SUPPLEMENTAL MATERIAL

The aberrant epigenome of *DNMT3B*-mutated ICF1 syndrome is amenable to correction in iPSCs, with the exception of regions with H3K4me3- and/or CTCF-based epigenetic memory

Varsha Poondi Krishnan[1], Barbara Morone[1], Shir Toubiana[2], Monika Krzak[3], Salvatore Fioriniello[1], Floriana Della Ragione[1,4], Maria Strazzullo[1], Claudia Angelini[3]*, Sara Selig[2,5]* and Maria R. Matarazzo[1]*

SUPPLEMENTAL METHODS

Quantitative Real time PCR

1 μg of RNA derived from iPSCs was reverse-transcribed using 100ng of Random Primers (48190011, Invitrogen™) and 100U of SuperScript™ II Reverse Transcriptase (18064022, Invitrogen™) in a T100™ Thermal Cycler (Biorad), according to manufacturer's protocol (5' at 65°C, 2'+10' at 25°C, 40' at 42°C and 15' at 70°C). Real-time quantitative PCR (RT-qPCR) was performed using SsoAdvanced™ universal SYBR® Green supermix (1725270, Bio-Rad) in a Bio-Rad iCycler, according to manufacturer's protocols. Expression levels were normalized to the *GAPDH* gene by the ΔΔCt method. Immunoprecipitated samples and corresponding mock samples (negative controls to measure background) were used for ChIP-qPCR. The enrichment of DNA was calculated in terms of % input = $2^{-\Delta Ct} \times 100$, where ΔCt (threshold cycle) was determined by Ct ChIP sample – Ct Input. Primer sequences for gene expression and ChIP analyses appear in Supplemental Table S2.

Western blot analysis

Total cell extracts were obtained by resuspending the cells in lysis buffer (100 mM Tris-HCl, pH 8, 140 mM NaCl, 20 mM EDTA, 0.2% SDS, 1% NP-40) supplemented with protease inhibitors, and protein lysates were quantified by Bradford protein assay (Bio-Rad), according to manufacturer instructions. Twenty-five micrograms of protein lysates were separated by SDS-PAGE and transferred to PVDF membranes (Millipore), which were, subsequently, blocked with 5% BSA in TBS buffer with 0.2% Tween20 for 30 min at room temperature (RT), and then incubated with an anti-DNMT3A antibody (1:1000, Abcam ab2850) overnight at 4°C, or an anti-Actin antibody (1:3500, Sigma-Aldrich A2066) for 1h at RT. Incubation with an HRP-conjugated secondary antibody was carried out for 1h at RT.

Genome-wide DNA methylation analysis (WGBS)

Identification of hypo-DMRs. We assessed the quality of sequenced PE reads using FastQC prior and post-trimming. Using cutadapt we filtered out the raw read pairs with Phred score < 30, read length < 40 and trimmed the Illumina adapter sequences by applying the following parameters: -u 7; -U 7; -q > 30; -m > 40; -a AGATCGGAAGAG -A AGATCGGAAGAG. Summarized reports of pre- and post-trimmed PE reads were obtained using multiQC software (Ewels et al. 2016). We used the Bismark program with default parameters for the bisulfite-conversion and indexing of the hg38 reference genome (canonical chromosomes including Chr 1-22; X and Y), alignment of PE reads to the reference index, and the removal of PCR duplicates. We obtained the per-base cytosine methylation report (CG, CHG, CHH) using *methylation extractor* command (parameters: -p --no_overlap --bedGraph --counts --zero_based –cutoff 2 --cytosine_report –CX_context). We evaluated the methylation level of

cytosines (Cs) from both strands in the context of CG, whereas we measured the methylation level of strand-specific Cs for the non-CGs. We tiled the genome into 1kb windows with at least 2 Cs and minimum coverage of 5 reads per tile, identified the Differentially Methylated Regions (DMRs) using the MethylKit R package (meth.diff score>25%, qvalue=0.01) and selected the hypo-DMRs with type="hypo" where the meth.diff score is the difference between the methylation percentages. Our global DNA methylation analysis included the repetitive fraction of the genome.

To filter the hypo-DMRs in the comparison between ICF1 and WT1 iPSCs, we downloaded the single-end (SE) FASTQ file of the additional control iPSC, WT2, (see Methods) using the *fasterq-dump* tool from SRA Toolkit (Katz et al. 2022) and aligned them to the hg38 reference genome using the Bismark aligner with default parameters for SE mapping. We performed the extraction of methylation content, and coverage calculation for WT2. Then, we filtered out the ICF1 hypo-DMRs when we observed a difference > 0.2 between the methylation levels (expressed as the ratio of the number of Cs over the total number of Cs and Ts) in WT1 and WT2 iPSCs, as described in Methods (Whole Genome Bisulfite Sequencing and data processing). The hypo-DMRs identified in Chromosome Y were removed from pR and its corrected clones. The ICF1 hypo-DMRs were defined as rescued in the corrected clones if compared with the internal control WT1 they had i) meth.diff score < 25%, level=”hypo” and/or q > 0.01; or ii) meth.diff > 25% and level=”hyper”.

Characterization of hypo-DMRs. We divided the ICF1 hypo-DMRs into four Groups 1-4 using k-means clustering based on the mean methylation levels across the iPSCs samples. Group1 consisted of hypo-DMRs with the lowest mean methylation level, while Group 4 had the highest mean methylation level across all analyzed iPSCs (Fig. 1A).

We annotated the hypo-DMRs to genes using the *annotatePeak* function from the ChIPseeker R package (Yu et al. 2015) and reported only the nearest gene feature per region. The gene features were classified into promoters (up to 2kb upstream of the transcription start sites (TSS) to 500bp downstream to the TSS), 5' UTR, Exons, Introns, 3' UTR, downstream (< 3kb downstream transcription end sites (TES)) and distal intergenic (> 2kb distal from TSS and > 3kb from TES) (Fig. 2A).

We subsetted the common pR and pG hypo-DMRs annotated to promoter and gene-body features and provided the associated genes as input for functional enrichment analysis (GO) on the PANTHER database web interface applying the following settings: Annotation set: "PANTHER GO-Slim Biological Process"; Reference: "Homo Sapiens genes"; Test Type and Correction set to default (Mi et al. 2021) (Fig. 2B).

To assess the frequency of ICF1 hypo-DMRs in regions enriched in CGI and GH promoters/enhancers (from GeneHancer database, GH) across the chromosomes (Supplemental Fig. S2), we binned the genome into 5Mb windows, and counted the number of CGI and GH promoters/enhancers present in each window. Based on this count, we defined the following four groups: "high_CGI/GH prom. and enh." (>10), "mid_CGI/GH prom. and enh." (5 -10), "low_CGI/GH prom. and enh." (1-5) and "no_CGI/GH prom. and enh.". We then assigned the hypo-DMRs to the windows they belong to, based on their position, we calculated the frequency of each group and visualized them as a stacked barplot (Fig. 1B). To assess the statistical significance of the correlation between the hypo-DMRs and CGI/GH promoters and enhancers, we used Poisson Regression with the number of hypo-DMRs as response variable, and the number of CGI or GH promoters/enhancers in each window and

the number covered tiles in each window as explanatory variables. To this purpose, we used the *glm* function in R setting family="poisson".

To compare our results with the microarray-based profiles of DNA methylation in ICF1 patients' whole blood (Velasco et al. 2018), we downloaded the methylation arrays from GSE95040, calculated the median beta (β) values (i.e., the methylation level expressed in the [0,1] range) of the control samples ($\beta_{Control}$) and ICF1 samples (β_{ICF1}), and identified the hypomethylated regions in ICF1 blood as those in which $\beta_{ICF1} - \beta_{Control} < -0.2$ (Supplemental Fig. S4 A,B).

We measured the overlap between two lists of observed regions by computing the number of regions ($nobs$) where they overlap by at least 1bp. We assessed the significance of such observed overlap by using the Shuffle test. Similar to the resampling procedures, the Shuffle test simulates the null hypothesis that the overlap is due to chance by randomly permuting the observed regions over the entire genome and computing the overlap among them. The random regions are of the same size and numerosity compared to the observed regions. The process is repeated $nshuffle$ times and after each repetition, the random overlap ($robs_i$) is evaluated. Then, the *P*-value of the observed overlap is estimated by comparing the observed overlap with the overlaps obtained from random regions (with the same size and numerosity). To perform this analysis, we used the *enrichPeakOverlap* function from ChIPseeker with the following parameters: TxDb=hg38_ensembl, pAdjustMethod="BH", nShuffle=4000, where TxDb denotes the genome assembly where the random regions are located and nshuffle denotes the number of times the simulation is repeated and pAdjustMethod denotes the post-hoc adjustment procedure on the obtained *P*-value. We applied such approach to ICF1 hypo-DMRs and the Regions of Interest (ROIs) (i.e., CGI, GH regulatory regions, DMV, hypo-

DMRs LCLs, ChIP-seq DERs, hypomethylated genomic domains (Huang et al. 2014) and DNMT3B KO/KD HMRs).

For Transcription Factor (TF) motif enrichment analysis, first, we selected the subsets of the hypo-DMRs overlapping with decreased DNMT3B DERs in pR and pG iPSCs. Next, the subsets of hypo-DMRs from Groups 1 and 2 (which showed low methylation recovery in corrected iPSCs) were combined. Similarly, the subsets of hypo-DMRs belonging to Groups 3 and 4 were combined. Then, we performed motif enrichment analysis using *findMotifsGenome.pl* from the HOMER suite (parameters: --size given and genome "hg38"), on i) Hypo-DMRs overlapping DNMT3B DERs in ICF1 (Fig. 2C), ii) the combined subset from Groups 1 and 2, and iii) the combined subset from Groups 3 and 4 (Fig 5A).

Data visualization. We used the ViewBS toolkit (Huang et al. 2018) for obtaining genome-wide methylation coverage (*GlobalMethLev*) for each cytosine context (Supplemental Fig. S1A-C) across all iPSC lines. We generated the profile-plots of the average weighted methylation level using *MethOverRegion function* (Supplemental Fig. S1A-C) and we used the *MethHeatmap* output to obtain the methylation levels at each ROIs (Fig. 1C, Fig. 4C and Supplemental Fig. S3B).

