Università degli studi di Napoli "Federico II"

PhD in Computational Biology and Bioinformatics
XXV cycle

"Integrated bioinformatics analysis of epigenomic and transcriptomic data from ICF syndrome patient's cells"

Tutor:
Dr. Maurizio D'Esposito

PhD candidate:
Sole Gatto

Co-tutor:
Prof. Sandro Banfi

Academic year 2012/2013

Index

"Integrated bioinformatics analysis of epigenomic and transcriptomic data from ICF syndrome patient's cells"

Aknowledgments

During this period of my formation many people have been important for my scientific growth and for the improvement of my research skills. Foremost, I would like to thank my tutor, Dr. Maurizio D'Esposito for his encouragement, insightful comments, and continuous presence. I would also like to express my sincere gratitude to Dr. Maria Matarazzo, for the continuous support of my Ph.D study and research, for her patience, motivation, enthusiasm, and knowledge. Her guidance helped me in all the time of research and writing of this thesis. Moreover, I would like to thank Dr. Claudia Angelini for her fundamental support and teaching that introduced me to the computational biology field.

Over the years, I also have enjoyed the aid of several fellowships, which have supported me while I completed my PhD. I received a short-term mobility fellowship by the CNR in 2011, a summer fellowship from FEBS in 2011 and a short-term fellowship from EMBO in 2011-2012. These fellowships allowed me to visit Dr. Sarah Teichmann's lab at the MRC-LMB in Cambrige, UK, for three months and Dr. Hendrik Stunnenberg's lab at the NCMLS in Nijmegen, NL for five months.

I would like to thank also my lab colleagues, Dr. Maria Strazzullo, Romina Francioso, Miriam Gagliardi, Sylwia Leppert, Dr. Floriana Della Ragione, Eva Csukonyi and Arianna Brancaccio.

Finally, I couldn't have gone through this difficult path of the PhD without my family. I will never thank enough my parents, who never stopped encouraging me with their high expectations and, very importantly, my beloved husband, Antonio, who always staid by my side and got to learn every day better the meaning of the word "patience".

1. Preface and aim of the thesis

"The concept of epigenetics includes those heritable changes that do not involve an alteration of the genome at the level of nucleotide sequences" (Guil and Esteller 2009).

The credit for coining the term epigenetics in 1942 goes to Conrad Waddington (1905-1975). He defines it as "the branch of biology which studies the casual interactions between genes and their products, which bring the phenotype into being".

Many are the fields in which a key role for epigenetics has been proved in the last years, spanning from cancer biology, personalized health care and drug response, embryogenesis, behavioral studies, environmental effects on human health, biological processes like imprinting, X chromosome inactivation and the definition and maintenance of cell identity.

The main actors in the epigenetic regulation of cell functions are DNA methylation, histone modifications, non-coding RNAs (lincRNA) and the tridimensional structure of the chromatin in the nucleus. All of those factors contribute to the gene expression regulation, activating and repressing it in specific temporal windows of the cell life and in response to specific stimuli. Consequently a disruption of this regulation can cause as much damage as a single gene mutation, but with less distinctive and identifiable pathways. Moreover, a single mutation in a gene that codify for one of the epigenetic regulators can cause extensive damage in the cell, because it has more than one target. These genetic pathologies are called "chromatin diseases" as the whole chromatin structure is disrupted.

Until about 10 years ago all the studies on those epigenetic marks have been performed with genomic regions-specific techniques, like bisulfite conversion of DNA and single molecule sequencing, to identify non-converted sites marking DNA methylation, or ChIP (Chromatin Immuno-Precipitation) coupled to real-time PCR (Polymerase Chain Reaction) or microarrays (ChIP on chip), to identify single or multiple binding sites for transcription regulators or histone modifications. On the other side, real-time PCR and, more recently, microarrays, have been the sole techniques

supporting us in the study of the effects of those variations on the regulation of gene expression. All those methods have a characteristic in common, that is the need to be target-specific. Even the microarrays, despite allowing the study of the expression of all the annotated genes, show their intrinsic limit in the need of knowing the sequence of the genes themselves.

The recent development of the Next Generation Sequencing (NGS) helped the biologists interested to epigenetics to overcome the problems of addressing the strong theories at the basis of their experiments and of limiting their action field on single, specific targets. The multiple applications of these technologies marked the beginning of the "-omic" era, including the epigenomic one.

NGS is a general term for describing a set of different techniques with different aims. The basis of the system (translated in different chemistries from the bunch of companies competing the market) involves the sequencing from scratch of any DNA or RNA sequence in massive scale and in surprisingly short time, compared to the older technologies.

The applications range from transcriptome analysis to ChIP-sequencing, from Single Nucleotide Polimorphism (SNP) genotyping and Copy Number Variations (CNV) analysis to whole-genome sequencing. All these applications will be described later in the introduction on NGS systems.

It is clear, then, how this technology boosts the chances for new discoveries in the epigenetics field, but, as all novelties, creates a new, even more challenging problem, for the biologists involved: the data analysis. All the DNA sequences produced by those machines need to be assembled, mapped, compared, reconstructed, therefore analyzed. This process involves more than a simple desktop computer, and the competences needed to face it span the computational and the statistics areas, not fully covered by the classic biologic formation. This novel necessities led to new exciting collaborations among the different areas of studies and to the birth of new professional figures, the computational biologists, who, forming their competences in the different fields of biology, statistics and computational sciences, try to integrate the knowledge to the purpose of facing the -omics era.

During my PhD I studied the epigenomic alterations occurring in ICF patients (Immunodeficency, Centromere instability and Facial anomalies) derived lymphoblastoid cell lines. These studies put me in front of a new challenge, which is the data analysis and interpretation. ICF cells are particularly interesting because they are mutated in the de novo DNA methyltransferase 3B gene (DNMT3B) and show a peculiar pattern of hypomethylation only in pericentromeric chromatin. The literature available until now was focused on specific portions of the genome, targeted by specific changes in expression. These studies led us to understand that big changes not only happen in the DNA methylation, but that the interplay between this epigenetic mark and histone modifications was altered. Thanks to the NGS technologies we could then perform whole-genome studies of histone marks enrichment linked to gene expression (H3K4me3 and H3K27me3) and to highly repetitive sequences (H3K9me3). Moreover, we could integrate this information with gene expression through RNA-seq and to DNA methylation with bisulphite-seq (performed by our collaborators). At the same time, the analysis of these data marked for me the opportunity to develop new and appropriate pipelines for the analysis. I will show in this work how I analyzed and integrated microRNA expression (from microarrays), ChIP-sequencing, RNA-sequencing and bisulphite-seq data from SOLiD and Illumina platforms.

2. Introduction

Next generation sequencing

DNA sequencing technologies help biologists in a broad range of applications such as molecular cloning, breeding, finding pathogenic genes, studying of gene regulations and comparative and evolutionary studies. DNA sequencing technologies ideally should be fast, accurate, easy-to-operate, and cheap. In the past thirty years, DNA sequencing technologies and applications have undergone tremendous development and act as the engine of the genome era which is characterized by vast amount of genome data and subsequently broad range of research areas. It is necessary to look back on the history of sequencing technology development to understand the utility and innovation of NGS systems (454, GA/HiSeq, and SOLiD) and to discuss the various applications (Liu, Li et al. 2012).

Since DNA discovery and characterization between the end of the XIX century and the beginning of the XX (Church 1984) many drastic improvements have been done in the DNA sequencing field (a simplified roadmap is tracked in Fig 1).

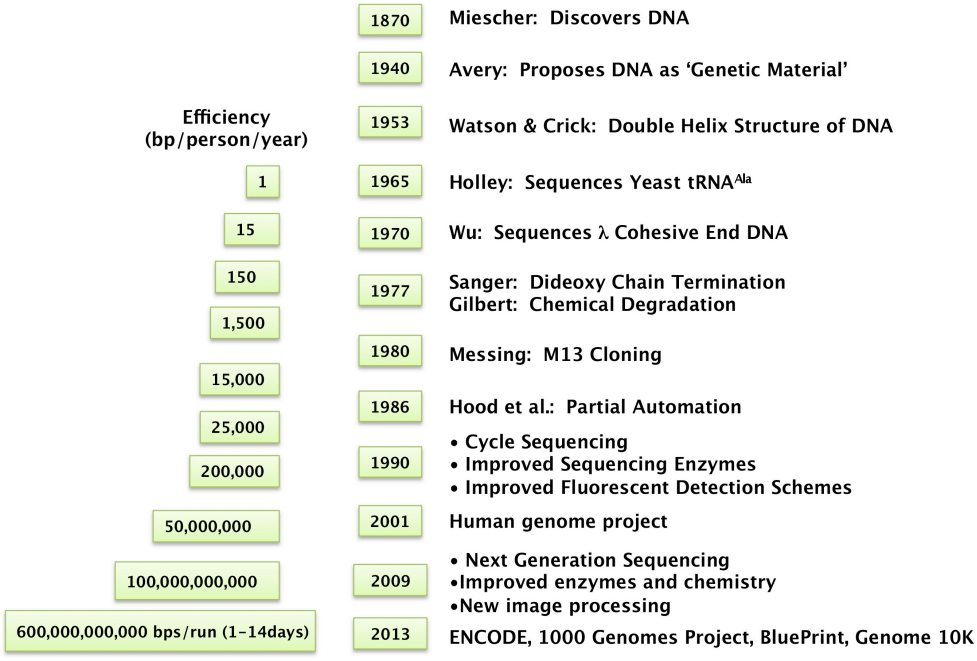

Figure 1. Brief history of DNA sequencing. Adapted from (Llaca and Messing 1998).

First Sanger introduced the possibility to sequence specific pieces of DNA

with a simple principle but a quite complex and laborious technique (Sanger method) (Maxam and Gilbert 1977; Sanger, Nicklen et al. 1977). Then the automated Sanger method (Hutchison 2007), using the capillary electrophoresis dominated the industry for almost two decades and led to a number of monumental accomplishments, including the completion of the only finished-grade human genome sequence (Lander, Linton et al. 2001; Venter, Adams et al. 2001). This method, though, still relied on big libraries of sheared DNA cloned into plasmids and fosmid subclones, requiring long time of sample preparation and analysis and also was not efficient enough to cover the gaps in highly polymorphic or repeated genomes. Despite many technical improvements during this era, the limitations of automated Sanger sequencing showed a need for new and improved technologies for sequencing large numbers of human genomes.

The Next Generation Sequencing technologies (NGS) replaced in many fields the use of the automated Sanger sequencing because of their ease of use, rapidity and sensitivity. Basically, they allowed parallelizing the sequencing process, producing thousands or millions of sequences at once.

With the introduction in the market of these new instruments the number of projects aiming to cover the entire genomic and epigenomic characteristics of all the different cell types, organs and organisms multiplied. The National Human Genome Research Institute (NHGRI) launched a public research consortium named ENCODE, the Encyclopedia Of DNA Elements, in September 2003, to carry out a project to identify all functional elements in the human genome sequence, that reached a productive phase in 2007, with the help on the new technologies (http://www.genome.gov/10005107). On its side many other projects came out, like the 1000 Genomes Project (http://www.1000genomes.org/) on human genetic variations, or the Genome 10K project (https://genome10k.soe.ucsc.edu/) to characterize genomes from 100 vertebrate genuses, or the BluePrint (http://www.blueprint-epigenome.eu/) to provide around 100 hematopoietic epigenome, or the Italian Epigen (http://www.epigen.it/) to characterize epigenomes from different human pathologies.

The first NGS platforms came out some years after the human genome project starting from 454 company in 2005, that launched the 454 platform (now Roche). Solexa released Genome Analyzer the next year, followed by

(Sequencing by Oligo Ligation Detection) SOLiD provided from Agencourt (now respectively Illumina and Life technologies) These are three most typical massively parallel sequencing systems in the NGS that shared good performance on throughput, accuracy, and cost.

These three platforms rely on different chemistries for sequencing and different outputs in terms of throughput and applications, each one with its pros and cons. They also have different outputs and accuracy (Fig 2).

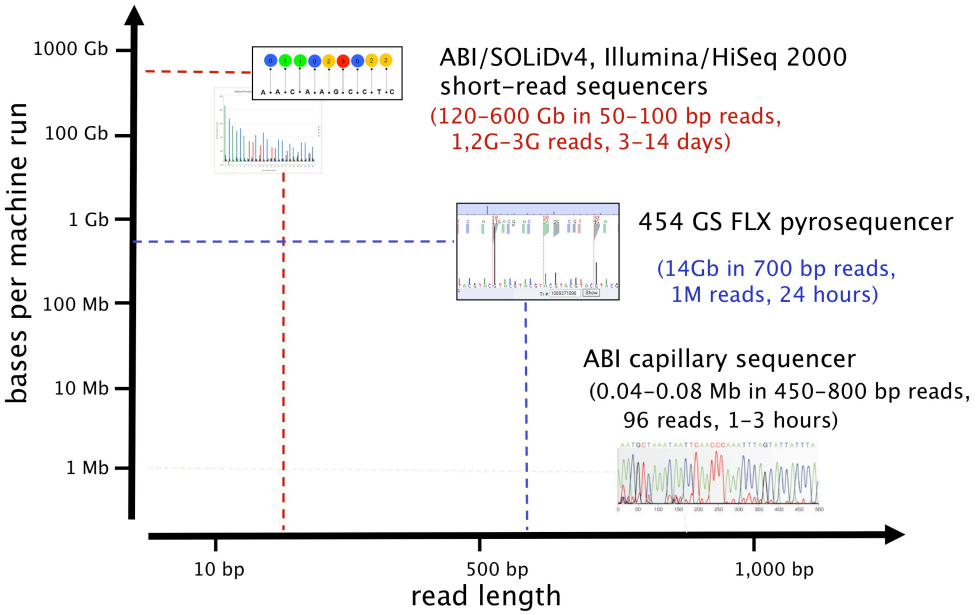

Figure 2. Representation of different techniques for DNA sequencing, based on amount of output and read length.

The most powerful and popular platforms available on the market today are the ones that came out first, Roche/454 FLX Pyrosequencer, Illumina/Genome Analyzer – HiSeq and Life/SOLiD.

The Illumina HiSeq 2000 features the biggest output, that was 200G per run initially, improved to 600G per run currently which can be finished in 8 days. It can have a 2% error rate and it is also the cheapest system in the market at the moment. With multiplexing incorporated in P5/P7 primers and adapters, it can handle thousands of samples simultaneously, that is another advantage of this system.

The SOLiD system has the highest accuracy among the others. The last version, SOLiD 5500xl, realized improved read length, accuracy, and data output of 85–100 bp, 99.99%, and 120G per run, respectively. A complete run can be finished within 7 days.

The Roche 454 system has the longest read length and fastest machine time. 454 GS FLX Titanium system give a read length of 700bps with accuracy 99.9% after filtering and outputs 14G data per run in 24 hours. One of the shortcomings is that it has relatively high error rate in terms of poly-bases longer than 6 bp.

454 system is the most used in applications where sequence coverage is highly important, for example in large genomes and de novo sequencing. On the contrary, the other two are mainly used for resequencing and in applications where the accuracy is important, as for mutation detection. This results in uses of the systems for different applications.

After years of evolution, these three systems exhibit better performance and their own advantages in terms of read length, accuracy, applications, consumables, man power requirement and informatics infrastructure, and so forth (Liu, Li et al. 2012).

In the present research project, only Illumina and SOLiD platforms have been used, specifically reflecting the scientific aims to pursue. In fact, the two experiments performed, the Chromatin Immuno-Precipitation (ChIP)-sequencing and RNA-sequencing, both need good accuracy and do not necessarily require long reads.

- **New and future approaches in sequencing**

NGS technologies are nowadays spreading more and more in labs all over the world, and the number of platforms is increasing year after year (Fig 3, omicsmaps.com).

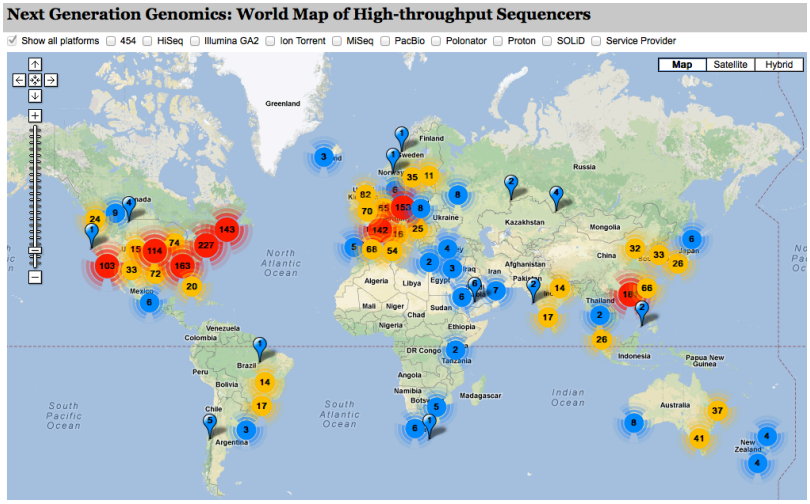

Figure 3. World distribution of NGS sequencers, which are becoming more diffused every day.

More compact sequencer, like Ion/Ion Personal Genome Machine (PGM) and Illumina/MiSeq came recently on the market featuring small size and fast turnover rates but limited data throughput. They are targeted to clinical applications and small labs. HeliScope is working with the technology true single molecule sequencing (tSMS). Pacific Bioscience introduced technology called single molecule real-time detection (SMRT). Both of them use some sort of nucleotide microscope, which is directly detecting incorporated nucleotide and thus avoiding many types of possible bias produced by other methods. Also it is supposed to be much much faster. Ion Torrent Systems, one of the newest companies, has developed technology based on a detection of hydrogen ions that are released during DNA polymerization (there is no need for optical detection systems). An advantage of this technique is a low cost of its reagents. A disadvantage is that only small fragments can be sequenced (Carneiro, Russ et al. 2012; Ginolhac, Vilstrup et al. 2012).

The applications to these technologies are as many as one can think (Table 1) and nowadays are the most diffuse methods for large-scale sequencing.

Table 1. Applications of NGS

Genome	*de novo*-seq: assembly of bacteria and lower eukaryotic genomes, metagenomics studies
	re-seq: Copy Number Variations (CNV) analysis, new variants discovery, Single Nucleotide Polimorphism (SNP) genotyping, association studies, cancer genomes
	targeted re-seq: exosome, closing of gaps
Transcriptome	mRNA-seq or total RNA-seq: quantitative and qualitative method for transcriptome analysis
	small RNA-seq
Epigenome	ChIP-seq: (Chromatin Immuno-Precipitation) transcription factors binding, motif discovery, histone modifications binding, effector proteins binding
	DNA methylation: bisulfite-converted DNA sequencing, methylation-sensitive restriction digest-enriched fragments, anti-methyl C-precipitated fragments, chromatin immunoprecipitates of methyltransferases trapped to aza-labeled DNA
	Higher order chromatin structures identification: MNase-seq, FAIRE-seq, DNase-seq, Hi-C

In the next section I will introduce the biological concepts underlying the working hypothesis of my research project.

Epigenetics and epigenomics

Epigenetics is one of the most rapidly expanding fields thanks to the recent technological breakthroughs in next generation sequencing. The advances of NGS technology made it possible to assess epigenetic marks at genome-wide scale, unveiling obscure aspects of gene expression regulation.

Epigenetic marks are classically defined as modifying factors of the chromatin, the highly structured DNA-protein complex organizing the genome of multicellular organisms (Espada and Esteller 2007). The main examples of modifying factors are the DNA methylation and histone H3 and H4 methylation and acetylation (Fig 4).

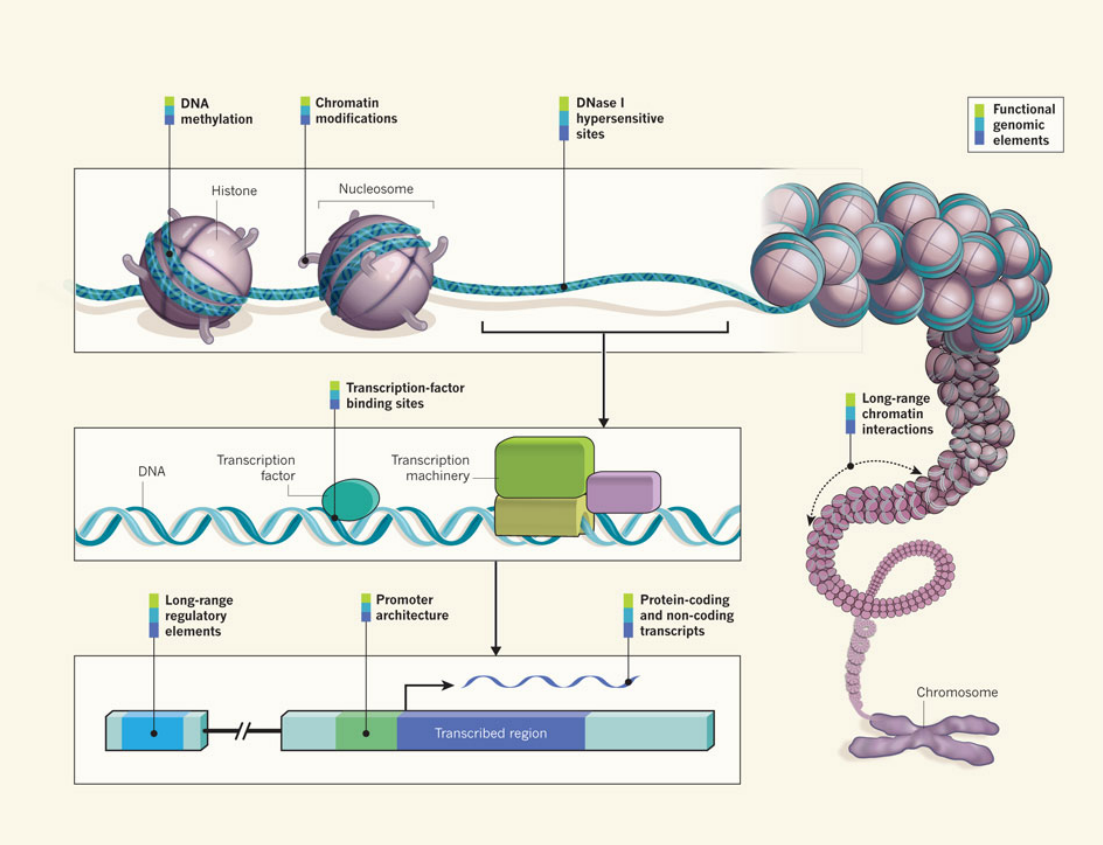

Figure 4. Beyond the sequence. DNA methylation and chemical modifications to histones can influence the rate of transcription of DNA into RNA molecules. Long-range chromatin interactions, such as looping, alter the relative proximities of different chromosomal regions in three dimensions and also affect transcription. Furthermore, the binding activity of transcription-factor proteins and the architecture (location and sequence) of gene-regulatory DNA elements, and more distant (long-range) regulatory elements play a role in transcription regulation. Accessible regions, called DNase I hypersensitive sites, are thought to indicate specific sequences at which the binding of transcription factors and transcription-machinery proteins has caused nucleosome displacement. From ENCODE explained (Ecker, Bickmore et al. 2012).

The levels of organization of the chromatin depend on the tridimensional positioning of the nucleosomes, the basic repetitive unit of the chromatin. Each nucleosome is formed by an octamer of proteins, composed of 4 groups of histones: one H3-H4 tetramer and two H2A-H2B dimers (Luger, Mader et al. 1997). All the proteins are wrapped in two turns of DNA filament (around 147 base pairs). A fifth histone type, H1, is the linker histone that connects each nucleosome to the next.

A simplistic model of the activity of the chromatin involves two basic states: the euchromatin, open and transcriptionally active, and the heterochromatin, highly condensed and transcriptionally repressed. In the genome we can find structures of constitutive heterochromatin (condensed mainly in centromeres) and regions that can undergo a transition from active to inactive state and vice versa.

These changes of state are fundamental in the regulation of the different transcriptional programs during the embryonic life, the development and the adult life. Moreover, they depend mainly from the epigenetic control mediated by histone modifications and DNA methylation.

Aberrant establishment of DNA methylation patterns is associated with several human disorders including chromatin diseases (Matarazzo, De Bonis et al. 2009), imprinting syndromes (Hirasawa and Feil 2010), psychiatric and neurodevelopmental defects, and immunological diseases (Portela and Esteller 2010). It also contributes both to the initiation and to the progression of various cancers (Jones 2002; Scarano, Strazzullo et al. 2005).

- **DNA methylation**

DNA methylation is present in almost all living organisms, from bacteria to plants and fungi, from invertebrates to vertebrates (Scarano, Strazzullo et al. 2005). Its abundance and its role vary markedly among the genomes, from the unmethylated genome of C. elegans to the heavily methylated genome of vertebrates. Different profiles of methylation in different species reveal the different role this DNA modification covers in their genomes. At an evolutionary level it has been proposed that DNA methylation developed as a generalized mechanism of repression in complex genomes (Bird 1995).

In mammals, DNA methylation represents a key layer of the transmitted epigenetic information mostly correlated with transcriptional gene silencing.

Cytosine methylation is required for embryonic development during which it plays a critical role in maintaining genomic integrity and regulating gene expression programs (Bird 2002; Li 2002; Mohn and Schubeler 2009). X chromosome inactivation, genomic imprinting, and the control of lineage specificity and pluripotency programs all represent processes for which proper DNA methylation is essential (Oda, Yamagiwa et al. 2006; Mohn, Weber et al. 2008; Borgel, Guibert et al. 2010). The role of DNA methylation in the tissue-specific expression of genes in somatic cells has more recently been uncovered.

Additionally, 5-hydroxymethylcytosine (5-hmC), which arises from the oxidation of the methyl group of 5-hmC, has recently been discovered in the mammalian genome (Kriaucionis and Heintz 2009; Tahiliani, Koh et al. 2009). Mechanisms and biological roles of non-CpG methylation and 5-hydroxymethylation remain unclear.

In the mammalian genome, DNA methylation occurs predominantly at the CpG dinucleotides and only occasionally at non-CpG sites. However, only certain CpG sites are methylated, resulting in the generation of a tissue- and cell-type-specific pattern of methylation.

CpGs are normally underrepresented in the genome, being usually quite rare. However, they can be found at a frequency closer to the statistical expectation in specific genomic regions, termed CpG islands (Gardiner-Garden and Frommer 1987). These represent 1% of the genome and are found in promoter regions of about 70% of all human genes and are usually unmethylated in normal cells. CpG islands generally show a relaxed chromatin without histone H1 and associate to nucleosomes with acetylated forms of histones H3 and H4 (Robertson and Wolffe 2000). However, about 6% of them become methylated in a tissue-specific program during early development or differentiation (Straussman, Nejman et al. 2009) (Fig 5a).

DNA methylation does not occur exclusively at CpG islands. Regions of lower CpG density lying in close proximity (~2 kb) of CpG islands, defined as CpG island shores, are methylated when associated with transcriptional inactivation (Fig 5b). Most of the tissue-specific DNA methylation seems to occur at CpG island shores (Doi, Park et al. 2009), which are also conserved between human and mouse. Furthermore, 70% of the differentially methylated regions during reprogramming are associated with CpG island

shores (Ji, Ehrlich et al. 2010). DNA methylation is less frequently correlated with permissive transcription, and in that case, it occurs at gene bodies (Ball, Li et al. 2009) (Fig 5c). Gene body methylation is common in housekeeping genes (Hellman and Chess 2007), and it is thought to be related to elongation efficiency, prevention of spurious initiations of transcription (Zilberman, Gehring et al. 2007) and to splicing regulation (Shukla, Kavak et al. 2011). A significant fraction of deeply methylated CpGs is also found in repetitive elements (Fig 5d).

