Universidade de Aveiro	Departamento de Química
2018

CAROLINA
NEVES
CONCEIÇÃO

METILAÇÃO DIFERENCIAL DE DNA NO ENVELHECIMENTO: EXPLORAÇÃO IN SILICO UTILIZANDO DADOS DE ELEVADO RENDIMENTO

DIFFERENTIAL DNA METHYLATION IN AGING: IN SILICO EXPLORATION USING HIGH-THROUGHPUT DATASETS

Universidade de Aveiro **Departamento de Química**
2018

CAROLINA NEVES CONCEIÇÃO

METILAÇÃO DIFERENCIAL DE DNA NO ENVELHECIMENTO: EXPLORAÇÃO IN SILICO UTILIZANDO DADOS DE ELEVADO RENDIMENTO

DIFFERENTIAL DNA METHYLATION IN AGING: IN SILICO EXPLORATION USING HIGH-THROUGHPUT DATASETS

Dissertação apresentada à Universidade de Aveiro para cumprimento dos requisitos necessários à obtenção do grau de Mestre em Biotecnologia, realizada sob orientação científica da Doutora Gabriela Maria Ferreira Ribeiro de Moura, Professora Auxiliar do Departamento de Ciências Médicas da Universidade de Aveiro.

Financiado pelo Programa Operacional de Competitividade e Internacionalização COMPETE, através do Fundo Europeu de Desenvolvimento Regional FEDER (POCI-01-0145-FEDER-007628), Fundação para Ciência e Tecnologia (UID/BIM/04501/2013) e através do projeto MEDPERSYST.

o júri

presidente	Doutora Mara Guadalupe Freire Martins Investigadora Coordenadora do Departamento de Química da Universidade de Aveiro
	Doutor Bruno Miguel Bernardes de Jesus Professor Auxiliar do Departamento de Ciências Médicas da Universidade de Aveiro
	Doutora Gabriela Maria Ferreira Ribeiro de Moura Professora Auxiliar do Departamento de Ciências Médicas da Universidade de Aveiro

agradecimentos

Ao Professor Doutor Manuel Santos pela oportunidade de elaboração deste projeto integrado no iBiMED;

À Professora Gabriela Moura por toda a orientação e disponibilidade para ajudar a melhorar este projeto, mas também por todo o apoio e compreensão que demonstrou na elaboração desta tese.

À minha mãe por ser o meu maior-exemplo, pelo esforço e preocupação, pela dedicação que tem em mim. Ao meu pai, a minha estrelinha mais forte, por toda a força e pelo exemplo que foi e é no meu dia-a-dia. À minha irmã e ao meu irmão por serem as melhores pessoas que conheço; por toda a amizade, presença, preocupação e alegria que me dão. À minha família. Por ter unido forças. Por todo o apoio e todas as alegrias que me trazem todos os dias;

Ao Pedro pelo apoio incessante. Por ter sido sempre o primeiro a dizer que isto não era impossível. Por todo o amor mas principalmente por toda a amizade e alegria que traz ao meu dia, todos os dias. Ao Vasco, à Sara e às divas pela amizade crescente.

Ao BEST Aveiro por ter sido um dos meus maiores pilares de crescimento nesta academia. Por tudo o que me ensinou; por todos os projetos que me deram oportunidade de desenvolver e por me deixarem ver o meu maior limite e aquilo que mais gosto de fazer;

À Andreia pelo incansável apoio técnico, pelos conselhos e ajuda na definição deste projeto. À Sonya, à Vera, à Rita, à Margarida, ao Miguel e ao Hugo por toda a disponibilidade. Ao Vasco Cluny por toda a partilha e preocupação na concretização deste projeto.

palavras-chave

Tecnologia *microarray*, sequenciação de última geração, epigenómica, bioinformática, metilação de DNA, genómica, reação de bisulfito, programação epigenética, regulação genética.

resumo

O aparecimento de metodologias de sequenciação de elevado rendimento após a conclusão do Projeto do Genoma Humano foi um avanço fundamental para a pesquisa biológica e biomédica na área da genómica. Embora as mutações genéticas tenham sido durante décadas o foco principal na causa de certas desordens, atualmente demonstrou-se que os mecanismos epigenéticos estão envolvidos na programação celular e na regulação genética, providenciando variações adaptativas do mesmo gene a um determinado ambiente e possuindo ainda uma associação direta com a diferenciação celular.

O desenvolvimento científico no campo da metilação de DNA revela atualmente factos essenciais na biologia molecular, como a existência de metilação nas ilhas CpG e em contextos alternativos que influenciam a expressão genética nos diferentes tecidos humanos. Para além disso, a influência dos estilos de vida no processo de envelhecimento já demonstrou estar relacionada com o estado do epigenoma, nomeadamente com as variações no metiloma humano. No caso do cancro, a cooperação dos fatores genéticos e epigenéticos é essencial para a compreensão do desenvolvimento desta patologia no organismo humano nomeadamente através do silenciamento de genes reguladores essenciais. Uma hipometilação global no genoma do cancro conduz geralmente a uma ativação de oncogenes enquanto que hipermetilações localizadas estão associadas com o silenciamento de genes supressores de tumores. Por estes motivos, o desenvolvimento de novas terapias para o cancro ou o envelhecimento torna-se um tópico de interesse pela comunidade científica da área da epigenómica.

Com o objetivo de desenvolver estes temas e melhorar a determinação de variações globais no epigenoma humano, esta investigação desenvolveu-se com base na utilização de dados de bases de dados públicas de indivíduos saudaveis de forma a extrair marcadores de metilação diferenciada em variados tecidos ao longo do envelhecimento saudável. O projeto foi validado através da utilização de amostras saúdaveis e de indivíduos com boas ou más performances cognitivas disponíveis no iBiMED. Em ambas as situações os genes ELOVL2 (cg16867657) e FHL2 (cg06639320) foram identificados como bons marcadores da idade dos indivíduos.

keywords

Microarray tecnhnology, next-generation sequencing, epigenomics, bioinformatics, DNA methylation, genome-wide analysis, bisulfite conversion, epigenetic programming, genetic regulation.

abstract

The emergence of high-throughput methodologies after the conclusion of the Human Genome Project has brought genomic and epigenomic wide studies to the forefront of current research of biological and biomedical knowledge. Currently, the focus in genetic mutations as primary cause of certain disorders is not so relevant as before, since it was demonstrated that epigenetic mechanisms are involved in cellular programming and gene regulation providing adaptive variants of a given gene to a changing environment with an association to cellular differentiation.

The research in the DNA methylation field has already revealed essential facts as the existence of methylation in CpG islands and alternative contexts that influence gene expression in tissue-specific manner. The influence of lifestyle choices in aging processes has also been related to methylome variations. And, in the case of cancer, the cooperation of epigenetic and genetic information is essential to understand the progress of cancer development as well as the silencing of key regulatory genes. An overall hypomethylation in cancer genome leads to oncogene activation whereas hypermethylation in specific regions is associated with silencing of tumour suppressor genes. For that reason, the research for new therapeutic approaches to cancer and aging is a current issue of the scientific community that work in the epigenomic field.

In order to contribute to the study of mammalian epigenomes during lifespans, this research focused on the usage of public databases datasets to further investigation about DNA methylation across aged individuals in order to extract tissue-specific markers related with healthy aging. The validation of results was made through the usage of samples, form healthy individuals with good or bad cognitive performances, available in iBiMED. In both situations the genes ELOVL2 (cg16867657) and FHL2 (cg06639320) were identified as good markers of age.

CONTENTS

FIGURE AND TABLE INDEX

ABBREVIATIONS AND ACRONYMS

5caC	5-carboxylcytosine
5fC	5-formylcytosine
5hmC	5-hydroxymethylcytosine
5mC	5-methylcytosine
a-DMRs	Age-related differentially methylated regions
ABI	Applied Biosystems
ADP	Adenosine diphosphate
ATAC	Assay for transposable acessible chromatin
BER	Base excision repair
BGZF	Blocked GNU zip format
BiMP	Bisulfite methylation profiling
BMIQ	β-mixture quantile dilation
Bp	Base pair
BS	Bisulfite
BSC	Bisulfite Crick strand
BSCRC	Bisulfite Crick reverse complementary strand
BSW	Bisulfite Watson strand
BSWRC	Bisulfite Watson reverse complementary strand
CCD	Charge-coupled device
cDNA	Complementary DNA
CGI	CpG island
ChAMP	Chip analysis methylation pipeline
CHARM	Comprehensive high-throughput arrays for relative methylation
ChIP	Chromatin immunoprecipitation
CpG	Cytosine-Phosphate-Guanine
CRISPR	Clustered regularly interspaced short palindromic repeats
Da	Dalton
DAVID	Database for annotation, visualization and integrated discovery
DDBJ	DNA data bank of japan
ddNTPs	Dideoxynucleotides
DHEAS	5-dehydroepiandrosterone
DMH	Differential methylation hybridization
DMR	Differentially methylated region
DNA	Deoxyribonucleic Acid
DNase	DNase i digestion nuclease
DNMT	DNA methyltransferase
dNTP	Deoxynucleoside triphosphate
dsDNA	Double-stranded DNA
ENA	European nucleotide archive
EVORA	Epigenetic variable outlier for risk prediction analysis
EWAS	Epigenome-wide association study
FAIRE	Formaldehyde-assisted isolation of regulatory elements
FDR	False discovery rate
Gb	Gigabyte
GEO	Gene Expression Omnibus
GO	Gene ontology
GPL	GEO platform
GSEA	Gene Set Enrichment Analysis
GSM	GEO sample
HAT	Histone acetyltransferase
HDAC	Histone deacetylase
HELP	*HpaII* tiny fragment enrichment by ligation-mediated PCR assay
HGP	Human Genome Project
HMT	Histone methyltransferase
HPLC	High-performance liquid chromatography
HTML	HyperText markup language
IGV	Integrative genomics viewer
IMA	Illumina methylation analyzer
INSD	International Nucleotide Sequence Database
ISVA	Independent Surrogate Variable Analysis
IUPAC	International Union of Pure and Applied Chemistry
KLF14	Krüppel-like factor 14

lnRNA	Long non-coding RNA
MAD	Methylation amplification DNA chip
MBD2	Methyl-CpG-binding domain protein 2
MBD	Methyl Binding Domain
MCA	Methylated CpG island amplification
MCAM	Methylated CpG island amplification microarray
MDS	Multidimensional scaling
MeCIP	Methyl-CpG immunoprecipitation
MECP2	Methyl-CpG-binding protein 2
MeDIP	Methylated DNA Immunoprecipitation
MIRA	Methylated CpG island recovery assay
miRNA	Micro-RNA
MRE	Methylation-sensitive restriction enzymes
MNase	Micrococcal Nuclease
NCBI	National Center for Biotechnology Information
ncRNA	Non coding RNA
NGS	Next-Generation Sequencing
NIH	National Institute of Health
NLM	National Library of Medicine
nt	Nucleotide
NuRD	Nucleosomal Remodelling Complex
PBAT	Post-bisulfite adapter tagging
PBMC	Peripheral blood mononuclear cell
PCR	Polymerase chain reaction
PE	Paired-end read
PGM	Personal genome machine
PMAD	Promoter-associated methylated DNA amplification DNA-chip assay
PMD	Partially methylated domain
$\mathbf{PP_i}$	Pyrophosphate
PTP	Picotiter Plate
QC	Quality control
RNA	Ribonucleic Acid
RRBS	Reduced Representation Bisulfite Sequencing
rRNA	Ribosomal RNA
SAM	Sequence Alignment Map/Sentrix Array Matrix
SBL	Sequencing-by-ligation
SBS	Sequencing-by-synthesis
scPBAT	Single-cell post-bisulfite adapter tagging
scRRBS	Single-cell reduced representation bisulfite sequencing
ssWGBS	Single-cell whole genome bisulfite sequencing
SE	Single-end read
siRNA	Small interfering RNA
SMRT	Single-molecule real time
snoRNA	Small nucleolar RNA
SNP	Single nucleotide polymorphism
SNV	Single nucleotide variant
SOFT	Simple Omnibus Format in Text
SOLiD	Sequencing by Oligonucleotide Ligation and Detection
SRA	Sequence Read Archive
ssDNA	Single-stranded DNA
SST	Gene Somatostatin
SVA	Surrogate Variable Analysis
SWI/SNF	Switch/Sucrose Non-Fermentable
TALE	Transcription activator-like effectors
Tc cells	Cytotoxic T cells
TCGA	The Cancer Genome Atlas
TDG	Thymine-DNA glycosylase
T-DMR	Tissue-specific Differentially Methylated Regions
TET	Ten eleven translocation enzymes
Th cells	T-helper cells
TLC	Thin-layer chromatography
tRNA	Transfer RNA
UCSC	University of California, Santa Cruz
uiDMR	Undefined intragenic DMRs
UTR	Untranslated region
SWAN	Subset-quantile within array normalization
UV	Ultraviolet light
WGBS	Whole Genome Bisulfite Sequencing

ZFP	Zinc Finger Protein
ZMW	Zero Mode Waveguide

GENERAL OVERVIEW

Since the beginning of the Human Genome Project (HGP) in 1990 until 2003 (Wilson & Nicholls 2015), a global effort was made in order to sequence and map the majority of the euchromatic portion of the human genome(International Human Genome Sequencing Consortium 2004). The advances achieved in the HGP were one of the major scientific endeavours in modern scientific research(Wilson & Nicholls 2015) since they gave access to a large domain of important biological and biomedical knowledge(International Human Genome Sequencing Consortium 2004). Previously, few genes were used to investigate the patterns of genetic variation among individuals, but the advances in sequencing technology made it possible to study genome wide variation among individuals relatively to several biological conditions(Borevitz et al. 2015). These achievements were possible due to the development of next generation sequencing (NGS) technologies, which are now available to the scientific world(Wilson & Nicholls 2015).

One of the most recent fields in the genome wide analysis has been the epigenomics where is made an analysis of the global patterns of cytosine methylation, chromatin state and non-coding RNA abundance(Friedman & Rando 2015). For many decades much focus was placed on genetic mutations as primary cause of certain disorders. However, in the last years the study of epigenetic mechanisms in the mammalian genome has demonstrated its influence on several cellular events as gene expression regulation, cell programming and differentiation as well as at the organism level, such as development, disease and aging (Bell et al. 2012)(Bollati et al. 2010).

Therefore, there is not only a current need in analyzing the epigenetic patterns among individuals in order to determine its influence in a specific biological condition but also a demand on high-throughput technologies that are able to tackle this problem at low cost, short time and with effective alignment and variant call tools.

This Master's Thesis aimed to:

a) Perform a rigorous literature review on the field of epigenomics, focusing on different available methodologies in a comparative perspective and recapitulating major findings relative to the evaluation of epigenetical marks with age;
b) Build a genomic map of age-dependent epigenetic markers in mammals, based on public data and state-of-the-art high-throughput methodologies that would allow to

reduce the costs of routine analysis at iBiMED, by turning from whole genome to targeted-based analysis, focusing only on those relevant regions of the genome;

c) Validate the bioinformatics pipeline build by us for methylome analysis, using in-house generated data, so that this methodology can become available at iBiMED.

