Inaugural dissertation for
obtaining the doctoral degree
of the
Combined Faculty of Mathematics, Engineering
and Natural Sciences of the
Ruprecht-Karls-Universität Heidelberg

Presented by

M.Sc. Ana Luísa Simões Costa
born in Coimbra, Portugal
Oral examination: June 23, 2023

Computational approaches to the study of chromatin in disease: examples from cancer and infection

Referees: Prof. Dr. Carl Herrmann

Dr. Marina Lusic

Declaration of Authorship

I hereby confirm that I have authored this dissertation independently and without the use of other sources than the ones indicated. I have not yet presented this thesis or parts thereof to a university as part of an examination or degree. This work was carried out in the Health Data Science Unit at the Center for Quantitative Analysis of Molecular and Cellular Biosystems (BioQuant) in the group of Prof. Dr. Carl Herrmann.

Heidelberg, April 17, 2023

. . . . . .

Ana Luísa Simões Costa

"Wenn die Nachtigallen aufhören zu schlagen,

fangen die Grillen an zu zirpen."

— Marie von Ebner-Eschenbach

"All we have to decide is what to do with the time that is given us."

— Gandalf, in *The Fellowship of the Ring,* J.R.R. Tolkien

Acknowledgements

First and foremost, to my supervisor **Prof. Dr. Carl Herrmann**, for all his guidance, support, and trust during these formative years. This journey helped me learn and grow extensively in many ways, and that was due to the opportunities he has offered me over these past years. Working alongside him and the team he has formed was inspiring and life-changing for me. Thank you so much, Carl, I am really grateful for all your ideas, your enthusiasm, and all the great discussions we had along the way.

To **Dr. Marina Lusic** for her time on the revision of this thesis, her perceptive suggestions on the course of our collaboration, and her advice both in the context of our project and as a member of the Thesis Advisory Committee.

To **Dr. Bernhard Radlwimmer and Dr. Michael Fletcher** for all the fruitful discussions we had in the past years, for all their time, dedication, and guiding advice which made our project grow.

To **Dr. Bojana Lucic, Mona Rheinberger, and Martin Kampmann** for all the work they accomplished for our project and for all the effort into the incredible collaboration they helped build.

To **Prof. Dr. Frank Lyko and Dr. Marco Binder** for their time and availability in the revision of this thesis as well as their participation in my thesis defense as examiners.

To **Prof. Dr. Dirk Grimm and PD Dr. Christian Fufezan** for their valuable par-

ticipation, insight, and helpful feedback over the course of this thesis as members of the Thesis Advisory Committee.

To my current and former colleagues and friends from the Biomedical Genomics group, particularly: **Dr. Andrés Quintero, Daria Doncevic, (almost Dr.) Lin Yang, Youcheng Zhang, Dr. Ashwini K. Sharma, Nils Mechtel, Qian-Wu Liao, and Dr. Carlos Ramirez**. Also, to our extended HDSU members **Pablo Naranjo and Dr. Nelida Palau**. This would not have been the same without all their companionship, kind words and encouragement. I am grateful to Lin, **Elizaveta Chernova, and Katharina Mikulik** for the work they dedicated into the projects found in this thesis and all the amazing perspective they brought with them.

To **Cathrin Hollenbach and Manuela Schäfer** for all their help on the administrative side of things and for always being there to answer my silly bureaucratic questions.

To **Dr. Carolina Gámez**, for all her kindness, companionship, and advice, to **Michael Ludäscher and Paula González**, for their friendship and in honour of our amazing D&D sessions (which brought me so much joy during the pandemic), to **Pedro Mateus and Inês Lorga**, for all their support, and to **Dr. Jeremy de Sousa and Dr. Hélène Ma** for the amazing times spent together between Frankfurt and Heidelberg.

À minha mãe, à minha irmã, e aos meus avós Luís e Maria, por todos os momentos em que me apoiaram e estiveram do meu lado, não só nestes últimos anos, mas desde sempre. Por todas as vezes em que me trouxeram uma palavra de conforto e boa disposição enquanto estive longe (e perto).

To **Tiago** for all the countless times he was there for me, even when he was far away. For all his encouragement, patience, and for believing in me more than I believe in myself.

List of publications

Thesis-related publications

- Rheinberger, **Costa** et al. (2022) *Genomic profiling of HIV-1 integration in microglia cells links viral integration to the topologically associated domains*, *Cell Reports*

Manuscripts in preparation / under review

- **Costa**, Yang et al. under preparation

Zusammenfassung

Hintergrund: Ansätze des maschinellen Lernens werden in der biologischen Forschung immer häufiger eingesetzt, da sie ein besseres Verständnis der komplexen Zelldynamik ermöglichen. Die Epigenetik umfasst Prozesse, die die Genexpression modulieren können und nicht von der Genomsequenz abhängen. Oftmals werden epigenetische Veränderungen mit Krankheiten in Verbindung gebracht. In dieser Arbeit haben wir mehrere computergestützte Ansätze angewandt, um die epigenetische Landschaft von Krankheitszuständen zu charakterisieren, die durch eine Infektion mit dem Humanen Immundefizienz-Virus und Krebs im Gehirn verursacht werden.

Ergebnisse: Im ersten Teil dieser Arbeit haben wir die nicht-negative Matrixfaktorisierung angewandt, um eine epigenetische Zustandskarte für die C20-Mikroglia-Zelllinie zu erstellen und den Zusammenhang zwischen Integration und Epigenetik im Zusammenhang mit der HIV-1-Infektion zu untersuchen. Mithilfe von Random-Forest-Modellen konnten wir feststellen, dass genomische Ziele der HIV-1-Integration von der ursprünglichen epigenetischen Landschaft beeinflusst werden und dass die Infektion zu Veränderungen der Chromatin-Zugänglichkeit und der TF-Bindung führt. Darüber hinaus fanden wir heraus, dass Regionen, die häufig von der viralen Integration betroffen sind, mit Chromatinstrukturen höherer Ordnung verbunden sind, insbesondere mit topologisch assoziierten Domänen. Im zweiten Teil dieser Arbeit haben wir die CGI von vier Glioblastom-Subtypen charakterisiert und einen neuen Phänotyp der CGI-Hypermethylierung identifiziert, der mit dem RTK-II-Subtyp assoziiert ist und sich von dem für den IDH-Subtyp

beobachteten Phänotyp unterscheidet. Wir verglichen die CGI-Hypermethylierungsphänotypen, die mit den IDH- und RTK-II-Subtypen assoziiert sind, unter Verwendung von Zufallswäldern und verwenden Vorläuferzustände, um die Tendenz innerhalb jeder CGI zu bewerten, hypermethyliert zu werden. Wir haben festgestellt, dass die CGI, die bei Krebs am ehesten hypermethyliert werden, bereits in undifferenzierten Zellstadien markiert sind. Außerdem haben wir festgestellt, dass die RTK-II-CGI-Hypermethylierung das Gleichgewicht zwischen astrogenem und neurogenem Schicksal stört.

Schlussfolgerungen: Diese Arbeit liefert neue Einblicke in die Epigenetik der HIV-1-Integration und der CGI-Hypermethylierung im Glioblastom. Durch einen genomischen und epigenomischen datengesteuerten Ansatz betonen wir die Bedeutung rechnerischer Ansätze wie nicht-negative Matrixfaktorisierung, Random Forest und Bayes'sche Netzwerke für die epigenetische Forschung, da diese einen ganzheitlichen Blick auf die globalen Auswirkungen der viralen Integration und CpG-Insel-Hypermethylierung in menschlichen Zellen ermöglichen.

Abstract

Background: Machine learning approaches are becoming increasingly common in biological research, as these allow for a better understanding of the complex cell dynamics. Epigenetics encompasses processes able to modulate gene expression that do not depend on genomic sequence. Oftentimes, epigenetic alterations have been linked to disease. In this thesis, we applied several computational approaches to characterise the epigenetic landscape of diseased states caused by Human Immunodeficiency Virus infection and cancer in the brain.

Results: On the first part of this thesis, we applied non-negative matrix factorisation to build an epigenetic state map for the C20 microglial cell line and assessed the connection between integration and epigenetics in the context of HIV-1 infection. Through random forest models, we observed that genomic targets of HIV-1 integration are influenced by the initial epigenetic landscape and that infection leads to changes in the chromatin accessibility and TF binding. Furthermore, we found that regions often targeted by viral integration are associated to higher order chromatin structures, in particular topologically associated domains. On the second part of this thesis, we characterised CpG islands (CGI) of four glioblastoma subtypes and identified a new phenotype of CGI hypermethylation associated to RTK-II subtype, different from the one observed on the IDH subtype. We compared the CGI hypermethylation phenotypes associated to the IDH and RTK-II subtypes using random forests and use progenitor states to assess the tendency within each CpG island to become hypermethylated. We observed that CGI

most likely to become hypermethylated in cancer are marked already on undifferentiated cell states. Moreover, we observed that RTK-II CGI hypermethylation disturbs the astrogenic/neurogenic fate balance.

Conclusions: This thesis provides novel insights into the epigenetics of HIV-1 integration and CGI hypermethylation in glioblastoma. Through a genomic and epigenomic data-driven approach, we emphasise the importance of computational approaches like non-negative matrix factorisation, random forest, and bayesian networks into epigenetic research, as these provided an hollistic view of the global effects of viral integration and CpG island hypermethylation in human cells.

Table of Contents

List of Abbreviations

AIDS	Acquired Immune Deficiency Syndrome
AML	Acute myeloid leukaemia
ART	Antiretroviral therapy
ATAC	Assay for Transposase-Accessible Chromatin
AUC	Area Under the Curve
CGI	CpG island
ChIP	Chromatin immunoprecipitation
CIMP	CpG island methylator phenotype
CNS	Central nervous system
DNA	Deoxyribonucleic acid
GO	Gene ontology
HAND	HIV-associated neurocognitive disorders
HIV	Human Immunodeficiency Virus
HMM	Hidden Markov Model
H3K27ac	Acetylation of lysine 27 on histone H3
H3K27me3	Tri-methylation of lysine 27 on histone H3
H3K36me3	Tri-methylation of lysine 36 on histone H3
H3K4me1	Mono-methylation of lysine 4 on histone H3
H3K4me3	Tri-methylation of lysine 4 on histone H3
H3K79me	Methylation of lysine 79 on histone H3
H3K9ac	Acetylation of lysine 9 on histone H3

H3K9me2	Di-methylation of lysine 9 on histone H3
H3K9me3	Di-methylation of lysine 9 on histone H3
IDH	Isocitrate Dehydrogenase
iPSC	Induced Pluripotent Stem Cell
IS	Integration site
KD	Knock-down
LM	Linker-mediated
MDM	Monocyte-derived macrophage
MES	Mesenchymal
NMF	Non-negative matrix factorization
NOMe	Nucleosome occupancy and methylome
NP	Neural progenitors
PCA	Principal Component Analysis
PCR	Polymerase chain reaction
PRC	Polycomb repressive complex
RF	Random Forest
RNA	Ribonucleic acid
ROC	Receiver Operating Characteristic
RPKM	Reads per Kilobase of exon per million
RRBS	Reduced representation bisulfite sequencing
RTK	Receptor tyrosine kinase
SE	Super-Enhancer
TAD	Topologically-associated domain
TET	Ten-eleven Translocation
TF	Transcription factor
TFBS	Transcription factor binding site
t-SNE	t-distributed Stochastic Neighbor Embedding
WGBS	Whole-genome bisulfite sequencing
WT	Wild type

List of Figures

List of Tables

Thesis outline

Firstly, the background of the findings can be found in the **Introduction**. This thesis is divided into two main sections, both focused on the study of epigenomic changes in two conditions: HIV-1 infection and cancer. The chapter on ***Epigenomics of HIV-1 integration in microglial cell model hints on viral-driven changes in 3D genome structure*** summarises the main project, on the chromatin interplay with HIV-1 integration[1]. Then, ***Characterisation of distinct CpG island methylator phenotypes in glioblastoma***, revolves around the study of the epigenomic landscape leading to the CpG island methylator phenotype. Each of the two chapters is divided into *Motivation*, which integrates the main background for each project, *Data*, where the datasets used and their sources are described, *Methodology*, including all the methodology applied, *Results*, where the findings are reported, and *Discussion*, which includes the interpretation of the results by sub-sections. Finally, **Conclusion** focuses on the commonalities between both projects.

Notes on the text:

- Over the course of the text, I use the first-person singular *"I"* on all the contributions I am the main source of or where I independently generated results;

- The first-person plural *"we"* refers to any results or analysis where I was not the only source of the ideas or results, referring to analysis suggested by any collaborators, such as Dr. Marina Lusic, Dr. Bojana Lucic, and Mona Rheinberger (Chapter 2), or Dr. Bernhard

[1] This work has been published in Cell Reports, as *Genomic profiling of HIV-1 integration in microglia cells links viral integration to the topologically associated domains* (Rheinberger et al. 2023)

Radlwimmer and Dr. Michael Fletcher (Chapter 3), analysis performed by, together with, or upon suggestion of my supervisor, Prof. Dr. Carl Herrmann (both Chapter 2 and 3), or analysis performed by Lin Yang (Chapter 3);

- In the cases where the analysis or data generation can be fully attributed to another person or group, this is indicated as such in the footnotes or main text.

Chapter 1

Introduction

1.1 Epigenetics and gene regulation in eukaryotes

Together with the genome, the epigenome is a dynamic key-modulator of gene expression in the cells, encompassing sequence-independent processes involved in defining transcriptional cell identity (Allis and Jenuwein 2016; Rivera and Ren 2013; Waddington 2012). DNA methylation, histone modifications, ATP-dependent chromatin-remodeling, and various RNA-mediated mechanisms are able to modulate gene expression. These processes lead to alterations individually or synergistically, mainly at the level of transcription, through the differential access of transcription factors (TFs) to regulatory elements such as promoters (proximally) and enhancers (distally) **(Figure 1.1)** (Li, Carey, and Workman 2007; Carter and Zhao 2021; Schoenfelder and Fraser 2019). In turn, these elements lead to alterations on the chromatin structure.

Chromatin is a DNA-protein complex organised into nucleosomes (Li, Carey, and Workman 2007). Chromatin structure is nonuniform and highly dynamic throughout the genome (Li, Carey, and Workman 2007). It ranges from compacted, as facultative or constitutive heterochromatin, to accessible, when in active regulatory loci or genes, as euchromatin (Thurman et al. 2012; Klemm, Shipony, and Greenleaf 2019). The dynamics of this landscape are susceptible to both environmental or developmental cues, as

context-specific gene expression depends on epigenetic control (Zhu et al. 2013).

1.1.1 Histone modifications

Histones are alkaline proteins constituting the nucleosome, the basic unit of chromatin. The nucleosome is composed of approximately 146 bp of DNA wrapped around a histone octamer with four positively charged core histones (H2A, H2B, H3 and H4) pairs (Peterson and Laniel 2004). Separated by 10 to 60 bp of 'linker' DNA, these form a 'beads-on-a-string' composition (Peterson and Laniel 2004). Albeit not part of the nucleosome itself, a linker histone (H1) binds to it and is essential for its organization (Graziano et al. 1994). Modifications to the histone N-terminal tails occur post-translationally, covalently, and are able to modulate chromatin structure and to recruit enzymes to influence transcription, replication and recombination (Bannister and Kouzarides 2011; Allfrey, Faulkner, and Mirsky 1964; Barth and Imhof 2010). Histone acetylation, phosphorylation, and methylation are the most well-studied modifications (Bannister and Kouzarides 2011). The modifications alter the chromatin structure by affecting the interaction between DNA and the histone, making DNA more or less accessible (Bannister and Kouzarides 2011). Histone acetylation depends on the action of histone acetyltransferases (*writers*) and histone deacetylases (*erasers*). The addition of an acetyl group neutralizes the positive charge of the lysine, decreasing the interaction between the histones and the DNA, leading to more accessibility (Bannister and Kouzarides 2011). Phosphorylation is controlled by kinases and phosphatases, and histone methylation depends on the action of methyltransferases (*writers*) and demethylases (*erasers*) (Eberharter and Becker 2002; Bannister and Kouzarides 2011; Youn 2017). While lysine can be mono-, di-, or trimethylated, arginine can be mono- or dimethylated (Youn 2017).

Combinations of different histone modifications are typically linked to defined states of gene activation/repression **(Figure 1.1)**. Active chromatin is characterised by the presence of euchromatic histone modifications, such as mono- and tri-methylation of lysine 4 on histone H3 (H3K4me1/3), acetylation of lysine 9 on histone H3 (H3K9ac), monomethylation of lysine 20 on histone H4, or acetylation of lysine 27 on histone H3

(H3K27ac). Transcriptional activation within gene bodies is associated to enrichment in tri-methylation of lysine 36 on histone H3 (H3K36me3) and methylation of lysine 79 on histone H3 (H3K79me) (Lim, Shannon, and Hardy 2010). On the other hand, repressed chromatin is typically linked to the presence of tri-methylation of lysine 27 on histone H3 (H3K27me3), dependent on the Polycomb repressive complex 2, or di- and tri-methylation of lysine 9 on histone H3 (H3K9me) (Montavon et al. 2021).

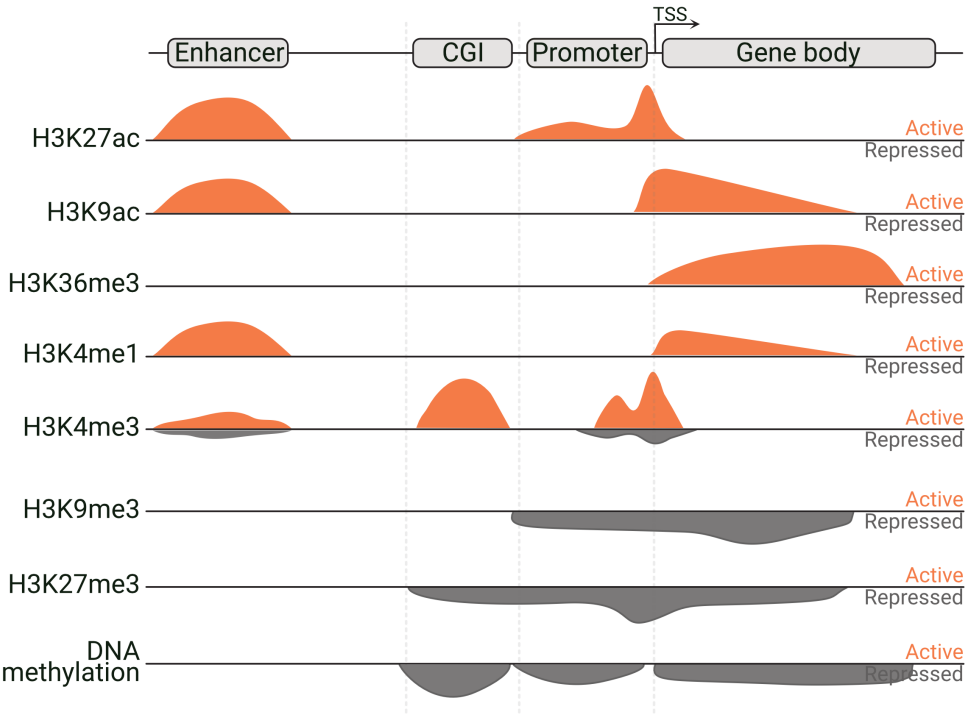


Figure 1.1: Diagram of the most studied histone modifications, comparison with DNA methylation, and their influence on transcription. Based on Barth and Imhof (2010), Lim et al (2010), and Jiang and Mortazavi (2018).

1.1.2 DNA methylation

DNA methylation is an important modulator of gene expression. However, it does not only serve as an important transcriptional regulator, but also as a central player in normal development, imprinting, genome stability (mainly through DNA mismatch repair), or in-

activation of the X chromosome (Li, Beard, and Jaenisch 1993; Zhou and Robertson 2016; Csankovszki, Nagy, and Jaenisch 2001). In mammals, DNA methylation occurs mostly at the C5-cytosine (5-methylcytosine (5mC)) in CpG dinucleotides. The addition of methyl groups is carried out by DNA methyltransferases. In humans, these are DNMT1, DNMT2, DNMT3A, DNMT3B and DNMT3L (Lyko 2018). DNMT3A and DNMT3B are both involved into *de novo* methylation, DNMT1 is associated with the maintenance of the existing DNA methylation, DNMT2 is known to methylate several transfer RNAs, and DNMT3L serves as an accessory protein to DNMT3A and DNMT3B (Lyko 2018; Gujar, Weisenberger, and Liang 2019). Removal of methyl groups, or demethylation, of 5mC converts it into hydroxymethylcytosine (5hmC), in a process catalised by Ten-eleven translocation (TET) enzymes (Tahiliani et al. 2009).

As self-reinforcers, the DNA methylation and histone modifications interplay is indispensable for trancription control in development, as these cooperate to mediate gene silencing. It is suggested that DNA methylation drives the histone modifications and vice versa, as histone modifications can also recruit DNA methyltransferases to certain loci (Vaissière, Sawan, and Herceg 2008). Methylcytosine-binding proteins can recruit histone deacetylases and DNA methyltransferases interact with the Polycomb repressive complex 2 (PRC2) protein EZH2, the histone methyltransferase which catalises the trimethylation of H3K27 (Nan et al. 1998; Cedar and Bergman 2009). Furthermore, the presence of DNA methylation is known to inhibit the activating H3K4 methylation and direct H3K9me2, the latter through the interaction between DNMT1 and the G9a histone methyltransferase, evidencing a cooperative role of both epigenetic players in the maintenance of gene repression (Cedar and Bergman 2009).

While single CpGs are typically methylated, clusters of CpGs organized into CpG islands (CGIs), frequently located on gene promoters, are mostly demethylated (Zemach et al. 2010). More than half of the human promoters contain a CGI in the 5' end (Shen et al. 2007). Methylation at CGI promoters is classically associated with gene silencing, and globally CGI methylation is variable according to differentiation and tissue-specificity (Bird 1986; Deaton et al. 2011; Meissner et al. 2008). Nevertheless, an activating role

for DNA methylation in CGIs was observed previously, challenging this notion (Yu et al. 2013). CpG island shores (2 KB regions bordering the islands) are generally found to be also hypomethylated (Nishiyama and Nakanishi 2021). Recently, the concept of conserved DNA methylation *canyons* has also been introduced, as long portions of the genome are exceptionally hypomethylated (Jeong et al. 2014). These regions are typically associated to enrichments of H3K27me3 and H3K4me3, denoting an interplay between DNA methylation and histone modifications (Jeong et al. 2014).

1.1.3 Transcription factors

TFs are a highly-conserved protein class with affinity for specific DNA sequence motifs found throughout the whole genome within regulatory regions (Spitz and Furlong 2012). TF access to regulatory regions, such as core promoters, enhancers, silencers or insulators, holds one of the most impactful effects of epigenetic alterations, as it defines cell type-specific gene expression (Haberle and Stark 2018; Pang and Snyder 2020; Burgess-Beusse et al. 2002). Enhancers and their associated TFs can distally activate or increase transcription on a promoter (Banerji, Rusconi, and Schaffner 1981; Spitz and Furlong 2012). On the other hand, silencers and insulators hold repressive and insulator abilities respectively (Pang and Snyder 2020; Burgess-Beusse et al. 2002).

In addition to their regulatory function, TFs are essential elements on the establishment of transcriptional programmes essential for cell response, identity, differentiation, and development (Carter and Zhao 2021; Vaquerizas et al. 2009). Thus, it is not surprising that the mutations directly in the coding genes for TFs, in TF motifs, or leading to putative TF motifs are underlying causes of disease, as it is the case for heart conditions, mental disorders, or cancer (Lee and Young 2013; Schott et al. 1998; Bell et al. 2015; Bae et al. 2022). Nevertheless, only a few human TFs have been annotated to a regulatory function, so further research is necessary (Vaquerizas et al. 2009).

1.1.4 Chromatin structure and nuclear organization

The higher order chromatin structure inside the nucleus is an important aspect of biological function, as it is linked to gene regulation, DNA repair, and replication (Bickmore 2013; Mirabella, Foster, and Bartke 2016). From larger to smaller scale, the genome is folded from chromosome territories, chromatin compartments, topologically associating domains (TADs), subTADs (or intra-TADs), and finally chromatin loops **(Figure 1.2)**. These morphologies are meant to ease the contact of regulatory elements with their targets. Epigenetic mechanisms are also involved in chromosomal organization (Dai, Ramesh, and Locasale 2020; Jiang and Mortazavi 2018).

Chromosome conformation capture methods, such as capture-on-chip (4C), capture carbon copy (5C), and more recently Hi-C, have been used to assess cross-linked contacts in the genome, allowing the construction of 3D profiles for many cell types (Belton et al. 2012; Simonis et al. 2006; Dostie et al. 2006). Hi-C represents an important advance in the field as it allows an unbiased assessment of the genome for interactions (all with all). Later on, promoter capture Hi-C was also developed to refine Hi-C for the detection of distal promoter-interacting regions (Schoenfelder et al. 2018).

Topologically associating domains

TADs were discovered through low-resolution Hi-C heatmaps as megabase-scale regions where DNA sequences exhibit a higher interaction frequency with each other within their own domain in comparison with external DNA sequences (Dixon et al. 2012; Beagan and Phillips-Cremins 2020). These structures modulate transcriptional regulation, constraining the interactions between cis-regulatory elements. Furthermore, the biological importance of the TADs is evidenced by their conservation among cell types and species (Dixon et al. 2012). TADs are also known to correlate with other genomic and epigenomic features, like histone modifications or DNA replication (Yang et al. 2019; McArthur and Capra 2021).

The regions limiting TADs, which act as insulatory elements, are known as TAD boundaries (McArthur and Capra 2021). H3K36me3, transcription start sites (TSSs),

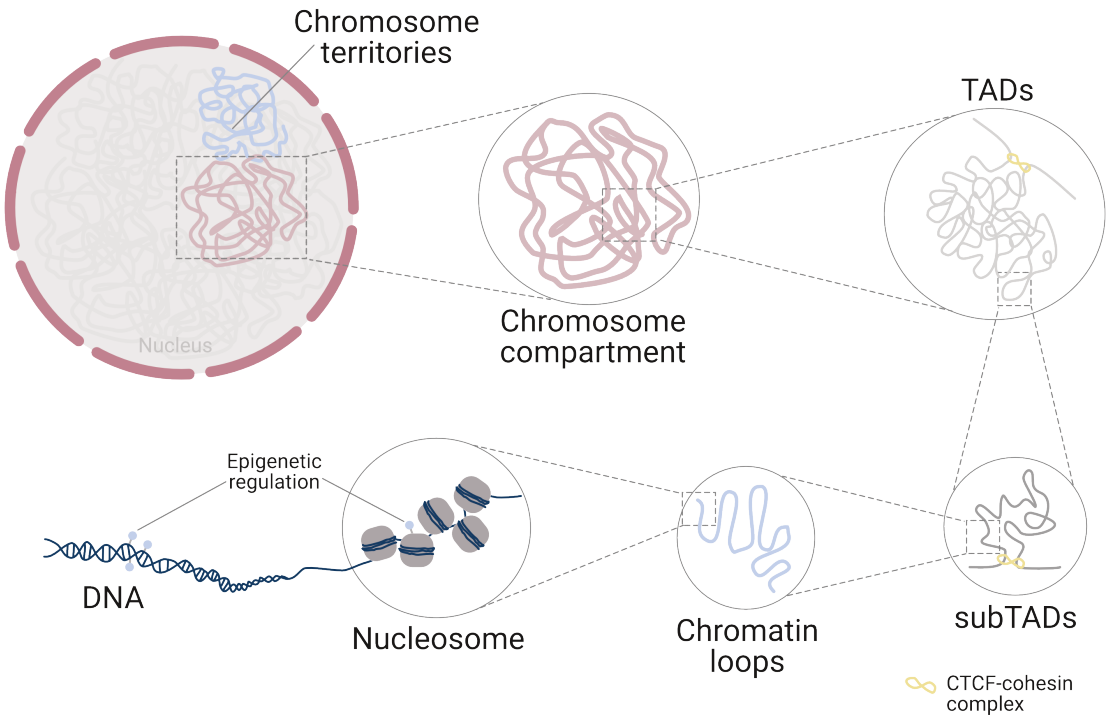


Figure 1.2: 3D genome organization inside the nucleus. Adapted from Wang et al (2021).

RNA polymerase II, retrotransposons, and housekeeping genes are among the features which are typically observed in the vicinity of TAD boundaries (Gan et al. 2019; Hong and Kim 2017). TAD reorganization or removal of TAD boundaries can also lead to medical conditions, such as human limb malformations or neurological disorders (Lupiáñez, Spielmann, and Mundlos 2016).

The chromatin looping is dependent on the concerted action of CCCTC-binding factor (CTCF) and cohesin (Splinter et al. 2006) **(Figure 1.2)**. CTCF interacts with a conserved region of the cohesin subcomplex to stabilise loops (Li et al. 2020). CTCF loss leads to an increase in the interdomain contacts, although it is not present in all TAD boundaries (Zuin et al. 2014; Dixon et al. 2012). Hong & Kim performed a full characterization of TAD boundary-associated elements and observed binding sites from TFs other than CTCF, such as Zinc finger protein (ZNF)143 and Yin Yang (YY)1 (Hong and Kim 2017).

1.2 Epigenome changes in disease

Epigenetic deregulation, as consequence of direct changes into epigenetic modifications or genetic mutations on the epigenetic modifiers, can lead to several diseases **(Figure ??)** (Zoghbi and Beaudet 2016). As an example, fragile X syndrome is linked to abnormal DNA methylation, Kabuki and Sotos syndromes are caused by mutations in genes encoding histone methylation enzymes, and CHARGE syndrome is caused by mutations into the *CHD7* gene, leading to chromatin remodeling defects (Pieretti et al. 1991; Hannibal et al. 2011; Douglas et al. 2003; Bergman et al. 2011). Epigenetics, particularly DNA methylation, has similarly been linked to ageing, as the loss of epigenetic information is considered a cause of ageing in mammals and yeast (Wilson and Jones 1983; Horvath 2013; Yang et al. 2023). Hypermethylation of CGIs upstream of tumour supressor genes has also been observed in several tumours, such as retinoblastoma, kidney cancer, breast cancer, or lung cancer, in processes not always related to mutations (Battagli et al. 2003; Dulaimi et al. 2004; Sakai et al. 1991; Esteller et al. 1999).

1.2.1 The interplay of viral infection with the epigenome

In the cell, host chromatin structure can act as an entry obstacle or an unexpected ally upon infection. In viruses, chromatin can contribute to innate immunity after infection, as viral expression can be repressed through chromatin modulation (Lieberman 2006). However, viruses may also opportunistically appropriate host chromatin and use it to activate viral expression (Zhang and Cao 2019). Non-integrating viruses, like the herpes simplex virus, have been shown to mimic host epigenomic features while establishing their own intra-cellular structures after entry (Kent et al. 2004). On the other hand, integrating viruses (such as HIV-1 or simian immunodeficiency virus) are known to perturb normal chromatin structure (Lieberman 2006).

Inside the host cell, viruses commonly hijack cellular processes to auspiciously replicate and evade immune response (Paschos and Allday 2010). Although research into host chromatin modifiers as viral targets is not very extensive, it is known that histone acetyl-

transferases can be recruited to activate viral expression by viral proteins, as in HIV-1 or adenovirus (Caron, Col, and Khochbin 2003). Some viruses, as the Epstein–Barr virus, appear to lead to an increase in the expression of the repressive proteins of the Polycomb group (Allday 2013). A similar link between HIV-1 and DNA methyltransferases has been hinted before, as DNA methylation levels between HIV-1 infected and uninfected cells present differences in immune-related loci (Mikovits et al. 1998). Host chromatin modulation might play an important role in viral latency as well. In latent HIV-1 integration, low proviral chromatin accessibility and DNA methylation act as mechanisms of viral repression, allowing viruses to evade immune response and retroviral therapy over long periods of time (Jefferys et al. 2021; Blazkova et al. 2009).

Chromatin is known to both influence viral integration targets and to be altered upon integration. In HIV-1, human papillomavirus, and other integrating viruses, multiple epigenomic features have been shown to drive integration site selection (Wang et al. 2007; Singh, Bedwell, and Engelman 2022; Lusic and Siliciano 2017; Marini et al. 2015; Doolittle-Hall et al. 2015; Miklı'k, Šenigl, and Hejnar 2018). On the other hand, after integration, viruses have also shown the ability to alter the host epigenome. The human papillomavirus is known to alter chromatin structure due to alterations in TADs and to the insertion of a new binding site for CTCF (the latter is unpublished proof) (Groves et al. 2021; Karimzadeh et al. 2022). Similarly, HIV-1 integration may also lead to alterations in higher-order chromatin structure (Shah et al. 2022). Even non-integrating viruses, like Epstein–Barr virus, have been shown to trigger 3D chromatin rearrangements (Okabe et al. 2020). Thus, it is likely that the role of the epigenome and chromatin in viral infection surpasses integration, as viruses can directly influence chromatin through the modulation of chromatin modifiers expression and host response is deeply influenced by chromatin features.

HIV-1 infection

Globally, millions of people are infected with Human Immunodeficiency Virus (HIV) every year and it is estimated that approximately 37.7 millions still live with the infection

(WHO 2021). In addition to Acquired Immune Deficiency Syndrome (AIDS), HIV-1 infection is associated to multiple debilitating conditions, such as nephropathies or neurocognitive disorders, and it can similarly increase susceptibility to cancer and to other infections (such as *Mycobacterium tuberculosis* or Hepatitis C virus) (Phillips, Neaton, and Lundgren 2008; Bell and Noursadeghi 2018; Gobran, Ancuta, and Shoukry 2021). Thus, HIV-1 infection remains a major burden for healthcare (WHO 2021). The introduction of antiretroviral therapy (ART) improved the survival rate of HIV-1 infection, but lifelong treatment is still necessary (Deeks et al. 2021).

HIV-1 entry on the host cell relies on co-receptor tropism, as it attaches to a CD4 receptor and a co-receptor, which is either CCR5 or CXCR4 (Clapham and McKnight 2001; John M Coffin and Varmus 1997). Thus, HIV-1 is able to infect primarly CD4+ T lymphocytes, its main target cell, along with cells from the monocyte/macrophage lineage, such as microglia and dendritic cells (John M Coffin and Varmus 1997).

The life cycle of HIV-1 is well documented on CD4+ T lymphocytes and on macrophages. After receptor binding and membrane fusion, the viral RNA is reversely transcribed into double-stranded DNA, trafficked to the nucleus, and integrated into the host chromatin (McLaren and Fellay 2021; Lusic and Siliciano 2017). In productive infection, these steps are followed by viral gene expression, splicing and replication (Lusic and Siliciano 2017). However, a state of latent infection can also be established when the HIV-1 provirus persists within the genome without being immediately transcribed (Mbonye and Karn 2014; John M Coffin and Varmus 1997).

HIV-1 integration and latency

While the use of ART downsized the once fatal impact of HIV-1 infection, eradication is still unachievable due to the existence of cell and tissue reservoirs on a state of reversible nonproductive infection, able to harbour replication-competent virus (Churchill et al. 2016). These constitute latent reservoirs of HIV-1 which can lead to rebound viraemia when treatment is interrupted in patients (Siliciano and Greene 2011).

While latency is traditionally linked to cells harboring transcriptionally inactive viral

genomes, defective proviruses are also able to lead to the production of viral proteins (Chun et al. 1995; Blankson, Persaud, and Siliciano 2002; Imamichi et al. 2020). Essentially, the location of the integration site (IS) is the first factor affecting viral transcription. IS have been frequently found in the introns of actively transcribed genes (Schröder et al. 2002; Wagner et al. 2014). HIV-1 integration favours open chromatin regions and histone modifications associated to it, such as H3 acetylation, H4 acetylation, and H3 or K4 methylation (Wang et al. 2007; Schröder et al. 2002; Scherdin, Rhodes, and Breindl 1990). As many of the genomic features of active transcription units are correlated with each other, it is challenging to identify the main determinant of integration targeting (Craigie and Bushman 2012). On the cellular scale, it has been observed that HIV-1 favours the nuclear periphery for integration and targets regions near speckle-associated genomic domains (Di Primio et al. 2013; Francis et al. 2020).

Long-lived memory CD4+ T cells are the most well-studied cell reservoir for HIV-1, yet the wide range of factors that have an effect in HIV-1 integration and in the latency establishment suggest that this is a complex process with dynamic players that might vary according to cell type and tissue (Dahabieh, Battivelli, and Verdin 2015). Together with blood, it is known that the lymphoid tissue and gut mucosa are important sites for viral replication (Pantaleo et al. 1993; Embretson et al. 1993; Poles et al. 2001). Cells from the bone marrow, liver, testis, and brain have also been found to be infected in patients, although it is not clear if all represent reservoirs (Wout et al. 1998; Carter et al. 2010; Wong and Yukl 2016).

HIV-1 in the central nervous system

In the brain, cells of the macrophage lineage and astrocytes can be infected by HIV-1 (Meulendyke, Croteau, and Zink 2014). While ART is generally effective, the blood-brain barrier offers an treatment-isolated environment for persistent viral replication in the brain (Osborne et al. 2020). HIV-1 infected patients often develop HIV-associated neurocognitive disorders (HAND), inflammatory conditions characterised by cognitive and motor dysfunction (Eggers et al. 2017). HAND has been linked to the neurotoxic

activity of microglia when responding to HIV-1 infection (Branton et al. 2022). Along with microglia, perivascular macrophages, and astrocytes are also known to be targeted by HIV-1 in the brain (Garcia-Mesa et al. 2017; Farhadian et al. 2018). However, only microglia and macrophages are ultimately considered to be the main latency reservoir in the brain.

Microglia are brain-resident macrophages, accounting for 0.5 to 16.6% of the total cell population in the human brain (depending on the region) (Mittelbronn et al. 2001). Microglial cells are long-lived, renew slowly (at a rate of 28% per year), and are known to play important roles in the innate immunity of the central nervous system (CNS) (Réu et al. 2017; H. Liu et al. 2020). Other functions are often attributed to microglia in the CNS, as these cells also play roles into normal brain development and homeostasis (Gosselin et al. 2017). Microglia disregulation contributes to multiple neurodegenerative and psychiatric diseases, such as Alzheimer's, Parkinson's, or schizophrenia (Bachiller et al. 2018; Gosselin et al. 2017). In the context of infection, microglia become highly activated through the upregulation of diverse cytokine and chemokine pathways (Colonna and Butovsky 2017). While their fundamental role in the defense of the CNS has been demonstrated in early stages of viral infection, microglial loss of function and chronic inflammation has been considered to be the cause of neuropathogenesis in the brain of HIV-1 infected patients (Ginsberg et al. 2018; Branton et al. 2022).

Chromatin and viral integration

Location features of IS are considered a defining factor to the persistence of HIV-1 on infected cells, specially when chromatin state comes into play (Maldarelli et al. 2014). The impact of the host chromatin on the integration of HIV-1 is widely documented (De Crignis and Mahmoudi 2017; Battivelli et al. 2018b; Vansant et al. 2020; Lelek et al. 2015). HIV-1 integration is influenced by nuclear architecture, genomic sequence, cell phase, chromatin structure, route of nuclear entry, among other factors (Lusic and Siliciano 2017; Marini et al. 2015; G. J. Bedwell and Engelman 2021). Moreover, research on IS selection concluded that integration is more frequent on open chromatin regions,

transcriptionally active, and neighboring enhancers, super-enhancers (SE), and nuclear speckle-associated genomic domains, within highly transcribed genes, on high GC content regions, and regions with high CpG island density (Francis et al. 2020; Lucic et al. 2019; Wang et al. 2007; Schröder et al. 2002; Maldarelli et al. 2014; Brady et al. 2009).

Viral integration seems to be influenced by histone modifications and other epigenetic players. A large-scale study on 40,569 unique IS found that locations with H3 acetylation, H4 acetylation, and H3 K4 methylation, typically characterizing acessible chromatin are targeted more often than the rest of the genome (Wang et al. 2007). Additionally, HIV-1 integration is frequently found in H3K36me3-enriched regions (Vansant et al. 2020). H3K36me3 is quite relevant in HIV-1 research, as it is associated to LEDGF/p75, an epigenetic reader for this histone modification, and its role as an host factor for HIV integrase (Cherepanov et al. 2003; Vansant et al. 2020). On the other hand, a negative association of IS targeting with repressive modifications H3 K27 trimethylation and DNA methylation was observed (Wang et al. 2007; Blazkova et al. 2009). However, there are still HIV-1 proviruses associated to transcriptional repression, possibly leading to post-integration latency (Debyser et al. 2018).

1.2.2 Epigenomic changes and cancer

Many studies have highlighted the central role of epigenetics in tumourigenesis (Michalak et al. 2019). Cancer development is linked to the abnormal activation of oncogenes or inactivation of tumour suppressor genes (Lee and Muller 2010). The transcriptional changes induced at these loci by changes in DNA methylation or histone modifications can often be attributed to mutations, but epigenetic origin is also deemed possible (Shanmugam et al. 2018).

Many of the epigenetic alterations observed in tumours result from mutations in genes encoding epigenetic enzymes, such as the ones encoding histone demethylases or DNA methyltransferases (Plass et al. 2013). *IDH1* or *DNMT3A* mutations are very common and lead to DNA methylation alterations (Tatton-Brown et al. 2014; Turcan et al. 2012). Mutations in *EZH2*, encoding a histone methyltransferase and component of the

PRC2, have been known to lead to H3K27me3 alterations in different cancer types, such as B-cell lymphomas, acute myeloid leukaemia (AML), or melanoma (Sneeringer et al. 2010; Stasik et al. 2020; Han et al. 2019). Mutations in many of these genes have also been consistently linked to proliferation, migration, survival, and other clinical features, denoting the importance of epigenetics into the presentation of the tumours and patient outcomes (Han et al. 2019).

DNA methylation is frequently altered in cancer, as both global DNA hypomethylation and CpG island hypermethylation (CpG island methylator phenotype (CIMP)) occur often upon tumourigenesis **(Figure 1.3)** (Nishiyama and Nakanishi 2021). Locus-specific aberrant methylation of CGIs has been observed in glioblastoma (GBM), AML, lung cancer, colorectal cancer, among others (Turcan et al. 2012; Costello et al. 2000; Baylin et al. 1986; Rijnsoever et al. 2002; Toyota et al. 2001; Yates and Boeva 2022). Aberrant DNA methylation has also been deemed as a cause for the transcriptional disregulation of tumour suppressor genes and oncogenes (Nishiyama and Nakanishi 2021; Esteller et al. 1999). Other genes, like the Homeobox genes, whose disregulation is often linked to tumourigenesis, can become activated through DNA methylation disruption (Su et al. 2018). Locations in the normally unmethylated *methylation canyons* have been likewise found to be preferentially hypermethylated in cancer (Jeong et al. 2014; Xie et al. 2013). *Canyons* typically encompass developmental regulators repressed by H3K27me3, and it has been suggested that stably bound DNA methylation replaces the more dynamic H3K27me3 in order to retain gene inactivation in these loci (Nishiyama and Nakanishi 2021).

Lastly, the alteration of regulatory regions in tumours can disrupt TF-binding sites and contact domains, changing the higher-order chromatin structure (Jia et al. 2020; Wang et al. 2022). Altered chromatin looping can often be found around multiple oncogenes, possibly holding an important role in their activation (Ahn et al. 2021). "Enhancer hijacking", a result of chromatin rearrangements, has been described in multiple cancers as a driver of aberrant oncogene expression (Northcott et al. 2014; Helmsauer et al. 2020). It has been suggested that mutational status could influence chromatin struc-

ture, as is the case for isocitrate dehydrogenase (IDH)-mutated tumours (Flavahan et al. 2016). In GBM, aberrant chromatin structure has been found in tumours harbouring *EGFR* amplification (common in the RTK-II subtype) (Yang et al. 2022).

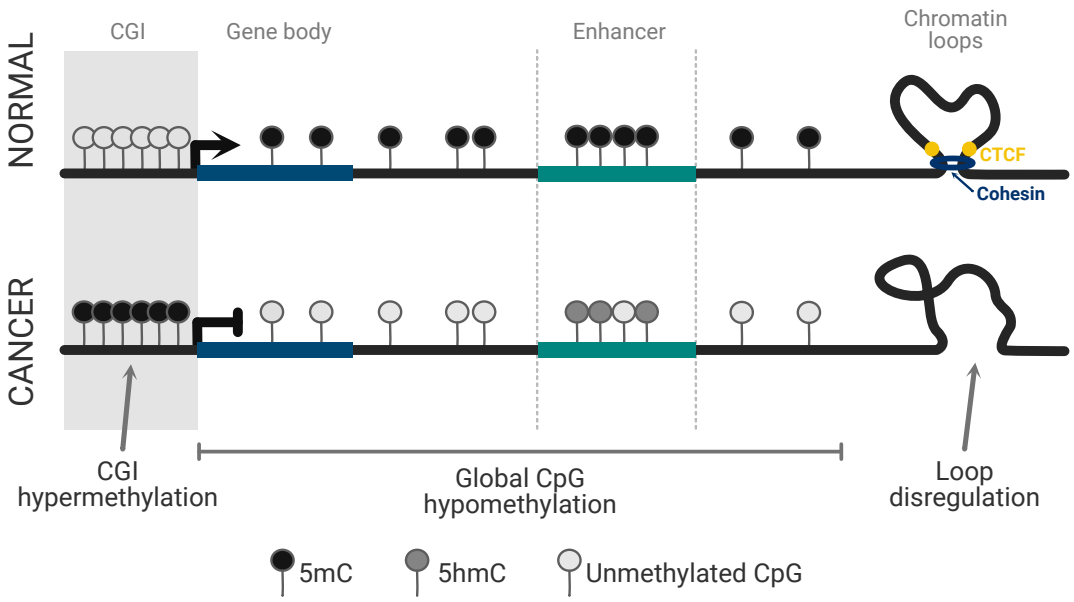


Figure 1.3: Effects of cancer on the transcription regulation of genes and chromatin structure. CGI hypermethylation and global CpG hypomethylation is very common in tumour cells, as DNA methylation regulators become impaired upon tumourigenesis. Adapted from Michalak et al. (2019).