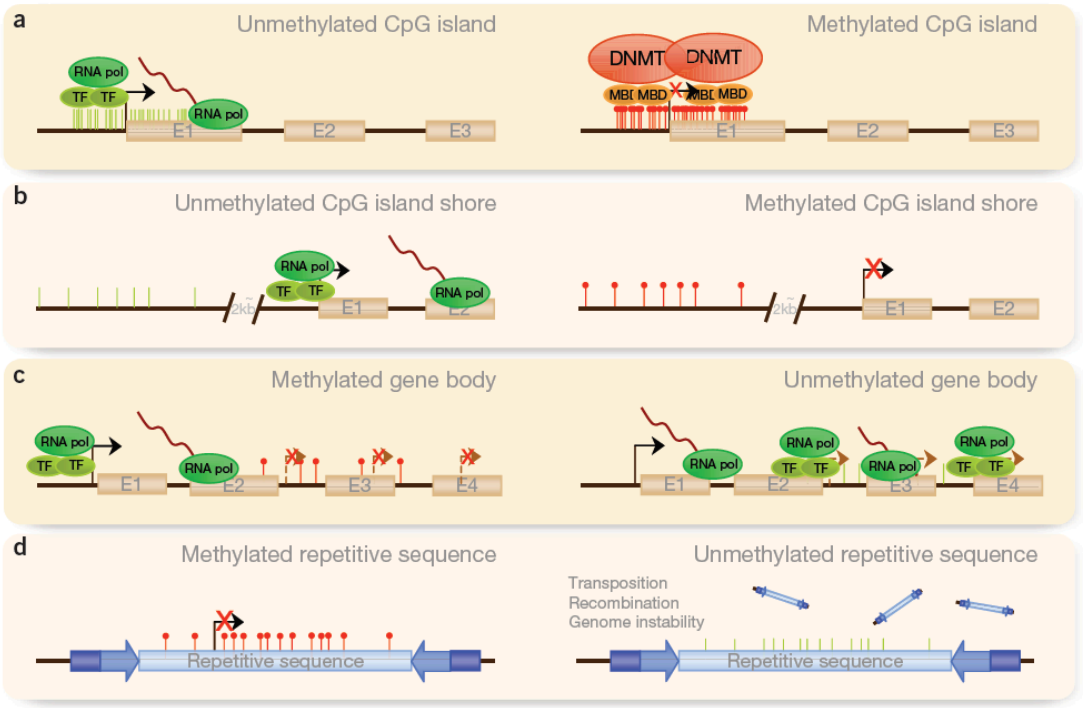

Figure 5. DNA methylation patterns. DNA methylation can occur in different regions of the genome. The alteration of these patterns leads to disease in the cells. In a, b, c, d are depicted the different methylation states with methylated cytosine in red and unmethylated in green. Refer to the text for a more detailed description. (From Esteller & Portela 2010)

This DNA methylation is necessary to protect chromosomal integrity, which is achieved by preventing reactivation of endoparasitic sequences that cause chromosomal instability, translocations and gene disruption (Esteller 2007).

The enzymes responsible for DNA methylation patterns are grouped in a family of cytosine C5-DNA methyltransferases (DNMTs) which act by transferring a methyl group from the universal methyl group donor, S-adenosyl-l-methionine (SAM), onto DNA (Fig 6) (Bestor 2000; Jurkowska, Jurkowski et al. 2011). In mammals, three enzymatically active members of the DNMT family have been reported (DNMT1, 3A, and 3B) and one related regulatory protein, DNMT3L, which lacks catalytic activity. DNMT3A and

DNMT3B have been considered as mainly devoted to the de novo methylation, being responsible for establishing the pattern of DNA methylation during embryonic development, whereas DNMT1, with preferential activity for hemimethylated DNA, acts mainly as maintenance methyltransferase. Null mutations of the three DNA methyltransferases are lethal in mice (Li, Bestor et al. 1992; Okano, Bell et al. 1999), clearly demonstrating that DNA methylation is essential for mammalian survival. Moreover, the recently produced triple KO mouse embryos (Dnmt1, 3A, and 3B mutant; TKO) unveiled the need of those enzymes for tissue-specific survival (Sakaue, Ohta et al. 2010).

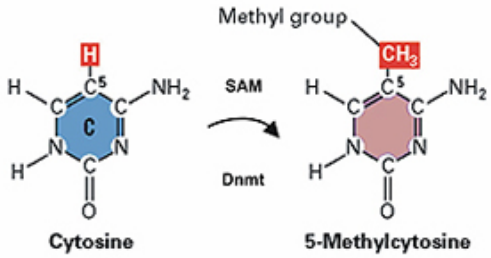

Figure 6 Cytosine methylation mediated by DNMT proteins, with SAM as methyl group donor.

The de novo DNMTs are highly expressed in embryonic tissue and stem (ES) cells and become downregulated in differentiated cells (Esteller 2007). Both DNMT3A and DNMT3B are stably associated with chromatin containing methylated DNA (Jeong, Liang et al. 2009) and localize to pericentromeric heterochromatin (Hansen, Wijmenga et al. 1999). DNMT3L acts as a stimulatory factor for DNMT3A and DNMT3B and interacts with them, being co-localized in the nucleus (Chen, Mann et al. 2005; Holz-Schietinger and Reich 2010).

The maintenance methyltransferase, DNMT1, shows a strong preference for hemimethylated DNA (Jeltsch 2006) due to its SET- and RING-associated (SRA) domain or the PHD (Bostick, Kim et al. 2007; Achour, Jacq et al. 2008). It is the most abundant DNMT in the cell and is localized at DNA replication foci during the S phase of the cell cycle; it is mostly required to methylate hemimethylated sites that are produced during semiconservative DNA replication. However, it also has de novo DNMT activity; in this latter function, DNMT1 might support DNMT3A and DNMT3B by using hemimethylated CpG sites produced by the DNMT3 enzymes as substrates (Fatemi, Hermann et al. 2002).

However, the distinction of functions between de novo and maintenance methylation is not always so clear, and several observations suggested an

active involvement of DNMT3 enzymes in the preservation of DNA methylation after DNA replication, especially in densely methylated or repetitive sequences. Accordingly, a revised and updated model has recently been proposed (Fig 7). This model still sustains the idea that the bulk of DNA methylation in replicating cells would be maintained by DNMT1 together with UHRF1 and PCNA. However, it also proposes that DNMT3A and DNMT3B, which have been shown to anchor strongly to nucleosomes containing methylated DNA, contribute to the maintenance of methylation at heterochromatic regions, de novo methylating the sites missed by DNMT1 at the replication fork (Jones and Liang 2009).

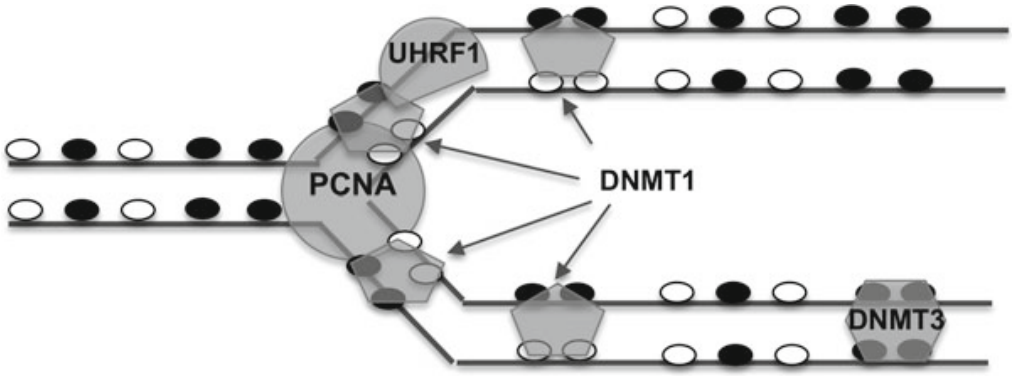

Figure 7. Up-to-date model for the maintenance of DNA methylation patterns after replication. DNMT1 localizes at the replication fork, and its methyltransferase activity on hemimethylated cytosines is promoted through its interaction with PCNA and UHRF1 proteins. DNMT3 enzymes actively participate also in the maintenance process of heavily methylated regions, ensuring methylation at CpG sites, which are missed by DNMT1. (From Gatto et al., 2011)

As is also emerging from the genome-wide methylome studies, the novel view is that "maintenance DNA methylation" implies the preservation of average levels of DNA methylation at certain regions rather than the accurate copy of individual CpG sites. That would be sufficient to ensure the inheritance of the epigenetic information in a stable manner (Gatto 2012).

- **Histone modifications**

The histones are small basic proteins formed by a globular domain and a flexible and charged NH2-terminal tail that hangs out of the nucleosomal structure. Histone tails can be subject to a great number of reversible enzymatic modifications in specific positions, mainly acetylation, methylation and phosphorylation (Margueron, Trojer et al. 2005; Nightingale, O'Neill et al. 2006).

These modifications alter DNA-histones interactions and have a strong impact on chromatin structure. In particular, lysine acetylation is catalyzed by histone acetyl-transferases (HAT) and marks the opening of the chromatin, while deacetylation of those residues from histone deacetylases (HDAC) is associated with transcriptional repression. Methylation of arginine and lysine occurs in histones H3 and H4 in the mono-di and tri-methylated form and histone methyltransferases (HMTs) catalyze this reaction. Depending on the histone type and the specific methylation site this particular modification can have different functional meanings. H3K9, H4K20 and H3K27 methylation is generally connected to heterochromatin formation and gene silencing respectively, while H3K4, H3K36 and H3K79 methylation is generally associated to euchromatin and transcriptionally active regions (Barski, Cuddapah et al. 2007; Portela and Esteller 2010).

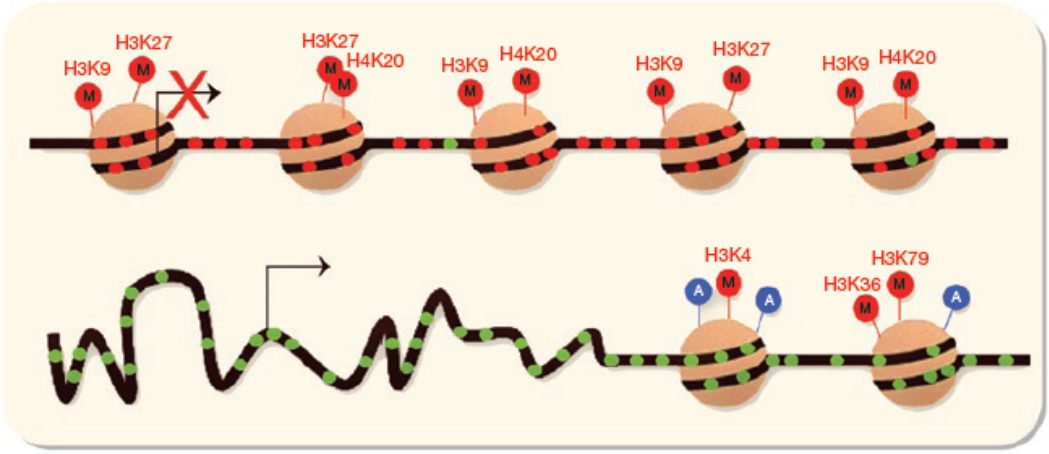

Figure 8. Histone modifications, DNA methylation and nucleosome positioning patterns. Transcriptionally active gene promoters (below) possess a nucleosome-free region at the 5′ and 3′ untranslated region, providing space for the assembly and disassembly of the transcription machinery. Methylated DNA (red dots) seems to be associated with 'closed' chromatin domains, where DNA is condensed into strictly positioned nucleosomes, thereby impeding transcription. Conversely, unmethylated DNA (green dots) is associated with 'opened' chromatin domains, which allow transcription. Histone acetylation (A) and methylation (M) have specific roles in the opening and closing of the chromatin. From (Portela and Esteller 2010).

Histone modifications can influence each other and interact with DNA methylation and drive the nucleosome repositioning (Fig 8). This combination of information is finely tuned in time and space and aims to appropriately program the expression profile in each single cell of the organism.

Two types of protein complexes participate with different roles to histone code regulation: one contains proteins of Trithorax group (TrxG) and the

other has Polycomb group proteins (PcG). Some of the components of the two groups have histone methyl-tranferase activity, while others have a reader role, interpreting the histone signals playing a central role in the epigenetic regulation of gene expression. Those complexes coordinate DNA accessibility during development and differentiation modulating the balance between silenced heterochromatin (bound by PcG) and transcriptionally competent euchromatin (bound by TrxG) (Schuettengruber, Chourrout et al. 2007).

One of the most heavily characterized markers of heterochromatin is trimethylated lysine 9 on H3 (H3K9me3). H3K9 can exist in a mono- (H3K9me1), di- (H3K9me2), or trimethylated state, in which multiple methyltransferase and demethylase enzymes act in concert to control distinct methylation profiles. Di- and trimethylation of H3K9 create binding sites for chromodomain containing proteins, including those of the heterochromatin protein 1 (HP1) family and are believed to promote transcriptional repression and genomic silencing through alterations in higher order chromatin structure throughout euchromatic and constitutively heterochromatic genomic loci (Bannister, Zegerman et al. 2001; Lachner, O'Carroll et al. 2001).

Although rare exceptions exist, the H3K9me3 mark, unlike H3K9me2 or H3K9me1, is thought to primarily reside in silenced, noncoding regions of the genome (Rosenfeld, Wang et al. 2009). Recent ChIP-Sequencing analyses have demonstrated that H3K9me3 is prevalent at many non-genic regions including the repetitive satellite DNA, centromeric and pericentromeric DNA and long terminal repeats of transposons (Mikkelsen, Ku et al. 2007; Rosenfeld, Wang et al. 2009).

- **Epigenetic cross-talk between DNA and histone methylation**

All the epigenetic factors, besides having a specific role defined by their intrinsic functions, have the capacity to interact to modulate each other's activity. DNA methylation, for example, can express its repressive activity through different mechanisms. Delivery of DNMTs to target genes through interaction with sequence-specific transcription factors or chromatin-interacting proteins has already been demonstrated in several examples.

DNMT3A has been reported to interact with several transcription factors, such as PU.1 (Suzuki, Yamada et al. 2006), Myc (Brenner, Deplus et al. 2005), and p53 (Fuks, Burgers et al. 2001; Wang, Kamarova et al. 2005). Additionally, the mammalian H3K9/H3K27 histone methyl transferase (HMT), G9a, is required for the recruitment of de novo DNMTs to gene promoters during mouse ES cells differentiation (Feldman, Gerson et al. 2006), whereas EZH2 (enhancer of zeste homologue 2), an H3K27-specific HMT, is involved in the recruitment of DNMT3A and 3B in cancer cells (Vire, Brenner et al. 2006). Variable interactions between H3K27me3 and DNA methylation have been also recently found in bisChIP-seq studies (ChIP followed by bisulfite-sequencin) (Brinkman, Gu et al. 2012; Statham, Robinson et al. 2012), where the histone mark is found alternatively in association with fully methylated or unmethylated DNA, depending on the sequence characteristics and the cell type. Finally, histone deacetylases (HDACs) and heterochromatin protein 1 (HP1) directly interact with DNMTs, and it has been suggested that they participate in the delivery of DNMTs to silenced chromatin regions (Fuks, Hurd et al. 2003).

Moreover, DNMT enzymes are also interacting with other histone modifications (Ooi, Qiu et al. 2007; Tachibana, Matsumura et al. 2008; Jeong, Liang et al. 2009) that influence their activity. Recent data have reported that DNMTs can directly read histone modifications through their N-terminal domains and apparently could be recruited to the nucleosomes containing unmethylated H3K4 (Ooi, Qiu et al. 2007; Otani, Nankumo et al. 2009; Zhang, Jurkowska et al. 2010). Because methylation of H3K4 is a chromatin mark associated to transcribed genes, the absence of this modification in specific regions could be read as a signal for their inactivation, whereas its presence could reject DNA methyltransferases. Moreover, targeting of DNA methylation by H3K36me3 is consistent with many studies indicating that this histone mark accumulates in the bodies of active genes (Vakoc, Sachdeva et al. 2006; Barski, Cuddapah et al. 2007), accordingly to the observation that active gene bodies are strongly methylated compared to inactive ones. Besides, more results suggest that DNA methylation and H3K36 methylation might have a role in regulating the splicing, with exons having increased levels of both H3K36me3 and DNA methylation compared to introns. Overall this suggests that the targeting of DNMTs by DNA- or chromatin-binding proteins is a widespread and general

mechanism for the generation of specific DNA methylation patterns within a cell.

Viceversa, DNA methylation can itself influence the binding of proteins that can modify the chromatin according to the transmitted signal. For instance, DNA methylation on specific cytosines recruits regulatory proteins, such as methyl group binging proteins (MBP). Methyl-CpG-binding proteins (MBPs) directly recognize methylated DNA and recruit co-repressor molecules to silence transcription and to modify surrounding chromatin (Klose and Bird 2006). These MBDs belong to three structural families: the MBD family, the SRA family (SET and RING-finger associated domain) and the zinc finger family (Buck-Koehntop and Defossez 2013).

The MBD is a family of seven proteins cointaining a methyl-CpG binding domain (MBD) that convert the methylation signal of DNA to a repressed state of the chromatin recruiting in turn other big regulatory complexes (Jones, Veenstra et al. 1998; Nan, Ng et al. 1998).

The first identified MBD, the methyl-CpG binding protein 2 (MECP2) selectively recognizes methylated DNA (Lewis, Meehan et al. 1992), and directly interacts with mSin3A, a co-repressor complexed with histone deacetylases (HDAC) (Jones, Veenstra et al. 1998; Nan, Ng et al. 1998). It also directly binds the histone methyltransferases (Yu, Thiesen et al. 2000; Fuks, Hurd et al. 2003) and interacts with transcriptional factors (as TFIIB) (Yu, Thiesen et al. 2000).

The second family of MBPs, the SRA family, includes UHRF1 and UHRF2, two related proteins that are thought to bind methylated DNA via their SRA domains. UHRF1 is an essential protein that binds hemimethylated DNA and recruits DNMT1 to facilitate maintenance DNA methylation; in the absence of UHRF1, there is a precipitous loss of DNA methylation (Unoki, Nishidate et al. 2004; Bostick, Kim et al. 2007; Sharif, Muto et al. 2007).

The third, and currently last, family of MBPs includes the zinc finger protein Kaiso, which is able to discriminate methylated from unmethylated DNA (Prokhortchouk, Hendrich et al. 2001). Kaiso has two close paralogs in mammalian genomes: Zbtb4 and Zbtb38 (Sasai and Defossez 2009). These proteins, like Kaiso, bind methylated DNA but can also bind a non-methylated consensus (Filion, Zhenilo et al. 2006; Sasai, Nakao et al. 2010).

Very recently, another zinc finger protein, ZFP57, was also shown to bind methylated DNA and to act in DNA methylation-dependent maintenance of imprinted genes (Quenneville, Verde et al. 2011).

Chromatin diseases and ICF syndrome

The functional meaning of the role of epigenetics is even clearer studying cells where these mechanisms are disrupted. Many human pathologies are caused by impairment of gene expression; DNA and chromatin modifications, epigenetics signals, take care of the regulation and control of the correct functioning of gene expression in mammalian cells. Somatic mutations of chromatin structural components or regulatory proteins can cause cancer in many different tissues. Germinal mutations, instead, can be inherited and transmitted to all the cells in the body and are therefore causing chromatin genetic diseases.

The study of this category of diseases allows us to understand more about the epigenetic regulation of gene expression and its direct effect on development. Moreover, this type of studies allows us to understand more about the molecular mechanisms underneath the illnesses and to uncover new therapeutic approaches to improve the pathological phenotypes.

As I mentioned before, the epigenetic regulation machinery is formed by a complex and entangled structure in which specific components combinations have specific roles. Depending on which piece of the network is impaired it can result on a different phenotype. Moreover, in some diseases, different mutations in the same protein cause high complexity and variability of the phenotype from a subject to the other and this can reflect the complex function of these proteins, whose impairment can have effects on multiple downstream targets.

The genetic chromatin disease I focused my studies on is the ICF syndrome (Immunodeficiency, Centromere instability and Facial anomalies, OMIM #242860). The ICF syndrome is a very rare autosomal recessive disease that severely damages the immune system of the affected subjects and exhibits a diffuse hypomethylation of specific heterochromatic regions of the DNA (Tiepolo, Maraschio et al. 1979; Maraschio, Zuffardi et al. 1988).

So far, around 60 ICF patients have been reported worldwide, and they have been classified in two distinct disease classes, ICF types 1 and 2, due to their genetic and epigenetic features (ICF1 and ICF2, Fig 9) (Hansen, Wijmenga et al. 1999; Jiang, Rigolet et al. 2005). Both classes present the same clinical phenotype, and until early 2011, their distinction criteria were the presence of mutations in the DNA methyltransferase 3B gene (DNMT3B) for ICF1 and hypomethylation of alpha satellites in centromeric heterochromatin for ICF2 patients (Jiang, Rigolet et al. 2005). Recently, de Greef, Wang et al. (2011) and Chouery, Abou-Ghoch et al. (2012) identified several mutations in the zinc-finger- and BTB (bric-a-bric, tramtrack, broad complex)-domain-containing 24 (ZBTB24) gene at 6q21 highly associated to ICF phenotype in some ICF2 patients (Fig 9d). With this finding, the ICF type 2 is now split in two subcategories, where alpha satellite hypomethylation is present, but ZBTB24 can either be mutated or not (Fig 9e).

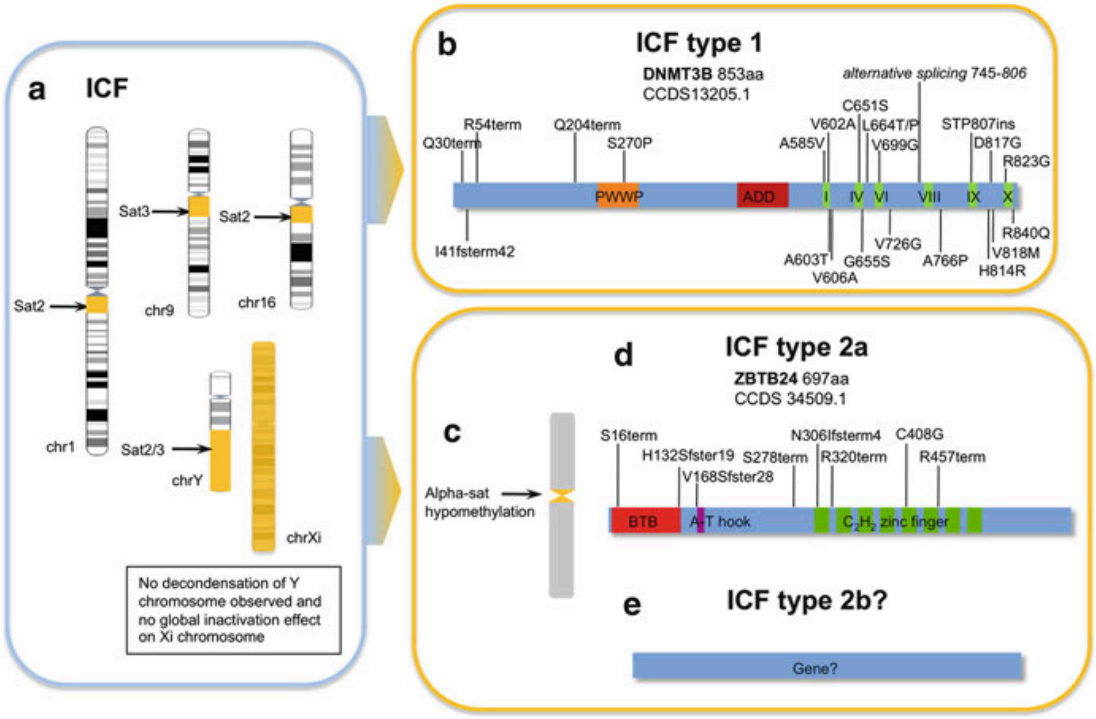

Figure 9. ICF syndrome molecular features. **a.** Hypomethylation of juxtacentromeric heterochromatin of chromosomes 1, 9, 16, and Y and of the inactive X chromosome. Regions of interest are marked in yellow. **b.** Mutations in the DNMT3B gene causing ICF type 1. In green are the active sites of the catalytic domain. **c.** Alpha satellite of centromeric heterochromatin is hypomethylated only in ICF type 2 on all chromosomes. **d.** Mutations in ZBTB24 are mostly nonsense and represent the hallmark of ICF type 2a. **e.** ICF type 2b has yet to be well characterized. It can be only de fi ned as neither type 1 nor 2.

ICF1 subjects present biallelic mutations in the DNMT3B gene at chromosomal locus 20q11.2, all leading to the hypofunctioning of the

protein. Twenty-three mutations have been reported until now, and they are listed in Fig 9b (Jiang, Rigolet et al. 2005; Hagleitner, Lankester et al. 2008). DNMT3B mutations are mainly missense and mostly concentrated in the C-terminal portion where they partially affect the catalytic function of the protein. All the major mutations, like the nonsense ones, appear in the N-terminal regulatory part of the protein and are always found as compound heterozygous, as the complete loss of function in the homozygous state is probably incompatible with life, analogous to the situation in mice. ZBTB24 (also known as ZNF450, BIF1, or PATZ2) is a member of the ZBTB family of transcriptional factors with a prominent role in hematopoiesis (Edgar, Dover et al. 2005; de Greef, Wang et al. 2011). Mutations of this protein in ICF2 are always biallelic and mostly nonsense, leading to the loss of function of the protein (Fig 9d). Up to now, eight mutations have been identified, only one missense, and only two of ten mutated patients are compound heterozygous, with the rest being homozygous. Both DNMT3B and ZBTB24 are ubiquitously expressed and apparently have different functions in the cell, but mutations in both lead to the same phenotype. The effects of DNMT3B mutations have been studied more in depth, and more information is available on their pathogenic effects, while, due to the only recent discovery of ZBTB24 mutations, their pathogenic mechanisms are still obscure.

- **Clinical and cytological phenotype**

ICF patients are mostly diagnosed during childhood due to recurrent infections, the characteristic symptom of the syndrome. In the blood biochemical analysis, they all show a combined immunodeficiency with reduction or absence of serum immunoglobulins of all subtypes (in different combinations) with a normal number of B and T cells (Blanco-Betancourt, Moncla et al. 2004). ICF patients, thus, are prone to recurrent severe respiratory and gastrointestinal infections that often cause death at young age. To complete the heterogeneous picture of the ICF phenotype, only some patients show facial anomalies and the other symptoms have an even more reduced penetrance, being present only in few individuals. Few ICF patients present congenital defects, hematological abnormalities, or malignancies (see Hagleitner, Lankester et al. (2008) for a complete description of the range of phenotypes).