CHAPTER I INTRODUCTION

Epigenomics and gene regulation

1. EPIGENOMICS

1.1. Overview

Despite an identical genetic background, different cell types execute distinct programmes of gene expression highly influenced by developmental, physiological and environmental stimulus, which means that the marks of developmental history are unlikely to be caused by widespread somatic mutations(McGowan & Szyf 2010),(Bird 2002). This evidence brings us the concept of epigenetics that, by definition, is *the study of mitotically and/or meiotically heritable changes in gene function that cannot be explained by changes in DNA sequence and that are potentially reversible*(Bird 2002).

One of the main functions of epigenetic processes is the packaging of genetic information in the nucleus of eukaryotic cells. Each diploid cell with 46 chromosomes contains approximately 6 billion base pairs (bp) of DNA that are condensed by histone proteins, the main characters of the organization of genetic information in cells. Additionally, the covalent modifications of histone proteins and DNA cytosine methylation-state are associated with different forms and functions of chromatin that regulate gene expression. Non-coding RNAs (ncRNAs) have also emerged as important epigenetic regulators in crucial biological processes as differentiation and development(Falahi et al. 2015).

Currently, the scientific community has been focused on the study of the phenotypic differences across humans by crossing it with genetic and epigenetic variations(McGowan & Szyf 2010). Until the emergence of epigenomics, the researchers were able to identify the genes that contributed to a particular trait or phenotype. However, if the variant were located in a non-coding region of DNA, sometimes there weren't insights into the regulatory mechanisms underlying the association. In these cases, providing the missing connection between genomic variation and cellular phenotype was essential(Romanoski et al. 2015). The role that epigenetics has in gene regulation it's important to study the adaptive variants of a given gene to a changing environment but it might also be associated with cellular differentiation(Mensaert et al. 2014),(Bird 2002). Since cellular differentiation is a prerequisite for complex multicellular organisms, than it is expected that almost all species in this category can benefit from epigenetic control(Mensaert et al. 2014).

Therefore, it is crucial that the studies related with phenotypic diversity consider epigenetic variations in addition to genetic sequence polymorphisms, not only to identify pathology associated epigenetic aberrations but also to understand how these marks are patterned across the genome and how these mechanisms can control biological processes(Mensaert et al. 2014),(McGowan & Szyf 2010).

1.2. DNA methylation

One example of an essential epigenetic process involved in cellular development, differentiation and regulation is methylation of cytosines (5mC) in DNA. This enzymatic reaction (Figure 1) is mediated by DNMTs (DNA methyltransferases) that transfer a methyl group from SAM (S-adenosyl-L-methionine) to the carbon 5 of a cytosine. Since carbon 5 is a weak nucleophile unable to interact with SAM on its own, a nucleophile from DNMT can attack the carbon 6 of the cytosine, covalently binding the enzyme to the DNA which will activate the nucleophilic character of the carbon 5, facilitating the transfer of the methyl group from SAM. After this, the enzyme nucleophile is eliminated and deprotonation at the carbon 5 separated the nucleotide-DNMT complex(Johnson et al. 2012).

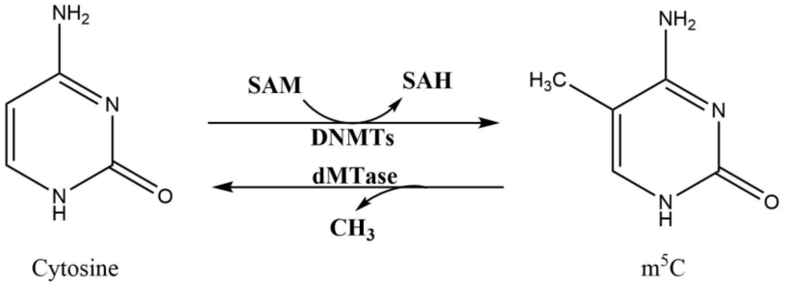


Figure 1 - Reversible enzymatic reaction of cytosine methylation. 5-methylcytosine (5mC) is obtained using DNMTs and a methyl donor (SAM) while cytosine can be obtained from 5mC through dMTases. Adapted from (McGowan & Szyf 2010)

The most common form of DNA methylation is present in C-G dinucleotides, referred to as CpGs(Jones et al. 2015). The regions with high density of CpGs are known as CpG islands (CGI), regions with more than 200 base pairs (bp), with a percentage of guanine and cytosine above 50% and 0.6 observed/expected ratio of CpGs (Jones et al. 2015). The regions immediately surrounding CGIs are referred to as shores followed by shelves (Jones et al. 2015). Previously, the scientific community thought that methylated CGI were unmethylated in normal cells, with the exception of those that were associated with imprinted genes and genes on the inactive X chromosomes. However, it was shown that non-imprinted autosomal CpG islands are methylated in normal cells and might use this mechanism for the control of gene expression (Laird 2003). Additionally, most methylated cytosine residues are found in CpG dinucleotides that are located outside of CpG islands and although the methylation in some CGI increase with age, the global genomic content of 5mc decreases with age (Laird 2003). In mammalian somatic cells, 5mC accounts for 1% of total DNA bases and affects 70-80% of CpG dinucleotides of the genome(Bird 2002).

1.2.1 Mechanisms and machinery

By definition, we know that the epigenetic features can be inherited, however methylation patterns are not copied by the DNA replication machinery (McGowan & Szyf 2010)'(Toyota et al. 2009). There are three DNMTs (DNMT1, DNMT3A, DNMT3B) which catalyze the methylation of a variety of genes, including genes involved in cell-cycle checkpoints, apoptosis, DNA repair, cell adhesion and signal transduction (Toyota et al. 2009).

This enzymatic machinery can be involved in two different methylation processes (Figure 2): maintenance of methylation patterns and *de novo* methylation. DNMT1 is the main enzyme responsible for the post-replicative restoration of the full methylation sites using a process called maintenance methylation. This procedure allows the reproduction of DNA methylation patterns through cell generations since it depends on semiconservative copying of the parental strand methylation pattern to the offspring DNA strand (Holliday & Pugh 1975)'(A.D. Riggs 1975).

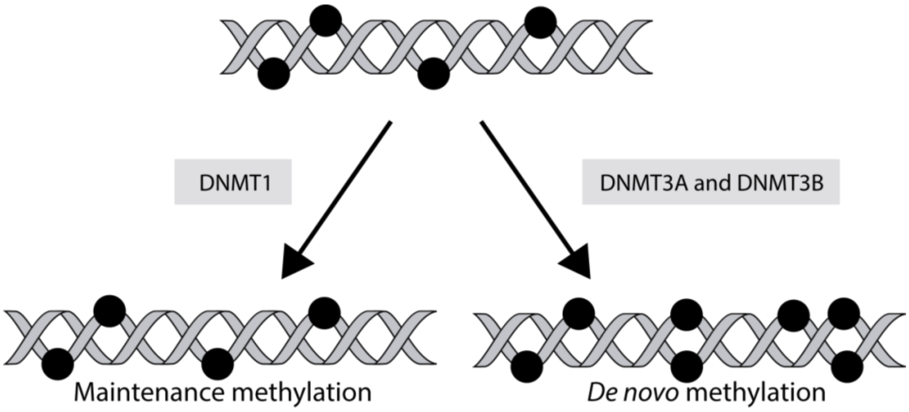


Figure 2 - DNA methylation mechanisms. Maintenance methylation by DNMT1 and de novo methylation by DNMT3A and DNMT3B. Adapted from (McGowan & Szyf 2010)

On the other hand, *de novo* methylation is characterized by the appearance of new DNA methylated spots by DNMT3A and DNMT3B(McGowan & Szyf 2010)'(Okano et al. 1999). *De novo* methylation events occur in germ or early embryonic cells but it can also be present in adult somatic cells and not all regions of the genome are equally accessible to DNA methyltransferases(Bird 2002).

1.2.2 DNA methylation patterns

Traditionally, the majority of genomic 5mC lies in CpG sites within CGI located in transposable repetitive elements and also in promoters(Schroeder et al. 2011). However there are recent evidences suggesting that methylation can also be found in alternate contexts including CHG and CHH (where H indicate non-G nucleotides)(Lister et al. 2009).. The human genome has about 60% of human genes associated with CGI(Antequera & Bird 1993).

For that reason, the driving force in DNA methylation studies has been particularly focused on CpG islands methylation in view of its demonstrated ability to silence genes in mammalian cells(Jones & Baylin 2007). However, it is worth noting that about 40% of human genes do not contain CpG islands in their promoters(Takai & Jones 2002). The most recent genome-wide analysis has been investigating the role of methylation in non-CpG islands because its mechanistic links have not been so well demonstrated and recent work have shown strong correlations between tissue-specific expression and methylation of non-CpG islands(Jones & Baylin 2007).

Tissue-specific differential methylated regions (T-DMR) have been reported in several human tissues along with partially methylated domains (PMD) and allele-specific methylation(Schultz et al. 2015). Therefore, since it is known that the methylation pattern is a balance of methylation and demethylation events which are responsible for a relationship between gene expression and environmental signals(McGowan & Szyf 2010); the functional consequences of DNA methylation as well as its interactions with the transcriptional machinery have been investigated in order to understand the diversity of human tissues and its relation with disease(Laird 2003),(Schultz et al. 2015).

As expected, DNA methylation of promoter regions is negatively associated with gene expression(Schroeder et al. 2011), whereas gene-body methylation has been reported to positively correlate with gene-expression levels(Tsai et al. 2012),(Parle-McDermott & Ozaki 2011). Often, DNA methylation also functions to repress repetitive elements, such as Alu and LINE-1, which are generally highly methylated in the human genome(Jones et al. 2015).

Cancer is an example of DNA methylation patterning that has medical interest and is associated with a number of genome-wide alterations. In this case, a global hypomethylation is related with oncogene activation(Wu et al. 2005) whereas hypermethylation is associated with tumor suppressor gene silencing(Esteller 2002). Then, since the epigenetic alterations are more readily reversible than genetic events, DNA methylation markers might be a promising future in both clinical diagnostics and therapeutics and also in the area of molecular diagnosis and early detection(Laird 2003).

1.2.3 DNA Demethylation

The DNA demethylation process, or loss of DNA methylation (concept known as hypomethylation), is also important to study the global DNA methylation patterns since it has been already observed in different biological contexts like mammalian embryogenesis or in specific loci in rapid response to environmental stimuli or in post-mitotic cells(Kohli & Zhang 2013). DNA demethylation can occur actively through an enzymatic process that removes or modifies the methyl-group in 5mC or passively through subsequent rounds of replication that does not replicate the 5mC in previous generation(Kohli & Zhang 2013).

The study of active demethylation, however, has been technically challenging and there are several proposed mechanisms to study it. Currently, the most convincing method involves the study of 5-hydroxymethylcytosine (5hmC), the key intermediate in active demethylation pathways. Briefly, the ten eleven translocation (TET) enzymes are known to be responsible for the oxidation of 5mC to 5hmC, to 5-formylcytosine (5fC) and finally to 5-carboxylcytosine (5caC) that can be descarboxylated by thymine-DNA glycosylase (TDG) regenerating normal cytosine. On the other hand, others have used the base excision repair (BER) mechanism that remove an entire modified base replacing it by an unmodified cytosine(Johnson et al. 2012),(Kohli & Zhang 2013).

1.3. Histone modifications

The base element of chromatin is the nucleosome that is the basis for packaging of genetic information in the nucleus of eukaryotic cells. This chromatin structure is made up of two copies of each of the four core histones (H3, H4, H2A, H2B) around which 146 bp of DNA are wrapped(Kornberg & Lorch 1999). The histone proteins are modified by methylation(McGowan & Szyf 2010),(Lehninger et al. 2005), phosphorylation(McGowan & Szyf 2010),(Lehninger et al. 2005), acetylation(McGowan & Szyf 2010),(Lehninger et al. 2005), ubiquitination(McGowan & Szyf 2010) and ADP-ribosylation(Lehninger et al. 2005) with consequences in the accessibility of the DNA wrapped around the nucleosome core(McGowan & Szyf 2010).

The histone proteins are evolutionarily conserved proteins characterized by molecular weights between 11 000 and 21 000 Da(Lehninger et al. 2005). They have an accessible amino terminal tail and a histone fold domain that mediates interactions between histones to form the nucleosome scaffold(Luger et al. 1997). These proteins are very rich in the basic amino acids arginine and lysine(Lehninger et al. 2005) and the ones involved in DNA compaction and chromatin remodeling are H1, H2A, H2B, H3 and H4(Falahi et al. 2015). The DNA backbone, negatively charged, interacts with these proteins, positively charged, thus blocking the interaction of transcription factors with the DNA(McGowan & Szyf 2010). Histone post-transcriptional modifications are reversible and are added by several enzymes like HATs (histone acetyltransferases), HDACs (histone deacetylases) and HMTs (histone methyltransferases)(McGowan & Szyf 2010). In general, gene repression is associated with H3K27me3 (designates 3 methylation groups on lysine 27 in the histone H3 tail) and H3K9me2/3 while active gene expression is associated with H3K4me3 and H3/H4 acetylation(Falahi et al. 2015),(Barski et al. 2007). Although the vast majority of these modifications remain poorly understood, the histone code, which postulates that a specific combination of modifications affects gene expression, is becoming unveiled in order to understand its roles in transcriptional regulation(Hon et al. 2009),(Suganuma & Workman 2011).

Histone modifications and DNA methylation are not independent events since global hypomethylation might lead to global alterations in histone acetylation and vice versa. It was already shown that cytosine methylation could attract methylated DNA binding proteins and histone deacetylases to methylated CpG islands during chromatin compaction and gene silencing(Jones & Baylin 2007),(Jones et al. 1998). The interplay between DNA methylation, histone covalent modifications and nucleosomal remodeling is involved in heritable gene repression at the start site of several genes, resulting in gene silencing. As an example, it is known that the nucleosomal remodeling complex (NuRD) and the SWI/SNF chromatin remodeling complex interact with DNA methylation binding proteins(Zhang et al. 1999).

1.4. Non-coding RNAs

Although only a small percent of the total amount of RNA is protein coding, up to 75% of the human genome is known to be transcribed into RNA(Djebali et al. 2012). Until recently, most of the known ncRNAs were associated with cell functions, i.e. rRNAs and tRNAs were involved in translation, snRNAs were involved in splicing and snoRNAs were involved in the modification of rRNAs(Mattick & Makunin 2006). Currently, there is an increasing number of non-coding RNA that function in association with introns and UTR(Daniel et al. 2015) that have been shown to regulate gene expression in response to stress and environmental stimuli as miRNAs, siRNAs and lnRNAs, among others(Kaikkonen et al. 2011).

Several researchers have identified functional and important roles of ncRNAs in diverse biological processes, such as in the recruitment of chromatin regulatory proteins to genomic DNA locations, or in the organization of distinct nuclear structures. Additionally, it was already determined the role of several ncRNAs in shaping aspects of 3D nuclear organization and on the emerging mechanisms to regulate gene expression(Quinodoz & Guttman 2014).

