

# Exploring methods to measure DNA methylation in the context of HIV-1

Sam Kint



GHENT  
UNIVERSITY

FACULTY OF MEDICINE  
AND HEALTH SCIENCES



FACULTY OF  
BIOSCIENCE ENGINEERING

# Exploring methods to measure DNA methylation in the context of HIV-1

Sam Kint



GHENT  
UNIVERSITY



FACULTY OF MEDICINE  
AND HEALTH SCIENCES



FACULTY OF  
BIOSCIENCE ENGINEERING

**Promoters: Prof. dr. ir. Wim Van Criekinge**

Laboratory for Bioinformatics and Computational Genomics (BioBix), Department of Data Analysis and Mathematical Modelling, Faculty of Bioscience Engineering, Ghent University

**Prof. dr. Linos Vandekerckhove**

**Dr. Wim Trypsteen**

HIV Cure Research Center (HCRC), Department of Internal Medicine and Pediatrics, Faculty of Medicine and Health Sciences, Ghent University

**Dean: Prof. dr. ir. Marc Van Meirvenne**

**Rector: Prof. dr. ir. Rik Van de Walle**

# **Exploring methods to measure DNA methylation in the context of HIV-1**

Sam Kint

Thesis submitted to fulfill the requirements for the degree of Doctor (PhD) of Bioscience  
Engineering: Biotechnology

## Dutch translation of the title

Evaluatie van methoden voor DNA methylatie analyse met een focus op HIV-1

## To refer to this thesis:

Kint S. (2019). Exploring methods to measure DNA methylation in the context of HIV-1. PhD thesis. Ghent university, Ghent, Belgium.

## Cover image:

The 'epigenetic landscape' as described by Waddington in 'The Strategy of The Genes' (1957). **Front:** the original illustration shows a biological system (a cell) at the start of its developmental journey taking one of the paths towards a new, more differentiated state. The cell is on the front cover replaced with an HIV-1 virion: after infection of a cell, every virus takes its own path within the host cell. **Back:** the view from the behind of the epigenetic landscape illustrates the complexity of the interactions that shape the landscape. The pegs in the ground (genes) are attached to interconnected guy-ropes (interacting gene products), whose tension determines the shape of the slopes within the landscape.

The author and the promotors give the authorization to consult and to copy parts of this work for personal use only. Every other use is subject to copyright laws. Permission to reproduce any material contained in this work should be obtained from the author.

## Author:

ir. Sam Kint

## Promoters:

Prof. dr. ir. Wim Van Criekinge

Prof. dr. Linos Vandekerckhove

Dr. Wim Trypsteen

# Members of the examination committee

## Promoters

### **Prof. dr. ir. Wim Van Criekinge**

Department of Data Analysis and Mathematical Modelling, Faculty of Bioscience Engineering, Ghent University, Belgium

### **Prof. dr. Linos Vandekerckhove**

Department of Internal Medicine and Pediatrics, Faculty of Medicine and Health Sciences, Ghent University, Belgium

### **Dr. Wim Trypsteen**

Department of Internal Medicine and Pediatrics, Faculty of Medicine and Health Sciences, Ghent University, Belgium

## Members of the Jury

### **Prof. dr. Godelieve Gheysen**

Department of Biotechnology, Faculty of Bioscience Engineering, Ghent University, Belgium

### **Prof. dr. ir. Tina Kyndt**

Department of Biotechnology, Faculty of Bioscience Engineering, Ghent University, Belgium

### **Prof. dr. ir. Tim De Meyer**

Department of Data Analysis and Mathematical Modelling, Faculty of Bioscience Engineering, Ghent University, Belgium

### **Prof. dr. Ward de Spiegelaere**

Department of Morphology, Faculty of Veterinary Medicine, Ghent University, Belgium

### **Prof. dr. Oleg Denisenko**

Department of Medicine, Division of Allergy & Infectious Diseases, University of Washington, Seattle, Washington

### **Prof. dr. Zeger Debyser**

Department of Pharmaceutical and Pharmacological Sciences, Faculty of Medicine, Catholic University Leuven, Belgium



# Table of Contents

<b>Abbreviations.....</b>	<b>I</b>
<b>Preface.....</b>	<b>VII</b>
<b>I – Introduction.....</b>	<b>1</b>
<b>I – Epigenetics: DNA methylation.....</b>	<b>1</b>
Historical setting .....	1
Epigenetic regulation mechanisms .....	3
DNA methylation .....	5
DNA methylation process .....	5
CpG suppression .....	6
Transcriptional regulation by DNA methylation .....	7
DNA methylation analysis methods .....	9
Global DNA methylation assessment .....	11
DNA methylation assessment using protein-DNA interactions .....	11
Single-molecule real-time sequencing.....	12
Oligonucleotide probes.....	13
Chemical methods.....	13
Bisulfite treatment.....	14
Single cell DNA methylation methods.....	16
<b>II – HIV/AIDS .....</b>	<b>21</b>
Historical setting .....	21
HIV classification .....	24
HIV-1 genomic structure .....	25
HIV-1 virion structure .....	26
HIV-1/AIDS disease progression .....	27
I – acute phase.....	29
II – asymptomatic phase .....	29
III – symptomatic phase .....	30
HIV-1 life cycle .....	30
HIV-1 life cycle revised .....	33
I – pre-integrative latency .....	34
II – post-integrative latency .....	34
III – replication-deficiency.....	36
Viral reservoir.....	37
I – cellular reservoirs .....	37
II – anatomical reservoirs .....	37



HIV-1 treatment .....	38
HIV-1 cure.....	40
<b>III – DNA methylation in HIV-1 proviral transcription regulation.....</b>	<b>45</b>
Epigenetic features of the HIV-1 genome .....	45
Nucleosome binding at the HIV-1 5'LTR.....	45
CpGIs in the HIV-1 genome .....	46
DNA methylation in HIV-1 .....	47
Challenges of DNA methylation analysis of the HIV-1 provirus.....	48
<b>II – Research Objectives .....</b>	<b>53</b>
<b>Objective I – Bisulfite-based HIV-1 proviral DNA methylation assay .....</b>	<b>54</b>
Objective Ia – Bisulfite treatment optimization .....	54
Objective Ib – Amplification of HIV-1 CpGIs.....	54
Objective Ic – Analysis of DNA methylation in HIV-1 patient cohorts .....	55
<b>Objective II – Single cell DNA methylation assay .....</b>	<b>56</b>
<b>III – Results.....</b>	<b>61</b>
<b>Objective I – Bisulfite-based DNA methylation assay .....</b>	<b>61</b>
Objective Ia – Bisulfite treatment optimization .....	62
Objective Ib & Ic – Amplification of HIV-1 CpGIs in vivo .....	90
<b>Objective II – Single cell DNA methylation assay .....</b>	<b>111</b>
Single cell epigenetic visualization assay, EVA.....	112
<b>IV – Discussion and future perspectives .....</b>	<b>137</b>
<b>Understanding HIV-1 latency: epigenetics .....</b>	<b>139</b>
Bisulfite-based HIV-1 proviral DNA methylation assay .....	140
Limitations of the HIV-1 methylation assay.....	141
Assay development .....	141
Study design.....	142
<b>Single cell epigenetic visualization assay.....</b>	<b>143</b>
EVA methodology .....	143
Versatility of EVA.....	143
Future assay improvements .....	144
<b>Impact of the assay development .....</b>	<b>146</b>
Bisulfite-based assay optimization.....	146
In vivo HIV-1 proviral DNA methylation.....	146
EVA development .....	147

Insights in HIV-1 latency regulation by DNA methylation .....	147
<b>Epigenetics in HIV-1 (cure) research.....</b>	<b>149</b>
Shock and lock .....	149
Linking different epigenetic modifications .....	149
Targeting epigenetic interventions .....	150
Changing HIV-1 proviral DNA methylation in vivo.....	151
<b>DNA methylation research – next steps .....</b>	<b>152</b>
<b>Conclusion .....</b>	<b>153</b>
<b>Bibliography.....</b>	<b>154</b>
<b>Summary .....</b>	<b>172</b>
<b>Samenvatting .....</b>	<b>174</b>
<b>Dankwoord .....</b>	<b>178</b>



# Abbreviations

A	Adenine	hmC	Hydroxymethylcytosine
AID	Activation-induced cytosine deaminase	HMTi	Histone methyl transferase inhibitor
AIDS	Acquired immunodeficiency syndrome	HSCT	Hematopoietic stem cell transplantation
APOBEC	Apolipoprotein B mRNA editing enzyme, catalytic polypeptide-like	HTLV	Human T-cell Leukemia Virus
ARV	AIDS-associated retrovirus	INSTI	Integrase strand transfer inhibitor
ASORF	Antisense open reading frame	IS	Integration site
ASP	Antisense protein	LAV	Lymphadenopathy-associated virus
BER	Base excision repair	lncRNA	Long non-coding RNA
bp	Base pair	LRA	Latency reversing agent
C	Cytosine	LTNP	Long-term non-progressor
caC	Carboxylcytosine	LTR	Long terminal repeat
cART	Combination antiretroviral therapy	MBD	Methyl-CpG-binding domain
cDNA	Complementary DNA	mC	Methylated cytosine
ChIP	Chromatin immunoprecipitation	MDA	Multiple displacement amplification
COBRA	Combined bisulfite conversion and restriction analysis	MeDIP	Methylated DNA immunoprecipitation
CpGI	CpG island	MID-RRBS	Microfluidic diffusion-based reduced representation bisulfite sequencing
CRE	Cis-regulatory element	MSP	Methylation specific PCR
CRISPR/	clustered regularly interspaced short	NCR	Non-coding region
dCas9	palindromic repeats/dead CRISPR-associated protein 9	NELF	Negative elongation factor
DNMT	DNA methyltransferases	NGS	Next-generation sequencing
dsDNA	Double Stranded DNA	NNRTi	Non-nucleoside reverse transcriptase inhibitor
DSIF	DRB-sensitive inducing factor	NRTi	Nucleoside reverse transcriptase inhibitor
EC	Elite controller	NuRD	Nucleosome remodeling and deacetylase complex
ETR	Env-Tat-Rev	P-TEFb	Positive transcription elongation factor b
EZH2	HMT enhancer of zeste homolog 2	PBAT	Post-bisulfite adaptor tagging
fC	Formylcytosine	PBMC	Peripheral blood mononuclear cell
FISH	Fluorescence in situ hybridization	PcG	Polycomb group
G	Guanine	PCR	Polymerase chain reaction
GALT	Gut-associated lymphoid tissue	PI	Protease inhibitor
HDAC	Histone deacetylase	PIC	Pre-integration complex
HDACi	Histone deacetylase inhibitor	PKC	Protein kinase C
HIV	Human immunodeficiency virus	PrEP	Pre-exposure prophylaxis
		qPCR	Quantitative polymerase chain reaction

qRRBS	Quantitative reduced representation bisulfite sequencing	Tr	Regulatory T cell
rDNA	Ribosomal DNA	Trm	Tissue resident memory T cell
RP-HPLC	Reversed-phase high-performance liquid chromatography	TrxG	Trithorax groups
RRBS	Reduced representation bisulfite sequencing	Tscm	Stem cell memory T cell
RRE	Rev responsive element	TSS	Transcription start site
RT	Reverse transcriptase	Ttd	Terminally differentiated memory T cell
RTC	Reverse transcription complex	Ttm	Transitional memory T cell
SAH	S-adenosyl-homocysteine	U	Uracil
SAM	S-adenosyl methionine	usRNA	Unspliced RNA
sci-MET	Single-cell combinatorial indexing for methylation analysis	VC	Viremic controller
scPBAT	Single cell post-bisulfite adaptor tagging	VL	Viral load
scRRBS	Single cell reduced representation bisulfite sequencing	WGBS	Whole genome bisulfite sequencing
scWGBS	Single cell whole genome bisulfite sequencing	WHO	World health organization
SIV	Simian immunodeficiency virus	ZMW	Zero-mode waveguides
SLBS	sSingle-cell locus-specific bisulfite sequencing		
SMRT	Single-molecule real-time sequencing		
snmC-seq	Single-nucleus mC sequencing		
ssDNA	Single stranded DNA		
T	Thymine		
TALE	Transcription activator-like effectors		
Tcm	Central memory T cell		
TDG	Thymine DNA glycosylase		
TE	Transposable element		
Tem	Effector memory T cell		
TET	Ten-eleven translocation		
TF	Transcription factor		
Tfh	Follicular T helper cell		
Th	T helper cell		
TLR	Toll-like receptor		
Tmm	Migratory memory T cell		
Tn	Naïve CD4+ T cell		





# PREFACE





# Preface

Epigenetics is the combination of mechanisms that control and regulate gene transcription in every cell without altering the genome itself: it adds a layer of information to the fixed genetic code within every cell. This regulation happens through a whole set of reversible modifications at the chromatin level. These modifications alter the chromatin conformation, which brings genes from an active state into an inactive state and vice versa. Studying these modifications can provide insights in a lot of diseases and in development of organisms.










One of these diseases is the infection with human immunodeficiency virus. This virus has claimed over 37 million lives and 36.9 million people are living with it today. Although the research community has done a lot of efforts to end this pandemic, still no effective cure strategy has been developed. The current therapy can stop the viral replication, but infected individuals are doomed to lifelong antiretroviral treatment to prevent the virus from being reactivated and eventually kill them. This chronic character of the infection is due to the formation of a latent viral reservoir, which resides all over the body of the patient. In this latent reservoir, the viral genome, which is integrated stably into the host genome, is not transcribed – no viral products are produced despite the presence of the genetic information.

Studying the epigenetic influences that target the viral genome, and help silencing the transcription, will provide insights in this viral latency, and might open perspectives towards a cure: interacting with these mechanisms might help reversing the latency, or forcing the virus into a deep latent state in patients.

To be able to study these modifications, decent analysis methods have to be developed. During my PhD, I have focused on the development of methods to analyze a specific epigenetic modification: DNA methylation. I contributed to the development of two different methodologies: a targeted bisulfite-based next generation sequencing method to measure viral DNA methylation and a targeted single-cell epigenetic visualization assay to measure both human and viral methylation.



CHAPTER I  
INTRODUCTION

- 1925**  Cytosine methylation is described
- 1942**  Introduction of the term 'epigenetics' by Waddington
- 1951**  mC is found in animal and plant DNA
- 1962**  CpG as main target for methylation in mammalian cells
- 1975**  mC is an epigenetic mark: X-chromosome inactivation
- 1987**  DNA methylation as a modulator of HIV expression
- 1990**  Holliday: epigenetics as an additional layer of information upon the genome
- 1994**  Holliday: phenotypic inheritability not based on changes in DNA sequences
- 2009**  Berger: operational definition for epigenetics  
Epigenator, epigenetic initiator, epigenetic maintainer

# I – Introduction

## I – Epigenetics: DNA methylation

### Historical setting

Epigenetics was first described by Conrad Hal Waddington in 1942 as “*the branch of biology which studies the causal interactions between genes and their products, which bring the phenotype into being*” (1,2). He referred to embryological development that must involve networks of gene interactions: a single cell (a fertilized egg) can develop to any organ, by following its predestined route of development (1,2). The course of this development is controlled by genes, but more importantly, by complex interactions between different genes, by plasticity of the genes and by environmental influences, showing that there is no simple relationship between a gene and its phenotypic effects, which is the basis of epigenetics (2,3).

Epigenetics results in plasticity: genetically identical cells differ heavily in structure and function. Liver cells, blood cells or skin cells are phenotypically completely different, and their variation is epigenetic, not genetic. This epigenetic variation originates from early in the differentiation from a stem cell and is stably inherited to the daughter cells. 48 years after Waddington, Holliday described epigenetics as “*The study of mechanisms of temporal and spatial control of the gene during the development of complex organisms*“, which is in line with the definition of Waddington (2,4). However, he added the concept that epigenetics is an additional layer of information which is reversible, but very stably present upon the genome, hence the name ‘*epi*’-genetics or ‘*on*’-, ‘*upon*’-, ‘*over*’-genetics. Later, he redefined epigenetics as “*The study of changes in gene expression, which occur in organisms with differentiated cells, and of mitotic inheritance of specific patterns of gene expression*” (2,5). It covers the regulation of the transcription of the genetic code, which is a fixed code. This transcription has to be heavily controlled and regulated within every cell of every organism to regulate and maintain the plasticity of the transcript, and thus protein production of the fixed set of genes (2). Holliday added a supplementary definition: “*Nuclear inheritance which is not based on differences in DNA sequence*” (2).

In 2009, an operational definition for epigenetics was proposed by Berger and colleagues: “*An epigenetic trait is a stably heritable (through either mitosis or meiosis) phenotype resulting from changes in a chromosome without alterations in the DNA sequence*” (6). Moreover, Berger proposed three separated signal categories which are involved in the establishment of an epigenetic state: the ‘**epigenator**’, or environmental trigger that can form the base of the epigenetic signaling pathway (e.g. differentiation signals, environmental factors); the ‘**epigenetic initiator**’, which is the receiver of the epigenator and determines the exact chromatin location to establish an epigenetic pathway (e.g. DNA binding factors, non-coding RNAs); and the ‘**epigenetic maintainer**’, the mechanisms to sustain the epigenetic environment (chromatin conformation) in multiple generations which is the

'additional layer of information upon the genome' in the form of epigenetic modifications such as histone modifications or DNA methylation (methylated cytosines (mC)) (6).

The existence of mC was first shown in 1925 by Johnson and Coghill (7), and in 1951, Gerard Wyatt demonstrated its presence in plant and animal DNA using a simple chromatographic method (8). Eleven years later, it was shown that mainly Cs in CpG dinucleotides were methylated in mammalian cells (9) and only in 1975, it was proposed that this mC was an epigenetic mark, capable of causing X-chromosome inactivation (10,11). At the moment, DNA methylation is one of the most studied epigenetic modifications which are shown to be involved in several biological processes associated with development and disease. Next to X-chromosome inactivation, this modification plays a role in cell differentiation, regulation of gene expression and genomic imprinting (12–15). In disease, DNA methylation is heavily involved in the development of diseases as Parkinson's, Alzheimer, carcinogenesis and silencing of intracellular viruses such as the human immunodeficiency virus (HIV) (12,16–24).

Today, epigenetics is still described as the bridge between genotype and phenotype: a cellular regulatory mechanism that changes the final expression of a locus or chromosome, and thus the phenotype of a cell, without changing the underlying sequence by reversible, mitotically inheritable modifications (25–27). These mechanisms are crucial in an organism during development and during cell differentiation: stem cells can be altered epigenetically to form specialized cells that undergo a specific, well controlled amount of transcription of every gene, and thus produce the perfect concentration of specific proteins in order to survive and perform their function (2,28). Epigenetic alterations are influenced by environmental factors such as food, temperature, lifestyle, presence of pathogens, cell differentiation signals, hormones or cell density (e.g. during development) (26,27).

## Epigenetic regulation mechanisms

The DNA in an eukaryotic cell is not freely moving in the cell nucleus, but it is packed into a dense structure: chromatin (**Figure 1A**) (29–33). This is a complex of histone proteins and DNA, with nucleosomes as base unit. Nucleosomes are octamers of histones, binding DNA strands of 147 base pairs (bp) and separated by linker DNA (29–32). The DNA is tightly wrapped around these nucleosomes (**Figure 1A**). Chromatin architecture can be remodeled by chromatin modifiers, making the DNA accessible (euchromatin) or inaccessible (heterochromatin) for transcription factors (TF) (**Figure 1B**) (34–36). Heterochromatin can be constitutive or facultative: the former type is permanently inaccessible for transcription factors, causing constitutive silencing of the affected regions (e.g. centromeres, telomeres, some repetitive elements). The latter can change depending on environmental factors that influence the epigenetic environment (e.g. cell type, tissue type, stage of development or differentiation, lifestyle, hormones). The presence or absence of these specific epigenetic modifications is responsible to alter the accessibility of the chromatin and recruit proteins that alter transcriptional control, DNA repair and RNA processing at the affected gene. This is performed by changing the density of these regions and switch them from hetero- to euchromatin or vice versa (37,38).

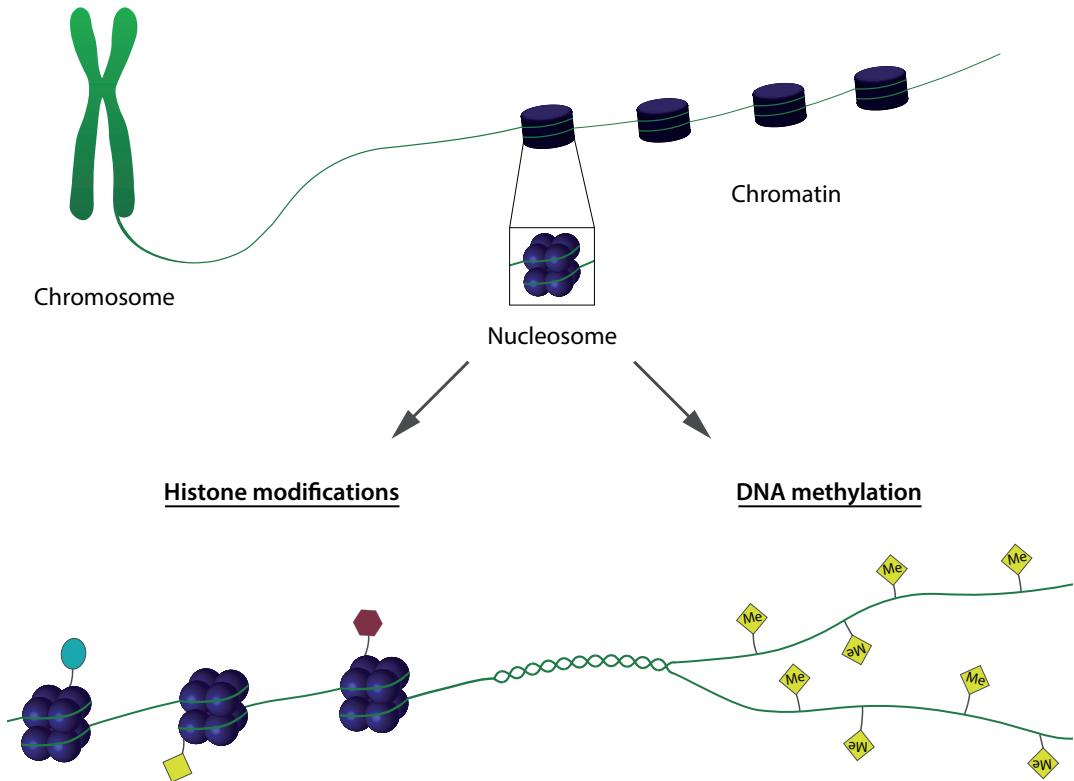
The DNA accessibility is regulated by two distinct classes of enzymes: (i) mediators of epigenetic modification such as covalent histone tail and DNA modifications, and (ii) ATP-consuming enzymes that mobilize nucleosomes and alter the chromatin architecture (38). Chromatin modifier complexes, which include the antagonistic polycomb groups (PcG) and trithorax groups (TrxG), induce respectively repressive and active epigenetic marks at genomic targets which contain polycomb/trithorax response elements (such as CpGs, pre-installed histone modifications, TF or long non-coding RNA (lncRNA)) (38–40). Several ATP-dependent chromatin remodeling complexes then catalyze energetically unfavorable reactions (as breaking bonds between negatively charged DNA and positively charged histones) to facilitate chromatin access by repositioning of nucleosomes, evicting histone dimers or ejecting histone octamers (38).

The most described epigenetic regulating mechanisms in mammalian genomes are stable, but reversible modifications of the histone amino-terminal domains (e.g. acetyl, methyl, phosphate, ubiquitin), the DNA (mainly DNA methylation) and the RNA (e.g. methylation) (41). By adding or removing small chemical groups at specific locations of proteins, RNA or DNA, their properties will alter to facilitate the regulation of gene transcription or translation processes (e.g. histone tail acetylation neutralizes the positive charge of the histone, which decreases the strength of the DNA-histone bond, facilitating DNA release from histones) (29,41).

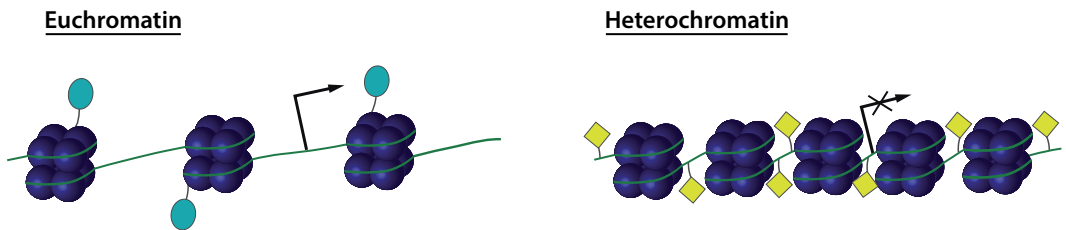
All these epigenetic modifications together form the epigenome, which is defined as a collaborative mechanism of different genome and protein modifications that change the genomic conformation (41). In this thesis, the main focus is on DNA methylation.



A



B



**Figure 1. The epigenome.** **A.** Genomic organisation within the nucleus. DNA, organised into chromosomes, is wrapped tightly around nucleosomes (histone octamers) to form chromatin. Epigenetic modifications happen both on DNA (e.g. DNA methylation) or on the histone tails (histone modifications). **B.** DNA methylation or histone modifications influence the density of the chromatin to form either euchromatin (open, accessible for transcription factors) or heterochromatin (dense, inaccessible for transcription factors).

## DNA methylation

DNA methylation is a well-studied epigenetic modification which implies the addition of a methyl group (CH<sub>3</sub>) to the number 5 carbon of the cytosine (C) pyrimidine ring (**Figure 2**). This modification usually happens on C in cytosine-guanine (CpG) dinucleotides and is shown to be crucial for normal cellular functioning, and it is heavily involved in development and aging of multicellular organisms by regulating crucial mechanisms as X-chromosome inactivation, cell differentiation, gene expression, genomic imprinting and suppression of parasitic and other repeat sequences (12–15,27,42–45). The methylation pattern is highly cell type dependent and irregularities in the methylation profile of a cell lead to dysregulation of normal cellular transcription, resulting in a variety of diseases, including cancer, autoimmune diseases, metabolic disorders, neurological disorders and Alzheimer. Moreover, it is shown to be involved in the regulation of transcription of (integrated) proviral genomes (e.g. in HIV or Herpes viruses) (10,11,46,16,17,19–24).

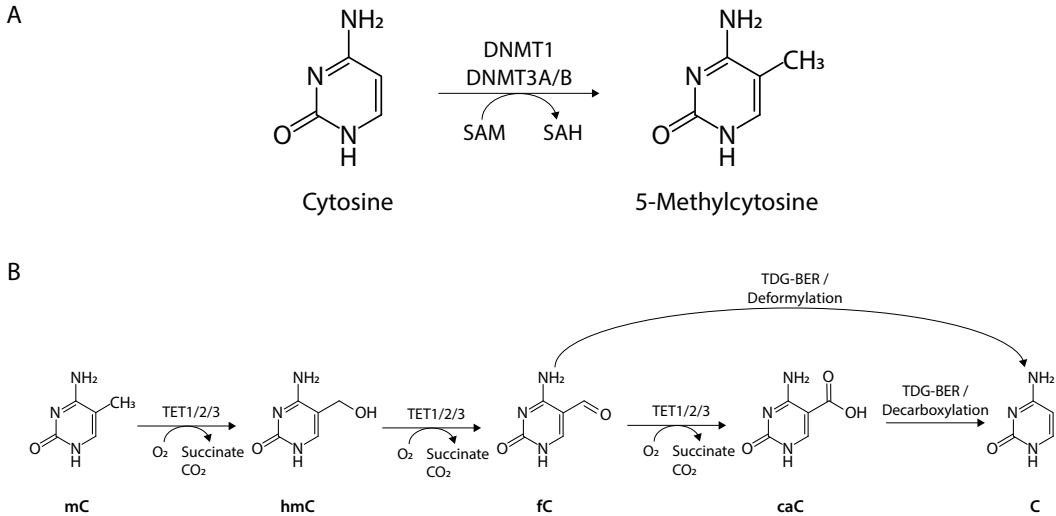
## DNA methylation process

The modification process of methylating CpG dinucleotides is enabled by DNA methyltransferases (DNMT) (**Figure 2A**) (42,46,47). This enzyme catalyzes the transfer of a methyl group from its substrate, S-adenosyl methionine (SAM) to C, resulting in 5-methylcytosine (mC) and S-adenosyl-homocysteine (SAH) (42,46–48). DNMT1 is responsible for maintenance methylation: it methylates the newly formed DNA strand after replication, if the original strand was methylated (15,42,49). *De novo* methylation is established by the DNMT3 family (DNMT3A and DNMT3B), and this process is fulfilled independently from DNA replication (14,42).

The process of demethylation is more complex. Loss of methyl groups from the CpG dinucleotides can happen actively or passively.

The passive demethylation process is caused by a lack of DNMT1 maintenance methylation after DNA replication, diluting the methylation pattern on the newly synthesized DNA strand (47). **Figure 2B** shows the process of active demethylation, which is initiated by ten-eleven translocation (TET) proteins (TET1, TET2 and TET3) (46,47,50). These enzymes oxidize mC in the presence of water, oxygen and  $\alpha$ -ketoglutarate resulting first in hydroxymethylcytosine (hmC), and subsequently in formylcytosine (fC) and carboxylcytosine (caC) with the production of CO<sub>2</sub> and succinate as result. TET1 interacts with thymine DNA glycosylase (TDG), an enzyme that can excise fC or caC to form an abasic site, and replace it by an unmodified C: TDG-dependent base excision repair (TDG-BER) (46,47,50). Other active demethylation processes include the direct conversion of fC and potentially caC by deformylation and decarboxylation respectively, or deamination of mC to thymine (T) by activation-induced cytosine deaminase (AID) and apolipoprotein B mRNA editing enzyme, catalytic polypeptide-like (APOBEC) enzymes, resulting in T:G mismatches that are corrected by BER mechanisms (47). However, these alternative mechanisms are rather exceptional.

These oxidized mC forms are not only intermediate forms of the demethylation process, but they also serve as epigenetic modifications with their impact on gene transcription, showing that the cell-type specific TET regulation plays an important role in transcription control.



**Figure 2. DNA methylation and demethylation processes.** **A.** DNA methylation process: a methyl group is added at the C-5 position of cytosine, catalyzed by DNMT enzymes. These enzymes use S-adenosyl methionine (SAM) as substrate, which results in the formation of methylcytosine (mC) and S-adenosyl-homocysteine (SAH). **B.** The active demethylation process is initiated by ten-eleven translocation (TET) proteins. TET proteins catalyze oxidation of mC in the presence of  $O_2$ , water and  $\alpha$ -ketoglutarate, resulting in hydroxymethylcytosine (hmC), formylcytosine (fC) and carboxylcytosine (caC) consecutively. fC and caC can consequently be excised from the DNA by thymine DNA glycosylase (TDG), resulting in a replacement by an unmodified C: TDG-dependent base excision repair (TDG-BER).

## CpG suppression

DNA methylation in mammals occurs mainly at CpG dinucleotides. The amount of non-CpG methylation in most differentiated somatic cells, and especially in human peripheral blood mononuclear cells (PBMCs), is negligible (<0.02%) (51–54). However, in specific cell types such as embryonic stem cells, (induced) pluripotent stem cells, non-dividing cells (neurons), glial cells, brain tissue, skeletal muscle cells or oocytes, non-CpG methylation occurs relatively frequent (51,54–60). Moreover, in specific disease contexts as Parkinson's, Alzheimer, Rett syndrome or amyotrophic lateral sclerosis, the amount of non-CpG methylation is also significantly increased.

In general, in vertebrate genomes, CpG suppression is observed (61–63). CpG suppression means that the amount of CpG dinucleotides throughout the genome is much smaller than expected: CpG only represents less than 1% of the dinucleotides in the human genome, while this genome has a CG content of about 42%. This implies that, statistically, about 4.41% ( $0.21 \times 0.21 = 0.0441$ ) of the

genome should be a CpG dinucleotide. This CpG suppression is a result of several factors (61): (i) C and guanine (G) have higher stacking energy than adenine (A) and T, so increased C and G bases in a DNA strand, result in more structural strain in the double DNA helix (64). (ii) The human innate immune response is activated by unmethylated CpGs, so increased amount of CpGs could lead to an increased risk of auto-immunity (65–68). (iii) Accumulation of C to T mutations due to CpG-specific DNMTs that accumulate mC, which might spontaneously deaminate into T (69–71). The subsequent DNA repair mechanisms recognize U (deaminated C) more easily compared to T, resulting in insufficient repair of T and consequently to an increased mutation rate (CpG to TpG shift).

Interestingly, the density of CpGs is significantly higher in gene-rich regions compared to gene-poor regions, implicating the importance of DNA methylation in the regulation of gene transcription in evolution (72). However, in some regions we observe a clear increase in the concentration of CpG dinucleotides: CpG islands (CpGIs) (61,73). The formal definition of a CpGI is a genomic region of at least 200 bp that has a GC content of at least 50% and an observed/expected CpG ratio of at least 60% (61,74). The observed CpG content over the human genome is 1%, and the expected CpG content is 4.41% (as described above), resulting in an observed/expected ratio of 23% ( $1\% / 4.41\% = 22.68\%$ ) for the complete human genome. Therefore, a CpGI in the human genome needs to contain at least 2.65% CpG dinucleotides ( $2.65\% / 4.41\% = 60\%$ ) over at least 200 bp. Only 1% of the human genome fulfills these criteria, due to the CpG suppression. The majority of these regions is found in and around promoter regions or the first exon of genes (75–77). Moreover, about 70% of the human promoters are associated with a CpGI, indicating the evolutionary importance of the transcriptional regulation by DNA methylation at promoter regions (61). Overall, the level of methylation is considered to be inversely correlated with CpG density: these promoter-associated CpGIs tend to be often hypomethylated, in contrast with the general degree of methylation, which is about 80% of all CpGs (78–81). Still, in mammalian cells, DNA methylation mainly occurs at CpG dinucleotides, making the CpGIs often hotspots for DNA methylation.

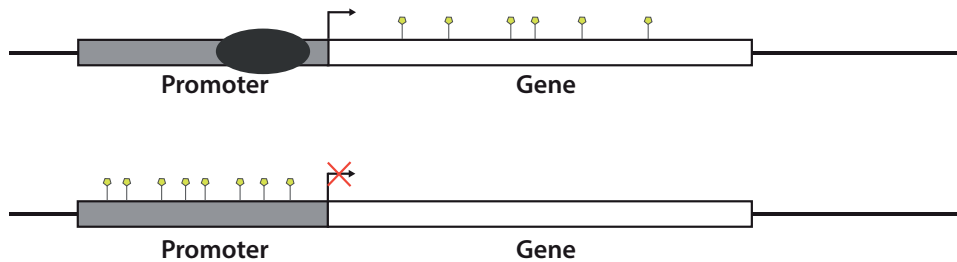
## Transcriptional regulation by DNA methylation

CpGI methylation status follows a bimodal distribution, where most of islands are either hypo- or hyper-methylated (81). In promoter regions, the effect of DNA methylation in CpGIs is well described: low DNA methylation (hypomethylation) is linked with active gene transcription (**Figure 3**). Hypomethylated CpGIs are often located in or near the promoter region of housekeeping genes. High DNA methylation densities (hypermethylation) of CpGIs in or near promoter regions is linked to stable inactivation of the transcription of the affected gene. This methylation (and transcriptional regulation) is highly tissue- and cell type dependent (13,42,61,63,82–86).

This transcriptional regulation is a result of a combination of several mechanisms inhibiting or enabling the binding of a TF to the promoter region of the affected gene. These TF carry biological inputs, that are converted to physiological responses (changes in gene expression) by mediator complexes. These complexes relay signals of the TF directly to the RNA polymerase enzymes to

alter the transcriptional activity (87,88). A first mechanism to inhibit TF binding is by the methylation itself: some TFs can recognize unmethylated CG rich motifs, but they cannot bind the methylated counterpart (61,89). However, most transcriptional regulation happens indirectly, as a result of changing chromatin conformation (such as density or histone modifications) due to the specific DNA methylation patterns (34,90). In a second mechanism, DNA hypermethylation is recognized and bound by methyl-CpG-binding domain (MBD) proteins, that, in turn, can recruit the nucleosome remodeling and deacetylase complex (NuRD). This complex contains histone deacetylase (HDAC), which cause a transcriptional repressing signal (deacetylated histone tails) (61,91–97). When this NuRD complex is recruited to methylated DNA, heterochromatin (dense chromatin) is formed, preventing the recruitment and binding of transcription complexes that can activate RNA polymerase II (78). This mechanism has been thought to be effective in approximately half of all genes (72,78).

- ◆ Methyl group
- Transcription factor



**Figure 3. Transcriptional regulation of by DNA methylation.** Hypomethylation of the promoter is linked to euchromatin, which allows transcription factors to bind. Hypermethylation of the promoter causes heterochromatin, and inhibits transcription factor binding. Gene body methylation shows an opposite relation to transcription activity of the gene.

The function of DNA methylation in non-promoter regions is less explored compared to promoter methylation. Consequently, the effects of methylation at intragenic CpGs (CpGs within gene bodies, both within as outside CpGIs) and intergenic CpGs (regions between genes) on transcription are less clear (98–102). The latter is shown to be involved in gene transcription regulation, especially when TF binding regions are affected (103). Intragenic DNA methylation has been shown to be involved in regulation of intragenic promoters, alternative splicing, cellular differentiation, activation of retroviruses or repetitive elements, prevention of aberrant transcript production. However, the exact mechanisms are still to be determined (78,98–102). A first hypothesis states that affecting the nucleosome binding places is a main mechanism: different histone variants with different histone tail modifications have different affinities for specific sequence motifs, including methylated DNA. This results in altered effects depending on the histone variant: some variants enhance transcription, others repress it (78,104). Other hypotheses for these transcription regulating mechanisms are that intragenic DNA methylation modifies the transcriptional efficiency of RNA polymerase II and

the transcriptional complex (78,105), or that methylation along the sense strand could suppress expression of the antisense strand mRNA (as full-length non-coding RNA or as microRNAs), which is complementary to sense RNA. A higher concentration of these antisense copies would result in sense-antisense mRNA complexes, reducing the amount of free sense RNA, and thus reducing the protein translation (78,106).

It is suggested that intragenic methylation – in particular in exonic regions – has the opposite effect of promoter DNA methylation: higher intragenic methylation is linked to increased gene expression (**Figure 3**) (78,107). Moreover, by studying the transcriptome and the DNA methylome simultaneously in single neurons, it is proposed that, in genes where a CpG is found in the promoter region, intragenic methylation is an even better indicator of transcription than the well-described and studied promoter methylation (107). In genes without CpG in the promoter region however, promoter methylation would still be the best indicator. In this study, single cells were analyzed, which has as advantage that the methylation and transcription signal could be compared directly. This enabled the researchers to demonstrate that genes with constitutively hypomethylated promoters had only active transcription in about 50% cells.

Transposable elements (TE) affect gene structure, expression patterns, and chromosome organization and may have deleterious consequences when released (108). DNA methylation in TE is an ancient mechanism which induces heterochromatin to stably silence these TE to prevent them to affect the genomic integrity (42,108,109).

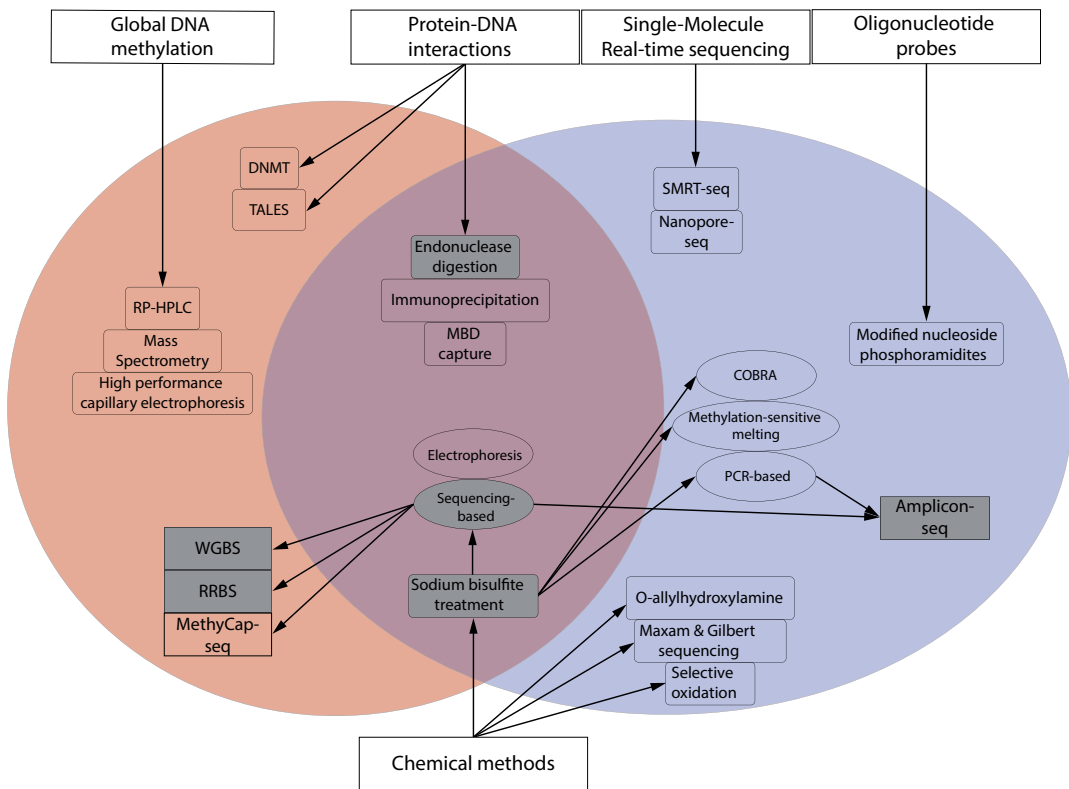
## DNA methylation analysis methods

Since the discovery of mC, a multiplicity of different analysis techniques has been developed. These methods are evolving continuously to become more sensitive and accurate, cheaper and with higher resolution (42,47,110,111). Several innovative methods are described, and some can measure methylation up to single base pair resolution, while others have a rather low resolution, but can easily estimate the global or genome-wide methylation density (110,112). Innovations are both in the techniques used for mC labeling, targeting or chemically altering, and in the detection methods (e.g. quantitative polymerase chain reaction (qPCR), optical and colorimetric readout, surface plasmon resonance, electrochemiluminescence, or electrochemical readout) (47,110,113). Assays can also be label-free, relying on the differences in the physicochemical properties of C and mC.

Most methods are associated with fragmentation of the DNA, resulting in downstream analysis difficulties, or lack of target specific genomic information. Moreover, the majority of methods use enrichment strategies, and therefore provide only useful information about the methylation hotspots, and not about the intragenic DNA methylation. In most cases, the amount of sample needed is relatively high and since methylation is lost after amplification, a lot of these analyses are limited to *in vitro* or animal models. This, combined with the need for special lab equipment for some of these techniques, results in difficulties of clinical translation of DNA methylation analysis with these methods (110). Moreover, most of the techniques do not discriminate between mC, hmC, fC or caC.

However, as reviewed by Booth et al. and Berney et al., most of these techniques can be slightly adapted to include analysis of these other modifications (47,113).

Some of the methods to analyze DNA methylation provide genome-wide methylation information, while others mainly provide location specific methylation information (**Figure 4**). **Global DNA methylation assessment** provides an estimate of DNA methylation in a sample, without sequence information. Some proteins or oligonucleotides have methylation-dependent action or binding properties, enabling DNA methylation methods using **protein-DNA interactions** and **oligonucleotide probes**. **Single-molecule real-time sequencing** measures directly the differences between C and mC, providing direct sequencing information. **Chemical methods** chemically alter the DNA on a methylation-specific manner. Finally, several **single cell methods** have been described, using similar techniques at a single cell resolution.



**Figure 4. DNA methylation analysis methods.** Methods can be used to analyze genome-wide DNA methylation (left, red ellipse), or to analyze targeted DNA methylation of a genomic region of interest (right, blue ellipse). Methods in the intersection are being used for both, depending on the detection method. Methods that have been shown to be adaptable to single cell analysis have a grey background. RP-HPLC: Reversed-phase high-performance liquid chromatography, DNMT: DNA methyl transferase, TALES: transcription activator-like effectors, MBD: methyl-CpG-binding domain, SMRT: single-molecule real-time sequencing, COBRA: combined bisulfite conversion and restriction analysis, WGBS: whole genome bisulfite sequencing, RRBS: reduced representation bisulfite sequencing.

## *Global DNA methylation assessment*

Reversed-phase **high-performance liquid chromatography** (RP-HPLC) combined with UV detection was one of the earliest methods used to study DNA methylation in hydrolyzed samples (110,111). Other methods as **mass spectrometry**-based methods, or **high-performance capillary electrophoresis** are also used to measure global DNA methylation levels. All these methods do require specialized lab equipment and most of them require high amounts of DNA (e.g. RP-HPLC requires 3-10 µg DNA) (47,110). Combinations and improvements of these methods (e.g. liquid chromatography coupled with mass spectrometric detection) are often more sensitive, but still expensive and technical challenging (47).

Therefore, other cheaper and more sensitive methods, preferentially providing sequence information, have higher resolution and that only require standard lab equipment, are desirable.

## *DNA methylation assessment using protein-DNA interactions*

Several proteins have different properties towards mC compared to C, enabling the use of these proteins in methylation analysis strategies.

**Endonuclease digestion-based** methods use nucleases with methylation-dependent site-specific restriction activity (47,110,113–115). In these reactions, restriction of unmethylated target sequences, but no restriction of methylated DNA is performed. Gel electrophoresis profiles can be analyzed to obtain a general idea of the methylation (113). By combining a methylation dependent with a methylation independent endonuclease carrying the same restriction site specificity, this method can be used to measure methylation quantitatively. Therefore, a qPCR can be performed targeting this specific restriction site to compare the amount of cleaved sequences. This strategy is mainly used for global or genome-wide DNA methylation assessment using qPCR, microarray or sequencing (113,116).

**Immunoprecipitation-based** methods recognizing methylated DNA can be used to measure DNA methylation density or enrich methylated DNA strands (47,72,110,113,116). Therefore, antibodies targeting mC are commercially available, and can be used to overcome the restriction-site bias related to the restriction-based methods. Methylated DNA immunoprecipitation (MeDIP) is a chromatin immunoprecipitation (ChIP)-derived assay where antibodies bind single stranded fragments of methylated DNA sequences to retain those and wash the not-methylated sequences away. Methylated sequences can subsequently be (q)PCR amplified and/or sequenced. Sequencing reads are mapped to the complete genome, enabling genome-wide methylation profiling (113). Quantification of specific methylated DNA strands by qPCR enables gene-specific DNA methylation levels, but this method will not provide absolute quantitative data (47,110). The resolution of the MeDIP depends on the original fragment length (47,110,113).



**Methyl-CpG-binding domain (MBD) capture** is a similar method to enrich for methylated DNA strands: MBD-polypeptides (e.g. coated on magnetic beads) can be used to retain methylated DNA from a genomic DNA sample (47,110). The downstream analysis possibilities to measure the degree of DNA methylation are similar in MeDIP and MBD (conventional methods as (q)PCR or sequencing). MBD peptides can also be fused to a green fluorescent protein to enable direct detection, which can even be made sequence specific by fusing an additional zinc finger protein (47). An important factor in the use of antibodies to enrich for methylated DNA is the possibility of varying qualities of antibodies. Low specificity or cross-reactivity with off-target sites might result in substantial background noise (113).

**Transcription activator-like effectors (TALEs)** are a family of bacterial proteins that have a DNA-binding domain with programmable sequence selectivity (117–119). These proteins can be modified to recognize not only the four canonical nucleobases, but also modified C nucleotides as mC (47,117).

The methylation activity of some **DNMTs** are used to tag unmethylated CpGs, which can be linked with biotin to retain only the unmethylated sequences. These sequences can subsequently be detected and quantified using PCR amplification (47,120).

These protein-based methods do meet most of the abovementioned criteria (cheaper, higher resolution, standard lab equipment suffices and possibilities of sequence information). However, they are never absolutely quantitative, and they never reach single nucleotide resolution.

### *Single-molecule real-time sequencing*

In contrary to high throughput next-generation sequencing (NGS) approaches (e.g. Illumina sequencing methods), single-molecule real-time sequencing (also called third-generation sequencing) approaches do not measure cell population DNA sequences, averaging the read-out from this population, but rather measure every single DNA strand directly. By sequencing every oligonucleotide separately, absolute quantification of the signal is possible. Moreover, no amplification of the DNA is needed, so epigenetic modifications are not removed from the sequences. Consequently, these methods can theoretically measure the DNA modifications such as mC. An additional advantage of these methods is that they can sequence very long reads (up to 100 kb) (121).

**Single-molecule real-time sequencing (SMRT-seq)** uses zero-mode waveguides (ZMWs) which are illuminated. Polymerases will incorporate complementary phospholinked nucleotides containing a base-specific fluorophore in the genome (122). By measuring the light pulse obtained after every incorporation of a nucleotide, the genome is sequenced. This method is capable of measuring the pausing of the polymerase due to chemical tags on the genome. These kinetics enable the detection of modified bases as mC or hmC (113,123).

**Nanopore-sequencing** is another long-read method that uses solid state nanopores. These are proteins with a nanometer-scaled pore which enables continuous passage of ions, and single stranded DNA (ssDNA) strands, if an electric potential is applied. The sequencing of DNA is performed by monitoring the ionic current over this pore: alterations of this current during DNA passage is depending on the structure of the nucleotide. Modified nucleotides will cause a slightly different current change compared to the unmodified nucleotide (47,121,124,125).

These methods are relatively new and are under continuous improvement. The error rate for these techniques is relatively high, however, for both techniques, the (analysis) software is improving rapidly, enabling even more accuracy and sensitivity. Moreover, new nanopores are being developed at a high rate with increased throughput, sensitivity, specificity and accuracy.

### *Oligonucleotide probes*

Oligonucleotide probes can be modified with a multitude of nucleotide-variants (**modified nucleoside phosphoramidites**) that bind much more efficient to mC compared to C (47). Main advantages include that these probes are cost-effective and that the probe selectivity is easy to modify by altering the probe sequence. Different strategies for the mC detection have been developed using alternative modifications for the probes: based on photochemical reactions using DNA-templated photoligations, on fluorescence using fluorophore-oligonucleotides, on electrochemical detection methods or by selective oxidation of mC. The use of a quartz crystal microbalance is another way to detect mC with immobilized probes targeting DNA after endonuclease treatment to eliminate all unmethylated regions. Furthermore, approaches based on non-covalent binding of oligonucleotides are described (47).

### *Chemical methods*

The basic principle of chemical DNA methylation detection methods is to alter specific nucleotides at a different rate than their modified counterpart. These changed nucleotides can subsequently be detected in order to measure the amount of DNA methylation. Sequencing of these altered DNA strands gives high resolution DNA methylation profiles, and chemical agents are less prone to low quality or aspecific action compared to antibody-based methods.

Standard **Maxam & Gilbert sequencing** uses four treatment reactions of 5' radiolabeled DNA, creating base-selective strand cleavage: DNA breaks at G/purines (A and G)/pyrimidines (C and T)/C, and is subsequently cleaved by treatment with hot piperidine. Sequencing is then performed by interpreting a gel electrophoresis ladder after these four treatments (47,113). This method can be used to detect mC since the treatment of DNA with hydrazine (to cleave at pyrimidine sites) is inefficient at mC, resulting in a gap on a ladder, which can be identified as mC with the detection of a G in the opposite strand. This method has further been optimized for the detection of mC: two different N-halogeno-N-sodiobenzenesulfonamide reagents show differential reactivity toward C and

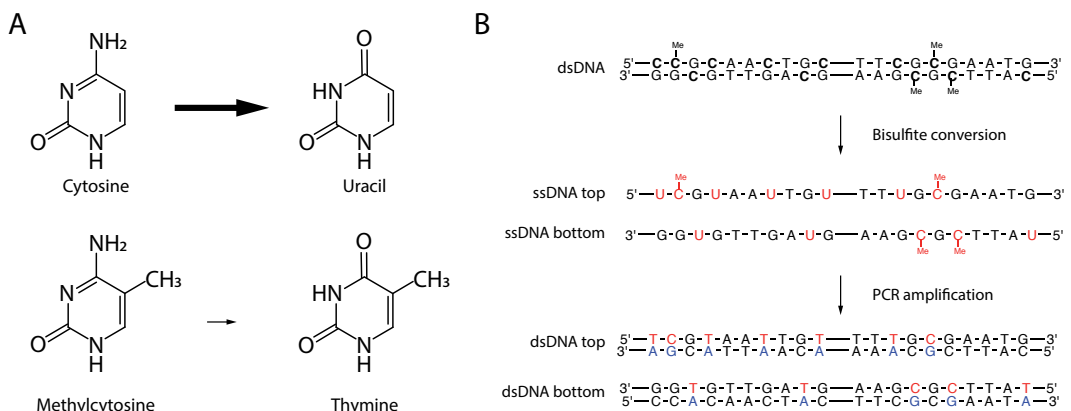
5mC resulting in either cleavage at C or cleavage at C and mC (47,113,126). These cleaved DNA strands can subsequently be analyzed by gel electrophoresis, and if wanted, sequenced to obtain robust, high resolution, sequence-specific DNA methylation levels.

**O-allylhydroxylamine** is a compound that reacts with C and mC, resulting in an E-isomeric and a Z-isomeric form of (m)C – O – allylhydroxylamine respectively. This difference is due to steric hindrance by the methyl group. The Z-isomer, will cause a polymerase block, while the E-isomer codes like a T, and will thus induce C-T transitions at unmodified Cs, which is easily detectable by sequencing (113,127).

**Selective oxidation methods** use several oxidizing agents ( $\text{OsO}_4$ ,  $\text{V}_2\text{O}_5$ ,  $\text{NaIO}_4 + \text{LiBr}$ ) to treat DNA, and obtain differential oxidation between C and mC. Additional DNA treatment with hot piperidine causes cleavage at the oxidized locations, which is detectable by gel electrophoresis (47).

## Bisulfite treatment

In 1992, Frommer et al. described sodium bisulfite treatment, which first enabled the analysis of DNA methylation at single base pair level (128). This method translates the epigenetic differences to genetic differences: by treating DNA with sodium-bisulfite ( $\text{NaHSO}_3$ ), C deaminates to uracil (U) (**Figure 5**) (129–132). mCs are deaminated to T at a 100-fold slower rate, making the technique suitable to discriminate methylated from unmethylated C if the bisulfite treatment was correctly balanced (128,133). This converted DNA can be amplified using conventional amplification techniques as PCR, where U will anneal with A (U and T molecules have the same binding behavior with A), resulting in a C to T conversion, while mC will be amplified as C (**Figure 5B**). Since the treated DNA strands are not complementary after treatment, the end result is two different ssDNA molecules (134). Sodium bisulfite method is still considered as the gold standard method for methylation analysis (48).



**Figure 5. Principle of DNA methylation analysis using sodium bisulfite treatment. A.** Cytosine(C) and methylcytosine (mC) deamination using sodium bisulfite: C deaminates to uracil (U), while mC deaminates to thymine, with a 100-fold lower rate. **B.** Mapping protocol: After PCR amplification and sequencing, mC will be read as C, while unmethylated C will be read as T, enabling differentiation of C and mC.

The methylation of a specific locus can be measured by (q)PCR of the region of interest, and the amplicons can be sequenced by conventional sequencing techniques. Bisulfite treated DNA can be used to analyze genome-wide methylation profiles (whole genome bisulfite sequencing (WGBS)), location specific methylation, and even with single bp resolution. In either case, the treated DNA needs some downstream analysis (110).

**PCR based analysis** methods encompass PCR amplification of bisulfite treated DNA. Methylation specific PCR (MSP) uses primers targeting the CpG of interest (116,135). The amplification of bisulfite treated DNA is performed using two primer sets and depends on the presence of methylation in the original DNA strand: primers targeting TpG will amplify unmethylated targets, while CpG-targeting primers are used to amplify methylated targets. This can be qualitatively analyzed using gel electrophoresis, and a quantitative alternative for this assay is MethyLight, making use of qPCR instead of normal PCR (116,136). Technically, other PCR methods as digital PCR can also be used to quantify DNA methylation after bisulfite treatment (12).

**Combined bisulfite conversion and restriction analysis (COBRA)** can be used to analyze the methylation after bisulfite treatment. Bisulfite treatment can cause the DNA at a specific location to either lose or create a specific restriction site of a restriction enzyme. By restricting before and after treatment, differences, and thus DNA methylation state can be visualized using gel electrophoresis (116,137).

**Sequencing** (e.g. direct Sanger sequencing, cloning and sequencing, pyrosequencing, NGS) of the treated DNA gives single base pair information of every sequenced CpG (128,138–140). Sequencing can be used to analyze specific regions of interest (targeted amplicon sequencing). A classical targeted bisulfite sequencing assay can essentially be divided in three different steps: DNA treatment with bisulfite, enrichment of the DNA using PCR and sequencing of the obtained product. This sequencing can be performed with a variety of sequencing methods (such as Sanger sequencing or NGS). Sequencing based methods are also used to measure genome-wide DNA methylation (WGBS or BS-seq), where complete genomes are sequenced using state of the art NGS techniques (52,141–144). However, as an alternative to WGBS, CpG-rich regions are often enriched since WGBS usually only provides 1X coverage for 95% of the CpGs with the latest deep sequencing methods (145). This enrichment can be performed using restriction of DNA (reduced representation bisulfite sequencing (RRBS) (146)) or through affinity-enrichment using antibodies targeting mC or MBD-containing proteins (MethylCap-seq (147)). These methods ensure the researcher a more cost-efficient analysis of several regions of interest: RRBS uses restriction with the methylation-insensitive restriction enzyme MspI to restrict DNA at 5'-CCGG-3' motifs (146,148). This results in fragments containing CpG sites at both ends. After size selection, this procedure reduces the whole genome to more or less 1% of the original size, still containing the majority of the cellular promoter regions and repeat regions (however, only 10% of the total CpGs are covered, resulting in poor coverage of regions with low CpG density) (145,148).

**Methylation-sensitive melting** curve analysis or methylation-sensitive high-resolution melting is based on the differences in melting properties due to composition changes: high CG content will increase the melting temperature (149,150).

**Electrophoresis-based analysis** uses differences in base composition of every ssDNA strand, which leads to different conformation formations. These conformations all have variable migration in electrophoresis. In single strand conformation analysis, CpG sites are amplified, made ssDNA and electrophoresed to detect methylation based on these migration patterns (151).

Analysis using sodium bisulfite has as advantage that no specialized lab equipment is needed, it is relatively cheap (except for potential sequencing), and it provides reliable single base-pair resolution data. However, treatment with sodium bisulfite also has some important drawbacks (12,110). (i) It needs to be performed in harsh conditions (e.g. low pH and high temperatures), resulting in high degree of fragmentation (typically fragments are <500 bp) and DNA loss (up to 95-99%) (152). This hinders the analysis of longer fragments and makes high amounts of DNA input necessary. Therefore, bisulfite-based analysis is often unsuitable for clinical applications (152–154). (ii) The treated DNA mainly consists of three bases instead of four (heavily depleted in C). This hampers primer design for following PCR, and it impedes subsequent mapping of sequencing reads, especially if the starting DNA is very variable (as in the HIV-1 genome). (iii) The variability of starting DNA can also lead to biased PCR, where one dominant PCR product can be formed. (iv) C conversion can be incomplete (often as a result of hairpin structure formation during conversion), resulting in an overestimation of the methylation. Inappropriate conversion on the other hand can lead to underestimation of methylation. If studying differentiated mammal cells such as PBMCs, non-CpG conversion is a very decent measure of the total conversion efficiency, since the non-CpG methylation is showed to be < 0.02%, meaning that all these Cs should be converted (51–54). Inappropriate conversion (conversion of mC to T) is more difficult to measure, but it can be estimated using analysis of DNA sequences with known methylation profile. (v) This method does not discriminate between 5mC or 5mC oxidation products like hmC, fC and caC (110,155).

### *Single cell DNA methylation methods*

In line with the increasing possibilities in single cell (multi-)omics' analyses, increased efforts are made to develop single cell epigenomic assays (156). Developing methods which create possibilities to measure single cell DNA methylation reliably, and are cost-effective and easy to perform, will open new research perspectives. This would open the research towards two important questions about DNA methylation (145). First, are the current models of DNA methylation regulation, that are predominantly built on analysis of bulk DNA methylation from heterogeneous populations, true on single cell level? Secondly, how does DNA methylation within a single cell exactly steer the biology of the cell? By using single cell technologies, it is indeed proposed that in neuron cells, DNA methylation in promoter regions is a good predictor of transcription, but that gene body methylation is actually a better predictor if a CpGI is present in the gene (107). Moreover, single cell methylation methods

have as important advantage that they can be used to analyze cells with limited availability, that could not be studied using bulk methods (145). This will lead to exciting new insights, both in fundamental research as in clinical research.

