-

Table S2.6. Calculated F1 scores per Model (as referred in the manuscript) for each sample's test dataset (continued).

Samples	FCDNN	Whole	Immuno	ESC	Subgroup
Male_Hypothalamus_Rep1	0.850	0.816	0.803	0.823	-
Male_Liver_Rep1	0.784	0.745	0.769	0.727	-
Male_Thalamus_Rep1	0.845	0.825	0.803	0.831	-
Mast_Cells_BMMC_Rep1	0.631	0.649	0.570	0.642	Immuno
Mesenchymal_StemProge_Rep1	0.664	0.539	0.381	0.525	-
Mesenchymal_StemProge_Rep2	0.672	0.585	0.485	0.581	-
Myeloid_Progenitors_Rep1	0.814	0.817	0.826	0.804	Immuno
Myeloid_Progenitors_Rep2	0.821	0.816	0.826	0.812	Immuno
Myeloid_Progenitors_Rep3	0.816	0.817	0.826	0.804	Immuno
Myeloid_Progenitors_Rep4	0.821	0.816	0.826	0.812	Immuno
Naive_T_Cell_Rep1	0.876	0.875	0.883	0.821	Immuno
Naive_T_Cell_Rep2	0.884	0.883	0.885	0.838	Immuno
P70_10w_F_Cortex_Rep1	0.838	0.826	0.792	0.832	-
Purkinje_Cells_Rep1	0.829	0.830	0.801	0.820	-
Purkinje_Cells_Rep2	0.816	0.822	0.793	0.819	-
Regulatory_T_Cell_Rep1	0.879	0.870	0.887	0.816	Immuno
Regulatory_T_Cell_Rep2	0.865	0.857	0.870	0.802	Immuno
Th1_Rep1	0.593	0.538	0.565	0.545	Immuno
Th2_Rep1	0.695	0.618	0.630	0.591	Immuno
V6.5_mESC_00h_wAA_2i_Rep1	0.708	0.700	0.586	0.727	ESC
V6.5_mESC_12h_wAA_2i_Rep1	0.805	0.791	0.672	0.811	ESC
V6.5_mESC_72h_wAA_2i_Rep1	0.797	0.714	0.434	0.760	ESC
V6.5_mESC_Rep1	0.737	0.675	0.645	0.691	ESC
V6.5_mESC_Rep10	0.804	0.782	0.695	0.807	ESC
V6.5_mESC_Rep11	0.805	0.778	0.637	0.799	ESC
V6.5_mESC_Rep12	0.795	0.749	0.589	0.785	ESC
V6.5_mESC_Rep2	0.736	0.655	0.626	0.647	ESC
V6.5_mESC_Rep3	0.758	0.717	0.671	0.736	ESC
V6.5_mESC_Rep4	0.764	0.718	0.662	0.739	ESC
V6.5_mESC_Rep5	0.766	0.738	0.705	0.741	ESC
V6.5_mESC_Rep6	0.737	0.666	0.523	0.689	ESC
V6.5_mESC_Rep7	0.791	0.735	0.563	0.772	ESC
V6.5_mESC_Rep8	0.775	0.714	0.510	0.747	ESC
V6.5_mESC_Rep9	0.817	0.787	0.708	0.812	ESC
WholeBrain_Rep1	0.832	0.803	0.692	0.806	-
iNK_T_Cell_Rep1	0.872	0.868	0.876	0.806	Immuno
iNK_T_Cell_Rep2	0.871	0.857	0.882	0.810	Immuno
mESC_UnspecBkGrnd_Rep1	0.756	0.745	0.647	0.754	-
pre_iPSC_Day00_wDMSO_Rep1	0.725	0.746	0.639	0.736	ESC
pre_iPSC_Day00_wDMSO_Rep2	0.734	0.740	0.686	0.726	ESC
pre_iPSC_Day00_wDMSO_Rep3	0.749	0.688	0.641	0.680	ESC
pre_iPSC_Day02_wAA_2i_Rep1	0.827	0.820	0.770	0.834	ESC
pre_iPSC_Day02_wAA_2i_Rep2	0.831	0.817	0.774	0.823	ESC
pre_iPSC_Day02_wAA_2i_Rep3	0.826	0.819	0.772	0.823	ESC
pre_iPSC_Day10_wAA_2i_Rep1	0.813	0.811	0.718	0.821	ESC
pre_iPSC_Day10_wAA_2i_Rep2	0.806	0.806	0.703	0.823	ESC
pre_iPSC_Day10_wAA_2i_Rep3	0.812	0.805	0.676	0.821	ESC
Mixed Models' Test dataset		0.748	0.802	0.738	-

Table S2.7. AUC and F1 scores of the specialized models with and without holding out randomly selected cell types.

AUC [F1] scores		Whole	Immuno	ESC
All Samples	Mean	0.86 [0.8]	0.89 [0.83]	0.83 [0.74]
	Median	0.82 [0.74]	0.87 [0.79]	0.8 [0.74]
Withheld	Mean	0.86 [0.78]	0.90 [0.81]	0.83 [0.74]
	Median	0.80 [0.71]	0.88 [0.79]	0.81 [0.72]
Fraction withheld		21/153	6/46	5/35

2.9 Author Contributions

E.G.-A., F.A and A.R. conceived and designed the project. E.G.-A. designed and performed experiments and analyzed data with F.A. and D.S.-C., E.G.-A. performed data collection and analysis. E.G.-A., F.A. and A.R. wrote the manuscript, with all authors contributing and providing feedback and advice.

2.10 Acknowledgements

Chapter 2, in full, is a reformatted presentation of the material currently being prepared for submission for publication as “Linking proximal and distal 5hmC enrichment to cell-specific gene regulation with graph convolutional networks” by Edahí González-Avalos, Daniela Samaniego-Castruita, Anjana Rao, and Ferhat Ay. The dissertation author was the primary investigator and first author of this material.

2.11 References

Alberge, J.-B., Magrangeas, F., Wagner, M., Denié, S., Guérin-Charbonnel, C., Campion, L., Attal, M., Avet-Loiseau, H., Carell, T., Moreau, P., Minvielle, S. and Sérandour, A.A. (2020). DNA hydroxymethylation is associated with disease severity and persists at enhancers of oncogenic regions in multiple myeloma. Clinical Epigenetics *12*, 163.

An, J., González-Avalos, E., Chawla, A., Jeong, M., López-Moyado, I.F., Li, W., Goodell, M.A., Chavez, L., Ko, M. and Rao, A. (2015). Acute loss of TET function results in aggressive myeloid cancer in mice. Nature Communications *6*, 10071.

An, J., Rao, A., Ko, M. (2017). TET family dioxygenases and DNA demethylation in stem cells and cancers. Experimental & Molecular Medicine *49*, 323–323.

Agarwal, V. and Shendure, J. (2020). Predicting mRNA Abundance Directly from Genomic Sequence Using Deep Convolutional Neural Networks. Cell Reports *31*, 107663.

Avsec, Ž., Agarwal, V., Visentin, D., Ledsam, J.R., Grabska-Barwinska, A., Taylor, K.R., Assael, Y., Jumper, J., Kohli, P. and Kelley, D.R. (2021). Effective gene expression prediction from sequence by integrating long-range interactions. Nature Methods *18*, 1196–1203.

Beer, M.A. and Tavazoie, S., 2004. Predicting gene expression from sequence. Cell *117*, 185–198.

Cartron, P.-F., Nadaradjane, A., LePape, F., Lalier, L., Gardie, B. and Vallette, F.M. (2013). Identification of TET1 Partners That Control Its DNA-Demethylating Function. Genes & Cancer *4*, 235–241.

Cheng, C., Yan, K.-K., Yip, K.Y., Rozowsky, J., Alexander, R., Shou, C. and Gerstein, M. (2011). A statistical framework for modeling gene expression using chromatin features and application to modENCODE datasets. Genome Biology *12*, R15.

Cheung, P., Vallania, F., Warsinske, H.C., Donato, M., Schaffert, S., Chang, S.E., Dvorak, M., Dekker, C.L., Davis, M.M., Utz, P.J., Khatri, P. and Kuo, A.J. (2018). Single-Cell Chromatin Modification Profiling Reveals Increased Epigenetic Variations with Aging. Cell *173*, 1385-1397.e14.

Chu, Y., Zhao, Z., Sant, D.W., Zhu, G., Greenblatt, S.M., Liu, L., Wang, J., Cao, Z., Tho, J.C., Chen, S., Liu, X., Zhang, P., Maciejewski, J.P., Nimer, S., Wang, G., Yuan, W., Yang, F.-C. and Xu, M. (2018). Tet2 Regulates Osteoclast Differentiation by Interacting with Runx1 and Maintaining Genomic 5-Hydroxymethylcytosine (5hmC). Genomics, Proteomics & Bioinformatics *16*, 172–186.

Coluccio, A., Ecco, G., Duc, J., Offner, S., Turelli, P. and Trono, D. (2018). Individual retrotransposon integrants are differentially controlled by KZFP/KAP1-dependent histone methylation, DNA methylation and TET-mediated hydroxymethylation in naïve embryonic stem cells. Epigenetics & Chromatin *11*, 7.

Delatte, B., Jeschke, J., Defrance, M., Bachman, M., Creppe, C., Calonne, E., Bizet, M., Deplus, R., Marroquí, L., Libin, M., Ravichandran, M., Mascart, F., Eizirik, D.L., Murrell, A., Jurkowski, T.P. and Fuks, F. (2015). Genome-wide hydroxymethylcytosine pattern changes in response to oxidative stress. Scientific Reports *5*, 12714.

Dominguez, P.M., Ghamlouch, H., Rosikiewicz, W., Kumar, P., Béguelin, W., Fontan, L., Rivas, M.A., Pawlikowska, P., Armand, M., Mouly, E., Torres-Martin, M., Doane, A.S., Calvo Fernandez, M.T., Durant, M., Della-Valle, V., Teater, M., Cimmino, L., Droin, N., Tadros, S. and Motanagh, S. (2018). TET2 deficiency causes germinal center hyperplasia, impairs plasma cell differentiation and promotes B-cell lymphomagenesis. Cancer Discovery *8*, 1632–1653.

Dong, X., Greven, M.C., Kundaje, A., Djebali, S., Brown, J.B., Cheng, C., Gingeras, T.R., Gerstein, M., Guigó, R., Birney, E. and Weng, Z. (2012). Modeling gene expression using chromatin features in various cellular contexts. Genome Biology *13*, 53.

Eraslan, G., Avsec, Ž., Gagneur, J. and Theis, F.J. (2019). Deep learning: new computational modelling techniques for genomics. Nature Reviews Genetics *20*, 389–403.

Fernández-Delgado, M., Cernadas, E., Barro, S. and Amorim, D. (2014). Do we Need Hundreds of Classifiers to Solve Real World Classification Problems? Journal of Machine Learning Research *15*, 3133–3181.

Ficz, G., Branco, M.R., Seisenberger, S., Santos, F., Krueger, F., Hore, T.A., Marques, C.J., Andrews, S. and Reik, W. (2011). Dynamic regulation of 5-hydroxymethylcytosine in mouse ES cells and during differentiation. Nature *473*, 398–402.

Gabrieli, T., Sharim, H., Nifker, G., Jeffet, J., Shahal, T., Arielly, R., Levi-Sakin, M., Hoch, L., Arbib, N., Michaeli, Y. and Ebenstein, Y. (2018). Epigenetic Optical Mapping of 5-Hydroxymethylcytosine in Nanochannel Arrays. ACS Nano *12*, 7148–7158.

Greco, C.M., Kunderfranco, P., Rubino, M., Larcher, V., Carullo, P., Anselmo, A., Kurz, K., Carell, T., Angius, A., Latronico, M.V.G., Papait, R. and Condorelli, G. (2016). DNA hydroxymethylation controls cardiomyocyte gene expression in development and hypertrophy. Nature Communications *7*, 12418.

Gu, T., Lin, X., Cullen, S.M., Luo, M., Jeong, M., Estecio, M., Shen, J., Hardikar, S., Sun, D., Su, J., Rux, D., Guzman, A., Lee, M., Qi, L.S., Chen, J.-J., Kyba, M., Huang, Y., Chen, T., Li, W. and Goodell, M.A. (2018). DNMT3A and TET1 cooperate to regulate promoter epigenetic landscapes in mouse embryonic stem cells. Genome Biology *19*, 88.

Hagihara, Y., Asada, S., Maeda, T., Nakano, T. and Yamaguchi, S. (2021). Tet1 regulates epigenetic remodeling of the pericentromeric heterochromatin and chromocenter organization in DNA hypomethylated cells. PLOS Genetics *17*, e1009646.

Han, D., Lu, X., Shih, Alan H., Nie, J., You, Q., Xu, M., Melnick, Ari M., Levine, Ross L. and He, C. (2016). A Highly Sensitive and Robust Method for Genome-wide 5hmC Profiling of Rare Cell Populations. Molecular Cell *63*, 711–719.

He, Y.-F., Li, B.-Z., Li, Z., Liu, P., Wang, Y., Tang, Q., Ding, J., Jia, Y., Chen, Z., Li, L., Sun, Y., Li, X., Dai, Q., Song, C.-X., Zhang, K., He, C. and Xu, G.-L. (2011). Tet-Mediated Formation of 5-Carboxylcytosine and Its Excision by TDG in Mammalian DNA. Science *333*, 1303–1307.

Hrit, J., Goodrich, L., Li, C., Wang, B.-A., Nie, J., Cui, X., Martin, E.A., Simental, E., Fernandez, J., Liu, M.Y., Nery, J.R., Castanon, R., Kohli, R.M., Tretyakova, N., He, C., Ecker, J.R., Goll, M. and Panning, B. (2018). OGT binds a conserved C-terminal domain of TET1 to regulate TET1 activity and function in development. eLife *7*, e34870.

Huang, Y., Chavez, L., Chang, X., Wang, X., Pastor, W.A., Kang, J., Zepeda-Martinez, J.A., Pape, U.J., Jacobsen, S.E., Peters, B. and Rao, A. (2014). Distinct roles of the methylcytosine oxidases Tet1 and Tet2 in mouse embryonic stem cells. Proceedings of the National Academy of Sciences *111*, 1361–1366.

Huang, Y., Pastor, W.A., Zepeda-Martínez, J.A. Rao, A. (2012). The anti-CMS technique for genome-wide mapping of 5-hydroxymethylcytosine. Nature Protocols *7*, 1897–1908.

Ito, S., Shen, L., Dai, Q., Wu, S.C., Collins, L.B., Swenberg, J.A., He, C. and Zhang, Y. (2011). Tet Proteins Can Convert 5-Methylcytosine to 5-Formylcytosine and 5-Carboxylcytosine. Science *333*, 1300–1303.

Ivanov, M., Kals, M., Kacevska, M., Barragan, I., Kasuga, K., Rane, A., Metspalu, A., Milani, L. and Ingelman-Sundberg, M. (2013). Ontogeny, distribution and potential roles of 5-hydroxymethylcytosine in human liver function. Genome Biology *14*, 83.

Jeong, M., Sun, D., Luo, M., Huang, Y., Challen, G.A., Rodriguez, B., Zhang, X., Chavez, L., Wang, H., Hannah, R., Kim, S.-B., Yang, L., Ko, M., Chen, R., Göttgens, B., Lee, J.-S., Gunaratne, P., Godley, L.A., Darlington, G.J. and Rao, A. (2013). Large conserved domains of low DNA methylation maintained by Dnmt3a. Nature Genetics *46*, 17–23.

