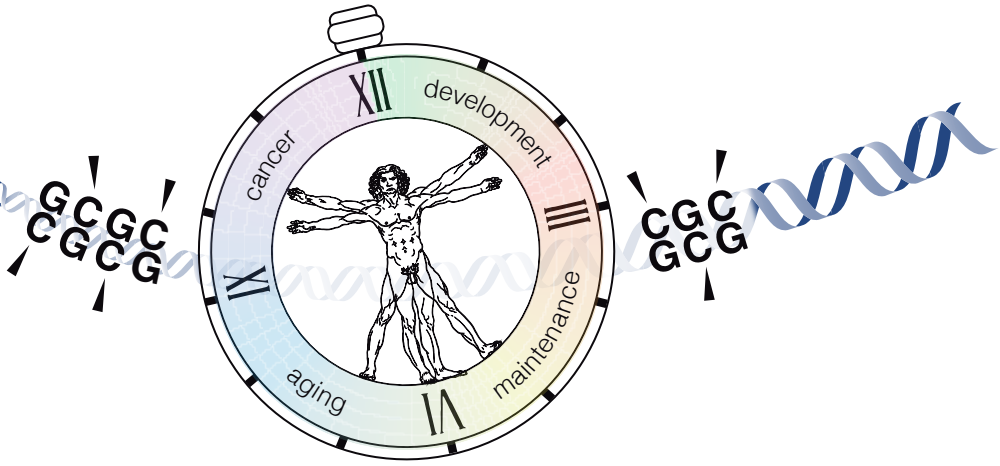
UNIVERSITAT DE BARCELONA

Tracing the developmental history of B-cell tumors by DNA methylation

Martí Duran Ferrer

Tracing the developmental history of B-cell tumors by DNA methylation

Martí Duran Ferrer



DEPARTMENT OF BASIC CLINICAL PRACTICE
FACULTY OF MEDICINE, UNIVERSITY OF BARCELONA
INSTITUT D'INVESTIGACIONS BIOMÈDIQUES AUGUST PI I SUNYER (IDIBAPS)

Tracing the developmental history of B-cell tumors by DNA methylation

Doctoral Thesis presented by **Martí Duran Ferrer**
to obtain the degree of Doctor in Biomedicine
PhD Program in Biomedicine
FACULTY OF MEDICINE, UNIVERSITY OF BARCELONA

Doctoral advisors:
Prof. **José Ignacio Martín Subero**
Prof. **Elías Campo Güerri**

Barcelona, 2020

Doctoral thesis presented by:

Martí Duran Ferrer

Doctoral advisors:

Director: Prof. **José Ignacio Martín Subero**

Tutor: Prof. **Elías Campo Güerri**

The work presented in this thesis has been funded by the Spanish Ministry of Economy and Competitivity through the Instituto de Salud Carlos III (Proyecto Genoma de la Leucemia Linfocítica Crónica) and by project number SAF2017-86126-R), as well as the EU-funded Blueprint Consortium (grant agreement 282510), the Fundació La Marató de TV3 (grant agreement 20132130) and the World Wide Cancer Research Foundation (grant agreement 16-1285). This work has been carried out at the Center Esther Koplowitz (CEK).

In the middle of difficulty lies opportunity

-Albert Einstein

ABSTRACT

The DNA from our cells carries the genetic information to build a human being. This information is interpreted through epigenetic marks that lead to tightly coordinated and cell type-specific gene expression programs. One of these marks is DNA methylation, which has been widely reported to regulate gene expression both in physiological and pathological conditions. Initial studies in cancer identified gene promoter methylation as an alternative to genetic alterations in the silencing of tumor suppressor genes. Nonetheless, genome-wide studies published in the last decade uncovered that the majority of DNA methylation changes in cancer are not directly related to gene regulation, and thus do not have an apparent functional impact. Some of these studies focused on the DNA methylome during B-cell development and malignant transformation to major B-cell neoplasms. These reports unveiled a highly dynamic DNA methylome during B-cell development and provided new insights into the cellular origin, pathogenic mechanisms and clinical behavior of B-cell neoplasms.

Despite these previous studies targeting specific cancers, a holistic perspective of the DNA methylome during the entire normal cell differentiation program and derived neoplasms was missing. This holistic view was neither available for B cells nor for any other human cell lineage, and therefore was the main goal of this thesis. I exploited previously and newly generated data to dissect the sources of DNA methylation variability of major B-cell tumors spanning the whole maturation spectrum in the context of the entire normal B-cell development. These included B-cell acute lymphoblastic leukemia (ALL), mantle cell lymphoma (MCL) (Study 1), chronic lymphocytic leukemia (CLL), diffuse large B-cell lymphoma (DLBCL) and multiple myeloma (MM). This comprehensive approach using over 2,000 samples showed that the human DNA methylome is more dynamic than previously conceived and uncovered new clinico-biological insights (Study 2). I observed that B-cell tumors display both DNA methylation imprints of normal development and *de novo* DNA methylation aberrations, which set the basis to build an accurate diagnostic tool for 14 B-cell tumor subtypes with different clinical management. In line with previous knowledge, I identified that most of the DNA methylation changes taking place in individual patients were located in silent chromatin. Remarkably, I could relate this phenomenon to the proliferative history of normal and neoplastic B cells, whereby each cell division seemed to accumulate

transcriptionally-inert epigenetic imprints into the genome (Study 3). In general, mitotic activity simultaneously left hyper- and hypomethylation imprints, but some B-cell neoplasms preferentially gained or lost DNA methylation. Based on this data, I built the epiCMIT epigenetic mitotic clock considering both hyper- and hypomethylation imprints related to cell division, which significantly improved the performance of previously reported mitotic clocks. Noticeably, the proliferative history traced by the epiCMIT before treatment was highly predictive of future patient outcome, not only in the B-cell tumors studied but also in other human neoplasias. The accumulation of genetic alterations with positive selection increased the epiCMIT, but some specific drivers seemed to confer a particular proliferative advantage to CLL and MCL cells and distinguished patients with a marked adverse outcome (Study 4). Finally, I compared the epiCMIT mitotic clock with another type of epigenetic clock that estimates the chronological age of a person, the so-called Horvath clock. Interestingly, the epiCMIT was strongly associated with an accelerated epigenetic aging in B-cell tumors measured by the Horvath clock, suggesting a crosstalk between mitotic activity and the aging process (Study 3).

In summary, the wealth of data presented in this doctoral thesis uncovers DNA methylation as a holistic tracer of B-cell tumor developmental history and provides new clinico-biological insights for B-cell tumors and cancer in general.

RESUMEN

El ADN de nuestras células contiene la información genética necesaria para crear un ser humano. Esta información se interpreta mediante marcas epigenéticas que permiten regular con precisión y de manera coordinada la expresión de los genes en cada tipo celular. Una de estas marcas es la metilación del ADN, la cual se ha descrito ampliamente como reguladora de la expresión génica tanto en condiciones fisiológicas como en patológicas. Las primeras investigaciones en cáncer identificaron la metilación en los promotores de los genes como método alternativo a las mutaciones genéticas para silenciar los genes supresores de tumores. No obstante, estudios del genoma completo publicados en la última década han revelado que la mayoría de cambios en la metilación del ADN no están asociados a la regulación génica, y consecuentemente en apariencia no tienen un impacto funcional. Algunos de estos estudios se han centrado en la metilación del ADN durante el desarrollo normal de células B y su transformación hacia los principales tipos de neoplasias de células B. Estas investigaciones han revelado la presencia de un metiloma muy cambiante durante el desarrollo normal de células B y han proporcionado nuevos conocimientos sobre la célula de origen, los mecanismos patogénicos y el comportamiento clínico de los tumores de células B.

A pesar de la relevancia de estas investigaciones previas focalizadas en cada tumor, faltaba una visión holística de cómo se modula el metiloma del ADN durante un programa entero de desarrollo celular normal y sus neoplasias derivadas. Esta visión integral no estaba disponible ni en células B ni en ningún otro linaje celular humano, y consecuentemente ha constituido el objetivo principal de esta tesis doctoral. Aprovechando tanto datos previos como nuevos desgrané las fuentes de variabilidad de la metilación del ADN tanto en la diferenciación normal de las células B como en las principales neoplasias de células B derivadas de células B en diferentes estadios de maduración. Estas neoplasias incluían la leucemia linfoblástica aguda de células B (LLA), el linfoma de células del manto (LCM) (Estudio 1), la leucemia linfática crónica (LLC), el linfoma difuso de células B grandes (LBDCG) y el mieloma múltiple (MM). Este enfoque integrador de más de 2.000 muestras de pacientes mostró que el metiloma humano es considerablemente más dinámico de lo concebido previamente y desveló nuevos conocimientos biológicos y clínicos (Estudio 2). Observé que los tumores de célula B presentan tanto huellas de metilación del desarrollo normal de las células B sanas como

aberraciones de metilación del ADN adquiridas *de novo*. Estos dos patrones de metilación del ADN permitieron la creación de una herramienta diagnostica que clasifica con precisión un total de 14 subtipos de tumores de célula B que requieren un manejo clínico diferencial. En consonancia con investigaciones previas, identifiqué que la mayoría de los cambios en la metilación del ADN de los pacientes ocurren en regiones silenciadas de la cromatina. Notablemente, pude relacionar este fenómeno con la historia proliferativa de las células B normales y tumorales, dónde cada división celular parecía que dejaba huellas epigenéticas en el genoma sin repercusiones transcripcionales (Estudio 3). En general, observé que la actividad mitótica deja simultáneamente ganancias y pérdidas en la metilación del ADN, aunque algunos tipos de tumores de célula B mostraban un sesgo en una u otra dirección. Basado en estos datos, desarrollé un reloj mitótico epigenético llamado epiCMIT que considera tanto ganancias como pérdidas en la metilación del ADN asociadas a la división celular, lo cual representa una mejora considerable respeto a otros relojes mitóticos propuestos con anterioridad. Cabe destacar que la historia proliferativa trazada mediante el reloj mitótico epiCMIT antes del tratamiento de los pacientes fue altamente predictiva de su comportamiento clínico futuro no sólo en los tumores de células B, sino en otras neoplasias humanas. Observé que la acumulación de alteraciones genéticas con selección positiva incrementaba la magnitud del epiCMIT. Sin embargo, algunas de estas alteraciones parecían conferir una ventaja proliferativa mayor en las células de LLC y LCM, y permitían distinguir pacientes con un desenlace clínico muy adverso (Estudio 4). Finalmente, comparé el reloj mitótico epiCMIT con otro tipo de reloj epigenético que predice la edad cronológica de una persona de manera muy precisa, el llamado reloj de Horvath. Curiosamente, el epiCMIT estaba claramente asociado con una edad epigenética muy avanzada en los tumores de células B calculada mediante el reloj de Horvath, sugiriendo así una posible relación entre la actividad mitótica y el envejecimiento (Estudio 3).

En conclusión, los resultados presentados en esta tesis doctoral revelan que la metilación del ADN representa un trazador holístico del desarrollo de los tumores de célula B y proporciona nuevos conocimientos biológicos y clínicos, no solo en el contexto de estos tumores sino también para el cáncer en general.

RESUM

L'ADN de les nostres cèl·lules porta la informació genètica necessària per crear un ésser humà. Aquesta informació és interpretada a través de marques epigenètiques que permeten l' expressió diferencial i altament coordinada dels gens en cada tipus cel·lular. La metilació de l'ADN representa una d'aquestes marques, i ha estat amplament descrita com a reguladora gènica tan en condicions fisiòlogues com en patològiques. Les primeres investigacions en el càncer varen identificar la metilació als promotor dels gens com un mètode alternatiu a les mutacions genètiques per silenciar els gens supressors de tumors. No obstant, estudis del genoma complet durant la darrera dècada han revelat que la majoria de canvis de metilació de l'ADN no estan directament relacionats amb la regulació gènica ,i conseqüentment no tenen aparentment un impacte funcional. Alguns d'aquests estudis s'han centrat en la metilació de l'ADN durant el desenvolupament normal de les cèl·lules B i la seva transformació cap als tipus principals de tumors de cèl·lula B. Aquestes investigacions han descrit un metiloma molt dinàmic durant el desenvolupament de cèl·lules B sanes i han proporcionat nous coneixements sobre la cèl·lula d'origen, els mecanismes patogènics i el comportament clínic de les neoplàsies de cèl·lules B.

Malgrat la rellevància d'aquests estudis previs enfocats en cada tumor, faltava una visió holística de la metilació de l'ADN durant un programa sencer de desenvolupament de cèl·lules sanes y les seves neoplàsies derivades. Aquesta visió no estava disponible ni per les cèl·lules B ni per cap altre llinatge humà, i per tant era l'objectiu principal d'aquesta tesis doctoral. Emprant dades prèvies i generades expressament, vaig explorar les fonts de variabilitat en la metilació de l'ADN de les principals neoplàsies de cèl·lula B sorgides al llarg del desenvolupament complet de cèl·lules B sanes. Aquestes neoplàsies varen incloure la leucèmia linfoblàstica aguda de cèl·lules B (LLA), el limfoma de cèl·lules del mantell (LCM) (**Estudi 1**), la leucèmia limfocítica crònica (LLC), el limfoma difús de cèl·lules B grans (LDCB), i el mieloma múltiple (MM). Aquest enfocament integrador amb més de 2.000 mostres de pacients va desxifrar que el metiloma humà és notablement més dinàmic del que el concebíem, i va revelà nous coneixements biològics i clínics de les neoplasias de cèl·lules B (**Estudi 2**). Vaig identificar que els tumors de les cèl·lules B presenten empremtes de metilació derivades del desenvolupament normal i canvis adquirits de *novo*. Ambdós tipus de

canvis de metilació de l'ADN varen permetre crear una eina diagnòstica epigenètica molt precisa per 14 subtipus de neoplàsies de cèl·lules B amb diferent abordatge clínic. En consonància amb coneixements previs, vaig identificar que la majoria dels canvis de metilació en l'ADN en el pacients tenien lloc en regions de la cromatina silenciades. Cal destacar que vaig poder relacionar aquest fenomen amb la història proliferativa de les cèl·lules B normals i tumorals, on cada divisió cel·lular semblava que deixava traces epigenètiques en el genoma sense repercussions transcripcionals (Estudi 3). En general, vaig veure que l'activitat mitòtica deixava simultàniament guanys i pèrdues de metilació en l'ADN, però algunes neoplàsies mostraven un biaix cap una direcció o l'altre. Basat en aquestes dades, vaig crear el rellotge mitòtic epigenètic epiCMIT considerant tant guanys com pèrdues de metilació en l'ADN relacionats amb la divisió cel·lular, la qual cosa representa una millora considerable respecte altres rellotges mitòtics proposats prèviament. Cal destacar que la història proliferativa recollida per l'epiCMIT abans del tractament dels pacients va ser altament predictiva del seu futur comportament clínic no només en tumors de cèl·lules B sinó en altres tipus de neoplàsies. Vaig observar que l'acumulació d'alteracions genètiques amb selecció positiva augmentaven l'epiCMIT, però algunes en particular semblaven que conferien una avantatge proliferativa significativa a les cèl·lules de LLC i LCM i distingien pacients amb un comportament clínic molt advers (Estudi 4). Finalment, vaig comparar el rellotge mitòtic epiCMIT amb un altre rellotge epigenètic que identifica de manera molt precisa l'edat cronològica de les persones, l'anomenat rellotge de Horvath. Curiosament, l'epiCMIT estava fortament associat amb una edat accelerada en les neoplàsies de cèl·lula B, suggerint una relació entre l'activitat mitòtica i l'envelliment (Estudi 3).

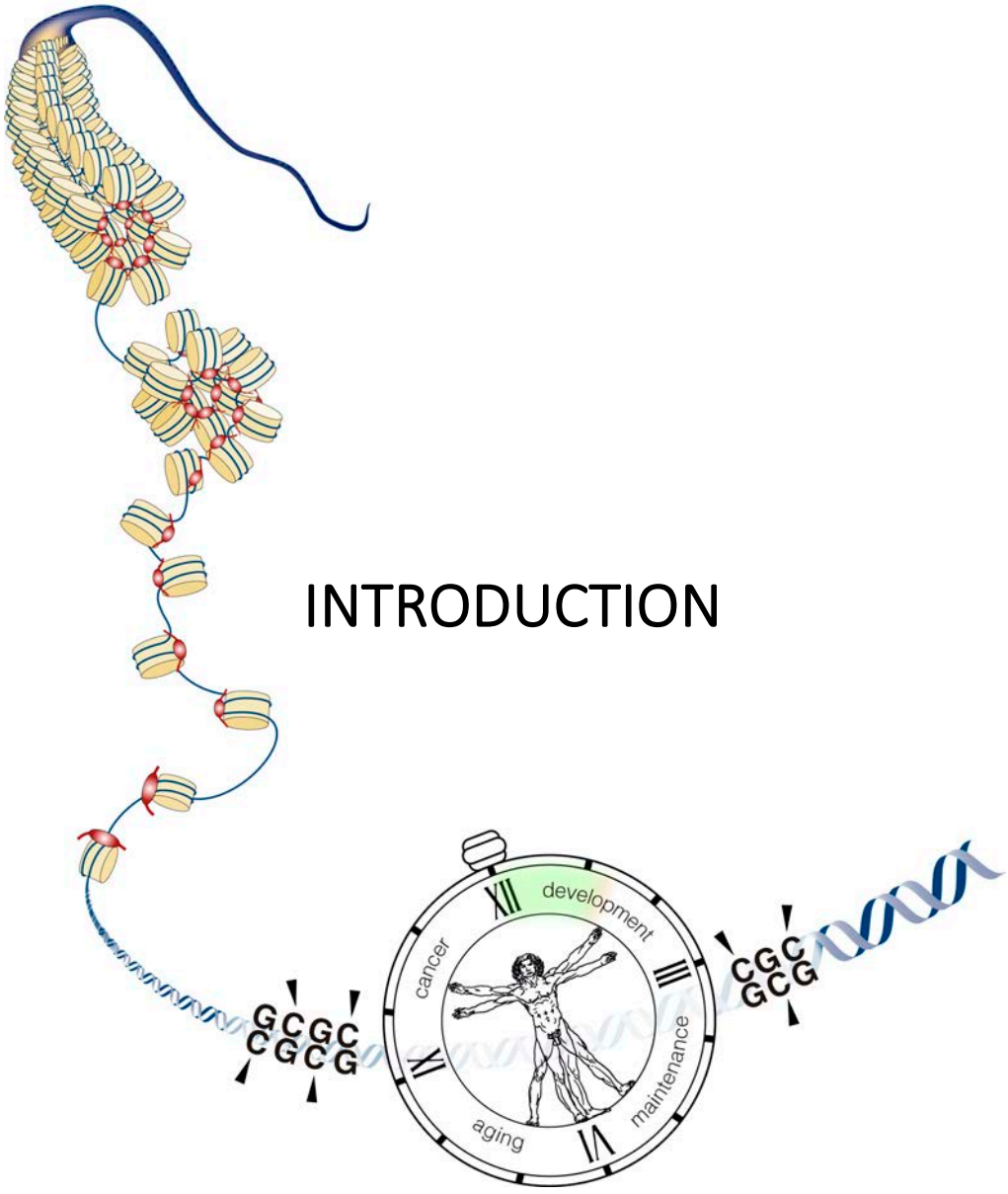
En conclusió, la riquesa de dades presentades en aquesta tesis doctoral revelen la metilació de l'ADN com un traçador holístic del desenvolupament tumoral en les neoplàsies de cèl·lules B, i proporcionen nous coneixements biològics i clínics pels tumors de cèl·lula B i el càncer en general.

TABLE OF CONTENTS

LIST OF SELECTED ABREVIATIONS

AID	Activation-induced cytidine deaminase
ALL	Acute lymphoblastic leukemia
BFB	Breakage fusion bridge
BCR	B-cell receptor
BM	Bone marrow
bmPC	Bone marrow plasma cells
bp	Base pair
CGI	CpG island
CFSE	Carboxyfluorescein succinimidyl ester
CIMP	CpG hypermethylator phenotype
CMP	Common myeloid progenitor
CSR	Class switch recombination
GC	Germinal center B cell
ChIP	Chromatin immunoprecipitation
CLL	Chronic lymphocytic leukemia
cMCL	Conventional MCL
CNA	Copy number alteration
DLBCL	Diffuse large B-cell lymphoma
DLBC-ABC	DLBCL activated B cell type
DLBCL-GCB	DLBCL germinal center B cell type
DMR	Differentially methylated region
DMC	Differentially methylated cytosine
DMV	DNA methylation valleys
DNMT	DNA methyltransferase
epiCMIT	Epigenetically-determined cumulative mitoses
epiTOC	Epigenetic timer of cancer risk
Fig	Figure
HPC	Hematopoietic progenitor cells
HSC	Hematopoietic stem cells
HR	Hazard ratio
ICGC	International Cancer Genome Consortium
i-CLL	Intermediate CLL

IG/Ig	Immunoglobulin
IGHV	Immunoglobulin heavy chain variable region
LMPP	Lymphoid-primed multipotent progenitor
LN	Lymph node
m-CLL	Memory-like CLL
MBC	Memory B cell
MBL	Monoclonal B-cell lymphocytosis
MCL	Mantle cell lymphoma
M-CLL	IGHV mutated CLL
MGUS	Monoclonal gammopathy of undetermined significance
MiAge	Mitotic age mitotic clock
MM	Multiple myeloma
NBC	Naïve B cell
n-CLL	Naive-like CLL
nnMCL	Leukemic non-nodal MCL
OS	Overall survival
PB	Peripheral blood
PCA	Principal component analysis
PMD	Partially methylated domain
preB	Precursor B cells
RAG	Recombination activating gene
RFS	Relapse-free survival
RRBS	Reduced representation bisulfite sequencing
SAM	S-Adenosylmethionine
SBS	Single base substitution
SHM	Somatic hypermutation
SLC	Surrogate light chain
SNP	Single nucleotide polymorphism
TET	Ten-eleven translocation
TF	Transcription factor
TTT	Time to first treatment
U-CLL	IHGV unmutated CLL
WGBS	Whole-genome bisulfite sequencing
WHO	World health organization



INTRODUCTION

1. Epigenetics.

The founder cell of our body is formed when a spermatozoid fertilizes one egg. How this single cell with a unique DNA molecule gives rise to all cell types of our body to finally become a complete organism remains one of the most remarkable phenomena in biology. In fact, proper organismal development implies the generation of a diversity of cell types with distinct but stable gene expression programs. These varying gene expression programs are orchestrated from the same DNA molecule and depend on different developmental and environmental cues, which are interpreted through the epigenetic language. Conrad Hal Waddington (1905-1975) is often given credit for being the first that introduced the term epigenetics in 1942. He defined epigenetics as "the branch of biology which studies the causal interactions between genes and their products, which bring the phenotype into being"[1]. In a broad sense, epigenetics can be defined as a phenomenon that influences the gene expression output without changing the underlying DNA sequence. Thus, epigenetics constitutes a bridge between the genotype and the phenotype, that is, from our genetic identity to our visible traits. Despite the importance of Waddington's work, the concept of epigenetics is constantly evolving thanks to high throughput technologies able to interrogate molecular information with genome-wide coverage [2,3]. So far, four main epigenetic layers at different genomic scales have been thoroughly mapped across the entire genome: i) the 3D structure of the genome, ii) chromatin accessibility, iii) the histone modifications and iv) DNA methylation (**Fig. 1**). The interplay among these epigenetic layers is crucial to give rise to stable and consistent transcriptional programs necessary for all cellular processes.

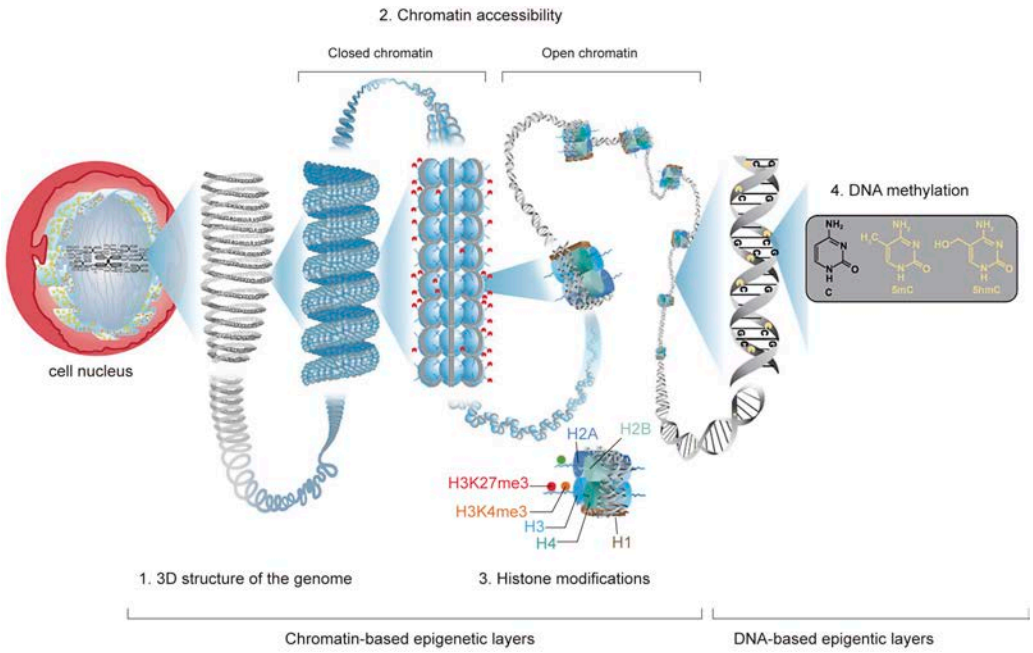


Fig. 1| The main epigenomic layers in human cells. 2 meters of DNA are tightly packaged inside human nucleated cells in the form of 22 autosome chromosomes and 2 sexual chromosomes, including the chrX and chrY forming the so-called chromatin. The DNA is organized thanks to histone proteins, which can be chemically modified at cytosine nucleotides, a process which is known as DNA methylation. The i) 3D structure of the genome, the ii) chromatin accessibility and iii) the histone modifications are commonly referred as chromatin-based epigenetic layers, while the DNA methylation affects the primary DNA sequence. Adapted from Aguilar and Craighead, 2013 [4].

1.1. Chromatin architecture.

The DNA is the genetic material that codifies our identity. Every cell of our body contains approximately 6 billion bps (A, C, G or T) forming the double strand DNA molecule, which would measure 2 meters long if stretched linearly. Considering that we have approximately $3x10^{12}$ nucleated cells [5] and there are about $7.8x10^{9}$ people worldwide, there are approximately $5x10^{22}$ meters of human DNA on Earth! This astronomical number is in fact similar to the estimated diameter of our home galaxy, the Milky Way. Thus, the 2 meters of DNA from each cell has to be extremely packaged to fit into the tiny cell nucleus. The structure of DNA inside the nucleus of human cells is known as chromatin, and this packaging is achieved thanks to a complex cellular machinery that involves histone proteins. In particular, the histones H2A, H2B, H3 and H4 interact together to form a histone octamer (**Fig. 2**) [6,7]. When this histone octamer is formed,

DNA can wrap around it and the nucleosome is assembled. This nucleosome constitutes the primary functional unit of the chromatin, and contain approximately 147 bps of DNA wrapped around the histone octamer [8]. Then, linker DNA joins different nucleosomes creating the so-called "beads on string", a metaphor commonly used to refer to the primary chromatin structure. An additional histone, the H1, is positioned at DNA entry and exit sides and confers the required stability to the nucleosome [9] allowing higher order chromatin structures, which are introduced in the next section.

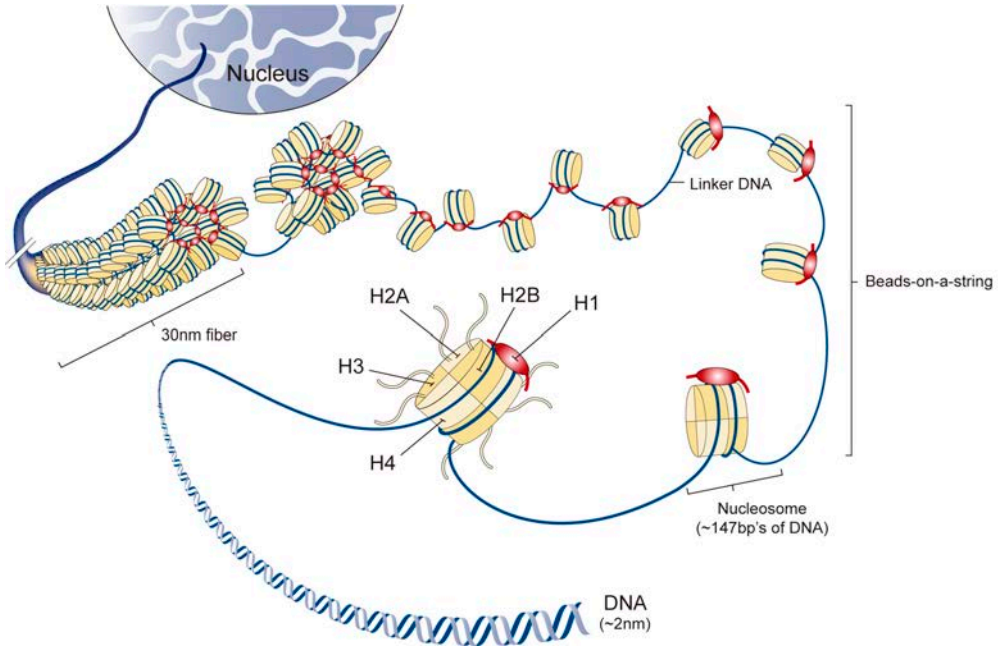


Fig. 2 | The primary structure of the DNA, the chromatin. 147 bp's of DNA are wrapped around the histone octamer -formed by two copies of the histone proteins H2A, H2B, H3 and H3- and constitute the nucleosome. The nucleosomes represent the basic unit of the chromatin structure and are separated by linker DNA, forming a structure which is known as the "beads-on-a-string". The H1 histone protein confers stability to the nucleosome and allows the subsequent packaging of nucleosomes in DNA in fibers, a more complex and packaged DNA structure. Adapted from Fyodorov 2018 [9].

1.2. Chromatin-based epigenetic layers.

1.2.1. 3D structure of the genome.

At the beginning of the century, a seminal work from Dekker and colleagues [10] paved the way for the development of a variety of different techniques to study the 3D structure of the DNA using next generation sequencing approaches [11]. Each of these techniques rely on linking the chromatin of neighboring regions at the 3D level to later create matrices of contact frequencies across genomic regions. Since then, multiple studies have shown that chromosomes are not randomly placed into the nucleus but in chromosome territories, with larger chromosomes found at the periphery (**Fig. 3**) [12–17]. The genomic regions inside chromosome territories that tend to interact more with each other, and represent the actively transcribed fraction of the genome or the so-called A compartment, while the remaining silent fraction is known as B compartment. In addition to this, chromosome territories might be further divided into structural domains known as TADs, which are greatly preserved thanks to some proteins such as CTCF and cohesins [18]. These TADs structures are insulated from each other, which functionally means that regulatory elements, i.e. enhancers, interact in the 3D space with their target genes within a TAD (**Fig. 3**). For these interactions to take place as well as the binding of regulatory molecules such as TFs, chromatin needs to be accessible.

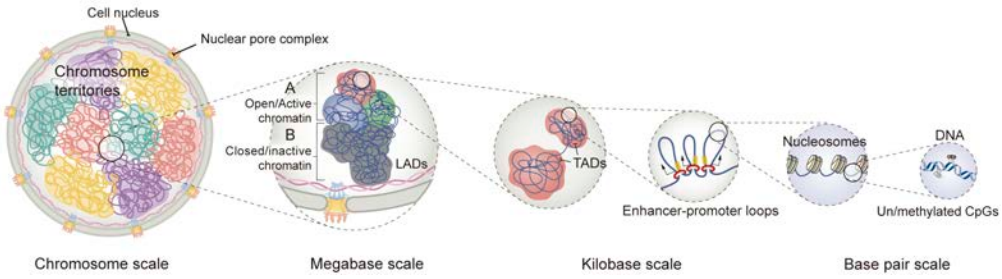


Fig. 3 | The 3D organization of the genome. The DNA is not randomly packaged in the human cell nucleus but in chromosome territories. Genomic regions with an open chromatin and actively transcribed genes are referred as the A compartment, while the remaining closed regions are inactive and referred as B compartment. Small substructures know as TADs are believed to represent the structures from which enhancers can interact with their corresponding genes to control their expression. Adapted from Uhler 2017 [19].

1.2.2. Chromatin accessibility.

The term chromatin accessibility refers to the extent to which nuclear complexes can physically contact chromatinized DNA. Thus, it can be understood as the fraction of the genome that represents the network of permissible physical interactions through which regulatory elements bind the DNA and cooperatively regulate gene expression [20]. This network therefore may reflect the regulatory capacity rather than a static biophysical state of the chromatin. The most used genome-wide approaches to measure the accessibility of the chromatin are DNase I enzymatic reaction [21,22], ATAC-seq [23], NOMe-seq [24,25] and MNase-seq [26,27]. Open chromatin has been found to be enriched in cell type-specific TFs binding sites [28–30], and accessible enhancers and promoters are necessary for gene transcription [31]. However, poised enhancers and promoters related to some silent genes are often open, suggesting that chromatin accessibility is needed but not sufficient for enhancer or promoter activity [30,32,33]. Thus, chromatin accessibility sets the stage for gene expression regulation, and other marks such as the presence of certain histone modifications are more informative about the regulatory and transcriptional states of a particular genomic region.

1.2.3. Histone modifications

As previously mentioned, histone proteins H2A, H2B, H3, H4 and H1 are essential for packing the DNA into nucleosomes, the structural functional unit of the chromatin. Nonetheless, far from being simple scaffolds for DNA packaging, histone proteins are dynamic structures that influence all DNA-based processes such as chromatin compaction, nucleosome dynamics or/and transcription [34]. To achieve this wide spectrum of functions, histone proteins can acquire chemical modifications both at their tails [35] and at their core regions [36] (**Fig. 4**), although the former are the most studied. These chemical modifications include methylation, acetylation, phosphorylation, ubiquitylation, sumoylation, ADP ribosylation, deamination, proline isomerization, β-N-acetylglucosamine at different histone residues, and the list keeps growing [37]. To measure all these chemical modifications, ChIP-Seq have become the gold-standard genome-wide approach [38–41]. This technology relies on three basic steps: i) the formation of cross-links between proteins and DNA, ii) the selective co-immunoprecipitation with an antibody of protein-bound DNA fragments and iii) the recovered DNA is sequenced and mapped to the genome to determine the location sequences bound by the protein [42]. Histone modifications have been shown to be highly cell-type specific and to be

associated with different cellular functions [43]. For instance, the H3K27ac marks active regulatory regions, which represent enhancers if in addition H3K4me1 is present, or promoters if instead, H3K4me3 is present. Other marks are related to silencing, like H3K27me3 or H3K9me3, while others to transcriptional elongation such as H3K36me3. To better characterize the functional output of this histone code, pioneer computational approaches integrated several histone modifications to define chromatin functions called "chromatin states" [44,45] (**Fig. 4**). This approach allows the segmentation of the genome into functional states, such as active promoters, or active enhancer, and facilitates the genome-wide read-out of histone functions and the integration with other epigenetic layers such as DNA methylation.

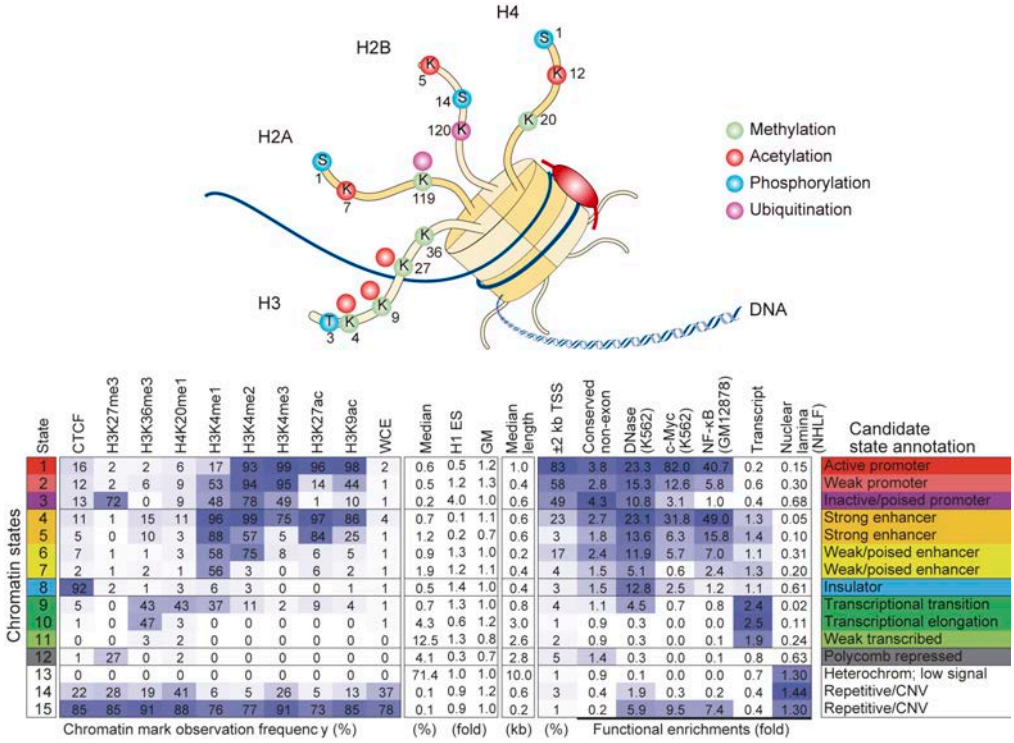


Fig. 4 | The histone code. Histone proteins can be chemically modified at their core structures and at their tails. Depending on the chemical modification as well as the amino acid residue which is modified, the histone proteins regulate different cellular processes. The combination of different histone proteins allows the segmentation of the genome into chromatin states, which are informative about the underlying genomic function [44].

1.3. DNA-based epigenetic layers: DNA methylation.

1.3.1. Introduction.

In the previous **Introduction section 1.2**, epigenetic layers related to the chromatin structure have been introduced. However, epigenetic control is also present at the DNA level and involves the covalent modification of nucleotides. So far, these modifications have been found in three out of four DNA bases, namely thymine, adenine and cytosine [46–48]. Nevertheless, the most prevalent and well-studied DNA modification in the human genome is cytosine methylation, which is the main focus of this doctoral thesis. Cytosine methylation is commonly referred to as DNA methylation, and this will be the term used throughout this thesis. DNA methylation takes place at the fifth carbon of a cytosine and at different nucleotide contexts, mostly in a CG context but also in a non-CG context including CHG or CHH (where H can be an A, C or a T) [49]. Methylation occurring at non-CG contexts has been observed mainly in embryonic stem-cells and mammalian brain tissue, with negligible levels in more differentiated cells [49–52]. A detailed summary of the molecular mechanisms underlying DNA methylation as well as its impact on physiological and pathological conditions is presented in the following sections.

1.3.2. DNA methylation acquisition and maintenance.

DNA methylation homeostasis involves multiple molecular complexes including enzymes and metabolites that together establish the proper dynamics of DNA methylation, including addition, maintenance, and removal (**Fig. 5**). The addition of a methyl group to the fifth carbon of the cytosine is carried out by the DNA methyltransferase (*DNMT*) family of enzymes [53], which use a methyl group donated from the S-adenosyl methionine (SAM) metabolite. So far, five *DNMTs* have been identified in the human genome: *DNMT1*, *DNMT2* (tRNA methylase), *DNMT3A*, *DNMT3B* and *DNMT3L*. *DNMT1* is primarily but not exclusively responsible for methylation maintenance. During DNA replication, newly synthesized DNA lacks methylation marks, and thus hemi-methylated CpGs are abundant at replication forks. These hemi-methylated CpGs are recognized by UHRF1 that helps to recruit DNMT1, which then catalyze the addition of a methyl group to the corresponding cytosines ensuring to pass the epigenetic inheritance to the newly synthesized DNA strand [54–56]. On the other hand, *DNMT3A* and *DNMT3B* have been attributed to *de novo* DNA methylation [57], and may need to form a structure with the catalytically inactive DNMT3L to mediate their methyl-transferase activity at corresponding CpG dinucleotides [58].

1.3.3. Mechanisms of DNA demethylation.

Contrary to the well-described methylation process, the demethylation pathway has remained more elusive. Nowadays, it is known that DNA demethylation process can take place either actively or passively.

1.3.3.1. Active DNA demethylation.

The main active pathway of DNA demethylation involves iterative oxidations of the methylated cytosine by the ten-eleven translocation (TET) family of enzymes, including *TET1*, *TET2* and *TET3* that give rise subsequently to 5-hydroxymethylcytosine (5hmC), 5-formilcytosine (5fC) and 5-carboxycytosine (5caC) [59,60]. Afterwards, these oxidized forms can be efficiently removed by thymine DNA glycosilase (TDG), a member of the base excision repair (BER) glycosylases [61]. This creates an abasic side (apurinic/apyrimidinic site) which is subsequently replaced by a new unmethylated cytosine. Alternatively, DNA demethylation can be caused by the activity of *AID*/*APOBEC* enzyme families. Although *AID*/*APOBEC* have a well-known role in deaminating cytosines and trigger somatic hypermutation in B cells (please be referred to **Introduction section 2.2** for further details), they also have been shown to have affinity for methylated cytosines [62]. The activity of these enzymes create oxidized derivatives of methylcytosines, including 5-hydroxymethyluracil (5hmU) and 5-formyluracil (5fU) [61,63], which can then be effectively removed by *SMUG1*, *TDG*, *MBD4* enzymes and finally replaced by an unmethylated cytosine via the BER pathway. Worthwhile mentioning, 5hmC, 5fC and 5caC were initially thought as intermediates of DNA demethylation pathway. However, there are some evidences that these oxidized forms themselves could have some biological functions, since they have been shown to be stably present in mammalian genomes [64,65]. Furthermore, specific binding proteins for 5hmC, 5fC and 5caC have been identified [66–68], suggesting that these cytosine modifications could have some important roles, including chromatin remodeling, transcriptional regulation or even increasing the flexibility of DNA [69].

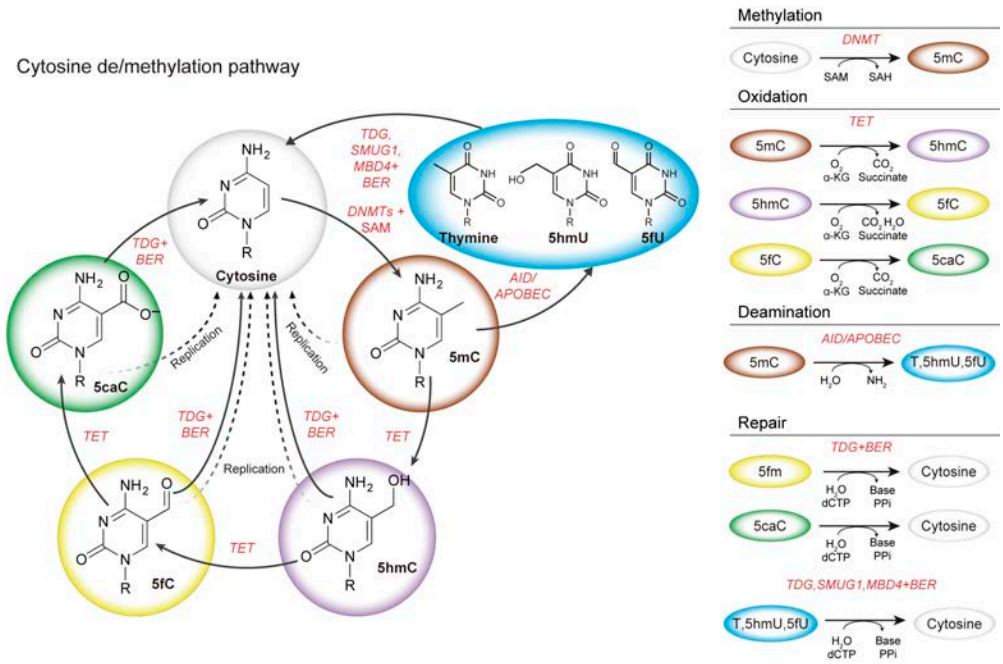


Fig. 5 | The DNA methylation pathway. The process of cytosine or DNA methylation is carried out by DNA methyl-transferases enzymes *DNMT1*, *DNMT3A* and *DNMT3B*. This process requires the methyl group from the S-adenosyl methionine (SAM) metabolite. Conversely, the process of DNA demethylation can take place actively of passively, with TET proteins or replication-dependent dilution, respectively. Adapted from reference [59].

1.3.3.2. Passive DNA demethylation.

In addition to the active pathway, DNA methylation can also be lost passively. Methylated cytosine and all their derivatives can be diluted passively in a replication-dependent manner. During DNA replication, unmodified cytosines are incorporated into the newly synthesized strand, creating hemi-modified 5mC:C, 5hmC:C, 5fC:C and 5caC:C dyads. These dyads have poor affinity for DNMT1 and UHRF1 [70–72], thus resulting in a dilution of DNA methylation during replication. Furthermore, these enzymes as well as their substrates may be limited under some circumstances such as late-replicating regions, further promoting the loss of DNA methylation [73].

1.3.4. DNA methylation in normal physiology.

1.3.4.1. Historical view of DNA methylation functions.

DNA methylation was first discovered in bacteria as early as 1898 [74,75] and then in eukaryotes [76]. Afterwards, seminal works from Holliday and Pugh as well as Riggs

proposed for the first time that DNA methylation at CpG dinucleotide context may mediate gene silencing [77,78]. Few years later, multiple studies added new functions to DNA methylation, such as genomic imprinting [79] or cell development and differentiation [80–82]. All these pioneer studies focused on CpG islands (CGIs) at TSS, and contributed to shape the general perception that the main DNA methylation function was gene silencing [83].

1.3.4.2. New concepts of DNA methylation functions.

In spite of the importance of these initial studies focused on promoter regions, genome-wide approaches developed over the last decade are certainly changing our perception of the functions of DNA methylation [43,47–49,84–86]. In fact, it seems that DNA methylation has actually a genomic and chromatin-dependent function (**Fig. 6**). For instance, DNA methylation at promoters is tightly associated with gene repression, but DNA methylation at gene bodies is needed for gene expression [87], prevents spurious transcription initiation [88] and regulates alternative splicing events [89] in part thanks to MeCP2 protein [90]. Beyond coding genes, DNA methylation has been reported to play a crucial role in silencing transposable elements, maintaining the overall genomic stability [91]. Furthermore, CpG poor regions distant to genes function as enhancers for gene expression and are lowly methylated and often bound by TFs [92]. However, the methylation status is not distinctive for active gene expression and may represent an imprint of past union of DNA-binding factors and gene activation [93] (**Fig. 6**). Notably, DNA methylation can block DNA-binding factors including CTCF, causing loss of insulation between topological domains and aberrant gene activation [94]. In line with this, it is thought that DNA methylation landscapes are shaped in part by TFs binding, causing demethylation to their targeted regions [95]. Nonetheless, this may not be the general rule, since it has recently been shown that there are some TFs that actually prefer methylated DNA for their binding [96]. Finally, other kilo/megabase-scale DNA methylation landscapes have been reported. These include DNA methylation canyons [97]/DNA methylation valleys (DMVs) [98], which are conserved and insulated self-interacting hypomethylation domains across vertebrates maintained by polycomb [99]; and Partially Methylated Domains (PMDs) [49]. Such PMDs have been found in differentiated somatic tissues [49,100] and cancers [101], and represent heterochromatic and transcriptionally inactive regions.

1.3.4.3. DNA methylation in normal organismal and cellular development.

Given the variety of functions presented above, it is not surprising that alterations DNA methylation patterns causes profound developmental and organismal defects. In fact, the DNA methylation landscape is widely reprogrammed early after fertilization until the post-implantation embryo, in which DNA methylome forms the basis for the entire body plan [102–106]. This reprogramming involves the erasure of DNA methylation in the paternal and maternal DNA, and the posterior gain of DNA methylation in the post-implantation embryo. Then, each cell starts to acquire its own specific DNA methylation signature, which is associated with different transcriptional outputs necessary for proper development [84,107,108]. Not surprisingly, knocking-out *DNMTs* widely alter DNA methylation patterns and cause embryonic (*DNMT1* and *DNMT3A*) [109] or postnatal lethality (*DNMT3B*) [109] in mice. *DNMT1* has been shown to be essential for progenitor cells with self-renewing capacity, including human embryonic stem cells [110], hematopoietic stem cells [111] and somatic tissues [112]. Additionally, murine hematopoietic stem cells with reduced *DNMT1* activity cannot suppress key myeloerythroid regulators and thus can differentiate into myeloerythroid but not lymphoid progeny. In addition to this, *DNMT3A* and *DNMT3B* loss has been reported to impair mouse embryonic stem cell differentiation [113], promote hematopoietic stem cell expansion with impaired differentiation [114,115] and immortalization *in vivo* [116]. Finally, induced pervasive loss of DNA methylation in adult mice causes a broad epigenetic deregulation and apoptosis via *TP53* [117]. Collectively, all these studies highlight the importance of maintaining the integrity and dynamics of the DNA methylome in cell differentiation and organismal development.

1.3.5. DNA methylation in cancer.

In the previous section, compelling evidences regarding the fundamental role of DNA methylation in normal physiology have been presented. Not surprisingly thus, DNA methylation alterations have been widely reported in human disease, most notably in cancer [86,118].

1.3.5.1. Historical view.

The initial identification of DNA methylation aberrations in cancer was independently discovered by two laboratories in 1983. The Ehrlich's laboratory detected a global genomic hypomethylation in various cancers versus a wide variety of normal tissues [119],

and Feinberg and Vogelstein found hypomethylation of some genes in colon adenocarcinomas versus normal colonic epithelium, [120]. Both studies showed that hypomethylation was more pronounced in metastatic cancer, already suggesting that DNA methylation is dynamically modulated during cancer progression. In addition to DNA methylation loss, aberrant hypermethylation in cancer was also observed for the first time in 1989 in retinoblastoma patients, and the DNA methylation of gene promoter as silencing event was then proposed [121,122]. Indeed, this was subsequently demonstrated to represent an oncogenic mechanism for multiple other cancer types [123,124].

1.3.5.2. New concepts of DNA methylation in cancer.

The first studies on DNA methylation used enzymatic approaches, and the initial findings have been broadly confirmed by more precise methods based on the sodium bisulfite reaction (please, be referred to **Introduction, section 1.4** for further details), which now include genome-wide approaches based on next generation sequencing and arrays [2,47,84,86,107,125–127]. A variety of studies published in the last decade have uncovered that cancer DNA methylomes are characterized by focal hypermethylation embedded in the context of a global hypomethylation. This global hypomethylation has been associated with genomic instability, increased activity of transposable elements, and demonstrated to promote tumorigenesis in mice [91,128–130] (**Fig. 6, bottom**). In addition to this, WGBS has allowed the detection of hypomethylation at low CpG-content regions showing activation marks (H3K27ac) that may contribute to malignant transformation and progression by altering the expression of downstream genes [131]. Likewise, additional studies showed that aberrant DNA hypermethylation also favors malignant transformation and progression by at least two means: by hypermethylation of promoters and subsequent silencing of tumor suppressor genes [123]; and by hypermethylation at TF binding sites, which avoids their union and can potentially cause an aberrant reconfiguration of the 3D topological organization, causing gene expression outcomes which favor malignant transformation [94,132]. Finally, several studies integrating DNA hypermethylation with histone marks have uncovered that H3K27me3 or bivalent (H3K27me3, H3K4me3) regions tend to gain methylation in cancer [133–135], although the underlying mechanisms remain to be fully elucidated.

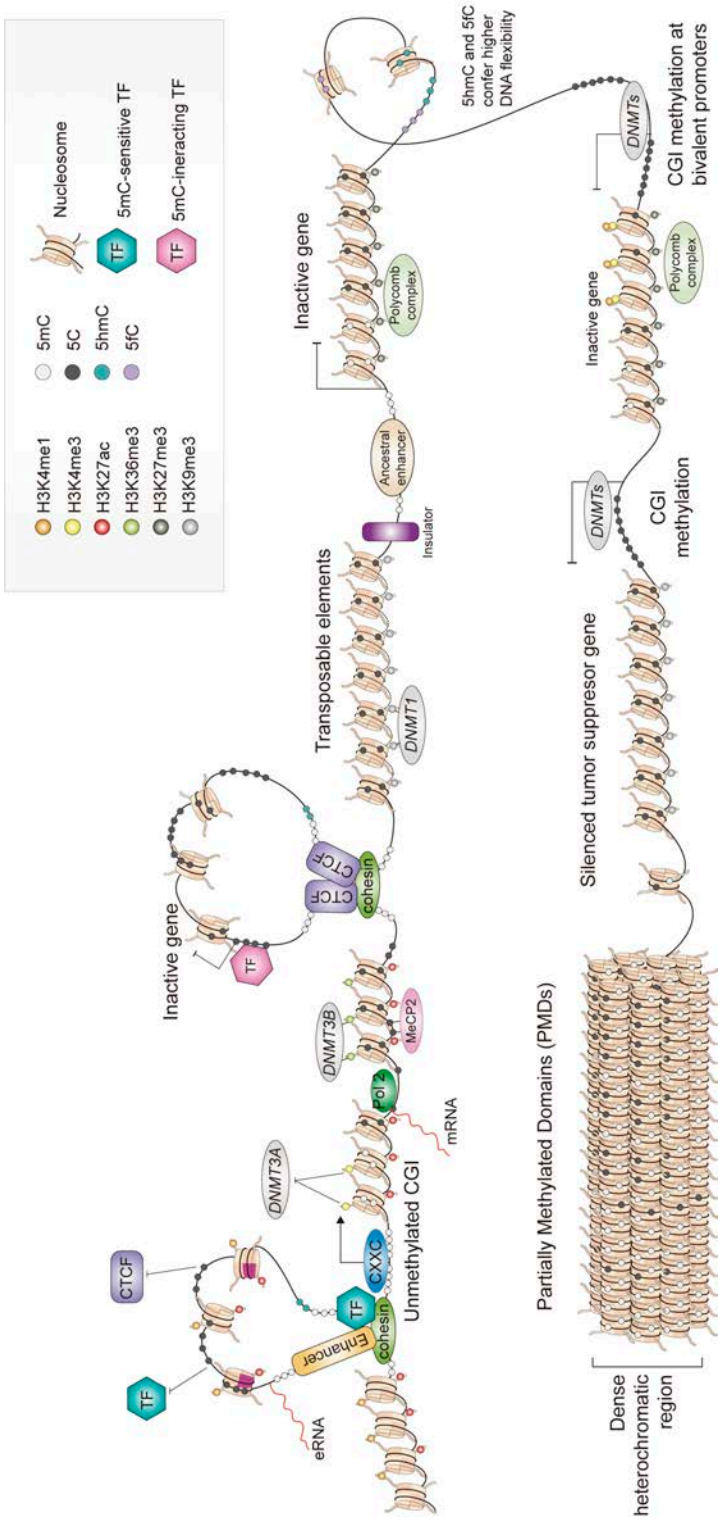


Fig. 6 | The DNA methylation functions in different genomic regions. DNA methylation has a genomic and chromatin-dependent functions. DNA methylation at promoters is associated with gene silencing, a relatively frequent alteration found in many cancer types. Conversely, it seems that the DNA methylation at gene body is necessary for transcription. Nonetheless, DNA methylation is in general poorly correlated with gene expression, and other epigenetic marks such as histone modifications are far more informative about gene activity. Beyond the coding genome fraction, other methylation landscapes have been described, such as PMD, which are present in differentiated and cancerous cells.

1.3.5.3. Interplay between DNA methylation and genomic aberrations.

Genomic and epigenomic aberrations undeniably contribute to human disease, with a prominent involvement in cancer [118,136]. In spite of this fact, the relationship between genome-wide genetic and epigenetic alterations driving malignant transformation and progression remains poorly characterized [137]. The study of this relationship requires simultaneous genome-wide genetic and epigenetic profiles of large patient cohorts, which until recently remained challenging to obtain due to limited technology and elevated costs. Early studies profiling the DNA methylome of tumors with specific mutations revealed the presence of the so-called CpG hypermethylator phenotype or CIMP, characterized by the hypermethylation of CpG-rich promoters [138]. This phenotype is particularly associated with the V600E mutation of the *BRAF* oncogene in colorectal cancer [139,140], whereas in glioma and glioblastoma multiforme it is tightly associated with mutations of the *IDH1* gene [141,142]. In fact, a subsequent study suggested a causative role of *IDH1* mutation for the CIMP phenotype in glioma [143]. In addition to *IDH1*, *IDH2* has been found mutated also in acute myeloid leukemia (AML) leading to a CIMP phenotype [144]. *IDH1/2* mutations produce an oncometabolite which competes with the natural substrate of *TET* enzymes and impairs normal DNA demethylation dynamics (**Fig. 5**) and hematopoietic differentiation. Furthermore, *DNMT3A* mutations seems to accumulate in HSC during aging [145–147] and may prelude acute myeloid leukemia (AML) [148,149], which show focal hypomethylation loses at CpG islands which would be methylated otherwise [150]. Finally, a study from Oakes and colleagues in CLL showed that DNA methylation changes might coevolve with certain genetic alterations during CLL progression and are associated with inferior patent outcome. Despite these perhaps anecdotal and mainly focal associations, the relationship between genetic and epigenetic alterations remains challenging to decipher.

1.4. DNA methylation measurement techniques.

Since the first cancer epigenetic studies in the 80's, multiple technologies with different genomic resolutions have been developed to detect and quantify cytosine methylation [46,85,151–153]. The selection of a particular technological approach for a given study will depend on the number of samples to be analyzed, the genome coverage, and the economic cost (**Fig. 7**).

1.4.1. General aspects.

The main methodologies to detect DNA methylation can be grouped into three main categories depending on the genomic DNA treatment, including i) enzymatic digestion, ii) affinity enrichment and iii) sodium bisulphite treatment (**Fig. 7**):

i) The enzymatic digestion of genomic DNA depending on the methylation status of the target sequence is based on the use of methylation-sensitive restriction enzymes such as HpaII and SmaI or MspI. The resulting DNA fragments are enriched in either methylated or unmethylated DNA associated with the target sequences recognized by the enzymes. Several methods have been developed using this strategy, such as the differential methylation hybridization (DMH) methodology and the HpaII tiny fragment enrichment by ligation-mediated PCR assay (HELP), among others [154–156].

ii) Methodologies performing affinity enrichments rely on the affinity of antibodies such as MECP2 against 5mC. Some of the caveats related to methylation enrichment methods are typically the sequence-dependent specificity of the selecting agents or the interpretation of the data, which is interpreted as the relative abundance of 5mC in genomic windows instead of quantifying the methylation status at single-cytosine resolution.

iii) The sodium bisulphite treatment of DNA is probably the most widely used approach. It induces the deamination of unmethylated cytosine resulting in the conversion to uracil, while leaving methylated cytosines unmodified. Thus, this creates a change in the DNA sequence that can be subsequently used to map the methylation status of each cytosine in the genome. Afterwards, the resulting processed DNA can be subjected to a variety of methodologies to interrogate the methylation status of DNA, including locus-specific and genome-wide approaches, such as Illumina Infinum arrays and next-generation sequencing approaches (**Fig. 7**). The selection of these technologies thus will depend on the scientific question to address as well as the budget available. In the course of this doctoral thesis, DNA methylation profiles have been obtained through Infinium DNA methylation arrays and the next generation sequencing approach WGBS, and therefore will be introduced in further detail below.

Pretreatment	Analytical step			
	Locus-specific	Gel-based	Array-based	NGS-based
Enzyme digestion	• HpaII-PCR	• Southern blot • RLGS • MS-AP-PCR • AIMS	• DMH • MCAM • HELP • MethylScope • CHARM • MMASS	• Methyl–seq • MCA–seq • HELP–seq • MSCC
Affinity enrichment	• MeDIP-PCR		• MeDIP • mDIP • mCIP • MIRA	• MeDIP–seq • MIRA–seq
Sodium bisulphite	• MethyLight • EpiTYPER • Pyrosequencing	• Sanger BS • MSP • MS-SNuPE • COBRA	• BiMP • GoldenGate • Infinium	• RRBS • BC–seq • BSPP • WGSBS

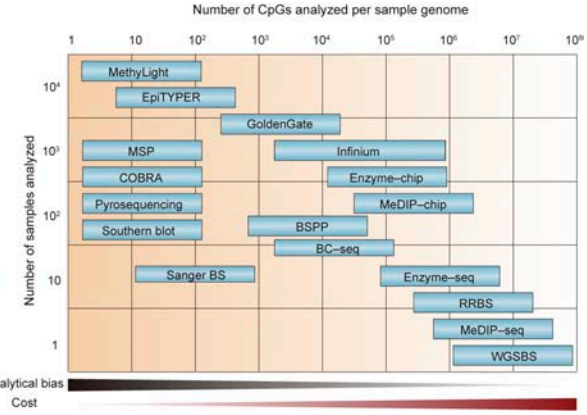


Fig. 7 | The DNA methylation measurement techniques. The main methodologies to detect DNA methylation can be grouped into three main categories depending on the genomic DNA treatment, including i) enzymatic digestion, ii) affinity enrichment and iii) sodium bisulphite treatment. Each respective category contains several analytical steps which can be grouped broadly into gel-based, array-based and NGS-based approaches, with each of them having different genome coverages and biases. Adapted from Laird 2010 [152].

1.4.2. Infinium DNA methylation arrays.

Over the last decade, Illumina arrays have become the gold standard methodology to measure DNA methylation thanks to their cost-effective DNA methylation measure of multiple CpGs, including virtually all protein coding genes and also CpGs present at regulatory and heterochromatic regions. Although initial array designs were severely biased towards promoters [157], the Infinium HumanMethylation450k BeadChip and the more recent Infinium HumanMethylationEPIC BeadChip (Illumina, San Diego, USA) allows the interrogation of DNA methylation across the whole genome. They have been shown to represent an efficient, robust, reproducible and affordable platform that allows the quantification of DNA methylation levels of individual CpG sites in regions spanning the entire genome with only 500ng of genomic DNA. Specifically, the Infinium HumanMethylation450k array measures the methylation status of 485,512 CpG in different genomic locations, including intergenic regions, 96% of all known CGIs, 5' regions, gene bodies and 3' prime regions, covering nearly all RefSeq genes and also ncRNAs and microRNAs. The newly incorporated InfiniumMethylationEPIC BeadChip builds on the Infinium HumanMethylation450 BeadChip with > 90% of the original CpGs plus an additional 350,000 CpGs in enhancer regions. These two arrays show excellent correlations between their methylation values. They are based on bisulfite converted

DNA, and employ two different chemistries, the Infinium I and II (**Fig. 8**). The Infinium I contain two probes per CpG, one designed for the unmethylated and the other for the methylated state, while the Infinium II requires only one probe and employs an extension step with differentially marked bases to measure DNA methylation [158]. The bisulphite treated DNA is hybridized to the array according to manufacturer's instructions and subsequently scanned with the iScan Illumina system. Afterwards, raw IDAT files are created and can be analyzed bioinformatically with some commercial software such as GenomeStudio, or with custom pipelines in the R statistical language with devoted packages such as *minfi* [159].

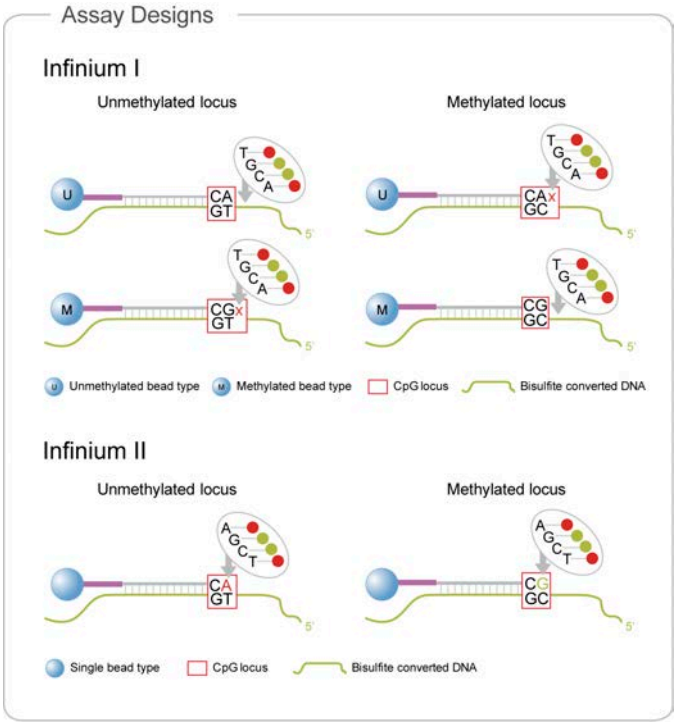


Fig. 8 | Infinium chemistry used in Illumina BeadChips 450k and EPIC arrays. The Infinium I contain two probes per CpG to measure the methylation status of one CpG, one designed for the unmethylated and the other for the methylated state. In contrast, the Infinium II requires only one probe and employs an extension step with differentially marked bases to measure DNA methylation. Adapted from https://www.illumina.com/.

1.4.3. Whole-genome bisulphite sequencing (WGBS).

Although Infinium arrays are widely used to measure DNA methylation in large patient cohorts, the most unbiased and comprehensive approach is still WGBS. WGBS allows the interrogation of the methylation status at every single cytosine over the whole genome (with the sequencing limitations typically associated to next-generation sequencing approaches). Although the price since the first human WGBS has significantly dropped [49], it is still an expensive technique that allows the analysis of a reduced number of samples (**Fig. 7**). Typically, the amount of required input DNA needed is less than 1 microgram, although high sequencing depth (30-50X) is needed for reaching good coverage for trusty methylation calling. It is worth mentioning that WGBS allows the study of DNA methylation at a non-CpG context, which has been observed in undifferentiated cell types and neurons [43,50–52,100]. Additionally, it permits to study CpG poor regions, that may have an important role for gene expression [92,160]. An important aspect of WGBS is that the analyses of the resulting FASTQ files from sequencing devices are computationally intensive, and are usually performed at sequencing facilities with specialized computational servers. Both the computational cost and the price per sample represent a limitation for a wider implementation in epigenetic studies. An alternative to WGBS is the reduced representation bisulfite sequencing (RRBS) approach [161–163], which sequences only a fraction of the genome and therefore it can be applied to larger series of patients due to its more affordable economic cost (**Fig. 7**).

2. Normal B-cell differentiation.

B-cell neoplasms are clonal tumors of immature and mature B cells at various stages of B-cell differentiation [164]. Therefore, although normal B-cell differentiation is not a major topic of this doctoral thesis, it is necessary to introduce it, as it is the basis to classify B-cell tumors and interpret their epigenetic landscapes.

2.1. Hematopoiesis.

Hematopoiesis is defined as the process that gives rise to all the cellular components of the hematopoietic system. This process starts in the bone marrow, which is the home of hematopoietic stem cells (HSC), which represent the pool of cells with the potential to divide and differentiate into all hematopoietic cellular types. These cells include red blood cells, megakaryocytes, basophils, eosinophils, neutrophils, mast cells, monocytes, macrophages, NK cells, T cells and also B cells (**Fig. 9**). Altogether, these cells constitute

the immune system and vary in their percentages and numbers among and within individuals upon different factors such as age or environmental stimuli [165,166]. Cells from the immune system act in coordination to defend our organism from external agents, invading pathogens and cancer cells, among other threats [165].

2.1.1 The HSC niche.

The HSC niche is a highly complex ecosystem that sustains HSC function by promoting survival and long-term maintenance of the HSC pool [167,168]. It is thought that the number of human HSCs increase from birth to adolescence, where it reaches a plateau that is made up of few to hundred thousand cells [169,170]. HSCs possibly divide at a maximum rate of 1-3 divisions per year. The total number of estimated mature blood cells that a human produces per year is about $1.4x10^{14}$, and thus the HSC pool must include a strong transient-amplifying compartment. In this context, the HSC niche is a dynamic system that changes in a time and space to adapt to possible changes of demand upon developmental and environmental cues. The differentiation of HSCs to the different immune cell types depends greatly on the coordinated action of key developmental TFs that ultimately give rise to lineage-specific gene expression programs [171–176] (**Fig. 9**). A fraction of the HSC pool undergoes a commitment into multipotent progenitors (MPP), which subsequently give rise to the main branches of the hematopoietic system: the common myeloid precursor and lymphoid-primed multipotent precursor (CMP and LMPP, respectively). Then, CLP and LMPP can subsequently give rise to more differentiated precursors with gradually decreased differentiation potential, including the megakaryocyte precursors (MEP), the eosinophil-basophil progenitor (EoBP), the granulocyte-monocyte progenitors (GMP) and the common lymphoid precursor (CLP). Afterwards, these committed precursors can give rise to the different mature cells of the immune system such as B cells.

Fig. 9 | The hematopoiesis process (see figure on next page). The HSC pool is thought to be formed by distinct types of stem cell subpopulations with different self-renewal and differentiation properties, which collectively ensure blood production upon different developmental end environmental cues throughout the whole human lifespan. Cell differentiation is achieved thanks to the coordinated action of key transcription factors that drive the differentiation process from myeloid (CMP) and lymphoid progenitors (LMPP) to the terminally differentiated blood cell types. MEP, megakaryocyte precursors; RBC, Red blood cells; EoBP, eosinophil-basophil progenitor; (GMP); granulocyte-monocyte progenitors (GMP). Adapted from Sankaran and Weiss, 2015 [177].

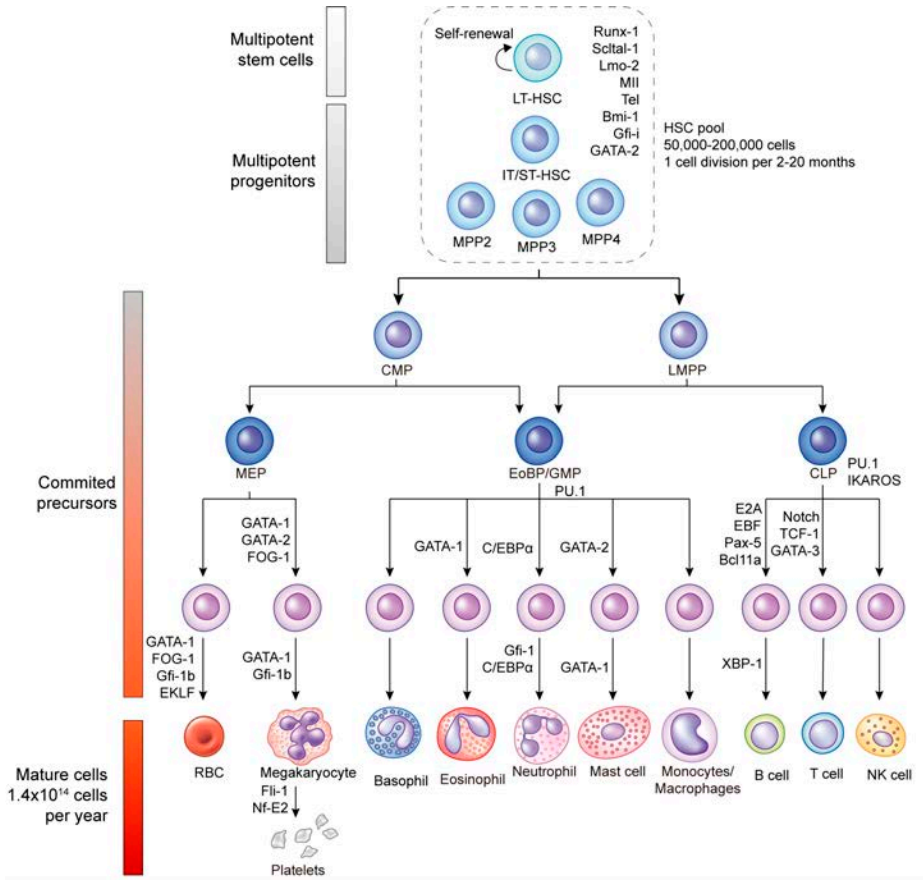


Fig. 9 | See caption on previous page.

2.2. B-cell differentiation stages and functions.

The generation of B cells, also known as B-cell lymphopoiesis, is perhaps one of the most complex processes in the human body, and it has been a subject of intense study since decades [178] (**Fig. 10**). One of the most important players in B-cell differentiation is the immunoglobulin gene that ultimately will constitute the B cell receptor (BCR), a key component for the survival and differentiation of B cells as well as their biological functions. This differentiation process starts in the bone marrow, where the HSC differentiates into CLP, which subsequently gives rise to the pro-B cell, the first primed cell to the B-cell lineage. Pro-B cells starts a key and B cell-specific molecular process known as the immunoglobulin heavy chain (IGH) recombination, in which the immunoglobulin gene is rearranged thanks to the action of the recombination-activating genes (RAG). Productive heavy (H)-chain gene assembly leads to the association of IgM H-chain (μ chain) with surrogate light chain (SLC) components and the subsequent

expression of the pre-B cell receptor (**Fig. 10a, b**). Afterwards, SLC components are downregulated and RAG-mediated rearrangement of the light (L) chain locus is produced, giving rise to pre-B cell with a complete BCR. L-chains that pair with H-chains trigger tonic BCR signaling, which promotes positive selection or receptor editing when the BCR is or not autoreactive, ensuring central tolerance (i.e, non-reactive B cells against autoantigens) [179] (**Fig. 10a**). When the BCR is correctly formed, the B cell is considered an immature B cell, which subsequently leaves the bone marrow to populate secondary lymphoid organs such as the spleen or the lymph nodes thanks to the homing properties of chemokines such as CXCR4 [180]. At these anatomical sites, B cells encounter a cognate antigen and become activated [181,182]. The encounter and activation of B cells is mainly mediated by specialized cell types of the lymph node, such as $CD4^+$ T cells, follicular helper T cells, and follicular dendritic cells (**Fig. 10c**). After their activation, B cells can undergo class switch recombination (CSR), a process by which B cells can change their isotype into another class (from IgM and IgD to IgG, IgE or IgA) (**Fig. 10b**) and differentiate into short-lived antibody secreting plasmablasts or into germinal-center-independent memory B cells (MBC) [183] (**Fig. 10c**). CSR allows the acquisition of different effector functions without changing the antibody specificity associated with the rearranged variable region. This process has been classically linked to the germinal center reaction but recent data suggests that CSR occurs mostly outside the germinal center [184]. Alternatively, activated B cells can differentiate into germinal center (GC) B cells, proliferate and form the GC compartment, which contains the dark zone (DZ) and the light zone (LZ). GC B cells re-circulate between DZ and LZ thanks to the coordinated expression of different genes, such as BCL-6, MYC and REL, or IRF4. In the dark zone, B cells undergo extensive proliferation and somatic hypermutation (SHM) (**Fig. 10b, c**) that leads to the mutation of the rearranged variable region, a process that allows to increase the affinity for the cognate antigen. In the light zone, B cells that carry a BCR with an increased affinity for the cognate antigen are selected, and those with auto-reactive BCR may be eliminated. After affinity maturation, B cells can subsequently differentiate into MBC and antibody secreting plasma cells (PC) that eventually will home to the peripheral blood and bone marrow, respectively, allowing long-term immunity [185] (**Fig. 10d**). These two cell types will ensure a rapid and efficient immune response against a future possible encounter with their cognate antigen. In addition to the secretion of high affinity antibodies with varying functions, B cells have other biological functions including cytokine production with varying functions and antigen presentation [186].

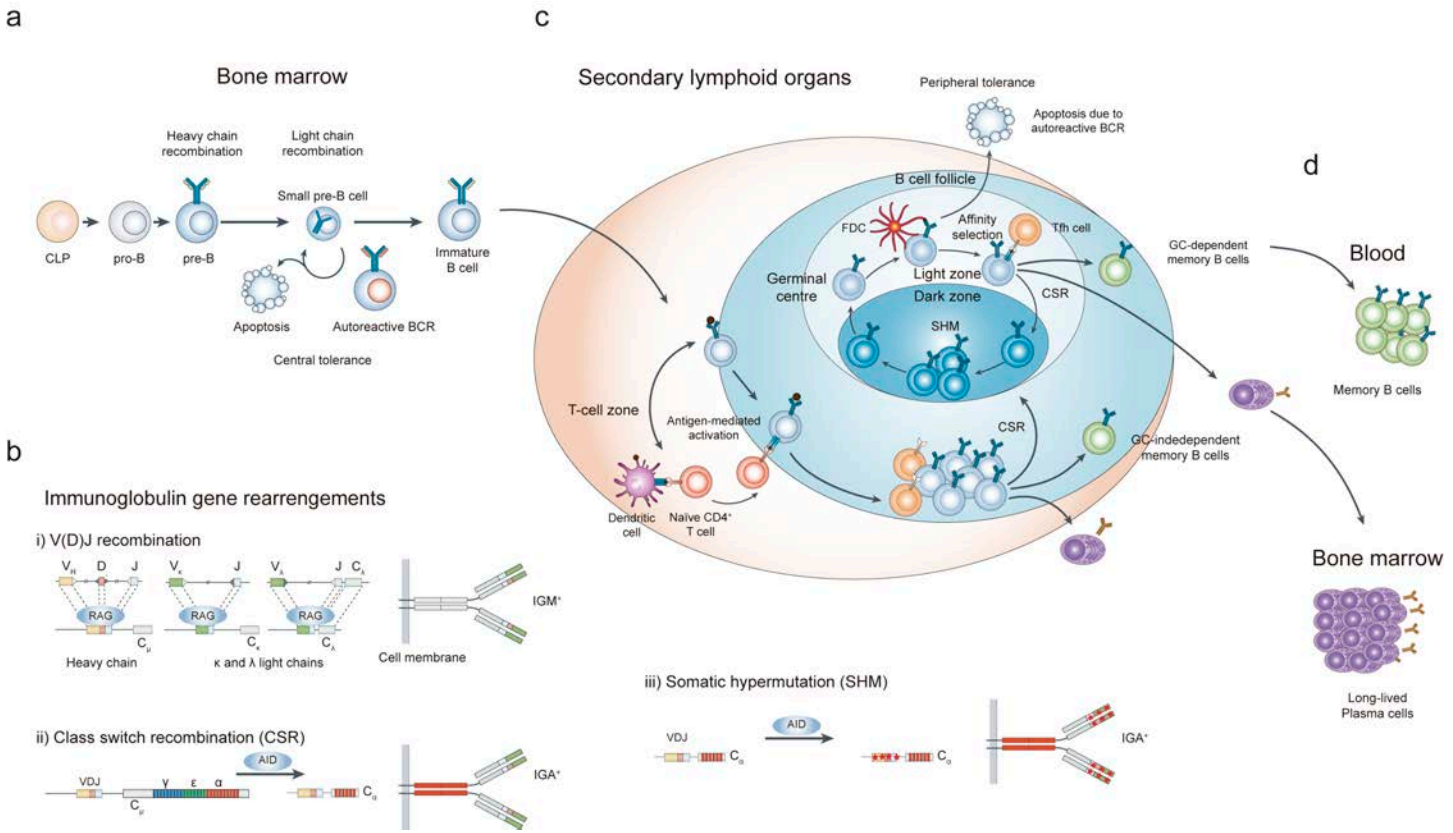


Fig. 10 | The B-cell differentiation stages and the main molecular processes involved. HPC differentiates into CLP, which eventually will give rise to precursor B cells (preB). Those preB with productive BCR rearrangements that do not recognize self-antigens will be selected and become immature B cells, which will encounter the cognate antigen in the secondary lymphoid organs such as lymph node. Upon antigen recognition, B cells will experience a clonal expansion and an affinity maturation through somatic hypermutation and class switch recombination processes to become terminally differentiated memory or bone-marrow plasma cells.

2.3 DNA methylation dynamics during B-cell differentiation.

As previously introduced, B-cell differentiation entails dramatic phenotypic changes from pro-B cells to terminally differentiated bone marrow plasma cells. These phenotypic changes are manifested in the acquisition of new cellular functions and are achieved thanks to molecular processes such as immunoglobulin gene rearrangements, CRS, SHM, as well as changes in the transcriptional profile orchestrated by key developmental TFs in every step of the maturation process. The epigenetic landscape from B-cell subpopulations at every step of the differentiation process remained poorly characterized until recently [50,157,187–189]. In 2015, we deciphered the whole DNA methylation landscape of the entire human B-cell lineage and uncovered that approximately 25% is dynamic during the entire maturation process. More precisely, a global loss of DNA methylation is observed during B-cell development, which is particularly prominent from germinal center B cells to bone marrow plasma cells. This DNA methylation loss takes place at late replicating heterochromatic regions without affecting gene expression. Conversely, DNA methylation loss at early stages of B-cell differentiation is related to enhancer commissioning and to the binding of key B-cell developmental TFs, such as EBF1, PAX5 or SP1. In addition to hypomethylation, DNA hypermethylation at polycomb-repressed regions takes place during B-cell development, particularly at late-stages of differentiation such as memory and plasma cells. Remarkably, these changes occurring in long-lived B cells were also observed in cancer DNA methylomes, suggesting a link between cancer, aging and normal B-cell development [50]. The study from my host lab was complemented by another study in which more subpopulations of mature B cells were investigated [187].

The precise characterization of DNA methylomes of normal B-cell subpopulations is of great importance for this doctoral thesis, since it represents the starting point for an improved analytical framework to understand the DNA methylation dynamics in B-cell tumors. As a co-author of this study, a reprint of this article appears in the appendix of this thesis (**appendix, manuscript 1**).

3. Hematological malignancies.

3.1. General aspects.

Hematological neoplasms comprise a large number of entities with different biological and clinical features that affect blood, bone marrow and secondary lymphoid organs. The classification of these diseases is established by the World Health Organization (WHO) and is based on the use of morphology, immunophenotype, genetic alterations and clinical features [164]. HNs originate from different cell lineages of the hematopoietic system at distinct maturation stages and through multiple pathogenetic mechanisms. This leads to a large number of entities and subtypes that can be broadly classified into myeloid, lymphoid, Hodgkin lymphomas, immunodeficiency-associated lymphoproliferative disorders and histiocytic and dendritic cell neoplasms [164]. Particularly, lymphoid neoplasms derive from mature and immature cell types of the lymphoid branch, which includes B cells, T cells and NK cells. Since this doctoral thesis has been focused on tumors derived from B cells, general concepts of B-cell neoplasias will be introduced in the next section.

3.2. B-cell neoplasias.

B-cell tumors comprise an heterogenous group of entities that can originate during the entire B-cell maturation process (**Fig. 10**). In some cases, the primary event leading to malignant transformation has been identified and is related to processes involved in normal-B-cell differentiation. For instance, normal B-cell differentiation involves mutational processes such as V(D)J recombination, CSR and SHM, and errors during these events may lead to chromosomal rearrangements that predispose to malignant transformation [190]. The most usual example is the translocation of oncogenes under the control of the IG regulatory regions, which are highly active in B cells and lead to deregulated expression of the juxtaposed oncogenes. Some examples of IG translocations are the t(11;14)(q13;q32) in mantle cell lymphoma (MCL) affecting the *CCND1* cell cycle gene, the t(8;14)(q24;q32) in Burkitt lymphoma (BL) affecting the *MYC* oncogene, or the t(14;18)(q32;q21) in Follicular Lymphoma involving the *BCL2* anti-apoptotic gene, among many others [191,192]. These genomic translocations are also informative of the cellular origin of B-cell tumors, and usually lead to an arrest in the maturation stages that ultimately lead to the immortalization and clonal expansion of malignant cells. In addition to these genomic aberrations, the cell of origin of these neoplasms is determined by further means including the morphology,

immunophenotype, histological appearance and/or gene expression patterns together with other particularities of B-cell development such as the rearrangements of the IGHV locus as well as CSR and SHM [193].For instance, the presence of ongoing or stable SHM indicates that the tumor's cellular origin is germinal or post-germinal center, whereas if SHM is absent the cell of origin is likely a pre-germinal center B cell. The importance to establish the cell of origin of B-cell tumors is not solely related to the diagnosis, but usually to their prognosis. For instance, diffuse large B-cell lymphoma (DLBCL) can be classified into two groups with different clinical behavior according to their similarity to the transcriptional profile to germinal center B cells (GCB type) or activated B cells (ABC type) [194]. Another example is the classification of chronic lymphocytic leukemia (CLL), which can be broadly classified into mutated and unmutated CLL based on the IGHV mutational status [195,196]. These groups likely represent CLL subtypes with different cellular origins of this disease, as shown by the different degrees of SHM.

Although general concepts of B-cell tumors have been briefly introduced, each of the B-cell tumors that have been studied in the course of this doctoral, namely acute lymphoblastic leukemia (ALL), MCL, CLL, DLBCL and multiple myeloma (MM), has its own particularities, and thus are introduced in greater detail in the next sections.

3.2.1. Acute lymphoblastic leukemia.

3.2.1.1. Epidemiological, biological and clinical features.

Acute lymphoblastic leukemia (ALL) is a neoplasm of precursor lymphoid cells committed to either T cells or B cells. In the United States, ALL is the most common cancer among children and the most frequent cause of death from cancer before 20 years of age [197]. Approximately 80-85% of the cases are of B-cell origin (B-ALL), and this group is the one that has been studied in this doctoral thesis. B-ALL or simply ALL from this point onwards, is primarily a disease of children, with 75% cases occurring in children younger than 6 years old [164]. The estimated annual incidence world-wide is 1-4.75 cases per 100,000 population. Although the etiology is unknown, there are known factors that predispose to a higher disease risk such as Down syndrome or some particular SNPs of genes like *IKZF1*, *CEBPE* or *CDKN2A*/*B* [198]. By definition, bone marrow is involved in all forms of ALL, and the peripheral blood is usually affected. Extramedullary involvement is common, with particular predilection for the central nervous system (CNS), lymph nodes, spleen, liver and testes. Most patients with ALL present with bone marrow

failure, including thrombocytopenia, anemia, and/or neutropenia. ALL has an overall good prognosis in children, but a less favorable prognosis in adults. The overall complete remission rate is more than 95% in children, versus 60-85% in adults. Nonetheless, there is a fraction that may progress and develop resistance to treatment. Cytogenetic abnormalities are seen in most of ALL patients and in many cases, they discriminate specific subtypes with unique phenotypic and prognostic features [199,200].

3.2.1.2. Recurrent Cytogenetic groups.

ALL with recurrent genetic abnormalities is a group of diseases which include balanced translocations and abnormalities involving entire chromosomes. The last version of the WHO classification in 2017 included nine genetically defined ALL with distinctive biological and clinical features. During this doctoral thesis, the DNA methylation of the main cytogenetic subtypes has been studied and thus are being introduced:

3.2.1.2.1. B-ALL with t(9;22)(q34.1;11.2); BCR-ABL1.

This subgroup of ALL is composed by lymphoblasts that harbor the translocation between *BCR* on chromosome 22 and the *ABL1* oncogene on chromosome 9. This disease subtype is more frequent in adults than kids, and historically *BCR-ABL1* have been considered the worst prognostic group of the major cytogenetic subtypes of ALL. Nonetheless, therapy with tyrosine kinase inhibitors has led to a significant improvement in the outcome for these B-ALL patients [197].

3.2.1.2.2. B-ALL with t(11q23.3); *KMT2A*-rearranged.

This group of ALL harbors the translocation between *KMT2A* (also called *MLL*) at band 11q23.3 and any gene from a large number of fusion partners. It is the most common leukemia in infants younger than one year, and it is thought that the translocation may already occur in utero. This ALL subtype has particularly poor clinical outcome [201].

3.2.1.2.3. B-ALL with t(12;21)(p13.2;q22. 1); *ETV6-RUNX1*.

The ALL cells of this group harbor a translocation between *ETV6* (also called *TEL*) on chromosome 12 and *RUNX1* (also called *AML1*) on chromosome 21. This leukemia is not seen in infants and infrequent in adults, but represent 25% of all ALL in children. ALL with the *ETV6-RUNX1* translocation has a very favorable prognosis, with cure seen in >90% of children [202].

3.2.1.2.4. B-ALL with hyperdiploidy (HeH).

This subgroup of ALL is characterized by blasts containing more than 50 chromosomes (but usually less than 66), typically without translocations or other structural alterations [203]. It is not seen in infants and is infrequent in adults (7-8%), but represent 25% of ALL in children. Hyperdiploid ALL has a very favorable prognosis, with cure seen in >90% of children overall, and even more commonly among children with a favorable risk profile.

3.2.1.2.5 B-ALL with t(1;19)(q23;p13.3); *TCF3-PBX1*.

The blasts of this group harbor a translocation between *TCF3* (also known as *E2A*) on chromosome 19 and *PBX1* on chromosome 1. It represents 6% in children, and is infrequent in adults. In early studies, ALL with *TCF3-PBX1* was associated with a poor prognosis, but intensive therapy introduced over the last years is improving its prognosis [204].

3.2.1.2.6 B-ALL with iAMP21.

This group is characterized by amplification of a portion of chromosome 21, typically detected by FISH with a probe for *RUNX1* that reveals 5 or more copies of the gene (or 3 or more extra copies on a single abnormal chromosome 21) [205]. It is present in children, accounting for 2% of ALL patients. ALL with iAMP21 has a relatively poor prognosis among children with cases that would otherwise be classified and treated as standard-risk ALL, although it appears that treatment of these children with more intensive therapy overcomes this adverse risk.

3.2.1.3 Epigenetic abnormalities.

DNA methylation aberrations have been broadly reported in tumors, and ALL is not an exception [206]. Several studies have generated genome-wide DNA methylation profiles with 450K arrays and WGBS, and revealed a widely altered epigenome compared with normal bone marrow [207,208]. A study using WGBS revealed a *de novo* methylation of small functional compartments [208], which was subsequently reported to be acquired stochastically [209], and a demethylation of large intercompartmental backbones in lamina-associated domains [208]. Furthermore, DNA methylation have been shown to hold diagnostic [210] and prognostic values not only using samples obtained at diagnosis [207], but also in samples at relapse, with the CpG hypermethylator phenotype (CIMP) associated with good outcome [211]. Nonetheless, the biological meaning of this CIMP in ALL remains

largely unexplored. Although the epigenetic language comprises different layers of information, DNA methylation has been the only one that have been thoroughly mapped. Other layers of the epigenome such as histone modifications and the chromatin accessibility remain to be fully characterized in this disease.

3.2.2. Mantle cell lymphoma.

3.2.2.1. Epidemiological, biological and clinical features.

Mantle cell lymphoma (MCL) is an overall aggressive B-cell lymphoma of monoclonal mature B cells with a tendency to disseminate throughout the body, infiltrate the lymphoid tissues, bone marrow, peripheral blood and extranodal sites [164]. Mantle cell lymphoma accounts for approximately 3-10% of non-Hodgkin lymphomas. It occurs in middle-aged to older individuals, with a median age of about 60 years. Its presentation is usually at the lymph nodes, but spleen and bone marrow with or without peripheral blood involvement can occur. The t(11;14)(q13;q32) translocation between an *IGH* gene and *CCND1* (encoding cyclin D1) is present in more than 95% of cases and is considered to be the primary genetic event [212]. The remaining 5% may include *CCND2*, *CCND3* translocations or the upregulation of *CCNE1* and *CCNE2* [213]. This translocation is assumed to take place in the bone marrow at the pre-B maturation stage as an error of the V(D)J recombination process. *CCND1* is a major regulator of the cell cycle by inducing G1 to S phase transition. Thus, upon its overexpression in MCL, the cell cycle becomes deregulated and leads to cell immortalization [214]. Then malignant transformation occurs later in more mature B cells, although the precise cellular origin remains elusive.

3.2.2.2. Conventional versus Leukemic non-nodal MCL.

Although traditionally MCL has been considered an aggressive lymphoma, recent studies have identified a group of patients with markedly better clinical course, with an overall survival of over 7-10 years even without the necessity of treatment at diagnosis. This MCL group has been recently termed Leukemic non-nodal mantle cell lymphoma [164,215]. This MCL subtype present distinctive clinical and molecular features compared with the classical aggressive MCL. They are characterized by a leukemic presentation with splenomegaly and absence of lymphadenopathy, together with less complex karyotypes and evidences for SHM in the IGHV locus [216,217]. Furthermore, this indolent group present also a transcriptional profile that is different form the classical MCL [218]. Remarkably, *SOX11* gene is highly expressed in the more aggressive forms of MCL, whereas lowly or

absent in the indolent forms and in normal B cells. This appreciation has led to additional studies that manifested the importance of *SOX11* in the pathogenesis of aggressive forms of MCL [219–223]. *SOX11* promotes tumor growth of MCL cells *in vivo* and regulates a broad transcriptional program that includes B-cell differentiation, cell proliferation, apoptosis and broad interactions with the microenvironment that ultimately leads to angiogenesis and tumor survival. In spite of the importance of this gene for MCL pathogenesis, no genetic alterations in *SOX11* have been reported. However, activating histone marks have been observed at the promoter region [224], but the causes leading to the gene activation are still a matter of debate and investigation.

3.2.2.3. Genetic abnormalities.

As introduced in the previous section, the primary genetic lesion of MCL is the t(11;14)(q13;q32) translocation between an *IGH* gene and *CCND1*, although other cyclin genes are also alternatively translocated/upregulated in a low proportion of the patients [213]. Thus, it seems that cyclin deregulation plays a main role in the pathogenesis of MCL. Nonetheless, some other oncogenic events seem to be needed for the generation of a MCL, since *CCDN1* upregulation alone is not sufficient to induce lymphomagenesis in transgenic mouse models [225,226]. In fact, with the exception of leukemic non-nodal MCL, MCL represents one of the most genetically complex lymphoid neoplasms, with more than 90% of patients showing secondary genetic lesions. Neoplastic cells from MCL cases contain numerous copy number alterations (CNA) including gains such as 3q, 7p and 8q, and losses such as of 1p, 6q, 9p, 11q, 13q and 17p, among others [227]. These CNA usually target oncogenes or tumor suppressor genes, e.g. 17p13 affects *TP53* and 13q14 affects *RB1*. Furthermore, next generation sequencing (NGS) initiatives have revealed that MCL contains a high complex mutational landscape [213,228,229]. Overall, *SOX11*-positive cases contain a larger number of recurrently mutated genes than those cases lacking *SOX11* expression. The most recurrent mutated gene is *ATM*, which is found approximately in 42-55% of the patients and normally associated with 11q deletions and *SOX11*-positive cases. Furthermore, other mutations may coexist, especially in the *SOX11*-positive patients, such as *NOTCH1* and *NOCTH2* (10%), *WHSC1* (10%), *MLL2* (14%) and *MEF2B* (3%).

3.2.2.4. Epigenetic abnormalities.

Several studies have profiled DNA methylation at CpG rich regions in MCL patients [230–232]. The promoter region of some genes has been found to lose DNA methylation in MCL and gain expression such as *CD37*, *HDAC1*, *NOTCH1* and *CDK5* [231]. Contrary, other genes including *SOX9*, *HOXA9*, *AHR*, *NR2F2* and *ROBO1* showed a gain in methylation at their promoters and were associated with a CIMP phenotype, high proliferation, higher number of genomic alterations and a poor prognosis [230,233]. Despite these associations, the underlying biological mechanisms leading to this CIMP phenotype remained to be fully characterized. Nevertheless, a complete and unbiased characterization of the DNA methylome of MCL was still missing and therefore constituted one of the aims of this doctoral thesis.

3.2.3. Chronic lymphocytic leukemia.

3.2.3.1. Epidemiological, biological and clinical features.

Chronic lymphocytic leukemia/small lymphocytic lymphoma (CLL/SLL) is a neoplasm composed of monomorphic small mature B cells expressing CD5 and CD23 [164]. CLL diagnosis implies a monoclonal B-cell count of more than $5x10^9$/L, with the characteristic morphology and phenotype of CLL in the peripheral blood. Individuals with a clonal CLL-like cell count lower than $5x10^9$/L and without lymphadenopathy, organomegaly, or other extramedullary disease are considered to have monoclonal B-cell lymphocytosis (MBL). About 5% healthy individuals over 40 years old present this condition, but only a small fraction will progress to CLL/SLL [234]. CLL represents the most common leukemia of adults in western countries, with an annual incidence rate about 5 cases per 100,000 population, which dramatically increases with age to more than 20 cases per 100,000 in individuals over 70 years. The median patient age at diagnosis of CLL is approximately 70 years, but it can also present in younger adults. Despite the homogeneous morphological and immunological phenotype, the clinical outcome of CLL is highly variable, with patients that remain untreated after years from diagnosis and others that require immediate treatment or even transform to a large B-cell lymphoma, a clinically-aggressive phenomenon known as Richter syndrome. This wide spectrum of clinical behaviors can be partly attributed to different disease subtypes characterized by the mutational status of the IGHV, with unmutated (≥98% identity with the germline) CLL (U-CLL) showing a worse clinical outcome than mutated (≤98% identity with the germline) CLL (M-CLL) [235,236]. These groups not only show differences in terms of clinical

outcome, but also in their genetic and epigenetic make-up (see next sections). The presence of unmutated and mutated IGHV in different samples indicates different cellular origins of this disease, although published evidence may suggest that hematopoietic stem cells could already carry pre-leukemic molecular lesions [237]. Other molecular markers have been shown to correlate with IGHV mutational status, such as the expression of *ZAP70* and CD38 [238–241]. Other studies cemented the idea of antigen selection in CLL disease ontogeny and evolution due to the discovery of almost identical or 'stereotyped' B-cell receptors among unrelated patients [242,243]. These stereotypes show different clinical behaviors and sometimes are formed by U-CLL and M-CLL patients, as it happens with stereotyped subset #2.

3.2.3.2. Genetic abnormalities.

CLL has a very heterogeneous mutational landscape [244–249]. About 80-90% of the cases have cytogenetic abnormalities detected by FISH or copy-number arrays. The most common alterations are deletions in 13q14.3 (present in approximately 50% of patients), which seems to confer a small proliferation advantage to cells [250], trisomy 12 (present in around 20%) and less commonly the deletion in 11q22-23 which usually affects *ATM* and *BIRC3*, and 17p13 deletion affecting the *TP53* tumor suppressor gene. In addition, the mutational landscape of CLL shows few genes mutated in 5-15% of the patients, such as *NOTCH1*, *SF3B1*, *ATM*, *BIRC3*, *CHD2*, *TP53* and *MYD88*, and a long tail of genes mutated in few patients [245,246]. Some of these alterations have been postulated to frequently be early events, such as trisomy 12 or *MYD88* mutations, whereas others seem to be more frequently acquired during disease progression, such as *TP53* alterations. In general, it has been shown that U-CLL tend to accumulate more CNA and mutations than M-CLLs, a feature that is associated with different clinical outcome of this subset of patients. In this context, it has been shown that particular genetic alterations confer a worse clinical outcome, such as mutations in *NOTCH1*, *SF3B1* or *EGR2*, among others. However, it seems that the total number of the driver genetic alterations may be a more potent prognostic variable rather than individual genetic changes [245]. Finally, the subclonal architecture of CLL has also been shown to be important for the clinical outcome of patients [246,251].

3.2.3.3. Epigenetic abnormalities.

CLL not only displays genetic changes but also epigenetic aberrations [244,252]. Initial studies focused on CpG islands at gene promoter [253]. Some of the genes with differential expression between U-CLL and M- CLL, such as *LPL*, *ZAP70*, *CRY1*, *SPG20*, *CLLU1* or *LAG1*, also show different DNA methylation levels that seem to be related to their distinct cell of origin. More recently, genome-wide approaches analyzed the DNA methylome of CLL and revealed that it is characterized by a widespread hypomethylation of DNA in gene bodies and enhancer elements, combined with local hypermethylation [187,254–256]. Furthermore, genetic and epigenetic aberrations may coevolve [255], and CLL methylation heterogeneity has been proposed to represent intra-tumor heterogeneity and to predispose to an adverse clinical outcome [256]. In addition to this, the genome-wide mapping of normal naïve and memory B cells together with U-CLLs and M-CLLs allowed to relate methylation patterns of U-CLL and M-CLL to normal B-cell differentiation, leading to the discovery of three epigenetic groups with different cellular origin and clinicobiological features [187,254]. These groups were termed naïve-like/low programmed CLL (n-CLL), intermediate/intermediate programed CLL (i-CLL) and memory-like/high programmed CLL (m-CLL), formed by different mutational load of the IGHV. Importantly, these groups have subsequently been validated and identified with few CpGs [257–259]. Finally, other epigenetic layers are being mapped in CLL [33,260], and are revealing the active chromatin landscape of CLL which might be mediated by some key TFs including NFAT, FOXO or TCF/LEF.

3.2.4. Diffuse large B-cell lymphoma.

3.2.4.1. Epidemiological, biological and clinical features.

Diffuse large B-cell lymphoma NOS (DLBCL) is a neoplasm of medium to large B lymphoid cells whose nuclei are usually more than twice the size of those of normal lymphocytes, and show a diffuse growth pattern [164]. The presentation is nodal or extranodal. DLBCL normally arise *de novo* (referred as primary tumor), but it can also represent a transformation from a less aggressive lymphoma (referred as secondary tumor), such as CLL (Richter syndrome), follicular lymphoma (FL), marginal zone lymphoma or nodular lymphocyte predominant Hodgkin lymphoma. These forms of DLBCL usually have a worse clinical outcome. DLBCL, NOS, constitutes 25-35% of adult non-Hodgkin lymphomas in developed countries with a median patient age around 70, but it can also occur in children and young adults. It is considered an aggressive lymphoma, and the 5-

year progression-free and overall survival rates are approximately 60% and 65% in the R-CHOP treatment era, respectively. Nonetheless, the presence of particular genetic alterations including *MYC* and *BCL2* and different subgroups confer particularly bad outcomes [193,261]. These DLBCL subdivisions include morphological variants (Centroblastic, Immunoblastic, Anaplastic, Other rare variants) and molecular subtypes. The most widely accepted molecular subtypes of DLBCL are called germinal center B-cell (GCB) and activated B-cell (ABC). These subgroups have a fundamentally different biology which is manifested in different gene expression profiles, chromosomal aberrations, and recurrent mutations and clinical outcome [193,194]. The most accepted cellular origin of DLBCL is germinal center B cells (GCB subtype) or germinal center exit/early plasmablastic or post- germinal center origin (ABC sub- type) which shows overall a worse clinical outcome. More recently, new classifications based of genetic alterations have also been proposed [262–264] (see next section).

3.2.4.2. Genetic abnormalities.

DLBCL is, together with MCL, one of the most complex B-cell tumors in terms of genetic alterations. It shows abundant CNA and mutations in a variety of genes implicated in different biological pathways [262,263,265]. As previously mentioned, DLBCL has been divided in two groups based on transcriptional profiles, namely GBC and ABC, which show different clinical outcomes and different prevalence of genetic alterations. For instance, GCB subgroup has high prevalence for *REL* amplifications, *BCL2* translocations, and *EZH2* mutations, while being rare in the ABC subgroup. Conversely, other genetic alterations such as *CDKN2A* mutations, *SPIB* amplifications or *CD79B* are very frequent in ABC subgroup while rare in GCB DLBCLs. More recently, two independent studies demonstrated that ABC and GCB may not reflect the entire genetic spectrum of DLBCL [262,263]. These two studies found different genetic clusters, which are characterized by different genetic landscapes. Importantly, these clusters add new prognostic value to the more established ABC-GCB subgrouping and allows better understanding of the biology of DLBCL. However, these clusters are not totally overlapping, with Chapuy et al defining 5 (C1-C5) and Schmitz et al four (MCD, BN2, N1 and EZB). MCD group, is characterized mainly by genetic alterations in *MYD88*, *CD79A*/*CD79B*, *PIM1*, *CDKN2A*, *HLA-A*, *HLA-B* and is similar to the C5 cluster. BN2 has genetic alterations in *NOTCH2*, *BCL6*, *TNFAIP3*, *BCL10*, *PRKCB* and can be compared with C1 cluster form Chapuy et al. N1 shows alterations in *NOTCH1*, *IRF4*, *ID3*, *BCOR* and *KLHL6*, and do not have any

cognate from Chapuy et al. Two clusters from Chapuy et al do not have any correspondence in Schmitz et al, which are C2, which has alterations in *TP53*, *CDKN2A* and abundant CNA, and C4 cluster, which is mainly composed by GCB DLBCLs that lack mutations in chromatin modifiers and *BCL2* translocations. Although these studies help to understand better DLBCL pathogenesis at the molecular level, the benefit at the clinical practice still remains to be demonstrated.

3.2.4.3. Epigenetic abnormalities.

The DNA methylation landscape of DLBCL has been also a subject of investigation during the last years. It has been shown that DLBCL with *TET2* mutations, a key regulator of active the DNA demethylation pathway, show hypermethylation at CpG-rich promoters and overlap with the bivalent mark (H3K27me3/H3K4me3) in human embryonic stem-cells [266]. Other studies have focused on the DNA methylation heterogeneity in DLBCL using both HELP assays and ERRBS. DNA methylation variability has been found to define different DLBCL subgroups that are associated with different clinical outcomes [267]. Greater magnitude of DNA methylation changes was found to be associated to a worse clinical outcome. Another study found that epigenetic heterogeneity is initiated in normal GC B-cells, increases markedly with DLBCL aggressiveness, and is associated with unfavorable clinical outcome [268]. The patterns of DNA methylation aberrations were also described to be associated with inferior clinical outcome. Lastly, a less biased approach using ERRBS with paired diagnosis-relapse DLBCL patients revealed that DNA methylation heterogeneity is informative of DLBCL relapse [269]. DLBCL patients at diagnosis showing lower DNA methylation heterogeneity have less probabilities to relapse compared with patients showing high DNA methylation heterogeneity. The authors also found a decrease in DNA methylation heterogeneity in relapsed patients compared with paired diagnosis samples, consistent with clonal evolution and selection of DLBCL tumors. How the aforementioned DLBCL subgroups defined through DNA methylation variability as well as the DNA methylation heterogeneity is associated with the genetic groups remains to be clarified. Furthermore, other layers of the epigenome remain to be fully mapped in DLBCL patients.

3.2.5 Multiple myeloma.

3.2.5.1. Epidemiological, biological and clinical features.

Multiple myeloma (MM) is a neoplasm derived from the neoplastic proliferation of plasma cells in the bone marrow, usually accompanied by an M protein in serum and/or urine and evidence of organ damage [164]. Bone marrow is the site of origin of virtually all MM, and in most of the cases there is disseminated bone marrow involvement, although other organs may be involved, especially at advanced stages of the disease. Furthermore, MM can progress to bone marrow-independent diseases, such as extramedullary myeloma and plasma cell leukemia. Although the exact cell of origin has not been established, the postulated normal counterparts of MM are post-germinal center long-lived plasma cells in which the IG genes have undergone CSR and SHM. The diagnosis of MM is based on clinical, morphological, immunological and imaging features [270]. MM accounts for 10-15% of hematopoietic neoplasms, with a higher prevalence in male. It is almost never found in children and very infrequent in adults younger than 30 years. The incidence increases with patient age thereafter, with about 90% of cases occurring in patients older than 50 years (the median patient age at diagnosis is around 70 years). Almost all MM arise from a precursor condition called precursor monoclonal gammopathy of undetermined significance (MGUS) with or without an identified intervening stage, referred to as smoldering multiple myeloma (SMM). MGUS is asymptomatic and present in more than 5% of individuals older than 70 years. In a nearly 15% of patients MGUS progress and give rise to a MM, which shows a markedly poorer clinical outcome. MM is considered an incurable progressive disease, but newer therapeutic approaches have significantly improved quality of life and survival of the patients [271,272].

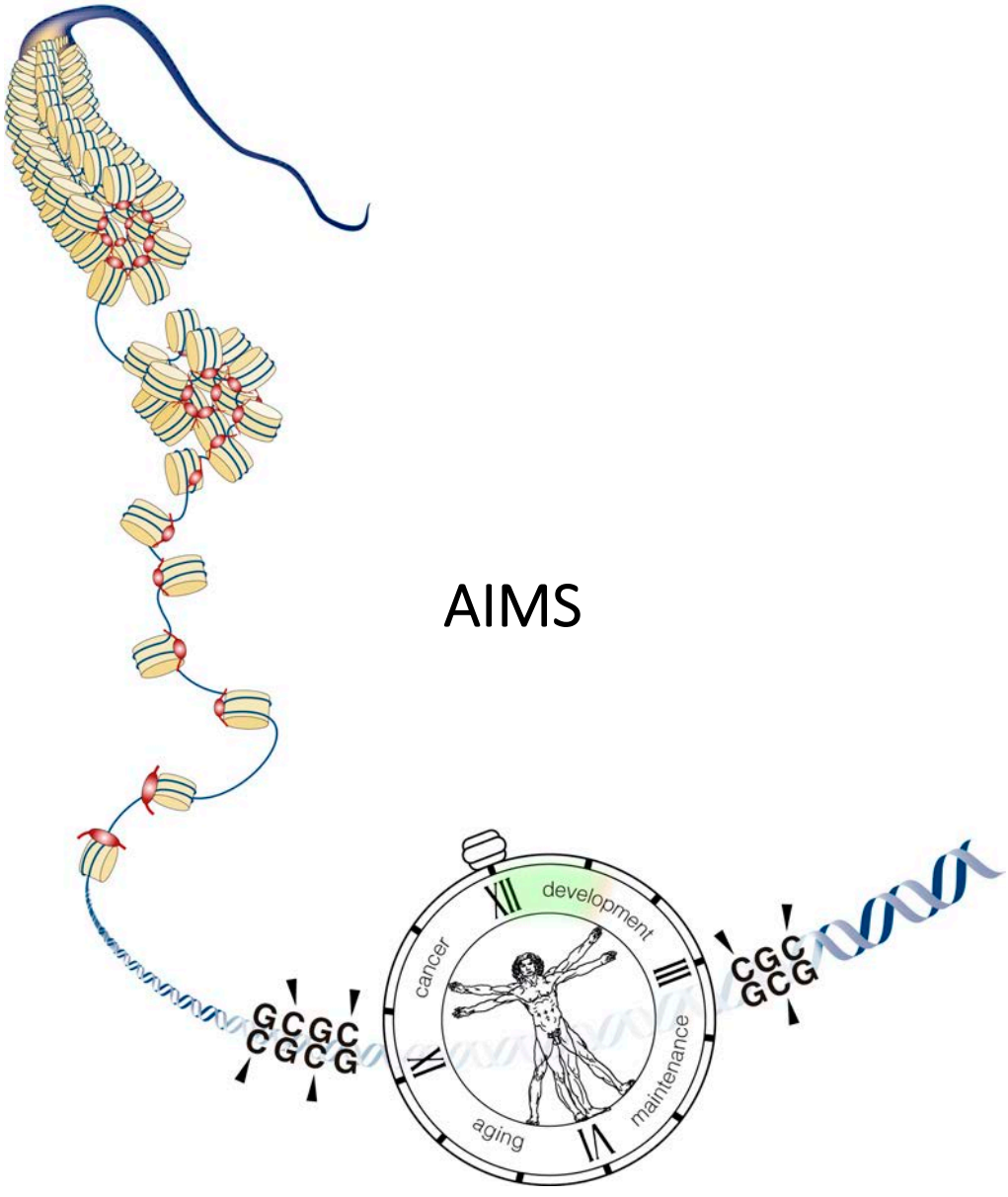
3.2.5.2. Genetic abnormalities.

MM is clinically and biologically heterogeneous with several genetic alterations proposed as driving events of myelomagenesis [273]. It is though that chromosomal translocations (involving the *IGH* gene) and aneuploidy in post GC B cells represent the primary genetic events leading to the development of MGUS. Next, other genetic and epigenetic alterations may accumulate and give rise to clonal evolution and disease progression [274,275]. Some particular genetic aberrations have been shown to manifest in distinct transcriptional profiles [276,277], which further manifests the genetic heterogeneity of MM cells. Although having a complex genetic landscape, no primary event in multiple

myeloma has been identified, and MM patients usually show the presence of coexistent subclones of malignant plasma cells with partially overlapping and unique mutations [273,278]. The most frequently mutated genes in MM patients are *KRAS* (23%), *NRAS* (20%), *FAM46C* (11%), *DIS3* (11%) and *TP53* (8%), but others may be present with less frequency, including *BRAF*, *TRAF3*, *PRDM1*, *CYLD*, *RB1* and *IRF4*, among others. Furthermore, no leading genetic alteration has been found to promote the progression from MGUS to MM, although some genetic and epigenetic alterations have been found more frequent in a subset of patients with higher probabilities to progress to MM. The molecular mechanisms responsible for the progression from MGUS to MM is of particular interest, because implies the transition from an indolent and asymptomatic state without the need for treatment to a life-threatening condition that urgently needs treatment.

3.2.5.3. Epigenetic abnormalities.

The epigenomic landscape of MM is less characterized compared with its genomic profile. Nonetheless, it seems that MM do show epigenomic alterations. In particular, global DNA methylation was shown to be variable in MM, with some patients showing a global hypomethylation and others showing a global hypermethylation, compared with normal plasma cells. Levels of hypermethylation are similar in MGUS and MM, whereas levels of hypomethylation are increased in MM, suggesting that this might play a part in disease development [279,280]. Furthermore, specific B-cell enhancers that are demethylated in normal B-cell development become aberrantly methylated in MM, a finding that seems to reflect the fact that this neoplasia losses the B-cell epigenetic identity [280]. In addition to this, aberrant methylation of promoter of tumor suppressor genes leading to a worse clinical outcome has been identified [281]. In particular, methylation of the promoter regions of *GPX3*, *RBP1*, *SPARC* and *TGFBI* genes was found to be associated with significantly shorter overall survival independent of other established prognostic factors, such as age and certain cytogenetic lesions. Finally, ChIP-seq and chromatin accessibility have been recently mapped in primary MM patients [282,283]. These studies manifest the first signs of an altered epigenome in MM, which overall involves a widespread decompaction of heterochromatin leading to an altered TF network and an overall chromatin activation as a unifying pathogenetic mechanism of MM patients [284].


AIMS
development
maintenance
aging
cancer

Over the last years, the DNA methylome of normal B-cell cell development and multiple B-cell neoplasms has been thoroughly mapped. Nonetheless, a comprehensive view of the DNA methylation dynamics of B-cell neoplasms from different maturation states in the context of the entire B-cell differentiation program has not been performed yet. Therefore, the main goal of this doctoral thesis was to dissect the sources of DNA methylation variability across B-cell neoplasias and normal B cells, with the ultimate aim to obtain new biological and clinical insights with potential clinical translation.

