School of Medicine

Department of Health Sciences

PhD Program in Medical Sciences and Biotechnology

RNA-mediated Correction of Aberrantly Methylated Genes

Supervisor:

Prof. Claudio Santoro

PhD candidate:

Marta Borchiellini

Cycle XXXII
Academic year 2016-2019

Index

Summary/Sommario

Promoter DNA methylation, which is carried by members of the DNA methyltransferase (DNMT) family, negatively regulates genes transcription. In most cancers, silencing of tumor suppressor genes by aberrant promoter DNA methylation is considered the initiating event to uncontrolled cell growth and tumorigenesis. Currently, the approved hypomethylating protocols are based on two drugs: Azacytidine (AC) and Decitabine (DAC). Unfortunately, cytotoxic and global non-specific demethylation effects limit their clinical application.

The identification of thousands of noncoding RNAs (ncRNAs), for long time deemed as merely "transcriptional noise", has provided an additional layer of control to many biological processes. The ncRNA revolution has not only impacted our understanding on gene regulation but has laid the foundations for a new "RNA-based" therapeutics era.

Short activating RNAs (saRNAs) are small noncoding RNAs able to trigger an RNA activation mechanism. They can be endogenously expressed or artificially designed to recognize targeted sequences and promote gene-specific activation upon ectopic delivery. Thus, saRNAs might represent a novel strategy to selectively induce gene expression with profound implications for basic research and clinical applications. Yet, their mechanism of action remains elusive.

Our group has previously identified a novel class of functional RNAs, named DNMT1-interacting RNAs (DiRs), able to inhibit the DNMT1 activity and to prevent DNA methylation in a *locus* specific manner. On the example of the *CEBPA* locus, we showed that a specific DiR, the extra-coding *CEBPA* (*ecCEBPA*), prevented downstream *CEBPA* methylation by forming complexes with DNMT1. Expression of *ecCEBPA* in cells in which *CEBPA* is methylated resulted in promoter demethylation and gene activation. Collectively, our data suggested that RNAs similar to *ecCEBPA* could protect their related gene *loci* from DNA methylation.

Provided their gene-specific mode of action, we investigated whether saRNAs could function as DiR-mimicking molecules, thereby modifying DNA methylation and restoring expression of silenced gene *loci*. We modeled our study on two genes, frequently undergoing aberrant DNA methylation in lung cancer and/or hematopoietic malignancies: *CEBPA* and *P15,* respectively. The former is a critical transcription factor that controls tissue-specific gene expression and proliferation arrest, the latter is a cell cycle regulator and tumor suppressor gene.

By using number of cell lines displaying varied *CEBPA* or *P15* expression levels, from undetectable to low, paralleled by similar promoter DNA methylation profiles, from fully- to hemi-methylated, we showed that transcriptional activation by saRNAs promotes, to different extent, DNA methylation changes of the targeted *loci*. Thus, saRNAs may represent the first demethylating tool to correct DNA methylation in a gene-specific fashion.

In conclusion, this study delineates a new mechanistic action of saRNAs and suggest a novel RNA-based therapeutic strategy, aiming at reestablishing proper expression of genes aberrantly silenced by DNA methylation.

La metilazione del promotore, meccanismo catalizzato dalla famiglia di enzimi chiamati DNA metiltrasferasi (DNMT), è associata ad una ridotta trascrizione genica. Nella maggior parte dei tumori, il silenziamento di soppressori tumorali a motivo di una metilazione aberrante del promotore è considerato il meccanismo iniziale che porta a crescita cellulare incontrollata ed a tumorigenesi. Attualmente, i protocolli ipometilanti si basano sull'utilizzo di due farmaci: Azacitidina (AC) e Decitabina (DAC). Purtroppo, la loro citotossicità e l'effetto ipometilante non specifico limitano il loro uso clinico.

Il riconoscimento di migliaia di RNA non codificanti, per molto tempo considerati semplicemente come "rumore trascrizionale di fondo", ha offerto un ulteriore livello di controllo in molti processi biologici. La rivoluzione degli RNA non codificanti non solo ha cambiato la nostra comprensione della regolazione genica ma ha anche posto le fondamenta per un'era terapeutica basata sull'RNA. Gli *short activating* RNA (saRNAs) sono piccoli RNA non codificanti capaci di attivare meccanismi trascrizionali. Possono essere endogeni oppure disegnati artificialmente al fine di riconoscere sequenze target e favorire l'attivazione specifica di geni in seguito alla loro trasfezione. Tuttavia, il loro meccanismo di azione rimane non del tutto compreso. Il nostro gruppo ha identificato una nuova classe di RNA funzionali, definiti *DNMT1-interacting RNAs* (DiRs), in grado di inibire l'attività di DNMT1 e prevenire la metilazione del DNA tramite un meccanismo *locus* specifico. Usando il *locus* di *CEBPA* come esempio, abbiamo dimostrato che uno specifico DiR, l'extra-coding *CEBPA* (*ecCEBPA*), è in grado di prevenire la metilazione del gene *CEBPA* grazie alla formazione di un complesso con DNMT1. L'espressione di *ecCEBPA* nelle cellule in cui *CEBPA* è metilato risulta in una demetilazione del promotore ed attivazione genica. Globalmente, i nostri dati suggeriscono che RNA simili ad *ecCEBPA* potrebbero proteggere i rispettivi *loci* genici dalla metilazione del DNA.

Considerando il loro meccanismo di azione gene-specifico, abbiamo valutato se gli saRNAs potessero agire in maniera simile ai DiRs, modificando la metilazione del DNA e ristabilendo l'espressione di geni non trascritti. Abbiamo condotto la nostra ricerca su due geni che spesso sono caratterizzati da metilazione aberrante nel tumore al polmone e/o in malattie ematopoietiche: *CEBPA* e *P15*, rispettivamente. *CEBPA* è un fattore di trascrizione chiave che controlla l'espressione genica tessuto-specifica ed arresta la proliferazione cellulare, *P15* invece è un regolatore del ciclo cellulare ed anche un soppressore tumorale.

Usando linee cellulari con diversa espressione di *CEBPA* e *P15*, da non rilevabile a bassa, insieme anche ai profili di metilazione, da completamente a parzialmente metilato, abbiamo dimostrato che l'attivazione trascrizionale tramite saRNAs favorisce, in misura diversa, cambiamenti nel profilo di metilazione dei geni target. Di conseguenza, gli saRNAs possono rappresentare il primo approccio demetilante per correggere difetti di metilazione in maniera gene-specifica.

In conclusione, tale studio identifica un nuovo meccanismo di azione degli saRNAs e suggerisce una nuova strategia terapeutica basata sull'RNA, al fine di ristabilire una corretta espressione di geni la cui trascrizione è soppressa a motivo di una metilazione difettiva.

Introduction

1 Epigenetics

The term epigenetics was originally coined in the 17th century by the physician and physiologist William Harvey to indicate the gradual development of the embryo from a homogeneous to a heterogeneous material, referred to as "epigenesis" [1]. Later on, in the 1940s, Conrad Waddington used the term "epigenetics" to explain the relationship between the genotype, defined as the whole genetic system of an organism, and the phenotype, indicating the entire set of characteristics that an organism develops over time [2]. Waddington established the first causal relationship between genes and their outcomes by introducing the concept of the "epigenetic landscape" as "the various developmental pathways that undifferentiated cells (sharing identical genotype) might take toward differentiation" (Figure 1) [3]. In other words, he described how the static information written in the form of nucleotide sequences is dynamically translated into tissues and organs, thus driving cell fate decisions [4]. In the last two decades, the definition of epigenetics has evolved from "the study of mitotically and/or meiotically heritable changes in gene function that cannot be explained by changes in DNA sequence" [5] to "the structural adaptation of chromosomal regions so as to register, signal or perpetuate altered activity states", which is inclusive of all the stable or transient chromosomal markers arising in response to different stimuli [6].

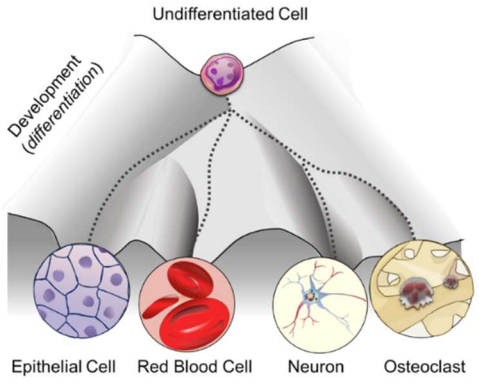

Figure 1. An outline depicting cell-fate plasticity according to the Waddington's epigenetic landscape (inspired by the model proposed by Waddington [3]).

2 DNA methylation

DNA methylation is a key epigenetic signature implicated in regulation of gene expression that occurs predominantly within CpG dinucleotides. DNA methylation can also regulate other mechanisms, such as DNA binding of transcription factors, nucleosome positioning, i.e. limiting the access of protein complexes to DNA regulatory regions, and gene splicing [7]. CpG dinucleotides are under-represented in the mammalian genome (1%), but tend to cluster in CpG-rich regions called CpG islands (CGIs), located in the proximity of the transcription start sites (TSSs) of the majority (70%) of human protein-coding genes [8, 9]. CGIs are stretches of DNA sequences of 200 nucleotides or greater [10], with the GC ratio observed/expected to be greater than 0.6. Although the bulk of genome is methylated at 70–80% of its CpGs, CGIs are mostly unmethylated in somatic cells [10, 11].

DNA methylation is mediated by members of the DNA methyltransferase (DNMT) family that can covalently transfer a methyl group (CH_3) from the universal donor S-adenosyl-L-methionine (SAM) to the carbon 5-position of the cytosine ring. Conventionally, DNMTs are classified as *de novo* (DNMT3a and DNMT3b) or maintenance (DNMT1) enzymes (Figure 2) [12, 13].

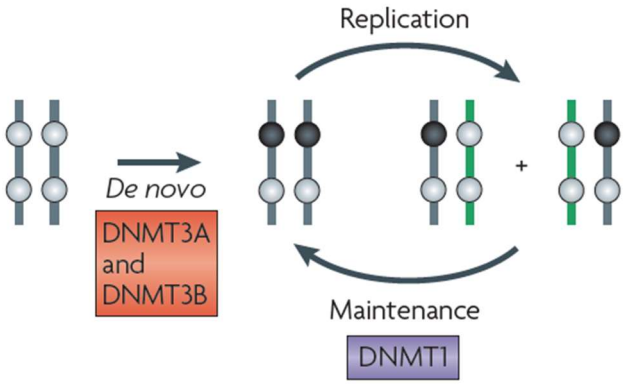

Figure 2. *De novo* and maintenance DNA methylation schema. *De novo* DNA methyltransferases (DNMTs) DNMT3A and DNMT3B establish DNA methylation in germ line and developing embryos, (the black circles represent methylated CpG dinucleotides and the grey circles represent unmethylated CpG dinucleotides), whereas DNMT1, which has a preference for hemimethylated DNA, maintains the methylation pattern throughout replication (Jones and Liang, 2009) [13].

Of note, DNMT2, a member of the methyltransferase family, catalyzes methylation of RNA at position 38 in tRNAAsp GUC [14-16] (Figure 3). DNMT3-like (DNMT3L) is a DNMT3-associated protein lacking an enzymatic domain and interacting with DNMT3a/3b to modulate its activity [17, 18]. In mice, Dnmt1 and Dnmt3b are essential for embryonic development, as DNA methylation changes dramatically. An overall demethylation after fertilization is followed by *de novo* methylation of discrete regions upon implantation [19]. *Dnmt1* knockout

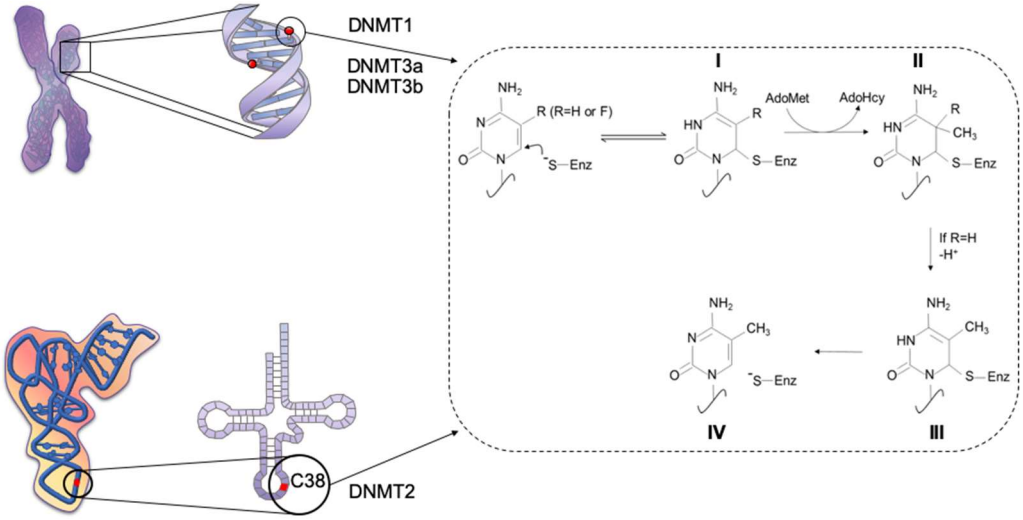

Figure 3. Schematic representation of the methylation reaction catalyzed by the DNA methyltransferases (DNMTs) (adapted from [14]). Shown are the mechanisms proposed for methylation of cytosine by DNMT1, 3a and 3b on DNA (upper left panel) or by DNMT2 on RNA (lower left panel). Briefly, a thiol group (SH) from the binding site of the enzyme provides the nucleophilic attack to position 6 of the cytosine heterocycle, to activate position 5 towards one-carbon transfer (I). The methyl group on position 5 is donated by the coenzyme AdoMet (II). A proton in position 5 of the 5,6-dihydropyrimidine is then removed (II–III), and a consequent β-elimination generate 5-methylcitosyne and free enzyme (IV).

mice show early lethality at embryonic day (E) 9.5, whereas Dnmt3b depletion induces death at E 14.5–18.5, due to developmental impairment. On the contrary, Dnmt3a knockout mice do not display defects in embryonic development, but they do die at 4 weeks of age. Although this binary classification is convenient, the function of the de novo and maintenance DNMTs overlaps in many instances [13, 20, 21].

A number of studies have shown that DNA methylation is not randomly distributed across the genome, but displays regional specificity [22]. Methyl groups promote conformational changes in the major groove of DNA, thus altering protein-DNA binding [23] and, as a result, gene

expression. Most studies have initially focused on the effect of CGI methylation within the promoter and TSS of protein-coding genes. Recently [24], however, more comprehensive genome-wide methylation analyses have started elucidating the role of DNA methylation at CpG clusters within exons, introns, and intergenic sequences, expanding on the previous knowledge of CGIs and leading to the identification of CG shores (regions up to 2 kb from CGI), shelves (regions from 2 to 4 kb from CGI), and open sea (the rest of the genome) regions [25-27].

Hence, DNA methylation needs to be framed in the context of the genomic location. As a proof of concept, methylation of TSS-associated CGIs negatively correlates with gene expression, leading to long-term gene silencing [8], whilst gene-body methylation positively correlates to gene expression [28-31]. Another interesting finding emerging from bulk methylome studies is that non-CGI-CpGs are mostly methylated and therefore less stable than CGIs, due to the tendency of 5-methylcytosine (5mC) to undergo spontaneous or enzymatic deamination to T [32]. The C-to-T transition causes germ-line or somatic mutations, resulting in the depletion of CpGs dinucleotides in the human genome. Methylation at other genomic regions, such as enhancers and insulators, does not follow a specific pattern and may vary in different settings. Enhancers and insulators are long-range regulatory elements able to alter gene expression or protect gene promoters from inappropriate signals, respectively [33, 34]. Aran *et al.* have shown that distal methylation sites in estrogen receptor (ER)-positive breast tumors associates with breast cancer-related gene expression better than promoter methylation [35]. Moreover, Tatetsu *et al*. demonstrated that aberrant methylation of the 17-kb 5′ upstream enhancer of *PU.1* is required for myeloma cell growth [36]. Insulators are bifunctional instead, acting either as a blocking enhancer, by preventing enhancer-mediated transcription, or as barriers, by limiting the advance of nearby heterochromatin that would otherwise silence expression [37]. CCCTC-binding factor (CTCF), an enhancer-blocking protein, does not bind to its DNA

consensus sequence if methylated, as demonstrated for the imprinted *IGF2-H19* locus and *CD45* gene [38]. It follows that DNA methylation can regulate gene expression indirectly by controlling access of enhancers to gene promoters [39]. Finally, CGI shores, with a lower CpG density, have recently emerged as critical regulatory elements affecting gene expression depending on their DNA methylation profile [40].

3 Non coding RNAs

The genome sequencing project revealed that complex organisms have a lower number of protein coding genes than expected. Differences between individuals and species cannot account for variation in protein-coding sequences, but rather for variation in gene expression regulated by a control architecture [41]. In this regard, perhaps one of the biggest surprise of the post-genome era is indeed the great number and diversity of transcripts arising from the previously assumed wastelands of non-coding regions. The first efforts in mapping transcriptional events to the genome led to the identification of thousands of long noncoding RNA (lncRNA) transcripts resulting in potential novel RNA regulatory elements [42]. Later on, other classes of regulatory noncoding RNAs were discovered, whose biological functions are still largely unknown. Thus, the growing number and increasing pace of discovery of new noncoding RNAs are followed by the challenge of their definition and annotation [43]. Noncoding RNAs (ncRNAs) can be classified either by function, as housekeeping (ribosomal RNA and transfer RNA) or regulatory RNAs, either by length, as lncRNAs or small RNAs (snRNAs).

3.1 Long noncoding RNAs (lncRNAs)

The general term long noncoding RNAs (lncRNA) defines a highly heterogeneous and poorly conserved group of transcripts >200 nucleotides (nt) in length that do not contain protein-coding sequences. The size threshold represents an arbitrary but convenient cutoff that excludes other classes of small infrastructural and regulatory RNAs [43, 44]. Some annotated lncRNAs

are not functional and do not confer any fitness advantage to the cell, e.g., spurious RNAs produced during transcription and showing tissue-specificity due to variations in chromatin states within different cellular contexts [45]. Despite the idea that the majority of lncRNAs are not functional, there exists a percentage of functional polyadenylated and non-polyadenylated lncRNAs, including sense or antisense, intronic, intergenic or enhancer, and overlapping transcripts with respect to protein-coding *loci*. They have been proposed to accomplish a variety of functions, such as gene regulation. lncRNAs indeed can be categorized into transcripts that regulate chromatin structure and gene expression in *cis* versus those that, by leaving the site of transcription, act in *trans* [46]. *Cis*-acting lncRNAs can regulate gene expression through three potential different mechanisms: **(**a**)** the lncRNA itself regulates the expression of neighboring genes by recruiting activating or repressing factors, **(**b**)** transcriptional or splicing events of the lncRNA acquire a gene-regulation functionality that is independent of the sequence of the transcripts, or (c) the recruitment of DNA elements within

the lncRNA's promoter is eventually responsible for the neighbor gene's regulation and is lncRNA's sequence independent (Figure 4).

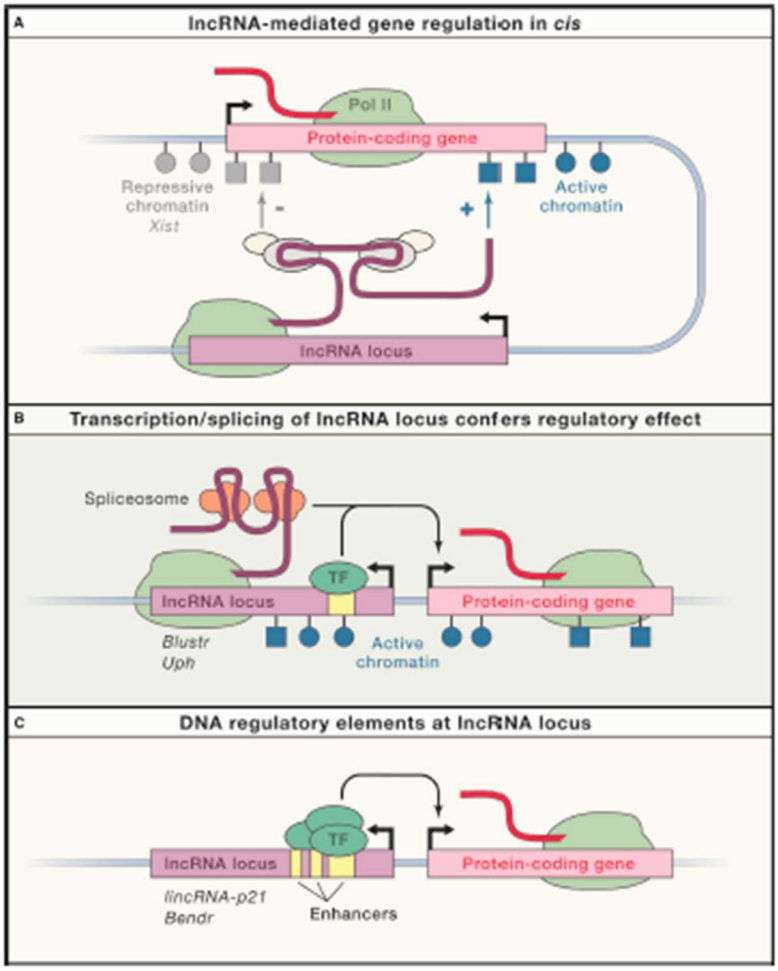

Figure 4. LncRNA's function in local gene regulation . lncRNAs can control the expression of nearby genes in *cis* (A) by sequence-specific functions of the mature lncRNA transcript, (B) by mechanisms of transcription or splicing of an RNA, but the lncRNA itself is not functional or (C) by DNA elements within the lncRNA promoter or gene body that function independently of the transcribed RNA . Pol II, RNA polymerase II; TF, transcription factor (Kopp and Mendell, 2018) [46].

(a) The most known and well-studied example of *cis*-acting lncRNA is the X-inactive-specific-transcript (*Xist*) that orchestrates X chromosome inactivation. In female mammals, during early embryonic development, one of the two X chromosomes is transcriptionally silenced for dosage compensation. The silencing process depends upon transcription of *Xist*, that is transcribed from only one X chromosome, which will be later inactivated. Despite aspects of the Xist-silencing mechanism are still elusive, recent development of focused and innovative

methods able to capture RNA-binding proteins (RBPs) has allowed a step forward in understanding *Xist*'s interactors as well as in re-evaluating previously proposed mechanisms [47]. Therefore, studies have demonstrated that *Xist* directly recruits SMART/histone deacetylase 1 (HDAC1)-associated repressor protein (SHARP or SPEN). This further leads to the recruitment of the SHARP/SPEN-interacting protein silencing mediator for retinoid and thyroid hormone receptors (SMRT) and its interactor histone deacetylase 3 (HDAC3), ultimately inducing histone deacetylation, one of the earliest event in X inactivation [48]. Chen *et colleagues* also demonstrated that *Xist* entails remodeling of the three-dimensional (3D) structure of the X chromosome by recruiting it to the nuclear lamina and, by doing so, *Xist* is able to spread to actively transcribed genes across the X [49].

(**b**) In contrast with the activity of *Xist*, where sequence recognition plays a key role in recruiting protein interactors, Engreitz *et colleagues* demonstrated that other *cis*-acting lncRNAs exert local effects on neighboring genes without sequence-specific function [50]. By analyzing 12 lncRNA *loci* in mouse embryonic stem cells (mES cells), 5 of these 12 lncRNA *loci* influence the expression of adjacent genes, by ultimately acting in *cis*. As one example, the *linc1536*, also called *Bendr* (Bend4-regulating effects not dependent on the RNA), regulates expression of the adjacent *Bend4* gene by DNA regulatory elements within the 750 bp promoter-proximal region. Thus, knocking-out the *Bendr* promoter reduces the expression of the neighboring gene by 57%, whereas introducing polyadenylating signals (pAS) into the first intron of the locus does not exert any effect.

The process of transcription and the 5' splice in the *Blustr* locus (bivalent locus (*Sfmbt2)* is upregulated by the splicing and transcription of an RNA, also known as linc1319) are both necessary for regulating the neighboring gene *Sfmbt2*, located 5 kb upstream. Hence, *Blustr* is essential for *Sfmb2* activation, but the entire process does not require precise RNA sequence beyond the initial splice signals.

(**c**) As for the *Bendr* locus, in most of lncRNA *loci*, effects on adjacent genes are mediated by enhancer-like function of DNA elements located within promoters and functioning independently of the production or sequence of the transcript [50]. *lincRNA-p21* is a nuclear intergenic noncoding RNA that neighbors the *CDKN1A* in human and mouse. An initial study defines *lincRNA-p21* as a p53-dependent trans-acting lncRNA, able to mediates apoptosis upon p53 signal. However, more recent studies suggest a *cis*-acting regulatory function. Indeed, even in tissues with no detectable *lincRNA-p21* transcript, deletion of the *locus* dramatically affects the expression of neighboring genes, as *Cdkn1a.* Further analysis has defined DNA enhancer elements within promoter of the locus, which are responsible of its regulatory activity. Therefore, the *lincRNA-p21* locus can exert an RNA-independent function on gene expression [51].

lncRNAs can also act in *trans* therefore regulating the expression of genes located at distant regions from their transcriptional start site (TSS), influencing chromatin structure and, interacting with proteins and other RNAs. *Cis* and *trans* lncRNA's activity are not mutually exclusive. The *Xist* has indeed the ability to shape X chromosome structure during X chromosome inactivation process, in addition to *cis*-acting gene silencing mechanism [49, 52]. *Firre*, a lncRNA transcribed from the active X chromosome and escaping X chromosome inactivation, has been detected, in addition to its transcriptional site, at other autosomal *loci* using RNA antisense purification (RAP) and RNA fluorescent in situ hybridization (FISH) in murine embryonic stem cells. Moreover, *Firre* expression is essential to maintaining the Xi structure and its epigenetic features [53].

3.2.1 DNMT1-interacting RNAs (DiRs)

Previous work has shown the existence of noncoding RNAs able to regulate methylation patterns of their respective coding transcripts [54]. These novel class of noncoding RNAs, namely DNMT1-interacting RNAs (DiRs), interact with and inhibit DNMT1, hence preventing

DNA methylation. Using as example the *CEBPA locus*, Di Ruscio *et al.* demonstrated the existence of an overlapping, sense-orientated transcript, the extra coding *CEBPA* (*ecCEBPA*), whose transcription regulated the expression of the corresponding coding *CEBPA* mRNA (Figure 5).

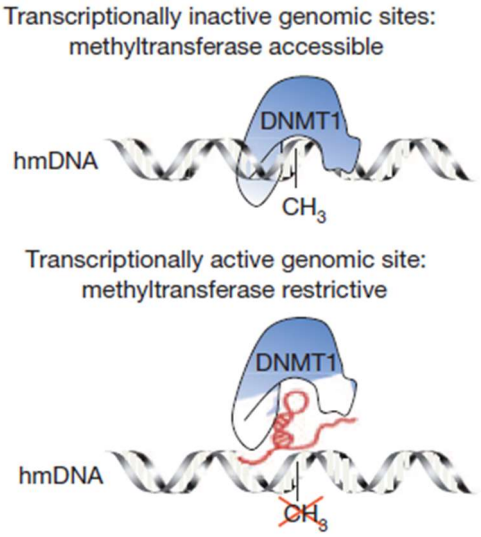

Figure 5. Schematic of DNMT1 sequestration. Top panel, DNMT1 can access transcriptionally inactive hemi-methylated sequences. Bottom panel, DNMT1 cannot access transcriptionally active hemi-methylated sequences. (Di Ruscio *et al,* 2013) [54].

Characterization of *ecCEBPA* demonstrated that DiRs were non-polyadenylated, enriched in the nuclear fraction, and overlapped the gene in the sense orientation. From genome-wide analysis, they found that, besides *CEBPA,* DiRs regulated other *loci* and were associated with hypomethylated regions. The discovery of DiRs provided a new mechanism showing how genes can be regulated by noncoding transcripts arising around a gene body..

3.3 Small noncoding RNAs (sncRNAs)

All noncoding RNA species smaller than 200 nt can be classified as small noncoding RNAs. However, what defines eukaryotic small noncoding RNAs in the RNA silencing pathway is their limited size (20-30 nt) and their association with Ago-family proteins [55]. At least three classes of snRNAs can be identified, based on their mechanism of biogenesis and the type of Ago-protein they are associated with: microRNAs (miRNAs), endogenous short interfering

RNAs (endo-siRNA) and piwi-interacting RNAs (piRNAs). An additional subgroup, the short activating RNAs (saRNAs), that positively regulate gene expression, will be also described.

3.3.1 MicroRNAs (miRNAs)

MicroRNAs (miRNAs or miRs) are short noncoding RNAs able to control expression by targeting specific messenger RNAs (mRNAs). Canonically, miRNAs are encoded by intronic regions of noncoding and coding transcripts, however, some miRNAs can also be encoded by exons. miRNAs are transcribed by RNA Pol II as poly- or mono-cistronic primary miRNAs, also named pri-miRNAs, and containing an m^7G cap at 5' end in addition to a poly(A) tail at 3' untranslated region (3'UTR) [56]; alternatively, some miRNAs requires Pol III, e.g. in the case of the human chromosome 19 miRNA cluster (C19MC) [57]. The miRNA precursors can be several thousand base pairs in length and are processed in the nucleus by the Drosha/DGCR8 complex to produce 70 bp stem-loops (pre-miRNAs). The small hairpin-shape is then exported to the cytoplasm by means of the nuclear export receptor exportin5 and cleaved by Dicer. Dicer's cleavage results in an RNA duplex of about 21 nt in length. Once the RNA duplex is unwound, it is loaded onto an Argonaute protein complex, named RNA-induced silencing complex (miRNP/RISC complex), which identifies the sense and passenger strands in the RNA duplex and guides mature miRNA (single-stranded) to its target mRNA (Figure 6) [58, 59]. miRNAs recognize their target mRNAs by Watson-Crick base pairing: they contain a seed region, centered on

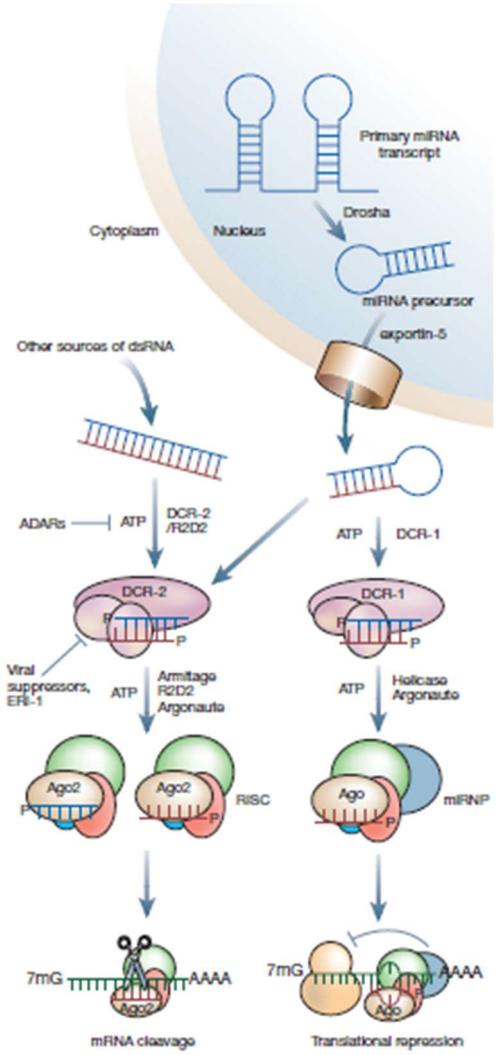

Figure 6. miRNAs' mediated expression regulation. Primary miRNA transcripts are processed to miRNA precursors in the nucleus by Drosha. The pre-miRNA is then exported to the cytoplasm by means of the nuclear export receptor exportin5 and cleaved by Dicer in 21-nt RNA duplex. The RNA duplex is resolved and loaded onto miRNP/RISC complex. Mature miRNAs bind to Ago proteins, which mediate translational repression or cleavage of target mRNAs. Other sources of long dsRNA in the cytoplasm of a cell are processed by Dicer into 21–23 nucleotide dsRNA intermediates. Following unwinding of the dsRNA (mediated by Armitage and R2D2) the single-stranded siRNA-containing RISC is formed. ADARs and exonuclease ERI-1 can regulate dsRNA stability during processing (Meister and Tuschl, 2004) [58].

nucleotides 2-7, that binds within the 3' UTR of the targeted RNA. However, miRNAs responsive elements (MREs) can be found

within 5' UTR sequences as well as coding regions of the target mRNAs. In addition, non-canonical seed pairing can also drive miRNA-mRNA recognition and annealing [60]. The degree of miRNA-mRNA complementarity has been considered a pivotal element of the regulatory mechanism. While perfect complementarity is responsible for Ago-catalyzed cleavage of the targeted mRNA, central mismatches prevent cleavage of the mRNA and promote translational repression of the transcript [61]. The canonical function of miRNAs is that these small RNA molecules repress gene expression by targeting the 3'UTR of targeted mRNAs, thus inducing mRNA's cleavage and/or transcriptional gene silencing in the cytoplasm. However, since the discovery of the first miRNA, lin-4, in the 1993 [62], more recent studies have demonstrated that miRNAs can localize into the nucleus and are able to regulate transcription [63-65]. MiRNAs nuclear localization is strictly dependent on tissue, cell line and condition analyzed. The regulation can be either repression or activation and the miRNA's *modus operandi* can be affected by the presence of TATA box motif, CpG island region and the epigenetic status of the promoter [66]. However, whether miRNAs function depends on DNA methylation status of the target sequence or whether miRNAs can determine changes in DNA methylation profile is still a matter of debate [67, 68].

3.3.2 Short interfering RNAs (siRNAs)

Short interfering RNAs partake, together with miRNAs, in the RNA interference mechanism [69]. Initially, siRNAs were thought to be primarily exogenous, that is directly derived from the virus, transposon or transgene trigger. Subsequently, they were identified upon transgene- and virus-induced silencing in plants, persistent with a natural role in genome defense. Further studies of small-RNA profiling in mice reveal the presence of various types of endogenous siRNAs (endo-siRNAs) in oocytes and, to a lesser extent, in embryonic stem (ES) cells [70]. Endo-siRNAs are slightly shorter than miRNAs: they are 21 nt in length and derive from different sources of double stranded RNAs (dsRNAs), such as transposable elements (TEs), *cis*-natural antisense transcripts (*cis*-NATs), *trans*NATs and hairpin RNA transcripts [55]. Some siRNAs originate from expressed pseudogenes, suggesting that pseudogenes can regulate levels of the founding mRNA by means of the RNAi machinery [71]. Long dsRNA precursors are processed by Dicer into discrete siRNAs duplex whose guide strand will direct the RNA interference (RNAi) complex at the targeted mRNA. During the canonical RNAi, a perfectly complementary mRNA to the siRNA's guide strand is recognized and the targeted mRNA is cleaved at a single site within the duplex siRNA-mRNA. After cleavage, the mRNA fragments are further degraded. In some cases, siRNAs can also recognize targets with imperfect complementarity therefore acting as miRNA-like molecules. Finally, in plants, siRNAs can activate or engage DNA methyltransferases (DNMTs), leading to an increase in DNA methylation and heterochromatin formation [61, 72, 73]. Therefore, siRNAs mediate gene specific silencing resulting in an exquisitely specific mRNA suppression. As a consequence, synthetic siRNAs can be used as a powerful tool for regulating exogenous and endogenous gene expression, as well as innovative therapeutic approach [74]. Conventional synthetic siRNAs consists of 19-21 nt with two nucleotide overhangs at 3' end, usually TT or UU [75]. siRNAs longer than 21 nt require Dicer cleavage to obtain shorter and active

noncoding transcripts. Increasing the length of siRNAs has been demonstrated to potentiate siRNAs' silencing effect. However, dsRNAs longer than 30 nt can activate the interferon (IFN) response, and should be avoided for therapeutic applications [76, 77].

3.3.3 PIWI-interacting RNAs (piRNAs)

PIWI-interacting RNAs (piRNAs) are a class of animal-specific small single-stranded RNAs that associate with PIWI proteins. Piwi proteins are a clade within the larger family of Argonaute proteins and are mainly expressed in the germline. In general, Argonaute proteins silence their target RNAs through RNA degradation, inhibition of translation or chromatin modification [78]. Biogenesis, as well as function, differentiate piRNAs from miRNAs and siRNAs. PiRNAs originate from long single-stranded RNAs independently from RNase III enzymes, whose role is essential in miRNAs and siRNAs expression. Single-stranded precursors are then processed into discrete piRNAs' units. PiRNAs precursors arise from genomic *loci* known as piRNAs clusters, from which long and single-stranded RNAs are produced. They usually consist of more than 100.000 bases and include transposable DNA elements. The majority of them have an antisense orientation to transposons, thus inducing their silencing. Loss-of-function mutations in piRNAs and PIWI proteins activate transposons that are able to randomly relocate or insert copies within the genome: these events can often impair gonadal development and lead to infertility [79, 80]. Indeed, the ancestral function attributed to piRNAs relates to transposon-induced genomic instability in the germline, where piRNAs repress transposable elements thus preventing transposons mobilization. However, some piRNAs correspond to unique genomic sequences unrelated to transposons. Therefore, there exist evidence suggesting that these piRNAs might regulate expression of host mRNAs [81, 82]. In addition, it has been demonstrated that piRNAs are not only expressed in the germline but, to a lesser extent, can also be detected in somatic tissues where regulate gene expression [83]. Indeed, disruption of piRNAs involved in transposable elements silencing

might results in genomic instability, transposons mobilization and hence, contribute to tumorigenesis. Besides being involved in silencing of transposable elements, piRNAs can also modulate DNA methylation. In this regard, genome-wide methylation profile revealed changes in DNA methylation in Farage and MCF7 cell lines upon transfection with single copy piRNAs. The study showed that genomic regions close to differentially methylated CpG sequences were enriched for sequences recognized by the transfected synthetic piRNA. In conclusion, piRNAs could induce DNA methylation of non-transposable elements *loci* by means of directly binding to genomic DNA or nascent mRNA [84].

3.3.4 Short activating RNAs (saRNAs)

Small double-stranded RNAs (dsRNAs) molecules, such as miRNAs, siRNAs and piRNAs, have been originally identified as the trigger for RNA silencing, where dsRNAs inhibit translation or degrade complementary target mRNA sequences. However, if RNAi is considered a mechanism to regulate gene expression, it is reasonable to suggest that RNA-mediated gene regulation can also positively control target sequences [85]. RNA activation (RNAa) is indeed a mechanism of enhancing gene expression at transcriptional level by means of short dsRNAs and is triggered by both endogenous and artificially designed small RNAs. dsRNAs usually target gene promoter [86, 87], however, there are examples of short activating RNAs (saRNAs) recognizing sequences within coding regions of target genes, e.g. the intronless gene *CEBPA* [88]. Therefore, saRNAs offer a new method to promote gene overexpression and represent a powerful laboratory tool to enhance transcription by a more natural approach. Moreover, RNAa can also be a new option for gene therapies: the ability to specifically upregulate transcription in the absence of exogenous DNA can have profound impacts either in basic research and therapeutics [87]. RNAa has been recently identified in mammals as endogenous mechanism for upregulating transcription. For example, Kuwabara *et al* identified a small double-stranded noncoding RNA able to induce transcription of genes

containing NRSE/RE1 sequences and to promote neuronal differentiation in adult stem cells [89]. Moreover, in the liver, mir-122 facilitates replication of hepatitis C viral RNAs by interacting with 5' noncoding region of the viral genome; hence, it may represent a new target for antiviral therapies [90]. To further study the role of dsRNA in RNAa, synthetic dsRNAs can be designed and tested *in vitro*. Based on empirical observations, a set of rules for saRNAs design has been generated in order to improve the chance to identify functional saRNAs targets, including (i) RNA duplex size of about 19 nt, (ii) 2-nt overhangs on the 3' ends of both RNA strands, (iii) GC content between 40-60 % and (iv) thermodynamic stability in base pairing within the 5' end, whose modification can interfere with Ago2 processing [91]. In regards to target location, saRNAs are usually designed between -100 to -1000 bp relative to the TSS, with the most responsive site usually occurring within -200 and -500 bp region of the promoter. Moreover, avoidance of GC islands, lack of repeat elements (Alu sequences) or inverted sequences are also considered during the design [87]. Due to their nuclear nature, high concentrations of saRNAs are required in order to compensate for the potential nuclear exclusion of duplex RNAs [92]. Therefore, concentrations between 10 to 100 nM can be tested for initial studies to identify optimal conditions for gene activation.

