UNIVERSITA' DEGLI STUDI DI NAPOLI "FEDERICO II"

DIPARTIMENTO DI MEDICINA MOLECOLARE E BIOTECNOLOGIE MEDICHE

DOTTORATO in

BIOLOGIA COMPUTAZIONALE E BIOINFORMATICA

Ciclo XXVIII

Integrative analyses of multi-omic data applied to the study of a rare human disease, the ICF syndrome

MIRIAM GAGLIARDI

Tutor:

Dr.ssa Maria R. Matarazzo

Coordinatore:

Prof. Sergio Cocozza

March 2015

Table of contents

Abstract

Application of the next-generation sequencing (NGS) technology has transformed epigenetic studies, generating large datasets that can be analyzed in different ways to answer a multitude of questions. Data integration is an essential step to understand intricate biological processes, such as the epigenetic control of gene regulation.

Recent "multi-omic" studies proposed the intriguing possibility that the intragenic DNA methylation would play a role in processing of transcripts during transcription modulating the elongation or splicing. Indeed a kinetic model, in which epigenetic modifications affect the rate of transcriptional elongation, and/or a recruitment model, in which adaptor proteins bind to epigenetic modifications recruiting splicing factors have been proposed.

Moreover, it was demonstrated that the intragenic methylation in highly transcribed genes is exclusively dependent on the DNMT3B function.

However, whether a DNMT3B-dependent epigenetic regulatory network modulates exon usage and transcription of alternative isoforms remains to be determined.

Through a large-scale integrative study we show that human DNMT3B germline mutations perturb its intragenic methyltransferase activity, affecting the relative abundance of transcript isoforms in the context of Immunodeficiency, Centromeric instability, Facial anomalies syndrome type-1 (ICF1). This correlates with changes of H3K4me3 and H3K27me3 at isoform-specific transcription start sites. Notably, the newly identified DNMT3B target genes might significantly contribute to ICF1 phenotype.

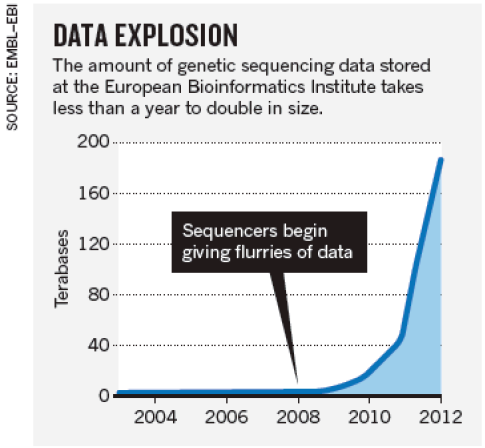
Chapter 1

Introduction

Within the last decade we have witnessed an important scientific metamorphosis. Indeed, the huge progresses in the field of molecular biology and in the same time technical engineering have led to the introduction of Next-Generation Sequencing (NGS) technologies (Margulies et al., 2005) changing the manner to study biological phenomena.

In particular, the traditional approach to study a single isolated gene to understand the logic of complex systems, such as development and response to environmental stimulus, has been rapidly replaced by the use of genome-scale data (Hawkins et al., 2010). This revolutionary process was already induced by "array" methods but it has been further speeded up in the era of NGS. Moreover, thanks to the new technology intrinsic advantages, making the NGS data not limited by a priori knowledge of the query genome or genomic features, the highly precise characterization and quantification of novel regions or isoforms has been possible (Hurd and Nelson, 2009).

A second advantage is the reduction of the background signal in NGS experiment when compared with the hybridization methods, associated to higher levels of reproducibility for technical replicates. Moreover, in experimental terms, to perform NGS analysis nanograms of sample are sufficient to obtain good data, allowing the reduction or the elimination of the reliance on PCR amplification (Hurd and Nelson, 2009).



All these advantages combined to the reduction of sequence costs are contributing to increase the number of single laboratories and big consortia that are using routinely the high-throughput sequencing to address their research. However, this "democratization" of genome-wide techniques led to the generation of a huge amount of large data sets (Berger et al., 2013). In this view, it is impressive that in the past decade the amount of published sequences outstrips the prediction based on the Moore's law (Kahn, 2011) with a new sequence data grown exponentially each year (Figure1, (Marx, 2013)).

The big explosion of sequence data has introduced new challenges in the biology field, such as the analysis and the integration of all the big-data produced.

At present, HiSeq by Illumina and Ion Torrent by Life Technologies represent the two main platforms for deep sequencing. Their output in terms of read number per run is 600Gb (read length 2x100bp) and 1.5-2Gb (read length 200-400bp) respectively. Moreover, other sequencing technologies have been released, the Heliscope (15Gb and reads length ~30bp) from Helicos BioScience and the PacBio SMRT (5Gb and 50% reads > 10kb) from Pacific Biosciences. Nevertheless, there is an emerging sequencing platform that has the potential to make this field a step forward, the nanopore technology. It implies long read lengths, of up to 10kb, minimal requirements of reagents and sample preparation and high sequencing rate at low cost. Despite these advantages, several technical problems relative to the nanopore sequencing remain to be solved.

1.1 Quality control and mapping of sequencing data

Raw sequencing data derive from the acquisition and processing of several images; this process can influence the quality of raw data. In order to evaluate the sequencing quality, several tools have been developed, allowing the evaluation of read (i.e., a consecutive sequence of nucleotides) quality, read duplication rate, GC content, nucleotide composition bias, etc.

The base-calling quality from a Sanger sequencing was measured using the Phred quality score (Q), in which Q is depending by the probability P of erroneous call, according to the equation $Q = -10 \times \log10(P)$. This suggests that if the quality score is 30 the probability of incorrect call is 1/1000.

In most of the sequencing output files the Q value is not reported in number format but in ASCII code (e.g., 33–126 or "!" to "V").

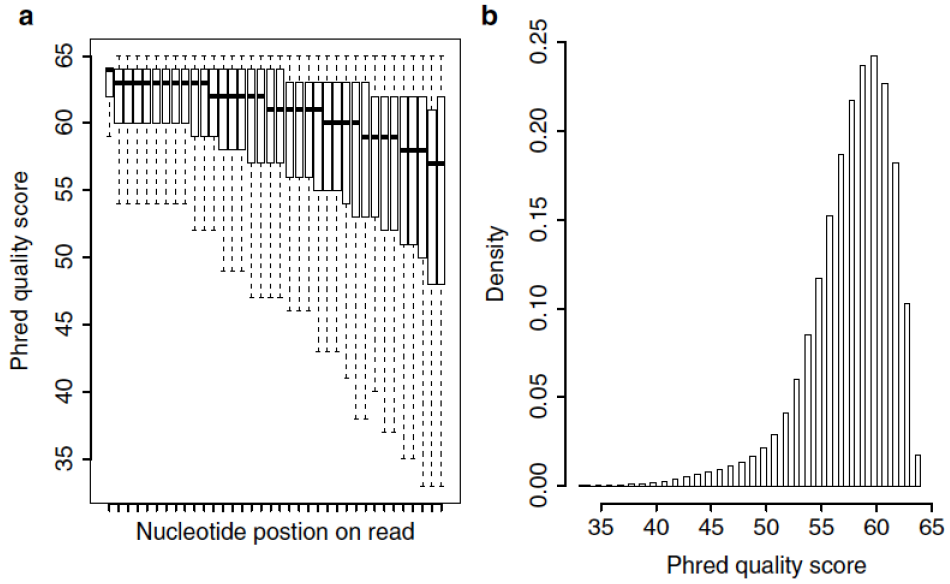


Fig. 2 (a) The boxplot displays "per nucleotide quality score." The Phred quality score distribution of all the reads is (Y-axis) is shown per each nucleotide position (X-axis). (b) "per sequence quality score" distribution, in this case the Phred score is calculated for each reads.

The Figure 2 reports two types of graph showing the Phred quality. In the graph a the Phred is summarized per read position while in b the distribution of total tags based on the Phred average per read is reported. In general, it is

recommended to use reads with scores over 20. Phred score higher than 30 indicates very good quality of the nucleotide sequencing, between 20 and 30 the quality is still acceptable while <20 indicates poor quality.

GC content (or guanine-cytosine content) is a further parameter that can be assayed after confirming the quality of raw data. This is a way to measure the DNA composition. Analysis tools often report a diagram, similar to what showed in Figure 3, in which it is reported for each base of the reads the nucleotide frequency (a) or the distribution of the sequences per GC content (b) . In the graph a, the lines should be roughly flat around 0.25, even though in the first 12 nucleotides usually a large deviation from 0.25 is observed because these positions represent the random hexamer priming during PCR amplification. A serious bias in these analyses is caused by the overrepresentation of some sequences, and it could influence the coverage uniformity. The low-quality reads defined by the read parameter described above are unlikely to be informative and therefore should be removed.

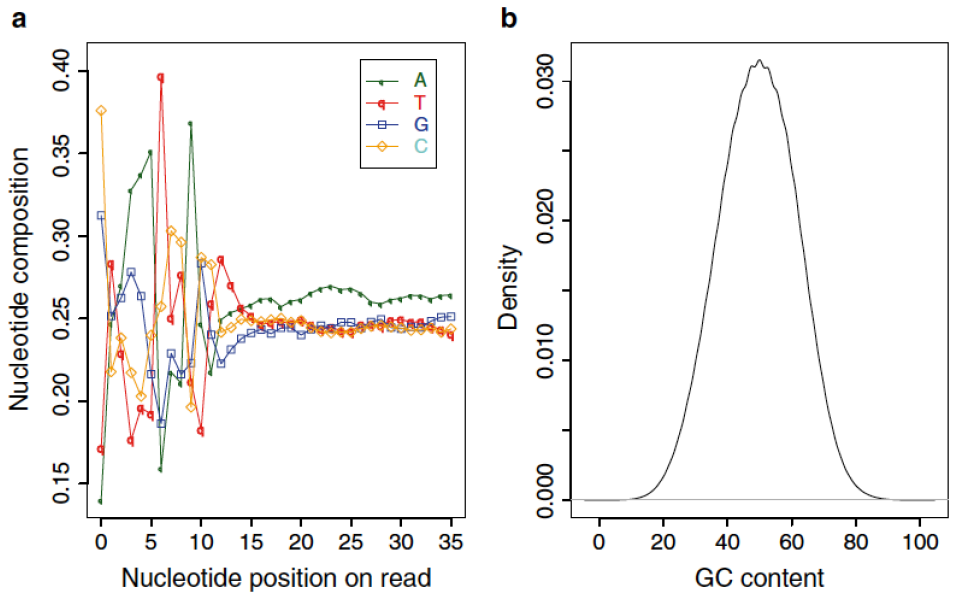


Fig. 3(a) Diagram displays nucleotide composition (expressed as nucleotide frequency) of all overlaid reads (b) In the graph is reported the "per sequence GC content" distribution.

In order to extrapolate a biological meaning from millions of good quality reads produced by sequencing experiments, the short-sequences have to be aligned against the reference genome, to identify their genomic origin. The classical mapping approach, in which the single read is compared to the whole-genome sequence for each base position, would have high computational costs. Currently the running time of the mapping is reduced through an approach involving the pre-processing of the reference genome into a flexible and compact indexed format. The most common short-reads alignment tools, such as Burrows-Wheeler Aligner (BWA), Bowtie and SOAP software references, use as core-technique an indexed and compressed reference genome in FM-index format, which is a compressed data structure for sequence data obtained by the combination of two diverse algorithms: Burrows-Wheeler transformation (BWT) and the suffix array (Figure 4).

The BWT (Burrows and Wheeler, 1994; Margulies et al., 2005) is a string transformation that converts highly redundant sequences producing an output string which can be easily compressed (Nelson, 1996). To obtain the transformation, the input sequence should be rotated in a matrix in which each position shifts in the start position exactly once, then the matrix is sorted in the lexicographical way, redundant rows are grouped together and the last column is extracted as output sequence (Manzini, 2001).

On the other hand, the suffix array method indexes all possible suffixes of a string (Manber and Myers, 1990). They are reported and ranked in the alphabetical order. The suffix array can be used as an index to quickly locate every occurrence of a substring the clustered pattern within the sequence, allowing the efficiently finding with two binary searches.

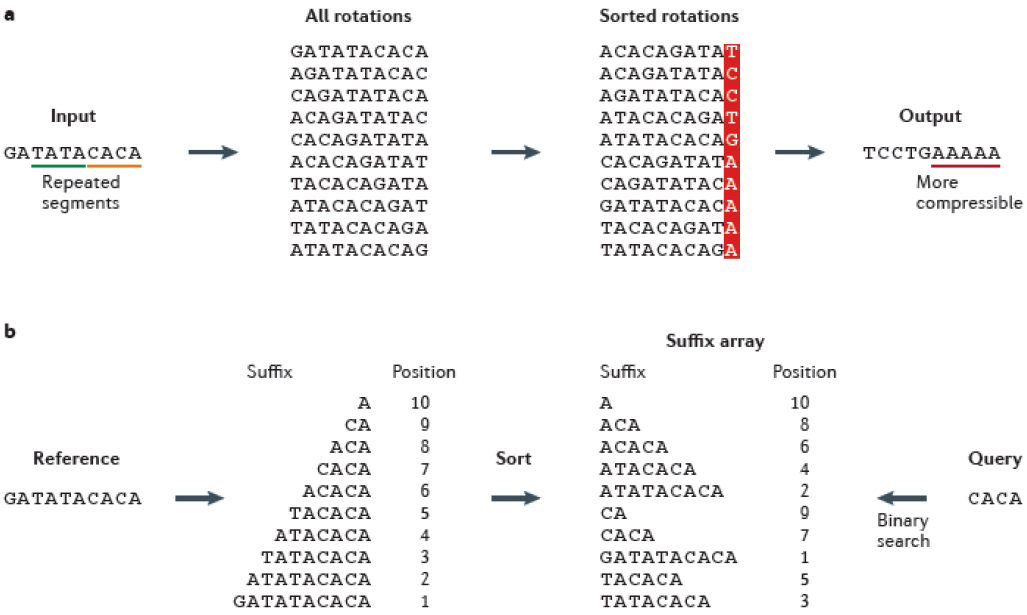


Fig. 4 The figure summarizes (a) Burrows-Wheeler transformation and (b) suffix array steps.

From the combination of BWT and suffix array it has been extrapolate an hybrid method ideal to index the short-reads, the FM-index (Ferragina and Manzini, 2005). The FM-index is a suffix-array-like format build starting from a BWT transformed reference sequence. This transformation and compression allows significant reduction in terms of storage space and of mapping time costs (Berger et al., 2013).

To speed up the mapping process, the parallel dynamic programming in Bowtie2 software has been implemented few years ago. This implementation allows improving the powerful of the long-reads mapping reducing the used time (Langmead and Salzberg, 2012).

The mapped reads contain a magnitude of potential information that needs to be pulled out in order to answer specific biological questions. The analysis pipeline that follows the mapping is strictly linked both to the feature of the sequenced sample and the used high-throughput technology. In fact, the NGS approach has been combined with different experimental protocols to have a genome-wide view in different studies of cellular processes, including expression analysis, epigenome and interactome.

However, even though improvements in the mapping algorithm development have been carried out, it is still not possible to map all the reads to the reference genome. This is caused perhaps by sequencing errors, structural rearrangements or insertions in the query genome, or deletions in the reference. Indeed, the unmapped reads can be highly informative in studies about structural variants and non-reference insertions.

Other unsolved challenges in alignment field is the ambiguous mappability of some reads derived from regions containing low-degeneracy repeats or low-complexity sequences. The ‘mappability’ (also known as uniqueness) of a sequence within a genome has a major influence on the average mapped depth and is an important source of false-negative single-nucleotide variant calls. Mappability improves with increased read length or using paired-end libraries, which increases the chance of one read of the pair mapping to a unique region outside the repeats.

1.2 Transcriptomics

The first application of NGS to experimental protocol was in transcription studies. Indeed, the introduction of RNA sequencing (RNA-seq) allowed the estimation of the abundance level or the relative changes of each possible transcript. The RNA-seq is the set of experimental procedures that generates sequences starting from cDNA derived from RNA molecules. The RNA-seq is a very powerful technology because of the low background noise, the single base-pair resolution and the dynamic detection range. Moreover, the transcriptome analyses performed using this technology are not limited to the interrogation of a priori knowledge about specific sequences; this allows their application to identify completely novel and not annotated transcripts.

This is the reason why the RNA-seq is also used to catalogue new different categories of non-coding RNAs (ncRNAs); to date, it allowed the annotation of long intergenic non-coding RNAs (lincRNAs, length>=200bp), microRNA (miRNA, length ~22bp), short interference RNAs (siRNAs) and other classes of small ncRNAs, such as snRNAs and piRNAs. Each of these ncRNA classes has its specific role in the regulation of molecular processes, including RNA stability and chromatin structure conformation. Moreover, strand-specific RNA-seq (which retain the orientation of original RNA molecules) studies have shown a much more complicated and not completely explained level of regulation mediated by antisense transcripts. Furthermore, using RNA-seq it is possible to get information about the expression of pseudogenes, retrotransposons and other repeats.

In this perspective, the transcription process represents a tightly regulated system that allows the fine-tuning expression in time-, cell-type- and stymulus-dependent manner. The RNA-seq studies performed to evaluate this phenomenon provided us with information about the transcription of diverse isoforms from the same genomic locus. These events can derive by the selection of alternative transcriptional start sites (TSS), events of alternative splicing or premature transcription end sites (TES).

Although the RNA-seq is becoming a routinely analysis, it is an ever-evolving process that require powerful and updated algorithms able to answer all these biological questions. A variety of tools is available to analyze the RNA-seq data sets.

The commonly used short-read alignment programs discussed above are not sufficient to an exhaustive RNA-seq mapping. In fact, they are not suitable to map reads located in poly(A) tails or exon-intron splice junctions. The reads originated from exon-intron boundaries are very helpful to identify splicing variant patterns. Accordingly, a number of tools as for instance TopHat and MapSplice are able to identify the junction from where the reads were originated.

To measure the gene expression levels, the mapped reads in the gene locus are counted. The most common pipeline for the analysis of RNA-seq considers the expression of the full gene and counts the reads mapped on all the associated gene exons. A simple procedure for read counting is to use coverage commands available in several packages for NGS analysis, such as the coverageBed in the Bedtools utility; these algorithms starting from the position of mapped reads in bed (tab-delimited file in which are reported in a prefixed order the chromosome, the start and end position of each mapped read) or bam files (binary version of a tab-delimited text file, SAM file, that contains the complete sequence alignment data) are able to count reads mapped in specific genic regions (e. g. genes or exons) available in bed format file. Specialized tools for read counting are available, as for instance HTseq. HTseq is a very famous and accurate counter; it offers to the user the opportunity to control several parameter settings, in particular it is able to recognize reads covering more than one gene. This tool requires a reference list of genic features available in a General Feature Format (GFF) file (tab-delimited file containing all the information about gene features). In the last years several other free counter-tools were published, most of them are available as packages for the statistical integrated suite R, such as

summarizeOverlap and featureCount; each tool presents specificity and settable parameters, thereby influencing the final read estimation. An other source of counting variability is the annotation used; indeed, the transcript annotation on which is based the gene model set is important for the expression analysis. Moreover, feature GFF from diverse databases (e. g. RefSeq or Ensembl) can include different structures for the same gene, that will result in discordant read estimations. Zhao and Zhang have recently analyzed this aspect (Zhao and Zhang, 2015) evaluating the weight of gene model on the estimation of gene expression. They found that the influence of the annotation is dependent on the region feature; in particular, the mapping of junction reads is more affected than the not-junction read mapping. Indeed, only the 53% of junction reads were mapped to exactly the same genomic location, while the non-junction reads aligned in the same position were the 95%. From this perspective, it appears even more evident the importance to validate the results obtained from in silico analyses through independent "wet" experiments.

To ensure accurate inference to differential expression analysis, counted reds need to be normalized before comparing them between different conditions. Multiple normalization methods have been developed to this purpose. For instance, the quantile normalization can improve the mRNA-seq data quality including those from low amounts of RNA.