This global aim is in turn divided into the following specific aims:

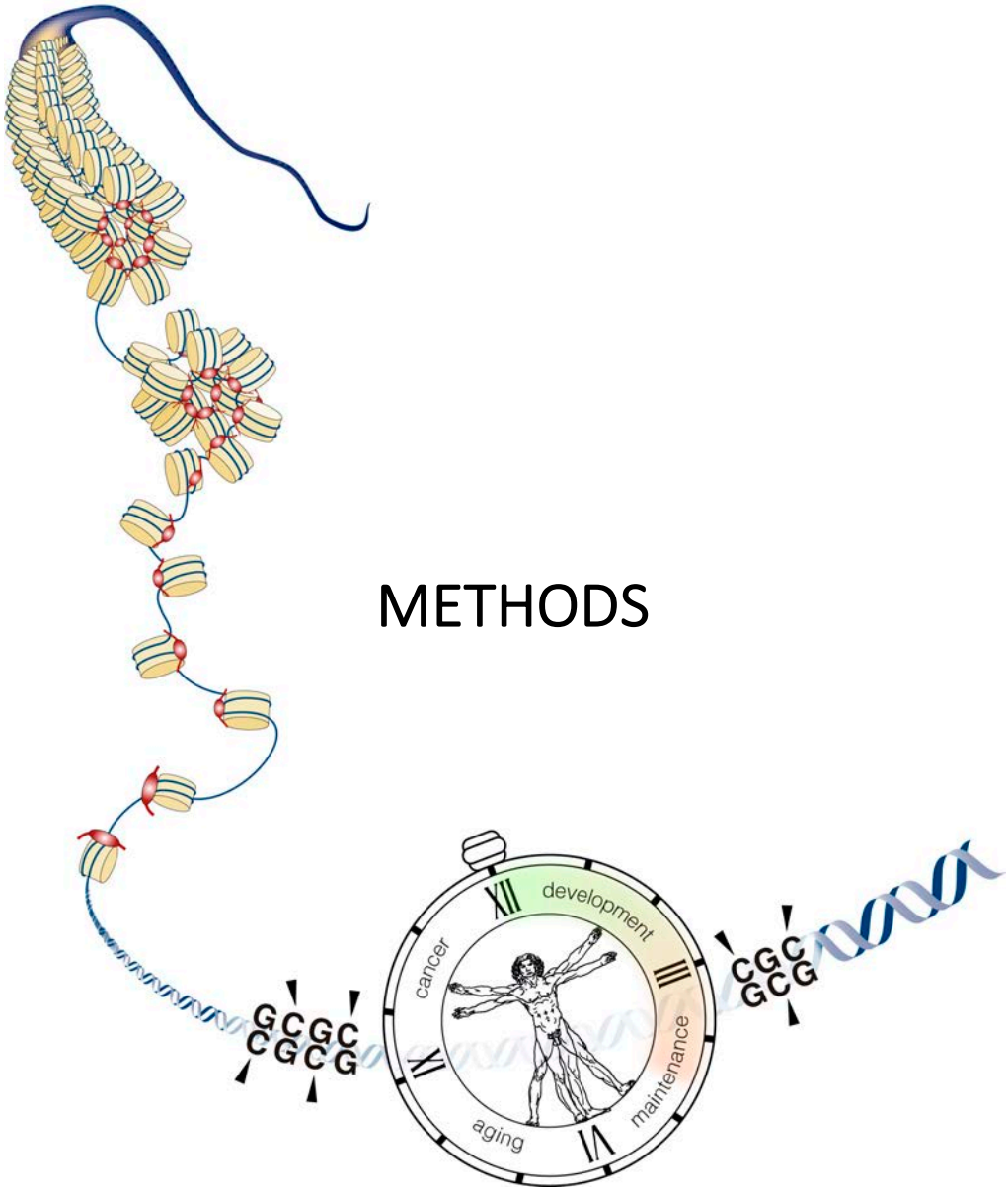
1. Study the entire DNA methylome of MCL in the context of normal B-cell development and establish an analytical framework to integrate all B-cell neoplasias (**Study 1**).

2. Dissect the sources of DNA methylation variability across B-cell neoplasias spanning the entire B-cell maturation spectrum using a systematic approach that includes cancer-specific, entity-specific, subtype-specific and individual-specific variation (**Studies 2 and 3**).

3. Identify differential DNA methylation patterns among B-cell tumor entities and subtypes with different clinical management (**Study 2**).

4. Determine the causes and clinico-biological implications of patient-specific DNA methylation changes in B-cell tumor entities and subtypes (**Studies 1, 2 and 3**).

5. Study the relationship between genetic and epigenetic aberrations in CLL and MCL to better understand the heterogeneous clinical behavior of individual patients (**Studies 3 and 4**).


GCGC
CGCG
XII
development
III
maintenance
VI
aging
IX
cancer
CGC
GCG

METHODS

1. Methods Study 1.

1.1. DNA methylation analysis with 450K arrays in MCL.

The EZ DNA Methylation Kit (Zymo Research) for bisulfite conversion was used to convert 500 ng genomic DNA per sample. Bisulfite-converted DNA was hybridized onto the HumanMethylation 450K BeadChip kit (Illumina) which covers 99% of RefSeq genes and 96% of CpG islands. The Infinium methylation assay was carried out as previously described [158]. Data from the 450k arrays were analyzed in R using the *minfi* package (version 1.18.2)[159], available through the Bioconductor open source software. To exclude technical and biological biases that might produce false results in further analyses, we developed and optimized an analysis pipeline with several filters (i.e. discarding CpGs with low detection P values, sex-specific CpGs, CpGs showing individual-specific methylation and CpGs overlapping with SNPs). Considering the different performance of Infinium I and Infinium II assays we used the subset-quantile within array normalization called SWAN [285] that corrects for the technical differences between the Infinium I and II assay designs and produces a smoother overall beta value distribution.

1.2. Deconvolution and *in-silico* purification of DNA methylation in MCL.

In order to estimate the purity for all MCL samples we used a statistical framework formally presented by Houseman and coworkers in 2012 (Houseman et al., 2012) and later adapted for 450k Illumina arrays by Jaffe and Irizarry in 2014 (Jaffe and Irizarry, 2014), made accessible through the function "estimateCellCounts" in the *minfi* package (version 1.18.2) available at Bioconductor. This function allows to determine the proportion of the main cell fractions of leukocytes in peripheral blood (B cells, CD8+ T cells, CD4+ T cells, NK cells, Monocytes and Granulocytes). As I report in **Study 1**, MCL has a highly disrupted DNA methylome, which in part affects CpGs whose methylation is modulated during B-cell differentiation, and therefore, the function developed for normal peripheral blood samples may not be accurate in our experimental setting. I initially compared the B-cell fraction estimated *in silico* and FCM data from 32 MCLs as gold standard, and observed an underestimation of B cell purity (data not shown). Therefore, we modified the published algorithm [286], and added the functionality to select a priori a set of CpGs for the deconvolution process and the capacity to remove unwanted CpGs, as well as the possibility to extract which set of CpGs are being used for the deconvolution. I proceeded selecting MCLs samples with high purity (≥99, based on FCM data) and comparing them to flow-sorted purified B cells [287]. I retained those CpGs

showing similar methylation values (i.e. a difference ≤ 0.1 in methylation levels, n=184,547 CpGs). These CpGs represent those whose methylation levels globally remain stable in MCL lymphomagenesis and therefore are useful for the deconvolution process. I next provided those CpGs to the modified version of the aforementioned algorithm and obtained a more precise tumor B-cell content estimate benchmarked against FCM data (R=0.95, pvalue<0.001) than with the initial approach [286](Results, Section 1.2.1, Fig.10c). It is important to note that I improved the accuracy when predicting tumor B-cell content while we maintained the precision in discriminating the non-B cells of the microenvironment. Afterwards, I assumed that the methylation estimates of MCL samples were the sum of the methylation estimates of the different cell types averaged by their calculated proportions:

$$Y = \sum_{1}^{n} \beta\theta = \beta_1\theta_1 + \beta_2\theta_2$$

where *Y* is the measured DNA methylation of the MCL samples, *n* is the total number of cell type fractions (6), *B1* is the *in-silico* purified MCL DNA methylation estimates, *O1* is the tumor B cell fraction, *B2* is the DNA methylation estimates of the non-tumor B cell and *O2* is the proportions of non-tumor B cell, respectively. I proceeded with all downstream analysis in Study 1 with this computationally purified methylation estimates (*B1*) in all MCL samples. As a proof of principle, I showed that the DNA methylation variability associated with sample cellular compositions that was present with the initial methylation estimates was minimized after adjusting for MCL purities.

1.3. Whole-genome bisulfite sequencing (WGBS).

WGBS was performed on two MCL samples, one from each of the defined epigenetic groups (C1 MCL and C2 MCL, please be referred to the next section for further details on these groups). Genomic DNA (1–2 µg) was spiked with unmethylated λ DNA (5 ng of λ DNA per µg of genomic DNA) (Promega). The DNA was shared by sonication to 50–500 bp using a Covaris E220 and fragments of size 150-300 bp were selected using AMPure XP beads (Agencourt Bioscience Corp.). Genomic DNA libraries were constructed using the Illumina TruSeq Sample Preparation kit (Illumina Inc.) following the lllumina standard protocol: end repair was performed on the DNA fragments, an adenine was added to the 3' extremities of the fragments and Illumina TruSeq adapters were ligated at each extremity. After adaptor ligation, the DNA was treated with sodium bisulfite using the

EpiTexy Bisulfite kit (Qiagen) following the manufacturer's instructions for formalin-fixed and paraffin- embedded (FFPE) tissue samples. Two rounds of bisulfite conversion were performed to ensure a conversion rate of over 99%. Enrichment for adaptor-ligated DNA was carried out through 7 PCR cycles using the PfuTurboCx Hotstart DNA polymerase (Stratagene). Library quality was monitored using the Agilent 2100 BioAnalyzer (Agilent), and the concentration of viable sequencing fragments (molecules carrying adaptors at both extremities) estimated using quantitative PCR with the library quantification kit from KAPA Biosystem. Paired-end DNA sequencing (2x100bp) was then performed using the Illumina Hi-Seq 2000.

Read mapping was carried out using the GEM aligner (v1.242)[288] against a composite reference containing two copies of the human GRCh37 reference and two copies of the NCBI viral genome database (v35). For both the human and viral references, one copy had all C bases replaced by T and the other had all G bases replaced by A. The names of the contigs in the combined reference FASTA file were modified by adding #C2T or #G2A to the end of the contig names depending on the conversion performed. Before mapping was performed the original sequence of each read was stored. The first read of each pair then had all C bases replaced by T, and the second read had all G bases replaced by A. Read mapping with GEM was performed allowing up to 4 mismatches per read from the reference. After read mapping the original sequence of each read was restored. Estimation of cytosine levels was carried out on read pairs where both members of the read mapped to the same contig with consistent orientation, and there was no other such configuration at the same or less edit distance from the reference. After mapping, we restored the original read data in preparation for the inference of genotype and methylation status. We estimated genotype and DNA methylation status simultaneously using software developed at the Centro Nacional de Análisis Genómico (CNAG, Barcelona, Spain)[289], considering the observed bases, base quality scores and the strand origin of each read pair. For each genome position, estimates for the most likely genotype and the methylation proportion were produced (for genotypes containing a C on either strand). A phred scaled likelihood ratio for the confidence in the genotype call was estimated for the called genotype at each position. For each sample, CpG sites were selected where both bases were called as homozygous CC followed by GG with a Phred score of at least 20, corresponding to an estimated genotype error level of <= 1%. Sites with >500x coverage depth were excluded to avoid centromeric/telomeric repetitive regions. A common set of called CpG sites for all

analyzed samples was generated, and all subsequent analyses used this common set (n=11,384,077 CpGs)

1.4. Identification of MCL subgroups.

First, a PCA was performed using the "prcomp" function (*Stats* package, R software). Then using linear discriminant analysis at each component individually a cut-off was identified comparing the germinal center-inexperienced B cells versus germinal center-experienced B cells. This cut-off value was used to identify which MCL samples were more similar to germinal center-inexperienced B cells and which were more similar to germinal center-experienced B cells.

1.5. Differential DNA methylation analyses.

To define regions with differential methylation between different MCL subgroups or between MCLs and normal B cell controls, using the 450K array data, the following two steps were used: (i) an absolute difference of mean DNA methylation levels of at least 0.25 between the compared groups and (ii) a false discovery rate (FDR) of less than 0.05 using a Wilcoxon test for independent samples. In the case of WGBS, another two different strategies were used. On the one hand, differentially methylated CpGs (DMCs) were identified in a pair-wise comparison of each MCL sample versus HPCs, and the 2 MCL samples versus each other. Statistical significance difference in DNA methylation was estimated based in beta-binomial distribution using the "bdiff" algorithm [290], and a DNA methylation difference of > 0.25. Annotation of CpGs was performed using the UCSC Table Browser GRCh37/hg19 version and considering each feature related to a gene. On the other hand, differentially methylated regions (DMRs) were calculated using a Hidden Markov Model (HMM) to segment the methylation values in the two samples under consideration. The Markov Model had 3 states corresponding to low, intermediate and high methylation; the transition probabilities were 0.9 for staying in the same state and 0.1 to change state. The emission probabilities of the HMM were the probabilities of obtaining the observed count of non- converted and converted reads assuming an underlying methylation value of less than 0.3 (low state), between 0.3 and 0.7 (intermediate state) and higher than 0.7 (high state). Stretches in the genome which correspond to the first samples being in "high" state and the second sample being in "low" state (or vice versa) were candidates for being DMRs. An additional filtering was

that these regions should contain more than 3 CpGs and a difference in average methylation larger than 0.25.

1.6. Genomic and functional annotation of CpGs.

Both CpGs analyzed by WGBS and 450K array data were annotated using the UCSC Genome Browser database (hg19). For the location relative to a gene, the following categories were used: (i) TSS 1500 (from 201 to 1,500 bp upstream of the transcriptional start site (TSS)), (ii) TSS 200 (from 1 to 200 bp upstream of the TSS), (iii) 5′ UTR, (iv) first exon, (v) gene body (from the first intron to the last exon), (vi) 3′ UTR and (vii) intergenic regions. Owing to the presence of alternative transcription start sites and regions containing more than one gene, some of the CpGs were assigned multiple annotations. For the location relative to a CpG island (CGI), the following groups were used: (i) within CGI, (ii) in CGI shore (0–2 kb from the CGI edge), (iii) in CGI shelf (>2 kb to 4 kb from the CGI edge) and (iv) outside CGI.

All CpG probes were also annotated according to chromatin activity by mapping chromatin states in normal naive (n=1) and memory B cells (n=1), as well as in 2 representative MCL cases (for which also the WGBS and 4C-sequencing was performed). To that end the chromHMM software was used [45] using ChIP-seq data of 6 histone marks (H3K4me1, H3K4me3, H3K27ac, H3K36me3, H3K27me3, H3K9me3) as input. The following regions were considered, adapted from the previously published segmentation: Active promoter (H3K4me3, H3K27ac), Weak Promoter (H3K4me1, H3K4me3), Poised Promoter (H3K4me1, H3K4me3, H3K27me3), Strong Enhancer 1 (H3K4me1, H3K4me3, H3K27ac), Strong Enhancer 2 (H3K4me1, H3K27ac), Weak Enhancer (H3K4me1), Transcription Transition (H3K36me3, H3K4me1, H3K27ac), Transcription Elongation (H3K36me3+), Weak Transcription (H3K36me3), H3K9me3 Repressed (H3K9me3), H3K27me3 Repressed (H3K27me3), Heterochromatin Low signal (none of the 6 marks).

1.7. ChIP-seq experiments and analysis.

ChIP-Seq of H3K4me1, H3K4me3, H3K27ac, H3K36me3, H3K27me3, and H3K9me3 were performed using standard protocols generated within the Blueprint Consortium. Protocol details can be found at http://www.blueprint-epigenome.eu/index.cfm?p=7BF8A4B6-F4FE-861A-2AD57A08D63D0B58. Catalog

numbers of antibodies (Diagenode) used are H3K27ac: C15410196/pAb-196-050 (LOT: A1723-0041D), H3K4me1: C15410194/pAb-194-050 (LOT: A1863-001P), H3K4me3: C15410003-50/pAb-003-050 (LOT: A5051-001P), H3K36me3: C15410192/(pAb-192-050 (LOT: A1847-001P), H3K9me3: C15410193/pAb-193-050 (LOT: A1671-001P), H3K27me3: C15410195/pAb-195-050 (LOT: A1811-001P). Regions significantly enriched for H3K27ac were determined using MACS2 using default setting [291].

1.8. 4C-sequencing experiments and analysis.

4C templates were prepared as previously described [292,293]. Briefly, $4x10^6$ cells (naive and memory B cells) were crosslinked with 2% (MCL primary cases) or 1% (naive and memory B cells) formaldehyde, chromatin was digested with a first restriction enzyme (NlaIII) followed by ligation. Next, chromatin was decrosslinked and DNA was digested with a second restriction enzyme (DpnII) and ligated. PCR amplification of viewpoint regions and its ligated fragments was performed using the 4C templates. The alternative protocol for naive and memory B cells was used due to the availability of lower cell numbers of these samples. It was tested using Z-138 ($2x10^6$ cells) and we observed that with the new protocol the interaction between the *SOX11* promoter and its distant enhancer was well detectable in Z-138. Samples were sequenced with the MiSeq instrument (Illumina) using 50bp single-reads, adding 5% PhiX. Restriction enzymes and primers used are indicated in the table below. 4C-seq analysis was performed using the 4C-seq pipeline 4cseqpipe, which can be found here: http://compgenomics.weizmann.ac.il/tanay/?page_id=367 .

1.9. Analysis of the proliferation signature in MCL.

Gene expression data of 25 MCL patients for which DNA methylation data was available was mined [218]. The expression levels of 18 genes were used to calculate the proliferation signature [294]. To that end, the mean of the “rma” normalized expression values of these 18 genes was calculated per sample. To calculate the relative difference of the proliferation signature compared to the group average, the average proliferation signature value of the group was subtracted from the proliferation signature per case. This average proliferation signature was then correlated with the number of DNA methylation changes in cases from C1 and C2 MCLs

1.10. Statistical analyses of clinicobiological variables.

The relationships between MCL subgroups and clinical and biological variables of was evaluated using the Fisher's exact test in the case of qualitative variables and the t-test for independent samples in the case of quantitative variables with normal distribution (a corrected p-value was used if the two groups had unequal variances). Statistical significance was defined as a p-value ≤ 0.05. The sample size in each of the comparisons varied depending on the available data. The relationship between the number of epigenetic changes and overall survival was examined by plotting the number of changes (x-axis) vs. the linear predictor plus martingale residuals (y-axis). The local regression line suggests a linear relationship between the number of changes and the log hazard. To detect MCL groups with different clinical behavior based on their DNA methylation changes, maxstat R package was used. This analysis allows to detect the most suitable threshold to separate MCL groups with differences in overall survival. Overall survival Kaplan-Meier plots and long-rank tests were performed with the IBM-SPSS Statistics version 20.

2. Methods Studies 2, 3 and 4.

2.1. Initial processing of 450k DNA methylation data.

I collected 450k DNA methylation array data for 913 ALL [207,208], 82 MCL [295], 491 CLL[245], and 104 MM [280]. I also used 450K data from our previous publication of normal B-cell subpopulations [50] totaling 67 samples as well as normal microenvironmental cells including 6 granulocytes, 5 $CD8^+$ and 5 $CD4^+$ T cells, 6 monocytes, 6 NK cells 6 whole blood samples and 6 peripheral blood mononuclear cells [287], 6 macrophages [296] and 16 endothelial cells [297]. These microenvironmental cells were used to infer B-cell tumor purities through DNA methylation data. In addition, genome-wide DNA methylation profiles measured by HumanMethylation450 BeadChip (Illumina) for 80 DLBCL samples with available genomic data were generated [265]. The EZ DNA Methylation Kit (Zymo Research) for bisulfite conversion was used to convert 500 ng genomic DNA per sample. Bisulfite-converted DNA was hybridized onto the HumanMethylation 450K BeadChip kit (Illumina) as previously reported [158]. In total, the DNA methylation of 1,799 samples was analyzed. I used a custom pipeline to analyze DNA methylation data using R packages and core Bioconductor packages, with special use of *minfi* package exclusively devoted to analyze DNA methylation data [159]. From the total of 485,512 probes present in the 450k array, I sequentially removed probes using the next steps: I initially removed 3,091 non-CpGs probes, 17,534 CpGs representing SNPs, 7,715 CpGs with individual-specific methylation [50], and 4,493 CpGs present in sexual chromosomes. All the remaining 452,679 CpGs had a detection p-value ≤0.01 in more than 10% of the samples. I then removed samples with bad intensity signal and/or bad probe conversions as well as those with a tumor percentage below 60% (please, see next section). In total, I removed 104 ALL samples, 8 MCL samples, 1 CLL sample, 25 DLBCL samples and 4 MM samples. I also removed microenvironmental cells to perform all the analyses in normal and neoplastic B cells. After all filtering criteria, I retained 1,595 samples with DNA methylation values for 452,679 CpGs, which were normalized using SWAN algorithm [285]. Some CpGs showed missing values in some samples and were removed from all the subsequent analyses (with the exception of biomarker discovery, Results section 2.2.5) and finally 437,182 CpGs were used. I used *lluminaHumanMethylation450kanno.lmn12.hg19* and *IlluminaHumanMethylationEPICanno.ilm10b4.hg19* R packages to annotate all CpGs. B-cell related and B-cell independent CpGs classification was used from our previous study to separate CpGs that are significantly modulated or not during B cell differentiation,

respectively [50]. The same pipeline was used to curate and normalize the data for the *in-vitro* model of B-cell differentiation previously published [298] and for all the DNA methylation data in the validation series used for the pan-B-cell tumor classifier. This pipeline was also used to establish clinical associations, which included new DNA methylation generation and DNA methylation data collection from previously published studies. Specifically, DNA methylation profiles for 70 MCL (58 non-overlapping with those in Study 1) and 12 DLBCL patients using Infinium MethylationEPIC BeadChip (Illumina) following manufacturer's instructions. I additionally collected DNA methylation profiles including 450K and EPIC data for 380 CLL from external collaborators. Finally, for ALL validations, I used 183 samples included in the initial analysis (Fig. 15a)[208] but not used to construct any classifier, and I also downloaded DNA methylation data from GSE76585 [299] and GSE69229 [300].

2.2. Inferring tumor purity through DNA methylation data.

DNA methylation has been shown to represent an appropriate biological layer to infer the proportions of blood cell types in peripheral blood [301–303]. I have previously implemented successfully this statistical framework in **Study 1** to infer tumor purity in MCL patient samples [295,304]. I have extended this strategy to all B-cell tumors using additional cell types to deconvolute DNA methylation data into cellular proportions including tumor cell content. I validated this approach using flow cytometry (FCM) and genetic data in MCL and CLL samples. Briefly, I assume that B-cell tumors retain a B-cell signature from its cell of origin and also have negligible proportion of normal B cells [305]. Thus, the percentage of neoplastic B cells in a sample can be inferred by the presence of a DNA methylation signature of B cells. This B-cell methylation signature was identified by two sequential steps: 1) I selected CpGs with shared methylation values during the entire B-cell maturation process (from early committed B cells to terminally-differentiated bone marrow plasma cells), and 2) from those CpGs selected above, I performed a differential DNA methylation analysis to identify CpGs whose methylation level was significantly different between B cells and the major non-neoplastic cells accompanying B cell tumors [306], namely granulocytes, T cells, monocytes, macrophages and endothelial cells. Then, with this set of CpGs representative of all major cell types present in tumor samples, I apply a linear constrained projection [301,302], also known as *reference-based* approach [307,308], to find the proportions of each cell type. Thus, since I did not have FCM and genetic data for all the cases to know tumor cell content, I relied

on DNA methylation-based predictions. However, in the case of MM and DLBCL samples this approach did not seem to be valid, since the number of samples predicted with tumor cell content equal or greater that 60% was clearly underestimated compared with FCM and genetic predictions (Fig. 18b, c). In the case of MM, it has been shown that it loses the DNA methylation signature related to B cells [280], and thus, as expected, my strategy using the B-cell methylation signature to infer tumor purity was not adequate. Intriguingly, I found a similar phenomenon in some DLBCL cases, in which the tumor cell content was high according to genetic events and very low according to DNA methylation-based prediction. Thus, I retained patient samples showing at least 60% tumor cell content according to DNA methylation-based predictions in ALL, MCL and CLL samples, to FCM in MM and to genetic data in DLBCL samples.

2.3. Tumor cell content estimation in DLBCL using genetic data

The 80 samples included have been analyzed in the context other studies by whole-genome copy number (CN) arrays (Cytoscan HD, Affymetrix) and gene mutations by targeted next generation sequencing of 106 genes [265]. The CNA data and gene mutations found in these samples were used to infer their potential tumor purity. The Allele-Specific Copy Number Analysis of Tumors (ASCAT) algorithm available at Nexus Copy Number (BioDiscovery, version 7) was used to infer the tumor purity directly from the Cytoscan HD array. The percentage of cells (or cancer cell fraction, CCF) carrying each somatic mutation found in loci not affected by a copy number alterations was calculated as CCF = 2xVAF, where VAF is the variant allele frequency of the mutation. Out of all the mutations, the highest CCF was considered as the best estimate of tumor purity of the samples based on gene mutations. As a final step, the maximum tumor purity detected by ASCAT or gene mutations was considered as the estimated tumor cell purity and used in downstream analyses. Using this approach, I retained 55 DLBCL samples showing at least 60% tumor cell content.

2.4. Gene expression data integration.

Gene expression profiles using hgu219 array for normal B cells was obtained from a previous study [50] and included 3 hematopoietic precursor cells, 7 pre-B cells, 10 naïve B cells, 11 germinal center B cells, 5 tonsillar plasma cells, 5 memory B cells and 1 bone marrow plasma cell. Additionally, I downloaded previously reported gene expression data for 56 ALL samples profiled with 133 plus 2 array [207], including several ALL subtypes,

namely 18 HeH, 5 11q23/MLL, 16 t(12;21), 6 t(1;19), 5 t(9;22) and 6 dic(9;20). I also used 15 MCL samples profiled with 133 plus 2 arrays [216] including 10 C1 and 5 C2 MCLs from the initial MCL cohort, and 23 C1 MCL and 14 C2 MCL profiles with the hgu219 array in the MCL validation series. I also used previously generated gene expression data with hgu219 array for 455 CLL samples [245]. For DLBCL samples, gene expression data was generated using 133 plus 2 arrays following the manufacturer's instructions for 43 DLBCL samples, including 17 GCB, 15 ABC, and 11 unclassified. Finally, I downloaded gene expression data for 328 MM samples from analyzed with the 133 plus 2 array platform [277]. I normalized all the data using "rma" function available in the *affy* R package. I am aware of potential incorrigible batch effects that are present when analyzing all these gene expression data together, where different entities come from different studies and also from different array platforms. Thus, I adopted several measures to minimize possible effects on our biological associations. In particular, I used gene expression data to strengthen the interpretations of previous results and not for primary and discovery analyses. Furthermore, I have prioritized qualitative instead of quantitative analyses. To do so, I have transformed all normalized gene expression values per sample to gene expression percentiles, and compared if genes are more or less expressed.

2.5. Shared DNA methylation dynamics in normal and neoplastic B cells.

To define CpGs whose methylation values do not change in normal and neoplastic B cells, I obtained CpGs showing differences of less than 0.25 across all normal and neoplastic B cells. Then, I classified them in hyper-, partially- and hypomethylated CpGs calculating the median of each CpGs for all the samples. This analysis revealed that approximately 12% of the human DNA methylome is stably methylated under normal B-cell development and malignant transformation and progression. Thus, this indicates that the majority, i.e. 88% of the DNA methylome is labile under normal and/or neoplastic conditions. This 88% includes the DNA methylation modulation during normal B-cell differentiation, which I and other studies estimated to be around 25% of the entire DNA methylome [50,187] Thus, the number of DNA methylation changes taking place exclusively in the context of B-cell tumor malignant transformation and progression should be around 63% (88%-25%).

2.6. ChIP-seq data collection, analysis and integration.

ChIP-seq data available from the Blueprint consortium [3] and from a previous study in ALL [309] was downloaded and processed. Particularly, Blueprint ChIP-seq data of six non-redundant histone marks was used, including H3K4me1, H3K4me3, H3K27ac, H3K36me3, H3K27me3 and H3K9me3 for 15 normal B cells (6 NBC, 3 GC, 3MBC and 3tPC), 5 MCLs, 7 CLLs and 4 MMs, as well as two DLBCL cell lines, i.e. KARPAS-422 and SUDHL-5 DLBCL. The antibodies used in the context of the Blueprint project were thoroughly validated http://www.blueprint-epigenome.eu/index.cfm?p=D6F8811F-DACF-7979-CEAC0B9034C28037). Next, these ChIP-seq data was integrated by using chromHMM software [45] as previously described [33]. Briefly, ChIP-seq data of the six histone marks was used to generate a B-cell specific chromatin state model with 12 emission states using the 15 normal B cells, corrected for their corresponding input. These 12 chromatin states were ActProm (active promoter, with H3K27ac and H3K4me3 marks), WkProm (weak promoter, with H3K4me1 and H3K4me3 marks), PoisProm (poised promoter, with H3K27me3, H3K4me1 and H3K4me3 marks), StrEnh1 (strong enhancer 1, with H3K27ac, H3K4me1 and H3K4me3 marks), StrEnh2 (strong enhancer 2, with H3K27ac and H3K4me1 marks), WkEnh (weak enhancer, with H3K4me1 mark), TxnTrans (transcription transition, with H3K36me3, H3K27ac and H3K4me1 marks), TxnElong (transcription elongation, with H3K36me3 mark), WkTxn (weak transcription, with low H3K36me3 mark), H3K9me3 (H3K9me3-repressed heterochromatin), H3K27me3 (H3K27me3-repressed heterochromatin) and Het;Low;Sign (low signal heterochromatin, with the absence of all the six histone marks). Next, this model was used to assign the chromatin states in the remaining primary B-cell tumors, namely 5 MCL, 7 CLL, 5 MM, and the 2 DLBCL cell lines. In the case of ALL, I downloaded H3K27ac ChIP-seq data (generated with the ChIP-grade ab4729 from Abcam) from the NALM6 ALL cell line [309]. I followed the Blueprint pipeline to find H3K27ac peaks (http://dcc.blueprint-epigenome.eu/#/md/chip_seq_grch37). Briefly, I checked data quality and peaks using *fastqc* (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/) and *igv* genome browser [310]. Next, I used bwa version 0.7.17 to align reads to GRCh37d5 reference genome with parameters -q 15, -t 8. BAM files were sorted and duplicates were marked using PICARD tools 2.20.2 (http://broadinstitute.github.io/picard/). Then, low quality and duplicate reads were removed using SAMTOOLS v1.9 with parameters -b -q 15 - @7. I next modelled fragment sizes using PhantomPeakQualTools R script and used it for

peak calling using MACS2 version 2.1.2 with parameters –nomodel -q 0.01 and –extsize 'with previously estimated fragment size'. To define regulatory regions in MCL, CLL, MM and DLBCL, I used the CHMM genome segmentation. Particularly, I used chromatin states containing H3K27ac, namely ActProm, StrEnh1, StrEnh2 and TxnTrans chromatin states. For ALL, regulatory regions were defined simply as regions showing H3K27ac peaks. These active regulatory regions were not merged but used in a disease-specific manner in this thesis. To calculate CHMM enrichments of CpGs sets, I used the CpGs present in the 450k Illumina DNA methylation array as a background. To select genes associated with regulatory regions, I obtained gene annotation for all CpGs within regulatory regions using the *lluminaHumanMethylation450kanno.lmn12.hg19* R package.

2.7. Gene Ontology Analysis.

Gene ontology analyses were performed using the "gometh" function within the *missMethyl* R package available at Bioconductor, which considers the differing number of probes per gene present on the 450k array.

2.8. Tumor specific DNA methylation signatures.

I performed Truncated Principal Component Analysis (PCA) using *irlba* package available at CRAN. Next, to find specific DNA methylation signatures in each B cell tumor, I filtered out all CpGs showing extensive modulation in B-cell differentiation (i.e. B-cell related changes)[50]. Afterwards, I used the *limma* package to perform pair-wise comparisons between each B-cell tumor entity. For each B-cell neoplasia as compared to other B-cell tumors, we retained CpGs that showed at least ≥0.25 methylation difference and FDR<0.05 in the same direction in all comparisons. I next classified the identified CpGs as hyper- or hypomethylated considering the methylation status of normal B cells.

2.9. Transcription factor binding analysis.

I used the PWMEnrich package available at Bioconductor. I focused on CpGs showing specific hypomethylation in each B-cell tumor entity overlapping with regions showing H3K27ac in primary samples of MCL, CLL or MM, and cell lines in the case of ALL (NALM6) and DLBCL (KARPAS-422 and SUDHL-5) (Fig. 2c). I next extended the DNA sequence 100bps (50bps to each side) for each CpG using *Bsgenome.Hsapiens.UCSC.hg19* annotation package available at Bioconductor. As a background sequences, I used

100,000 random B-cell independent CpGs, since those CpGs represent the starting point for tumor specific DNA methylation changes. We then calculated the frequency of A, T, C and G bases in the background sequences. Next, we obtained the 537 CORE JASPAR 2018 TFs for Homo sapiens and transformed motifs to Position Weight Matrices (PWM) using previously calculated frequencies of each base to account for biases in the 450k array. We then calculated a lognormal background distribution with tiles of 100 bps to finally perform TFs binding predictions. We retrieved enrichments per group of sequences and the frequency of each TF that belongs to the Top 5% enrichment TFs, i.e. how often a TF is among the top 5% enriched TFs in all the interrogated sequences. This is of particular interest since the abundance of TFs may represent a more accurate biological measure rather than considering solely p-values. I considered TF as relevant when being within the top 5% TFs in at least 10% of the sequences, showing an FDR ≤0.025 and consistently expressed in each respective B-cell tumor.

2.10. Construction of the classifier algorithm for B cell tumor subtypes.

DNA methylation data for 1,345 samples of B-cell neoplasms was used to build a two-step classifier for the classification of the 5 main B-cell tumor entities (first step) followed by the classification B-cell tumor subtypes (second step, out of the 1,345, 1,013 samples with subtype diagnosis were available). The DNA methylation values of 452,679 CpGs were used, including B-cell related and B-cell independent CpGs [50]. Specifically, the second step consisted of four independent predictors, each one discriminating the subtypes within one specific entity (ALL, MCL, CLL or DLBCL). The five predictors are highlighted as 1, 2, 3 ,4 and 5 in **Results, section 2.2.5, Fig. 21**, depicts the classifier algorithm (made out of the 5 predictors) in conjunction with the heatmap of the final set of CpGs. Finally, **Results, section 2.2.5, Fig. 21b**, shows the number of training and validation samples per entity/subtype (class). Of note, to build the classifier CpGs only present in both methylation array platforms were used (450k and EPIC arrays), so that it can be applied to any sample regardless of the platform. CpGs with minimal variation (interquartile range below 0.07) were removed in the training series of each one of the five predictors.

The following strategy was used to build the predictor for the main B-cell tumor entities as well as for ALL, MCL and DLBCL tumor subtypes (predictors 1, 2, 3 and 5). In the case of CLL, we used another strategy, which is described thereafter:

1) For every class k,
    i. Rank the CpGs according to the Mann–Whitney U test p-value resulting from the comparison of samples of class k against the samples of all other classes.
    ii. Define the signature of class k as the mean of the methylation values of the top M_k CpGs (or one minus the value for hypomethylated CpGs in class k). In case of ties in the p-value ranking, prioritize the CpG with higher mean DNA methylation change.

2) Train a support vector machine model with the signatures of the k classes, using a linear kernel and optimizing the cost C by cross-validation. In the case of only two classes (such as MCL or DLBCL, e.g. C1 vs C2, and ABC vs GCB subtypes), the two signatures are redundant and only one is retained.

The number of CpGs included in the signature of each class in 1) ii, vector M = {M_{ALL}, …, M_{GCB}}, was chosen by 10-fold stratified cross-validation. Specifically, the above algorithm was repeated at each fold where all combinations of possible M_k values were tested and the values that maximized the balanced accuracy were selected. The tested values ranged from 1 to a different quantity depending on the predictor (20 for the main entities, 30 for the ALL subtypes, 20 for MCL, 20 for CLL and 20 for DLBCL). **Results, section 2.2.5, Fig. 21b**, shows, for each predictor, the balanced accuracy and sensitivities of the best performing vector M for each predictor.

For the classification of the three CLL subtypes (m-CLL, i-CLL, n-CLL), another scenario was present, since there already exists a 5-CpG classifier extensively validated that robustly distinguishes these CLL subtypes [254,257,258]. The problem with the published classifier is that it uses a CpG not present in the EPIC array platform (cg09637172), so we trained a new predictor with CpGs present in both platforms. In this case, the strategy that was used to build the predictor was the following:

1) Select the 50 CpGs with the lowest Mann–Whitney U test p-value for each pairwise comparison between the three subtypes.
2) Apply the SVM-RFE algorithm [311] to the subset of CpGs selected in step 1.

3) Train a support vector machine model with the top M_{CLL} CpGs of step 2, cost C, and a linear kernel.

A similar cross-validation strategy as the previous algorithm was used to optimize the M_{CLL} and C parameters. The tested values were M_{CLL} = {1, 2, …, 20} CpGs and C = $10^{\{-3, -2, ..., 3\}}$ cost. **Results section 2.2.5, Fig. 21b**, shows the balanced accuracy and sensitivities of the best performing cost for each number of CpGs.

Finally, we used two strategies to estimate the accuracy of the five predictors: (1) with nested cross-validation in the training series and (2) with a validation series. For the training series, a 10-fold stratified cross-validation was used, where the optimization of the M and C parameters was independently performed at each fold using an inner stratified cross-validation step. For the validation series, we used the following data:
For ALL , 183 samples already included in the Study 2 [208] but not used to construct any classifier nor in any of the other analyses of the manuscript were used. Additionally, I downloaded the following DNA methylation data: GSE76585 [299] and GSE69229 [300]. For MCL validations, I used 58 non-overlapping MCL samples out of the 70 MCL samples generated with the EPIC DNA methylation arrays. For CLL validation, II collected 450k methylation data for 109 CLL samples from a previous study [187] (EGAD00010000871), and 145 CLL with 450k data and 126 CLL with EPIC data kindly provided by Dr. Thorsten Zenz and partially deposited in [312] (EGAD00010000948). Finally, for DLBCL validation 12 DNA profiles with EPIC arrays was generated (EGAS00001004640).
To more accurately represent indetermination in newly obtained samples, not all cross-validated training samples nor validation samples were assigned to an entity/subtype. Specifically, the *svm* function of the e1071 R package was used to obtain a probability for each entity/subtype in each one of the samples. Next, samples where the maximum probability was below 50% or multiple entities/subtypes (including the true entity) had a probability above 35% were considered unclassified. **Results, section 2.2.5, Fig. 21b**, shows for each predictor and class, the two series estimated percentage of unclassified samples and the estimated sensitivity in the classified ones.

In the case of MCL, there was not a gold-standard for the cell-of-origin classification in C1 and C2 MCL subtypes (**Study 1**). The MCL training samples included in this Study 1 were classified into C1 and C2 by performing a principal component

analysis (PCA) of those MCL samples together with normal samples spanning the entire B-cell lineage [295]. In order to classify the validation MCL samples into C1 or C2 (**Study 2**), a strategy that mirrored the original was used. Specifically, a PCA space was first created using all of the unfiltered methylation information in the training samples, and identified that the two first components contained most of the information related to the subtype. Then, these two components were used to fit a quadratic discriminant analysis (QDA) model that distinguished the two cell-of-origin subtypes in this new space. Finally, the validation samples were projected into the training PCA space and the fitted QDA model was applied to them. Only samples with either C1 or C2 probability ≥85% were assigned to one of the subtypes. This strategy allowed to define a cell-of-origin subtype for the validation series using the methylation information as a whole.

2.11. Inter-patient DNA methylation heterogeneity.

To analyze the variability of DNA methylation data among patients, I identify CpGs with differential methylation in each patient individually. To do this, I compared data from each single patient with the mean in HPC samples, and considered a DNA methylation change for a given CpG when a difference ≥0.25 was reached. Next, to define all the DNA methylation changes occurring in patients diagnosed with a specific B-cell tumor subtype, I selected all the CpGs meeting these two criteria; 1) in at least one patient of a specific B cell tumor subtype showing an absolute methylation difference ≥0.25 as compared to HPC, and 2) all other patients in the B cell tumor subtype show the same trend, i.e. towards hypomethylation or hypermethylation.

2.12. Construction of the epiCMIT score.

To create the epiCMIT score, I first collected ChIP-seq data for H3Kme1, H3Kme3, H3K27ac, H3K36me3, H3K9me3 and H3K27me3 from the Blueprint consortium [3] in 15 normal B cells, 5 MCLs, 7 CLLs, 4 MMs (**Results, section 3.2.2, Fig. 25a**). They were used to generate a B-cell specific chromatin state model with 12 emission states using the chromHMM software[45] in the 15 normal B cells, corrected for their corresponding input. Next, this model was used to assign chromatin states in the remaining primary B cell tumors, namely 5 MCL, 7 CLL and 5 MM. The 12 states are represented in the second step of epiCMIT construction figure at **Results, section 3.2.2, Fig 25a**. Next, I selected all CpGs from 450k array of our entire DNA methylation matrix of normal and neoplastic B-cells located in inactive regions, particularly in poised promoters (PoisProm, with

H3K27me3, H3K4me1 and H3M4me3 marks), in H3K27me3 regions, in H3K9me3 regions, and in low signal heterochromatin (Het;LowSign, absence of any of the six marks analyzed). I divided this set of CpGs into two distinct sets, CpGs located in H3K27me3-repressed regions or PoisProm, and CpGs located in H3K9me3-repressed regions or Het;Low;Sign heterochromatin. I next performed differential DNA methylation analysis between normal B-cells with the lowest and the high proliferative histories, namely HPC and bmPC (**Results, section 3.2.2, step 3 of Fig. 25a**) and I retained CpGs gaining DNA methylation in bmPC in H3K27me3 regions or PoisProm, and CpGs losing DNA methylation in bmPC in H3K9me3 and Het;Low;Sign heterochromatin. In addition, I imposed two key restrictions to these two sets of CpGs. First, CpGs gaining methylation during cell division must showed a very low DNA methylation level (<=0.1) in lowly divided cells, i.e. HPC, and CpGs losing DNA methylation during cell division must showed high DNA methylation levels (>=0.9) in the lowly divided HPC cells. This ensured to have a clear baseline of DNA methylation changes occurring during cell division in B cells. Second, I retained only those CpGs showing extensive modulation between the lowly divided HPC and highly divided bmPC cells. This second condition was imposed to maximize the differences in the DNA methylation values upon cell division. With all these restrictions, I ended with 184 CpGs unmethylated in HPC, gaining at least 0.5 methylation in bmPC, and falling in H3K27me3 or PoisProm regions (**Results, section 3.2.2, step 3 Fig. 25a**). These 184 CpGs constitute the CpGs used to construct the epiCMIT-hyper mitotic score. Conversely, I retained 1,164 CpGs methylated in HPC, losing at least 0.5 methylation in bmPC and falling in H3K9me3 or Het;Low;Sign regions. These 1,164 CpGs represent the CpGs used to construct the epiCMIT-hypo mitotic score. To construct the epiCMIT-hyper and epiCMIT-hypo mitotic clock scores, I calculated the mean of all respective set of CpGs per sample (i.e, mean of the 184 CpGs for epiCMIT-hyper and of the 1,164 CpGs for epiCMIT-hypo). In the case of epiMIT-hypo I inverted the score to have the same scale (**Results, section 3.2.2, step 4 of Fig. 25a**). Particularly, I calculated each score with the following formulas:

$$\text{epiCMIT} - \text{hyper} = \frac{\sum_1^{184} \text{DNA methyaltion epiCMIT} - \text{hyper CpGs}}{184}$$

$$\text{epiCMIT} - \text{hypo} = 1 - \frac{\sum_1^{1164} \text{DNA methylation epiCMIT} - \text{hypo CpGs}}{1164}$$

Finally, to construct the epiCMIT score, I evaluated per sample both epiCMIT-hyper and epiCMIT-hypo scores, and selected the higher of the two (**Results, section 3.2.2, step 5 Fig. 25a**):

$$\text{epiCMIT} = \max\{\text{epiCMIT} - \text{hyper}, \text{epiCMIT} - \text{hypo}\} \text{ per sample}$$

As the epiCMIT score was built with 450k array data, there are 84 CpGs that are not present in the currently available EPIC array from Illumina (10 epiCMIT-hyper and 74 epiCMIT-hypo). Nonetheless, I observed high correlations between epiCMIT scores calculated with all the original CpGs with those exclusively present in both 450K and EPIC arrays).

2.13. Determination of published mitotic clocks and the Horvath aging clock

To determine epiTOC [313], MiAge [314], CIMP [315], PMDsoloWCGW [316] and Horvath [317] DNA methylation clocks I used their underlying CpGs overlapping with those present in our whole DNA methylation matrix. Specifically, the number of CpGs from these clocks present in my DNA methylation matrix after all normalization and quality control procedures were the following: 377 out of the 385 epiTOC CpGs, 261 out of the 268 MiAge CpGs, 88 out of the 89 pan-cancer CIMP CpGs [315], 5,595 out of the 6,214 PMDsoloWCGW CpGs and 351 out of the 353 Horvath CpGs. For the epiTOC and MiAge scores, we calculated them as previously indicated [313,314]. For CIMP score, I used a set of previously proposed CpGs [315] and used the same strategy than the epiCMIT-hyper (explained in the previous section). In the case of the PMDsoloWCGW mitotic clocks, I applied the same strategy previously used for the epiCMIT-hypo score (explained in the previous section). Finally, I used Horvath to predict age using R as previously reported [317].

2.14. Somatic mutations and mutational signature analysis in CLL.

The somatic mutations found in the CLL samples were previously reported [245]. I considered driver alterations those reported as such in two landmark genomic publications [245,246]. In addition to this, a new recurrent driver mutation has been recently added to CLL, namely the U1 spliceosomal RNA [318]. The U1 mutational status for 318 CLL patients already published. For the remaining 172 CLL patients from our analyses, the U1 mutational status was evaluated using rhAmp SNP Assay (Integrated DNA

Technology) as previously described [318]. Next, the mutational signature analysis was performed following a similar framework as the one described by Alexandrov and coworkers [319]. Briefly, *de novo* signature extraction was performed using a hierarchical Dirichlet process (*hdp* R package, https://github.com/nicolaroberts/hdp), and extracted signatures were matched to the recently described list of mutational signatures [319] based on cosine similarity and the biological knowledge of each mutational process. Signatures identified through this approach were signature SBS1, SBS5, SBS8, SBS9, SBS17b, and SBS18. Finally, the contribution of each of the previously identified signatures for each sample was measured using a fitting approach (*MutationalPatterns* R package). To avoid signature bleeding between samples, we iteratively removed one signature after another and the least contributing signature was censored if removal reduced the cosine similarity <0.005, with the exception of signature SBS1 and SBS5, which were always included based on their reported presence in all normal and tumor samples.

2.15. Gene Set Enrichments Analysis (GSEA) in CLL.

In order to perform GSEA analysis in CLLs with different epiCMIT score, we took CLLs samples separated by their cellular origin [254,258] (epigenetic groups) above 85% percentile and below 15% percentile of epiCMIT. CLLs belonging to the i-CLL were excluded due to smaller sample size. We performed differential gene expression analysis using *limma*. We then used *fgsea* package [320] to perform GSEA analyses using log FC as summary statistic to rank genes. We downloaded 5,501 curated (C2) gene signatures from Molecular Signatures Database v7.0 https://www.gsea-msigdb.org/gsea/index.jsp. We performed *GSEA* analysis with all these pathways filtering those with less than 5 genes and more than 5,000. We used 10,000 permutations to obtain p-values. We next selected 118 gene expression signatures related to cell proliferation and MYC in an unbiased way. These 118 expression signatures were found in R by regular expression matching with *grep* R function using the following expression: grep("CELL_CYCLE|prolifer|divi|mitotic|_CYCLING|M_PHASE|_MYC_",names(gene_expression_signatures_names)).

2.16. Whole-genome/exome sequencing in MCL.

WES data of 21 tumor/normal pairs and WGS data of 4 tumor/normal pairs was described in a previous study (European Genome-Phenome Archive, EGA accession number EGAS00001000510) [228]. Additional WGS were performed on 57 pairs using a

PCR-free library preparation protocol (TruSeq DNA PCR Free), sequenced in an Illumina HiSeq X Ten (2x150bp), and deposited at EGA (EGAS00001004165). Tumor and normal WGS raw reads were mapped to the human reference genome (GRCh37) using the BWA-mem algorithm (v0.7.15). BAM files were sorted, indexed, and optical and PCR duplicates were flagged using biobambam2 (v2.0.65, https://gitlab.com/german.tischler/biobambam2). Quality metrics were extracted using Picard (https://broadinstitute.github.io/picard, v2.10.2). All algorithms were executed with default options.

Somatic single nucleotide variants (SNVs) and indels were called using Sidrón, as previously described [245]. Kataegis was defined by the presence of 5 or more SNVs with a mutational inter-distance below 1000 bp. Short insertions/deletions (indels) were detected integrating Sidrón and Pindel (cgpPindel v2.2.3, with default arguments) algorithms. All coding indels were manually reviewed on Integrative Genomic Viewer (IGV) http://software.broadinstitute.org/software/igv/. Mutational burden was calculated based on the number of non-N bases in the reference genome used. Germline variant calling was performed following Genome-Analysis Toolkit (GATK) best practices. Briefly, germline SNPs and indels from WGS and WES data were called using HaplotypeCaller (GATK, v4.0.2.0), and filtering of low-quality variants was performed using VariantRecalibrator and ApplyRecalibration [321,322]. SNVs were functionally predicted using CADD, Mutation Assessor, PolyPhen and SIFT (dbNSFP v3.5a).

Detection of copy number alterations (CNA) was analyzed using Battenberg algorithm (cgpBattenberg, v3.2.2) [323] in the 61 samples analyzed by WGS. CNA were also analyzed using the Affymetrix Genome-wide Human SNP Array 6.0 (n=78) and Cytoscan (n=4) (Thermo Fisher Scientific) in the 70 MCL cases as previously described [213]. CNA from SNP-arrays were used to verify and complement the results obtained from Battenberg. CNA smaller than 100 Kb and physiological CNAs at immunoglobulin (IG) and T-cell receptor (TCR) loci were filtered out. To allow the integration of the 70 cases, CNA present in <15% of the tumor cells called by Battenberg (cases with WGS) were removed as a result of their low sensitivity of the SNP-array (cases with WES and SNP-array). CNA located at terminal regions were obtained from SNP-arrays due to the inherent difficulty to study these regions from WGS data. Tumor purity and ploidy were extracted from Battenberg in the 61 cases with WGS. Predicted ploidies were manually compared with

SNP-array and FISH data, and adjusted when necessary. The tumor cell content measured by flow cytometry analyses was used in the 21 cases analyzed by WES.

Detection of structural variants Structural variants (SV) were extracted from WGS data using SMuFin [324] and LUMPY [325] (v0.2.13), and were visually inspected on IGV. The integration of the two algorithms was performed using in-house scripts merging SV with breakpoints closer than 150 base pairs. SV were annotated using ANNOVAR19 at breakpoint positions and classified as "disrupted gene" if the breakpoint was located inside the gene. SV detected by a single pipeline that were covered by less than 7 paired-end reads and/or less than 3 split-reads, and did not overlap with CNA were removed. Small SV (<500 bp) not overlapping with CNA were also removed. Physiological SV at the IG and TCR loci were removed. Chromothriptic events were defined by the presence of seven or more oscillating changes between two different copy number states and/or the presence of more than six breakpoints occurring in a single chromosome. Chromoplexic events, defined by the ligation of fragments from three or more chromosomes by translocations, were inferred from WGS data. Similar to chromoplexia, templated insertions were considered when, in addition to ligation of 3 or more chromosome fragments, concomitant small duplications were found in all breakpoints. Breakage-fusion bridge cycles were inferred from WGS and SNP-array data based on the identification of patterns of focal copy number increases of chromosomal segments as well as fold-back inversions, together with the presence of telomeric deletions. Telomere length was determined from WGS data using *qMotif* (https://sourceforge.net/p/adamajava/wiki/qMotif/, v1.2) in the 61 tumor and normal samples.

Mutational signatures were analyzed for SNV and were first classified in the 96 classes according to their 5' and 3' base. Next, they were extracted *de novo* using the *MutationalPatterns* R package (v1.4.3) using a non-negative matrix factorization (NMF). The extracted signatures were compared to the signatures described in COSMIC (https://cancer.sanger.ac.uk/cosmic/signatures/SBS), and those with a cosine similarity >0.7 were considered. To measure the contribution of each signature in each tumor we used a fitting approach (*MutationalPatterns*) and iteratively removed the less contributing signature if removal of the signature decreases the cosine similarity between the original and reconstructed 96-profile <0.01, as previously described [319].

Finally, to identify driver alterations were identified using the integration of the data with an in-house workflow with several steps: i) mutated driver genes were defined considering type of mutation, gene size, coverage, and local density of mutations as previously described[245]; ii) GISTIC algorithm [326] (2.0.23) was used to identify driver CNA from the 61 cases from WGS and the 21 cases with WES; iii) Driver altered genes and CNA were considered when they were recurrently altered in 3 or more cases, and statistically significant (false discovery rate 18%). Genes mutated at lower frequency but carrying known driver or truncating mutations were also considered as drivers. CNA not significant by GISTIC but with high frequency (>20%) or with focal homozygous deletion or amplification and previously published in more than one MCL series were also considered as drivers. Recurrently mutated driver genes in the present series but that were not expressed in any normal B cell or in MCL were flagged as "not expressed" and not considered as driver alterations in subsequent analyses. These genes corresponded to large genes located in late-replicating regions.

2.17. epiCMIT clinical associations.

I performed univariate analyses of the epiCMIT score for relapse-free survival (RFS), overall survival (OS), and OS after relapse in ALL; OS and Time to First Treatment (TTT) for CLL and OS with different clinical MCL patients depending on their clinical management at diagnosis using Kaplan Meyer curves with maxstat statistics to define groups with high and low epiCMIT. The hazard ratios and their corresponding p-values are shown when epiCMIT categorization was performed. Finally, epiCMIT was assessed in OS together with ABC and GCB DLBCL transcriptomic subtypes [194]. Furthermore, the epiCMIT prognostic value was assessed in presence of other well-established prognostic factors in all diseases using multivariate Cox regression models. In ALL, this includes including Hyperdiploid ALLs (HeH), Others (including non-recurrent, undefined, <45chr,>67chr and iAMP21), t(1;19), t(12,21), dic(9;20), t(9;22) and 11q23/MLL. In MCL, I performed the multivariate Cox regression model for OS with epiCMIT together with epigenetic groups C1 and C2 and with age. Then, I performed multivariate Cox regression models with CNA, BFB and epiCMIT. Finally, in CLL I performed multivariate Cox regression models for TTT and OS with epiCMIT together with age at sampling, epigenetic groups and the total number of driver alterations considering mutations in both studies [245,246]. Furthermore, epiCMIT prognostic value was assessed in 15 tumor TCGA samples (n=3,995) for OS together with all the other mitotic clocks. All mitotic

clocks were scaled when comparing the prognostic among them. Finally, driver alterations strongly associated with epiCMIT levels were analyzed in the context of other driver alterations including the driver alterations not associated with epiCMIT in the case of CLL, and CNA, and epiCMIT in the case of MCL. Clinical analyses were done using *survival* and *survminer* R packages.

2.18. Finding genetic driver alterations related with increased epiCMIT.

In the case of CLL, I used previously published data [245,318] as well as newly generated for the mutational status of U1 for 172 CLL cases. In the case of MCL, 70 newly generated samples with concurrent WGS /WES and DNA methylation data were used [229]. I analyzed the association of each genetic alteration with epiCMIT in the whole cohort of CLL and MCL patients separately, and in each disease subtypes. In the whole cohort analyses, I modelled epiCMIT score with each genetic alteration using linear regression correcting by each disease subtype. On the other hand, I used t-tests between the levels of epiCMIT in mutated and unmutated patients for each genetic alteration within each disease subtype. I derived point estimates and 95% confident intervals in both types of analyses (p-values were corrected using FDR). Finally, genetic alterations significantly associated with epiCMIT were grouped by pathways implicated in the pathogenesis of CLL and MCL.

2.19. Data availability.

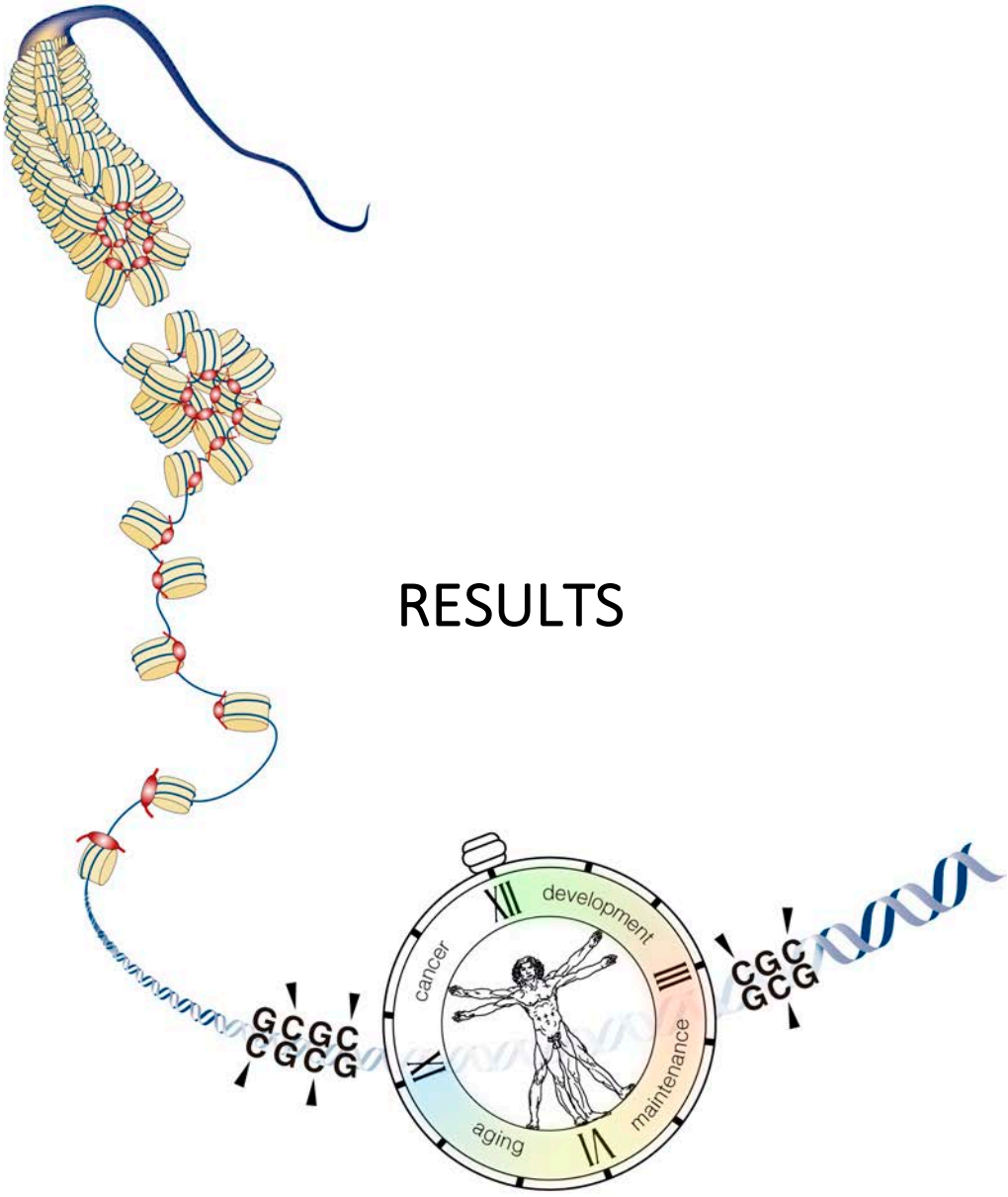
DNA methylation and gene expression data that support the findings of this doctoral thesis have been deposited at the European Genome-phenome Archive (EGA) under accession number EGAS00001004640. Previously published DNA methylation data re-analyzed can be found under accession codes: B cells, EGAS00001001196; ALL, GSE16368, GSE47051, GSE7658515, GSE6922916; MCL, EGAS00001001637, EGAS00001004165; CLL, EGAD00010000871, EGAD00010000948; MM, EGAS00001000841; In vitro B-cell differentiation model of naïve B cells from human primary samples, GSE72498. Normalized DNA methylation matrices used for all the analyses in this doctoral thesis are available at: http://resources.idibaps.org/paper/the-proliferative-history-shapes-the-DNA-methylome-of-B-cell-tumors-and-predicts-clinical-outcome. Published gene expression datasets can be found under the accession codes: B cells, EGAS00001001197; ALL, GSE47051; MCL, GSE36000; CLL, EGAS00000000092, EGAD00010000254; MM, GSE19784; In vitro B-cell differentiation model of naïve B cells from human primary samples, GSE72498. ChIP-seq datasets that

were re-analyzed here can be found under the accession codes: GSE109377 (NALM6 ALL cell line, n=1) and EGAS00001000326 (15 normal B cells donors, and 5 MCL, 7 CLL and 4 MM patients) available from Blueprint https://www.blueprint-epigenome.eu/.

2.20. Code availability.

The source code for the DNA methylation classifier of B-cell tumors entities and subtypes and for the calculation of the epiCMIT mitotic clock can be found at https://github.com/Duran-FerrerM/Pan-B-cell-methylome.

RESULTS



Study 1. The whole DNA methylome of MCL in the light of normal B-cell differentiation.

The results of this study have been published (**appendix, manuscripts 2 and 3**).

*****Queirós AC, *****Beekman R, *****Vilarrasa-Blasi R, **Duran-Ferrer M**, Clot G, Merkel A, Raineri E, Russiñol N, Castellano G, Beà S, Navarro A, Kulis M, Verdaguer-Dot N, Jares P, Enjuanes A, Calasanz MJ, Bergmann A, Vater I, Salaverría I, van de Werken HJG, Wilson WH, Datta A, Flicek P, Royo R, Martens J, Giné E, Lopez-Guillermo A, Stunnenberg HG, Klapper W, Pott C, Heath S, Gut IG, Siebert R, Campo E, **#**Martín-Subero JI. Decoding the DNA Methylome of Mantle Cell Lymphoma in the Light of the Entire B Cell Lineage. Cancer Cell. 2016 Nov 14;30(5):806-821. doi:10.1016/j.ccell.2016.09.014. PMID: 27846393; PMCID: PMC5805090. (***** Contributed equally, **#** Corresponding author).

Duran-Ferrer M, Beekman R, **#**Martín-Subero JI. *In silico* deconvolution and purification of cancer epigenomes. Oncoscience. 2017 Apr 14;4(3-4):25-26. doi: 10.18632/oncoscience.346. PMID: 28540331; PMCID: PMC5441471. (**#** Corresponding author).

1.1. Introduction.

In spite of the publication of several articles focusing on the MCL epigenome, they all focused on the analysis of promoter regions, and the whole DNA methylome of the disease remains poorly characterized [230–233]. Thus, to obtain new insights into the MCL epigenome, a whole-genome DNA methylome analysis together with chromatin marks was performed in this thesis. As normal B-cell differentiation already entails a huge modulation of the DNA methylome [50,187], the MCL DNA methylation data was analyzed considering DNA methylation changes happening during all B-cell differentiation process. This strategy has the advantage of clearly separating those MCL methylation changes related to normal B-cell maturation and those that take place only in the context of lymphomagenesis, and thus are MCL-specific. This analytical approach stands in sharp contrast to what has been routinely done in previous DNA methylation studies [327], where tumor-specific changes are found by comparing tumor cells with their assumed normal counterpart, or by comparing longitudinal tumor samples. Applying this strategy, it was identified that a great fraction of DNA methylation changes happening in MCL relate to imprints of normal B-cell differentiation, which in turn led to the identification of two MCL subgroups with different cellular origin and clinico-biological features. In addition, the integration of multiple epigenetic layers including DNA methylation, histone marks and the 3D structure of the genome unrevealed a previously unrecognized epigenetic mechanism leading to *de novo SOX11* expression seen on some subsets of MCLs. Finally, the inter-patient heterogeneity among MCL patients was characterized and related to the clinical outcome of the patients.

1.2. Results.

1.2.1. *In-silico* deconvolution and purification of MCL DNA methylome.

DNA methylation profiles using the Infinium HumanMethylation450k BeadChip were generated in 82 MCL patients. As normal controls, 67 samples from ten different cell subpopulations spanning the entire B-cell lineage were used [50]. Before analyzing the MCL DNA methylome, two potential confounding variables that could affect DNA methylation analyses were considered, including the biological origin of the samples (lymph node or peripheral blood) and the tumor cell content. No consistent differential DNA methylation patterns between lymph node and peripheral blood samples could be identified. However, despite the generally high tumor cell content of the selected MCL samples (median, 89%; range, 39%–100%), some MCL samples with the lowest tumor cell content showed a distinct DNA methylation profile, suggesting that tumor purity could affect the bulk DNA methylome. To avoid the removal of those samples with the lowest tumor cell content, I developed a strategy to deconvolute the DNA methylation signal of mixed subpopulations and to isolate *in silico* the DNA methylation levels of the tumor cells (**Fig. 11**, **Methods Study 1, section 1.2**). To that end, I used a previously proposed statistical framework [301] to estimate the proportion of six different blood cell types [287] in the MCL samples (**Fig. 11a**). The normal B-cell fraction in MCL patients is negligible [305] and therefore the blood $CD19^{+}$ cell signature present in MCL samples was taken as a measure of tumor cell content (**Fig. 11b**). I validated this approach by comparing the tumor cell content predicted through MCL DNA methylation data with the tumor cell content of 32 blood MCL samples measured by flow cytometry (R=0.95) (**Fig. 11c**). Finally, I removed the DNA methylation coming from microenvironmental cells in MCL samples and retained the *in-silico* purified DNA methylation from MCL cells (**Fig. 11d, Methods Study 1, section 1.2**). Using a previously applied statistical approach [286], I observed a reduction of the influence of varying MCL tumor cell content in MCL DNA methylation variability (**Fig. 11e**), and thus the *in-silico* purified DNA methylation values were used for all downstream analyses.

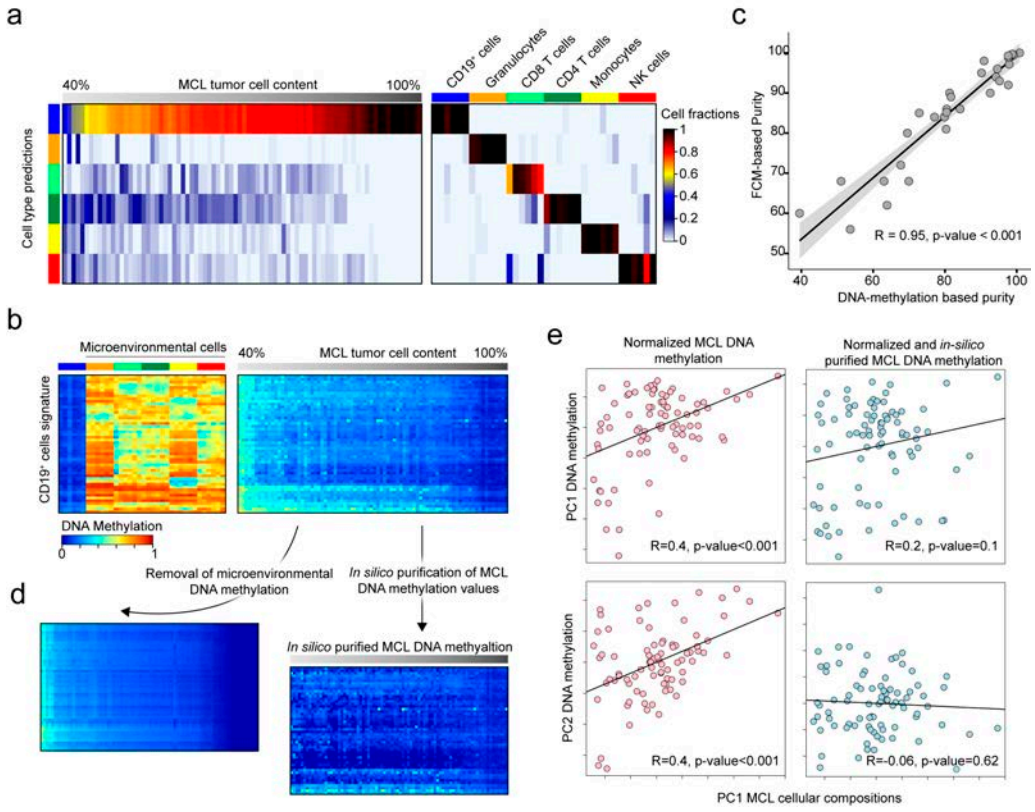


Fig. 11 | MCL tumor cell content estimation and *in-silico* MCL DNA methylation purification. a, MCL tumor cell content and microenvironmental cell predictions. **b,** B-cell signature used to deconvolute and estimate the MCL tumor cell content. **c,** Validation of the approach with flow cytometry data (FCM). **d,** *In-silico* MCL DNA methylation purification and removal of DNA methylation contamination from the microenvironmental cells. **e,** Reduction of DNA methylation variability due to different MCL tumor cell content in primary samples as a statistical proof of principle for the *in-silico* purification approach.

1.2.2. Identification of two epigenetic MCL subgroups.

An unsupervised principal component analysis (PCA) was initially performed with MCL samples and normal B-cell spanning the entire maturation process (**Fig. 12a**). The two first components separated normal B cells according to their maturation stage, clearly separating germinal-center-inexperienced B cells (uncommitted precursors, pre-B cells, and naïve B cells) from germinal-center-experienced B cells (germinal-center B cells, memory B cells, and bone-marrow plasma cells). Component 1 showed that all MCLs are globally more similar to germinal-center-experienced B cells and thus to antigen

experienced cells. In contrast, component 2 divided MCL samples into two subgroups which were named cluster 1 (C1) (n = 62) and cluster 2 (C2) (n = 20). C1 MCL showed a DNA methylation imprints from germinal-center-inexperienced B cells whereas C2 MCL showed imprints from germinal-center-experienced B cells. These subgroups showed distinct clinico-biological differences including IGHV mutation levels, *SOX11* expression, number of copy number alterations, nodal presentation, and requirement of treatment at diagnosis (**Fig. 12b**). Furthermore, C1 cases showed a significantly worse overall survival than C2 cases at the time of diagnosis (p = 0.026) (**Fig. 12c**). Collectively, these results indicate the presence of two distinct clinico-biological subgroups of MCL, which can be distinguished by distinct DNA methylation profiles.

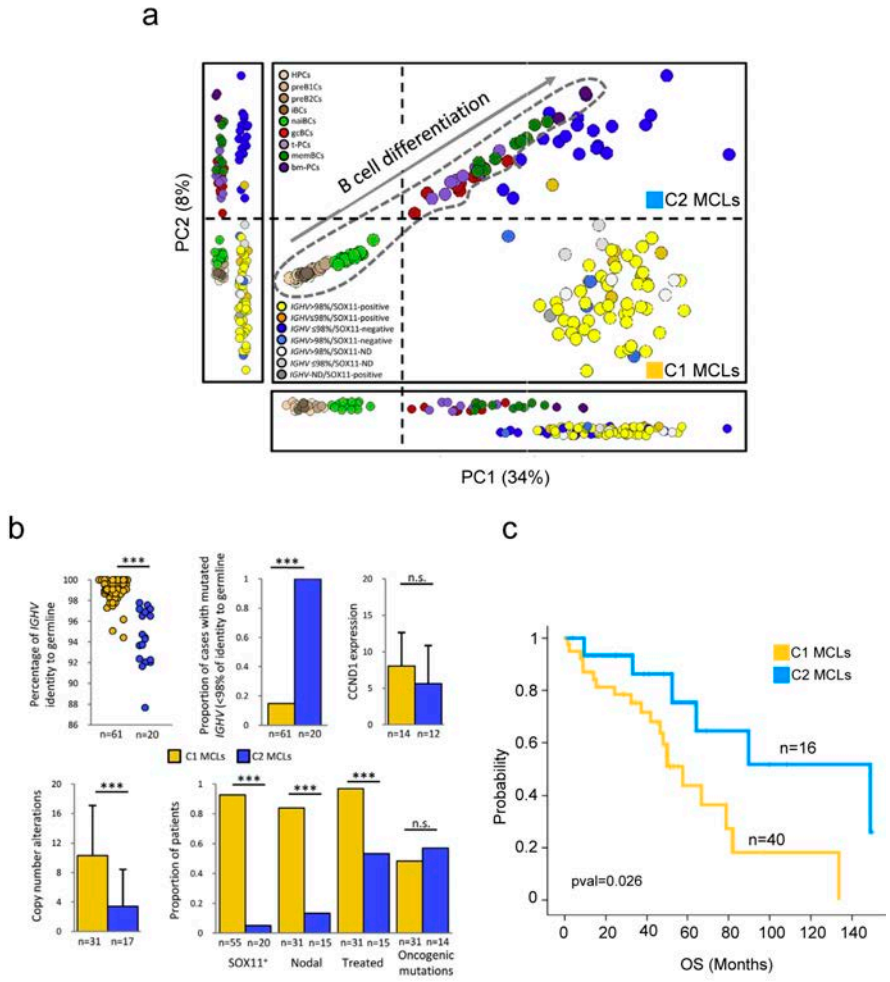


Fig. 12 | See caption on next page.

Fig. 12 | Identification of two MCL subtypes with different clinico-biological features (see figure on previous page). a, PCA analysis with normal B cell subpopulation and MCL primary samples. Two MCL clusters with imprints from pre- and post-germinal center B cells were identified and named C1 and C2 MCL, respectively. **b,** Different clinico-biological features associated with the MCL clusters identified. **c,** Differential overall survival (OS) for C1 and C2 MCL clusters.

1.2.3. Deep characterization of MCL DNA methylome using WGBS.

Next, the entire DNA methylome of two representative MCL samples with high tumor-cell content, and previously analyzed by 450K microarrays, were sequenced at a single base-pair resolution (one C1 and one C2 MCL, at 48x mean coverage). The WGBS data from these two MCL WGBS were analyzed in the context of normal B-cell differentiation using 12 previously generated WGBS profiles of normal B cells [50]. A global DNA methylation loss in the two MCL subtypes was identified as compared to the normal B cells (**Fig. 13a**). Next, a differential DNA methylation analyses including DMCs and DMRs comparing first each MCL subtype against HPC and then C1 versus C2 MCLs was performed. Determining DMRs increased the detection of regulatory regions compared with detecting DMCs , and therefore all downstream analyses were performed following the DMR strategy (please, be referred to a previous publication [308] for an extensive review of dealing with DMRs versus DMCs). As expected based on the previous 450K array data, a great fraction of DNA methylation changes in MCL were related to normal B-cell maturation. Few MCL-specific DMRs were detected, and the majority of detected DMRs were modulated (B-cell related DMR) or partially modulated (Mixed DMR) during normal B-cell maturation (**Fig. 13b**). Focusing on DMRs between the two MCL subtypes, a cluster of CpGs hypomethylated in the *SOX11*-expressing C1 MCL overlapping with an enhancer region located 624–653 kb's downstream of *SOX11* gene was observed. This hypomethylation cluster correlated with the DNA methylation levels of the *SOX11* promoter. Furthermore, this distal region was looping to the *SOX11* promoter at the 3D level only in the *SOX11*-expressing MCL C1 case (**Fig 13c**). Remarkably, the same scenario was found in *SOX11*-expressing MCL cell lines, whereas not in *SOX11*-negative MCL cell lines or in normal B cells (**appendix, manuscript 2**). These findings suggest that the DNA hypomethylation of this distal region together with activating histone marks and the looping to the *SOX11* promoter is needed for *SOX11* expression in primary MCL patients.

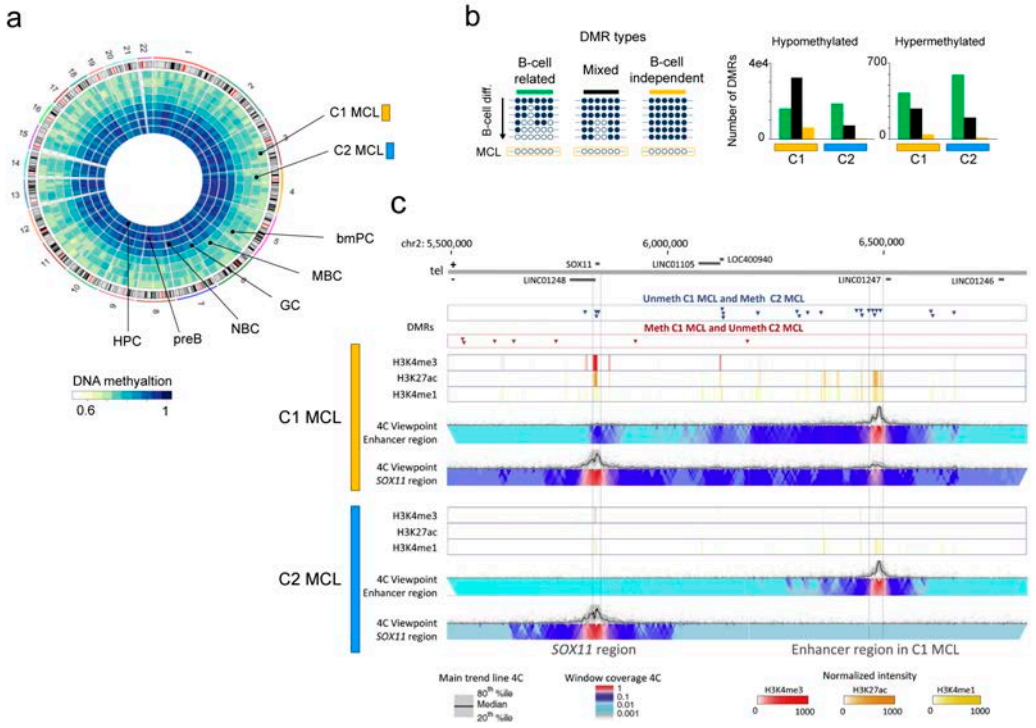


Fig. 13 | WGBS characterization of MCL subtypes. a, Circos plot from WGB showing the overall DNA methylation levels of normal B cells subpopulations from Kulis et.al 2015 and one C1 and C2 MCL subtypes. **b,** Number of different types of DMRs in MCL subtypes compared to HPC. **c,** Identification of an hypomethylated distant enhancer region looping to SOX11 promoter in C1 but not C2 MCL primary patient.

1.2.4. High DNA methylation heterogeneity in MCL.

At the unsupervised level, MCL display two distinct DNA methylation clusters showing imprints from germinal-center inexperienced B cells (C1 MCL) and germinal center-experienced B cells (**Fig. 12a**). Nonetheless, each MCL subgroup showed high intra cluster-variability, suggesting the presence of inter-patient DNA methylation heterogeneity in both MCL subgroups. To further analyze this phenomenon, the total number of DNA methylation changes per MCL patient was computed as compared to HPC, since C1 and C2 presented DNA methylation imprints from different cellular origins and thus likely derive from distinct stages of normal B-cell differentiation. The DNA methylation changes were classified then into four categories, hypo- or hypermethylation, and B-cell related or B-cell independent, depending on whether CpGs were extensively modulated or not during normal B-cell differentiation [50]. MCL patients

showed varying levels of DNA methylation changes, with some samples showing few, while other many DNA methylation changes (**Fig. 14a**). Interestingly, the number of B-cell related changes and B-cell independent changes was highly correlated, suggesting that the overall DNA methylation burden of the tumor in each individual patient may be shaped by a similar underlying phenomenon (**Fig. 14b**). Next, to better characterize all patient-specific DNA methylation changes, the recurrence of each CpG change within each MCL subgroup was computed. This analysis showed that the great majority of DNA methylation changes detected in MCL are present in one or few patients, whereas recurrent DNA methylation changes are in fact relatively rare events (**Fig. 14c, d, left panels**). Furthermore, these distributions were accompanied by different chromatin environments in DNA hypomethylation changes in both C1 and C2 MCL subtypes (**Fig. 14c, d, upper panels**). Less recurrent hypomethylation changes were located at heterochromatic regions while highly recurrent changes at regulatory regions (presence of H3K27ac mark) (**Fig. 14d**). On the contrary, DNA hypermethylation changes were located mainly at H3K27me3 marked regions regardless of their recurrence in MCL samples (**Fig. 14c, d, bottom panels**). Collectively, this data suggest that the majority of DNA methylation changes target non-regulatory regions and that few recurrent hypomethylation changes target regulatory sites with potential impact on MCL pathogenesis, as it is exemplified with the *SOX11* enhancer previously identified.

Fig. 14 | MCL patient-specific DNA methylation changes and their related clinico-biological features (see figure on next page). a, Number of DNA methylation changes per individual in normal B cells and MCL sample subtypes. DNA methylation changes are classified into four categories, as being hyper- or hypomethylated and B-cell related or B-cell independent. **b,** The number of B-cell related and B-cell independent changes is strongly correlated in both MCL subtypes. **c, d** Number of DNA methylation changes shared by patients in C1 and C2 MCL separately. The chromatin state enrichment according to increasing quantiles of recurrence. **e,** Overall survival in C1 and C2 MCL patients divided into low or high number of DNA methylation changes through the maxstat statistic. **f,** Multivariate Cox regression model of the number of DNA methylation changes with previously recognized important prognostic variable in MCL. **g,** Number of DNA methylation changes in C1 and C2 MCL patients and the presence of any of a set of 6 driver alterations. **h,** Heatmaps showing the number of DNA methylation changes in relation to the proliferative signatures in C1 and C2 MCL samples. **i,** Correlation between the number of DNA methylation changes and the average proliferation signature relative to MCL groups.

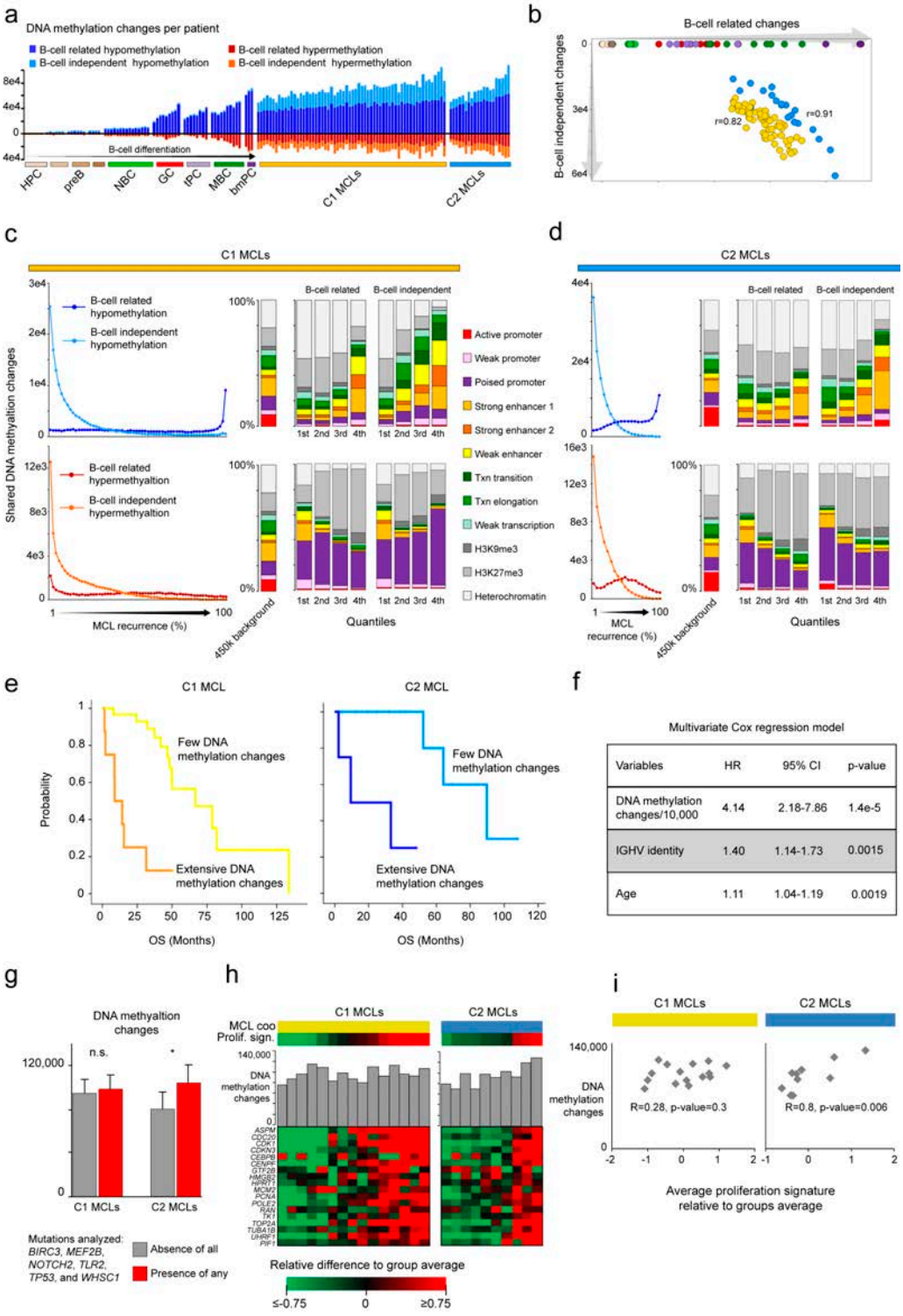


Fig. 14 | See caption on previous page.

1.2.5. DNA methylation heterogeneity and MCL patient clinical outcome.

MCL shows a wide spectrum of clinical behaviors, which in part may be related to their cellular origin and *SOX11* expression [216]. To further identify new biological mechanisms leading to different MCL patient outcomes, the DNA methylation variability observed in MCL patients was associated with their clinical outcome. This analysis clearly revealed that the higher the number of DNA methylation changes, the worse clinical outcome of MCL patients (**Fig 14e**). This association was also significant in a multivariate Cox regression model with previously recognized variables including IGHV status and age [216] (**Fig. 14f**). Furthermore, using available mutational data of a set of 6 genes [228], the number of DNA methylation changes was found to be correlated with the presence of any of a set of 6 genetic alterations (**Fig. 14g**) and increased proliferation in C2 MCLs (**Fig. 14h, i**). Collectively, these data suggest that patients with more DNA methylation changes have a worse clinical outcome and that, in C2 MCLs, this correlates with the acquisition of particular mutations and increased cell proliferation.

1.3 Conclusions.

- Cell lineage-specific DNA methylation profiles can be used to accurately estimate the proportion of neoplastic and microenvironmental cells in MCL samples.

- DNA methylation analysis identifies two MCL clusters, named C1 and C2, which respectively show imprints from germinal center-independent and germinal center-experienced B cells.

- C1 and C2 MCL present distinct clinicobiological features including IGHV mutation levels, SOX11 expression, number of copy number alterations, nodal presentation, requirement of treatment at diagnosis and clinical outcomes.

- MCL patients show varying numbers of DNA methylation changes, with the majority of changes being patient-specific and located in silent regions of the genome.

- The number of DNA methylation changes is associated with poorer overall survival in C1 and C2 MCL subtypes independently from IHGV status and age.

- WGBS combined with chromatin profiling allowed the identification of a distant hypomethylated active enhancer that loops to the *SOX11* gene only in *SOX11*-expressing MCL.

Study 2. Dissecting the DNA methylation variability in B-cell tumors.

The results of this study have been published (**appendix, manuscript 4**).

 # Duran-Ferrer M, Clot G, Nadeu F, Beekman, R, Baumann, T, Nordlund, J, Marincevic-Zuniga, Y, Lönnerholm G, Rivas-Delgado, A, Martín, S, Ordoñez, R, Castellano, G, Kulis, M, Queirós, A, Seung-Tae, L, Wiemels, J, Royo, R[7], Puiggrós, M, Junyan, L, Giné, E, Beà, S, Jares, P, Agirre, X, Prosper, F, López-Otín, C, Puente, XS, Oakes, CC, Zenz, T, Delgado, J, López-Guillermo, A, Campo, E, **#**Martín-Subero, J.I. The proliferative history shapes the DNA methylome of B-cell tumors and predicts clinical outcome. Nature Cancer (in press). (**#** Corresponding authors).

2.1. Introduction.

During the last decade, several studies have focused on the analyses of DNA methylation during the entire B-cell maturation program [50,187] and in various B-cell neoplasms spanning the maturation spectrum. These include ALL [207,208], derived from precursor B cells, our previous **Study 1** in MCL [295] and CLL [187,254] derived from germinal center inexperienced and germinal center experienced B cells, diffuse large B-cell lymphoma (DLBCL) [328,329] derived from germinal center B cells, and multiple myeloma (MM) [280,281] derived from terminally-differentiated plasma cells. These studies have revealed a dynamic DNA methylome during B-cell maturation as well as novel insights into the cellular origin, pathogenic mechanisms and clinical behavior of B-cell neoplasms (extensively reviewed in [188,189]). However, a global analysis of the entire normal cell differentiation program and derived neoplasms is neither available for B cells nor for any other human cell lineage. Thus, in this **Study 2** of my thesis, I exploited both previously generated DNA methylation datasets as well newly generated data totaling over 2,000 samples to systematically decipher the sources of DNA methylation variability across B-cell neoplasms. This comprehensive dataset allowed us to perform a holistic DNA methylation analysis to dissect the DNA methylation variability at previously unexplored scales, including cancer-specific, tumor entity-specific, tumor subtype-specific and patient-specific variability.

2.2. Results.

2.2.1. Evaluating the impact of tumor cell content on DNA methylation data.

2.2.1.1. *In-silico* DNA methylation purification strategies.

In the **Study 1**, I detected that varying tumor cell content of MCL samples was reflected on the DNA methylation variability observed in a PCA analysis (**Fig. 11e**). After applying an *in-silico* DNA methylation purification strategy, the association between MCL DNA methylation variability and MCL tumor cell content decreased (**Fig. 11e**), and thus this strategy was considered valid. This decision was further supported by higher number of methylation changes between MCL subtypes as well as clear clinico-biological associations. This type of evaluation of DNA methylation variability was previously implemented to assess the impact of p-value overfitting in EWAS analyses due to cellular compositions of blood [286]. Although my initial statistical approach showed a reduction of the association between DNA methylation variability and cellular composition variability (**Fig. 11e**) (i.e., MCL with varying tumor cell contents), an experimental approach was needed. To properly study the validity of the *in-silico* purification of DNA methylation data, I obtained DNA methylation profiles using Infinium MethylationEPIC BeadChip (Illumina) of synchronic unpurified and purified MCL (5 pairs) and CLL (3 pairs) samples from peripheral blood and lymph node, respectively. The final goal of this analysis was to *in silico* purify the unsorted samples and assess whether a DNA methylation profile similar to the purified samples could be obtained. As a first step in the *in-silico* purification of DNA methylation data, I assumed that the major microenvironmental cells that could contaminate the DNA methylation data of B-cell tumor samples are HPC, granulocytes, T cells, monocytes, macrophages and endothelial cells. This was an extension of the initial strategy applied in **Study 1**, in which I assumed that MCL samples could only be contaminated with blood cell types, including granulocytes, monocytes, NK cells and T cells (please, be referred to **Results, section 2.2.2.** for an in-depth explanation on how to estimate the tumor cell content as well as that of microenvironmental cells by DNA methylation data). Once I obtained the tumor cell content and the proportions of microenvironmental cells, I next applied two strategies to *in silico* purify the DNA methylation data based on the cellular compositions present in each sample, i.e. the tumor cell content and the microenvironmental cells. I applied the previous methodology from the **Study 1** [295,304], and also a new linear modelling strategy I developed which I initially thought it could be more appropriate for B-cell tumors arising from different niches. I first assessed the linear modelling approach and

found, to my disappointment, that the majority of the *in-silico* purified samples clustered together regardless of their case identity, indicating that this method was overfitting the data and thus not adequate (**Fig. 15a**). Next, I analyzed the results with my initial methodology [295,304] (**Methods Study1, section 1.2., appendix, manuscript 2, 3**) and found that it was working well for some samples, while not for others (**Fig. 15b**). In those unpurified samples in which the *in-silico* purification strategy worked, namely M001, M004 and M360, the *in-silico* purified sample was clustering with the FACS-sorted sample (**Fig. 15b**). None of the 3 CLL samples achieved satisfactory results. Notably, in those samples in which the method did not work, the bias introduced was not as high as in the case of the linear modelling approach, since all the *in-silico* purified samples clustered together with their corresponding sample identity. Nonetheless, for the samples that the initial approach was working, the added benefit of the *in-silico* purification strategy was really small (very close branches in the hierarchical clustering with respect to the unpurified counterpart), suggesting that this complex bioinformatic labor may not be justified even in those samples where a subtle benefit is achieved (**Fig. 15b, c**). Collectively, none of the studied statistical approaches achieved satisfactory results, and thus I proceeded to carefully quantify the effect of varying tumor cell content in DNA methylation analyses. Worthwhile mentioning, the evaluation of all these strategies, although at the end disappointing, took nearly 1 year of computational analyses.

Fig. 15 | Experimental validation of *in-silico* DNA methylation purification approaches (see figure on next page). a, b Hierarchical clustering of synchronic MCL and CLL samples with low purity and their respective *in-silico* purified counterpart together with the same samples but with high purity based on flow cytometry data. All MCL samples were from blood, while CLL unpurified samples were from lymph nodes and their purified counterpart from peripheral blood. Two approaches to *in-silico* purify are shown, based on linear modelling **(a)** and on my previous developed strategy from **Study 1** (**Methods**) **(b). c,** Genome-wide DNA methylation comparison between the unpurified M360 MCL sample with its respective FACS-sorted sample and with the *in-silico* purified samples with the linear modelling strategy and the developed strategy in **Study 1** (Avg, Queiros 2016) [295].

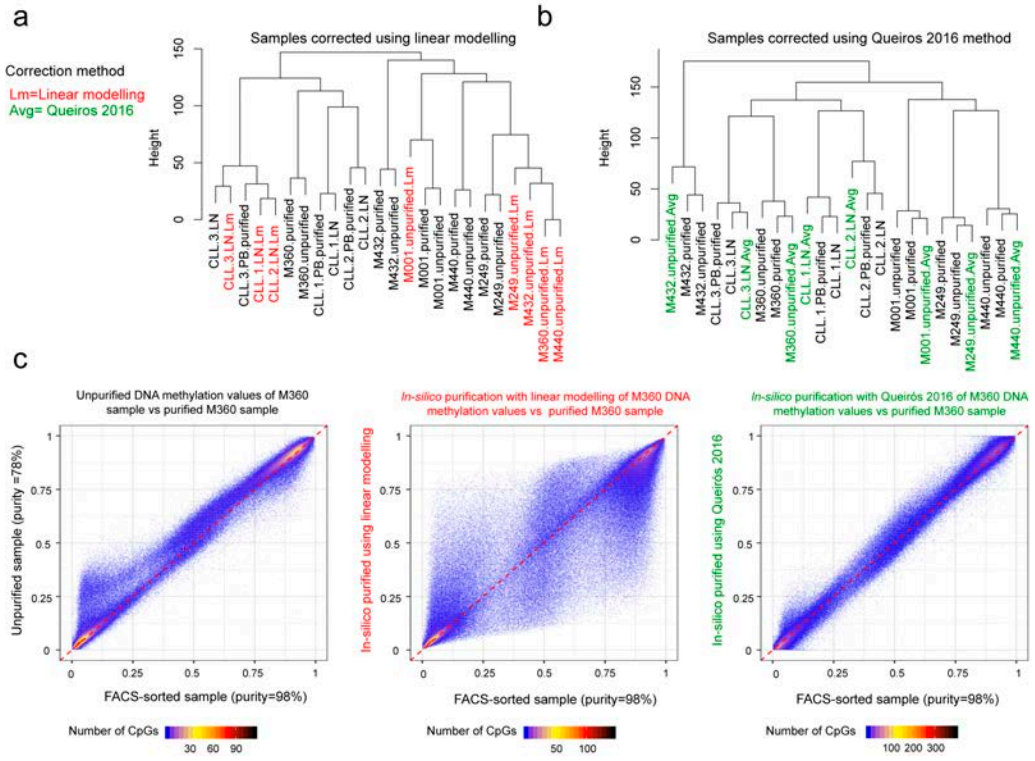


Fig. 15 | See caption on previous page.

2.2.1.2. Assessing the DNA methylation variability related to tumor cell content.

After the previous discouraging results, I studied the previous DNA methylation data of synchronic unpurified and purified MCL and CLL samples to explore to which extent differences in tumor cell content affected DNA methylation data. To dissect the DNA methylation variability related to varying tumor cell content, I first performed a PCA analysis with 5 pairs of synchronic MCL samples from peripheral blood (**Fig. 16a**). I represented all components of the PCA analysis to distill all sources of DNA methylation variability. To my surprise, the influence of different tumor cell content among samples was not clearly manifested until the last component, which explained roughly 2% of the total DNA methylation variability. Accordingly, samples were clustered in all the remaining components based on their biological origin and not on their tumor cell content (**Fig. 16a, b**). This finding suggests that the effect of varying tumor cell content over the DNA methylation data is in general almost negligible if tumor samples have a

tumor cell content of approximately 60%. Indeed, this value coincides with the last recommendations for sample selection of the TCGA consortium: https://www.cancer.gov/about-nci/organization/ccg/blog/2018/bcr-tips. In line with these results, a PCA analysis with all the synchronic MCL and CLL samples with varying purities showed that samples clustered according to their biological origin, and no apparent effect of varying tumor cell content was observed (**Fig. 16c**). Collectively, all these data indicate that tumor cell content barely affects DNA methylation data compared with the tumor biology if patient samples have tumor cell content above approximately 60%. Thus, I subsequently excluded samples whose tumor cell content was below 60%.

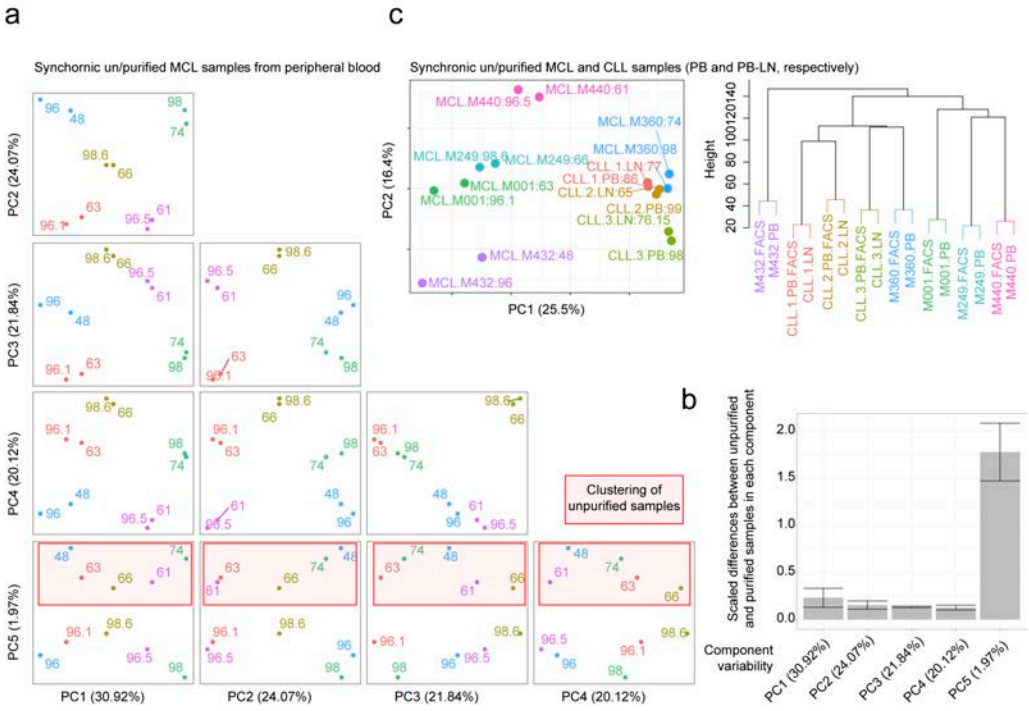


Fig. 16 | Tumor cell content effect on DNA methylation variability in MCL and CLL samples. a, PCA analysis including until the 5th component of synchronic un/purified MCL samples. **b,** Scaled differences component values between unpurified and purified MCL samples. **c,** PCA analysis and hierarchical clustering of MCL and CLL synchronic samples. The same samples are represented with the same color and the tumor cell content is labelled.

2.2.2. Inferring tumor cell content through DNA methylation.

In the **Results section 1.2.1. from Study 1**, I previously estimated the tumor cell content of MCL samples using the DNA methylation profile of each patient sample. There, I relied on a previous statistical framework which is able to accurately predict the cellular composition from blood samples [301,302]. This strategy relies on the fact that the main cell types present in blood are known and can be distinguished by their DNA methylation profile [287]. Thus, this approach requires the DNA methylomes of the cell types forming the complex sample to be available (i.e., blood). This is commonly known as a *reference-based* approach [307,308], which differs from "reference-free" approaches in the necessity of previously generated DNA methylation profiles from the sample under study. Notably, it has been shown that the *reference-based* approach outperforms reference-free strategies if DNA methylation profiles are available [330]. As previously mentioned, I adapted this strategy to calculate MCL tumor cell content (**Fig. 11b**). In the initial strategy, I assumed that the DNA methylation signature of $CD19^+$ cells from healthy donors present in tumor samples could be used to infer the tumor cell content, since the presence of normal B cells in tumor samples is negligible, and tumor samples preserve a DNA methylation signature from their cellular origin. These assumptions were proved correct, as I obtained highly accurate tumor cell content predictions using this strategy (**Fig. 11b**). To adapt this initial strategy to all B-cell tumors regardless of their cellular origin, I substituted the DNA methylation signature of $CD19^+$ cells from blood by a Pan-B-cell DNA methylation signature created with purified B cells, including precursor B cells to terminally differentiated bmPC. To do so, I used purified subpopulations previously generated in my host lab [50]. This ensured that the tumor cell content of B-cell tumors originated throughout the entire B-cell maturation spectrum could be accurately predicted. In addition, I increased the number of potential reference cell types present in the B-cell tumor samples. This responds to the fact that in the present study B-cell tumors could be derived mainly from three different niches including bone marrow, peripheral blood and lymph node. Thus, in addition to the tumor cells, I assumed that samples could contain the main microenvironmental cells from any of these niches, including HPC, granulocytes, T cells, monocytes, macrophages and endothelial cells. With this new strategy, I calculated the tumor cell content in the entire dataset. To assess the accuracy of this methodology, I compared the obtained tumor cell content predictions with gold standard methods, including FCM and genetic data. In line with the previous results in MCL (**Results section 1.2.1. Fig. 11b**), I obtained high correlations

between tumor cell content predictions using DNA methylation profiles and FCM/genetic data in MCL and CLL samples, which included two independent cohorts (**Fig. 17a**). As expected, I did not obtain such high correlations in MM (**Fig. 17b, c**), which are known to lose the B-cell DNA methylation signature [280]. Surprisingly, this was also the case for some DLBCL samples, where estimated tumor cell content based on somatic mutations was above to 80% but very low based on DNA methylation (**Fig. 17b, c**). The precise mechanisms of this B-cell DNA methylation identity loss in some DLBCL samples is a novel finding serendipitously observed in the course of this doctoral thesis, and warrants further investigation. Based on these findings, I used tumor cell content predictions based on DNA methylation in ALL, MCL and CLL, whereas in MM and DLBCL I used available FCM and genetic data, respectively. After filtering samples without a tumor cell content above 60%, I retained 1,595 samples including normal and neoplastic B cells for primary analyses. This complete dataset allowed me to interrogate the DNA methylation dynamics in normal and neoplastic B-cell from a previously unexplored scales, including cancer-specific, entity-specific, subtype-specific and patient-specific levels (**Fig. 18a, b**).

Fig. 17 | Tumor cell content estimation through DNA methylation in B-cell tumors (see figure on next page). a, Passing Bablok regression fits between flow cytometry or genetic data and DNA methylation-based tumor cell content estimates in MCL and CLL. **b**, Passing Bablok regression fits between flow cytometry and DNA methylation-based tumor cell content estimates in MM and DLBCL. **c,** DNA methylation signature used to estimate the purity of B-cell tumors in MM and DLBCL, as well as their purity according to FMC and genetic data. The DNA methylation values of the B-cell signature for the microenvironmental cell types is also shown.

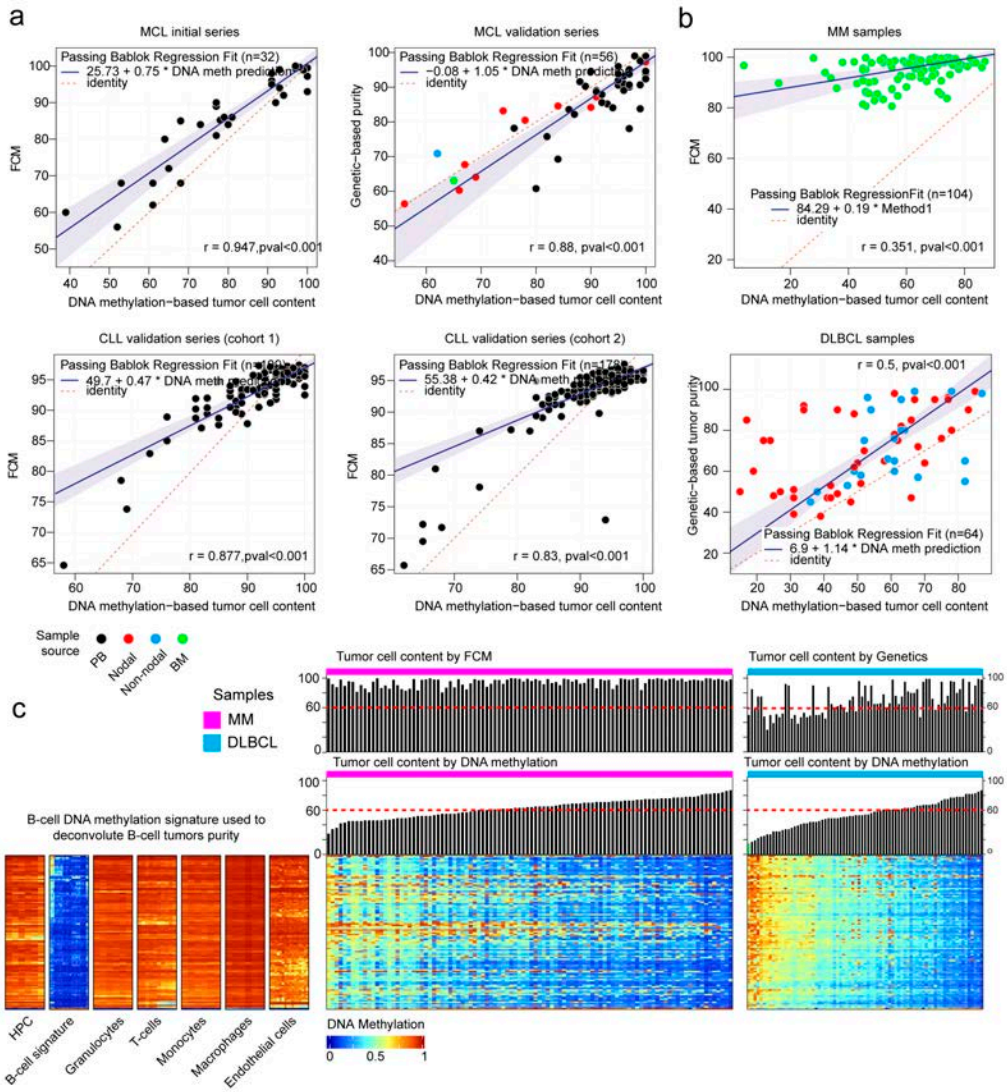


Fig. 17 | See caption on previous page.

2.2.3. Shared DNA methylation dynamics in normal and neoplastic B cells.

Previous DNA methylation studies focused on the identification of DNA methylation changes taking place in health and disease [52,84,123], but none of them analyzed the features of the genome fraction whose DNA methylome remains stable. To shed light into this aspect, I used the comprehensive dataset including normal B cell and derived neoplasms spanning the entire B-cell differentiation spectrum (**Fig. 18a, b**). I identified that only 12% of the studied CpGs present in the 450k array show stable DNA

methylation levels in normal and neoplastic B cells (**Fig. 18c, d**). I characterized these CpGs based on genetic location, CpG content and chromatin states from primary samples of the Blueprint consortium (**Methods Study 2, section 2.6**) [3]. I identified that stably methylated regions are located at gene-bodies of actively transcribed genes whereas stably unmethylated sites are prevalent in CGI of promoter regions showing active chromatin marks (**Fig. 18e, f, g**). Remarkably, stably methylated and unmethylated CpGs mainly converge into the same genes (**Fig. 18h**), which are highly expressed in normal and neoplastic B cells (**Fig. 18i**) and related to essential cellular functions such as cell cycle, RNA processing and energy metabolism (**Fig. 18j**).

Finally, I wondered whether all B-cell tumors, regardless their specific diagnosis, share a unifying DNA methylation signature related to their neoplastic nature. As the 12% of the DNA methylome is stable in the studied series and 25% is modulated in normal B-cell development [50,187], I took the remaining 63 % of the DNA methylome that could contain tumor-specific DNA methylation changes. Comparing the DNA methylomes of normal and neoplastic B cells, I did not identify any consistent *de novo* DNA methylation signature shared by all B-cell neoplasms. Instead, DNA methylation variability seems to be related to differences among B-cell tumor entities and subtypes as well as patient-specific variability, as will be shown in the following sections.

Fig. 18 | Dataset of B-cell tumors used for primary analyses and shared DNA methylation dynamics in normal and neoplastic B cells (see figure on next page). a, Dataset of normal and neoplastic B-cells which passed all filtering criteria and used for all primary analyses. **b,** DNA methylation variability levels explored in the **Study 2**. **c,** The DNA methylation dynamics in normal and neoplastic B cells. **d,** Heatmaps showing representative stably un/methylated CpGs in normal and neoplastic B cells. The total number of CpGs affected is shown on the left. **e,** Chromatin state genome segmentation used in the whole thesis with six non-redundant histone marks shown at bottom. **f,** Chromatin state enrichment for the stably un/methylated CpGs. Fold change enrichments were calculated over the 450k background. **g,** Location for the stably methylated and unmethylated CpGs with respect to the background 450k. **h,** Number of genes affected by any CpGs stably methylated (red circle) or unmethylated (blue circle) CpGs and their overlap. **i,** Gene expression percentiles in normal and neoplastic B cells of genes affected by both stably methylated and unmethylated CpGs. The number of samples analyzed is shown at bottom. **j,** Gene ontology terms enriched for the genes in panel **i.**

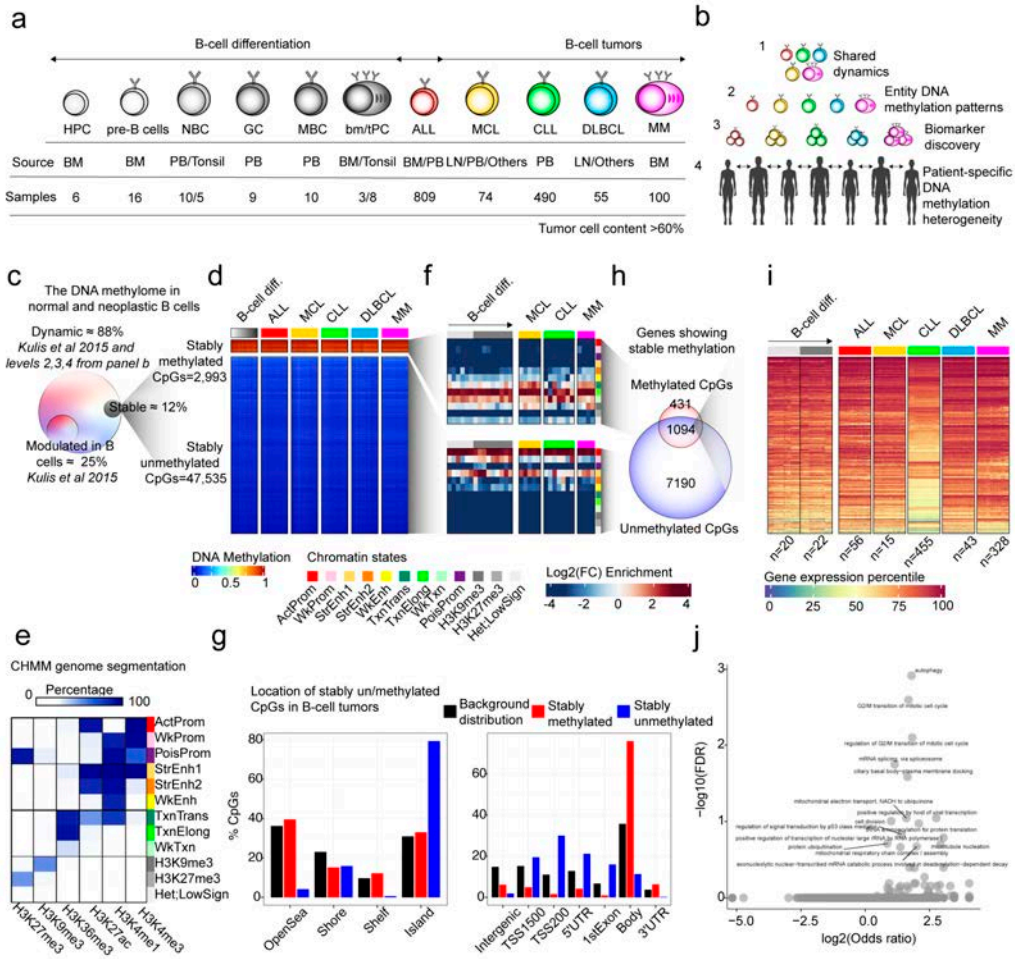


Fig. 18 | See caption on previous page.