As RNai, RNAa depends on argonaute proteins but with a different kinetics. During RNAa, gene activation is delayed by about 48 hours and, even though the underlying reason is not entirely known yet, it might depend on cell division and accessibility of the nuclear membrane. Moreover, since duplex RNAs are loaded by Ago2 in the cytoplasm before entering the nucleus, it is possible that this process is a passive mechanism that occurs during cell division. Therefore, gene expression analysis should be performed between 48 and 96 hours post transfection in order to validate saRNAs activity. However, saRNA-mediated gene activation is a stable mechanism whose effects can be detected up to 13 days after a single transfection [86].

Thus far, the mechanism responsible for RNAa describes Ago2 protein loading the guide strand (antisense to the sense target gene) of the duplex saRNAs and forming an active saRNA-Ago2 complex. The complex binds to promoter of target genes and induces open chromatin structure and transcription by recruiting CTR9 and RHA proteins. Activation of transcription is also associated with phosphorylation at serine 2 of RNA polymerase II (RNAP II) and monoubiquitination of H2B (H2Bub1) (Figure 7) [93].

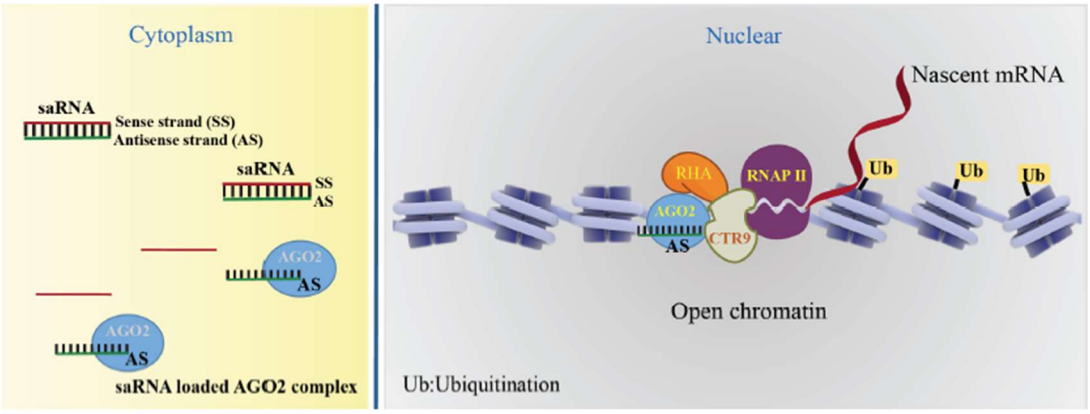

Figure 7. Accepted working model of the molecular mechanisms of saRNA-induced gene activation. The saRNA:AGO2 complex enters the nucleus, binds at promoter of target genes and induces open chromatin structure. AGO2 recruits CTR9, RHA and promotes transcript initiation associating with phosphorylation of RNAP II on Ser2 and H2Bub1 (Yoon and Rossi, 2018) [93].

Even though there is a sound understanding of the saRNAs mechanism, there exist some unresolved areas regarding RNAa effect on DNA methylation and whether both saRNAs strands might function in transcriptional activation.

4 CEBPA

The CCAAT/enhancer-binding protein alfa (CEBPA) is a decisive transcription factor that regulates tissue-specific gene expression by coupling lineage commitment to terminal differentiation and cell cycle arrest [94]. CEBPA is a member of the bZIP family and consists of N-terminal transactivating domains, a basic region necessary for specific DNA sequence binding, and a leucine-zipper region necessary for dimerization at the C-terminal end [95]. Moreover, evidences demonstrated that it acts as tumor suppressor, supporting the

general view that disruption of normal differentiation and its uncoupling with cell cycle arrest are key features for tumor development [96]. The intronless *CEBPA* gene localizes on chromosome 19q13.11 and its mRNA is transcribed into different protein isoforms generated from two different but consecutive in-frame AUGs within the *CEBPA* mRNA. The shorter isoform, namely p30, lacks the amino-terminal 117 amino acids whereas retains the same carboxyl terminus as the full-length form, known as p42. Differences in the N-terminal sequences confer distinct functions to the two isoform in the regulation of differentiation and proliferation. The p30 isoform maintains the DNA-binding domain but miss the N-terminal transactivation domain and is a dominant-negative regulator of the full-length p42 isoform. Since p30 fails to induce differentiation and promote proliferation of myeloid progenitors, the ratio p30/p40 is critical for granulopoiesis [97]. An additional CEBPA protein isoform, named extended-CEBPA, have been identified. This isoform, translated from an alternative non-AUG initiation codon and having an extended N-terminal sequence, occupies the ribosomal DNA promoter in the nucleoli and stimulates rRNA synthesis [98].

4.1 CEBPA in normal hematopoiesis

Fine mechanisms of differentiation regulate formation of terminally differentiated hematopoietic cells from hematopoietic stem cells (HSCs). HSCs maintain the ability to self-renew throughout the entire life of an organism and also to generate all types of mature blood cells [99, 100]. The interplay between tissue-specific transcription factors is a decisive element driving the expression of genes defining a precise cell type. As examples, CEBPA synergistically cooperates with PU.1, c-Myb, and RUNX1 to regulate transcription of genes such as myeloperoxidase, neutrophil elastase, lysozyme, lactoferrin, granulocyte colony

stimulating factor receptor (G-CSFr), macrophage colony stimulating factor receptor (M-CSFr), and granulocyte-macrophage colony stimulating factor receptor (GM-CSFr) (Figure 8).

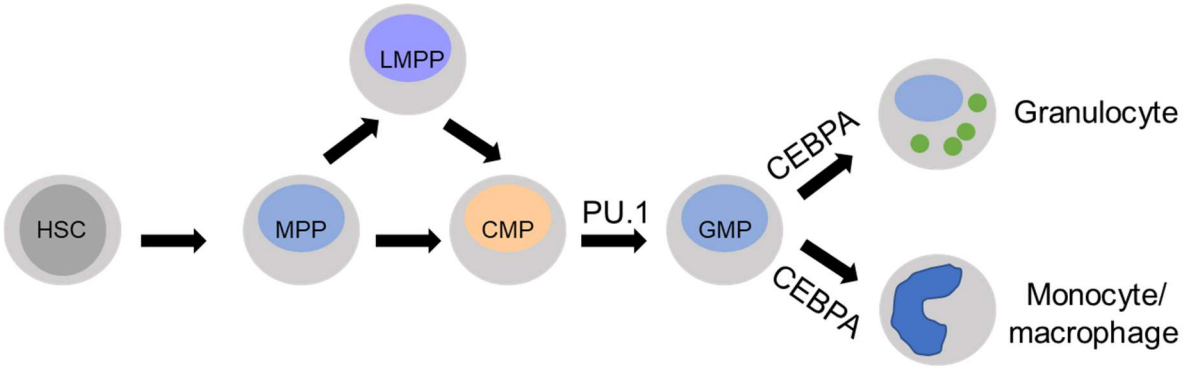

Figure 8. Development of granulocytes and monocytes/macrophages cells from hematopoietic stem cell (HSC). CEBPA drives differentiation towards granulocytic and mononcytic/macrophagic population. MPP, multipotent progenitor; LMPP, lymphoid-primed multipotent progenitor; CMP, common myeloid progenitor; GMP, granulocyte–macrophage progenitor (adapted from Imperato *et al*, 2015) [101].

CEBPA is largely expressed in several cell lineages, including adipocytes, hepatocytes, and type II pneumocytes. Within hematopoiesis, CEBPA is primarily expressed in granulocyte, monocyte, and eosinophil as compared to the lymphoid or megakaryocyte/erythroid lineages [102]. Reducing CEBPA expression promotes monopoiesis by forming heterodimers with AP-1 proteins, as c-Jun or c-Fos. On the contrary, formation of active CEBPA homodimers is essential for granulopoiesis. Indeed CEBPA induces transcription of various proteins indispensable for subsequent lineage maturation, as CEBPE, Gfi-1 and KLF5 [103]. Moreover, CEBPA induces mir-223 transcription that leads to *NFI-A* mRNA degradation thus enhancing granulopoiesis. CEBPA-deficient mice have normal numbers of common myeloid progenitors (CMPs) but lack granulocyte-monocyte progenitors (GMPs) and all subsequent granulocytic stages [104]. Conditional CEBPA knockout in GMPs allows for normal granulopoiesis *in vitro*, indicating that CEBPA is not necessary for differentiation towards granulocytes beyond the GMP stage [96].

In hematopoietic malignancies *CEBPA* expression can be altered by promoter

hypermethylation. Methylation of the core promoter, located -600 bp from the ATG, was found to be an infrequent event in AML [105], whereas methylation of the distal promoter (-1600 to -600 from ATG), also known as upstream promoter, occurs with higher frequency in AML patients [106, 107]. Surprisingly, Hackanson *et al* demonstrated that, while treatment with DAC of AML cell lines induced upregulation of *CEBPA* mRNA *in vitro,* the CEBPA protein levels decreased. Further investigation proved that mir124a, often silenced by epigenetic mechanism and targeting *CEBPA* transcript, was upregulated upon epigenetic treatment and responsible for *CEBPA* degradation [106]

4.2 CEBPA in lung development

CEBPA is expressed in a number of epithelial tissues, including the respiratory epithelium, breast, colon, and prostate [108]. In the respiratory epithelium, CEBPA is involved in lung development and airway epithelial cells differentiation. CEBPA, indeed, regulates the expression of several genes during lung differentiation, as surfactant B and uteroglobin [109]. Since CEBPA abnormalities were originally observed in CEBPA (-/-) knockout mice, where aberrant proliferation of type II pneumocytes was detected, it has been demonstrated that CEBPA plays also a critical role in lung cancer development. CEBPA has growth inhibition properties, therefore its downregulation in lung cancer leads to uncontrolled cell proliferation, whereas its induction promotes growth arrest, morphological changes characteristic of differentiation, and apoptosis [110]. Reduced expression of CEBPA in lung cancer might occur because of promoter hypermethylation. As in AML, methylation of the upstream promoter is a critical event regulating the *CEBPA* gene in lung cancer [111].

5 P15

5.1 P15: a cell-cycle regulator

P15^{INK4b}, also known as *P15*, *CDKN2B* or *MTS2*, and its functional homologue *P16*^{INK4a} (*P16* and *CDKN2A*) are genes located in close proximity within a small 25 kb fragment on chromosome 9p21 and they are both transcribed in the centromere to telomere direction. The *P16* locus, in addition to *P16*, encodes for the alternative transcript *P14ARF* (*p19ARF* in mice). *P16* and *P14ARF* share common exons 2 and 3, but have different first exons, 1a for *P16* and 1b for *P14* [112]. The three genes are tandemly linked to each other within 42 Kb of genomic locus on the *9p21* chromosome. The *9p21* locus also contains *the Antisense RNA in the INK4 Locus* (*ANRIL*), a long-noncoding antisense transcript that overlaps the entire *P15* gene body and shares a bidirectional promoter with *P16* (Figure 9) [113, 114].

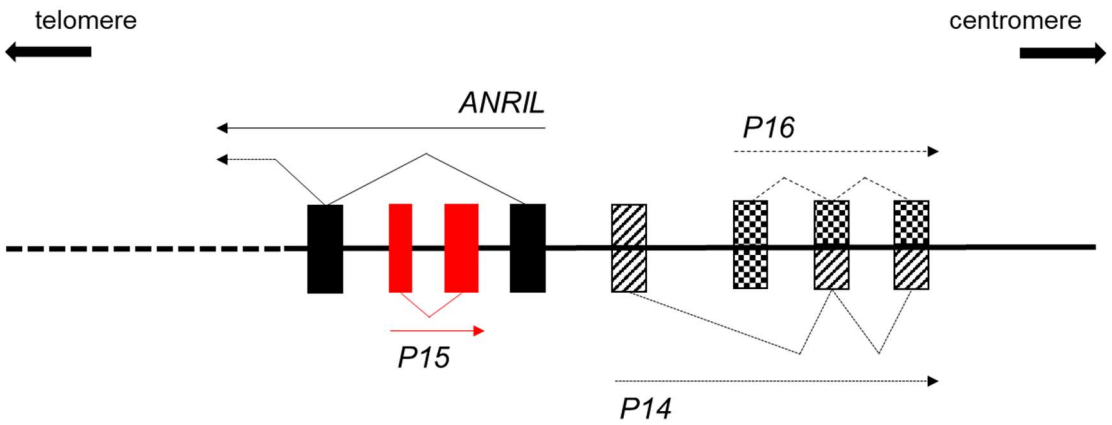

Figure 9. Schematic of the *9p21 locus*. Red rectangles indicate *P15*, striped-pattern rectangles indicate *P14* and squared-pattern rectangles indicate *P16* transcripts. *P15*, *P14* and *P16* mRNAs are transcribed in the direction telomere to centromere. Black rectangles depict *ANRIL* transcript, whose transcription follows centromere to telomere direction.

P15 and *P16* encode for cyclin-dependent kinase inhibitors (CDKIs) that negatively regulate the cell cycle by inhibition of cyclin-dependent kinases (CDKs) 4 and 6, which control G1 to S phase progression through regulation of the Rb pathway [115, 116]. Once the cell pass the G1/S phase checkpoint, it is irreversibly committed to the next cell division. CDKs are a family of multifunctional enzymes able to phosphorylate target proteins involved in the cell cycle and

require the presence of cyclins to be activated. There exist multiple cyclins acting during specific stage of the cell cycle; in this respect, CDK4 and 6 only partner D-type cyclins which regulate the early G1 phase. P15 and P16, acting as negative regulators of cell-cycle progression, directly bind CDKs thus inducing allosteric changes that propagate to the cyclin binding site. Since interaction with cyclins is essential for CDK's activation, conformational changes within cyclins' binding site prevent cyclins from binding hence causing CDK's inhibition [117]. *P15* and *P16* genes share 90% homology at coding exon 2, and isoform 1 of either proteins show 83 % identity. Therefore, the high degree of structural identity explains the high degree of functional homology. However, despite their similarity, in a variety of tumors, genetic alterations of sole *P16* expression due to deletions and/or mutations are observed. On the contrary, deletions of *P15* in lymphomas, carcinomas and sarcomas, occurred in conjunction with *P16* and *P14ARF loci* [118].

5.2 P15: a tumor suppressor gene

P15 is important during differentiation of early hematopoietic progenitors: it indeed drives normal CD34+ towards granulocytic and megakaryocytic commitment. P15 mediates antimitotic signals and it is the effector of transforming growth factor β1 (TGF-β1), which induces arrest at G1 phase [119]. Aberrant *P15* methylation is a major gene silencing mechanism in hematological malignancies, whereas *P16* and *P14* methylation more often occurs in solid tumors as well as in leukemias and lymphomas [120]. *P15* promoter hypermethylation occurs at high frequency in acute myeloid leukemia (AML) and myelodysplastic syndrome (MDS), 80 and 50 % respectively [121, 122]. Aberrant methylation of the *P15* locus is associated with poor prognosis in AML and with increased risk of progression to AML in MDS patients [118]. *P15* aberrant methylation has also been reported in up to 60 % of chronic myelomonocytic leukemia (CMML) cases and is associated with a higher degree of disease's severity [123]. Due to its important role in hematopoietic

differentiation, P15 has been suggested to be a tumor suppressor. To assess whether inactivation of *P15* is a critical event for premalignant myeloid disorder and development of leukemia, a *p15* (-/-) mouse model, in which the second coding exon of the gene was eliminated by homologous recombination, was developed [124]. The *p15* knockout mouse did not show a correlation between loss of p15 and AML outcome, even though the *p15 (-/-)* mice have a competitive advantage over wild-type cells in myeloid cell formation, therefore suggesting an important role in common myeloid progenitors' (CMPs) differentiation [125]. A more informative mouse model generated to define the role of p15 as tumor suppressor is a conditional one, the p15Ink4b$^{fl/fl}$LysMcre mouse, where *P15* was silenced specifically in myeloid cells. Even though the p15Ink4b$^{fl/fl}$LysMcre mice did not develop acute myeloid leukemia, inactivation of the *p15* locus increased susceptibility when additional oncogenic hits were provided. Moreover, the mice developed non-reactive monocytosis in the peripheral blood and expansion of myeloid blast progenitors in the bone marrow (BM), either symptoms resembling CMML in human patients [123]. These data demonstrate that loss of *P15* contribute to preleukemic conditions and can be defined as tumor suppressor gene for AML [118, 126]. As in AML, P15 is not expressed in acute promyelocytic leukemias (APL), hence favoring blasts expansion, but it is progressively upregulated by all-trans retinoic acid (ATRA) treatment, able to drive differentiation [121]. Restoring normal methylation is therefore a rational therapeutic approach when aberrant methylation occurs.

5-azaytidine (5AC, trade name; Vidaza®, Azadine) and 5-aza-2'-deoxycytidine (5-aza-CdR; DAC; also known as Decitabine, trade name; Dacogen®) are nucleoside analogs of cytidine used as demethylating agents (Figure 10) [127].

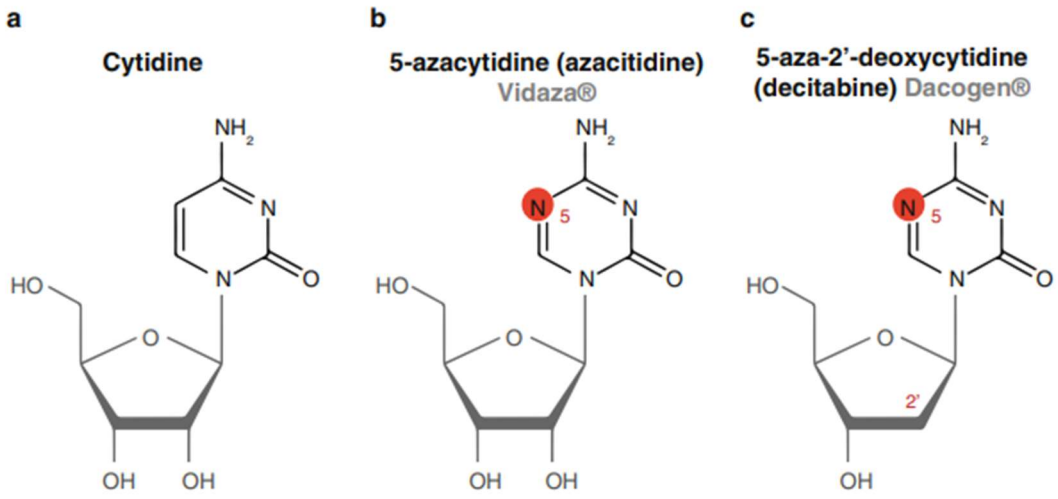

Figure 10. Chemical structures of cytidine nucleoside (a) and azanucleoside (b, c). Sugar moieties are indicated in grey whereas chemical changes between cytidine nucleoside and azanucleosides are highlighted in red (Diesch *et al*, 2016) [128]

Later, they both were approved for the treatment of MDS and AML. 5AC and DAC act differently within the cell: 5AC is mainly incorporated into the RNA (80-90%) and a small percentage into DNA; DAC, instead, incorporate into DNA only. 5AC and DAC are cell-cycle dependent drugs able to exert their activity during S phase, when irreversibly and covalently trap DNMTs at C-6 position [129]. DNMTs' inhibition has disparate consequences: it can induce mutagenesis within CpG dinucleotides, inhibit methylation during the following DNA replication, and cause cytotoxic effects when additional drug molecules do not bind any further DNMTs. Moreover, 5AC and DAC can trigger genomic instability resulting from demethylation of repetitive sequences. It has long been debated whether 5AC and DAC's anticancer effect depends on their cytotoxicity or on the analog itself once incorporated into nucleic acids [129]. However, it is emerging that low doses of analogs inhibit DNMTs, whereas high doses induce cell-cycle arrest because of cytotoxic effects. Besides their efficacy, epigenetic therapies induce off-target effects that limit their clinical use. Moreover, little

success has been achieved thus far in treatment of solid tumors [130]. Therefore, these evidences suggest the need for more specific therapies to treat hematological malignancies, where aberrant methylation occurs.

Aims of the thesis

CEBPA and *P15* are genes frequently silenced by aberrant promoter DNA methylation in hematological malignancies and/or lung cancer.

These past years, our laboratory has been testing a double stranded RNA platform, the short activating RNAs (saRNAs) to reactivate expression of aberrantly methylated genes. SaRNAs are double stranded RNAs shown to upregulate transcription in a gene-specific manner, yet the mechanism behind this activation remains elusive.

This study is built on the hypothesis that saRNAs may promote gene expression by acting as DiR mimicking molecules and changing DNA promoter methylation.

In collaboration with the U.K. biotech firm MiNA Therapeutics [131, 132], saRNAs targeting *CEBPA* and *P15* gene *loci* were tested.

The CEBPA saRNA AW1-51, developed by Voutila *et al.* [88] and currently under evaluation in patients with advanced liver cancer in a phase 1/2a trial (ClinicalTrials.gov: NCT02716012), was delivered by nucleofection in the AML cell line K562 and by lipid-based transfection in the lung cancer cell line A549, which have low-to-medium *CEBPA* expression and fully- to-hemi methylated promoter profile, respectively.

For the *P15 locus* three saRNAs targeting: the promoter (saP15-PR313); the first exon (saP15-PR11) or the intron (saP15-PR56) were designed. All three saRNAs were delivered by nucleofection into the AML cell lines KG1a and Raji, which display low-to- medium *P15* expression and fully- to-hemi-methylated promoter profile, respectively.

The effects of RNA-mediated activation was assayed by qRT-PCR and strand-specific qRT-PCR analyses. RNA sequencing was performed to account for potential off-target effects and activation of specific pathways correlated to *P15* and *CEBPA* re-expression. DNA methylation changes, upon gene reactivation, were assessed for the respective *loci* by Bisulfite sequencing PCR and comprehensive genome-scale (in progress), to exclude additional off-target effects, by the Infinium EPIC arrays platform which covers 850k CpG sites across the genome.

Material and methods

1. Cell culture

KG1a, K562, Raji, U937 and A549 cell lines were growth in RPMI medium supplemented with 10% fetal bovine serum (FBS). HEK293 cell line was cultured in DMEM High glucose medium supplemented with 10% FBS. All cell lines were purchased from ATCC and grown at 37°C, 5% CO_2 in the absence of antibiotics. For all cell lines, media was replaced every 2/3 days; KG1a, K562, Raji and U937 were passaged 1:3/1:4, whereas A549 and HEK293 were passaged 1:5/1:7 every 48/72 hours.

In contrast to the parental KG1, the KG1a population, a male-derived acute myelogenous leukemia cell line, is mainly composed of undifferentiated promyeloblasts, thus being morphologically, cytochemically and functionally less mature. The KG-1a cell is insensitive to colony stimulating factor (CSF) and does not spontaneously differentiate to granulocyte and macrophage like cells.

K562 is a highly undifferentiated population and has been classified as human erythroleukemia line. K562 blasts are able to spontaneously differentiate into progenitors of the erythrocytic, granulocytic and monocytic series. K562 cells are positive for BCR-ABL fusion gene whose oncoprotein leads to transcriptional inhibition of the granulocyte colony–stimulating factor receptor (G-CSF-R), through down-regulation of CEBPA [133].

U937 was derived from histiocytic lymphoma and can be differentiated into macrophages or myeloid lineages dendritic cells (DC) by supernatants from human mixed lymphocyte cultures.

The Raji cell line was derived from B-lymphocytes of a Burkitt's lymphoma and holds the t(8;14) translocation, which results in the juxtaposition of the *MYC* gene to the *IGH* enhancers, which leads to its activation and oncogenic transformation [134, 135]. KG1a, K562 and Raji are suspension cell lines. A549 is an epithelial-like cell line derived from a lung carcinomatous tissue from a Caucasian male. HEK293 is a human embryonic kidney cell line. Either A549 and HEK293 are adherent cell lines.

2. RNA isolation

Total RNA was isolated by phenol chloroform purification using the commercially available Trizol (Invitrogen) according to manufacturer's instructions. Briefly, cells are harvested by centrifugation and supernatant is carefully removed, leaving just a drop of liquid. Cells are briefly vortexed and Trizol is added to the cells solution. Chloroform is added to Trizol (1:5 v:v, respectively) and tube is inverted several times. Follows centrifugation at 13200 rpm, 15 minutes at 4°C. The suspension is now a lower red phenol-chloroform phase (organic), an interphase and a colorless upper phase (aqueous). RNA is exclusively in the aqueous phase. The aqueous solution is pipetted out and isopropyl alcohol is added (ratio 1:1 v:v). Sample is incubated at -20°C O/N (low amount) or 2 hours (high amount). The sample is centrifuged at 13200 rpm, 4°C and pellet is resuspended in 178 µl of RNase free water. DNase I (Roche) is added to sample in the presence of RNase inhibitors (Promega) and solution is incubated 1 hour at 37°C. EDTA 0.5 M and NaCl 5 M are added to sample at the end of incubation to stop DNase I reaction and precipitate RNA, respectively. Cold phenol solution saturated with 0.1M citrate buffer, pH 4.3 (Sigma) is added to sample and tube is centrifuged at 13200 rpm 15 min at 4°C. The supernatant is moved to a new tube and 2.5 volumes of cold 100% ethanol is added. Sample is incubated at -20°C, O/N (low amount) or 2-3hrs (high amount). Pellet is recovered by centrifugation, 13200 rpm at 4°C for >20 minutes. Supernatant is carefully removed and pellet is resuspended in 30-40 µl of RNase free water.

For one-step quantitative reverse transcription PCR (qRT-PCR) using Taqman probes, 50 ng of total RNA was used per 14 µl reaction (Affymetrix USB; HotStart-IT Probe One-Step qRT-PCR Master Mix Kit). *GAPDH* was used as normalization control (Applied Biosystems). Taqman qRT-PCR conditions were: 50°C for 10 minutes (to generate cDNA), 95°C for 2 minutes (to inactivate the reverse transcriptase and activate the Hot Start DNA polymerase), followed by 40 cycles at 95°C (denaturation) for 15 seconds and 60°C (annealing and

elongation) for 1 minute with fluorescence acquisition during the final step. qRT-PCR was performed using a StepOnePlus Real-Time PCR System (Applied Biosystems) using the standard run protocol. For relative expression by qRT-PCR, target gene amplification was calculated using the formula 2^-ΔΔDDCt as described by Livak *et al* [136].

To accurately assess expression and re-expression of low level transcripts, we verified transcript levels using strand specific qRT-PCR (ssqRT-PCR). For ssqRT-PCR 250 or 1000 ng of total RNA were retrotranscribed using SuperScript IV RT, according to the manufactuer's instructions (Invitrogen). In place of random hexamers, specific primers were used for *P15* (KG1a and Raji) and *ecCEBPA* (K562) detection. cDNA was purified using the High Pure PCR Product Purification Kit (Roche) and 60 ng of purified cDNA was used for 14 µl reaction. For *P15*, iTaq Universal SYBR Green Supermix (BioRad) was used; for *ecCEBPA,* One-Step qRT-PCR Master Mix Kit was used and conditions of reaction were modified (95°C for 2 minutes, followed by 40 cycles at 95°C for 15 seconds and 60°C for 1 minute). SYBR Green conditions were: 95°C for 5 minutes (to activate Taq polymerase), followed by 40 cycles at 95 °C for 15 seconds (denaturation) and 60 °C for 1 minutes (elongation and annealing). To measure number of copy of *P15* transcript per each sample, we used a reference standard with known number of copies. The standard curve was generated by serial dilutions of a reference plasmid, pGEM-T easy vector (Promega), containing the amplified PCR product. SsqRT-PCR was performed using a StepOnePlus Real-Time PCR System (Applied Biosystems) using the standard curve protocol. qRT-PCR primer set for the *CEBPA* mRNA is located in the coding region and after the poly(A) signal for *ecCEBPA* (Fig1f and g). qRT-PCR primer set for *P15* is located across exon-exon junctions, as depicted in Fig3d and 4d.

Primer sequences are as the following:

Human *CEBPA*:

Forward `5'- TCG GTG GAC AAG AAC AG-3';`

Reverse 5′-GCA GGC GGT CAT TG -3′;

TaqMan Probe 5′-ACA AGG CCA AGC AGC GC-3′

Human *ecCEBPA*:

Forward 5′-GGT TGT CTG TGG GCC AGG TCA-3′;

Reverse 5′-AGA GCT CAT GAA AGT CAG GAT TG-3′;

TaqMan Probe 5′-AAT AAT ACA GCA TTT TCC CTG GCG G-3′

Human *P15*:

Forward 5′-CGG GGA CTA GTG GAG AAG G-3′

Reverse 5′-GTG AGA GTG GCA GGG TCT G-3′

P15 specific primer for RT reaction:

Reverse 5′-CCT GTG AAC CTT TAA CAT TTC TCA-3′

ecCEBPA specific primer for RT reaction:

Reverse 5′-GGT AGG GTG TAG CCA CAT GGT CTA-3′

3. DNA isolation

DNA was isolated by phenol:chloroform:isoamyl alcohol (25:24:1) purification. Briefly, cells were resuspended in lysis buffer (0.5% SDS, 25 mM EDTA pH 8, 10 mM TRIS pH 8, 200 mM NaCl) and treated with RNAse A (Roche) for 20 minutes at 37°C. The cell lysate was digested with proteinase K (Roche) at 65˚C overnight. Phenol solution (pH 8) was added to cell lysate (v:v 1:1) and sample centrifuged 15 minutes at 13200 rpm. The aqueous phase was recovered with a cut tip and transferred to a new tube. Isopropyl alcohol 1:1 v:v was added and sample incubated at -20°C for 20 minutes (high amount) or O/N (low amount). Pellet was recovered by centrifugation and resuspended in TE (pH8).

4. Sodium bisulfite conversion

Bisulfite genomic sequencing is defined as the gold-standard for detection of DNA methylation at single base-pair resolution. This method was first developed by Frommer *et al* and it is based

on the finding that the amination reactions of cytosine and 5-methylcytosine (5mC) proceeds with very different consequences after the treatment of sodium bisulfite [137]. Therefore, cytosines will be converted into uracil residues and recognized as thymine in subsequent PCR amplification and sequencing, whereas 5mCs are immune to this conversion and remain as cytosines allowing 5mCs to be distinguished from unmethylated cytosines. A subsequent PCR process is performed to define the methylation status of the *loci* of interest [138]. From 200 to 1000 ng of genomic DNA were bisulfite converted according to manufacturer's instructions (EZ DNA Methylation kit, Zymo Research) and eluted in 20 µl of warm RNase free water. The conversion was performed as follows: 98˚C for 8 minutes, 64˚C for 7.5 hour, 4˚C storage.

5. Combined Bisulfite Restriction Analysis (COBRA)

Combine Bisulfite Restriction Analysis (COBRA) is a qualitative approach to detect the methylation status of CpG-containing regions. Sodium bisulfite treatment introduces methylation-dependent sequence modifications and new restriction enzyme (RE) sites can be generated or pre-existing ones can be lost or retained depending on the methylation status [139]. As an example, the BstUI enzyme recognizes the CGCG sequence, which can be altered upon bisulfite conversion depending on the methylation status of the Cs. This property can be exploited to study DNA methylation in a qualitative way, by digesting the PCR amplicon with different REs and looking at the resulting digestion patterns, that is the basic principle of Combined Bisulfite Restriction Analysis (COBRA).

4 µl of bisulfite converted DNA were used in 50 µl of PCR reaction (FastStart Taq DNA Polymerase, Roche). The PCR reaction was performed as follows: 95°C for 6 minutes, followed by 35 cycles at 95°C for 30 seconds, 53-57°C for 1 minute, 72°C for 1 minute, and a final step at 72°C for 7 minutes. PCR product was gel-purified (QIAEXII Gel Extraction Kit, Qiagen) in a 2% TAE agarose gel. After gel purification, about 400 ng of gel-purified product

were digested at 60°C for 2 hours with 10 units of BstUI enzyme (NEB). Digested and undigested product were run on a 3% TAE agarose gel. Primer sequences are as the following:

CEBPA upstream promoter, -1.4 kb from the TSS:

Forward 5′-GGT GTT TTT AGT TGT GTT TTT TT-3′

Reverse 5′-AAA CCC TAA AAC CCC TTA-3′

CEBPA upstream promoter, -1.1 kb from the TSS:

Forward 5′-TAT TTA AGG GGT TTT AGG-3′

Reverse 5′-AAA AAC AAA CTT AAC TCT AA-3′

CEBPA upstream promoter, -0.8 kb from the TSS:

Forward 5′-TAG TTT YGT TAG TTT GGG GGG TTT-3′

Reverse 5′-AGG TTA AGG YGG TTG TGG GTT TTA-3′

P15 promoter:

Forward 5′-GAT ATT TAG YGA GTA GTG TAG TTA GTA TTT TTG G-3′

Reverse 5′-CCT YGC TCT AAC AAA ATA AAA AAC CAA-3′

6. Bisulfite Sequencing PCR (BSP)

After bisulfite treatment of genomic DNA and PCR amplification of the *CEBPA* upstream promoter or *P15* promoter, the PCR product was gel purified as previously described. The amplicon was cloned into the pGEM-T Easy Vector (Promega) as manufacturer's instructions. *E.coli* competent cells (bacterial strain JM109, Promega) were transformed with 3 µl of ligase and plated on LB agar plates containing ampicillin (Sigma Aldrich), IPTG (Sigma Aldrich) and X-Gal (Sigma Aldrich) for blue/white screening and standard selection. White positive colonies were amplified and extracted using Pure yield plasmid Miniprep system Kit (Promega). Plasmid DNA was sent for sequencing (Quintara Biosciences, Boston, MA) and results analyzed by QUMA software (http://quma.cdb.riken.jp/). Clones with percentage of unconverted CpGs higher than 95 and with percentage of identity lower than 90 were excluded

from the final analysis. For DNA methylation analysis of the *CEBPA* upstream promoter, 42 nt, preceding the reverse primer 5'-AAAAACAAACTTAACTCTAA-3' and containing 6 CpGs dinucleotides, were excluded from the analysis. The 6 CpGs were unmethylated both in the CEBPA-saRNA treated samples and in the NC-treated control. Moreover, the region where the primer was designed is the only one without CpGs, whose presence within primer sequence can affect primers' efficiency.

7. DNA Fluorescence in Situ Hybridization (DNA-FISH)

DNA Fluorescence In Situ Hybridization (DNA-FISH) takes advantage of fluorophore labeled Bacterial Artificial Chromosome (BAC) probes to visualize the presence or absence of genomic DNA fragments on interphase DNA. BAC probes (Empire Genomics) covering selected areas of the *P16/P15* locus were used to assess the presence of the locus. The manufacturer's recommended hybridization protocol was followed with minor modifications. Briefly, cells were fixed with methanol acetic acid (3:1 ratio) for more than 2 hours at -20°C, dropped onto uncharged glass slides, and then left for at least 20 minutes at 45°C. To increase accessibility of the chromatin to the BAC probe, dried slides were then washed for 1 minute in 2X SSC (Cold Spring Harbor Protocols) and digested with 1 μg/ml proteinase K (Roche) in proteinase K buffer (Cold Spring Harbor Protocols) for 3 minutes at room temperature. Slides were then washed twice in 70% ethanol, 85% ethanol, 100% ethanol, and then allowed to dry. Hybridization was carried out with 1 uL of probe and 9 μL of hybridization buffer for at least 16 hours in a humidified Hybrite chamber: 85°C for 3 minutes, followed by 37°C for at least 16 hours. The following day, excess probe was removed in Wash 1 (0.4X SSC/ 0.3% NP-40 warmed to 73°C), 2 minutes; Wash 2 (2X SSC/ 0.1% NP-40 at room temperature), 1 minute; and dried in the dark. Prolong Gold antifade with DAPI (Thermo Fisher Scientific) was used to counterstain and cover the slide. A chromosome 9 control probe (9q21) was used as hybridization control. BAC probe coverage is depicted in Figure 3b.

8. RNA Fluorescence In Situ Hybridization (RNA-FISH)

Single molecule RNA FISH is a methodology for detecting and localizing particular RNA. Stellaris RNA FISH technology is based on the use of fluorescently labeled tiling oligonucleotides complementary to a sequence of interest. Unlike DNA FISH, RNA FISH is strand-specific, enabling the visualization of sense and antisense RNA strands. The use of tiling probes also has the added benefit of increasing the chance of hybridization success within the oligo pool and reducing the effect of nonspecific hybridization as multiple co-localizing probe signals are needed to visualize a single RNA molecule. The *P15* probe set was designed against nucleotides 1-516, 517-1650, 1651-2198 of NM_004936.3 covering the coding sequences only, and detects both *P15* variants, P15 NM_004936.3 and NM_078487.2 (www.biosearchtech.com). Visualization of RNAs was carried out following the Modified Fluorescent In Situ Hybridization protocol (Stellaris, BioSearch Technologies). Briefly, 1x10^6 cells/100 mL were cytospun onto glass slides and fixed using 3:1 methanol-acetic acid for 10 minutes at room temperature. Slides were first rinsed twice with freshly prepared Wash Buffer A. Hybridization was carried out with 250 nM (2 ml of the 1.25 mM stock) of Ready Designed P15 probe set in 10% formamide (Sigma) for 4 hours at 37°C in a humidified chamber. Next, slides were incubated with Wash Buffer A twice, 30 minutes each wash, to remove unbound probe. Hoechst counterstain solution was added at 1:1000 concentration in the second wash. Finally, slides were rinsed and then incubated for 5 minutes at room temperature with Wash Buffer B and mounted with Vectashield H-1000 Mounting Medium (Vectashield). The Ready Designed Xist probe set was used for control of nuclear permeabilization and compartment-specific hybridization. All solutions were supplemented with vanadyl complex RNAse inhibitor (New England Biolabs). RNA FISH images are generated from a composite of z-stack images. Raw images were acquired on a Leica DM 5500B Microscope with a 100W high-pressure mercury lamp. Autofluorescence was checked

for by simultaneous acquisition of images in the opposite filter channel. Images were taken with a Leica DFC 350 FX camera (CCD with a Peltier cooling system). 10 μM of z-stack images were acquired in 0.3 mM steps. Images were assembled and contrast-enhanced using FIJI (FIJI is Just Image J; https://fiji.sc/) as per manufacturer's recommendations.

9. Cell synchronization

During synchronization, cells representing different phases of the cell cycle are brought to the same stage, resulting in a synchronous population. Cell cycle consists of the synthesis of genetic materials (S-phase) and the further distribution of the genetic materials, together with other cellular components, onto daughter cells (M-phase). Progression from S- to M-phase is intermitted by two gap phases, G1 and G2, that establish the cell cycle. An additional phase, named G0, defines quiescent cells that is cells that have stopped dividing and have left the cell cycle [140].

Thymidine is a chemical method used for reversely perturbating cell cycle progression and synchronizing cells in G1/early S-phase [141]. As Xeros *et al* previously demonstrated, excess of thymidine prevents DNA synthesis by inhibiting the formation of deoxycytidine triphosphate (dCTP) from cytidine-5'-phosphate [142]. The reduced size of dCTP pool compared to other deoxynucleosides triphosphates might have a regulatory function in the rate of DNA synthesis [143]. To obtain a population of cells uniformly arrested in G1/early S-phase, cells were treated with various concentrations of thymidine, according to the cell line used. Thymidine (Sigma) was dissolved in cell culture-grade water at 50 mM concentration and aliquots stored at -20°C. KG1a cells were seeded at 1*10^6 cells/ml in complete RPMI medium (Corning) in the presence of 2.5 mM thymidine for 18 hours (first block). Following the first block, thymidine was removed and cells were washed twice with fresh medium and grown in thymidine-free medium for 8 hours to release cells from the block. Eight-hours release was followed by a second thymidine block by the addition of 2.5 mM thymidine for 18 hours.

K562 cells were seeded at 0.5*10^6 cells/ml in complete RPMI medium and grown in the presence of 2.5 mM thymidine for 18 hours. As for KG-1a cells, K562 cell line was washed twice with fresh medium and incubated with thymidine-free medium for 8 hours. A second 18-hours thymidine block (2.5 mM) followed the 8-hours release.

Raji cells were seeded at 0.75*10^6 cells/ml and incubated with 3 mM thymidine for 18 hours. As for KG-1a and K562 cell lines, the first block was followed by 8 hours incubation in fresh medium to release cells from the block and by a second thymidine-block for the same duration as the first block.

A549 cell line was seeded at 0.07*10^6 cells/ml 24 hours before incubation with thymidine. After 24 hours cultivation in fresh medium, cells were incubated with 3.5 mM thymidine for 11 hours. A one-thymidine block was sufficient for synchronizing the cells. After thymidine incubation, cell cycle was assessed by flow cytometry prior transfection.