CpG island methylator phenotype

CIMP was first documented on colorectal cancer, where it promoted the inactivation of tumour suppressor genes *CDKN2A* (encoding the p16 protein) and *THBS1* (Toyota et al. 1999). This effect implied that CIMP acted as a cancer-inducing mechanism, as it was later observed in other tumour types (Dulaimi et al. 2004; Battagli et al. 2003). Nevertheless, it has also been shown that the genes targeted by CIMP are oftentimes already repressed in the tissue where the phenotype originates (Sproul et al. 2012). CIMP has similarly been observed in lowly expressed genes or genes in a *bivalent* state, harbouring both active H3K4 methylation and the repressive mark H3K27me3 (Teodoridis, Hardie, and Brown 2008; Ohm et al. 2007).

Systematic pan-cancer analyses have suggested that CIMP is present in many cancer types other than colorectal cancer, such as sarcoma, adrenocortical carcinoma, GBM, kidney carcinoma, lung adenocarcinoma, among others (Moarii, Reyal, and Vert 2015; Yates and Boeva 2022). In some tumour types, CIMP has been attributed to mutations in *IDH1* and *SETD2*, while in colorectal cancer it has been associated to microsatellite instability, *KRAS* mutation, and *BRAF* V600E activating mutation (Yates and Boeva 2022; Weisenberger et al. 2006; Ogino et al. 2006). Recently, *BRAF* V600E activation has been connected to ageing, in line with the conception that CIMP is mostly driven by age-related methylation (Tao et al. 2019; Christensen et al. 2009). In AML, CIMP can be traced to *DNMT3A* mutations but it is considered a consequence of tumour progression and cellular proliferation (Spencer et al. 2017). Recently, Yates and Boeva have classified four possible origins for CIMP: (i) caused by mutations in genes associated to DNA demethylation, (ii) caused by mutations in genes not associated to the maintenance of DNA methylation, (iii) caused by mutations in histone methyltransferases, (iv) or derived from microsatellite instability (Yates and Boeva 2022).

The hypermethylation associated to tumourigenesis can be either due to *de novo* methylation events, disregulation of the TET-dependent demethylation, or to the abnormal increase of already existing methylation. Functionally, it has been suggested that hypermethylation at CGIs could be due to an accumulation of DNA methyltransferases driven by DNA damage. While gene repression done upon DNA damage is usually transitory, some lowly expressed genes retain promoter hypermethylation despite the subsequent DNA repair (Nishiyama and Nakanishi 2021). Unsurprisingly, the increase in DNA methylation has also been attributed to higher expression of DNA methyltransferases (Teodoridis, Hardie, and Brown 2008). This increase has been subsequently linked to other factors, such as *BRAF* or *KRAS* mutational status, single nucleotide polymorphisms (in the promoter of *DNMT3B6* for example), and even infection (Teodoridis, Hardie, and Brown 2008; Chang et al. 2006). On the other hand, disregulation of the epigenetic regulators from the TET family is also seen as a possible CIMP-inducing mechanism (Tulstrup et al. 2021). TET enzymes hold an important role on 5hmC production,

a major form of DNA demethylation, and production of 5hmC by TET becomes affected through the *IDH1/2* mutations found in many tumour types (Figueroa et al. 2010). The product of non-mutated *IDH* is the isocitrate dehydrogenase, involved in glucose metabolism as part of the Krebs cycle (Ye et al. 2013). When mutated in tumours, *IDH* produces 2-hydroxylglutarate instead of α-ketoglutarate, leading to a deleterious accumulation of 2-hydroxylglutarate. In turn, hypermethylation arises as a result of the competitive action of 2-hydroxylglutarate over two histone demethylases and the DNA demethylase TET2, both α-ketoglutarate-dependent (Lu et al. 2012; Ye et al. 2013). 2-hydroxylglutarate has also been found to block cell differentiation, evidencing its role as a oncometabolite (Losman et al. 2013).

Glioblastoma

GBM is the most lethal and common type of primary brain tumour in adults, currently assigned as a grade IV brain tumor based on its histological and molecular features (Wirsching, Galanis, and Weller 2016; Louis et al. 2021). GBM is further divided into 4 distinct subtypes: IDH (characterised by presenting a CpG island methylator phenotype), MES (or *mesenchymal*), RTK-I (Receptor tyrosine kinase (RTK), previously named *proneural*), and RTK-II (previously named *classical*) **(Table 1.1)** (Wang et al. 2017; Verhaak et al. 2010; Sturm et al. 2012). The distinction between IDH-mutant and IDH-wild type tumors is also frequently made, as IDH-mutant GBM evolves from IDH-mutant astrocytoma[2]. The subtypes present distinct epigenetic characteristics which are frequently linked to different outcomes, survival rates, and treatment options (Filbin and Suvà 2016). Moreover, subtype transitions have been observed in patients (Phillips et al. 2006).

GBM is characterised by an extremely heterogeneous genetic landscape, both inter-

[2] The recently published 2021 WHO Classification of Tumors of the CNS eliminates the term "Glioblastoma IDH-mutant" and replaces it with "Astrocytoma, IDH-mutant". As this work preceded the latter change in nomenclature, I will refer to "Astrocytoma, IDH-mutant" using the former "Glioblastoma IDH-mutant" nomenclature

tumorally and intratumorally (Filbin and Suvà 2016). These differences often lead to clinical variability. For example, IDH mutations which are present in about 10% of the cases and mostly in younger patients, usually indicate a more favorable outcome (Kleihues and Ohgaki 1999; Noushmehr et al. 2010). Mutations in *TERT* (composing one of the units of telomerase), are common in about 80% of GBM tumours and are linked to abnormal cell proliferation caused by *TERT* activation (Filbin and Suvà 2016).

Table 1.1: GBM subtypes and correspondent genetic features. Source: Verhaak et al (2010) and Wu et al (2020).

Subtype	Common mutations	Chromosomal aberrations	Highly affected pathways
IDH	IDH1/IDH2 mutation		
MES	NF1 mutations	Chromosome 7 gain and chromosome 10 loss	Akt signaling pathway
RTK-I	PDGFRA gene amplification and TP53 mutations	Chromosome 10 loss	
RTK-II	EGFR gene amplification	Chromosomes 7 and 19 gain, along with chromosome 10 loss	Retinoblastoma pathway

Epigenetic alterations in glioblastoma

Epigenetic heterogeneity is a feature of GBM, accompanying the genetic and transcriptional heterogeneity associated to these tumours (Klughammer et al. 2018). Over the years, several studies approached the epigenetic alterations associated to glioblastoma (Filbin and Suvà 2016). These alterations have been used for pharmaceutical research, as epigenetic modulation can be targeted as a GBM therapy and is often linked to clinical features like tumour progression and survival (Phillips et al. 2006). These therapies include the use of histone deacetylase or DNA methyltransferase inhibitors (Uddin et al. 2022).

DNA methylation aberrations are common and often found to be responsible for the inactivation of multiple tumour suppressor genes in GBM, such as *NDRG2*, *CDKN2A*,

KLF4, among others (Uddin et al. 2022). CIMP is one of the most well-documented DNA methylation aberrations. In GBM, it has been described in the IDH subtype and it is indirectly caused by *IDH1* or *IDH2* mutations (Noushmehr et al. 2010; Turcan et al. 2012; Figueroa et al. 2010).

Histone modificators, like histone deacetylases or histone methyltransferases, have been found altered in GBM cell lines (Uddin et al. 2022). Abnormally high expression of the H3K27me3-mediator *EZH2* has similarly been observed in GBM, being connected with tumour features like metastasis or progression (Orzan et al. 2011). Lastly, abnormal chromatin structure has been observed to contribute to changes in gene expression and clinical features of GBM, as chromatin state has been associated to drug tolerance and persistence (Liau et al. 2017).

1.3 Sequencing approaches to chromatin research

Recently, the advent of next-generation sequencing allowed for the development of high-throughput methods to profile epigenomic mechanisms **(Figure 1.4)**. Examples of these methods are ChIP-seq (Chromatin immunoprecipitation sequencing), WGBS (Whole-genome bisulfite sequencing), RRBS (Reduced representation bisulfite sequencing), MeDIP-seq (Methylated DNA immunoprecipitation sequencing), ATAC-seq (Assay for Transposase-accessible chromatin sequencing), among others **(Table 1.2)** (Gu et al. 2011; Park 2009; Cazaly et al. 2019; Buenrostro et al. 2015). ChIP-seq has been a widely adopted assay to assess TF binding and histone modification enrichments (Park 2009). Similarly, WGBS and RRBS have provided many advances in the study of DNA methylation (Gu et al. 2011; Lister et al. 2009). These methods have allowed for vast advances into epigenome-wide association studies and into the research of epigenetics in disease and development.

Chromatin segmentation using Hidden Markov models (HMM) helped to make the combinations of different modifications more interpretable (Ernst and Kellis 2012). Now, the evolution of multiomics allows the combination of epigenomic with transcriptomic, proteomic, or genomic data, providing valuable knowledge into the full understanding of cell machinery.

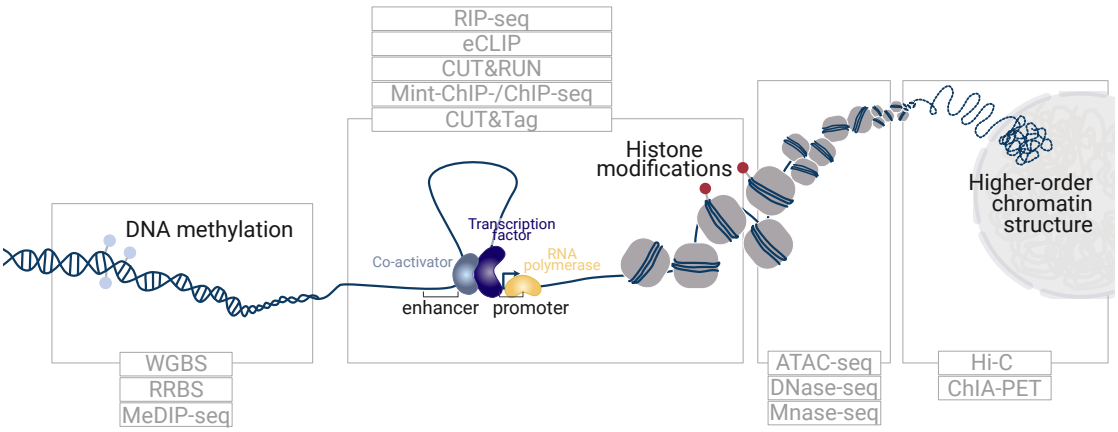


Figure 1.4: Most advanced sequencing methods applied to epigenomic research. Methods are labelled in the boxes.

Table 1.2: Commonly used sequencing assays for epigenomics, grouped by respective targets.

Assay	Method	Features
DNA methylation		
WGBS	Treatment with sodium bisulfite	Highest coverage and unbiased
RRBS	Restriction enzymes combined with bisulfite treatment	More coverage at promoters and CGIs than single CpGs
MeDIP-seq	Treatment with anti-5mC antibodies and DNA purification	Fragment-based and largely qualitative
Histones/TFs		
ChIP-seq	Crosslinking of DNA-protein complexes, fragmentation, immunoprecipitation	Lower signal-to-noise ratio than others and requiring extensive optimisation
Mint-ChIP-seq	Same as ChIP-seq but multiplexed and indexed	Better for low input samples
RIP-seq	Immunoprecipitation of RNA-protein complexes and reverse transcription to cDNA	More accurate results than eCLIP
eCLIP	UV-crosslinking of RNA-protein complexes, immunoprecipitation, and reverse transcription to cDNA	Only crosslinked RNAs are used as input
CUT&Tag	Treatment with Tn5 transposase, and simultaneous fragmentation and adapter insertion	Less input material needed and high throughput
CUT&RUN	Treatment with pAG-MNase, DNA cleavage, and extraction	Less input material needed, high throughput, but more prone to errors than CUT&Tag
Chromatin accessibility		
ATAC-seq	Treatment with transposase and fragmentation of open chromatin sites	Simple to setup and requiring less input material
DNAse-seq	DNase l treatment for DNA-protein complexes, followed by DNA extraction	Large number of cells are needed and more sensitive at promoters
Mnase-seq	MNase treatment for DNA-protein complexes, followed by DNA extraction	Large number of cells are needed
Interactions		
ChIA-PET	Crosslinking of DNA-protein complexes, fragmentation, immunoprecipitation, and proximity-based ligation	More protein-biased than Hi-C
Hi-C	Crosslinking of DNA-protein complexes, fragmentation, and extraction	Unbiased genome-wide coverage

1.4 Computational methods and methodological concepts

In parallel with the recent developments in genome-wide sequencing, many computational methodologies have been applied to biological research to better understand cell biology. Data integration currently allows for an approximation of systems-level knowledge on the genomic, transcriptomic, epigenomic, and proteomic layers of cell function. The recent emergence of single-cell techniques led to the a superior understanding of the interplay between cells, helping understand gene regulation at a higher resolution (Stuart et al. 2019).

Over the recent years, machine learning has been increasingly applied in biology and medicine to model complex biological systems. Machine learning can usually be divided into four main categories: supervised, unsupervised, semi-supervised, and reinforcement learning (Sarker 2021). *Supervised learning* is "task-driven" and comprises all techniques which rely on a defined labelled input and output (Sarker 2021). Classification tasks are very often performed through supervised learning. As an example, a classifier used for tuberculosis diagnosis and trained using X-rays obtained from patients with tuberculosis and healthy individuals would be an example of a supervised learning method. On the other hand, *unsupervised learning* is applied in the analysis of unlabelled data (Sarker 2021). It is very often used in clustering or dimensionality reduction as it can be useful for the identification of meaningful distinguishable features in the input data. *Semi-supervised learning* is defined as an interfusion of the previous two approaches, as it is used on both labelled and unlabelled data (Sarker 2021). It was created to overcome the lack of labelled data and it is used in fraud detection or text classification (Sarker 2021). Lastly, *reinforcement learning* allows for the machine to define its own optimal performance through a *reward/punishment* approach (Kaelbling, Littman, and Moore 1996). It is mostly used in robotics or automation (Sarker 2021).

In this section, I will only focus on the machine learning techniques applied in the methodology of this work. In both projects, we have applied random forest (RF) in feature selection and stratification, non-negative matrix factorization (NMF) in feature

selection and dimensionality reduction, and bayesian networks in the prediction of mutual influences between epigenomic players.

1.4.1 Non-negative matrix factorization

In recent years, the combination between the inherent complexity of biology and the amount of data being produced by high-throughput sequencing required the development of dimensionality reduction stategies (Eckmann and Tlusty 2021). While principal component analysis (PCA) is useful for certain tasks, it can be limiting to apply it to high-dimensional data (Pearson 1901). Other linear and non-linear techniques can be used for dimensionality reduction, such as linear discriminant analysis, NMF, autoencoders, or uniform manifold approximation and projection. These methods have thus been applied to provide a natural mathematical simplification of the biological system (Eckmann and Tlusty 2021).

NMF is a widely used unsupervised method of dimensionality reduction and feature extraction, which has been applied in image analysis, text classification, artificial intelligence, signal processing, among others (Lee and Seung 1999; Lin and Boutros 2020). Unlike PCA, NMF learns a parts-based data representation of the initial data (Lee and Seung 1999). NMF aims to find approximate k factorizations so that **(Equation 1.1)**:

$$V \approx WH \tag{1.1}$$

where $V \in \mathbb{R}^{n \times m}$ is the input (data) matrix, $W \in \mathbb{R}^{n \times k}$ is a signature matrix, and $H \in \mathbb{R}^{m \times k}$ is an exposure matrix **Figure 1.5**.

The factors k are usually selected so that $(n+m)k < nm$ (Lee and Seung 1999). NMF is influenced by its nonnegativity, a constraint on matrices W and H which allows only additive combinations and makes the results more easily interpretable (Lee and Seung 1999). In biology, NMF has been used to define molecular signatures from expression profiles or *de novo* identification of copy number signatures (Devarajan 2008; Gartlgruber et al. 2021; Steele et al. 2022). Recently, integrative NMF also emerged, allowing the integration of multiple omics datasets (Gao et al. 2021). Examples of the structure of

V, W, and H in biological systems are shown in **Figure 1.5**.

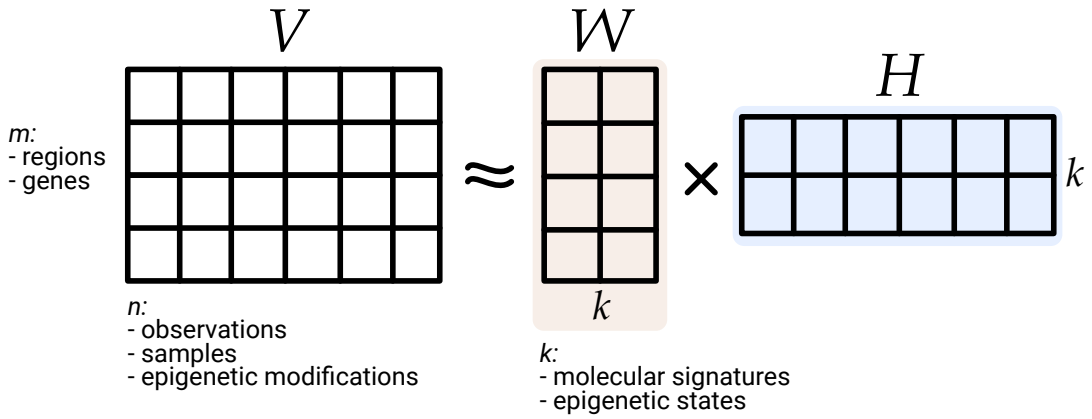


Figure 1.5: Basic concept of NMF and applications of NMF in biology. Nonnegative V matrix is factored into matrices W and H. Possible representations of the V, W and H matrixes in NMF are labelled.

1.4.2 Random forest

Unlike regression tasks, which assume that the prediction (or response variable) is quantitative, in classification tasks the prediction is a categorical variable. Classification of an observation into a category can be achieved through multiple classification techniques. Some commonly used algorithms are logistic regression, linear discriminant analysis, decision trees, or k-nearest neighbors (James et al. 2014). Often, multiple individual models can be combined in order to enhance global predictions, in what is defined as ensemble learning (Hastie, Tibshirani, and Friedman 2001).

RF is an parallel ensemble supervised algorithm which combines independently generated decision trees in random subspaces (Hastie, Tibshirani, and Friedman 2001). First proposed in 1995 by Ho and developed further by Breiman, RF can be used in both regression and classification tasks, much like its base model (Breiman 2001; Ho 1998, 1995). This model, the decision tree, is highly interpretable, fast to execute and accurate, but its use is limited in larger tasks (Quinlan 1986). Each tree has high variance

and low bias, but RF builds on the simple decision tree models and profits from their advantages using *bagging* or *bootstrap* aggregation, as multiple trees are bootstrapped on different training samples to achieve lower variance on the estimated predictions (James et al. 2014). The predictors (p) and observations are both used for training. Yet, the predictor subset in a tree, considered at every split, is similarly a random sample of p (usually $\sqrt{p}$), ensuring the trees are *decorrelated* (James et al. 2014).

While in regression, bagging implies the resulting predictions are averaged, in classification tasks, the *majority vote* approach applies. This means that the final prediction is the most frequent one **(Figure 1.6)**. This way, this method overcomes over-fitting and leads to an increased performance (Sarker 2021).

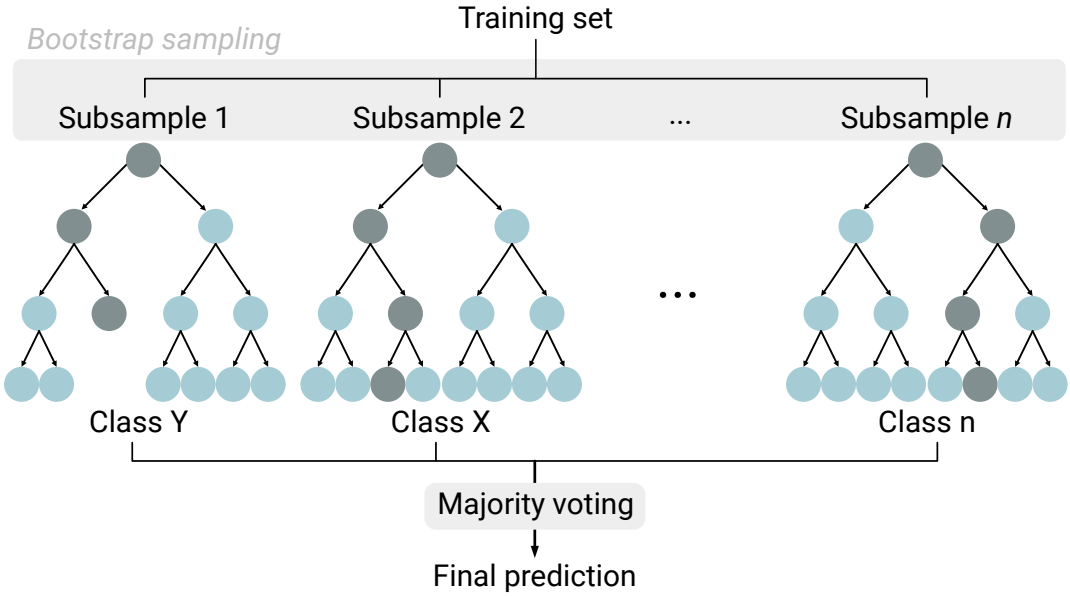


Figure 1.6: RF model diagram. Each individual bootstrap sampled decision tree outputs one prediction, which is later combined with the predicitions obtained on all the decision trees. The final prediction is the most overly represented one, decided through majority voting.

1.4.3 Bayesian networks

In biology, processes are oftentimes represented as graphs. Probabilistic graphical models can be easily applied in biology, as these allow the representation of the interactions

between variables, easily comparable to a cause-effect relationship (Su et al. 2013). Bayesian networks are one of the most commonly used graphical models in biology.

Bayesian networks are probabilistical graphical models whose structure is characterised by an underlying directed acyclic graph (Ni et al. 2018). In the bayesian network, each node (V) represents continuous or discrete random variables ($V = \{X_1, X_2, \ldots, X_{\mathrm{v}}\}$), while edges ($A$) represent the probabilistic dependencies between them. Each node is associated with a conditional probability distribution **(Equation 1.2)**, given by its parent nodes. Such that the joint distribution is given by:

$$P(X_1, \ldots, X_v) = \prod_{i=1}^{v} P(X_i | parents(X_i)) \tag{1.2}$$

Bayesian networks represent joint distributions graphically. In **Figure 1.7**, the chain rule of probability for the graph **(Equation 1.3)** would be:

$$P(FH, S, LC) = P(FH)P(S)P(LC|FH, S) \tag{1.3}$$

allowing an estimation of the probability that a certain patient will develop lung cancer. Independence can be deduced from the graph representation. In this example, family history and smoking status are independent, even though they both link to a probability of developing lung cancer. These networks can represent both linear and non-linear,

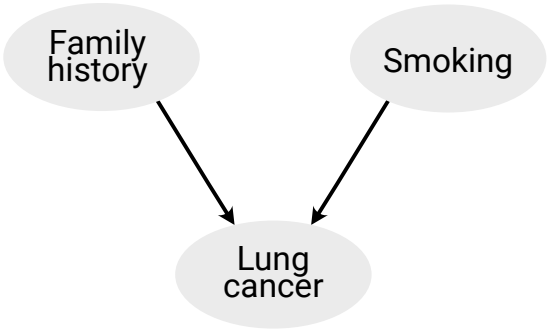


Figure 1.7: Simple Bayesian network example.

along with combinatorial and stochastic relationships, making them useful for the study of biological systems. Additionally, they can handle noisy data. In biology, bayesian

networks may represent molecules, epigenetic modifications, or genes. These models have been applied in gene regulatory networks, predictions of protein-protein interactions, identification of cancer driver events, among others (Yu et al. 2004; Su et al. 2013; Angelopoulos et al. 2022).

Chapter 2

Epigenomics of HIV-1 integration in microglial cell model hints on viral-driven changes in 3D genome structure

The results found in this section have been published in Cell Reports (Rheinberger et al. 2023) under the Creative Commons CC-BY-NC-ND license and are reproduced here in accordance with the rights of open-access publishing from Elsevier.

2.1 Motivation

The integration of HIV-1 in the human genome leads to several effects on the host cell and it is the key-event behind viral latency establishment and persistence (Chun et al. 1995; Siliciano and Greene 2011). Integration is very often studied on the resting memory CD4+ T cells, the main HIV-1 cell reservoirs (Finzi et al. 1999). However, other cell types have been shown to be important HIV-1 reservoirs, such as epithelial cells,

dendritic cells, or tissue-resident macrophages, like Kupffer cells in the liver or microglia in the brain. HIV-1 enters the CNS within the first 2 weeks of infection (Valcour et al. 2012). In the brain, it infects microglia, astrocytes, and perivascular macrophages, although only macrophages and microglia are considered to hold potential to become proviral reservoirs (Sreeram et al. 2022). HIV-1 infection is known to significantly affect the CNS, as neurocognitive disorders associated to HIV-1 impact the lives of patients undergoing ART (Eggers et al. 2017). ART is not very effective in the brain, possibly due to the existence of the blood–brain barrier, and it has been suggested that these neurocognitive disorders arise from the the neurotoxic and inflammatory activity of microglia when actively infected (H. Liu et al. 2020). Microglia also offers a potential latent reservoir for HIV-1 in the brain, as it is a long-lived cell and might allow productive HIV-1 replication after activation of the proviral promoter (Sreeram et al. 2022).

After HIV-1 infection, it is known that certain regions are more likely to lead to post-integration latency of the integrated provirus, evidencing that IS selection is directly related to the subsequent proviral state (Debyser et al. 2018). In CD4+ T cells, HIV-1 favours introns of actively transcribed genes, near enhancers, super-enhancers (SE), and nuclear speckle-associated genomic domains (Schröder et al. 2002; Wang et al. 2007; Lucic et al. 2019; Francis et al. 2020). High GC content regions, and regions with high CpG island density also tend to be targeted often (Brady et al. 2009). Linker mediated (LM)-PCR has been developed to determine IS in HIV-1 targeted cells (Serrao, Cherepanov, and Engelman 2016). In this work, we aimed to understand which regions are targeted by HIV-1 in the microglial cell and to compare targeted regions to the ones observed in other cell targets for HIV-1, like other macrophages and CD4+ T cells. We have used LM-PCR obtained on an infected microglial cell model (C20), given the limited access to brain tissue of HIV-1 patients, which can only occur postmortem. C20 is a human microglia cell line, derived from adult brain tissue, which was immortalized using a combination of SV40 T antigen and human telomerase reverse transcriptase (Garcia-Mesa et al. 2017).

Viral integration is known to affect the host transcriptional programmes and lead to

cellular proliferation, to the production of virus-host chimeric RNA, activation of criptic splice sites, or promoter/enhancer insertions, as observed in other integration-capable virus (R. Liu et al. 2020; Yoon et al. 2020; Mellors et al. 2021; Cesana et al. 2017; Linden and Jones 2022). In some, such as human papillomavirus, human leukemia virus, or human T-lymphotropic virus 1, effects of the viral integration on the chromatin have been reported (Melamed et al. 2018, 2022; Groves et al. 2021; Satou et al. 2016). In HIV-1, it was also observed that target regions are linked to specific histone modifications, like H3K36me3 and H3K27ac, hinting on chromatin landscape as an important factor for integration permissibility (Albanese et al. 2008; G. J. Bedwell et al. 2021; Singh, Bedwell, and Engelman 2022; Vansant et al. 2020; Wang et al. 2007). Thus, we aimed to determine how the transcriptional and epigenetic landscape in the microglial cell act as integration determinants. We used RNA-seq and ChIP-seq data obtained from histone modifications linked to both heterochromatin and euchromatin to assess the transcriptional state of the healthy cell before infection. Next, we aimed to understand the impact of different proviral states in the chromatin accessibility of microglia after HIV-1 infection, to find if the integration leads to chromatin alterations and whether these alterations depend on the HIV-1 state. Thus, we have used ATAC-seq data obtained on two cell populations according to the proviral state (active or latent) and compared these with the uninfected population.

In brief, in this chapter I document the genomic profiling of viral integration on a microglia cell model, compare it with other cell targets of HIV-1, and assess both the effect of the chromatin upon viral integration and the effect of viral integration on the chromatin. I decomposed epigenomic data into two scales of integration permissibility signatures and generated a model of integration permissibility for this cell type using random forest classification. Next, we associated IS with higher-order chromatin structures and I identified TF linked to these structures, particularly CTCF, as altered on distinct HIV-1 infection states. Lastly, we assessed the links between the HIV-1 integration and units of 3D nuclear organization.

2.2 Data

The two main aims were to assess which genetic and epigenetic features influence HIV-1 integration and how chromatin is affected by integration in microglia. To understand this, next-generation sequencing data was produced for both uninfected and infected C20 microglial cells **(Figure 2.1)**.

All the sequencing data used on this work is fully described into **Appendix A** and was generated either by the Lusic lab (CIID, Heidelberg) or obtained from public datasets.

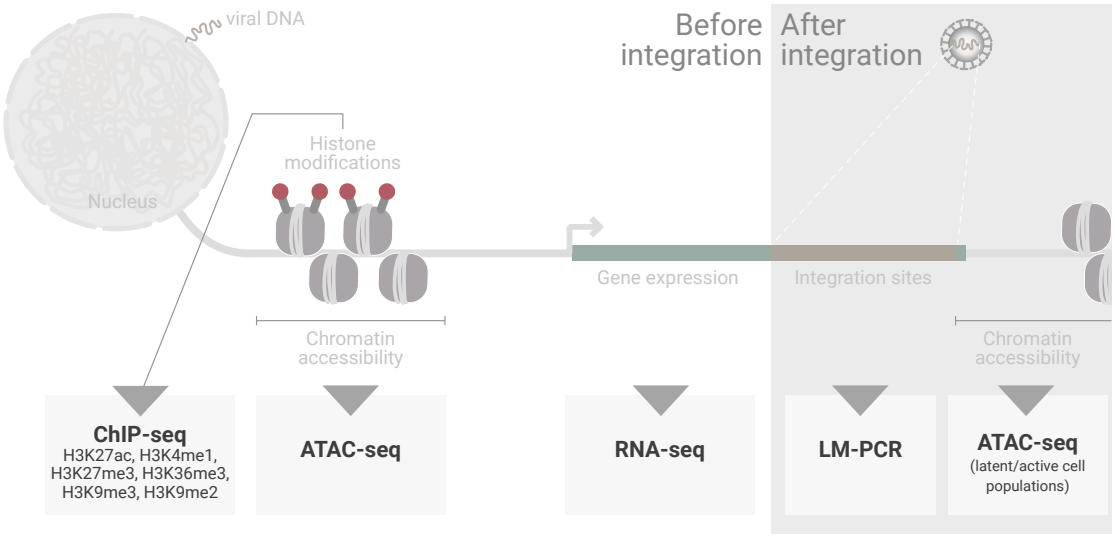


Figure 2.1: Diagram of the C20-derived data used in this work. Sequencing assays are included in boxes. Before integration, ChIP-seq, ATAC-seq, and RNA-seq were obtained on the uninfected cells. After infection, IS were determined using LM-PCR and ATAC-seq was applied to assess the chromatin accessibility on cell populations sorted according to HIV-1 status (in active or latent transcription).

2.2.1 Microglia (inhouse datasets)

To determine HIV-1 IS in the C20 microglial cell line, LM-PCR was applied on infected cells. LM-PCR is a next-generation sequencing method used for amplification and sequencing of the retroviral IS in the host genome (Serrao, Cherepanov, and Engelman 2016). In this method, sequencing reads include virus-host junctions, identifiable through

the long terminal repeat (LTR), which allow the determination of HIV-targeted loci after alignment to the human genome. The LTR is a repeat region found at each terminal end of the provirus. We have used both single-end and paired-end libraries.

We have also used LM-PCR data on infected human induced pluripotent stem cell (iPSC)-derived microglia to compare the C20 insertion profiles with another cell model used in microglia studies.

To assess the epigenetic determinants of HIV-1 integration and to build a chromatin state model for the C20 cell line, ChIP-seq on histone modifications H3K27ac, H3K36me3, H3K4me1, H3K9me2/3, and H3K27me3 was obtained from uninfected C20 cells. RNA-seq was also performed on uninfected cells to assess the impact of transcription level into the likehood of integration.

To understand the impact of the two infection states (active and latent) into the chromatin, C20 microglial cells were sequenced through ATAC-seq after infection with a HIV Green Kousubira Orange reporter virus (Jefferys et al. 2021). This dual-labeled virus allows the distinction of the proviral status, as it includes the fluorescent protein eGFP under the control of the HIV-1 promoter and the fluorescent protein mKO2 under the control of a housekeeping gene (Battivelli et al. 2018a). This system allows the sorting of cells into three populations using fluorescence-activated cell sorting: uninfected (eGFP-mKO2-), active infection (eGFP+mKO2+), and latent infection (eGFP-mKO2+). These three cell populations were sequenced through ATAC-seq.

Lastly, to understand the impact of CTCF, an important TF for chromatin structure, on the HIV-1 insertion patterns, we have analysed LM-PCR libraries obtained on both wild type (WT) and CTCF knock-down (KD). In parallel, CTCF ChIP-seq was generated for both of these conditions.

2.2.2 Public datasets

To compare the microglia-derived IS with other cell types, I have used publicly available IS datasets from CD4+ T cells and MDMs (Kok et al. 2016; Lucic et al. 2019). Moreover, to compare microglia with other HIV-1 targets regarding epigenetic and transcriptomic

features, I have used publicly available ChIP-seq data from histone modifications and RNA-seq on CD4+ T cells (ENCODE Project Consortium 2012; Lucic et al. 2019). To compare C20 with primary microglia cells, I have used published proximity ligation-assisted ChIP-seq on chromatin contacts, ATAC-seq, RNA-seq, and ChIP-seq on histone modifications obtained from primary microglia (Gosselin et al. 2017; Nott et al. 2019).

In order to discover TFs with an important role in TAD boundaries, we have used footprint-derived TFBS (using ATAC-seq) and TAD boundaries from 9 different cells and tissues. ATAC-seq files (in BAM format) used for the TFBS footprinting applied as input for the TAD boundaries RF model were obtained from ENCODE (ENCODE Project Consortium 2012). TAD boundaries used as ground truth for the RF model training (hg38 reference genome assembly) (Wang et al. 2018; Dixon et al. 2012).

In order to assess the connection between IS and TAD boundaries in microglia, we have used Neu- TAD boundaries (Hu et al. 2021). The Neu- population includes non-neuronal cells in the brain, such as oligodendrocytes, microglia, and astrocytes (Hu et al. 2021). This data was generated as part of the PsychENCODE Consortium (accession number *syn4921369*).

2.3 Methodology

IS discovery pipeline from LM-PCR

The LM-PCR processing pipeline generated was based on published protocols for IS determination and created considering the read structure obtained after sequencing (Wells et al. 2020; Ciuffi et al. 2009). The method was tested in a small set of reads and tuned to accomodate single-end and paired-end.

LM-PCR reads include two primers (one on each end), a linker, a host genome portion, and the LTR **(Figure 2.2)**. Thus, for the IS determination pipeline, reads with the LTR sequence (first mate in paired-end and unique mate in single-end) and linker (second mate in paired-end) were filtered while allowing for 2 mismatches. To generate higher quality alignments, LTR and linker were trimmed out using Cutadapt (v3.2)

(Martin 2011). If resulting trimmed reads were shorter than 15 bp, these were excluded. Trimmed reads were converted to FASTA format for BLAT alignment (parameters: *-stepSize=6 -minIdentity=97 -maxIntron=0 -minScore=15*) (Kent 2002).

BLAT resulting entries were filtered as follows: (i) aligned portions must be longer than 30 bp (for single-end reads) or 10 bp (for paired-end reads); (ii) alignment start position must be between the 1st and 5th base pair; (iii) alignment must be on standard chromosomes; (iv) for single-end multi-mapped reads, the difference between the longest aligned BLAT result and the second longest aligned BLAT result must be $\geq$ 25bp; (v) for paired-end multi-mapped reads, only mates aligning closer than 1KB were considered properly paired. IS were considered duplicates if distance to the nearest was shorter than 10bp. We obtained 1,771 IS from the paired-end reads and 2,822 from the single-end reads. These were merged into one set (N=4,590). IS were annotated to genes using ChIPpeakAnno (v3.24.2) (Zhu et al. 2010).

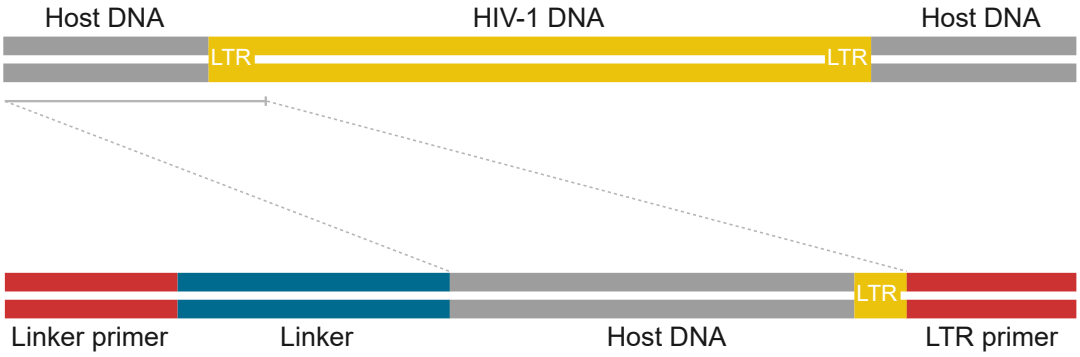


Figure 2.2: Structure of a LM-PCR read. After the virus has integrated, the LTR can be used to infer where the IS can be found in the genomic DNA (top). After the DNA sonication, the location of the IS can be determined using the viral-host junction. Between 2 rounds of PCR, an LTR primer, the linker, and a linker primer are added before sequencing (bottom).

RNA-seq analysis

RNA-seq reads were processed using the *nf-core* pipeline (alignment to the hg38 reference genome was performed by Martin Kampmann) (Ewels et al. 2020). I transformed

TPM values into logTPM and stratified genes according to expression levels by defining the top 10% expressed genes as *"high expression"*, bottom 10% expressed genes as *"low expression"*, genes between high and low expression as *"mid-expression"*, and the genes with logTPM = 0 as *"non-expressed"*.

ChIP-seq analysis

Sequencing data from ChIP-seq was processed using the HDSU pipeline, available in the HDSU GitHub repository (https://github.com/hdsu-bioquant). The pipeline uses *TrimGalore* (v0.4) for trimming (with a maximum allowed error rate 0.3), *Bowtie2* (v2.3) for genome alignment (hg38 reference genome), *MACS2*(v2.1) for peak calling with *broad cut-off=0.1* (Martin 2011; Langmead and Salzberg 2012; Zhang et al. 2008). RPKM-normalized BigWigs were generated using the input control file in *bamCompare* (Ramirez et al. 2014).

In the comparison between CTCF-KD and WT, differential peak analysis was performed using *DiffBind* (Wu et al. 2015). Change of binding associated to the TAD boundaries was determined by averaging the CTCF log2 fold change as computed by *DiffBind*. Assignment of CTCF peaks to TAD boundaries was done considering the overlap of CTCF peaks with the +/-50KB region around the TAD boundary midpoint. All ChIP-seq-derived profile plots and metagene plots were produced using *soGGi* (v1.20).

Super-enhancer determination

Super-enhancers were defined through single-end and paired-end peaks obtained from H3K27ac. I applied *HOMER* (v4.10) findPeaks function under the *-style super -o auto* parameters, as suggested by the authors (Heinz et al. 2010).

ATAC-seq analysis

Sequencing data from ATAC-seq was processed using the HDSU pipeline, available in the HDSU GitHub repository. The pipeline applies *TrimGalore* (v0.4) for trimming (with a maximum allowed error rate 0.3), *Bowtie2* (v2.3) for genome alignment (hg38 reference

genome), *MACS2*(v2.1) for peak calling (Martin 2011; Langmead and Salzberg 2012; Zhang et al. 2008). Peaks with a MACS score ≥ 30 are kept. RPKM-normalized BigWigs were generated in *bamCoverage* (Ramirez et al. 2014).

TFBS footprinting: Transcription factor footprinting was performed using the TOBIAS toolbox (v0.11.6) with motifs of TFs identified as part of the microglia TF signature (Bentsen et al. 2020; Gosselin et al. 2017; Nott et al. 2019). Scores directly derived from the *BINDetect* function output (TOBIAS toolbox) were used to infer binding dynamics of TFs.

Generation of matched phantom sites

For the profile plots and the chromatin state expected IS locations, I generated a set of control sites (termed *matched phantom sites*). Using TSS as baseline, I sampled a number of TSS corresponding to the original number of IS (N=4,590) in a chromosome-controlled manner, ensuring a similar IS chromosomal distribution to the real set. To ensure a similar distance to the closest TSS, I generated a pair of IS per TSS. This set was subsequently sampled considering an equal genic/intergenic balance to the real IS set.

Definition of IS-permissible windows (HMM- and NMF-based)

NMF: The genome was fully partitioned into 50KB windows, resulting into 57,238 windows (Quinlan and Hall 2010). All merged replicate files were converted to BigWig through *bamCompare* (ChIP-Seq) and *bamCoverage* (ATAC-Seq and RNA-Seq), and summarised over the windows using *multiBigwigSummary* (Ramirez et al. 2014). The resulting matrix (57,238x8) was decomposed into signatures through NMF using the *ButchR* package (v1.0) (Quintero et al. 2020). Computation was carried out over 10^4 iterations, 20 initializations, and rank factorization tested from 2 to 7. Final factorisation rank was 4. H- and W-matrix heatmaps were generated using R package *ComplexHeatmap* (v2.6.2) (Gu, Eils, and Schlesner 2016).

HMM: *ChromHMM* (v1.22) was used to generate a nucleosome-scale chromatin model

of the C20 cell line (Ernst and Kellis 2010, 2012, 2017). This model integrated data on histone modifications H3K27ac, H3K36me3, H3K4me1, H3K27me3, H3K9me3, and H3K9me2, along with chromatin accessibility (as obtained from ATAC-Seq). *Binarize-Bam* function was used to binarise the input data (bin size = 200 bp as default). *Learn-Model* function was used to train 5-state to 15-state models. The final model comprised 10 chromatin states. Chromatin states were identified and annotated using published information (Ernst and Kellis 2010; Hoffman et al. 2013).

Random-Forest for IS-targeted windows

Windows (N=57,238) used for the NMF analysis were separated and labeled (IS-targeted vs non-targeted). To the 8 epigenetic features included in the input data for NMF, further features were added (Expression, GC content, overlap with repeats, overlap with genes, overlap with CGIs, overlap with CTCF footprints, and overlap with SE). RF training was performed using the *caret* package with 70% of the initial set using 500 trees, and 10-fold cross validation (Kuhn 2008). Class imbalance was corrected through downsampling. Validation was performed with 30% of the initial set. Receiver operating characteristic (ROC) plots and calculation of area under the curve (AUC) were performed using *caret*.

Random-Forest for TF linked to TAD boundaries

To identify TFs linked to TAD boundaries, I trained a random forest classification model using footprinting-derived TFBS from ATAC-seq data on 9 biological samples (ENCODE Project Consortium 2012; Luo et al. 2020). ATAC-seq datasets used for TF footprinting can be found in **Appendix B**. Classes (TAD boudary or non-TAD boundary) were labelled using TADs obtained from the 3D Genome Browser **(Appendix C)** (Wang et al. 2018). A panel of 68 well-documented TFs was used for the TF footprinting. Model training was performed using *caret* package (mtry=2 after tuning, 500 trees, 10-fold cross-validation and downsampling for class imbalance correction) (Kuhn 2008).

Association between IS and TAD boundaries

Density plots for CTCF footprints, IS and histone modification peaks were produced using TADs from published glial cells (NeuN- cells composed of: oligodendrocytes, astrocytes, and microglia) (Hu et al. 2021). TAD boundaries are defined as the midpoint between two consecutive TADs, resulting into 2,077 TAD boundaries. Histone modification density plots were produced using MACS2-called peaks from C20. IS density plots combined IS from C20 (N=4,590), phantom IS, and bootstrapped IS, which were obtained by subsampling 80% of the integration sites.

Conservation score assessment

Conservation score of the TAD boundaries was determined by comparing the TAD boundaries from the NeuN- cells with a reference set of TAD boundaries (N=44 cells/tissues). The reference set of TAD boundaries was generated using TADs from undisturbed biological samples of the 3D Genome Browser (Wang et al. 2018). Score was computed as the fraction of sets from the reference where the same TAD boundary is found (defined as overlapping the +/-50KB vicinity of the Neu- TAD boundary).

2.4 Results

2.4.1 LTR-based IS discovery pipeline from LM-PCR

In retroviruses, IS discovery is mainly based on the presence of the LTR, a repeat region located at each terminal end of the provirus (Mandell, Bennett, and Dolin 2010; Sherman et al. 2017). Upon sequencing, the LM-PCR reads include: (i) a LTR region, marking the start of the integrated HIV-1 provirus, (ii) a portion of genomic DNA from the host, (iii) a linker sequence, used for amplification, and (iv) two sequencing primers, one for the LTR, and another one for the linker (Wells et al. 2020). Upon processing, the aim is to yield only host genomic sequence in order to correctly locate the IS.

I developed a blat-based workflow to determine IS from LM-PCR sequencing data, similarly to a published methodology **(Figure 2.3)** (Ciuffi et al. 2009; Wells et al. 2020).

The methodology developed can be applied to LM-PCR data from distinct human cell types targeted by HIV-1 or other retroviruses, although this was not done here. This approach can also be applied to both single-end and paired-end reads. The LTR is used for read filtering and is then trimmed to increase the efficacy of alignment. Blat, a BLAST-like alignment tool, is used for alignment because the trimmed reads are shorter and it directly produces a set of possible results for each read which can be filtered according to the user's criteria (see *Methodology*) (Wells et al. 2020).