2.2.4. Entity-specific DNA methylation landscapes.

I next aimed at studying whether disease-specific DNA methylation patterns may unravel differential pathogenic mechanisms underlying each B-cell neoplasm. Overall, a principal component analysis (PCA) showed that all B-cell neoplasms cluster separately (**Fig. 19a, b**). The first component was related to B-cell development, separating neoplasms according to the maturation stage of their cellular origin, i.e. ALL together with pre-germinal center B cells and mature B-cell neoplasms together with germinal center B cells, memory B cells and plasma cells. The remaining components showed tumor-specific patterns, such as PC3, PC7 and PC9 related to MM, DLBCL and MCL-specific variability, respectively. Next, in order to identify DNA methylation signatures specifically

associated with malignant transformation, I focused the next analysis on the genome fraction potentially changing only in tumor entities, which as mentioned above accounts for 63% of the studied CpGs. (**Fig. 19c, top panel**). I detected varying numbers of tumor-specific DNA methylation (tsDNAm) changes, including hyper- and hypomethylation changes which ranged from 616 in CLL to 49,279 in MM (**Fig. 19c, bottom panel**). Overall, the total number of tsDNAm represented roughly 15% of the entire DNA methylome (**Fig. 19d**). In general, hypermethylation was enriched at CGI and promoter related regions, whereas hypomethylation was related to low CpG content regions such as open sea, shore and shelves (**Fig. 19e**). Notably, I observed that DNA methylation changes manifested differently in distinct neoplasms. ALL and DLBCL showed more tumor-specific DNA hypermethylation (tsDNAm-hyper) whereas MCL, CLL and MM acquired more tumor-specific DNA hypomethylation (tsDNAm-hypo), being this skew towards hypomethylation remarkable in MM (**Fig. 19c, bottom panel**). In this context, I evaluated whether this different trends of acquiring aberrant DNA methylation were related to the expression levels of *DNMTs*. However, it seems that *DNMT1*, *DNMT3A* and *DNMT3B* expression does not explain these differences, since their expression is rather homogenous in all B-cell tumors (with the exception of *DNMT3B* in MCL and CLL samples) (**Fig. 19f**). Furthermore, in cases with concurrent DNA methylation and expression profiling, I could not identify any clear correlation between the hypermethylation/hypomethylation ratio and the *DNMT1*, *DNMT3A* or *DNMT3B* expression levels (please, be referred to **Results, Study 3, section 3.2.2** for further details). Additional studies are needed to elucidate the mechanisms underlying the different trends of aberrant DNA methylation in B-cell tumors.

Fig. 19 | B-cell-tumor-specific DNA methylation changes characterization (see figure on next page). a, b, First two (panel **a**) and nine (panel **b**) components from a PCA analysis with the complete dataset of normal and neoplastic B cells. **c,** Number of *de novo* tumor-specific DNA methylation changes. **d,** The DNA methylation dynamics in normal and neoplastic B cells and the percentage that *de novo* DNA methylation changes represent over the whole DNA methylome. **e,** Location of the *de novo* tumor-specific DNA methylation changes in B-cell tumors with respect to 450k background. Colors at the bottom of each barplot represent B-cell tumor samples as in panel **a. f,** Gene expression percentiles of the main *DNMTs* in normal and neoplastic B cells.

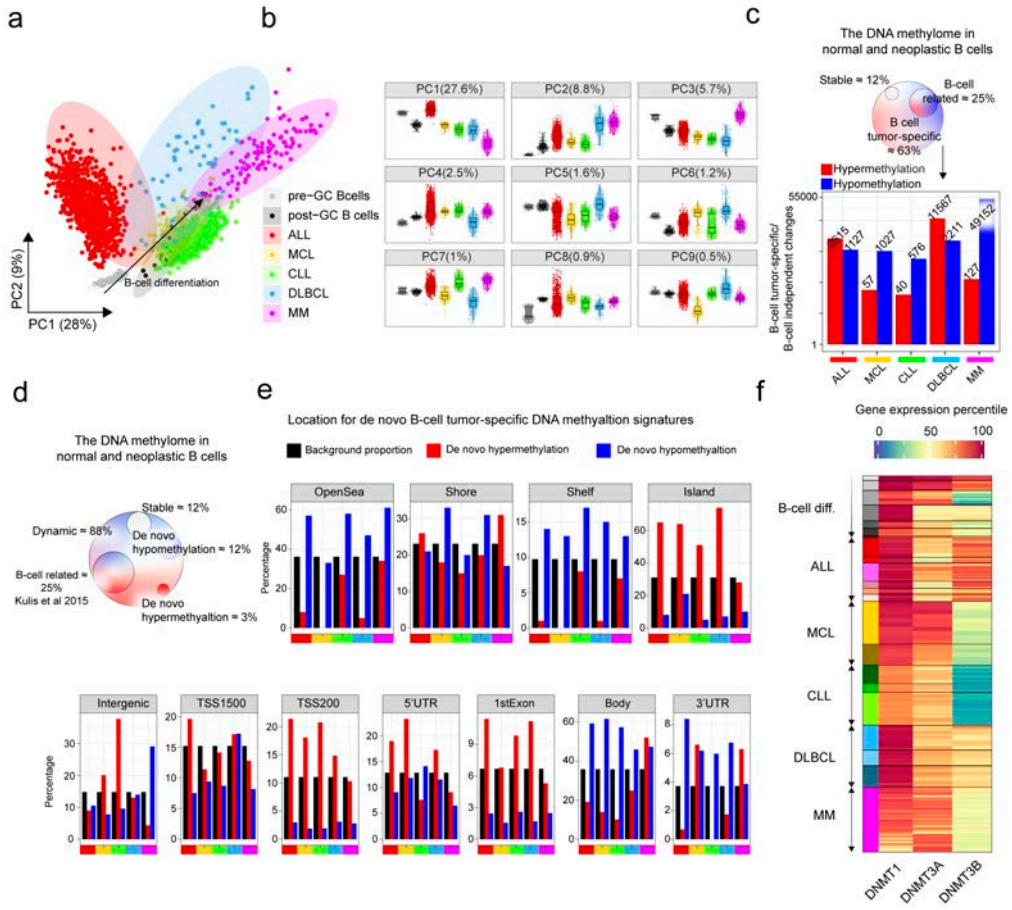


Fig. 19 | See caption on previous page.

Next, I assessed the potential regulatory function of both tsDNAm-hyper and tsDNAm-hypo signatures and their possible relation to the pathogenesis of each respective B-cell tumor. As widely described in the literature [47,86], tsDNAm-hyper signatures affected CGI, polycomb repressed regions/poised promoters of silenced genes, with the exception of MM, which was related to silencing of B cell-specific enhancers [280]. Interestingly, tsDNAm-hypo signatures were strongly enriched for regulatory regions in ALL, MCL and CLL, whereas depleted in DLBCL and MM (**Fig. 20a**). Transcription factors (TFs) have been reported to reduce DNA methylation levels of regulatory regions upon their binding to the DNA [92,160]. Therefore, I performed TFs binding site prediction analysis in active regulatory regions containing tsDNAm-hypo CpG (**Fig. 20b**). Remarkably, the entities in which tsDNAm-hypo was predominantly

located in H3K27ac regions (**Fig. 20a**) showed enrichments for binding sites of TFs expressed in each respective entity and previously linked to their pathogenesis, such as SPI1/SPIB and EBF1 in ALL, TCF/ZEB in MCL, and NFAT in CLL (**Fig. 20b, c**) [331–333]. In the case of DLBCL and MM, their associated tsDNAm-hypo CpGs were actually depleted of regulatory elements (**Fig. 20a**), suggesting that TF binding may not be a major factor leading to their tumor-specific DNA methylation signatures in these B-cell tumors. However, focusing only on the underrepresented H3K27ac-containing tsDNAm-hypo CpGs, I could also detect significant relationships with TFs potentially involved with their pathogenesis, such as FOX family in DLBCL [334], and NRL (a member of the oncogenic MAF family), ISL1, TEAD, and YY1 in MM [335–338]. Finally, beyond the potential role of TFs in shaping tumor-specific DNA methylation signatures, I also investigated the downstream transcriptional levels of genes related to tsDNAm-hypo signatures. An analysis of the transcriptional profiles of cases from all five diseases revealed a total of 94 genes associated with tsDNAm-hypo genes expressed in a disease-specific manner (**Fig. 20d**). Although some of the identified genes have been shown to be specifically expressed in a particular disease, such as *CTLA4* and *KSR2* in CLL [339], this comprehensive analysis further provides a rich resource of disease-specific candidate genes in which differential DNA methylation may play a role in their deregulation.

Fig. 20 | B-cell-tumor-specific DNA hypomethylation changes (see figure on next page). a, Heatmap showing *de novo* B-cell tumor-specific DNA hypomethylation and the number of CpGs in regulatory. **b,** Significantly (FDR<0.01) enriched TF binding site prediction analyses in regions containing CpGs in regulatory regions from panel **a**. **c,** Gene expression percentiles in B-cell tumors of the most significantly enriched TF for each B-cell tumor. **d,** Gene expression percentile in normal and neoplastic B cells for genes related to *de novo* DNA hypomethylation for each B-cell tumor from panel **a.**

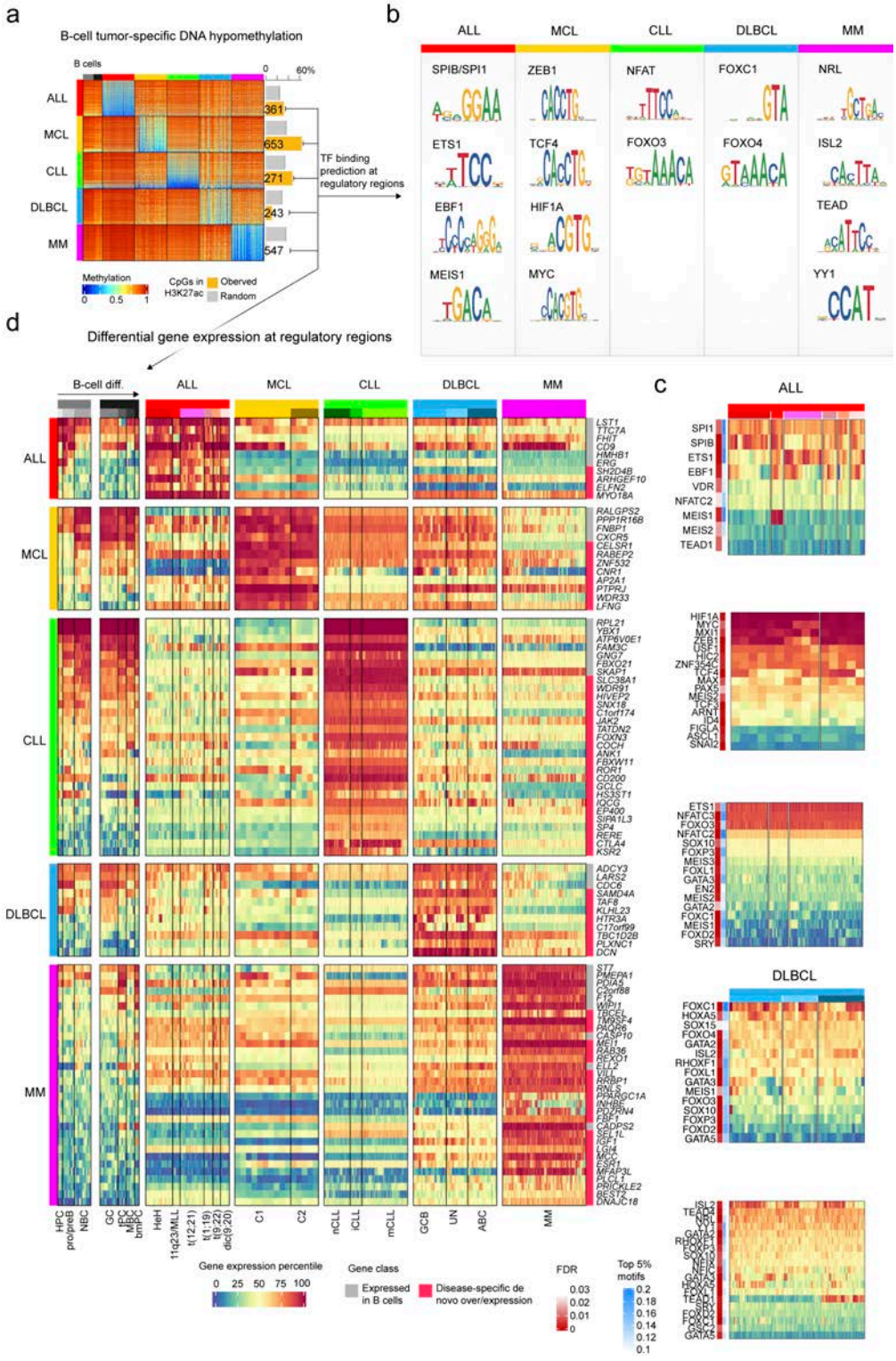


Fig. 20 | See caption on previous page.

2.2.5. Construction of a Pan-B-cell tumor classifier.

The B-cell neoplasms under study (**Fig. 18a**) represent broad categories which are further classified into subtypes with different clinico-biological features based on genetic, transcriptional or epigenetic features [164]. These include high-hyperdiploid (HeH) ALLs, and ALLs with structural variants: rearrangements affecting 11q23/MLL, three different chromosomal translations, i.e. t(12;21), t(1;19), and t(9;22), as well as the dicentric chromosome dic(9;20) [207]; Cluster 1 (C1, DNA methylation patterns related to germinal center-inexperienced cells) and Cluster 2 (C2, DNA methylation patterns related to germinal center-experienced cells) (**Study 1**) MCLs [295] which mostly reflect conventional and leukemic non-nodal MCLs [229] (**Study 4**); naïve-like/low-programmed, intermediate/intermediate programmed and memory-like/high-programmed CLLs [187,254], and finally germinal center B cell (GCB) and activated B cell (ABC) DLBCLs [194]. In MM, a previous report did not show DNA methylation differences among the distinct cytogenetic subtypes [280] and thus MM subgrouping was not included in the analyses. All these studies focused on the particularities of DNA methylation patterns within each respective disease subtypes and led to new insights into the cellular origin and pathogenic mechanisms of each B-cell tumor. Thus, my goal here was the identification of epigenetic biomarkers that may allow a comprehensive diagnosis of B-cell tumors entities and subtypes at once. I devised a strategy to construct a classifier algorithm that yielded 56 CpGs as the optimal number distributed along 5 predictors to accurately discriminate the main B-cell tumor entities as a first step (predictor 1), and subsequently B-cell tumor subtypes as a second step (predictors 2, 3, 4 or 5) (**Fig. 21a, b and Methods Study 2, section 2.10**). The accuracy of the five predictors was evaluated using nested 10-fold stratified cross-validation in the training series (n=1,345) and with independent validation series (n=711) (**Fig. 21c**). Overall, I obtained very high accuracies in the predictions in both main B-cell tumor entities (mean sensitivity was 97% for training series and 99% for validation series) and B-cell tumor subtypes (mean sensitivity was 90% for training series and 97% for validation series). This epigenetic classifier (**Methods Study 2, section 2.20**) may represent the basis for a simple and accurate diagnostic tool for main B-cell tumors and B-cell tumor subtypes with more complicated differential diagnosis such as MCL or CLL subtypes.

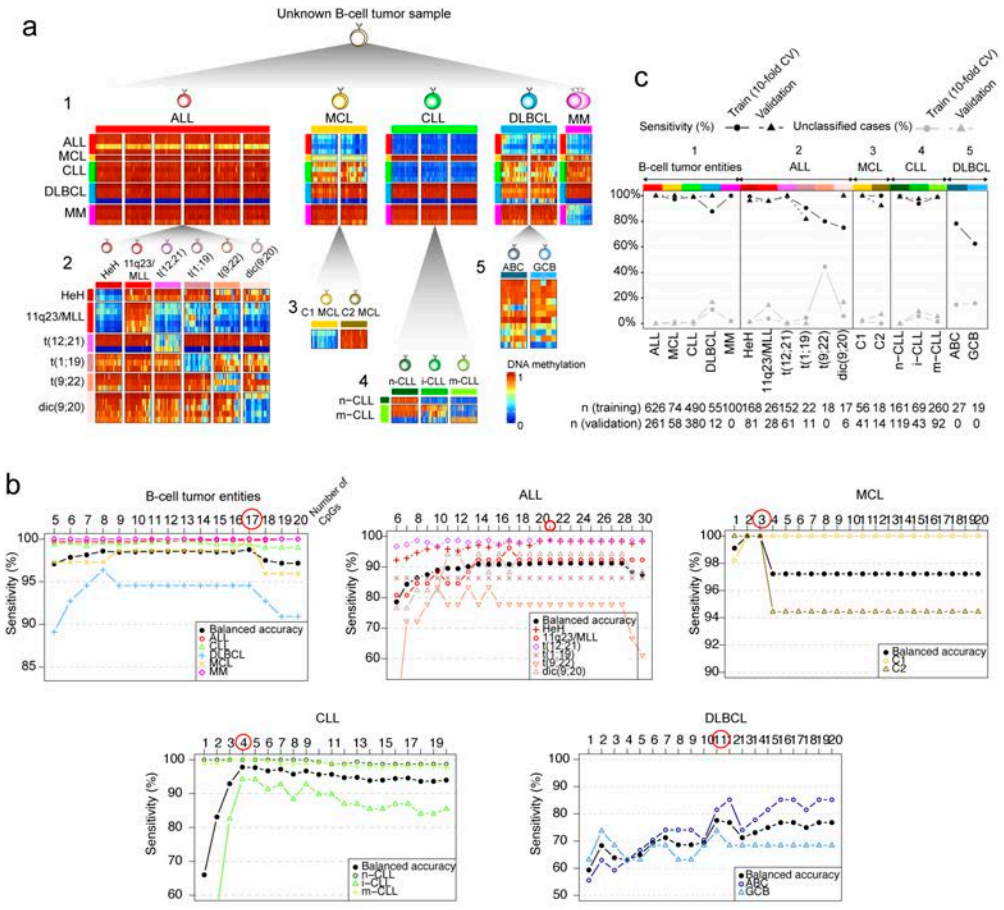


Fig. 21 | Pan-B-cell diagnostic algorithm. a, Heatmap showing the CpGs used for the pan B-cell tumor diagnostic classifier. Each of the 5 predictors that compose the classifier are shown separately. B-cell tumor entities are predicted in the first step (predictor 1), and B-cell tumor subtypes in the second step (predictors 2, 3, 4 and 5). **b,** Sensitivity of each of the 5 predictors according to an increasing number of CpG used in the training series. The optimal number of CpGs per each predictor (red circle) was selected maximizing the balanced accuracy in the cross-validation. **c,** Sensitivity of the final 5 predictors as well as the percentage of the unclassified cases in the training and validation series.

2.2.6. Inter-patient DNA methylation variability.

2.2.6.1. Normal and neoplastic DNA methylation changes are highly correlated.

After having characterized entity-based sources of DNA methylation variability, I next aimed at studying DNA methylation changes in individual patients within each tumor subtype (**Fig. 18b, level 4**). To that end, I computed the total number and the number of hyper- and hypomethylation changes in every single patient within each B-cell tumor

subtype as compared to HPC (**Fig. 22a**), as it was previously done in the **Study 1** for MCL patients. As each B-cell tumor entity is derived from a distinct cellular origin, this approach has the advantage of fixing a reference point for all B-cell tumors, and the changes observed can be subsequently dissected into those mainly modulated in normal B-cell maturation (i.e. B cell-related changes) and those taking place mainly in the context of neoplastic transformation (i.e. B cell-independent changes). Overall, I found large differences in the numbers of DNA methylation changes per patient (**Fig. 22a**), with some patients displaying few and others extensive DNA methylation changes. To analyze whether some B-cell neoplasms show an intrinsically more variable epigenome, I assessed the degree of DNA methylation variability as a function of the number of DNA methylation changes in each group, but no differences were observed (**Fig. 22b**). The total number of altered CpGs per case was classified into four categories, depending whether DNA methylation was gained or lost and whether it was modulated or not during normal B-cell development [50]. I found striking correlations between the degree of B-cell related and B-cell independent DNA methylation changes regardless of B-cell tumor entities and subtypes (**Fig. 22c**), a finding that was maintained when hypermethylation and hypomethylation were studied separately (**Fig. 22d**). This association suggests that the overall DNA methylation burden of the tumor in each individual patient may be shaped by a similar underlying phenomenon, which take place simultaneously in the form of hyper- or hypomethylation changes in both normal and neoplastic B cells.

Fig. 22 | B-cell related and B-cell independent DNA methylation changes are highly correlated in B-cell tumors (see figure on next page). a, Individual-specific DNA methylation changes in normal and neoplastic B cells. Outer, middle and inner tracks represent the total number of DNA methylation changes, and the total number of hypo- and hypermethylation changes, respectively. Dark and light colors represent B-cell related and B-cell independent DNA methylation changes. **b,** Correlation between the interquartile range (IQR) and the median of DNA methylation changes for each samples subtype. **c,** Correlation between the number of B-cell related and B-cell independent DNA methylation changes in normal and neoplastic B cells. Correlations for each B-cell tumor subtype are shown below. **d,** Correlations between B-cell related and B-cell independent DNA methylation changes with hyper- and hypomethylation changes analyzed separately.

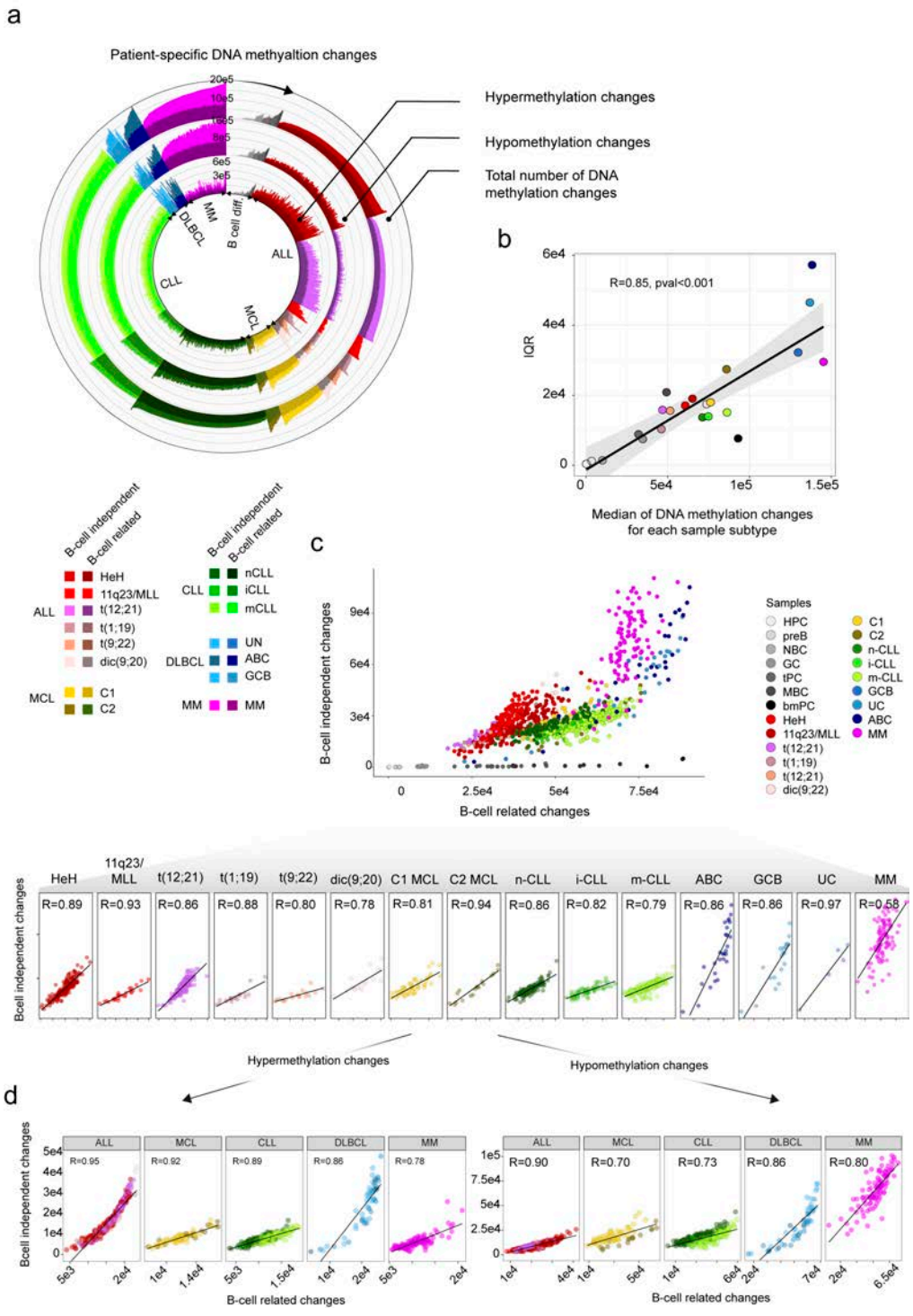


Fig. 22 | See caption on previous page.

2.2.6.2. Individual-specific DNA methylation changes target silent chromatin.

Next, I proceeded to characterize the striking correlations previously found in B-cell tumors between DNA methylation changes taking place during normal B-cell development and malignant transformation. To do so, I found CpGs with consistent DNA methylation dynamics in all individual patients, separating hyper- or hypomethylation and B-cell related or B-cell independent DNA methylation (**Fig.23a and Methods Study 2, section 2.11**). This was done by retaining CpGs with any of the aforementioned four behaviors in any patient and not showing the opposite trend in any other individual patient. Doing so, I ensured to capture a more homogenous set of CpGs to better understand the correlations in individual patients between B-cell-related and B-cell independent changes (**Fig. 22c, d**). I found over 200,000 CpGs fulfilling this criterion, with hypomethylated changes being more abundant than hypermethylated changes (72% vs 28%), and with a slightly a greater number of B-cell independent than B-cell related changes (58% vs 42%) (**Fig. 23a**). A thorough annotation of the four CpGs types based on genomic location, chromatin states and gene expression revealed two main tendencies regardless of the number of DNA methylation changes taking place during normal differentiation or neoplastic transformation. Patient-specific CpGs that underwent hypomethylation in both B-cell related and B-cell independent fractions were consistently located in low CpG-content (open sea) low-signal heterochromatin, and were associated with silent genes in both normal and neoplastic B cells (**Fig. 23b, c**). In the case of patient-specific hypermethylation, CpGs in both fractions were located in promoter regions and CGIs in regions containing the H3K27me3 mark and poised-promoter chromatin states, and affect genes that also remained silent across normal differentiation and neoplastic transformation of B cells (**Fig. 23b, c**).

Collectively, all these findings indicate that most of DNA methylation changes in B-cell tumor patients occur in silent chromatin regions in the absence of concurrent phenotypic changes, suggesting that a mechanism independent from gene regulation may shape their overall DNA methylation landscape.

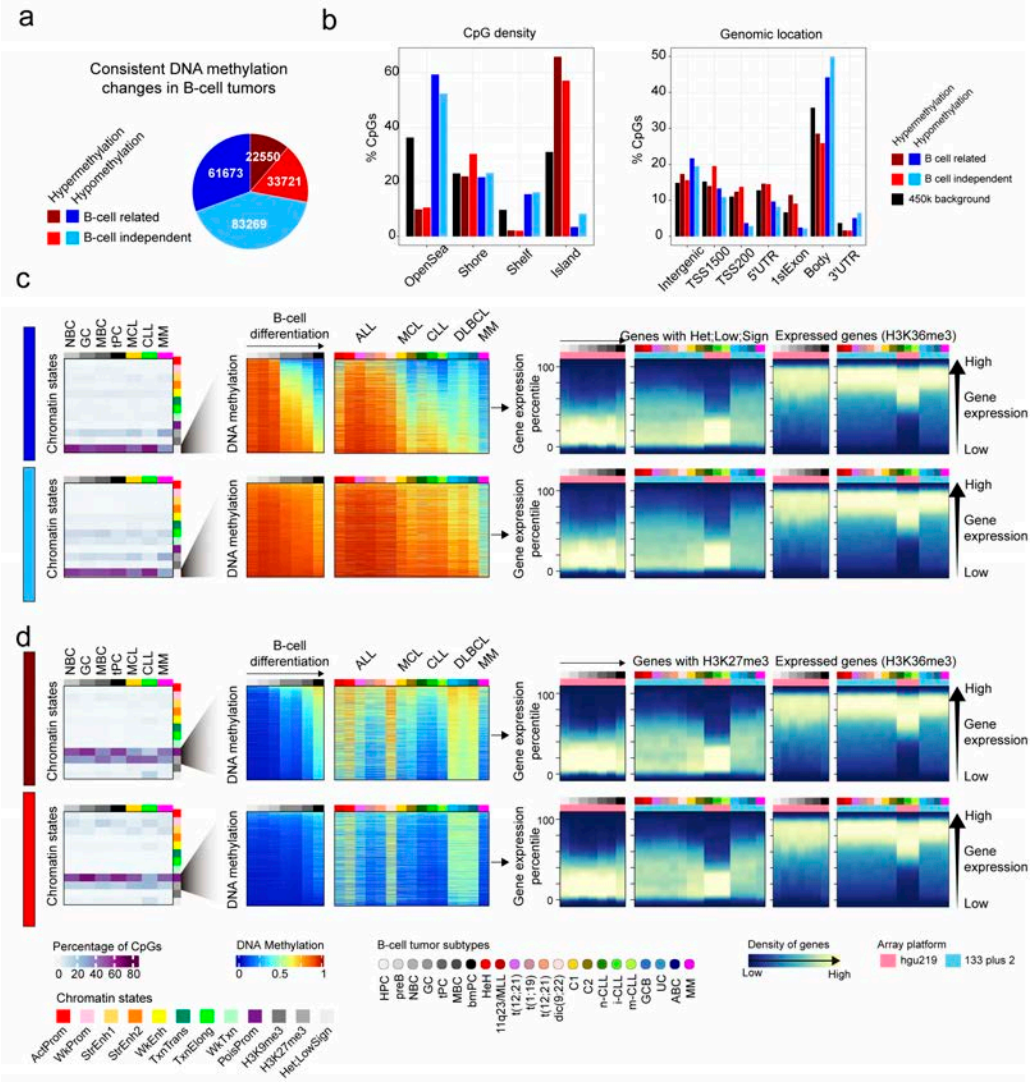


Fig. 23 | Chromatin and transcriptomic characterization of individual-specific DNA methylation changes in normal and neoplastic B cells samples. a, Consistent DNA methylation changes in B-cell tumors separated by hyper or hypomethylation and by B-cell related or B-cell independent changes. **b,** Genomic location for the CpGs in panel **a**. **c,** Left, percentage of B-cell related (top) or B-cell independent (bottom) hypomethylated CpGs from panel **a** in each respective chromatin state in normal and neoplastic B cells. Middle, heatmaps with some representative CpGs from Het;LowSign regions. Right, gene expression density in gene expression percentiles of genes containing any CpG in Het;LowSign regions. **d,** Left, percentage of B-cell related (top) or B-cell independent (bottom) hypermethylated CpGs from panel **a** in each respective chromatin sates in normal and neoplastic B cells. Middle, heatmaps with some representative CpGs from H3K27me3-marked regions. Right, gene expression density in gene expression percentiles of genes containing any CpG in H3K27me3-marked regions. Expressed genes with the H3K36me3 mark are also shown.

2.3. Conclusions.

- The overall DNA methylome of B-cell tumor samples is minimally affected by microenvironmental cells if tumor cell content is above 60%, and bioinformatic corrections may result in subtle benefits or even lead to data overfitting.

- Tumor cell content can be accurately predicted through the DNA methylome with a *reference-based* approach in ALL, MCL and CLL, while in MM and in a subset of DLBCL patients this approach is not valid due to their B-cell epigenetic signature loss.

- A great fraction (88%) of the DNA methylome is modulated under normal and neoplastic conditions. The remaining fraction shows stable DNA methylation levels in normal and neoplastic B cells, affecting mainly the same expressed genes which are involved in fundamental cellular processes.

- The extensive DNA methylation variability among different B-cell neoplasms is in part related to imprints of normal B-cell development.

- Disease-specific hypomethylation signatures are in part related to binding sites of particular TFs, and are associated with the expression of genes related to the pathogenesis of each B-cell tumor.

- B-cell tumor entities and subtypes can be accurately identified with few CpGs through the implementation of a Pan-B-cell diagnostic algorithm.

- The number of B-cell related and B-cell independent changes in individual B-cell tumor patients is heterogeneous but highly correlated, suggesting that these epigenetic changes are caused by a similar underlying phenomenon.

- The great majority of DNA methylation changes in individual patients are located in silent chromatin without any impact on gene expression, and therefore do not seem to have any apparent regulatory function.

Study 3. Understanding patient-specific DNA methylation changes in B-cell tumors.

The results of this study have been partly published (**appendix, manuscript 4**), while others are in preparation.

 # Duran-Ferrer M, Clot G, Nadeu F, Beekman, R, Baumann, T, Nordlund, J, Marincevic-Zuniga, Y, Lönnerholm G, Rivas-Delgado, A, Martín, S, Ordoñez, R, Castellano, G, Kulis, M, Queirós, A, Seung-Tae, L, Wiemels, J, Royo, R[7], Puiggrós, M, Junyan, L, Giné, E, Beà, S, Jares, P, Agirre, X, Prosper, F, López-Otín, C, Puente, XS, Oakes, CC, Zenz, T, Delgado, J, López-Guillermo, A, Campo, E, **#**Martín-Subero, J.I. The proliferative history shapes the DNA methylome of B-cell tumors and predicts clinical outcome. Nature Cancer (in press). (**#** Corresponding authors).

3.1. Introduction.

It is increasingly accepted that the main features of cancer DNA methylomes are focal hypermethylation at CGI and a global loss of DNA methylation [47,84,86]. Despite this general finding, fewer studies have analyzed patient-specific DNA methylation, and none of them considering the DNA methylation changes related to normal cell development prior to malignant transformation. Some previous works linked the DNA methylation variability among individuals to human chronological age [317,340]. These findings have led to the concept of epigenetic clocks, which aim at capturing the chronological age of individuals [341,342]. In addition to this link to aging, several studies published during the last years suggest that DNA methylation changes are accumulated during cell division [73,343–345]. These studies, performed in fibroblasts and hematopoietic stem cells, reported that DNA methylation loss in late-replicated regions and DNA methylation gain at CGI increase as cells proliferate without an apparent impact on gene expression. In fact, these studies may relate to earlier observations reporting preferential DNA methylation of CGI marked by H3K27me3 in cancer cells [133–135]. Based on these previous findings, and considering the availability of a large dataset of normal and neoplastic B cells, the **Study 3** of my thesis is aimed at obtaining a deeper insight into the relationship among individual DNA methylation variability, cell division and aging.

3.2. Results.

3.2.1. Silent DNA methylation changes accumulate during cell division.

First, inspired by experimental evidence published in the literature [73,343,344], I explored whether DNA methylation changes in silent regions reflect the proliferative history of primary human B cells (**Fig. 24a**). To do so, I took advantage of previous independent DNA methylation data available in my host lab from a previous collaboration with the group of Thierry Fest [298]. This dataset consisted of 450K DNA methylation data and gene expression arrays of an *in-vitro* differentiation model of primary NBCs into plasma cells which involves varying levels of cell division (**Fig. 24b**). At days 4 and 6, different B-cell subpopulations were separated based on their proliferation history measured by carboxyfluorescein succinimidyl ester (CFSE) dilution. In these cells, I evaluated the DNA methylation profile in repressed chromatin regions and I detected the presence of hypermethylation at repressed H3K27me3-containing regions and hypomethylation of low signal/H3K9me3-containing heterochromatic regions (**Fig. 24c**). To map these repressive regions, I took the constitutively silent chromatin states from primary naïve B cells and germinal center B cells. Interestingly, the three B-cell subpopulations with different proliferative histories sorted at day 6 showed a gradual accumulation of DNA methylation changes in silent regions, a finding that was directly associated with their proliferative history (from less divided P3 to highly divided P1 cells). The same phenomenon was observed at day 4 but was less marked. The genes associated with the measured CpGs did not show any change of expression levels regardless of their DNA methylation status, and thus were unlikely to be related to the phenotype of the cells (**Fig. 24d**). Collectively, these data support the concept that cell division leaves transcriptionally-inert epigenetic imprints in silent chromatin of B cells. Therefore, the findings in **Study 2, section 2.2.6**. showing that great fraction of the patient-specific DNA methylation changes in B-cell tumors accumulate at H3K27me3 and low signal/H3K9me3 regions without affecting gene expression levels, are consistent with an epigenetic mitotic clock reflecting the accumulation of DNA methylation changes during cell division.

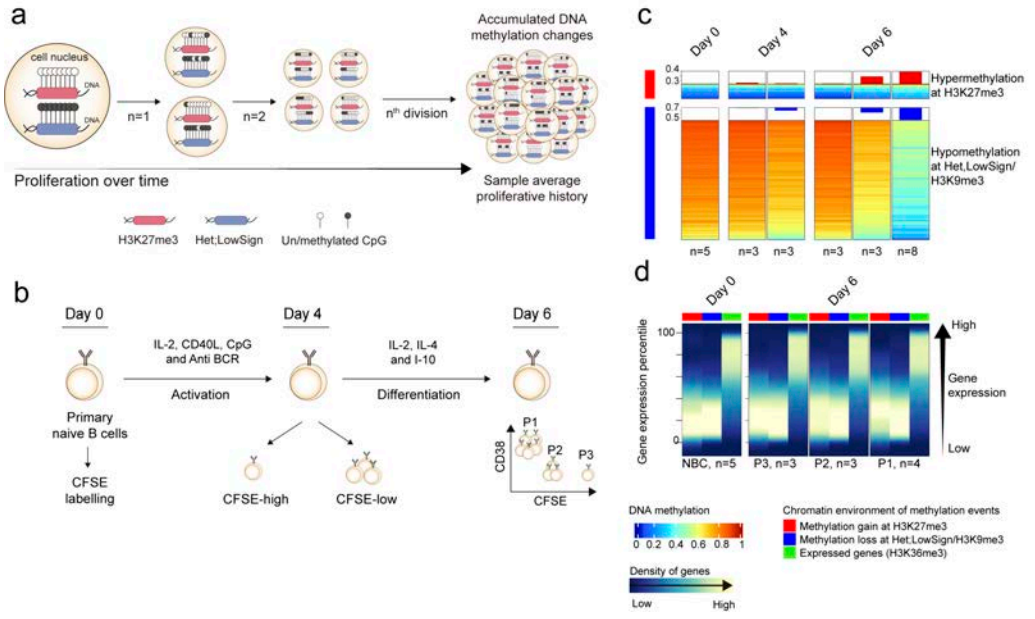


Fig. 24 | DNA methylation changes are accumulated in silent chromatin regions in primary naïve B cells during cell division. a, Hypothetical model derived from previous data from literature of DNA methylation changes taking place during mitotic cell division. **b,** *In-vitro* B-cell differentiation model of human primary naïve B cells to plasmablasts in 6 days. **c,** Heatmaps showing the mean DNA methylation in H3K27me3 and heterochromatic regions of NBC at day 0, 4 and 6 of cells with different levels of proliferative history based on CFSE dilution. **d,** Gene expression percentiles of genes in H3K27me3-marked and heterochromatic regions from panel **c** in NBC at day 0 and day 6 with different proliferative histories.

3.2.2. Development of an epigenetic mitotic clock: the epiCMIT score.

Based on the previous results, I next sought to reduce the DNA methylation signature related to cell division. Thus, I performed a step-wise selection of CpGs whose methylation change would reflect the cell mitotic history (**Fig. 25a**). First, I used primary ChIP-seq data to segment the genome into chromatin states and select CpGs located in constitutively silenced/poised chromatin in normal and neoplastic B cells. Second, I identified CpGs methylated (≥0.9) or unmethylated (≤0.1) in HPCs that extensively lose or gain methylation (a difference of at least 0.5) in bmPCs. This difference was used to capture CpGs undergoing extensive methylation changes between cells in the B-cell lineage with the lowest and highest proliferative histories. Third, I obtained 184 CpGs located at constitutive H3K27me3-containing regions and 1,164 CpGs at constitutive heterochromatin which, based on previous evidence [73,150,313,316,343–345] as well as the

previous *in-vitro* data, gain and lose DNA methylation upon cell division, respectively. Fourth, I constructed two mitotic clocks with these two sets of CpGs, one gaining DNA methylation upon cell division called epiCMIT-hyper (184 CpGs) and one losing DNA methylation upon cell division called epiCMIT-hypo (1,164 CpGs) (**Fig. 25a, b**), with epiCMIT standing for epigenetically-determined Cumulative MIToses. Both mitotic clocks range from 0 to 1 depending on low or high relative proliferative histories, respectively. I next evaluated the relationship of these two mitotic clocks with the total number of DNA methylation changes in normal and neoplastic B cells. I found strikingly high correlations between the total number of hypermethylation changes and epiCMIT-hyper (mean R=0.89, range 0.72-0.95), and with the total number of DNA hypomethylation changes and epiCMIT-hypo (mean R=0.84, range 0.63-0.96) (**Fig. 25c**). These high correlations suggest that the overall DNA methylation landscape of normal and neoplastic B cells is greatly influenced by the accumulated mitotic activity of the cells. To further study the relationship between hyper- and hypomethylation changes related to cell division, I initially evaluated both mitotic clocks in normal B cells and observed an expected but strikingly high correlation between them (R=0.96, p-value<0.001). These B-cell subpopulations were distributed according to their maturation state, and thus according to their accumulated proliferative history during B-cell differentiation and not to the proliferation status of B-cell subpopulations (**Fig. 25d, left panel**). A clear example of this phenomenon are bmPCs, which in spite of being non-proliferative, show the highest epiCMIT and thus the highest proliferative history. The association between the degree of hyper- and hypomethylation supports previous observations in colorectal cancer [346] and indicates that mitotic cell division in normal B cells leaves simultaneously both hyper- and hypomethylated imprints. Although this high correlation between the two mitotic clocks was also observed for MCL, CLL and DLBCL (**Fig. 25d**), it does not seem to be a universal phenomenon, as ALL and MM do not show any correlation. The epiCMIT-hyper was notably greater than the epiCMIT-hypo in ALL samples (up to 5 times), and the opposite scenario was observed in MM (up to 3 times). Upon neoplastic transformation, dividing precursor B cells do not seem to acquire broad hypomethylation in heterochromatin but rather hypermethylation in polycomb-repressed regions. In contrast, neoplastic plasma cells acquire widespread hypomethylation in heterochromatin and virtually lack hypermethylation of polycomb-repressed regions beyond the level observed in normal plasma cells. This distinct tendencies in gaining or losing DNA methylation in ALL and MM can also be observed in our previous analyses, including the differences in the number of *de novo* tsDNAm (**Study**

2, section 2.2.4., Fig. 19) or the distribution of DNA methylation changes in individual patients (Study 2, section 2.2.6., Fig. 22a, c, d). As previously discussed, these differences do not seem to relate to differential expression of DNA methyltransferases among B-cell tumors, as they show a rather similar expression levels in all B-cell tumors (Study 2, section 2.2.4., Fig. 19f). In addition to this previous analysis, I next assessed the epiCMIT-hyper/epiCMIT-hypo ratio as a function of gene expression of the main DNA methyltransferase enzymes, namely *DNMT1*, *DNMT3A* and *DNMT3B* in ALL, CLL and MCL cases that I had DNA methylation and gene expression data for the same patients. The rationale of this analysis is that cases with higher expression levels of *DNMT* enzymes should presumably acquire more hypermethylation during cell division, and thus the epiCMIT-hyper/epiCMIT-hypo ratio above 1. Conversely, ALL cases with lower expression levels of *DNMT* enzymes should show more hypomethylation, and thus the epiCMIT-hyper/epiCMIT-hypo ratio below 1. This analysis however showed that these two variables were neither correlated in ALL as a whole nor in ALL subtypes (Fig. 25e, top panels). Identical results were obtained in CLL and MCL samples (Fig. 25e, middle and bottom panels). In the case of MM samples, concurrent DNA methylation and gene expression data were not available and therefore I performed a complementary analysis using 3 MM cell lines, namely U266, JJN3 and H929. More specifically, I analyzed the epiCMIT-hyper/epiCMIT-hypo ratio of these 3 MM cell lines and compared them to the ratio of bmPC and MM patients (Fig. 25f). This analysis showed that bmPC contain already an epiCMIT-hyper/epiCMIT-hypo ratio below 1, a finding that is exacerbated in MM patients and MM cell lines. Bone marrow plasma cells show a subtle reduction of the main *DNMTs* [50] (Study 2, section 2.2.4, Fig. 19f) which could potentially be behind the epiCMIT-hyper/epiCMIT-hypo ration below 1. However, this certainly cannot be the case for MM patients or MM cell lines, which have been shown to have an upregulation of *DNMT1* of 1.4 and 7.7 compared with normal bone marrow plasma cells, respectively, and in spite of it, they continue losing methylation [347]. Collectively, these analyses again support the notion that the differences in the acquisition of DNA hyper- or hypomethylation changes in B-cell tumors cannot be explained solely by the expression levels of *DNMTs*.

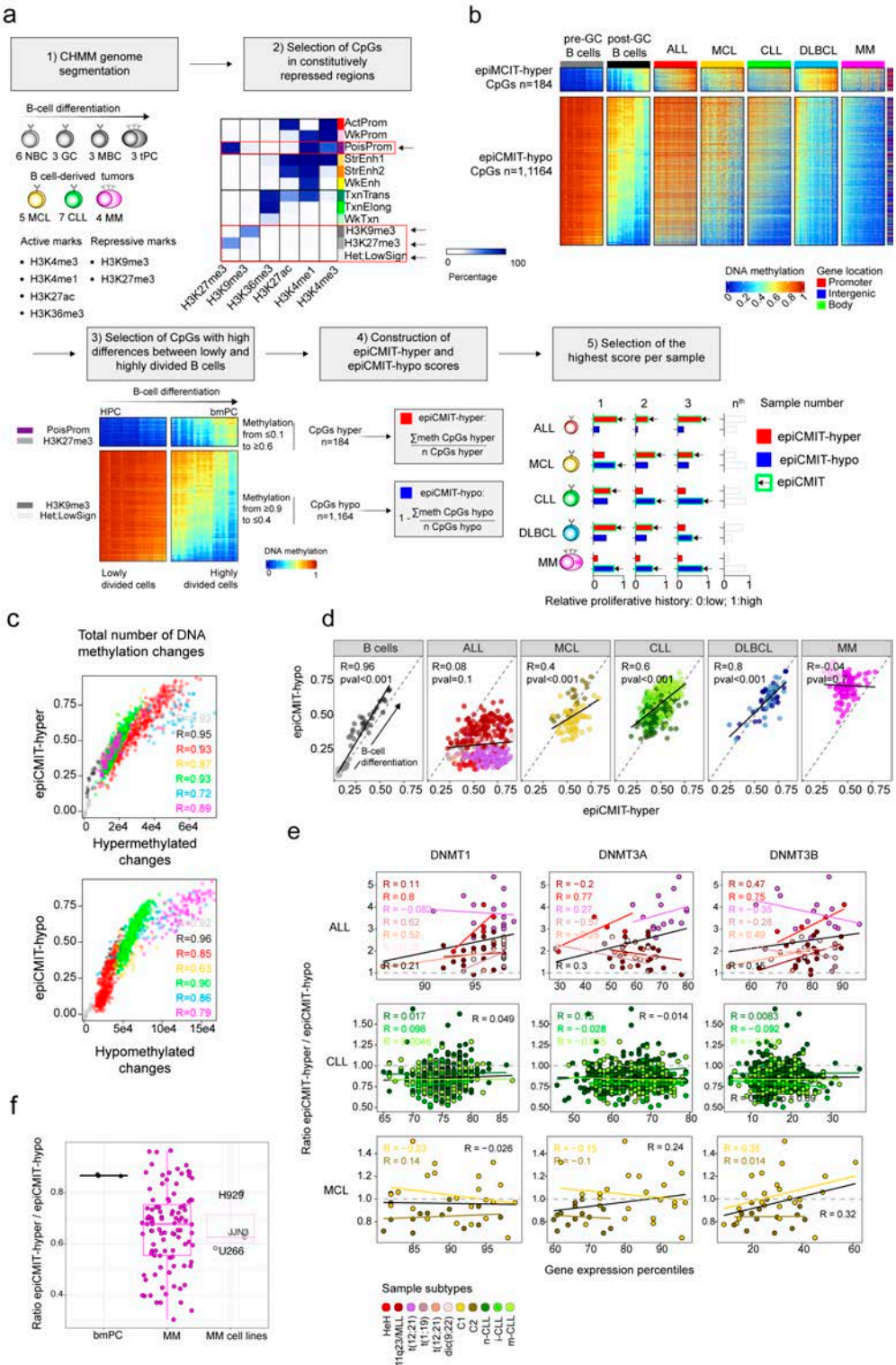


Fig. 25 | See caption on next page.

Fig. 25 | Development of the epiCMIT mitotic clock (see figure on previous page). a, Schematic representation of the developmental steps for the epiCMIT mitotic clock construction. **b,** Heatmap showing the DNA methylation status of CpGs for the epiCMIT-hyper and epiCMIT-hypo mitotic clocks in normal and neoplastic B cell samples. The location of the CpGs is shown on the right. **c,** Correlation between the total number of DNA hyper- and hypomethylated changes with the epiCMIT-hyper and epiCMIT-hypo mitotic clocks in normal and neoplastic B cell samples. **d,** Correlation between the epiCMIT-hyper and epiCMIT-hypo mitotic clocks in normal and neoplastic B cells samples. **e,** Correlation between epiCMIT-hyper/epiCMIT-hypo ratio and the levels of the main *DNMTs* in ALL, MCL and CLL samples with concurrent DNA methylation and gene expression data. **f,** Distribution of the epiCMIT-hyper/epiCMIT-hypo ratio in bmPC, MM patients and 3 MM cell lines with high expression of *DNMT1*.

Finally, to circumvent possible underestimations of the true proliferative history of samples with different tendencies in losing or gaining DNA methylation, I then selected the highest score from the epiCMIT-hyper and epiCMIT-hypo per sample to derive a unique epiCMIT value (**Fig. 25a, Fig. 26a**). The epiCMIT shall then reflect the relative accumulation of mitotic cell divisions of a particular sample, including the mitotic history associated with normal cell development and with malignant transformation and progression. Remarkably, the epiCMIT is able to capture different tendencies in gaining or losing DNA methylation during mitotic cell division, as it respectively happens in ALL and MM patients. Notably, the epiCMIT cannot be affected by a different distribution of cell cycle phases in tumor samples, since the DNA methylome remains rather stable during the whole cell cycle [348].

3.2.3. Validation of the epiCMIT mitotic clock.

The validity and applicability of the epiCMIT as mitotic clock was validated through several perspectives (**Fig. 26a**). First, I experimentally validated it using the previously presented *in-vitro* B-cell differentiation model of primary NBCs into plasma cells [298] (**Results, section 3.2.1, Fig. 24**). This analysis, although a bit redundant, allowed us to confirm that the reduced set of CpGs used to build the epiCMIT was able to precisely trace proliferative histories. As expected, I observed that epiCMIT increased with cell proliferative history both at day 4 and at day 6, in which cells with the lower CFSE concentration (i.e. highest proliferation history) showed the highest epiCMIT values (**Fig. 26b, left panel**). Consistent with the lack of regulatory function of the CpGs associated with cell division and their location in silent chromatin, genes associated to the epiCMIT-CpGs remained silenced in all conditions regardless of the cell phenotype and proliferative history (**Fig. 26b, right panel**). Second, I studied the link between the

epiCMIT and genetic changes using WGS data from 138 CLL and 70 MCL patients [229,245]. In particular, I measured the activity of know mutational processes through the analysis of single base substitution (SBS) signatures [319], as recently described [349]. I observed significant correlations between our epiCMIT and the mutational signatures SBS5 and SBS1, which have been previously described as mitotic-like mutational processes [350] (**Fig. 26c, d**). I also identified a significant link between the epiCMIT and the non-canonical AID signature (ncAID, SBS9) [247,319] in IGHV mutated CLL and MCL cases with ncAID footprint [229], possibly reflecting rounds of cell division during the BCR affinity maturation in the germinal center of the ancestor B cell prior to its malignant transformation to CLL or MCL (**Fig. 26e**). Third, although the epiCMIT reflects the cell proliferative history, a relationship between epiCMIT and cell proliferation is expected in tumors (more proliferative history implies higher proliferation, although it also depends on time). Accordingly, MCL cases showing higher Ki-67 (a proliferation marker) also had higher epiCMIT than cases with lower Ki-67 expression (**Fig. 26f**). Furthermore, a gene set enrichment analysis in CLL cases with high and low epiCMIT revealed that cases with high epiCMIT showed higher expression of genes related with cell proliferation and MYC activity (**Fig. 26g**). This is remarkable, as leukemic CLL samples are not considered proliferative, but the epiCMIT score is able to distinguish CLL patients with higher expression of genes related to proliferation. Thus, these data suggest that cases with higher proliferative history also seem to have higher proliferative capacity at the time of sampling. Fourth, I performed a benchmarking of the epiCMIT with all the mitotic clocks previously reported in the literature (**Methods Study 2, section 2.13**). I compared the epiCMIT with two previously reported hypermethylation-based mitotic clocks called epiTOC and MiAge [313,314]. In addition, I calculated a new hypomethylation-based mitotic clock using a previously defined pan-cancer set of CpGs losing DNA methylation upon cell division called PMDsoloWCGW CpGs [316]. Focusing on hypermethylation-based mitotic clocks, the epiCMIT showed excellent correlations with epiTOC and MiAge in B-cell neoplasms that tend to acquire polycomb-related hypermethylation (e.g. mostly ALL, but also DLBCL and MCL); a moderate correlation in the case of CLL, which acquires more hypo- than hypermethylation, and a total lack of correlation in the case of MM, which mostly loses DNA methylation (**Fig. 26h upper panels**). Interestingly, identical observations were obtained comparing the epiCMIT and the widely-reported CpG island methylator phenotype (CIMP) in human cancer [138] calculated as previously proposed [315]. Therefore, this analysis also reveals that the pan-cancer CIMP score in reality may represent a measure of the cell mitotic history. Interestingly, the opposite scenario was

found when comparing epiCMIT with the hypomethylation-based mitotic clock PMDsoloWCGW. I observed excellent correlations between epiCMIT and PMDsoloWCGW in tumors with extensive DNA hypomethylation (mostly MM and CLL, but also MCL and DLBCL) and a null correlation in ALL, where PMDsoloWCGW was poorly modulated (**Fig. 26h, bottom panel**). In spite of these striking discrepancies in ALL and MM, mitotic clocks were in general highly correlated in normal and neoplastic B cells, even though the generally poor overlap of their underlying CpGs (**appendix, manuscript 4**), indicating that cell proliferative history can be accurately traced with different sets of CpGs. Remarkably, epiCMIT outperformed all mitotic clocks to identify cells with different proliferative histories using the controlled setting of the in vitro B-cell differentiation model (**appendix, manuscript 4**), a finding that suggests its higher accuracy to trace cell divisions in B cells. In addition to these correlations with other mitotic clocks, and in line with the previous analysis with epiCMIT-hyper and epiCMIT-hypo, I observed that epiCMIT is highly correlated with the total number of DNA methylation changes accumulated in each sample since the HPC stage, suggesting that the overall DNA methylation landscape seems to be strongly influenced by the cell proliferative history (**Fig. 25h bottom right**). When interpreting the epiCMIT in tumor samples, it is worth clarifying that it represents the total cumulative proliferative history of samples. This is of paramount importance when interpreting the epiCMIT score in B-cell tumors, which contains the proliferative history related to the ancestor normal B cell, and the proliferative history associated with malignant transformation and progression (**Fig. 26i**). For example, ALL derives from precursor B cells which have a low proliferative history (**Fig. 26a**). Then, during tumor progression ALL acquire both B-cell related and B-cell independent changes in silent chromatin (**Study 2, section 2.2.6, Fig. 22 and Fig 23**), which collectively are captured by the epiCMIT score. From this perspective, ALL has similar DNA methylation changes compared with bmPC due to its extensive proliferative history targeting silent regions that become modulated in normal B-cell differentiation (**Study 2, section 2.2.6, Fig. 21a, c**). Thus, it is imperative to compare the relative epiCMIT across tumor samples with a similar cellular origin, otherwise the different proliferative history of the ancestor B cell could influence the readout of the epiCMIT score and lead to misleading associations (**Fig. 26i)**. Finally, as a last step to proof the applicability of the epiCMIT as a mitotic clock in B-cell tumors, it is logical to assume that it should increase during disease progression (**Fig. 26j**). To address this, I compared the epiCMIT in two paradigmatic transitions between two precursor conditions and overt cancer, including the monoclonal gammopathy of undetermined

significance (MGUS) and MM, as well as the monoclonal B cell lymphocytosis (MBL) and CLL categorized according to their cellular origin. This analysis showed an overall lower epiCMIT in precursor lesions as compared with overt cancer (**Fig. 26j, upper panels**), as would be expected due to their increased leukemia-specific proliferative history. In line with this finding, the epiCMIT increased in paired CLL samples at diagnosis and progression before treatment as well as in sequential ALL samples at diagnosis, first relapse and second relapse (**Fig. 26j, lower panels**). These results suggest that the epiCMIT evolves together with clinical progression.

Collectively, all these lines of evidence strongly suggest that the epiCMIT represents a relative measure of the number of cumulative mitotic cell divisions that normal and neoplastic B cells have undergone since the uncommitted hematopoietic cell stage. Remarkably, the epiCMIT may represent a more universal mitotic clock than previously reported mitotic clocks exclusively based on hyper- or hypomethylation, since it is able to capture the proliferative history of neoplasms regardless of their tendency to gain or lose DNA methylation during mitotic cell division.

Fig. 26 | Validation of the epiCMIT mitotic clock (see figure on next page). a, The epiCMIT mitotic clock in normal and neoplastic B cells samples. **b,** Experimental validation of the epiCMIT mitotic clock in an independent *in-vitro* model of B-cell differentiation of human primary naïve B cells to plasmablasts in six days. **c, d** Correlation between the epiCMIT and the mitotic-like mutational signatures SBS1 and SBS5 in CLL and MCL patients with available WGB. **e,** Correlation between epiCMIT and the non-canonical AID mutational signature in CLL and MCL. **f,** MCL patients with high epiCMIT values also show higher Ki67 staining levels. **g,** Gene set enrichment analysis (GSEA) in n-CLL samples with low and high epiCMIT (15 and 85% percentiles, respectively). On the left, heatmap showing the gene of to the most significant gene expression signatures related to proliferation and MYC signaling. On the right, example GSEA plots of some representative gene expression signatures from the heatmap. **h,** Correlation between epiCMIT and previously reported hypermethylation-based mitotic clocks epiTOC and MiAge as well as the pan-cancer CIMP score, the PMDsoloWCGW CpGs and the total number of DNA methylation changes in normal and neoplastic B cells samples. **i,** epiCMIT interpretation in normal neoplastic B cells, **j,** epiCMIT in precursor conditions as compared to overt cancer in MM and CLL, and in trios of ALL at diagnosis, first and second relapses.

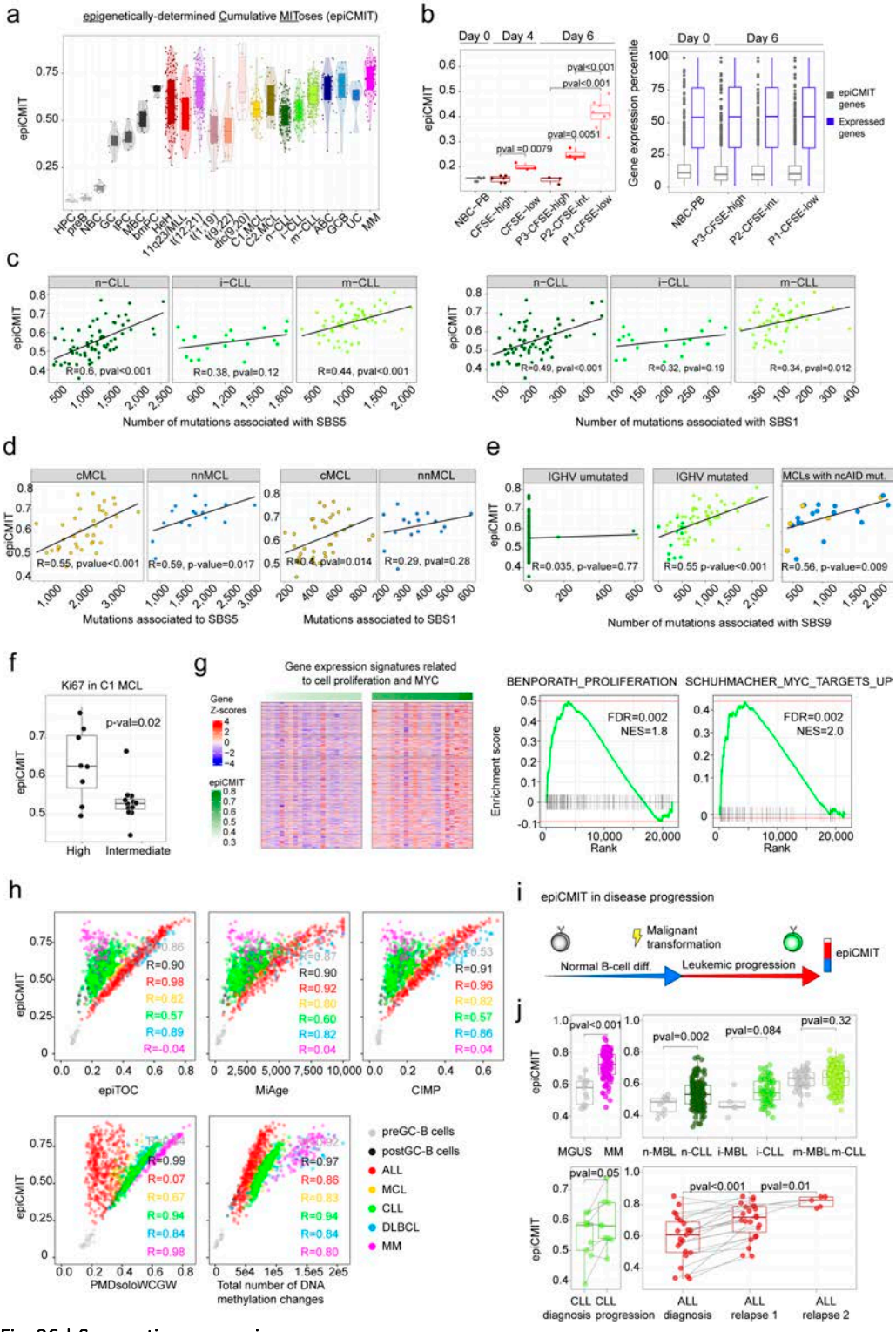


Fig. 26 | See caption on previous page.

3.2.4. epiCMIT represents a strong prognostic factor in B-cell tumors.

Based on my previous observations, I next wondered whether the epiCMIT could be useful to predict the clinical behavior of B-cell neoplasms. I analyzed the relationship of epiCMIT with clinical outcome in specific B-cell tumor subtypes based on chromosomal alterations (i.e. ALL) or cell of origin (i.e. MCL, CLL and DLBCL), and thus having a similar ground state B-cell specific/cell of origin proliferative history (**Fig. 26i**). In ALL, high epiCMIT was consistently associated with longer overall survival (OS), OS after relapse (**Fig. 27a, b**) and relapse-free survival (RFS) (**appendix, manuscript 4**) of the patients and thus with better clinical outcome. These epiCMIT associations maintained an independent statistical significance from the well-established ALL cytogenetic groups as prognostic variable in RFS and OS, and a marginal significance in OS after relapse. In contrast to ALL, the opposite clinical scenario was observed in mature B-cell neoplasms. In each of the CLL subtypes with different cellular origin, a high epiCMIT was strongly associated with a worse prognosis using time to first treatment (TTT) as end-point variable, both from sampling time (**Fig. 27c**) and in cases whose sample was obtained close to diagnosis (**appendix, manuscript 4**). Additionally, the epiCMIT as continuous variable showed a highly significant independent prognostic impact in the context of the main and well-established prognostic factors in CLL, including the IGHV status and *TP53* alterations (deletion and mutation) (**appendix, manuscript 4**). Overall, considering genetic and epigenetic variables together, it seems that the epiCMIT, CLL epigenetic subgroups [187,254,258] , and genomic complexity measured by the total number of driver alterations [245,246] are the most significant independent variables associated with prognosis in CLL (**Fig. 27c**). In addition, despite the variability of treatments in our initial CLL series, the epiCMIT showed marginal significance in OS (**appendix, manuscript 4**). All these findings were widely confirmed in an additional series of 210 CLLs provided by Prof. Thorsten Zenz from Zurich, which were treated mainly with chemo-immunotherapy (**Fig. 27d**). In the case of MCL, the epiCMIT showed an independent prognostic impact from the two cell-of-origin subtypes, namely C1 and C2, an observation that was confirmed in an extended series in the more aggressive and prevalent C1 group (**Fig. 28a, b**). Furthermore, in the complete series used for validation I also had WGS or WES data available to evaluate the epiCMIT score in relationship with the genetic alterations predisposing to an adverse clinical outcome (n=70 cases, 58 non-overlapping cases with the initial series). Using the whole series, I also showed epiCMIT as a strong and independent prognostic factor in a multivariate Cox model with copy

number alterations (CNA) and breakage-fusion bridge (BFB) cycles [229](please, be referred to **Study 4, Results section 4.2.3** for further details). Lastly, although the sample size was limited and requires further studies, our data suggest that high epiCMIT could also represent a poor prognostic variable within the two cell-of-origin DLBCL subtypes (**appendix, manuscript 4**). Finally, the epiCMIT score showed, as compared with all the other DNA methylation-based mitotic clocks, an overall superior prognostic value in the B-cell tumors with the largest number of patients, further supporting its value beyond previously developed mitotic clocks (**appendix, manuscript 4**). Collectively, these results suggest that an increased proliferative history of the neoplastic clone in precursor B-cell neoplasms at diagnosis predicts for a better disease-free survival. In sharp contrast, in mature B-cell neoplasms, which are overall less proliferative than ALL, neoplastic clones with high proliferation history seem to predict for future proliferative capacity and consistently show worse clinical outcomes.

Fig. 27 | epiCMIT as independent prognostic variable in ALL and CLL (see figure on next page). a, b Kaplan Meyer curves for ALL patients with low or high epiCMIT values using maxstat statistic for overall survival (OS) **(a)** or OS after relapse **(b)**. epiCMIT retains statistical significance in multivariate Cox model from the well-established ALL cytogenetic groups for OS and marginal significance for OS after relapse. **c,** Kaplan Meyer curves for CLL subtypes divided into low or high epiCMIT based on maxstat statistic. epiCMIT retains statistical significance in multivariate Cox model with well-established clinical variables. **d,** CLL validation series from **c**.

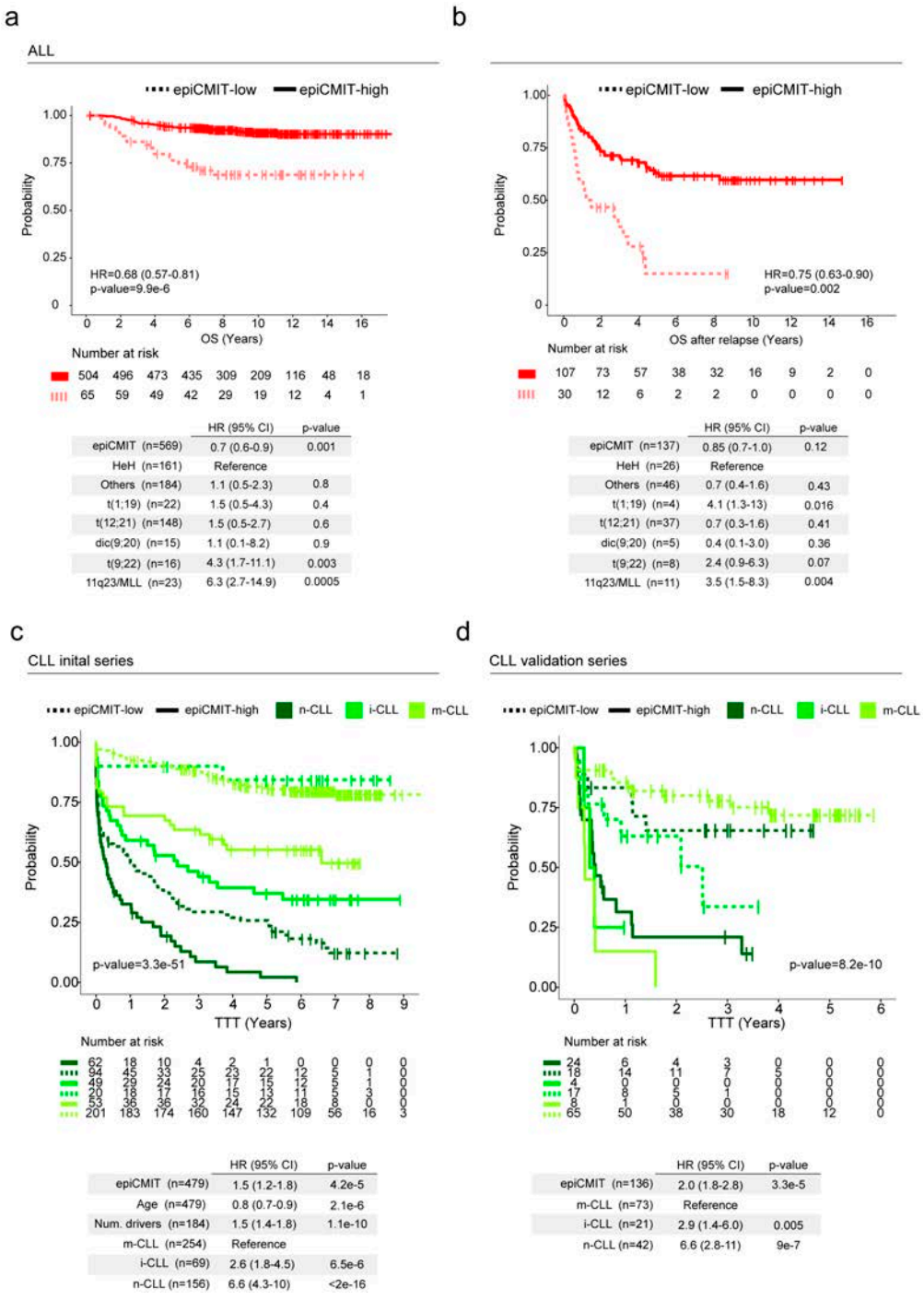


Fig. 27 | See caption on previous page.

In addition to these important prognostic implications of the epiCMIT in B-cell tumors, I next wondered whether it could represent a more general and perhaps a pan-cancer mitotic clock with prognostic impact for other human neoplasias. Although I developed the epiCMIT score with epigenetic data from normal and neoplastic B cells, I reasoned that the stringent filtering criteria used for the epiCMIT-CpGs could make the epiCMIT score cell type-independent. This notion is supported by the overall good correlation between epiCMIT and the other previously developed mitotic clocks, which were not biased towards any particular cell type (**Results, section 3.2.3**) [313–316]. These correlations were even higher for the epiCMIT-hyper and epiCMIT-hypo compared with hyper- and hypomethylation clocks separately, which showed a mean correlation of R=0.92 (range 0.62-0.99) and R=0.97 (range=0.87-0.99), respectively (**appendix, manuscript 4**). To test my hypothesis, I next downloaded DNA methylation data from 15 distinct TCGA tumor types (n=3995) (https://portal.gdc.cancer.gov/repository) and calculated the epiCMIT and all the other mitotic clocks to assess their prognostic value. Overall, I found that epiCMIT correlated well with all mitotic clocks in the TCGA tumors analyzed, and that in 12 out of the 15 TCGA tumor types analyzed mitotic clocks were able to predict patient clinical outcome (**Fig. 28c**). Remarkably, in the majority of tumor types higher values of mitotic clocks were associated with an adverse patient clinical outcome, whereas in few of them high clock values were associated with better patient clinical outcome, including acute myeloid leukemia (TCAG-LAML), lung squamous cell carcinoma (TCGA-LUSC) and glioblastoma multiforme (TCGA-GBM) (**Fig. 28c**). These findings support my previous observations in B-cell tumors, where epiCMIT was associated with better survival in ALL and worse clinical outcomes in MCL, CLL and DLBCL.

Altogether, these data indicate that the epiCMIT may indeed represent a pan-cancer mitotic clock showing prognostic value in human neoplasias other than B-cell tumors.

Fig. 28 | epiCMIT as prognostic variable in MCL and TCGA human neoplasias (see figure on next page). **a, b** Kaplan Meyer curves for MCL subtypes with low or high epiCMIT values using maxstat statistic for overall survival (OS) for initial **(a)** and validation series **(b)**. epiCMIT show independent prognostic impact from the cell of origin C1 and C2 MCL subtypes and from age. **c,** Prognostic value of epiCMIT and the other mitotic clocks in 15 human neoplasias from the TCGA dataset. P-value<0.001***; p-value<0.01**; p-value<0.05*; p-value<0.1·.

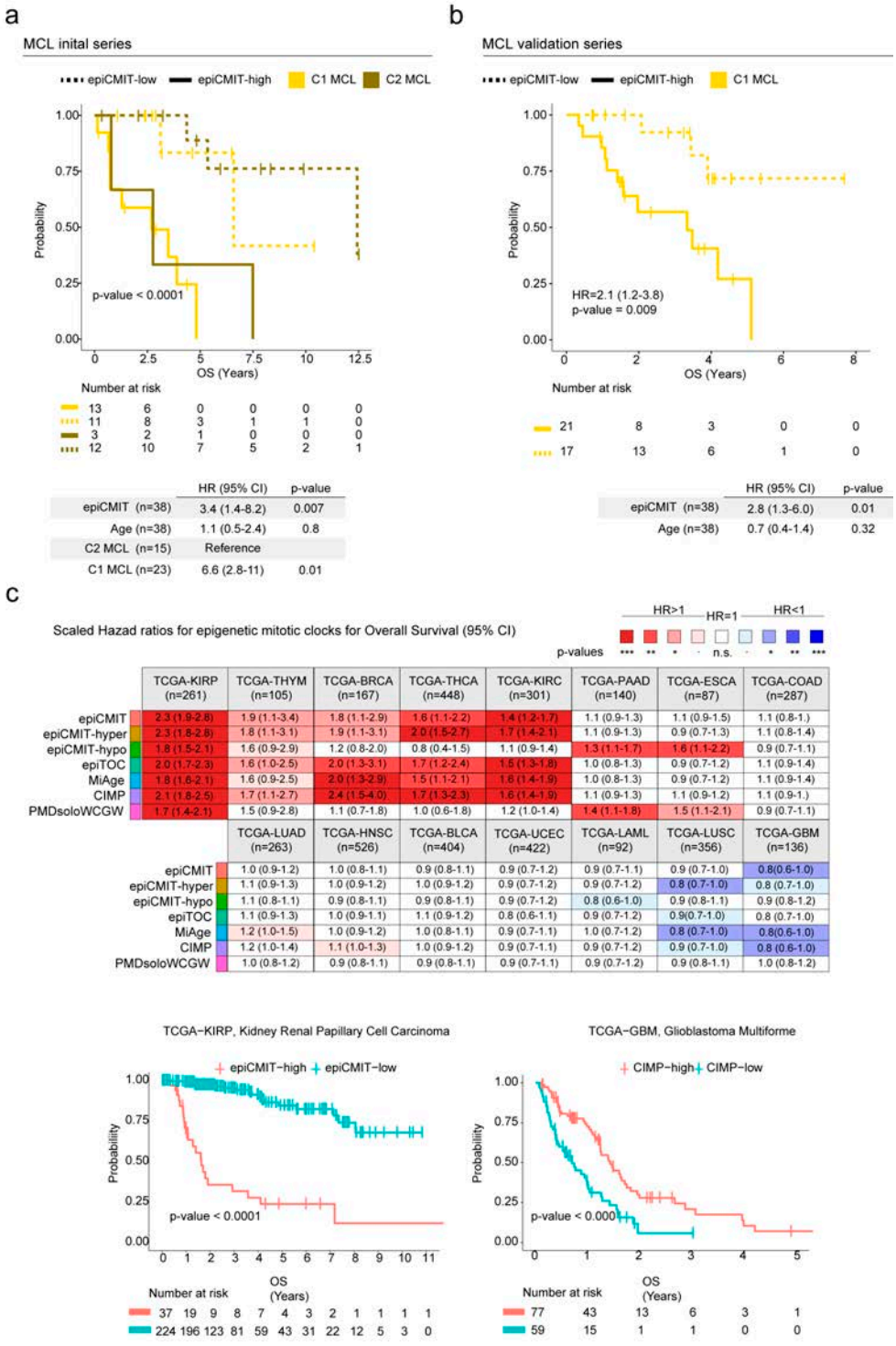


Fig. 28 | See caption on previous page.

3.2.5. epiCMIT and its relationships with the aging process.

Despite the data presented above, a potential confusing aspect related to epiCMIT is the fact that DNA methylation in polycomb-repressed regions and heterochromatin has also been reported to be modulated during physiological aging [317,351–353]. In fact, this phenomenon has led to the successful use of DNA methylation data to accurately predict human chronological age [341,342,354]. The most widely used epigenetic chronological clock is the Horvath's clock, which has been shown to represent an accurate multi-tissue age predictor [317]. To study the potential relationship between mitotic activity and the aging process, I first analyzed the epiCMIT in normal B cells with low (NBC) and high (MBC) epiCMIT values in samples from infants, young adults and elderly donors (**Fig. 29a, left**). I found similar epiCMIT levels in the same B-cell subtypes regardless of donor's age, indicating that the epiCMIT reflects the proliferative history. Although the sample size is low, apparently this analysis did not reveal any evidence linking the epiCMIT with the chronological age of healthy donors, which indeed is accurately predicted by the Horvath's aging clock (**Fig. 29a left-middle panel**). In the case of B-cell tumors, I observed the same general tendency. Although the ALL samples studied in this thesis are from pediatric patients and thus show a very low age span, they display the highest epiCMIT range, and thus I extrapolate that the association between epiCMIT and age in this disease is negligible. In DLBCL I observed a similar scenario, since 30 and 90-year-old patients showed similar epiCMIT levels. Only in MCL and CLL patients I detected minor correlations between epiCMIT and patient's age (**Fig. 29a, right**). I then applied the Horvath's clock to patient samples and, as previously shown in other cancers [317], I found significant epigenetic age acceleration, with some pediatric ALL patients reaching an impressive predicted age over 200 years. Interestingly, I found that the epiCMIT shows a highly significant correlation with the increased epigenetic age predicted by Horvath's clock in the majority of B-cell tumors subtypes ($R=0.62$, p-value<2e16). I believe that this is a remarkable finding, as it suggests that the epigenetic age acceleration previously seen in cancer patients may be related to the increased cell proliferation of cancer cells (**Fig. 29a, bottom**). In spite of this intriguing correlation that deserves further investigation, the epiCMIT and Horvath's clocks seem to be targeting different molecular features. To study in depth these two epigenetic clocks, I divided the epiCMIT-CpGs into those gaining and losing DNA methylation upon cell division, namely the epiCMIT-hyper and epiCMIT-hypo CpGs. In the case of the Horvath CpGs, I divided those CpGs positively (gain of DNA methylation) and negatively (loss of DNA methylation) associated with

aging [317]. In addition, I further subdivided these CpG categories into those with CpGs extensively modulated during B-cell differentiation and those showing a greater modulation during malignant transformation and progression, i.e. B-cell related and B-cell independent CpGs. Interestingly, the majority of Horvath-CpGs were classified as B-cell independent CpGs, indicating that subtle DNA methylation changes are accumulated during aging, which are mathematically captured by the Horvath clock. First, Horvath and epiCMIT clocks do not share any CpG in common, and their corresponding CpGs showed distinct genomic enrichments (**Fig. 29b, c**). The epiCMIT-hypo CpGs were abundant at low CpG content regions, whereas epiCMIT-hyper CpGs were prominent at promoter-related regions. Conversely, Horvath-CpGs were generally distributed at shores and promoter-related regions regardless of their positive or negative association with age and their B-cell related or B-cell independent classification (**Fig. 29c**). Second, Horvath-CpGs and epiCMIT-CpGs showed distinct DNA methylation dynamics in normal and neoplastic B cells. epiCMIT-CpGs are by definition greatly modulated in normal B cells and also in B-cell tumors. In contrast, the majority of CpGs used to predict chronological age in the Horvath model do not significantly change (at least with the classical parameters of 0.25 difference) in normal and neoplastic B cells (**Fig. 29d**). Third, epiCMIT and Horvath CpGs were enriched in distinct chromatin states (**Fig. 29e**). By definition, epiCMIT-CpGs gaining DNA methylation are enriched at H3K27me3 regions, whereas epiCMIT-CpGs losing DNA methylation are enriched at low signal heterochromatin and H3K9me3-repressed regions. Conversely, Horvath CpGs were generally enriched at regulatory regions including enhancers and promoters in normal and neoplastic B cells. In addition to this general enrichment, I could detect that CpGs showing a positive association with age also displayed enrichments in H3K27me regions, with those B-cell related associated with poised promoters and H3K27me3 regions, and with B-cell independent CpGs enriched at poised promoters but not in H3K27me3 regions (**Fig. 29e**). Thus, this shared enrichments for H3K27me3 regions for the epiCMIT-hyper and Horvath CpGs positively related with aging may suggest that the DNA methylation changes occurring in cell division and aging share in part the same chromatin environment. Interestingly, CpGs undergoing DNA methylation loss in the Horvath model were not enriched in heterochromatin. Finally, I analyzed the transcriptional levels of genes containing epiCMIT and Horvath CpGs. As previously showed in our *in-vitro* system, genes containing epiCMIT-CpGs were silent both in normal and neoplastic B cells, whereas genes containing Horvath CpGs were generally

expressed in normal and neoplastic B cells, with genes containing CpGs negatively associated with age showing the highest expression.

Altogether, all these data indicate that despite mitotic and aging processes may share some epigenetic features, epiCMIT and Horvath epigenetic clocks are in general related to distinct biological phenomena, with epiCMIT capturing the proliferative history of cells while the Horvath clock reflects the chronological age.

Fig. 29 | Relationship between the epiCMIT mitotic clock and Horvath chronological clock (see figure on next page). a, epiCMIT and Horvath clock in normal and neoplastic B cells. On left, epiCMIT and Horvath clocks in NBC and MBC from infants, young adults and elderly people. Middle, epiCMIT and Horvath clocks in MCL, CLL and DLBCL patients. On right, correlation between the epiCMIT and Horvath clocks in all B-cell tumors. **b,** Null overlap between epiCMIT and Horvath CpGs. epiCMIT CpGs are divided into the those gaining or losing DNA methylation during cell division, namely epiCMIT-hyper and epiCMIT-hypo, respectively. Horvath clock CpGs are divided according to their correlation with age, i.e. positive or negative, gain or loss of DNA methylation, respectively, and B-cell related or B-cell independent. These categories were maintained during all subsequent analyses. **c,** Genomic location for epiCMIT and Horvath CpGs. **d,** epiCMIT and Horvath CpGs modulation in normal and neoplastic B cells. **e,** Chromatin state enrichment for epiCMIT and Horvath CpGs over the 450k background. **f,** Gene expression percentiles of genes containing any epiCMIT or Horvath CpGs.

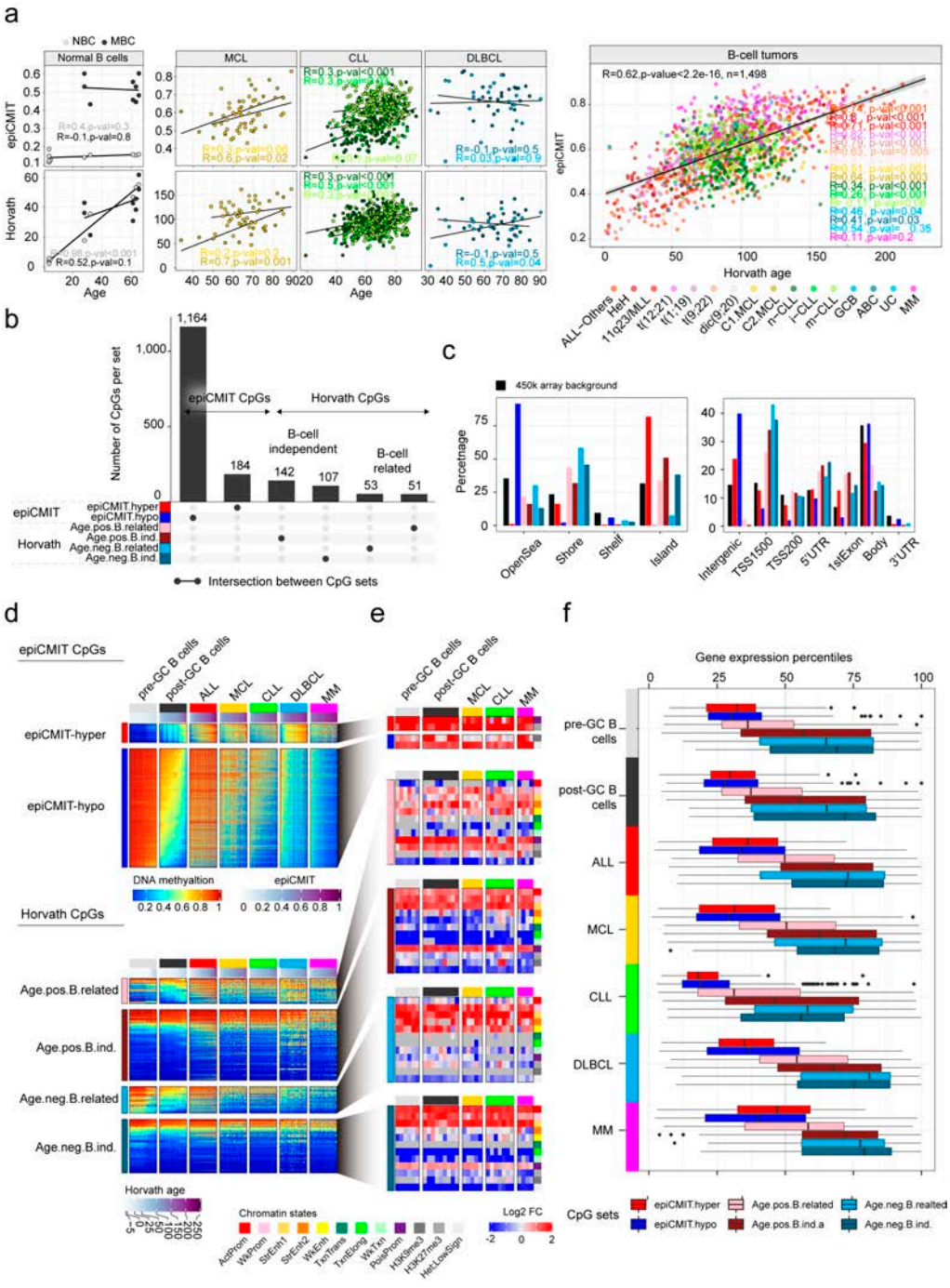


Fig. 29 | See caption on previous page.

3.3 Conclusions.

- Hypermethylation of H3K27me3 regions and hypomethylation of heterochromatic regions account for most of the epigenetic changes associated with inter-patient DNA methylation variability, do not have an impact on gene expression and are related to the proliferative history of the cells.

- Hypermethylation and hypomethylation associated with cell division are in general highly correlated in normal and neoplastic B cells, although acute lymphoblastic leukemia cells are biased towards hypermethylation and multiple myeloma cells towards hypomethylation.

- The proliferative history of normal and neoplastic B-cells can be captured by the thoroughly validated epiCMIT score, which considers both hypermethylation and hypomethylation, and represents a relative measure of the accumulated proliferative history.

- epiCMIT represents a strong independent prognostic variable in ALL, MCL, and CLL, being associated with good clinical outcome in ALL and with an adverse clinical outcome in CLL and MCL.

- epiCMIT may represent a pan-cancer mitotic clock with significant prognostic value in other human neoplasias.

- DNA methylation changes associated with mitotic and aging processes share some molecular features, although the epiCMIT clock mainly captures the proliferative history of samples, while the Horvath aging clock estimates the sample chronological age.

Study 4. Interplay between DNA methylation and genetic alterations in CLL and MCL.

The results of this study have been partly published (**appendix, manuscripts 4 and 5**), while others are in preparation.

 # Duran-Ferrer M, Clot G, Nadeu F, Beekman, R, Baumann, T, Nordlund, J, Marincevic-Zuniga, Y, Lönnerholm G, Rivas-Delgado, A, Martín, S, Ordoñez, R, Castellano, G, Kulis, M, Queirós, A, Seung-Tae, L, Wiemels, J, Royo, R[7], Puiggrós, M, Junyan, L, Giné, E, Beà, S, Jares, P, Agirre, X, Prosper, F, López-Otín, C, Puente, XS, Oakes, CC, Zenz, T, Delgado, J, López-Guillermo, A, Campo, E, **#**Martín-Subero, J.I. The proliferative history shapes the DNA methylome of B-cell tumors and predicts clinical outcome. Nature Cancer (in press). (**#** Corresponding authors).

*****Nadeu F, *****Martin-Garcia D, Clot G, Díaz-Navarro A, **Duran-Ferrer M**, Navarro A, Vilarrasa-Blasi R, Kulis M, Royo R, Gutiérrez-Abril J, Valdés-Mas R, López C, Chapaprieta V, Puiggròs M, Castellano G, Costa D, Aymerich M, Jares P, Espinet B, Muntañola A, Ribera-Cortada I, Siebert R, Colomer D, Torrents D, Gine E, López-Guillermo A, Küppers R, Martin-Subero I, Puente XS, **#**Beà S & **#**Campo E. Genomic and epigenomic insights into the origin, pathogenesis and clinical behavior of mantle cell lymphoma subtypes. Blood. 2020 Jun 25:blood.2020005289. doi:10.1182/blood.2020005289. Epub ahead of print. PMID: 32584970. (***** Contributed equally, **#** Corresponding authors).

4.1. Introduction.

Although MCL is genetically more complex than CLL, both diseases present a highly heterogeneous genetic landscape, with abundant driver genetic alterations, including mutations and CNA, in relatively low percentage of the patients [228,245,246]. In addition to this, both MCL and CLL show an altered DNA methylome compared to normal B cells [254,295]. Early studies showed that the CLL DNA methylome remains fairly similar over time [355], although DNA methylation evolution can also be observed during CLL progression upon the acquisition of particular genetic alterations [255]. This DNA methylation evolution was defined as the number of DNA methylation changes between time points in serial CLL samples and seemed to be associated to inferior patient clinical outcome. A subsequent study from my host group using WGS and WES reported differential enrichment of driver genetic alterations within the three CLL epigenetic subtypes [245] initially identified in 2012 [254]. Besides, a correlation between the number of DNA methylation changes and the total number of mutations detected by WGS could be identified, but the underlying mechanisms for these correlations remained widely unknown. In the case of MCL, as shown in the **Study 1**, a significant correlation between the number of DNA methylation changes and the presence of genetic alterations in patients from the C2 MCL subtype was observed (**section 1.2.5., Fig 14g**). Despite all these findings, how genome-wide genetic and epigenetic aberrations may cooperate in the pathogenesis of MCL and CLL remains poorly characterized. In the present chapter of my doctoral thesis, I extensively mine DNA methylation and mutational data obtained from 490 CLL and 70 MCL patients to shed light into the association between genetic and epigenetic changes.

4.2. Results.

4.2.1. Establishing the basis for an (epi)genetic integration in CLL and MCL.

Before analyzing in depth the relationship between genomic and epigenomic aberrations in CLL and MCL, I assessed if the presence of any particular mutation in epigenetic genes could affect the downstream analyses. It is also worth mentioning that, as compared to other hematological neoplasms, epigenetic genes are in general rarely mutated in CLL and MCL, and are not involved directly to DNA methylation homeostasis such as *DNMT* and *TET* families [228,229,245,246]. These genes are *ZMYM3*, *ASXL1*, *SETD2*, *ARID1A*, *SETD1A*, *HIST1H1B*, *SYNE1*, *CHD2* and *MLL2* in CLL, and *KMT2D*, *SP140*, *NSD2*, *SMARCA4*, SMARCB1 and SYNE1 in MCL. With few exceptions, such as *KMT2D* and *NSD2* in MCL that respectively reach approximately 20% and 10% of the patients [229], these epigenetic genes individually show a low recurrence. Thus, I examined whether the presence of any of them could influence the DNA methylome in CLL and MCL subtypes, namely n-CLL i-CLL and m-CLL and C1 MCL and C2 MCLs [254,295]. An unsupervised PCA analysis showed that patients with mutations in any epigenetic gene were distributed evenly across the PCA space (**Fig. 30a**). This was also the case analyzing each epigenetic gene separately. These analyses suggest that the presence of mutations in epigenetic genes do not seems to alter to a significant extend the DNA methylation profile of CLL and MCL patients. Instead, the main source of DNA methylation variability in both CLL and MCL is related to their cellular origin and proliferative history (**Fig. 30b, c**). In the case of CLL, both the cellular origin and the proliferative history were mixed in the first two principal components of a PCA analysis (**Fig. 30b**), whereas in MCL the cellular origin was prominent in the first component while the proliferative history in the second component (**Fig. 30c**). In the case of MCL, I would like to emphasize here that the previous classification into C1 and C2 from the **Studies 1-3** [295] is highly concordant with the last WHO classification of MCL [164], showing C1 and C2 MCL a great overlap with cMCL and nnMCL, respectively [229] (**Fig. 30d**). In summary, the main sources of DNA methylation variability in CLL and MCL are the tumor cellular origin and proliferative history, and where used for all subsequent genomic integration analyses.

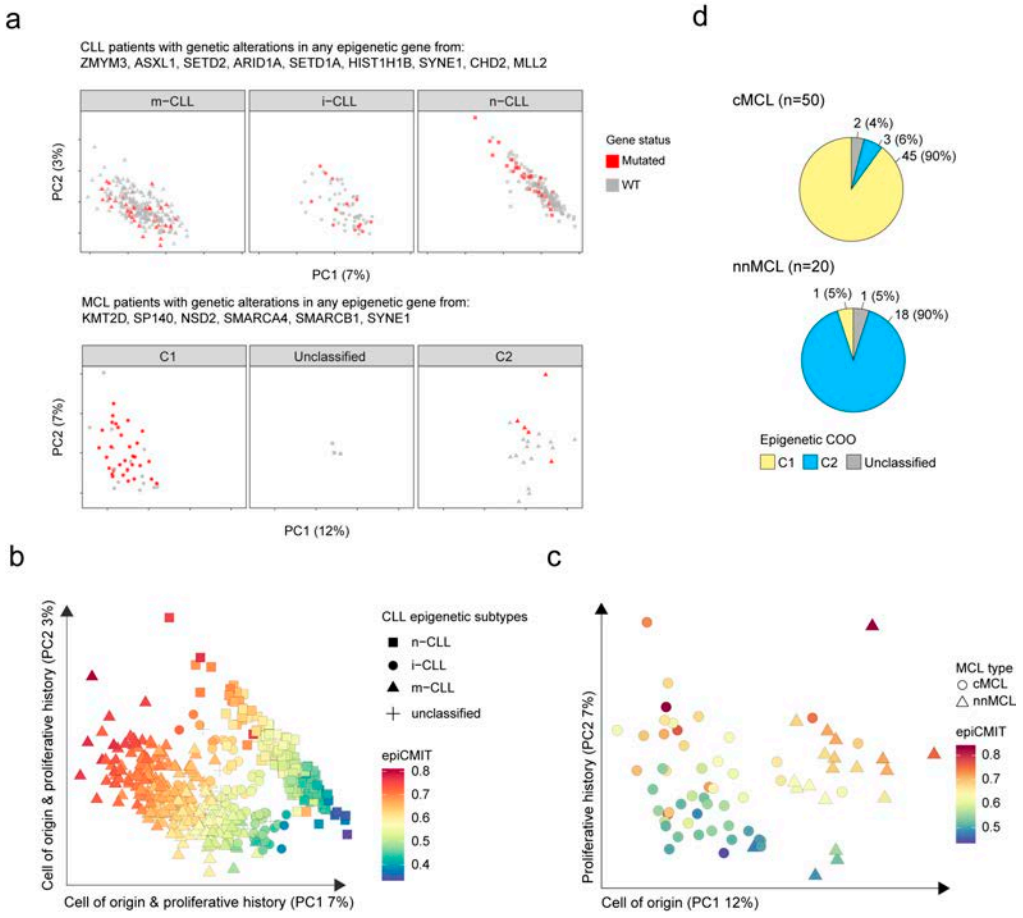


Fig. 30 | The tumor cell of origin and proliferative history are the main sources of DNA methylation variability in CLL and MCL. a, PCA analysis of the whole DNA methylome of CLL and MCL patients with patient showing mutations in any epigenetic gene highlighted. **b,** PCA analysis of the whole CLL 450k DNA methylome with the epiCMIT levels displayed for each individual patient with CLL subtypes differentiated. **c,** PCA analysis of the whole MCL DNA methylome with the epiCMIT levels displayed for each individual MCL patient with cMCL and nnMCL MCL subtypes differentiated. **d,** Overlap between C1 and cMCL and C2 MCL and nnMCL subtypes, respectively. The Pan B-cell tumor classifier from **Study 2** was used to classify patients.

4.2.2. (Epi)genomic crosstalk in CLL.

4.2.2.1. (Epi)genomic crosstalk related to CLL cellular origin.

I reanalyzed CLL genetic data from my host group [245] considering also driver alterations described by Landau and colleagues [246]. Altogether, 89 driver alterations where considered in the context of the three epigenetic CLL subtypes [254,258] as well as each patient proliferative history measured by our epiCMIT score (**Fig. 31a**). As previously described, CLL subgroups showed statistically different enrichments in different genetic alterations, such as *NOTCH1*, *ATM* or del(11q) in n-CLL and *SF3B1* or *MYD88* in i-CLL. Interestingly, the newly identified non-coding driver U1 spliceosomal RNA [318] was almost exclusively mutated in the bad prognostic group n-CLL. Furthermore, I found that CLL subtypes present different number of driver alterations, with n-CLL displaying a mean of 5.5 driver alterations (range 0-10), i-CLL of 2.1 (range 0-7) and m-CLL of 1.4 (range 0-5). Collectively, these analyses indicate that the cellular origin of CLL seems to predispose to the selection of distinct genetic landscapes characterized by incremental numbers of driver genetic alterations, a finding that may be related to the differences in patient clinical outcome of the three CLL subtypes [187,254,257–259].

Fig. 31 | (Epi)genomic crosstalk related to cellular origin in CLL patients (see figure on next page). a, Oncoprint showing the entire CLL driver genetic landscape as considered by the two landmark genomic studies in CLL [245,246]. CLL patients are divided according to the three CLL subtypes and ordered by increasing epiCMIT levels. On top, multiple clinico-biological variables are depicted, including the levels of monoclonal cells in blood (i.e, MBL versus CLL), the IGHV status, the epiCMIT groups according to maxstat statistic, the epiCMIT levels, the need for treatment and the patient status at last follow-up. The total number of driver alterations (considering all types) are shown as barplots at top. All possible types of genetic alteration for each gene are depicted with different colors and shapes. **b,** Correlation between the epiCMIT and the total number of mutations detected by WGS in CLL patients separated by epigenetic subtypes. **c,** Correlation between the epiCMIT mitotic clock and the accumulation of driver alterations in CLL separated by each epigenetic subtype.

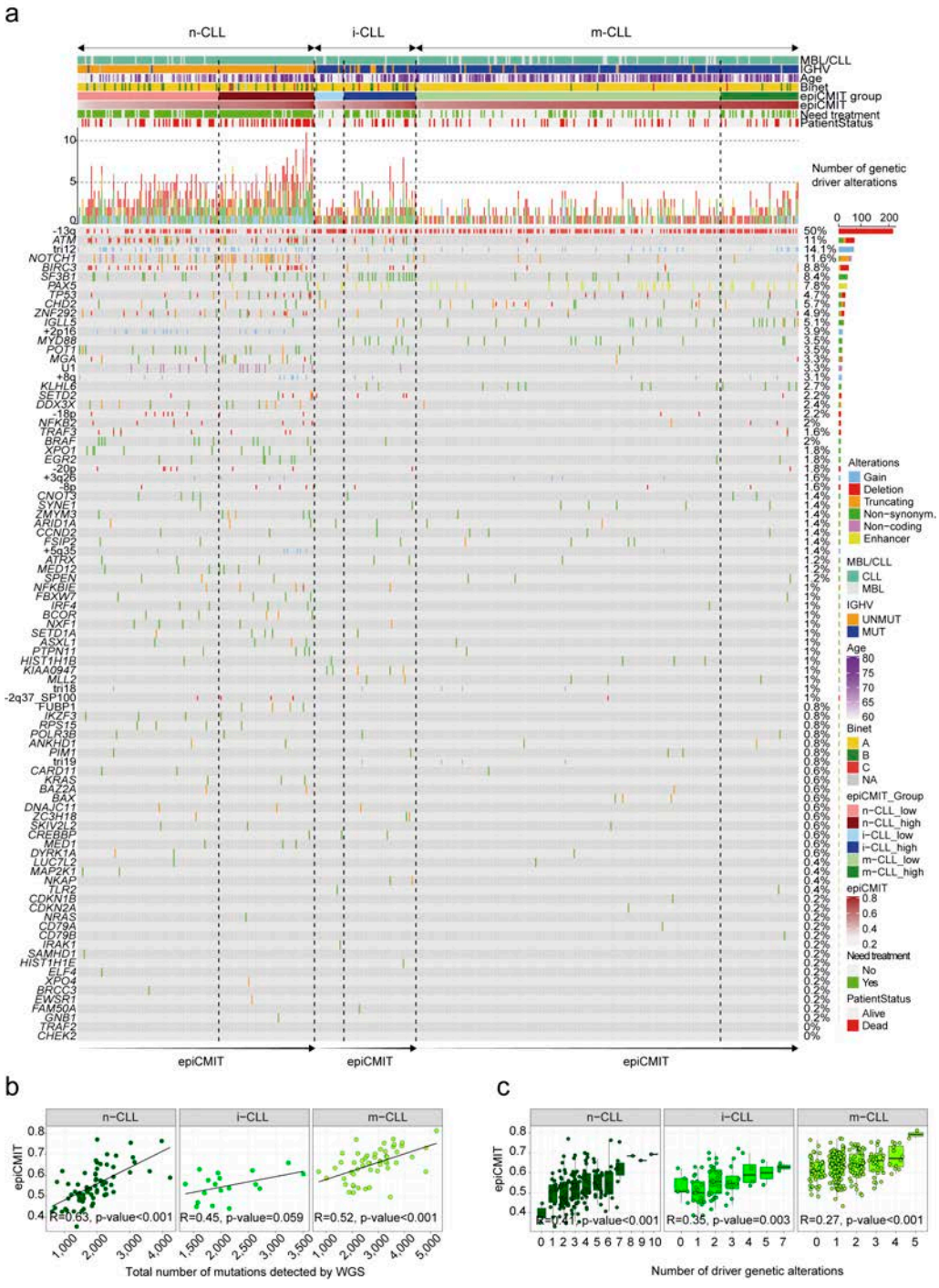


Fig. 31 | See caption on previous page.

4.2.2.2. (Epi)genomic crosstalk related to CLL proliferative history.

Next, I sought to determine the relationship between genetic aberrations and the CLL proliferative history measured by the epiCMIT score. I detected a correlation between the number of mutations detected by WGS and the epiCMIT score (R=0.63, p-value<0.001 for n-CLL, R=0.45; p-value=0.06 for i-CLL, and R=0.52; p-value<0.001 for m-CLL) (**Fig. 31b**). This finding is related to the fact that the majority of mutations in CLL can be attributed to SBS1, SBS5 and SBS9 mutational signatures in IGHV mutated CLL, which altogether represent mitotic-like mutations and thus are expected to be correlated with the epiCMIT score. These results explain the correlation previously seen between the number of somatic alterations and the number of DNA methylation changes [245], as the latter is strongly correlated with the epiCMIT (**Study 3, section 3.2.3, Fig. 26h**). In addition, the epiCMIT score correlated with the total number of driver alterations within each CLL subtype, suggesting that the epiCMIT score is significantly increased by mutations with positive selection compared with random or passenger mutations (**Fig. 31c**). Beyond this general association, I next assessed which individual driver alterations could predispose CLL cells to a higher proliferative capacity and thus a higher epiCMIT (**Fig. 32a**). To do so, 59 driver alterations showing at least 4 mutated patients in the whole cohort were retained for the analyses, which were performed in CLL as a whole and within each CLL epigenetic subtype separately (**Fig. 32b**). I found significant and positive associations between the epiCMIT score and the presence of 23 individual genetic driver alterations affecting the main signaling pathways deregulated in CLL (**Fig. 32b**) [245,246]. The majority of these genetic alterations have been previously linked to an adverse clinical behavior of CLL patients, such as *NOTCH1*, *TP53*, *SF3B1*, *ATM*, *BIRC3* or *EGR2*. Interestingly, epiCMIT showed an association with the new and recently identified non-coding genetic driver in CLL, the U1 spliceosomal RNA, a finding that may explain its suggested poor prognostic impact [318]. Remarkably, the presence of some genetic alterations was associated with high epiCMIT indistinctly in all patients, such as *TP53*, while others were particularly associated with high epiCMIT within CLL subtypes, such as *NOTCH1* in n-CLL and i-CLL, and *SF3B1* or *ATM* in i-CLL. Collectively, these results suggest that the well-established clinical impact of particular genetic alterations in CLL may be explained by their capacity to confer a higher proliferative advantage to CLL cells, which seems to be different depending on CLL cellular origin.

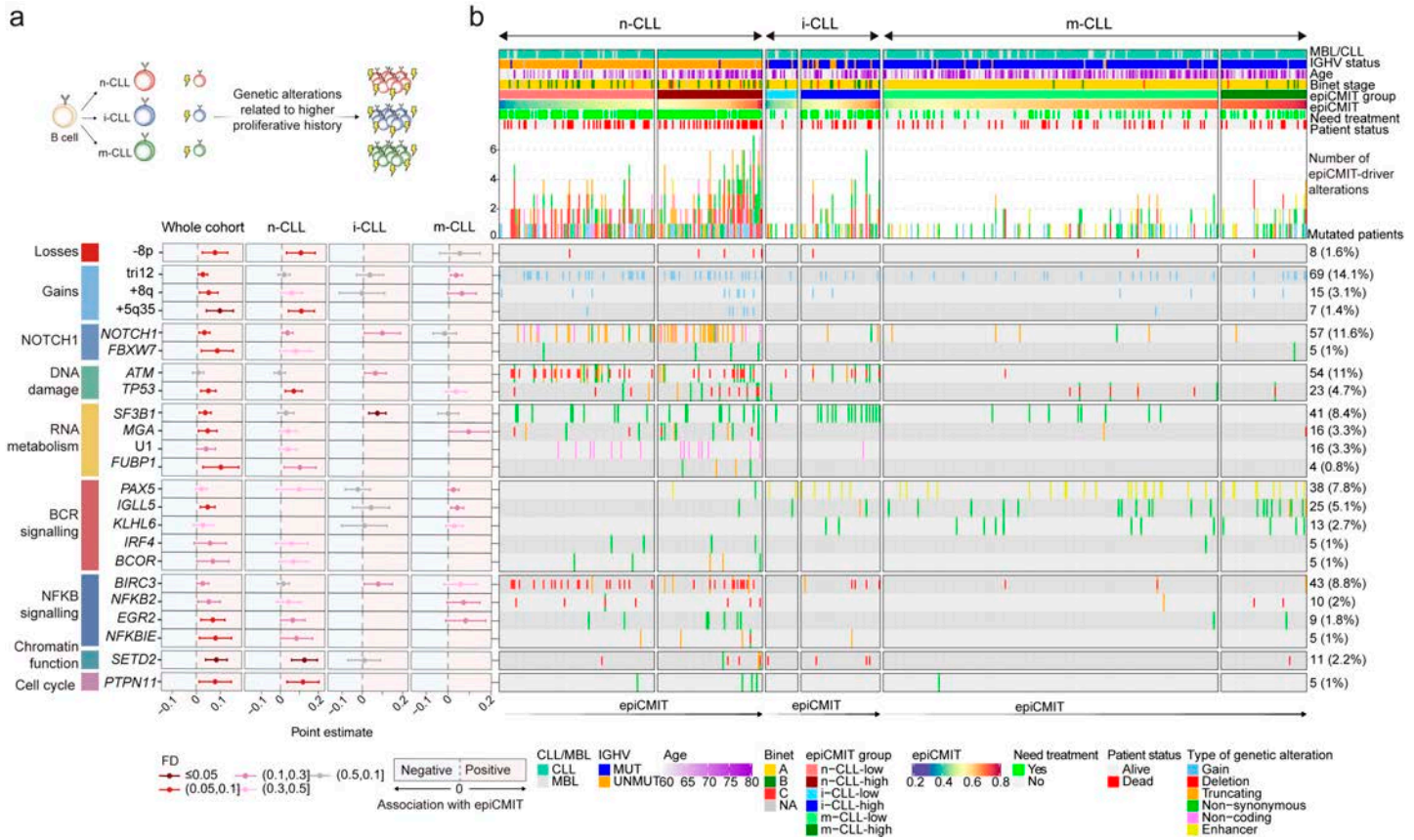


Fig. 32 | Genetic driver alterations associated with high epiCMIT levels in CLL. a, b Genetic driver alterations associated with high epiCMIT levels are grouped by signaling pathways. The analyses are done separately for the whole cohort and for each CLL subtype. The oncoprint shows CLL subtypes ordered by increasing epiCMIT levels. The same information as in **Fig. 31** is shown. Point estimates and 95% confidence intervals were derived from linear modelling and t-tests between mutated and wild type patietns (**Methods Study 4, section2.18**).

4.2.2.3. Prognostic impact of (epi)genetic features in CLL.

During the last years, it has been shown that the clinical behavior of CLL patients can be better predicted by the assessment of genomic complexity instead of single genetic alterations [245,251,356]. At the time of writing this thesis, the only genetic alteration modifying treatment decisions in CLL affects the *TP53* tumor suppressor gene, which discriminates patients that should be treated with novel agents including ibrutinib, venetoclax or idelalisib instead of standard chemotherapy regimens. Nonetheless, its prognostic importance needs to be understood in the context of the full CLL genomic landscape. *TP53* alterations are frequently associated with higher genetic complexity in the form of larger number of driver genes and higher chromosomal complexity, both of which representing very important prognostic variables regardless of the underlying alterations. In fact, cases with *TP53* alteration can be stratified into different prognostic groups based on the number of chromosomal aberrations, and those cases with TP53 alteration and 0-2 additional changes do not show a particularly poor prognosis [356]. Even more, a recent study has revealed that some CLL cases which spontaneously regress without treatment actually show *TP53* mutations [357]. Despite all these studies showing a better CLL prognostication using genomic complexity, the evaluation of all CLL driver alterations is still far from the clinical routine. Thus, a reduction of the whole driver gene list while maintaining the prognostic power is relevant from a clinically-oriented perspective. In this context, I reasoned that the association of genetic driver alterations with an increased epiCMIT level may represent an appropriate rationale to optimize the number of driver alterations assessed while keeping the prognostic power. With this rationale in mind, I first analyzed 89 driver alterations considered by the two seminal genetic papers in CLL [245,246] and as previously found, the accumulation of genetic driver alterations conferred a progressively worse effect on patient clinical outcome (**Fig. 33a**). Next, I retained 59 driver alterations present in at least 4 patients in the whole cohort, 23 of which were those associated with higher levels of epiCMIT (**Fig. 32b**), which I called CLL-epiCMIT-drivers, while the remaining 36 were not associated at all with the epiCMIT and were termed CLL-non-epiCMIT-drivers. Interestingly, considering only the 23 CLL-epiCMIT-drivers, all patients could be stratified into 4 groups with markedly different clinical outcomes (**Fig. 33b**). This patient stratification was less clear in the case of the 36 CLL-non-epiCMIT-drivers despite analyzing 13 drivers more, and patients with one driver had the same clinical outcome than patients without any driver (**Fig. 33c**). In fact, a multivariate Cox regression model with these two sets of driver alterations showed

that CLL-epiCMIT-drivers confer a remarkably higher risk to receive treatment compared to CLL-non-epiCMIT-drivers (0.64 for each driver, and thus almost 2 times higher risk in the group of 3 or more drivers) (**Fig. 33d, left panel**). Since I could not validate this results in an external series, I permuted the data to know whether this HR was frequent considering other combinations of 23 driver alterations. To do so, I took 10,000 times 23 driver alterations randomly selected and obtained the distribution of HRs of this set of drivers in a multivariate Cox model together with the remaining 36 possible drivers not selected (**Fig. 33d, right panel**). This analysis showed that out of the total 59 driver alterations analyzed, it is unlikely to obtain a higher HR in the multivariate Cox model selecting other sets of 23 driver alterations. These results suggest that the CLL-epiCMIT-drivers are probably among the alterations conferring the worst clinical outcome. Next, I reasoned that CLL samples with a higher proliferative history may be those with the worst clinical outcome. I classified each CLL patient according to its proliferative history burden considering both genetic (CLL-epiCMIT-drivers) and epigenetic (epiCMIT score itself) variables in the context of the 3 CLL epigenetic subtypes. Thus, each CLL epigenetic subtype was further classified as having both variables low (low epiCMIT and low CLL-epiCMIT-drivers), one of them high, or both high, i.e. 9 CLL groups in total. This analysis indeed showed that the higher the proliferative history the worse patient clinical outcome regardless of CLL epigenetic subtypes (**Fig. 33e, f**). Interestingly, despite CLL-epiCMIT-drivers and epiCMIT being associated with CLL proliferative history, a multivariate Cox regression model with these two variables together with age and CLL epigenetic subtypes revealed them as independent from each other. This is an interesting result which suggests that even though both variables are related to the CLL proliferative history, they still seem to give complementary prognostic information (**Fig. 33g**).