Several statistical tests can be applied on normalized count matrix to evaluate the significant changes in gene expression across different conditions. In 2008, Mortazavi (Mortazavi et al., 2008) proposed the RPKM (Reads Per Kilobase per Million of mapped reads) as a method of quantifying gene/transcript expression from RNA sequencing data by normalizing on the total number of mapped reads and length of the analized gene (or feature). In case of pair-end reads it is instead used the FPKM (Fragments Per Kilobase per Million mapped fragments, where a fragment), which describes the relative abundance of transcripts in terms of the expected biological objects (fragments) observed

from the RNA-Seq experiment. Recently, it has been shown that RPKM/FPKM do not represent the best approach for data normalization because they handle all the reads almost uniformly, without considering the gene composition (Dillies et al., 2013). In case of differential expression analysis, other normalization approaches included in the differential expression model are more indicated. The simplest analysis includes the application of count-based probability (e.g., Poisson) distributions followed by Fisher's exact test. However, this strategy has as bias that it does not consider the biological variability and this is the most important task in the differential expression analysis. Indeed, to statistically highlight the differences across samples it is mandatory to have a large number of replicates to build an efficient and powerful statistical model to describe the data. Working with biological and technical replicates allows the assessment of biological variability and the statistical measurement of differentially expressed genes can be performed simply applying an extended Poisson distribution. However, it has been noted that the Poisson distribution underestimates the variation seen in the data (Nagalakshmi et al., 2008), a problem known as over-dispersion. Unfortunately, "wet" biologists know how it is often difficult to have a large number of replicates because of both the still high sequencing costs and, even more important, the scarcity of samples available for the sequencing. Therefore, several tools for the detection of significantly differentially expressed genes have overcome this problem applying the Negative Binomial distribution because of its ability to trade with the over-dispersion problem. The improvements in the differential gene expression studies introduced by Negative Binomial distribution allowed this methodology to become the dominant one in the modeling of RNA-Seq counts (Zhang et al., 2014), Examples of tools based on the Negative Binomial distribution are Cuffdiff (tool included in Cufflinks package), EdgeR and DESeq (R packages). They are the most used algorithms in the biomedical and clinical published research. In the expression analysis a further level of

complexity is added by the expression of different isoforms from the same genic locus.

The isoform composition influences and modulates the efficiency of a large number of pivotal molecular processes; furthermore, its alteration is strictly associated to pathological conditions.

The expression of different isoform is caused by the alternative splicing of not constitutive exons or by the selection of different TSS. The alternative splicing affects around 95% of multi-exonic genes; this phenomenon allows the expression of isoforms that differ for the inclusion/exclusion of one or more specific exons or just an exon portion, or the retention of the intron. The splicing is a very complex multifactorial process and it is regulated by RNA-RNA molecule, RNA-protein and protein-protein interactions. Tissue-specific expression of RNAs and/or proteins that take part to the splicing machinery, also known as spliceosome, led to the contest-dependent expression of specific isoforms.

RNA-seq technology allows the estimation of the isoform abundance in a genome-wide perspective. In this view, many methods and tools have been developed to quantify the isoform expression and most of them are based on Bayesian inference methodology, as for instance BitSeq and MISO (Mixture of Isoforms). BitSeq estimates the expression of individual transcripts from different RNA-seq experiments while MISO is able to discriminate the exon inclusion level, returning as output both the information about isoform abundance and exon usage. Among other algorithms, MATS (Multivariate Analysis of Transcript Splicing) is focused on the statistical detection of differential alternative splicing events from RNA-seq data. However, the field of RNA-seq usage in alternative splicing studies is still in the early stage of its development and certainly it will benefit of new methods and strategies.

1.3 Epigenomics

NGS technology found a very broad application in the epigenetic field. Epigenetic studies focus their attention on the heritable features that modulate the genome–environment interaction without DNA sequence alterations. The epigenetic traits often reveal extensive flexibility, enabling the cell to adapt to environmental changes and leading to acquisition of appropriate gene expression profile.

The accurate measurement of epigenetic marks is of crucial importance, not only to identify pathologic conditions associated to epigenetic aberrations, but also to deepen our insights into the mechanisms by which epigenetic drivers control biological processes.

A widely studied subset of epigenetic marks is composed by the histone modifications. Histone modifications mark the genome and play a key role in regulation of its accessibility and spatial organization. In fact, the DNA molecule is too long to be stored in a nucleus without any superior organization, and therefore a complex structure called chromatin evolved, in which the DNA is wrapped around the nucleosomes. Nucleosomes are composed by protein complexes, which are formed by one H3-H4 tetramer and two H2A-H2B histone dimers. The histone N-terminal tails are exposed outside the scaffold core of the nucleosome and can be affected by post translational modifications (e.g., methylation, acetylation, ubiquitination) leading to different expositions of the DNA sequences depending on the nature of the chemical change and the amino acid residue that is modified (Park, 2009).

In the late 1920s, two different chromatin condensed states were microscopically identified as heterochromatin and euchromatin. It was early supposed that structural differences might correspond to functional distinctions. Later, the euchromatic regions have been described as permissive open regions associated to active transcription, while the heterochromatin has been described as highly condensed and transcriptionally repressed. The

heterochromatic status can be constitutive (as for instance the condensation of centromeres and telomeres) or can be temporary undergoing a transition from active to inactive state and vice versa.

Up to 15 years ago, studies linking gene expression and chromatin organization were limited to a few genes. However, the novel genome-wide techniques have now radically increased the number of information that it is possible to obtain from a single experiment, providing a much more detailed view of functional regulation of genome accessibility and its role in the modulation of gene expression.

The chromatin immunoprecipitation (ChIP) is the main used tool to go inside the study of the role of histone modifications. The ChIP is a technique for assaying protein–DNA interactions in vivo, in which the binding between protein and DNA is fixed using formaldehyde or UV and the chromatin is sheared by sonication into small fragments (around 200–500 bp). All the regions interacting with the target protein (or enriched for a specific histone modification) are precipitated using a specific antibody able to recognize the target protein or a particular histone modification. The DNA fragments enriched in the immunoprecipitation step can be analyzed in gene specific manner using the polymerase chain reaction (PCR) both quantitative and semi-quantitative (ChIP-(q)PCR) or in a whole genome view using the sequencing technology (ChIP-seq) (Comes et al., 2013).

Data derived from ChIP experiments are crucially influenced by several aspects, first of all the quality of antibody-antigen recognition. A specific and sensitive antibody will result in a more clear enrichment signal.

Other problems in ChIP-seq data can be the artefacts, which may derive from several potential sources. The open chromatin regions, for example, are more sensible to the shearing creating an unbalanced distribution of sequence tags across the genome (Park, 2009). Also, repetitive sequences can show enrichments because of imprecisions in the number of copies of the repeats in the assembled genome. Therefore, the ChIP-seq profile should be normalized

against a control sample to determine the significance of each single peak. Usually, in ChIP experiments two types of control are used: positive and negative control. The positive control is always the input sample, which represents the starting chromatin. By contrast, the negative control is achieved after immunoprecipitating a parallel sample with a not specific antibody, generally the rabbit IgG, or a mock sample, immunoprecipitated in absence of antibody.

In ChIP-qPCR, the input sample is used to normalize for the amount of starting DNA the amplification obtained from the IP sample (defined as input percentage; %input), while IgG and mock can be used to estimate and correct for the background noise.

In ChIP-seq, there is no consensus on which of these controls is the most appropriate. A huge fraction of all published ChIP-seq studies use the normalization against the Input DNA; in ChIP-seq context, this normalization corrects for bias related to the variable solubility of different regions or the differences in the shearing sensibility or the amplification artefacts. The use of mock or IgG IP as ChIP-seq controls is less common, because a very little DNA amount can be pulled down in the absence of antibody or with a not specific one. Therefore the results of multiple mock or IgG IPs may not be consistent.

To obtain a good control for ChIP-seq a large amount of sequencing reads is required because many of the sequenced tags for input DNA are spread evenly across the genome. To obtain accurate peak detection throughout the genome, sufficient numbers of tags are needed at each point; this will improve the power of peak detection reducing errors due to sampling bias. Therefore, the total number of tags to be sequenced is potentially very large. Alternatively, if the study involves only the detection of differential binding patterns between two or more conditions, starting from the same genomic background, it is possible to avoid the control sample sequencing (Park, 2009).

The sequencing depth is important not only for the controls but also for the

analyzed samples. It is hard to define a priori the optimal amount of reads needed for a specific experiment. Intuitively, if the study is focused on a protein that binds the genome in a large number of sites or on an histone modification that covers spread genomic domain much higher number of tags will be needed to cover each bound region at the same tag density; conversely if the protein binds few and very defined target region in the genome the amount of needed reads for the binding detection will be smaller.

To define the sufficient depth of ChIP-seq samples one considerable criteria would be the finding of a "saturation point" binding sites in which the number of identified peaks does not change adding more reads.

In fact, as demonstrated in a recent simulation study (Kharchenko et al., 2008), when the minimum sequence depth is achieved, the rate at which new sites were being discovered using more reads slows down, (described as plateau in the graph read number versus detected peaks) (Park, 2009).

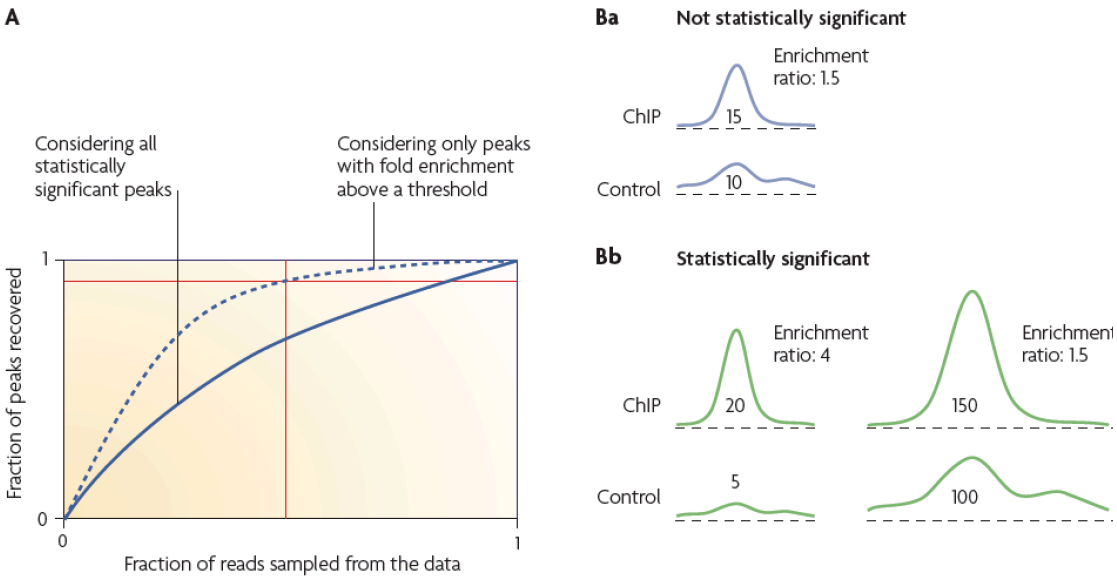


Fig. 5 (A) Diagram displays the rate of peak detection based increasing the sequencing depth. (B) Sequencing depth affects the peaks detection; the peak in panel Ba is not statistically significant even has the same Enrichment ratio of significant peak in panel Bb.

However, these considerations are true only if a threshold on the minimum fold enrichment between the peaks in the ChIP experiment and the peaks in the control is imposed (Figure5).

After read alignment, it is important to detect genomic loci that are significantly highly enriched in the ChIP sample than in the control. To find these regions, the common pipelines include an analysis step in which the coverage level is typically transformed to count data for predefined DNA regions, for example, genes or promoters or equally sized bins or extracted from the data using peak detection algorithms.

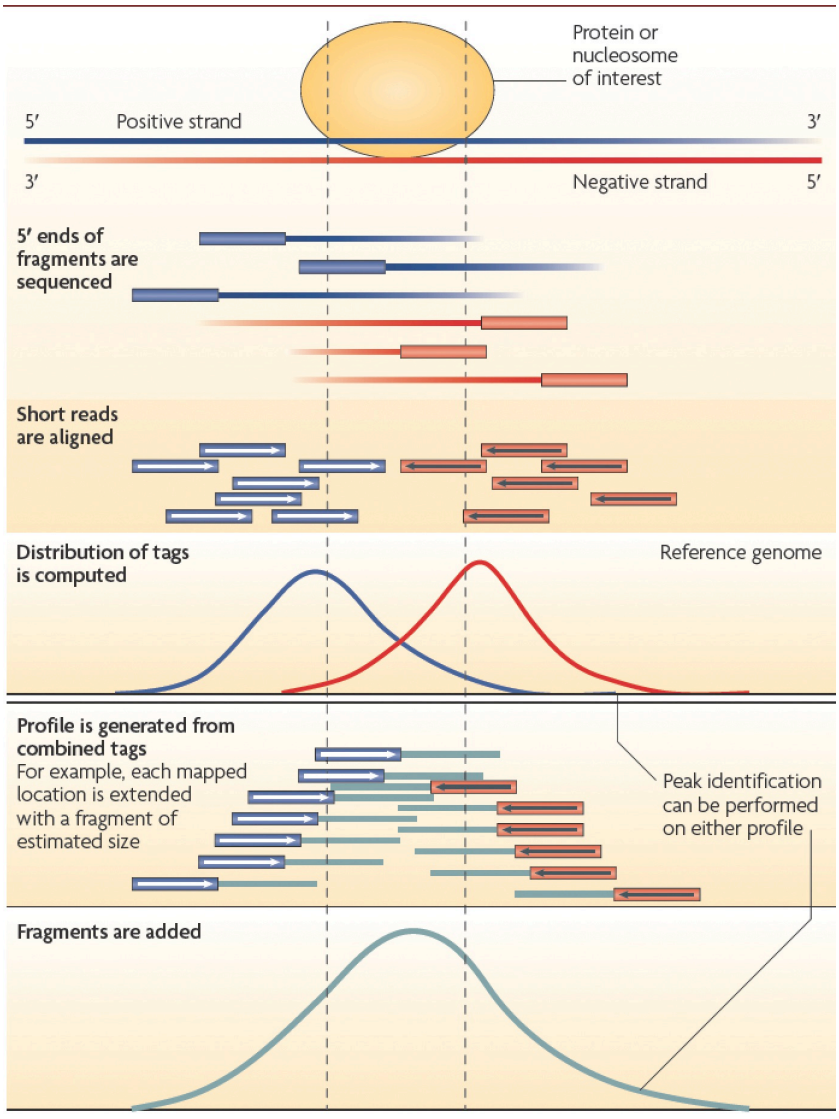


Fig. 6 Schematic representation of strand-specific profiles at enriched sites of DNA fragments from 5'-sequenced chromatin immunoprecipitated sample.

Algorithms able to scan the ChIP-seq enrichment along the genome are commonly known as "peak callers". The first type of tools for the peak detection was based on the scoring of predefined regions by the number of tags in a window and then assessed by a set of criteria based on factors such as enrichment over the control and minimum tag density, whereas the more recent tools are based on the read directionality. As shown in figure, mapped reads form two distributions, one on the positive and one on the negative strand, because of the 5'end sequencing of the fragment (Figure6) (Park, 2009). The two distributions are combined in a single distribution using an inclusion profile algorithm. This step combines the distributions by shifting each of them towards the center or by extending each mapped position into an appropriately oriented fragment length. This approach is more accurate in the profile description than the width of the binding, but it requires the estimation and the uniformity of the fragment size. Also in this type of analysis it is important to have a statistical estimation of the peak quality. For this reason, the putative peaks are compared with the background distribution, either by simulation (nonparametric) or by a statistical modeling approach.

As already mentioned above, the binding profile of histone modification and/or transcription factors to the DNA might be very different depending on the role of the studied protein; indeed, transcription factors and histone modification that marks regulatory elements, such as promoters and/or enhancer, give rise to very sharp peaks, whereas broad signals are often associated with histone modifications that mark domains. Although specific algorithms for the detection of both these types of peaks have been published and widely used, tools able to discover enrichment of proteins or histone modifications that have mixed binding profiles are still not optimized.

In order to assay significant differences in ChIP-seq experiment carried out in various conditions, basic statistical methods can be applied, as for instance Wilcoxon rank–sum and t-test. However, the distribution of reads in ChIP-seq

experiment can also be modelled as a negative binomial distribution. This allows to take advantage of RNA-seq data processing tools, such as DESeq or EdgeR, especially in experiments with a low number of replicates. These types of experiments can take advantage also from the statistical power of Hidden Markov models (Rabiner, 1989), as for instance the ChIPDiff tool in which the correlation between consecutive bins is adapted an Hidden Markov model (HMM). Here, the transmission probabilities were automatically instructed in an unsupervised way, followed by the inference of the states of histone modification changes using the trained HMM parameters (Park, 2009).

The downstream analyses from peak detection include the discovery of binding motif and/or the positional association with genomic features.

The search for specific binding sequences is a characteristic step in the description of protein-DNA interactions; in particular it is recommended in studies of proteins with very sharp peaks. The sequence can be obtained using top-scoring peaks as input data into motif-finding tools, such as meme, weeder and webmOTIF. These tools return potential motifs associated to their statistical significance.

After the identification of the peaks, their localization in the genomic context can be analyzed. This step is also known as "peak annotation", which means to annotate the proximity of each peak to specific genomic features, as transcriptional start site, exon–intron boundaries and the 3'ends of genes . However, it is also possible to customize these analyses looking for other features of interest, such as repetitive elements or enhancer regions. A common analyses, that is performed in ChIP-seq experiments for the histone H3 trimethylated at lysine 4 (H3K4me3), is the identification of TSS marked by this histone modification. In fact, it is known that H3K4me3 is enriched on the TSS of active genes.

DNA methylation is an additional important epigenetic player (Bird, 2002); it has an impact on gene regulation, chromatin structure, development and disease. Generally, most mammalian genomes are largely methylated except at active or "poised" promoters, enhancers and CpG islands, where DNA methylation has a repressive effect (Jones, 2012). Nevertheless, gene bodies DNA methylation has been associated with high expression levels. DNA methylation is established and maintained by the combined function of three active DNA methyltransferases DNMT3A, DNMT3B, and DNMT1 (Goll and Bestor, 2005). In mammalian cells, these enzymes catalyze the covalent addition of a methyl group preferentially to cytosine of CpG dinucleotides, and most CpG sites in the genome are methylated. DNA methylome studies took advantages from the introduction of NGS technologies. In particular, the single-base resolution in a genome-wide view allowed to discover that approximately 25% of methylated cytosine in stem cells was in a non-CG context in contrast to what observed in differentiated cells (Bock, 2012).

These studies have been carried out using the bisulfite sequencing technology, in which genomic DNA is treated with the sodium bisulfite allowing the conversion of unmethylated cytosines (Cs) in thymines(Ts) whereas methylated Cs are largely protected from bisulfite-induced conversion. The bisulfite treated DNA molecules are sequenced and then aligned to the reference genome with short-read aligners that have to take into account the depletion of Cs after the bisulfite treatment. Two alternative approaches have been developed. Aligners as BSMAP, GSNAP and RRBSMAP replace Cs in the genomic DNA sequence by the letter Y, which matches both Cs and Ts in the read sequence, or they modify the alignment-scoring matrix in a way that mismatches between Cs and Ts in the read sequence are not penalized. Conversely, tools as Bismark and BS-Seeker are defined three-letter aligners because they use to convert all Cs into Ts in the reads and for both strands of the genomic DNA sequence (Bock, 2012). This way, they can carry out the

alignment exclusively on a three-letter alphabet (namely, A, G and T) using a standard aligner, such as Bowtie.

The three-letter approach is more accurate in the definition of the methylation level even though the decrease of the sequence complexity increases the possibility of ambiguous alignments.

After the alignment, the methylation level of each covered genomic Cs can be estimate. The most common and simple way to perform this analysis is to calculate the methylation percentage of each C of the reference genome by dividing the sequenced Cs number on the total number of aligned nucleotides (Cs and Ts). However, to improve the accuracy additional steps can be added to the analysis, such as local realignment, analysis of sequence quality scores and statistical modeling of allele distributions. The use of variant caller, such as Bis-SNP, can reduce a common error source in the analysis of DNA methylation data, helping to distinguish bisulfite-induced changes from genetic variants. This is possible considering that the nucleotide variations induced by bisulfite treatment exhibit a G on the opposing strand, whereas genetic C-to-T variants exhibit an A.

The sequencing of genomic DNA can be performed on genomic scale or on enriched regions and these two different strategies are defined as "whole-genome bisulfite sequencing" (WGBS-seq) and "reduced-representation bisulfite sequencing" (RRBS-seq), respectively.

The key advantages of the WGBS-seq are the higher coverage, the quantitative accuracy and the reproducibility. However, the WGBS-seq is still too much expensive. In order to reduce the costs of analysing the methylation status of CpG-rich regions, the RRBS-seq is extensively used. This strategy combines the bisulfite sequencing with enrichment strategies using restriction enzymes. However, methylation studies can be performed also with enrichment-based strategies. In particular, methylated DNA can be enriched using methylation-specific antibodies (MeDIP–seq) or methyl-CpG-binding

domain (MBD) proteins (MBD-seq) and the pulled-down mehylated DNA can be sequenced.

The enrichment-based DNA methylation derived tags are aligned to reference genome using a standard aligner, such as Bowtie. Further analysis is the estimation of enrichment scores by counting the number of unique reads that overlap with each CpG or with the genomic regions of interest. A bias can derive from this kind of analysis if the enrichment score is not normalized on the CpG density of the region. To correct this bias, several algorithms have been developed. For example, the BATMAN algorithm uses a Bayesian method, which provides accurate results but it is too slow when applied to large data sets. The MEDME method is based on a logistic regression model for data normalization, but it is rarely used owing to the need for calibration using a fully methylated reference sample. The MEDIPS software combines the previous reported tools into a data normalization and analysis pipeline that is sufficiently fast and easy to use to be practical for routine processing of MeDIP–seq and related data types (Bock, 2012).

After the estimation of the absolute methylation level, the typical next step is the identification of differentially methylated regions (DMRs) between samples (for example, cases versus controls).