**Bisulfite** based methods need in general a high amount of input DNA, due to high DNA degradation. However, some bisulfite based single cell DNA methylation assays have been described. First, single cell RRBS (scRRBS), and subsequently the more sensitive and less biased quantitative RRBS (qRRBS) were developed, resulting in a coverage of 40% of the bulk RRBS counterpart (145,157–159). To enhance the coverage, several additions to the bulk methods have been described: single cell bisulfite sequencing (scBS-seq) and single cell WGBS (scWGBS) use post-bisulfite adaptor tagging (PBAT) before performing deep sequencing of the bisulfite treated DNA fragments. This results in high throughput, but relatively low coverage DNA methylation information of the complete genome. Single cell PBAT (scPBAT) is a whole genome shotgun bisulfite sequencing method, capable of sequencing mainly repetitive regions (160–165).

With the potential of single cell bisulfite-based methods demonstrated, several adaptations are being developed to increase coverage and throughput of the methods (145). Single-nucleus mC sequencing (snmC-seq) optimizes recovery after bisulfite treatment, while single-cell combinatorial indexing for methylation analysis (sci-MET) enables multiplexing of several cells in scWGBS (or scRRBS). Moreover, microfluidic approaches are being developed (e.g. microfluidic diffusion-based RRBS, MID-RRBS) (166–168).

The abovementioned methods are all genome-wide approaches, but to study gene- or locus-specific DNA methylation patterns, single-cell locus-specific bisulfite sequencing (SLBS) has been developed (169). This method uses whole genome amplification using a multiple displacement amplification (MDA) method after bisulfite treatment, followed by conversion specific nested PCR and sequencing.




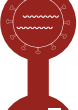











Bisulfite treatment degrades most of the DNA, and potential problems related to the conversion efficiency can occur as described earlier. Consequently, a variety of single cell bisulfite-free methods is described. **Endonuclease digestion** with methylation-dependent restriction enzymes, coupled to (multiplexed) (q)PCR amplification in one reaction, enables high throughput single cell analysis. However, the sensitivity of these methods is rather low. Consequently, genome-wide single cell CpG methylation sequencing (scCGI-seq) has been developed using MDA after the restriction step, to selectively amplify and deep sequence CpG-containing sequences. Endonuclease digestion-based methods have lower resolution than bisulfite-based methods.

An important advantage of single cell analysis is that specific methylation patterns can directly be linked to specific phenotypes. Moreover, the ultimate method to get insight into the effects of DNA methylation would be able to push the technology to the next level and link methylation profiles to other epigenetic modifications and/or RNA transcription of the analyzed genes (multi-omic approaches). Ideally, these methods are also capable of allele-specific DNA methylation analysis to link epi-haplotypes (allele-specific epigenetic state) to the transcription of the allele.

Several single cell techniques have been developed for single cell epigenomic and transcriptomic analysis, other than DNA methylation (sequencing-based and imaging based) (156,170). Moreover, methods have been described that enable simultaneous analysis of the methylome with transcriptomic information (107,171), nucleosome occupancy (172), or even multiple information layers including genomic copy number variations, genotyping, chromatin state, nucleosome positioning, etc. (173–175).





- ± 1900**  West-Central Africa: first infection from Chimpanzee
- 1981**  USA: Kaposi's sarcomas and opportunistic infections  
Disease is called GRID
- 1982**  GRID becomes AIDS  
First suggestion of a viral infection
- 1983**  Virus isolated: LAV and HTLV-III
- 1984**  Heterosexual transmission of the AIDS causing virus  
Additional virus isolated: ARV
- 1985**  LAV, HTLV-III and ARV are same virus, causing AIDS  
First commercial blood HIV test: ELISA
- 1986**  AIDS causing virus is officially named HIV
- 1987**  DNA methylation as a modulator of HIV expression  
First antiretroviral drug AZT is approved
- 1995**  Triple therapy is introduced
- 1996**  Foundation of UNAIDS
- 1997**  Triple therapy becomes standard in HIV care
- 2008**  Berlin patient is cured
- 2012**  PrEP is FDA approved
- 2014**  90-90-90 goals
- 2019**  London patient is cured

## II – HIV/AIDS

### Historical setting

The first cases of human immunodeficiency virus (HIV) infected individuals are traced back to the late 1800s - early 1900s in West-Central Africa (176,177). The most commonly accepted theory is that the virus came into humans through zoonotic transmission from several non-human primates. In these primates, several strains of simian immunodeficiency virus (SIV) are found, and different independent events have occurred where the species barrier is crossed from primate to human via blood-to-blood contact during hunting and butchering primates, or by keeping them captive or as a pet (176–179). At least seven events of independent zoonotic transfers from chimpanzees (SIV<sub>cpz</sub>), gorillas (SIV<sub>gor</sub>) and sooty mangabeys (SIV<sub>sm</sub>) have led to different HIV subtypes and groups with following relations discovered by phylogenetic analysis: HIV-1 group M and N from SIV<sub>cpz</sub>; HIV-1 group O and P from SIV<sub>gor</sub>; HIV-2 group A-H from SIV<sub>sm</sub> (176–182). The current global epidemic of HIV-1 is mainly caused by the HIV-1 group M: more than 95% of HIV infections worldwide are from this group. This virus strain originates from SIV<sub>cpz</sub> and has started in West-Central Africa, and spread from the border of Cameroon and the Republic of Congo to the cities Kinshasa and Brazzaville (176,180,183). Here, urban development, increasing population levels and densities, combined with international connections, facilitated the fast spread of this deadly HIV virus through sexual interactions. Moreover, international vaccination programs using unsterile needles boosted the spread of this virus exponentially in these cities (184,185). This cocktail paved the way for the virus to travel outside Africa and these first infections have most probably happened in Haiti: healthcare workers which have been infected in Africa during humanitarian missions, brought the virus back to Haiti. There it could further spread into the American population through blood donor centers installed on the island collecting blood for USA-based clinics. Subsequently, the virus spread all over the world establishing the current pandemic (176).

In the early 1980's, multiple cases of rare deaths were reported in large cities in the United States (New York, San Francisco, Miami, Los Angeles) (186–189). Physicians were confronted with several homosexual men that were dying from aggressive Kaposi's sarcomas in skin, lymph nodes, mucosa and viscera and opportunistic infections such as mucosal candidiasis, *Pneumocystis carinii* and cytomegalovirus. These diseases are all caused by otherwise self-limiting infections, unless patients had a comprised immunity (186,187). Indeed, these patients had a clear imbalance of peripheral-blood T-cell subsets and especially the CD4+ T cells were depleted. Hence, it was immediately suggested that these patients were suffering from a cell-mediated immune suppression and it quickly became clear that the hospitals had to deal with a deadly epidemic with no treatment options other than a symptomatic relief (186).

Without knowing the cause of this disease, the risk groups were identified: the 4 H's (hemophiliacs, heroin addicts, homosexuals and Haitians) (190). The U.S. Center for Disease Control suggested that the transmission of the putative microbial cause of the disease happened via blood, sex and

breast feeding, leading to the suggestion of a retroviral infection as a cause of the disease by Robert Gallo (190). This suggestion was mainly based on the similarity of the symptoms to two other recently discovered retroviruses (Human T-cell Leukemia Virus (HTLV)-I and -II).

This disease was first named gay-related Immune deficiency (GRID) but in 1982, the disease was officially defined as acquired immunodeficiency syndrome (AIDS) and described as ‘a disease at least moderately predictive of a defect in cell-mediated immunity, occurring in a person with no known case for diminished resistance to that disease’. Two years after the initial AIDS reports, the causal agent of this strange disease was identified in three independent laboratories: Montagnier and colleagues from the Pasteur Institute (Paris, France) called the retrovirus isolated from patients suffering from lymphadenopathy ‘lymphadenopathy-associated virus’ (LAV) (191), but Gallo and his colleagues in the National Cancer Institute (Bethesda, Maryland), used tissue sent to them by Montagnier, and called the isolated virus ‘Human T-cell Leukemia Virus-III’ (HTLV-III) (192,193). The third group, led by Levy, simply called it AIDS-associated retrovirus (ARV) (194). It would take two more years until the researchers proved that these LAV, HTLV-III and ARV were the same virus (1.5% sequence difference between LAV and HTLV-III; 6% sequence difference between ARV and the two others) (195,196). This resulted in the revision of the definition of AIDS to note that the cause of this defect in cell-mediated immunity is this newly identified virus.

In 1984, two Belgian publications showed that there was a high AIDS prevalence in Africa (Rwanda and in the Democratic Republic of the Congo) and these studies demonstrated that the virus could also be transmitted by heterosexual contact, and thus that women could become infected too (197,198). One year later, the dynamics of this viral infection were clearly described, showing two latency phases during disease progression (188). A first latency phase between initial infection and antibody development and a second latency phase between antibody development and disease (AIDS) outcome.

In 1986, the name ‘Human Immunodeficiency Virus’ (HIV) was proposed by Coffin and colleagues, which became the only official name for the virus causing AIDS (195,199). The same year, the first commercial blood test for HIV was licensed by the Food and Drug Administration (FDA), resulting in HIV screening in US blood banks. In 1987, the first anti-retroviral drug azidothymidine (AZT) was approved by the FDA to treat HIV in the USA. Due to resistance problems of this single drugs therapy strategy, the first triple combination therapy (combination antiretroviral therapy (cART)) as treatment was introduced in 1995, together with the FDA approval of the first protease inhibitor as antiretroviral drug. This cART regimen became standard in HIV care two years later.

In 1996, a joint United Nations Programme on HIV and AIDS (UNAIDS) was founded. Its goal was to organize and coordinate the global HIV/AIDS research to end the pandemic. Up to today, it provides the strategic direction, advocacy, coordination and technical support needed to catalyze and connect leadership from governments, the private sector and communities to deliver life-saving HIV services.

In 2008, the first patient ever was declared cured from HIV infection: the Berlin patient, an HIV-infected patient, was diagnosed with acute myeloid leukemia, for which he received a bone marrow

transplantation from a donor with a CCR5 $\Delta$ 32 mutation, causing resistant cells for most HIV infections. The same procedure was successfully repeated in a second patient in 2019: the London patient.

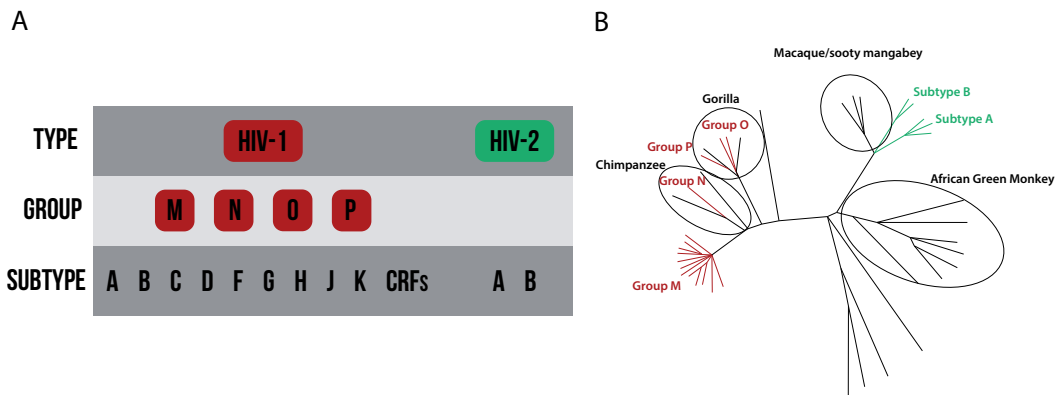
In 2012, pre-exposure prophylaxis (PrEP) was approved to protect individuals with substantial risk for contracting HIV and two years later, in 2014, UNAIDS developed the 90-90-90 targets, which aimed for 90% of people living with HIV to be diagnosed, 90% of those diagnosed to be accessing medical treatment, and 90% of those accessing treatment to achieve viral suppression by 2020. In 2015, the world health organization (WHO), recommended HIV treatment initiation at time of diagnosis.

In a recent report, the UNAIDS documented a decrease in AIDS related deaths over the last ten years, with in 2018, less than 1 million people died (770 000, from which 100 000 children (<15 years)). In addition, it is estimated that about ¾ of the people living with HIV knew their status, and 21.7 million people were on treatment in 2017 (+2.3 million compared to end of 2016) (200,201). The deaths of HIV patients declined faster than new HIV infections, resulting in an increase of amount of infected individuals: at the end of 2018, it was estimated that 37.9 million people were infected with the virus (+1 million compared to 2017), from which 18.8 million were women, and 1.7 million are children (<15 years). Of those infected patients, 1.7 million people became infected in 2018, from which 100 000 were children (< 15 years) (200,201).

Today, much more is known of this virus and the consequences of an infection with HIV. The medical world has made it possible to control the virus on long-term, and, especially in western countries such as the USA, Europe, Canada, made the life expectancy of HIV infected patients similar to non-infected patients (202,203). However, there is still no effective vaccination or cure available to end patients' burden of living the rest of their lives with HIV. Hence, patients are still condemned to lifelong therapy (cART, usually triple therapy) in order to suppress HIV and prevent viral rebound. This viral rebound happens very quickly after treatment interruption due to the reversal of a latent reservoir that is formed right after infection with HIV (204–209). This latent reservoir forms the main hurdle towards a cure, and insight in the mechanisms of latency initiation, maintenance and reversal is needed to find an effective HIV cure.

## HIV classification

The HIV virus is member of the order of the Ortervirales (210,211). ‘Orter’, being an inversion of ‘retro’, is derived from reverse transcription, indicating the need of reverse transcriptase to complement RNA to complementary DNA (cDNA) during the life cycle. The Ortervirales order encloses five families of reverse-transcribing viruses: *Belpaoviridae*, *Caulimoviridae*, *Metaviridae*, *Pseudoviridae* and *Retroviridae*, with HIV belonging to the latter (210–212). This family consists of enveloped viruses with an RNA genome, infecting vertebrate hosts and integrating their reverse transcribed genome stably into the genome of the host (212,213). Retroviruses originate about 460-550 million years ago, alongside their vertebrate hosts, infecting all kinds of vertebrates from the beginning of their existence, leading to the huge variety of endogenous retroviral elements found in all vertebrate genomes (212,214).



**Figure 6. HIV classification.** **A.** Overview of the HIV virus types, groups and subtypes. **B.** phylogenetic tree displaying the intermingling of SIV and HIV groups. Red arms refer to HIV-1 groups, green arms refer to HIV-2 subtypes. HIV-2 subtypes C-H are not included since they were only found in single individuals. Based on (215).

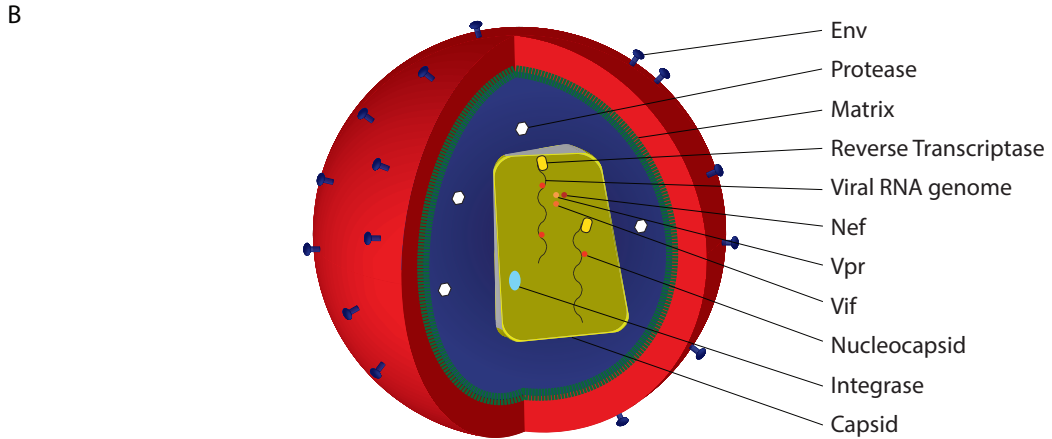
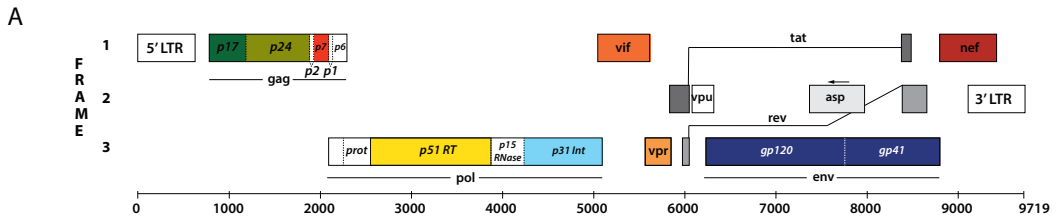
Retroviruses are divided in two subfamilies: *Orthoretrovirinae* and *Sumaretrovirinae* (210,211). The former, *Orthoretrovirinae*, consists of six different genera, from which the genus of the *Lentivirinae* is the one wherein HIV is classified (211). Lenti, Latin for slow, implicates that an infection with the viruses of this genus results in slow progressing inflammatory diseases (216). Two HIV species are described: HIV-1 and HIV-2 (**Figure 6**). Within these species, several groups can be found (HIV-1 groups M, N, O and P; HIV-2 groups A-B (and C-H which are only found in single individuals)), originating from different zoonotic transfers (176,179–182,217). HIV-1 group M is most prevalent: the current global epidemic of HIV-1 is mainly caused by this group, consequently, most research is performed studying HIV-1 group M (176). This is also the group that will be described in this thesis

## HIV-1 genomic structure

Due to the high variability, a multitude of HIV-1 variants are found, however, the HXB2 genome has been chosen as the reference genome in order to facilitate communication about genomic regions in HIV-1 (218). This genome has a length of 9719 bp, and it is flanked by two identical long terminal repeat (LTR) regions of 634 bp, formed during reverse transcription. These regions are segmented into U3, R and U5 regions, and multiple transcription regulating regions and TF binding sites are found (e.g. enhancer region, poly-adenylation signaling region, TCF1- $\alpha$  binding site, NF- $\kappa$ B binding site, SP1 binding site, TATAA box) (219). In addition, the 5'LTR contains the single promoter region and the transcription start site (bp 456) of the HIV-1 genome. This viral genome is slightly more complicated than the typical retroviral genome: the HIV-1 genome has nine genes: *gag*, *pol*, *vif*, *vpr*, *vpu*, *env*, *tat*, *rev* and *nef* (**Figure 7A**). By the use of different open reading frames (ORFs), alternative RNA splicing, cleavage of precursor polyproteins and ribosome frameshifting, it codes for 18 proteins (216,218,220,221). As for all retroviruses, the structural genes are *gag*, *pol* and *env*, coding for the essential structural proteins needed for formation of new infectious viral particles (222). *Gag* and *pol* genes are translated together into large Gag-Pol precursor polyproteins, that are cleaved by a viral encoded protease: *gag* codes for p55 (Assemblin), which is cleaved to p17 (Matrix), p24 (Capsid), p7 (Nucleocapsid), p6 and two spacer proteins p1 and p2 (223). *Pol* codes for protease, p51 (reverse transcriptase), p31 (integrase) and p15 (RNase H) (**Figure 7**) (224).

*Env* is translated as one full length env protein (gp160), which is transported through the trans-Golgi network to become glycosylated and cleaved by host proteases (furin) into gp120 (surface glycoprotein unit) and gp41 (transmembrane glycoprotein unit), which are consequently directed to the plasma membrane at the surface of the infected cell (216,224,225). At this point, the infection becomes visible for the immune system by expressing foreign proteins at the cell membrane.

The six other genes code for primary translational products: *tat* and *rev* are involved in the regulation of viral gene expression and are produced through alternative splicing. Tat is a transcriptional transactivator, essential for HIV replication by interacting with the TAR RNA element (located at the 5' end of all HIV-1 RNA strands, forming a transcription block with host cellular proteins) to dramatically increase the transcription rate (216,226). Rev binds to the rev responsive element (RRE), located in the *env* gene (bp 7710 to 8061) to facilitate export and promote stabilization and translation of unspliced and incompletely spliced viral RNA to the cytoplasm (216,226). *Nef*, *vif*, *vpu* and *vpr* are four accessory genes, playing a role in HIV-1 virulence by modifying the cellular environment to optimal conditions for viral replication, and counteracting viral transcription inhibiting cellular mechanisms (216,227). The HIV-1 genome also codes for an antisense open reading frame (ASORF), starting in the 3'LTR, through *env* where the *antisense protein* (*asp*) is located (bp 7373 to 7942) (228).



**Figure 7. HIV-1 viral structure.** **A.** Structure of the viral genome, based on the reference genome HXB2. The genome consists of three open reading frames, and contains nine genes, coding for 18 proteins. Based on (229) **B.** Virion structure of HIV-1. The colors of the genes in (A) correspond with the proteins they code for in the virion structure (B).

## HIV-1 virion structure

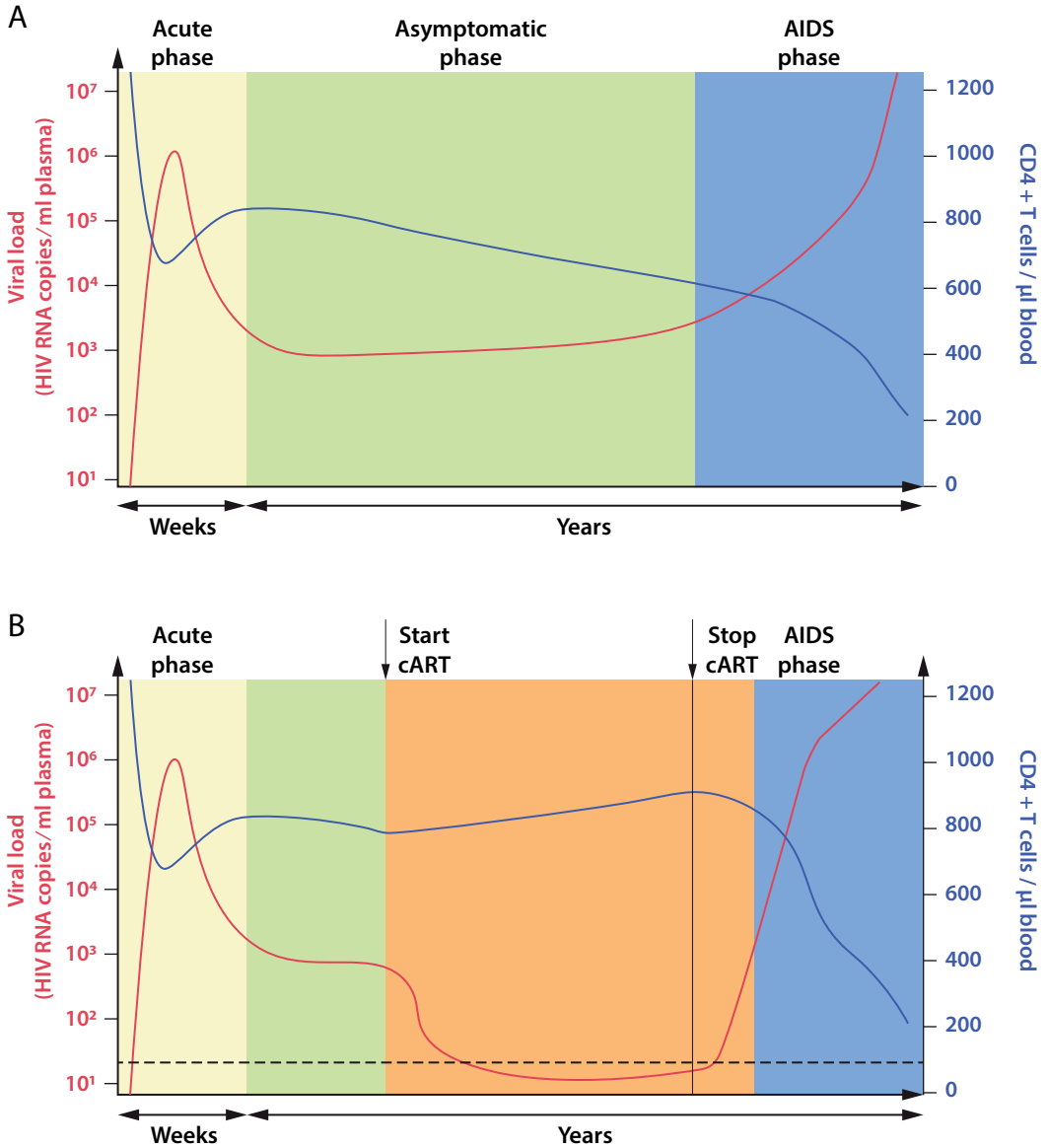
The HIV-1 virions are 100-120 nm-sized spherical particles, surrounded by an outer lipid bilayer envelope obtained from the host cell plasma membrane containing spikes made from surface glycoproteins gp120 (head) and transmembrane glycoproteins gp41 (spike) (**Figure 7B**) (230–232). Within this outer membrane, the matrix, consisting of *gag*-proteins p17, protects and provides structural support to the inner viral capsid. This capsid (mainly *gag*-proteins p24) is a cone structure and forms the inner viral core: it surrounds the two copies of the positive sense, single stranded RNA (+ssRNA) genome of the HIV-1 virus, which is additionally protected by p7 nucleocapsid proteins. This inner capsid also consists of several copies of protease, integrase, reverse transcriptase and accessory proteins (Vif, Nef, Vpu), and cellular factors as proteins, ncRNA and tRNA that serves as primer for reverse transcription (233,234).

## HIV-1/AIDS disease progression

Next to HIV viruses, the genus of the *lentiviridae* covers viruses infecting several vertebrates: SIV (monkeys/great apes), *Visna-maedi* virus (sheep), *equine infectious anemia virus* (horse), *feline immunodeficiency virus* (cat) or *caprine arthritis-encephalitis virus* (goat), *puma lentivirus* (mountain lions) and *Jebrana disease virus* (cow). This genus is characterized by infections causing slowly progressing inflammatory diseases. They are associated with long incubation times and persistent infections: typically in the range of months to years (216). In the case of HIV-1, the time from the moment of infection until AIDS development (final stage) in normal progressors is around 6-10 years (235,236).

Before HIV-1 infected individuals develop AIDS, they go through three disease phases, monitorable by CD4+ T cell count and HIV-1 RNA plasma levels: (i) the acute phase, (ii) the asymptomatic phase, (iii) the symptomatic phase or the AIDS phase (**Figure 8A**) (216,237).





**Figure 8. Disease progression during HIV-1 infections.** **A.** HIV-1 disease progression without antiretroviral therapy (cART). **B.** Effect of cART on HIV-1 disease progression. cART restores the amount of CD4+ T cells and reduces the viral load (VL) drastically: the dashed line represent the limit of detection for the VL. Interruption of cART results in a viral rebound, eventually leading to AIDS. The red lines represent the VL in the plasma of the infected individual. Blue lines represent the amount of CD4+ T cells in the blood of the infected individual.

## *I – acute phase*

The acute phase is during the first weeks to months after initial infection, and starts with the eclipse phase: a local inflammation around the infection area (typically the gut mucosa upon sexual contact) where HIV-1 particles are captured by dendritic cells and macrophages (237,238). The virus stays undetectable in blood plasma and no immune response or symptoms are observed, but foci of infection are formed near these transmission sites due to freely replicating viruses (238). Subsequently, gut associated CD4+ T cells are attracted to this inflammation site, and will become infected and quickly depleted by HIV-1 (238). The dendritic cells with captured HIV will migrate to the lymph nodes in order to initiate the adaptive immune response by presenting HIV-1 antigens for B and T lymphocytes. This migration of HIV-1 viral particles to their target cells results in a rapid spread throughout the body of the infected individual, and the creation of the latent HIV-1 viral reservoir (238). Due to the lack of acquired immunity to HIV-1 antigens, the viral load (VL) peaks at this point, which is three to four weeks after infection. This peak is observed in blood, central nervous system and lymphatic system (up to  $10^7$  HIV-1 RNA copies per ml blood) (216,238). At the same time, the amount of CD4+ T cells decreases drastically from 1200 to less than 800 cells per  $\mu$ l blood (239). Consequently, about 50% of the infected individuals will experience flu-like symptoms during this phase (e.g. fever, sore throat, skin rash or nausea). In the other 50%, the acute phase of the infection will be symptomless (237,238,240–242). Six to eight weeks after the initial infection, the cellular adaptive immune response succeeds in producing the first HIV-1 specific neutralizing antibodies and CD8+ cytotoxic lymphocytes (216,238). This leads to a decrease of the VL, and a temporal and partial increase of the CD4+ T cell concentration. Approximately 12 weeks after initial infection, a balance is reached between immune response and ongoing replication: the VL set point. Due to the lack of symptoms, or the vague nature of the symptoms, HIV-1 infections can be missed in the acute phase.

## *II – asymptomatic phase*

The second phase is the asymptomatic phase, or the clinical latency phase. The patient does not experience any HIV-1 related symptoms. During this phase, the virus does actively replicate in the lymph nodes at a constant rate, and a fast turnover of plasma virions and CD4+ T cells takes place (238). The continuous viral replication in the CD4+ T cells causes persistent immune activation and exhaustion of HIV-1 specific T cells. In addition, the infection of these immune cells leads to cell death of CD4+ T cells. Consequently, it is impossible for the immune system to regenerate the CD4+ T cell pool completely, and the amount of CD4+ T cells drops from near normal (1000 cells/ $\mu$ l) to very low levels (200 cells/ $\mu$ l) (238,243,244). Resulting in a gradual decrease of the functionality of the immune system, together with a constant increase of the VL (up to  $10^5$  copies / mL). The immune system can, however, slow down viral replication, helped by several mechanisms that stimulate the clinical latency: low CCR5 expression (crucial cofactor for fusion of the virus to the host cell), low cellular dNTP levels (needed for reverse transcription) and low cellular ATP levels (crucial for cDNA transport to the nucleus) (245–247).

### *III – symptomatic phase*

Patients enter the third, symptomatic or AIDS phase, at the moment their CD4+ T cell count drops below the critical 200 cells/ $\mu$ l blood (237,238). Consequently, the immune system fails to combat otherwise harmless opportunistic infections (238). This immunodeficiency leads to uncontrolled HIV-1 replication (resulting in heavily increased VL), combined with typical opportunistic infections by *Pneumocystis carinii*, (mucosal) candidiasis and *cytomegalovirus*, and patients develop Kaposi's sarcomas (as a result of the opportunistic virus *Kaposi's sarcoma-associated herpes virus*). In the end, the patient will die from these infections (248).

Although the typical time to develop AIDS is 6 to 10 years, there are 10-15% of infected individuals who will develop AIDS after 2 to 5 years (rapid progressors). On the other hand, a small fraction (around 5% of the infected individuals), can control the viral infection without therapy over ten years (235,236,249). This group can be divided in three subgroups: (i) long-term non-progressors (LTNPs) are infected individuals that maintain CD4+ T cell counts  $> 500$  cells per  $\mu$ l blood and viral RNA load (VL)  $< 1000$  copies / ml plasma for more than 10 years. (ii) Viremic controllers (VC) are characterized by VL  $< 1000$  copies / ml plasma and (iii) elite controllers (EC) controlling viremia to undetectable levels (VL  $< 50$  copies / ml plasma) (236,249–251). These patients form a very interesting study population, since the understanding of the mechanisms in these individuals that control the viral infection might provide insights in the research towards an HIV-1 (functional) cure (252). The mechanisms that drive patients to control the virus long-time are not yet completely understood, but they involve sustained immunity via several host factors such as broadly neutralizing antibody response (249,250). Moreover, the distribution of infected cells shifts towards shorter-lived cell populations in EC (253).

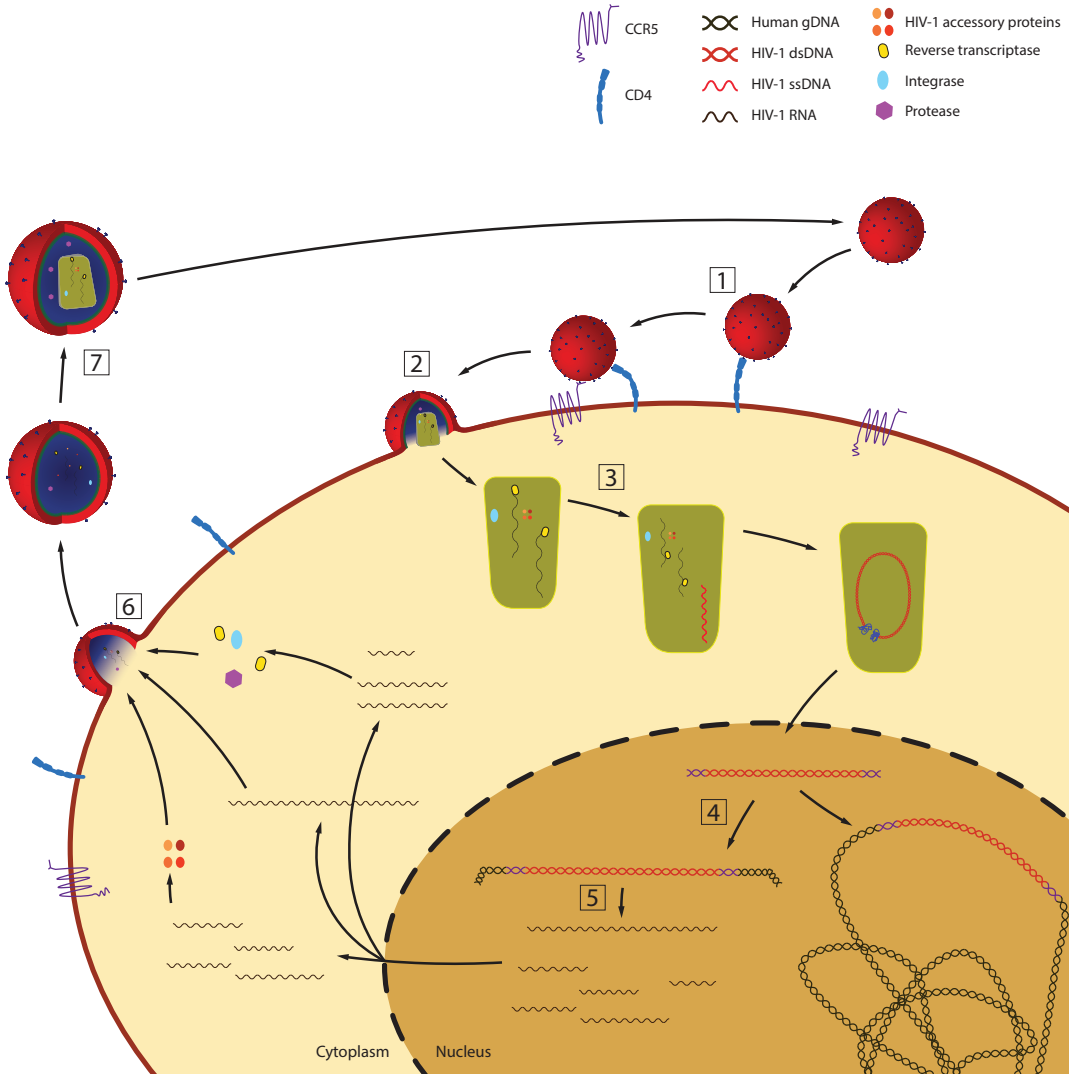
## **HIV-1 life cycle**

To infect a host cell, the gp120 surface proteins of the viral envelope needs to attach to a CD4 receptor on the target cells, which are typically immune cells (CD4+ T lymphocytes, monocytes, natural killer (NK) cells, macrophages, dendritic cells (DC), etc.) (**Figure 9**) (216,254). This will cause a conformational change of the gp120, resulting in the release of the coreceptor (CCR5 or CXCR4) binding site at gp120 and exposure of the gp41 subunit. This leads to the anchoring of the gp41 hydrophobic fusion peptide into the cellular membrane to bring virus and host cell in close proximity. Through a second conformational change, the viral and cellular plasma membranes fuse, resulting in the release of the viral capsid into the cellular cytoplasm by uncoating of the matrix (216,255). The uncoating of the viral core (capsid) was long thought to happen immediately after entry, however, more recent data suggest that this core protects the viral elements during the next step: reverse transcription. The process of reverse transcription is started by generating a reverse transcription complex (RTC), containing viral RNA, reverse transcriptase (RT) protein, tRNA and the capsid surrounding this complex. Other viral and cellular proteins assist the HIV-1 RT in generating double stranded cDNA: first, single stranded DNA (ssDNA) is the negative sense DNA strand, which becomes

complemented to double stranded DNA (dsDNA) and flanked by two identical LTR regions. This RTC is transformed to a pre-integration complex (PIC) when it is transported into the cellular nucleus by cellular mechanisms and it starts interacting with host factors as LEDGF/p75 (256,257). Once in the nucleus, the linear dsDNA is integrated into the host genome, coordinated by the viral encoded integrase enzyme together with several cellular proteins (258,259). Therefore, host chromosomal DNA is opened, and viral DNA, containing overhanging ends, and host DNA are linked by repairing the DNA breaks (258). Most of the HIV-1 genomes do integrate near active transcription sites, since the genomic structure in these regions is accessible (open chromatin, euchromatin) and LEDGF/p75 facilitates integration into these regions, where viral replication is enabled (258,260,261). However, some genomes are integrated into dense, less accessible, non-genic chromosome regions further from active transcription start sites, or in opposite orientation relative to host genes, both hindering subsequent proviral transcription (260).

Once the viral DNA is integrated, this provirus is treated as a cellular gene: it can be transcribed or (epigenetically) silenced by the normal cellular mechanisms. These mechanisms recognize the proviral promoter region and transcription factor binding sites in the HIV-1 LTR region.

The transcription of the HIV-1 genome is regulated by one single promoter, located in the 5' side of the genome, the LTR region (262,263). This LTR region codes for several cis-regulatory elements (CREs), both up- and downstream of the transcription start site (TSS) such as CREB, NF- $\kappa$ B, AP-1, TCF-1 $\alpha$ , SP1 binding sites (264,265). Various agents (chemicals, ultraviolet radiation, bacterial and viral products or cytokines) activate HIV-1 expression using these enhancer elements, and these elements are essential to obtain any significant viral replication during normal HIV-1 progression (264,266). However, in order to replicate, HIV-1 uses several host transcription factors, and both viral as host factors interfere in order to alter the transcription activity. Examples of factors that influence this activity are the integration site of the provirus, transcriptional interference, insufficient levels of cellular transcriptional activators, failing of correct splicing, presence of transcription repressors, chromatin conformation changes (by epigenetic modifications as histone tail modifications or DNA methylation), fluctuations of Tat protein expression (viral transactivator protein) (20,267,268).



**Figure 9. Life cycle of the HIV-1 virus.** 1. Attachment of the virion to the CD4 receptor on the cell surface. 2. Fusion of the viral and cellular plasma membranes and release of the viral capsid into the cytoplasm of the cell. 3. Formation of the reverse transcription complex, and cDNA generation. 4. Pre-integration complex, interacting with host cellular factors and integrase to integrate into the host genome. 5. Integrated HIV-1 DNA in the host genome that is actively transcribed into spliced and unspliced HIV-1 RNA, that is exported towards the cytoplasm and translated into HIV-1 proteins. 6. Assembly of new virions at the cell membrane: HIV-1 proteins, viral unspliced (usRNA) and several host proteins and RNA molecules. Viral budding can occur by the encapsulation of the virion cores in a cellular lipid bilayer containing viral Env proteins. 7. Maturation of the virion: rearrangement of the viral proteins and formation of new infectious particles.

Following active transcription initiation and translation of the mRNA, early viral encoded proteins (typically Tat, Rev, Nef) accumulate. However, further transcriptional elongation is hindered at the RNA TAR region at 5' LTR by two host cellular factors: negative elongation factor (NELF) and DRB-sensitive inducing factor (DSIF) (269). Viral Tat is a crucial protein to overcome this block: it recruits the cellular positive transcription elongation factor b (P-TEFb) enabling elongation of viral RNA and enhancing the transcription drastically. After transcription, Rev transports both full-length HIV-1 RNA and mRNA strands to the cytoplasm, allowing for translation of Env, Vpu, Vpr, Vif and Gag-Pol (226). Consequently, Gag proteins will orchestrate the assembly of new viral particles at the cellular membrane: HIV-1 proteins and viral full-length (unspliced) RNA (usRNA), together with several host proteins and RNA molecules are gathered to form new virions. Eventually, viral budding can occur by the encapsulation of the virion cores in the cellular plasma membrane, which is enriched in viral Env proteins (gp41 and gp120). The final maturation step happens after the viral budding, outside the host cell. It implies cleavage of the structural Gag-Pol polyproteins by HIV-1-encoded protease. This maturation is finished when these proteins are rearranged in the viral core and form a mature infectious particle (254,270). These particles can start a new round of infection of CD4+ cells.

## **HIV-1 life cycle revised**

In the previously described HIV-1 life cycle, which is the productive cycle, new virions are formed, ensuring that the virus can infect new cells and spread throughout its host. However, a subset of HIV-1 proviruses is found to be not contributing to the viral spread in the host. A portion of these proviruses has a complete, non-defective (replication-competent) genome, including a normal promoter region, and they are stably integrated into the host genome, however, they are not actively transcribed. Consequently, these proviruses do not participate in new virion production (271). This phenomenon is called viral (or cellular) latency and it is one of the most important aspects in the quest to understand all aspects of HIV-1 infections (272). This proviral DNA (integrated HIV-1 DNA) is continuously present, even in patients on cART, and can at any moment be (re)activated to (re)start viral replication. This activation of transcription results in a viral rebound in absence of cART (264,272–274). This viral latency is substantially different from the clinical latency as first described by Blattner in 1985, which is the asymptomatic disease phase of the HIV-1 infection. Clinical latency is the absence of clinical symptoms, despite continuous ongoing replication of the virus, while viral latency is absence of viral replication, despite the intactness, and thus the replication competence of the provirus (272). Consequently, during clinical latency, new cells become infected with HIV-1, either productively (supporting further spread of the virus) or latently (viral latency).

Essentially, this viral latency is caused by multiple blocks in transcription after the stable integration of the proviral DNA (275). Both cellular and viral factors are involved in this silencing process, which is beneficial for the host as well as the virus on short-term: the cell will not die as a result of cytotoxicity, and no additional cells will be infected from virions produced by this particular cell. For the virus it is beneficial since a part of the population can hide from the host immune system (when HIV-1 is silenced, the infection is invisible for the immune system) and can proliferate through cell divisions.

On the long term, a transcriptional reactivation can initiate virion production in these cells, which is unfavorable for the host, especially if the HIV-1 infection did exhaust the immune system.

Two forms of viral latency can be discriminated: **pre-integrative** and **post-integrative latency** (**Figure 10**) (253,264,272,273).

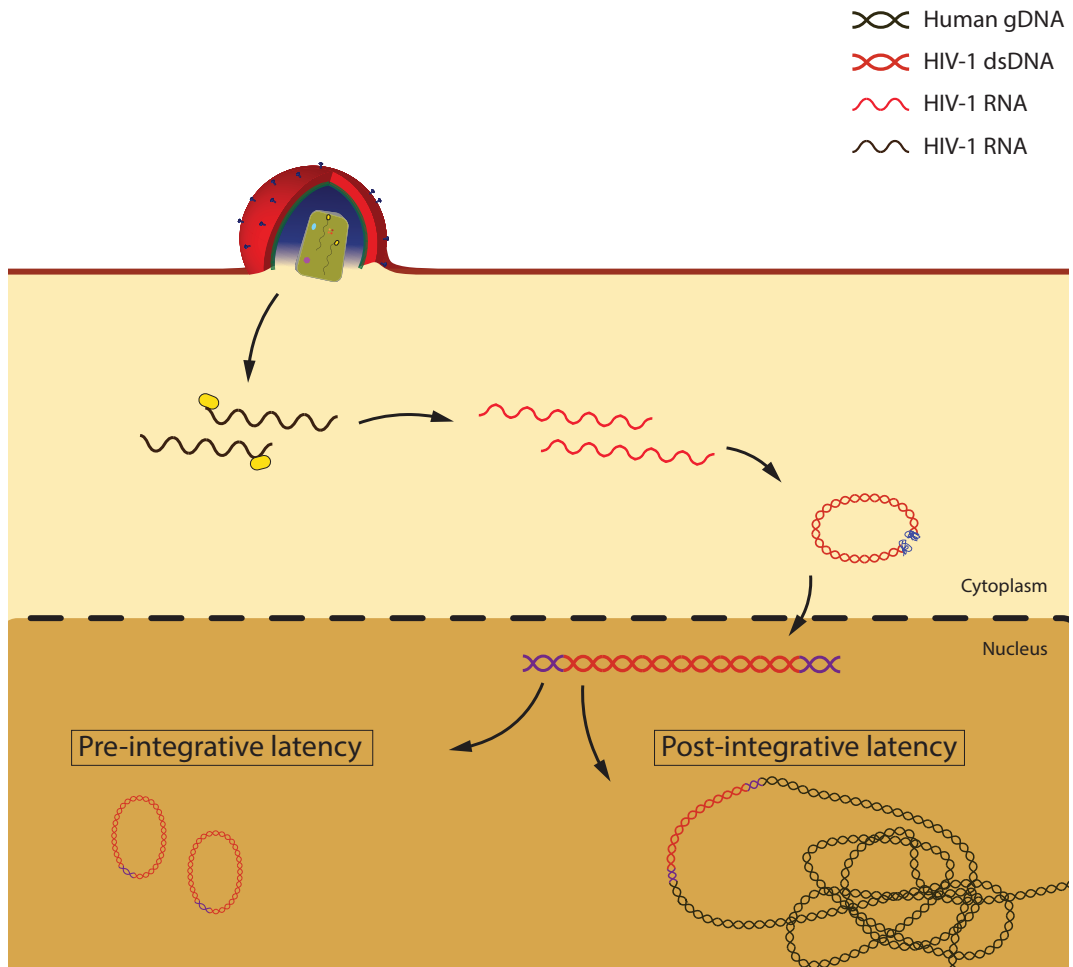
### *I – pre-integrative latency*

Due to several blocks in the HIV-1 viral life cycle, the infection process is very inefficient (276). Pre-integrative latency happens when a virus infects a resting cell: the reverse transcription of the viral RNA genome will be performed and subsequently, the cDNA can be transported to the nucleus, however, it will not always be integrated into the host genome (**Figure 10**) (277–279). These pre-integration complexes (PIC) are found in the nucleus or the cytoplasm of cells in G<sub>0</sub> phase (272). They have short life spans (1-6 days) and the dsDNA strand be either linear or circular (containing one (1-LTR circles) or two (2-LTR circles) copies of the LTR) (253). Due to their short life span, these DNA forms are found to reflect a state of ongoing replication. In some cases, this DNA could be used as a short lived competent template for viral replication (272,280). If the cells are activated, the viral DNA can still be integrated, making this pre-integrated DNA a potential, but rather weak and transient source of the viral reservoir (246,253,272,281).

### *II – post-integrative latency*

Post-integrative latency on the other hand happens after stable integration of a replication-competent provirus into the host genome. Despite this integration, transcription of the proviral genome is inhibited and viral replication is blocked (**Figure 10**) (272,275,282,283). It is this stable latency mechanism which is responsible for the maintenance of the cellular latent reservoir. This is also the main hurdle to find a cure for infections with HIV-1 (274,284). The exact latency establishment and maintenance mechanisms are not known, however, it is assumed that a multiplicity of (transcriptional) regulating mechanisms are involved in this process, all contributing to the long-lived latent reservoir. It is proposed that this latency originates from infections during reversion of an activated cell to a resting (memory) state (253,273,285,286), or that it could be a survival mechanism of activated cells to revert to a resting state after infection (253,272,273,287). However, studies have predicted that latency is a stochastic event, which can happen in both activated and resting T cells (253,275,284).

Moreover, in individuals with prolonged cART treatment, an enrichment in integrated replication-competent proviruses is found in the non-genic regions, with increased distance from actively transcribed regions (260). This suggests a selection of these proviruses that are more difficult to activate, and that the chromosomal density and the transcriptional activity of the integration site (IS) are among the main drivers of viral latency.



**Figure 10. HIV-1 pre- and post-integrative latency.** Reverse transcribed dsDNA enters the nucleus, where it can stay present either as a linear or a circular pre-integration complex. In resting cells, these complexes do often fail to be integrated into the host genome, causing pre-integrative HIV-1 latency. Post-integrative latency happens in cells with successfully integrated HIV-1 genomes with a stable transcriptional block.

It is shown that the frequency of cells with latently integrated HIV-1 DNA is 10 to 100 times higher than cells with active replicating proviral DNA (272). Some infected cells seem to be unresponsive to activating signals, which results in better maintenance of the amount latently infected cells (288).

The current anti-HIV therapy focuses on blocking ongoing replication of the virus, but it does not affect the latently infected cells. Although, the current treatment guidelines are to start treatment as soon as the infection is diagnosed, in order to minimize viral spread, and thus the amount of latently infected cells (289). Interacting with these latently infected cells will be crucial to reduce the latent reservoir, but this requires a better understanding of every single part of the complex multifactorial latency regulating mechanisms.



### *III – replication-deficiency*

Of all integrated proviruses found in an infected individual, 88 - 93% is replication-deficient, which implicates that they, due to mutations, do not encode for replication-competent virion particles (207,290,291). The majority of the mutations are substitution mutations, often leading to stop codons in the genome, and to lesser extent insertions and deletions. Cells carrying replication-deficient proviruses are not capable of fueling a viral rebound (no infectious virions can be produced), so they do not contribute to the latent reservoir (253). Some proviruses will undergo transcription, leading eventually to translation of the viral RNA into viral proteins. This presence of viral antigens can activate and exhaust the immune system of the infected individual. Other proviruses, however, have more substantial defects, and are harmless to the infected individual. These could be accounted for as endogenous retroviral sequences, which are a significant part of all vertebrate genomes (212). The reason for these incomplete retroviral sequences is partly due to a main characteristic of retroviruses: reverse transcription (complementing of viral RNA to cDNA) (212,213). The responsible enzyme, reverse transcriptase (RT), does not have any proofreading capabilities, resulting in a high error rate during cDNA production:  $3 \times 10^{-5}$  mutations per base per round of copying (290,292–296). However, despite this high mutation rate, this RT is estimated to only account for 2% of the mutations found in HIV-1 sequences. The other 98% of mutations is caused by cellular cytidine deaminases of the APOBEC3 family. These enzymes deaminate cytidine (ribose-cytosine) to uridine (ribose-uracil) in the negative viral cDNA strand, ultimately resulting in G to A substitutions in the proviral genomic DNA (290,297–299). These mechanisms combined cause  $4.1 \pm 1.7 \times 10^{-3}$  mutations per bp per cell in PBMCs, which is the highest reported for any biological entity (290). Interestingly, plasma-derived HIV-1 sequences show a 44 times lower mutation frequency, indicating that most of the mutated PBMC-derived HIV-1 genomes are replication-deficient, and that these proviruses fail to reach the plasma (290). Unless most mutations result in replication incompetence of the provirus, they can, combined with a high viral replication rate (up to  $10^9$  virions per day), and potential recombination events of different viral variants, result in HIV-1 populations within an infected individual to become quasispecies (complex heterogenic population of non-identical, but closely related viral genomes): millions of HIV-1 variants are found within a single patient (242,272,300). This viral diversity increases continually during the infection time. By mutating genes of structural proteins in such a way that they do not result in replication-deficiency of the provirus, mutant versions of these protein will be expressed on the cellular surface. These will not yet be recognized by the immune system (301). Consequently, these cells stay under the radar of the adapted immune system, which is an important factor why the immune system cannot clear an HIV-1 infection.

## Viral reservoir

### *I – cellular reservoirs*

The combination of all the latently infected cells (cells with transcriptionally silenced replication-competent proviruses) forms the latent reservoir. This reservoir is established very early during infection, and is the last hurdle towards a complete HIV-1 cure: as long as it is present in the patient, it can be (re)activated and cause a viral rebound (207,253,284). Due to the transcriptional silence of these proviruses, no antigens are produced, making the infection invisible for the immune system.

Since HIV-1 needs CD4 as a receptor at the cellular surface to bind, the reservoir consists of different types of CD4+ T cells (253). These T cells are classified based on their differentiation and their memory status or their effector functions (253,284,302). Naïve CD4+ T cells (Tn) are formed in the bone marrow, and migrate to the thymus, where they differentiate to specialized effector cells: T helper cells (Th), follicular T helper cells (Tfh) and regulatory T cells (Tr) (253,284). These short-lived cells are preferred targets of HIV-1, because of their high CD4 expression. Most of these infected cells die because of the cytotoxicity of the infection. Some of these activated cells can differentiate into long-lived memory T cells: stem cell memory (Tscm), central memory (Tcm), transitional memory (Ttm), effector memory (Tem), migratory memory (Tmm), tissue resident memory (Trm) and terminally differentiated (Ttd) cells (284). By this reversion, infected cells can become long-lived, becoming a part of the latent reservoir. However, non-infected long-lived cells can also become infected by HIV-1, and also contribute to the viral reservoir (284,287).

Innate immune cells as  $\gamma\delta$  T cells, monocytes and macrophages are shown to contain replication-competent HIV-1 DNA in infected individuals. Monocytes originate in the bone marrow, and can differentiate into macrophages, which are not killed by HIV-1 infection *in vitro*, making them likely a portion of the viral reservoir (284,303,304). Dendritic cells do not become infected by HIV-1, but these cells retain virions on their cell surface, making them a potential source of viral rebound (284).

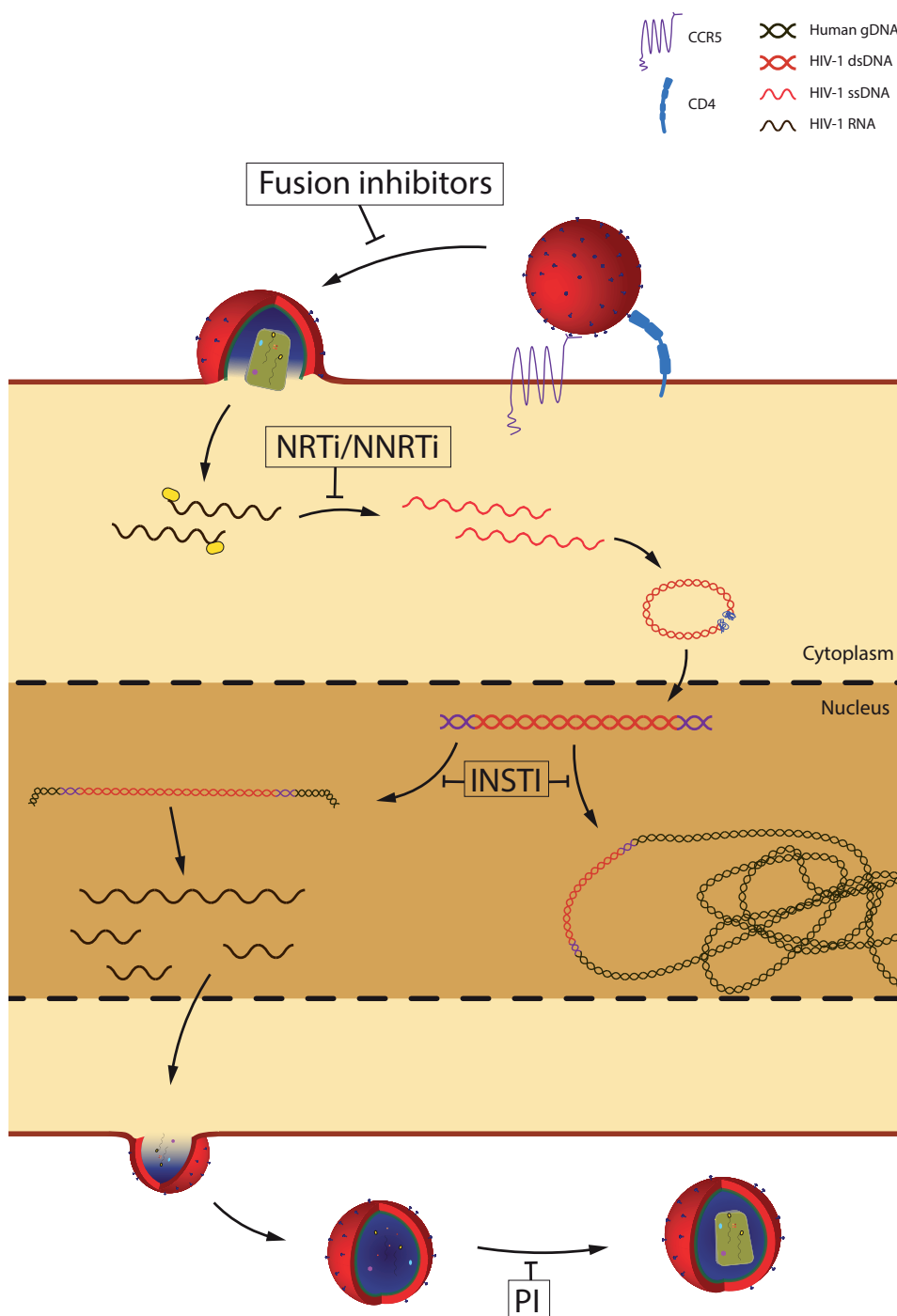
### *II – anatomical reservoirs*

In addition to using infected cell types to define the viral reservoir, it can also be described by infected tissue (anatomical reservoirs). After entering the body, the virus quickly infects lymph nodes, followed by the peripheral blood. From the bloodstream, the virus spreads throughout the whole body by means of transport by immune cells and cell-to-cell transmission. Lymphoid tissue (spleen, thymus, gut associated lymphoid tissue (GALT), lymph nodes) are the most important sites of viral replication, since they contain a lot of tissue-specific cells that can be infected (epithelial cells, microglia, astrocytes or podocytes) (253). Infected cells are also found in several other tissues such as brain (and cerebrospinal fluid), lungs, kidneys, liver, adipose tissue, gastrointestinal tract, male and female genitourinary systems and bone marrow (253,305,306).

## HIV-1 treatment

With current cART, it is possible to prolong the asymptomatic – or clinical latency – phase of the infection drastically: HIV-1 infected patients have similar life expectancy as non-infected individuals, on condition that they adhere strictly to their therapy (202,203). Currently, it is advised to start therapy as early as a patient is diagnosed as this minimizes the size of the reservoir (289). This therapy generally consists of a combination of three active compounds: a backbone of 2 nucleoside reverse transcriptase inhibitors (NRTi), combined with a non-nucleoside reverse transcriptase inhibitors (NNRTi), a boosted protease inhibitor (PI), an integrase strand transfer inhibitor (INSTI) or a fusion inhibitor (**Figure 11**) (307,308).

The combination of three compounds decreases the chances of virus resistance dramatically. As a result of this treatment, patients will typically suppress the blood VL stably to undetectable levels (VL < 50 copies / ml plasma), and CD4+ T cell counts will increase to > 500 cells per  $\mu$ l blood (**Figure 8B**) (237). However, this therapy cannot prevent all ongoing replication in some sanctuary sites, an event that contributes to the viral persistence (253,271). Once this treatment is interrupted, a viral rebound will very quickly be fueled (204,205). The subsequent viral replication will result in cell-free virions, which can again spread over the entire body, activate (latently infected) CD4+ T cells as immune response, and simultaneously infect new cells and stimulate viral spread. This results in a rapid rebound of VL and decrease of CD4+ T cell counts. It is shown that this viral rebound can originate from diverse reservoir compartments (309).



**Figure 11. Antiretroviral therapy in HIV-1 infections.** Several types of antiretroviral therapy (ART) are being used to inhibit viral replication at various locations in the replication cycle of HIV-1. Usually, three compounds are combined (combinational ART, cART) to stably block viral replication and prevent viral resistance. (N)NRTi: (non-)nucleoside reverse transcription inhibitor, INSTI: integrase strand transfer inhibitor, PI: protease inhibitor.