10. Cell cycle analysis

The simultaneous detection of RNA and DNA molecules can be determined by the means of Pyronin Y staining dye (PY) and double-stranded DNA staining, respectively. PY is indeed able to distinguish quiescent cells (cells in G0 phase) from active cycling cells through their RNA content: cells in G0 phase have a lower content of RNA, thus resulting in a lower PY signal. On the contrary, a double-stranded DNA staining, as Vybrant DyeCycle Violet dye is (VV, Invitrogen), allows to reconstruct cells' cycle according to their amount of DNA. Cell-cycle analysis on thymidine and non-thymidine treated cells was performed as follows. Cells were harvested and washed once with phosphate buffer saline (PBS). Cell pellet was resuspended in phosphate-citrate buffer solution (see recipe below) and incubated 20 minutes at room temperature (RT). Four ml PBS was added to the cell solution followed by a second wash with 1 ml PBS. Cell pellet was washed once with 1 ml PY (1.5 mg/ml in PBS, Sigma) and resuspended with 400 µl of staining solution (PY 1.5 mg/ml and 5 µM VV). PY is able to

stain RNA molecules whereas VV is a cell permeable double-stranded DNA staining. Samples were incubated 15/30 minutes at 37 °C in the dark prior analysis by flow cytometry. CytoFLEX LX (Beckman Coulter) instrument was used for detecting PY and VV signal. PY was detected by the 488 nm laser (Blue) and B610-ECD filter; VV signal was detected by the 405 nm laser (Violet) and V450-PB filter.

For 100 ml of phosphate-citrate buffer solution:

Prepare phosphate buffer stock solution by dissolving 8.903 g of Na_2HPO_4 in 250 ml of H_2O (**A**). Prepare the citric acid buffer stock solution by dissolving 2.11 g of citric acid in 100 ml of H_2O (**B**). To 35 ml of **A** add **B** until pH 4.8; then add 0.87 g NaCl, 0.5 g bovine serum albumin (BSA), 0.02 g saponin and 1 ml 500 mM sodium EDTA. Adjust volume to 100 ml with H_2O.

11. Short-activating RNAs (saRNAs)

The short-activating RNAs (saRNAs) targeting the *P15* and the *CEBPA* locus were designed by MiNa Therapeutics and purchased from Sigma and IDT. saRNAs were reconstituted at 100 μM stock with cell culture-grade water and diluted to ready-to-use concentration of 50 μM. Either 100 μM stock and diluted aliquots were kept at -20 °C. The saRNA's sequences were selected among bioinformatically identified hotspots at the gene of interest. Hits against the *P15* gene include: saP15-PR11 (targeting the first exon), saP15-PR56 (targeting the intron) and saP15-PR313 (targeting the promoter). In addition, a scramble control against the firefly luciferase, named FLUC, were designed as negative control. Since the P15-saRNAs had never been previously tested, in order to assess their transfection efficiency as well as their localization within cellular compartments, the 3' end of either strands was labeled with different fluorophores to distinguish sense and antisense strand. The guide, or sense, strand was labeled with Alexa Fluor 488 dye, whereas the passenger, or antisense, strand was labeled with ATTO590 dye. Transfection efficiency and cellular localization were assessed by flow

cytometry and fluorescent microscopy, respectively. The CEBPA-saRNA, the AW1-51 saRNA targeting the coding region of the gene, was chosen according to published data [88, 144]. In parallel to AW1-51 saRNA, a negative control, NC, was used in each transfection.

12. saRNAs delivery

12.1 Nucleofection

Nucleofection is an electroporation-based methods that, by combining cell-specific nucleofector solution and electrical pulses, enables DNA and/or RNA molecules to directly access the nucleus. Therefore, in contrast with other transfection methods, nucleofection does not depend on cell division.

KG1a cells were transfected with Amaxa Nucleofector Device II, Kit L, program V-001, according to manufacturer's instructions (Lonza). More specifically, 1*10^6 cells were nucleofected in the presence of 4.5 μl saRNA (50 μM stock). After nucleofection, cells were diluted into 4.5 ml fresh medium and distributed into 3 wells in a 12-well plate, 1.5 ml per well. Two μg of a pmax green fluorescent protein plasmid (pmaxGFP, provided by the manufacturer) was used as transfection-efficiency control and detected either by fluorescent microscopy or flow cytometry. Raji and K562 cells were also transfected with Amaxa Nucleofector II, Kit V, program M-013 and T-016, respectively, according to manufacturer's instructions and the protocol previously described for KG1a cell line. For all cell lines, 1*10^6 cells per cuvette were used. Seventy-two hours post nucleofection, cells were re-transfected with saRNAs with lipofectamine2000 (L2000, Invitrogen), a cationic liposome based reagent. 0.33*10^6 cells were plated in 1.3 ml fresh medium (12-well plate). The complex lipofectamine-RNA was prepared as follows:

1) 4 μl L2000 were added to 100 μl optimem (Gibco) and the complex incubated 5 minutes at RT (mix-**A**);

2) 1.5 µl saRNA (50 µM stock) was added to 100 µl optimem and the mix incubated 5 minutes at RT (mix-**B**);

3) mix-**A** was added to mix-**B** and incubated 20 minutes at RT (mix-**C**);

4) mix-**C** (200 µl) was distributed to each well (1.3 ml cell suspension) dropwise;

The ratio saRNA:L2000 was the following: per 1.5 µl saRNA (50 µM stock), 4 µl L2000 were used. In addition to saRNAs, pmaxGFP (2 µg) was used to monitor the transfection efficiency. Cells were collected 24 hours post lipofectamine transfection.

12.2 Lipid-based transfection

The CEBPA-saRNA was delivered into A549 cells by DharmaFECT 1 (DF1, Horizon Discovery), a lipid-based transfection reagent. Two consecutive transfections were performed: 1) a reverse transfection followed by a 2) forward transfection. The first reverse transfection was prepared as follows: 10 µl DF1 were added to 490 µl optimem; 1 µl saRNA (50 µM stock) was added to the mix. The complex DF1-optimem-saRNA was transferred into wells (12-well plate) and incubated 30 minutes at RT. Meanwhile, synchronized A549 were diluted to 0.13*10^6 cells/0.5 ml and 0.5 ml was added, per well, to the complex DF1-optimem-saRNA at the end of incubation. Twenty-four hours post transfection, a second transfection (forward) was performed as follows:

1) 1 ul saRNA (50 µM stock) was added to 99 µl optimem and the mix incubated 5 minutes at RT (mix-**A**);

2) 5 µl DF1 were added to 95 µl optimem and the complex incubated 5 minutes at RT (mix-**B**);

3) mix-**A** was added to mix-**B** and incubated 20 minutes at RT (mix-**C**);

4) Medium was removed from wells and replaced with 800 µl of fresh medium;

5) mix-**C** was distributed to wells dropwise;

Medium was changed 48 hours after forward transfection.

13. saRNAs transfection efficiency

Labelled P15-saRNAs were transfected into KG1a and Raji cell lines and fluorescence intensity was determined by flow cytometry. Twenty-four, forty-eight and seventy-two hours post nucleofection, samples were collected and washed once with PBS. Finally, cell pellet was resuspended in 400 µl PBS and DAPI was added to each sample in order to exclude dead cells during analysis. CytoFLEX LX was used for assessing fluorescence intensity. Alexa Fluor 488 dye (green) was detected by the 488 nm laser (Blue) and B-525 filter; ATTO590 dye (red) intensity was assessed by the 561 nm laser (Yellow Green) and Y-610 filter. Dapi was measured by the 375 nm laser (Near UV) and 450 filter. Forward and side scatter, FSC and SSC respectively, were used to identify the cell population and exclude doublets. Dapi positive cells, whose cut-off was first defined in the non-transfected control, were excluded from the analysis. By plotting B-525 (green) against Y-610 (red) and tracking a quadrant gate, 4 sub-populations were identified: a) green negative and red negative, b) green positive and red negative, c) green positive and red positive and d) green negative and red positive.

14. Fluorescent microscopy

P15-saRNAs localization was assessed by fluorescent microscopy in KG1a and Raji cell lines. Images were acquired with LSM 880 Inverted Live-cell Laser Scanning confocal microscope (Zeiss) and either live and fixed cells were analyzed. Samples were collected 30 minutes, 1 hour, 1.5 hour, 24 hours and 48 hours post transfection. For live-cell imaging, about 0.05*10^6 cells were re-suspended in 150 µl fresh and complete RPMI and transferred into a multi-chamber. Hoechst 33342 was added to cells 10 minutes before acquisition. Whilst live-cell imaging did not require any processing of the sample, confocal microscopy imaging did. Hence, about 0.06*10^6 cells were collected, washed twice with cold PBS and re-suspended in 200 µl PBS. Samples were cyto-spun at 400 rpm for 5 minutes onto microscope slides and incubated 20 minutes at -20 °C in 100% methanol. Slides were washed twice with PBS, 5

minutes each wash, and incubated 5 minute at RT with DAPI to visualize cell nuclei. Following DAPI incubation, a PBS wash and a final rinse with water were performed. Excess of liquid was removed from the sample by gently tapping the edge of the slide on a wipe. A single drop of mounting medium (ProLong Glass Antifade Mountant, Invitrogen) was applied to each cell spot and a coverslip was lowered onto the mountant avoiding trapping any air bubbles. Images were acquired after curing of the mountant.

15. Western blotting

Cultured cells were washed with PBS and lysed with reduced 2x sample buffer (0.1 M TRIS pH 6.8, 4% SDS, 5% glycerol, 0.2 M DTT and 4% β-mercaptoethanol), 1:1 ratio PBS: reduced 2x sample buffer. Samples were boiled 5 minutes at 95°C. Equivalent amounts of protein were resolved on a NuPage 4-12% bis-tris gel (Invitrogen) with MES buffer (Invitrogen) at 130 V, and transferred to polyvinylidene difluoride (PVDF, 0.45 μM) membrane at constant voltage (12 V) on ice. The resulting blot was blocked 1 hour at RT with 5% nonfat dry milk in 0.1% TBS-T. CEBPA protein was detected by overnight incubation at 4°C with the appropriate antibody (1:500, 5% nonfat dry milk in 0.1% TBS-T) (Abcam) followed by enhance chemiluminescence detection (Thermo Scientific). CEBPA protein signal was normalized to Beta-Actin protein (Sigma).

Results

CEBPA-saRNA AW1-51 induced re-expression of *CEBPA* and *ecCEBPA* mRNA in myeloid and lung cancer cell lines

Previous evidence revealed the efficacy of CEBPA-saRNAs for treatment of hepatocellular carcinoma [88, 144]. Voutila *et al*, by means of bioinformatically directed nucleotide walk around the *CEBPA locus*, identified the CEBPA-saRNA AW1-51 as the one able to upregulate *CEBPA* mRNA in HepG2 cells and in patients with advanced liver cancer in a phase 1/2a trial (ClinicalTrials.gov: NCT02716012) [88, 144] (Fig.1a). To investigate whether CEBPA-saRNA could also upregulate expression of *CEBPA* in myeloid and lung cancer cells, we tested CEBPA-saRNA AW1-51 in two cell lines, A549 and K562, expressing *CEBPA* mRNA at low and low-to-undetectable levels, respectively [111, 145, 146]. Prior to treatment with 50 nM CEBPA-saRNA, cells were synchronized in G1/early S phase, when the synthesis phase of the cell cycle starts and DNMT1's activity peaks [54, 147] (Fig.1b and d). Synchronization, indeed, allowed to enhance potential direct and/or indirect interactions between CEBPA-saRNA and DNMT1 at S phase [54]. The CEBPA-saRNA reduced proliferation in both cell lines (Fig.1c and e), data consistent with an antiproliferative and differentiating role of CEBPA in myeloid lineages and lung cells [108, 110, 111]. Twenty-four hours post transfection, cell-growth arrest was detected in NC-treated samples as well, indicative of an initial cellular toxicity of the transfection reagent, which was more significant upon lipid-based transfection method (Fig1.c). Commonly, the extent of cellular toxicity caused by different transfection techniques is influenced by the reagent and also by the nature of the cells, as we observed.

Data from the published literature show that gene activation by means of saRNAs is delayed by about 48 hours. Since saRNAs exert their activity in the nucleus, the effect on gene activation depends on cell division and accessibility of the nuclear membrane, resulting in a cell-cycle dependent mechanism. Moreover, as duplex RNAs are loaded by Ago2 in the cytoplasm before entering the nucleus, it is possible that this process is a passive mechanism

that occurs during mitosis [86]. Therefore, gene expression analysis should be performed between 48 and 96 hours post transfection in order to validate saRNAs activity.

Standard qRT-PCR analysis demonstrated upregulation of *CEBPA* in A549 and K562 since 48 hours post transfection, with peak of expression (up to 4- and 2-fold in A549 and K562, respectively) 72 hours post treatment (Fig.2a and b, left panel).

Gene expression can be affected by noncoding RNAs arising from the promoter and downstream region of the corresponding gene [47, 148, 149]. The *extra-coding CEBPA* (*ecCEBPA*) is an overlapping, sense-orientated noncoding transcript arising from the *CEBPA locus,* whose transcription controls the expression of the corresponding coding mRNA *CEBPA*. e*cCEBPA* is able to interact with DNMT1 and prevent DNA methylation of the *CEBPA* gene [54]. To assess whether CEBPA-saRNA promotes transcription of noncoding transcripts, we evaluated expression of *ecCEBPA* upon treatment. While in A549 the highest *ecCEBPA* re-expression was detected at 72 hours, in K562 peak of *ecCEBPA* activation was observed 48 hours post transfection.

The delayed re-expression of *ecCEBPA* in A549 might account for a different detection method. While in A549 the expression of *ecCEBPA* was evaluated by standard qRT-PCR, in K562 it was necessary to perform strand specific qRT-PCR (ssqRT-PCR) as a result of low basal levels of expression (Fig.2a and b, right panel). Therefore, since ssqRT-PCR is a more sensitive methods compared to standard qRT-PCR, the different approach might explain the delay in the re-expression of *ecCEBPA* in A549 cell line.

We also evaluated CEBPA protein expression in A549 upon treatment with saRNA (Fig.2c). CEBPA expression increased 48 and 72 hours post transfection in saRNA-treated samples compared to the negative control (Fig.2c, upper panel). The expression of CEBPA was normalized to Beta Actin and an increment of 2 fold was observed 72 hours post treatment (Fig.2c, bottom panel). On the contrary, in K562 cells, CEBPA protein expression, either

before and post treatment, was undetectable (data not shown). K562 cell line expresses the oncogenic fusion protein BCR-ABL, which downregulates CEBPA protein by inducing hnRNP E2, responsible for inhibiting *CEBPA* mRNA translation [133].

Re-expression of CEBPA protein in A549 led to morphological changes, as axonal growth (Fig.2d, left panel, red arrows), 120 hours post transfection. This result confirmed previous data showing that ectopic expression of CEBPA induced differentiation in lung cancer cell lines, as Calu-I, squamous cancer cells, and H358, adenocarcinoma bronchoalveolar type cells [110].

Collectively, these data demonstrate that CEBPA-saRNA upregulates *CEBPA* and *ecCEBPA* transcripts in synchronized myeloid and lung cancer cells and re-expression of CEBPA protein induces morphological changes, suggestive of differentiation.

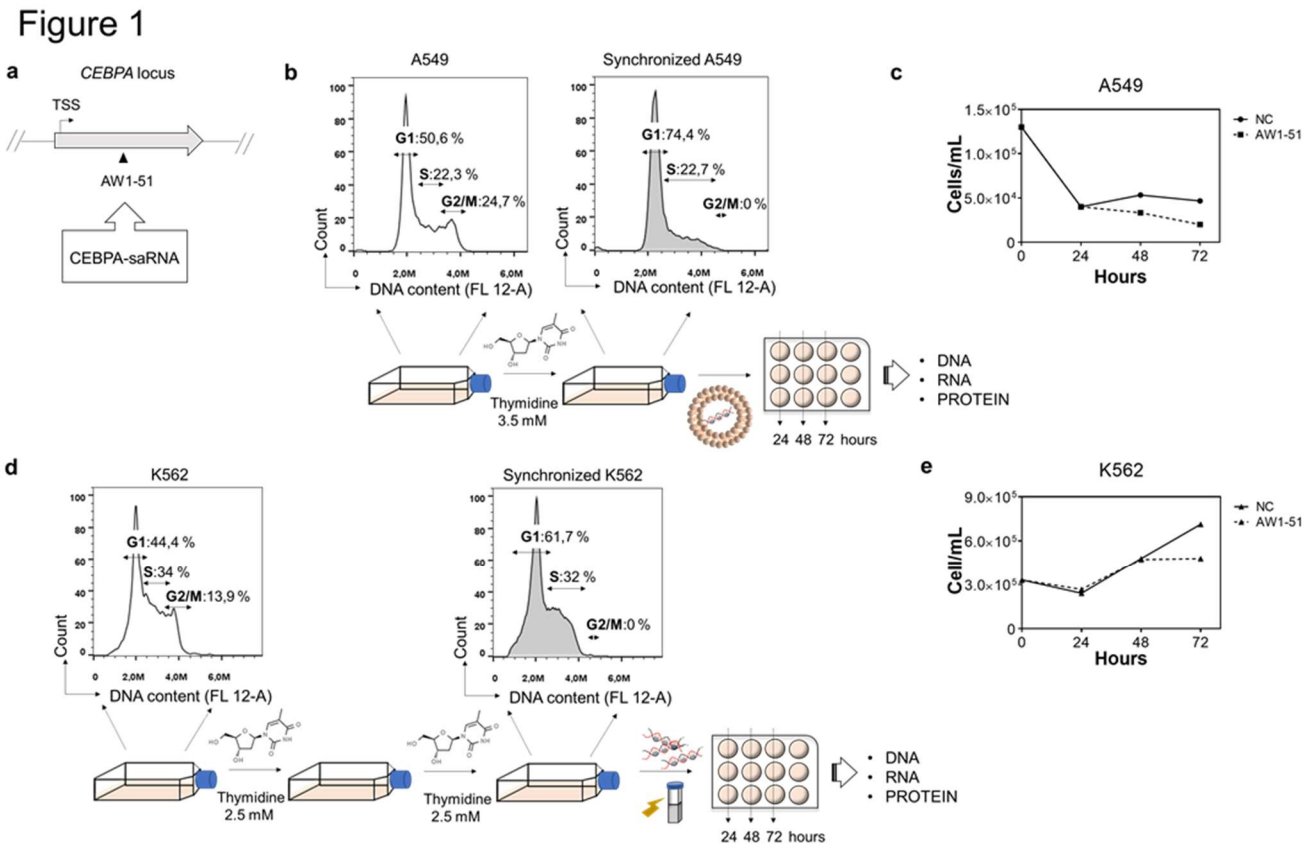

Figure 1 Schematic of CEBPA-saRNA delivery in A549 and K562 cell lines. **a**, Schematic of CEBPA-saRNA (AW1-51) targeting the *CEBPA* coding region. **b**, Schematic of saRNA's delivery to A549 cells. Top panel, A549 cells are seeded 24 hours prior synchronization and then synchronized in

G1 phase by one-thymidine block (11 hours, 3.5 mM). Bottom panel, synchronized cells are reverse-transfected with 50 nM CEBPA-saRNA or 50 nM NC and, after 24 hours, forward-transfected with CEBPA-saRNA or NC at 50 nM final concentration. DNA, RNA and proteins are collected from 24 to 72 hours post forward-transfection. **c**, A549 cell-growth curve upon 50 nM CEBPA-saRNA treatment, from 0 to 72 hours post forward-transfection. **d**, Schematic of saRNA's delivery to K562 cells. K562 cells are synchronized in G1 phase by double-thymidine block (2.5 mM); after synchronization cells are electroporated by Amaxa nucleofection with 50 nM CEBPA-saRNA or NC. DNA, RNA and proteins are collected from 48 to 72 hours post nucleofection. **e**, K562 cell growth curve upon 50 nM CEBPA-saRNA treatment, from 0 to 72 hours post nucleofection.

Figure 2

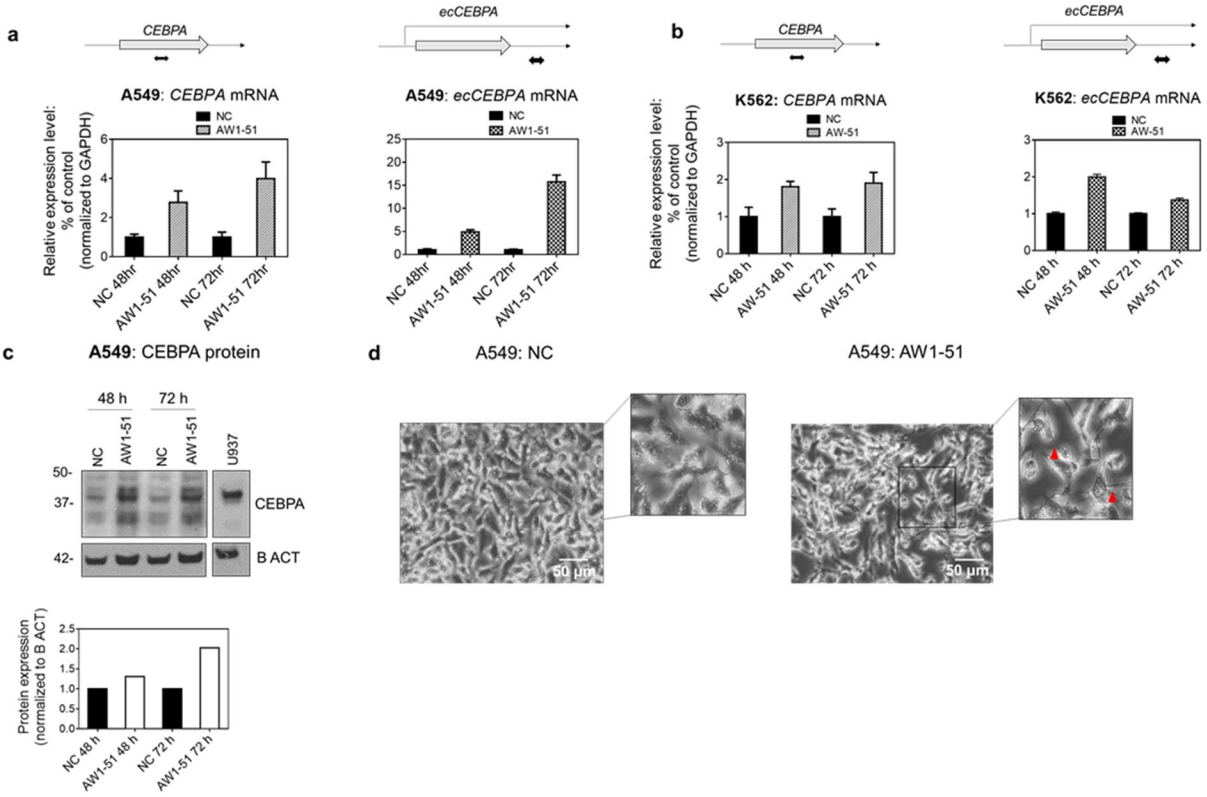

Figure 2 CEBPA-saRNA reactivates transcription of *CEBPA* mRNA in A549 and K562 cell lines. **a**, Top panel, diagram of *CEBPA locus* and corresponding qRT.PCR primers (black arrows). *CEBPA* (bottom left panel) and *ecCEBPA* (bottom right panel) re-expression upon 50 nM CEBPA-saRNA (AW1-51) treatment, from 48 to 72 hours post transfection, in A549 cells (n=2). **b**, Top panel, diagram of *CEBPA locus* and corresponding qRT.PCR primers (black arrows). *CEBPA* (bottom left panel) and *ecCEBPA* (bottom right panel) re-expression upon 50 nM CEBPA-saRNA, from 24 to 72 hours post transfection, in K562 cells (n=2). **c**, Upper panel indicates CEBPA protein re-expression upon CEBPA-saRNA, at 48 and 72 hours, in A549 cells; bottom panel shows densitometric analysis of CEBPA protein relative to B ACT signal. **d**, A549 morphological changes upon CEBPA-saRNA treatment 120 hours post nucleofection. 20x magnification. NC: negative control; *ecCEBPA*: extra-coding CEBPA; B ACT: Beta Actin.

The CEBPA-saRNA induced, to different extent, demethylation of the *CEBPA* upstream promoter in A549 and K562 cell lines

DNA methylation is a key epigenetic signature that is not randomly distributed across the genome, but exhibits regional specificity. Evidences demonstrated that methylation of promoter associated CGIs negatively correlates with gene expression, and this mechanism often occurs in cancer [8, 9].

CEBPA expression can be altered by promoter hypermethylation, as in AML and lung cancer. A 2883 bp long CGI, with 281 CpG dinucleotides and coordinates chr19:33792062-33794944 (hg19), extends throughout the promoter and part of the coding region of the *CEBPA* gene. In the literature, the mentioned CGI was arbitrarily divided into two sequences, the upstream or distal promoter (-1.6 to -0.6 kb from the ATG), which is the one often altered in AML and lung cancer, and the core promoter (-0.6 and 0 kb from the ATG). A schematic of the CGI is shown in Figure 3a. In order to evaluate the methylation profile of A549 and K562 within the *CEBPA* upstream promoter, COBRA was performed on bisulfite treated DNA and two regions were analyzed (from -1.4 to -1.1 kb and from -0.8 to -0.55 kb from the TSS) (Fig.3b). Aware that COBRA is not a quantitative method, we used this technique only as preliminary approach. As COBRA demonstrated, A549 displayed high methylation level (fully digested) within the sequence located at -1.4 kb from the TSS, whereas displayed a non-methylated sequence (fully undigested) at -0.8 kb from the TSS. On the contrary, the K562 cell line displayed a hemi-methylated region (partially digested) at -0.8 kb from TSS and high methylation profile (fully digested) at -1.4 kb from the TSS. Therefore, since we wanted to assess potential DNA methylation changes upon CEBPA-saRNA treatment, we decided to analyze further the following regions: in A549 cells the sequence located -1.4 kb from the TSS, whereas in the K562 cells the sequence located -0.8 kb from the TSS. Indeed, previous studies demonstrated that the upstream promoter region located between -0.8 and -0.55 kb from the TSS was the most responsive in terms of demethylation in K562 cells [54]. Another region, located between

-1.1 and -0.8 kb, was analyzed in A549 according to the study of Tada *et al* where the methylation profile of the *CEBPA* promoter was evaluated in lung cancer cells [111]. The methylation profile of saRNA-treated A549 cells was analyzed by bisulfite sequencing (Fig.3c). The genomic sequence located between -1.4 and -0.8 kb from the TSS is shown in Figure 3c, upper panel (double pointed black arrow), and comprises 45 CpGs. The middle panel shows the methylation trend, plotted as ratio of methylated versus unmethylated CpGs relative to the respective NC, from 24 to 72 hours post transfection. A demethylation effect was detected at 48 hours post transfection (10%) and became more significant 72 hours post treatment, with a decrease in methylation of about 30% compared to the control. The lollipop representation is shown in the bottom panel, where an increase in demethylated CpGs (white dots) can be easily observed 72 hours post transfection. Bisulfite sequencing analysis was also performed in CEBPA-saRNA treated K562 (Fig.3d). The region between -0.8 and -0.55 kb from the TSS was divided in two sequences, having 16 CpG dinucleotides each (upper panel). Within the 16 CpGs located further from the TSS, a significant demethylation effect was observed 24 and 48 hours post treatment, which then vanished at 72 hours. Remarkably, the wave of demethylation transferred to the second 16 CpGs, located closer to the TSS, where a statistical significant demethylation effect was observed 48 and 72 hours post treatment. Altogether, these data demonstrated that saRNAs, besides inducing re-activation of gene expression, can also induce demethylation within the *CEBPA* promoter in cell lines displaying a different methylation profile.

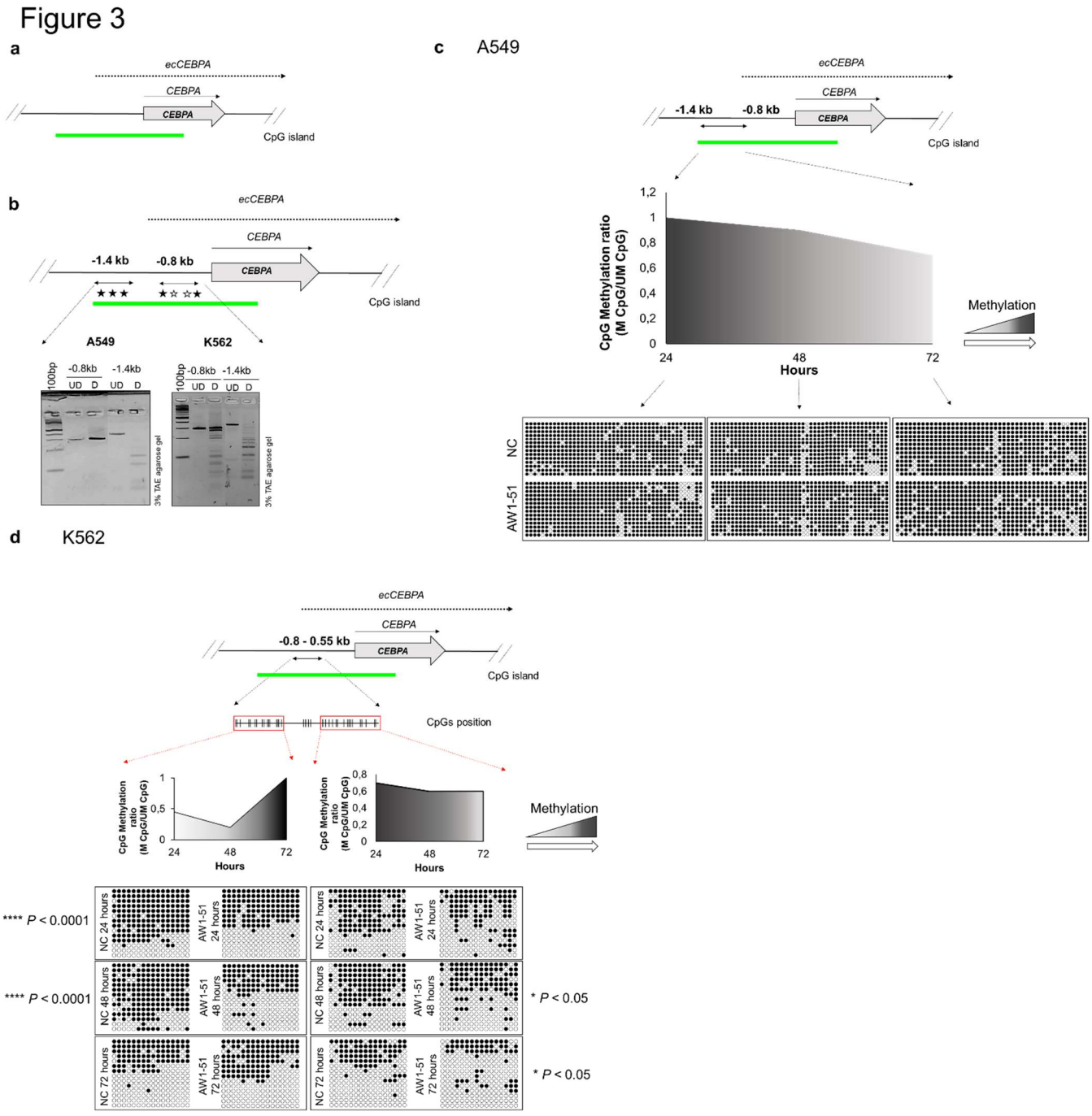

Figure 3 CEBPA-saRNA treatment induces, to different extent, demethylation of the *CEBPA* gene in A549 and K562 cell lines. a, Schematic profile of the CpG island (green rectangle) extending within the promoter and the coding region of the *CEBPA* gene. **b**, Combined Bisulfite Restriction Analysis (COBRA) on bisulfite-converted genomic DNA from A549 and K562 cells. Solid black arrows indicate the analyzed sequences localized at -1.4 and -0.8 kb from the TSS of *CEBPA* and containing 36 and 29 CpG dinucleotides (CpGs), respectively. Stars show position of the BstUI recognition site: black stars indicate one site, white stars indicate two adjacent sites. **c**, Bisulfite Sequencing PCR (BSP) on CEBPA-saRNA-treated samples from A549 cells. The sequence within the CEBPA's promoter and located between -1.4 and -0.8 kb from the TSS is evaluated (upper panel). Middle panel shows relative ratio of methylated (M) to unmethylated (UM) CpGs: dark color indicates high DNA methylation level whereas

light grey color indicates lower DNA methylation level. Lollipop representation (bottom panel) shows methylated (black dot) and unmethylated (white dot) CpGs upon CEBPA-saRNA treatment 24, 48 and 72 hours post transfection (n ≥ 13 clones). **d**, BSP on CEBPA-saRNA-treated K562. Methylation profile of sequence located between -0.8 and -0.55 Kb from the TSS of the gene is evaluated (upper panel). Red rectangles show the CpGs analyzed within the sequence. DNA methylation changes are shown either as the relative ratio of methylated (M) to unmethylated (UM) CpGs in all clones analyzed per each sample (middle panel) and as lollipop representation (bottom panel). Color scale in middle panels can be interpreted as in c. (n = 14 clones). All bisulfite sequenced clones were analyzed by Fisher's exact test. UD: undigested; D: digested; NC: negative control; x: single nucleotide polymorphism (SNP); **** : $P < 0.0001$; * $P < 0.05$.

Characterization of the *P15* locus in various cell lines

In order to choose the appropriate cell models for testing P15-saRNAs, characterization of the *P15* locus (Fig.4a) was performed in K562, KG1a, Raji and HEK293 cell lines. Standard PCR analysis (data not shown) on wild type cells demonstrated the presence or otherwise of the *P15* locus among the mentioned cell lines. Interestingly, only the K562 cell line did not show any amplification, indicating absence of the analyzed sequence. DNA FISH (Fig.4b) on KG1a and K562 cell lines confirmed the results previously obtained, suggesting that K562 cells, lacking the entire *P15* locus, could be considered a natural P15-knockout cell model and a negative control in the study. Since regulatory mechanisms, as DNA methylation, can influence gene expression, levels of the *P15* mRNA were also evaluated (Fig.4c). By means of ssqRT-PCR, the *P15* transcript was measured in K562, KG1a, Raji and HEK293 cell lines. Intriguingly, the KG1a cells, despite having an intact *P15* locus, showed low-to-undetectable levels of the *P15* mRNA, expressed as input of copy number. On the contrary, the Raji and HEK293 cell lines showed medium-to-high expression of the *P15* mRNA. RNA FISH analysis on K562, KG1a and HEK293 validated the results showed by means of ssqRT-PCR, indicating HEK293 as a P15-high expressing cell line whereas KG1a as a P15-non expressing cell line (Fig.4d). Therefore, except for K562, the other cell lines had an intact *P15* locus but different levels of *P15* expression. According to the results, we decided to test the P15-saRNAs in KG1a and Raji cell lines, and used HEK293 as positive control.

Different levels of *P15* mRNA measured in KG1a and Raji might account for different DNA methylation profiles. Early studies on methylation-dependent *P15* silencing in AML demonstrated that *P15* promoter is fully methylated in KG1a whereas is partially methylated in Raji cell line [150, 151]. Li *et al*, while testing saRNAs targeting the *CDH1* gene, realized that epigenetically silenced genes, as *CDH1* in HeLa cells, could not be re-activate by saRNAs alone. Therefore, prior to saRNAs treatment, low doses of 5AC (between 1 to 10 μM) were used to partially demethylate the *locus* and allow gene re-activation upon saRNAs [86]. However, since even low doses of demethylating agent can induce minimal demethylation and gene upregulation throughout the genome [152], we did not treat KG1a cells with 5AC before saRNAs transfection, in order to avoid additional off-targets and unspecific demethylating effects. However, as we did in K562 and A549 prior to CEBPA-saRNA treatment, we synchronized KG1a and Raji cell lines in G1/early S phase as to obtain a homogeneous cell population (Fig5b).

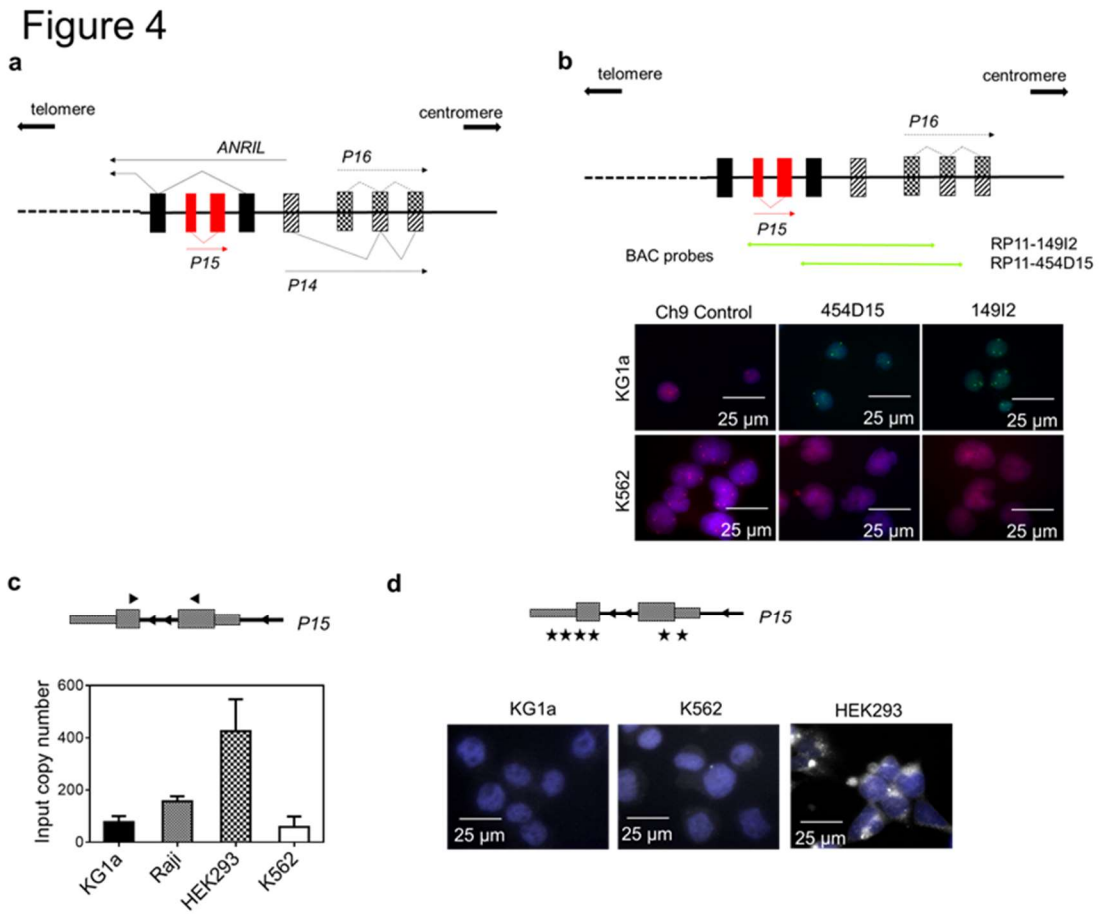

Figure 4 *P15 locus* characterization. **a,** Schematic of the *9p21 locus* and corresponding transcripts. **b,** DNA fluorescent in situ hybridization (FISH) to detect the *9p21 locus* on KG1a and K562 cell lines. The BAC probe RP11-149I2 has coordinates on chr9:21909258-22010413 (hg19); the BAC probe RP11-454D15 has coordinates on chr9: 21819270-21971266 (hg18) (www.empiregenomics.com). **c,** Strand-specific qRT-PCR (ss qRT-PCR) from total RNA to measure copy number of *P15* mRNA in KG1a, Raji, HEK293 and K562 cell lines, expressed as input of copy number. **d,** Detection of *P15* mRNA by RNA FISH on HEK293, KG1a and K562. Black stars indicate position of Stellaris FISH probes in *P15* transcript. In the picture, white dots indicate *P15* mRNA.

