IRIS AperTO

UNIVERSITÀ DEGLI STUDI DI TORINO

AperTO - Archivio Istituzionale Open Access dell'Università di Torino

Intragenic DNA methylation prevents spurious transcription initiation.

This is the author's manuscript

Original Citation:

Availability:

This version is available http://hdl.handle.net/2318/1634241 since 2017-05-17T20:15:43Z

Published version:

DOI:10.1038/nature21373

Terms of use:

Open Access

Anyone can freely access the full text of works made available as "Open Access". Works made available under a Creative Commons license can be used according to the terms and conditions of said license. Use of all other works requires consent of the right holder (author or publisher) if not exempted from copyright protection by the applicable law.

(Article begins on next page)

28 April 2024

Intragenic DNA methylation prevents spurious transcription initiation

Francesco Neri[1,3], Stefania Rapelli[2], Anna Krepelova[1,2], Danny Incarnato[1], Caterina Parlato[1], Giulia Basile[1], Mara Maldotti[1,2], Francesca Anselmi[1,2], and Salvatore Oliviero[1,2].

1) Human Genetics Foundation (HuGeF), via Nizza 52, 10126 Torino, Italy.

2) Dipartimento di Scienze della Vita e Biologia dei Sistemi, Università di Torino, via Accademia Albertina 13, 10123 Torino, Italy.

3) Leibniz Institute on Aging – Fritz Lipmann Institute (FLI), Beutenbergstrasse 11, 07745 Jena, Germany.

Corresponding Authors:

Salvatore Oliviero

Email: salvatore.oliviero@unito.it

Francesco Neri

Email: francesco.neri@leibniz-fli.de

SUMMARY

In mammals, DNA methylation occurs mainly at CpG dinucleotides. The methylation on the promoter leads to the suppression of gene expression, while the functional role of gene body DNA methylation in highly expressed genes has yet to be clarified.

Here, we show that the Dnmt3b-dependent intragenic DNA methylation protects the gene body from RNA Polymerase II spurious entry and cryptic transcription initiation. Using different genome-wide approaches, we demonstrate that this Dnmt3b function is dependent on its enzymatic activity and recruitment on the gene body by H3K36me3. Furthermore, the spurious transcripts can be either degraded by the RNA exosome complex or capped, polyadenylated, and delivered to the ribosome to produce aberrant proteins.

Thus, elongating RNA Polymerase II triggers an epigenetic crosstalk that involves SetD2, H3K36me3, Dnmt3b, and DNA methylation to ensure gene transcription initiation fidelity with implications for intragenic hypomethylation in cancer.

DNA methylation of cytosine residues on CpGs is a heritable epigenetic modification crucial for mammalian development that involves the coordinated processes of DNA methylation, demethylation, and maintenance of the methylated cytosine[1-4]. *De novo* establishment of DNA methylation is regulated by the DNA methyltransferases Dnmt3a and Dnmt3b alone or in a complex with Dnmt3l, whereas DNA methylation maintenance is mediated by Dnmt1[5-9].

The methylation of gene promoters is associated with gene silencing, while the function of gene body DNA methylation has not yet been clarified. Recent studies have reported that Dnmt3b binds preferentially to the gene bodies by interacting with the histone modification H3K36me3[10,11]. In this study, we took advantage of Dnmt3b specificity to target intragenic DNA methylation with the aim to clarify the function DNA methylation within the gene body.

Dnmt3b loss reduces gene body DNA methylation

To gain insights into the functional role of Dnmt3b-dependent intragenic DNA methylation, we generated two independent Dnmt3b knockout cell lines from mouse embryonic stem cell (ESC) line E14 to exclude experimental artefacts due to different genetic backgrounds, or prolonged cell culturing (Extended Data Fig. 1a-f). Next we investigated the distribution of the endogenous Dnmt3b. ChIP-seq analysis revealed the intragenic binding of the endogenous Dnmt3b especially on the genes belonging to the third and fourth quartiles (q3-q4) of gene expression and correlated with H3K36me3 histone modification (Fig. 1a-c and Extended Data Fig 1g-q). Whole genome bisulfite sequencing (WGBS) in WT or Dnmt3b$^{-/-}$ ESCs revealed a global reduction of genomic DNA methylation, with a significant decrease of the level of 5-methylcytosine (5mC) on exons and introns (Fig. 1d and Extended Data Fig. 1r). Genome splitting in deciles of H3K36me3 occupancy showed that both Dnmt3b binding and DNA methylation loss in Dnmt3b$^{-/-}$ cells markedly correlate with the abundance of H3K36me3 (Fig. 1e).

Analysis of intragenic spurious transcription

During the transcriptional elongation, the RNA polymerase II (RNA Pol II) recruits the enzyme SetD2 across the transcribed regions for H3K36 trimethylation[12,13], which has been demonstrated to maintain a repressive chromatin environment to prevent spurious entry of the RNA Pol II[14]. In yeast, the repressive action of H3K36me3 is mediated by the recruitment of histone deacetylases (HDACs) that make the chromatin inaccessible[15]. Silencing of SetD2 in mammalian cells leads to cryptic transcription initiation of a large fraction of active genes. However, no change in histone acetylation levels occurs after SetD2 knockdown[12,14,16].

To investigate the functional role of the Dnmt3b-dependent DNA methylation on gene bodies, we performed a high-coverage total RNA-seq analysis in WT and Dnmt3b$^{-/-}$ ESCs (Extended Data Fig. 2a, b). The occurrence of cryptic intragenic transcription initiation was measured by the ratio between the RPKM of the intermediate exons and the RPKM of the first exon. Dnmt3b$^{-/-}$ cells displayed a significantly higher ratio starting from the second exon (Fig. 2a). 1445 genes (18% of total genes having RPKM > 1) had a log2 ratio of intermediate exons versus first exon > 1 (Fig. 2b). Analysis performed on technical and biological replicates, and validation by RT-qPCR on a subset of genes confirmed this result (Extended Data Fig. 2c-e). This indicates that in Dnmt3b$^{-/-}$ cells, a significant amount of RNAs are transcribed starting within the gene body.

Intragenic DNA methylation can regulate alternative promoters[17]. To test this hypothesis we investigated the promoter usage of genes having two or more annotated alternative promoters. Even though some alternative promoters'

reactivation events take place in Dnmt3b$^{-/-}$ cells, the Dnmt3b loss did not show evidence of a general reactivation of these events at genome-wide level (Extended Data Fig. 3).

Activation of cryptic intragenic transcription initiation events can be a consequence of RNA Pol II spurious entry on the gene body. To test this, we performed ChIP-seq analysis of RNA Pol II and H3K36me3 in WT and Dnmt3b$^{-/-}$ cells. To distinguish between the engaged and the elongating RNA Pol II, we treated the cells with 5,6-Dichloro-1-β-D-ribofuranosylbenzimidazole (DRB), a chemical compound that inhibits transcription elongation[18 19]. We then performed ChIP-seq using two different antibodies recognizing pan RNA Pol II or the RNA Pol II phosphorylated on Serine 5 (Ser5), both of which are able to immunoprecipitate stalled RNA Pol II (Extended Data Fig. 4a-c). We did not observe any significant changes in H3K36me3 and RNA Pol II genomic profiles when comparing untreated WT to Dnmt3b$^{-/-}$ cells (Fig. 2c, d and Extended Data Fig. 4d-f). In contrast, we observed a significant increase of RNA Pol II binding on intragenic regions when comparing Dnmt3b$^{-/-}$ to WT cells treated with DRB. This was especially true on introns and exons of q3-q4 genes using both antibodies (Fig. 2c, d and Extended Data Fig. 4f) with a small but detectable increment of H3K4me3 and H3ac on the q4 genes (Extended Data Fig. 4g-j). These data strongly support the hypothesis that Dnmt3b is necessary to prevent intragenic spurious transcription initiation by repressing RNA Pol II entry downstream the canonical promoters.

To confirm the increase of intragenic transcription activation in Dnmt3b$^{-/-}$ cells, we performed an RNA immunoprecipitation with a CAP-specific antibody[20] that was followed by high-throughput sequencing (CAPIP-seq) (Extended

Data Fig. 5a-e). Application of this technique to total RNA, showed a significant enrichment of CAP signal in intermediate intronic and exonic regions of Dnmt3b$^{-/-}$ transcripts (Extended Data Fig. 5f-g).

Since the Pol II ChIP-seq and CAPIP-seq are affinity-purification methods, which suffer of high backgrounds and lack the resolution power to map the start sites at single-base resolution, we performed a high-throughput identification of the transcriptional start sites at single-base resolution in WT and Dnmt3b$^{-/-}$ cells using the RNA 5' Pyrophosphohydrolase (RppH) enzyme to decap eukaryotic mRNAs, leaving a 5´ monophosphate group[21] that was selectively used for adapter ligation to map the CAP signals, transcriptome-wide (DECAP-seq) (Extended Data Fig. 6a-e). DECAP-seq analysis in WT and Dnmt3b$^{-/-}$ cells revealed a significant increase of the transcription start sites (TSSs) in Dnmt3b$^{-/-}$ cells with respect to WT cells on the gene bodies of the q3 and q4 genes (Fig. 3a, b and Extended Data Fig. 6f, g). Since the average RPM value on each single-base TSS on annotated TSSs is around 6 (Extended Data Fig. 6h-j), we further analysed the single-base intragenic TSSs having RPM > 6. This analysis identified 2627 highly expressed TSSs (RPM > 6) specific of the Dnmt3b$^{-/-}$ cells, 780 TSSs common in both the cell lines and 936 specific of WT cells (Fig. 3c). The percentage of the total mapped reads specific for Dnmt3b$^{-/-}$ TSSs was 2.76%, while the percentage of common TSSs was 5.22% suggesting that these latter represent canonical TSSs not present in the used gene annotation (Fig. 3d). The number of intragenic TSSs and the reads distribution was similar in both the DECAP-seq replicates showing high overlap (Extended Data Fig. 6k-m) and were confirmed by significant enrichment of CAPIP-seq and Pol II ChIP-seq as well

as by an increased RNA-seq ratio between downstream vs upstream exons in the KO cells (Extended Data Fig. 16n-o).

The intragenic TSSs identified in Dnmt3b$^{-/-}$ cells were within genomic regions with significantly higher H3K36me3 and Dnmt3b binding, showed loss of methylation upon Dnmt3b depletion, and lower nucleosome occupancy with respect to randomly chosen intragenic regions (Fig. 3e). The latter observation is in agreement with the recent finding on Dnmt3b enzymatic preference to the linker DNA[10,11].

To get an insight of the mechanism that generates the cryptic intragenic transcripts, we next analysed the sequence context where the spurious intragenic TSSs are generated. It has been previously shown that mammalian transcription preferentially starts with a pyrimidine (C/T) at position -1 and with a purine (A/G) at position +1[22]. Our data confirm this observation on canonical TSSs, while on intragenic TSSs we observed the loss of the pyrimidine enrichment at position -1 and a reduced enrichment of the purine at position +1 (Fig. 3f). Interestingly, within a region of 50 bp centred at the Dnmt3b-dependent intragenic TSSs, we found a statistically significant enrichment of the CpG dinucleotide and several transcription factor binding motifs containing CpG sequence, including SP1 and members of the ETS family (Fig. 3g, h and Extended data Fig. 7a). Since both SP1 and ETS proteins can recruit the transcription machinery in TATA-less promoters, and their binding to the DNA is affected by CpG methylation[23-28], we verified the presence of their motif on some example TSSs identified by the DECAP-seq analysis (Fig. 3i and Extended Data Fig. 7b).

First, we confirmed the presence of these cryptic TSSs with targeted CAP-immunoprecipitation and RNA Pol II ChIP experiments as well as with RT-qPCR by measuring the RNA level of the downstream/upstream exon of the intragenic TSSs and validated the differential methylation between WT and Dnmt3b$^{-/-}$ cells by using bisulfite Sanger sequencing (Extended Data Fig. 7c, d). We observed the enrichment of Tbp, TfIIb, Sp1, Elk1 and/or Elf1 by ChIP experiments on the cryptic TSSs, but not in the control region (Extended Data Fig. 7e).