2. ROADMAP FOR REGULATION

2.1. Epigenome-wide analysis

Evidences suggest that exposure to particular environmental factors like nutrition during early development, may affect susceptibility to certain chronic diseases(Mensaert et al. 2014),(Ozanne & Constância 2007). The stimulus applied at a critical period of development that result in long-term effects on the structure or function of an organism is called programming(Ozanne & Constância 2007). It is known that these kinds of changes in gene expression are maintained in spite of cell division, which means that a mechanism which allows the stable propagation of gene activity-states from one generation of cells to the next is required(Ozanne & Constância 2007). Epigenetic mechanisms are one such possibility, since it is known that epigenetic arrangements are important for gene regulation providing variants to a changing environment(Mensaert et al. 2014),(Bird 2002),(Ozanne & Constância 2007).

Several studies revealed that the analysis and comparison of epigenomes is essential for detecting and understanding the drivers of certain diseases and traits. In order to understand the role of epigenetics in developmental programming, it is necessary to measure the epigenetic marks throughout the genome using robust and sensitive quantification methods(Ozanne & Constância 2007). Once the role of epigenetics in programming is known, it will be possible to understand the correlations between chromatin components and therefore the scientific community will be closer to improve the prevention, detection and therapy of certain chronic diseases like cancer(Jones & Baylin 2007).

However, it is known that epigenetic variants are often located in tissue-specific regulatory regions. For that reason, each cell type must be analyzed in several individuals to assess the effect of genetic variation on personal cell-type specific epigenomes in normal and disease states(Mensaert et al. 2014),(Hirst & Marra 2011),(Romanoski et al. 2015). One example of a relevant epigenetic mark in human disease and in the biological processes is gene silencing which is essential for the life of eukaryotic organisms and can be mediated by DNA methylation and covalent modification of histones(Jones & Baylin 2007).

Although epigenome-wide association studies (EWAS) have already been focusing on characterizing of genome-wide DNA methylation, currently it aims to examine additional epigenetic marks in order to analyze the association between epigenetic variants and disease(Tsai et al. 2012). This approach already identified DMRs for several traits, but many aspects still require careful consideration owing to the unique features of DNA methylation(Tsai et al. 2012).

2.2. Tissue-specific differential methylated regions

Although epigenetic variants can be tissue-specific or shared across tissues, there have been identified more dissimilarities in different tissues from the same individual than in the same type of cell from the same tissue from unrelated individuals(Tsai et al. 2012). For that reason, it is fundamental to link genetic information, which is identical in most of an individual’s cells with epigenetic mechanisms that have tissue-specific roles in order to understand the diversity of human tissues(Schultz et al. 2015).

The determination of tissue-specific differential methylation (T-DMRs), partially methylated domains, allele-specific methylation and transcription and also the presence of non-CpG methylation were already analyzed in all contexts of the major human organ systems(Schultz et al. 2015). Transcription is strongly associated with intragenic DMRs in tissues and it was suggested before that these intragenic methylation differences mark intragenic CpG islands. However, additional data suggests that predicted enhancers and putative promoters only accounted for 23% and 22% of intragenic DMRs, which means that the remaining DMRs represent an unrecognized set of functional elements. It is also known that the methylation level of uiDMRs (undefined intragenic DMRs) is strongly correlated with the expression of the genes containing them(Schultz et al. 2015).

Schultz et al. 2015 examined whether variation in methylation is associated with genetic variation across individuals. In this study, the tissue-specific methylation from DNA motifs was predicted and the motif groups were clustered by their tissue hypo and hypermethylation specificities. Additionally, evidence was found about the existence of methylation outside of the CG contexts, as in CH contexts. This analysis revealed a negative correlation between expression and methylated CH.

The partially methylated domains (PMDs) have not yet been extensively studied but its presence is known to involve several organ systems and it is suggested that they could mark transcriptionally repressive genomic domains. The IMR90 human fetal fibroblast cells and the human SH-SY5Y neuronal cells are an example of cells which have large regions of their genome, 41% and 19% respectively, with PMDs. In (Schroeder et al. 2011), autism candidate genes were also enriched within PMDs and the largest one showed a strong genetic association to autism.

On the other hand, although CpG methylation has been thought to disrupt the interactions between trancription factors and DNA, it was already shown that the transcription factors preferentially bind to methylated CpG sites. Wan, J. *et al.* (2015) characterised T-DMRs and correlated them with the expression levels of associated genes. It was found that genes whose expression was negatively correlated with T-DMRs were enriched for functions carried out in adult tissues, while the positively correlated genes were enriched for negative regulators such as transcriptional repressors (Wan et al. 2015).. Additionally, only 14% of the predicted motifs associated with negative gene regulation contain a CpG site, while 78% of the positive gene regulation motifs contained at least one CpG (Wan et al. 2015). For the positively associated motifs that contain a CpG site, it may be the methylation of that specific CpG, which allows the binding of a particular transcription factor that only binds to methylated DNA and promotes transcription (Wan et al. 2015). On the other hand, on negative T-DMRs, generalized methylation of the T-DMR may be more likely to inhibit transcription by the binding of methyl-binding proteins rather than a specific transcription factor that only binds to methylated DNA(Wan et al. 2015).

2.3. The age-associated epigenome

Remarks about the influence of lifestyle choices in the aging process have led to the search for biological markers involved in the aging process that can be used to provide some insights into age decline and disease(Hannum et al. 2013). The progression of multiple degenerative processes and the progressive loss of regenerative capacity and tissue function within an individual is the key to understand the molecular mechanisms of normal and premature aging(Bell et al. 2012),(Bewerunge-Hudler et al. 2014),(Winnefeld & Lyko 2012).

The factors that contribute to the rate of healthy aging within an individual were already identified by several studies(Bell et al. 2012). It was also shown that stress may affect gene expression patterns through specific changes in DNA methylation. However, spontaneous epigenetic changes may also occur without environmental stress, leading to unpredictable differences in the epigenome between individuals. These ones are caused by chemical agents that disrupt methyl groups or through errors in copying methylations states during replication(Hannum et al. 2013).

Besides the expression of genes especially involved in metabolic and DNA repair pathways, telomere length is also an aging-marker that shows an accelerated rate of decay under environmental stress.(Hannum et al. 2013) Blood pressure, lung function, bone mineral density and serum levels of 5-dehydroepiandrosterone (DHEAS), cholesterol, albumin and creatinine were also considered biomarkers of aging.(Bell et al. 2012)

DNA methylation is the most commonly studied epigenetic modification in humans and has been linked to complex age-associated diseases like metabolic disease, cancer, diabetes and cardiovascular disease(Hannum et al. 2013),(Bewerunge-Hudler et al. 2014). Additionally, it was already observed a phenomenon called *epigenetic drift* where increasing differences in DNA methylation marks were observed in identical twins as a function of age(Hannum et al. 2013).

For these reasons, the scientific community has been focusing the attention on the associations between age and the state of the epigenome even though the rate of change and contribution to biological aging are poorly understood(Tsai et al. 2012),(Bell et al. 2012). The determination of a quantitative measurement of methylome states in order to identify relevant factors and to detect different rates of human aging is essential in order to stablish relations to clinical or environmental variables(Hannum et al. 2013).

Since aging is associated with multifactorial changes that are beginning to be understood(Winnefeld & Lyko 2012), the determination of differential methylated regions during lifetime is an important goal that requires the characterization of methylation patterns in large (Bewerunge-Hudler et al. 2014). Although global changes in DNA methylation may be due to a progressive loss of methylation in repetitive sequences throughout the genome, individual CpG sites that specifically change with age have already been reported (Tsai et al. 2012),(Bewerunge-Hudler et al. 2014). Several epigenome-wide scans have identified age-associated changes in the methylome at some CpG sites and also at non-island *loci*, with a positive and negative correlation between methylation in CpG and non-CpG contexts being found, respectively (Bewerunge-Hudler et al. 2014).

These kind of studies are made by determining the genome localization of age related-DMRs (a-DMRs) and their functional role (Bell et al. 2012). Comparisons have been made between the epigenetic variations and aging-related traits that are essential biomarkers of aging (Bell et al. 2012). Distinct studies showed that the aging rate is influenced by several parameters like gender, body mass index and specific genetic variants. In the case of gender, it is known that the methylome of men appears to age approximately 4% faster than that of women (Hannum et al. 2013).

On the other hand, the majority of a-DMRs lay within genes with aging-related functions. For example, methylation markers have been related to the gene for somatostatin (SST), a key regulator of endocrine and nervous systems, and transcription factor KLF14, an important regulator of obesity and other metabolic traits. These have highlighted an association between aging, longevity, metabolic activity and have been implicated in obesity and metabolism (Hannum et al. 2013).

Additionally, it was already shown that although a-DMRs do not appear to be random events, the majority of observed a-DMRs may either be neutral to measures of biological age at later stages of life, or may relate to yet unknown pathways that correlate with biological aging(Bell et al. 2012). However, the timing of the age-related trigger at each CpG site remains unclear although it is known that DNA methylation plays a key role in development and tissue (Bell et al. 2012).

The methylation levels of DMRs and the expression of the closest genes also showed a negative correlation that was stronger closer to the transcription start site (Schultz et al. 2015). These tend to be associated with epigenetic marks targeting low levels of transcription and gene expression in samples of middle-aged individuals and present in tissues functionally linked to development and aging (Bell et al. 2012). Several results indicate that a proportion of a-DMRs are conserved across tissues in samples of different ages and genders which suggests that there the epigenetic mechanisms represent a potential pathway for mediating healthy aging and age-related traits (Bell et al. 2012).

3. METHODOLOGIES TO STUDY THE METHYLOME

One of the unique contributions of epigenomic data to the study of genomic sciences is its quantitative nature in contrast to the sequence itself, which is discrete (Callinan & Feinberg 2006). The measurement of these epigenetic marks is crucial to identify those that are associated with pathology or to the control of biological processes. The emergence of high-throughput methodologies as microarrays and next-generation sequencing coupled to innovative molecular and computational techniques has brought epigenomic studies to the forefront of current research (Mensaert et al. 2014),(Hirst & Marra 2011).

The research community has been improving these techniques at an exponential rate and has revolutionized molecular biology with genomic studies focused on mRNA abundance and on fields ranging from cancer genome sequencing to systematic dissection of protein structure and function (Friedman & Rando 2015)·(Hirst & Marra 2011). Currently, the NGS technology enables parallel sequencing of millions of DNA fragments in a short time and provides accurate information on the composition of DNA samples, making it the method of choice for genomics and epigenomics (Dijk et al. 2014).

3.1. DNA Methylation Profiling

The profiling of DNA methylation since the recognition of its importance to gene expression in 1975(Holliday & Pugh 1975)·(A.D. Riggs 1975) has evolved a lot, firstly with the development of early non-specific methods and differential gene methylation analysis and then with the appearance of the microarray technology and NGS methods (Harrison & Parle-McDermott 2011).

The earliest breaches were based on the separation of methylated and unmethylated deoxynucleosides using HPLC or TLC, enzymatical incorporation of tritium-labelled methyl groups to unmethylated cytosines and posterior radioactivity measurement, quantification of radiolabeled DNA retained by polyclonal antibodies followed by visualization by electron microscopy or even usage of anti-5mC monoclonal antibody and secondary antibodies labelled with fluorescent isothyocianate (Harrison & Parle-McDermott 2011). Then, the differential gene methylation analysis emerged where methylation-sensitive restriction enzymes are used followed by radiolabeling, TLC, Southern-blot or even methylated-sensitive PCR methods (Harrison & Parle-McDermott 2011). Currently, the categorization of DNA methylation profiling methods can be made into three main methods: restriction enzyme, affinity enrichment and bisulfite conversion-based methods; all of them followed by microarray or next-generation sequencing techniques (Yong et al. 2016).

The restriction enzyme-based methods take advantage of the differential digestion properties of isoschizomers and neoschizomers since it exhibit different sensitivities to DNA methylation state (Yong et al. 2016). The cleavage is made by methylation-sensitive restriction enzymes (MRE) like *Bst*UI, *Hpa*II, *Not*I or *Sma*I that leave the methylated DNA intact, cleaving only the unmethylated one (Yong et al. 2016). On the other hand, the affinity enrichment-based methods use proteins like methyl CpG-binding domains (MBDs) or antibodies with specificity to methylated cytosines to enrich methylated DNA sequences (Yong et al. 2016).

The particular case of bisulfite conversion was described simultaneously by the Shapiro and Hayatsu groups in the early 1970s (Hirst & Marra 2011) and is based on the ability of sodium bisulfite to deamine the unmethylated cytosine residue to uracil in single-stranded DNA which is read as thymidine, whereas 5mC remains non-reactive (Mensaert et al. 2014),(Clark et al. 1994). This process is made through a multistep process with 1) an reversible addition of bisulfite to the 5-6 double bound of cytosine; 2) hydrolytic deamination of the resulting cytosine derivative to give an uracil-bisulfite derivative; 3) the sulphonate group is removed by a subsequent alkali treatment to give uracil (Figure 3)(Mensaert et al. 2014). Then, the sequence is amplified using PCR in which all uracil and thymine residues are amplified as thymine and only 5mCs are amplified as cytosines(Mensaert et al. 2014).

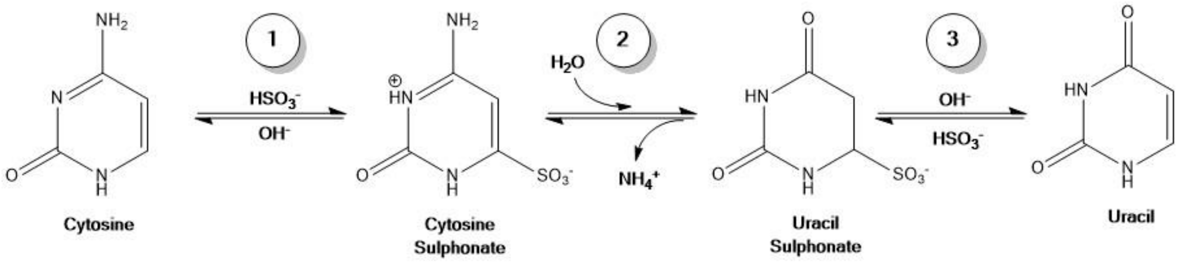


Figure 3 – Bisulfite conversion of cytosine to uracil. The reaction is a multistep procedure where firstly is a sulphonation followed by a hydrolytic deamination and finally an alkali desulphonation.