## HIV-1 cure

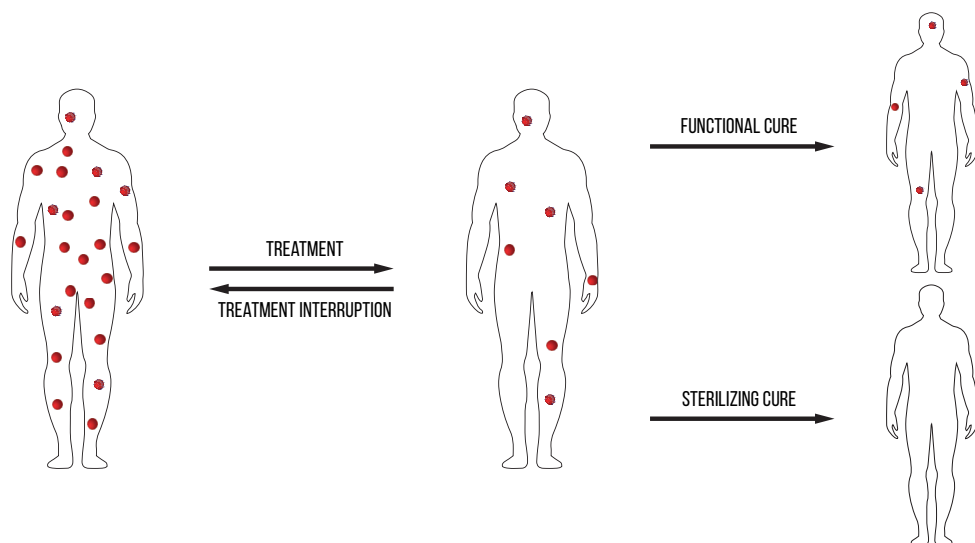
At the moment, HIV-1 infected individuals cannot be cured. A cure is obtained when viral rebound is inhibited, without the need for any additional therapy. The main hurdle to cure infected individuals is the presence of the latent viral reservoir in resting cells, and in body compartments where cART is not as highly concentrated as in the blood stream (e.g. genital tract, brain). Moreover, cART effectively prevents infection of new cells, but it does not prevent a small amount of ongoing viral replication, nor clonal proliferation of infected cells (310–312).

The most certain way to call a patient cured, is when the patients' viral reservoir is completely eradicated, as performed for the first time in the Berlin patient (313,314). However, such a 'sterilizing cure' is quasi impossible to obtain, and extremely difficult to verify within a patient (314). Prediction models showed that not all infected cells should be eradicated, since some cells may never be able to fuel a rebound, and will eventually die (315). Still, dramatic reductions of the reservoir in all reservoir compartments are necessary for a sterilizing cure, which has been found to be a real challenge. Consequently, a 'functional cure', where some proviruses might still be present, but under stable control of the host immune system, is generally thought to be a more feasible strategy (**Figure 12A**) (314).

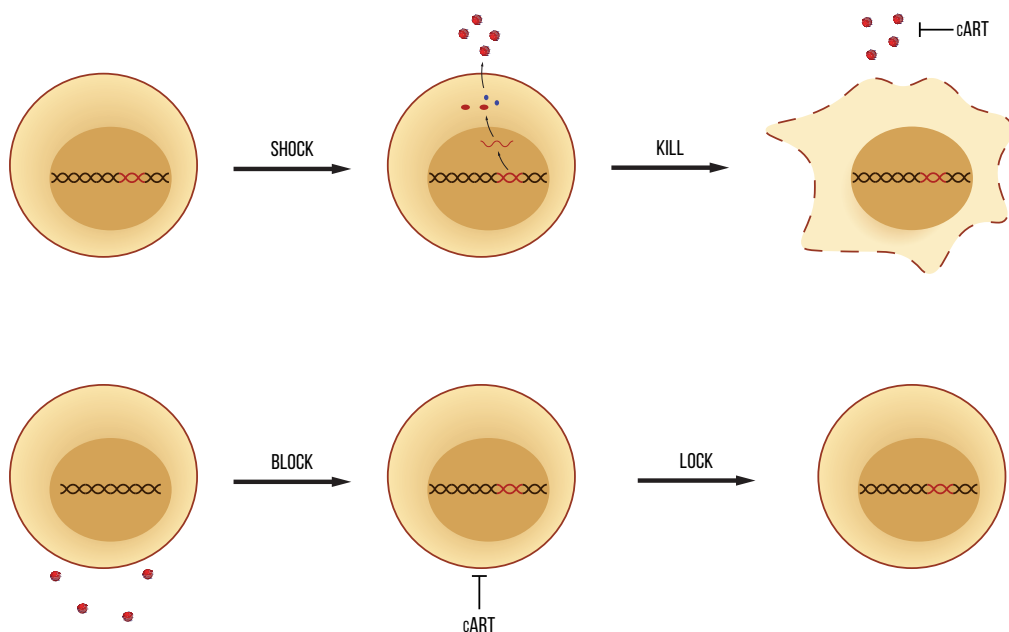
In either case, the size of the reservoir is preferentially as small as possible since this factor delays viral rebound and offers the possibility to control HIV-1 (316). Early cART treatment can minimize the viral reservoir, which can prolong the time to rebound (317). Although, it is shown that only very small reservoirs or early treatment initiation is not sufficient to cure infected individuals.

To date, only one cure strategy has been shown to be successful: an allogenic hematopoietic stem cell transplantation (HSCT) from a CCR5 $\Delta$ 32 homozygous donor. The importance of the donor cells carrying the CCR5 $\Delta$ 32 mutation, is that these cells are resistant to infections with R5-tropic HIV-1 virus (318,319). This strategy was shown to result in long-term HIV-1 remission in two individuals, called the Berlin and the London patient, named to the city where the intervention took place. An HSCT is performed to treat cancers in blood, lymph node or bone marrow such as acute myeloid leukemia or Hodgkin lymphoma. Both the Berlin and the London patients were diagnosed with a cancer that required a HSCT, and they underwent the allogenic CCR5 $\Delta$ 32 HSCT (318,319). This intervention combined with multiple rounds of chemotherapy, graft-versus-host disease and total body irradiation in one of the two patients, eradicated (almost) all of the patients' CD4+ immune cells from the body, which are replaced with the HIV-1 resistant CCR5 $\Delta$ 32 CD4+ T cells. Unfortunately, this strategy is not scalable to all HIV-1 infected individuals since this intervention has a high mortality rate and because of the difficulty to find a matching CCR5 $\Delta$ 32 donor (316).

A



B



**Figure 12. HIV-1 cure strategies.** **A.** Sterilizing cure and functional cure of HIV-1 infections. Treatment initiation in infected individuals results in a significant decline of viral load (VL) within the patient, treatment interruption leads to a reversal to the state before initial treatment start: viral rebound. A functional cure aims to achieve a state where the patients' immune system can control the viral replication (block viral rebound) without the need for cART. A sterilizing cure aims to eliminate all the replication competent proviruses (eliminate the complete reservoir) from a patient. **B.** Shock and kill vs block and lock. Shock and kill aims to reactivate latent proviruses, and simultaneously inhibit further viral spread using cART. Infected cells will either die from cytopathic effects of viral replication, or they will need to be targeted to be killed. In block and lock, infected cells are pushed into a deep latent state, in order to lock them into the cell, preventing them to (re-)initiate viral replication.

Therefore, other cure strategies are highly investigated. The two most studied strategies are 'shock and kill' and 'block and lock' (**Figure 12B**) (317,320,321). The shock and kill strategy aims on minimizing the viral reservoir by a shock intervention, causing reversal of the latency of the integrated proviruses by latency reversing agents (LRAs) (shock). This results in new virion production causing either cytopathic effects or antigen presentation, leading to immune system-based clearance of infected cells (320,322). In addition to the LRA, a kill intervention is needed to eliminate all the cells with reactivated proviruses (317,323). New infections are blocked by cART, resulting eventually in a reduction of the reservoir size, theoretically until no infected cells are present anymore (sterilizing cure). Several types of LRA has been tested to obtain a shock: (i) histone post-translational modification modulators (e.g. histone deacetylase inhibitors (HDACi), histone methyl transferase inhibitors (HMTi)); (ii) Non-histone chromatin modulators (e.g. DNA methylation inhibitors as 5-aza-2'-deoxycytidine (5-aza-CdR, decitabine)); (iii) NF- $\kappa$ B stimulators (protein kinase C (PKC) pathway agonists as PMA, prostatin or bryostatatin); (iv) toll-like receptor (TLR) agonists; (v) extracellular stimulators (e.g. TNF- $\alpha$ , PHA); and (vi) miscellaneous compounds modulating unique cellular mechanisms or have an unknown/unconfirmed mechanism of action (320). In order to obtain an effective shock, combinations of these compounds (shocktails) will most probably be needed, as most LRAs have altering efficacies in different cell types (320,324). Kill strategies are also studied extensively, and include inhibiting pathways that promote cell survival, or stimulate pathways inducing cell death by apoptosis (325). New advances in therapeutic vaccinations and in chimeric antigen receptor T-cells (genetically engineered T-cells that recognize HIV-1 antigens and induce cytotoxic effects of the infected cells) show promising results (326,327).

At the moment, several clinical trials were performed to reactivate latent HIV-1, but no strategy has been developed that succeeds in reactivating the complete reservoir (325,328–334). More active reactivating compounds/shocktails and more effective kill strategies should be developed. Consequently, the interest in exploring alternative strategies as the 'block and lock' strategy is growing (317,320). This cure strategy has been proposed more recently than the shock and kill, and it sets as goal to obtain a functional cure (321). It is essentially the opposite strategy of a shock and kill: in block and lock, latent proviruses are pushed towards a permanent deep latent state, to lock them up into the infected cells, inhibiting reactivation. It is shown that proviruses integrated in non-genic regions with very dense, suppressing chromatin states, are heavily resistant to reactivation (260). Indeed, viral integration site and its chromatin density are suggested to be among the main drivers of viral latency. Most of the locking strategies have not yet been tested *in vivo*, but this approach is theoretically very promising (317). Some examples of HIV-1 locking strategies are to block the viral transcription through increased epigenetic repression, guide HIV integration towards dense non-genic chromatin regions, or to interfere with the viral Tat protein, since this protein is necessary for transcriptional elongation of the viral RNA (317,321). Interfering with the interaction between viral integrase and the host cofactor LEDGF/p75 using small molecule inhibitors as LEDGINS has shown to be a successful strategy to reduce viral integration and relocate the integration site of HIV-1 towards reactivation-resistant genomic regions (261,335). Additionally, these compounds also inhibit efficient viral replication by affecting the produced viral particles morphologically (261,336).

The most crucial step in both cure strategies is the interference with the viral latency. The intervening actions should target only HIV-1 infected cells, and they should target all of these cells in all reservoir compartments. This is heavily complicated by the identical phenotype of non- and latently- infected cells, and by the viral spread throughout the whole body. Before researchers can successfully design novel latency interference strategies, all aspects of the latency regulation should be understood completely. Several latency regulating mechanisms are determined, but the exact mechanisms of this cooperation towards transcriptionally silencing of the proviral genome are still unclear. This multifactorial, and potentially cell type specific regulation of HIV-1 latency complicates the quest to understanding it and should be unraveled in order to advance the search towards an HIV-1 cure.





## III – DNA methylation in HIV-1 proviral transcription regulation

HIV-1 latency could essentially be defined as transcriptional silencing of the proviral genes caused by multiple transcriptional blocks after the stable integration of the proviral DNA into the host genome (275). When the HIV-1 provirus is integrated in the host genome, its transcription is regulated by host-specific mechanisms. Transcriptional regulation is heavily steered by several epigenetic mechanisms as DNA methylation and histone modifications. These mechanisms alter the chromatin architecture, making specific genomic regions accessible or inaccessible for TFs. It can thus be assumed that HIV-1 transcription, and thus HIV-1 latency is regulated, at least partially, by epigenetic modifications. Indeed, several studies showed that targeting epigenetic modifications (e.g. DNA methylation or histone modifications) alters the HIV-1 transcriptional state, indicating that these mechanisms play a role in latency regulation (20,61,337–343). Further studying these mechanisms is crucial to understand their impact on latency and to touch upon their potential in HIV-1 cure strategies.

### Epigenetic features of the HIV-1 genome

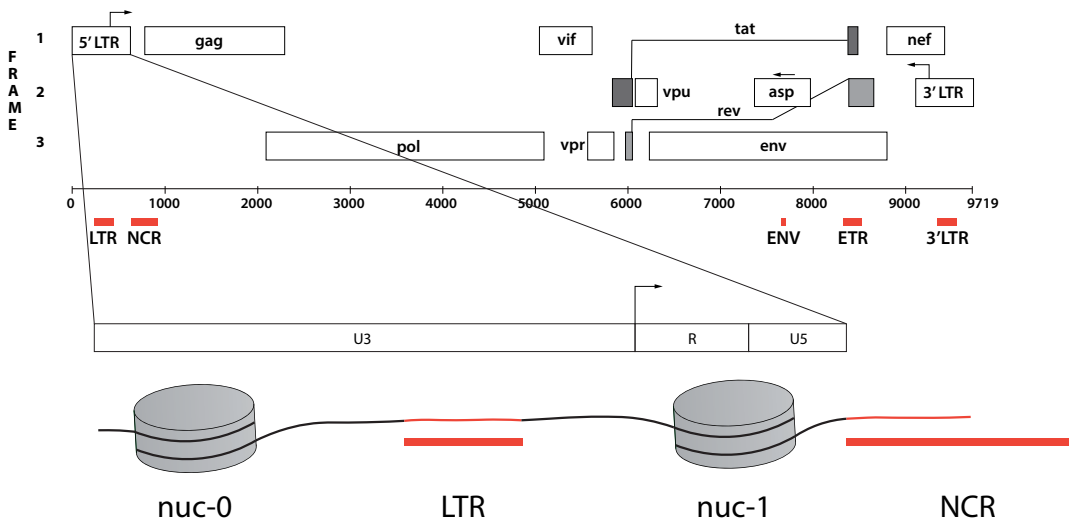
Major epigenetic modifications are histone modifications and DNA methylation. Histone modifications can only happen if nucleosomes can bind the targeted DNA, and DNA methylation will only be effective with sufficient CpG density. The HIV-1 genome contains all of these characteristics: nucleosome binding regions and CpGs are present on this genome.

#### *Nucleosome binding at the HIV-1 5'LTR*

The 5'-LTR of the HIV-1 provirus contains 2 nucleosome binding regions, on which nucleosomes are stably bound in case of a latent infection: nuc-0 and nuc-1 (**Figure 13**) (344–346). Nuc-0 is located at the beginning of the HIV-1 genome in the U3 region of LTR (nt 40-200) (344). No nucleosomes will bind the subsequent region of U3, leaving a gap of roughly 250 bp without nucleosome binding. The nuc-1 binding region starts at the TSS (nt 452-596) (344). Nuc-1 binding causing a transcription elongation block, and proviral reactivation results in nuc-1 displacement (344–347). The predicted nucleosome affinity of nuc-0 and nuc-1 is lower than the predicted affinity of the region between them (348). Therefore, it is suggested that the region between nuc-0 and nuc-1 is an additional nucleosome binding site, where a nucleosome can be positioned rather loosely (345,348). To suppress viral transcription, this nucleosome is displaced towards the energetically less favorable nuc-1 position, in order to hinder the transcriptional elongation. Proviral reactivation induces an opposite nucleosome displacement (348). In region downstream of nuc-1, another nucleosome-free region of 124 nt is found. This linker region is longer than the usual region between two nucleosomes, but the subsequent nucleosomes (nuc-2, nuc-3 and nuc-4) are tightly positioned, starting at nt 720 (344,346).

## CpGIs in the HIV-1 genome

In line with eukaryotic genomes and many other RNA viruses, CpG suppression is also observed throughout the HIV-1 genome (61,71). Moreover, the CpG dinucleotides occur to be located in clusters (CpGIs). CpGIs in a proviral genome can be identified as CpGIs in the human genome, however, this genome is extremely short in comparison with the human genome. The definition of a CpGI has to be interpreted, and can be flexible in length, CG% and/or observed/expected CpG ratio, resulting in the identification of the 5 CpGIs in the HIV genome (**Figure 13**) (61,337). Two CpGIs are located at the 5' LTR region (the promoter region): CpGI LTR and CpGI Non-Coding Region (NCR). These CpGIs are flanking HIV-1 transcription start site and several transcription factor binding sites (e.g. TCF-1 $\alpha$ , NF- $\kappa$ B, SP1), and they are located at regions where nucleosomes do not bind (344,349). These two characteristics are often associated with bona fide CpGIs (84). The *env* gene does also contain 2 CpGIs: CpGI ENV and CpGI *env-tat-rev* (ETR) (61). These two CpGI are flanking the start of the HIV-1 antisense open reading frame (ASORF) of the antisense protein (*asp*) (61,228). However, the presence of the ENV CpGIs is only found in 35% of the sequences in the Los Alamos National Library (LANL) database, and is only conserved in subtypes D, F2 and AE. The final CpGI is located in the 3' LTR (CpGI 3' LTR) and is positioned at the antisense transcription start site (61).



**Figure 13. CpGIs in the HIV-1 genome and nucleosome binding targets at the 5'LTR.** The HIV-1 genome codes for 5 CpGIs, which are indicated by red bars: LTR (long terminal repeat) and NCR (non-coding region) in the 5' LTR region, ENV (enveloppe) and ETR (*env-tat-rev*) at the *env* gene, and an additional 3'LTR CpGI. Nuc-0 and nuc-1 binding regions are located in the 5' LTR, starting at the beginning of the genome and at the transcription start site respectively. HIV-1 gene map is based on (229).

## DNA methylation in HIV-1

The presence of the CpGIs in the HIV-1 proviral genome suggests that DNA methylation is one of the factors of the multifactorial HIV-1 latency regulation. Indeed, since the involvement of DNA methylation in HIV-1 latency was first described in 1987 (17), it was confirmed in several latency models and cultured primary cells that methylation density in the CpGIs of the promoter region of the HIV-1 provirus is associated with silencing stability: inducing DNA methylation can initiate/stabilize HIV-1 latency (19,20,61,337,338). Moreover, methylation inhibitors (e.g. 5-aza-CdR) cause HIV-1 reactivation and show clear synergistic effects with other LRAs as HDACis (18–21,61,337,338,350–352). These studies were mainly focused on LTR methylation and showed that DNA methylation in the promoter region was a crucial regulator of latency. The findings of these studies were in line with the general idea of transcription regulation by DNA methylation: hypermethylation of the promoter region suppresses both basal promoter activity and responses to activating stimuli, hypomethylation is a transcription mark (352).

However, the DNA methylation studies on DNA obtained from infected individuals' samples showed contrasting results. Most studies reported low LTR methylation, while some studies reported high DNA methylation in a subset of LTNPs (19,20,61,337,350,353,354). However, several correlations were shown, such as increasing DNA methylation with increasing time on treatment (20), or with increasing time of infection in LTNPs (354). These contrasting results can be explained in part by the observation of Cortés-Rubio et al., the first study to measure LTR methylation in patients longitudinally (355). They showed that over time, in CD4+ T cells, the methylation varies heavily. These *in vivo* data are not consistent with the rather straightforward *in vitro* LTR methylation data. Additionally, these contrasting results can also be explained in part by the fact that different studies used different cellular subsets (PBMCs, CD4+ T cells, resting CD4+ T cells or memory CD4+ T cells). Different CD4+ T cell subsets do react differently on (epigenetic) LRAs as HDACis, which indicates epigenetic differences between cell types (324).

The role of intragenic DNA methylation – methylation within the gene body – is, in contrary to the promoter methylation, less straightforward (98–102). Initial studies suggested that this modification could have a role in activation of retroviruses, repetitive elements, alternative splicing, or in transcription initiation in canonical promoters in embryonic stem cells, preventing the production of aberrant transcripts (100–102). Some previous *in vivo* HIV-1 methylation studies did include *env* methylation analysis (20,207,350,353). Three out of four of these studies, showed higher methylation in *env* compared to LTR, however, they did not report it as a biological relevant results, but as a control that methylation could happen on the proviral genome (20,207,353).

At the moment, very little is known about the role of HIV-1 proviral DNA methylation in latency regulation *in vivo*. This is partly due to the limited data available at the moment: all previously mentioned studies combined provide data from 115 patients. This makes it difficult to compare, especially since most of the data is generated using low throughput techniques (clonal Sanger sequencing), using incomparable patient cohorts and different cellular subsets.

## Challenges of DNA methylation analysis of the HIV-1 provirus

In general, analysis of epigenome or epigenetic profiles (e.g. DNA methylation studies) is more challenging than classical genome analysis. Standard molecular biological analysis methods as PCR or cloning, erase the methylation profiles from the DNA (356). Therefore, most DNA methylation analysis methods make use of a pretreatment or an enrichment of the methylated DNA. These procedures result generally in a loss of genetic material due to harsh chemical procedures or washing steps. These methodological difficulties of DNA methylation analysis have as consequence that these methylation profiling methods are usually not used in routine clinical screening (110). Next to the technical limitations, the interpretation of DNA methylation results is often impeded due to (i) need of conversion of the DNA, (ii) bulk DNA needed as input, impeding cell specific – or even cell type specific differential methylation analysis and (iii) impossibility to combine the DNA methylation profile with other linked/related biological pathways (e.g. RNA transcription, histone modifications or nucleosome density).

To study HIV-1 genomes *in vivo*, the low abundance of HIV-1 infected cells in the patient is a huge stumbling block: a typical HIV-1 patient harbors about 50-200 total HIV-1 proviral copies per million PBMCs, measured by PCR, including integrated proviral DNA (from which 88-93% is replication-deficient) and episomal DNA (1- and 2-LTR circles) (207,291,357). Due to this low abundance of infected cells in patient, signal amplification is crucial, which is mostly done by PCR amplification of the target DNA. Therefore, to study the methylation profiles of HIV-1 proviral genomes at single bp resolution, (bisulfite) conversion of the DNA is necessary to retain the methylation information after PCR amplification. Previous *in vitro* and *in vivo* HIV-1 proviral DNA methylation studies did use targeted bisulfite sequencing of specific proviral regions (17,18,351–355,358–361,19–21,61,207,264,337,350). As mentioned before, sodium bisulfite treatment degrades most of the DNA, reducing the amount of already limited HIV-1 DNA drastically. Therefore, the use of bulk DNA samples is difficult to avoid. This does hamper interpretation of the data heavily: epigenetic profiles are cell type-, tissue- and environment dependent. Different methylation patterns in different cells or cell types, could mask the methylation signal of other cells. Environmental influences include treatment regimen, diet, age, sex, duration of infection and other diseases.

In addition, the integration site (IS) of a provirus (and thus its epigenetic environment) and the replication competence are two more factors that might influence the proviral epigenetic profile. The DNA degradation does also impede combined analysis of DNA methylation state, IS or replication competence of the HIV-1 genome. Moreover, the fragmentary character of the DNA and the use of bulk samples impedes separation between data from active replicating, latent, or replication-deficient proviruses.

Moreover, the high genomic variability both between patients as within every infected individual, hinders the development of one single uniform assay for a complete patient cohort: this genomic

variability, restricts a single assay to an undefined subset patients or proviral genomes within the patient. This high variability of HIV-1 proviruses also biases bisulfite-based analysis: amplification of the region of interest can be biased due to mismatches, suboptimal annealing or deletions, making primer design and mapping of sequencing reads very hard.

Another hurdle to overcome in HIV-1 (epi)genomic analysis is the spread throughout the body (HIV-1 resides in several body compartments and in several cellular subsets). It is impossible to sample all compartments routinely or isolate all separate cellular subsets, so every study is limited to a part of the reservoir.

All these technical issues have led to the development of several easily manageable *in vitro* models to study HIV-1 characteristics. These models are ideal for developing and optimizing (epi)genetic assays, however, the translation to *in vivo* information is often not straightforward. Especially for epigenetic alterations, *in vitro* settings are mostly not representative for *in vivo* situations (362–364). Epigenetic modifications are influenced by the environment, and it is shown that even the change of culture medium can alter the DNA methylation profiles. Moreover, some models undergo immortalization, changing the methylome completely (362,363). These observations, together with the uniformity in cell lines, which is not found in patients, make cell lines often poor predictors for *in vivo* DNA methylation regulation. Indeed, in HIV-1 studies, it is shown that proviral methylation patterns, and their effects in cultured cells differ drastically from these in patients.

Consequently, in order to understand the regulatory role of HIV-1 proviral DNA methylation in latency, it has to be studied in patient samples. Therefore, methodologies will have to be optimized to make the information interpretable. When using targeted bisulfite sequencing, minimizing the DNA fragmentation, optimizing primer design and optimizing patient sampling are the crucial elements for successful DNA methylation assessment.

Also, development of alternative methods to analyze the role of DNA methylation in HIV-1 latency are desirable. A single cell DNA methylation analysis method would be a useful tool, especially if the methylation profile could be linked to a transcriptional state, an integration site or the replication competence of the provirus. However, due to the low abundance of infected cells in a patient, the impossibility to phenotypically identify latently infected cells without activating HIV-1 expression (which would alter the methylation profile), and the high genomic variability, this development has not resulted in a workable assay yet.



# CHAPTER II

## RESEARCH OBJECTIVES





## II – Research Objectives

The HIV-1 proviral genome integrates stably into the host genome, where it is subjected to the human gene regulating mechanisms such as DNA methylation. The involvement of DNA methylation in the HIV-1 latency process is clearly indicated in *in vitro* models, but the underlying mechanisms *in vivo* are hardly understood. In most infected individuals, the number of infected cells is very low, making the availability of target sequences very low. Standard molecular techniques to amplify these sequences (e.g. PCR) do not retain the epigenetic modifications in DNA. Consequently, epigenetic and epigenomic methods require a more complex workflow compared to genetic analyses. The gold standard method for DNA methylation analysis is sodium bisulfite conversion of the DNA coupled with PCR amplification. Using standard lab equipment and basic techniques, it provides single base-pair resolution methylation data. However, this conversion implies degradation of most of the treated DNA. This, combined with the low abundance of target sequences and the high genetic variance of the HIV-1 genomes *in vitro* hinders epigenetic analysis of HIV-1 proviral DNA drastically. Consequently, no standardized HIV-1 DNA methylation assay is available. However, it is important to understand every piece of the multifactorial puzzle of this latency regulation. In order to get a better insight in the role of proviral DNA methylation in HIV-1 latency regulation, we focused on developing a **bisulfite-based** DNA methylation assay of the HIV-1 genome to measure *in vivo* proviral DNA methylation signatures in several patient cohorts (objective I).

This bisulfite-based method uses bulk cellular DNA, which potentially averages signals out and masking the real signal and regulating effect. Cell-to-cell variation can be measured using single cell DNA methylation assays. The number of single cell DNA methylation analysis methods is increasing, however, no reliable targeted single cell DNA methylation assay is described yet. In order to be conclusive about regulating effects of this modification, a combinatorial assay, capable of analyzing DNA methylation simultaneously with associated gene transcription, histone modifications or nucleosome density is crucial. We developed a **single cell epigenetic visualization assay**, which can be used to measure host gene methylation, including integrated genes such as HIV-1 proviral genes (objective II). This assay also allows combinatorial analyses with RNA transcription.

## Objective I – Bisulfite-based HIV-1 proviral DNA methylation assay

Since bisulfite treatment provides methylation information at single base-pair resolution, is not expensive, fast, and it requires no specialized lab equipment, it is the gold standard method to measure DNA methylation. However, in the case of HIV-1 DNA methylation analysis, researchers have to deal with the high mutation rate and low amounts of target DNA.

We optimized a targeted bisulfite treatment-based DNA methylation assay to measure the HIV-1 proviral DNA methylation, based on methods used in other studies. It should ideally be useful for both cultured cells and (HIV-1) patient cellular material. To be generally applicable, the assay needs to cover the high variability of the HIV-1 genomes found in patients, and it needs to have a high specificity in order to prevent analysis of endogenous retroviral elements (or other retrovirus-like DNA sequences). It also has to be extremely sensitive and it has to minimize the DNA losses to handle the very low concentration of target (HIV-1) genomes.

The assay basically consisted of three steps, from which the first two needed optimization: (i) **bisulfite treatment** of the DNA, (ii) **PCR amplification** of the bisulfite treated samples and (iii) **next-generation sequencing** (NGS) of the amplicons.

### Objective Ia – Bisulfite treatment optimization

Optimization of the bisulfite treatment to

- Minimize the DNA fragmentation
- Maximize the DNA recovery
- Maintain high conversion efficiency

### Objective Ib – Amplification of HIV-1 CpGs

Development of PCR-assays to analyze four out of five CpGs of the HIV-1 genome, taking into account

- The high HIV-1 genomic variability
- Low amount of target DNA

## Objective Ic – Analysis of DNA methylation in HIV-1 patient cohorts

Measuring proviral DNA methylation profiles in a large, well-characterized patient cohort in order to get insight in the latency regulating role of this modification.

- 72 patients, divided into 4 patient cohorts
  - Late treated individuals
  - Early treated individuals
  - Acute seroconverters
  - Long-term non-progressors

## Objective II – Single cell DNA methylation assay

To measure cell-to-cell epigenetic variation, single cell epigenetic profiles needs to be studied. Moreover, in order to understand the effects of epigenetic modifications on transcriptional activity, this single cell epigenetic profile should be linked directly to the RNA transcriptional profile. The epigenetic visualization assay (EVA) enables us to measure single cell epigenetic profiles of single genes, together with transcriptional activity of that gene. This FISH-based method uses target specific probes and fluorescence microscopy for visualization. This assay will complement the expected emerging (multi-)’omic’ single cell approaches to study specific epigenetic alterations in genomic targets, both in promoter and intragenic regions.

Single cell epigenetic visualization assay development and optimization in cultured cells:

- Multiple targets: rDNA methylation in HEK 293 cells
- Assay validation: Xist DNA methylation in HUVEC and GM-5505 cells
- One bi-allelic target: EGR1 DNA methylation in Jurkat cells
- Single-allelic target: HIV-1 DNA methylation in J-Lat cells
- Combinational EVA: EVA + RNA/DNA FISH





# CHAPTER III

## RESULTS



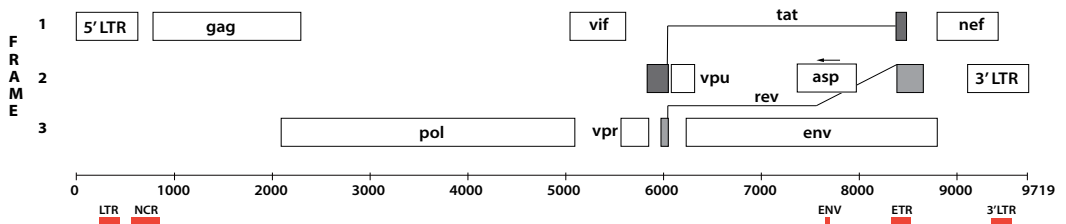


# III – Results

## Objective I – Bisulfite-based DNA methylation assay

The HIV-1 proviral bisulfite treatment-based DNA methylation assay is a targeted methylation assessment, which only provides information of the selected target genomic region. In this case, we were interested in the DNA methylation in four of the five CpGIs found in the HIV-1 genome: two CpGIs are located in the promoter (LTR) region: CpGI LTR and CpGI NCR. Two others are intragenic CpGIs located in the *env* gene: CpGI ENV and CpGI ETR. The fifth CpGI (3'LTR) was not analyzed, since at the moment the assay was designed, the 3'LTR was thought to be an irrelevant copy of the 5'LTR (promoter region). The assay consists of three separate steps: bisulfite treatment, followed by PCR amplification of the regions of interest, and finished by next-generation sequencing of the amplicons.

The bisulfite treatment procedure of human genomic DNA and the HIV-1 proviral CpGI amplification of the treated DNA were optimized during this thesis (**objectives Ia and Ib**). This optimized protocol could then be used in patient samples to measure HIV-1 proviral DNA methylation in PBMCs of infected individuals (**objective Ic**) (**Figure 14**).



**Figure 14.** The HIV-1 gene map indicating the four CpGIs (red bars) that are assayed in the bisulfite-based DNA methylation assay. Based on (229).

## Objective Ia – Bisulfite treatment optimization

For the optimization of the bisulfite treatment, we started with an in-depth comparison of commercially available bisulfite kits, focused on DNA fragmentation, DNA recovery and conversion efficiency.

The work was presented in “*Kint S, De Spiegelaere W, De Kesel J, Vandekerckhove L, Van Criekinge W. Evaluation of bisulfite kits for DNA methylation profiling in terms of DNA fragmentation and DNA recovery using digital PCR. Albertini E, editor. PLoS One. 2018 Jun 14;13(6):e0199091*”, (12) and is adapted to fit in this thesis.

### Contribution of the doctorandus:

Study design

Experimental work

Data analysis, visualization and interpretation

Manuscript writing

# **Evaluation of bisulfite kits for DNA methylation profiling in terms of DNA fragmentation and DNA recovery using digital PCR**

Sam Kint<sup>1,2\*</sup>, Ward De Spiegelaere<sup>3</sup>, Jonas De Kesel<sup>4</sup>, Linos Vandekerckhove<sup>2</sup>, Wim Van Criekinge<sup>1</sup>

<sup>1</sup>Department of Data Analysis and Mathematical Modelling, Faculty of Bioscience Engineering, Ghent University, Ghent, Belgium.

<sup>2</sup>HIV Cure Research Center, Department of Internal Medicine, Faculty of Medicine and Health Sciences, Ghent University and Ghent University Hospital, Ghent, Belgium.

<sup>3</sup>Department of Morphology, Faculty of Veterinary Medicine, Ghent University, Merelbeke, Belgium.

<sup>4</sup>Department of Biotechnology, Faculty of Bioscience Engineering, Ghent University, Ghent, Belgium.

\* Corresponding author

E-mail: Sam.Kint@UGent.be

## *Abstract*

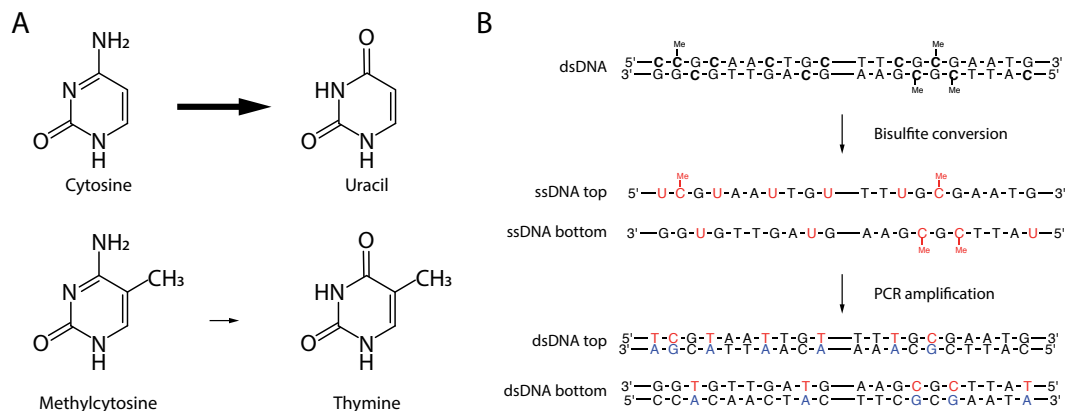
DNA methylation is one of the most important epigenetic modifications in the regulation of gene transcription. The current gold standard to study this modification is bisulfite sequencing. Although multiple commercial bisulfite treatment kits provide good conversion efficiencies, DNA loss and especially DNA fragmentation remain troublesome. This hampers DNA methylation profiling of long DNA sequences. Here, we explored the performance of twelve commercial bisulfite kits by an in-depth comparison of DNA fragmentation using gel electrophoresis, qPCR and digital PCR, DNA recovery by spectroscopic measurements and digital PCR and conversion efficiency by next generation sequencing. The results show a clear performance difference between the bisulfite kits, and depending on the specific goal of the study, the most appropriate kit might differ. Moreover, we demonstrated that digital PCR is a valuable method to monitor both DNA fragmentation as well as DNA recovery after bisulfite treatment.

## *Introduction*

DNA methylation (5-methylcytosine) is an epigenetic modification that is typically associated with stable transcriptional silencing (13,82,83,90). This modification plays an important role in several biological processes associated with development and disease. Examples are cell differentiation, regulation of gene expression, X-chromosome inactivation and genomic imprinting (13–15). In disease, DNA methylation is heavily involved in the development of diseases as Alzheimer's, Parkinson, carcinogenesis and silencing of intracellular viruses (10,11,24,16–23). Consequently, DNA methylation provides a promising diagnostic tool in medicine. Previous studies to understand the exact role of DNA methylation in these disease settings have already resulted in several clinically validated biomarkers (e.g. MGMT promoter methylation in patients with glioblastoma (PredictMDx, MDxHealth, Inc.); methylation of GSTP1, APC and RASSF1 genes for prostate cancer testing (ConfirmMDx, MDxHealth, Inc.); methylation of PITX2 in Formalin-Fixed, Paraffin-Embedded prostatectomy specimens for identifying patients who are at high risk to suffer from prostate-specific antigen recurrence after radical prostatectomy; free-circulating methylated SEPT9 gene copies in plasma as a screening biomarker for colorectal cancer; SHOX2 DNA methylation as a plasma based biomarker for detection of lung cancer) (365,366,375,367–374).

The analysis of DNA methylation profiles is typically done through methylation-specific PCR or bisulfite sequencing of DNA (128,135). Bisulfite (HSO<sub>3</sub><sup>-</sup>) treatment is first described by Frommer et al. (128) in 1992, and it is still the gold standard to analyze DNA methylation. This chemical deaminates unmethylated cytosines (C), but not methylated cytosines (mC) to uracil (U), enabling the analysis of the methylation profile through sequencing (**Figure 15**) (128,376). Despite its wide use, bisulfite treatment has its disadvantages. Bisulfite conversion causes DNA fragmentation, resulting in small sequences, typically smaller than 500 nucleotides (nt) (153,154,377,378). This is a result of the aggressive reaction condition of this conversion: pH 5 and temperatures up to 90°C (152–154,369). The DNA fragmentation is mainly caused by depyrimidation followed by alkali

treatment, which leads to abasic sites, resulting by DNA cleavage of the DNA phosphodiester bond (i.e. DNA degradation) (152). Consequently, the analysis of the methylation of large CpG-islands (CpGIs) is hampered (379).



**Figure 15. Principle of bisulfite-mediated methylcytosine (mC) mapping.** **A.** Deamination of cytosine (C) and mC. Sodium bisulfite deaminates C to uracil (U) (upper row) and mC to thymine (T) (lower row). The rate of mC deamination is two orders of magnitude less than that of C. **B.** The mapping protocol: after bisulfite treatment, mC will remain C where unmethylated C will be deaminated to U. During subsequent PCR amplification, the U deamination product templates adenine (A), which then templates T, resulting in a C to T transition at unmethylated C. By sequencing, mC can be identified as bases that remained C after bisulfite treatment (61).

It is assumed that longer reaction times and higher conversion temperatures cause relatively more degradation (153,154,369). However, if the temperature is too low or the reaction time is too short, the conversion might be incomplete, resulting in an overestimation of the methylation in the analyzed fragments. Therefore, the final conversion reaction has to be balanced out between desired (conversion of Cs) and undesired effects (DNA fragmentation and inappropriate conversion (conversion of mC)) (369). A number of studies have compared the performance of different commercially available bisulfite kits, however, fragmentation or degradation of the treated DNA was never the focus of these studies (369,380–382). Yet, this issue seems very important for some studies where the analysis of long fragments is necessary (61,146). For example, in the HIV genome, an important CpGI is situated in the 5' Long Terminal Repeat (LTR) region, and the genome is flanked by two identical LTRs (61). Yet, mainly the 5' LTR is of interest, since this LTR contains the promoter region (19,20,61). To determine the methylation pattern of the promoter region of the HIV genome, we have to be able to discriminate the 5' from the 3' LTR. This can only be performed by amplifying DNA fragments covering the LTR as well as its flanking region. As a result, DNA fragments have to be longer than 600 nt, which is problematic for the analysis after bisulfite treatment. The issue of DNA fragmentation has also been observed by Meissner et al. (146) in the development of an approach for a large-scale high-resolution DNA methylation analysis termed reduced representation bisulfite sequencing. In this protocol, DNA is digested, and fragments of 500-600 nt are selected, and analyzed via bisulfite sequencing.

DNA yield analysis of bisulfite kits is typically performed with spectroscopic measurements and DNA fragmentation analysis is performed by gel electrophoresis or with quantitative real-time PCR (qPCR). In these cases, multiple analysis methods are necessary. Digital PCR (dPCR) would enable us to assess the DNA yield and fragmentation with one single method. This absolute quantification method bypasses bias from PCR inhibition since it provides an end-point measurement of the positive signal enabling us to directly compare the DNA product prior to bisulfite treatment with the product after treatment. Because of the absolute quantification, it can replace multiple measurements with qubit, gel electrophoresis and qPCR.

In the current study, we provide a comprehensive procedure using state of the art technologies as dPCR to evaluate bisulfite treatment protocols by comparing twelve commercially available bisulfite kits, with focus on DNA fragmentation. The evaluation procedure consists of spectroscopic measurements and dPCR (DNA yield), gel electrophoresis, qPCR and dPCR (DNA degradation) and next generation sequencing (conversion efficiency). We show that dPCR is a valuable method to investigate the quality of bisulfite treated DNA. By using dPCR, our workflow provides a method for differentiation of DNA loss from DNA fragmentation.

## *Material and methods*

### **Human blood samples**

Blood samples from healthy individuals were purchased from the Belgian Red Cross. All donors signed a medical questionnaire containing an informed consent. This consent states that the donated blood can be used for scientific and epidemiologic research if the blood was refused for transfusion. The use of this blood was approved by the Ethical committee of Ghent University Hospital with reference B670201317826.

### **DNA samples**

Peripheral blood mononuclear cells (PBMCs) from five healthy human donors were isolated using a lymphoprep centrifugation. DNA from aliquots of  $10^7$  PBMCs was isolated using the DNeasy<sup>®</sup> Blood & Tissue Kit (Qiagen, 69504). The DNA samples were all obtained on the same moment, and they were immediately aliquoted into 20 aliquots to ensure the same amount of freeze-thaw cycles for every sample. The DNA concentrations of the samples were determined with the Qubit dsDNA BR (broad range) Assay Kit (Q32850, ThermoFisher Scientific) on a Qubit 2.0 fluorometer, and the exact concentration was verified using the HS (high sensitivity) Assay Kits (Q32851, ThermoFisher Scientific).

## **Bisulfite treatment**

### *First analysis - selection of the best kits*

DNA from the five donor samples was bisulfite treated with twelve different commercially available kits (**Table 1**). Every sample was converted in duplicate, and after conversion, the duplicates were pooled to provide sufficient material for subsequent testing. For all the kits, the standard protocols provided by the manufacturer with the suggested input/elution volumes were used, and if not exactly provided, the average of the indicated minimal and maximal input/elution were used.

### *Second analysis - effect of time and temperature on fragmentation*

Based on the first analysis, the Epiect Bisulfite kit (Qiagen, 59110) was selected as the least fragmenting kit. Subsequently, this kit was used to perform the protocol identically as described above on the five PBMC samples, except for the conversion time or temperature which were modified (**Table 2** and **Table 3**). 1 µg of PBMC DNA from donor 3 (160 ng/µl gDNA) was used, and every conversion was performed in duplicate. The converted DNA was eluted twice in 20 µl elution buffer, and the duplicates were pooled.



**Table 1. Overview over the kits and their main characteristics and performance results.**

Kit	Name	Short name	Conversion temperature (°C) <sup>a</sup>	Conversion time (min) <sup>a</sup>	Input (ng DNA)	Elution volume (µl)	Recovery		Fragmentation			Conversion efficiency (% ± SD)
							Qubit (ranking)	Gel electrophoresis (Integrity; ++ = high; - = low)	qPCR (ranking)	dPCR (ranking)		
1	Bisulfiteash™ DNA Modification Kit (Epigenetek, P-1026)	Bisulfiteash	95	20	350	15	9	-	11	10	10	Not Assayed
2	Bisulfiteash™ DNA Bisulfite Conversion Easy Kit (Epigenetek, P-1054)	Bisulfiteash Easy	80	45	135.8 <sup>b</sup>	15	10	-	12	12	12	Not Assayed
3	Premium Bisulfite Kit (Diagenode, C02030030)	Premium	98 + 54	8 + 60	350	10	3	-	9	9	9	Not Assayed
4	Imprint® DNA Modification Kit (Sigma-Aldrich, MOD50)	Imprint	99 + 65	6 + 90	350	16	2	+/-	8	7	7	93.2 ± 9.3
5	EZ DNA Methylation-Gold™ Kit (Zymo Research, D5005)	EZ Gold	98 + 64	10 + 150	350	10	1	+/-	6	5	5	99.7 ± 0.1
6	EZ DNA Methylation-Lightning™ Kit (Zymo Research, D5030)	EZ Lightning	98 + 54	8 + 60	350	10	6	+	7	8	8	99.5 ± 0.1
7	Fast Bisulfite Conversion Kit (Abcam®, ab1127127)	Fast	95	20	135.8 <sup>b</sup>	15	5	-	10	11	11	Not Assayed
8	ImmuCONVERT Bisulfite Basic kit (Analytic Jena, 845-IC-1000008)	ImmuCONVERT	85	45	1500	50	4	+	3	6	6	99.4 ± 0.4
9	Epitect® Fast DNA Bisulfite Kit (Qiagen, 59824)	Epitect Fast	95 +60	10 + 20	1000	15	8	+	3	4	4	98.0 ± 2.2
10	Epitect® Bisulfite Kit (Qiagen, 59110)	Epitect	95 +60	15 + 285	1000	20	7	++	1	2	2	98.3 ± 0.7
11	CpGenome™ Turbo Bisulfite Modification Kit (Merck Millipore, S7847)	CpGenome	37 + 70	10 + 40	500	35	11	++	2	1	1	No data
12	MethylEasy™ Xceed (Human Genetic Signatures, ME001)	MethylEasy	37 + 80	15 + 45	2146 <sup>b</sup>	55	12	+	5	3	3	97.7 ± 2.0

<sup>a</sup> If two values are given, the pre-incubation step (denaturation) was at another temperature than the incubation step. The first value indicates the pre-incubation temperature or time, the second value indicates the incubation temperature or time.

<sup>b</sup> Recommended amount of input DNA in the protocol was given in volume. Since the recommended protocol was used, input varied for every donor sample. The input in this table is the average input of the samples from the different donors.

**Table 2: Conversion protocol for different temperatures for Epiect (kit 10).** The different protocols followed a fixed time schedule as provided by the manual from the manufacturer. Temperature protocol 3 is the same as the protocol provided by the manufacturer.

Step	Time (min)	Temperature protocol 1 (°C)	Temperature protocol 2 (°C)	Temperature protocol 3 (°C)	Temperature protocol 4 (°C)
Denaturation1	5	95	95	95	95
Conversion1	25	40	50	60	75
Denaturation2	5	95	95	95	95
Conversion2	85	40	50	60	75
Denaturation3	5	95	95	95	95
Conversion3	175	40	50	60	75

**Table 3: Conversion protocol for different time schedules for Epiect (kit 10).** The different protocols always followed the temperature scheme as provided by the manual from the manufacturer. Time protocol 3 is the same as the protocol provided by the manufacturer.

Step	Temp (°C)	Time protocol 1 (min)	Time protocol 2 (min)	Time protocol 3 (min)	Time protocol 4 (min)
Denaturation1	95	5	5	5	5
Conversion1	60	10	20	25	35
Denaturation2	5	5	5	5	5
Conversion2	60	50	75	85	100
Denaturation3	5	5	5	5	5
Conversion3	60	100	135	175	200

## DNA recovery

After conversion, 9 µl of every sample was diluted 1:2 in nuclease free H<sub>2</sub>O. From this 18 µl, 4 µl was used to determine the recovery after each conversion reaction. The Qubit ssDNA Assay Kit (Q10212, ThermoFisher Scientific) was used on a Qubit 2.0 fluorometer to perform duplicate spectrometric measurements on 2 µl of the diluted single-stranded, bisulfite-converted DNA sample.

## Fragmentation

### *Gel electrophoresis*

Visualization of the fragmentation was performed by gel electrophoresis using 40 ng converted DNA on an E-Gel® EX Agarose Gel, 2% (Invitrogen, G401002). Since the bisulfite treated DNA was single stranded, and the SYBR Safe dye in the gel only binds dsDNA, a five-minute incubation on an ice bath enabled partial hybridization and visualization of the DNA. The electrophoresis experiments were repeated on an Agilent 2100 Bioanalyzer instrument with an Agilent RNA 6000 Pico Kit.

### *qPCR*

To evaluate the amount of intact DNA copies of increasing length after bisulfite treatment, the converted DNA was amplified using qPCR with 6 different primer pairs, resulting in amplicons of increasing length (88 – 476 bp). Two types of primers were used: cytosine free (CF)-primers and cytosine containing (CC)-primers (**Table 4**), which are similar to the Methylation Independent Primers and MethPrimers as previously described by Fuso et al. (383). The CF-primers target genomic regions without C residues, with the same PCR efficiency before and after treatment. The CC-primers target genomic regions that contain non-CpG C residues. In the bisulfite converted samples, CF-primers amplify both converted and unconverted DNA strands, but the CC-primers can only amplify converted DNA strands.

qPCR was performed using the LightCycler® 480 SYBR Green I Master PCR mix (Roche, 04707516001): 2.5 ng of bisulfite converted DNA was added to the qPCR mix containing SYBR Green Mix (2x) and 500 nM forward and reverse primers, in a final volume of 20 µl. PCR amplification reactions consisted of initial denaturation at 95°C for 10 min, followed by 50 cycles (60 cycles for primer pair CCP1) of denaturation at 95°C for 30 sec, annealing at specific temperature (**Table 4**) for 30 sec and elongation at 72°C for 45 sec, and ended by a melting curve analysis from 72°C to 95°C. Every qPCR was performed in triplicate on all five bisulfite treated samples for each kit. The geometric mean of the C<sub>q</sub> values among all replicates, which corresponds with the amount of intact fragments, was used to rank all the kits. If the melting curve analysis was not conclusive, amplicons were analyzed based on length, using a HT DNA 5K chip (CLS760675, Perkin Elmer) on a Labchip® GX (Perkin Elmer) (e.g. **Figure 16** and **Figure 17**) (384).

**Table 4. Primer sequences and annealing temperatures for qPCR and dPCR.** Primers CFF, CFP1 and CFP2 are obtained from Holmes et al. (369); primers CCP1, CCP2 and CCP3 are obtained from Ehrich et al. (379).

Primer pair (amplicon length)	Sequence	Binding region (Ensembl assembly GRCh37)	Annealing temperature qPCR (°C)	Annealing temperature ddPCR (°C)
CFF (88 bp)	F:TAAGAGTAATAATGGATGGATGATG R:CCTCCCATCTCCCTTCC	Chr13 : 19555120 Chr13 : 19555208	58 <sup>a</sup> – 57 <sup>b</sup>	58 <sup>a</sup> – 59 <sup>b</sup>
CFP1 (227 bp)	F:TGGGTAAAGTGATTGAGTAA R:TATTCATCCTTCAACTTACCCT	Chr2 : 21454728 Chr2 : 21454955	52 <sup>b</sup> – 53 <sup>a</sup>	53 <sup>b</sup> – 54 <sup>a</sup>
CFP2 (414 bp)	F:ATGGGTAAGGATATGAAGTTAAT R:TATCACTTAATCACCTCCTAAACTA	Chr2 : 21557568 Chr2 : 21557982	51 <sup>a</sup> – 60 <sup>b</sup>	54 <sup>b</sup> – 56 <sup>a</sup>
CCP1_before (175 bp)	F:GCTGAGGGGCAGAGGGAAGTGC R:GCTTCAGACAGGAAAGTGGCC	Chr11 : 2019526 Chr11 : 2019701	63 – 64	65 – 66
CCP2_before (361 bp)	F:GCTGAGGGGCAGAGGGAAGTGC R:CTCACCAAAGGCCAAGGTGGTGACC	Chr11 : 2019526 Chr11 : 2019887	63	66
CCP3_before (476 bp)	F:TGCACATGGCTGGGGGCCAGCTG R:CCCTCACCAAAGGCCAAGGTGGTGAC	Chr11 : 2019413 Chr11 : 2019889	64	66
CCP1_after (175 bp)	F:GTTGAGGGGTAGAGGGAAGTGT R:ATCTTCAAACAAAAAATAACC	Chr11 : 2019526 Chr11 : 2019701	60	/
CCP2_after (361 bp)	F:GTTGAGGGGTAGAGGGAAGTGT R:CTCACCAAACCAAATAATAACC	Chr11 : 2019526 Chr11 : 2019887	62	/
CCP3_after (476 bp)	F:TGTATATGGTTGGGGTTAGTTG R:CCCTCACCAAACCAAATAATAAC	Chr11 : 2019413 Chr11 : 2019889	63	/

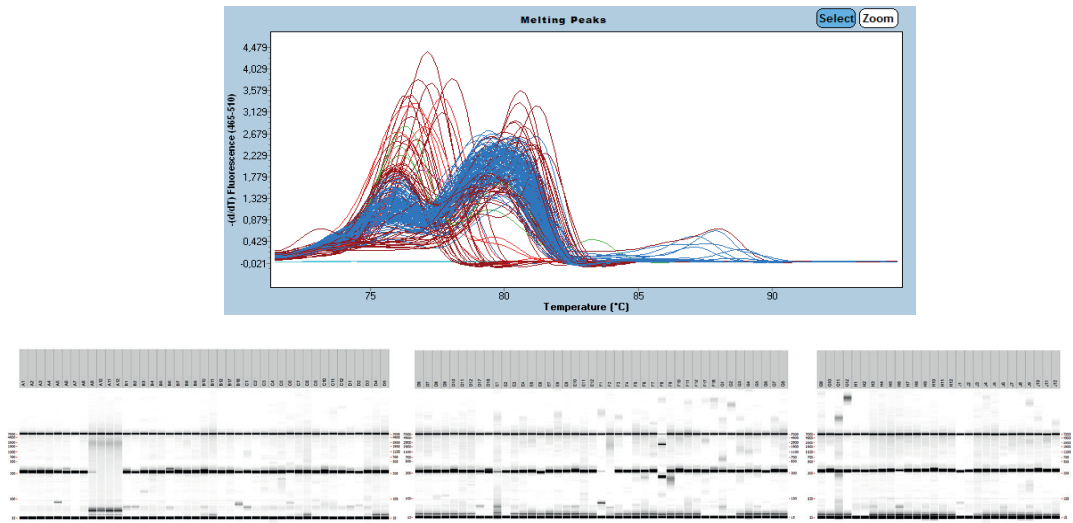
<sup>a</sup> Optimal annealing temperature for DNA before bisulfite treatment<sup>b</sup> Optimal annealing temperature for DNA after bisulfite treatment

F: Forward primer

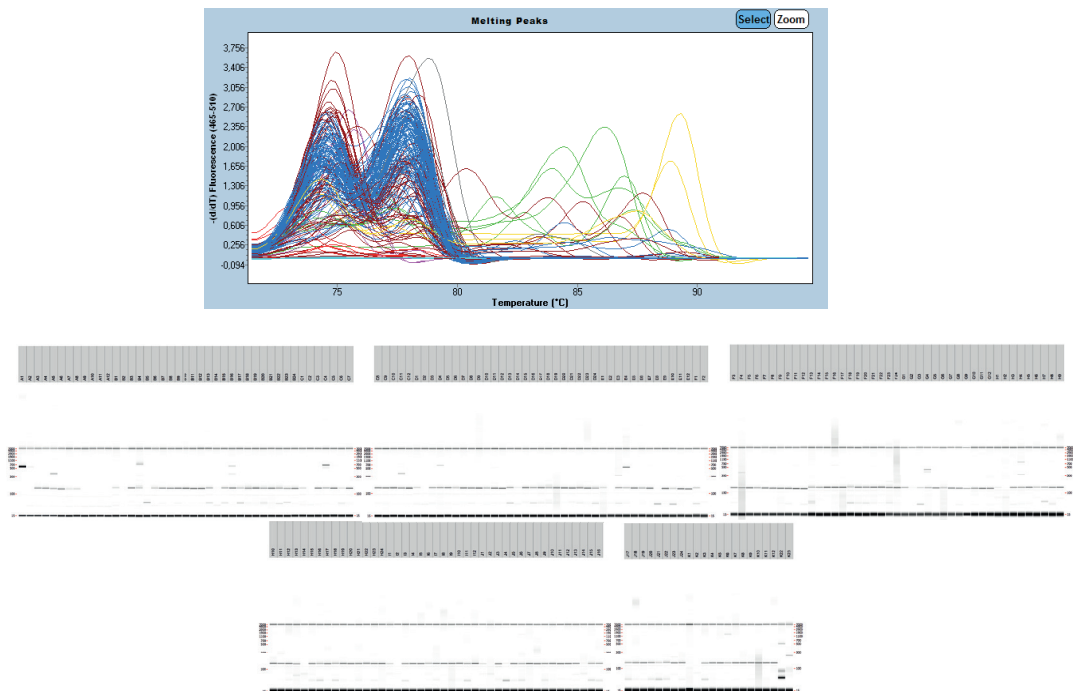
R: Reverse primer

\_before: primer targeting genomic DNA before bisulfite treatment

\_after: primer targeting bisulfite converted DNA



**Figure 16. Melting curves and electrophoresis plots of amplicons CCP1\_after.** qPCR analysis did not always show conclusive melting curves (upper panel). In these cases, the amplicons were analyzed to measure the amplicon length using Caliper LabChip GX results (lower panel) after qPCR. The Caliper only shows specific bands with the correct length for the reaction CCP1\_after (175 bp).



**Figure 17. Melting curves and electrophoresis plots of amplicons CCP2\_after.** qPCR analysis did not always show conclusive melting curves (upper panel). In these cases, the amplicons were analyzed to measure the amplicon length using Caliper LabChip GX results (lower panel) after qPCR. The Caliper only shows specific bands with the correct length for the reaction CCP2\_after (361 bp).

## *Digital PCR*

To eliminate the potential inhibitory effect of elution buffers or impurities on the qPCR reaction, fragmentation was also analyzed using dPCR. Because of the 'digital' end-point measurement in dPCR, effects of PCR efficiencies are eliminated in the quantification. Moreover, with dPCR, it is possible to perform a direct absolute quantification of intact DNA copies in the PCR mix (385,386). Similar as in the qPCR, several CF primers resulting in amplicons of different lengths (88 to 414 bp), were used to investigate the difference in fragmentation between the twelve kits.

2  $\mu$ l DNA of all 60 bisulfite treated samples (1:2 diluted) was added to the dPCR mix containing QX200 ddPCR EvaGreen Supermix (2x) (186-4033, Bio-Rad) with 100 nM forward and reverse primers, in a final volume of 20  $\mu$ l. PCR reactions consisted of initial denaturation at 95°C for 5 min, followed by 40 cycles of denaturation at 95°C for 30 sec and annealing/elongation at specific temperature (**Table 4**) for 2 min, and ended with a signal stabilization at 4°C for 5 min and 95°C for 5 min. Each sample was analyzed in duplicate. Moreover, 8.65  $\mu$ l of untreated DNA samples from all donors was restricted with EcoRI enzyme (Promega, R6011) according to the manufacturer's instructions. 2  $\mu$ l of this restricted DNA was also used for quantification using the same dPCR mix as the treated DNA samples. Based on the DNA concentration (absolute amount of intact copies per ng bisulfite treated input DNA), a ranking of the 12 kits was made for each primer pair.

## **Conversion efficiency**

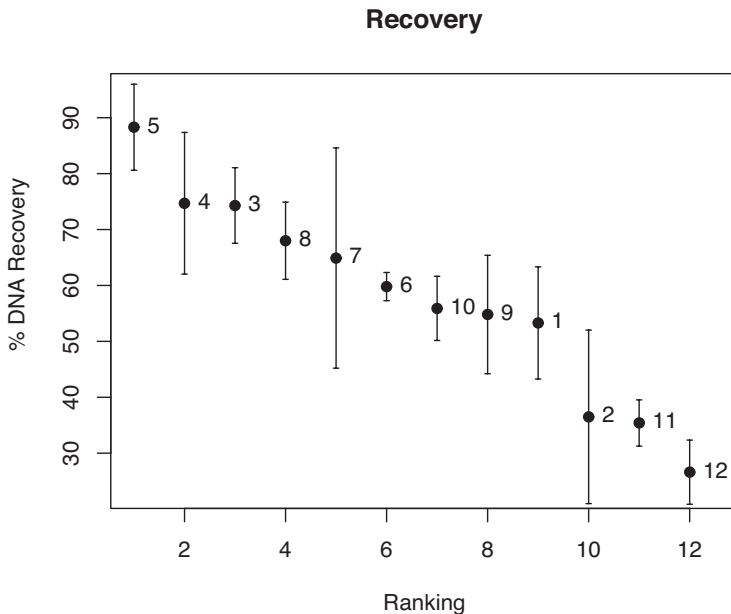
The conversion efficiency was calculated as the amount of converted to non-converted non-CpG Cs, which are usually not methylated in human PBMCs (51–54,387,388). Almost all these Cs should thus be converted during bisulfite treatment, providing us with a method to calculate the conversion efficiency.