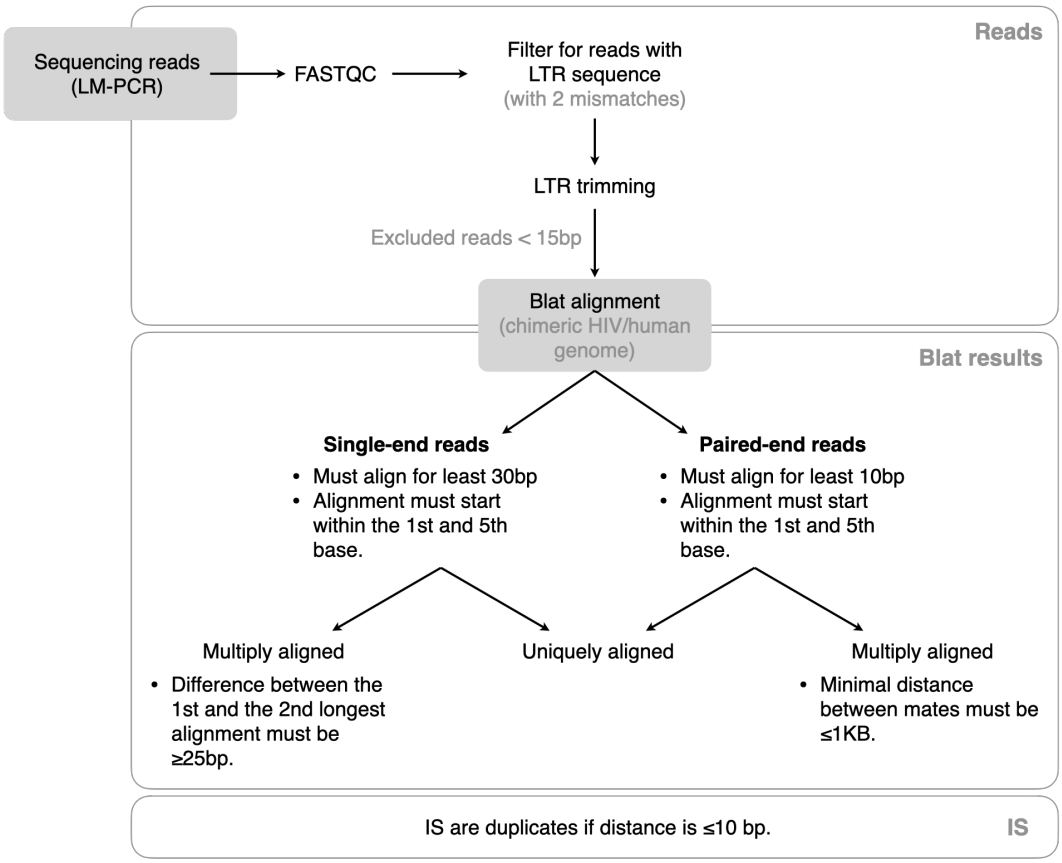


Figure 2.3: Diagram of the LM-PCR processing pipeline for IS.

2.4.2 Location-based comparison of the IS found on microglial cells with IS from other cell types

With the pipeline previously described, I recovered a set of IS (N=4,590) from the microglial cell line C20 (Garcia-Mesa et al. 2017), after infection with HIV-1[3]. The C20 cell line model was used on this work as the study of HIV-1 latency with real human brain data is difficult and obtention of CNS cells is dependent on invasive procedures (Farhadian et al. 2018). C20 has been used in other HIV-1 research work and it expresses typical microglial markers (Garcia-Mesa et al. 2017; Alvarez-Carbonell et al. 2019; H. Liu et al. 2020). Nevertheless, we have compared HIV-1 insertion patterns from C20 with another cell model commonly used in microglial studies, the iPSC-derived microglia. On this analysis, we have concluded that HIV-1 targets similar regions in both cells **(Appendix D)**.

Microglial IS were annotated to the human genome to assess genomic features of integration on this cell type **(Figure 2.4a)**. Microglial IS are mostly found within gene bodies, particularly in intronic regions (58%). The genomic distribution was also compared with the one observed into CD4+ T cells (N=13,544) and MDM (N=987) from previous publications **(Figure 2.4a-b)** (Kok et al. 2016; Lucic et al. 2019). Overall, chromosomal IS distribution is similar to the other cell types used for the comparison. Similarly, IS targeted mostly gene bodies, in particular intronic regions, on all the cell types. These commonalities suggest the integration profiles in microglia are similar to the other HIV-1 cell targets, but more to CD4+ T cells than to MDMs.

We next compared the genes targeted on the 3 cell types. Similarly, there seems to be more commonality with CD4+ T cells (Jaccard index = 0.209) than with MDMs (Jaccard index = 0.096) **(Figure 2.4c)**. MDM is more closely related to microglia, so the observation that the genes targeted by HIV-1 IS are more similar to CD4+ T cells targets is surprising. Considering the similarities between microglia and CD4+ T cells and the vast amount of data available in the context of HIV-1 infection and integration

[3] This and the remaining inhouse experimental data used on this chapter was entirely generated by the Lusic group (CIID, Heidelberg)

on the latter, we focused on CD4+ T cells as a baseline of comparison for the microglia cell model in the next sections.

Next, we assessed the impact of gene expression on IS targeting. The integration of the IS with expression of their gene targets (by levels, from *"no expression"* to *"high expression"*) revealed that IS on microglia are mostly found on medium or highly expressed genes (91%), similarly to IS in CD4+ T cells, albeit this effect appears stronger on microglia **(Figure 2.4d)**. I conducted a GO analysis (Biological Processes) to verify functions and pathways these genes are involved in, as it is possible that viral integration hinders the normal gene function **(Figure 2.4e)** (R. Liu et al. 2020). Interestingly, IS-genes seem to be associated to maintenance processes of epigenetics and chromatin.

2.4.3 Linking IS with specific histone modifications and transcription levels

HIV-1 integration has been associated with active transcription and several histone modifications in other cell types (Imai, Togami, and Okamoto 2010; Méndez et al. 2018; Lange et al. 2020). Thus, we assessed the epigenetic landscape which would make a region permissible for integration in microglia using ChIP-seq and ATAC-seq. Data from H3K4me1 (poised enhancers), H3K36me3 (active transcription and gene bodies), H3K27ac (active enhancers), H3K9me2 (facultative heterochromatin), H3K27me3 (Polycomb-mediated repression), and H3K9me3 (heterochromatin) on the uninfected C20 cell line was used. Chromatin accessibility was also assessed on a sorted uninfected cell population. The epigenetic landscape of the entire set of IS was averaged over its vicinity for each dataset **(Figure 2.5a)**. A set of randomly generated IS with similar chromosomal distribution and similar distance to the closest gene is used as a baseline of comparison (see *Methodology*).

Chromatin is normally accessible around the IS but not at its location, where it is closer to the expected (as indicated by P IS in **(Figure 2.5a)**), following a trend that is similar to the one observed from active enhancer modification H3K27ac. This suggests that the IS could be located near actively transcribed open-chromatin regions. The

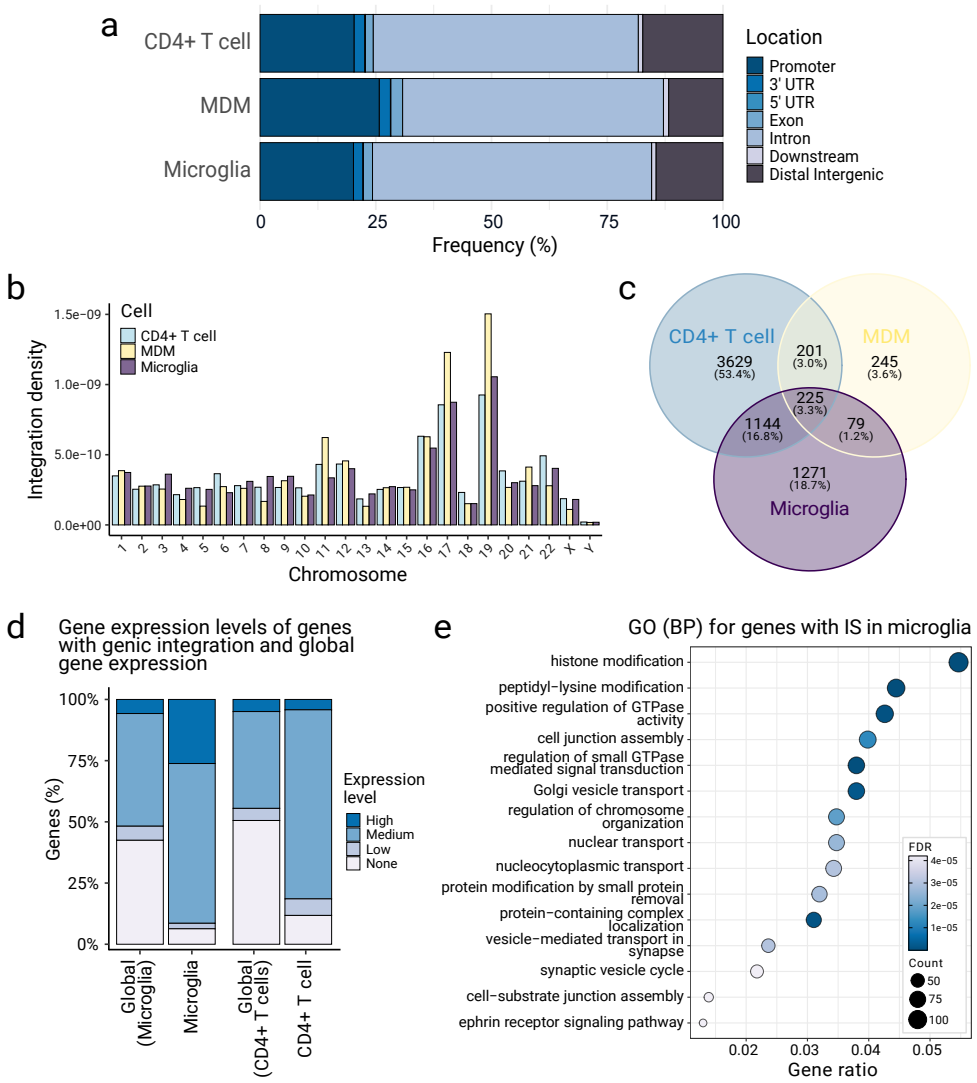


Figure 2.4: Genomic features of integration in microglia in comparison with other HIV-1 cell targets. [a] Genomic locations of IS from microglia annotated in comparison with other HIV-1 cell targets. [b] Normalised chromosomic distribution of IS on microglia and comparison with other cell types. [c] Intersection of genes where genic integration is observed on microglia, MDM, and CD4+ T cells. [d] Expression of genes (dicretized into 4 levels) where genic integration is observed in microglia and comparison with CD4+ T cells. Global ratios are shown. [e] Gene ontology enrichment (*Biological Processes*) for genes where genic integration is observed on microglia. Color represents significance (FDR) and dot size indicates number of genes in each ontology term. *Panels of this figure were adapted from Rheinberger et al. (2023).*

profiles of H3K4me1 and H3K9me2 indicate these are likely not important integration determinants. Repressive histone modifications H3K27me3 and H3K9me3 are possibly avoided by HIV-1 upon integration. On the other hand, the gene body-associated H3K36me3 seems to favour HIV-1 targeting. It was also found that H3K36me3-marked regions are associated with HIV-1 targetting regardless of their location (as genic or intergenic) **(Appendix E)**. Globally, these trends are also similar to the ones observed in CD4+ T cells **(Appendix F)**.

As gene expression and H3K36me3 could be confounding, we looked back to genes that are targeted in microglia, stratified them by expression level, and compare them with non-targeted genes **(Figure 2.5b)**. An example of the highly targeted gene *NPLOC4* is also shown **(Figure 2.5c-d)**. It is clear that while HIV-1 targets open-chromatin and transcribing regions, the main epigenetic driver for integration in microglia could be H3K36me3.

2.4.4 Defining integration-permissible windows through epigenomics clustering (HMM- and NMF-based)

While individual histone modifications provide important insights into the epigenetic landscape of the genome, combinations of chromatin modifications are often used to infer effects on a systematic manner for multiple cell types (Ernst and Kellis 2012). Histone modification or chromatin accessibility data produced through ChIP-seq and ATAC-seq can be integrated into chromatin states or reduced into epigenomic signtures (Ernst and Kellis 2012; Stuart et al. 2021; Roadmap Epigenomics Consortium et al. 2015). Chromatin states are highly cell-type–specific (Mikkelsen et al. 2007; Hawkins et al. 2010). Thus, I unbiasedly integrated our chromatin accessibility, histone modification, and transcriptome data to infer context-dependent relationships between the different epigenomic layers of the regions targeted by HIV-1. I applied two different approaches on different scales. First, on the nucleosomal scale (200bp), I used *ChromHMM* to infer chromatin states of integration-permissible regions **(Figure 2.6a)** (Ernst and Kellis 2017, 2012). Then, I used NMF to assess the preferential epigenetic landscape of integration on a

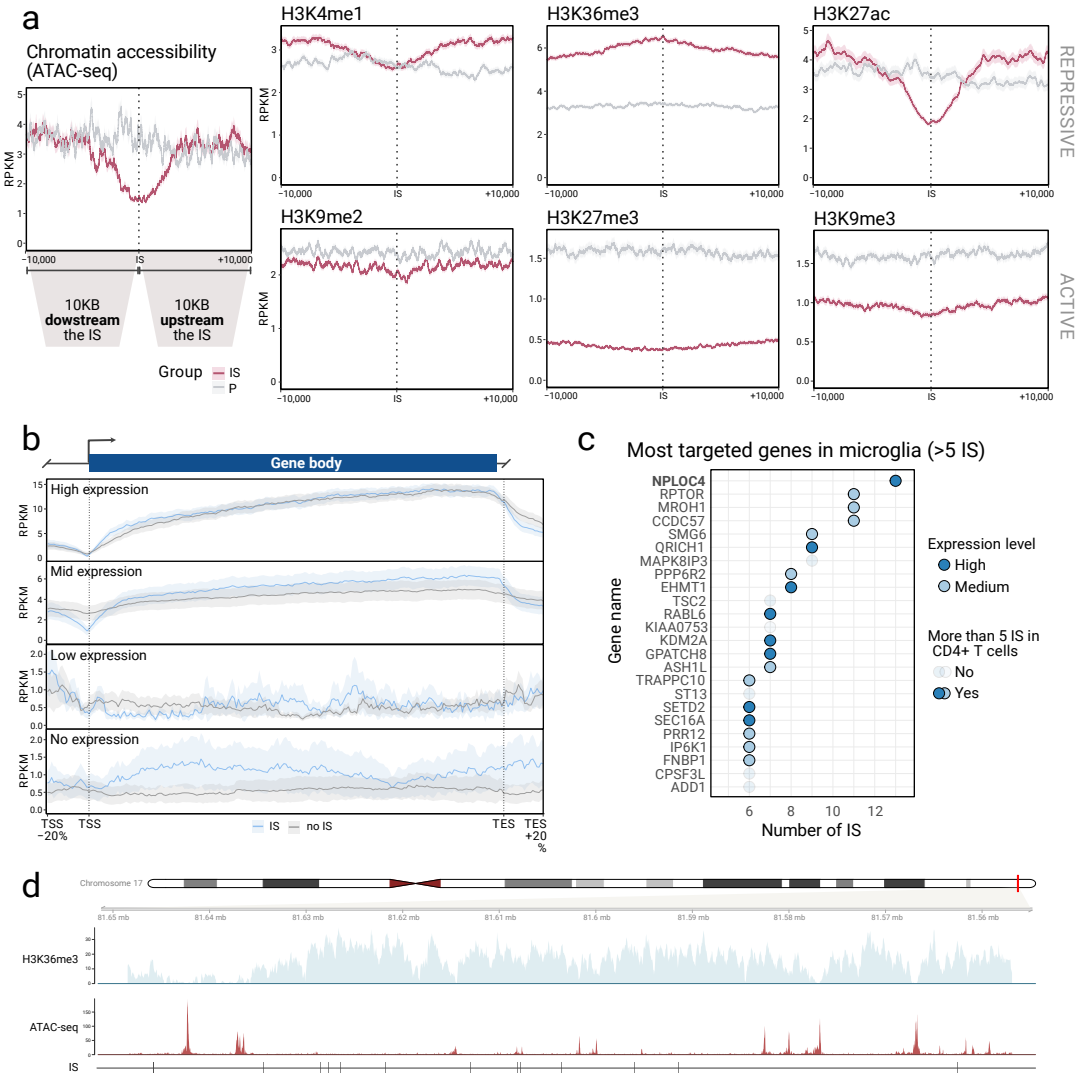


Figure 2.5: Epigenomic characterisation of IS-associated genes and regions in microglia. [a] Epigenetic profiles (in reads per kilobase of exon per million (RPKM)) for chromatin accessibility, H3K4me1, H3K36me3, H3K27ac, H3K9me2, H3K27me3, and H3K9me3 on the IS vicinity (10KB upstream and 10KB downstream) as before integration. In each panel, profiles are shown for the averaged signal over the IS set (IS, in red) and over a matched phantom IS set (P, in grey). Confidence interval (95%) is shown in shaded color. [b] H3K36me3 signal (RPKM) over the 4 gene expression groups previously displayed in Figure D. Genes targeted by IS are shown in blue, while genes without IS from the same expression level are shown in grey. Confidence interval (95%) is shown in shaded color. [c] Representation of the most targeted genes in microglia (more than 5 IS per gene) stratified by number of IS (x-axis) per gene (y-axis) and colored by expression level. Transparency indicates if the gene is also frequently targeted in CD4+ T cells. [d] Epigenetic profile (H3K36me3 and chromatin accessibility, in RPKM) of the NPLOC4 gene and the set of ISs (N=13, at the bottom) found in its gene body. *Panels of this figure were adapted from Rheinberger et al. (2023).*

larger resolution (50KB), to allow the assessment of histone modifications influencing targeted regions distally **(Figure 2.6b-c)** (Quintero et al. 2020).

On the nucleosomal scale, we observed that the IS are frequently found on the *Strong Transcription* state (35.1%), characterised by an enrichment in H3K36me3. *Genic enhancers* (5.7%), *weak enhancers* (5.6%), and a *H3K27ac/ H3K9me2-enriched* state (4.3%) are more targeted than expected (One-sided binomial test, expectation is displayed as *All states*, p.value ≤ 0.05) **(Figure 2.6a)**. Although there seems to be a high overlap with the *Quiescent* state (44.8%), this is still significantly lower than expected (61.3%), implying this state is associated with avoidance, as are *Heterochromatin* (0.9%) and *Polycomb* (both high and low) (0.3% and 0.5%) (One-sided binomial test, expectation is displayed as *All states*, p.value ≤ 0.05).

On a larger scale, I generated 4 distinct NMF-derived signatures with the same data while additionally including gene expression (see *Methodology*) (Quintero et al. 2020). We found comparable results to the nucleosomal scale using this approach **(Figure 2.6b-d)**. Signature 1, mostly characterised by H3K36me3 and expression, is the most permissible to integration, as the *Strong Transcription* and *Genic Enhancers* in the ChromHMM. Signature 4, mostly associated to H3K27ac, high chromatin accessiblity, and H3K4me1, is also targeted, but much less. Similarly to the repression-associated states from the ChromHMM, signatures 2 and 3 seem to be avoided by HIV-1 upon integration. The larger scale NMF-derived signatures can be compared with larger scale elements, such as SE, which were frequently found in the vicinity of IS in CD4+ T cells (Lucic et al. 2019) **(Figure 2.6d)**. However, in microglia, SE do not seem to be found in the same signature as most IS, while in CD4+ T cells they do **(Appendix G)**.

Although IS locations seem to be related with specific features of the regions, it was still unclear what really distinguishes HIV-1 targeted windows. To understand this, I trained a RF-based model to classify HIV-1 targeted and non-targeted windows using 14 genomic features (see *Methodology*). The model is able to identify each class (AUC=0.785) and it mostly relies on gene expression and H3K36me3, as suggested by the remaining analysis **(Figure 2.6e-f)**. However, other features known to be important in other cell types,

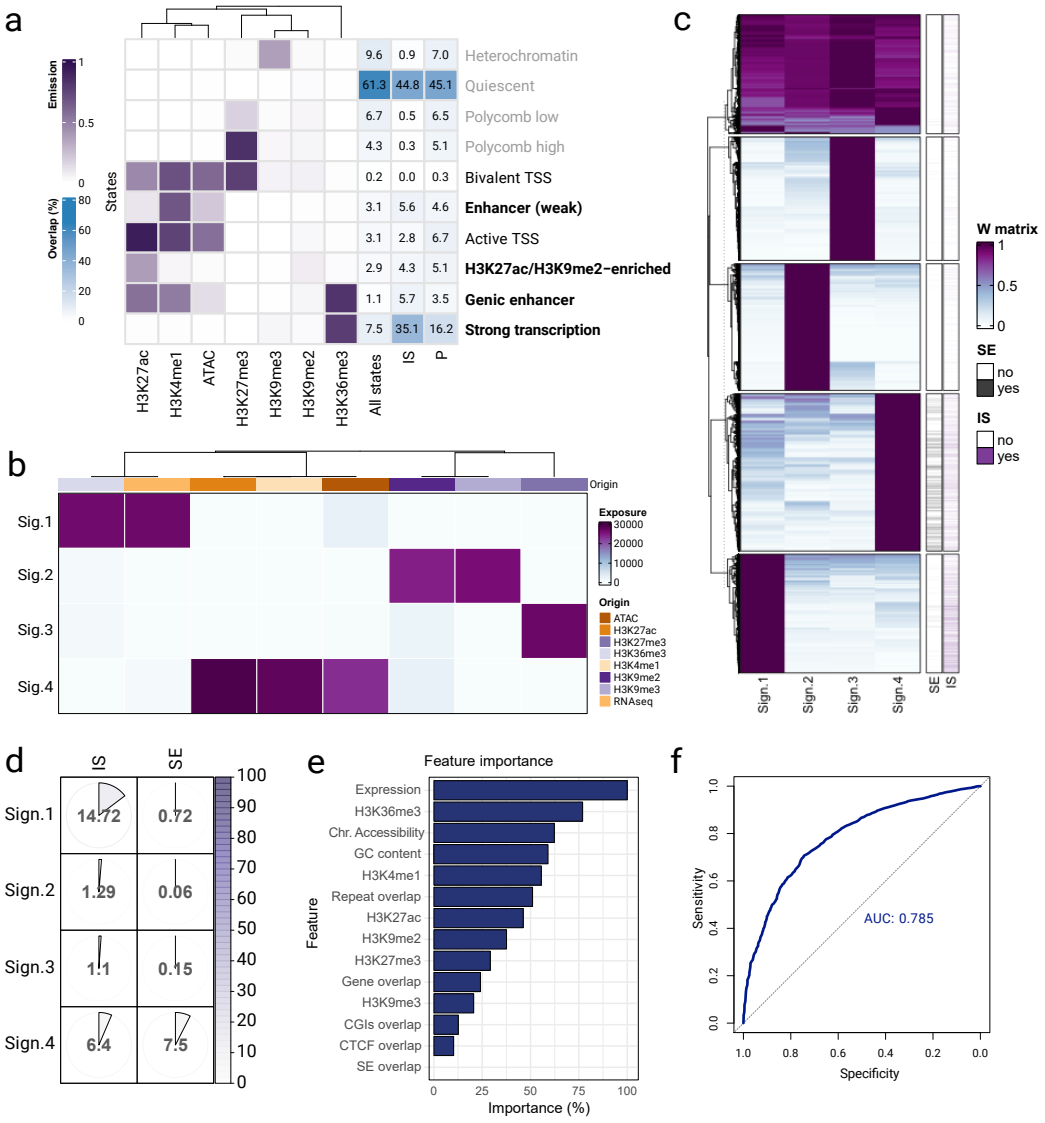


Figure 2.6: Signatures of HIV-1 integration on the microglia cell model. [a] Emission map of the ChromHMM-derived 10-state model (left, in purple) for the 7 features (ATAC-seq and ChIP-seq for 6 histone modifications) used was used to infer each chromatin state using published references. IS percentages (right, in blue) are represented for each state in comparison with all states and random matched phantom sites (P). Chromatin states targeted by HIV-1 more than expected are labelled in bold, while states targeted less than expected are labelled in grey. [b] Exposure matrix H for NMF-derived signatures (k=4, in rows) based on expression (RNA-seq), chromatin accessibility (ATAC-seq) and ChIP-seq for 6 histone modifications. [c] Exposure matrix W for NMF-derived signatures (in columns) on all genome windows. Colour indicates if the window is assigned to one signature. Bars on the right indicate whether each window overlaps with IS (purple) and SE (black). [d] Representation of the overlap between each NMF-derived signature and the IS and SE set in microglia. Both colour and angle represent the overlap (%). [e] Feature importance of the RF model used to classify HIV-1 targeted or non-targeted windows. Features are ordered by importance (%). [f] ROC curve and AUC for the RF used to classify HIV-1 targeted or non-targeted genome windows. *Panels of this figure were adapted from Rheinberger et al. (2023).*

such as SE overlap, are not as important for HIV-1 targeting in microglia, as suggested by the comparison between **Figure 2.6d** and **Appendix G (c)** (Lucic et al. 2019).

2.4.5 Assessing differential TF binding on distinct HIV-1 infection states

While the host chromatin landscape influences HIV-1 integration, the integration could also lead to alterations on the chromatin landscape. Therefore, chromatin accessibility was profiled through ATAC-seq on sorted C20 cell populations (see *Data*) from the three possible states after HIV-1 infection: uninfected (eGFP-mKO2-), active infection (eGFP+mKO2+), and latent infection (eGFP-mKO2+)[4].

While differences between the three conditions were subtle **(Table 2.1)**, TF footprinting revealed significant changes in the predicted binding between conditions. We observed that some TF were differentially bound between conditions in at least 2 comparisons, such as CTCF, FOS, NFKB1, SMAD2-4, and MAFF **(Figure 2.7a)**. As the set of TF selected for footprinting was based on a microglia-specific TF signature, these observations could imply that different infected states could, to some extent, lead to alterations on the microglial transcriptional identity (Gosselin et al. 2017).

Table 2.1: ATAC-seq peaks on the three cell populations (MACS2 q-value < 0.001)

Peaks	Condition
92812	Uninfected
95401	Active
54744	Latent

[4] Infection with the HIV-GKO reporter virus and sorting were performed by the Lusic group (CIID, Heidelberg)

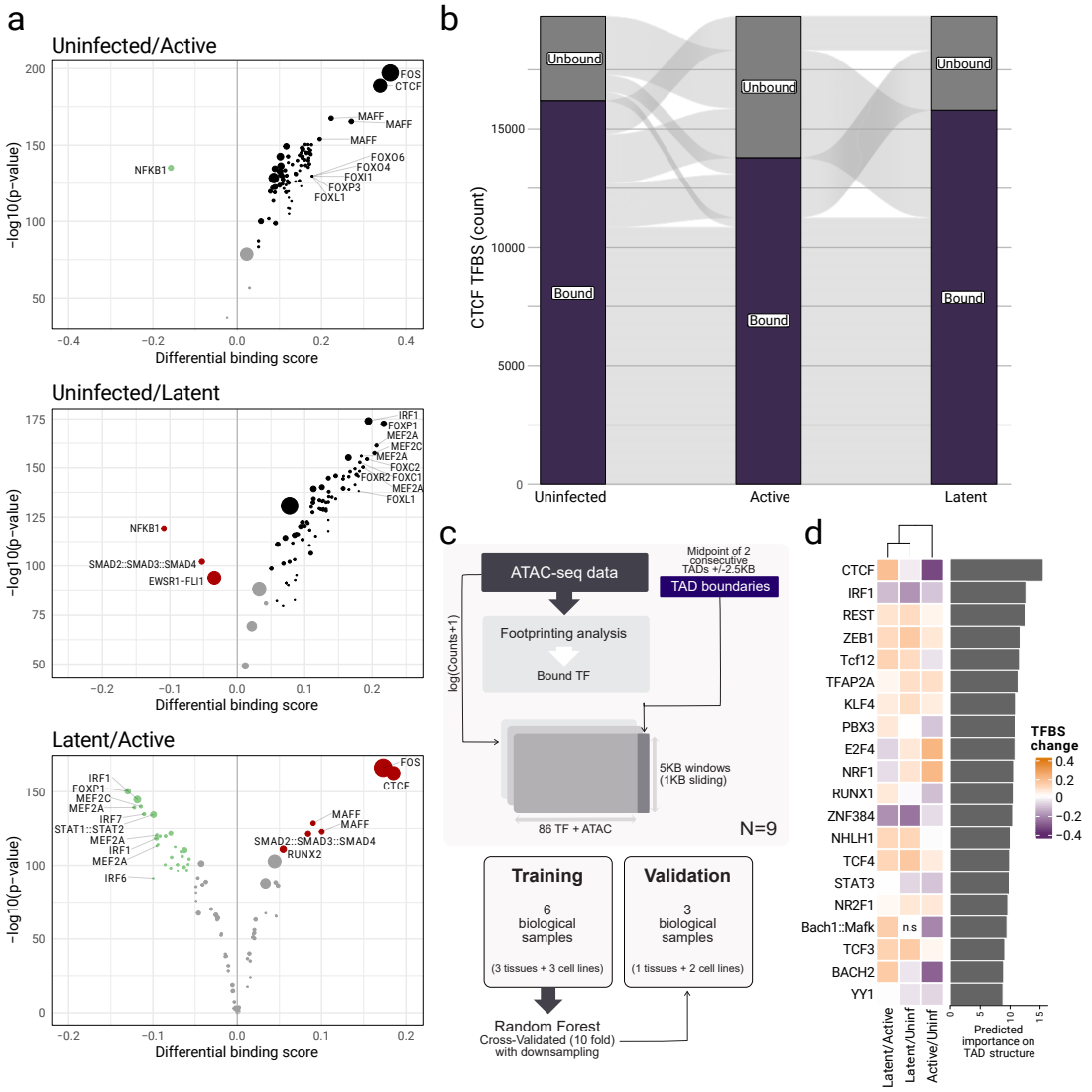


Figure 2.7: TF binding dynamics between the different cell states. [a] Predicted differential binding between uninfected and active infection (top), uninfected and latent infection (middle), and latent and active infection (bottom). Coloured dots (red for latent/green for active/black for uninfected) indicate more differences in level of TF binding for the respective condition comparison. Dot size indicates the ratio of TF binding sites bound for one TF per total of bound TF. [b] 20 TFs most likely to contribute to TAD boundary formation or function ordered by their predicted importance. Heatmap represents the three condition comparison pairs (columns) and the direction of binding over the baseline (colour) for each TF. [c] Diagram of the RF model generated to identify TF linked to TAD boundaries. [d] Binding dynamics of all predicted CTCF TFBS over the three conditions. For each condition, both bound (purple) and unbound (grey) TFBS are shown. *Panels of this figure were adapted from Rheinberger et al. (2023).*

2.4.6 Random forest classifier defines TFs most linked to TAD boundaries

CTCF, linked to chromatin structure in the context of the formation and maintenance of TADs and chromatin loops, was observed to be one of the most altered TFs between the three conditions **(Figure 2.7a-b)** (Splinter et al. 2006; Wit et al. 2015; Wutz et al. 2017). TADs have been found to be altered upon integration of human papillomavirus (Groves et al. 2021). Overall, given that active infection leads to a decrease in CTCF binding, while latent infection leads to an increase of CTCF binding in comparison with active infection, we speculated whether these changes could be hinting on a proviral-driven alteration on higher-order chromatin structure.

Considering the potential of TF footprinting and the conserved nature of TAD boundaries, we hypothesised that other TFs linked to these functions could also be altered. Thus, I trained a RF model on an independent set of cell lines and tissues to find differences on chromatin accessibility and on the predicted binding of several TFs between TAD boundaries and non-TAD boundaries (see *Methodology*). Chromatin accessibility was the most important feature for the TAD boundary distinction while, unsurprisingly, CTCF was the top TF feature **(Figure 2.7c; Appendix H)**. Along with CTCF, the model suggests that TFs such as IRF1, REST, ZEB1, Tcf12, among others, could be linked to TAD boundaries. These TFs, consistently found to be associated to TAD boundaries on other cell types, were also found to be altered between the conditions, suggesting that the HIV-1 infection state could alter the delicate TF-mediated balance necessary for TAD boundary maintenance or formation (Hong and Kim 2017).

2.4.7 Associating HIV-1 integrations with TAD boundaries

To better understand the interplay between the HIV-1 integrations, TAD boundaries, and CTCF, we compared the CTCF footprints on the uninfected cell state with TAD boundaries (N=2,077) obtained from a non-neuronal (Neu-) glial population (Hu et al. 2021) **(Figure 2.8a)**. However, the Neu- population includes other cells beyond microglia, such as astrocytes and oligodendrocytes, making the comparison sligtly indirect

(Hu et al. 2021). Upon verification that the CTCF footprints obtained in C20 are located near the TAD boundaries of Neu- and most promoter-enhancer contacts from primary microglia (Nott et al. 2019) are also within the TADs from Neu- **(Appendix I)**, we proceeded to cautiously use these data for our comparison.

Given that the integration of the viral genome could be triggering the alterations on the chromatin landscape observed on the actively and latently cell states, we compared the locations of the IS with the TADs from Neu-, a sorted population of brain cells which countains microglial cells **(Figure 2.8b)**. Notably, IS seem to distribute close to the TAD boundaries. However, this pattern seems to be mirrored by co-localisation of the histone modification H3K36me3 at the TAD boundary, implying that this modification could still be acting as the main integration driver **(Figure 2.8c)**. Histone modifications which were previously found to be avoided by HIV-1 upon integration, such as H3K9me3, are also less common at the TAD boundaries **(Figure 2.8c)**.

To clarify if the enrichment of H3K36me3 is confounding, we compared the locations of 4 classes of IS: (i) overlapping H3K36me3 peaks (N=2,450), (ii) not overlapping H3K36me3 peaks (N=2,140), (iii) located within gene bodies (N=3,862), and (iv) located outside genes (N=728) **(Figure 2.8d)**. IS overlapping H3K36me3 peaks are more often located at the TAD boundary, while IS not overlapping H3K36me3 and intergenic IS tend to be located within the TAD, closer to its midpoint. H3K36me3 is also enriched on TADs harbouring IS when compared to TADs which do not **(Figure 2.8e)**.

The role of H3K36me3 as an important driver for HIV-1 integration is widely documented, as the celullar factor LEDGF/p75, which can recognize this histone modification, interacts with HIV integrase (Cherepanov et al. 2003; Lapaillerie et al. 2021). Moreover, CTCF has been shown to modulate the action of other histone modifications, such as H3K27me3 (Weth et al. 2014). Nevertheless, the landscape at TAD boundaries appears to be different from the majority of the genome, as suggested by **Figure 2.8**. While the role of CTCF in looping is well understood, we postulated whether its local co-existence with H3K36me3 and other histone modifications at the TAD boundaries could imply these players form functionally relevant interactions for the maintenance of

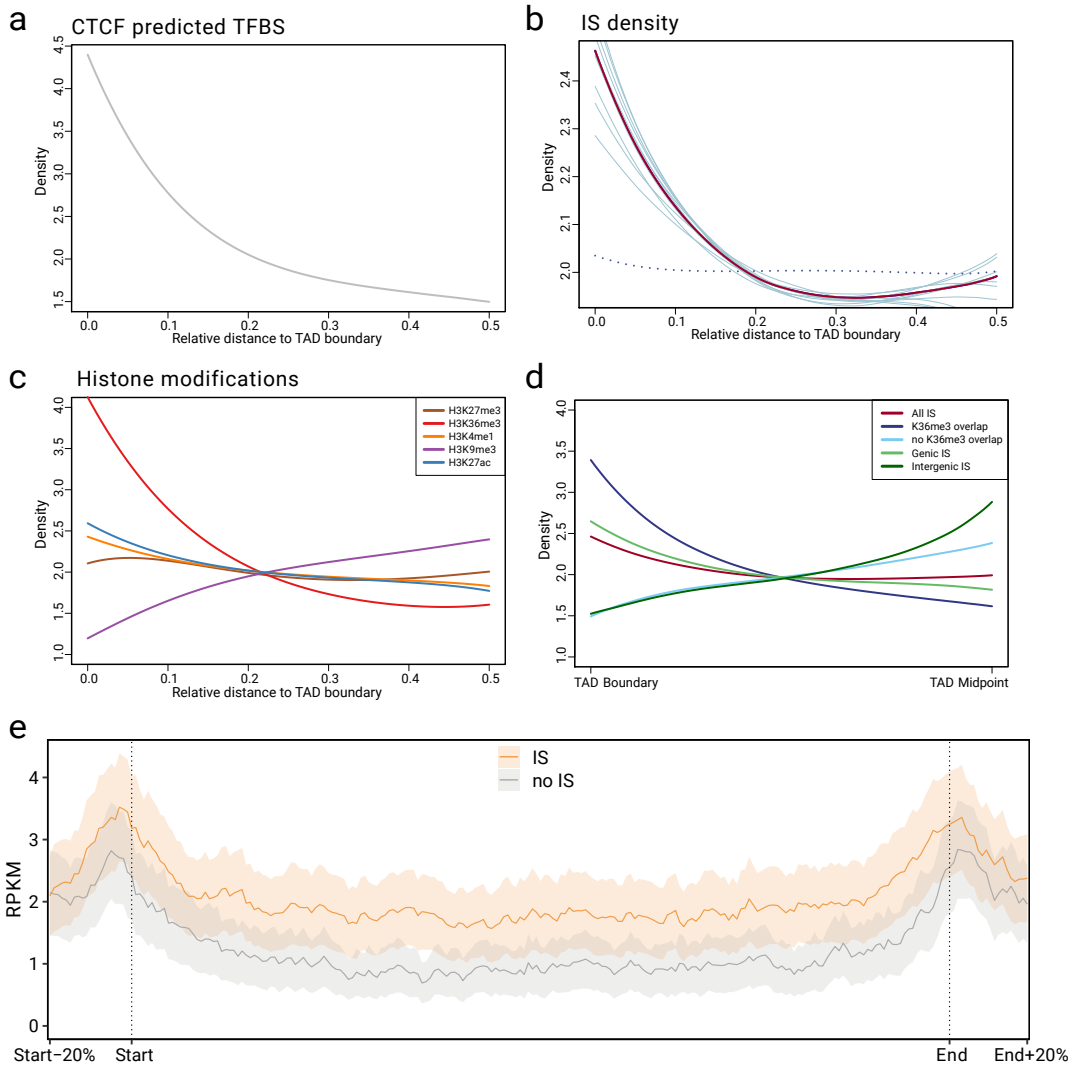


Figure 2.8: IS distribution over the TADs from Neu- and the potential effect of the potential . [a] Distribution of the predicted CTCF TFBS over the TADs. Relative distance to boundary (0) is shown (x-axis), where 0.5 represents the scaled midpoint of the TADs (N=2,077). [b] Distribution of the C20 IS (in red) over the TADs. Relative distance to boundary is shown as in panel a. Blue lines represent 10 random subsamplings of the IS and the dotted line represents a random set of IS. [c] Distribution of the peaks from H3K27me3, H3K36me3, H3K4me1, H3K9me3, and H3K27ac (coloured lines) on C20 over the TADs. Relative distance to boundary is shown as in panel a. [d]] Distribution of the C20 IS over the TADs by category (established using H3K36me3 overlap or genic/intergenic locations) in comparison with all IS (in red). [e] Averaged H3K36me3 profiles (RPKM) for IS-targeted (in orange) and non-targetted (in grey) TADs (including 20% upstream and 20% downstream). Confidence interval (95%) is shown in shaded color. *Panels of this figure were adapted from Rheinberger et al. (2023).*

the TAD boundary (Li et al. 2020). To understand if histone modifications might have a role in the establishment of the TAD boundaries through the interplay with CTCF, we have trained a Bayesian network model over the TAD boundaries using both histone modifications and CTCF predicted-TFBS on the C20 cell line **(Figure 2.9)**. Using this approach, we can also infer if epigenetic players like H3K36me3 could have a driving role into the establishment of the provirus at these locations. While the directionality of some of the connections is deemed unresolved (as indicated by the bi-directional arrows in many of the edges), it emphasises the central role of CTCF and H3K27me3 at these locations in microglia, a finding which can translate into other cell types and explain how these higher-order chromatin structures are formed. Moreover, these observations can also imply that H3K27me3 and CTCF are, together with H3K36me3, important drivers of IS targeting in the host cell.

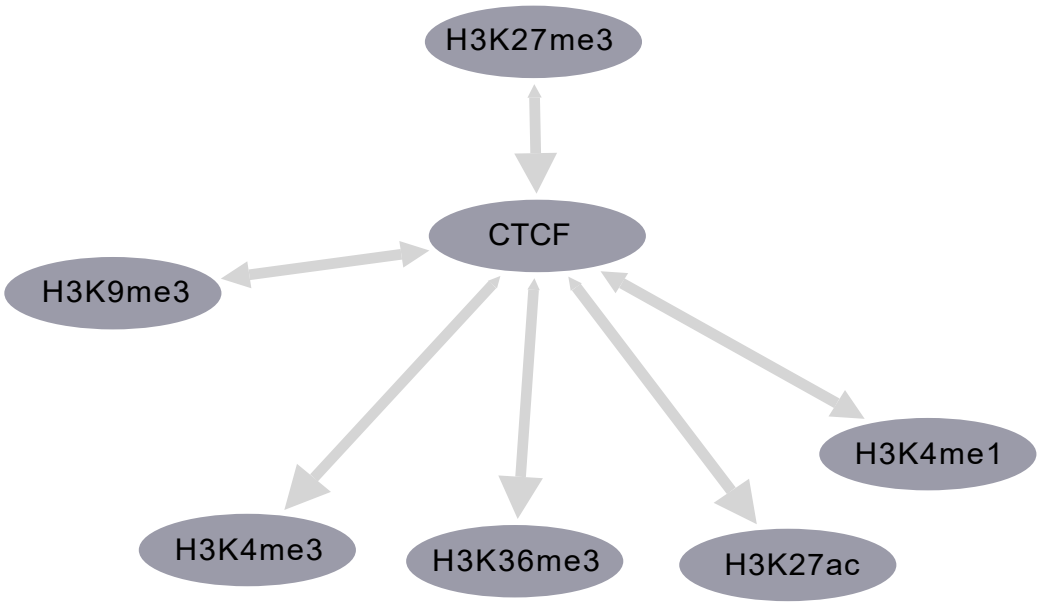


Figure 2.9: Epigenomic Bayesian network on the TAD boundaries in microglia. Edges represent the connections between the different nodes (epigenomic features) which are predicted during training. *This figure was adapted from Rheinberger et al. (2023).*

2.4.8 Comparing TAD boundary conservation levels with infection-driven TF binding alterations

TAD boundaries are known to be conserved between cell types and species (Dixon et al. 2012; Rao et al. 2014). I assessed the level of TAD boundary conservation in the Neu- population we are using in our comparative analysis. To do so, I combined TAD boundaries from 44 undisturbed cell lines and tissues and intersected it with the Neu- TAD boundaries. Most Neu- TAD boundaries are present in 25% to 50% of the reference set **(Figure 2.10a)**.

The maintenance of TAD boundaries is highly dependent on CTCF, so we hypothesised that CTCF binding dynamics could be related with conservation level of the TAD boundaries associated to it. Thus, I assigned CTCF-footprinted TFBS to TAD boundaries and compared the binding dynamics observed in uninfected, latent, and active conditions **(Figure 2.10b)**. We observed that most dynamic CTCF-footprints on latently infected cells are associated to less conserved TAD boudaries, more specific to Neu- and, potentially, to microglia.

2.4.9 Verifying the effects of CTCF loss into HIV-1 integration

As the CTCF is an important factor in the establishment of TAD boundaries, our collaborators generated a CTCF knock-down (KD) of the microglial cell model[5]. ChIP-seq was performed on both WT and CTCF-KD conditions **(Figure 2.11a)**.

The CTCF-KD led to a significant drop of CTCF, as expected (42,472 peaks: FDR ≤ 0.05, Log2 fold change ≤ -1). In some regions, binding is increased by the CTCF-KD (1,119 peaks: FDR ≤ 0.05, Log2 fold change ≥ 1) although this is mostly observed in high binding regions, impling an over-compensation in locations where binding could be essential. Simultaneously, IS were obtained under the same CTCF-KD and -WT conditions to understand the impact of CTCF loss in the integration **(Figure 2.11b)**. CTCF-KD led to a decrease in the total IS numbers (2,814 versus 2,326 IS). Most importantly,

[5] This experiment was performed by the Lusic group (CIID, Heidelberg) using CTCF-targeting siRNA.

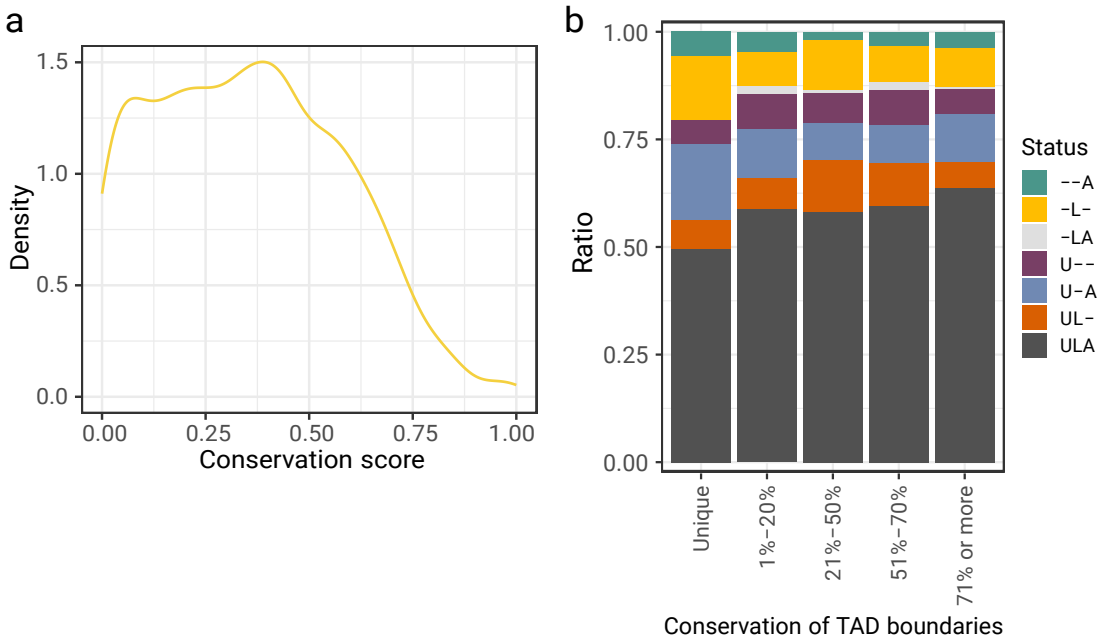


Figure 2.10: Comparison between conservation levels and infection-driven CTCF binding dynamics. [a] Conservation score for each TAD boundary on the Neu- cell population, calculated using a TAD boundary reference set (N=44 cell lines/tissues). [b] Comparison between conservation levels and the dynamics of CTCF predicted-TFBS on the uninfected, active, and latent C20 populations. *Panels of this figure were adapted from Rheinberger et al. (2023).*

integration in the CTCF-KD was directed into locations other than the WT **(Figure 2.11c)**. While it is not possible to infer that this CTCF-KD experiment resulted into TAD boundary alterations, it has been shown in other studies that CTCF-KD experiment lead to TAD disruption (Khoury et al. 2020). I compared the log2 fold change of CTCF associated to TAD boudaries (CTCF peaks found in the +/-50KB region of the TAD boundary midpoint) with non-TAD boundary CTCF and we observed that there appears to be a tendency to retain TAD boundary associated CTCF unchanged.