Altogether, it seems that when considering the number of driver alterations to estimate prognosis in CLL, not all driver genetic alterations may have the same weight. Particularly, those associated with higher epiCMIT levels may represent driver alterations conferring a higher proliferative capacity to CLL cells and thus a worse patient clinical outcome. The integration of CLL-epiCMIT-drivers and the epiCMIT score in the context of CLL epigenetic subtypes represents a reduced CLL (epi)genetic patient stratification with high prognostic power.

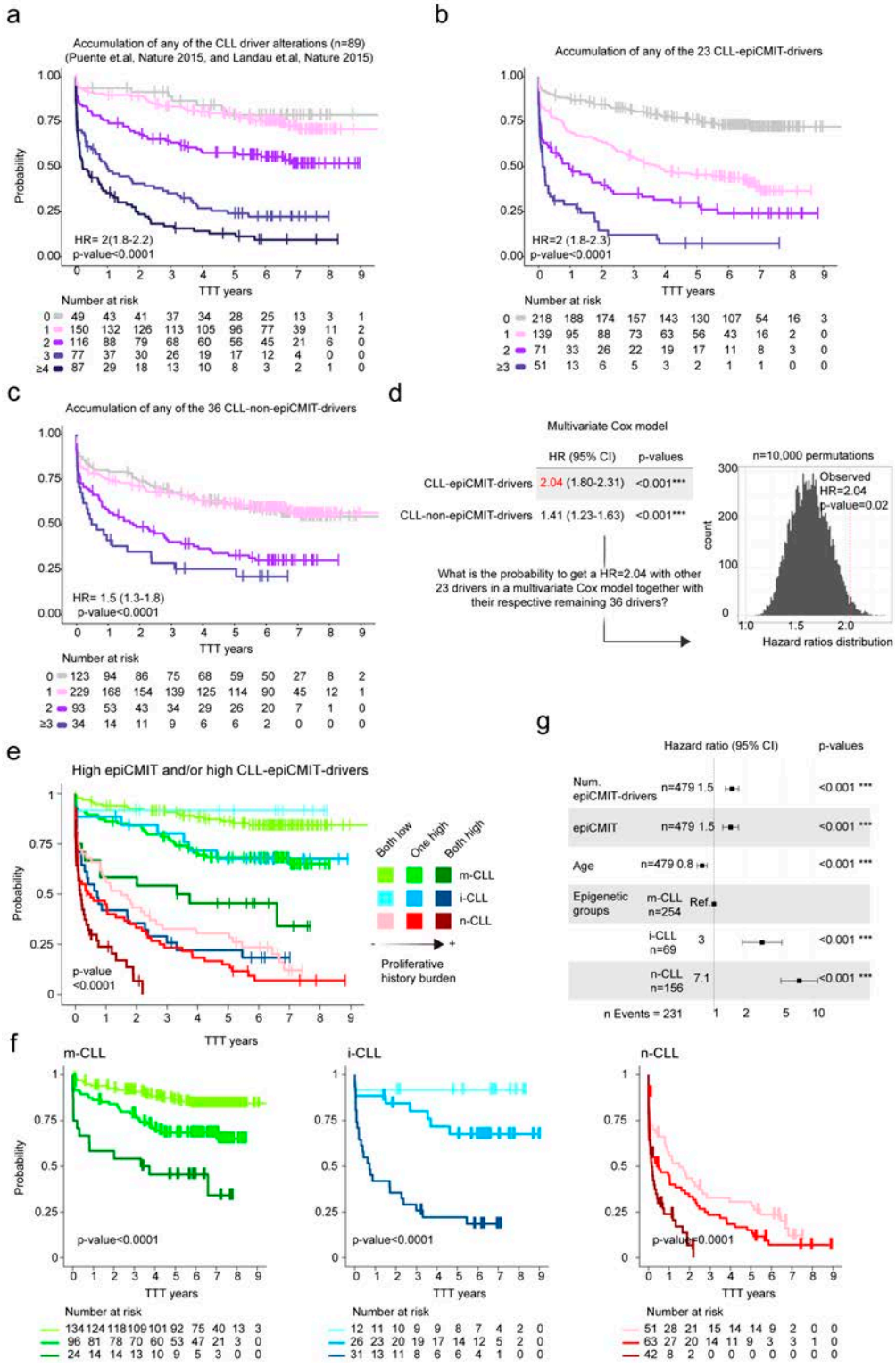


Fig. 33 | See caption on next page.

Fig. 33 | The prognostic impact of genetic driver alterations associated with high epiCMIT in CLL patients (see figure on previous page). a, Kaplan Meyer curves for TTT in CLL patients according to the accumulation of any of the 89 genetic driver alterations considered by the two landmark genomic studies in CLL [245,246].**b,** Kaplan Meyer curves for TTT in CLL patients according to the accumulation of any of the 23 driver alterations associated with high epiCMIT (CLL-epiCMIT-drivers). **c,** Kaplan Meyer curves for TTT in CLL patients according to the accumulation of any of the 36 driver alterations not associated with epiCMIT (CLL-non-epiCMIT-drivers). Genetic drivers with at least 4 patients affected were considered for panels **b** and **c. d,** Multivariate Cox regression model for TTT with CLL-epiCMIT-drivers and CLL-non-epiCMIT-drivers. On the right, a permutation analysis comparing the hazard ratios of 23 random drivers and the respective remaining 36. **e,** Kaplan Meyer curves for TTT in CLL patients divided in CLL subtypes and according to the proliferative burden, defined as the accumulation of CLL-epiCMIT-drivers and high epiCMIT. **f,** CLL epigenetic subtypes from **e** are shown separately. **g,** Multivariate Cox regression model for TTT with CLL-epiCMIT-drivers and epiCMIT with other well-established clinical variables in CLL

4.2.3. (Epi)genomic crosstalk in MCL.

4.2.3.1. (Epi)genomic crosstalk related to MCL cellular origin.

In the case of MCL, I considered the two MCL subtypes defined by the last version of the WHO, namely cMCL and nnMCL, which greatly overlap with C1 and C2 MCL epigenetic subtypes, respectively (**Fig. 30d**) [229]. Using WGS and WES in the whole cohort, 43 high-confidence driver genetic alterations were detected, including 30 genes and 13 chromosomal regions without a defined target gene. Remarkably, 99% of MCL cases presented further genetic events in addition to the primary genetic change, the t(11;14) translocation (median 6, range:0-14) (**Fig. 34a**). The most frequently altered genes have been previously described in MCL and were *ATM* (48%), *CCND1* (44%) with exon1/intron1 somatic mutations (26%) and/or 3' UTR activating alterations (21%), *TP53* (26%), *KMT2D* (23%), *RB1* (23%), *BIRC3* (22%), *CDKN2A* (21%), *SP140* (13%), *NSD2* (12%), *BMI1* (11%), *MIR17HG* (10%), and *UBR5* (6%). Furthermore, seven novel MCL driver genes altered by missense or truncating mutations and deletions were identified, including *CDKN1B* (12%), *SAMHD1* (10%), *BCOR* (9%), *SYNE1* (6%), *HNRNPH1* (6%), *SMARCB1* (4%) and *DAZAP1* (4%) (**Fig. 34a**). Some alterations were exclusively seen in cMCL, including *ATM* alterations, deletions of 1p, 10p and 19p, gain of 7p and breakage-fusion bridge (BFB) cycles [358], a novel and frequent genomic aberration in cMCL, whereas other alterations were slightly enriched (*TP53* and *TERT*) or manly found (SHM in *CCND1*) in nnMCL (**Fig. 34a**). Altogether, cMCL cases had a significant higher number of genetic drivers (median 7 vs. 2, p-value<0.001) and complex alterations (52% vs. 18%, p-value=0.02) than nnMCL (**Fig. 34a**), a finding that may be related to the distinct clinical outcomes between cMCL and nnMCL patients.

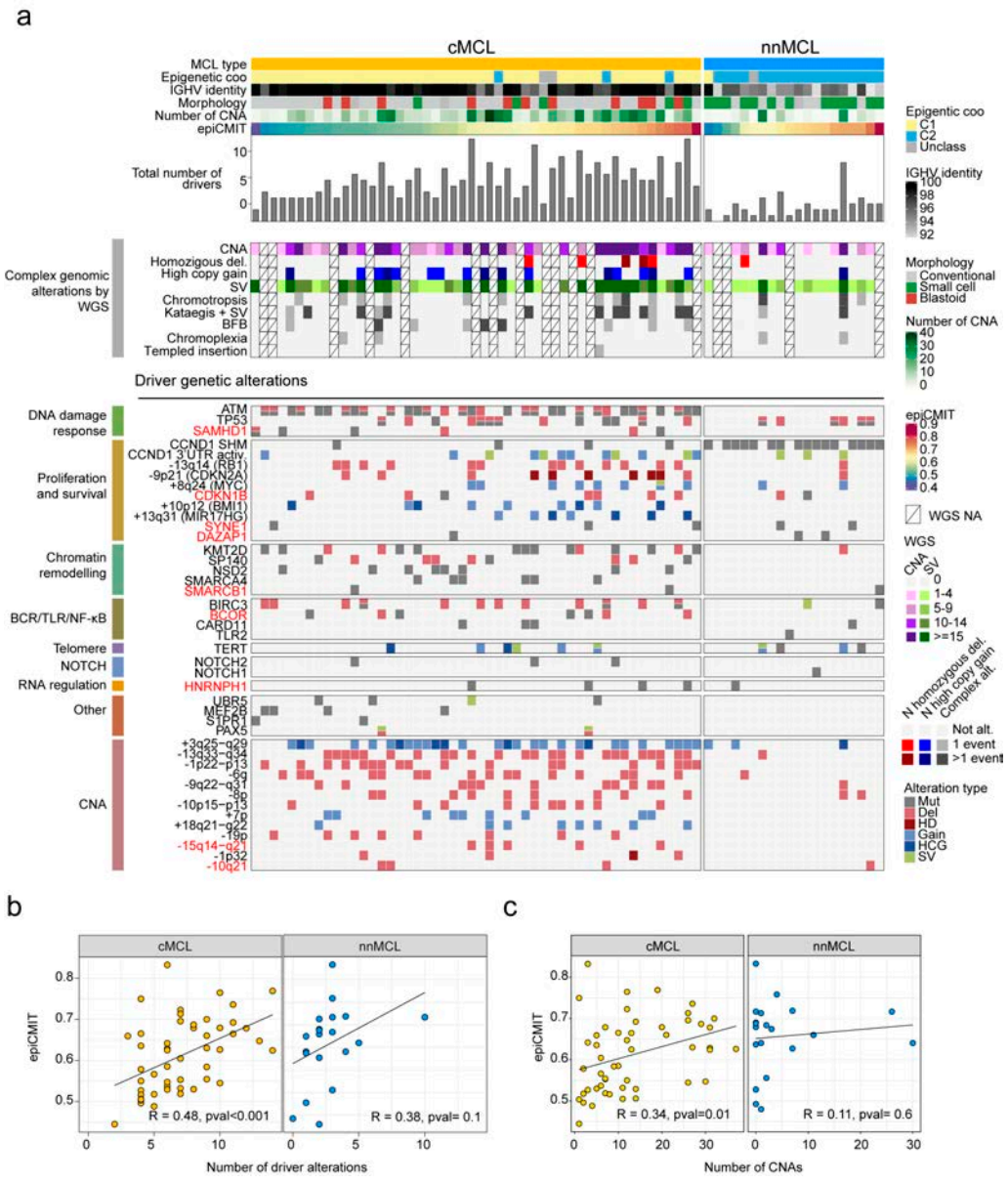


Fig. 34 | (Epi)genomic crosstalk related to cellular origin in MCL patients. a, Oncoprint representing the entire driver genetic landscape and complex genomic alterations detected by WGS in MCL subtypes ordered by increasing levels of epiCMIT within each MCL subtype. Genetic alterations are grouped by pathways. Other clinico-biological features are represented on top, including MCL epigenetic subtypes, IHGV identity, morphology, CNA (based on array data) and epiCMIT values. **b**, Correlation between the epiCMIT and the number of driver alterations in each MCL subtype. **c,** Correlation between the epiCMIT and the number of CNA in each MCL subtype.

4.2.3.2. (Epi)genomic crosstalk related to MCL proliferative history.

I next studied the relationship between genetic alterations and MCL proliferative histories. As previously seen in CLL, I detected a high correlation between the total number of mutations and the epiCMIT, a finding consistent with the fact that the majority of mutations in MCL are also related to mitotic-like mutational processes (SBS1, SBS5 and SBS9). Besides, the epiCMIT correlated with genomic complexity in cMCL, both in the form of number of driver alterations and number of CNA (**Fig. 34b, c**). These correlations again suggest that the presence of genetic alterations with positive selection and not random mutations increase the epiCMIT score. In addition to this general association, I next analyzed the impact of individual genetic alterations on the epiCMIT score, and found some of them particularly linked to differential epiCMIT levels (**Fig. 35**). Specifically, 8 driver genetic alterations were associated with high epiCMIT levels and affected pathways involved in MCL pathogenesis including DNA damage response (*TP53*, del(13q33), +18q21) proliferation and survival (*CCND1* 3'UTR activ., -9p21-*CDKN2A*, +8q24-*MYC*, +13q31-MIR17HG) and RNA regulation (*HNRNPH1*). Notably, *TP53* was the only driver alteration significantly associated with a higher proliferative history in nnMCL. Interestingly, the newly identified genetic driver *HNRNPH1* and the well-known driver *MEF2B* were associated with a higher and lower proliferative histories, respectively, a finding which may explain the opposite clinical outcomes recently seen of these two alterations in MCL patients [359].

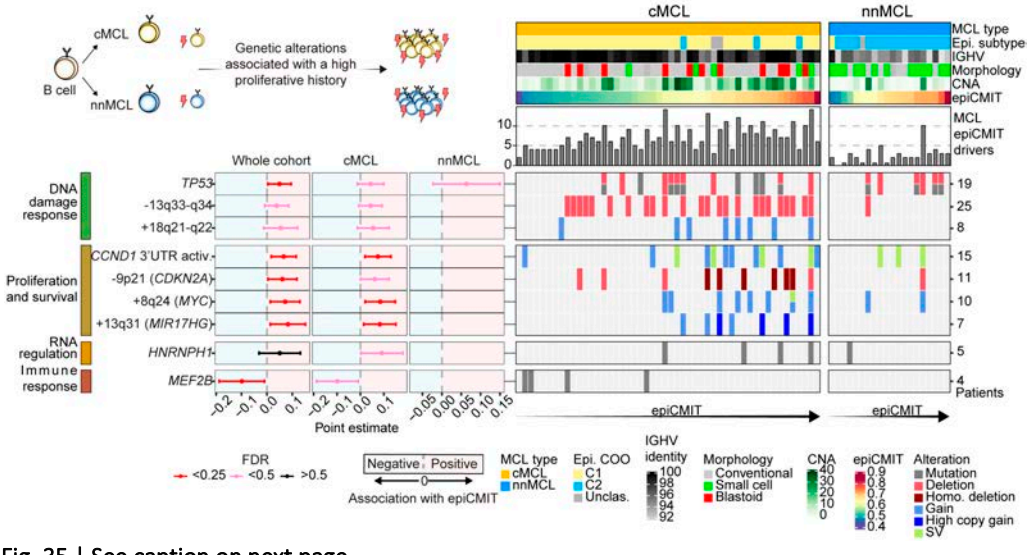


Fig. 35 | See caption on next page.

Fig. 35 | Genetic driver alterations associated with high epiCMIT levels in MCL patients (see figure on previous page). Genetic driver alterations associated with high epiCMIT levels are grouped by different signaling pathways (MCL-epiCMIT-drivers). The same clinico-biological information as in Fig. 34 is represented at top. For MCL-epiCMIT-drivers, alterations with at least 2 mutated patients were included for the analyses. Point estimates and 95% confidence intervals were derived from linear modelling and t-tests (**Methods Study 4, section 2.18**).

4.2.3.3. Prognostic impact of (epi)genetic features in MCL.

To analyze the clinical relevance of the identified genetic and epigenetic features, I considered both MCL as a whole and two groups of patients, those that needed urgent treatment at diagnosis (100% cMCL cases), and patients that were initially managed with a watch-and-wait approach (90% nnMCL cases). I found some already known individual alterations impairing the survival of patients, such as *TP53* alterations, 9p21/*CDKN2A*, −13q14/*RB1*, and −9q22−q31 (**Fig. 36a, b**) [360–364]. Interestingly, six novel drivers were also associated with significant shorter OS, including −15q14−q21, +8q24/*MYC*, *SP140*, +13q31/MIR17HG, +18q21-q22, and −13q33-q34 (**Fig. 36a, b**). The results in patients treated at diagnosis were similar to those observed in the whole cohort, whereas TERT was the only driver associated with TTT in the group of patients with deferred treatment (**Fig. 36c**). In this small subgroup of cases with few events, *TP53* alterations had a marginal effect.

Next, I assessed the potential clinical relevance of genomic complexity and epigenetic changes captured by the epiCMIT score. The increasing number of CNA, presence of BFB, and epiCMIT, but not the number of driver mutations and SV, were associated with shorter OS both in the whole cohort and the subgroup of patients treated at diagnosis (**Fig. 36d**). Despite the small number of events in patients with deferred treatment, I found that the number of CNA and chromothripsis conferred a shorter TTT (**Fig. 36d**). Remarkably, only *TP53*, *MYC* and the newly identified driver gene *HNRNPH1* had prognostic value independently of CNA (**Fig. 36a, b**, highlighted in red and with an asterisk), indicating that the clinical value of all other individual drivers may be due to their association with genomic complexity measured by CNA. Worthy of note, the number of CNA, presence of BFB, and epiCMIT retained independent prognostic value for OS for the whole series (**Fig, 36e, left panel**) whereas for the patients treated at diagnosis, only epiCMIT and BFB retained independent prognostic impact for OS, and only epiCMIT for progression-free survival (PFS) (**Fig. 36 middle panels**). Finally, only CNA

had prognostic value in the smaller group of patients with deferred treatment. The integration of genomic and epigenomic complexity measured by CNA and BFB and the epiCMIT score stratified patients with a markedly distinct clinical outcome (**Fig. 36f**), indicating that the complex evolution of MCL may be better recognized by the integration of genomic and epigenomic parameters.

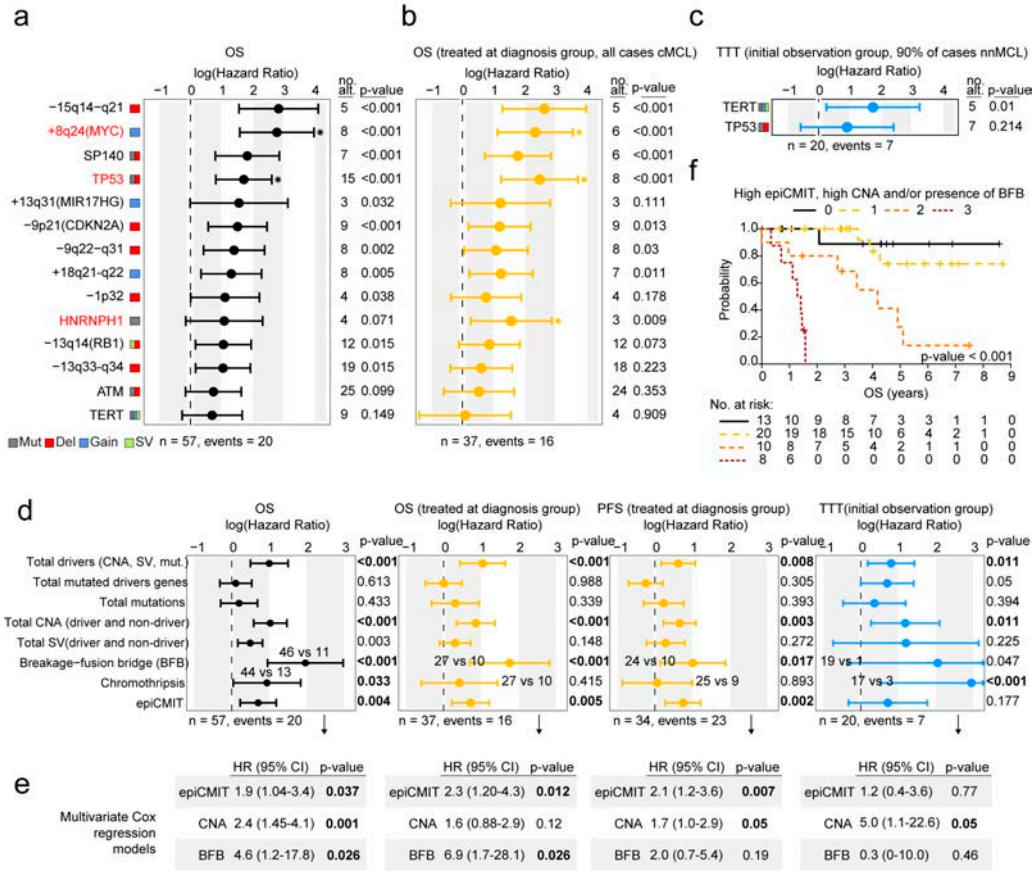


Fig. 36 | The prognostic impact of genetic and epigenetic aberrations in MCL patients. a, Individual genetic driver alterations with impact on overall survival of the whole cohort of MCL patients. Asterisks mark driver alterations with independent prognostic impact from CNA. **b**, Same in **a**, but for MCL patients with urgent treatment at diagnosis (all cases were cMCL). **c,** Same in **a** but for patients with deferred treatment. **d,** Univariate Cox models for overall survival for variables capturing genomic complexity and the epiCMIT in the whole cohort and patients treated at diagnosis (first two panels). Third panel, univariate Cox model for progression-free survival (PFS) with the same variables in the treated at diagnosis group, and in the fourth panel, TTT with the same variables in patients with deferred treatment. **e,** Multivariate Cox models with the most potent prognostic variables from **e**. **f,** Kaplan Meyer curves for OS considering the CNA, BFB and the epiCMIT.

Despite the accurate MCL prognostication using genetic and epigenetic variables together (**Fig. 36f**), the analysis of the entire MCL (epi)genomic landscape is still far from being part of the clinical routine. Thus, as in the case of CLL, the reduction of these (epi)genetic variables to few biomarkers while keeping the high prognostic power is relevant for the clinical practice. I reasoned again that the genetic alterations linked to an increased epiCMIT score could represent good candidates. As an initial observation further supporting this notion, all 3 genetic driver alterations with independent prognostic impact from the CNA in our series, a very potent prognostic variable, were strongly associated with higher epiCMIT levels (**Fig. 36a, b**, highlighted in red and with an asterisk), representing 3 out of the 8 drivers associated with higher epiCMIT levels or MCL-epiCMIT-drivers (**Fig. 35**). Thus, I analyzed patient clinical outcome in relation to the accumulation of only the 8 MCL-epiCMIT-drivers. Remarkably, I found that their incremental acquisition conferred a progressively worse effect on patient outcome (**Fig. 37a**). The prognostic power of the MCL-epiCMIT drivers was indeed comparable with that of previously recognized genetic and epigenetic variables, including the total number of CNA (**Fig. 37b**), the epiCMIT score (**Fig. 37c**), or even with the previous (epi)genetic-based MCL stratification including CNA, BFB and epiCMIT (**Fig. 36f**). In fact, it seems that the acquisition of the 8 MCL-epiCMIT-drivers is associated with both MCL genomic complexity and proliferative history. The accumulation of MCL-epiCMIT-drivers strongly correlated with the total number of CNA both in cMCL and nnMCL subtypes (**Fig. 37d**), with the epiCMIT score (mainly in cMCL) (**Fig. 37e**), and with Ki67 expression in cMCL (**Fig. 37f**). Furthermore, bivariate Cox regression models with MCL-epiCMIT-drivers, epiCMIT and the total number of CNA revealed MCL-epiCMIT-drivers as an overall superior prognostic variable, with epiCMIT losing its prognostic power and the number of CNA being barely significant (**Fig. 37f, left panel**). After 10,000 random permutations, this finding is very unlikely with other combinations of 8 drivers in the MCL series analyzed (**Fig. 37f, right panel**).

Collectively, it seems that MCL clinical evolution can be better determined by integrating genetic and epigenetic variables, including the number of CNA, BFB and the epiCMIT score. More interestingly, the evaluation of only 8 genetic drivers alterations associated with higher epiCMIT levels seems to capture both MCL genomic complexity and proliferative history. These 8 MCL-epiCMIT-drivers shows a remarkable prognostic value and thus they may represent the basis for a reduced set of alterations to be used in the clinical setting.

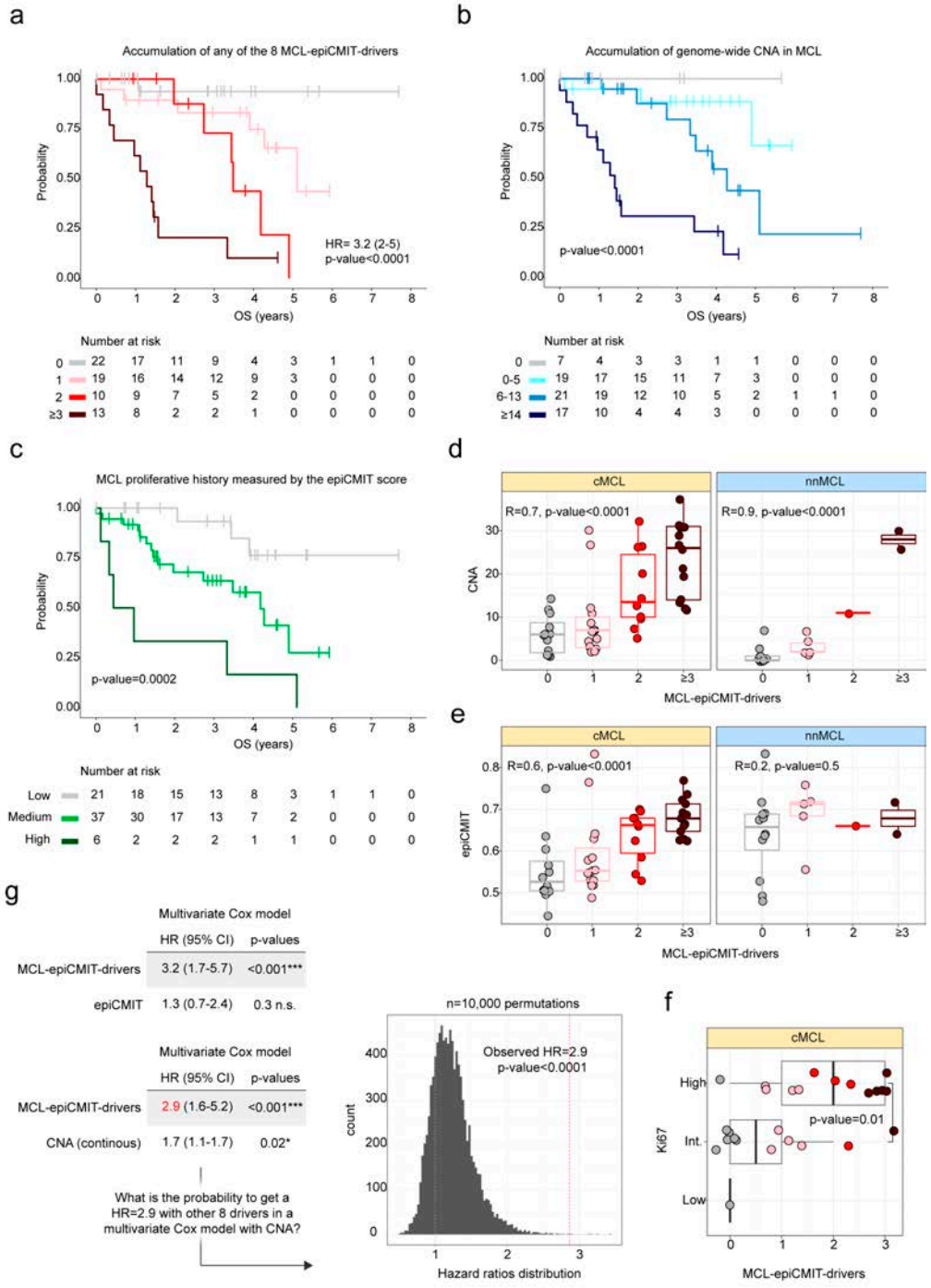


Fig.37 | See caption on next page.

Fig.37 | The prognostic impact of genetic driver alterations associated with high epiCMIT in MCL patients (see figure on previous page). a, Kaplan Meyer curves for overall survival according to the accumulation of any of the 8 genetic driver alterations associated with high epiCMIT in MCL patients (MCL-epiCMIT-drivers). **b,** Kaplan Meyer curves for overall survival according to the accumulation of CNA. Cutoffs were defined as in Clot, G 2018 [360]. **c,** Kaplan Meyer curves for overall survival according to the MCL cells proliferative history. Cutoffs were based on epiCMIT distribution in MCL samples. **d,** The accumulation of MCL-epiCMIT-drivers correlates with the acquisition of CNA. **e,** The accumulation of MCL-epiCMIT-drivers correlated with the epiCMIT in cMCL and with Ki67expression **(f)**. **g,** Bivariate Cox regression models with the MCL-epiCMIT-drivers together with the epiCMIT and the CNA. A permutation analysis of bivariate Cox regression models with 8 random driver alterations with CNA. All clinical analyses were done in the whole cohort of patients. For MCL-epiCMIT-drivers, alterations with at least 2 mutated patients were included for the analyses.

4.3. Conclusions.

- The main sources of DNA methylation variability in CLL and MCL are their cellular origin and their proliferative histories detected by the epiCMIT score.

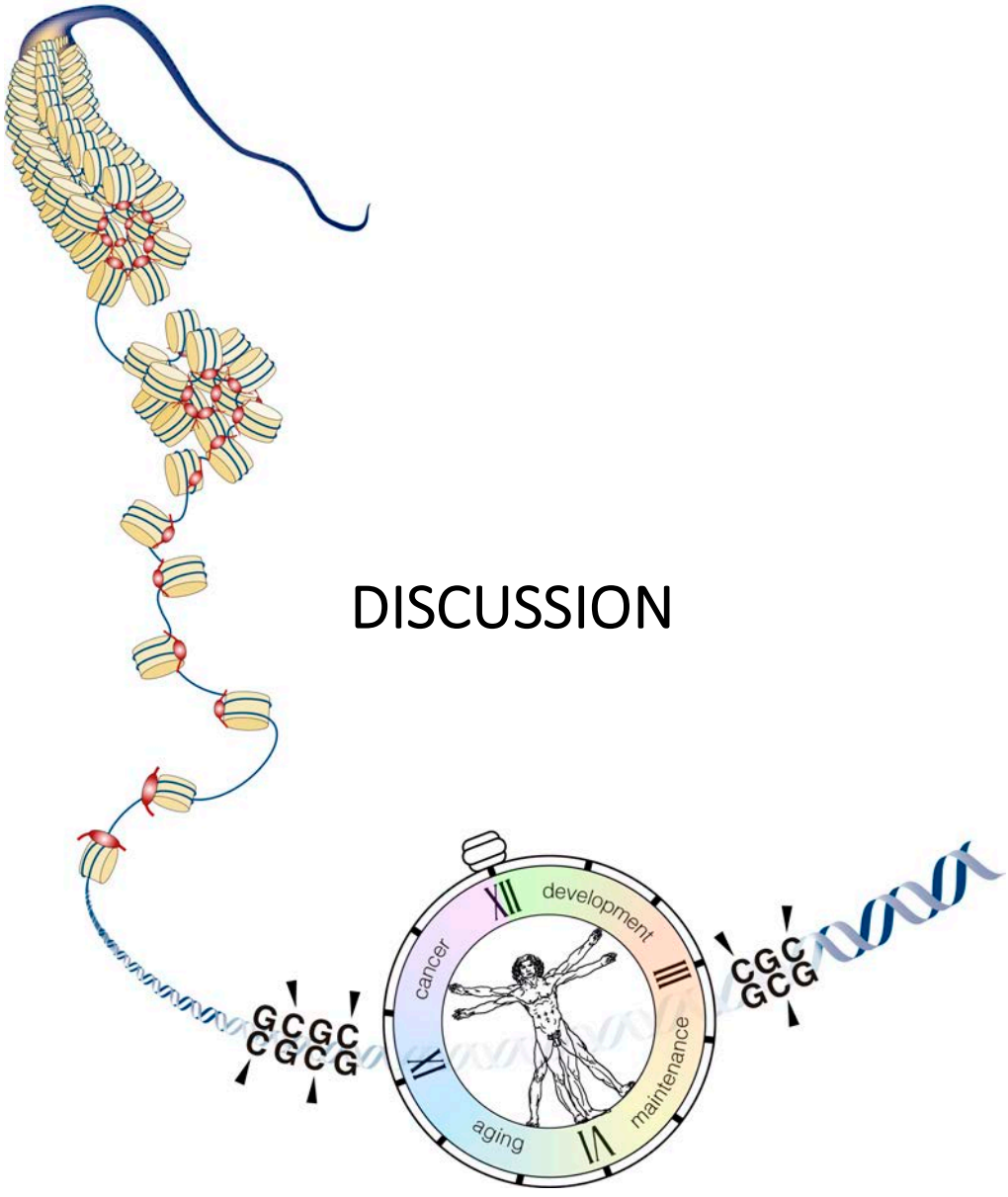
- CLL and MCL samples with a germinal center-independent origin display higher genomic complexity and number of driver alterations, which are correlated with the epiCMIT score.

- The epiCMIT score seems to be significantly increased by the accumulation of genetic alterations with positive selection in both CLL and MCL patients.

- Some driver genetic alterations are particularly associated with higher epiCMIT levels and target pathways important for CLL and MCL pathogenesis. These alterations may confer a higher proliferative capacity to tumor cells and are generally related to a worse patient clinical outcome.

- The complex clinical evolution of CLL and MCL patients may be better determined by the integration of genomic and epigenomic aberrations, including genomic complexity and the epiCMIT score.

- The accumulation of drivers significantly associated with higher epiCMIT levels confer a progressively worse clinical outcome. The evaluation of these drivers may represent the basis for a reduced set of driver alterations appropriate for the clinical routine.


GCGC
CGCG
XII
development
III
maintenance
VI
aging
IX
cancer
CGC
GCG

DISCUSSION

During the course of this doctoral thesis, I have followed a systematic approach to dissect the sources of DNA methylation variability of normal cell development and 14 tumor subtypes spanning the entire human B-cell lineage. The complete DNA methylome of MCL was first analyzed in the context of normal B-cell differentiation (**Study 1**), a study that laid the foundations for the subsequent comprehensive characterization of DNA methylation variability among the main B-cell tumors and their subtypes (**Study 2**), which allowed me to reveal new insights into the biology and the clinical behavior of this heterogeneous group of neoplasms (**Study 3 and Study 4**).

1. Differential DNA methylation patterns in normal and neoplastic B cells.

Overall, I found that the DNA methylation levels of 88% of the studied CpGs can be modulated in normal and/or neoplastic B cells, suggesting that the human DNA methylome is even more dynamic than previously conceived [49,50,52]. The extensive DNA methylation variability among different B-cell neoplasms is in part related to imprints of normal cell development. This imprints has been recently used to classify not only B-cell neoplasms [187,254,258,295], but also solid tumors [365–367]. Beyond this epigenetic memory tracing the tumor cellular origin, B-cell neoplasias acquire additional methylation changes naturally occurring during normal B-cell development (**Study 2)** and reference [50]. Based on the findings derived from this thesis, these methylation changes are mostly related to the extensive proliferative history of cancer cells (**Study 3**). As B cells also proliferate during their normal maturation program, it is expected that physiological and pathological proliferation can lead similar DNA methylation imprints. For instance, highly proliferative ALL cells deriving from precursor B cells acquire a number of B-cell related methylation changes similar to that observed in terminally-differentiated bmPC, which accumulate in themselves the entire proliferative history of normal B-cell development (**Fig 38**). In addition to this epigenetic link to normal cell maturation, each B-cell neoplasm also shows disease-specific hyper- and hypomethylation. Of particular interest are the disease-specific *de novo* hypomethylation signatures in active regulatory regions, which are associated with TF binding sites and related to disease-specific transcriptional profiles. This phenomenon is particularly marked in ALL, MCL and CLL, whose *de novo* hypomethylation is enriched in regulatory elements and TF binding sites previously linked to their pathogenesis [331–333]. Unexpectedly, although DLBCL and MM pathogenesis has been related to TFs, *de novo* hypomethylation in these diseases is depleted of regulatory elements containing TF binding sites. In these two malignancies,

I detected few binding sites of TFs at regulatory regions with potential involvement in the diseases. However, I did not detect classical TFs such as BCL6 in DLBCL or IRF4 in MM, possibly because they are key players during B-cell differentiation and their binding sites may be already hypomethylated in the normal B-cell counterparts of DLBCL and MM.

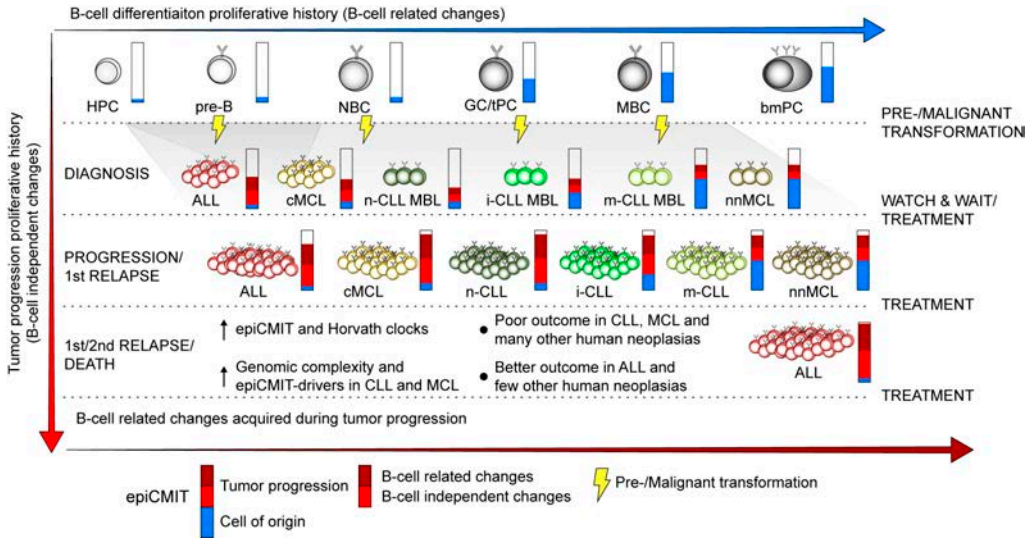


Fig.38 | The proliferative history shapes the DNA methylome of B-cell tumors and predicts clinical outcome. Mitotic cell division leaves transcriptionally-inert DNA methylation changes in silent chromatin in both physiological and malignant conditions. These changes are captured by the epiCMIT mitotic clock, which therefore contain cell divisions associated with normal B-cell differentiation and with malignant transformation and progression. The epiCMIT is related to genomic complexity and particular genetic driver alterations in CLL and MCL and predict patient clinical outcome not only in B-cell tumors, but also in other human neoplasia from the TCGA dataset. The higher proliferative history of B-cell tumors correlates with an increased epigenetic aging measured by Horvath clock, suggesting a crosstalk between mitotic activity and aging.

All in all, I can conclude that the DNA methylome of B-cell tumors show a) imprints of the maturation stage from which they are derived, b) changes shared with normal B-cell differentiation, and c) *de novo* disease-specific alterations, some of which target regulatory regions and relate to specific transcriptional profiles. Using an unbiased machine-learning approach with all types of differential methylation patterns, an accurate Pan B-cell tumor diagnostic tool that discriminates the main B-cell tumor

entities and subtypes with different clinical management was built (Study 2). This epigenetic classifier may represent the basis for a simple and accurate diagnostic tool to be used in the clinical routine, which can complement current diagnostic procedures and clarify cases with unclear differential diagnosis. More importantly, it can be of special relevance for B-cell tumor subtypes with a challenging diagnosis with current methods and with a marked distinct outcome, such as cMCL and nnMCL MCL subtypes [360] or n-CLL, i-CLL and m-CLL CLL subtypes [187,254,258,259,368].

2. Epigenetic mitotic history and related clinico-biological features.

In spite of the widely-reported importance of DNA methylation at regulatory regions, I identified that the majority of DNA methylation changes in individual patients are located in inactive chromatin (Study 2). These DNA methylation changes are manifested mainly in the form of concomitant hypomethylation of heterochromatin and hypermethylation of H3K27me3-containing regions, a phenomenon previously observed in colorectal cancer [346]. Compelling published evidences over the last years [73,150,313,316,343–345] and the data presented in this thesis support the notion that mitotic cell division leaves transcriptionally-inert epigenetic imprints onto the DNA located in repressive chromatin environments (Study 3). More recently, this knowledge has led to the concept of using these hyper- or hypomethylation changes as a mitotic clock [313,314,316], which in turn have been recently confirmed at the single cell level [369,370]. Nonetheless, I identified that using only hyper- or hypomethylation to build a mitotic clock may be insufficient to capture the entire mitotic history of cancer cells, as some neoplasms seem to preferentially gain or lose DNA methylation upon cell division. For instance, ALL seems to acquire broad hypermethylation upon cell division, whereas I consistently observed the opposite scenario in MM. Thus, using exclusively hyper- or hypomethylation [313,314,316] to determine the mitotic history of MM or ALL cells would incongruently lead to the conclusion that they have not proliferated beyond their cellular origin. The epiCMIT mitotic clock circumvents this limitation by selecting carefully both hyper- and hypomethylation at key CpGs, which altogether capture the entire mitotic history of normal and neoplastic B cells regardless of their tendencies to gain or lose methylation. In B-cell tumors, the epiCMIT includes cell divisions associated with both normal development as well as neoplastic transformation and progression (Fig. 38). Thus, the epiCMIT should not be compared among B-cell tumors arising from different normal counterparts but its relative magnitude must be studied in those arising

from a particular maturation stage. Within each of these subgroups, the relative epiCMIT has a superior prognostic value than previous mitotic clocks and a profound independent prognostic value from other well-established clinical variables in B-cell tumors. Increased epiCMIT is associated with worse clinical outcome in CLL and MCL, suggesting that superior proliferative history before treatment seems to determine future proliferative capacity of mature B-cell neoplasms. Strikingly, I consistently found the opposite scenario in ALL, a finding in line with recent reports showing CIMP (which is highly correlated with epiCMIT in ALL) as a good prognostic marker in this disease [211,371]. This result may suggest that the high proliferative ALL cells of children at diagnosis (and thus having a larger proliferative history) are more efficiently eliminated by high intensive chemotherapy regimens [372], which cannot be administrated in elderly patients such as in the case of CLL and MCL. Beyond all the B-cell tumors analyzed, epiCMIT also shows a potent prognostic value for a number of other human neoplasias from the TCGA dataset, suggesting that it may represent a pan-cancer mitotic clock with clinical utility. This notion is further supported by the exquisite correlation between the epiCMIT-hyper and epiCMIT-hypo with all the previous cell-type independent hyper- or hypomethylation-based mitotic clocks, respectively. Moreover, I found some neoplasias in which high proliferative history measured by all mitotic clocks before treatment is associated with better clinical outcome, such as acute myeloid leukemia and glioblastoma multiforme, indicating that ALL may not be an exception.

Collectively, it seems that tracing the past proliferative history of neoplasms is highly predictive of their future clinical behavior. Therefore, these data strongly argue to incorporate this epigenetic information in the clinical practice to better monitor cancer evolution. This will require more automated and cost-effective methodologies [259] beyond those offered by current array-based technologies. These may include sequencing-based approaches of plasma-cell free DNA methylomes, as they have been proven successfully in recent studies to classify several non-B-cell tumor types [373–375]. Moreover, integrating the detection of genetic and DNA methylation changes in a single technology, such as nanopore sequencing [376], or a novel approach invented by the Catalan start-up Aniling may represent ideal strategies to be introduced into the clinics (https://www.aniling.com/m/technology).

3. Genomic determinants of increased mitotic history and downstream clinical implications in CLL and MCL.

The comprehensive molecular characterization of CLL and MCL carried out in this thesis allowed me to investigate in depth the complex relationship between genetic and epigenetic aberrations and derive relevant clinical implications (**Study 4**).

Over the last years, multiple studies contributed to characterize the genetic and epigenetic landscapes and their clinical associations in both CLL and MCL [50,187,256,257,259,295,360,368,377,378,228,229,245–247,251,254,255]. Beyond individual alterations, these studies revealed that the tumor cellular origin and genomic complexity are probably the best variables to stratify patients with markedly distinct clinical outcomes. Within each CLL and MCL subtypes based on the IGHV hypermutation status, a more complex genetic makeup manifested in a higher number of driver genetic alterations, CNA, structural variants and/or the presence of BFB cycles and related to worse clinical outcomes [229,245,246,251,360]. The accumulation of genetic lesions with positive selection and not random mutations seems to increase the epiCMIT, as driverless patients show an overall lower epiCMIT. Nonetheless, the epiCMIT is a strong and independent prognostic variable from the tumor cellular origin and genomic complexity in both CLL and MCL, improving the prognostic power of these two well-established prognostic variables. The epiCMIT probably captures all the events leading to a proliferative advantage of malignant cells, which in part are intrinsic lesions such as genetic alterations, and in part are extrinsic factors such as microenvironmental interactions. In fact, some genetic driver alterations seem particularly associated with high epiCMIT (**Fig. 38**) and target important pathways related to CLL and MCL pathogenesis, such as DNA damage or survival and proliferation. Notably, these epiCMIT-driver alterations are among the most informative ones for patient outcome. In CLL, the accumulation of epiCMIT-drivers showed an independent prognostic impact from the epiCMIT itself, which probably indicates that collectively they have other biological effects on the fitness of CLL cells beyond that of conferring a proliferative advantage. Interestingly, the CLL-epiCMIT-drivers *SF3B1* and *ATM* are significantly associated with high epiCMIT only in the poorly characterized i-CLL subtype and define a distinct i-CLL subgroup with high levels of epiCMIT and poor clinical outcome. This group of patients may overlap with a recently reported subset of i-CLL patients with a bias towards the λ light chain IGLV3-21 [259], a immunoglobulin gene rearrangement previously related to an inferior outcome [379,380].

Taken together, these observations may suggest the presence of a unique i-CLL subgroup with poor clinical outcome characterized by high epiCMIT, the presence of *SF3B1* and *ATM* alterations, and IGLV3-21 usage. In the case of MCL, the accumulation of any of the 8 epiCMIT-drivers seems to capture both genetic complexity and the proliferative history of MCL cells, and overcome the potent prognostic impact of both CNA and epiCMIT. Although the prognostic impact of epiCMIT-drivers in both CLL and MCL should be validated in extended validation series, based on the permutation analyses as well as on their strong clinical associations I strongly believe that they will represent important assets for a more reduced gene panels with high prognostic power.

In addition to these new insights that point to future research directions, the identification of driver alterations linked to high epiCMIT may also help to better understand previous results in both CLL and MCL patients. In CLL, a previous study found a coevolution between the acquisition of certain genetic alterations and methylation changes [255]. The authors noticed a prominent increase of DNA methylation changes between time points in serial CLL, but only in patients with clinical progression and acquiring particular genetic alterations such as *TP53* or *SF3B1*. Interestingly, these two genetic alterations are strongly associated with high epiCMIT in the studied CLL cohort (Study 4). Given the extreme correlation between the epiCMIT with the total number of DNA methylation changes in all B-cell tumors, these previous results most likely reflect an increased proliferative history of CLL cells after acquiring genetic lesions conferring a proliferative advantage to CLL cells. In line with these results, a collaborative study with the group of Kostas Stamatopoulos in which I participated (shown in appendix, others), recently demonstrated that a shorter time to treatment for CLL patients with the highest number of DNA methylation changes [381]. In turn, this finding is congruent with my results in MCL in the Study 1, where patients with higher number of DNA methylation changes show an impaired clinical outcome. Finally, previous studies from my host lab detected aberrant DNA methylation of polycomb genes and linked it to the presence of a CIMP in some MCL patients [230,233]. This group of patients showed high genomic complexity, associations with the *CDKN2A* locus, high Ki67 expression, and an adverse clinical outcome, which altogether represent clinico-biological features related to high epiCMIT levels. The epiCMIT is highly correlated with the CIMP in MCL. Thus, the DNA methylation patterns previously observed in a subset of MCL patients most likely relate to an increased proliferative history of MCL cells upon the acquisition of certain genetic lesions such as *CDKN2A* or *TP53*, which are strongly associated with high epiCMIT [229].

Collectively, these results indicate that the complex clinical evolution of CLL and MCL patients is better recognized by the integration of genomic and epigenomic aberrations, representing the epiCMIT mitotic clock an important new piece of the puzzle. Furthermore, the epiCMIT may highlight driver alterations which confer a particularly poor clinical outcome to both CLL and MCL patients and thus represent strong candidates for extended analyses in large patient cohorts.

4. Mitotic activity and aging, two sides of the same coin?

The data reported in this doctoral thesis describe that the majority of DNA methylation changes in normal and neoplastic B cells, and in particular those of the individual-specific dimension, are accumulated upon rounds of cell division at polycomb-repressed and heterochromatic regions without affecting gene expression. Most remarkably, previous studies also related the very same DNA methylation changes to the process of chronological aging [316,317,351–353]. Some of these initial studies were done in blood and may be affected by different cellular compositions [286], which indeed are modulated as individuals get older. For instance, the ratio memory/naive B cells in newborns is very low and increases with age [382]. Thus, as memory B cells have an extensive proliferative history behind them, it is likely to speculate that the observed link between DNA methylation and aging is in part related to cell composition. However, the large sample size of these studies as well as the high number of cell types analyzed also support that these DNA methylation changes also relate to the aging process itself and not to aging-associated changes in cell composition. In the **Study 3**, I extensively analyzed the relationship between mitotic history and aging in normal and neoplastic B cells comparing the epiCMIT mitotic clock and the highly accurate Horvath chronological aging clock [317], respectively. Overall, epiCMIT and Horvath clocks seem to reflect broadly different epigenetic information imprinted onto the DNA methylome. This notion is supported by multiple perspectives, including the similar epiCMIT levels in the same normal B-cell subpopulations regardless of donor's age, the different (epi)genomic and transcriptomic features between the epiCMIT and the Horvath clocks, and the fact that the prognostic value of the epiCMIT in B-cell tumors is independent from age. Despite this seemingly independence of mitotic and aging clocks, I did observe a remarkable association between the epiCMIT and the age predicted by the Horvath clock in B-cell tumors. This finding suggests that the accelerated age reported in human cancer [317] may

actually reflect an enhanced mitotic activity of cancer cells, which in turn may explain the increased predicted age of cells upon *in vitro* passages [317]. Thus, this finding may also indicate that the DNA methylation dynamics during mitotic activity and aging are probably interconnected [383], and cancer cells may emerge as a paradigmatic model of this relationship (**Fig. 38**).

During heathy aging, the organism needs to ensure tissue maintenance and cellular function through a constant rate of cellular regeneration from an expanding compartment derived from stem cells (**Fig. 39**). In fact, even though the biological definition of aging is an active area of research, it seems established that stem cell exhaustion is one of the hallmarks of aging [384–387]. Related to this, it has been reported that a higher number of stem cell divisions accumulates during aging and is related to higher cancer incidence [388–390]. In turn, the number of stem cell divisions correlates with aberrant CGI methylation at polycomb regions [391], and PMD hypomethylation begins during embryonic development and is increased in differentiated, aged, and cancerous tissues [316,353]. Taken together, these data suggest that the epiCMIT mitotic clock captures more extreme methylation changes related to cell division in differentiated and cancerous tissues. In contrast, the Horvath clock may capture more subtle methylation changes related to stem cell divisions to ensure tissue maintenance, which are then manifested in all the distinct differentiated cell types of the organism and relate to human chronological age (**Fig. 39**). Under physiological conditions, there may be a normal and constant rate of cellular demand during the whole human lifespan. This demand can potentially be increased upon different developmental and environmental cues or pathological conditions such as chronic infections and inflammation, cardiovascular disease, smoking, frailty, stress, obesity, cancer, Down and Werner syndromes, among others, which ultimately may lead to a stem cell exhaustion and an aging phenotype (**Fig. 39**). In fact, all these previous conditions show an increased age predicted by the Horvath clock (extensively reviewed in reference [341]), whose acceleration is not only related to a higher cancer risk and shorter cancer survival [392], but also to all-cause mortality when it is accelerated in blood after adjusting for known cell types [393,394]. This observation in blood may respond to a higher demand of blood production from the HPC -a truly aged cell type- upon challenging conditions that may lead to a clonal expansion of precursor cells to reach the cellular requirements, which ultimately associate with an aging phenotype and an adverse outcome. Supporting this, the over-estimation of the chronological age in bone marrow failure syndromes is

indicative for exhaustion of the hematopoietic stem cell pool [395]. Collectively, these findings may also be linked to several studies showing that clonal hematopoiesis taking place in aging relates to higher cancer incidence and to all-cause mortality [145–147,396,397].

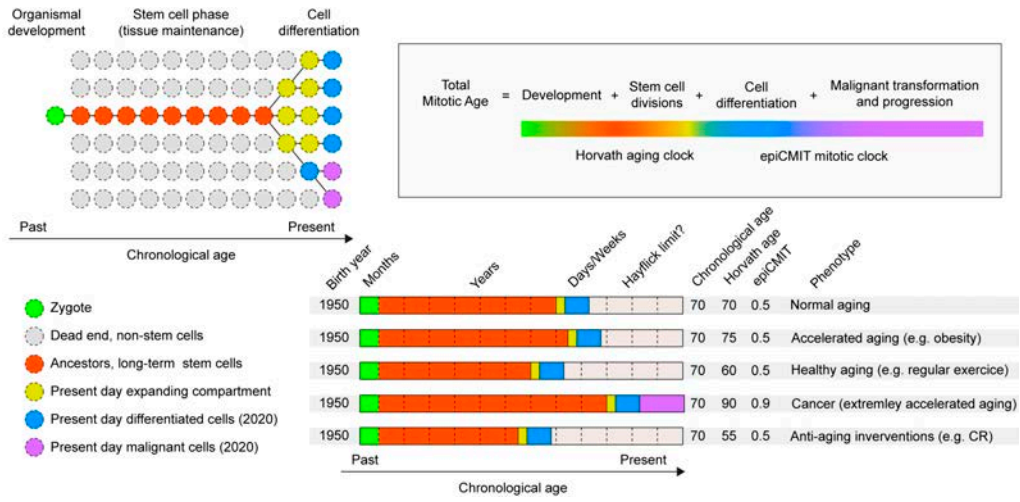
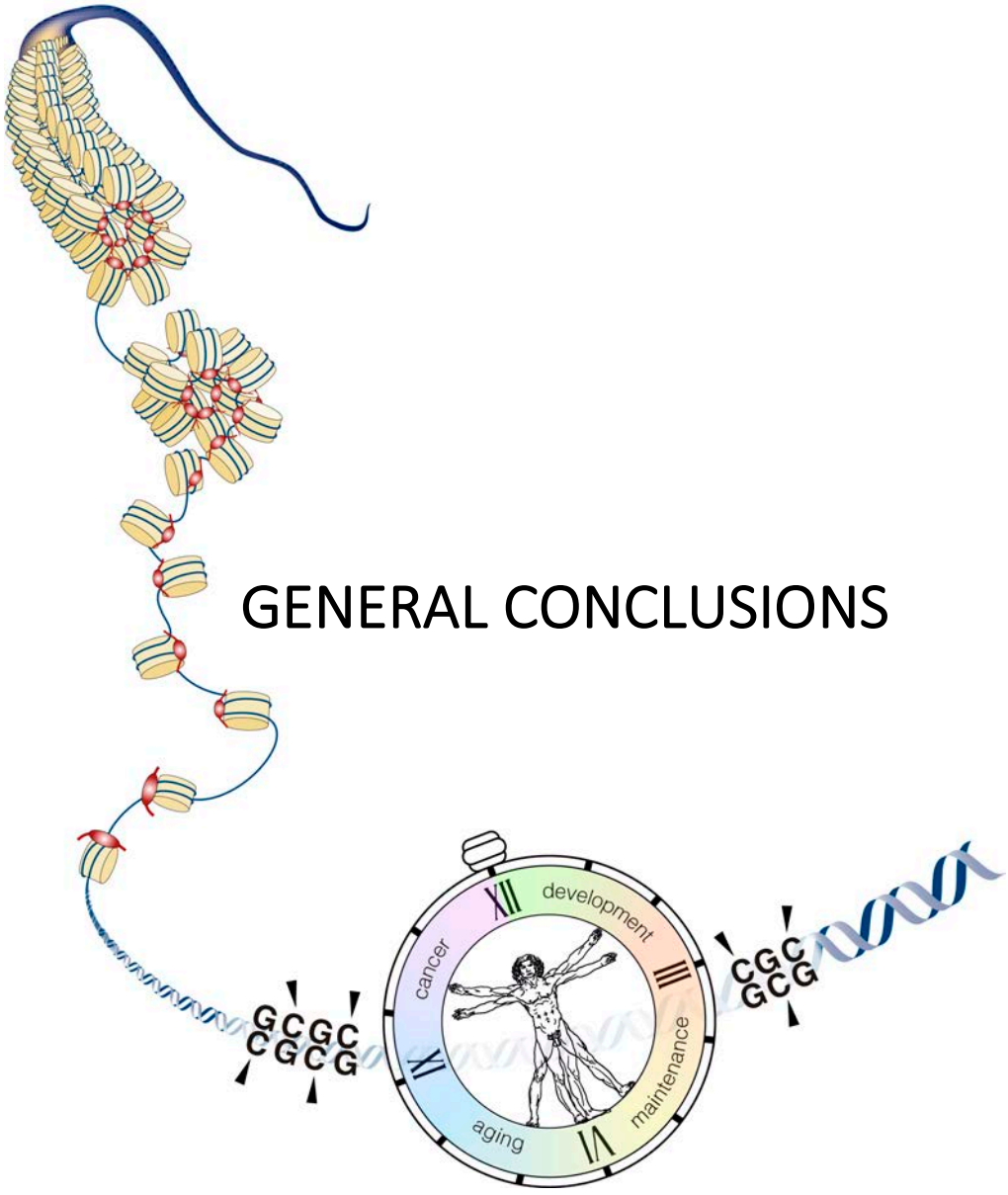


Fig.39 | Crosstalk between mitotic activity and human aging. The total mitotic age of a given cell type include the cell divisions accumulated from embryonic development, stem cell divisions, cell differentiation and eventually from the malignant transformation and progression. The stem cells represent truly aged cells, which ensure tissue maintenance during human lifespan. This maintenance must imply a constant rate of stem cell divisions (which may be limited, i.e. Hayflick limit) followed by an expanding compartment to give rise to more differentiated and effector cell types. The DNA methylation changes related to this constant rate of stem cell divisions may be captured by the Horvath clock, and relates to human health- and lifespan. In contrast, differentiated and malignant cells show most notorious DNA methylation changes in some genomic locations shared with the Horvath clock, which are captured by the epiCMIT mitotic clock. Five different donors with the same age and cellular type analyzed but with different phenotypes are shown. An illustrative value for Horvath and epiCMIT clocks is shown for each phenotype condition, including healthy, normal and accelerated aging, pathological aging due to cancer, and one anti-aging intervention such as caloric restriction (CR). Inspired by Shibata 2009 [398].

In agreement with the data shown in the previous paragraph, the DNA methylome of HPC in old mice can be recapitulated in HPC from young mice by enforced proliferation, which ultimately associates with a decline in HPC function similar to that observed in HPC from old mice [343]. Similarly, two independent seminal studies demonstrated that replicative stress is a potent driver of functional decline in aging HPC [399], which possibly is associated with an increased DNA damage [400]. In fact, it has been

shown that repressive complexes formed by polycomb and *DNMTs* target sites of both DNA damage, in the form of oxidative damage or double strand breaks, and CpG islands [401–404], which possibly become methylated thereafter [405]. Mechanistically, this DNA methylation may be the result of an active, energetically-demanding process that requires to stop transcription or replication upon DNA damage of the surrounding genomic region. An immediate side effect of this could be the methylation at promoters of tumor suppressor genes [123], which may be prevented by TET proteins under normal physiological conditions [406,407] or facilitated under harsh conditions such as high glucose levels [408]. Thus, the hypermethylation at these sites may represent an epigenetic memory of past DNA damage that may have taken place in the context of cell division, regardless if the damage was successfully repaired or not. These methylation changes have been referred in literature as epigenetic drift or aging-like methylation changes, and have been shown to prelude disease onset and predict disease course [342,354,383,409]. This epigenetic drift probably defines lifespan and is conserved within species [410], and more interestingly and perhaps as a proof of principle, it is ameliorated by pro-longevity interventions such as caloric restriction in mice [411–415] and rhesus monkeys [410] (Fig. 39). To the best of my knowledge, although caloric restriction may also induce this phenomenon in humans [416], compelling data are not yet available.

In summary, the data shown above indicate that although, broadly-speaking, the epiCMIT and Horvath clocks seem to capture different biological phenomena, a possible crosstalk between DNA methylation changes observed in proliferating and aging cells seems to exist. The complete understanding of this interconnexion is of high relevance for geroscience and biomedicine as a whole, and will need the thorough evaluation of the DNA methylome of truly aged cells such as HPC under controlled settings. The calculation of epiCMIT and Horvath clocks in extended experimental settings in young and aged cells upon enforced proliferation with different DNA damage levels will undoubtedly provide new valuable insights. Finally, multi-omic single-cell approaches [417] may also help to distill particular genetic and epigenetic features which may underlie mitotic activity and aging processes.



GENERAL CONCLUSIONS

This doctoral thesis is divided into four different studies, whose specific conclusions appear at the end of each respective chapter. Nonetheless, here I present the major general conclusions that can be drawn from the four studies as a whole:

1. The DNA methylome of B-cell tumors shows DNA methylation imprints from the normal cellular origin, DNA methylation changes shared with normal B-cell differentiation, and *de novo* disease-specific alterations, some of which target regulatory regions and relate to specific transcriptional profiles.

2. DNA methylation is a sound molecular mark to classify human neoplasias, as it is exemplified with the Pan B-cell tumor diagnostic tool which accurately identifies B-cell tumors entities and subtypes with different clinical management.

3. DNA methylation changes in individual patients take place mainly in silent chromatin without transcriptional consequences and relate to the normal and neoplastic cell proliferative history.

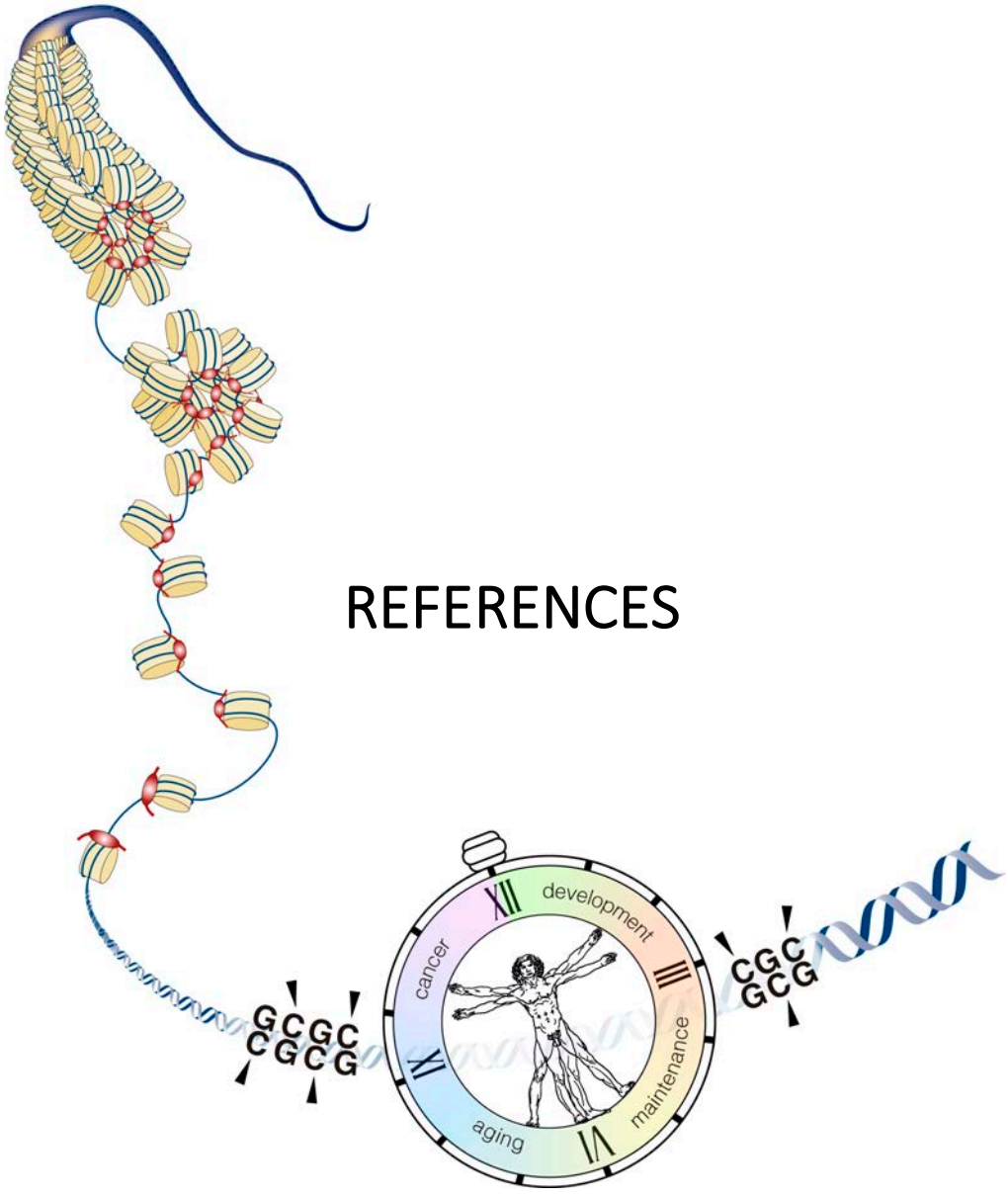
4. The epiCMIT epigenetic mitotic clock was developed and validated. It uses both hyper- and hypomethylation to accurately trace the proliferative history of B-cell neoplasms with different preferences to gain or lose DNA methylation.

5. The epiCMIT predicts patient outcome in B-cell tumors and in other human neoplasias.

6. Overall, DNA methylation changes related to the proliferative history seem to increase by genetic alterations with positive selection in CLL and MCL.

7. Genetic driver alterations associated with high proliferative history in CLL and MCL samples may represent robust candidates for extended clinical analyses in large patient cohorts.

8. Mitotic and aging epigenetic clocks generally reflect distinct biological phenomena, but they may be related in tumors, as the accelerated age observed in patient samples correlates with their proliferative history.


GCGC
CGCG
XII
development
III
maintenance
VI
aging
IX
cancer
CGC
GCG

REFERENCES

1. Waddington, C. H. The epigenotype. 1942. *Int. J. Epidemiol.* **41**, 10–3 (1942).
2. Baylin, S. B. & Jones, P. A. A decade of exploring the cancer epigenome - biological and translational implications. *Nat. Rev. Cancer* **11**, 726–34 (2011).
3. Stunnenberg, H. G., International Human Epigenome Consortium & Hirst, M. The International Human Epigenome Consortium: A Blueprint for Scientific Collaboration and Discovery. *Cell* **167**, 1145–1149 (2016).
4. Aguilar, C. A. & Craighead, H. G. Micro- and nanoscale devices for the investigation of epigenetics and chromatin dynamics. *Nat. Nanotechnol.* **8**, 709–18 (2013).
5. Sender, R., Fuchs, S. & Milo, R. Revised Estimates for the Number of Human and Bacteria Cells in the Body. *PLoS Biol.* **14**, e1002533 (2016).
6. Luger, K., Mäder, A. W., Richmond, R. K., Sargent, D. F. & Richmond, T. J. Crystal structure of the nucleosome core particle at 2.8 A resolution. *Nature* **389**, 251–60 (1997).
7. Olins, D. E. & Olins, A. L. Chromatin history: our view from the bridge. *Nat. Rev. Mol. Cell Biol.* **4**, 809–14 (2003).
8. Venkatesh, S. & Workman, J. L. Histone exchange, chromatin structure and the regulation of transcription. *Nat. Rev. Mol. Cell Biol.* **16**, 178–189 (2015).
9. Fyodorov, D. V., Zhou, B. R., Skoultchi, A. I. & Bai, Y. Emerging roles of linker histones in regulating chromatin structure and function. *Nat. Rev. Mol. Cell Biol.* **19**, 192–206 (2018).
10. Dekker, J., Rippe, K., Dekker, M. & Kleckner, N. Capturing chromosome conformation. *Science* **295**, 1306–11 (2002).
11. Denker, A. & de Laat, W. The second decade of 3C technologies: detailed insights into nuclear organization. *Genes Dev.* **30**, 1357–82 (2016).
12. Lieberman-Aiden, E. *et al.* Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326**, 289–293 (2009).
13. Dixon, J. R. *et al.* Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* **485**, 376–380 (2012).
14. Lupiáñez, D. G. *et al.* Disruptions of Topological Chromatin Domains Cause Pathogenic Rewiring of Gene-Enhancer Interactions. *Cell* **161**, 1012–1025 (2015).
15. Spielmann, M., Lupiáñez, D. G. & Mundlos, S. Structural variation in the 3D genome. *Nat. Rev. Genet.* **19**, 453–467 (2018).
16. Rao, S. S. P. *et al.* A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* **159**, 1665–80 (2014).
17. Pombo, A. & Dillon, N. Three-dimensional genome architecture: Players and mechanisms. *Nat. Rev. Mol. Cell Biol.* **16**, 245–257 (2015).
18. Diao, Y. *et al.* Chromatin architecture reorganization during stem cell differentiation. *Nature* **518**, 331–336 (2015).
19. Uhler, C. & Shivashankar, G. V. Regulation of genome organization and gene expression by nuclear mechanotransduction. *Nat. Rev. Mol. Cell Biol.* **18**, 717–727 (2017).
20. Klemm, S. L., Shipony, Z. & Greenleaf, W. J. Chromatin accessibility and the regulatory epigenome. *Nat. Rev. Genet.* **20**, 207–220 (2019).
21. Sabo, P. J. *et al.* Genome-scale mapping of DNase I sensitivity in vivo using tiling DNA microarrays. *Nat. Methods* **3**, 511–8 (2006).
22. Crawford, G. E. *et al.* DNase-chip: A high-resolution method to identify DNase I hypersensitive sites using tiled microarrays. *Nat. Methods* **3**, 503–509 (2006).
23. Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y. & Greenleaf, W. J. Transposition

of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat. Methods* **10**, 1213–8 (2013).
24. Kelly, T. K. *et al.* Genome-wide mapping of nucleosome positioning and DNA methylation within individual DNA molecules. *Genome Res.* **22**, 2497–506 (2012).
25. Krebs, A. R. *et al.* Genome-wide Single-Molecule Footprinting Reveals High RNA Polymerase II Turnover at Paused Promoters. *Mol. Cell* **67**, 411-422.e4 (2017).
26. Mueller, B. *et al.* MNase titration reveals differences between nucleosome occupancy and chromatin accessibility. *Nat. Commun.* **7**, 1–11 (2016).
27. Mueller, B. *et al.* Widespread changes in nucleosome accessibility without changes in nucleosome occupancy during a rapid transcriptional induction. *Genes Dev.* **31**, 451–462 (2017).
28. Thurman, R. E. *et al.* The accessible chromatin landscape of the human genome. *Nature* **489**, 75–82 (2012).
29. Heinz, S., Romanoski, C. E., Benner, C. & Glass, C. K. The selection and function of cell type-specific enhancers. *Nat. Rev. Mol. Cell Biol.* **16**, 144–154 (2015).
30. Corces, M. R. *et al.* Lineage-specific and single-cell chromatin accessibility charts human hematopoiesis and leukemia evolution. *Nat. Genet.* **48**, 1193–203 (2016).
31. Li, B., Carey, M. & Workman, J. L. The Role of Chromatin during Transcription. *Cell* **128**, 707–719 (2007).
32. Dogan, N. *et al.* Occupancy by key transcription factors is a more accurate predictor of enhancer activity than histone modifications or chromatin accessibility. *Epigenetics and Chromatin* **8**, 1–21 (2015).
33. Beekman, R. *et al.* The reference epigenome and regulatory chromatin landscape of chronic lymphocytic leukemia. *Nat. Med.* **24**, 868–880 (2018).
34. Kouzarides, T. Chromatin Modifications and Their Function. *Cell* **128**, 693–705 (2007).
35. Bannister, A. J. & Kouzarides, T. Regulation of chromatin by histone modifications. *Cell Res.* **21**, 381–395 (2011).
36. Lawrence, M., Daujat, S. & Schneider, R. Lateral Thinking: How Histone Modifications Regulate Gene Expression. *Trends Genet.* **32**, 42–56 (2016).
37. Farrelly, L. A. *et al.* Histone serotonylation is a permissive modification that enhances TFIID binding to H3K4me3. *Nature* **567**, 535–539 (2019).
38. Solomon, M. J., Larsen, P. L. & Varshavsky, A. Mapping protein-DNA interactions in vivo with formaldehyde: evidence that histone H4 is retained on a highly transcribed gene. *Cell* **53**, 937–47 (1988).
39. Barski, A. *et al.* High-Resolution Profiling of Histone Methylations in the Human Genome. *Cell* **129**, 823–837 (2007).
40. Jaffe, D. B. *et al.* Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature* **448**, 553–560 (2007).
41. Furey, T. S. ChIP-seq and beyond: New and improved methodologies to detect and characterize protein-DNA interactions. *Nat. Rev. Genet.* **13**, 840–852 (2012).
42. Mardis, E. R. & Ris, K. ChIP-seq: welcome to the new frontier. **4**, 613–614 (2007).
43. Roadmap Epigenomics Consortium *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317–30 (2015).
44. Ernst, J. *et al.* Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature* **473**, 43–49 (2011).
45. Ernst, J. & Kellis, M. ChromHMM: automating chromatin-state discovery and characterization. *Nat. Methods* **9**, 215–216 (2012).

46. Raiber, E.-A., Hardisty, R., van Delft, P. & Balasubramanian, S. Mapping and elucidating the function of modified bases in DNA. *Nat. Rev. Chem.* **1**, 0069 (2017).
47. Jones, P. a. Functions of DNA methylation: islands, start sites, gene bodies and beyond. *Nat. Rev. Genet.* **13**, 484–492 (2012).
48. Schübeler, D. Function and information content of DNA methylation. *Nature* **517**, 321–6 (2015).
49. Lister, R. *et al.* Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature* **462**, 315–322 (2009).
50. Kulis, M. *et al.* Whole-genome fingerprint of the DNA methylome during human B cell differentiation. *Nat. Genet.* **47**, 746–756 (2015).
51. Lister, R. *et al.* Global epigenomic reconfiguration during mammalian brain development. *Science* **341**, 1237905 (2013).
52. Ziller, M. J. *et al.* Charting a dynamic DNA methylation landscape of the human genome. *Nature* **500**, 477–81 (2013).
53. Lyko, F. The DNA methyltransferase family: A versatile toolkit for epigenetic regulation. *Nat. Rev. Genet.* **19**, 81–92 (2018).
54. Hermann, A., Goyal, R. & Jeltsch, A. The Dnmt1 DNA-(cytosine-C5)-methyltransferase Methylates DNA Processively with High Preference for Hemimethylated Target Sites. *J. Biol. Chem.* **279**, 48350–48359 (2004).
55. Bostick, M. *et al.* UHRF1 plays a role in maintaining DNA methylation in mammalian cells. *Science* **317**, 1760–4 (2007).
56. Sharif, J. *et al.* The SRA protein Np95 mediates epigenetic inheritance by recruiting Dnmt1 to methylated DNA. *Nature* **450**, 908–912 (2007).
57. Okano, M., Bell, D. W., Haber, D. A. & Li, E. DNA methyltransferases Dnmt3a and Dnmt3b are essential for de novo methylation and mammalian development. *Cell* **99**, 247–57 (1999).
58. Zhang, Z.-M. *et al.* Structural basis for DNMT3A-mediated de novo DNA methylation. *Nature* **554**, 387–391 (2018).
59. Kohli, R. M. & Zhang, Y. TET enzymes, TDG and the dynamics of DNA demethylation. *Nature* **502**, 472–479 (2013).
60. Wu, X. & Zhang, Y. TET-mediated active DNA demethylation: mechanism, function and beyond. *Nat. Rev. Genet.* **18**, 517–534 (2017).
61. Bauer, N. C., Corbett, A. H. & Doetsch, P. W. The current state of eukaryotic DNA base damage and repair. *Nucleic Acids Res.* **43**, 10083–10101 (2015).
62. Nabel, C. S. *et al.* AID/APOBEC deaminases disfavor modified cytosines implicated in DNA demethylation. *Nat. Chem. Biol.* **8**, 751–758 (2012).
63. Conticello, S. G. The AID/APOBEC family of nucleic acid mutators. *Genome Biol.* **9**, (2008).
64. Bachman, M. *et al.* 5-Formylcytosine can be a stable DNA modification in mammals. *Nat. Chem. Biol.* **advance on**, (2015).
65. Yang, X. *et al.* 5-Hydroxymethylcytosine is a predominantly stable DNA modification. *Nat. Chem.* **6**, 1049–1055 (2014).
66. Wang, D. *et al.* MAX is an epigenetic sensor of 5-carboxylcytosine and is altered in multiple myeloma. *Nucleic Acids Res.* **45**, 2396–2407 (2017).
67. Spruijt, C. G. *et al.* Dynamic readers for 5-(hydroxy)methylcytosine and its oxidized derivatives. *Cell* **152**, 1146–59 (2013).
68. Iurlaro, M. *et al.* A screen for hydroxymethylcytosine and formylcytosine binding

proteins suggests functions in transcription and chromatin regulation. *Genome Biol.* **14**, R119 (2013).
69. Ngo, T. T. M. *et al.* Effects of cytosine modifications on DNA flexibility and nucleosome mechanical stability. *Nat. Commun.* **7**, 10813 (2016).
70. Upadhyay, A. K. *et al.* Recognition and potential mechanisms for replication and erasure of cytosine hydroxymethylation. *Nucleic Acids Res.* **40**, 4841–4849 (2012).
71. Otani, J. *et al.* Cell cycle-dependent turnover of 5-hydroxymethyl cytosine in mouse embryonic stem cells. *PLoS One* **8**, 1–11 (2013).
72. Ji, D., Lin, K., Song, J. & Wang, Y. Effects of Tet-induced oxidation products of 5-methylcytosine on Dnmt1- and DNMT3a-mediated cytosine methylation. *Mol. Biosyst.* **10**, 1749–1752 (2014).
73. Aran, D., Toperoff, G., Rosenberg, M. & Hellman, A. Replication timing-related and gene body-specific methylation of active human genes. *Hum. Mol. Genet.* **20**, 670–680 (2011).
74. Ruppel, W. G. Zur chemie der tuberkelbacillen. *Z. Physiol. Chem* (1898).
75. Johnson, T. B. & Coghill, R. D. RESEARCHES ON PYRIMIDINES. C111. THE DISCOVERY OF 5-METHYL-CYTOSINE IN TUBERCULINIC ACID, THE NUCLEIC ACID OF THE TUBERCLE BACILLUS 1. *J. Am. Chem. Soc.* **47**, 2838–2844 (1925).
76. SRINIVASAN, P. R. & BOREK, E. ENZYMATIC ALTERATION OF NUCLEIC ACID STRUCTURE. *Science* **145**, 548–53 (1964).
77. Holliday, R. & Pugh, J. E. DNA modification mechanisms and gene activity during development. *Science* **187**, 226–32 (1975).
78. Riggs, A. D. X inactivation, differentiation, and DNA methylation. *Cytogenet. Cell Genet.* **14**, 9–25 (1975).
79. Elhamamsy, A. R. Role of DNA methylation in imprinting disorders: an updated review. *J. Assist. Reprod. Genet.* **34**, 549–562 (2017).
80. Razin, A. & Riggs, A. D. DNA methylation and gene function. *Science* **210**, 604–10 (1980).
81. Stein, R., Gruenbaum, Y., Pollack, Y., Razin, A. & Cedar, H. Clonal inheritance of the pattern of DNA methylation in mouse cells. *Proc. Natl. Acad. Sci. U. S. A.* **79**, 61–5 (1982).
82. Reik, W., Collick, A., Norris, M. L., Barton, S. C. & Surani, M. A. Genomic imprinting determines methylation of parental alleles in transgenic mice. *Nature* **328**, 248–251 (1987).
83. Bird, A. DNA methylation patterns and epigenetic memory. *Genes Dev.* **16**, 6–21 (2002).
84. Dor, Y. & Cedar, H. Principles of DNA methylation and their implications for biology and medicine. *Lancet* **392**, 777–786 (2018).
85. Plongthongkum, N., Diep, D. H. & Zhang, K. Advances in the profiling of DNA modifications: cytosine methylation and beyond. *Nat. Rev. Genet.* **15**, 647–61 (2014).
86. Luo, C., Hajkova, P. & Ecker, J. R. Dynamic DNA methylation: In the right place at the right time. *Science* **361**, 1336–1340 (2018).
87. Baubec, T. *et al.* Genomic profiling of DNA methyltransferases reveals a role for DNMT3B in genic methylation. *Nature* **520**, 243–247 (2015).
88. Neri, F. *et al.* Intragenic DNA methylation prevents spurious transcription initiation. *Nature* **543**, 72–77 (2017).
89. Maunakea, A. K. *et al.* Conserved role of intragenic DNA methylation in regulating alternative promoters. *Nature* **466**, 253–257 (2010).

90. Maunakea, A. K., Chepelev, I., Cui, K. & Zhao, K. Intragenic DNA methylation modulates alternative splicing by recruiting MeCP2 to promote exon recognition. *Cell Res.* **23**, 1256–1269 (2013).
91. Deniz, Ö., Frost, J. M. & Branco, M. R. Regulation of transposable elements by DNA modifications. *Nat. Rev. Genet.* **20**, 417–431 (2019).
92. Stadler, M. B. *et al.* Correction: Corrigendum: DNA-binding factors shape the mouse methylome at distal regulatory regions. *Nature* **484**, 550–550 (2012).
93. Hon, G. C. *et al.* Epigenetic memory at embryonic enhancers identified in DNA methylation maps from adult mouse tissues. *Nat. Genet.* **45**, 1198–206 (2013).
94. Flavahan, W. A. *et al.* Insulator dysfunction and oncogene activation in IDH mutant gliomas. *Nature* **529**, 110–4 (2016).
95. Zhu, H., Wang, G. & Qian, J. Transcription factors as readers and effectors of DNA methylation. *Nat. Rev. Genet.* **17**, 551–65 (2016).
96. Yin, Y. *et al.* Impact of cytosine methylation on DNA binding specificities of human transcription factors. *Science* **356**, (2017).
97. Jeong, M. *et al.* Large conserved domains of low DNA methylation maintained by Dnmt3a. *Nat. Genet.* **46**, 17–23 (2013).
98. Xie, W. *et al.* Epigenomic analysis of multilineage differentiation of human embryonic stem cells. *Cell* **153**, 1134–1148 (2013).
99. Li, Y. *et al.* Genome-wide analyses reveal a role of Polycomb in promoting hypomethylation of DNA methylation valleys. *Genome Biol.* **19**, 1–16 (2018).
100. Schultz, M. D. *et al.* Human body epigenome maps reveal noncanonical DNA methylation variation. *Nature* **523**, 212–6 (2015).
101. Salhab, A. *et al.* A comprehensive analysis of 195 DNA methylomes reveals shared and cell-specific features of partially methylated domains. *Genome Biol.* **19**, 9–11 (2018).
102. Smith, Z. D. *et al.* DNA methylation dynamics of the human preimplantation embryo. *Nature* **511**, 611–615 (2014).
103. Gu, H. *et al.* A unique regulatory phase of DNA methylation in the early mammalian embryo. *Nature* **484**, 339–344 (2012).
104. Krueger, F. *et al.* The Dynamics of Genome-wide DNA Methylation Reprogramming in Mouse Primordial Germ Cells. *Mol. Cell* **48**, 849–862 (2012).
105. Zhang, Y. *et al.* Dynamic epigenomic landscapes during early lineage specification in mouse embryos. *Nat. Genet.* **50**, 96–105 (2018).
106. Guo, H. *et al.* The DNA methylation landscape of human early embryos. *Nature* **511**, 606–10 (2014).
107. Bergman, Y. & Cedar, H. DNA methylation dynamics in health and disease. *Nat. Struct. Mol. Biol.* **20**, 274–281 (2013).
108. Smith, Z. D. & Meissner, A. DNA methylation: Roles in mammalian development. *Nat. Rev. Genet.* **14**, 204–220 (2013).
109. Li, E., Bestor, T. H. & Jaenisch, R. Targeted mutation of the DNA methyltransferase gene results in embryonic lethality. *Cell* **69**, 915–26 (1992).
110. Liao, J. *et al.* Targeted disruption of DNMT1, DNMT3A and DNMT3B in human embryonic stem cells. *Nat. Genet.* **47**, 469–78 (2015).
111. Trowbridge, J. J., Snow, J. W., Kim, J. & Orkin, S. H. DNA methyltransferase 1 is essential for and uniquely regulates hematopoietic stem and progenitor cells. *Cell Stem Cell* **5**, 442–9 (2009).
112. Sen, G. L., Reuter, J. A., Webster, D. E., Zhu, L. & Khavari, P. A. DNMT1 maintains

progenitor function in self-renewing somatic tissue. *Nature* **463**, 563–567 (2010).
113. Chen, T., Ueda, Y., Dodge, J. E., Wang, Z. & Li, E. Establishment and maintenance of genomic methylation patterns in mouse embryonic stem cells by Dnmt3a and Dnmt3b. *Mol. Cell. Biol.* **23**, 5594–605 (2003).
114. Challen, G. a *et al.* Dnmt3a is essential for hematopoietic stem cell differentiation. *Nat. Genet.* **44**, 23–31 (2011).
115. Challen, G. A. *et al.* Dnmt3a and Dnmt3b have overlapping and distinct functions in hematopoietic stem cells. *Cell Stem Cell* **15**, 350–364 (2014).
116. Jeong, M. *et al.* Loss of Dnmt3a Immortalizes Hematopoietic Stem Cells In Vivo. *Cell Rep.* **23**, 1–10 (2018).
117. Jaenisch, R. *et al.* Loss of genomic methylation causes p53-dependent apoptosis and epigenetic deregulation. *Nat. Genet.* **27**, 31–39 (2001).
118. Feinberg, A. P. The Key Role of Epigenetics in Human Disease Prevention and Mitigation. *N. Engl. J. Med.* **378**, 1323–1334 (2018).
119. Gama-Sosa, M. A. *et al.* The 5-methylcytosine content of DNA from human tumors. *Nucleic Acids Res.* **11**, 6883–94 (1983).
120. Feinberg, A. P. & Vogelstein, B. Hypomethylation distinguishes genes of some human cancers from their normal counterparts. *Nature* **301**, 89–92 (1983).
121. Greger, V., Passarge, E., Höpping, W., Messmer, E. & Horsthemke, B. Epigenetic changes may contribute to the formation and spontaneous regression of retinoblastoma. *Hum. Genet.* **83**, 155–8 (1989).
122. Sakai, T. *et al.* Allele-specific hypermethylation of the retinoblastoma tumor-suppressor gene. *Am. J. Hum. Genet.* **48**, 880–8 (1991).
123. Esteller, M. Epigenetics in cancer. *N. Engl. J. Med.* **358**, 1148–59 (2008).
124. Baylin, S. B. & Ohm, J. E. Epigenetic gene silencing in cancer - A mechanism for early oncogenic pathway addiction? *Nat. Rev. Cancer* **6**, 107–116 (2006).
125. Zoghbi, H. Y. & Beaudet, A. L. Epigenetics and human disease. *Cold Spring Harb. Perspect. Biol.* **8**, 1–28 (2016).
126. Portela, A. & Esteller, M. Epigenetic modifications and human disease. *Nat. Biotechnol.* **28**, 1057–1068 (2010).
127. Feinberg, A. P. & Tycko, B. The history of cancer epigenetics. *Nat. Rev. Cancer* **4**, 143–153 (2004).
128. Esteller, M. Cancer epigenomics: DNA methylomes and histone-modification maps. *Nat. Rev. Genet.* **8**, 286–298 (2007).
129. Xu, G. L. *et al.* Chromosome instability and immunodeficiency syndrome caused by mutations in a DNA methyltransferase gene. *Nature* **402**, 187–91 (1999).
130. Gaudet, F. *et al.* Induction of tumors in mice by genomic hypomethylation. *Science* **300**, 489–92 (2003).
131. Heyn, H. *et al.* Epigenomic analysis detects aberrant super-enhancer DNA methylation in human cancer. *Genome Biol.* **17**, 11 (2016).
132. Domcke, S. *et al.* Competition between DNA methylation and transcription factors determines binding of NRF1. *Nature* **528**, 575–579 (2015).
133. Schlesinger, Y. *et al.* Polycomb-mediated methylation on Lys27 of histone H3 pre-marks genes for de novo methylation in cancer. *Nat. Genet.* **39**, 232–236 (2007).
134. Ohm, J. E. *et al.* A stem cell–like chromatin pattern may predispose tumor suppressor genes to DNA hypermethylation and heritable silencing. *Nat. Genet.* **39**, 237–242 (2007).

135. Widschwendter, M. *et al.* Epigenetic stem cell signature in cancer. *Nat. Genet.* **39**, 157–158 (2007).
136. Campbell, P. J. *et al.* Pan-cancer analysis of whole genomes. *Nature* **578**, 82–93 (2020).
137. Shen, H. & Laird, P. W. Interplay between the cancer genome and epigenome. *Cell* **153**, 38–55 (2013).
138. Issa, J. CpG island methylator phenotype in cancer. *Nat. Rev. Cancer* **4**, 988–93 (2004).
139. Weisenberger, D. J. *et al.* CpG island methylator phenotype underlies sporadic microsatellite instability and is tightly associated with BRAF mutation in colorectal cancer. *Nat. Genet.* **38**, 787–793 (2006).
140. Hinoue, T. *et al.* Genome-scale analysis of aberrant DNA methylation in colorectal cancer. *Genome Res.* **22**, 271–282 (2012).
141. Noushmehr, H. *et al.* Identification of a CpG Island Methylator Phenotype that Defines a Distinct Subgroup of Glioma. *Cancer Cell* **17**, 510–522 (2010).
142. Sturm, D. *et al.* Hotspot mutations in H3F3A and IDH1 define distinct epigenetic and biological subgroups of glioblastoma. *Cancer Cell* **22**, 425–37 (2012).
143. Turcan, S. *et al.* IDH1 mutation is sufficient to establish the glioma hypermethylator phenotype. *Nature* **483**, 479–483 (2012).
144. Figueroa, M. E. *et al.* Leukemic IDH1 and IDH2 mutations result in a hypermethylation phenotype, disrupt TET2 function, and impair hematopoietic differentiation. *Cancer Cell* **18**, 553–67 (2010).
145. Genovese, G. *et al.* Clonal hematopoiesis and blood-cancer risk inferred from blood DNA sequence. *N. Engl. J. Med.* **371**, 2477–87 (2014).
146. Jaiswal, S. *et al.* Age-related clonal hematopoiesis associated with adverse outcomes. *N. Engl. J. Med.* **371**, 2488–2498 (2014).
147. Xie, M. *et al.* Age-related mutations associated with clonal hematopoietic expansion and malignancies. *Nat. Med.* **20**, 1472–1478 (2014).
148. Shlush, L. I. *et al.* Identification of pre-leukaemic haematopoietic stem cells in acute leukaemia. *Nature* **506**, 328–333 (2014).
149. Yang, L., Rau, R. & Goodell, M. A. DNMT3A in haematological malignancies. *Nat. Rev. Cancer* **15**, 152–165 (2015).
150. Spencer, D. H. *et al.* CpG Island Hypermethylation Mediated by DNMT3A Is a Consequence of AML Progression. *Cell* **168**, 801-816.e13 (2017).
151. Berney, M. & McGouran, J. F. Methods for detection of cytosine and thymine modifications in DNA. *Nat. Rev. Chem.* **2**, 332–348 (2018).
152. Laird, P. W. Principles and challenges of genomewide DNA methylation analysis. *Nat. Rev. Genet.* **11**, 191–203 (2010).
153. Liu, Y. *et al.* Bisulfite-free direct detection of 5-methylcytosine and 5-hydroxymethylcytosine at base resolution. *Nat. Biotechnol.* **37**, 424–429 (2019).
154. Huang, T. H. *et al.* Identification of DNA methylation markers for human breast carcinomas using the methylation-sensitive restriction fingerprinting technique. *Cancer Res.* **57**, 1030–4 (1997).
155. Yan, P. S., Potter, D., Deatherage, D. E., Huang, T. H.-M. & Lin, S. Differential methylation hybridization: profiling DNA methylation with a high-density CpG island microarray. *Methods Mol. Biol.* **507**, 89–106 (2009).
156. Khulan, B. Comparative isoschizomer profiling of cytosine methylation: The HELP assay. *Genome Res.* **16**, 1046–1055 (2006).
157. Martin-Subero, J. I. *et al.* A comprehensive microarray-based DNA methylation study of

367 hematological neoplasms. *PLoS One* **4**, e6986 (2009).
158. Bibikova, M. *et al.* High density DNA methylation array with single CpG site resolution. *Genomics* **98**, 288–95 (2011).
159. Aryee, M. J. *et al.* Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. *Bioinformatics* **30**, 1363–9 (2014).
160. Burger, L., Gaidatzis, D., Schübeler, D. & Stadler, M. B. Identification of active regulatory regions from DNA methylation data. *Nucleic Acids Res.* **41**, (2013).
161. Akalin, A. *et al.* Base-Pair Resolution DNA Methylation Sequencing Reveals Profoundly Divergent Epigenetic Landscapes in Acute Myeloid Leukemia. *PLoS Genet.* **8**, e1002781 (2012).
162. Gu, H. *et al.* Preparation of reduced representation bisulfite sequencing libraries for genome-scale DNA methylation profiling. *Nat. Protoc.* **6**, 468–81 (2011).
163. Meissner, A. *et al.* Reduced representation bisulfite sequencing for comparative high-resolution DNA methylation analysis. *Nucleic Acids Res.* **33**, 5868–77 (2005).
164. Swerdlow SH, Campo E, Harris NL, Jaffe ES, Pileri SA, Stein H, T. J. *WHO Classification of Tumours of Haematopoietic and Lymphoid Tissues*. (International Agency for Research on Cancer (IARC), 2017).
165. Parkin, J. & Cohen, B. An overview of the immune system. *Lancet (London, England)* **357**, 1777–89 (2001).
166. Brodin, P. *et al.* Variation in the human immune system is largely driven by non-heritable influences. *Cell* **160**, 37–47 (2015).
167. Morrison, S. J. & Scadden, D. T. The bone marrow niche for haematopoietic stem cells. *Nature* **505**, 327–34 (2014).
168. Laurenti, E. & Göttgens, B. From haematopoietic stem cells to complex differentiation landscapes. *Nature* **553**, 418–426 (2018).
169. Catlin, S. N., Busque, L., Gale, R. E., Guttorp, P. & Abkowitz, J. L. The replication rate of human hematopoietic stem cells in vivo. *Blood* **117**, 4460–4466 (2011).
170. Lee-Six, H. *et al.* Population dynamics of normal human blood inferred from somatic mutations. *Nature* **561**, 473–478 (2018).
171. Matthias, P. & Rolink, A. G. Transcriptional networks in developing and mature B cells. *Nat. Rev. Immunol.* **5**, 497–508 (2005).
172. Orkin, S. H. & Zon, L. I. Hematopoiesis: an evolving paradigm for stem cell biology. *Cell* **132**, 631–44 (2008).
173. Cedar, H. & Bergman, Y. Epigenetics of haematopoietic cell development. *Nat. Rev. Immunol.* **11**, 478–488 (2011).
174. Kurosaki, T., Shinohara, H. & Baba, Y. B Cell Signaling and Fate Decision. *Annu. Rev. Immunol.* **28**, 21–55 (2010).
175. Recaldin, T. & Fear, D. J. Transcription factors regulating B cell fate in the germinal centre. *Clin. Exp. Immunol.* **183**, 65–75 (2016).
176. Nutt, S. L. & Kee, B. L. The Transcriptional Regulation of B Cell Lineage Commitment. *Immunity* **26**, 715–725 (2007).
177. Sankaran, V. G. & Weiss, M. J. Anemia: Progress in molecular mechanisms and therapies. *Nat. Med.* **21**, 221–230 (2015).
178. Cooper, M. D. The early history of B cells. *Nat. Rev. Immunol.* **15**, 191–7 (2015).
179. Nemazee, D. Mechanisms of central tolerance for B cells. *Nat. Rev. Immunol.* **17**, 281–294 (2017).
180. Okada, T. *et al.* Chemokine Requirements for B Cell Entry to Lymph Nodes and Peyer's

Patches. *J. Exp. Med.* **196**, 65–75 (2002).
181. Mesin, L., Ersching, J. & Victora, G. D. Germinal Center B Cell Dynamics. *Immunity* **45**, 471–482 (2016).
182. De Silva, N. S. & Klein, U. Dynamics of B cells in germinal centres. *Nat. Rev. Immunol.* **15**, 137–48 (2015).
183. Kurosaki, T., Kometani, K. & Ise, W. Memory B cells. *Nat. Rev. Immunol.* **15**, 149–59 (2015).
184. Roco, J. A. *et al.* Class-Switch Recombination Occurs Infrequently in Germinal Centers. *Immunity* **51**, 337-350.e7 (2019).
185. Nutt, S. L., Hodgkin, P. D., Tarlinton, D. M. & Corcoran, L. M. The generation of antibody-secreting plasma cells. *Nat. Rev. Immunol.* **15**, 160–171 (2015).
186. Cyster, J. G. & Allen, C. D. C. B Cell Responses: Cell Interaction Dynamics and Decisions. *Cell* **177**, 524–540 (2019).
187. Oakes, C. C. *et al.* DNA methylation dynamics during B cell maturation underlie a continuum of disease phenotypes in chronic lymphocytic leukemia. *Nat. Genet.* **48**, 253–64 (2016).
188. Oakes, C. C. & Martin-Subero, J. I. Insight into origins, mechanisms & utility of DNA methylation in B cell malignancies. *Blood* **132**, blood-2018-02-692970 (2018).
189. Martin-Subero, J. I. & Oakes, C. C. Charting the dynamic epigenome during B-cell development. *Semin. Cancer Biol.* **51**, 139–148 (2018).
190. Küppers, R. & Dalla-Favera, R. Mechanisms of chromosomal translocations in B cell lymphomas. *Oncogene* **20**, 5580–5594 (2001).
191. Küppers, R. Mechanisms of B-cell lymphoma pathogenesis. *Nat. Rev. Cancer* **5**, 251–262 (2005).
192. Küppers, R. & Dalla-Favera, R. Mechanisms of chromosomal translocations in B cell lymphomas. *Oncogene* **20**, 5580–94 (2001).
193. Basso, K. & Dalla-Favera, R. Germinal centres and B cell lymphomagenesis. *Nat. Rev. Immunol.* **15**, 172–184 (2015).
194. Alizadeh, A. A. *et al.* Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature* **403**, 503–511 (2000).
195. Bosch, F. & Dalla-Favera, R. Chronic lymphocytic leukaemia: from genetics to treatment. *Nat. Rev. Clin. Oncol.* **16**, 684–701 (2019).
196. Kipps, T. J. *et al.* Chronic lymphocytic leukaemia. *Nat. Rev. Dis. Prim.* **3**, (2017).
197. Hunger, S. P. & Mullighan, C. G. Acute Lymphoblastic Leukemia in Children. *N. Engl. J. Med.* **373**, 1541–52 (2015).
198. Inaba, H., Greaves, M. & Mullighan, C. G. Acute lymphoblastic leukaemia. *Lancet* **381**, 1943–1955 (2013).
199. Iacobucci, I. & Mullighan, C. G. Genetic basis of acute lymphoblastic leukemia. *J. Clin. Oncol.* **35**, 975–983 (2017).
200. Roberts, K. G. & Mullighan, C. G. Genomics in acute lymphoblastic leukaemia: Insights and treatment implications. *Nat. Rev. Clin. Oncol.* **12**, 344–357 (2015).
201. Mann, G. *et al.* Improved outcome with hematopoietic stem cell transplantation in a poor prognostic subgroup of infants with mixed-lineage-leukemia (MLL)-rearranged acute lymphoblastic leukemia: results from the Interfant-99 Study. *Blood* **116**, 2644–50 (2010).
202. Vora, A. *et al.* Treatment reduction for children and young adults with low-risk acute lymphoblastic leukaemia defined by minimal residual disease (UKALL 2003): a

randomised controlled trial. *Lancet. Oncol.* **14**, 199–209 (2013).
203. Paulsson, K. *et al.* The genomic landscape of high hyperdiploid childhood acute lymphoblastic leukemia. *Nat. Genet.* **47**, 672–676 (2015).
204. Felice, M. S. *et al.* Prognostic impact of t(1;19)/ TCF3-PBX1 in childhood acute lymphoblastic leukemia in the context of Berlin-Frankfurt-Münster-based protocols. *Leuk. Lymphoma* **52**, 1215–21 (2011).
205. Li, Y. *et al.* Constitutional and somatic rearrangement of chromosome 21 in acute lymphoblastic leukaemia. *Nature* **508**, 98–102 (2014).
206. Nordlund, J. & Syvänen, A.-C. Epigenetics in pediatric acute lymphoblastic leukemia. *Semin. Cancer Biol.* **51**, 129–138 (2018).
207. Nordlund, J. *et al. Genome-wide signatures of differential DNA methylation in pediatric acute lymphoblastic leukemia. Genome biology* **14**, (2013).
208. Lee, S.-T. *et al.* Epigenetic remodeling in B-cell acute lymphoblastic leukemia occurs in two tracks and employs embryonic stem cell-like signatures. *Nucleic Acids Res.* **43**, 2590–602 (2015).
209. Wahlberg, P. *et al.* DNA methylome analysis of acute lymphoblastic leukemia cells reveals stochastic de novo DNA methylation in CpG islands. *Epigenomics* **8**, 1367–1387 (2016).
210. Nordlund, J. *et al.* DNA methylation-based subtype prediction for pediatric acute lymphoblastic leukemia. *Clin. Epigenetics* **7**, 11 (2015).
211. Borssén, M. *et al.* DNA methylation holds prognostic information in relapsed precursor B-cell acute lymphoblastic leukemia. *Clin. Epigenetics* **10**, 31 (2018).
212. Jares, P., Colomer, D. & Campo, E. Molecular pathogenesis of mantle cell lymphoma. *J. Clin. Invest.* **122**, 3416–23 (2012).
213. Martín-Garcia, D. *et al.* CCND2 and CCND3 hijack immunoglobulin light-chain enhancers in cyclin D1- mantle cell lymphoma. *Blood* **133**, 940–951 (2019).
214. Jares, P., Colomer, D. & Campo, E. Genetic and molecular pathogenesis of mantle cell lymphoma: perspectives for new targeted therapeutics. *Nat. Rev. Cancer* **7**, 750–62 (2007).
215. Campo, E. & Rule, S. Mantle cell lymphoma: Evolving management strategies. *Blood* **125**, 48–55 (2015).
216. Navarro, A. *et al.* Molecular subsets of mantle cell lymphoma defined by the IGHV mutational status and SOX11 expression have distinct biologic and clinical features. *Cancer Res.* **72**, 5307–5316 (2012).
217. Royo, C. *et al.* Non-nodal type of mantle cell lymphoma is a specific biological and clinical subgroup of the disease. *Leukemia* **26**, 1895–1898 (2012).
218. Fernàndez, V. *et al.* Genomic and gene expression profiling defines indolent forms of mantle cell lymphoma. *Cancer Res.* **70**, 1408–1418 (2010).
219. Palomero, J. *et al.* SOX11 promotes tumor angiogenesis through transcriptional regulation of PDGFA in mantle cell lymphoma. *Blood* **124**, 2235–2247 (2014).
220. Balsas, P. *et al.* SOX11 promotes tumor protective microenvironment interactions through CXCR4 and FAK regulation in mantle cell lymphoma. *Blood* **130**, 501–513 (2017).
221. Palomero, J. *et al.* SOX11 defines two different subtypes of mantle cell lymphoma through transcriptional regulation of BCL6. *Leukemia* **30**, 1580–1599 (2016).
222. Palomero, J. *et al.* SOX11 defines two different subtypes of mantle cell lymphoma through transcriptional regulation of BCL6. *Leukemia* **519**, 1–3 (2015).