Johnson, D.S., Mortazavi, A., Myers, R.M. and Wold, B., 2007. Genome-wide mapping of in vivo protein-DNA interactions. Science *316*, 1497–1502.

Kirasich, K., Smith, T. and Sadler, B. (2018). Random Forest vs Logistic Regression: Binary Classification for Heterogeneous Datasets. SMU Data Science Review *1*, 9.

Kirigin, F.F., Lindstedt, K., Sellars, M., Ciofani, M., Low, S.L., Jones, L., Bell, F., Pauli, F., Bonneau, R., Myers, R.M., Littman, D.R. and Chong, M.M.W. (2012). Dynamic MicroRNA Gene Transcription and Processing during T Cell Development. The Journal of Immunology *188*, 3257–3267.

Kriaucionis, S. and Heintz, N. (2009). The Nuclear DNA Base 5-Hydroxymethylcytosine Is Present in Purkinje Neurons and the Brain. Science *324*, 929–930.

Ko, M., Huang, Y., Jankowska, A.M., Pape, U.J., Tahiliani, M., Bandukwala, H.S., An, J., Lamperti, E.D., Koh, K.P., Ganetzky, R., Liu, X.S., Aravind, L., Agarwal, S., Maciejewski, J.P.

and Rao, A. (2010). Impaired hydroxylation of 5-methylcytosine in myeloid cancers with mutant TET2. Nature *468*, 839–843.

Kweon, S.-M., Zhu, B., Chen, Y., Aravind, L., Xu, S.-Y. and Feldman, D.E. (2017). Erasure of Tet-Oxidized 5-Methylcytosine by a SRAP Nuclease. Cell Reports *21*, 482–494.

Lara-Astiaso, D., Weiner, A., Lorenzo-Vivas, E., Zaretsky, I., Jaitin, D.A., David, E., Keren-Shaul, H., Mildner, A., Winter, D., Jung, S., Friedman, N. and Amit, I. (2014). Immunogenetics. Chromatin state dynamics during blood formation. Science *345*, 943–949.

LeCun, Y., Bengio, Y. and Hinton, G. (2015). Deep Learning. Nature *521*, 436–444.

Leung, D., Du, T., Wagner, U., Xie, W., Lee, A.Y., Goyal, P., Li, Y., Szulwach, K.E., Jin, P., Lorincz, M.C. and Ren, B. (2014). Regulation of DNA methylation turnover at LTR retrotransposons and imprinted loci by the histone methyltransferase Setdb1. Proceedings of the National Academy of Sciences *111*, 6690–6695.

Li, J., Ching, T., Huang, S. and Garmire, L.X. (2015). Using epigenomics data to predict gene expression in lung cancer. BMC Bioinformatics *16*, 10.

Li, R., Zhou, Y., Cao, Z., Liu, L., Wang, J., Chen, Z., Xing, W., Chen, S., Bai, J., Yuan, W., Cheng, T., Xu, M., Yang, F.-C. and Zhao, Z. (2018). TET2 Loss Dysregulates the Behavior of Bone Marrow Mesenchymal Stromal Cells and Accelerates Tet2-Driven Myeloid Malignancy Progression. Stem Cell Reports *10*, 166–179.

Li, W., Yin, Y., Quan, X. and Zhang, H. (2019). Gene Expression Value Prediction Based on XGBoost Algorithm. Frontiers in Genetics *10*, 1077.

Lin, I-Hsuan., Chen, Y.-F. and Hsu, M.-T. (2017). Correlated 5-Hydroxymethylcytosine (5hmC) and Gene Expression Profiles Underpin Gene and Organ-Specific Epigenetic Regulation in Adult Mouse Brain and Liver. PLoS ONE *12*, e0170779.

Lio, C.-W., Zhang, J., González-Avalos, E., Hogan, P.G., Chang, X. and Rao, A. (2016). Tet2 and Tet3 cooperate with B-lineage transcription factors to regulate DNA modification and chromatin accessibility. eLife *5*, e18290.

Lio, C.-W.J., Shukla, V., Samaniego-Castruita, D., González-Avalos, E., Chakraborty, A., Yue, X., Schatz, D.G., Ay, F. and Rao, A. (2019). TET enzymes augment activation-induced deaminase (AID) expression via 5-hydroxymethylcytosine modifications at the Aicda superenhancer. Science Immunology *4*, eaau7523.

Lio, C.-W.J., Yue, X., López-Moyado, I.F., Tahiliani, M., Aravind, L. and Rao, A. (2020). TET methylcytosine oxidases: new insights from a decade of research. Journal of Biosciences *45*, 21.

Lister, R., Mukamel, E.A., Nery, J.R., Urich, M., Puddifoot, C.A., Johnson, N.D., Lucero, J., Huang, Y., Dwork, A.J., Schultz, M.D., Yu, M., Tonti-Filippini, J., Heyn, H., Hu, S., Wu, J.C., Rao, A., Esteller, M., He, C., Haghighi, F.G. and Sejnowski, T.J. (2013). Global Epigenomic Reconfiguration During Mammalian Brain Development. Science *341*, 1237905.

Lloret-Llinares, M., Karadoulama, E., Chen, Y., Wojenski, L.A., Villafano, G.J., Bornholdt, J., Andersson, R., Core, L., Sandelin, A. and Jensen, T.H. (2018). The RNA exosome contributes to gene expression regulation during stem cell differentiation. Nucleic Acids Research *46*, 11502–11513.

López-Moyado, I.F., Tsagaratou, A., Yuita, H., Seo, H., Delatte, B., Heinz, S., Benner, C. and Rao, A. (2019). Paradoxical association of TET loss of function with genome-wide DNA hypomethylation. Proceedings of the National Academy of Sciences *116*, 16933–16942.

Mathur, R., Alver, B.H., San Roman, A.K., Wilson, B.G., Wang, X., Agoston, A.T., Park, P.J., Shivdasani, R.A. and Roberts, C.W.M. (2016). ARID1A loss impairs enhancer-mediated gene regulation and drives colon cancer in mice. Nature Genetics *49*, 296–302.

Mellén, M., Ayata, P. and Heintz, N. (2017). 5-hydroxymethylcytosine accumulation in postmitotic neurons results in functional demethylation of expressed genes. Proceedings of the National Academy of Sciences *114*, e7812–e7821.

Montagner, S., Leoni, C., Emming, S., Della Chiara, G., Balestrieri, C., Barozzi, I., Piccolo, V., Togher, S., Ko, M., Rao, A., Natoli, G. and Monticelli, S. (2016). TET2 Regulates Mast Cell Differentiation and Proliferation through Catalytic and Non-catalytic Activities. Cell Reports *15*, 1566–1579.

Moore, L.D., Le, T. and Fan, G. (2012). DNA Methylation and Its Basic Function. Neuropsychopharmacology *38*, 23–38.

Mortazavi, A., Williams, B.A., McCue, K., Schaeffer, L. and Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nature Methods *5*, 621–628.

Natarajan, A., Yardimci, G.G., Sheffield, N.C., Crawford, G.E. and Ohler, U. (2012). Predicting cell-type-specific gene expression from regions of open chromatin. Genome Research *22*, 1711–1722.

Neri, F., Incarnato, D., Krepelova, A., Rapelli, S., Pagnani, A., Zecchina, R., Parlato, C. and Oliviero, S. (2013). Genome-wide analysis identifies a functional association of Tet1 and Polycomb repressive complex 2 in mouse embryonic stem cells. Genome Biology *14*, 91.

Nestor, C.E., Ottaviano, R., Reddington, J., Sproul, D., Reinhardt, D., Dunican, D., Katz, E., Dixon, J.M., Harrison, D.J. and Meehan, R.R. (2012). Tissue type is a major modifier of the 5-hydroxymethylcytosine content of human genes. Genome Research *22*, 467–477.

Pan, F., Wingo, T.S., Zhao, Z., Gao, R., Makishima, H., Qu, G., Lin, L., Yu, M., Ortega, J.R., Wang, J., Nazha, A., Chen, L., Yao, B., Liu, C., Chen, S., Weeks, O., Ni, H., Phillips, B.L., Huang, S. and Wang, J. (2017). Tet2 loss leads to hypermutagenicity in haematopoietic stem/progenitor cells. Nature Communications *8*, 15102.

Pan, W., Zhu, S., Qu, K., Meeth, K., Cheng, J., He, K., Ma, H., Liao, Y., Wen, X., Roden, C., Tobiasova, Z., Wei, Z., Zhao, J., Liu, J., Zheng, J., Guo, B., Khan, S.A., Bosenberg, M., Flavell,

R.A. and Lu, J. (2017). The DNA Methylcytosine Dioxygenase Tet2 Sustains Immunosuppressive Function of Tumor-Infiltrating Myeloid Cells to Promote Melanoma Progression. Immunity *47*, 284–297, e5.

Pastor, W.A., Aravind, L. and Rao, A. (2013). TETonic shift: biological roles of TET proteins in DNA demethylation and transcription. Nature Reviews Molecular Cell Biology *14*, 341–356.

Pastor, W.A., Huang, Y., Henderson, H.R., Agarwal, S. Rao, A. (2012). The GLIB technique for genome-wide mapping of 5-hydroxymethylcytosine. Nature Protocols *7*, 1909–1917.

Pastor, W.A., Pape, U.J., Huang, Y., Henderson, H.R., Lister, R., Ko, M., McLoughlin, E.M., Brudno, Y., Mahapatra, S., Kapranov, P., Tahiliani, M., Daley, G.Q., Liu, X.S., Ecker, J.R., Milos, P.M., Agarwal, S. and Rao, A. (2011). Genome-wide mapping of 5-hydroxymethylcytosine in embryonic stem cells. Nature *473*, 394–397.

Raab, J.R., Tulasi, D.Y., Wager, K.E., Morowitz, J.M., Magness, S.T. and Gracz, A.D. (2019). Quantitative classification of chromatin dynamics reveals regulators of intestinal stem cell differentiation. Development *147*, dev181966.

Sardina, J.L., Collombet, S., Tian, T.V., Gómez, A., Di Stefano, B., Berenguer, C., Brumbaugh, J., Stadhouders, R., Segura-Morales, C., Gut, M., Gut, I.G., Heath, S., Aranda, S., Di Croce, L., Hochedlinger, K., Thieffry, D. and Graf, T. (2018). Transcription Factors Drive Tet2-Mediated Enhancer Demethylation to Reprogram Cell Fate. Cell Stem Cell *23*, 727–741.e9.

Sekhon, A., Singh, R. and Qi, Y. (2018). DeepDiff: DEEP-learning for predicting DIFFerential gene expression from histone modifications. Bioinformatics *34*, i891–i900.

Shrikumar, A., Greenside, P. and Kundaje, A. (2019). Learning Important Features Through Propagating Activation Differences. arXiv:1704.02685.

Singh, R., Lanchantin, J., Robins, G. and Qi, Y. (2016). DeepChrome: deep-learning for predicting gene expression from histone modifications. Bioinformatics *32*, i639–i648.

Singh, R., Lanchantin, J., Sekhon, A. and Qi, Y. (2017). Attend and Predict: Understanding Gene Regulation by Selective Attention on Chromatin. arXiv:1708.00339

Song, C.-X., Diao, J., Brunger, A.T. and Quake, S.R. (2016). Simultaneous single-molecule epigenetic imaging of DNA methylation and hydroxymethylation. Proceedings of the National Academy of Sciences *113*, 4338–4343.

Song, C.-X., Szulwach, K.E., Fu, Y., Dai, Q., Yi, C., Li, X., Li, Y., Chen, C.-H., Zhang, W., Jian, X., Wang, J., Zhang, L., Looney, T.J., Zhang, B., Godley, L.A., Hicks, L.M., Lahn, B.T., Jin, P. He, C. (2011). Selective chemical labeling reveals the genome-wide distribution of 5-hydroxymethylcytosine. Nature Biotechnology *29*, 68–72.

Song, C.-X., Yin, S., Ma, L., Wheeler, A., Chen, Y., Zhang, Y., Liu, B., Xiong, J., Zhang, W., Hu, J., Zhou, Z., Dong, B., Tian, Z., Jeffrey, S.S., Chua, M.-S., So, S., Li, W., Wei, Y., Diao, J. and

Xie, D. (2017). 5-Hydroxymethylcytosine signatures in cell-free DNA provide information about tumor types and stages. Cell Research *27*, 1231–1242.

Stoyanova, E., Riad, M., Rao, A. and Heintz, N. (2021). 5-Hydroxymethylcytosine-mediated active demethylation is required for mammalian neuronal differentiation and function. eLife *10*, e66973.

Szulwach, K.E., Li, X., Li, Y., Song, C.-X., Han, J.W., Kim, S., Namburi, S., Hermetz, K., Kim, J.J., Rudd, M.K., Yoon, Y.-S., Ren, B., He, C. and Jin, P. (2011). Integrating 5-Hydroxymethylcytosine into the Epigenomic Landscape of Human Embryonic Stem Cells. PLoS Genetics *7*, e1002154.

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

Taiwo, O., Wilson, G.A., Morris, T., Seisenberger, S., Reik, W., Pearce, D., Beck, S. and Butcher, L.M. (2012). Methylome analysis using MeDIP-seq with low DNA concentrations. Nature Protocols *7*, 617–636.

Tekpli, X., Urbanucci, A., Hashim, A., Vågbø, C.B., Lyle, R., Kringen, M.K., Staff, A.C., Dybedal, I., Mills, I.G., Klungland, A. and Staerk, J. (2016). Changes of 5-hydroxymethylcytosine distribution during myeloid and lymphoid differentiation of CD34+ cells. Epigenetics & Chromatin *9*, 21.

Tran, K.A., Dillingham, C.M. and Sridharan, R. (2019). Coordinated removal of repressive epigenetic modifications during induced reversal of cell identity. The EMBO Journal *38*, e101681.

Tsagaratou, A., Aijo, T., Lio, C.-W. J., Yue, X., Huang, Y., Jacobsen, S.E., Lahdesmaki, H. and Rao, A. (2014). Dissecting the dynamic changes of 5-hydroxymethylcytosine in T-cell development and differentiation. Proceedings of the National Academy of Sciences *111*, E3306–E3315.

Tsagaratou, A., González-Avalos, E., Rautio, S., Scott-Browne, J.P., Togher, S., Pastor, W.A., Rothenberg, E.V., Chavez, L., Lähdesmäki, H. and Rao, A. (2017). TET proteins regulate the lineage specification and TCR-mediated expansion of iNKT cells. Nature Immunology *18*, 45–53.

Uribe-Lewis, S., Carroll, T., Menon, S., Nicholson, A., Manasterski, P.J., Winton, D.J., Buczacki, S.J.A. and Murrell, A. (2020). 5-hydroxymethylcytosine and gene activity in mouse intestinal differentiation. Scientific Reports *10*, 546.

van Os, H.J.A., Ramos, L.A., Hilbert, A., van Leeuwen, M., van Walderveen, M.A.A., Kruyt, N.D., Dippel, D.W.J., Steyerberg, E.W., van der Schaaf, I.C., Lingsma, H.F., Schonewille, W.J., Majoie, C.B.L.M., Olabarriaga, S.D., Zwinderman, K.H., Venema, E., Marquering, H.A. and Wermer, M.J.H. (2018). Predicting Outcome of Endovascular Treatment for Acute Ischemic Stroke: Potential Value of Machine Learning Algorithms. Frontiers in Neurology *9*, 784.