The vast majority of interesting DMRs fall within a size range of a few hundred to a few thousand bases, although a single methylated CpG may occasionally modulate the expression of a gene. In the most basic form of DMR detection, T-tests or Wilcoxon rank sum tests compare the DNA methylation levels of each C between two sample groups. Several more advanced methods have been described that aim to improve DMR detection using mixture models, feature selection, aggregation of genomic regions by type.

1.4 Data integration

Nowadays, all these different "omic" approaches are widely used to measure gene expression or to obtain genome-wide maps of transcription factor and epigenetic signature profiles. Although several computational tools have been developed for their analysis demonstrating the interplay between transcriptomic and epigenomic profiles, efficient pipelines for complete multi-omic analyses are still limited. However, performing integrative analyses is helpful to address comprehensive studies to investigate some long-standing questions related to fundamental mechanisms of genome function and disease (Hawkins et al., 2010).

Data integration can be achieved in several ways, starting from the simple multidimensional view on Genome browser for the visualization of sequencing data. Looking at the expression and histone modification and/or DNA methylation profiles at few loci of interest can help to formulate new functional hypothesis. Anyway, to take advantage from the whole-genome approach, visualization tools are not sufficient. The first and easily used approach is the analysis of the overlap between lists of differentially expressed genes (resulted by RNA-seq experiments) and of genes enriched for a protein binding or for specific epigenetic changes. The enrichment for genes associated to interesting pathways or gene ontology can be calculated from the resulting subsets.

In one of the first ChIP-seq and RNA-seq integration studies, the expression was considerated as a response variable to different interaction of transcription factor (TF) and a log-linear regression model was proposed (Ouyang 2009). The use of this regression strategy on a small number of histone modifications has shown its high precision also with these epigenetic marks (Karlic 2010).

Cheng and collaborators used a different approach in which they measured epigenetic signals directly on genic features, such as TSS and TTS. Without *a priori* assumption on the relation between expression and specific profiles of

TF and histone modifications, they were able to capture much more complex functional interactions (Cheg 2011, 2012).

The used strategies can be also applied in the evaluation how the epigenetic alteration impact on expression changes comparing two conditions. For example, Althammer and colleagues in 2012 analyzed the relationship between expression profiles in two conditions with the status of 13 features. In fact, they classified the genes as upregulated, downregulated and unchanged expression and associated to each of them a resuming vector with a value for each feature (included TF, histone modification and DNA methylation). Recently, a Bayesan mixture model has been proposed to estimate the weight of epigenetic variations on differential gene expression (Klein et al., 2014).
Although realistic quantitative models of genome-wide regulatory networks are still missing, it is possible to discover the main interactions and the most relevant players combining in a unique pipeline supervised analysis to addressed biological questions. Therefore, from a biological point of view the integration step open up several possibilities to growth up new hypothesis (Angelini and Costa, 2014).

Taking advantages from the integrative analyses, it has been reported that the DNA methylation is not only restricted to promoter regions but it is present at the intragenic regions, supporting the hypothesis of a more complex function for this epigenetic modification. Indeed, in transcribed regions DNA methylation might potentially silence alternative promoters, enhancers, transcription factor binding sites, retrotransposon elements, and other functional elements to ensure the efficiency of transcription (Maunakea et al., 2010; Wolff et al., 2010; Kulis et al., 2012). Even more interestingly, this multi-omic approach led to discovery that DNA methylation plays a role in the processing of mRNAs during transcription modulating the elongation or splicing (Chodavarapu et al., 2010; Anastasiadou et al., 2011; Gelfman et al.,

2013). Accordingly, DNA methylation has been recently found positively and/or negatively correlated with inclusion level of alternative exons (Maunakea et al., 2010; Shukla et al., 2011; Yearim et al., 2015).

Moreover, integration studies can be very helpful for the molecular defects underlying the chromatin diseases, which represent group of human genetic disorder in which proteins modifying the chromatin and/or histone marks are mutated.

In this thesis, an integrative analysis to heterogeneous genome-wide datasets of samples derived from patients of the rare chromatin disease, the Immunodeficiency Centromeric instability and Facial anomalies (ICF) syndrome, has been carried out. The multi-omic approach will support the discovery and the comprehension of still unknown defective molecular processes in this human disease (Angelini and Costa, 2014).

Chapter 2

MATERIALS AND METHODS

2.1 Cell culture

Epstein Barr Virus immortalized lymphoblast cell lines (B-LCL) were derived from peripheral blood mononuclear cells (PBMCs) of two unrelated ICF patients with missense mutations in DNMT3B, ICF1p1 (female, heterozygous A603T and intron 22 G to A mutation resulting in insertion of three amino acids (STP) in DNMT3B, Coriell Cell Repository) and ICF1p2 (male, heterozygous V699G and R54X mutation in DNMT3B, provided by Dr. R.S. Hansen). Control EBV-immortalized B-LCLs were from unaffected unrelated individuals (Ctrl1 and Ctrl4) and phenotypically normal parents of ICF1p1 patient (Ctrl2 and Ctrl3). The B-LCLs from the ICF patients' parents had a similar passage history as the patients' B-LCLs. We confirmed the genome-wide results in additional B-LCLs (kindly provided by Dr. Francastel, INSERM) deriving from ICF1 patients (ICF1pT and ICF1pY, with homozygous mutations D817G/D817G and T775I/T775I, respectively). DNA extraction was performed using Wizard Genomic DNA (Promega). For drug treatment, $1x10^6$ control cells were plated and treated with 1uM 5-AzaC (Sigma) for 24h. Then, the medium containing 5-AzaC was removed and replaced with RPMI+FBS10% and cells were harvested at different time points (48h, 72h and 120h) for further analyses. Total RNA was extracted using the TRIzol reagent (Life Technologies) according to the manufacturer's instructions. RNA quality was checked on the Agilent 2100 Bioanalyzer and quantity was measured on a Qubit instrument (Life Technologies).

2.2 Reference genome and transcriptome

For sequencing alignment we used the human reference genome assembly GRCh37/hg19 (http://ftp.ensembl.org/Homo_sapiens.GRCh37.75.dna_sm.primary_assembly.fa.gz), while for transcriptome annotation the version 82 of the GRCh37 (http://ftp.ensembl.org/pub/grch37/release-82/gtf/homo_sapiens/Homo_sapiens.GRCh37.82.chr.gtf.gz) was used.

2.3 Reduced representation bisulfite sequencing (RRBS) and data processing

1-10ug of DNA was used for RRBS library preparation according to Illumina's instructions. Libraries were generated and sequenced at IGA Technology Services (Italy), by using the NuGEN Ovation Ultralow Methyl-Seq Library System and 50bp single-end sequencing on the Illumina HiSeq2500 platform. Sequence reads were processed by adaptor trimming (Illumina Pipeline Casava) and filtering for low quality reads and subjected to quality control (FastQC).

Two technical replicates for each DNA sample with independent bisulfite conversion and library preparation were produced.

Reads were aligned to the reference genome using the Bismark aligner (Krueger and Andrews, 2011) and methylation call was performed with methylation extractor script. A summary of mapped reads using the BisMark aligner against the reference genome is presented in TableS2. The overall DNA methylation correlation between the two technical replicates was 0.95 for both Ctrl1 and Ctrl2, 0.96 for the ICF1p1 and 0.91 for the ICF1p2. The highly correlated replicates were pooled for further analyses. We used the R library package methylKit (Akalin et al., 2012) to calculate methylation percentages per each single CpG, by dividing the number of methylated Cs by the total coverage on that base. CpGs with at least 10X read coverage were retained for calling CpG methylation. We calculated DNA methylation state at

1,291,564 CpG sites of genomic regions including promoters, exons, introns, intergenic regions and a high percentage of CpG islands (FigS1A). Differentially methylated regions (DMRs) between samples were defined as sequences with minimum 10 (less than 2Kb apart) common CpGs with concordant methylation difference (≥ |25%|). The whole genome bisulfite sequencing (WGBS) data were downloaded from the Gene Expression Omnibus (GSE37578)(Heyn et al., 2012). DMR annotation to genes was performed using the R library package ChIPpeakAnno v.2.16.4 (Zhu et al., 2010). Distances between peaks and genes were calculated with a home-made script and TSS-2Kb/TTS+2Kb was considered as distance.

Gene-specific DNA methylation level was evaluated by MethylMiner Methylated DNA Enrichment Kit (Invitrogen) according to the manufacturer's instructions. Primer sequences are reported in **TableS3.**

2.3 RNA sequencing and data processing

RNA isolation and library construction was performed according to Illumina's instructions. After mapping the reads to the reference genome using TopHat2 (Kim et al., 2013), differential gene expression between two different cell conditions was calculated using DEseq implemented in R (Anders and Huber, 2010), while the relative isoform abundance estimation was performed using BitSeq (Glaus et al., 2012).

Two independent RNA-seq experiments were carried out and the results were compared after performing the analysis of the two datasets separately. Reads were mapped to the reference genome using TopHat2 v.2.0.13 (Kim et al., 2013). In the first step we aligned the reads against the transcriptome. We used the following non-default TopHat2 parameters: -r 250, -m 2, --min-coverage-intron 50, --max-coverage-intron 100000, —mate-std-dev 50, —segment-

length 17. The total tag number obtained for each sample is listed in **TableS2.** The counts of sequenced reads per annotated gene were derived with the use of htseq-count script distributed with HTSeq (Anders et al., 2015). We used the R library package DESeq v.1.20.0 (Anders and Huber, 2010) for measuring differential gene expression between two different cell conditions. DESeq treats gene expression data as count data modeled under a negative binomial distribution. We picked out genes with p-value <0.05. We considered as list of differentially expressed genes (DE-genes) only the overlapping group of genes between the two independent RNA-seq experiments.

The relative isoform abundance estimation was performed using BitSeq (Glaus et al., 2012), which is a tool for inferring transcript expression with a probabilistic model of the read generation process based on a Markov chain Monte Carlo (MCMC) algorithm for Bayesian inference over the model. We used the default settings (-p) and we extract as optput the RPKM (-outType). From BitSeq output we selected only the genes shared by two independent replicates, which were associated to isoforms with: $(\sigma^2_{ICF} + \sigma^2_{wt}) <$ 95th percentile, RPKM ≥ 0.1 (at least in one condition), $\log_2(RPKM_{ICF}/RPKM_{wt}) \geq |1.5|$.

Gene ontology analysis was performed using DAVID Bioinformatics Resource (Huang da et al., 2009), while enrichment of specific pathways was analysed using Ingenuity pathway analysis (IPA; http://www.ingenuity.com).

2.4 Quantitative Real time PCR

Total RNA from B-LCLs was reverse-transcribed using iScript cDNA Synthesis kit (Bio-Rad San Diego, California). Quantitative real-time PCR (qRT-PCR) was performed using SsoAdvanced™ universal SYBR® Green supermix (Bio-Rad) on Bio-Rad iCycler according to the manufacturer's

protocols. The ΔΔCt method was used to determine relative quantitative levels. *GAPDH* was used to normalize the data. Primer sequences for gene expression analysis are shown in TableS3.

2.5 ChIP sequencing and data processing

Chromatin immunoprecipitation was performed as previously described (Matarazzo et al., 2004). Suitable amount of chromatin was incubated with 5 μg of the indicated antibodies against H3K27me3, H3K4me3, H3K9me3, H3K36me3, H3K4me2, Pol II, DNMT1 (Abcam) and anti-DNMT3B (Diagenode). Immunoprecipitated complexes were recovered with protein A sepharose (Pharmacia), washed with low and high salt buffers, reverse-crosslinked, and purified. Primers sequences are reported in **TableS3.**

Immunoprecipitated samples were used for preparing libraries and sequenced at NGS Core facility (IGB, Naples) or at IGA Technology Services (Italy).

DNMT3B immunoprecipitated samples were sequenced at IGA Technology Services (Udine, Italy). 50bp single end reads were aligned on the reference genome using Bowtie v. 1.1.1 (Li and Durbin, 2009).

The parameters used for Bowtie were –a, –m3, –-best, and –-strata.

H3K4me3 and H3K27me3 immunoprecipitated samples were sequenced at IGB NGS Core facility (Naples, Italy) with SOLiD System 4.0 (Applied Biosystems). Reads were first analyzed by the Applied's pipeline software for quality filtering and aligned to the reference genome by using Bowtie; we selected as parameters -C, -k1, -m3, --best, and --strata.

The total tag number obtained for each sample is listed in **TableS2.** We selected for further analysis only uniquely mapped reads and removed the PCR amplification artefacts and the technical replicates were pooled together. All the reported results were obtained by normalizing for the total library size.

DNMT3B peak calling was performed using MACS v.1.4.2 (Feng et al.,

2012). The sample enrichments were normalized for cell line specific input. MACS parameter settings were: band width of 200bp (--bw), p-value 1e-4 (-p) and the range of high-confidence enrichment ratio against background including setting between 8 and 30 (-m). Moreover, we detected large peaks using EDD (Lund et al., 2014), setting a bin size of 2Kb (log_ratio_bin_size), a gap penalty of 5 (-g) and 20000 Monte Carlo trials (-n). Confidence intervals were calculated using the normal approximation method for binomial proportions. The SICER v.1.1 (Zang et al., 2009) peak-finding algorithm was used to identify the H3K4me3- and H3K27me3-enriched sites throughout the genome. For all these histone marks we selected peaks with a false discovery rate (FDR) of 1e-5. We used as window size and gap different values accordingly with the profile of analyzed histone marks. In particular, for H3K4me3 we used window size 200 bp and gap size 200 bp, while for H3K27me3 500 bp and 1000 bp. The gap size selection was performed as described (Zang et al., 2009). We counted the reads on each detected histone mark peak in all conditions using Bedtools (Quinlan and Hall, 2010) and we detected differentially enriched domain using DESeq. We picked out differentially enriched regions with p-value <0.01. Density plots were obtained using NGSplot tool (Shen et al., 2014). We calculated the read count per million mapped reads using fragment length (-FL) 300, as reference region (-R) the position of interesting regions extended of (-L) 2000 and smoothing window (-MW) 10 were used.

DNMT3B peak annotation to genes was performed using the R library package ChIPpeakAnno v.2.16.4 (Zhu et al., 2010). The distances between peaks and genes were calculated with a home-made script and TSS-2Kb/TTS+2Kb was considered. The pericentromeric domains were defined as the regions between the nearest gene to the centromere and the centromere, and filtering out the 10 kb proximal to the gene. Bedtools and Bedops (Neph et al., 2012) toolkits were used for the above described data analysis.

Gene-specific ChIP assays were carried out by quantitative Real Time PCR, by using SYBR Green quantitative PCR (SsoAdvanced Universal SYBR Green Supermix, Biorad) according to CFX96TM Real Time PCR Detection Systems. The enrichment of DNA was calculated in terms of % input = $2^{-\Delta Ct\ x\ 100}$, where ΔCt (threshold cycle) is determined by $Ct_{IP\ sample}$ - Ct_{Input} and 100 refers to the input, which is 1% of the starting chromatin.

2.6 Statistical analysis

Gene expression, ChIP-qPCR, RIP-qPCR and gene specific methylation analysis were presented as means ± standard deviations (SD) from at least 3 independent experiments. Statistical analyses were performed using T-student test (two-tails). P-values were adjusted with BH method and we generally considered the following values as statistically significant: *p-adj<0.05; **p-adj<0.005; ***p-adj<0.0005. The significance of the overlapping between two and/or three gene lists was calculated using the hypergeometric test available in R. Statistics relative to DNA methylation comparison between samples was analyzed using Kolmogorov-Smirnov test and values were corrected with Bonferroni method.

2.7 RNA-Immunoprecipitation (RIP)

RNA immunoprecipitation experiments were performed using the Magna RIP™ RNA-Binding Protein Immunoprecipitation Kit (Millipore) according to the manufacturer's instructions. Antibodies for RIP assays were anti-DNMT3B (Diagenode) and anti-hnRNPLL (Aviva). Immunoprecipitated fractions were retrotranscribed and cDNAs were used for qPCR with gene-specific primers.

2.8 Flow cytometry analysis

Determination of cell surface expression of CD45RABC and CD45RO molecules was carried out by cytofluorimetric analysis using the FACS ARIA cell-sorting system and DIVA software (BD Biosciences). Direct immunofluorescence was performed using PerCP and FITC mouse anti-human CD45RABC and CD45RO antibodies respectively, along with the appropriate mouse IgG isotype controls (BioLegend). Staining, washing and analysis were performed following the manufacturer's recommendations.

2.9 Co-immunoprecipitation

Co-immunoprecipitation experiments were performed using Nuclear Complex Co-IP kit (Active Motif) and following the manufacturer’s instructions. The antibodies used were the following: DNMT3B (Diagenode), Suz12 (Abcam), hnRNPLL (Aviva).

Chapter 3

RESULTS

3.1 ICF1-specific *DNMT3B* mutations mainly affect CpG methylation at intragenic regions

In humans, hypomorphic DNMT3B mutations are sufficient to cause majority of the rare autosomal recessive disorder Immunodeficiency, Centromere instability and Facial anomalies (ICF) syndrome (MIM 242860) cases, reported as ICF type1 (Hagleitner et al., 2008; Matarazzo et al., 2009; Weemaes et al., 2013). Patients are characterized by DNA hypomethylation and decondensation of specific heterochromatic and euchromatic regions, and show alterations in tissue-specific gene silencing (Jin et al., 2007; Matarazzo et al., 2007). ICF1-specific DNA methylation defects give rise to severe chromosomal rearrangements only in lymphocytes, probably acting in the onset of immunological phenotype. Defective steps of B-cell terminal differentiation might contribute to the agammaglobulinemia in ICF syndrome, given that ICF peripheral blood only contain naive B cells, while memory and gut plasma cells are absent (Blanco-Betancourt et al., 2004).

Most ICF1 patients carry missense mutations in or near the catalytic domain of DNMT3B (Weemaes et al., 2013). Nonsense mutations always occur as compound heterozygous, highlighting that the DNMT3B protein is essential for life, according to mouse models (Okano et al., 1999); (Ueda et al., 2006; Velasco et al., 2010). Mutations dramatically perturb the DNA methylation profile at satellite 2 and 3 of juxtacentromeric heterochromatin and at telomeric/subtelomeric repeats, where it associates with chromosomal instability and abnormal shortening of telomeres, respectively (Jeanpierre et al., 1993; Gisselsson et al., 2005; Yehezkel et al., 2008).

Previous candidate gene approaches were unsuccessful in identifying ICF1-

specific DNA hypomethylation at CpG island-associated promoters of DE-genes, indicating either that only few of them are direct target of DNMT3B or that differences in DNA methylation level are outside promoters (Ehrlich et al., 2001; Jin et al., 2007). Methylome studies carried out in one ICF1 B-LCL suggested that DNA methylation defects might be more extended than what supposed until now (Heyn et al., 2012; Simo-Riudalbas et al., 2015). However, the study did not explore how DNA methylation defects functionally impact on gene expression regulation and whether this effect occurs through the modulation of other epigenetic marks.

In this light, we carried out Reduced Representation Bisulfite Sequencing [RRBS; (Smith et al., 2009)] to quantify methylation differences at single CpG sites in B-LCLs derived from peripheral blood of two unrelated ICF patients with DNMT3B missense mutations, ICF1p1 and ICF1p2, compared to control B-LCLs. These cells represent a suitable model for ICF studies because of the central role of B cells in abnormal immunoglobulin production in ICF cases and the scarcity of primary B cells from patients, of which only few reach adulthood. Despite cell culturing, B-LCLs maintain highly significant ICF-specific differences in mRNA levels of immunoglobulin genes [(Ehrlich et al., 2001; Gatto et al., 2010) and the present study)] and high frequencies of karyotypic anomalies described in mitogen-stimulated ICF lymphocytes (Gisselsson et al., 2005).

We first quantified the methylation level of 1,291,564 CpGs spanning a number of different functional regions in the genome and belonging to different CpG contexts **(FigS1A)**, and with sequencing depth of at least 10 reads in each condition (the median depth was >23 reads among these CpGs in each condition). We identified 79,521 and 174,030 hypo-methylated CpGs in ICF1p1 and ICF1p2 respectively, as compared to the mean CpG methylation percentage of control samples (**Fig1A,B).** CpG clustering and PCA analysis showed that samples from patients and controls are distinguishable by their methylome (**FigS1B,C**).

When we compared DNA methylation values from RRBS and the whole-genome bisulfite sequencing (WGBS), previously reported for one ICF1 sample (Heyn et al., 2012), we found that they were concordant (**FigS1D**). Moreover, because the DNA methylation profile may be heavily influenced by immortalization and culture condition, we validated our results by comparing the RRBS results with DNA methylation assessment by Infinium HumanMethylation 450K in PBMCs of ICF1 patients (Dr. Francastel, personal communication). We found that the examined CpGs corresponding to the hypomethylated probes in 450K at all the genic features (promoter, TSS, gene body and 3'UTR) were significantly hypomethylated also in both our ICF1 samples (**FigS1E).** As a whole, we observed that PBMCs and B-LCLs (ICF1p1 and ICF1p2) shared 52% and 57% of hypomethylation sites respectively, confirming the validity of B-LCLs to study the ICF1 DNA methylation defects.