Heatmap representation of methylation levels across ICF1 hypo-DMR groups and their rescue status (Fig. 1A) was generated using the ComplexHeatmap R package. The integrated heatmaps was constructed using the *heatmap* function with the following parameters: split=k-means_row_order; color (methylation level)=colorRamp2(c(0,25), c("yellow","blue")) or color (rescue status)=c("green3", "orange", "red). We kept the labels "Groups 1-4" obtained here for the subsequent heatmaps.

Boxplots indicating the methylation levels across rescued and unrescued hypo-DMRs associated with CGI and promoters/enhancers from GeneHancer GH database were generated using ggplot2 (Fig. 1E and Supplemental Fig. S3E).

We plotted the chromosomal distribution (22 autosomes; Supplemental Fig. S2) of ICF1 hypo-DMRs in pR, cR7, cR35 and pG, cG13 and cG50 followed by CpG islands (CGI) as density line plots, using the gtrellis_layout function from the gtrellis R package (Gu et al. 2016). We set the window.size parameter to 2Mb for hypo-DMRs and 1kb for CGI, respectively.

We represented the genic distribution of the annotated hypo-DMRs and those rescued in corrected clones by gradient donut charts (Fig. 2A) generated with the *patternpie* and *patternring1* functions from the patternplot R package. We visualized the enriched Biological Processes (GO-BP) associated with the pR and pG common hypo-DMR annotated genes (ENSGs) as a Multidimensional Scaling (MDS) scatter-plot (Fig. 2B) using REVIGO web-server (Supek et al. 2011). In addition, we represented the proportion of the pR and pG common hypo-DMRs rescued in genes associated with top enriched GO terms as a stacked barplot using the *geom_bar* function from ggplot2 (Fig. 2B). We created Venn diagrams to depict the overlaps of ICF1 hypo-DMRs through the *draw.pairwise.venn* function from the VennDiagram R package (Chen and Boutros 2011) (Fig. 2D and Supplemental Fig. S4A-B).

We converted the bedgraph files from Bismark methylation caller to BigWig tracks using *bedGraphtoBigWig* utility and hosted them at Cyverse Discovery Environment (Merchant et al. 2016). The methylation coverage files across the genome were visualized on the UCSC genome browser (Navarro Gonzalez et al. 2021) with the following settings: track type=bigWig, visibility=full, viewLimits=default, windowingFunction=mean,

smoothingWindow=10, color=100,0,0). The filtered hypo-DMRs were uploaded as BED tracks with differential methylation scores indicated by gradient grey-scale boxes. Four shades of grey ranging from light grey to dark grey correspond to -25 to -39, -40 to -59, -60 to -79, and -80 to -100 hypo-DMRs score.

We visually compared the hypo-DMRs obtained in ICF1 iPSCs to hypomethylated regions (HMR) in wild-type human embryonic stem cells (hESCs) and hESCs in which DNMT3B was knocked-out and knocked-down, available in UCSC genome browser database. HMRs are defined as regions across the genome with low-methylation levels identified using the MethPipe package (Song et al. 2013). We displayed the following HMRs available in MethBase track hub: three wild-type hESCs - H1 (Lister et al. 2009); H9 (Martins-Taylor et al. 2012); HUES6 (Lister et al. 2013), two DNMT3B knock-out hESCs - early (2–7) and late (17–22) passage (Liao et al. 2015) (3BKO) and one shRNA DNMT3B knock-down hESCs (Martins-Taylor et al. 2012) (3BKD).

The DNMT3B-KO/KD HMRs were filtered for those that were unique to KO/KD hESCs lines or those at least 100bp longer than the overlapping HMRs in wild-type hESCs. Then, the shortest distance between the filtered KO/KD HMRs and ICF1 hypo-DMRs was obtained using *annotatePeakInBatch* function from ChIPpeakAnno R package and the proportion of hypo-DMRs in each bin was calculated and visualized using *geom_hist* function from ggplot2 (Supplemental Fig. S1E).

RNA-seq data analysis

We sequenced pR-related and pG-related iPSC samples in two separate batches, with the WT1 replicates sequenced in both batches. First, we filtered out the low-quality reads and

trimmed the adapters in the strand-specific PE reads using cutadapt by setting the following parameters: -q 30 -m 40 -a AGATCGGAAGAG -A AGATCGGAAGAG. Then, we aligned the trimmed reads to the hg38 reference genome (canonical chromosomes only) using HISAT2 with the following options: -p 8 --dta --rna-strandness RF. Next, we quantified the gene expression by counting the reads mapping to genes (ENSGs) using the *featureCounts* function from the Rsubread package with the following settings (annot.ext="hg38.v85" gtf file, useMetaFeatures=TRUE, allowMultiOverlap=FALSE, strandSpecific=2, CountMultiMappingReads=FALSE). After that, we filtered out zero count or low count genes (i.e., those with CPM < 0.5) using the Proportion Test Method from the NOIseq R package. Overall, we obtained a list of expressed genes (n=18077) excluding the genes expressed from Chromosome Y since the iPSCs were derived from individuals of different biological sex (pR- female; pG-male; WT1- male). Then, we normalized the raw counts across the samples using the Upper Quartile (UQUA) method.

The inspection of the Principal Component Analysis (PCA) on the normalized read count matrix revealed a batch effect between the samples due to the two different sequencing runs. Therefore, we performed the batch-effect removal using the *ArySyNseq* function with parameters: factor="run", batch=TRUE, norm="n", where the WT1 counts from the two runs were henceforth considered as one control sample with four replicates while every other sample (ICF1 and corrected iPSCs) had two replicates.

Identification of differentially expressed (DE) genes. We used the *noiseq* function from the NOIseq package. We compared the expression of ICF1 iPSCs and corrected iPSCs versus WT1 iPSCs, and we defined the DE genes in each comparison as those genes with a posterior probability (pp) > 0.9. As a further quality control of the gene expression profile in our WT1

iPSCs, we downloaded the RNA-seq FASTQ reads of additional WT iPSCs (Huang et al. 2014) (Ma et al. 2014; WT2) and ESCs (Tan et al. 2019) and processed them as described above to obtain the Log_2 of the UQUA normalized counts (i.e., log_2.UQUA external control). The DE genes in the comparisons of pR and pG vs WT1 were filtered out by removing those genes with abs(log_2.UQUA WT1 - log_2.UQUA external control) > 2 and showing a fold change with opposite sign when compared to patient iPSCs.

RNA-seq Total RNA from ICF1 and WT HPCs was extracted with RNAeasy plus mini kit (74134, Qiagen) according to the manufacturer protocol including a gDNA elimination step. For the RNA-seq experiment, strand-specific TruSeq libraries were prepared according to Illumina's instructions and the libraries were sequenced using the Illumina NovaSeq6000 platform (paired end reads of 150bp length). The RNA-seq data analysis was performed in collaboration with Sequentia Biotech, Barcelona, Spain. First, quality control of PE reads was done by using the fastp tool by setting the following parameters: -f 10 -q 25 -l 35 -w 4 (Chen et al. 2018). Then, the sequences were aligned to the hg38 reference assembly using STAR aligner with following settings: --outSAMtype BAM SortedByCoordinate –alignEndsProtrude 50 ConcordantPair --chimOutType WithinBAM --chimSegmentMin 10 (Dobin et al. 2013). Next, gene expression was quantified as read counts using the FeatureCounts function from Rsubread package and the differentially expressed genes (DE) were identified using the NOIseq R package, for samples with no replicates, after TMM normalization using HTSFilter (Rau et al. 2013) R package. A stringent threshold of posterior probability (pp) > 0.95 and $log_2FC \pm 1.2$ was applied and only genes commonly deregulated in both pR and pG HPCs vs WT1 HPCs were considered. We performed the functional enrichment analysis (GO) of these deregulated genes using *gost* function from gProfiler2 R

package with the following settings: correction_method = FDR < 0.01; user_threshold=0.01 and custom_bg=”expressed_ENSG_HPCs”

***Data visualization*.** Scatterplots showing the log_{10} normalized counts of all expressed genes (Supplemental Fig. S5A) or the log_2FC of the expressed genes associated with ICF1 hypo-DMRs (Supplemental Fig. S5D) or log_2FC of the expressed genes in ICF1 HPCs compared to WT1 (Supplemental Fig. S5E) were generated using the ggplot2 package. We used Upset plots (Supplemental Fig. S5B) to visualize the distribution of DE genes in pR and pG ICF1 iPSCs and their differential expression status in the corrected iPSCs vs WT1 using the UpsetR package (order.by=“freq", keep.order=“TRUE”, group.by="degree", decreasing="TRUE”). We depicted the changes in the expression level of DE genes (pp > 0.8) associated with ICF1 hypo-DMRs and annotated to specific gene features (promoter, exon, intron, 3’UTR) as boxplots (Fig. 3A) using the *geom_boxplot* function (ggplot2). We grouped the DE genes in ICF1 and their corrected clones vs WT1 in each gene feature category. Then, we performed a pairwise comparison (Pairwise Wilcoxon test with two-sided alternative) between the log_2FC of the ICF1 vs WT1 genes associated with promoter hypo-DMRs and those annotated to other genomic feature categories.

To identify the methylation status of deregulated genes in ICF1 iPSCs that were fully/partially/slightly rescued in corrected clones, we subsetted the ICF1 hypo-DMRs that were annotated to these genes and constructed a ComplexHeatmap depicting their methylation level and whether these hypo-DMRs were fully, partial or not rescued (Supplemental Fig. S5C).

Aligned BAM files of each iPSC lines were sorted by chromosome position, converted to bedGraph, and bigWig files and then hosted at Cyverse Discovery Environment. The bigWig

files were visualized as normalised coverage tracks [(Number of reads×1M) / mapped library size)] on UCSC genome browser using the following setting: (tracktype=bigWig, viewLimits=0:1, windowingFunction=mean, smoothingWindow=10, visibility=full).

ChIP-seq data analysis

We assessed the quality of the sequenced single-end (SE) reads from IP and input using FastQC prior and post-trimming. Using cutadapt, we filtered and adaptor-trimmed the SE reads using the following parameters: -q > 30; -m > 40; -a AGATCGGAAGAG. We aligned the retained reads to the custom hg38 reference genome with canonical chromosomes only (GRCh38/hg38 primary assembly) using Bowtie 2 with the default parameters. Furthermore, we filtered out aligned SE reads with MAPQ score < 20 using the *samtools view* -q 20 setting (Danecek et al. 2021), then removed PCR duplicates using the *samtools markdup* command.