Vanheer, L., Song, J., De Geest, N., Janiszewski, A., Talon, I., Provenzano, C., Oh, T., Chappell, J. and Pasque, V. (2019). Tox4 modulates cell fate reprogramming. Journal of Cell Science *132*, jcs232223.

Wei, G., Abraham, Brian J., Yagi, R., Jothi, R., Cui, K., Sharma, S., Narlikar, L., Northrup, Daniel L., Tang, Q., Paul, William E., Zhu, J. and Zhao, K. (2011). Genome-wide Analyses of Transcription Factor GATA3-Mediated Gene Regulation in Distinct T Cell Types. Immunity *35*, 299–311.

Welte, T., Tuck, A.C., Papasaikas, P., Carl, S.H., Flemr, M., Knuckles, P., Rankova, A., Bühler, M. and Großhans, H. (2019). The RNA hairpin binder TRIM71 modulates alternative splicing by repressing MBNL1. Genes & Development *33*, 1221–1235.

Wu, H., D'Alessio, A.C., Ito, S., Xia, K., Wang, Z., Cui, K., Zhao, K., Eve Sun, Y. and Zhang, Y. (2011). Dual functions of Tet1 in transcriptional regulation in mouse embryonic stem cells. Nature 473, 389–393.

Yue, X., Samaniego-Castruita, D., González-Avalos, E., Li, X., Barwick, B.G. and Rao, A. (2021). Whole-genome analysis of TET dioxygenase function in regulatory T cells. EMBO reports *22*, e52716.

Zeng, W., Wang, Y. and Jiang, R. (2019). Integrating distal and proximal information to predict gene expression via a densely connected convolutional neural network. Bioinformatics *36*, 496–503.

Zhang, Jingli A., Mortazavi, A., Williams, Brian A., Wold, Barbara J. and Rothenberg, Ellen V. (2012). Dynamic Transformations of Genome-wide Epigenetic Marking and Transcriptional Control Establish T Cell Identity. Cell *149*, 467–482.

Zhao, Z., Chen, L., Dawlaty, Meelad M., Pan, F., Weeks, O., Zhou, Y., Cao, Z., Shi, H., Wang, J., Lin, L., Chen, S., Yuan, W., Qin, Z., Ni, H., Nimer, Stephen D., Yang, F.-C., Jaenisch, R., Jin, P. and Xu, M. (2015). Combined Loss of Tet1 and Tet2 Promotes B Cell, but Not Myeloid Malignancies, in Mice. Cell Reports *13*, 1692–1704.

Zhu, G., Li, Y., Zhu, F., Wang, T., Jin, W., Mu, W., Lin, W., Tan, W., Li, W., Street, R. Craig, Peng, S., Zhang, J., Feng, Y., Warren, Stephen T., Sun, Q., Jin, P. and Chen, D. (2014). Coordination of Engineered Factors with TET1/2 Promotes Early-Stage Epigenetic Modification during Somatic Cell Reprogramming. Stem Cell Reports *2*, 253–261.

Zrimec, J., Börlin, C.S., Buric, F., Muhammad, A.S., Chen, R., Siewers, V., Verendel, V., Nielsen, J., Töpel, M. and Zelezniak, A., 2020. Deep learning suggests that gene expression is encoded in all parts of a co-evolving interacting gene regulatory structure. Nature communications *11*, 1–16.

CHAPTER 3: Integrating 3D genome structure with 5hmC enrichment to predict gene expression and long-distance regulatory regions.

3.1 Abstract

5-hydroxymethylcytosine (5hmC) is an epigenetic mark generated from 5-methylcytosine (5mC) by TET enzymes, which deposit the highest levels of 5hmC at the most highly transcribed genes and the most active enhancers. Here, we use data on 5hmC distribution genome-wide, in conjunction with genome-wide Hi-C chromosome conformation capture data, to predict gene expression and to identify novel enhancers important for cell-specific gene regulation. We accurately detected previously validated enhancers with short range as well as other important enhancers with long-range interactions to the *Aicda* in activated and resting B cells. We also identified novel enhancer regions for the *Il4* locus in Th2 cells, and the *Cd8ab1* locus in CD4 and CD8 T cells. In the *Aicda* locus, we found previously unknown, putative distal regulatory regions whose time-course of 5hmC enrichment was reminiscent to that of known Tet-dependent enhancers TetE2 and TetE1, offering a system for prioritization of enhancers for further experimental validation. Our work demonstrates that the integration of 5hmC with 3D chromatin structure can be used to predict gene expression and to identify novel regulatory regions.

3.2 Introduction

We showed in chapter 1 that TET enzymes convert 5-methylcytosine (5mC) to 5-hydroxymethylcytosine (5hmC) at active enhancers, which bind transcription factors that regulate the expression of the genes controlled by those enhancers. These transcription factors recruit a variety of chromatin-based proteins that regulate expression of the corresponding genes, including chromatin remodelling complexes that change the accessibility of chromatin in the vicinity, histone acetyltransferases such as p300 and CBP that deposit H3K27Ac and other acetylated histone marks, and TET proteins that convert 5mC (usually present in the CpG sequence context) to its oxidized forms including 5hmC. Whereas DNA methylation is a stable and generally heritable mark that is quickly restored at most sites in the genome after DNA replication, enhancers that are newly activated during cellular activation or differentiation show a slow, progressive deposition of 5hmC during the activation/differentiation process. 5hmC can be a highly stable modification in the eventual, non-proliferating differentiated cells (Bachman et al. 2014), and 5hmC genomic enrichment can be assessed by multiple assay techniques, such as GLIB-seq (Pastor et al. 2012), CMS-IP (Huang et al. 2012), hMe-Seal (Song et al. 2017), nano-hmC-Seal (Han et al. 2016; Gabrieli et al. 2018), and hMEDIP (Song et al. 2011; Taiwo et al. 2012).

Numerous studies have used epigenetic marks as tools to link regulatory regions to their target gene. Most of these studies have typically focused on signals such as histone marks (H3K27Ac, H3K9me3, etc.) or accessible genomic regions (Assay for Transposase-Accessible Chromatin, Buenrostro et al. 2015) and analyzes of these in conjunction with chromosome conformation capture methods such as Hi-C or its variants. 5hmC is an epigenetic modification embedded in the DNA molecule, that has been the focus of multiple studies confirming its

association with lineage-specific enhancers (as seen in chapter 1), and a comprehensive analysis of joint profiles of chromatin organization and 5hmC DNA mark is missing in the literature.

Analysis of long-range interactions, and their dynamic association with epigenetic modifications during cell activation and differentiation, has been the focus of many genomics studies. In many cases these studies have led to the discovery and validation of novel enhancers. Examples include Enformer (Avsec et al. 2021), which makes gene expression predictions from DNA sequences by integrating information from interactions in the genome that are up to 100 kb away; Activity by Contact model (Fulco & Nasser et al. 2019), which constructs enhancer-gene connection maps to predict enhancers and their target genes by the use of chromatin accessibility or acetylation and chromatin conformation; TargetFinder (Whalen et al. 2016), which models the interaction status of predefined pairs of enhancers and promoters by integration of multiple genomic features; JEME (Cao et al. 2017), which considers the joint effect of multiple enhancers on a given TSS that can be up to 1 Mb away; GraphReg (Karbalayghareh et al. 2021), which utilizes either epigenomic data only (Epi-GraphReg), or integrates DNA sequences (Seq-GraphReg) with chromatin interaction data to predict gene expression; and GC-MERGE (Bigness et al. 2020), which integrates multiple epigenetic modifications associated with repression and activation with the 3D genome organization to predict gene expression.

A key component of many of these models is the use of machine learning to process the massive datasets, capture dependencies among the data, and train a model to predict the enhancer-promoter associations. This is because machine learning algorithms are designed to automatically detect patterns in data (Libbrecht et al. 2015). Thus, machine learning algorithms are ideal for data-driven sciences such as genomics (Eraslan et al. 2019), where current frontiers lie in the creative integration of deep neural networks (DNNs). DNNs are models composed of successive

elementary operations, computing increasingly complex features from the results of preceding operations as input (Greener & Kandathil et al. 2021). An example of creative integration was the adaptation of convolutional operations into graph-structured data to produce 'graph convolutional networks' (GCNs; Kipf et al. 2017) that can produce node representations that encode both local graph structure and features of nodes, known as vector embedding (or simply "embeddings") that can be fed to downstream machine learning systems. Although efficient, these embeddings were limited to fixed graphs, meaning that the goal was to generate representations of the nodes themselves. A more recent framework, known as GraphSAGE (Hamilton et al. 2018) proposed a different strategy that, instead of learning the node representation themselves, learns a function to aggregate feature information from a node's local neighborhood, that can efficiently generate the embeddings. This trained function can then be used in previously unseen data.

Many of the state-of-the-art enhancer-gene linkage or gene expression predictors use a vast amount of data (Bigness et al. 2020; Avsec et al. 2021; Karbalayghareh et al. 2021; Whalen et al. 2016). Given the robust predictions of gene expression that we were able to make using only 5hmC signal as a 1D epigenetic mark and rather simple neural network structures (discussed in Chapter 2), and considering the observed 5hmC enrichment in cell-specific distal enhancers, we were interested in integrating 5hmC with 3D chromatin structure for the task of predicting both gene expression and putatively functional enhancer regions for each gene. We used the graphical convolutional network structure developed by Bigness and colleagues (Bigness et al. 2020). This structure makes use of the GraphSAGE framework (Hamilton et al. 2018), which allowed us to train an embedding-generator function in a cell-type, and then to used this function in a previously unseen cell type. We hypothesize that, as long as the graphs for each of our samples, as well as the

node attributes (such as the 5hmC enrichment and Input signal), are generated the same way, the trained function may retain predictive value across different cell-types.

Using our previously processed 5hmC, input and gene expression datasets (from chapter 2), and integrating publicly available chromatin contact maps for specific cell types (**Table S3.1**), we trained our graphical 5hmC convolutional networks ("GhmCNs") for the task of predicting gene expression. We obtained models with high predictive ability as calculated by our unbiased metrics. We demonstrated the power of our approach in specific cell-types by showing that the top interactions defining the expression state of key genes containing 5hmC-rich enhancers have been validated in the literature. In the course of this analysis, we decoded our trained GhmCN models with GNNExplainer, thereby discovering new potential regulatory regions that bore several hallmarks of *bona fide* enhancers: they were enriched for H3K27ac, were highly accessible and contained multiple transcription factor binding sites (TFBS), identified both from sequence only (Castro-Mondragon et al. 2021) and from physical binding measured by ChIP-seq data for transcription factors known to be relevant for gene expression.

3.3 Results

Cell-specific predictions. As input data for our models we used our previously processed RNA-seq, 5hmC and Input enrichment datasets, and in addition integrated multiple publicly available Hi-C contact maps at 10 kb resolution. Details of the input generation can be found in the methods section; briefly, for each sample we build a graph based on the strongest Hi-C contacts per window, where the nodes are the 10 kb windows, and the edges are drawn between each window and its top 10 interactors. For each node, we obtained 5hmC and Input signal; if a node overlapped a gene's TSS, that gene's expression label was assigned to the node. We trained all of our graphical 5hmC convolutional network (GhmCN) models based on the extensive hyperparameter tuning performed by Bigness and colleagues (Bigness et al. 2020). Details of the parameters can also be found in the Methods section. For each cell type, we collected and calculated the area under the curve (AUC) score for the gene expression prediction task, based on the test set, and plotted the respective true positive versus the false positive rates in **Fig. 3.1A**. Across the six different cell types/conditions, all the models we generated displayed a powerful ability to discriminate between positive and negative cases, with activated B cells showing the highest and the DP thymocyte model showing the lowest AUC scores of 0.8593 and 0.8046, respectively.

To test the relevance of long-range interactions, as well as to establish a baseline of our predictions, we repeated our cell-specific models by using only the nearest bins (1D genomic distance) (for a total of 10 interactions per bin, 5 upstream and 5 downstream each) as interaction partners for each bin and regenerated our models (**Fig S3.1**). We observed that all of our models suffered loss of both AUC and AUPR scores when using only the 10 nearest bins (+/-5) to the

promoter region, pointing to the importance of long-range interactions in improving gene expresion predictions, presumably by capturing key distal regulatory regions that act as enhancers.

Cross-cell type comparisons. One of the properties of graphical convolutional networks is that they are not tied to a specific graph structure. In our study, the graph structure is composed of the Hi-C contacts (observed interactions between genomic regions). This plasticity means that the weights of a trained GhmCN model can be used to process a different cell-type's input features and to make predictions. Given this property, we were interested in the cross-cell prediction ability of each of our models to obtain evidence of how generalizable our cell-specific models are. We took the weights (training parameters) from the embedding-generating function of a model trained in a given cell-type, predicted the gene expression of a different cell-type using its input features, and calculated the predictive ability of the pre-trained model on this cross-cell gene expression prediction task. We repeated this process for each of our 6 models. **Fig. 3.1B** shows the cross-cell-type AUC scores, ranging from 0.81 when predicting gene expression in Activated B cells by using a model trained on resting B cells, to 0.54, when predicting gene expression in resting B cells using a model trained on data from Naïve CD4 T cells (detailed numbers can be found in **Fig S3.2A)**. Overall, we saw that the closer a cell-type is to the model being used to process its features, the higher the predictive ability of that combination of model and cell-type. We corroborated this observation by observing the grouping of each of the 6 cell-types' expression profiles through principal component analysis (**Fig S3.2B)**.

Mixed-Hi-C model. Given our observations that the models trained in one cell-type and tested in a different cell-type depend on the similarity between the cell types, we asked if using an aggregate set of Hi-C interactions such that the graph structure is identical across samples, could be a viable option. To do this, we generated an "averaged" 3D contact map, based on the known

similarity of Hi-C contacts patterns across cell types, largely determined by linear genomic distance (Sanborn et al. 2015; Yardımcı et al. 2019). A similar approach of using an aggregate Hi-C signal has been employed by the ABC score (Fulco et al. 2019). Our motivation was that the use of an aggregate Hi-C map would benefit the analysis of cell types where maps of 3D contacts are not available.

To this end, we downsampled all of our Hi-C datasets to 183M randomly selected valid interactions, using only samples with more that 183M valid interactions (**Table S3.1**, methods; DP and Th2 cells were excluded due to low coverage) and obtained the normalized interactions thereafter using the iterative correction (ICE) technique (Imakaev et al. 2012). Similar to using the complete datasets, we took the top 10 interactions for each bin to generate the interaction graph. We re-processed each of the cell-types using this new graph structure based on the averaged Hi-C values and obtained the AUC scores shown in **Fig. 3.1C**. Except naïve CD8 T cells, all cell-types tested have performed better with cell-specific contact maps rather than the averaged contact map, suggesting the importance of cell-specific chromatin organization in gene regulation. In a few cases, the AUC score for averaged maps was not much different from that of the cell-specific graph structure; this was particularly true for DP and Th2 cell-types, which had the lowest sequencing depth, and, hence low power to delineate specific enhancer-promoter contacts in the cell-specific model. The cell-types that showed a noticeable drop in their AUC score when the averaged Hi-C data is used (from 0.8593 to 0.7954 in activated B cells) were the ones with deeply sequenced Hi-C maps. Overall, our results suggest that, while it is ideal to use cell-specific and sufficiently sequenced Hi-C contact maps, the averaged Hi-C data and the graph structure inferred from it can be used in conjunction with cell-specific 5hmC data to reasonably predict gene expression of cell-types without available Hi-C data. We then repeated the averaged-Hi-C experiment withholding

the Hi-C interactions of the datasets with a matching 5hmC enrichment profile, and then testing this averaged map with said sample's 5hmC profiling. Thus, we further tested the *in-silico* generalizability of our averaged maps (*e.g.* exclude Hi-C from B cells and then test this B-cell-withheld average-Hi-C with the B cell's 5hmC dataset). We found small differences between the average map containing all of the available Hi-C datasets and those with the sample of interest being withheld in both AUC and AUPR scores (**Table 3.1**).