Based on the results from the DNA recovery and the fragmentation analysis, we selected the eight least fragmenting kits for subsequent sequencing. We used qPCR amplicons of donor 2 generated with primer pairs CCP3 (149 Cs in amplicon, of which 32 CpG) and CFP2 (62 Cs in amplicon, of which 2 CpG) for subsequent sequencing. The second primer pair can be used to assess the overall conversion efficiency, while analysis with CCP3 can potentially result in an overestimation of this efficiency. To assess the variability between the different donors, converted DNA samples of a second donor (donor 3) obtained by three of the kits were also analyzed. The amplicons were purified with the High Pure PCR Product Purification Kit (11732668001, Roche) and sequenced in a MiSeq sequencing system (Illumina). The sequencing reads were aligned using the Bismark package (version 0.10.1), and further analysis for the conversion efficiency was done by the MethylKit package (version 0.9.5) in R.

## Results

### Evaluation of DNA recovery

The absolute concentration of the DNA samples was measured on a Qubit 2.0 fluorometer before bisulfite treatment (dsDNA assay) and after bisulfite treatment (ssDNA assay). The starting DNA concentration was on average 136 ng/μl. The DNA recovery ranged from 26.6% (Methyleasy (kit 12)) to 88.3% (EZ Gold (kit 5)) of the maximum theoretical concentration based on the amount of input DNA (**Figure 18**, **Table 5** and **Table 6**). Of note: different quantification kits (Qubit dsDNA and ssDNA assays) were used before and after bisulfite treatment. The quantification of these two assays was compared by measuring the DNA concentration before and after heat denaturation of a DNA sample. This indicated a similar quantification for both assays (data not shown). Therefore, only a relative comparison of the total DNA loss during the bisulfite conversion of the different kits can be made by this analysis.



**Figure 18. DNA recovery of the twelve bisulfite kits.** DNA recovery is shown as percentages  $\pm$  SD and is calculated by the ratio of the measured output concentration of each bisulfite kit (Qubit ssDNA) to the maximal theoretical output concentration based on the input in every kit (Qubit dsDNA). The labels in the figure refer to the kit numbers in **Table 1**.

**Table 5. Concentration of the DNA samples before bisulfite treatment.** The concentration of the untreated DNA samples obtained from PBMCs of five donors before bisulfite treatment measured by Qubit dsDNA BR (broad range) Assay Kit.

Donor	Concentration (ng/μl)
1	100
2	116
3	160
4	148
5	158

**Table 6. Amount of input DNA for bisulfite treatment and concentration of the DNA samples after bisulfite treatment.** Two aliquots of all five DNA samples were treated two independent times, yielding in ten bisulfite treated samples, and subsequently, the duplicate samples were pooled. The quantification measurements are done in duplicate with the Qubit ssDNA Assay kit, and the data shown are averages  $\pm$  SD of these ten concentrations.

Kit	DNA input (ng)	Elution volume (μl)	Maximal theoretical concentration (ng/μl)	Concentration (ng/μl $\pm$ SD)	Recovery (% $\pm$ SD)	Ranking
Bisulflash	350	15	23.3	12.2 $\pm$ 2.4	53.3 $\pm$ 10.1	<b>9</b>
Bisulflash Easy	135.8 <sup>a</sup>	15	9.1	3.22 $\pm$ 1.41	36.5 $\pm$ 15.6	<b>10</b>
Premium	350	10	35.0	26.0 $\pm$ 2.4	74.3 $\pm$ 6.8	<b>3</b>
Imprint	350	16	21.9	16.3 $\pm$ 2.8	74.7 $\pm$ 12.7	<b>2</b>
EZ Gold	350	10	35.0	30.9 $\pm$ 2.7	88.3 $\pm$ 7.7	<b>1</b>
EZ Lightning	350	10	35.0	20.9 $\pm$ 0.9	59.8 $\pm$ 2.6	<b>6</b>
Fast	135.8 <sup>a</sup>	15	9.1	5.88 $\pm$ 1.78	64.9 $\pm$ 19.7	<b>5</b>
InnuCONVERT	1500	50	30.0	20.4 $\pm$ 2.1	68.0 $\pm$ 7.0	<b>4</b>
Epitect Fast	1000	15	66.7	36.6 $\pm$ 7.1	54.8 $\pm$ 10.6	<b>8</b>
Epitect	1000	20	50.0	28.0 $\pm$ 2.9	55.9 $\pm$ 5.8	<b>7</b>
CpGenome	500	35	14.3	5.06 $\pm$ 0.59	35.4 $\pm$ 4.2	<b>11</b>
Methyleasy	2146 <sup>a</sup>	55	39.0	10.4 $\pm$ 2.3	26.6 $\pm$ 5.8	<b>12</b>

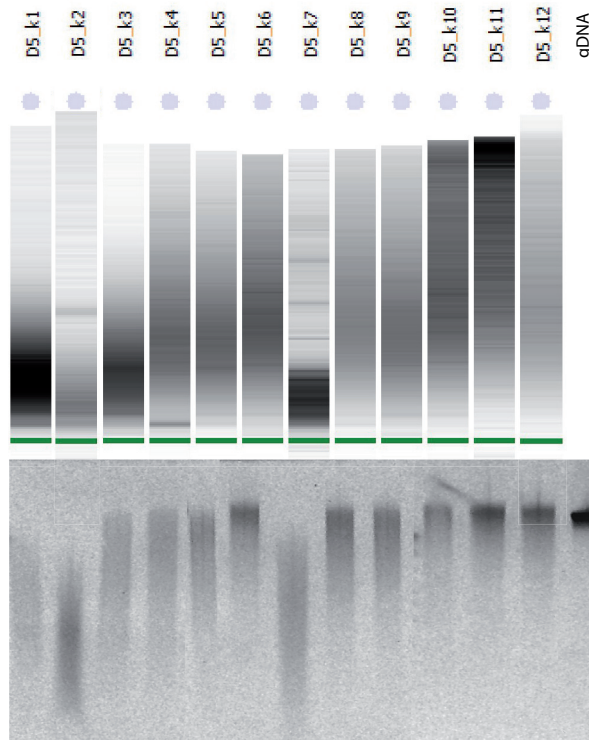
<sup>a</sup> Recommended DNA input was given in volume DNA sample. Values are the mean of the actual input



## Evaluation of fragmentation

### *Gel electrophoresis*

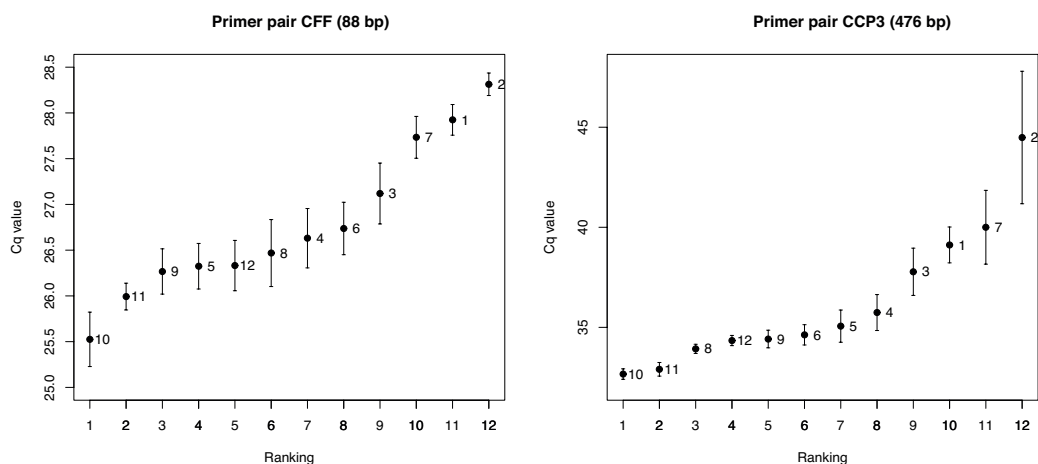
DNA fragmentation by the different bisulfite kits was visualized by gel electrophoresis and Bioanalyzer analysis of the bisulfite treated DNA samples. These methods do not provide quantifiable data about the fragmentation, but provide a rough estimate to compare the best and the worst kits. The untreated gDNA was used as a control showing a clear band of minimally fragmented DNA after the DNA isolation procedure. A smear of larger fragments was detected for the Epitect (kit 10) and CpGenome (kit 11) kits, indicating that these are the least fragmenting kits (**Table 1, Figure 19**).



**Figure 19.** *Bioanalyzer (upper plot) and gel electrophoresis (lower plot) analysis of the different kits. The plots show the electrophoresis data of one of the five donor samples (donor five). gDNA was analyzed along the samples using gel electrophoresis. D5: donor 5; k: kit.*

## qPCR

The comparison of fragmentation of the kits by qPCR included different primer pairs which were developed to amplify amplicons with lengths varying from 88 base pairs (bp) to 476 bp. Overall, the smallest Cq values were found in the Epitect (kit 10) and CpGenome (kit 11) indicating that the concentration of intact DNA strands of the analyzed length is the highest in these kits (**Figure 20**).



**Figure 20.** Cq values  $\pm$  SD of the smallest and the largest amplicons. The data used are the geometric means of the average values from the five donor samples as shown in **Table 7**. The data labels refer to the kit number as provided in **Table 1**.

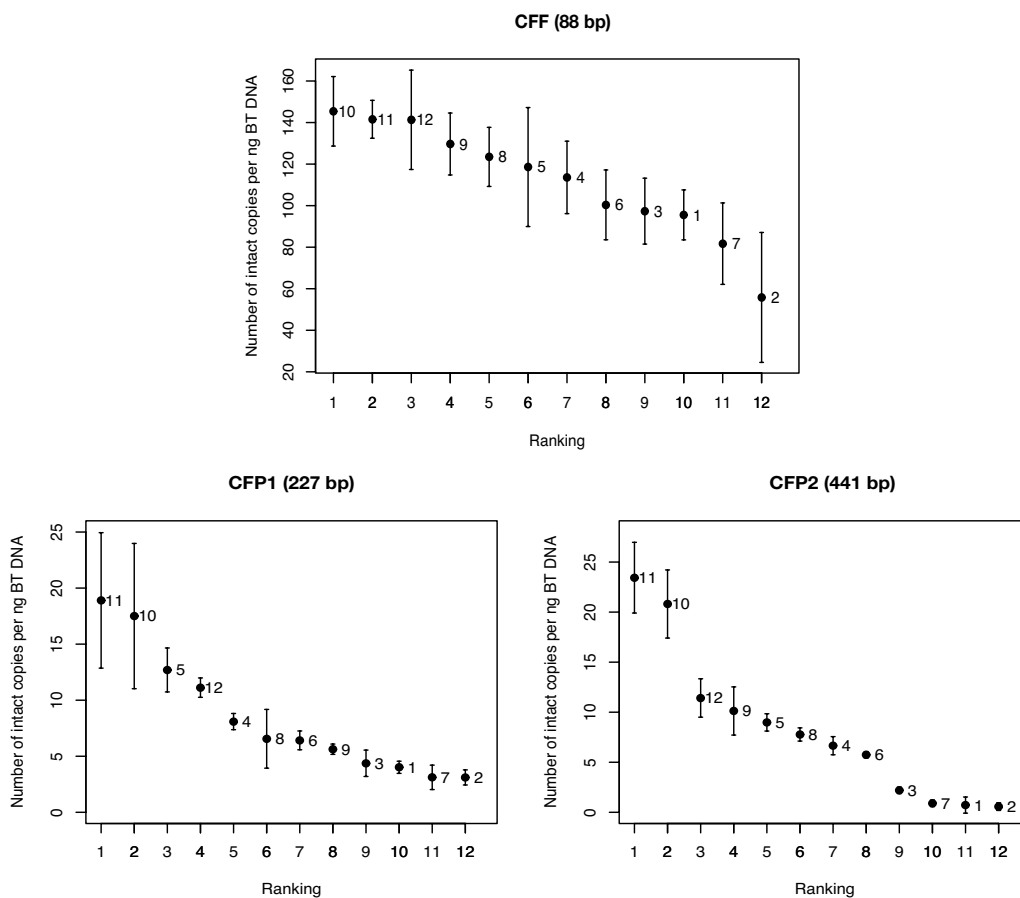
**Table 7: Cq values and ranking of the qPCR experiments from the six used primer pairs.** First the average of the three technical replicates was calculated for all five samples. The data given is the geometric mean of the average values from the five donor samples. Genomic DNA is the measurement of untreated DNA, which is only conducted for the cytosine free primers since they have the same efficiency before and after treatment. The different kits are ranked by the Cq values for every primer pair (Rank in the table). Subsequently, the median of these rankings is calculated to assess a final ranking. This is the ranking given in **Table 1**.

Kit	CFP (Cq ± SD)	Rank CFP	CFP1 (Cq ± SD)	Rank CFP1	CFP2 (Cq ± SD)	Rank CFP2	CCP1 (Cq ± SD)	Rank CCP1	CCP2 (Cq ± SD)	Rank CCP2	CCP3 (Cq ± SD)	Rank CCP3	Median Rank	Final Rank
<b>Bisuiflash</b>	27.92 ± 0.20	11	29.69 ± 0.12	11	37.60 ± 0.83	10	50.24 ± 1.66	11	38.68 ± 1.37	11	39.06 ± 2.30	10	11	11
<b>Bisuiflash Easy</b>	28.31 ± 0.23	12	30.07 ± 0.11	12	48.96 ± 2.24	12	57.19 ± 3.90	12	39.25 ± 1.43	12	44.40 ± 3.04	12	12	12
<b>Premium</b>	27.12 ± 0.13	9	28.75 ± 0.18	9	34.39 ± 0.40	8	43.35 ± 1.57	8	34.47 ± 0.47	9	37.77 ± 0.62	9	9	9
<b>Imprint</b>	26.63 ± 0.23	7	28.12 ± 0.21	8	34.23 ± 2.56	7	43.70 ± 2.81	9	32.72 ± 0.32	8	35.73 ± 0.70	8	8	8
<b>EZ Gold</b>	26.32 ± 0.25	4	28.00 ± 0.27	6	32.16 ± 0.24	2	40.85 ± 1.11	4	32.37 ± 0.39	7	35.05 ± 0.95	7	5	6
<b>EZ Lightning</b>	26.74 ± 0.24	8	28.08 ± 0.21	7	33.55 ± 1.21	5	41.30 ± 1.00	7	31.56 ± 0.48	5	34.63 ± 0.16	6	6	7
<b>Fast</b>	27.73 ± 0.16	10	29.43 ± 0.21	10	38.16 ± 0.61	11	47.31 ± 1.17	10	38.18 ± 0.99	10	39.98 ± 1.61	11	10	10
<b>InnuCONVERT</b>	26.47 ± 0.18	6	27.89 ± 0.13	5	32.81 ± 0.48	3	41.01 ± 0.95	6	31.24 ± 0.21	3	33.93 ± 0.25	3	4	3*
<b>EpiTECT Fast</b>	26.27 ± 0.29	3	27.85 ± 0.22	4	33.40 ± 1.82	4	40.87 ± 1.17	5	31.55 ± 0.41	4	34.42 ± 0.46	4	4	3*
<b>EpiTECT</b>	25.52 ± 0.39	1	27.14 ± 0.26	2	31.28 ± 1.09	1	39.68 ± 1.61	1	30.40 ± 0.39	1	32.67 ± 0.21	1	1	1
<b>CpGenome</b>	25.99 ± 0.13	2	26.81 ± 0.26	1	36.16 ± 0.67	9	39.72 ± 0.73	2	30.80 ± 0.33	2	32.90 ± 0.13	2	2	2
<b>MethylEasy</b>	26.33 ± 0.31	5	27.56 ± 0.20	3	33.89 ± 1.69	6	39.94 ± 1.50	3	31.71 ± 0.38	6	34.34 ± 0.34	4	4.5	5
<b>Genomic DNA</b>	20.74 ± 0.17		21.79 ± 0.13		23.93 ± 0.58		NA		NA		NA			

\*The same overall ranking was given if two kits obtained the same median of rankings.

## Digital PCR

To test whether digital PCR (dPCR) can be used to assess both DNA recovery as well as DNA fragmentation we used several primers pairs resulting in amplicons of different lengths (88 to 414 bp) to investigate the DNA recovery and the difference in fragmentation between the twelve kits with dPCR. Since dPCR provides direct absolute quantification, the exact amount of DNA strands lost during conversion is measured. Overall, the highest DNA concentrations were found in the Epitect (kit 10) and CpGenome (kit 11): respectively 143.7 and 141.1 intact DNA strands of 88 nt; 16.5 and 14.1 intact DNA strands of 227 nt and 20.8 and 23.3 intact DNA strands of 414 nt per ng input DNA (**Figure 21**, **Table 1** and **Table 8**).



**Figure 21.** Visual representation of the rankings of the dPCR experiments. Results are given as the geometric means of the average number of intact copies per ng bisulfite treated (BT) DNA measured by dPCR  $\pm$  SD of all the five donor samples as shown in **Table 8**. The data labels refer to the kit number as provided in **Table 1**.

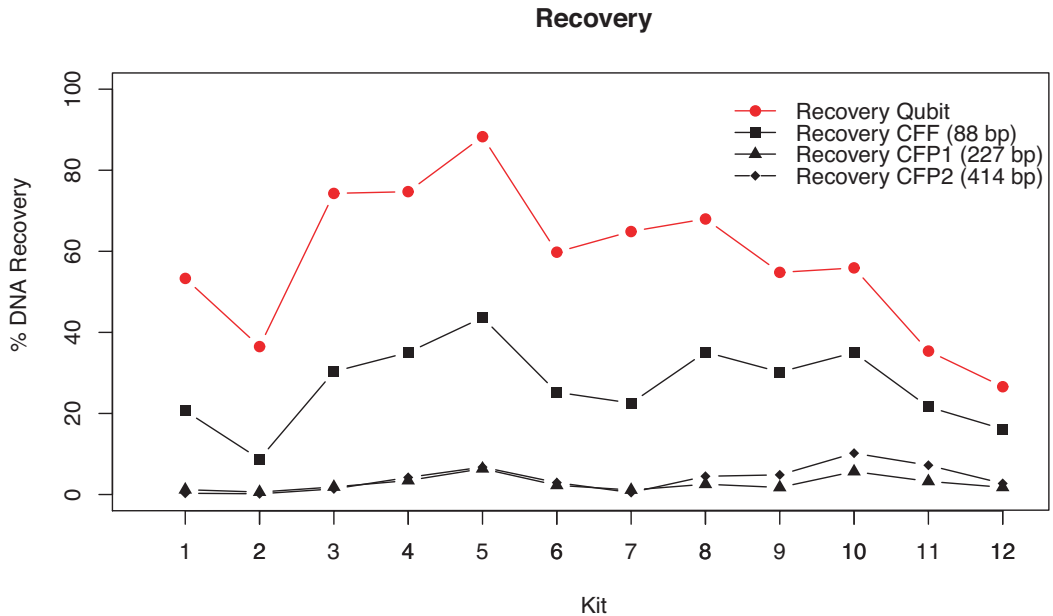
**Table 8: Results and ranking of the dPCR experiments from the three primer pairs that were used.** First, the average of the two technical replicates was calculated and normalized to the DNA input for all five samples. The data given is the geometric mean of the average values from all five donor samples. The different kits are ranked by the amount of copies per ng bisulfite treated DNA for every primer pair (Rank in the table). Subsequently, the median of these rankings is calculated to assess a final ranking. This is the final ranking given in Table 1.

Kit	CFF (copies per ng DNA $\pm$ SD)	Rank	CFP1 (copies per ng DNA $\pm$ SD)	Rank	CFP2 (copies per ng DNA $\pm$ SD)	Rank	Median Rank	Final Rank
Bisulfash	92.5 $\pm$ 12.1	10	3.4 $\pm$ 0.7	10	0.7 $\pm$ 0.9	11	10	<b>10</b>
Bisulfash Easy	51.7 $\pm$ 31.3	12	3.0 $\pm$ 0.7	12	0.6 $\pm$ 0.4	12	12	<b>12</b>
Premium	96.5 $\pm$ 15.9	9	4.1 $\pm$ 1.3	9	2.2 $\pm$ 0.4	9	9	<b>9</b>
Imprint	112.7 $\pm$ 17.5	7	7.4 $\pm$ 0.9	5	6.5 $\pm$ 1.0	7	7	<b>7</b>
EZ Gold	117.1 $\pm$ 28.7	6	12.2 $\pm$ 3.0	3	9.0 $\pm$ 0.9	5	5	<b>5</b>
EZ Lightning	98.6 $\pm$ 16.8	8	5.5 $\pm$ 0.9	7	5.7 $\pm$ 0.4	8	8	<b>8</b>
Fast	80.3 $\pm$ 19.7	11	3.0 $\pm$ 1.4	11	0.8 $\pm$ 0.4	10	11	<b>11</b>
InnuCONVERT	122.7 $\pm$ 14.3	5	5.8 $\pm$ 2.7	6	7.6 $\pm$ 0.7	6	6	<b>6</b>
Epitect Fast	129.0 $\pm$ 15.0	4	5.1 $\pm$ 0.7	8	10.0 $\pm$ 2.5	4	4	<b>4</b>
Epitect	143.7 $\pm$ 16.8	1	16.5 $\pm$ 7.0	2	20.8 $\pm$ 3.4	2	2	<b>2</b>
CpGenome	141.1 $\pm$ 9.2	2	17.8 $\pm$ 6.7	1	23.3 $\pm$ 3.6	1	1	<b>1</b>
MethylEasy	138.8 $\pm$ 24.0	3	10.6 $\pm$ 1.3	4	11.3 $\pm$ 2.0	3	3	<b>3</b>

Absolute quantification by dPCR before and after treatment showed a significant decrease in number of intact copies in all kits when the length increased from 88 to 227 nt (on average 72.4% DNA lost vs 97.3% DNA lost, Wilcoxon rank sum test, p-value = 7.40e-07), but there was no significant effect of the increase in length from 227 to 414 nt (on average 97.3% DNA lost vs. 96.1% DNA lost, Wilcoxon rank sum test, p-value = 0.478) (**Table 9**). Between the different bisulfite kits, the Epitect (kit 10) retained 10.2% of the input DNA with length of 414 nt, while the Bisulflash easy (kit 2) retained only 0.2%. The same trend of DNA recovery (on average over all the kits) was found with the Qubit analysis (**Figure 22**).

**Table 9. The percentages of overall DNA loss in the samples.** The DNA loss is assessed by dPCR before and after bisulfite treatment. This dPCR measures the intact copies of specific lengths present in the samples. To obtain the averages given in the table, the average of the two technical replicates was calculated both before and after bisulfite treatment. These averages were used to calculate the average loss per kit of all the five donor samples. Subsequently, the geometric means and standard deviations from these averages were calculated.

Kit	Average loss intact copies CFF (88 bp) (% ± SD)	Average loss intact copies CFP1 (227 bp) (% ± SD)	Average loss intact copies CFP2 (414 bp) (% ± SD)
Bisulflash	79.4 ± 5.2	98.8 ± 0.9	99.7 ± 0.1
Bisulflash Easy	91.2 ± 4.4	99.4 ± 0.2	99.8 ± 0.2
Premium	69.7 ± 4.8	98.1 ± 1.0	98.6 ± 0.3
Imprint	65.0 ± 4.1	96.6 ± 1.8	95.8 ± 1.2
EZ Gold	56.3 ± 5.8	93.6 ± 2.3	93.2 ± 1.2
EZ Lightning	74.8 ± 5.3	97.7 ± 1.6	97.1 ± 0.6
Fast	77.5 ± 4.9	98.9 ± 0.3	99.5 ± 0.2
InnuCONVERT	64.9 ± 4.7	97.5 ± 1.6	95.5 ± 1.4
Epitect Fast	69.8 ± 2.5	98.3 ± 0.8	95.2 ± 1.3
Epitect	65.0 ± 5.2	94.3 ± 2.1	89.8 ± 1.3
CpGenome	78.4 ± 2.5	96.8 ± 2.2	92.8 ± 0.6
Methyleasy	83.9 ± 4.0	98.2 ± 0.9	97.3 ± 1.1
Average loss per primer pair	<b>72.4 ± 9.7</b>	<b>97.3 ± 1.8</b>	<b>96.1 ± 3.2</b>



**Figure 22. Comparison of different recovery analyzes (red: Qubit vs black: dPCR).** This recovery is based on the amount of overall DNA loss in the samples as shown in **Table 9**.

## Evaluation of conversion efficiency and inappropriate conversion

Since correct estimation of the methylation in the samples is the main goal after bisulfite treatment, both appropriate conversion (i.e. conversion of C to U) and inappropriate conversion (i.e. conversion of mC to thymine (T)) are crucial characteristics of a bisulfite kit. Therefore, all kits were excluded from the comparison if the appropriate conversion was lower than 95%. This conversion efficiency of the different bisulfite kits was analyzed by the amount of unconverted Cs situated outside CpGs: in principle, 100% conversion of these Cs is expected, since non-CpG Cs are seldom (<0.02%) methylated in human PBMCs (51–54,387,388). Bisulfite treated samples from one donor obtained with eight kits with best scores in the evaluation of the fragmentation and recovery were analyzed (**Table 10**). Two kits (Imprint (kit 4) and CpGenome (kit 11)) showed lower conversion efficiencies for the non-CpG Cs in amplicon CFP2 (62 Cs, 2 CpG) compared to amplicon CCP3 (159 Cs, 32 CpG), indicating that the bisulfite conversion was incomplete. However, the coverage of the sequencing reaction of CpGenome (kit 11) was low (only 14 reads) (**Table 10**). Consequently, these kits were excluded from the final ranking. Moreover, a relative comparison of the inappropriate conversion was made based on the methylation percentages obtained with the 34 CpGs (CFP2 + CCP3) analyzed. Gold (kit 5) showed highest methylation (56.2% methylation), indicating that this kit shows the least inappropriate conversion, followed by Epitect (kit 10) (46.4% methylation) and Imprint (kit 4) (42.5% methylation) (**Table 10**). Of note: as a control, the variability of conversion efficiency between different donors was analyzed. Therefore, the conversion efficiency of samples from a second donor obtained

with three kits (innuCONVERT (kit 8), Epitect (kit 10) and Methyleasy (kit 12)) were analyzed and a low variability was observed: standard deviation between 0.01 and 2.42 % (data not shown).

**Table 10. Conversion efficiencies and overall methylation percentage for the different kits.** The data is obtained by sequencing of the 8 kits that performed best in the previous fragmentation assessments. One out of five donor samples was used, and two separate PCR products were sequenced: amplicons CFP2 and CCP3. CFP2 (414 bp) counts 62 Cs, of which 2 CpGs; CCP3 (476 bp) counts 159 Cs, of which 32 CpGs.

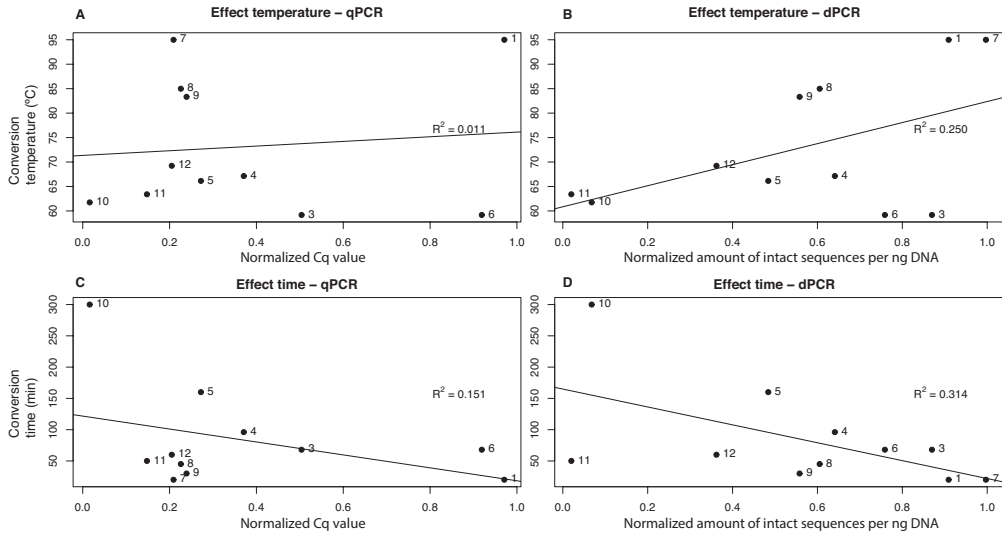
Kit	Conversion efficiency			Overall methylation			Coverage	
	CFP2 (%)	CCP3_after (%)	Mean (% ± SD)	CFP2 (%)	CCP3_after (%)	CFP2 + CCP3_after (%)	CFP2	CCP3_after
Imprint	86.6	99.7	93.2 ± 9.3	92.6	39.4	42.5	678	117
EZ Gold	99.7	99.6	99.7 ± 0.1	90.7	54.0	56.2	464	111
EZ Lightning	99.5	99.5	99.5 ± 0.1	93.2	26.0	30.0	659	126
InnuCONVERT	99.2	99.7	99.4 ± 0.4	92.1	38.8	41.9	482	123
Epitect Fast	96.5	99.5	98.0 ± 2.2	91.2	32.3	35.8	952	118
Epitect	98.1	98.5	98.3 ± 0.7	95.1	43.4	46.4	314	42
CpGenome	52.1	99.8	75.9 ± 33.8	100	14.3	19.6	14	14
Methyleasy	96.0	99.4	97.7 ± 2.0	91.3	31.5	35.0	288	62

\_after: primer targeting bisulfite converted DNA

## Evaluation of effect of conversion time and temperature

To analyze the effect of increasing temperature and time of the bisulfite conversion on the fragmentation, the degree of fragmentation of the different kits was linked with the time and conversion temperature of each kit (**Figure 23**). The kits were ranked by fragmentation performance based on the qPCR and dPCR results. No significant correlation between fragmentation and time or temperature was found (Spearman's rank correlation p-values: 0.521 (qPCR – temperature), 0.155 (dPCR – temperature), 0.374 (qPCR – time) and 0.079 (dPCR – time)). To exclusively analyze the effect of conversion temperature and time within a single kit, the fragmentation was compared by changing the conversion temperature and time in Epitect (kit 10). We found no significant difference in fragmentation by these alterations (**Table 2**, **Table 3** and **Table 11** and **Figure 24** and **Figure 25**).

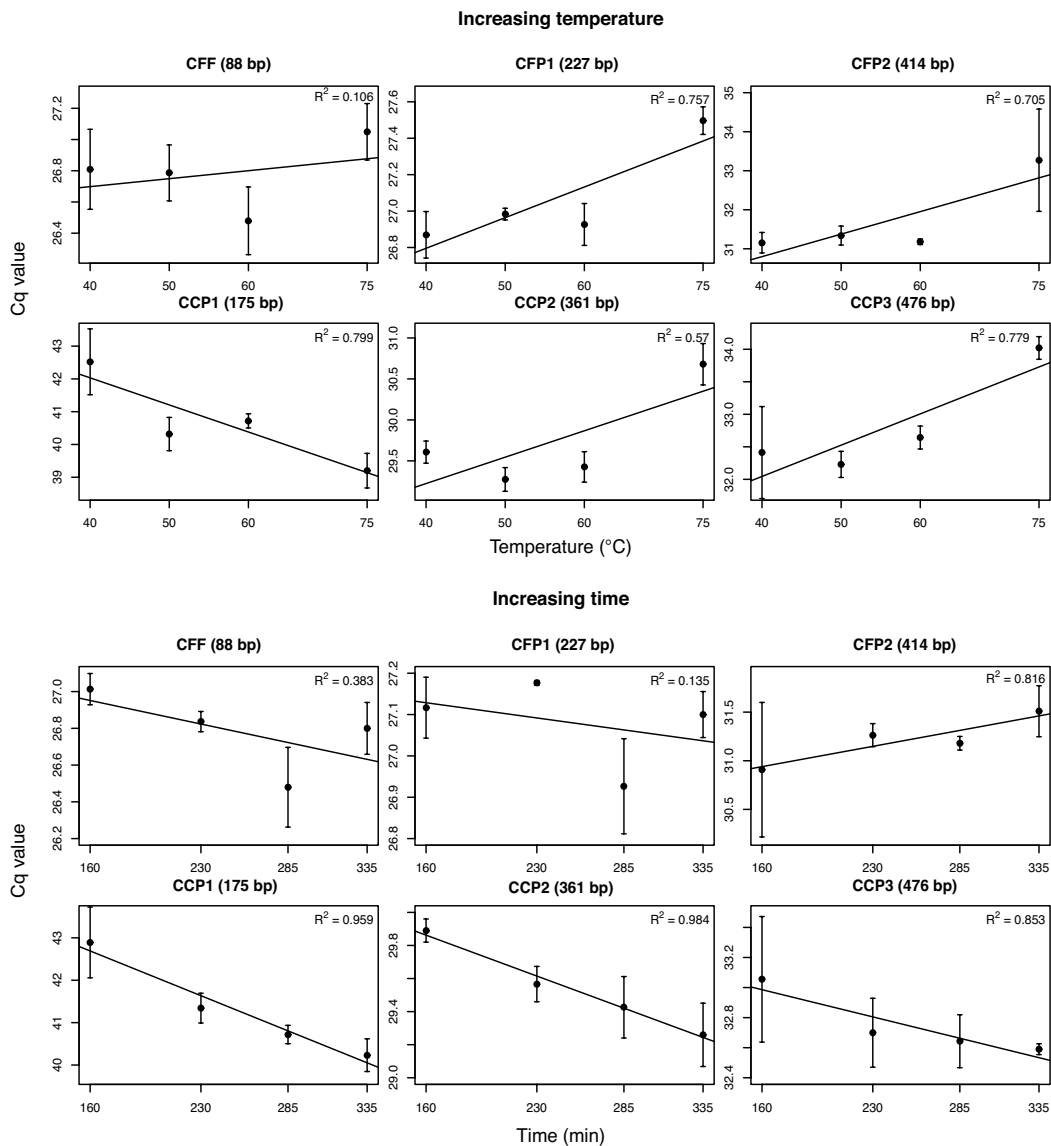




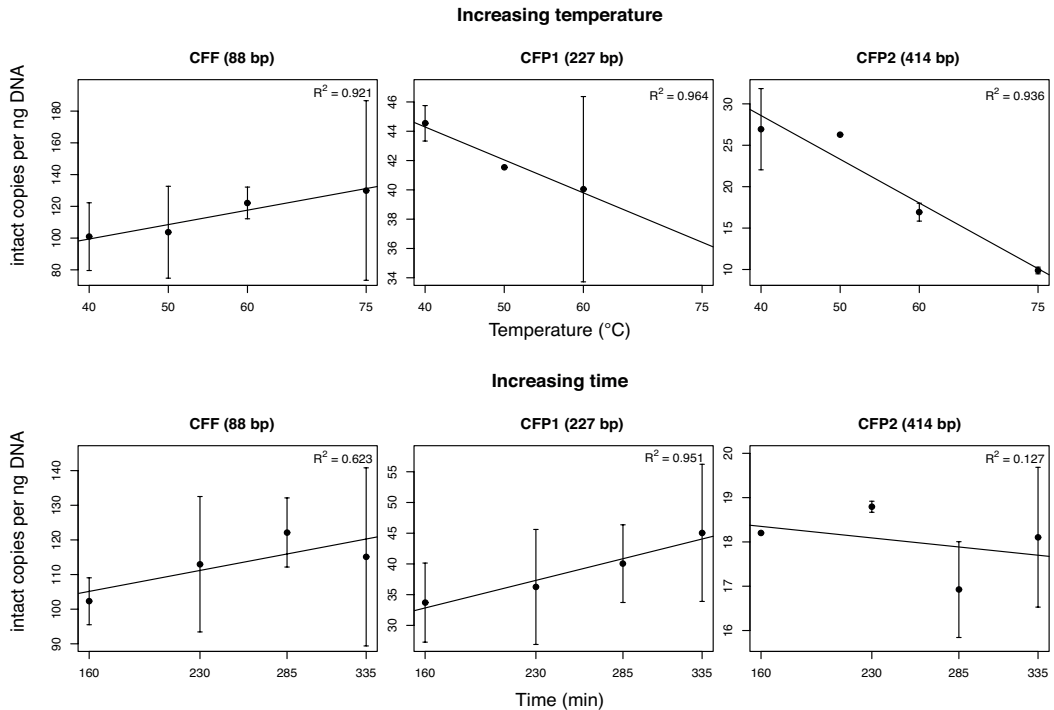
**Figure 23. Effect of time and temperature.** Effect of time and temperature of the conversion protocol on fragmentation between the different bisulfite conversion kits. Labels refer to the kit numbers in Table 1. The upper panels (A-B) show the effect of different conversion temperatures between the kits measured by qPCR (A) and dPCR (B). The lower panels (C-D) show the effect of different conversion times between the kits measured by qPCR (C) and dPCR (D). The values on the x-axis are normalized to show the relative amount of fragmentation: 0 is least fragmenting and 1 is most fragmenting.

**Table 11. Concentration after elution with Epitect (kit 10) with the different time and temperature protocols as depicted in Table 2 and Table 3.** In the end of the protocol, 2 elutions of the same sample were performed (as recommended by the manufacturer to maximize DNA yield). The data given is a single measurement of the donor used in these time and temperature experiments.

Protocol	Elution1 (ng/μl)	Elution2 (ng/μl)	% Recovery (elution1)
Time 1 (160')	33.4	9.48	66.8
Time 2 (230')	34.0	8.56	68.0
Time 3 (285')	34.0	10.6	68.0
Time 4 (335')	24.2	14.3	48.4
Temperature 1 (40°C)	22.2	8.66	44.4
Temperature 2 (50°C)	29.8	7.84	59.6
Temperature 3 (60°C)	34.0	10.6	68.0
Temperature 4 (75°C)	35.0	6.20	70.0



**Figure 24. qPCR analysis of the alternative protocols from the Epitect Bisulfite kit as shown in Table 2 and Table 3. Upper panel: protocol changed in conversion temperature (Table 2). Lower panel: protocol changed in conversion time (Table 3). Results are given as Cq values  $\pm$  SD.**



**Figure 25. dPCR analysis of the alternative protocols from the Epitect Bisulfite kit as shown in Table 2 and Table 3. Upper panel: protocol changed in conversion temperature (Table 2). Lower panel: protocol changed in conversion time (Table 3). Results are given as number of intact copies per ng bisulfite treated DNA measured by dPCR  $\pm$  SD.**

## Discussion

The importance of DNA methylation in different biological processes and the need for an easy-to-use technology have resulted in a wide range of commercially available bisulfite conversion kits. To select the most appropriate bisulfite conversion kit for specific applications, several parameters have to be taken into account. In the current study, we provide a comprehensive workflow for analyzing bisulfite kit performance by comparing twelve bisulfite kits based on DNA recovery and DNA fragmentation. Since dPCR provides a direct absolute quantification with minimal influence of variations in PCR efficiency (385), we can use one method to directly compare samples before and after bisulfite conversion, enabling us to investigate both DNA recovery and fragmentation with more ease and higher accuracy compared to classical methods. The present work shows the strength of dPCR for analysis of the quality of bisulfite treated DNA.

The most essential parameter of a bisulfite kit is the conversion rate, for which appropriate conversion should be maximized, and inappropriate conversion minimized. Our results showed slightly higher estimation of the conversion rates for the cytosine containing (CC) primers (or MethPrimers as called by Fuso et al. (383)) than for the cytosine free (CF) primers (or Methylation Insensitive Primers

(MIT) as in the manuscript of Fuso et al. (383)) in seven of the eight samples, which was expected since the reactions with the CF primers amplify both converted as unconverted DNA strands. Since non-CpG methylation is negligible in human PBMCs, it is very unlikely that this bias is similar to the bias described in Fuso et al. (383) which is due to methylated non-CpG Cs. During our comparison, CpGenome (kit 11) and Imprint (kit 4) showed a large difference between the two primer types, indicating a low overall conversion efficiency, and were therefore excluded from the comparison. However, CpGenome (kit 11) had no reliable coverage in the sequencing reaction, for which it also had to be excluded. In other similar studies performed by Holmes et al. (369), Leontiou et al. (380), Izzi et al. (381) and Bryzgunova et al. (382), the conversion efficiency was >95% for all the analyzed kits.

In the current comparison, we showed clear differences between the DNA recovery of the bisulfite kits. This recovery is measured as the overall loss of DNA during the bisulfite treatment, including DNA loss and fragmentation. Remarkably, the best recovery was found with EZ Gold (kit 5), while EZ Lightning (kit 6) was ranked on the sixth place. Since both kits appear to use the same DNA clean up procedure, the conversion and desulphonation buffers and conditions may have impacted the DNA recovery. Our ranking of the recovery of the kits is in accordance with previous comparative studies (369,380–382). However, in these studies, DNA recovery and fragmentation was never evaluated in depth: Leontiou et al. (380) evaluated both DNA fragmentation and recovery by only measuring DNA concentration via spectroscopic measurements, assuming that fragmentation leads to degraded DNA to such extent that it cannot be analyzed by spectroscopic means. Consequently, our ranking for fragmentation is completely different. Moreover, our findings show that their assumption may not hold for all kits as we observed that kits with low fragmentation can have low recoveries (e.g. CpGenome (kit 11) and Methyleasy (kit 12)).

To get a more detailed insight of how DNA is lost during bisulfite treatment, we used dPCR to distinguish between DNA loss during clean-up and DNA fragmentation during conversion. This method combines the insights obtained by spectroscopic measurements, gel electrophoresis and qPCR to analyze the DNA recovery and the fragmentation. The exact amount of DNA strands lost during the complete bisulfite treatment can be measured by comparing the concentration before and after treatment. By quantifying the exact amount of DNA fragments of different lengths, a fragmentation assessment can be made. The conclusions made by the dPCR about the fragmentation and the DNA loss are in accordance with the results of the Qubit, qPCR and gel electrophoresis, showing the strength of dPCR to combine recovery and fragmentation analysis in a single platform. In this work, we did not evaluate whether the DNA fragmentation and recovery are differently affected by the conversion step and the clean-up step. Future work could help to decouple the effects of both separate parts of bisulfite treatment on the resulting distribution of fragment sizes to completely understand the fragmentation kinetics and further optimize methylation kit performance.

In the current study, Epitect (kit 10) was selected as the most appropriate kit for methylation studies of long DNA fragments since this kit appears to induce least fragmentation. Its conversion time of approximately 300 minutes is much longer than the average (80 minutes). Of note: we could not

exclude a potential bias in favor for Epiect (kit 10) since the DNA extraction and bisulfite treatment kits are provided by Qiagen. Earlier studies by Grunau et al., Raizis et al. and Holmes et al. indicated that increased conversion temperature or prolonged conversion time cause more fragmentation (153,154,369). In the current study, we found no evidence for a significant correlation between time and fragmentation or between temperature and fragmentation for different kits. Interestingly, Grunau et al. found that fragmentation mainly occurs during the first time period of the conversion (after 5 minutes, they observed >90% DNA loss). This could explain the lack of effect on fragmentation by changing the conversion protocol, and it could implicate that the first denaturation step is the most important factor for the fragmentation during conversion. This step was not altered in the alternative protocols of Epiect (kit 10), and could explain the lack of effect on the fragmentation of these alterations. Although, this long incubation could potentially influence the conversion efficiency and it does hamper high throughput methylation analysis with this kit (389). For high throughput studies, several kits provide a protocol with an incubation time less than 60 minutes. In analysis of shorter fragments, other factors might be decisive: if the amount of DNA is limited, a high DNA recovery is needed, and thus EZ Gold (kit 5) appears to be a better option. Next to this, a kit with a premade bisulfite mix (e.g. EZ Lightning (kit 6)) would fit better to minimize pipetting variation and is ideal for high-throughput analyzes.

In conclusion, our study provides a comprehensive workflow to monitor bisulfite conversion kits based on different parameters such as fragmentation, recovery or conversion efficiency. By using dPCR to analyze the DNA recovery, we are able to analyze the DNA loss and DNA fragmentation with one method. This comprehensive method enables researchers to select the most appropriate kit, depending on the application in which DNA methylation should be analyzed. This test could be important for every study, since the manufacturers can improve the reagents and protocols of the kits over time. In this study, we limited the analysis to standard conditions for the most kits and we only used genomic DNA obtained with one method. Future studies could be performed with equalized DNA input, elution volume, conversion conditions and DNA clean up in order to further understand the dynamics of the bisulfite treatment, or use different types of input DNA in order to assess the influence of the buffers of the extraction method on these parameters. There, our workflow would perfectly suit to be used to monitor the effect of changes during optimizations of a bisulfite kit.



## Objective Ib & Ic – Amplification of HIV-1 CpGs in vivo

The second step of the assay, the amplification of the CpGs of interest was optimized in order to increase the **compatibility** to the high genomic variability of HIV-1, without losing **specificity** and to increase the **sensitivity** to be compatible with patient samples carrying a low amount of infected cells.

The accuracy of this optimized protocol was tested by *in vitro* methylation, and subsequently used to measure methylation (intragenic (*env*) and promoter (LTR) methylation) in PBMCs of a well-characterized patient cohort consisting of 72 patients, divided in four groups: (i) LTNPs, (ii) acute seroconverters, (iii) patients on therapy that started treatment early and (iv) patients that started therapy late.

The following work is submitted to Clinical Epigenetics for publication on December 27th 2019: “*Kint S, Trypsteen W, Van Hecke C, Malatinkova E, De Spiegelaere W, Kinloch-de Loes S, De Meyer T, Van Criekinge W, Vandekerckhove L. Underestimated effect of intragenic HIV-1 DNA methylation on viral transcription in HIV infected patients*”, and is adapted to fit in this thesis.

### Contribution of the doctorandus:

Study design

Experimental work assay optimization and methylation profiling

Data analysis, visualization and interpretation

Manuscript writing

# Underestimated effect of intragenic HIV-1 DNA methylation on viral transcription in infected individuals

Sam Kint<sup>1,2</sup>, Wim Trypsteen<sup>1</sup>, Ward De Spiegelaere<sup>3</sup>, Eva Malatinkova<sup>1</sup>, Sabine Kinloch-de Loes<sup>4</sup>, Tim De Meyer<sup>2</sup>, Wim Van Criekinge<sup>2,\*</sup>, Linos Vandekerckhove<sup>1,\*§</sup>

\*Authors contributed equally

<sup>1</sup>HIV Cure Research Center, Department of Internal Medicine and Pediatrics, Faculty of Medicine and Health Sciences, Ghent University and Ghent University Hospital, Corneel Heymanslaan 10, 9000 Ghent, Belgium

<sup>2</sup>Biobix, Department of Data Analysis and Mathematical Modelling, Faculty of Bio-science Engineering, Ghent University, Coupure Links 653, 9000 Ghent, Belgium

<sup>3</sup>Department of Morphology, Faculty of Veterinary Medicine, Ghent University, Salisburylaan 133, 9820 Merelbeke, Belgium

<sup>4</sup>Division of Infection and Immunity, Royal Free Hospital, Royal Free Campus, University College London, Pont St, Hampstead, London NW3 2QG, United Kingdom

**Corresponding author (§):** Linos Vandekerckhove; Corneel Heymanslaan 10, Medical Research Building 2, Ghent University, 9000 Ghent, Belgium; Tel: +3293323398; [linos.vandekerckhove@ugent.be](mailto:linos.vandekerckhove@ugent.be)



## Abstract

**Background:** The HIV-1 proviral genome harbors multiple CpG islands (CpGIs), both in the promoter and intragenic regions. DNA methylation in the promoter region has been shown to be heavily involved in HIV-1 latency regulation in cultured cells. However, its exact role in proviral transcriptional regulation in infected individuals is poorly understood or characterized. Moreover, methylation at intragenic CpGIs has never been studied in depth.

**Results:** A large, well-characterized HIV-1 patient cohort (n=72), consisting of 17 long-term non-progressors and 8 recent seroconverters (SRCV) without combination antiretroviral therapy (cART), 15 early cART-treated and 32 late cART-treated patients, was analyzed using a next-generation bisulfite sequencing DNA methylation method. In general, we observed low level of promoter methylation and higher levels of intragenic methylation. Additionally, SRCV showed increased promoter methylation and decreased intragenic methylation compared to the other patient groups. This data indicates that increased intragenic methylation could be involved in proviral transcriptional regulation.

**Conclusions:** Contrasting *in vitro* studies, our results indicate that intragenic hypermethylation of HIV-1 proviral DNA is an underestimated factor in viral control in HIV-1 infected individuals, showing the importance of analyzing the complete proviral genome in future DNA methylation studies.

**Keywords:** HIV-1; DNA methylation; HIV-1 latency; Epigenetics; Next-Generation Sequencing; Bisulfite sequencing, intragenic DNA methylation

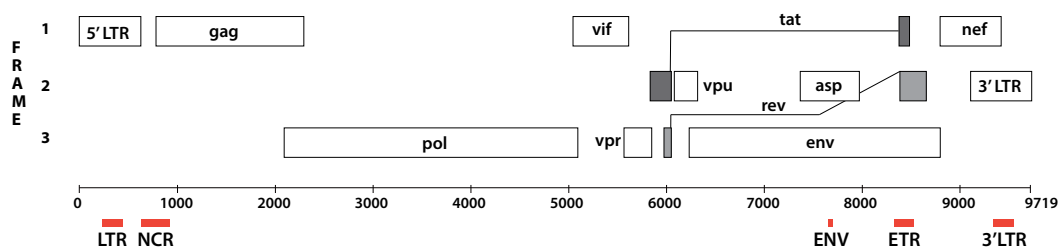
## Background

Current combination antiretroviral therapy (cART) can successfully control human immunodeficiency virus type 1 (HIV-1) infection and prevent disease progression to the acquired immunodeficiency syndrome (AIDS). However, a cure is not generally achievable due to the establishment of a latent reservoir of proviral HIV-1 DNA which remains dormant and fuels viral rebound upon treatment interruption (204–206,309). Therefore, better insight into the mechanisms regulating HIV-1 latency is crucial in order to interfere with this latency state and to develop cure strategies. The state of HIV-1 latency can be defined as the transcriptional silencing of proviral genes caused by multiple transcriptional blocks after the stable integration of proviral DNA into the host genome (275). Some of the major silencing mechanisms consist of epigenetic modifications, which have led to several clinical trials investigating the latent viral reservoir reactivation with histone deacetylase inhibitors, albeit with limited success (320,328–330,334). Other epigenetic modifications such as HIV-1 proviral DNA methylation have also been described in HIV-1 transcriptional silencing and have been explored as targets for HIV-1 latency reversing strategy (20,21,61,355).

DNA methylation is a well-described epigenetic modification in which a methyl group is added at the number five carbon of the cytosine pyrimidine ring in CpG dinucleotides (13,42). This modification

plays a role in genome transcription regulation and is crucial in processes such as the development of multicellular organisms, cell differentiation, regulation of gene expression, X-chromosome inactivation, genomic imprinting and in the suppression of parasitic and other repeat sequences (12–15,27,42–45). In general, reliable and stable transcriptional silencing is caused if CpG islands (CpGIs) – stretches of DNA that contain an increased frequency of CpG dinucleotides (CG content > 50% and observed/expected CpG ratio > 60%) – in promoter regions are hypermethylated (13,42,61,63,85). Methylation of CpGIs within gene bodies (intragenic methylation) has been shown to be involved in regulation of intragenic promoters, alternative splicing and cellular differentiation, but also in the activation of retroviruses, repetitive elements and prevention of aberrant transcript production (98–102).

The HIV-1 genome encodes five CpGIs (61): two surrounding the promoter region and flanking the HIV-1 transcription start site and several transcription factor binding sites (e.g. TCF-1 $\alpha$ , NF- $\kappa$ B, SP1) at the 5' long terminal repeat (LTR) region (CpGI LTR in the U3 region of the 5'LTR and CpGI Non-Coding Region (NCR), downstream the HIV-1 5' LTR (**Figure 26**)) (61). Two other CpGIs are located in the *env* gene (CpGI ENV (35% conserved) and CpGI *env-tat-rev* (ETR)), surrounding the HIV-1 antisense open reading frame (**Figure 26**) (61,228). The fifth CpGI is located in the 3'LTR, where the antisense transcription start site is located (61,228). In cultured HIV-1 infected cells, the regulatory role of proviral promoter methylation in viral transcriptional activity is clearly demonstrated: hypermethylation stabilizes HIV-1 latency and demethylating agents can induce activation of HIV-1 transcription (19,20,61,337,338). However, studies performed on DNA methylation in infected individuals could not reproduce these findings indicating that this *in vitro* regulation does not apply *in vivo* (19,207,350,353–355).



**Figure 26. Location of the 5 CpGIs in the HIV-1 genome.** The location of the 5 CpGIs as described by Chavéz *et al.* (61). CpGIs long terminal repeat (LTR) and non-coding region (NCR) are located around the HIV-1 promoter location. CpGIs ENV and *env-tat-rev* are located in the *env* gene. The fifth CpGI (3'LTR) is located in the 3' LTR region, where the antisense promoter region is found.

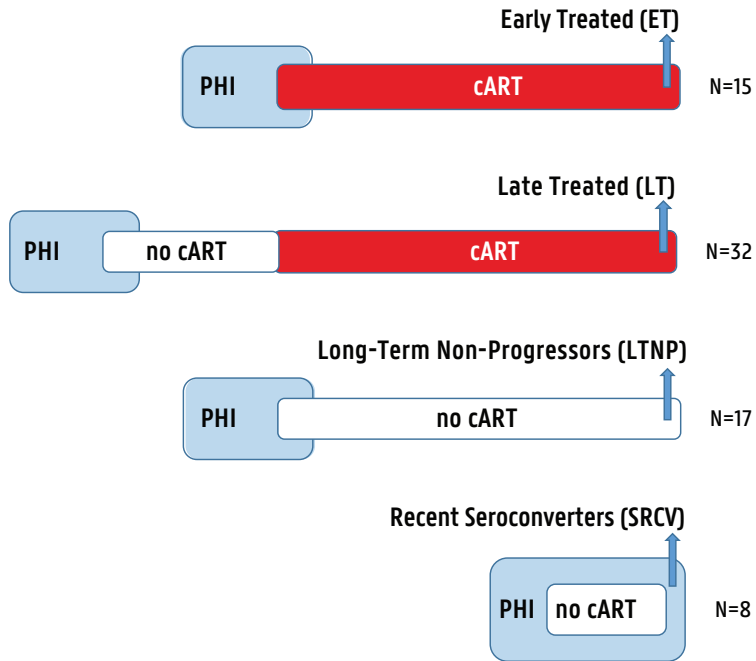
To further understand the role of proviral HIV-1 DNA methylation in infected individuals, an NGS-based bisulfite assay was developed to characterize HIV-1 proviral DNA methylation profiles of both promoter and intragenic regions in the context of a large, well-characterized patient cohort (n=72). This cohort comprises four different patient groups as described by Malatinkova *et al.* (390): 15 early cART-treated individuals (ET), 32 late cART-treated individuals (LT), 17 long-term non-progressors (LTNP) and 8 acute seroconverters (SRCV).

## Methods

### Patient cohorts and DNA samples

HIV-1 positive patients were recruited from two clinical centers, the Ian Charleson Day Centre (Royal Free Hospital, London, United Kingdom) and the AIDS Reference Center (Ghent University Hospital, Ghent, Belgium) during the study performed by Malatinkova *et al.* (390). Seventy-two HIV-1 positive PBMC samples from that study were selected. Patients were divided into four cohorts based on their disease status (**Figure 27**). The detailed study design and inclusion criteria have been described previously (390). Briefly, 1) long-term cART-treated individuals (median treatment time of 10.77 years (interquartile range (IQR): 6.46-12.34 years)) who had initiated treatment during HIV-1 seroconversion (early treated (ET); n=15) or 2) during the chronic phase of the infection (late treated (LT); n=32); 3) cART-naïve long-term non-progressors (LTNPs, n=17) who had maintained HIV-1 viral load (VL)  $\leq$  1000 copies/ml and CD4+ T cells  $>$ 500 cells/mm<sup>3</sup> over  $>$ 7 years post-infection or 4) cART-naïve seroconverters (SRCV, n=8), who were sampled during the acute phase of the infection. Baseline characteristics and clinical parameters of these cohorts are summarized in **Table 12**. The Ethical Committees of Ghent University Hospital and the Royal Free Hospital had approved this study (reference numbers: B670201317826 (Ghent) and 13/LO/0729 (London)) with all study subjects giving their written informed consent.

DNA from aliquots of 10<sup>7</sup> PBMCs was isolated using the DNeasy® Blood & Tissue Kit (Qiagen, The Netherlands, 69504). Sample DNA concentration was determined with the Qubit dsDNA BR (broad range) Assay Kit (Thermo Fisher Scientific, MA, USA, Q32850) on a Qubit 2.0 fluorometer according to the manufacturer's instructions.



**Figure 27. Overview of patient cohorts included in this study.** Patients are divided into four groups based on their disease state: early treated, late treated, Long-Term Non-Progressor and acute seroconverter. Arrows depict moment of sampling. PHI = Primary HIV-1 Infection; cART = combination Anti-Retroviral Therapy

**Table 12:** Clinical characteristics and viral reservoir markers of the patient cohorts

	Cohort 1 = ET	Cohort 2 = LTNP	Cohort 3 = LT	Cohort 4 = SRCV
# patients	15	17	32	8
<b>Clinical characteristics</b>				
Age (yrs)	45 (43 - 54.5)	49 (38 - 51)	48 (45 - 53.25)	37 (27 - 44.75)
Total cART (yrs)	11.65 (10.39 - 11.97)	0 (0 - 0)	9.80 (6.09 - 14.73)	0 (0 - 0)
Total VL suppression (yrs)	11.18 (9.82 - 11.37)	9.72 (0 - 14.67)	6.53 (5 - 10.42)	0 (0 - 0)
log VL zenith (copies/ml)	5.74 (5.31 - 5.88)	2.24 (1.79 - 2.76)	4.93 (4.24 - 5.52)	6.15 (5.14 - 6.31)
CD4 nadir (cells/ $\mu$ l)	413.5 (274.5 - 539.75)	624 (562 - 693)	154.5 (51.25 - 266.25)	483.5 (393.75 - 520.25)
CD4 at collection (cells/ $\mu$ l)	961 (737 - 1129.5)	793 (685 - 1010)	624.5 (484 - 885.5)	534 (393.75 - 617.50)
CD4/CD8	1.12 (0.8 - 1.47)	0.91 (0.82 - 1.47)	0.74 (0.6 - 0.93)	0.62 (0.37 - 0.87)
<b>Viral reservoir markers</b>				
Total HIV-1 DNA (c/M PBMC)	88.14 (46.19 - 124.02)	48.01 (20.16 - 56.50)	137.01 (56.08 - 219.20)	1290.48 (519.63 - 4428.60)
Integrated HIV-1 DNA (c/M PBMC)	158.00 (122.70 - 388.55)	28.16 (0 - 158.41)	586.65 (315.12 - 918.15)	1802.68 (272.19 - 3966.55)
CA HIV-1 usRNA (c/M PBMC)	0.79 (0.28 - 3.12)	0.44 (0.27 - 3.51)	6.12 (1.80 - 10.08)	15.47 (0.62 - 77.60)
2-LTR circles (c/M PBMC)	1.48 (0 - 3.03)	0.77 (0.65 - 2.70)	1.32 (0.57 - 2.18)	15.35 (4.82 - 24.12)

Values are reported as median (Interquartile Range); SRCV: seroconverters; LTNP: long-term non-progressors; PBMCs: peripheral blood mononuclear cells; CA: cell-associated; usRNA: unspliced RNA; cART: combination antiretroviral therapy; VL: viral load.

## Cell culture

Jurkat cells (human T cell leukemia line) and J-Lat 8.4 (Jurkat cells infected with one HIV-1 copy per cell (391)) were cultured in a humidified atmosphere of 37°C and 5% CO<sub>2</sub> in RPMI 1640 medium with GlutaMAX™ Supplement (Thermo Fisher Scientific, MA, USA, 61870-010), supplemented with 10% FCS and 100 µg/mL penicillin/streptomycin. The culture medium was renewed every two to three days. DNA was isolated as described in the previous section.