The P15-saRNAs induced upregulation of the aberrantly methylated *P15 locus* in KG1a and Raji cell lines

Three saRNAs targeting the *P15* locus were designed and tested in KG1a and Raji cell lines: PR313 (at -375 bp from the TSS, targeting the promoter), PR11 (at +624 bp from the TSS, targeting the first exon) and PR56 (at +1241 bp from the TSS, targeting the intron) (Fig.5a). A negative control, the saRNA-FLUC was also used. 50 nM of each saRNA, both as single saRNA and as mix of all three RNA molecules, were delivered into KG1a and Raji cells

(Fig.5b). Before nucleofection, cells were synchronized by double thymidine block in G1/early S phase. Samples were collected from 24 to 96 hours post nucleofection. Since P15 is a cell cycle regulator and its activation is able to arrest cell cycle progression, cell growth was monitored upon treatment from 24 to 72 hours post nucleofection. Cells growth was not evaluated 96 hours post nucleofection because at 72 hours cells were re-plated at a specific concentration and then re-transfected with lipofectamine. There was not significant change in cell growth upon saRNA treatment, compared to the scramble control FLUC, in KG1a cells. On the contrary, in Raji cell line, treatment with mix of all three saRNAs partially reduced cell growth compared to other treatments and FLUC (Fig.5b).

To address whether P15-saRNAs were efficiently delivered into cells, tagged-saRNAs were designed and tested. As described in Material and Methods, sense and antisense strand of double-strand RNAs were labeled with different fluorophores: Alexa488 identified the sense strand, whereas the ATTO590 identified the antisense strand. This approach allowed us to

evaluated potential differences in transfection efficiency between sense and antisense strand (Fig.6a). Cytofluorimetric analysis was performed 24 hours post transfection in live cells from KG1a and Raji. Only live cells were gated and considered for further analysis. FLUC-treated samples did not show any fluorescence in either cell lines, thus excluding mechanism of autofluorescence due to transfection. In KG1a cells the efficiency of transfection was greater than 50% in each samples, reaching almost 100% in the Mix (Fig.6a, middle panel). In Raji, the efficiency of transfection was about 100% in each sample. Differences in fluorescence from sense and antisense strand could also be observed. In KG1a, except for PR313 and Mix, green fluorescence associated with sense strand was more intense than red fluorescence associated with antisense strand. This phenomenon could depend on processing on the double-stranded RNA and degradation of the fluorophore inside the cells. As in KG1a, in Raji green fluorescence was more intense than red fluorescence, which is barely detectable. Interestingly, cell population showing either sense and antisense fluorescence (yellow color in the graph) reached 60% in PR313 and almost 100% in Mix. On the contrary, PR11 and PR56 showed a higher green fluorescence. Yellow fluorescence identified double positive events (cells), having either green and red fluorescence. However, whether sense (green) and antisense (red) strand were still annealed and yet to be processed or whether, after processing, they were still labeled and fluorescent could not be addressed by standard cytofluorimetric analysis. Overall, these results demonstrated that P15-saRNAs were efficiently transfected in either cell lines.

We further examined *P15* RNA activation by means of ssqRT-PCR. KG1a and Raji cells were transfected with 50 nM P15-saRNAs, as single saRNAs (PR313, PR11 and PR56) and as mix of all three, and mRNA level was measured 48, 72 and 96 hours after transfection. In KG1a, re-expression of the transcript was observed at each time point; the highest increased in *P15* mRNA was detected upon mix treatment, with peak of expression at 96 hours (4.5-fold

increase) (Fig.6b, left panels). On the contrary, in Raji, *P15* mRNA re-expression was observed only at 48 hours upon PR56 treatment, with a 2-fold increase compared to scramble control (Fig.6b, right panels). Altogether, these results indicate P15-saRNAs as innovative and potential tool to re-activate the aberrantly methylated *P15* gene.

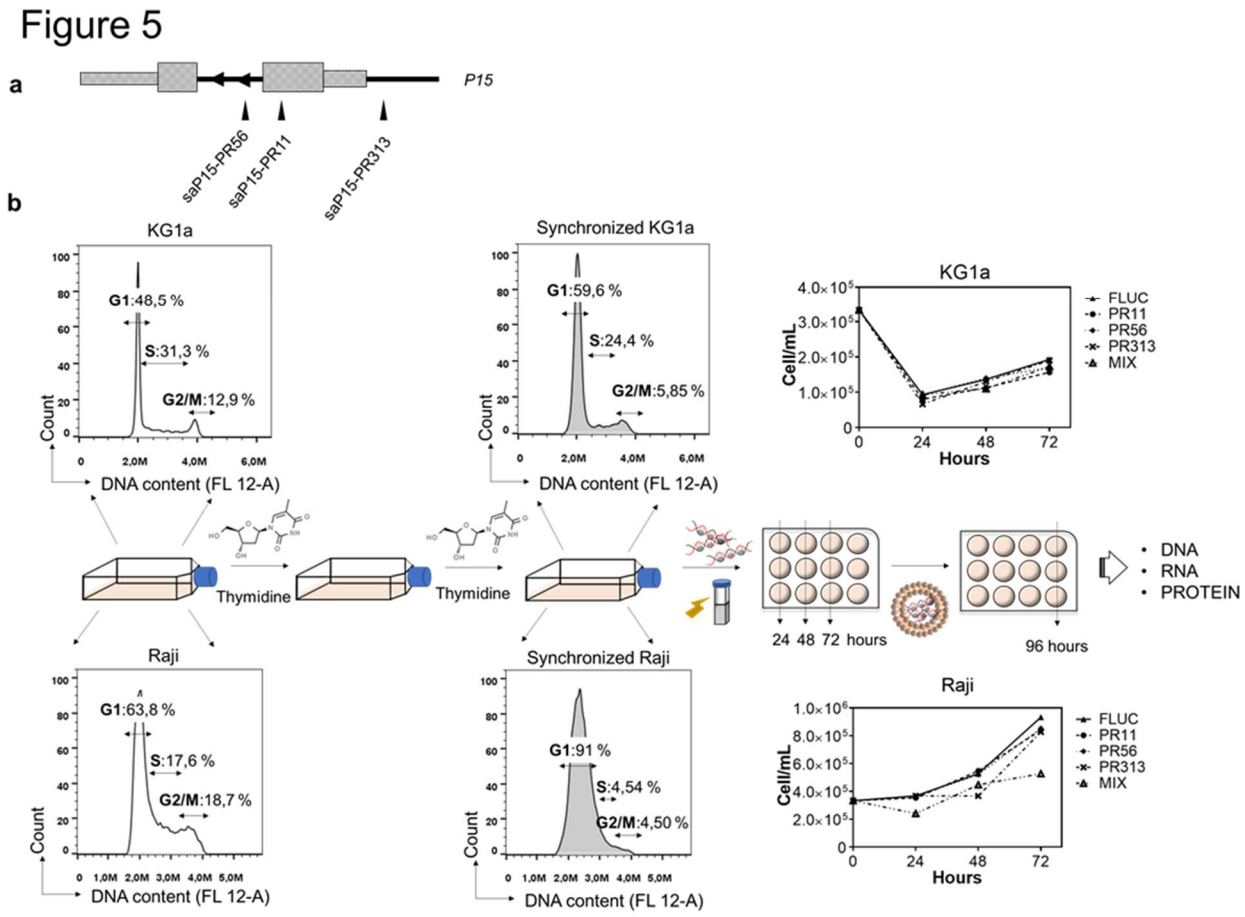

Figure 5 Schematic of P15-saRNA delivery a, Schematic of saRNAs' hotspots targeting the *P15 locus*; from left to right: saP15-PR56 targets the intron, saP15-PR11 targets the first exon, saP15-PR313 targets the promoter. **b,** Schematic of saRNAs' delivery. KG1a and Raji cell lines are first synchronized in G1 phase by double thymidine block (2.5 and 3 mM, respectively); synchronized cells are then electroporated by nucleofection with 50 nM saRNA, as single saRNAs (PR313, PR11 and PR56) or as mix of all three RNAs. DNA and RNA are collected from 24 to 72 hours post nucleofection. 72 hours post-nucleofection cells are transfected with lipofectamine 2000 and 50 nM saRNAs. Samples are collected 24 hours post-lipofectamine transfection. Cell growth is monitored from 0 to 72 hours after transfection.

Figure 6

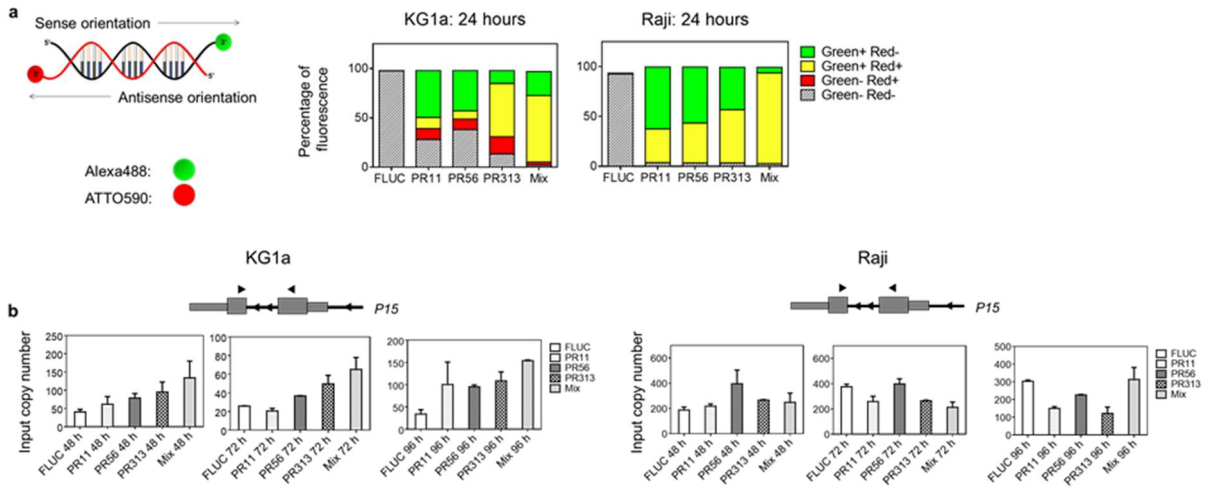

Figure 6 *P15* mRNA reactivation upon saRNAs treatment a, saRNAs' transfection efficiency is evaluated by means of labeled saRNAs in both KG1a and Raji cell lines. Sense and antisense strands are labeled at 3' end with Alexa488 (green) and ATTO590 (red), respectively (left panel). Fifty nM of labeled and double stranded RNAs are delivered into cells by nucleofection and fluorescence is measured by standard cytofluorimetric analysis in live cells (right panel). Yellow color identifies double positive cells, whereas striped pattern double negative cells. **b**, Top left and right, diagram of *P15 locus* with corresponding qRT-PCR primers (black arrows). Bottom left and right, *P15* copy number from total RNA upon saRNAs' treatment 48, 72 and 96 hours post nucleofection (n=2).

P15-saRNAs induce demethylation, to different extent, in KG1a and Raji cell lines, respectively

To examine demethylation effects upon P15-saRNA treatment, we performed bisulfite sequencing analysis on KG1a and Raji cell lines. The *P15* locus has promoter-associated CGI that undergoes aberrant methylation in myeloid malignancies and sporadic parathyroid adenomas [118, 123, 153]. As mentioned before, KG1a cell line is heavily methylated whereas Raji is hemi-methylated within the promoter-associated CGI. Additionally, hypermethylation of the *P15* 3'UTR region is frequently observed in tumors with loss of heterozygosity but without methylation of the promoter-associated CGI, as in primary lymphomas [154]. Therefore, we decided to analyze 16 CpG dinucleotides, flanking the promoter and 5'UTR of the *locus*. In more details, 13 CpGs out of 17 localize within the *P15* promoter, the remaining 4 CpGs localize within the 5'UTR of the gene. Since we observed re-activation of the *locus* at

72/96 hours and 48 hours after transfection in KG1a and Raji, respectively, promoter DNA methylation was evaluated at 48 and 72 hours in either cell lines (Fig.7a and b). In KG1a, a decrease in DNA methylation level was observed 48 hours after transfection upon mix treatment (Fig.7a, right panel), consistent with data of expression. However, the same result was not observed 72 hours after treatment, when also re-expression of the transcripts was detected. This result might be explained with a different DNA modification associated with RNAa, the 5-hydroxymethylcytosine (5hmC), produced by the ten-eleven translocation (TET) family of dioxygenases during active DNA demethylation. Besides being a chemical intermediate, 5hmC is a stable DNA modification that binds to specific regulatory proteins and is mainly found within actively transcribed genes. 5hmC marks decline in cancer tissue, suggesting its potential regulatory role in the mammalian genome [155]. Therefore, to overcome the limitations associated with bisulfite-based approaches, which do not segregate 5mC and 5hmC marks, new bisulfite-free techniques [156], as well as other strategies, have been developed that preserve the genomic DNA integrity [157, 158]. Hence, a different approach able to evaluate 5hmC mark should be considered for evaluating DNA methylation changes in KG1a upon saRNA treatment.

On the contrary, bisulfite sequencing analysis on DNA from PR56-saRNA-treated Raji cells demonstrated a statistical significant demethylation both 48 and 72 hours after treatment (Fig.7b), which, however, did not correspond to an equal gene re-activation. In this regard, Raji cells were derived from B-lymphocytes of a Burkitt's lymphoma and comprise structural aberrations as the t(8;14) translocation, resulting in the juxtaposition of the *MYC* gene to the *IGH* enhancers, which leads to its activation and oncogenic transformation [134, 135]. MYC negatively regulates *P15* expression by inhibiting the *P15* gene's activator MIZ-1, resulting in *P15* transcriptional repression [159]. Therefore, even though we observed demethylation 72

hours post treatment at the *P15* promoter, *P15* mRNA transcription could be inhibited by the activity of MYC, resulting in inconsistent changes in *P15* expression.

Overall these data demonstrated that saRNAs can induce demethylation of targeted genes and, in the presence of heavily methylated *loci*, other epigenetic modifications, as 5hmC could be evaluated.

Figure 7

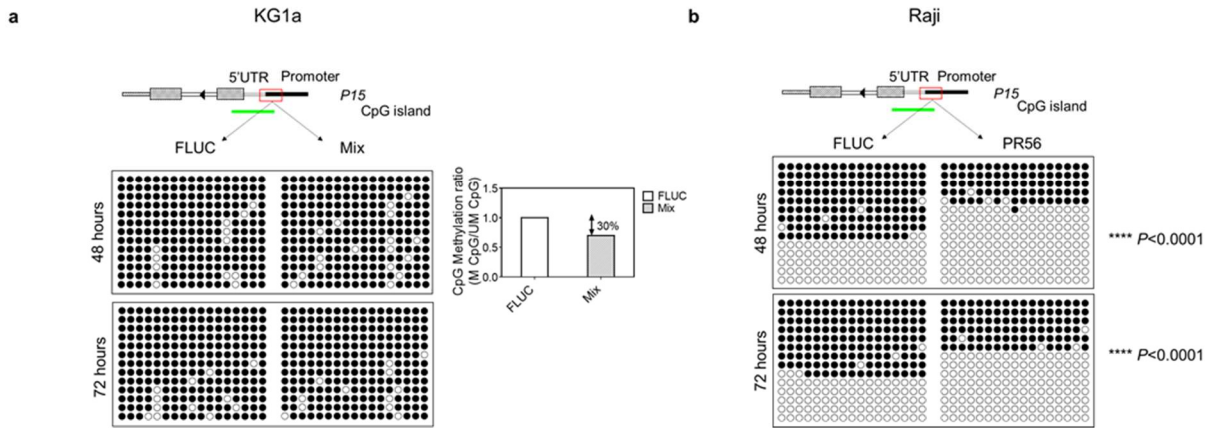

Figure 7 saRNAs mediated demethylation of the *P15* promoter. a, Bisulfite sequencing analysis on KG1a DNA upon Mix treatment 48 and 72 hours after transfection. Upper panel shows schematic of the *P15* locus and the P15 promoter-associated CGI. DNA methylation profile of FLUC and Mix samples is shown as lollipop representation, where black dots indicate methylated CpG dinucleotides whereas white dots unmethylated CpGs (bottom panel). Histogram shows percentage of demethylation 48 hours post treatment, expressed as ratio of methylated (M) to unmethylated (UM) normalized to scramble control (FLUC). **b**, Bisulfite sequencing analysis on Raji DNA upon PR56-saRNA treatment, 48 and 72 hours after transfection. Upper panel shows schematic of the *P15* locus and the P15 promoter-associated CGI. DNA methylation profile of FLUC and PR56 samples is shown as lollipop representation and can be interpreted as in b (bottom panel) ($n \geq 13$ clones). All bisulfite sequenced clones were analyzed by Fisher's exact test.

Discussion

This study explores the role of a recent class of noncoding RNAs, the short activating RNAs (saRNAs) [86, 89, 93]. Using *CEBPA* and *P15* as gene models, we want to investigate whether saRNAs, while re-activating gene transcription, are also able to induce DNA methylation changes thus acting as DNMT1-interacting RNAs mimicking molecules. To increase the chance of interaction between the saRNAs and DNMT1 enzyme, we synchronize all cell lines in G1/early S phase (Fig.1b and d, Fig.5b), when DNMT1's activity peaks.

We show that the CEBPA-saRNA AW1-51 promotes *CEBPA* re-expression in A549 and K562 cell lines (Fig.2a and b, left panels), suggesting its potential therapeutic role in lung cancer and leukemia, respectively. Importantly, we also show that CEBPA-saRNA induces transcription of the extra coding *CEBPA* (*ecCEBPA*), whose expression negatively correlates with methylation of the *CEBPA* promoter and, hence, positively associates with expression of the corresponding coding gene (Fig.2a and b, right panels) [54]. We also investigate DNA methylation profile of either cell lines upon saRNA treatment and we demonstrate that RNA activation (RNAa) induces changes in DNA methylation (Fig.3c and d). Though, in A549, where re-expression of coding and noncoding transcript significantly increases upon treatment, we do not observe an equal significant demethylation within the *CEBPA* promoter. This result can be explained by considering the basal methylation level of the *CEBPA* upstream promoter in A549 and the method we use to analyze DNA methylation profile.

Promoter hypermethylation has been recognized as a hallmark of cancer and frequently marks promoter-associated CGIs of tumor suppressor genes [160, 161]. DNA methylation is an epigenetic feature of heterochromatinic structures, which, together with other regulatory mechanisms [162], contributes in determining genes transcription [163]. As a result, the higher is the methylation level within a promoter-associated CGI, the lower is the accessibility of the *locus* [164]. *CEBPA* upstream promoter is heavily methylated in A549, therefore longer treatment with saRNA might be necessary in order to observe a more significant demethylation.

DNA methylation shows heterogeneity across clones, either in A549 and K562, ranging from 20 to 80% of demethylation compared to control (Fig.3c and d). In K562 the demethylating effect is also dependent on the CpGs' location, closer to or further from the TSS of the *locus* (Fig.3d). It is therefore difficult to predict an accurate demethylating effect if saRNAs would be used in treatment of patients. However, in addition to longer treatment to potentiate the saRNA effect on DNA demethylation, it would be of great help the use of a single cell approach to analyze DNA methylation profile and evaluate clones distribution. Moreover, Li *et al* suggest that, in the presence of highly methylated *loci* as *CDH1* in HeLa cells, the use of low doses of demethylating agents prior to saRNAs treatment will promote RNAa thus resulting in a more accessible *locus* and initial demethylation that might be maintained by saRNAs [86]. A similar trend is observed in KG1a cell line, where re-activation of the *P15* mRNA is detected upon treatment with mix of all three saRNAs at 48, 72 and 96 hours post transfection (Fig.6b, left panel), but demethylation is observed only at 48 hours (Fig.7a). Indeed, as for *CEBPA* in A549, *P15* promoter is highly methylated in KG1a cell line thus resulting in a very close gene *locus* and low-to-undetectable mRNA levels. Besides DNA methylation, more recent studies have highlighted DNA hydroxymethylation as epigenetic modification produced during active DNA demethylation and involved in transcriptional activation. As previously mentioned, 5-hydroxymethylcytosine (5hmC), besides being a chemical intermediate, is also a stable DNA modification that is mainly found within actively transcribed genes. Indeed, 5hmC marks decrease in cancer cells thus suggesting its potential regulatory role [155]. Bisulfite sequencing PCR (BSP) is not able to discriminate between 5-methylcytosine (5mC) and 5hmC, therefore, by detecting only 5mC we might ignore essential information regarding DNA epigenetic modifications upon saRNAs treatment. To overcome limitations associated with bisulfite-based methods, e.g. BSP, bisulfite-free techniques, able to distinguish 5mC and 5hmC, should be used [156-158].

Single-cell analysis is a powerful approach that allows detection of new and potentially unsurprising biological discoveries relative to traditional techniques based on bulk population. Tumor cells comprise heterogeneous populations showing distinct morphology, expression and epigenetic signatures [165, 166]. Some characteristics confer a selective advantage to the cell, such as more rapid growth, and the descendants of a cell bearing such a mutation will consequently become dominant within the tumor population. This *phenomenon* occurs also upon specific treatments as, for examples, saRNAs. saRNA-mediated re-activation of tumor suppressor genes, as *P15*, or transcription factors regulating terminal differentiation and cell cycle arrest, as *CEBPA*, represents a disadvantage for a cell. DNA methylation analysis on bulk population might overestimate methylated clones relative to demethylated ones, thus resulting in unmodified DNA methylation profile. In view of that, results shown in Fig.3c (A549) and in Fig.7a (KG1a) might represent methylated clones with selective advantage over unmethylated ones. Therefore, single clone selection will help to elucidate percentage of partially or fully unmethylated clones within the analyzed population.

Strikingly, DNA methylation analysis in Raji cell line upon PR56-saRNA treatment demonstrates a significant decrease in DNA methylation level at 48 and 72 hours post transfection (Fig.7b), supporting our hypothesis that saRNAs affect, besides gene expression, also DNA methylation. However, we do not observe increase in *P15* transcript expression at 72 hours, as methylation data would suggest. Gene expression, other than by epigenetic marks, is regulated by transcription factors. Raji cell line contains the t(8;14) translocation, which leads to aberrant activation of the *MYC* oncogene [134, 167]. MYC negatively regulates *P15* expression by interacting with MIZ-1, whose activity is essential for transcriptional activation of the *P15 locus* [159]. Therefore, even though PR56-saRNA exerts a significant demethylating effect within the *P15* promoter, transcriptional reactivation is antagonized by *MYC*, that limits *P15* re-expression. Cells doubling time may also be considered in order to interpret the results.

Raji cells have a doubling time of about 18 hours, therefore, if DNMT1's activity is impaired upon saRNA treatment, a demethylating effect might be detected by 48 hours post transfection, as we observed (Fig.7b). Besides DNA methylation profile, cell-cycle too regulates gene expression. A comprehensive study has systematically investigated the transcriptional landscape across cell-cycle, identifying different cell-cycle regulated genes and suggesting a lag between transcription and steady-state expression during the cycle of cells [168]. Ross *et al*, by exploring the expression of 8000 unique genes in 60 cell lines, concluded that gene expression pattern is related to physiological features of the cells, as their doubling time in culture [169]. Therefore, the discrepancy between gene re-expression and DNA demethylation in Raji can be attributed to the characteristics, as transcription factors accessibility and doubling time, of the cell line analyzed.

The causal relationship between DNA methylation pattern and gene expression is not always obvious. While, in some contexts, DNA methylation and transcription are inversely related, in others they might have a direct correlation. In this regard, as already stated in the Introduction, DNA methylation has to be framed in the context of genomic location, resulting in a more complex *scenario* than expected [8]. Moreover, regulatory elements are tissue-specific, even within a promoter [170]. Therefore, considering the *P15 locus* and its complexity, DNA demethylation may occur in regions other than those analyzed in the study and might also be tissue-specific. Indeed, a comprehensive study has shown that differential methylation often occurred in regions surrounding the TSS of tissue-specific *loci*, as first exon [171].

Besides direct and indirect correlation between DNA methylation and gene expression, even temporal relationship between the two events should be investigated. Pacis and colleagues demonstrate that, during infection with *Mycobacterium tuberculosis*, re-expression of specific genes in dendritic cells precedes active demethylation of regulatory elements [172], hence indicating that transcriptional activation itself contributes to DNA demethylation. Therefore,

longer treatment with saRNAs accompanied with longer time points would help unveiling mechanisms of demethylation that are difficult to observe otherwise.

In cancer, either hyper- and hypomethylation play a role in silencing of tumor suppressor genes and in activating genes required for metastasis and invasion, respectively [173]. Therefore, the use of specific agent affecting DNA methylation profile is necessary in order to prevent off-target effects. In this regard, current hypomethylating agents, as 5AC and DAC, regardless their efficacy, demonstrate significant unspecific and toxic effects which limit their clinical use [128-130]. Hence, saRNAs might represent a new therapeutic tool able to modify local DNA methylation profile of target genes. Aiming at evaluating genome-wide methylation effects upon saRNAs, we are currently analyzing data from Infinium MethylationEPIC arrays and RNA sequencing that will reveal unspecific demethylation and activation of pathways correlated to *CEBPA* and *P15* expression, respectively.

Besides gene expression, DNA methylation is essential in controlling mechanisms of splicing. [7]. Small nuclear RNA (snRNA) species play an important role in the splicing process and recognize the sequence, within the nascent RNA, constituting splice and branch site [174]. Therefore, in view of that, saRNAs, by interacting with DNMT1 and affecting DNA methylation, might prevent and/or repair wrongly spliced genes.

Taken together, these data support the hypothesis that short activating RNAs participate in mechanisms of demethylation and pave the way for site-specific adjustments of aberrantly methylated genes in cancer.

References

1. Deichmann, U., Epigenetics : The origins and evolution of a fashionable topic. *Dev. Biol.*, **2016**, *416*, 249-254.
2. Nicoglou, A. and F. Merlin, Epigenetics : A way to bridge the gap between biological fi elds. *Stud. Hist. Philos. Biol. Biomed. Sci.*, **2017**, *66*, 73-82.
3. Goldberg, A.D., C.D. Allis, and E. Bernstein, Epigenetics : A Landscape Takes Shape. *Cell*, **2007**, *128*, 635-638.
4. Moris, N., C. Pina, and A.M. Arias, Transition states and cell fate decisions in epigenetic landscapes. *Nat. Rev. Genet.*, **2016**, *17*, 693-703.
5. Russo, V.E.A., R.A. Martienssen, and A.D. Riggs, in *Epigenetic Mechanisms of Gene Regulation*. 1996, CSHL Press.
6. Bird, A., Perceptions of epigenetics. *Nature*, **2007**, *447*, 396-398.
7. Tirado-Magallanes, R., et al., Whole genome DNA methylation: beyond genes silencing. *Oncotarget*, **2016**, *8*, 5629-5637.
8. Jones, P.A. and D. Takai, The Role of DNA Methylation in Mammalian Epigenetics. *Science*, **2001**, *293*, 1068-1071.
9. Illingworth, R.S. and A.P. Bird, CpG islands - 'A rough guide'. *FEBS Lett.*, **2009**, *583*, 1713-1720.
10. Fazzari, M.J. and J.M. Greally, Epigenomics: beyond CpG islands. *Nat Rev Genet*, **2004**, *5*, 446-455.
11. Saxonov, S., P. Berg, and D.L. Brutlag, A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters. *PNAS*, **2006**, *103*, 1412-1417.
12. Klose, R.J. and A.P. Bird, Genomic DNA methylation : the mark and its mediators. *Trends Biochem. Sci.*, **2006**, *31*.
13. Jones, P.A. and G. Liang, Rethinking how DNA methylation patterns are maintained. *Nat Rev Genet*, **2009**, *10*, 805-811.
14. Jeltsch, A., et al., Mechanism and biological role of Dnmt2 in Nucleic Acid Methylation. *RNA Biol.*, **2017**, *14*, 1108-1123.
15. Goll, M.G., et al., Methylation of tRNAAsp by the DNA methyltransferase homolog Dnmt2. *Science*, **2006**, *311*, 395-398.
16. Santi, Y.L.a.D.V., m5C RNA and m5C DNA methyl transferases use different cysteine residues as catalysts. *PNAS*, **2000**, *97*, 8263–8265.
17. Chedin, F., M.R. Lieber, and C.L. Hsieh, The DNA methyltransferase-like protein DNMT3L stimulates de novo methylation by Dnmt3a. *PNAS*, **2002**, *99*, 16916-16921.
18. Neri, F., et al., Dnmt3L antagonizes DNA methylation at bivalent promoters and favors DNA methylation at gene bodies in ESCs. *Cell*, **2013**, *155*, 121-134.
19. Robertson, K.D., DNA methylation and chromatin – unraveling the tangled web. *Oncogene*, **2002**, *21*, 5361-5379.
20. Yarychkivska, O., et al., BAH domains and a histone-like motif in DNA methyltransferase 1 (DNMT1) regulate de novo and maintenance methylation in vivo. *J. Biol. Chem.*, **2018**, *293*, 19466-19475.
21. Elliott, E.N., K.L. Sheaffer, and K.H. Kaestner, The 'de novo' DNA methyltransferase Dnmt3b compensates the Dnmt1-deficient intestinal epithelium. *Elife*, **2016**, *5*.
22. Song, Y., et al., Dynamic Enhancer DNA Methylation as Basis for Transcriptional and Cellular Heterogeneity of ESCs. *Mol. Cell*, **2019**, *75*, 905-920.e906.

23. Liebl, K. and M. Zacharias, How methyl-sugar interactions determine DNA structure and flexibility. *NAR*, **2019**, *47*, 1132-1140.
24. Skvortsova, K., et al., DNA Hypermethylation Encroachment at CpG Island Borders in Cancer Is Predisposed by H3K4 Monomethylation Patterns. *Cancer Cell*, **2019**, *35*, 297-314.e298.
25. Ollikainen, M., et al., Genome-wide blood DNA methylation alterations at regulatory elements and heterochromatic regions in monozygotic twins discordant for obesity and liver fat. *Clin Epigenetics*, **2015**, *7*, 39.
26. Visone, R., et al., DNA methylation of shelf, shore and open sea CpG positions distinguish high microsatellite instability from low or stable microsatellite status colon cancer stem cells. *Epigenomics*, **2019**, *11*, 587-604.
27. Irizarry, R.A., et al., The human colon cancer methylome shows similar hypo- and hypermethylation at conserved tissue-specific CpG island shores. *Nat. Genet.*, **2009**, *41*, 178-186.
28. Hellman, A. and A. Chess, Gene body-specific methylation on the active X chromosome. *Science*, **2007**, *315*, 1141-1143.
29. Teissandier, A. and D. Bourc'his, Gene body DNA methylation conspires with H3K36me3 to preclude aberrant transcription. *EMBO J.*, **2017**, *36*, 1471-1473.
30. Arechederra, M., et al., Hypermethylation of gene body CpG islands predicts high dosage of functional oncogenes in liver cancer. *Nat. Commun.*, **2018**, *9*, 3164.
31. Sartor, R.C., et al., Identification of the expressome by machine learning on omics data. *PNAS*, **2019**, *116*, 18119-18125.
32. Sleutels, F. and D.P. Barlow, The Origins of Genomic. *Adv. Genet.*, **2002**, *46*, 119-163.
33. Gaszner, M. and G. Felsenfeld, Insulators: exploiting transcriptional and epigenetic mechanisms. *Nat Rev Genet*, **2006**, *7*, 703-713.
34. Alasoo, K., et al., Shared genetic effects on chromatin and gene expression indicate a role for enhancer priming in immune response. *Nat. Genet.*, **2018**, *50*, 424-431.
35. Aran, D. and A. Hellman, DNA Methylation of Transcriptional Enhancers and Cancer Predisposition. *Cell*, **2013**, 11-13.
36. Tatetsu, H., et al., Down-regulation of PU.1 by methylation of distal regulatory elements and the promoter is required for myeloma cell growth. *Cancer Res.*, **2007**, *67*, 5328-5336.
37. West, A.G., M. Gaszner, and G. Felsenfeld, Insulators : many functions , many mechanisms. *Genes Dev.*, **2002**, *16*, 271-288.
38. Wang, H., et al., Widespread plasticity in CTCF occupancy linked to DNA methylation. *Genome Res.*, **2012**, *22*, 1680-1688.
39. Bell, A.C. and G. Felsenfeld, Methylation of a CTCF-dependent boundary controls imprinted expression of the Igf2 gene. *Nature*, **2000**, *405*, 2-5.
40. Rao, X., et al., CpG island shore methylation regulates caveolin-1 expression in breast cancer. *Oncogene*, **2013**, 4519-4528.
41. Mattick, J.S., Non-coding RNAs: the architects of eukaryotic complexity. *EMBO reports*, **2001**, *2*, 986–991.
42. Goff, L.A. and J.L. Rinn, Linking RNA biology to lncRNAs. *Genome Res*, **2015**, *25*, 1456-1465.
43. Mattick, J.S. and J.L. Rinn, Discovery and annotation of long noncoding RNAs. *Nat Struct Mol Biol*, **2015**, *22*, 5-7.
44. Ma, L., V.B. Bajic, and Z. Zhang, On the classification of long non-coding RNAs. *RNA Biol*, **2013**, *10*, 925-933.

45. Ulitsky, I. and D.P. Bartel, lincRNAs: genomics, evolution, and mechanisms. *Cell*, **2013**, *154*, 26-46.
46. Kopp, F. and J.T. Mendell, Functional Classification and Experimental Dissection of Long Noncoding RNAs. *Cell*, **2018**, *172*, 393-407.
47. Da Rocha, S.T. and E. Heard, Novel players in X inactivation: insights into Xist-mediated gene silencing and chromosome conformation. *Nat Struct Mol Biol*, **2017**, *24*, 197-204.
48. Chu, C., et al., Systematic discovery of Xist RNA binding proteins. *Cell*, **2015**, *161*, 404-416.
49. Chen, C.K., et al., Xist recruits the X chromosome to the nuclear lamina to enable chromosome-wide silencing. *Science*, **2016**, *354*, 468-472.
50. Engreitz, J.M., et al., Local regulation of gene expression by lncRNA promoters, transcription and splicing. *Nature*, **2016**, *539*, 452-455.
51. Groff, A.F., et al., In Vivo Characterization of Linc-p21 Reveals Functional cis-Regulatory DNA Elements. *Cell Rep*, **2016**, *16*, 2178-2186.
52. Guttman, J.R.a.M., RNA and dynamic nuclear organization. *Science*, **2014**, *345*.
53. Fang, H., et al., Trans- and cis-acting effects of the lncRNA Firre on epigenetic and structural features of the inactive X chromosome. *bioRxiv*, **2019**, 687236.
54. Di Ruscio, A., et al., DNMT1-interacting RNAs block gene-specific DNA methylation. *Nature*, **2013**, *503*, 371-376.
55. Kim, V.N., J. Han, and M.C. Siomi, Biogenesis of small RNAs in animals. *Nat. Rev. Mol. Cell Biol.*, **2009**, *10*, 126-139.
56. Romano, G., et al., Small non-coding RNA and cancer. *Carcinogenesis*, **2017**, *38*, 485-491.
57. Borchert, G.M., W. Lanier, and B.L. Davidson, RNA polymerase III transcribes human microRNAs. *Nat Struct Mol Biol*, **2006**, *13*, 1097-1101.
58. Bartel, D.P., MicroRNAs: Genomics, Biogenesis, Mechanism, and Function. *Cell*, **2004**, *116*, 281-297.
59. Tuschl, G.M.a.T., Mechanisms of gene silencing by double-stranded RNA. *Nature*, **2004**, *431*, 343-349.
60. Chi, S.W., G.J. Hannon, and R.B. Darnell, An alternative mode of microRNA target recognition. *Nat Struct Mol Biol*, **2012**, *19*, 321-327.
61. Carthew, R.W. and E.J. Sontheimer, Origins and Mechanisms of miRNAs and siRNAs. *Cell*, **2009**, *136*, 642-655.
62. Lee, R.C., R.L. Feinbaum, and V. Ambros, The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. *Cell*, **1993**, *75*, 843-854.
63. Hwang, H.W., E.A. Wentzel, and J.T. Mendell, A hexanucleotide element directs microRNA nuclear import. *Science*, **2007**, *315*, 97-100.
64. Catalanotto, C., C. Cogoni, and G. Zardo, MicroRNA in Control of Gene Expression: An Overview of Nuclear Functions. *Int J Mol Sci*, **2016**, *17*.
65. Turunen, T.A., et al., Changes in nuclear and cytoplasmic microRNA distribution in response to hypoxic stress. *Sci Rep*, **2019**, *9*, 10332.
66. Liu, H., et al., Nuclear functions of mammalian MicroRNAs in gene regulation, immunity and cancer. *Mol Cancer*, **2018**, *17*, 64.
67. Morris, K.V., et al., Small Interfering RNA–Induced Transcriptional Gene Silencing in Human Cells. *Science*, **2004**, *305*, 1289-1291.
68. Ting, A.H., et al., Short double-stranded RNA induces transcriptional gene silencing in human cancer cells in the absence of DNA methylation. *Nat. Genet.*, **2005**, *37*, 906-910.

69. Karunanithi, S., et al., Exogenous RNAi mechanisms contribute to transcriptome adaptation by phased siRNA clusters in Paramecium. *NAR*, **2019**, *47*, 8036-8049.
70. Watanabe, T., et al., Endogenous siRNAs from naturally formed dsRNAs regulate transcripts in mouse oocytes. *Nature*, **2008**, *453*, 539-543.
71. Tam, O.H., et al., Pseudogene-derived small interfering RNAs regulate gene expression in mouse oocytes. *Nature*, **2008**, *453*, 534-538.
72. Lai, K.O.a.E.C., Endogenous small interfering RNAs in animals. *Nat. Rev. Mol. Cell Biol.*, **2008**, *9*, 673-678.
73. Castanotto, D., et al., Short hairpin RNA-directed cytosine (CpG) methylation of the RASSF1A gene promoter in HeLa cells. *Mol Ther*, **2005**, *12*, 179-183.
74. Sørensen, D.R., M. Leirdal, and M. Sioud, Gene Silencing by Systemic Delivery of Synthetic siRNAs in Adult Mice. *Journal of Molecular Biology*, **2003**, *327*, 761-766.
75. Lam, J.K., et al., siRNA Versus miRNA as Therapeutics for Gene Silencing. *Mol Ther Nucleic Acids*, **2015**, *4*, e252.
76. Sui, H., et al., siRNA containing a unique 5-nucleotide motif acts as a quencher of IFI16-mediated innate immune response. *Mol Immunol*, **2019**, *114*, 330-340.
77. Maillard, P.V., et al., Inactivation of the type I interferon pathway reveals long double-stranded RNA-mediated RNA interference in mammalian cells. *EMBO J.*, **2016**, *35*, 2505-2518.
78. Luteijn, M.J. and R.F. Ketting, PIWI-interacting RNAs: from generation to transgenerational epigenetics. *Nat Rev Genet*, **2013**, *14*, 523-534.
79. Kalmykova, A.I., M.S. Klenov, and V.A. Gvozdev, Argonaute protein PIWI controls mobilization of retrotransposons in the Drosophila male germline. *NAR*, **2005**, *33*, 2052-2059.
80. Khurana, J.S. and W. Theurkauf, piRNAs, transposon silencing, and Drosophila germline development. *J Cell Biol*, **2010**, *191*, 905-913.
81. Castel, S.E. and R.A. Martienssen, RNA interference in the nucleus: roles for small RNAs in transcription, epigenetics and beyond. *Nat Rev Genet*, **2013**, *14*, 100-112.
82. Ozata, D.M., et al., PIWI-interacting RNAs: small RNAs with big functions. *Nat Rev Genet*, **2019**, *20*, 89-108.
83. Tan, L., et al., PIWI-interacting RNA-36712 restrains breast cancer progression and chemoresistance by interaction with SEPW1 pseudogene SEPW1P RNA. *Mol Cancer*, **2019**, *18*, 9.
84. Fu, A., D.I. Jacobs, and Y. Zhu, Epigenome-wide analysis of piRNAs in gene-specific DNA methylation. *RNA Biol*, **2014**, *11*, 1301-1312.
85. Portnoy, V., et al., Small RNA and transcriptional upregulation. *Wiley Interdiscip Rev RNA*, **2011**, *2*, 748-760.
86. Li, L.-C., et al., Small dsRNAs induce transcriptional activation in human cells. *PNAS*, **2006**, *103*, 17337–17342.
87. Wang, J., et al., Inducing gene expression by targeting promoter sequences using small activating RNAs. *J Biol Methods*, **2015**, *2*.
88. Voutila, J., et al., Development and Mechanism of Small Activating RNA Targeting CEBPA, a Novel Therapeutic in Clinical Trials for Liver Cancer. *Molecular Therapy*, **2017**, *25*, 2705-2714.
89. Kuwabara, T., et al., A Small Modulatory dsRNA Specifies the Fate of Adult Neural Stem Cells. *Cell*, **2004**, *116*, 779–793.
90. Jopling, C.L., et al., Modulation of hepatitis C virus RNA abundance by a liver-specific MicroRNA. *Science (New York, N.Y.)*, **2005**, *309*, 1577-1581.
91. Place, F.R., et al., Defining Features and Exploring Chemical Modifications to Manipulate RNAa Activity. *Curr Pharm Biotechnol*, **2010**, *11*, 518-526.