223. Vegliante, M. C. *et al.* Plenary Paper SOX11 regulates PAX5 expression and blocks terminal B-cell differentiation in aggressive mantle cell lymphoma. *Blood* **121**, 2175–2186 (2014).
224. Vegliante, M. C. *et al.* Epigenetic activation of SOX11 in Lymphoid Neoplasms by Histone modifications. *PLoS One* **6**, (2011).
225. Bodrug, S. E. *et al.* Cyclin D1 transgene impedes lymphocyte maturation and collaborates in lymphomagenesis with the myc gene. *EMBO J.* **13**, 2124–2130 (2018).
226. Lovec, H., Grzeschiczek, A., Kowalski, M. B. & Möröy, T. Cyclin D1/bcl-1 cooperates with myc genes in the generation of B-cell lymphoma in transgenic mice. *EMBO J.* **13**, 3487–95 (1994).
227. Royo, C. *et al.* Seminars in Cancer Biology The complex landscape of genetic alterations in mantle cell lymphoma. *Semin. Cancer Biol.* **21**, 322–334 (2011).
228. Beà, S. *et al.* Landscape of somatic mutations and clonal evolution in mantle cell lymphoma. *Proc. Natl. Acad. Sci. U. S. A.* **110**, 18250–5 (2013).
229. Nadeu, F. *et al.* Genomic and epigenomic insights into the origin, pathogenesis, and clinical behavior of mantle cell lymphoma subtypes. *Blood* **136**, 1419–1432 (2020).
230. Enjuanes, A. *et al.* Identification of methylated genes associated with aggressive clinicopathological features in mantle cell lymphoma. *PLoS One* **6**, e19736 (2011).
231. Leshchenko, V. V. *et al.* Genomewide DNA methylation analysis reveals novel targets for drug development in mantle cell lymphoma. *Blood* **116**, 1025–1034 (2010).
232. Halldórsdóttir, A. M. *et al.* Mantle cell lymphoma displays a homogenous methylation profile: A comparative analysis with chronic lymphocytic leukemia. *Am. J. Hematol.* **87**, 361–367 (2012).
233. Enjuanes, A. *et al.* Genome-wide methylation analyses identify a subset of mantle cell lymphoma with a high number of methylated CpGs and aggressive clinicopathological features. *Int. J. Cancer* **133**, 2852–2863 (2013).
234. Strati, P. & Shanafelt, T. D. Monoclonal B-cell lymphocytosis and early-stage chronic lymphocytic leukemia: Diagnosis, natural history, and risk stratification. *Blood* **126**, 454–462 (2015).
235. Hamblin, T. J., Davis, Z., Gardiner, A., Oscier, D. G. & Stevenson, F. K. Unmutated Ig V(H) genes are associated with a more aggressive form of chronic lymphocytic leukemia. *Blood* **94**, 1848–54 (1999).
236. Damle, R. N. *et al.* Ig V gene mutation status and CD38 expression as novel prognostic indicators in chronic lymphocytic leukemia. *Blood* **94**, 1840–7 (1999).
237. Kikushige, Y. *et al.* Self-Renewing Hematopoietic Stem Cell Is the Primary Target in Pathogenesis of Human Chronic Lymphocytic Leukemia. *Cancer Cell* **20**, 246–259 (2011).
238. Rassenti, L. Z. *et al.* ZAP-70 Compared with Immunoglobulin Heavy-Chain Gene Mutation Status as a Predictor of Disease Progression in Chronic Lymphocytic Leukemia. *N. Engl. J. Med.* **351**, 893–901 (2004).
239. Wiestner, A. ZAP-70 expression identifies a chronic lymphocytic leukemia subtype with unmutated immunoglobulin genes, inferior clinical outcome, and distinct gene expression profile. *Blood* **101**, 4944–4951 (2003).
240. Crespo, M. *et al.* ZAP-70 Expression as a Surrogate for Immunoglobulin-Variable-Region Mutations in Chronic Lymphocytic Leukemia. *N. Engl. J. Med.* **348**, 1764–1775 (2003).
241. Ghia, P. *et al.* The pattern of CD38 expression defines a distinct subset of chronic lymphocytic leukemia (CLL) patients at risk of disease progression. *Blood* **101**, 1262–

1269 (2003).
242. Stamatopoulos, K., Agathangelidis, A., Rosenquist, R. & Ghia, P. Antigen receptor stereotypy in chronic lymphocytic leukemia. *Leukemia* **31**, 282–291 (2017).
243. ten Hacken, E., Gounari, M., Ghia, P. & Burger, J. A. The importance of B cell receptor isotypes and stereotypes in chronic lymphocytic leukemia. *Leukemia* **33**, 287–298 (2019).
244. Fabbri, G. & Dalla-Favera, R. The molecular pathogenesis of chronic lymphocytic leukaemia. *Nat. Rev. Cancer* **16**, 145–62 (2016).
245. Puente, X. S. *et al.* Non-coding recurrent mutations in chronic lymphocytic leukaemia. *Nature* **526**, 519–24 (2015).
246. Landau, D. A. *et al.* Mutations driving CLL and their evolution in progression and relapse. *Nature* **526**, 525–30 (2015).
247. Puente, X. S. *et al.* Whole-genome sequencing identifies recurrent mutations in chronic lymphocytic leukaemia. *Nature* **475**, 101–105 (2011).
248. Landau, D. a. *et al.* Evolution and impact of subclonal mutations in chronic lymphocytic leukemia. *Cell* **152**, 714–726 (2013).
249. Quesada, V. *et al.* Exome sequencing identifies recurrent mutations of the splicing factor SF3B1 gene in chronic lymphocytic leukemia. *Nat. Genet.* **44**, 47–52 (2012).
250. Klein, U. *et al.* The DLEU2/miR-15a/16-1 Cluster Controls B Cell Proliferation and Its Deletion Leads to Chronic Lymphocytic Leukemia. *Cancer Cell* **17**, 28–40 (2010).
251. Nadeu, F. *et al.* Clinical impact of the subclonal architecture and mutational complexity in chronic lymphocytic leukemia. *Leukemia* **32**, 645–653 (2018).
252. Guièze, R. & Wu, C. J. Genomic and epigenomic heterogeneity in chronic lymphocytic leukemia. *Blood* **126**, 445–53 (2015).
253. Cahill, N. & Rosenquist, R. Uncovering the DNA methylome in chronic lymphocytic leukemia. *Epigenetics* **8**, 138–148 (2013).
254. Kulis, M. *et al.* Epigenomic analysis detects widespread gene-body DNA hypomethylation in chronic lymphocytic leukemia. *Nat. Genet.* **44**, 1236–1242 (2012).
255. Oakes, C. C. *et al.* Evolution of DNA Methylation Is Linked to Genetic Aberrations in Chronic Lymphocytic Leukemia. *Cancer Discov.* **4**, 348–361 (2014).
256. Landau, D. A. *et al.* Locally disordered methylation forms the basis of intratumor methylome variation in chronic lymphocytic leukemia. *Cancer Cell* **26**, 813–25 (2014).
257. Bhoi, S. *et al.* Prognostic impact of epigenetic classification in chronic lymphocytic leukemia: The case of subset #2. *Epigenetics* **11**, 449–455 (2016).
258. Queirós, a C. *et al.* A B-cell epigenetic signature defines three biologic subgroups of chronic lymphocytic leukemia with clinical impact. *Leukemia* **29**, 598–605 (2015).
259. Giacopelli, B. *et al.* Developmental subtypes assessed by DNA methylation-iPLEX forecast the natural history of chronic lymphocytic leukemia. *Blood* **134**, 688–698 (2019).
260. Ott, C. J. *et al.* Enhancer Architecture and Essential Core Regulatory Circuitry of Chronic Lymphocytic Leukemia. *Cancer Cell* **34**, 982-995.e7 (2018).
261. Li, L. *et al.* Prognostic significances of overexpression MYC and/or BCL2 in R-CHOP-treated diffuse large B-cell lymphoma: A Systematic review and meta-analysis. *Sci. Rep.* **8**, 1–9 (2018).
262. Chapuy, B. *et al.* Molecular subtypes of diffuse large B cell lymphoma are associated with distinct pathogenic mechanisms and outcomes. *Nat. Med.* **24**, 679–690 (2018).
263. Schmitz, R. *et al.* Genetics and Pathogenesis of Diffuse Large B-Cell Lymphoma. *N. Engl.*

J. Med. **378**, 1396–1407 (2018).
264. Wright, G. W. *et al.* A Probabilistic Classification Tool for Genetic Subtypes of Diffuse Large B Cell Lymphoma with Therapeutic Implications. *Cancer Cell* **37**, 551-568.e14 (2020).
265. Karube, K. *et al.* Integrating genomic alterations in diffuse large B-cell lymphoma identifies new relevant pathways and potential therapeutic targets. *Leukemia* **32**, 675–684 (2018).
266. Asmar, F. *et al.* Genome-wide profiling identifies a DNA methylation signature that associates with TET2 mutations in diffuse large B-cell lymphoma. *Haematologica* **98**, 1912–1920 (2013).
267. Chambwe, N. *et al.* Variability in DNA methylation defines novel epigenetic subgroups of DLBCL associated with different clinical outcomes. *Blood* **123**, 1699–708 (2014).
268. De, S. *et al.* Aberration in DNA methylation in B-cell lymphomas has a complex origin and increases with disease severity. *PLoS Genet.* **9**, e1003137 (2013).
269. Pan, H. *et al.* Epigenomic evolution in diffuse large B-cell lymphomas. *Nat. Commun.* **6**, 6921 (2015).
270. Zamagni, E., Tacchetti, P. & Cavo, M. Imaging in multiple myeloma: How? When? *Blood* **133**, 644–651 (2019).
271. Moreau, P., Attal, M. & Facon, T. Frontline therapy of multiple myeloma. *Blood* **125**, 3076–84 (2015).
272. Pawlyn, C. & Davies, F. E. Toward personalized treatment in multiple myeloma based on molecular characteristics. *Blood* **133**, 660–676 (2019).
273. Morgan, G. J., Walker, B. A. & Davies, F. E. The genetic architecture of multiple myeloma. *Nat. Rev. Cancer* **12**, 335–348 (2012).
274. Palumbo, A. & Anderson, K. Multiple myeloma. *N. Engl. J. Med.* **364**, 1046–60 (2011).
275. Kumar, S. K. *et al.* Multiple myeloma. *Nat. Rev. Dis. Prim.* **3**, 17046 (2017).
276. Kuiper, R. *et al.* A gene expression signature for high-risk multiple myeloma. *Leukemia* **26**, 2406–2413 (2012).
277. Broyl, A. *et al.* Gene expression profiling for molecular classification of multiple myeloma in newly diagnosed patients Gene expression profiling for molecular classification of multiple myeloma in newly diagnosed patients. *October* **116**, 2543–2553 (2011).
278. Walker, B. A. *et al.* Mutational spectrum, copy number changes, and outcome: Results of a sequencing study of patients with newly diagnosed myeloma. *J. Clin. Oncol.* **33**, 3911–3920 (2015).
279. Heuck, C. J. *et al.* Myeloma Is Characterized by Stage-Specific Alterations in DNA Methylation That Occur Early during Myelomagenesis. *J. Immunol.* **190**, 2966–2975 (2013).
280. Agirre, X. *et al.* Whole-epigenome analysis in multiple myeloma reveals DNA hypermethylation of B cell-specific enhancers. *Genome Res.* **25**, 478–87 (2015).
281. Kaiser, M. F. *et al.* Global methylation analysis identifies prognostically important epigenetically inactivated tumor suppressor genes in multiple myeloma. *Blood* **122**, 219–226 (2013).
282. Fulciniti, M. *et al.* Non-overlapping Control of Transcriptome by Promoter- and Super-Enhancer-Associated Dependencies in Multiple Myeloma. *Cell Rep.* **25**, 3693-3705.e6 (2018).
283. Jin, Y. *et al.* Active enhancer and chromatin accessibility landscapes chart the regulatory

network of primary multiple myeloma. *Blood* **131**, 2138–2150 (2018).
284. Ordoñez, R. *et al.* Chromatin activation as a unifying principle underlying pathogenic mechanisms in multiple myeloma. *Genome Res.* **30**, 1217–1227 (2020).
285. Maksimovic, J., Gordon, L. & Oshlack, A. SWAN: Subset-quantile within array normalization for illumina infinium HumanMethylation450 BeadChips. *Genome Biol.* **13**, R44 (2012).
286. Jaffe, A. E. & Irizarry, R. A. Accounting for cellular heterogeneity is critical in epigenome-wide association studies. *Genome Biol.* **15**, R31 (2014).
287. Reinius, L. E. *et al.* Differential DNA methylation in purified human blood cells: implications for cell lineage and studies on disease susceptibility. *PLoS One* **7**, e41361 (2012).
288. Marco-Sola, S., Sammeth, M., Guigó, R. & Ribeca, P. The GEM mapper: fast, accurate and versatile alignment by filtration. *Nat. Methods* **9**, 1185–1188 (2012).
289. Lei, M., Xu, J., Huang, L.-C., Wang, L. & Li, J. Network module-based model in the differential expression analysis for RNA-seq. *Bioinformatics* **33**, 2699–2705 (2017).
290. Raineri, E., Dabad, M. & Heath, S. A note on exact differences between beta distributions in genomic (Methylation) studies. *PLoS One* **9**, e97349 (2014).
291. Zhang, Y. *et al.* Model-based Analysis of ChIP-Seq (MACS). *Genome Biol.* **9**, R137 (2008).
292. Simonis, M., Kooren, J. & de Laat, W. An evaluation of 3C-based methods to capture DNA interactions. *Nat. Methods* **4**, 895–901 (2007).
293. van de Werken, H. J. G. *et al.* Robust 4C-seq data analysis to screen for regulatory DNA interactions. *Nat. Methods* **9**, 969–972 (2012).
294. Rosenwald, A. *et al.* The proliferation gene expression signature is a quantitative integrator of oncogenic events that predicts survival in mantle cell lymphoma. *Cancer Cell* **3**, 185–197 (2003).
295. Queirós, A. C. *et al.* Decoding the DNA Methylome of Mantle Cell Lymphoma in the Light of the Entire B Cell Lineage. *Cancer Cell* **30**, 806–821 (2016).
296. Vento-Tormo, R. *et al.* IL-4 orchestrates STAT6-mediated DNA demethylation leading to dendritic cell differentiation. *Genome Biol.* **17**, 4 (2016).
297. Brönneke, S. *et al.* DNA methylation regulates lineage-specifying genes in primary lymphatic and blood endothelial cells. *Angiogenesis* **15**, 317–329 (2012).
298. Caron, G. *et al.* Cell-Cycle-Dependent Reconfiguration of the DNA Methylome during Terminal Differentiation of Human B Cells into Plasma Cells. *Cell Rep.* **13**, 1059–71 (2015).
299. Bergmann, A. K. *et al.* DNA methylation profiling of pediatric B-cell lymphoblastic leukemia with KMT2A rearrangement identifies hypomethylation at enhancer sites. *Pediatr. Blood Cancer* **64**, 1–5 (2017).
300. Gabriel, A. S. *et al.* Epigenetic landscape correlates with genetic subtype but does not predict outcome in childhood acute lymphoblastic leukemia. *Epigenetics* **10**, 717–726 (2015).
301. Houseman, E. A. *et al.* DNA methylation arrays as surrogate measures of cell mixture distribution. *BMC Bioinformatics* **13**, 86 (2012).
302. Accomando, W. P., Wiencke, J. K., Houseman, E. A., Nelson, H. H. & Kelsey, K. T. Quantitative reconstruction of leukocyte subsets using DNA methylation. *Genome Biol.* **15**, R50 (2014).
303. Houseman, E. A., Kelsey, K. T., Wiencke, J. K. & Marsit, C. J. Cell-composition effects in the analysis of DNA methylation array data: a mathematical perspective. *BMC*

Bioinformatics **16**, 95 (2015).
304. Duran-Ferrer, M., Beekman, R. & Martín-Subero, J. I. In silico deconvolution and purification of cancer epigenomes. *Oncoscience* **4**, 25–26 (2017).
305. Saba, N. S. *et al.* Pathogenic role of B-cell receptor signaling and canonical NF-κB activation in mantle cell lymphoma. *Blood* **128**, 82–92 (2016).
306. Scott, D. W. & Gascoyne, R. D. The tumour microenvironment in B cell lymphomas. *Nat. Rev. Cancer* **14**, 517–534 (2014).
307. Teschendorff, A. E. & Zheng, S. C. Cell-type deconvolution in epigenome-wide association studies: a review and recommendations. *Epigenomics* **9**, epi-2016-0153 (2017).
308. Teschendorff, A. E. & Relton, C. L. Statistical and integrative system-level analysis of DNA methylation data. *Nat. Rev. Genet.* **19**, 129–147 (2018).
309. Debaize, L. *et al.* Interplay between transcription regulators RUNX1 and FUBP1 activates an enhancer of the oncogene c-KIT and amplifies cell proliferation. *Nucleic Acids Res.* **46**, 11214–11228 (2018).
310. IGV (Integrative Genomic Viewer). 1b - Dataset normalization. *Broad Inst.* **29**, 24–26 (2013).
311. Le Thi, H. A., Nguyen, V. V. & Ouchani, S. *Advanced Data Mining and Applications*. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* **5139**, (Springer Berlin Heidelberg, 2008).
312. Dietrich, S. *et al.* Drug-perturbation-based stratification of blood cancer. *J. Clin. Invest.* **128**, 427–445 (2017).
313. Yang, Z. *et al.* Correlation of an epigenetic mitotic clock with cancer risk. *Genome Biol.* **17**, 205 (2016).
314. Youn, A. & Wang, S. The MiAge Calculator: a DNA methylation-based mitotic age calculator of human tissue types. *Epigenetics* **13**, 192–206 (2018).
315. Sánchez-Vega, F., Gotea, V., Margolin, G. & Elnitski, L. Pan-cancer stratification of solid human epithelial tumors and cancer cell lines reveals commonalities and tissue-specific features of the CpG island methylator phenotype. *Epigenetics and Chromatin* **8**, 1–24 (2015).
316. Zhou, W. *et al.* DNA methylation loss in late-replicating domains is linked to mitotic cell division. *Nat. Genet.* **50**, 591–602 (2018).
317. Horvath, S. DNA methylation age of human tissues and cell types. *Genome Biol.* **14**, R115 (2013).
318. Shuai, S. *et al.* The U1 spliceosomal RNA is recurrently mutated in multiple cancers. *Nature* **574**, 712–716 (2019).
319. Alexandrov, L. B. *et al.* The repertoire of mutational signatures in human cancer. *Nature* **578**, 94–101 (2020).
320. Sergushichev, A. An algorithm for fast preranked gene set enrichment analysis using cumulative statistic calculation. *bioRxiv* (2016).
321. McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–303 (2010).
322. DePristo, M. A. *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* **43**, 491–8 (2011).
323. Nik-Zainal, S. *et al.* The life history of 21 breast cancers. *Cell* **149**, 994–1007 (2012).
324. Moncunill, V. *et al.* Comprehensive characterization of complex structural variations in

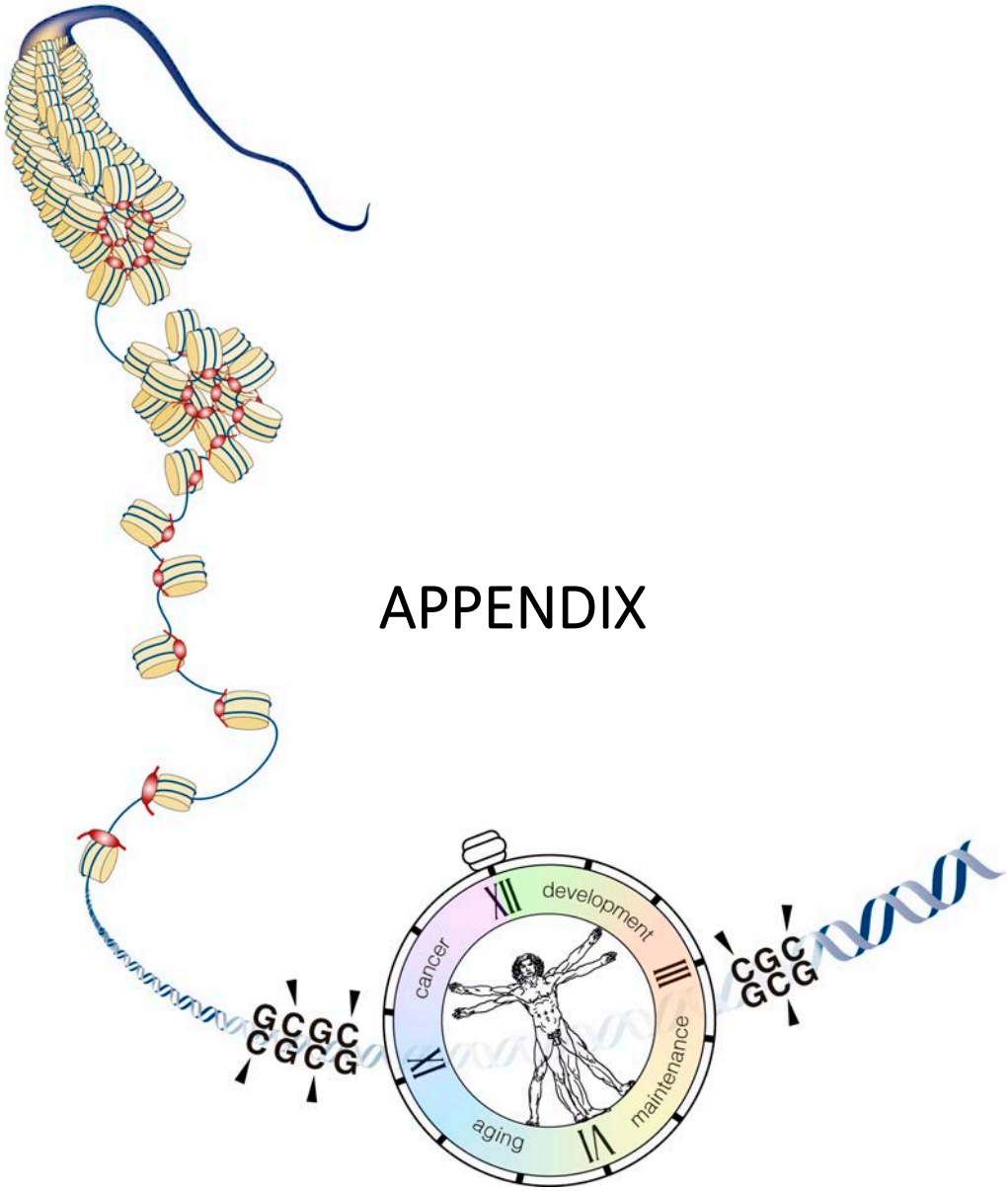
cancer by directly comparing genome sequence reads. *Nat. Biotechnol.* **32**, 1106–12 (2014).
325. Layer, R. M., Chiang, C., Quinlan, A. R. & Hall, I. M. LUMPY: a probabilistic framework for structural variant discovery. *Genome Biol.* **15**, R84 (2014).
326. Mermel, C. H. *et al.* GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. *Genome Biol.* **12**, R41 (2011).
327. Bock, C. Analysing and interpreting DNA methylation data. *Nat. Rev. Genet.* **13**, 705–19 (2012).
328. Shaknovich, R. *et al.* DNA methylation signatures define molecular subtypes of diffuse large B-cell lymphoma. *Blood* **116**, e81-9 (2010).
329. Kretzmer, H. *et al.* DNA methylome analysis in Burkitt and follicular lymphomas identifies differentially methylated regions linked to somatic mutation and transcriptional control. *Nat. Genet.* **47**, 1316–1325 (2015).
330. Hattab, M. W. *et al.* Correcting for cell-type effects in DNA methylation studies: reference-based method outperforms latent variable approaches in empirical studies. *Genome Biol.* **18**, 24 (2017).
331. Somasundaram, R., Prasad, M. A. J., Ungerbäck, J. & Sigvardsson, M. Transcription factor networks in B-cell differentiation link development to acute lymphoid leukemia. *Blood* **126**, 144–152 (2015).
332. Sánchez-Tilló, E. *et al.* The EMT activator ZEB1 promotes tumor growth and determines differential response to chemotherapy in mantle cell lymphoma. *Cell Death Differ.* **21**, 247–257 (2014).
333. Wolf, C. *et al.* NFATC1 activation by DNA hypomethylation in chronic lymphocytic leukemia correlates with clinical staging and can be inhibited by ibrutinib. *Int. J. Cancer* **142**, 322–333 (2018).
334. Blonska, M. *et al.* Jun-regulated genes promote interaction of diffuse large B-cell lymphoma with the microenvironment. *Blood* **125**, 981–991 (2015).
335. Huerta-Yepez, S. *et al.* Overexpression of Yin Yang 1 in bone marrow-derived human multiple myeloma and its clinical significance. *Int. J. Oncol.* **45**, 1184–1192 (2014).
336. Sprynski, A. C. *et al.* Insulin is a potent myeloma cell growth factor through insulin/IGF-1 hybrid receptor activation. *Leukemia* **24**, 1940–1950 (2010).
337. Riz, I. & Hawley, R. G. Increased expression of the tight junction protein TJP1/ZO-1 is associated with upregulation of TAZ-TEAD activity and an adult tissue stem cell signature in carfilzomib-resistant multiple myeloma cells and high-risk multiple myeloma patients. *Oncoscience* **4**, 79–94 (2017).
338. Herath, N. I., Rocques, N., Garancher, A., Eychène, A. & Pouponnot, C. GSK3-mediated MAF phosphorylation in multiple myeloma as a potential therapeutic target. *Blood Cancer J.* **4**, e175–e175 (2014).
339. Navarro, A. *et al.* Improved classification of leukemic B-cell lymphoproliferative disorders using a transcriptional and genetic classifier. *Haematologica* **102**, 360–363 (2017).
340. Hannum, G. *et al.* Genome-wide methylation profiles reveal quantitative views of human aging rates. *Mol. Cell* **49**, 359–367 (2013).
341. Horvath, S. & Raj, K. DNA methylation-based biomarkers and the epigenetic clock theory of ageing. *Nat. Rev. Genet.* **19**, 371–384 (2018).
342. Field, A. E. *et al.* DNA Methylation Clocks in Aging: Categories, Causes, and

Consequences. *Mol. Cell* **71**, 882–895 (2018).
343. Beerman, I. *et al.* Proliferation-dependent alterations of the DNA methylation landscape underlie hematopoietic stem cell aging. *Cell Stem Cell* **12**, 413–25 (2013).
344. Landan, G. *et al.* Epigenetic polymorphism and the stochastic formation of differentially methylated regions in normal and cancerous tissues. *Nat. Genet.* **44**, 1207–14 (2012).
345. Siegmund, K. D., Marjoram, P., Woo, Y.-J., Tavaré, S. & Shibata, D. Inferring clonal expansion and cancer stem cell dynamics from DNA methylation patterns in colorectal cancers. *Proc. Natl. Acad. Sci. U. S. A.* **106**, 4828–4833 (2009).
346. Berman, B. P. *et al.* Regions of focal DNA hypermethylation and long-range hypomethylation in colorectal cancer coincide with nuclear lamina-associated domains. *Nat. Genet.* **44**, 40–6 (2011).
347. Bollati, V. *et al.* Differential repetitive DNA methylation in multiple myeloma molecular subgroups. *Carcinogenesis* **30**, 1330–5 (2009).
348. Vandiver, A. R., Idrizi, A., Rizzardi, L., Feinberg, A. P. & Hansen, K. D. DNA methylation is stable during replication and cell cycle arrest. *Sci. Rep.* **5**, 1–8 (2015).
349. Maura, F. *et al.* A practical guide for mutational signature analysis in hematological malignancies. *Nat. Commun.* **10**, 2969 (2019).
350. Alexandrov, L. B. *et al.* Clock-like mutational processes in human somatic cells. *Nat. Genet.* **47**, 1402–7 (2015).
351. Rakyan, V. K. *et al.* Human aging-associated DNA hypermethylation occurs preferentially at bivalent chromatin domains. *Genome Res.* **20**, 434–439 (2010).
352. Teschendorff, A. E. *et al.* Age-dependent DNA methylation of genes that are suppressed in stem cells is a hallmark of cancer. *Genome Res.* **20**, 440–6 (2010).
353. Heyn, H. *et al.* Distinct DNA methylomes of newborns and centenarians. *Proc. Natl. Acad. Sci. U. S. A.* **109**, 10522–7 (2012).
354. Bell, C. G. *et al.* DNA methylation aging clocks: challenges and recommendations. *Genome Biol.* **20**, 249 (2019).
355. Cahill, N. *et al.* 450K-array analysis of chronic lymphocytic leukemia cells reveals global DNA methylation to be relatively stable over time and similar in resting and proliferative compartments. *Leukemia* **27**, 150–158 (2013).
356. Baliakas, P. *et al.* Cytogenetic complexity in chronic lymphocytic leukemia: Definitions, associations, and clinical impact. *Blood* **133**, 1205–1216 (2019).
357. Kwok, M. *et al.* Integrative analysis of spontaneous CLL regression highlights genetic and microenvironmental interdependency in CLL. *Blood* **135**, 411–428 (2020).
358. Tanaka, H. & Watanabe, T. Mechanisms Underlying Recurrent Genomic Amplification in Human Cancers. *Trends in cancer* **6**, 462–477 (2020).
359. Pararajalingam, P. *et al.* Coding and non-coding drivers of mantle cell lymphoma identified through exome and genome sequencing Title : Coding and non-coding drivers of mantle cell lymphoma identified through exome and genome sequencing Affiliations : 1 Department of Molecular Biol. *Blood* (2020).
360. Clot, G. *et al.* A gene signature that distinguishes conventional and leukemic nonnodal mantle cell lymphoma helps predict outcome. *Blood* **132**, 413–422 (2018).
361. Aukema, S. M. *et al.* Expression of TP53 is associated with the outcome of MCL independent of MIPI and Ki-67 in trials of the European MCL Network. *Blood* **131**, 417–420 (2018).
362. Delfau-Larue, M.-H. *et al.* High-dose cytarabine does not overcome the adverse prognostic value of CDKN2A and TP53 deletions in mantle cell lymphoma. *Blood* **126**,

604–11 (2015).
363. Eskelund, C. W. *et al.* TP53 mutations identify younger mantle cell lymphoma patients who do not benefit from intensive chemoimmunotherapy. *Blood* **130**, 1903–1910 (2017).
364. Ferrero, S. *et al.* KMT2D mutations and TP53 disruptions are poor prognostic biomarkers in mantle cell lymphoma receiving high-dose therapy: a FIL study. *Haematologica* **105**, 1604–1612 (2020).
365. Hoadley, K. A. *et al.* Cell-of-Origin Patterns Dominate the Molecular Classification of 10,000 Tumors from 33 Types of Cancer. *Cell* **173**, 291-304.e6 (2018).
366. Rodríguez-Paredes, M. *et al.* Methylation profiling identifies two subclasses of squamous cell carcinoma related to distinct cells of origin. *Nat. Commun.* **9**, (2018).
367. Bormann, F. *et al.* Cell-of-Origin DNA Methylation Signatures Are Maintained during Colorectal Carcinogenesis. *Cell Rep.* **23**, 3407–3418 (2018).
368. Wojdacz, T. K. *et al.* Clinical significance of DNA methylation in chronic lymphocytic leukemia patients: Results from 3 UK clinical trials. *Blood Adv.* **3**, 2474–2481 (2019).
369. Gaiti, F. *et al.* Epigenetic evolution and lineage histories of chronic lymphocytic leukaemia. *Nature* **569**, 576–580 (2019).
370. Meir, Z., Mukamel, Z., Chomsky, E., Lifshitz, A. & Tanay, A. Single-cell analysis of clonal maintenance of transcriptional and epigenetic states in cancer cells. *Nat. Genet.* **52**, 709–718 (2020).
371. Sandoval, J. *et al.* Genome-wide DNA methylation profiling predicts relapse in childhood B-cell acute lymphoblastic leukaemia. *Br. J. Haematol.* **160**, 406–9 (2013).
372. Rhein, P. *et al.* Gene expression shift towards normal B cells, decreased proliferative capacity and distinct surface receptors characterize leukemic blasts persisting during induction therapy in childhood acute lymphoblastic leukemia. *Leukemia* **21**, 897–905 (2007).
373. Shen, S. Y. *et al.* Sensitive tumour detection and classification using plasma cell-free DNA methylomes. *Nature* **563**, 579–583 (2018).
374. Nuzzo, P. V. *et al.* Detection of renal cell carcinoma using plasma and urine cell-free DNA methylomes. *Nat. Med.* **38**, 728–728 (2020).
375. Nassiri, F. *et al.* Detection and discrimination of intracranial tumors using plasma cell-free DNA methylomes. *Nat. Med.* **26**, 1044–1047 (2020).
376. Giesselmann, P. *et al.* Analysis of short tandem repeat expansions and their methylation state with nanopore sequencing. *Nat. Biotechnol.* **37**, 1478–1481 (2019).
377. Nadeu, F. *et al.* Clinical impact of clonal and subclonal TP53, SF3B1, BIRC3, NOTCH1, and ATM mutations in chronic lymphocytic leukemia. *Blood* **127**, 2122–30 (2016).
378. Brieghel, C. *et al.* The Number of Signaling Pathways Altered by Driver Mutations in Chronic Lymphocytic Leukemia Impacts Disease Outcome. *Clin. Cancer Res.* **26**, clincanres.4158.2018 (2020).
379. Maity, P. C. *et al.* IGLV3-21∗01 is an inherited risk factor for CLL through the acquisition of a single-point mutation enabling autonomous BCR signaling. *Proc. Natl. Acad. Sci. U. S. A.* **117**, 4320–4327 (2020).
380. Stamatopoulos, B. *et al.* The light chain IgLV3-21 defines a new poor prognostic subgroup in chronic lymphocytic leukemia: Results of a multicenter study. *Clin. Cancer Res.* **24**, 5048–5057 (2018).
381. Tsagiopoulou, M. *et al.* DNA methylation profiles in chronic lymphocytic leukemia patients treated with chemoimmunotherapy. *Clin. Epigenetics* **11**, 177 (2019).

382. Blanco, E. *et al.* Age-associated distribution of normal B-cell and plasma cell subsets in peripheral blood. *J. Allergy Clin. Immunol.* **141**, 2208-2219.e16 (2018).
383. Issa, J. Aging and epigenetic drift: a vicious cycle. *J. Clin. Invest.* **124**, 24–9 (2014).
384. López-Otín, C., Blasco, M. a., Partridge, L., Serrano, M. & Kroemer, G. The Hallmarks of Aging. *Cell* **153**, 1194–1217 (2013).
385. Campisi, J. *et al.* From discoveries in ageing research to therapeutics for healthy ageing. *Nature* **571**, 183–192 (2019).
386. Signer, R. A. J. & Morrison, S. J. Mechanisms that regulate stem cell aging and life span. *Cell Stem Cell* **12**, 152–165 (2013).
387. Zhang, W., Qu, J., Liu, G. H. & Belmonte, J. C. I. The ageing epigenome and its rejuvenation. *Nat. Rev. Mol. Cell Biol.* **21**, 137–150 (2020).
388. Tomasetti, C., Li, L. & Vogelstein, B. Stem cell divisions, somatic mutations, cancer etiology, and cancer prevention. *Science* **355**, 1330–1334 (2017).
389. Tomasetti, C. & Vogelstein, B. Variation in cancer risk among tissues can be explained by the number of stem cell divisions. *Science (80-.).* **347**, 78–81 (2015).
390. Wu, S., Powers, S., Zhu, W. & Hannun, Y. A. Substantial contribution of extrinsic risk factors to cancer development. *Nature* **529**, 43–7 (2016).
391. Klutstein, M., Moss, J., Kaplan, T. & Cedar, H. Contribution of epigenetic mechanisms to variation in cancer risk among tissues. *Proc. Natl. Acad. Sci. U. S. A.* **114**, 2230–2234 (2017).
392. Dugué, P. A. *et al.* DNA methylation-based biological aging and cancer risk and survival: Pooled analysis of seven prospective studies. *Int. J. Cancer* **142**, 1611–1619 (2018).
393. Marioni, R. E. *et al.* DNA methylation age of blood predicts all-cause mortality in later life. *Genome Biol.* **16**, 25 (2015).
394. Zhang, Y. *et al.* DNA methylation signatures in peripheral blood strongly predict all-cause mortality. *Nat. Commun.* **8**, 1–11 (2017).
395. Weidner, C. I. *et al.* Aging of blood can be tracked by DNA methylation changes at just three CpG sites. *Genome Biol.* **15**, (2014).
396. Jaiswal, S. *et al.* Clonal Hematopoiesis and risk of atherosclerotic cardiovascular disease. *N. Engl. J. Med.* **377**, 111–121 (2017).
397. Zink, F. *et al.* Clonal hematopoiesis, with and without candidate driver mutations, is common in the elderly. *Blood* **130**, 742–752 (2017).
398. Shibata, D. Inferring human stem cell behaviour from epigenetic drift. *J. Pathol.* **217**, 199–205 (2009).
399. Flach, J. *et al.* Replication stress is a potent driver of functional decline in ageing haematopoietic stem cells. *Nature* **512**, 198–202 (2014).
400. Walter, D. *et al.* Exit from dormancy provokes DNA-damage-induced attrition in haematopoietic stem cells. *Nature* **520**, 549–552 (2015).
401. Ha, K. *et al.* Rapid and transient recruitment of DNMT1 to DNA double-strand breaks is mediated by its interaction with multiple components of the DNA damage response machinery. *Hum. Mol. Genet.* **20**, 126–140 (2011).
402. Xia, L. *et al.* CHD4 Has Oncogenic Functions in Initiating and Maintaining Epigenetic Suppression of Multiple Tumor Suppressor Genes. *Cancer Cell* **31**, 653-668.e7 (2017).
403. O'Hagan, H. M. *et al.* Oxidative Damage Targets Complexes Containing DNA Methyltransferases, SIRT1, and Polycomb Members to Promoter CpG Islands. *Cancer Cell* **20**, 606–619 (2011).
404. Ding, N. *et al.* Mismatch repair proteins recruit DNA methyltransferase 1 to sites of

oxidative DNA damage. *J. Mol. Cell Biol.* **8**, 244–254 (2016).
405. O'Hagan, H. M., Mohammad, H. P. & Baylin, S. B. Double strand breaks can initiate gene silencing and SIRT1-dependent onset of DNA methylation in an exogenous promoter CpG island. *PLoS Genet.* **4**, (2008).
406. Zhang, Y. W. *et al.* Acetylation Enhances TET2 Function in Protecting against Abnormal DNA Methylation during Oxidative Stress. *Mol. Cell* **65**, 323–335 (2017).
407. Kong, L. *et al.* A primary role of TET proteins in establishment and maintenance of De Novo bivalency at CpG islands. *Nucleic Acids Res.* **44**, 8682–8692 (2016).
408. Wu, D. *et al.* Glucose-regulated phosphorylation of TET2 by AMPK reveals a pathway linking diabetes to cancer. *Nature* **559**, 637–641 (2018).
409. Zheng, S. C., Widschwendter, M. & Teschendorff, A. E. Epigenetic drift, epigenetic clocks and cancer risk. *Epigenomics* **8**, 705–19 (2016).
410. Maegawa, S. *et al.* Caloric restriction delays age-related methylation drift. *Nat. Commun.* **8**, 539 (2017).
411. Cole, J. J. *et al.* Diverse interventions that extend mouse lifespan suppress shared age-associated epigenetic changes at critical gene regulatory regions. *Genome Biol.* **18**, 1–16 (2017).
412. Hahn, O. *et al.* Dietary restriction protects from age-associated DNA methylation and induces epigenetic reprogramming of lipid metabolism. *Genome Biol.* **18**, 56 (2017).
413. Wang, T. *et al.* Epigenetic aging signatures in mice livers are slowed by dwarfism, calorie restriction and rapamycin treatment. *Genome Biol.* **18**, 1–11 (2017).
414. Stubbs, T. M. *et al.* Multi-tissue DNA methylation age predictor in mouse. *Genome Biol.* **18**, 1–14 (2017).
415. Petkovich, D. A. *et al.* Using DNA Methylation Profiling to Evaluate Biological Age and Longevity Interventions. *Cell Metab.* **25**, 954-960.e6 (2017).
416. Redman, L. M. *et al.* Metabolic Slowing and Reduced Oxidative Damage with Sustained Caloric Restriction Support the Rate of Living and Oxidative Damage Theories of Aging. *Cell Metab.* **27**, 805-815.e4 (2018).
417. Nam, A. S., Chaligne, R. & Landau, D. A. Integrating genetic and non-genetic determinants of cancer evolution by single-cell multi-omics. *Nat. Rev. Genet.* (2020). doi:10.1038/s41576-020-0265-5



APPENDIX

Manuscript 1

***** These authors contributed equally to this work
# Corresponding author

Kulis M, Merkel A, Heath S, Queirós AC, Schuyler RP, Castellano G, Beekman R, Raineri E, Esteve A, Clot G, Verdaguer-Dot N, **Duran-Ferrer M**, Russiñol N, Vilarrasa-Blasi R, Ecker S, Pancaldi V, Rico D, Agueda L, Blanc J, Richardson D, Clarke L, Datta A, Pascual M, Agirre X, Prosper F, Alignani D, Paiva B, Caron G, Fest T, Muench MO, Fomin ME, Lee ST, Wiemels JL, Valencia A, Gut M, Flicek P, Stunnenberg HG, Siebert R, Küppers R, Gut IG, Campo E, **#**Martín-Subero JI. Whole-genome fingerprint of the DNA methylome during human B cell differentiation. Nat Genet. 2015 Jul;47(7):746-56. doi: 10.1038/ng.3291. Epub 2015 Jun 8. PMID: 26053498; PMCID: PMC5444519.

ARTICLES

nature genetics

Whole-genome fingerprint of the DNA methylome during human B cell differentiation

Marta Kulis[1], Angelika Merkel[2], Simon Heath[2], Ana C Queirós[1], Ronald P Schuyler[2], Giancarlo Castellano[1], Renée Beekman[1], Emanuele Raineri[2], Anna Esteve[2], Guillem Clot[1], Núria Verdaguer-Dot[1], Martí Duran-Ferrer[1,2], Nuria Russiñol[1], Roser Vilarrasa-Blasi[1], Simone Ecker[3], Vera Pancaldi[3], Daniel Rico[3], Lidia Agueda[2], Julie Blanc[2], David Richardson[4], Laura Clarke[4], Avik Datta[4], Marien Pascual[5], Xabier Agirre[5], Felipe Prosper[5,6], Diego Alignani[7], Bruno Paiva[6,7], Gersende Caron[8], Thierry Fest[8], Marcus O Muench[9,10], Marina E Fomin[9,10], Seung-Tae Lee[11], Joseph L Wiemels[12], Alfonso Valencia[3], Marta Gut[2], Paul Flicek[4], Hendrik G Stunnenberg[13], Reiner Siebert[14], Ralf Küppers[15], Ivo G Gut[2], Elías Campo[1] & José I Martín-Subero[1]

We analyzed the DNA methylome of ten subpopulations spanning the entire B cell differentiation program by whole-genome bisulfite sequencing and high-density microarrays. We observed that non-CpG methylation disappeared upon B cell commitment, whereas CpG methylation changed extensively during B cell maturation, showing an accumulative pattern and affecting around 30% of all measured CpG sites. Early differentiation stages mainly displayed enhancer demethylation, which was associated with upregulation of key B cell transcription factors and affected multiple genes involved in B cell biology. Late differentiation stages, in contrast, showed extensive demethylation of heterochromatin and methylation gain at Polycomb-repressed areas, and genes with apparent functional impact in B cells were not affected. This signature, which has previously been linked to aging and cancer, was particularly widespread in mature cells with an extended lifespan. Comparing B cell neoplasms with their normal counterparts, we determined that they frequently acquire methylation changes in regions already undergoing dynamic methylation during normal B cell differentiation.

The multitude of cell types and tissues of an organism can be defined by their particular epigenetic makeup[1,2]. DNA methylation is an important component of the epigenome, which is extensively modulated during regulatory and developmental processes, both in the context of physiological and pathological conditions[3–5]. Although recent reports have analyzed the DNA methylation profiles of various cell types on a whole-genome scale[1,6–16], the DNA methylome of a single human cell type throughout its complete differentiation process has not been defined thus far. The B cell lineage represents a paradigmatic cellular model to study the dynamic epigenome during cell development and specification because major B cell maturation stages have distinct phenotypic and gene expression features and can be isolated in sufficient numbers from hematopoietic tissues[17–19].

B cell lymphopoiesis is a complex and tightly coordinated process guided by the hierarchical expression of different stage-specific transcription factors and microenvironmental influences[20,21]. The process starts in the bone marrow, where hematopoietic stem cells differentiate into multipotent progenitors and common lymphoid progenitors, which then commit to the B cell lineage and give rise to precursor B cells. These precursors gradually rearrange their immunoglobulin genes and differentiate into mature naive B cells, which leave the bone marrow to enter the bloodstream. Resting naive B cells transit through lymph nodes, and they are eventually activated by specific antigens via activation of the B cell receptor, which induces the germinal center reaction. Germinal center B cells further rearrange and mutate their immunoglobulin genes, rapidly proliferate and differentiate. Finally, the germinal center reaction gives rise to plasma cells producing large amounts of high-affinity antibodies and memory B cells. Plasma cells exiting the lymph nodes migrate to the bone marrow, where they can reside for extended periods of time, and long-lived

[1]Institut d'Investigacions Biomédiques August Pi i Sunyer (IDIBAPS), Department of Anatomic Pathology, Pharmacology and Microbiology, University of Barcelona, Barcelona, Spain. [2]Centro Nacional de Análisis Genómico (CNAG), Parc Científic de Barcelona, Barcelona, Spain. [3]Structural Biology and Biocomputing Program, Spanish National Cancer Research Centre (CNIO), Spanish National Bioinformatics Institute, Madrid, Spain. [4]European Molecular Biology Laboratory–European Bioinformatics Institute (EMBL-EBI), Wellcome Trust Genome Campus, Hinxton, UK. [5]Area de Oncología, Centro de Investigación Médica Aplicada (CIMA), Universidad de Navarra, Pamplona, Spain. [6]Servicio de Hematología, Clínica Universidad de Navarra, Pamplona, Spain. [7]Flow Cytometry Core, Centro de Investigación Médica Aplicada (CIMA), Universidad de Navarra, Pamplona, Spain. [8]Université de Rennes 1, INSERM U917, Hematology Laboratory, University Hospital of Rennes, Rennes, France. [9]Blood Systems Research Institute, San Francisco, California, USA. [10]Department of Laboratory Medicine, University of California, San Francisco, San Francisco, California, USA. [11]Department of Laboratory Medicine, Yonsei University College of Medicine, Seoul, Republic of Korea. [12]Department of Epidemiology and Biostatistics, University of California, San Francisco, San Francisco, California, USA. [13]Molecular Biology, Nijmegen Centre for Molecular Life Sciences (NCMLS), Faculties of Science and Medicine, Radboud University, Nijmegen, the Netherlands. [14]Institute of Human Genetics, Christian Albrechts University, Kiel, Germany. [15]Institute of Cell Biology (Cancer Research), Medical School, University of Duisburg-Essen, Essen, Germany. Correspondence should be addressed to J.I.M.-S. (imartins@clinic.ub.es).

Received 20 October 2014; accepted 3 April 2015; published online 8 June 2015; doi:10.1038/ng.3291

memory B cells recirculate through the blood and lymphoid organs, providing the basis for enduring humoral immunity[22,23]. Hence, an interesting feature of the B cell maturation process is that it entails a variety of cell types with different functional features, proliferation abilities, microenvironmental influences and lifespans, providing an exceptional opportunity to study the epigenome in the context of different biological processes and to provide insights into the fields of cell differentiation, B cell biology, cancer and aging.

RESULTS

Whole-genome DNA methylation maps of B cell subpopulations

We generated unbiased DNA methylation maps of sorted human cell populations—uncommitted hematopoietic progenitor cells (HPCs) and five B cell–lineage subpopulations, including pre-BII cells, naive B cells from peripheral blood, germinal center B cells, memory B cells from peripheral blood and plasma cells from bone marrow—by whole-genome bisulfite sequencing (WGBS) (**Fig. 1a** and **Supplementary Table 1**). We sequenced two biological replicates of each subpopulation for a total of 2,217 billion base pairs, of which 85–95% could be mapped (mean depth of 54-fold per sample) (**Supplementary Table 2**). On average, we measured the methylation levels of 22.7 million CpG sites per sample (range of 21 to 25 million). In unsupervised principal-component analysis (PCA) of CpG methylation levels, B cell subpopulations were segregated according to their developmental stage (**Fig. 1b**). Globally, B cell differentiation was accompanied by gradual, widespread demethylation of the genome, with this demethylation more pronounced at late differentiation stages, such as in memory B cells and bone marrow plasma cells (**Fig. 1c–e**). The global methylation status of CpGs was largely bimodal in all sorted cell populations, and the level of partially methylated regions increased to 19–24% in advanced maturation stages (**Fig. 1e**). This result is in contrast to other WGBS studies using whole tissues, in which the proportion of partially methylated regions is usually high[24], and highlights the importance of using purified cell subpopulations for DNA methylation studies.

The results obtained by WGBS were complemented with the analysis of three to nine replicates of ten different B cell subpopulations by high-density DNA methylation microarray[25] (**Fig. 1a**). These subpopulations included those analyzed by WGBS as well as pre-BI cells, immature B cells, tonsillar naive B cells and tonsillar plasma cells (**Supplementary Table 1**). The biological replicates of each subpopulation analyzed by WGBS or microarray showed high reproducibility (correlation coefficient > 0.95; **Supplementary Fig. 1**), and we further validated both high-throughput techniques by bisulfite pyrosequencing (BPS; **Supplementary Fig. 2**). In line with the WGBS data, PCA of the microarray data separated B cell subpopulations mostly according to their developmental stage (**Fig. 1f**), and we observed gradual global methylation loss throughout B cell maturation (**Fig. 1g**). Interestingly, samples clustered into 'antigen-inexperienced' and 'antigen-experienced' cells (**Supplementary Fig. 3**). This finding indicates that proliferative germinal center B cells start a massive reconfiguration of the DNA methylome[18,19], which continues in cell subpopulations with extended lifespan, such as memory B cells and bone marrow plasma cells.

Demethylation of non-CpG sites upon B cell commitment

Cytosine methylation in mammals can occur outside CpGs, a phenomenon commonly observed in embryonic stem cells (ESCs) and neurons[7,8,16]. In our study, when considering non-CpG sites with high-confidence methylation estimates and having ruled out suboptimal bisulfite conversion and sequence variants, we did observe non-CpG methylation, which was primarily confined to the most undifferentiated cells (HPCs) and occurred mainly in a CpApC sequence context (**Fig. 2a–c**). We detected substantial levels of non-CpG methylation in HPCs by WGBS at 25,763 sites in replicate 1 and at 16,838 sites in replicate 2, with mean methylation levels of 25.1% and 24.7%, respectively. Non-CpG methylation in HPCs frequently targeted the same sites in the two biological replicates, a finding that was also confirmed by BPS in independent samples (**Fig. 2d** and **Supplementary Fig. 4**). Non-CpG methylation in HPCs was preferentially located in gene bodies (both introns and exons) and was depleted in lamina-associated domains (**Supplementary Fig. 5**). Although the methylation microarray only measured methylation at 3,091 non-CpG sites, we could confirm the WGBS results (**Fig. 2e,f**). Similarly to reports in ESCs[26], methylated non-CpGs were flanked by methylated CpGs. However, demethylation of non-CpGs and CpGs was not simultaneous. We observed dramatic demethylation of non-CpGs upon B cell commitment in pre-BII cells, but 97% of these demethylated non-CpGs remained flanked by methylated CpGs (**Supplementary Fig. 6**). These data were confirmed by BPS (**Supplementary Fig. 4**) and indicate that non-CpG methylation is passively erased in the transition from HPCs to pre-BII cells, without simultaneous demethylation of flanking CpGs (**Fig. 2g**). Although previous studies have reported that high expression of *DNMT3A* and *DNMT3B* is associated with non-CpG methylation[26,27], we did not identify consistent differences in the expression levels of *de novo* methyltransferases (DNMTs) in HPCs and pre-B cells (**Supplementary Fig. 7**).

Identification of dynamic DNA methylation patterns

We next focused our analysis on the modulation of CpG methylation throughout the complete B cell maturation program. We identified dynamic methylation levels at 4.93 million CpGs (>0.25 methylation change in one set of samples and >0.1 methylation change in the other), which represent 30.6% of the 16.1 million CpGs with methylation estimates in all 12 samples analyzed. The cell subtypes showing the most pronounced methylation changes in comparison to the preceding stage were germinal center B cells, memory B cells and bone marrow plasma cells (**Fig. 3a** and **Supplementary Fig. 8**). Interestingly, multiple genes directly involved in B cell differentiation (for example, *ARID3A*, *BCL2*, *BLK*, *EBF1* and *IRF4*) showed complex modulation of their DNA methylation profiles across the gene length, with different regulatory elements losing methylation at distinct maturation stages (**Fig. 3b** and **Supplementary Fig. 9**). In a previous study, 5.6 million dynamic CpGs were found to be differentially methylated in a wide range of human cell types and tissues[1]. Only one-third of the dynamic CpGs in B cell differentiation defined in our study overlapped with these CpGs (**Supplementary Fig. 10**). Although this result may in part have been caused by differences in coverage and the bioinformatics pipelines used in the two studies, it suggests that the majority of dynamic methylation in B cells may entail CpGs not previously detected to be differentially methylated in other cell types and tissues.

Similarly to WGBS data, although on a smaller scale, microarray data showed that 22.4% of the measured CpGs had variable methylation levels during B cell differentiation ($n = 106,562$; Online Methods). Because the set of samples was more comprehensive for these analyses, we used these data to define sets of CpGs showing similar DNA methylation dynamics during B cell differentiation. To this end, we based our next analysis on a linear model of B cell maturation, from HPCs to the terminal differentiation of germinal center B cells into tonsillar plasma cells and long-lived bone marrow plasma cells. We defined 20 major modules containing at least 500 CpGs each that could be

classified according to 4 general DNA methylation modulation patterns during differentiation: (i) methylation levels decreased (9 modules), (ii) methylation levels first decreased and then increased (3 modules), (iii) methylation levels first increased and then decreased (2 modules) and (iv) methylation levels increased (6 modules) (**Supplementary Data Set 1**). Three of these 20 modules covered 57.8% of all dynamic CpGs: module 8 defined by CpG demethylation starting in germinal center B cells and continuing in bone marrow

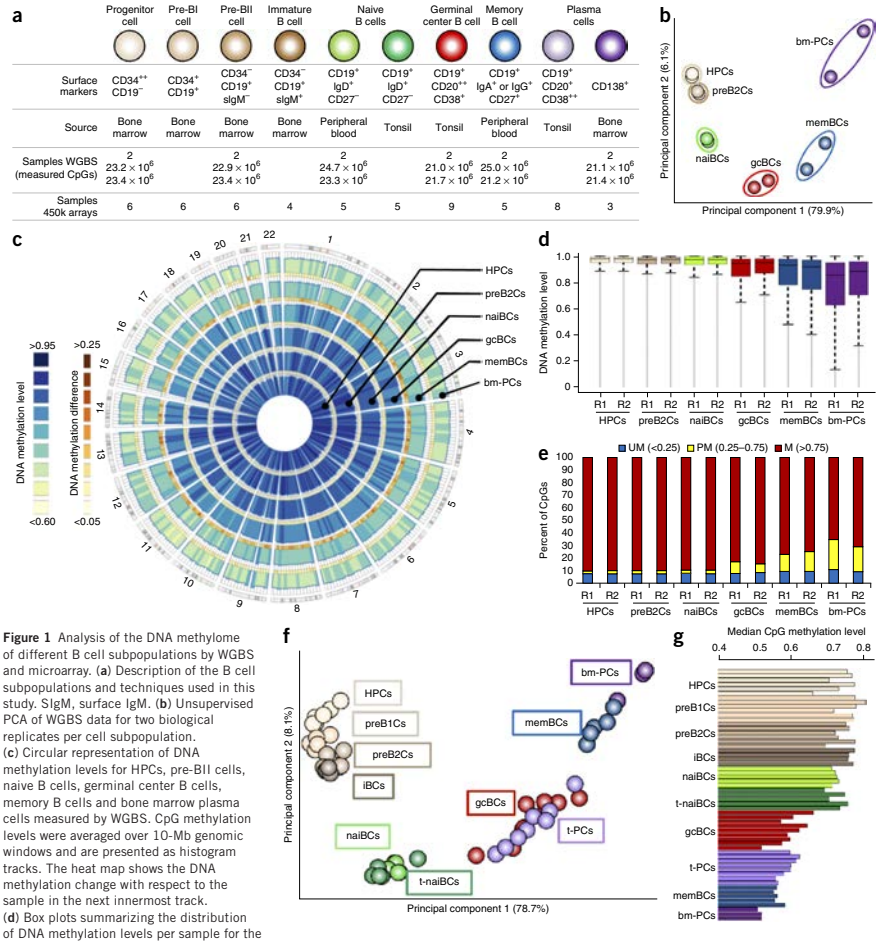


Figure 1 Analysis of the DNA methylome of different B cell subpopulations by WGBS and microarray. (**a**) Description of the B cell subpopulations and techniques used in this study. SIgM, surface IgM. (**b**) Unsupervised PCA of WGBS data for two biological replicates per cell subpopulation. (**c**) Circular representation of DNA methylation levels for HPCs, pre-BII cells, naive B cells, germinal center B cells, memory B cells and bone marrow plasma cells measured by WGBS. CpG methylation levels were averaged over 10-Mb genomic windows and are presented as histogram tracks. The heat map shows the DNA methylation change with respect to the sample in the next innermost track. (**d**) Box plots summarizing the distribution of DNA methylation levels per sample for the 16.1 million CpGs with methylation estimates in all 12 samples. The lower and upper limits of the colored boxes represent the first and third quartiles, respectively, and the black horizontal line is the median. Whiskers indicate the variability outside the upper and lower quartiles, and outliers are plotted as gray points. (**e**) Global methylation status of samples measured by WGBS. Percentages of methylated (M; red), partially methylated (PM; yellow) and unmethylated (UM; blue) CpGs are shown. (**f**) Unsupervised PCA of microarray methylation data for all samples used in the study. (**g**) Median values of DNA methylation data measured by microarray. HPCs, hematopoietic progenitor cells; preB1Cs, pre-BI cells; preB2Cs, pre-BII cells; iBCs, immature B cells; naiBCs, naive B cells from peripheral blood; t-naiBCs, naive B cells from tonsil; gcBCs, germinal center B cells; t-PCs, plasma cells from tonsil; memBCs, memory B cells from peripheral blood; bm-PCs, plasma cells from bone marrow. In **d** and **e**, R1 and R2 refer to the two biological replicates.

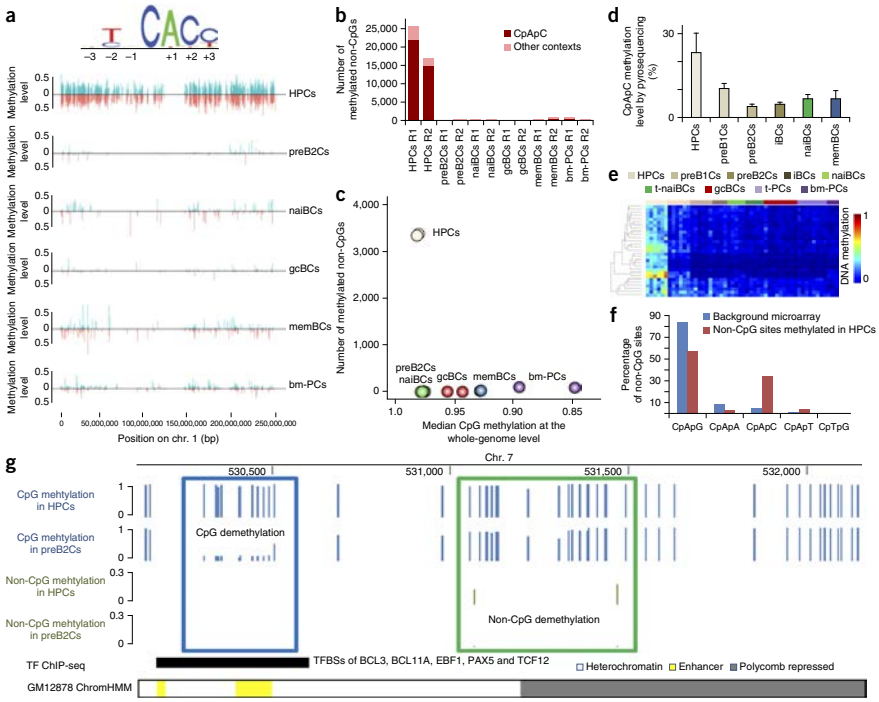


Figure 2 Non-CpG methylation detected during B cell differentiation. (**a**) Browser representation of non-CpG methylation, which takes place mostly in the CpApC sequence context (only chromosome 1 is shown). Methylation on the reverse strand is marked in blue, whereas that on the forward strand appears in red. (**b**) Number of non-CpG sites with nonzero methylation in different B cell subpopulations detected by WGBS. Methylated cytosines in the CpApC context are marked in dark red, and those in other contexts are marked in pale red. (**c**) Scatter plot showing the numbers of methylated non-CpG sites (using only the 3,437 non-CpGs with methylation estimates in all 12 samples and methylated in at least one of them) and median CpG methylation levels. In this analysis, 99% of the non-CpGs methylated in one HPC sample were also methylated in the biological replicate. (**d**) Validation of non-CpG methylation by BPS in two independent biological replicates of each subpopulation. For this analysis, we used a CpApC site (chr. 2: 85,933,406) shown to be methylated in HPCs by WGBS. Error bars, s.d. (**e**) Heat-map representation of 26 methylated non-CpGs measured by microarray (mean methylation = 34.7%). (**f**) Percentage of methylated non-CpGs in distinct sequence contexts detected by microarray. (**g**) Representation of CpG and non-CpG dynamics upon B cell commitment. CpG methylation is marked in blue, and non-CpG methylation is marked in green. Regions with CpG methylation loss (enhancer region, blue box) and non-CpG methylation loss (heterochromatin and Polycomb-repressed region, green box) are not coupled. The ChromHMM track, available at the UCSC Genome Browser, shows the chromatin states of the displayed genomic region in the lymphoblastoid B cell line GM12878. TF, transcription factor.

plasma cells (n = 34,604), module 9 showing demethylation only in bone marrow plasma cells (n = 13,044) and module 20 characterized by hypermethylation in only bone marrow plasma cells (n = 13,949) (**Fig. 3c**). These data confirmed that DNA hypomethylation mainly occurs in germinal center B cells and bone marrow plasma cells. Furthermore, it could be appreciated that hypermethylation is a late event, mainly occurring in bone marrow plasma cells. Overall, we may conclude that 84.5% of dynamic CpGs either gain or lose methylation (DNA methylation modulation patterns (i) and (iv); n = 90,070) as B cell differentiation progresses. Hence, each B cell differentiation stage has its specific DNA methylation pattern but furthermore retains an epigenetic memory of the previous stages. Interestingly, although B cells can be separated by maturation stage on the basis of the methylation levels of the 20 modules, we performed a complexity reduction step and selected 5 CpGs in genes important for B cell differentiation, such as *BLK*, *SEMA4B*, *ARID3A*, *AICDA* and *PRDM1*, whose methylation levels could accurately classify B cells into each maturation stage (**Supplementary Fig. 11**).

Functional analysis of dynamically methylated regions

In general, the CpGs losing methylation at any B cell maturation stage were preferentially located in introns, intergenic regions and repetitive elements (for example, long interspersed nuclear elements (LINEs), short interspersed nuclear elements (SINEs) and long

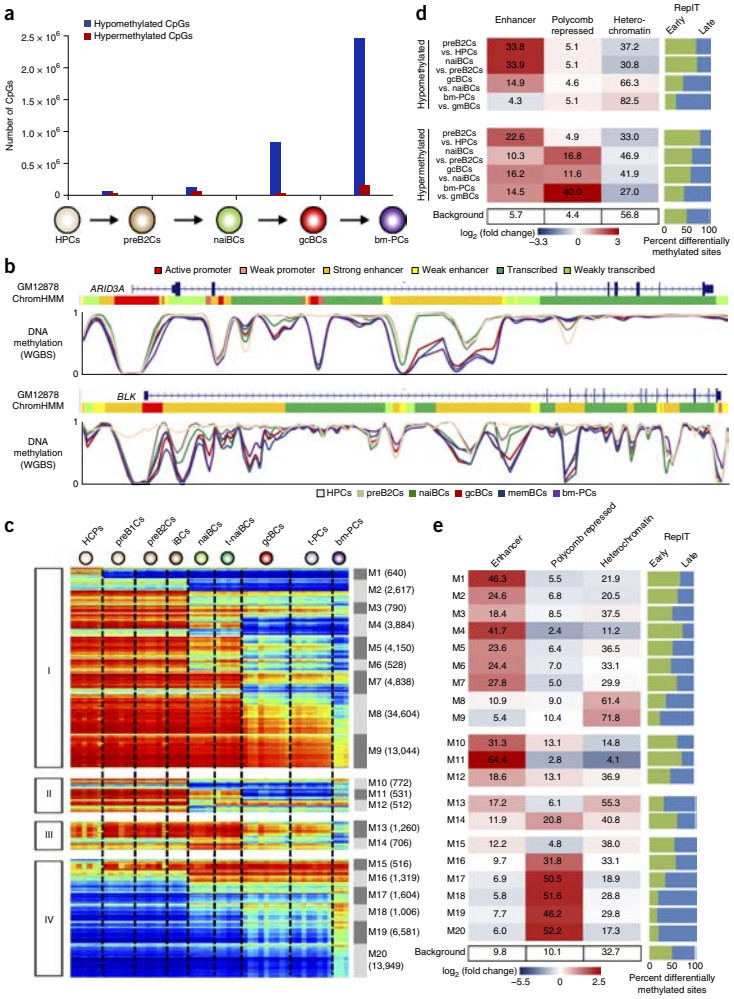


Figure 3 Dynamic DNA methylation during B cell differentiation. (**a**) Differentially methylated CpGs detected by WGBS considering the two replicates per cell subpopulation (see the Online Methods for an explanation of the criteria). (**b**) Smoothed DNA methylation data generated by WGBS across the promoter regions and gene bodies of *ARID3A* and *BLK*. The DNA methylation patterns for these genes are differentially modulated in different B cell subpopulations, especially in enhancer regions. (**c**) Heat-map representation of 20 major modules of dynamic CpGs, divided into 4 different patterns, detected by microarray. The number of CpGs within each module is given in parentheses. (**d**) Chromatin state characterization of differentially methylated CpGs identified by WGBS. (**e**) Chromatin state characterization of the 20 major modules detected by microarray. In **d** and **e**, numbers indicate the percentage of sites located in enhancers, Polycomb-repressed regions or heterochromatin. The blue-to-red color scale represents $\log_2$ (fold change), with respect to the background. Green and blue bars represent the percentages of differentially methylated sites that reside in early- or late-replicating regions, respectively. RepIT, replication timing.

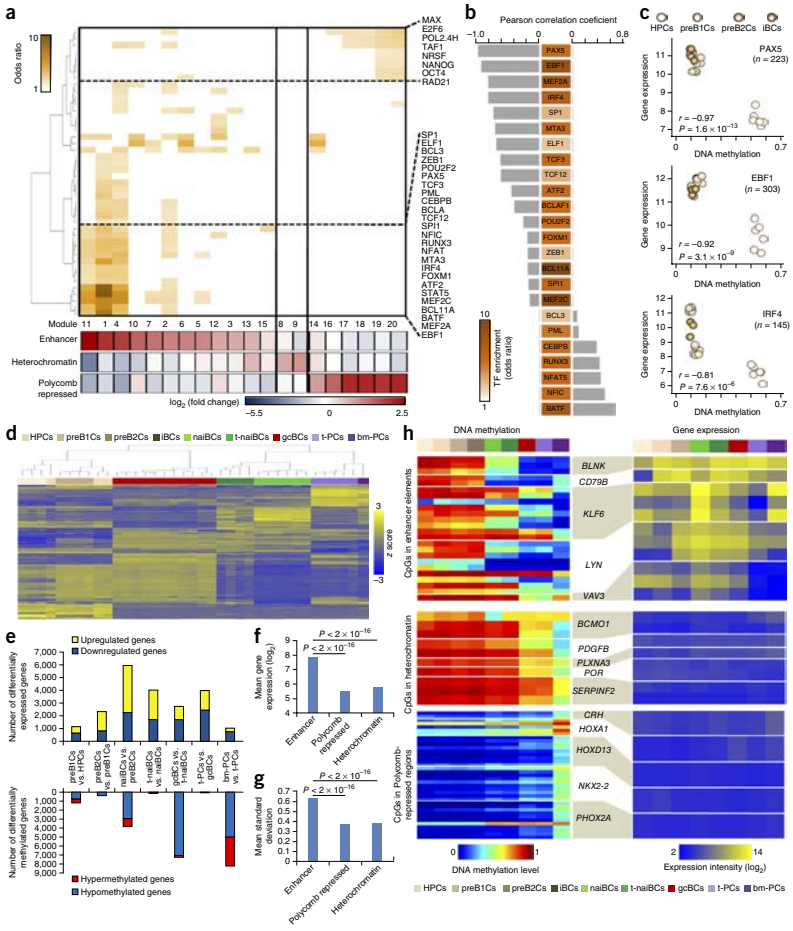


Figure 4 Association between DNA methylation and gene expression in different chromatin states. (**a**) Heat map presenting significant ($P < 0.01$) enrichments for TFBSs in the different methylation modules identified by methylation array. Below the heat map, $\log_2$ (fold change) for enhancer, heterochromatin and Polycomb-repressed regions in each differentially methylated group is presented. (**b**) Correlation between the expression levels of transcription factors and the mean methylation levels of their binding sites using samples with available expression and methylation data for the same donors. We applied this analysis to methylation data for module 1 (demethylation upon B cell commitment) and gene expression data for precursor cells. The white-to-brown color scale represents the odds ratio for TFBS enrichment. (**c**) Scatter plots showing correlation of the expression levels of PAX5, EBF1 and IRF4 with the mean methylation levels of their binding sites in each sample (the number of TFBSs associated with CpGs belonging to module 1 is shown below the transcription factor name). (**d**) Unsupervised clustering analysis of gene expression data using the 687 tags (439 genes) with the highest variability (s.d. >2) across the B cell differentiation process. (**e**) Differentially expressed genes (top) and differentially methylated genes (bottom) in each comparison of adjacent cell subpopulations. (**f**,**g**) Mean expression levels (**f**) and expression variability (**g**) during B cell differentiation for genes containing dynamic CpGs targeting enhancers, Polycomb-repressed regions and heterochromatin. (**h**) Heat maps showing the DNA methylation levels (left) and gene expression levels (right) of representative genes with dynamic methylation in enhancers, heterochromatin and Polycomb-repressed regions.

terminal repeats (LTRs)) and were enriched for genomic areas lacking CpG islands (CGIs) (**Supplementary Figs. 12** and **13**). In contrast, CpGs gaining methylation were enriched for CGIs and promoter regions (**Supplementary Fig. 12**). Additionally, we classified differentially methylated sites using categorization of the genome into the different chromatin states observed in immortalized mature B cells[28] (**Supplementary Fig. 14**). Both WGBS and microarray data showed that the majority of dynamic CpGs during B cell differentiation were enriched for enhancer regions (mainly intragenic), Polycomb-repressed regions or heterochromatin (**Fig. 3d**,**e** and **Supplementary Figs. 15** and **16**). Demethylation in precursor B cells was mostly related to enhancer elements, whereas that occurring exclusively from germinal center B cells onward was preferentially located in heterochromatic regions. Gain of CpG methylation was a rare event in early B cell differentiation but was rather frequent in mature B cells, especially in bone marrow plasma cells. Such CpG hypermethylation preferentially targeted Polycomb-repressed regions (**Fig. 3d**,**e**).

We next studied the mechanisms underlying enhancer demethylation in the B cell differentiation process. We identified a significant enrichment (fold change > 2, $P < 0.01$) in these enhancers of the transcription factor binding sites (TFBSs) of key B cell transcription factors such as BCL11A, EBF1, IRF4, MEF2A, MEF2C, PAX5 or TCF3 (E2A) (**Fig. 4a**, **Supplementary Fig. 17** and **Supplementary Data Sets 2** and **3**). As B cell commitment is associated with the expression of lineage-specific transcription factors[20], we analyzed the transition from HPCs to pre-BI cells (module 1) in detail and observed globally an inverse correlation between the expression of transcription factors and the methylation levels of their binding sites (**Fig. 4b**,**c**). We further investigated the association between expression of transcription factors and the methylation status of their binding sites over the entire differentiation program; in general, once a TFBS became demethylated at any B cell differentiation stage, it remained unmethylated in subsequent stages, suggesting an epigenetic memory of transcription factor binding[24] (**Supplementary Fig. 18**).

At the functional level, genes within microarray-based methylation modules enriched for enhancer elements were involved in multiple immune system–related functions (**Supplementary Fig. 19** and **Supplementary Data Set 4**). In contrast, CpGs in modules enriched for heterochromatin or Polycomb-repressed regions did not target genes involved in the immune system but rather affected terms such as development, locomotion or behavior (**Supplementary Fig. 19** and **Supplementary Data Set 4**). On the basis of these observations, we hypothesized that differential methylation in enhancer elements might be globally associated with gene expression, whereas that affecting inactive elements (heterochromatin or Polycomb-repressed regions) might not. We initially explored the transcriptome of B cell subpopulations and observed that they clustered separately using an unsupervised approach (**Fig. 4d**). We then calculated the numbers of genes differentially expressed and differentially methylated by comparing adjacent cell subpopulations. Globally, there was poor association between these numbers, and large transcriptional changes could be related to minor modulation of the DNA methylome and vice versa (**Fig. 4e**). We further explored the association between DNA methylation patterns and gene expression by focusing on genes with dynamic methylation in enhancers, heterochromatin and Polycomb-repressed regions. We observed that both the variability in expression and mean expression levels of genes containing dynamic CpGs in enhancer elements were much higher than for genes showing modulation of CpG methylation in non-functional chromatin states such as heterochromatin and Polycomb-repressed regions ($P < 0.001$; **Fig. 4f**–**h**). However, regardless of the DNA methylation pattern throughout B cell maturation, dynamic CpGs targeted genes with higher expression levels and greater variation in expression throughout the entire B cell differentiation process. Thus, these data suggest that there is no direct correlation between DNA methylation and gene expression, as previously shown elsewhere[6,10], but rather that dynamic CpGs affecting functional elements target immune system–related genes whose expression is modulated during B cell differentiation.

We next sought to analyze the functions of genes with dynamic enhancer methylation in more detail. As B cell receptor signaling is a key

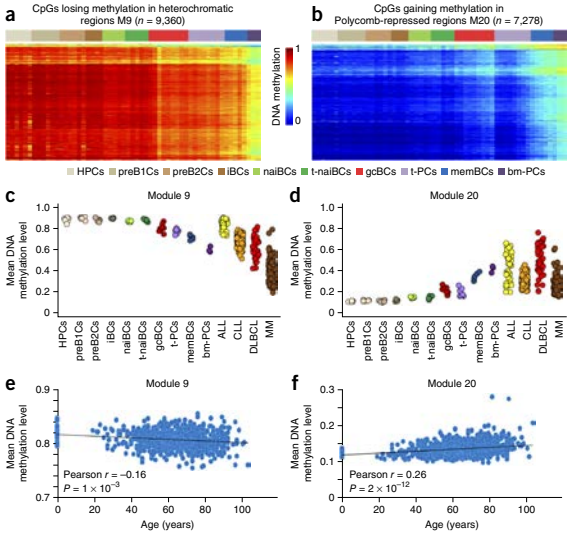


Figure 5 DNA methylation changes during B cell differentiation in the context of cancer and aging. (**a**) Heat map of a subset of CpGs from module 9 that lose methylation in heterochromatic regions. (**b**) Heat map of a subset of CpGs from module 20 that gain methylation in Polycomb-repressed regions. (**c**,**d**) Scatter plots presenting the mean methylation levels of CpGs in heterochromatin from module 9 (**c**) and in Polycomb-repressed regions from module 20 (**d**) in different B cell subsets and four types of hematological neoplasm. (**e**,**f**) Mean methylation levels of CpGs in heterochromatin from module 9 (**e**) and in Polycomb-repressed regions from module 20 (**f**) in whole-blood samples isolated from donors of different age. ALL, acute lymphoblastic leukemia; CLL, chronic lymphocytic leukemia; DLBCL, diffuse large B cell lymphoma; MM, multiple myeloma.

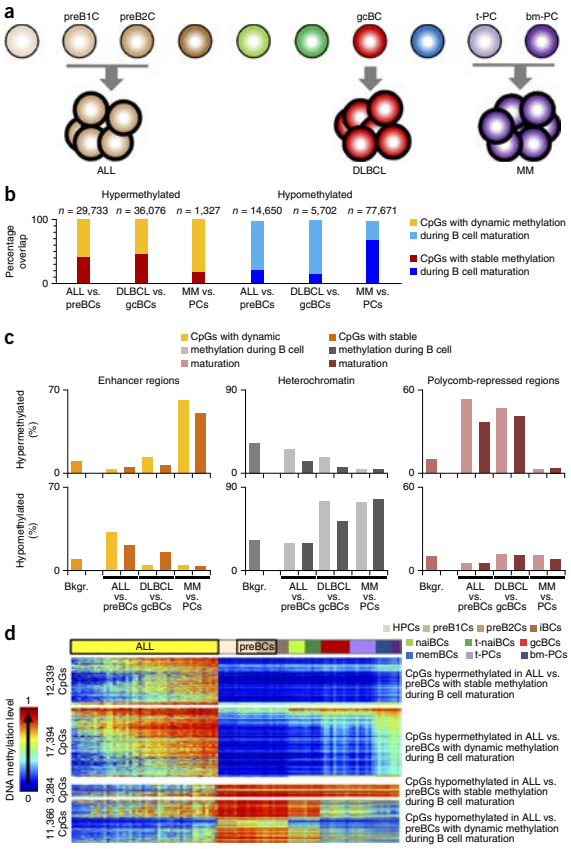


Figure 6 DNA methylation changes in various B cell neoplasms in comparison to their normal counterparts. (**a**) Analysis of differential methylation was performed for three models of lymphoid neoplasm that arise from three distinct maturation stages in B cell development: ALL versus precursor B cells (pre-BI and pre-BII cells), the GCB subgroup of DLBCL versus germinal center B cells, and MM versus plasma cells (from the tonsil and bone marrow). (**b**) Bar plots showing the proportion of dynamically methylated CpGs in B cell differentiation that are also differentially methylated in hematological neoplasias in comparison to their normal counterparts. (**c**) Percentages of hypermethylated (top) and hypomethylated (bottom) CpGs located in enhancers (left), heterochromatin (middle) and Polycomb-repressed regions (right). (**d**) Heat maps showing differentially methylated CpGs in ALL in comparison to precursor B cells in the context of normal B cell differentiation. ALL, acute lymphoblastic leukemia; DLBCL, diffuse large B cell lymphoma; MM, multiple myeloma; Bkgr., background for 450k microarray data.

element specific for B cell differentiation, we studied the methylation patterns of 41 genes involved in this function. Of these, 38 (93%) had a total of 234 CpGs with dynamic methylation that were preferentially located in enhancers (**Supplementary Fig. 20**). Furthermore, we observed that 38% of all enhancers with dynamic methylation ($P < 0.001$) belonged to a B cell–specific functional gene network[29] (**Supplementary Fig. 21**).

As our study comprised cell subpopulations isolated from three different compartments (bone marrow, peripheral blood and tonsil), we aimed to detect particular DNA methylation imprints related to these different locations, but we did not identify any consistent pattern. However, we did observe that naive B cells isolated from different compartments (peripheral blood or tonsils) showed a pronounced change in their transcriptomes, whereas their methylomes remained virtually identical (**Supplementary Figs. 22** and **23**). On the basis of this observation, we may hypothesize that the gene expression changes in naive tonsillar B cells are essential to optimize antigen recognition in the tonsil, followed by successful B cell activation. However, when no antigen is found, the naive tonsillar B cells will reenter the bloodstream. Hence, the naive B cell state in tonsils has to be reversible, which may be the reason why no changes were observed at the level of the DNA methylome.

Epigenetic link among long-lived B cells, cancer and aging

We observed that long-lived B cells such as memory B cells and bone marrow plasma cells[23] showed extensive perturbation of their DNA methylomes in comparison to germinal center B cells (**Supplementary Fig. 8**). Remarkably, although bone marrow plasma cells and newly generated tonsillar plasma cells had similar transcriptomes, their methylomes were different (**Supplementary Fig. 22**). The major DNA methylation changes in memory B cells and bone marrow plasma cells affected presumably non-functional elements such as heterochromatin and Polycomb-repressed regions. Thus, our findings suggest that part of the epigenetic makeup of memory B cells and bone marrow plasma cells is caused by epigenetic drift associated with their long lifespans[23]. To determine whether such drift might be related to the expression of DNMTs, we analyzed *DNMT1*, *DNMT3A* and *DNMT3B* transcript levels by quantitative RT-PCR (qRT-PCR) in sorted germinal center B cells, memory B cells, tonsillar plasma cells and bone marrow plasma cells. There were low levels of *DNMT1* in memory B cells, tonsillar plasma cells and bone marrow plasma cells (**Supplementary Fig. 7**). As these cell types are considered to be non-proliferative, our finding implies that demethylation occurs either through an active mechanism or passively if these cell types still proliferate at a low rate[30]. Gain of methylation in Polycomb-repressed regions has been

linked to the activities of DNMT3A and DNMT3B[31,32]. We detected that, in comparison to germinal center B cells, tonsillar plasma cells and bone marrow plasma cells, and to a lesser extent memory B cells, had upregulated *DNMT3A* levels. These results may suggest that minor increases in the levels of this enzyme in the context of long-lived cells could result in hypermethylation of Polycomb-repressed areas. To obtain further insights into the mechanisms of chromatin repression by hypermethylation in Polycomb-repressed regions (defined by the presence of histone H3 trimethylated at lysine 27, H3K27me3), we performed bisulfite sequencing of immunoprecipitated chromatin (ChIP-BS) using an antibody to H3K27me3 (refs. 33,34). This experiment suggested that, in memory B cells, H3K27me3 and DNA methylation coexist both in CpG-rich and CpG-poor regions and that the DNA methylation levels within nucleosomes containing H3K27me3 are lower than in regions outside such nucleosomes (**Supplementary Fig. 24**).

Remarkably, demethylation of heterochromatin (in part bound to the nuclear lamina) and hypermethylation of Polycomb-repressed regions are among the most frequent epigenetic changes in solid and hematological tumors[11,35–38]. To determine whether the CpGs showing differential methylation in long-lived B cells overlapped with those becoming differentially methylated in cancer, we used methylation data from various lymphoid neoplasms, including acute lymphoblastic leukemia[39] (ALL; $n = 46$), chronic lymphocytic leukemia[10] (CLL; $n = 139$), diffuse large B cell lymphoma (DLBCL; $n = 40$) and multiple myeloma[40] (MM; $n = 104$). We analyzed the methylation levels of CpGs losing methylation (modules 8 and 9) and those gaining methylation in memory B cells and bone marrow plasma cells (modules 19 and 20). Indeed, we observed that, for these CpGs, neoplastic cells showed a DNA methylation profile similar to those of memory B cells and bone marrow plasma cells (**Fig. 5** and **Supplementary Fig. 25**). To further evaluate the epigenetic link between normal B cell differentiation and neoplastic transformation, we compared the DNA methylomes of B cell neoplasms with those of their normal cell counterparts, for example, ALL[39] versus pre-B cells, germinal center B cell–like DLBCL versus germinal center B cells and MM[40] versus plasma cells (**Fig. 6a** and **Supplementary Data Set 5**). The results indicate that a large fraction of the CpGs differentially methylated in cancers are dynamically methylated during normal B cell differentiation, with the overlap ranging from 53 to 82% for hypermethylated sites and from 29 to 84% for hypomethylated sites (**Fig. 6b**). Interestingly, we found that hypomethylation in ALL was enriched for CpGs in enhancers, whereas hypomethylation in DLBCL and MM predominantly affected CpGs in heterochromatin (**Fig. 6c**). Additionally, although ALL cells are arrested at the pre-B cell stage, they acquired hypermethylation in Polycomb-repressed regions, which is characteristic of more mature differentiation stages (**Fig. 6d**). MM cells, in contrast, did not acquire hypermethylation of Polycomb-repressed regions, as their cell of origin already shows this feature, but, as they downregulate the B cell program, they acquire hypermethylation of CpGs in B cell–specific enhancers[40].

Finally, global hypomethylation accompanied by local hypermethylation of Polycomb targets is also a molecular hallmark of aging[41–45]. We analyzed the DNA methylation values of CpGs within modules 8, 9, 19 and 20 using data from 694 peripheral blood samples obtained from donors ranging in age from 0 to 101 years[43,44]. Indeed, we detected a significant correlation with age ($P < 0.01$), following the same trend in long-lived cells and cancer cells—that is, methylation of heterochromatin diminished with age, whereas methylation of Polycomb targets increased (**Fig. 5e**,**f** and **Supplementary Fig. 25**). We next compared B cells with short and long lifespans (naive B cells and memory B cells, respectively) isolated simultaneously from individuals of different age, and we found that memory B cells acquired hypomethylation of heterochromatin and hypermethylation of Polycomb targets regardless of the chronological age of the donor (**Supplementary Fig. 26**). As the cellular composition of blood changes with age, our results imply that a relative increase in the fraction of long-lived cells in older individuals may represent a confounding variable in age-related methylation studies[46].

DISCUSSION

The B cell maturation process is an orchestrated program integrating internal and environmental signals to finally give rise to plasma cells and memory B cells that have an essential role in adaptive immunity. Although previous reports have studied epigenetic changes in the context of B cell differentiation, they only studied partial DNA methylomes either of precursor or mature B cells[17–19,47]. With the exception of a few cell subpopulations such as transitional B cells, $CD5^+$ B cells and splenic marginal zone B cells, our study comprises all major B cell differentiation stages and represents the first whole-genome epigenetic characterization of a complete human cell lineage from progenitor to terminally differentiated cells. The comprehensive nature of our study has allowed us to provide epigenetic insights into different scientific fields and offers a resource for researchers working in different areas of cell differentiation, B cell biology and related diseases, cancer and aging, both at single-gene and genome-wide levels.

Our study points to a massive perturbation of the DNA methylome during B cell differentiation, affecting 30% of all autosomal CpG sites. These changes follow an accumulative pattern in which each B cell maturation stage, although characterized by a particular signature, keeps an epigenetic memory of previous differentiation stages. In contrast to other reports for hematopoietic precursors[48,49], we did observe non-CpG methylation in HPCs, which virtually disappeared upon B cell commitment in regions lacking simultaneous CpG demethylation. Precursor B cells showed relatively small losses in CpG methylation, which mostly affected enhancers containing binding sites for B cell–specific transcription factors. The functional link between transcription factor binding, CpG demethylation and enhancer activation has recently been analyzed during stem cell differentiation[50,51] as well as in hematopoietic cells[52,53], non-hematopoietic cells[54] and cancer[40,55].

Interestingly, more than half of all enhancers defined in immortalized mature B cells show dynamic DNA methylation throughout the B cell differentiation process, and 38% of all genes with dynamically methylated enhancers are included in a regulatory network associated with human B cells[29]. Although mainstream research on DNA methylation still remains centered on promoter regions, our results imply that DNA methylation changes in enhancers seem to be more closely related to cell specification and maturation[1,9,56]. However, similarly to other recent studies[6,10,57], we rarely observed a direct correlation between gene expression and DNA methylation, even in regulatory elements. Our study also suggests that, at later stages of B cell differentiation (from naive B cells onward), DNA methylation changes are guided more by other mechanisms than by the intrinsic program of B cell transcription factors. Upon antigen encounter, the germinal center reaction is induced and, at this stage, germinal center B cells start experiencing a wave of global demethylation, mostly affecting late-replicating regions such as heterochromatin and DNA repeats, and local hypermethylation of Polycomb-repressed regions. This finding can be partially explained by the high proliferation rate of germinal center B cells, as normal proliferative tissues tend to lose

methylation at late-replicating regions[58]. However, downstream B cell subpopulations derived from germinal center B cells, such as non-proliferative memory B cells and plasma cells, which recirculate through the body and reside in bone marrow, respectively, acquire additional epigenetic changes in heterochromatin and Polycomb-repressed regions. We postulate that these additional changes may be related to potential epigenetic drift in the context of longevity that may be mediated by downregulation of *DNMT1* and slight upregulation of *DNMT3A*.

Hypomethylation of heterochromatin and hypermethylation of Polycomb-repressed regions has previously been described as an epigenetic hallmark of organismal aging, cellular senescence and cancer[4,41–44,59–61]. Here we observe that this signature starts in proliferating germinal center B cells and becomes particularly enhanced in non-proliferative, long-lived B cells. On the basis of our results, we hypothesize that not all cells in an organism are subject to epigenetic drift as a consequence of time, only those with a long lifespan. Finally, one of the most relevant implications of our study is related to the field of cancer. We demonstrate that B cell tumors and long-lived cells have similar DNA methylation signatures. Furthermore, comparing various B cell neoplasms with their normal cellular counterparts, we observe that a large proportion of the differentially methylated sites in cancers overlap with those undergoing dynamic methylation during normal differentiation, especially with those altered in memory B cells and bone marrow plasma cells. Interestingly, as in pre-B cells, hypomethylation in ALL is enriched for enhancer elements, whereas, in line with germinal center B cells and bone marrow plasma cells, hypomethylation in DLBCL and MM is mostly enriched for heterochromatin. In general, these findings suggest that the epigenetic configuration of a cell from a particular maturation stage influences the DNA methylation changes acquired during its clonal expansion and neoplastic transformation. This new strategy of analyzing the DNA methylome of B cell tumors in the context of the entire differentiation program may allow new insights into the role of DNA methylation in cancer. We postulate that methylation changes shared by neoplastic transformation and normal differentiation may represent epigenetic passengers, whereas those exclusively taking place in tumor cells should constitute epigenetic drivers with a potential functional impact in the disease.

URLs. European Genome-phenome Archive (EGA), http://www.ebi.ac.uk/ega/studies/; Blueprint project, http://www.blueprint-epigenome.eu/; R Project for Statistical Computing, http://www.r-project.org/; UCSC Genome Browser, http://genome.ucsc.edu/.

METHODS

Methods and any associated references are available in the online version of the paper.

Accession codes. WGBS data have been deposited in the European Genome-phenome Archive (EGA) under accessions EGAD00001001304 and EGAS00001000272. DNA methylation and gene expression microarray data are available from the EGA under accessions EGAS00001001196 and EGAS00001001197, respectively.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

We thank C. López-Otín for critical reading of this manuscript, M. Dabad Castellà for his assistance with the WGBS data analysis and M.A. Peinado (Institute of Predictive and Personalized Medicine of Cancer, Barcelona) for providing RNA from HCT116 DKO cells. This work was funded by the European Union's Seventh Framework Programme through the Blueprint Consortium (grant agreement 282510) and the Spanish Ministry of Economy and Competitivity (MINECO; project SAF2009-08663). Methylation microarrays were outsourced to the Spanish Centro Nacional de Genotipado (CEGEN-ISCIII). We are indebted to the Genomics core facility of the Institut d'Investigacions Biomèdiques August Pi i Sunyer (IDIBAPS) for technical help. This work was partially developed at the Centro Esther Koplowitz (CEK; Barcelona, Spain). M.K. is supported by Agència de Gestió d'Ajuts Universitaris i de Recerca (AGAUR; Generalitat de Catalunya), E.C. is an Academia Researcher of the Institució Catalana de Recerca i Estudis Avançats and J.I.M.-S. is a Ramón y Cajal researcher of MINECO.

AUTHOR CONTRIBUTIONS

M.K., A.C.Q., N.R., M.P., X.A., F.P., D.A., B.P., G. Caron, T.F., M.O.M., M.E.F., S.-T.L. and J.L.W. provided samples from healthy donors and/or purified B cell subpopulations. M.K., A.C.Q., G. Castellano, R.B. and G. Clot analyzed DNA methylation and gene expression arrays. L.A., J.B. and M.G. performed WGBS library preparation and sequencing. A.M., S.H., R.P.S., E.R., A.E. and M.D.-F. processed and analyzed WGBS data. M.K., N.V.-D. and R.V.-B. performed validation experiments. M.K., G. Castellano. S.E., V.P., D. Rico and A.V. functionally characterized dynamically methylated genes. D. Richardson, L.C., A.D. and P.F. were in charge of data management. I.G.G. and H.G.S. coordinated sequencing efforts and performed primary data analysis. H.G.S., R.S., R.K. and E.C. participated in the study design and data interpretation. J.I.M.-S. conceived the study. J.I.M.-S. led the experiments and wrote the manuscript with predominant assistance from M.K. and R.B.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

1. Ziller, M.J. *et al.* Charting a dynamic DNA methylation landscape of the human genome. *Nature* **500**, 477–481 (2013).
2. Bernstein, B.E., Meissner, A. & Lander, E.S. The mammalian epigenome. *Cell* **128**, 669–681 (2007).
3. Smith, Z.D. & Meissner, A. DNA methylation: roles in mammalian development. *Nat. Rev. Genet.* **14**, 204–220 (2013).
4. Bergman, Y. & Cedar, H. DNA methylation dynamics in health and disease. *Nat. Struct. Mol. Biol.* **20**, 274–281 (2013).
5. Bird, A. DNA methylation patterns and epigenetic memory. *Genes Dev.* **16**, 6–21 (2002).
6. Hovestadt, V. *et al.* Decoding the regulatory landscape of medulloblastoma using DNA methylation sequencing. *Nature* **510**, 537–541 (2014).
7. Lister, R. *et al.* Global epigenomic reconfiguration during mammalian brain development. *Science* **341**, 1237905 (2013).
8. Lister, R. *et al.* Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature* **462**, 315–322 (2009).
9. Gifford, C.A. *et al.* Transcriptional and epigenetic dynamics during specification of human embryonic stem cells. *Cell* **153**, 1149–1163 (2013).
10. Kulis, M. *et al.* Epigenomic analysis detects widespread gene-body DNA hypomethylation in chronic lymphocytic leukemia. *Nat. Genet.* **44**, 1236–1242 (2012).
11. Berman, B.P. *et al.* Regions of focal DNA hypermethylation and long-range hypomethylation in colorectal cancer coincide with nuclear lamina-associated domains. *Nat. Genet.* **44**, 40–46 (2012).
12. Habibi, E. *et al.* Whole-genome bisulfite sequencing of two distinct interconvertible DNA methylomes of mouse embryonic stem cells. *Cell Stem Cell* **13**, 360–369 (2013).
13. Hansen, K.D. *et al.* Increased methylation variation in epigenetic domains across cancer types. *Nat. Genet.* **43**, 768–775 (2011).
14. Li, Y. *et al.* The DNA methylome of human peripheral blood mononuclear cells. *PLoS Biol.* **8**, e1000533 (2010).
15. Xie, W. *et al.* Epigenomic analysis of multilineage differentiation of human embryonic stem cells. *Cell* **153**, 1134–1148 (2013).
16. Varley, K.E. *et al.* Dynamic DNA methylation across diverse human cell lines and tissues. *Genome Res.* **23**, 555–567 (2013).
17. Lee, S.T. *et al.* A global DNA methylation and gene expression analysis of early human B-cell development reveals a demethylation signature and transcription factor network. *Nucleic Acids Res.* **40**, 11339–11351 (2012).
18. Lai, A.Y. *et al.* DNA methylation profiling in human B cells reveals immune regulatory elements and epigenetic plasticity at Alu elements during B cell activation. *Genome Res.* **23**, 2030–2041 (2013).
19. Shaknovich, R. *et al.* DNA methyltransferase 1 and DNA methylation patterning contribute to germinal center B-cell differentiation. *Blood* **118**, 3559–3569 (2011).
20. Matthias, P. & Rolink, A.G. Transcriptional networks in developing and mature B cells. *Nat. Rev. Immunol.* **5**, 497–508 (2005).

21. Kurosaki, T., Shinohara, H. & Baba, Y. B cell signaling and fate decision. *Annu. Rev. Immunol.* **28**, 21–55 (2010).
22. Manz, R.A., Thiel, A. & Radbruch, A. Lifetime of plasma cells in the bone marrow. *Nature* **388**, 133–134 (1997).
23. Slifka, M.K., Antia, R., Whitmire, J.K. & Ahmed, R. Humoral immunity due to long-lived plasma cells. *Immunity* **8**, 363–372 (1998).
24. Hon, G.C. *et al.* Epigenetic memory at embryonic enhancers identified in DNA methylation maps from adult mouse tissues. *Nat. Genet.* **45**, 1198–1206 (2013).
25. Bibikova, M. *et al.* High density DNA methylation array with single CpG site resolution. *Genomics* **98**, 288–295 (2011).
26. Ziller, M.J. *et al.* Genomic distribution and inter-sample variation of non-CpG methylation across human cell types. *PLoS Genet.* **7**, e1002389 (2011).
27. Arand, J. *et al. In vivo* control of CpG and non-CpG DNA methylation by DNA methyltransferases. *PLoS Genet.* **8**, e1002750 (2012).
28. Ernst, J. *et al.* Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature* **473**, 43–49 (2011).
29. Lefebvre, C. *et al.* A human B-cell interactome identifies MYB and FOXM1 as master regulators of proliferation in germinal centers. *Mol. Syst. Biol.* **6**, 377 (2010).
30. Tooze, R.M. A replicative self-renewal model for long-lived plasma cells: questioning irreversible cell cycle exit. *Front. Immunol.* **4**, 460 (2013).
31. Viré, E. *et al.* The Polycomb group protein EZH2 directly controls DNA methylation. *Nature* **439**, 871–874 (2006).
32. Cedar, H. & Bergman, Y. Linking DNA methylation and histone modification: patterns and paradigms. *Nat. Rev. Genet.* **10**, 295–304 (2009).
33. Brinkman, A.B. *et al.* Sequential ChIP-bisulfite sequencing enables direct genome-scale investigation of chromatin and DNA methylation cross-talk. *Genome Res.* **22**, 1128–1138 (2012).
34. Statham, A.L. *et al.* Bisulfite sequencing of chromatin immunoprecipitated DNA (BisChIP-seq) directly informs methylation status of histone-modified DNA. *Genome Res.* **22**, 1120–1127 (2012).
35. Martín-Subero, J.I. *et al.* New insights into the biology and origin of mature aggressive B-cell lymphomas by combined epigenomic, genomic, and transcriptional profiling. *Blood* **113**, 2488–2497 (2009).
36. Ohm, J.E. *et al.* A stem cell–like chromatin pattern may predispose tumor suppressor genes to DNA hypermethylation and heritable silencing. *Nat. Genet.* **39**, 237–242 (2007).
37. Schlesinger, Y. *et al.* Polycomb-mediated methylation on Lys27 of histone H3 pre-marks genes for *de novo* methylation in cancer. *Nat. Genet.* **39**, 232–236 (2007).
38. Widschwendter, M. *et al.* Epigenetic stem cell signature in cancer. *Nat. Genet.* **39**, 157–158 (2007).
39. Busche, S. *et al.* Integration of high-resolution methylome and transcriptome analyses to dissect epigenomic changes in childhood acute lymphoblastic leukemia. *Cancer Res.* **73**, 4323–4336 (2013).
40. Agirre, X. *et al.* Whole-epigenome analysis in multiple myeloma reveals DNA hypermethylation of B cell–specific enhancers. *Genome Res.* **25**, 478–487 (2015).
41. López-Otín, C., Blasco, M.A., Partridge, L., Serrano, M. & Kroemer, G. The hallmarks of aging. *Cell* **153**, 1194–1217 (2013).
42. Horvath, S. *et al.* Aging effects on DNA methylation modules in human brain and blood tissue. *Genome Biol.* **13**, R97 (2012).
43. Hannum, G. *et al.* Genome-wide methylation profiles reveal quantitative views of human aging rates. *Mol. Cell* **49**, 359–367 (2013).
44. Heyn, H. *et al.* Distinct DNA methylomes of newborns and centenarians. *Proc. Natl. Acad. Sci. USA* **109**, 10522–10527 (2012).
45. Maegawa, S. *et al.* Widespread and tissue specific age-related DNA methylation changes in mice. *Genome Res.* **20**, 332–340 (2010).
46. Jaffe, A.E. & Irizarry, R.A. Accounting for cellular heterogeneity is critical in epigenome-wide association studies. *Genome Biol.* **15**, R31 (2014).
47. Deaton, A.M. *et al.* Cell type–specific DNA methylation at intragenic CpG islands in the immune system. *Genome Res.* **21**, 1074–1086 (2011).
48. Hodges, E. *et al.* Directional DNA methylation changes and complex intermediate states accompany lineage specificity in the adult hematopoietic compartment. *Mol. Cell* **44**, 17–28 (2011).
49. Jeong, M. *et al.* Large conserved domains of low DNA methylation maintained by Dnmt3a. *Nat. Genet.* **46**, 17–23 (2014).
50. Tsankov, A.M. *et al.* Transcription factor binding dynamics during human ES cell differentiation. *Nature* **518**, 344–349 (2015).
51. Stadler, M.B. *et al.* DNA-binding factors shape the mouse methylome at distal regulatory regions. *Nature* **480**, 490–495 (2011).
52. Schmidl, C. *et al.* Lineage-specific DNA methylation in T cells correlates with histone methylation and enhancer activity. *Genome Res.* **19**, 1165–1174 (2009).
53. Tagoh, H. *et al.* Dynamic reorganization of chromatin structure and selective DNA demethylation prior to stable enhancer complex formation during differentiation of primary hematopoietic cells *in vitro*. *Blood* **103**, 2950–2955 (2004).
54. Wiench, M. *et al.* DNA methylation status predicts cell type–specific enhancer activity. *EMBO J.* **30**, 3028–3039 (2011).
55. Taberlay, P.C., Statham, A.L., Kelly, T.K., Clark, S.J. & Jones, P.A. Reconfiguration of nucleosome-depleted regions at distal regulatory elements accompanies DNA methylation of enhancers and insulators in cancer. *Genome Res.* **24**, 1421–1432 (2014).
56. Schlesinger, F., Smith, A.D., Gingeras, T.R., Hannon, G.J. & Hodges, E. *De novo* DNA demethylation and noncoding transcription define active intergenic regulatory elements. *Genome Res.* **23**, 1601–1614 (2013).
57. Aran, D., Sabato, S. & Hellman, A. DNA methylation of distal regulatory sites characterizes dysregulation of cancer genes. *Genome Biol.* **14**, R21 (2013).
58. Aran, D., Toperoff, G., Rosenberg, M. & Hellman, A. Replication timing–related and gene body–specific methylation of active human genes. *Hum. Mol. Genet.* **20**, 670–680 (2011).
59. Teschendorff, A.E. *et al.* Age-dependent DNA methylation of genes that are suppressed in stem cells is a hallmark of cancer. *Genome Res.* **20**, 440–446 (2010).
60. Cruickshanks, H.A. *et al.* Senescent cells harbour features of the cancer epigenome. *Nat. Cell Biol.* **15**, 1495–1506 (2013).
61. Rakyan, V.K. *et al.* Human aging-associated DNA hypermethylation occurs preferentially at bivalent chromatin domains. *Genome Res.* **20**, 434–439 (2010).

ONLINE METHODS

Isolation of B cell subpopulations. Precursor B cells were isolated from fetal bone marrow (22-week fetuses) using flow cytometry sorting. Early progenitors were isolated on the basis of high levels of CD34 protein expression ($CD34^{hi}$) and lack of expression of the B cell marker CD19. This population, designated uncommitted HPCs, contained predominantly multipotent progenitors before lineage commitment and also common lymphoid progenitors and hematopoietic stem cells. B cell–committed progenitors were isolated on the basis of their expression of CD19 and CD34 ($CD19^+CD34^+$) and were predominantly pre-BI cells. Two immature B cell populations expressing CD19 and lacking CD34 were isolated and differentiated on the basis of sIgM expression: pre-BII cells that were $sIgM^-CD19^+$ and immature B cells that were $sIgM^+CD19^+$. DNA methylation and gene expression data for these four subpopulations have been published previously[17].

Peripheral blood B cell subpopulations—naive B cells and memory B cells—were obtained from buffy coats for healthy adult donors ranging in age between 28 and 66 years. After Ficoll-Isopaque density centrifugation, $CD19^+$ B cells were isolated by positive magnetic cell separation using the AutoMACS system (Miltenyi Biotec, 130-050-301). $CD19^+$ cells were labeled with antibodies to CD27 (BD Biosciences, clone M-T271), IgD (BD Biosciences, clone IA6-2), IgM (BD Biosciences, clone G20-27), IgG (BD Biosciences, clone G18-145) and IgA (DacoCytomation, F0188) for 15 min at room temperature in staining buffer (PBS with 0.5% BSA). Naive B cells ($CD19^+CD27^-IgD^+$) and memory B cells ($CD19^+CD27^+IgA^+$ or $CD19^+CD27^+IgG^+$) were obtained by FACS sorting on a FACSAria II (BD Biosciences).

Plasma cells, germinal center B cells and naive B cells were isolated from the tonsils of children undergoing tonsillectomy (ranging in age between 2 and 13 years), obtained from the Clínica Universidad de Navarra (Pamplona, Spain) or Clinique Mutualiste La Sagesse (Rennes, France). Tonsils were minced extensively; after Ficoll-Isopaque density centrifugation, enrichment of B cells was performed with the AutoMACS system either by positive selection of $CD19^+$ cells or using B Cell Isolation Kit II (Miltenyi Biotec). Tonsillar plasma cells ($CD20^{med}CD38^{hi}$), germinal center B cells ($CD20^{hi}CD38^{med}$) and naive B cells ($CD20^+CD23^+$) were separated by FACS sorting using antibodies from BD Biosciences (CD20, clone 2H7; CD38, clone HIT-2; CD23, clone M-L233). In part, naive B cells were also selected using a slightly different marker combination ($CD19^+CD27^-IgD^+$ or $IgD^+CD38^{low}CD27^-$). Germinal center B cells were also selected by the marker combination $IgD^-CD38^{hi}CD10^+CXCR4^+$ (CD38 and CD10 antibodies from Beckman Coulter, clones LS198-4-3 and ALB1, respectively; CXCR4 antibody from BD Biosciences, clone 51505). Naive B cells and germinal center B cells isolated with different markers constituted the same cell subpopulations, as evidenced by the fact that each subpopulation showed homogeneous DNA methylation and transcriptional profiles (**Supplementary Fig. 22**).

Bone marrow plasma cells were selected from healthy donors ranging from 20 to 30 years of age. After density gradient centrifugation, we performed selective depletion of $CD3^+$, $CD14^+$ and $CD15^+$ cells by immunomagnetic selection (Miltenyi Biotec) followed by flow cytometry cell sorting for $CD45^+CD138^+CD38^+$ cells (CD38 and CD45 antibodies from Miltenyi Biotec, clones HIT-2 and 5B1, respectively; CD138 antibody from BD Biosciences, clone 44F9) using a FACSAria II device.

The purity of each of the isolated B cell subpopulations exceeded 90% in all samples. DNA was extracted from purified samples using a Qiagen kit (QIAamp DNA Mini kit), following the manufacturer's instructions, and was quantified using a Nanodrop ND-100 spectrophotometer. DNA samples for WGBS, 450k array and BPS experiments were derived from individual donors with the exception of those from bone marrow plasma cells, which were pooled from four different donors. Total RNA was extracted with TRIzol (Invitrogen) following the manufacturer's recommendations. RNA quality was assessed with the Agilent 2100 Bioanalyzer (Agilent Technologies). The use of the samples analyzed in the present study was approved by the ethics committees of the Hospital Clinic de Barcelona, Hospital Universidad de Navarra and the University Hospital of Rennes as well as the University of California San Francisco Committee on Human Research.

Whole-genome bisulfite sequencing. We performed WGBS on two independent sets of biological replicates for six B cell differentiation stages. Briefly, genomic DNA (1–2 µg) was spiked with unmethylated λ DNA (5 ng of λ DNA per microgram of genomic DNA; Promega). DNA was shared by sonication to 50–500 bp in size using a Covaris E220 sonicator, and fragments of 150–300 bp were selected using AMPure XP beads (Agencourt Bioscience). Genomic DNA libraries were constructed using the Illumina TruSeq Sample Preparation kit following Illumina's standard protocol: end repair was performed on the DNA fragments, an adenine was added to the 3′ end of each fragment and Illumina TruSeq adaptors were ligated to both ends. After adaptor ligation, DNA was treated with sodium bisulfite using the EpiTexy Bisulfite kit (Qiagen), following the manufacturer's instructions for formalin-fixed, paraffin-embedded tissue samples. Two rounds of bisulfite conversion were performed to ensure a conversion rate of over 99%. Enrichment for adaptor-ligated DNA was carried out through seven PCR cycles using PfuTurboCx Hot-Start DNA polymerase (Stratagene). Library quality was monitored using the Agilent 2100 Bioanalyzer, and the concentration of viable sequencing fragments (molecules carrying adaptors at both ends) was estimated using quantitative PCR with the library quantification kit from Kapa Biosystems. Paired-end DNA sequencing (2×100 bp) was then performed using the Illumina HiSeq 2000 platform. The amounts of sequence reads and proportions of aligned reads are shown in **Supplementary Table 2**.

Read mapping and estimation of cytosine methylation levels. Read mapping was carried out using the GEM aligner (v1.242)[62] against a composite reference containing two copies of the human GRCh37 reference genome and two copies of the NCBI viral genome database (v35). For both the human and viral references, one copy had all cytosine bases replaced by thymine bases and the other had all guanine bases replaced by adenine bases. The names of the contigs in the combined reference FASTA file were modified by adding "C2T" or "G2A" to the end of the contig names depending on the conversion performed. Before mapping was performed, the original sequence for each read was stored. The first read from each pair then had all cytosine bases replaced by thymine bases, and the second read had all guanine bases replaced by adenine bases. Read mapping with GEM was performed, allowing up to four mismatches per read with respect to the reference. After read mapping, the original sequence for each read was restored.

Estimation of cytosine levels was carried out on read pairs where both members of the pair mapped to the same contig with consistent orientation and there was no other such configuration at the same or a smaller edit distance from the reference. After mapping, we restored the original read data in preparation for the inference of genotype and methylation status. We estimated genotype and DNA methylation status simultaneously using software developed at the Centro Nacional de Análisis Genómico, taking into account the observed bases, base quality scores and the strand origin of each read pair. For each genome position, we produced estimates of the most likely genotype and the methylation proportion (for genotypes containing a cytosine base on either strand). A Phred-scaled likelihood ratio for the confidence in the genotype call was estimated for the called genotype at each position. For each sample, CpG sites were selected where both bases were called as homozygous CC followed by GG with a Phred score of at least 20, corresponding to an estimated genotype error level of ≤1%. Sites with >500× coverage depth were excluded to avoid centromeric or telomeric repetitive regions. A common set of called CpG sites for all analyzed samples was generated, and all subsequent analyses used this common set.

Microarray-based DNA methylation analysis with 450k arrays. We used the EZ DNA Methylation kit (Zymo Research) for bisulfite conversion of 500 ng of genomic DNA. Bisulfite-converted DNA was hybridized to the HumanMethylation 450k BeadChip (Illumina), which covers 99% of RefSeq genes and 96% of CpG islands. The Infinium methylation assay was carried out as described previously[25,63]. Data from the HumanMethylation 450k array were analyzed in R using the minfi package[64] (version: 1.6.0), available through Bioconductor open source software. To exclude technical and biological biases that might produce false results in further analyses, we developed and optimized an analysis pipeline with several filters (i.e., removing CpGs with low detection *P* values, with sex-specific or individual-specific methylation, or overlapping with SNPs). Taking into account the different performance of Infinium I and Infinium II assays we used subset-quantile within-array

normalization (SWAN)[65], which corrects for technical differences between the Infinium I and Infinium II assay designs and produces a smoother overall β-value distribution.

Detection of non-CpG methylation and differential methylation analysis. Cytosines in a non-CpG context were defined as two adjacent nucleotides where the genotype of the first nucleotide was called with high confidence as homozygous C and the second nucleotide was called with high confidence as a genotype other than G. Non-CpG cytosines were called methylated if they had at least two non-converted reads, at least six reads informative for methylation status and a methylation probability greater than twice its standard deviation. The significance of the change in methylation levels between samples was assessed using the numbers of converted and non-converted reads in both samples with a χ^2 test or with a Fisher's exact test when the χ^2 approximation was not appropriate.

The difference in methylation levels between different stages of B cell differentiation was calculated using 16.1 million CpGs with methylation estimates in all 12 samples analyzed by WGBS. The normal approximation to the binomial was used to test for significant differences of individual CpGs between samples. As we sequenced two biological replicates per cell subpopulation, we defined consistent DNA methylation changes between two differentiation stages by one set of samples having a methylation difference above 0.25 and the second set of samples having a difference of at least 0.1 in the same direction (hyper- or hypomethylation). CpGs with dynamic methylation were defined as those with differential methylation in comparisons of adjacent stages and in comparison of HPCs and bone marrow plasma cells.

To calculate the overlap between dynamically methylated CpGs in B cells and those identified by Ziller *et al.*[1], we downloaded all the differentially methylated regions (DMRs) with dynamic CpGs from that study (Gene Expression Omnibus (GEO), GSE46644) and determined how many of the 4.93 million dynamic CpGs in B cells were located within DMRs identified by Ziller *et al.*

We also defined CpGs that showed variable methylation levels throughout B cell development using 450k microarray data for HPCs, pre-BI cells, pre-BII cells, immature B cells, naive B cells, tonsillar naive B cells, germinal center B cells, tonsillar plasma cells and bone marrow plasma cells. From the germinal center B cell stage on, B cell differentiation is branched into memory B cells or plasma cells, and memory B cells were therefore not included in this linear analysis. We performed pairwise comparisons between all these subsets of B cell differentiation. We defined as dynamic those CpGs that presented a mean DNA methylation difference above 0.25 in at least one comparison (false discovery rate (FDR) < 0.1, Wilcoxon test). Furthermore, we detected CpGs whose methylation showed a similar modulation pattern over the entire B cell differentiation process, and we grouped them into distinct modules. To allow for the identification of gradual changes (but consistent in different replicates) throughout the differentiation program, we applied a mean methylation difference between adjacent subpopulations of 0.1. The modules were placed into four groups depending on their methylation tendency during differentiation (decreased, increased, decreased then increased, and increased and then decreased). Differentially methylated sites between specific B cell tumor entities and their normal cellular counterpart were identified by a mean DNA methylation difference above 0.25 and FDR < 0.05.

Bisulfite pyrosequencing studies. The DNA methylation levels generated by WGBS and 450k microarray were validated by BPS. Briefly, 500 ng of genomic DNA was bisulfite converted using the EpiTect 96 Bisulfite kit or the EpiTect Plus Bisulfite Conversion kit (Qiagen) according to the manufacturer's instructions. PCR amplification of the bisulfite-treated DNA was performed using specific primers for each of the selected CpGs and non-CpGs (**Supplementary Table 3**). These primers were selected using PyroMark Assay Design software (Qiagen). BPS and DNA methylation data analysis were performed with the PyroMark Q96 ID pyrosequencer and PyroMark CpG software (Qiagen).

Genomic and functional annotation of CpG sites. Both WGBS and 450k microarray data were annotated using the UCSC Genome Browser database (hg19). For location of a site relative to a gene, we used these categories: TSS 1,500 (from 201 to 1,500 bp upstream of the TSS), TSS 200 (from 1 to 200 bp upstream of the TSS), 5′ UTR, first exon, exon (all exons excluding exon 1), intron, 3′ UTR and intergenic. Owing to the presence of alternative TSSs and regions containing more than one gene, some of the CpGs were assigned multiple annotations. For location of a site relative to a CGI, we used these groups: within CGI, in CGI shore (0–2 kb from the CGI edge), in CGI shelf (>2 kb to 4 kb from the CGI edge) and outside CGI.

We also annotated all CpG probes using a recent categorization of chromatin and transcriptional states from the lymphoblastoid B cell line GM12878 (ref. 28; ChromHMM track of the UCSC Genome Browser), which has a DNA methylome similar to memory B cells and plasma cells (**Supplementary Fig. 14**). Regions with chromatin states 1–3 (active promoter, weak promoter and poised promoter) defined as 'promoter regions', states 4–7 (strong enhancer and weak enhancer) 'enhancer regions', state 8 an 'insulator', state 9 a 'transcriptional transition', state 10 'transcriptional elongation', state 11 'weak transcription', state 12 'Polycomb-repressed regions' and state 13 'heterochromatin (nuclear lamina)'.

Replication timing in GM12878 data was obtained from the UW Repli-seq track of the UCSC Genome Browser. Replication timing values for all sites from the background was divided into three bins: early-, mid- and late-replicating regions. Only early- and late-replicating regions were used for the analysis.

Annotation of repeat elements was carried out on the basis of RepeatMasker Annotation, available at the UCSC Genome Browser.

B cell network analysis. We used the B cell–specific functional interaction network of Lefebvre *et al.*[29], containing 5,748 nodes (genes) and 64,600 unique edges (interactions) based on Entrez gene identifiers. We selected the 5,668 genes with dynamically methylated enhancers and mapped them to Entrez gene identifiers, resulting in 5,658 unique Entrez gene identifiers. Of these genes, 2,154 are contained in the B cell network and 1,993 are directly connected in the network by 11,741 edges. This subnetwork of 1,993 nodes and 11,741 edges was investigated further. We identified 9 communities in the subnetwork using Gephi[66] and Louvain's method[67].

Chromatin immunoprecipitation coupled with bisulfite sequencing. B cells were cross-linked with 1% formaldehyde for 8 min (room temperature) before FACS separation of memory B cells. Chromatin preparation and ChIP were performed according to the Blueprint histone ChIP protocol with an antibody to H3K27me3 (C15410195, Diagenode). For whole-genome bisulfite library construction and sequencing, the immunoprecipitated DNA (50 ng) was sheared on a Covaris E220 to a fragment size of 50–500 bp and size selected for 150- to 500-bp fragments using AMPure XP beads (Agencourt Bioscience). Unmethylated λ DNA (500 ng; Promega) was treated in parallel on a Covaris E220 and also size selected with AMPure XP beads to the same fragment sizes as for the DNA sample. The unmethylated λ DNA was spiked into fragmented and size-selected immunoprecipitated DNA (5 ng of λ DNA per 1 µg of DNA), and the TruSeq Sample Preparation kit (Illumina) was used to prepare the Illumina library by adding platform-specific adaptors. After adaptor ligation, 450 ng of fragmented and size-selected unmethylated λ DNA was added to the library. Two rounds of bisulfite conversion were performed to obtain >99% conversion, following the manufacturer's instructions for formalin-fixed, paraffin-embedded tissue samples (EpiTect Bisulfite kit, Qiagen). Adaptor-ligated DNA was enriched through ten cycles of PCR with the Kapa HiFi Uracil$^+$ polymerase (Kapa Biosystems). The library was run on a fraction of a lane of a HiSeq 2000 flow cell (to generate 35 million paired-end reads) with read length of 2 × 100 bp, according to standard Illumina operation procedures. Primary data analysis was carried out with the standard Illumina pipeline.

The sequence reads were passed through the same read mapping and genotype/methylation calling pipeline as the conventional WGBS samples. In addition, aligned reads were analyzed with the NucHunter package[68], to provide predictions of the positions of H3K27me3-modified nucleosomes. The average methylation of cytosines within 500 bp of each peak was calculated, taking into account the strand on which the cytosine was present. The same analysis was performed on WGBS data from the same cell type (memory B cells), to allow comparison of the enriched (ChIP-BS) and non-enriched

(WGBS) results. This analysis was repeated using only predicted nucleosome peaks that fell within predicted Polycomb-repressed regions in the Broad ChromHMM analysis of the ENCODE cell line GM12878.

Gene ontology analysis. The GOstat package[69] available through Bioconductor was used to determine the enrichment of individual ontology terms in the different methylation modules as compared to all genes analyzed on the 450k array. The top 20 most significant terms for each module ($P < 0.001$) are shown in **Supplementary Data Set 4**.

Analysis of transcription factor binding sites. TFBS information was obtained using ChIP-seq data from the ENCODE Project and available through the UCSC Genome Browser. A total of 79 TFBSs were used for the analysis. The relative enrichment of each TFBS was calculated in comparison to background. A Fisher's exact test was used to assign an odds ratio and *P* value to each comparison.

Gene expression analyses. RNA samples from HPCs, pre-BI cells, pre-BII cells, naive B cells, tonsillar naive B cells, germinal center B cells, memory B cells, tonsillar plasma cells and bone marrow plasma cells were hybridized to Affymetrix Human Genome U219 arrays according to Affymetrix standard protocols. Analysis of scanned images for each probe set of the array was performed with GeneChip Operating Software (GCOS, Affymetrix). Raw CEL files were processed, and signals were normalized with the robust multichip average (RMA) algorithm using R statistical software in conjunction with the affy library[70] available through Bioconductor. GeneChip Human Gene 1.0 ST array data for progenitor B cells were downloaded from GEO (GSE45461) and normalized using RMA (these data were only used for **Fig. 4b**,**c**).

To evaluate the variability in gene expression among B cell subpopulations, we calculated the standard deviation between cell subtypes for each of the Affymetrix tags. A global measure of the variability for a particular set of genes (within modules and chromatin states) was then calculated as the average of all the standard deviations. Differential expression between tonsillar naive B cells and naive B cells was calculated using the limma library available through Bioconductor, requiring fold change > 1 ($\log_2$) between two groups and adjusted $P < 0.05$.

Reverse transcription and quantitative PCR for DNMT expression was carried out as follows. cDNA was synthesized from 100 ng of RNA sample in 20 µl of reaction mix using oligo(dT) primers and SuperScript III enzyme according to the manufacturer's recommendations (Invitrogen). Primer sequences for the DNMTs were taken from Fang *et al.*[71], whereas the sequences for those amplifying the housekeeping gene *EEF2* are provided in **Supplementary Table 3**. PCR amplification was carried out with 1 µl of the reverse transcription sample diluted 1:2 using Power SYBR Green PCR Master Mix according to the manufacturer's recommendations (Applied Biosystems). PCR reactions were run in triplicate on a StepOne System (Applied Biosystems).