Crosstalk between H3K36me3 and DNA methylation

To demonstrate the involvement of the H3K36me3 histone mark in this molecular event, we silenced SetD2 and measured the activation of cryptic TSSs in WT and Dnmt3b$^{-/-}$ cells. SetD2 knockdown resulted in a drastic reduction of H3K36me3 both in WT and Dnmt3b$^{-/-}$ cells and loss of Dnmt3b intragenic binding in WT cells (Fig. 4a-c and Extended Data Fig. 8a-f). The intermediate/first exon ratio in SetD2 silenced with respect to the control cells estimated by total RNA-seq showed an increase of spurious transcripts that was comparable to the increase measured in Dnmt3b$^{-/-}$ versus WT cells (Fig. 4d, e and Extended Data Fig. 8 g, h). DECAP-seq revealed a significant increase of TSSs in SetD2 silenced cells with respect to control cells on the gene bodies of the transcripts belonging to the genes q3 and q4 (Extended Data Fig. 8i-k). This analysis revealed 3560 high confident (RPM > 6) intragenic TSSs in SetD2 silenced cells, of which 2759 were specific of the SetD2 knockdown and more than 2000 were in common with the TSSs

identified in Dnmt3b$^{-/-}$ cells (Extended Data Fig. 8l-o and 9a). Thus, Dnmt3b-mediated inhibition of spurious transcription initiation is dependent on the presence of H3K36me3.

Next we investigate whether this mechanism depends on the catalytic activity of Dnmt3b, we expressed either the full-length WT Dnmt3b enzyme or the catalytically inactive mutant (V725G)[29] form into Dnmt3b$^{-/-}$ cells (Fig. 4f). Both the Dnmt3b proteins showed similar intragenic binding profiles (Fig. 4g, h and Extended Data Fig. 9b, c), although only the cells expressing the WT protein showed an increase of DNA methylation levels (Extended Data Fig. 9d, e). Total RNA-seq (Extended Data Fig. 9f, g) revealed a significant reduction of the intermediates/first exon ratio in the cells expressing the WT Dnmt3b enzyme, but not in those expressing Dnmt3bV725G mutant (Fig. 4i, j). To further confirm the importance of the H3K36me3-mediated recruitment of the active Dnmt3b on the gene bodies we also expressed mutants (S277P and VW-RR) unable to bind H3K36me3 (Extended Data Fig. 9h)[10]. These Dnmt3b mutants showed lower intragenic binding with reduced DNA methylation activity (Extended Data Fig. 9i, j) and were unable to reduce the intermediates/first exon ratio (Extended Data Fig. 10a-d). Thus, only the re-expression of the WT Dnmt3b was able to rescue the loss of spurious intragenic transcription initiation in Dnmt3b$^{-/-}$ cells demonstrating that Dnmt3b-dependent DNA methylation is responsible for preventing the occurrence of cryptic intragenic transcription.

Analysis of spurious transcripts

Next, we analysed the fate of the cryptic RNAs generated by spurious intragenic initiation events. To this end we first analysed whether these RNAs are degraded. Silencing of the Dis3 and Rrp6 component of the RNA exosome complex did not show changes in the total number of intragenic TSSs (Extended Data Fig. 10e-g), but a significant increase of their expression level (Extended Data Fig. 10h-k). Thus, indicating that a fraction of the cryptic transcripts generated in Dnmt3b$^{-/-}$ cells is degraded by the RNA exosome complex. Although analysis of the polyA+ transcripts by RNA-seq (Extended Data Fig. 10l, m) showed an increase of the intermediate/first exon ratio in Dnmt3b$^{-/-}$ cells (Fig. 5a, b and Extended Data Fig. 10n, o) showing that the intragenic transcripts are, at least in part, polyadenylated.

To investigate the stability of these aberrant RNAs, and to follow their fate, we performed DECAP-seq from PolyA-enriched and cytoplasmic RNA. We found a significant reduced number of intragenic TSSs in PolyA+ and cytosolic DECAP-seq with respect to DECAP-seq starting from total RNA (Extended Data Fig. 10p, q) confirming that a fraction of the spurious transcripts is actually degraded and indicating that not all cryptic transcripts undergo polyadenylation and nuclear export. Interestingly, almost all the highly expressed TSSs identified in Dnmt3b$^{-/-}$ cells can be found in the PolyA+ and cytosolic DECAP-seq fractions (Fig. 5c), even though, with a significantly lower level of expression with consequent reduction of the percentage of the total mapped reads (Fig. 5d, e), while highly expressed common TSSs do not globally change their expression among the different cellular compartments (Extended Data Fig. 10r, s).

To further measure the stability of the PolyA+ cryptic RNAs, we estimated the mRNA half-life in WT and Dnmt3b$^{-/-}$ cells by performing polyA+ RNA-seq analyses in DRB-treated cells at different time points[30] (Extended Data Fig. 11a-b). The intronic RNA inclusion in Dnmt3b$^{-/-}$ cells remained always higher than in WT cells along the time course (Fig. 5f) and the intronic RNA half-life slightly increased in Dnmt3b$^{-/-}$ cells (Extended Data Fig. 11d, f) whereas the whole transcriptome half-life did not show any significant difference between WT and Dnmt3b$^{-/-}$ cells (Extended Data Fig. 11c, e, g). Moreover, the intermediates/first exon ratio was higher in Dnmt3b$^{-/-}$ with respect to WT cells at all time points of DRB treatment (Fig. 5g, h), revealing that the spurious RNAs, once polyadenylated, are as stable as the full length RNAs transcribed from canonical TSSs.

To characterize the biological impact of the cryptic transcripts we investigated their post-transcriptional processing by performing mRNA ribosome profiling[31] (Extended Data Fig. 11h, i). Active mRNA Translation sequencing (ART-seq) revealed that Dnmt3b$^{-/-}$ cells displayed a significant reduction of the ribosome occupancy on the 5'-UTR region together with a marked increase in the occupancy of RNA intronic regions, especially on q3-q4 genes (Fig. 5i, j and Extended Data Fig. 11j, k), while we did not observe significant changes in the ribosome enrichment of the coding region of the transcripts between WT and Dnmt3b$^{-/-}$ cells (Fig. 5i and Extended Data Fig. 11j). Thus, suggesting that the increment of RNA Pol II spurious entry and consequent transcription of cryptic RNAs might generate aberrant proteins.

Discussion

DNA methylation takes place both at the promoters and within the gene body. While it is well documented that DNA methylation at the promoters is associated with gene silencing, the function of gene body DNA methylation remains elusive. We now demonstrate that Dnmt3b-dependent DNA methylation on the gene body is responsible for the prevention of aberrant transcription initiation events necessary to guarantee the mRNA transcription initiation fidelity (Extended Data Fig. 12a). Previous studies reported examples of alternative splicing or activation of retroviruses due to intragenic DNA methylation[32-34]. We tested these hypotheses and, in agreement with the literature, we found few events of alternative splicing and activation of repetitive elements actually occurring in Dnmt3b$^{-/-}$ cells (data not shown). However, these regulations were sporadic in our model.

Our work, unveiling the functional role of the epigenetic crosstalk between RNA Pol II, H3K36me3, and DNA methylation (Extended Data Fig. 12b), can explain physiological gene regulation as well as the occurrence of abnormal transcripts in cancer. Indeed, the global DNA hypomethylation, especially occurring on intragenic regions, is a general feature of most tumours[35-37]. Moreover, several recent studies reported the loss/mutation of SETD2 and the consequent loss of H3K36me3 histone mark as a key event in promoting cancer growth and malignancy[38-40]. The present study reveals a novel function of intragenic DNA methylation and provides new aspects that have to be considered in evaluating the roles of SETD2 and DNMT3B in cancer establishment and progression. Partial internal RNAs generated by the loss of transcription initiation fidelity, could impact on multiple biological processes

impairing molecular mechanisms such as regulation of the gene expression, miRNA targeting, truncated proteins generation and others, thus favouring (stochastic) tumour cell heterogeneity and neoplastic predisposition.

METHODS

Cell culture

E14 mouse Embryonic Stem Cells (ESCs) were cultured in high-glucose DMEM (Invitrogen) supplemented with 15% FBS (Millipore), 0.1 mM nonessential amino acids (Invitrogen), 1 mM sodium pyruvate (Invitrogen), 0.1 mM 2-mercaptoethanol, 1500 U/ml LIF (Millipore), 25 U/ml penicillin, and 25 µg/ml streptomycin. The cells were mycoplasma free.

Generation of Dnmt3b$^{-/-}$ ESCs was performed using TALEN technology. Cells were transfected with the two TALEN constructs targeting Exon 17 of murine Dnmt3b (corresponding to the start of the catalytic domain) and after 16 hours were seeded as a single cell. After ten days, clones were screened by western blot analysis. Positive clones were analysed by genomic sequencing.

For half-life measurements and RNA Pol II elongation inhibition, WT and Dnmt3b$^{-/-}$ ESCs were treated with DRB at the concentration of 75 µM for the indicated times.

Protein extraction and Western blotting

For total cell extracts, cells were resuspended in F-buffer (10mM Tris-HCl pH 7.0, 50mM NaCl, 30mM Na-pyrophosphate, 50mM NaF, 1% Triton X-100, anti-proteases) and sonicated for 3 pulses. Extracts were quantified using BCA assay (Pierce) and were run on SDS-polyacrylamide gels at different percentages, transferred to nitrocellulose membranes and incubated with specific primary antibodies overnight.

Nuclear proteins extract were performed as described in[41]. Briefly, cells were harvested in PBS 1x and resuspended in Isotonic Buffer (20mM HEPES pH 7.5, 100mM NaCl, 250mM Sucrose, 5mM $MgCl_2$, 5µM $ZnCl_2$). Successively, cells were resuspended in isotonic buffer supplemented with 1% NP-40 to isolate nuclei. The isolated nuclei were resuspended in Digestion Buffer (50mM Tris-HCl pH 8.0, 100mM NaCl, 250mM Sucrose, 0.5mM $MgCl_2$, 5mM $CaCl_2$, 5µM $ZnCl_2$) and treated with Micrococcal Nuclease (NEB) at 30°C for 10 min.

Immunoprecipitation

Nuclear proteins from about 10 x 10^6 cells were incubated with 3 µg of specific antibody overnight at 4° C. Immunocomplexes were incubated with protein-G- conjugated magnetic beads (DYNAL, Invitrogen) for 2 hours at 4° C. Samples were washed four times with digestion buffer supplemented with 0.1% NP-40 at RT. Proteins were eluted by incubating with 0.4M NaCl TE buffer for 30 min and were analysed by western blotting.

DNA construct and shRNA

Custom shRNAs against SetD2, Dis3 and Rrp6 were constructed using the TRC hairpin design tool (http://www.broadinstitute.org/rnai/public/seq/search), and designed to target the 3'-UTR. shRNAs with more than 14 consecutive matches to non-target transcripts were avoided. Hairpins were cloned into pLKO.1 vector (Addgene: 10878) and each construct was verified by sequencing. Dnmt3b construct was obtained by PCR amplification and cloned into pEF6/V5-His vector (Invitrogen). The Dnmt3b V725G, S277P, and VW-

RR, constructs were generated by introducing a site-specific mutation in the DNA sequence corresponding to Val725 to mutate it into a Glycine, or Ser277 to mutate in Prolin, or Val236Trp237 to mutate in Arg-Arg, using QuickChange XL Site-Directed Mutagenesis Kit (Agilent Technologies).

Transfections

Transfections of mouse ESCs were performed using Lipofectamine™ 2000 Transfection Reagent in according to manufacturer's protocol using equal amounts of each plasmid in multiple transfections. For the SetD2 knockdown, cells were transfected with 5 µg of the specific shRNA construct, and maintained in medium with puromycin selection (1µg/ml) for 48h.