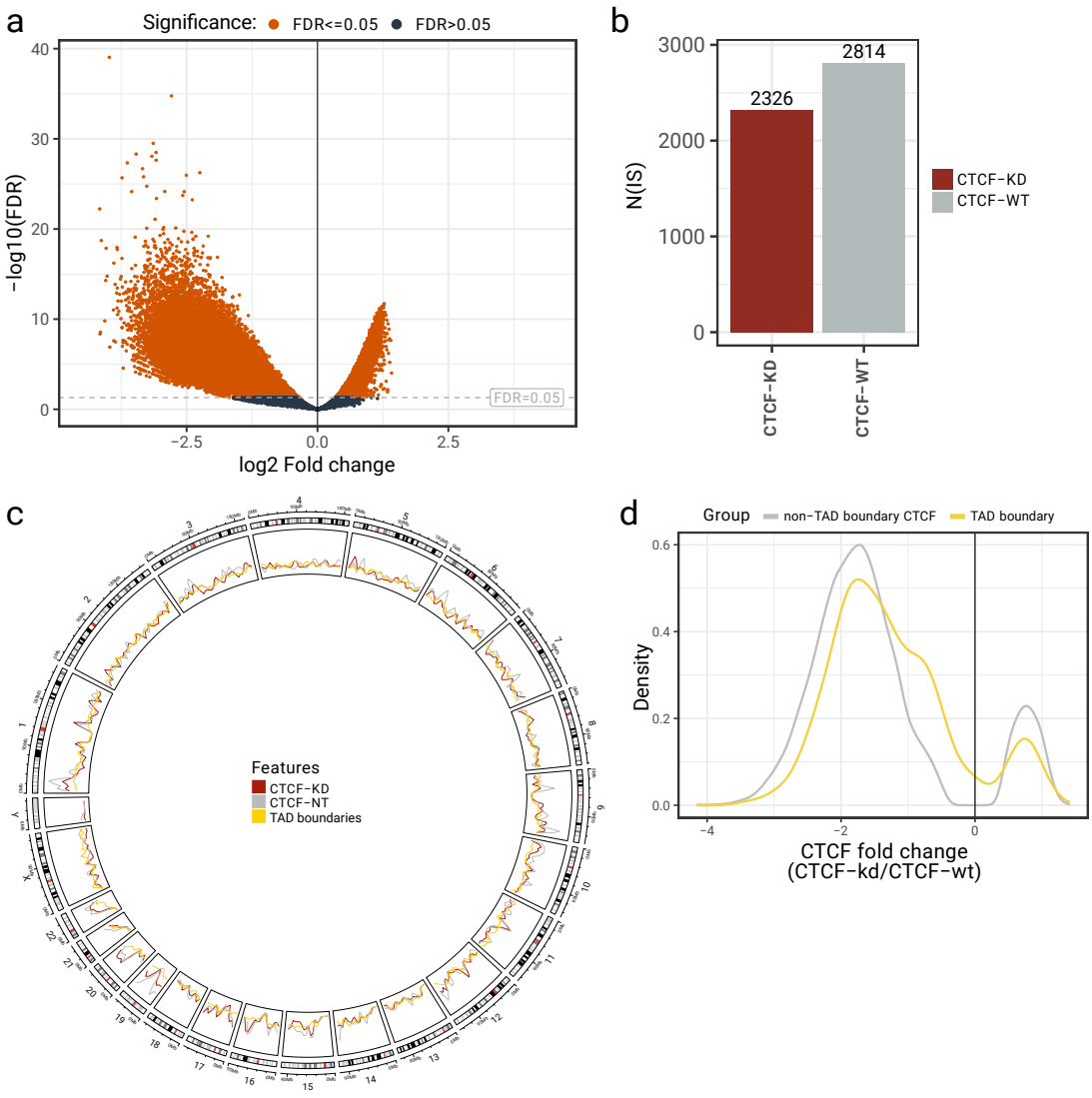


Figure 2.11: Comparison of the CTCF-KD with WT. [a] Volcano plot illustrating the effect of the CTCF-KD in CTCF. Each dot represents one peak, while dark orange represents peaks significantly altered (increased upon CTCF-KD if located on the right, decreased if located on the left). Grey dotted line borders the significance threshold. [b] IS numbers in the CTCF-KD (red) and -WT (grey) conditions. [c] Comparison of the IS (in both CTCF-KD and -WT) and TAD boundary distribution (in yellow) in the human genome. CTCF-KD IS are shown in red, while control CTCF-WT is shown in grey. [d] Density of CTCF (FDR ≤ 0.05) log2 fold change (x-axis) by location in relation to TAD boundaries. Averaged CTCF associated to TAD boundaries (in yellow) and CTCF non-associated to TAD boundaries (in grey) is shown as distinct density lines. *Panels of this figure were adapted from Rheinberger et al. (2023).*

2.5 Discussion

2.5.1 Genomic features of HIV-1 integration in microglia

HIV-1 integration is a biological process known to be influenced by the genetic and epigenetic landscape in the host cell (Vansant et al. 2020; Lelek et al. 2015). In this project, we explored the integration in a microglial cell model. Microglia are one of the target cells for HIV-1 in the brain and an important player to its persistence, latency, and immune evasion (Alvarez-Carbonell et al. 2019).

While it is known that viable HIV-1 proviruses remain in the brain even after ART, the usage of the C20 cell model in place of microglia obtained from the brain of HIV-1 infected patients, could limit some of the conclusions about viral integration and its features (Cochrane et al. 2022; Rai et al. 2020). To ensure that C20 represents microglia on our studies, we have also compared its chromatin accessibility, gene expression, and H3K27ac with the ones observed in primary microglia **(Appendix J-L)** (Gosselin et al. 2017; Nott et al. 2019). These comparisons aimed to show that C20 resembles primary microglia cells in both epigenetic and transcriptomic features. While the studies on microglia alone could prove meaningful for the understanding of the effects of HIV-1 infection in the brain, Alvarez-Carbonell and colleagues showed that the effect of other cells, like neurons, in the HIV-1 infected brain should not be disregarded (Alvarez-Carbonell et al. 2019). In patients, it is possible that HIV-1 affects only a small amount of microglial cells, as suggested by studies performed on cerebrospinal fluid (Farhadian et al. 2018). A recent analysis of viral integration in the brains of HIV-infected patients with encephalitis recovered 1,221 IS, mostly in microglia cell clusters (Plaza-Jennings et al. 2022). Of the microglial cells found, 4.8% harboured HIV-1 proviral sequences (Plaza-Jennings et al. 2022). Although the C20 cell line holds its limitations for the study of infection in the brain, mostly in terms of proliferation rate and viral replication, its application in integration research could still be meaningful (Rai et al. 2020; Garcia-Mesa et al. 2017).

Here, we explored the differences and similarities between IS found on the microglial cell model with IS from CD4+ T cells and MDMs, known cell targets and latent reser-

voirs for HIV-1. Interestingly, the genomic features of integration between these three cell types are similar, albeit CD4+ T cells and microglia share more similarities than microglia and MDMs. This observation is surprising, given that MDMs and microglia are both monocytic-lineage cells (Kok et al. 2016). Other unidentified factors could influence IS selection in these cell types. Moreover, the IS set size on MDMs was small, limiting any definitive conclusions (Kok et al. 2016). Commonly to CD4+ T cells, we observed that there is a tendency for highly-expressed genes to be targets of integration. This falls in line with the multiple studies linking IS hotspots to open chromatin and active transcription (Singh, Bedwell, and Engelman 2022; Lucic et al. 2019; Wang et al. 2007). As most genes targeted by IS are highly expressed, GO results could be biased by the frequent occurrence of essential genes, although this did not seem to be the case. Frequently targeted genes (like *NPLOC4* or *RPTOR*) were alternatively long genes with multiple and long intronic portions, suggesting that while the expression level is a critical factor in IS selection, intronic integration is also favored, as observed in other cell types (Singh et al. 2015). This tendency can be due to a high intron/exon size ratio or other spatial features, as suggested by others (Anderson and Maldarelli 2018; Singh, Bedwell, and Engelman 2022).

In microglia, HIV-1 integration gene targeting might also be influencing immune function, much like what has been observed for human papillomavirus or hepatitis B virus, where integration is seen as a player influencing cell proliferation and immune response (Linden and Jones 2022). These viruses were also shown to alter chromatin upon integration (Karimzadeh et al. 2022; Linden and Jones 2022). Thus, the abundance of chromatin-related gene sets among the GO results for IS-targeted genes could imply a opportunistic relationship between HIV-1 and epigenomic modifications (Lange et al. 2020).

2.5.2 Epigenomic features as determinants of HIV-1 integration in microglia

We explored the host chromatin features influencing HIV-1 integration in microglia and compared these with the ones observed into CD4+ T cells. To do this, we generated a set of "phantom" IS, comparable to real IS both in chromosome distribution and distance to genes. This procedure helped outline distinctive epigenomic features of the IS in the microglial cell model individually. To understand the impact of these features in combination, we have used two distinct scales: a smaller nucleosomal scale, resulting in ChromHMM-derived chromatin states, and a larger scale, performed by NMF to define IS-permissible loci. These two approaches aimed to cover both local and distal influences of the IS selection.

In microglia, we found that IS are typically found in H3K36me3-enriched loci, both distally and locally. Around the ISs, the profiles in active histone modifications (particularly H3K27ac) mirror chromatin accessibility, although the same cannot be said of H3K4me1. In CD4+ T cells, IS were strongly associated to the H3K27ac-enriched SEs, but an equal tendency was not observed in microglia (Lucic et al. 2019). Although some IS could be found in the NMF-derived windows most associated to SE, microglial IS were instead co-localising almost strictly with H3K36me3, independently of gene expression level. This apparent microglia-specific tendency to strongly favour introns of highly expressed genes, H3K36me3-marked, or genic enhancers rather than H3K27ac-marked, fully active, open chromatin regions could hint on the success of HIV-1 integration and latency in microglia, as it would allow the viral proviruses to remain unnoticed for longer, albeit not fully repressed. The impact of LEDGF/p75 and its role as an host factor for the HIV integrase can not be ruled out as impactful driving IS selection into H3K36me3-enriched regions in microglia.

H3K27me3, H3K9me2, and H3K9me3 are markers of non-permissible windows and are avoided by the virus, as in T cells (Wang et al. 2007). However, H3K27me3 has been linked to latency establishment, as it helps build a heterochromatin-like landscape, more suitable for proviral repression (Lange et al. 2020; Friedman et al. 2011). H3K9me2

has also been linked to latency establishment before, making its occurrence with some IS also not surprising (Imai, Togami, and Okamoto 2010).

Although hard to quantify, only a small fraction of CD4+ T cells persist in latency (Crooks et al. 2015). In brain, more research is necessary to understand the maintenance of latency, its interplay with chromatin, and impact in proviral reactivation. Intact proviruses found in the brain imply that the brain retains a competent HIV-1 reservoir in spite of ART (Cochrane et al. 2022). Closed chromatin serves as a latency sustainer, and KD of Polycomb machinery associated to H3K27me3 maintenance has been shown to reactivate silenced HIV-1 proviruses (Méndez et al. 2018; Friedman et al. 2011). In microglia, we found a small fraction of NMF windows linked to closed chromatin (mostly in signature 2 and 3) overlap IS, making it more likely that these IS would be mantained as latent proviruses. However, it remains speculative how the histone modification landscape changes after integration in microglia.

2.5.3 Effects of HIV-1 integration in chromatin in microglia

Through ATAC-seq, we assessed the changes in the chromatin accessibility of microglia upon HIV-1 integration. Here, we have also observed that the host chromatin is affected in the microglial cell model, as suggested by studies in CD4+ T cells (Jefferys et al. 2021). When assessing chromatin accessibility of latent, active, and uninfected cells, we did not observe very strong differences. A comparable study on CD4+ T cells showed that the latent provirus is characterised by a reduced accessibility in comparison with the actively infected cells (Jefferys et al. 2021). However, some of the results obtained in the latent condition can be influenced by the smaller amount of cells in this state and by the slightly lower read quality.

Through TF footprinting, we observed a distinct change in the predicted binding of multiple TFs between the three cell states, such as CTCF, FOS, NFKB1, IRF1, FOXP1, or MAFF. The TFs included for footprinting analysis were a part of a microglia-derived TF signature, indicating that the progression to active and latent state of infection could strongly impact the transcriptional programs of microglia (Gosselin et al. 2017).

TFs from the NF-KB family have been shown to be chromatin modifiers which repress the proviral sequence, and it is possible that some of the TFs found here play similar roles (Chan and Greene 2011). CTCF, an important factor in chromatin looping, has been associated to latency establishment in CD4+ T cells before (Jefferys et al. 2021). Nevertheless, it is important to note that these results are not a direct assessment of TF activity, as TF footprinting is an indirect and theoretical measurement of TF binding, dependent on data and motif quality, and the individual TF intricacies (Bentsen et al. 2020). To be certain that these changes are being accompanied by corresponding TF or gene regulatory network alterations, integration of ATAC-seq footprinting together with RNA-seq or ChIP-seq targeting specific TFs would be required.

CTCF footprinting is often extremely reliable, so we compared individual motifs and their predicted dynamics between the three cell conditions in microglia. CTCF appears to be enriched in latently infected cells, but it gets depleted on the active condition. The roles of CTCF as a modulator of TAD formation, an insulator, and its potential ability to repress the provirus suggests that this TF serves as an important TF for HIV-1 latency, independently of host cell (Splinter et al. 2006; Jefferys et al. 2021).

The depletion of CTCF later allowed the assessment of the effect of this TF in HIV-1 integration. CTCF depletion is known to cause changes in the 3D chromatin architecture (Khoury et al. 2020). In microglia, CTCF-KD leads to a subtle drop in the IS numbers. IS were depleted at the TAD boundaries (associated to healthy Neu- cells), implying these are targeting other regions in the CTCF-KD. These regions could be either newly formed TAD boundaries or, in case the presence of CTCF is the driving force of alterations, random genomic locations. Nevertheless, these observations remain speculative without further research on the interplay between CTCF, IS, and 3D chromatin.

2.5.4 Other players involved in TAD boundary establishment

The RF model we built here to find other TFs linked to higher-order chromatin structures serves as an important starting point for a unbiased larger-scale analysis on the multiple regulatory players involved in the formation of 3D chromatin elements. More-

over, the usage of ATAC-seq here makes this assessment cost-effective, as it avoids the generation of Hi-C and ChIP-seq for all the TFs. Nevertheless, this approach extends on the same limitations of TF footprinting, being dependent on the quality of each motif and specificities of TFs. The relationship between TADs and other genomic elements has been assessed by others, and addition of TFs, like ZNF143 or YY1, found at the vicinity of TAD boundaries, to the RF model I generated could be informative (Hong and Kim 2017).

Beyond CTCF, training of this model highlighted the potential roles of IRF1, REST, ZEB1, or Tcf12 in the formation of TADs. The model is heavily dependent on chromatin accessibility itself, hinting on the possibility of using ATAC-seq alone to map TADs (Tan et al. 2023). Extreme class imbalance presents itself as a significant challenge to the model conception, as only approximately 0.279% of the training data was part of the positive class (*TAD boundary*). Other advanced methods could be used for training, like neural networks, in order to detect underlining non-linear effects between predictors or other complex interactions.

Developing computational methods as alternatives to Hi-C in the prediction of chromatin structure can be very useful, as is allows for a cost-effective and swift way to assess higher-order chromatin. Histone modifications have been used in the prediction of TAD boundaries before, although the existing methods are not optimal. *TAD-Lactuca* is a supervised method which uses histone modifications and DNA sequence to predict the location of TAD boundaries (Gan et al. 2019). However, this approach uses simply a limited subset of pre-selected regions during training, making its application to other datasets biased and challenging. ATAC-seq can be used as a baseline, as it assesses chromatin accessibility and it has already been used in other tools, together with CTCF ChIP-seq and DNA sequence (Tan et al. 2023). Later on, we have used histone modifications to generate Bayesian networks and characterise the epigenomic dynamics between histone modifications and CTCF at TAD boundaries. Although these have highlighted the confirmed central role of CTCF, they open questions regarding other known and unknown players which could have an impact in chromatin loop and TAD formation, like

H3K36me3 or H3K27me3 (Hong and Kim 2017).

2.5.5 3D chromatin dynamics in HIV-1 integration

TAD boundaries are part of a highly dynamic environment, characterised by open chromatin, multiple TF binding sites, housekeeping genes, and TSSs (Hong and Kim 2017; Dixon et al. 2012). We have found that IS fall into the vicinity of TAD boundaries, implying that IS could disturb the 3D chromatin structure. This is accompanied by the tendency of H3K36me3 to be enriched at these locations. Thus, it is hard to identify the driver of integration between H3K36me3 and TAD boundaries. Nevertheless, co-occurrence of IS at these locations represents a highly disruptive potential to 3D chromatin structure and, ultimately, gene regulation (Dixon et al. 2012). In human papillomavirus, a comparable event induces disregulation of gene expression and genome interactions in the host cell (Groves et al. 2021).

While we have infered that IS could be disrupting the host chromatin of HIV-1 infected cells in the two infected states in comparison with the healthy cell from changes in CTCF binding and proximity to the TAD boundaries , it is not clear whether this is directly caused by the IS itself, or indirectly by the changes caused by infection in its latent or active form, or immune response leads to changes in the TADs (Plaza-Jennings et al. 2022). To fully tackle this question, a recently developed tool (*C.Origami*), allows for the reconstruction of Hi-C interaction matrixes using DNA sequence, CTCF ChIP-seq, and ATAC-seq alone (Tan et al. 2023). This approach can thus be applied to generate a Hi-C interaction matrix for each IS obtained in this study, and thus predict the potential structural consequences of IS when found in diferent genomic structures (introns, exons, promoters, among others).

2.6 Chapter summary

In conclusion:

- IS in microglia are similar in location and features to IS in CD4+ T cells;

- IS are located in highly transcribed regions and the vicinity of H3K36me3-enriched domains both inside and outside gene bodies;

- Infection states (latent or active) tend to alter binding of TFs, such as CTCF, which are linked to chromatin organisation;

- IRF1, REST, ZEB1 or Tcf12 are TF which potentially hold important roles on the maintenance of TAD boundaries and would be worth further research;

- IS are usually located into the vicinity of TAD boundaries;

- CTCF seem to have a connection with IS which was previously undescribed.

Chapter 3

Characterisation of distinct CpG island methylator phenotypes in glioblastoma

3.1 Motivation

Cancer leads to alterations in the DNA methylation of affected cells. While global hypomethylation at single CpGs is frequent, CGIs are often found to be hypermethylated in multiple cancer types, such as colorectal cancer, gastric cancer, or glioma (Toyota et al. 1999; Chang et al. 2006; Malta et al. 2018). CGIs are at sites of transcription initiation and their dysregulation impairs gene expression. Promoters of tumour suppressor genes can be inactivated through hypermethylation, as it was found in colorectal cancer, emphasising the importance of these alterations in the formation of tumours (Toyota et al. 1999).

Hypermethylation at CGIs, which is known as CIMP, can be caused by mutations in genes encoding proteins holding epigenetic functions, such as histone methyltransferases, or microsatellite instability (Yates and Boeva 2022). In colorectal cancer, CIMP has been

associated to microsatellite instability and BRAF mutations (Weisenberger et al. 2006). CIMP can also be caused by IDH mutations (Turcan et al. 2012). The mutated IDH enzymes (*IDH1* and *IDH2*) lead to the production of a metabolite which competes with histone demethylases and the DNA demethylase TET2, ultimately leading to impaired histone and DNA demethylation (Noushmehr et al. 2010; Turcan et al. 2012; Lu et al. 2012). Origins of CIMP in other tumours are often unknown. It has been suggested that aging-like epigenetic abnormalities could also play a role into the origins of CIMP and it is known that the enrichment in H3K27me3 can make CGIs more prone to hypermethylation (Tao et al. 2019; Court and Arnaud 2017).

GBM is a common aggressive brain tumour which is linked to a poor prognosis and long-term survival. GBM is classified into 4 distinct molecular subtypes: IDH (characterised by the existence of the IDH mutations), MES (or *mesenchymal*), RTK-I (previously named *proneural*), and RTK-II (previously named *classical*) (Wang et al. 2017; Verhaak et al. 2010; Sturm et al. 2012). The subtypes are characterised by epigenetic and molecular differences. IDH-mutant subtype in GBM has been linked to CIMP, but so far none of the remaining subtypes have. In this work, we aimed to characterise the CGIs in the four GBM subtypes according to DNA methylation and other epigenetic features to understand the impact of the epigenetic variability underlining the GBM subtypes. To do this, we have integrated data from DNA methylation, histone modification, and expression obtained on the 4 GBM subtypes.

Cancer is characterised by the existence of cancer stem cells, which can drive tumour formation and recurrence (Dirks 2010). These cells imply a malignant transformation of normal stem/progenitor cells before or during tumour formation (Dirks 2010). Cancer stem cells has been observed in GBM, and able to propagate tumours between hosts, self-renew, and produce differentiated progeny (Galli et al. 2004; Couturier et al. 2020). It has been shown that most differentiated cells are least susceptible to tumorigenesis in GBM, but its precise cell-of-origin is still a matter of debate. Different studies have pointed out that astrocytes, interneurons, neural stem cells, or other early progenitor cells could serve as cells-of-origin for GBM (Chen et al. 2020; Alcantara Llaguno et

al. 2019; Lee et al. 2018). In our analysis, we have included healthy cells as a comparison baseline, in order to understand if and how malignant transformation is related to epigenetic alterations. We also aimed to determine which epigenetic features before tumourigenesis are more likely to be affected by DNA methylation aberrations. To do this, I integrated epigenetic data from neural progenitors (NPs) and used these cells as a healthy counterpart to the GBM stem cell (Couturier et al. 2020).

In this chapter, I explored the DNA methylation landscape of all GBM subtypes and characterise a new CIMP, associated to the RTK-II subtype and independent from the IDH mutations. Using histone modifications and DNA methylation from both tumour and healthy cells, we compare RTK-II associated CIMP with IDH associated CIMP. At the progenitor states, I evaluate the capability of each CGI of becoming affected with any of the two CIMP when in tumour state, linking malignant methylation alterations with normal development. Further, I assessed the overlap between CIMP-targeted CGIs and downstream genes to identify functional effects of CIMP. I have also identified cell development trajectories which are more likely to become affected by the epigenetic alterations found in GBM.

3.2 Data

In this project, our three main goals were to characterise DNA methylation abnormalities over CGIs in the different GBM subtypes, understand how these alterations can affect gene function and normal cell development, and assess which features are linked to the occurrence of DNA hypermethylation before tumourigenesis. To understand this we integrated DNA methylation and histone modification data obtained from the different GBM subtypes and matching data from the NPs to assess (representing the epigenetic landscape before tumourigenesis). Further, we have used single-cell expression data from developing and adult brain to verify the potential impact of the DNA methylation alterations in brain development.

3.2.1 Glioblastoma

A published inhouse GBM dataset (Wu et al. 2020) is used to characterise the DNA methylation abnormalities on the CGIs and to determine CIMP. The dataset includes 60 samples obtained from patients classified under the four GBM subtypes (Wu et al. 2020). Description of the samples used for the analysis in this section is published (Supplementary Data 1 from Wu et al. (2020)). The dataset includes DNA methylation assayed through WGBS, ChIP-seq on histone modifications, and gene expression assayed through RNA-seq.

3.2.2 Healthy cells and tissues

DNA methylation and histone modifications from NPs were used as a healthy comparison baseline to assess which features are linked to the occurrence of CIMP before tumourigenesis. Processed DNA methylation (WGBS) for NP was obtained from *GSE156723* record (Choi et al. 2020). ChIP-seq data for histone modifications on NPs was obtained on ENCODE (ENCODE Project Consortium 2012; Luo et al. 2020).

To infer how these alterations can potentially affect normal cell development in the brain, we have used a brain cell development dataset (Kanton et al. 2019) and a adult brain dataset from the Allen Human Brain Atlas (Tasic et al. 2018), both obtained through single-cell RNA-seq.

3.2.3 Acute myeloid leukemia

Lastly, we have also used published DNA methylation data from AML to compare the CIMP-affected CGIs from GBM with another tumour type displaying a CIMP which is not caused by the IDH mutations.

3.3 Methodology

Sequencing data analysis

Inhouse histone modification (ChIP-seq on H3K27ac, H3K36me3, H3K4me3, H3K4me1, H3K27me3, and H3K9me3), DNA methylation (WGBS), and expression (RNA-seq) data was processed as published previously (Wu et al. 2020). Subtype classification used for publication was used just as described. ChIP-seq data from histone modifications from NP and HPSC was obtained on ENCODE in BigWig format (ENCODE Project Consortium 2012; Luo et al. 2020). DNA methylation (WGBS) from HPSC were also obtained on ENCODE as BED files (ENCODE Project Consortium 2012). Processed DNA methylation (WGBS) for NP was obtained from GSE156723 record (Choi et al. 2020).

Definition of CIMP-CGIs

CIMP definition in GBM was based on published thresholds (Issa 2004). CpG methylation was averaged over all CGIs (N=26,268). Then, CGIs were compared within their corresponding CIMP-related subtype (IDH or RTK-II) with CIMP-negative subtypes (MES and RTK-I) and normal brain. CIMP-CGIs must:

- Display low DNA methylation levels in normal brain (methylation beta $<.75$);

- Present a significantly higher DNA methylation in CIMP-subtype on subtype-specific testing with CIMP-negative subtypes (Kruskal-Wallis test, adjusted p.value < 0.001);

- Present a high DNA methylation methylation difference in comparison with CIMP-negative subtypes ($|$methylation beta difference$| > 0.2$).

CGI-based NMF analysis

NMF analysis was performed using butchR package (v1.0) on all CGIs (Quintero et al. 2020). The analysis was performed independently from CIMP status and it aimed to define CGI signatures and assess histone modifications linked to CIMP. For ChIP-seq, CGI-averaged outputs of *multiBigwigSummary* were used, along with CGI-based averages

from WGBS on CpGs (Ramirez et al. 2014). NMF computation was carried out over 10^4 iterations, 20 initializations, and rank factorization tested on a range of 3 to 10. Final factorisation rank, based on observation and metrics, was 4. The resulting H- and W-matrix were visualised as heatmaps with *ComplexHeatmap* (v2.6.2) (Gu, Eils, and Schlesner 2016).

On signature-uniquely assigned CGIs, signature assignment was performed using the default *SignatureSpecificFeatures* function in butchR (Quintero et al. 2020). When assigned by *SignatureSpecificFeatures* to more than one signature, CGIs were instead re-assigned to the signature with maximal W-matrix normalized exposure (when exposure ≥ 0.8). This way, all CGIs are assigned to at least one signature. If the W-matrix normalized exposure value on a given CGI is over this threshold in multiple instances, CGIs are considered multiply assigned.

Random forest classification for CIMP groups

To distinguish between CIMP groups (RTK2- vs IDH-CIMP, IDH-CIMP vs non-CIMP, and RTK2-CIMP vs non-CIMP) using CGI-based epigenomic and genomic features, a RF model was generated (9 features). When present (RTK2- vs IDH-CIMP), CIMP category overlap was excluded. Training was performed on 70% of CGIs per group while testing was perfomed on 30%. Model training was performed using the caret R package with 2000 trees, cross-validation (10-fold, repeated 5 times), and down-sampling for the smaller classes.

Generation of the bayesian networks

Bayesian networks were generated with summarised ChIP-seq and DNA methylation over the CIMP-affected (RTK2 or IDH) CGIs using the bnlearn R package (Scutari 2010; Nagarajan and Scutari 2013). Training was performed using bootstrapping.

Single-cell RNA-seq analysis

A scRNA-seq dataset obtained in stem cell-derived cerebral organoids over development was used to compare CIMP-affected genes between CIMP classes over different stages of celullar development (Kanton et al. 2019). In paralel, the same comparison was performed in adult brain using the data from the Allen Human Brain Atlas (Tasic et al. 2018). A CIMP module score was calculated for RTK2-CIMP and IDH-CIMP on all cells using Seurat package function *AddModuleScore* (Hao et al. 2021).

3.4 Results

3.4.1 Definition of CIMP in the RTK-II subtype

While DNA methylation alterations at CpGs are an important feature of many cancers, I focused on understanding CGI methylation alterations on GBM. Using a previously published WGBS dataset, we assessed the global DNA methylation landscape of all CGIs (N=26,268) in the four GBM subtypes: IDH (harbouring mutations on isocitrate dehydrogenase genes and linked to CIMP previously), MES, RTK-I (often associated to a *PDGFRA* gene amplification), and RTK-II (often connected to an *EGFR* gene amplification) (Noushmehr et al. 2010; Wang et al. 2017; Wu et al. 2020). We averaged CGI methylation over all subtype samples and used normal brain as a comparison baseline for healthy tissue **(Figure 3.1a)**. In GBM, we generally observed a CGI hypermethylation in relation to normal brain. CIMP has been documented in the IDH subtype, making the CGI hypermethylation observed on this subtype expected (Noushmehr et al. 2010). However, the CGI methylation levels of the RTK-II subtype are closer to the ones observed in IDH. On the other hand, MES and RTK-I seem closer to normal brain regarding their CGI methylation level.

We applied a published criteria to assess if CGIs are affected by CIMP in the RTK-II subtype (Issa 2004). The CIMP classification requires that DNA methylation at a high number of CGIs is significantly higher than both normal tissue (methylation beta in

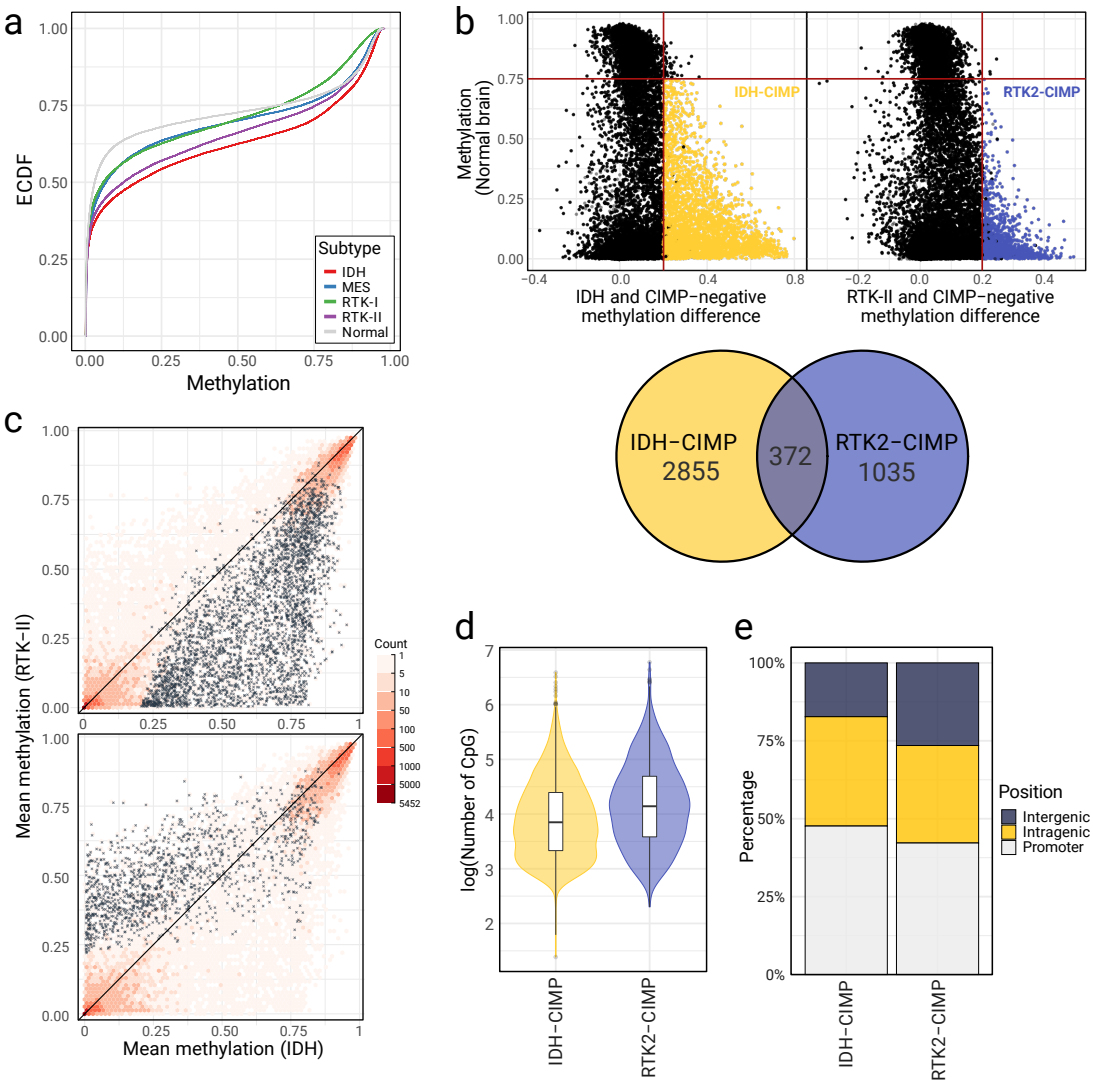


Figure 3.1: Definition and features of CIMP in RTK-II and in IDH subtypes. [a] Global CGI methylation by GBM subtype (in colours) compared to normal brain (in grey). [b] *(top)* Diagram of the thresholds (red lines) used to define CIMP-CGIs in GBM. Methylation beta difference of CIMP-associated subtypes is shown in the x-axis and methylation in normal brain is shown in the y-axis. Each dot represents one CGI. Coloured dots indicate IDH-CIMP (in yellow) and RTK2-CIMP (in blue). *(bottom)* Numbers of CIMP-CGIs by group and respective overlap. [c] Hexbin plot of the DNA methylation on RTK-II (y-axis) and on IDH (x-axis) subtypes. Amount of CGIs per hexbin is displayed on the left (red shades). CIMP-CGIs are shown as crosses (top: IDH-CIMP; bottom: RTK2-CIMP). [d] Distribution of CpG numbers by CIMP group (RTK2-CIMP as blue; IDH-CIMP as yellow). [e] Locations of CGIs associated to each CIMP group in relation to genes.

normal tissue must be $< .75$) and CIMP-negative subtypes (for GBM, MES and RTK-I; |methylation beta difference| > 0.2) (Issa 2004). Fulfilling this criteria, we obtained 3,227 CIMP-affected CGIs in IDH (henceforth referred to as IDH-CIMP) and 1,407 hypermethylated CGIs in RTK-II (henceforth referred to as RTK2-CIMP) **(Figure 3.1b)**. Unlike the remaining CGIs, RTK2-CIMP and IDH-CIMP CGIs presented a corresponding subtype-specific pattern of hypermethylation **(Figure 3.1c)**. Although overlapping, IDH-CIMP and RTK2-CIMP CGIs displayed distinct features. RTK2-CIMP usually harbour more CpGs (66±59 vs 88±82 CpGs) and tend to more often located outside gene bodies (one-sided Binomial test, p<2.2e-16) in comparison with IDH-CIMP **(Figure 3.1d-e)**.

3.4.2 Effects of CIMP in gene expression

Given the role of DNA methylation at CGIs in the gene expression, I evaluated the impact of these alterations in gene transcription at the corresponding tumour subtypes. I used a expression dataset obtained in the same samples used for WGBS (Wu et al. 2020). First, I performed a differential gene expression analysis between IDH and RTK-II samples and normal brain. Then, to exclude tumour-led effects, I compared IDH and RTK-II to CIMP-negative subtypes MES and RTK-I **(Figure 3.2a)**. I defined a set of CIMP-associated genes, which harbour CIMP-CGIs at promoter or intragenic regions, as these are the most directly affected by CIMP-driven transcriptional alterations. We observed that while not all CIMP-associated genes seem to be alterated, many are shown to be disregulated. Most disregulated CIMP-genes in both normal brain and CIMP-negative comparisons are shown in **Table 3.1**. Furthermore, RTK2-CIMP seems to be more associated to gene repression than IDH-CIMP, regardless of the affected CIMP-CGI location **(Figure 3.2b)**.

To understand the functional effects of CIMP in the RTK-II and IDH tumours, I performed a gene ontology analysis on CIMP-affected genes located in promoter regions **(Figure 3.2c)**. Targets of RTK2-CIMP are located upstream from developmental genes, as the ones associated to terms such as *pattern specification process*, *regionalization* or *cell*

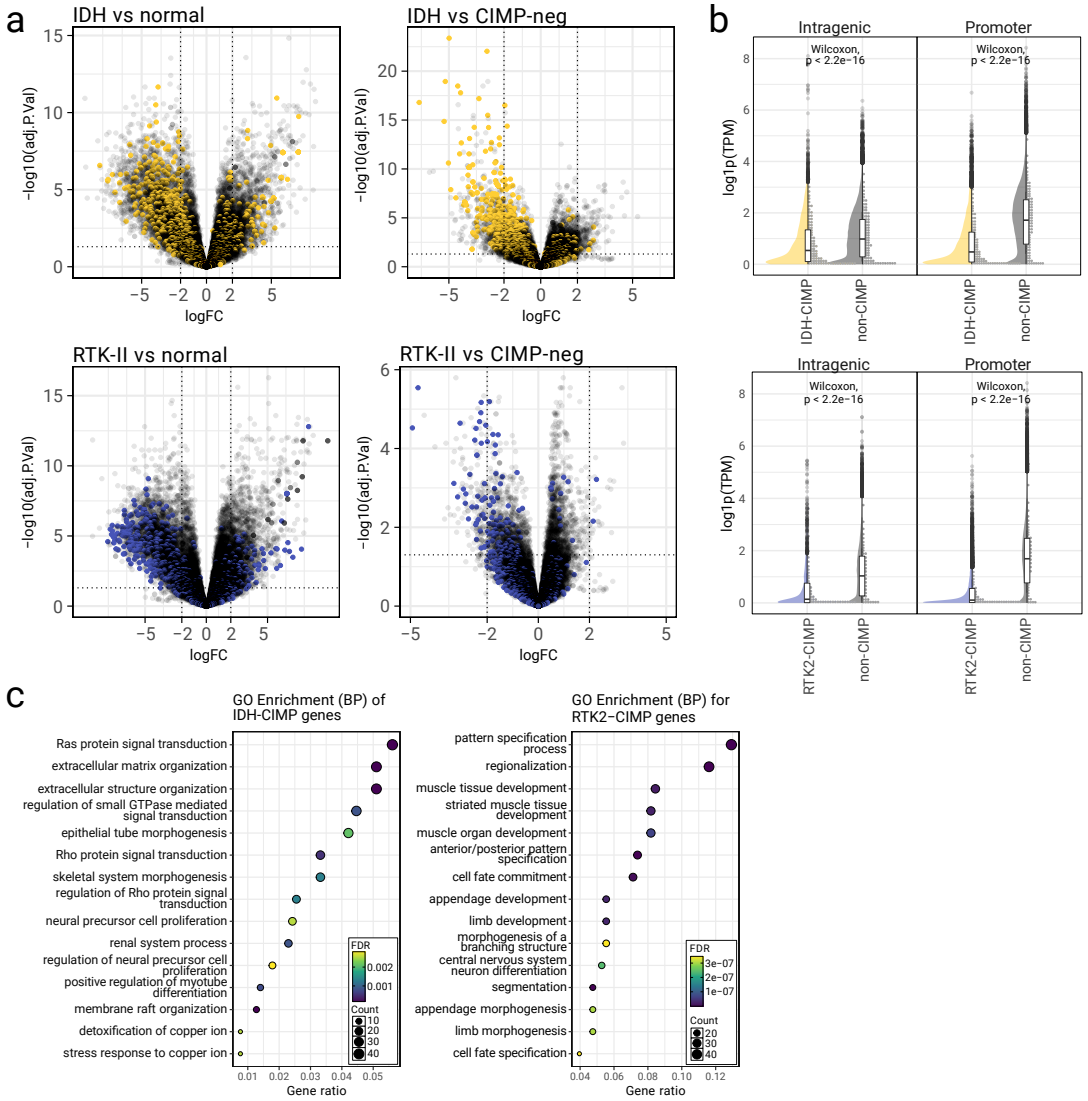


Figure 3.2: Effects of CIMP in gene expression. [a] Gene expression volcano plots of the differential gene expression analysis between each CIMP-associated subtype (top: IDH; bottom: RTK-II) and normal brain (left) or CIMP-negative subtypes (right). Log fold change (x-axis) and -log10 of adjusted p.value (y-axis) represent significance. Coloured dots (yellow: IDH-CIMP; blue: RTK2-CIMP) highlight CIMP-associated genes. [b] Raincloud plot of the expression distribution (in log1pTPM) for CIMP-associated genes and CIMP group (yellow for IDH-CIMP, blue for RTK2-CIMP, as indicated on the x-axis). Genes are separated according to location in relation to genes. The non-CIMP category (grey) represents genes harbouring non-CIMP CGIs in corresponding positions. [c] Gene ontology analysis (top 15 terms) results from CIMP-associated genes (on the *left*: RTK2-CIMP; on the *right*: IDH-CIMP. False-discovery rate (FDR) is encoded by dot colour. Gene number from the set in each term is proportional to dot size.

Table 3.1: Top 10 most downregulated CIMP-genes (intersection of CIMP-negative and normal brain comparisons)

Gene symbol	Gene description	log fold change	adj. p-value
IDH-CIMP alterated genes			
DES	desmin	-6.618	1.594e-17
AQP5	aquaporin 5	-5.259	1.385e-15
RBP1	retinol binding protein 1	-5.197	1.14e-19
RARRES2	retinoic acid receptor responder 2	-5.024	2.476e-11
FBXO17	F-box protein 17	-4.986	4.246e-24
PDLIM4	PDZ and LIM domain 4	-4.896	1.309e-08
LECT1		-4.618	9.032e-08
TOM1L1	target of myb1 like 1 membrane trafficking protein	-4.514	3.319e-19
C2orf70		-4.417	2.891e-12
NSUN7	NOP2/Sun RNA methyltransferase family member 7	-4.364	1.648e-18
RTK2-CIMP alterated genes			
SFRP2	secreted frizzled related protein 2	-3.3	0.0007682
NHLH2	nescient helix-loop-helix 2	-3.154	0.001694
SMOC1	SPARC related modular calcium binding 1	-3.058	2.276e-05
POPDC3	popeye domain containing 3	-2.928	0.001075
VAX1	ventral anterior homeobox 1	-2.854	0.002451
KY	kyphoscoliosis peptidase	-2.831	0.001417
BEND4	BEN domain containing 4	-2.73	0.006122
KCNH1	potassium voltage-gated channel subfamily H member 1	-2.676	0.0004523
NETO1	neuropilin and tolloid like 1	-2.592	0.003053
TRAM1L1	translocation associated membrane protein 1 like 1	-2.483	6.12e-05

fate commitment, while IDH-CIMP are located upstream from genes related to cellular processes and signalling, such as *Ras protein signal transduction* or *regulation of small GTPase mediated signal transduction.*

3.4.3 NMF-based assessment of CGI signatures and effects on CIMP

After assessing the effect of CIMP in gene expression, we aimed to survey the epigenetic crosstalk of DNA methylation with other epigenomic modifications. Different histone modificatins have so far been associated to CIMP, such as H3K27me3 and H3K4me3

(Court and Arnaud 2017; Dunican et al. 2020). To unbiasedly assert epigenetic patterns linked to CIMP, I started by combining epigenomic data from both DNA methylation and histone modifications on the GBM subtypes. I used ChIP-seq on histone modifications obtained from the GBM tumours analysed so far (Wu et al. 2020). The dataset includes activating (H3K27ac, H3K4me1, and H3K4me3), gene transcription/body (H3K36me3), and repressive histone marks (H3K27me3 and H3K9me3). To include a baseline for comparison, the same biological targets obtained from NP cells were also included, as this is meant to be a close healthy counterpart to the GBM cell-of-origin. With all the NP and GBM subtypes data, I generated a combined CGI-based matrix (44 x 26,268) which was then reduced into unique epigenetic signatures using a NMF-based decomposition **(Figure 3.3a-b)**.