## Primer design

Primers targeting the four major HIV-1 CpGIs were designed using two online available primer design tools (Methprimer (392) and Bisulfite primer seeker (Zymo Research, CA, USA, <https://www.zymoresearch.com/pages/bisulfite-primer-seeker>)). LTR primers were obtained from Trejbalova *et al.* (20) and ETR\_1 primers from Weber *et al.* (353). To evaluate primers *in silico*, the bio-informatics tool developed by Rutsaert *et al.* (393), estimating the complementarity of each primer combination to all full length HIV-1 sequences in the Los Alamos National Laboratory (LANL) database, which is the largest public collection for HIV sequences ([www.hiv.lanl.gov](http://www.hiv.lanl.gov)) (394), was adapted: the database was transformed to the bisulfite treated variant (C → T; CG → CG), nested primer combination analysis was included, as well as analysis of combinations of multiple PCR assays. First, the *in silico* analysis was used to evaluate primer combinations that were obtained from literature as well as in-house designed. Primer combinations matching at least 50% of the LANL database and nested combinations with an overlap of at least 2/3 of the matched sequences were retained. Selected primers were *in vitro* tested using DNA from J-Lat 8.4 (391), diluted in Jurkat DNA at different concentrations to mimic patient samples (10000, 5000, 1000, 500, 250, 100 HIV-1 copies per 10<sup>6</sup> cells). Finally, an additional *in silico* analysis was used to select four or less primer combinations per CpGI that targeted at least 60% of the LANL database. These final primer sequences are listed in **Additional File 1**.

## Bisulfite treatment

A minimum of 5x1µg of DNA per patient was bisulfite treated using the Epiect Bisulfite kit (Qiagen, The Netherlands, 59110), which is the least fragmenting commercial bisulfite kit available, according to a previous in-house comparison (12). We used the standard protocol as provided by the manufacturer. The five aliquots per patient were pooled, and immediately stored at -20°C.

## Bisulfite specific PCR

All PCR reactions were performed in triplicate to reduce the probability of preferential amplification of one specific amplicon that would dominate the output. Nested PCR reactions were performed using the FastStart™ Taq DNA Polymerase, 5 U/µl (Roche Applied Science, Belgium, 12032953001). A volume containing theoretically at least ten bisulfite-treated HIV-1 copies (based on the droplet digital PCR measurements as in Malatinkova *et al.* (390)) was added to the PCR mix containing 10x PCR

buffer, 2.5U polymerase, 400 nM forward and reverse primers and 3% DMSO in a final volume of 25  $\mu$ l. Each CpG was amplified with one nested primer combination, and after a failed PCR reaction, the subsequent primer combination was used (**Additional File 1**). Amplicons were visualized using 3% agarose gel electrophoresis. Depending on the selected primer, we used an in-house optimized PCR amplification protocol or one of the two previously published protocols (20,353), as described in **Additional File 1**.

## Sequencing

Bisulfite-treated amplicons were pooled equimolarly and libraries were prepared using the NEBNext Ultrall DNA Library Prep Kit for Illumina (NEB, MA, USA, #E7645L/#E7103L). These libraries were sequenced on a MiSeq sequencing system (MiSeq® Reagent Kit v3 (600 cycle), MS-102-3003, Illumina). Sequencing reads were trimmed using Trimmomatic (version 0.38), quality controlled using FastQC (version 0.11.8), and subsequently mapped to an in-house developed HIV-1 consensus genome using the Bismark package (version 0.10.1) (395), providing a conversion efficiency estimation and methylation state of all analyzed CpGs.

## Statistical analysis

HIV-1 specific amplicons with coverage > 250 were normalized and divided into tiles (blocks of the HIV-1 genome containing the region of interest (LTR or *env*)). Differential methylation analysis per region was performed using the MethylKit package (version 1.6.3) in R (version 3.5.1)(396,397), including correction for overdispersion. P-value calculation was performed using the Chi-square test and p-value correction for multiple testing was performed within each comparison using false discovery rate (FDR) (398,399).

Spearman rank correlation analysis was performed to explore correlations between DNA methylation (LTR and *env*) and patient characteristics (HIV-1 reservoir and immunological parameters, obtained from Malatinkova *et al.* (390)). Therefore, methylation data of both regions of every individual was summarized by calculating an M-value over all CpGs using the formula as described by Du *et al* (400). Using stepwise regression model selection, linear regression models were developed for LTR and *env* methylation densities to determine which independent variables may explain variable DNA methylation in both regions.

Visualization was performed using R (version 3.5.1) with the following packages: PMCMR (version 4.3), Hmisc (version 4.2-0), graphics (version 3.5.1), ggplot2 (version 3.1.0) and corrplot (version 0.84) (397).

## Results

### *In silico*, *in vitro* and *in vivo* HIV-1 DNA methylation assay development

338 different nested primer combinations (assays) (13 LTR, 303 NCR, one ENV and 21 ETR) were subjected to an *in silico* analysis using an adapted version of the bioinformatics tool developed by Rutsaert *et al.* (393) to estimate the complementarity to the LANL database, resulting in 70 nested PCR assays (two LTR, 46 NCR, one ENV and 21 ETR, **Figure 28A**). The performance of these assays was subsequently tested by PCR amplification in undiluted and diluted J-Lat 8.4 DNA (up to 100 infected cells/10<sup>6</sup> cells), resulting in 36 assays (two LTR, 15 NCR, one ENV and 18 ETR) that were capable of generating PCR products at the lowest dilutions (**Figure 28A**). After a final *in silico* analysis, a set of nine primer combinations (two LTR, three NCR, one ENV and three ETR; **Figure 28** and **Additional File 1**) was selected.

These nine assays were used to determine the HIV-1 methylation profile of HIV-1 positive blood samples. The percentage of patients for whom the primer combinations generated PCR amplicons is listed in **Table 13**. This data demonstrates a similar trend as expected based on the *in silico* analysis, being that a certain percentage of HIV-1 sequences would not be detected in patients for certain primer combinations due to HIV-1 sequence variation. The difference between expected amplification percentage and the actual amplification percentage was 7.85%, 1.57%, 10.58% and 3.57% for LTR, NCR, ENV and ETR, respectively (**Table 13**).

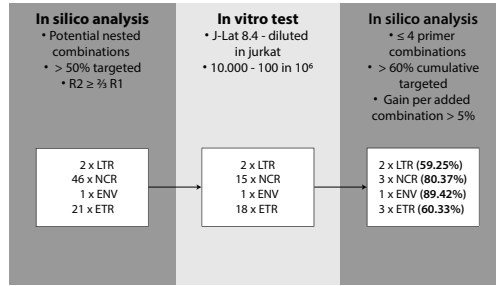
**Table 13.** Performance of the nine final assays compared to the predicted performance using *in silico* analysis of the primer complementarity.

	LTR	NCR	ENV	ETR
# primer combinations	2	3	1	3
% of patients expected to generate amplicons based on <i>in silico</i> analysis*	59.25	80.37	89.42	60.33
% of patients in which DNA is amplified	51.40	81.94	100	63.90
% of patients from which the DNA had sufficient quality to be mapped to HIV-1	48.61	75.00	41.67	63.90

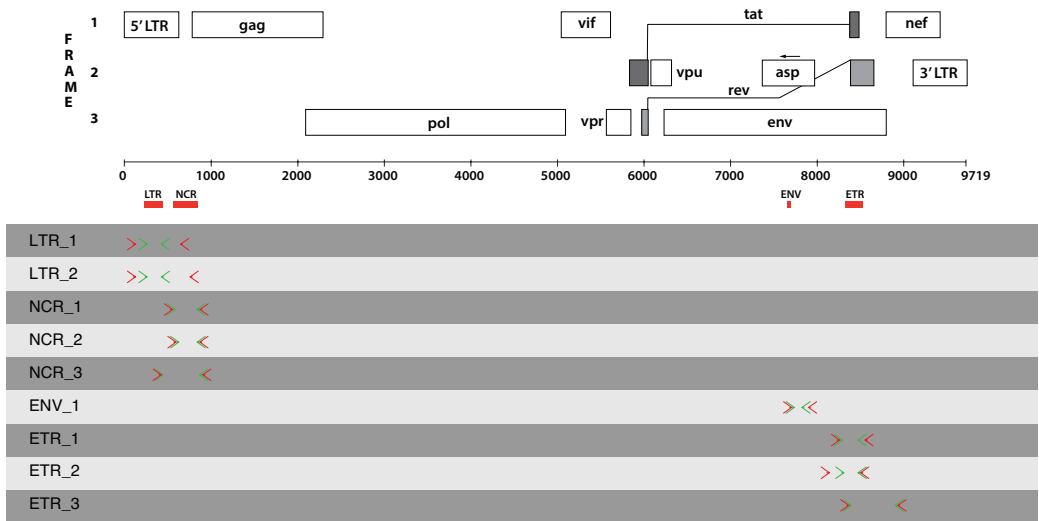
\**In silico* analysis is based on the bioinformatics primer evaluation tool as described by Rutsaert *et al.* (382)



**A**



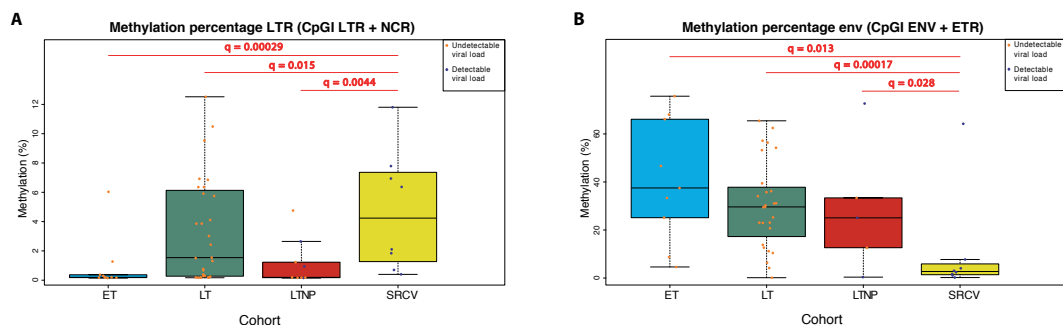
**B**



**Figure 28. Primer selection procedure.** **A.** Workflow used for the development of our DNA methylation assay determining HIV-1 DNA methylation in HIV-1 infected patient samples. **B.** Location of the nine different assays on the HIV-1 genome. Red arrows depict first round PCR primer location; green arrows show second round PCR primer location; Red bars indicate the location of the four analysed CpGIs based on Chavez et al. (61).

## SRCV show increased LTR methylation and decreased *env* methylation

In all four patient cohorts together, average methylation of all CpGs within the LTR region was 2.94% (IQR: 0.19-5.5%). When comparing patient cohorts, we observed significantly higher LTR methylation in SRCV as compared to all the other cohorts (ET, LT and LTNP) ( $\Delta = 6.48\%$ ;  $q = 0.00029$ ,  $\Delta = 4.15\%$ ;  $q = 0.015$  and  $\Delta = 5.94\%$ ;  $q = 0.0044$ , respectively) (**Figure 29A**).

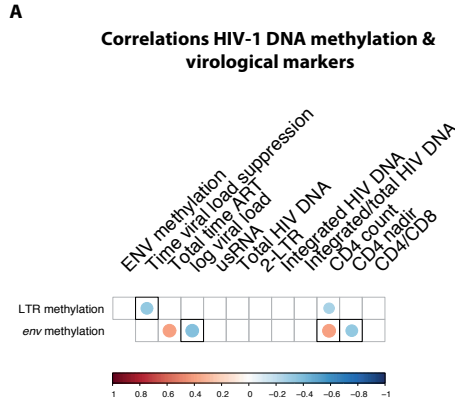


**Figure 29. HIV-1 proviral DNA methylation comparison between patient cohorts.** **A.** Summary of the methylation data in the LTR region (CpGI LTR + CpGI NCR) using average methylation over all CpGs in the region. **B.** Summary of the methylation data in the *env* region (CpGI ENV + CpGI ETR) using average methylation over all CpGs in the region.  $q$  = FDR-corrected  $p$ -values for multiple testing. LT = late treated; ET = early treated; SRCV = acute seroconverter; LTNP = long-term non-progressor.

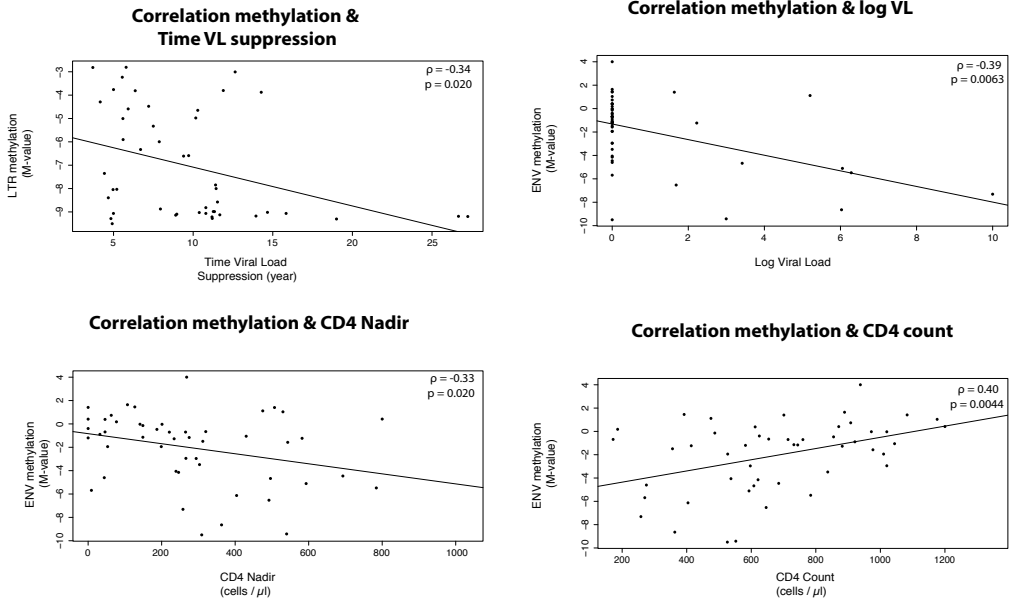
Higher CpG methylation was observed in the *env* region as compared to LTR, averaging 28.86% (IQR: 8.73-39.44%). All cohorts (ET, LT and LTNP) showed a significantly higher methylation density compared to SRCV ( $\Delta = 33.47\%$ ;  $q = 0.013$ ,  $\Delta = 35.32\%$ ;  $q = 0.00017$  and  $\Delta = 35.26\%$ ;  $q = 0.028$ , respectively) (**Figure 29B**).

## Correlations between HIV-1 methylation status and reservoir markers

During the explorative correlation analysis, negative correlations were found between the DNA methylation density in the LTR region and the duration of viral suppression ( $\rho = -0.34$ ;  $p = 0.020$ ) and CD4+ T cell count at time of collection ( $\rho = -0.27$ ;  $p = 0.043$ ) (**Figure 30A**). However, we observed a significantly positive association for DNA methylation in the *env* region and the CD4 T cell count ( $\rho = 0.40$ ;  $p = 0.0045$ ) and cART duration ( $\rho = 0.39$ ;  $p = 0.0055$ ) (**Figure 30A**). Moreover, *env* methylation decreased with increasing VL levels ( $\rho = -0.39$ ;  $p = 0.0063$ ) and higher CD4+ T cell nadir ( $\rho = -0.33$ ;  $p = 0.020$ ) (**Figure 30A**). Based on the linear regression models, the only variable that was independently associated with DNA methylation in the LTR was the duration of VL suppression. Three variables were independently associated with the *env* methylation: VL, CD4 nadir and CD4 count at time of sampling (**Figure 30B**).



**B**



**Figure 30. Spearman correlations between HIV-1 proviral DNA methylation and patient characteristics. A.** Correlation of DNA methylation with several virological and viral reservoir markers in HIV-1 infected individuals. Positive and negative correlations are depicted in red and blue, respectively. Non-significant correlations are left blank. Correlations with covariates that independently explained methylation in the linear regression models are depicted with a black frame. **B.** Correlation plots between DNA methylation (M-value) and the independent variables from the linear models. Upper left: LTR methylation vs. duration of VL suppression. Upper right: env methylation vs. log VL. Lower left: env methylation vs CD4 nadir. Lower right: env methylation vs CD4 count.

## Discussion

The lack of consensus about the role of proviral DNA methylation in HIV-1 transcriptional regulation illustrates the need for a reliable and widely applicable methylation assessment method. In this study, we first described an *in silico* procedure to accurately predict the complementarity of PCR assays to the HIV LANL database, and an *in vitro* validation protocol to test the sensitivity of the designed assays. This procedure resulted in nine functional DNA methylation assays, designed against the four most common CpGs of the HIV-1 provirus, which were consequently used to characterize HIV-1 DNA methylation in a large, well-characterized patient cohort. The *in silico* analysis was predictive of the number of patient samples leading to successfully amplified PCR products (**Table 13**), indicating that this is an effective approach to prioritize testing of primer sets in the context of HIV-1 or other pathogens with a high sequence variability. In addition, as shown in the study of Cortés-Rubio et al. (355), by using an NGS-based approach, our method fulfils the need to analyze a large number of proviruses for each patient when compared to the established Sanger sequencing-based methods (401).

Across our four patient cohorts, we have found that the HIV-1 provirus had low amounts of DNA methylation in the promoter region (average 2.94%, IQR: 0.19-5.5%) but substantially higher levels of intragenic (*env*) methylation (average 28.86%, IQR: 8.73-39.44%). When comparing the differential methylation between the cohorts, only SRCV showed distinct methylation profiles, with increased LTR, and decreased *env* methylation. Similarly, if patients were divided based on their VL status (detectable VL (VL > 40 HIV-1 copies / ml plasma), comprising all SRCV and 6/17 LTNPs. vs undetectable VL (VL < 40 HIV-1 copies/ml plasma), comprising ET, LT and 11/17 LTNPs), individuals with a detectable VL had higher DNA methylation density in the HIV-1 LTR region and a lower density in the *env* region compared to those with an undetectable VL. These observations might indicate that specific methylation profiles may be associated with *in vivo* HIV-1 transcriptional control and latency maintenance.

Indeed, since the involvement of DNA methylation in HIV-1 latency was first described in 1987 (17), it has been confirmed in HIV-1 infected cultured cells and latency models that promoter methylation density is associated with silencing stability: DNA methylation induction can initiate/stabilize HIV-1 latency, while methylation inhibitors as 5-aza-2'-deoxycytidine (5-aza-CdR) cause HIV-1 reactivation and display clear synergistic effects with other latency reversing agents (18–21,61,337,338,350–352). These studies reported a major role of promoter DNA methylation in latency regulation, which was in line with the general concept of transcription regulation by DNA methylation: hypermethylation of the promoter region suppresses both basal promoter activity and responses to activating stimuli, and hypomethylation is a transcription mark (352). However, DNA methylation studies on patient-derived samples have shown – with the exception of some LTNPs – the same trend as in our present observation: low level of DNA methylation in the promoter region, even in patients suppressing VL successfully, therefore not following the predictions from the *in vitro* experiments (353,354). It has been shown that DNA methylation behavior in cell lines is often drastically different from that of

*in vivo* cells due to completely different epigenetic environments and immortalization, sometimes producing unreliable results in terms of predicting *in vivo* DNA methylation events (362,363).

In previous DNA methylation studies in HIV-1 patients, the focus was on promoter methylation assessment (19,20,350,353–355). In contrast to promoter methylation, the role of intragenic DNA methylation in general transcriptional regulation is less clearly described (98–102). Studies outside of the HIV-1 field, have suggested that intragenic methylation could have a role in the activation of retroviruses, repetitive elements, alternative splicing, transcription initiation in canonical promoters of embryonic stem cells and prevention of aberrant transcript production (100–102). Moreover, intragenic methylation has been shown to be a robust predictor of gene transcription in genes with a CpGI containing promoter (107). In our study, decreased *env* methylation levels in individuals with active ongoing replication (SRCVs) suggests that intragenic methylation increases in the case of proviral transcriptional silencing, leading to higher methylation in latently infected cells or in those in which viral replication is blocked. Indeed, cART-treated patients and LTNP have lower viral transcription (measured as cell-associated unspliced RNA (CA usRNA)) than SRCV (**Table 12**) and *env* methylation shows an inverse correlation with CA usRNA within the SRCV cohort ( $\rho = -0.81$ ;  $p = 0.014$ ). Furthermore, intragenic methylation did correlate positively with the CD4+ T cell count, linking high intragenic methylation with viral control. Intragenic methylation was also negatively associated with the VL, a measure that indicates ongoing replication.

In contrast to what was proposed by LaMere *et al.* (401), we have found no statistical difference between proviral methylation in LTNP with undetectable VL (latent infection) and treated patients (cART-induced suppression) (LTR:  $\Delta = 0.85\%$ ,  $q = 0.74$ ; *env*:  $\Delta = 2.29\%$ ,  $q = 0.94$ ). This could be due to the low number of LTNPs with undetectable VL.

This study shows that intragenic DNA methylation could be of importance in HIV-1 transcriptional regulation, however, it has some limitations that could be addressed in similar future studies. Whereas the cohort size was much larger than previous studies (19,20,350,353–355), the patient groups described here were not balanced, not in size, nor for sex and age. Additionally, we did only sample one single time point, and did no specific CD4+ T cells selection. The use of PBMCs could potentially mask differential methylation since it is shown that LRAs have cell type specific effects, indicating cell type specific epigenetic profiles (324). Moreover, due to the targeted nature of the methodology, it does not allow to provide information about integration site methylation or replication competence of the analyzed provirus.

## Conclusions

Altogether, our study illustrates the underestimation of the role of intragenic proviral DNA methylation in patient samples. Previous studies have mainly focused on LTR methylation, and have interpreted LTR methylation as a transcriptional regulatory factor, ignoring any potential role of *env* methylation (20,207,354). We suggest that both *env* and LTR methylation are involved in HIV-1 transcription regulation and that *env* methylation could be an important predictor of viral transcription *in vivo*.

However, we also suggest that proviral promoter methylation is hindered/inhibited in all HIV-1 positive patients, especially those on cART, but that its density still influences viral transcription rate.

The exact functions of DNA methylation of these two regions should be clarified by performing additional experiments using longitudinal follow-up studies to monitor proviral DNA methylation dynamics within patients, starting early during infection, and ideally continuing over a period of multiple years of cART. Different CD4+ T cell types should be analyzed separately to avoid cell-type dependent bias of the data. If HIV-1 positive patients were to undergo treatment interruption, DNA methylation profiles should also be monitored in order to understand the methylation dynamics during viral rebound. Moreover, proviral intragenic non-CpGI methylation analysis could also provide a better understanding of HIV-1 latency regulation by DNA methylation. Here, we do provide a useful tool to help design and estimate the sample size needed in these studies. Altogether these insights should be of paramount importance when looking at the various strategies to control HIV-1 after discontinuation of cART and for the HIV-1 cure field.

## ***Abbreviations***

cART: combination antiretroviral therapy; HIV-1: human immunodeficiency virus type 1; AIDS: acquired immunodeficiency syndrome; CpGIs: CpG islands; LTR: long terminal repeat; NCR: Non-Coding Region; ETR: *env-tat-rev*; PBMC: peripheral blood mononuclear cells; LTNP: long-term non-progressors; VL: viral load; NGS: next-generation sequencing; ET: early cART-treated individuals; LT: late cART-treated individuals; IQR: interquartile range; CA usRNA: cell-associated unspliced RNA; LANL: Los Alamos National Laboratory; qPCR: quantitative real-time PCR; 5-aza-CdR: 5-aza-2'-deoxycytidine

## ***Acknowledgements***

The following reagent was obtained through the NIH AIDS Reagent Program, Division of AIDS, NIAID, NIH: J-Lat Full Length Clone (clone 8.4) from Dr. Eric Verdin. This study was in part performed with the support of The Foundation for AIDS Research (AmfAR), United States (Grant ID: 108314-51-RGRL), HIV-ERA (130442 SBO, EURECA) and by The National Institute of Health (NIH), United States (Grant R01 AI134419). Linos Vandekerckhove was supported by the Research Foundation Flanders (FWO), Belgium and he received a fundamental clinical mandate (1.8.020.09.N.00). Eva Malatinkova was funded by the Agency for Innovation by Science and Technology of the Flemish Government, Belgium (IWT, grant ID: 111286) and is currently funded by Special Research Fund Ghent University, Belgium (BOF, grant ID: 01P06816). The study did not have pharmaceutical firm sponsorship.

## ***Declaration of interests***

The authors declare no competing interests.

## *Author contributions*

SK designed the study, performed experiments, analyzed, interpreted and visualized data, and drafted the manuscript. SKdeL and LV recruited patients. EM, SKdeL and WDS contributed to data collection. WDS contributed to study design. WT contributed to data interpretation and manuscript writing. TDM contributed to the statistical data analysis. LV and WVC designed the study, supervised experiments and contributed to manuscript writing. All co-authors revised and edited the manuscript.

## *Ethics*

Human subjects: Patient written informed consent was obtained from all the study participants. The study was approved by the Ethical Committee of Ghent University Hospital (Reference number: B670201317826) and Royal Free Hospital (Reference number: 13/LO/0729).

## Additional File 1: Primers PCR experiments

	Sequence Round 1	Binding region (HXB2)	Name Sam	Protocol round 1
<b>LTR</b>				
<b>Round1</b>				
LTR_1_f (a)	TAGATATTTATTGATTTTTGGATGGTG			
LTR_1_r (a)	AAAAAACTCTCTAATTTYHCTTC	110-686	ART2-ART3	Trejbalova_R1 (a)
LTR_2_f (a)	TAGATATTTATTGATTTTTGGATGGTG			
LTR_2_r (a)	CACCCATCTCTCCTCTAACCTC	110-796	ART2-ART2	Trejbalova_R1 (a)
<b>Round2</b>				
LTR_1_f (a)	AGTGTTAGTGTGGAGGTTTGATA			
LTR_1_r (a)	CAAAAAACCCAATACAAACAAAAA	248-464	ART2-ART3 – ART3-EXTRA	Trejbalova_R2 (a)
LTR_2_f (a)	AGTGTTAGTGTGGAGGTTTGATA			
LTR_2_r (a)	CAAAAAACCCAATACAAACAAAAA	248-464	ART2-ART2 – ART3-EXTRA	Trejbalova_R2 (a)
% expected amplicon ( <i>in silico</i> analysis)		59.25		
% amplified		51.4		
<b>NCR</b>				
<b>Round1</b>				
NCR_1_f	TGTGTGATTTGGTAATTAGAGATT			
NCR_1_r	ACTCCCTACTTACCCTACTATAT	572-910	LTR02	Meth56
NCR_2_f	TAGTGTGAAAAATTTTAGTAGTGG			
NCR_2_r	ATTATTTCTTTCCCCTAAC	614-874	M3L7	Meth55
NCR_3_f	TGTATATAAGTAGTGTTTTTTGTGTTGT			
NCR_3_r	TCTAACTCCCTACTTACCCTACTATA	423-914	M2L4	Meth56
<b>Round2</b>				
NCR_1_f	TGTGTGATTTGGTAATTAGAGATT			
NCR_1_r	ATTTCTTTCCCCTAACCTTAAC	572-871	LTR02 -> L2L15	Meth55
NCR_2_f	TGTGAAAAATTTTAGTAGTGG			
NCR_2_r	ATTATTTCTTTCCCCTAAC	617-874	M3L7 -> LTR07	Meth54
NCR_3_f	TGTATATAAGTAGTGTTTTTTGTGTTGT			
NCR_3_r	TAACCTCCCTACTTACCCTACTATATA	423-912	M2L4 -> Z2M2	Meth57
% expected amplicon ( <i>in silico</i> analysis)		80.37		
% amplified		81.94		
<b>ETR</b>				
<b>Round1</b>				
ETR_1_f (b)	AGTGAATAGAGTTAGGTAGGGATATT			
ETR_1_r (b)	AATTCCTAACTCCAATACTATAAAAAA	8336-8644	ETR02	Weber (b)
ETR_2_f	GTAAAAATAGTAAGAAAAAGAAATGAA			
ETR_2_r	ACAATCAAAAATAAATCTCTCAAAC	8171-8557	ETR11	Weber (b)
ETR_3_f	GTGGAGAGAGATAGAGATAGAT			
ETR_3_r	AAAACCCACCTCTCCTC	8434-9000	ETR07	Trejbalova_R1 (a)
<b>Round2</b>				
ETR_1_f (b)	GAGTTAGTAGGGATATTATTATT			
ETR_1_r (b)	TTACAATCAAAAATAAATCTCTCAA	8344-8559	ETR02 -> ETR06	Weber (b)
ETR_2_f	TTGTATTTTTTATAGTGAATAGAGTTAC			
ETR_2_r	ACAATCAAAAATAAATCTCTCAAAC	8323-8557	ETR11 -> ETR10	Weber (b)
ETR_3_f	GTGGAGAGAGATAGAGATAGAT			
ETR_3_r	CCTCCTCTTATACTTCTAAC	8434-8989	ETR07 -> ETR08	Trejbalova_R2 (a)
% expected amplicon ( <i>in silico</i> analysis)		60.32		
% amplified		63.9		
<b>ENV</b>				
<b>Round1</b>				
ENV_1_f	AGTTTTGTTTTTGGGTTTTTG			
ENV_1_r	ATCTTCCACAAACCAAAATTCT	7772-7980	ENV1	Trejbalova_R1 (a)
<b>Round2</b>				
ENV_1_f	AGTTTTGTTTTTGGGTTTTTG			
ENV_1_r	CCTCAATAACCCCTCAACAAATTA	7772-7905	ENV1 -> ENV2	Trejbalova_R2 (a)
% expected amplicon ( <i>in silico</i> analysis)		89.42		
% amplified		100		



**Additional File 1: Primers PCR experiments (continued)****Trejbalova\_R1 (a)**

	Temperature (°C)	Time (min:sec)	Cycles
Initial denaturation	94	5:00	1
Denaturation	94	0:15	
Annealing	65 - 1 / 2 cycles	1:30	20
Elongation	72	1:00	
Denaturation	94	0:15	
Annealing	57	1:30	20
Elongation	72	1:00	
Final elongation	72	7:00	1

**Trejbalova\_R2 (a)**

	Temperature (°C)	Time (min:sec)	Cycles
Initial denaturation	94	5:00	1
Denaturation	94	0:15	
Annealing	58	1:30	20
Elongation	72	1:00	
Denaturation	94	0:15	
Annealing	65	1:30	20
Elongation	72	1:00	
Final elongation	72	7:00	1

**Weber (b)**

	Temperature (°C)	Time (min:sec)	Cycles
Initial denaturation	95	5:00	1
Denaturation	95	1:00	
Annealing	58	2:00	40
Elongation	72	2:00	
Final elongation	72	2:00	1

**Meth**

	Temperature (°C)	Time (min:sec)	Cycles
Initial denaturation	95	5:00	1
Denaturation	95	0:15	
Annealing	Temp + 4	1:30	5
Elongation	72	1:00	
Denaturation	95	0:15	
Annealing	Temp + 2	1:30	5
Elongation	72	1:00	
Denaturation	95	0:15	
Annealing	Temp	1:30	40
Elongation	72	1:30	
Final elongation	72	7:00	1

(a) Trejbalová et al., 2017 (20)

(b) Weber et al., 2014 (353)

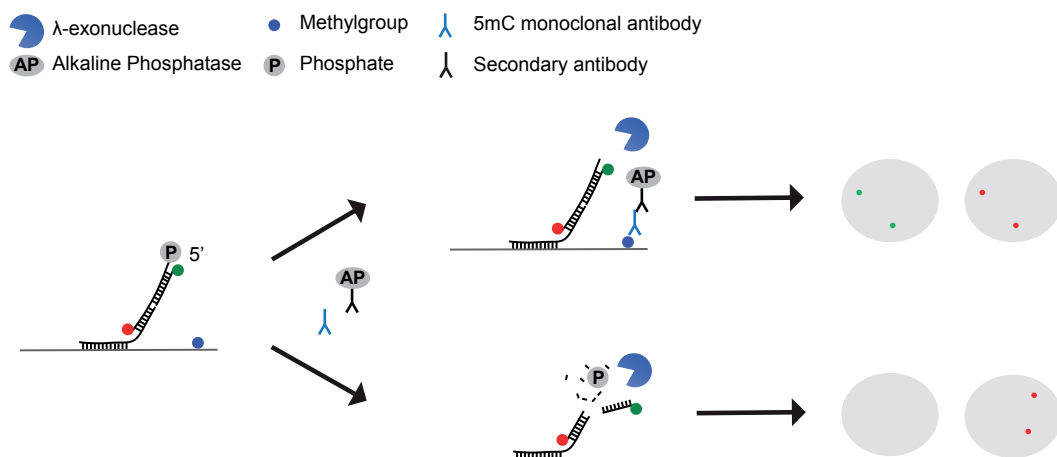




## Objective II – Single cell DNA methylation assay

We developed a single cell targeted DNA methylation assay that provides information about global DNA methylation over a genomic region of 1-5 kb. The region can, but does not need to contain CpGs.

Single cell Epigenetic Visualization assay (EVA) is a method developed in the lab of professor Oleg Denisenko. The principle of the assay is shown in **Figure 31**. In this fluorescent in situ hybridization (FISH)-based technique, the density of an epigenetic mark at a specific region of interest in single cells is quantified using probes, targeting a region of interest. These probes are hybridized to a red, 5'-phosphorylated sensor oligo and a green detector oligo. Alkaline phosphatase-conjugated antibodies recognizing the epigenetic mark of interest (DNA methylation) remove the 5' phosphate from the red sensor oligo, depending on the presence of mC near the probe binding site. This removal results in protection of the oligo against degradation by  $\lambda$ -exonuclease. This degradation can be visualized and used to quantify methylation of the target region.



**Figure 31. The principle of single cell epigenetic visualization assay (EVA).** The 5'-phosphorylated sensor oligo (black, containing red fluorophore) and the detector oligo (black, containing green fluorophore) are tethered to the gene of interest (grey) by a gene-specific oligo (black). In the presence of the epigenetic mark of interest (blue, upper part), the 5'-phosphate will be removed by the alkaline phosphatase (AP) recruited as an antibody conjugate to the epigenetic mark. Dephosphorylated sensor oligo is resistant to the 5' phosphate-dependent  $\lambda$ -exonuclease, yielding green and red signals detected at the locus. In the absence of epigenetic mark,  $\lambda$ -exonuclease degrades sensor oligo to the nearest gap, resulting in the loss of green signal. By measuring the green to red signal ratio, the amount of epigenetic marks at the targeted location can be quantified.

## Single cell epigenetic visualization assay, EVA

The following work is submitted to Nucleic Acids Research for publication on January 22nd, 2020: “**Kint S., Van Crieke W., Vandekerckhove L., De Vos W., Bormzstyk K., Krause D., Denisenko O. Single cell epigenetic visualization assay, EVA**”, and is adapted to fit in this thesis.

### **Contribution of the doctorandus:**

Experimental work – data generation, probe design, image acquisition

Image analysis

Manuscript writing – drafting and editing of text and figures

## Single cell epigenetic visualization assay, EVA

Sam Kint<sup>1,2</sup>, Wim Van Criekinge<sup>1,\*</sup>, Linos Vandekerckhove<sup>2,\*</sup>, Winnok De Vos<sup>3</sup>, Karol Bomzstyk<sup>4</sup>, Diane Krause<sup>5</sup>, Oleg Denisenko<sup>4,#</sup>

\*Authors contributed equally

<sup>1</sup>Department of Data analysis and mathematical modelling, Ghent University, Ghent, Belgium

<sup>2</sup>Department of Medicine, Ghent University, Ghent, Belgium

<sup>3</sup>Department of Veterinary Sciences, University of Antwerp, Antwerp, Belgium

<sup>4</sup>Department of Medicine, University of Washington, Seattle, WA 98109

<sup>5</sup>Yale University, New Haven, CT 06520

#Corresponding author: [odenis@uw.edu](mailto:odenis@uw.edu)

## Abstract

Characterization of epigenetic status of individual cells remains a challenge. Current sequencing approaches have limited coverage and assigning epigenetic status to transcription state of individual gene alleles in the same cell is hardly possible. To address these limitations, a targeted microscopy-based epigenetic visualization assay, EVA, was developed for detection and quantification of epigenetic marks at genes of interest in single cells. The assay is based on *in situ* biochemical reaction between an antibody-conjugated alkaline phosphatase bound to the epigenetic mark of interest, and a 5'-phosphorylated fluorophore-labeled DNA oligo tethered to a target gene by gene-specific oligonucleotides. When epigenetic mark is present at the gene, phosphate group removal by phosphatase protects the oligo from  $\lambda$ -exonuclease activity providing quantitative fluorescent epigenetic readout. We applied EVA to measure 5-methylcytosine (5mC) levels at multiple copy rDNA loci, single copy genes EGR1 and XIST, and the HIV-1 provirus in human cell lines. To link 5mC density to gene transcription, EVA was combined with RNA-FISH. Higher 5mC levels observed at XIST gene promoter on active compared to inactive X chromosomes in female somatic cells validated this approach and demonstrated that EVA can be used to relate epigenetic profiles to the transcription status of individual gene alleles.

## Introduction

Epigenetic programs specify cell phenotypes through covalent modifications of histones and DNA, nucleosome position and density, and substitution by histone variants (25,41,402). While epigenetic alterations drive normal organism development, aberrations in these processes have emerged as hallmarks of cancer and other diseases where particular cell types and individual cell may play critical roles. Conventional approaches used to study epigenetic events, such as chromatin immunoprecipitation (ChIP) and bisulfite sequencing (BS-seq), measure averaged gene epigenetic states in bulk cell populations or tissue fragments and therefore cannot be used to estimate the contribution of individual cells to the epigenetic profile of a specimen.

To assess epigenetic states in individual cells, sequencing-based and imaging-based techniques have been recently introduced for single cell epigenetic analyses (156). A microscopy-based approach has been developed to measure histone modifications at individual genes in single cells by using a proximity ligation assay (ISH-PLA (403)). The advantage of ISH-PLA is that it works *in situ*, however, it involves two enzymatic steps (ligation and rolling circle replication) that yield qualitative rather than quantitative results. Also, this approach works only in a small fraction of cells (<1%). Most of the other epigenetic approaches are focused on genome-wide DNA methylation measurements as these have higher sensitivity compared to histone modification analytical tools. DNA methylation is described as the addition of a methyl group at the C-5 position of Cytosine (5mC) mostly in CpG dinucleotides. There are 28 million CpGs in the human genome and most of them are methylated in somatic cells (404–406). Locus-specific DNA methylation can be involved in gene silencing, which is an important event in processes such as cell differentiation, X-chromosome inactivation, genomic imprinting, and

suppression of repeat sequences (12–14,25,27,42–45). Across genome, the level of methylation is considered to be inversely correlated with CpG density (80,81). Hypermethylation of dense CpG clusters, called CpG-Islands (CpGIs), near promoter regions of mammalian genes is related to transcriptional silencing (13,61,63,82,85,90). CpGI methylation status follows a bimodal distribution, where most of islands are either hypo- or hyper-methylated (81). Less studied methylation of sparse intragenic CpGs is most likely involved in the regulation of cryptic intragenic promoters, alternative splicing and cellular differentiation (98,99,407–409). Generally, levels of intragenic methylation positively correlate with gene transcription rates.

Low sequencing coverage and resolution remain the key limitations of sequencing-based DNA methylation and histone modification analyses in a single cell (156). Also, these approaches cannot be used to assign epigenomic sequencing data to individual gene alleles in a cell, and more importantly, to the gene transcription status. These technologies have also limited application for analysis of rare genomic/epigenomic events, such as HIV-1-infected cells present at frequencies  $10^{-4}$  or lower in infected patients. As a consequence, very large numbers of cells need to be sequenced to cover rare events. To enable simultaneous analysis of epigenetic and transcription states of genes of interest in single cells, we developed a novel Epigenetic Visualization Assay (EVA). This method is based on an *in situ* proximity reaction that generates fluorescent signal proportional to the density of an epigenetic mark at the gene of interest. EVA was tested and validated to quantitate levels of DNA methylation of several target genes in single cells. Combining EVA with RNA FISH allows for simultaneous analysis of DNA methylation and RNA transcription levels at individual gene alleles.

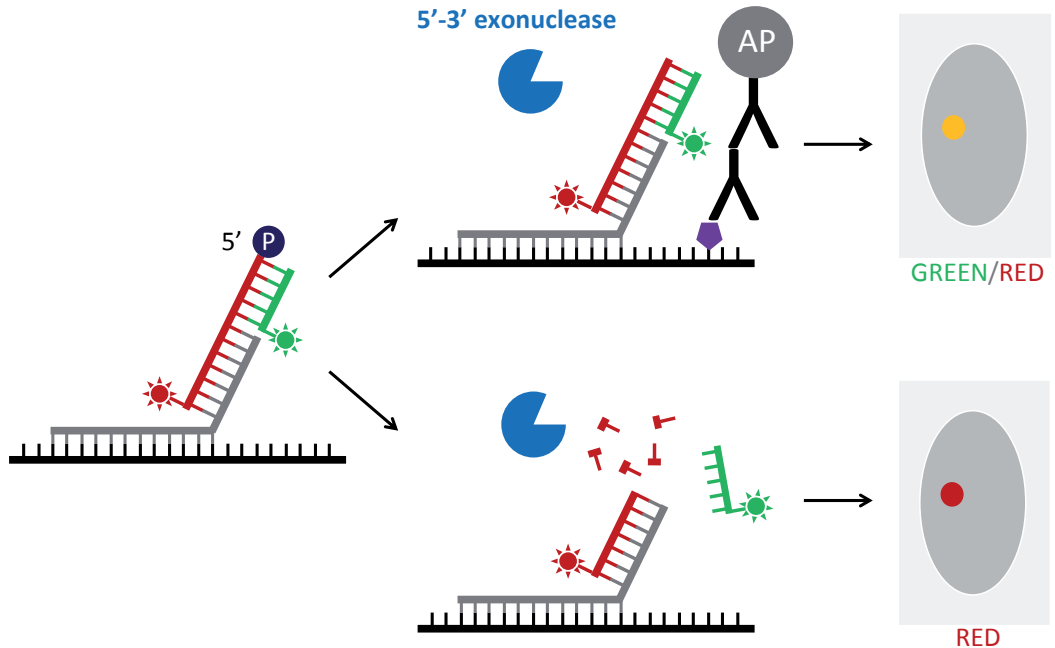


## Results

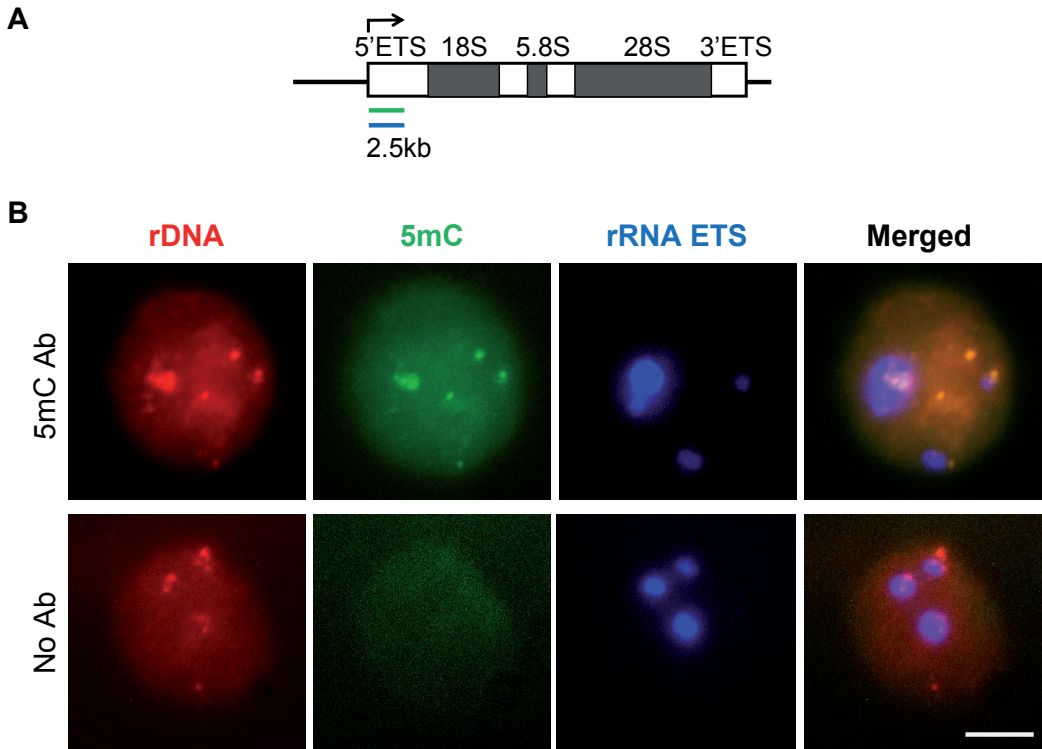
### Epigenetic Visualization Assay

EVA is based on *in situ* proximity reaction where the substrate, 5'-phosphorylated sensor (red) and detector (green) oligonucleotides are tethered to the gene of interest via a series of 30nt-long gene-specific oligos that have a common 20nt sequence added to 3' ends (**Figure 32**). Cognate enzyme, alkaline phosphatase (AP) is recruited as an antibody conjugate to the epigenetic mark. This assay takes advantage of the ability of 5'-3'  $\lambda$ -exonuclease to selectively degrade 5'-phosphorylated strands of double-stranded DNA, with substantially lower activity at un-phosphorylated strands (or single-stranded DNA). Therefore, in the absence of AP (no epigenetic mark),  $\lambda$ -exonuclease is expected to degrade the 5' half of the sensor oligo because it cannot proceed beyond the nick between the green and gene-specific oligos (that have no 5'-phosphate). This event is detected as a low detector/sensor (green/red in **Figure 32**) fluorescent signal ratio. In the presence of AP, 5' phosphate is removed, protecting it from the exonuclease. This event is detected as a high detector/sensor signal ratio. By virtue of the internal signal normalization, the assay thus allows for quantitative analysis of epigenetic marker density at selected genomic sites.

To test this design, we began with the detection of DNA methylation (5-methylcytosine, 5mC) at ribosomal DNA loci, which have, due to the multiple copies of rDNA in mammalian genomes, lower sensitivity requirements. rDNA probe was a mixture of 50 oligonucleotides (50-mers), where 5' 30 nt were specific to the human gene, and 3' 20 nt were common for all oligos and were used to recruit sensor and detector oligos (**Figure 32**) The probe covered 2.5 kb of the beginning of rDNA transcription unit (**Figure 33A**). This probe was hybridized to fixed human HEK-293 cells overnight, then sensor and detector oligos were added. After incubation with 5mC primary and AP- conjugated secondary antibodies followed by washes, phosphorylated red sensor oligo was added to the specimens in buffer supplemented with phosphatase inhibitors to block the reaction between the bound AP and sensor oligo in solution. Specimens treated without 5mC antibody were used as a control. There are 5 chromosome arms in human cells that contain rDNA gene clusters (50-80 gene copies per cluster), therefore we expected to see maximum ~15 foci per cell in HEK-293 cells (hypotriploids with modal chromosome number 64) if they were not clustered. We observed on average ten rDNA foci of diverse sizes per nucleus and all of them also contain the detector green signal which most likely reflects the presence of methylated CpGs at the locus (**Figure 33B**). This suggestion is further supported by the absence of detectable green signal at red foci in control cells treated without 5mC antibodies. No difference in green signal intensity can be seen between rDNA loci that are transcribed (cluster around 5' ETS rRNA transcript, blue in **Figure 33B**) and those that are not transcribed. However, confirmation of specificity of green signal as a readout for 5mC at rDNA gene copies cannot be done by other methods because the methylation status of rDNA copies within and between rDNA clusters is poorly understood. Therefore, to validate EVA, next we analyzed a gene with a known distribution of DNA methylation between the silenced and expressed alleles, XIST.



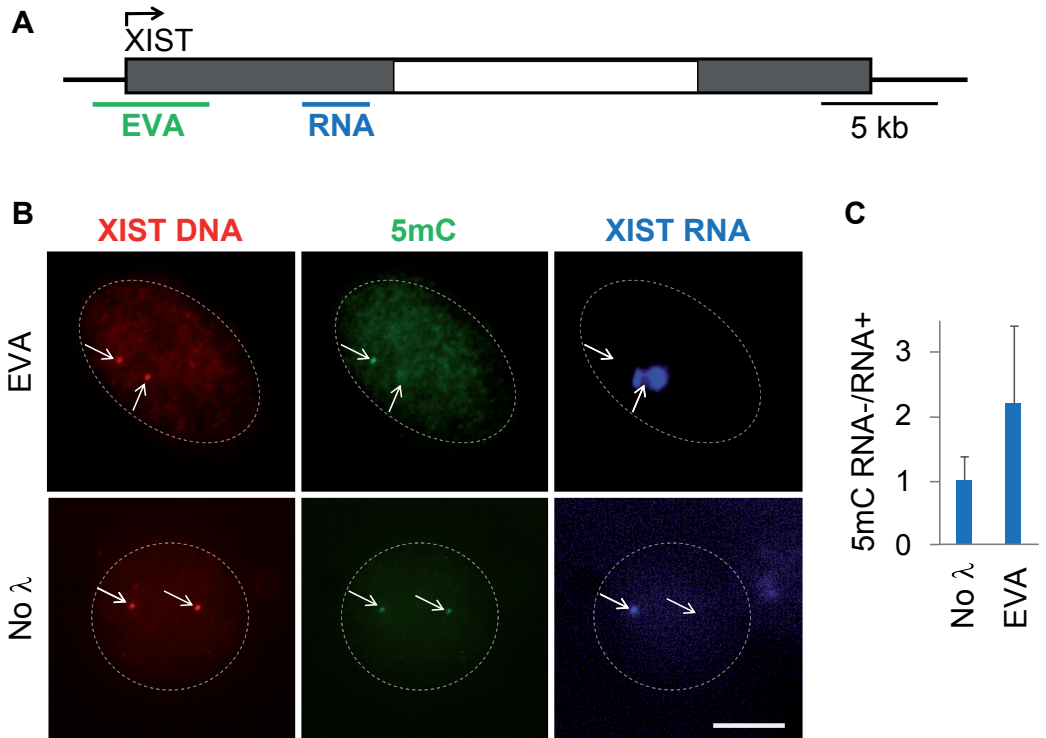
**Figure 32. Epigenetic Visualization Assay, EVA.** Dual-label fluorescent oligonucleotide construct is used as follows. The 5'-phosphorylated sensor oligo (red) and the detector oligo (green) are tethered to the gene of interest (black) by gene-specific oligos (grey). In the presence of the epigenetic mark of interest (purple, upper part), the 5'-phosphate is removed by the alkaline phosphatase (AP) recruited as an antibody conjugate to the epigenetic mark. Dephosphorylated sensor oligo is resistant to the 5' phosphate-dependent  $\lambda$ -exonuclease, yielding green and red signals detected at the locus. In the absence of epigenetic mark,  $\lambda$ -exonuclease degrades phosphorylated sensor oligo to the nearest gap, resulting in the loss of green signal.



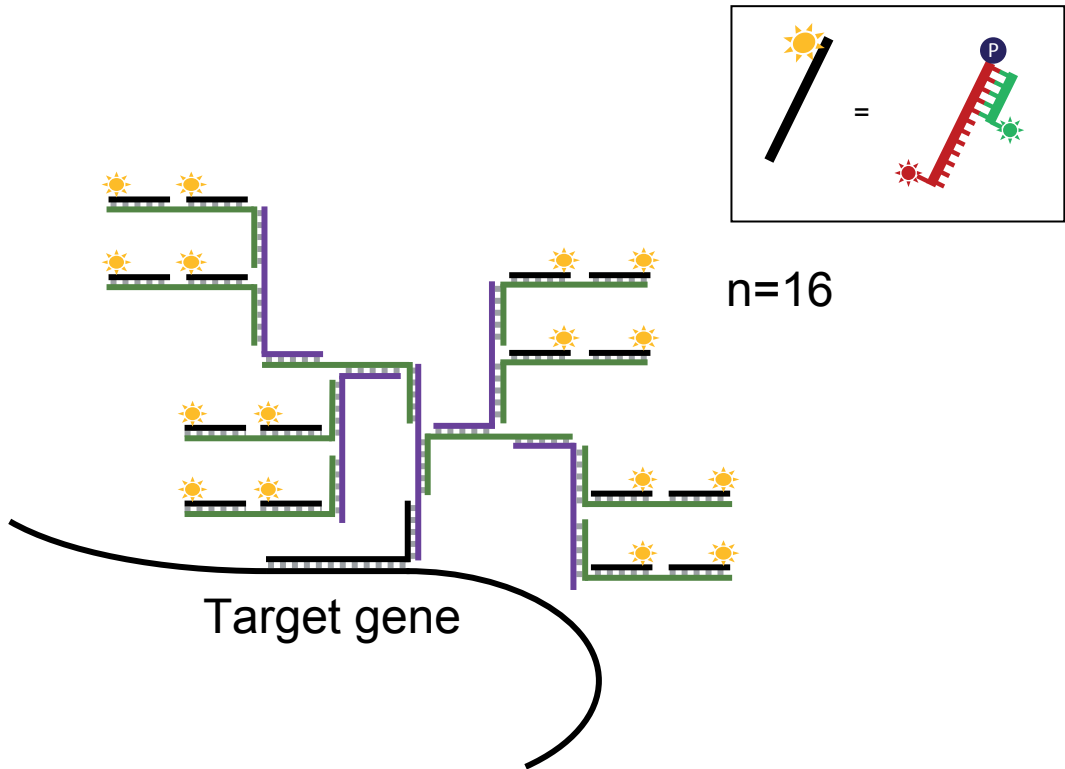
**Figure 33. RNA-EVA analysis of rDNA locus. A.** rDNA transcription unit and probe design. EVA probe (green) consisted of 50 oligos that cover 2.5kb of the 5' External Transcribed Spacer (5' ETS). RNA probe (blue) was to the opposite strand of the same region of 5' ETS. **B.** Representative images of rDNA EVA performed in HEK293 cells showing signals in red (rDNA), green (5mC), and blue (rRNA ETS) channels. The upper panel shows RNA-EVA images (5mC Ab), the lower panel shows a no-antibody control images (No Ab). Projections of z-stack images of representative cells are shown. Scale bar 5 $\mu$ m.

## XIST EVA

In human female somatic cells, the *XIST* gene, located on the inactive X chromosome (Xi), is transcribed and its promoter is hypo-methylated, whereas its copy on the active X chromosome (Xa) is silenced and hyper-methylated (410,411). Non-coding *XIST* transcript binds to Xi and participates in its inactivation (412,413). Thus, experimentally, Xi can be visualized by *XIST* RNA FISH (412). Therefore, we set out to perform combined *XIST* RNA - anti-5mC EVA assay in human cells. We designed an EVA probe to ~4.5 kb region including the gene promoter and part of the first exon. The probe boundaries were determined using Human Methylome Browser DNA methylation maps (<http://neomorph.salk.edu>) as a contiguous stretch of DNA that is half- methylated in female somatic cells compared to the left and right flanking regions. The *XIST* RNA FISH probe consisted of a mix of oligos to the end of exon 1 (both probes are shown in **Figure 34A**). In test experiments, DNA FISH signal obtained with *XIST* EVA probe in HUVEC cells was very weak in some cells and undetectable in others. To make this signal detectable in the majority of cells, we introduced a branched oligo-based signal amplification step (**Figure 35**).



**Figure 34. XIST gene – monoallelic expression.** **A.** Human XIST gene locus and probe design. Grey boxes – exons, white box – intron. EVA probe (green) consisted of 85 oligos that cover the ~5kb which include gene promoter and the beginning of exon 1. RNA probe (blue) was designed to the opposite strand of the end of exon 1. **B.** Representative image of XIST RNA-EVA performed in human female HUVEC. The upper panel shows RNA-EVA images (5mC Ab, EVA), the lower panel shows a no  $\lambda$ -exonuclease control (No  $\lambda$ ). Arrows point to XIST DNA foci (red), Scale bar 5 $\mu$ m. **C.** Quantitative analysis of images. To measure 5mC density at XIST loci, green and red signals were background corrected, and green-to-red signal intensity ratios were taken for XIST RNA- positive (RNA+) and negative (RNA-) foci. The ratio of 5mC signal at RNA- to RNA+ foci in each cell was calculated. Mean values  $\pm$  SD are shown in the graph,  $p < 0.05$ ,  $n = 25$ .



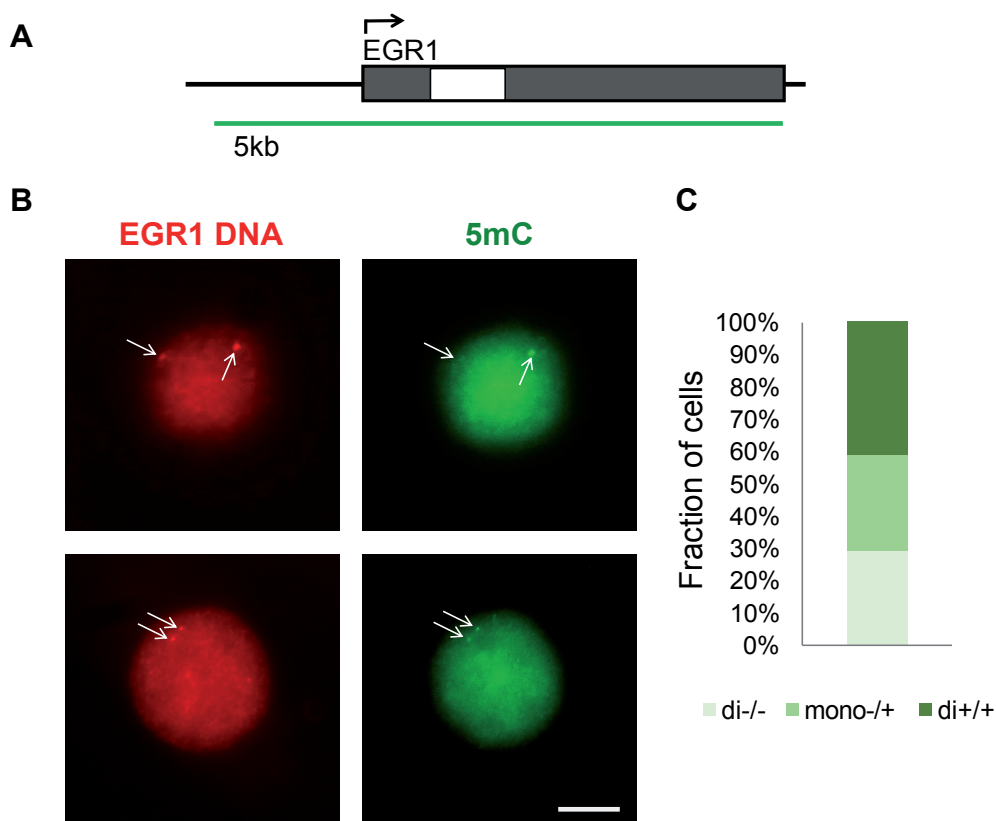
**Figure 35. Signal amplification in EVA.** To amplify EVA signal, two oligos (purple and green) that form a tree-like structure were consecutively used. 16x amplification was sufficient to detect a single copy gene in human cells. Estimated size of the 16x “tree” is 25 nm.

Combined *XIST* RNA-EVA assay in human female cell lines GM-5505 (**Figure 34**) and HUVEC (data not shown) showed a two-fold lower DNA methylation signal (measured as green-to-red signal ratio) at the *XIST* RNA(+) locus compared to the RNA(-) locus ( $p < 0.05$ ) (**Figure 34B** upper panel, and **C**). In cells treated without  $\lambda$ -exonuclease, green signal was equal at both the RNA positive and negative *XIST* foci (**Figure 34B** lower panel), and green signal was undetectable in cells treated without 5mC antibody (not shown). These observations provide further support to the interpretation that green signal quantitatively reflects the density of 5mC at the locus. Some cells revealed more than two *XIST* EVA signals per nucleus, most likely due to DNA replication. Such cells were excluded from the analysis.