The hallmark of this syndrome lays in the karyotype of the affected subjects, where chromosomes 1, 9, and 16 show evident decondensation of juxtacentromeric heterochromatin causing chromosome breaks and rearrangements in radial structures only in phytohemagglutinin-stimulated peripheral blood lymphocytes. The molecular basis of this phenomenon has mainly been addressed to the loss of DNA methylation within classical satellites (Sat 2 and 3) at the juxtacentromeric heterochromatin of the long arms of chromosomes 1, 16, sometimes 9 and Y in males (Fig 9a). DNA hypomethylation is also present in the nonsatellite repeats NBL2 on acrocentric chromosomes and D4Z4 in the subtelomeres of the long arms of chromosomes 4 and 10 (Jeanpierre, Turleau et al. 1993; Kondo, Bobek et al. 2000; Tuck-Muller, Narayan et al. 2000). Additional hypomethylation, localized in the alpha-satellite repeats of the centromeres, is found only in ICF2 patients (Miniou, Jeanpierre et al. 1997; Jiang, Rigolet et al. 2005) (Fig 9c). This DNA hypomethylation is present in all analyzed cell types, but it gives rise to rearrangements only in lymphoblasts, probably playing a specific role in the onset of the immunologic phenotype (Jeanpierre, Turleau et al. 1993).

- **Molecular phenotype**

Despite the disease has been described more than twenty years ago, the ICF syndrome pathogenesis is not clear yet. Particularly, it is not known why the impaired DNMT3B activity mainly leads to an immune-specific phenotype and to what extent the activity of DNMT3B on its specific genomic targets in lymphocytes is altered. Besides its biomedical interest, ICF syndrome represents an ideal model system to study the intricate interactions between chromatin regulating layers.

Bioinformatics analysis of gene expression microarrays on lymphoblastoid cell lines (LCLs) showed that most of the affected genes were critical for immune function, development and neurogenesis, which are highly relevant to the ICF phenotype (Ehrlich, Buchanan et al. 2001; Jin, Tao et al. 2008). To better understand the molecular derangement observed in ICF syndrome, a number of epigenetic aspects have been analyzed by our group over the years, ranging from the contribution of the higher-order nuclear organization to the microRNA epigenetic regulation. At the present, the

genome-scale chromatin modifications in ICF cells is the next challenging step that needs to be explored.

The dysregulated genes in ICF cell lines have been found mainly associated to changes in chromatin modifying proteins and rarely to aberrant methylation, as it would be expected (Jin, Tao et al. 2008). Some other genes did not show any change at all by ChIP-qPCR and targeted bisulfite sequencing analyses. On another level of regulation, miRNA expression has been tested with a microarray, finding eighty-nine dysregulated miRNAs, some of which involved in immune function, development and neurogenesis. Again, significant DNA hypomethylation of miRNA CpG islands was not observed in all cases of miRNA up-regulation in ICF cells, suggesting a more subtle effect of DNMT3B deficiency on their regulation; however, a modification of histone marks, especially H3K27 and H3K4 trimethylation, was observed concomitantly with changes in microRNA expression. Functional correlation between miRNA and their target gene expression suggested a regulation either at mRNA level or at protein level (Gatto, Della Ragione et al. 2010).

Although doubtless perturbed, how broadly the histone modifications change as a consequence of the impaired DNMT3B activity and the entity of the damages that they cause to gene expression is still unclear. With Next-Generation Sequencing it is finally possible to dig deeper into the multiform molecular phenotype of this syndrome. The first whole-genome experiment performed on cells from one ICF patient has been the bisulfite-seq from (Heyn, Vidal et al. 2012). They detected a decrease of methylation level of 42% (much higher than the 7% detected with the old techniques), with the most profound changes occurring in inactive heterochromatic regions, satellite repeats and transposons. Interestingly, transcriptional active loci and ribosomal RNA repeats escaped global hypomethylation. Despite a genome-wide loss of DNA methylation the epigenetic landscape and crucial regulatory structures were conserved. Remarkably, a mislocated activity of mutant DNMT3B to H3K4me1 loci was detected resulting in hypermethylation of active promoters.

3. Results – Part I – The Pipeline

As I mentioned in the preface the aims of my work during my PhD were two. The first was to build a reliable pipeline for the analysis and the integration of ChIP- and RNA-seq data; the second was to uncover new large-scale aspects of the epigenetic perturbation in ICF lymphoblastoid cell lines through the comparison of histone modifications (with ChIP-seq), gene expression (RNA-seq and miRNA array) and DNA methylation (bisulfite-seq). In this section of the results I will first introduce the NGS data characteristics and some general concepts regarding the analysis, then I will describe the definition of an appropriate pipeline for my purposes.

Data analysis for ChIP-seq and RNA-seq is a bottom-up process that begins with mapped sequence reads and proceeds upward to produce increasingly abstracted layers of information. In Fig 10 there is a generalized pipeline (Pepke, Wold et al. 2009) that describes the numerous passages that lead to a complete analysis of sequencing data. A pipeline can be defined as a combined set of instructions and programs that connects raw data (input) with a certain results. It can be easily described by a direct graph whose nodes are programs and whose arcs are intermediate results. Intermediate results are the partial output of a program that become the input of another program, sometimes intermediate results are by themselves of interest. A pipeline can be then applied automatically to perform analogous actions and analysis on several dataset. The workflow has to be adapted then to the characteristics of the data obtained and to the

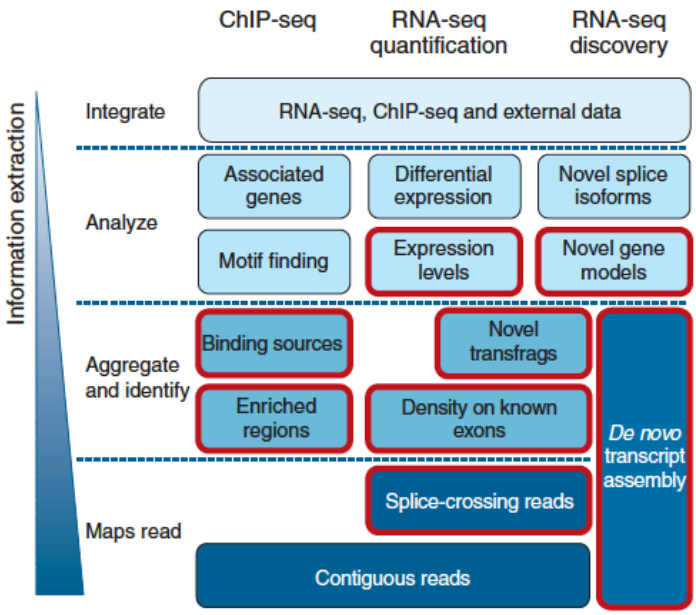

Figure 10. A hierachical overview of ChIP-seq and RNA-seq data analyses. The bottom-up analysis of ChIP-seq and RNA-seq data typically involves the use of several software packages whose output serves as the input of the higher level analyses. (From Pepke, Wold et al. (2009)).

specific purposes driven by the biological questions of the project. Each passage has to be carefully performed and tuned to avoid repercussions on the final interpretation of the data; a careful evaluation of each step will lead to a correct and reliable result.

The data analysis can be divided in three main steps, that can be named primary, secondary and tertiary analysis. The primary analysis is performed by the sequencing machines, which output the raw sequence files and perform the first quality control. In the secondary analysis the short sequences need to be mapped to a reference genome or transcriptome (for species with a fully-sequenced genome). It is not a small task to optimally align tens or even hundreds of millions of sequences to multiple gigabases for the typical mammalian genome, and this early step remains one of the most computationally intensive in the entire process. At last, the tertiary analysis concerns the actual information extraction and can vary significantly from one experiment to the other.

In this work ChIP-seqs and an RNA-seqs have been performed on human-derived cell lines. Before mapping them on the human genome the quality assessment of all the raw files has been performed.

All data analyses have been performed on Lilligrid, a cluster of 20 64 bits dual-processor nodes in rack configuration from the Istituto per le Applicazioni del Calcolo "Mauro Picone" (IAC-CNR, "Institute for Calculus Applications").

All the tools used in this work are listed in Table 2.

	Version	Application	Link
BioScope	1.2	ChIP-seq mapper	New site: http://www.lifescopecloud.com/
bowtie	0.12.7	ChIP-seq mapper	http://bowtie-bio.sourceforge.net/index.shtml
SAMtools	0.1.17	SAM-BAM manipulation tools	http://samtools.sourceforge.net/
bedtools	2.9.9	BED manipulation tools	http://code.google.com/p/bedtools/
FastQC	0.10.0	Quality control tool	http://www.bioinformatics.babraham.ac.uk/projects/fastqc
SICER	1.1	Unsupervised peak finding for ChIP-seq	http://home.gwu.edu/~wpeng/Software.htm
EpiCHiP	0.9.7-e	Supervised peak finding for ChIP-seq	http://epichip.sourceforge.net/
DESeq	1.10.1	Differential gene expression and peak enrichment	http://bioconductor.org/packages/2.11/bioc/html/DESeq.html
ChIPpeakAnno	2.6.0	Sequence annotator	http://www.bioconductor.org/packages/2.11/bioc/html/ChIPpeakAnno.html

PeakAnalyzer	1.4	Sequence annotator	http://www.bioinformatics.org/peakanalyzer/wiki/
TopHat	1.3.1	RNA-seq mapper	http://tophat.cbcb.umd.edu/
HTseq	0.5.3.p1	ChIP-seq data manipulation	http://www-huber.embl.de/users/anders/HTSeq/doc/index.html

Table 2. List of all programs used in this work, with version number, application and link.

Data formats

Several specific formats have been proposed for storing genomic data, few of them become de-facto standard and are now commonly used. Being able of converting and manipulating different data format is fundamental for building the computational pipeline.

Indeed a pipeline need to define at each step the input data format and the output data format, data conversion is required to make an intermediate output suitable for the next input. The advantages of having few de-facto standard for data format and some tools for their manipulation and conversion significantly facilitate the construction of efficient pipeline.

- **Raw data -fastq, csfasta, qual**

Before facing the issue of the pipeline construction for the secondary analysis it is important to briefly introduce the format of the output produced by the NGS machines. Image acquisition and processing is a fundamental process that can vary depending on the technology (see chapter 5, materials and methods); from the processed image the software outputs a raw results file containing the sequences with the quality assessment of every single base read.

This information is stored in different file formats, depending on the machine. Illumina uses the format fastq that is a simple fasta file with the integration of the quality scores. An example is shown here:

.fastq

```
@DBV2SVN1:119:C1BFAACXX:7:1101:1226:2107 1:N:0:GGCTAC
AGGATTAATATAGTAAAAATGGCCATTTTCCAAAAGCAATCTAAAGATTCA
+
@CCFFFFFHHHGHIIIJJJGCGGIJJJJJIJJJJJGAFHIJIJJJJJJJJI
```

Example	Description
DBV2SVN1	unique instrument name
119	run id
C1BFAACXX	flowcell id
7	flowcell lane
1101	tile number within the flowcell lane
1226	'x'-coordinate of the cluster within the tile
2107	'y'-coordinate of the cluster within the tile
1	member of a pair, 1 or 2 (paired-end or mate-pair reads only)
N	Y if the read fails filter (read is bad), N otherwise
0	0 when none of the control bits are on, otherwise is an even number
GGCTAC	index sequence

Table 3. Fastq format

The first line represents the description of the sequence. This description has changed with the different versions of the analysis software and this is how it appears now with the current version (v1.8, Table 3).

The second line contains the proper sequence, while the fourth line encodes the quality values for the sequence in line two, and must contain the same number of symbols as letters in the sequence.

SOLiD systems also use a similar format, but, like 454, it stores the quality information in a different file. The main difference between the two systems, though, resides in the nature of the data, as 454 reads the sequence in base-space, while SOLiD uses the color space, where colors represent unique couples of bases (see Materials and methods for more information on color space). Consequently, 454 stores the sequence in a fasta file, with the actual sequence in bases, while SOLiD uses the csfasta, with numbers representing colors. csfasta and _QV.qual files are here described:

.csfasta

```
>931_29_9_F3
T011213122200221123032111221021 21.131332222101
```

_QV.qual

```
>931_29_9_F3
31 32 29 31 31 31 33 31 31 33 30 22 31 30 30 32 29 30 31 26 33
33 31 24 33 26 22 24 32 21 29 30 31 30 29 29 27 -1 28 26 29 22
31 26 29 32 29 25 29 24
```

The first line, starting with > indicates the description of the sequence, with position coordinates and primer used for sequencing (F3 in this case).

If a read has a "." like in the example it means that the color calling was ambiguous (this would have been an N if it was in base space). In this case, the workflow simply cuts off the rest of the read, since there is no way to know the right phase of the rest of the colors in the read. If the read starts with a dot, it is not imported. In the quality file, the equivalent value is -1, and this will also cause the read to be clipped.

- **Mapping Output – SAM, BAM, BED**

The mapping pipeline can output different kinds of files, depending on the software used. The SAM/BAM format (Sequence Alignment/Map or Binary AM) was created by a team at the Sanger center to support 1000 Genome Project and now is becoming the most diffuse file format for mapped sequences. The SAM and BAM files contain the same information but in different formats. While SAM is a tab-delimited text file storing sequence data in a series of tab delimited ASCII columns, BAM is its binary form. SAM files can have headers and must have alignment information. Each header line begins with character @ followed by a two-letter record type code (@HD, @SQ, @RG, @PG, @CO). In the header, each line is TAB-delimited and except the @CO lines, each data field follows a format `TAG:VALUE' where TAG is a two-letter string that defines the content and the format of VALUE. It contains information about the alignment. The alignment section is mandatory and has to contain at least 11 tab-separated fields.

An example of the alignment section of a SAM file is below. In table 4 the details of the sequence format.

```
1303_20_178_F3 0       chr2    120416180       255     48M     *       0       0
AGCCCCCCTCAACACGCACACACACACACACACACACACACATTTTCA
        PGLMU?5AB67PSW]YLKSXVSSUYVSS:?Z<<V;8POTTL)!-<1.A    XA:i:2  MD:Z:48
        NM:i:0  CM:i:7
```

Field	Alignment	Description
QNAME	1303_2_221_F3	Query template NAME
FLAG	4	bitwise FLAG
RNAME	chr2	Reference sequence NAME (chromosome)
POS	120416180	1-based leftmost mapping POSition
MAPQ	255	MAPping Quality

CIGAR	48M	CIGAR string
RNEXT	*	Ref. name of the mate/NEXT segment
PNEXT	0	Position of the mate/ NEXT segment
TLEN	0	observed Template LENght
SEQ	AGCCCCCCTCAACACGCACACACACACACACACACACACACATTTTCA	segment SEQuence
QUAL	PGLMU?5AB67PSW]YLKSXVSSUYVSS:?Z<<V;8POTTL)!-<1.A	ASCII of Phred-scaled base QUALity
TAG	XA:i:2 MD:Z:48 NM:i:0 CM:i:7	Tags with format TAG:TYPE:VALUE

Table 4. SAM format

More information on SAM/BAM formats can be found in The SAM Format Specification (v1.4-r985; http://samtools.sourceforge.net/SAM1.pdf).

The set of tools that allows the manipulation of these data is SAMtools (Li, Handsaker et al. 2009). SAMtools provides various utilities for manipulating alignments in the SAM format, including sorting, merging, indexing and generating alignments in a per-position format.

Most of the event along the genome can be described in terms of intervals (e.g., mapped reads, peaks position, genes, or other annotations such as CpG islands, etc); the BED format is a simple tab-delimited text file suitable for containing this type of information. Each line of the file represent an interval, described in terms of chromosome, start and end position of the interval along the sequence. Additionally the interval can then have a name describing the annotated feature, a score and strand information.

More in general BED format provides a flexible way to define the data lines that are displayed in an annotation track in a genome browser. BED lines have three required fields and nine additional optional fields. The number of fields per line must be consistent throughout any single set of data in an annotation track. The order of the optional fields is binding: lower-numbered fields must always be populated if higher-numbered fields are used.

Below there is an example. In Table 5 a detailed description.

```
chr7    127471196  127472363  Pos1  0  +  127471196  127472363  255,0,0

chr7    127472363  127473530  Pos2  0  +  127472363  127473530  255,0,0
```

Field	Example	Description
Chrom	chr1	chromosome
ChromStart	1000	start position of feature

ChromEnd	`2000`	end position of the feature
Name	`seq1`	name of the feature
Score	`960`	a score between 0 and 1000
Strand	`+`	strand, can be + or -
thickStart	`1000`	starting position at which the feature is drawn thickly
thickEnd	`2000`	ending position at which the feature is drawn thickly
itemRgb	`255,0,0`	RGB color of the feature to display
blockCount	`2`	Number of blocks in the line
blockSizes	`345,45`	Sizes of blocks in the line
blockStarts	`500,1300`	Starts of blocks in the line

Table 5. BED format

In order to deal with bed files and therefore with genome data, Bedtools is the most used, fast and flexible toolset for genome arithmetic (Quinlan and Hall 2010). It supports a wide range of operations for interrogating and manipulating genomic features. One of its tools is called bamToBed and converts files from bam to bed format.

- **Coverage – wig**

Another, very meaningful way to summarize the data is to calculate the coverage, the enrichment of the reads on windows, or portions, of the genome. This information is stored in the wiggle format (wig), which is compact and displays data at regular intervals. The bigWig file is a derived version of the wiggle, is an indexed binary file and its main advantage is relative to the visualization of sequencing tracks on the UCSC browser. In fact, only the portions of the files needed to display a particular region are transferred to UCSC, so for large data sets bigWig is considerably faster than regular wiggle files. A two step format conversion is needed to make a bigWig file from mapped data, first with the help of BEDtools with genomeCoverageBed (-split -bg -ibam), that converts a bam file to a bedgraph file. Similar to Wiggle files, bedgraph are used to display quantitative data across genomic regions. They use variable length intervals instead of constant intervals found in wiggle files, and are usually a little bigger in size. Subsequently, a utility provided from UCSC Browser itself, called bedGraphToBigWig converts the bedgraph file in bigWig format, ready to be uploaded.

Visualization of mapped reads and peaks

One step that should never be skipped and that is common to all kinds of applications is the visual inspection of the data (Landt, Marinov et al. 2012). Local inspection of the profile of the mapped reads over a genome looking for positive controls (i.e. regions already known to be enriched or expressed in some conditions) helps to assess the quality of the sequencing experiment. Moreover, after the peak finding in ChIP-seq, it is important to look at the actual peaks pointed out by the software. There are many available browsers to visualize the data, but the most used ones are the UCSC Browser (Kent, Sugnet et al. (2002); http://genome.ucsc.edu/) and the IGV browser (Integrative Genomics Viewer; Robinson, Thorvaldsdottir et al. (2011); Thorvaldsdottir, Robinson et al. (2013)). The main difference between the two is that the first is on-line, virtually unlimited in terms of hosted data (from many international projects) and works at a good speed. The second, instead, is handy because is simple to use and runs locally on personal computers. Its limit is that it needs some RAM to run smoothly and it only contains user's uploaded data. On both browsers it is possible to upload many different file types, but the most used for ChIP-seq and RNA-seq are the bed file (for peaks) and the wiggle (or bigWig) for a graphic dense distribution of sample enrichment. Thie bigWig file is big enough that UCSC browser doesn't allow to upload it directly on their servers, thus an http, https or an ftp server is necessary to store the data in order to be visualized. UCSC Browser will only temporarily store the data that is visualized from the user. Small bed files, like the ones of the peaks, can be uploaded directly from the personal computer to the server.

The bigWig files have to be uploaded with a track line like this on the upload page of the browser:

```
track type=bigWig name="ctrlSample" description=" ctrl Sample"
bigDataUrl=http://123.456.78.910/~FTP/ctrl_sample.bw
color=255,0,0
```

This indicates the type of track, the description and name of the track that will be visualized on top and side of the track, the place where the file is stored and the color that we want to assign to the track (in RGB scale). For the bed files, instead, a track line can be pasted as header of the file and then the file can be uploaded from the upload button. Example:

```
track name=sample1_peaks description="Sample1 peaks"
color=255,0,255
```

In Fig 11 there is an example of a bigWig file and a bed file representing a histone modification.

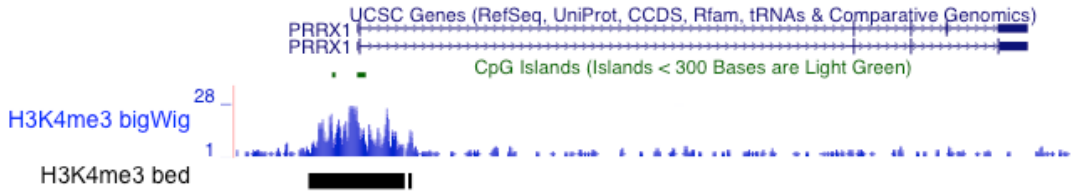

Figure 11. Screenshot of H3K4me3 profile on UCSC browser. From the top the representation of a gene, with arrows indicating the direction of transcription. In green bars predicted CpG island from UCSC database. In blue the bigWig file representing the enrichment of the histone modification. In black the BED file depicting the peak detected on the gene TSS.

Quality control

Assessing the quality of the sequence files is important in order to be able to rely on the results extracted from them. For this process two different tools have been used, one developed at the Istituto per le Applicazioni del Calcolo (IAC-CNR), for SOLiD data, and the other for Illumina data is the FastQC tool (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/). The quality assessment is done using the quality codes in the raw files, translated to quality statistic reports and plotted as boxplots (see examples in Fig 12 for SOLiD (a) and Illumina (b) data).

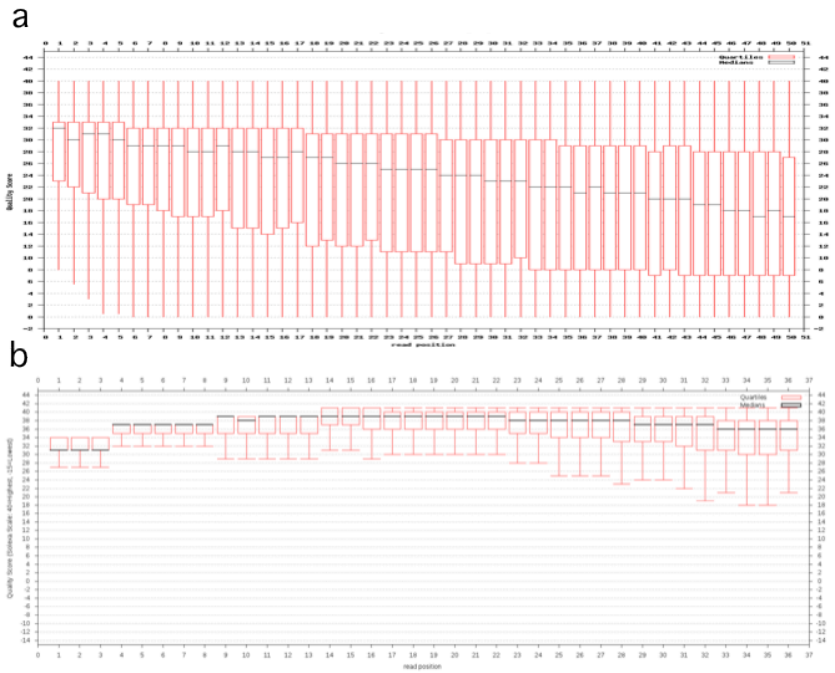

Figure 12. Boxplot representation of reads quality from **a.** SOLiD, **b.** Illumina raw files.

The range is different because of the different scales used by the machines; both the results shown are considered to be in a good range.

The utility of looking at the single base quality in the reads stays in the ability we have to trim the reads if the quality drops suddenly towards the end of the sequence for most of them (i.e. due to a technical problem). This helps increasing the mapping quality.

The quality control is the same for all kind of sequencing data (even if technology-dependent), but the consequent steps, starting from the mapping, become application-specific.

ChIP-Seq

RNA- and ChIP-seq follow two different data analysis paths that can be downstream integrated. The pipeline in Fig 13 represents the principal steps of the analysis performed in this work, which will be further discussed below more in depth. For the analysis of the repetitive sequences the workflow deviates from the principal one and it will be discussed in the dedicated section.

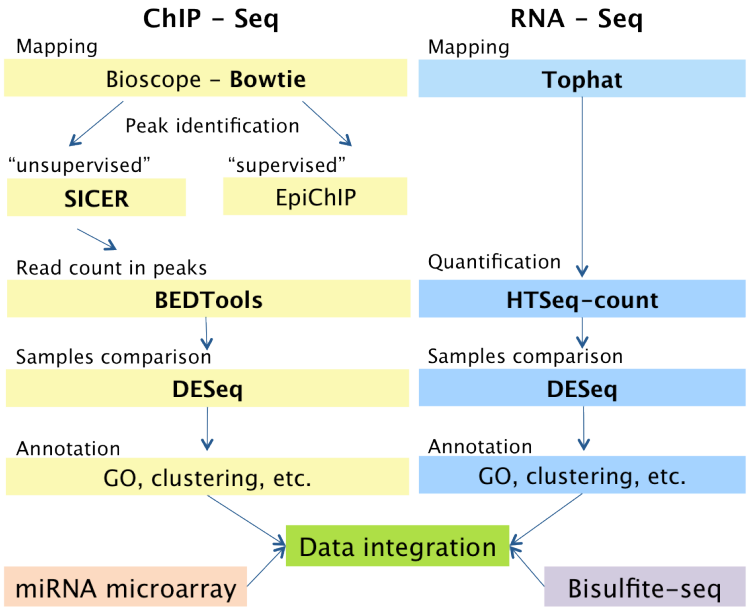

Fig 13. Data analysis pipeline for ChIP-seq and mRNA-seq. In the boxes, in bold are indicated the tools of choice.

- **Mapping**

The alignment of the reads to the genome is not a trivial process. It is the fundamental step that has to accurately assign the sequenced reads to the right position in the genome where they were generated.

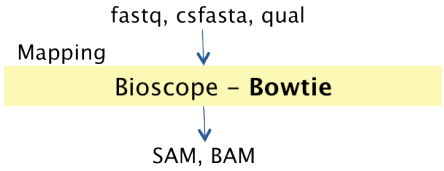

Fig 14. ChIP-seq mapping step. The inputs for Bowtie and BioScope are fastq or csfasta and qual samples and in output are SAM or BAM files.

There are many programs nowadays with the sole purpose to map short sequences to the genomes. Old aligners, like BLAST, do not ensure adequate performance anymore, as they are not efficient enough to map millions of short sequences (50bps) to very large genomes (many Gb). Moreover possible mismatches have to be considered and with the new tools there may be some loss of sensitivity, with gain of mapping capacity. Some of those new alignment tools are provided by the companies that produce the platforms (i.e. BioScope from Life technologies and ELAND from Illumina), and come with the machines; some others are free software developed by users that can be downloaded from the Internet and freely used (like bowtie, BWA, MAQ, SOAP and many others).

Two programs were tested and compared to map the ChIP-seq reads to the reference genome (GRCh37/hg19) (Fig 14). The test was performed on one of the ChIP-seq performed on SOLiD machine to be able to choose the best one for further experiments. The first used was BioScope™ Software (SOLiD system integrated mapper, Life technology), which consists of a framework and a group of tools. The advantage of BioScope™ is that it is specifically designed to map color space reads to the genome, using all information from this technology. It has a mismatch control that takes advantage of the SOLiD chemistry, considering two neighboring mismatches as one. On the other side, it can't map base space sequences. The other software used was bowtie (Langmead, Trapnell et al. 2009), a short read aligner designed to be ultrafast and memory-efficient. It aligns short DNA sequences (reads) to the human genome at a rate of over 25 million 35-bp reads per hour. bowtie indexes the genome with a Burrows-Wheeler index to keep its memory footprint small: typically about 2.2 GB for the human genome (2.9 GB for paired-end). There are many advantages of the use of this tool, from the speed, to the number of different input formats you can submit, to the frequent updates that are performed by the maintenance team. Moreover, nowadays, it is one of the most used mapping software in literature.