Given its features, NMF is able to recognise the complex epigenetic crosstalk between the histone modifications and DNA methylation. The NMF-based analysis resulted into 4 unique CGI-based chromatin signatures (Sig.1: 14.26%, Sig.2: 22.88%, Sig.3: 19.0%, Sig.4: 30.21%, and 13.60% multiply assigned). Multiply-assigned CGIs showed strong exposure ($\geq$ 0.80) to more than one CGI signature, implying these are highly variable **(Appendix M)**. Signature 1, possibly representing CGIs associated to poised enhancers in both GBM and NPs, was characterised by a sharp enrichment in H3K4me1, particularly in GBM. On the other hand, signature 2 is associated to both DNA methylation and H3K36me3, implying an association to intragenic CGIs amidst transcribed gene bodies. Signature 3 exhibits an enrichment of repressive histone modifications H3K27me3 and H3K9me3, albeit also a small enrichment for H3K4me1 in NPs. Signature 3 CGIs could be located into heterochromatic regions or associated to a bivalent CGI state. Lastly, signature 4 seems denotative of promoter CGIs of actively transcribed genes, given its association to active marks H3K27ac and H3K4me3. Some CGIs signatures seem stable on both GBM and the NPs, as signature 4. However, other signatures present more variability between tumour and healthy cells, as signature 1 and 3. These highly variable signatures are also shown to be more often targeted by CIMP **(Figure 3.3c)**. RTK2-CIMP massively targets CGIs assigned to signature 3, while IDH-CIMP distributes over

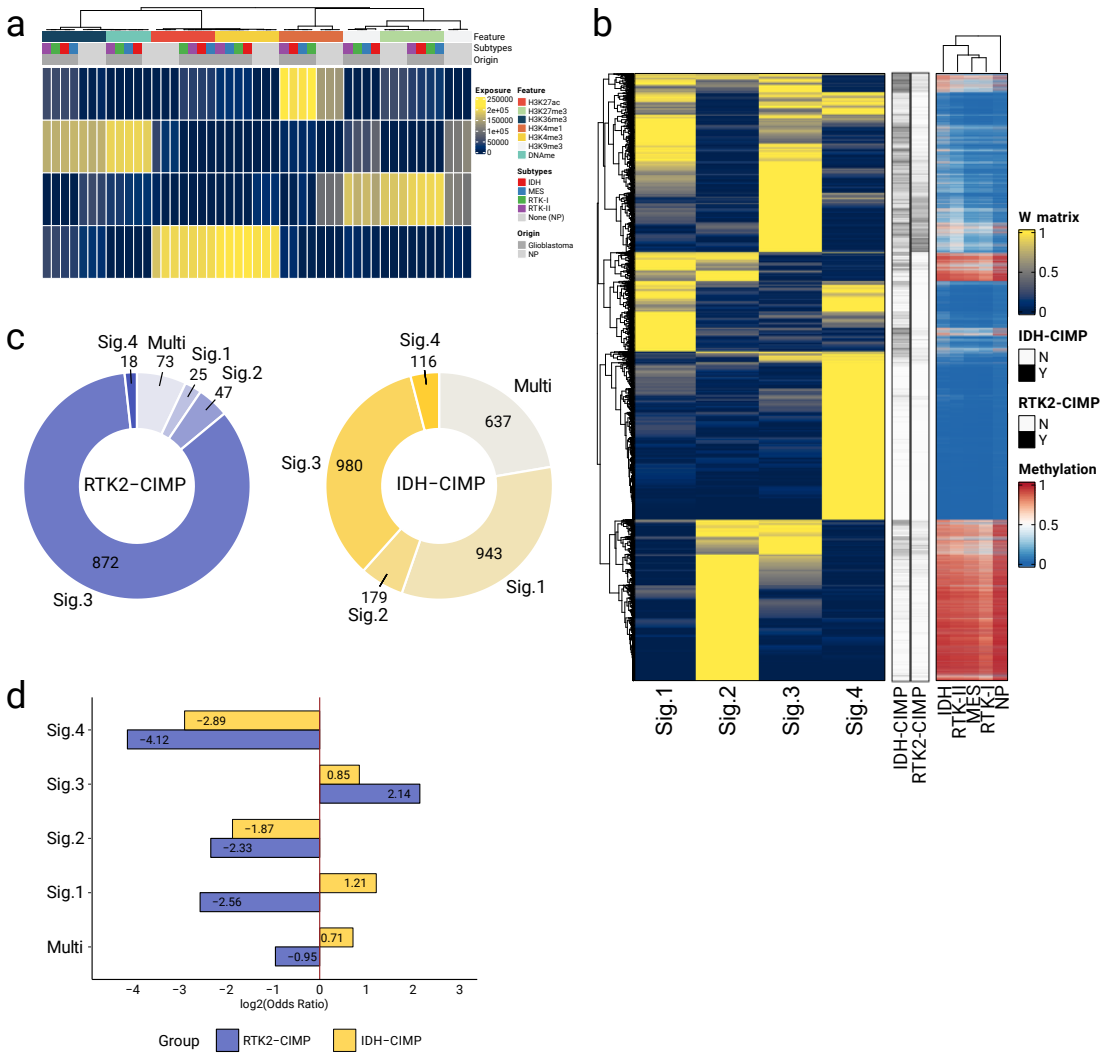


Figure 3.3: Chromatin signatures of CGIs in GBM and NPs. [a] NMF H-matrix of the signature exposures obtained from all CGIs for GBM and NPs (grey annotations). Gradient indicates exposure/contribution (yellow as most contributing to navy as least contributing) of each feature (column) to each signature (row). Each epigenomic feature and its origin (GBM subtype or NPs) is annotated above the heatmap. [b] NMF W-matrix *(left heatmap)* of each CGI (row) exposure to each signature (column) and CGI DNA methylation by condition *(right heatmap)*. Gradient can be interpreted as signature assignment *(left heatmap)*. DNA methylation is represented from high (red) to low (blue). Middle bars indicate whether each CGI is targeted by CIMP in IDH (IDH-CIMP) or RTK-II (RTK2-CIMP). [c] CIMP-CGIs assignment to the NMF-derived CGI signatures. [d] Log2 odds ratio representing the deviation from expected CGI signature assignment (red line) on RTK2-CIMP (blue) and IDH-CIMP (yellow) CGIs.

signature 1, 3, and multiply-assigned CGIs. Signature 3 is actually more targeted than expected by RTK2-CIMP (log2 odds ratio = 2.14), as revealed by its signature log2 odds ratio **(Figure 3.3d)**. Similarly, it avoids highly active CGIs, characteristic of the predominant signature 4. IDH-CIMP also targets signature 1, 3, and multiply-assigned CGIs more than the remaining (log2 odds ratio = 0.85, 1.21, and 0.71 respectively), albeit less expressively. These results show that epigenomic variability related to histone modifications (mostly in H3K27me3, H3K9me3, and H3K4me1) in the CGIs can be linked to CIMP in GBM.

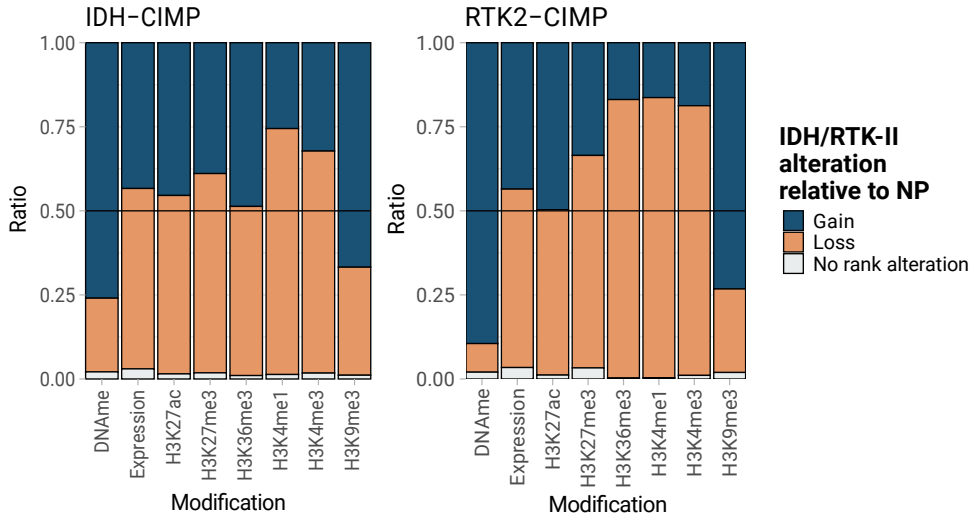


Figure 3.4: Rank-based comparison between NPs and GBM subtypes affected by CIMP within CIMP-CGIs. Loss, gain, or no rank alteration of epigenomic and transcriptomic features between NP and GBM subtypes IDH and RTK-II is shown for IDH-CIMP (*right*) and RTK2-CIMP (*right*) respectively.

When directly performing a rank-based comparison of the signatures between NPs and GBM (all subtypes combined), we observed that signature 1 CGIs tend to gain active transcription marks (H3K27ac, H3K36me3, H3K4me1, and H3K4me3) in GBM, although gene expression does not increase comparably. Signature 2 loses H3K27ac and H3K9me3, while signature 3, linked to CIMP, tends to lose H3K36me3, H3K4me1, and H3K4me3 **(Appendix N)**. Signature 4 tends to gain H3K36me3. Overall, alterations of multiple histone modifications within CGIs, particularly in H3K36me3, seem to accompany

tumourigenesis, even if these alterations do not plainly match alterations in gene expression. The same rank-based comparison was performed in IDH-CIMP and RTK2-CIMP to understand whether these CGIs displayed highly evident changes between healthy and tumour state (on their respective subtype) **(Figure 3.4)**. Unsurprisingly, both IDH- and RTK2-CIMP were associated to a gain in DNA methylation relatively to NPs. RTK2-CIMP CGIs display a sharp loss of H3K36me3, H3K4me1, H3K4me3, and H3K27me3, albeit the latter to a lower extent. On the other hand, IDH-CIMP was also associated to CGIs losing H3K4me1 and H3K4me3. Lastly, both IDH- and RTK2-CIMP CGIs gained H3K9me3 in tumour relatively to NPs. All these observations point to epigenomic differences in both CIMP groups, suggesting a distinct underlying mechanism.

3.4.4 Prediction of CIMP occurrence in GBM using epigenomic features of precursor cells

Given the tendency of CIMP to affect CGIs characterised by specific combinations of histone modifications which are variable between NP and tumour state, we speculated whether the CGI landscape in healthy NPs could be directly used to infer that a CGI is likely to become hypermethylated. To this end, I combined CGI epigenetic information obtained from NPs to define what distinguishes (i) IDH-CIMP from RTK2-CIMP, (ii) RTK2-CIMP from non-CIMP, and (iii) IDH-CIMP from non-CIMP in a healthy cell before tumourigenesis **(Appendix O)**. Using 9 features in total, I trained a RF model to classify the CGIs accordingly and identify the features that distinguish them in NPs **(Figure 3.5a)**.

The RF classifier is able to distinguish RTK2-CIMP from IDH-CIMP (AUC=0.796). The performance of this model is mostly reliant on DNA methylation and, to a lesser degree, repressive modifications H3K27me3 and H9K9me3 **(Figure 3.5b)**. The distinction between IDH- and RTK2-CIMP from the non-CIMP class was mostly dependent on H3K27me3 and H3K4me1 respectively. Next, I compared the classes in NPs and concluded that IDH-CIMP CGIs have already a higher DNA methylation and that RTK2-CIMP shows a higher enrichment in H3K27me3 when compared to IDH-CIMP **(Figure**

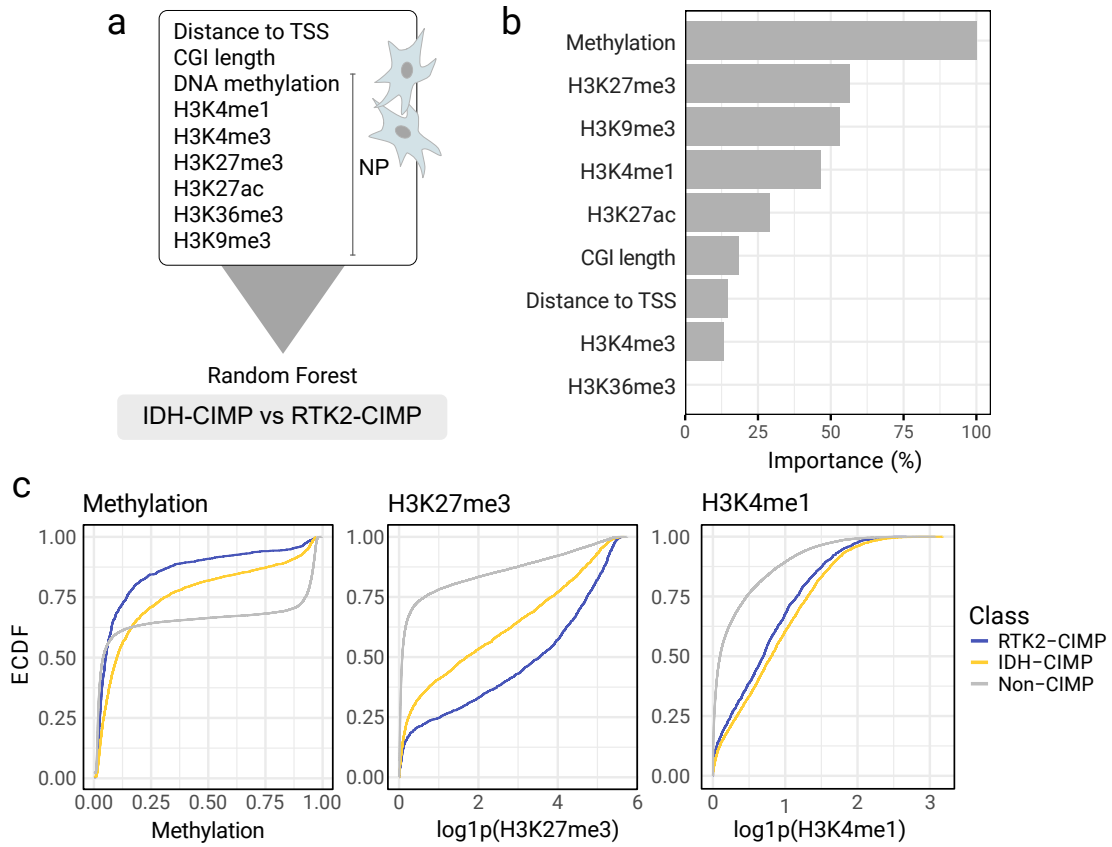


Figure 3.5: RF model for CIMP classification and features in NPs. [a] Diagram of the RF model and respective features used to classify CIMP-affected CGIs in the NPs. [b] Ordered importance (x-axis; most to least important) for features used on the IDH- vs RTK2-CIMP model as exported by *caret*. [c] ECDF of each CIMP class for DNA methylation, H3K27me3, and H3K4me1 in NPs. CIMP class is represented by colours and non-CIMP CGIs are shown in grey as comparison baseline.

3.5**c)**. Overall, CIMP-prone CGIs could be identified using histone modifications and DNA methylation before tumourigenesis on healthy cells. This further suggests that CIMP-prone CGIs hold epigenetic features recognisable from other CGIs.

To understand how the histone modifications might be dynamically driving the formation of CIMP-specific DNA methylation patterns, we trained and generated a bayesian network model on NP-derived data for RTK2-CIMP and IDH-CIMP independently **(Figure 3.6)**. Using a bootstrapped methodology, we observed that the formation of RTK2-CIMP could be deeply influenced by H3K27me3. H3K4me1 appears to be the driving force underlining IDH-CIMP. However, IDH-CIMP DNA methylation appears to influence other modifications itself, unlike its RTK2-CIMP counterpart.

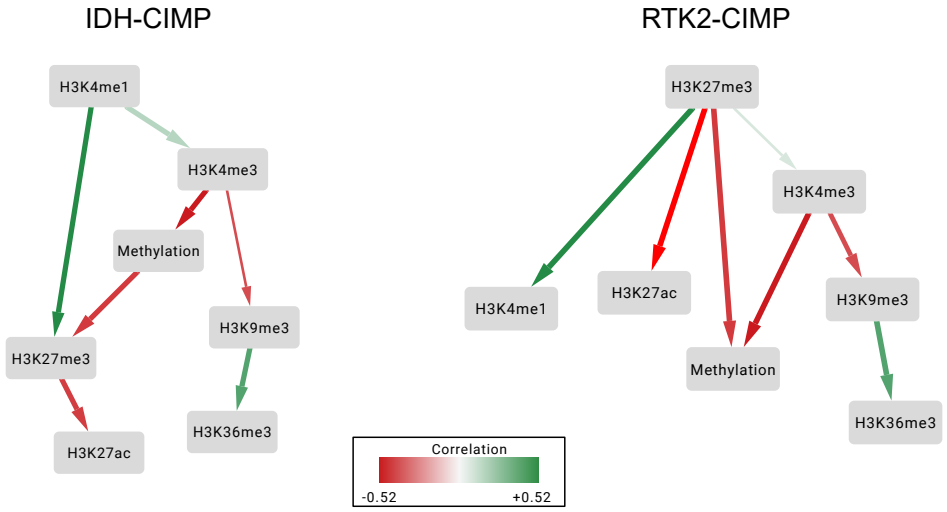


Figure 3.6: Bayesian network representations on epigenomic features of IDH- and RTK2-CIMP. Each epigenomic feature is represented by one node. Edges represent the existence and directionality of significant connections between features, while their colour (as represented in the scale) and width indicate correlation and strength of the connections respectively.

3.4.5 Association of CIMP with cell populations and differentiation tracks

As CIMP targets CGIs which are enriched in H3K27me3, an important histone modification for cell fate and differentiation, and often upstream from genes associated to

development **(Figure 3.2c)**, I assessed the possible impact of CIMP through normal brain development. Developmental genes are also characterised by bivalent states, associated to both H3K27me3 and H3K4me3, much like the ones being targeted by RTK2-CIMP (Michalak et al. 2019). I have used a published single-cell RNA-seq dataset from brain development to assign scores to CIMP-associated genes, pinpointing stages or cell types potentially affected by the effects of CIMP **(Figure 3.7a)** (Kanton et al. 2019). The dataset, obtained from organoids, was generated to infer differentiation trajectories from pluripotency into neuronal fates. We found that the expression of CIMP-associated genes in both RTK2- and IDH-CIMP is overall low **(Figure 3.7b)**. However, genes associated to RTK2-CIMP display a slightly higher expression in the more differentiated neuronal cells, both excitatory (Glutamatergic) and inhibitory (GABAergic) neurons **(Figure 3.7c)**. In contrast, IDH-CIMP does not seem particularly associated to any developmental stage in particular. Considering the absence of other differentiated non-neuronal cells in this dataset, I compared genes associated to RTK2-CIMP with two sets of adult cell markers, in order to infer if CIMP regions are found upstream or in the vicinity of genes important for neuronal function or maintenance (Couturier et al. 2020; McKenzie et al. 2018). Similarly, I observed that RTK2-CIMP overlaps mostly with neuron markers, while IDH-CIMP distributed broadly over different cell markers **(Figure 3.7d)**. Therefore, I repeated the previous analysis (in **Figure 3.7a-c**) with adult brain cells. As before, I have generated a CIMP score for each cell. These scores are represented in the t-distributed Stochastic Neighbor Embedding (t-SNE) plot found in **Figure 3.8a (right)**. Here, we have observed that the genes targeted by RTK2-CIMP are genes mostly active in neuron populations, regardless of their subtype **(Figure 3.8b)**. We have also assessed tumour cell composition using bulk RNA-seq deconvolution. In this approach, we have used cell markers obtained from brain populations to estimate which and how many cells of a single population could be present in the bulk RNA-seq GBM samples. RTK-II appears to be depleted of neurons in comparison with both normal brain and other GBM subtypes **(Figure 3.8c)**. Together, these results would imply that neuron fates could be repressed by RTK2-CIMP.

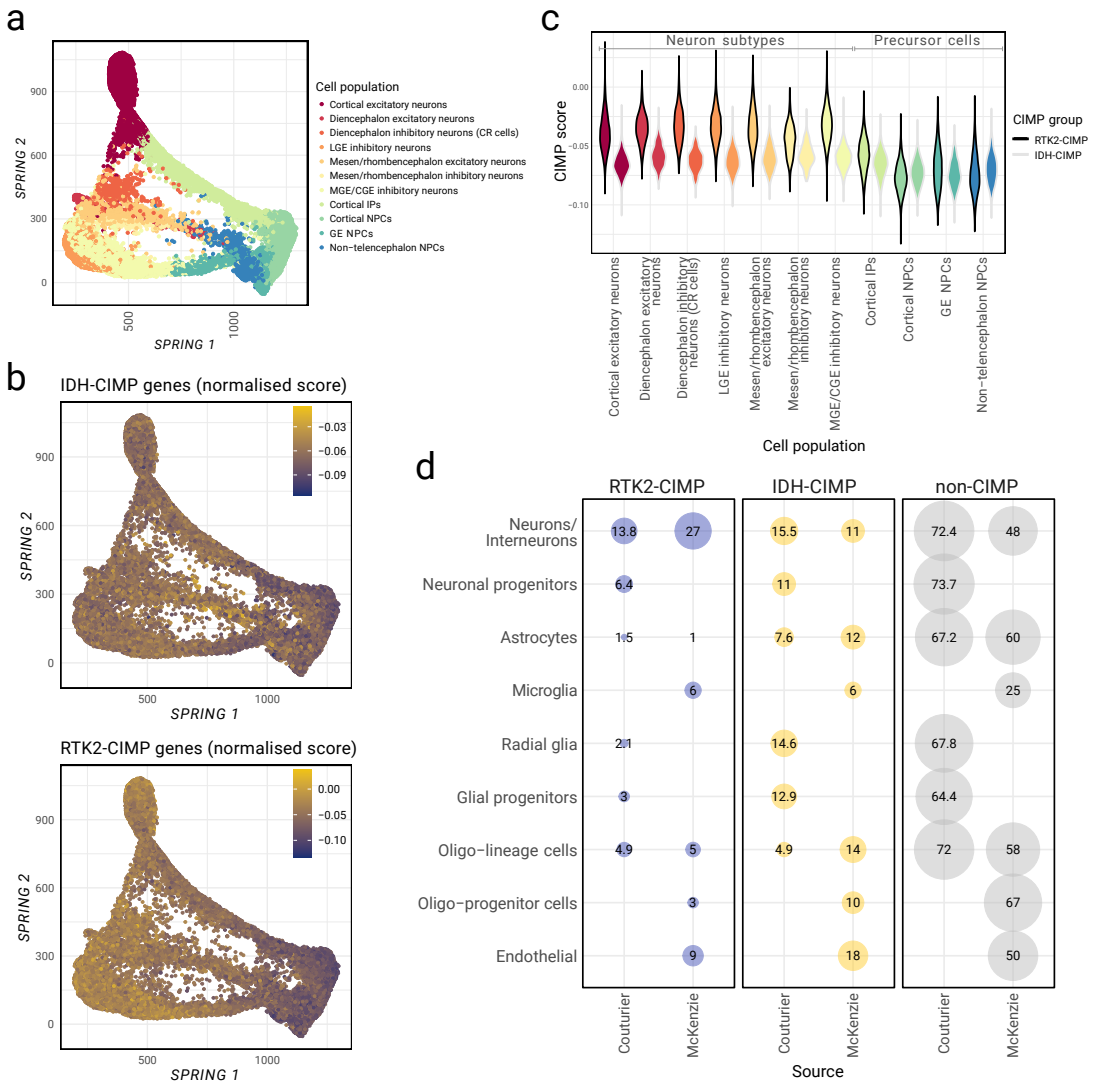


Figure 3.7: Locating CIMP effects into brain development. [a] SPRING reconstruction of all the cell populations present in the brain development dataset from Kanton et al. 2019. [b] CIMP-score derived from CIMP-associated gene expression in IDH-CIMP (*top*) and RTK2-CIMP (*bottom*) over the SPRING reconstruction. Scores are represented using the gradient (yellow to purple) in the top right corner. Non-neuronal cells are highlighted. [c] CIMP scores (y-axis) distribution by cell population (x-axis and fill colour as in panel **a**) and CIMP group. CIMP group is distinguished through the outline (black as RTK2-CIMP or grey as IDH-CIMP). Neuronal cells (*left*) were separated from precursor cells (*right*). [d] Percentage of cell markers (labelled and as dot size) found to be CIMP-associated genes in the three CIMP categories (panel titles and colours). Empty spaces denote cell type for which the gene set did not have markers for. Cell markers source is indicated on the x-axis.

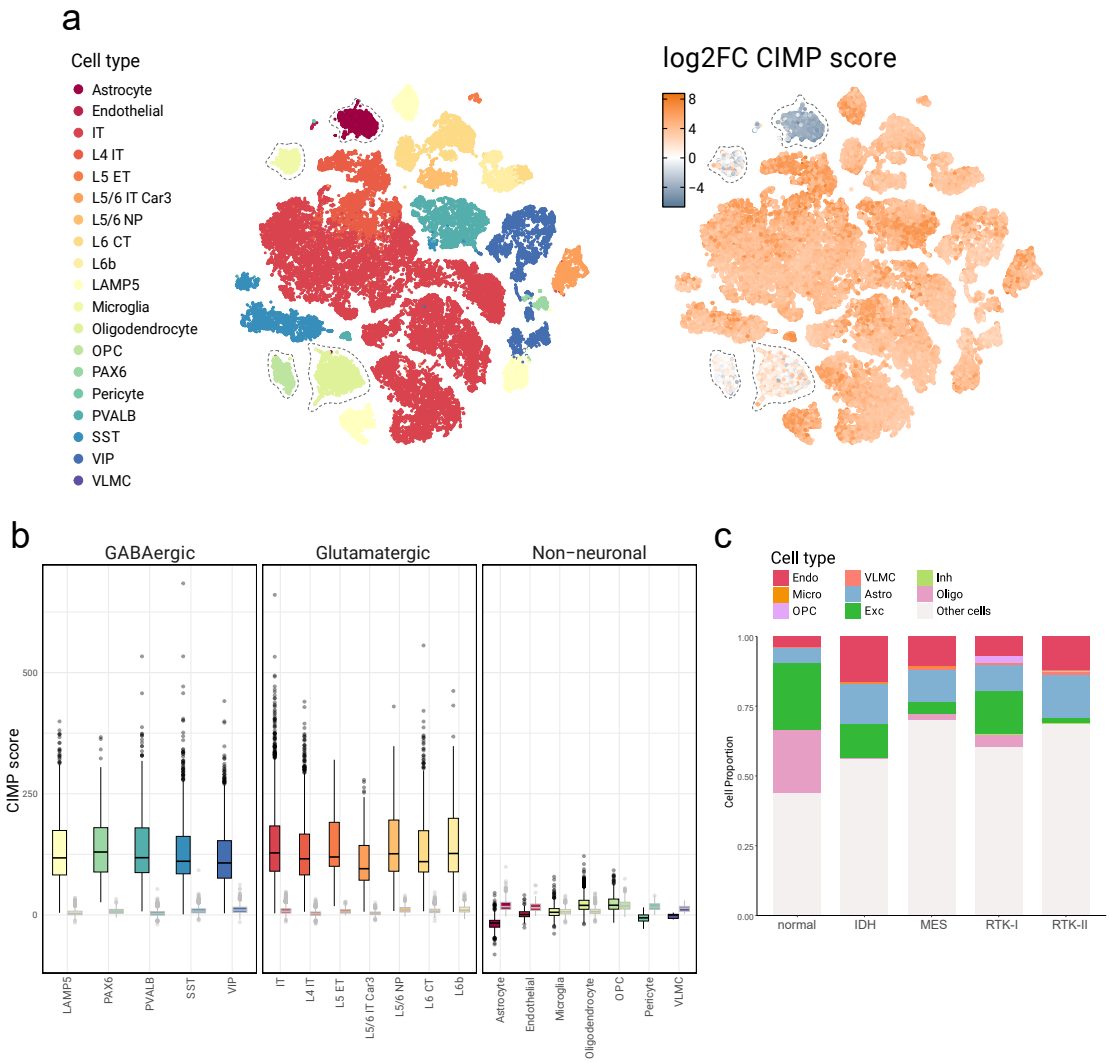


Figure 3.8: Assessing CIMP into adult brain cells. [a] t-SNE of all the cell populations present in the adult brain dataset from Tasic et al. 2018 (*left*). Log2 fold change of the CIMP-scores derived from CIMP-associated gene expression in IDH-CIMP and RTK2-CIMP over the t-SNE (*right*). Scores are represented using the diverging gradient (orange to blue, with 0 as white) in the top left corner. [b] CIMP scores (y-axis) distribution by cell population (x-axis and fill colour as in panel **a (right)**) and CIMP group. CIMP group is distinguished through the outline (black as RTK2-CIMP or grey as IDH-CIMP). The two types of neuronal cells (GABAergic or Glutamatergic) were separated from non-neuronal cells (facet). [c] Deconvolution of the RNA-seq into cell types for the samples of the four GBM subtypes and the normal brain (Wu et al. 2020). The latter analysis was performed by Lin Yang, generated using the EPIC tool, and based on marker genes obtained from the Allen Brain Atlas. (Endo = endothelial; Micro = microglia; OPC = Oligodendrocyte progenitor cell; VLMC = Vascular and leptomeningeal cells; Astro = Astrocyte; Exc = Excitatory neuron; Inh = Inhibitory neuron; Oligo = Oligodendrocyte).

3.4.6 Comparison with A-CIMP in AML

Having assessed the CGI landscape associated to the RTK2-CIMP and distinguished it from the IDH-CIMP, it is clear that RTK2-CIMP does not share the same origin and mechanism found on IDH-CIMP, as the RTK2-CIMP is not associated to IDH1/2 somatic mutations. Therefore, we looked for other tumours where CGI hypermethylation was also found and were not associated to IDH mutations. A CIMP with such features and not associated to mutations was described in acute myeloid leukemia (AML) (Kelly et al. 2017). This phenotype, termed A-CIMP, was also different from IDH-driven CIMP, which is also present in AML. When comparing both groups **(Figure 3.9)**, we observed that there is a substantial overlap between the IDH-CIMP in both tumour types and A-CIMP and RTK2-CIMP CGIs[6]. Notably, 3 CGIs are found on all CIMP groups. One of these CGIs is intergenic, while the remaining two are located within the *SPACA6* and *TBX1* gene bodies. The latter is associated to developmental processes (Moraes et al. 2005).

3.4.7 Tracing CIMP back to HSCs and other organs

Considering the traceability of CIMP through the developmental course, the connection with histone modifications, and the overlap between the A-CIMP and RTK2-CIMP groups, we assessed whether the common A-/RTK2-CIMP and IDH-CIMP targets in AML and GBM could be associated to any particular epigenomic features, just as RTK2-CIMP are. We ran these CGIs through a multi-tissue analysis tool, *i-cisTarget*, to assess whether there is a common association to any particular histone modification (Herrmann et al. 2012). Interestingly, we observed that common A-/RTK2-CIMP targets are strongly associated to H3K27me3, while common IDH-CIMP targets seem to be mostly associated to H3K4me1, H3K79me2, and H3K36me3 **(Figure 3.10)**. Considering a specific RTK2-CIMP epigenomic landscape is observed in parallel for multiple tissues, we wondered whether we could trace the common A-/RTK2-CIMP targets to earlier

[6] Analysis and definition of CIMP in AML was performed by Lin Yang (Molecular Pathology Research Center, Chinese Academy of Medical Sciences, Beijing, China), visiting PhD student at HDSU.

stages of differentiation. Therefore, I compared AML-specific A-CIMP, GBM-specific RTK2-CIMP, and A-/RTK2-CIMP intersecting CGIs in hESCs. Here, we observed that A-/RTK2-CIMP intersecting CGIs are already highly enriched into H3K27me3 in hESCs both in comparison with other CGIs which are not affected by CIMP and tumour-specific CIMP-CGIs.

3.5 Discussion

3.5.1 Epigenomics of the CIMP in RTK-II and IDH

While tumours are often associated to global epigenomic alterations like DNA hypomethylation or chromatin rearrangements, the hypermethylation of CGIs can be highly disruptive (Nishiyama and Nakanishi 2021; Plass et al. 2013). Here, we characterised CIMP in the GBM subtype RTK-II for the first time. We have also observed that this particular CIMP affects CGIs differently compared to the CIMP observed on the IDH subtype. In colorectal cancer, distinct forms of CIMP translate into differences in clinical outcomes, as CIMP-high tumours are often correlated with lower colon cancer-specific mortality rate (Ogino et al. 2009). A pan-cancer analysis restated CIMP as a survival factor, now in both GBM and low-grade glioma (Yates and Boeva 2022). It also revealed its correlation to specific tumour microenvironment features in other cancer types, such as macrophage regulation, lymphocyte infiltration (Yates and Boeva 2022). Thus, it is likely that the newly-described CIMP in RTK-II could hold clinical importance as well, making these findings meaningful for GBM prognosis.

Although characterised by multiple genetic aberrations, RTK-II is highly responsive to therapy and known for a better prognosis (Wu et al. 2020; Zhang et al. 2020). Nevertheless, genetic and epigenetic cancer-related changes alike result in an increase in the heterogeneity of the tumour, which is accompanied by clinical variability and other phenotypic alterations (Feinberg and Levchenko 2023; Hansen et al. 2011). This affirms the need of further studies into the origins of the CIMP found in this subtype, as epigenetic modifications or enzymes might serve as targets for therapy research.

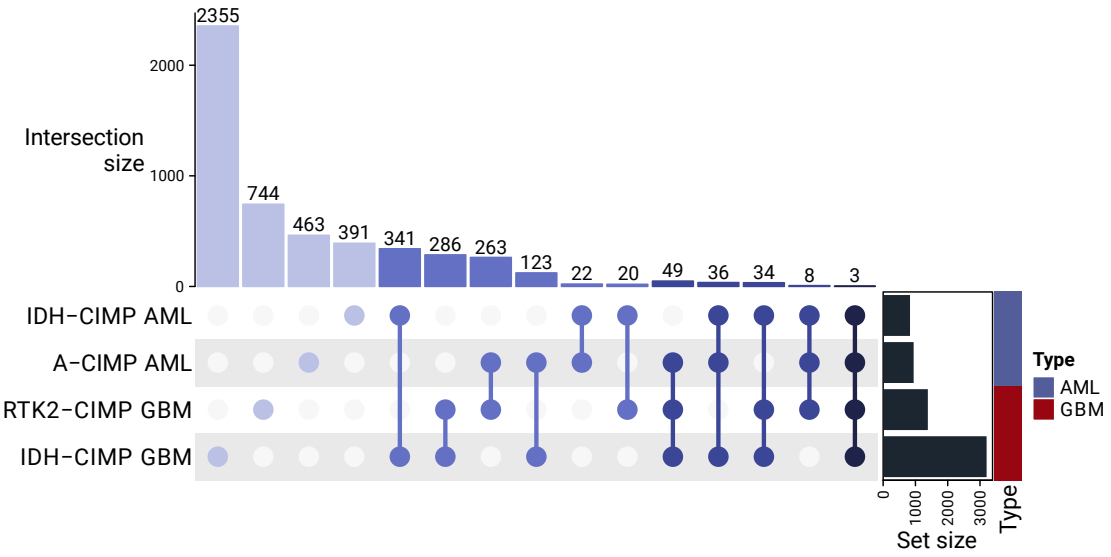


Figure 3.9: Comparison of CIMP in GBM with CIMP in AML. Upset plot representing the intersections between the CIMP associated to GBM (IDH- and RTK2-CIMP) to the one associated to AML (IDH- and A-CIMP).

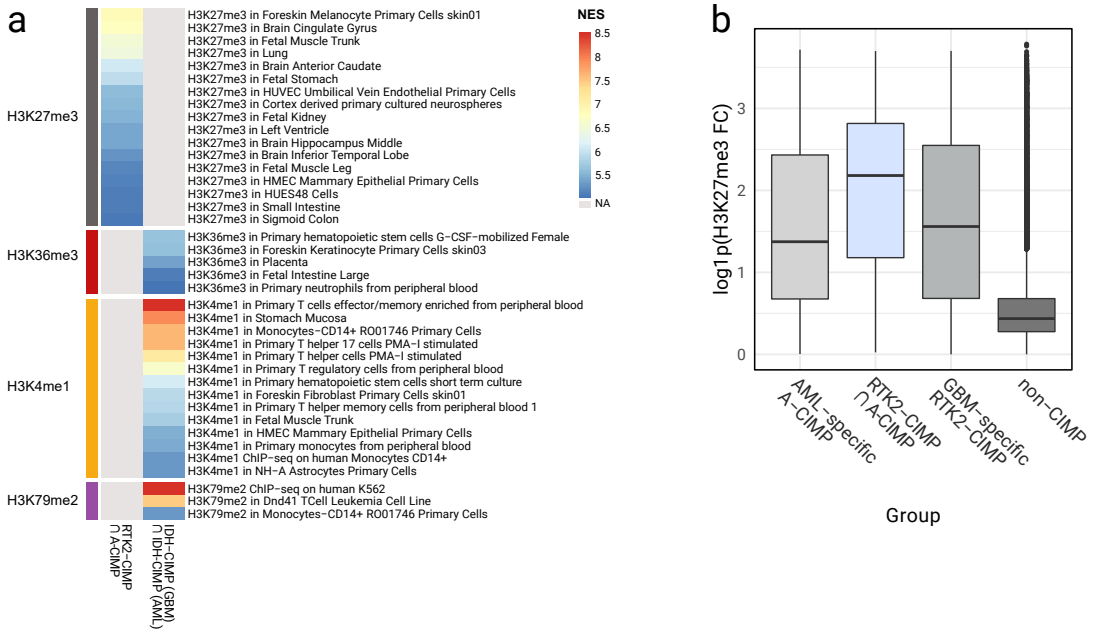


Figure 3.10: Assessing enrichments of CIMP-CGIs in other organs and on HSCs. [a] Heatmap representing the enrichment of promoter regions at shared A-/RTK2-CIMP (left column) or at shared IDH-CIMP targets (right column) in AML/GBM across several publicly available epigenomic datasets. [b] Average H3K27me3 log1pFC (over the input) from HSCs on the different classes of alternative CIMP for GBM and AML.

3.5.2 Causes and consequences of CIMP in GBM

In CGIs, DNA methylation usually increases upon tumourigenesis (Issa 2004). Genetic aberrations, like *IDH1/2* or *TET2* mutations have been deemed as the cause for CGI hypermethylation (Noushmehr et al. 2010; Tulstrup et al. 2021). We use IDH subtype for the comparison with CIMP in RTK-II as *IDH1/IDH2* mutations are not the cause of CIMP in RTK-II. Causes for CIMP were delineated by Yates and Boeva, but it is not yet clear which mechanism unequivocally fits into the origin of RTK2-CIMP (Yates and Boeva 2022). We verified the effect of chromosomes 19 and 20 gain and co-gain, having concluded this does not impact the CGI methylation of RTK2-CIMP. *EGFR* mutations have been linked to epigenomic remodeling in GBM before, but it is possible the CIMP we observed in RTK-II is due to another yet undiscovered alteration (Liu et al. 2015). While epigenomic origin is possible, it is imaginable that RTK2-CIMP could be caused by a mutational event affecting modifiers of DNA methylation or other chromatin modifications directly or indirectly, two of the CIMP-causing mechanisms identified before (Yates and Boeva 2022).

On the other hand, the existence of CIMP in RTK-II brings out questions on the consequences of the phenotype. In the IDH subtype, the CIMP-causing IDH mutation leads indirectly to defects in the glial cell differentiation, increases in H3K9 and H3K27 methylation, and disturbs CTCF binding on TAD boundaries (Lu et al. 2012; Sturm et al. 2012; Flavahan et al. 2016). In RTK-II, we found that RTK2-CIMP leads to gene repression and, potentially, to alterations in the neural cell differentiation. It is important to note that, while RTK2-CIMP appears to lead to a strong decrease in gene expression, post-transcriptional variability can still be attenuated through other mechanisms. Nevertheless, it is fair to assume that developmental programmes in general and the neural differentiation programmes in particular might be affected in RTK-II development. This is evidenced by the deconvolution analysis found in **Figure 3.8c** and by the observation that RTK2-CIMP targets are mostly active in neuron populations in the healthy brain **(Figure 3.8b)**.

The definite assumption that RTK2-CIMP affects the neuronal differentiation trajec-

tory still requires experimental validation. It is also possible that the existence of CGI hypermethylation at the RTK2-CIMP loci holds other consequences that were out of the scope of this study, like higher-order chromatin rearrangements.

3.5.3 CIMP in the tumourigenesis and development of GBM

We have used NPs to determine epigenetic priming that would allow identification of CGIs prone to become hypermethylated in healthy tissue. This finding implies that certain histone modification and DNA methylation patterns are associated to the formation of CIMP in GBM (Court and Arnaud 2017). Similarly, this observation also hinted on CIMP as a process affecting cell development or differentiation. In the case of RTK2-CIMP, the dynamics of these CGIs could also suggest that CIMP is more likely to emerge in CGIs of high epigenetic plasticity, characterised by the occurrence of both repressive and active histone modifications, as observed in the NMF analysis. In comparison with all CGIs, CIMP-targeted CGIs are linked to high H3K27me3, a histone modification often mentioned in the context of CIMP (Court and Arnaud 2017; Dunican et al. 2020). In ependymoma, CIMP has been found to target genes from the Polycomb repressive complex 2 (Mack et al. 2014). According to the Bayesian network we have generated, in the cell-of-origin, RTK2-CIMP might be the result of a similar action affecting H3K27me3 (or other histone modifications upstream from DNA methylation) to drive a epigenome-wide deregulation which ultimately leads to CGI hypermethylation. We were however not able to validate this hypothesis.

Previously, Alcantara Llaguno and colleagues observed that the potential for GBM development is anti-correlated with cell differentiation, meaning differentiated cells are less likely to be cells-of-origin (Alcantara Llaguno et al. 2019). Interneurons, oligodendrocyte precursor cells, astrocytes, or even neural stem cells have been suggested as a potential cell-of-origin for GBM (Dirks 2010; Chen et al. 2020; Zong, Parada, and Baker 2015). In this research, we hinted on the possibility of RTK2-CIMP being a tumourigenesis mechanism which leads to a smaller fraction of neuronal cells in the tumour, at least in the RTK-II subtype. This also implies that the cell-of-origin is not a differenti-

ated neuron, but rather a neural progenitor or another more undifferentiated cell type, common ancestor to both neurons and glial cells (Dirks 2010).

3.6 Chapter summary

In conclusion:

- The RTK-II subtype in GBM is associated to a definite CIMP (RTK2-CIMP). RTK2-CIMP is different from the IDH-CIMP, independent from the IDH mutations, and affects distinct CGIs;

- The two CIMP types affect gene expression in GBM, albeit not to the same extent;

- It is possible to associate CIMP-affected CGIs to specific epigenomic signatures, mostly characterised by an enrichment in H3K27me3 and H3K4me1. These epigenomic modifications can be predictive of CIMP already before tumourigenesis;

- RTK2-CIMP could be affecting the differentiation of specific cell lineages, like neurons, and affect the cell fate balance in tumours;

- Commonalities for between RTK2-CIMP and IDH-CIMP in GBM have been found in AML;

- AML and GBM IDH-CIMP and A-/RTK2-CIMP are linked to specific epigenomic traits, like H3K27me3 and H3K4me1 respectively, which can be traced to undifferentiated cells.

Chapter 4

Conclusion

Epigenomic-wide alterations play an important role in many conditions (Wang et al. 2022; Lieberman 2006). Here, we have assessed the chromatin landscape in both viral infection and cancer. In the analyses performed, it was evident that the process of integration in HIV-1 infection and its impact in the host cell requires further research. It is possible that these alterations might have clinical significance as well, particularly in the neurological syndromes which are associated to HIV-1 infection. While we have used a microglial cell line here, the impact of HIV-1 infection and integration in the brain environment requires further research as well. Single-cell studies from brain tissue would allow a more comprehensive analysis of the brain cells and can provide insights on the impact of HIV-1 even if the number of cells harbouring a latent provirus is low.

Tumourigenesis has also been connected to epigenetic alterations which hold clinical significance (Malta et al. 2018). We have generated results that still require experimental validation, but their value is still meaningful in GBM research, as the study of aberrant DNA methylation could bring us one step closer into understanding tumourigenesis. The impact of CIMP in the cell differentiation in RTK-II implies the need to study the tumour holistically, through the use of multi-omics for example. Although this was out of the scope of this work, it is possible that the DNA methylation alterations we observed in GBM could also translate into higher-order chromatin alterations.

In this work, we applied NMF, RF, and bayesian networks to the study of epigenomic modifications in the context of two conditions: HIV-1 infection and GBM. All three approaches made findings more interpretable and concise and opened further questions for future research. We applied NMF to define large-scale integration permissible windows in HIV-1 infection and CGI signatures in GBM. The same chromatin segmentation approach can be applied at smaller (nucleosomal) scales, as in the ChromHMM-derived chromatin states done in HIV-1 (Ernst and Kellis 2012). ChromHMM works with binary emission probability, as it converts the signal from 200bp-long windows to binary values (Ernst and Kellis 2017, 2012). However, this results in a loss of magnitude. NMF works with values directly, ensuring a more accurate representation of the data while still retaining biological interpretability. Gandolfi & Tramontano developed a NMF-based method for chromatin profile identification which proved to identify a larger fraction of functional regions (Gandolfi and Tramontano 2017). Furthermore, ChromHMM does not account for the fact that cells share common genetic information (Zhang and Hardison 2017). While integrative NMF would solve this issue on a multi-cell analysis, this problem was not encountered in this project given the small amount of biological data analysed. Nevertheless, NMF proved to be an useful tool in dimensionality reduction for the datasets used, allowing the selection of important epigenomic features associated to both cases in study.

We used RF to stratify features distinguishing the IDH- and RTK2-CIMP from other CGIs, locations most often targeted by HIV-1, and TF most likely to be involved in the formation of TAD boundaries. These analysis allowed for the selection of important features for downstream analysis and opened new questions which can be further explored. Unlike neural networks or other deep learning methods, it is possible to apply it even with small sample sizes, it is interpretable, and non-parametric (Breiman 2001). While RF was not used here for classification, but rather as a method of feature selection and importance, it holds immense potential for genomic studies in both.

Lastly, bayesian networks were trained to understand the interplay between epigenomic modifications over TAD boundaries and CIMP-affected CGIs. In both cases,

bayesian networks were applied to understand the connections between these different epigenetic players and to identify one or multiple master player driving the others in a restricted space. Epigenomic interactions are intricate and complex, and interactions can be location-dependent. Thus, bayesian networks can be an efficient and interpretable way to represent such processes, identify features of interest, and even predict outcomes. Genome-wide approaches have been done previously to determine targeting interactions in chromatin and have shown that bayesian networks are able to robustly predict novel interactions, provide insights into already known interactions, and its findings are translatable from experimental validation (Steensel et al. 2010).

In recent years, bioinformatics has been a reliable way to use computational methods to solve biological and medical problems. The usage of high-throughput sequencing allowed research to grow closer to a systems-level understanding of the cell. This work provided novel insights into the dynamics between epigenetic modifications in the context of HIV-1 infection and GBM and allowed inference on the chromatin alterations caused by disease through the usage of computational methods. In addition to proving targets for the experimental setting, these approaches proved useful in the understanding of cell response to cancer and infection, making them valuable for future research.

References

Ahn, Jeong Hyun, Eric S Davis, Timothy A Daugird, Shuai Zhao, Ivana Yoseli Quiroga, Hidetaka Uryu, Jie Li, et al. 2021. "Phase Separation Drives Aberrant Chromatin Looping and Cancer Development." *Nature* 595 (7868): 591–95. https://doi.org/10.1038/s41586-021-03662-5.

Albanese, Alberto, Daniele Arosio, Mariaelena Terreni, and Anna Cereseto. 2008. "HIV-1 Pre-Integration Complexes Selectively Target Decondensed Chromatin in the Nuclear Periphery." *PLoS One* 3 (6): e2413. https://doi.org/10.1371/journal.pone.0002413.