Peak calling and DERs detection. We identified the narrow H3K4me3 peaks (enriched regions) and differentially enriched regions (DERs) for each IP replicate using the *sicer_df* command (SICER2), setting the following arguments: -f 200, -fdr 0.00001, -fdr_df 0.01, -egf 0.88, -w 200, -g 200, -s hg38. For broad H3K36me3 peaks (enriched regions), we increased the gap size to -g 1400 (optimized by plotting a curve with the sum of island counts at different gap sizes). Since DNMT3B peak structure reflects H3K36me3 peak enrichment, we also set -g to 1400 with -fdr and -fdr_df set to 0.0001. After that, we removed the enriched peaks and DERs overlapping the ENCODE hg38 blacklist regions (Amemiya et al. 2019) with the *intersectBed* function from the BEDtools suite (Quinlan et al. 2010). We defined the consensus peaks and DERs as those overlapping by at least 1bp in both IP replicates. DERs with contradictory fold change (FC) direction (increased or decreased) between IP replicates

were removed. We also defined the increased H3K4me3 DERs identified in ICF1 iPSCs as rescued in corrected iPSCs if not detected in the list of increased H3K4me3 DERs compared to WT1 or if called as decreased H3K4me3 DERs in this comparison. We used the same criterion to define the rescue of decreased DNMT3B DERs in corrected iPSCs.

Calculation of the ChIP-seq enrichment. To calculate the enrichment of an IP across ROIs, we first counted the reads in both IP and input using the *featureCounts* function with the following settings: annotation=custom SAF file, useMetaFeatures=FALSE, allowMultiOverlap=FALSE, strandSpecific=0, CountMultipleMappingReads=TRUE). Then, we performed batch-effect removal using the *ArySyNseq* function (factor="run", batch=TRUE, norm="n") to remove the bias likely introduced by the different sequencing runs (if necessary). Subsequently, we normalized the read counts by the ChIP-seq library size and computed the fold change (FC) = Normalized read count of IP/Normalized read count of input (IP/input) (Fig. 4A and Fig. 5C).

We first identified the DNMT3B peaks in WT1 and then computed the FC of these peaks across all the iPSCs (Supplemental Fig. S6A). Next, we calculated the FC ratio of WT1 to pR/pG and identified WT1 peaks showing lower DNMT3B enrichment in ICF1 iPSCs by applying a threshold of FC ratio > 1.2. From these peaks, we further subsetted those showing higher enrichment (FC) scores in corrected clones compared to ICF1 to compute the rescue percentage described in the results**.**

We obtained the CTCF motif containing ICF1 hypo-DMRs using the HOMER suite (Fig. 5A) and calculated the CTCF FC (IP/input) across pR vs WT1 hypo-DMRs (WT1, pR, cR7) and pG vs WT1 hypo-DMRs (WT1, pG, cG50) (Fig. 5B). To quantify the statistical significance of the increased CTCF enrichment in Groups 1 and 2 compared to Groups 3 and 4, we

calculated the difference between ICF1(FC) – WT(FC) and corrected clone(FC) – WT1(FC) for each group and applied non-parametric paired Wilcoxon test with one-sided alternative and BH-FDR correction. Finally, we computed the effect size using *rank_biserial* function from the effect size R package (Ben-Shachar et al. 2022) between the mentioned differences.

Data visualization. Following the TF motif enrichment analysis at ICF1 hypo-DMRs intersecting with DNMT3B decreased DERs, the TF binding at the hypo-DMRs was confirmed by calculating the significance of the overlap between the hypo-DMRs and the TF ChIP-seq enriched peaks (ENCODE Accession number for E2A: ENCFF658WIO; EBF1: ENCFF249SVT; derived from GM12878 LCLs) and represented as Venn diagrams (Fig. 2C). We generated the multi-omics integrated heatmaps (Fig. 4A) displaying the DNA methylation levels, histone marks and DNMT3B enrichment, genomic annotation, and the rescue status of hypo-DMRs associated with CGI and GH promoters and enhancers in pR and pG iPSCs using the *heatmap* function from the ComplexHeatmap R package. We divided the genomic features annotated to the hypo-DMRs into three groups: “Promoter”, “Gene body” (Intron/Exon/5’UTR/3’UTR) and “Distal intergenic”. The hypo-DMR status and the rows of the heatmap were divided into four clusters, based on Groups 1-4 derived from the heatmap in Fig.1A. The color scale for methylation levels and ChIP-seq enrichment (FC) was produced using the *RcolorBrewer* package.

We plotted the H3K4me3 fold enrichment (FC) density plot (Fig. 4B) at H3K4me3 DERs overlapping pR and pG hypo-DMRs using the geom_density_ridges function from the ggridges R package. For constructing the density plot, we calculated the H3K4me3 FC at the ROIs across pR and pG iPSCs and grouped them into three categories based on the ranking of FC from lowest to highest in ICF1 iPSCs.

Using *ViewBS MethOverRegion*, we plotted the average methylation profile of WT1, ICF1 and corrected clones across the hypo-DMRs overlapping with ICF1 H3K4me3 increased DERs, which are distinguished into the “H3K4me3 corrected” and “H3K4me3 uncorrected” subgroups (Fig. 4C). Using *MethHeatmap* function, we computed the methylation level for each of the above hypo-DMRs and we measured for each region the difference in DNA methylation levels between ICF1 and WT1, as well as between corrected clones and WT1. Then, we estimated for each region the regain of methylation between corrected clones and ICF1 [(corr-WT1) - (ICF1-WT1)]. After that, we tested that the regain of methylation was higher in the “H3K4me3 corrected” subgroup than in the “H3K4me3 uncorrected” using the Welch's *T*-test with a one-side alternative, and we estimated the effect size as the mean regain of methylation within the subgroup.

We generated the hybrid plots (Dot, box and violin plot; Supplemental Fig. S6A) to represent the fold enrichment of DNMT3B in pR-related (pR, cR7, cR35) and pG-related (pG, cG13, cG50) samples using the ggstatsplot R package (Patil 2021) with the following parameters: plot.type="boxviolin", type="nonparametric", p.adjust.method="BH", centrality.type=“parametric” (to denote the mean enrichment score). Then, we performed a Pairwise Wilcoxon test with two-sided alternative to compare the WT1 DNMT3B peaks fold enrichment (FC) to the other samples in the plot using the *Stat_compare_means* function from the ggpubr R package (method=”wilcox.test”, paired=TRUE and p.adjust.methods="fdr"). The genes expressed in WT1 were ranked by their $\log_2$ FPKM normalized counts and divided into four quartiles with Q1 and Q4 denoting the lowest and highest expressed genes in WT1 iPSCs, respectively. The ChIP-seq enrichment of WT1 for the entire list of ranked expressed genes was represented using a heatmap, while the average enrichment across each quartile

was visualized as a profile plot. We used the *ngs.plot.r* command with the following settings: "-G hg38, -R gene body (DNMT3B/H3K36me3) or TSS (H3K4me3), -E custom_gene_list, -GO none, -YAS 0,0.10/0.15/1.5 (DNMT3B/H3K36me3/H3K4me3), -L 2000, -LEG 1, -SE 0" to obtain the heatmaps and profile plots (Supplemental Fig. S6B).

The DNMT3B peaks in wild-type human ESCs were obtained from public datasets (Verma et al. 2018 and Tan et al. 2019) and uploaded along with our datasets. We assessed the statistical significance of the overlap between WT1 DNMT3B peaks and the hESC wild-type DNMT3B peaks by shuffling using *enrichPeakOverlap* function, as described previously (*P*-adj < 0,0001; shuffle test).

We represented the CTCF enrichment (FC) across ICF1 hypo-DMR groups as violin-boxplots using the *geom_violin* and *geom_boxplot* function from ggplot2 (Fig. 5B). We utilized the *ComplexHeatmap* R package to simultaneously visualize the methylation level, CTCF enrichment (FC), and presence of ICF1 H3K4me3 peaks along with the rescue category of the CTCF motif containing hypo-DMRs (Fig. 5C).

To evaluate the correlation between the DNA methylation level of hypo-DMRs given in Fig 4A (7125 in pR and 7495 in pG) or in Fig.5C (1566 in pR and 1584 in pG) and their H3K4me3 enrichment level (FC) or their CTCF binding level (FC), respectively, we first binned the DNA methylation level (0-100%) into 200 bins and calculated the average FC (IP/input) of ICF1 iPSCs for each bin. Then, we plotted the methylation level against average FC per bin using *ggscatterstats* function from ggstatsplot R package and extracted summary statistics, which included the Pearson correlation coefficient (r) and *P*-value.

We obtained the bigWig coverage tracks of ChIP-seq replicates and input by converting BAM to BED files, extending the read length by 100bp (fragment size ~ 200bp) using *slopBed* and

computing the read coverage using the *genomecov* function in the BEDtools suite. We normalized the bedGraph tracks to the mapped library size, sorted and converted to bigWig files, and hosted them at Cyverse Discovery Environment and uploaded on a UCSC genome browser session with the following setting: (track type=bigWig, visibility=full,viewLimits=default, windowingFunction=mean, smoothingWindow=10, color=0,0,255[DNMT3B]/ 128,0,128[H3K36me3]/ 0,100,0[H3K4me3]).
The enriched consensus peaks and DERs were also uploaded as bedGraph tracks.

Reference genome assembly for sequencing alignment and annotation

We used the Ensembl GRCh38/hg38 human reference genome assembly release v102 for iPSCs and v104 for HPCs (Homo_sapiens.GRCh38.dna_sm.primary_assembly.fa.gz) for performing sequence alignment. For transcriptome annotation, we used the *Homo_sapiens.GRCh38_canon_chr_header.gtf* (a filtered version of GRCh38/hg38.v85/v105 annotation file that includes only canonical chromosomes) downloaded from the Ensembl FTP server.
The hg38 chromosome sizes were downloaded via the UCSC Genome Browser FTP server. We downloaded the CpG islands (cpgislandExt), Genehancer regulatory elements, and the ENCODE cCREs (hg38) tracks using the UCSC table browser. We obtained the DNA-methylation valleys (DMVs) BED file from (Xie et al. 2013). For the DMRs annotation to the genomic features, we built the TxDb object from the hg38 Ensembl database using the *makeTxDbFromEnsembl* function in the *ChIPseeker* R package. The hg19 tracks downloaded from public databases were converted to hg38 using the *LiftOver* command-line tool.