Each of the two programs has many parameters to set, but, while bowtie only requires some options passed through command line to be launched,

BioScope use includes the manipulation of four main files (globalC.ini, mappingClassic.ini, matobam.ini and Classic.plan) in order to set its parameters. This, of course, makes BioScope usage more complex.

BioScope outputs a ".ma" file, that is immediately converted to a ".bam" file. bowtie, on the other hand, has its own output format (one alignment per line - each line being collection of 8 fields separated by tabs) but can also convert the output into SAM format (declaring the option -S/--sam).

In table 6 is a comparison of mappings of the same samples with bowtie or BioScope.

H3K4me3			Bowtie		BioScope	
Cell line	IP/input	# usable reads	reads with at least 1 reported alignment	% mapped	reads with at least 1 reported alignment	% mapped
RC	A1	43.404.761	20.339.822	46,86	17.635.608	40,63
	A2	41.228.459	18.075.321	43,84	16.100.500	39,05
ICF	B1	33.862.814	18.416.154	54,38	16.069.572	47,45
	B2	42.081.420	17.428.286	41,42	15.130.599	35,96
	C1	26.938.754	14.924.792	55,40	12.603.483	46,79
	C2	40.045.574	19.052.267	47,58	16.400.228	40,95
UC	D1	38.449.035	19.762.821	51,40	16.822.183	43,75
	D2	31.703.327	16.460.339	51,92	14.183.943	44,74

Table 6. Mapping results from bowtie and BioScope tools for the same samples, sequenced in duplicates (A1-2, B1-2, etc). Percentages of uniquely mapped reads over the total reads are in % mapped column.

In the "# usable reads" column I reported the reads that passed the quality control in the primary analysis and are considered usable for subsequent analyses. The mapped reads indicated are the uniquely mapped reads and the percentage of the usable reads.

It is noticeable, then, that bowtie gives a higher percentage of mapped reads than BioScope, probably because of its more relaxed definition of uniquely mapped reads. For the purposes of our analyses we prefer to have more aligned reads with some mismatches due to technical errors than less, perfectly aligned reads. Moreover bowtie is significantly faster in analysis than BioScope. All experiments have been performed in duplicates and both replicates have been sequenced.

To compare the results from the two aligners, the correlation between the counted the reads in windows of 1000bps in the duplicates within one analysis and the same sample in the two analyses was calculated. The results are shown in Fig 15.

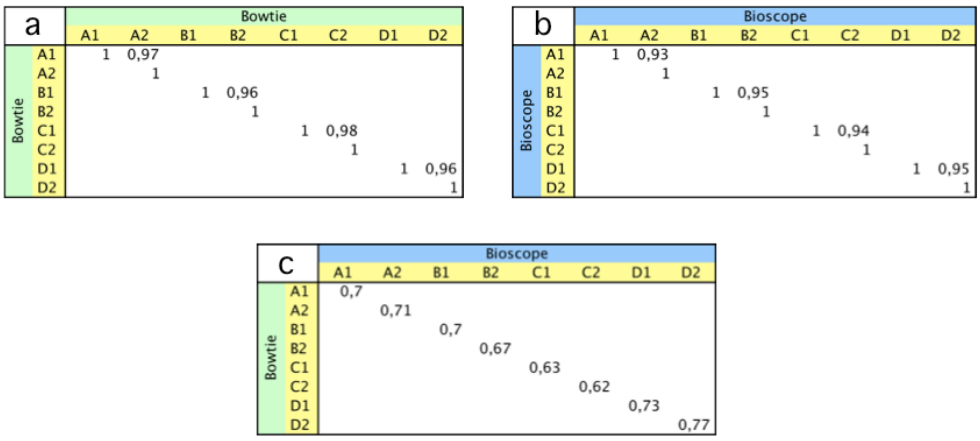

Figure 15. Measurement of the consistency of two different mapping tools on real data. The numbers in tables are the Pearson's correlation values between the counts of mapped reads in 1Kb windows of the genome coming from BioScope or bowtie. **a.** Pearson's correlation between the same samples (r=1) and the duplicates (r<1) analyzed with bowtie. **b.** Same as a. but with BioScope. **c.** Comparison of the two tools.

A good consistency of the results among the duplicates within the same analysis was observed (Fig 15 a-b), but the correlation degree reduced when comparing the two analyses (Fig 15 c). This was of course expected, and it does not tell that the results are not comparable, but it shows the algorithm variations between the two tools. bowtie was, therefore, the selected tool for mapping mainly because of the need to compare results from Illumina and SOLiD sequencing, not allowed by BioScope.

Clearly, the comparison of the tools could have been done in a more comprehensive and sensitive way, using, for example, SEAL (SEquence ALignment evaluation suite), a comprehensive sequencing simulation and alignment tool evaluation suite (Ruffalo, LaFramboise et al. 2011). This kind of more intensive and specific analysis, though, was not considered among the aims of this work, so we made a choice based exclusively on our data and also general practice from the literature.

In bowtie the "-n" mode of alignment was used, with maximum 2 mismatches allowed in the first 28 bases of the sequence. The software was then restricted first to report only those reads having up to 3 alignments on the genome and then only the first best mapped read, with the best "stratum" (those having mismatches just in the "seed" portion of the alignment). In this way the output contains only those reads defined as uniquely mapped with a less stringent criteria.

All results are listed in Table 7.

H3K4me3

Cell line	IP/input	# usable reads	reads with at least 1 reported alignment	% mapped
RC	IP	43.404.761	20.339.822	46,86
	IP	41.228.459	18.075.321	43,84
	input	34.933.878	17.442.718	49,93
ICF	IP	33.862.814	18.416.154	54,38
	IP	42.081.420	17.428.286	41,42
	input	26.938.754	14.924.792	55,40
	input	40.045.574	19.052.267	47,58
UC	IP	38.449.035	19.762.821	51,40
	IP	31.703.327	16.460.339	51,92
	input	36.149.97[illegible]		

H3K27me3

Cell line	IP/input	# usable reads	reads with at least 1 reported alignment	% mapped
RC	IP	45.801.442	23.469.598	51,24
	IP	51.179.186	25.779.123	50,37
	IP	53.801.720	27.053.659	50,28
	input	42.543.895	21.259.411	49,97
ICF	IP	53.125.153	30.371.864	57,17
	IP	43.986.693	24.074.622	54,73
	IP	52.187.358	29.981.036	57,45
	input	44.965.575	22.329.826	49,66
UC	IP	53.846.306	31.546.397	58,59
	IP	48.344.629	27.132.776	56,12
	input	46.809.807	23.164.958	49,49
		[illegible]2	20.491.355	44,47

H3K9me3

Cell line	IP/input	# usable reads	reads with at least 1 reported alignment	% mapped
RC	IP	27.119.386	18.801.538	69,33
	IP	32.162.002	22.588.033	70,23
	input	35.857.042	26.676.985	74,40
ICF	IP	28.379.927	20.800.309	73,29
	IP	31.705.829	21.581.134	68,07
	input	33.806.273	29.330.207	86,76
UC	IP	32.907.679	24.444.756	74,28
	IP	31.702.806	22.080.095	69,65
	input	33.378.782	28.507.670	85,41

Table 7. Mapping results from bowtie for ChIP-seq data for H3K4me3 and H3K27me3 from SOLiD and H3K9me3 from Illumina. Reads with at least 1 reported alignment are the uniquely mapped reads in this case.

Two more steps are necessary to proceed to the peak calling after the mapping: the first step eliminates duplicate reads, leaving to the following analyses only one read for each mapping start site. With this expedient, miscalculations due to amplification biases of the PCR step of the library preparation can be avoided (see Materials and methods for libraries preparation details). This can be easily accomplished with SAMtools with the tools view -bS (to convert from SAM to BAM) and rmdup -s (to remove duplicate reads in BAM file). Once obtained this last, polished, file, the second step is to convert the bamfile to BED file, which most alignment tools require (even though some use BAM, too) (Fig 16).

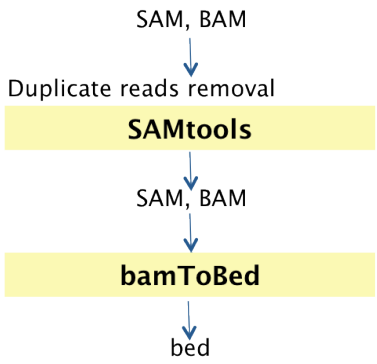

Fig 16. File polishing and conversion prior to peak calling.

The H3K4me3 and H3K27me3 ChIP-seq have been performed with the SOLiD4 machine at the Institute of Genetics and Biophysics (IGB-ABT) in Napoli, while the H3K9me3 ChIP-seq has been sequenced with Illumina HiSeq in NCMLS in Nijmegen (Nijmegen Center for Molecular Life Sciences) and at the IGA (Institute of Applied Genomics) in Udine within the Epigen consortium.

- **Peak calling**

The second and most important step of the ChIP-seq analysis is the detection of regions of the genome significantly enriched with sequences due to the protein binding (commonly called "peak finding"). The peak will represent the site of the genome that is bound by the immunoprecipitated protein and therefore carries more sequences than it would if it was just casual enrichment. The critical point of the analysis of these data is to be able to distinguish the significant increase of sequence read tag density along the genome of these regions compared to the random enrichment that can occur in the background. The background can either be the input from the ChIP, or the negative IP, performed in parallel with an aspecific antibody. It is also possible to estimate the background in silico, but this is not the best direction to take for clean data. During the last years many developers have produced their own tools to make the peak calling operation automated. There are many difficulties in standardizing this process, among which are the different profiles of the peaks characterizing different protein bindings on the DNA. Transcription factors usually show more peaked regions, spanning few hundreds of bases, while histone proteins, mainly the ones correlated to heterochromatin, cover bigger regions, with less defined profiles. Moreover, depending on the goal of the experiment, the "supervised" and "unsupervised" approaches can be taken into account. The unsupervised approach is the one I have described before, that is the search for peaks without previous assumptions about where the peaks should be located. The supervised approach, instead, looks for enrichments in specific positions of the genome, for example, transcription start sites of genes (TSS). To see which could be revealed to be the best approach for me I took into closer consideration two programs: EpiChIP (Hebenstreit, Gu et al. 2011) and SICER (Zang, Schones et al. 2009). Both these tools are focused on the analysis of histone modifications, therefore seemed to be appropriate for my analysis workflow.

EpiChIP is centered on the correlation of the histone modifications with the genes and their expression. It quantifies the enrichment of the histone binding on a defined portion of the gene (TSS, exon, intron, etc.) and also makes a distribution-based distinction between noise and signal.

On the other side, SICER, based on the biological observation that histone modifications tend to cluster to form domains, identifies spatial clusters of signals unlikely to appear by chance. SICER computes probability scores in non-overlapping windows, and then aggregates windows into 'islands' of sub-threshold windows separated by gaps in order to capture broad enrichment regions (Fig 17). The advantage of using SICER is more than one. First of all, it takes into consideration the fact that the peaks are not supposed to be narrow and separated one from each other (it allows to define the gap in which the peaks have to be considered joint). Moreover, it does not look for peaks only in defined regions, as genes TSS, but outputs a complete list of statistical significant peaks in all genome. These peaks can later be associated to the closest genes. The cons of both methods is that they do not consider duplicates in the analyses, thus each sample has to be analyzed separately or the two samples have to be pooled together. SICER, anyway, takes into account the possibility of pooled samples, allowing a redundancy threshold of the reads equal to two. It also allows samples comparisons, but without the use of duplicates.

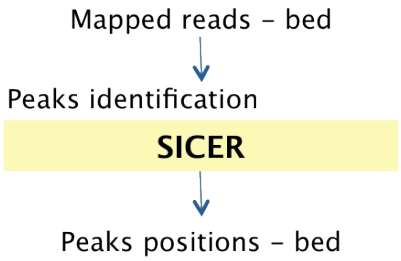

Fig 17. Peak calling with SICER. Input files contain the mapped reads of the ChIP sample and the negative control sample (input). The output is a bed file of the significantly enriched peaks in the ChIP compared to the input.

SICER was the software of choice to identify the enriched islands in H3K4me3, H3K27me3 and H3K9me3 samples for its major pertinence to the analysis of histone domains. The fragment size we selected while preparing the libraries was used as a parameter, and different windows and gaps were chosen depending on the histone modification. FDR was equal to 1e-5 for all of them (False Discovery Rate).

Histone modification	Fragment lenght (bps)	Window	Gap	FDR
H3K4me3	200	200	200	1e-5
H3K27me3	200	1000	3000	1e-5
H3K9me3	350	200	400	1e-5

Table 8. SICER peak calling used parameters for different histone modifications.

In table 8 are the parameters I used, in the Results – Part II chapter are the results.

SICER detected an approximate average of 7000 peaks for H3K27me3, 29000 for H3K4me3 and 25000 for H3K9me3. Only the peaks present in both duplicates were considered for further analyses.

Both output and input from SICER are bed files, so either BEDTools or R (R Development Core Team (2012). R: A language and environment for statistical computing. URL http://www.R-project.org/) were used for all further manipulations.

- **Peaks comparison with DESeq**

In this project it was very important to be able to distinguish quantitative increases or decreases of the same histone modification in the patient's sample compared to the control, due to the nature of the pathology and the previous knowledge in literature (see results – Part II). Most projects only require identifying peaks that are appearing of disappearing in different biological conditions in their ChIP-seq experiments. For this reason there are not many tools that support a statistical approach to verify the quantitative differences between the samples and there is no consensus on how it should be accomplished.

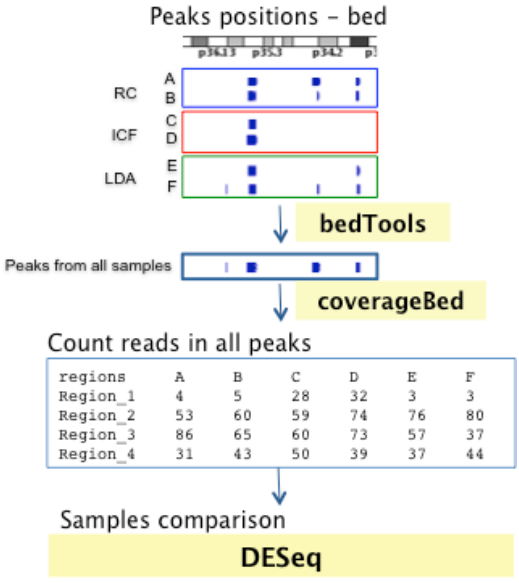

Fig 18. Peak enrichment quantification and comparison between samples.

To assess the enrichment differences among the different samples DESeq (Anders and Huber 2010) is a tools that seems to have a good performance. It is an R package for analyzing count data from high-throughput sequencing assays and tests for differential expression or enrichment. This package needs a list of regions (or genes) as input and the count of reads that map in them. It calculates the variance for each gene (or region) using the duplicates variability and the uncertainty in measuring a concentration by counting reads, known as Poisson noise, which is the

dominating noise source for lowly expressed genes. The sum of both, shot noise and dispersion, is considered in the differential expression inference. The method to test for differential expression or enrichment uses the negative binomial distribution.

For the experiments described in this work the pipeline in Fig 18 was developed. A bed file was created with all the possible peaks for each modification, merging all peaks from all the samples (`intersectBed` for the duplicates, then `cat` for all the intersected files, `sortBed` and `mergeBed` to eliminate replicate regions). Then the reads spanning the regions for all samples were counted (with `coverageBed`). Those count files were assembled into a matrix of counts to input to DESeq on R (see code in Appendix A). Running DESeq and only differentially enriched regions with p-value<0.05 were considered. In Fig 19 there is the MA-plot, representing in red the differentially enriched genes comparing two samples.

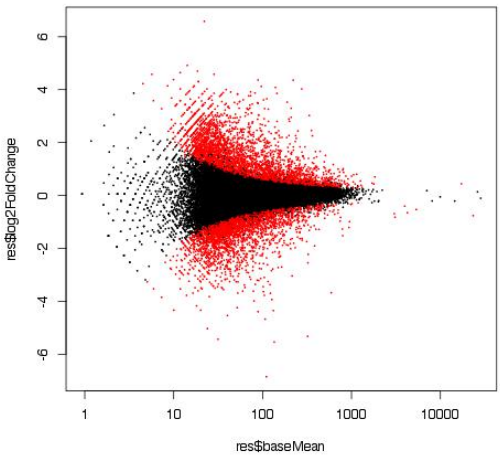

Fig 19. Example of MA-plot of differentially enriched peaks. On x axis there is the mean of the enrichment, on y axis the log2 of the fold change between the two compared samples. Red dots have pval<0,01.

After this first selection of regions other two filters were applied, to make the list more reliable and eliminate some false positives and false negatives. Among the regions more enriched in one sample compared to the control in DESeq analysis were eliminated those that in the first place were not considered enriched from SICER in that specific sample (3-10%). Then I added to these regions those that, making a simple intersection of the peaks data were considered present in one sample and not in the other from the firt SICER analysis. Thus the final list of differentially enriched peaks was obtained.

- **Peaks annotation and Gene Ontology**

Once the enriched regions are defined and differences between the samples spotted, it is necessary to correlate those regions to specific features of the genome, as genes, CpG islands, and so on.

Again, for this task, many tools are available and the most diffused will be described. The first is again an R package, ChIPpeakAnno (Zhu, Gazin et al. 2010), that is very flexible, allowing to associate not only regions to genes, but also to any list of features the user is interested in comparing. Another, more simple, tool is PeakAnnotator from PeakAnalyzer (Salmon-Divon, Dvinge et al. 2010), developed in Java and C++ in the Sanger institute and widely used in literature; it allows very simple operations with graphical outputs and with a user-friendly interface. The third tool that is worth a mention is GREAT (Genomic Regions Enrichment of Annotations Tool; McLean, Bristor et al. (2010)), that integrates the association of regions to genes and the gene ontology of the subset of genes detected. The basic and innovative idea of this tool is that it models the cis-regulatory landscape through the use of long-range regulatory domains and a genomic region-based enrichment test, allowing analyses that take into consideration the large number of binding events that occur far beyond proximal promoters.

ChIPpeakAnno was mainly used for analyses in this work, as there was more interest in correlating peaks to genomic features other than genes. A simple script I wrote in R allows invoking ChIPpeakAnno directly from the command line to annotate peaks on hg19 genome (see Appendix A). For other annotations features, like CpG islands, sites of hyper- and hypo-methylation and miRNAs this script was modified removing the step of data extraction of the human genome and creating IRange objects from other specific lists of features.

The association between genes and histone modifications peaks was observed with respect to the TSS of the genes. For H3K4me3 were considered associated those peaks within 5Kb upstream or downstream of the TSS. For H3K27me3 and H3K9me3 this number was increased to 10Kb as those modifications show broader regions on the TSS.

The distribution of the peaks on specific features of the genome, like 3' UTR, 5'UTR, introns, exons and not genic regions was performed with

PeakAnnotator from PeakAnalyzer, with all coding and non coding genes. The results from annotation will be discussed in the next chapter, Results-Part II.

The following step, after correlating the peaks to the genes, is to infer the functional classes in which these genes are more enriched. This information can be crucial to understand which biological functions are altered in the studied system and which gene categories. There are many tools that, given a list of genes, can calculate their enrichment in cellular components, functional categories or biological processes (see Huang da, Sherman et al. (2009) for a complete overview of the available tools).

DAVID (the Database for Annotation, Visualization and Integrated Discovery) (Huang da, Sherman et al. 2009), is the tool used in this work for gene ontology. DAVID is able to extract biological features/meaning associated with large gene lists. It provides typical batch annotation and gene-GO term enrichment analysis to highlight the most relevant GO terms associated with the gene list. It assembles data from more than 40 annotation categories and calculates the enrichment of gene lists cross-referencing the different categories. It evaluates the enrichment with a modified, more stringent, Fisher's exact test (EASE).

mRNA-seq

- **Mapping - TopHat**

Transcriptome analysis has multiple functions, broadly divided between transcript discovery and mapping on the one hand and RNA quantification on the other. This work is centered on transcriptomes from human cell lines and my interest is mainly focused on RNA quantification and comparison among samples. While mapping for ChIP-seq sequence tags is fairly simple, because it is made of fragments of genomic DNA, RNA-seq mapping is more complicated, being the mRNA subject to introns removal during the splicing process. For this reason it is necessary to take into account those reads covering the junctions between two exons, which result not mappable on the genome. Known splice junctions, based on gene

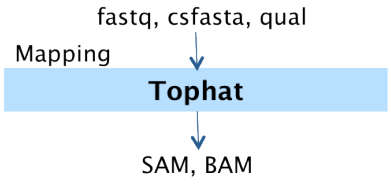

Fig 20. RNa-seq mapping step. The input for TopHat can be fastq or csfasta and qual samples and in output are SAM or BAM files.

models and ESTs can be handled by incorporating them computationally in the primary read mapping, whereas newly inferred junctions are considered later.

TopHat (Trapnell, Pachter et al. 2009) is a fast splice junction mapper. It aligns mRNA-Seq reads to mammalian-sized genomes using the ultra high-throughput short read aligner bowtie, and then analyzes the mapping results to identify splice junctions between exons (Fig). By first mapping RNA-Seq reads to the genome, TopHat identifies potential exons, since many RNA-Seq reads will contiguously align to the genome. Using this initial mapping information, TopHat builds a database of possible splice junctions and then maps the reads against these junctions to confirm them. It is widely used and cited in literature and updates are released with high frequency.

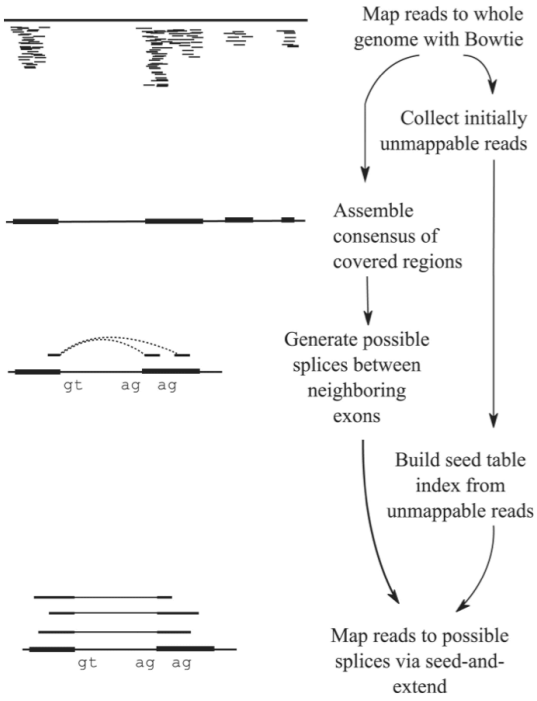

Fig 21. TopHat pipeline. RNA-Seq reads are mapped against the reference genome; reads that do not map are set aside. An initial consensus of mapped regions is assembled. Sequences flanking potential donor/acceptor splice sites within neighboring regions are joined to form potential splice junctions. The IUM reads are indexed and aligned to these splice junction sequences. From Trapnell, Pachter et al. (2009)

TopHat outputs three result files, a bam file with all the mapped reads (accepted_hits.bam) and three UCSC bed track files with the junctions, the insertions and the deletions. Moreover it outputs summary files:

```
less left_kept_reads.info
        min_read_len=50
        max_read_len=50
        reads_in =74262816
        reads_out=74139815

less bowtie.left_kept_reads.fixmap.log
        # reads processed: 74139815
        # reads with at least one reported alignment: 37156205 (50.12%)
        # reads that failed to align: 27858377 (37.58%)
        # reads with alignments suppressed due to -m: 9125233 (12.31%)
        Reported 37156205 alignments to 1 output stream(s)
```

With some default parameters except segment length = 17 (minimum length of the segments in which the read is cut in order to map it correctly) and max multi-hits = 1 (number of mapping sites allowed for the read to be considered uniquely mapped). In Table 9 are summarized the results from the samples used in this work.

RNA-seq

Cell line		reads	# left reads with at least one reported alignment	% mapped
RC	Single end	31.993.108	18539711	57,95
	Single end	36.967.986	20804183	56,28
ICF	Single end	30.156.656	18694863	61,99
	Single end	45.385.629	25242706	55,62
UC	Single end	37.244.678	18701934	50,21
	Single end	50.060.476	27319311	54,57

Table 9. Mapping results from TopHat on ICF, RC and UC samples.

- **Estimating differential gene expression – HTseq and DESeq**

One of the most diffused aims of works with transcriptomics data is to detect genes with differential expression in different samples or treatment conditions. To address that, there is the need to correctly quantify the expression of each gene in order to be able to compare them in an unbiased manner. The normalization of these data is still a matter in discussion; many methods have been proposed to correctly account for samples technical and biological variations. One of the first proposed methods was the calculation of the RPKM (Reads Per Kilobase per Million mappable reads; Mortazavi, Williams et al. (2008)). It takes into account that the number of reads from a gene is a function of the length of the mRNA as well as its molar concentration. RPKMs for genes are then directly comparable within the sample by providing a relative ranking of expression. Other methods such as Tags/Transcripts per million (TPM) and new metrics such as FPKM (Trapnell, Williams et al. 2010), per-lane upper quartile correction metric (UQUA) (Bullard, Purdom et al. 2010), and trimmed mean of M values (TMM) (Robinson and Oshlack 2010) have been developed to compare expression levels both between and within samples.

The count nature of the next-generation data has necessitated the development of new algorithms (or rediscovery of SAGE analysis techniques)

to accurately estimate differential expression. The early RNA-seq papers frequently used the Poisson model to identify differentially expressed genes. This approach has increasingly been recognized as inappropriate. The most commonly used methods have been parametric methods utilizing variants of the negative binomial distribution such as edgeR (Robinson, McCarthy et al. 2010), DESeq (Anders and Huber 2010), and bayseq (Hardcastle and Kelly 2010). Nonparametric methods such as NOISeq (Tarazona, Garcia-Alcalde et al. 2011) and Samseq (Li and Tibshirani 2011) and expectation-maximization methods such as RSEM (Li and Dewey 2011) have also been applied to this problem. The Fisher Exact Test (FET) also performs well in some comparisons. No consensus has yet emerged as to the best algorithm or pipeline to use (Bullard, Purdom et al. 2010; McGettigan 2013). Therefore in this work DESeq was chosen for the consensus it receives in the field and for the ability to efficently use duplicates.

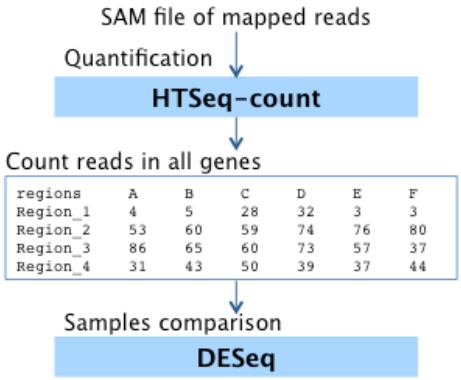

Fig 22. Gene expression quantification and comparison between samples.