Besides the expected hypomethylation, ICF1 samples also showed 75,573 and 35,872 hypermethylated CpGs (ICFp1 and ICFp2 respectively), as compared to the mean CpG methylation percentage of control samples (**Fig1B)**. In contrast to observations in ICF1 iPSCs (Huang et al., 2014) and similar to a recent report in cancer (Berman et al., 2012), our results did not show substantial levels of DNA methylation at non-CpG sites in ICF1 or control B-LCLs (data not shown). According to RRBS results, we calculated differentially methylated regions (DMRs) as sequences with at least 10 common CpGs (less than 2Kb apart) with concordant methylation difference ($D \geq |25\%|$). We identified 2,381 in ICF1p1 (1,833 hypo-DMRs and 548 hyper-DMRs) and 3,520 DMRs in ICF1p2 (3,053 hypo-DMR and 467 hyper-DMRs), which affected all the chromosomes **(FigS2A).** Of note, 74% of hypomethylated DMRs in ICF1p1 were shared with ICF1p2, while 13% of hypermethylated DMRs were shared between the two ICF1 samples. As previously reported, pericentromeric heterochromatin and other repetitive sequences (satellite, LINE, SINE, LTRs, etc) were considerably

hypomethylated in ICF1 samples **(FigS2B,C)**.

Remarkably, most hypo and hyper-DMRs in both ICF1 samples affected CpG-dense regions at gene-bodies, particularly at exons and introns rather than at promoters **(Fig1C,D).** To gain insights into the functions of DMR associated genes (DMR-genes) we performed gene ontology (GO) analysis for each ICF1 sample. Interestingly, hyper-DMR and hypo-DMR genes were significantly enriched at genes involved in developmental and transcriptional regulation processes (**Fig1E and TableS1**). Moreover, they were enriched for transcription factors with homeobox and DNA binding domains (**Fig1E**), which were consistently differentially expressed in RNA-seq experiments and qPCR suggesting that these genes are preferential DNMT3B target genes. This may explain the transcriptional deregulation of other, DNMT3B independent, downstream genes. For instance, hypomethylated and aberrantly expressed DNMT3B target genes known to act as transcription factors and/or as modulator of RNA Polymerase II (Pol II) mediated transcription were *BCL11B*, *PRRX1*, *NR2F2*, *TCF12* and *SATB1* (**FigS2D**). Intriguingly, all of them are functionally involved in developmental processes altered in ICF syndrome. Overall, we found that DMRs associated genes and genes bound by DNMT3B, identified by ChIP-Seq experiment, in ICF1 and control samples significantly overlapped (p-value: $<10^{-10}$; $<<10^{-10}$; **Fig1F and TableS1**), confirming that DNA methylation defects occur mostly at DNMT3B target genes and that are dependent on DNMT3B mutations. However, when we compared the DMR and/or DNMT3B associated genes with DE-genes identified by RNA-Seq we found overlaps extending over approximately 10% of DE-genes, indicating that the transcription of most genes is indirectly affected by DNMT3B dysfunction (**Fig1F**). Moreover, we also identified DNMT3B target genes, which were transcriptionally deregulated without showing a clear difference in DNA methylation profile, confirming that DNMT3B may modulate the gene expression independently from DNA methyltrasferase activity (**Fig1F;**(Bachman et al., 2001).

Overall, we found that nearly 35-38% of DE-genes were shared by the two ICF1 samples when compared to controls. Consistently with previous findings, we identified some germline genes as target of DNMT3B, hypomethylated at CpG islands and derepressed, like *SYCE1*, *MAEL, TDRD1 and TDRD9* (Velasco et al., 2014); **FigS3A**). Moreover, up- and down-regulation of ICF1-specific genes previously reported were confirmed in our RNA-Seq datasets [(Ehrlich et al., 2001; Jin et al., 2007); **TableS1**].

Figure 1

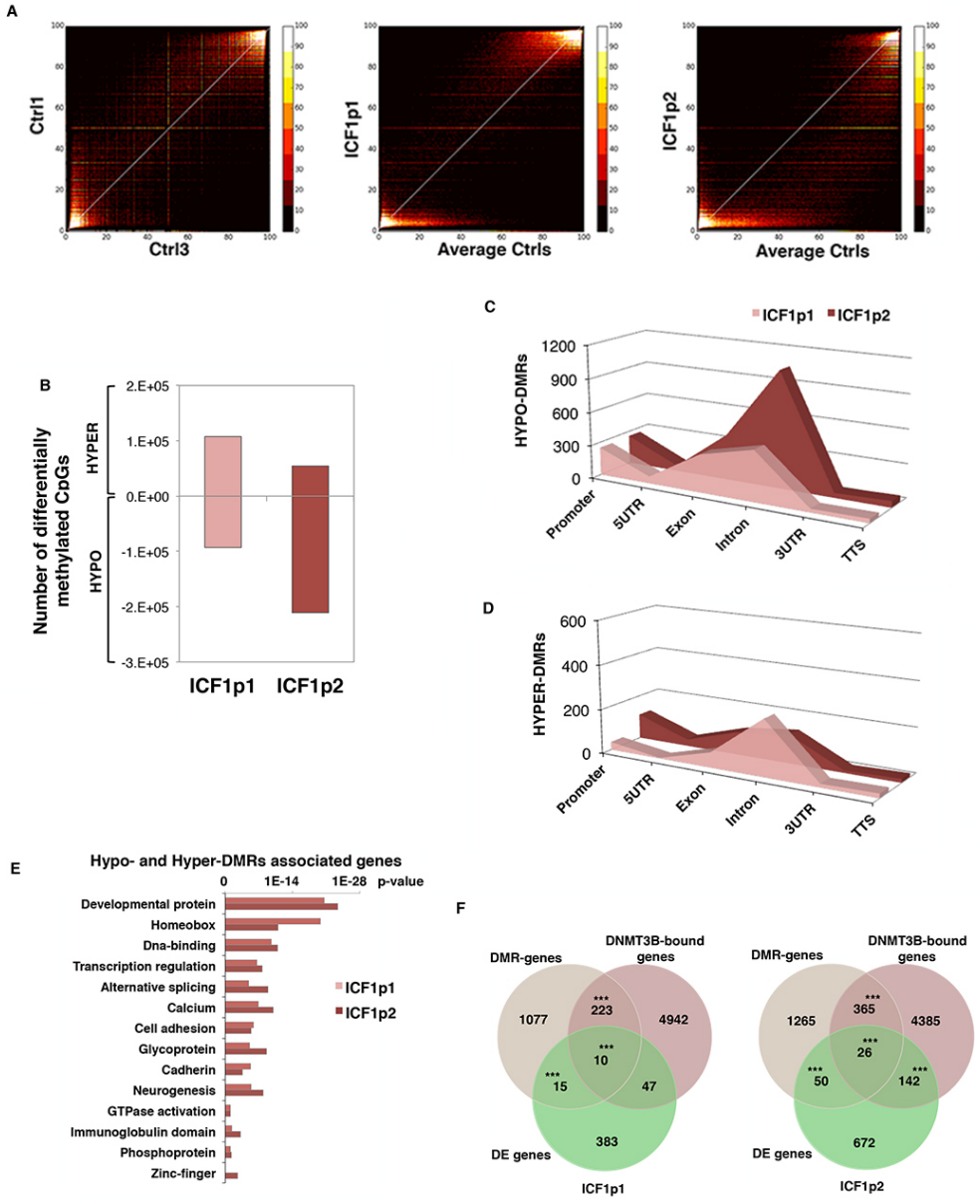


Fig1. Large-scale DNA methylation profile shows CpG hypo- and hypermethylation at intragenic regions. A, Scatter plots and density color codes for DNA methylation data of all autosomes. Pairwise comparisons of methylation percentage between controls (Ctrls) and between ICF1p1 or ICF1p2 and average Ctrls are shown; B, Histogram showing total number of hypo- and hypermethylated CpGs in ICF1p1 and ICF1p2 compared to controls; C,D, Distribution of hypo and hyper-DMRs (D≥ |25%|) along the gene features; E, Gene Ontology (DAVID) of hypo- and hyper-DMRs associated genes (from TSS-2kb to TTS+2kb; p-values corrected by BH method were considered; F, Venn Diagram showing overlaps between DMR-associated genes, DE-genes and DNMT3B-bound genes (p-values were calculated with hypergeometric test (one-tail); for the three lists overlap we considered the DMR-genes/DE-genes and DNMT3B-genes/DE-genes subsets).

3.2 DNMT3B deficient activity correlates with H3K27me3 redistribution at genic and intergenic regions

As mentioned above, we found that most CpG methylation alterations were not associated with marked differences in expression of the associated genes. This suggests that DNA methylation defects alone are prevalently not sufficient to perturb gene expression. In these cases, proper transcription might be preserved by other epigenetic signals. For instance, H3K27me3 is a repressive histone mark cooperating with DNA methylation in ensuring proper gene silencing in somatic cells. Therefore, we hypothesized that the DNMT3B deficient activity might influence the genome wide signature of these histone marks, by igniting compensatory mechanisms. We analyzed H3K27me3 profile in ICF1p1 and control sample by ChIP-Seq, finding a redistribution of this mark, with a general increase of H3K27me3-enriched, which prevalently overlap to genic regions **(Fig2A and FigS3B).**

To determine the direct relationship between histone and DNA methylation, we integrated RRBS and ChIP-Seq datasets to calculate the normalized average density of histone mark enrichment at differentially methylated genomic regions. We found a significant increase of H3K27me3 at hypomethylated DMRs, whereas it was not enriched at hyper-DMRs in all samples **(Fig2B and FigS3C)**. This indicates that DNMT3B mutations alter the H3K27me3 profile at the hypomethylated target regions, likely affecting the recruitment of PRC2 complex (Vire et al., 2006). Interestingly, a significant number of regions increasing in H3K27me3 were also targeted by mutant-DNMT3B in ICF1p1 sample (**Fig2C)**. In the context of transcriptional regulation, most hypomethylated genes showed unaffected and/or increased H3K27me3 (Log2FC ICF1/Ctrl H3K27me3 >= 0) and genes associated to this mark did not change their expression, suggesting that the repressive H3K27me3 mark is presumably balancing the loss of DNA methylation (**Fig2D**).

One interesting case is given by the *HOXC* gene cluster, which belongs to the

hypomethylated subset of genes **(Fig2E, left panel).** Here, the H3K27me3 mark profile in ICF1p1 sample was not affected in most part of the cluster area **(Fig2E, left panel)**, thereby preserving the silencing of most *HOXC* genes. Interestingly, *HOXC4* gene that is supposed to be normally expressed in B-LCLs, being essential for immunoglobulin class switch recombination, resulted abnormally silenced in disease samples and this correlated with CpG hypomethylation and H3K27me3 increase mediated by PRC2 complex binding **(Fig2E, right panels).** Intriguingly, mutant-DNMT3B efficiently bound the H3K27me3-enriched region of *HOXC4* in ICF1 samples, suggesting that it may influence the recruitment of PRC2 complex to the neighboring nucleosomes. Consistently, SUZ12 binding at *HOXC4* was preferentially enriched in ICF1 samples when compared to controls **(Fig2E, right panels)**. Notably, we found that mutant-DNMT3B, which is known to have a correct folding according to our and previous observations (Geiman et al., 2004) was able to interact with SUZ12 protein **(FigS3D).**

Overall, we found that the ability of mutant-DNMT3B to bind DNA was not impaired in terms of total number of peaks, but rather in terms of their extension depending on the DNMT3B variant. Indeed, DNMT3B peaks larger than 100kb were increased in ICF1p1 sample, compared to ICF1p2 sample and two controls (data not shown). Remarkably, by examining the distribution of DNMT3B binding sites along the genic features, we found that wild-type and mutant-DNMT3B preferentially bound gene bodies (Fig2F), where most DNA methylation defects occurred (Fig1D,E). CpG methylation level at mutant-DNMT3B genomic targets clearly decreased compared to controls, as demonstrated in both RRBS and WGBS, indicating that ICF1-specific mutations prevalently cause hypomethylation (FigS3E), while the hypermethylation probably reflects the compensatory activity of DNMT3A and/or DNMT1. Consistently, the ICF1p1 sample showing higher level of DNA hypermethylation than ICF1p2 sample (Fig1E) also presented a lower number of DNMT3B peaks at gene body regions (Fig2F).

Given that the extensive loss of DNA methylation at pericentromeric heterochromatin represents an hallmark of ICF molecular phenotype, we examined the DNMT3B binding at repetitive regions, finding that it was unaltered in ICF1 samples compared to controls, as for genes (FigS4A-C). Taken together, these evidences indicate that ICF1-specific mutations do not affect the DNMT3B targeting at genomic regions, which are nonetheless hypomethylated because of the deficient methyltransferase activity.

When examining the DNMT3B target genes in terms of gene ontology analysis, we observed that most of them were associated to developmental processes and in particular to neurogenesis, presumably accounting for the variable cognitive defects reported in ICF1 patients **(FigS4D)**. However, among them we found genes, functions of which were linked to ICF-specific immune system phenotype, such as genes contributing to the immunoglobulin production (e.g. *FAS, FOXP1*) and to the humoral immune response mediated by circulating immunoglobulin (e.g. *BCL11, PTPRC, CR1, FYN, CD247*). Moreover, we identified genes functionally involved in chromosome condensation and segregation at centromeric regions (e.g. *KIF2B, CENPU, KNTC1, SPC25, ZWINT),* which may account for the chromosomal instability and the mitotic missegregation (Gisselsson et al., 2005). Genes involved in chromatin binding and remodeling, and genes regulating histone deacetylation (e.g. *HDAC9*) and methylation (e.g. *MLL5, SMYD3, MLL3)* were also detected as DNMT3B target genes, suggesting the occurrence of an even more complex regulatory cross-talk between DNA methylation and histone modifications. Nevertheless, our finding from DESeq analysis was that only few DNMT3B bound genes resulted differentially expressed **(Fig1F)**. This prompted us to better and deeper investigate the potential defects of the ICF1 transcriptome.

Figure 2

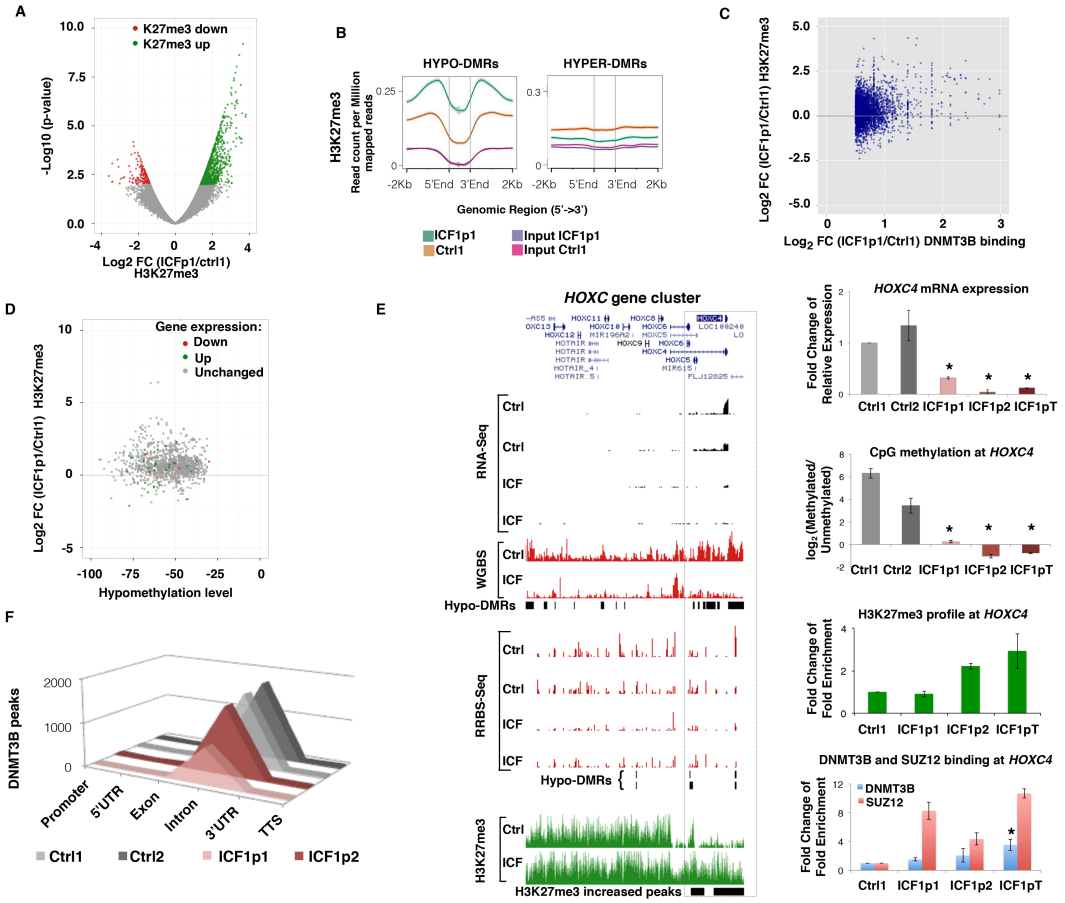


Fig2. CpG hypomethylation correlates with H3K27me3 increase. A, Volcano plot showing that DNMT3B mutations associate with H3K27me3 redistribution in ICF1p1 sample, with up and down regions; B, Density plot of H3K27me3 (read count per million mapped reads) at hypo- and hyper-DMRs (-2kb upstream/+2kb downstream); C, Scatter plot showing the correlation between differential DNMT3B-enriched regions (Log2 Fold change ICF1p1/Ctrl1) and differential H3K27me3-enriched regions (Log2 Fold change ICF1p1/Ctrl1); D, Scatter plot showing the expression status of hypomethylated genes and differential H3K27me3-enrichment (Log2 Fold change ICF1p1/Ctrl1); E, *HOXC* gene cluster is represented in the genome browser UCSC screenshot, showing RNA-Seq, WGBS, RRBS and H3K27me3 ChIP-Seq (left panel). The setting of vertical viewing range is the same for ICF1 and control samples in each experiment. Right panels show mRNA and CpG methylation level (qPCR and DNA methylation enrichment assay), H3K27me3 enrichment and DNMT3B or SUZ12 binding (ChIP-qPCR) at *HOXC4* gene in ICF1 samples compared to controls (p-values were calculated with T-student test (two-tails) and adjusted with BH method); F, Distribution of DNMT3B peaks along the genic features (promoter, 5'UTR, exon, intron, 3'UTR and TTS) calculated using Homer tool (Heinz et al., 2010).

3.3 Isoform-specific transcriptional regulation is severely impaired in DNMT3B deficient cells

By analyzing the gene ontology of DNMT3B ChIP-Seq and RRBS global profiles, we surprisingly found that DNMT3B target genes and DMR-genes were significantly enriched for the alternatively splicing functional category in both control and disease samples [~45% of DNMT3B targets and DMR-genes (**Fig1E and FIgS4D**)]. This suggested that DNMT3B might play a role in the regulation of alternative isoform expression. To address this possibility and investigate whether this process is perturbed by ICF1-specific DNMT3B mutations, we analyzed the RNA-Seq datasets by using BitSeq tool, which evaluates the exon-exon junction usage to measure the transcript isoform abundance (Glaus et al., 2012). We found significant differences among the isoforms expressed in disease and control cell lines revealing major alterations at a larger number of genes not identified from previous RNA-Seq analysis **(Fig3A, upper panel and TableS1)**. We found that nearly 55% of DE-isoform associated genes were shared by the two ICF1 samples when compared to controls. The newly identified genes belonged to functional categories relevant for ICF immunological phenotype, such as phosphoproteins, proteins with GTPase regulatory activity and chromatin modifiers **(Fig3B).**

Remarkably, among the genes with the most altered isoform abundance we identified key players of immunoglobulin production regulation and immunoglobulin mediated immune response, such as *FOXP1* and *IL10,* which were not detected before based on the classical gene expression analysis. The most significant pathway affected in both ICF1 samples was the B cell receptor signaling, as identified based on IPA analysis (36% of genes within the pathway showed differential isoform expression; Benjamini- Hochberg (BH) adjusted p-value: $1x10^{-3}$ and $1x10^{-7}$ for ICF1p1 and ICF1p2, respectively) **(FigS4E).** Furthermore, critical genes modulating chromosome condensation especially at centromeric regions, such as the centromeric protein *CENPU,*

components of the spindle-assembly checkpoint *MAD1L1, KNTC1 and SPDL1, the* component of the condensin complex *SMC4,* were identified. Intriguingly, the newly identified genes with differential isoform abundance were significantly enriched in DNMT3B-bound and differentially methylated genes (hypo- and hyper-DMR associated genes) compared to DESeq-derived dataset, suggesting a functional effect of DNA methylation changes on isoform selection **(Fig3A, lower panel**).

Coherently, CpG hypomethylation (-500bp/TSS covered by RRBS) was observed at about 80% of up-regulated isoform-specific intragenic TSS (i-TSS; **FigS5A).** Moreover, the enriched gene ontology categories of DE-isoforms and DNMT3B target genes were largely overlapping, supporting the idea that these defects are possibly direct consequence of DNMT3B deficient activity.