Chromatin Immunoprecipitation assay

To investigate the distribution of the endogenous Dnmt3b we tested different antibodies and found one that was able to immunoprecipitate the endogenous Dnmt3b cross-linked to chromatin, which showed no background signal in Dnmt3b$^{-/-}$ (Extended Data Fig. 1g-i). The ChIP-seq data were validated by ChIP-qPCR, using several biological replicates, on target genomic regions and by not cross-linked co-immunoprecipitation experiments between Dnmt3b and H3K36me3 in WT or Dnmt3b$^{-/-}$ ESCs (Extended Data Fig.1o, p). For Dnmt3b ChIP-seq, approximately 2 x 10^7 cells were cross-linked by addition of formaldehyde to 1% for 10 min at RT, quenched with 0.125 M glycine for 5 min at RT, and then washed twice with cold PBS. The cells were resuspended in Lysis Buffer 1 (50 mM Hepes-KOH pH 7.5, 140 mM NaCl, 1 mM EDTA, 10% Glycerol, 0.5% NP-40, 0.25% Triton X-100 and protease inhibitor) to

disrupt the cell membrane and in Lysis Buffer 2 (10 mM Tris-HCl pH 8.0, 200 mM NaCl, 1 mM EDTA, 0.5 mM EGTA and protease inhibitor) to isolate nuclei. The isolated nuclei were then resuspended in SDS ChIP Buffer (20 mM Tris-HCl pH 8.0, 10 mM EDTA, 1% SDS and protease inhibitors). Extracts were sonicated using the BioruptorH Twin (Diagenode) for 2 runs of 10 cycles [30 sec ‘‘ON’’, 30 sec ‘‘OFF’’] at high power setting. Cell lysate was centrifuged at 12,000 g for 10 min at 4°C. The supernatant was diluted with ChIP Dilution Buffer (20 mM Tris-HCl pH 8.0, 150 mM NaCl, 2 mM EDTA, 1% Triton) before immunoprecipitation step. Magnetic beads (Dynabeads® Rat Anti-Mouse IgM for anti-PolII-phospho-S5, Dynabeads®Protein G for all the other ChIPs, Life Technologies) were saturated with PBS/1% BSA and the samples were incubated with 2 µg of antibody overnight at 4°C on a rotator. Next day samples were incubated with saturated beads for two hours at 4°C on a rotator. Successively immunoprecipitated complexes were washed five times with RIPA buffer (50 mM Hepes-KOH pH7.6, 500 mM LiCl, 1 mM EDTA, 1% NP-40, 0,7% Na-Deoxycholate) at 4°C for 5 min each on a rotator. For other ChIP-seq, ChIP-seq was performed as described[42]. Elution Buffer was added and incubated at 65°C for 15 min. The decrosslinking was performed at 65°C overnight. Decrosslinked DNA was purified using QiaQuick PCR Purification Kit (Quiagen) according to the manufacture’s instruction

MeDIP was performed using MeDIP kit (Active Motif), according to manufacturer’s protocol.

DNA was analysed by quantitative real-time PCR by using SYBR GreenER kit (Invitrogen). All experiment values were normalized to input. The data shown represent triplicate real-time quantitative PCR measurements of the

immunoprecipitated DNA. The data are expressed as (%) 1/100 of the DNA inputs. Error bars represent standard deviation determined from triplicate experiments. Oligonucleotide sequences are reported in TableS1.

DNA extraction and dot-blot analysis

Genomic DNA was extracted from cells using DNeasy Blood and Tissue kit (Qiagen). For dot-blot analysis, extracted genomic DNA was sonicated using the BioruptorH Twin (Diagenode) for 2 runs of 10 cycles [30 sec ‘‘ON’’, 30 sec ‘‘OFF’’] at high power setting, in order to obtain 300 bp fragments, denatured with 0.4 M NaOH and incubated for 10 min at 95°C prior to being spotted onto HybondTM-N+ (GE Healthcare). Membranes were saturated with 5% milk and incubated with the specific antibodies overnight.

ChIP-seq library preparation

Approximately 10 ng of purified ChIP DNA were end-repaired, dA-tailed, and adapter-ligated using the NEBNext ChIP-seq Library Prep Master Mix Set (NEB), following manufacturer instructions.

Whole genome bisulfite-seq library preparation

For whole-genome Bisulfite-Seq library preparation, 2.5 μg of ESCs genomic DNA, were spiked-in with 1ng of Escherichia coli genomic DNA, and sheared using a Bioruptor® Twin sonicator (Diagenode) for 3 runs of 10 cycles [30 sec “ON,” 30 sec “OFF”] at high power setting. Fragmented/digested DNA was then end-repaired, dA-tailed, and ligated to methylated adapters, using the Illumina TruSeq DNA Sample Prep Kit, following manufacturer instructions.

DNA was loaded on EGel Size select 2% agarose pre-cast gel (Invitrogen), and a fraction corresponding to fragments ranging from 180bp to 350bp was recovered. Purified DNA was then subjected to bisulfite conversion using the EpiTect Bisulfite Kit (Qiagen). Bisulfite-converted DNA was finally enriched by 15 cycles of PCR using Pfu Turbo Cx HotStart Taq (Agilent).

RNA extraction and RT-PCR analysis

Total RNA was extracted as previously described[43] by using TRIzol reagent (Invitrogen). Real-time PCR was performed using the SuperScript III Platinum One-Step Quantitative RT-PCR System (Invitrogen) following the manufacturer's instructions.

RNA-seq library preparation

Ribo- RNA-seq library preparation was performed as described[44]. Briefly, 2.5 µg of total RNA were depleted of ribosomal RNA using the RiboMinus Eukaryote System v2 kit (Invitrogen), following manufacturer instructions. Ribo- RNA was resuspended in 17 µl of EFP buffer (Illumina), heated to 94°C for 8 min, and used as input for First Strand synthesis, using the TruSeq RNA Sample Prep kit, following manufacturer instructions. PolyA RNA-seq library was performed by using the TruSeq RNA Sample Prep kit, following manufacturer instructions.

CAPIP-seq immunoprecipitation and library preparation

For immunoprecipitation of mRNA for CAP-Seq experiments, 30 µg of total RNA were fragmented by alkaline hydrolysis in ~200nt fragments and

incubated with 5 µg of mouse anti-CAP antibody (Anti-m3G-cap, m7g-cap, Clone H20, Millipore MABE419) (or IgG) overnight at 4°C in 0.5 ml of IP Buffer (10mM Tris-HCl pH 7.5; 150mM NaCl; 0,1% Triton X-100) supplemented with 50 U/ml RNaseOUT (Invitrogen), 50 U/ml SuperaseIN (Invitrogen), and 50 U/ml RNase Inhibitor (Ambion). 25 µl of Dynabeads Protein G (Invitrogen) were saturated overnight at 4°C in IP Buffer supplemented with 150 µg of Sonicated Salmon Sperm DNA (Qiagen). Following incubation, beads were washed two times in IP Buffer and incubated with the preformed RNA-antibody complexes at 4°C. After 3hrs, beads were washed 4 times with IP Buffer. Specific elution of recovered fragments were obtained by incubation of beads with 100 µl Elution Buffer (5mM Tris pH7.5; 1mM EDTA; 0.05% SDS; 0.3 mg/ml Proteinase K) for 1h 30 min at 50°C. Fragments were then purified by addition of 1ml of TRIzol reagent (Invitrogen), and subjected to random-primed reverse transcription using the SuperScript III Reverse Trancriptase (Invitrogen) at 50°C for 1 h. Resulting cDNAs were then used as input for the TruSeq RNA Sample Prep kit (Illumina), starting from the “Second Strand Synthesis” step, to produce the sequencing library, following manufacturer instruction.

DECAP-seq library preparation

To map the transcriptional start sites at single-base resolution we used an enzymatic-based approach by the use of the RNA 5' Pyrophosphohydrolase (RppH) enzyme to decap eukaryotic mRNAs[21]. We validated the specificity of this technique in a pilot experiment by comparing RppH treated RNA versus

untreated or T4 Polynucleotide Kinase (PNK) treated RNA (Extended Data Fig. 6a-e).

When required the total RNA was depleted from small nuclear RNAs (snRNAs) by using the following protocol. 5 µg of Total RNA was resuspended in snRNA-depletion buffer (20mM HEPES pH 7.5, 80mM KCl, 1 mM DTT), 1 µl RNase Inhibitor (Ambion), 2 µM oligo mix (designed against snRNAs sequences, primers sequences in TABLE 1) in a final volume of 50 µl, heated to 70°C for 5 min and immediately put in ice. After that it was added 25 µl snRNA-depletion buffer 2x (40mM HEPES pH 7.5, 160mM KCl, 10 mM $MgCl_2$, 2 mM DTT), supplemented with 1 µl RNase Inhibitor (Ambion) and 1 µl of RnNAse H (NEB) to a final volume of 100 µl. Incubated for 30 min at 37°C. snRNA-depleted RNA were purified by RNA Clean and Concentration kit (Zymo Research) and DNAseI digestion was performed following manufacturer instructions. snRNA-depleted RNAs were further depleted from ribosomal RNA by using the RiboMinus Eukaryote System v2 kit (Invitrogen).

The RNA obtained from previous depletions (or polyA+ RNA enriched using The NEBNext Poly(A) mRNA Magnetic Isolation Module kit (NEB), following manufacturer instructions) was chemically fragmented by using First Strand buffer of the SuperScript® II Reverse Transcriptase (Invitrogen). The fragmented RNA was Dephosphorylated of natural 5' and fragmentation-derived 3' phosphate by using Antarctic Phosphatase (AP, NEB). Dephosphorylated RNA was then treated with RNA 5´ Pyrophosphohydrolase (RppH, NEB) in 1X Thermopol buffer (NEB) (for decapping and pyrophosphate removal from the 5´ end of RNA to leave a 5´ monophosphate

RNA). For positive and negative control, the dephosphorylated RNA was treated with the T4 Polynucleotide Kinase (PNK, NEB) (for 5' phosphorylation of all RNA fragments) or was performed without adding the enzyme. 5' RNA adapter ligation was carried out by using the TruSeq Small RNA Sample Preparation Kit (Illumina). Reverse transcription was performed with SuperScript III enzyme (Invitrogen) and Illumina 3' Adapter Rev-Comp Random Hexamers (RC3N6). The RNA was size selected on TBE-Urea 10% PAGE gel and PCR amplification was carried out by using the TruSeq Small RNA Sample Preparation Kit (Illumina).

ART-seq library preparation

Ribosome profiling was performed using the ARTseq/TruSeq Ribo Profile (Illumina), with minor changes to the manufacturer protocol. Briefly, ~30x10^6 cells were treated with 0.1µg/µl final Cycloheximide for 5 min at 37°C. Cells were then washed twice and harvested with ice-cold PBS (supplemented with 0.1 µg/µl final Cycloheximide). Cells were lysed in 1ml of Mammalian Lysis Buffer (supplemented with 0.5 % final concentration of NP-40) at 4°C for 10 min on a rotator. The lysate was then treated with 50 U of ART-seq Nuclease for 45 min at 25°C, with moderate shaking. 400 µl of the digested lysate were then layered on the top of a 2.5 ml sucrose cushion, and centrifuged at 265,000 x *g* for 5 h at 4°C. After completely removing the supernatant, the pellet was resuspended in 100µl nuclease-free water, and purified on RNA Clean & Concentrator™-5 columns (Zymo Research). 5 µg of the recovered monosomes RNA was then subjected to 2 consecutive rounds of rRNA depletion using the Ribo-Zero™ Gold Kit (Human/Mouse/Rat, Epicentre), and

then run on a 10% TBE-Urea PAGE gel for 25 min at 200 V. A gel slice corresponding to 28-30 nt was then cut, crushed, and RNA was recovered by passive diffusion at 4°C for 16 h. The eluted RNA fragments were then end-repaired, ligated to the 3' adapter, and reverse transcribed. The cDNA was run on 10 % TBE-Urea PAGE gel for 30 min at 180 V, and a gel slice corresponding to fragments of ~70-80 nt was cut, crushed, and cDNA was recovered by passive diffusion at 37°C for 16h with vigorous shaking. The eluted cDNA was then subjected to circularization, and the final library was obtained by 10 cycles of PCR. The final library was inspected on the Fragment Analyzer™ (Advanced Analytical), revealing a single sharp peak around 150 bp.