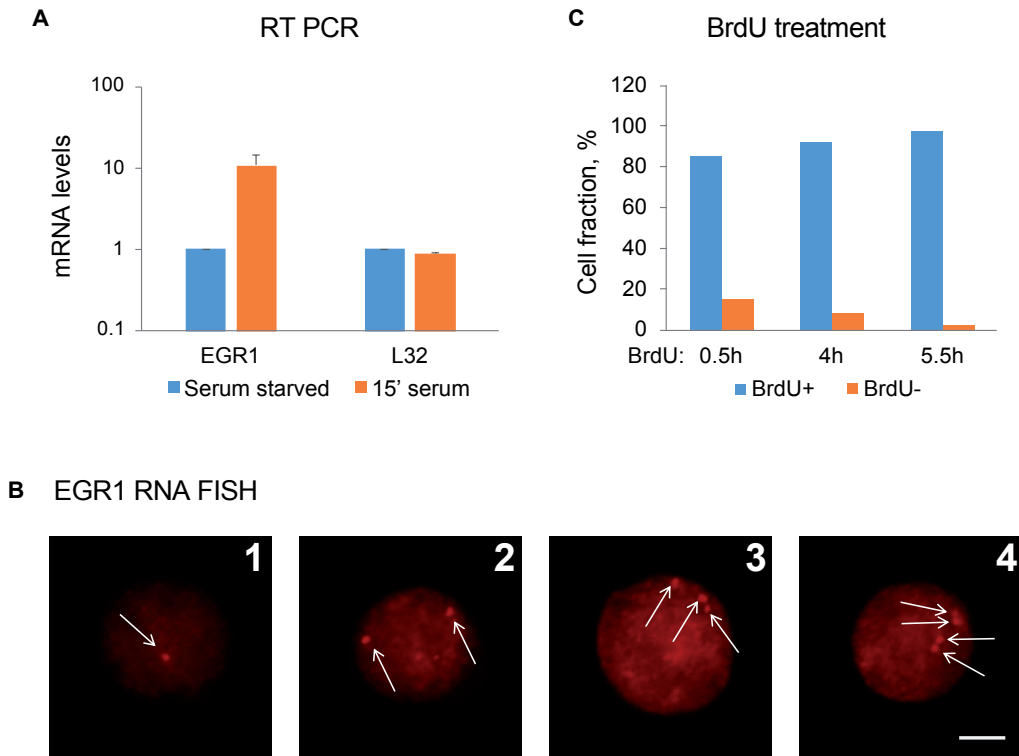
## EGR1 EVA

We applied EVA to examine *EGR1*, a well-studied gene that is transcriptionally inducible by mitogens. The *EGR1* EVA probe covers gene promoter and transcribed regions (**Figure 36A**). As before, results of EVA analysis demonstrated that the green signal was undetectable without antibodies (not shown), but with AP recruited to 5mC, there was detectable green signal (**Figure 36B**). Green

signal was normalized to red signal, providing an estimate of 5mC density at the locus. We found that qualitatively there are three cell subpopulations, (i) cells with both *EGR1* foci containing green signal (green, di+/+), (ii) cells with one of the foci with green signal (light green, mono+/-), and (iii) with both foci without green signal (light-light green, di-/-) (**Figure 36C**). These data show that methylation of *EGR1* alleles is not coordinated with each other. Treatment of these serum-starved cells with mitogenes, such as serum or TPA, rapidly activates *EGR1* transcription. This event can be detected by either RT-PCR or RNA FISH (**Figure 37A and B**) which allows for detection of nascent RNA transcripts at the gene (414). Therefore, the RNA FISH-EVA combined assay can be used to examine kinetics of epigenetic changes at the locus during gene activation (**Figure 38**).

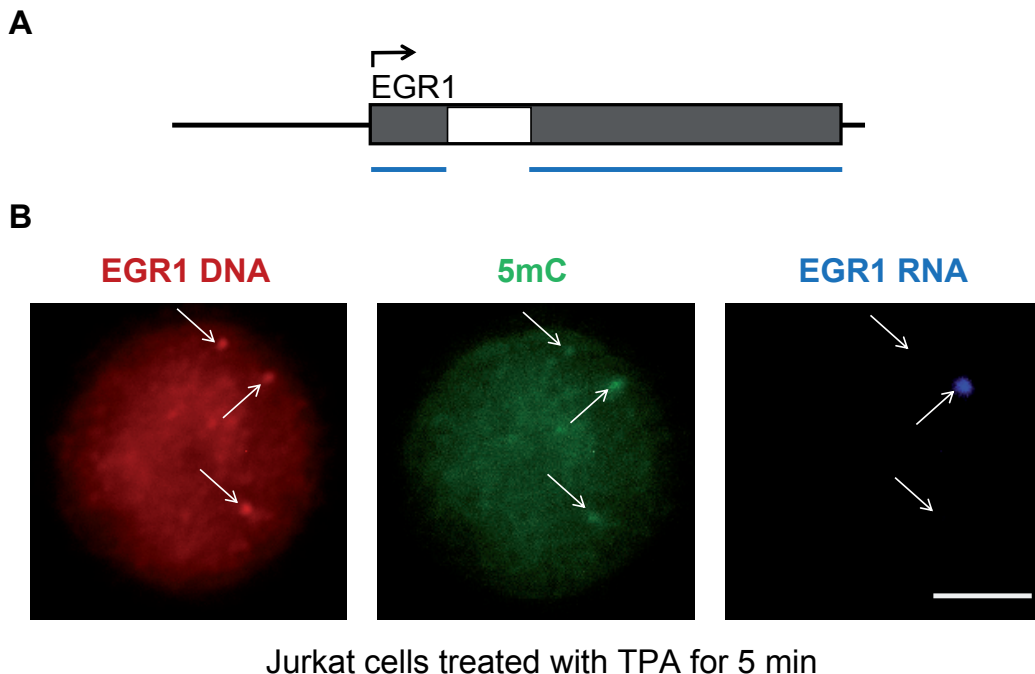


**Figure 36. Single copy *EGR1* gene.** **A.** Human *EGR1* gene locus and probe design. Grey boxes are exons, white box is intron. EVA probe (green) consisted of 96 oligos that cover the 5kb including gene promoter and transcribed regions. **B.** EVA images of *EGR1* locus in Jurkat cells showing DNA methylation in the *EGR1* gene. The upper and lower rows show images of two representative cells. Arrows indicate positions of red foci. Scale bar 5µm. **C.** Fraction of cells with both *EGR1* foci containing green signal (green, di+/+), one of the foci with green signal (light green, mono+/-), and both foci without green signal (light-light green, di-/).



**Figure 37. *EGR1* gene expression in Jurkat cells.** **A.** RT qPCR analysis of *EGR1* transcript levels in Jurkat cells either serum starved (blue bars) or treated with serum for 15 minutes (orange bars). Housekeeping gene *RPL32* expression was used as a control (*L32*). Transcript levels were normalized to  $\beta$ -actin mRNA, fold difference is shown as mean  $\pm$  SD,  $n$ =three independent experiments. **B.** *EGR1* RNA FISH analysis reveals cells with 1 to 4 foci per nucleus in Jurkat cells treated with serum for 15 min. Representative nuclei images are shown. Arrows indicate positions of *EGR1* transcript foci. Scale bar 5 $\mu$ m. **C.** *EGR1* RNA FISH analysis was done in cells treated with BrdU. Jurkat cells were exposed to BrdU for 0.5, 4, or 5.5 hours, then co-stained with anti-BrdU antibody and *EGR1* RNA FISH probe. Graph represents fractions of BrdU (+) (blue bars) and BrdU (-) (orange bars) cells that have 3 or 4 *EGR1* foci. Representative cell images are shown.

In a fraction of Jurkat cells both EVA and RNA FISH revealed more than two foci (3 or 4) (**Figure 37A and B**). To test the hypothesis that additional EVA signals were related to DNA replication during the cell cycle, we used *EGR1* RNA FISH staining combined with 5-bromo-2'-deoxyuridine (BrdU) labeling to detect proliferating Jurkat cells. BrdU is used as a substrate by DNA polymerase and can be detected by antibody staining, providing a readout for cells that undergo DNA replication. This approach allowed for relating *EGR1* foci number to cell cycle phases. This revealed that BrdU(+) cells have 3 or 4 *EGR1* foci (S/G<sub>2</sub> phases), whereas most of the BrdU(-) cells (G<sub>0</sub>/G<sub>1</sub>) had only two foci (**Figure 37C**). These observations support our suggestion that the number of *EGR1* foci reflects cell's DNA replication status. These data also imply that EVA can be used to monitor epigenetic changes during the cell cycle.



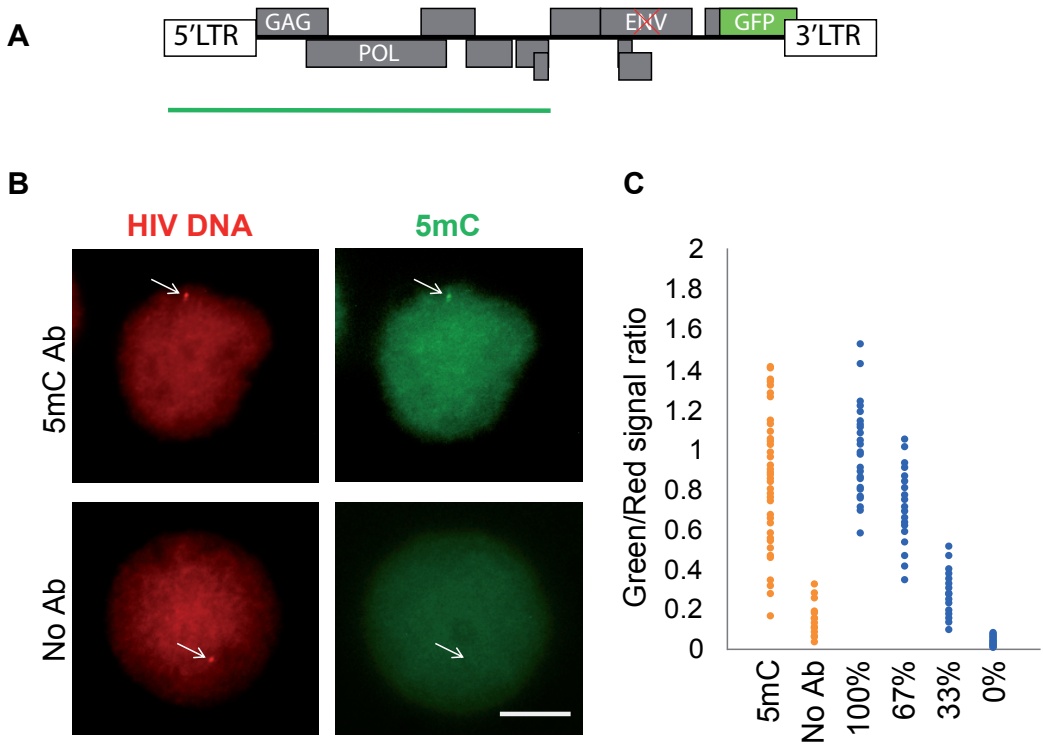
**Figure 38. EGR1 EVA combined with RNA FISH in Jurkat cells treated with TPA.** **A.** EGR1 gene locus and RNA FISH probe design. Grey boxes represent exons, white box is intron. RNA FISH probe (blue) consisted of 48 oligos that cover the 2.5kb that include both exons. **B.** RNA FISH – EVA image of one Jurkat cell treated with TPA for 5 min showing EGR1 locus (red, left panel), DNA methylation (green, middle panel,) and RNA transcript (blue, right panel) signals. Scale bar 5 $\mu$ m.

## HIV provirus EVA

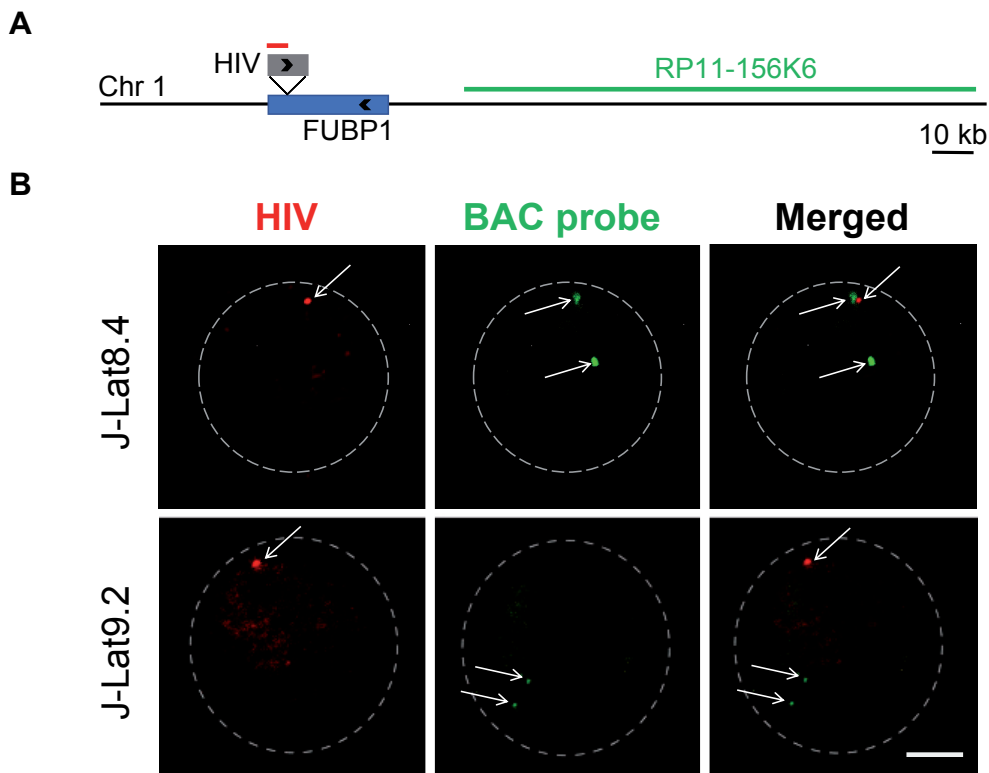
Next, DNA methylation of the HIV-1 provirus integrated into the human genome was examined. In these experiments we used J-Lat 8.4 and J-Lat 9.2 cell lines derived from HIV-1 infected Jurkat cells that were selected to contain one copy of provirus per genome (391). Integrated provirus carries inactivating mutation in *env* gene and *nef* gene was replaced with the GFP gene (**Figure 39A**). The HIV EVA probe was designed to target the 5' half of the virus (5 kb, including LTR region, *gag*, and *pol* genes, **Figure 39A**) and entailed a mix of 85 oligos aligned along the consensus sequence of the HIV-1 genome derived by alignment of all complete HIV-1 genomes available in the database ([www.hiv.lanl.gov](http://www.hiv.lanl.gov)). The HIV RNA FISH probe (48 oligos) was targeting *gag* and *pol* RNA. To test if the HIV EVA signal was specific to the HIV-1 locus, we performed a FISH assay with a mixture of two hybridization probes, one probe to the HIV genome (same as used in EVA) and another BAC clone-based DNA FISH probe to the genomic region adjacent to the insertion site (**Figure 40A**). In J-Lat 8.4 cells, HIV-1 is integrated into the FUBP1 gene on chromosome 1, therefore the adjacent BAC clone RP11-156K6 (143 kb) was used. Co-FISH revealed juxtaposition of the HIV probe focus and one of two BAC DNA FISH foci. Additionally, none of the J-Lat 9.2 cells, in which the HIV-1 genome is integrated elsewhere, showed overlap between the HIV probe and the FUBP-1 BAC FISH



probe (**Figure 40B**). No HIV genome signal was detected in uninfected Jurkat cells (not shown), all together supporting specificity of the HIV DNA probe used in EVA. This probe produced measurable 5mC EVA signal at the HIV locus (**Figure 39B**).



**Figure 39. HIV provirus in latently infected cells.** **A.** HIV locus and the EVA probe (green bar) that covers 5' 5kb region of the HIV-1 genome. ENV gene is mutated (red cross), and NEF gene is replaced with GFP (green box). **B.** Representative EVA Images of HIV locus in J-Lat 8.4 cell line that contains one copy of HIV-1 provirus inserted in FUBP-1 gene. EVA was done with 5mC antibody (upper row) or without antibody (lower row). Arrows indicate positions of red foci. Scale bar 5 $\mu$ m. **C.** Quantitation of DNA methylation levels of the HIV-1 proviral genome and dynamic range of the HIV EVA signal measurements. Orange: green-to-red EVA signal ratios of HIV locus with (5mC) and without 5mC antibody (No Ab). Blue: HIV EVA was done in the same cells without  $\lambda$ -exonuclease treatment to estimate maximum green signal intensity and technical noise. Different proportions of green oligo were mixed with the same unlabeled oligo (0.00, 0.33, 0.67, 1.00 labeled to unlabeled oligo ratios). Each dot represents one cell.

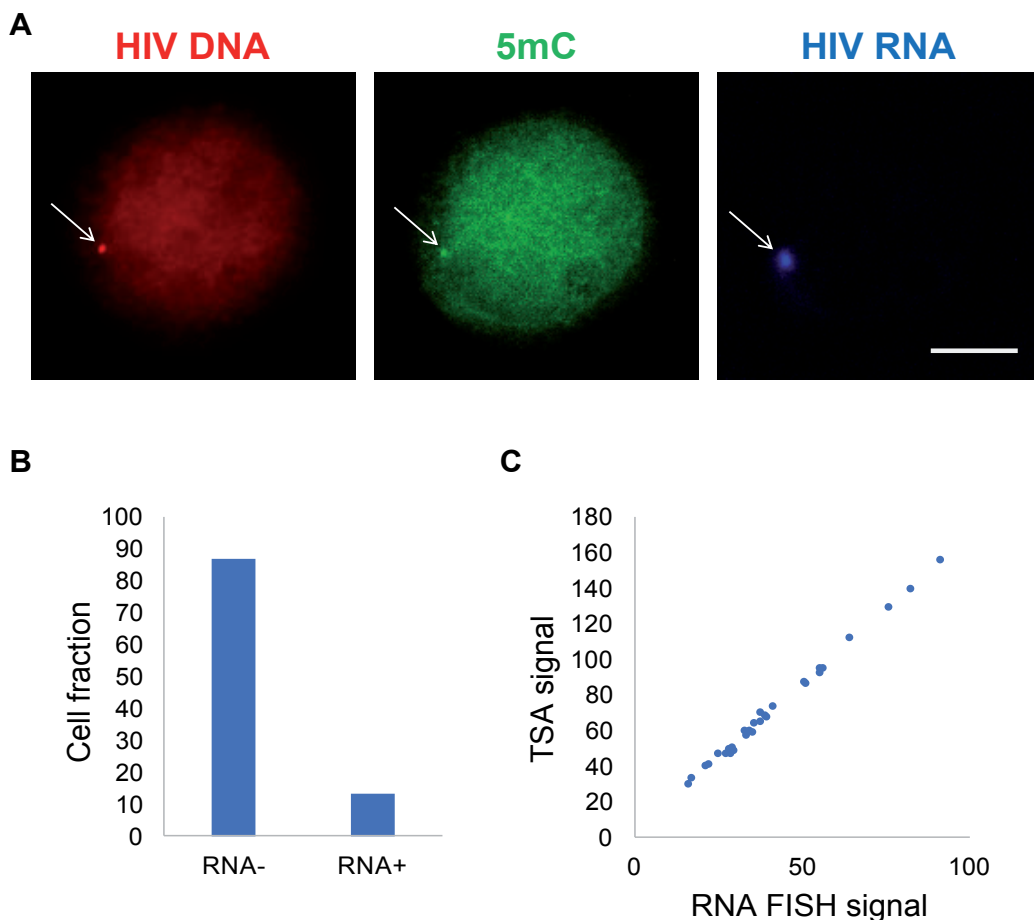


**Figure 40. Testing HIV EVA probe for specificity.** **A.** FISH probe design. In J-Lat 8.4 cells, HIV-1 is integrated in FUBP1 gene. The HIV EVA probe (red) covers 5' 5kb region of the HIV-1 genome, and the BAC probe RP11-156K6 (143 kb) (green) covers the genomic region adjacent to the FUBP1 gene. **B.** Representative images of FISH assay done with HIV probe (red) combined with RP11-156K6 BAC probe (green) in J-Lat 8.4 (upper part) and J-Lat 9.2 (lower part) cells. J-Lat 9.2 cell contain the same virus integrated at different genomic site. Cell nuclei are encircled by dashed lines. Scale bar 5 $\mu$ m.

To estimate the dynamic range of the assay, we performed an EVA experiment where green detector oligomer was mixed with unlabeled oligomer in different proportions and no  $\lambda$ -exonuclease was used. This approach allowed for estimation of the lowest (0%) and highest (100%) possible EVA green-to-red signal ratios. The DNA methylation levels at HIV-1 measured by EVA as estimated by this approach are at the level of 87% of the highest signal that could be obtained with this probe (**Figure 39C**). These results indicate that on average 87% of CpGs are methylated at the HIV locus in J-Lat 8.4 cells, which is close to the levels of HIV-1 DNA methylation in this cell line estimated by using other approaches (86.4% (337)). Importantly, this type of analysis allows one to estimate contribution of technical noise to the cell-to-cell variation in EVA measurements. Thus, the variation in green-to-red signal ratio measured in cells that were treated without  $\lambda$ -exonuclease is due to technical noise, whereas in  $\lambda$ -exonuclease treated cells such variation is a result of both technical noise and biological variation. **Figure 39C** shows that technical noise in HIV-1 DNA methylation

measurements is smaller than EVA data variation. This observation agrees with previously described substantial cell-to-cell variation in HIV-1 DNA methylation detected as a wide variation between bisulfite sequencing reads where each read is derived from one cell (337).

J-Lat cell lines were selected as cells where integrated HIV-1 is transcriptionally silenced (391). Various agents, including TPA, can activate HIV-1 transcription in these cells (61,337). We used RT-PCR and RNA FISH to examine HIV-1 transcription in cells treated with TPA. Unlike rapid EGR1 activation (**Figure 38**), no HIV-1 transcription was detected after 30 min of TPA treatment in J-Lat 8.4 cells. However, after 8 hours of treatment, there was robust RT PCR HIV-1 RNA signal (more than 100-fold induction compared to zero time point, data not shown), whereas RNA FISH detected only ~10% HIV RNA positive cells (**Figure 41A and B**). This observation was confirmed by FACS analysis that revealed 12.8% of GFP-positive cells in the same cells treated with TPA for 8hrs compared to 0.25% at 0 time point (not shown). To enable the assignment of epigenetic changes at the HIV-1 locus to the transcription status of the provirus, we combined EVA assay with RNA FISH. Results of the combined assay (**Figure 41A**) show that we can detect both, HIV-1 transcript (blue) at the provirus locus and DNA methylation (green) at the same locus in cells that transcribe HIV-1. RNA-FISH signal intensity is proportional to the transcript levels at the locus (**Figure 41C**). These data demonstrate that EVA analysis of proviral DNA methylation in relation to its transcription status allows studying the role of DNA methylation (separately within the promoter and gene body) in HIV-1 latency and transcriptional re-activation more directly than with the current bisulfite sequencing-based technologies.



**Figure 41. Combined HIV EVA-RNA FISH assay.** **A.** RNA FISH – EVA image of a representative J-Lat 8.4 cell treated with TPA for 8 hrs showing HIV-1 locus (red, left panel), DNA methylation (green, middle panel), and RNA transcript (blue, right panel) signals. Scale bar 5 $\mu$ m. **B.** Graph represents fractions of cells with detectable HIV RNA FISH signal after 8 hrs of TPA treatment. Results of one representative experiment are shown. **C.** Simultaneous RNA FISH HIV transcript analysis with fluorescein-conjugated oligo probe and TSA in RNA-positive cells treated with TPA for 8 hrs. Fluorescein and TSA signal intensities were measured for each focus and plotted on the graph. These data show that TSA can be used for quantitative analysis of HIV RNA transcript levels in RNA- EVA experiments.

## Method limitations

In EVA, signal intensity is determined by the probe length (number of fluorophore groups). For single copy genes, we currently use mixtures of 85-96 gene-specific oligonucleotides. As a result, epigenetic measurements are averaged for the region covered by EVA probe (currently 4-5 kb), and detected changes reflect coordinated behavior of all CpGs present within this region, masking contribution of individual CpGs. Narrowing the EVA probes down to single CpGs will require additional signal amplification or improvements in hybridization conditions.

For a single copy gene, the EVA signal is amplified using branched oligos (**Figure 35**). Such amplification is associated with increased physical distance between the epigenetic marks and the sensor oligos, and therefore decreased resolution. Currently, signal is amplified 16-fold, and the estimated size of the branched oligonucleotide “tree” is 25 nm, which is comparable to the size of complex between the primary and secondary antibodies, and on the same scale with 30 nm chromatin fibers. Further EVA signal amplification using this approach may not be practical.

EVA is designed for analysis of one epigenetic mark at a time. The same limitation applies to all other available single cell epigenetic technologies. Simultaneous analysis of different epigenetic marks (e.g. DNA hydroxymethylation and methylation) by EVA is feasible by using a different enzyme-substrate pair for each epigenetic mark and additional colors.

## Conclusions

We have developed EVA, an *in situ* proximity biochemical reaction-based method for quantitative analysis of epigenetic marks at genes of interest in a single cell. This method utilizes inexpensive readily available reagents and standard lab equipment, and it is as simple as the conventional FISH assay.

Despite limitations listed above, EVA have several advantages over sequencing-based methods. First, it allows monitoring single gene alleles in a cell. Second, if combined with RNA-FISH, it provides means to link changes in gene expression to epigenetic marks at individual alleles in the same cell. Third, if applied to tissues, such as developing embryo or tumor, EVA measurements can be combined with histological information to advance data interpretation. Finally, this method can also be used to visualize other epigenetic modifications (e.g. histone modifications). The only requirement is to use validated antibodies and to optimize the cell fixation protocol for other epigenetic marks.

## *Methods*

### **Cell culture**

J-Lat clones 8.4 and 9.2 from Dr. Verdin laboratory were obtained through the NIH AIDS Reagent Program, Division of AIDS, NIAID, NIH (391). Jurkat and J-Lat cells were grown in RPMI1640, 10% FBS. Serum starvation was performed by incubating cells overnight in RPMI1640 supplemented with 0.1% FBS. Serum starved cells were treated with 12-O- Tetradecanoylphorbol-13-acetate (TPA, 50 nM) for indicated periods of time, then chilled on ice, washed once with ice cold PBS, and either fixed with cold methacarn solution, or dissolved in Trizol for RNA/DNA isolation.

### **Genomic and RT qPCR**

RNA was isolated using TRIzol reagent (Invitrogen) according to the standard protocol. After cleaning with DNase I (1 U/10µg DNA in 20µl final volume, 15 min at RT) (Epicentre), 1 µg of RNA was reverse transcribed with SuperScript IV Reverse Transcriptase (4 U/µl, Invitrogen) in a 10 µl reaction, 45 min at 37°C. After reverse transcription, cDNA was diluted 100-fold with TE, boiled for 5 min, chilled on ice and quantified using qPCR with primers targeting specific genes of interest (primer sequences are shown in **Table 14**). PCR was performed in triplicates using an in-house qPCR mix. 2 ng cDNA was added to 2.5µl of in-house 2x qPCR mix containing SYBR Green and 500 nM forward and reverse primers, in a final volume of 5 µl. Amplification (three steps, 40 cycles), data acquisition and analysis were carried out using the 7900HT Real Time PCR system and SDS Enterprise Database software (Applied Biosystems). Standard dilutions of genomic DNA (for genomic targets) or dilutions of pooled RT reactions (for cDNA targets) were included in each PCR run. Transcript levels were normalized to both LAMC1 and ribosomal protein RPL32 mRNAs, on which the effect of TPA is minimal.

**Table 14.** Oligonucleotide sequences

	Name	Location	Sequence
1	hLamc1	LAMC1 exon 1	<b>F:</b> CCT TCA ACG TGA CTG TGG TG <b>R:</b> GTC GGC CTG GTT GTT GTA GT
2	hL32	RPL32 exon 2-3	<b>F:</b> AGT TCC TGG TCC ACA ACG TC <b>R:</b> TTG GGG TTG GTG ACT CTG AT
3	hEgr1	EGR1 exon 2	<b>F:</b> ACT CCT CTG TTC CCC CTG CT <b>R:</b> GTC CTG GGA GAA AAG GTT GCT
4	CCB		TGC TAT GGC ATG CTT GAC AAT ATG CTA TGG CAT GCT TGA CAA TAG TTG CGG AAA GCT GAA ACT A
5	DAA		TTG TCA AGC ATG CCA TAG CAT ATA GTT TCA GCT TTC CGC AAC TAT AGT TTC AGC TTT CCG CAA C
6	RF42		5'-P-TTG TCA AGC ATG CCA TAG CAT AGT TGC GGA AAG CTG AAA CTA-3'-HEX315
7	DF		TGC TAT GGC ATG CTT GAC AA-3'-FAM
8	RN1-Bio		TAC AGG TGG GAT TCG GAT TC-3'-Biotin

## Probe design

EVA oligo probes for human genes were designed using sequences obtained from the assembly GRCh38/hg38, UCSC genome browser (<https://genome.ucsc.edu/cgi-bin/hgGateway>). We filtered out repeats and regions with similarity to other genomic sites and designed 50 to 96 perfect matching oligo probes (30-mers) with spaces between them of about 20bp, so that the probe mix covers a genomic region up to 5kbp. All oligos had the same 3' common sequence (TAG TTT CAG CTT TCC GCA AC) attached, to be used for signal detection. A list of gene-specific oligos is available upon request. Probes for the HIV target were designed similarly, using a reference HIV-1 genome consensus sequence derived by alignment of all complete HIV genomes available in the database from the Los Alamos National Laboratory website ([www.hiv.lanl.gov](http://www.hiv.lanl.gov)). Oligonucleotides were synthesized by Eurofins Genomics.

## EVA assay

EVA detection/signal amplification oligonucleotides (**Table 14**) were HPLC-purified and dissolved in water at 100 $\mu$ M.

### **Day 1: hybridization.**

Cells were placed on ice, collected by centrifugation, washed once with cold PBS (20 mM K-phosphate pH 7.2, 150 mM NaCl), and the pellet was resuspended in 1 ml of cold methacarn fixative (three parts of methanol mixed with one part of glacial acetic acid), stored at -80°C. For specimen preparation, 10  $\mu$ l of cell suspension in methacarn were dropped on a 22x22 mm cover slip and allowed to spread. After air drying, cover slips were incubated at 65°C for 10 min, then cooled to RT. 10  $\mu$ l of hybridization solution (2xSSC, 50% formamide, 10% dextran sulfate, 1% Tween-20, oligo mix (96 oligos, 100 nM each)) was pipetted on a glass slide, cover slips (cells down) were placed on top, and sealed with rubber cement. Slides were incubated at 95°C for 3 min, transferred to a humidified Petri dish (10 cm), and incubated at 37°C overnight.

### **Day 2: signal amplification and alkaline phosphatase (AP) treatment.**

Slides were covered by PBS, rubber cement was removed, and coverslips were transferred to a Petri dish with 10 ml of wash solution (2xSSC, 50% formamide, 0.1% Tween-20), and shaken for 30 min at RT. After washing 4 times with TBST (10 mM Tris-HCl pH7.5, 150 mM NaCl, 0.1% Tween-20), coverslips were incubated with 10 ml blocking buffer (5% BSA in TBST) for 30 min at RT. 100  $\mu$ l of blocking buffer with 1:200 anti-5mC monoclonal antibody (clone 33D3) and 150 nM CCB was pipetted on a piece of parafilm in a humidified Petri dish, coverslips were placed on the drop (cells down), and incubated for 30 min at RT. After 3x5 min washes with TBST, cover slips were incubated with 100  $\mu$ l of 150 nM DAA/TBST for 30 min, washed with TBST 3x5 min, incubated in 100  $\mu$ l TBST with 1:200 AP-anti-mouse antibody and 150 nM CCB for 30 min, washed with TBST 3x5 min, incubated with 100  $\mu$ l of 150 nM DAA/TBST for 30 min, washed with TBST 3x5 min. Then coverslips were incubated with 100  $\mu$ l of 150 nM RF42/TBST with phosphatase inhibitors (PI, *p*-nitrophenyl phosphate 30 mM,  $\beta$ -glycerophosphate 10 mM, NaF 10 mM, Na<sub>3</sub>VO<sub>4</sub> 0.1 mM, Na<sub>2</sub>MoO<sub>4</sub> 0.1 mM) for 30 min, washed with TBST/PI 3x5 min, rinsed once with TBST and once with AP buffer (50 mM Tris-HCl pH8.8, 100 mM NaCl, 2.5 mM MgCl<sub>2</sub>), and incubated in AP buffer overnight at RT in the dark.

### **Day 3: exonuclease treatment.**

Cover slips were washed with TBST 3x5 min, incubated with 100  $\mu$ l of 150 nM DF for 30 min, washed 3x5 min with TBST, rinsed with AP buffer, incubate with  $\lambda$ -exonuclease (0.2U/ $\mu$ l) (M0262S, New England Biolabs) (45  $\mu$ l H<sub>2</sub>O, 5  $\mu$ l 10x buffer, 10U exonuclease per coverslip) for 1.5 hrs at RT, washed 3x10 min with TBST, rinsed with Tris-HCl 50 mM pH7.5 with or without DAPI, mounted in 8  $\mu$ l Vectashield (H-1000, Vector) on a slide and sealed with colorless nail polish. Slides were store at 4°C in the dark.



## RNA FISH

10  $\mu$ l of cell suspension in methacarn were dropped on 22x22 mm cover slip and spread. After drying, cover slips were incubated at 65°C for 10 min, then allowed to cool to RT. 10  $\mu$ l of hybe solution (2xSSC, 50% formamide, 10% dextran sulfate, 1% Tween-20, oligo mix (48 oligos to RNA of interest, 100 nM each)) was dropped on a glass slide, cover slips (cells down) were placed on top, and sealed with rubber cement. After air drying, slides were transferred to a humidified Petri dish and incubated at 37°C overnight.

For tyramide signal amplification (TSA), coverslips were washed once for 30 min in hybe washing solution at RT, rinsed 4 times with TBST, incubated with 100  $\mu$ l of 150 nM RN1-Bio for 30 min, washed 3x5 min with TBST, rinsed once with 4xSSC 0.1% Tween 20, incubated with 100  $\mu$ l of Avidin-HRP conjugate (A106, Leinco) in TBST (1:300 dilution) for 45 min. After washes, coverslips were incubated with 100 mM Na-Borate pH8.5, 0.5 mM H<sub>2</sub>O<sub>2</sub>, 1  $\mu$ g/ml Biotin-xx-tyramide (Biotium) (0.1  $\mu$ g/ml for rRNA ETS probe) for 30 min at RT. After washes, coverslips were incubated in 100  $\mu$ l of streptavidin conjugated to desired fluorophore, washed and mounted in Vectashield (H-1000, Vector) with or without DAPI. To combine RNA FISH with EVA, coverslips after TSA reaction were incubated in hybridization washing solution for 1 hour at RT, then excess of solution was drained on paper towel, coverslips were placed on 10  $\mu$ l of hybridization solution containing EVA probe on a slide, sealed with rubber cement, and processed according to the EVA protocol as described above.

## Image analysis

Images were collected using Zeiss Axiovert 200M microscope with 100x objective. The images were analyzed in Fiji image analysis freeware (415) using a dedicated image analysis script (LociDetector.ijm, available upon request). The analysis essentially detects individual loci within single nuclei and calculates for each locus the ratio of signal intensities between two channels (markers). It uses the (smoothened) background signal of the reference channel to segment the nucleus and then exploits a Laplacian spot detector to segment the loci of interest. Per locus, a concentric band of 5 pixels (rim) is generated to correct for local variations in background signal. Then, the signal intensity is measured for each locus and corresponding concentric band per channel, and the ratio of the corrected intensities is calculated as follows:  $[\text{Locus\_Ch2} - \text{Rim\_Ch2}] / [\text{Locus\_Ch1} - \text{Rim\_Ch1}]$ , with Ch1 the reference channel and Ch2 the measurement channel. Before analysis, images were registered (by translation) to correct for chromatic shift.

## BrdU staining

Tumor cell lines cannot be efficiently arrested in G1 phase by serum deprivation (416), therefore Jurkat cell line is expected to continue growth under serum starvation conditions. 10  $\mu$ M BrdU (5-bromo-2'-deoxyuridine) was added to culture medium to serum starved Jurkat cells (0.1% FBS overnight) for 0.5, 4, 5.5 hours (417). For the last 5 min of BrdU treatment, serum was added to 10%.

Cells were chilled on ice, washed in cold PBS, fixed in ice-cold methacarn and stored at -80°C. 10  $\mu$ l of cell suspension in methacarn was pipetted onto a 22x22 mm cover slip and spread. After drying, cover slips were incubated at 65°C for 10 min and cooled to RT. RNA FISH was done with EGR1 probe as described above using TSA and Biotin-xx-tyramide. Cells were incubated in 1 M HCl for 20 min at RT, neutralized in 0.1 M sodium phosphate buffer pH7.4, washed three times in PBS and stained with UltraAvidin DyLight 594 at 1:500 dilution (Leinco) and rabbit anti-BrdU antibody at 1:500 dilution (Rockland) followed by fluorescein-conjugated goat anti-rabbit IgG (Vector). Specimens were counterstained with DAPI and mounted in antifade solution on slides. The number of EGR1 RNA foci (red) per nucleus was recorded for BrdU- positive (green) and negative cells.

## Statistical analysis

An estimated minimum sample size to detect 1.5-fold difference between two samples (for standard deviation 0.5, power 0.8,  $\alpha$  0.05) was 16. To measure differential methylation, EVA signal of at least 25 cells per specimen were analyzed as  $[\text{Locus\_Ch2} - \text{Rim\_Ch2}] / [\text{Locus\_Ch1} - \text{Rim\_Ch1}]$  ratio as described above. These ratios were normalized using the average green-to-red ratio of the negative controls (cells treated without  $\lambda$ -exonuclease). Wilcoxon rank-sum test was used to calculate p-values.

## FACS sorting

J-Lat 8.4 cells were latently infected with an HIV-1 defective virus encoding GFP (391). Cells were either untreated or treated with TPA for 8hrs, chilled on ice, washed with cold PBS, and kept on ice before sorting. FACS Aria II cell sorter (BD Biosciences) was used to sort GFP+ and GFP- J-Lat 8.4 cells. Singlets were selected by forward versus side scatter profiles.



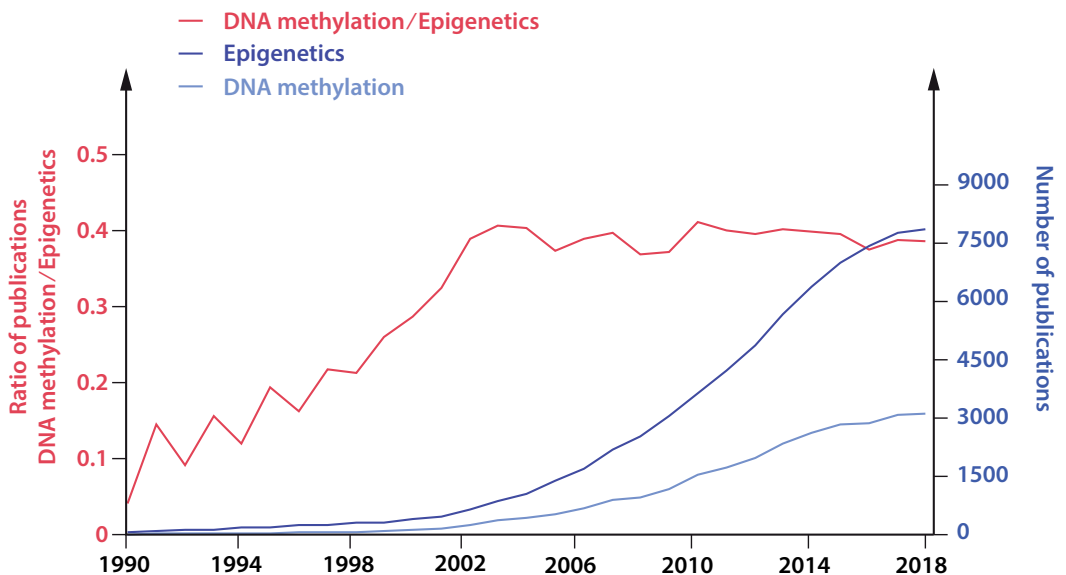
## CHAPTER IV

# DISCUSSION – FUTURE PERSPECTIVES



# IV – Discussion and future perspectives

DNA methylation is first described in 1925 and recognized as an epigenetic modification being involved in the silencing of one of the X-chromosomes in 1975. However, based on information of the publication database Web of Science (Clarivate), the amount of scientific publications with topic 'DNA methylation' did not increase significantly until 1990, the year where Holliday proposed a more molecular definition of epigenetics (**Figure 42**). During the 10 subsequent years, the ratio of publications about DNA methylation to those about epigenetics gradually rose to 40%, and since then, this ratio is stable, with up to 8000 publications about 'epigenetics' per year, from which about 3200 per year about 'DNA methylation' and 'epigenetics'.



**Figure 42. Scientific publications about DNA methylation and epigenetics.** Based on the information found on the publication database Web of Science (Clarivate), a historical overview of the studies with as topic 'DNA methylation' and 'epigenetics'.

For all these studies, different DNA methylation assays have been developed, each having their limitations. The strategies can be divided in chemical methods (usually sodium bisulfite-based), methods using DNA-protein interactions and methods measuring physical differences of mC and C. Some methods are capable of measuring single cell DNA, while others are less sensitive, and need bulk cellular material.

Most of the current methylation analysis methods have technical limitations: the interpretation of the results is often impeded due to (i) need of conversion of the DNA, (ii) bulk DNA needed, hindering cell specific – or even cell type specific – differential methylation analysis, (iii) difficult to combine

with analysis of other linked/related biological pathways (e.g. RNA transcription) and/or (iv) low resolution. Moreover, (pre)treatment of sample, needed for the methylation specific conversion or enrichment of methylated DNA, often leads to DNA losses and fragmentation. This is due to washing steps, harsh conditions of the treatment, insufficient affinity or enzyme activity. Therefore, the amount of starting material needed is often high and makes epigenetic analysis in general more complicated than standard genetic tests. Consequently, these epigenetic analyses are usually not used in routine clinical screening, in contrast to routine genetic screenings: around 1800 DNA methylation markers have been described in several publications, but only 14 (0.8%) DNA methylation-based (cancer) prediction/detection/monitoring platforms have been commercialized in 2018 (110,418).

## Understanding HIV-1 latency: epigenetics

The state of HIV-1 latency can be defined as the transcriptional silencing of replication competent proviral genes caused by multiple transcriptional blocks after the stable integration of proviral DNA into the host genome (275). This proviral silencing is reversible and results in viral rebound upon treatment interruption. Epigenetics is described as a cellular regulatory mechanism that changes the final expression of a locus or chromosome, without changing the underlying sequence by stable, reversible and mitotically inheritable modifications (25–27). Therefore, HIV-1 latency could essentially be described as an epigenetic problem, which involves both host and proviral epigenetic alterations.

Interacting with latency (both reversal or stabilization) is an important study subject in HIV-1 research. LRAs that are currently studied include compounds that affect the epigenetic state of the cell directly (e.g. HDACi, HMTi, DNA methylation inhibitors), while others are relying on T cell activation, which does involve epigenetic changes in the affected cells (320,419). Stimulating latency to induce a deep latent state in infected cells can be obtained through increased epigenetic repression (e.g. HDAC, HMT, HMT enhancer of zeste homolog 2 (EZH2), DNMT) or by guiding HIV-1 integration towards dense non-genic chromatin regions (e.g. LEDGINS) (261,317,335).

A lot of these compounds are found to be effective *in vitro*, but have either limited effects *in vivo*, or do induce side effects (419). These problems are at least partially caused by the untargeted nature of most of these compounds: they change the overall epigenetic state of the cell, resulting in altered proviral, but also host gene transcription. Moreover, it is shown that LRAs act cell type-dependent, indicating that viral transcription is (epigenetically) different regulated per cell type (324). Together, this illustrates that an approach to efficiently interact with the HIV-1 latency should be targeted to specific genomic regions, in specific cell types. Importantly, this strategy should not solely be targeted to the provirus, but also towards the integration site (IS) of the provirus, since it is suggested that the viral integration site and its chromatin density are to be among the main drivers of viral latency (260). Since HIV-1 latency is a multifactorial transcriptional regulatory problem, which involves both viral and human factors, efficient latency interaction strategies will have to be combinatorial approaches.

In order to find effective strategies, researchers need to find clarity in the multiplicity of (epigenetic) modifications that are involved in this complex latency regulation. All these mechanisms are linked, yet all they have their own specific function. Therefore, understanding every single aspect of the latency regulation is of utmost importance. We focused 1 of them: proviral DNA methylation. DNA methylation has been shown to be a stabilizing epigenetic mark, and the *in vivo* effect of proviral methylation, especially intragenic, is not understood. This information should, in the end, be one piece of the (epigenetic) puzzle to understand HIV-1 latency mechanisms, which is another step towards an HIV-1 cure.



## Bisulfite-based HIV-1 proviral DNA methylation assay

To study DNA methylation regulation, the use of cultured cells is proven to be biased: epigenetic modifications are regulated by environmental factors, so the use of *in vitro* models to study DNA methylation does often not represent their role in *in vivo* situations. Therefore, to understand epigenetic regulation in disease, the use of precious patient samples is unavoidable.

Previous *in vivo* HIV-1 studies analyzed the methylation state of the proviral promoter using low throughput bisulfite Sanger sequencing (NGS was only used in one study). (19,20,207,350,353–355). All these studies used relatively small patient cohorts: over 7 studies, 105 patients were analyzed, ranging from 8 to 25 patients per study (19,20,207,350,353–355). Over these studies, different cell types were used: PBMCs (2 studies), total CD4+ T cells (1 study), memory CD4+ T cells (1 study) or resting CD4+ T cells (3 studies). In addition, mostly, only one timepoint per patient was analyzed. The varying nature of all these studies has provided contrasting information, however, based on these studies, it was suggested that latency and cART-induced suppression might have different epigenetic signatures in patients (401). Although, using bisulfite treatment, distinguishing latently infected cells from actively transcribed or replication-deficient proviruses is not possible. These suggestions are based on DNA methylation observations that are being generalized to differences between specific patient groups (e.g. elite controllers, LTNPs, acute seroconverters, non-controllers, patients on therapy), rather than to phenotypical differences between individual patients or cells.

All *in vivo* HIV-1 studies suffer from low proviral target DNA concentration and high proviral mutation rate (401). Combined with the DNA fragmentation, degradation and loss during bisulfite treatment, the starting material of *in vivo* HIV-1 proviral DNA methylation studies is always very scarce. Nevertheless, the use of sodium bisulfite to study DNA methylation is still the preferred method. Pretreatment of DNA with sodium bisulfite alters the DNA sequence (deamination of unmethylated cytosines to uracil), transforming epigenetic differences into genetic differences, enabling the use of general molecular technologies as PCR, cloning or sequencing to analyze the epigenetic profile. This method is easy to perform, no specialized lab equipment is needed, and it provides single base-pair resolution methylation data. We aimed to develop a better bisulfite-based HIV-1 proviral DNA methylation assay that was broadly applicable, more informative, but still as easy to perform as in the previous studies.

Technical difficulties linked to the bisulfite conversion as DNA fragmentation, DNA degradation and conversion efficiency were tackled by an **in-depth comparison** of these parameters in twelve commercially available bisulfite treatment protocols. By using dPCR, this comparison enabled us to select the most appropriate kit for the HIV-1 proviral bisulfite assay. Additional adaptations of the conversion protocol of the most appropriate kit did not significantly decrease the loss of longer fragments. Consequently, the best performing kit (Epitect® Bisulfite Kit (Qiagen)), using the manufacturer's instructions, was selected as the optimal kit to generate as much information as possible out of the precious patient samples.

The high mutation rate and the low prevalence of the proviral DNA sequences in HIV-1 patient samples require an optimal primer design. Primers need to be highly sensitive, HIV-1 specific (to avoid endogenous retroviral sequences to be amplified) and cover as many genomic variants as possible. The former two aspects were thoroughly tested during our rigorous *in vitro* assay optimization. In this optimization, the PCR conditions were optimized, and the performance of the primer combinations was tested in dilution series up to 50 HIV-1 copies per million cells. The latter was countered by developing an *in silico* primer design strategy to select primer combinations by predicting the proportion of patients that do harbor HIV-1 sequences that each combination would target. This strategy was used to design PCR assays to, in contrast to most previous studies, measure both promoter and intragenic methylation. An additional improvement to most previous studies was the use of NGS instead of clonal Sanger sequencing. Moreover, since the starting material for the assay is purified DNA, methylation could be assessed from any cellular reservoir.

By using the assay in a large and well-characterized patient cohort (n=72), we aimed to measure the differential methylation signatures to link specific methylation profiles to transcriptional states, and to get a better understanding of the role of DNA methylation in HIV-1 latency regulation.

### *Limitations of the HIV-1 methylation assay*

#### **Assay development**

The data analysis used in our study was performed using a consensus HIV-1 genome to map the sequenced reads. This consensus genome was created using all full-length sequences from the LANL database, taking into account as much of the genomic variation as possible. However, if sequencing data of every specific patient (e.g. proviral full-length genomic data) was available, this would improve the data processing: these sequences could be used to make a patient-specific consensus genome.

In the current study, for every CpGI, only one single amplicon was analyzed. If a region was targeted with a first primer combination, subsequent combinations were not used. More data would have been generated if all primer combinations would have been used on all samples. This would be a less economic, yet more informative approach.

By measuring the DNA methylation with this bisulfite-based DNA methylation assay, we start from a DNA sample where all RNA is destroyed, and most of the DNA is fragmented or degraded. Therefore, additional assays such as transcriptional activity, integration site analysis, full-length proviral genome sequencing or analysis of other epigenetic modifications can only be performed in parallel starting from separate samples. This hinders interpretation of DNA methylation results obtained by our approach: proviral integration in euchromatin might influence the methylation in the integrated provirus other than integration in heterochromatin. Similarly, a cell might respond differently to a replication-competent provirus compared to a replication-deficient provirus incapable of producing specific antigens.

Due to the DNA fragmentation, information of separate CpGs cannot be traced back to one provirus, which impedes linking promoter- and intragenic methylation directly. Analyzing this link using bisulfite treatment could be possible after limiting dilution of the sample to one HIV-1 sequence per well before performing bisulfite treatment. Subsequent multiplex analysis of the different CpGs could then provide an estimate of the number of full-length sequences, however, this strategy would often not be conclusive due to fragmentation and partial targeting of the genome. This strategy of limited dilution would also drastically decrease the throughput and increase the cost of the assay. Such a whole proviral genome methylation analysis combined with integration site methylation analysis would theoretically be possible using novel technologies (e.g. SMRT sequencing or nanopore sequencing).

### **Study design**

Despite our comprehensive approach (promoter and intragenic DNA methylation) and the large number of patients (n=72), this study design should be improved in order to clarify the exact role of DNA methylation in latency regulation. A next step could be the expansion of the current assay to a full proviral genome DNA methylation assay, measuring all 5 CpGs and non-CpG intragenic methylation. This would provide an even more comprehensive overview.

In our study, patients were only sampled at one timepoint, and only one body compartment (blood) was analyzed. Therefore, we did only provide a snapshot of the DNA methylation profile in the patient, ignoring the dynamics of this modification (both temporal and spatial) (355,401). The use of PBMCs, which is a mix of different infected and non-infected mononuclear blood cells, was perfect to illustrate the sensitivity of the assay: the proportion of infected cells in these samples is lower compared to blood-derived sorted cells (e.g. CD4+ T cell subsets). However, the use of sorted cells, might provide more informative data: different LRAs do have varying effect on CD4+ T cell subsets (324). Therefore, it can be assumed that the (epigenetic) regulation of the proviral transcription varies per cell type. Indeed, in previous HIV-1 proviral methylation studies, conclusions varied for studies using different cell types (memory CD4+ T cells vs resting CD4+ T cells vs PBMCs) (19,20,350,353,354). Longitudinal follow-up of patients and sampling of multiple reservoir and cellular compartments would result in more interpretable data about the methylation dynamics, especially during treatment initiation, interruption or regimen change

# Single cell epigenetic visualization assay

The increasing interest and the recognition of the importance of cell-to-cell (epigenetic) variation within tissues has led to a high number of single-cell '-omics' technologies (156). Single cell epigenetic methods did already reveal new insights in epigenetic regulatory mechanisms, measuring signals that would be masked by being averaged over different cells using bulk cell analysis (107,156). These '-omics' methods are becoming crucial in basic scientific research, however, to study the role of DNA methylation of targeted location in single cells, novel combination methods are needed. These techniques should be capable to link gene methylation signatures directly to RNA transcription, nucleosome occupancy or other epigenetic modifications.

## EVA methodology

To fulfill the need for a targeted, single cell epigenetic assay, the 'epigenetic visualization assay' (EVA) was developed. This fluorescence in situ hybridization (FISH) based epigenetic assay is an improvement compared to the previously developed microscopy-based epigenetic assay ISH-PLA (403). It takes advantage of the phosphate dependency of  $\lambda$ -exonuclease: fluorescent probes that bind in proximity of 5mC are protected from exonuclease activity by 5mC-antibody-alkaline phosphatase conjugates to provide a quantitative fluorescent epigenetic readout. This method is unprecedented and will stimulate the understanding of the role of epigenetic modifications in several research domains drastically. By using a visualization approach to measure the DNA methylation of a specific target in single cells, EVA provides allele-specific epigenetic information. This epi-haplotyping is a big advantage over other DNA methylation analysis methods, especially if it is combined with analysis of RNA transcription, or other epigenetic marks.

## *Versatility of EVA*

The EVA method is suitable to study DNA methylation in all genomic regions, including transposable elements (TE), promoter, intragenic or intergenic regions. Several EVA probes were tested: ribosomal DNA (rDNA) methylation was measured first, providing several signals (5 chromosome arms with rDNA gene clusters in human cells) to proof that this visualization method worked. Subsequently, human genes (*EGR1* and *Xist*) were targeted to measure methylation in the two alleles per cell. Both alleles were measured separately, which enables analysis of allele specific methylation profiles. *Xist* methylation served as a control, with a known distribution of one hypermethylated and one hypomethylated allele in female cells. Finally, HIV-1 methylation, providing only 1 signal per infected cell, was measured. In theory, EVA can be used to study any epigenetic modification by using a primary antibody targeting the modification of interest.

Since the EVA method is based on standard FISH methods and readily available reagents, it can be used in most scientific labs that have standard lab equipment. Moreover, FISH assays can be performed in most cell types, making the applicability of EVA reaching to different research areas

where epigenetics are involved (e.g. development, cancer research, ageing, infectious diseases), both in human samples as in other organisms. An additional advantage of using a FISH-based method, is the possibility to adapt it to or combine it with additional methods: DNA FISH was used to prove that the observed signal was correct (EVA probes targeting HIV-1 + FISH probes targeting the HIV-1 IS). To link an epigenetic state of a gene directly to its transcriptional activity, EVA + RNA FISH can be combined, as successfully performed for the *Xist* gene, where increased 5mC at the promoter is linked to decreased *Xist* expression. To study methylation in infectious diseases such as HIV-1, this method opens perspectives to differentiate the epigenetic differences between latently infecting proviruses (no transcription) and proviruses with active transcription but with (cART-) inhibited viral replication.

In HIV-1 DNA methylation research, this method could be of huge interest. By linking RNA transcription directly to the methylation density, it enables us to measure differential methylation between latently infecting proviruses and actively transcribed proviruses. Additionally, if an individual has a lot of clonal proliferated infected cells, the methylation of the clone of interest can be measured by colocalizing the integration site (DNA FISH) and the provirus (EVA).

### *Future assay improvements*

Several improvements to the EVA method would be desirable to increase the **resolution**, optimize the **cell fixation** and make it suitable for **high-throughput, large-scale** studies.

The resolution of EVA is restricted by the probe size, which depends on the signal intensity. In order to create sufficient signal, the probe spans usually 1-5 kb of the target gene by using 20-96 probes, providing averaged methylation data over this complete region. These probes might target CpG dense regions, often together with CpG deserts (regions with clear CpG suppression). Ideally, to make EVA widely applicable, the resolution should be reduced to the span of a CpGI, which is defined as a region of at least 200 bp in the human genome (61,74). Therefore, alternative, stronger signal amplification method could solve this issue. Expanding the current amplification method (branched oligo tree) might result in oligonucleotides being out of reach of the alkaline phosphatase activity: the current 16x amplification already uses a construct which is has a size comparable to the antibody complex. Alternatively, the use of brighter fluorophores, attachment of other amplification mechanisms to the detection oligos (e.g. Bromo-dU and digoxigenin and run consecutive tyramide signal amplification (TSA)), or the use of a more sensitive microscope could suffice. Additionally, optimization of buffer compositions (e.g. lowering the concentration of dextran sulfate in the hybridization buffer) has already been shown to increase signal intensity.

Next, with the current methodology, the adaptability of the method to measure other epigenetic modifications is only valid for DNA modifications (e.g. hmC, caC, fC): the fixation protocol used (methacarn-based fixation) does not allow for staining histones in the nucleus. Promising steps have already been taken to use EVA on formaldehyde-fixed cells, which allows for both DNA- as histone modification visualization (unpublished data). Moreover, improved fixation could allow analysis of

tissue slides. This would open perspectives to visualize the location-dependent histological DNA methylation information (e.g. difference in the center of a tumor compared to the tumor edges, region around the tumor containing epigenetically altered cells (the so-called 'halo') or the healthy tissue around the tumor). Additionally, cell fixation optimization might be necessary to analyze specific cell types, especially in other organisms (e.g. plant cells that contain cell walls).

In the current EVA protocol, microscopic analysis, signal detection and imaging is performed manually. An automated microscopic signal detection and imaging would eliminate potential subjectivity during imaging. This automatized cell detection would be a crucial element to increase the throughput, and scalability of the assay, and would be necessary to study rare events as HIV-1 proviral genomes in patient cells.

# Impact of the assay development

## Bisulfite-based assay optimization

By comparing twelve commercial bisulfite treatment kits using state-of the art dPCR technologies, we did provide a clear overview of the performance differences of these kits. The dPCR approach allowed us to directly measure the actual loss of intact DNA fragments of specific lengths. Using this workflow provided in **objective Ia**, researchers can easily select the most appropriate bisulfite kit for their application. Our HIV-1 proviral DNA methylation study required the presence of long fragments (>500 bp) in order to be able to discriminate between the 5'LTR and the 3' LTR. Therefore, Epitect® Bisulfite Kit was chosen as the most appropriate kit. However, if a study aims to analyze shorter fragments (e.g. 200 bp), the EZ DNA Methylation Gold™ Kit shows less loss of intact fragments.

The primer design procedure in **objective Ib** that was used for the HIV-1 proviral DNA methylation assay included an *in silico* prediction of the complementarity of primers and an *in vitro* assay optimization using dilution series of target DNA (HIV-1) in background DNA (human). This primer development procedure could serve as a useful workflow for future HIV-1 studies using PCR. Moreover, the approach could be used to better predict the number of patient samples needed by taking into account the proportion of patients that harbor proviruses that are complementary to the primers. Moreover, it would minimize the amount of primer combinations needed for a maximal data output.

Bisulfite sequencing is the most used method to study HIV-1 proviral DNA methylation. Although, in almost all previous studies, low throughput sequencing methods were used to focus only on promoter methylation. We adapted this approach in two ways to obtain a more comprehensive overview of the DNA methylation compared to previous studies. (i) By using an **NGS-based approach**, data output is maximized: all amplicons generated will be sequenced on a high throughput sequencing system. This approach was proposed by LaMere et al. (401) and only used in one other study (355). (ii) Analyzing **both promoter and intragenic DNA methylation** as a biological relevant modification, rather than use the latter as a control of the methylation assessment (which was done in three out of four studies that did include *env* methylation assessment) (20,207,350,353).

## In vivo HIV-1 proviral DNA methylation

To study the role of proviral DNA methylation in the latency regulation (**objective Ic**), we used the optimized assay in 72 patients divided into four groups (LTNP, ET, LT and SRCVs), from which different virological characteristics were previously determined. We found, in line with most of the previous studies on PBMCs, low levels of promoter (LTR) DNA methylation. SRCV showed higher LTR DNA methylation compared to the other groups, and similarly, patients with detectable VL had higher promoter methylation than patients controlling the VL. Although these differences were statistically significant, the low levels of promoter DNA methylation does raise questions about the

relevance of these differences. Moreover, we could not confirm previous correlations between LTR methylation and time of cART (Trejbalova et al.) or time of infection within LTNPs (Palacios et al.). (20,354). Previous studies indicated that the LTR methylation density changes over time, and that LTR methylation as such is not strong enough for transcriptional regulation (19,20,61,350,353–355).

Interestingly, we showed substantially higher intragenic DNA methylation compared to the promoter region of HIV-1. The presence of increased *env* DNA methylation was also observed in previous studies, yet its importance was never studied in depth (20,107,207,350,353). Additionally, we showed that in SRCV, the *env* methylation was lower compared to all other patient groups, indicating that low intragenic methylation is associated with active replicating proviruses. Our data suggest that this intragenic proviral DNA methylation is wrongly ignored in previous studies and should thus be included in future HIV-1 proviral DNA methylation studies.