Alcantara Llaguno, Sheila, Daochun Sun, Alicia M Pedraza, Elsa Vera, Zilai Wang, Dennis K Burns, and Luis F Parada. 2019. "Cell-of-Origin Susceptibility to Glioblastoma Formation Declines with Neural Lineage Restriction." *Nat Neurosci* 22 (4): 545–55. https://doi.org/10.1038/s41593-018-0333-8.

Allday, Martin J. 2013. "EBV Finds a Polycomb-Mediated, Epigenetic Solution to the Problem of Oncogenic Stress Responses Triggered by Infection." *Front Genet* 4 (October): 212. https://doi.org/10.3389/fgene.2013.00212.

Allfrey, V G, R Faulkner, and A E Mirsky. 1964. "Acetylation and Methylation of Histones and Their Possible Role in the Regulation of RNA Synthesis." *Proc Natl Acad Sci U S A* 51 (May): 786–94. https://doi.org/10.1073/pnas.51.5.786.

Allis, C. David, and Thomas Jenuwein. 2016. "The Molecular Hallmarks of Epigenetic Control." *Nature Reviews Genetics* 17 (8): 487–500. https://doi.org/10.1038/nrg.2016.59.

Alvarez-Carbonell, David, Fengchun Ye, Nirmala Ramanath, Yoelvis Garcia-Mesa, Pamela E Knapp, Kurt F Hauser, and Jonathan Karn. 2019. "Cross-Talk Between Microglia and Neurons Regulates Hiv Latency." *PLoS Pathog* 15 (12): e1008249. https://doi.org/10.1371/journal.ppat.1008249.

Anderson, Elizabeth M, and Frank Maldarelli. 2018. "The Role of Integration and Clonal Expansion in HIV Infection: Live Long and Prosper." *Retrovirology* 15 (1): 71. https://doi.org/10.1186/s12977-018-0448-8.

Angelopoulos, Nicos, Aikaterini Chatzipli, Jyoti Nangalia, Francesco Maura, and Peter J Campbell. 2022. "Bayesian Networks Elucidate Complex Genomic Landscapes in Cancer." *Commun Biol* 5 (1): 306. https://doi.org/10.1038/s42003-022-03243-w.

Bachiller, Sara, Itzia Jiménez-Ferrer, Agnes Paulus, Yiyi Yang, Maria Swanberg, Tomas Deierborg, and Antonio Boza-Serrano. 2018. "Microglia in Neurological Diseases: A Road Map to Brain-Disease Dependent-Inflammatory Response." *Front Cell Neurosci* 12: 488. https://doi.org/10.3389/fncel.2018.00488.

Bae, Taejeong, Liana Fasching, Yifan Wang, Joo Heon Shin, Milovan Suvakov, Yeongjun Jang, Scott Norton, et al. 2022. "Analysis of Somatic Mutations in 131 Human Brains Reveals Aging-Associated Hypermutability." *Science* 377 (6605): 511–17. https://doi.org/10.1126/science.abm6222.

Banerji, J, S Rusconi, and W Schaffner. 1981. "Expression of a Beta-Globin Gene Is Enhanced by Remote Sv40 DNA Sequences." *Cell* 27 (2 Pt 1): 299–308. https://doi.org/10.1016/0092-8674(81)90413-x.

Bannister, Andrew J, and Tony Kouzarides. 2011. "Regulation of Chromatin by Histone Modifications." *Cell Res* 21 (3): 381–95. https://doi.org/10.1038/cr.2011.

22.

Barth, Teresa K, and Axel Imhof. 2010. "Fast Signals and Slow Marks: The Dynamics of Histone Modifications." *Trends Biochem Sci* 35 (11): 618–26. https://doi.org/10.1016/j.tibs.2010.05.006.

Battagli, Cristina, Robert G Uzzo, Essel Dulaimi, Inmaculada Ibanez de Caceres, Rachel Krassenstein, Tahseen Al-Saleem, Richard E Greenberg, and Paul Cairns. 2003. "Promoter Hypermethylation of Tumor Suppressor Genes in Urine from Kidney Cancer Patients." *Cancer Res* 63 (24): 8695–9.

Battivelli, Emilie, Matthew S Dahabieh, Mohamed Abdel-Mohsen, J Peter Svensson, Israel Tojal Da Silva, Lillian B Cohn, Andrea Gramatica, et al. 2018a. "Distinct Chromatin Functional States Correlate with HIV Latency Reactivation in Infected Primary Cd4+ T Cells." *Elife* 7 (May). https://doi.org/10.7554/eLife.34655.

———. 2018b. "Distinct Chromatin Functional States Correlate with HIV Latency Reactivation in Infected Primary $Cd4^+$ T Cells." Edited by Viviana Simon. *eLife* 7 (May). eLife Sciences Publications, Ltd: e34655. https://doi.org/10.7554/eLife.34655.

Baylin, S B, J W Höppener, A de Bustros, P H Steenbergh, C J Lips, and B D Nelkin. 1986. "DNA Methylation Patterns of the Calcitonin Gene in Human Lung Cancers and Lymphomas." *Cancer Res* 46 (6): 2917–22.

Beagan, Jonathan A, and Jennifer E Phillips-Cremins. 2020. "On the Existence and Functionality of Topologically Associating Domains." *Nat Genet* 52 (1): 8–16. https://doi.org/10.1038/s41588-019-0561-1.

Bedwell, Gregory J, and Alan N Engelman. 2021. "Factors That Mold the Nuclear Landscape of HIV-1 Integration." *Nucleic Acids Res* 49 (2): 621–35. https://doi. org/10.1093/nar/gkaa1207.

Bedwell, Gregory J, Sooin Jang, Wen Li, Parmit K Singh, and Alan N Engelman.

2021. “Rigrag: High-Resolution Mapping of Genic Targeting Preferences During HIV-1 Integration in Vitro and in Vivo.” *Nucleic Acids Res* 49 (13): 7330–46. https://doi.org/10.1093/nar/gkab514.

Bell, Lucy C K, and Mahdad Noursadeghi. 2018. “Pathogenesis of HIV-1 and Mycobacterium Tuberculosis Co-Infection.” *Nat Rev Microbiol* 16 (2): 80–90. https://doi.org/10.1038/nrmicro.2017.128.

Bell, Robert J A, H Tomas Rube, Alex Kreig, Andrew Mancini, Shaun D Fouse, Raman P Nagarajan, Serah Choi, et al. 2015. “The Transcription Factor Gabp Selectively Binds and Activates the Mutant Tert Promoter in Cancer.” *Science* 348 (6238): 1036–9. https://doi.org/10.1126/science.aab0015.

Belton, Jon-Matthew, Rachel Patton McCord, Johan Harmen Gibcus, Natalia Naumova, Ye Zhan, and Job Dekker. 2012. “Hi-c: A Comprehensive Technique to Capture the Conformation of Genomes.” *Methods* 58 (3): 268–76. https://doi.org/10.1016/j.ymeth.2012.05.001.

Bentsen, Mette, Philipp Goymann, Hendrik Schultheis, Kathrin Klee, Anastasiia Petrova, René Wiegandt, Annika Fust, et al. 2020. “ATAC-Seq Footprinting Unravels Kinetics of Transcription Factor Binding During Zygotic Genome Activation.” *Nat Commun* 11 (1): 4267. https://doi.org/10.1038/s41467-020-18035-1.

Bergman, J E H, N Janssen, L H Hoefsloot, M C J Jongmans, R M W Hofstra, and C M A van Ravenswaaij-Arts. 2011. “CHD7 Mutations and Charge Syndrome: The Clinical Implications of an Expanding Phenotype.” *J Med Genet* 48 (5): 334–42. https://doi.org/10.1136/jmg.2010.087106.

Bickmore, Wendy A. 2013. “The Spatial Organization of the Human Genome.” *Annu Rev Genomics Hum Genet* 14: 67–84. https://doi.org/10.1146/annurev-genom-091212-153515.

Bird, A P. 1986. “CpG-Rich Islands and the Function of DNA Methylation.” *Nature* 321 (6067): 209–13. https://doi.org/10.1038/321209a0.

Blankson, Joel N, Deborah Persaud, and Robert F Siliciano. 2002. "The Challenge of Viral Reservoirs in HIV-1 Infection." *Annu Rev Med* 53: 557–93. https://doi.org/10.1146/annurev.med.53.082901.104024.

Blazkova, Jana, Katerina Trejbalova, Françoise Gondois-Rey, Philippe Halfon, Patrick Philibert, Allan Guiguen, Eric Verdin, et al. 2009. "CpG Methylation Controls Reactivation of HIV from Latency." *PLoS Pathog* 5 (8): e1000554. https://doi.org/10.1371/journal.ppat.1000554.

Brady, Troy, Luis M Agosto, Nirav Malani, Charles C Berry, Una O'Doherty, and Frederic Bushman. 2009. "HIV Integration Site Distributions in Resting and Activated CD4 + T Cells Infected in Culture." *AIDS* 23 (12): 1461–71. https://doi.org/10.1097/QAD.0b013e32832caf28.

Branton, William G, Jason P Fernandes, Nazanin Mohammadzadeh, Mathew A L Doan, Jon D Laman, Benjamin B Gelman, Zahra Fagrouch, et al. 2022. "Microbial Molecule Ingress Promotes Neuroinflammation and Brain Ccr5 Expression in Persons with HIV-Associated Neurocognitive Disorders." *Brain Behav Immun* 107 (October): 110–23. https://doi.org/10.1016/j.bbi.2022.09.019.

Breiman, Leo. 2001. "Random Forests." *Machine Learning* 45 (1): 5–32. https://doi.org/10.1023/A:1010933404324.

Buenrostro, Jason D, Beijing Wu, Howard Y Chang, and William J Greenleaf. 2015. "ATAC-Seq: A Method for Assaying Chromatin Accessibility Genome-Wide." *Curr Protoc Mol Biol* 109 (January): 21.29.1–21.29.9. https://doi.org/10.1002/0471142727.mb2129s109.

Burgess-Beusse, Bonnie, Catherine Farrell, Miklos Gaszner, Michael Litt, Vesco Mutskov, Felix Recillas-Targa, Melanie Simpson, Adam West, and Gary Felsenfeld. 2002. "The Insulation of Genes from External Enhancers and Silencing Chromatin." *Proc Natl Acad Sci U S A* 99 Suppl 4 (Suppl 4): 16433–7. https://doi.org/10.1073/pnas.162342499.

Caron, Cécile, Edwige Col, and Saadi Khochbin. 2003. "The Viral Control of Cellular Acetylation Signaling." *Bioessays* 25 (1): 58–65. https://doi.org/10.1002/bies.10202.

Carter, Benjamin, and Keji Zhao. 2021. "The Epigenetic Basis of Cellular Heterogeneity." *Nat Rev Genet* 22 (4): 235–50. https://doi.org/10.1038/s41576-020-00300-0.

Carter, Christoph C, Adewunmi Onafuwa-Nuga, Lucy A McNamara, James Riddell 4th, Dale Bixby, Michael R Savona, and Kathleen L Collins. 2010. "HIV-1 Infects Multipotent Progenitor Cells Causing Cell Death and Establishing Latent Cellular Reservoirs." *Nat Med* 16 (4): 446–51. https://doi.org/10.1038/nm.2109.

Cazaly, Emma, Joseph Saad, Wenyu Wang, Caroline Heckman, Miina Ollikainen, and Jing Tang. 2019. "Making Sense of the Epigenome Using Data Integration Approaches." *Front Pharmacol* 10: 126. https://doi.org/10.3389/fphar.2019.00126.

Cedar, Howard, and Yehudit Bergman. 2009. "Linking DNA Methylation and Histone Modification: Patterns and Paradigms." *Nat Rev Genet* 10 (5): 295–304. https://doi.org/10.1038/nrg2540.

Cesana, Daniela, Francesca R Santoni de Sio, Laura Rudilosso, Pierangela Gallina, Andrea Calabria, Stefano Beretta, Ivan Merelli, et al. 2017. "HIV-1-Mediated Insertional Activation of Stat5b and Bach2 Trigger Viral Reservoir in T Regulatory Cells." *Nat Commun* 8 (1): 498. https://doi.org/10.1038/s41467-017-00609-1.

Chan, Jonathan K L, and Warner C Greene. 2011. "NF-Kb/Rel: Agonist and Antagonist Roles in Hiv-1 Latency." *Curr Opin HIV AIDS* 6 (1): 12–18. https://doi.org/10.1097/COH.0b013e32834124fd.

Chang, Moon-Sung, Hiroshi Uozaki, Ja-Mun Chong, Tetsuo Ushiku, Kazuya Sakuma, Shunpei Ishikawa, Rumi Hino, et al. 2006. "CpG Island Methylation Status in Gastric Carcinoma with and Without Infection of Epstein-Barr Virus." *Clin Cancer Res* 12 (10): 2995–3002. https://doi.org/10.1158/1078-0432.CCR-05-1601.

Chen, Carol C L, Shriya Deshmukh, Selin Jessa, Djihad Hadjadj, Véronique Lisi, Augusto Faria Andrade, Damien Faury, et al. 2020. "Histone H3.3g34-Mutant Interneuron Progenitors Co-Opt Pdgfra for Gliomagenesis." *Cell* 183 (6): 1617–1633.e22. https://doi.org/10.1016/j.cell.2020.11.012.

Cherepanov, Peter, Goedele Maertens, Paul Proost, Bart Devreese, Jozef Van Beeumen, Yves Engelborghs, Erik De Clercq, and Zeger Debyser. 2003. "HIV-1 Integrase Forms Stable Tetramers and Associates with Ledgf/P75 Protein in Human Cells." *J Biol Chem* 278 (1): 372–81. https://doi.org/10.1074/jbc.M209278200.

Choi, Won-Young, Ji-Hyun Hwang, Ann-Na Cho, Andrew J Lee, Inkyung Jung, Seung-Woo Cho, Lark Kyun Kim, and Young-Joon Kim. 2020. "NEUROD1 Intrinsically Initiates Differentiation of Induced Pluripotent Stem Cells into Neural Progenitor Cells." *Mol Cells* 43 (12): 1011–22. https://doi.org/10.14348/molcells.2020.0207.

Christensen, Brock C, E Andres Houseman, Carmen J Marsit, Shichun Zheng, Margaret R Wrensch, Joseph L Wiemels, Heather H Nelson, et al. 2009. "Aging and Environmental Exposures Alter Tissue-Specific DNA Methylation Dependent Upon CpG Island Context." *PLoS Genet* 5 (8): e1000602. https://doi.org/10.1371/journal.pgen.1000602.

Chun, T W, D Finzi, J Margolick, K Chadwick, D Schwartz, and R F Siliciano. 1995. "In Vivo Fate of HIV-1-Infected T Cells: Quantitative Analysis of the Transition to Stable Latency." *Nat Med* 1 (12): 1284–90. https://doi.org/10.1038/nm1295-1284.

Churchill, Melissa J, Steven G Deeks, David M Margolis, Robert F Siliciano, and Ronald Swanstrom. 2016. "HIV Reservoirs: What, Where and How to Target Them." *Nat Rev Microbiol* 14 (1): 55–60. https://doi.org/10.1038/nrmicro.2015.5.

Ciuffi, Angela, Keshet Ronen, Troy Brady, Nirav Malani, Gary Wang, Charles C Berry, and Frederic D Bushman. 2009. "Methods for Integration Site Distribution Analyses in Animal Cell Genomes." *Methods* 47 (4): 261–8. https://doi.org/10.1016/j.

ymeth.2008.10.028.

Clapham, P R, and A McKnight. 2001. "HIV-1 Receptors and Cell Tropism." *Br Med Bull* 58: 43–59. https://doi.org/10.1093/bmb/58.1.43.

Cochrane, Catherine R, Thomas A Angelovich, Sarah J Byrnes, Emily Waring, Aleks C Guanizo, Gemma S Trollope, Jingling Zhou, et al. 2022. "Intact HIV Proviruses Persist in the Brain Despite Viral Suppression with ART." *Ann Neurol* 92 (4): 532–44. https://doi.org/10.1002/ana.26456.

Colonna, Marco, and Oleg Butovsky. 2017. "Microglia Function in the Central Nervous System During Health and Neurodegeneration." *Annu Rev Immunol* 35 (April): 441–68. https://doi.org/10.1146/annurev-immunol-051116-052358.

Costello, J F, M C Frühwald, D J Smiraglia, L J Rush, G P Robertson, X Gao, F A Wright, et al. 2000. "Aberrant CpG-Island Methylation Has Non-Random and Tumour-Type-Specific Patterns." *Nat Genet* 24 (2): 132–8. https://doi.org/10.1038/72785.

Court, Franck, and Philippe Arnaud. 2017. "An Annotated List of Bivalent Chromatin Regions in Human ES Cells: A New Tool for Cancer Epigenetic Research." *Oncotarget* 8 (3): 4110–24. https://doi.org/10.18632/oncotarget.13746.

Couturier, Charles P, Shamini Ayyadhury, Phuong U Le, Javad Nadaf, Jean Monlong, Gabriele Riva, Redouane Allache, et al. 2020. "Single-Cell RNA-Seq Reveals That Glioblastoma Recapitulates a Normal Neurodevelopmental Hierarchy." *Nat Commun* 11 (1): 3406. https://doi.org/10.1038/s41467-020-17186-5.

Craigie, Robert, and Frederic D Bushman. 2012. "HIV DNA Integration." *Cold Spring Harb Perspect Med* 2 (7): a006890. https://doi.org/10.1101/cshperspect.a006890.

Crooks, Amanda M, Rosalie Bateson, Anna B Cope, Noelle P Dahl, Morgan K Griggs, JoAnn D Kuruc, Cynthia L Gay, et al. 2015. "Precise Quantitation of the Latent HIV-1 Reservoir: Implications for Eradication Strategies." *J Infect Dis* 212 (9): 1361–5.

https://doi.org/10.1093/infdis/jiv218.

Csankovszki, G, A Nagy, and R Jaenisch. 2001. "Synergism of Xist RNA, DNA Methylation, and Histone Hypoacetylation in Maintaining X Chromosome Inactivation." *J Cell Biol* 153 (4): 773–84. https://doi.org/10.1083/jcb.153.4.773.

Dahabieh, Matthew S, Emilie Battivelli, and Eric Verdin. 2015. "Understanding HIV Latency: The Road to an HIV Cure." *Annu Rev Med* 66: 407–21. https://doi.org/ 10.1146/annurev-med-092112-152941.

Dai, Ziwei, Vijyendra Ramesh, and Jason W. Locasale. 2020. "The Evolving Metabolic Landscape of Chromatin Biology and Epigenetics." *Nature Reviews Genetics* 21 (12): 737–53. https://doi.org/10.1038/s41576-020-0270-8.

Deaton, Aimée M, Shaun Webb, Alastair R W Kerr, Robert S Illingworth, Jacky Guy, Robert Andrews, and Adrian Bird. 2011. "Cell Type-Specific DNA Methylation at Intragenic CpG Islands in the Immune System." *Genome Res* 21 (7): 1074–86. https: //doi.org/10.1101/gr.118703.110.

Debyser, Zeger, Gerlinde Vansant, Anne Bruggemans, Julie Janssens, and Frauke Christ. 2018. "Insight in HIV Integration Site Selection Provides a Block-and-Lock Strategy for a Functional Cure of HIV Infection." *Viruses* 11 (1). https://doi.org/10.3390/v11010012.

De Crignis, E., and T. Mahmoudi. 2017. "Chapter Six - the Multifaceted Contributions of Chromatin to HIV-1 Integration, Transcription, and Latency." In, edited by Lorenzo Galluzzi, 328:197–252. International Review of Cell and Molecular Biology. Academic Press. https://doi.org/https://doi.org/10.1016/bs.ircmb.2016.08. 006.

Deeks, Steven G, Nancie Archin, Paula Cannon, Simon Collins, R Brad Jones, Marein A W P de Jong, Olivier Lambotte, et al. 2021. "Research Priorities for an HIV Cure: International Aids Society Global Scientific Strategy 2021." *Nat Med*, December.

https://doi.org/10.1038/s41591-021-01590-5.

Devarajan, Karthik. 2008. "Nonnegative Matrix Factorization: An Analytical and Interpretive Tool in Computational Biology." *PLoS Comput Biol* 4 (7): e1000029. https://doi.org/10.1371/journal.pcbi.1000029.

Di Primio, Cristina, Valentina Quercioli, Awatef Allouch, Rik Gijsbers, Frauke Christ, Zeger Debyser, Daniele Arosio, and Anna Cereseto. 2013. "Single-Cell Imaging of HIV-1 Provirus (Scip)." *Proc Natl Acad Sci U S A* 110 (14): 5636–41. https://doi. org/10.1073/pnas.1216254110.

Dirks, Peter B. 2010. "Brain Tumor Stem Cells: The Cancer Stem Cell Hypothesis Writ Large." *Mol Oncol* 4 (5): 420–30. https://doi.org/10.1016/j.molonc.2010.08.001.

Dixon, Jesse R, Siddarth Selvaraj, Feng Yue, Audrey Kim, Yan Li, Yin Shen, Ming Hu, Jun S Liu, and Bing Ren. 2012. "Topological Domains in Mammalian Genomes Identified by Analysis of Chromatin Interactions." *Nature* 485 (7398): 376–80. https://doi.org/10.1038/nature11082.

Doolittle-Hall, Janet M, Danielle L Cunningham Glasspoole, William T Seaman, and Jennifer Webster-Cyriaque. 2015. "Meta-Analysis of DNA Tumor-Viral Integration Site Selection Indicates a Role for Repeats, Gene Expression and Epigenetics." *Cancers (Basel)* 7 (4): 2217–35. https://doi.org/10.3390/cancers7040887.

Dostie, Josée, Todd A Richmond, Ramy A Arnaout, Rebecca R Selzer, William L Lee, Tracey A Honan, Eric D Rubio, et al. 2006. "Chromosome Conformation Capture Carbon Copy (5C): A Massively Parallel Solution for Mapping Interactions Between Genomic Elements." *Genome Res* 16 (10): 1299–1309. https://doi.org/10.1101/gr.5571506.

Douglas, Jenny, Sandra Hanks, I Karen Temple, Sally Davies, Alexandra Murray, Meena Upadhyaya, Susan Tomkins, Helen E Hughes, Trevor R P Cole, and Nazneen Rahman.

2003. “NSD1 Mutations Are the Major Cause of Sotos Syndrome and Occur in Some Cases of Weaver Syndrome but Are Rare in Other Overgrowth Phenotypes.” *Am J Hum Genet* 72 (1): 132–43. https://doi.org/10.1086/345647.

Dulaimi, Essel, Jeanne Hillinck, Inmaculada Ibanez de Caceres, Tahseen Al-Saleem, and Paul Cairns. 2004. “Tumor Suppressor Gene Promoter Hypermethylation in Serum of Breast Cancer Patients.” *Clin Cancer Res* 10 (18 Pt 1): 6189–93. https://doi.org/10.1158/1078-0432.CCR-04-0597.

Dunican, Donnchadh S, Heidi K Mjoseng, Leanne Duthie, Ilya M Flyamer, Wendy A Bickmore, and Richard R Meehan. 2020. “Bivalent Promoter Hypermethylation in Cancer Is Linked to the H3K27me3/H3k4me3 Ratio in Embryonic Stem Cells.” *BMC Biol* 18 (1): 25. https://doi.org/10.1186/s12915-020-0752-3.

Eberharter, Anton, and Peter B Becker. 2002. “Histone Acetylation: A Switch Between Repressive and Permissive Chromatin. Second in Review Series on Chromatin Dynamics.” *EMBO Rep* 3 (3): 224–9. https://doi.org/10.1093/embo-reports/kvf053.

Eckmann, Jean-Pierre, and Tsvi Tlusty. 2021. “Dimensional Reduction in Complex Living Systems: Where, Why, and How.” *Bioessays* 43 (9): e2100062. https://doi.org/10.1002/bies.202100062.

Eggers, Christian, Gabriele Arendt, Katrin Hahn, Ingo W Husstedt, Matthias Maschke, Eva Neuen-Jacob, Mark Obermann, et al. 2017. “HIV-1-Associated Neurocognitive Disorder: Epidemiology, Pathogenesis, Diagnosis, and Treatment.” *J Neurol* 264 (8): 1715–27. https://doi.org/10.1007/s00415-017-8503-2.

Embretson, J, M Zupancic, J L Ribas, A Burke, P Racz, K Tenner-Racz, and A T Haase. 1993. “Massive Covert Infection of Helper T Lymphocytes and Macrophages by HIV During the Incubation Period of AIDS.” *Nature* 362 (6418): 359–62. https: //doi.org/10.1038/362359a0.

ENCODE Project Consortium. 2012. "An Integrated Encyclopedia of DNA Elements in the Human Genome." *Nature* 489 (7414): 57–74. https://doi.org/10.1038/nature11247.

Ernst, Jason, and Manolis Kellis. 2010. "Discovery and Characterization of Chromatin States for Systematic Annotation of the Human Genome." *Nat Biotechnol* 28 (8): 817–25. https://doi.org/10.1038/nbt.1662.

———. 2012. "ChromHMM: Automating Chromatin-State Discovery and Characterization." *Nat Methods* 9 (3): 215–6. https://doi.org/10.1038/nmeth.1906.

———. 2017. "Chromatin-State Discovery and Genome Annotation with ChromHMM." *Nat Protoc* 12 (12): 2478–92. https://doi.org/10.1038/nprot.2017.124.

Esteller, M, M Sanchez-Cespedes, R Rosell, D Sidransky, S B Baylin, and J G Herman. 1999. "Detection of Aberrant Promoter Hypermethylation of Tumor Suppressor Genes in Serum DNA from Non-Small Cell Lung Cancer Patients." *Cancer Res* 59 (1): 67–70.

Ewels, Philip A, Alexander Peltzer, Sven Fillinger, Harshil Patel, Johannes Alneberg, Andreas Wilm, Maxime Ulysse Garcia, Paolo Di Tommaso, and Sven Nahnsen. 2020. "The Nf-Core Framework for Community-Curated Bioinformatics Pipelines." *Nat Biotechnol* 38 (3): 276–78. https://doi.org/10.1038/s41587-020-0439-x.

Farhadian, Shelli F, Sameet S Mehta, Chrysoula Zografou, Kevin Robertson, Richard W Price, Jenna Pappalardo, Jennifer Chiarella, David A Hafler, and Serena S Spudich. 2018. "Single-Cell RNA Sequencing Reveals Microglia-Like Cells in Cerebrospinal Fluid During Virologically Suppressed HIV." *JCI Insight* 3 (18). https://doi.org/10. 1172/jci.insight.121718.

Feinberg, Andrew P, and Andre Levchenko. 2023. "Epigenetics as a Mediator of Plasticity in Cancer." *Science* 379 (6632): eaaw3835. https://doi.org/10.1126/science.aaw3835.

Figueroa, Maria E, Omar Abdel-Wahab, Chao Lu, Patrick S Ward, Jay Patel, Alan Shih, Yushan Li, et al. 2010. "Leukemic *Idh1* and *Idh2* Mutations Result in a Hypermethylation Phenotype, Disrupt Tet2 Function, and Impair Hematopoietic Differentiation." *Cancer Cell* 18 (6): 553–67. https://doi.org/10.1016/j.ccr.2010.11.015.

Filbin, Mariella G, and Mario L Suvà. 2016. "Gliomas Genomics and Epigenomics: Arriving at the Start and Knowing It for the First Time." *Annu Rev Pathol* 11 (May): 497–521. https://doi.org/10.1146/annurev-pathol-012615-044208.

Finzi, D, J Blankson, J D Siliciano, J B Margolick, K Chadwick, T Pierson, K Smith, et al. 1999. "Latent Infection of Cd4+ T Cells Provides a Mechanism for Lifelong Persistence of HIV-1, Even in Patients on Effective Combination Therapy." *Nat Med* 5 (5): 512–7. https://doi.org/10.1038/8394.

Flavahan, William A, Yotam Drier, Brian B Liau, Shawn M Gillespie, Andrew S Venteicher, Anat O Stemmer-Rachamimov, Mario L Suvà, and Bradley E Bernstein. 2016. "Insulator Dysfunction and Oncogene Activation in *Idh* Mutant Gliomas." *Nature* 529 (7584): 110–4. https://doi.org/10.1038/nature16490.

Francis, Ashwanth C, Mariana Marin, Parmit K Singh, Vasudevan Achuthan, Mathew J Prellberg, Kristina Palermino-Rowland, Shuiyun Lan, et al. 2020. "HIV-1 Replication Complexes Accumulate in Nuclear Speckles and Integrate into Speckle-Associated Genomic Domains." *Nat Commun* 11 (1): 3505. https://doi.org/10.1038/s41467-020-17256-8.

Friedman, Julia, Won-Kyung Cho, Chung K Chu, Kara S Keedy, Nancie M Archin, David M Margolis, and Jonathan Karn. 2011. "Epigenetic Silencing of HIV-1 by the Histone H3 Lysine 27 Methyltransferase Enhancer of Zeste 2." *J Virol* 85 (17): 9078–89. https://doi.org/10.1128/JVI.00836-11.

Galli, Rossella, Elena Binda, Ugo Orfanelli, Barbara Cipelletti, Angela Gritti, Simona De Vitis, Roberta Fiocco, Chiara Foroni, Francesco Dimeco, and Angelo Vescovi. 2004. "Isolation and Characterization of Tumorigenic, Stem-Like Neural Precursors from Human Glioblastoma." *Cancer Res* 64 (19): 7011–21. https://doi.org/10.1158/

0008-5472.CAN-04-1364.

Gan, Wei, Juan Luo, Yi Zhou Li, Jia Li Guo, Min Zhu, and Meng Long Li. 2019. "A Computational Method to Predict Topologically Associating Domain Boundaries Combining Histone Marks and Sequence Information." *BMC Genomics* 20 (Suppl 13): 980. https://doi.org/10.1186/s12864-019-6303-z.

Gandolfi, Francesco, and Anna Tramontano. 2017. "A Computational Approach for the Functional Classification of the Epigenome." *Epigenetics Chromatin* 10: 26. https://doi.org/10.1186/s13072-017-0131-7.

Gao, Chao, Jialin Liu, April R Kriebel, Sebastian Preissl, Chongyuan Luo, Rosa Castanon, Justin Sandoval, et al. 2021. "Iterative Single-Cell Multi-Omic Integration Using Online Learning." *Nat Biotechnol* 39 (8): 1000–1007. https://doi.org/10.1038/s41587-021-00867-x.

Garcia-Mesa, Yoelvis, Taylor R Jay, Mary Ann Checkley, Benjamin Luttge, Curtis Dobrowolski, Saba Valadkhan, Gary E Landreth, Jonathan Karn, and David Alvarez-Carbonell. 2017. "Immortalization of Primary Microglia: A New Platform to Study HIV Regulation in the Central Nervous System." *J Neurovirol* 23 (1): 47–66. https: //doi.org/10.1007/s13365-016-0499-3.

Gartlgruber, Moritz, Ashwini Kumar Sharma, Andrés Quintero, Daniel Dreidax, Selina Jansky, Young-Gyu Park, Sina Kreth, et al. 2021. "Super Enhancers Define Regulatory Subtypes and Cell Identity in Neuroblastoma." *Nat Cancer* 2 (1): 114–28. https://doi.org/10.1038/s43018-020-00145-w.

Ginsberg, Stephen D, Melissa J Alldred, Satya M Gunnam, Consuelo Schiroli, Sang Han Lee, Susan Morgello, and Tracy Fischer. 2018. "Expression Profiling Suggests Microglial Impairment in Human Immunodeficiency Virus Neuropathogenesis." *Ann Neurol* 83 (2): 406–17. https://doi.org/10.1002/ana.25160.

Gobran, Samaa T, Petronela Ancuta, and Naglaa H Shoukry. 2021. "A Tale of Two

Viruses: Immunological Insights into HCV/HIV Coinfection." *Front Immunol* 12: 726419. https://doi.org/10.3389/fimmu.2021.726419.

Gosselin, David, Dylan Skola, Nicole G Coufal, Inge R Holtman, Johannes C M Schlachetzki, Eniko Sajti, Baptiste N Jaeger, et al. 2017. "An Environment-Dependent Transcriptional Network Specifies Human Microglia Identity." *Science* 356 (6344). https://doi.org/10.1126/science.aal3222.

Graziano, V, S E Gerchman, D K Schneider, and V Ramakrishnan. 1994. "Histone H1 Is Located in the Interior of the Chromatin 30-Nm Filament." *Nature* 368 (6469): 351–4. https://doi.org/10.1038/368351a0.

Groves, Ian J, Emma L A Drane, Marco Michalski, Jack M Monahan, Cinzia G Scarpini, Stephen P Smith, Giovanni Bussotti, et al. 2021. "Short- and Long-Range Cis Interactions Between Integrated Hpv Genomes and Cellular Chromatin Dysregulate Host Gene Expression in Early Cervical Carcinogenesis." *PLoS Pathog* 17 (8): e1009875. https://doi.org/10.1371/journal.ppat.1009875.

Gu, Hongcang, Zachary D Smith, Christoph Bock, Patrick Boyle, Andreas Gnirke, and Alexander Meissner. 2011. "Preparation of Reduced Representation Bisulfite Sequencing Libraries for Genome-Scale Dna Methylation Profiling." *Nat Protoc* 6 (4): 468–81. https://doi.org/10.1038/nprot.2010.190.

Gu, Zuguang, Roland Eils, and Matthias Schlesner. 2016. "Complex Heatmaps Reveal Patterns and Correlations in Multidimensional Genomic Data." *Bioinformatics* 32 (18): 2847–9. https://doi.org/10.1093/bioinformatics/btw313.

Gujar, Hemant, Daniel J Weisenberger, and Gangning Liang. 2019. "The Roles of Human Dna Methyltransferases and Their Isoforms in Shaping the Epigenome." *Genes (Basel)* 10 (2). https://doi.org/10.3390/genes10020172.

Haberle, Vanja, and Alexander Stark. 2018. "Eukaryotic Core Promoters and the Functional Basis of Transcription Initiation." *Nat Rev Mol Cell Biol* 19 (10): 621–37.

https://doi.org/10.1038/s41580-018-0028-8.

Han, Mei, Lina Jia, Wencai Lv, Lihui Wang, and Wei Cui. 2019. "Epigenetic Enzyme Mutations: Role in Tumorigenesis and Molecular Inhibitors." *Front Oncol* 9: 194. https://doi.org/10.3389/fonc.2019.00194.

Hannibal, Mark C, Kati J Buckingham, Sarah B Ng, Jeffrey E Ming, Anita E Beck, Margaret J McMillin, Heidi I Gildersleeve, et al. 2011. "Spectrum of Mll2 (Alr) Mutations in 110 Cases of Kabuki Syndrome." *Am J Med Genet A* 155A (7): 1511–6. https://doi.org/10.1002/ajmg.a.34074.

Hansen, Kasper Daniel, Winston Timp, Héctor Corrada Bravo, Sarven Sabunciyan, Benjamin Langmead, Oliver G McDonald, Bo Wen, et al. 2011. "Increased Methylation Variation in Epigenetic Domains Across Cancer Types." *Nat Genet* 43 (8): 768–75. https://doi.org/10.1038/ng.865.

Hao, Yuhan, Stephanie Hao, Erica Andersen-Nissen, William M Mauck 3rd, Shiwei Zheng, Andrew Butler, Maddie J Lee, et al. 2021. "Integrated Analysis of Multimodal Single-Cell Data." *Cell* 184 (13): 3573–3587.e29. https://doi.org/10.1016/j.cell.2021.04.048.

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman. 2001. *The Elements of Statistical Learning*. Springer Series in Statistics. New York, NY, USA: Springer New York Inc.

Hawkins, R David, Gary C Hon, Leonard K Lee, Queminh Ngo, Ryan Lister, Mattia Pelizzola, Lee E Edsall, et al. 2010. "Distinct Epigenomic Landscapes of Pluripotent and Lineage-Committed Human Cells." *Cell Stem Cell* 6 (5): 479–91. https://doi.org/10.1016/j.stem.2010.03.018.

Heinz, Sven, Christopher Benner, Nathanael Spann, Eric Bertolino, Yin C Lin, Peter Laslo, Jason X Cheng, Cornelis Murre, Harinder Singh, and Christopher K Glass. 2010. "Simple Combinations of Lineage-Determining Transcription Factors Prime

Cis-Regulatory Elements Required for Macrophage and B Cell Identities." *Mol Cell* 38 (4): 576–89. https://doi.org/10.1016/j.molcel.2010.05.004.

Helmsauer, Konstantin, Maria E Valieva, Salaheddine Ali, Rocı'o Chamorro González, Robert Schöpflin, Claudia Röefzaad, Yi Bei, et al. 2020. "Enhancer Hijacking Determines Extrachromosomal Circular Mycn Amplicon Architecture in Neuroblastoma." *Nat Commun* 11 (1): 5823. https://doi.org/10.1038/s41467-020-19452-y.

Herrmann, Carl, Bram Van de Sande, Delphine Potier, and Stein Aerts. 2012. "I-cisTarget: An Integrative Genomics Method for the Prediction of Regulatory Features and Cis-Regulatory Modules." *Nucleic Acids Res* 40 (15): e114. https://doi.org/10.1093/nar/gks543.

Ho, Tin Kam. 1995. "Random Decision Forests." In *Proceedings of 3rd International Conference on Document Analysis and Recognition*, 1:278–82 vol.1. https://doi.org/10.1109/ICDAR.1995.598994.

———. 1998. "The Random Subspace Method for Constructing Decision Forests." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 20 (8): 832–44. https://doi.org/10.1109/34.709601.

Hoffman, Michael M, Jason Ernst, Steven P Wilder, Anshul Kundaje, Robert S Harris, Max Libbrecht, Belinda Giardine, et al. 2013. "Integrative Annotation of Chromatin Elements from Encode Data." *Nucleic Acids Res* 41 (2): 827–41. https://doi.org/10.1093/nar/gks1284.

Hong, Seungpyo, and Dongsup Kim. 2017. "Computational Characterization of Chromatin Domain Boundary-Associated Genomic Elements." *Nucleic Acids Res* 45 (18): 10403–14. https://doi.org/10.1093/nar/gkx738.

Horvath, Steve. 2013. "DNA Methylation Age of Human Tissues and Cell Types." *Genome Biol* 14 (10): R115. https://doi.org/10.1186/gb-2013-14-10-r115.

Hu, Benxia, Hyejung Won, Won Mah, Royce B Park, Bibi Kassim, Keeley Spiess, Alexey

Kozlenkov, et al. 2021. "Neuronal and Glial 3D Chromatin Architecture Informs the Cellular Etiology of Brain Disorders." *Nat Commun* 12 (1): 3968. https://doi.org/10.1038/s41467-021-24243-0.

Imai, Kenichi, Hiroaki Togami, and Takashi Okamoto. 2010. "Involvement of Histone H3 Lysine 9 (H3k9) Methyltransferase G9a in the Maintenance of HIV-1 Latency and Its Reactivation by Bix01294." *J Biol Chem* 285 (22): 16538–45. https://doi.org/ 10.1074/jbc.M110.103531.

Imamichi, Hiromi, Mindy Smith, Joseph W Adelsberger, Taisuke Izumi, Francesca Scrimieri, Brad T Sherman, Catherine A Rehm, et al. 2020. "Defective HIV-1 Proviruses Produce Viral Proteins." *Proc Natl Acad Sci U S A* 117 (7): 3704–10. https: //doi.org/10.1073/pnas.1917876117.

Issa, Jean-Pierre. 2004. "CpG Island Methylator Phenotype in Cancer." *Nat Rev Cancer* 4 (12): 988–93. https://doi.org/10.1038/nrc1507.

James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani. 2014. *An Introduction to Statistical Learning: With Applications in R*. Springer Publishing Company, Incorporated.

Jefferys, Stuart R, Samuel D Burgos, Jackson J Peterson, Sara R Selitsky, Anne-Marie W Turner, Lindsey I James, Yi-Hsuan Tsai, et al. 2021. "Epigenomic Characterization of Latent HIV Infection Identifies Latency Regulating Transcription Factors." *PLoS Pathog* 17 (2): e1009346. https://doi.org/10.1371/journal.ppat.1009346.

Jeong, Mira, Deqiang Sun, Min Luo, Yun Huang, Grant A Challen, Benjamin Rodriguez, Xiaotian Zhang, et al. 2014. "Large Conserved Domains of Low Dna Methylation Maintained by Dnmt3a." *Nat Genet* 46 (1): 17–23. https://doi.org/10.1038/ng.2836.

Jia, Qunying, Shuhua Chen, Yuan Tan, Yuejin Li, and Faqing Tang. 2020. "Oncogenic Super-Enhancer Formation in Tumorigenesis and Its Molecular Mechanisms." *Exp*

Mol Med 52 (5): 713–23. https://doi.org/10.1038/s12276-020-0428-7.

Jiang, Shan, and Ali Mortazavi. 2018. "Integrating Chip-Seq with Other Functional Genomics Data." *Brief Funct Genomics* 17 (2): 104–15. https://doi.org/10.1093/bfgp/ely002.

John M Coffin, Stephen H Hughes, and Harold E Varmus, eds. 1997. *Retroviruses*. Cold Spring Harbor Laboratory Press.

Kaelbling, L. P., M. L. Littman, and A. W. Moore. 1996. "Reinforcement Learning: A Survey." *Journal of Artificial Intelligence Research* 4 (May). AI Access Foundation: 237–85. https://doi.org/10.1613/jair.301.

Kanton, Sabina, Michael James Boyle, Zhisong He, Malgorzata Santel, Anne Weigert, Fátima Sanchı's-Calleja, Patricia Guijarro, et al. 2019. "Organoid Single-Cell Genomic Atlas Uncovers Human-Specific Features of Brain Development." *Nature* 574 (7778): 418–22. https://doi.org/10.1038/s41586-019-1654-9.

Karimzadeh, Mehran, Christopher Arlidge, Ariana Rostami, Mathieu Lupien, Scott V. Bratman, and Michael M. Hoffman. 2022. "Human Papillomavirus Integration Transforms Chromatin to Drive Oncogenesis." *bioRxiv*. Cold Spring Harbor Laboratory. https://doi.org/10.1101/2020.02.12.942755.

Kelly, A D, H Kroeger, J Yamazaki, R Taby, F Neumann, S Yu, J T Lee, et al. 2017. "A Cpg Island Methylator Phenotype in Acute Myeloid Leukemia Independent of Idh Mutations and Associated with a Favorable Outcome." *Leukemia* 31 (10): 2011–9. https://doi.org/10.1038/leu.2017.12.

Kent, J R, P-Y Zeng, D Atanasiu, J Gardner, N W Fraser, and S L Berger. 2004. "During Lytic Infection Herpes Simplex Virus Type 1 Is Associated with Histones Bearing Modifications That Correlate with Active Transcription." *J Virol* 78 (18): 10178–86. https://doi.org/10.1128/JVI.78.18.10178-10186.2004.

Kent, W James. 2002. "BLAT–the Blast-Like Alignment Tool." *Genome Res* 12 (4):

656–64. https://doi.org/10.1101/gr.229202.

Khoury, Amanda, Joanna Achinger-Kawecka, Saul A Bert, Grady C Smith, Hugh J French, Phuc-Loi Luu, Timothy J Peters, et al. 2020. "Constitutively Bound Ctcf Sites Maintain 3D Chromatin Architecture and Long-Range Epigenetically Regulated Domains." *Nat Commun* 11 (1): 54. https://doi.org/10.1038/s41467-019-13753-7.

Kleihues, P, and H Ohgaki. 1999. "Primary and Secondary Glioblastomas: From Concept to Clinical Diagnosis." *Neuro Oncol* 1 (1): 44–51. https://doi.org/10.1093/neuonc/1.1.44.

Klemm, Sandy L, Zohar Shipony, and William J Greenleaf. 2019. "Chromatin Accessibility and the Regulatory Epigenome." *Nat Rev Genet* 20 (4): 207–20. https://doi.org/10.1038/s41576-018-0089-8.

Klughammer, Johanna, Barbara Kiesel, Thomas Roetzer, Nikolaus Fortelny, Amelie Nemc, Karl-Heinz Nenning, Julia Furtner, et al. 2018. "The Dna Methylation Landscape of Glioblastoma Disease Progression Shows Extensive Heterogeneity in Time and Space." *Nat Med* 24 (10): 1611–24. https://doi.org/10.1038/s41591-018-0156-x.

Kok, Yik Lim, Valentina Vongrad, Mohaned Shilaih, Francesca Di Giallonardo, Herbert Kuster, Roger Kouyos, Huldrych F Günthard, and Karin J Metzner. 2016. "Monocyte-Derived Macrophages Exhibit Distinct and More Restricted HIV-1 Integration Site Repertoire Than Cd4(+) T Cells." *Sci Rep* 6 (April): 24157. https://doi.org/10. 1038/srep24157.

Kuhn, Max. 2008. "Building Predictive Models in R Using the Caret Package." *Journal of Statistical Software* 28 (5): 1–26. https://doi.org/10.18637/jss.v028.i05.

Lange, Ulrike C, Roxane Verdikt, Amina Ait-Ammar, and Carine Van Lint. 2020. "Epigenetic Crosstalk in Chronic Infection with HIV-1." *Semin Immunopathol* 42 (2): 187–200. https://doi.org/10.1007/s00281-020-00783-3.

Langmead, Ben, and Steven L Salzberg. 2012. "Fast Gapped-Read Alignment with Bowtie 2." *Nat Methods* 9 (4): 357–9. https://doi.org/10.1038/nmeth.1923.

Lapaillerie, Delphine, Benoıˆt Lelandais, Eric Mauro, Floriane Lagadec, Camille Tumiotto, Csaba Miskey, Guillaume Ferran, et al. 2021. "Modulation of the Intrinsic Chromatin Binding Property of HIV-1 Integrase by Ledgf/P75." *Nucleic Acids Res* 49 (19): 11241–56. https://doi.org/10.1093/nar/gkab886.

Lee, D D, and H S Seung. 1999. "Learning the Parts of Objects by Non-Negative Matrix Factorization." *Nature* 401 (6755): 788–91. https://doi.org/10.1038/44565.

Lee, Eva Y H P, and William J Muller. 2010. "Oncogenes and Tumor Suppressor Genes." *Cold Spring Harb Perspect Biol* 2 (10): a003236. https://doi.org/10.1101/cshperspect.a003236.

Lee, Joo Ho, Jeong Eun Lee, Jee Ye Kahng, Se Hoon Kim, Jun Sung Park, Seon Jin Yoon, Ji-Yong Um, et al. 2018. "Human Glioblastoma Arises from Subventricular Zone Cells with Low-Level Driver Mutations." *Nature* 560 (7717): 243–47. https://doi.org/10.1038/s41586-018-0389-3.