62. Marco-Sola, S., Sammeth, M., Guigo, R. & Ribeca, P. The GEM mapper: fast, accurate and versatile alignment by filtration. *Nat. Methods* **9**, 1185–1188 (2012).
63. Bibikova, M. *et al.* Genome-wide DNA methylation profiling using Infinium® assay. *Epigenomics* **1**, 177–200 (2009).
64. Aryee, M.J. *et al.* Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. *Bioinformatics* **30**, 1363–1369 (2014).
65. Maksimovic, J., Gordon, L. & Oshlack, A. SWAN: subset-quantile within array normalization for Illumina Infinium HumanMethylation450 BeadChips. *Genome Biol.* **13**, R44 (2012).
66. Bastian, M., Heymann, S. & Jacomy, M. in *Proc. 3rd Int. AAAI Conf. Weblogs and Social Media* 361–362 (2009).
67. Blondel, V.D., Guillaume, J.L., Lambiotte, R. & Lefebvre, E. Fast unfolding of communities in large networks. *J. Stat. Mech. Theory Exp.* **10**, P10008 (2008).
68. Mammana, A., Vingron, M. & Chung, H.R. Inferring nucleosome positions with their histone mark annotation from ChIP data. *Bioinformatics* **29**, 2547–2554 (2013).
69. Falcon, S. & Gentleman, R. Using GOstats to test gene lists for GO term association. *Bioinformatics* **23**, 257–258 (2007).
70. Gautier, L., Cope, L., Bolstad, B.M. & Irizarry, R.A. affy—analysis of Affymetrix GeneChip data at the probe level. *Bioinformatics* **20**, 307–315 (2004).
71. Fang, J. *et al.* Epigenetic changes mediated by microRNA miR29 activate cyclooxygenase 2 and λ-1 interferon production during viral infection. *J. Virol.* **86**, 1010–1020 (2012).

doi:10.1038/ng.3291

Manuscript 2

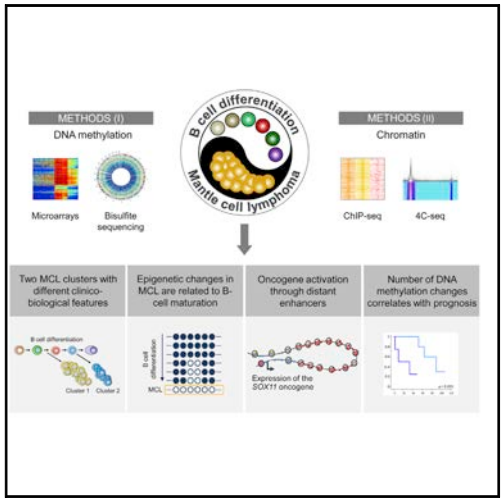
*Queirós AC, *Beekman R, *Vilarrasa-Blasi R, **Duran-Ferrer M**, Clot G, Merkel A, Raineri E, Russiñol N, Castellano G, Beà S, Navarro A, Kulis M, Verdaguer-Dot N, Jares P, Enjuanes A, Calasanz MJ, Bergmann A, Vater I, Salaverría I, van de Werken HJG, Wilson WH, Datta A, Flicek P, Royo R, Martens J, Giné E, Lopez-Guillermo A, Stunnenberg HG, Klapper W, Pott C, Heath S, Gut IG, Siebert R, Campo E, #Martín-Subero JI. Decoding the DNA Methylome of Mantle Cell Lymphoma in the Light of the Entire B Cell Lineage. Cancer Cell. 2016 Nov 14;30(5):806-821. doi:10.1016/j.ccell.2016.09.014. PMID: 27846393; PMCID: PMC5805090.

Article

Cancer Cell

Decoding the DNA Methylome of Mantle Cell Lymphoma in the Light of the Entire B Cell Lineage

Graphical Abstract



Authors

Ana C. Queirós, Renée Beekman, Roser Vilarrasa-Blasi, ..., Reiner Siebert, Elías Campo, José I. Martín-Subero

Correspondence

imartins@clinic.ub.es

In Brief

As part of the IHEC consortium, Queirós et al. perform an epigenomic analysis of mantle cell lymphoma, which reveals two major subtypes with distinct clinicobiological features and identifies distant enhancers as potential epigenetic drivers. Explore the Cell Press IHEC web portal www.cell.com/consortium/IHEC.

Highlights

- Epigenetic imprints of normal B cell subpopulations identify two subtypes of MCL
- Dynamically methylated regions in normal B cells acquire additional changes in MCL
- Identification of distant enhancers of oncogenes as potential epigenetic drivers
- Epigenetic heterogeneity in MCL is linked to genetic changes and clinical behavior

Queirós et al., 2016, Cancer Cell *30*, 806–821
November 14, 2016
http://dx.doi.org/10.1016/j.ccell.2016.09.014

CellPress

CellPress

Cancer Cell
Article

Decoding the DNA Methylome of Mantle Cell Lymphoma in the Light of the Entire B Cell Lineage

Ana C. Queirós,[1,19] Renée Beekman,[2,19] Roser Vilarrasa-Blasi,[1,19] Martí Duran-Ferrer,[1] Guillem Clot,[2] Angelika Merkel,[3] Emanuele Raineri,[3] Nuria Russiñol,[1,2] Giancarlo Castellano,[2] Sílvia Beà,[2] Alba Navarro,[2] Marta Kulis,[1] Núria Verdaguer-Dot,[1,2] Pedro Jares,[1,4] Anna Enjuanes,[4] María José Calasanz,[5] Anke Bergmann,[6,7] Inga Vater,[6] Itziar Salaverría,[2] Harmen J.G. van de Werken,[8,9,10] Wyndham H. Wilson,[11] Avik Datta,[12] Paul Flicek,[12] Romina Royo,[13] Joost Martens,[14] Eva Giné,[15] Armando Lopez-Guillermo,[15] Hendrik G. Stunnenberg,[14] Wolfram Klapper,[16] Christiane Pott,[17] Simon Heath,[3] Ivo G. Gut,[3] Reiner Siebert,[6] Elías Campo,[1,2,18] and José I. Martín-Subero[1,2,20,*]

[1]Departamento de Fundamentos Clínicos, Universitat de Barcelona, Barcelona 08036, Spain
[2]Institut d'Investigacions Biomèdiques August Pi i Sunyer (IDIBAPS), Barcelona 08036, Spain
[3]Centro Nacional de Análisis Genómico, Parc Cientific de Barcelona, Barcelona 08028, Spain
[4]Unidad de Genómica, IDIBAPS, Barcelona 08036, Spain
[5]Department of Genetics, University of Navarra, Pamplona 31008, Spain
[6]Institute of Human Genetics, Christian-Albrechts University, Kiel 24105, Germany
[7]Department of Pediatrics, Christian-Albrechts University & University Hospital Schleswig-Holstein, Kiel 24105, Germany
[8]Department of Cell Biology
[9]Cancer Computational Biology Center
[10]Department of Urology
Erasmus MC, Rotterdam 3015 CN, the Netherlands
[11]Lymphoid Malignancies Branch, Center for Cancer Research, National Cancer Institute, Bethesda, MD 20892, USA
[12]European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Wellcome Trust Genome Campus, Hinxton CB10 1SD, UK
[13]Joint Program on Computational Biology, Barcelona Supercomputing Center (BSC) and Institute of Research in Biomedicine (IRB), Barcelona Science Park, Barcelona 08034, Spain
[14]Molecular Biology, NCMLS, FNWI, Radboud University, Nijmegen 6500 HB, the Netherlands
[15]Servicio de Hematología, Hospital Clínic, IDIBAPS, Barcelona 08036, Spain
[16]Hematopathology Section and Lymph Node Registry, Christian-Albrecht University, Kiel 24105, Germany
[17]Second Medical Department, University Hospital Schleswig-Holstein, Kiel 24116, Germany
[18]Unidad de Hematopatología, Servicio de Anatomía Patológica, Hospital Clínic, Barcelona 08036, Spain
[19]Co-first author
[20]Lead Contact
*Correspondence: imartins@clinic.ub.es
http://dx.doi.org/10.1016/j.ccell.2016.09.014

SUMMARY

We analyzed the in silico purified DNA methylation signatures of 82 mantle cell lymphomas (MCL) in comparison with cell subpopulations spanning the entire B cell lineage. We identified two MCL subgroups, respectively carrying epigenetic imprints of germinal-center-inexperienced and germinal-center-experienced B cells, and we found that DNA methylation profiles during lymphomagenesis are largely influenced by the methylation dynamics in normal B cells. An integrative epigenomic approach revealed 10,504 differentially methylated regions in regulatory elements marked by H3K27ac in MCL primary cases, including a distant enhancer showing de novo looping to the MCL oncogene *SOX11*. Finally, we observed that the magnitude of DNA methylation changes per case is highly variable and serves as an independent prognostic factor for MCL outcome.

Significance

Recent studies on the DNA methylome of cancer cells are reshaping our perception on the pathogenic role of this epigenetic mark, including reports that suggest a major link between the dynamic DNA methylation landscape of normal cell differentiation and neoplastic transformation. Here, we performed a detailed epigenomic analysis of mantle cell lymphoma (MCL), a heterogeneous B cell tumor, in the context of the DNA methylome of the entire normal B cell maturation program. Our results provide insights into the cellular origin, pathogenetic mechanisms, and clinical behavior of MCL, and we highlight that integrative analyses of the DNA methylome, histone modifications and three-dimensional interactions in cancer cells can identify potential epigenetic drivers at distant regulatory elements of key oncogenes.

INTRODUCTION

The existence of alterations in the DNA methylome of cancer cells has been known since the early 1980s (Feinberg and Vogelstein, 1983; Gama-Sosa et al., 1983). Despite the widely reported role of DNA methylation in cancer (Baylin and Jones, 2011; Esteller, 2008), the analyses of whole DNA methylomes are now questioning the accepted view of DNA methylation as a major player in gene deregulation. These analyses are revealing that the roles of this epigenetic mark are more variable and context dependent than previously appreciated (Jones, 2012; Kulis et al., 2013) and that a large fraction of the DNA methylation changes in cancer do not seem to have any apparent functional effect (Agirre et al., 2015; Keshet et al., 2006; Ziller et al., 2013). So far, cancer epigenomics studies have detected tumor-specific changes by comparing tumor cells with their normal counterparts, or by comparing longitudinal tumor samples. However, it is becoming increasingly evident that regions with dynamic DNA methylation levels in normal cells seem to be prone to be altered upon neoplastic transformation (Feinberg, 2014; Hansen et al., 2011; Kulis et al., 2015). Thus, a detailed analysis of a specific tumor type in the context of the entire differentiation program of its normal cellular counterpart will reveal new insights into the role of DNA methylation in tumorigenesis.

Mantle cell lymphoma (MCL) is a B cell lymphoma that shows a broad spectrum of clinical behaviors and biological features (Jares et al., 2012). Despite the heterogeneity, the unifying factor in MCL is the t(11;14) (q13;q32) translocation leading to cyclin D1 gene (*CCND1*) deregulation, which is considered to be a primary driver event in this disease (Jares et al., 2007). Most MCLs have an aggressive clinical behavior with poor survival rates. However, some cases classified as leukemic non-nodal MCLs show a rather indolent clinical course even in the long-term absence of chemotherapy (Royo et al., 2012). Aggressive cases are highly proliferative and seem to be associated with a lack or low levels of somatic mutations in the *IGHV* locus and de novo expression of SOX11, which is not expressed in normal B cells. In contrast, leukemic non-nodal MCLs show a very low proliferation index, have high levels of somatic mutations in the *IGHV* locus, and lack SOX11 expression. However, some of these SOX11-negative MCLs can acquire oncogenic mutations and progress toward a fatal clinical outcome (Jares et al., 2012).

The DNA methylome of MCL remains largely unknown, as it has only been analyzed in promoter regions (Enjuanes et al., 2013; Halldorsdottir et al., 2012; Leshchenko et al., 2010; Rahmatpanah et al., 2006). To obtain deeper insights into MCL epigenetics, we have applied an analytic strategy to deconstruct the DNA methylome of MCL in the light of the complete normal B cell differentiation program (Kulis et al., 2015).

RESULTS

Deconvolution and In Silico Purification of MCL DNA Methylation Signatures

We generated genome-wide DNA methylation profiles of 82 MCL samples using the HumanMethylation450 BeadChip (Illumina) (Bibikova et al., 2011). Biological and clinical information of the analyzed cases is shown in Table S1. As normal controls, we used 67 samples from ten different cell subpopulations spanning the entire B cell lineage (Kulis et al., 2015). We considered two potential confounding variables that may affect our epigenomic analyses, i.e., the biological origin of the samples (lymph node versus peripheral blood) and the tumor cell content. We did not identify any consistent differential methylation pattern between lymph node and peripheral blood samples (data not shown). However, despite the generally high tumor cell content of the selected MCL samples (median, 89%; range, 56%–100%; Table S1), purity affected the DNA methylation analyses (Figure S1). Therefore, we developed a strategy to deconvolute the DNA methylation signal of mixed subpopulations and to isolate in silico the DNA methylation levels of the tumor cells (Figure 1A). To that end, we adapted a published algorithm (Houseman et al., 2012; Jaffe and Irizarry, 2014) to estimate the fractions of six different hematopoietic cell types (Reinius et al., 2012) in our tumor samples (Figure 1B). The normal B cell fraction in MCL samples is estimated to be very low (0%–0.3%) (Saba et al., 2016), therefore the total B cell fraction was taken as a measure for the tumor fraction. Using the adapted algorithm, we calculated the proportion of each cell type in our samples. We validated the approach by comparing the in silico estimated tumor B cell fraction with the sample purity measured by flow cytometry in 32 MCL samples (Pearson $r = 0.947$; Figure 1C). Finally, we used the DNA methylation estimates of the normal non-B cell subtypes together with their respective proportions to extract the DNA methylation signature derived from the tumor B cells in each MCL sample (Figure 1D). These pure DNA methylation estimates of the tumor fraction were used for all downstream analyses.

Genome-wide DNA Methylation Analysis Reveals Two Major MCL Subgroups with Distinct Clinicobiological Features

We performed an unsupervised principal component analysis (PCA) of DNA methylation data from normal B cell subpopulations and MCL samples (Figure 2A). The two first components ordered normal B cells according to their maturation stage, mainly separating germinal-center-inexperienced B cells (uncommitted precursors, pre-B cells, and naive B cells) from germinal-center-experienced B cells (germinal-center B cells, memory B cells, and plasma cells). Principal component 1 showed that all MCLs are globally more similar to germinal-center-experienced B cells (i.e., antigen experienced). In contrast, principal component 2 split MCLs into two subgroups: cluster 1 (C1) ($n = 62$) and cluster 2 (C2) ($n = 20$), which respectively showed a DNA methylation pattern more similar to germinal-center-inexperienced B cells and germinal-center-experienced B cells. These subgroups showed significant clinicobiological differences ($p < 0.001$) in, for example, *IGHV* mutation levels, SOX11 expression, number of copy number alterations, nodal presentation, and requirement of treatment at diagnosis (Figure 2B). Furthermore, C1 cases showed a significantly worse overall survival than C2 cases ($p = 0.026$) (Figure S2A).

Next, we compared C1 and C2 MCLs, and identified 13,691 differentially methylated CpGs (Figure 2C). Most CpGs hypomethylated in C2 MCLs linked C1 cases to germinal-center-inexperienced cells and C2 cases to germinal-center-experienced B cells (Figure 2C), further supporting the concept shown in the

CellPress

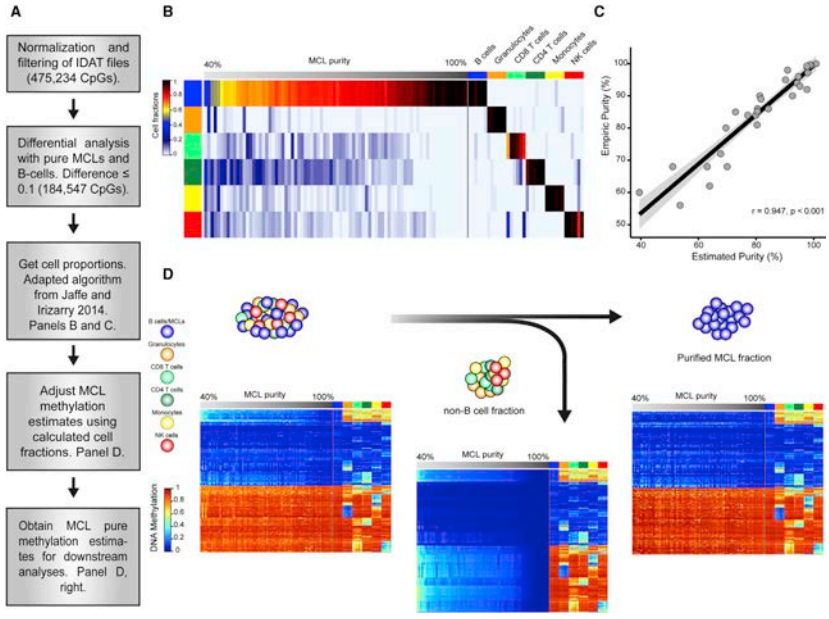


Figure 1. Deconvolution of DNA Methylation Data and In Silico Purification of MCL Methylation Estimates

(A) Work flow of the deconvolution process in MCL samples.

(B) Estimation of the proportion of hematopoietic cell subpopulations in MCL samples and in sorted B cells, $CD8^+$ T cells, $CD4^+$ T cells, natural killer (NK) cells, monocytes, and granulocytes. Sorted cell subpopulations (right part of the heatmap) are correctly predicted and MCLs show a gradient from lower to higher proportion of B cells (left part of the heatmap).

(C) The proportion of B cells in MCL samples as detected by flow cytometry and by the in silico prediction are highly correlated.

(D) Heatmaps of the CpGs representative of each cell type (n = 580) showing the initial methylation estimates from the MCL samples (left), the extraction of the DNA methylation signature from contaminating non-B cells (middle), and the final in silico purification of the DNA methylation estimates from MCL cells (right).

See also Figure S1 and Table S1.

second component of the PCA analysis (Figure 2A). In contrast, hypomethylation in C1 was predominantly a de novo event targeting regions that are highly methylated both in C2 MCLs and normal B cells (Figure 2C). These regions frequently targeted CpG island shores and gene bodies (Figure 2D). Furthermore, we performed chromatin immunoprecipitation sequencing (ChIP-seq) with six histone marks and generated chromatin states from sorted naive and memory B cells from healthy donors; we observed that hypomethylated regions in C1 MCLs were enriched for enhancers and transcribed regions (Figure 2E). Naive and memory B cells have been previously suggested as potential cells of origin of *IGHV* unmutated and mutated MCLs, respectively (Navarro et al., 2012). Therefore, in this study we used them as normal counterparts of C1 MCLs and C2 MCLs, respectively. Interestingly, the genes affected by hypomethylation in C1 MCLs were significantly enriched (adjusted $p < 0.05$) in several pathways (Table S2), such as NOTCH signaling (Figure S2B), which has been previously linked to MCL pathogenesis of the *IGHV* unmutated/SOX11-positive subgroup (i.e., C1) (Bea et al., 2013; Kridel et al., 2012).

Comparing MCL Groups with Their Normal Cell Counterparts Reveals a Major Epigenetic Link with Normal B Cell Differentiation

Next, we sought to detect epigenetic differences in the MCL subgroups compared with their respective putative normal counterparts. We observed 60,622 differentially methylated CpGs in C1 MCLs (78% hypomethylated) in comparison with naive B cells and 5,469 CpGs in C2 MCLs (84% hypomethylated) in comparison with memory B cells (Figure 3A). Interestingly, we found

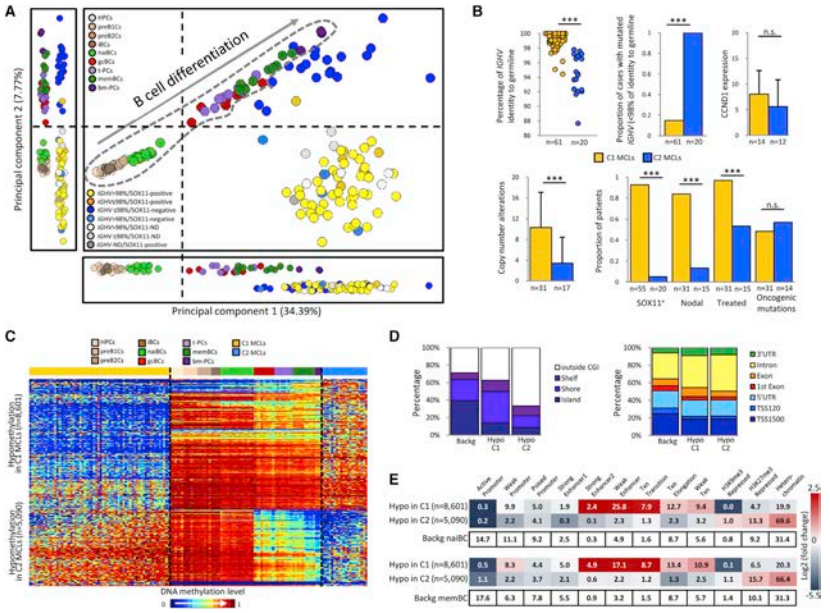


Figure 2. Identification of Two MCL Subgroups Based on DNA Methylation Profiling

(A) Unsupervised PCA of 82 MCLs and 67 normal B cell subpopulations using the adjusted methylation values of all CpGs analyzed with the 450K array. The two main principal components are shown together in a two-dimensional plot and separately. Vertical and horizontal dotted lines point to the cut-off value separating germinal-center-inexperienced and -experienced B cells. Normal B cells are surrounded by a dotted gray line.

(B) Comparison of biological and clinical features between the two epigenetic subgroups (i.e., C1 and C2). The presence of oncogenic mutations is defined as having a mutation in at least one of the following genes: *BIRC3*, *MEF2B*, *NOTCH2*, *TLR2*, *TP53*, and *WHSC1*. Data show means ± SD. ***p < 0.001; n.s., not significant (Fisher's exact test or t test for independent samples).

(C) Heatmap of the CpGs differentially methylated in C1 compared with C2.

(D) Location of the hypo- and hypermethylated CpGs between C1 and C2 MCLs in the context of CpG islands (CGI) and gene-related regions.

(E) Chromatin states of naive (upper panel) and memory (lower panel) B cells of the differentially methylated CpGs between C1 and C2 MCLs. The numbers inside each cell point to the percentage of CpGs belonging to a particular chromatin state. The differentially methylated CpGs annotated in (D) and (E) are the same as those shown in (C). Backg, background; bm-PCs, plasma cells from bone marrow; C1 MCLs, germinal-center-inexperienced MCLs; C2 MCLs, germinal-center-experienced MCLs; gcBCs, germinal-center B cells; HPCs, hematopoietic progenitor cells; Hyper, hypermethylation; Hypo, hypomethylation; iBCs, immature B cells; memBCs, memory B cells from peripheral blood; naiBCs, naive B cells from peripheral blood; preB1Cs, pre-BI cells; preB2Cs, pre-BII cells; t-PCs, plasma cells from tonsil; TSS, transcriptional start site; UTR, untranslated region.

See also Figure S2 and Table S2.

that 61%–79% of these CpGs overlapped with those previously described to show variable DNA methylation levels during normal B cell differentiation (Kulis et al., 2015) (Figures 3B and 3C). This finding suggests that only a fraction of the DNA methylation changes in MCLs compared with their normal counterparts is unrelated to normal B cell differentiation and thus, strictly tumor specific. Those CpGs dynamically methylated both in MCL and B cell differentiation (from now on called B cell-related CpGs) and those exclusively changing in MCL (from now on called B cell-independent CpGs) were in part enriched in different chromatin states defined in naive and memory B cells (Figure 3D). Overall, hypomethylation in MCL in both the B cell-related and independent fractions was enriched for enhancer elements. On the contrary, B cell-related hypermethylated CpGs in MCL were located both in H3K27me3-repressed and poised promoters, whereas those in the B cell-independent fraction were mostly associated with poised promoters (Figure 3D). To identify chromatin state transitions in relationship with DNA methylation changes, we generated ChIP-seq profiles and chromatin states from two MCL cases representative for C1

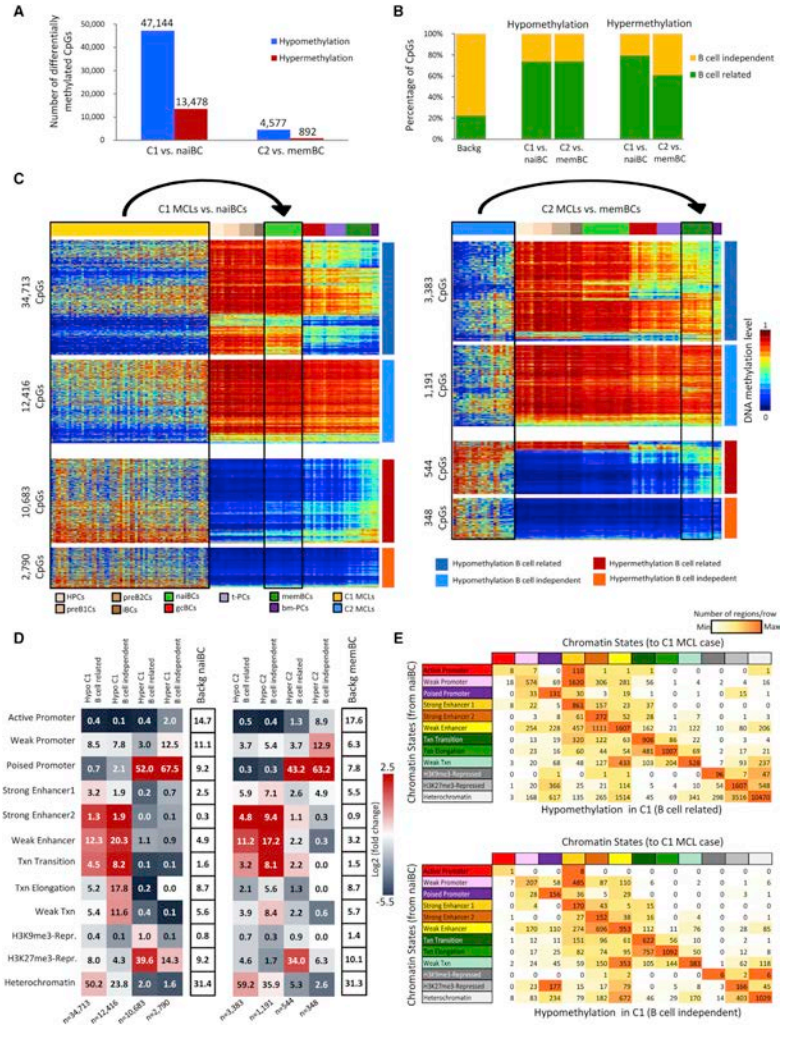


Figure 3. DNA Methylation of MCL Subgroups versus Their Respective Normal B Cell Counterpart

(A) Number of differentially methylated CpGs between C1 and naiBCs, and between C2 and memBCs.

(B) Percentage of B cell-related and B cell-independent CpGs differentially methylated in each comparison.

(legend continued on next page)

and C2, and compared them with those from naive and memory B cells, respectively (Figures 3E, S3A, and S3B). Overall, 56% of the regions did not seem to change their chromatin state in MCL upon DNA methylation alteration. However, we observed that repressed regions losing DNA methylation tend to change toward chromatin states related to activating histone modifications (especially H3K4me3 in poised promoters and H3K4me1 in weak enhancers) (Figure 3E); this phenomenon is more prominent in the B cell-independent fraction than in the B cell-related fraction in C1 (52% versus 18%, $p < 0.001$). In the case of hypermethylated regions in C1 MCLs, we observed that active and weak promoters in naive B cells turn into poised promoters in MCL and that poised promoters turn into H3K27me3-repressed regions (Figure S3A).

Individual Epigenetic Heterogeneity in MCL

The data presented in Figures 2 and 3 suggest that both MCL groups are epigenetically heterogeneous. Based on these observations, we have applied a second analytic strategy to tackle individual epigenetic variation of MCL cases in the context of the entire B cell maturation program. We compared the DNA methylome of each individual MCL case with the hematopoietic progenitor cells (HPCs) (using as cut-off an absolute difference of methylation values of at least 0.25). This seemingly unorthodox approach has the advantage that it uses a fixed reference point for B cell neoplasms with different normal counterparts, and allows us not only to precisely dissect but also to compare the DNA methylation modulation of each individual MCL sample from the moment of B cell commitment up to and beyond its cell of origin. We observed that the total number of changes per case is highly variable both in C1 and C2 MCLs (ranging from 62,888 to 143,925 CpGs) (Figure 4A). Furthermore, we identified that the DNA methylation levels of the MCLs correlate less among each other than within normal B cells, showing that the inter-sample heterogeneity is much higher in MCLs than in normal B cells (Figure 4B). In addition, we saw that 318,659 unique CpGs (98% of the 106,552 B cell-related and 53% of the 368,442 B cell-independent CpGs measured by the 450K array) showed a DNA methylation change compared with HPCs in at least one MCL case, suggesting that a large fraction of the human methylome can be modulated in normal and neoplastic B cells.

Next, to identify regions that may play a role in MCL development, we sought to identify B cell-independent CpGs with recurrent differential methylation in C1 and C2 MCLs. We detected that the majority of the differentially methylated sites between MCL and HPC were present in one or few MCLs, and that highly recurrent changes were rare events (Figure 4C). Furthermore, the relative proportion of differentially methylated regions marked by particular chromatin states (as defined in primary MCL cases), such as heterochromatin and enhancers, was related to the level of recurrence of the DNA methylation changes (Figures 4D, S4A, and S4B). These findings suggest that most B cell-independent changes in individual MCLs seem to target non-functional regions (i.e., heterochromatin) while commonly altered CpGs, although rare, target regulatory elements (i.e., enhancers).

An additional interesting aspect of this analysis of individual variation was that the number of B cell-related and B cell-independent differentially methylated CpGs per MCL case were linearly related (Pearson $r = 0.82$ and 0.91 for C1 and C2 MCLs, respectively; $p < 0.001$) (Figure 4E). This association suggests that the mechanisms underlying differentially methylation in B cell-related and B cell-independent CpGs are shared, even though different cases show different degrees of epigenetic changes. In addition, in C1 MCLs, we detected 6,245 CpGs with an inverse correlation between their DNA methylation levels and the percentage of *IGHV* somatic hypermutation (SHM) (Figures S4C–S4F), a phenomenon not observed in C2 cases. The fact that some C1 MCLs concurrently show some degree of SHM and DNA demethylation suggests that C1 MCLs may be derived from germinal-center-inexperienced B cells at different maturation stages, ranging from those lacking SHM to those showing low but variable degrees of SHM, which correlate with epigenetic changes (Kolar et al., 2007; Sims et al., 2005).

Deep Characterization of the MCL Methylome by Whole-Genome Bisulfite Sequencing

We sequenced the entire DNA methylome of two highly pure (95% and 99% tumor cells) representative MCLs previously analyzed by 450K microarrays (one from MCL C1 and one from MCL C2; Figures S5A–S5C) at a single base pair resolution (48× mean coverage; Table S3), and we analyzed them in the context of the DNA methylome of the B cell lineage (Kulis et al., 2015) (Figures 5A, S5D, and S5E). The methylation estimates obtained by the two methods were highly comparable (Pearson $r = 0.97$ for both cases; Figure S5C). We compared each MCL with HPCs as fixed reference, and we defined both differentially methylated CpGs (DMCs) and differentially methylated regions (DMRs). Determining DMRs increased the detection of regulatory regions compared with detecting DMCs, and therefore we continued our analyses using the DMR strategy (Figures S5F–S5I). Subsequently, we split the CpGs within DMRs into B cell-related and B cell-independent CpGs, and we observed that 55%–92% overlapped with those modulated during normal B cell differentiation (Figure 5B). Intriguingly, most DMRs in MCL either contained only B cell-related or a mixture of both B cell-related and B cell-independent CpGs, and few were exclusively B cell independent (Figures 5C–5E). More specifically, in the C1 MCL case, only 9.3% of the hypomethylated and 5.6% of the hypermethylated DMRs were B cell-independent, and these numbers dropped to 1% and 1.2%, respectively, in the C2 MCL case (Figure 5D). This analysis suggests that those

(C) Heatmaps of differentially methylated CpGs in C1 MCLs compared with naiBCs (left) and in C2 MCLs compared with memBCs (right) in the context of normal B cell differentiation.
(D) Chromatin states in naiBCs and memBCs of the differentially methylated CpGs between C1 and naiBCs (left panel), and between C2 and memBCs (right panel), respectively. The numbers inside each cell point to the percentage of CpGs belonging to a particular chromatin state.
(E) Transition of the chromatin states from naiBCs to a C1 MCL case in the B cell-related and B cell-independent hypomethylated CpGs. The numbers inside each cell point to the total number of CpGs in each transition.
See also Figure S3.

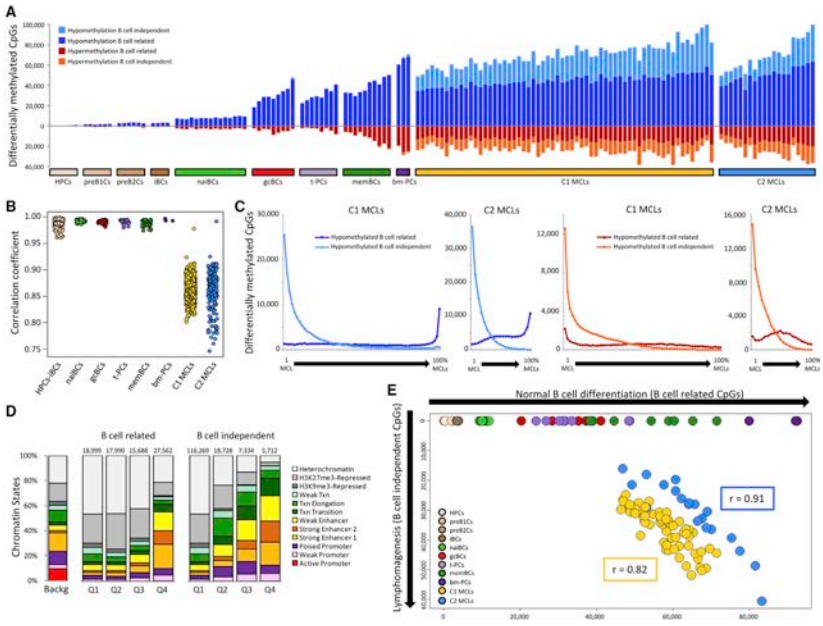


Figure 4. Association between B Cell-Related and B Cell-Independent DNA Methylation Changes in MCL

(A) Number of differentially methylated CpGs for each individual normal B cell subpopulation and MCL compared with HPCs.

(B) Correlation coefficient among samples of the different groups.

(C) Number of B cell-related and B cell-independent differentially methylated CpGs based on their level of recurrence in C1 (1st and 3rd panel) and C2 (2nd and 4th panel).

(D) Chromatin states, defined in an MCL primary case representative of C1 cases, of the hypomethylated CpGs between C1 and naiBCs divided into quartiles based on their level of recurrence. Q1, recurrent in 0%–25% of patients; Q2, recurrent in 25%–50% of patients; Q3, recurrent in 50%–75% of patients; Q4, recurrent in 75%–100% of patients.

(E) Scatterplot showing the number of B cell-related (x axis) and B cell-independent (y axis) CpGs differentially methylated in individual MCLs and normal B cells compared with HPCs.

See also Figure S4.

regions prone to acquire differential methylation during normal B cell differentiation seem to be predisposed to be further altered in the context of malignant transformation, and that regions with pure tumor-specific DMRs seem to be a rare phenomenon.

Identification of Potential Epigenetic Drivers in MCL and Detection of Distant *SOX11* Enhancers

Next, we aimed to study whether DMRs between C1 and C2 MCLs can lead to the detection of potential functional regulatory regions that are differentially active in these two groups. By comparing them, we detected 26,603 DMRs hypomethylated in C1 and 4,457 DMRs hypomethylated in C2. Approximately 60% of these DMRs contained a mixed pattern of B cell-related and B cell-independent CpGs (Figure 6A). Subsequently, we generated ChIP-seq profiles of the same MCL cases studied by whole-genome bisulfite sequencing (WGBS) and overlapped the detected DMRs with the genomic regions simultaneously containing H3K27ac, which marks active regulatory elements (Heintzman et al., 2009). We observed that hypomethylated DMRs in the C1 MCL case had a substantial overlap (39%) with H3K27ac peaks, which were predominantly present either in the MCL C1 case only or in both MCL cases (Figures 6B and 6C, and Table S4).

We then analyzed the chromatin architecture of the DMRs within H3K27ac peaks in further detail by taking into account H3K4me1, mostly marking enhancers, and H3K4me3, marking promoters (Figures 6D–6F). The hypomethylated DMRs in the C2 MCL case that are located within H3K27ac peaks (n = 118,

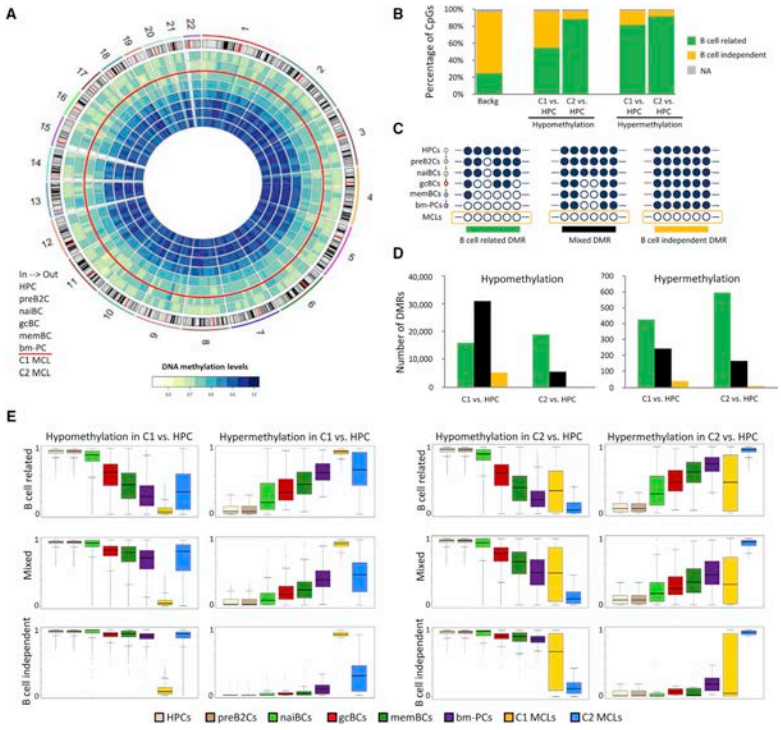


Figure 5. Analysis of the MCL Methylome by WGBS

(A) Circular representation of the DNA methylation levels for HPC, preB2C, naiBC, gcBC, memBC, and bm-PC, as well as two MCLs representative for C1 and C2, respectively. CpG methylation levels are averaged over 10 Mb genomic windows.

(B) Percentage of B cell-related and B cell-independent CpGs differentially methylated in C1 MCL and C2 MCL versus HPC.

(C) Graphical representation of the different DMR types: DMRs with only B cell-related CpGs are defined as B cell-related DMRs (left), DMRs containing both B cell-related and B cell-independent CpGs are defined as mixed DMRs (middle), and DMRs with only B cell-independent CpGs are defined as B cell-independent DMRs (right). Filled and empty circles represent methylated and unmethylated CpGs, respectively.

(D) Number of B cell-related, mixed, and B cell-independent DMRs between C1 versus HPC and between C2 versus HPC.

(E) Distribution of DNA methylation levels for the different DMRs types defined between C1 MCL and HPC and between C2 MCL and HPC. Boxplots show upper and lower quartiles and the median, and whiskers represent minimum and maximum, with outer points indicating outliers.

See also Figure S5 and Table S3.

2.6%) in the corresponding MCL case but not in the C1 MCL case, normal naive or memory B cells, showed simultaneous presence of H3K4me1 and H3K4me3 (Figure 6D), suggesting that these regions represent de novo active promoters. The hypomethylated DMRs (n = 4,452, 16.7%) in the C1 MCL case within H3K27ac peaks only in the corresponding MCL case and not in MCL C2, normal naive or memory B cells, showed enrichment for H3K4me1 (Figure 6E), pointing toward de novo activation of enhancers at these regions. Furthermore, the DMRs within H3K27ac peaks appeared to be significantly enriched ($p < 0.001$) in mixed and B cell-independent DMRs (Figures 6D–6F). Similar results were obtained analyzing the overlap between DMRs and super enhancers (Figure S6 and Table S5). Overall, these results show that DMRs between C1 and C2 MCLs may point toward differential active enhancers and promoters in these samples, especially when they contain

CellPress

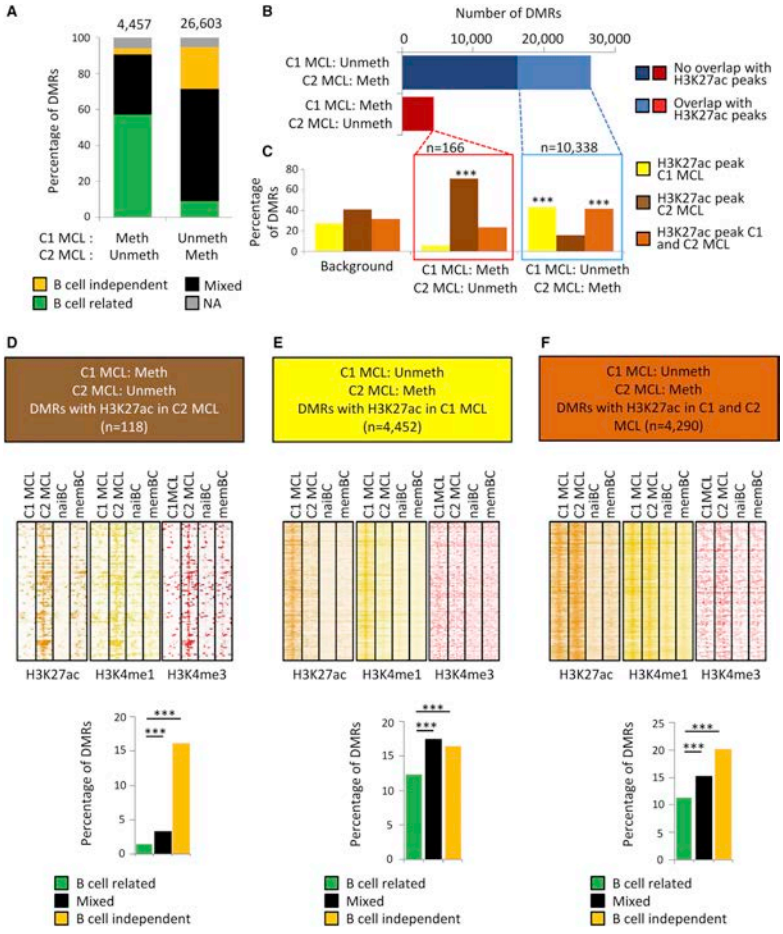


Figure 6. Integrative Analysis of Differentially Methylated Regions and Histone Modifications

(A) Distribution of DMRs defined by WGBS between the MCL cases representative of C1 (SOX11-positive) and C2 (SOX11-negative) into three different DMR types (B cell-dependent, B cell-independent, or mixed DMRs; NA, non-assigned).

(B) Number of DMRs between the C1 and C2 MCL cases and their overlap with H3K27ac peaks in these MCL cases.

(C) Distribution of the DMRs showing an overlap with H3K27ac peaks in the C1 MCL case only, the C2 MCL case only or in both cases. The background represents all H3K27ac peaks in the C1 and C2 MCL case, and shows which percentage is unique for these cases (yellow and dark brown) and which percentage overlaps (light brown). ***$p < 0.001$ (Fisher's test).

(D–F) Heatmaps showing the read density of H3K27ac, H3K4me1, and H3K4me3 ChIP-seq in the C1 MCL case, C2 MCL case, naive B cells (NBC), and memory B cells (MBC) at selected DMRs (±10 kb). Only the DMRs showing significant differences versus the background in (C) were used for these heatmaps,

(legend continued on next page)

CpGs that change only in MCL (i.e., mixed/B cell-independent DMRs).

One striking example is a cluster of mixed DMRs hypomethylated in the C1 MCL case overlapping with an enhancer region located 624–653 kb downstream of *SOX11* only in the SOX11-expressing MCL C1 case (Figure 7A). A set of 4C-seq analyses (Simonis et al., 2007; van de Werken et al., 2012) showed that this region presents high contact frequencies with the *SOX11* gene in three-dimensional (3D) space in the representative C1 primary MCL case and three SOX11-positive MCL cell lines but not in the C2 MCL case, in the SOX11-negative MCL cell line JVM-2, or in normal naive and memory B cells (Figures 7A and 7B). To investigate whether the association between DNA hypomethylation of this distant enhancer and the expression of SOX11 is a recurrent phenomenon in MCL primary cases, we analyzed the DNA methylation status of this region by bisulfite pyrosequencing in additional primary SOX11-positive (n = 12) and SOX11-negative MCL cases (n = 10). In this way, we confirmed that the identified regulatory region is de novo demethylated in SOX11-positive (average methylation level 14%–21%) compared with SOX11-negative cases (average methylation level 63%–85%, $p < 0.01$) or naive B cells (average methylation level 79%–91%) (Figure 7C and Table S6). However, whether this demethylation is a cause or a consequence of the enhancer activation and SOX11 expression remains to be elucidated. These data suggest a model in which aberrant SOX11 expression in MCL is associated with a de novo activation of a distant enhancer element that interacts with the *SOX11* locus in 3D space (Figure 7D).

Link among Epigenetic Burden, Genetic Changes, and Clinical Outcome of MCL Patients

In addition to the significant survival difference between C1 and C2 MCLs (Figure S2A), we postulated that the epigenetic burden (i.e., number of differentially methylated sites regardless of their relationship to normal B cells) may also be associated with clinical behavior. Indeed, in both MCL subgroups, we found that the number of DNA methylation changes compared with HPCs showed a significant linear association with the clinical outcome, approximately doubling the risk of death with each 10,000 methylation changes (Figures 8A and S7A). Beyond this quantitative association, we also calculated the threshold of DNA methylation changes that maximizes the difference in clinical outcome between two subsets of patients (Figures 8B and 8C as well Figures S7B–S7D). Furthermore, we compared DNA methylation changes with the presence of mutations using a set of six recurrent driver genes in MCL (Bea et al., 2013). We observed that cases with gene mutations in C2 MCLs, but not C1 MCLs, displayed a significantly higher number of CpG methylation changes (Figures 8D and 8E). To determine whether these observations can be linked to cell proliferation, we calculated the proliferation signature in 25 of our MCL cases (Navarro et al., 2012). As expected, MCL C1 cases are in general more proliferative than C2 cases (Figure S7E), but the proliferation signature was positively correlated with the number of epigenetic changes only in C2 MCLs (Figure S7F). Finally, we performed a multivariate Cox regression model with six variables related to MCL prognosis (Supplemental Experimental Procedures) and identified that the number of DNA methylation changes was the strongest independent prognostic factor in our MCL series ($p = 1.4 \times 10^{-5}$) followed by *IGHV* identity levels ($p = 0.0015$) and age ($p = 0.0019$) (Figure 8F). These data suggest that patients with more epigenetic changes have a worse clinical outcome and that, in C2 MCLs, this correlates with the acquisition of genetic changes and increased cell proliferation.

DISCUSSION

Recent reports using unbiased genome-wide approaches are reshaping our perception of the role of DNA methylation in cancer (Agirre et al., 2015; Berman et al., 2012; Hansen et al., 2011; Jones, 2012; Kulis et al., 2012; Ziller et al., 2013). Here, we have analyzed the DNA methylome of MCL, a heterogeneous B cell neoplasm, and decoded its clinicobiological impact in the context of the DNA methylome of the entire B cell lineage. This analytic strategy has allowed us to obtain insights into not only the pathogenesis and clinical behavior of MCL but also the general role of DNA methylation and its significance in cancer. An important aspect of our study was the initial deconvolution of the methylation estimates and in silico extraction of the methylation levels of tumor cells. Thus, the results obtained were not influenced by the composition of non-tumoral cells within the MCL samples. We believe that this strategy can be highly valuable for other epigenetic studies in which purified tumor cells cannot be obtained.

Our results indicate a major link between the dynamic DNA methylome during B cell maturation and MCL tumorigenesis from various perspectives. From the biological point of view, our findings put together two previous observations in MCL. First, most MCLs are derived from antigen-experienced B lymphocytes (Hadzidimitriou et al., 2011; Xochelli et al., 2015), which is reflected by the fact that all MCLs in our study have a DNA methylation profile more similar to antigen-experienced cells. Second, MCLs with unmutated and mutated *IGHV* may actually reflect a different cellular origin (Navarro et al., 2012). In the case of unmutated MCLs (C1), they retain an imprint of B cells preceding the germinal center, and its cellular origin may range from naive B cells lacking somatic hypermutation to pro-germinal-center B cells with modest somatic hypermutation (Kolar et al., 2007). In contrast, mutated MCLs (C2) clearly show an imprint of B cells that have experienced the germinal-center reaction. This phenomenon has also been observed in chronic lymphocytic leukemia (CLL), in which three distinct clinicobiological entities can be defined based on DNA methylation patterns of B cell subpopulations at different maturation stages (Bhoi et al., 2016; Kulis et al., 2012; Oakes et al., 2016; Queiros et al., 2015). Overall, we propose an (epi)genetic model of MCL pathogenesis (Figure 8G) in which C1 MCL cases derive from a

i.e., unmethylated regions in the C2 case that overlap with H3K27ac peaks in the C2 case only (D), unmethylated regions in the C1 case that overlap with H3K27ac peaks in the C1 case only (E), or with H3K27ac peaks in both the C1 and C2 case (F). In the lower part of these panels, the percentage of these respective DMRs within the B cell-related, mixed, and B cell-independent DMRs is represented (***$p < 0.001$).
See also Figure S6, and Tables S4 and S5.

CellPress

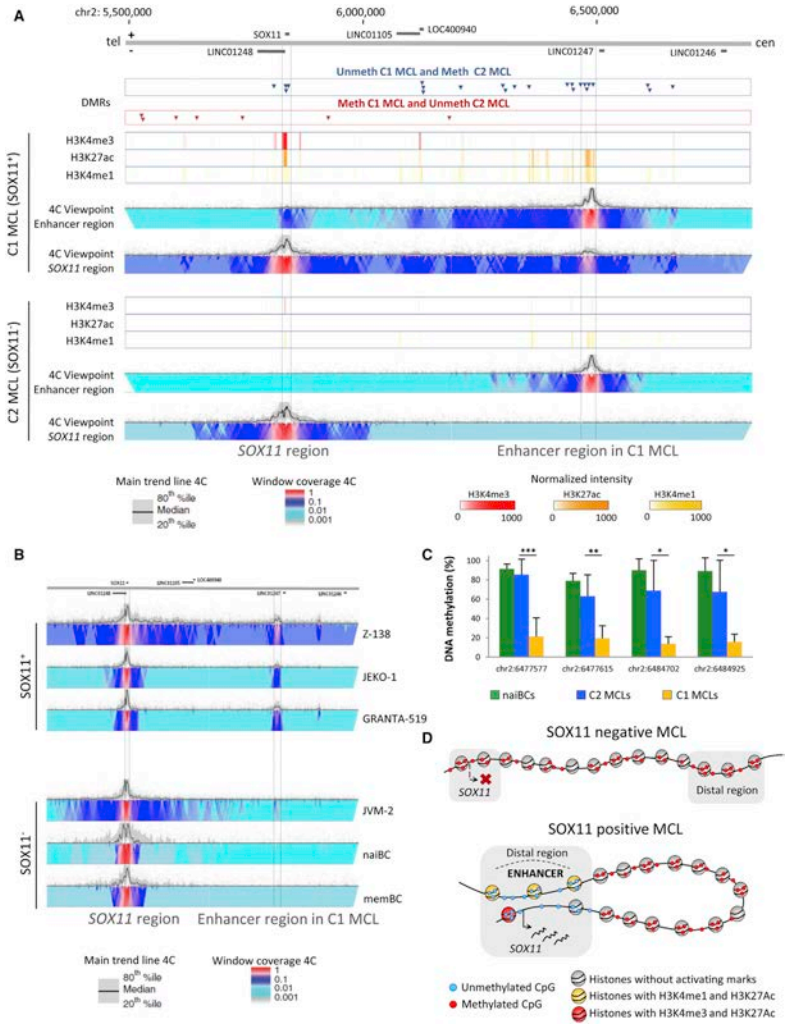


Figure 7. Analysis of the Epigenetic and 3D Structure of the *SOX11* Locus

(A) DMRs, ChIP-seq levels, and 4C-seq signals around the *SOX11* locus. The represented region covers chr2:5,492,778-6,834,378 (hg19). Unmethylated DMRs in the C1 (SOX11-positive) and C2 (SOX11-negative) MCL cases, respectively, are represented in the upper part of the panel by the blue and red arrows. In the

(legend continued on next page)

range of germinal-center-inexperienced B cells that carry the t(11;14) translocation and show absence or low levels of *IGHV* somatic hypermutation (Navarro et al., 2012). Early during transformation, these cells acquire genetic and epigenetic changes and show expression of SOX11, which prevents these cells from entering the germinal center (Palomero et al., 2015). C2 MCLs also carry the t(11;14) translocation but, in contrast to cases from C1, they lack SOX11 expression and show high levels of *IGHV* somatic hypermutation. This fits with the finding that they seem to be derived from germinal-center-experienced B cells, most likely memory B cells (Navarro et al., 2012). C2 MCLs with an indolent clinical course lack oncogenic mutations and acquire few epigenetic changes, whereas C2 MCLs with a more aggressive clinical behavior acquire mutations and present extensive DNA methylation changes. The accumulation of DNA methylation changes may suggest the presence of an epigenetic drift derived from enhanced proliferation induced by oncogenic mutations. However, this finding may also point to a co-evolution of genetic and epigenetic aberrations, as previously reported in CLL (Oakes et al., 2014).

Like within the genetics field, a major question in cancer epigenomics is how to detect potential drivers within a widespread alteration of the DNA methylation landscape. Extrapolating from recent cancer genomic studies, in which the number of potential driver mutations is low compared with the entire mutational burden (Puente et al., 2015; Schuster-Bockler and Lehner, 2012), the proportion of epigenetic drivers may also be low. This is also supported by our data comparing chromatin states between normal B cells and primary cases, in which overall 56% of the regions that undergo DNA methylation changes maintain a stable chromatin environment, and therefore, the function of these regions is most likely not altered. Furthermore, we showed that the majority of CpGs with methylation changes in MCL are affected in only one or few cases. Most likely, this low frequency of recurrent patterns highlights the epigenetic heterogeneity of cancer and reflects that DNA methylation changes globally follow a stochastic model, as previously observed (Landan et al., 2012; Landau et al., 2014; Shipony et al., 2014).

Despite the above-mentioned heterogeneity, an integrative approach combining the DNA methylome and histone modification patterns in primary MCL cases allowed us to identify DMRs with potential functional impact. We propose that epigenetic drivers should be searched in recurrent DMRs containing at least some B cell-independent CpGs and showing a concurrent change in the chromatin activation state. This approach is exemplified by our findings related to the *SOX11* oncogene. With the exception of activating histone marks in its promoter region (Vegliante et al., 2011), no other epigenetic or genetic alterations have been described to account for its de novo upregulation in MCL. Here, we have identified a connection between SOX11 expression and a cluster of hypomethylated DMRs located 650 kb downstream of *SOX11*, a phenomenon that has been observed previously for other cancer-related genes (Aran and Hellman, 2013; Aran et al., 2013). This region showed the canonical elements of an enhancer element such as the presence of nucleosomes containing H3K4me1 and H3K27ac. Furthermore, this region showed high interaction frequencies with the *SOX11* promoter at the 3D level exclusively in SOX11-expressing MCLs, strongly suggesting that it represents an important *SOX11* regulatory region in MCL.

From the clinical perspective, our results suggest that the magnitude of DNA methylation changes is the most relevant independent prognostic factor in our MCL series. However, a more clinically oriented study with a better characterized and homogeneously treated series is required to validate our findings. The extensive epigenetic changes observed in MCL suggest that patients may benefit from the administration of epigenetic drugs (Fiskus et al., 2012). However, the epigenetic heterogeneity observed in MCL may influence efficacy, and it should be taken into account as a potential means to stratify patients.

In conclusion, the analytic strategy presented in this study highlights the significance of taking into account the dynamics of the DNA methylome during normal differentiation to better understand the cancer epigenome and its clinical implications. Furthermore, our study underlines the importance of performing an integrative whole-genome analysis, as proposed by international consortia (Adams et al., 2012; Bernstein et al., 2010), combining DNA methylation, histone modifications, 3D looping, and gene expression to detect distant regulatory elements associated with cancer.

EXPERIMENTAL PROCEDURES

The following experimental procedures represent a succinct summary of the extensive materials and methods applied in the present study. Please refer to the Supplemental Experimental Procedures for further details.

Samples Studied

A total of 82 MCL samples, 4 MCL cell lines (Z-138, JVM-2, JEKO-1, and Granta-519) and 67 samples from B cell subpopulations at different maturation stages (Kulis et al., 2015) were used in the present study. Clinical and biological features of the MCL patients are shown in Table S1. Patients gave their written informed consent, and the study was approved by the clinical research ethics committee of the Hospital Clinic of Barcelona (number 2009/5069) and the internal review board of the University of Kiel (number 447/10).

Deconvolution of DNA Methylation Values

We estimated the proportion of B cells, $CD8^+$ T cells, $CD4^+$ T cells, natural killer cells, monocytes, and granulocytes in the MCL samples (algorithm adapted from Houseman et al., 2012; Jaffe and Irizarry, 2014), and purified the

lower two panels, normalized ChIP-seq intensities for H3K4me3, H3K4me1, and H3K27ac are depicted for the C1 and C2 MCL case. Furthermore, normalized 4C-seq intensities are indicated using the enhancer in MCL C1 (chr2:6,465,559-6,496,708, hg19) or the *SOX11* region as viewpoint. tel, telomere; cen, centromere.

(B) Normalized 4C-seq intensities taking the *SOX11* region as viewpoint in three SOX11-positive MCL cell lines (Z-138, JEKO-1, GRANTA-519), one SOX11-negative MCL cell line (JVM-2), and in normal naive and memory B cells (naiBCs and memBCs).

(C) Mean methylation levels of four CpGs within the SOX11-positive MCL enhancer region in naive B cells (green, n = 4), SOX11-negative (blue, n = 10), and SOX11-positive (orange, n = 12) MCLs as analyzed by bisulfite pyrosequencing. Data show means ± SD. $^*p < 0.01$, $^{**}p < 0.001$, $^{***}p < 0.0001$ (Wilcoxon test for independent samples).

(D) Model of the *SOX11* locus in SOX11-negative MCL (upper) and SOX11-positive MCL (lower).

See also Table S6.

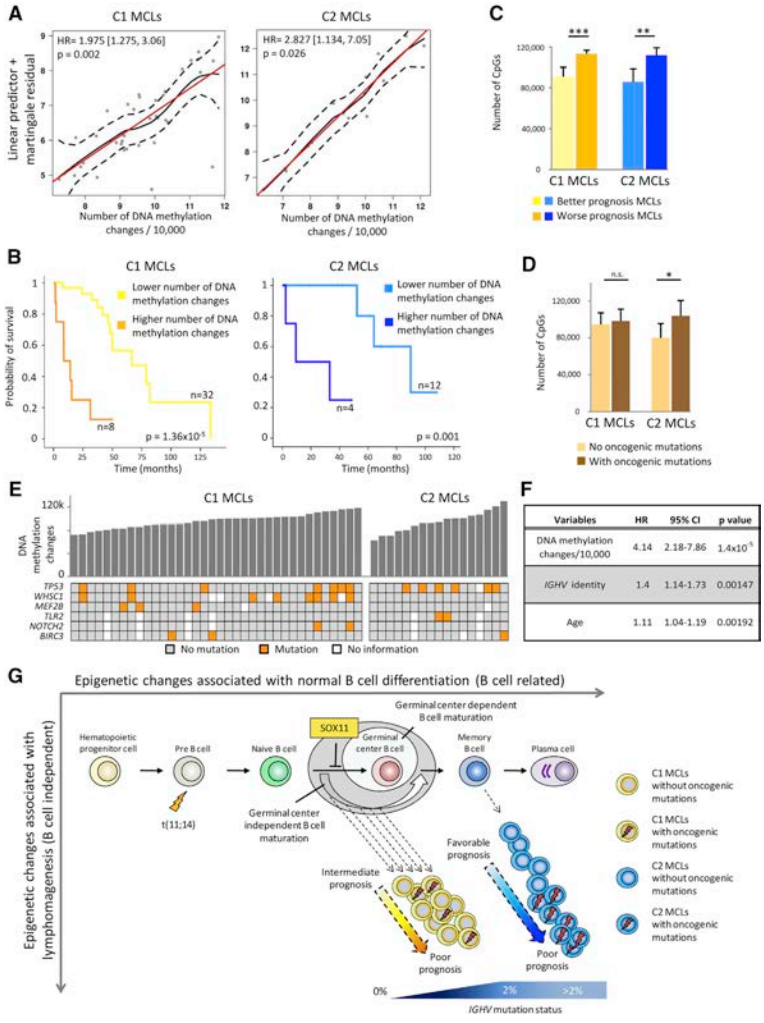


Figure 8. Link between the Number of DNA Methylation Changes and Prognosis

(A) Relationship between the number of epigenetic changes and overall survival through a linear predictor. Red line, perfect linear relationship; black line, observed regression line; dashed line, 95% confidence interval of observed regression.

(B) Kaplan-Meier plots of MCLs with lower versus higher number of differentially methylated CpGs compared with HPCs in C1 and C2 MCLs.

(legend continued on next page)

methylation values of the B cell (i.e., tumor) fraction by subtracting the methylation estimates of the non-B cell fractions.

Epigenomic Analyses

We applied a range of different epigenomic methods, including the HumanMethylation BeadChip (Illumina) in 82 primary MCLs and 67 normal B cell subpopulations (Kulis et al., 2015); WGBS in two MCLs; bisulfite pyrosequencing in 22 MCL cases and naive B cells; ChIP-seq with six different histone modifications in the two MCLs analyzed by WGBS as well as naive and memory B cells, and finally 4C-seq in the two MCLs analyzed by WGBS, four MCL cell lines as well as naive and memory B cells.

Statistical Analysis

The relationships between MCL subgroups and clinical and biological variables of patients was evaluated using Fisher's exact test or t tests, and statistical significance was defined as $p < 0.05$ (corrected for multiple testing if necessary). Univariate and multivariate survival analyses were used to measure the impact of DNA methylation changes in the clinical behavior of MCL patients. All analyses were performed with IBM-SPSS Statistics version 20 or various packages within the R software.

ACCESSION NUMBERS

WGBS, ChIP-seq, and microarray data have been deposited in the European Genome-phenome Archive (EGA) under accession numbers EGAS00001001638, EGAD00001002397, and EGAS00001001637, respectively.

SUPPLEMENTAL INFORMATION

Supplemental Information includes Supplemental Experimental Procedures, seven figures, and six tables and can be found with this article online at http://dx.doi.org/10.1016/j.ccell.2016.09.014.

AUTHOR CONTRIBUTIONS

A.C.Q., R.B., M.D.-F., and M.K. analyzed DNA methylation arrays and integrated the data. R.V.-B., R.B., N.V.-D., A.C.Q., and H.J.G.v.d.W performed and/or analyzed 4C-seq experiments. A.C.Q., R.B., A.M., E.R., M.D.-F., and S.H. processed and analyzed WGBS data. N.R., A.B., and J.M. performed ChIP-seq experiments. S.B., A.N., P.J., A.E., M.J.C., I.V., I.S., W.H.W., W.K., C.P., R.S., and E.C. provided study materials and biological information. A.C.Q., R.B., G.Ca., and G.Cl. functionally characterized differentially methylated regions and performed statistical analysis. E.G., A.L.-G. W.K., C.P., and E.C. reviewed the pathologic and clinical data and confirmed diagnoses. A.D., P.F., and R.R. were in charge of data management. S.H. and I.G.G. coordinated sequencing efforts and performed primary data analysis. M.K., H.G.S., R.S., and E.C. participated in the study design and data interpretation together with A.C.Q., R.B., R.V.-B., and J.I.M.-S. E.C. and J.I.M.-S. conceived the study and led the experiments. A.C.Q., R.B., R.V.-B., and J.I.M.-S. wrote the manuscript.

ACKNOWLEDGMENTS

This work was funded by the European Union's Seventh Framework Program through the Blueprint Consortium (grant agreement 282510), the European Hematology Association (Non-Clinical Advanced Research Fellowships to J.I.M.-S.), the Worldwide Cancer Research (grant number 16-1285, to J.I.M.-S.), Spanish Ministerio de Economía y Competitividad (MINECO), grant no. SAF2015-64885-R (to E.C.) and Generalitat de Catalunya Suport Grups de Recerca AGAUR 2014-SGR-795 (to E.C.). Methylation microarrays were outsourced to the Spanish Centro Nacional de Genotipado (CEGEN-ISCIII). We are indebted to the Genomics core facility of the Institut d'Investigacions Biomèdiques August Pi i Sunyer (IDIBAPS) for technical help. This work was partially developed at the Centro Esther Koplowitz (CEK; Barcelona, Spain). J.I.M.-S. is a Ramón y Cajal researcher of MINECO, E.C. is an Academia Researcher of the "Institució Catalana de Recerca i Estudis Avançats" (ICREA) of the Generalitat de Catalunya, A.C.Q. is supported by a Portuguese Fundação para a Ciência e a Tecnologia (FCT) fellowship, R.B. by fellowships from the Netherlands Organisation for Scientific Research (NWO) and the EU (Marie Curie), and R.V.-B by a pre-doctoral fellowship of the MINECO. Paul Flicek is a member of the Scientific Advisory Board for Omicia, Inc.

Received: February 22, 2016
Revised: July 18, 2016
Accepted: September 19, 2016
Published: November 14, 2016

REFERENCES

Adams, D., Altucci, L., Antonarakis, S.E., Ballesteros, J., Beck, S., Bird, A., Bock, C., Boehm, B., Campo, E., Caricasole, A., et al. (2012). BLUEPRINT to decode the epigenetic signature written in blood. Nat. Biotechnol. *30*, 224–226.

Agirre, X., Castellano, G., Pascual, M., Heath, S., Kulis, M., Segura, V., Bergmann, A., Esteve, A., Merkel, A., Raineri, E., et al. (2015). Whole-epigenome analysis in multiple myeloma reveals DNA hypermethylation of B cell-specific enhancers. Genome Res. *25*, 478–487.

Aran, D., and Hellman, A. (2013). DNA methylation of transcriptional enhancers and cancer predisposition. Cell *154*, 11–13.

Aran, D., Sabato, S., and Hellman, A. (2013). DNA methylation of distal regulatory sites characterizes dysregulation of cancer genes. Genome Biol. *14*, R21.

Baylin, S.B., and Jones, P.A. (2011). A decade of exploring the cancer epigenome - biological and translational implications. Nat. Rev. Cancer *11*, 726–734.

Bea, S., Valdes-Mas, R., Navarro, A., Salaverria, I., Martin-Garcia, D., Jares, P., Gine, E., Pinyol, M., Royo, C., Nadeu, F., et al. (2013). Landscape of somatic mutations and clonal evolution in mantle cell lymphoma. Proc. Natl. Acad. Sci. USA *110*, 18250–18255.

Berman, B.P., Weisenberger, D.J., Aman, J.F., Hinoue, T., Ramjan, Z., Liu, Y., Noushmehr, H., Lange, C.P., van Dijk, C.M., Tollenaar, R.A., et al. (2012). Regions of focal DNA hypermethylation and long-range hypomethylation in colorectal cancer coincide with nuclear lamina-associated domains. Nat. Genet. *44*, 40–46.

Bernstein, B.E., Stamatoyannopoulos, J.A., Costello, J.F., Ren, B., Milosavljevic, A., Meissner, A., Kellis, M., Marra, M.A., Beaudet, A.L., Ecker, J.R., et al. (2010). The NIH roadmap epigenomics mapping consortium. Nat. Biotechnol. *28*, 1045–1048.

Bhoi, S., Ljungstrom, V., Baliakas, P., Mattsson, M., Smedby, K.E., Juliusson, G., Rosenquist, R., and Mansouri, L. (2016). Prognostic impact of epigenetic classification in chronic lymphocytic leukemia: the case of subset #2. Epigenetics *11*, 449–455.

(C) Number of differentially methylated CpGs between the subgroups with different prognosis defined in (B). Data show means ± SD. $^{**}p < 0.01$, $^{***}p < 0.001$ (t test for independent samples).
(D) Association between the number of differentially methylated CpGs and the presence of oncogenic mutations (in *BIRC3*, *MEF2B*, *NOTCH2*, *TLR2*, *TP53*, and *WHSC1* genes) for both C1 and C2 MCLs. Data show means ± SD. $^{*}p < 0.05$; n.s., not significant (t test for independent samples). For cases without or with mutations, the sample sizes are, respectively: C1 (n = 16 and n = 15) and C2 (n = 6 and n = 8).
(E) Representation of epigenetic changes and the presence of oncogenic mutations in both C1 and C2 subgroups.
(F) Results of the multivariate Cox regression model.
(G) Proposed epi(genetic) model of MCL pathogenesis.
See also Figure S7.

CellPress

Bibikova, M., Barnes, B., Tsan, C., Ho, V., Klotzle, B., Le, J.M., Delano, D., Zhang, L., Schroth, G.P., Gunderson, K.L., et al. (2011). High density DNA methylation array with single CpG site resolution. Genomics *98*, 288–295.

Enjuanes, A., Albero, R., Clot, G., Navarro, A., Bea, S., Pinyol, M., Martin-Subero, J.I., Klapper, W., Staudt, L.M., Jaffe, E.S., et al. (2013). Genome-wide methylation analyses identify a subset of mantle cell lymphoma with a high number of methylated CpGs and aggressive clinicopathological features. Int. J. Cancer *133*, 2852–2863.

Esteller, M. (2008). Epigenetics in cancer. N. Engl. J. Med. *358*, 1148–1159.

Feinberg, A.P. (2014). Epigenetic stochasticity, nuclear structure and cancer: the implications for medicine. J. Intern. Med. *276*, 5–11.

Feinberg, A.P., and Vogelstein, B. (1983). Hypomethylation distinguishes genes of some human cancers from their normal counterparts. Nature *301*, 89–92.

Fiskus, W., Rao, R., Balusu, R., Ganguly, S., Tao, J., Sotomayor, E., Mudunuru, U., Smith, J.E., Hembruff, S.L., Atadja, P., et al. (2012). Superior efficacy of a combined epigenetic therapy against human mantle cell lymphoma cells. Clin. Cancer Res. *18*, 6227–6238.

Gama-Sosa, M.A., Slagel, V.A., Trewyn, R.W., Oxenhandler, R., Kuo, K.C., Gehrke, C.W., and Ehrlich, M. (1983). The 5-methylcytosine content of DNA from human tumors. Nucleic Acids Res. *11*, 6883–6894.

Hadzidimitriou, A., Agathangelidis, A., Darzentas, N., Murray, F., Delfau-Larue, M.H., Pedersen, L.B., Lopez, A.N., Dagklis, A., Rombout, P., Beldjord, K., et al. (2011). Is there a role for antigen selection in mantle cell lymphoma? Immunogenetic support from a series of 807 cases. Blood *118*, 3088–3095.

Halldorsdottir, A.M., Kanduri, M., Marincevic, M., Mansouri, L., Isaksson, A., Goransson, H., Axelsson, T., Agarwal, P., Jernberg-Wiklund, H., Stamatopoulos, K., et al. (2012). Mantle cell lymphoma displays a homogenous methylation profile: a comparative analysis with chronic lymphocytic leukemia. Am. J. Hematol. *87*, 361–367.

Hansen, K.D., Timp, W., Bravo, H.C., Sabunciyan, S., Langmead, B., McDonald, O.G., Wen, B., Wu, H., Liu, Y., Diep, D., et al. (2011). Increased methylation variation in epigenetic domains across cancer types. Nat. Genet. *43*, 768–775.

Heintzman, N.D., Hon, G.C., Hawkins, R.D., Kheradpour, P., Stark, A., Harp, L.F., Ye, Z., Lee, L.K., Stuart, R.K., Ching, C.W., et al. (2009). Histone modifications at human enhancers reflect global cell-type-specific gene expression. Nature *459*, 108–112.

Houseman, E.A., Accomando, W.P., Koestler, D.C., Christensen, B.C., Marsit, C.J., Nelson, H.H., Wiencke, J.K., and Kelsey, K.T. (2012). DNA methylation arrays as surrogate measures of cell mixture distribution. BMC Bioinformatics *13*, 86.

Jaffe, A.E., and Irizarry, R.A. (2014). Accounting for cellular heterogeneity is critical in epigenome-wide association studies. Genome Biol. *15*, R31.

Jares, P., Colomer, D., and Campo, E. (2007). Genetic and molecular pathogenesis of mantle cell lymphoma: perspectives for new targeted therapeutics. Nat. Rev. Cancer *7*, 750–762.

Jares, P., Colomer, D., and Campo, E. (2012). Molecular pathogenesis of mantle cell lymphoma. J. Clin. Invest. *122*, 3416–3423.

Jones, P.A. (2012). Functions of DNA methylation: islands, start sites, gene bodies and beyond. Nat. Rev. Genet. *13*, 484–492.

Keshet, I., Schlesinger, Y., Farkash, S., Rand, E., Hecht, M., Segal, E., Pikarski, E., Young, R.A., Niveleau, A., Cedar, H., and Simon, I. (2006). Evidence for an instructive mechanism of de novo methylation in cancer cells. Nat. Genet. *38*, 149–153.

Kolar, G.R., Mehta, D., Pelayo, R., and Capra, J.D. (2007). A novel human B cell subpopulation representing the initial germinal center population to express AID. Blood *109*, 2545–2552.

Kridel, R., Meissner, B., Rogic, S., Boyle, M., Telenius, A., Woolcock, B., Gunawardana, J., Jenkins, C., Cochrane, C., Ben-Neriah, S., et al. (2012). Whole transcriptome sequencing reveals recurrent NOTCH1 mutations in mantle cell lymphoma. Blood *119*, 1963–1971.

Kulis, M., Heath, S., Bibikova, M., Queiros, A.C., Navarro, A., Clot, G., Martinez-Trillos, A., Castellano, G., Brun-Heath, I., Pinyol, M., et al. (2012). Epigenomic analysis detects widespread gene-body DNA hypomethylation in chronic lymphocytic leukemia. Nat. Genet. *44*, 1236–1242.

Kulis, M., Queiros, A.C., Beekman, R., and Martin-Subero, J.I. (2013). Intragenic DNA methylation in transcriptional regulation, normal differentiation and cancer. Biochim. Biophys. Acta *1829*, 1161–1174.

Kulis, M., Merkel, A., Heath, S., Queiros, A.C., Schuyler, R.P., Castellano, G., Beekman, R., Raineri, E., Esteve, A., Clot, G., et al. (2015). Whole-genome fingerprint of the DNA methylome during human B cell differentiation. Nat. Genet. *47*, 746–756.

Landan, G., Cohen, N.M., Mukamel, Z., Bar, A., Molchadsky, A., Brosh, R., Horn-Saban, S., Zalcenstein, D.A., Goldfinger, N., Zundelevich, A., et al. (2012). Epigenetic polymorphism and the stochastic formation of differentially methylated regions in normal and cancerous tissues. Nat. Genet. *44*, 1207–1214.

Landau, D.A., Clement, K., Ziller, M.J., Boyle, P., Fan, J., Gu, H., Stevenson, K., Sougnez, C., Wang, L., Li, S., et al. (2014). Locally disordered methylation forms the basis of intratumor methylome variation in chronic lymphocytic leukemia. Cancer Cell *26*, 813–825.

Leshchenko, V.V., Kuo, P.Y., Shaknovich, R., Yang, D.T., Gellen, T., Petrich, A., Yu, Y., Remache, Y., Weniger, M.A., Rafiq, S., et al. (2010). Genomewide DNA methylation analysis reveals novel targets for drug development in mantle cell lymphoma. Blood *116*, 1025–1034.

Navarro, A., Clot, G., Royo, C., Jares, P., Hadzidimitriou, A., Agathangelidis, A., Bikos, V., Darzentas, N., Papadaki, T., Salaverria, I., et al. (2012). Molecular subsets of mantle cell lymphoma defined by the IGHV mutational status and SOX11 expression have distinct biologic and clinical features. Cancer Res. *72*, 5307–5316.

Oakes, C.C., Claus, R., Gu, L., Assenov, Y., Hullein, J., Zucknick, M., Bieg, M., Brocks, D., Bogatyrova, O., Schmidt, C.R., et al. (2014). Evolution of DNA methylation is linked to genetic aberrations in chronic lymphocytic leukemia. Cancer Discov. *4*, 348–361.

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

Palomero, J., Vegliante, M.C., Eguileor, A., Rodriguez, M.L., Balsas, P., Martinez, D., Campo, E., and Amador, V. (2015). SOX11 defines two different subtypes of mantle cell lymphoma through transcriptional regulation of BCL6. Leukemia *30*, 1596–1599.

Puente, X.S., Bea, S., Valdes-Mas, R., Villamor, N., Gutierrez-Abril, J., Martin-Subero, J.I., Munar, M., Rubio-Perez, C., Jares, P., Aymerich, M., et al. (2015). Non-coding recurrent mutations in chronic lymphocytic leukaemia. Nature *526*, 519–524.

Queiros, A.C., Villamor, N., Clot, G., Martinez-Trillos, A., Kulis, M., Navarro, A., Penas, E.M., Jayne, S., Majid, A., Richter, J., et al. (2015). A B-cell epigenetic signature defines three biologic subgroups of chronic lymphocytic leukemia with clinical impact. Leukemia *29*, 598–605.

Rahmatpanah, F.B., Carstens, S., Guo, J., Sjahputera, O., Taylor, K.H., Duff, D., Shi, H., Davis, J.W., Hooshmand, S.I., Chitma-Matsiga, R., and Caldwell, C.W. (2006). Differential DNA methylation patterns of small B-cell lymphoma subclasses with different clinical behavior. Leukemia *20*, 1855–1862.

Reinius, L.E., Acevedo, N., Joerink, M., Pershagen, G., Dahlen, S.E., Greco, D., Soderhall, C., Scheynius, A., and Kere, J. (2012). Differential DNA methylation in purified human blood cells: implications for cell lineage and studies on disease susceptibility. PLoS One *7*, e41361.

Royo, C., Navarro, A., Clot, G., Salaverria, I., Gine, E., Jares, P., Colomer, D., Wiestner, A., Wilson, W.H., Vegliante, M.C., et al. (2012). Non-nodal type of mantle cell lymphoma is a specific biological and clinical subgroup of the disease. Leukemia *26*, 1895–1898.

Saba, N.S., Liu, D., Herman, S.E., Underbayev, C., Tian, X., Behrend, D., Weniger, M.A., Skarzynski, M., Gyamfi, J., Fontan, L., et al. (2016). Pathogenic role of B-cell receptor signaling and canonical NF-kappaB activation in mantle cell lymphoma. Blood *128*, 82–92.

Schuster-Bockler, B., and Lehner, B. (2012). Chromatin organization is a major influence on regional mutation rates in human cancer cells. Nature *488*, 504–507.

Shipony, Z., Mukamel, Z., Cohen, N.M., Landan, G., Chomsky, E., Zeliger, S.R., Fried, Y.C., Ainbinder, E., Friedman, N., and Tanay, A. (2014). Dynamic and static maintenance of epigenetic memory in pluripotent and somatic cells. Nature *513*, 115–119.

Simonis, M., Kooren, J., and de Laat, W. (2007). An evaluation of 3C-based methods to capture DNA interactions. Nat. Methods *4*, 895–901.

Sims, G.P., Ettinger, R., Shirota, Y., Yarboro, C.H., Illei, G.G., and Lipsky, P.E. (2005). Identification and characterization of circulating human transitional B cells. Blood *105*, 4390–4398.

van de Werken, H.J., Landan, G., Holwerda, S.J., Hoichman, M., Klous, P., Chachik, R., Splinter, E., Valdes-Quezada, C., Oz, Y., Bouwman, B.A., et al. (2012). Robust 4C-seq data analysis to screen for regulatory DNA interactions. Nat. Methods *9*, 969–972.

Vegliante, M.C., Royo, C., Palomero, J., Salaverria, I., Balint, B., Martin-Guerrero, I., Agirre, X., Lujambio, A., Richter, J., Xargay-Torrent, S., et al. (2011). Epigenetic activation of SOX11 in lymphoid neoplasms by histone modifications. PLoS One *6*, e21382.

Xochelli, A., Sutton, L.A., Agathangelidis, A., Stalika, E., Karypidou, M., Marantidou, F., Lopez, A.N., Papadopoulos, G., Supikova, J., Groenen, P., et al. (2015). Molecular evidence for antigen drive in the natural history of mantle cell lymphoma. Am. J. Pathol. *185*, 1740–1748.

Ziller, M.J., Gu, H., Muller, F., Donaghey, J., Tsai, L.T., Kohlbacher, O., De Jager, P.L., Rosen, E.D., Bennett, D.A., Bernstein, B.E., et al. (2013). Charting a dynamic DNA methylation landscape of the human genome. Nature *500*, 477–481.

Manuscript 3

Duran-Ferrer M, Beekman R, **#**Martín-Subero JI. *In silico* deconvolution and purification of cancer epigenomes. Oncoscience. 2017 Apr 14;4(3-4):25-26. doi: 10.18632/oncoscience.346. PMID: 28540331; PMCID: PMC5441471.

www.impactjournals.com/oncoscience/ Oncoscience, Vol. 4(3-4), March 2017

Editorial

In silico deconvolution and purification of cancer epigenomes

Martí Duran-Ferrer, Renée Beekman, and José I. Martín-Subero

Large scale epigenomic initiatives are generating vast amounts of epigenomic profiles from normal and neoplastic cells, including whole genome maps of DNA methylation, histone modifications, chromatin accessibility and the 3D chromatin structure [1]. The effect of tumor cell content is well taken into consideration in fields like genomics, but its impact on epigenomic data remains poorly understood. Tumor samples frequently contain an admixture of neoplastic cells and a non-negligible proportion of healthy cell subtypes from the tumor microenvironment (Figure 1, top), which themselves also show distinct epigenomic profiles. Thus, epigenomic analyses using samples with varying tumor purity in combination with varying proportions of microenvironmental cells can eventually lead to inaccurate biological and clinical interpretations.

We recently performed a detailed analysis of the DNA methylome of mantle cell lymphoma (MCL), a clinically heterogeneous B-cell tumor [2]. Our initial experimental design included samples with a high tumor cell content (median 89%, range 56-100%). However, two referees of the submitted manuscript questioned whether our results and interpretation were influenced by tumor cell content. This criticism triggered a series of analyses and indeed, we observed that tumor cell content greatly affected the DNA methylation estimates and all subsequent downstream biological and clinical associations. In MCL, the major sites of presentation are lymph node (LN) and peripheral blood (PB). Tumor cells from PB can be purified by cell sorting (Figure 1, right). However, if the lymphoma presents in solid tissues, samples are routinely processed in pathology departments as formalin-fixed paraffin-embedded blocks or cryoblocks, which hampers any possibility of cell sorting (Figure 1, left). To remove the effect of contamination of non-tumoral cells from the MCL samples, we adapted the statistical framework proposed by Houseman and coworkers [3] to estimate the proportions of cell subpopulations based on DNA methylation patterns. Next, we used the *in silico* estimated percentages of tumor and microenvironmental cells of each MCL sample, and removed the contribution of the microenvironment to end up with purified DNA methylation values from tumor cells. We repeated all the downstream analyses with corrected DNA methylation values, and all our biological and clinical interpretations became more clear than with uncorrected values (Figure 1, bottom). These included a better separation of two clinico-biological MCL subgroups with different cellular origin and a better estimation of the epigenetic drift (i.e. number of changes per case as compared to hematopoietic precursor cells), which in turn is highly associated with clinical outcome.

In order to correct the DNA methylation values *in silico*, few requirements should be met. First, to predict the purity of tumor samples, a DNA methylation signature of tumor cells related to its cellular origin should be available. This allows to distinguish tumor cells from the microenvironment. In our case, all MCLs contained a clear signature of their B-cell origin. Of note, in some instances, like in multiple myeloma, tumor cells tend to erase the B-cell signature [4] which would make the *in silico* deconvolution and purification inaccurate. Second, the proportion of tumor cells should be much higher than that of their normal cell counterparts. In MCL the proportion of normal B cells is negligible as compared to neoplastic B cells, and therefore the total estimated B-cell fraction could be taken as a surrogate for the tumor cell fraction. Third, we observed that the tumor cell fraction should be sufficiently high, i.e. more than 50%, to accurately correct DNA methylation values. Fourth, reference DNA methylation patterns of purified normal cell types from the tumor microenvironment should be available to allow for proper estimations of their respective fractions and to accurately remove their contribution to the methylation signal of the unpurified sample. In this context, it is worth mentioning that cellular composition can also be estimated using reference-free methods, which have the virtue of allowing deconvolution of methylation signals into its constituents in complex tissues with unknown cell types [5]. However, reference-based methods are superior to reference-free ones if the cellular composition of the samples is known and DNA methylomes for the main cell types are available [6]. In our study, PB samples could be accurately deconvoluted due to the availability of DNA methylation profiles of the major cell subtypes in PB. However, in the case of LN samples, some cell subpopulations were not available, such as macrophages and endothelial cells, and we recognize that deconvolution was less accurate for those samples.

In this editorial, we aimed at highlighting the importance of tumor purity for DNA methylation studies. Ideally, tumor cells should be sorted in the laboratory or, if not possible, adequate bioinformatic pipelines should be applied to deconvolute and purify tumor cell DNA

methylation estimates *in silico*. The benefit of the proposed analytic strategy is multiple. On the one hand, it allows to estimate the composition of the tumor microenvironment, which has been shown to be biologically and clinically important in many tumor entities [7]. On the other hand, this analysis allows for a more accurate characterization of DNA methylation profiles of tumors cells and leads to more robust biological interpretations and clinical associations. Finally, it paves the way for the identification of diagnostic and/or prognostic biomarkers that are not influenced by tumor sample composition.

CONFLICTS OF INTEREST

The authors declare no conflicts of interest.

José I. Martín-Subero: Biomedical Epigenomics Group, Institut d'Investigacions Biomèdiques August Pi i Sunyer (IDIBAPS), Universitat de Barcelona, Barcelona, Spain

Correspondence: *José I. Martín-Subero,* ***email*** *imartins@clinic.cat*

Keywords: *deconvolution, epigenetic profiling, DNA methylation, cancer, tumor purity*

Received: *March 14, 2017*
Published: *April 14, 2017*

REFERENCES

1. Stunnenberg HG, et al. Cell. 2016; 167: 1145-9.
2. Queirós AC, et al. Cancer Cell. 2016; 30: 806-21.
3. Houseman EA, et al. BMC Bioinformatics. 2012; 13: 86.
4. Agirre X, et al. Genome Res. 2015; 25: 478-87.
5. Houseman EA, et al. Bioinformatics. 2014; 30: 1431-9.
6. Hattab MW, et al. Genome Biol. 2017; 18: 24.
7. Lenz G, et al. N Engl J Med. 2008; 359: 2313-23.

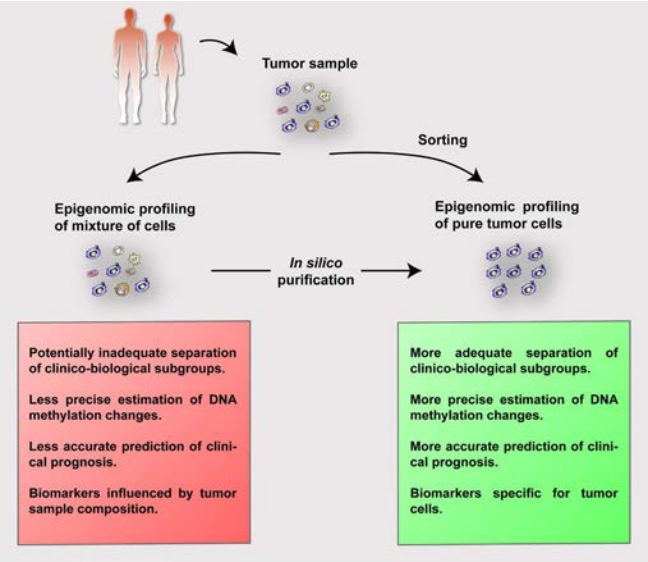


Figure 1: Summary of the process and advantages of analyzing purified (either by cell sorting or *in silico*) epigenomic signatures of tumor cells.

Manuscript 4

 # Duran-Ferrer M, Clot G, Nadeu F, Beekman, R, Baumann, T, Nordlund, J, Marincevic-Zuniga, Y, Lönnerholm G, Rivas-Delgado, A, Martín, S, Ordoñez, R, Castellano, G, Kulis, M, Queirós, A, Seung-Tae, L, Wiemels, J, Royo, R[7], Puiggrós, M, Junyan, L, Giné, E, Beà, S, Jares, P, Agirre, X, Prosper, F, López-Otín, C, Puente, XS, Oakes, CC, Zenz, T, Delgado, J, López-Guillermo, A, Campo, E, **#**Martín-Subero, J.I. The proliferativa history shapes the DNA methylome of B-cell tumors and predicts clinical outcome Nat Cancer, in press.

nature cancer

ARTICLES

https://doi.org/10.1038/s43018-020-00131-2

Check for updates

The proliferative history shapes the DNA methylome of B-cell tumors and predicts clinical outcome

Martí Duran-Ferrer[1,2]✉, Guillem Clot[1,2], Ferran Nadeu[1,2], Renée Beekman[1,2], Tycho Baumann[2,3], Jessica Nordlund[4], Yanara Marincevic-Zuniga[4], Gudmar Lönnerholm[5], Alfredo Rivas-Delgado[1,3], Silvia Martín[1,2], Raquel Ordoñez[2,6], Giancarlo Castellano[1], Marta Kulis[1], Ana C. Queirós[1], Seung-Tae Lee[7], Joseph Wiemels[8], Romina Royo[9], Montserrat Puiggrós[9], Junyan Lu[10], Eva Giné[1,2,3], Sílvia Beà[1,2,11], Pedro Jares[1,2,11], Xabier Agirre[2,6], Felipe Prosper[2,6,12], Carlos López-Otín[2,13], Xosé S. Puente[2,13], Christopher C. Oakes[11], Thorsten Zenz[11,14], Julio Delgado[1,2,3], Armando López-Guillermo[1,2,3], Elías Campo[1,2,15] and José Ignacio Martín-Subero[1,2,15,16]✉

We report a systematic analysis of the DNA methylation variability in 1,595 samples of normal cell subpopulations and 14 tumor subtypes spanning the entire human B-cell lineage. Differential methylation among tumor entities relates to differences in cellular origin and to de novo epigenetic alterations, which allowed us to build an accurate machine learning-based diagnostic algorithm. We identify extensive individual-specific methylation variability in silenced chromatin associated with the proliferative history of normal and neoplastic B cells. Mitotic activity generally leaves both hyper- and hypomethylation imprints, but some B-cell neoplasms preferentially gain or lose DNA methylation. We construct a DNA-methylation-based mitotic clock, called epiCMIT, whose lapse magnitude represents a strong independent prognostic variable in B-cell tumors and is associated with particular driver genetic alterations. Our findings reveal DNA methylation as a holistic tracer of B-cell tumor developmental history, with implications in differential diagnosis and the prediction of clinical outcome.

The process of neoplastic transformation implies a dramatic alteration of cellular identity[1]. However, cancer cells partially maintain molecular imprints of the cellular lineage and maturation stage from which they originate[2]. B-cell neoplasms are paradigmatic of this model, as the maturation stage of different B-cell neoplasms is the main principle behind the World Health Organization's classification of these tumors[3]. Multiple studies have analyzed the DNA methylome, a bona fide epigenetic marker related to cellular identity and gene regulation[1,4], during the entire B-cell maturation program[5] and in various B-cell neoplasms spanning the whole maturation spectrum. These neoplasms include B-cell acute lymphoblastic leukemia (ALL)[6,7] derived from precursor B cells, mantle cell lymphoma (MCL)[8,9] and chronic lymphocytic leukemia (CLL)[10,11] derived from pre- and post-germinal center mature B cells, diffuse large B-cell lymphoma (DLBCL)[12] derived from germinal center B cells and multiple myeloma (MM)[13,14] derived from terminally differentiated plasma cells. These studies, as reviewed in ref. [15], have revealed a dynamic DNA methylome during B-cell maturation as well as novel insights into the cellular origin, pathogenic mechanisms and clinical behavior of B-cell neoplasms. However, a global analysis of the entire normal cell differentiation program and derived neoplasms is not available for B cells or for any other human cell lineage. Here we use both previously generated DNA methylation datasets and newly generated data to systematically decipher the sources of DNA methylation variability across B-cell neoplasms. This comprehensive approach using over 2,000 samples, including training and validation series, indicates that the human DNA methylome is more dynamic than previously appreciated[5,11,16] and reveals hidden biological insights and clinical associations. In particular, de novo disease-specific hypomethylation in active regulatory

[1]Institut d'Investigacions Biomèdiques August Pi i Sunyer (IDIBAPS), Barcelona, Spain. [2]Centro de Investigación Biomédica en Red de Cáncer, CIBERONC, Madrid, Spain. [3]Servicio de Hematología, Hospital Clínic, IDIBAPS, Barcelona, Spain. [4]Department of Medical Sciences, Molecular Medicine and Science for Life Laboratory, Uppsala University, Uppsala, Sweden. [5]Department of Women's and Children's Health, Pediatrics, Uppsala University, Uppsala, Sweden. [6]Centro de Investigación Médica Aplicada (CIMA), IDISNA, Pamplona, Spain. [7]Department of Laboratory Medicine, Yonsei University College of Medicine, Seoul, Korea. [8]Center for Genetic Epidemiology, University of Southern California, Los Angeles, CA, USA. [9]Programa Conjunto de Biología Computacional, Barcelona Supercomputing Center (BSC), Institut de Recerca Biomèdica (IRB), Spanish National Bioinformatics Institute, Universitat de Barcelona, Barcelona, Spain. [10]European Molecular Biology Laboratory (EMBL), Heidelberg, Germany. [11]Division of Hematology, Department of Internal Medicine, The Ohio State University, Columbus, OH, USA. [12]Hematology and Cell Therapy Department, Clínica Universidad de Navarra, Universidad de Navarra, Pamplona, Spain. [13]Departamento de Bioquímica y Biología Molecular, Instituto Universitario de Oncología (IUOPA), Universidad de Oviedo, Oviedo, Spain. [14]Department of Medical Oncology and Hematology, University Hospital Zürich and University of Zürich, Zurich, Switzerland. [15]Departament de Fonaments Clinics, Facultat de Medicina, Universitat de Barcelona, Barcelona, Spain. [16]Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Spain. ✉e-mail: maduran@clinic.cat; imartins@clinic.cat

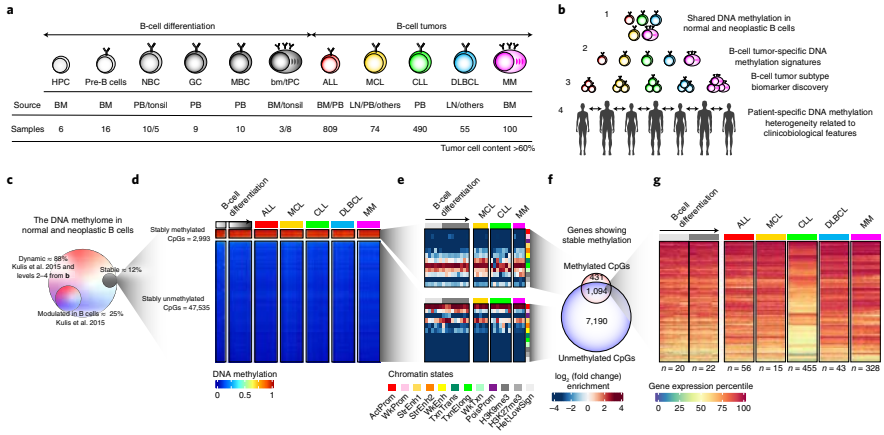


Fig. 1 | Experimental design and characterization of stably methylated regions. a, Experimental design, including normal B-cell subpopulations, B-cell tumors under study, the source of the samples and the number of samples from affected individuals included in the study with a tumor cell content greater than 60%. HPC, hematopoietic precursor cells; pre-B, precursor B cells and immature B cells; NBC, naive B cells; GC, germinal center B cells; MBC, memory B cells; tPC, tonsillar plasma cells; bmPC, bone marrow plasma cells; PB, peripheral blood; LN, lymph node. **b**, Different levels of DNA methylation variability addressed in the study. **c**, Percentage of CpGs whose methylation is either stable or modulated in normal and neoplatic B cells. Percentages were calculated over the total number of CpGs analyzed. **d**, Heat maps showing stably methylated CpGs (top) and stably demethylated CpGs (bottom) in normal and neoplastic B cells. **e**, Chromatin state enrichments for stably unmethylated and methylated CpGs in normal and neoplastic B cells. All CpGs analyzed were used as background. ActProm, active promoter; WkProm, weak promoter; StrEnh1, strong enhancer 1 (promoter related); StrEnh2, strong enhancer 2; WkEnh, weak enhancer; TxnTrans, transcription transition; TxnElong, transcription elongation; WkTxn, weak transcription; PoisProm, poised promoter; H3K27me3, Polycomb-repressed region; H3K9me3, H3K9me3 heterochromatin; Het;LowSign, low-signal heterochromatin. **f**, Overlap between the target genes of the stably methylated and unmethylated CpGs. **g**, Gene expression percentiles in normal and neoplastic B cells of genes showing stable hyper- and hypomethylation.

regions is associated with differential transcription factor (TF) binding, and it targets genes important for disease-specific pathogenesis. From the clinical perspective, we define a set of epigenetic biomarkers that can accurately classify B-cell neoplasms requiring differential clinical management and construct a DNA-methylation-based mitotic clock, called epiCMIT, as a personalized predictor of clinical behavior within each B-cell neoplasm.

Results

Initial data processing and global DNA methylation dynamics in normal and neoplastic B cells. We analyzed previously published DNA methylation profiles of samples corresponding to normal and neoplastic B cells spanning the entire B-cell differentiation spectrum, all generated with the 450K microarray platform from Illumina. These included ten normal B-cell subpopulations[5] as well as the five main categories of B-cell neoplasms: ALL[6,7], MCL[8], CLL[10,17], DLBCL (our own unpublished series) and MM[13] (Fig. 1a and Supplementary Table 1). Following the guidelines of the TCGA Consortium (https://www.cancer.gov/about-nci/organization/ccg/blog/2018/bcr-tips), we selected samples containing a tumor cell content greater than 60%. The validity of this percentage was experimentally confirmed by analyzing the methylation profiles of sorted and unsorted tumor cells from MCL and CLL samples (Extended Data Fig. 1a). Tumor cell content was estimated by flow cytometry (FCM)[5,8,10,13,17], genetic data[18] and/or lineage-specific DNA methylation patterns (Supplementary Table 2) and was highly concordant (Extended Data Fig. 1b). However, in MM samples, the DNA-methylation-based estimation of tumor cell content was far lower than that obtained by FCM (Extended Data Fig. 1c,d), as expected due to their loss of B-cell identity[13]. Interestingly, some DLBCL samples also showed a similar effect (Extended Data Fig. 1c,d). Therefore, in MM and DLBCL samples, tumor cell content was estimated with FCM and genetic data. After filtering the data (Methods), we generated a curated data matrix containing 1,595 high-quality samples (Fig. 1a and Supplementary Table 1) with DNA methylation values for 437,182 CpGs, which were used in all downstream analyses.

This comprehensive dataset was used to dissect the DNA methylation variability of normal and neoplastic B cells at different levels, including cancer-specific, tumor-entity-specific, tumor-subtype-specific and individual-specific variability (Fig. 1b). Of all the studied CpGs, only 12% showed stable DNA methylation levels in normal and neoplastic B cells and they targeted constitutively expressed genes (Fig. 1c–g, Extended Data Fig. 1e–h and Supplementary Table 3), indicating that the majority of the DNA methylome (88%) is labile during normal B-cell development and neoplastic transformation. We could not identify any de novo epigenetic signature that was shared by all B-cell tumors. Therefore, the observed DNA methylation variability was related to differences among B-cell tumor entities and subtypes as well as individual-specific variability.

Disease-specific hypomethylation targeting regulatory regions is associated with TF binding and differential gene expression.

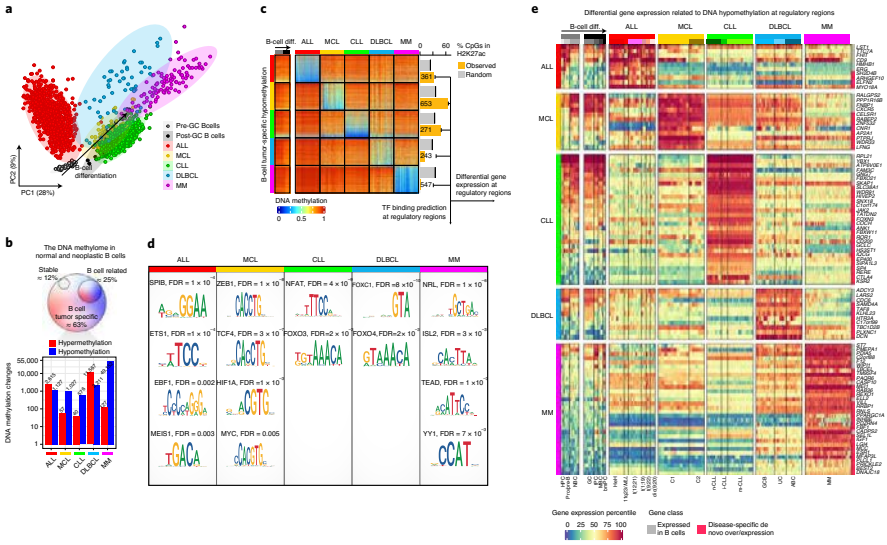


Fig. 2 | Disease-specific DNA methylation signatures. a, PCA of normal and neoplastic B cells. Sample sizes are the same as in Fig. 1a. **b**, Number of de novo DNA methylation changes in each B-cell tumor entity. Percentages were calculated over the total of 437,182 CpGs analyzed. Bar plots represent single data values. **c**, Heat map showing de novo B-cell tumor-specific hypomethylation and the number of CpGs at active regulatory regions marked by H3K27ac. **d**, Enrichment of TF binding sites expressed in B-cell tumors and in regions with de novo hypomethylated CpGs located in active regulatory elements from **c**. **e**, Differential gene expression percentiles for genes showing B-cell tumor-specific hypomethylation in active regulatory regions.

An unsupervised principal-component analysis (PCA) showed that different B-cell neoplasms clustered separately (Fig. 2a and Extended Data Fig. 2a), with neoplasms grouped according to the maturation stage of their cells of origin; for example, ALL samples grouped together with pre-germinal center B cells and mature B-cell neoplasms grouped together with germinal center-experienced B cells. Next, to identify DNA methylation signatures associated with malignant transformation, we focused on the 63% of the genome with potential tumor-specific DNA methylation signatures (Fig. 2b). We detected varying numbers of de novo tumor-specific DNA methylation (tsDNAm) changes, ranging from 616 in CLL to 49,279 in MM (Fig. 2b,c, Extended Data Fig. 2b–d, Supplementary Tables 4 and 5 and Methods). Overall, hypermethylation was enriched at CpG islands and promoter-related regions, whereas hypomethylation occurred at regions with low CpG content (Extended Data Fig. 2e). Remarkably, we observed that DNA methylation changes manifested differently in distinct neoplasms. ALL and DLBCL showed more tumor-specific DNA hypermethylation (tsDNAm-hyper), whereas MCL, CLL and MM acquired more tumor-specific DNA hypomethylation (tsDNAm-hypo), making this trend toward hypomethylation remarkable in MM (Fig. 2b,c). These distinct preferences among neoplasms were apparently not related to differential expression of DNA methyltransferase (*DNMT*) genes, as we could not identify any clear association between the hypermethylation/hypomethylation ratio and *DNMT1*, *DNMT3A* or *DNMT3B* expression levels (Supplementary Fig. 1).

Next, we sought to identify potential upstream mediators for de novo DNA methylation signatures in each B-cell tumor. As TF binding has been reported to induce hypomethylation at regulatory regions[19], we performed a TF binding site prediction analysis in active regulatory elements, which are marked by H3K27ac, containing tsDNAm-hypo CpGs (Methods). Interestingly, the entities that we identified where tsDNAm-hypo was predominantly located within H3K27ac-marked regions (Fig. 2c) showed enrichments for the binding sites of TFs that are expressed in each respective neoplasm and have a previously reported association with their pathogenesis, such as SPI1/SPIB and EBF1 in ALL, TCF/ZEB in MCL and NFAT in CLL (Fig. 2d, Extended Data Fig. 2f and Supplementary Table 6; refs. [20–22]). In the cases of DLBCL and MM, the tsDNAm-hypo CpGs were depleted of active regulatory elements (Fig. 2c), suggesting that TF binding may not be a major factor leading to tumor-specific DNA methylation signatures. However, the fraction of tsDNAm-hypo CpGs located in regulatory regions was enriched in binding sites for TFs potentially involved in the respective diseases, such as the FOX family in DLBCL[23] and NRL (a member of the oncogenic MAF family), ISL1, TEAD and YY1 in MM[24–27] (Fig. 2d).

Beyond the potential role of TFs in shaping tumor-specific DNA methylation signatures, we also investigated the downstream transcriptional associations of tsDNAm-hypo signatures. An analysis of the transcriptional profiles of samples from all five diseases revealed a total of 94 genes associated with tsDNAm-hypo genes that were expressed in a disease-specific manner (Fig. 2e). Although some of the identified genes have been shown to be specifically expressed in a particular disease, such as *CTLA4* and *KSR2* in CLL[28], this comprehensive analysis provides a rich resource of disease-specific candidate genes in which differential DNA methylation may play a role in deregulation.

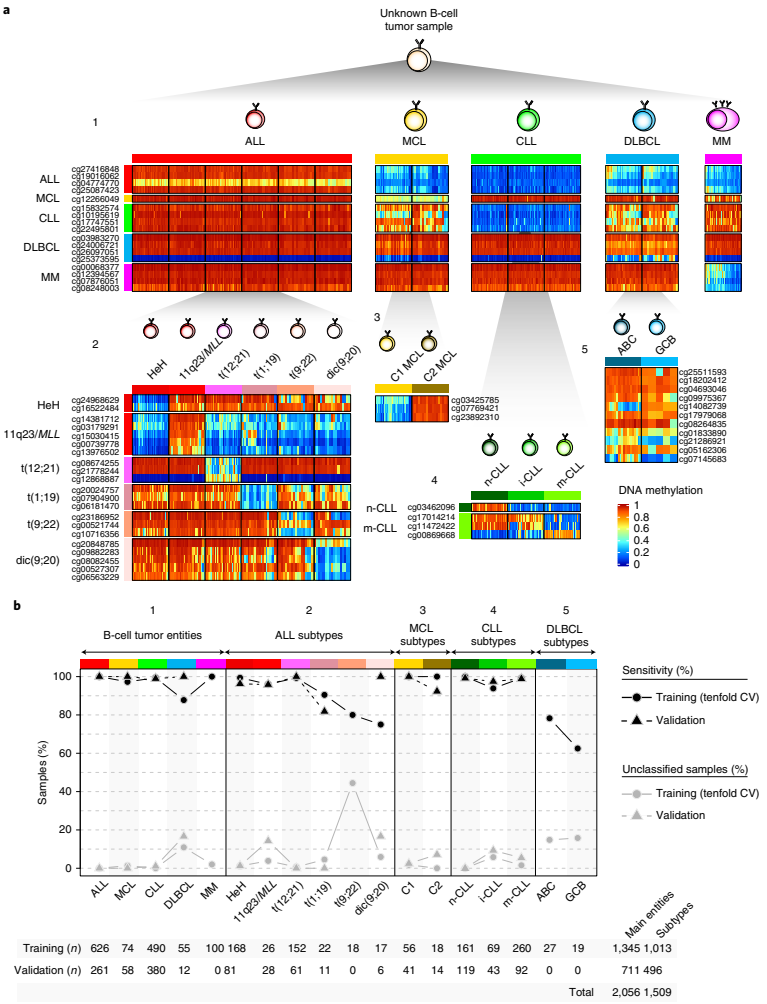


Fig. 3 | Development and validation of a DNA-methylation-based diagnostic classifier of different subtypes of B-cell neoplasms. a, Heat map showing DNA methylation values of the CpGs used for the two-step pan-B-cell cancer classifier. The training samples from **b** are represented. **b**, Accuracy for the pan-B-cell cancer diagnostic classifier composed of the five predictors in **a** in both training and validation series. Sensitivity is represented as black circles or triangles for training or validation series, respectively. The percentage of cases without a clear prediction (unclassified) is represented in gray. The total number of samples used for training and validation is shown at the bottom.

Accurate classification of 14 clinicobiological subtypes of B-cell neoplasms using epigenetic biomarkers. The B-cell neoplasms shown in Fig. 1a represent broad categories that are further classified into subtypes with different clinicobiological features based on genetic, transcriptional or epigenetic features[3]. These features include high-hyperdiploid (HeH) ALLs and ALLs with

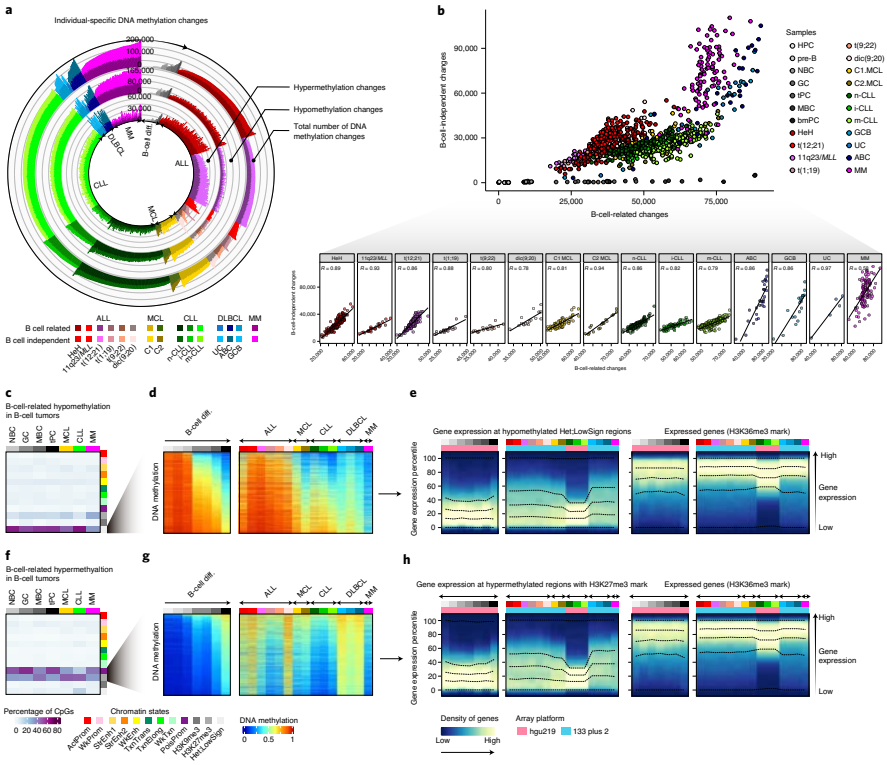


Fig. 4 | Identification and characterization of individual-specific DNA methylation changes. a, Number of DNA methylation changes in individuals for normal and neoplastic B cells as compared to the HPC stage. The total number of DNA methylation changes, hypo- and hypermethylation changes is depicted in the outer, middle and inner tracks, respectively. Changes in DNA methylation are further classified and color coded as B cell related or B cell independent. Sample sizes are the following: HPC, $n=6$; pre-B, $n=16$; NBC, $n=15$; GC, $n=9$; tPC, $n=8$; MBC, $n=10$; bmPC, $n=3$; HeH ALL, $n=168$; 11q23/*MLL* ALL, $n=26$; t(12;21) ALL, $n=152$; t(1;19) ALL, $n=22$; t(9;22) ALL, $n=18$; dic(9;20) ALL, $n=17$; C1 MCL, $n=56$; C2 MCL, $n=18$; n-CLL, $n=161$; i-CLL, $n=69$; m-CLL, $n=260$; GCB DLBCL, $n=19$; ABC DLBCL, $n=27$; UC DLBCL, $n=5$; and MM, $n=100$. The same sample size applies to **b**, **d** and **g**. **b**, Correlation between B-cell-related changes and B-cell-independent changes in normal and neoplastic B cells. *R* values derived from linear models are shown. **c**, B-cell-related CpGs losing DNA methylation in B-cell tumors and the percentages in genomic regions with each chromatin state in normal and neoplastic B cells. The mean of the percentages for each sample type is shown. Sample sizes are the following: NBC, $n=6$; GC, $n=3$; MBC, $n=3$ and tPC, $n=3$ donors; MCL, $n=5$; CLL, $n=7$ and MM, $n=4$ affected individuals. The same sample size applies for **f**. **d**, The mean of 2,000 representative CpGs for each sample subtype from **c**. **e**, Gene density distributed along the expression percentiles of genes associated with B-cell-related CpGs losing DNA methylation in B-cell tumors at Het;LowSign regions. Expressed genes with H3K36me3 are displayed at the right as a positive control. Data are shown as the mean for each sample type. Lines represent the 0, 25th, 50th, 75th and 100th percentiles. **f**, B-cell-related CpGs gaining DNA methylation in B-cell tumors and the percentages in genomic regions with each chromatin state in normal and neoplastic B cells. Data are shown as the mean of the percentages for each sample type. **g**, The mean of 2,000 representative CpGs for each sample subtype from **f**. **h**, Gene density distributed along the expression percentiles of genes associated with B-cell-related CpGs gaining DNA methylation in B-cell tumors in regions containing H3K27me3. Expressed genes with H3K36me3 are displayed at the right as a positive control. Means within each B-cell subpopulation as well as B-cell tumors are shown. Sample subtypes from **d**, **e**, **g** and **h** are color coded as in **b**. Sample sizes for gene expression analyses in **e** and **h** are the following: HPC, $n=3$; pre-B, $n=7$; NBC, $n=10$; GC, $n=11$; tPC, $n=5$ donors; MBC, $n=5$ and bmPC, $n=1$ donor(s); HeH ALL, $n=18$; 11q23/*MLL* ALL, $n=5$, t(12;21) ALL, $n=16$, t(1;19) ALL, $n=6$, t(9;22) ALL, $n=5$, dic(9;20) ALL, $n=6$; C1 MCL, $n=10$; C2 MCL, $n=5$; n-CLL, $n=142$, i-CLL, $n=64$; m-CLL, $n=249$; GCB DLBCL, $n=17$, UC DLBCL, $n=11$, ABC DLBCL, $n=15$ and MM, $n=328$.

structural variants, including rearrangements affecting 11q23/*MLL*, the three different chromosomal translocations t(12;21), t(1;19) and t(9;22) and the dicentric chromosome dic(9;20)[6]; cluster 1 (C1, DNA methylation patterns related to germinal center-inexperienced cells) and cluster 2 (C2, DNA methylation patterns related to germinal center-experienced cells) MCLs, which mostly reflect conventional and leukemic non-nodal MCLs[8,9]; naive-like/low-programmed, intermediate/intermediate-programmed and memory-like/high-programmed CLLs[10,11]; and finally, DLBCLs categorized according to the well-established cell-of-origin classification into germinal center B cell (GCB) and activated B cell (ABC)[29], but not according to the most recent genetic classifications[30,31] whose link with epigenetic profiles deserves further investigation. In MM, a previous report did not show robust methylation differences among the distinct cytogenetic subtypes[13], and thus MM subgrouping was not included in our analyses. Here we focused on the identification of epigenetic biomarkers that may allow a comprehensive diagnosis of B-cell tumor entities and subtypes. We built a classifier algorithm that yielded 56 CpGs as the optimal number, distributed along five predictors (Extended Data Fig. 3a,b, Supplementary Table 7 and Methods), to accurately discriminate the main B-cell tumor entities as a first step (predictor 1) and B-cell tumor subtypes as a second step (predictor 2, 3, 4 or 5; Fig. 3a). The accuracy of the five predictors was evaluated using nested tenfold stratified cross-validation in the training series ($n=1,345$) and with external validation series ($n=711$) (Fig. 3b). Overall, we obtained very high accuracies in the predictions in both main B-cell tumor entities (mean sensitivity was 97% for the training series and 99% for the validation series) and B-cell tumor subtypes (mean sensitivity was 90% for the training series and 97% for the validation series). This epigenetic classifier may represent the basis for a simple and accurate diagnostic tool for B-cell tumor subtypes with different clinical management (see "Code availability" section).