92. Yi, R., et al., Exportin-5 mediates the nuclear export of pre-microRNAs and short hairpin RNAs. *Genes Dev.*, **2003**, *17*, 3011-3016.
93. Yoon, S. and J.J. Rossi, Therapeutic Potential of Small Activating RNAs (saRNAs) in Human Cancers. *Curr Pharm Biotechnol*, **2018**, *19*, 604-610.
94. Porse, B.T., et al., Loss of C/EBP alpha cell cycle control increases myeloid progenitor proliferation and transforms the neutrophil granulocyte lineage. *J Exp Med*, **2005**, *202*, 85-96.
95. Smith, M.L., et al., Mutation of CEBPA in Familial Acute Myeloid Leukemia. *The New England Journal of Medicine*, **2004**, *351*, 2403–2407.
96. Koschmieder, S., et al., Dysregulation of the C/EBPalpha differentiation pathway in human cancer. *J Clin Oncol*, **2009**, *27*, 619-628.
97. Hughes, J.M., et al., C/EBPα-p30 protein induces expression of the oncogenic long non-coding RNA UCA1 in acute myeloid leukemia. *Oncotarget*, **2015**, *6*.
98. Muller, C., et al., Nucleolar retention of a translational C/EBPalpha isoform stimulates rDNA transcription and cell size. *EMBO J.*, **2010**, *29*, 897-909.
99. Gorgens, A., et al., Revision of the human hematopoietic tree: granulocyte subtypes derive from distinct hematopoietic lineages. *Cell Rep*, **2013**, *3*, 1539-1552.
100. Paul, F., et al., Transcriptional Heterogeneity and Lineage Commitment in Myeloid Progenitors. *Cell*, **2015**, *163*, 1663-1677.
101. Imperato, M.R., et al., The RUNX1-PU.1 axis in the control of hematopoiesis. *Int J Hematol*, **2015**, *101*, 319-329.
102. Friedman, A.D., C/EBPalpha in normal and malignant myelopoiesis. *Int J Hematol*, **2015**, *101*, 330-341.
103. Ma, O., et al., Granulopoiesis requires increased C/EBPalpha compared to monopoiesis, correlated with elevated Cebpa in immature G-CSF receptor versus M-CSF receptor expressing cells. *PLoS One*, **2014**, *9*, e95784.
104. Zhang, D.-E., et al., Absence of granulocyte colony-stimulating factor signaling andneutrophil development in CCAAT enhancer binding protein a-deficient mice. *PNAS*, **1997**, *94*, 569–574,.
105. Chim, C.S., A.S.Y. Wong, and Y.L. Kwong, Infrequent hypermethylation of CEBPA promotor in acute myeloid leukaemia. *Br. J. Haematol.*, **2002,** , *119* 988–990.
106. Hackanson, B., et al., Epigenetic modification of CCAAT/enhancer binding protein alpha expression in acute myeloid leukemia. *Cancer Res.*, **2008**, *68*, 3142-3151.
107. Poetsch, A.R. and C. Plass, Transcriptional regulation by DNA methylation. *Cancer Treat Rev*, **2011**, *37 Suppl 1*, S8-12.
108. Lourenco, A.R. and P.J. Coffer, A tumor suppressor role for C/EBPalpha in solid tumors: more than fat and blood. *Oncogene*, **2017**, *36*, 5221-5230.
109. LI, F., et al., Correlation of expression of transcription factor C/EBPa and surfactant protein genes in lung cells *American Journal of Physiology*, **1995**, *269*, 241-247.
110. Halmos, B., et al., Down-Regulation and Antiproliferative Role of C/EBPa in Lung Cancer. *Cancer Res.*, **2002**, *62*, 528–534.
111. Tada, Y., et al., Epigenetic modulation of tumor suppressor CCAAT/enhancer binding protein alpha activity in lung cancer. *J Natl Cancer Inst*, **2006**, *98*, 396-406.
112. Brown, V.L., et al., p16INK4a and p14ARF Tumor Suppressor Genes Are Commonly Inactivated in Cutaneous Squamous Cell Carcinoma. *Journal of Investigative Dermatology*, **2004**, *122*, 1284-1292.
113. Holdt, L.M. and D. Teupser, Long Noncoding RNA ANRIL: Lnc-ing Genetic Variation at the Chromosome 9p21 Locus to Molecular Mechanisms of Atherosclerosis. *Front Cardiovasc Med*, **2018**, *5*, 145.

114. Pasmant, E., et al., Characterization of a germ-line deletion, including the entire INK4/ARF locus, in a melanoma-neural system tumor family: identification of ANRIL, an antisense noncoding RNA whose expression coclusters with ARF. *Cancer Res.*, **2007**, *67*, 3963-3969.
115. Fuxe, J., E. Raschperger, and R.F. Pettersson, Translation of p15.5INK4B, an N-terminally extended and fully active form of p15INK4B, is initiated from an upstream GUG codon. *Oncogene* **2000** *19*, 1724-1728.
116. Tien, H.-F., et al., Methylation of the p15INK4B gene in myelodysplastic syndrome: it can be detected early at diagnosis or during disease progression and is highly associated with leukaemic transformation. *Br. J. Haematol.*, **2001**, *112*, 148-154.
117. Pavletich, N.P., Mechanisms of Cyclin-dependent Kinase Regulation: Structures of Cdks, their Cyclin Activators, and Cip and INK4 Inhibitors. *Journal of Molecular Biology*, **1999**, *287*, 821-828.
118. Wolff, L. and J. Bies, p15Ink4b Functions in determining hematopoietic cell fates: implications for its role as a tumor suppressor. *Blood Cells Mol Dis*, **2013**, *50*, 227-231.
119. Teofili, L., et al., Expression of p15ink4b gene during megakaryocytic differentiation of normal and myelodysplastic hematopoietic progenitors. *Blood*, **2001**, *98*, 495-497.
120. Kusy, S., C.J. Larsen, and J. Roche, p14ARF, p15INK4b and p16INK4a methylation status in chronic myelogenous leukemia. *Leuk Lymphoma*, **2004**, *45*, 1989-1994.
121. Teofili, L., et al., Expression of cyclin-dependent kinase inhibitor p15INK4B during normal and leukemic myeloid differentiation. *Experimental Hematology*, **2000**, *28*, 519–526.
122. Figueroa, M.E., et al., MDS and secondary AML display unique patterns and abundance of aberrant DNA methylation. *Blood*, **2009**, *114*, 3448-3458.
123. Fares, J., L. Wolff, and J. Bies, CDKN2B (cyclin-dependent kinase inhibitor 2B (p15, inhibits CDK4)). *Atlas of Genetics and Cytogenetics in Oncology and Haematology*, **2012**.
124. Latres, E., et al., Limited overlapping roles of P15INK4b and P18INK4c cell cycle inhibitors in proliferation and tumorigenesis. *The EMBO Journal*, **2000**, *19*, 3496-3506.
125. Rosu-Myles, M., B.J. Taylor, and L. Wolff, Loss of the tumor suppressor p15Ink4b enhances myeloid progenitor formation from common myeloid progenitors. *Exp Hematol*, **2007**, *35*, 394-406.
126. Bies, J., et al., Myeloid-specific inactivation of p15Ink4b results in monocytosis and predisposition to myeloid leukemia. *Blood*, **2010**, *116*, 979-987.
127. Mehdipour, P., T. Murphy, and D.D. De Carvalho, The role of DNA-demethylating agents in cancer therapy. *Pharmacol Ther*, **2019**, 107416.
128. Diesch, J., et al., A clinical-molecular update on azanucleoside-based therapy for the treatment of hematologic cancers. *Clin Epigenetics*, **2016**, *8*, 71.
129. Shen, H. and P.W. Laird, In epigenetic therapy, less is more. *Cell Stem Cell*, **2012**, *10*, 353-354.
130. Kelly, T.K., D.D. De Carvalho, and P.A. Jones, Epigenetic modifications as therapeutic targets. *Nat Biotechnol*, **2010**, *28*, 1069-1078.
131. Reebye, V., et al., Gene activation of CEBPA using saRNA: preclinical studies of the first in human saRNA drug candidate for liver cancer. *Oncogene*, **2018**, *37*, 3216-3228.
132. Voutila, J., et al., Gene Expression Profile Changes After Short-activating RNA-mediated Induction of Endogenous Pluripotency Factors in Human Mesenchymal Stem Cells. *Mol Ther Nucleic Acids*, **2012**, *1*, e35.

133. Perrotti, D., et al., BCR-ABL suppresses C/EBPalpha expression through inhibitory action of hnRNP E2. *Nat. Genet.*, **2002**, *30*, 48-58.
134. Masauzi, N., et al., A translocation t(8;14) and c-myc gene rearrangement associated with the histological transformation of B-cell acute lymphocytic leukemia (FAB-L2) into Burkitt's type (FAB-L3) leukemia. *Leuk Lymphoma*, **1997**, *27*, 357-363.
135. Karpova, M.B., et al., Raji revisited: cytogenetics of the original Burkitt's lymphoma cell line. *Leukemia*, **2005**, *19*, 159-161.
136. Livak, K.J. and T.D. Schmittgen, Analysis of relative gene expression data using real-time quantitative PCR and the 2(-Delta Delta C(T)) Method. *Methods*, **2001**, *25*, 402-408.
137. Frommer, M., et al., A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual DNA strands. *PNAS*, **1992**, *89*, 1827-1831.
138. Li, Y. and T.O. Tollefsbol, DNA methylation detection: bisulfite genomic sequencing analysis. *Methods Mol. Biol.*, **2011**, *791*, 11-21.
139. Xiong, Z. and P.W. Laird, COBRA : a sensitive and quantitative DNA methylation assay. *NAR*, **1997**, *25*, 2532-2534.
140. Banfalvi, G., Overview of cell synchronization. *Methods Mol. Biol.*, **2011**, *761*, 1-23.
141. Ma, H.T. and R.Y. Poon, Synchronization of HeLa cells. *Methods Mol. Biol.*, **2011**, *761*, 151-161.
142. Xeros, N., Deoxyriboside Control and Synchronization of Mitosis. *Nature*, **1962**, *194*, 682-683.
143. Bjursell, G.a.R., P., Effects of Thymidine on Deoxyribonucleoside Triphosphate Pools and Deoxyribonucleic Acid Synthesis in Chinese Hamster Ovary Cells. *Journal of Biological Chemistry*, **1973**, *248*, 3904-3909.
144. Reebye, V., et al., Novel RNA oligonucleotide improves liver function and inhibits liver carcinogenesis in vivo. *Hepatology*, **2014**, *59*, 216-227.
145. Koschmieder, S., et al., CDDO induces granulocytic differentiation of myeloid leukemic blasts through translational up-regulation of p42 CCAAT enhancer binding protein alpha. *Blood*, **2007**, *110*, 3695-3705.
146. Radomska, H.S., et al., CAAT/Enhancer Binding Protein Alpha Is a Regulatory Switch Sufficient for Induction of Granulocytic Development from Bipotential Myeloid Progenitors. *MOLECULAR AND CELLULAR BIOLOGY*, **1998**, 4301-4314.
147. Leonhardt, H., et al., A targeting sequence directs DNA methyltransferase to sites of DNA replication in mammalian nuclei. *Cell*, **1992**, *71*, 865-873.
148. Preker, P., et al., RNA Exosome Depletion Reveals Transcription Upstream of Active Human Promoters. *Science*, **2008**, *322*, 1851-1854.
149. Ebisuya, M., et al., Ripples from neighbouring transcription. *Nat Cell Biol*, **2008**, *10*, 1106-1113.
150. Herman, J.G., et al., Hypermethylation-associated Inactivation Indicates a Tumor Suppressor Role for p15INK4B1. *Cancer Res.*, **1996**, *56*, 722-727.
151. Martel, V., et al., De novo methylation of tumour suppressor genes CDKN2A and CDKN2B is a rare finding in B-cell chronic lymphocytic leukaemia. *British Journal of Haematology,* , **1997**, *99*, 320–324.
152. Cameron, E.E., et al., Synergy of demethylation and histone deacetylase inhibition in the re-expression of genes silenced in cancer. *Nat. Genet.*, **1999**, *21*, 103-107.
153. Arya, A.K., et al., Promoter hypermethylation inactivates CDKN2A, CDKN2B and RASSF1A genes in sporadic parathyroid adenomas. *Sci Rep*, **2017**, *7*, 3123.
154. Malumbres, M., et al., Hypermethylation of the cell cycle inhibitor p15INK4b 3'-untranslated regioninterferes with its transcriptional regulation in primary lymphomas. *Oncogene*, **1999**, *18*, 385-396.

155. Bachman, M., et al., 5-Hydroxymethylcytosine is a predominantly stable DNA modification. *Nat. Chem.*, **2014**, *6*, 1049-1055.
156. Liu, Y., et al., Bisulfite-free direct detection of 5-methylcytosine and 5-hydroxymethylcytosine at base resolution. *Nat. Biotechnol.*, **2019**, *37*.
157. Skvortsova, K., et al., Comprehensive evaluation of genome - wide 5 - hydroxymethylcytosine profiling approaches in human DNA. *Epigenetics Chromatin*, **2017**, *10:16*, 1-20.
158. Song, C.-X., et al., Simultaneous single-molecule epigenetic imaging of DNA methylation and hydroxymethylation. *PNAS*, **2016**, 3-8.
159. Staller, P., et al., Repression of p15INK4b expression by Myc through association with Miz-1. *Cell Bio.*, **2001**, *3*.
160. Esteller, M., *CANCER EPIGENETICS: DNA METHYLATION AND CHROMATIN ALTERATIONS IN HUMAN CANCER*, in *New Trends in Cancer for the 2/" Century*, L.-B.a. Felipo, Editor. 2003, Kluwer Academic/Plenum Publishers. p. 39-49.
161. Watanabe, Y. and M. Maekawa, *Methylation of DNA in Cancer*. 2010. p. 145-167.
162. Baylin, S.B., DNA methylation and gene silencing in cancer. *Nat. Clin. Pract. Oncol.*, **2005**, *2*, 4-11.
163. Vignaux, P.A., C. Bregio, and N.A. Hathaway, Contribution of promoter DNA sequence to heterochromatin formation velocity and memory of gene repression in mouse embryo fibroblasts. *PLoS One*, **2019**, *14*, e0217699.
164. Liu, G., et al., Inherited DNA methylation primes the establishment of accessible chromatin during genome activation. *Genome Res*, **2018**, *28*, 998-1007.
165. Hwang, B., J.H. Lee, and D. Bang, Single-cell RNA sequencing technologies and bioinformatics pipelines. *Exp Mol Med*, **2018**, *50*, 96.
166. Gross, A., et al., Technologies for Single-Cell Isolation. *Int J Mol Sci*, **2015**, *16*, 16897-16919.
167. Haluska, F.G., et al., The t(8;14) chromosomal translocation occurring in B-cell malignancies results from mistakes in V-D-J joining. *Nature*, **1986**, *324*, 158-161.
168. Liu, Y., et al., Transcriptional landscape of the human cell cycle. *PNAS*, **2017**, *114*, 3473-3478.
169. Douglas T. Ross, U.S., Michael B. Eisen, Charles M. Perou, Christian Rees, Paul Spellman, Vishwanath Iyer, Stefanie S. Jeffrey, Matt Van de Rijn, Mark Waltham, Alexander Pergamenschikov, Jeffrey C.F. Lee, Deval Lashkari, Dari Shalon, Timothy G. Myers, John N. Weinstein, David Botstein &Patrick O. Brown, Systematic variation in gene expression patterns in human cancer cell lines. *Nat. Genet.*, **2000**, *24*.
170. Suh, H.C., et al., LPS independent activation of the pro-inflammatory receptor Trem1 by C/EBPepsilon in granulocytes. *Sci Rep*, **2017**, *7*, 46440.
171. Hodges, E., et al., Directional DNA methylation changes and complex intermediate states accompany lineage specificity in the adult hematopoietic compartment. *Mol. Cell*, **2011**, *44*, 17-28.
172. Pacis, A., et al., Gene activation precedes DNA demethylation in response to infection in human dendritic cells. *PNAS*, **2019**, *116*, 6938-6943.
173. Szyf, M., P. Pakneshan, and S.A. Rabbani, DNA demethylation and cancer: therapeutic implications. *Cancer Lett*, **2004**, *211*, 133-143.
174. Sharp, P.A., The centrality of RNA. *Cell*, **2009**, *136*, 577-580.

Appendix

cells

MDPI

Review

The Bright and Dark Side of DNA Methylation: A Matter of Balance

Marta Borchiellini [1,2], Simone Ummarino [3,*] and Annalisa Di Ruscio [2,3,*]

[1] Department of Health Sciences, University of Eastern Piedmont, 28100 Novara, Italy; marta.borchiellini@uniupo.it
[2] Department of Translational Medicine, University of Eastern Piedmont, 28100 Novara, Italy
[3] Harvard Medical School Initiative for RNA Medicine, Harvard Medical School, Boston, MA 02115, USA
* Correspondence: summarin@bidmc.harvard.edu (S.U.); adirusci@bidmc.harvard.edu (A.D.R.); Tel.: +1-617-735-2228 (S.U.); +1-617-735-2218 (A.D.R.)

Received: 20 September 2019; Accepted: 10 October 2019; Published: 12 October 2019

Abstract: DNA methylation controls several cellular processes, from early development to old age, including biological responses to endogenous or exogenous stimuli contributing to disease transition. As a result, minimal DNA methylation changes during developmental stages drive severe phenotypes, as observed in germ-line imprinting disorders, while genome-wide alterations occurring in somatic cells are linked to cancer onset and progression. By summarizing the molecular events governing DNA methylation, we focus on the methods that have facilitated mapping and understanding of this epigenetic mark in healthy conditions and diseases. Overall, we review the bright (health-related) and dark (disease-related) side of DNA methylation changes, outlining how bulk and single-cell genomic analyses are moving toward the identification of new molecular targets and driving the development of more specific and less toxic demethylating agents.

Keywords: epigenetics; DNA methylation; DNMTs; imprinting; cancer

1. DNA Methylation: More than One Purpose

DNA methylation is a key epigenetic signature (Box 1, Figure 1) implicated in regulation of gene expression that occurs predominantly within CpG dinucleotides [1–6]. CpG dinucleotides are under-represented in the mammalian genome (1%), but tend to cluster in CpG-rich regions called CpG islands (CGIs), located in the proximity of the transcription start sites (TSSs) of the majority (70%) of human protein-coding genes [7,8]. CGIs are stretches of DNA sequences of 200 nucleotides or greater [9], with the GC ratio observed/expected to be greater than 0.6. Although the bulk of genome is methylated at 70–80% of its CpGs, CGIs are mostly unmethylated in somatic cells [9,10].

DNA methylation is mediated by members of the DNA methyltransferase (DNMT) family that can covalently transfer a methyl group (CH_3) from the universal donor S-adenosyl-L-methionine (SAM) to the carbon 5-position of the cytosine ring. Conventionally, DNMTs are classified as de novo (DNMT3a and DNMT3b) or maintenance (DNMT1) enzymes [11]. Of note, DNMT2, a member of the methyltransferase family, catalyzes methylation of RNA at position 38 in tRNAAsp GUC [12–14] (Figure 2). DNMT3-like (DNMT3L) is a DNMT3-associated protein lacking an enzymatic domain and interacting with DNMT3a/3b to modulate its activity [15,16]. In mice, Dnmt1 and Dnmt3b are essential for embryonic development, as DNA methylation changes dramatically. An overall demethylation after fertilization is followed by de novo methylation of discrete regions upon implantation [17]. *Dnmt1* knockout mice show early lethality at embryonic day (E) 9.5, whereas Dnmt3b depletion induces death at E 14.5–18.5, due to developmental impairment. On the contrary, Dnmt3a knockout mice do not display defects in embryonic development, but they do die at 4 weeks of age. Although this binary

classification is convenient, the function of the de novo and maintenance DNMTs overlaps in many instances [18–20].

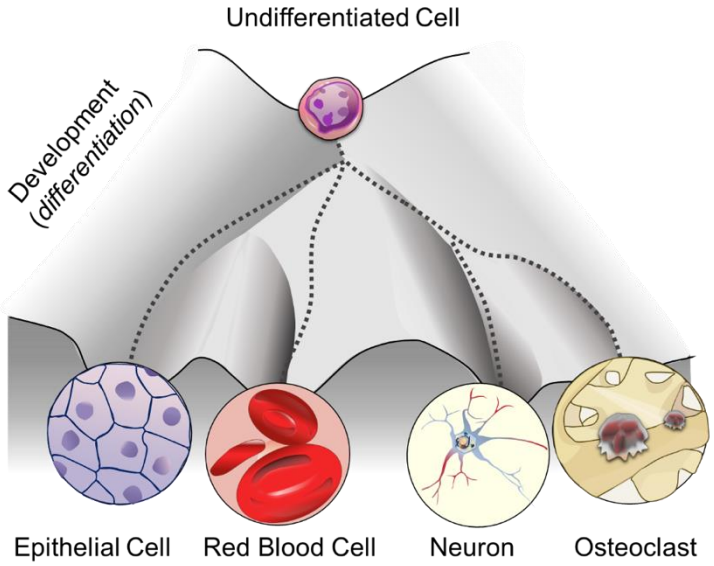

Figure 1. An outline depicting cell-fate plasticity according to the Waddington's epigenetic landscape (inspired by the model proposed by Waddington [3]).

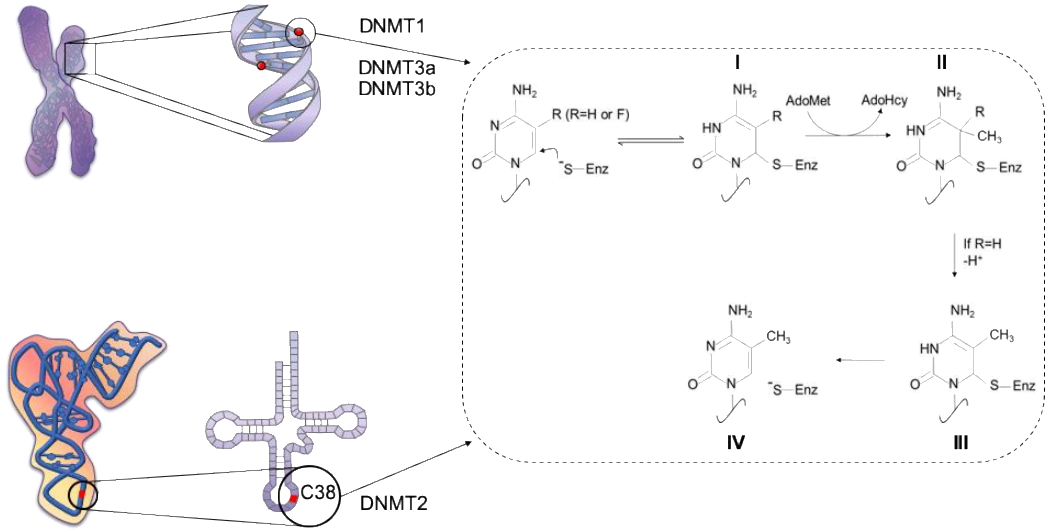

Figure 2. Schematic representation of the methylation reaction catalyzed by the DNA methyltransferases (DNMTs) (adapted from [14]). Shown are the mechanisms proposed for methylation of cytosine by DNMT1, 3a and 3b on DNA (upper left panel) or by DNMT2 on RNA (lower left panel). Briefly, a thiol group (SH) from the binding site of the enzyme provides the nucleophilic attack to position 6 of the cytosine heterocycle, to activate position 5 towards one-carbon transfer (I). The methyl group on position 5 is donated by the coenzyme AdoMet (II). A proton in position 5 of the 5,6-dihydropyrimidine is then removed (II–III), and a consequent β-elimination generate 5-methylcitosyne and free enzyme (IV).

Box 1. History of epigenetics.

The term epigenetics was originally coined in the 17th century by the physician and physiologist William Harvey to indicate the gradual development of the embryo from a homogeneous to a heterogeneous material, referred to as "epigenesis" [1]. Later on, in the 1940s, Conrad Waddington used the term "epigenetics" to explain the relationship between the genotype, defined as the whole genetic system of an organism, and the phenotype, indicating the entire set of characteristics that an organism develops over time [2]. Waddington established the first causal relationship between genes and their outcomes by introducing the concept of the "epigenetic landscape" as "the various developmental pathways that undifferentiated cells (sharing identical genotype) might take toward differentiation" (Figure 1) [3]. In other words, he described how the static information written in the form of nucleotide sequences is dynamically translated into tissues and organs, thus driving cell fate decisions [4]. In the last two decades, the definition of epigenetics has evolved from "the study of mitotically and/or meiotically heritable changes in gene function that cannot be explained by changes in DNA sequence" [5] to "the structural adaptation of chromosomal regions so as to register, signal or perpetuate altered activity states", which is inclusive of all the stable or transient chromosomal markers arising in response to different stimuli [6].

A number of studies have shown that DNA methylation is not randomly distributed across the genome, but displays regional specificity [21]. Methyl groups promote conformational changes in the major groove of DNA, thus altering protein-DNA binding [22] and, as a result, gene expression. Most studies have initially focused on the effect of CGI methylation within the promoter and TSS of protein-coding genes. Recently [23], however, more comprehensive genome-wide methylation analyses have started elucidating the role of DNA methylation at CpG clusters within exons, introns, and intergenic sequences, expanding on the previous knowledge of CGIs and leading to the identification of CG shores (regions up to 2 kb from CGI), shelves (regions from 2 to 4 kb from CGI), and open sea (the rest of the genome) regions [24–26].

Hence, DNA methylation needs to be framed in the context of the genomic location. As a proof of concept, methylation of TSS-associated CGIs negatively correlates with gene expression, leading to long-term gene silencing [7], whilst gene-body methylation positively correlates to gene expression [27–30]. Another interesting finding emerging from bulk methylome studies is that non-CGI-CpGs are mostly methylated and therefore less stable than CGIs, due to the tendency of 5-methylcytosine (5mC) to undergo spontaneous or enzymatic deamination to T [31]. The C-to-T transition causes germ-line or somatic mutations, resulting in the depletion of CpGs dinucleotides in the human genome. Methylation at other genomic regions, such as enhancers and insulators, does not follow a specific pattern and may vary in different settings. Enhancers and insulators are long-range regulatory elements able to alter gene expression or protect gene promoters from inappropriate signals, respectively [32,33]. Aran et al. have shown that distal methylation sites in estrogen receptor (ER)-positive breast tumors associates with breast cancer-related gene expression better than promoter methylation [34]. Moreover, Tatetsu et al. demonstrated that aberrant methylation of the 17-kb 5′ upstream enhancer of PU.1 is required for myeloma cell growth [35]. Insulators are bifunctional instead, acting either as a blocking enhancer, by preventing enhancer-mediated transcription, or as barriers, by limiting the advance of nearby heterochromatin that would otherwise silence expression [36]. CCCTC-binding factor (CTCF), an enhancer-blocking protein, does not bind to its DNA consensus sequence if methylated, as demonstrated for the imprinted *IGF2-H19* locus and the *CD45* gene [37]. It follows that DNA methylation can regulate gene expression indirectly by controlling access of enhancers to gene promoters [38]. Finally, CGI shores, with a lower CpG density, have recently emerged as critical regulatory elements affecting gene expression depending on their DNA methylation profile [39].

This review will address the impact of cutting-edge next-generation sequencing technologies on our perception and interpretation of DNA methylation in health conditions and diseases.

2. DNA Methylation Analysis: Think Globally, Act Locally

DNA methylation has a crucial role in various biological processes, such as development, differentiation, and gene expression [40–42]. Thus, comprehensive mapping becomes critical for addressing the functional role of this modification [43]. Various strategies have been developed to differentiate methylated and non-methylated C residues [44]. The initial lack of genome-wide approaches has restricted DNA methylation profiling to gene-specific evaluation using polymerase chain reaction (PCR) amplification of the target sequence. As DNA polymerases do not discriminate between C and 5mC, all potential differences in methylation are lost during classical PCR amplification [44,45]. To overcome this challenge, Frommer et al. developed a locus-specific method based on DNA treatment with sodium-bisulfite (SB) that leads to the conversion of all unmodified cytosine to uracil [45]. 5mC are resistant to deamination induced by SB, and are preserved during the PCR amplification by primers designed on the converted DNA. The resulting product is then analyzed by Sanger sequencing. Although very laborious, bisulfite sequencing PCR (BSP) is considered the gold standard to quantitatively study gene-locus specific DNA methylation [46].

The possibility to interrogate the entire genome using next generation sequencing technologies has expanded DNA methylation analyses [47]. The application of massive parallel sequencing to bisulfite-treated DNA has resulted in new genome-scale and -wide protocols, such as reduced representation bisulfite sequencing (RRBS) and whole genome bisulfite sequencing (WGBS), respectively. RRBS uses MspI, a methylation insensitive enzyme, to produce small fragments with CpG dinucleotides at both ends. Digested products are bisulfite converted and then sequenced [48]. Even though RRBS is more capable of covering a higher number of CpG loci within a given region than array-based techniques [47,49], the coverage across corresponding CpG-rich sequences might change among the samples tested, therefore introducing higher inter-sample variability and altering reproducibility [50]. As RRBS is biased for CpG-rich regions, such as CGIs, the coverage drops for CG shores, shelves, and open sea regions [50].

WGBS couples bisulfite-conversion of genomic DNA with high-throughput sequencing [51] and therefore provides a comprehensive and quantitative analysis of DNA methylome at single nucleotide resolution, without relying on restriction enzyme enrichment. Until few years ago, the high amount of DNA required as input, as well as the cost, limited its use. The recent advancement in sequencing platforms and sample preparation have now made WGBS more accessible and feasible, especially cost-wise, to allow projects with large sample sizes [52].

Likewise, non-bisulfite-based methods have also benefited from the introduction of high-throughput sequencing platforms. For example, immunoprecipitation of methylated DNA with an antibody recognizing 5mC residues coupled with sequencing (methylated DNA immunoprecipitation sequencing [MeDIP-Seq]) [47], is commonly used as an alternative approach to RRBS (Box 2).

MeDIP-Seq covers a higher number of regions than those normally screened by RRBS, but its efficiency is based on the specificity of the antibody, which can be biased toward hypermethylated sequences. Yet, neither RRBS nor MeDIP-Seq can determine the profile of virtually all CpG dinucleotides throughout the genome. Therefore, WGBS is considered the gold standard for full methylome analysis. More recently, the establishment of enzymatic methyl-Seq (EM-Seq), has added another free-bisulfite-based method to profile the entire methylome in instances when BS treatment is not suitable (i.e., fragmented DNA) or the DNA input is too low. EM-Seq relies on two sequential enzymatic reactions protecting 5mC and 5hmC from downstream deamination. The enzymatically converted DNA can then be processed for sequencing similar to WGBS [53].

Despite their sensitivity, all bulk sequencing methods are unable to dissect intra-cellular and intra-tumoral epigenetic heterogeneity within a specific cell population [54]. Single-cell technology has emerged as an invaluable tool to ascertain this heterogeneity [55]. WGBS and RRBS protocols have been optimized to carry out single-cell bisulfite sequencing (scBS) and single-cell RRBS (scRRBS), respectively [56]. While scBS is able to cover a higher number of CpG sites than scRRBS, the latter enables a better coverage of CGIs [57] (Figure 3).

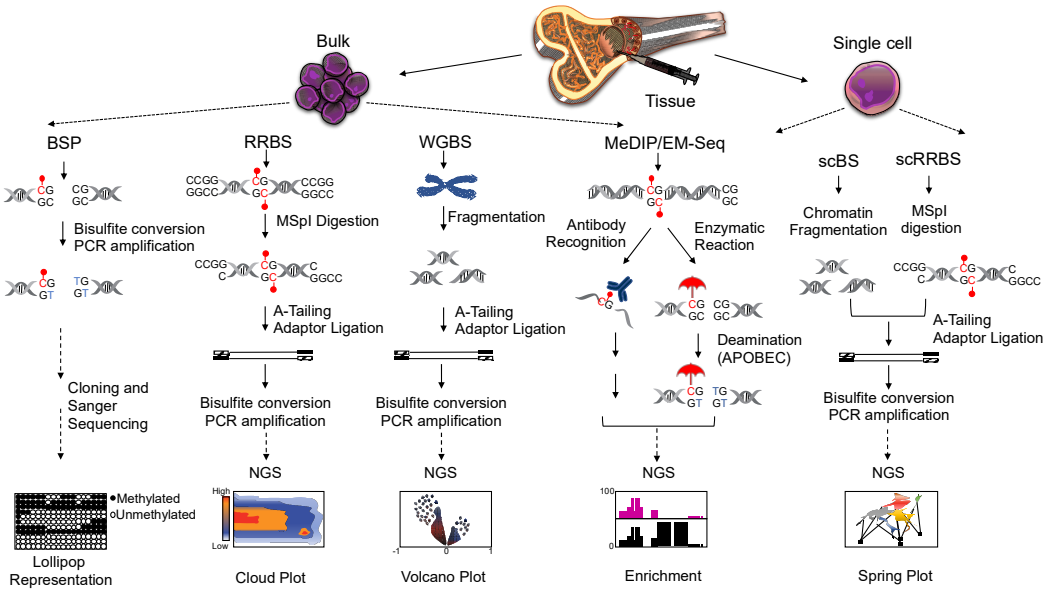

Figure 3. Workflows of bisulfite and non-bisulfite-based methods applied to either bulk- or single—cell population for DNA methylation analyses. On the left hand-side, bulk-bisulfite-based methods: bisulfite sequencing PCR (BSP), reduced representation bisulfite sequencing (RRBS), and whole genome bisulfite sequencing (WGBS). Following bulk-non-bisulfite-based methods: methylated DNA immunoprecipitation sequencing (MeDIP-Seq) and enzymatic methyl sequencing (EM-Seq), in which an enzymatic reaction protects [☂] mC and/or hmC from the deamination by APOBEC. On the right hand-side, single-cell bisulfite-based methods as indicated: single-cell bisulfite sequencing (scBS) and single-cell reduced representation bisulfite sequencing (scRRBS). The red dot (●) indicates methylated cytosine.

Notably, these new approaches are shedding light on the impact of epigenetic alterations in specific cellular subsets with respect to developmental processes and cancer, filling the gaps arising from bulk studies [58].

Box 2. More about sodium bisulfite treatment.

Although bisulfite sequencing approaches are considered the gold standard for DNA methylation analysis, all bisulfite-based methods share common limitations. Sodium bisulfite is a harsh chemical treatment that degrades template DNA, thus producing a poor quality product. An incomplete conversion of cytosine to uracil may also introduce artifacts into the analysis [59]. In addition, bisulfite-based methods are not able to discriminate between 5-methylcitocine (5mC) and its oxidative derivative 5-hydroxymethylcytosine (5hmC), produced by the ten-eleven translocation (TET) family of dioxygenases during active DNA demethylation. Other than being a chemical intermediate, 5hmC is a stable DNA modification that binds to specific regulatory proteins and is mainly found within actively transcribed genes. 5hmC marks a decrease in cancer tissues, suggesting its potential regulatory role in the mammalian genome [60]. To overcome the limitations associated with bisulfite-based approaches, new bisulfite-free techniques [47,59] have been developed that preserve genomic DNA integrity, as well as other strategies able to distinguish 5mC and 5hmC marks [61,62].

3. DNA Methylation in Health: A Matter of Location and Timing

DNA methylation plays a critical role in the regulation of early development in humans and other mammals. In contrast to DNA sequences, DNA methylation is not inherited from gametes, as the parental DNA methylation pattern is erased at an early embryo stage [63]. During implantation, the DNA methylation profile is re-established, and the entire genome undergoes de novo methylation with the exclusion of CpG island-like regions, which elude this epigenetic modification due to the presence of RNA polymerase complexes that prevent access of de novo methylation machinery to the DNA. The resulting bimodal pattern is conserved throughout development and preserved for the

whole lifespan of the organism unless unexpected alterations [64,65]. Yet, a group of genes, so-called "imprinted" genes, can escape this extensive reprogramming process occurring during embryonic development. Genomic imprinting is defined as the monoallelic, parental-specific expression of a gene in diploid cells, and as such, it is considered a form of gene regulation. Imprinted genes are epigenetically marked, i.e., "imprinted", within differentially methylated regions (DMRs) in gametes, and such imprints are conserved after fertilization [66]. As a result, only one parental copy of the imprinted gene will be expressed, whereas the other copy will be silenced by DNA methylation [67–69]. The human insulin-like growth factor II (*IGF-II*)/*H19* region is an example of paternally imprinted gene, in which methylation of the imprinting control sequence (ICR) regulates binding of the CTCF [70,71]. In mammals, the majority of imprinted genes affects growth of the embryo, placenta, and neonate. In that regard, paternally imprinted genes function as growth enhancers, whereas maternally imprinted genes function as growth repressors [72]. The loss of maternal DNA methyltransferases results in post-implantation lethality [16,72], suggesting the essential role of genomic imprinting in embryonic development and differentiation. Another category of imprinted genes include those involved in neurologic and behavioral regulation [73].

DNA methylation also partakes in protecting the structural integrity of the genome. It has been proposed that methylation of CpGs within parasitic DNA elements and retrotransposons, which account for 40% of the entire genome, operates as a genome defense system, in order to prevent the expression of these elements and preserve genomic stability [74].

Another puzzling and intriguing matter is whether the dynamic changes observed in response to biological and non-biological stimuli are paralleled by DNA methylation changes during the different phases of the cell cycle in physiological conditions. Substantial data with respect to this question are still lacking, and the few reported studies do not seem to agree with one another. Single-cell RNA sequencing results from murine embryonic stem cells (mESCs) point to a higher methylation rate in G1/S compared to other phases of the cycle, in line with previous observations in HeLa cells and human primary foreskin fibroblasts [75,76]. Similarly, Desjorbet et al. have demonstrated that the inheritance of DNA methylation marks efficiently occurs in late S phase, and cytosine methylation is completed in G2/M phase [77]. However, a different study analyzing DNA methylation levels in G0, G1, and G2 phases in low passage primary dermal fibroblasts did not confirm these cell cycle changes in DNA methylation patterns [78]. The specific cell type interrogated, along with not including the S phase in the analysis, could possibly account for these contrasting results, since hemi-methylated DNA sequences are challenging to evaluate with low coverage sequencing. Further analyses will need to delve into this question.

4. DNA Methylation and Disease: Too Little, too Much, or Both?

Given its critical role in many biological processes, it is not surprising that dysregulation of DNA methylation is frequently linked to either germ-line or somatic diseases [79,80]. The study of monozygotic (MZ) twins has offered an exceptional tool to investigate the contribution of the epigenetic load to phenotypic variations, including predisposition to pathological conditions, in individuals with identical genetic background [81–84]. MZ twin studies have shown in many instances that phenotypic variations can be ascribed to differences in DNA methylation [84]. Following up, studies from the recent NASA research on MZ twins demonstrated that global DNA methylation was altered in response to different environmental stimuli, namely the absence of gravity for the one subject who was sent into space, versus the gravity the sibling who remained on earth was exposed to, affecting specific pathways transiently or permanently [85] (Box 3).

Box 3. Environmental factors "another variation on the theme".

Major clues have been brought about by the evidence that the environment and environmental factors can profoundly affect DNA methylation. To list few examples, prenatal or chronic exposure to chemicals, including air pollution and tobacco smoke, or drugs associated with either global or site-specific DNA methylation alterations [86,87]. Nutritional elements, also considered under environmental factors, can impact developmental programming and result in later-life health outcomes [86,87]. Furthermore, prenatal and early life social conditions might lead to DNA methylation alterations involving immune functions and inflammatory pathways. Maternal depression in the prenatal period can affect DNA methylation of the brain-derived neurotrophic factor (BDNF) and maternal anxiety can induce loss of methylation in the *IGF2/H19* locus [71].

4.1. Germ-Line Associated Diseases

Epimutations within DMRs of imprinting genes are responsible for up to 50% of imprinting diseases (IDs), leading to alteration in the imprinted gene dosage. Defects in DNA methylation, as hyper- or hypomethylation, mainly localize within specific DMRs of the imprinted gene, and epigenetic mutations at further DMRs can account for the severity and outcome of the pathology [88]. Examples of IDs are Silver-Russell syndrome, displaying hypomethylation of the paternally imprinted locus *H19/IGF2* in the 11p15 region, together with other alterations at chromosome 7 and 11 [89–92], Fragile X syndrome, most commonly caused by a trinucleotide repeat expansion "CGG", within the promoter of the fragile X mental retardation 1 protein (FMR1), which leads to aberrant DNA methylation and gene silencing [93–95], and Angelman syndrome, with a frequency of 2–3%, caused by an imprinting defect in the expression of maternally expressed 15q11-q13 genes. Additional examples of IDs are listed in Table 1.