The R package DESeq is a parametric method, derived from edgeR, that uses negative binomial distribution as a model for differential expression analysis (check *Peak comparison with DESeq* paragrah). This package expects count data in the form of a matrix of integer values. Each column corresponds to a sample, the rows correspond to the entities for which you want to compare coverage, in this case genes. The counts can be derived with the use of htseq-count script distributed with HTSeq (http://www-huber.embl.de/users/anders/HTSeq/doc/count.html#count). The script used to quantify differential gene expression is the same used for ChIP-seq samples (see *Peak comparison with DESeq* paragraph and Appendix A).

Association of mRNA-seq, ChIP-seq, Bis-seq and miRNA microarray data

Given the complex nature of the ICF phenotype and the interplay among all the epigenomic regulators that is clearly disrupted in this pathology, it

resulted interesting to integrate all epigenetic and transcriptional data from these cell lines. Differentially expressed genes were checked for differences in enrichment in histone modifications around their TSS. Differentially enriched peaks for H3K4me3 were searched +/-5Kb around the TSS of genes, while broader modifications like H3K27me3 and H3K9me3 were searched +/-10Kb around the TSS.

CpG islands are mostly positioned close to gene TSS, influencing gene expression, therefore genes were considered CpG rich when a UCSC annotated CpG island was found +/-5Kb around its TSS. DNA methylation changes detected in Heyn et al., 2012, were integrated to genes expression and histone modifications changes using the differentially methylated regions (DMRs) defined as regions of at least five consistently differentially methylated CpG sites between the control and ICF sample. These hyper- and hypo-methylated regions were physically correlated to genes TSS and histone modifications binding sites as described in the next chapter.

From previous miRNA expression profiling carried out through microarray platform (Gatto et al., 2010), 40 microRNAs were detected as differentially expressed comparing the same ICF cell line and control (RC) used in this work; here, their predicted or validated TSS have been associated to changes in DNA methylation and histone modifications enrichment. The results are shown in Results -Part II.

Transcription and histone methylation at repetitive sequences

In order to analyze the H3K4me3, H3K27me3 and H3K9me3 profile at repetitive regions and their transcription it was not possible to follow the "classic" pipeline, because of the intrinsic nature of these genomic regions. It is important to take into account also the multiple mapped reads, left out from standard analysis. Few experiments of this type have been performed and there is no standard for this kind of analysis (Maze, Covington et al. 2010), therefore a new pipeline was built (Fig 23).

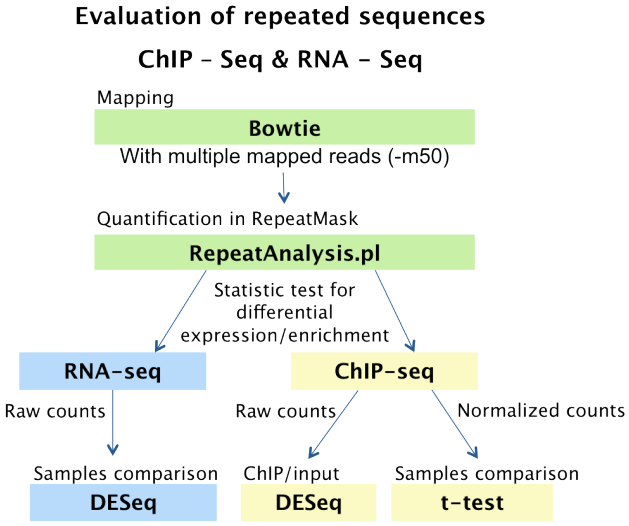

Figure 23. Data analysis pipeline for ChIP-seq and mRNA-seq on genomic repetitive sequences. In the boxes, in bold are indicated the tools of choice.

- **H3K4me3, H3K27me3 and H3K9me3 enrichment**

It is known that in ICF syndrome cell lines the most striking characteristic is the hypomethylation and decondensation of the pericentromeric regions, mainly in Sat2 and 3 (Jeanpierre, Turleau et al. 1993), that leads to chromosomal rearrangments. The telomeric regions too are affected in ICF syndrome, being hypomethylated, failing to establish proper heterochromatin and influencing transcription (Sawyer, Swanson et al. 1995; Deng, Campbell et al. 2010). For these reasons it resulted very interesting to analyze all kinds of repetitive regions with differential enrichment of histone modifications.

As the repeated sequences are interspersed in the whole genome, this causes all the reads coming from those regions to fall into the multiple mapped reads and to be excluded from the analysis. Therefore re-mapping the sequenced reads with bowtie was necessary, this time raising the number of allowed multiple mapped reads from 3 to 50. Adjusting this parameter, were included also those reads probably falling in repetitive regions (in this case, mapping in up to 50 places in the genome) that could be missed in the first place. In any case, I only kept the best mapping site in order to be able to correctly quantify them. Subsequently, the reads falling in repetitive regions were counted. A perl script that was developed in Henk Stunnenberg's lab in Nijmegen (RepeatCount.pl, see Appendix A) counts the reads in the UCSC RepeatMasker collection of interspersed repeats and low-complexity DNA sequences (http://www.repeatmasker.org). The repetitive

regions annotated in RepeatMasker are classified as Repeats, Families and Classes. Each category groups elements of the previous one. The script outputs the counts for each element in all the categories, either raw or normalized by the total number of reads. To assess whether these categories were differentially enriched among the samples two parallel paths have been followed. First it was tested whether the repeat (or family or class) was more enriched in the sample than in the input. Then it was verified if those repeats showed differential enrichments comparing the samples with a t-test.

The raw count nature of the repeats allowed the use of DESeq once again, testing the differences between the ChIP samples and the inputs, using the duplicates. Then a Student's t-test was performed between the log ratio of the normalized counts between the ChIP sample and its input, using the duplicates. Those repeats, families or classes whose pvalue in the t-test was < 0.05 and log ratio $> |1.5|$ were considered differentially enriched. Moreover, these features had to show a p-value from DESeq analysis < 0.05.

Those differences were represented plotting the mean log ratio for each sample and its standard deviation.

- **Transcriptional profile**

Roughly half of the human genome is comprised of repetitive elements; these elements range from the 6Kb LINE1 to micro and minisatellites (Richard, Kerrest et al. 2008). The biological role of repetitive elements is not known and in general they are believed to be nonfunctional sequences. It is known that repetitive elements in the DNA are expressed, but their profile remains largely uncharacterized. Very few transcriptomic experiments have been performed to analyze the expression of repetitive elements (Tyekucheva, Yolken et al. 2011) and no best practice exists for those. The mapping of the RNA-seq reads was performed with bowtie, again (as for ChIP-seq reads) with 50 allowed multiple mapped reads. The reads falling in the RepeatMasker repeats, families and classes where then counted and DESeq was used to assess the statistical significance of the difference between the enrichments in the different samples. Results are represented as mean and standard deviation of normalized counts (see results - Part II).

4. Results – Part II – ICF cells epigenomic profile

In this work we characterized the transcriptomic and epigenomic landscape of lymphoblastoid cell lines (LCLs) isolated from one ICF (Immunodeficiency, Chromosomal instability and Facial anomalies) patient (GM08714, from now on called ICF) and two control subjects, GM08728 (ICF patient's mother, related control, RC) and LDA (normal subject, unrelated control, UC).

High-throughput ChIP sequencing (ChIP-seq) was performed on those cells to observe the binding of three different histone modifications, such as H3K4me3, H3K27me3 and H3K9me3. Moreover, RNA-seq has been performed on them. Each experiment has been performed at least in duplicates on Illumina and Life technologies platforms. We obtained an average of 16-31·10^6 mapped reads for the ChIP experiments and 18-27·10^6 for the RNA-seq.

For ChIP-seq analysis of enriched regions we mapped the reads with bowtie (Langmead, Trapnell et al. 2009), then we run SICER (Zang, Schones et al. 2009) for peak finding and counted the reads in each peak with bedtools. Subsequently we used DESeq (Anders and Huber 2010), an R package that tests for differential expression by use of the negative binomial distribution and a shrinkage estimator for the distribution's variance, to detect the differentially enriched peaks among the samples (see Results – Part I for a more detailed explanation of the analysis). Out of the 25-29·10^3 peaks of H3 trimethylated in K4, and K9, and 4-10·10^3 of H3 trimethylated K27 detected in the samples we found many of them differentially enriched comparing the samples (Fig 24). To quantify the differential enrichment we calculated the percentage of the

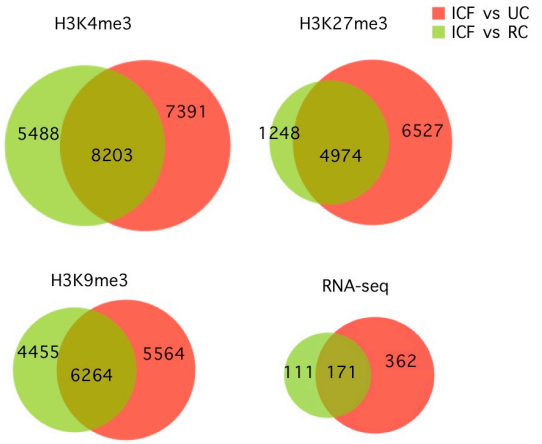

Figure 24. ChIP-seq and RNA-seq differences among the samples. For ChIP-seq (H3K4me3, H3K27me3, H3K9me3) are indicated the number of peaks differentially enriched comparein ICF and the controls, for RNA-seq the genes differentially expressed. In red is the comparison ICF/unrelated controls, in green the ICF/related control.

bases enriched in all samples that is covered by differentially enriched peaks (as we don't consider comparable the numbers of peaks). Following this, we see that 32-36% of the area covered by H3K4me3 peaks is differentially enriched in the two comparisons, and the common differentially enriched peaks represent the 18,7% of the total. For H3K27me3 the changes cover 81-47% of the area and share 37,9% of it. H3K9me3 differences in binding emerge in 61-48% of the total area and the commonly deregulated are the 28%.

Interestingly, the related control (RC) showed always fewer differences with the ICF samples that the unrelated control (UC) as expected. Many differentially enriched peaks, moreover, were shown to be common to the two comparisons (Fig 24), indicating the partially common nature of the controls. RNA-seq differentially expressed genes are also shown in Fig 24, highlighting again the higher similarity of the ICF sample to the RC than to the UC.

Genomic distribution of H3K4me3, H3K27me3 and H3K9me3 in ICF and control cells

The H3K4me3 genomic distribution (Fig 25a), as expected, results more enriched at gene promoters than at other regions (promoter regions defined as regions flanking +/-5Kb the transcription start site - TSS).

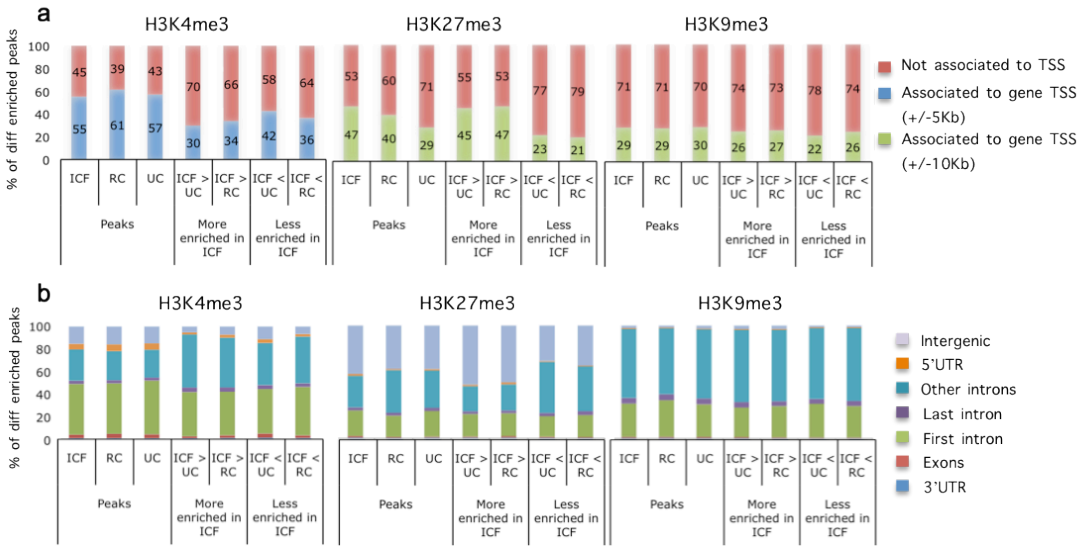

Figure 25. Genomic distribution of methylated histones binding sites. a. Measure of the association of the peak to the TSS of the genes. b. Detailed distribution in 3' or 5' UTR, introns, exons and intergenic regions.

Noticeable is that the differentially enriched H3K4me3 peaks among the samples exhibit a skewed distribution towards the non-promoter regions. In a more detailed classification of the distribution of the peaks (Fig 25b) the regions that score as differentially enriched peaks in ICF compared to the controls mainly reside in the intronic regions. The explanation might be that some of these regions are misclassified and there are TSS not yet reported in the reference genome. Moreover, genes characterized by H3K4me3 peaks, are also associated to CpG islands for the 70% in all samples, which is the distribution in the total genome (Fig 26). The differentially enriched peaks also maintain the same casual distribution, demonstrating that there is no change in distribution of this histone mark correlated to CpG islands methylation.

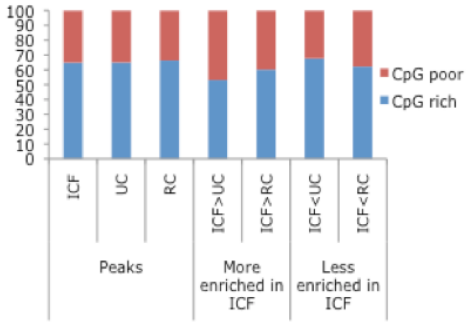

Figure 26. Percentage of CpG rich genes associated with H3K4me3 marks or showing increase or decrease of this mark. CpG rich genes have a CpG island +/-5Kb from their TSS.

Differently from H3K4me3, the repressive histone mark H3K27me3 shows a differential distribution already among the three samples (Fig 25a), where the unrelated control has fewer peaks associated to genes (here we picked a broader region to associate peaks to the TSS, +/-10Kb, due to the nature of the histone modification broader peaks). At the same time, the peaks more enriched in ICF compared to the controls keep the same distribution of peaks in ICF sample alone, while the ones more enriched in the controls are more associated to non-promoter regions. Moreover, the annotation in Fig 25b shows that H3K27 in ICF seems to re-localize from intragenic regions to intergenic regions. In fact, the majority of the peaks more enriched in ICF resides in the intergenic portion. Another interesting variation observed regarding H3K27me3 peaks consists in the slight reduction of peaks width in ICF sample compared to controls, highlighted by the shift of the means of the distributions in Fig 27. Changes in the size of large regions covered by H3K27me3 were also recently described in triple knock out cells for DNMTs (DNMT1, 3a and 3b), where instead they appear to broaden (Brinkman, Gu et al. 2012). Regardless the functional meaning of these two opposite results

reported in two different cell contexts, these findings clearly indicate that there is interdependence between DNA methylation and H3K27me3.

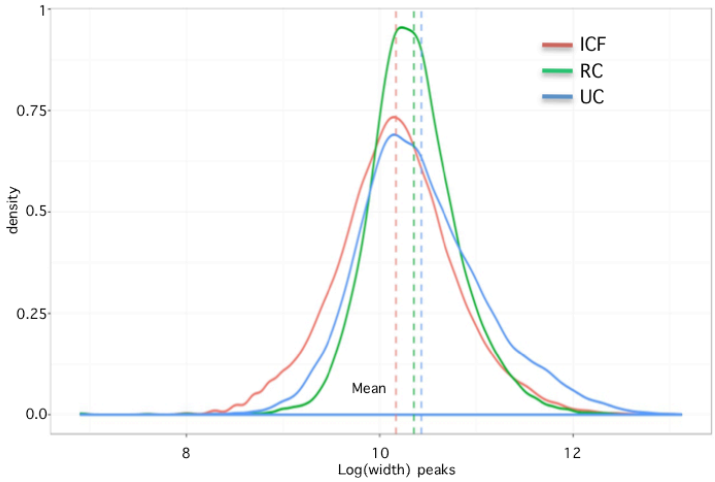

Figure 27. Distribution of H3K27me3 peaks width. On x-axis is the log of the width of the peaks, on y-axis is the density of peaks showing that width. The variation of the width of the peaks among ICF and controls is shown by the shift of the means of the distributions, represented by dashed lines.

Concerning H3K9me3 modification, we found that it is poorly associated to gene TSS (+/-10Kb around it) and covers mainly the gene body, as expected. The localization of this histone modification does not appear to be affected in DNMT3B hypomorphic cells compared to the controls, even when its enrichment changes among the samples. The unexpected, apparent absence of H3K9me3 in intergenic portions of the genome is probably due to the fact that it mainly binds highly repetitive sequences and with this mapping this portion is lost. The extent of H3K9me3 enrichment at repetitive sequences is described below.

Correlation between gene expression and histone methylation profile

By integrating the results of ChIP-Seq and RNA-Seq, we found that the differences in H3K4 and K27 trimethylation level at genes transcriptional start sites (TSS) correlates with the expression of the genes. Indeed, many genes show an increase or decrease of expression that mirrors the increase or decrease of the corresponding histone variation. Remarkable is that only a small percentage of histone marks variations at genes reflects a change in expression (1-6%), while a more consistent fraction of genes with altered transcription shows a modified chromatin marker (5-50%, Fig 28a-b)).

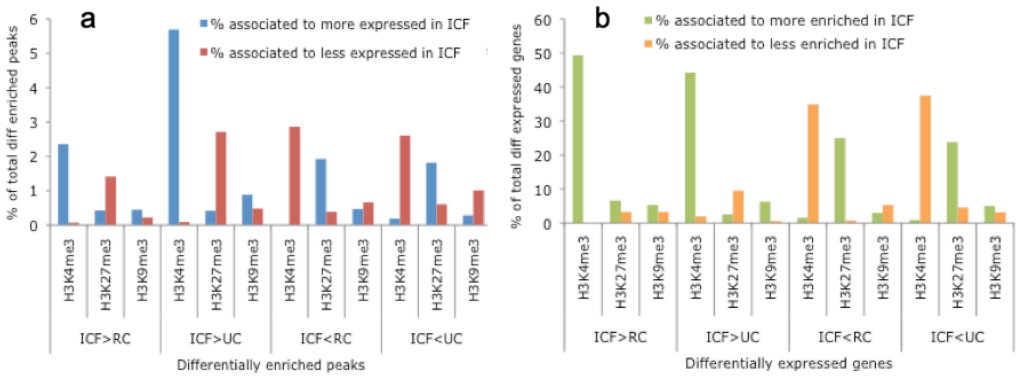

Figure 28. Correlation between variation of histone marks enrichment and variation of expression of the marked genes. a. On y-axis is the percentage of differentially enriched peaks associated to genes with differential expression. b. On y-axis is the percentage of differentially expressed genes associated with differentially enriched peaks.

Functional gene ontology reveals that transcriptionally deregulated genes show enrichment in biological processes like cell adhesion, neuron development, cell-cell signaling, cell morphogenesis involved in differentiation, immune response and cell motility (Fig 29) and the enrichment of these categories does not vary in those marked by changes in histone modifications. The most enriched gene functional categories among the deregulated genes are the immunoglobulins and proteins for cell adhesion and inflammation response (Fig 30). However, the subgroup of genes differentially expressed and marked by variations in H3K4me3 is enriched in the category of zinc finger proteins (ZF), which is also marked by changes in H3K9me3 even though to a lesser extent. It is known that in double knock out cells for DNMT1 and DNMT3B, ZF genes show increased expression and lower levels of H3K9me3 in promoter regions and gene bodies (Hahn, Wu et al. 2011).

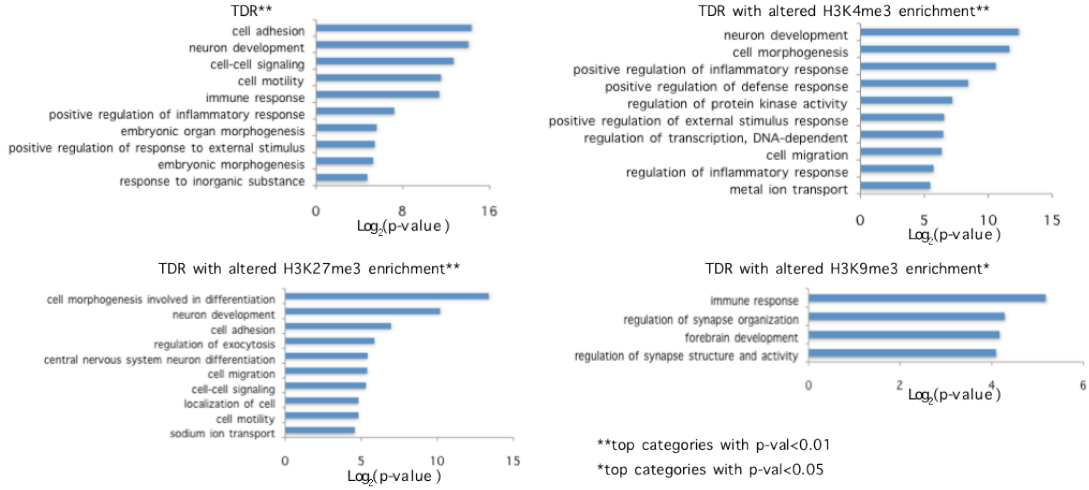

Figure 29. Biological processes of transcriptionally deregulated genes (TDR) and of TDR differentially enriched of histone marks.

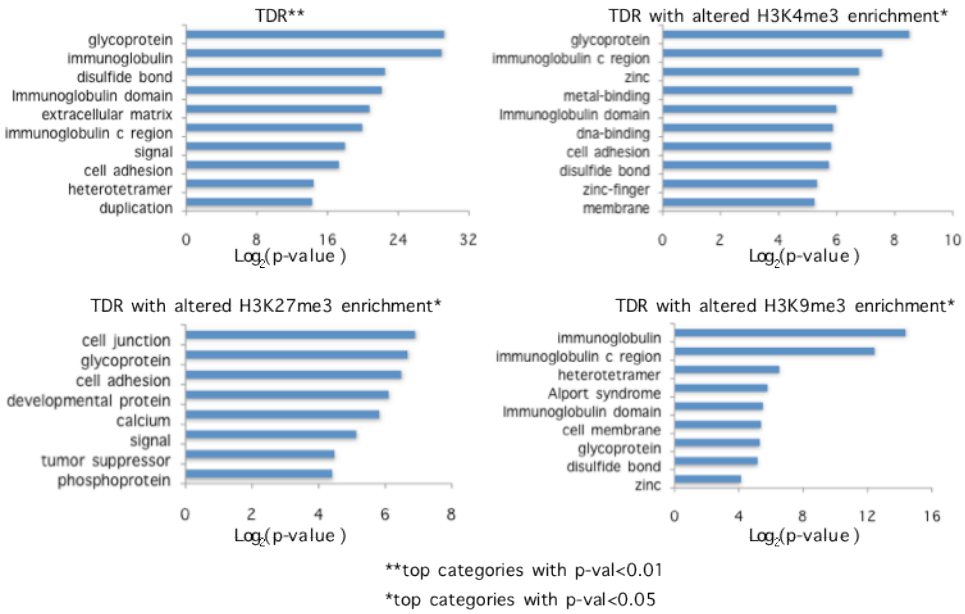

Figure 30. Gene functional categories of transcriptionally deregulated genes (TDR) and of TDR differentially enriched of histone marks.

In line with this finding, our data suggest that the effect on the trascription and histone marks of this specific family of genes is directly regulated by DNMT3B. Moreover, the H3K4me3 appears to play an even more important role than the H3K9me3 for what specifically concerns the ZF gene family. This functional category seems to be particularly important in the pathogenesis of the ICF disease, as it is known now that the ICF type 2, phenotypically similar to the ICF1 studied in this work, is mutated in a zinc finger protein (ZBTB24, see Chapter 2). Furthermore, it is well known how this protein family plays an important role in interpreting the DNA methylation signals and maybe also affected by changes in the epigenetic regulation of DNMT3B mutated cells.

Correlation between gene expression, DNA methylation and histone methylation

DNA methylation in ICF cells has been studied for long time and only recently a genome-wide profile has been performed (Heyn, Vidal et al. 2012). DNA methylation levels result significantly lowered in ICF cells and this reduction occurs mainly at heterochromatic regions. We found that 22% differentially methylated regions (DMR) are associated to genes (9% hypermethylated, 13% hypomethylated, data not shown). Of the genes associated to the hypo- and hyper-methylated regions, only 0,4-0,9% shows differential expression compared to the controls (Fig 31a). From a

different point of view, of the total of the differentially expressed genes, only 7-15% show differential methylation on the promoter (Fig 31b). This brings us again to the conclusion that the defects associated to deficient activity of the DNA-methyltransferase DNMT3B directly target predominantly intergenic regions, which have then secondary effects on gene expression.

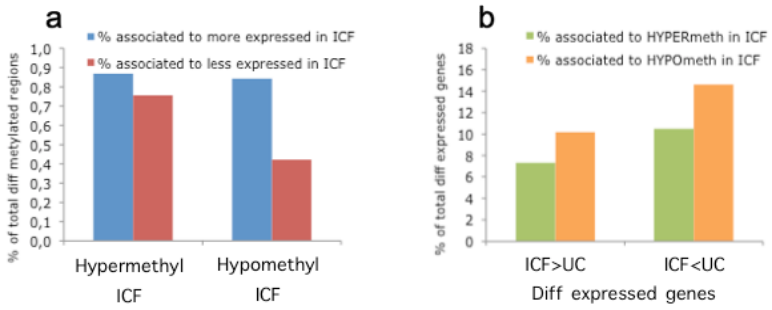

Figure 31. Correlation of gene expression and changes in DNA methylation on the TSS. a. On y-axis is the percentage of genes hyper- or hypo-methylated on TSS that show differential expression. b. On y-axis is the percentage of differentially expressed genes associated to hyper- or hypo-methylated DMRs on the TSS.

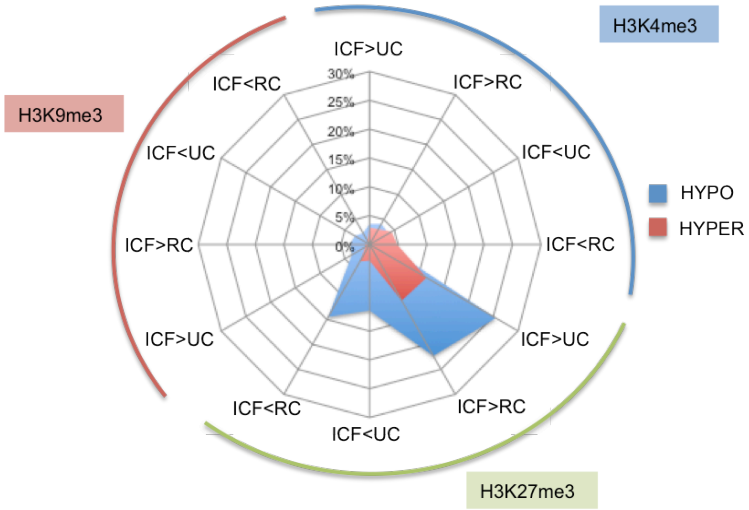

Figure 32. Overlap between differential histone marks enrichment and DMRs. ICF>UC and RC are peaks more enriched in ICF than in control and vice versa ICF<UC and RC. The red and blue areas on the graph represent the percentage of overlap between hyper- or hypo-methylated regions and the differentially enriched histone marks.