List of website links of data, software and tools utilized in the present article

Sofware/Package	Version	Source
BEDtools	v2.29.2	https://github.com/arq5x/bedtools2
Bismark	v0.19.1	https://github.com/FelixKrueger/Bismark
Bowtie 2	v2.3.4.3	https://github.com/BenLangmead/bowtie2
ChIPseeker	v1.29.1	https://bioconductor.org/packages/ChIPseeker/
ChIPpeakAnno	v3.27.6	https://bioconductor.org/packages/release/bioc/html/ChIPpeakAnno.html
ComplexHeatmap	v2.9.4	https://bioconductor.org/packages/ComplexHeatmap/
Circlize	v0.4.13	https://cran.r-project.org/package=circlize
Cutadapt	v1.9.1	https://pypi.org/project/cutadapt/1.9.1/
DeepTools	v3.3.2	https://github.com/deeptools/deepTools
Effect size	v0.8.2	https://cran.r-project.org/package=effectsize
FastQC	v0.11.5	https://www.bioinformatics.babraham.ac.uk/projects/fastqc/
fastp	v0.23.1	http://opengene.org/fastp/fastp.0.23.1
GenomicRanges	v1.45.2	https://bioconductor.org/packages/release/bioc/html/GenomicRanges.html
gtrellis	v1.22.0	https://bioconductor.org/packages/release/bioc/html/gtrellis.html
ggplot2	v3.3.3	https://cran.r-project.org/package=ggplot2
ggfortify	v0.4.12	https://cran.r-project.org/package=ggfortify
ggridges	v0.5.3	https://github.com/cran/ggridges
ggstatsplot	v0.8.0	https://indrajeetpatil.github.io/ggstatsplot/
g:Profiler	v0.2.0	https://cran.r-project.org/web/packages/gprofiler2/

GRCh38/hg38 GTF	v85	http://ftp.ensembl.org/pub/release85/gtf/homo_sapiens/Homo_sapiens.GRCh38.85.chr.gtf.gz
GRCh38/hg38 GTF	v105	http://ftp.ensembl.org/pub/release-105/gtf/homo_sapiens/Homo_sapiens.GRCh38.105.chr.gtf.gz
GRCh38/hg38 Chromosome sizes		http://hgdownload.cse.ucsc.edu/goldenpath/hg38/bigZips/hg38.chrom.sizes
HISAT2	v2.1.0	https://github.com/DaehwanKimLab/hisat2
HOMER	v4.11	http://homer.ucsd.edu/homer/
HTSFilter	v1.38.0	https://bioconductor.org/packages/HTSFilter/
methylKit	v1.16.0	https://bioconductor.org/packages/release/bioc/html/methylKit.html
MultiQC	v1.9	https://pypi.org/project/multiqc/1.9/
ngs.plot	v2.63	https://github.com/shenlab-sinai/ngsplot
NOIseq	v2.34.0	https://bioconductor.org/packages/release/bioc/html/NOISeq.html
PANTHER	v16.0	http://www.pantherdb.org/
patternplot	v1.0.0	https://cran.r-project.org/package=patternplot
plotrix	v3.8-2	https://cran.r-project.org/package=plotrix
primer-blast		https://www.ncbi.nlm.nih.gov/tools/primer-blast/
Python	v.2.7/3.7.6	https://www.python.org/downloads/release/python-376/
R	v3.6.3	https://cran.r-project.org/bin/linux/ubuntu/#install-r
Rstudio	v1.2.1335	https://www.rstudio.com/products/rstudio/release-notes/rstudio-1-2/
ReviGO		http://revigo.irb.hr/
Rsubread	v2.7.3	https://bioconductor.org/packages/release/bioc/html/Rsubread.html

SICER2	v2.0	https://zanglab.github.io/SICER2/
SRAToolKit	v2.10.4	https://github.com/ncbi/sra-tools
STAR	v2.7.9a	https://github.com/alexdobin/STAR/archive/2.7.9a.tar.gz
upsetR	v1.4.0	https://cran.r-project.org/package=UpSetR
VennDiagram	v1.6.20	https://cran.r-project.org/package=VennDiagram
ViewBS	v0.1.11	https://github.com/xie186/ViewBS

SUPPLEMENTAL REFERENCES

Amemiya HM, Kundaje A, Boyle AP. 2019. The ENCODE Blacklist: Identification of Problematic Regions of the Genome. *Sci Rep* **9**: 9354.

Ben-Shachar M, Lüdecke D, Makowski D 2020. effectsize: Estimation of Effect Size Indices and Standardized Parameters. *J. Open Source Software*, **5**: 2815.

Chen H, Boutros PC. 2011. VennDiagram: a package for the generation of highly-customizable Venn and Euler diagrams in R. *BMC Bioinformatics* **12**: 35.

Chen S, Zhou Y, Chen Y, Gu J. 2018. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **34**: i884–i890.

Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, Whitwham A, Keane T, McCarthy SA, Davies RM, et al. 2021. Twelve years of SAMtools and BCFtools. *Gigascience* **10**. Giab008.

Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. 2013. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**:15-21.

Ewels P, Magnusson M, Lundin S, Käller M. 2016. MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics* **32**: 3047–3048.

Gu, Z., Eils, R. & Schlesner, M. 2016. gtrellis: an R/Bioconductor package for making genome-level Trellis graphics. *BMC Bioinform.* **17**, 169.

Huang X, Zhang S, Li K, Thimmapuram J, Xie S, Wren J. 2018. ViewBS: a powerful toolkit for visualization of high-throughput bisulfite sequencing data. *Bioinformatics* **34**: 708–709.

Katz K, Shutov O, Lapoint R, Kimelman M, Brister JR, O'Sullivan C. 2022. The Sequence Read Archive: a decade more of explosive growth. *Nucleic Acids Res* **50**: D387–D390.

Lister R, Mukamel EA, Nery JR, Urich M, Puddifoot CA, Johnson ND, Lucero J, Huang Y, Dwork AJ, Schultz MD, et al. 2013. Global epigenomic reconfiguration during mammalian brain development. *Science* **341**: 1237905.

Merchant N, Lyons E, Goff S, Vaughn M, Ware D, Micklos D, Antin P. 2016. The iPlant Collaborative: Cyberinfrastructure for Enabling Data to Discovery for the Life Sciences. *PLoS Biol* **14**: e1002342.

Navarro Gonzalez J, Zweig AS, Speir ML, Schmelter D, Rosenbloom KR, Raney BJ, Powell CC, Nassar LR, Maulding ND, Lee CM, et al. 2021. The UCSC Genome Browser database: 2021 update. *Nucleic Acids Res* **49**: D1046–D1057.

Patil, I. 2021. Visualizations with statistical details: The 'ggstatsplot' approach. *J. Open Source* Software, **6**, 3167.

Quinlan AR, Clark RA, Sokolova S, Leibowitz ML, Zhang Y, Hurles ME, Mell JC, Hall IM. 2010. Genome-wide mapping and assembly of structural variant breakpoints in the mouse genome. *Genome Res* **20**: 623–635.

Rau A, Gallopin M, Celeux G, Jaffrézic F. 2013.Data-based filtering for replicated high-throughput transcriptome sequencing experiments. *Bioinformatics* **29**: 2146–2152.

Song Q, Decato B, Hong EE, Zhou M, Fang F, Qu J, Garvin T, Kessler M, Zhou J, Smith AD. 2013. A reference methylome database and analysis pipeline to facilitate integrative and comparative epigenomics. *PLoS One* **8**: e81148.

Supek F, Bošnjak M, Škunca N, Šmuc T. 2011. REVIGO summarizes and visualizes long lists of gene ontology terms. *PLoS One* **6**: e21800.