Explaining the predictions of our GCN model. In our model, the Hi-C datasets are used only to generate the graph structure (define the nodes and what nodes are interacting), and are not used downstream in generating the models, or making predictions based on the models. The predictions from the graphical convolutional network models are based on the 5hmC IP and Input signal's embeddings. The graphical convolutional network structure does not have a module (inner function) to obtain the relative importance of the produced embeddings per node (defined as the 10-Kb windows, the Hi-C resolution used in this study), or the effect (additive or not) that each embedding from a node's connections (the neighborhood) has in the prediction itself. To overcome these limitations, we used GNNExplainer (Ying et al. 2019), a tool that assigns a relative significance score to a node and its interacting neighbors in the prediction result for said node. GNNExplainer takes the predicted label (High or Low), the trained GhmCN model's weights, and the graph structure, and converts each interaction between any two nodes to a significance score. Here we focused on the interactions between the node of interest (i.e., a specific gene) and its neighboors, and did not focus on the interactions among the neighbors themselves. Below we analyze our models with GNNExplainer to prioritize interactions of the nodes containing gene transcription start sites (TSS). We found that the top ranked nodes for each gene showed a high probability of containing regulatory elements with biological significance (see below).

Case study A: New and reported regions for *Aicda* regulation in B cell activation. The nodes conserved after filtering top 10 interactions with *Aicda* promoter in resting and activated B cells are listed in **Table S3.2**. As discussed in Chapter 1, AID, encoded by the *Aicda* gene, is crucial for class switch recombination (CSR) in B cells activated with LPS and IL-4. We also reported two Tet-dependent enhancers located ~10-kb (TetE1) and ~26-kb (TetE2) 5' of the *Aicda* TSS, which showed a progressive increase in 5hmC signal with time after stimulation. We found that in both resting and activated B cells, these two experimentally validated regions were among the top 10 candidates reported by GNNExplainer, suggesting our model's ability to capture putative functional enhancers. Among the other top-ranked interactions in activated B cells were the 10-kb window harbouring the *Apobec* TSS as well as the region between TetE2 and TetE1; these regions all contain known Aicda regulators discussed in Chapter 1.

Notably, we also observed two long-distance interactions (more than 100-kb away from the *Aicda* TSS) that were not found in resting B cells. These regions were located ~260-kb (6:122290000-122300000) and ~160-kb (6:122390000-122400000) 5' of the *Aicda* TSS (**Fig. 3.2A**, **Table S3.2**, 1st and 2nd row respectively), and have not previously been reported to have regulatory roles in *Aicda* expression. Since 5hmC is enriched in lineage-specific enhancers as seen in **Fig. 1.1G**, we tested the hypothesis that these two new regions may harbor unreported long-distance regulators of *Aicda* expression by exploring the dynamics of 5hmC enrichment within these 10-kb windows (**Fig. 3.2B-D**). Using our previously published 5hmC mapping data (by CMS-IP) obtained from WT and double Tet2/3-deficient B cells, resting or 24, 48 and 72 hours after activiation with LPS and IL-4, as discussed in Chapter 1, we observed that a region inside each node significantly gained (p-value <0.1) 5hmC signal after 72 hours of stimulation (chr6:122,293,509-122,294,342 and chr6:122,393,397-122,393,996 respectively, **Fig. 3.2E**). This

pattern is reminiscent of the 5hmC gain that is observed in the known Tet-dependent *Aicda* regulators TetE2 and TetE1 (**Fig. 3.2D**). Using Remap2022, a manually curated, high-quality catalog of regulatory regions (Hammal et al. 2021), and UniBind, collection of high-confidence direct TF-DNA interactions (Puig et al. 2021), we found additional evidence of relevant DNA binding proteins such as SMARCA4, CHD4, NIPBL, KMT2A, HDAC2 and P300, all associated with chromatin remodelling, that were present in the TetE1 and TetE2 known regulatory regions as well as in the two novel regions we report here.

Taken together, we found that the top ranked interacting regions identified by GNNExplainer captures the validated *Aicda* enhancers TetE2 and TetE1 discovered in Chapter 1. In addition, our model predicts two novel regions that are also likely to be *Aicda* enhancers, since they share the stimulation-responsive 5hmC pattern with TetE2 and TetE1 as well as the binding sites of similar chromatin remodelling proteins. Experimental validation of these new regions as *Aicda* enhancers in B cells both in culture and *in vivo* are needed to fully understand their causal role in regulation of *Aicda* during B cell activation.

Case study B: Novel and reported Th2 interactions. Type 2 helper T (Th) cells (Th2 cells) are generated by polarization of naïve CD4 T cells in the presence of interleukin (IL)-4, a potent inducer that directs differentiation of naive CD4+ T cells into CD4+ Th2 effector cells (Chen et al. 2004). Many studies have focused on *Il4* gene regulatory networks: key regions within the the last exons of *Rad50* (Lee et al. 2004; Fields et al. 2004), a gene located 5' of *Il4*; between the TTS of *Il4* and *Kif3a* (Harada et al. 2012), and in the intergenic space between *Il4* and *Il13* (Loots et al. 2000; Baguet et al. 2004), have been reported as *Il4* enhancers (Ansel et al. 2006).

We found that, among the top 10 interacting regions associated to the *Il4* TSS, 5 covered reported regulatory regions: CNS2, also known as hypersensitive site V (HS V) (Agarwal & Rao

1998; Vijayanand et al. 2012; Harada et al. 2012), CNS1 (Loots et al. 2000; Harada et al. 2012; Baguet et al. 2004; Guo et al 2002) located between *Il4* and *Il13*, CGRE 1.6 kbp upstream from the IL-13 gene (Harada et al. 2012; Yamashita et al. 2002), RHS6/7 and RHS5 located in the last exon of the *Rad50* gene (Lee G.R. et al. 2004; Lee, D.U. et al. 2004; Fields et al. 2004) (between the coordinates chr11:53600000-53670000 in **Fig. 3.3A**). Of the other 5 interacting regions, two (here termed Kif-A and Kif-B for convenience) appeared particularly relevant based on their proximity to the *Il4* gene and that none of the other T cell samples (DP, CD4 and CD8 naive T cells) had these two regions in their top interactions (**Table S3.4**, see regions demarked by the black box). Inside the Kif-A and Kif-B regions we observed clear 5hmC signal peaks and strong presence of transcription factor binding sites in or near the 5hmC signal peaks (chr11:53580000-53600000, **Fig. 3.3B**). The predominant TFBSs found by Remap2022 (Hammal et al. 2021) and UniBind (Puig et al. 2021), and analysis of public ChIP-seq datasets in our regions of interest, are associated with binding of Foxo1, NFAT1, 2 and 4, CREB, STAT, MYC, Fos, JunD/B, BATF, MAFF, IRF4 and more bZip-related TFs, all important for IL-4 production (Sahoo et al. 2015). Although Malik and colleagues (Malik et al. 2017) showed that inhibition of Foxo1 had no effect on *Il4* expression, several reports have shown evidence of the crucial role of BATF and other bZIP factors both in Th2 cell generation and *Il4* expression both in mouse and human (Kuwahara et al. 2016; Bao et al. 2016; Sahoo et al. 2015s).

To explore the potential roles of the Kif-A and Kif-B regions in regulating *Il4* expression, we downloaded accessibility data, chromatin immunoprecipitation (ChIP-seq) data of multiple epigenetic marks associated to regulatory regions (Histone 3 K27ac, K27me3, K4me1, K4me3 & K79me2), as well as ChIP-seq data for the transcription factors BACH, BAFF, p300 and IRF4 for Th2 cells (**Table S3.3**). We found that subregions inside each of these two nodes (**Fig 3.3B**, pink

highlights) have clear 5hmC peaks with strong co-binding of key transcription factors such as BATF and IRF4, also that this binding of IRF4 is lost in BATF KO and BATF/BATF3 DKO Th2 cells. Moreover the Kif-A and Kif-B regions were accessible and displayed H3K27ac enrichment in Th2 cells, and contained one perfect match (chr11:53585651-53585753) to the the activating protein 1 (AP-1) binding consensus sequence (TGASTCA), and the other (chr11:53593319-53593416) a very close match to the AP-1–IRF composite elements (AICE2; TacCnnnnTGASTCA), known to enable IRF4/8-dependent transcription by cooperative binding with BATF, resulting in expression of genes associated with activation and differentiation for Th17, B, and dendritic cells, and are also used in Th2 cells (Glasmacher et al. 2012l; Yosef et al. 2013). Kuwahara and colleagues (Kuwahara et al. 2016) showed that there is a positive feed-forward (amplification) loop between *Il4* and *Batf* to induce Th2 cell differentiation, where the BATF:IRF4 complex is key for IL-4 expression, and overexpression of IL-4 further augments BATF expression. Both ReMap 2022 and UniBind provided further evidence for BATF and IRF4 binding as well as general bZIP TF binding. Taking these observations, it is possible that the Kif-A and Kif-B regions may be unreported *Il4* enhancers mediated through bZIP TF family members, such as the BATF:IRF4 complex; this hypothesis is currently being investigated.

Case study C: Short- and long-range enhancer predictions for naïve CD4 and CD8 T cells. As a final case, we wanted to examine the *Cd8ab1* gene complex in the T cell lineage. The dynamic and complex pattern of CD8 expression (encoded by the *Cd8b1* gene) has been reported to be regulated by at least six Cd8 enhancers, designated as Enhancer of CD8 (E8)-I to E8VI, and found within the *Cd8ab1* gene complex (Gülich et al. 2019). We observed that in naïve CD8 T cells the top 10 selected interactions not only contained the E8II, E8VI, E8I, and E8V enhancers (**Table S3.5**, **Fig. 3.4A**), but also the nodes containing these regions were among the top candidates

explaining *Cd8b1* gene expression as categorized by GNNExplainer (**Fig. 3.4B**). On the other hand, the top 10 interactions selected in naïve CD4 T cells only contained the enhancer E8I (**Table S3.5**, **Fig. 3.4C**). However, it is worth mentioning that this enhancer E8I contributes to CD8 expression under HDAC inhibitor conditions in naïve CD4 T cells (Gülich et al. 2019). When followed by our GNNExplainer analysis, we observe that the ranking assigned to each of the nodes was associated to the evidence we had for those regions being important for the regulation in naïve CD8 T cells, for example the top three most relevant nodes (position 1st to 3rd in the rank, **Table S3.5**, under Naïve CD8 T "GNNExplainer rank") are associated to the E8 enhancers I, II, and VI, whose interplay has been observed to regulate CD8 expression: E8I-core and E8VI double deletion leads to CD8 expression reduction during activation (Gülich et al. 2019); E8I is key to maintain transcription of CD8a during activation (Ellmeier et al. 2002); E8II deletion disrupted CD8a expression in both DP thymocytes and CD8+ T cells (Hostert et al. 1998).

Overall, our study has brought multiple data modalities altogether to not only use the 5hmC signal as a means of predicting gene expression, as observed in Chapter 2, but also to extend this to further explain, and prioritize putative functional enhancers with respect to their 3D promixty to the promoters rather than their linear genomic distance. With these analyzes we obtained a high ability to label gene expression prediction in High and Low categories and additionally, were able to identify regions with the potential for regulating expression of particular genes. Finally, we show that we can use averaged Hi-C data for the sole purpose of predicting gene expression, although the discovery of novel regulatory regions may be compromised since the averaged Hi-C data lose key cell-type-specific interactions.

3.4 Discussion

5hmC is a DNA modification mark that has been associated with active regions in the genome, such as highly expressed genes and lineage-specific enhancers with high activity, as defined by both H3K27 acetylation and monomethylation marks. Here we have shown that this 5hmC genomic distribution can be used to link enhancers to their target genes by integrating 3D chromatin structure (as obtained by Hi-C-derived genomic interaction matrices) into the task of predicting gene expression. In our GhmCN machine learning models, we used the 3D chromatin structure to connect 5hmC signal levels, within each genomic region of size 10-kb, with their top 10 interacting genomic regions. By doing this, we integrated the spatial structure and the 5hmC signal distribution to predict gene expression and obtained cell-specific models with high predictive power, having an AUC score above 0.81 in all our tested models. By only using the 5 interactions next to each bin when constructing the graphs to integrate the 5hmC signal, we demonstrated the importance of long-range interactions since the models without these interactions performed worse as indicated by our unbiased metrics. When we tried to use cell-type-specific trained models with data from other cell-types, the prediction accuracy dropped proportional to the distance between the cell types used for training and testing. However, when we generated an averaged Hi-C interaction map from subsampled multiple Hi-C datasets (cell types included naïve and activated B cells; DP and $CD4^+$ naïve T cells; $CD8^+$ naïve, effector and exhausted T cells; LSK, Th2 and BMDMs), we showed that these models conserved strong predictive ability with a minimum AUC score of 0.78. This provided us evidence that cell-type-specific 5hmC enrichment signals can be a powerful way to predict gene expression when integrated with averaged 3D chromatin structure data. Since GCN models do not inform what nodes, or edges, are the most important in making the classification for a given node (prediction of binary expression class), we

used GNNExplainer, a tool that assigns relative importance to each edge and node feature in a graph. Our GNNExplainer analysis proved to be a reliable way to interpret which nodes (genomic regions) were the most important among those interacting with a gene's TSS, in terms of making gene expression predictions.

Validating our methods, we found that the top candidates for exemplar genes were consistent with observed roles for those regions that had been reported in the literature, such as TetE1- and TetE2-containing nodes being ranked in the top 5 important interactions in activated B cells. Moreover, our prioritization of the top ten candidate interacting regions allowed us to focus deeper on regions that GNNExplainer deemed important, but that might not yet have been validated as having regulatory roles (for novel putative enhancers). We believe the analysis strategy we followed here can be an important tool to prioritize putative functional enhancers that regulate key genes of each studied cell type. Finally, it is important to keep in mind that, while Hi-C and 5hmC signal enrichment constitute a powerful pair, Hi-C is substantially more expensive and requires intact nuclei, compared to just 5hmC. Even though this is a limitation, our results showing that an averaged Hi-C contact map from an ensemble of cell types provides reasonable predictions when combined with cell-specific 5hmC signals making it possible to generalize our model to a broader set of samples. It would be interesting to eliminate the use of Hi-C to link 5hmC-derived enhancers to their targer genes, however, this would require a study surveying multiple cell-types through multiple differentiation steps to have a dynamic high-resolution map of 5hmC changes pointing to possible putative enhancers.