On the other hand, independently from gene proximity, there is a good correlation between changes in H3K27me3 and changes in methylation (Fig 32). A significant portion (22-25%) of H3K27me3 peaks more enriched in ICF than in the controls overlap at least one hypomethylated region in ICF, whereas 11% of H3K27me3 peaks overlap hypermethylated sites. Also in regions showing decrease of this histone mark in ICF compared to controls there is an 11-14% of hypomethylation and 2-3% of hyper. The association

between the methylation and this histone mark in ICF cells was only predicted from Heyin et al., but we show here that there is an interplay between the two, probably causing part of the ICF phenotype. For what concerns the other two histone modifications there seems to be no overlap between their enrichment variations and DNA methylation increase or decrease in presence of DNMT3B mutations.

Correlation between miRNA expression, DNA methylation and histone methylation

An important role in ICF syndrome is probably played by microRNAs, of which a number around one hundred change their expression in ICF compared to controls. Their epigenetic profile, though, does not vary significantly for those tested for changes in methylation and histone modifications enrichment (Gatto, Della Ragione et al. 2010). In order to compare the previous expression results with the new epigenomic profiles described in this work, only the miRNAs differentially expressed between the ICF line 8714 (ICF in this work) and the control line 8728 (RC) were considered (40 genes).

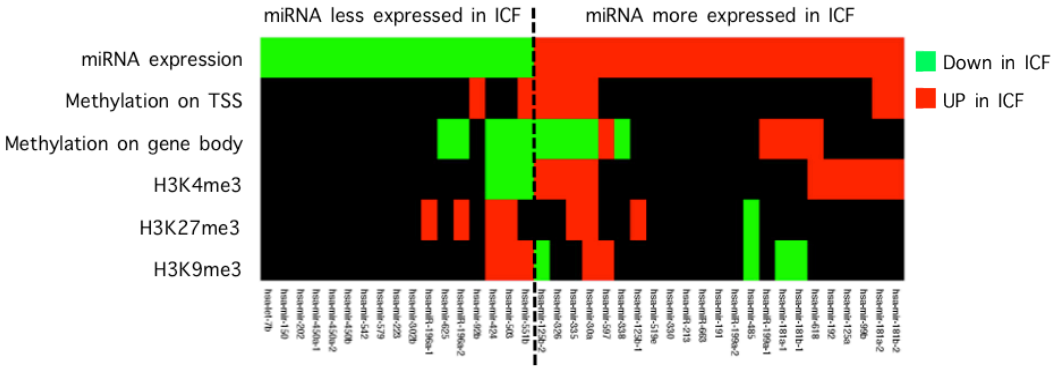

Figure 33. Heatmap representing the epigenetic profile of differentially expressed microRNAs. Each modification is indicated as "UP" or "DOWN", the color is not indicative of the amount of increase or decrease.

Taking advantage of the new data provided by these new NGS experiments, we could confirm, that the previously analyzed regions with targeted sequencing after bisulfite conversion and found unchanged were actually not marked by changes in methylation. However, we found that many of the differentially expressed miRNAs show changes in methylation in other sites, which are not corresponding with the previously identified CpG islands. Interestingly only less than half of the miRNAs shows differences in DNA

methylation and histone modifications (Fig 33). H3K4me3 changes always correlate with expression changes and most H3K27-K9me3 variations fit with the expression. Surprisingly, there is no strict correlation between changes of expression of microRNAs and their methylation profile, with many over-expressed genes showing hypermethylation on the predicted TSS and hypomethylation in the gene body (in case of intragenic miRNAs).

Epigenomic and transcriptomic alterations at repetitive regions

To better address the possible function or meaning of histones enrichment in regions depleted of genes, we analyzed alterations occurring in pericentromeric and gene desert regions in the ICF patient in more detail. The enrichment of all histone modifications in these regions compared to the background, represented by the input sample, genomic DNA, is shown in Fig 34. Among the studied histone marks, H3K9me3 is best known for its abundance in intergenic regions; in pericentromeres there seems not to be enrichment of H3K4me3 and H3K27me3 (ratio<1), while the enrichment of H3K9me3 does not vary among the samples (Fig 34a). In gene deserts, instead, the enrichment of H3K9me3 seems to be decreased in the ICF sample compared to the controls (Fig 34b). Although it needs to be verified by qPCR, this result might be interesting to define the connection of H3K9me3 and DNMT3B methylation in these cells.

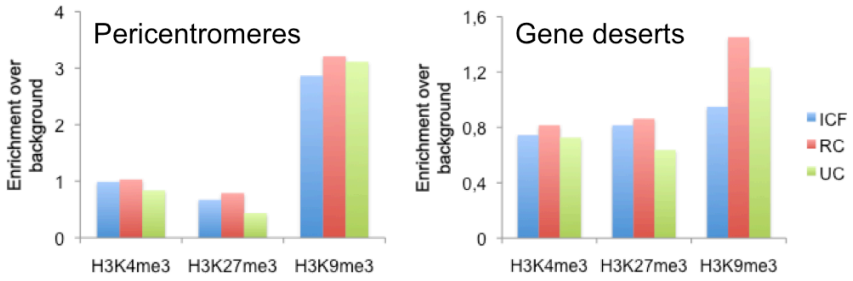

Figure 34. Pericentromeres and gene desert histone marks enrichment over the background (log of ChIP/input ratio).

Many of the intergenic sites correspond to regions containing a high abundance of genomic repeats, therefore we assessed systematically the occurrence of differential histone binding at repetitive elements genome-wide. For this purpose we calculated the relative enrichment of histone modification binding over the background on all repetitive elements, families and classes from RepeatMasker (http://www.repeatmasker.org).

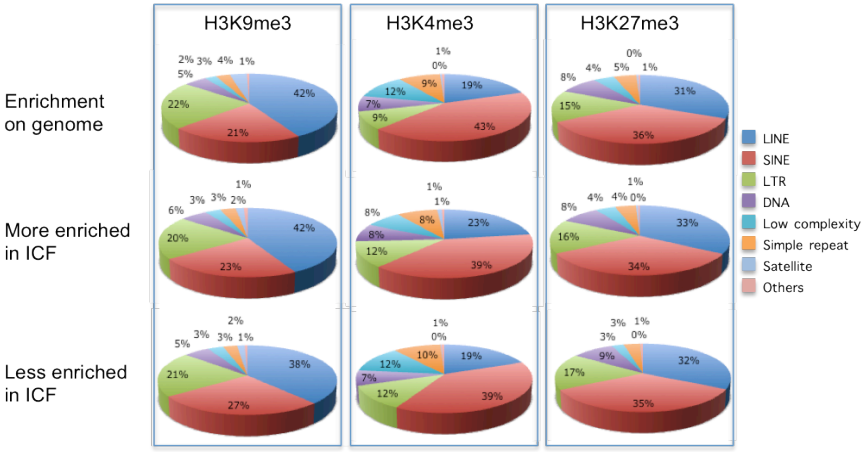

Figure 35. Histone marks enrichment at repetitive sequences in the genome

Although we couldn't identify any difference the distribution of the reads over the different classes of repeats among the samples (as shown in Fig 35), some differential enrichment was found in specific repeats.

Histone modifications enrichment and RNA expression were measured as described in Results – Part I section. For the ChIP-seq the mean fold enrichment of the modification over the background was calculated, while for RNA-seq the mean number of the reads mapping in repetitive regions, normalized by the total number of reads, are a measure of the expression of the region (Fig 36).

H3K4me3 is mainly increased in ICF compared to control cells in simple repeats, satellite, low complexity repeats, LTR, snRNA and tRNA. H3K27me3 changes mostly in satellites, tRNAs and simple repeats, but we also show that entire families and classes are differentially enriched of this modification, such as tRNA, rRNA, acro and SINE. H3K9me3 as well shows differences in simple repeats and LTRs, but also in SVA repeats. A strong reduction of DNA methylation has been found in the same categories, like satellites, LTRs, tRNA, rRNA, low complexity, simple repeats, SINE, SVA (Heynes et al).

The effect of these variations is reflected in RNA expression as well. Single repeats seem to highly increase in controls compared to ICF, while satellites are more enriched in ICF sample, particularly in Tar1, the repeat characterizing the telomeric sequences, particularly affected in ICF syndrome (Deng, Campbell et al. 2010).

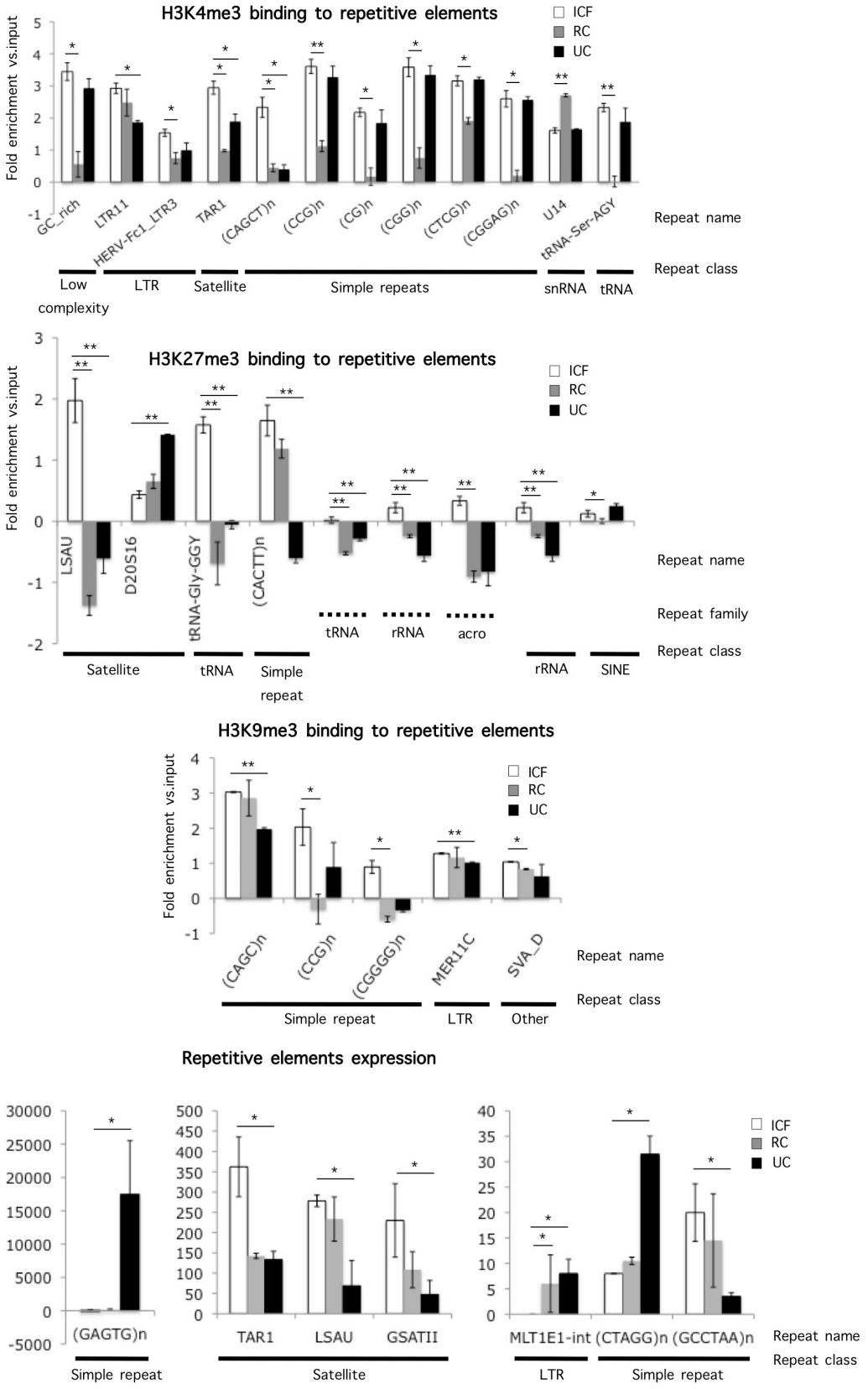

Figure 36. Histone marks differential enrichment in repetitive regions and their differential expression. From the top is represented the enrichment of H3K4me3, H3K27me3 and H3K9me3 over the background (as log of ChIP/input ratio). RNA expression is measured in average number of reads in the regions. Significant differences are marked with ** (p-val<0.01) and ** (p-val<0.05).

5. Materials and methods

Cell lines

The ICF syndrome cell line used in this study is the Epstein-Barr virus-transformed lymphoblastoid cell line (LCL) GM08714 (ICF). These cells come from the heterozygous ICF patient P4 [A603T and intron 22 G to A mutation resulting in insertion of three amino acids (STP) in DNMT3B]. Control cells include LCLs derived from normal individuals, GM08728 and LDA (Unrelated Control, UC). GM08728 is heterozygous for one ICF mutation (wt/A603T), being ICF patient's mother (Related Control, RC). LCLs were grown in RPMI1640 media (Euroclone) supplemented with 2 mM L-glutamine and 10% heat-inactivated fetal bovine serum (Euroclone).

Chromatin Immuno-Precipitation (ChIP)

The Chromatin Immuno-Precipitation (ChIP) is used to identify the DNA binding sites of a specific protein of interest. This technique is based on the cross-linking, that is the formation of reversible bonds among primary aminic groups in close proximity to one each other in the proteins (mainly lysines) and DNA and RNA bases (cytosine, adenine and guanine) through the activity of formaldehyde. The cross-linked DNA is then sonicated to small fragments (300-600bps) and one part (generally the 1%) of it is saved as input, to be used as control, and the rest is immunoprecipitated with a specific antibody and an unspecific one as a negative control. The antibody is precipitated adding protein A/G PLUS-AGAROSE beads to the mix; these beads bind the constant portion of the antibody and are separated through centrifugation from the rest of unbound chromatin. After the precipitation all the bonds are reversed through de-crosslinking at high temperatures (65°C) and the DNA is then purified from the proteins.

The fraction of DNA bound by the protein of interest can then be used in two applications. It can be amplified by Real Time-PCR to target specific regions that we are interested in studying, or it can be sequenced by Next Generation Sequencing, to be able instead to scan all the regions bound by that specific protein (Fig 37).

For each ChIP 10^6 cells (or $1,5x10^6$ for H3K9me3) were fixed with Formaldehyde 1% at a concentration of $5x10^5$ cell/ml.

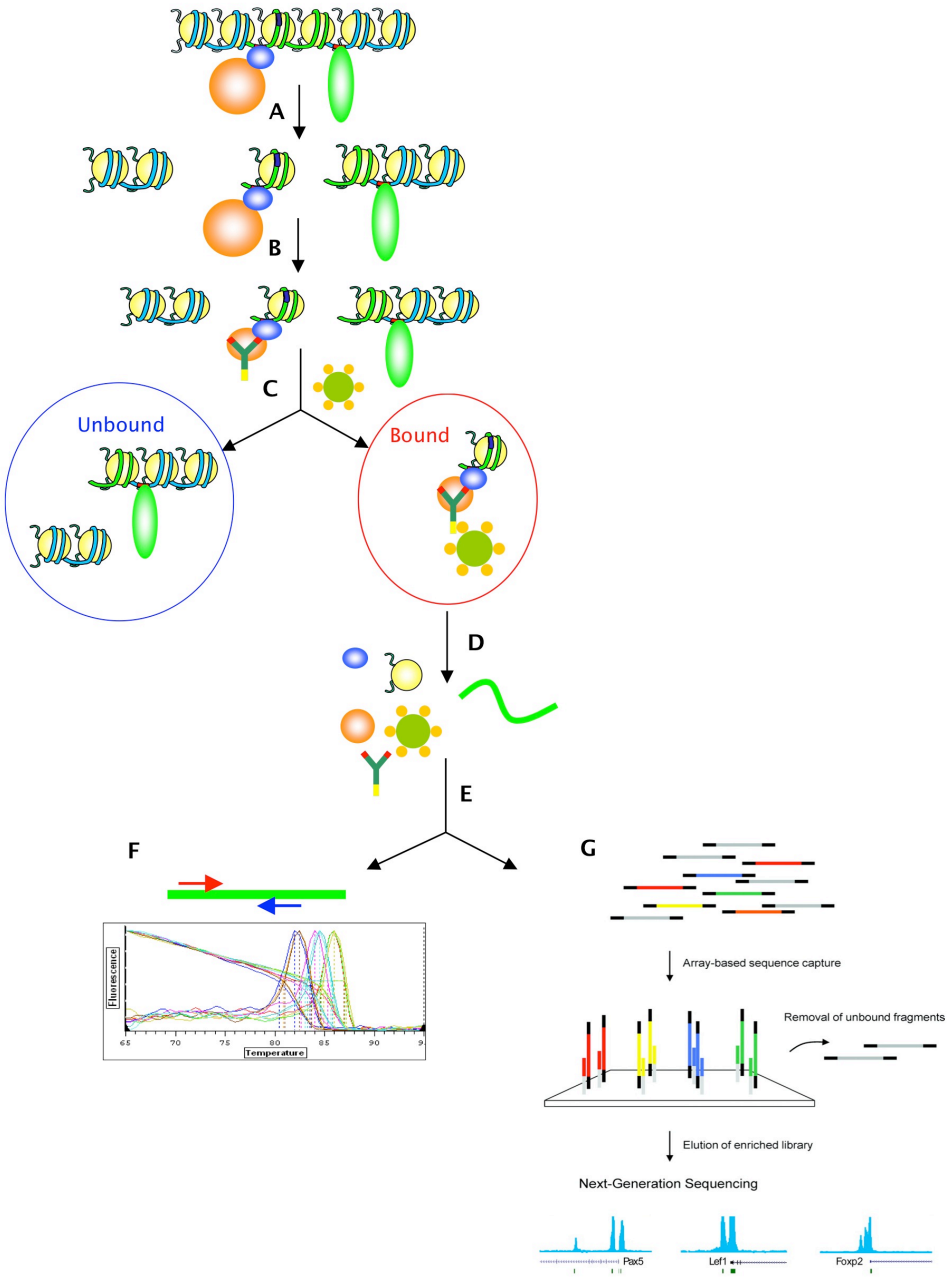

Figure 37. Chromatin Immuno-Precipitation. A. Formaldehyde-crosslinked DNA is sonicated. B. Specific antibody binds the protein of interest. C. Protein A/G beads bind the antibody and precipitate the DNA-protein complex. D. The immunoprecipitated complex is de-crosslinked at high temperature. E. DNA is purified. F. Real Time PCR is used to evaluate the protein binding on specific DNA regions. G. Adapters are ligated to DNA fragments and amplified on solid supports to perform Next Generation Sequencing for the genome-wide assessment of DNA binding sites for the specific protein.

After crosslinking the chromatin wa sonicated either with COVARIS (30 cycles of 30sec ON and 30 sec OFF, intensity 6) or Bioruptor (10 cycles of 30sec ON and 30 sec OFF, high power) instrumentations, depending on the availability. The antibodies used for the ChIPs are Anti-H3K27me3 (Abcam, mAb, ab6002, 10ug), Anti-H3K4me3 (Abcam, ab8580, 10ug) and Anti-H3K9me3 (Abcam, ab8898, 2ug). Protein A/G PLUS-Agarose beads from

Santa Cruz (sc-2003) was used for antibody precipitation. The eluted and purified DNA was sequenced with alternatively SOLiD and Illumina Next Generation Sequencing.

RNA extraction

RNA extraction was performed with a standard QIAzol (Lysis reagent from QIAGEN) protocol. RNA from 10^6 cells was isolated and resuspended in nuclease-free water. This RNA was used to perform libraries by the IGA facility in Udine.

For the RNA-seq performed in Cambridge the poly-A mRNA was extracted directly from cells. Oligotex kit from QIAgen was used. 10^6 lymphoblastoid cells were first lysed and homogenized in the presence of a highly denaturing guanidine-isothiocyanate (GITC) buffer, which immediately inactivates RNases to ensure isolation of intact mRNA. Oligotex Suspension was added, and hybridization took place between the oligo dT_{30} of the Oligotex particle and the poly-A tail of the mRNA. Contaminants are then washed away, and high-quality poly A+ RNA was eluted.

NGS Platforms - basics

NGS platforms rely on a combination of template preparation, sequencing, imaging and data analysis. In this study I only used the Illumina/HiSeq and the Life/SOLiD for sequencing in this work. The last step, the data analysis, is described in detail in the Result - Part I section.

The principal steps of the functioning of the machines are summarized here:

a. *Template preparation*

All the three systems clonally amplify the templates to obtain a higer representation of the single DNA molecules. For the 454 and SOLiD platforms the DNA libraries with specific adaptors are denatured into single strand and captured by amplification beads followed by emulsion PCR (Fig 38a). In Solexa protocol instead, the library with fixed adaptors is denatured to single strands and grafted to the flowcell, followed by bridge amplification to form clusters that contains clonal DNA fragments (Fig 38b).

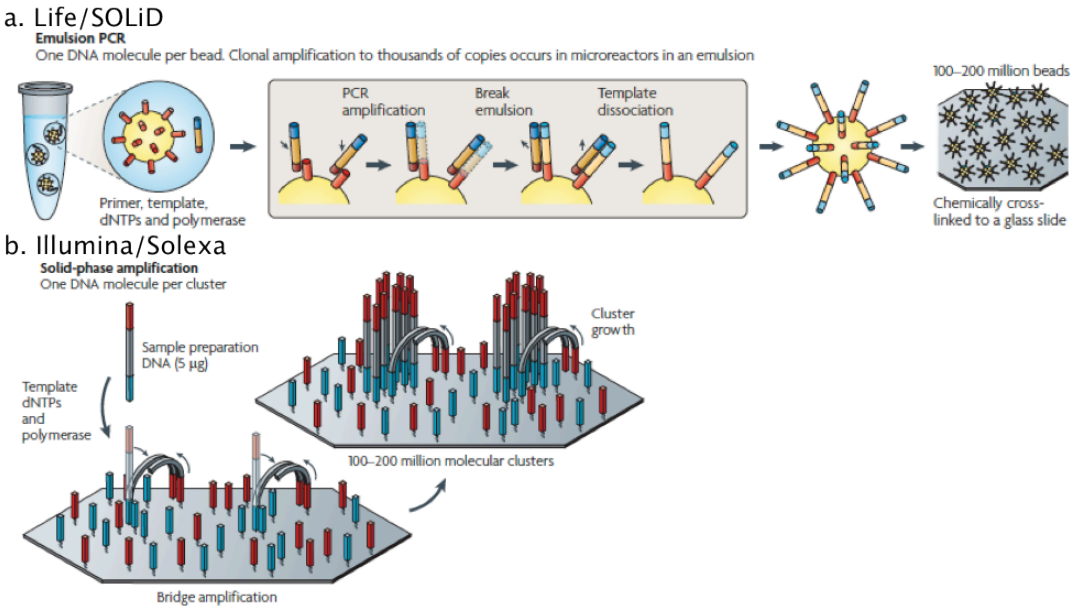

Figure 38. Sequencing library preparation for Life (a) and Illumina (b) technologies. From Metzker (2010).

b. Sequencing and imaging

This is the step that mainly differs among the three systems; here are the basics of sequencing and imaging for Roche/Life/Illumina sequencers.

- Cyclic reversible termination (HiSeq)

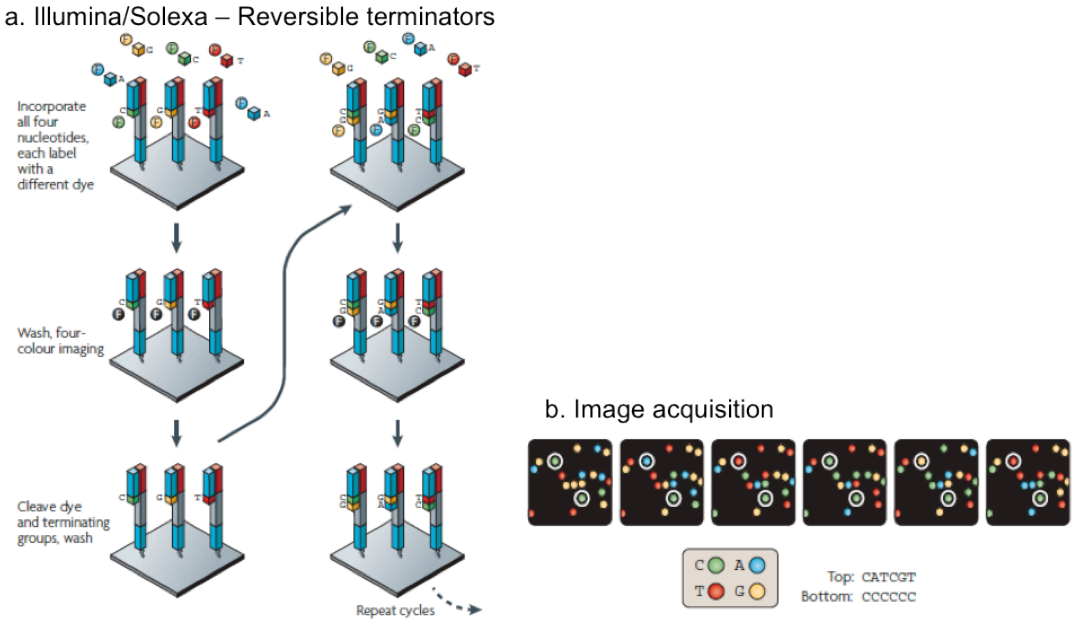

Figure 39. Sequencing through cyclic reversible terminators from Illumina. From Metzker (2010).

This technique somehow recalls the automated Sanger sequencing. Before sequencing, the library splices into single strands with the help of a linearization enzyme, and then four kinds of nucleotides

(dideoxyATP/ddATP, ddGTP, ddCTP, ddTTP) which contain different cleavable fluorescent dye and a removable blocking group complement the template one base at a time, and the signal is captured by a (charge-coupled device) CCD. One image is acquired at each cycle of addition of nucleotide, then the block is removed and a new base can be added to the sequence (Fig 39a-b).

- Sequencing by ligation (SOLiD)

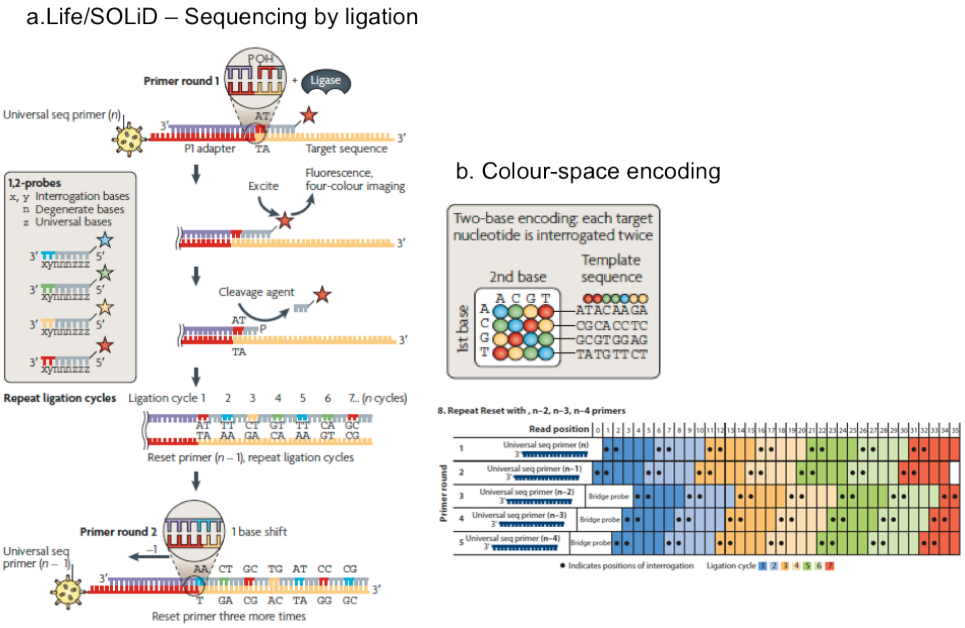

Figure 40. Sequencing by ligation from SOLiD. From Metzker (2010).