SUPPLEMENTAL FIGURES

Supplemental Figure S1

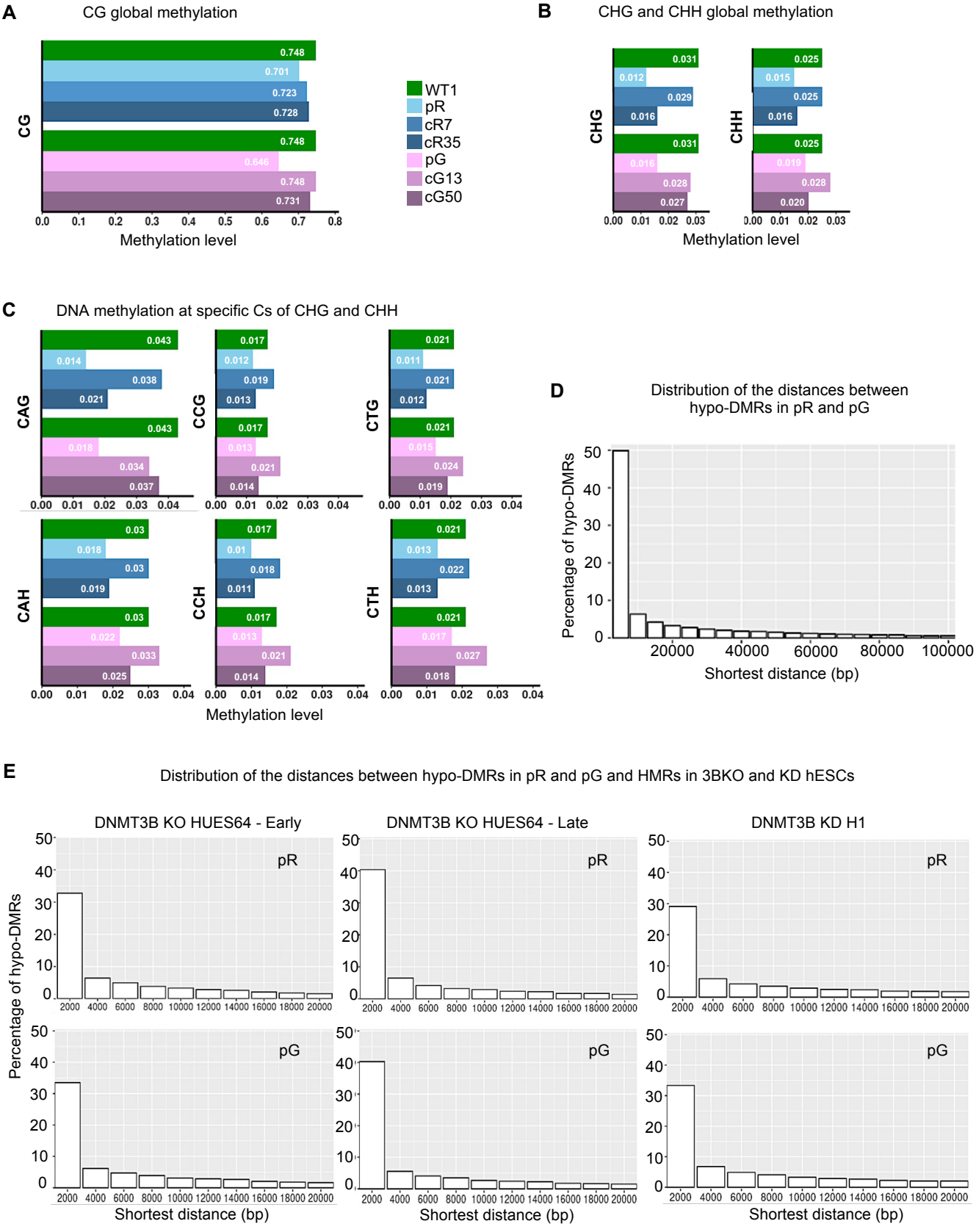
Supplemental Figure S2

Supplemental Figure S3

Supplemental Figure S4

Supplemental Figure S5

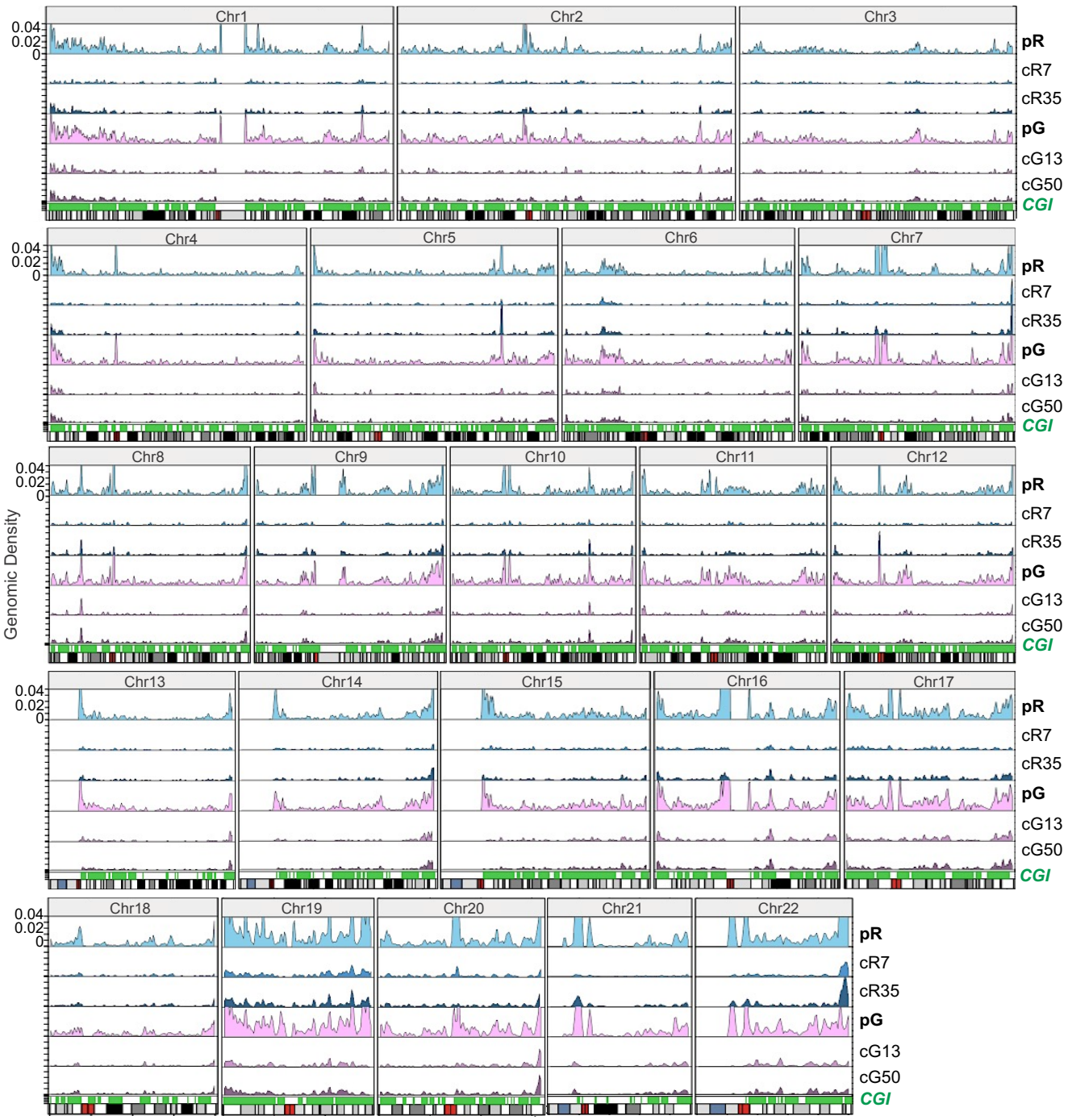
Supplemental Figure S6


A
CG global methylation
CG
0.748
0.701
0.723
0.728
0.748
0.646
0.748
0.731
WT1
pR
cR7
cR35
pG
cG13
cG50
Methylation level
B
CHG and CHH global methylation
CHG
0.031
0.012
0.029
0.016
0.031
0.016
0.028
0.027
CHH
0.025
0.015
0.025
0.016
0.025
0.019
0.028
0.020
Methylation level
C
DNA methylation at specific Cs of CHG and CHH
CAG
CCG
CTG
CAH
CCH
CTH
Methylation level
D
Distribution of the distances between hypo-DMRs in pR and pG
Percentage of hypo-DMRs
Shortest distance (bp)
E
Distribution of the distances between hypo-DMRs in pR and pG and HMRs in 3BKO and KD hESCs
DNMT3B KO HUES64 - Early
DNMT3B KO HUES64 - Late
DNMT3B KD H1
pR
pG
Percentage of hypo-DMRs
Shortest distance (bp)

Supplemental Figure S1

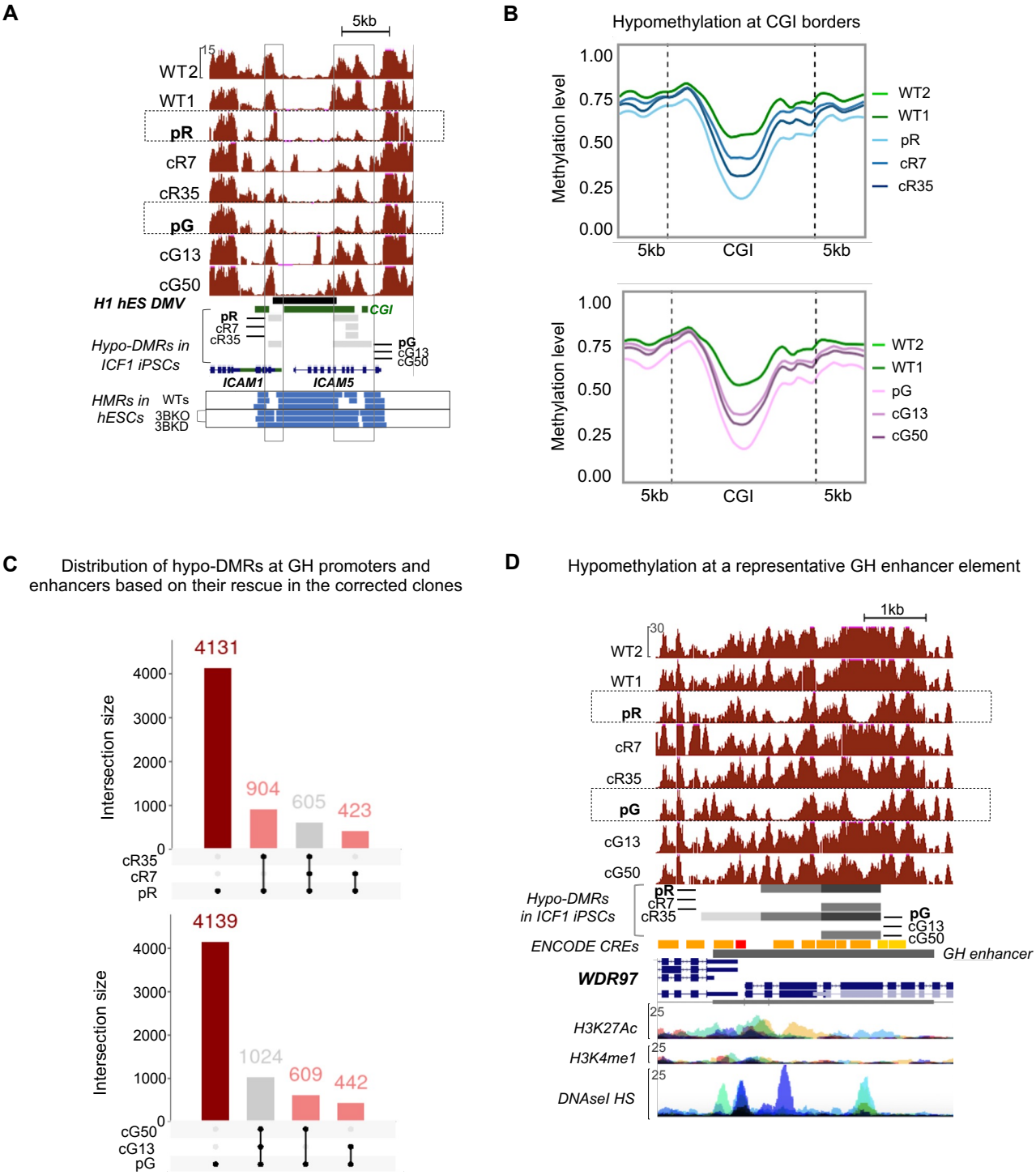
(A,B) Average weighted global methylation levels of cytosines as determined by WGBS analysis (weighted based on region size and number of CGs and CHs) at the context of CG, CHG, and CHH (H=A/G/T), in WT1, ICF1, and corrected iPSCs, cR7, cR35, cG13 and cG50. The X-axis denotes the methylation level expressed as the ratio of the number of Cs over the total number of Cs and Ts. The Y-axis indicates the cytosine context. (*C)* Average weighted global methylation levels of mCA, mCC, and mCT at the context of CHH and CHG, expressed as the ratio of the number of Cs over the total number of Cs and Ts in WT1, ICF1, and corrected iPSCs. *(D)* Histogram of the distances between individual hypo- DMRs in pR and pG iPSCs. Each bin in the X-axis denotes the shortest distance (bp) between the compared regions, while the Y-axis indicates the percentage of hypo- DMRs in each bin of 5kb size. *(E)* Histogram of the distances between ICF1 hypo-DMRs and hypomethylated regions (HMR) in early/late DNMT3B knock-out (KO) and DNMT3B knock-down (KD) hESCs. Each bin in the X-axis denotes the shortest distance (bp) between the compared regions, while the Y-axis indicates the percentage of hypo- DMRs in each bin of 2kb size. The overlap between pR and pG hypo-DMRs and between ICF1 hypo- DMRs and HMRs in 3BKO (early and late) is significant (*P*-value < 0.0001; shuffle test).