Lee, Tong Ihn, and Richard A Young. 2013. "Transcriptional Regulation and Its Misregulation in Disease." *Cell* 152 (6): 1237–51. https://doi.org/10.1016/j.cell.2013.02.014.

Lelek, Mickaël, Nicoletta Casartelli, Danilo Pellin, Ermanno Rizzi, Philippe Souque, Marco Severgnini, Clelia Di Serio, et al. 2015. "Chromatin Organization at the Nuclear Pore Favours HIV Replication." *Nat Commun* 6 (March): 6483. https://doi.org/10.1038/ncomms7483.

Li, Bing, Michael Carey, and Jerry L Workman. 2007. "The Role of Chromatin During Transcription." *Cell* 128 (4): 707–19. https://doi.org/10.1016/j.cell.2007.01.015.

Li, E, C Beard, and R Jaenisch. 1993. "Role for Dna Methylation in Genomic Imprint-

ing." *Nature* 366 (6453): 362–5. https://doi.org/10.1038/366362a0.

Li, Yan, Judith H I Haarhuis, Ángela Sedeño Cacciatore, Roel Oldenkamp, Marjon S van Ruiten, Laureen Willems, Hans Teunissen, et al. 2020. "The Structural Basis for Cohesin-Ctcf-Anchored Loops." *Nature* 578 (7795): 472–76. https://doi.org/10.1038/s41586-019-1910-z.

Liau, Brian B, Cem Sievers, Laura K Donohue, Shawn M Gillespie, William A Flavahan, Tyler E Miller, Andrew S Venteicher, et al. 2017. "Adaptive Chromatin Remodeling Drives Glioblastoma Stem Cell Plasticity and Drug Tolerance." *Cell Stem Cell* 20 (2): 233–246.e7. https://doi.org/10.1016/j.stem.2016.11.003.

Lieberman, Paul M. 2006. "Chromatin Regulation of Virus Infection." *Trends Microbiol* 14 (3): 132–40. https://doi.org/10.1016/j.tim.2006.01.001.

Lim, Pek Siew, M Frances Shannon, and Kristine Hardy. 2010. "Epigenetic Control of Inducible Gene Expression in the Immune System." *Epigenomics* 2 (6): 775–95. https://doi.org/10.2217/epi.10.55.

Lin, Xihui, and Paul C Boutros. 2020. "Optimization and Expansion of Non-Negative Matrix Factorization." *BMC Bioinformatics* 21 (1): 7. https://doi.org/10.1186/s12859-019-3312-5.

Linden, Noemi, and R Brad Jones. 2022. "Potential Multi-Modal Effects of Provirus Integration on HIV-1 Persistence: Lessons from Other Viruses." *Trends Immunol* 43 (8): 617–29. https://doi.org/10.1016/j.it.2022.06.001.

Lister, Ryan, Mattia Pelizzola, Robert H Dowen, R David Hawkins, Gary Hon, Julian Tonti-Filippini, Joseph R Nery, et al. 2009. "Human Dna Methylomes at Base Resolution Show Widespread Epigenomic Differences." *Nature* 462 (7271): 315–22. https://doi.org/10.1038/nature08514.

Liu, Feng, Gary C Hon, Genaro R Villa, Kristen M Turner, Shiro Ikegami, Huijun Yang, Zhen Ye, et al. 2015. "EGFR Mutation Promotes Glioblastoma Through

Epigenome and Transcription Factor Network Remodeling." *Mol Cell* 60 (2): 307–18. https://doi.org/10.1016/j.molcel.2015.09.002.

Liu, Hang, Run-Hong Zhou, Yu Liu, Le Guo, Xu Wang, Wen-Hui Hu, and Wen-Zhe Ho. 2020. "HIV Infection Suppresses Tlr3 Activation-Mediated Antiviral Immunity in Microglia and Macrophages." *Immunology* 160 (3): 269–79. https://doi.org/10.1111/imm.13181.

Liu, Runxia, Yang-Hui Jimmy Yeh, Ales Varabyou, Jack A Collora, Scott Sherrill-Mix, C Conover Talbot Jr, Sameet Mehta, et al. 2020. "Single-Cell Transcriptional Landscapes Reveal HIV-1-Driven Aberrant Host Gene Transcription as a Potential Thera-peutic Target." *Sci Transl Med* 12 (543). https://doi.org/10.1126/scitranslmed. aaz0802.

Losman, Julie-Aurore, Ryan E Looper, Peppi Koivunen, Sungwoo Lee, Rebekka K Schneider, Christine McMahon, Glenn S Cowley, David E Root, Benjamin L Ebert, and William G Kaelin Jr. 2013. "(R)-2-Hydroxyglutarate Is Sufficient to Promote Leukemogenesis and Its Effects Are Reversible." *Science* 339 (6127): 1621–5. https://doi.org/10.1126/science.1231677.

Louis, David N, Arie Perry, Pieter Wesseling, Daniel J Brat, Ian A Cree, Dominique Figarella-Branger, Cynthia Hawkins, et al. 2021. "The 2021 Who Classification of Tumors of the Central Nervous System: A Summary." *Neuro Oncol* 23 (8): 1231–51. https://doi.org/10.1093/neuonc/noab106.

Lu, Chao, Patrick S Ward, Gurpreet S Kapoor, Dan Rohle, Sevin Turcan, Omar Abdel-Wahab, Christopher R Edwards, et al. 2012. "IDH Mutation Impairs Histone Demethylation and Results in a Block to Cell Differentiation." *Nature* 483 (7390): 474–8. https://doi.org/10.1038/nature10860.

Lucic, Bojana, Heng-Chang Chen, Maja Kuzman, Eduard Zorita, Julia Wegner, Vera Minneker, Wei Wang, et al. 2019. "Spatially Clustered Loci with Multiple Enhancers Are Frequent Targets of HIV-1 Integration." *Nat Commun* 10 (1): 4059. https:

//doi.org/10.1038/s41467-019-12046-3.

Luo, Yunhai, Benjamin C Hitz, Idan Gabdank, Jason A Hilton, Meenakshi S Kagda, Bonita Lam, Zachary Myers, et al. 2020. "New Developments on the Encyclopedia of Dna Elements (Encode) Data Portal." *Nucleic Acids Res* 48 (D1): D882–D889. https://doi.org/10.1093/nar/gkz1062.

Lupiáñez, Darı'o G, Malte Spielmann, and Stefan Mundlos. 2016. "Breaking Tads: How Alterations of Chromatin Domains Result in Disease." *Trends Genet* 32 (4): 225–37. https://doi.org/10.1016/j.tig.2016.01.003.

Lusic, Marina, and Robert F Siliciano. 2017. "Nuclear Landscape of HIV-1 Infection and Integration." *Nat Rev Microbiol* 15 (2): 69–82. https://doi.org/10.1038/nrmicro.2016.162.

Lyko, Frank. 2018. "The Dna Methyltransferase Family: A Versatile Toolkit for Epigenetic Regulation." *Nat Rev Genet* 19 (2): 81–92. https://doi.org/10.1038/nrg.2017.80.

Mack, S C, H Witt, R M Piro, L Gu, S Zuyderduyn, A M Stütz, X Wang, et al. 2014. "Epigenomic Alterations Define Lethal Cimp-Positive Ependymomas of Infancy." *Nature* 506 (7489): 445–50. https://doi.org/10.1038/nature13108.

Maldarelli, F, X Wu, L Su, F R Simonetti, W Shao, S Hill, J Spindler, et al. 2014. "HIV Latency. Specific HIV Integration Sites Are Linked to Clonal Expansion and Persistence of Infected Cells." *Science* 345 (6193): 179–83. https://doi.org/10.1126/science.1254194.

Malta, Tathiane M, Camila F de Souza, Thais S Sabedot, Tiago C Silva, Maritza S Mosella, Steven N Kalkanis, James Snyder, Ana Valeria B Castro, and Houtan Noushmehr. 2018. "Glioma Cpg Island Methylator Phenotype (G-Cimp): Biological and Clinical Implications." *Neuro Oncol* 20 (5): 608–20. https://doi.org/10.1093/neuonc/nox183.

Mandell, Gerald L, John E Bennett, and Raphael Dolin. 2010. *Mandell, Douglas, and Bennett's Principles and Practice of Infectious Diseases*. 7th ed. Philadelphia, PA: Churchill Livingstone/Elsevier.

Marini, Bruna, Attila Kertesz-Farkas, Hashim Ali, Bojana Lucic, Kamil Lisek, Lara Manganaro, Sandor Pongor, et al. 2015. "Nuclear Architecture Dictates HIV-1 In-tegration Site Selection." *Nature* 521 (7551): 227–31. https://doi.org/10.1038/nature14226.

Martin, Marcel. 2011. "Cutadapt Removes Adapter Sequences from High-Throughput Sequencing Reads." *EMBnet.journal* 17 (1): 10–12. https://doi.org/10.14806/ej.17.1.200.

Mbonye, Uri, and Jonathan Karn. 2014. "Transcriptional Control of HIV Latency: Cellular Signaling Pathways, Epigenetics, Happenstance and the Hope for a Cure." *Virology* 454-455 (April): 328–39. https://doi.org/10.1016/j.virol.2014.02.008.

McArthur, Evonne, and John A Capra. 2021. "Topologically Associating Domain Boundaries That Are Stable Across Diverse Cell Types Are Evolutionarily Constrained and Enriched for Heritability." *Am J Hum Genet* 108 (2): 269–83. https://doi.org/10.1016/j.ajhg.2021.01.001.

McKenzie, Andrew T, Minghui Wang, Mads E Hauberg, John F Fullard, Alexey Kozlenkov, Alexandra Keenan, Yasmin L Hurd, et al. 2018. "Brain Cell Type Specific Gene Expression and Co-Expression Network Architectures." *Sci Rep* 8 (1): 8868. https://doi.org/10.1038/s41598-018-27293-5.

McLaren, Paul J, and Jacques Fellay. 2021. "HIV-1 and Human Genetic Variation." *Nat Rev Genet* 22 (10): 645–57. https://doi.org/10.1038/s41576-021-00378-0.

Meissner, Alexander, Tarjei S Mikkelsen, Hongcang Gu, Marius Wernig, Jacob Hanna, Andrey Sivachenko, Xiaolan Zhang, et al. 2008. "Genome-Scale Dna Methylation Maps of Pluripotent and Differentiated Cells." *Nature* 454 (7205): 766–70. https:

//doi.org/10.1038/nature07107.

Melamed, Anat, Tomas W Fitzgerald, Yuchuan Wang, Jian Ma, Ewan Birney, and Charles R M Bangham. 2022. "Selective Clonal Persistence of Human Retroviruses in Vivo: Radial Chromatin Organization, Integration Site, and Host Transcription." *Sci Adv* 8 (17): eabm6210. https://doi.org/10.1126/sciadv.abm6210.

Melamed, Anat, Hiroko Yaguchi, Michi Miura, Aviva Witkover, Tomas W Fitzgerald, Ewan Birney, and Charles Rm Bangham. 2018. "The Human Leukemia Virus Htlv-1 Alters the Structure and Transcription of Host Chromatin in Cis." *Elife* 7 (June). https://doi.org/10.7554/eLife.36245.

Mellors, John W, Shuang Guo, Asma Naqvi, Leah D Brandt, Ling Su, Zhonghe Sun, Kevin W Joseph, et al. 2021. "Insertional Activation of Stat3 and Lck by HIV-1 Proviruses in T Cell Lymphomas." *Sci Adv* 7 (42): eabi8795. https://doi.org/10. 1126/sciadv.abi8795.

Meulendyke, Kelly A, Joshua D Croteau, and M Christine Zink. 2014. "HIV Life Cycle, Innate Immunity and Autophagy in the Central Nervous System." *Curr Opin HIV AIDS* 9 (6): 565–71. https://doi.org/10.1097/COH.0000000000000106.

Méndez, Catalina, Scott Ledger, Kathy Petoumenos, Chantelle Ahlenstiel, and Anthony D Kelleher. 2018. "RNA-Induced Epigenetic Silencing Inhibits HIV-1 Re-activation from Latency." *Retrovirology* 15 (1): 67. https://doi.org/10.1186/s12977-018-0451-0.

Michalak, Ewa M., Marian L. Burr, Andrew J. Bannister, and Mark A. Dawson. 2019. "The Roles of Dna, Rna and Histone Methylation in Ageing and Cancer." *Nature Reviews Molecular Cell Biology* 20 (10): 573–89. https://doi.org/10.1038/s41580-019-0143-1.

Mikkelsen, Tarjei S, Manching Ku, David B Jaffe, Biju Issac, Erez Lieberman, Georgia Giannoukos, Pablo Alvarez, et al. 2007. "Genome-Wide Maps of Chromatin State

in Pluripotent and Lineage-Committed Cells." *Nature* 448 (7153): 553–60. https://doi.org/10.1038/nature06008.

Miklı'k, Dalibor, Filip Šenigl, and Jiři' Hejnar. 2018. "Proviruses with Long-Term Stable Expression Accumulate in Transcriptionally Active Chromatin Close to the Gene Regulatory Elements: Comparison of Aslv-, HIV- and Mlv-Derived Vectors." *Viruses* 10 (3). https://doi.org/10.3390/v10030116.

Mikovits, J A, H A Young, P Vertino, J P Issa, P M Pitha, S Turcoski-Corrales, D D Taub, C L Petrow, S B Baylin, and F W Ruscetti. 1998. "Infection with Human Immunodeficiency Virus Type 1 Upregulates Dna Methyltransferase, Resulting in de Novo Methylation of the Gamma Interferon (Ifn-Gamma) Promoter and Subsequent Downregulation of Ifn-Gamma Production." *Mol Cell Biol* 18 (9): 5166–77. https://doi.org/10.1128/MCB.18.9.5166.

Mirabella, Anne C, Benjamin M Foster, and Till Bartke. 2016. "Chromatin Deregulation in Disease." *Chromosoma* 125 (1): 75–93. https://doi.org/10.1007/s00412-015-0530-0.

Mittelbronn, M, K Dietz, H J Schluesener, and R Meyermann. 2001. "Local Distribution of Microglia in the Normal Adult Human Central Nervous System Differs by up to One Order of Magnitude." *Acta Neuropathol* 101 (3): 249–55. https://doi.org/10.1007/s004010000284.

Moarii, Matahi, Fabien Reyal, and Jean-Philippe Vert. 2015. "Integrative Dna Methylation and Gene Expression Analysis to Assess the Universality of the Cpg Island Methylator Phenotype." *Hum Genomics* 9 (October): 26. https://doi.org/10.1186/s40246-015-0048-9.

Montavon, Thomas, Nicholas Shukeir, Galina Erikson, Bettina Engist, Megumi Onishi-Seebacher, Devon Ryan, Yaarub Musa, et al. 2021. "Complete Loss of H3k9 Methylation Dissolves Mouse Heterochromatin Organization." *Nat Commun* 12 (1): 4359. https://doi.org/10.1038/s41467-021-24532-8.

Moraes, Filipa, Ana Nóvoa, Loydie A Jerome-Majewska, Virginia E Papaioannou, and Moisés Mallo. 2005. "Tbx1 Is Required for Proper Neural Crest Migration and to Stabilize Spatial Patterns During Middle and Inner Ear Development." *Mech Dev* 122 (2): 199–212. https://doi.org/10.1016/j.mod.2004.10.004.

Nagarajan, Radhakrishnan, and Marco Scutari. 2013. *Bayesian Networks in R with Applications in Systems Biology.* New York: Springer. https://doi.org/10.1007/978-1-4614-6446-4.

Nan, X, H H Ng, C A Johnson, C D Laherty, B M Turner, R N Eisenman, and A Bird. 1998. "Transcriptional Repression by the Methyl-Cpg-Binding Protein Mecp2 Involves a Histone Deacetylase Complex." *Nature* 393 (6683): 386–9. https://doi.org/10.1038/30764.

Ni, Yang, Peter Müller, Lin Wei, and Yuan Ji. 2018. "Bayesian Graphical Models for Computational Network Biology." *BMC Bioinformatics* 19 (S3). Springer Science; Business Media LLC. https://doi.org/10.1186/s12859-018-2063-z.

Nishiyama, Atsuya, and Makoto Nakanishi. 2021. "Navigating the Dna Methylation Landscape of Cancer." *Trends Genet* 37 (11): 1012–27. https://doi.org/10.1016/j.tig.2021.05.002.

Northcott, Paul A, Catherine Lee, Thomas Zichner, Adrian M Stütz, Serap Erkek, Daisuke Kawauchi, David J H Shih, et al. 2014. "Enhancer Hijacking Activates Gfi1 Family Oncogenes in Medulloblastoma." *Nature* 511 (7510): 428–34. https://doi.org/10.1038/nature13379.

Nott, Alexi, Inge R Holtman, Nicole G Coufal, Johannes C M Schlachetzki, Miao Yu, Rong Hu, Claudia Z Han, et al. 2019. "Brain Cell Type-Specific Enhancer-Promoter Interactome Maps and Disease-Risk Association." *Science* 366 (6469): 1134–9. https://doi.org/10.1126/science.aay0793.

Noushmehr, Houtan, Daniel J Weisenberger, Kristin Diefes, Heidi S Phillips, Kanan

Pujara, Benjamin P Berman, Fei Pan, et al. 2010. "Identification of a Cpg Island Methylator Phenotype That Defines a Distinct Subgroup of Glioma." *Cancer Cell* 17 (5): 510–22. https://doi.org/10.1016/j.ccr.2010.03.017.

Ogino, Shuji, Takako Kawasaki, Gregory J Kirkner, Massimo Loda, and Charles S Fuchs. 2006. "CpG Island Methylator Phenotype-Low (Cimp-Low) in Colorectal Cancer: Possible Associations with Male Sex and Kras Mutations." *J Mol Diagn* 8 (5): 582–8. https://doi.org/10.2353/jmoldx.2006.060082.

Ogino, Shuji, Katsuhiko Nosho, Gregory J Kirkner, Takako Kawasaki, Jeffrey A Meyerhardt, Massimo Loda, Edward L Giovannucci, and Charles S Fuchs. 2009. "CpG Island Methylator Phenotype, Microsatellite Instability, Braf Mutation and Clinical Outcome in Colon Cancer." *Gut* 58 (1): 90–96. https://doi.org/10.1136/gut.2008.155473.

Ohm, Joyce E, Kelly M McGarvey, Xiaobing Yu, Linzhao Cheng, Kornel E Schuebel, Leslie Cope, Helai P Mohammad, et al. 2007. "A Stem Cell-Like Chromatin Pattern May Predispose Tumor Suppressor Genes to Dna Hypermethylation and Heritable Silencing." *Nat Genet* 39 (2): 237–42. https://doi.org/10.1038/ng1972.

Okabe, Atsushi, Kie Kyon Huang, Keisuke Matsusaka, Masaki Fukuyo, Manjie Xing, Xuewen Ong, Takayuki Hoshii, et al. 2020. "Cross-Species Chromatin Interactions Drive Transcriptional Rewiring in Epstein-Barr Virus-Positive Gastric Adenocarcinoma." *Nat Genet* 52 (9): 919–30. https://doi.org/10.1038/s41588-020-0665-7.

Orzan, F, S Pellegatta, P L Poliani, F Pisati, V Caldera, F Menghi, D Kapetis, C Marras, D Schiffer, and G Finocchiaro. 2011. "Enhancer of Zeste 2 (Ezh2) Is up-Regulated in Malignant Gliomas and in Glioma Stem-Like Cells." *Neuropathol Appl Neurobiol* 37 (4): 381–94. https://doi.org/10.1111/j.1365-2990.2010.01132.x.

Osborne, Olivia, Nadia Peyravian, Madhavan Nair, Sylvia Daunert, and Michal To-borek. 2020. "The Paradox of HIV Blood-Brain Barrier Penetrance and Antiretroviral Drug Delivery Deficiencies." *Trends Neurosci* 43 (9): 695–708. https://doi.org/

10.1016/j.tins.2020.06.007.

Pang, Baoxu, and Michael P Snyder. 2020. "Systematic Identification of Silencers in Human Cells." *Nat Genet* 52 (3): 254–63. https://doi.org/10.1038/s41588-020-0578-5.

Pantaleo, G, C Graziosi, J F Demarest, L Butini, M Montroni, C H Fox, J M Orenstein, D P Kotler, and A S Fauci. 1993. "HIV Infection Is Active and Progressive in Lymphoid Tissue During the Clinically Latent Stage of Disease." *Nature* 362 (6418): 355–8. https://doi.org/10.1038/362355a0.

Park, Peter J. 2009. "ChIP-Seq: Advantages and Challenges of a Maturing Technology." *Nat Rev Genet* 10 (10): 669–80. https://doi.org/10.1038/nrg2641.

Paschos, Konstantinos, and Martin J Allday. 2010. "Epigenetic Reprogramming of Host Genes in Viral and Microbial Pathogenesis." *Trends Microbiol* 18 (10): 439–47. https://doi.org/10.1016/j.tim.2010.07.003.

Pearson, Karl. 1901. "LIII. On Lines and Planes of Closest Fit to Systems of Points in Space." Doi: 10.1080/14786440109462720. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 2 (11). Taylor & Francis: 559–72. https://doi.org/10.1080/14786440109462720.

Peterson, Craig L, and Marc-André Laniel. 2004. "Histones and Histone Modifications." *Curr Biol* 14 (14): R546–51. https://doi.org/10.1016/j.cub.2004.07.007.

Phillips, Andrew N, James Neaton, and Jens D Lundgren. 2008. "The Role of HIV in Serious Diseases Other Than Aids." *AIDS* 22 (18): 2409–18. https://doi.org/10.1097/QAD.0b013e3283174636.

Phillips, Heidi S, Samir Kharbanda, Ruihuan Chen, William F Forrest, Robert H Soriano, Thomas D Wu, Anjan Misra, et al. 2006. "Molecular Subclasses of High-Grade Glioma Predict Prognosis, Delineate a Pattern of Disease Progression, and Resemble Stages in Neurogenesis." *Cancer Cell* 9 (3): 157–73. https://doi.org/10.1016/j.ccr.2006.02.019.

Pieretti, M, F P Zhang, Y H Fu, S T Warren, B A Oostra, C T Caskey, and D L Nelson. 1991. "Absence of Expression of the Fmr-1 Gene in Fragile X Syndrome." *Cell* 66 (4): 817–22. https://doi.org/10.1016/0092-8674(91)90125-i.

Plass, Christoph, Stefan M Pfister, Anders M Lindroth, Olga Bogatyrova, Rainer Claus, and Peter Lichter. 2013. "Mutations in Regulators of the Epigenome and Their Connections to Global Chromatin Patterns in Cancer." *Nat Rev Genet* 14 (11): 765–80. https://doi.org/10.1038/nrg3554.

Plaza-Jennings, Amara L, Aditi Valada, Callan O'Shea, Marina Iskhakova, Benxia Hu, Behnam Javidfar, Gabriella Ben Hutta, et al. 2022. "HIV Integration in the Human Brain Is Linked to Microglial Activation and 3D Genome Remodeling." *Mol Cell* 82 (24): 4647–4663.e8. https://doi.org/10.1016/j.molcel.2022.11.016.

Poles, M A, J Elliott, P Taing, P A Anton, and I S Chen. 2001. "A Preponderance of Ccr5(+) Cxcr4(+) Mononuclear Cells Enhances Gastrointestinal Mucosal Susceptibility to Human Immunodeficiency Virus Type 1 Infection." *J Virol* 75 (18): 8390–9. https://doi.org/10.1128/jvi.75.18.8390-8399.2001.

Quinlan, Aaron R, and Ira M Hall. 2010. "BEDTools: A Flexible Suite of Utilities for Comparing Genomic Features." *Bioinformatics* 26 (6): 841–2. https://doi.org/10.1093/bioinformatics/btq033.

Quinlan, J. R. 1986. "Induction of Decision Trees." *Machine Learning* 1 (1): 81–106. https://doi.org/10.1007/BF00116251.

Quintero, Andres, Daniel Hübschmann, Nils Kurzawa, Sebastian Steinhauser, Philipp Rentzsch, Stephen Krämer, Carolin Andresen, et al. 2020. "ShinyButchR: Interactive Nmf-Based Decomposition Workflow of Genome-Scale Datasets." *Biol Methods Protoc* 5 (1): bpaa022. https://doi.org/10.1093/biomethods/bpaa022.

Rai, Mohammad A, Jason Hammonds, Mario Pujato, Christopher Mayhew, Krishna Roskin, and Paul Spearman. 2020. "Comparative Analysis of Human Microglial

Models for Studies of HIV Replication and Pathogenesis." *Retrovirology* 17 (1): 35. https://doi.org/10.1186/s12977-020-00544-y.

Ramirez, Fidel, Friederike Dündar, Sarah Diehl, Björn A Grüning, and Thomas Manke. 2014. "DeepTools: A Flexible Platform for Exploring Deep-Sequencing Data." *Nucleic Acids Res* 42 (Web Server issue): W187–91. https://doi.org/10.1093/nar/gku365.

Rao, Suhas S P, Miriam H Huntley, Neva C Durand, Elena K Stamenova, Ivan D Bochkov, James T Robinson, Adrian L Sanborn, et al. 2014. "A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping." *Cell* 159 (7): 1665–80. https://doi.org/10.1016/j.cell.2014.11.021.

Réu, Pedro, Azadeh Khosravi, Samuel Bernard, Jeff E Mold, Mehran Salehpour, Kanar Alkass, Shira Perl, et al. 2017. "The Lifespan and Turnover of Microglia in the Human Brain." *Cell Rep* 20 (4): 779–84. https://doi.org/10.1016/j.celrep.2017.07.004.

Rheinberger, Mona, Ana Luisa Costa, Martin Kampmann, Dunja Glavas, Iart Luca Shytaj, Sheetal Sreeram, Carlotta Penzo, et al. 2023. "Genomic Profiling of HIV-1 Integration in Microglia Cells Links Viral Integration to the Topologically Associated Domains." *Cell Rep* 42 (2): 112110. https://doi.org/10.1016/j.celrep.2023.112110.

Rijnsoever, M van, F Grieu, H Elsaleh, D Joseph, and B Iacopetta. 2002. "Characterisation of Colorectal Cancers Showing Hypermethylation at Multiple Cpg Islands." *Gut* 51 (6): 797–802. https://doi.org/10.1136/gut.51.6.797.

Rivera, Chloe M, and Bing Ren. 2013. "Mapping Human Epigenomes." *Cell* 155 (1): 39–55.

Roadmap Epigenomics Consortium, Anshul Kundaje, Wouter Meuleman, Jason Ernst, Misha Bilenky, Angela Yen, Alireza Heravi-Moussavi, et al. 2015. "Integrative Anal-

ysis of 111 Reference Human Epigenomes." *Nature* 518 (7539): 317–30. https://doi.org/10.1038/nature14248.

Sakai, T, J Toguchida, N Ohtani, D W Yandell, J M Rapaport, and T P Dryja. 1991. "Allele-Specific Hypermethylation of the Retinoblastoma Tumor-Suppressor Gene." *Am J Hum Genet* 48 (5): 880–8.

Sarker, Iqbal H. 2021. "Machine Learning: Algorithms, Real-World Applications and Research Directions." *SN Computer Science* 2 (3). Springer Science; Business Media LLC. https://doi.org/10.1007/s42979-021-00592-x.

Satou, Yorifumi, Paola Miyazato, Ko Ishihara, Hiroko Yaguchi, Anat Melamed, Michi Miura, Asami Fukuda, et al. 2016. "The Retrovirus Htlv-1 Inserts an Ectopic Ctcf-Binding Site into the Human Genome." *Proc Natl Acad Sci U S A* 113 (11): 3054–9. https://doi.org/10.1073/pnas.1423199113.

Scherdin, U, K Rhodes, and M Breindl. 1990. "Transcriptionally Active Genome Regions Are Preferred Targets for Retrovirus Integration." *J Virol* 64 (2): 907–12. https://doi.org/10.1128/JVI.64.2.907-912.1990.

Schoenfelder, Stefan, and Peter Fraser. 2019. "Long-Range Enhancer-Promoter Contacts in Gene Expression Control." *Nat Rev Genet* 20 (8): 437–55. https://doi.org/10.1038/s41576-019-0128-0.

Schoenfelder, Stefan, Biola-Maria Javierre, Mayra Furlan-Magaril, Steven W Wingett, and Peter Fraser. 2018. "Promoter Capture Hi-c: High-Resolution, Genome-Wide Profiling of Promoter Interactions." *J Vis Exp*, no. 136 (June). https://doi.org/10.3791/57320.

Schott, J J, D W Benson, C T Basson, W Pease, G M Silberbach, J P Moak, B J Maron, C E Seidman, and J G Seidman. 1998. "Congenital Heart Disease Caused by Mutations in the Transcription Factor Nkx2-5." *Science* 281 (5373): 108–11. https://doi.org/10.1126/science.281.5373.108.

Schröder, Astrid R W, Paul Shinn, Huaming Chen, Charles Berry, Joseph R Ecker, and Frederic Bushman. 2002. "HIV-1 Integration in the Human Genome Favors Active Genes and Local Hotspots." *Cell* 110 (4): 521–9. https://doi.org/10.1016/s0092-8674(02)00864-4.

Scutari, Marco. 2010. "Learning Bayesian Networks with the bnlearn R Package." *Journal of Statistical Software* 35 (3): 1–22. https://doi.org/10.18637/jss.v035.i03.

Serrao, Erik, Peter Cherepanov, and Alan N Engelman. 2016. "Amplification, Next-Generation Sequencing, and Genomic Dna Mapping of Retroviral Integration Sites." *J Vis Exp*, no. 109 (March). https://doi.org/10.3791/53840.

Shah, Raven, Christian M Gallardo, Yoonhee H Jung, Ben Clock, Jesse R Dixon, William M McFadden, Kinjal Majumder, et al. 2022. "Activation of HIV-1 Proviruses Increases Downstream Chromatin Accessibility." *iScience* 25 (12): 105490. https://doi.org/ 10.1016/j.isci.2022.105490.

Shanmugam, Muthu K, Frank Arfuso, Surendar Arumugam, Arunachalam Chinnathambi, Bian Jinsong, Sudha Warrier, Ling Zhi Wang, et al. 2018. "Role of Novel Histone Modifications in Cancer." *Oncotarget* 9 (13): 11414–26. https://doi.org/10.18632/oncotarget.23356.

Shen, Lanlan, Yutaka Kondo, Yi Guo, Jiexin Zhang, Li Zhang, Saira Ahmed, Jingmin Shu, Xinli Chen, Robert A Waterland, and Jean-Pierre J Issa. 2007. "Genome-Wide Profiling of Dna Methylation Reveals a Class of Normally Methylated Cpg Island Promoters." *PLoS Genet* 3 (10): 2023–36. https://doi.org/10.1371/journal.pgen.0030181.

Sherman, Eric, Christopher Nobles, Charles C Berry, Emmanuelle Six, Yinghua Wu, Anatoly Dryga, Nirav Malani, et al. 2017. "INSPIIRED: A Pipeline for Quantitative Analysis of Sites of New Dna Integration in Cellular Genomes." *Mol Ther Methods Clin Dev* 4 (March): 39–49. https://doi.org/10.1016/j.omtm.2016.11.002.

Siliciano, Robert F, and Warner C Greene. 2011. "HIV Latency." *Cold Spring Harb Perspect Med* 1 (1): a007096. https://doi.org/10.1101/cshperspect.a007096.

Simonis, Marieke, Petra Klous, Erik Splinter, Yuri Moshkin, Rob Willemsen, Elzo de Wit, Bas van Steensel, and Wouter de Laat. 2006. "Nuclear Organization of Active and Inactive Chromatin Domains Uncovered by Chromosome Conformation Capture-on-Chip (4C)." *Nat Genet* 38 (11): 1348–54. https://doi.org/10.1038/ng1896.

Singh, Parmit Kumar, Gregory J. Bedwell, and Alan N. Engelman. 2022. "Spatial and Genomic Correlates of HIV-1 Integration Site Targeting." *Cells* 11 (4). https: //doi.org/10.3390/cells11040655.

Singh, Parmit Kumar, Matthew R Plumb, Andrea L Ferris, James R Iben, Xiaolin Wu, Hind J Fadel, Brian T Luke, et al. 2015. "LEDGF/P75 Interacts with mRNA Splicing Factors and Targets HIV-1 Integration to Highly Spliced Genes." *Genes Dev* 29 (21): 2287–97. https://doi.org/10.1101/gad.267609.115.

Sneeringer, Christopher J, Margaret Porter Scott, Kevin W Kuntz, Sarah K Knutson, Roy M Pollock, Victoria M Richon, and Robert A Copeland. 2010. "Coordinated Activities of Wild-Type Plus Mutant Ezh2 Drive Tumor-Associated Hypertrimethylation of Lysine 27 on Histone H3 (H3k27) in Human B-Cell Lymphomas." *Proc Natl Acad Sci U S A* 107 (49): 20980–5. https://doi.org/10.1073/pnas.1012525107.

Spencer, David H, David A Russler-Germain, Shamika Ketkar, Nichole M Helton, Tamara L Lamprecht, Robert S Fulton, Catrina C Fronick, et al. 2017. "CpG Island Hypermethylation Mediated by Dnmt3a Is a Consequence of Aml Progression." *Cell* 168 (5): 801–816.e13. https://doi.org/10.1016/j.cell.2017.01.021.

Spitz, François, and Eileen E M Furlong. 2012. "Transcription Factors: From Enhancer Binding to Developmental Control." *Nat Rev Genet* 13 (9): 613–26. https://doi.org/10.1038/nrg3207.

Splinter, Erik, Helen Heath, Jurgen Kooren, Robert-Jan Palstra, Petra Klous, Frank

Grosveld, Niels Galjart, and Wouter de Laat. 2006. "CTCF Mediates Long-Range Chromatin Looping and Local Histone Modification in the Beta-Globin Locus." *Genes Dev* 20 (17): 2349–54. https://doi.org/10.1101/gad.399506.

Sproul, Duncan, Robert R Kitchen, Colm E Nestor, J Michael Dixon, Andrew H Sims, David J Harrison, Bernard H Ramsahoye, and Richard R Meehan. 2012. "Tissue of Origin Determines Cancer-Associated Cpg Island Promoter Hypermethylation Patterns." *Genome Biol* 13 (10): R84. https://doi.org/10.1186/gb-2012-13-10-r84.

Sreeram, Sheetal, Fengchun Ye, Yoelvis Garcia-Mesa, Kien Nguyen, Ahmed El Sayed, Konstantin Leskov, and Jonathan Karn. 2022. "The Potential Role of HIV-1 Latency in Promoting Neuroinflammation and HIV-1-Associated Neurocognitive Disorder." *Trends Immunol* 43 (8): 630–39. https://doi.org/10.1016/j.it.2022.06.003.

Stasik, Sebastian, Jan M Middeke, Michael Kramer, Christoph Röllig, Alwin Krämer, Sebastian Scholl, Andreas Hochhaus, et al. 2020. "EZH2 Mutations and Impact on Clinical Outcome: An Analysis in 1,604 Patients with Newly Diagnosed Acute Myeloid Leukemia." *Haematologica* 105 (5): e228–e231. https://doi.org/10.3324/haematol.2019.222323.

Steele, Christopher D, Ammal Abbasi, S M Ashiqul Islam, Amy L Bowes, Azhar Khandekar, Kerstin Haase, Shadi Hames-Fathi, et al. 2022. "Signatures of Copy Number Alterations in Human Cancer." *Nature* 606 (7916): 984–91. https://doi.org/10.1038/s41586-022-04738-6.

Steensel, Bas van, Ulrich Braunschweig, Guillaume J Filion, Menzies Chen, Joke G van Bemmel, and Trey Ideker. 2010. "Bayesian Network Analysis of Targeting Interactions in Chromatin." *Genome Res* 20 (2): 190–200. https://doi.org/10.1101/gr.098822.109.

Stuart, Tim, Andrew Butler, Paul Hoffman, Christoph Hafemeister, Efthymia Papalexi, William M Mauck 3rd, Yuhan Hao, Marlon Stoeckius, Peter Smibert, and Rahul

Satija. 2019. "Comprehensive Integration of Single-Cell Data." *Cell* 177 (7): 1888–1902.e21. https://doi.org/10.1016/j.cell.2019.05.031.

Stuart, Tim, Avi Srivastava, Shaista Madad, Caleb A Lareau, and Rahul Satija. 2021. "Single-Cell Chromatin State Analysis with Signac." *Nat Methods* 18 (11): 1333–41. https://doi.org/10.1038/s41592-021-01282-5.

Sturm, Dominik, Hendrik Witt, Volker Hovestadt, Dong-Anh Khuong-Quang, David T W Jones, Carolin Konermann, Elke Pfaff, et al. 2012. "Hotspot Mutations in H3f3a and Idh1 Define Distinct Epigenetic and Biological Subgroups of Glioblastoma." *Cancer Cell* 22 (4): 425–37. https://doi.org/10.1016/j.ccr.2012.08.024.

Su, Chengwei, Angeline Andrew, Margaret R Karagas, and Mark E Borsuk. 2013. "Using Bayesian Networks to Discover Relations Between Genes, Environment, and Disease." *BioData Min* 6 (1): 6. https://doi.org/10.1186/1756-0381-6-6.

Su, Jianzhong, Yung-Hsin Huang, Xiaodong Cui, Xinyu Wang, Xiaotian Zhang, Yong Lei, Jianfeng Xu, et al. 2018. "Homeobox Oncogene Activation by Pan-Cancer Dna Hypermethylation." *Genome Biol* 19 (1): 108. https://doi.org/10.1186/s13059-018-1492-3.

Tahiliani, Mamta, Kian Peng Koh, Yinghua Shen, William A Pastor, Hozefa Bandukwala, Yevgeny Brudno, Suneet Agarwal, et al. 2009. "Conversion of 5-Methylcytosine to 5-Hydroxymethylcytosine in Mammalian Dna by Mll Partner Tet1." *Science* 324 (5929): 930–5. https://doi.org/10.1126/science.1170116.

Tan, Jimin, Nina Shenker-Tauris, Javier Rodriguez-Hernaez, Eric Wang, Theodore Sakellaropoulos, Francesco Boccalatte, Palaniraja Thandapani, et al. 2023. "Cell-Type-Specific Prediction of 3D Chromatin Organization Enables High-Throughput in Silico Genetic Screening." *Nat Biotechnol*, January. https://doi.org/10.1038/s41587-022-01612-8.

Tao, Yong, Byunghak Kang, Daniel A Petkovich, Yuba R Bhandari, Julie In, Genevieve Stein-O'Brien, Xiangqian Kong, et al. 2019. "Aging-Like Spontaneous Epigenetic

Silencing Facilitates Wnt Activation, Stemness, and Brafv600e-Induced Tumorigenesis." *Cancer Cell* 35 (2): 315–328.e6. https://doi.org/10.1016/j.ccell.2019.01.005.

Tasic, Bosiljka, Zizhen Yao, Lucas T Graybuck, Kimberly A Smith, Thuc Nghi Nguyen, Darren Bertagnolli, Jeff Goldy, et al. 2018. "Shared and Distinct Transcriptomic Cell Types Across Neocortical Areas." *Nature* 563 (7729): 72–78. https://doi.org/10.1038/s41586-018-0654-5.

Tatton-Brown, Katrina, Sheila Seal, Elise Ruark, Jenny Harmer, Emma Ramsay, Silvana Del Vecchio Duarte, Anna Zachariou, et al. 2014. "Mutations in the Dna Methyltransferase Gene Dnmt3a Cause an Overgrowth Syndrome with Intellectual Disability." *Nat Genet* 46 (4): 385–8. https://doi.org/10.1038/ng.2917.

Teodoridis, Jens M, Catriona Hardie, and Robert Brown. 2008. "CpG Island Methylator Phenotype (Cimp) in Cancer: Causes and Implications." *Cancer Lett* 268 (2): 177–86. https://doi.org/10.1016/j.canlet.2008.03.022.

Thurman, Robert E, Eric Rynes, Richard Humbert, Jeff Vierstra, Matthew T Maurano, Eric Haugen, Nathan C Sheffield, et al. 2012. "The Accessible Chromatin Landscape of the Human Genome." *Nature* 489 (7414): 75–82. https://doi.org/10.1038/nature11232.

Toyota, M, N Ahuja, M Ohe-Toyota, J G Herman, S B Baylin, and J P Issa. 1999. "CpG Island Methylator Phenotype in Colorectal Cancer." *Proc Natl Acad Sci U S A* 96 (15): 8681–6. https://doi.org/10.1073/pnas.96.15.8681.

Toyota, M, K J Kopecky, M O Toyota, K W Jair, C L Willman, and J P Issa. 2001. "Methylation Profiling in Acute Myeloid Leukemia." *Blood* 97 (9): 2823–9. https://doi.org/10.1182/blood.v97.9.2823.

Tulstrup, Morten, Mette Soerensen, Jakob Werner Hansen, Linn Gillberg, Maria Needhamsen, Katja Kaastrup, Kristian Helin, Kaare Christensen, Joachim Weischenfeldt,

and Kirsten Grønbæk. 2021. "TET2 Mutations Are Associated with Hypermethylation at Key Regulatory Enhancers in Normal and Malignant Hematopoiesis." *Nat Commun* 12 (1): 6061. https://doi.org/10.1038/s41467-021-26093-2.

Turcan, Sevin, Daniel Rohle, Anuj Goenka, Logan A Walsh, Fang Fang, Emrullah Yilmaz, Carl Campos, et al. 2012. "IDH1 Mutation Is Sufficient to Establish the Glioma Hypermethylator Phenotype." *Nature* 483 (7390): 479–83. https://doi.org/10.1038/nature10866.

Uddin, Md Sahab, Abdullah Al Mamun, Badrah S Alghamdi, Devesh Tewari, Philippe Jeandet, Md Shahid Sarwar, and Ghulam Md Ashraf. 2022. "Epigenetics of Glioblastoma Multiforme: From Molecular Mechanisms to Therapeutic Approaches." *Semin Cancer Biol* 83 (August): 100–120. https://doi.org/10.1016/j.semcancer.2020.12.015.

Vaissière, Thomas, Carla Sawan, and Zdenko Herceg. 2008. "Epigenetic Interplay Between Histone Modifications and Dna Methylation in Gene Silencing." *Mutat Res* 659 (1-2): 40–48. https://doi.org/10.1016/j.mrrev.2008.02.004.

Valcour, Victor, Thep Chalermchai, Napapon Sailasuta, Mary Marovich, Sukalaya Lerdlum, Duanghathai Suttichom, Nijasri C Suwanwela, et al. 2012. "Central Nervous System Viral Invasion and Inflammation During Acute HIV Infection." *J Infect Dis* 206 (2): 275–82. https://doi.org/10.1093/infdis/jis326.

Vansant, Gerlinde, Heng-Chang Chen, Eduard Zorita, Katerina Trejbalová, Dalibor Miklı'k, Guillaume Filion, and Zeger Debyser. 2020. "The Chromatin Landscape at the HIV-1 Provirus Integration Site Determines Viral Expression." *Nucleic Acids Res* 48 (14): 7801–17. https://doi.org/10.1093/nar/gkaa536.

Vaquerizas, Juan M, Sarah K Kummerfeld, Sarah A Teichmann, and Nicholas M Luscombe. 2009. "A Census of Human Transcription Factors: Function, Expression and Evolution." *Nat Rev Genet* 10 (4): 252–63. https://doi.org/10.1038/nrg2538.

Verhaak, Roel G W, Katherine A Hoadley, Elizabeth Purdom, Victoria Wang, Yuan Qi, Matthew D Wilkerson, C Ryan Miller, et al. 2010. "Integrated Genomic Analysis Identifies Clinically Relevant Subtypes of Glioblastoma Characterized by Abnormalities in Pdgfra, Idh1, Egfr, and Nf1." *Cancer Cell* 17 (1): 98–110. https://doi.org/10.1016/j.ccr.2009.12.020.

Waddington, C H. 2012. "The Epigenotype. 1942." *Int J Epidemiol* 41 (1): 10–13. https://doi.org/10.1093/ije/dyr184.

Wagner, Thor A, Sherry McLaughlin, Kavita Garg, Charles Y K Cheung, Brendan B Larsen, Sheila Styrchak, Hannah C Huang, Paul T Edlefsen, James I Mullins, and Lisa M Frenkel. 2014. "HIV Latency. Proliferation of Cells with HIV Integrated into Cancer Genes Contributes to Persistent Infection." *Science* 345 (6196): 570–3. https://doi.org/10.1126/science.1256304.

Wang, Gary P, Angela Ciuffi, Jeremy Leipzig, Charles C Berry, and Frederic D Bushman. 2007. "HIV Integration Site Selection: Analysis by Massively Parallel Pyrosequencing Reveals Association with Epigenetic Modifications." *Genome Res* 17 (8): 1186–94. https://doi.org/10.1101/gr.6286907.

Wang, Meng, Benjamin D Sunkel, William C Ray, and Benjamin Z Stanton. 2022. "Chromatin Structure in Cancer." *BMC Mol Cell Biol* 23 (1): 35. https://doi.org/10.1186/s12860-022-00433-6.

Wang, Qianghu, Baoli Hu, Xin Hu, Hoon Kim, Massimo Squatrito, Lisa Scarpace, Ana C deCarvalho, et al. 2017. "Tumor Evolution of Glioma-Intrinsic Gene Expression Subtypes Associates with Immunological Changes in the Microenvironment." *Cancer Cell* 32 (1): 42–56.e6. https://doi.org/10.1016/j.ccell.2017.06.003.

Wang, Yanli, Fan Song, Bo Zhang, Lijun Zhang, Jie Xu, Da Kuang, Daofeng Li, et al. 2018. "The 3D Genome Browser: A Web-Based Browser for Visualizing 3D Genome Organization and Long-Range Chromatin Interactions." *Genome Biol* 19 (1): 151. https://doi.org/10.1186/s13059-018-1519-9.

Weisenberger, Daniel J, Kimberly D Siegmund, Mihaela Campan, Joanne Young, Tiffany I Long, Mark A Faasse, Gyeong Hoon Kang, et al. 2006. "CpG Island Methylator Phenotype Underlies Sporadic Microsatellite Instability and Is Tightly Associated with Braf Mutation in Colorectal Cancer." *Nat Genet* 38 (7): 787–93. https://doi.org/10.1038/ng1834.

Wells, Daria W, Shuang Guo, Wei Shao, Michael J Bale, John M Coffin, Stephen H Hughes, and Xiaolin Wu. 2020. "An Analytical Pipeline for Identifying and Mapping the Integration Sites of HIV and Other Retroviruses." *BMC Genomics* 21 (1): 216. https://doi.org/10.1186/s12864-020-6647-4.