3.2. Next-Generation Sequencing

Since the conclusion of the Human Genome Project, substantial changes have occurred in NGS methods especially at the whole-genome sequencing scale. These technologies have a major impact on the ability to explain and study genome-wide biological questions since they not only change our sequencing approaches but also accelerated the process.(Mardis 2008) The availability of NGS techniques to study genomic DNA is transforming the biological and medical science in several fields.(Ansorge 2009)

These methodologies enhanced the epigenomics studies in several organisms. The quantification of the expression level and its correlation with changes in environmental factors will intensify the annotation of sequenced genomes while the impact of mutations will become more broadly interpretable across the genome.(Mardis 2008) The research about ancient genomes has also been a difficult task in the genomics research since the characterization of ancient DNAs has been limited by the degraded state of samples. However, the NGS technology made it possible to directly sample the nuclear genomes of the cave bear, mammoth and the Neanderthal and with NGS evolution we expect to increase this sort of evidences.(Mardis 2008) In the field of Microbial Genomics, the goal is to measure the genetic diversity encoded by microbial life in organisms inhabiting a common environment. This research has been supported by the Human Microbiome Project where comparative analysis of the collection of microbes in and on the human body is being made that could contribute to further understanding human health and disease.(Ansorge 2009)

Therefore, the benefits from recent advances of the NGS technology should burst several multidisciplinary fields as epigenomics, genomics, proteomics, microbiology, medical research and anthropology.(Ansorge 2009)

3.2.1. Genomic DNA preparation

The NGS technology requires the conversion of the nucleic acid material to be sequenced into standard libraries suitable for loading onto a sequencing instrument (Figure 4).(Dijk et al. 2014) For that purpose, it is necessary to carry out a library preparation process that can be divided into two distinct steps: the fragmentation of genomic DNA and the preparation of the fragments for sequencing. (Hirst & Marra 2011)All of these processes are dependent of the kind of sequencing that is made.(Hirst & Marra 2011)

The starting material for epigenomic studies is generally double-stranded DNA in the form of isolated genomic DNA or chromatin (ChIP-Seq) which should be fragmented using one of the third available methods (Dijk et al. 2014). First, physical fractionation methods as sonication, apply force to break chromatin. Second, nuclease-susceptibility methods as MNase-Seq (*Micrococcal* Nuclease Sequencing), DNase-Seq (DNase I digestion Sequencing) and ATAC-Seq (Assay for Transposase Accessible Chromatin Sequencing) are based in the susceptibility of certain regions of the genome to enzymatic attack and to separate accessible regions from compact ones. Third, the chemical susceptibility of chromatin could be used such as hydroxyl radical cleavage of DNA backbone or bisulfite treatment to distinguish between cytosine and 5mC (Friedman & Rando 2015).

Then, if the used method requires it, methods of separation or enrichment are used to enrich the sample in specific classes of chromatin and to provide insight into packaging of genomic regions (Friedman & Rando 2015),(Mensaert et al. 2014). It can be made using physical methods as solubility or FAIRE (Formaldehyde-Assisted Isolation of Regulatory Elements)(Friedman & Rando 2015), or immunoprecipitation as ChIP where antibodies are attached to the post-translational modification of histones under study (Hirst & Marra 2011).

Finally, further processing is made according to the correspondent sequencing protocol described in the next sections. In order to obtain homogenous and blunt-end fragments, a previous end-repair is made, together with adapter ligation and usually a size selection step to remove free adapters and to select molecules in the desired size range. It is performed a PCR amplification to generate sufficient quantities of template DNA to allow accurate quantification (Dijk et al. 2014). Although the PCR amplification is known to introduce bias in sample composition due to the fact that not all fragments in the mixture are amplified with the same efficiency, currently several DNA Polymerases or additives minimize amplification bias which is crucial to rise the quality of the library preparation (Dijk et al. 2014). After all these processes the DNA is ready to be sequenced.

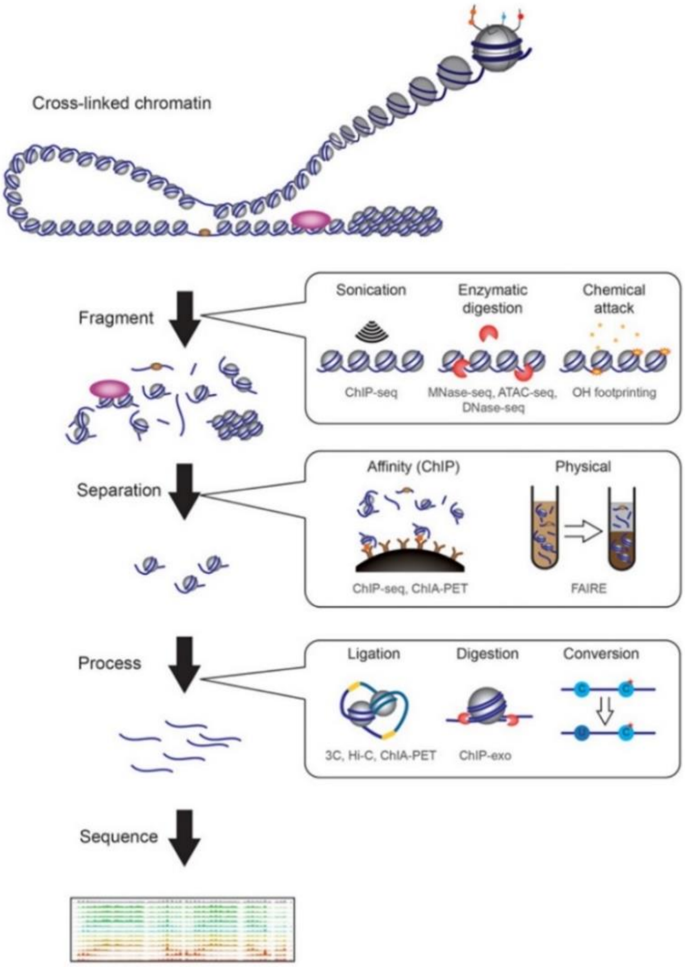


Figure 4 - Library preparation using as starting material a chromatin sample. First, chromatin is fragmented using one of several methods – sonication, enzyme digestion or chemical attack. Second, methods of enrichment or affinity should be used to enrich the sample in specific classes of chromatin and to provide insight into packaging of genomic regions. Third, the DNA is processed according to the correspondent treatment protocol of each study and after end-repair, adapter ligation and PCR the DNA is sequenced. Each sequencing method has a specific and more appropriate protocol according to the aim of the study. Adapted from (Friedman & Rando 2015).

3.2.1.1. Whole Genome Bisulfite Sequencing

Whole Genome Bisulfite sequencing, also called MethylC-Seq or BS-Seq, is useful to determine the methylation status of cytosines at single nucleotide level but at genome scale.

This method was first tested in *Arabidopsis thaliana* and was recently adapted to the human genome. Despite all the advantages of bisulfite conversion, in the original methodology the results of the reaction were amplified by PCR and subjected to Sanger sequencing which does not scale well and cannot be applied to whole genome studies. Currently, with the advent of next generation sequencing it is possible to directly shotgun sequence bisulfite treated genomic DNA (Hirst & Marra 2011).

The library preparation of this particular sequencing method is made using a similar approach to the one explained before. After the fragmentation of genomic DNA, the ends of sheared-DNA are repaired to ensure that each molecule is free of overhangs and contains 5'-phosphate and 3'-hydroxyl groups and to allow the addition of sequencing adapters. Klenow polymerase is used to remove 3'-overhands, T4 DNA polymerase to fill in 5'-overhangs and T4 polynucleotide kinase to phosphorylate 5'-OH(Pomraning et al. 2009). Then, genomic DNA is ligated to sequence adapters, artificial sequences where DNA sequencing will be initiated(Goodwin et al. 2016), where all cytosines are methylated.(Hirst & Marra 2011) Bisulfite conversion only takes place after all these processes and since adapters do not have unmethylated cytosines, they are protected of deamination (Lister & Ecker 2009).

After bisulfite conversion, the library preparation can be made in a directional or non-directional manner.(Krueger et al. 2012) In the first one, adapters are attached to the DNA fragments such that only the original top or bottom strands will be sequenced, resulting in either BSW (Bisulfite Watson) or BSC (Bisulfite Crick) reads that correspond to a bisulfite converted version of either the original strands (Krueger et al. 2012),(Hackenberg et al. 2012). Only one round of PCR is performed, in which the primers should be complemented to adapter sequences producing the final library that can be sequenced(Lister & Ecker 2009). On the other hand, the non-directional libraries consist of two PCR rounds with two different adapter sequences and consequently two primers complemented to them (Lister & Ecker 2009). This results in either BSW and BSC reads plus their reverse complementary strands (BSWRC and BSCRC) which does not preserve strand identity (Hackenberg et al. 2012),(Krueger & Andrews 2011). Therefore, all four DNA strands that arise through bisulfite treatment and amplification can be sequenced with the same frequency(Krueger et al. 2012) (Figure 5).

The paired-end read concept is also important for bisulfite sequencing since a considerable number of wrong mappings can be detected and removed. In this methodology both end of the DNA fragment are sequenced and since the distance between them is known, it is possible to determine a region where both reads must map (Hackenberg et al. 2012). Paired-end reads contain one read from one original strand (BSW or BSC) and one complementary strand. In the case of a directional paired-end library, the first read always come from either the BSW or BSC strand, while in a nondirectional paired-end library, the first read may originate from any of the four possible bisulfite strands (Babraham Bioinformatics 2013).

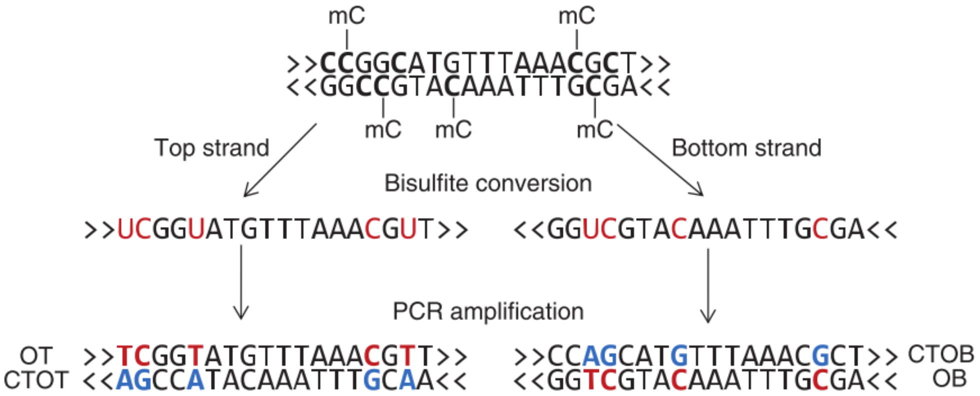


Figure 5 – Representation of a non-directional library. After the DNA genomic fragments are end-repaired and have adapter sequences, the bisulfite conversion is made in both top and bottom strands. Then, at least two rounds of PCR initialized by primers complementary to adapter sequences are performed. In the first round of PCR after bisulfite treatment, all uracil and thymine residues are amplified as thymine and only the methylated cytosine is amplified as cytosine. In the second round of amplification, both the BSWRC (in the figure CTOT – strand complementary to the original top strand) and BSCRC (in the figure CTOB – strand complementary to the original bottom strand) are produced and therefore, all four DNA strands can be sequenced with the same frequency. Adapted from (Krueger et al. 2012).

3.2.1.2. Reduced Representation Bisulfite Sequencing

Reduced Representation Bisulfite Sequencing (RRBS) is a method with prior enrichment for genomic regions of interest and was introduced to reduce sequence redundancy associated to a whole genome sequencing method (Hirst & Marra 2011). In the case of CpG island methylation, the cost of bisulfite sequencing data can be reduced using RRBS that enriches the library on CpG-dense regions (Krueger et al. 2012). The RRBS technique is started by a fragmentation of genomic DNA using a restriction enzyme like BglII or more frequently *MspI* (Laird 2010). In the case of the methylation insensitive MspI restriction enzyme, the phosphodiester bond upstream of the CpG dinucleotide is cleaved in its CCGG recognition element (Lister & Ecker 2009). Often, the library is size-selected using gel-electrophoresis to generate a fragment library within the range of next-generation sequencing platforms, allowing the creation of a reproducible but reduced representation of the DNA methylome (Mensaert et al. 2014)'(Hirst & Marra 2011). The process follows the workflow of WGBS in library preparation being the size selection followed by end-repair, A-tailing, adapter ligation, bisulfite conversion and amplification by PCR with primers complementary to the adapter sequences (Babraham Bioinformatics 2013).

3.2.1.3. Single-Cell Methylome

The currently most used DNA methylation profiling techniques requires large amounts of cells per experiment, making it difficult to study rare cell populations and hetegogeneity among individual cells (Farlik et al. 2015). For these reasons, the development of single-cell epigenome mapping had to depend on very small amount of initial DNA and originated scWGBS, scRRBS and scPBAT (post-bisulfite adapter tagging) that are particularly useful for specific cell types that play important roles in early development such as sperm cells, oocytes, primordial germ cells and embryonic stem cells (Yong et al. 2016).

The scRRBS provides information in one individual mouse or human cell by using multifluidics or emulsion-based single-cell lysates in which the MspI digestion is carried out directly, in order to minimize DNA loss (Yong et al. 2016). On the other hand, scPBAT uses the PBAT protocol where it is made an adapter tagging process after bisulfite conversion treatment, eliminating the need of an amplification step (Yong et al. 2016). As to scWGBS it is the method of choice for analyzing large number of single cells at low sequencing coverage (Farlik et al. 2015).

3.2.1.4. Methylated DNA Immunoprecipitation Sequencing

The Methylated DNA Immunoprecipitation Sequencing (MeDIP-Seq) is an immunoprecipitation technique that aims at enriching the fragmented DNA pool according to its methylation content, by using an antibody specific for methylated cytosines. The technique should be used in a denatured state since the antibodies might be raised against a single-stranded methylated cytosine and the library preparation is made prior to immunoprecipitation step to avoid over representation of high methylated genomic repeats.

3.2.1.5. Methylated DNA Binding Domain Sequencing

The Methylated DNA Binding Domain Sequencing (MBD-Seq) is a method with high similarity to MeDIP-Seq in which bead immobilized high affinity methyl-binding proteins MECP2 or MBD2 are used to enrich methylated DNA fragments from a pool of genomic DNA fragments. In the MBD-Seq, while the weakly methylated DNA fragments are eluted at lower salt concentrations, the densely methylated DNA fragments are eluted at high salt concentrations (Hirst & Marra 2011).

3.2.1.6. Methyl-sensitive restriction sequencing (MRS-Seq)

The Methyl-sensitive Restriction Sequencing (MRS-Seq) also known as Methyl-sensitive Restriction Enzyme Sequencing (MRE-Seq) is a restriction enzyme-based method to profile the unmethylated fraction of the genome using restriction enzymes that are sensible to the

CpG methylation state (Hirst & Marra 2011). After the cleavage of unmethylated target sequenced for enzymes like *BstUI*, *HpaII*, *NotI* and *SmaI*, the resulting DNA fragments are selected by size, it is made a library construction step and then a NGS method is performed (Yong et al. 2016).

The NGS methods are divided between first-generation sequencing platforms, also called basic sequencing methods, mainly based on fluorescence methods, second-generation sequencing that introduced the whole-genome and high-throughput concept and third-generation sequencing, mainly baised on real-time and single-cell detections. Currently, a fourth-generation concept is emerging. This technology enable highly spatially resolved transcriptomics regardless of the specimen by sequencing nucleic acids directly in cells and tissues (Ke et al. 2016).

3.2.2. Basic Sequencing Methods

3.2.2.1. Sanger Method

The Sanger method (Figure 6) was developed by Frederick Sanger in 1977 and was based on chain-termination method also known as Sanger sequencing (Sander et al. 1975)'(Sanger & Nicklen 1977). This method was adopted as the primary technology in laboratory sequencing applications and suffered a gradual improvement yielding capillary-based(Swerdlow et al. 1990), semi-automated implementations of the Sanger biochemistry.