3.5 Materials and Methods

5hmC Enrichment datasets. All enrichment datasets (and its input) were processed with the same pipeline as follows. We downloaded the raw reads and mapped them to the mm10 genome reference assembly using Bsmap (Xi et al. 2009). Unmapped reads were remapped after using TrimGalore (Krueger 2015) and added to the mapping results after both files were sorted with SAMtools (Li et al. 2009). PCR duplicates were estimated and removed using Picard Toolkit's MarkDuplicates (Broad Institute. Picard Toolkit 2018). Mapping results aligned to ENCODE's blacklisted regions (Amemiya et al. 2019) were removed before further analysis. We generated HOMER's TagDirectories followed by HOMER's makeMultiWig tracks for visualization in the genome browser (Heinz et al. 2010). The 5hmC (and input) signal in the graph's nodes was obtained using GenomicAlignments's summarizeOverlaps function (Lawrence et al. 2013).

ATAC-seq datasets. Paired raw reads were aligned to the Mus musculus genome (mm10) using Bowtie (Langmead et al. 2009). Unmapped reads were trimmed to remove adapter sequences and clipped by one base pair with TrimGalore (Krueger 2015) before being aligned again. Sorted alignments from the first and second alignments were merged together with SAMtools (Li, Handsaker et al. 2009), followed by removal of reads aligned to the mitochondrial genome. Duplicated reads were removed with Picard Toolkit's MarkDuplicates (Broad Institute. Picard Toolkit 2018). Reads aligning to the blacklisted regions (Amemiya et al. 2019) were removed using bedtools intersect (Quinlan et al. 2010). Final mapping results were processed using HOMER's makeTagDirectory program followed by the makeMultiWigHub.pl program (Heinz et al. 2010) to produce normalized bigWig genome browser tracks.

ChIP-seq datasets. All downloaded ChIP-seq datasets were processed similarly to the 5hmC enrichment datasets with the only difference being the use of BWA mem (Li et al. 2009) opposed to Bsmap for the mapping steps, Bsmap is specific to reduced genomes such as bisulphite-treated samples.

Hi-C-seq analysis. All datasets were processed using HiC-Pro (Servant et al. 2015). We downloaded the raw reads and mapped them to the UCSC genome annotation database for the Dec. 2011 (GRCm38/mm10) assembly of the mouse genome. We obtained the appropriate restriction enzyme per sample from their corresponding manuscript's published methods, required for HiC-Pro's configuration file. For samples with either multiple lanes or multiple replicates, we generated a merged sample folder and re-compute the ICE (Imakaev et al. 2012) normalized matrices by running HiC-Pro and the steps "-s merge_persample -s build_contact_maps -s ice_norm". To generate the contacts without long-range interactions, we filter out the interactions beyond 60-kb, thus keeping only the neighborhood of 5 interactions of each bin.

RNA-seq analysis. All expression profiles datasets were processed using STAR (Dobin et al. 2012). We downloaded the raw reads and mapped them to the UCSC genome annotation database for the Dec. 2011 (GRCm38/mm10) assembly of the mouse genome. Counts per gene were obtained using FeatureCounts (Liao et al. 2013). Similar results were obtained when using STAR's count algorithm. For the generation of the output labels, we RPKM-normalized the RNA signal expression and took the median gene expression as the threshold to divide and label genes as "High" and "Low" (above or below threshold, respectively).

Graph Convolution Networks. We followed the same strategy as reported by Bigness and colleagues (Bigness et al. 2020). Briefly, we followed the GraphSAGE framework (Hamilton et al. 2018) formulation as the structure for our GCNs due to its portability and lack of restrictions

to a specific graph. The window size we used to capture the 5hmC signal enrichment, and used in the convolution embeddings was equal to the size of the Hi-C nodes, 10-kb. The model layers consisted of a series of convolutions (convolutions = {2}) interconnected by a ReLU operational unit, followed by a multi-layered perceptron (layers = {3}) preceeded by a dropout chance of (ds = {0.5}) to avoid overfitting. Based on the ICE (Imakaev et al. 2012) normalized Hi-C signal, we filtered a total of top 10 neighboors (k = {10}) per node. With this construction it is possible that some nodes will have more than 10 edges/neighbors because our network is undirected and a gene node can be within the top 10 neighbors of another gene. It is worth mentioning that we tried to use 15 neighboors instead, but we also faced the problem discussed in Bigness et al., 2021. Our NVIDIA GPU ran out of memory to hold this bigger network structure. To assign genes to the nodes, we used as anchor point the TSS coordinates of genes. When a node had more than one TSS (overlapping genes), the mean expression was taken. A gene was marked as either being "On" or "Off" based on the median gene expression of the sample: gene expression above the median indicates that a gene was "On". Training the network made use of mask to consider only the nodes with at least one TSS, therefore, a valid prediction. The Train, Dev and Test fold datasets per sample were split into 70/15/15% form the total. For GNNExplainer we used (e = {1500}) epochs, and explained the queried nodes up to 1-hop away (num_hops = {1}).

3.6 Figures

Figure 3.1. Evaluation of GhmCN models on cell-specific and cross-cell-type gene expression prediction tasks.
(A) Receiver operator characteristic (ROC) curves for each of the six models trained and tested using matching set of Hi-C, 5hmC signal and expression information for each cell-type. The best performers are the B-cell associated samples, which may be due to the high sequencing depth of the Hi-C contact maps compared to the DP and the Th2. **(B)** Calculated AUROC curve for cross-cell analysis of our models. We took a model trained in one cell-type to make gene expression predictions for another cell-type. The more closely related the samples used for training and testing, the higher were the AUC scores obtained. **(C)** ROC curves for each of the six models trained and tested using an averaged set of Hi-C contacts but with cell-specific 5hmC signal. The averaged Hi-C was obtained from subsampling multiple Hi-C datasets from different cell-types to the same sequencing depth and obtaining the ICE normalized matrices from the contacts. Aside from a slight improvement for CD8 T cells, we observed a descrease in prediction accuracies for all other cell types when averaged Hi-C data is used instead of the cell-specific Hi-C data (e.g., a substantial drop from 0.86 to 0.81 for activated B cells).

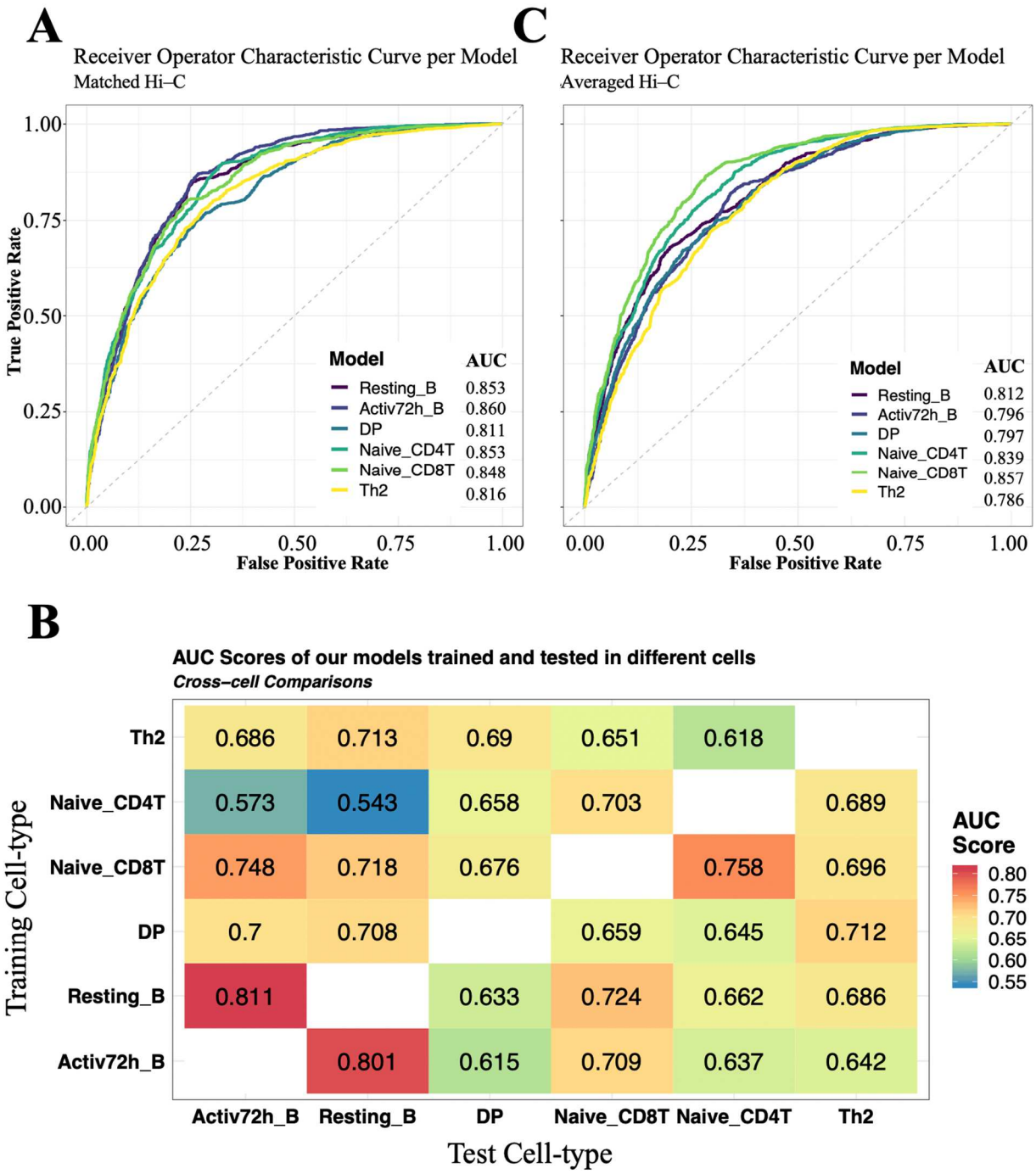

A
Receiver Operator Characteristic Curve per Model
Matched Hi–C
True Positive Rate
False Positive Rate
Model
AUC
Resting_B 0.853
Activ72h_B 0.860
DP 0.811
Naive_CD4T 0.853
Naive_CD8T 0.848
Th2 0.816
C
Receiver Operator Characteristic Curve per Model
Averaged Hi–C
False Positive Rate
Model
AUC
Resting_B 0.812
Activ72h_B 0.796
DP 0.797
Naive_CD4T 0.839
Naive_CD8T 0.857
Th2 0.786
B
AUC Scores of our models trained and tested in different cells
Cross–cell Comparisons
Training Cell-type
Th2 0.686 0.713 0.69 0.651 0.618
Naive_CD4T 0.573 0.543 0.658 0.703 0.689
Naive_CD8T 0.748 0.718 0.676 0.758 0.696
DP 0.7 0.708 0.659 0.645 0.712
Resting_B 0.811 0.633 0.724 0.662 0.686
Activ72h_B 0.801 0.615 0.709 0.637 0.642
Activ72h_B Resting_B DP Naive_CD8T Naive_CD4T Th2
Test Cell-type
AUC Score
0.80
0.75
0.70
0.65
0.60
0.55

Figure 3.2. Novel *Aicda* gene regulatory regions reminiscent of Tet-dependent enhancers. **(A)** Genome browser overview of the top 10 interactions used to predict Aicda gene expression in resting and activated B cells. Red and blue lines below the Placental conservation track indicate the 10-kb windows used to process Hi-C data and generate the graph structure. Arcs above and below the 10-kb windows represent the actual interactions for activated (top, red) and resting (bottom, green) B cells. Multiple resting B cell interactions from Aicda to beyond the TSS of Apobec were omitted; these data can be accessed in **Table S3.2**. Red highlights near the *Aicda* gene show the validated *Aicda* enhancers, TetE1 and TetE2. The two blue highlights at left represent the two novel putative enhancers, which make interactions with the Aicda promoter only in activated B cells. **(B, C)** A zoomed in view of the two long-distance nodes containing the two novel putative enhancers that interact with *Aicda* only in activated B cells, 260-kb (**B**) and 160-kb (**C**) away from *Aicda*'s TSS respectively. The highlighted regions indicate our regions of interest with clear enrichment of 5hmC peaks in activated B cells. **(D-E)** Analysis of 5hmC-signal enrichment 24, 48 and 72 hours after activation of WT (blue lines) and TET2/3 DKO (red lines) B cells in (**D**) the known TetE2 and TetE1 enhancers and **(E)** in the novel putative enhancers (highlighted yellow regions shown in **(B)** and **(C)** respectively. Error bars represent the standard error of the mean. These data were obtained from our previously published work, discussed in chapter 1. There is a significant 5hmC increase (*) in all four regions of WT cells, but not *Tet2/3 DKO* B cells, after 72h of activation compared to resting B cells (Welch Two Sample t-test's p values of 0.09366 and 0.0814 for **(D)** left and right panel respectively; and 0.08347 and 0.06413 for (**E**) left and right panel, respectively) (* p-value < 0.1).