Chromosomal distribution of hypo-DMRs in patient and corrected iPSCs



Supplemental Figure S2

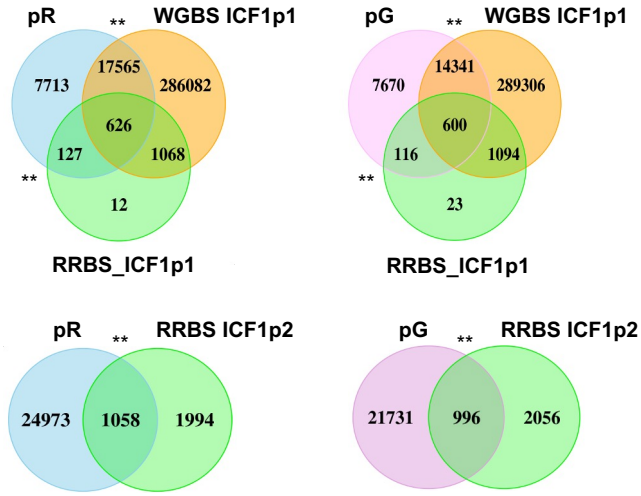
Genomic density line plots of hypo-DMRs in ICF1 pR and pG iPSCs and their respective corrected clones. The 22 autosomes are depicted as ideograms in the X-axis with the red band denoting the centromere position. For each chromosome from top to bottom: pR and pG hypo-DMRs line plots followed by CGI distribution (green). The Y-axis of the line plots represents the density of hypo-DMRs defined as the proportion of the regions of interest present in each defined genomic window. Hypo-DMRs and CGI are partitioned into genomic windows of 2Mb and 1kb respectively. The statistical significance of the correlation between the number of hypo-DMRs and the number of CGI in the same genomic window was analyzed using the Poisson regression *(P*-value < 10-5).


A
5kb
WT2
WT1
pR
cR7
cR35
pG
cG13
cG50
H1 hES DMV
CGI
Hypo-DMRs in ICF1 iPSCs
ICAM1
ICAM5
HMRs in hESCs
WTs
3BKO
3BKD
B
Hypomethylation at CGI borders
Methylation level
5kb
CGI
5kb
C
Distribution of hypo-DMRs at GH promoters and enhancers based on their rescue in the corrected clones
Intersection size
4131
904
605
423
4139
1024
609
442
D
Hypomethylation at a representative GH enhancer element
1kb
Hypo-DMRs in ICF1 iPSCs
ENCODE CREs
GH enhancer
WDR97
H3K27Ac
H3K4me1
DNaseI HS
E
Methylation level of GH promoters and enhancers associated hypo-DMRs
Methylation level
Not rescued n=605
Rescued n= 4131
Not rescued n=1024
Rescued n= 4139

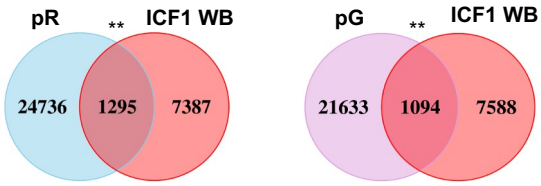
Supplemental Figure S3

(A) A genome browser view of Differentially Methylated Valleys (DMVs) in the *ICAM5* gene representing an example of methylation loss at the borders of a DMV carrying a CGI in ICF1 iPSCs compared to WT counterparts. Dark tracks denote methylation coverage measured by WGBS in all iPSCs. The six tracks at the bottom illustrate the hypomethylated regions (HMRs) in WT human embryonic stem cells (hESCs) from H1, HUES9 and H9 lines, followed by DNMT3B- KO hESCs (early and late passage DNMT3B-KO; 3BKO) and shRNA DNMT3B-KD hESCs (3BKD). *(B)* Average methylation level across pR (n=3374) and pG (n=3162) hypo-DMRs overlapping CGIs and 5kb flanking regions in WT1, ICF1 and corrected iPSCs. *(C)* Upset plots displaying the hypo-DMRs at GH promoters and enhancers (+/-2kb) in pR (top) and pG (bottom) iPSCs and their corrected counterparts. Vertical bars represent the intersection size between the hypo-DMRs present in each iPSC sample. Connected black dots below each plot represent the hypo-DMRs present in each intersection. Dark red bars correspond to hypo-DMRs present in pR or pG iPSCs (black dot) and absent, therefore rescued, in both cR7 and cR35 (n=4131) or in cG13 and cG50 (n=4139). Pink bars correspond to hypo-DMRs present in pR or pG iPSCs and in one corrected clone, but absent in the second corrected clone. Grey bars represent the hypo-DMRs that are present in patients and corrected iPSCs, and therefore are resistant to de novo methylation following the correction of the DNMT3B mutations. *(D)* A genome browser view of the *WDR97* gene, representing an example of methylation loss at enhancer regions (ENCODE cCRE and GeneHancer, GH) in pR and pG iPSCs, and ICF1 LCLs. Red tracks display methylation coverage measured by WGBS and grey boxes represent hypo-DMRs detected in pR, pG and corrected iPSCs compared to WT1 and WT2 iPSCs. The tracks underneath indicate the gene regulatory elements including enhancers reported in ENCODE cis Regulatory Elements (cCRE) and in GH databases, as well as the H3K4me1, H3K27Ac enrichment and DNAseI sites. *(E)* Boxplots representing the distribution of methylation levels following WGBS analysis (expressed as the ratio of the number of Cs over the total number of Cs and Ts) at GH promoters and enhancers associated hypo-DMRs in ICF1 iPSCs, which either remain hypomethylated (n=605 for pR and 1024 for pG), or that are rescued (n=4131 for pR and 4139 for pG) in the corresponding isogenic clones.

A Hypo-DMRs in ICF1 iPSCs intersecting with hypomethylated regions in ICF1p1 and ICF1p2 vs WT LCLs



B Hypo-DMRs in ICF1 iPSCs intersecting with hypomethylated regions in ICF1 patient vs WT whole blood samples

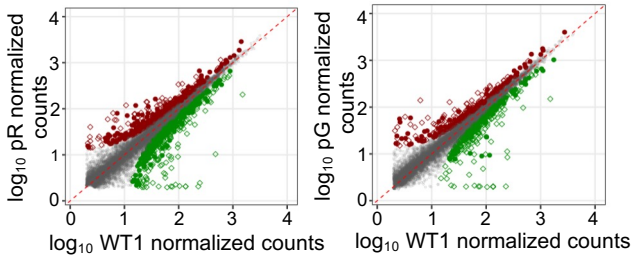


C

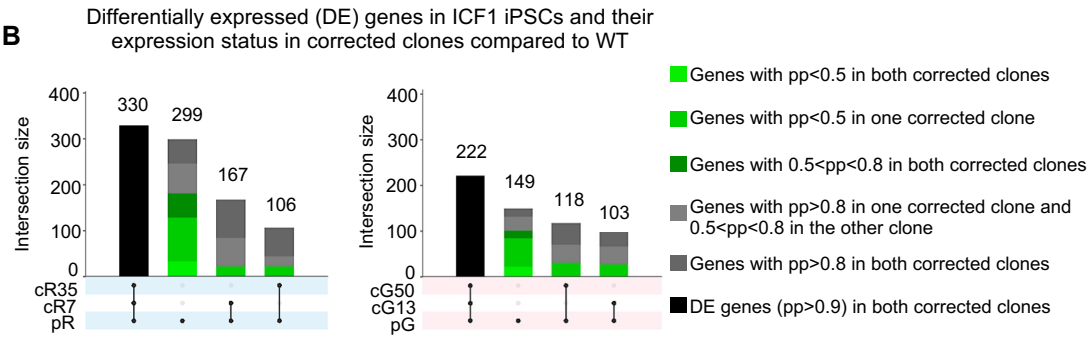
Supplemental Figure S4

(A,B) Venn diagrams representing the intersection of hypo-DMRs in pR and pG iPSCs with *(A)* hypo-DMRs in ICF1p1 LCL compared to WT LCL, obtained from RRBS and WGBS experiments (top), and ICF1p2 LCL, obtained from RRBS experiment (Gatto et al.2017) (bottom), *(B)* hypo-DMRs in ICF1 patient whole blood compared to WT whole blood (ICF1 WB) (39). Statistical significance of the overlap was calculated using the shuffle method (P-adj < 0.001). *(C)* A genome browser view of the hypomethylated genomic regions across *IGH* cluster genes of Chromosome 14 in WT iPSCs compared with ICF1 iPSCs and corrected counterparts. The grey boxes represent the hypo-DMRs in ICF1 iPSCs and their respective corrected clones. Blue tracks below illustrate the hypomethylated regions (HMRs) in WT human embryonic stem cells (hESCs) from H1, HUES9 and H9 lines, followed by DNMT3B-KO hESCs (early and late passage DNMT3B-KO; 3BKO) and shRNA DNMT3B-KD hESCs; 3BKD).