Table 1. List of the best-known imprinting diseases and associated epigenetic lesions.

Imprinting Diseases	Epigenetic Lesions	Reference
Transient Neonatal Diabetes Mellitus Type 1 (TNDM1)	Hypomethylation of the maternally imprinted genes *PLAGL1* and *HYMAI*	[96,97]
Silver-Russell Syndrome	Hypomethylation of the paternally imprinted locus *H19/IGF2* and promoter hypomethylation of *HOXA4*	[91,92]
Beckwith-Wiedemann	Imprinting defects within two imprinted domains, *IGF2/H19* and *CDKN1C/KCNQ1OT1*	[98,99]
Fragile X Syndrome	De novo methylation of the *FMR1* gene	[93,95]
Angelman Syndrome	Imprinting defects within chromosome 15q11-q13 that alter the expression of the maternally inherited *UBE3A*	[100,101]
Prader-Willi Syndrome	Loss of expression of the paternally inherited chromosome 15q11.2-q13 due to imprinting defects	[102,103]
Pseudohypoparathyroidism	Epigenetic defects in the imprinted *GNAS* cluster on chromosome 20q13.3	[104,105]

4.2. Somatic Diseases

According to the "two-hits" model of cancer proposed by Knudson in 1971, a dominantly inherited germ-line mutation (first hit) and a second somatic mutation (second hit) predisposes and causes tumorigenesis, respectively [106]. However, evidence later proved that methylation-associated gene silencing can act as cancer-predisposing event (first hit) as effectively as genetic mutation [107–109]. Two distinct alterations of normal DNA methylation patterns occur in cancer: global hypo-methylation and gene-specific hypermethylation [110–113]. Genome-wide analyses evidenced that only 40 to 60% of a cancer cell genome is methylated, versus 80% of methylated genome in healthy controls. This global demethylation in cancer cells contributes to genomic instability, aneuploidy, an increased mutation rate, and could result from deregulation of passive and active demethylation processes mediated by DNMTs and TET family proteins, respectively [114–116]. Conversely, aberrant methylation of CGIs within 5′ regions of cancer-related genes is a hallmark of nearly all tumors and correlates with changes of chromatin structure that lead to silencing of tumor suppressor genes (TSGs) [117,118]. Colorectal cancer

(CRC) and myelodysplastic syndrome (MDS) offer two good examples for understanding how aberrant loss or gain of DNA methylation may contribute to tumorigenesis, but they are not the only ones [119,120].

4.2.1. Colorectal Cancer

CRC results from the accumulation of genetic and epigenetic alterations, both leading to genomic instability. Aberrant DNA methylation is considered a potential driver in CRC and understanding its establishment might contribute to developing new therapeutic strategies in the treatment of the tumor [121–123]. At the genome level, CRC is characterized by loss in DNA methylation (10–40% lower levels of absolute methylation than normal colonic tissue), mainly occurring within repetitive sequences and resulting in genomic instability and potential initiation of CRC [124,125]. In contrast, hypermethylation has been observed in regions corresponding to CGIs, histone H3 trimethylated on lysin 4 (H3K4me3), and open chromatin in normal controls [126]. A subgroup of CRCs, termed CpG-island methylator phenotype (CIMP), is characterized by aberrant promoter DNA methylation of critical genes involved in the WNT, P53, and RAS signaling pathways, although the role of component of these pathways, as initiators or drivers in the progression of the disease remains elusive [124,125].

4.2.2. Myelodysplastic Syndrome

MDS is a clonal hematologic disorder often leading to acute myeloid leukemia (AML) in approximately 30% of cases. In contrast, with the common finding of global hypomethylation in cancer, MDS is characterized by a global hypermethylation associated with poor prognosis [127,128]. Aberrant promoter methylation of transcription factors, such as the CCAAT/enhancer-binding protein alpha (*CEBPA*), or TSG as *CDKN2B* (*P15*) and *CDKN2A* (*P16*), the adhesion molecule e-cadherin (*CDH1*) and the estrogen receptor (*ER*) [129] have been reported in either MDS and/or AML [129–131]. MDS displays extensive epigenetic reprogramming that explains why the number of hypermethylated genes is greater than in de novo AML or normal CD34+ cells. Hence, DNA methylation aberrations are linked to MDS progression, along with epigenetic regulators including *TET2*, *DNMT3a*, *IDH1*, *ASXL1*, and *EZH2* which are among the group of genes most frequently mutated in MDS [132,133]. In summary, genetic and epigenetic alterations often coexist, leading to distinct DNA methylation signatures in cancer cell genomes [134]. In terms of MDS treatment, the heterogenous nature of the disease warrants a complex combined therapy. Thus far, the DNMT inhibitors (DNMTi) 5-azacytidine (AZA) and decitabine (DAC), are the most successful epigenetic drugs approved for MDS and AML treatment. DNMTi modulate the epigenome of cancer cells by reversing aberrant DNA methylation patterns and re-establishing transcription of epigenetically silenced genes [135,136]. Additionally, a recent study performed in different AML cell lines has demonstrated that AZA is able to induce down-regulation of cell metabolism and up-regulation of immune defense-related genes. However, despite their efficacy, the severe side-effects of DNMTi, together with no or partial responses to treatment, do limit their clinical application [137].

5. Conclusions: The Best Is yet to Come

Epigenetics, defined as the interplay of DNA methylation, histone modifications, and expression of non-coding RNAs, govern numerous biological mechanisms from early development to old age, including biological response to endogenous or exogenous stimuli contributing to disease transition. In this review, we focused on DNA methylation, as alterations of this epigenetic mark are considered a triggering event in the pathogenesis of several human diseases affecting either germ-line imprinting disorders or somatic malignant transformation. The identification and characterization of different DNMT family members has led to a better understanding of mechanisms behind the establishment of DNA methylation in health and disease, and to the discovery of new therapeutic targets. DNMTs inhibitors have indeed been approved for the treatment of MDS and AML, while they

are being tested in combination with other anticancer drugs, as a therapeutic approach for multiple solid cancers such as colon, ovarian, and lung cancer [138]. Unfortunately, the lack of specificity and the high cytotoxicity of the currently approved demethylating drugs hamper their clinical application.

The study of DNA methylation at single-cell level resolution is providing deeper insights into the epigenetic events driving disease onset and progression, and will likely shed light on other unknown features of DNA methylation. By means of new advanced technologies, such as WGBS and scBS, novel molecular targets will be identified and, more specific and less toxic therapeutic molecules might be developed with the ultimate goal of overcoming the downside effects of conventional hypomethylating protocols.

Funding: Our studies were supported by the National Cancer Institute of the National Institutes of Health under Award Number R00CA188595 (the content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health); the Fondazione Cariplo N. 2016-0476; the Giovanni Armenise-Harvard Foundation Career Development Award and the Italian Association for Cancer Research (AIRC) Start-up Grant N.15347 to A.D.R. M.B. was supported by a PhD Fellowship from the Italian Ministry of Education, Universities, and Research to the University of Eastern Piedmont.

Acknowledgments: We thank Laura Buffa and Daniel G. Tenen for reading the manuscript.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Deichmann, U. Epigenetics: The origins and evolution of a fashionable topic. *Dev. Biol.* **2016**, *416*, 249–254. [CrossRef] [PubMed]
2. Nicoglou, A.; Merlin, F. Epigenetics: A way to bridge the gap between biological fi elds. *Stud. Hist. Philos. Biol. Biomed. Sci.* **2017**, *66*, 73–82. [CrossRef] [PubMed]
3. Goldberg, A.D.; Allis, C.D.; Bernstein, E. Epigenetics: A Landscape Takes Shape. *Cell* **2007**, *128*, 635–638. [CrossRef] [PubMed]
4. Moris, N.; Pina, C.; Arias, A.M. Transition states and cell fate decisions in epigenetic landscapes. *Nat. Rev. Genet.* **2016**, *17*, 693–703. [CrossRef] [PubMed]
5. Russo, V.E.A.; Martienssen, R.A.; Riggs, A.D. *Epigenetic Mechanisms of Gene Regulation*; CSHL Press: Cold Spring Harbor, NY, USA, 1996.
6. Bird, A. Perceptions of epigenetics. *Nature* **2007**, *447*, 396–398. [CrossRef] [PubMed]
7. Jones, P.A.; Takai, D. The Role of DNA Methylation in Mammalian Epigenetics. *Science* **2001**, *293*, 1068–1071. [CrossRef]
8. Illingworth, R.S.; Bird, A.P. CpG islands—'A rough guide'. *FEBS Lett.* **2009**, *583*, 1713–1720. [CrossRef]
9. Fazzari, M.J.; Greally, J.M. Epigenomics: Beyond CpG islands. *Nat. Rev. Genet.* **2004**, *5*, 446–455. [CrossRef]
10. Saxonov, S.; Berg, P.; Brutlag, D.L. A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters. *Proc. Natl. Acad. Sci. USA* **2006**, *103*, 1412–1417. [CrossRef]
11. Klose, R.J.; Bird, A.P. Genomic DNA methylation: The mark and its mediators. *Trends Biochem. Sci.* **2006**, *31*, 89–97. [CrossRef]
12. Jeltsch, A.; Ehrenhofer-murray, A.; Jurkowski, T.P.; Lyko, F.; Reuter, G.; Ankri, S. Mechanism and biological role of Dnmt2 in Nucleic Acid Methylation. *RNA Biol.* **2017**, *14*, 1108–1123. [CrossRef] [PubMed]
13. Goll, M.G.; Kirpekar, F.; Maggert, K.A.; Yoder, J.A.; Hsieh, C.L.; Zhang, X.; Golic, K.G.; Jacobsen, S.E.; Bestor, T.H. Methylation of tRNAAsp by the DNA methyltransferase homolog Dnmt2. *Science* **2006**, *311*, 395–398. [CrossRef] [PubMed]
14. Liu, Y.; Santi, D.V. m5C RNA and m5C DNA methyl transferases use different cysteine residues as catalysts. *Proc. Natl. Acad. Sci. USA* **2000**, *97*, 8263–8265. [CrossRef] [PubMed]
15. Chedin, F.; Lieber, M.R.; Hsieh, C.L. The DNA methyltransferase-like protein DNMT3L stimulates de novo methylation by Dnmt3a. *Proc. Natl. Acad. Sci. USA* **2002**, *99*, 16916–16921. [CrossRef] [PubMed]
16. Neri, F.; Krepelova, A.; Incarnato, D.; Maldotti, M.; Parlato, C.; Galvagni, F.; Matarese, F.; Stunnenberg, H.G.; Oliviero, S. Dnmt3L antagonizes DNA methylation at bivalent promoters and favors DNA methylation at gene bodies in ESCs. *Cell* **2013**, *155*, 121–134. [CrossRef]

17. Robertson, K.D. DNA methylation and chromatin—Unraveling the tangled web. *Oncogene* **2002**, *21*, 5361–5379. [CrossRef]
18. Yarychkivska, O.; Shahabuddin, Z.; Comfort, N.; Boulard, M.; Bestor, T.H. BAH domains and a histone-like motif in DNA methyltransferase 1 (DNMT1) regulate de novo and maintenance methylation in vivo. *J. Biol. Chem.* **2018**, *293*, 19466–19475. [CrossRef]
19. Jones, P.A.; Liang, G. Rethinking how DNA methylation patterns are maintained. *Nat. Rev. Genet.* **2009**, *10*, 805–811. [CrossRef]
20. Elliott, E.N.; Sheaffer, K.L.; Kaestner, K.H. The 'de novo' DNA methyltransferase Dnmt3b compensates the Dnmt1-deficient intestinal epithelium. *Elife* **2016**, *5*, e12975. [CrossRef]
21. Song, Y.; van den Berg, P.R.; Markoulaki, S.; Soldner, F.; Dall'Agnese, A.; Henninger, J.E.; Drotar, J.; Rosenau, N.; Cohen, M.A.; Young, R.A.; et al. Dynamic Enhancer DNA Methylation as Basis for Transcriptional and Cellular Heterogeneity of ESCs. *Mol. Cell* **2019**, *75*, 905–920. [CrossRef]
22. Liebl, K.; Zacharias, M. How methyl-sugar interactions determine DNA structure and flexibility. *Nucleic Acids Res.* **2019**, *47*, 1132–1140. [CrossRef] [PubMed]
23. Skvortsova, K.; Masle-Farquhar, E.; Luu, P.L.; Song, J.Z.; Qu, W.; Zotenko, E.; Gould, C.M.; Du, Q.; Peters, T.J.; Colino-Sanguino, Y.; et al. DNA Hypermethylation Encroachment at CpG Island Borders in Cancer Is Predisposed by H3K4 Monomethylation Patterns. *Cancer Cell* **2019**, *35*, 297–314. [CrossRef] [PubMed]
24. Ollikainen, M.; Ismail, K.; Gervin, K.; Kyllonen, A.; Hakkarainen, A.; Lundbom, J.; Jarvinen, E.A.; Harris, J.R.; Lundbom, N.; Rissanen, A.; et al. Genome-wide blood DNA methylation alterations at regulatory elements and heterochromatic regions in monozygotic twins discordant for obesity and liver fat. *Clin. Epigenetics* **2015**, *7*, 39. [PubMed]
25. Visone, R.; Bacalini, M.G.; Di Franco, S.; Ferracin, M.; Colorito, M.L.; Pagotto, S.; Laprovitera, N.; Licastro, D.; Di Marco, M.; Scavo, E.; et al. DNA methylation of shelf, shore and open sea CpG positions distinguish high microsatellite instability from low or stable microsatellite status colon cancer stem cells. *Epigenomics* **2019**, *11*, 587–604.
26. Irizarry, R.A.; Ladd-Acosta, C.; Wen, B.; Wu, Z.; Montano, C.; Onyango, P.; Cui, H.; Gabo, K.; Rongione, M.; Webster, M.; et al. The human colon cancer methylome shows similar hypo- and hypermethylation at conserved tissue-specific CpG island shores. *Nat. Genet.* **2009**, *41*, 178–186.
27. Hellman, A.; Chess, A. Gene body-specific methylation on the active X chromosome. *Science* **2007**, *315*, 1141–1143. [CrossRef]
28. Teissandier, A.; Bourc'his, D. Gene body DNA methylation conspires with H3K36me3 to preclude aberrant transcription. *EMBO J.* **2017**, *36*, 1471–1473. [CrossRef]
29. Arechederra, M.; Daian, F.; Yim, A.; Bazai, S.K.; Richelme, S.; Dono, R.; Saurin, A.J.; Habermann, B.H.; Maina, F. Hypermethylation of gene body CpG islands predicts high dosage of functional oncogenes in liver cancer. *Nat. Commun.* **2018**, *9*, 3164. [CrossRef]
30. Sartor, R.C.; Noshay, J.; Springer, N.M.; Briggs, S.P. Identification of the expressome by machine learning on omics data. *Proc. Natl. Acad. Sci. USA* **2019**, *116*, 18119–18125. [CrossRef]
31. Sleutels, F.; Barlow, D.P. The Origins of Genomic. *Adv. Genet.* **2002**, *46*, 119–163.
32. Gaszner, M.; Felsenfeld, G. Insulators: Exploiting transcriptional and epigenetic mechanisms. *Nat. Rev. Genet.* **2006**, *7*, 703–713. [CrossRef] [PubMed]
33. Alasoo, K.; Rodrigues, J.; Mukhopadhyay, S.; Knights, A.J.; Mann, A.L.; Kundu, K.; Hale, C.; Dougan, G.; Gaffney, D.J. Shared genetic effects on chromatin and gene expression indicate a role for enhancer priming in immune response. *Nat. Genet.* **2018**, *50*, 424–431. [CrossRef] [PubMed]
34. Aran, D.; Hellman, A. DNA Methylation of Transcriptional Enhancers and Cancer Predisposition. *Cell* **2013**, 11–13. [CrossRef] [PubMed]
35. Tatetsu, H.; Ueno, S.; Hata, H.; Yamada, Y.; Takeya, M.; Mitsuya, H.; Tenen, D.G.; Okuno, Y. Down-regulation of PU.1 by methylation of distal regulatory elements and the promoter is required for myeloma cell growth. *Cancer Res.* **2007**, *67*, 5328–5336. [CrossRef]
36. West, A.G.; Gaszner, M.; Felsenfeld, G. Insulators: Many functions, many mechanisms. *Genes Dev.* **2002**, *16*, 271–288. [CrossRef]
37. Wang, H.; Maurano, M.T.; Qu, H.; Varley, K.E.; Gertz, J.; Pauli, F.; Lee, K.; Canfield, T.; Weaver, M.; Sandstrom, R.; et al. Widespread plasticity in CTCF occupancy linked to DNA methylation. *Genome Res.* **2012**, *22*, 1680–1688. [CrossRef]

38. Bell, A.C.; Felsenfeld, G. Methylation of a CTCF-dependent boundary controls imprinted expression of the Igf2 gene. *Nature* **2000**, *405*, 2–5. [CrossRef]
39. Rao, X.; Evans, J.; Chae, H.; Pilrose, J.; Kim, S.; Yan, P.; Huang, R.l.; Lai, H.c.; Lin, H.; Liu, Y.; et al. CpG island shore methylation regulates caveolin-1 expression in breast cancer. *Oncogene* **2013**, 4519–4528. [CrossRef]
40. Bestor, T.H.; Edwards, J.R.; Boulard, M. Notes on the role of dynamic DNA methylation in mammalian development. *Proc. Natl. Acad. Sci. USA* **2015**, *112*, 6796–6799. [CrossRef]
41. Schubeler, D. Function and information content of DNA methylation. *Nature* **2015**, *517*, 321–326. [CrossRef]
42. Cholewa-Waclaw, J.; Shah, R.; Webb, S.; Chhatbar, K.; Ramsahoye, B.; Pusch, O.; Yu, M.; Greulich, P.; Waclaw, B.; Bird, A.P. Quantitative modelling predicts the impact of DNA methylation on RNA polymerase II traffic. *Proc. Natl. Acad. Sci. USA* **2019**, *116*, 14995–15000. [CrossRef] [PubMed]
43. Taiwo, O.; Wilson, G.A.; Morris, T.; Seisenberger, S.; Reik, W.; Pearce, D.; Beck, S.; Butcher, L.M. Methylome analysis using MeDIP-seq with low DNA concentrations. *Nat. Protoc.* **2012**, *7*, 617–636. [CrossRef] [PubMed]
44. Dahl, C.; Guldberg, P. DNA methylation analysis techniques. *Biogerontology* **2003**, *4*, 233–250. [CrossRef] [PubMed]
45. Frommer, M.; McDonald, L.E.; Millar, D.S.; Collist, C.M.; Wattt, F.; Griggt, G.W.; Molloyt, P.L.; Paul, C.L. A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual DNA strands. *Proc. Natl. Acad. Sci. USA* **1992**, *89*, 1827–1831. [CrossRef]
46. Zhang, Y.; Rohde, C.; Tierling, S.; Stamerjohanns, H.; Reinhardt, R.; Walter, J.; Jeltsch, A. DNA methylation analysis by bisulfite conversion, cloning, and sequencing of individual clones. *Methods Mol. Biol.* **2009**, *507*, 177–187.
47. Harris, R.A.; Wang, T.; Coarfa, C.; Nagarajan, R.P.; Hong, C.; Downey, S.L.; Johnson, B.E.; Fouse, S.D.; Delaney, A.; Zhao, Y.; et al. Analysis Comparison of sequencing-based methods to profile DNA methylation and identification of monoallelic epigenetic modifications. *Nat. Biotechnol.* **2010**, *28*, 1097–1105. [CrossRef]
48. Gu, H.; Smith, Z.D.; Bock, C.; Boyle, P.; Gnirke, A.; Meissner, A. Preparation of reduced representation bisulfite sequencing libraries for genome-scale DNA methylation profiling. *Nat. Protoc.* **2011**, *6*, 468–481. [CrossRef]
49. Sant, K.E.; Nahar, M.S.; Dolinoy, D.C. DNA methylation screening and analysis. *Methods Mol. Biol.* **2012**, *889*, 385–406.
50. Carmona, J.J., Jr.; Accomando, W.P.; Binder, A.M.; Hutchinson, J.N.; Pantano, L.; Izzi, B.; Just, A.C.; Lin, X.; Schwartz, J.; Vokonas, P.S.; et al. Empirical comparison of reduced representation bisul fi te sequencing and In fi nium BeadChip reproducibility and coverage of DNA methylation in humans. *NPJ Genom. Med.* **2017**, *2*, 13. [CrossRef]
51. Li, N.; Ye, M.; Li, Y.; Yan, Z.; Butcher, L.M.; Sun, J.; Han, X.; Chen, Q.; Wang, J. Whole genome DNA methylation analysis based on high throughput sequencing technology. *Methods* **2010**, *52*, 203–212. [CrossRef]
52. Suzuki, M.; Liao, W.; Wos, F.; Johnston, A.D.; Degrazia, J.; Ishii, J.; Bloom, T.; Zody, M.C.; Germer, S.; Greally, J.M. Whole-genome bisulfite sequencing with improved accuracy and cost. *Genome Res.* **2018**, *28*, 1364–1371. [CrossRef] [PubMed]
53. Williams, L.; Bei, Y.; Church, H.E.; Dai, N.; Dimalanta, E.T.; Ettwiller, L.M.; Evans, T.C., Jr.; Langhorst, B.W.; Borgaro, J.G.; Guan, S.; et al. *Enzymatic Methyl-Seq: The Next Generation of Methylome Analysis*. 2019. Available online: https://www.neb.com/tools-and-resources/feature-articles/enzymatic-methyl-seq-the-next-generation-of-methylome-analysis (accessed on 15 September 2019).
54. Clark, S.J.; Smallwood, S.A.; Lee, H.J.; 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**, *12*, 534–547. [CrossRef] [PubMed]
55. Smallwood, S.A.; Lee, H.J.; Angermueller, C.; Krueger, F.; Saadeh, H.; Peat, J.; Andrews, S.R.; Stegle, O.; Reik, W.; Kelsey, G. Single-cell genome-wide bisulfite sequencing for assessing epigenetic heterogeneity. *Nat. Methods* **2014**, *11*, 6–11. [CrossRef] [PubMed]
56. Karemaker, I.D.; Vermeulen, M. Single-Cell DNA Methylation Profiling: Technologies and Biological Applications. *Trends Biotechnol.* **2018**, *36*, 952–965. [CrossRef]
57. Guo, H.; Zhu, P.; Guo, F.; Li, X.; Wu, X.; Fan, X.; Wen, L.; Tang, F. Profiling DNA methylome landscapes of mammalian cells with single-cell reduced-representation bisulfite sequencing. *Nat. Protoc.* **2015**, *10*, 645–659. [CrossRef]

58. Gaiti, F.; Chaligne, R.; Gu, H.; Brand, R.M.; Kothen-hill, S.; Schulman, R.C.; Grigorev, K.; Risso, D.; Kim, K.-T.; Pastore, A.; et al. Epigenetic evolution and lineage histories of chronic lymphocytic leukaemia. *Nature* **2019**, 576–599. [CrossRef]
59. Liu, Y.; Siejka-Zielińska, P.; Velikova, G.; Bi, Y.; Yuan, F.; Tomkova, M.; Bai, C.; Chen, L.; Schuster-Böckler, B.; Song, C.-X. Bisulfite-free direct detection of 5-methylcytosine and 5-hydroxymethylcytosine at base resolution. *Nat. Biotechnol.* **2019**, *37*, 424–429. [CrossRef]
60. Bachman, M.; Uribe-Lewis, S.; Yang, X.; Williams, M.; Murrell, A. 5-Hydroxymethylcytosine is a predominantly stable DNA modification. *Nat. Chem.* **2014**, *6*, 1049–1055.
61. Skvortsova, K.; Zotenko, E.; Luu, P.-L.; Gould, C.M.; Nair, S.S.; Clark, S.J.; Stirzaker, C. Comprehensive evaluation of genome-wide 5-hydroxymethylcytosine profiling approaches in human DNA. *Epigenetics Chromatin* **2017**, *10*, 1–20. [CrossRef]
62. Song, C.-X.; Diao, J.; Brunger, A.T.; Quake, S.R. Simultaneous single-molecule epigenetic imaging of DNA methylation and hydroxymethylation. *PNAS* **2016**, 3–8. [CrossRef]
63. Bergman, Y.; Cedar, H. DNA methylation dynamics in health and disease. *Nat. Struct. Mol. Biol.* **2013**, *20*, 274. [CrossRef] [PubMed]
64. Greenfield, R.; Tabib, A.; Keshet, I.; Moss, J.; Sabag, O.; Goren, A.; Cedar, H. Role of transcription complexes in the formation of the basal methylation pattern in early development. *Proc. Natl. Acad. Sci. USA* **2018**. [CrossRef] [PubMed]
65. Dor, P.Y.; Cedar, P.H. Review Principles of DNA methylation and their implications for biology and medicine. *Lancet* **2018**, *392*, 777–786. [CrossRef]
66. Gigante, S.; Gouil, Q.; Lucattini, A.; Keniry, A.; Beck, T.; Tinning, M.; Gordon, L.; Woodruff, C.; Speed, T.P.; Blewitt, E.; et al. Using long-read sequencing to detect imprinted DNA methylation. *Nucleic Acids Res.* **2019**, *47*, e46. [CrossRef]
67. Bourc'his, D.; Xu, G.-L.; Lin, C.-S.; Bollman, B.; Bestor, T.H. Dnmt3L and the Establishment of Maternal Genomic Imprints. *Science* **2001**, *294*, 2536–2540. [CrossRef]
68. Ferguson-smith, A.C. Genomic imprinting: The emergence of an epigenetic paradigm. *Nat. Rev. Genet.* **2011**, *12*. [CrossRef]
69. Barlow, D.P.; Bartolomei, M.S. Genomic Imprinting in Mammals. *Cold Spring Harb. Perspect. Biol.* **2014**, *31*, 493–525. [CrossRef]
70. Giannoukakis, N.; Deal, C.; Paquette, J.; Goodyer, C.G.; Polychronakos, C. Parental genomic imprinting of the. *Nat. Genet.* **1993**, *4*, 98–101. [CrossRef]
71. Park, K.-S.; Mitra, A.; Rahat, B.; Kim, K.; Pfeifer, K. Loss of imprinting mutations define both distinct and overlapping roles for misexpression of IGF2 and of H19 lncRNA. *Nucleic Acids Res.* **2017**, *45*, 12766–12779. [CrossRef]
72. Ishida, M.; Moore, G.E. The role of imprinted genes in humans. *Mol. Asp. Med.* **2013**, *34*, 826–840. [CrossRef]
73. Xu, Q.; Xiang, Y.; Wang, Q.; Wang, L.; Brind'Amour, J.; Bogutz, A.B.; Zhang, Y.; Zhang, B.; Yu, G.; Xia, W.; et al. SETD2 regulates the maternal epigenome, genomic imprinting and embryonic development. *Nat. Genet.* **2019**, *51*, 844–856. [CrossRef] [PubMed]
74. Robertson, K.D.; Wolffe, A.P. DNA Methylation in Health and Disease. *Nat. Rev. Genet.* **2000**, *1*, 11–19. [CrossRef] [PubMed]
75. Liu, Z.; Lou, H.; Xie, K.; Wang, H.; Chen, N.; Aparicio, O.M.; Zhang, M.Q.; Jiang, R.; Chen, T. Reconstructing cell cycle pseudo time-series via single-cell transcriptome data. *Nat. Commun.* **2017**. [CrossRef] [PubMed]
76. Brown, S.E.; Fraga, M.F.; Weaver, I.C.G.; Berdasco, M.; Brown, S.E.; Fraga, M.F.; Weaver, I.C.G.; Berdasco, M.; Szyf, M. Variations in DNA Methylation Patterns During the Cell Cycle of HeLa Cells. *Epigenetics* **2007**, *2*, 54–65. [CrossRef] [PubMed]
77. Desjobert, C.; Maï, M.E.; Gérard-hirne, T.; Guianvarc, D.; Carrier, A.; Pottier, C.; Arimondo, P.B.; Riond, J.; Desjobert, C.; Maï, M.E.; et al. Combined analysis of DNA methylation and cell cycle in cancer cells Combined analysis of DNA methylation and cell cycle in cancer cells. *Epigenetics* **2015**, *2294*, 82–91. [CrossRef] [PubMed]
78. Vandiver, A.R.; Idrizi, A.; Rizzardi, L.; Feinberg, A.P.; Hansen, K.D. DNA methylation is stable during replication and cell cycle arrest. *Sci. Rep.* **2015**. [CrossRef] [PubMed]
79. Robertson, K.D. Dna methylation and human disease. *Nat. Rev. Genet.* **2005**, *6*, 597–610. [CrossRef]
80. Greenberg, M.V.C.; Bourc, D. The diverse roles of DNA methylation in mammalian development and disease. *Nat. Rev. Mol. Cell Biol.* **2019**. [CrossRef]

81. Fraga, M.F.; Ballestar, E.; Paz, M.F.; Ropero, S.; Setien, F.; Ballestar, M.L.; Heine-Suner, D.; Cigudosa, J.C.; Urioste, M.; Benitez, J.; et al. Epigenetic differences arise during the lifetime of monozygotic twins. *Proc. Natl. Acad. Sci. USA* **2005**, *102*, 10604–10609. [CrossRef]
82. Li, C.; Zhao, S.; Zhang, N.; Zhang, S.; Hou, Y. Differences of DNA methylation profiles between monozygotic twins' blood samples. *Mol. Biol. Rep.* **2013**, *40*, 5275–5280. [CrossRef]
83. Souren, N.Y.; Gerdes, L.A.; Lutsik, P.; Gasparoni, G.; Beltran, E.; Salhab, A.; Kumpfel, T.; Weichenhan, D.; Plass, C.; Hohlfeld, R.; et al. DNA methylation signatures of monozygotic twins clinically discordant for multiple sclerosis. *Nat. Commun.* **2019**, *10*, 2094. [CrossRef] [PubMed]
84. Castillo-fernandez, J.E.; Spector, T.D.; Bell, J.T. Epigenetics of discordant monozygotic twins: Implications for disease. *Genome Med.* **2014**. [CrossRef] [PubMed]
85. Garrett-Bakelman, F.E.; Manjul, D.; Stefan, J.G. The NASA Twins Study: A multidimensional analysis of a year-long human spaceflight. *Science* **2019**, *144*, eaau8650.
86. Beayno, A.; Hayek, S.E.; Noufi, P.; Tarabay, Y.; Shamseddee, W. The Role of Epigenetics in Addiction: Clinical Overview and Recent Updates. In *Psychiatric Disorders*, 2nd ed.; Human Press: Hertfordshire, UK, 2019; pp. 609–631.
87. Martin, E.M.; Fry, R.C. Environmental Influences on the Epigenome: Exposure-Associated DNA Methylation in Human Populations. *Annu. Rev. Public Health* **2018**, *39*, 309–333. [CrossRef] [PubMed]
88. Reik, W.; Walter, J. Genomic imprinting: Parental influence on the genome. *Nat. Rev. Genet.* **2001**, *2*, 21–32. [CrossRef]
89. Horsthemke, B. In Brief: Genomic imprinting and imprinting diseases. *J. Patol.* **2014**, 485–487. [CrossRef]
90. Eggermann, T.; Nanclares, G.P.D.; Maher, E.R.; Temple, I.K.; Tümer, Z.; Monk, D.; Mackay, D.J.G.; Grønskov, K.; Riccio, A.; Linglart, A.; et al. Imprinting disorders: A group of congenital disorders with overlapping patterns of molecular changes affecting imprinted loci. *Clin. Epigenet.* **2015**. [CrossRef]
91. Muurinen, M.; Hannula-jouppi, K.; Reinius, L.E.; Söderhäll, C. Hypomethylation of HOXA4 promoter is common in Silver-Russell syndrome and growth restriction and associates with stature in healthy children. *Sci. Rep.* **2017**. [CrossRef]
92. Wakeling, E.L.; Brioude, F.; Lokulo-sodipe, O.; Connell, S.M.O.; Salem, J.; Bliek, J.; Canton, A.P.M.; Chrzanowska, K.H.; Davies, J.H.; Dias, R.P.; et al. Diagnosis and management of Silver–Russell syndrome: First international consensus statement. *Nat. Rev. Endocrinol.* **2017**, *13*, 105–124. [CrossRef]
93. Gerhardt, J. Epigenetic modi fi cations in human fragile X pluripotent stem cells; Implications in fragile X syndrome modeling. *Brain Res.* **2017**, *1656*, 55–62. [CrossRef]
94. Korb, E.; Herre, M.; Zucker-scharff, I.; Gresack, J.; Allis, C.D.; Darnell, R.B.; Korb, E.; Herre, M.; Zucker-scharff, I.; Gresack, J.; et al. Excess Translation of Epigenetic Regulators Contributes to Fragile X Syndrome and Is Alleviated by Brd4 Inhibition Article Excess Translation of Epigenetic Regulators Contributes to Fragile X Syndrome and Is Alleviated by Brd4 Inhibition. *Cell* **2017**, *170*, 1209–1210. [CrossRef] [PubMed]
95. Hecht, M.; Tabib, A.; Kahan, T.; Orlanski, S.; Gropp, M.; Tabach, Y.; Yanuka, O.; Benvenisty, N.; Keshet, I.; Cedar, H. Epigenetic mechanism of FMR1 inactivation in Fragile X syndrome. *Int. J. Dev. Biol.* **2017**, *61*, 285–292. [CrossRef] [PubMed]
96. Baglivo, I.; Esposito, S.; De, C.L.; Sparago, A.; Anvar, Z.; Riso, V.; Cammisa, M.; Fattorusso, R.; Grimaldi, G.; Riccio, A.; et al. Genetic and epigenetic mutations affect the DNA binding capability of human ZFP57 in transient neonatal diabetes type 1. *FEBS Lett.* **2013**, *587*, 1474–1481. [CrossRef] [PubMed]
97. Bak, M.; Boonen, S.E.; Dahl, C.; Hahnemann, J.M.D.; Mackay, D.J.D.G.; Tümer, Z.; Grønskov, K.; Temple, I.K.; Guldberg, P.; Tommerup, N. Genome-wide DNA methylation analysis of transient neonatal diabetes type 1 patients with mutations in ZFP57. *BMC Med. Genet.* **2016**, *17*, 29. [CrossRef]
98. Fontana, L.; Bedeschi, M.F.; Maitz, S.; Cereda, A.; Faré, C.; Motta, S.; D'Ursi, P.; Orro, A.; Pecile, V. Characterization of multi-locus imprinting disturbances and underlying genetic defects in patients with chromosome 11p15. 5 related imprinting disorders. *Epigenetics* **2018**, *13*, 897–909. [CrossRef]
99. Krzyzewska, I.M.; Alders, M.; Maas, S.M.; Bliek, J.; Venema, A.; Henneman, P.; Rezwan, F.I.; Lip, K.V.D.; Mul, A.N.; Mackay, D.J.G.; et al. Genome-wide methylation profiling of Beckwith-Wiedemann syndrome patients without molecular confirmation after routine diagnostics. *Clin. Epigenet.* **2019**, *11*, 53. [CrossRef]
100. Margolis, S.S.; Sell, G.L.; Zbinden, M.A.; Bird, L.M. Angelman Syndrome. *Neurotherapeutics* **2015**, *12*, 641–650. [CrossRef]

101. Buiting, K.; Williams, C.; Horsthemke, B. Angelman syndrome—Insights into a rare neurogenetic disorder. *Nat. Rev. Neurol.* **2016**, *12*, 584–593. [CrossRef]
102. Angulo, M.A.; Butler, M.G.; Cataletto, M.E. Prader-Willi syndrome: A review of clinical, genetic, and endocrine findings. *J. Endocronol. Investig.* **2015**, *38*, 1249–1263. [CrossRef]
103. Wang, S.E.; Jiang, Y.-h. Potential of Epigenetic Therapy for Prader-Willi Syndrome. *Trends Pharmacol. Sci.* **2019**, *40*, 605–608. [CrossRef]
104. Rochtus, A.; Martin-Trujillo, A.; Izzi, B.; Elli, F.; Garin, I.; Linglart, A.; Mantovani, G.; Perez de Nanclares, G.; Thiele, S.; Decallonne, B.; et al. Genome-wide DNA methylation analysis of pseudohypoparathyroidism patients with GNAS imprinting defects. *Clin. Epigenet.* **2016**, *8*, 10. [CrossRef] [PubMed]
105. Tafaj, O.; Jüppner, H. Pseudohypoparathyroidism: One gene, several syndromes. *J. Endocronol. Investig.* **2017**, *40*, 347–356. [CrossRef] [PubMed]
106. Knudson, A.G., Jr. Mutation and cancer: Statistical study of retinoblastoma. *Proc. Natl. Acad. Sci. USA* **1971**, *68*, 820–823. [CrossRef] [PubMed]
107. Boultwood, J.; Wainscoat, J.S.; Molecular, L.R.F.; Unit, H.; Hospital, J.R. Gene silencing by DNA methylation in haematological malignancies. *Br. J. Haematol.* **2007**, *138*, 3–11. [CrossRef] [PubMed]
108. Erola, P.; Torabi, K.; Miró, R.; Camps, J. The non-random landscape of somatically-acquired uniparental disomy in cancer. *Oncotarget* **2019**, *10*, 3982–3984. [CrossRef]
109. Gelli, E.; Pinto, A.M.; Somma, S.; Imperatore, V.; Mari, F.; Cannone, M.G.; Galimberti, D.; Renieri, A.; Hadjistilianou, T.; Currò, A.; et al. Evidence of predisposing epimutation in retinoblastoma. *Hum. Mutat.* **2019**, *40*, 201–206. [CrossRef]
110. Issa, J.P.; Baylin, S.B.; Herman, J.G. DNA methylation changes in hematologic malignancies: Biologic and clinical implications. *Leukemia* **1997**, *11* (Suppl. 1), S7–S11.
111. Costello, J.F.; Fruhwald, M.C.; Smiraglia, D.J.; Rush, L.J.; Robertson, G.P.; Gao, X.; Wright, F.A.; Feramisco, J.D.; Peltomaki, P.; Lang, J.C.; et al. Aberrant CpG-island methylation has non-random and tumour-type-specific patterns. *Nat. Genet.* **2000**, *24*, 132–138. [CrossRef]
112. Herman, J.G.; Baylin, S.B. Gene silencing in cancer in association with promoter hypermethylation. *N. Engl. J. Med.* **2003**, *349*, 2042–2054. [CrossRef]
113. Jones, P.A.; Baylin, S.B. The fundamental role of epigenetic events in cancer. *Nat. Rev. Genet.* **2002**, *3*, 415–428. [CrossRef]
114. Weber, M.; Davies, J.J.; Wittig, D.; Oakeley, E.J.; Haase, M.; Lam, W.L.; Schubeler, D. Chromosome-wide and promoter-specific analyses identify sites of differential DNA methylation in normal and transformed human cells. *Nat. Genet.* **2005**, *37*, 853–862. [CrossRef] [PubMed]
115. Robert, M.F.; Morin, S.; Beaulieu, N.; Gauthier, F.; Chute, I.C.; Barsalou, A.; MacLeod, A.R. DNMT1 is required to maintain CpG methylation and aberrant gene silencing in human cancer cells. *Nat. Genet.* **2003**, *33*, 61–65. [CrossRef] [PubMed]
116. Huang, Y.; Rao, A. Connections between TET proteins and aberrant DNA modification in cancer. *Trends Genet.* **2014**, *30*, 464–474. [CrossRef] [PubMed]
117. Kim, M.; Costello, J. DNA methylation: An epigenetic mark of cellular memory. *Exp. Mol. Med.* **2017**, *49*, e322–e328. [CrossRef]
118. Baylin, S.B. DNA methylation and gene silencing in cancer. *Nat. Clin. Pract. Oncol.* **2005**, *2*, 4–11. [CrossRef]
119. Klutstein, M.; Nejman, D.; Greenfield, R.; Cedar, H. DNA Methylation in Cancer and Aging. *Cancer Res.* **2016**, *76*, 3446–3450. [CrossRef]
120. Kurkjian, C.; Kummar, S.; Murgo, A.J. DNA methylation: Its role in cancer development and therapy. *Curr. Probl. Cancer* **2008**, *32*, 187–235. [CrossRef]
121. Grady, W.M.; Carethers, J.M. Genomic and epigenetic instability in colorectal cancer pathogenesis. *Gastroenterology* **2008**, *135*, 1079–1099. [CrossRef]
122. Markowitz, S.D.; Bertagnolli, M.M. Molecular origins of cancer: Molecular basis of colorectal cancer. *N. Engl. J. Med.* **2009**, *361*, 2449–2460. [CrossRef]
123. Baretti, M.; Azad, N.S. The role of epigenetic therapies in colorectal cancer. *Curr. Probl. Cancer* **2018**, *42*, 530–547. [CrossRef]
124. Baylin, S.B.; Jones, P.A. Epigenetic Determinants of Cancer. *Cold Spring Harb. Perspect. Biol.* **2016**, *8*, a019505. [CrossRef] [PubMed]