Reads mapping and data analysis

Samples were sequenced on the HiScanSQ or Next500 platforms (Illumina). All of the analysed datasets, were mapped to a recently published variant of the mm9 genome assembly, that includes single nucleotide variants from E14 ESCs[45]. Prior mapping, sequencing reads were trimmed from low quality score basis and clipped from the adapter sequence by using FASTX toolkit (http://hannonlab.cshl.edu/fastx_toolkit/). For RNA-seq data analysis, reads were mapped using TopHat v2.0.6[46] and mRNA quantification was performed using Cuffdiff v2.0.2[47]. For ChIP-seq data analysis, reads were mapped Bowtie version 0.12.7[48], reporting only unique hits with up to two mismatches (parameters: -m 1 –v 2). For Bisulfite-seq data analysis, reads were mapped using BSMAP v2.74[49]. Unmapped reads from the first mapping round were trimmed by 10nt at their 5'-end, and 15nt at their 3'-end using fastx_trimmer

tool from the FASTX Toolkit, and subjected to a second round of mapping. Reads failing this second mapping round were mapped to the Escherichia coli str. K-12 substr. DH10B genome (NCBI Accession: NC_010473), in order to estimate bisulfite conversion efficiency.

RNA-seq analysis

RNA-seq correlation analyses were performed by using Pearson correlation coefficient and by plotting RPKM value calculated on RefFlat gene annotation. Intragenic transcription initiation analysis was performed on a non-redundant gene annotation built starting from the RefFlat annotation, by keeping only the longest isoform for each gene, with at least 1 RPKM of expression and at least 5 exons. RPKM on each exon was calculated by counting reads falling in the exon (normalizing on the exon length in kb and on the total mapped reads of the experiment in millions) using custom script and then the ratio was calculated as the Log2 fold-change of second, third and last exons RPKM over the first exon RPKM for each gene. For intermediate/first exons ratio, averages of the RPKM value of all the other exons (from 4^{th} to penultimate) were used. Alternative promoter analysis was performed on a non-redundant gene annotation built starting from the RefFlat annotation by keeping only the genes having at least 2 isoforms transcribed from known different promoters. RPKM of the first exon of the isoforms transcribed from 1^{st} or 2^{nd} alternative promoter and, for those having, from the 3^{rd}, 4^{th} or all the others (from 5^{th} to 12^{th}) alternative promoters, was calculated by custom script. Log2 ratio between the first exons transcribed from the 1^{st} over the 2^{nd} promoter was plotted by using heatscatter function (on R) and correlation was calculated by

using Pearson coefficient. Alternative promoter analysis was calculated on the same reference as above. Log2 ratio was calculated as RPKM value of the first exon transcribed from each classes of different alternative promoter over the RPKM of the whole transcript, in order to normalize differentially expressed genes in WT and Dnmt3b$^{-/-}$ cells.

DECAP-seq analysis

For DECAP-seq only intragenic mapped reads were used for further analysis. We used a RefSeq-based genic reference containing only the annotated longest isoforms and deprived from all the genes overlapping other genes or ncRNAs on the same strand. Since DECAP-seq is a single base resolution technique and the first base of the sequenced reads corresponds to the base having the cap signal, only the first position of the mapped read was used to calculate a count per million of mapped reads (RPM). All the analyses were performed on the genes belonging to the q3 or q4 quartiles of expression. Venn diagram overlap is calculated at single-base resolution. Logo analysis of the sequence enrichment was performed by using WebLogo (http://weblogo.berkeley.edu/). Motif discovery was performed by using HOMER Motif Analysis (http://homer.salk.edu/homer/motif/).

CAPIP-seq analysis

For CAPIP-seq only intragenic mapped reads were used for further analysis. RPKM of each genomic feature were calculated as described above by using custom script. Enrichment was calculated as the Log2 fold change of RPKM value from CAP immunoprecipitated samples over the RPKM from Input

samples for each genomic feature. As for DECAP-seq, Intragenic CAPIP-seq signal ratio between wt and Dnmt3b$^{-/-}$ cells was calculated as the fold change of the intragenic enrichment (from 2kb downstream TSS to TES) in WT over Dnmt3b$^{-/-}$ cells. Ratio gene-body/TSS was defined as the Log2 fold change of gene-body enrichment (derived from intronic and intermediate exonic regions) over the enrichment calculated on the first 200nt of the transcripts. All the analyses were performed on the genes belonging to the q3 or q4 quartiles of expression.

Half-life analysis

PolyA enriched RNA-seq analyses were performed from RNA derived from DRB-treated wt and Dnmt3b$^{-/-}$ ESCs. For half-life calculation, gene quantifications performed with CuffDiff (see above) were normalized on the average of the top ten genes showing less degradation rate following DRB treatment having at least 10 RPKM in ESCs. Degradation rate has been defined as the ratio of RPKM value of the sample at time 0h of DRB treatment over the average RPKM value of the samples treated for 3, 6 and 12 hours with DRB. The top ten genes are Tmsb10, Mt1, Mt2, Rps14, Rplp2, 4930412F15Rik, Rpl38, Rplp1, Tomm7 and Cox6a1. Only genes having RPKM > 1 were used for further analysis and a constant of 0.1 pseudo-RPKM was introduced to reduce sampling noise. Half-life ($t_{1/2}$) was calculated by using the following formula[50]:

$$t_{1/2} = \frac{\ln 2}{k_{decay}}$$

where k_{decay} is the decay rate constant obtained by fitting data (gene RPKM value for each time points) with an exponential function. Half-life on introns was measured as calculated for mature mRNAs, but gene quantification (RPKM) was performed counting the reads on introns and normalizing for introns length (kb) and for the number of total intragenic mapped reads (milions). For introns and exons quantification, reads were treated as above (see *RNA-seq analysis*).

Ribosome profiling analysis (ART-seq)

Analysis of ART-seq experiments were performed as previously described[31]. Differently from the other sequencing data, for ribosome profiling, only adapter containing reads were used in order to avoid total RNA contamination. Reads were clipped from adapters and mapped on rRNAs and tRNAs. Only reads not mapping on rRNA/tRNA genes were used for downstream analysis. Quantification (RPKM) of the reads derived from different transcript parts or genomic features was performed as described above.

ChIP-seq and WGBS analysis

Following mapping, reads with the same start mapping coordinates were collapsed using custom Perl scripts, and peak calling was performed using MACS version 1.4.1[51].

ChIP-seq signal Log2 enrichment was calculated as previously described[10], with some modifications. Briefly, the mouse genome was partitioned into 500 bp bins. Bins overlapping with satellite repeats and with an insufficient coverage in WGBS (less than 50% of all CpGs covered at least 10x) were

removed resulting in 2,708,724 bins. Signal enrichment was calculated as the Log2 of ChIP-seq/Input RPKM. These whole-genome Log2 enrichment values were used for clustering, correlation, boxplots and scatterplots analysis by using custom scripts. For genomic binning by H3K36me3, the above bins were divided in ten equal-size groups rank-ordered by their Log2 enrichment for H3K36me3.

Heatmap representations of ChIP-seq peaks and plots were performed with respect to annotated RefSeq genes, sorted by their expression level, according to RNA-seq data. Plots of Dnmt3b and H3K36me3 distribution on genes clustered in quartiles of expression revealing an almost identical distribution for both features.

For the analysis of Dnmt3b intragenic binding in Setd2 KD ESCs and Dnmt3b-reexpressing Dnmt3b$^{-/-}$ ESCs, a non-redundant gene annotation was built starting from the RefFlat annotation, by keeping only the longest isoform for each gene. After calling H3K36me3 peaks in WT ESCs using MACS 1.4.1 (parameters: -p 1e-8 --nolambda), the genes from the RefFlat annotation that overlap an H3K36me3 peak were marked as H3K36me3-positive, while genes lacking any overlap were marked as H3K36me3-negative. For each gene in the two datasets, the normalized Dnmt3b signal (RPKM) in control and treated ESCs was calculated as:

$$\mathrm{RPKM} = 10^9 \times \frac{n}{(TES - TSS) \times N}$$

where n is the number of Dnmt3b reads overlapping gene's coordinates, TSS and TES are respectively the start and end coordinate of the gene annotation, and N is the total number of mapped reads in the ChIP-seq experiment. P-

values were calculated using a one-tailed paired Wilcoxon rank sum test statistics.

Methylation calling was performed using the methratio.py script provided with the BSMAP tool and comparative analyses were performed by using only CpG covered at least 5x in both wt and Dnmt3b$^{-/-}$ cells.

Heatmaps and comparative analysis were performed using custom Perl scripts. Datasets used for comparative analysis were obtained from Gene Expression Omnibus by downloading the following datasets GSE12241, GSE11172, GSE31039, GSE44642, GSE44566, GSE55660, GSE57413, GSE44566).

Antibodies

Antibodies were purchased from Abcam (anti-Dnmt3b; anti-H3K36me3; anti-single strand DNA; anti-H3 pan; anti-Tbp; anti-TIIb), from Imgenex (anti-Dnmt3a; anti-Dnmt3b; anti-Dnmt1), from Diagenode (anti-5-methylcytidine), from Millipore (anti-H3K27me3; Anti-m3G-cap, m7G-cap; anti-Elk1), from Upstate (anti-H3K4me3), from Covance (RNA PolII-phosphoSer5), from SantaCruz (pan RNA PolII, anti-Sp1; anti-Elf1), from Upstate (anti-H3K4me3; anti-H3ac)*.* Antibody anti-Dnmt3L was kindly provided from Dr. S. Yamanaka (Kyoto University, Japan).

References

1. Robertson, K. D. & Robertson, K. D. DNA methylation, methyltransferases, and cancer. *Oncogene* **20,** 3139–3155 (2001).
2. Chen, Z. X. & Riggs, A. D. DNA Methylation and Demethylation in Mammals. *Journal of Biological Chemistry* **286,** 18347–18353 (2011).
3. Neri, F. *et al.* Single-Base Resolution Analysis of 5-Formyl and 5-Carboxyl Cytosine Reveals Promoter DNA Methylation Dynamics.