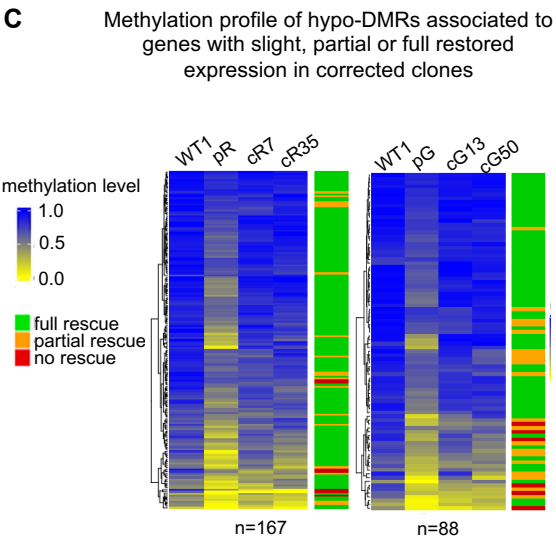
A



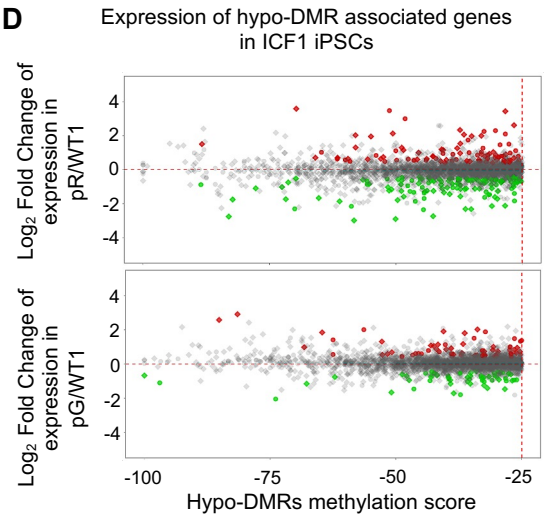
B



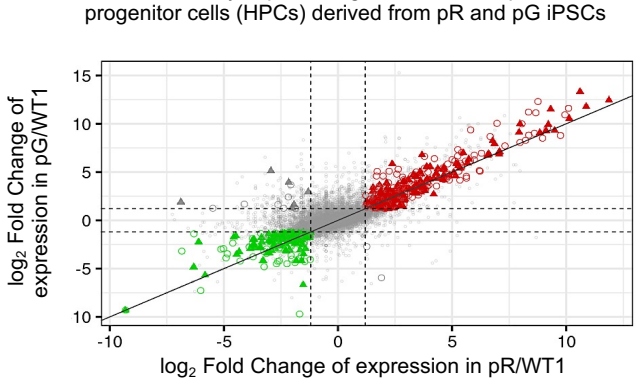
C



D



E



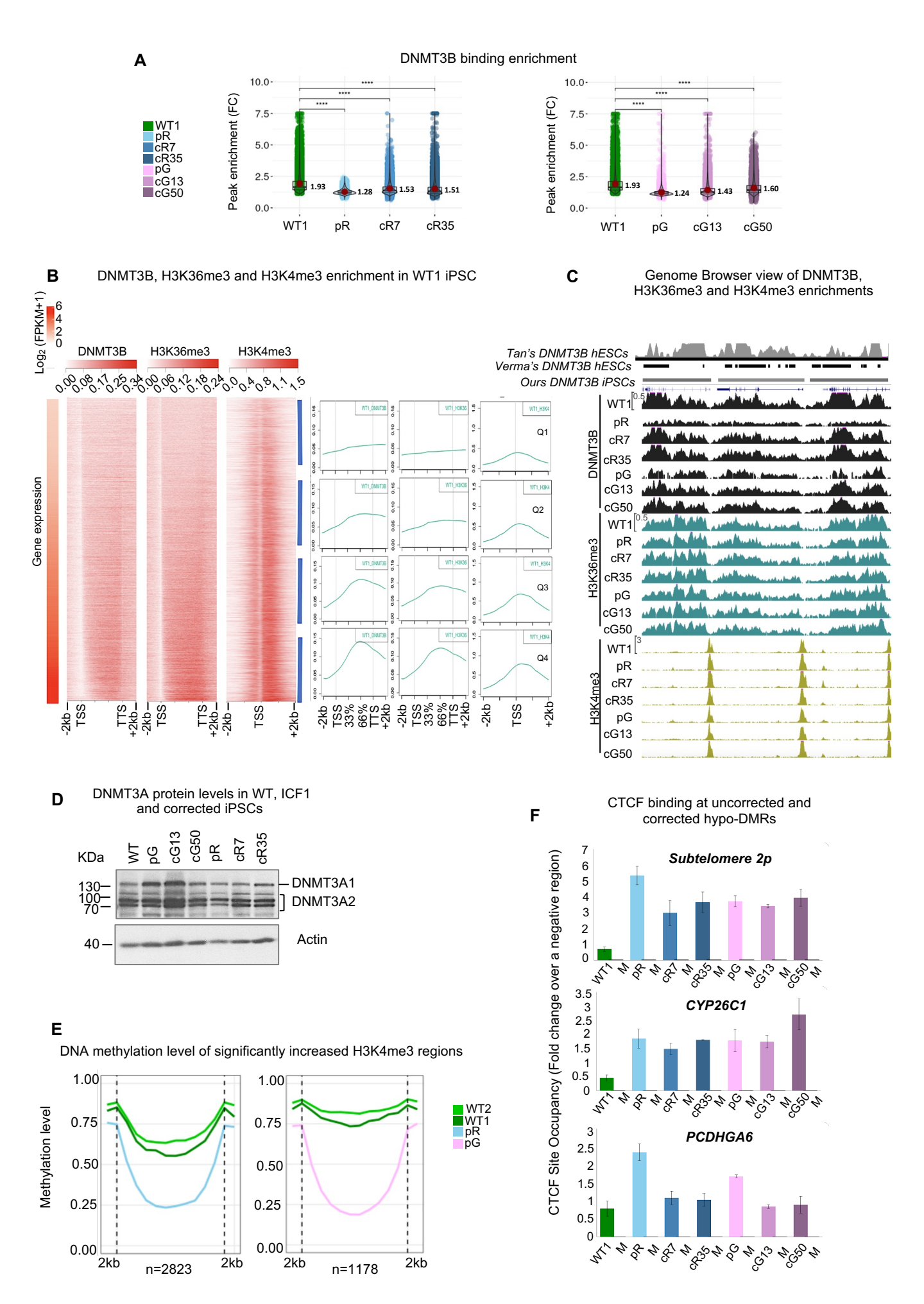
F

Gene Ontology analysis

Source	Term name	*P*-adj	genes
GO:BP	immune response	4.95×10^{-23}	129
GO:BP	immune system process	1.38×10^{-21}	174
GO:BP	inflammatory response	1.73×10^{-21}	83
GO:BP	defense response to other organism	3.03×10^{-17}	89
GO:BP	response to external stimulus	5.79×10^{-17}	163
GO:BP	biological process involved in interspecies interaction between organisms	2.96×10^{-14}	109
KEGG	Neutrophil extracellular trap formation	3.37×10^{-7}	27
KEGG	Staphylococcus aureus infection	4.7×10^{-7}	13
KEGG	Phagosome	2.37×10^{-6}	21
KEGG	Cytokine-cytokine receptor interaction	2.37×10^{-6}	22
KEGG	Viral protein interaction with cytokine and cytokine receptor	3.15×10^{-6}	13
KEGG	Hematopoietic cell lineage	7.86×10^{-6}	14
REAC	Immune System	4.62×10^{-19}	147
REAC	Innate Immune System	2.90×10^{-12}	85
REAC	Neutrophil degradation	5.04×10^{-12}	56
REAC	Interleukin-10 signaling	4.44×10^{-7}	12
REAC	Immunoregulatory interactions between Lymphoid and a non-Lymphoid cell	6.49×10^{-7}	18
REAC	Cytokine Signaling in Immune system	6.49×10^{-7}	55

Supplemental Figure S5

(A) Scatterplots of gene expression levels depicted as log10 of UQUA (Upper Quartile) normalized counts in pR (left) and pG (right) iPSCs (Y-axis) compared to WT1 iPSCs (X-axis). The differentially expressed (DE) genes with posterior probability (pp) > 0.9 in ICF1 compared to WT1 iPSCs are displayed in red (up-regulated: 392 in pR and 260 in pG) and green (down-regulated: 510 in pR and 332 in pG), while the statistically non-significant genes are shown in grey. The DE genes in both or in either pR or pG iPSCs are denoted as diamonds and circles, respectively. *(B)* Upset plots showing the distribution of DE genes (pp > 0.9) in pR (left) and pG (right) iPSCs and their corrected counterparts. Vertical bars represent the intersection size of DE genes between the depicted iPSCs. Connected black dots at the bottom panel represent DE genes present in each intersection. In detail, the black bar indicates DE genes in ICF1 iPSCs and both their corrected clones compared to WT, whereas the second bar to the right includes DE genes only in patient iPSCs and shows full (light green; pp < 0.5 in both clones), partial (medium green; pp < 0.5 in only one clone) or slight (dark green; 0.5 < pp < 0.8 in both clones) rescue in their respective corrected clones. Light grey represents genes with pp > 0.8 in one clone and 0.5 < pp < 0.8 in the second clone, while dark grey denotes genes with pp > 0.8 in both corrected clones. The last two bars on the right represent DE genes in pR or pG and in one of the two corrected clones, with the second clone being partially or slightly rescued (medium green and dark green). *(C)* A heatmap illustration of methylation level across hypo-DMRs annotated to deregulated genes that are slightly, partially or fully rescued in both corrected clones of pR (n=167) and pG (n=88). The heatmap denotes methylation level across iPSC lines along with the rescue category for each hypo-DMR indicated. *(D)* Scatterplots describing distribution of genes annotated to hypo-DMR associated regulatory elements in pR (n=3181) and pG iPSCs (n=3319). The X-axis denotes the differential methylation score (expressed as the difference in the methylation percentages) of the hypo-DMRs associated with each gene (average in the case of multiple hypo-DMRs in one gene) and the Y-axis denotes the log2 fold change of the expressed genes. The genes common to both patient iPSCs or unique to pR/pG iPSCs are indicated as diamonds and circles, respectively. Up-regulated and down-regulated genes (pp > 0.9) are depicted in red and green respectively. *(E)* A scatterplot of expressed genes, depicted as log2 Fold Change, in hematopoietic progenitor cells (HPCs) derived from ICF1 iPSCs compared to HPCs derived from WT iPSCs (pp > 0.95 and |log2FC| > 1.2). Out of 658 commonly deregulated genes in pR and pG HPCs, 422 are up- regulated (red) and 220 are down-regulated (green). Triangles indicate deregulated genes associated with hypo-DMRs ICF1 in patient iPSCs. *(F)* Top five enriched terms from Gene Ontology (GO) analysis of the deregulated genes in pR and pG HPCs including Biological Processes (GO:BP), KEGG and REACTOME Pathways. *P*-adj of the GO terms correspond to Benjamini-Hochberg False Discovery Ratio (BH-FDR) < 0.01.