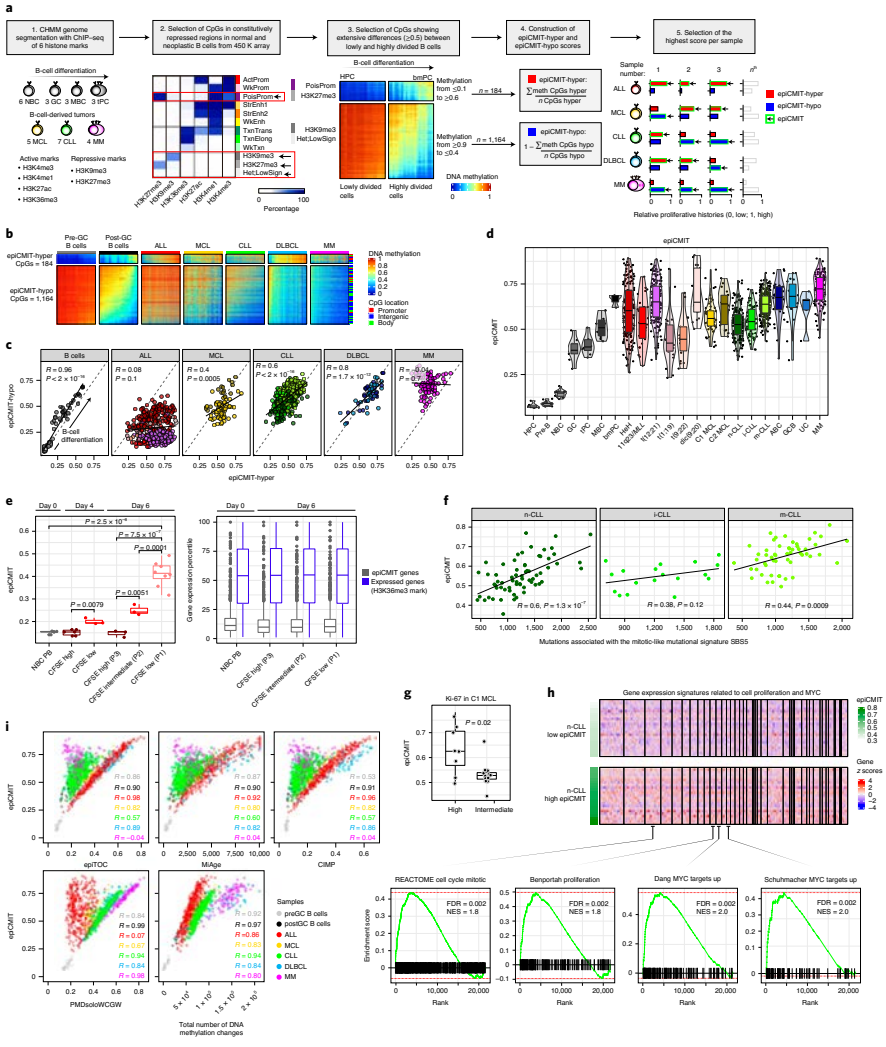
Individual-specific DNA methylation changes are associated with silent chromatin without an impact on gene expression. To determine individual-specific changes within each tumor subtype (Fig. 1b, level 4), we calculated the total number of changes and the number of hyper- and hypomethylation changes in each individual within each B-cell tumor subtype as compared to hematopoietic precursor cells (HPCs). As each B-cell tumor entity has a distinct cellular origin, this approach has the advantage of fixing a reference point for all B-cell tumors. Furthermore, each methylation change was further classified as being extensively modulated or not during normal B-cell development (B cell related or B cell independent, respectively)[5] (Fig. 4a). Overall, we found large differences in the numbers of DNA methylation changes between individuals (Fig. 4a and Supplementary Table 8), and all B-cell tumors showed a similar degree of DNA methylation variability (Extended Data Fig. 4a). We also detected strikingly high correlations between the degree of B-cell-related and B-cell-independent DNA methylation changes (Fig. 4b, Extended Data Fig. 4b and Supplementary Table 8). This association suggests that the overall DNA methylation burden of the tumor in each individual may be shaped by a similar underlying phenomenon. In support of this concept, we observed that hypomethylated CpGs, in both the B-cell-related and B-cell-independent fractions, were mainly located in low-CpG-content, low-signal heterochromatin (Het;LowSign), and the associated genes are constitutively silent in both normal and neoplastic B cells (Fig. 4c–e and Extended Data Fig. 4c–f). In the case of hypermethylation, CpGs in both fractions were located mainly in promoter regions and CpG islands with H3K27me3-repressed and poised-promoter chromatin states and affected genes that remain silent across the normal differentiation and neoplastic transformation of B cells (Fig. 4f–h and Extended Data Fig. 4c,g–i).

Collectively, these findings indicate that most DNA methylation changes in individuals with B-cell tumors occur in silent chromatin regions in the absence of concurrent phenotypic changes, suggesting that a mechanism independent from gene regulation may underlie the overall DNA methylation landscape.

Development of an epigenetic mitotic clock that reflects the proliferative history of normal and neoplastic B cells. Beyond the classical role of DNA methylation as a gene regulator, an accumulating body of published evidence supports the concept that hypomethylation of low-CpG-content heterochromatin and hypermethylation of high-CpG-content Polycomb target regions accumulate during cell division in a way consistent with an epigenetic mitotic clock[32–39]. Here we observe that the DNA methylation variability between samples of B-cell tumors mainly affects inactive chromatin, including hypomethylation of heterochromatin and hypermethylation of regions marked with H3K27me3-containing chromatin states (Fig. 4c–h and Extended Data Fig. 4d–i). Based on these data, these DNA methylation changes most likely reflect the different tumor cell proliferative histories of individuals. We next performed a stepwise selection of CpGs whose methylation change would reflect cell mitotic history (Fig. 5a, Extended Data Fig. 5a and Methods). First, we selected CpGs within constitutively silenced/poised chromatin. Second, we identified CpGs that were methylated (≥0.9) or unmethylated (≤0.1) in HPC samples that extensively lost or gained

Fig. 5 | Development and validation of the epiCMIT. a, Steps to construct the epiCMIT-hyper, epiCMIT-hypo and epiCMIT mitotic clocks. **b**, CpGs constituting the epiCMIT-hyper (184 CpGs) and epiCMIT-hypo (1,164 CpGs) mitotic clocks. **c**, Correlation between the epiCMIT-hyper and epiCMIT-hypo clocks in normal and neoplastic B cells. *R* and *P* values were derived from linear models. **d**, Box plot showing the distribution of epiCMIT values in normal and neoplastic B cells. **e**, Experimental validation of epiCMIT scores with an in vitro B-cell differentiation model of primary human NBCs into plasma cells. The epiCMIT was calculated at day 0, day 4 and day 6 in B cells with distinct proliferative histories based on CFSE dilution. Sample sizes are the following: NBC-PB, $n=5$; CFSE-high at day 4, $n=6$; CFSE-low at day 4, $n=3$; P3 (CFSE-high, CD38$^-$) cells at day 6, $n=3$; P2 (CFSE-intermediate, CD38$^-$) cells at day 6, $n=3$ and P1 (CFSE-low, CD38$^+$) cells at day 6, $n=8$. Each dot within each category is derived from a different donor and represents a biologically independent sample. *P* values were derived from two-sided *t*-tests. On the right, the gene expression profiles of genes containing CpGs belonging to epiCMIT are depicted. The number of genes containing epiCMIT genes analyzed is $n=1,278$, and the number of genes with CpGs mapping at H3K36me3 is $n=14,598$. **f**, epiCMIT correlates with the mitotic-like mutational signature SBS5 in CLL. *R* and *P* values were derived from linear models. Individuals with CLL ($n=138$) with whole-genome sequencing and DNA methylation data are shown. Sample sizes for CLL subtypes are the following: n-CLL, $n=66$; i-CLL, $n=18$ and m-CLL, $n=54$. **g**, epiCMIT is associated with high Ki-67 staining in C1 MCL samples. The number of samples was $n=8$ and $n=12$ for high and intermediate Ki-67 values, respectively. Two-sided *t*-tests were used to assess statistical significance. **h**, GSEA showing that epiCMIT is associated with gene expression signatures related to cell proliferation and MYC activity in CLL. n-CLL samples ($n=142$) were analyzed, and 22 n-CLL samples with low and high epiCMIT scores are shown (15th and 85th percentiles, respectively). At the top, *z*-scores for each gene are represented. At the bottom, representative gene expression signature enrichments are shown. **i**, Correlation between the epiCMIT and previously reported mitotic clocks, including epiTOC, MiAge and PMDsoloWCGW, the pan-cancer CIMP and the total number of DNA methylation changes accumulated since the HPC stage in each individual. *R* values correspond to linear regression models. The sample for **b**–**d** and **i** is the same as that used in Fig. 4a. For all box plots, the center line, box limits, whiskers and points represent the median, 25th and 75th percentiles, 1.5× interquartile range and individual samples, respectively.

methylation (a difference of at least 0.5) in bmPC samples. This difference was used to capture CpGs undergoing extensive methylation changes between cells with the lowest and highest proliferative histories in the B-cell lineage. Third, we obtained 184 CpGs located at constitutive H3K27me3-containing regions and 1,164 CpGs at constitutive heterochromatin, which gain and lose DNA methylation upon cell division, respectively (Fig. 5a,b, Supplementary Table 9 and Methods). Fourth, we constructed two mitotic clocks with these two sets of CpGs, one gaining DNA methylation upon cell division called epigenetically determined cumulative mitoses



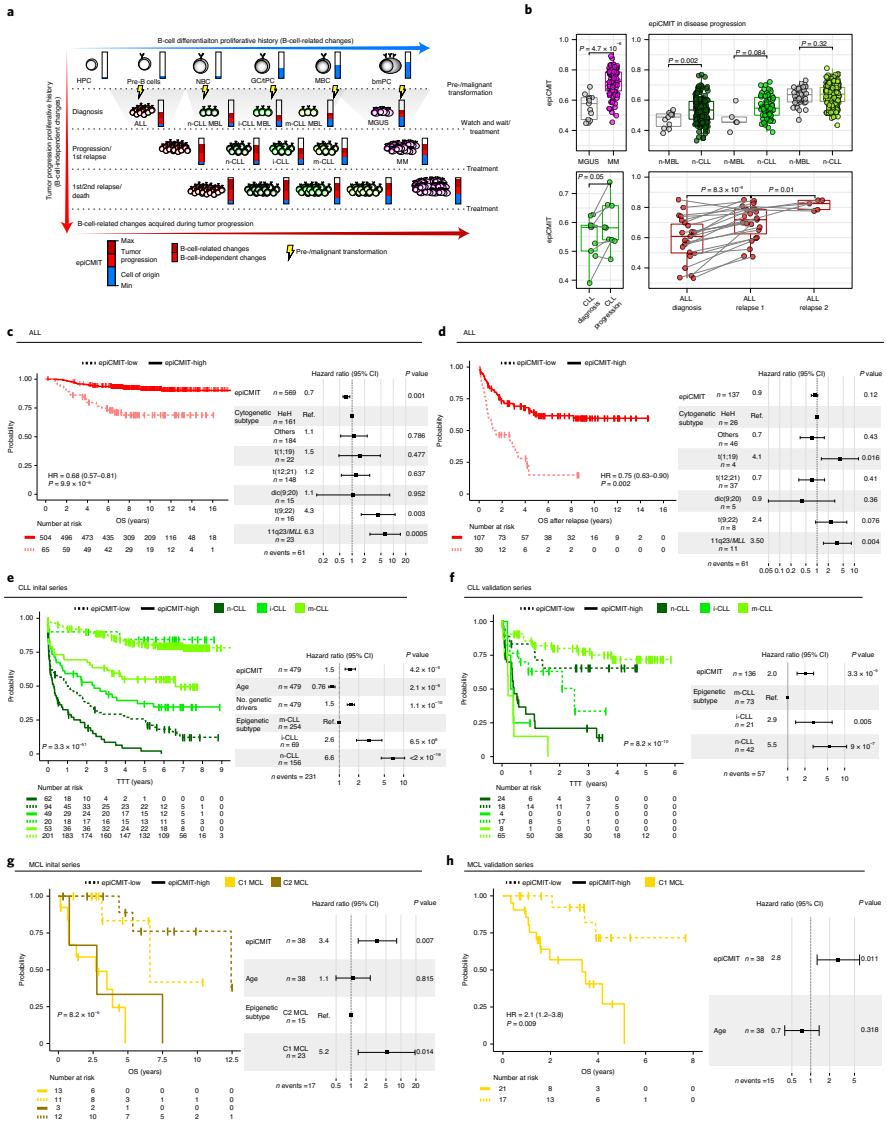
(epiCMIT)-hyper and one losing DNA methylation called epiCMIT-hypo (Fig. 5a,b and Methods). We initially evaluated both mitotic clocks in normal B cells and observed a high correlation ($R=0.96$, $P<2\times10^{-16}$), indicating that B-cell subpopulations are distributed according to their accumulated proliferative history during B-cell differentiation and not to their current proliferation status (Fig. 5c, left). This association between the degree of hyper- and hypomethylation supports previous observations in colorectal cancer[40] and indicates that mitotic cell division in normal B cells leaves both hyper- and hypomethylated imprints. Although this high correlation between the two mitotic clocks was also observed for MCL, CLL and DLBCL (Fig. 5c), it does not seem to be a universal phenomenon, as no correlation was observed in ALL and MM. In line with the overall trend to gain methylation in ALL and to lose methylation in MM (Fig. 2b), we observed that the epiCMIT-hyper was greater than the epiCMIT-hypo in ALL samples, and the opposite was true in MM. These differences do not seem to arise from differential expression of *DNMT* genes (Supplementary Fig. 1). As a final step in development of the epiCMIT mitotic clock, we selected the highest score from the epiCMIT-hyper and epiCMIT-hypo for each sample to derive a unique epiCMIT value (Fig. 5a,d, Supplementary Table 9 and Methods). The epiCMIT then reflected the relative accumulation of mitotic cell divisions of a particular sample, including the mitotic history associated with normal cell development, malignant transformation and progression. Moreover, the epiCMIT cannot be affected by a different distribution of cell cycle phases in tumor samples, as the DNA methylome remains rather stable during the whole cell cycle[41].

Validation of the epiCMIT score as a mitotic clock in normal and neoplastic B cells. The applicability of the epiCMIT as a mitotic clock was validated through the use of several methods. First, we used an independent in vitro B-cell differentiation model of primary NBCs into plasma cells[42] in which cell divisions were controlled by carboxyfluorescein succinimidyl ester (CFSE) staining (Extended Data Fig. 5b). At days 4 and 6, different B cells were separated based on their proliferation history as measured by CFSE dilution, and we observed that epiCMIT increased in cells with lower CFSE concentration, indicating a higher proliferative history (Fig. 5e, left). The genes related to epiCMIT-CpGs remained silenced in all conditions regardless of the cell phenotype and proliferative history (Fig. 5e, right). Second, we studied the link between the epiCMIT and genetic changes using whole-genome sequencing data of 138 individuals with CLL from our cohort[17]. We observed that the epiCMIT was correlated with the total number of somatic mutations and with genomic complexity as measured by the number of driver genetic alterations, that is, mutations with positive selection (Extended Data Fig. 5c,d). Additionally, we measured the activity of known mutational processes through the analysis of single-base substitution (SBS) signatures[43] (Extended Data Fig. 5e). We detected significant correlations between our epiCMIT and signatures SBS5 and SBS1, which have been previously described as mitotic-like mutational processes (Fig. 5f and Extended Data Fig. 5f). We also identified a significant link between the epiCMIT and the non-canonical activation-induced cytidine deaminase (AID) signature (SBS9)[17,43] in *IGHV*-mutated CLL, possibly reflecting accumulated rounds of cell divisions in the germinal center of the ancestor B cell prior to its transformation to CLL (Extended Data Fig. 5g). Third, although the epiCMIT is aimed at capturing the proliferative history of the cell, a relationship with cell proliferation is expected in tumors (more proliferative history implies higher proliferation, although it also depends on time). Accordingly, the epiCMIT was higher in MCL cases showing high Ki-67 expression (a proliferation marker) than in cases with lower Ki-67 expression (Fig. 5g). Furthermore, leukemic CLL cases with high epiCMIT, although not considered to be proliferative, showed higher expression of genes related to cell proliferation and MYC activity (Fig. 5h and Supplementary Table 10). Thus, these data suggest that cases with higher proliferative history also seem to have higher proliferative capacity at the time of sampling.

We next compared the epiCMIT with two previously proposed hypermethylation-based mitotic clocks called epiTOC and MiAge[37,39] (Supplementary Table 8 and Methods). In addition, we calculated a hypomethylation-based mitotic clock using a previously defined pan-cancer set of CpGs losing methylation called PMDsoloWCGW CpGs[38] (Supplementary Table 8 and Methods). Focusing on hypermethylation-based mitotic clocks, the epiCMIT showed correlations with epiTOC and MiAge in B-cell neoplasms acquiring Polycomb-related hypermethylation (mostly ALL, but also DLBCL and MCL). Additionally, the epiCMIT showed a moderate correlation in the case of CLL, which acquires more hypo- than hypermethylation, and a total lack of correlation in the case of MM, which mostly loses DNA methylation (Fig. 5i, top, and Extended Data Fig. 5h). Interestingly, identical observations were obtained comparing the epiCMIT and the widely reported CpG island methylator phenotype (CIMP) in human cancer[44], suggesting that the pan-cancer CIMP score may also represent a measure of the cell mitotic history. Interestingly, the opposite scenario was found when comparing epiCMIT with the hypomethylation-based mitotic clock PMDsoloWCGW. We showed correlations between epiCMIT and PMDsoloWCGW in tumors with extensive DNA hypomethylation (mostly MM and CLL, but also MCL and DLBCL) and a null correlation in ALL (Fig. 5i, bottom). In spite of these discrepancies in ALL and MM, mitotic clocks were, in general, highly correlated

Fig. 6 | Clinical impact of the epiCMIT in B-cell tumors. a, The epiCMIT in neoplastic B cells includes the proliferative history associated with normal B-cell development and with malignant transformation and progression (blue and red components of the epiCMIT bar, respectively). B-cell tumors derive from different maturation stages and contain different normal B-cell baseline epiCMIT. Most of the B-cell-related DNA methylation changes occurring in B-cell tumors relate to cell division. **b**, epiCMIT evolves during disease progression. epiCMIT is lower in precursor conditions, such as monoclonal gammopathy of undetermined significance (MGUS; $n=13$) and MBL ($n=53$), than their respective cancer conditions CLL ($n=437$) and MM ($n=100$), paired CLL samples from diagnosis to progression ($n=9$) and trios of samples obtained from individuals with ALL at diagnosis and first relapse ($n=23$) and second relapse ($n=5$). *P* values were obtained from two-sided *t*-tests, and paired *t*-tests in the case of paired samples. For the box plots, center line, box limits, whiskers and points represent the median, 25th and 75th percentiles, 1.5× interquartile range and individual samples, respectively. **c,d**, Kaplan–Meier curves for OS (**c**) and OS after relapse (**d**) in individuals with ALL with low or high epiCMIT according to the maxstat rank statistics-based cutoff. Hazard ratios and *P* values for the univariate Cox regression models are shown on the left panels. Multivariate Cox regression models with epiCMIT as a continuous variable and ALL cytogenetic groups are shown on the right. Hazard ratio for epiCMIT corresponds to 0.1 increments and also in **e**–**h**. **e**, Kaplan–Meier curves for CLL epigenetic groups based on different cellular origin divided into low and high epiCMIT according to the maxstat rank statistics-based cutoff. A multivariate Cox regression model for TTT with epiCMIT as a continuous variable together with age, number of driver alterations and epigenetic groups based on different cellular origin is shown on the right. The results obtained with the independent validation series are shown in **f**. **g**, Kaplan–Meier curves for MCL epigenetic groups based on different cellular origin divided into low and high epiCMIT according to the maxstat rank statistics-based cutoff. A multivariate Cox regression model for OS with epiCMIT as a continuous variable together with epigenetic groups and age is shown on the right. Validation series for C1 MCL are shown in **h**.

despite the poor overlap of their underlying CpGs, indicating that cell proliferative history can be traced with different sets of CpGs (Extended Data Fig. 5i). Additionally, we observed that epiCMIT is highly correlated with the total number of DNA methylation changes that are accumulated in all samples since the HPC stage, suggesting that the overall DNA methylation landscape seems to



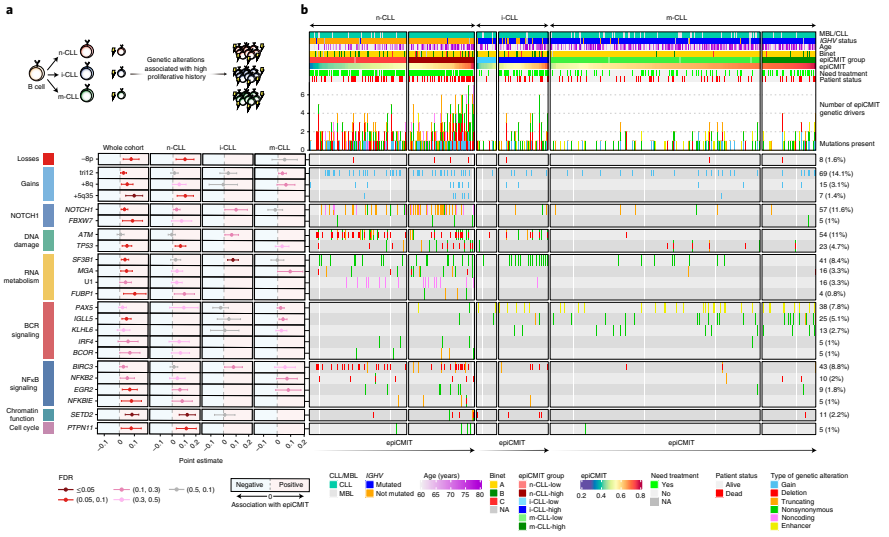


Fig. 7 | Association between the epiCMIT and genetic driver alterations in CLL. a, Schematic representing which potential genetic driver alterations may confer a higher proliferative capacity to CLL cells. **b**, Analysis of the association between the epiCMIT levels and the presence of specific driver genes grouped by signaling pathways. Point estimates with 95% confidence intervals were calculated for the whole cohort using linear modeling between epiCMIT values, and alterations were adjusted for CLL subtypes and with two-sided *t*-tests within CLL subtypes. Point estimates represent the coefficient of each respective alteration in each corresponding linear model in the whole-cohort analysis and the difference between the mean of individuals with CLL with and without each corresponding alteration for the analysis within each CLL subtype. Point estimates are color coded according to false discovery rate (FDR) correction. The Oncoprint shows genetic driver alterations that are significantly associated with higher epiCMIT with CLL epigenetic groups shown separately. Other clinicobiological features, including MBL or CLL, *IGHV* status, age, Binet stage, epiCMIT subgroups based on maxstat rank statistic cutoff, need for treatment and patient status are shown. Samples are ordered within each CLL subgroup from lower to higher epiCMIT values. Genetic driver alterations are depicted with different colors and shapes depending on the alteration type. The number of samples with mutations as well as the percentage of mutated samples over the whole cohort is shown on the right. The whole CLL initial series was used for these analyses and is represented ($n = 490$ samples).

be strongly influenced by the cell proliferative history (Fig. 5i, bottom, and Extended Data Fig. 5h). Finally, epiCMIT outperformed all mitotic clocks to identify cells with different proliferative histories using the controlled setting of the in vitro B-cell differentiation model (Extended Data Fig. 5b,j), a finding that suggests its higher accuracy to trace the B-cell proliferative history. Collectively, these analyses suggest that the epiCMIT is a more universal mitotic clock than previously reported mitotic clocks exclusively based on hyper- or hypomethylation.

A potential confusing aspect related to epiCMIT is the fact that DNA methylation changes take place during aging[45,46] and can be used to predict chronological age[47–49], as exemplified by Horvath's epigenetic clock[50]. To study the potential relationship between mitotic activity and the aging process, we first analyzed the epiCMIT in normal B cells with low (NBC) and high (MBC) epiCMIT values in samples from infants, young adults and elderly donors (Extended Data Fig. 6a, left). This analysis did not reveal any evidence linking the epiCMIT with the chronological age of healthy donors, which indeed is accurately predicted by the Horvath's aging clock (Extended Data Fig. 6a). In the case of B-cell tumors, we observed the same general tendency. Pediatric ALL samples showed the highest epiCMIT range despite the very low age range and a negligible association between the epiCMIT and age. In DLBCL, we observed a similar scenario, as 30- and 90-year-old individuals showed similar epiCMIT levels. Only in individuals with MCL and CLL did we observe minor correlations between epiCMIT and the individual's age (Extended Data Fig. 6a, right). We then applied Horvath's clock to samples from individuals with B-cell tumors and, as previously shown in other cancers[50], we found significant epigenetic age acceleration, reaching an impressive predicted age of over 200 years for some individuals with pediatric ALL. Interestingly, we found that the epiCMIT showed a highly significant correlation with epigenetic age that was predicted by Horvath's clock in the majority of B-cell tumor subtypes ($R = 0.62$, $P < 2 \times 10^{-16}$), suggesting that epigenetic age acceleration may be related to the increased proliferation of cancer cells (Extended Data Fig. 6a, bottom). Despite this intriguing correlation that deserves further investigation, the epiCMIT and Horvath's clocks seem to be targeting different molecular features, as their underlying CpGs show markedly distinct genomic locations, DNA methylation dynamics in normal and neoplastic B cells, chromatin enrichments and gene expression of their associated genes (Extended Data Fig. 6b–f).

The epiCMIT is a strong independent variable that predicts clinical behavior in B-cell tumors. In normal B-cell maturation, the epiCMIT gradually augments as B cells proliferate, an increase that is particularly marked in highly proliferative germinal center B cells (Fig. 5d). In neoplastic B cells, however, the interpretation of the epiCMIT is less trivial and must be divided into two components, (1) the epiCMIT of the cell of origin and (2) the epiCMIT acquired in the course of neoplastic transformation and progression (Fig. 6a). Therefore, the relative epiCMIT must be compared between individuals with entities arising from the same B-cell maturation stage and should be a dynamic variable during cancer progression. We therefore compared the epiCMIT in two paradigmatic transitions between precursor conditions and overt cancer, that is, monoclonal gammopathy of undetermined significance (MGUS) and MM, as well as monoclonal B-cell lymphocytosis (MBLs) and CLLs that were categorized according to their cellular origin. This analysis showed an overall lower epiCMIT in precursor lesions than in overt cancer (Fig. 6b, top). In line with this finding, the epiCMIT increased in paired CLL samples at diagnosis and progression before treatment as well as in sequential ALL samples obtained at diagnosis, first relapse and second relapse (Fig. 6a, bottom).

Based on these observations, we next wondered whether the epiCMIT could be useful to predict the clinical behavior of B-cell neoplasms. We analyzed specific B-cell tumor subtypes based on cytogenetic subtypes (ALL) or cell of origin (MCL, CLL and DLBCL), and, therefore, all subtypes have a similar ground state proliferative history (Fig. 6a). In ALL, high epiCMIT was consistently associated with longer overall survival (OS), OS after relapse and relapse-free survival (RFS; Fig. 6c,d and Extended Data Fig. 7a). These epiCMIT associations maintained an independent statistical significance from the well-established ALL cytogenetic groups as prognostic variable in RFS and OS and a marginal significance in OS after relapse. In contrast to ALL, the opposite clinical scenario was observed in mature B-cell neoplasms. In each of the CLL subtypes, a high epiCMIT was strongly associated with a worse prognosis using time to first treatment (TTT) as an endpoint variable, both from sampling time (Fig. 6e) and in cases whose sample was obtained close to diagnosis (Extended Data Fig. 7b). Additionally, the epiCMIT as a continuous variable showed a highly significant independent prognostic impact in the context of major prognostic factors in CLL, including the *IGHV* status and *TP53* alterations (deletion and mutation; Extended Data Fig. 7c). Overall, it seems that the epiCMIT, CLL epigenetic subgroups[10,11,51] and genomic complexity measured by the total number of driver alterations[17,52] are the most significant independent variables associated with prognosis in CLL. In addition, despite the variability of treatments in our initial CLL series, the epiCMIT also showed marginal significance in OS (Extended Data Fig. 7d). These findings were widely confirmed in an additional series of 210 individuals with CLL treated mainly with chemo-immunotherapy (Fig. 6f and Extended Data Fig. 7b,d). In the case of MCL, the epiCMIT showed an independent poor prognostic impact in the two cell-of-origin subtypes (C1 and C2), an observation that was confirmed in an extended series in the more aggressive and prevalent C1 group (Fig. 6g,h). In the case of the two cell-of-origin DLBCL subtypes, our data suggest that high epiCMIT could also represent a poor prognostic variable (Extended Data Fig. 7e). Finally, our epiCMIT score showed an overall superior prognostic value compared to the other DNA-methylation-based mitotic clocks in all B-cell tumors with the largest number of affected individuals (Extended Data Fig. 8).

epiCMIT is associated with specific genetic driver alterations in CLL. Despite the independent prognostic impact of epiCMIT and genetic alterations in CLL, we next assessed which CLL driver alterations could potentially confer a proliferative advantage to neoplastic cells and a higher epiCMIT (Fig. 7a). To that end, we exploited CLL samples for which we had DNA methylation data and whole-exome sequencing data[17] and depicted all driver genetic changes in each CLL subtype divided into those with high and low epiCMIT (Extended Data Fig. 9a). Next, we interrogated the levels of epiCMIT in samples with each driver genetic alteration both in the whole cohort and in each epigenetic subgroup separately (Fig. 7b, Extended Data Fig. 9b and Methods). We showed significant and positive associations of epiCMIT with 23 genetic driver alterations (Fig. 7b)[17,52]. The majority of these genetic alterations have been previously linked to adverse clinical outcomes in patients and included genes such as *NOTCH1*, *TP53*, *SF3B1*, *ATM*, *BIRC3 or EGR2*. Interestingly, epiCMIT showed an association with a recently identified noncoding genetic driver associated with poor prognosis in CLL, the U1 spliceosomal RNA[53]. Remarkably, the presence of some genetic alterations was associated with high epiCMIT indistinctly in all samples, such as *TP53*, while others were particularly associated with epiCMIT within CLL subgroups, such as *SF3B1* and *ATM* in i-CLL.

Collectively, these results suggest that the well-established clinical impact of certain genetic alterations in CLL may be explained by their association with a high proliferative potential, and this association is different for certain genetic alterations depending on the maturation state of the cellular origin.

Discussion

Here we have followed a systematic approach to dissect the sources of DNA methylation variability of B-cell neoplasms in the context of the normal B-cell differentiation program. Overall, we found that the methylation levels of 88% of the studied CpGs are modulated in normal and/or neoplastic B cells, suggesting that the human DNA methylome is even more dynamic than previously appreciated[5,16]. The extensive DNA methylation variability among different B-cell neoplasms is, in part, related to imprints of normal cell development, a phenomenon that has been recently used to classify not only B-cell neoplasms[8,10,11,51] but also solid tumors[2,54]. In addition, each B-cell neoplasm also shows de novo disease-specific hyper- and hypomethylation, and the latter is possibly related to differential binding of disease-specific TFs and subsequent changes in disease-specific gene expression profiles.

In spite of the widely reported importance of DNA methylation at regulatory regions, we identified that the majority of DNA methylation changes in B-cell neoplasms are located in inactive chromatin. These DNA methylation changes are manifested mainly in the form of concurrent hypomethylation of heterochromatin and hypermethylation of H3K27me3-containing regions, a phenomenon previously observed in colorectal cancer[40]. Compelling published evidence[32–38] and our data support the notion that mitotic cell division leaves transcriptionally inert epigenetic imprints on the DNA located in repressive chromatin environments. More recently, this knowledge has led to the concept of using DNA methylation as a mitotic clock[37–39] and also has been confirmed at the single-cell level[55,56]. Here we identified that using only hyper- or hypomethylation to build a mitotic clock may be insufficient to capture the mitotic history of cancer cells, as some neoplasms seem to preferentially gain or lose DNA methylation upon cell division. For instance, ALL seems to acquire broad hypermethylation upon cell division, whereas we consistently observed the opposite scenario in MM. Thus, using exclusively hyper- or hypomethylation[37–39] to determine the mitotic history of MM or ALL cells would erroneously lead to the conclusion that they have not proliferated beyond their cellular origin. Therefore, to circumvent these limitations, our epiCMIT uses several filters to carefully select both hyper- and hypomethylated CpGs. The strict filtering criteria together with the high correlation with previous cell-type-independent mitotic clocks suggest that the epiCMIT may represent a pan-cancer mitotic clock. Here we showed that epiCMIT captures the entire mitotic history of B cells,

including cell division associated with normal development as well as neoplastic transformation and progression. Thus, the epiCMIT should not be compared among B-cell tumors arising from different normal counterparts, but its relative magnitude must be studied in tumors arising from a particular maturation stage. Within each of these subgroups, the relative epiCMIT has a superior prognostic value in comparison to previous mitotic clocks and a profound independent prognostic value relative to other well-established clinical variables in B-cell tumors. Increased epiCMIT is associated with worse clinical outcome in CLL and MCL, suggesting that superior proliferative history before treatment seems to determine the future proliferative capacity of CLL and MCL cells. Strikingly, we consistently found the opposite pattern in ALL, a finding in line with recent reports showing that the presence of CIMP is associated with better clinical outcome[57,58]. This result may suggest that the highly proliferative ALL cells of children at diagnosis (which have a larger proliferative history) are more efficiently killed by high-intensity chemotherapy regimens[59], which cannot be administrated in elderly patients such as in the case of CLL and MCL.

DNA methylation has also been used as a clock to predict the chronological age of healthy donors[47–49]. The epiCMIT and aging clocks, such as that developed by Horvath[50], seem to broadly reflect different layers of epigenetic information imprinted onto DNA. This notion is supported by multiple perspectives, including the similar levels of epiCMIT in the same normal B-cell subpopulations regardless of donor age, the differential (epi)genomic and transcriptomic features between Horvath's clock and epiCMIT and the independent prognostic value of epiCMIT and age in B-cell tumors. In spite of this overall independence of mitotic and aging clocks, we did observe a remarkable association between the epiCMIT and the epigenetic age predicted by Horvath's clock in B-cell tumors. This finding suggests that the accelerated epigenetic age reported in human cancer[50] may actually reflect the mitotic activity of cancer cells. This concept is further supported by previous results indicating that the predicted age of a sample increases with increased number of in vitro cell passages[50].

Finally, we found that epiCMIT is enhanced by the presence of some mutations with positive selection (that is, those in driver genes) and not by random mutations, as individuals with CLL lacking driver mutations show an overall lower epiCMIT than those with abundant genetic driver alterations. We identified 23 driver genetic alterations that are associated with higher epiCMIT levels or methylation evolution[60] and may represent genetic alterations conferring a higher proliferative capacity to CLL cells. These genetic alterations were distributed throughout the main altered signaling pathways in CLL and manifested differently in distinct CLL subgroups based on their cellular origin (Fig. 7b). This finding suggests that specific alterations may predispose cells to a higher proliferative advantage depending on the maturation stage and (epi)genetic makeup of the CLL cellular origin.

In summary, our comprehensive epigenetic evaluation of normal and neoplastic B cells spanning the entire human B-cell lineage uncovers multiple insights into the biological roles of DNA methylation in cancer, an analytic approach that may also benefit understanding of other cancers. From a clinical perspective, DNA methylation may provide a holistic diagnostic and prognostic approach to B-cell neoplasms. We defined an accurate and easy-to-implement pan-B-cell tumor diagnostic tool and generated a mitotic clock that reflects the proliferative history of the neoplastic cells of each patient to estimate their clinical risk, which represents a valuable asset in the precision medicine era.

Methods

Quality control, normalization, filtering and annotation of DNA methylation data. We collected 450K DNA methylation array data from 913 ALL[5,7], 82 MCL[8], 491 CLL[17] (EGAS00001004640) and 104 MM[13] samples (Supplementary Table 1). We also collected data from 67 normal B-cell subpopulations[5] and data from normal microenvironmental cells, including 6 granulocytes, 5 $CD8^+$ and 5 $CD4^+$ T cells, 6 monocytes, 6 natural killer cells, 6 whole-blood samples, 6 peripheral blood mononuclear cell samples[61], 6 macrophage samples[62] and 16 endothelial cell samples[63]. These microenvironmental cells were used to infer B-cell tumor purities through DNA methylation data. In addition, we generated genome-wide DNA methylation profiles for 80 and 12 DLBCL samples using 450K and EPIC BeadChips (Illumina) (EGAS00001004640), respectively, with partially available genomic data[18]. The analysis of these DLBCL samples was approved by the Institutional Review Board of Hospital Clínic (Barcelona, Spain), and informed consent was obtained from all patients in accordance with the Declaration of Helsinki. In total, 1,799 samples were profiled with the 450K DNA methylation microarrays. We used a custom pipeline to analyze DNA methylation data using R (version 3.6.3) packages and core Bioconductor (version 3.10) packages with special use of the minfi package (version 1.32), which was exclusively used to analyze DNA methylation data[64]. From the total of 485,512 probes present in the 450K array, we sequentially removed probes. Specifically, we initially removed 3,091 non-CpG probes, 17,534 CpGs representing SNPs, 7,715 CpGs with individual-specific methylation[5] and 4,493 CpGs present in sex chromosomes. All of the remaining 452,679 CpGs had a detection P value of $\leq$0.01 in more than 10% of the samples. We then removed samples with poor intensity signal and/or poor probe conversions as well as those with a tumor percentage below 60% (see "Inferring tumor purity through DNA methylation data"). In total, we removed 104 ALL samples, 8 MCL samples, 1 CLL sample, 25 DLBCL samples and 4 MM samples. We also removed microenvironmental cells to perform all the analyses in normal and neoplastic B cells. After applying all filtering criteria, we retained 1,595 samples (Supplementary Table 1 and Fig. 1a) with DNA methylation values for 452,679 CpGs and normalized the values using the SWAN algorithm[65]. Some CpGs showed missing values in some samples and were then removed from all subsequent analyses (with the exception of biomarker discovery; Fig. 3) A final number of 437,182 CpGs were used for all subsequent analyses. We used the IlluminaHumanMethylation450kanno.ilmn12.hg19 (version 0.6) and IlluminaHumanMethylationEPICanno.ilm10b4.hg19 (version 0.6) R packages to annotate all CpGs. Respective B-cell-related and B-cell-independent CpG classification was performed as described previously to separate CpGs that are significantly modulated or not during B-cell differentiation[5]. The same pipeline was used to curate and normalize the data from the previously published[42] in vitro model of B-cell differentiation shown in Extended Data Fig. 5b, all the DNA methylation data for validation series used for the pan-B-cell tumor classifier and clinical associations, including our newly generated EPIC DNA methylation data for the 12 DLBCL samples, as well as other EPIC and previously published 450K DNA methylation data. In particular, we collected EPIC DNA methylation profiles for 70 individuals with MCL[8] and 450K and EPIC data for 380 individuals with CLL from external collaborators (EGAD00010000871, EGAD00010000948).

Inferring tumor purity through DNA methylation data. DNA methylation has been shown to represent an appropriate biological layer to infer the proportions of blood cell types in peripheral blood[66]. We have previously successfully implemented this statistical framework to infer tumor purity in samples from individuals with MCL[8]. We have extended this strategy to all B-cell tumors using additional cell types to deconvolute DNA methylation data into cellular proportions, including tumor cell content. We validated this approach using FCM and genetic data in MCL and CLL samples (Extended Data Fig. 1b). Briefly, we assumed that B-cell tumors retain a B-cell signature from their cell of origin and also have a negligible proportion of normal B cells. Thus, the percentage of neoplastic B cells in a sample can be inferred by the presence of a DNA methylation signature of B cells. This B-cell methylation signature was identified by two sequential steps: (1) we selected CpGs with shared methylation values during the entire B-cell maturation process (from early committed B cells to terminally differentiated bmPCs) and (2) from those CpGs selected above, we performed a differential DNA methylation analysis to identify CpGs whose methylation levels were significantly different between B cells and the major non-neoplastic cells accompanying B-cell tumors[69], namely granulocytes, T cells, monocytes, macrophages and endothelial cells. Then, with this set of CpGs that represented all major cell types present in tumor samples, we applied a linear constrained projection[68], also known as a reference-based approach[70], to find the proportions of each cell type.

As a final filtering step, we retained patient samples showing at least 60% tumor cell content according to DNA-methylation-based predictions (ALL, MCL and CLL samples), FCM (MM samples) and genetic data (DLBCL samples).

Purity estimation from mutational and copy number variation data in DLBCL. The 80 DLBCL samples included in this study were previously analyzed by whole-genome copy number arrays (Cytoscan HD, Affymetrix), and gene mutations were analyzed by targeted next-generation sequencing of 106 genes[18]. The allele-specific copy number analysis of tumors algorithm available at Nexus Copy Number (BioDiscovery, version 7) was used to infer the tumor purity directly from the Cytoscan HD array. The percentage of cells (or the cancer cell fraction, CCF) carrying each somatic mutation found in loci not affected by copy number alterations was calculated as CCF = 2 × VAF, where VAF is the variant allele

frequency of the mutation. Of all the mutations, the highest CCF was considered as the best estimate of sample tumor purity based on gene mutations. As a final step, the maximum tumor purity detected by allele-specific copy number analysis or gene mutations was considered as the estimated tumor cell purity.

Gene expression data. Gene expression profiles using hgu219 array data for normal B cells were obtained from ref. [5] (3 HPCs, 7 pre-B cells samples, 10 NBCs, 11 germinal center B cells samples, 5 tPCs, 5 MBCs and 1 bmPC). Additionally, we downloaded gene expression data for 56 ALL samples profiled with 133 plus 2 arrays from ref. [6], including several ALL subtypes, namely 18 HeH, 5 11q23/*MLL*, 16 t(12;21), 6 t(1;19), 5 t(9;22) and 6 dic(9;20). We also used 15 MCL samples profiled with 133 plus 2 arrays[71], including 10 C1 and 5 C2 MCLs. We used previously generated gene expression data from hgu219 arrays for 455 CLL samples[17]. For DLBCL samples, we generated gene expression data using 133 plus 2 arrays following the manufacturer's instructions for 43 DLBCL samples, including 17 GCBs, 15 ABCs and 11 unclassified (UC) cells. Finally, we downloaded gene expression data for 328 MM samples from ref. [72], which were analyzed with the 133 plus 2 array platform. We normalized all the data using the rma function available in the affy (version 1.64) R package. As the gene expression data came from different studies and different array platforms, we transformed all normalized gene expression values for each sample to gene expression percentiles to minimize batch effects. Also, we generally used expression data to strengthen the interpretation of previous results and not for primary and discovery analyses.

Shared DNA methylation dynamics in normal and neoplastic B cells. To define CpGs whose methylation values do not change in normal and neoplastic B cells, we obtained CpGs showing differences of less than 0.25 across all normal and neoplastic B cells. We then classified them into hyper-, partial and hypomethylated CpGs by calculating the median of each CpG for all samples.

ChIP–seq data collection, analysis and integration. We downloaded and processed ChIP–seq data available from Blueprint[73] and from a previous study in ALL[74]. Specifically, we used Blueprint ChIP–seq data of six histone marks, including H3K4me1, H3K4me3, H3K27ac, H3K36me3, H3K27me3 and H3K9me3, available for 15 normal B cells (6 NBCs, 3 GCs, 3 MBCs and 3 tPCs), 5 MCLs, 7 CLLs and 4 MMs as well as two DLBCL cell lines, KARPAS-422 and SUDHL-5. We next integrated these ChIP–seq data using chromHMM software[75] as previously described[76]. Briefly, we generated a B-cell-specific chromatin state model with 12 emission states using the 15 normal B cells and corrected for their corresponding input. These 12 chromatin states were ActProm (active promoter, with H3K27ac and H3K4me3), WkProm (weak promoter, with H3K4me1 and H3K4me3), PoisProm (poised promoter, with H3K27me3, H3K4me1 and H3K4me3), StrEnh1 (strong enhancer 1, with H3K27ac, H3K4me1 and H3K4me3), StrEnh2 (strong enhancer 2, with H3K27ac and H3K4me1), WkEnh (weak enhancer, with H3K4me1), TxnTrans (transcription transition, with H3K36me3, H3K27ac and H3K4me1), TxnElong (transcription elongation, with H3K36me3), WkTxn (weak transcription, with low H3K36me3), H3K9me3 (H3K9me3-marked repressed heterochromatin), H3K27me3 (H3K27me3-marked repressed heterochromatin) and Het;LowSign (low-signal heterochromatin, with the absence of all six histone marks). Next, this model was used to assign the chromatin states in the remaining primary B-cell tumors, namely five MCL, seven CLL and four MM samples and the two DLBCL cell lines. In the case of ALL, we downloaded H3K27ac ChIP–seq data (generated with the ChIP-grade ab4729 from Abcam) from the NALM6 ALL cell line[74]. We followed the Blueprint pipeline to find H3K27ac peaks (http://dcc.blueprint-epigenome.eu/#/md/chip_seq_grch37). To define regulatory regions in MCL, CLL, MM and DLBCL, we used the CHMM genome segmentation. Specifically, we used chromatin states containing H3K27ac, namely ActProm, StrEnh1, StrEnh2 and TxnTrans chromatin states. For ALL, regulatory regions were defined as regions showing H3K27ac peaks. These active regulatory regions were not merged but were used in a disease-specific manner in the manuscript. To calculate CHMM enrichments of CpG sets, we used the CpGs present in the 450K Illumina DNA methylation array as a background. To calculate CpG enrichments in the regulatory regions in Fig. 2c and Extended Data Fig. 2c, the number of CpGs falling in regulatory regions was compared to the same number of de novo CpGs chosen 10,000 times at random from the DNA methylome fraction with potential tumor-specific signatures falling in regulatory regions. To select genes associated with regulatory regions (Fig. 2), we obtained gene annotations for all CpGs within regulatory regions using the IlluminaHumanMethylation450kanno.ilmn12.hg19 R package.

Gene ontology analysis. Gene ontology analyses were performed using the gometh function within the missMethyl R package available from Bioconductor, which takes into account the differing number of probes per gene present on the 450K array.

Tumor-specific DNA methylation signatures. We performed a truncated PCA using the irlba package available from the CRAN. To find specific DNA methylation signatures in each B-cell tumor, we filtered out all CpGs that showed extensive modulation in B-cell differentiation[7]. Afterwards, we used the limma package to perform pairwise comparisons between each B-cell tumor entity. For each B-cell neoplasia as compared to other B-cell tumors, we retained CpGs that showed at least a $\geq$0.25 methylation difference and an FDR of <0.05 in the same direction in all comparisons. We next classified the identified CpGs as hyper- or hypomethylated by considering the methylation status of normal B cells.

Transcription factor binding analysis. For the TF binding analysis, we used the PWMEnrich package available from Bioconductor. We focused on CpGs showing specific hypomethylation in each B-cell tumor entity that overlapped with regions showing H3K27ac in primary samples of MCL, CLL or MM and in cell lines in the case of ALL (NALM6) and DLBCL (KARPAS-422 and SUDHL-5; Fig. 2c). We next extended the DNA sequence 100 base pairs (bp; 50 bp to each side) for each CpG using the Bsgenome.Hsapiens.UCSC.hg19 annotation package available from Bioconductor. As background sequences, we used 100,000 random B-cell-independent CpGs. We then calculated the frequency of A, T, C and G bases in the background sequences. Next, we obtained the 537 CORE JASPAR 2018 TFs for *Homo sapiens* and transformed the motifs to position weight matrices using previously calculated frequencies of each base to account for biases in the 450K array. We then calculated a lognormal background distribution with tiles of 100 bp to predict TF binding. We retrieved enrichments per group of sequences and the frequency of each TF that belonged to the top 5% of enriched TFs, that is, how often a TF was among the top 5% enriched TFs in all the interrogated sequences. We considered a TF as relevant when it was within the top 5% of TFs in at least 10% of the sequences, had an FDR $\leq$ 0.025 and was consistently expressed in each respective B-cell tumor.

Construction of the classifier algorithm for B-cell tumor subtypes. DNA methylation data for 1,345 samples of B-cell neoplasms were used to build a two-step classifier for the classification of the five main B-cell tumor entities (first step) followed by the classification of B-cell tumor subtypes (second step; of the 1,345 samples, 1,013 samples with a subtype diagnosis were available). We ussed the DNA methylation values of 452,679 CpGs, including B-cell-related and B-cell-independent CpGs[5]. Of note, to build the classifier, we only used CpGs present in both methylation array platforms (450K and EPIC arrays). CpGs with minimal variation, defined by having an interquartile range below 0.07, were removed in the training series of each one of the five predictors.

The following strategy was used to build the predictor for the main B-cell tumor entities as well as for ALL, MCL and DLBCL tumor subtypes (predictors 1, 2, 3 and 5). In the case of CLL, we used another strategy, which is subsequently described.

1. For every class *k*,
   i. Rank the CpGs according to the Mann–Whitney *U*-test *P* value resulting from the comparison of samples of class *k* against the samples of all other classes.
   ii. Define the signature of class *k* as the mean of 13methylation values of the top M_k CpGs (or one minus the value for hypomethylated CpGs in class *k*). In case of ties in the *P*-value ranking, the CpG with a higher mean DNA methylation change is prioritized.
2. Train a support vector machine model with the signatures of the *k* classes using a linear kernel and optimize the cost *C* by cross-validation. In the case of only two classes, such as MCL or DLBCL (for example, C1 versus C2 and ABC versus GCB subtypes), the two signatures are redundant and only one is retained.

The number of CpGs included in the signature of each class in (1) ii, vector $M=\{M_{ALL}, \ldots, M_{GCB}\}$ was chosen by tenfold stratified cross-validation. Specifically, the above algorithm was repeated at each fold where all combinations of possible M_k values were tested and the values that maximized the balanced accuracy were selected. The tested values ranged from 1 to a different quantity depending on the predictor (20 for the main entities, 30 for the ALL subtypes, 20 for CLL, 20 for MCL and 20 for DLBCL).

For the classification of the three CLL subtypes (m-CLL, i-CLL and n-CLL), the described 5-CpG classifier[10,51] could not be applied as one CpG (cg09637172) is not present in the EPIC array, and, therefore, we reanalyzed the data to obtain a new predictor using the following steps:

1. Select the 50 CpGs with the lowest Mann–Whitney *U*-test *P* value for each pairwise comparison between the three subtypes.
2. Apply the SVM-RFE algorithm[77] to the subset of CpGs selected in step 1.
3. Train a support vector machine model with the top M_{CLL} CpGs of step 2, cost *C* and a linear kernel.

A similar cross-validation strategy as the previous algorithm was used to optimize the M_{CLL} and *C* parameters. The tested values were $M_{CLL}=\{1, 2, \ldots, 20\}$ CpGs and $C=10^{\{-3, -2, \ldots, 3\}}$ cost. Extended Data Fig. 3d shows the balanced accuracy and sensitivities of the best performing cost for each number of CpGs.

Finally, we used two strategies to estimate the accuracy of the five predictors: (1) with nested cross-validation in the training series and (2) with a validation series. For the training series, we used tenfold stratified cross-validation, where the optimization of the *M* and *C* parameters was independently performed at each fold

using an inner stratified cross-validation step. For the validation series, we used the following data.

For ALL, we used 183 samples that were already included in the initial analysis (Figs. 1 and 2; ref. [7]) but were not used to construct any classifier or used in any of the other analyses of the manuscript. Additionally, we downloaded the following DNA methylation data: GSE76585 (ref. [66]) and GSE69229 (ref. [67]). For MCL validations, we used DNA methylation data from 58 non-overlapping MCL cases[9] (EGAS00001004165). For CLL validation, we collected 450K methylation data for 109 CLL samples from a previous study[13] (EGAD00010000871), and 145 CLL samples with 450K data and 126 CLL samples with EPIC data were kindly provided by T. Zenz and were partially deposited in ref. [74] (EGAD00010000948). Finally, for DLBCL validation, we generated DNA profiles with EPIC arrays (EGAS00001004640).

To more accurately represent indetermination in newly obtained samples, not all cross-validated training samples or validation samples were assigned to an entity/subtype. Specifically, we used the svm function of the e1071 R package to obtain a probability for each entity/subtype in each one of the samples. Next, samples where the maximum probability was below 50% or where multiple entities/subtypes (including the true entity) had a probability above 35% were considered unclassified.

In the case of MCL, the classification of the training series into C1 and C2 subtypes was performed using a strategy that mirrored the previously described approach[9]. Specifically, we first created a PCA space using all of the unfiltered methylation information in the training samples and identified that the first two components contained most of the information related to the subtype. These two components were used to fit a quadratic discriminant analysis model that distinguished the two cell-of-origin subtypes in this new space. Finally, the validation samples were projected into the training PCA space, and the fitted quadratic discriminant analysis model was applied to them. Only samples with either a C1 or C2 probability ≥85% were assigned to one of the subtypes. This strategy allowed us to define a cell-of-origin subtype for the validation series using the methylation information as a whole.

Between-sample DNA methylation heterogeneity. To analyze the variability of DNA methylation data among samples, we identified CpGs with differential methylation in each individual. To do this, we compared data from each individual with the mean in HPC samples and considered a DNA methylation change for a given CpG when a difference of ≥0.25 in methylation was reached. Next, to define all the DNA methylation changes occurring in individuals diagnosed with a specific B-cell tumor subtype, we selected all CpGs that met the following criteria: (1) present in at least one individual with a specific B-cell tumor subtype showing an absolute methylation difference of ≥0.25 as compared to HPCs and (2) all other individuals in the B-cell tumor subtype show the same trend toward hypomethylation or hypermethylation.

Construction of the epiCMIT score. To create the epiCMIT score, we selected all CpGs from the 450K array of our entire DNA methylation matrix of normal and neoplastic B cells ($n = 1{,}595$) that were located in inactive regions, particularly in PoisProms (with H3K27me3, H3K4me1 and H3M4me3), in H3K27me3 regions, in H3K9me3 regions and in Het;LowSign heterochromatin (absence of any of the six marks analyzed). We divided this set of CpGs into two distinct sets, CpGs located in H3K27me3-repressed regions or PoisProm, and CpGs located in H3K9me3-repressed regions or Het;LowSign heterochromatin. We next performed a differential DNA methylation analysis between normal B cells with the lowest and the highest proliferative histories, namely HPCs and bmPCs (step 3; Extended Data Fig. 5a), and we retained CpGs that demonstrated increased DNA methylation in bmPCs in H3K27me3 regions or PoisProm and CpGs that demonstrated decreased DNA methylation in bmPCs in H3K9me3 and Het;LowSign heterochromatin. In addition, we imposed two key restrictions to these two sets of CpGs. First, CpGs gaining and losing methylation during cell division must respectively show very low (≤0.1) and very high (≥0.9) methylation levels in cells with low division, that is, HPCs. Second, we retained only those CpGs that showed extensive modulation between low-dividing HPCs and high-dividing bmPCs. This second condition was imposed to maximize the differences in the DNA methylation values upon cell division. With all these restrictions, we ended with 184 hypermethylated CpGs that were used to build the epiCMIT-hyper score. Conversely, we retained 1,164 hypomethylated CpGs to construct the epiCMIT-hypo mitotic score. These scores were generated using the following formulas:

$$epiCMIT-hyper = \frac{\sum_1^{184} DNA\ methylation\ epiCMIT-hyper\ CpGs}{184}$$

$$epiCMIT-hypo = 1 - \frac{\sum_1^{1164} DNA\ methylation\ epiCMIT-hypo\ CpGs}{1164}$$

Finally, to construct the epiCMIT score, we evaluated the epiCMIT-hyper and epiCMIT-hypo scores for each sample and selected the higher of the two:

$$epiCMIT = max\{epiCMIT-hyper, epiCMIT-hypo\}per\ sample$$

As the epiCMIT score was built with 450K array data, there are 84 CpGs that are not present in the currently available EPIC array from Illumina (10 epiCMIT-hyper and 74 epiCMIT-hypo). Nonetheless, we showed high correlations between epiCMIT scores that were calculated with all the original CpGs and those exclusively present in both the 450K and EPIC arrays (data not shown).

Determination of epiTOC, MiAge, CIMP and PMDsoloWCGW mitotic clocks and the Horvath chronological clock. To determine epiTOC[37], MiAge[39], CIMP[79], PMDsoloWCGW[38] and Horvath[80] DNA methylation clocks, we used their underlying CpGs that overlapped with those present in our curated DNA methylation matrix. Specifically, the numbers of CpGs were the following: 377 of the 385 epiTOC CpGs, 261 of the 268 MiAge CpGs, 88 of the 89 pan-cancer CIMP CpGs[79], 5,595 of the 6,214 PMDsoloWCGW CpGs and 351 of the 353 Horvath CpGs. We calculated the epiTOC and MiAge scores as previously indicated[37,39]. For the CIMP score, we used a set of previously proposed CpGs[79] and used the same strategy as with the epiCMIT-hyper score. In the case of the PMDsoloWCGW mitotic clock, we applied the same strategy that we used for the epiCMIT-hypo score (explained in the previous section). Finally, we used the Horvath chronological clock to predict age using R as previously reported[80].

Somatic mutations and mutational signature analysis in CLL. The somatic mutations found in the CLL samples used in this study were reported elsewhere[17]. We considered driver alterations as those that were reported in Puente et al.[17] and Landau et al.[52]. In addition, a new recurrent driver mutation has been recently added to CLL, namely a mutation in the U1 spliceosomal RNA[53]. We obtained the U1 mutational status for 318 individuals with CLL that were already published. For the remaining 172 individuals with CLL from our analyses, we evaluated the U1 mutational status using the rhAmp SNP Assay (Integrated DNA Technology) as previously described[53]. Next, the mutational signature analysis[80] was performed following a similar framework as the one described in Alexandrov et al.[43]. Briefly, de novo signature extraction was performed using a hierarchical Dirichlet process (hdp R package, https://github.com/nicolaroberts/hdp), and extracted signatures were matched to the recently described list of mutational signatures[43] based on cosine similarity and the biological knowledge of each mutational process. Signatures identified through this approach included SBS1, SBS5, SBS8, SBS9, SBS17b and SBS18. Finally, the contribution of each of the previously identified signatures for each sample was measured using a fitting approach (MutationalPatterns R package). To avoid signature bleeding between samples, we iteratively removed one signature after another, and the signature contributing the least was censored if removal reduced the cosine similarity by <0.005 with the exception of SBS1 and SBS5, which were always included based on their reported presence in all normal and tumor samples.

Gene Set Enrichments Analysis (GSEA). To perform GSEA analysis in CLLs with different epiCMIT scores, we took CLL samples separated by their cellular origin[18,51] (epigenetic groups) above the 85th percentile and below the 15th percentile of epiCMIT. i-CLL samples were excluded due to the smaller sample size. We performed differential gene expression analysis using limma. We then used the fgsea package to perform GSEA analyses using log fold change as the summary statistic to rank genes. We downloaded 5,501 curated (C2) gene signatures from the Molecular Signatures Database v7.0 (https://www.gsea-msigdb.org/gsea/index.jsp). We performed GSEA analysis with all pathways, filtering those with less than 5 genes and with more than 5,000 genes. We used 10,000 permutations to obtain P values. We next selected 118 gene expression signatures related to cell proliferation and MYC in an unbiased way. These 118 expression signatures were found in R by regular expression matching with the grep() R function using the following expression: grep("CELL_CYCLE|prolifer|divi|mitotic|_CYCLING|M_PHASE|_MYC_", names(gene_expression_signatures_names)).

epiCMIT clinical associations. We performed a univariate analysis of epiCMIT scores for RFS, OS and OS after relapse in ALL. OS and TTT for CLL and OS for MCL were calculated using Kaplan–Meier curves with maxstat statistics to define groups with high and low epiCMIT. The hazard ratios and their corresponding P values were shown when epiCMIT categorization was performed. Finally, epiCMIT was assessed in OS together with ABC and GCB DLBCL transcriptomic subtypes[59]. The epiCMIT prognostic value was assessed in the presence of other well-established prognostic factors in all diseases with multivariate Cox regression models. In ALL, this included HeH ALLs, others (non-recurrent, undefined, < 45 chr, > 67 chr and iAMP21), t(1;19), t(12,21), dic(9;20), t(9;22) and 11q23/MLL. In MCL, we performed the multivariate Cox regression model for OS with epiCMIT together with epigenetic groups C1 and C2 and with age. Finally, in CLL we performed multivariate Cox regression models for TTT and OS with epiCMIT together with age at sampling, epigenetic groups and the total number of driver alterations considering mutations in both studies[17,52]. We scaled all mitotic clocks when comparing the prognostic value among them.

Finding CLL driver alterations associated with increased epiCMIT. We analyzed the association of each genetic alteration with epiCMIT in all individuals with CLL and in those belonging to each epigenetic subgroup separately. When evaluating

all CLL samples together, we modeled epiCMIT scores with each genetic alteration using linear regression correcting by epigenetic subgroups. We used *t*-tests comparing the levels of epiCMIT in mutated and unmutated samples for each genetic alteration within each epigenetic subgroup. We derived point estimates and 95% confidence intervals in both the global analysis for all CLLs and within each epigenetic subgroup for all the tests performed (*P* values were corrected using FDR). We finally grouped the genetic alterations most significantly associated with epiCMIT with pathways implicated in the pathogenesis of CLL. Treated and untreated individuals at the time of sampling were used to perform these analyses.

Statistics and reproducibility. Sample size and data exclusion criteria are extensively explained in "Quality control, normalization, filtering and annotation of DNA methylation data". The experiments were not randomized. The investigators were not blinded to allocation during experiments and outcome assessment.

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

DNA methylation and gene expression data that support the findings of this study have been deposited at the European Genome-phenome Archive (EGA) under accession number EGAS00001004640. Previously published DNA methylation data that were reanalyzed in this study can be found under the following accession codes: B cells, EGAS00001001196; ALL, GSE16368, GSE47051, GSE7658515 and GSE6922916; MCL, EGAS00001001637 and EGAS00001004165; CLL, EGAD00010000871 and EGAD00010000948; MM, EGAS00001000841; in vitro B-cell differentiation model of NBCs from human primary samples, GSE72498. Normalized DNA methylation matrices used for the analyses in this study are available at http://resources.idibaps.org/paper/the-proliferative-history-shapes-the-DNA-methylome-of-B-cell-tumors-and-predicts-clinical-outcome. Published gene expression datasets can be found under the following accession codes: B cells, EGAS00001001197; ALL, GSE47051; MCL, GSE36000; CLL, EGAS00000000092 and EGAD00010000254; MM, GSE19784; in vitro B-cell differentiation model of NBCs from human primary samples, GSE72498. ChIP–seq datasets that were reanalyzed in this study can be found under the following accession codes: GSE109377 (NALM6 ALL cell line, $n = 1$) and EGAS00001000326 (15 normal B cell donors and 5 individuals with MCL, 7 individuals with CLL and 4 individuals with MM) available from Blueprint (https://www.blueprint-epigenome.eu/). All other data supporting the findings of this study are available from the corresponding author on reasonable request.

Code availability

The source code for the DNA methylation classifier of B-cell tumor entities and subtypes and for the calculation of the epiCMIT mitotic clock can be found at https://github.com/Duran-FerrerM/Pan-B-cell-methylome. All other source code supporting the findings of this study is available from the corresponding author on reasonable request.

Received: 7 February 2020; Accepted: 22 September 2020;
Published online: 02 November 2020

References

1. Roy, N. & Hebrok, M. Regulation of cellular identity in cancer. *Dev. Cell* **35**, 674–684 (2015).
2. Hoadley, K. A. et al. Cell-of-origin patterns dominate the molecular classification of 10,000 tumors from 33 types of cancer. *Cell* **173**, 291–304 (2018).
3. Swerdlow, S. H. et al. *WHO Classification of Tumours of Haematopoietic and Lymphoid Tissues* 4th edn, Vol. 2 (International Agency for Research on Cancer (IARC), 2017).
4. Luo, C., Hajkova, P. & Ecker, J. R. Dynamic DNA methylation: in the right place at the right time. *Science* **361**, 1336–1340 (2018).
5. Kulis, M. et al. Whole-genome fingerprint of the DNA methylome during human B cell differentiation. *Nat. Genet.* **47**, 746–756 (2015).
6. Nordlund, J. et al. Genome-wide signatures of differential DNA methylation in pediatric acute lymphoblastic leukemia. *Genome Biol.* **14**, r105 (2013).
7. Lee, S.-T. et al. Epigenetic remodeling in B-cell acute lymphoblastic leukemia occurs in two tracks and employs embryonic stem cell-like signatures. *Nucleic Acids Res.* **43**, 2590–2602 (2015).
8. Queirós, A. C. et al. Decoding the DNA methylome of mantle cell lymphoma in the light of the entire B cell lineage. *Cancer Cell* **30**, 806–821 (2016).
9. Nadeu, F. et al. Genomic and epigenomic insights into the origin, pathogenesis, and clinical behavior of mantle cell lymphoma subtypes. *Blood* **136**, 1419–1432 (2020).
10. Kulis, M. et al. Epigenomic analysis detects widespread gene-body DNA hypomethylation in chronic lymphocytic leukemia. *Nat. Genet.* **44**, 1236–1242 (2012).
11. Oakes, C. C. et al. DNA methylation dynamics during B cell maturation underlie a continuum of disease phenotypes in chronic lymphocytic leukemia. *Nat. Genet.* **48**, 253–264 (2016).
12. Shaknovich, R. et al. DNA methylation signatures define molecular subtypes of diffuse large B-cell lymphoma. *Blood* **116**, e81–e89 (2010).
13. Agirre, X. et al. Whole-epigenome analysis in multiple myeloma reveals DNA hypermethylation of B cell-specific enhancers. *Genome Res.* **25**, 478–487 (2015).
14. Kaiser, M. F. et al. Global methylation analysis identifies prognostically important epigenetically inactivated tumor suppressor genes in multiple myeloma. *Blood* **122**, 219–226 (2013).
15. Oakes, C. C. & Martin-Subero, J. I. Insight into origins, mechanisms, and utility of DNA methylation in B cell malignancies. *Blood* 132, 999–1006 (2018).
16. Ziller, M. J. et al. Charting a dynamic DNA methylation landscape of the human genome. *Nature* **500**, 477–481 (2013).
17. Puente, X. S. et al. Non-coding recurrent mutations in chronic lymphocytic leukaemia. *Nature* **526**, 519–524 (2015).
18. Karube, K. et al. Integrating genomic alterations in diffuse large B-cell lymphoma identifies new relevant pathways and potential therapeutic targets. *Leukemia* **32**, 675–684 (2018).
19. Stadler, M. B. et al. DNA-binding factors shape the mouse methylome at distal regulatory regions. *Nature* **480**, 490–495 (2011).
20. Somasundaram, R., Prasad, M. A. J., Ungerbäck, J. & Sigvardsson, M. Transcription factor networks in B-cell differentiation link development to acute lymphoid leukemia. *Blood* **126**, 144–152 (2015).
21. Sánchez-Tilló, E. et al. The EMT activator ZEB1 promotes tumor growth and determines differential response to chemotherapy in mantle cell lymphoma. *Cell Death Differ.* **21**, 247–257 (2014).
22. Wolf, C. et al. NFATC1 activation by DNA hypomethylation in chronic lymphocytic leukemia correlates with clinical staging and can be inhibited by ibrutinib. *Int. J. Cancer* **142**, 322–333 (2018).
23. Blonska, M. et al. Jun-regulated genes promote interaction of diffuse large B-cell lymphoma with the microenvironment. *Blood* **125**, 981–991 (2015).
24. Huerta-Yepez, S. et al. Overexpression of Yin Yang 1 in bone marrow-derived human multiple myeloma and its clinical significance. *Int. J. Oncol.* **45**, 1184–1192 (2014).
25. Sprynski, A. C. et al. Insulin is a potent myeloma cell growth factor through insulin/IGF-1 hybrid receptor activation. *Leukemia* **24**, 1940–1950 (2010).
26. Riz, I. & Hawley, R. G. Increased expression of the tight junction protein TJP1/ZO-1 is associated with upregulation of TAZ–TEAD activity and an adult tissue stem cell signature in carfilzomib-resistant multiple myeloma cells and high-risk multiple myeloma patients. *Oncoscience* **4**, 79–94 (2017).
27. Herath, N. I., Rocques, N., Garancher, A., Eychène, A. & Pouponnot, C. GSK3-mediated MAF phosphorylation in multiple myeloma as a potential therapeutic target. *Blood Cancer J.* **4**, e175 (2014).
28. Navarro, A. et al. Improved classification of leukemic B-cell lymphoproliferative disorders using a transcriptional and genetic classifier. *Haematologica* **102**, 360–363 (2017).
29. Alizadeh, A. A. et al. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature* **403**, 503–511 (2000).
30. Chapuy, B. et al. Molecular subtypes of diffuse large B cell lymphoma are associated with distinct pathogenic mechanisms and outcomes. *Nat. Med.* **24**, 679–690 (2018).
31. Schmitz, R. et al. Genetics and pathogenesis of diffuse large B-cell lymphoma. *N. Engl. J. Med.* **378**, 1396–1407 (2018).
32. Aran, D., Toperoff, G., Rosenberg, M. & Hellman, A. Replication timing-related and gene body-specific methylation of active human genes. *Hum. Mol. Genet.* **20**, 670–680 (2011).
33. Beerman, I. et al. Proliferation-dependent alterations of the DNA methylation landscape underlie hematopoietic stem cell aging. *Cell Stem Cell* **12**, 413–425 (2013).
34. Landan, G. et al. Epigenetic polymorphism and the stochastic formation of differentially methylated regions in normal and cancerous tissues. *Nat. Genet.* **44**, 1207–1214 (2012).
35. Siegmund, K. D., Marjoram, P., Woo, Y.-J., Tavaré, S. & Shibata, D. Inferring clonal expansion and cancer stem cell dynamics from DNA methylation patterns in colorectal cancers. *Proc. Natl Acad. Sci. USA* **106**, 4828–4833 (2009).
36. Spencer, D. H. et al. CpG island hypermethylation mediated by DNMT3A is a consequence of AML progression. *Cell* **168**, 801–816 (2017).
37. Yang, Z. et al. Correlation of an epigenetic mitotic clock with cancer risk. *Genome Biol.* **17**, 205 (2016).
38. Zhou, W. et al. DNA methylation loss in late-replicating domains is linked to mitotic cell division. *Nat. Genet.* **50**, 591–602 (2018).
39. Youn, A. & Wang, S. The MiAge Calculator: a DNA methylation-based mitotic age calculator of human tissue types. *Epigenetics* **13**, 192–206 (2018).
40. Berman, B. P. et al. Regions of focal DNA hypermethylation and long-range hypomethylation in colorectal cancer coincide with nuclear lamina-associated domains. *Nat. Genet.* **44**, 40–46 (2011).

41. Vandiver, A. R., Idrizi, A., Rizzardi, L., Feinberg, A. P. & Hansen, K. D. DNA methylation is stable during replication and cell cycle arrest. *Sci. Rep.* **5**, 1–8 (2015).
42. Caron, G. et al. Cell-cycle-dependent reconfiguration of the DNA methylome during terminal differentiation of human B cells into plasma cells. *Cell Rep.* **13**, 1059–1071 (2015).
43. Alexandrov, L. B. et al. The repertoire of mutational signatures in human cancer. *Nature* **578**, 94–101 (2020).
44. Issa, J. CpG island methylator phenotype in cancer. *Nat. Rev. Cancer* **4**, 988–993 (2004).
45. Rakyan, V. K. et al. Human aging-associated DNA hypermethylation occurs preferentially at bivalent chromatin domains. *Genome Res.* **20**, 434–439 (2010).
46. Teschendorff, A. E. et al. Age-dependent DNA methylation of genes that are suppressed in stem cells is a hallmark of cancer. *Genome Res.* **20**, 440–446 (2010).
47. Bell, C. G. et al. DNA methylation aging clocks: challenges and recommendations. *Genome Biol.* **20**, 249 (2019).
48. Field, A. E. et al. DNA methylation clocks in aging: categories, causes, and consequences. *Mol. Cell* **71**, 882–895 (2018).
49. Horvath, S. & Raj, K. DNA methylation-based biomarkers and the epigenetic clock theory of ageing. *Nat. Rev. Genet.* **19**, 371–384 (2018).
50. Horvath, S. DNA methylation age of human tissues and cell types. *Genome Biol.* **14**, R115 (2013).
51. Queirós, A.C. et al. A B-cell epigenetic signature defines three biological subgroups of chronic lymphocytic leukemia with clinical impact. *Leukemia* **29**, 598–605 (2015).
52. Landau, D. A. et al. Mutations driving CLL and their evolution in progression and relapse. *Nature* **526**, 525–530 (2015).
53. Shuai, S. et al. The U1 spliceosomal RNA is recurrently mutated in multiple cancers. *Nature* **574**, 712–716 (2019).
54. Rodríguez-Paredes, M. et al. Methylation profiling identifies two subclasses of squamous cell carcinoma related to distinct cells of origin. *Nat. Commun.* **9**, 577 (2018).
55. Gaiti, F. et al. Epigenetic evolution and lineage histories of chronic lymphocytic leukaemia. *Nature* **569**, 576–580 (2019).
56. Meir, Z., Mukamel, Z., Chomsky, E., Lifshitz, A. & Tanay, A. Single-cell analysis of clonal maintenance of transcriptional and epigenetic states in cancer cells. *Nat. Genet.* **52**, 709–718 (2020).
57. Borssén, M. et al. DNA methylation holds prognostic information in relapsed precursor B-cell acute lymphoblastic leukemia. *Clin. Epigenetics* **10**, 31 (2018).
58. Sandoval, J. et al. Genome-wide DNA methylation profiling predicts relapse in childhood B-cell acute lymphoblastic leukaemia. *Br. J. Haematol.* **160**, 406–409 (2013).
59. Rhein, P. et al. Gene expression shift towards normal B cells, decreased proliferative capacity and distinct surface receptors characterize leukemic blasts persisting during induction therapy in childhood acute lymphoblastic leukemia. *Leukemia* **21**, 897–905 (2007).
60. Oakes, C. C. et al. Evolution of DNA methylation is linked to genetic aberrations in chronic lymphocytic leukemia. *Cancer Discov.* **4**, 348–361 (2014).
61. Reinius, L. E. et al. Differential DNA methylation in purified human blood cells: implications for cell lineage and studies on disease susceptibility. *PLoS ONE* **7**, e41361 (2012).
62. Vento-Tormo, R. et al. IL-4 orchestrates STAT6-mediated DNA demethylation leading to dendritic cell differentiation. *Genome Biol.* **17**, 4 (2016).
63. Brönneke, S. et al. DNA methylation regulates lineage-specifying genes in primary lymphatic and blood endothelial cells. *Angiogenesis* **15**, 317–329 (2012).
64. Aryee, M. J. et al. Minfi: a flexible and comprehensive bioconductor package for the analysis of Infinium DNA methylation microarrays. *Bioinformatics* **30**, 1363–1369 (2014).
65. Maksimovic, J., Gordon, L. & Oshlack, A. SWAN: subset–quantile within array normalization for Illumina Infinium HumanMethylation450 BeadChips. *Genome Biol.* **13**, R44 (2012).
66. Bergmann, A. K. et al. DNA methylation profiling of pediatric B-cell lymphoblastic leukemia with *KMT2A* rearrangement identifies hypomethylation at enhancer sites. *Pediatr. Blood Cancer* **64**, 1–5 (2017).
67. Gabriel, A. S. et al. Epigenetic landscape correlates with genetic subtype but does not predict outcome in childhood acute lymphoblastic leukemia. *Epigenetics* **10**, 717–726 (2015).
68. Houseman, E. A. et al. DNA methylation arrays as surrogate measures of cell mixture distribution. *BMC Bioinformatics* **13**, 86 (2012).
69. Scott, D. W. & Gascoyne, R. D. The tumour microenvironment in B cell lymphomas. *Nat. Rev. Cancer* **14**, 517–534 (2014).
70. Teschendorff, A. E. & Relton, C. L. Statistical and integrative system-level analysis of DNA methylation data. *Nat. Rev. Genet.* **19**, 129–147 (2018).
71. Navarro, A. et al. Molecular subsets of mantle cell lymphoma defined by the *IGHV* mutational status and SOX11 expression have distinct biologic and clinical features. *Cancer Res.* **72**, 5307–5316 (2012).
72. Broyl, A. et al. Gene expression profiling for molecular classification of multiple myeloma in newly diagnosed patients. *Blood* **116**, 2543–2553 (2010).
73. Stunnenberg, H. G., International Human Epigenome Consortium & Hirst, M. The International Human Epigenome Consortium: a blueprint for scientific collaboration and discovery. *Cell* **167**, 1145–1149 (2016).
74. Debaize, L. et al. Interplay between transcription regulators RUNX1 and FUBP1 activates an enhancer of the oncogene c-KIT and amplifies cell proliferation. *Nucleic Acids Res.* **46**, 11214–11228 (2018).
75. Ernst, J. & Kellis, M. ChromHMM: automating chromatin-state discovery and characterization. *Nat. Methods* **9**, 215–216 (2012).
76. Beekman, R. et al. The reference epigenome and regulatory chromatin landscape of chronic lymphocytic leukemia. *Nat. Med.* **24**, 868–880 (2018).
77. Guyon, I., Weston, J., Barnhill, S. & Vapnik, V. Gene selection for cancer classification using Support Vector Machines. *Machine Learning* **46**, 389–422 (2002).
78. Dietrich, S. et al. Drug-perturbation-based stratification of blood cancer. *J. Clin. Invest.* **128**, 427–445 (2017).
79. Sánchez-Vega, F., Gotea, V., Margolin, G. & Elnitski, L. Pan-cancer stratification of solid human epithelial tumors and cancer cell lines reveals commonalities and tissue-specific features of the CpG island methylator phenotype. *Epigenetics Chromatin* **8**, 1–24 (2015).
80. Maura, F. et al. A practical guide for mutational signature analysis in hematological malignancies. *Nat. Commun.* **10**, 2969 (2019).

Acknowledgements

This research was funded by the European Union's Seventh Framework Programme through the Blueprint Consortium (grant agreement 282510); the European Research Council under the European Union's Horizon 2020 research and innovation program (Project BCLLATLAS, grant agreement 810287); Generalitat de Catalunya Suport Grups de Recerca AGAUR 2017-SGR-1142 (to E.C.) and 2017-SGR-736 (to J.I.M.-S.); Ministerio de Ciencia, Innovación y Universidades of the Spanish Government, grants RTI2018-094274-B-I00 (to E.C.) and SAF2017-86126-R (to J.I.M.-S.); Proyecto Medicina Personalizada PERMED (grant PMP15/00007), which is part of Plan Nacional de I+D+I and is cofinanced by the ISCIII-Sub-Directorate General for Evaluation and the European Regional Development Fund (FEDER; "Una manera de Hacer Europa"); CIBERONC (CB16/12/00225, CB16/12/00334, CB16/12/00236 and CB16/12/00489); the Accelerator award CRUK/AIRC/AECC joint funder-partnership; research funding from Fondo de Investigaciones Sanitarias, Instituto de Salud Carlos III PI17/01061 (S.B.); Ministerio de Ciencia, Innovación y Universidades, RTI2018-094274-B-I00, SAF2015-64885-R (E.C.); NIH grant number 1P01CA229100 (E.C.) and the European Regional Development Fund "Una manera de fer Europa", CERCA Programme/Generalitat de Catalunya. F.N. is supported by a predoctoral fellowship of the Ministerio de Economía y Competitividad (MINECO, BES-2016-076372). E.C. is an Academia Researcher of the Institució Catalana de Recerca i Estudis Avançats (ICREA) of the Generalitat de Catalunya. This work was partially developed at the Centro Esther Koplowitz (CEK, Barcelona, Spain). We thank F. Maura for his help with the analysis of mutational signatures.

Author contributions

M.D.-F., G. Clot, F.N. and R.B. performed DNA methylome, ChIP–seq, transcriptome, genetic, clinical and/or statistical analyses. T.B., J.N., Y.M.-Z., G.L., A.R.-D., S.M., R.O., G. Castellano, M.K., A.C.Q., S.-T.L., J.W., J.L., E.G., S.B., P.J., X.A., F.P., C.L.-O., X.S.P., C.C.O., T.Z., J.D., A.L.-G. and E.C. contributed to sample biological and/or clinical annotation. R.R. and M.P. provided computational support. M.D.-F. and J.I.M.-S. participated in study design. M.D.-F., F.N., R.B., T.B., J.D., A.L.-G., E.C. and J.I.M.-S. participated in data interpretation. J.I.M.-S. directed the research and wrote the manuscript together with M.D.-F.

Competing interests

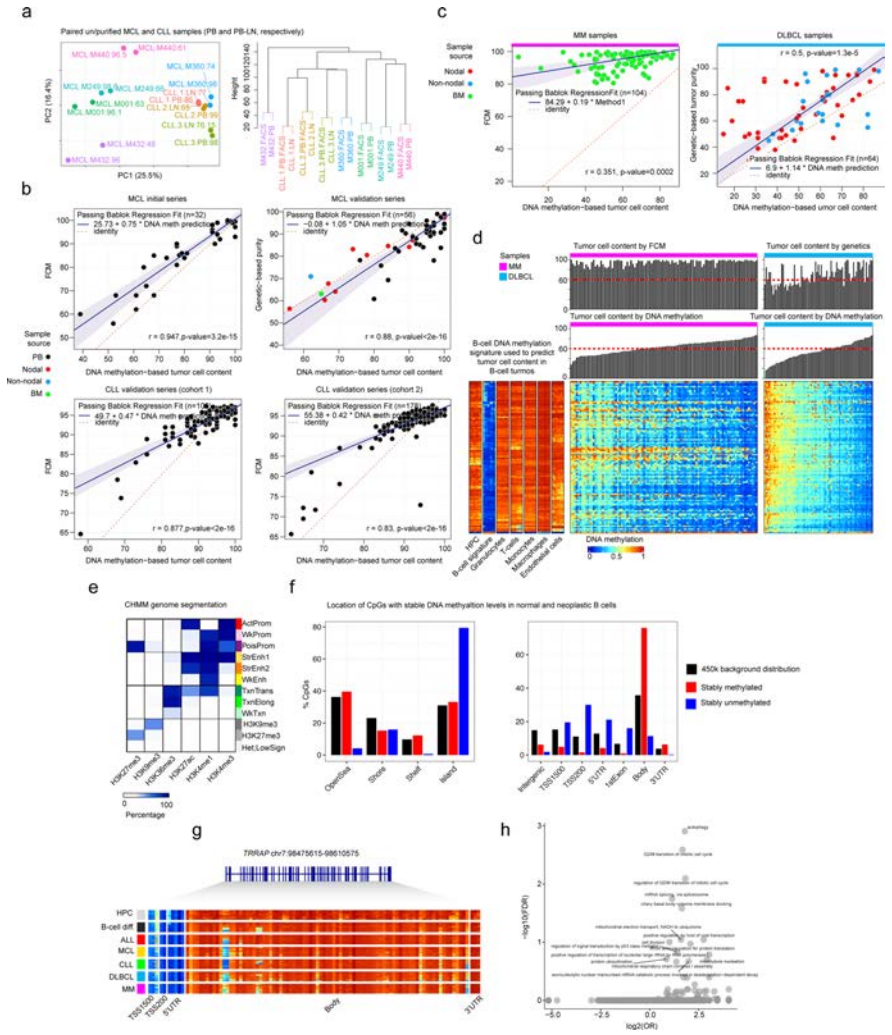
The authors declare no competing interests.

Additional information

Extended data is available for this paper at https://doi.org/10.1038/s43018-020-00131-2.

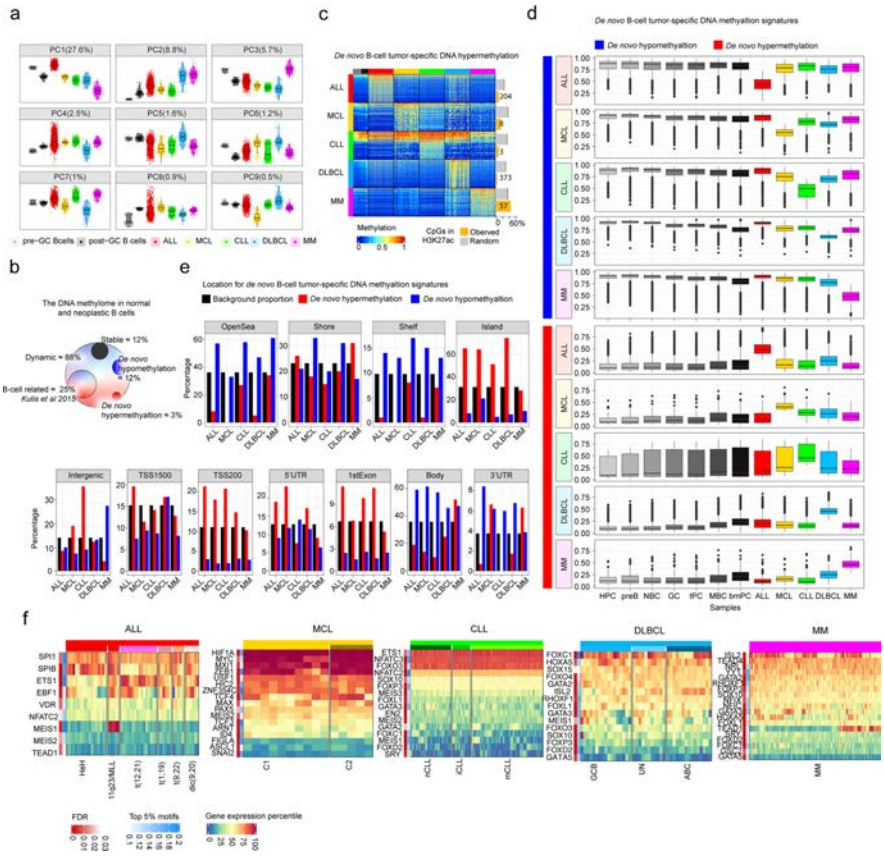
Supplementary information is available for this paper at https://doi.org/10.1038/s43018-020-00131-2.

Correspondence and requests for materials should be addressed to M.D.-F. or J.I.M.-S.

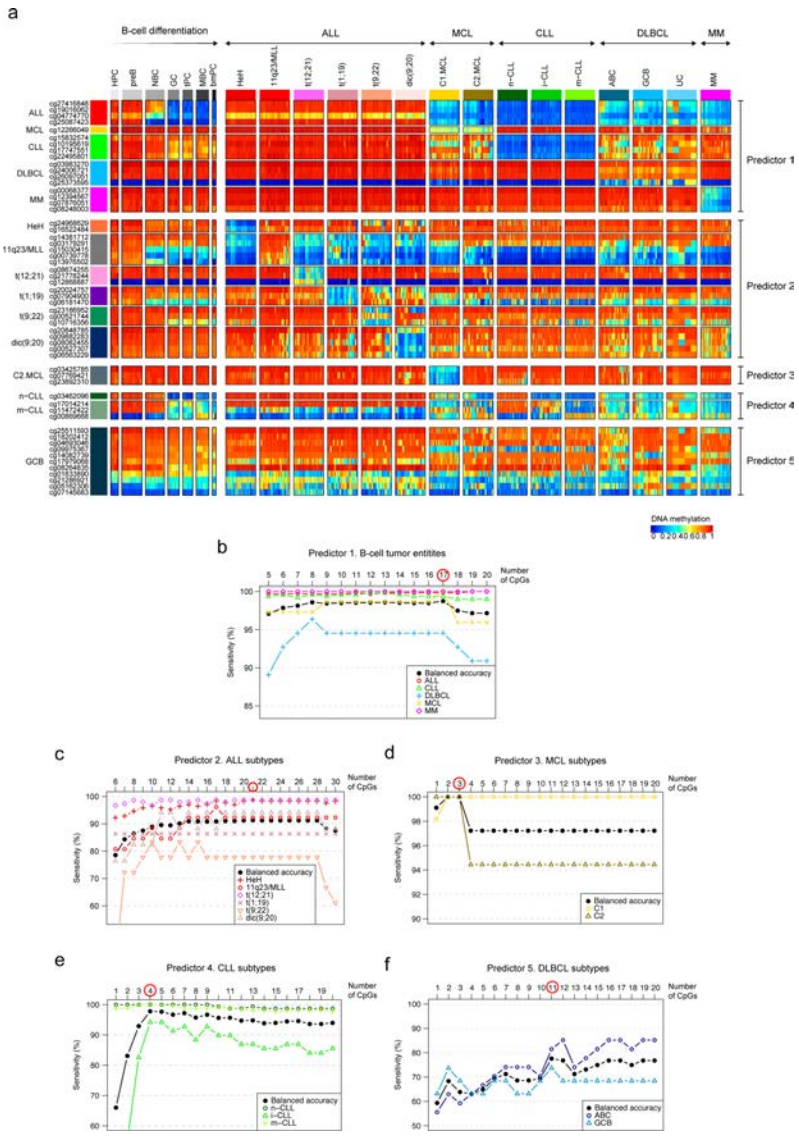


Extended Data Fig. 1 | See next page for caption.

Extended Data Fig. 1 | Analyses related to sample selection and annotation of stably-methylated CpGs. a, Principal component analysis and hierarchical clustering of synchronic unpurified/purified DNA methylation profiles obtained with EPIC array from MCL and CLL patients. Colors represent the same sample, with FCM-based purities highlighted in each sample. MCL, mantle cell lymphoma. CLL, chronic lymphocytic leukemia. **b**, Correlations and Passing Bablock regression fits of gold-standard methods for tumor purity prediction (FCM and genetic-based) against DNA methylation-based tumor purity prediction for MCL and CLL patients in initial and validation series. Samples sizes are: MCL initial series, n=32; MCL validation series, n=56; CLL cohort 1, n=109 and CLL cohort 2, n=178 patients. Shaded area represents 95% confidence intervals. Pearson correlation and derived p-values are also shown. **c**, Pearson correlations and Passing Bablock regression fits for gold-standard methods for tumor purity predictions (FCM and genetic-based) against DNA methylation-based tumor purity predictions in MM and DLBCL patients. Sample sizes are: MM, n=100 and DLBCL, n=55 patients and are the same as in panel **d**. Shaded area represents 95% confidence intervals. Pearson correlation and derived p-values are also shown. **d**, Pan-B cell DNA methylation signature used to deconvolute DNA methylation data and obtain B-cell tumor purities in B-cell tumors. The DNA methylation levels for the Pan-B-cell DNA methylation signature is shown for microenvironmental cells as well as MM and DLBCL. Bar plots representing DNA-methylation based predictions as well as gold standard-based predictions for MM and DLBCL are represented on the top of the heatmaps. **e**, Chromatin state genome segmentation with the CHMM software using the 6 histone marks used in the whole study for normal B cells, MCL, CLL and MM primary cases as well as for KARPAS-422 and SUDHL-5 DLBCL cells lines. **f**, Genomic distribution of stably methylated and unmethylated CpGs in normal and neoplastic B cell. Barplots represent single data values. **g**, Example gene showing stably unmethylated CpGs at promoters and stably methylated CpGs at gene body in normal and neoplastic B cells. A total of 98 CpGs are shown. **h**, Gene ontology analysis of genes showing both stably methylated and stably unmethylated CpGs in normal and neoplastic B cells.

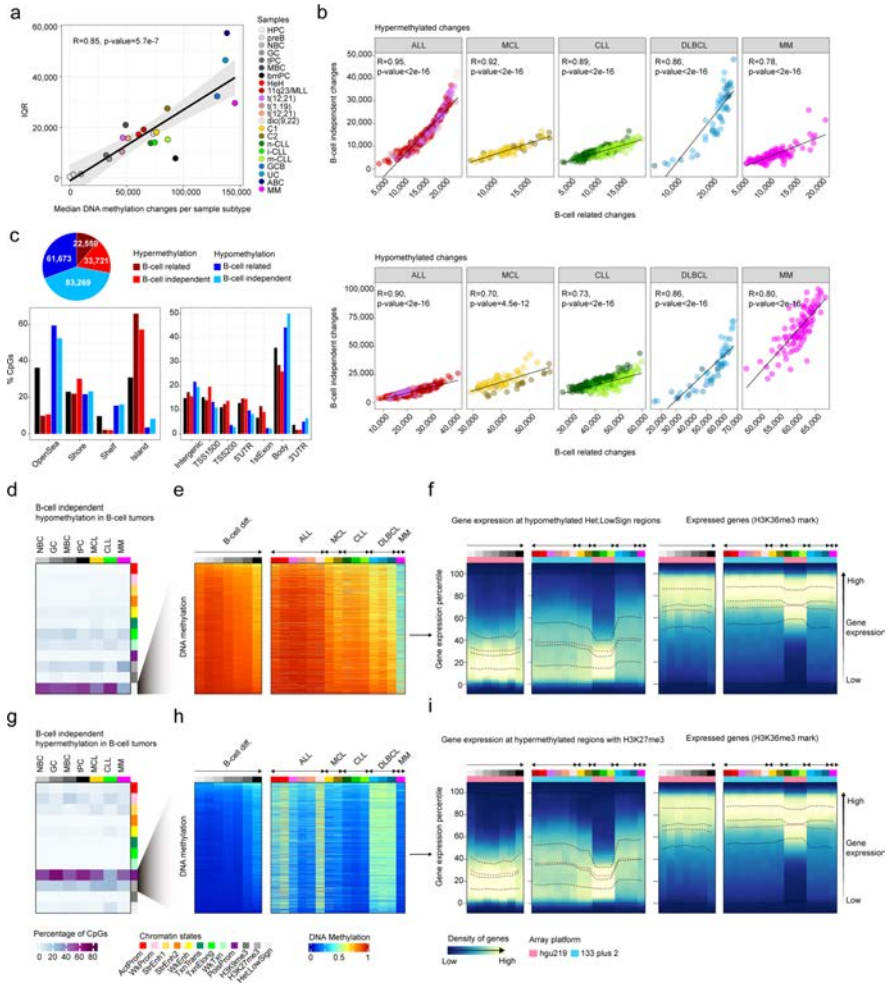


Extended Data Fig. 2 | Characterization of tumor-specific DNA methylation signatures. a, First 9 components of a Principal Component Analysis for normal and neoplastic B cells. Samples sizes are the same as in Fig. 1a. The same sample size applies also for panel **b**, **c** and **d**. **b**, Percentages of *de novo* DNA methylation signatures over the total DNA methylome. All *de novo* hyper- and hypomethylation from the five B-cell tumors analyzed are considered together to derive each respective percentage. **c**, Heatmap showing B-cell tumor-specific hypermethylation and the number of CpGs located at active regulatory regions (marked by H3K27ac). To calculate CpG enrichments in regulatory regions, the number of CpGs falling in regulatory regions were compared with the same number of *de novo* CpGs 10,000 times randomly chosen from the DNA methylome fraction with potential tumor-specific signatures falling in regulatory regions. **d**, Distribution of mean methylation levels of CpGs from *de novo* B-cell tumor-specific DNA methylation signatures across all normal and neoplastic B cell samples subtypes. The number of samples used to calculate the means is shown in Fig. 1a and the number of CpGs analyzed are those from Fig. 2b. **e**, Genomic distribution for *de novo* DNA methylation changes in B-cell tumors. Barplots represent single data values. **f**, Gene expression percentile of TFs showing the most significant p-values and frequencies for TFs binding site predictions (Methods) in *de novo* hypomethylation signatures in each B-cell tumor from Fig. 2d. Sample sizes for gene expression analyses in tumor samples are the same than in Fig. 4e. Center line, box limits, whiskers and points represent the median, 25th and 75th percentiles, 1.5x interquartile range and individual samples, respectively.



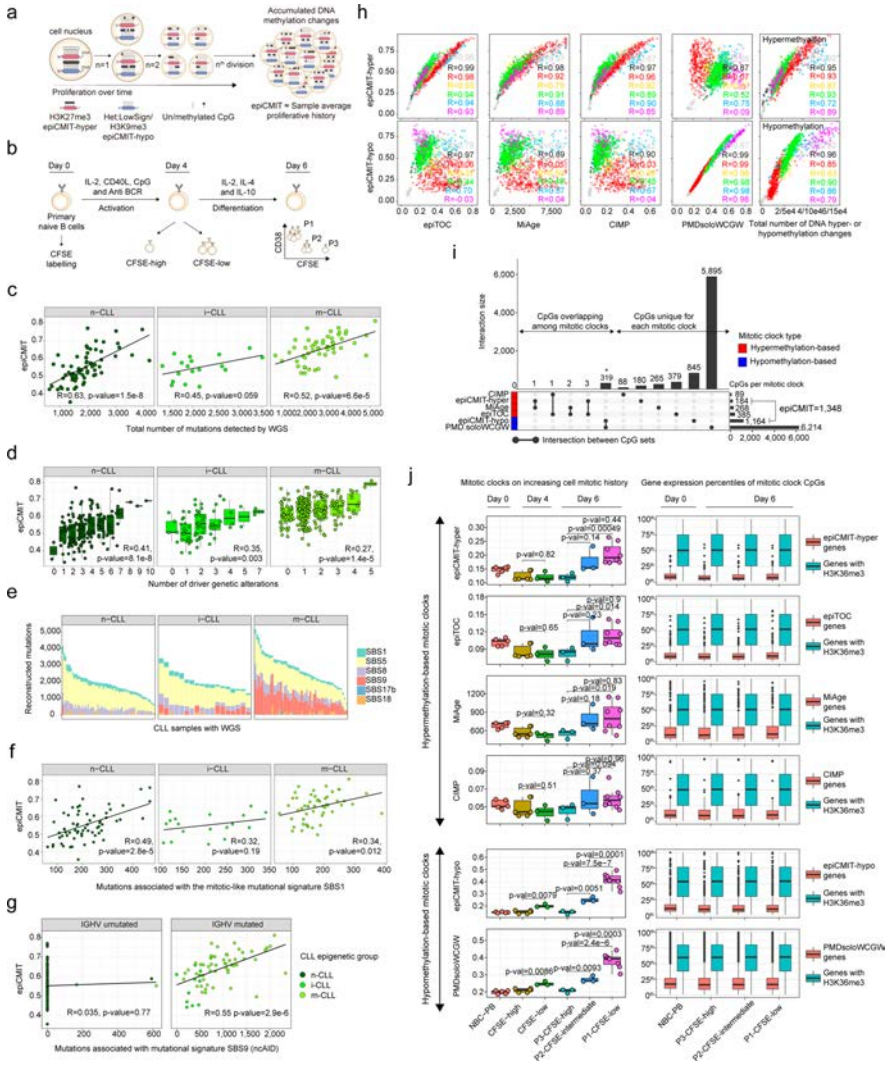
Extended Data Fig. 3 | See next page for caption.

Extended Data Fig. 3 | DNA methylation levels and analysis of the sensitivity of the epigenetic classifier of B cell neoplasms. a, DNA methylation levels of all CpGs from the pan-B-cell diagnostic algorithm in normal and neoplastic B cells. Sample sizes are the training samples shown in Fig. 3b. **b**, Estimated sensitivity according to the number of CpGs used in the pan-B-cell diagnostic algorithm for the classification of an unknown B-cell tumor into ALL, MCL, CLL, DLBCL or MM (first step of Fig. 3a, predictor 1). The number of CpGs selected for the predictor was chosen by maximizing the highest balanced accuracy and is indicated with a red circle. This strategy was applied also in the remaining 4 predictors to classify B-cell tumor subtypes in panels **c**–**f**, (second step of Fig. 3a). Each B-cell tumor is represented with different shapes and colors. **c**, Estimated sensitivity according to the number of CpGs used in the pan-B-cell diagnostic algorithm (predictor 2 of Fig. 3a) for the classification of ALL into the subtypes HeH, 11q23/MLL, t(12;21), t(1;19), t(9;22) and dic(9;20) while incrementing the number of CpGs (predictor 2 in Fig. 3a). **d**, Estimated sensitivity according to the number of CpGs used in the pan-B-cell diagnostic algorithm (predictor 3 of Fig. 3a) for the classification of MCL into the subtypes C1 or C2 while incrementing the number of CpGs (predictor 3 in Fig. 3a). **e**, Estimated sensitivity according to the number of CpGs used in the pan-B-cell diagnostic algorithm for the classification of CLL into the subtypes n-CLL, i-CLL or m-CLL while incrementing the number of CpGs (predictor 4 in Fig. 3a). **f**, Estimated sensitivity according to the number of CpGs used in the pan-B-cell diagnostic algorithm for the classification of DLBCL into the subtypes ABC and GCB while incrementing the number of CpGs (predictor 5 in Fig. 3a).



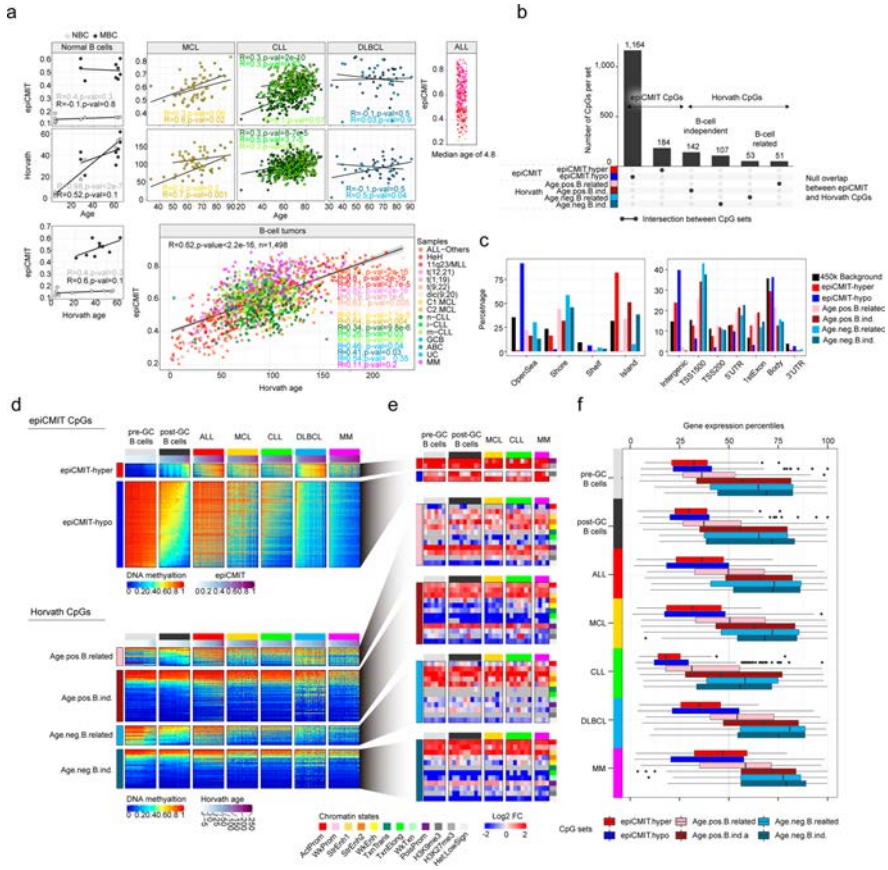
Extended Data Fig. 4 | See next page for caption.

Extended Data Fig. 4 | Further characterization of patient-specific DNA methylation changes. a, Variability of DNA methylation changes measured by the interquartile range (IQR) in normal and neoplastic B cells against the median number of DNA methylation changes per each subtype. R and p-values were derived from linear modelling. Shaded area represents 95% confidence interval. **b**, Correlations in all B cell tumors between B-cell independent DNA methylation changes and B-cell related changes for hypermethylation (top) and hypomethylation (bottom) changes. R and p-values were derived from linear models. **c**, Number of B-cell related or B-cell independent hyper- or hypomethylation in B-cell tumors showing consistent patterns (Methods). **d**, B-cell independent CpGs losing DNA methylation in B-cell tumors and the percentages of each chromatin state in normal and neoplastic B-cells. The mean of percentages per sample type is shown. The sample sizes are the same as in Fig. 4c and also apply for panel **g**. **e**, The mean of 2,000 representative CpGs per each sample subtype from panel **d** is represented. **f**, Gene density distributed along the expression percentiles of genes associated with B-cell independent CpGs losing DNA methylation at low signal heterochromatin in B-cell tumors. Expressed genes (H3K36me3) are displayed at right as control. Means within each B-cell subpopulation as well as B-cell tumors are represented. **g**, B-cell independent CpGs gaining DNA methylation in B-cell tumors and the percentages in each chromatin state in normal and neoplastic B-cells. **h**, The mean of 2,000 representative CpGs per each sample subtype from panel **g** is represented. **i**, Gene density distributed along the expression percentiles of genes associated with B-cell independent CpGs gaining DNA methylation at H3K27me3 regions in B-cell tumors. Expressed genes (H3K36me3) are displayed at right as control. Means within each B-cell subpopulation as well as B-cell tumors are represented. Sample size for DNA methylation analyzes in panels **a, b**, **c, e** and **h** are the same as in Fig. 4a. Samples sizes for gene expression analyses in panels **f** and **i** are the same as in Fig. 4e.

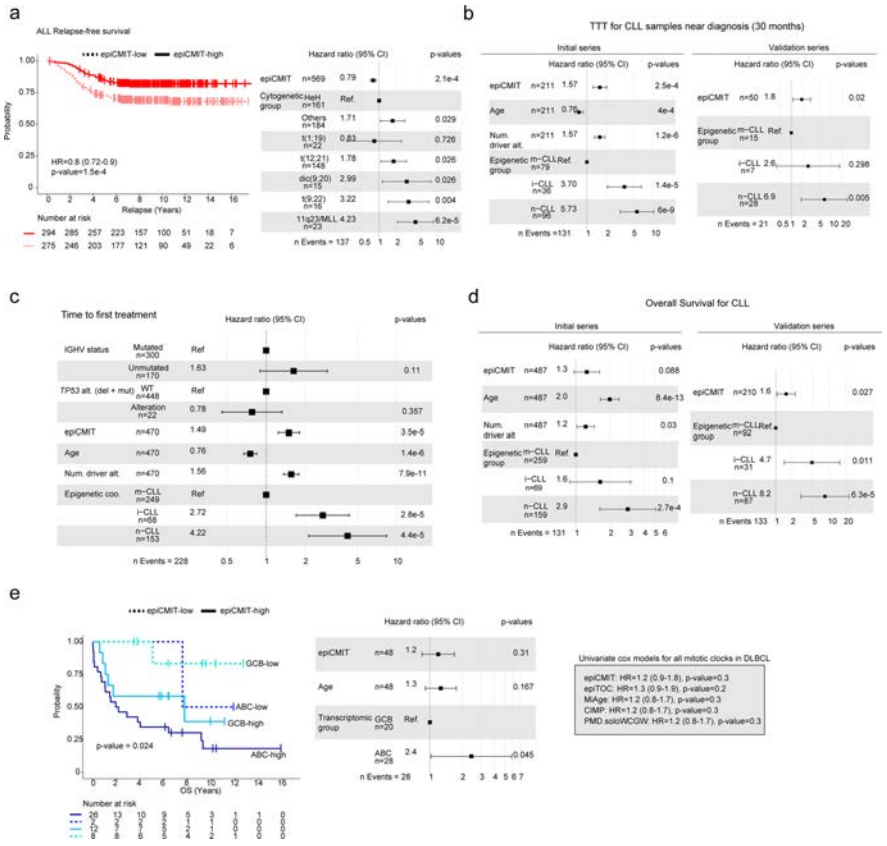


Extended Data Fig. 5 | See next page for caption.

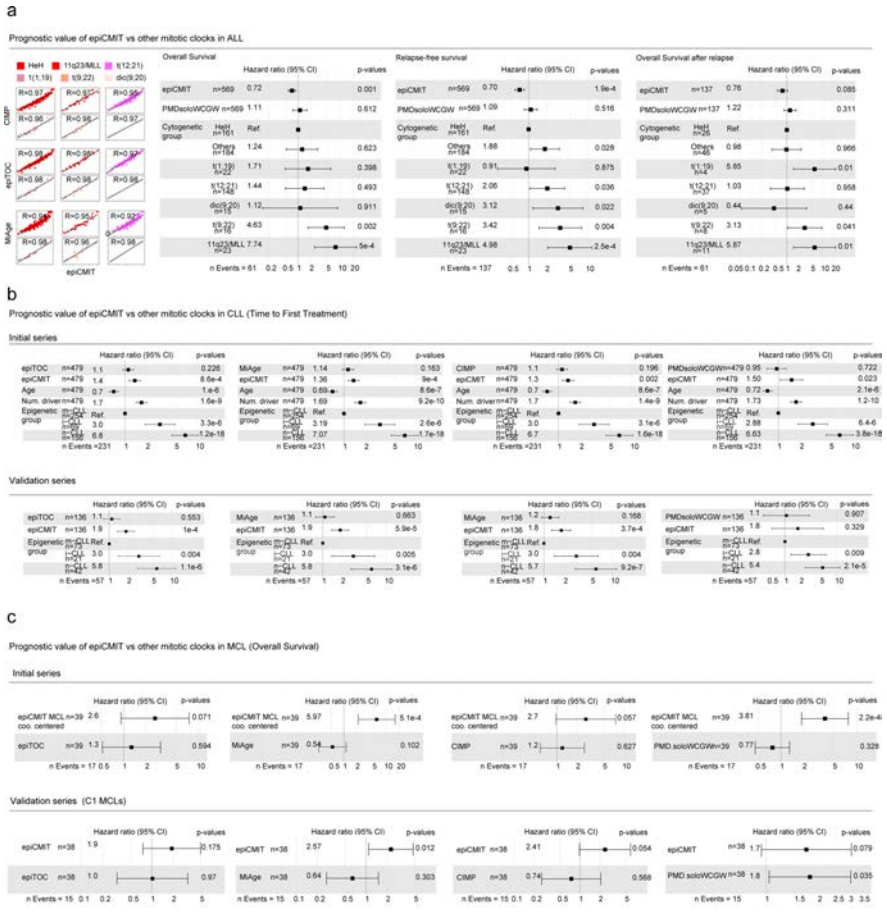
Extended Data Fig. 5 | Additional analyses performed to validate the epiCMIT. a, Illustrative scheme showing DNA methylation changes upon cell division and how they relate to epiCMIT scores. **b**, In vitro B-cell differentiation model used to experimentally validate the epiCMIT score. Primary naïve B cells are differentiated into plasma cells in 6 days. At day 0, primary human B cells are incubated with carboxyfluorescein succinimidyl ester (CFSE) and harvested with activation and proliferation cocktails necessary for plasma cell differentiation. The epiCMIT was calculated at day 0, day 4 and day 6 in B cells with different proliferative histories based on CFSE dilution. **c**, The epiCMIT is correlated with total number of mutations detected by WGS in each CLL epigenetic subtype. R and p-values are derived from linear modelling. 138 CLL patient samples with WGS and DNA methylation data are shown (66 n-CLL, 18 i-CLL and 54 m-CLL). The same sample size applies for panel **e**, **f** and **g**. **d**, The epiCMIT is correlated with CLL genomic complexity measured by the total number of driver alterations and thus with mutations with positive selection. Fitted linear regression models and derived R and p-values are shown for each group. The sample size for each number of driver alterations are: 0 drivers: n-CLL, n=2, i-CLL, n=5, m-CLL, n=44; 1 driver: n-CLL, n=14, i-CLL, n=19, m-CLL, n=119; 2 drivers: n-CLL, n=37, i-CLL, n= 25, m-CLL, n= 55; 3 drivers: n-CLL, n=38, i-CLL, n= 12, m-CLL, n=28; 4 drivers: n-CLL, n=27, i-CLL, n=4, m-CLL, n=12; 5 drivers: n-CLL, n=23, i-CLL, n=2, m-CLL, n=2; 6 drivers: n-CLL, n=10, i-CLL, n=0, m-CLL, n=0; 7 drivers: n-CLL, n=7, i-CLL, n=2, m-CLL, n=0; 8 drivers: n-CLL, n=1; 9 drivers: n-CLL, n=1; 10 drivers: n-CLL, n=1. **e**, Mutational signatures found in CLL with available WGS. CLL subtypes are shown separately. **f**, The epiCMIT is correlated with the mitotic-like mutational signature SBS1. CLL samples are divided in CLL epigenetic subgroups. R and p-values are derived from linear models. **g**, The epiCMIT is correlated with the mitotic-like mutational signatures SBS9. CLL samples are separated with the classical IGHV mutational status (98%). R and p-values are shown for each respective linear model. **h**, epiCMIT-hyper CpGs and epiCMIT-hypo mitotic clocks are compared with other hyper- or hypomethylation based mitotic clocks as well as the total number of hyper- (rightmost top) or hypomethylation (rightmost bottom) changes per sample since HPC stage. R from linear models are shown. Samples sizes are the same as in Fig. 4a. **i**, Overlap among the CpG used to build each mitotic clock. Barplots represent single data values. **j**, Performance of all mitotic clocks in the in vitro B-cell differentiation model from panel **c**. The fraction of epiCMIT which gain methylation (epiCMIT-hyper) and the fraction that lose DNA methylation (epiCMIT-hypo) were analyzed together with hyper- and hypomethylation-based mitotic clocks, respectively. Biological independent sample sizes are the same as in Fig. 5e. P-values are derived from two-sided t-tests and from biological independent experiments. On the right, expression of genes containing any CpG of each respective mitotic clock as well as genes containing CpGs in H3K36me3 regions are depicted (n=14,598). The number of genes analyzed per each mitotic clock are: epiCMIT-hyper, n=155; epiTOC, n=412; MiAge, n=298; CIMP, n=102; epiCMIT-hypo, n1,123; PMDsoloWCGW, n=4053. For the box plot, center line, box limits, whiskers and points represent the median, 25th and 75th percentiles, 1.5x interquartile range and individual samples, respectively.