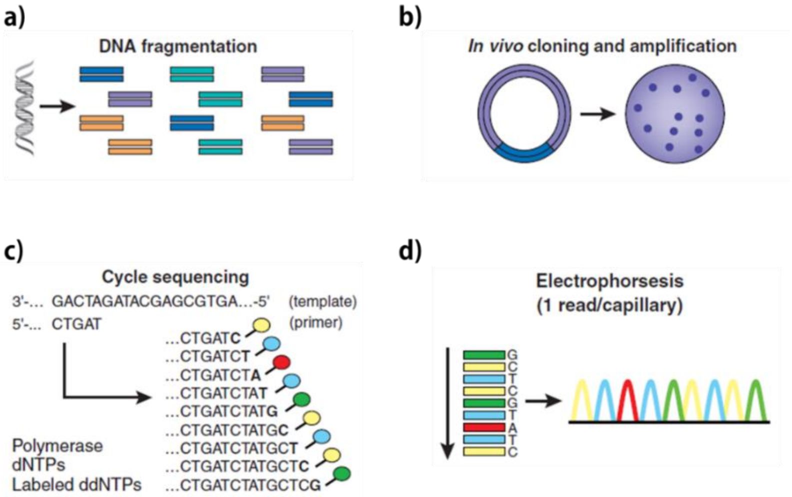


Figure 6 – High-throughput Sanger sequencing. Starting with a fragmentation of DNA (a), followed by an implication in vivo (b) and a cycle sequencing using ddNTPs (c) and finally a capillary-based electrophoresis (d)(Shendure & Ji 2008)

Firstly, DNA is prepared by one of two mechanisms. In the shotgun *de novo* sequencing mechanism, the fragmented DNA is cloned into a plasmid that is used to transform, for example, *Escherichia coli*. On the other hand, in the target resequencing approach, the fragmented DNA is amplified using primers that flank the target. After DNA preparation, the sequencing reaction takes place through several cycles of denaturation, primer annealing and primer extension. Since dideoxynucleotides (ddNTPs) fluorescently labelled are provided in the reaction media, each one being marked with a different fluorophore, the extension reaction is terminated when one of these molecules is incorporated in the sequence. The outcomes of the Sanger sequencing reaction will be DNA fragments with different lengths and a ddNTP in its end. Then, the end-labelled products are separated by size by capillary-based electrophoresis and through laser excitation of fluorescent labels, the DNA sequence can be reconstructed (Shendure & Ji 2008).

3.2.2.2. Maxam Gilbert Method

Allam Maxam and Walter Gilbert developed in 1977 a sequencing method called Maxam Gilbert sequencing(Maxam & Gilbert 1977) which was used in sequence cases which could not easily be resolved with Sanger technique (Ansorge 2009). This technology was based on chemical modification of DNA and subsequent cleavage at specific bases (Liu et al. 2012).

Firstly, the DNA molecules (double or single stranded) are labelled with ^{32}P at one end of one strand. Then, the DNA molecule is broke at guanine, adenine, cytosine and thymine with chemical agents, producing a nested set of radioactive fragments from the labelled end to each of the positions of that base. Finally, the fragments are separated according to its size and analyzed through an autoradiograph of the gel (Maxam & Gilbert 1977).

3.2.3. Second generation sequencing platforms

Although the automated Sanger sequencing has dominated the industry for several years, its limitations showed a need for new and improved technologies for DNA sequencing (Ansorge 2009). The second-generation sequencing platforms avoid the need for cloning of DNA fragments by the determination of the sequence data from amplified single DNA fragments without major increase of sequencing errors in comparison with Sanger sequencing technique. However, these technologies remain expensive for generating sequences with high-throughput and producing very short read lengths which are a challenge to developers of software computer algorithms and for resolving repetitive regions as CpG islands (Ansorge 2009).

The main second-generation sequencing platforms are Roche 454 Pyrosequencer Instrument released in 2005, Illumina Genome Analyser firstly known as Solexa Genome Analyser in 2006, SOLiD Applied Biosystems in 2007 and Life Technologies Ion Torrent in

2010. These technologies include a number of similar methods that can be grouped as template preparation, sequencing and imaging and data analysis. The combination of specific protocols distinguishes one technology from another and determines the data output on each platform (Michael L Metzker 2010). Although all chemistries were studied under the scope of this thesis, only Illumina protocol will be described, since this is the one that iBiMED has.

Table 1 - Comparison of second-generation sequencing techniques regarding sequencing mechanism, year and company of release (Liu et al. 2012) (Goodwin et al. 2016).

Platform	*Sequencing Mechanism*	*Year of release*	*Company releaser*
Roche 454 Pyrosequencer Instrument	Sequencing by synthesis	2005	Life Sciences
Illumina Sequencing Technology	Sequencing by synthesis	2006	Solexa
Life Technologies SOLiD System	Sequencing by ligation	2007	Applied Biosystems
Life Technologies Ion Torrent	Sequencing by synthesis	2010	Life Technologies

The second NGS platforms, also called short-read length NGS platforms(Goodwin et al. 2016), are divided into sequencing-by-synthesis (SBS) or sequencing-by-ligation (SBL) mechanisms (Table 1). In the SBS approach the methods are DNA-Polymerase dependent and its action is reported by a signal, whereas in the SBL techniques imaging depends on an hybridization of a probe to a DNA fragment (Goodwin et al. 2016).

3.2.3.1. Illumina Sequencing Technology

The original Illumina Sequencing Technology used four nucleotides that are reversibly labelled with a different fluorescent dye and added simultaneously to the surface of a flow cell channel along with DNA Polymerase responsible for DNA synthesis (Mardis 2008),(Ansorge 2009).

Firstly, the DNA fragments are ligated to their adapter and, after denaturation, the single-stranded chains are immobilized on a proprietary flow cell surface designed to facilitate access to enzymes while ensuring high stability of surface-bound template and low non-specific binding of fluorescently labelled nucleotides (Mardis 2008),(Illumina 2010). The single stranded fragments will perform bridge amplification where a bridge structure is created through hybridization of the free end to the complementary adapter on the surface of the support (Ansorge 2009). After the addition of the PCR amplification reagents, the DNA Polymerase will produce double-stranded bridges using as primers the adapters on the flow cell surface (Ansorge 2009). The denaturation of double-stranded bridges leaves single-stranded templates anchored to the substrate (Illumina 2010). When the amplification is complete, there are several million of dense clusters in each channel of the flow cell (Figure 7).

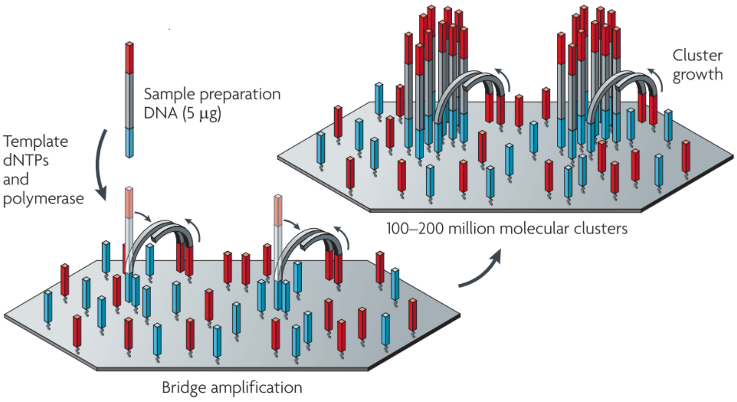


Figure 7 – Illumina immobilization strategy. After the sample preparation in which are obtained single stranded DNA fragments ligated to their adapters, the chains are immobilized in a flow cell surface where they will suffer bridge amplification though the addition of template dNTPs and DNA polymerase . The denaturation of double-stranded bridges leaves the single-stranded templates anchored to the substrate. Adapted from (Michael L Metzker 2010)

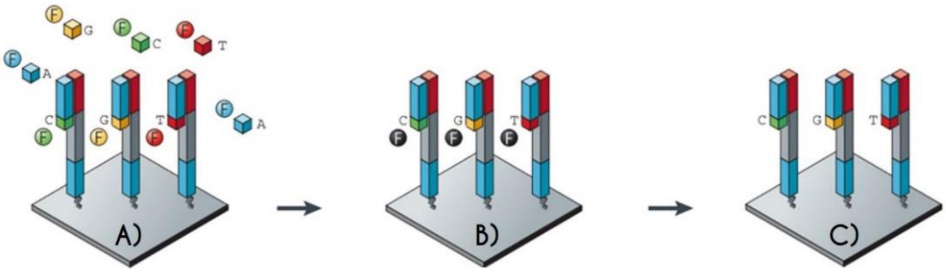


Figure 8 – Sequencing cycle of Illumina technology. (A) After the addition of DNA polymerase, dNTPs each one with a different dye and primers the amplification will start. (B) The addition of a dNTP is followed by an imaging step. (C) The nucleotides are washed and the terminating group of amplification is performed. The process will repeat in cycles. Adapted from (Michael L Metzker 2010).

This technology uses labelled nucleotides to sequence the clusters on the flow cell surface, DNA polymerase and primers (Ansorge 2009)'(Illumina 2010). In the sequencing cycle (Figure 8), a single labelled deoxynucleoside triphosphate (dNTP) is added to the nucleic acid chain (Illumina 2010). This dNTP is a reversible terminator of DNA sequencing and after its incorporation, an imaging step is performed followed by washing of the remain nucleotides and cleavage of the terminating group (Michael L Metzker 2010). This process is called cyclic reversible termination since after the cleavage step, the amplification continues through the following nucleotides (Michael L Metzker 2010). Finally, it is performed an imaging step where the images are subsequently analyzed to generate a focal map for each cluster (Hirst & Marra 2011).

3.2.4. Third-generation sequencing platforms

The Third-generation Sequencing Platforms, also called long-read sequencing methods, consist of single-molecule real time sequencing approaches or synthetic approaches which rely on existing short-read technologies to construct long read *in silico* (Goodwin et al. 2016). These technologies require an extremely sensitive light detection system capable of detecting and identifying signal from single molecules (Ansorge 2009). Particularly interesting for the scope of this thesis are the Pacific Biosciences SMRT system and Nanopore Sequencer, described bellow.

3.2.4.1. Pacific Biosciences SMRT System

This technology was introduced in 2010 by Pacific Biosciences and is called single-molecule real time (SMRT) DNA-sequencing platform (Shokralla et al. 2012). The method of real-time sequencing involves imaging the continuous incorporation of dye-labelled nucleotides during DNA synthesis and uses hairpin library structures (Michael L Metzker 2010).

The method is performed in individual picolitre wells with Zero Mode Waveguide (ZMW) detectors in their bottom along with a stationary DNA Polymerase (Goodwin et al. 2016). The natural capacity of DNA polymerase to incorporate nucleotides is used in this method followed by a fluorescence detection. However, in this case the fluorescent label is attached to the terminal phosphate group rather than in the nucleotide base (Shokralla et al. 2012). Then, the incorporated laser and camera system records the color and duration of emitted light as the nucleotide momentarily pauses during incorporation in the bottom of the ZMW. The polymerase cleaves the dNTP-bound fluorophore during incorporation, allowing it to diffuse away from the sensor area before the next labelled dNTP is incorporated(Goodwin et al. 2016).

This technology has already been reported as a possible sequencing method to detect directly DNA methylation without the use of bisulfite conversion method. Since it is known that SMRT sequencing polymerase synthesis rates are sensitive to DNA primary and secondary structure, the methylated bases in a DNA template might be detected directly on the principle that their presence affects polymerase kinetics during SMRT sequencing (Flusberg et al. 2010).

3.2.4.2. Nanopore sequencer

In the nanopore sequencing the nucleic acids are driven through a nanopore either a biological membrane protein or a synthetic pore and the detection is not made through incorporations of nucleotides during synthesis that cause variations of measurable parameters like light, color or pH (Goodwin et al. 2016). On the other hand, a direct detection of DNA composition is made using a ssDNA or ssRNA native molecules (Goodwin et al. 2016). The translocation of DNA through the nanopore induces fluctuations in DNA conductance through the pore or can cause interactions of individual bases with the pore which can be used to infer the nucleotide sequence (Shendure & Ji 2008). The first consumer prototype of this sequencer is the *MinION* from the Oxford Nanopore Technologies, released in 2014, that uses a hairpin library structure (Goodwin et al. 2016).

This method was also already reported as a detector of DNA methylation without chemical modifications of the strand (Simpson et al. 2017). Hidden Markov models (HMMs) were used to analyse nanopore sequencing data and there were clear differences in the electrical current distributions of methylated and unmethylated DNA (Simpson et al. 2017).

3.2.5. Comparison of several NGS platforms

The continuous emergence of new generation sequencing technologies has been due to the demand of efficient quantification of DNA sequence. For that reason, every method has their own advantages and disadvantages as well as differences in several factors: sequencing mechanism, read lengths, run time, output per run and machine cost (Table 2).

The Sanger sequencing turned automatic after years of improvement by Applied Biosystems and was adopted as the primary technology in the first generation era due to its high efficiency and low radioactivity when compared to Maxam Gilbert sequencing method. Although Sanger sequencing was the main tool for the development of the human genome project, the appearance of second and third next-generation sequencing platforms allowed a massively parallel analysis of genomes, high throughput and reduced cost (Liu et al. 2012).

Since 2005, Life Sciences and Roche have made significant improvements to the 454 Pyrosequencer instrument and currently it can generate about one million reads with about 700 bp read length in a 24 hours run (Liu et al. 2012),(Shokralla et al. 2012),(Escalante et al. 2014). The biggest advantages of this technology are the faster run times (Michael L. Metzker 2010), the large read lengths that improve mapping in repetitive regions (Michael L. Metzker 2010) and its automation possibility in library construction and emulsion PCR (Liu et al. 2012). However, this methodology is not used for epigenomics studies due to its high cost, limited number of reads and reduced reading accuracy (Liu et al. 2012),(Hirst & Marra 2011). The major limitation of 454 Pyrosequencer is related with homopolymers since the technology can’t properly interpret long stretches of the same nucleotide. This results in high error rates caused by base insertions or deletions during base calling (Mardis 2008),(Shendure & Ji 2008). In 2016 the 454 platform, due to its incapability of compete in yield and cost, was discontinued (GenomeWeb n.d.).