A
DNMT3B binding enrichment
WT1
pR
cR7
cR35
pG
cG13
cG50
Peak enrichment (FC)
10.0
7.5
5.0
2.5
0.0
****
****
****
1.93
1.28
1.53
1.51
WT1
pR
cR7
cR35
1.93
1.24
1.43
1.60
WT1
pG
cG13
cG50
B
DNMT3B, H3K36me3 and H3K4me3 enrichment in WT1 iPSC
Log2 (FPKM+1)
DNMT3B
H3K36me3
H3K4me3
Gene expression
Q1
Q2
Q3
Q4
-2kb
TSS
TTS
+2kb
33%
66%
C
Genome Browser view of DNMT3B, H3K36me3 and H3K4me3 enrichments
Tan's DNMT3B hESCs
Verma's DNMT3B hESCs
Ours DNMT3B iPSCs
DNMT3B
H3K36me3
H3K4me3
WT1
pR
cR7
cR35
pG
cG13
cG50
D
DNMT3A protein levels in WT, ICF1 and corrected iPSCs
KDa
WT
pG
cG13
cG50
pR
cR7
cR35
130
100
70
40
DNMT3A1
DNMT3A2
Actin
E
DNA methylation level of significantly increased H3K4me3 regions
Methylation level
1.00
0.75
0.50
0.25
0.00
2kb
n=2823
n=1178
WT2
WT1
pR
pG
F
CTCF binding at uncorrected and corrected hypo-DMRs
CTCF Site Occupancy (Fold change over a negative region)
Subtelomere 2p
CYP26C1
PCDHGA6
WT1
M
pR
cR7
cR35
pG
cG13
cG50

Supplemental Figure S6

(A) Hybrid plots (boxplots and dotplots) showing the fold enrichment of endogenous DNMT3B binding at its target regions (n=19706 binding sites observed in WT1) obtained by ChIP-seq analysis of WT1, ICF1 and their corrected iPSCs. FC indicates the fold change at DNMT3B peaks over input. Mean for each sample is indicated by the red dot and adjacent number, while the black line represents the median. Statistically significant differences in DNMT3B enrichment level (FC) between WT1 and other iPSCs were calculated using the non-parametric, paired Wilcoxon test with two-sided alternative with Benjamini-Hochberg False Discovery Ratio (BH-FDR) correction (****P*-adj < 0.0001). *(B)* A heatmap representation *(left)* and average plots *(right)* of ChIP-seq binding profiles of DNMT3B, H3K36me3 and H3K4me3 in WT1 iPSCs at the expressed genes in this iPSC line. The enrichment profiles are sorted with respect to the expression of the genes. Expression levels were determined by RNA-seq, calculated as log2 of FPKM counts in WT1. The highest levels of expression are at the bottom, the lowest ones are at the top. For plots on the right, DNMT3B and histone mark enrichments at expressed genes in WT1 are clustered into four quartiles based on gene expression levels (blue lines; Q1 – lowest expressed genes, Q4 – highest expressed genes). For DNMT3B and H3K36me3, the X-axis denotes genomic regions spanning +/-2kb of gene bodies. For H3K4me3 the X-axis denotes genomic regions spanning +/-2kb of the TSSs. The Y-axis for each ChIP-seq dataset indicates the number of counts per million mapped reads within regions of expressed genes. *(C)* Genome browser view of representative regions enriched for DNMT3B in WT1 iPSCs and hESC controls from public datasets (Verma et al. 2018 and Tan et al. 2019). The black bars at the top denote the DNMT3B peaks in control WT1 and hESCs. Below, the DNMT3B (black), H3K36me3 (dark green) and H3K4me3 (light green) coverage tracks are shown for WT1, ICF1 and their respective corrected iPSC lines. We detected a significant overlap between ours DNMT3B peaks and Verma et al. as well as Tan et al. DNMT3B peaks (shuffle test, *P*-adj < 0,0001). *(D)* Western analysis of DNMT3A expression was carried out on WT1 iPSCs and ICF1 patient iPSCs pR and pG, and their corrected clones. The actin protein was used as a protein loading control for the various samples. DNMT3A isoforms of different sizes are visible in the blot. As already shown for DNMT3B protein levels (Toubiana et al. 2019), Western blot analysis indicates that DNMT3A protein levels do not differ in patient iPSCs, and following DNMT3B editing in ICF1 iPSCs, from those of WT iPSCs. *(E)* Plots of average CG methylation levels of hypo-DMRs intersecting with increased H3K4me3 DERs (+/-2kb) in pR (n=1442, *left*) and in pG (n=1081, *right*) in comparison to the control WT1 and WT2 iPSCs. (F) ChIP-qPCR measuring CTCF binding levels at hypo-DMRs within subtelomere 2p, *CYP26C1* and *PCDHGA6* genes in WT, ICF1 and corrected iPSCs. For the validation of the CTCF enriched regions identified by ChIP-seq, we used both corrected clones for each patient iPSCs. Amplicon enrichment in immunoprecipitated and mock samples (M) is expressed as a fold change in site occupancy, determined as the ratio between percentage (%) of input of the region of interest and a negative control region amplified in the same samples (Neg_CTCF_Chr12). Bars and error bars represent means and SEM of at least three experimental repeats. Statistical analyses were performed using a one-tail two-sample Student's *t*-test compared to WT1 (***P*-value < 0.01, ****P*-value < 0.001).

SUPPLEMENTAL TABLES

Supplemental Table S1. Sequencing experimental design

Experiment	Sequencing reads	Read length	Number of sequenced reads per sample
WGBS	Paired-end (PE)	100bp	300-350 million
ChIP-seq	Single-end (SE)	100bp	
i) H3K4me3		100bp	25-30 million
ii) H3K36me3		100bp	40-50 million
iii) DNMT3B		100bp	50-60 million
iv) CTCF		65bp	20-25 million
RNA-seq	Paired-end (PE)	125bp (iPSC) 150bp (HPCs)	50-60 million

Supplemental Table S2. List of primer sequences and qPCR conditions used in gene expression and ChIP experiments

Gene/genomic region	Primer Sequence 5'→ 3'	Experiment; amplicon size (bp)	Thermocycling parameters
RNF212	F- TGCTTGATTTGTAAAGCTCCTTG R- TGGGAGGTTTCCCTGGAGTA	RT-qPCR; 141 bp	95°C,62°C,72°C 20s (35 cycles)
PTPN20	F- CCTGTTGGTCTGGGAAGCAT R- AGGCATGGCAAAAGTCTCCT	RT-qPCR; 148 bp	95°C-62°C-72°C 20s (35 cycles)
TSPYL5	F- CGTGTCTTTGAAGCTGCCTCC R- TACTGTGAAGGGTCCGGGTC	RT-qPCR; 152 bp	95°C-64°C-72°C 20s (35 cycles)
GAPDH	F- GAAGGTGAAGGTCGGAGTC R- GAAGATGGTGATGGGATTTC	Normalizing gene RT-qPCR; 234 bp	95°C-62°C/64°C-72°C 20s (35 cycles)
MYOD1	F- CCTCTTTCGGTCCCTCTTTC R- TTCCAAACCTCTCCAACACC	Control of genomic DNA in RT-qPCR; 223 bp	95°C-62°C-72°C 20s (39 cycles)
PCDHGA6	F- TAAGCCAGTAATGGCGCCTC R- CCAGTCCCAGATCCTTGACG	ChIP-qPCR; 172 bp	95°C-62°C-72°C 20s (39 cycles)
CYP26C1	F-TCAGTCTACGACGCCTCCAAAG R-AACGTCCAGAGGCAGTGAGAAG	ChIP-qPCR; 147bp	95°C-62°C-72°C 20s (39 cycles)
2p-subtelomere	F- GTGGAACCTCAATAATCCGAAAA R- GGACACCACTGTAAGCAAGATAGC	ChIP-qPCR; 150bp	95°C-62°C-72°C 20s (39 cycles)
Neg_CTCF_ Chr12	F - GGCCTCTCAAATCTCCTCCG R - GGAGTAAAGCTTCCGATAGAG (Chr12:7,715,183-7,715,283)	Negative region for CTCF binding; 101bp	95°C-62°C-72°C 20s (39 cycles)

Supplemental Table S3.xls

Lists of hypo-DMRs identified in ICF1 iPSCs and their categorization based on genomic annotation, rescue status in isogenic corrected clones and overlap with CGI and/or GH promoters/enhancers. Hypo-DMR score denotes the difference in methylation percentages between the ICF1 and WT1 iPSCs. Group assignment and rescue category (full, partial, no rescue) of each hypo-DMR is reported in Fig. 1A.

List of biological processes (PANTHER) enriched among the genes associated with hypo-DMRs identified in ICF1 iPSCs and annotated to promoter or gene body. The column IDs are indicated as provided by the PANTHER software. HOMER results showing the known Transcription Factor Motifs enriched at ICF1 hypo-DMRs and overlapping with decreased DNMT3B DERs (DNMT3B-Dec).

Supplemental Table S4.xls

Genes that are differentially expressed obtained from RNA-Seq data analysis of ICF1 iPSCs compared to WT1. Log_2FC and posterior probability for both ICF1 and the corresponding corrected clones compared with WT1 are provided for the listed genes.

Genes that are differentially expressed in both ICF1 hematopoietic progenitor cells (HPCs) compared to WT1 HPCs ($pp > 0.95$ and $|Log2\ FC| > 1.2$).

Supplemental Table S5.xls

HOMER results showing the known Transcription Factor Motifs enriched at the subset of hypo-DMRs belonging to Group1 and 2 or Group 3 and 4, based on the definition described in Fig. 1A. The column IDs are indicated as provided by the HOMER software.