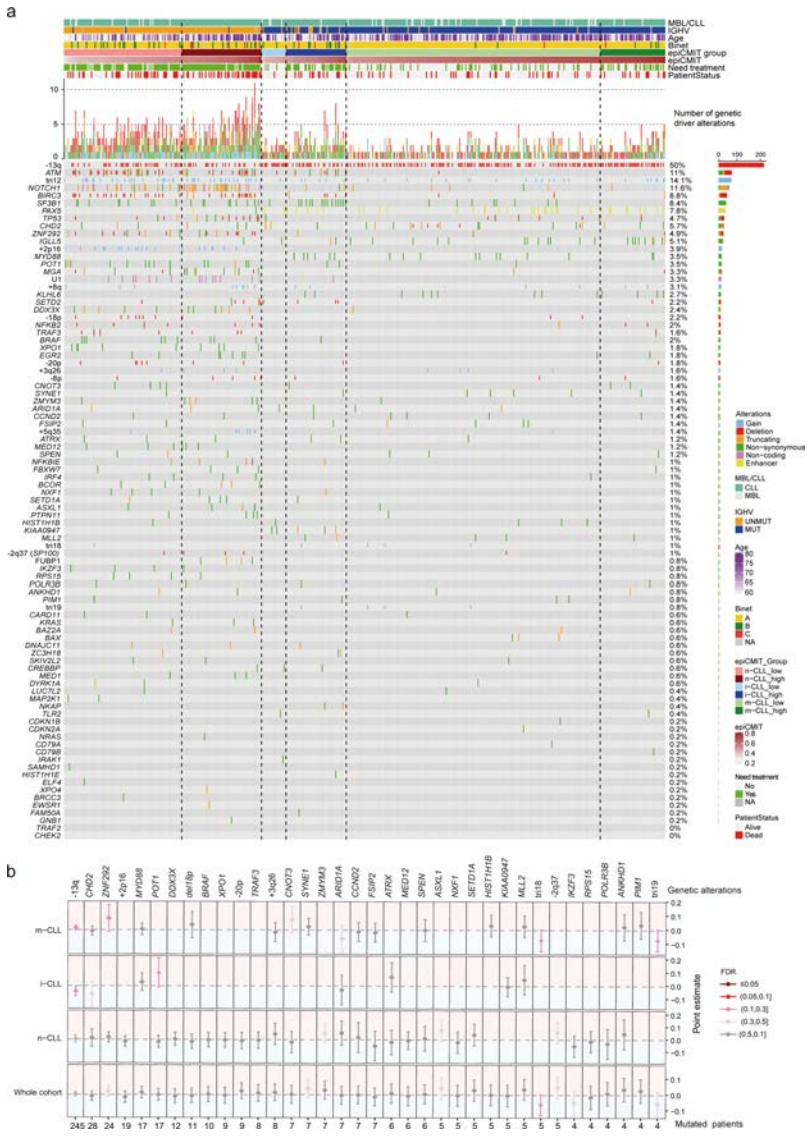
Extended Data Fig. 6 | Comparison between the epiCMIT mitotic clock and the Horvath aging clock. a, Correlations among epiCMIT, age and Horvath-predicted age in normal and neoplastic B cells. Samples sizes are: NBC, n=10 and MBC, n=9 donors; C1 MCL, n=40; C2 MCL, n=17; n-CLL, n=159; i-CLL, n=69; m-CLL, n=260; GCB DLBCL, n=20 and ABC DLBCL, n=28 patients. R and p-value are derived from linear models. Shaded areas represent 95% confidence intervals. **b**, epiCMIT and Horvath clocks do not have any CpG in common. CpGs of the Horvath model are divided into positively associated with age (gain of methylation) and negatively associated with age (loss of methylation). In addition, they are further classified into B-cell related or B-cell independent if they are extensively modulated or not during normal B-cell differentiation. Barplots represent single data values. **c**, The CpGs used to build the epiCMIT and Horvath clock show distinct genomic locations. Barplots represent single data values. **d**, DNA methylation levels of the CpGs from the epiCMIT and Horvath clocks in normal and neoplastic B cells. Sample sizes are the same as in Fig. 4a. **e**, The CpGs associated with the epiCMIT and Horvath clocks are located in markedly different chromatin states. Sample sizes are the same as in Fig. 4c. **f**, Genes associated with epiCMIT and Horvath CpGs show distinct transcriptional states in normal and neoplastic B cells. Gene probes shared across all normalized matrices from normal and neoplastic B cells were retained and were the following: epiCMIT-hyper, n=60; epiCMIT-hypo, n=327; Age positive B-cell related, n=44; Age positive B-cell independent, n=118; Age negative B-cell related, n=49; Age negative B-cell independent, n=101. For the box plot, center line, box limits, whiskers and points represent the median, 25th and 75th percentiles, 1.5x interquartile range and individual samples, respectively. Sample size are the same as in Fig. 4e.



Extended Data Fig. 7 | Additional characterization of the clinical impact of the epiCMIT in B cell tumors. a, Kaplan-Meier curves for relapse-free survival in ALL patients with low or high epiCMIT according to the maxstat rank statistics-based cutoff. Hazard ratio and p-value for the univariate Cox regression model are shown. A multivariate Cox regression model with epiCMIT as continuous variable and ALL cytogenetic groups is shown on the right. **b**, epiCMIT preserves its prognostic value in multivariate Cox regressions for time to first treatment in CLL patients whose samples were acquired at maximum 30 months after diagnosis both in initial and validation series. **c**, epiCMIT shows independent prognostic value from major prognostic variables in CLL including IGHV mutational status and *TP53* alterations (deletions and mutations) in multivariate Cox regressions for time to first treatment (TTT). **d**, Multivariate cox regression models in initial and validation CLL series for overall survival with epiCMIT and important prognostic variables. **e**, Kaplan-Meier curves for overall survival in GCB and ABC DLBCL patients with low or high epiCMIT according to the maxstat rank statistics-based cutoff. A multivariate Cox regression model with epiCMIT as continuous variable, the DLBCL subtype and age is shown on the right. On the right, univariate cox regression model for all mitotic clocks. All hazard ratios for epiCMIT correspond to 0.1 increments.



Extended Data Fig. 8 | Clinical impact of the epiCMIT as compared to other mitotic clocks. a, On the left, epiCMIT and hypermethylation-based mitotic clocks are highly correlated in ALL, creating a collinearity phenomenon in multivariate cox regression models with multiple mitotic clocks. On the right, multivariate Cox regression models with epiCMIT and PMDsoloWCGW mitotic clocks and ALL cytogenetic subgroups for overall survival, relapse-free survival and overall survival after relapse. **b**, In CLL, epiCMIT shows superior prognostic value in multivariate cox models for time to first treatment than all the other mitotic clocks in both initial and validation series. **c**, In MCL, epiCMIT shows an overall superior prognostic value in multivariate Cox models for overall survival in both initial series (with C1 and C2 MCL subtypes) and in the validation series, which only contain C1 MCL subtypes. In the initial series, MCL subtypes with different cellular origin were not introduced in multivariate Cox regression models due to few events, and thus the epiCMIT of each MCL patient was centered according to its cellular origin (C1 or C2) to account for normal B-cell development epiCMIT (Fig. 6a). Hazard ratios for the mitotic clocks correspond to scaled values and are comparable among them in each corresponding Cox model.



Extended Data Fig. 9 | See next page for caption.

Extended Data Fig. 9 | Additional data regarding the link between the epiCMIT and genetic changes in CLL. a, Oncoprint showing all genetic driver alterations considered in the whole CLL initial series composed by 490 CLL patient samples grouped by epigenetic subtypes and ordered according to increasing levels of epiCMIT (from left to right within each epigenetic subgroup). Other clinico-biological features including MBL or CLL, IGHV status, Age, Binet stage, epiCMIT subgroups based on maxstat rank statistic, need for treatment and patient status are shown. Distinct genetic driver alterations are depicted with different colors and shapes. The percentage of mutated patients and number of mutated patients for each alteration is shown at right. **b**, Driver genetic alterations without clear associations with epiCMIT. Analyses were done in the whole cohort as well as within each epigenetic subgroup. Point estimates with 95% confidence intervals were derived in the whole cohort using linear modelling between epiCMIT and alterations adjusted for CLL subtypes, and with two-sided t-tests within CLL subtypes. Point estimates then represent the coefficient of each respective alteration in each corresponding linear model (whole cohort analysis) or the difference between means (CLL subtypes analysis). Point estimates are color-coded according to FDR correction. Treated and untreated patients at the moment of sampling were considered for these analyses.

nature research

Corresponding author(s): M. Duran-Ferrer
J.I. Martin-Subero

Last updated by author(s): Aug 28, 2020

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection: No software was used to collect data.

Data analysis

All the analyses in the study have been performed using R statistical language (version 3.6.3) under linux (Ubuntu 14.04 LTS and Red Hat Enterprise Linux Server 7.5) and/or OS (10.14.6) environments.
Base core R and Bioconductor (version 3.1) packages as well as custom R scripts were used to mine, analyze and plot the data. ggplot2, circlize and ComplexHeatmap packages were used to plot data. Minfi (v. 1.32) R package was used to normalize DNA methylation data. missMethyl package was used to perform Gene Ontology analysis using CpGs. irlba and package and prcomp() function from stats package were used to perform PCA analyses for DNA methylation data. limma (3.42.2) package was used to perform differential methylation and expression analyses.
PWMEnrich was used to perform TF binding site prediction analyses.
e1071 package was used to built the diagnostic algorithm.
hdp and MutationalPatterns packages were used to extract mutational signatures.
Gene expression data was mined and normalized using affy package (version 1.64). fgsea package was used to perform GSEA analysis in R using C2 curated signatures from Molecular Signatures Database v7.0.
The Allele-Specific Copy Number Analysis of Tumors (ASCAT) algorithm available at Nexus Copy Number (BioDiscovery, version 7) was used to infer the tumor purity directly from the Cytoscan HD array.
NALM6 ALL cell line ChIP-seq data quality was checked using fastqc and igv, and analyzed using bwa (version 0.7.17), PICARD tools (version 2.20.2), SAMTOOLS (version 1.9), PhantomPeakQualTools and MACS2 (version 2.1.2). chromHMM was used to generate chromatin states for normal B cells, MCL, CLL, MM and KARPAS-422 and SUDHL-5 DLBCL cell lines by integrating ChIP-seq of 6 histone marks including H3K27ac, H3K4me3, H3K4me1, H3K36me3, H3K27me3 and H3K9me3.
Survival and survminer R packages were used to analyze and display clinical associations.

The source code for the DNA methylation classifier of B-cell tumors entities and subtypes as well as for the calculation of the epiCMIT mitotic clock can be found at https://github.com/Duran-FerrerM/Pan-B-cell-methylome. All other source code supporting the findings of this study are available from the corresponding author on reasonable request.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

The accession codes containing previously generated as well as newly generated data are the following:

450k and EPIC DNA methylation data:
Normal B cells: EGAS00001001196.
ALL: GSE16368, GSE47051, GSE7658515, GSE6922916.
MCL: EGAS00001001637, EGAS00001004165.
CLL: EGAS00001004640, EGAD00010000871, EGAD00010000948.
DLBCL: EGAS00001004640.
MGUS and MM: EGAS00001000841.
In-vitro B-cell differentiation model of naïve B cells from human primary samples: GSE72498.

Gene expression data:
Normal B cells: EGAS00001001197.
ALL:GSE47051.
MCL: GSE36000.
CLL: EGAS00000000092, EGAD00010000254.
DLBCL: EGAS00001004640.
MM:GSE19784.
In-vitro B-cell differentiation model of naïve B cells from human primary samples: GSE72498.

ChIP-seq data:
ALL: GSE109377.
Normal B cells, MCL, CLL and MM: EGAS00001000326.

The normalized DNA methylation matrices used for this study are available upon resonable request at: http://resources.idibaps.org/paper/the-proliferative-history-shapes-the-DNA-methylome-of-B-cell-tumors-and-predicts-clinical-outcome

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	Sample size was not determined by calculations, but by sample availability and data quality. 1,799 samples were initally considered for the inital analyses. After all filtering criteria, we removed 104 ALL samples, 8 MCL samples, 1 CLL sample, 25 DLBCL samples and 4 MM samplesas well as microenvironmental cells. We retained 1,595 samples , including 67 normal B cells spanning the entire maturation process, 809 ALL samples (including 2 independent cohorts), 74 MCLs, 490 CLLs, 55 DLBCLs and 100 MMs. For the pan-B-cell diagnostic algorithm, we use 1345 samples in the training series (626 ALL, 74 MCL, 490 CLL, 55 DLBCL, 100MM), and 711 samples for the validation series (261 ALL, 58 MCL, 380 CLL and 12 DLBCL). For the clinical associations, we use 569 ALL, 479 CLL, 39 MCL and 48 DLBCL samples. Additionally, we use as a validation series 210 CLL and 38 MCL samples.
Data exclusions	Samples with tumor cell content below 60% were excluded. Selection criterion was pre-established and experimentaly validated.
Replication	Experimental replication was not atempted.
Randomization	Due to the experimental design and goals of our article randomization was not meaningful.
Blinding	Due to the experimental design and goals of our article sample blinding was not meaningful.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology
☒	☐ Animals and other organisms
☐	☒ Human research participants
☒	☐ Clinical data

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Human research participants

Policy information about studies involving human research participants

Population characteristics	This is an analysis article using mostly epigenetic and clinical data reported in previous publications.
Recruitment	Individual studies from which data were obtained for this analysis article explain how patients were recruited. In general, all the studies recruited patients for methylation studies based on sample availability in their biobanks.
Ethics oversight	The study was approved by the etihcs committees of the institutions that generated the data used in the present study.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Check for updates

news & views

CANCER EPIGENETICS

A ticking clock for B cell tumors

Tumor-specific changes in DNA methylation are both acquired actively through transcription-coupled processes and passively accumulated over time. Analysis across B cell malignancies now shows that these changes provide insight into the cellular origin as well as the proliferative history of tumors and thereby have diagnostic value and prognostic value, respectively.

Paolo Strati and Michael R. Green

B cells undergo a complex multi-step process of differentiation that is regulated at the molecular level by dynamic epigenetic changes, including DNA methylation, that govern gene-expression programs that control cell fate decisions, proliferation and survival. Deregulation of these programs by somatic mutations and other factors can drive transformation of B cells at various stages of differentiation, which results in malignancies with molecular and immunophenotypic similarities to an inferred normal B cell counterpart. For example, B cell acute lymphoblastic leukemia (B-ALL) shares features with B cell progenitors, mantle-cell lymphoma (MCL) shares features with mantle-zone B cells, chronic lymphocytic leukemia (CLL) shares features with naive or memory B cells, diffuse large B cell lymphoma (DLBCL) shares features with germinal-center B cells, and multiple myeloma (MM) shares features with plasma cells[1]. Active changes in DNA methylation associated with differentiation are focal, are localized to regulatory elements and gene bodies and are accompanied by passive changes that are more diffuse, not linked with gene expression and, in part, associated with mitosis. The DNA-methylation landscape of B cell malignancies is a product of these active and passive changes and can therefore inform the differentiation state and proliferative history of the tumor, as well as encompassing other tumor-specific alterations[2]. In this issue of *Nature Cancer*, Duran-Ferrer et al. report their analysis of the DNA-methylation signatures of 1,595 samples across a spectrum of B cell malignancies showing that changes in DNA methylation have both diagnostic implications and prognostic implications[3]. Specifically, signatures of both hypermethylation and hypomethylation in tumors were able to accurately classify disease entities, as well as more-discrete molecular subtypes, and were leveraged to develop a DNA methylation–based mitotic 'clock', epigenetically determined cumulative mitoses (epiCMIT) (Fig. 1), that showed independent prognostic value in multiple B cell malignancies by capturing the proliferative history of tumor cells.

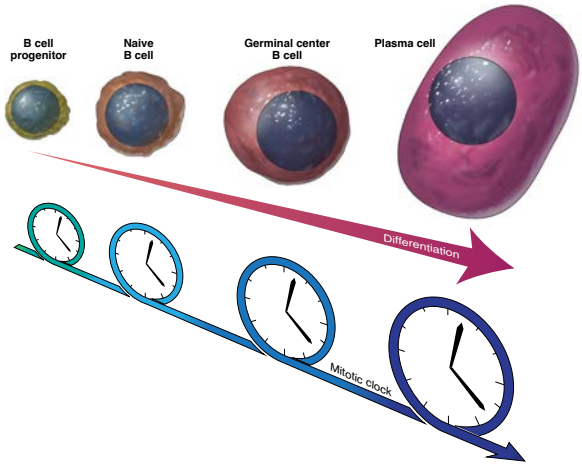


Fig. 1 | The epiCMIT clock. epiCMIT captures DNA methylation–based readouts of mitotic history associated with normal B cell differentiation and the proliferation of malignant B cells to derive a prognostically significant measure of cumulative mitoses. The contribution of normal B cell differentiation to methylation changes means that malignancies that align to different normal counterparts (for example, B-ALL with B cell progenitors, CLL with naive B cells, DLBCL with germinal-center B cells, and MM with plasma cells) have uniquely derived epiCMIT scores that should be compared only to tumors from the same maturation state.

Through analysis of high-density array data, the methylation states of 88% of measured CpGs were found to be labile in B cell malignancies, corresponding to 25% that were altered in association with normal B cell differentiation, reflective of the association of individual disease subtypes with distinct cellular origins[1]. The additional 63% of CpG islands were altered in a disease-specific fashion compared with CpG islands in their normal B cell counterparts. This striking variability can be mined for insight into the role of DNA methylation in the biology of this class of tumors. For example, assessing hypomethylated regions that overlapped B cell enhancers (marked by the active histone modification H3K27Ac), the authors defined disease-specific patterns of enrichment for motifs of transcription

factors known to be linked with disease pathogenesis, such as the transcription factor NFAT in CLL; this suggested that binding of transcription factors may contribute to altered methylation patterns. Furthermore, 94 genes with disease-specific patterns of DNA hypomethylation were shown to have corresponding changes in gene expression. These data support a role for actively acquired changes in focal DNA methylation, localized to regulatory elements and gene bodies in association with cellular cues and transitions, in the etiology of different B cell malignancies. However, further understanding of the processes involved in regulating the DNA-methylation landscape, and their implications for disease biology, will require integration of DNA-methylation patterns with tumor-matched measurements of other epigenetic marks, such as histone post-translational modifications. This will be particularly revealing for tumors with mutations in genes encoding chromatin-modifying molecules[4]. For example, one might expect that inactivating mutations of *CREBBP* (which encodes the acetyltransferase CREBBP) in germinal-center B cell–derived malignancies would be associated with increased DNA methylation at enhancers with lower H3K27Ac[5]. As different classes of *CREBBP* mutations confer varying magnitudes of H3K27Ac loss[6], DNA-methylation levels at CREBBP-regulated enhancers could conceivably represent an easily assayed readout for the severity of targetable mutation-associated changes in enhancer activity.

Differentially methylated regions associated with regulatory elements and gene bodies represent only a subset of the broad disease-specific changes in methylation. Furthermore, each B cell malignancy can be further subcategorized into molecular subtypes. Duran-Ferrer et al. therefore leveraged this large dataset to construct a two-step classifier that was able to discriminate between diseases (ALL, MCL, CLL, DLBCL or MM) and further subclassify cytogenetic subtypes of ALL, epigenetically defined subtypes of MCL (cluster 1 or cluster 2) and CLL (low programmed, intermediate programmed or high programmed), and transcriptionally defined subtypes of DLBCL (activated B cell–like (ABC-like) or germinal-center B cell–like (GCB like) DLBCL) exclusively on the basis of a subset of CpG biomarkers[3]. This classifier may have utility in cases in which features shared by diseases can make differential diagnosis difficult, such as CLL and MCL, or in the classification of more-complex molecular subtypes, such as ABC-like and GCB-like subtypes of DLBCL. Additional studies will be needed to define whether DNA methylation may also be informative for the discrimination of other prognostically important subtypes, such as the molecular high-grade signatures of B cell lymphoma[7], that were not included in this dataset. Furthermore, although this approach had good success in discriminating between cytogenetic subtypes of B-ALL, these subtypes also have distinctive transcriptional phenotypes[8]. It is therefore uncertain whether DNA-methylation signatures will be informative for etiological subtypes defined by a more-complex genetic architecture that is not accompanied by strong transcriptional phenotypes, such as the recently described genetic subtypes of DLBCL[9,10]. If they are informative in this context, DNA-methylation signatures may be a more streamlined approach for the classification of complex molecular subtypes and could potentially 'rescue' a fraction of the cases that remain unclassified by genetics alone.

The majority of changes in DNA methylation are not directly linked to changes in gene expression and are probably passively acquired or linked to processes that are yet to be defined. Duran-Ferrer et al. identified patient-specific patterns of DNA hypermethylation and hypomethylation by calculating changes with reference to hematopoietic stem cells[3]. This served to provide a common landmark for all B cell malignancies and to define four categories of methylation alterations that were linked either with B cell differentiation–associated changes or with tumor-specific changes. Regardless of their class, these changes were located mainly in silent chromatin regions and thus were not associated with transcription. Such DNA-methylation changes are accumulated during cell division and thus these patterns have been employed to derive methylation-based mitotic clocks[11]. Duran-Ferrer et al. first used data obtained by chromatin immunoprecipitation followed by deep sequencing to identify regions of interest that were constitutively silenced and/or marked with repressive histone modification H3K27me3, and then integrated those with DNA-methylation data from human hematopoietic stem cells and bone marrow–derived plasma cells to identify CpG islands that changed in methylation state between cells with the lowest (hematopoietic stem cells) and highest (bone marrow–derived plasma cells) proliferative histories[3]. These were used to derive two 'clocks' of epiCMIT: one representing hypermethylation and one representing hypomethylation. The clock with the highest value was used as the epiCMIT score for each tumor, which thereby enabled the capture of differential tendencies toward either gain of methylation or loss of methylation. Notably, the epiCMIT score not only can reveal the proliferative history of a particular tumor but also relates to the proliferative capacity at the time of analysis, as measured by staining of the proliferation marker Ki67 in MCL or gene-expression profiling in CLL. In line with this, the epiCMIT score was a strong independent prognostic factor in cohorts of patients with ALL, CLL or MCL. In CLL, epiCMIT was associated with the presence of genetic driver alterations, which suggests that this methylation-based proliferative score captures important biological variables that influence patients' outcomes. epiCMIT may therefore represent an important prognostic tool that could be broadly applied across B cell malignancies, but it still requires validation in subtypes such as DLBCL, follicular lymphoma and high-grade B cell lymphomas, for which data were not available in this study. Furthermore, recent advances have allowed the investigation of tumor-specific DNA-methylation patterns in the cell-free DNA compartment of peripheral blood[12]. Therefore, the signatures described in this study may serve as an important first step toward the future development of minimally invasive diagnostic and prognostic assays that use cell-free DNA-methylation patterns in patients with B cell malignancies.

Collectively, Duran-Ferrer et al. have provided important insights into the mechanisms of DNA-methylation changes in B cell malignancies, as well as their diagnostic and prognostic implications[3]. These findings provide a framework for the future application of DNA methylation as a clinically relevant biomarker for B cell malignancies. Although it is beyond the scope of the current study, actionable phenotypes may also be associated with DNA-methylation signatures that could be employed as biomarkers for therapy selection. For example, activation of transcription factors downstream of B cell receptor signaling may result in a unique DNA-methylation signature[13], which could support the use of inhibitors of the kinase BTK. Furthermore, polycomb-repressed genes are hypermethylated in B cell malignancies[5,14], which suggests that hyperactivity of the histone methyltransferase EZH2 may be accompanied by a DNA-methylation signature, and this could support the use of EZH2 inhibitors. Exploration of these more pathway- and/or target-oriented signatures could aid in therapeutic prioritization and could be a strong future addition

to the proven diagnostic ability of DNA methylation demonstrated in this study. ❒

Paolo Strati[1,2] and Michael R. Green[1,3]✉
[1]*Department of Lymphoma & Myeloma, University of Texas MD Anderson Cancer Center, Houston, TX, USA.* [2]*Department of Translational Molecular Pathology, University of Texas MD Anderson Cancer Center, Houston, TX, USA.* [3]*Department of Genomic Medicine, University of Texas MD Anderson Cancer Center, Houston, TX, USA.*
✉*e-mail: mgreen5@mdanderson.org*

Published online: 02 November 2020
https://doi.org/10.1038/s43018-020-00132-1

References

1. Seifert, M., Scholtysik, R. & Küppers, R. *Methods Mol. Biol.* **1956**, 1–33 (2019).
2. Oakes, C. C. & Martin-Subero, J. I. *Blood* **132**, 999–1006 (2018).
3. Duran-Ferrer, M. et al. *Nat. Cancer* https://doi.org/10.1038/s43018-020-00131-2 (2020).
4. Lunning, M. A. & Green, M. R. *Blood Cancer J.* **5**, e361 (2015).
5. Kulis, M. et al. *Nat. Genet.* **47**, 746–756 (2015).
6. Mondello, P. et al. *Cancer Discov.* **10**, 440–459 (2020).
7. Sha, C. et al. *J. Clin. Oncol.* **37**, 202–212 (2019).
8. Figueroa, M. E. et al. *J. Clin. Invest.* **123**, 3099–3111 (2013).
9. Chapuy, B. et al. *Nat. Med.* **24**, 679–690 (2018).
10. Wright, G. W. et al. *Cancer Cell* **37**, 551–568.e514 (2020).
11. Horvath, S. & Raj, K. *Nat. Rev. Genet.* **19**, 371–384 (2018).
12. Shen, S. Y. et al. *Nature* **563**, 579–583 (2018).
13. Oakes, C. C. et al. *Nat. Genet.* **48**, 253–264 (2016).
14. Kretzmer, H. et al. *Nat. Genet.* **47**, 1316–1325 (2015).

Competing interests

The authors declare no competing interests.

Manuscript 5

*Nadeu F, *Martin-Garcia D, Clot G, Díaz-Navarro A, **Duran-Ferrer M**, Navarro A, Vilarrasa-Blasi R, Kulis M, Royo R, Gutiérrez-Abril J, Valdés-Mas R, López C, Chapaprieta V, Puiggros M, Castellano G, Costa D, Aymerich M, Jares P, Espinet B, Muntañola A, Ribera-Cortada I, Siebert R, Colomer D, Torrents D, Gine E, López-Guillermo A, Küppers R, Martin-Subero JI, Puente XS, #Beà S, #Campo E. Genomic and epigenomic insights into the origin, pathogenesis, and clinical behavior of mantle cell lymphoma subtypes. Blood. 2020 Sep 17;136(12):1419-1432. doi: 10.1182/blood.2020005289. PMID: 32584970; PMCID: PMC7498364.

blood®
Regular Article

LYMPHOID NEOPLASIA

CME Article

Genomic and epigenomic insights into the origin, pathogenesis, and clinical behavior of mantle cell lymphoma subtypes

Ferran Nadeu,[1,2,*] David Martin-Garcia,[1,2,*] Guillem Clot,[1,2] Ander Díaz-Navarro,[2,3] Martí Duran-Ferrer,[1] Alba Navarro,[1,2] Roser Vilarrasa-Blasi,[1] Marta Kulis,[1] Romina Royo,[4] Jesús Gutiérrez-Abril,[3] Rafael Valdés-Mas,[3] Cristina López,[1,5] Vicente Chapaprieta,[1] Montserrat Puiggros,[4] Giancarlo Castellano,[6] Dolors Costa,[7] Marta Aymerich,[1,2,7] Pedro Jares,[1,7,8] Blanca Espinet,[9] Ana Muntañola,[10] Inmaculada Ribera-Cortada,[7,11] Reiner Siebert,[5] Dolors Colomer,[1,2,7,8] David Torrents,[4] Eva Gine,[2,7] Armando López-Guillermo,[1,2,7,8] Ralf Küppers,[12,13] Jose I. Martin-Subero,[1,2,8,14] Xose S. Puente,[2,3] Sílvia Beà,[1,2,7,8,†] and Elias Campo[1,2,7,8,†]

[1]Institut d'Investigacions Biomèdiques August Pi i Sunyer (IDIBAPS), Barcelona, Spain; [2]Centro de Investigación Biomédica en Red de Cáncer, Madrid, Spain; [3]Departamento de Bioquímica y Biología Molecular, Instituto Universitario de Oncología, Universidad de Oviedo, Oviedo, Spain; [4]Barcelona Supercomputing Center, Barcelona, Spain; [5]Institute of Human Genetics, Ulm University and Ulm University Medical Center, Ulm, Germany; [6]Unitat de Genòmica, IDIBAPS, Barcelona, Spain; [7]Hospital Clínic of Barcelona, Barcelona, Spain; [8]Departament de Fonaments Clinics, Universitat de Barcelona, Barcelona, Spain; [9]Laboratori de Citogenètica Molecular, Servei de Patologia, Hospital del Mar, Barcelona, Spain; [10]Servei d'Hematologia, Hospital Mútua de Terrassa, Terrassa, Spain; [11]Hospital Nostra Senyora de Meritxell, Escaldes-Engordany, Andorra la Vella, Andorra; [12]Institute of Cell Biology (Cancer Research), University of Duisburg-Essen, Essen, Germany; [13]German Consortium for Cancer Research, Heidelberg, Germany; and [14]Institució Catalana de Recerca i Estudis Avançats, Barcelona, Spain

KEY POINTS

- **The primary *CCND1* rearrangement is mediated by the same mechanisms in cMCL and nnMCL, but they differ in the (epi)genetic and driver makeup.**
- **Genomic complexity and DNA methylation changes related to proliferative cell history stratify patients with distinct clinical outcomes.**

Mantle cell lymphoma (MCL) is a mature B-cell neoplasm initially driven by *CCND1* rearrangement with 2 molecular subtypes, conventional MCL (cMCL) and leukemic non-nodal MCL (nnMCL), that differ in their clinicobiological behavior. To identify the genetic and epigenetic alterations determining this diversity, we used whole-genome (n = 61) and exome (n = 21) sequencing (74% cMCL, 26% nnMCL) combined with transcriptome and DNA methylation profiles in the context of 5 MCL reference epigenomes. We identified that open and active chromatin at the major translocation cluster locus might facilitate the t(11;14)(q13;32), which modifies the 3-dimensional structure of the involved regions. This translocation is mainly acquired in precursor B cells mediated by recombination-activating genes in both MCL subtypes, whereas in 8% of cases the translocation occurs in mature B cells mediated by activation-induced cytidine deaminase. We identified novel recurrent MCL drivers, including *CDKN1B, SAMHD1, BCOR, SYNE1, HNRNPH1, SMARCB1*, and *DAZAP1*. Complex structural alterations emerge as a relevant early oncogenic mechanism in MCL, targeting key driver genes. Breakage-fusion-bridge cycles and translocations activated oncogenes (*BMI1, MIR17HG, TERT, MYC*, and *MYCN*), generating gene amplifications and remodeling regulatory regions. cMCL carried significant higher numbers of structural variants, copy number alterations, and driver changes than nnMCL, with exclusive alterations of *ATM* in cMCL, whereas *TP53* and *TERT* alterations were slightly enriched in nnMCL. Several drivers had prognostic impact, but only *TP53* and *MYC* aberrations added value independently of genomic complexity. An increasing genomic complexity, together with the presence of breakage-fusion-bridge cycles and high DNA methylation changes related to the proliferative cell history, defines patients with different clinical evolution. (*Blood*. 2020;136(12):1419-1432)

Introduction

Mantle cell lymphoma (MCL) is a mature B-cell neoplasm with very heterogeneous behavior genetically characterized by the translocation t(11;14)(q13;q32), leading to *CCND1* overexpression.[1-4] The World Health Organization recognizes 2 molecular subtypes that differ in their clinical and biological features.[5-9] The most common conventional MCL (cMCL) derives from naive-like mature B cells, expresses the oncogenic transcription factor SOX11 and accumulates high numbers of genomic alterations. Patients usually have generalized lymphadenopathy and an adverse outcome. In contrast, leukemic non-nodal MCL (nnMCL) originates from memory-like B cells, is negative for SOX11, and is genetically stable. The disease usually involves peripheral blood and spleen but not lymph nodes in early stages, and follows an indolent behavior.[10] Both MCL subtypes carry the t(11;14) translocation, but the mechanisms leading to this rearrangement in both subtypes and the subsequent molecular alterations that drive their different evolution are not well defined.

The mutational profile of MCL has been previously studied in a small series of cases using whole-exome, transcriptome, or targeted sequencing, revealing a heterogeneous set of mutated genes, most of them at low frequencies.[11-15] However, these strategies did not explore the genome-wide mutational and structural alterations of the tumors, and have not properly addressed the differences between the 2 molecular subtypes of MCL. Furthermore, although the whole DNA methylome of MCL has been recently described, its relationship with genome-wide genetic events remains poorly characterized.[16]

To determine the influence of genome-wide (epi)genetic alterations in the heterogeneous behavior of MCL, we performed a combined analysis of whole-genome sequencing (WGS), transcriptome, and DNA methylome of a large cohort of MCL in the context of the reference epigenome of 5 representative cases from both MCL subtypes.

Methods

Patients and genomic studies

We performed WGS of paired tumor/normal samples from 61 MCL patients (44 cMCL, 17 nnMCL) (Table 1). We expanded the analysis with the whole-exome sequencing (WES) of 21 nonoverlapping cases previously reported (supplemental Table 1, available on the *Blood* Web site).[11] Cases were classified as cMCL or nnMCL on the basis of gene expression signatures detected by expression arrays, NanoString platform or reverse transcription quantitative polymerase chain reaction (PCR), and/or SOX11 immunohistochemistry, depending on the available material, as previously described (supplemental Methods).[6,17,18] Clinical status, morphological status, immunoglobulin heavy chain gene (IGHV) mutational status, genetic alterations, or epigenetic data were not used for the classification. WGS was performed using the TruSeq DNA PCR-free protocol and sequenced in a HiSeq X Ten (2 × 150 bp; Illumina) (supplemental Table 2). Raw reads were mapped

Table 1. Clinicopathological features of the 61 MCL patients with WGS analysis

Variable	Total (n)	cMCL (n = 44)	nnMCL (n = 17)	*P*
Clinical data (at diagnosis)				
Age, median (range), y	64 (38-85)	64 (38-85)	64 (51-80)	.477
Male/female, no.	43/18	33/11	10/7	.229
Nodal presentation, %	22/55 (40)	22/40 (55)	0/15 (0)	<.001
Splenomegaly, %	26/55 (47)	24/40 (60)	2/15 (13)	.002
LDH (>ULN), %	16/53 (30)	16/39 (41)	0/14 (0)	.005
MIPI high risk, %	42/46 (91)	32/36 (89)	10/10 (100)	.562
ECOG (≥2), %	8/48 (17)	8/37 (22)	0/11 (0)	.170
Pathological and molecular data				
Cyclin D1 positive, %	60/61 (98)	43*/44 (98)	17/17 (100)	1
Mutated IGHV genes				
Identity <98%, %	24/61 (39)	8/44 (18)	16/17 (94)	<.001
Identity, median (range)	99 (91-100)	99 (93-100)	95 (91-99)	<.001
Nanostring L-MCL16 assay, %†				<.001
cMCL	15/29 (52)	15/16 (94)	0/13 (0)	
nnMCL	11/29 (38)	0/16 (0)	11/13 (85)	
Undetermined	3/29 (10)	1/16 (6)	2/13 (15)	
Epigenetic COO, %‡				<.001
C1 (GC inexperienced)	35/54 (65)	34/37 (92)	1/17 (6)	
C2 (GC experienced)	17/54 (31)	2/37 (5)	15/17 (88)	
Undetermined	2/54 (4)	1/37 (3)	1/17 (6)	
Complex karyotype, %	17/37 (46)	12/21 (57)	5/16 (31)	.185
Morphology, %				<.001
Small cell	15/57 (26)	5/41 (12)	10/16 (62)	
Classic	30/57 (53)	24/41 (59)	6/16 (38)	
Blastoid	12/57 (21)	12/41 (29)	0/16 (0)	
Light chain restriction, %				.259
κ	35/61 (57)	23/44 (52)	12/17 (71)	
λ	26/61 (43)	21/44 (48)	5/17 (29)	
Sequenced sample, %				.023
Lymph node	12/61 (20)	12/44 (27)	0/17 (0)	
Other tissue§	2/61 (3)	2/44 (5)	0/17 (0)	
Peripheral blood	46/61 (75)	29/44 (66)	17/17 (100)	
Bone marrow	1/61 (2)	1/44 (2)	0/17 (0)	
Pretreatment sample, %	56/60 (93)	40/43 (93)	16/17 (94)	1
Time from diagnosis to pretreatment sample, median (range), mo	0.9 (0-101.6)	0.4 (0-14)	9.5 (0-101.6)	<.001
Treatment at diagnosis, %‖				<.001
High-dose therapy	17/58 (29)	17/41 (41)	0/17 (0)	
Immunochemotherapy	12/58 (21)	12/41 (29)	0/17 (0)	
Low-dose chemotherapy	6/58 (10)	6/41 (15)	0/17 (0)	
Observation	23/58 (40)	6/41 (15)	17/17 (100)	
Follow-up data				
Treated at 2 y, % (95% CI)	67 (52-78)	91 (76-97)	7 (0-18)	<.001
n treated, n censored, n missing	38, 3, 4	37, 2, 3	1, 1, 1	
2-y OS, % (95% CI)	81 (72-92)	73 (61-88)	100 (100-100)	.006
n dead, n censored, n missing	11, 4, 1	11, 4, 1	0, 0, 0	

CI, confidence interval; ECOG, Eastern Cooperative Oncology Group; LDH, lactate dehydrogenase; MIPI, MCL International Prognostic Index; ULN, upper level of normal.

*One case was negative for cyclin D1 expression and *CCND1* rearrangement but had MCL morphologic and phenotypic criteria (including SOX11 positivity) according to the WHO classification.[30]

†Clot et al.[17]

‡Queirós et al.[16]

§Corresponding to 1 skin and 1 tonsil.

‖The treatment information in 3 patients could not be obtained. High-dose therapy includes Cytarabine-based immunochemotherapy and/or autologous stem-cell transplantation; Immunochemotherapy includes *R*-CHOP-like regimens; and Low-dose therapy includes alkylating agents alone or in combination.

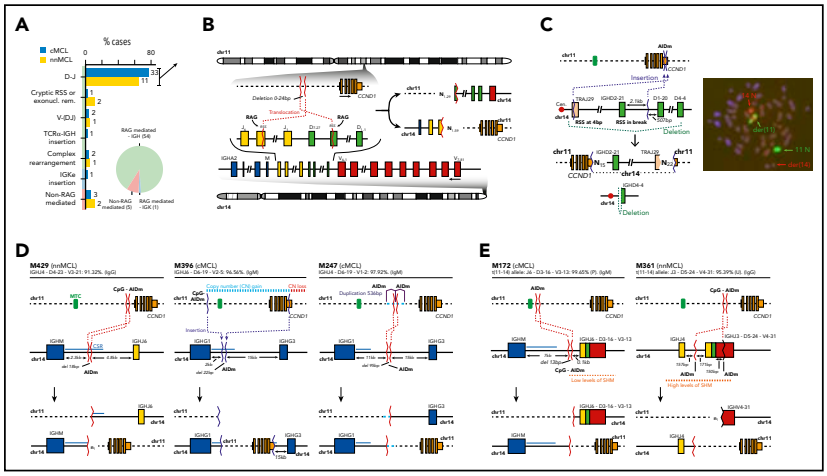


Figure 1. Characterization of the IG breaks of the IG/*CCND1* translocation. (A) Distribution of the IG breaks according to their underlying mechanism (RAG or non-RAG mediated) and location of the breakpoints in chr14. (B) Schema of the most recurrent translocation pattern with breaks at IGHD and IGHJ genes, near RSS, likely generated during the first step of an IGH D-J rearrangement. The presence of N-nucleotides in both breakpoints (N_N) supports a RAG-mediated process. (C) Unbalanced IG/*CCND1* translocation in which virtually the entire 14q arm is inserted at the 3′ UTR of *CCND1*. In this case, a single event truncated the *CCND1* 3′ UTR region and placed the IG enhancer near *CCND1*. The FISH whole-chromosome painting for chr11 and chr14 verifies the ins(11;14)(q13;q11q32) identified by WGS. For the sake of clarity, an interphasic nuclei present in the lower left part of the original picture was masked using Adobe Photoshop. (D) Representation of 3 cases in which the IG breakpoint was likely mediated by AID during CSR. The switch regions 5′ of the IGH constant genes are indicated by a blue line. The productive V(D)J rearrangement, IGHV identity (%), and isotype expression is specified. (E) Representation of 2 cases showing evidence for the involvement of the SHM machinery in the breakpoints of chr14 in already V(D)J-rearranged alleles. P, productive rearrangement; U, unproductive rearrangement.

to the human reference genome (GRCh37) using the BWA-mem algorithm,[19] and somatic single-nucleotide variants were called using Sidrón,[20] short insertions/deletions combining Sidrón and Pindel, and germline variants using HaplotypeCaller.[21] Copy number alterations (CNA) were extracted using Battenberg[22] and from Affymetrix Genome-wide Human SNP Array 6.0/Cytoscan using Nexus Biodiscovery.[20] Structural variants (SV) were analyzed using SMuFin and LUMPY.[23,24] Sanger sequencing, conventional cytogenetics, and fluorescence in situ hybridization (FISH) were used to verify selected mutations and rearrangements (supplemental Tables 3-6). Telomere length was determined using qMotif. Analysis of mutational signatures was performed as previously described.[25-27] Driver genes were identified as previously described,[20] whereas GISTIC was used to select driver CNA.[28] SV were integrated in driver discovery analyses. Timing of driver alterations was inferred from the clonality of each alteration as described elsewhere.[29]

Immunoglobulin gene rearrangements

Immunoglobulin gene rearrangements and identity were analyzed from WGS data using IgCaller[31] (supplemental Table 7). The primary IG/*CCND1* translocation was characterized from WGS data using a bespoke algorithm that mapped the breakpoints at base pair resolution and searched for evidences of aberrant V(D)J recombination (ie, recombination-activating gene [RAG] activity), class switch recombination (CSR), and somatic hypermutation (SHM) (supplemental Table 8).

Epigenetic and gene expression analyses

Infinium Methylation EPIC BeadChip was used to generate DNA methylation profiles (n = 70). The reference epigenome including the analysis of 6 histone marks (H3K4me3, H3K4me1, H3K27ac, H3K36me3, H3K9me3, and H3K27me3), chromatin accessibility (ATAC-seq), and RNA-seq of 15 normal B cells and 5 MCL (2 cMCL, 3 nnMCL) were generated within the Blueprint Consortium.[32] In situ Hi-C data for the 5 MCL as well as naïve and memory B cell subpopulations were obtained from our recent study.[33] Gene expression profiling was performed using Affymetrix U219 microarrays (n = 44). cMCL and nnMCL gene expression signatures were studied by the L-MCL16 assay (NanoString Technologies).[17]

Statistical analyses

The log-rank test (categorical) and Cox regression (continuous) were used to measure the association of the overall survival (OS) with clinicobiological variables. *P* values were adjusted using the Benjamini-Hochberg method.

Results

Whole-genome overview and active mutational processes

We detected a median of 3593 (range, 1691-6597) somatic mutations per case (1.2 mutations per megabase), including 33 (range, 13-56) coding mutations per tumor. MCL tumors carried a median

of 9 (range, 1-56) SV and 9 (range, 0-37) CNA (supplemental Figures 1-3; supplemental Tables 9-11). The mutational burden was similar in both MCL subtypes, but cMCL carried higher number of SV (median, 13 vs 3; $P < .001$) and CNA (median, 12 vs 1; $P < .001$) than nnMCL (supplemental Figures 1-3). A complex genomic landscape, defined by the presence of ≥15 SV and/or ≥15 CNA (mean values in cMCL), was observed in 20 (45%) cMCL and 2 (12%) nnMCL ($P = .018$).

We identified 6 predominant signatures of mutational processes operative in MCL: clock-like signatures 1 (SBS1) and 5 (SBS5) present in all MCL; the noncanonical (SBS9) and canonical (SBS84) activation-induced cytidine deaminase (AID) enriched in nnMCL, with SBS84 predominantly associated with IG genes and active promoters/enhancers; and 2 APOBEC-related signatures (SBS2 and SBS13) occurring in regions of kataegis, particularly those associated with SV and chromothripsis, and enriched in cMCL, which is in line with their higher number of SV (supplemental Figures 4-7).

Genomic/epigenomic characterization of *CCND1* rearrangement

CCND1 was rearranged with immunoglobulin (IG) genes (IG/*CCND1*) in the 60 cases overexpressing cyclin D1, 59 cases with IG heavy locus (IGH) and 1 with IG kappa locus (IGK) (Figure 1A; supplemental Table 12). One cMCL was a cyclin D1-negative MCL that overexpressed cyclin E1/E2.[30] This unusual case had MCL morphology and phenotype ($CD20^+$, $CD5^+$) with strong expression of SOX11, and that were maintained in different relapses. In 55 cases, the IG breaks had evidence of being mediated by RAG enzymes during V(D)J recombination. As expected,[34] this pattern was seen in 40 of 43 (93%) cMCL derived from naïve-like B cells, but also in 15 of 17 (88%) memory-like nnMCL. Intriguingly, the IG breaks in the remaining 5 cases (3 cMCL, 2 nnMCL) appeared to involve CSR in 3 cases and SHM in 2 cases (Figure 1A; supplemental Table 13).

In 44 of the 55 cases with RAG-mediated IG rearrangements, the IG breaks occurred in IGHD and IGHJ genes, likely during the initial IGHD–IGHJ recombination. We identified the RAG recombination signal sequence (RSS) at the IGHD and IGHJ breakpoints with the addition of 1 to 59 nontemplated nucleotides (N-nucleotides) at both derivative junctions (Figure 1A-B). These N-nucleotide segments were longer than in physiological V(D)J recombination of B cells, likely because of the absence of selection to retain a limited length of the IGH third complementarity-determining region required for a functional B-cell receptor in these IGH/*CCND1* rearrangements. The IGHJ and IGHD genes involved in these translocations were similar in cMCL and nnMCL and mainly related to those used in physiological recombination of normal B cells (supplemental Figure 8). The remaining 11 of 55 cases with RAG rearrangements (RSS and N-nucleotides) included 3 cases with breaks at canonical RSS of an IGHD and IGHV, suggesting that the translocation occurred during the second step of the V(D)J recombination; 3 cases with 1 of the breaks at an atypical cryptic RSS distant from the near IGHD or IGHJ segments, or missing a RSS potentially caused by exonucleolytic removal; 3 cases with complex rearrangements including small fragments of chromosomes other than chromosome (chr)11/14; and 2 cases with unbalanced translocations in which the IGK locus or virtually all 14q (TRAJ19-IGHD21 segment) were inserted upstream or at the 3′ untranslated region (UTR) of *CCND1*, respectively (Figure 1A-C; supplemental Figure 9).

Five MCL had a t(11;14) with breakpoints in the IGH gene consistent with the involvement of AID in CSR and SHM mechanisms and, therefore, generated in a mature B cell probably during a follicular germinal center reaction (Figure 1D-E; supplemental Figures 10 and 11). In 2 cases, the breakpoints were in the IGHM- and IGHG1-defined CSR regions, respectively. In 1 case, the breakpoint was between IGHG1 and IGHG3, but in the absence of N-nucleotides and RSS sites and the presence of point mutations, CSR was the most probable mechanism. In the last 2 cases, the mechanism seems to have involved the SHM machinery, with breakpoints at unusual sites of the IGHV/D/J genes. These translocations occurred in already V(D)J-rearranged alleles, reinforcing the idea of being acquired at a mature B-cell stage. The comparison of the percentage of cells carrying the translocation by FISH and the tumor cell content detected by flow cytometry available in 4 cases was relatively similar, suggesting that these translocations were clonal. These 5 translocations seem to trigger a similar overexpression of *CCND1* as in conventional RAG rearrangements and do not confer different clinical or biological features to the tumors (supplemental Figure 12; supplemental Table 13).

We next analyzed the breakpoints on chr11. Nineteen translocations (14 cMCL, 5 nnMCL) occurred at a small region of 89 bp previously recognized as a major translocation cluster (MTC). The remaining breakpoints were similarly scattered at both sides of the MTC in cMCL and nnMCL, and their distribution was not associated with IG κ/λ expression. As previously suggested, most 5′ and MTC breaks occurred near CpG sites and AID motifs, whereas 3′ breaks were only found near AID motifs (Figure 2A; supplemental Figures 13 and 14).[35,36]

Next, we exploited the analysis of the reference epigenomes to define the local chromatin features associated with the t(11;14) (Figure 2B).[33] The chr11 breakpoint was at the MTC locus in 2 cases, and 5′ or 3′ regions distant from the MTC in the other 3 cases. We observed that the MTC locus corresponds to an open chromatin region (defined by ATAC-seq) with histone marks of active regulatory elements (H3K4me1/H3K4me3/H3K27ac) in the 5 MCL and also in normal naïve and memory B cells, but not in germinal center or plasma cells. As the t(11;14) is mostly associated with the V(D)J rearrangement occurring at precursor B-cell stage, we analyzed the chromatin of B-cell acute lymphoblastic leukemias available through the Blueprint Consortium as a surrogate of precursor B cells.[32] We observed that the MTC locus also corresponds to an enhancer region (H3K4me1) in these cells (Figure 2B). Interestingly, breakpoints occurring at the MTC lead to an extension of the existing enhancer/promoter marks of this locus, whereas 2 breakpoints distant from the MTC seemed to generate novel active enhancer/promoters. This gain of regulatory marks in chr11 breakpoints is most likely caused by the fusion of these regions with the active enhancer/promoter region of the IG (Figure 2C). HiC-seq performed in tumor cells from these 5 MCL and normal naïve and memory B cells showed a reconfiguration of the 3-dimensional (3D) chromatin structure of this region in all MCL as compared with their normal counterparts. The chr11 breakpoints overlapped with novel tumor-specific topologically associating domain (TAD) borders. Of note, *CCND1* was always found at the distal border of the TAD, confirming that these IG-novel chr11 promoter/enhancer regions contribute to dysregulate *CCND1* by creating specific build blocks (Figure 2D; supplemental Figure 15).

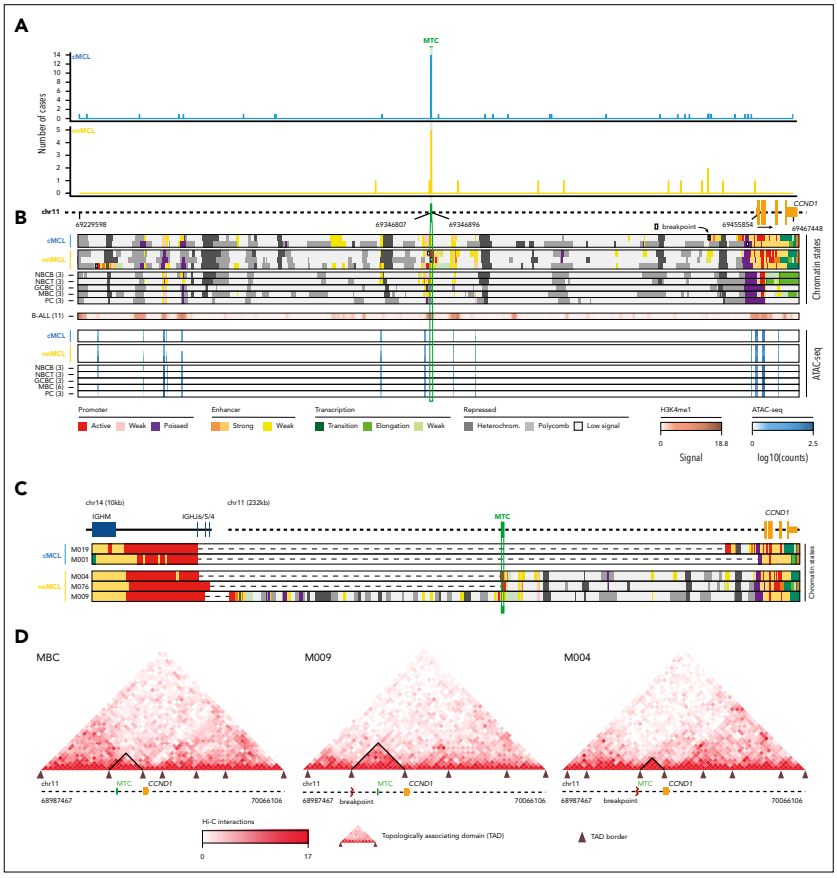


Figure 2. Integrative analysis of chr11 breaks of the IG/*CCND1* translocation. (A) Distribution of breakpoints observed in chr11 in cMCL (top) and nnMCL (bottom). The number of cases with breakpoints in close proximity is summarized using a sliding window of 89 bp (MTC length) starting from the MTC region (shown in green). (B) Representation of the reference epigenomes of 5 MCL cases and 15 normal B cell samples spanning the B-cell maturation program. Numbers in brackets indicate the number of samples considered to build the consensus chromatin map of each cell type. (top) The breakpoints of each case are highlighted by a white square. (middle) Signal of H3K4me1 in B-cell acute lymphoblastic leukemia (B-ALL). (bottom) Chromatin accessibility (ATAC-seq peaks) showing the presence of an open chromatin region near the MTC region. B cells are grouped, and a consensus is depicted. (C) Reconstruction of the chromatin states of the IG/*CCND1* translocated allele. Dashed lines indicate the junction of both chromosomes to maintain the breakpoints relative to the MTC and *CCND1*. (D) Hi-C contact matrices for memory B cells (MBC) and 2 nnMCL cases. The TAD associated with *CCND1* and modulated upon the IG/*CCND1* translocation is highlighted. GCBC, germinal center B cell; NBCB, naive B cell from peripheral blood; NBCT, naive B cell from tonsil; PC, plasma cell.

Landscape of structural alterations in MCL subtypes

The WGS analysis allowed the precise characterization of the multiple structural alterations in MCL. Complex alterations were significantly more frequent in cMCL (52%) than nnMCL (18%; $P = .02$; Figure 3A). Chromothripsis events were clonal in all but 1 case, mainly involving chromosomes 1, 5, 10, 12, and 13 and recurrently targeting *RB1* in 4 (9%) cases and *TERT* in 2 (12%) nnMCL (Figure 3B). Chromoplexia affected 14 different chromosomes (chromosomes 2, 6, 12, and 19 in 2 cases each), with *TERT* the only cancer gene affected in 1 case (Figure 3Ci). Chromothripsis and chromoplexia occurred in both MCL subtypes, but breakage-fusion-bridge (BFB) cycles, a novel and frequent finding in MCL, was only observed in cMCL (20%). BFB cycles generated recurrent high-level amplification of

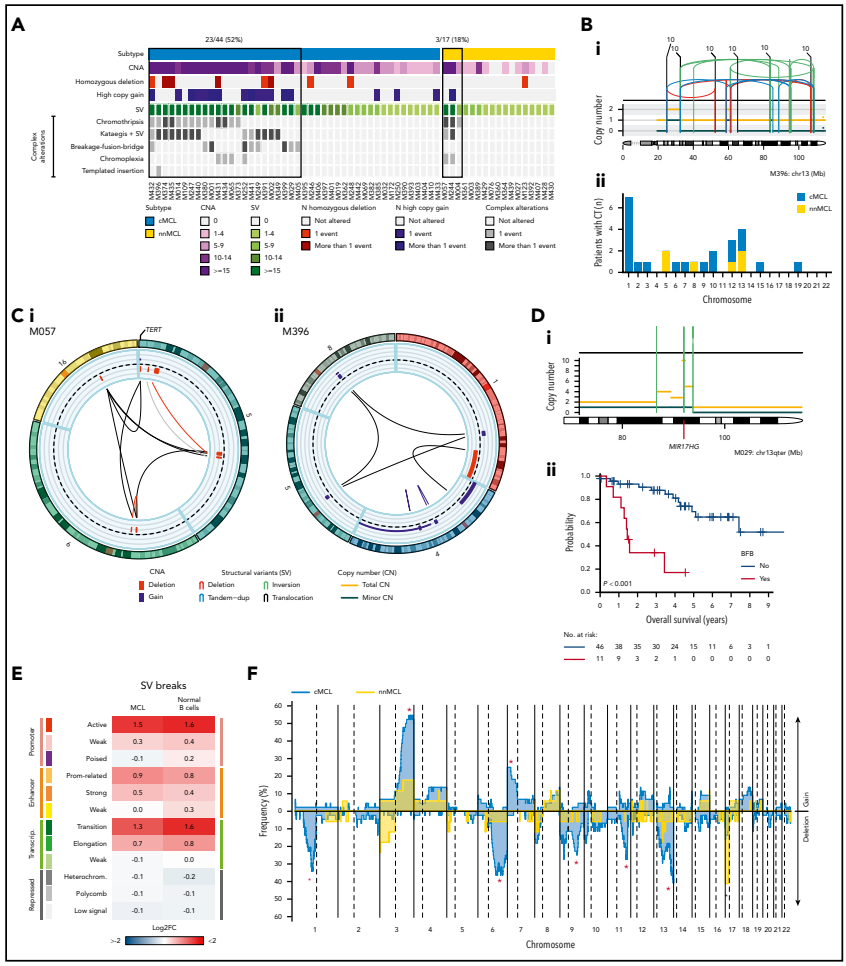


Figure 3. Complex genomic alterations in MCL identified by WGS. (A) Genomic complexity identified in both subtypes of MCL. Cases are depicted in columns. Rows illustrate CNA complexity, the number of homozygous deletions and amplifications, SV complexity, and different SV complex phenomena characterized by clustered SV (chromothripsis, kataegis overlapping with SV, BFB cycles, chromoplexia, and templated insertions). (B) Illustrative example of a whole chr13 affected by chromothripsis in a cMCL (i). (ii) Bar plot of chromosomes affected by chromothripsis (CT) in the 2 MCL subtypes. (C) Partial circos plot showing the 3 chromosomes involved in chromoplexia and *TERT* amplification and translocation in a nnMCL (i). (ii) Partial circos plot showing 4 chromosomes with crossed rearrangements and template insertions (focal gains) in 3 of them (chromosomes 1, 5, and 8) in a cMCL. (D) Illustrative example of BFB cycles resulting in amplification of *MIR17HG* in 13q accompanied by a terminal deletion (i). (ii) Kaplan-Meier curve of OS according to the presence of BFB cycles. (E) Enrichment of SV breakpoints in different chromatin states, as compared with background, in MCL and normal B cells. (F) Global profile of CNA in cMCL (blue) and nnMCL (yellow). The regions with different proportions of altered cases between subtypes ($Q < 0.15$) are indicated (*). Only regions with at least 6 altered cases were included in the comparison.

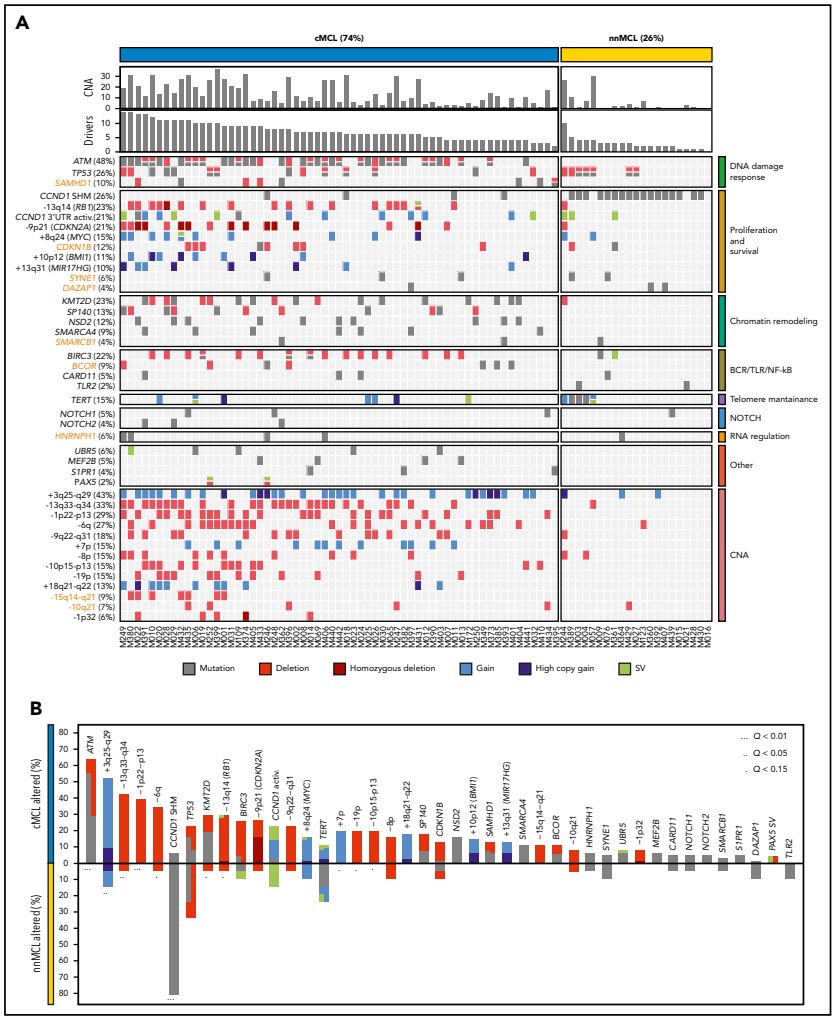


Figure 4. Significantly mutated genes/CNA and pathways. (A) Oncoprint representation of the 43 driver alterations identified in MCL. Drivers are depicted in rows, and cases are displayed in columns. Novel driver alterations identified in this study are highlighted in dark orange. (B) Distribution and type of driver alterations in the 2 different MCL subtypes. Drivers with a different proportion of altered cases between subtypes ($Q < 0.15$) are indicated.

BMI1 (4 cases) and *MIR17HG* (2 cases), and were associated with worse clinical outcome (Figure 3D). One cMCL presented features of a replication-based mechanism of templated insertions (Figure 3Cii).

Although chromosomal translocations and inversions were relatively frequent in MCL, they were not recurrent and very few were associated with known cancer genes. Two cMCL had SV that truncated *PAX5*. One cMCL had a balanced 2p

inversion that fused *MYCN* with IGK enhancer, leading to high overexpression of the gene, and the patient had central nervous system involvement.[37] One cMCL carried an activating *MYC* non-IG rearrangement that remodeled the adjacent *MYC* regulatory regions associated with *PVT1*[38] and hijacked distant enhancers in chr8 (supplemental Figures 16-18).

We analyzed the presence of SV signatures and identified 3 patterns mainly related to the presence of clustered structural alterations, high SV/CNA complexity, and small alterations with low genomic complexity (supplemental Figure 19). Remarkably, SV breaks were enriched in active promoters and enhancers and in transcription elongation–associated chromatin regions, suggesting that open/active chromatin might facilitate SV formation, and they may impact gene expression (Figure 3E).

The CNA detected by WGS confirmed the specific MCL profile previously characterized by frequent losses of 1p22-p13, 6q, 9p21/*CDKN2A*, 9q22-q31, 11q22-q23/*ATM*, 13q14/*RB1*, 13q33-q34, and 17p/*TP53*, and gains of 3q25-q29 and 7p. We also identified novel recurrent losses at 10q21.1 and 15q14-q21.1 and significant differences in the distribution of specific alterations in cMCL (losses of 1p22-p13, 6q, 9q22-q31, 11q22-q23/*ATM*, 13q33-q34, and gains of 3q25-q29 and 7p) and nnMCL (loss of 17p/*TP53*) (Figure 3F; supplemental Figures 20 and 21).

MCL driver alterations

To discover genomic alterations involved in MCL lymphomagenesis we integrated mutations, CNA and SV combining the results of the 61 WGS with the WES of 21 nonoverlapping cases (supplemental Tables 14 and 15).[11] The 82 cases encompassed 74% cMCL and 26% nnMCL. We identified 26 genes significantly altered in the whole cohort, and 4 genes mutated at lower frequency, but carrying known driver alterations (*NOTCH1*, *NOTCH2*, *TLR2*, and *PAX5*). In addition, we identified 13 chromosomal regions without a defined target gene significantly affected by CNA, resulting in deletions in most cases (Figure 4A; supplemental Figure 22; supplemental Tables 16-18). Overall, 81 of 82 (99%) MCL cases had at least 1 driver alteration in addition to the t(11;14) (median, 6; range, 0-14). The most frequently altered genes have been previously described in MCL and were *ATM* (48%), *CCND1* (44%) with exon1/intron1 somatic mutations (26%) and/or 3′ UTR activating alterations (21%), *TP53* (26%), *KMT2D* (23%), *RB1* (23%), *BIRC3* (22%), *CDKN2A* (21%), *SP140* (13%), *NSD2* (12%), *BMI1* (11%), *MIR17HG* (10%), and *UBR5* (6%). Furthermore, we identified 7 novel MCL driver genes altered by missense or truncating mutations and deletions, including *CDKN1B* (12%), *SAMHD1* (10%), *BCOR* (9%), *SYNE1* (6%), *HNRNPH1* (6%), *SMARCB1* (4%), and *DAZAP1* (4%) (Figure 4A). The integration of WGS, gene expression and FISH analysis identified the relevance of *TERT* in MCL (15%) affected by promoter mutations (3 cases), gain/amplification (6 cases), and translocations (3 cases) with high *TERT* overexpression without an apparent impact on telomere length (supplemental Figure 23). However, cases with *ATM* alterations had significant shorter telomeres (supplemental Figure 24). No other recurrent mutations in expressed or regulatory noncoding regions were found. We identified a significant co-occurrence of several alterations such as *ATM* with +10p12/*BMI1* or +8q24/*MYC* with −13q14/*RB1*, and recognized early driver events including *ATM*/11q alterations and −17p, or late events including +8q24/*MYC* and *BIRC3* mutations, among others (supplemental Figure 25; supplemental Table 19). Collectively, 8 main pathways were frequently altered in MCL including proliferation, cell survival, DNA damage response (DDR), telomere maintenance, chromatin remodeling, B-cell receptor/Toll-like receptor/NF-κB signaling, NOTCH signaling, and RNA regulation (Figure 4A).

We also searched whether the MCL drivers described in this study and additional known cancer predisposing genes[39,40] were recurrently mutated in the germline of the patients; we found 7 cases with *ATM* and 2 with *CHEK2* mutations (supplemental Table 20). In 2 cases with germline *ATM* mutation, the wild-type allele was lost in the tumor. The 2 *CHEK2* mutations had been previously recognized as pathogenic in cancer.[41-43]

Finally, most MCL drivers were found to be preferentially altered in cMCL cases, with the unique exception of SHM in *CCND1* mainly found in nnMCL (Figure 4B). Of note, *ATM* alterations (64%); deletions of 1p, 10p, and 19p; and gain of 7p were exclusively seen in cMCL, whereas *TP53* and *TERT* alterations were the only drivers slightly enriched in nnMCL, with all 5 cases with *TERT* alterations carrying concomitant *TP53* aberrations (Figure 4A-B). Altogether, cMCL cases had a significant higher number of driver alterations than nnMCL (median, 7 vs 2; $P < .001$).

DNA methylome of MCL subtypes and interplay with genomic alterations

We next studied the DNA methylome of MCL and its relationship with driver genetic alterations. We previously classified MCL cases into 2 clusters (C1 and C2) on the basis of DNA methylation imprints of their postulated cell of origin, ie, pregerminal and postgerminal center B cells, respectively.[16] We found that cMCL and nnMCL widely overlapped with C1 and C2 epigenetic subgroups, respectively ($P < .001$; Figure 5A). A principal component analysis showed that the main source of DNA methylation variability is related to the MCL cell of origin, and consequently is associated with its clinicobiological features, such as IGHV identity, total number of driver alterations, and *ATM* or *CCND1* mutations in cMCL and nnMCL, respectively (Figure 5B; supplemental Figure 26). We have previously observed that an additional source of DNA methylation variability among MCL was the accumulation of DNA methylation changes in the tumors, which, in addition, was related to clinical outcome.[16] The majority of these changes were located at transcriptionally silenced regions, particularly at low signal heterochromatin and H3K27me3 for hypomethylation and hypermethylation, respectively.[16] Mounting lines of evidence indicate that these DNA methylation changes do not play a regulatory role, but instead seem to gradually accumulate during rounds of cell divisions.[44-50] Based on this principle, we have recently built a DNA methylation-based mitotic clock called epiCMIT (epigenetically determined cumulative mitoses), which reflects the proliferative history of B-cell tumor samples (supplemental Methods).[51] Interestingly, in the present MCL cohort, the second main source of DNA methylation variability after the cell of origin was related to this epiCMIT score (Figure 5B). As expected, the epiCMIT correlated with mutational signatures related to cell division (SBS1, SBS5, and SBS9) and Ki67-index (Figure 5C; supplemental Figure 26).

We next explored whether the proliferative history (epiCMIT) was related to genetic changes. Indeed, we observed a significant association between the epiCMIT and the number of driver alterations and CNA burden (Figure 5D; supplemental Figure 26). At a single driver level, 9 genetic alterations were associated with

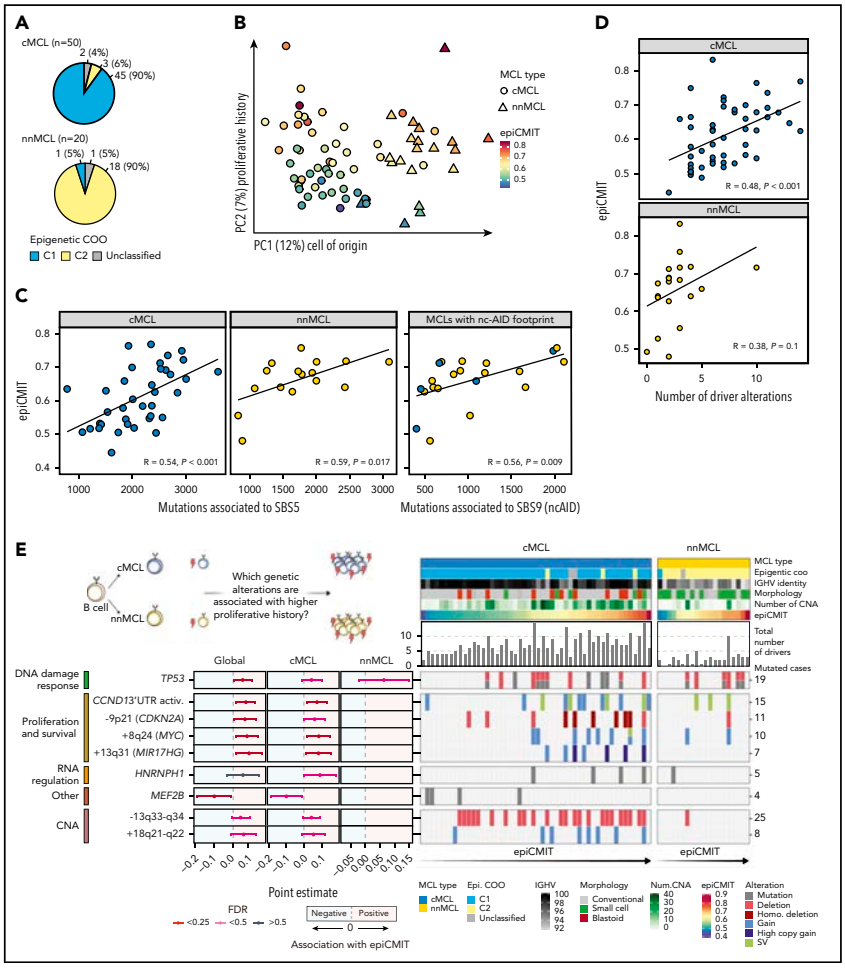


Figure 5. DNA methylome of MCL subtypes and its relationship with genomic alterations. (A) Overlap of cMCL and nnMCL with epigenetic subgroups on the basis of cell of origin methylation signature. (B) Principal component analysis of DNA methylation data for 70 MCL (first and second components are shown). MCL subtypes are represented as triangles or circles, whereas the color represents the proliferative history on the basis of DNA methylation of each MCL sample (epiCMIT score). (C) epiCMIT correlates with mutational signatures related to cell division, including SBS5 and SBS9 (ncAID). (D) epiCMIT correlates with the total number of driver alterations in MCL, particularly in cMCL. (E) Driver alterations associated with higher or lower epiCMIT. The 95% confidence intervals for the mean epiCMIT difference between the presence and absence of each alteration are shown. The effect in the whole cohort was adjusted by the cell of origin (C1/C2). (left) Colors depict different significant levels after false discovery rate (FDR) correction. Oncoprint with genetic alterations associated with the epiCMIT together with clinicobiological variables. (right) Patients are ordered according to the epiCMIT score, separately in the 2 MCL subtypes.

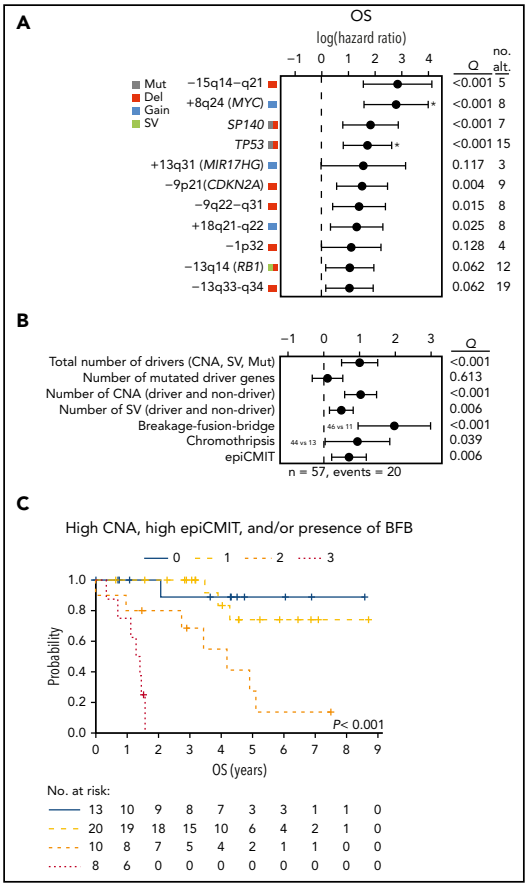


Figure 6. Clinical relevance of genomic and epigenomic alterations in MCL subtypes. (A) Impact of driver alterations on OS. The impact is quantified with the 95% confidence interval of the log hazard ratios. The *Q* value shown is the adjusted *P* value of the log-rank test. Only alterations with at least 3 altered cases and prognostic value are shown. Drivers with independent prognostic value ($Q < 0.05$) of the number of CNA are indicated (*). (B) Impact of the cumulative number of genetic and epigenetic changes to OS. The impact is quantified with the 95% CI of the log hazard ratios. Continuous variables were scaled. The *Q* value shown is the adjusted *P* value of the simple Cox regression for the continuous variables or the log-rank test for the binary variables (BFB and chromothripsis). The number of SV was available in 42 cases. The epiCMIT was available in 51 cases and its effect was adjusted by the cell of origin (C1/C2). (C) Kaplan-Meier curves of OS according to the number of risk features (high CNA, high epiCMIT, and/or presence of BFB). Number of CNA >7 (median) and epiCMIT >0.6 (valley of a bimodal distribution) were considered high. no. alt, number of altered cases.

higher proliferation histories of MCL samples, which were related to pathways linked to proliferation, survival, and DDR such as +8q24/*MYC*, −9p21/*CDKN2A*, and *TP53* alterations. Interestingly, MCL carrying mutations in *HNRNPH1* and *MEF2B* were associated with significantly higher and lower proliferation histories of MCL cells, respectively (Figure 5E; supplemental Figure 27).

Clinical implications

Finally, we analyzed the clinical relevance of the previous findings in 57 cases, excluding patients managed with palliative measures or studied at relapse. The main parameters associated with shorter OS were high lactate dehydrogenase and MCL International Prognostic Index (both with $P = .002$) and cMCL molecular subtype ($P = .02$). In addition, *TP53* alterations, −9p21/*CDKN2A*, −13q14/*RB1*, and −9q22−q31 impaired the OS of the patients. Interestingly, 5 novel drivers were also associated with significant shorter OS (−15q14−q21, +8q24/*MYC*, *SP140*, +18q21-q22, and −13q33-q34; $Q < 0.1$; Figure 6A).

Next, we assessed the potential clinical relevance of the genomic complexity and epigenetic changes. The total number of CNA, number of SV, presence of BFB, chromothripsis, and the epiCMIT, but not the number of driver mutations, were associated with shorter OS ($Q < 0.05$; Figure 6B). *TP53* and *MYC* alterations had prognostic value for OS independently of the total number of CNA ($Q < 0.05$; Figure 6A). Of note, the number of CNA, presence of BFB, and epiCMIT retained independent prognostic

value for OS in a multivariate analysis ($P < .05$), and stratified patients with distinct clinical outcomes (Figure 6C).

Discussion

In this study we have provided a comprehensive characterization of the molecular development and progression of MCL. The primary *CCND1* rearrangement is mediated by the same mechanisms in cMCL and nnMCL, but these subtypes differ in their epigenetic makeup, subsequent driver composition, and genomic complexity that influence their different clinical behavior.

The t(11;14) in both cMCL and nnMCL originates in precursor B cells mediated by RAG activity, even though the tumor cell expansion will arise in a pregerminal or postgerminal center cell, respectively, as supported by the different DNA methylation signatures related to naïve and memory-like B cells. The mechanism of diversification of these MCL subtypes may be determined in part by SOX11 overexpression in cMCL, as previously described.[5] In this comprehensive study, we have not identified any genetic alteration leading to SOX11 overexpression, emphasizing the relevance of its epigenetic dysregulation.[16] Intriguingly, the origin of the t(11;14) in 8% of the cases, including both cMCL and nnMCL, seems to occur in a mature B cell mediated by SHM/CSR mechanisms without apparent clinical or biological differences. This situation parallels the inverse findings in most cases of multiple myeloma, in which the t(11;14) occurs in a mature B cell mediated by AID/CSR, but it may also arise in a precursor B cell in a minority of tumors.[52] These findings suggest that the development of these different lymphoid and plasma cell neoplasms is independent of the B-cell differentiation stage in which the initial oncogenic event occurs. Our integrative (epi)genomic analysis also suggests that the open and active chromatin structure of the MTC region may allow the access of AID to mediate the DNA breaks and subsequent translocation,[53-56] and shows that a 3D chromatin reconfiguration of the *CCND1* locus upon the translocation favors *CCND1* overexpression.

We have identified novel driver genes involved in different mechanisms relevant for MCL pathogenesis such as cell cycle (*CDKN1B*), DNA replication and DDR (*SAMHD1*),[57] RNA processing (*HNRNPH1*),[58] or chromatin modification (*SMARCB1*).[59] The more aggressive behavior of cMCL is associated with higher number of drivers, particularly CNA, compared with nnMCL. Interestingly, *ATM* alteration seems a specific and early event driving cMCL (64% of cases) that may even occur in the germline of some patients. The inactivation of this element may promote the shortening of telomeres, facilitating the development of the frequent structural genomic complexity of these tumors.[37] On the contrary, nnMCL have very few driver alterations, with only *TERT* and *TP53* slightly more enriched in this subgroup. The high genomic complexity in a small subset of nnMCL carrying these 2 alterations suggests that they may be an alternative mechanism facilitating tumor progression.

The WGS of this cohort shows the complex genomic landscape of MCL at unprecedented high resolution and uncovers novel mechanisms of oncogenic activation including BFB cycles amplifying *BMI1* and *MIR17HG*,[60,61] or inversions and translocations reorganizing regulatory regions (*NMYC, MYC, TERT*).[62] Indeed, this genomic complexity emerges as one of the most relevant parameters determining the patient outcome, and is associated with increasing DNA methylation changes previously related to cell division.[44-50] Our epiCMIT score captures these changes and correlates with higher proliferation, and specific genetic alterations such as −9p21 (*CDKN2A*), *TP53*, *MYC*, or the newly identified *HNRNPH1*, which collectively suggests that these alterations may provide to MCL cells a higher proliferative capacity.

Several individual drivers have prognostic value. However, their impact is mainly because of their association with the global genomic complexity captured by the number of CNA. Only *TP53* and *MYC* alterations add prognostic information to CNA, reinforcing the prognostic relevance of these individual alterations observed in previous studies.[17,63-66] Interestingly, the epiCMIT score also has independent prognostic value, indicating that the complex evolution of MCL may be better recognized by the integration of genomic and epigenomic parameters.

In conclusion, the integration of whole-genome and epigenomic analyses of MCL reveals the complexity and relationship of these alterations that determine the different evolution of the 2 MCL subtypes and define subsets of patients with marked distinct outcomes.

Acknowledgments

The authors are grateful to have used the hematopathology collection registered at the Biobank of the Hospital Clinic–Institut d'Investigacions Biomèdiques August Pi i Sunyer (IDIBAPS) for sample procurement. The authors are indebted to the IDIBAPS Genomics core facility. We are grateful to Miriam Prieto, Silvia Martín, and Sílvia Ruiz for their technical and logistic assistance.

This work was supported by research funding from Fondo de Investigaciones Sanitarias, Instituto de Salud Carlos III (PI17/01061 to S.B.); National Institutes of Health, National Cancer Institute "Molecular Diagnosis, Prognosis, and Therapeutic Targets in Mantle Cell Lymphoma" (P01CA229100 to E.C.); Ministerio de Ciencia e Innovación (RTI2018-094274-B-I00 to E.C., SAF2017-87811-R to X.S.P., and SAF2017-86126-R to J.I.M.-S.); Generalitat de Catalunya Suport Grups de Recerca AGAUR (2017-SGR-1142 to E.C., and AGAUR 2017-SGR-736 to J.I.M.-S.); German Ministry of Science and Education in the framework of the ICGC MMML-Seq (01KU1002) and ICGC DE-Mining (01KU1505) to R.S. F.N. is supported by a predoctoral fellowship of the Ministerio de Ciencia e Innovación (BES-2016-076372). A.D.-N. is supported by the Department of Education of the Basque Government (PRE_2017_1_0100). E.C. is an Academia Researcher of the "Institució Catalana de Recerca i Estudis Avançats" of the Generalitat de Catalunya.

This work was mainly developed at the Centre Esther Koplowitz, Barcelona, Spain.

Authorship

Contribution: F.N. and D.M.-G. analyzed and interpreted data, wrote the manuscript, and contributed to the conception of the study; G. Clot performed statistical and clinical analyses; A.D.-N., A.N., R.R., J.G.-A., R.V.-M., C.L., M.P., G. Castellano, P.J., D. Costa, D.T., and R.K. analyzed and interpreted data; M.D.-F., R.V.-B., M.K., and V.C. analyzed and interpreted epigenetic data; A.N., M.A., B.E., A.M., I.R.-C., R.S., and D. Colomer performed sample preparation and/or provided clinical data; E.G. and A.L.-G. centralized clinical data and interpreted data; J.I.M.-S. and X.S.P. analyzed and interpreted data, and contributed to the conception of the study; S.B. and E.C. jointly designed the study, supervised the research, interpreted data, and wrote the manuscript; and all authors read, commented on, and approved the manuscript.

Conflict-of-interest disclosure: E.C. is named inventor on 2 patents filed by the National Institutes of Health, National Cancer Institute: "Methods for selecting and treating lymphoma types," licensed to NanoString Technologies, and "Evaluation of mantle cell lymphoma and methods related thereof." X.S.P. has ownership interest in DREAMgenics. The remaining authors declare no competing financial interests.

ORCID profiles: F.N., 0000-0003-2910-9440; A.D.-N., 0000-0001-6808-4800; M.D.-F., 0000-0003-1666-5819; J.G.-A., 0000-0001-8863-5876; C.L., 0000-0001-6644-1659; V.C., 0000-0002-4910-5394; M.P., 0000-0001-5034-7924; G. Castellano, 0000-0002-5715-7733; B.E., 0000-0002-4294-8145; R.K., 0000-0002-6691-7191; X.S.P., 0000-0001-9525-1483; S.B., 0000-0001-7192-2385; E.C., 0000-0001-9850-9793.

Correspondence: Elias Campo, Unitat Hematopatologia, Hospital Clínic, Villarroel 170, 08036 Barcelona, Spain; e-mail: ecampo@clinic.cat; and Sílvia Beà, Molecular Pathology of Lymphoid Neoplasms, Institut d'Investigacions Biomèdiques August Pi i Sunyer, Rosselló, 153, 08036 Barcelona, Spain; e-mail: sbea@clinic.cat.

Footnotes

Submitted 10 February 2020; accepted 14 April 2020; prepublished online on *Blood* First Edition 25 June 2020. DOI 10.1182/blood.2020005289.

*F.N. and D.M.-G. contributed equally to this work.

†S.B. and E.C. jointly supervised this work.

Our entire genomic data set, including whole-genome sequencing, SNP arrays, gene expression arrays, and DNA methylation arrays have been deposited at the European Genome-phenome Archive (accession number EGAS00001004165).

The online version of this article contains a data supplement.

There is a *Blood* Commentary on this article in this issue.

The publication costs of this article were defrayed in part by page charge payment. Therefore, and solely to indicate this fact, this article is hereby marked "advertisement" in accordance with 18 USC section 1734.

REFERENCES

1. In: Swerdlow SH, Campo E, Harris NL, eds., et al. WHO Classification of Tumours of Haematopoietic and Lymphoid Tissues (Revised 4th Edition), Lyon, France: IARC; 2017:
2. Martin P, Chadburn A, Christos P, et al. Outcome of deferred initial therapy in mantle-cell lymphoma. *J Clin Oncol.* 2009;27(8):1209-1213.
3. Martin P, Leonard J. Is there a role for "watch and wait" in patients with mantle cell lymphoma? *Semin Hematol.* 2011;48(3):189-193.
4. Campo E, Rule S. Mantle cell lymphoma: evolving management strategies. *Blood.* 2015;125(1):48-55.
5. Fernàndez V, Salamero O, Espinet B, et al. Genomic and gene expression profiling defines indolent forms of mantle cell lymphoma. *Cancer Res.* 2010;70(4):1408-1418.
6. Royo C, Navarro A, Clot G, et al. Non-nodal type of mantle cell lymphoma is a specific biological and clinical subgroup of the disease. *Leukemia.* 2012;26(8):1895-1898.
7. Ondrejka SL, Lai R, Smith SD, Hsi ED. Indolent mantle cell leukemia: a clinicopathological variant characterized by isolated lymphocytosis, interstitial bone marrow involvement, kappa light chain restriction, and good prognosis. *Haematologica.* 2011;96(8):1121-1127.
8. Jares P, Colomer D, Campo E. Molecular pathogenesis of mantle cell lymphoma. *J Clin Invest.* 2012;122(10):3416-3423.
9. Puente XS, Jares P, Campo E. Chronic lymphocytic leukemia and mantle cell lymphoma: crossroads of genetic and microenvironment interactions. *Blood.* 2018;131(21):2283-2296.
10. Swerdlow SH, Campo E, Pileri SA, et al. The 2016 revision of the World Health Organization classification of lymphoid neoplasms. *Blood.* 2016;127(20):2375-2390.
11. Beà S, Valdés-Mas R, Navarro A, et al. Landscape of somatic mutations and clonal evolution in mantle cell lymphoma. *Proc Natl Acad Sci USA.* 2013;110(45):18250-18255.
12. Wu C, de Miranda NF, Chen L, et al. Genetic heterogeneity in primary and relapsed mantle cell lymphomas: impact of recurrent CARD11 mutations. *Oncotarget.* 2016;7(25):38180-38190.
13. Zhang J, Jima D, Moffitt AB, et al. The genomic landscape of mantle cell lymphoma is related to the epigenetically determined chromatin state of normal B cells. *Blood.* 2014;123(19):2988-2996.
14. Meissner B, Kridel R, Lim RS, et al. The E3 ubiquitin ligase UBR5 is recurrently mutated in mantle cell lymphoma. *Blood.* 2013;121(16):3161-3164.
15. Kridel R, Meissner B, Rogic S, et al. Whole transcriptome sequencing reveals recurrent NOTCH1 mutations in mantle cell lymphoma. *Blood.* 2012;119(9):1963-1971.
16. Queirós AC, Beekman R, Vilarrasa-Blasi R, et al. Decoding the DNA methylome of mantle cell lymphoma in the light of the entire B cell lineage. *Cancer Cell.* 2016;30(5):806-821.
17. Clot G, Jares P, Giné E, et al. A gene signature that distinguishes conventional and leukemic nonnodal mantle cell lymphoma helps predict outcome. *Blood.* 2018;132(4):413-422.
18. Soldini D, Valera A, Solé C, et al. Assessment of SOX11 expression in routine lymphoma tissue sections: characterization of new monoclonal antibodies for diagnosis of mantle cell lymphoma. *Am J Surg Pathol.* 2014;38(1):86-93.
19. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics.* 2009;25(14):1754-1760.
20. Puente XS, Beà S, Valdés-Mas R, et al. Non-coding recurrent mutations in chronic lymphocytic leukaemia. *Nature.* 2015;526(7574):519-524.
21. McKenna A, Hanna M, Banks E, et al. The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* 2010;20(9):1297-1303.
22. Nik-Zainal S, Van Loo P, Wedge DC, et al; Breast Cancer Working Group of the International Cancer Genome Consortium. The life history of 21 breast cancers [published correction appears in *Cell*. 2015;162(4):924]. *Cell.* 2012;149(5):994-1007.
23. Moncunill V, Gonzalez S, Beà S, et al. Comprehensive characterization of complex structural variations in cancer by directly comparing genome sequence reads. *Nat Biotechnol.* 2014;32(11):1106-1112.
24. Layer RM, Chiang C, Quinlan AR, Hall IM. LUMPY: a probabilistic framework for structural variant discovery. *Genome Biol.* 2014;15(6):R84.
25. Maura F, Degasperi A, Nadeu F, et al. A practical guide for mutational signature analysis in hematological malignancies [published correction appears in *Nat Commun.* 2019;10(1):3431]. *Nat Commun.* 2019;10(1):2969.
26. Alexandrov LB, Kim J, Haradhvala NJ, et al; PCAWG Consortium. The repertoire of mutational signatures in human cancer. *Nature.* 2020;578(7793):94-101.
27. Shinde J, Bayard Q, Imbeaud S, et al. Palimpsest: an R package for studying mutational and structural variant signatures along clonal evolution in cancer. *Bioinformatics.* 2018;34(19):3380-3381.
28. Mermel CH, Schumacher SE, Hill B, Meyerson ML, Beroukhim R, Getz G. GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. *Genome Biol.* 2011;12(4):R41.
29. Landau DA, Tausch E, Taylor-Weiner AN, et al. Mutations driving CLL and their evolution in progression and relapse. *Nature.* 2015;526(7574):525-530.
30. Martín-Garcia D, Navarro A, Valdés-Mas R, et al. *CCND2* and *CCND3* hijack immunoglobulin light-chain enhancers in cyclin D1$^-$ mantle cell lymphoma. *Blood.* 2019;133(9):940-951.
31. Nadeu F, Mas-de-les-Valls R, Navarro A, et al. IgCaller for reconstructing immunoglobulin gene rearrangements and oncogenic translocations from whole-genome sequencing in lymphoid neoplasms. *Nat Commun.* 2020;11(1):3390.
32. Stunnenberg HG, Hirst M; International Human Epigenome Consortium. The International Human Epigenome Consortium: a blueprint for scientific collaboration and discovery [published correction appears in *Cell.* 2016;167(7):1897]. *Cell.* 2016;167(5):1145-1149.
33. Vilarrasa-Blasi R, Soler-Vila P, Verdaguer-Dot N, et al. Dynamics of genome architecture and chromatin function during human B cell

differentiation and neoplastic transformation. *bioRxiv*. 2019;764910.

34. Küppers R, Dalla-Favera R. Mechanisms of chromosomal translocations in B cell lymphomas. *Oncogene*. 2001;20(40):5580-5594.

35. Welzel N, Le T, Marculescu R, et al. Templated nucleotide addition and immunoglobulin JH-gene utilization in t(11;14) junctions: implications for the mechanism of translocation and the origin of mantle cell lymphoma. *Cancer Res*. 2001;61(4):1629-1636.

36. Greisman HA, Lu Z, Tsai AG, Greiner TC, Yi HS, Lieber MR. IgH partner breakpoint sequences provide evidence that AID initiates t(11;14) and t(8;14) chromosomal breaks in mantle cell and Burkitt lymphomas. *Blood*. 2012;120(14):2864-2867.

37. Wlodarska I, Dierickx D, Vanhentenrijk V, et al. Translocations targeting CCND2, CCND3, and MYCN do occur in t(11;14)-negative mantle cell lymphomas. *Blood*. 2008;111(12):5683-5690.

38. Cho SW, Xu J, Sun R, et al. Promoter of lncRNA gene PVT1 is a Ttmor-suppressor DNA boundary element. *Cell*. 2018;173(6):1398-1412.e22.

39. Porter CC. Germ line mutations associated with leukemias. *Hematology Am Soc Hematol Educ Program*. 2016;2016:302-308.

40. Tiao G, Improgo MR, Kasar S, et al. Rare germline variants in ATM are associated with chronic lymphocytic leukemia. *Leukemia*. 2017;31(10):2244-2247.

41. Tort F, Hernàndez S, Beà S, et al. CHK2-decreased protein expression and infrequent genetic alterations mainly occur in aggressive types of non-Hodgkin lymphomas. *Blood*. 2002;100(13):4602-4608.

42. Nones K, Johnson J, Newell F, et al; Brisbane Breast Bank (BBB). Whole-genome sequencing reveals clinically relevant insights into the aetiology of familial breast cancers. *Ann Oncol*. 2019;30(7):1071-1079.

43. Petridis C, Arora I, Shah V, et al. Frequency of pathogenic germline variants in BRCA1, BRCA2, PALB2, CHEK2 and TP53 in ductal carcinoma in situ diagnosed in women under the age of 50 years. *Breast Cancer Res*. 2019;21(1):58.

44. Aran D, Toperoff G, Rosenberg M, Hellman A. Replication timing-related and gene body-specific methylation of active human genes. *Hum Mol Genet*. 2011;20(4):670-680.

45. Beerman I, Bock C, Garrison BS, et al. Proliferation-dependent alterations of the DNA methylation landscape underlie hematopoietic stem cell aging. *Cell Stem Cell*. 2013;12(4):413-425.

46. Landan G, Cohen NM, Mukamel Z, et al. Epigenetic polymorphism and the stochastic formation of differentially methylated regions in normal and cancerous tissues. *Nat Genet*. 2012;44(11):1207-1214.

47. Siegmund KD, Marjoram P, Woo Y-J, Tavaré S, Shibata D. Inferring clonal expansion and cancer stem cell dynamics from DNA methylation patterns in colorectal cancers. *Proc Natl Acad Sci USA*. 2009;106(12):4828-4833.

48. Spencer DH, Russler-Germain DA, Ketkar S, et al. CpG island hypermethylation mediated by DNMT3A is a consequence of AML progression. *Cell*. 2017;168(5):801-816.e13.

49. Yang Z, Wong A, Kuh D, et al. Correlation of an epigenetic mitotic clock with cancer risk. *Genome Biol*. 2016;17(1):205.

50. Zhou W, Dinh HQ, Ramjan Z, et al. DNA methylation loss in late-replicating domains is linked to mitotic cell division. *Nat Genet*. 2018;50(4):591-602.

51. Duran-Ferrer M, Clot G, Nadeu F, et al. The proliferative history shapes the DNA methylome of B-cell tumors and predicts clinical outcome. *bioRxiv*. 2020;2020.02.06.937383.

52. Walker BA, Wardell CP, Johnson DC, et al. Characterization of IGH locus breakpoints in multiple myeloma indicates a subset of translocations appear to occur in pregerminal center B cells. *Blood*. 2013;121(17):3413-3419.

53. Hogenbirk MA, Heideman MR, de Rink I, et al. Defining chromosomal translocation risks in cancer. *Proc Natl Acad Sci USA*. 2016;113(26):E3649-E3656.

54. Daniel JA, Nussenzweig A. The AID-induced DNA damage response in chromatin. *Mol Cell*. 2013;50(3):309-321.

55. Lieber MR. Mechanisms of human lymphoid chromosomal translocations. *Nat Rev Cancer*. 2016;16(6):387-398.

56. Lu Z, Lieber MR, Tsai AG, et al. Human lymphoid translocation fragile zones are hypomethylated and have accessible chromatin. *Mol Cell Biol*. 2015;35(7):1209-1222.

57. Tramentozzi E, Ferraro P, Hossain M, Stillman B, Bianchi V, Pontarin G. The dNTP triphosphohydrolase activity of SAMHD1 persists during S-phase when the enzyme is phosphorylated at T592. *Cell Cycle*. 2018;17(9):1102-1114.

58. Yamazaki T, Liu L, Manley JL. TCF3 mutually exclusive alternative splicing is controlled by long-range cooperative actions between hnRNPH1 and PTBP1. *RNA*. 2019;25(11):1497-1508.

59. Agarwal R, Chan Y-C, Tam CS, et al. Dynamic molecular monitoring reveals that SWI-SNF mutations mediate resistance to ibrutinib plus venetoclax in mantle cell lymphoma. *Nat Med*. 2019;25(1):119-129.

60. Beà S, Tort F, Pinyol M, et al. BMI-1 gene amplification and overexpression in hematological malignancies occur mainly in mantle cell lymphomas. *Cancer Res*. 2001;61(6):2409-2412.

61. Hilton LK, Tang J, Ben-Neriah S, et al. The double-hit signature identifies double-hit diffuse large B-cell lymphoma with genetic events cryptic to FISH. *Blood*. 2019;134(18):1528-1532.

62. Nagel I, Szczepanowski M, Martín-Subero JI, et al. Deregulation of the telomerase reverse transcriptase (TERT) gene by chromosomal translocations in B-cell malignancies. *Blood*. 2010;116(8):1317-1320.

63. Aukema SM, Hoster E, Rosenwald A, et al. Expression of TP53 is associated with the outcome of MCL independent of MIPI and Ki-67 in trials of the European MCL Network. *Blood*. 2018;131(4):417-420.

64. Delfau-Larue M-H, Klapper W, Berger F, et al; European Mantle Cell Lymphoma Network. High-dose cytarabine does not overcome the adverse prognostic value of CDKN2A and TP53 deletions in mantle cell lymphoma. *Blood*. 2015;126(5):604-611.

65. Eskelund CW, Dahl C, Hansen JW, et al. *TP53* mutations identify younger mantle cell lymphoma patients who do not benefit from intensive chemoimmunotherapy. *Blood*. 2017;130(17):1903-1910.

66. Ferrero S, Rossi D, Rinaldi A, et al. KMT2D mutations and TP53 disruptions are poor prognostic biomarkers in mantle cell lymphoma receiving high-dose therapy: a FIL study. *Haematologica*. 2019;105(6):1604-1612.

Others

Other articles published during the doctoral thesis

1. *Ordoñez R, *Kulis M, Russiñol N, Chapaprieta V, Carrasco-Leon A, García-Torre B, Charalampopoulou S, Clot G, Beekman R, Meydan C, **Duran-Ferrer M**, Verdaguer-Dot N, Vilarrasa-Blasi R, Soler-Vila P, Garate L, Miranda E, José-Enériz ES, Rodriguez-Madoz JR, Ezponda T, Martínez-Turrilas R, Vilas-Zornoza A, Lara-Astiaso D, Dupéré-Richer D, Martens JHA, El-Omri H, Taha RY, Calasanz MJ, Paiva B, Miguel JS, Flicek P, Gut I, Melnick A, Mitsiades CS, Licht JD, Campo E, Stunnenberg HG, **#**Agirre X & **#**Prosper F & **#**Martin-Subero JI. Chromatin activation as a unifying principle underlying pathogenic mechanisms in multiple myeloma. Genome Res. 2020 Aug 20. doi: 10.1101/gr.265520.120. Epub ahead of print. PMID: 32820006.

2. Tsagiopoulou M, Chapaprieta V, **Duran-Ferrer M**, Moysiadis T, Psomopoulos F, Kollia P, Papakonstantinou N, Campo E, Stamatopoulos K, #Martin-Subero JI. Chronic lymphocytic leukemias with trisomy 12 show a distinct DNA methylation profile linked to altered chromatin activation. Haematologica. 2020 Feb 27:haematol.2019.240721. doi: 10.3324/haematol.2019.240721. Epub ahead of print. PMID: 32107330

3. ***Tsagiopoulou M, ***Papakonstantinou N, Moysiadis T, Mansouri L, Ljungström V, **Duran-Ferrer M**, Malousi A, Queirós AC, Plevova K, Bhoi S, Kollia P, Oscier D, Anagnostopoulos A, Trentin L, Ritgen M, Pospisilova S, Stavroyianni N, Ghia P, Martin-Subero JI, Pott C, Rosenquist R, **#**Stamatopoulos K. DNA methylation profiles in chronic lymphocytic leukemia patients treated with chemoimmunotherapy. Clin Epigenetics. 2019 Dec 2;11(1):177. doi: 10.1186/s13148-019-0783-1. PMID: 31791414; PMCID: PMC6889736.

4. Papakonstantinou N, Ntoufa S, Tsagiopoulou M, Moysiadis T, Bhoi S, Malousi A, Psomopoulos F, Mansouri L, Laidou S, Papazoglou D, Gounari M, Pasentsis K, Plevova K, Kuci-Emruli V, **Duran-Ferrer M**, Davis Z, Ek S, Rossi D, Gaidano G, Ritgen M, Oscier D, Stavroyianni N, Pospisilova S, Davi F, Ghia P, Hadzidimitriou A, Belessi C, Martin-Subero JI, Pott C, Rosenquist R, **#**Stamatopoulos K. Integrated epigenomic and transcriptomic analysis reveals TP63 as a novel player in clinically aggressive chronic lymphocytic leukemia. Int J Cancer. 2019 Jun 1;144(11):2695-2706. doi: 10.1002/ijc.31999. Epub 2019 Jan 15. PMID: 30447004.

5. Beekman R, Chapaprieta V, Russiñol N, Vilarrasa-Blasi R, Verdaguer-Dot N, Martens JHA, **Duran-Ferrer M**, Kulis M, Serra F, Javierre BM, Wingett SW, Clot G, Queirós AC, Castellano G, Blanc J, Gut M, Merkel A, Heath S, Vlasova A, Ullrich S, Palumbo E, Enjuanes A, Martín-García D, Beà S, Pinyol M, Aymerich M, Royo R, Puiggros M,

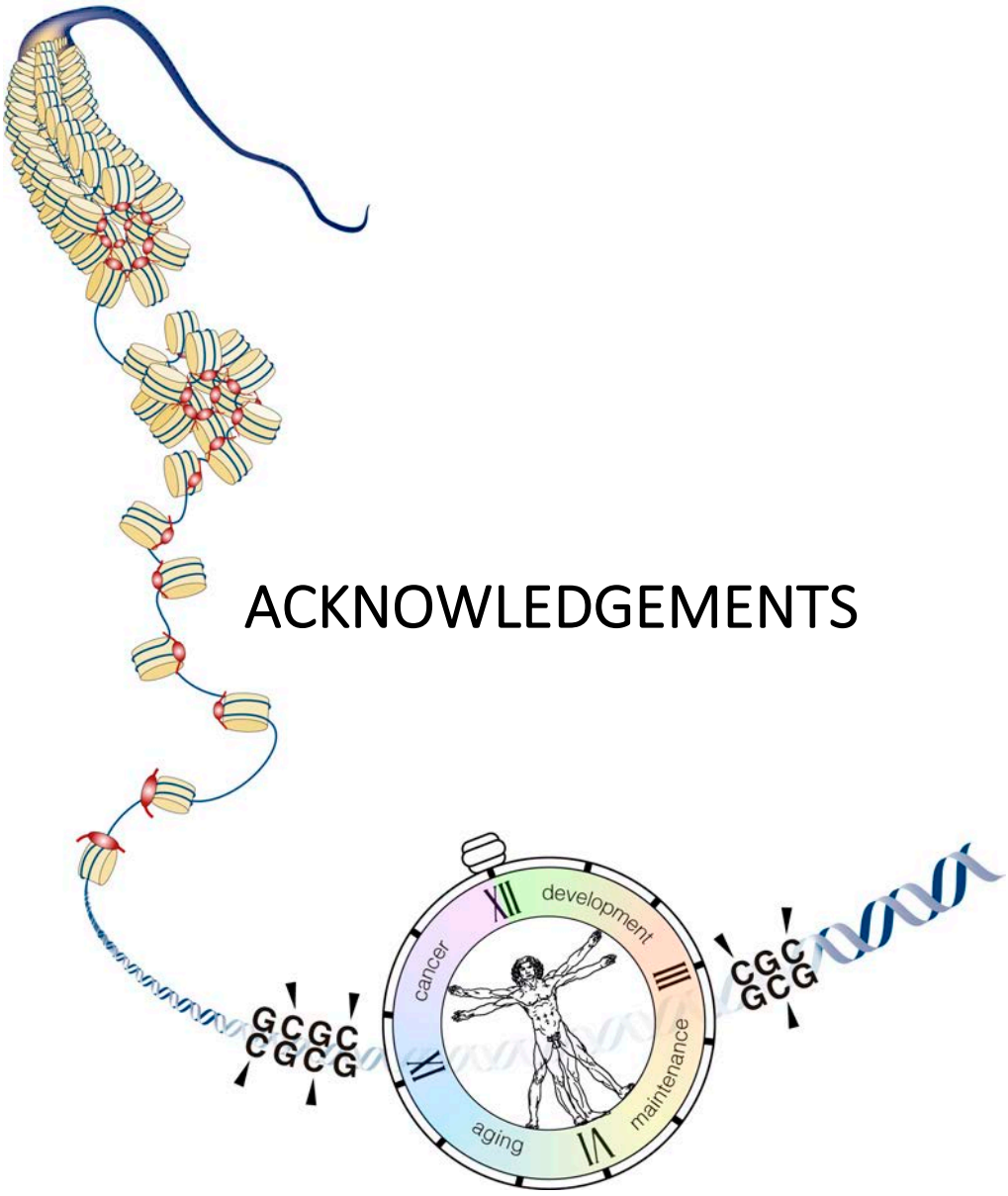
Torrents D, Datta A, Lowy E, Kostadima M, Roller M, Clarke L, Flicek P, Agirre X, Prosper F, Baumann T, Delgado J, López-Guillermo A, Fraser P, Yaspo ML, Guigó R, Siebert R, Martí-Renom MA, Puente XS, López-Otín C, Gut I, Stunnenberg HG, Campo E, #Martin-Subero JI. The reference epigenome and regulatory chromatin landscape of chronic lymphocytic leukemia. Nat Med. 2018 Jun;24(6):868-880. doi: 10.1038/s41591-018-0028-4. Epub 2018 May 21. PMID: 29785028; PMCID: PMC6363101.

Presentations at congresses during the doctoral thesis

1. Poster presentation at the 5th ESH International Conference on New Concepts in Lymphoid Malignancies Estoril, Portugal, from October 3th to 5th, 2019.
2. Oral presentation at 24th EHA annual Congress, 2019 Amsterdam.
3. Poster presentation at 60th ASH annual Meeting and Exposition, San Diego, USA, 2018.
4. Pitch poster presentation at 22nd European Hematology Association (EHA) Congress 2017, Madrid.

Attendance to congresses during the doctoral thesis

1. International Conference on New Concepts in Lymphoid Malignancies Estoril, Portugal, from October 3th to 5th, 2019.
2. ERIC international meeting: New frontiers in CLL Research, Barcelona, 2018.
3. ESH 4th Annual International Conference on New Concepts in Lymphoid Malignancies: Focus on Aggressive Lymphomas, Dublin, 2018.
4. Epigenetic mechanisms in health and disease. BCEC, Cosmocaixa. Barcelona 2017.
5. IV Symposium of Bioinformatics and Genomics, Catalan Society de Biologica, PRBB, Barcelona, 20th December 2016.
6. 5th CNAG Symposium on Genome Research: Single Cell Studies, Auditorium of the Barcelona Science Park 19th May 2016.
7. Coding and Non-Coding Functions of the Genome. 2015 BCEC, Cosmocaixa. Barcelona 2015.
8. The fourth annual BLUEPRINT consortium meeting. Hinxton, (UK). September 2015.



ACKNOWLEDGEMENTS

Sembla ahir quan vaig emprendre aquest viatge, i que el dia d'avui no havia d'arribar, però ja és aquí. Miro endarrere aquests cinc anys de tesi doctoral i podria escriure'n unes quantes amb les experiències viscudes. Al cap i la fi, una tesi doctoral és més que un exercici d'erudits, és una etapa vital de maduració personal i que allò que n'aprenem, més enllà de la part acadèmica, ens acompanya durant tota la vida.

Les persones que heu contribuït a que aquesta tesi doctoral arribi a bon port, front tota mena d'adversitats, sou moltes, i encara que no vos mencioni explícitament, espero que totes vos hi pugueu sentir identificades. Sense vosaltres, no hauria arribat aquí!

Encara que hagi realitzat el doctorat a l'IDIBAPS, no hauria estat possible sense el meu pas anterior pel CNAG. *Thanks to all CNAG people that in one way or another helped me in progressing towards this doctoral thesis. Thank you for giving me the opportunity to pursue the prelude of this doctoral thesis and to be passionate about DNA methylation!*

Mil gràcies a tots vosaltres, els de cada dia, *divos* i *divas*, els que hem compartit *meetings*, poiates i teclats, alegries i tristeses, en aquest món de la Ciència. Molt més que això, gràcies per compartir dinars, sushi, cases rurals, vídeos de tesi, disfresses, cafès, pastissos del *Milky*, jocs de taula, *memes*, infusions putrefactes, amic invisibles, concerts *barna rock*, escalades, caminades, cimes, cerveses, braves, olives, tast de vins i formatges amb fongs, *scape rooms*, congressos, viatges, festivals de terror de Sants, cures del càncer, figures *patchwork*, hamburgueses del *Foc i Oli*, concursos de galetes, ofertes de Groupon, San Fermines, etc, etc, etc. Gràcies per aquests moments i mil més que formen part d'aquesta tesi!

Gràcies per les interrupcions, les del mig dia i les de la tarda, buscant algú, buscant xocolata i, sobretot, converses de to pujat. Gràcies per deixar-me l'àngel estadístic de la guard, i gràcies a tu per ser-ho. Perquè al final, tot ja esta dit, i només importen els *CNAs. Thank you for being my neighbor these years, for being an example of perseverance and dedication, for the scientific discussions and for all the simple but grateful daily-life conversations.* A vosotros, gracias por los *meetings* y la puesta a punto de los datos clínicos, por mis visitas al despacho de l'Hospital Clínico, y por apoyar y guiar los análisis clínicos porque, al fin y al cabo, nuestras elucubraciones siempre han pretendido en última instancia mejorar la calidad de vida de los pacientes.

Gràcies a tu, admirador del meu compatriota Llorenç Villalonga, *the Boss,* gràcies pel teu preciosíssim temps, pels brindis amb cerveses *Guinness* a Dublín, per les meditacions científiques a Montserrat, al teu despatx de l'Hospital Clínic amb el microscopi al fons o al teu despatx de l'IDIBAPS, on mai han faltat ni xocolata ni cafè. Un despatx sense xocolata fa més por que una pedregada!

Compañeroooooooo, gracias a ti! Gracias por desear antes que nada mi bienestar. Por empatizar conmigo en todas las etapas del doctorado, algunas de ellas especialmente duras académica y personalmente. Por confiar en mi y darme las oportunidades que me has dado, de viajar, aprender, publicar, de ponernos cachondos con algún análisis, *heatmap* o figura cañera, de comer sushi hasta acelerar por momentos nuestra edad epigenética, de enseñarme ciencia de verdad en las *Young EHA parties*, en Chueca o en Ámsterdam, con o sin *tutu*, y de compartir todos estos momentos y visiones de la vida más allá del mundo académico.

Gràcies a tots vosaltres, als de fora, els qui no heu estat al laboratori però sí heu estat amb mi. Als que hi heu estat sempre, els que hi sou de fa més poc i als qui, per desgràcia, la distància, el temps, o la vida ens separa.

I sobretot, gràcies a vosaltres, aquells qui m'ho heu donat tot des del moment zero, que m'heu vist créixer i menjar molt, i m'heu ensenyat a caminar en aquesta vida. Ho heu fet tot possible, i perquè per a vosaltres, doctor o no, sempre seré en Martinet. I gràcies a tu, perquè sense en Zipi no hi ha en Zape, ni piruetes impossible a n'es Mondeo, ni joc de llums a l'estil OT, ni xalets de sofà, ni suport en hores baixes.

A tots, gràcies de tot cor!

Martí Duran Ferrer

Life is not easy for any of us. But what of that? We must have perseverance and above all confidence in ourselves. We must believe that we are gifted for something and that this thing must be attained.

-Marie Curie