## EVA development

In **objective II**, a novel method to measure single cell epigenetic signatures at genes of interest was developed: EVA. By measuring methylation density in larger genomic regions (1-5 kb), this method is complementary to the bisulfite-based targeted DNA methylation assays. It measures with lower resolution (the probe size determines the resolution), however, it can be used to link the epigenetic state directly to RNA transcription, and it enables the possibility to study both cell-to-cell epigenetic variability and allele-specific epigenetic variability.

## Insights in HIV-1 latency regulation by DNA methylation

Our data suggest that both promoter and intragenic HIV-1 proviral DNA methylation are involved in the regulation of proviral transcription, but that efficient promoter regulation is inhibited/hindered, resulting in overall low promoter methylation. It can be suggested that increased intragenic methylation and decreased promoter methylation are associated with proviral silencing stability and viral control. However, alternative hypotheses and explanations should be taken into account and additional research will be crucial to reveal the actual role of DNA methylation in latency regulation.

**First**, the observed differences between SRCV and other cohorts could also be explained by the increased number of 2-LTR circles observed in the SRCV. These episomal DNA circles can carry epigenetic modifications, but the regulation is potentially different from gDNA. This hypothesis could easily be tested by analyzing DNA samples where episomal DNA is removed. **Secondly**, controlling the HIV-1 infection will result in a shift in infected cell types. Indeed, the reservoir shifts towards shorter-lived cell populations in elite controllers (253). The epigenetic differences of these cellular subsets might be reflected into the proviral DNA methylation. This is also in line with the observations in Trejbalova et al. (2016) where a shift of proviral DNA methylation over time is shown (20). This hypothesis could be tested by analyzing DNA methylation longitudinally and by selecting specific CD4 cellular subsets on every time point. **Thirdly**, during infection, a selection of proviruses that



are integrated in heterochromatin is observed (shift towards difficult to reactivate proviruses). It is suggested that the chromatin conformation of the IS plays a crucial role in the latency. In this case, it is possible that proviral DNA is in an active epigenetic conformation, but that the transcription is blocked due to the surrounding epigenetic state. This might be tested in patients with a high amount of clonally proliferated infected cells. In these cells, proviral and IS epigenetic state should be measured simultaneously in the same cell, making this hypothesis technically harder to validate (probably impossible using bisulfite treatment). The use of a combination EVA assay to measure DNA methylation (or other epigenetic modifications) of both the provirus as the integration site might be a solution to test this hypothesis. **Finally**, intragenic methylation in the *env* region might affect splicing (it is shown that intragenic methylation is involved in alternative splicing regulation). Hypermethylation of this region might for example inhibit correct splicing of *tat* or *rev*, resulting in downregulation of transcriptional elongation or viral RNA export respectively. This might be tested by simultaneous mRNA sequencing. Although, as described above, bisulfite treatment is hardly combinable with these other tests due to its harsh reaction conditions and DNA destruction. Therefore, EVA combined with RNA FISH targeting mRNA splice junctions could be another strategy to test this hypothesis.

## Epigenetics in HIV-1 (cure) research

The complexity of HIV-1 latency regulation is expressed by both the multiplicity of mechanisms involved and by the different regulation per cell type. Understanding all the pieces of this multifactorial regulation will be crucial to efficiently take the next steps in HIV-1 monitoring and cure: if a specific epigenomic state is linked to a phenotypical outcome such as prolonged time to viral rebound or a stable viral control, this state could serve as a biomarker in patient follow-up. Moreover, interacting with these modifications could result in changing the proviral latency state.

### Shock and lock

Shock and kill or block and lock are two major HIV-1 cure strategies that both involve interaction with HIV-1 latency. Both strategies are theoretically promising, however, they also have their disadvantages. Several shock and kill strategies are being investigated in order to find a sterilizing cure, but none of them succeeds in reactivating the complete reservoir (325,328–334). More active reactivating compounds/shocktails and more effective kill strategies are being developed, but this approach also holds some risks, as the reservoir is spread over the entire body of a patient, including the brains and cerebrospinal fluid. Reactivating provirus in these reservoir components not only is technical hard (getting compounds to cross the blood/brain barrier) but it is also potentially harmful for the patient. The opposite strategy, block and lock, with a functional cure as goal is also highly investigated. In this strategy, replication competent proviruses could still be present in the patients, although their transcription would be stably (epigenetically) suppressed. Since all epigenetic marks are reversible, this strategy holds the risk that on every moment, a single provirus could escape from its latent state, and would fuel a viral rebound.

Therefore, combining both strategies might be a more powerful tool towards an HIV-1 cure. Reactivable proviruses are shocked and the reactivated cells are killed as much as possible. The non-reactivated cells will be good targets to push into a deep latent state. New integrations can be blocked by cART and/or guided towards the hard-to-reactivate (heterochromatin) regions within the host genome (e.g. LEDGINS). Multiple rounds of shock and kill and/or block and lock might be desirable in order to ensure only real, inactivatable deep latent proviruses are left in the body of the patient and to prevent any provirus to escape from its latent state.

### Linking different epigenetic modifications

In this study, we did only focus on one single aspect of the epigenetic regulation of proviral latency: DNA methylation. Foreign/exogenous DNA can be (permanently) silenced by hypermethylation, however, this silencing is in the end a collaboration between different epigenetic mechanisms: it can be mediated by repressor complexes, heterochromatinization and DNA methylation (90). DNA methylation is often seen as a stabilizing epigenetic mark (secondary event that provides long-term stability of an epigenetic state) and is often associated with gene silencing by direct either

hindering binding of TF, or by recruiting MBD proteins, that subsequently recruit other epigenetic reorganization complexes (e.g. NuRD complexes). The methylation pattern itself is also influenced by the presence of other (epigenetic) marks such as lncRNAs or histone modifications (e.g. methylation) (96,420–422). In the case of HIV-1 latency regulation, the role of DNA methylation is not straightforward, and when looking for strategies to intervene with the proviral (epigenetic) silencing, its interaction with these other epigenetic marks cannot be ignored. Indeed, changing the histone code (e.g. methylation, acetylation, crotonylation) is shown to be influencing the viral transcription (320). Moreover, it is shown that several lncRNAs are differentially expressed in HIV-1 infected cells and that mRNA modifications (e.g. m<sup>5</sup>C) affect viral gene expression, making them also a potential target for latency reversal/stimulation (423,424).

Reversal of proviral latency with single epigenetic compounds (e.g. HMT, HDACi, 5-aza-CdR) is shown to be effective to some extent, but in general, it is not strong enough to effectively have significant impact on the reservoir. This illustrates that these epigenetic modifications are all involved in the latency regulation, and that therefore targeting only one of them will not suffice in order to efficiently alter the latency state. Next to studying each of these epigenetic modifications separately, their collaborative mechanisms should also be taken into account. It is indeed shown that there is a synergistic effect of several epigenetic LRAs (337). By understanding every individual epigenetic effect, and its role in the complete latency regulation, the sequence of which epigenetic modifiers should be administered could be optimized to effectively manipulate HIV-1 latency.

With this information, both reversing and stimulating HIV-1 latency into an active or a deep latent state respectively, could be done more efficiently: block and lock strategies could first silence all proviral activity by targeting the more dynamic, less stable modifications before stimulate stabilization modifications. Proviral reactivation on the other hand should start with removal of the stabilizing silencing marks, followed by targeting the other dynamic silencing modifications.

## Targeting epigenetic interventions

Efficient intervening with proviral latency will probably include several epigenetic modulators that alter both host and proviral epigenetic state. A key aspect of these modulators is that they act targeted in order to avoid side effects. Despite the fact that HIV-1 infections do alter the global human epigenetic landscape of a patient, the interventions should only affect infected cells and only change the epigenetic environment of the relevant genomic location (e.g. provirus or IS). Targeting the infected cells might be a technical challenge due to the phenotypical identity of infected and uninfected cells. Targeted epigenetic editing using the clustered regularly interspaced short palindromic repeats (CRISPR)/dead CRISPR-associated protein 9 (CRISPR/dCas9) technology linked to an epigenetic modification compound could be way to avoid or minimize off-target effects of the epigenetic alterations. Moreover, it is suggested that the latency regulation could be cell type-specific, so even cell-type specific delivery of specific epigenetic modifiers could be necessary for an effective intervention strategy. Additionally, all infected cells should be affected by an intervention.

This includes cells in the easily accessible reservoir compartments as blood and lymphoid tissue, but also the harder to reach reservoir compartments (e.g. cerebrospinal fluid or brain for which the blood-brain/blood-cerebrospinal fluid barrier has to be penetrated).

## Changing HIV-1 proviral DNA methylation in vivo

It is shown in cell lines that DNMT inhibition (e.g. 5-Aza-Cdr) does induce some viral transcription. In these cells, the promoter methylation is substantially higher compared to the in vivo data generated by us and other groups. Using demethylation agents in vivo to alter proviral promoter methylation would probably not be very effective. Using demethylation compounds to alter intragenic methylation on the other hand might induce more biological relevant changes. Another strategy might be to induce proviral methylation. High proviral promoter methylation is only observed in cell lines, so the in vivo transcriptional result could go in two directions (activating/repressing).

Passive demethylating agents as 5-aza-Cdr inhibit the maintenance methylation enzyme DNMT1. Therefore, it only affects the methylation state after cell division. Using such compounds would not be advised in HIV-1 interventions, since HIV-1 infected cells are often resting cells. Therefore, it won't change the methylation state rapidly (the expected time for a memory CD4+ T cell to divide is 22 weeks, a naïve CD4+ T cell divides on average every 3.5 years). This would implicate that these cells should also be stimulated to increase proliferation rate. Consequently, it would be advisable to use active methylation/demethylation compounds should be used to alter the DNA methylation in vivo.

Since DNA methylation is often seen as a stabilizing epigenetic mark, in combinational interventions, altering the DNA methylation state will rather be an early phase for a shock strategy, but a late phase in a block/lock.

## DNA methylation research – next steps

To increase sensitivity and resolution of DNA methylation methods, the DNA methylation analysis will probably move towards two directions: **third generation sequencing** (single-molecule real-time (SMRT) sequencing) and **visualization methylation analysis methods**, both eliminating the need of PCR amplification of the DNA.

Third generation sequencing methods as Oxford Nanopore sequencing or Pacific Biosciences (PacBio) SMRT-seq are two technologies capable of real-time DNA (and RNA) sequencing. The methods used differ, but both technologies provide real-time read-out of the sequence. Since only DNA input is needed, this sequencing is possible in samples from all cell types, and single cells or bulk cells can be used as input material. In theory, DNA does not have to be amplified before sequencing, enabling these methods to sequence very long DNA strands (up to 100 kb). By eliminating PCR amplification before sequencing, these methods are capable of discriminating mC (and hmC, caC and fC) from C (425). With improving throughput, accuracy and sensitivity, these methods will be used to measure methylation profiles at specific genomic locations.

HIV-1 represents the ultimate challenge in DNA methylation research, mainly due to its low abundance in patient samples and high genetic variability. These technologies would enable to measure the complete HIV-1 genomic sequence, the surrounding genes (IS) and the methylation profile together. Consequently, these methods will be used to link methylation profiles to specific IS, to replication-competence, and measure the impact of the surrounding epigenetic environment. This would be an ultimate assay to measure impact of cure interventions on the viral reservoir, or to predict a viral rebound.

Visualization methylation analysis methods provide the possibilities to analyze DNA methylation profiles, taking into account the location of cells where specific methylation signatures are observed. These methods could have a targeted approach (e.g. EVA), or a more general approach to measure genome-wide methylation states.

Development of visualization-based DNA methylation assays, that are preferably combined with related analyses (nucleosome occupancy, histone modifications, RNA transcription), could be used to map epigenetic modifications and study their cell-specific regulation. This (co-)localization of epigenetic modifications within cells and cell tissue would be useful in several applications such as in oncology, developmental research or infectious diseases. In oncology, it could measure differences in DNA methylation within the tumor, but also the epigenetically altered tissue around a tumor could be visualized. In developmental research, the effect of DNA methylation on the moment of cellular differentiation could be visualized during embryologic development. In infectious diseases (e.g. HIV-1), the influence of the tissue and cells surrounding the infected cell on proviral methylation can be measured, as well as the effect of the IS of the provirus on its methylation state. Moreover, the influence of an infection on the (general) methylation state of surrounding cells can be visualized.

## Conclusion

Many diverse DNA methylation analysis methods have been developed and optimized in the past years. The number of these methods does not only illustrate the importance of this modification but also the complexity of cellular mechanisms involved in its regulation. The two approaches to measure DNA methylation that we have developed, EVA and the HIV-1 bisulfite-based assay, complement each other in the research of this epigenetic modification. The differences of these two assays reflect the multitude of research goals that can be investigated within the field of DNA methylation: single cell vs bulk analysis; high resolution vs low resolution; direct link to other cellular factors vs very specific DNA methylation analysis.

Our work aimed to unravel the role of DNA methylation in the complex regulatory mechanism the HIV-1 proviral transcription. The observations did highlight the underestimated role of intragenic DNA methylation of the provirus. The main contribution of this work to resolve this mystery is by providing new tools for future research. Using these tools in a well-designed, hypothesis-driven studies will provide important insights in the role of DNA methylation of both the promoter and intragenically in transcriptional regulation in HIV-1 latency, as well as in other research fields involving DNA methylation.

# Bibliography

1. Waddington CH. The Epigenotype. *Endeavour*. 1942;18–20.
2. Jablonka E, Lamb M. The Changing Concept of Epigenetics: *Annals of the New York Academy of Sciences*. *Ann N Y Acad Sci*. 2002;981(January):82–96.
3. Waddington CH. Towards a Theoretical Biology. *Nature*. 1968;218(May):525–7.
4. Holliday R. Mechanisms for the Control of Gene Activity During Development. *Biol Rev*. 1990 Nov;65(4):431–71.
5. Holliday R. Epigenetics: An overview. *Dev Genet*. 1994 Jan;15(6):453–7.
6. Berger SL, Kouzarides T, Shiekhattar R, Shilatifard A. An operational definition of epigenetics. *Genes Dev*. 2009 Apr 1;23(7):781–3.
7. Johnson TB, Coghill RD. Researches on pyrimidines. C111. The discovery of 5-methyl-cytosine in tuberculinic acid, the nucleic acid of the tubercle bacillus. *J Am Chem Soc*. 1925 Nov;47(11):2838–44.
8. Wyatt GR. Recognition and estimation of 5-methylcytosine in nucleic acids. *Biochem J*. 1951 May;48(5):581–4.
9. Doskocil J, Sorm F. Distribution of 5-methylcytosine in pyrimidine sequences of deoxyribonucleic acids. *Biochim Biophys Acta*. 1962 Jun 11;55:953–9.
10. Riggs AD. X inactivation, differentiation, and DNA methylation. *Cytogenet Genome Res*. 1975;14(1):9–25.
11. Holliday R, Pugh JE. DNA modification mechanisms and gene activity during development. *Sci New Ser*. 1975 Jul 23;187(4173):226–32.
12. Kint S, De Spiegelaere W, De Kesel J, Vandekerckhove L, Van Criekinge W. Evaluation of bisulfite kits for DNA methylation profiling in terms of DNA fragmentation and DNA recovery using digital PCR. Albertini E, editor. *PLoS One*. 2018 Jun 14;13(6):e0199091.
13. Attwood JT, Yung RL, Richardson BC. DNA methylation and the regulation of gene transcription. *Cell Mol Life Sci*. 2002 Feb;59(2):241–57.
14. Okano M, Bell DW, Haber DA, Li E. DNA methyltransferases Dnmt3a and Dnmt3b are essential for de novo methylation and mammalian development. *Cell*. 1999;99(3):247–57.
15. Pradhan S, Bacolla A, Wells RD, Roberts RJ. Recombinant human DNA (cytosine-5) methyltransferase. *J Biol Chem*. 1999;274(46):33002–10.
16. Shen H, Laird PW. Interplay between the cancer genome and epigenome. *Cell*. 2013 Mar 28;153(1):38–55.
17. Bednarik DP, Mosca JD, Raj NB. Methylation as a modulator of expression of human immunodeficiency virus. *J Virol*. 1987;61(4):1253–7.
18. Schulze-Forster K, Götz F, Wagner H, Kröger H, Simon D. Transcription of HIV1 is inhibited by DNA methylation. *Biochem Biophys Res Commun*. 1990;168(1):141–7.
19. Blazkova J, Trejbalova K, Gondois-Rey F, Halfon P, Philibert P, Guiguen A, et al. CpG methylation controls reactivation of HIV from latency. *PLoS Pathog*. 2009 Aug;5(8):e1000554.
20. Trejbalová K, Kovářová D, Blažková J, Machala L, Jilich D, Weber J, et al. Development of 5' LTR DNA methylation of latent HIV-1 provirus in cell line models and in long-term-infected individuals. *Clin Epigenetics*. 2016;8(1):19.
21. Gutekunst KA, Kashanchi F, Brady JN, Bednarik DP. Transcription of the HIV-1 LTR is regulated by the density of DNA CpG methylation. *J Acquir Immune Defic Syndr*. 1993;6:541–9.
22. Hayatsu H. The bisulfite genomic sequencing used in the analysis of epigenetic states, a technique in the emerging environmental genotoxicology research. *Mutat Res - Rev Mutat Res*. 2008;659(1–2):77–82.
23. Zhang Y, Li S-K, Yi Yang K, Liu M, Lee N, Tang X, et al. Whole genome methylation array reveals the down-regulation of IGFBP6 and SATB2 by HIV-1. *Sci Rep*. 2015;5:10806.
24. Milavetz BI, Balakrishnan L. Viral epigenetics. *Methods Mol Biol*. 2015;1238:569–96.
25. Goldberg AD, Allis CD, Bernstein E. Epigenetics: A Landscape Takes Shape. *Cell*. 2007;128(4):635–8.
26. Finer S, Holland ML, Nanty L, Rakyán VK. The Hunt for the Epiallele. *Environ Mol Mutagen*. 2011;52:1–11.

27. Reik W. Stability and flexibility of epigenetic gene regulation in mammalian development. *Nature*. 2007;447(7143):425–32.
28. Wolffe AP, Matzke MA. Epigenetics: regulation through repression. *Science* (80- ). 1999;286(5439):481–6.
29. Tsompana M, Buck MJ. Chromatin accessibility: a window into the genome. *Epigenetics Chromatin*. 2014;7(1).
30. Kornberg RD. Chromatin structure: a repeating unit of histones and DNA. *Science*. 1974;184(139):868–71.
31. Luger K, Mäder a W, Richmond RK, Sargent DF, Richmond TJ. Crystal structure of the nucleosome core particle at 2.8 Å resolution. *Nature*. 1997;389(6648):251–60.
32. Richmond TJ, Davey CA. The structure of DNA in the nucleosome core. *Nature*. 2003;423(6936):145–50.
33. Mcghee JD, Felsenfeld G, Eisenberg H. Nucleosome Structure and Conformational-Changes. *Biophys J*. 1980 Jan 28;32(1):261–70.
34. Ng H-H, Bird A. DNA methylation and chromatin modification. *Curr Opin Genet Dev*. 1999 Apr 1;9(2):158–63.
35. Knipe DM, Cliffe A. Chromatin control of herpes simplex virus lytic and latent infection. *Nat Rev Microbiol*. 2008 Mar;6(3):211–21.
36. Li B, Carey M, Workman JL. The Role of Chromatin during Transcription. *Cell*. 2007;128(4):707–19.
37. Peters AHFM, Mermoud JE, O'Carroll D, Pagani M, Schweizer D, Brockdorff N, et al. Histone H3 lysine 9 methylation is an epigenetic imprint of facultative heterochromatin. *Nat Genet*. 2002;30(1):77–80.
38. Decaestecker B. Core regulatory circuitries and epigenetic plasticity in neuroblastoma. 2019.
39. Müller J, Kassis JA. Polycomb response elements and targeting of Polycomb group proteins in *Drosophila*. Vol. 16, *Current Opinion in Genetics and Development*. 2006. p. 476–84.
40. Bauer M, Trupke J, Ringrose L. The quest for mammalian Polycomb response elements: are we there yet? *Chromosoma*. 2016 Jun 9;125(3):471–96.
41. Bernstein BE, Meissner A, Lander ES. The Mammalian Epigenome. *Cell*. 2007 Feb 23;128(4):669–81.
42. Law JA, Jacobsen SE. Establishing , maintaining and modifying DNA methylation patterns in plants and animals. *Nat Rev Genet*. 2010 Mar;11(MARCH):204–20.
43. Denisenko O, Mar D, Trawczynski M, Bomsztyk K. Chromatin changes trigger laminin genes dysregulation in aging kidneys. *Aging (Albany NY)*. 2018 May 29;10(5):1133–45.
44. Yoder JA, Walsh CP, Bestor TH. Cytosine methylation and the ecology of intragenomic parasites. *Trends Genet*. 1997 Aug 1;13(8):335–40.
45. Matzke MA, Mette MF, Aufsatz W, Jakowitsch J, Matzke AJM. Host defenses to parasitic sequences and the evolution of epigenetic control mechanisms. *Genetica*. 1999;107:271–87.
46. Jin Z, Liu Y. DNA methylation in human diseases. *Genes Dis*. 2018;5(1):1–8.
47. Berney M, McGouran JF. Methods for detection of cytosine and thymine modifications in DNA. *Nat Rev Chem*. 2018;2:332–48.
48. Beaulaurier J, Schadt EE, Fang G. Deciphering bacterial epigenomes using modern sequencing technologies. *Nat Rev Genet*. 2018;20:157–72.
49. Haig D. Huddling: Brown Fat, Genomic Imprinting and the Warm Inner Glow. *Curr Biol*. 2008;18(4):174–6.
50. Weber AR, Krawczyk C, Robertson AB, Kuśnierczyk A, Vågbo CB, Schuermann D, et al. Biochemical reconstitution of TET1–TDG–BER-dependent active DNA demethylation reveals a highly coordinated mechanism. *Nat Commun*. 2016 Apr 2;7(1):10806.
51. Sharma G, Sowpati DT, Singh P, Khan MZ, Ganji R, Upadhyay S, et al. Genome-wide non-CpG methylation of the host genome during *M. tuberculosis* infection. *Sci Rep*. 2016 Jul 26;6:25006.
52. Lister R, Pelizzola M, Downen RH, Hawkins RD, Hon G, Tonti-Filippini J, et al. Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature*. 2009 Nov 14;462(7271):315–22.
53. Li Y, Zhu J, Tian G, Li N, Li Q, Ye M, et al. The DNA methylome of human peripheral blood mononuclear cells. *PLoS Biol*. 2010;8(11):e1000533.



54. Jang HS, Shin WJ, Lee JE, Do JT. CpG and non-CpG methylation in epigenetic gene regulation and brain function. *Genes (Basel)*. 2017 May 23;8(6):2–20.
55. Lee J-H, Park S-J, Nakai K. Differential landscape of non-CpG methylation in embryonic stem cells and neurons caused by DNMT3s. *Sci Rep*. 2017 Dec 12;7(1):11295.
56. Yu B, Dong X, Gravina S, Kartal Ö, Schimmel T, Cohen J, et al. Genome-wide, Single-Cell DNA Methylomics Reveals Increased Non-CpG Methylation during Human Oocyte Maturation. *Stem Cell Reports*. 2017 Jul 11;9(1):397–407.
57. Koganti PP, Wang J, Cleveland B, Yao J. 17 $\beta$ -Estradiol Increases Non-CpG Methylation in Exon 1 of the Rainbow Trout (*Oncorhynchus mykiss*) MyoD Gene. *Mar Biotechnol (NY)*. 2017 Aug 3;19(4):321–7.
58. Saikia S, Rehman AU, Barooah P, Sarmah P, Bhattacharyya M, Deka MM, et al. Alteration in the expression of MGMT and RUNX3 due to non-CpG promoter methylation and their correlation with different risk factors in esophageal cancer patients. *Tumor Biol*. 2017 May 4;39(5):101042831770163.
59. Zhang D, Wu B, Wang P, Wang Y, Lu P, Nechiporuk T, et al. Non-CpG methylation by DNMT3B facilitates REST binding and gene silencing in developing mouse hearts. *Nucleic Acids Res*. 2017 Apr 7;45(6):3102–15.
60. Pietrzak M, Rempala GA, Nelson PT, Hetman M. Non-random distribution of methyl-CpG sites and non-CpG methylation in the human rDNA promoter identified by next generation bisulfite sequencing. *Gene*. 2016 Jul 1;585(1):35–43.
61. Chávez L, Kauder S, Verdin E. In vivo, in vitro, and in silico analysis of methylation of the HIV-1 provirus. *Methods*. 2011 Jan;53(1):47–53.
62. Burge C, Campbell AM, Karlin S. Over- and under-representation of short oligonucleotides in DNA sequences. *Proc Natl Acad Sci*. 1992;89(4):1358–62.
63. Bird AP. DNA methylation and the frequency of CpG in animal DNA. *Nucleic Acids Res*. 1980 Apr 11;8(7):1499–504.
64. Sponer J, Gabb HA, Leszczynski J, Hobza P. Base-base and deoxyribose-base stacking interactions in B-DNA and Z-DNA: a quantum-chemical study. *Biophys J*. 1997 Jul 1;73(1):76–87.
65. Bachmann MF, Kopf M. On the Role of the Innate Immunity in Autoimmune Disease: Figure 1. *J Exp Med*. 2001;193(12):F47–50.
66. Janeway CA, Medzhitov R. Innate Immune Recognition. *Annu Rev Immunol*. 2002 Apr 28;20(1):197–216.
67. Krieg AM. A role for Toll in autoimmunity. *Nat Immunol*. 2002 May;3(5):423–4.
68. Singal R, Ginder GD. DNA methylation. *Blood*. 1999 Jun 15;93(12):4059–70.
69. Fryxell KJ, Zuckerkandl E. Cytosine Deamination Plays a Primary Role in the Evolution of Mammalian Isochores. *Mol Biol Evol*. 2000 Sep 1;17(9):1371–83.
70. Sved J, Bird A. The expected equilibrium of the CpG dinucleotide in vertebrate genomes under a mutation model. *Proc Natl Acad Sci U S A*. 1990;87(12):4692.
71. Takata MA, Gonçalves-Carneiro D, Zang TM, Soll SJ, York A, Blanco-Melo D, et al. CG dinucleotide suppression enables antiviral defence targeting non-self RNA. *Nature*. 2017 Oct 27;550(7674):124–7.
72. Weber M, Davies JJ, Wittig D, Oakeley EJ, Haase M, Lam WL, et al. Chromosome-wide and promoter-specific analyses identify sites of differential DNA methylation in normal and transformed human cells. *Nat Genet*. 2005 Aug 10;37(8):853–62.
73. Bird AP. CpG-rich islands and the function of DNA methylation. *Nature*. 1986 May;321(6067):209–13.
74. Gardiner-Garden M, Frommer M. CpG Islands in vertebrate genomes. *J Mol Biol*. 1987 Jul 20;196(2):261–82.
75. Antequera F, Bird A. Number of CpG islands and genes in human and mouse. *Proc Natl Acad Sci U S A*. 1993 Dec 15;90(24):11995–9.
76. Caiafa P, Zampieri M. DNA methylation and chromatin structure: The puzzling CpG islands. *J Cell Biochem*. 2005;94(2):257–65.
77. McClelland M, Ivarie R. Asymmetrical distribution of CpG in an "average" mammalian gene. *Nucleic Acids Res*. 1982;10(23):7865–77.
78. Shenker N, Flanagan JM. Intragenic DNA methylation: implications of this epigenetic mechanism for cancer research. *Br J Cancer*. 2012 Jan 13;106(2):248–53.

79. Takai D, Jones PA. Comprehensive analysis of CpG islands in human chromosomes 21 and 22. *Proc Natl Acad Sci U S A*. 2002 Mar 19;99(6):3740–5.
80. Edwards JR, O'Donnell AH, Rollins RA, Peckham HE, Lee C, Milekic MH, et al. Chromatin and sequence features that define the fine and gross structure of genomic methylation patterns. *Genome Res*. 2010 Jul 1;20(7):972–80.
81. Zhang Y, Rohde C, Tierling S, Jurkowski TP, Bock C, Santacruz D, et al. DNA Methylation Analysis of Chromosome 21 Gene Promoters at Single Base Pair and Single Allele Resolution. Schübeler D, editor. *PLoS Genet*. 2009 Mar 27;5(3):e1000438.
82. Kass SU, Landsberger N, Wolffe a P. DNA methylation directs a time-dependent repression of transcription initiation. *Curr Biol*. 1997;7(3):157–65.
83. Siegfried Z, Eden S, Mendelsohn M, Feng X, Tsuberi BZ, Cedar H. DNA methylation represses transcription in vivo. *Nat Genet*. 1999 Jun 1;22(2):203–6.
84. Antequera F. Cellular and Molecular Life Sciences Structure , function and evolution of CpG island promoters. *Cell Mol life Sci*. 2003;60(3):1647–58.
85. Schorderet DF, Gartler SM. Analysis of CpG suppression in methylated and nonmethylated species. *Proc Natl Acad Sci U S A*. 1992 Feb 1;89(3):957–61.
86. Stelzer Y, Shivalila CS, Soldner F, Markoulaki S, Jaenisch R. Tracing Dynamic Changes of DNA Methylation at Single-Cell Resolution. *Cell*. 2015 Sep 24;163(1):218–29.
87. Soutourina J. Transcription regulation by the Mediator complex. Vol. 19, *Nature Reviews Molecular Cell Biology*. Nature Publishing Group; 2018. p. 262–74.
88. Poss ZC, Ebmeier CC, Taatjes DJ. The Mediator complex and transcription regulation. Vol. 48, *Critical Reviews in Biochemistry and Molecular Biology*. 2013. p. 575–608.
89. Bird AP, Wolffe AP. Methylation-Induced Repression— Belts, Braces, and Chromatin. *Cell*. 1999 Nov 24;99(5):451–4.
90. Cedar H, Bergman Y. Linking DNA methylation and histone modification: patterns and paradigms. *Nat Rev Genet*. 2009 May 1;10(5):295–304.
91. Boyes J, Bird A. DNA methylation inhibits transcription indirectly via a methyl-CpG binding protein. *Cell*. 1991 Mar 22;64(6):1123–34.
92. Hendrich B, Bird A. Identification and characterization of a family of mammalian methyl-CpG binding proteins. *Mol Cell Biol*. 1998 Nov 1;18(11):6538–47.
93. Jones PL, Jan Veenstra GC, Wade PA, Vermaak D, Kass SU, Landsberger N, et al. Methylated DNA and MeCP2 recruit histone deacetylase to repress transcription. *Nat Genet*. 1998 Jun;19(2):187–91.
94. Nan X, Ng H-H, Johnson CA, Laherty CD, Turner BM, Eisenman RN, et al. Transcriptional repression by the methyl-CpG-binding protein MeCP2 involves a histone deacetylase complex. *Nature*. 1998 May;393(6683):386–9.
95. Ng H-H, Zhang Y, Hendrich B, Johnson CA, Turner BM, Erdjument-Bromage H, et al. MBD2 is a transcriptional repressor belonging to the MeCP1 histone deacetylase complex. *Nat Genet*. 1999 Sep;23(1):58–61.
96. Zhang Y, Ng HH, Erdjument-Bromage H, Tempst P, Bird A, Reinberg D. Analysis of the NuRD subunits reveals a histone deacetylase core complex and a connection with DNA methylation. *Genes Dev*. 1999 Aug 1;13(15):1924–35.
97. Wade PA, Geggion A, Jones PL, Ballestar E, Aubry F, Wolffe AP. Mi-2 complex couples DNA methylation to chromatin remodelling and histone deacetylation. *Nat Genet*. 1999 Sep;23(1):62–6.
98. Shukla S, Kavak E, Gregory M, Imashimizu M, Shutinoski B, Kashlev M, et al. CTCF-promoted RNA polymerase II pausing links DNA methylation to splicing. *Nature*. 2011 Nov 3;479(7371):74–9.
99. Jeziorska DM, Murray RJS, De Gobbi M, Gaentzsch R, Garrick D, Ayyub H, et al. DNA methylation of intragenic CpG islands depends on their transcriptional activity during differentiation and disease. *Proc Natl Acad Sci U S A*. 2017 Sep 5;114(36):E7526–35.
100. Neri F, Rapelli S, Krepelova A, Incarnato D, Parlato C, Basile G, et al. Intragenic DNA methylation prevents spurious transcription initiation. *Nature*. 2017 Mar 22;543(7643):72–7.
101. Jones PA. Functions of DNA methylation: islands, start sites, gene bodies and beyond. *Nat Rev Genet*. 2012 Jul 29;13(7):484–92.
102. Teissandier A, Bourc'his D. Gene body DNA methylation conspires with H3K36me3 to preclude aberrant transcription. *EMBO J*. 2017 Jun 1;36(11):1471–3.

103. Almamun M, Kholod O, Stuckel AJ, Levinson BT, Johnson NT, Arthur GL, et al. Inferring a role for methylation of intergenic DNA in the regulation of genes aberrantly expressed in precursor B-cell acute lymphoblastic leukemia. *Leuk Lymphoma*. 2017 Sep 2;58(9):2156–64.
104. Segal E, Widom J. What controls nucleosome positions? *Trends Genet*. 2009 Aug 1;25(8):335–43.
105. Lorincz MC, Dickerson DR, Schmitt M, Groudine M. Intragenic DNA methylation alters chromatin structure and elongation efficiency in mammalian cells. *Nat Struct Mol Biol*. 2004 Nov 3;11(11):1068–75.
106. Tufarelli C, Stanley JAS, Garrick D, Sharpe JA, Ayyub H, Wood WG, et al. Transcription of antisense RNA leading to gene silencing and methylation as a novel cause of human genetic disease. *Nat Genet*. 2003 Jun 5;34(2):157–65.
107. Hu Y, Huang K, An Q, Du G, Hu G, Xue J, et al. Simultaneous profiling of transcriptome and DNA methylome from a single cell. *Genome Biol*. 2016 Dec 5;17(1):88.
108. Laporte M, Le Luyer J, Rougeux C, Dion-Côté AM, Krick M, Bernatchez L. DNA methylation reprogramming, TE derepression, and postzygotic isolation of nascent animal species. *Sci Adv*. 2019 Oct 16;5(10).
109. Huff JT, Zilberman D. Dnmt1-independent CG methylation contributes to nucleosome positioning in diverse eukaryotes. *Cell*. 2014 Mar 13;156(6):1286–97.
110. Bhattacharjee R, Moriam S, Umer M, Nguyen N-T, Shiddiky MJA. DNA methylation detection: recent developments in bisulfite free electrochemical and optical approaches. *Analyst*. 2018;143(20):4802–18.
111. Kuo KC, McCune RA, Gehrke CW, Midgett R, Ehrlich M. Quantitative reversed-phase high performance liquid chromatographic determination of major and modified deoxyribonucleosides in DNA. *Nucleic Acids Res*. 1980 Oct 24;8(20):4763–76.
112. Kurinomaru T, Kurita R. Bisulfite-free approaches for DNA methylation profiling. *Anal Methods*. 2017 Mar 9;9(10):1537–49.
113. Booth MJ, Raiber E-A, Balasubramanian S. Chemical Methods for Decoding Cytosine Modifications in DNA. *Chem Rev*. 2015 Mar 25;115(6):2240–54.
114. Ben-Hattar J, Jiricny J. Effect of cytosine methylation on the cleavage of oligonucleotide duplexes with restriction endonucleases HpaII and MspI. *Nucleic Acids Res*. 1988 May 11;16(9):4160–4160.
115. Hatada I, Fukasawa M, Kimura M, Morita S, Yamada K, Yoshikawa T, et al. Genome-wide profiling of promoter methylation in human. *Oncogene*. 2006 May 9;25(21):3059–64.
116. Umer M, Herceg Z. Deciphering the epigenetic code: an overview of DNA methylation analysis methods. *Antioxid Redox Signal*. 2013 May 20;18(15):1972–86.
117. Kubik G, Summerer D. TALEored Epigenetics: A DNA-Binding Scaffold for Programmable Epigenome Editing and Analysis. *ChemBioChem*. 2016 Jun 2;17(11):975–80.
118. Bogdanove AJ, Voytas DF. TAL Effectors: Customizable Proteins for DNA Targeting. *Science (80- )*. 2011 Sep 30;333(6051):1843–6.
119. Boch J, Bonas U. Xanthomonas AvrBs3 Family-Type III Effectors: Discovery and Function. *Annu Rev Phytopathol*. 2010 Jul 5;48(1):419–36.
120. Kriukienė E, Labrie V, Khare T, Urbanavičiūtė G, Lapinaitė A, Koncvičius K, et al. DNA unmethylome profiling by covalent capture of CpG sites. *Nat Commun*. 2013 Oct 23;4(1):2190.
121. Shendure J, Balasubramanian S, Church GM, Gilbert W, Rogers J, Schloss JA, et al. DNA sequencing at 40: past, present and future. *Nature*. 2017 Oct 11;550(7676):345–53.
122. Eid J, Fehr A, Gray J, Luong K, Lyle J, Otto G, et al. Real-time DNA sequencing from single polymerase molecules. *Science*. 2009 Jan 2;323(5910):133–8.
123. Flusberg BA, Webster DR, Lee JH, Travers KJ, Olivares EC, Clark TA, et al. Direct detection of DNA methylation during single-molecule, real-time sequencing. *Nat Methods*. 2010 Jun;7(6):461–5.
124. Jain M, Koren S, Miga KH, Quick J, Rand AC, Sasani TA, et al. Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nat Biotechnol*. 2018 Apr 29;36(4):338–45.
125. Bayley H. Nanopore sequencing: from imagination to reality. *Clin Chem*. 2015 Jan 1;61(1):25–31.
126. Wang T, Hong T, Tang T, Zhai Q, Xing X, Mao W, et al. Application of N -Halogeno- N -sodiobenzenesulfonamide Reagents to the Selective Detection of 5-Methylcytosine in DNA Sequences. *J Am Chem Soc*. 2013 Jan 30;135(4):1240–3.

127. Münzel M, Lercher L, Müller M, Carell T. Chemical discrimination between dC and 5Me dC via their hydroxylamine adducts. *Nucleic Acids Res.* 2010 Nov 1;38(21):e192–e192.
128. Frommer M, McDonald LE, Millar DS, Collis CM, Watt F, Grigg GW, et al. A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual DNA strands. *Proc Natl Acad Sci U S A.* 1992;89(5):1827–31.
129. Hayatsu H, Wataya Y, Kai K, Iida S. Reaction of sodium bisulfite with uracil, cytosine, and their derivatives. *Biochemistry.* 1970 Jul;9(14):2858–65.
130. Shapiro R, Servis RE, Welcher M. Reactions of Uracil and Cytosine Derivatives With Sodium Bisulfite. A Specific Deamination Method. *J Am Chem Soc.* 1970 Jan;92(2):422–4.
131. Shapiro R, Braverman B, Louis JB, Servis RE. Nucleic Acid Reactivity and Conformation. *J Biol Chem.* 1973;248(11):4060–4.
132. Wang RYH, Gehrke CW, Ehrlich M. Comparison of bisulfite modification of 5-methyldeoxycytidine and deoxycytidine residues. *Nucleic Acids Res.* 1980;8(20):4777–90.
133. Hayatsu H, Shiragami M. Reaction of bisulfite with the 5-hydroxymethyl group in pyrimidines and in phage DNAs. *Biochemistry.* 1979 Feb;18(1974):632–7.
134. Susan JCI, Harrison J, Paul CL, Frommer M. High sensitivity mapping of methylated cytosines. *Nucleic Acids Res.* 1994;22(15):2990–7.
135. Herman JG, Graff JR, Myohanen S, Nelkin BD, Baylin SB. Methylation-specific PCR: a novel PCR assay for methylation status of CpG islands. *Proc Natl Acad Sci.* 1996;93(18):9821–6.
136. Eads CA, Danenberg KD, Kawakami K, Saltz LB, Blake C, Shibata D, et al. MethyLight: a high-throughput assay to measure DNA methylation. *Nucleic Acids Res.* 2000 Apr 15;28(8):32e – 0.
137. Brena RM, Plass C. Bio-COBRA: Absolute Quantification of DNA Methylation in Electrofluidics Chips. In *Humana Press*; 2009. p. 257–69.
138. Tost J, Gut IG. Analysis of Gene-Specific DNA Methylation Patterns by Pyrosequencing. In: *Pyrosequencing Protocols.* New Jersey: Humana Press; 2007. p. 89–102.
139. Warnecke PM, Stirzaker C, Song J, Grunau C, Melki JR, Clark SJ. Identification and resolution of artifacts in bisulfite sequencing. *Methods.* 2002 Jun 1;27(2):101–7.
140. Paul CL, Clark SJ. Cytosine Methylation: Quantitation by Automated Genomic Sequencing and GENESCAN™ Analysis. *Biotechniques.* 1996 Jul 2;21(1):126–33.
141. Lister R, O'Malley RC, Tonti-Filippini J, Gregory BD, Berry CC, Millar AH, et al. Highly Integrated Single-Base Resolution Maps of the Epigenome in Arabidopsis. *Cell.* 2008 May 2;133(3):523–36.
142. Cokus SJ, Feng S, Zhang X, Chen Z, Merriman B, Haudenschild CD, et al. Shotgun bisulfite sequencing of the Arabidopsis genome reveals DNA methylation patterning. *Nature.* 2008 Mar 17;452(7184):215–9.
143. Feil R, Charlton J, Bird AP, Walter J, Reik W. Methylation analysis on individual chromosomes: improved protocol for bisulfite genomic sequencing. *Nucleic Acids Res.* 1994 Feb 25;22(4):695–6.
144. Berman BP, Weisenberger DJ, Aman JF, Hinoue T, Ramjan Z, Liu Y, et al. Regions of focal DNA hypermethylation and long-range hypomethylation in colorectal cancer coincide with nuclear lamina-associated domains. *Nat Genet.* 2012 Jan 27;44(1):40–6.
145. Karemaker ID, Vermeulen M. Single-Cell DNA Methylation Profiling: Technologies and Biological Applications. *Trends Biotechnol.* 2018;36(9):952–65.
146. Meissner A, Gnirke A, Bell GW, Ramsahoye B, Lander ES, Jaenisch R. Reduced representation bisulfite sequencing for comparative high-resolution DNA methylation analysis. *Nucleic Acids Res.* 2005;33(18):5868–77.
147. Brinkman AB, Simmer F, Ma K, Kaan A, Zhu J, Stunnenberg HG. Whole-genome DNA methylation profiling using MethylCap-seq. *Methods.* 2010;52(3):232–6.
148. Gu H, Smith ZD, Bock C, Boyle P, Gnirke A, Meissner A. Preparation of reduced representation bisulfite sequencing libraries for genome-scale DNA methylation profiling. *Nat Protoc.* 2011 Mar;6(4):468–81.
149. Worm J, Aggerholm A, Guldborg P. In-tube DNA methylation profiling by fluorescence melting curve analysis. *Clin Chem.* 2001;47(7):1183–9.

150. Wojdacz TK, Dobrovic A. Methylation-sensitive high resolution melting (MS-HRM): a new approach for sensitive and high-throughput assessment of methylation. *Nucleic Acids Res.* 2007 Mar 1;35(6):e41–e41.
151. Dobrovic A, Bianco T, Tan LW, Sanders T, Hussey D. Screening for and analysis of methylation differences using methylation-sensitive single-strand conformation analysis. *Methods.* 2002 Jun 1;27(2):134–8.
152. Tanaka K, Okamoto A. Degradation of DNA by bisulfite treatment. *Bioorg Med Chem Lett.* 2007;17(7):1912–5.
153. Grunau C, Clark SJ, Rosenthal A. Bisulfite genomic sequencing: systematic investigation of critical experimental parameters. *Nucleic Acids Res.* 2001;29(13):E65–5.
154. Raizis AM, Schmitt F, Jost JP. A bisulfite method of 5-methylcytosine mapping that minimizes template degradation. *Anal Biochem.* 1995 Mar;226(1):161–6.
155. Huang Y, Pastor WA, Shen Y, Tahiliani M, Liu DR, Rao A. The Behaviour of 5-Hydroxymethylcytosine in Bisulfite Sequencing. Liu J, editor. *PLoS One.* 2010 Jan 26;5(1):e8888.
156. Shema E, Bernstein BE, Buenrostro JD. Single-cell and single-molecule epigenomics to uncover genome regulation at unprecedented resolution. *Nat Genet.* 2019;51(1):19–25.
157. Guo H, Zhu P, Guo F, Li X, Wu X, Fan X, et al. Profiling DNA methylome landscapes of mammalian cells with single-cell reduced-representation bisulfite sequencing. *Nat Protoc.* 2015 May 2;10(5):645–59.
158. Guo H, Zhu P, Wu X, Li X, Wen L, Tang F. Single-cell methylome landscapes of mouse embryonic stem cells and early embryos analyzed using reduced representation bisulfite sequencing. *Genome Res.* 2013 Dec 1;23(12):2126–35.
159. Wang K, Li X, Dong S, Liang J, Mao F, Zeng C, et al. Q-RRBS: a quantitative reduced representation bisulfite sequencing method for single-cell methylome analyses. *Epigenetics.* 2015 Sep 2;10(9):775–83.
160. Miura F, Enomoto Y, Dairiki R, Ito T. Amplification-free whole-genome bisulfite sequencing by post-bisulfite adaptor tagging. *Nucleic Acids Res.* 2012 Sep 1;40(17):e136–e136.
161. Miura F, Ito T. Highly sensitive targeted methylome sequencing by post-bisulfite adaptor tagging. *DNA Res.* 2015 Feb 1;22(1):13–8.
162. Smallwood SA, Lee HJ, Angermueller C, Krueger F, Saadeh H, Peat J, et al. Single-cell genome-wide bisulfite sequencing for assessing epigenetic heterogeneity. *Nat Methods.* 2014 Aug 20;11(8):817–20.
163. Clark SJ, Smallwood SA, Lee HJ, Krueger F, Reik W, Kelsey G. Genome-wide base-resolution mapping of DNA methylation in single cells using single-cell bisulfite sequencing (scBS-seq). *Nat Protoc.* 2017 Mar 9;12(3):534–47.
164. Farlik M, Sheffield NCC, Nuzzo A, Datlinger P, Schönegger A, Klughammer J, et al. Single-Cell DNA Methylome Sequencing and Bioinformatic Inference of Epigenomic Cell-State Dynamics. *Cell Rep.* 2015 Mar 3;10(8):1386–97.
165. Kobayashi H, Koike T, Sakashita A, Tanaka K, Kumamoto S, Kono T. Repetitive DNA methylome analysis by small-scale and single-cell shotgun bisulfite sequencing. *Genes to Cells.* 2016 Nov 1;21(11):1209–22.
166. Luo C, Keown CL, Kurihara L, Zhou J, He Y, Li J, et al. Single-cell methylomes identify neuronal subtypes and regulatory elements in mammalian cortex. *Science.* 2017 Aug 11;357(6351):600–4.
167. Mulqueen RM, Pokholok D, Norberg SJ, Torkency KA, Fields AJ, Sun D, et al. Highly scalable generation of DNA methylation profiles in single cells. *Nat Biotechnol.* 2018 May 9;36(5):428–31.
168. Ma S, de la Fuente Revenga M, Sun Z, Sun C, Murphy TW, Xie H, et al. Cell-type-specific brain methylomes profiled via ultralow-input microfluidics. *Nat Biomed Eng.* 2018 Mar 7;2(3):183–94.
169. Gravina S, Ganapathi S, Vijg J. Single-cell, locus-specific bisulfite sequencing (SLBS) for direct detection of epimutations in DNA methylation patterns. *Nucleic Acids Res.* 2015 Aug 18;43(14):e93–e93.
170. Hwang B, Lee JH, Bang D. Single-cell RNA sequencing technologies and bioinformatics pipelines. *Exp Mol Med.* 2018 Aug 7;50(8):96.
171. Angermueller C, Clark SJ, Lee HJ, Macaulay IC, Teng MJ, Hu TX, et al. Parallel single-cell sequencing links transcriptional and epigenetic heterogeneity. *Nat Methods.* 2016 Mar 11;13(3):229–32.
172. Pott S. Simultaneous measurement of chromatin accessibility, DNA methylation, and nucleosome phasing in single cells. *Elife.* 2017 Jun 27;6.
173. Clark SJ, Argelaguet R, Kapourani C-A, Stubbs TM, Lee HJ, Alda-Catalinas C, et al. scNMT-seq enables joint profiling of chromatin accessibility DNA methylation and transcription in single cells. *Nat Commun.* 2018 Dec 22;9(1):781.

174. Guo F, Li L, Li J, Wu X, Hu B, Zhu P, et al. Single-cell multi-omics sequencing of mouse early embryos and embryonic stem cells. *Cell Res.* 2017 Aug 16;27(8):967–88.
175. Cheow LF, Courtois ET, Tan Y, Viswanathan R, Xing Q, Tan RZ, et al. Single-cell multimodal profiling reveals cellular epigenetic heterogeneity. *Nat Methods.* 2016 Oct 15;13(10):833–6.
176. Hemelaar J. The origin and diversity of the HIV-1 pandemic. *Trends Mol Med.* 2012 Mar 1;18(3):182–92.
177. Worobey M, Gemmel M, Teuwen DE, Haselkorn T, Kunstman K, Bunce M, et al. Direct evidence of extensive diversity of HIV-1 in Kinshasa by 1960. *Nature.* 2008 Oct 2;455(7213):661–4.
178. Hahn BH, Shaw GM, De Cock KM, Sharp PM. AIDS as a Zoonosis: Scientific and Public Health Implications. *Science (80- ).* 2000 Jan 28;287(5453):607–14.
179. Sharp PM, Hahn BH. Origins of HIV and the AIDS pandemic. *Cold Spring Harb Perspect Med.* 2011 Sep 1;1(1):1–22.
180. Gao F, Bailes E, Robertson DL, Chen Y, Rodenburg CM, Michael SF, et al. Origin of HIV-1 in the chimpanzee *Pan troglodytes troglodytes*. *Nature.* 1999 Feb;397(6718):436–41.
181. Foley BT, Leitner T, Paraskevis D, Peeters M. Primate immunodeficiency virus classification and nomenclature: Review. *Infect Genet Evol.* 2016;46:150–8.
182. Clavel F. HIV-2, the West African AIDS virus. *AIDS.* 1987 Sep;1(3):135–40.
183. Keele BF, Van Heuverswyn F, Li Y, Bailes E, Takehisa J, Santiago ML, et al. Chimpanzee reservoirs of pandemic and nonpandemic HIV-1. *Science (80- ).* 2006 Jul 28;313(5786):523–6.
184. Drucker E, Alcabas PG, Marx PA. The injection century: massive unsterile injections and the emergence of human pathogens. *Lancet.* 2001 Dec 8;358(9297):1989–92.
185. Gisselquist D. Emergence of the HIV Type 1 Epidemic in the Twentieth Century: Comparing Hypotheses to Evidence. *AIDS Res Hum Retroviruses.* 2003 Dec 5;19(12):1071–8.
186. Gottlieb MS, Schroff R, Schanker HM, Weisman JD, Fan PT, Wolf RA, et al. Pneumocystis carinii Pneumonia and Mucosal Candidiasis in Previously Healthy Homosexual Men. *N Engl J Med.* 1981 Dec 10;305(24):1425–31.
187. Friedman-Kien AE, Laubenstein LJ, Rubenstein P, Buimovici-Klein E, Marmor M, Stahl R, et al. Disseminated Kaposi's Sarcoma in Homosexual Men. *Ann Intern Med.* 1982 Jun 1;96(6\_part\_1):693.
188. Blattner WA, Biggar RJ, Weiss SH, Melbye M, Goedert JJ. Epidemiology of human T-lymphotropic virus type III and the risk of the acquired immunodeficiency syndrome. *Ann Intern Med.* 1985;103(5):665–70.
189. Hymes K, Greene J, Marcus A, William D, Cheung T, Prose N, et al. Kaposi's sarcoma in homosexual men - a report of eight cases. *Lancet.* 1981;318(8247):598–600.
190. Gallo RC. A reflection on HIV/AIDS research after 25 years. *Retrovirology.* 2006 Dec 20;3(1):72.
191. Barré-Sinoussi F, Chermann JC, Rey F, Nugeyre MT, Chamaret S, Gruest J, et al. Isolation of a T-lymphotropic retrovirus from a patient at risk for acquired immune deficiency syndrome (AIDS). *Science (80- ).* 1983 May 20;220(4599):868–71.
192. Gallo RC, Sarin PS, E.P. G, Robert-Guroff M, Richardson E, Kalyanaraman VS, et al. Isolation of Human T-Cell Leukemia Virus in Acquired Immune Deficiency Syndrome (AIDS). *Science (80- ).* 1983 May 20;220(4599):865–7.
193. Popovic M, Sarngadharan MG, Read E, Gallo RC, Rey M, Santos-Ferreira M, et al. Detection, isolation, and continuous production of cytopathic retroviruses (HTLV-III) from patients with AIDS and pre-AIDS. *Science (80- ).* 1984 May 4;224(4648):497–500.
194. Levy JA, Hoffman AD, Kramer SM, Landis JA, Shimabukuro JM, Oshiro LS. Isolation of Lymphocytotropic Retroviruses from San Francisco Patients with AIDS. *Science (80- ).* 1984 Aug 24;225(4664):840–2.
195. Marx J. A Virus by Any Other Name . . . *Science (80- ).* 1985 Mar 22;227(4693):1449–51.
196. Ratner L, Gallo RC, Wong-Staal F. HTLV-III, LAV, ARV are variants of same AIDS virus. *Nature.* 1985;313(6004):636–7.
197. Piot P, Taelman H, Bila Minlangu K, Mbendi N, Ndangi K, Kalambayi K, et al. Acquired Immunodeficiency Syndrome in a Heterosexual Population in Zaire. *Lancet.* 1984;324(8394):65–9.
198. Van De Perre P, Lepage P, Kestelyn P, Hekker A, Rouvroy D, Bogaerts J, et al. Acquired Immunodeficiency Syndrome in Rwanda. *Lancet.* 1984;324(8394):62–5.

199. Coffin J, Haase A, Levy J a., Montagnier L, Oroszlan S, Teich N, et al. What to call the AIDS virus? *Nature*. 1986 Jan;321(6065):10.
200. UNAIDS. UNAIDS Data 2018. 2018.
201. UNAIDS. UNAIDS Data 2019. 2019.
202. Samji H, Cescon A, Hogg RS, Modur SP, Althoff KN, Buchacz K, et al. Closing the Gap: Increases in Life Expectancy among Treated HIV-Positive Individuals in the United States and Canada. Okulicz JF, editor. *PLoS One*. 2013 Dec 18;8(12):e81355.
203. Trickey A, May MT, Vehreschild J-J, Obel N, Gill MJ, Crane HM, et al. Survival of HIV-positive patients starting antiretroviral therapy between 1996 and 2013: a collaborative analysis of cohort studies. *Lancet HIV*. 2017 Aug 1;4(8):e349–56.
204. Barré-Sinoussi F, Ross AL, Delfraissy J-F. Past, present and future: 30 years of HIV research. *Nat Rev Microbiol*. 2013;11(12):877–83.
205. Archin NM, Sung JM, Garrido C, Soriano-Sarabia N, Margolis DM. Eradicating HIV-1 infection: seeking to clear a persistent pathogen. *Nat Rev Microbiol*. 2014 Nov 1;12(11):750–64.
206. J Buzón M, Massanella M, Llibre JM, Esteve A, Dahl V, Puertas MC, et al. HIV-1 replication and immune dynamics are affected by raltegravir intensification of HAART-suppressed subjects. *Nat Med*. 2010 Apr 14;16(4):460–5.
207. Ho YC, Shan L, Hosmane NN, Wang J, Laskey SB, Rosenbloom DIS, et al. Replication-competent noninduced proviruses in the latent reservoir increase barrier to HIV-1 cure. *Cell*. 2013;155(3):540–51.
208. Richman DD, Margolis DM, Delaney M, Greene WC, Hazuda D, Pomerantz RJ. The Challenge of Finding a Cure for HIV Infection. *Science (80- )*. 2009 Mar 6;323(5919):1304–7.
209. Castro-Gonzalez S, Colomer-Lluch M, Serra-Moreno R. Barriers for HIV Cure: The Latent Reservoir. *AIDS Res Hum Retroviruses*. 2018 Sep 14;34(9):739–59.
210. Krupovic M, Blomberg J, Coffin JM, Dasgupta I, Fan H, Geering AD, et al. Ortervirales: New Virus Order Unifying Five Families of Reverse-Transcribing Viruses. *J Virol*. 2018 Jun 15;92(12):e00515-18.
211. International Committee on Taxonomy of Viruses (ICTV): taxonomy [Internet]. [cited 2019 Jul 8]. Available from: <https://talk.ictvonline.org/taxonomy/>
212. Hayward A. Origin of the retroviruses: when, where, and how? *Curr Opin Virol*. 2017 Aug 1;25:23–7.
213. Vogt V. Retroviral Virions and Genomes. Coffin JM, Hughes SH, Varmus HE, editors. *Retroviruses*. Cold Spring Harbor Laboratory Press; 1997.
214. Aiewsakun P, Katzourakis A. Marine origin of retroviruses in the early Palaeozoic Era. *Nat Commun*. 2017 Apr 10;8(1):13954.
215. Calef C, Mokili J, Connor DHO, Watkins DI, Korber B. Numbering Positions in SIV Relative to SIVMM239 ( revised \*). HIV Seq database. 2005;171–81.
216. Shors T. Understanding viruses. Jones & Bartlett Learning; 2013. 704 p.
217. Santiago ML, Range F, Keele BF, Li Y, Bailes E, Bibollet-Ruche F, et al. Simian Immunodeficiency Virus Infection in Free-Ranging Sooty Mangabeys (*Cercocebus atys atys*) from the Tai Forest, Cote d'Ivoire: Implications for the Origin of Epidemic Human Immunodeficiency Virus Type 2. *J Virol*. 2005 Oct 1;79(19):12515–27.
218. Korber BT, Foley BT, Kuiken CL, Pillai SK, Sodroski JG. Numbering Positions in HIV Relative to HXB2CG. 1998.
219. Krebs FC, Hogan TH, Quiterio S, Gartner S, Wigdahl B. Lentiviral LTR-directed Expression, Sequence Variation, and Disease Pathogenesis. In: *HIV sequence compendium 2001 Los Alamos (New Mexico): Theoretical Biology and Biophysics*, Los Alamos National Laboratory. 2001. p. 29–70.
220. Feinberg MB, Greene WC. Molecular insights into human immunodeficiency virus type 1 pathogenesis. *Curr Opin Immunol*. 1992 Jan 1;4(4):466–74.
221. Schwartz S, Felber BK, Benko DM, Fenyó EM, Pavlakis GN. Cloning and functional analysis of multiply spliced mRNA species of human immunodeficiency virus type 1. *J Virol*. 1990 Jun 1;64(6):2519–29.
222. Coffin JM. Structure and Classification of Retroviruses. In: *The Retroviridae The Viruses* Springer, Boston, MA. 1992. p. 19–49.