In addition, we evaluated the H3K4me3 and H3K27me3 pattern at TSS of DE-isoforms, finding changes of these histone marks at 57% and 25% of hypomethylated and hypermethylated TSS, respectively **(Fig3C)**. Namely, H3K4me3 increased or decreased based on TSS methylation status (hypo- or hypermethylation, respectively), while H3K27me3 increased at hypomethylated TSS**,** according to our previous results at genomic DMRs **(Fig2B)**. As an example, the longest isoform of *TCEA2* gene including the most upstream TSS was abnormally silenced in both ICF1 samples. Moreover, the memory B-cell marker *CD27* was downregulated, while the overlapping antisense transcript *CD27-AS1* resulted upregulated, as revealed by isoform-specific analysis**.** Both transcriptional deregulation events correlated with DNA methylation and H3K4me3 changes **(Fig3D)**. Of note, H3K4me3 level also changed at all genomic hypo- and hyper-DMRs consistently, suggesting that CpG methylation defects induce H3K4me3 changes **(Fig3E)**.

Intragenic DNA methylation has been proposed as mechanism potentially able to modulate alternative splicing events (Lev Maor et al., 2015). Therefore, to test whether ICF1-specific DNA methylation defects may affect this process,

we extrapolated the annotated splicing events (i.e. alternative 5' or 3' splice site, cassette exon, intron retention, etc) and searched for a significant association with hypo- and hypermethylated CpGs (at least 5 differentially methylated CpGs within 150bp from the alternative splice site) in ICF1 samples. We found 129 splicing events in proximity of regions showing decreased or increased CpG methylation (118 associated to hypo-CpGs and 11 to hyper-CpGs, respectively) common between ICF1p1 and ICF1p2 samples compared to controls. A significant percentage (17-20%; p-value adjusted $=3.2x10^{-6}$ and $3.6x10^{-4}$**,** respectively) of these identified splicing events overlapped with the differentially expressed transcript isoforms identified by BitSeq analysis. To confirm this result at wider scale and to examine these events in detail, we performed the same analysis taking advantage of WGBS. This allowed us to extensively measure the correlation between specific splicing events associated to DE-isoforms and differential CpG methylation. We found that the most affected splicing events were variations at the 5' and/or 3' splicing sites of introns in both ICF1 samples **(Fig3F)**. These results support a role for gene body DNA methylation in regulating alternative splicing events, which are perturbed in presence of DNMT3B-mediated CpG hypo- and hypermethylation in ICF1 B-LCLs.

Among the differentially spliced genes, we identified the Protein Tyrosine Phosphatase, Receptor type, C gene (*PTPRC*; also known as *CD45*), which is a trans-membrane protein tyrosine phosphatase essential for antigen receptor-mediated signaling in lymphocytes (Rhee and Veillette, 2012). Differences in exon skipping and glycosylation of the protein produce various isoforms of CD45 that are present in a cell-specific manner. CD45RO, the smallest isoform, distinguishes effector-memory T cells, whereas primary B cells are known to express predominantly the largest CD45 protein isoform (encoded by RABC, including exons 4-5-6). In line with the in vivo findings, the cultured B cells and B-LCLs express almost uniquely the RABC and RBC isoforms (Hermiston et al., 2003). When we carried out isoform-specific PCR,

we found that ICF1 cells exhibited significantly higher CD45RO isoform expression compared to controls, while the CD45RABC isoform decreased, suggesting that a proper DNMT3B activity is necessary to ensure the inclusion of the exons 4-5-6 **(Fig4A-C).** The altered isoform transcription led to a change from $CD45RABC^{high}/CD45RO^{low}$ population in control cells to $CD45RABC^{low}/CD45RO^{high}$ population in ICF1 cells, which is more or less pronounced depending on DNMT3B mutations **(Fig4D).**

Given the growing evidence for chromatin-mediated regulation of spliceosome assembly and alternative exon recognition, we tested the enrichment of various histone marks known to be involved in this regulation, as for instance H3K36me3, H3K4me3, H3K9me3 and acetylated histone H3 (Zhou et al., 2014). While H3K36me3 and H3K9me3 did not show significant differences at the exons 4-6, we observed a clear decrease of H3K4me3 in all ICF1 samples indicating that this histone mark is associated with the proper inclusion of alternative exons 4-6 **(Fig4E)**. Notably, very recent observations positively linked H3K4me3 enrichment with inclusion level of exons displaying cell type-specific splicing (Curado et al., 2015). In this light, H3K4me3 decrease in ICF1 samples may contribute to aberrant exclusion of alternative exons 4-6.

Conversely, the alternative exons were highly CpG methylated, either in control or ICF1 samples, without showing significant changes between the two conditions **(Fig4F)**. Considering that RNA polymerase II (Pol II) elongation rate may influence exon inclusion, we also examined its binding at skipped exons in ICF1 and control samples. We found that the disease associated splice variant correlated with a reduced binding of Pol II throughout the exons 4, 5 and 6, according to a faster transcriptional elongation rate **(Fig4E)**, thereby accounting for the increased exclusion of these exons. Taken together, our results identified an additional layer of epigenetic regulation of the transcription that is impaired in ICF1-specific condition, with unprecedented and interesting functional implications. In

CD45 gene specific case, chromatin changes may directly perturb the exon recognition presumably through modulating the kinetics of Pol II elongation, but do not explain how DNMT3B-mediated defects do alter the transcript splicing.

Figure 3

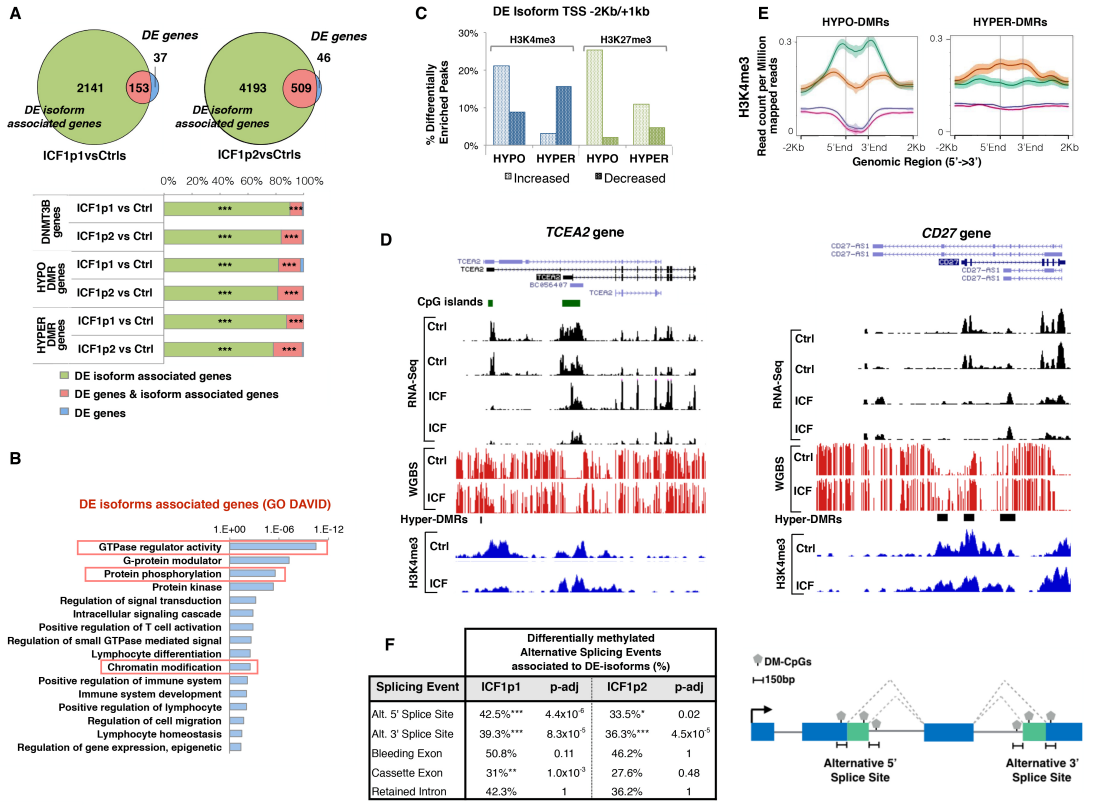


Fig3. Differential expression of transcript isoforms in ICF1 samples and alteration of intragenic epigenetic marks. A, Venn Diagram showing overlap between DE-genes obtained using DESeq and DE-isoform associated genes obtained using BitSeq tools (upper panel); Histograms showing the distribution of DE-genes and/or DE-isoform associated genes among DNMT3B bound and hypo- and hyper-DMRs genes (lower panel; p-values were calculated with hypergeometric test one-tail); B, Gene ontology (DAVID) showing enriched molecular functions and biological processes from the list of DE-isoforms associated genes (p-values corrected by BH method were considered); C, Histogram representing H3K4me3 and H3K27me3 changes (p-value <0.05) at hypo- and hypermethylated TSS of DE-isoforms; D, UCSC genome browser screenshots of *TCEA2* and *CD27* genes showing differentially expressed transcript isoforms, CpG methylation status and H3K4me3 enrichment at corresponding TSS E, Density plot of H3K4me3 (read count per million mapped reads) at hypo- and hyper-DMRs (-2kb upstream, +2kb downstream); F, Percentage of differentially methylated alternative splicing events (from UCSC list) associated to differentially expressed isoforms identified using BitSeq analysis (left panel; p-values were calculated with hypergeometric test one-tail and corrected with BH method); schematic representation of the most affected alternative splicing events linked to DNA methylation defects in ICF1 samples (right panel).

Figure 4

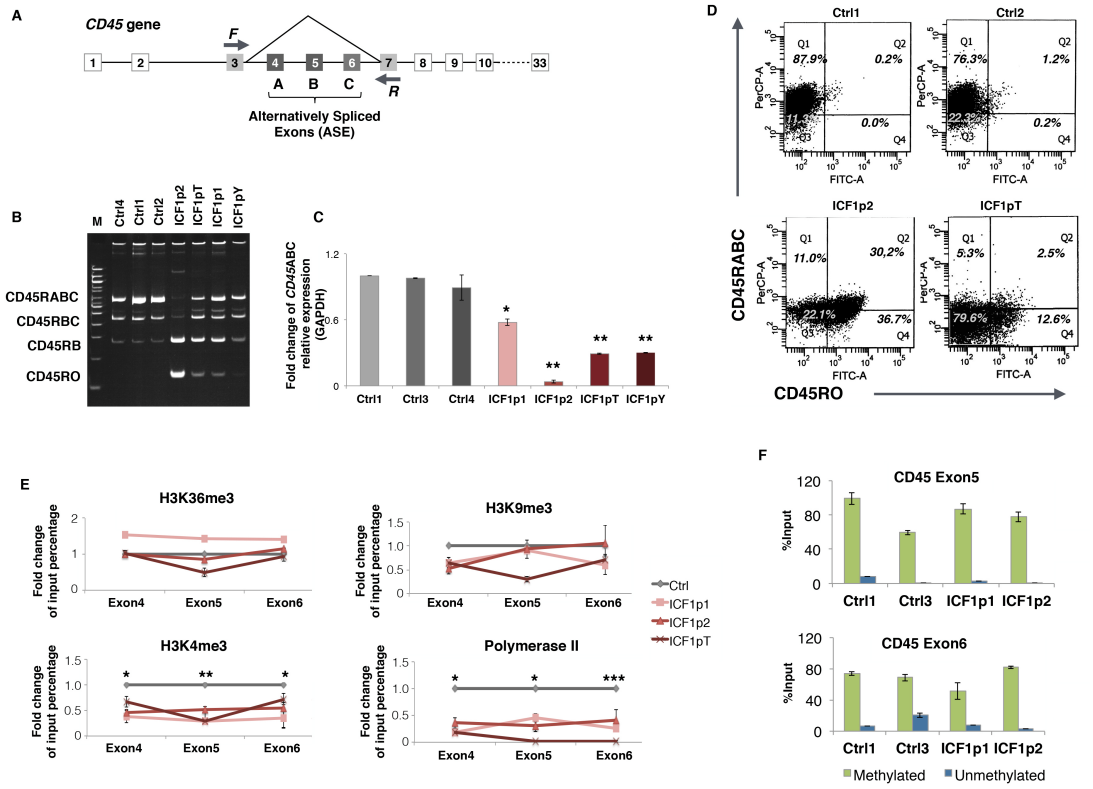


Fig4. Alternative exons 4-6 of *PTPRC (CD45)* gene are aberrantly excluded in ICF1 samples. A, Human *CD45* gene structure with constitutive and alternative (dark gray) exons reported; B, Semiquantitative PCR amplification of splicing isoforms of *CD45* gene in controls and ICF1 samples. Primers used are F and R as shown in A; C, Expression level (qPCR) of CD45RABC using isoform specific oligonucleotides; D, FACS analysis confirming that ICF1 samples express higher level of CD45RO than CD45RABC compared to controls; E, Enrichment of H3K36me3, H3K4me3, H3K9me3 marks and Pol II binding at alternative exons 4-6 by ChIP-qPCR; F, DNA methylation assay (methylated DNA enrichment method) at exons 5 and 6 of *CD45* gene. P-values in C, E and F were calculated with T-student test (two-tails) and adjusted with BH method.

3.4 Mutant-DNMT3B might affect the alternative splicing of *CD45* transcript by interacting with premRNA and hnRNP-LL

In one study DNMT3B has been reported capable to bind components of the Pol II-mediated transcriptional machinery, transcript elongation factors and heterogeneous ribonucleoprotein during in vitro hESC differentiation (Rigbolt et al., 2011), thus supporting a possible role in modulating mRNA processing. Therefore, in order to understand the mechanism leading to the altered *CD45* isoform expression we further examined whether and how mutant-DNMT3B was involved in the exon 4-6 skipping. We explored the hypothesis that DNMT3B modulates the splicing exons by directly interacting with the transcript, considering the well-known affinity of DNA methyltransferases for RNA molecules (Jeffery and Nakielny, 2004; Holz-Schietinger and Reich, 2012; Di Ruscio et al., 2013). By carrying out anti-DNMT3B RNA immunoprecipitation we found that the mutant protein significantly bound the *CD45* pre-mRNA. In contrast, this interaction was barely appreciable in control cells, or absent examining a different gene, thereby representing a negative control (e.g. *HOXC4*) **(Fig5A).**

We next sought to identify molecules potentially participating to this novel regulatory mechanism. *CD45* alternative splicing is tightly controlled by a tissue-specific ribonucleoprotein, the heterogeneous nuclear RNA-binding protein L-Like (hnRNP-LL). It is specifically induced in terminally differentiated lymphocytes, including effector T cells and plasma cells, where it mediates the transition from CD45RA or CD45RABC to CD45RO, respectively (Oberdoerffer et al., 2008; Chang et al., 2015).

Latest models of combinatorial alternative splicing propose that hnRNP-LL cooperates with the heterogeneous ribonucleoprotein hnRNP-L on *CD45* pre-mRNA, bridging exons 4 and 6 and looping out exon 5, thereby achieving full repression of the three variable exons (Preussner et al., 2012). Remarkably, *hnRNP-LL* was derepressed in ICF1 cells compared to controls, and this correlated with mutant-DNMT3B mediated CpG hypomethylation of the

regulatory region **(Fig5B-D and FigS5B)**. Conversely, the transcription of the protein interactor *hnRNP-L* was unaffected (data not shown). As further demonstration of the altered hnRNP-LL-mediated regulation of *CD45* splicing, we observed that hnRNP-LL upregulation was associated to an increased interaction with *CD45* pre-mRNA compared to controls, by performing anti-hnRNP-LL RNA immunoprecipitation **(Fig5E)**. Importantly, hnRNP-LL and mutant-DNMT3B physically interacted in ICF1 cells, suggesting the formation of a multi-protein complex presumably perturbing the correct inclusion of exons 4-6 **(Fig5F)**. To directly test whether DNA hypomethylation resulted in defects of *CD45* pre-mRNA splicing, we treated control cell lines with 5-AzaC and harvested cells at different times after treatment. Intriguingly, after modulating DNA methylation we observed a shift toward the exons 4-5-6 exclusion in the mature *CD45* transcript and a decrease of CD45RABC, as it occurs in mutant-DNMT3B ICF1 cells (**Fig5G,H).** This evidence linked DNA hypomethylation to the aberrant exclusion of *CD45* exons 4-5-6. However, CpG methylation profile at differently spliced exons did not exhibit dramatic defects **(Fig4F),** indicating that DNA hypomethylation may also indirectly alter the *CD45* alternative splicing, through affecting the regulators of this process. To highlight the mechanism and investigate whether it involved the heterogeneous ribonucleoproteins known to modulate the *CD45* alternative splicing, we further tested the *hnRNP-LL* expression after 5-AzaC treatments. Notably, we found an aberrant up-regulation of *hnRNP-LL* transcription, thereby explaining the increase of exon 4-5-6 exclusion in 5-AzaC treated cells **(Fig5I,J).** Overall, these results demonstrate that ICF1-specific DNMT3B mutations lead to CpG hypomethylation and derepression of *hnRNP-LL* gene, and to the recruitment of the protein to *CD45* pre-mRNA, where it is engaged in a protein complex through interacting with the mutant-DNMT3B.

Figure 5

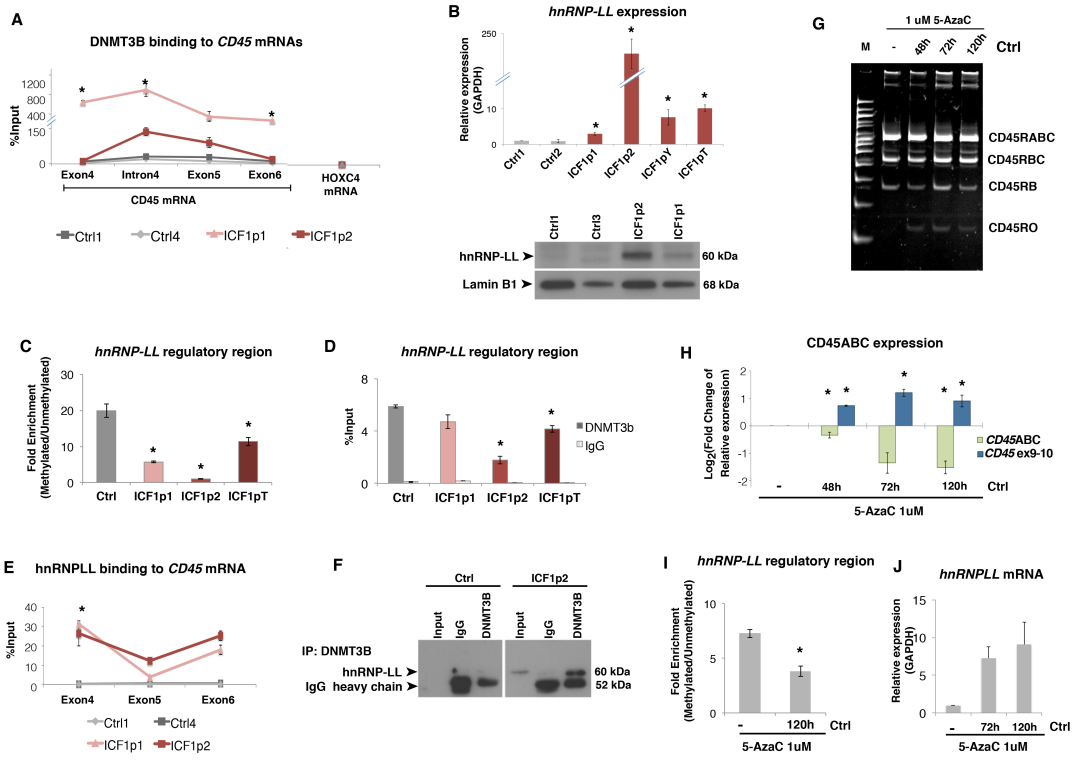


Fig5. Skipping of alternative exons 4-6 of *CD45* correlates with ectopic activity of hnRNPLL. A, RNA immunoprecipitation assay (RIP-qPCR) showing binding of DNMT3B to *CD45* pre-mRNA. Binding to *HOXC4* mRNA is reported as negative control; B, Expression level of *hnRNP-LL* in ICF1 samples and controls by qPCR (upper panel) and western blot (lower panel); C, DNA methylated enrichment assay reporting the hypomethylation at the *hnRNP-LL* regulatory region; D, DNMT3B binding to *hnRNP-LL* regulatory region by ChIP-qPCR; E, RIP-qPCR showing hnRNP-LL binding to *CD45* exons 4 and 6 exclusively in ICF1 samples; F, DNMT3B and hnRNP-LL physically interact in ICF1p2 sample as shown in co-immunoprecipitation experiments; G,H, Control B-LCLs exhibit increased CD45RO expression, with concomitant decrease of CD45RABC expression after 5-AzaC treatment (1uM 5-AzaC for 48h, 72h and 120h); I,J DNA hypomethylation and upregulation of *hnRNP-LL* gene induced by 5-AzaC treatment. P-values in A,B,C,D,E,H and I were calculated with T-student test (two-tails) and adjusted with BH method.