CellReports **10,** 674–683 (2015).
4. Neri, F. *et al.* TET1 is a tumour suppressor that inhibits colon cancer growth by derepressing inhibitors of the WNT pathway. *Oncogene* **34,** 4168–4176 (2015).
5. Okano, M., Bell, D. W., Haber, D. A. & Li, E. DNA methyltransferases Dnmt3a and Dnmt3b are essential for de novo methylation and mammalian development. *Cell* **99,** 247–257 (1999).
6. Bestor, T. H. The DNA methyltransferases of mammals. *Human Molecular Genetics* **9,** 2395–2402 (2000).
7. Schübeler, D. Function and information content of DNA methylation. *Nature* **517,** 321–326 (2015).
8. Jeltsch, A. & Jurkowska, R. Z. New concepts in DNA methylation. *Trends in Biochemical Sciences* **39,** 310–318 (2014).
9. Neri, F. *et al.* Dnmt3L Antagonizes DNA Methylationat Bivalent Promoters and FavorsDNA Methylation at Gene Bodies in ESCs. *Cell* **155,** 121–134 (2013).
10. Baubec, T. *et al.* Genomic profiling of DNA methyltransferases reveals a role for DNMT3B in genic methylation. *Nature* **520,** 243–247 (2015).
11. Morselli, M. *et al.* In vivo targeting of de novo DNA methylation by histone modifications in yeast and mouse. *eLife* **4,** e06205 (2015).
12. Edmunds, J. W., Mahadevan, L. C. & Clayton, A. L. Dynamic histone H3 methylation during gene induction: HYPB/Setd2 mediates all H3K36 trimethylation. *EMBO J.* **27,** 406–420 (2007).
13. Yoh, S. M. *et al.* The Iws1:Spt6:CTD complex controls cotranscriptional mRNA biosynthesis and HYPB/Setd2-mediated histone H3K36 methylation. *Genes & Development* **22,** 3422–3434 (2008).
14. Wagner, E. J. & Carpenter, P. B. Understanding the language of Lys36 methylation at histone H3. *Nature Reviews Molecular Cell Biology* **13,** 115–126 (2012).
15. Carrozza, M. J. *et al.* Histone H3 methylation by Set2 directs deacetylation of coding regions by Rpd3S to suppress spurious intragenic transcription. *Cell* **123,** 581–592 (2005).
16. Carvalho, S. *et al.* Histone methyltransferase SETD2 coordinates FACT recruitment with nucleosome dynamics during transcription. *Nucleic Acids Research* **41,** 2881–2893 (2013).
17. Maunakea, A. K. *et al.* Conserved role of intragenic DNA methylation in regulating alternative promoters. *Nature* **466,** 253–257 (2010).
18. Maderious, A. & Chen-Kiang, S. Pausing and premature termination of human RNA polymerase II during transcription of adenovirus in vivo and in vitro. *Proc. Natl. Acad. Sci. U.S.A.* **81,** 5931–5935 (1984).
19. Yankulov, K., Yamashita, K., Roy, R., Egly, J. M. & Bentley, D. L. The transcriptional elongation inhibitor 5,6-dichloro-1-beta-D-ribofuranosylbenzimidazole inhibits transcription factor IIH-associated protein kinase. *Journal of Biological Chemistry* **270,** 23922–23925 (1995).
20. Bochnig, P., Reuter, R., Bringmann, P. & Lührmann, R. A monoclonal antibody against 2,2,7-trimethylguanosine that reacts with intact, class U, small nuclear ribonucleoproteins as well as with 7-methylguanosine-capped RNAs. *European journal of biochemistry / FEBS* **168,** 461–467 (1987).

21. Deana, A., Celesnik, H. & Belasco, J. G. The bacterial enzyme RppH triggers messenger RNA degradation by 5' pyrophosphate removal. *Nature* **451,** 355–358 (2008).
22. Carninci, P. *et al.* Genome-wide analysis of mammalian promoter architecture and evolution. *Nat Genet* **38,** 626–635 (2006).
23. Butler, J. E. F. & Kadonaga, J. T. The RNA polymerase II core promoter: a key component in the regulation of gene expression. *Genes & Development* **16,** 2583–2592 (2002).
24. Clark, S. J., Harrison, J. & Molloy, P. L. Sp1 binding is inhibited by (m)Cp(m)CpG methylation. *Gene* **195,** 67–71 (1997).
25. Douet, V., Heller, M. B. & Le Saux, O. DNA methylation and Sp1 binding determine the tissue-specific transcriptional activity of the mouse Abcc6 promoter. *Biochemical and Biophysical Research Communications* **354,** 66–71 (2007).
26. Hogart, A. *et al.* Genome-wide DNA methylation profiles in hematopoietic stem and progenitor cells reveal overrepresentation of ETS transcription factor binding sites. *Genome Research* **22,** 1407–1418 (2012).
27. Uchiumi, F., Miyazaki, S. & Tanuma, S.-I. The possible functions of duplicated ets (GGAA) motifs located near transcription start sites of various human genes. *CMLS, Cell. Mol. Life Sci.* **68,** 2039–2051 (2011).
28. Yu, M. *et al.* GA-binding protein-dependent transcription initiator elements. Effect of helical spacing between polyomavirus enhancer a factor 3(PEA3)/Ets-binding sites on initiator activity. *Journal of Biological Chemistry* **272,** 29060–29067 (1997).
29. Gowher, H., Gowher, H., Jeltsch, A. & Jeltsch, A. Molecular Enzymology of the Catalytic Domains of the Dnmt3a and Dnmt3b DNA Methyltransferases. *Journal of Biological Chemistry* **277,** 20409–20414 (2002).
30. Tani, H. & Akimitsu, N. Genome-wide technology for determining RNA stability in mammalian cells: historical perspective and recent advantages based on modified nucleotide labeling. *RNA Biol* **9,** 1233–1238 (2012).
31. Ingolia, N. T., Lareau, L. F. & Weissman, J. S. Ribosome profiling of mouse embryonic stem cells reveals the complexity and dynamics of mammalian proteomes. *Cell* **147,** 789–802 (2011).
32. Maunakea, A. K., Chepelev, I., Cui, K. & Zhao, K. Intragenic DNA methylation modulates alternative splicing by recruiting MeCP2 to promote exon recognition. *Cell Res.* **23,** 1256–1269 (2013).
33. Yearim, A. *et al.* HP1 is involved in regulating the global impact of DNA methylation on alternative splicing. *CellReports* **10,** 1122–1134 (2015).
34. Jones, P. A. Functions of DNA methylation: islands, start sites, gene bodies and beyond. *Nat Rev Genet* **13,** 484–492 (2012).
35. Jones, P. A. & Baylin, S. B. The fundamental role of epigenetic events in cancer. *Nat Rev Genet* **3,** 415–428 (2002).
36. Gaudet, F. Induction of Tumors in Mice by Genomic Hypomethylation. *Science* **300,** 489–492 (2003).
37. Feinberg, A. P. & Vogelstein, B. Hypomethylation distinguishes genes of some human cancers from their normal counterparts. *Nature* **301,** 89–92 (1983).

38. Kanu, N. *et al.* SETD2 loss-of-function promotes renal cancer branched evolution through replication stress and impaired DNA repair. *Oncogene* **34,** 5699–5708 (2015).
39. Fontebasso, A. M. *et al.* Mutations in SETD2 and genes affecting histone H3K36 methylation target hemispheric high-grade gliomas. *Acta Neuropathol.* **125,** 659–669 (2013).
40. Duns, G. *et al.* Histone methyltransferase gene SETD2 is a novel tumor suppressor gene in clear cell renal cell carcinoma. *Cancer Research* **70,** 4287–4291 (2010).
41. Neri, F. *et al.* Genome-wide analysis identifies a functional association of Tet1 and Polycomb repressive complex 2 in mouse embryonic stem cells. *Genome Biol.* **14,** R91 (2013).
42. Mikkelsen, T. S. *et al.* Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature* **448,** 553–560 (2007).
43. Incarnato, D., Neri, F., Diamanti, D. & Oliviero, S. MREdictor: a two-step dynamic interaction model that accounts for mRNA accessibility and Pumilio binding accurately predicts microRNA targets. *Nucleic Acids Research* **41,** 8421–8433 (2013).
44. Genome-wide profiling of mouse RNA secondary structures reveals key features of the mammalian transcriptome. **15,** 491 (2014).
45. Incarnato, D., Krepelova, A. & Neri, F. High-throughput single nucleotide variant discovery in E14 mouse embryonic stem cells provides a new reference genome assembly. *Genomics* **104,** 121–127 (2014).
46. Kim, D. *et al.* TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol.* **14,** R36 (2013).
47. Trapnell, C. *et al.* Differential analysis of gene regulation at transcript resolution with RNA-seq. *Nat Biotechnol* **31,** 46–53 (2013).
48. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* **10,** R25 (2009).
49. Xi, Y. & Li, W. BSMAP: whole genome bisulfite sequence MAPping program. *BMC Bioinformatics* **10,** 232 (2009).
50. Chen, C.-Y. A., Ezzeddine, N. & Shyu, A.-B. Messenger RNA half-life measurements in mammalian cells. *Meth. Enzymol.* **448,** 335–357 (2008).
51. Zhang, Y. *et al.* Model-based analysis of ChIP-Seq (MACS). *Genome Biol.* **9,** R137 (2008).
52. Li, J. Y. *et al.* Synergistic Function of DNA Methyltransferases Dnmt3a and Dnmt3b in the Methylation of Oct4 and Nanog. *Mol. Cell. Biol.* **27,** 8748–8759 (2007).
53. Core, L. J., Waterfall, J. J. & Lis, J. T. Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. *Science* **322,** 1845–1848 (2008).
54. Sharova, L. V. *et al.* Database for mRNA half-life of 19 977 genes obtained by DNA microarray analysis of pluripotent and differentiating

mouse embryonic stem cells. *DNA Res.* **16,** 45–58 (2009).
55. Schwanhäusser, B. *et al.* Global quantification of mammalian gene expression control. *Nature* **473,** 337–342 (2011).

Figure Legends

Figure 1 | Dnmt3b colocalises with H3K36me3 on the gene body and its loss reduces gene body DNA methylation. a, Binding enrichment of control IgG and endogenous Dnmt3b in WT and Dnmt3b$^{-/-}$ ESCs on the specified genomic features. **b,** Box plot of the endogenous Dnmt3b in ESCs on the introns and exons partitioned in quartiles on the basis of the expression level (q4 = upper quartile, most expressed genes). **c,** Heatmap plots of Dnmt3b and H3K36me3 distribution on genes (see methods). **d,** Violin plot of the methylation level of all CpGs by WGBS. **e,** Genomic 500bp-bins divided in deciles depending on their H3K36me3 enrichment.

Figure 2 | Dnmt3b knockout increases intragenic RNA Pol II spurious entry and transcription initiation events on gene body. a, Box plot of the ratio between normalized RNA-seq read counts (RPKM) in WT and Dnmt3b$^{-/-}$ ESCs. Significance is given by Wilcoxon Rank Sum test statistics. **b,** Pie-chart showing the percentage of transcripts having (intermediate/first exon ratio in Dnmt3b$^{-/-}$ vs. intermediate/first exon ratio in WT ESCs) Log2 fold-change as indicated **c,** Plots of the pan (left panel) or phospho-Ser5 (right panel) RNA pol II distribution in WT and Dnmt3b$^{-/-}$ ESCs treated (or not, left panel) with DRB to inhibit RNA Pol II elongation. **d,** Binding enrichment of the pan (upper

panel) or phospho-Ser5 (bottom panel) RNA pol II in WT and Dnmt3b$^{-/-}$ ESCs treated (or not, only pan RNA Pol II) with DRB on the indicated genic features.

Figure 3 | Dnmt3b is required to maintain transcription initiation fidelity.

a, b, Total TSSs distribution along genes in Dnmt3b$^{-/-}$ compared with WT ESCs. **c,** Venn diagram of intragenic TSSs having DECAP-seq signal > 6 RPM. Overlap is calculated at single-base resolution. **d,** Pie-charts of the DECAP-seq reads distribution on TSSs > 6 RPM in WT and Dnmt3b$^{-/-}$ cells. **e,** Box plot distribution of the IgG, Dnmt3b, and H3K36me3 ChIP-seq signal enrichment as well as of the DNA methylation ratio between Dnmt3b$^{-/-}$ and WT ESCs (from WGBS) and nucleosome occupancy on the novel identified TSSs. **f,** Sequence logo for genomic region sequences of the TSSs identified by DECAP-seq experiments in Dnmt3b$^{-/-}$ and WT ESCs. **g, h,** Histogram plot of observed over expected ratio showing a specific enrichment in CpG dinucleotide in a region around 25nt of Dnmt3b$^{-/-}$ specific TSSs and a significant enrichment for CpG containing motifs in Dnmt3b$^{-/-}$ ESCs specific intragenic TSSs. **i**, Genomic view of the Fn1 and Phb2 genes showing intragenic transcription initiation events. Pie-charts indicate the DECAP-seq reads distribution in Dnmt3b$^{-/-}$ and WT ESCs. Bottom panel, schematic representation of CpG localization and putative transcription factors binding elements. Significance is given by Wilcoxon Rank Sum in panels b and e and by chi-square test statistics in panel g.