223. Göttinger HG. HIV-1 Gag: a Molecular Machine Driving Viral Particle Assembly and Release. In: HIV sequence compendium 2001 Los Alamos (New Mexico): Theoretical Biology and Biophysics, Los Alamos National Laboratory. 2001. p. 2–28.
224. Spreadsheets AG. Landmarks of the HIV genome [Internet]. Human Retroviruses and AIDS. 1998 [cited 2015 Mar 12]. p. 4–8. Available from: <http://www.hiv.lanl.gov/content/sequence/HIV/MAP/landmark.html>
225. Wyatt R, Kwong PD, Hendrickson WA, Sodroski JG. Structure of gp120 Structure of the Core of the HIV-1 gp120 Exterior Envelope Glycoprotein. 1998.
226. Karn J, Stoltzfus CM. Transcriptional and posttranscriptional regulation of HIV-1 gene expression. *Cold Spring Harb Perspect Med*. 2012 Feb 1;2(2):a006916.
227. Seelangari A, Maddukuri A, Berro R, de la Fuente C, Kehn K, Deng L, et al. Role of viral regulatory and accessory proteins in HIV-1 replication. *Front Biosci*. 2004 Sep 1;9:2388–413.
228. Cassan E, Arigon-Chifolleau A-M, Mesnard J-M, Gross A, Gascuel O. Concomitant emergence of the antisense protein gene of HIV-1 and of the pandemic. *Proc Natl Acad Sci U S A*. 2016 Oct 11;113(41):11537–42.
229. Landmarks of the HIV genome, Los Alamos National Library [Internet]. [cited 2019 Jul 23]. Available from: <https://www.hiv.lanl.gov/content/sequence/HIV/MAP/landmark.html>
230. Briggs JAGG, Kräusslich H-GG. The Molecular Architecture of HIV. *J Mol Biol*. 2011 Jul 22;410(4):491–500.
231. Turner BG, Summers MF. Structural biology of HIV. *J Mol Biol*. 1999 Jan 8;285(1):1–32.
232. Ganser-Pornillos BK, Yeager M, Sundquist WI. The structural biology of HIV assembly. *Curr Opin Struct Biol*. 2008 Apr;18(2):203–17.
233. Cantin R, Méthot S, Tremblay MJ. Plunder and Stowaways: Incorporation of Cellular Proteins by Enveloped Viruses. *J Virol*. 2005 Jun 1;79(11):6577–87.
234. Eckwahl MJ, Arnion H, Kharytonchik S, Zang T, Bieniasz PD, Telesnitsky A, et al. Analysis of the human immunodeficiency virus-1 RNA packageome. *RNA*. 2016 Aug 1;22(8):1228–38.
235. Pantaleo G, Fauci a S. Immunopathogenesis of hiv infection 1. *Annu Rev Microbiol*. 1996;50(50):825–54.
236. Levy JA. HIV and the Pathogenesis of AIDS, Third Edition. American Society of Microbiology; 2007.
237. Maartens G, Celum C, Lewin SR. HIV infection: epidemiology, pathogenesis, treatment, and prevention. *Lancet*. 2014 Jul 19;384(9939):258–71.
238. Coffin J, Swanstrom R. HIV pathogenesis: dynamics and genetics of viral populations and infected cells. *Cold Spring Harb Perspect Med*. 2013 Jan 1;3(1):a012526.
239. Janeway CA, Travers P, Walport M, Shlomchik M. Immunobiology: The Immune System In Health And Disease. 5th ed. *Immuno Biology* 5. Garland Publishing, New York; 2001. 892 p.
240. Sierra S, Kupfer B, Kaiser R. Basics of the virology of HIV-1 and its replication. *J Clin Virol*. 2005;34(4):233–44.
241. Schacker T, Collier AC, Hughes J, Shea T, Corey L. Clinical and epidemiologic features of primary HIV infection. *Ann Intern Med*. 1996;125(4):257–64.
242. Mcmichael a J, Borrow P, Tomaras GD, Goonetilleke N, Haynes BF. The immune response during acute HIV-1 infection: clues for vaccine development. *Nat Rev Immunol*. 2010;10(1):11–23.
243. Pantaleo G, Graziosi C, Demarest JF, Butini L, Montroni M, Fox CH, et al. HIV infection is active and progressive in lymphoid tissue during the clinically latent stage of disease. *Nature*. 1993 Mar 25;362(6418):355–8.
244. Ho DD, Neumann a U, Perelson a S, Chen W, Leonard JM, Markowitz M. Rapid turnover of plasma virions and CD4 lymphocytes in HIV-1 infection. *Nature*. 1995;373(6510):123–6.
245. Bleul CC, Wu L, Hoxie JA, Springer TA, Mackay CR. The HIV coreceptors CXCR4 and CCR5 are differentially expressed and regulated on human T lymphocytes. *Immunology*. 1997;94(March):1925–30.
246. Bukrinsky MI, Sharova N, Dempsey MP, Stanwick TL, Bukrinskaya AG, Haggerty S, et al. Active nuclear import of human immunodeficiency virus type 1 preintegration complexes. *Proc Natl Acad Sci U S A*. 1992;89(14):6580–4.
247. Meyerhans A, Vartanian JP, Hultgren C, Plikat U, Karlsson A, Wang L, et al. Restriction and enhancement of human immunodeficiency virus type 1 replication by modulation of intracellular deoxynucleoside triphosphate pools. *J Virol*. 1994;68(1):535–40.



248. Lackner AA, Lederman MM, Rodriguez B. HIV pathogenesis: the host. *Cold Spring Harb Perspect Med.* 2012 Sep 1;2(9):a007005.
249. Okulicz JF, Marconi VC, Landrum ML, Wegner S, Weintrob A, Ganesan A, et al. Clinical outcomes of elite controllers, viremic controllers, and long-term nonprogressors in the US Department of Defense HIV natural history study. *J Infect Dis.* 2009;200(11):1714–23.
250. Deeks SG, Walker BD. Human Immunodeficiency Virus Controllers: Mechanisms of Durable Virus Control in the Absence of Antiretroviral Therapy. *Immunity.* 2007;27(3):406–16.
251. Lambotte O, Boufassa F, Madec Y, Nguyen A, Meyer L, Rouzioux C, et al. HIV / AIDS HIV Controllers : A Homogeneous Group of HIV-1 – Infected Patients with Spontaneous Control of Viral Replication. *Curr Opin Microbiol.* 2005;41:13–6.
252. Klein F, Mouquet H, Dosenovic P, Scheid JF, Scharf L, Nussenzweig MC. Antibodies in HIV-1 vaccine development and therapy. *Science (80- ).* 2013 Sep 13;341(6151):1199–204.
253. Vanhamel J, Bruggemans A, Debysers Z. Establishment of latent HIV-1 reservoirs: what do we really know? *J virus Erad.* 2019 Jan 1;5(1):3–9.
254. Engelman A, Cherepanov P. The structural biology of HIV-1: mechanistic and therapeutic insights. *Nat Rev Microbiol.* 2012;10(4):279–90.
255. Wilen CB, Tilton JC, Doms RW. HIV: cell binding and entry. *Cold Spring Harb Perspect Med.* 2012 Aug 1;2(8):a006866.
256. Vandekerckhove L, Christ F, Van Maele B, De Rijck J, Gijsbers R, Van den Haute C, et al. Transient and Stable Knockdown of the Integrase Cofactor LEDGF/p75 Reveals Its Role in the Replication Cycle of Human Immunodeficiency Virus. *J Virol.* 2006 Nov 15;80(4):1886–96.
257. Piller S, Caly L, Jans D. Nuclear Import of the Pre-Integration Complex (PIC): The Achilles Heel of HIV ? *Curr Drug Targets.* 2003 Jul 1;4(5):409–29.
258. Craigie R, Bushman FD. HIV DNA integration. *Cold Spring Harb Perspect Med.* 2012 Jul 1;2(7):a006890.
259. Arhel N. Revisiting HIV-1 uncoating. *Retrovirology.* 2010 Dec 17;7(1):96.
260. Einkauf KB, Lee GQ, Gao C, Sharaf R, Sun X, Hua S, et al. Intact HIV-1 proviruses accumulate at distinct chromosomal positions during prolonged antiretroviral therapy. *J Clin Invest.* 2019;129(3):988–98.
261. Vansant G, Vranckx LS, Zurnic I, Van Looveren D, Van De Velde P, Nobles C, et al. Impact of LEDGIN treatment during virus production on residual HIV-1 transcription. *Retrovirology.* 2019 Apr 2;16(1).
262. Wain-Hobson S, Sonigo P, Danos O, Cole S, Alizon M. Nucleotide sequence of the AIDS virus, LAV. *Cell.* 1985 Jan 1;40(1):9–17.
263. Korber B, Kuiken C, Foley B, Hahn B, McCutchan F, Mellors J, et al. Human Retroviruses and AIDS 1998: A Compilation and Analysis of Nucleic Acid and Amino Acid Sequences. *Theor Biol Biophys Group, Los Alamos Natl Lab Los Alamos, NM.* 1998;
264. Tanaka J, Ishida T, Choi B, Yasuda J, Watanabe et al. T. Latent HIV-1 reactivation in transgenic mice requires cell cycle -dependent demethylation of CREB/ATF sites in the LTR. *Aids.* 2003;17(October 2002):167–75.
265. Pereira LA, Bentley K, Peeters A, Churchill MJ, Deacon NJ. A compilation of cellular transcription factor interactions with the HIV-1 LTR promoter. *Nucleic Acids Res.* 2000 Feb 1;28(3):663–8.
266. Levy JA. Pathogenesis of human immunodeficiency virus infection. *Microbiol Mol Biol Rev.* 1993 Mar 1;57(1):183–289.
267. Lassen K, Han Y, Zhou Y, Siliciano J, Siliciano RF. The multifactorial nature of HIV-1 latency. *Trends Mol Med.* 2004;10(11):525–31.
268. Kumar A, Darcis G, Van Lint C, Herbein G. Epigenetic control of HIV-1 post integration latency: implications for therapy. *Clin Epigenetics.* 2015;7(1):103.
269. Wei P, Garber ME, Fang S-M, Fischer WH, Jones KA. A Novel CDK9-Associated C-Type Cyclin Interacts Directly with HIV-1 Tat and Mediates Its High-Affinity, Loop-Specific Binding to TAR RNA. *Cell.* 1998 Feb 20;92(4):451–62.
270. Sundquist WI, Kräusslich H-G. HIV-1 assembly, budding, and maturation. *Cold Spring Harb Perspect Med.* 2012 Jul 1;2(7):a006924.

271. Eisele E, Siliciano RF. Redefining the Viral Reservoirs that Prevent HIV-1 Eradication. *Immunity*. 2012 Sep 21;37(3):377–88.
272. McCune JM. Viral latency in HIV disease. *Cell*. 1995;82(2):183–8.
273. Finzi D, Hermankova M, Pierson T, Carruth L, Buck C, Chaisson R, et al. Identification of a Reservoir for HIV-1 in Patients on Highly Active Antiretroviral Therapy. *Science* (80- ). 1997;278(5341):1295–300.
274. Siliciano JD, Kajdas J, Finzi D, Quinn TC, Chadwick K, Margolick JB, et al. Long-term follow-up studies confirm the stability of the latent reservoir for HIV-1 in resting CD4+ T cells. *Nat Med*. 2003;9(6):727–8.
275. Yuki SA, Kaiser P, Kim P, Telwate S, Joshi SK, Vu M, et al. HIV latency in isolated patient CD4+ T cells may be due to blocks in HIV transcriptional elongation, completion, and splicing. *Sci Transl Med*. 2018 Feb 28;10(430):1–16.
276. Zhou Y, Zhang H, Siliciano JD, Siliciano RF. Kinetics of Human Immunodeficiency Virus Type 1 Decay following Entry into Resting CD4+ T cells. *J Virol*. 2005 Feb;79(4):2199–210.
277. Bukrinsky MI, Haggerty S, Dempsey MP, Sharova N, Adzhubel A, Spitz L, et al. A nuclear localization signal within HIV-1 matrix protein that governs infection of non-dividing cells. *Nature*. 1993 Oct 14;365(6447):666–9.
278. von Schwedler U, Kornbluth RS, Trono D. The nuclear localization signal of the matrix protein of human immunodeficiency virus type 1 allows the establishment of infection in macrophages and quiescent T lymphocytes. *Proc Natl Acad Sci U S A*. 1994;91(15):6992–6.
279. Heinzinger NK, Bukrinsky MI, Haggerty SA, Ragland AM, Kewalramani V, Lee MA, et al. The Vpr protein of human immunodeficiency virus type 1 influences nuclear localization of viral nucleic acids in nondividing host cells. *Proc Natl Acad Sci U S A*. 1994;91(15):7311–5.
280. Stevenson M, Haggerty S, Lamonica CA, Meier CM, Welch et al. SK. Integration is not necessary for expression of human immunodeficiency virus type 1 protein products. *J Virol*. 1990;64(5):2421–5.
281. Pierson T, McArthur J, Siliciano RF. Reservoirs for HIV-1: Mechanisms for Viral Persistence in the Presence of Antiviral Immune Responses and Antiretroviral Therapy. *Annu Rev Immunol*. 2000;18(1):665–708.
282. Wu H. HIV-1 gene expression: lessons from provirus and non-integrated DNA. *Retrovirology*. 2004;1:13.
283. Coiras M, López-Huertas MR, Pérez-Olmeda M, Alcamí J. Understanding HIV-1 latency provides clues for the eradication of long-term reservoirs. *Nat Rev Microbiol*. 2009 Nov 1;7(11):798–812.
284. Barton K, Winckelmann A, Palmer S. HIV-1 Reservoirs During Suppressive Therapy. *Trends Microbiol*. 2016 May 1;24(5):345–55.
285. Chun T-W, Finzi D, Margolick J, Chadwick K, Schwartz D, Siliciano RF. In vivo fate of HIV-1-infected T cells: Quantitative analysis of the transition to stable latency. *Nat Med*. 1995 Dec;1(12):1284–90.
286. Chun T-W, Carruth L, Finzi D, Shen X, DiGiuseppe JA, Taylor H, et al. Quantification of latent tissue reservoirs and total body viral load in HIV-1 infection. *Nature*. 1997 May;387(6629):183–8.
287. Chavez L, Calvanese V, Verdin E. HIV Latency Is Established Directly and Early in Both Resting and Activated Primary CD4 T Cells. Emerman M, editor. *PLOS Pathog*. 2015 Jun 11;11(6):e1004955.
288. Seu L, Sabbaj S, Duverger A, Wagner F, Anderson JC, Davies E, et al. Stable Phenotypic Changes of the Host T Cells Are Essential to the Long-Term Stability of Latent HIV-1 Infection. *J Virol*. 2015 Jul;89(13):6656–72.
289. Ananworanich J, Dubé K, Chomont N. How does the timing of antiretroviral therapy initiation in acute infection affect Hiv reservoirs? *Curr Opin Hiv Aids*. 2015 Jan 1;10(1):18–28.
290. Cuevas JM, Geller R, Garijo R, López-Aldeguer J, Sanjuán R. Extremely High Mutation Rate of HIV-1 In Vivo. Rowland-Jones SL, editor. *PLOS Biol*. 2015 Sep 16;13(9):e1002251.
291. Bruner KM, Murray AJ, Pollack RA, Soliman MG, Laskey SB, Capoferri AA, et al. Defective proviruses rapidly accumulate during acute HIV-1 infection. *Nat Med*. 2016 Sep 8;22(9):1043–9.
292. Sanjuán R, Nebot MR, Chirico N, Mansky LM, Belshaw R. Viral mutation rates. *J Virol*. 2010 Oct 1;84(19):9733–48.
293. Abram ME, Ferris AL, Shao W, Alvord WG, Hughes SH. Nature, Position, and Frequency of Mutations Made in a Single Cycle of HIV-1 Replication. *J Virol*. 2010;84(19):9864–78.
294. Mansky LM. Retrovirus mutation rates and their role in genetic variation. *J Gen Virol*. 1998 Jun 1;79(6):1337–45.
295. Roberts JD, Bebenek K, Kunkel TA. The Accuracy of Reverse Transcriptase from HIV-1. *Science* (80- ). 1988 Jun 20;242(4882):1171–3.

296. Ji J, Loeb LA. Fidelity of HIV-1 reverse transcriptase copying RNA in vitro. *Biochemistry*. 1992 Feb 4;31(4):954–8.
297. Desimie BA, Delviks-Frankenberry KA, Burdick RC, Qi D, Izumi T, Pathak VK. Multiple APOBEC3 Restriction Factors for HIV-1 and One Vif to Rule Them All. *J Mol Biol*. 2014 Mar 20;426(6):1220–45.
298. Santa-Marta M, de Brito PM, Godinho-Santos A, Goncalves J. Host Factors and HIV-1 Replication: Clinical Evidence and Potential Therapeutic Approaches. *Front Immunol*. 2013 Oct 24;4:343.
299. Moris A, Murray S, Cardinaud S. AID and APOBECs span the gap between innate and adaptive immunity. *Front Microbiol*. 2014 Oct 13;5:534.
300. Perelson AS, Neumann AU, Markowitz M, Leonard JM, Ho DD. HIV-1 dynamics in vivo: virion clearance rate, infected cell life-span, and viral generation time. *Science (80- )*. 1996 Mar 15;271(5255):1582–6.
301. Leslie AJ, Pfafferoth KJ, Chetty P, Draenert R, Addo MM, Feeney M, et al. HIV evolution: CTL escape mutation and reversion after transmission. *Nat Med*. 2004;10(3):282–9.
302. Chang JT, Wherry EJ, Goldrath AW. Molecular regulation of effector and memory T cell differentiation. *Nat Immunol*. 2014 Dec 14;15(12):1104–15.
303. Stevenson M. Role of myeloid cells in HIV-1-host interplay. *J Neurovirol*. 2015 Jun 19;21(3):242–8.
304. Swingler S, Mann AM, Zhou J, Swingler C, Stevenson M. Apoptotic Killing of HIV-1-Infected Macrophages Is Subverted by the Viral Envelope Glycoprotein. *PLoS Pathog*. 2007;3(9):e134.
305. Mzingwane ML, Tiemessen CT. Mechanisms of HIV persistence in HIV reservoirs. *Rev Med Virol*. 2017 Mar 1;27(2):e1924.
306. Wong J, YuKL SA. Tissue reservoirs of HIV. *Curr Opin HIV Aids*. 2016 Jul 1;11(4):362–70.
307. World Health Organization. Consolidated guidelines on HIV prevention, diagnosis, treatment and care for key populations : 2016 update. 155 p.
308. World Health Organization, World Health Organization. Department of HIV/AIDS. Guideline on when to start antiretroviral therapy and on pre-exposure prophylaxis for HIV. 76 p.
309. De Scheerder M-A, Vrancken B, Dellicour S, Schlub T, Lee E, Shao W, et al. HIV Rebound Is Predominantly Fueled by Genetically Identical Viral Expansions from Diverse Reservoirs. *Cell Host Microbe*. 2019 Aug 27;
310. Wagner TA, McLaughlin S, Garg K, Cheung CYK, Larsen BB, Styrchak S, et al. Proliferation of cells with HIV integrated into cancer genes contributes to persistent infection. *Science (80- )*. 2014 Aug 1;345(6196):570–3.
311. Simonetti FR, Sobolewski MD, Fyne E, Shao W, Spindler J, Hattori J, et al. Clonally expanded CD4+ T cells can produce infectious HIV-1 in vivo. *Proc Natl Acad Sci U S A*. 2016 Feb 16;113(7):1883–8.
312. Maldarelli F, Wu X, Su L, Simonetti FR, Shao W, Hill S, et al. Specific HIV integration sites are linked to clonal expansion and persistence of infected cells. *Science (80- )*. 2014 Jul 11;345(6193):179–83.
313. Xu W, Li H, Wang Q, Hua C, Zhang H, Li W, et al. Advancements in Developing Strategies for Sterilizing and Functional HIV Cures. *Biomed Res Int*. 2017 Apr 26;2017:1–12.
314. Henrich TJ, Lelièvre J-D. Progress towards obtaining an HIV cure: slow but sure. *Curr Opin HIV AIDS*. 2018 Sep 1;13(5):381–2.
315. Hill AL, Rosenbloom DIS, Fu F, Nowak MA, Siliciano RF. Predicting the outcomes of treatment to eradicate the latent reservoir for HIV-1. *Proc Natl Acad Sci*. 2014 Sep 16;111(37):13475–80.
316. Hill AL, Rosenbloom DIS, Goldstein E, Hanhauser E, Kuritzkes DR, Siliciano RF, et al. Real-Time Predictions of Reservoir Size and Rebound Time during Antiretroviral Therapy Interruption Trials for HIV. Weinberger L, editor. *PLoS Pathog*. 2016 Apr 27;12(4):e1005535.
317. Darcis G, Van Driessche B, Van Lint C. HIV Latency: Should We Shock or Lock? *Trends Immunol*. 2017 Mar 1;38(3):217–28.
318. Gupta RK, Abdul-Jawad S, McCoy LE, Mok HP, Peppas D, Salgado M, et al. HIV-1 remission following CCR5Δ32/Δ32 haematopoietic stem-cell transplantation. *Nature*. 2019 Apr 5;568(7751):244–8.
319. Hütter G, Nowak D, Mossner M, Ganepola S, Müßig A, Allers K, et al. Long-Term Control of HIV by CCR5 Delta32/ Delta32 Stem-Cell Transplantation. *N Engl J Med*. 2009 Feb 12;360(7):692–8.
320. Abner E, Jordan A. HIV “shock and kill” therapy: In need of revision. *Antiviral Res*. 2019 Jun 1;166:19–34.

321. Kessing CF, Nixon CC, Li C, Tsai P, Takata H, Mousseau G, et al. In Vivo Suppression of HIV Rebound by Didehydro-Cortistatin A, a "Block-and-Lock" Strategy for HIV-1 Treatment. *Cell Rep.* 2017;21(3):600–11.
322. Hamer D. Can HIV be Cured? Mechanisms of HIV Persistence and Strategies to Combat It. *Curr HIV Res.* 2004 Apr 1;2(2):99–111.
323. Shan L, Deng K, Shroff NS, Durand CM, Rabi SA, Yang H-C, et al. Stimulation of HIV-1-Specific Cytolytic T Lymphocytes Facilitates Elimination of Latent Viral Reservoir after Virus Reactivation. *Immunity.* 2012 Mar 23;36(3):491–501.
324. Grau-Expósito J, Luque-Ballesteros L, Navarro J, Curran A, Burgos J, Ribera E, et al. Latency reversal agents affect differently the latent reservoir present in distinct CD4+ T subpopulations. Swanson R, editor. *PLOS Pathog.* 2019 Aug 19;15(8):e1007991.
325. Kim Y, Anderson JL, Lewin SR. Getting the Kill into Shock and Kill: Strategies to Eliminate Latent HIV. *Cell Host Microbe.* 2018 Jan 10;23(1):14–26.
326. Wagner TA. Quarter Century of Anti-HIV CAR T Cells. *Curr HIV/AIDS Rep.* 2018;15(2):147–54.
327. Mothe B, Manzardo C, Sanchez-Bernabeu A, Coll P, Morón-López S, Puertas MC, et al. Therapeutic Vaccination Refocuses T-cell Responses Towards Conserved Regions of HIV-1 in Early Treated Individuals (BCN 01 study). *EClinicalMedicine.* 2019 May 1;11:65–80.
328. Archin NM, Liberty AL, Kashuba AD, Choudhary SK, Kuruc JD, Crooks AM, et al. Administration of vorinostat disrupts HIV-1 latency in patients on antiretroviral therapy. *Nature.* 2012;487(7408):482–5.
329. Elliott JH, Wightman F, Solomon A, Ghneim K, Ahlers J, Cameron MJ, et al. Activation of HIV Transcription with Short-Course Vorinostat in HIV-Infected Patients on Suppressive Antiretroviral Therapy. Cullen BR, editor. *PLoS Pathog.* 2014 Nov 13;10(11):e1004473.
330. Rasmussen TA, Tolstrup M, Brinkmann CR, Olesen R, Erikstrup C, Solomon A, et al. Panobinostat, a histone deacetylase inhibitor, for latent virus reactivation in HIV-infected patients on suppressive antiretroviral therapy: A phase 1/2, single group, clinical trial. *Lancet HIV.* 2014 Oct;1(1):e13–21.
331. Elliott JH, McMahon JH, Chang CC, Lee SA, Hartogensis W, Bumpus N, et al. Short-term administration of disulfiram for reversal of latent HIV infection: a phase 2 dose-escalation study. *Lancet HIV.* 2015 Dec 1;2(12):e520–9.
332. Archin NM, Kirchherr JL, Sung JAM, Clutton G, Sholtis K, Xu Y, et al. Interval dosing with the HDAC inhibitor vorinostat effectively reverses HIV latency. *J Clin Invest.* 2017 Aug 1;127(8):3126–35.
333. Gutiérrez C, Serrano-Villar S, Madrid-Elena N, Pérez-Eliás MJ, Martín ME, Barbas C, et al. Bryostatins-1 for latent virus reactivation in HIV-infected patients on antiretroviral therapy. *AIDS.* 2016;30(9):1385–92.
334. Søgaard OS, Graversen ME, Leth S, Olesen R, Brinkmann CR, Nissen SK, et al. The Depsipeptide Romidepsin Reverses HIV-1 Latency In Vivo. Siliciano RF, editor. *PLOS Pathog.* 2015 Sep 17;11(9):e1005142.
335. Demeulemeester J, Vets S, Schrijvers R, Madlala P, De Maeyer M, De Rijck J, et al. HIV-1 integrase variants retarget viral integration and are associated with disease progression in a chronic infection cohort. *Cell Host Microbe.* 2014;16(5):651–62.
336. Christ F, Shaw S, Demeulemeester J, Desimie BA, Marchan A, Butler S, et al. Small-molecule inhibitors of the LEDGF/p75 binding site of integrase block HIV replication and modulate integrase multimerization. *Antimicrob Agents Chemother.* 2012 Aug;56(8):4365–74.
337. Kauder SE, Bosque A, Lindqvist A, Planelles V, Verdin E. Epigenetic regulation of HIV-1 latency by cytosine methylation. *PLoS Pathog.* 2009 Jun;5(6):e1000495.
338. Bouchat S, Delacourt N, Kula A, Darcis G, Van Driessche B, Corazza F, et al. Sequential treatment with 5-aza-2'-deoxycytidine and deacetylase inhibitors reactivates HIV-1. *EMBO Mol Med.* 2016;8(2):117–38.
339. Friedman J, Cho W-K, Chu CK, Keedy KS, Archin NM, Margolis DM, et al. Epigenetic Silencing of HIV-1 by the Histone H3 Lysine 27 Methyltransferase Enhancer of Zeste 2. *J Virol.* 2011;85(17):9078–89.
340. Wightman F, Ellenberg P, Churchill M, Lewin SR. HDAC inhibitors in HIV. *Immunol Cell Biol.* 2012;90(1):47–54.
341. Khan S, Iqbal M, Tariq M, Baig SM, Abbas W. Epigenetic regulation of HIV-1 latency: focus on polycomb group (PcG) proteins. *Clin Epigenetics.* 2018;10(1):14.
342. Narasipura SD, Kim S, Al-Harhi L. Epigenetic regulation of HIV-1 latency in astrocytes. *J Virol.* 2014;88(5):3031–8.

343. Hakre S, Chavez L, Shirakawa K, Verdin E. Epigenetic regulation of HIV latency. *Curr Opin HIV AIDS*. 2011;6(1):19–24.
344. Verdin E, Paras P, Van Lint C. Chromatin disruption in the promoter of human immunodeficiency virus type 1 during transcriptional activation. *EMBO J*. 1993;12(8):3249–59.
345. Röling MD, Stoszko M, Mahmoudi T. Molecular Mechanisms Controlling HIV Transcription and Latency – Implications for Therapeutic Viral Reactivation. *Adv Mol Retrovirology*. 2016;45–105.
346. Verdin E. DNase I-hypersensitive sites are associated with both long terminal repeats and with the intragenic enhancer of integrated human immunodeficiency virus type 1. *J Virol*. 1991 Dec 1;65(12):6790–9.
347. Van Lint C, Emiliani S, Ott M, Verdin E. Transcriptional activation and chromatin remodeling of the HIV-1 promoter in response to histone acetylation. *EMBO J*. 1996;15(5):1112–20.
348. Rafati H, Parra M, Hakre S, Moshkin Y, Verdin E, Mahmoudi T. Repressive LTR Nucleosome Positioning by the BAF Complex Is Required for HIV Latency. Emerman M, editor. *PLoS Biol*. 2011 Nov 29;9(11):e1001206.
349. El Kharroubi A, Verdin E. Protein-DNA interactions within DNase I-hypersensitive sites located downstream of the HIV-1 promoter. *J Biol Chem*. 1994;269(31):19916–24.
350. Blazkova J, Murray D, Justement JS, Funk EK, Nelson AK, Moir S, et al. Paucity of HIV DNA methylation in latently infected, resting CD4+ T cells from infected individuals receiving antiretroviral therapy. *J Virol*. 2012 May;86(9):5390–2.
351. Bednarik DP, Cook J a, Pitha PM. Inactivation of the HIV LTR by DNA CpG methylation: evidence for a role in latency. *EMBO J*. 1990;9(4):1157–64.
352. Ishida T, Hamano A, Koiki T, Watanabe T. 5' long terminal repeat (LTR)-selective methylation of latently infected HIV-1 provirus that is demethylated by reactivation signals. *Retrovirology*. 2006;3:69.
353. Weber S, Weiser B, Kemal KS, Burger H, Ramirez CM, Korn K, et al. Epigenetic analysis of HIV-1 proviral genomes from infected individuals: Predominance of unmethylated CpG's. *Virology*. 2014 Jan 20;449:181–9.
354. Palacios JA, Pérez-Piñar T, Toro C, Sanz-Minguela B, Moreno V, Valencia E, et al. Long-term nonprogressor and elite controller patients who control viremia have a higher percentage of methylation in their HIV-1 proviral promoters than aviremic patients receiving highly active antiretroviral therapy. *J Virol*. 2012 Dec;86(23):13081–4.
355. Cortés-Rubio CN, Salgado-Montes de Oca G, Prado-Galbarro FJ, Matías-Florentino M, Murakami-Ogasawara A, Kuri-Cervantes L, et al. Longitudinal variation in human immunodeficiency virus long terminal repeat methylation in individuals on suppressive antiretroviral therapy. *Clin Epigenetics*. 2019 Dec 13;11(1):134.
356. Laird PW. Principles and challenges of genome-wide DNA methylation analysis. *Nat Rev Genet*. 2010 Mar 3;11(3):191.
357. Avettand-Fénoël V, Hocqueloux L, Ghosn J, Cheret A, Frange P, Melard A, et al. Total HIV-1 DNA, a marker of viral reservoir dynamics with clinical implications. *Clin Microbiol Rev*. 2016 Oct;29(4):859–80.
358. Fang J, Mikovits JA, Bagni R, Cari L, Ruscetti et al. FW. Infection of Lymphoid Cells by Immunodeficiency Virus Type 1 Increases De Novo Methylation Infection of Lymphoid Cells by Integration-Defective Human Immunodeficiency Virus Type 1 Increases De Novo Methylation. *J Virol*. 2001;75(20):9753–61.
359. Jeeninga RE, Westerhout EM, van Gerven ML, Berkhout B. HIV-1 latency in actively dividing human T cell lines. *Retrovirology*. 2008;5:37.
360. Mok H-P, Lever AM. Chromatin, gene silencing and HIV latency. *Genome Biol*. 2007;8(11):228.
361. Pion MM, Jordan A, Biancotto A, Dequiedt F, Gondois-Rey FF, Rondeau S, et al. Transcriptional Suppression of In Vitro-Integrated Human Immunodeficiency Virus Type 1 Does Not Correlate with Proviral DNA Methylation. *J Virol*. 2003 Apr 1;77(7):4025–32.
362. Cruickshanks HA, McBryan T, Nelson DM, VanderKraats ND, Shah PP, van Tuyn J, et al. Senescent cells harbour features of the cancer epigenome. *Nat Cell Biol*. 2013 Dec 24;15(12):1495–506.
363. Nestor CE, Ottaviano R, Reinhardt D, Cruickshanks HA, Mjoseng HK, McPherson RC, et al. Rapid reprogramming of epigenetic and transcriptional profiles in mammalian culture systems. *Genome Biol*. 2015 Feb 4;16(1):11.
364. McEwen KR, Leitch HG, Amouroux R, Hajkova P. The impact of culture on epigenetic properties of pluripotent stem cells and pre-implantation embryos. *Biochem Soc Trans*. 2013 Jun 1;41(3):711–9.

365. Schatz P, Dietrich D, Koenig T, Burger M, Lukas A, Fuhrmann I, et al. Development of a diagnostic microarray assay to assess the risk of recurrence of prostate cancer based on PITX2 DNA methylation. *J Mol Diagn.* 2010 May;12(3):345–53.
366. Weiss G, Cottrell S, Distler J, Schatz P, Kristiansen G, Ittmann M, et al. DNA methylation of the PITX2 gene promoter region is a strong independent prognostic marker of biochemical recurrence in patients with prostate cancer after radical prostatectomy. *J Urol.* 2009;181(4):1678–85.
367. Dietrich D, Hasinger O, Bañez LL, Sun L, van Leenders GJ, Wheeler TM, et al. Development and clinical validation of a real-time PCR assay for PITX2 DNA methylation to predict prostate-specific antigen recurrence in prostate cancer patients following radical prostatectomy. *J Mol Diagn.* 2013;15(2):270–9.
368. Bañez LL, Sun L, van Leenders GJ, Wheeler TM, Bangma CH, Freedland SJ, et al. Multicenter clinical validation of PITX2 methylation as a prostate specific antigen recurrence predictor in patients with post-radical prostatectomy prostate cancer. *J Urol.* 2010;184(1):149–56.
369. Holmes EE, Jung M, Meller S, Leisse A, Sailer V, Zech J, et al. Performance evaluation of kits for bisulfite-conversion of DNA from tissues, cell lines, FFPE tissues, aspirates, lavages, effusions, plasma, serum, and urine. *PLoS One.* 2014 Jan;9(4):e93933.
370. Weller M, Stupp R, Reifenberger G, Brandes AA, van den Bent MJ, Wick W, et al. MGMT promoter methylation in malignant gliomas: ready for personalized medicine? *Nat Rev Neurol.* 2010 Jan 8;6(1):39–51.
371. Stewart GD, Van Neste L, Delvenne P, Delrée P, Delga A, McNeill SA, et al. Clinical utility of an epigenetic assay to detect occult prostate cancer in histopathologically negative biopsies: Results of the MATLOC study. *J Urol.* 2013;189(3):1110–6.
372. Kneip C, Schmidt B, Seegebarth A, Weickmann S, Fleischhacker M, Liebenberg V, et al. SHOX2 DNA methylation is a biomarker for the diagnosis of lung cancer in plasma. *J Thorac Oncol.* 2011 Oct;6(10):1632–8.
373. Church TR, Wandell M, Lofton-Day C, Mongin SJ, Burger M, Payne SR, et al. Prospective evaluation of methylated SEPT9 in plasma for detection of asymptomatic colorectal cancer. *Gut.* 2014 Feb;63(2):317–25.
374. Grützmann R, Molnar B, Pilarsky C, Habermann JK, Schlag PM, Saeger HD, et al. Sensitive detection of colorectal cancer in peripheral blood by septin 9 DNA methylation assay. *PLoS One.* 2008;3(11):e3759.
375. DeVos T, Tetzner R, Model F, Weiss G, Schuster M, Distler J, et al. Circulating methylated SEPT9 DNA in plasma is a biomarker for colorectal cancer. *Clin Chem.* 2009;55(7):1337–46.
376. Li Y, Tollefsbol TO. DNA methylation detection: bisulfite genomic sequencing analysis. *Methods Mol Biol.* 2011;791:11–21.
377. Tusnádý GE, Simon I, Váradi A, Arányi T. BiSearch: primer-design and search tool for PCR on bisulfite-treated genomes. *Nucleic Acids Res.* 2005;33(1):e9.
378. Darst RP, Pardo CE, Ai L, Brown KD, Kladde MP. Bisulfite sequencing of DNA. *Curr Protoc Mol Biol.* 2010;Chapter 7:Unit–7.917.
379. Ehrich M, Zoll S, Sur S, van den Boom D. A new method for accurate assessment of DNA quality after bisulfite treatment. *Nucleic Acids Res.* 2007;35(5):e29.
380. Leontiou CA, Hadjidaniel MD, Mina P, Antoniou P, Ioannides M, Patsalis PC. Bisulfite conversion of DNA: Performance comparison of different kits and methylation quantitation of epigenetic biomarkers that have the potential to be used in non-invasive prenatal testing. *PLoS One.* 2015;10(8):1–22.
381. Izzi B, Binder AM, Michels KB. Pyrosequencing evaluation of widely available bisulfite conversion methods: considerations for application. *Med Epigenetics.* 2014;2(1):28–36.
382. Bryzgunova O, Laktionov P, Skvortsova T, Bondar A, Morozkin E, Lebedeva A, et al. Efficacy of bisulfite modification and recovery of human genomic and circulating DNA using commercial kits. *Eur J Mol Bio.* 2013;1(1):1–8.
383. Fuso A, Ferraguti G, Scarpa S, Ferrer I, Lucarelli M. Disclosing bias in bisulfite assay: MethPrimers underestimate high DNA methylation. Chiariotti L, editor. *PLoS One.* 2015 Feb 18;10(2):e0118318.
384. Wojdacz TK, Dobrovic A, Hansen LL. Methylation-sensitive high-resolution melting. *Nat Protoc.* 2008 Nov 20;3(12):1903–8.
385. Vynck M, Trypsteen W, Thas O, Vandekerckhove L, De Spiegelaere W. The future of digital polymerase chain reaction in virology. *Mol Diagn Ther.* 2016 Oct 28;20(5):437–47.

386. Trypsteen W, Kiselina M, Vandekerckhove L, De Spiegelaere W. Diagnostic utility of droplet digital PCR for HIV reservoir quantification. *J virus Erad.* 2016;2(3):162–9.
387. Pelizzola M, Ecker JR. The DNA methylome. *FEBS Lett.* 2011;585(13):1994–2000.
388. Laurent L, Wong E, Li G, Huynh T, Tsigos A, Ong CT, et al. Dynamic changes in the human methylome during differentiation. *Genome Res.* 2010;20:320–31.
389. Genereux DP, Johnson WC, Burden AF, Stöger R, Laird CD. Errors in the bisulfite conversion of DNA: Modulating inappropriate- and failed-conversion frequencies. *Nucleic Acids Res.* 2008;36(22).
390. Malatinkova E, De Spiegelaere W, Bonczkowski P, Kiselina M, Vervisch K, Trypsteen W, et al. Impact of a decade of successful antiretroviral therapy initiated at HIV-1 seroconversion on blood and rectal reservoirs. *Elife.* 2015;4(OCTOBER2015):e09115.
391. Jordan A, Bisgrove D, Verdin E. HIV reproducibly establishes a latent infection after acute infection of T cells in vitro. *EMBO J.* 2003 Apr 15;22(8):1868–77.
392. Li L-Cc, Dahiya R. MethPrimer: designing primers for methylation PCRs. *Bioinformatics.* 2002 Nov 1;18(11):1427–31.
393. Rutsaert S, De Spiegelaere W, Van Hecke C, De Scheerder M-A, Kiselina M, Vervisch K, et al. In-depth validation of total HIV-1 DNA assays for quantification of various HIV-1 subtypes. *Sci Rep.* 2018;8(1):17274.
394. Foley B, Leitner T, Apetrei C, Hahn B, Mizrahi I, Mullins J, et al. HIV Sequence Compendium 2016. Theor Biol Biophys Group, Los Alamos Natl Lab NM, LA-UR-16-25625. 2016;
395. Krueger F, Andrews SR. Bismark: A flexible aligner and methylation caller for Bisulfite-Seq applications. *Bioinformatics.* 2011;27(11):1571–2.
396. Akalin A, Kormaksson M, Li S, Garrett-Bakelman FE, Figueroa ME, Melnick A, et al. methylKit: a comprehensive R package for the analysis of genome-wide DNA methylation profiles. *Genome Biol.* 2012 Oct 3;13(10):R87.
397. R Development Core Team. R: A Language and Environment for Statistical Computing. *R Found Stat Comput.* 2008;1(2.11.1):2673.
398. McCullagh P, Nelder JA. *Generalized Linear Models*. 2nd dition. London: Chapman and Hall; 1989.
399. Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple. Vol. 57, *Journal of the Royal Statistical Society. Series B (Methodological)*. 1995.
400. Du P, Zhang X, Huang C-C, Jafari N, Kibbe WA, Hou L, et al. Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis. *BMC Bioinformatics.* 2010 Dec 30;11(1):587.
401. LaMere SA, Chaillon A, Huynh C, Smith DM, Gianella S. Challenges in Quantifying Cytosine Methylation in the HIV Provirus. *MBio.* 2019 Feb 22;10(1):e02268-18.
402. Li E. Chromatin modification and epigenetic reprogramming in mammalian development. *Nat Rev Genet.* 2002 Sep;3(9):662–73.
403. Gomez D, Shankman LS, Nguyen AT, Owens GK. Detection of histone modifications at specific gene loci in single cells in histological sections. *Nat Methods.* 2013 Feb;10(2):171–7.
404. Bird A. DNA methylation patterns and epigenetic memory. *Genes Dev.* 2002;16(1):6–21.
405. Goll MG, Bestor TH. Eukaryotic cytosine methyltransferases. *Annu Rev Biochem.* 2005 Jun 13;74(1):481–514.
406. Chen Z, Riggs AD. DNA Methylation and Demethylation in Mammals. *J Biol Chem.* 2011 May 27;286(21):18347–53.
407. Deaton AM, Webb S, Kerr ARW, Illingworth RS, Guy J, Andrews R, et al. Cell type-specific DNA methylation at intragenic CpG islands in the immune system. *Genome Res.* 2011 Jul;21(7):1074–86.
408. Illingworth RS, Gruenewald-Schneider U, Webb S, Kerr ARW, James KD, Turner DJ, et al. Orphan CpG Islands Identify Numerous Conserved Promoters in the Mammalian Genome. Reik W, editor. *PLoS Genet.* 2010 Sep 23;6(9):e1001134.
409. Long HK, Sims D, Heger A, Blackledge NP, Kutter C, Wright ML, et al. Epigenetic conservation at gene regulatory elements revealed by non-methylated DNA profiling in seven vertebrates. *Elife.* 2013 Feb 26;2:e00348.
410. Csankovszki G, Nagy A, Jaenisch R. Synergism of Xist RNA, DNA methylation, and histone hypoacetylation in maintaining X chromosome inactivation. *J Cell Biol.* 2001 May 14;153(4):773–84.
411. Sharp AJ, Stathaki E, Migliavacca E, Brahmachary M, Montgomery SB, Dupre Y, et al. DNA methylation profiles of human active and inactive X chromosomes. *Genome Res.* 2011 Oct 1;21(10):1592–600.

412. Clemson CM, McNeil JA, Willard HF, Lawrence JB. XIST RNA paints the inactive X chromosome at interphase: evidence for a novel RNA involved in nuclear/chromosome structure. *J Cell Biol.* 1996 Feb 1;132(3):259–75.
413. Heard E. Delving into the diversity of facultative heterochromatin: the epigenetics of the inactive X chromosome. *Curr Opin Genet Dev.* 2005 Oct 1;15(5):482–9.
414. Levisky JM, Shenoy SM, Pezo RC, Singer RH. Single-cell gene expression profiling. *Science* (80- ). 2002 Aug 2;297(5582):836–40.
415. Schindelin J, Arganda-Carreras I, Frise E, Kaynig V, Longair M, Pietzsch T, et al. Fiji: an open-source platform for biological-image analysis. *Nat Methods.* 2012 Jul 1;9(7):676–82.
416. Holley RW. Control of growth of mammalian cells in cell culture. *Nature.* 1975;258(5535):487–90.
417. Gratzner HG. Monoclonal antibody to 5-bromo- and 5-iododeoxyuridine: A new reagent for detection of DNA replication. *Science* (80- ). 1982;218(4571):474–5.
418. Koch A, Joosten SC, Feng Z, De Ruijter TC, Draht MX, Melotte V, et al. Analysis of DNA methylation in cancer: Location revisited. *Nat Rev Clin Oncol.* 2018 Jul 17;15(7):459–66.
419. Spivak AM, Planelles V. Novel Latency Reversal Agents for HIV-1 Cure. *Annu Rev Med.* 2018 Jan 29;69(1):421–36.
420. Miller JL, Grant PA. The role of DNA methylation and histone modifications in transcriptional regulation in humans. *Subcell Biochem.* 2013 May 29;61:289–317.
421. Kooistra SM, Helin K. Post-translational modifications: Molecular mechanisms and potential functions of histone demethylases. Vol. 13, *Nature Reviews Molecular Cell Biology.* 2012. p. 297–311.
422. Zhao Y, Sun H, Wang H. Long noncoding RNAs in DNA methylation: New players stepping into the old game. Vol. 6, *Cell and Bioscience.* BioMed Central Ltd.; 2016.
423. Trypsteen W, Mohammadi P, Van Hecke C, Mestdagh P, Lefever S, Saeys Y, et al. Differential expression of lncRNAs during the HIV replication cycle: an underestimated layer in the HIV-host interplay. *Sci Rep.* 2016 Dec 26;6(1):36111.
424. Courtney DG, Tsai K, Bogerd HP, Kennedy EM, Law BA, Emery A, et al. Epitranscriptomic Addition of m5C to HIV-1 Transcripts Regulates Viral Gene Expression. *Cell Host Microbe.* 2019 Aug 14;26(2):217–227.e6.
425. Jain M, Olsen HE, Paten B, Akeson M. The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community. *Genome Biol.* 2016;17(1):239.



# Summary

Epigenetics can be described as the bridge between genotype and phenotype: a cellular regulatory mechanism that changes the final expression of a locus or chromosome, and thus the phenotype of a cell, without changing the underlying DNA sequence by reversible, mitotically inheritable modifications. One of the most studied epigenetic modifications is DNA methylation, the addition of a methyl group to cytosine residues. This modification is involved in development, cell differentiation, X-chromosome inactivation and several diseases as cancer, Alzheimer's and infectious diseases as HIV.

An individual infected with HIV-1 can successfully suppress viral replication and viral spread as long as antiretroviral therapy is consistently taken. Nevertheless, this virus causes incurable infections, due to the formation of a latent viral reservoir. This reservoir consists of cells, carrying replication competent proviruses, that are stably integrated into the genome of the host cell, but do not actively produce new infectious virions. This reservoir can be (re)activated to fuel a viral rebound once therapy is interrupted and is generally assumed to be the last hurdle to find a HIV-1 cure.

This proviral latency is at least partially epigenetically regulated, and several studies showed that DNA methylation of the provirus is one of the latency influencing factors in cultured cells. However, this involvement is never clearly shown in patients. Since the understanding of every piece of the multifactorial puzzle of the HIV-1 latency will be crucial in order to find a targeted approach for an HIV-1 cure, understanding the role of DNA methylation in latency initiation and stability will be crucial.

Despite the high amount of studies analyzing DNA methylation, methods to measure these epigenetic profiles suffer from technical issues. Consequently, a lot of questions about DNA methylation regulation in clinical settings remain unanswered, and analysis of this modification is barely used in routine clinical screening.

In this thesis, we describe the development of two DNA methylation analysis methods: a **sodium bisulfite-based HIV-1 proviral DNA methylation assay**, which was used to analyze HIV-1 proviral DNA methylation profiles in a **large, well-characterized patient cohort**, and a **single cell epigenetic visualization assay (EVA)**.

Sodium bisulfite treatment is the gold standard DNA methylation analysis method which converts epigenetic differences to genetic differences by modifying cytosine, but not methylated cytosine to uracil. This method suffers from DNA fragmentation and degradation, leading to high DNA loss. Since HIV-1 in patients on therapy infects a very low amount of cells, this DNA loss hinders this methylation profiling severely. Moreover, the genome of HIV-1 is heavily prone to mutations, impeding data analysis of this specific DNA methylation analysis method.

The method was optimized to address these drawbacks of bisulfite treatment, and was subsequently used to measure the methylation profile of 72 infected individuals, divided in 4 cohorts: early treated individuals, late treated individuals, long-term non-progressors and acute seroconverters. We showed that, in patients, intragenic methylation was much more abundant compared to promoter methylation, which is the opposite of what is found in most experiments using cultured cells. Moreover, we showed that acute seroconverters (patients with high viral replication activity), have decreased intragenic DNA methylation, and slightly increased promoter methylation compared to patients controlling viral replication.

Evidence is growing that cell-to-cell epigenetic variability is more abundant than previously thought. Moreover, studies showed that analysis of epigenetic modifications using bulk cell DNA masks some of the real signals, leading to wrong epigenetic regulation models. Therefore, single cell epigenetic assays, preferably able to measure several epigenetic modifications and/or link them to RNA transcription are needed. EVA, the single cell assay developed during this PhD, is a versatile single cell epigenetic analysis assay that is capable of combining an epigenetic measurement with RNA transcription analysis. It can analyze allele-specific epigenetic states of specific genes and can be used in different cell types.

Altogether, these assays open perspectives for future epigenetic research, not only for HIV-1, where it already showed the importance of intragenic DNA methylation, but also in several other research contexts.

# Samenvatting

Epigenetica kan gezien worden als de brug tussen het geno- en fenotype: cellulaire regulerende mechanismen die de expressie van een genomisch locus of chromosoom kunnen aanpassen om zo het fenotype van een cel te modifieren. Dit alles gebeurt zonder de onderliggende DNA-sequentie aan te passen, maar door reversibele, mitotisch overerfbare modificaties. Een van de meest bestudeerde epigenetische modificaties is DNA methylatie, de aanhechting van een methylgroep aan een cytosine-residu. Deze modificatie is betrokken in de ontwikkeling van organismen, cel differentiatie, X-chromosoom inactivatie, maar ook in verschillende ziekten zoals allerhande kankers, Alzheimer en infectieziekten als HIV.

Personen die geïnfecteerd zijn met HIV-1 kunnen de virale replicatie en verspreiding succesvol onderdrukken zolang ze de antiretrovirale medicatie zorgvuldig innemen. Desalniettemin veroorzaakt dit virus een ongeneeslijke infectie door de vorming van een latent viraal reservoir. Dit reservoir bestaat uit cellen die replicatie-competente provirusen bevatten die stabiel zijn geïntegreerd in het genoom van de gastheercel. Echter, deze provirusen produceren geen nieuwe infectieuze virions. Dit reservoir kan te allen tijde worden ge(re)activeerd en zo een virale heropflakking veroorzaken bij stopzetten van de therapie. De aanwezigheid van het virale reservoir wordt dan ook gezien als de laatste hindernis in de zoektocht naar een HIV-genezing.

Deze provirale latentie is op zijn minst gedeeltelijk gereguleerd door epigenetische processen: verschillende studies hebben aangetoond dat DNA methylatie van het provirale genoom één van de factoren is die latentie beïnvloedt *in vitro*. Echter, deze betrokkenheid is nooit duidelijk aangetoond *in vivo*. Het oplossen van de complexe en multifactoriële puzzel over HIV-latentie zal cruciaal zijn om een gerichte methode te vinden om HIV te genezen, en dus zal ook de rol van DNA methylatie in de latentie initiatie en stabiliteit moeten worden ontrafeld.

Ondanks het grote aantal studies over DNA methylatie, ondervinden de verschillende methoden nog talrijke nadelen en technische moeilijkheden. Bijgevolg blijven er nog veel vragen rond DNA methylatie regulatie in de klinische context onbeantwoord, en wordt de analyse van DNA methylatie nauwelijks gebruikt in routine klinische screening.

In deze thesis wordt de ontwikkeling van twee nieuwe DNA methylatie analysemethoden beschreven: een **natrium bisulfiet-gebaseerde HIV-1 provirale DNA methylatie assay**, die werd gebruikt om HIV-1 provirale DNA methylatie profielen te bepalen in **een grote, goed gekarakteriseerde patiënten cohorte**, en een **single-cell epigenetische visualisatie assay (EVA)**.

Natrium bisulfiet behandeling van DNA is de gouden standaardmethode om DNA methylatie te analyseren. Het verandert epigenetische verschillen naar genetische verschillen door cytosine om te zetten naar uracil, zonder gemethyleerd cytosine aan te passen. Deze methode veroorzaakt een enorme DNA-fragmentatie en degradatie, en dus heel veel verlies van stalen. Aangezien HIV-1 patiënten onder therapie weinig geïnfecteerde cellen hebben, bemoeilijkt deze behandeling het methylatie onderzoek van het provirus. Bovendien is HIV-1 heel vatbaar voor mutaties, wat de data-analyse van deze specifieke methode bemoeilijkt.

Onze methode is geoptimaliseerd om met deze nadelen van bisulfiet behandeling om te kunnen gaan en werd vervolgens gebruikt om het methylatieprofiel van 72 HIV-1 geïnfecteerde personen in kaart te brengen. Deze patiënten werden opgedeeld in 4 groepen: patiënten met een snelle opstart van therapie na infectie, patiënten met een late opstart van therapie na infectie, long-term non-progressors (individuen die de infectie onder controle hebben zonder therapie) en patiënten tijdens seroconversie (patiënten die de virale infectie (nog) niet onder controle hebben).

We toonden aan dat intragene DNA methylatie veel meer aanwezig was in patiënten in vergelijking met promotor methylatie, wat het tegenovergestelde resultaat is van vele experimenten *in vitro*. Bovendien toonden we aan dat seroconverters differentiële methylatieprofielen vertoonden in vergelijking met de andere groepen (die de infectie onder controle hadden).

Er is meer en meer bewijs dat intercellulaire epigenetische variabiliteit een belangrijkere rol speelt dan voorheen werd gedacht. Bovendien werd aangetoond dat de analyse van epigenetische modificaties op bulk cel DNA het effectieve signaal (gedeeltelijk) maskeert door deze variatie, en dat dit tot foutieve epigenetische regulatie modellen leidt. Daarom zijn single-cell epigenetische analyses cruciaal. Deze zijn in het beste geval ook in staat om verschillende epigenetische modificaties samen te meten en deze te linken aan RNA-transcriptie. EVA, de methode die werd ontwikkeld tijdens dit doctoraat, is een veelzijdige single-cell epigenetisch visualisatie assay die in staat is om de analyse van epigenetische modificaties rechtstreeks te linken aan RNA-transcriptie. Bovendien is het in staat om op een allel-specifieke manier naar bepaalde genen te kijken, en kan het ook worden gebruikt in verschillende celtypes.

Samengevat, deze assays openen perspectieven voor verder epigenetisch onderzoek, niet enkel voor HIV-1, waar ze reeds geleid hebben tot het inzicht dat intragene methylatie een belangrijke rol speelt, maar ook in verschillende andere onderzoeksvelden.

# CURRICULUM VITAE

## Sam Kint

### Education

2013 – 2015    **Master of Science:** Bioscience Engineering – Cell and Gene Biotechnology  
Ghent University

Master thesis: *“Methylation profiling of HIV-1 to assess the epigenetic regulation of the latency”*

2010 – 2013    **Bachelor of Science:** Bioscience Engineering – Cell and Gene Biotechnology  
Ghent University

### Scientific experience

2015 – 2019    **PhD researcher**  
Biobix & HIV Cure Research Center, Ghent University  
*Exploring methods to measure DNA methylation in the context of HIV-1*

June – August    **Visiting Student**  
2018    Division of Allergy & Infectious Diseases, University of Washington  
March 2019    *Single cell Epigenetic Visualization Assay development*

### Publications

Kint S., De Spiegelaere W., De Kesel J., Vandekerckhove L. & Van Criekinge W.

***Evaluation of bisulfite kits for DNA methylation profiling in terms of DNA fragmentation and DNA recovery using digital PCR.***

PLOS ONE 2018, 13(6).

Kint S., Trypsteen W., De Spiegelaere W., Malatinkova E., Kinloch-de-Loes S., De Meyer T., Van Criekinge W. & Vandekerckhove L.

***Underestimated effect of intragenic HIV-1 DNA methylation on viral transcription in HIV infected patients.***

Submitted to Clinical Epigenetics (December 27<sup>th</sup>, 2019)

Kint S., Van Crieke W., Vandekerckhove L., De Vos W.H., Bomzstyk K., Krause D.S. & Denisenko O.

***Single cell epigenetic visualization assay, EVA***

Submitted to Nucleic Acids Research (January 22<sup>nd</sup>, 2020)

Kint S., Van Hecke C., Cole B., Vandekerckhove L. & Sips M.

***Highlights from the HIV Cure and Reservoir Symposium, 11–12 September 2017, Ghent, Belgium***

Journal of Virus Eradication 2018, 4: 55–58

Rutsaert S., Steens J, Gineste P., Cole B., Kint S., Barrett P., Tazi J., Scherrer D., Ehrlich H. & Vandekerckhove L.

***Safety, tolerability and impact on viral reservoirs of the addition to antiretroviral therapy of ABX464, an investigational antiviral drug, in individuals living with HIV-1: a Phase IIa randomised controlled study***

Journal of Virus Eradication 2018, 5: e1–e13

Presentations

September 2018

**Frontiers in Retrovirology 2018**

Leuven, Belgium

Poster presentation: HIV-1 proviral DNA methylation profiles show differential methylation between patient groups

March 2019

**Conference on Retroviruses and Opportunistic Infections (CROI) 2019**

Seattle, Washington

Poster presentation: HIV proviral DNA methylation in seroconverters, controllers and ART-treated patients

# Dankwoord

*“Competitive rowing is an undertaking of extraordinary beauty preceded by brutal punishment.”*

De voorbije vier jaar doctoreren waren in mijn ogen het best te vergelijken met een roeiwedstrijd. In dit geval een lange afstandswedstrijd. Je schrijft je enthousiast in, maar kort na de start beseft je pas waaraan je bent begonnen. Na 500 meter is elke spier in je lichaam verzuurd, maar je bent nog niet aan een tiende van de wedstrijd. Halverwege krijg je nieuwe moed. De wind zit even mee, en je besluit de tweede helft harder te roeien dan je voordien deed. Na de volgende bocht blaast de wind echter alweer recht tegen je in. Er lijkt geen einde aan te komen. Na elke bocht hoop je dat het de laatste is, hoewel je weet dat er nog vele komen. Na elke gemiste slag laad je jezelf weer op om je snelheid terug op te bouwen. Aan de finale 500m is het einde eindelijk in zicht. Een laatste sprint om alle energie er nog één keer uit te persen. Eenmaal over de meet, overheerst het gevoel van voldoening. Een intens geluk, en meteen blik je positief terug op de wedstrijd. Je overweegt je zelfs in te schrijven voor een volgende, misschien zelfs langere wedstrijd. Uiteindelijk is ‘pijn’ is toch gewoon ‘fijn’ met een ‘p’.

Om deze inspanning tot een goed einde te brengen, heb ik onderweg heel wat steun gekregen. Want zonder team, coaching of supporters bereikt niemand de finish.

Lieselot, de stuurvrouw: de eerste taak van de stuurvrouw is het sturen van de boot. Echter naast deze niet evidente taak, speelt de stuurvrouw een nog belangrijkere rol: ze houdt de controle over alles wat in de boot gebeurt. Ze is de kapitein die de roeiers door en door kent, en ze gebruikt de sterktes en zwaktes om het beste uit elke roeier te halen. Ze weet als geen ander hoe een uitgeputte roeier net harder te laten doorgaan in plaats van te laten verzwakken, zelfs als alles al verloren lijkt.

*“In short, a good coxswain is a quarterback, a cheerleader, and a coach all in one. He or she is a deep thinker, canny like a fox, inspirational, and in many cases the toughest person in the boat.”*

Mijn familie, de meest fervente supporters: ze kijken toe, soms met verwondering, soms met medelijden. Hoewel ze je pijn niet kunnen voelen of begrijpen, blijven ze je onvoorwaardelijk steunen. Ook al mijn vrienden, de minder fervente supporters: ze volgen de wedstrijd niet op de voet, maar zijn desondanks wel geïnteresseerd in de uitslagen en tussentijden. Zij bezorgden me af en toe het nodige vleugje meewind, waardoor ik de kans kreeg om te recupereren, mijn pijn even te vergeten, en het volgende deel van de wedstrijd weer aan te vangen. Samen hebben we veel prachtige momenten kunnen meemaken in de afgelopen vier jaar. Bedankt aan de roeiers (zowel in Gent als in Seattle), lelemaal Gazpacho, bio-ingenieur vrienden, scouts vrienden, en vrienden in Seattle voor alle ontspannende momenten. Die afleiding is noodzakelijk, want je supporters beseffen dat het vaak lastiger is dan het lijkt.

*“Rowing is like a beautiful duck. On the surface it is all grace, but underneath the bastard’s paddling like mad!”*

Mijn collega's, de teamleden: in het UZ, bij Biobix, en in het FFW. De teamleden zijn een van de voornaamste redenen dat roeiers tijdens een wedstrijd niet stoppen. Samen starten, samen afzien en samen finishen. Deze wedstrijd mag dan nog een skiff-wedstrijd zijn geweest, de teamleden hebben alle voorbereidingen meegemaakt, en samen werd het afzien altijd een stuk draaglijker. Na iedere zware training of wedstrijd vergeet iedereen de geleden pijn en ben je blij dat je als team de finish bereikt hebt, en al snel plan je de volgende sessie.

*“One of the fundamental challenges in rowing is that when any one member of a crew goes into a slump the entire crew goes with him.”*

Mijn promotoren (Wim, Linos, Wim) en Oleg, de coaches: zij volgen mee langs de zijlijn en overzien elke training en elke wedstrijd. Vanaf de zijlijn geven ze instructies, maar jij bent uiteindelijk diegene die hiermee aan de slag moet, wat in realiteit niet altijd even gemakkelijk blijkt. Maar Ze zorgen ervoor dat de roeiers zich enkel moeten focussen op roeien, en sturen dit bij waar nodig.

*“Every good rowing coach, in his own way, imparts to his men the kind of self-discipline required to achieve the ultimate from mind, heart, and body.” — George Yeoman Pocock*

*“Rowing is perhaps the toughest of sports. Once the race starts, there are no time-outs, no substitutions. It calls upon the limits of human endurance.” — George Yeoman Pocock*