125. Tse, J.W.T.; Jenkins, L.J.; Chionh, F.; Mariadason, J.M. Aberrant DNA Methylation in Colorectal Cancer: What Should We Target ? *Trends Cancer* **2017**, *3*, 698–712. [CrossRef] [PubMed]
126. Bian, S.; Hou, Y.; Zhou, X.; Li, X.; Yong, J.; Wang, Y.; Wang, W.; Yan, J.; Hu, B.; Guo, H.; et al. Single-cell multiomics sequencing and analyses of human colorectal cancer. *Science* **2018**, *1063*, 1060–1063. [CrossRef] [PubMed]
127. Römermann, D.; Hasemeier, B.; Metzig, K.; Göhring, G.; Schlegelberger, B.; Länger, F.; Lehmann, U. Global increase in DNA methylation in patients with myelodysplastic syndrome. *Leukemia* **2008**, 1954–1956. [CrossRef]
128. Aleshin, A.; Greenberg, P.L. Molecular pathophysiology of the myelodysplastic syndromes: Insights for targeted therapy. *Blood Adv.* **2018**, *2*, 2787–2797. [CrossRef] [PubMed]
129. Nolte, F.; Hofmann, W.-K. Myelodysplastic syndromes: Molecular pathogenesis and genomic changes. *Ann. Hmatol.* **2008**, 777–795. [CrossRef]
130. Hu, X.-m.W.J.-b.; Yang, J.; Qian, W. CEBPA methylation and mutation in myelodysplastic syndrome. *Med. Oncol.* **2015**, *32*, 192.
131. Figueroa, M.E.; Skrabanek, L.; Li, Y.; Jiemjit, A.; Fandy, T.E.; Paietta, E.; Fernandez, H.; Tallman, M.S.; Greally, J.M.; Carraway, H.; et al. MYELOID NEOPLASIA MDS and secondary AML display unique patterns and abundance of aberrant DNA methylation. *Blood* **2009**, *114*, 3448–3459. [CrossRef]
132. Stosch, J.M.; Rothenberg-thurley, M.; Riba, J.; Renz, N.; Szarc, K.; Pfeifer, D.; Follo, M.; Pahl, H.L.; Zimmermann, S.; Duyster, J.; et al. Gene mutations and clonal architecture in myelodysplastic syndromes and changes upon progression to acute myeloid leukaemia and under treatment. *Br. J. Haematol.* **2018**, 830–842. [CrossRef]
133. Heuser, M.; Yun, H.; Thol, F. Epigenetics in myelodysplastic syndromes. *Semin. Cancer Biol.* **2018**, *51*, 170–179. [CrossRef]
134. Hasegawa, N.; Oshima, M.; Sashida, G.; Matsui, H.; Koide, S.; Saraya, A.; Wang, C.; Muto, T.; Takane, K.; Kaneda, A.; et al. Impact of combinatorial dysfunctions of Tet2 and Ezh2 on the epigenome in the pathogenesis of myelodysplastic syndrome. *Leukemia* **2017**, *31*, 861–871. [CrossRef] [PubMed]
135. Garcia-manero, G.; Montalban-bravo, G. Annual Clinical Updates in Hematological Myelodysplastic syndromes: 2018 update on diagnosis, risk-stratification and management. *Am. J. Hematol.* **2018**, 129–147. [CrossRef]
136. Platzbecker, U. Treatment of MDS. *Blood* **2019**, *133*, 1096–1108. [CrossRef] [PubMed]
137. Leung, K.K.; Nguyen, A.; Shi, T.; Tang, L.; Ni, X.; Escoubet, L.; Macbeth, K.J. Multiomics of azacitidine-treated AML cells reveals variable and convergent targets that remodel the cell-surface proteome. *Proc. Natl. Acad. Sci. USA* **2018**, *116*, 695–700. [CrossRef] [PubMed]
138. Takeshima, H.; Yamada, H.; Ushijima, T. Chapter 5 Cancer Epigenetics Aberrant DNA Methylation in Cancer Diagnosis and Treatment. In *Oncogenomics*; Academic Press: Cambridge, MA, USA, 2019; pp. 65–76. [CrossRef]

E-cadherin is regulated by GATA-2 and marks the early commitment of mouse hematopoietic progenitors to the basophil and mast cell fates

Authors: Anaïs Wanet[1], Mahmoud A. Bassal[1,2], Francisco Marchi[1], Samanta A. Mariani[3], Nouraiz Ahmed[4], Marta Borchiellini[5,6], Sisi Chen[1], Junyan Zhang[1], Annalisa Di Ruscio[1,6], Kensuke Miyake[7], Mindy Tsai[8], Anuya Paranjape[8], Shin-Young Park[9], Leslie E. Silberstein[9], Hajime Karasuyama[7], Timm Schroeder[4], Elaine Dzierzak[3], Stephen J. Galli[8,10], Daniel G. Tenen[1,2,*, †] and Robert S. Welner[11,*, †]

Affiliations:

[1]Harvard Stem Cell Institute, Cambridge, MA 02138, USA, [2]Cancer Science Institute of Singapore, National University of Singapore, Singapore, [3]Centre for Inflammation Research, The University of Edinburgh, Edinburgh, UK, [4]Department of Biosystems Science & Engineering, ETH Zurich, 4058 Basel, Switzerland, [5]University of Eastern Piedmont, Department of Health Sciences, Novara, 28100, Italy, [6]University of Eastern Piedmont, Department of Translational Medicine, Novara, 28100, Italy, [7]Department of Immune Regulation, Graduate School of Medical and Dental Sciences, Tokyo Medical and Dental University (TMDU), Tokyo, Japan, [8]Department of Pathology, Stanford University School of Medicine, Stanford, CA 94305, USA, [9]Transfusion Medicine, Boston Children's Hospital and Harvard Medical School, Boston, MA 02115, USA Harvard Medical School, Boston, MA 02115, USA, [10]Department of Microbiology and Immunology and Sean N. Parker Center for Allergy and Asthma Research, Stanford University School of Medicine, Stanford, CA 94305, USA, [11]Division of Hematology/Oncology,

The University of Alabama at Birmingham, Comprehensive Cancer Center, Birmingham, AL 35294-3300, USA.

*Corresponding authors:

Robert S. Welner. Address: UAB Hematology/Oncology, 1824 6th Ave S, Birmingham, AL, 35294, USA; Phone: +1 (617) 416-2349; email: rwelner@uab.edu

Daniel G. Tenen. Address: Center for Life Sciences Room 437, 3 Blackfan Circle, Boston, MA 02115, USA; Phone: +1 (617) 735-2235; Fax: +1 (617) 735-2222; email: csidgt@nus.edu.sg.

† Co-last authors

One Sentence Summary: Expression of E-cadherin on mouse hematopoietic progenitors marks their commitment to the basophil and mast cell fates from an early stage, downstream of the expression of the GATA-2 transcription factor.

Abstract: E-cadherin is a calcium-dependent cell-cell adhesion molecule extensively studied for its involvement in tissue formation, epithelial cell behavior and suppression of cancer. However, E-cadherin expression in the hematopoietic system has not been fully elucidated. Combining single-cell RNA sequencing analyses and immunophenotyping, we revealed that granulocyte-monocyte progenitors (GMPs) expressing high levels of E-cadherin have an enriched capacity to differentiate into basophils and mast cells. Importantly, we detected E-cadherin expression on committed progenitors prior to the expression of other reported markers of these lineages. We named such progenitors pro-BMPs (pro-basophil and mast cell progenitors). Using

RNA-sequencing, we observed transcriptional priming of pro-BMPs to the basophil and mast cell lineages. We also showed that GATA-2 directly regulates E-cadherin expression in the basophil and mast cell lineages, thus providing a mechanistic connection between the expression of this cell surface marker and the basophil and mast cell fate specification.

Introduction

Basophils and mast cells (MCs) are functionally related cells of the innate immune system, playing critical roles in allergic diseases and responses to multicellular pathogens, including helminths. Both cell types express the high affinity receptor for IgE (FcεRIα), responsible for their activation upon the binding of allergen to cell-bound IgE(*1*).

Several progenitor populations have been reported to differentiate into basophils and/or MCs. However, the heterogeneity of these populations, and the controversies regarding the existence of bipotent progenitors giving rise to basophils and MCs, have hampered development of a clear model of their ontogeny. Mast cell progenitors (MCPs, Lin^- c-Kit^+ Sca-1^- $Ly6c^-$ FcεRIα^- $CD27^-$ integrin β7$(β7)^+$ T1/$ST2^+$) with exclusive MC potential were identified in the bone marrow (BM)(*2*). In the spleen, basophil-mast cell progenitors (BMCPs), identified as Lin^- c-Kit^+ FcγRII/II^{hi} $β7^{hi}$ cells, were initially described as bipotential basophil and MC progenitors(*3*). However, two subsequent reports suggested that splenic BMCPs only have MC potential(*4, 5*). Multipotential development of erythroid, megakaryocyte, and MC lineages was observed in a progenitor fraction called SN-β7(*6*), defined as Lin^- (including FcεRIα^-) c-Kit^+

Sca-1$^-$ Flk-2$^-$ CD150$^-$ β7$^+$, which contains the previously described MCPs(*2*). A small fraction of GMPs with low levels of β7 reportedly generates MCs and basophils while still preserving granulocyte and monocyte potential(*3*). And yet another small fraction of GMPs, with expression of FcεRIα and designated pre-basophil and mast cell progenitors (pre-BMPs), displays enriched capacity to differentiate into basophils and MCs(*4*). Adding more complexity, a recent study identified bone marrow BMCPs, a population phenotypically identical to splenic BMCPs, which can differentiate toward basophil and MC lineages but falls outside of the GMP definition due to low expression of CD34(*7*).

In this study, we demonstrate that E-cadherin, a calcium-dependent cell-cell adhesion molecule involved in adherens junctions, and extensively studied for its involvement in epithelial cell behavior(*8*), comprehensively marks the mouse basophil and MC lineages. We show that E-cadherin is expressed by mature basophils and splenic MCs, and that expression of E-cadherin in progenitor cells correlates with their ability to differentiate into basophils and MCs. We also identified a new "pro-basophil and mast cell progenitor" (pro-BMP) stage, contained within the GMP fraction and lacking other reported markers of basophils and MCs, which displays a strong commitment to the basophil and MC fates. Finally, we showed that E-cadherin expression is regulated downstream of the GATA-2 transcription factor in these lineages, thus demonstrating a direct connection between the expression of this surface marker with the transcriptional program regulating their fate.

Results

E-cadherin is expressed in the basophil and MC lineages

Analysis of mouse hematopoietic stem and progenitor cells (HSPCs)(*9*) using the Gene Expression Commons(*10*), the Immunological Genome Project(*11*), and public dataset(*12*) indicated a moderate, yet higher, expression of E-cadherin in GMPs compared with other progenitors (Fig. 1A-B). We next exploited a public single cell RNA sequencing (scRNA-seq) dataset(*7*) to query E-cadherin expression in the bone marrow lineage$^-$ c-Kit$^+$ compartment containing HSPCs. We found a fraction of HSPCs expressed E-cadherin (gene symbol: cdh1) (Fig. 1C). To gain more insight into the identity of these E-cadherin-expressing progenitors, we performed differential gene expression (DGE) analysis and found 1113 differentially expressed genes (false discovery rate under 0.001). Among the top 20 most upregulated genes (Fig. 1D), we identified multiple basophil and MC markers, including prss34(*13*) (mast cell protease 11), gzmb (granzyme B)(*14*), mcpt8(*15*) (mast cell protease 8), ms4a2 (high affinity IgE receptor subunit β)(*16*), fcer1a(*16*), and cma1 (chymase 1)(*17*), whose expression profiles overlapped with that of E-cadherin (Fig. 1E). Next, we exploited the scRNA-seq dataset from the Tabula Muris consortium(*18*). This dataset was established from unfractionated bone marrow cells. Thirteen different clusters were identified (Fig. S1A-B), among which two (clusters 4 and 13) expressed E-cadherin (Fig. 1F). DGE analysis identified multiple basophil specific genes among the top upregulated genes of E-cadherin-expressing cells from cluster 4 (Fig. 1G), including prss34(*13*), mcpt8(*15*), fcer1a(*16*), cpa3 (carboxypeptidase A3)(*19, 20*), cd200r3(*21*), and ms4a2(*16*). Conversely, by computing a cell probability score based on the simultaneous

expression of multiple markers of the basophil and MC lineages, we found an overlap between E-cadherin expressing cells and the basophil and MC shared gene expression signatures, as well as with the basophil-specific gene expression signature (Fig. 1H). Similar analyses performed on E-cadherin expressing cells from cluster 13 identified them as plasmacytoid dendritic cells (pDCs) (Fig. S1C-D), a cell type recently reported to also express E-cadherin(*22*).

To confirm the scRNA-seq data, we quantified E-cadherin mRNA in multiple fluorescence activated cell sorting (FACS)-purified populations corresponding to hematopoietic progenitors, pre-BMPs, basophil progenitors (baPs), basophils, and splenic MCs (Fig. 2A). pDCs were used as positive controls. GMPs and pre-BMPs displayed a higher expression of E-cadherin compared with other multipotent progenitors, while the highest levels were observed in basophils, baPs, and MCs. Next, we quantified the frequency of E-cadherin expression by flow cytometry across multiple populations (Fig. 2B-C). E-cadherin expression was detected on 31.1% (± 1.8 % (SD)) of GMPs, and on virtually all pre-BMPs, baPs, basophils, BMCPs, and splenic MCs.

Importantly, all CD49b$^+$ E-cadherin$^+$ cells were positive for FcεRIα (Fig. 2D). Therefore, E-cadherin staining can replace FcεRIα staining (together with CD49b) for the identification of mature basophils. Similar observations were performed for peripheral blood and spleen basophils (Fig. S2). Focusing on the progenitor compartment (Lin$^-$ c-Kit$^+$ Sca-1$^-$), we found that E-cadherin expression mostly overlapped with the GMP compartment based on CD34 and CD16/32 expression (Fig. 2E). A minor fraction of cells with lower CD34 expression corresponded to the recently described BM-BMCPs(*7*).

GMPs expressing high levels of E-cadherin are committed to a basophil and MC fate

As E-cadherin was expressed in some GMPs and was uniformly expressed in basophils, MCs, and their reported progenitors, we hypothesized that E-cadherin expression on GMPs might correlate with their differentiation fate. Using FACS sorting and differentiation in culture in a medium supportive of multiple myeloid lineages, we compared the differentiation potential of GMPs with high (E-cad^{high}), intermediate (E-cad^{int}) or no E-cadherin (E-cad^{neg}) expression (Fig. 3A). Strikingly, most of the cells differentiated from E-cad^{high} GMP showed commitment to the basophil or MC lineages, while E-cad^{neg} and E-cad^{int} fractions essentially differentiated into neutrophils and monocytes (Fig. 3B). We then plotted the number of each differentiated cell type generated by 1,000 GMPs over time (Fig. 3C). E-cad^{high} GMPs generated significantly higher numbers of basophils and MCs, and significantly lower numbers of neutrophils and monocytes, compared to E-cad$^{int/neg}$ GMPs. E-cad^{int} GMPs generated a significantly higher number of eosinophils compared with E-cad$^{high/neg}$ GMPs. Stainings confirmed the immature morphology of E-cad^{high} GMPs, indistinguishable from E-cad^{neg} GMPs (Fig. 3D). Cytospins from day 5 and 7 cultures also confirmed the major prevalence of basophils and MC-committed cells from the E-cad^{high} GMP fraction and of neutrophils from E-cad^{neg} GMPs.

E-cadherin expression defines a new pro–BMP stage in the basophil/MC lineages

We next further investigated the expression of E-cadherin on progenitor cells as an early marker of the basophil and MC lineages. Pre-BMPs are a fraction of GMPs that express FcεRIα

and which have been shown to have an enriched capacity to generate basophils and MCs(*4*). We therefore tested whether E-cadherin expression in GMPs would be restricted to pre-BMPs or would define a new population of progenitors. We found that only half of the E-cad^{high} GMPs also expressed FcεRIα (Fig. 4A). We thus defined a new subpopulation of GMPs characterized by a high expression of E-cadherin (at the same intensity as in pre-BMPs) and no expression of FcεRIα, that we designated "pro-basophil and mast cell progenitors" (pro-BMPs).

We first compared phenotypically the pro-BMP population with pre-BMPs, baPs, bone marrow basophils and splenic BMCPs and MCs (Fig. 4B). Pro-BMPs and pre-BMPs expressed similar levels of the progenitor markers c-Kit and CD34 and both lacked the mature basophil marker CD49b. Most pro-BMPs lacked expression of integrin β7, which was expressed in pre-BMPs and baPs albeit at a lower level than in BMCPs and MCs. By definition, pro-BMPs displayed no expression of FcεRIα, which was expressed in all the other populations of the basophil/MC lineages.

We next investigated the fate of pro-BMPs, pre-BMPs, E-cad^{int} GMPs or double-negative (E-cad^{neg} FcεRIα^{neg}) GMPs (DN-GMPs) at the single-cell level. After 4 days of differentiation, colonies were analyzed by flow cytometry (FC) (Fig. 4C). Colonies were detected for >65% of single-sorted cells from all populations. DN-GMPs generated no basophils and MCs, and mostly generated cells of other myeloid lineages (primarily neutrophils and monocytes). A small fraction of E-cad^{int} GMPs demonstrated basophil potential, but most clones differentiated into eosinophils, neutrophils, and monocytes. Pro-BMPs generated mostly cells in the basophil and MC lineages. Pre-BMPs almost exclusively gave rise to basophils and MCs. Retrospective

analysis (Fig. 4D) of the index-sorted cells showed that GMPs that exhibited multipotential ability to generate basophils and other myeloid lineages were mostly contained within the E-cad^{int} and pro-BMP fractions. However, MC potential and basophil/MC bipotential cells were found mostly in pro-BMP and pre-BMP fractions. Interestingly, bipotent progenitors for the basophil/MC lineages were found almost as frequently as unipotent MC progenitors, confirming the existence of a common progenitor for these lineages (Fig. 4E). Time course analyses showed that pro-BMPs and pre-BMPs differentiated into baPs, basophils, and MCs with similar kinetics (Fig. 4F).

In order to determine the hierarchical relationship between pro-BMPs and pre-BMPs, we performed short-term adoptive transfer experiments with analyses at 42h post-transplantation, a time-point at which transplanted cells retained high CD34 expression (Fig. 4G), consistent with an immature, progenitor state. We found that the majority of pro-BMPs generated FcεRIα-expressing cells *in vivo*, consistent with a pre-BMP stage (Fig. 4H). A smaller fraction of pro-BMPs-derived cells remained E-cadherin^{+} and FcεRIα^{-}, while a small percentage lost E-cadherin expression, consistent with their commitment to other myeloid lineages (as observed during *ex vivo* culture). Pre-BMPs however, essentially generated double-positive (E-cadherin^{+} FcεRIα^{+}) cells, suggesting that they cannot generate pro-BMPs. These data support a hierarchical relationship between pro-BMPs and pre-BMPs, according to which a large fraction of pro-BMPs transition through a pre-BMP stage during their commitment toward the basophil and MC lineages.

Pro-BMPs show transcriptional commitment to the basophil lineage and the capacity to express type 2 cytokines despite lacking FcεRIα expression

To pursue further the potential commitment of pro-BMPs to the basophil and MC lineages, we performed bulk RNA-seq on FACS-purified DN-GMPs, pro-BMPs, pre-BMPs, and basophils. Multidimensional scaling plots confirmed that the 4 populations were distinct (Fig. S3A). To understand the dynamics of basophil commitment, we generated a heatmap based on the top 50 most variable genes across all samples (Fig. 5A). Strikingly, the heatmap revealed in pro-BMPs a transcriptomic profile intermediate between DN-GMPs and pre-BMPs. Four major trends were observed: genes that were progressively upregulated from pro-BMPs to basophils; genes that were increased from pre-BMPs to basophils; genes that were specific for basophils; and genes that were expressed in DN-GMPs and pro-BMPs, and repressed in pre-BMPs and basophils. Among the genes that showed gradual upregulation from pro-BMPs to basophils, we identified basophil and/or MC markers, including the G-protein coupled receptor 34 (gpr34)(*23*), the interleukin (il)-33 receptor il1rl1 (also known as ST2)(*24*), and prss34(*13*).

To better unveil differences between pro-BMPs and DN-GMPs, we performed DGE analysis. Only 262 genes were differentially expressed between pro-BMPs and DN-GMPs, confirming their transcriptional proximity (versus 3336 between pre-BMPs and DN-GMPs and 6507 between basophils and DN-GMPs, using a false discovery rate cutoff of 0.05) (Tables S1-3). Importantly, we identified multiple genes with reported basophil and/or MC functions among the upregulated genes (Fig. 5B). These genes were further increased in pre-BMPs and basophils with the exception of gzmb, which is downregulated in basophils, in agreement with its

specific expression in MCs(*7, 25*) (Fig. 5C). By contrast, downregulated DEGs included genes expressed by other myeloid cell types, such as neuropilin 1 (nrp1) and triggering receptor expressed on myeloid cells 2 (trem2) (Table S1), both expressed in macrophages and dendritic cells(*26-29*). To look for a broader basophil and MC signature in pro-BMPs, we generated a heatmap of genes that were significantly upregulated in pro-BMPs, pre-BMPs, and basophils in comparison with DN-GMPs. Using this approach, 118 significantly upregulated genes were identified (Fig. S3C). Importantly, many of these genes were also found specifically expressed in basophils, MCs, and/or their progenitors in other studies(*4, 7, 17*), confirming the existence of a broad basophil/MC signature in pro-BMPs.

As basophils and MCs are robust producers of type 2 cytokines, we next compared the ability of the different populations to express IL-4 and IL-13(*30*). IL-4 and IL-13 mRNAs were expressed at very low to undetectable levels in DN-GMPs (Fig. 5D). Pro-BMPs expressed IL-4 and IL-13, albeit at lower levels than in more differentiated cells. Importantly, pro-BMPs, pre-BMPs, baPs, and basophils all displayed the ability to significantly induce IL-4 mRNA upon PMA and ionomycin stimulation (Fig. 5E). As expected, pro-BMPs, which lacked FcεRIα expression, did not induce IL-4 expression upon IgE crosslinking, as was observed in pre-BMPs, baPs, and basophils.

E-cadherin marks the commitment toward the basophil and MC lineages downstream of GATA-2 expression

To better understand how pro-BMPs are primed to the basophil and mast cell lineages, we compared the expression of transcription factors involved in basophil and MC specification, including the zinc finger transcription factors GATA-1(*31, 32*) and GATA-2(*33, 34*), runt-related transcription factor 1 (Runx1)(*5*), and CCAAT/enhancer-binding protein alpha (C/EBPα)(*4*). GATA-1 and GATA-2 were significantly upregulated in pro-BMPs compared with DN-GMPs (Fig. 6A, Table S1, and Fig. S4A), which confirmed their transcriptional priming to the basophil/MC lineage. Runx1 and C/EBPα, however, were only upregulated at later stages, supporting the immature state of these progenitors.

We next used the upstream regulator analysis tool of the Ingenuity Pathway Analysis software to identify upstream regulators responsible for gene expression changes between pro-BMPs and DN-GMPs (Table S4). This analysis identified GATA-2 and Zinc Finger Protein, FOG Family Member 1 (zfpm1, also known as Friend Of GATA-1 (FOG-1)) as the two regulators with the most significant overlap with the differentially expressed genes. GATA-2 was in addition predicted to be activated (Z-score 1.39) and differentially expressed in pro-BMPs vs DN-GMPs (Table S1). These two regulators were, according to the IPA database, predicted to regulate multiple basophil and MC-specific genes increased in pro-BMPs vs DN-GMPs, such as cpa3, gzmb, itga2b and ms4a2 (Table S4).

Interestingly, in MCF7 breast cancer cells, GATA-3 induces E-cadherin expression by binding GATA-like motifs located in its promoter(*35*). Further, GATA and its coregulator, U-shaped, regulate E-cadherin expression in Drosophila hematopoietic progenitors(*36*). We thus hypothesized that E-cadherin expression in the basophil and MC lineages may depend on the expression of GATA-1 and/or -2, thus providing a mechanistic connection between the expression of this surface marker and the commitment of pro-BMPs to the basophil and MC lineages.

Using a GATA-1 protein reporter mouse (*37*), we found that most GATA-1-positive GMPs expressed E-cadherin, whereas only a minor fraction expressed FcεRIα (Fig. 6B). Therefore, E-cadherin expression in GMPs correlates better with GATA-1 expression than does FcεRIα. Consistently, the majority of pro-BMPs and pre-BMPs expressed GATA-1 at the protein level, whereas E-cad^{int} GMPs and DN-GMPs did not (Fig. 6C). We next investigated whether GATA-1 directly regulates E-cadherin expression. We found a significant increase in the number of basophils in the GATA-1^{low} mutant mice(*38*) compared with WT controls, although these were less mature based on their reduced expression of CD49b (Fig.6D, Fig. S4B). These basophils expressed higher levels of GATA-2 and E-cadherin mRNA compared with WT controls (Fig. 6E), suggesting that GATA-2, rather than GATA-1, regulates E-cadherin expression. Supporting this hypothesis, analysis of public GATA-2 ChIP-sequencing datasets (Fig. 6F) showed a strong enrichment of GATA-2 in the promoter region and in the intron 2 of E-cadherin in BMDMCs(*39*), a region previously demonstrated to be essential to initiate and maintain E-cadherin expression in different tissues(*40*). These peaks were also enriched in bone marrow hematopoietic progenitor cells (lineage-negative)(*41*) and in the multipotent

FDCPmixed (factor-dependent cells Paterson) cell line(*42*). Using a GATA-2 transcriptional reporter mouse(*43*), we found that most GMPs expressing high levels of GATA-2-Venus also expressed E-cadherin, whereas only a fraction expressed FcεRIα (Fig. 6G). Consistently, the majority of pro-BMPs and all pre-BMPs expressed GATA-2 Venus, while E-cad^{int} GMPs and DN-GMPs did not (Fig. 6H). Therefore, GATA-2 and E-cadherin are co-expressed in myeloid progenitors committed to the basophil/MC lineages.

To demonstrate that GATA-2 regulates E-cadherin expression, we transfected baPs with siRNA against GATA-2 and found that the knockdown of GATA-2 resulted in the decreased expression of E-cadherin mRNA (Fig. 6I). To confirm these findings, we deleted GATA-2 in FACS-sorted pre-BMPs from an inducible GATA-2 knockout mouse(*33*) (Fig. 6J). Accordingly, GATA-2 deletion resulted in a loss of E-cadherin expression at the mRNA and protein levels (Fig. 6J-K).

Discussion

We have demonstrated that mouse basophils express E-cadherin and can be identified using CD49b in conjunction with E-cadherin instead of FcεRIα. This opens the possibility of marking and sorting mature basophils while leaving the FcεRI untouched, which could be of great utility for functional studies involving IgE ligation. The expression of E-cadherin has been reported in mouse BMCMCs and peritoneal MCs(*44*). In agreement with these observations, we found expression of E-cadherin in splenic MCs. Surprisingly, E-cadherin mRNA is downregulated in connective tissue MCs when compared with basophils(*17*), suggesting that

E-cadherin expression in mature MCs may be heterogeneous depending on their anatomic location.

We also demonstrated that, in the bone marrow, E-cadherin comprehensively marks the basophil and MC lineages early in their specification. We identified a new progenitor population, "pro-BMPs", which displays an enriched capacity to differentiate into basophils and MCs and expresses high levels of E-cadherin, yet lacks other reported markers of the basophil and MC lineages. Pro-BMPs are phenotypically different from the previously described pre-BMPs(*4*), which express FcεRIα, and from MCPs(*2*), which express β7 and have no potential to differentiate into basophils. Pro-BMPs also differ from bone marrow BMCPs(*7*), which constitute a much rarer subset, expressing β7 but not CD34. Finally, pro-BMPs differ from the SN-β7 progenitors, which also express β7 and display multipotential development of erythroid, megakaryocyte and MC lineages(*6*). Although we did not analyze the erythroid and megakaryocyte potential of pro-BMPs, a small fraction of progenitors expressing E-cadherin overlaps with the CMP definition (due to lower CD16/32 expression) (Fig. 2E). Therefore, it remains possible that progenitors expressing E-cadherin have multipotential to differentiate toward the erythroid, megakaryocyte and MC lineages.

Using bulk RNA-seq, we demonstrated that pro-BMPs display a transcriptional profile consistent with their priming to the basophil/MC lineages. We identified a broad transcriptional signature in pro-BMPs that is enriched for genes specifically expressed in basophils, MCs, and their progenitors. We also found that the transcription factors GATA-1 and -2, involved in

basophil and MC differentiation and functions(*31, 32, 34*), were upregulated in pro-BMPs, pre-BMPs and basophils when compared with DN-GMPs. We demonstrated that E-cadherin expression was regulated by the transcription factor GATA-2, thus explaining a mechanistic connection between the expression of E-cadherin and the commitment of early hematopoietic progenitors to the basophil and MC lineages.

Differentiation assays at the single cell level demonstrated that pro-BMPs had a strong capacity to differentiate into basophils and MCs. The existence of bipotent progenitors of both the basophil and MC lineages has been controversial. Because basophil/MC bipotential was found almost as frequently as MC unipotential (Fig. 4E), the probability that the observed basophil/MC bipotential results from the sorting of cell doublets is extremely unlikely. Our data support the conclusion that a small fraction of progenitors displays both basophil and MC potential, at least in the culture conditions tested. We also observed that a fraction of pro-BMPs and GMPs with intermediate levels of E-cadherin expression displayed multipotential toward the basophil and other myeloid lineages (including neutrophils). Our data are therefore in contrast with the observation that human CD34+ CD38+ progenitors have either neutrophil/monocyte or eosinophil/basophil/MC potential(*45*). These differences are likely related to species-specific differences and/or culture conditions. Accordingly, mouse GATA-1-expressing pre-GMs can generate MC, neutrophils and eosinophils(*46*). Unlike previously reported observations(*4*), we found that the ability for bi/multipotential differentiation to basophil and other myeloid fates was restricted to cells lacking FcεRIα expression. These differences are likely due to the different culture conditions and sorting stringencies employed (we used single cell assays). Our observations support the conclusion that high levels of FcεRIα expression is acquired once cells

are strictly committed to the basophil or MC fate and have essentially lost the potential to generate other myeloid lineages.

Interestingly, a number of ligands for E-cadherin have been described, including E-cadherin itself, but also the integrinα2β1(*47*), integrinαEβ7(*48*), and the Killer Cell Lectin Like Receptor G1(*49*). These ligands are expressed by multiple immune and non-immune cell subsets, suggesting that E-cadherin expression on basophils and MCs might be important for their ability to signal to multiple other cell types. Further studies are now required to identify the cell interactions mediated by E-cadherin in basophils, MCs, and their progenitors, and to understand how E-cadherin may modulate their signaling in health and disease.

Materials and Methods

Study design

The aim of this study was to investigate the expression of E-cadherin in the mouse hematopoietic system. We used public microarrays and single-cell RNA-sequencing datasets to query E-cadherin expression in the hematopoietic system and identified expression in the mouse basophil lineage. We used flow cytometry and RNA-sequencing analyses to characterize a new progenitor population characterized by a high commitment towards the basophil and MC lineages that we named “pro-BMPs”. Using genetic approaches, we demonstrated a role for GATA-2 in the regulation of E-cadherin expression in the mouse basophil and MC lineages. All

data were collected from independent age and gender-matched biological replicates, i.e., from different mice or cells isolated from different mice.

Mice

Wild-type (WT) C57BL/6 mice and B6.SJL-Ptprca Pepcb/BoyJ were purchased from The Charles River Laboratory. Experiment were performed with 8-12 weeks-old mice. Short-term adoptive transfer experiments were performed by injecting in the tail vein 12,000 pro-BMPs or pre-BMPs FACS-sorted from C57BL/6 donor mice into B6.SJL-Ptprca Pepcb/BoyJ recipients, pre-conditioned by a sublethal irradiation at 450 rads 6h before transplantation. Mice were housed in a sterile barrier facility and procedures were approved by the Institutional Animal Care and Use Committee at the Beth Israel Deaconess Medical Center. Bone marrow cells from GATA-1^{low} mice(*38*) and from GATA-2 inducible KO mice (Gata2f/f RosaYfp/Yfp TgCreErt2hemi and its control Gata2+/+ RosaYfp/Yfp TgCreErt2hemi)(*33*) were provided by Stuart Orkin and Hua Huang, respectively. The GATA-1-mCherry knock-in mice(*37*) were housed and analyzed in Schroeder's laboratory at ETH Zurich and animal experiments were approved according to Institutional guidelines of ETH Zurich and Swiss Federal Law by veterinary office of Canton Basel-Stadt, Switzerland (approval number 2655). GATA-2 Venus mice(*43*) were housed and bred by Dzierzak's laboratory, in animal facilities at the University of Edinburgh, in compliance with the Home Office regulations. All procedures with animals were conducted by a Home Office UK Project License and approved by the University of Edinburgh Ethical Review Committee.

Flow cytometry (FC)

Single-cell suspensions from various organs were analyzed with an LSRII flow cytometer (BD Biosciences) and sorted using a FACSAria II (BD Biosciences). Phosphate buffered saline with 2% fetal bovine serum was used as a buffer at all times as the use of 2 mM EDTA interfered with the binding of the anti-E-cadherin antibody. Detailed procedures, antibodies and gating strategies are described in the Supplementary Material and Methods and Fig. S5-6.

qRT-PCR

RNA was extracted using TRIzol reagent (Thermo Fisher Scientific), reverse transcribed using the Transcriptor First Strand cDNA Synthesis Kit (Roche), and quantitatively assessed using Taq™ universal SYBR® Green (Bio-Rad) and a Rotor-Gene 6000 Real-time PCR machine (Corbett). For each sample, transcript levels of tested genes were normalized to RPL13A. Primers are listed in Supplementary Material and Methods.

Cell culture

Sorted cells were differentiated in StemSPAN SFEM (STEMCELL Technologies) supplemented with 10 ng/ml mIL-3, 10 ng/ml hIL-6, 10 ng/ml mM-CSF, 10 ng/ml mGM-CSF, 20 ng/ml mSCF, 50 ng/ml mFLT3 (all cytokines from Peprotech), and Pen-Strep (Gibco) (100U/mL). When indicated, sorted cells were stimulated for 4h with 50 ng/ml phorbol 12-myristate 13-acetate (PMA) and 1 µM ionomycin (P&I) and the corresponding vehicles, crosslinked (CL) with 1 µg/ml IgE (D8406, Sigma) and 1 µg/ml anti-IgE (Clone R35-72, BD) or not crosslinked (NCL). Nucleofection was performed 2 days post-plating using Amaxa Nucleofector kit L using Amaxa Nucleofector II program X-001 and 1 µM of siRNA against mouse GATA-2

(Dharmacon E-062114-00- 0005) or non-targeting control (Dharmacon D-001910-01-05). Following nucleofection, 900 ul of complete medium was added to cells, which were cultured for an additional 48h before analysis. For the induction of GATA-2 KO, sorted pre-BMPs were treated with 50 nM 4-hydroxytamoxifen (Sigma-Aldrich) for 5 days before analysis.

RNA-sequencing (RNA-seq)

The indicated FACS-sorted populations from 10-12-week-old C57BL/6 mice were lysed in TRIzol and RNA was isolated. Biological triplicates were prepared. RNA quality was assessed using Agilent 2100 Bioanalyzer (RIN ≥ 8.4 for all samples). Libraries were prepared using SmartSeq v4 (Takara) and Nextera XT DNA library kits (Illumina) and sequenced on a NextSeq 500 (Illumina) High flowcell (2x75). Data were deposited on Gene Expression Omnibus(*50*) under accession number GSE132122.

Bioinformatics analyses

Expression of E-cadherin was accessed via the Gene Expression Commons(*10*) (https://gexc.riken.jp/models/1038/genes/Cdh1?q=CDH1) and via the Immunological Genome Project(*11*) website for GSE15907 and GSE60101. For the Dahlin et al. (*7*) dataset, normalized counts for the Lin^- $c\text{-}Kit^+$ fraction were sourced from https://gottgens-lab.stemcells.cam.ac.uk/adultHSPC10X/. For the Tabula Muris Consortium(*18*), raw droplet counts were sourced from the mouse bone marrow compartment (GSM3040900 and GSM3040901). Detailed pipeline processing for scRNA-seq and bulk RNA-seq are described in Supplementary Material and Methods. Public GATA-2 ChIP-sequencing datasets were accessed

using ChIP-Atlas(*51*) (http://chip-atlas.org) and visualized with Integrative Genomics Viewer(*52*).

Statistical analyses

Statistical analyses were performed using the Prism 7 Software. Depending on the data being compared, we used t-tests, one-way or two-way ANOVAs followed by multiple comparisons using the Tukey's test or Dunnetts'test as indicated in the Table S6. To ensure variance homogeneity, data presented in Fig. 5D-E were log10 transformed.

Supplementary Materials

Supplementary Materials and Methods

Supplementary Figures

Supplementary Tables

References:

1. D. Voehringer, Protective and pathological roles of mast cells and basophils. *Nat Rev Immunol* **13**, 362-375 (2013).
2. C. C. Chen, M. A. Grimbaldeston, M. Tsai, I. L. Weissman, S. J. Galli, Identification of mast cell progenitors in adult mice. *Proc Natl Acad Sci U S A* **102**, 11408-11413 (2005).
3. Y. Arinobu *et al.*, Developmental checkpoints of the basophil/mast cell lineages in adult murine hematopoiesis. *Proc Natl Acad Sci U S A* **102**, 18105-18110 (2005).
4. X. Qi *et al.*, Antagonistic regulation by the transcription factors C/EBPalpha and MITF specifies basophil and mast cell fates. *Immunity* **39**, 97-110 (2013).
5. K. Mukai *et al.*, Critical role of P1-Runx1 in mouse basophil development. *Blood* **120**, 76-85 (2012).
6. C. B. Franco, C. C. Chen, M. Drukker, I. L. Weissman, S. J. Galli, Distinguishing mast cell and granulocyte differentiation at the single-cell level. *Cell Stem Cell* **6**, 361-368 (2010).