3.5 ICF1-specific DNA methylation defects associate with deregulation of alternative intragenic transcription initiation sites.

The current models describing the relationship between DNA methylation and gene expression report that promoter methylation is associated with gene silencing, and gene body methylation is associated with expression (Ball et al., 2009; Rauch et al., 2009; Aran et al., 2011; Varley et al., 2013). So far, our findings would support a role of DNMT3B methylation activity in regulating the transcription of alternative transcript isoforms from differentially methylated TSS or the alternative exons inclusion in the mature transcript.

Recent studies have proposed that cryptic or alternative promoters, marked by H3K4me3 and detected by the capped analysis of gene expression (CAGE) sequencing, may be characterized by promoter-like methylation in gene bodies (Illingworth et al., 2010; Maunakea et al., 2010; Deaton et al., 2011). To deeper investigate the status of these potential transcription initiations that could be affected by DNMT3B deficient activity, we predicted the annotation of potential i-TSS by integrating several datasets from ENCODE Project (from many different cell types). We overlapped CAGE tags, Polymerase II binding sites and H3K4me3-enriched sites filtering them for intragenic position and excluding the annotated TSS. First, we measured the differential expression of the predicted i-TSS between ICF1 samples and controls by DE-Seq (about 1,251 and 2,793 i-TSS for ICF1p1 and ICF1p2, respectively) and subsequently associated the DE-sites covered by RRBS (156 and 314 for ICF1p1 and ICF1p2, respectively) to differentially methylated CpGs (D> |25%|). Remarkably, we found that about 50-60% of differentially expressed i-TSS subset covered by RRBS was differentially methylated in both ICF1 samples (p-value< 10^{-10} and $8.2x10^{-7}$ for ICF1p1 and ICF1p2, respectively). We found consistent results after analyzing the WGBS, with 80% of putative

i-TSS which was differentially methylated (Fig6A; p-value< 10^{-10} and $8.2x10^{-7}$ for ICF1p1 and ICF1p2, respectively). These results indicate that the activity of putative alternative i-TSS is compromised in the context of DNMT3B-mediated DNA methylation defects. Indeed, CpG island hypomethylation within the exon 5 of *NEURL* gene, corresponding to an intragenic Polymerase II binding site, resulted in the activation of a cryptic H3K4me3-enriched i-TSS in ICF1 samples (Fig6B). The activation of this spurious i-TSS was linked to the expression of an aberrant short isoform including the last three exons of *NEURL* gene (Fig6B).

Furthermore, the large second intron of *CD45* gene including a putative i-TSS was silenced in both ICF1 samples (Fig6C). Consistently, this event associated with CpG hypermethylation and decrease of H3K4me3 mark as indicated by WGBS, RRBS and H3K4me3 ChIP-Seq and confirmed by qPCR (Fig6C; FigS5C,D). Interestingly, DNMT1 binding was observed only in ICF1 samples, indicating that the hypermethylation might be mediated by the catalytic activity of this enzyme (FigS5E). We then looked at the local enrichment of additional histone marks, such as H3K9me3, H3K27me3 and H3K4me2, finding a decrease of permissive chromatin structure and an increase of repressive histone marks (FigS5D). Of note, we found the H3K36me3 increase together with the H3K4me3 decrease, in line with the reported mechanisms of cryptic transcription repression (Hayakawa et al., 2007; Carvalho et al., 2013).

Intriguingly, when we looked at H3K4me3 profile at global level, we observed that it was clearly affected in ICF1 samples, with increased and decreased peaks mostly distributed along the genic features, as 5'-UTR, CpG island, exons, introns and 3'-UTR **(Fig6D,E)**. These changes might influence not only spurious transcription but also several other aspects of the transcriptional process. Indeed, recent findings indicate that trimethylated H3K4 serves to facilitate the competency of pre-mRNA maturation through the bridging of spliceosomal components to H3K4me3 via Chromodomain protein 1 (CHD1).

Through recognizing H3K4me3, CHD1 recruits the complex FACT/PAF1C enhancing the efficiency of pre-mRNA splicing and transcript elongation of Pol II (Sims et al., 2007; Lenasi and Barboric, 2010). Intriguingly, DNMT3B interacted with components of PAF1C complex during hESC differentiation (Rigbolt et al., 2011). Therefore, we evaluated this interaction in B-LCLs finding that only the mutant-DNMT3B co-immunoprecipitated with PAF1 protein **(Fig6F).** This finding, together with the evidence that mutant-DNMT3B binds pre-mRNAs, would imply a functional cross-talk between DNA methylation machinery and transcriptional regulators, which is dependent on ICF1-specific mutations **(Fig6G).**

All together, our studies provide the first evidence in human cells that ICF1-specific mutations of DNMT3B may affect mRNA splicing and the transcriptional activity from alternative i-TSS.

Figure 6

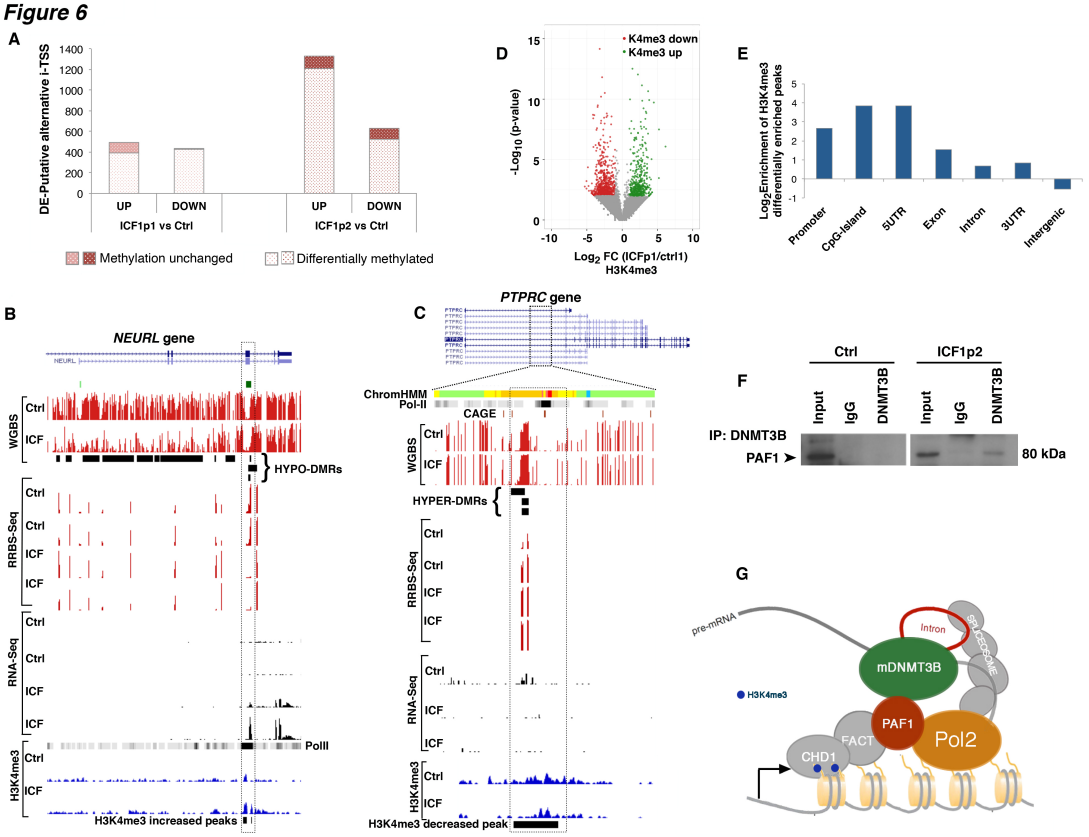


Fig6. Intragenic DNA methylation changes affect the activity of putative alternative transcription initiation sites (i-TSS). A, Histograms showing the methylation status of differentially expressed putative i-TSS (p-value<0.05 from DE-Seq); B, UCSC genome browser screenshot representing putative alternative i-TSS at *NEURL* gene associated to CpG hypomethylation, H3K4me3 increase and transcriptional upregulation as shown by WGBS, RRBS, RNA-Seq and H3K4me3 ChIP-Seq; C, UCSC genome browser screenshot representing the large second intron of *CD45*, including an alternative i-TSS which displays CpG hypermethylation, H3K4me3 decrease and transcriptional downregulation; D, Volcano plot showing H3K4me3 mark redistribution in ICF1p1 sample, with up and down regions; E, Distribution of H3K4me3 increased and decreased regions along the genomic features calculated using the Homer tool); F, Mutant-DNMT3B and PAF1 physically interact in ICF1 samples as shown in co-immunoprecipitation experiments; G, Summarizing model showing the engagement of mutant-DNMT3B in the altered regulation of transcript processing.

Chapter 4

DISCUSSION

The diffusion of next-generation technology opened up several possibilities to study molecular biology phenomena at whole-genome level. The continuous decrease of sequencing costs is making of the NGS a standard tool to investigate biological questions. However, even though the increasing number of sequencing data and of developed tools for each specific analysis, the informative power of whole genome data is still not completely used.

In particular is recommended to include in the analysis pipelines integration steps that are slowly loosing their merely descriptive function acquiring an inferential role. Indeed, the integration of heterogeneous data can cover uncountable gap still present in gene-specific studies.

Our integrative large-scale study demonstrated that ICF1-specific mutations in DNMT3B catalytic domain slightly affect its ability to bind DNA at genomic target regions, while the methyltransferase activity is rather impaired. The fact that mutant-DNMT3B is able to bind DNA is expected, considering that ICF1-specific mutations mainly affect the catalytic domain, while the DNA binding capability depends on the PWWP domain within the N-terminal region (Qiu et al., 2002).

Most DNMT3B binding occur at intragenic positions of transcribed genes, exactly where DNA methylation defects largely take place. This is a novel finding in the context of ICF syndrome studies and explains why previous attempts to identify profound differences in DNA methylation profile at CpG island-associated promoters of deregulated genes failed (Ehrlich et al., 2001; Jin et al., 2007; Gatto et al., 2010). These results paved the way to the identification of unprecedented aspects related to DNMT3B deficient activity, such as the alteration of isoform transcription by modulating the epigenetic

signature at exons and introns or by directly interacting with the pre-mRNA, as observed in the specific case of CD45 gene. Notably, we found deregulation of alternative isoforms at genes important for the ICF1-specific phenotype only after using dedicated tool for RNA-Seq analysis. For instance, we identified transcriptional defects in glycoproteins, members of Rho GTPase family and interleukins involved in the regulation of immunoglobulin production and immunoglobulin mediated immune response, which is the main defect in B lymphocytes, leading to the failure to cope with infections. The affected genes were in large part direct targets of DNMT3B and differentially methylated, suggesting that proper binding and methylating activity of this protein is required for the appropriate relative abundance of transcript isoform. Importantly, we demonstrate a significant correlation between intragenic DNA methylation dependent on DNMT3B and alternative splicing events, especially the selection of 5' and/or 3' alternative donor/acceptor, changing the boundaries of upstream or downstream exons. These results sustain the current view that CpG methylation may modulate the alternative splicing process (Lev Maor et al., 2015).

DNMT3B variants here examined mainly determine intragenic hypomethylation with events of hypermethylation, the extension of which depends on specific mutations. The hypermethylation might reflect the compensatory activity of DNMT3A and/or DNMT1, which are not affected in ICF1 cells. We confirmed this hypothesis at the CD45 intron2 hypermethylation site, while we cannot rule out that other mechanisms are responsible for CpG hypermethylation events or that these are hyper-hydroxymethylation sites, thus implying the TET enzymes involvement. Strikingly, DNMT3B knockout caused CpGs ‘‘hypermodification’’ events in gene bodies, which overlapped with CpGs hypomethylated under DNMT1 depletion, thus revealing an antithetical regulatory interaction between DNMT1, DNMT3B, and the TETs (Tiedemann et al., 2014). Conversely, some DNMT3B target genes were not associated to DMRs, indicating that

DNMT3B deficient activity may result in a more complex epigenetic and transcriptional deregulated network than that dependent on the methyltransferase activity. Consistently, the repressive role of DNMT3B was previously highlighted even in absence of DNA methylation changes (Bachman et al., 2001).

ICF patients' peripheral blood contains only naive B cells presenting an immature phenotype, with an accumulation of bone marrow B-cell emigrants and a lack of memory B and plasma cells. These data indicate a terminal B-cell differentiation blockage in ICF patients at the transitional B-cell stage (Blanco-Betancourt et al., 2004). Consistently, we found that ICF samples exhibited transcriptional defects at genes involved in the B cell receptor signalling pathway, which when altered typically cause the absence of circulating mature B cells and of all immunoglobulin isotypes, accompanied by the accumulation of pre-B cells in the bone marrow (Durandy et al., 2013).

Of note, among the newly identified differentially expressed genes we found either receptor-type PTPs, (e.g. *PTPRC*, *PTPRJ* and *PTPRB*) and nonreceptor-type PTPs (e.g. *PTPN13*). These are known to positively or negatively regulate lymphocyte activation and development, thus providing novel candidate genes contributing to the immune response defects in ICF1 patients. The most striking example was *CD45*, which main role in B and T lymphocytes promotes cells activation. Specifically, CD45 has a synergistic action with PTPRJ (also known as CD148) in B cells. In line with that, mice lacking both CD45 and CD148 have a greater block in B cell development and BCR signaling than mice lacking CD45 or CD148 alone (Rhee and Veillette, 2012). Unlike *PTPRJ*, which is downregulated in ICF1 samples, *PTPRC* is affected by an aberrant skipping of alternative exons 4-6 generating higher amount of the smallest isoform CD45RO in disease samples. Remarkably, CD45 activity is negatively modulated by dimerization (Xu and Weiss, 2002). Larger CD45RABC isoforms exist predominantly as monomeric active phosphatase, while the smaller size of the CD45RO extracellular domain

facilitates dimerization, rendering it less active and increasing the signal transduction threshold. In T or B cells this would contribute to cessation of the primary immune response (Hermiston et al., 2003). In this light, the aberrant exon skipping detected in ICF cells might perturb the CD45 protein activity by increasing the abundance of the CD45RO form. Our findings suggest that DNMT3B deficiency may also affect T-lymphocyte function potentially contributing to ICF1-specific immunodeficiency. For instance, the transcription factor *FOXP1* is an hypomethylated DNMT3B target gene and it is overexpressed in both ICF1 samples. It has been recently shown that Foxp1 is a critical negative regulator of CD4+ follicular helper T cells (TFH cell) differentiation (Wang et al., 2014). Help provided by TFH cells to B cells is essential for the formation of germinal centers (GCs) and to differentiate into memory B cells and plasma cells for the generation of long-lived high-affinity antibodies.

A growing body of evidence correlate chromatin signature to co-transcriptional regulation of alternative splicing (Luco and Misteli, 2011). Epigenetic modifications can affect chromatin structure, which in turn influences the Pol II elongation rate. Alternatively, histone modifications can directly recruit splicing factors to pre-mRNAs via a chromatin-binding protein reading the histone marks. For instance, direct modulation of either H3K4me3 or the CHD1 influences the association of splicing factors with chromatin and the efficiency of pre-mRNA splicing in vivo (Sims et al., 2007). Defects in these processes might explain the alternative exons splicing defects observed in this study not only at *CD45* but also at the other genes, considering the broad intragenic alterations of H3K4me3 detected in disease samples. The direct implication of the mutant-DNMT3B in protein complexes acting in the transcriptional elongation process is suggested by the evidence that it interacts with the known regulator PAF1. Further investigations will be required to dissect the contribution of mutant-DNMT3B to these mechanisms.

This finding together with the evidence that alternative putative i-TSS are illegitimately expressed in ICF1 samples would suggest that a proper DNMT3B-mediated methylation profile is necessary to regulate the spurious initiation of transcription. Indeed, gene body methylation has been associated to repression of cryptic transcription from alternative promoters, retrotransposon elements and/ or antisense transcript in an increasing number of evidence. This would be critical to achieve an efficient elongation of mRNAs during Pol II-mediated transcription (Maunakea et al., 2010; Kulis et al., 2012; Varley et al., 2013). All together, our findings support the concept that perturbation of intragenic CpG methylation significantly associates with alteration of other epigenetic signals, e.g. H3K4me3 and H3K27me3, in ICF1 cells suggesting the presence of a tight cross-talk between these epigenetic marks in modulating the mRNA processing.

From a mechanistic point of view, we were surprised to find no DNA methylation defects at skipped exons of *CD45* gene, given that 5-AzaC treatment increased the exon skipping. However, we clarified the molecular mechanism underlying the illegitimate exon exclusion finding that *hnRNP-LL* was overexpressed as consequence of CpG hypomethylation in ICF1 samples. HnRNP-LL cooperates with hnRNP-L on the *CD45* pre-mRNA, bridging exons 4 and 6 and looping out exon 5, thereby blocking the inclusion of the three exons (Preussner et al., 2012). Depletion or overexpression of hnRNP-LL in B and T cell lines result in reciprocal alteration of CD45RABC and CD45RO isoform expression (Oberdoerffer et al., 2008). In line with this finding, we demonstrated that the derepressed hnRNP-LL interacted to exon 4 and exon 6 of pre-mRNA, thus causing their exclusion from the mature mRNA. Strikingly, mutant-DNMT3B was recruited on the same exons of *CD45* pre-mRNA and physically interacted with hnRNP-LL, suggesting that it is involved in the increased expression of CD45RO isoform.

This novel finding would imply that endogenous DNMT3B is able to interact with mRNA molecules and that ICF1-specific mutations increase this ability.

The RNA binding property of DNMTs was predicted based on several evidences, particularly involving non-coding RNAs (Schmitz et al., 2010; Di Ruscio et al., 2013). We describe for the first time a functional interaction between DNMT3B and pre-mRNA molecules in a disease context, hypothesizing that this aberrant binding prevent a proper sequence of events during alternative exon inclusion. Notably, in one study DNMT3B has been shown able to interact with proteins involved in various aspects of Pol II-mediated transcription (Rigbolt et al., 2011). Whether this mechanism occurs in physiological condition representing a general DNMT3B-mediated mechanism of mRNA processing and alternative exon splicing needs to be further investigated. However, from a disease perspective these findings are important considering the latest evidences functionally implying hnRNP-LL in B-cell to plasma cell differentiation (Chang et al., 2015).

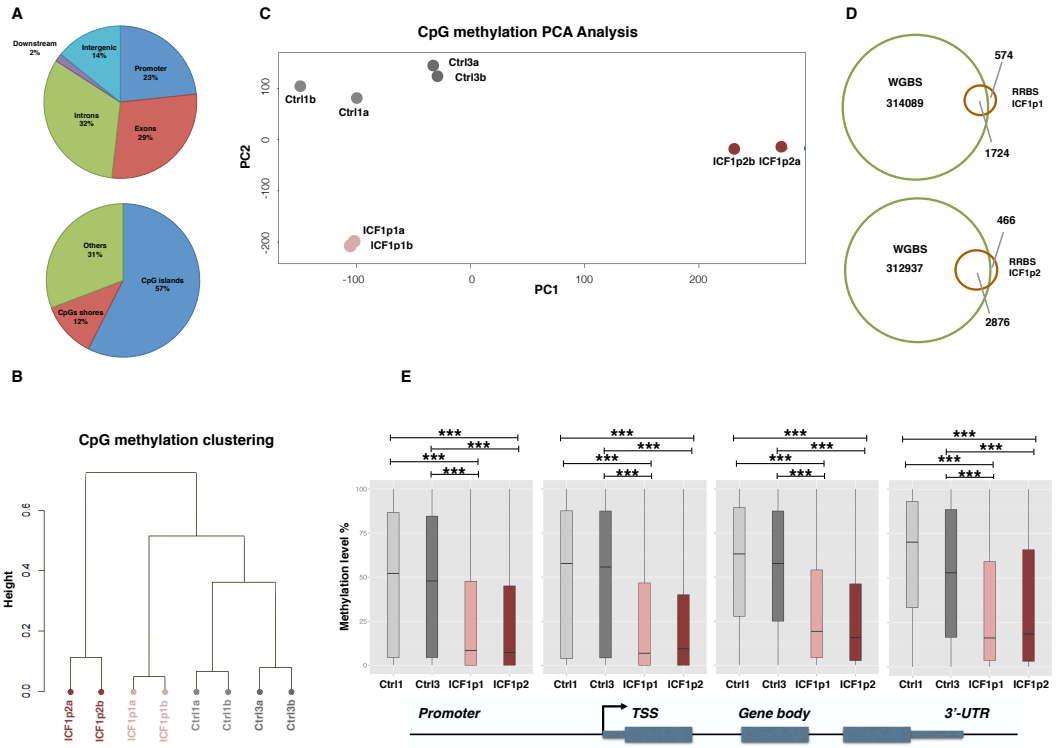
One of the hallmark of ICF mitogen-stimulated lymphocytes and B-LCLs is the chromosome decondensation linked to interphase chromosomal abnormalities and mitotic missegregation of hypomethylated sequences (Gisselsson et al., 2005). Although a direct link between the machinery regulating DNA methylation and mitotic chromosome condensation was previously proposed, no insights were provided to explain ICF1-specific defects (Geiman et al., 2004). For the first time, we identified several candidate genes for these altered processes, which are targeted by DNMT3B and transcriptionally deregulated at isoform level. The mechanism by which these defects were more pronounced at chromosomes with largest pericentromeric domains (e.g. chr1, 16 and 9) in ICF1-specific conditions needs further studies.