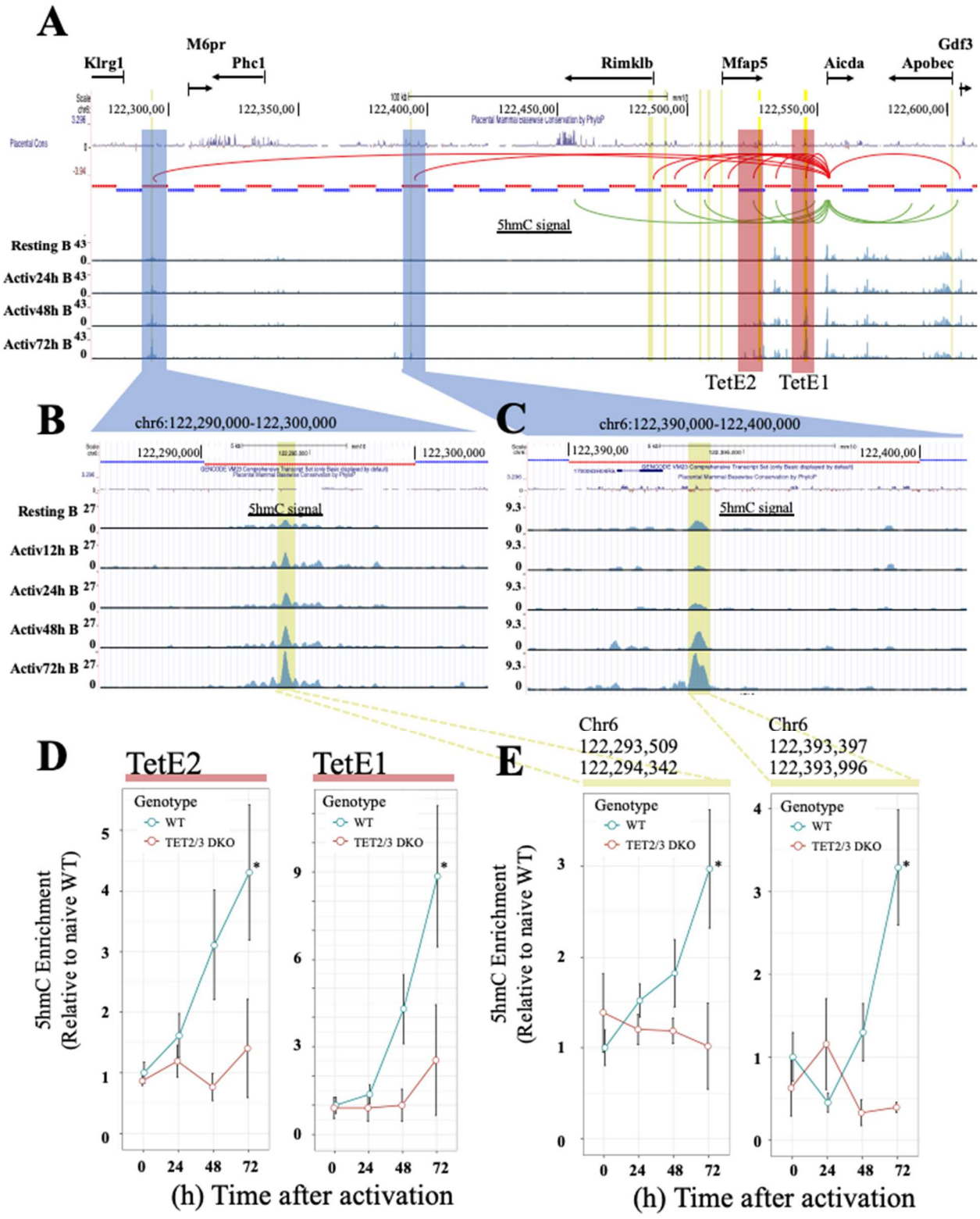
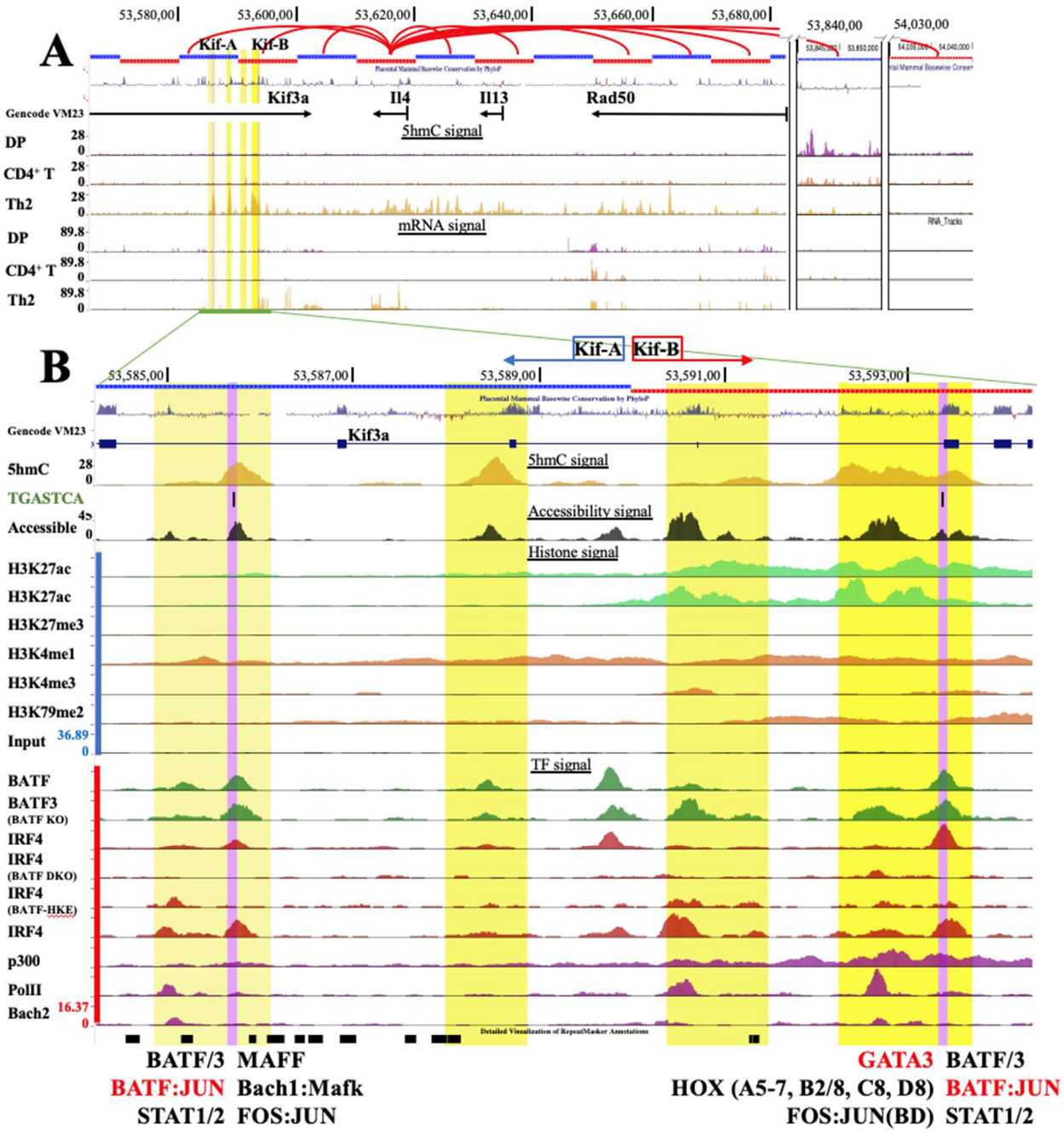

A
Klrg1
M6pr
Phc1
Rimklb
Mfap5
Aicda
Apobec
Gdf3
122,300,00
122,350,00
122,400,00
122,450,00
122,500,00
122,550,00
122,600,00
5hmC signal
Resting B
Activ24h B
Activ48h B
Activ72h B
TetE2
TetE1
B
chr6:122,290,000-122,300,000
122,290,000
122,300,000
5hmC signal
Resting B
Activ12h B
Activ24h B
Activ48h B
Activ72h B
C
chr6:122,390,000-122,400,000
122,390,00
122,400,00
5hmC signal
D
TetE2
TetE1
Genotype
WT
TET2/3 DKO
5hmC Enrichment
(Relative to naive WT)
(h) Time after activation
E
Chr6
122,293,509
122,294,342
Chr6
122,393,397
122,393,996
Genotype
WT
TET2/3 DKO
5hmC Enrichment
(Relative to naive WT)
(h) Time after activation

Figure 3.3. Novel Th2 gene regulatory regions with strong BATF:IRF4 features.
(A) Genome browser overview of the top 10 interactions used to predict *Il4* gene expression in Th2 cells. Red and blue lines above the Placental conservation track represent the 10-kb windows used to process Hi-C data and generate the graph structure. Red arcs on top of the 10-kb windows represent the used interactions for Th2 cells. For continuity, the distance to the two interactions that were far apart (right side of the panel) was omitted. From top to bottom tracks are 10-kb scaffolds, placental conservation, GENCODE VS23 gene annotations, 5hmC signal for DP, CD4 T naïve and Th2 cells, followed by RNA-seq signal in same cellular order. The yellow highlight above the green segment represent two *Il4*-interacting nodes (here termed *Kif-A* and *Kif-B*) that haven't been analyzed in the literature yet for roles in *Il4* gene regulation. **(B)** is a zoomed in version of both *Kif-A* and *Kif-B* regions (the green segment). From top to bottom tracks are 10-kb scaffolds, placental conservation, GENCODE VS23 gene annotations, Th2-specific 5hmC signal, perfect match to AICE sequence, accessibility signal, H3K27ac, H3K27ac, H3K27me3, H3K4me1, H3K4me3, H3K79me2, input, BATF, BATF3 in BATF-KO condition, IRF4, IRF4 in BATF/BATF3 DKO, IRF4 in BATF/BATF3 DKO plus BATF-HKE, IRF4 in BATF KO condition, p300, PolIII and Bach2 ChIP-seqs, followed by RepeatMasker annotations. The yellow highlights represent the same regions as in **(A)**, pink highlight represents the 5hmC peaks that had co-binding of BATF, BATF and IRF4; IRF4 whose binding is lost in BATF KOs; have strong accesssibility and H3K27ac enrichment signal; and have a perfect match for the AP-1 sequence (left) and a very close match to the AP-1–IRF composite elements (AICE2; TacCnnnnTGASTCA). These two regions, one per interacting node, represent what we suspect to be previously unreported *Il4* regulatory regions.


A
53,580,00
53,600,00
53,620,00
53,640,00
53,660,00
53,680,00
53,840,00
54,030,00
Kif-A
Kif-B
Kif3a
Il4
Il13
Rad50
Gencode VM23
5hmC signal
DP
CD4+ T
Th2
mRNA signal
B
Kif-A
Kif-B
53,585,00
53,587,00
53,589,00
53,591,00
53,593,00
Placental Mammal Basewise Conservation by PhyloP
Kif3a
5hmC
TGASTCA
Accessible
Accessibility signal
Histone signal
H3K27ac
H3K27ac
H3K27me3
H3K4me1
H3K4me3
H3K79me2
Input
TF signal
BATF
BATF3
(BATF KO)
IRF4
IRF4
(BATF DKO)
IRF4
(BATF-HKE)
IRF4
p300
PolII
Bach2
Detailed Visualization of RepeatMasker Annotations
BATF/3 MAFF
BATF:JUN Bach1:Mafk
STAT1/2 FOS:JUN
GATA3 BATF/3
HOX (A5-7, B2/8, C8, D8) BATF:JUN
FOS:JUN(BD) STAT1/2

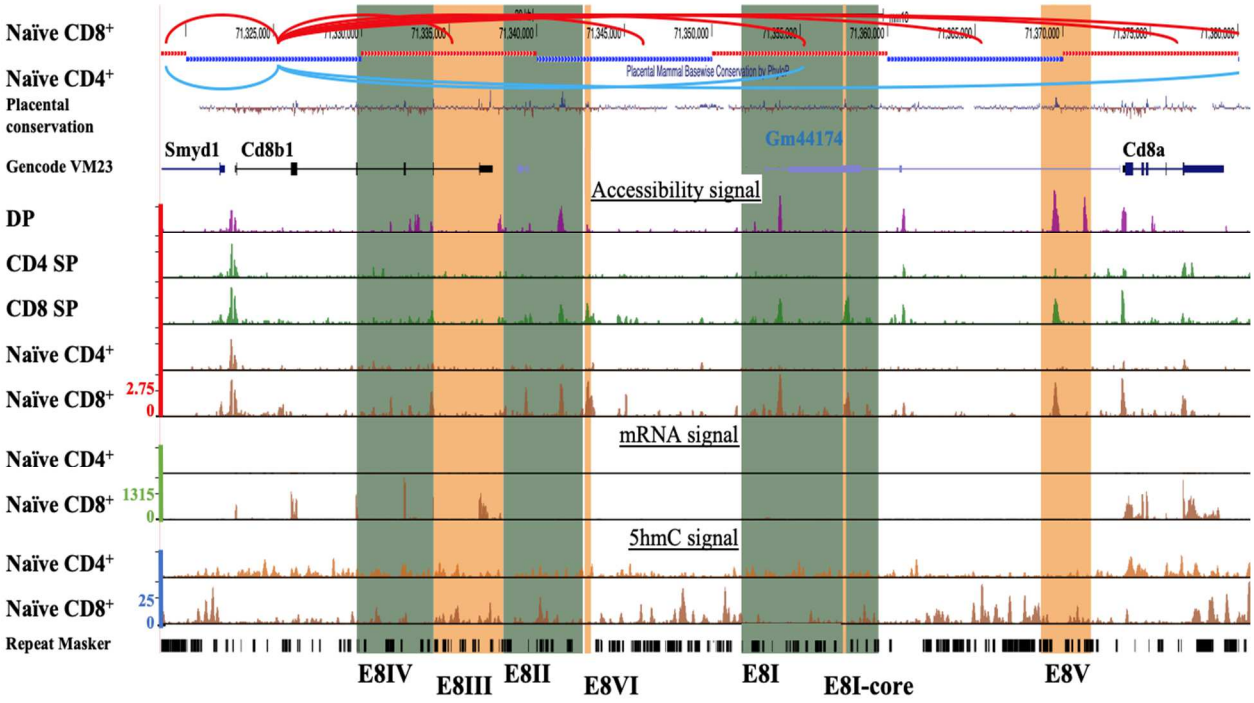


Figure 3.4. Effective selection of close- and long-range interactions in CD4/8 naïve T cells. Genome browser overview of the *Cd8ab1* gene complex in the T cell lineage. Interleaved red and blue lines represent each 10-kb window. Arcs above or below the 10-kb windows represent the *Cd8ab1* locus' interactions for CD8 (top, red) and CD4 (bottom, blue) naïve T cells. From top to bottom tracks are 10-kb scaffolds, placental conservation, GENCODE VS23 gene annotations, accessibility signal for DP, SP CD4, SP CD8, Naïve CD4 and Naïve CD8, mRNA signal for CD4 and CD8 naïve T cells followed by 5hmC signal for CD4 and CD8 naïve T cells. The orange and green highlights represent the regions corresponding to [OR containing] the annotated enhancer E8 at the bottom of the tracks. The E8I-core is located inside the E8I region. Unlike CD4 T cells, naïve CD8 T cells interact heavily with many regions in the vicinity, a feature reported to be key for upregulating Cd8b1.

3.7 Tables

Table 3.1 AUC and AUPR scores when a sample was withheld from making the averaged contact maps.

Metric Scores		Resting B	Activ72h B	Naïve CD4T	Naïve CD8T
All Available Hi-C	AUROC	0.812	0.796	0.839	0.857
	AUPR	0.772	0.743	0.797	0.815
Sample withheld	AUROC	0.808	0.79	0.84	0.851
	AUPR	0.771	0.75	0.8	0.829

3.7 Supplemental Tables and Figures

Table S3.1. Samples used in this study and their valid interactions.

Valid Interactions	Cell Type	Matched 5hmC	Used Individual	Used in averaged HiC	GEO	Citation
1,314,772,344	Resting B	X	X	X	GSE82144	Vian et al. 2018
1,060,738,232	(72h) Activated B	X	X	X	GSE82144	Vian et al. 2018
433,592,619	Effector CD8T			X	GSE158375	Lu et al. 2021
248,466,091	Naïve CD4T	X	X	X	GSE158375	Lu et al. 2021
225,882,913	Exhausted CD8T			X	GSE158375	Lu et al. 2021
213,565,710	Naïve CD8T	X	X	X	GSE158375	Lu et al. 2021
183,924,654	LSK			X	GSE99151	Johanson et al. 2018
169,355,943	DP	X	X		GSE182995	Feng et al. 2021
68,556,096	Th2	X	X		GSE66343	Ren et al. 2017

Table S3.2. Set of nodes found among resting and activated B cell samples.

		Absent in	
GeneName/NodeID	Coordinates	Resting	Activated
94953	6:122290000-122300000	X	
94963	6:122390000-122400000	X	
94969	6:122450000-122460000		X
Rimklb	6:122480000-122490000	X	
94973	6:122490000-122500000		
94974	6:122500000-122510000		
Mfap5	6:122510000-122520000	X	
94976	6:122520000-122530000		
94977	6:122530000-122540000		
94978	6:122540000-122550000		
Aicda	6:122550000-122560000		
94982	6:122580000-122590000		X
94983	6:122590000-122600000	—	X
Apobec1	6:122600000-122610000	—	
Gdf3	6:122610000-122620000		X
Dppa3	6:122620000-122630000		X
94989	6:122650000-122660000		X

Table S3.3. All data downloaded for the integrative analysis of Th2's *Il4* gene.