On a SOLiD flowcell, the libraries are sequenced by 8 base-probe ligation (Fig 40a, box), which contains one ligation site (the first base), one cleavage site (the fifth base), and four different fluorescent dyes (linked to the last base). The fluorescent signal is recorded when the probes are complementary to the template strand and removed by the cleavage of probes' last 3 bases. The sequence of the fragment can be deduced after 5 round of sequencing using ladder primer sets (Fig 40b).

Library preparation

All the experiments, ChIP- and RNA-seq, were performed in duplicates. H3K4me3 and K27me3 samples were sequenced using the SOLiD4 system, whereas H3K9me3 samples were sequenced using the Illumina GAII or HiSeq1000 system. The two RNA-seq duplicates were performed both with Illumina platforms (HiSeq1000). Here is a list of the samples we sequenced (Table 10, in red the samples sequenced in this work).

Cell system			Epigenomic regulation				Expression		
B cells (LCLs)	*DNMT3B mutations*	*Relatedness*	*Chip-seq 3meK4H3*	*Chip-seq 3meK27H3*	*Chip-seq 3meK9H3*	*Methylation (Hein et al., 2012)*	*mRNA seq*	*Gene array (Jin et al, 2007)*	*microRNA array (Gatto et al, 2010)*
8714 (ICF)	A603T/ STP807ins	patient	✓	✓	✓	✓	✓	✓	✓
8728 (ctrl1)	wt/A603T	patient's mother (RC)	✓	✓	✓		✓	✓	✓
LDA (ctrl2)	wt	unrelated control (UC)	✓	✓	✓	✓	✓		

Table 10. Genome wide experiments performed on ICF and control cell lines for the study of gene expression and epigenomic alterations. In red are the experiments performed within this work.

The relevant steps of the library preparation are summarized below.

- **ChIP-seq**

1. End repair. This step converts the overhangs into phosphorylated blunt ends, using T4 DNA polymerase, E. coli DNA Pol I large fragment (Klenow polymerase), and T4 polynucleotide kinase (PNK). The 3' to 5' exonuclease activity of these enzymes removes 3' overhangs and the polymerase activity fills in the 5' overhangs. It includes the purification with QIAquick columns.

2. Add 'A' Bases to the 3' End of the DNA Fragments. This step adds an 'A' base to the 3' end of the blunt phosphorylated DNA fragments, using the polymerase activity of Klenow fragment (3' to 5' exo minus). This prepares the DNA fragments for ligation to the adapters, which have a single 'T' base overhang at their 3' end. The DNA is then purified on MinElute columns.

3. Ligate adapters to DNA fragments. This is a fundamental step of the library prep. Adapters are ligated to the ends of the DNA fragments, preparing them to be hybridized to a flow cell. The DNA is again purified on MinElute columns.

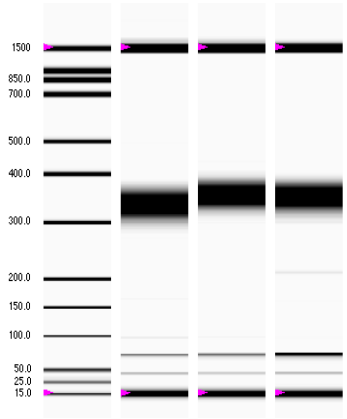

Figure 41. Bioanalyzer run. On the left column there is the size marker, the other three are the libraries. The color black indicates higher DNA amount.

4. Size selection. This step removes excess adaptors and selects a size range of templates that have to be sequenced. In this work a size of 350bps was chosen. 2% E-Gel SizeSelect Agarose Gels are used for size selection, with pre-cut wells for collection of DNA.

5. Amplification of the DNA library with PCR. Afterwards a qPCR with specific primer was performed as a quality control. The samples were tested with the BioAnalyzer (capillary electrophoresis) for size (Fig).

- **RNA-seq**

Library preparation for RNA-seq for Illumina platforms was performed in 6 steps:

1. RNA Sonication. The RNA was then sonicated in order to have smaller fragments for the library preparation.

2. Synthesis of double-strand cDNA. This cDNA was synthesized with the Just cDNA Double-Stranded cDNA Synthesis Kit from Agilent.

3. Steps 1-3 and 5 from ChIP-seq library synthesis were performed.

6. Discussion and conclusions

In this work the development of customized pipelines for Next Generation Sequencing (NGS) data analysis and integration was described. These workflows were tested on ChIP-seq, RNA-seq, Bisulfite-seq and microRNA (microarray) expression data from lymphoblastoid cell lines deriving from individuals affected by the human genetic disease ICF and from healthy individuals.

Next generation sequencing revolutionized the field of genomics in the last ten years. These new high-throughput sequencing technologies made it possible to apply sequence-based approaches in an unanticipated number of fields. In the epigenomic field for example, where the interests are mainly focused on transcriptional regulation in biological systems mediated by non-genomic factors, NGS technology allowed to observe from a novel, wider point of view differenti epigenetic marks. Some relevan examples are the distribution of DNA binding proteins, as transcription factors or histones isoforms, or the DNA methylation landscape. In their infancy, these new technologies have been largely used to profile the epigenomic patterns characterizing specific cell types, as histone modifications binding in T cells (Barski, Cuddapah et al. 2007) and pluripotent and committed cells (Mikkelsen, Ku et al. 2007; Hawkins, Hon et al. 2010). These works have shed new light on the specific roles of each histone modification on transcriptional regulation and on their distribution pattern across the genome. Lately, an enormous work has been done to correlate these marks with DNA methylation changes in cancer (Jin, Ernst et al. 2012) and in human pathologies (Heyn, Vidal et al. 2012). Moreover, studies on microRNAs roles in cancer are evolving rapidly towards a better understanding of the complex alterations leading to the pathology thanks to NGS techniques (Lopez-Serra and Esteller 2012).

Nowadays, many new research branches are developing due to the introduction of NGS technologies. One of these branches is driven by computational sciences applied to biologic concepts, to support the novel, overwhelming, amount of sequencing data available and to give biologists the right instruments to confidently extract information from such data;

another, completely innovative, branch bases its studies on published and freely available data, supporting the idea that there is no limit to the information that can be extracted from high-throughput sequencing. Moreover, other groups are still producing new data, which will then enrich our knowledge and our databases.

One of the first problems that biologists face when approaching sequencing is merely technical, and it concerns the infrastructures and the knowledge needed to handle data. Raw data have significant sizes and can occupy many Gb of hard disk, not generally handled by a common desktop computer. Moreover, the first mapping step requires a consistent amount of space, memory and time to run. This is the reason why a researcher approaching such a task has to be prepared and evaluate the solutions to the problem. Nowadays many companies that produce sequencing data are also offering data analysis and storage service for customers, requiring, of course, an additional fee to the already not indifferent one paid for the sequencing itself. Many institutions (or even single labs) also start to consider the global interest of people working on NGS data and begin providing internal services. Another, very often used, way of dealing with this issue is to train an internal person to handle the data and either rent the computers externally or start a collaboration with already equipped laboratories.

In this work the epigenomic regulation of transcription in a rare genetic monogenic disease (ICF syndrome) was investigated by integrating published data of bisulfite-sequencing and microRNA expression with newly produced data of ChIP-sequencing for histone H3 trimethylated in lysines K4, K27 and K9 and gene expression data from mRNA-seq. In order to efficiently analyze this data many existing tools have been tested and linked by custom scripts or data manipulation mediated by bedTools and SAMtools. The two pipelines here produced and discussed (Fig 13 and 23) have been built around the data, hence tested through the biological validation of the data (part of which was already published in previous works and part is currently ongoing).

Bioinformatics tools for NGS data analysis are continuously in development and every month new tools are available. Moreover, all the current international sequencing projects (like ENCODE, 1K genome project, etc.)

require every day more and more standardized workflows for seek of consistency and reproducibility of the data (Landt, Marinov et al. 2012; Dobin, Davis et al. 2013). Subsequently, one of the future steps that should be considered is to test each single part of the pipelines with synthetic data sets and compare the used tools with the new, more advanced and efficient ones. Moreover, it would be important to improve the pipeline stream in order to make it easier for all users in a lab to take advantage of it and to integrate this part of work in the commonly used tools of the lab.

The use of high throughput sequencing in this work had a very specific biological aim. To our knowledge, this is the first time that the global distribution and enrichment of histone modifications in lymphoblastoid cell lines from patients with a mutation in the DNA methyltranferase 3B is reported. It was already known for some specific loci that the histone H3 trimethylated in K4, K27 and K9 marks vary mainly in terms of enrichment and are not completely erased or rewritten, and it was confirmed at a genome-wide scale. The novel observation we could make is that the differentially enriched peaks for trimethylated H3K4 and K27 have a skewed genomic distribution with respect to the one observed in all detected peaks in the samples. This may mean, for example, that peaks of H3K4me3 in gene bodies are more targeted by DNMT3B hypofunctioning than the ones in intergenic regions; conversely, the intergenic regions show an increase of H3K27me3 compared to the control. Only H3K9me3 distribution does not change its genomic distribution, even when its enrichment is altered.

The changes in enrichment are reflected only partially in gene expression. Of the genes bound by these marks on their transcription start sites only a small percentage (1-6%) changes its expression; but many differentially expressed genes (5-50%) are marked by changes in histone modifications. This can have different explanations. One of these is that probably not all changes in enrichment are directly effective on gene expression and may be balanced from other factors we did not analyze in this work. On the other side, is also true that not all changes in expression occur in presence of changes of the analyzed histone modifications, therefore also these alterations could be due to other factors. Moreover, we did a very stringent RNA-seq analysis, considering as duplicated sequencing data from two different places, ending with only 544 total differentially expressed genes.

This factor could have led to the underestimation of the deregulated genes, restricting the group to the ones with the biggest difference in expression, thus resulting also in an underestimation of genes with changes in epigenetic profile.

The annotation of all differentially expressed genes (TDR, transcriptionally deregulated) and of those associated to differentially enriched peaks of histone modifications revealed important enriched gene categories. The biological processes and gene functional categories enriched in TDR genes and in TDR genes associated to changes in histone marks are, as expected, mainly enriched in immunological regulation, cell motion and migration, development and neuronal pathways. Only the TDR genes with alterations in H3K4me3 show enrichment in two novel categories, the regulation of transcription and zinc-finger protein family. This is a novel finding, as it was known that the general transcription pathway is altered in these cells, but the link between DNMT3B function, the zinc-finger proteins, the H3K4me3 mark and transcription was still uncovered.

H3K9me3 plays an important role in zinc-finger proteins transcription regulation (Hahn, Wu et al. 2011). However, in our cells only a weak enrichment has been found at these genes, although consistent with their expression changes. The correlation between this mark and this gene family, though, has been proven at the level of the gene body, while the association displayed in Fig 30 is done with gene TSS. It is possible that extending the correlation to the gene body of the differentially expressed genes could also increase the number of differentially expressed zinc fingers marked by changes in H3K9me3.

A very interesting link between the ICF syndrome and the zinc-finger protein family has arisen when the ICF type 2 genetic origin was described (de Greef, Wang et al. 2011) and the ZBTB24 gene was pointed out as responsible for another form of ICF, characterized by hypomethylation in alpha satellites. This gene is part of the ZBTB family, whose involvement in hematopoiesis, and in particular in regulation of lymphoid development and function, mainly in GC B cells (Germinal Center), has lately been characterized (Lee and Maeda 2012). ZBTB24 is not deregulated in ICF (type 1) cells, but another member of the family, ZBTB32 (among many others) is less expressed compared to control cells. This protein is a key regulator of

the differentiation of B cell to plasma cells, acting as a repressor of CIITA in complex with another protein, Blimp1, to stimulate the progression of state of the B cells to plasma cells (Yoon, Scharer et al. 2012). It has been proven that the reduction of this protein delays kinetics in B cells progression toward active, immunoglobulin-producing cells. The connection with the ICF phenotype is clear, as the main symptoms of this disease are agammaglobulinemia and lack of mature cells and plasmacells in patients. With this observation we have probably moved a step forward towards the better understanding of the phenotype of the syndrome.

A very interesting observation coming from the correlation of bisulfite-sequencing data and RNA-seq data is that DNA methylation can be both increased or decreased on the promoters of upregulated genes or microRNAs, and the same is observed in less expressed genes or microRNAs in ICF versus controls. This may mark a misguided activity of the mutated DNMT3B, or a compensatory mechanism, that does not seem to have a direct and clear effect on genes transcription.

Defects in the proper DNMT3B targeting would be compatible with the specific mutations described in ICF patients. Indeed, in the ICF variants the binding of DNMT3L regulating the targeting of DNMT3B is disrupted, thus presumably explaining the defects in methylation profile.

Moreover, DNA hypomethylation in ICF syndrome was before correlated to H3K27me3 and H3K9me3 histone marks, but it was never observed in correlation with their changes in the same cells. We clearly observe that only changes in H3K27me3 co-localize with changes of DNA methylation, while H3K9me3 alterations are poorly overlapping with hyper- or hypo-methylation in ICF cells. This observation may reinforce the hypothesis that these two epigenetic factors are strictly interconnected and that the mutated DNMT3B can influence both of them.

Overall, of all the disregulated genes, 60-62% of them showed alterations in DNA methylation or variations in enrichment of at least one of the histone marks on the TSS, while the rest did not show any alteration in this analysis. It is possible that these genes are direct target of altered transcription factors or target of disregulated microRNAs. In Gatto, Della Ragione et al. (2010), 40 microRNAs were detected as differentially expressed comparing ICF cell line and the control (RC); their predicted or validated TSS were here

associated to changes in DNA methylation and histone modifications enrichment. Again, 60% of differentially expressed microRNAs showed alterations in DNA methylation or histone marks enrichment, some to attribute to mistaken prediction of the TSS and some probably due to factors not included in this study. At the time the miRNA expression microarray was performed there was little information about the real TSS of the miRNAs and also a small number of them were discovered, compared to now. In 2009 there were 735 human microRNA on the expression microarray (Sanger miRNA db v 9.0-9.1, Berezikov, van Tetering et al. (2006)), while now only MirBase (Kozomara and Griffiths-Jones (2011), www.mirbase.org, release 19, Aug 2012) contains 1600 sequences. Moreover, thanks to the NGS technologies more detailed information is now available to identify the TSS of those genes and specific databases are born to collect data from the literature (mirT, http://www.isical.ac.in/~bioinfo_miu/miRT/miRT.php, Bhattacharyya, Das et al. (2012); miRStart, http://mirstart.mbc.nctu.edu.tw/). For this reason, we probably miss some part of information about those miRNAs that were not tested for expression.

ICF syndrome is mainly affected by hypomethylation in heterochromatin, such as pericentromeric regions and satellites, and repetitive regions (Heyn, Vidal et al. 2012). This hypomethylation is not reflected into changes in histones enrichment in pericentromeric regions, but a decrease of H3K9me3 is observed in gene deserts, where repetitive regions are mostly concentrated. A deeper analysis of such repetitive sequences in the genome showed that many repeat sequences are differentially enriched in H3K4me3, H3K27me3 and H3K9me3 and some are also differentially expressed, like telomere repeats TAR1. Some classes and families, like satellites, LTR, simple repeats, rRNA and tRNA are differentially enriched in more than one histone modification, are differentially expressed and are hypomethylated. DNMT3B defective activity seems to have a strong impact on those regions and it would be highly interesting to study in greater detail the effect that those changes cause in the cells.

In conclusion, the use of ChIP-seq and RNA-seq in this work, joint with bisulfite-seq and microRNA microarray data, allowed to create a wider picture of the epigenetic landscape in cells affected by mutations in

DNMT3B. This new information will provide new insights on ICF syndrome pathogenesis, better dissecting its molecular phenotype. Moreover, the big amount of data produced is a long-lasting source of information, and will serve to answer even more questions than the ones already answered here.

7. References

Achour, M., X. Jacq, et al. (2008). "The interaction of the SRA domain of ICBP90 with a novel domain of DNMT1 is involved in the regulation of VEGF gene expression." Oncogene **27**(15): 2187-2197.

Anders, S. and W. Huber (2010). "Differential expression analysis for sequence count data." Genome Biol **11**(10): R106.

Ball, M. P., J. B. Li, et al. (2009). "Targeted and genome-scale strategies reveal gene-body methylation signatures in human cells." Nature biotechnology **27**(4): 361-368.

Bannister, A. J., P. Zegerman, et al. (2001). "Selective recognition of methylated lysine 9 on histone H3 by the HP1 chromo domain." Nature **410**(6824): 120-124.

Barski, A., S. Cuddapah, et al. (2007). "High-resolution profiling of histone methylations in the human genome." Cell **129**(4): 823-837.

Berezikov, E., G. van Tetering, et al. (2006). "Many novel mammalian microRNA candidates identified by extensive cloning and RAKE analysis." Genome Res **16**(10): 1289-1298.

Bestor, T. H. (2000). "The DNA methyltransferases of mammals." Human molecular genetics **9**(16): 2395-2402.

Bhattacharyya, M., M. Das, et al. (2012). "miRT: a database of validated transcription start sites of human microRNAs." Genomics Proteomics Bioinformatics **10**(5): 310-316.

Bird, A. (2002). "DNA methylation patterns and epigenetic memory." Genes & development **16**(1): 6-21.

Bird, A. P. (1995). "Gene number, noise reduction and biological complexity." Trends Genet **11**(3): 94-100.

Blanco-Betancourt, C. E., A. Moncla, et al. (2004). "Defective B-cell-negative selection and terminal differentiation in the ICF syndrome." Blood **103**(7): 2683-2690.

Borgel, J., S. Guibert, et al. (2010). "Targets and dynamics of promoter DNA methylation during early mouse development." Nature genetics **42**(12): 1093-1100.

Bostick, M., J. K. Kim, et al. (2007). "UHRF1 plays a role in maintaining DNA methylation in mammalian cells." Science **317**(5845): 1760-1764.

Brenner, C., R. Deplus, et al. (2005). "Myc represses transcription through recruitment of DNA methyltransferase corepressor." The EMBO journal **24**(2): 336-346.

Brinkman, A. B., H. Gu, et al. (2012). "Sequential ChIP-bisulfite sequencing enables direct genome-scale investigation of chromatin and DNA methylation cross-talk." Genome Res **22**(6): 1128-1138.

Buck-Koehntop, B. A. and P. A. Defossez (2013). "On how mammalian transcription factors recognize methylated DNA." Epigenetics : official journal of the DNA Methylation Society **8**(2): 131-137.

Bullard, J. H., E. Purdom, et al. (2010). "Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments." BMC Bioinformatics **11**: 94.

Carneiro, M. O., C. Russ, et al. (2012). "Pacific biosciences sequencing technology for genotyping and variation discovery in human data." BMC Genomics **13**: 375.

Chen, Z. X., J. R. Mann, et al. (2005). "Physical and functional interactions between the human DNMT3L protein and members of the de novo methyltransferase family." Journal of cellular biochemistry **95**(5): 902-917.

Chouery, E., J. Abou-Ghoch, et al. (2012). "A novel deletion in ZBTB24 in a Lebanese family with immunodeficiency, centromeric instability, and facial anomalies syndrome type 2." Clin Genet **82**(5): 489-493.

Church, G. M. (1984). "Genomic Sequencing." Proceedings of the National Academy of Sciences of the United States of America **81**(7): 5.

de Greef, J. C., J. Wang, et al. (2011). "Mutations in ZBTB24 are associated with immunodeficiency, centromeric instability, and facial anomalies syndrome type 2." Am J Hum Genet **88**(6): 796-804.

Deng, Z., A. E. Campbell, et al. (2010). "TERRA, CpG methylation and telomere heterochromatin: lessons from ICF syndrome cells." Cell Cycle **9**(1): 69-74.

Dobin, A., C. A. Davis, et al. (2013). "STAR: ultrafast universal RNA-seq aligner." Bioinformatics **29**(1): 15-21.

Doi, A., I. H. Park, et al. (2009). "Differential methylation of tissue- and cancer-specific CpG island shores distinguishes human induced pluripotent stem cells, embryonic stem cells and fibroblasts." Nature genetics **41**(12): 1350-1353.

Ecker, J. R., W. A. Bickmore, et al. (2012). "Genomics: ENCODE explained." Nature **489**(7414): 52-55.

Edgar, A. J., S. L. Dover, et al. (2005). "Bone morphogenetic protein-2 induces expression of murine zinc finger transcription factor ZNF450." Journal of cellular biochemistry **94**(1): 202-215.

Ehrlich, M., K. L. Buchanan, et al. (2001). "DNA methyltransferase 3B mutations linked to the ICF syndrome cause dysregulation of lymphogenesis genes." Hum Mol Genet **10**(25): 2917-2931.

Espada, J. and M. Esteller (2007). "Epigenetic control of nuclear architecture." Cell Mol Life Sci **64**(4): 449-457.

Esteller, M. (2007). "Cancer epigenomics: DNA methylomes and histone-modification maps." Nat Rev Genet **8**(4): 286-298.

Esteller, M. (2007). "Epigenetic gene silencing in cancer: the DNA hypermethylome." Human molecular genetics **16 Spec No 1**: R50-59.

Fatemi, M., A. Hermann, et al. (2002). "Dnmt3a and Dnmt1 functionally cooperate during de novo methylation of DNA." European journal of biochemistry / FEBS **269**(20): 4981-4984.

Feldman, N., A. Gerson, et al. (2006). "G9a-mediated irreversible epigenetic inactivation of Oct-3/4 during early embryogenesis." Nature cell biology **8**(2): 188-194.

Filion, G. J., S. Zhenilo, et al. (2006). "A family of human zinc finger proteins that bind methylated DNA and repress transcription." Mol Cell Biol **26**(1): 169-181.

Fuks, F., W. A. Burgers, et al. (2001). "Dnmt3a binds deacetylases and is recruited by a sequence-specific repressor to silence transcription." The EMBO journal **20**(10): 2536-2544.

Fuks, F., P. J. Hurd, et al. (2003). "The methyl-CpG-binding protein MeCP2 links DNA methylation to histone methylation." J Biol Chem **278**(6): 4035-4040.

Gardiner-Garden, M. and M. Frommer (1987). "CpG islands in vertebrate genomes." Journal of molecular biology **196**(2): 261-282.

Gatto, S., D'Esposito, M., Matarazzo, M.R. (2012). The role of DNMT3B mutations in the pathogenesis of ICF syndrome. Patho-Epigenetics of Disease. S. S. B. Media.

Gatto, S., F. Della Ragione, et al. (2010). "Epigenetic alteration of microRNAs in DNMT3B-mutated patients of ICF syndrome." Epigenetics : official journal of the DNA Methylation Society **5**(5): 427-443.

Ginolhac, A., J. Vilstrup, et al. (2012). "Improving the performance of true single molecule sequencing for ancient DNA." BMC Genomics **13**: 177.

Guil, S. and M. Esteller (2009). "DNA methylomes, histone codes and miRNAs: tying it all together." Int J Biochem Cell Biol **41**(1): 87-95.

Hagleitner, M. M., A. Lankester, et al. (2008). "Clinical spectrum of immunodeficiency, centromeric instability and facial dysmorphism (ICF syndrome)." J Med Genet **45**(2): 93-99.

Hahn, M. A., X. Wu, et al. (2011). "Relationship between gene body DNA methylation and intragenic H3K9me3 and H3K36me3 chromatin marks." PLoS One **6**(4): e18844.

Hansen, R. S., C. Wijmenga, et al. (1999). "The DNMT3B DNA methyltransferase gene is mutated in the ICF immunodeficiency syndrome." Proc Natl Acad Sci U S A **96**(25): 14412-14417.

Hardcastle, T. J. and K. A. Kelly (2010). "baySeq: empirical Bayesian methods for identifying differential expression in sequence count data." BMC Bioinformatics **11**: 422.

Hawkins, R. D., G. C. Hon, et al. (2010). "Distinct epigenomic landscapes of pluripotent and lineage-committed human cells." Cell Stem Cell **6**(5): 479-491.

Hebenstreit, D., M. Gu, et al. (2011). "EpiChIP: gene-by-gene quantification of epigenetic modification levels." Nucleic Acids Res **39**(5): e27.

Hellman, A. and A. Chess (2007). "Gene body-specific methylation on the active X chromosome." Science **315**(5815): 1141-1143.

Heyn, H., E. Vidal, et al. (2012). "Whole-genome bisulfite DNA sequencing of a DNMT3B mutant patient." Epigenetics : official journal of the DNA Methylation Society **7**(6): 542-550.

Hirasawa, R. and R. Feil (2010). "Genomic imprinting and human disease." Essays Biochem **48**(1): 187-200.

Holz-Schietinger, C. and N. O. Reich (2010). "The inherent processivity of the human de novo methyltransferase 3A (DNMT3A) is enhanced by DNMT3L." The Journal of biological chemistry **285**(38): 29091-29100.

Huang da, W., B. T. Sherman, et al. (2009). "Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists." Nucleic Acids Res **37**(1): 1-13.

Huang da, W., B. T. Sherman, et al. (2009). "Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources." Nat Protoc **4**(1): 44-57.

Hutchison, C. A., 3rd (2007). "DNA sequencing: bench to bedside and beyond." Nucleic Acids Res **35**(18): 6227-6237.

Jeanpierre, M., C. Turleau, et al. (1993). "An embryonic-like methylation pattern of classical satellite DNA is observed in ICF syndrome." Hum Mol Genet **2**(6): 731-735.

Jeltsch, A. (2006). "On the enzymatic properties of Dnmt1: specificity, processivity, mechanism of linear diffusion and allosteric regulation of the enzyme." Epigenetics : official journal of the DNA Methylation Society **1**(2): 63-66.

Jeong, S., G. Liang, et al. (2009). "Selective anchoring of DNA methyltransferases 3A and 3B to nucleosomes containing methylated DNA." Mol Cell Biol **29**(19): 5366-5376.

Jeong, S., G. Liang, et al. (2009). "Selective anchoring of DNA methyltransferases 3A and 3B to nucleosomes containing methylated DNA." Molecular and cellular biology **29**(19): 5366-5376.

Ji, H., L. I. Ehrlich, et al. (2010). "Comprehensive methylome map of lineage commitment from haematopoietic progenitors." Nature **467**(7313): 338-342.

Jiang, Y. L., M. Rigolet, et al. (2005). "DNMT3B mutations and DNA methylation defect define two types of ICF syndrome." Hum Mutat **25**(1): 56-63.

Jin, B., J. Ernst, et al. (2012). "Linking DNA methyltransferases to epigenetic marks and nucleosome structure genome-wide in human tumor cells." Cell Rep **2**(5): 1411-1424.

Jin, B., Q. Tao, et al. (2008). "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 **17**(5): 690-709.