Table 2 - Comparison of NGS platforms regarding read lengths, number of reads, run time, output per run, machine cost and cost of service per Gb (adapted from (Liu et al. 2012),(Goodwin et al. 2016),(Escalante et al. 2014),(Liu et al. 2012))

Type	*Platform*	*Read Length (bp)*	*Number of reads*	*Run Time*	*Output per run (Gb)*	*Machine cost*	*Cost of service (per Gb)*
1st NGS	Sanger 3730xl	400-900	N/A	20-180 min	1.9-84 (x10^{-9})	$95 000	N/A
2nd NGS	454 GS FLX Titanium	700	1 M	24 hours	0.7	$500 000	$9500-$12000
	Illumina HiSeqX	150 PE	2.6 – 3 B	<3 days	800-900	$1000	$7
	Illumina MiSeqv3	75 PE	44 – 55 B	21-56 hours	3.3-3.8	$99 000	$142-1000
		300 PE			13.2-15		
	SOLiD 5500xl	50 SE	1.4 B	10 days	160	$251 000	$70
		75 SE			240		
		50 SE			320		
	Ion Torrent PGM318	200 SE	4 – 5.5 M	4 hours	0.6-1	$700 – $1 000	$450-$800
		400SE		7.3 hours	1 - 2		
	Ion Torrent S5 540	200 (SE)	60 – 80 M	2.5 hours	10-15	$65	$300
3rd NGS	HeliScope	30-35	1 B	<1 day	20-28	N/A	N/A
	PacBio RS II	20 000	55 000	4 hours	0.5-1	$695	$1000
	Mk1 MinION	<200 000	>100 000	<48 hours	<1.5	$1000	$750

PE – Paired-end; SE – Single-end; B – Billions; M - Millions

About Illumina Sequencing Technology, the evolution is also remarkable and has dominated the short-read sequencing industry in the last years (Quail et al. 2012) due to its high quality sequences and high throughput range (Escalante et al. 2014). The HiSeqX can generate about 2.6 to 3 billion paired-end (PE) reads with up to 150 bp PE read length in a 3 days maximum run time. Although HiSeqX is the highest-throughput device available, its acquisition is limited for an all-purpose instrument since it can only be efficiently used for Whole Genome Sequencing (WGB) or Whole Genome Bisulfite Sequencing (WGBS) (Goodwin et al. 2016). MiSeq, a limited data throughput sequencer, was designed to clinical applications and small labs and that has special interest in bacterial sequencing(Liu et al. 2012). The main errors of Illumina platforms are related to the underrepresentation of AT and GC-rich regions and substitution errors (Goodwin et al. 2016).

Concerning Life Technologies SOLiD system, its relevance arises from a high accuracy level caused by multiple-time base probing (Goodwin et al. 2016) and a low error rate due to an inherent error correction(Michael L Metzker 2010). The biggest disadvantages of this platform are the long run times (Michael L Metzker 2010), the underrepresentation of AT and GC-rich regions (Michael L Metzker 2010) and probably the short read lengths (Goodwin et al. 2016). SOLiD 5500xl, the most recent update of SOLiD technology, could generate about 1.4 billion reads with up to 75 bp read length in a 10 days run time (Goodwin et al. 2016). However, as well as the Roche 454 Pyrosequencer, in 2016 the manufacture and sale was discontinued and is now only available as a service platform for human whole genome sequencing (Thermo Fisher Scientific n.d.).

Finally, Life Technologies currently commercialize an Ion Torrent Personal Genome Machine (PGM), similar to MiSeq of Illumina at output level (Liu et al. 2012), with three available ion chips (Shokralla et al. 2012). Its popularity in the market is explained by the higher speed, lower cost and smaller instrument size that doesn’t need many technical requirements or maintenance (Escalante et al. 2014). The Ion Torrent PGM 318 can generate 4 to 5.5 million reads with up to 400 bp of read length in 7.3 hours maximum run time (Goodwin et al. 2016). However, as well as the Roche 454 Pyrosequencer, it also has a higher error rate caused by difficulties in homopolymer detection and by insertions and deletions (Goodwin et al. 2016).

It is known that although PCR amplification has revolutionized DNA analysis, it may introduce base sequence errors into the copied DNA strands, disturbing their abundance levels (Ansorge 2009). For that reason, the third generation sequencing methods offer a much simplified library generation process, long-read and real-time detection. Among them, the most widely used instrument is the PacBio RS II, although its limited throughput and high cost place it out of the reach of many small laboratories. The instrument has also high error rates for longer reads (Goodwin et al. 2016).

3.3. Microarray Technologies

As for methylome analysis, this technique was initially used together with a methyl-sensitive digestion of DNA and later with immunoprecipitation and bisulfite-conversion techniques (Table 3)(Hackenberg et al. 2012). The principle of the microarray technique is that methylated and unmethylated fragments of the genome are separated and analyzed using single-stranded DNA probes that are immobilized on a substrate (Harrison & Parle-McDermott 2011). The targeted DNA from the sample is labelled with a fluorophore and hybridized to the array and the intensity of the signal will determine the number of bound molecules (Goodwin et al. 2016).

3.3.1. Microarray-based methylation profiling

The three main categories of DNA microarrays – endonuclease restriction, bisulfite conversion and affinity based analyses – are consistent with the DNA methylation profiling techniques described previously (Huang et al. 2010). The several methods are described in Table 3 and in the following paragraphs.

Table 3 – Distribution of several microarrays used for DNA methylation profiling across years and according to its pre-treatment protocol. Adapted from (Harrison & Parle-McDermott 2011)·(Laird 2010).

Abbreviation	*Method*	*Year*	*Pre-treatment*
DMH	Differential methylation hybridization	1999	Endonuclease digestion
PMAD	Promoter-associated methylated DNA amplification DNA chip	2004	
HELP	*HpaII* tiny fragment enrichment by ligation-mediated PCR	2006	
CHARM	Comprehensive high-throughput arrays for relative methylation	2008	
MeDIP	Methylated-DNA immunoprecipitation	2005	Immunoprecipitation
MeCIP	Methyl-CpG immunoprecipitation	2006	
MIRA	Methylated-CpG island recovery assay	2005	
BiMP	Bisulfite methylation profiling	2008	Bisulfite treatment
Infinium	*Illumina Infinium*	2011	

The endonuclease restriction-based microarrays analyses started with differential methylation hybridization (DMH) and methylated CpG island amplification (MCA) that evolved to methylated CpG island amplification microarray (MCAM), methylation amplification DNA chip (MAD) and promoter-associated methylated DNA amplification DNA-chip assay (PMAD). All of these methods can be differentiated by the type of enzymes used and its implications on the resulting DNA (Huang et al. 2010). In 2006 a most reliable method, updated in 2009, emerged. *Hpa*II tiny fragment enrichment by ligation-mediated PCR assay (HELP) was reported to measure 28-34% methylated CpG islands and to identify T-DMRs (Bibikova & Fan 2010). Currently, comprehensive high-throughput arrays for relative methylation (CHARM) is the most known method in this category of microarrays since in 2008 it was created due to a need of a new platform of original array design strategies and statistical procedures involving genome-weighted averages from larger genomic areas (Harrison & Parle-McDermott 2011). With this method, it has been possible to discover that the highest differences in methylation between cells from colon cancer and its adjacent normal cells were located on CpG islands shores and that DMRs in CpG islands shores have a strong inverse relationship with differential gene expression (Yong et al.

2016). This method uses the McrBC enzyme that cleaves half of the methylated DNA and all the methylated CpG islands. The unmethylated DNA is size-selected and hybridized to DNA similarly processed but no cut with the enzyme, on high density arrays (Yong et al. 2016).

The affinity-based microarray analyses rely in an enrichment of the methylated or unmethylated fraction of the genome.(Huang et al. 2010) Methylated DNA immunoprecipitation (MeDIP-chip), similarly to MeDIP-Seq, uses an anti-methylcytosine antibody to immunoprecipitate DNA with methylated CpG sites.(Yong et al. 2016) The genomic DNA is sheared to produce random fragments, denatured and incubated with the antibody (Bibikova & Fan 2010). This is followed by purification of the enriched fraction of the genome and the immunoprecipitated fraction is hybridized to a microarray (Huang et al. 2010). An alternative approach is methyl-CpG immunoprecipitation (MeCIP) that is similar to MeDIP in terms of techniques but uses a recombinant protein complex with the same properties of the antibody. In methylated CpG island recovery assay (MIRA) another protein complex, MBD3LI bound to MBD2, uses the high-binding affinity of its methyl-binding domain to double-stranded DNA. This binding domain is not sequence specific except for methylated CpG (Huang et al. 2010).

The bisulfite conversion-based microarray analyses are based on a bisulfite treatment followed by a special PCR, as described in the beginning of section 3. In a methylation-specific oligonucleotide assay, the PCR amplicons generated function as probes to hybridize targets corresponding to the methylated or unmethylated regions of the genome (Huang et al. 2010). This hybridization produces a fluorescent signal that is measured and analysed (Huang et al. 2010).

3.3.2. Microarray platforms

The modern microarray platforms are classified into three basic types of arrays: printed arrays or spotted arrays on glass, in situ synthesized oligonucleotide arrays and high-density bead arrays or self-assembled arrays (Huang et al. 2010)·(Bumgarner 2013). This methods can be distinguished based upon characteristics such as the nature of the probe, the solid-surface support used and the specific method used for probe addressing or target detection (Miller & Tang 2009). The emergence of these platforms progressed rapidly as new methods of production and fluorescence detection and the increasing knowledge of multiple genomes provided the raw information necessary to ensure that arrays could be made that represented a large fraction of genes in a genome (Bumgarner 2013).

3.3.2.1. Printed or Spotted Arrays

The printed arrays, or spotted arrays on glass, emerged in 1996 and were made in poly-lysine-coated glass microscope slides that provided a good binding of DNA (Bumgarner 2013). In order to spot multiple glass slide arrays from DNA stored in microtiter dishes, it is used a robotic spotter in this technology (Bumgarner 2013). The spotting process made on glass allows for a fluorescence labeling of the sampling, which brings several advantages in comparison to the radioactive or chemiluminescent labels, such as higher sensitivity, larger dynamic range, lower costs and simplicity (Bumgarner 2013).

Arrays can be divided into double-stranded DNA or oligonucleotide microarrays, depending on the nature of the probes (Miller & Tang 2009). The dsDNA probes are amplification products obtained by PCR, shotgun library clones or cDNA that are denatured and attached to the glass slide surface through an electrostatic interaction between the negative phosphate backbone of DNA and the positive charged coating of glass surface or by UV cross linked covalent bonds between thymine bases of DNA and amine groups on slides.(Miller & Tang 2009) The dsDNA probes are typically 200 to 600 bp long and each one represents a different gene. Although dsDNA probes have a high sensitivity and hybridization strength, they suffer in specificity because they have higher melting temperatures and greater mismatch tolerance.(Miller & Tang 2009) Even though the decreased specificity in the study of a genomic sequence rich in natural polymorphisms can be beneficial, it is disadvantageous when trying to discriminate among highly similar target sequences and unacceptable for clinical diagnostic applications.(Miller & Tang 2009)

On the other hand, the oligonucleotide probes range from 25 to 80 bp to studies out of gene expression field.(Miller & Tang 2009) The probes are attached to the glass slides by covalent linkage and the probes are coupled to the microarray surface by 5’ or 3’ ends on aldehyde or epoxy functional groups provided by coated slides.(Miller & Tang 2009) With shorter length than the dsDNA probes, the oligonucleotide probes introduce fewer errors during probe synthesis and facilitates the interrogation of small genomic regions, including polymorphisms.(Miller & Tang 2009) However, they need comparable melting temperatures and lack palindromic regions, which forces more a careful design.(Miller & Tang 2009)

In general, the printed microarrays are distinguished by their simplicity, cost accessibility and flexibility, being useful in study organisms that are not fully sequenced.(Miller & Tang 2009) However, their use in clinical diagnostics is limited to specific research applications since they need complex monitoring tasks to ensure reproducibility and quality of data.(Miller & Tang 2009)

3.3.2.2. In-situ Synthesized Oligonucleotide Arrays

The in-situ synthesized oligonucleotide arrays were introduced in 1991 and later optimized by several public commercial microarrays like Affymetrix GeneChips, Roche NimbleGen or Agilent.(Miller & Tang 2009) In all of the methods, the oligonucleotide probes are synthesized directly on the surface of the microarray and multiple probes per target are included to improve sensitivity, specificity and statistical accuracy.(Miller & Tang 2009) The probes are grouped in sets that include one perfect-match probe and one mismatch probe with a single-nucleotide difference in the middle of the probe, allowing the identification of possible nonspecific cross-hybridization events.(Miller & Tang 2009)

The Affymetrix GeneChips technology, with typically more than 10^6 features(Miller & Tang 2009), uses probes that are synthesized using semiconductor-based photochemical method in which the nucleotides are protected by light-sensitive protecting groups and the microarray surface is chemically protected from nucleotide addition until deprotected by light.(Bumgarner 2013) When the array surface is exposed to UV light, the nucleotides are deprotected and can be added to the growing oligonucleotide chain.(Miller & Tang 2009) The photolithographic masks are used to determine the specific nucleotides to probe sites because each mask has a defined pattern of windows that act as a filter that block or transmit light. This feature provides a pattern of windows in each mask that directs the order of nucleotide addition. (Miller & Tang 2009)

On the other hand, Agilent technologies, are able to go up to 244 000 features(Huang et al. 2010), acquired a method developed in 1996 that uses inkjet printing technology, standard oligonucleotide synthesis chemistry and longer oligonucleotide probes.(Bumgarner 2013) The glass slide of this technology is adapted to contain hydrophilic regions surrounded by hydrophobic regions that will receive the inkjet printer heads with the four different nucleotides phosphoramidites.(Bumgarner 2013) The presence of both hydrophilic and hydrophobic regions will provide a surface to which the phosphoramidites will couple and where the droplets emitted by the inkjets will be, respectively.(Bumgarner 2013)

The Roche NimbleGen technology, that can contain up to 2.1 million features per slide(Huang et al. 2010), is similar the Affymetrix GeneChip platform but in this case the photolithographic masks are replaced by virtual or digital masks in a maskless synthesizer technology.(Miller & Tang 2009) This technology generates probes that use an array of programmable micromirrors to create digital masks that reflect the desired pattern of UV light to deprotect the features where the next nucleotide will be coupled.(Miller & Tang 2009)

In both cases of NimbleGen and Agilent platforms, the hybridization are multicolour and use longer oligonucleotide probes (≈ 60 bp) while Affymetrix is limited to one label and shorter probes (20-25 bp) are used.(Huang et al. 2010),(Bumgarner 2013)

One of the biggest challenges of in-situ synthesized oligonucleotide microarrays is related to the complex nature of its chemical synthesis and expenses involved in production, which turns the synthesized microarrays not conducive to user-defined development.(Miller & Tang 2009) For these reasons, the usage of this method relies on its customization to the specific study of interest.(Miller & Tang 2009) Although this was the biggest disadvantage of Affymetrix, in 2002 the technology was improved with a micro-mirror system coupled to the photo-deprotection step to direct light at the pixels on the array, which allows a lower manufacturing cost of custom arrays.(Bumgarner 2013) On the other hand, the Agilent and Roche NimbleGen technologies can be easily customized with unique oligonucleotide sequence content.(Miller & Tang 2009) These systems are advantageous because of their reproducibility, standardization of reagents, instrumentation, data analysis and improved the accuracy and reproducibility of data through time, due to their ability of standardize probe concentrations and hybridization temperatures while controlling the nonspecific hybridization.(Miller & Tang 2009)

3.3.2.3. High-density Bead or Self-assembled Arrays

The third platform type was created in 2000 and lately adopted by Illumina.(Bumgarner 2013) In this case, instead of glass slides or silicon wafers as substrate, 3 μm silica beads are assembled to one of two available subtracts (SAM – Sentrix Array Matrix or Sentrix BeadChip) with a fiber-optic composition.(Miller & Tang 2009) (Figure 9) However, BeadChips are more appropriate for very-high density applications like whole-genome genotyping, which require up to 10^5 to 10^6 features.(Miller & Tang 2009) Since each manufactured microarray will not be identical, the BeadArrays have the built-in redundancy advantage, a crucial experimental control for intermicroarray comparative data. Additionally, altering the bead pattern helps identifying spatial biases.(Miller & Tang 2009)