Figure 4 | H3K36me3-dependent maintaining of transcription initiation fidelity is mediated by Dnmt3b through its DNA methylation activity.

a, Western blotting analysis of control or SetD2 knockdowns in WT and Dnmt3b$^{-/-}$ ESCs. **b,** Box-plot of normalized Dnmt3b ChIP-seq read counts (RPKM) in control or SetD2 knockdown WT cells on H3K36me3 negative and positive genes. **c,** Dnmt3b distribution on genes in control or SetD2 knockdown cells. **d,** Box plot of the ratio between normalized RNA-seq RPKM in WT and Dnmt3b$^{-/-}$ control or SetD2 knockdown ESCs. **e,** Pie-charts showing the percentage of transcripts having (intermediate/first exon ratio in the indicated treatments vs. intermediate/first exon ratio in WT control knockdown ESCs). **f,** Western blot of Dnmt3b$^{-/-}$ ESCs transfected with mock, Dnmt3b (WT) or inactive Dnmt3b (V725G). **g,** Box-plot of normalized Dnmt3b ChIP-seq RPKM in Dnmt3b$^{-/-}$ ESCs transfected as indicated. **h,** Dnmt3b distribution on genes in Dnmt3b$^{-/-}$ ESCs transfected as indicated. **i,** Box plot of the ratio between normalized RNA-seq RPKM transfected as indicated. **j,** Pie-chart showing the percentage of transcripts having (intermediate/first exon ratio in Dnmt3b (WT or mutant) expressing Dnmt3b$^{-/-}$ ESCs vs. intermediate/first exon ratio in mock Dnmt3b$^{-/-}$ ESCs) Log2 fold-change (FC) as indicated. Significance is given by Wilcoxon Rank Sum in panels b, d. g, and e.

Figure 5 | Transcripts produced from intragenic cryptic starting sites are polyadenylated, stable and ribosome-associated RNAs.

a, Box plot of the ratio between normalized polyA RNA-seq RPKM in WT and Dnmt3b$^{-/-}$ ESCs. **b,** Pie-chart showing the percentage of transcripts having (intermediate/first exon ratio in Dnmt3b$^{-/-}$ vs. intermediate/first exon ratio in WT ESCs) Log2 fold-change as indicated. **c,** Venn diagram of the Dnmt3b$^{-/-}$

specific intragenic TSSs (RPM > 6) in the indicated RNA compartments. **d,** Box plot of the normalized DECAP-seq RPM on the Dnmt3b$^{-/-}$ specific intragenic TSSs in the indicated RNA compartments. **e,** Pie-chart of the DECAP-seq reads distribution. **f,** Box plot of the RPKM counts on introns in time-course cells (WT and Dnmt3b$^{-/-}$) treated with DRB. **g,** Box plot of the ratio between normalized polyA+ RNA-seq RPKM in WT and Dnmt3b$^{-/-}$ ESCs treated 12h with DRB. **h,** Pie-chart showing the percentage of transcripts having (intermediate/first exon ratio in Dnmt3b$^{-/-}$ vs. intermediate/first exon ratio in WT ESCs) Log2 fold-as indicated in cells treated 12h with DRB. **i,** Box plot of the normalized ART-seq RPKMon the indicated RNA regions. **j,** Box plot of the normalized ART-seq RPKM on the indicated classes of introns divided in quartiles in according to the belonged gene expression.

Significance is given by Wilcoxon Rank Sum in panels a, d, g f, i, and j.

Acknowledgments

We thank Tuncay Baubec and Dirk Schübeler (FMI, Basel) for providing the Dnmt3b construct. We thank Shinya Yamanaka (CiRA, Kyoto) for anti-Dnmt3l antibody. We thank Ernesto Guccione (IMCB, Singapore), Raffaele Calogero (MCB, Torino), and Thomas Bates (FLI, Jena) for helpful suggestions and critical reading the manuscript. This work was supported by the Associazione Italiana Ricerca sul Cancro (AIRC) IG 2014 Id15217.

Author Contribution

F.N. and S.O. conceived the study; S.R. and A.K. performed genome-wide experiments, cloning and cell treatments; F.N. and D.I. performed genome-wide experiments and data analysis; M.M performed cloning and cell treatments; C.P and G.B. performed RNA-seq; F.A. performed CAPIP-seq experiments; F.N and S.O. wrote the paper with input from all authors.

Author information

Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper.

Accession Number

Raw and analysed data for this study are deposited at Geo Datasets under the reference GSE72856.

Supplemental Information

This file contains uncropped blots for Figures 1 and 3, Extended Data Figures 1, 7, 9, and Supplementary Tables 1-2.

Extended Data Figures

Extended Data Figure 1: Generation of Dnmt3b$^{-/-}$ and mapping of the endogenous Dnmt3b in ESCs.

Dnmt3b$^{-/-}$ ESC clones (B126 and B77) showed normal cell growth and alkaline phosphatase (AP) staining as well as impaired silencing, by promoter DNA methylation of Nanog expression during the differentiation into embryonic bodies (EBs) with respect to the WT cell line, thus indicating the *bona fide* nature of the transgenic cell lines. **a,** Scheme of the region of the Dnmt3b gene targeted by TALEN zinc-fingers, and representative sequences on the two alleles of two Dnmt3b$^{-/-}$ clones, compared to WT (KO #1 = B77; KO #2 = B126). **b,** Western blot analysis of Dnmt3b protein in the two Dnmt3b$^{-/-}$ clones compared to WT. Dnmt1 and Dnmt3a2 levels are not affected by loss of Dnmt3b. Actin is used as loading control. Interestingly, also the mRNA level (data not shown) of the Dnmt3b gene is almost completely lost in Dnmt3b$^{-/-}$ cells. **c,** Growth curve of WT and Dnmt3b$^{-/-}$ ESCs in a 3-days' time course. **d,** Alkaline phosphatase staining of WT and Dnmt3b$^{-/-}$ ESC colonies. **e,** RT-qPCR of Nanog levels in embryoid bodies (EBs) derived from Dnmt3b$^{-/-}$ clones, compared to WT ESCs, and EBs. Error bars represent the standard deviation of at least 3 independent experiments. **f,** Sanger sequencing of bisulfite-treated genomic DNA from WT ESCs and EBs, and Dnmt3b$^{-/-}$ ESCs-derived EBs, at the region of the Nanog promoter previously shown to be target of Dnmt3b-mediated methylation upon differentiation[52]. **g,** Histogram showing the quantity of the DNA recovered in ChIP experiments performed with different antibodies directed against Dnmt3b protein. **h,** Western blot analysis of Dnmt3b protein in WT and Dnmt3b$^{-/-}$ ESCs. Actin was used as

loading control. **i,** Histogram showing the quantity (ng) of the DNA recovered in ChIP experiments performed with anti-Dnmt3b antibody (Ab122932) in WT and Dnmt3b$^{-/-}$ ESC. **j,** Genomic views of the mapped reads from different ChIP-seq datasets in ESCs. IgG and Dnmt3b ChIP-seq and WGBS are from the present work, bio-Dnmt3b from GSE57413, MeDIP-seq from GSE44644, GLIB-seq from GSE44566, histone modifications from GSE12241. **k,** (Left) Heatmap representations of Dnmt3b binding and relevant histone modifications on a window of ±3kb centred on the TSS of RefSeq genes, sorted by their expression level, according to RNA-seq data. (Right) Plots of Dnmt3b binding and relevant histone modifications on a window of ±3kb centred on the TSS of RefSeq genes, clustered in the 4 quartiles of expression (q4 = upper quartile, the most expressed genes). **l, m,** Binding enrichment of IgG (in WT ESCs) as well as IgG and Dnmt3b (in Dnmt3b$^{-/-}$ ESCs) on the exons or introns partitioned in quartiles on the basis of the expression of the related gene. These figures represent control experiments for the Figure 1b. **n,** Hierarchical clustering of pairwise Pearson correlation of Dnmt3b, and third-party ChIP-seq datasets in ESCs, reveals a strong genome-wide association of Dnmt3b with H3K36me3 histone mark. **o,** Scatterplots comparing intragenic H3K36me3 and IgG/Dnmt3b enrichments (Log2) in WT and Dnmt3b$^{-/-}$ cells. r = Pearson correlation. **p,** qPCR of ChIP analysis of Dnmt3b on the indicated regions. A specific enrichment can be observed on gene body of active genes. Error bars represent the standard deviation of at least 3 independent experiments. Primers used are reported in Table S1. **q,** Immunoprecipitation experiment (using a different antibody for Dnmt3b, Ab2851) in ESCs reveals the

interaction of Dnmt3b with H3K36me3, but not H3K4me3, in agreement with ChIP-seq data. **r,** Violin plots of the methylation level of all CpGs covered in both WT and Dnmt3b$^{-/-}$ ESCs WGBS on the indicated genomic features. Significance is given by Wilcoxon Rank Sum test statistics.

Extended Data Figure 2: Dnmt3b loss increases intragenic RNA transcription initiation.

a, Scatter plots of the Log2 RPKM gene values in the indicated samples. r = Pearson correlation. **b,** Genomic views of the RNA-seq mapped reads from the indicated samples. **c,** Box plots of the ratio between normalized RNA-seq read counts (RPKM) for the second and the first exon (left-top), the third and the first exon (left-bottom), the average of the intermediate exons (from the 4th to penultimate) and the first exon (right-top), the last and the first exon (right-bottom), in WT (rep #2) and Dnmt3b$^{-/-}$ (rep #2 and clone B77) ESCs. Significance is given by Wilcoxon Rank Sum test statistics. **d,** Pie-chart showing the percentage of transcripts having (intermediate/first exon ratio in Dnmt3b$^{-/-}$ (rep #2 and clone B77) vs. intermediate/first exon ratio in WT (rep 2) ESCs) Log2 fold-change > 1, < -1 or between -1 and 1. **e,** RT-qPCR analysis of Ints2, Nodal, Gabpa and XpoI transcripts by using primers targeting different exons to discriminate different isoforms in WT, Dnmt3b$^{-/-}$ (cl. B126) and Dnmt3b$^{-/-}$ (cl. B77) ESCs. All the PCR were normalized on β-Actin and on the WT condition. Error bars represent the standard deviation of at least 3 independent experiments. P-value was calculated against WT condition for

each experiment by using t-test. (** p-value < 0.001, * p-value < 0.01). Primers used are reported in TableS1.

Extended Data Figure 3: Dnmt3b loss does not extensively affect alternative promoter activation.

We investigate the activation or repression of alternative promoters on the subset of genes showing at least two annotated alternative promoters. On these genes, we measured the RPKM value of the first exon of all the isoforms transcribed from the first, second, third, fourth or 5-12th promoter in WT or Dnmt3b$^{-/-}$ cells. We observed that the Dnmt3b$^{-/-}$ cells showed a general trend to have genes with the first exon less expressed independently from the isoform, thus suggesting a non-global general activation of intragenic promoters. Analysis of the ratio of the expression between the 1st promoter and the 2nd downstream promoter identified 4 genes (on a total of 2563 genes) with a reactivation of the intragenic promoter in Dnmt3b$^{-/-}$ cells. **a,** Scheme of the gene dataset used for alternative promoter analysis. The dataset is composed of total 2563 gene showing at least two annotated alternative promoters, including 713 genes having at least three, 195 genes at least four and other 189 genes with multiple alternative promoters (from 5 to max 12). **b,** RPKM value of the first exon of all the isoforms transcribed from the first, second, third, fourth or 5-12th promoter in WT or Dnmt3b$^{-/-}$ ESCs. Dnmt3b$^{-/-}$ cells showed a general trend to have genes with the first exon less expressed independently from the isoform, and none of putative intragenic promoters (from the 2nd to the 12th) showed general activation. **c,** Analysis of the ratio of the expression of the 1st promoter over the 2nd downstream

promoter displayed high correlation between replicates and WT or Dnmt3b$^{-/-}$ ESCs. Only 4 genes (on a total of 2563 genes) showed a reactivation of the intragenic promoter in Dnmt3b$^{-/-}$ ESCs. Further analysis of the ratio between RPKM of the first exon and of the whole transcript for each class of alternative promoters transcribed genes did not show any evidence for possible reactivation of any class of transcript isoforms derived from intragenic promoters. **d, e,** Analysis of the ratio of the RPKM value of the first exon over the whole transcript for each class of alternative promoters transcribed genes showed high correlation between WT and Dnmt3b$^{-/-}$ ESCs and did not reveal evidence for possible reactivation of any class of transcript isoforms derived from intragenic promoters.