Jones, P. A. (2002). "DNA methylation and cancer." Oncogene **21**(35): 5358-5360.

Jones, P. A. and G. Liang (2009). "Rethinking how DNA methylation patterns are maintained." Nature reviews. Genetics **10**(11): 805-811.

Jones, P. L., G. J. Veenstra, et al. (1998). "Methylated DNA and MeCP2 recruit histone deacetylase to repress transcription." Nat Genet **19**(2): 187-191.

Jurkowska, R. Z., T. P. Jurkowski, et al. (2011). "Structure and function of mammalian DNA methyltransferases." Chembiochem : a European journal of chemical biology **12**(2): 206-222.

Kent, W. J., C. W. Sugnet, et al. (2002). "The human genome browser at UCSC." Genome Res **12**(6): 996-1006.

Klose, R. J. and A. P. Bird (2006). "Genomic DNA methylation: the mark and its mediators." Trends Biochem Sci **31**(2): 89-97.

Kondo, T., M. P. Bobek, et al. (2000). "Whole-genome methylation scan in ICF syndrome: hypomethylation of non-satellite DNA repeats D4Z4 and NBL2." Hum Mol Genet **9**(4): 597-604.

Kozomara, A. and S. Griffiths-Jones (2011). "miRBase: integrating microRNA annotation and deep-sequencing data." Nucleic Acids Res **39**(Database issue): D152-157.

Kriaucionis, S. and N. Heintz (2009). "The nuclear DNA base 5-hydroxymethylcytosine is present in Purkinje neurons and the brain." Science **324**(5929): 929-930.

Lachner, M., D. O'Carroll, et al. (2001). "Methylation of histone H3 lysine 9 creates a binding site for HP1 proteins." Nature **410**(6824): 116-120.

Lander, E. S., L. M. Linton, et al. (2001). "Initial sequencing and analysis of the human genome." Nature **409**(6822): 860-921.

Landt, S. G., G. K. Marinov, et al. (2012). "ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia." Genome Res **22**(9): 1813-1831.

Langmead, B., C. Trapnell, et al. (2009). "Ultrafast and memory-efficient alignment of short DNA sequences to the human genome." Genome Biol **10**(3): R25.

Lee, S. U. and T. Maeda (2012). "POK/ZBTB proteins: an emerging family of proteins that regulate lymphoid development and function." Immunol Rev **247**(1): 107-119.

Lewis, J. D., R. R. Meehan, et al. (1992). "Purification, sequence, and cellular localization of a novel chromosomal protein that binds to methylated DNA." Cell **69**(6): 905-914.

Li, B. and C. N. Dewey (2011). "RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome." BMC Bioinformatics **12**: 323.

Li, E. (2002). "Chromatin modification and epigenetic reprogramming in mammalian development." Nat Rev Genet **3**(9): 662-673.

Li, E., T. H. Bestor, et al. (1992). "Targeted mutation of the DNA methyltransferase gene results in embryonic lethality." Cell **69**(6): 915-926.

Li, H., B. Handsaker, et al. (2009). "The Sequence Alignment/Map format and SAMtools." Bioinformatics **25**(16): 2078-2079.

Li, J. and R. Tibshirani (2011). "Finding consistent patterns: A nonparametric approach for identifying differential expression in RNA-Seq data." Stat Methods Med Res.

Liu, L., Y. Li, et al. (2012). "Comparison of next-generation sequencing systems." J Biomed Biotechnol **2012**: 251364.

Llaca, V. and J. Messing (1998). "Amplicons of maize zein genes are conserved within genic but expanded and constricted in intergenic regions." Plant J **15**(2): 211-220.

Lopez-Serra, P. and M. Esteller (2012). "DNA methylation-associated silencing of tumor-suppressor microRNAs in cancer." Oncogene **31**(13): 1609-1622.

Luger, K., A. W. Mader, et al. (1997). "Crystal structure of the nucleosome core particle at 2.8 A resolution." Nature **389**(6648): 251-260.

Maraschio, P., O. Zuffardi, et al. (1988). "Immunodeficiency, centromeric heterochromatin instability of chromosomes 1, 9, and 16, and facial anomalies: the ICF syndrome." J Med Genet **25**(3): 173-180.

Margueron, R., P. Trojer, et al. (2005). "The key to development: interpreting the histone code?" Curr Opin Genet Dev **15**(2): 163-176.

Matarazzo, M. R., M. L. De Bonis, et al. (2009). "Lessons from two human chromatin diseases, ICF syndrome and Rett syndrome." Int J Biochem Cell Biol **41**(1): 117-126.

Maxam, A. M. and W. Gilbert (1977). "A new method for sequencing DNA." Proc Natl Acad Sci U S A **74**(2): 560-564.

Maze, I., H. E. Covington, 3rd, et al. (2010). "Essential role of the histone methyltransferase G9a in cocaine-induced plasticity." Science **327**(5962): 213-216.

McGettigan, P. A. (2013). "Transcriptomics in the RNA-seq era." Curr Opin Chem Biol **17**(1): 4-11.

McLean, C. Y., D. Bristor, et al. (2010). "GREAT improves functional interpretation of cis-regulatory regions." Nat Biotechnol **28**(5): 495-501.

Metzker, M. L. (2010). "Sequencing technologies - the next generation." Nat Rev Genet **11**(1): 31-46.

Mikkelsen, T. S., M. Ku, et al. (2007). "Genome-wide maps of chromatin state in pluripotent and lineage-committed cells." Nature **448**(7153): 553-560.

Miniou, P., M. Jeanpierre, et al. (1997). "alpha-satellite DNA methylation in normal individuals and in ICF patients: heterogeneous methylation of constitutive heterochromatin in adult and fetal tissues." Hum Genet **99**(6): 738-745.

Mohn, F. and D. Schubeler (2009). "Genetics and epigenetics: stability and plasticity during cellular differentiation." Trends Genet **25**(3): 129-136.

Mohn, F., M. Weber, et al. (2008). "Lineage-specific polycomb targets and de novo DNA methylation define restriction and potential of neuronal progenitors." Molecular cell **30**(6): 755-766.

Mortazavi, A., B. A. Williams, et al. (2008). "Mapping and quantifying mammalian transcriptomes by RNA-Seq." Nat Methods **5**(7): 621-628.

Nan, X., H. H. Ng, et al. (1998). "Transcriptional repression by the methyl-CpG-binding protein MeCP2 involves a histone deacetylase complex." Nature **393**(6683): 386-389.

Nightingale, K. P., L. P. O'Neill, et al. (2006). "Histone modifications: signalling receptors and potential elements of a heritable epigenetic code." Curr Opin Genet Dev **16**(2): 125-136.

Oda, M., A. Yamagiwa, et al. (2006). "DNA methylation regulates long-range gene silencing of an X-linked homeobox gene cluster in a lineage-specific manner." Genes & development **20**(24): 3382-3394.

Okano, M., D. W. Bell, et al. (1999). "DNA methyltransferases Dnmt3a and Dnmt3b are essential for de novo methylation and mammalian development." Cell **99**(3): 247-257.

Ooi, S. K., C. Qiu, et al. (2007). "DNMT3L connects unmethylated lysine 4 of histone H3 to de novo methylation of DNA." Nature **448**(7154): 714-717.

Otani, J., T. Nankumo, et al. (2009). "Structural basis for recognition of H3K4 methylation status by the DNA methyltransferase 3A ATRX-DNMT3-DNMT3L domain." EMBO reports **10**(11): 1235-1241.

Pepke, S., B. Wold, et al. (2009). "Computation for ChIP-seq and RNA-seq studies." Nat Methods **6**(11 Suppl): S22-32.

Portela, A. and M. Esteller (2010). "Epigenetic modifications and human disease." Nat Biotechnol **28**(10): 1057-1068.

Prokhortchouk, A., B. Hendrich, et al. (2001). "The p120 catenin partner Kaiso is a DNA methylation-dependent transcriptional repressor." Genes Dev **15**(13): 1613-1618.

Quenneville, S., G. Verde, et al. (2011). "In embryonic stem cells, ZFP57/KAP1 recognize a methylated hexanucleotide to affect chromatin and DNA methylation of imprinting control regions." Mol Cell **44**(3): 361-372.

Quinlan, A. R. and I. M. Hall (2010). "BEDTools: a flexible suite of utilities for comparing genomic features." Bioinformatics **26**(6): 841-842.

Richard, G. F., A. Kerrest, et al. (2008). "Comparative genomics and molecular dynamics of DNA repeats in eukaryotes." Microbiol Mol Biol Rev **72**(4): 686-727.

Robertson, K. D. and A. P. Wolffe (2000). "DNA methylation in health and disease." Nat Rev Genet **1**(1): 11-19.

Robinson, J. T., H. Thorvaldsdottir, et al. (2011). "Integrative genomics viewer." Nat Biotechnol **29**(1): 24-26.

Robinson, M. D., D. J. McCarthy, et al. (2010). "edgeR: a Bioconductor package for differential expression analysis of digital gene expression data." Bioinformatics **26**(1): 139-140.

Robinson, M. D. and A. Oshlack (2010). "A scaling normalization method for differential expression analysis of RNA-seq data." Genome Biol **11**(3): R25.

Rosenfeld, J. A., Z. Wang, et al. (2009). "Determination of enriched histone modifications in non-genic portions of the human genome." BMC Genomics **10**: 143.

Ruffalo, M., T. LaFramboise, et al. (2011). "Comparative analysis of algorithms for next-generation sequencing read alignment." Bioinformatics **27**(20): 2790-2796.

Sakaue, M., H. Ohta, et al. (2010). "DNA methylation is dispensable for the growth and survival of the extraembryonic lineages." Curr Biol **20**(16): 1452-1457.

Salmon-Divon, M., H. Dvinge, et al. (2010). "PeakAnalyzer: genome-wide annotation of chromatin binding and modification loci." BMC Bioinformatics **11**: 415.

Sanger, F., S. Nicklen, et al. (1977). "DNA sequencing with chain-terminating inhibitors." Proc Natl Acad Sci U S A **74**(12): 5463-5467.

Sasai, N. and P. A. Defossez (2009). "Many paths to one goal? The proteins that recognize methylated DNA in eukaryotes." Int J Dev Biol **53**(2-3): 323-334.

Sasai, N., M. Nakao, et al. (2010). "Sequence-specific recognition of methylated DNA by human zinc-finger proteins." Nucleic Acids Res **38**(15): 5015-5022.

Sawyer, J. R., C. M. Swanson, et al. (1995). "Centromeric instability of chromosome 1 resulting in multibranched chromosomes, telomeric fusions, and "jumping translocations" of 1q in a human immunodeficiency virus-related non-Hodgkin's lymphoma." Cancer **76**(7): 1238-1244.

Scarano, M. I., M. Strazzullo, et al. (2005). "DNA methylation 40 years later: Its role in human health and disease." J Cell Physiol **204**(1): 21-35.

Schuettengruber, B., D. Chourrout, et al. (2007). "Genome regulation by polycomb and trithorax proteins." Cell **128**(4): 735-745.

Sharif, J., M. Muto, et al. (2007). "The SRA protein Np95 mediates epigenetic inheritance by recruiting Dnmt1 to methylated DNA." Nature **450**(7171): 908-912.

Shukla, S., E. Kavak, et al. (2011). "CTCF-promoted RNA polymerase II pausing links DNA methylation to splicing." Nature **479**(7371): 74-79.

Statham, A. L., M. D. Robinson, et al. (2012). "Bisulfite sequencing of chromatin immunoprecipitated DNA (BisChIP-seq) directly informs methylation status of histone-modified DNA." Genome Res **22**(6): 1120-1127.

Straussman, R., D. Nejman, et al. (2009). "Developmental programming of CpG island methylation profiles in the human genome." Nature structural & molecular biology **16**(5): 564-571.

Suzuki, M., T. Yamada, et al. (2006). "Site-specific DNA methylation by a complex of PU.1 and Dnmt3a/b." Oncogene **25**(17): 2477-2488.

Tachibana, M., Y. Matsumura, et al. (2008). "G9a/GLP complexes independently mediate H3K9 and DNA methylation to silence transcription." Embo J **27**(20): 2681-2690.

Tahiliani, M., K. P. Koh, et al. (2009). "Conversion of 5-methylcytosine to 5-hydroxymethylcytosine in mammalian DNA by MLL partner TET1." Science **324**(5929): 930-935.

Tarazona, S., F. Garcia-Alcalde, et al. (2011). "Differential expression in RNA-seq: a matter of depth." Genome Res **21**(12): 2213-2223.

Thorvaldsdottir, H., J. T. Robinson, et al. (2013). "Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration." Brief Bioinform **14**(2): 178-192.

Tiepolo, L., P. Maraschio, et al. (1979). "Multibranched chromosomes 1, 9, and 16 in a patient with combined IgA and IgE deficiency." Hum Genet **51**(2): 127-137.

Trapnell, C., L. Pachter, et al. (2009). "TopHat: discovering splice junctions with RNA-Seq." Bioinformatics **25**(9): 1105-1111.

Trapnell, C., B. A. Williams, et al. (2010). "Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation." Nat Biotechnol **28**(5): 511-515.

Tuck-Muller, C. M., A. Narayan, et al. (2000). "DNA hypomethylation and unusual chromosome instability in cell lines from ICF syndrome patients." Cytogenet Cell Genet **89**(1-2): 121-128.

Tyekucheva, S., R. H. Yolken, et al. (2011). "Establishing the baseline level of repetitive element expression in the human cortex." BMC Genomics **12**: 495.

Unoki, M., T. Nishidate, et al. (2004). "ICBP90, an E2F-1 target, recruits HDAC1 and binds to methyl-CpG through its SRA domain." Oncogene **23**(46): 7601-7610.

Vakoc, C. R., M. M. Sachdeva, et al. (2006). "Profile of histone lysine methylation across transcribed mammalian chromatin." Molecular and cellular biology **26**(24): 9185-9195.

Venter, J. C., M. D. Adams, et al. (2001). "The sequence of the human genome." Science **291**(5507): 1304-1351.

Vire, E., C. Brenner, et al. (2006). "The Polycomb group protein EZH2 directly controls DNA methylation." Nature **439**(7078): 871-874.

Wang, Y. A., Y. Kamarova, et al. (2005). "DNA methyltransferase-3a interacts with p53 and represses p53-mediated gene expression." Cancer biology & therapy **4**(10): 1138-1143.

Yoon, H. S., C. D. Scharer, et al. (2012). "ZBTB32 is an early repressor of the CIITA and MHC class II gene expression during B cell differentiation to plasma cells." J Immunol **189**(5): 2393-2403.

Yu, F., J. Thiesen, et al. (2000). "Histone deacetylase-independent transcriptional repression by methyl-CpG-binding protein 2." Nucleic Acids Res **28**(10): 2201-2206.

Zang, C., D. E. Schones, et al. (2009). "A clustering approach for identification of enriched domains from histone modification ChIP-Seq data." Bioinformatics **25**(15): 1952-1958.

Zhang, Y., R. Jurkowska, et al. (2010). "Chromatin methylation activity of Dnmt3a and Dnmt3a/3L is guided by interaction of the ADD domain with the histone H3 tail." Nucleic acids research **38**(13): 4246-4253.

Zhu, L. J., C. Gazin, et al. (2010). "ChIPpeakAnno: a Bioconductor package to annotate ChIP-seq and ChIP-chip data." BMC Bioinformatics **11**: 237.

Zilberman, D., M. Gehring, et al. (2007). "Genome-wide analysis of Arabidopsis thaliana DNA methylation uncovers an interdependence between methylation and transcription." Nature genetics **39**(1): 61-69.

Appendix A

Here listed are some examples of codes in R or Perl that have been used in data analysis for this work.

Differential peaks enrichment evaluation using DESeq package.

```
library("preprocessCore")

setwd("/Users/path/m3_Counts")

#Load counts files
ICF1K27=read.table("Counts_ICF1_K27_all_regions.coverage")
ICF2K27=read.table("Counts_ICF2_K27_all_regions.coverage")
ICF3K27=read.table("Counts_ICF3_K27_all_regions.coverage")
LDA1K27=read.table("Counts_LDA1_K27_all_regions.coverage")
LDA2K27=read.table("Counts_LDA2_K27_all_regions.coverage")
MOM1K27=read.table("Counts_MOM1_K27_all_regions.coverage")
MOM2K27=read.table("Counts_MOM2_K27_all_regions.coverage")
MOM3K27=read.table("Counts_MOM3_K27_all_regions.coverage")

#Create reference file of regions with fictious names and matrix of counts
regK27= paste("Region",1:nrow(ICF1K27), sep="")
regionsK27=data.frame(ICF1K27[,1:3], row.names=regK27)
write.table(regionsK27, file="regionsK27.txt", sep="\t", dec=",", quote=FALSE)

M_K27=data.frame(ICF1=ICF1K27[,4],         ICF2=ICF2K27[,4],         ICF3=ICF3K27[,4],
 LDA1=LDA1K27[,4],      LDA2=LDA2K27[,4],      MOM1=MOM1K27[,4],      MOM2=MOM2K27[,4],
 MOM3=MOM3K27[,4] row.names=regK27)
write.table(N_K27, file="MatrixK27.txt", sep="\t", dec=",", quote=FALSE)

# Evaluate enrichment differences:
library("DESeq")
print("Running DESeq...")

# estimate variance - differs based on if you have replicates or not...
countsTableMatrixK27 <- as.matrix(M_K27)
condsK27 = c("I", "I", "I", "L", "L", "M", "M","M")
dlistK27 = unique(condsK27)
cdsK27 <- newCountDataSet(countsTableMatrixK27, condsK27)
cdsK27 <- estimateSizeFactors(cdsK27)
cdsK27 <- estimateDispersions(cdsK27, method="pooled")

# Calculate differential enrichment between all pairs of samples
tmax = length(dlistK27)
for (i in 1:tmax)
{
 for (j in 1:tmax)
 {
      if (i != j)
      {
            a = dlistK27[i]
            b = dlistK27[j]
#First create the output folder where to save file!
            tname = paste("DESeq24Nov12/K27/", a, "-", b, ".deseq", sep="")
            print(tname)
            res = nbinomTest(cdsK27, a, b)
            write.table(file = tname, res, quote = F, sep="\t", row.names=F)
            res_Pvalue0.01 <- res[res$pval < 0.01, ]
            tname2  =  paste("DESeq24Nov12/K27/",  a,  "-",  b,  "pValue0.01.deseq",
 sep="")
            print(tname2)
            #Save only regions with p-value<0.01
```

```
            write.table(file = tname2, res_Pvalue0.01, quote = F, sep="\t",
 row.names=F)
            res_Pvalue0.05 <- res[res$pval < 0.05, ]
            tname4 = paste("DESeq24Nov12/K27/", a, "-", b, "pValue0.05.deseq",
 sep="")
            print(tname4)
            #Save only regions with p-value<0.05
            write.table(file = tname4, res_Pvalue0.05, quote = F, sep="\t",
 row.names=F)
            tname3 = paste("DESeq24Nov12/K27/", a, "-", b, ".jpg", sep="")
            jpeg(file=tname3)
            plot(res$baseMean, res$log2FoldChange, log="x", pch=20, cex=.1, col =
 ifelse( res$pval < .01, "red", "black" ) )
            dev.off()
      }
 }
}

print("Done K27!")
```

Annotation of peaks on genomic features.

```
# RunChIPpeakAnno.R
################################################
### Script for annotation
################################################
# Parse and use command line arguments
# Invoke % R --slave --args path/fileame.bed < RunChIPpeakAnno.R

#take arguments
Args <- commandArgs()
regions <- read.delim(Args[4], header=FALSE, stringsAsFactors=TRUE)
library(ChIPpeakAnno)

#extract regions information
starts = regions$"V2"
ends = regions$"V3"
chrs = regions$"V1"

#create an IRange object
myPeak = RangedData(IRanges(start=starts, end=ends), space=chrs)

#Extract data from human genome GRCh37/hg19
data(TSS.human.GRCh37)

print("running annotation")
annotatedPeak = annotatePeakInBatch(myPeak, AnnotationData=TSS.human.GRCh37,
 PeakLocForDistance="middle", FeatureLocForDistance="TSS")

print(head(as.data.frame(annotatedPeak)))

print("writing tables and images")
input=as.character(Args[4])
name=unlist(strsplit(input, "\\."))

jpeg(paste(name[1],'_pie.jpg'))
pie(table(as.data.frame(annotatedPeak)$insideFeature))
dev.off()

write.table(as.data.frame(annotatedPeak), file=
 (paste(name[1],"_annotatedPeakList.xls")), sep="\t", dec=",", row.names=FALSE)

anno=as.data.frame(addGeneIDs(annotatedPeak, "org.Hs.eg.db", "symbol"))
```

```
write.table(anno[anno$distancetoFeature < 5000 & anno$distancetoFeature> -5000,],
 file= (paste(name[1],"_associatedPeaks5000.xls")), sep="\t", dec=",",
 row.names=FALSE)
write.table(anno[anno$distancetoFeature < 10000 & anno$distancetoFeature> -10000,],
 file= (paste(name[1],"_associatedPeaks10000.xls")), sep="\t", dec=",",
 row.names=FALSE)

jpeg((paste(name[1],'_pie_associated.jpg')))
pie(table(as.data.frame(anno[anno$distancetoFeature < 5000 & anno$distancetoFeature>
 -5000,])$insideFeature))
dev.off()

jpeg((paste(name[1],'_hist_TSS.jpg')))
y1=annotatedPeak$distancetoFeature
 [!is.na(annotatedPeak$distancetoFeature)&annotatedPeak$fromOverlappingOrNearest ==
 "NearestStart"]
hist(y1, xlab="Distance To Nearest TSS", main="", breaks=1000, xlim=c(min(y1)-100,
 max(y1)+100))
dev.off()

jpeg((paste(name[1],'_hist_TSS_zoom.jpg')))
hist(y1, xlab="Distance To Nearest TSS", main="", breaks=10000, xlim=c(-2e+04,
 2e+04))
dev.off()

print("Done!!!")
```

Perl script for reads count in repetitive regions from RepeatMask - Adapted from A.Brinkman

```
#!/usr/bin/perl
#
use warnings;
use strict;
use Getopt::Long;
use File::Basename;

our $title = "repeatAnalysisSole.pl v1.0 Sun Feb 13 23:13:21 CET 2011 -- Sole Gatto -
- Adapted from Arjen Brinkman";
our $scriptname = "repeatAnalysisSole.pl";

#set some general variables
my $pathRefGenome = "/home/gatto/rseg/hg19/hg19.fasta";
my $pathToRepeatClasses = "/home/gatto/RepeatMask_hg19";
my $name;
my $infile;
my $outfile;
my $totalReadsNoDuplicates;

#Process the options
our($opt_f, $opt_r);
&Process_Options;

my $totalReadsMapped = `cat $opt_f | wc -l`;
chomp $totalReadsMapped;

print "$totalReadsMapped\n";

#create temporary storage place
our $tmp=`mktemp -d`;
chomp $tmp;

#count reads within repeatclasses
```

```
&getName;

system `cat $pathToRepeatClasses |grep -v "#" |sort -k1,1 -k2g,2 |cut -f1,2,3
>"$tmp"/repeatclasses`;
system `cat $pathToRepeatClasses |grep -v "#" |sort -k1,1 -k2g,2 |cut -f4-
>"$tmp"/repeatclassNames`;
system `/usr/bin/peakstats.py -p "$tmp"/repeatclasses -d $opt_f -z -f number |cut -f4
>"$tmp"/tagcountInRepeats.peakstats`;
system `paste "$tmp"/repeatclassNames "$tmp"/tagcountInRepeats.peakstats
>"$tmp"/countedRepeats`;

#compile counted data
my %nameCount;
my %classCount;
my %familyCount;
$infile = $tmp . "/countedRepeats";
open IN, "<$infile" or die "Could not open $infile:$!\n";
while (<IN>) {
        chomp $_;
        my @line = split(/\t/, $_);
        my $repName = $line[0];
        $repName =~ s/\?//g;
        $repName =~ s/_$//;
        my $repClass = $line[1];
        $repClass =~ s/\?//g;
        $repClass =~ s/_$//;
        my $repFamily = $line[2];
        $repFamily =~ s/\?//g;
        $repFamily =~ s/_$//;
        my $readCount = $line[3];
        $nameCount{$repName} += $readCount;
        $classCount{$repClass} += $readCount;
        $familyCount{$repFamily} += $readCount;

}
close(IN);

#divide readcount by total mapped reads / 1 million
my $corrFactor = $totalReadsMapped/1000000;

#print name count
&getName;
$outfile = $name . ".repeatCount";
open OUT, ">$outfile" or die "Could not open $outfile:$!\n";
my @names = keys %nameCount;
@names = sort(@names);
foreach(@names) {
        my $corrCount = sprintf("%.2f", ($nameCount{$_}/$corrFactor));
      print OUT "$_\t$corrCount\n";
}
close(OUT);

#print class count
&getName;
$outfile = $name . ".repeatClassCount";
open OUT, ">$outfile" or die "Could not open $outfile:$!\n";
my @classes = keys %classCount;
@classes = sort(@classes);
foreach(@classes) {
      my $corrCount = sprintf("%.2f", ($classCount{$_}/$corrFactor));
      print OUT "$_\t$corrCount\n";
}
close(OUT);

#print family count
```

```
&getName;
$outfile = $name . ".repeatFamilyCount";
open OUT, ">$outfile" or die "Could not open $outfile:$!\n";
my @families = keys %familyCount;
@families = sort(@families);
foreach(@families) {
      my $corrCount = sprintf("%.2f", ($familyCount{$_}/$corrFactor));
        print OUT "$_\t$corrCount\n";
}
close(OUT);

system `rm -r $tmp`;

# Process the flags
sub Process_Options {
        my $stop;
        my @flags;
        my $tempopt;
        my $flag;

        GetOptions(
                "b=s" => \$opt_f,
       "r=s" => \$opt_r,
        );

        # Check if the essential options are set
        if ( !$opt_f ) {
                $stop .= "--- Please specify input file   ---\n";
        }
        if ($opt_f && !-e $opt_f) {
                $stop .= "--- ERROR: $opt_f does not exist   ---\n";
        }
      if (!$opt_r) {
        $opt_r = 1e06;
      }
      if (!-e $pathRefGenome) {
        $stop .= "--- ERROR: $pathRefGenome does not exist   ---\n";
      }
      if (!-e $pathToRepeatClasses) {
        $stop .= "--- ERROR: $pathToRepeatClasses does not exist   ---\n";
      }
        if ( $stop ) {
                print "$stop\n";
                &Usage("I");
        }
}

# Print the program usage
sub Usage {
        my $flag = shift @_;
        if ( $flag eq "I" ){
                print "$title\n\n";
                print "usage: $scriptname -f inputfile\n\n";
                print "\t-b bed file containing aligned reads (chr start end)\n\n";
       print "\t-r number of randomly selected reads to process (default = 1
million)\n\n";
                print "3 output files are produced, giving readcounts per
repeat/class/family\n";
       print "output represents readcounts per million mapped reads\n\n\n";
        }
        exit;
}


#get a name for outputfiles
```

```
sub getName {
        $name = basename($opt_f);
        $name =~ s/\.gz//i;
        $name =~ s/\.sequence\.txt//i;
        $name =~ s/_sequence\.txt//i;
        $name =~ s/sequence\.txt//i;
        $name =~ s/\.fastq//i;
}
```