In early versions of these arrays, the beads were encoded with different fluorophore combinations that were used for the decoding process.(Bumgarner 2013) However, this method limited the total number of unique beads that could be distinguished.(Bumgarner 2013) Currently, each bead is covered with several copies of a specific oligonucleotide that capture specific sequences and these beads are deposited in the end of the fiber-optic array in which the ends of the fibers were etched to provide a well that is slightly larger than one bead.(Bumgarner 2013) Unlike the known locations of printed and in-situ hybridized microarray features, the beads in BeadArrays randomly assort to their final location on the array.(Miller & Tang 2009) Since the specific oligonucleotide attached to each bead is unique, the bead location is decoded through the identification of this sequence.(Miller & Tang 2009) Then, the mapping of Illumina beads is made by a series of hybridization and washing steps, allowing fluorescently labelled complementary oligonucleotides to bind to their specific bead sequence and track the location of the bead type.(Miller & Tang 2009) Later versions of BeadArrays, used a pitted glass surface to contain the beads instead of fiber-optic arrays.(Bumgarner 2013)

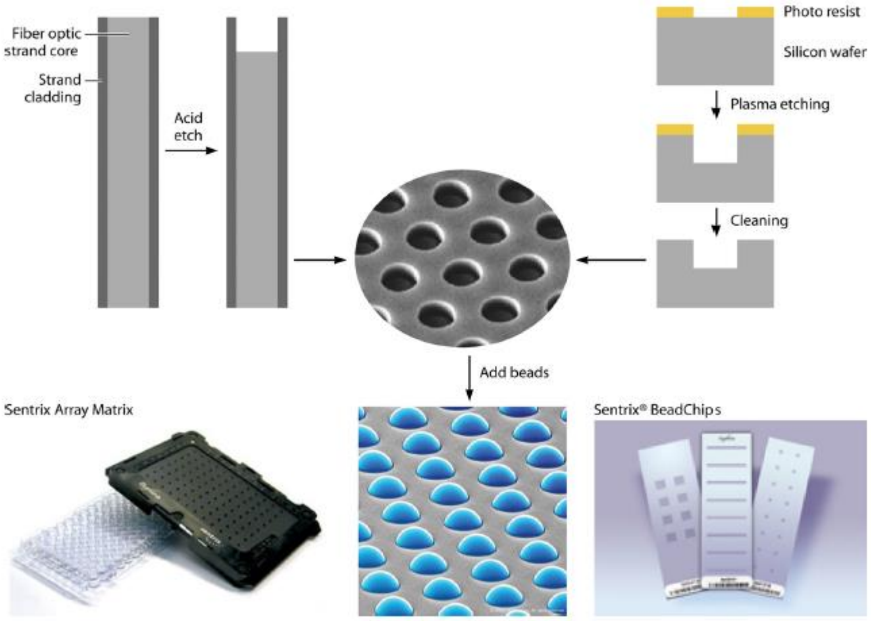


Figure 9 – Structure of Illumina BeadArray. The Sentrix Array Matrix has several fiber-optic bundles with a microwell for a single bead with a specific oligonucleotide attached to it. After several hybridization and washing steps, the fluorescence is determined.

The particular case of Illumina Methylation 450k technology is an example of bisulfite-conversion microarray(Huang et al. 2010) and BeadChip technology that allows to assess the methylation status of 485 577 cytosines, specifically 482 421 CpG sites, 3091 non-CpG sites and 65 random SNPs(Bibikova et al. 2011); with a 99% coverage of RefSeq genes and all the differential epigenetically important genomic regions such as CpG island, island shore and shelf, 5' and 3' UTRs and promoter, gene body and intergenic regions.(Dedeurwaerder et al. 2011) This technology emerged in 2011 as an extension of Illumina Methylation 27k technology(Dedeurwaerder et al. 2011), however whereas Illumina 27k only includes one type of assay currently known as Infinium I or Type I, the 450k technology includes two different probes, with about 50 bases long, referred to as Infinium I and Infinium II or Type I and Type II, which differ at the end-nucleotide that matches the cytosine position of a CpG.(Bibikova et al. 2011) In both cases, it is made a single-base extension step using fluorescent-labeled nucleotides which originates the signal.(Chen et al. 2013)

The Type I methylation-specific assay design, used in 28% of the cases, uses methylated or unmethylated paired probes, located on two different bead types, that measure the methylated and unmethylated DNA, respectively.(Dedeurwaerder et al. 2011) The 3' terminus of the probe match either the protected cytosine or the thymine base resulting from bisulfite-conversion.(Bibikova et al. 2011) Since both bead types will incorporate the same labeled nucleotide that precedes the interrogated cytosine in the CpG locus, the signal will be detected in the same color channel, using either red or green signal (Figure 10).(Bibikova et al. 2011)

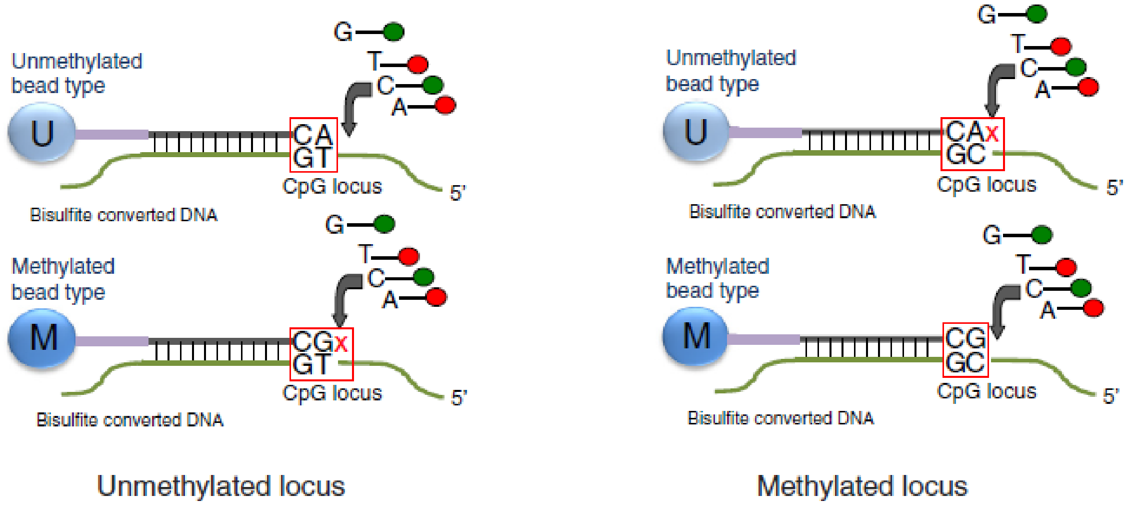


Figure 10 – Infinium I Methylation Assay scheme. Uses two bead types that correspond to the methylated and unmethylated state of the site. The nucleotide incorporated is the same for both locus and the detection is made using the same color channel.

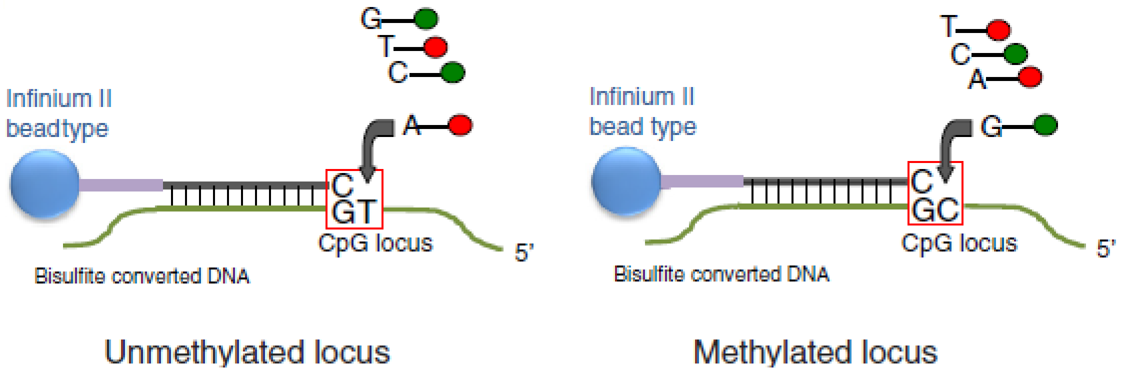


Figure 11 – Infinium II Methylation Assay scheme. Uses one bead typesto both methylated and unmethylated state of the site and the methylation state is detected by single-base extension and the detection is made using two color channels.

On the other hand, the Type II methylation-specific assay design, used on 72% of the cases, uses an unique probe that complements the 3' terminal last base of the bisulfite-converted DNA and that after a single-base extension result in the addiction of a guanine complementary to a methylated cytosine which results in a green signal or an adenine complementary to a thymine which results in a red signal (Figure 11).(Bibikova et al. 2011)

Although the Infinium II type of probe is used to determine more methylated cytosines in the 450k technology, it is known that this assay is less accurate, reproductible and sensitive than the Infinium I.(Dedeurwaerder et al. 2011) There are several differences between the data produced by both techniques: a difference between the β-values and an average probe-variance between replicates.(Dedeurwaerder et al. 2011) This means that Infinium I and Infinium II data cannot be comparable before a complex downstream bioinformatics analysis.(Dedeurwaerder et al. 2011) However, with this problem solved, the technology turns into one of the most attractive technologies to study the human methylome.(Dedeurwaerder et al. 2011)

Illumina Infinium HumanMethylation450 BeadChip is a user-friendly DNA methylation microarray that has reached a predominant place in the market and the scientific arena.(Moran et al. 2016) This methodology has already been used at The Cancer Genome Atlas (TCGA) and for projects focusing on the aging process or interindividual variability.(Moran et al. 2016) Additionally, the versatibility of this technique has also been shown by its capacity to determine 5mC DNA patterns from formalin-fixed paraffin-embedded samples and for the 5hmC mark.(Moran et al. 2016) Currently, Illumina Infinium MethylationEPIC BeadChip is the standard method used for methylation variations in enhancers with a good ovelap with the 450k DNA methylation data.

3.4. Comparison between profiling methods

The comparison between the several available DNA methylation profiling techniques is essential to choose correctly the most appropriate method according to the research goals, available amount of samples, available bioinformatics tools and desired coverage and resolution.(Laird 2010) The DNA microarray technology provides cheap and accessible insights into the DNA methylation status of a sample or a large number of samples and initially was the leading platform to profile the DNA methylation status.(Harrison & Parle-McDermott 2011) However, the appearance of NGS methods allowed allele-specific DNA methylation analysis, can cover more of the genome with less input DNA and avoids hybridization artefacts although it is subjected to sequence library preparation and does not need an appropriately designed microarray.(Laird 2010)

Inside the NGS methods, WGBS is the golden standard method for genome-wide DNA methylation and hydroxymethylation analysis due to its capacity to capture information in all cytosine positions, to profile the methylation state across all the genome and to be accurate and reproducible. One of the biggest disadvantages of WGBS is the facility of failure of bisulfite conversion which can lead to an incomplete conversion of DNA easy to reach due to the restricted laboratory conditions that this method needs. The appearance of RRBS was a big advance compared to the bisulfite conversion methods since it allows the study of targeted regions with high density of CpG sites while the WGBS, due to the number of sequences that yield no relevant information, its complexity and high costs, would not be efficient in this kind of study. However, WGBS is still the standard profiling method for studies interested in regions outside of CpG islands such as in major epigenome consortiums like NIH Roadmap, ENCODE, Blueprint and IHEC.(Yong et al. 2016)

Additionally, MeDIP-Seq and MBD-Seq also revealed advantageous in the DNA methylation estimation for single CpG resolution but its lack of adaptability to low CpG content regions is a big disadvantage. Although the MeDIP-Seq has also a low resolution due to the limited size of fragments from immunoprecipitation and the efficiency of its affinity purification assay can be affected by CpG density and GC content of samples, this method is unique in its application to the study of 5-formylcytosines and 5-carboxylcytosines, has a low cost per CpG and is more tolerant to DNA impurity and integrity. Similarity to the enrichment based methods, the MRE-Seq is also limited in the coverage and resolution due to the sequence type specificity of enzymes. Currently, it is already quite common to combine individual methods to increase coverage or efficiency. A combination of MeDIP-Seq and MRE-Seq can be made to provide the appropriate balance among genomic CpG coverage, resolution, quantitative accuracy, and cost. (Li et al. 2015) In 2016, it was also shown that coupling MeDIP-Seq with bisulfite treatment in a process called methylated DNA immunoprecipitation Bisulfite Sequencing (MeDIP-BS) remarkably improves cost-effectiveness but also enhances analysis resolution when compared to WGBS, Targeting-BS, RRBS and MeDIP-Seq. In this case, the immunoprecipitation step is made before the bisulfite conversion that is followed by library preparation and NGS.(Jeong et al. 2016)

The development of Illumina Infinium Methylation Assay allowed user-friendly feature to the measurement of DNA methylation, cost-effective experiments and with low amounts of input DNA, which is a big advantage for small institutes.(Yong et al. 2016) Despite the coverage of the method being highly dependent on the array design, the Illumina Infinium Methylation Array 450k already has a big CpG coverage that further evolved to the 850k in the last years.(Yong et al. 2016) For all of these reasons, we can conclude that the Illumina Infinium Methylation Array is a great alternative to NGS to study the DNA methylation in humans.(Yong et al. 2016)

3. BIOINFORMATICS

4.1. Bioinformatics in NGS-based methods for methylome analysis

Most NGS technology produces tens of millions of short reads in a single run(Li et al. 2008) which makes the analysis of this data a significant challenge to the genomics' research.(Hirst & Marra 2011), requiring highly efficient and accurate algorithms.(Li et al. 2008) The analysis of genomic data is made using several Bioinformatics' Tools optimized for certain types of information and can be broken down into four steps (Figure 12) starting with the NGS output data as strings of base pairs(Hirst & Marra 2011) or color space base transitions(Hirst & Marra 2011).