The integrative analysis here described provided us with novel insights into DNMT3B-dependent functional cross-talk between DNA methylation and other epigenetic determinants. Remarkably, the genome-wide H3K27 trimethylation profiles demonstrated that the hypo-DMR associated genes were significantly increased in this repressive mark, which might balance the

DNA methylation loss. Consistently, most H3K27me3 increase at hypomethylated regions did not lead to gene expression changes. Nevertheless, H3K27me3 gain associated with events of aberrant gene silencing, as it occurred at *HOXC4* gene. Here, mutant-DNMT3B binding persists in disease cells recruiting components of PRC2 complex, such as Suz12, which in turn might repress the gene through depositing the H3K27me3 mark. Although we do not know whether this silencing is directly caused by H3K27me3 gain, it is interesting to note that H3K27me3 increase upon DNA hypomethylation was also observed in mouse somatic cells associated to events of de novo repression of transcription (Reddington et al., 2013). In line with the mouse models, where DNA hypomethylation is genetically or pharmacologically induced (Brinkman et al., 2012; Reddington et al., 2013), we observed a concomitant decrease of H3K27me3 mark in other regions of the genome, implying a potential dilution of PRC2 away from its normal targets and its retention at hypomethylated regions. Importantly, these finding contribute to clarify the functional cross-talk between DNA methylation and H3K27 methylation in modulating gene silencing in human cells.

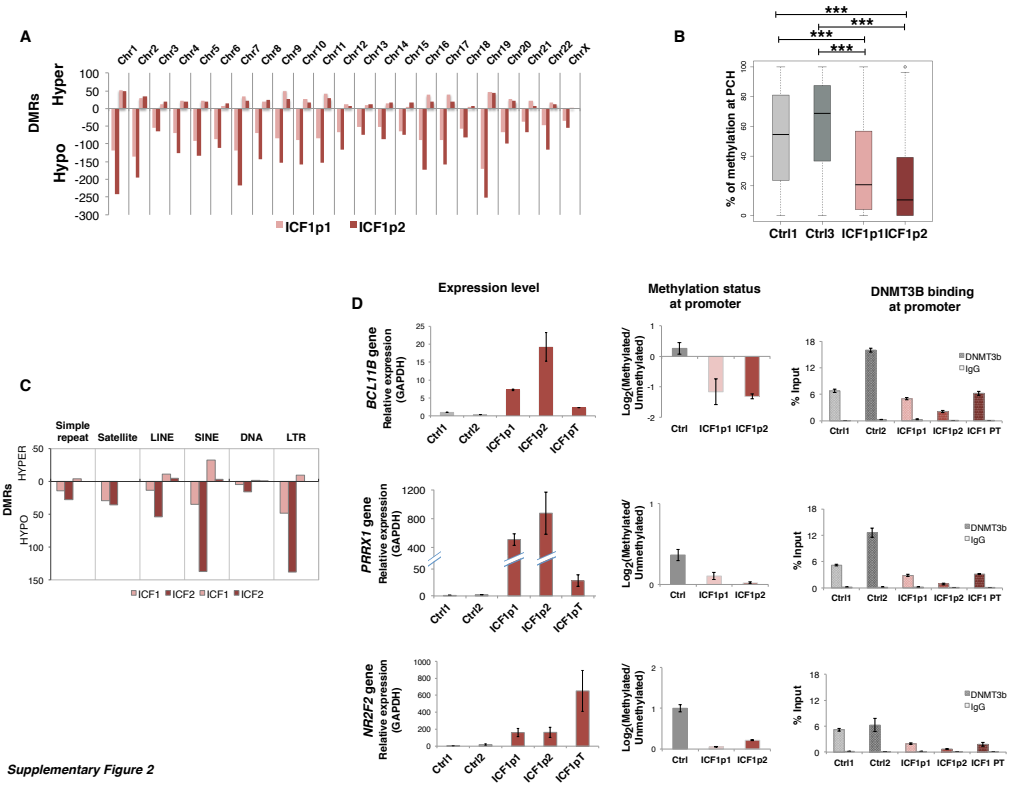
In summary, we have performed a large-scale analysis of the epigenomic and transcriptional profile in ICF1-derived B-LCL. Compared to previous classical transcriptional studies, we identified a new level of DNMT3B-mediated transcriptional regulation, which highlighted defects in gene pathways and functions predicted and long sought, thus representing a step forward towards a more careful characterization of the ICF1 immune phenotype. Furthermore, our study highlights the power of integrative analyses to clarify the intricate regulatory epigenetic network modulating the proper tissue-specific transcriptome.

Supplementary figures



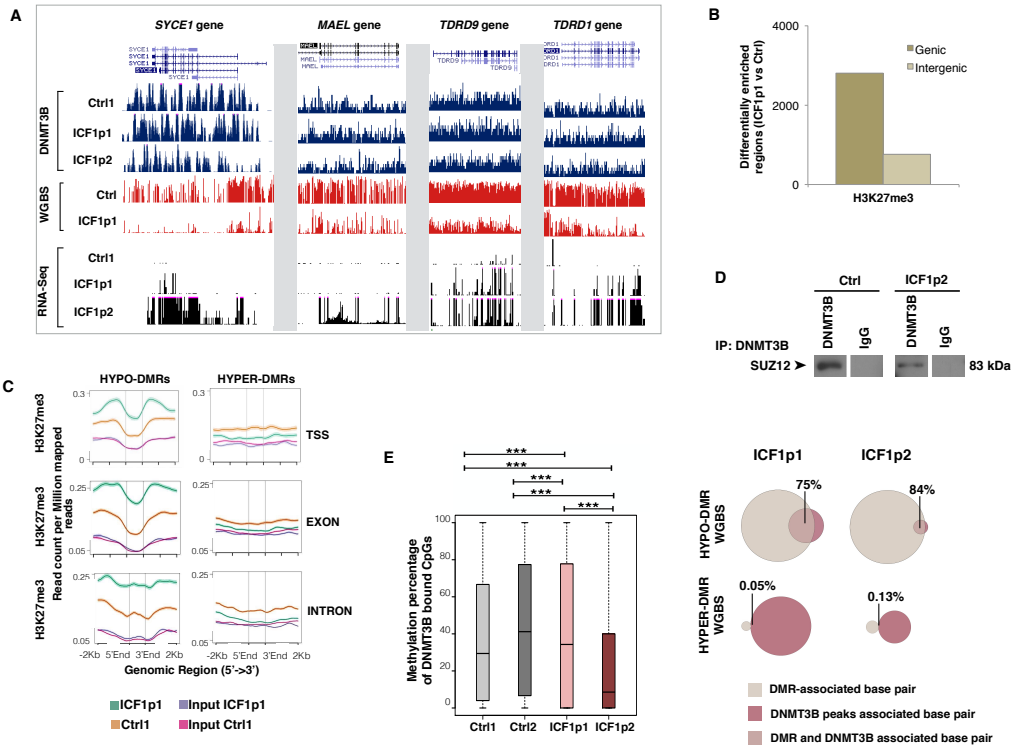
Supplementary Figure 1

FigS1 A, Distributions of covered CpGs (>10X read coverage) by RRBS in promoters, exons, introns, downstream and intergenic regions are shown in the pie chart. B, CpG methylation clustering diagram between samples. Distance method: "correlation", linkage method: "Ward". Samples analyzed, ICF1p1, ICF1p2, Ctrl1 and Ctrl3. C, Principle component analysis (PCA) of mean methylation levels from ICF1 samples and controls. The PCA revealed a clear separation between ICF1 samples and controls by the first two components. As expected the controls are homogeneous in terms of methylome, while ICF1 samples are heterogeneous because of the diverse DNMT3B mutations. Conversely, technical replicates among the samples (a,b) are highly concordant. D, Comparison of CpG-specific DNA methylation data measured by whole-genome bisulfite sequencing [WGBS; (Heyn et al., 2012)] and RRBS (the present study). E, DNA methylation level measured by RRBS in ICF1 and control B-LCLs at CpGs hypomethylated in patients' PBMCs (Infinium 450K Illumina microarray), occurring at promoters, TSS, gene bodies and 3'UTR (p-value was calculated with Kolmogorov-Smirnov method and adjusted with Bonferroni method).



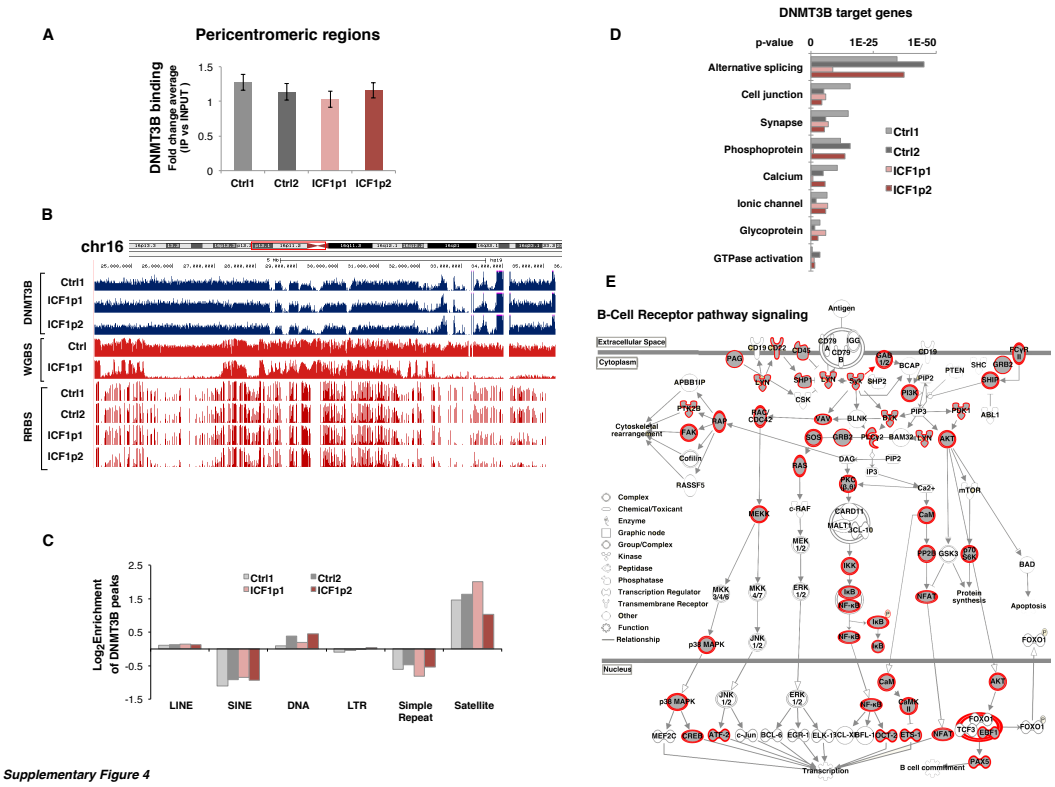
Supplementary Figure 2

FigS2 A, Distributions of hypo- and hyper-DMRs at all chromosomes in ICF1p1 and ICF1p2 compared to controls; B, CpG mehylation level at pericentromeric heterochromatin regions (PCH) in ICF1 and control samples (p-value was calculated with Kolmogorov-Smirnov method and adjusted with Bonferroni method); C, Distribution of hypo- and hyper-DMRs at repetitive sequence families by using the Homer tool (Heinz et al., 2010); D, mRNA expression level (by qPCR), CpG methylation status (by methylated DNA enrichment assay) and DNMT3B binding (by ChIP assay) at homeobox genes and transcription factors identified as differentially expressed by RNA-Seq using DESeq.



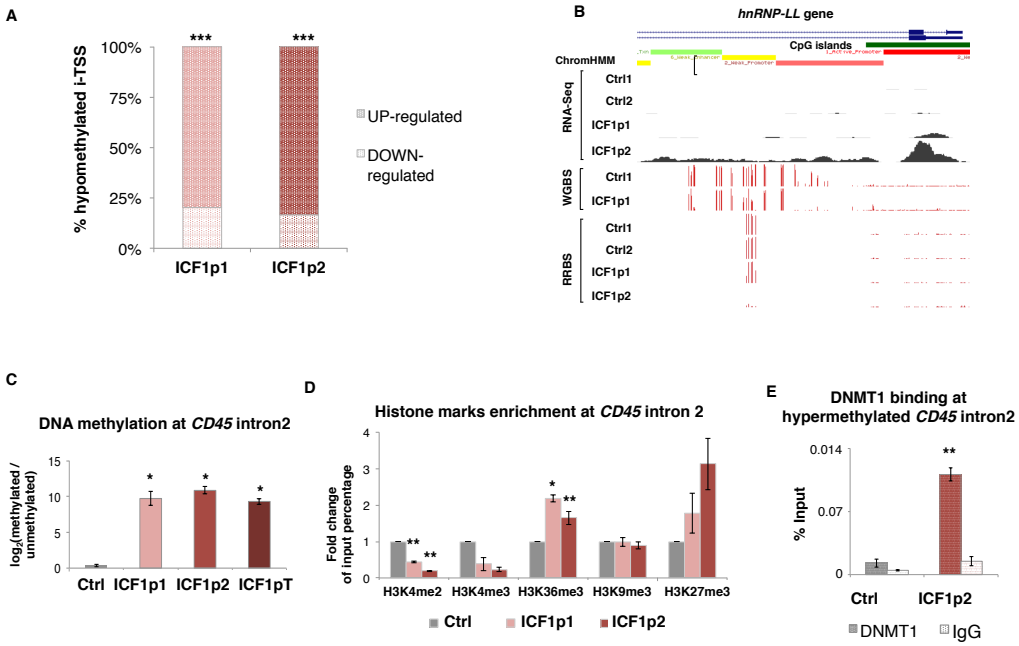
Supplementary Figure 3

FigS3 A, UCSC genome browser screenshots showing DNMT3B binding (ChIP-Seq), CpG methylation (WGBS) and expression profile (RNA-Seq) at "germline genes", previously reported as illegitimately transcribed in ICF1 somatic cells; B, Genomic distribution of H3K27me3 differentially enriched regions in ICF1p1 sample compared to control; C, Density plot of H3K27me3 (read count per million mapped reads) at hypo- and hyper-DMRs (-2kb upstream, +2kb downstream) located at TSS, exons and introns; D, Co-immunoprecipitation performed with anti-DNMT3B and detected with anti-Suz12 in ICF1p2 and control; E, Measurement of the methylation level (percentage) in DNMT3B bound CpGs showing significant reduction in ICF1p1 and ICF1p2 samples (left panel; p-value was calculated with Kolmogorov-Smirnov method and adjusted with Bonferroni method); Venn diagram showing the overlap between DMRs from WGBS and DNMT3B peaks in ICF1 samples (right panel).



Supplementary Figure 4

FigS4 A, DNMT3B binding at pericentromeric heterochromatin regions (PCH) in ICF1 and control samples reported as average of fold change (IP/INPUT) calculated in 200 bins; B, UCSC genome browser screenshot showing DNMT3B, WGBS and RRBS profiles at pericentromeric and centromeric regions of Chromosome 16; C, Distribution of DNMT3B peaks at repetitive sequence families using the Homer tool (Heinz et al., 2010); D, Gene Ontology (DAVID) of DNMT3B target genes (from TSS-2kb to TTS+2kb; p-values corrected by BH method were considered); E, B-cell receptor signaling pathway detected using IPA Ingenuity Pathway analysis. Red highlights indicate DE-isoform associated genes.



Supplementary Figure 5

FigS5 A, Histogram showing that CpG hypomethylation at isoform specific i-TSS mostly associates with their up-regulation (p-values $<<10^{-10}$ were calculated using 50% binomial test one-tail); B, UCSC genome browser screenshot showing hypomethylation, and upregulation of *hnRNP-LL* gene in both ICF1 samples compared to controls. Results from RNA-Seq, WGBS and RRBS are reported; C, CpG methylation level at *CD45* intron2 measured in ICF1 samples and control; D, H3K4me2, H3K4me3, H3K27me3, H3K36me3, H3K9me3 mark enrichment obtained by ChIP assay; E, DNMT1 binding at hypermethylated CD45 intron2 examined by ChIP assay in ICF1p2 and control samples. P-values in D and E were calculated using T-student test (two-tails) adjusted with Benjamini-Hochberg (BH) method.

References

Akalin A, Kormaksson M, Li S, Garrett-Bakelman FE, Figueroa ME, Melnick A, Mason CE. 2012. methylKit: a comprehensive R package for the analysis of genome-wide DNA methylation profiles. Genome Biol 13:R87.

Anastasiadou C, Malousi A, Maglaveras N, Kouidou S. 2011. Human epigenome data reveal increased CpG methylation in alternatively spliced sites and putative exonic splicing enhancers. DNA Cell Biol 30:267-275.

Anders S, Huber W. 2010. Differential expression analysis for sequence count data. Genome Biol 11:R106.

Anders S, Pyl PT, Huber W. 2015. HTSeq--a Python framework to work with high-throughput sequencing data. Bioinformatics 31:166-169.

Angelini C, Costa V. 2014. Understanding gene regulatory mechanisms by integrating ChIP-seq and RNA-seq data: statistical solutions to biological problems. Front Cell Dev Biol 2:51.

Aran D, Toperoff G, Rosenberg M, Hellman A. 2011. Replication timing-related and gene body-specific methylation of active human genes. Hum Mol Genet 20:670-680.

Bachman KE, Rountree MR, Baylin SB. 2001. Dnmt3a and Dnmt3b are transcriptional repressors that exhibit unique localization properties to heterochromatin. J Biol Chem 276:32282-32287.

Ball MP, Li JB, Gao Y, Lee JH, LeProust EM, Park IH, Xie B, Daley GQ, Church GM. 2009. Targeted and genome-scale strategies reveal gene-body methylation signatures in human cells. Nat Biotechnol 27:361-368.

Berger B, Peng J, Singh M. 2013. Computational solutions for omics data. Nat Rev Genet 14:333-346.

Berman BP, Weisenberger DJ, Aman JF, Hinoue T, Ramjan Z, Liu Y, Noushmehr H, Lange CP, van Dijk CM, Tollenaar RA, Van Den Berg D, Laird PW. 2012. Regions of focal DNA hypermethylation and long-range hypomethylation in colorectal cancer coincide with nuclear lamina-associated domains. Nat Genet 44:40-46.

Bird A. 2002. DNA methylation patterns and epigenetic memory. Genes Dev 16:6-21.

Blanco-Betancourt CE, Moncla A, Milili M, Jiang YL, Viegas-Pequignot EM, Roquelaure B, Thuret I, Schiff C. 2004. Defective B-cell-negative selection and terminal differentiation in the ICF syndrome. Blood 103:2683-2690.

Bock C. 2012. Analysing and interpreting DNA methylation data. Nat Rev Genet 13:705-719.

Brinkman AB, Gu H, Bartels SJ, Zhang Y, Matarese F, Simmer F, Marks H, Bock C, Gnirke A, Meissner A, Stunnenberg HG. 2012. Sequential ChIP-bisulfite sequencing enables direct genome-scale investigation of chromatin and DNA methylation cross-talk. Genome Res 22:1128-1138.