Assay	Antibody	Sample Name	GEO	Citation
ATAC-seq	—	ATAC_Th2	GSE159505	Kiuchi et al. 2021
ChIP-seq	H3K27ac	H3K27ac	GSE144586	Maqbool et al 2020
ChIP-seq	H3K27ac	H3K27ac_	GSE63380	Kuwahara et al. 2016
ChIP-seq	H3K27me3	H3K27me3	GSE144586	Maqbool et al 2020
ChIP-seq	H3K4me1	H3K4me1	GSE144586	Maqbool et al 2020
ChIP-seq	H3K4me3	H3K4me3	GSE144586	Maqbool et al 2020
ChIP-seq	H3K79me2	H3K79me2	GSE144586	Maqbool et al 2020
ChIP-seq	Input	Input	GSE144586	Maqbool et al 2020
ChIP-seq	anti-IRF4	IRF4	GSE64749	Bruchard et al. 2015
ChIP-seq	p300	p300	GSE40463	Adamson et al. 2013
ChIP-seq	PolII	PolII	GSE144586	Maqbool et al 2020
ChIP-seq	anti-BATF	BATF	GSE85172	Iwata et al. 2017
ChIP-seq	anti-BATF3	BATFKO BATF3	GSE85172	Iwata et al. 2017
ChIP-seq	anti-IRF4	IRF4	GSE85172	Iwata et al. 2017
ChIP-seq	anti-IRF4	DKO_IRF4	GSE85172	Iwata et al. 2017
ChIP-seq	anti-IRF4	Batf-HKE IRF4	GSE85172	Iwata et al. 2017
ChIP-seq	anti-IRF4	BATFKO IRF4	GSE85172	Iwata et al. 2017
ChIP-seq	anti-BACH2	Bach2_Thn	GSE63380	Kuwahara et al. 2016

Table S3.4. Th2 unique interactions across our T cell lineage data.

Set of interactions found among T cells		Present in			
GeneName/NodeID	Coordinates	DP	CD4 T Naive	CD8T Naive	Th2
154414	11:52800000-52810000			X	
154442	11:53080000-53090000	X			
154467	11:53330000-53340000	X			
154475	11:53410000-53420000	X			
154480	11:53460000-53470000		X	X	
154481	11:53470000-53480000	X	X	X	
Sowaha	11:53480000-53490000	X		X	
154483	11:53490000-53500000		X	X	
154485	11:53520000-53530000		X		
154487	11:53530000-53540000	X		X	
154488	11:53540000-53550000			X	
154489	11:53550000-53560000		X	X	
Kif3a	11:53560000-53570000			X	
154492	11:53580000-53590000				X
154493	11:53590000-53600000				X
154494	11:53600000-53610000	X	X		X
Il4	11:53610000-53620000	-	-	-	-
154496	11:53620000-53630000		X	X	X
Il13	11:53630000-53640000			X	X
154499	11:53650000-53660000		X		X
154500	11:53660000-53670000				X
154501	11:53670000-53680000				X
154518	11:53840000-53850000				X
154526	11:53920000-53930000	X			
154529	11:53950000-53960000			X	
154537	11:54030000-54040000				X
154569	11:54340000-54350000		X		
154598	11:54640000-54650000	X			
154658	11:55240000-55250000			X	
154667	11:55320000-55330000		X		
155836	11:67010000-67020000		X		
157501	11:83670000-83680000	X			
158046	11:89120000-89130000	X			
160526	11:113910000-113920000		X		
160589	11:114550000-114560000			X	

Table S3.5. CD4 and CD8 Naïve T cell interactions, their GNNExplainer ranking for Cd8b1 prediction, and coordinates that have a E8 enhancer.

Interactions among Naive CD4/8 T cells		**Present in**				**E8 enhancers**
GeneName/NodeID	Coordinates	CD4 T Naive		CD8 T Naive		
			GNNExplainer rank		GNNExplainer rank	
Fabp1	6:71190000-71200000	X	4			
Mir8112+Krcc1	6:71270000-71280000	X	7			
89854	6:71290000-71300000	X	2			
89855	6:71300000-71310000	X	3	X	7	
89856	6:71310000-71320000	X	1	X	6	
Cd8b1	6:71320000-71330000	-		-		
89857	6:71330000-71340000			X	8	IV, III, II
89858	6:71340000-71350000			X	3	II, VI
89859	6:71350000-71360000	X	10	X	2	I, I-core
89860	6:71360000-71370000			X	1	V
Cd8a	6:71370000-71380000			X	5	V
89862	6:71380000-71390000	X	8	X	9	
89863	6:71390000-71400000			X	10	
89864	6:71400000-71410000			X	4	
89866	6:71410000-71420000	X	6			
89868	6:71430000-71440000	X	5			
89872	6:71470000-71480000	X	9			

	AUC Score		AUPR Score	
	Top 10 3D interacting bins	Nearest 10 1D bins	Top 10 3D interacting bins	Nearest 10 1D bins
Resting B	0.8475	0.8162	0.8146	0.7820
Activ72h B	0.8588	0.8133	0.8052	0.7776
DP	0.8047	0.7845	0.7787	0.7571
Naïve CD4T	0.8492	0.8114	0.8146	0.7731
Naïve CD8T	0.8444	0.8042	0.8089	0.7674
Th2	0.8062	0.7668	0.7640	0.7392

Figure S3.1. AUC and AUPR scores of using only 10 interactions around each bin.
Every cell-specific model performed better integrating all of the long-range interactions when evaluated with either AUC or AUPR unbiased metric scores.

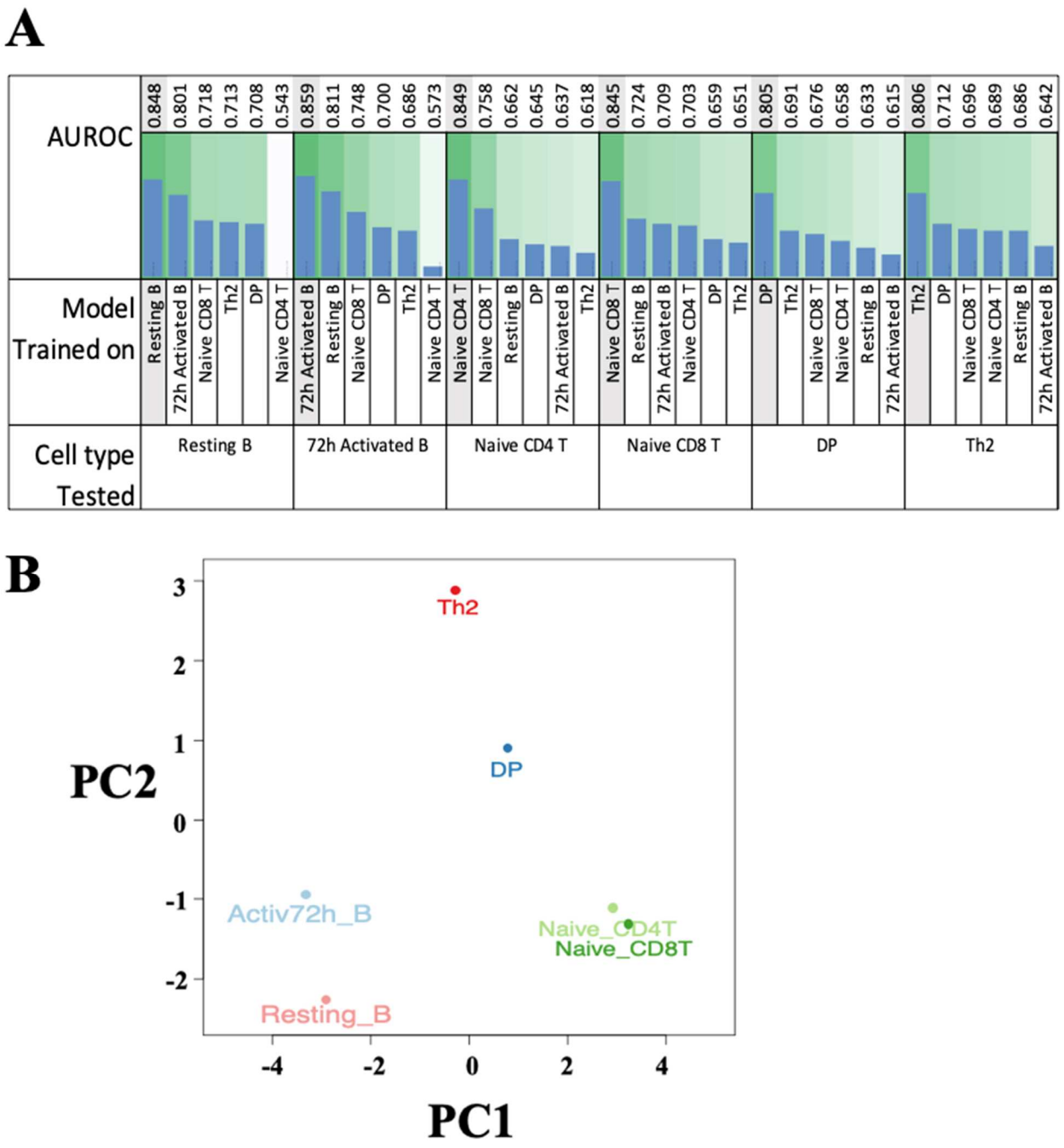


Figure S3.2. Cross-Cell AUC scores and PCA plot of RNA-seq data used in this study. **(A)** contains all of the individual AUC scores per comparison, whereas **(B)** is the PCA drawn from the top 1000 most variable genes among all six samples.

3.8 Author Contributions

E.G.-A., A.R. and F.A. conceived and designed the project. E.G.-A. designed and performed experiments and analyzed data with F.A. and D.S.-C., E.G.-A. performed data collection and analysis, E.G.-A., A.R., and F.A wrote the manuscript, with all authors contributing and providing feedback and advice.

3.9 Acknowledgements

Chapter 3, in full, is a reformatted presentation of the material currently being prepared for submission for publication as “Linking proximal and distal 5hmC enrichment to cell-specific gene regulation with graph convolutional networks” by Edahí González Avalos, Daniela Samaniego-Castruita, Anjana Rao, and Ferhat Ay. The dissertation author was the primary investigator and first author of this material.

3.10 References

Adamson, A., Ghoreschi, K., Rittler, M., Chen, Q., Sun, H.-W., Vahedi, G., Kanno, Y., Stetler-Stevenson, W.G., O'Shea, J.J. and Laurence, A. (2013). Tissue Inhibitor of Metalloproteinase 1 Is Preferentially Expressed in Th1 and Th17 T-Helper Cell Subsets and Is a Direct Stat Target Gene. PLoS ONE *8*, e59367.

Agarwal, S. and Rao, A. (1998). Modulation of Chromatin Structure Regulates Cytokine Gene Expression during T Cell Differentiation. Immunity *9*, 765–775.

Agarwal, R., and Kaye, S.B. (2003). Ovarian cancer: strategies for overcoming resistance to chemotherapy. Nature Reviews Cancer *3*, 502–516.

Amemiya, H.M., Kundaje, A. and Boyle, A.P., (2019). The ENCODE blacklist: identification of problematic regions of the genome. Scientific reports *9*, 1–5.

Ansel, K.M., Djuretic, I., Tanasa, B. and Rao, A. (2006). Regulation of Th2 Differentiation and *Il4* Locus Accessibility. Annual Review of Immunology *24*, 607–656.

Avsec, Ž., Agarwal, V., Visentin, D., Ledsam, J.R., Grabska-Barwinska, A., Taylor, K.R., Assael, Y., Jumper, J., Kohli, P. and Kelley, D.R. (2021). Effective gene expression prediction from sequence by integrating long-range interactions. Nature Methods *18*, 1196–1203.

Baguet, A. and Bix, M. (2004). Chromatin landscape dynamics of the Il4-Il13 locus during T helper 1 and 2 development. Proceedings of the National Academy of Sciences *101*, 11410–11415.

Bao, K., Carr, T., Wu, J., Barclay, W., Jin, J., Ciofani, M. and Reinhardt, R.L. (2016). BATF Modulates the Th2 Locus Control Region and Regulates CD4+ T Cell Fate during Antihelminth Immunity. The Journal of Immunology *197*, 4371–4381.

Bigness, J., Loinaz, X., Patel, S., Larschan, E. and Singh, R. (2020). Integrating long-range regulatory interactions to predict gene expression using graph convolutional networks.

Broad Institute. Picard Toolkit (2018); http://broadinstitute.github.io/picard/

Bruchard, M., Rebé, C., Derangère, V., Togbé, D., Ryffel, B., Boidot, R., Humblin, E., Hamman, A., Chalmin, F., Berger, H., Chevriaux, A., Limagne, E., Apetoh, L., Végran, F. and Ghiringhelli, F. (2015). The receptor NLRP3 is a transcriptional regulator of TH2 differentiation. Nature Immunology *16*, 859–870.

Cao, Q., Anyansi, C., Hu, X., Xu, L., Xiong, L., Tang, W., Mok, M.T.S., Cheng, C., Fan, X., Gerstein, M., Cheng, A.S.L. and Yip, K.Y. (2017). Reconstruction of enhancer–target networks in 935 samples of human primary cells, tissues and cell lines. Nature Genetics *49*, 1428–1436.

Castro-Mondragon, J.A., Riudavets-Puig, R., Rauluseviciute, I., Berhanu Lemma, R., Turchi, L., Blanc-Mathieu, R., Lucas, J., Boddie, P., Khan, A., Manosalva Pérez, N., Fornes, O., Leung, T., Aguirre, A., Hammal, F., Schmelter, D., Baranasic, D., Ballester, B., Sandelin, A., Lenhard, B.

and Vandepoele, K. (2021). JASPAR 2022: the 9th release of the open-access database of transcription factor binding profiles. Nucleic Acids Research *50*, D165–D173.

Chen, L., Grabowski, K.A., Xin, J.-P., Coleman, J., Huang, Z., Espiritu, B., Alkan, S., Xie, H.B., Zhu, Y., White, F.A., Clancy, J. and Huang, H. (2004). IL-4 induces differentiation and expansion of Th2 cytokine-producing eosinophils. Journal of Immunology *172*, 2059–2066.

Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M. and Gingeras, T.R. (2012). STAR: ultrafast universal RNA-seq aligner. Bioinformatics *29*, 15–21.

Ellmeier, W., Sunshine, M.J., Maschek, R. and Littman, D.R. (2002). Combined Deletion of CD8 Locus cis-Regulatory Elements Affects Initiation but Not Maintenance of CD8 Expression. Immunity *16*, 623–634.

Eraslan, G., Avsec, Ž., Gagneur, J. and Theis, F.J. (2019). Deep learning: new computational modelling techniques for genomics. Nature Reviews Genetics *20*, 389–403.

Feng, D., Chen, Y., Dai, R., Bian, S., Xue, W., Zhu, Y., Li, Z., Yang, Y., Zhang, Y., Zhang, J., Bai, J., Qin, L., Kohwi, Y., Shi, W., Kohwi-Shigematsu, T., Liao, S. and Hao, B. (2021). Chromatin organizer SATB1 controls the cell identity of CD4+ CD8+ double-positive thymocytes by compacting super-enhancers.

Fields, P.E., Lee, G.R., Kim, S.T., Bartsevich, V.V. and Flavell, R.A. (2004). Th2-Specific Chromatin Remodeling and Enhancer Activity in the Th2 Cytokine Locus Control Region. Immunity *21*, 865–876.