Extended Data Figure 4: Dnmt3b loss does not globally affect elongating RNA Pol II or H3K36me3 deposition on the gene bodies, but increases intragenic RNA Pol II spurious entry.

a, Genomic views of the mapped reads from the indicated different ChIP-seq datasets in WT and Dnmt3b$^{-/-}$ ESCs normally cultured or treated with the RNA Pol II elongation inhibitor DRB. **b,** Hierarchical clustering of pairwise Pearson correlation of ChIP-seq experiments performed in this work, and third-party ChIP-seq datasets in ESCs. **c,** Heatmap representations of the indicated ChIP-seq (in WT and Dnmt3b$^{-/-}$ ESCs) peaks with respect to annotated RefSeq genes, sorted by their expression level, according to RNA-seq data. Each gene was extended by 3kb upstream of its TSS, and downstream of its TES. **d,** Plots of the H3K36me3 distribution in WT and Dnmt3b$^{-/-}$ ESCs. **e,** Binding enrichment of the H3K36me3 on intermediate exons and introns in

WT and Dnmt3b$^{-/-}$ ESCs. **f,** Binding enrichment of the indicated ChIP-seq experiments in WT and Dnmt3b$^{-/-}$ ESCs treated (or not) with DRB on the intermediate exons and introns subdivided in quartiles of expression. This result demonstrated that only not elongating RNA Pol II is enriched on the bodies of the most expressed genes (q3 and q4) in Dnmt3b$^{-/-}$ ESCs. Significance is given by Wilcoxon Rank Sum test statistics (** p-value < 2.2E-16). **g,** Genomic views of the mapped reads from the ChIP-seq analyses for H3K4me3 and H3ac in WT and Dnmt3b$^{-/-}$ ESCs. **h,** Hierarchical clustering of pairwise Pearson correlation of ChIP-seq experiments performed in this work, compared with ENCODE ChIP-seq datasets. **i,** Heatmap representations of the indicated ChIP-seq (in WT and Dnmt3b$^{-/-}$ ESCs) peaks with respect to annotated RefSeq genes, sorted by their expression level, according to RNA-seq data. Each gene was extended by 3kb upstream of its TSS, and downstream of its TES. **j,** Binding enrichment of the indicated ChIP-seq experiments in WT and Dnmt3b$^{-/-}$ ESCs on the first exons, intermediate exons and introns subdivided in quartiles of expression. This result demonstrated that H3K4me3 and H3ac distribution are enriched on the intermediate exons and introns of the most expressed genes of the Dnmt3b$^{-/-}$ ESCs. Significance is given by Wilcoxon Rank Sum test statistics.

Extended Data Figure 5: CAPIP-seq enrichment of the 5' of the RNAs shows that Dnmt3b loss increases intragenic spurious transcription initiation.

a, Schematic view of the CAPIP-seq protocol used. Total RNA is chemically fragmented and then subjected to immunoprecipitation by using specific anti-

CAP antibody or control anti-IgG antibody. Eluted RNA (as well as 1/10 of the starting material for Input) is subjected to random primer reverse transcription. Library is then completed starting from second strand generation. **b,** Scatter plots of the Log2 RPKM of CAPIP-seq data (anti-CAP antibody) and Input in WT and Dnmt3b$^{-/-}$ ESCs. r = Pearson correlation. **c,** Hierarchical clustering of pairwise Pearson correlation of CAPIP-seq related sequencings in WT and Dnmt3b$^{-/-}$ ESCs. **d,** Genomic views of the total mapped reads from the indicated CAPIP-seq related sequencings. Enrichment of the CAP signal is present on the 5' of the RNA as a peak of about 150bp broader with respect to the signal obtained by performing DECAP-seq. **e,** Plots of the CAPIP-seq mapped reads distribution in WT and Dnmt3b$^{-/-}$ ESCs with respect to annotated RefSeq genes, extended by 5kb upstream of its TSS, and downstream of its TES. **f,** Box-plots of the Log2 enrichment of the CAPIP-seq signal rep#2 (CAP immunoprecipitation signal over input in WT and Dnmt3b$^{-/-}$ ESCs on the indicated genic features. Significance is given by Wilcoxon Rank Sum test statistics. **g.** Further analysis showing the increase of CAP localization from intragenic regions of the RNA. Intragenic ratio is calculated as the Log2 ratio of cap signal gene-body enrichment in Dnmt3b$^{-/-}$ versus WT cells. The correlation between the two replicates is shown (r = Pearson correlation).

Extended Data Figure 6: DECAP-seq method maps, at single base resolution, TSSs on the gene body in ESCs.

a, Schematic representation of the workflow of the DECAP-seq technique that is based on the RNA 5´ Pyrophosphohydrolase (RppH) enzymatic activity that

in Thermopol buffer is able to mediate decapping and pyrophosphate removal from the 5´ end of RNA to leave a 5´ monophosphate RNA (5'-P). 5'-P RNA is then used for selective adapter ligation by T4 RNA ligase to the originally capped RNA fragments allowing single base resolution mapping of the RNA capping sites. Treating sample in the same way, but without RppH enzyme generates a negative control (to detect technical background). Positive control is generated by treating sample with T4 Polynucleotide Kinase (PNK) for 5' phosphorylation of all RNA fragments. This method represents an affordable alternative to the use of the Tobacco Acid Pyrophosphatase (TAP) enzyme that has been used in several high-throughput techniques as GRO-seq, CAP-seq, CIP-TAP[53] because the EpiCentre Technologies (so far the only company producing commercial TAP) has discontinued TAP and all kits containing it. **b,** Total RNA fragmentation was verified by using Fragment Analyzer™ (Advanced Analytical). **c,** Final DECAP-seq libraries were inspected on Fragment Analyzer before gel size selection. RppH treated and untreated samples showed a double peak around 130bp corresponding to the dimers of adapter, but only the RppH treated sample showed a higher enrichment (in the red box) corresponding to the decapped RNA fragments. PNK treated sample displayed a large peak around 200 bp. **d,** Final DECAP-seq libraries were quantified on Qubit (Invitrogen) after gel size selection and PCR enrichment. Library generated by treating RNA with RppH showed a fifty-fold higher concentration with respect to the library generated without RppH treatment (5 ng/µl vs. 0.1 ng/µl). **e,** Genomic views of the total DECAP-seq mapped reads from the indicated treatment on a gene (Actb) on the Crick DNA strand (- strand) and a gene (Rpl5) on the Watson DNA strand (+

strand). A pronounced sharp peak (red arrow) is present on the TSS only on the respective gene strand, thus reflecting both the cap- and strandness-specificity of the method. Unstranded RNA-seq is shown as reference example. **f,** Plot of total TSSs (identified by using DECAP-seq rep #2) distribution along genes in Dnmt3b$^{-/-}$ (blue line) compared with WT (red line) ESCs. **g,** Box plots showing the number of total TSS / gene on RefSeq annotated TSSs and on gene body in WT and Dnmt3b$^{-/-}$ ESCs. Significance is given by Wilcoxon Rank Sum test statistics. **h,** Histogram plot showing the average RPM of novel identified TSSs by DECAP-seq in both replicates of WT ESCs. **i, j,** Scatter plots of the Log2 RPM values on canonical annotated TSSs(±100bp) in both replicate of DECAP-seq samples in WT and Dnmt3b$^{-/-}$ ESCs. r = Pearson correlation. **k,** Venn diagrams of intragenic TSSs having DECAP-seq signal > 6 RPM showing the single-base resolution overlap between the DECAP-seq experiments replicates. Significance is given by the Hypergeometric Distribution test statistics. **l,** Venn diagram of intragenic TSSs having DECAP-seq signal > 6 RPM showing single-base resolution overlap between Dnmt3b$^{-/-}$ and WT ESCs (rep #2). **m,** Pie-charts of the DECAP-seq reads distribution on TSSs > 6 RPM in WT (on the left) and Dnmt3b$^{-/-}$ (on the right) cells (rep #2). In green are shown the novel TSSs that overlap with RefSeq annotated TSS, in yellow all the common TSSs distributed on gene body and in pink the sample-specific TSSs on gene body. **n,** Box-plot distribution of the enrichment of the CAPIP-seq and Pol II ChIP-seq signals calculated as the Log2 ratio in Dnmt3b$^{-/-}$ versus WT cells on the novel identified TSSs (in green those overlapping with RefSeq annotated TSSs and in pink those specifically found on gene bodies of Dnmt3b$^{-/-}$ ESCs) and on an

intragenic random dataset. Significance is given by Wilcoxon Rank Sum test statistics. **o,** Box-plot distribution of the ratio between downstream and upstream exon expression levels with respect to the novel identified intragenic TSSs or an intragenic random dataset in Dnmt3b$^{-/-}$ cells. The exon expression levels were calculated by counting the reads from the RNA-seq experiments in Dnmt3b$^{-/-}$ or WT cells. Significance is given by Wilcoxon Rank Sum test statistics.

Extended Data Figure 7: DECAP-seq maps in Dnmt3b$^{-/-}$ ESCs the internal TSSs revealing their correlation with the binding of methylation sensitive transcription factors.

a, Sequence binding logo of the indicated transcription factors. **b,** Schematic representation of CpG localization and putative transcription factors binding elements on the regions (±50bp) of some Dnmt3b$^{-/-}$ ESCs specific intragenic TSSs is shown. **c**, RT-qPCR analysis of CAPIP (top panel) and qPCR analysis of ChIP (middle panel) experiments on the indicated genomic regions in WT and Dnmt3b$^{-/-}$ ESCs. For CAPIP RT-qPCR the primers were designed downstream the novel identified TSSs. Bottom panel represents the fold difference of the ratio between downstream and upstream exon expression levels with respect to the novel identified intragenic TSSs. For TSSs falling on exon, the downstream or upstream part of the same exon was considered as downstream or upstream exon if longer than 200bp. P-value was calculated against WT condition by using t-test. (** p-value < 0.01; * p-value < 0.05; n.s. = not significant). **d**, Sanger bisulfite sequencing of WT and Dnmt3b$^{-/-}$ ESCs

on intragenic TSSs previously described. **e**, qPCR analysis of ChIP experiments on the indicated genomic regions.