Burrows M, Wheeler DJ. 1994. A block sorting lossless data compression algorithm. Technical Report 124,

Digital Equipment Corporation, Palo Alto, California.
Carvalho S, Raposo AC, Martins FB, Grosso AR, Sridhara SC, Rino J, Carmo-Fonseca M, de Almeida SF. 2013. Histone methyltransferase SETD2 coordinates FACT recruitment with nucleosome dynamics during transcription. Nucleic Acids Res 41:2881-2893.
Chang X, Li B, Rao A. 2015. RNA-binding protein hnRNPLL regulates mRNA splicing and stability during B-cell to plasma-cell differentiation. Proc Natl Acad Sci U S A 112:E1888-1897.
Chodavarapu RK, Feng S, Bernatavichute YV, Chen PY, Stroud H, Yu Y, Hetzel JA, Kuo F, Kim J, Cokus SJ, Casero D, Bernal M, Huijser P, Clark AT, Kramer U, Merchant SS, Zhang X, Jacobsen SE, Pellegrini M. 2010. Relationship between nucleosome positioning and DNA methylation. Nature 466:388-392.
Comes S, Gagliardi M, Laprano N, Fico A, Cimmino A, Palamidessi A, De Cesare D, De Falco S, Angelini C, Scita G, Patriarca EJ, Matarazzo MR, Minchiotti G. 2013. L-Proline induces a mesenchymal-like invasive program in embryonic stem cells by remodeling H3K9 and H3K36 methylation. Stem Cell Reports 1:307-321.
Curado J, Iannone C, Tilgner H, Valcarcel J, Guigo R. 2015. Promoter-like epigenetic signatures in exons displaying cell type-specific splicing. Genome Biol 16:236.
Deaton AM, Webb S, Kerr AR, Illingworth RS, Guy J, Andrews R, Bird A. 2011. Cell type-specific DNA methylation at intragenic CpG islands in the immune system. Genome Res 21:1074-1086.
Di Ruscio A, Ebralidze AK, Benoukraf T, Amabile G, Goff LA, Terragni J, Figueroa ME, De Figueiredo Pontes LL, Alberich-Jorda M, Zhang P, Wu M, D'Alo F, Melnick A, Leone G, Ebralidze KK, Pradhan S, Rinn JL, Tenen DG. 2013. DNMT1-interacting RNAs block gene-specific DNA methylation. Nature 503:371-376.
Dillies MA, Rau A, Aubert J, Hennequet-Antier C, Jeanmougin M, Servant N, Keime C, Marot G, Castel D, Estelle J, Guernec G, Jagla B, Jouneau L, Laloe D, Le Gall C, Schaeffer B, Le Crom S, Guedj M, Jaffrezic F, French StatOmique C. 2013. A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. Brief Bioinform 14:671-683.
Durandy A, Kracker S, Fischer A. 2013. Primary antibody deficiencies. Nat Rev Immunol 13:519-533.
Ehrlich M, Buchanan KL, Tsien F, Jiang G, Sun B, Uicker W, Weemaes CM, Smeets D, Sperling K, Belohradsky BH, Tommerup N, Misek DE, Rouillard JM, Kuick R, Hanash SM. 2001. DNA methyltransferase 3B mutations linked to the ICF syndrome cause dysregulation of lymphogenesis genes. Hum Mol Genet 10:2917-2931.
Feng J, Liu T, Qin B, Zhang Y, Liu XS. 2012. Identifying ChIP-seq enrichment using MACS. Nat Protoc 7:1728-1740.
Ferragina P, Manzini G. 2005. Indexing compressed text. Journal of the Acm 52:552-581.

Gatto S, Della Ragione F, Cimmino A, Strazzullo M, Fabbri M, Mutarelli M, Ferraro L, Weisz A, D'Esposito M, Matarazzo MR. 2010. Epigenetic alteration of microRNAs in DNMT3B-mutated patients of ICF syndrome. Epigenetics 5:427-443.

Geiman TM, Sankpal UT, Robertson AK, Zhao Y, Zhao Y, Robertson KD. 2004. DNMT3B interacts with hSNF2H chromatin remodeling enzyme, HDACs 1 and 2, and components of the histone methylation system. Biochem Biophys Res Commun 318:544-555.

Gelfman S, Cohen N, Yearim A, Ast G. 2013. DNA-methylation effect on cotranscriptional splicing is dependent on GC architecture of the exon-intron structure. Genome Res 23:789-799.

Gisselsson D, Shao C, Tuck-Muller CM, Sogorovic S, Palsson E, Smeets D, Ehrlich M. 2005. Interphase chromosomal abnormalities and mitotic missegregation of hypomethylated sequences in ICF syndrome cells. Chromosoma 114:118-126.

Glaus P, Honkela A, Rattray M. 2012. Identifying differentially expressed transcripts from RNA-seq data with biological variation. Bioinformatics 28:1721-1728.

Goll MG, Bestor TH. 2005. Eukaryotic cytosine methyltransferases. Annu Rev Biochem 74:481-514.

Hagleitner MM, Lankester A, Maraschio P, Hulten M, Fryns JP, Schuetz C, Gimelli G, Davies EG, Gennery A, Belohradsky BH, de Groot R, Gerritsen EJ, Mattina T, Howard PJ, Fasth A, Reisli I, Furthner D, Slatter MA, Cant AJ, Cazzola G, van Dijken PJ, van Deuren M, de Greef JC, van der Maarel SM, Weemaes CM. 2008. Clinical spectrum of immunodeficiency, centromeric instability and facial dysmorphism (ICF syndrome). J Med Genet 45:93-99.

Hawkins RD, Hon GC, Ren B. 2010. Next-generation genomics: an integrative approach. Nat Rev Genet 11:476-486.

Hayakawa T, Ohtani Y, Hayakawa N, Shinmyozu K, Saito M, Ishikawa F, Nakayama J. 2007. RBP2 is an MRG15 complex component and down-regulates intragenic histone H3 lysine 4 methylation. Genes Cells 12:811-826.

Heinz S, Benner C, Spann N, Bertolino E, Lin YC, Laslo P, Cheng JX, Murre C, Singh H, Glass CK. 2010. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. Mol Cell 38:576-589.

Hermiston ML, Xu Z, Weiss A. 2003. CD45: a critical regulator of signaling thresholds in immune cells. Annu Rev Immunol 21:107-137.

Heyn H, Vidal E, Sayols S, Sanchez-Mut JV, Moran S, Medina I, Sandoval J, Simo-Riudalbas L, Szczesna K, Huertas D, Gatto S, Matarazzo MR, Dopazo J, Esteller M. 2012. Whole-genome bisulfite DNA sequencing of a DNMT3B mutant patient. Epigenetics 7:542-550.

Holz-Schietinger C, Reich NO. 2012. RNA modulation of the human DNA methyltransferase 3A. Nucleic Acids Res 40:8550-8557.

Huang da W, Sherman BT, Lempicki RA. 2009. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc 4:44-57.

Huang K, Wu Z, Liu Z, Hu G, Yu J, Chang KH, Kim KP, Le T, Faull KF, Rao N, Gennery A, Xue Z, Wang CY, Pellegrini M, Fan G. 2014. Selective demethylation and altered gene expression are associated with ICF syndrome in human-induced pluripotent stem cells and mesenchymal stem cells. Hum Mol Genet 23:6448-6457.

Hurd PJ, Nelson CJ. 2009. Advantages of next-generation sequencing versus the microarray in epigenetic research. Brief Funct Genomic Proteomic 8:174-183.

Illingworth RS, Gruenewald-Schneider U, Webb S, Kerr AR, James KD, Turner DJ, Smith C, Harrison DJ, Andrews R, Bird AP. 2010. Orphan CpG islands identify numerous conserved promoters in the mammalian genome. PLoS Genet 6:e1001134.

Jeanpierre M, Turleau C, Aurias A, Prieur M, Ledeist F, Fischer A, Viegas-Pequignot E. 1993. An embryonic-like methylation pattern of classical satellite DNA is observed in ICF syndrome. Hum Mol Genet 2:731-735.

Jeffery L, Nakielny S. 2004. Components of the DNA methylation system of chromatin control are RNA-binding proteins. J Biol Chem 279:49479-49487.

Jin B, Tao Q, Peng J, Soo HM, Wu W, Ying J, Fields CR, Delmas AL, Liu X, Qiu J, Robertson KD. 2007. DNA methyltransferase 3B (DNMT3B) mutations in ICF syndrome lead to altered epigenetic modifications and aberrant expression of genes regulating development, neurogenesis, and immune function. Hum Mol Genet.

Jones PA. 2012. Functions of DNA methylation: islands, start sites, gene bodies and beyond. Nat Rev Genet 13:484-492.

Kahn SD. 2011. On the future of genomic data. Science 331:728-729.

Kharchenko PV, Tolstorukov MY, Park PJ. 2008. Design and analysis of ChIP-seq experiments for DNA-binding proteins. Nat Biotechnol 26:1351-1359.

Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL. 2013. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol 14:R36.

Krueger F, Andrews SR. 2011. Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. Bioinformatics 27:1571-1572.

Kulis M, Heath S, Bibikova M, Queiros AC, Navarro A, Clot G, Martinez-Trillos A, Castellano G, Brun-Heath I, Pinyol M, Barberan-Soler S, Papasaikas P, Jares P, Bea S, Rico D, Ecker S, Rubio M, Royo R, Ho V, Klotzle B, Hernandez L, Conde L, Lopez-Guerra M, Colomer D, Villamor N, Aymerich M, Rozman M, Bayes M, Gut M, Gelpi JL, Orozco M, Fan JB, Quesada V, Puente XS, Pisano DG, Valencia A, Lopez-Guillermo A, Gut I, Lopez-Otin C, Campo E, Martin-Subero JI. 2012. Epigenomic analysis detects widespread gene-body DNA hypomethylation in chronic lymphocytic leukemia. Nat Genet 44:1236-1242.

Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. Nature Methods 9:357-U354.

Lenasi T, Barboric M. 2010. P-TEFb stimulates transcription elongation and pre-mRNA splicing through multilateral mechanisms. RNA Biol 7:145-150.
Lev Maor G, Yearim A, Ast G. 2015. The alternative role of DNA methylation in splicing regulation. Trends Genet 31:274-280.
Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25:1754-1760.
Luco RF, Misteli T. 2011. More than a splicing code: integrating the role of RNA, chromatin and non-coding RNA in alternative splicing regulation. Curr Opin Genet Dev 21:366-372.
Lund E, Oldenburg AR, Collas P. 2014. Enriched domain detector: a program for detection of wide genomic enrichment domains robust against local variations. Nucleic Acids Res 42:e92.
Manber U, Myers G. 1990. Suffix Arrays: A New Method for On-Line String Searches. Proceeding

SODA '90 Proceedings of the first annual ACM-SIAM symposium on Discrete algorithms.
Manzini G. 2001. An analysis of the Burrows-Wheeler Transform. Journal of the Acm 48:407-430.
Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, Berka J, Braverman MS, Chen YJ, Chen Z, Dewell SB, Du L, Fierro JM, Gomes XV, Godwin BC, He W, Helgesen S, Ho CH, Irzyk GP, Jando SC, Alenquer ML, Jarvie TP, Jirage KB, Kim JB, Knight JR, Lanza JR, Leamon JH, Lefkowitz SM, Lei M, Li J, Lohman KL, Lu H, Makhijani VB, McDade KE, McKenna MP, Myers EW, Nickerson E, Nobile JR, Plant R, Puc BP, Ronan MT, Roth GT, Sarkis GJ, Simons JF, Simpson JW, Srinivasan M, Tartaro KR, Tomasz A, Vogt KA, Volkmer GA, Wang SH, Wang Y, Weiner MP, Yu P, Begley RF, Rothberg JM. 2005. Genome sequencing in microfabricated high-density picolitre reactors. Nature 437:376-380.
Marx V. 2013. Biology: The big challenges of big data. Nature 498:255-260.
Matarazzo MR, Boyle S, D'Esposito M, Bickmore WA. 2007. Chromosome territory reorganization in a human disease with altered DNA methylation. Proc Natl Acad Sci U S A 104:16546-16551.
Matarazzo MR, De Bonis ML, Vacca M, Della Ragione F, D'Esposito M. 2009. Lessons from two human chromatin diseases, ICF syndrome and Rett syndrome. Int J Biochem Cell Biol 41:117-126.
Matarazzo MR, Lembo F, Angrisano T, Ballestar E, Ferraro M, Pero R, De Bonis ML, Bruni CB, Esteller M, D'Esposito M, Chiariotti L. 2004. In vivo analysis of DNA methylation patterns recognized by specific proteins: coupling CHIP and bisulfite analysis. Biotechniques 37:666-668, 670, 672-663.
Maunakea AK, Nagarajan RP, Bilenky M, Ballinger TJ, D'Souza C, Fouse SD, Johnson BE, Hong C, Nielsen C, Zhao Y, Turecki G, Delaney A, Varhol R, Thiessen N, Shchors K, Heine VM, Rowitch DH, Xing X, Fiore C, Schillebeeckx M, Jones SJ, Haussler D, Marra MA, Hirst M, Wang T, Costello JF. 2010. Conserved role of intragenic DNA methylation in regulating alternative promoters. Nature 466:253-257.

Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. 2008. Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat Methods 5:621-628.

Nagalakshmi U, Wang Z, Waern K, Shou C, Raha D, Gerstein M, Snyder M. 2008. The transcriptional landscape of the yeast genome defined by RNA sequencing. Science 320:1344-1349.

Nelson MR. 1996. Data compression with the Burrows Wheeler transform - Transformation opens the door to new data-compression techniques. Dr Dobbs Journal 21:46-+.

Neph S, Kuehn MS, Reynolds AP, Haugen E, Thurman RE, Johnson AK, Rynes E, Maurano MT, Vierstra J, Thomas S, Sandstrom R, Humbert R, Stamatoyannopoulos JA. 2012. BEDOPS: high-performance genomic feature operations. Bioinformatics 28:1919-1920.

Oberdoerffer S, Moita LF, Neems D, Freitas RP, Hacohen N, Rao A. 2008. Regulation of CD45 alternative splicing by heterogeneous ribonucleoprotein, hnRNPLL. Science 321:686-691.

Okano M, Bell DW, Haber DA, Li E. 1999. DNA methyltransferases Dnmt3a and Dnmt3b are essential for de novo methylation and mammalian development. Cell 99:247-257.

Park PJ. 2009. ChIP-seq: advantages and challenges of a maturing technology. Nat Rev Genet 10:669-680.

Preussner M, Schreiner S, Hung LH, Porstner M, Jack HM, Benes V, Ratsch G, Bindereif A. 2012. HnRNP L and L-like cooperate in multiple-exon regulation of CD45 alternative splicing. Nucleic Acids Res 40:5666-5678.

Qiu C, Sawada K, Zhang X, Cheng X. 2002. The PWWP domain of mammalian DNA methyltransferase Dnmt3b defines a new family of DNA-binding folds. Nat Struct Biol 9:217-224.

Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26:841-842.

Rabiner LR. 1989. A tutorial on hidden Markov models and selected applications in speech recognition. AT&T Bell Lab., Murray Hill, NJ, USA

.

Rauch TA, Wu X, Zhong X, Riggs AD, Pfeifer GP. 2009. A human B cell methylome at 100-base pair resolution. Proc Natl Acad Sci U S A 106:671-678.

Reddington JP, Perricone SM, Nestor CE, Reichmann J, Youngson NA, Suzuki M, Reinhardt D, Dunican DS, Prendergast JG, Mjoseng H, Ramsahoye BH, Whitelaw E, Greally JM, Adams IR, Bickmore WA, Meehan RR. 2013. Redistribution of H3K27me3 upon DNA hypomethylation results in de-repression of Polycomb target genes. Genome Biol 14:R25.

Rhee I, Veillette A. 2012. Protein tyrosine phosphatases in lymphocyte activation and autoimmunity. Nat Immunol 13:439-447.

Rigbolt KT, Prokhorova TA, Akimov V, Henningsen J, Johansen PT, Kratchmarova I, Kassem M, Mann M, Olsen JV, Blagoev B. 2011. System-wide temporal characterization of the proteome and phosphoproteome of human embryonic stem cell differentiation. Sci Signal 4:rs3.

Schmitz KM, Mayer C, Postepska A, Grummt I. 2010. Interaction of noncoding RNA with the rDNA promoter mediates recruitment of DNMT3b and silencing of rRNA genes. Genes Dev 24:2264-2269.

Shen L, Shao N, Liu X, Nestler E. 2014. ngs.plot: Quick mining and visualization of next-generation sequencing data by integrating genomic databases. BMC Genomics 15:284.

Shukla S, Kavak E, Gregory M, Imashimizu M, Shutinoski B, Kashlev M, Oberdoerffer P, Sandberg R, Oberdoerffer S. 2011. CTCF-promoted RNA polymerase II pausing links DNA methylation to splicing. Nature 479:74-79.

Simo-Riudalbas L, Diaz-Lagares A, Gatto S, Gagliardi M, Crujeiras AB, Matarazzo MR, Esteller M, Sandoval J. 2015. Genome-Wide DNA Methylation Analysis Identifies Novel Hypomethylated Non-Pericentromeric Genes with Potential Clinical Implications in ICF Syndrome. PLoS One 10:e0132517.

Sims RJ, 3rd, Millhouse S, Chen CF, Lewis BA, Erdjument-Bromage H, Tempst P, Manley JL, Reinberg D. 2007. Recognition of trimethylated histone H3 lysine 4 facilitates the recruitment of transcription postinitiation factors and pre-mRNA splicing. Mol Cell 28:665-676.

Smith ZD, Gu H, Bock C, Gnirke A, Meissner A. 2009. High-throughput bisulfite sequencing in mammalian genomes. Methods 48:226-232.

Tiedemann RL, Putiri EL, Lee JH, Hlady RA, Kashiwagi K, Ordog T, Zhang Z, Liu C, Choi JH, Robertson KD. 2014. Acute depletion redefines the division of labor among DNA methyltransferases in methylating the human genome. Cell Rep 9:1554-1566.

Ueda Y, Okano M, Williams C, Chen T, Georgopoulos K, Li E. 2006. Roles for Dnmt3b in mammalian development: a mouse model for the ICF syndrome. Development 133:1183-1192.

Varley KE, Gertz J, Bowling KM, Parker SL, Reddy TE, Pauli-Behn F, Cross MK, Williams BA, Stamatoyannopoulos JA, Crawford GE, Absher DM, Wold BJ, Myers RM. 2013. Dynamic DNA methylation across diverse human cell lines and tissues. Genome Res 23:555-567.

Velasco G, Hube F, Rollin J, Neuillet D, Philippe C, Bouzinba-Segard H, Galvani A, Viegas-Pequignot E, Francastel C. 2010. Dnmt3b recruitment through E2F6 transcriptional repressor mediates germ-line gene silencing in murine somatic tissues. Proc Natl Acad Sci U S A 107:9281-9286.

Velasco G, Walton EL, Sterlin D, Hedouin S, Nitta H, Ito Y, Fouyssac F, Megarbane A, Sasaki H, Picard C, Francastel C. 2014. Germline genes hypomethylation and expression define a molecular signature in peripheral blood of ICF patients: implications for diagnosis and etiology. Orphanet J Rare Dis 9:56.

Vire E, Brenner C, Deplus R, Blanchon L, Fraga M, Didelot C, Morey L, Van Eynde A, Bernard D, Vanderwinden JM, Bollen M, Esteller M, Di Croce L, de Launoit Y, Fuks F. 2006. The Polycomb group protein EZH2 directly controls DNA methylation. Nature 439:871-874.

Wang H, Geng J, Wen X, Bi E, Kossenkov AV, Wolf AI, Tas J, Choi YS, Takata H, Day TJ, Chang LY, Sprout SL, Becker EK, Willen J, Tian L, Wang X, Xiao C,

Jiang P, Crotty S, Victora GD, Showe LC, Tucker HO, Erikson J, Hu H. 2014. The transcription factor Foxp1 is a critical negative regulator of the differentiation of follicular helper T cells. Nat Immunol 15:667-675.

Weemaes CM, van Tol MJ, Wang J, van Ostaijen-ten Dam MM, van Eggermond MC, Thijssen PE, Aytekin C, Brunetti-Pierri N, van der Burg M, Graham Davies E, Ferster A, Furthner D, Gimelli G, Gennery A, Kloeckener-Gruissem B, Meyn S, Powell C, Reisli I, Schuetz C, Schulz A, Shugar A, van den Elsen PJ, van der Maarel SM. 2013. Heterogeneous clinical presentation in ICF syndrome: correlation with underlying gene defects. Eur J Hum Genet 21:1219-1225.

Wolff EM, Byun HM, Han HF, Sharma S, Nichols PW, Siegmund KD, Yang AS, Jones PA, Liang G. 2010. Hypomethylation of a LINE-1 promoter activates an alternate transcript of the MET oncogene in bladders with cancer. PLoS Genet 6:e1000917.

Xu Z, Weiss A. 2002. Negative regulation of CD45 by differential homodimerization of the alternatively spliced isoforms. Nat Immunol 3:764-771.

Yearim A, Gelfman S, Shayevitch R, Melcer S, Glaich O, Mallm JP, Nissim-Rafinia M, Cohen AH, Rippe K, Meshorer E, Ast G. 2015. HP1 Is Involved in Regulating the Global Impact of DNA Methylation on Alternative Splicing. Cell Rep 10:1122-1134.

Yehezkel S, Segev Y, Viegas-Pequignot E, Skorecki K, Selig S. 2008. Hypomethylation of subtelomeric regions in ICF syndrome is associated with abnormally short telomeres and enhanced transcription from telomeric regions. Hum Mol Genet 17:2776-2789.

Zang C, Schones DE, Zeng C, Cui K, Zhao K, Peng W. 2009. A clustering approach for identification of enriched domains from histone modification ChIP-Seq data. Bioinformatics 25:1952-1958.

Zhang ZH, Jhaveri DJ, Marshall VM, Bauer DC, Edson J, Narayanan RK, Robinson GJ, Lundberg AE, Bartlett PF, Wray NR, Zhao QY. 2014. A comparative study of techniques for differential expression analysis on RNA-Seq data. PLoS One 9:e103207.

Zhao S, Zhang B. 2015. A comprehensive evaluation of ensembl, RefSeq, and UCSC annotations in the context of RNA-seq read mapping and gene quantification. BMC Genomics 16:97.

Zhou HL, Luo G, Wise JA, Lou H. 2014. Regulation of alternative splicing by local histone modifications: potential roles for RNA-guided mechanisms. Nucleic Acids Res 42:701-713.

Zhu LJ, Gazin C, Lawson ND, Pages H, Lin SM, Lapointe DS, Green MR. 2010. ChIPpeakAnno: a Bioconductor package to annotate ChIP-seq and ChIP-chip data. BMC Bioinformatics 11:237.