7. J. S. Dahlin *et al.*, A single-cell hematopoietic landscape resolves 8 lineage trajectories and defects in Kit mutant mice. *Blood* **131**, e1-e11 (2018).
8. F. van Roy, G. Berx, The cell-cell adhesion molecule E-cadherin. *Cell Mol Life Sci* **65**, 3756-3788 (2008).
9. R. Gazit *et al.*, Transcriptome analysis identifies regulators of hematopoietic stem and progenitor cells. *Stem Cell Reports* **1**, 266-280 (2013).
10. J. Seita *et al.*, Gene Expression Commons: an open platform for absolute gene expression profiling. *PLoS One* **7**, e40321 (2012).
11. T. S. Heng, M. W. Painter, C. Immunological Genome Project, The Immunological Genome Project: networks of gene expression in immune cells. *Nat Immunol* **9**, 1091-1094 (2008).
12. D. Lara-Astiaso *et al.*, Immunogenetics. Chromatin state dynamics during blood formation. *Science* **345**, 943-949 (2014).
13. T. Ugajin *et al.*, Basophils preferentially express mouse Mast Cell Protease 11 among the mast cell tryptase family in contrast to mast cells. *J Leukoc Biol* **86**, 1417-1425 (2009).
14. C. M. Tschopp *et al.*, Granzyme B, a novel mediator of allergic inflammation: its induction and release in blood basophils and human asthma. *Blood* **108**, 2290-2299 (2006).
15. M. Poorafshar, H. Helmby, M. Troye-Blomberg, L. Hellman, MMCP-8, the first lineage-specific differentiation marker for mouse basophils. Elevated numbers of potent IL-4-producing and MMCP-8-positive cells in spleens of malaria-infected mice. *Eur J Immunol* **30**, 2660-2668 (2000).
16. H. Metzger, A. Goetze, J. Kanellopoulos, D. Holowka, C. Fewtrell, Structure of the high-affinity mast cell receptor for IgE. *Fed Proc* **41**, 8-11 (1982).
17. D. F. Dwyer, N. A. Barrett, K. F. Austen, C. Immunological Genome Project, Expression profiling of constitutive mast cells reveals a unique identity within the immune system. *Nat Immunol* **17**, 878-887 (2016).
18. C. Tabula Muris *et al.*, Single-cell transcriptomics of 20 mouse organs creates a Tabula Muris. *Nature* **562**, 367-372 (2018).
19. W. E. Serafin, E. T. Dayton, P. M. Gravallese, K. F. Austen, R. L. Stevens, Carboxypeptidase A in mouse mast cells. Identification, characterization, and use as a differentiation marker. *J Immunol* **139**, 3771-3776 (1987).
20. D. Voehringer, K. Shinkai, R. M. Locksley, Type 2 immunity reflects orchestrated recruitment of cells committed to IL-4 production. *Immunity* **20**, 267-277 (2004).
21. D. Voehringer, D. B. Rosen, L. L. Lanier, R. M. Locksley, CD200 receptor family members represent novel DAP12-associated activating receptors on basophils and mast cells. *J Biol Chem* **279**, 54117-54123 (2004).
22. E. Bi *et al.*, E-cadherin expression on multiple myeloma cells activates tumor-promoting properties in plasmacytoid DCs. *J Clin Invest* **128**, 4821-4831 (2018).
23. T. Sugo *et al.*, Identification of a lysophosphatidylserine receptor on mast cells. *Biochem Biophys Res Commun* **341**, 1078-1087 (2006).
24. D. R. Moritz, H. R. Rodewald, J. Gheyselinck, R. Klemenz, The IL-1 receptor-related T1 antigen is expressed on immature and mature mast cells and on fetal blood mast cell progenitors. *J Immunol* **161**, 4866-4874 (1998).
25. J. Pardo *et al.*, Granzyme B is expressed in mouse mast cells in vivo and in vitro and causes delayed cell death independent of perforin. *Cell Death Differ* **14**, 1768-1779 (2007).

26. X. Dai *et al.*, Ablation of Neuropilin 1 in Myeloid Cells Exacerbates High-Fat Diet-Induced Insulin Resistance Through Nlrp3 Inflammasome In Vivo. *Diabetes* **66**, 2424-2435 (2017).
27. R. Tordjman *et al.*, A neuronal receptor, neuropilin-1, is essential for the initiation of the primary immune response. *Nat Immunol* **3**, 477-482 (2002).
28. A. Bouchon, C. Hernandez-Munain, M. Cella, M. Colonna, A DAP12-mediated pathway regulates expression of CC chemokine receptor 7 and maturation of human dendritic cells. *J Exp Med* **194**, 1111-1122 (2001).
29. D. H. Chung, W. E. Seaman, M. R. Daws, Characterization of TREM-3, an activating receptor on mouse macrophages: definition of a family of single Ig domain receptors on mouse chromosome 17. *Eur J Immunol* **32**, 59-66 (2002).
30. C. M. Yuk *et al.*, Basophil-derived IL-6 regulates TH17 cell differentiation and CD4 T cell immunity. *Sci Rep* **7**, 41744 (2017).
31. A. R. Migliaccio *et al.*, GATA-1 as a regulator of mast cell differentiation revealed by the phenotype of the GATA-1low mouse mutant. *J Exp Med* **197**, 281-296 (2003).
32. Y. Nei *et al.*, GATA-1 regulates the generation and function of basophils. *Proc Natl Acad Sci U S A* **110**, 18620-18625 (2013).
33. Y. Li, X. Qi, B. Liu, H. Huang, The STAT5-GATA2 pathway is critical in basophil and mast cell differentiation and maintenance. *J Immunol* **194**, 4328-4338 (2015).
34. F. Y. Tsai, S. H. Orkin, Transcription factor GATA-2 is required for proliferation/survival of early hematopoietic cells and mast cell formation, but not for erythroid and myeloid terminal differentiation. *Blood* **89**, 3636-3643 (1997).
35. W. Yan, Q. J. Cao, R. B. Arenas, B. Bentley, R. Shao, GATA3 inhibits breast cancer metastasis through the reversal of epithelial-mesenchymal transition. *J Biol Chem* **285**, 14042-14051 (2010).
36. H. Gao, X. Wu, N. Fossett, Drosophila E-cadherin functions in hematopoietic progenitors to maintain multipotency and block differentiation. *PLoS One* **8**, e74684 (2013).
37. P. S. Hoppe *et al.*, Early myeloid lineage choice is not initiated by random PU.1 to GATA1 protein ratios. *Nature* **535**, 299-302 (2016).
38. M. A. McDevitt, Y. Fujiwara, R. A. Shivdasani, S. H. Orkin, An upstream, DNase I hypersensitive region of the hematopoietic-expressed transcription factor GATA-1 gene confers developmental specificity in transgenic mice. *Proc Natl Acad Sci U S A* **94**, 7976-7981 (1997).
39. F. J. Calero-Nieto *et al.*, Key regulators control distinct transcriptional programmes in blood progenitor and mast cells. *EMBO J* **33**, 1212-1226 (2014).
40. M. P. Stemmler, A. Hecht, R. Kemler, E-cadherin intron 2 contains cis-regulatory elements essential for gene expression. *Development* **132**, 965-976 (2005).
41. L. Li *et al.*, Nuclear adaptor Ldb1 regulates a transcriptional program essential for the maintenance of hematopoietic stem cells. *Nat Immunol* **12**, 129-136 (2011).
42. G. May *et al.*, Dynamic analysis of gene expression and genome-wide transcription factor binding during lineage specification of multipotent progenitors. *Cell Stem Cell* **13**, 754-768 (2013).
43. P. Kaimakis *et al.*, Functional and molecular characterization of mouse Gata2-independent hematopoietic progenitors. *Blood* **127**, 1426-1437 (2016).
44. T. Tegoshi *et al.*, E-cadherin and cadherin-associated cytoplasmic proteins are expressed in murine mast cells. *Lab Invest* **80**, 1571-1581 (2000).

45. R. Drissen, S. Thongjuea, K. Theilgaard-Monch, C. Nerlov, Identification of two distinct pathways of human myelopoiesis. *Sci Immunol* **4**, (2019).
46. R. Drissen *et al.*, Distinct myeloid progenitor-differentiation pathways identified through single-cell RNA sequencing. *Nat Immunol* **17**, 666-676 (2016).
47. J. D. Whittard *et al.*, E-cadherin is a ligand for integrin alpha2beta1. *Matrix Biol* **21**, 525-532 (2002).
48. U. Cavallaro, G. Christofori, Cell adhesion and signalling by cadherins and Ig-CAMs in cancer. *Nat Rev Cancer* **4**, 118-132 (2004).
49. M. Ito *et al.*, Killer cell lectin-like receptor G1 binds three members of the classical cadherin family to inhibit NK cell cytotoxicity. *J Exp Med* **203**, 289-295 (2006).
50. R. Edgar, M. Domrachev, A. E. Lash, Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res* **30**, 207-210 (2002).
51. S. Oki *et al.*, ChIP-Atlas: a data-mining suite powered by full integration of public ChIP-seq data. *EMBO Rep* **19**, (2018).
52. J. T. Robinson *et al.*, Integrative genomics viewer. *Nat Biotechnol* **29**, 24-26 (2011).

Acknowledgments:

We thank Hua Huang and Yapeng Li (National Jewish Health, Denver, CO, USA) for sharing material from the inducible GATA-2 KO mice, Stuart H. Orkin (Dana Farber and Boston Children's Hospital, Boston, MA, USA) for sharing material from the GATA-1^{low} mutant mouse strain and P.P. Pandolfi (BIDMC) for access to equipment. We thank also Samuel L. Wolock and Allon M. Klein (Department of Systems Biology, Harvard Medical School, Boston, MA, USA) for their help with the analyses of scRNA-seq datasets, the Bauer Core Facility at Harvard University for expertise in RNA-seq, the HSCI/Joslin flow cytometry facility for expertise in cell sorting. **Funding**: A.W. was supported by the B.A.E.F. and WBI.World fellowships. This work was supported by NIH grants P01HL131477 (D.G.T. and R.S.W.), 5P01CA66996 (D.G.T.), R35CA197697 (D.G.T.), P01HL095489 (L.E.S.), R01AR067145, R01AI132494 (S.J.G.) and #CF-0003-11-04 and P30DK036836 (Joslin flow cytometry facility). M.B. was supported by a PhD Fellowship from the Italian Ministry of Education, University and Reasearch. A.D.R. was

supported by NIH/NCI Award Number R00CA188595, the Italian Association for Cancer Research (AIRC) and the Giovanni Armenise-Harvard Foundation. N.A. and T.S. acknowledge support from SNF grant 179490. E.D. acknowledge support from ERC Advanced grant 341096 and Bloodwise 18010. This research is supported by the Singapore Ministry of Health's National Medical Research Council under its Singapore Translational Research (STaR) Investigator Award, and by the National Research Foundation Singapore and the Singapore Ministry of Education under its Research Centre of Excellence initiative. **Author contributions**: Contribution: A.W. conceived the project, designed the analyses, performed most experiments and wrote the manuscript. M.A.B. designed and performed the bioinformatics analyses. F. M., S.A.M., N.A., M. B.., S. C., J. Z., A.D.R., S.P., K.M. contributed to data and analyses. L.E.S, H.K., T.S. and E.D. supervised experiments performed in their respective labs. M.T., A.P. and S.J.G. provided critical feedback. D.G.T. and R.S.W. supervised the study. All authors discussed and approved the manuscript. **Competing interests**: The authors declare no conflict of interest. **Data and materials availability**: RNA-sequencing data from the pro-BMPs, pre-BMPs, DN-GMPs and basophil populations were deposited on Gene Expression Omnibus(*50*) under accession number GSE132122.

Figures:

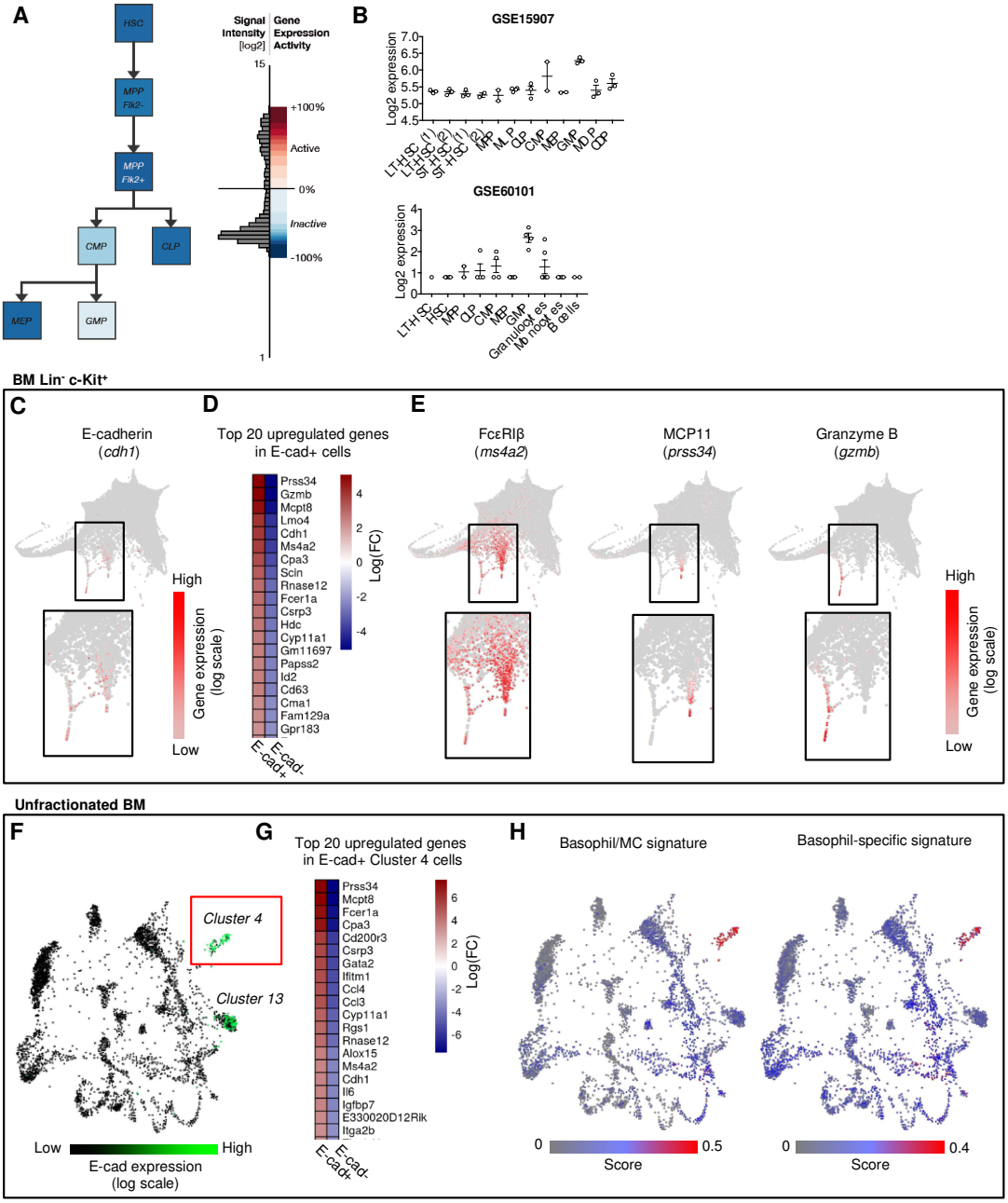

Fig. 1. Transcriptomic analyses uncover expression of E-cadherin in GMPs and in the basophil/MC lineage. **(A)** Graphical representation of the probeset meta profile for E-cadherin in the Gene Expression Commons

(https://gexc.riken.jp/models/1038/genes/Cdh1?q=CDH1). The axis represents expression intensity in log2 scale. The distribution of expression levels in the database is displayed by the histogram on the left side of the axis. **(B)** E-cadherin expression in progenitor populations from ImmGen GSE15907 (microarray) and GSE60101 (RNA-seq). **(C-E)** Analysis of a public dataset on BM Lin$^-$ c-Kit$^+$ HSPCs. **(C, E)** Expression pattern of the selected genes was accessed at https://gottgens-lab.stemcells.cam.ac.uk/adultHSPC10X/ and is visualized in the force-directed graph layout of the single-cell profiles (ms4a2, basophil/MC-specific gene; prss34, basophil-specific gene; gzmb, MC-specific gene). **(D)** Top 20 most upregulated genes in E-cadherin-expressing BM Lin$^-$ c-Kit$^+$ HSPCs versus non-E-cadherin-expressing cells. **(F-H)** Analysis of the Tabula Muris dataset on unfractionated BM. **(F)** Springplot showing E-cadherin expression. **(G)** Top 20 most upregulated genes in E-cadherin-expressing cells from cluster 4 versus all non-E-cadherin-expressing cells. **(H)** Cell probability scores were calculated for the Tabula Muris Dataset on the basis of previously published gene expression signatures shared by basophils and MCs or specific for basophils(*17*). Abbreviations: LT-HSC (long term hematopoietic stem cell), ST-HSC (short term HSC), MPP (multipotent progenitor), MLP (multilymphoid progenitor), CLP (common lymphocyte progenitor), CMP (common myeloid progenitor), MEP (megakaryocyte erythrocyte progenitor), MDP (monocyte dendritic cell progenitor), CDP (common dendritic cell progenitor).

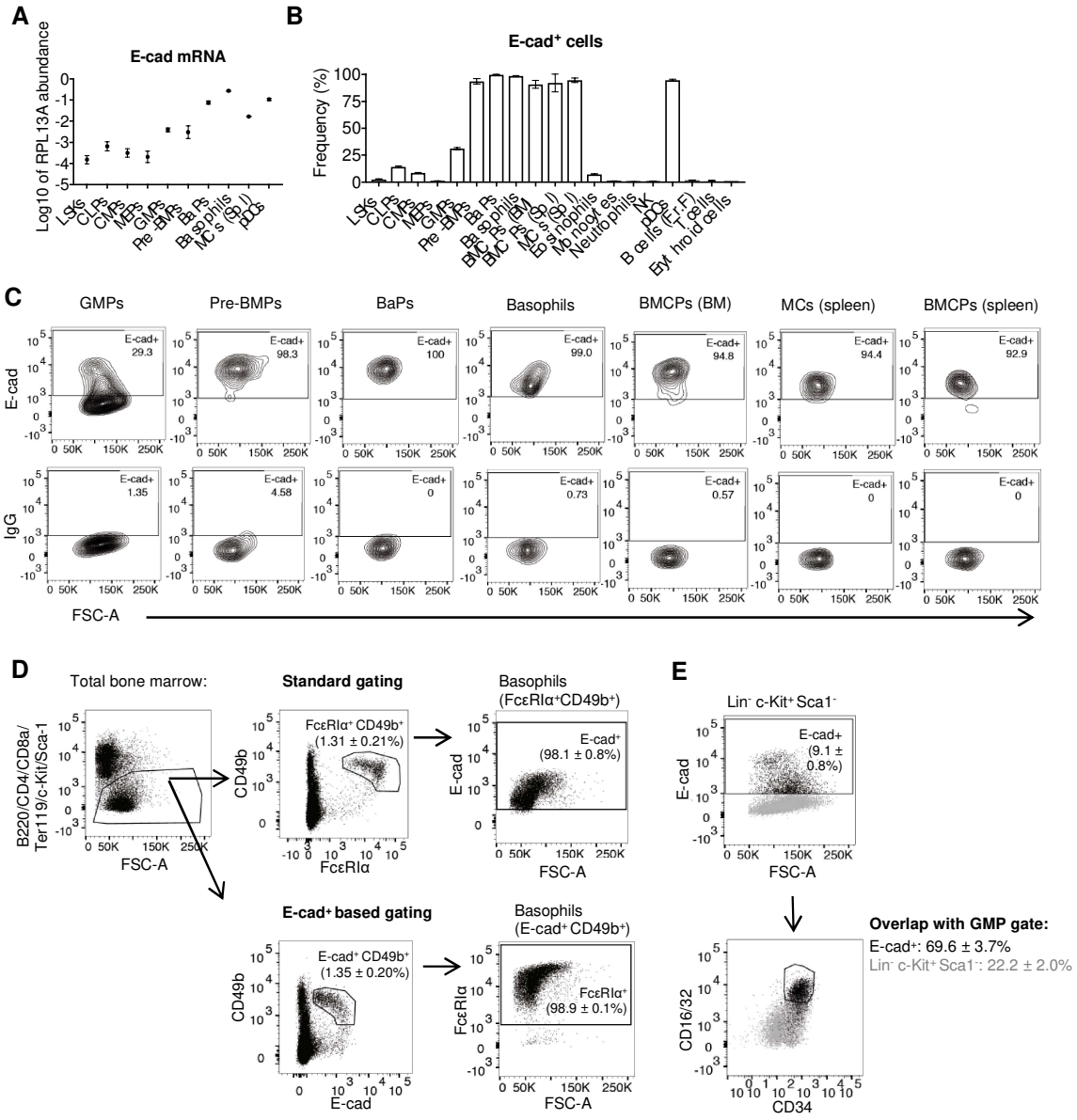

Fig. 2. E-cadherin is expressed in the mouse basophil and MC lineages. **(A)** RNA was isolated from the indicated populations (sorted from the bone marrow, unless otherwise indicated); E-cadherin mRNA abundance was determined by RT-qPCR and is expressed relative to RPL13A levels. **(B)** Frequency of E-cadherin expression for each hematopoietic cell as determined by FC. **(C)** Representative contour plots showing E-cadherin expression. **(D)** Dot plots showing that gating of Lin$^-$ Sca-1$^-$ c-Kit$^-$ CD49b$^+$

E-cadherin^{+} cells identifies basophils as efficiently as gating of Lin^{-} Sca-1^{-} c-Kit^{-} CD49b^{+} FcεRIα^{+} cells. **(E)** Dot plots showing expression of E-cadherin on Lin^{-} Sca-1^{-} c-Kit^{+} cells and the overlap with the GMP gate. Dots on the left of the GMP gate, with reduced CD34 expression, correspond to BM-BMCPs. **(A-E)** Data are the mean ± SEM of 3 C57BL/6 mice. Abbreviations: LSKs (Lin^{-} Sca-1^{+} c-Kit^{+}), CLPs (common lymphocyte progenitors), CMPs (common myeloid progenitors), MEP (megakaryocyte erythrocyte progenitors), NK (natural killer), Fr. F (Hardy's fraction F).

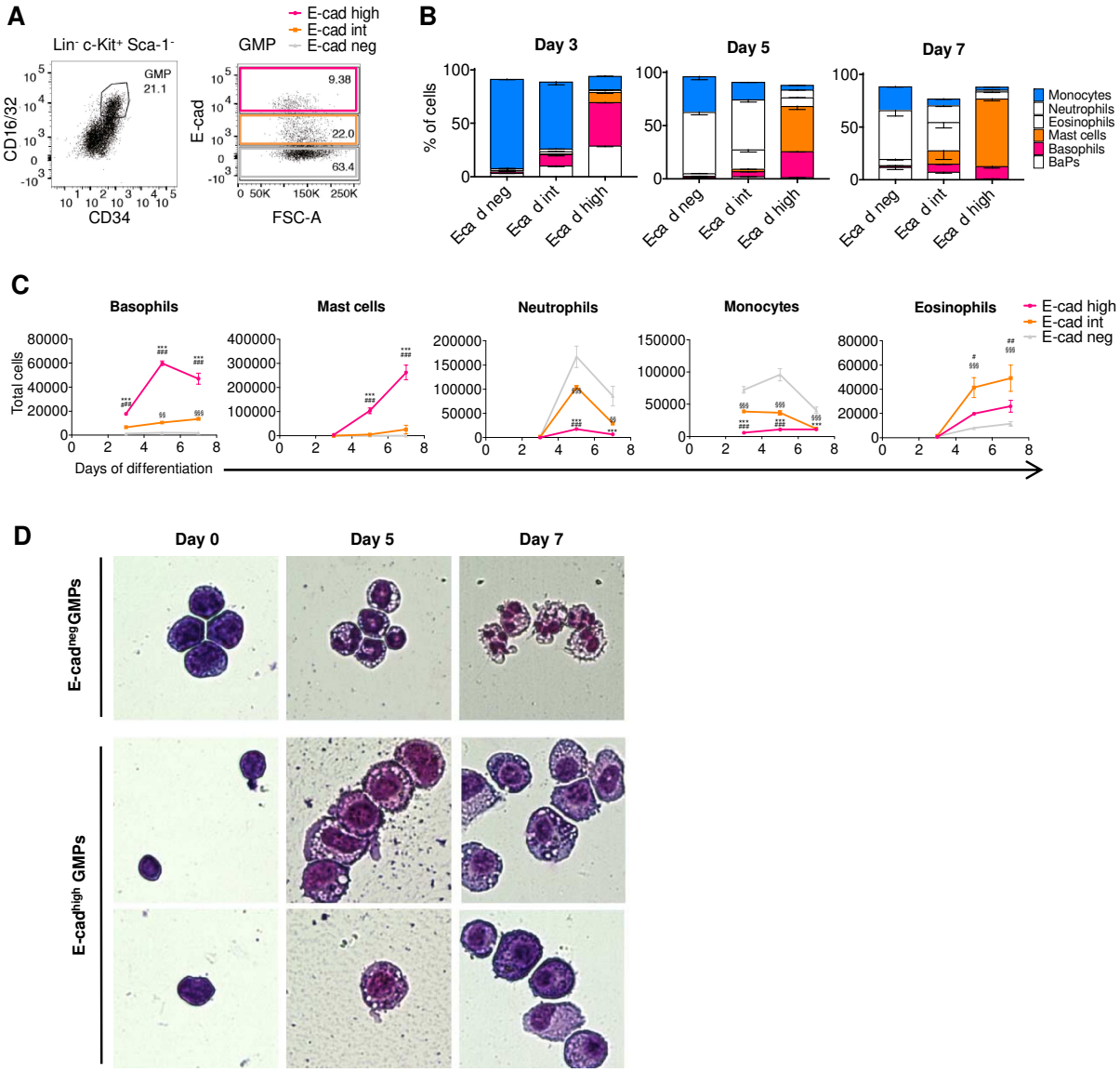

Fig. 3. E-cadherin-expressing GMPs generate more basophils and MCs in culture. **(A)** Representative dot plot for the sorting of GMPs expressing different levels of E-cadherin. **(B-D)** GMPs were differentiated in liquid culture for 7 days. **(B)** Differentiation was assessed by FC and the frequency of each cell type is expressed as percentage of live cells. **(C)** Total number of differentiated cells obtained when plating 1,000 GMPs. Statistics were calculated using two-ways ANOVA followed by Tukey's multiple comparison tests. **(D)** Wright-Giemsa stainings before and after 5 and 7 days of

differentiation in culture. Magnification: 600x. Data are the mean ± SEM of 3 experiments.

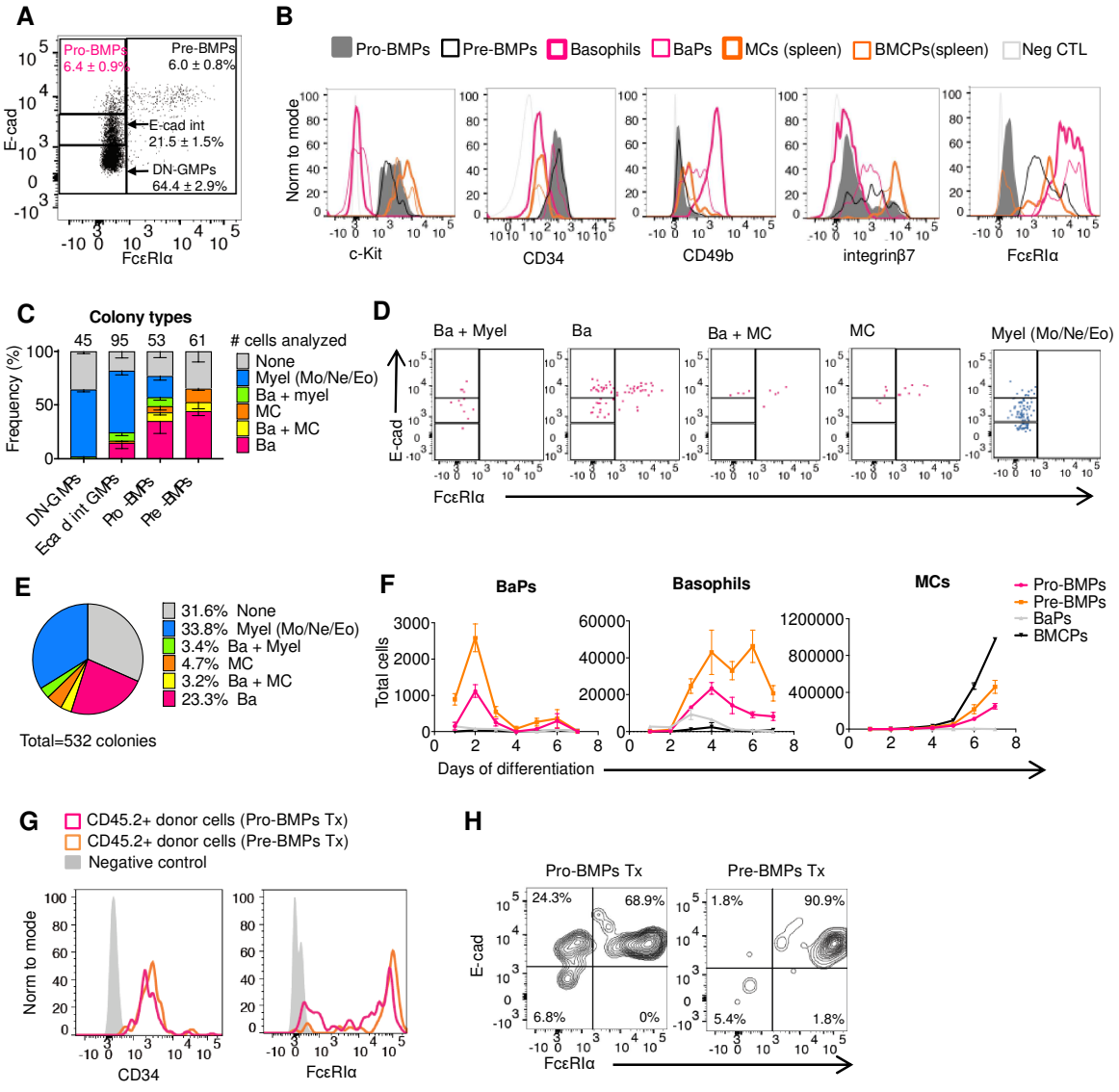

Fig. 4. E-cadherin marks the commitment to the basophil and MC lineages at the pro-BMP stage. **(A)** Dot plot showing distribution and frequency of E-cadherin and FcεRIα on GMPs (mean ± SEM of n=6). **(B)** Histograms comparing the expression of the indicated markers on the different populations. **(C-E)** Single GMPs were index-sorted based on the expression of E-cadherin and FcεRIα and differentiated for 4 days. Differentiation was

assessed by FC. Data are the mean ± SEM of cells sorted from 3 different mice and are representative of 2 independent experiments. **(C)** Frequency of each colony type on day 4. **(D)** Dot plots from day 0 showing the expression of E-cadherin and FcεRIα of single cells that generated the indicated type of colony (listed on top) on day 4. **(E)** Frequency of each colony type obtained on day 4 after pooling the data from all single progenitors analyzed (2 independent experiments, for a total of progenitors sorted from 6 mice). **(F)** Pro-BMPs, pre-BMPs, baPs and BMCPs (depicted by lines of different colors) were FACS-sorted from the BM (pro-BMPs, pre-BMPs and baPs) or spleen (BMCPs) and differentiated in liquid culture for 7 days. Their differentiation into baPs, basophils and MC was assessed by FC until day 7. The number of differentiated cells was calculated based on the plating of 1,000 cells/well (mean ± SEM of n=3). **(G-H)** Pro-BMPs and pre-BMPs were FACS-sorted from donor mice (expressing CD45.2) and transplanted into recipients (expressing CD45.1). Expression of the indicated cell surface markers on the transplanted cells (CD45.2+) were analyzed 42h later by FACS. Data are representative of 2 independent experiments.

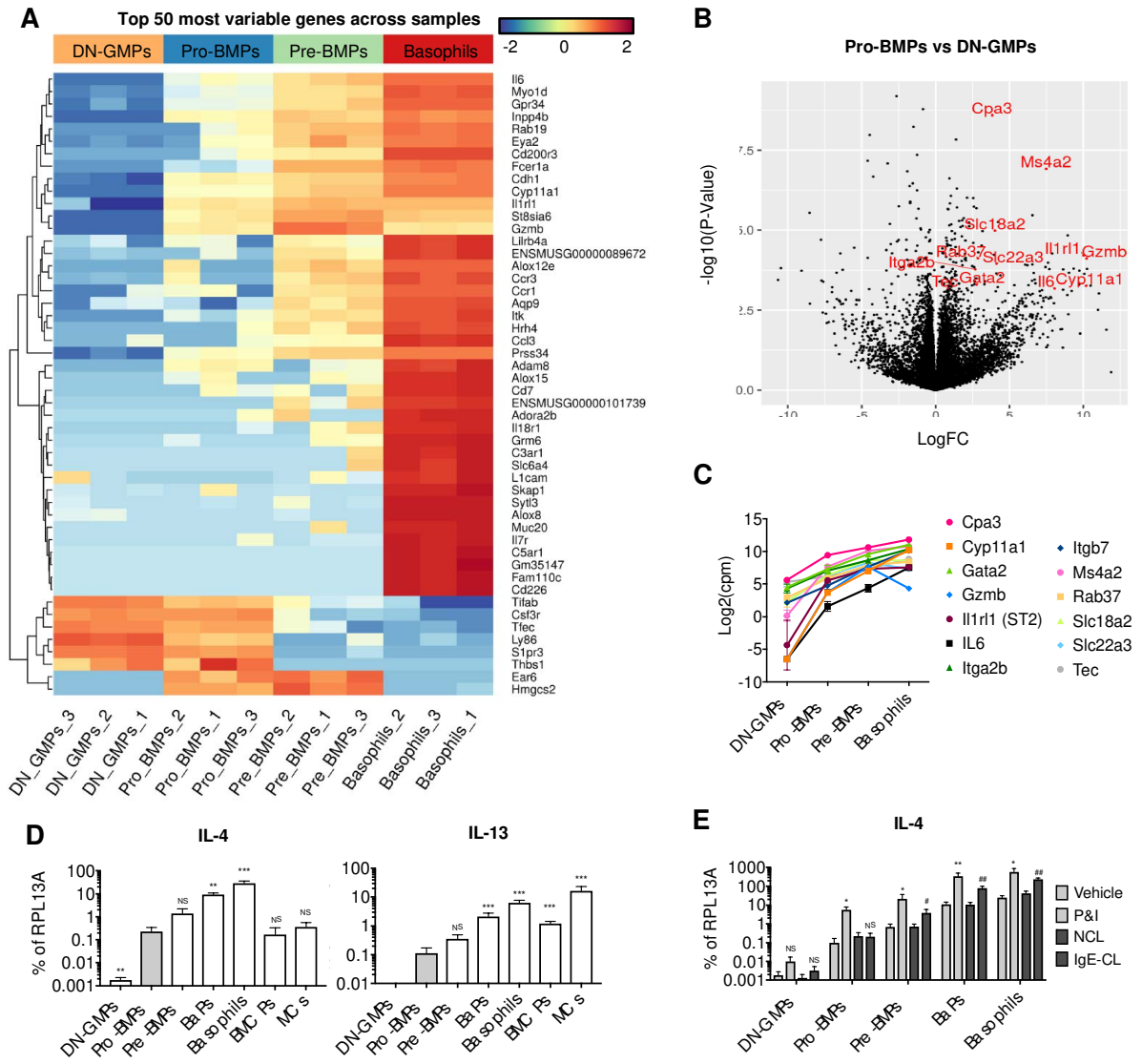

Fig. 5. Pro-BMPs are transcriptionally primed toward a basophil/MC fate and express Th2 cytokines but are not responsive to IgE stimulation. **(A-C)** RNA-seq was performed on FACS-sorted DN-GMPs, pro-BMPs, pre-BMPs, and mature basophils. **(A)** Heatmap showing the top 50 most variable genes between the compared populations. **(B)** Volcano plot showing differential gene expression between pro-BMPs and DN-GMPs. Selected basophil and MC markers are highlighted in red. **(C)** Expression profile of selected genes that were significantly induced in pro-BMPs vs DN-GMPs. **(D-E)** The indicated cell

populations were FACS-sorted. **(D)** mRNA abundance of the indicated cytokines was determined by RT-qPCR. Statistics were calculated using one-way ANOVA followed by multiple comparisons using Dunnett's tests. **(E)** Cells were stimulated for 4 h with PMA and ionomycin (P&I) (or vehicle) or with IgE crosslinking (or not crosslinked (NCL)) and IL-4 mRNA abundance was measured (mean ± SEM of n=3-5). Statistics were calculated using t-tests for the P&I vs vehicle and CL vs NCL comparisons.

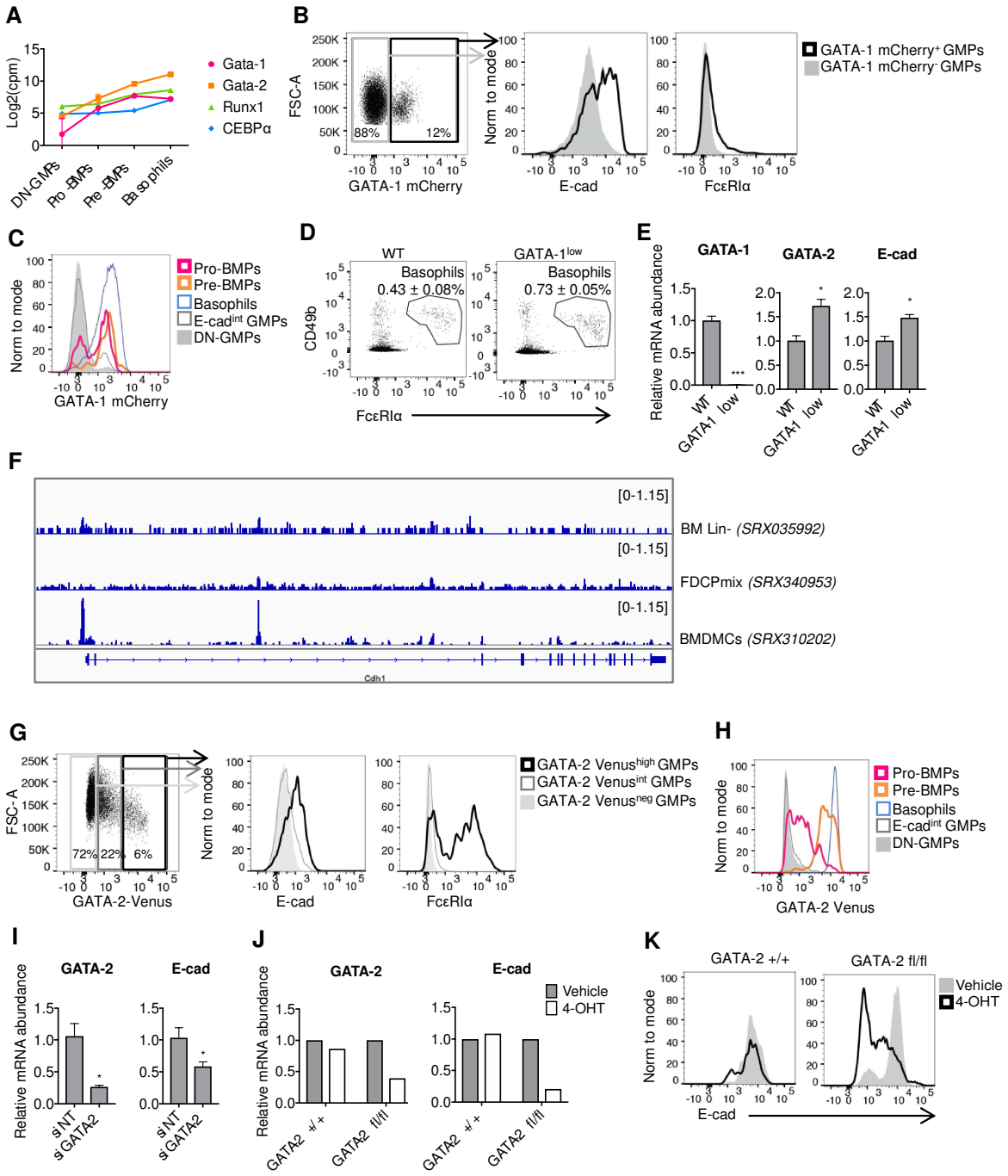

Fig. 6. E-cadherin is induced in the basophil and MC lineages downstream of the GATA-2 transcription factor. **(A)** Expression profile of selected transcription factors involved in basophil and MC differentiation (RNA-seq). **(B)** Flow cytometry analysis of GATA-1mCherry, E-cadherin and FcεRIα on gated GMPs from a GATA-1mCherry

reporter mouse. **(C)** Expression intensity of GATA-1mCherry in the indicated populations. **(D)** Representative dot plots showing the bone marrow basophil populations from wild-type and GATA-1^{low} mutant mice (numbers are the average of n=3, SEM). **(E)** Relative mRNA abundance of GATA-1, GATA-2 and E-cadherin from FACS-sorted basophils from wild-type and GATA-1^{low} mutant mice (n=3, SEM). **(F)** GATA-2 ChIP-seq profiles at the E-cadherin locus. **(G)** Flow cytometry analysis of GATA-2Venus, E-cadherin and FcεRIα on gated GMPs from a GATA-2Venus reporter mouse. **(H)** Expression intensity of GATA-2Venus in the indicated populations. **(I)** BaPs were FACS-sorted and nucleofected with non-targeting siRNA (siNT) or siRNA against GATA-2 (siGATA-2). Relative mRNA abundance of GATA-2 and E-cadherin was analyzed 48h post-nucleofection (n=4, SEM). **(J-K)** Pre-BMPs were FACS-sorted from the bone marrow of GATA-2+/+ TgCreErt2hemi and GATA-2fl/fl TgCreErt2hemi mice and GATA-2 deletion was induced by the addition of 50 nM 4-hydroxytamoxifen (4-OHT) for 5 days before analysis of **(J)** GATA-2 and E-cadherin mRNA abundance and **(K)** E-cadherin protein expression by FACS. Statistics were calculated using t-tests.