Extended Data Figure 8: SetD2 knockdown reduces H3K36me3 mark, Dnmt3b binding, intragenic DNA methylation, and spurious TSSs on the gene bodies.

a, RT-qPCR of SetD2 knock-down in WT and Dnmt3b$^{-/-}$ ESCs, using two independent shRNAs. Error bars represent the standard deviation of at least 3 independent experiments. **b,** Venn diagram showing genome-wide the number of H3K36me3 peaks in control and SetD2 knockdown ESCs. **c,** Plots of H3K36me3 distribution on genes in control and SetD2 knockdown cells show a decrease of H3K36me3 distribution on gene bodies of the SetD2 silenced cells. **d,** Histograms of the percentage of Dnmt3b ChIP-seq peaks overlapping intronic, and exonic regions of genes clustered in quartiles of expression (q1-4) in control or SetD2 knockdown cells. P-value was calculated by using chi-squared test. (** p-value < 0.001). **e,** Genomic views of the mapped reads from H3K36me3 and Dnmt3b ChIP-seq datasets in control and two different SetD2 knockdowns ESCs. **f,** qPCR analysis of H3K36me3 and Dnmt3b ChIP experiments and MeDIP analysis in control and SetD2 knockdown cells for the indicated genomic regions. A specific loss of Dnmt3b and DNA methylation is observed only on gene body of active genes. Error bars represent the standard deviation of at least 3 independent experiments. P-value was calculated against WT condition for each experiment by using t-test. (** p-value < 0.001). Primers used are reported in TableS1. **g,** Scatter plots of the Log2 RPKM gene values in the indicated

samples. r = Pearson correlation. **h,** Genomic views of the RNA-seq mapped reads from the indicated samples. **i,** Plot of total TSSs (identified by using DECAP-seq) distribution along genes in SetD2 knockdown (yellow line) compared with control knockdown (red line) ESCs. **j,** Box plots showing the number of total TSS / gene on RefSeq annotated TSSs and on gene body in control and SetD2 knockdown ESCs. Significance is given by Wilcoxon Rank Sum test statistics. **k,** Scatter plots of the Log2 RPM values on canonical annotated TSSs (±100bp) in control and Setd2 knockdown ESCs. r = Pearson correlation. **l,** Venn diagram of intragenic TSSs having DECAP-seq signal > 6 RPM showing single-base resolution overlap between control and SetD2 knockdown ESCs. **m, n,** Venn diagrams of intragenic TSSs having DECAP-seq signal > 6 RPM showing single-base resolution overlap between the indicated samples. Significance is given by the Hypergeometric Distribution test statistics. **o,** Pie-charts of the DECAP-seq reads distribution on TSSs > 6 RPM in control knockdown (on the top) and Setd2 (on the bottom) ESCs. In green are shown the novel TSSs that overlap with RefSeq annotated TSS, in yellow all the common TSSs distributed on gene body and in pink the sample-specific TSSs on gene body.

Extended Data Figure 9: Internal transcription activation in Dnmt3b$^{-/-}$ ESCs show the same intragenic TSSs also in SetD2 knockdown cells.

a, Genomic view of the indicated genes showing intragenic transcription initiation increase in Dnmt3b$^{-/-}$ and in shSetD2 WT cells. Lower part, sanger bisulfite sequencing of shCTR and shSetD2 WT ESCs on intragenic TSSs

previously described. **b,** Genomic views of the mapped reads from H3K36me3 (in WT ESCs) and Dnmt3b ChIP-seq datasets (in Mock or Dnmt3b-transfected Dnmt3b$^{-/-}$ ESCs). Both the WT or the catalytically inactive mutant (V725G) Dnmt3b showed intragenic binding enrichment. **c,** qPCR analysis of IgG and Dnmt3b ChIP experiments in Mock or Dnmt3b (WT and mutant V725G) transfected Dnmt3b$^{-/-}$ ESCs for the indicated intragenic regions. Error bars represent the standard deviation of at least 3 independent experiments. P-value was calculated against Mock condition by using t-test. (** p-value < 0.001). Primers used are reported in TableS1. **d,** Dot-blot analysis of genomic DNA isolated from Mock or Dnmt3b (WT and mutant V725G) transfected Dnmt3b$^{-/-}$ ESCs. Dot intensity quantification from three biological replicates revealed that Dnmt3b WT (but not the V725G mutant) significantly (p=0.003) increases global DNA 5mC. P-value was calculated against Mock condition by using t-test. **e,** qPCR analysis of MeDIP experiments in Mock or Dnmt3b (WT and mutant V725G) transfected Dnmt3b$^{-/-}$ ESCs for the indicated intragenic regions. An intragenic significant increase of DNA methylation has been observed in Dnmt3b WT (but not mutant) transfected Dnmt3b$^{-/-}$ ESCs. Error bars represent the standard deviation of at least 3 independent experiments. P-value was calculated against Mock condition by using t-test. (** p-value < 0.001). Primers used are reported in Table S1. **f,** Genomic views of the RNA-seq mapped reads from the indicated samples. **g,** Scatter plots of the Log2 RPKM gene values in the indicated samples. r = Pearson correlation. Of note, Mock ESCs showed higher correlation with Dnmt3b mutant transfected ESCs (r = 0.99) with respect that Dnmt3b WT transfected ESCs (r = 0.95), thus suggesting that DNA

methylation enzymatic activity is the major driver of the Dnmt3b-dependent transcriptome alterations. **h,** Western blot of Dnmt3b$^{-/-}$ ESCs transfected with mock, Dnmt3b WT, Dnmt3b S277P, and Dnmt3b VW-RR. β-Actin was used as protein loading control. I, **j,** qPCR analysis of ChIP and MeDIP experiments of the indicated regions in Dnmt3b mutant conditions. Specific impairment of Dnmt3b binding and DNA methylation is observed in both the mutants compared to the rescue by using WT Dnmt3b enzyme. Error bars represent the standard deviation of at least 3 independent experiments. Primers used are reported in TableS1**.**

Extended Data Figure 10: Cryptic RNA transcripts are degraded in part by the RNA exosome complex.

a, RNA-seq profile of Dnmt3b$^{-/-}$ cells transfected with mock or Dnmt3b mutants. **b,** Scatter plots of the Log2 RPKM gene values in the indicated samples. r = Pearson correlation. **c.** Box plot of the ratio between normalized RNA-seq read counts (RPKM) for the second, the third, the intermediate (average) and the last over the first exons, in Dnmt3b$^{-/-}$ ESCs transfected with mock, Dnmt3b (WT) or mutants Dnmt3b (S277P and VW-RR). Significance is given by Wilcoxon Rank Sum test statistics. **d,** Pie-chart showing the percentage of transcripts having (intermediate/first exon ratio in Dnmt3b (WT or mutants) transfected Dnmt3b$^{-/-}$ ESCs vs. intermediate/first exon ratio in mock Dnmt3b$^{-/-}$ ESCs) Log2 fold-change (FC) > 1, < -1 or between -1 and 1.

e, f, Histogram plot and western blot showing mRNA and protein levels of Dis3 and Rrp6 genes in control or Dis3/Rrp6 double knockdown (dKD) in

Dnmt3b$^{-/-}$ ESCs. β-Actin was used as protein loading control. **g,** Box plots showing the number of total TSS / gene on RefSeq annotated TSSs and on gene body in control or Dis3/Rrp6 dKD Dnmt3b$^{-/-}$ ESCs. Significance is given by Wilcoxon Rank Sum test statistics. **h,** Box plot of the normalized DECAP-seq read counts (RPM) on the intragenic TSSs in the indicated samples. Significance is given by Wilcoxon Rank Sum test statistics. **i,** Scatter plots of the Log2 RPM values on canonical annotated TSSs(±100bp) of the indicated samples. r = Pearson correlation. **j,** Venn diagrams of intragenic TSSs having DECAP-seq signal > 6 RPM showing the single-base resolution overlap between the DECAP-seq experiments replicates performed in Dnmt3b$^{-/-}$ ESCs. Significance is given by the Hypergeometric Distribution test statistics. **k,** Pie-charts of the DECAP-seq reads distribution on TSSs > 6 RPM in control (on the left) and Dis3/Rrp6 KD (on the right) Dnmt3b$^{-/-}$ ESCs. In green are shown the novel TSSs that overlap with RefSeq annotated TSSs, in yellow all the common TSSs distributed on gene body and in pink the sample-specific TSSs on gene body. **l,** Scatter plots of the Log2 RPKM gene values in the indicated samples. r = Pearson correlation. **m,** Genomic views of the RNA-seq mapped reads from the indicated samples. **n,** Box plots of the ratio between normalized polyA RNA-seq read counts (RPKM) for the second and the first exon, the third and the first exon, the average of the intermediate (from the 4th to penultimate) and the first exon, the last and the first exon in WT (rep #2) and Dnmt3b$^{-/-}$ (rep #2 and cl. B77) ESCs. Significance is given by Wilcoxon Rank Sum test statistics. **o,** Pie-chart showing the percentage of transcripts having intermediate/first exon ratio (in Dnmt3b$^{-/-}$ rep #2 and clone B77 polyA RNA-seq) vs. intermediate/first exon ratio (in WT rep #2 polyA RNA-seq) Log2

fold-change > 1, < -1 or between -1 and 1. **p,** Box plots showing the number of total TSS / gene on RefSeq annotated TSSs and on gene body identified by DECAP-seq in the indicated RNA compartments. Significance is given by Wilcoxon Rank Sum test statistics. **q,** Scatter plots of the Log2 RPM values on canonical annotated TSSs (±100bp) in the indicated RNA compartments. r = Pearson correlation. **r,** Venn diagram of the common intragenic TSSs (defined as having RPM > 6 in both Dnmt3b$^{-/-}$ and WT ESCs) in the indicated RNA compartments. **s,** Box plot of the normalized DECAP-seq read counts (RPM) on the common (defined as having RPM > 6 in both Dnmt3b$^{-/-}$ and WT ESCs) intragenic TSSs in the indicated RNA compartments.

Extended Data Figure 11: Loss of Dnmt3b generates partial intragenic starting RNAs that are as stable as canonical mRNAs.

a, Genomic views of the RNA-seq mapped reads from the indicated samples. Genes with slow, medium and fast decay are shown. **b,** Gene Ontology (GO) analysis of the subsets of the mRNAs having fast decay (half-life lower of four hours) or slow decay (half-life higher than nine hours) in WT ESCs. The analysis revealed that fast decay mRNAs are mainly involved in cell cycle and transcription biological processes, while slow decay mRNAs are more related to metabolism and translation. This result is in agreement with that previously observed in mESCs[54,55], thus suggesting the *bona fide* of the experiment. **c,** Scatter plot of the mRNAs half-life (in hours) in WT and Dnmt3b$^{-/-}$ ESCs. **d, e,** Box plots of the introns half-life (in hours) in WT and Dnmt3b$^{-/-}$ ESCs. Introns half-life is estimated by considering only the reads mapped on intronic regions. Introns half-life is generally lower with respect to the mRNAs half-life,

thus suggesting a less stability of the RNAs containing intronic parts. Introns half-life calculated in Dnmt3b$^{-/-}$ ESCs is significantly ($p = 0.0016$) higher with respect to that calculated in WT ESCs. Significance is given by Wilcoxon Rank Sum test statistics. **f, g**, Frequency distribution of introns and mRNA half-life among all introns in WT and Dnmt3b$^{-/-}$ ESCs. **h,** Genomic views of the RNA-seq mapped reads from the indicated samples. ART-seq reads derived only from the coding sequences (CDS) of the mRNAs. RNA-seq is shown as reference example. **i,** Scatter plots of the Log2 RPKM gene values in the indicated samples. r = Pearson correlation. **j,** Box plot of the normalized ART-seq rep #2 read counts (RPKM) on the indicated RNA regions in WT and Dnmt3b$^{-/-}$ ESCs. **k,** Box plot of the normalized ART-seq read counts (RPKM, for both the biological replicates) on the introns in WT and Dnmt3b$^{-/-}$ ESCs. Significance is given by Wilcoxon Rank Sum test statistics

Extended Data Figure 12: Models from of the obtained results.

a, Scheme of the functional role of the Dnmt3b-dependent intragenic DNA methylation in ESCs. In WT cells the Dnmt3b is able to methylate gene bodies to favour a repressive chromatin inhibiting spurious entries of the RNA Polymerase II. In absence of Dnmt3b, gene bodies are hypomethylated, thus leading to RNA Pol II intragenic entries generating intragenic transcription initiation. **b,** Epigenetic crosstalk between RNA Pol II, SetD2 and Dnmt3b and relative H3K36me3 and 5mC chromatin modifications, unveils how the RNA

Pol II, through the transcription elongation process, triggers a safety mechanism to ensure its transcription initiation fidelity.