Fulco, C.P., Nasser, J., Jones, T.R., Munson, G., Bergman, D.T., Subramanian, V., Grossman, S.R., Anyoha, R., Doughty, B.R., Patwardhan, T.A., Nguyen, T.H., Kane, M., Perez, E.M., Durand, N.C., Lareau, C.A., Stamenova, E.K., Aiden, E.L., Lander, E.S. and Engreitz, J.M. (2019). Activity-by-contact model of enhancer–promoter regulation from thousands of CRISPR perturbations. Nature Genetics *51*, 1664–1669.

Glasmacher, E., Agrawal, S., Chang, A.B., Murphy, T.L., Zeng, W., Vander Lugt, B., Khan, A.A., Ciofani, M., Spooner, C.J., Rutz, S., Hackney, J., Nurieva, R., Escalante, C.R., Ouyang, W., Littman, D.R., Murphy, K.M. and Singh, H. (2012). A Genomic Regulatory Element That Directs Assembly and Function of Immune-Specific AP-1–IRF Complexes. Science *338*, 975–980.

Gülich, A.F., Preglej, T., Hamminger, P., Alteneder, M., Tizian, C., Orola, M.J., Muroi, S., Taniuchi, I., Ellmeier, W. and Sakaguchi, S. (2019). Differential Requirement of Cd8 Enhancers E8I and E8VI in Cytotoxic Lineage T Cells and in Intestinal Intraepithelial Lymphocytes. Frontiers in Immunology *10*, 409.

Guo, L., Hu-Li, J., Zhu, J., Watson, C.J., Difili ntonio, M.J., Pannetier, C. and Paul, W.E. (2002). In TH2 cells the Il4 gene has a series of accessibility states associated with distinctive probabilities of IL-4 production. Proceedings of the National Academy of Sciences *99*, 10623–10628.

Greener, J.G., Kandathil, S.M., Moffat, L. and Jones, D.T. (2021). A guide to machine learning for biologists. Nature Reviews Molecular Cell Biology *23*, 40–55.

Hamilton, W.L., Ying, R. and Leskovec, J. (2018). Inductive Representation Learning on Large Graphs. arXiv:1706.02216.

Hammal, F., de Langen, P., Bergon, A., Lopez, F. and Ballester, B. (2021). ReMap 2022: a database of Human, Mouse, Drosophila and Arabidopsis regulatory regions from an integrative analysis of DNA-binding sequencing experiments. Nucleic Acids Research *50*, D316–D325.

Harada, Y., Tanaka, S., Motomura, Y., Harada, Y., Ohno, S., Ohno, S., Yanagi, Y., Inoue, H. and Kubo, M. (2012). The 3′ Enhancer CNS2 Is a Critical Regulator of Interleukin-4-Mediated Humoral Immunity in Follicular Helper T Cells. Immunity *36*, 188–200.

Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y.C., Laslo, P., Cheng, J.X., Murre, C., Singh, H. and Glass, C.K., 2010. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. Molecular cell *38*, 576–589.

Hostert, A., Garefalaki, A., Mavria, G., Tolaini, M., Roderick, K., Norton, T., Mee, P.Joseph., Tybulewicz, V.L.J., Coles, M. and Kioussis, D. (1998). Hierarchical Interactions of Control Elements Determine CD8α Gene Expression in Subsets of Thymocytes and Peripheral T Cells. Immunity *9*, 497–508.

Imakaev, M., Fudenberg, G., McCord, R.P., Naumova, N., Goloborodko, A., Lajoie, B.R., Dekker, J. and Mirny, L.A. (2012). Iterative correction of Hi-C data reveals hallmarks of chromosome organization. Nature Methods *9*, 999–1003.

Iwata, A., Durai, V., Tussiwand, R., Briseño, C.G., Wu, X., Grajales-Reyes, G.E., Egawa, T., Murphy, T.L. and Murphy, K.M. (2017). Quality of TCR signaling determined by differential affinities of enhancers for the composite BATF–IRF4 transcription factor complex. Nature Immunology *18*, 563–572.

Johanson, T.M., Lun, A.T.L., Coughlan, H.D., Tan, T., Smyth, G.K., Nutt, S.L. and Allan, R.S. (2018). Transcription-factor-mediated supervision of global genome architecture maintains B cell identity. Nature Immunology *19*, 1257–1264.

Karbalayghareh, A., Sahin, M. and Leslie, C.S. (2021). Chromatin interaction aware gene regulatory modeling with graph attention networks.

Kipf, T.N. and Welling, M. (2017). Semi-Supervised Classification with Graph Convolutional Networks. arXiv:1609.02907.

Kiuchi, M., Onodera, A., Kokubo, K., Ichikawa, T., Morimoto, Y., Kawakami, E., Takayama, N., Eto, K., Koseki, H., Hirahara, K. and Nakayama, T. (2021). The Cxxc1 subunit of the Trithorax complex directs epigenetic licensing of CD4+ T cell differentiation. Journal of Experimental Medicine *218*, e20201690.

Krueger, F., 2015. Trim galore. A wra r tool around Cutadapt and FastQC to consistently apply quality and adapter trimming to FastQ files, 516–517.

Kuwahara, M., Ise, W., Ochi, M., Suzuki, J., Kometani, K., Maruyama, S., Izumoto, M., Matsumoto, A., Takemori, N., Takemori, A., Shinoda, K., Nakayama, T., Ohara, O., Yasukawa, M., Sawasaki, T., Kurosaki, T. and Yamashita, M. (2016). Bach2–Batf interactions control Th2-type immune response by regulating the IL-4 amplification loop. Nature Communications *7*, 12596.

Langmead, B., Trapnell, C., Pop, M. and Salzberg, S.L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome biology, 10, 1–10.

Lawrence, M., Huber, W., Pages, H., Aboyoun, P., Carlson, M., Gentleman, R., Morgan, M.T. and Carey, V.J. (2013). Software for computing and annotating genomic ranges. PLoS computational biology *9*, e1003118.

Lee, D.U. and Rao, A. (2004). Molecular analysis of a locus control region in the T helper 2 cytokine gene cluster: A target for STAT6 but not GATA3. Proceedings of the National Academy of Sciences *101*, 16010–16015.

Lee, G.R., Spilianakis, C.G. and Flavell, R.A. (2004). Hypersensitive site 7 of the TH2 locus control region is essential for expressing TH2 cytokine genes and for long-range intrachromosomal interactions. Nature Immunology *6*, 42–48.

Li, H. and Durbin, R., (2009). Fast and accurate short read alignment with Burrows–Wheeler transform. bioinformatics *25*, 1754–1760.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G. and Durbin, R., (2009). The sequence alignment/map format and SAMtools. Bioinformatics *25*, 2078–2079.

Liao, Y., Smyth, G.K. and Shi, W. (2013). featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics *30*, 923–930.

Libbrecht, M.W. and Noble, W.S. (2015). Machine learning a ications in genetics and genomics. Nature Reviews Genetics *16* .321–332.

Loots, G.G., Locksley, R.M., Blankespoor, C.M., Wang, Z.E., Miller, W., Rubin, E.M. and Frazer, K.A. (2000). Identification of a coordinate regulator of interleukins 4, 13, and 5 by cross-species sequence comparisons. Science *288*, 136–140.

Lu, J., Wang, X., Sun, K. and Lan, X., (2021). Chrom-Lasso: a lasso regression-based model to detect functional interactions using Hi-C data. Briefings in Bioinformatics *22*, bbab181.

Malik, S., Sadhu, S., Elesela, S., Pandey, R.P., Chawla, A.S., Sharma, D., Panda, L., Rathore, D., Ghosh, B., Ahuja, V. and Awasthi, A. (2017). Transcription factor Foxo1 is essential for IL-9 induction in T helper cells. Nature Communications *8*, 815.

Maqbool, M.A., Pioger, L., El Aabidine, A.Z., Karasu, N., Molitor, A.M., Dao, L.T.M., Charbonnier, G., van Laethem, F., Fenouil, R., Koch, F., Lacaud, G., Gut, I., Gut, M., Amigorena, S., Joffre, O., Sexton, T., Spicuglia, S. and Andrau, J.-C. (2020). Alternative Enhancer Usage and Targeted Polycomb Marking Hallmark Promoter Choice during T Cell Differentiation. Cell Reports *32*, 108048.

Martin, M., (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet. journal *17*, 10–12.

Puig, R.R., Boddie, P., Khan, A., Castro-Mondragon, J.A. and Mathelier, A. (2021). UniBind: maps of high-confidence direct TF-DNA interactions across nine species. BMC Genomics *22*, 482.

Quinlan, A.R. and Hall, I.M., (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics *26*, 841–842.

R Core Team (2017). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/

Ren, G., Jin, W., Cui, K., Rodrigez, J., Hu, G., Zhang, Z., Larson, D.R. and Zhao, K. (2017). CTCF-Mediated Enhancer-Promoter Interaction Is a Critical Regulator of Cell-to-Cell Variation of Gene Expression. Molecular Cell *67*, 1049–1058.e6.

Sahoo, A., Alekseev, A., Tanaka, K., Obertas, L., Lerman, B., Haymaker, C., Clise-Dwyer, K., McMurray, J.S. and Nurieva, R. (2015). Batf is important for IL-4 expression in T follicular helper cells. Nature Communications *6*, 7997.

Sanborn, A.L., Rao, S.S.P., Huang, S.-C., Durand, N.C., Huntley, M.H., Jewett, A.I., Bochkov, I.D., Chinna n, D., Cutkosky, A., Li, J., Geeting, K.P., Gnirke, A., Melnikov, A., McKenna, D., Stamenova, E.K., Lander, E.S. and Aiden, E.L. (2015). Chromatin extrusion explains key features of loop and domain formation in wild-type and engineered genomes. Proceedings of the National Academy of Sciences *112*, E6456–E6465.

Servant, N., Varoquaux, N., Lajoie, B.R., Viara, E., Chen, C.-J., Vert, J.-P., Heard, E., Dekker, J. and Barillot, E. (2015). HiC-Pro: an optimized and flexible pipeline for Hi-C data processing. Genome Biology *16*, 259.

Vian, L., Pękowska, A., Rao, S.S.P., Kieffer-Kwon, K.-R., Jung, S., Baranello, L., Huang, S.-C., El Khattabi, L., Dose, M., Pruett, N., Sanborn, A.L., Canela, A., Maman, Y., Oksanen, A., Resch, W., Li, X., Lee, B., Kovalchuk, A.L., Tang, Z. and Nelson, S. (2018). The Energetics and Physiological Impact of Cohesin Extrusion. Cell *175*, 292–294.

Vijayanand, P., Seumois, G., Simpson, Laura J., Abdul-Wajid, S., Baumjohann, D., Panduro, M., Huang, X., Interlandi, J., Djuretic, Ivana M., Brown, Daniel R., Sharpe, Arlene H., Rao, A. and Ansel, K. Mark (2012). Interleukin-4 Production by Follicular Helper T Cells Requires the Conserved Il4 Enhancer Hypersensitivity Site V. Immunity *36*, 175–187.

Whalen, S., Truty, R.M. and Pollard, K.S. (2016). Enhancer–promoter interactions are encoded by complex genomic signatures on looping chromatin. Nature Genetics *48*, 488–496.

Wickham, H., 2009. Elegant graphics for data analysis. Media *35*, 10–1007.

Xi, Y. and Li, W. (2009). BSMAP: whole genome bisulfite sequence MAPping program. BMC Bioinformatics *10*, 232.

Yamashita, M., Ukai-Tadenuma, M., Kimura, M., Omori, M., Inami, M., Taniguchi, M. and Nakayama, T. (2002). Identification of a Conserved GATA3 Response Element Upstream Proximal from the Interleukin-13 Gene Locus. Journal of Biological Chemistry *277*, 42399–42408.

Yardımcı, G.G., Ozadam, H., Sauria, M.E.G., Ursu, O., Yan, K.-K., Yang, T., Chakraborty, A., Kaul, A., Lajoie, B.R., Song, F., Zhan, Y., Ay, F., Gerstein, M., Kundaje, A., Li, Q., Taylor, J., Yue, F., Dekker, J. and Noble, W.S. (2019). Measuring the reproducibility and quality of Hi-C data. Genome Biology *20*, 57.

Ying, R., Bourgeois, D., You, J., Zitnik, M. and Leskovec, J. (2019). GNNExplainer: Generating Explanations for Graph Neural Networks. arXiv:1903.

Yosef, N., Shalek, A.K., Gaublomme, J.T., Jin, H., Lee, Y., Awasthi, A., Wu, C., Karwacz, K., Xiao, S., Jorgolli, M., Gennert, D., Satija, R., Shakya, A., Lu, D.Y., Trombetta, J.J., Pillai, M.R., Ratcliffe, P.J., Coleman, M.L., Bix, M. and Tantin, D. (2013). Dynamic regulatory network controlling TH17 cell differentiation. Nature *496*, 461–468.

Conclusion

Here we have explored the role 5hmC enrichment plays in maintaining gene expression in the B cell activation context, as a predictor of gene expression, and also as a marker for putative distal enhancers that can be linked to their target genes through the 3D genome structure. Our work deleting TET proteins, responsible for 5hmC deposition through 5mC oxidation, showed that the presence of 5hmC in the key regulatory element TetE1 is important for sustained expression of the gene *Aicda*, required for CSR during B cell activation. Furthermore, we showed that 5hmC signal enrichment alone (throughout the gene body and around the promoter) can effectively be used to train models in the binary classification task of gene expression, yielding sample-specific models that reach high performance (median AUC scores above 0.87); we also showed that 5hmC models can be generalized with conserved prediction ability (model trained in all data got an AUC score of 0.86). Finally, we demonstrated that putative enhancers can be found by the integration of 3D structure when solving the task of binary classification of gene expression by 5hmC signal enrichment. The windows likely to contain the putative enhancers are found by GNNExplainer, ranking the most important 10-kb interactions used by the predictive models. The putative enhancers can be further defined by locating the 5hmC peaks inside these 10-kb windows.

Our research in the use of 5hmC signal enrichment for the predicting, both gene expression and distal enhancers, highlights the importance of this specific modification in gene regulation. Combined with our predictive models, 5hmC signal that can be obtained from DNA alone can allow us to study samples whose viability is compromised, such as FFPE preserved tissues. The goal when predicting the gene expression is to explore mechanisms controlling (or driving) the expression of genes. When using a set of features that represent the 5hmC signal enrichment in the gene body and throughout the promoter, we found that the most important fixed-size bins (features)

are those closely located both upstream (200-bp) and downstream (600-bp) of the gene's transcription start site as suggested by our DeepLift results from the Promoter-related bins. From the variable-sized bins representing the gene body, our DeepLift analysis indicated that the first couple of bins were the most significant, further corroborating our observations that the most important bins associated with promoters occur just 3' of the promoter and just within the gene body. These observations were confirmed by analysing either sample-specific trained models, or across the specialized models "Immuno" and "Embryo". Although ESC-related models did not achieve as high performance as the models trained using differentiated cells, the most important features are shared, suggesting a similar mechanism for 5hmC deposition and its subsequent association with gene expression. Our results emphasize that the further study of 5hmC deposition dynamics in and around transcription start sites may provide substantial insights into how TET and 5hmC dysregulation may drive cells toward oncogenic transformation.