Weth, Oliver, Christine Paprotka, Katharina Günther, Astrid Schulte, Manuel Baierl, Joerg Leers, Niels Galjart, and Rainer Renkawitz. 2014. "CTCF Induces Histone Variant Incorporation, Erases the H3k27me3 Histone Mark and Opens Chromatin." *Nucleic Acids Res* 42 (19): 11941–51. https://doi.org/10.1093/nar/gku937.

WHO. 2021. *Global Progress Report on HIV, Viral Hepatitis and Sexually Transmitted Infections, 2021.* Edited by WHO. Global HIV, Hepatitis and Sexually Transmitted Infections Programmes. WHO.

Wilson, V L, and P A Jones. 1983. "DNA Methylation Decreases in Aging but Not in Immortal Cells." *Science* 220 (4601): 1055–7. https://doi.org/10.1126/science.6844925.

Wirsching, Hans-Georg, Evanthia Galanis, and Michael Weller. 2016. "Glioblastoma." *Handb Clin Neurol* 134: 381–97. https://doi.org/10.1016/B978-0-12-802997-8.00023-2.

Wit, Elzo de, Erica S M Vos, Sjoerd J B Holwerda, Christian Valdes-Quezada, Marjon J A M Verstegen, Hans Teunissen, Erik Splinter, Patrick J Wijchers, Peter H L Krijger, and Wouter de Laat. 2015. "CTCF Binding Polarity Determines Chromatin Looping." *Mol Cell* 60 (4): 676–84. https://doi.org/10.1016/j.molcel.2015.09.023.

Wong, Joseph K, and Steven A Yukl. 2016. "Tissue Reservoirs of HIV." *Curr Opin HIV AIDS* 11 (4): 362–70. https://doi.org/10.1097/COH.0000000000000293.

Wout, A B van't, L J Ran, C L Kuiken, N A Kootstra, S T Pals, and H Schuitemaker. 1998. "Analysis of the Temporal Relationship Between Human Immunodeficiency Virus Type 1 Quasispecies in Sequential Blood Samples and Various Organs Obtained at Autopsy." *J Virol* 72 (1): 488–96. https://doi.org/10.1128/JVI.72.1.488-496.1998.

Wu, Dai-Ying, Danielle Bittencourt, Michael R Stallcup, and Kimberly D Siegmund. 2015. "Identifying Differential Transcription Factor Binding in Chip-Seq." *Front Genet* 6: 169. https://doi.org/10.3389/fgene.2015.00169.

Wu, Yonghe, Michael Fletcher, Zuguang Gu, Qi Wang, Barbara Costa, Anna Bertoni, Ka-Hou Man, et al. 2020. "Glioblastoma Epigenome Profiling Identifies Sox10 as a Master Regulator of Molecular Tumour Subtype." *Nat Commun* 11 (1): 6434. https://doi.org/10.1038/s41467-020-20225-w.

Wutz, Gordana, Csilla Várnai, Kota Nagasaka, David A Cisneros, Roman R Stocsits, Wen Tang, Stefan Schoenfelder, et al. 2017. "Topologically Associating Domains and Chromatin Loops Depend on Cohesin and Are Regulated by Ctcf, Wapl, and Pds5 Proteins." *EMBO J* 36 (24): 3573–99. https://doi.org/10.15252/embj.201798004.

Xie, Wei, Matthew D Schultz, Ryan Lister, Zhonggang Hou, Nisha Rajagopal, Pradipta Ray, John W Whitaker, et al. 2013. "Epigenomic Analysis of Multilineage Differentiation of Human Embryonic Stem Cells." *Cell* 153 (5): 1134–48. https://doi.org/10.1016/j.cell.2013.04.022.

Yang, Jae-Hyun, Motoshi Hayano, Patrick T Griffin, João A Amorim, Michael S Bonkowski, John K Apostolides, Elias L Salfati, et al. 2023. "Loss of Epigenetic Information as a Cause of Mammalian Aging." *Cell*, January. https://doi.org/10.1016/j.cell.2022.12.027.

Yang, Qi, Nian Jiang, Han Zou, Xuning Fan, Tao Liu, Xi Huang, Siyi Wanggou, and Xuejun Li. 2022. "Alterations in 3D Chromatin Organization Contribute to Tumorigenesis of Egfr-Amplified Glioblastoma." *Comput Struct Biotechnol J* 20: 1967–78. https://doi.org/10.1016/j.csbj.2022.04.007.

Yang, Yang, Yang Zhang, Bing Ren, Jesse R Dixon, and Jian Ma. 2019. "Comparing 3D Genome Organization in Multiple Species Using Phylo-Hmrf." *Cell Syst* 8 (6): 494–505.e14. https://doi.org/10.1016/j.cels.2019.05.011.

Yates, Josephine, and Valentina Boeva. 2022. "Deciphering the Etiology and Role in Oncogenic Transformation of the Cpg Island Methylator Phenotype: A Pan-Cancer Analysis." *Brief Bioinform* 23 (2). https://doi.org/10.1093/bib/bbab610.

Ye, Dan, Shenghong Ma, Yue Xiong, and Kun-Liang Guan. 2013. "R-2-Hydroxyglutarate as the Key Effector of Idh Mutations Promoting Oncogenesis." *Cancer Cell* 23 (3): 274–6. https://doi.org/10.1016/j.ccr.2013.03.005.

Yoon, John K, Joseph R Holloway, Daria W Wells, Machika Kaku, David Jetton, Rebecca Brown, and John M Coffin. 2020. "HIV Proviral Dna Integration Can Drive T Cell Growth Ex Vivo." *Proc Natl Acad Sci U S A* 117 (52): 32880–2. https://doi.org/10.1073/pnas.2013194117.

Youn, Hong-Duk. 2017. "Methylation and Demethylation of DNA and Histones in Chromatin: The Most Complicated Epigenetic Marker." *Exp Mol Med* 49 (4): e321. https://doi.org/10.1038/emm.2017.38.

Yu, Da-Hai, Carol Ware, Robert A Waterland, Jiexin Zhang, Miao-Hsueh Chen, Manasi Gadkari, Govindarajan Kunde-Ramamoorthy, Lagina M Nosavanh, and Lanlan Shen. 2013. "Developmentally Programmed 3' CpG Island Methylation Confers Tissue-and Cell-Type-Specific Transcriptional Activation." *Mol Cell Biol* 33 (9): 1845–58. https://doi.org/10.1128/MCB.01124-12.

Yu, Jing, V Anne Smith, Paul P Wang, Alexander J Hartemink, and Erich D Jarvis.

2004. “Advances to Bayesian Network Inference for Generating Causal Networks from Observational Biological Data.” *Bioinformatics* 20 (18): 3594–3603. https://doi.org/10.1093/bioinformatics/bth448.

Zemach, Assaf, Ivy E McDaniel, Pedro Silva, and Daniel Zilberman. 2010. “Genome-Wide Evolutionary Analysis of Eukaryotic DNA Methylation.” *Science* 328 (5980): 916–9. https://doi.org/10.1126/science.1186366.

Zhang, Pei, Qin Xia, Liqun Liu, Shouwei Li, and Lei Dong. 2020. “Current Opinion on Molecular Characterization for GBM Classification in Guiding Clinical Diagnosis, Prognosis, and Therapy.” *Front Mol Biosci* 7: 562798. https://doi.org/10.3389/fmolb.2020.562798.

Zhang, Qian, and Xuetao Cao. 2019. “Epigenetic Regulation of the Innate Immune Response to Infection.” *Nat Rev Immunol* 19 (7): 417–32. https://doi.org/10.1038/s41577-019-0151-6.

Zhang, Yong, Tao Liu, Clifford A Meyer, Jérôme Eeckhoute, David S Johnson, Bradley E Bernstein, Chad Nusbaum, et al. 2008. “Model-Based Analysis of Chip-Seq (Macs).” *Genome Biol* 9 (9): R137. https://doi.org/10.1186/gb-2008-9-9-r137.

Zhang, Yu, and Ross C Hardison. 2017. “Accurate and Reproducible Functional Maps in 127 Human Cell Types via 2D Genome Segmentation.” *Nucleic Acids Res* 45 (17): 9823–36. https://doi.org/10.1093/nar/gkx659.

Zhou, D., and K.D. Robertson. 2016. “Chapter 24 - Role of DNA Methylation in Genome Stability.” In *Genome Stability*, edited by Igor Kovalchuk and Olga Kovalchuk, 409–24. Boston: Academic Press. https://doi.org/https://doi.org/10.1016/B978-0-12-803309-8.00024-0.

Zhu, Jiang, Mazhar Adli, James Y Zou, Griet Verstappen, Michael Coyne, Xiaolan Zhang, Timothy Durham, et al. 2013. “Genome-Wide Chromatin State Transitions Associated with Developmental and Environmental Cues.” *Cell* 152 (3): 642–54.

https://doi.org/10.1016/j.cell.2012.12.033.

Zhu, Lihua J, Claude Gazin, Nathan D Lawson, Hervé Pagès, Simon M Lin, David S Lapointe, and Michael R Green. 2010. "ChIPpeakAnno: A Bioconductor Package to Annotate Chip-Seq and Chip-Chip Data." *BMC Bioinformatics* 11 (May): 237. https://doi.org/10.1186/1471-2105-11-237.

Zoghbi, Huda Y, and Arthur L Beaudet. 2016. "Epigenetics and Human Disease." *Cold Spring Harb Perspect Biol* 8 (2): a019497. https://doi.org/10.1101/cshperspect.a019497.

Zong, Hui, Luis F Parada, and Suzanne J Baker. 2015. "Cell of Origin for Malignant Gliomas and Its Implication in Therapeutic Development." *Cold Spring Harb Perspect Biol* 7 (5). https://doi.org/10.1101/cshperspect.a020610.

Zuin, Jessica, Jesse R Dixon, Michael I J A van der Reijden, Zhen Ye, Petros Kolovos, Rutger W W Brouwer, Mariëtte P C van de Corput, et al. 2014. "Cohesin and Ctcf Differentially Affect Chromatin Architecture and Gene Expression in Human Cells." *Proc Natl Acad Sci U S A* 111 (3): 996–1001. https://doi.org/10.1073/pnas.1317788111.

Appendix

Appendix A. Datasets used for the analysis present in Chapter 2

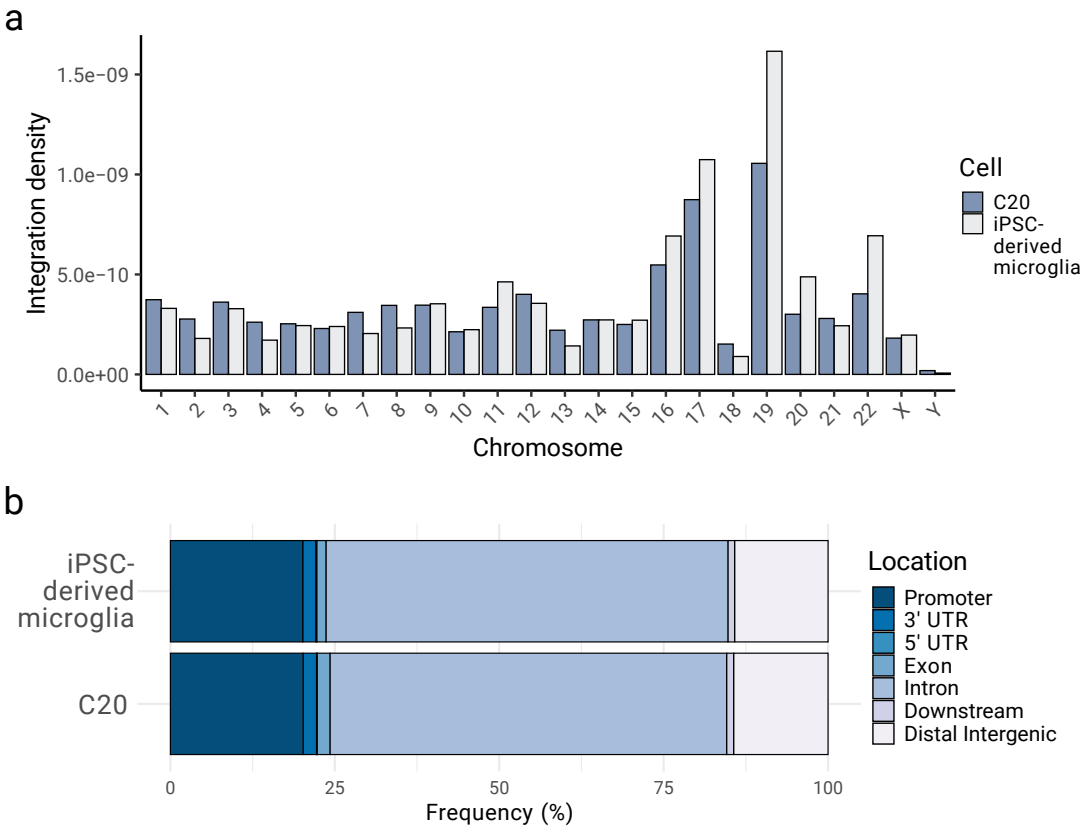
Type	Cell	N or Samples	Assay	Source
Combined dataset of HIV-1 integration sites	CD4+ T cell	13544	LM-PCR/LAM-PCR	Lucic et al, 2019 Brady et al, 2009 Wagner et al, 2014 Maldarelli et al, 2014 Han et al, 2004 Cohn et al, 2015 Ikeda et al, 2007
HIV-1 integration sites	Macrophages	987	nrLAM-PCR	Kok et al, 2016
HIV-1 integration sites	Microglia	4590	LM-PCR (paired-end and single-end)	Produced for this work
ChIP-seq of uninfected cells	Microglia	2	H3K36me3	Produced for this work
ChIP-seq of uninfected cells	Microglia	1+1	H3K27ac	Produced for this work
ChIP-seq of uninfected cells	Microglia	2	H3K4me1	Produced for this work
ChIP-seq of uninfected cells	Microglia	2	H3K27me3	Produced for this work
ChIP-seq of uninfected cells	Microglia	2	H3K9me3	Produced for this work
ChIP-seq of uninfected cells	Microglia	1+1	H3K9me2	Produced for this work
ChIP-seq of uninfected cells	Microglia	1+1+1	Input	Produced for this work
ATAC-seq of uninfected cells	Microglia	2	ATAC-seq/Chromatin accessibility	Produced for this work
ATAC-seq of infected cells (latent)	Microglia	2	ATAC-seq/Chromatin accessibility	Produced for this work
ATAC-seq of infected cells (active)	Microglia	2	ATAC-seq/Chromatin accessibility	Produced for this work
RNA-seq of uninfected cells	Microglia	3	Expression	Produced for this work
RNA-seq of uninfected cells	CD4+T cells	3	Expression	Lucic et al, 2019

Appendix B. ATAC-seq files used for training of the TAD boundary RF model (source: ENCODE)

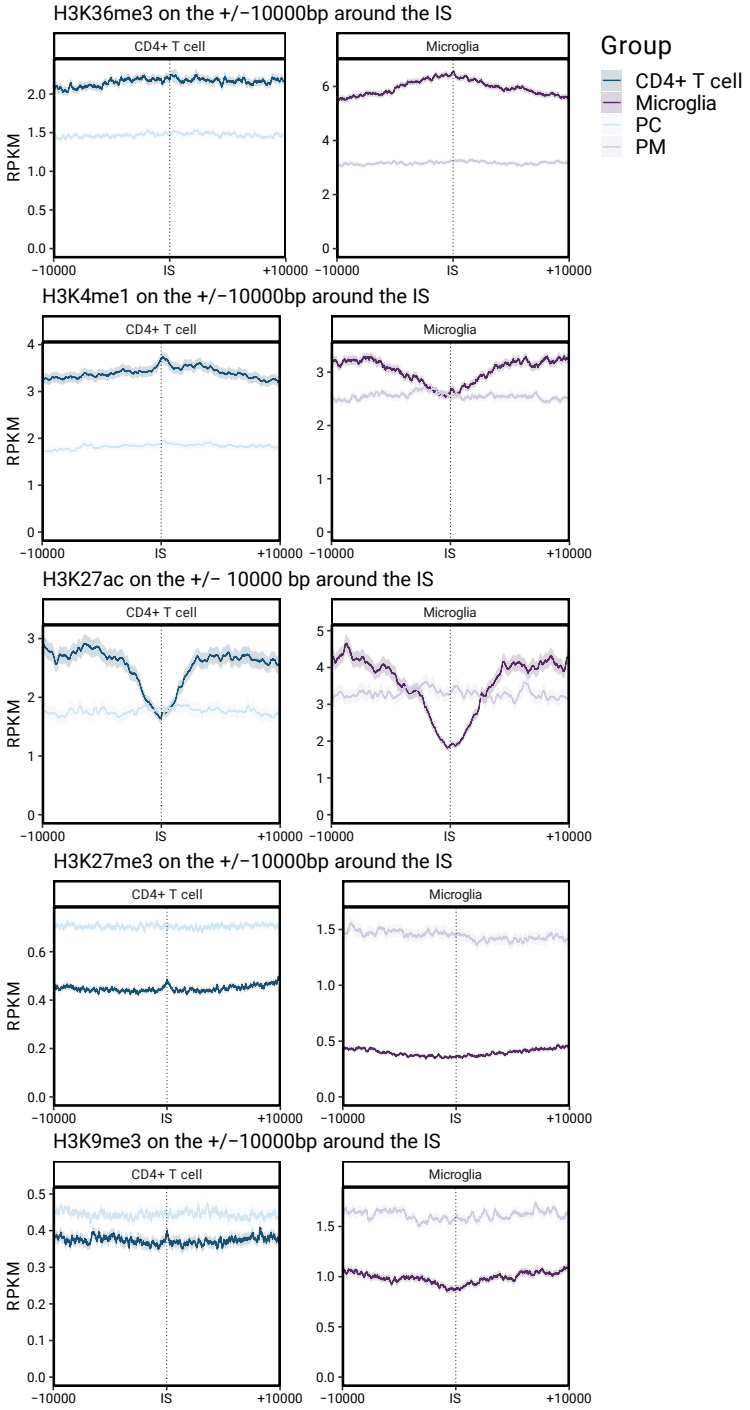
ENCODE ID	File ID BAM	Cell Tissue	Isogenic replicate index	Isogenic replicates	Assay	Genome
ENCSR032RGS	ENCFF607DTB	A549	1	Y	ATAC-seq	hg38
	ENCFF701BDT	A549	2	Y	ATAC-seq	hg38
	ENCFF616DYV	A549	3	Y	ATAC-seq	hg38
ENCSR637XSC	ENCFF981FXV	GM12878	1	Y	ATAC-seq	hg38
	ENCFF962FMH	GM12878	2	Y	ATAC-seq	hg38
	ENCFF440GRZ	GM12878	3	Y	ATAC-seq	hg38
ENCSR291GJU	ENCFF990VCP	HepG2	1	Y	ATAC-seq	hg38
	ENCFF624SON	HepG2	2	Y	ATAC-seq	hg38
	ENCFF926KFU	HepG2	3	Y	ATAC-seq	hg38
ENCSR868FGK	ENCFF534DCE	K562	1	Y	ATAC-seq	hg38
	ENCFF128WZG	K562	2	Y	ATAC-seq	hg38
	ENCFF077FBI	K562	3	Y	ATAC-seq	hg38
ENCSR200OML	ENCFF848XMR	IMR-90	1	Y	ATAC-seq	hg38
	ENCFF715NAV	IMR-90	2	Y	ATAC-seq	hg38
ENCSR996ZCR	ENCFF454SNX	ovary	1	N	ATAC-seq	hg38
ENCSR392UJM	ENCFF925ACE	ovary	1	N	ATAC-seq	hg38
ENCSR227FVE	ENCFF615QSS	ovary	1	N	ATAC-seq	hg38
ENCSR286STX	ENCFF440WVI	leftVentricle	1	N	ATAC-seq	hg38
ENCSR846VPV	ENCFF456PVX	leftVentricle	1	N	ATAC-seq	hg38
ENCSR078EBD	ENCFF159BUD	spleen	1	N	ATAC-seq	hg38
ENCSR647AOY	ENCFF810ZGD	lung	1	N	ATAC-seq	hg38

Appendix C. TADs used for class labels in the TAD boundary RF model (source: 3D Genome Browser)

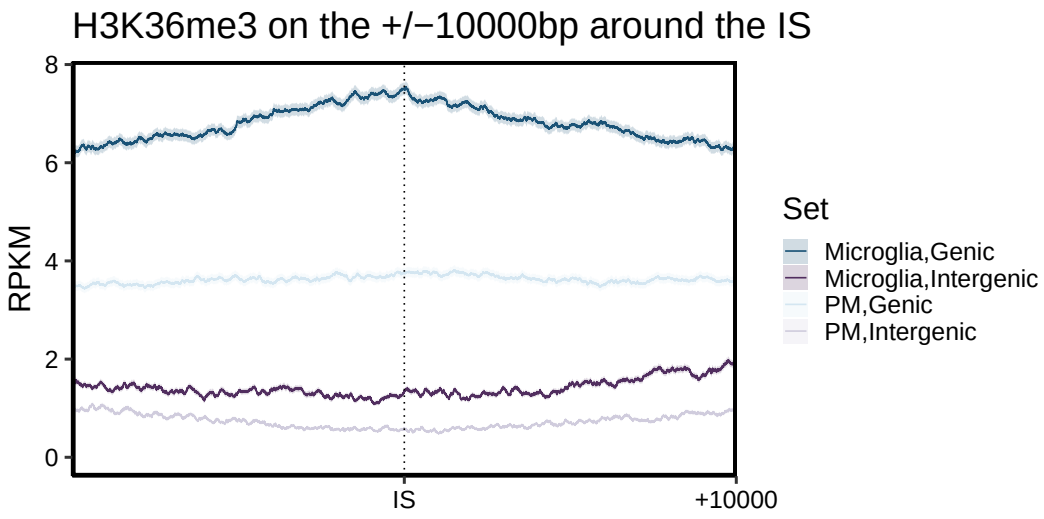
Cell/Tissue	Number of pooled TADs
A549	1797
GM12878	548
HepG2	2878
K562	244
IMR-90	493
ovary	1180
left ventricle	1096
spleen	2336
lung	1392



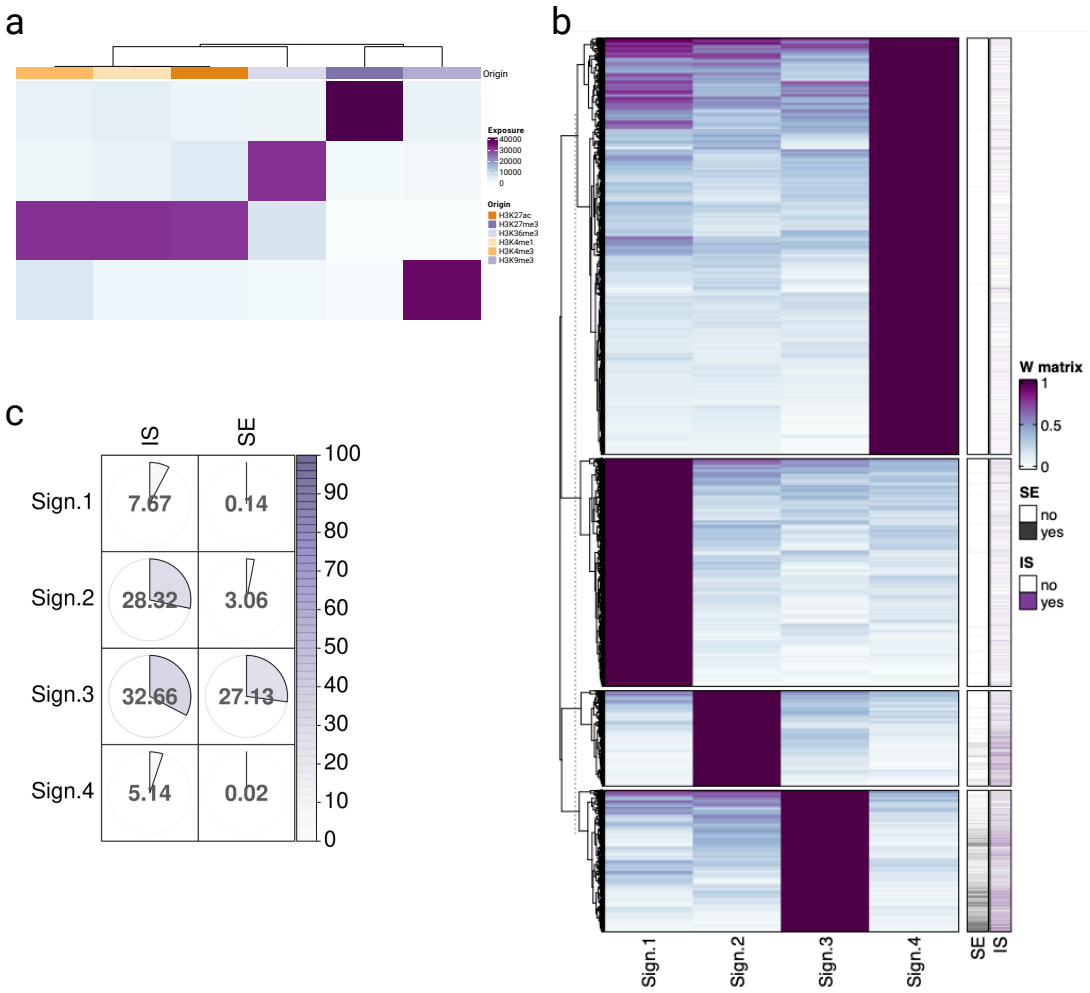
Appendix D. Comparison between integration patterns in the C20 microglial cell line and iPSC-derived microglia. [a] Normalised chromosomic distribution of IS on the C20 microglial cell line in comparison with iPSC-derived microglia. [b] Genomic features of integration in C20 in comparison with iPSC-derived microglia. Locations are labelled by colour.



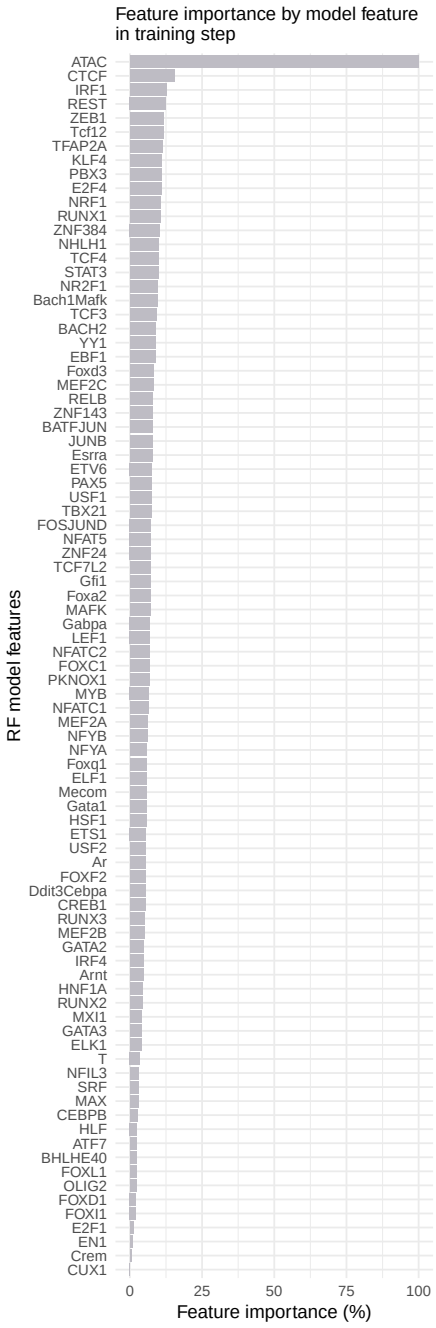
Appendix E. Epigenetic profile for different histone modifications (RPKM) on the IS vicinity (10KB upstream and 10KB downstream) in both microglia and CD4+ T cells. Averaged signal over the IS set (cell labelled, in full colour) and over a matched phantom IS set (PM for microglia and PC for CD4+ T cells, in transparency). Confidence interval (95%) is shown in shaded color.



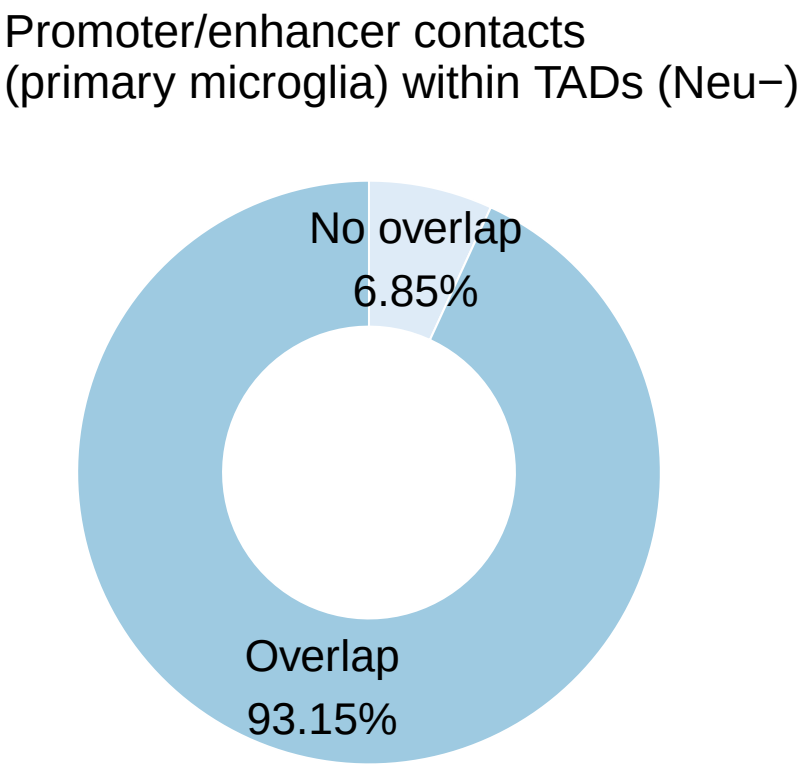
Appendix F. Epigenetic profile for H3K36me3 (RPKM) on the IS vicinity (10KB upstream and 10KB downstream) as before integration. Averaged signal over the IS set (IS, in full colour) and over a matched phantom IS set (PM, in transparency). IS inside genes (in blue) and in the intergenic space (in purple) are compared. Confidence interval (95%) is shown in shaded color.



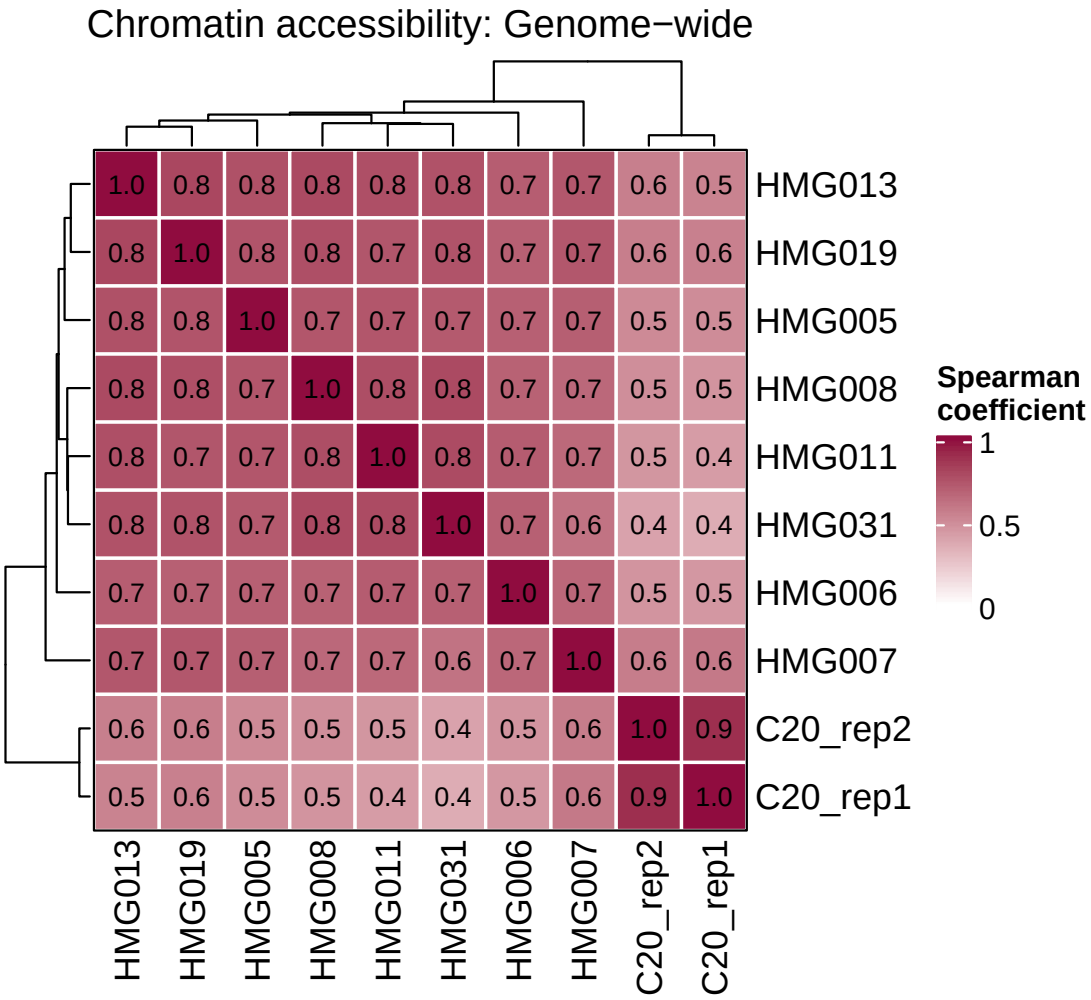
Appendix G. Signatures of HIV-1 integration on the CD4+ T cell model. [a] Exposure matrix H for NMF-derived signatures (k=4, in rows) based on ChIP-seq for 6 histone modifications. [b] Exposure matrix W for NMF-derived signatures (in columns) on all genome windows. Colour indicates if the window is assigned to one signature. Bars on the right indicate whether each window overlaps with IS (purple) and SE (black). [c] Representation of the overlap between each NMF-derived signature and the IS and SE set in CD4+ Y cells. Both colour and angle represent the overlap (%).



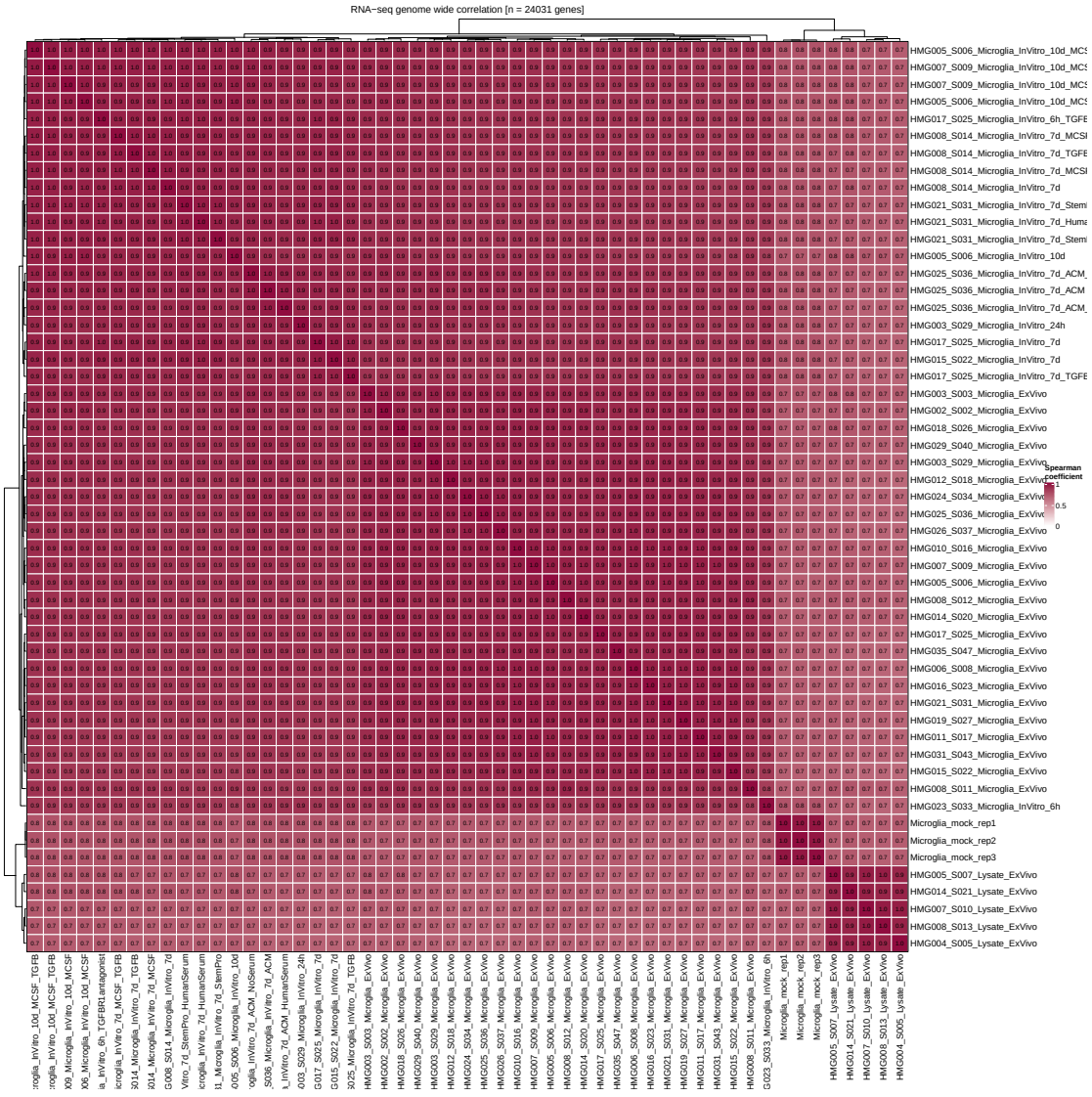
Appendix H. Feature importance of the RF model used to identify TFs most associated to TAD boundaries. Features are ordered by importance (%) as determined by *caret*.



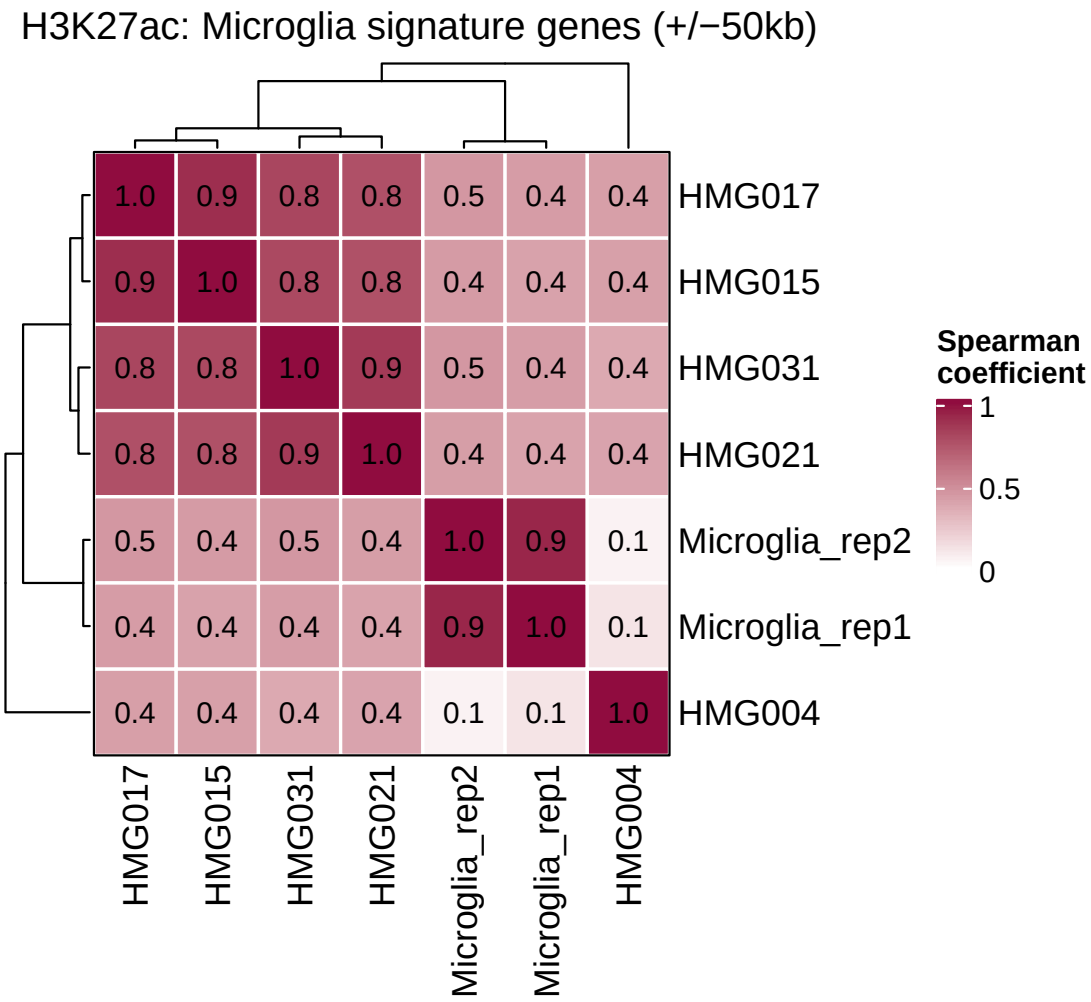
Appendix I. Fraction of promoter-enhancer contacts from primary microglia located within TADs from the Neu- cell population.



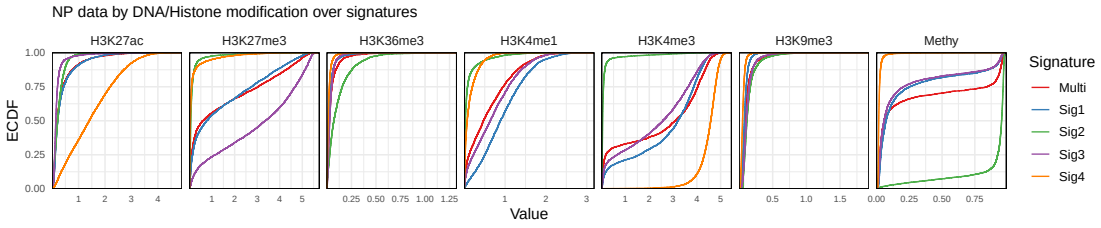
Appendix J. Correlation between the genome-wide chromatin accessibility in the C20 microglial cell line samples with primary microglia. Spearman correlation values are labelled by colour gradient (right).



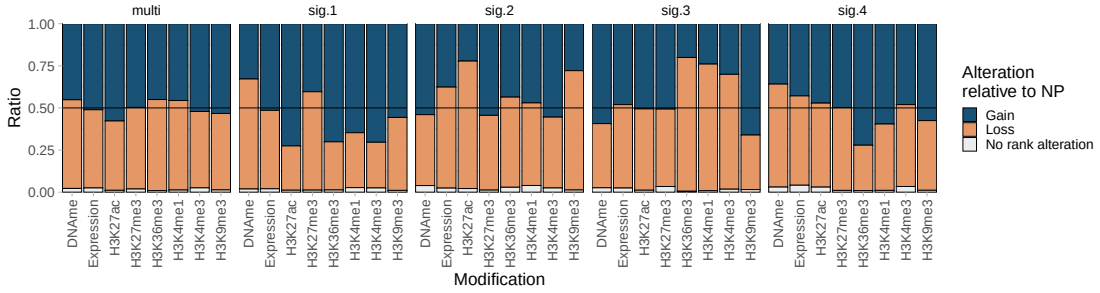
Appendix K. Correlation between the expression of protein-coding genes in the C20 microglial cell line samples with primary microglia. Spearman correlation values are labelled by colour gradient (right).



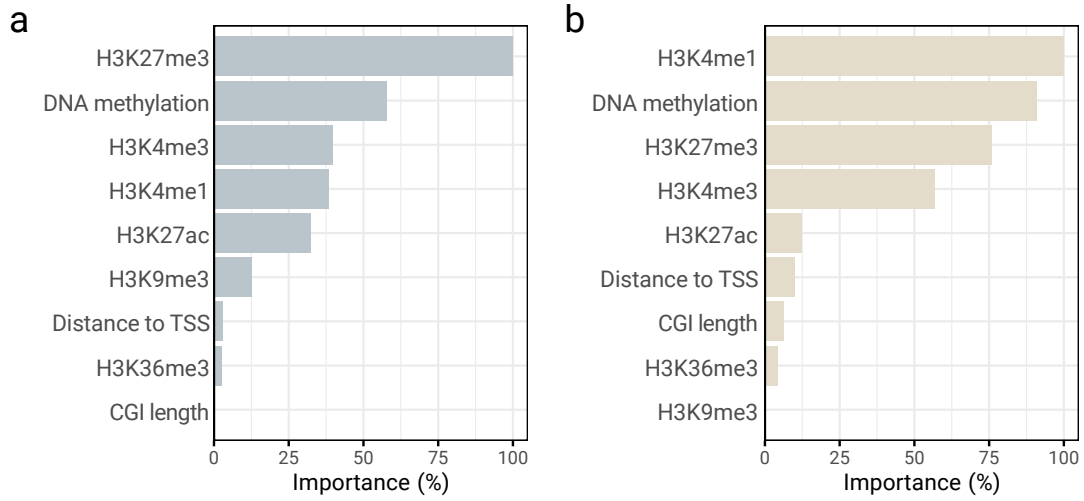
Appendix L. Correlation between the genome-wide H3K27ac in the C20 microglial cell line samples with primary microglia. Spearman correlation values are labelled by colour gradient (right).



Appendix M. Epigenetic modifications on all CGIs for NPs by signature. Colours represent NMF signatures while each facet is relative to one epigenetic modification.



Appendix N. Rank-based comparison between NPs and GBM subtypes affected by CIMP within all CGIs by signature. Loss, gain, or no rank alteration of epigenomic and transcriptomic features between NP and GBM subtypes for each CGI signature.



Appendix O. RF for the IDH-CIMP and RTK2-CIMP distinction from non-CIMP CGIs. Ordered importance (x-axis; most to least important) for features used on the IDH- vs RTK2-CIMP model as exported by *caret*.