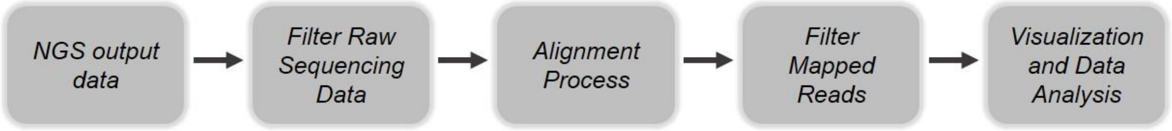


Figure 12 – Typical procedure of a bioinformatics NGS protocol. Begins with NGS output data and ends in visualization and data analysis. (Hirst & Marra 2011)

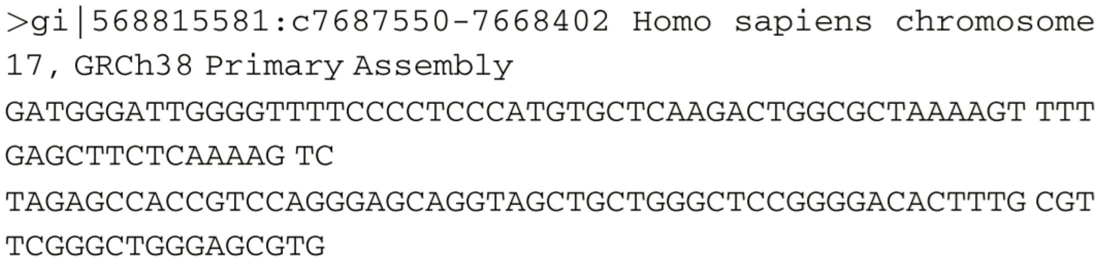
Firstly, a filtration is made where the NGS output data (fastq file(Hirst & Marra 2011)) is scanned according to the reads quality score, length and ambiguity level. Then, the alignment process is developed where the output filtered data is aligned to a reference genome in order to generate a data set consisting of the genomic coordinates of the aligned reads to the reference genome. The obtained files are on the SAM (Sequence Alignment Map) or BAM (Binary Alignment Map) file format(Hirst & Marra 2011).

Since mutations or sequencing errors may lead to read mapping to the wrong location or mapping of reads equally well in multiple positions, it is necessary to filter the mapped reads according to mapping quality, sequence identity and insert size. Then, it is made an extraction process where the filtered data is submitted to a variant calling according to the genomic variant under study. Finally, the filtered data may be viewed directly by converting the read alignments into read density maps and displaying the result on a genome browser.(Hirst & Marra 2011)

4.1.1. NGS Output Data

To obtain genomic position data it is necessary to adopt several procedures that manipulate the raw reads and uncover the genomic structures and variations of interest.(Zhang 2016) These processes might produce files from few gigabytes to terabytes in size that need to be efficiently stored, parsed and analysed(Pavlopoulos et al. 2013). For this reason, several file formats have been developed during the last years.(Zhang 2016)

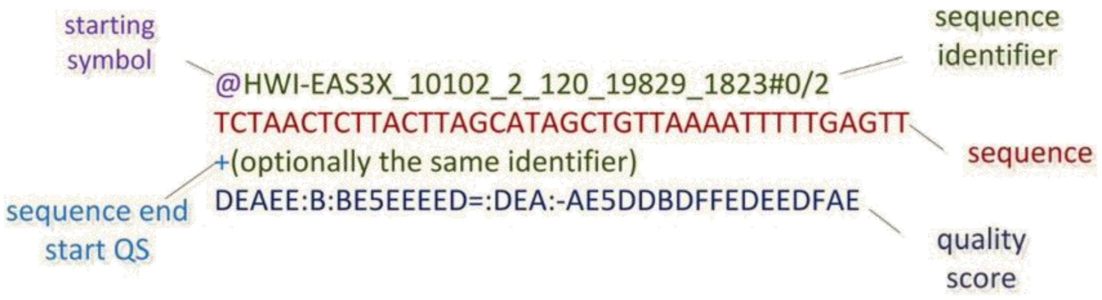
The FASTA sequence file format (commonly ".fa" or ".fasta") was originally invented by Bill Pearson(Cock et al. 2009) and has been the standard format for nucleotide sequence since the first generation sequencing.(Zhang 2016) This file format is a text-based format where the sequencing data represented by a single letter code is preceded by a title line that begins with a ">" symbol followed by a summary description of the sequence containing its accession number, organism designation and sequence location.(Zhang 2016) (Figure 13)



```
>gi|568815581:c7687550-7668402 Homo sapiens chromosome
17, GRCh38 Primary Assembly
GATGGGATTGGGGTTTTCCCCTCCCATGTGCTCAAGACTGGCGCTAAAAGT TTT
GAGCTTCTCAAAAG TC
TAGAGCCACCGTCCAGGGAGCAGGTAGCTGCTGGGCTCCGGGGACACTTTG CGT
TCGGGCTGGGAGCGTG
```

Figure 13 – Representation of a FASTA sequence file format. In the first line it's a summary description of the sequence, beginning with its accession number and preceded by an optional additional information section and after that the sequencing data is present. (Adapted from (Zhang 2016))

Furthermore, FASTQ file format (commonly ".fq" or ".fastq") was originally invented by Jim Mullikin at the Wellcome Trust Sanger Institute. Its simplicity, interchangeable file format and ability to store a numeric quality score (PHRED) associated with each nucleotide in a sequence, makes the FASTQ the most common file format used in NGS.(Cock et al. 2009) This file is also a text-based format where each sequence is defined by four lines of text: 1) a header line with a sequence identifier and optional additional information with no length limit that starts with a "@" symbol; 2) the whole sequence nucleotides in IUPAC nomenclature (A,T,G,C and N for unknown) and uppercase letters and without spaces or tabs; 3) a finisher line with a "+" symbol that represents the end of the sequence and that can be followed by a full repeat of the header line; 4) a quality line that must contain the same number of symbols that letters in the sequence based on ASCII printable representation.(Zhang 2016)-(Cock et al. 2009) (Figure 14)



```
@HWI-EAS3X_10102_2_120_19829_1823#0/2
TCTAACTCTTACTTAGCATAGCTGTTAAAATTTTTGAGTT
+(optionally the same identifier)
DEAEE:B:BE5EEEED=:DEA:-AE5DDBDFFEDEEDFAE
```

Figure 14 – Representation of an Illumina FASTQ file. The file starts with a “@” symbol before a sequence identification code. In the next line it is presented the sequencing data that is terminated in the third line with the “+” symbol and an optional additional sequence identifier reference. In the fourth and last line the symbols follow the ASCII printable representation and present the quality score of the sequencing data. (Adapted from (Pavlopoulos et al. 2013))

FASTQ file format has several variants that depend on the NGS technology used and that affects the structure and format of the quality line, but all of them are based in the PHRED quality score (Q_{PHRED}), defined in terms of estimated probability of error (P_e), where the higher the Q_{PHRED} is, the more reliable the base is.(Zhang 2016),(Cock et al. 2009) The quality score format varies according to the used type of FASTQ file (Table 4) but their quality values can be converted between them using specific formulas available on the literature.(Cock et al. 2009)

Table 4 - Representation of the differences between the several variants of FASTQ – quality score range and type and ASCII characters range and correspondent string. (Pavlopoulos et al. 2013)(Cock et al. 2009)

<table>
<tr><th rowspan="2"><i>FASTQ format</i></th><th colspan="2"><i>Quality score</i></th><th colspan="2"><i>ASCII characters</i></th></tr>
<tr><th><i>Range</i></th><th><i>Type</i></th><th><i>Range</i></th><th><i>String</i></th></tr>
<tr><td>Sanger</td><td>0-93</td><td>PHRED</td><td>33-126</td><td>!"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrstuvwxyz{|}~</td></tr>
<tr><td>Illumina 1.0</td><td>-5-62</td><td>Solexa</td><td>59-126</td><td>;<=>?@ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrstuvwxyz{|}~</td></tr>
<tr><td>Illumina 1.3+</td><td>0-62</td><td>PHRED</td><td>64-126</td><td>@ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrstuvwxyz{|}~</td></tr>
</table>

4.1.2. Pre-processing of the reads

The pre-processing of the reads is a multistep process that aims to make a thorough assessment of the raw sequence data.(Krueger et al. 2012) Although the risk of losing some valid information exists, the increased confidence level of alignments and methylation calls resulting from this pre-processing step overcome this problem.(Krueger et al. 2012) Firstly the data is filtered for the elimination or manipulation of low quality reads and then preparation for the alignment step is made through quality and adapter trimming of the reads.(Krueger & Andrews 2012),(Hackenberg et al. 2012) Since read with errors are infrequent and random, low-frequency reads are candidates for error correction algorithms.(El-Metwally et al. 2013)

4.1.3. Genome Assembly

The genome assembly process is the natural step after a pre-processing of NGS reads and consists of two different approaches: the comparative approach and the *de novo* approach.(El-Metwally et al. 2013) While the first one aligns the NGS reads to a reference genome of the same organism or a similar one, the second reconstructs genomes that are not similar to any available.(El-Metwally et al. 2013) The choice of an approach is based on the intended biological application, cost and time available(Michael L Metzker 2010).

Genome assembly has a lot of associated challenges that makes the process really complex.(Escalante et al. 2014) The increased number of reads to be assembled that increases the complexity in placing each read in the correct position is one of the first obstacles.(El-Metwally et al. 2013) The genome sequences can also be repetitive and since the sequencing errors may impact in this, it is necessary to know how to distinguish both.(Escalante et al. 2014) The rate of overlapping between reads, that happens when there is a sequence match between reads that is long enough to be reliably distinguished from a random event, has also consequences in the accuracy of the assembly(Escalante et al. 2014). Lastly, coverage and read length increase confidence levels of the assembly process but even with high coverage, overcoming the problem of repeats and derived assembly gaps sometimes needs to be spanned by paired-reads sequencing.(Escalante et al. 2014) These facts make genome assembly a process that requires major computational capacities and a lot of software evolution.(Escalante et al. 2014)

In the particular case of bisulfite-sequencing, since all cytosines with the exception of 5mC become converted into thymines, it is not known whether a thymine base call is actually a thymine or a bisulfite converted cytosine.(Pomraning et al. 2009) This decreases the complexity of the sequence as the concentration of methylated cytosines increases, challenging the alignment process.(Hackenberg et al. 2012) Consequently, the NGS reads cannot be aligned directly to the reference genome otherwise the bisulfite converted cytosines would be mismatched and this would make the alignments less specific and the CpG highly methylated dense regions unsequenced.(Hackenberg et al. 2012)

For these reasons, it is necessary to adopt a special alignment method for bisulfite-sequencing data (Figure 15): 1) the wild card aligner that replaces all the Cs in the reference genome by the wild-card letter Y or modify the scoring matrix in such a way that mismatches are not penalized;(Bock 2012) 2) the three letter aligner which use two difference reference genomes that substitute all Cs by Ts or all Gs by As.(Hackenberg et al. 2012) While the wild-card aligner can achieve a higher genomic coverage, it is a slower method with an increased risk of introducing bias towards higher methylation levels, the three letter aligner has a large percentage of reads discarded due to ambiguous alignment positions.(Hackenberg et al. 2012)'(Bock 2012)

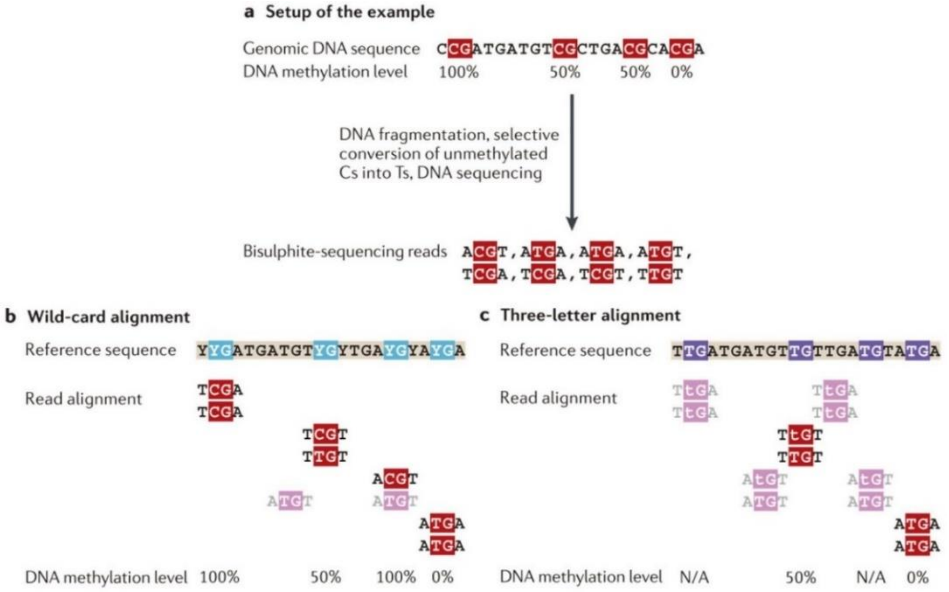


Figure15 – Alignment of several bisulfite-sequencing reads. a) Representation of eight reads obtained from a bisulfite sequencing method originated by a genomic DNA sequence with known DNA methylation content at four CpG sites. b) Wild-card alignment approach where each C in the reference sequence is replaced by a wild-card letter Y and with a consequent increasing of DNA methylation level. c) Three letter alignment approach where each C is replaced by a T and with a consequent reduced sequence complexity that clearly affects the alignment. In both methods the reads with more than one perfect alignment are discarded (represented in grey). Adapted from (Bock 2012)

As before, the output file of the genome assembly process has a SAM format that when compressed turns into a binary representation of SAM, named BAM, and both serve as inputs for various downstream analysis such as feature counts and variant calling.(Zhang 2016) The BAM files are used in a BGZF format and hold the same information as SAM but with a higher store efficiency and lower intensive data processing.(Pavlopoulos et al. 2013) The BAM/SAM file format support both short and long reads produced by sequencing platforms and are tab-delimited text files with a header and an alignment section.(Pavlopoulos et al. 2013) Like the other file formats explained before, the header section contains generic information divided according to its specifications in four types - @HD, @SQ, @RG and @PG - delimited by tabs. On the other hand, the alignment section contains sequences with genomic position and other descriptive information. Each sequence is present in one line text and each-line consists of at least eleven mandatory tab-delimited text fields.(Zhang 2016) (Figure 16)

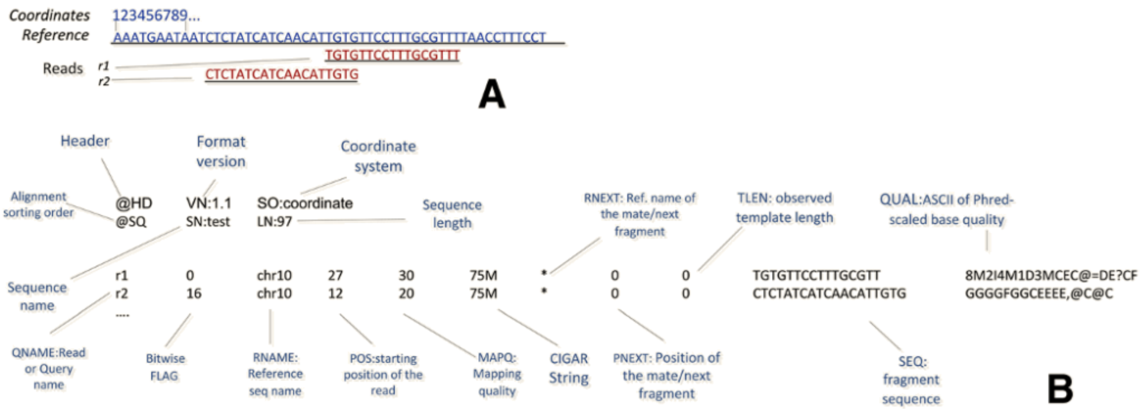


Figure 16 – Representation of a SAM/BAM file. a) Alignment of two reads to a reference genome; b) SAM/BAM files correspondent to the representation in a) with both the header section with its specifications and the alignment section with its tab-delimited text fields. Adapted from (Pavlopoulos et al. 2013)

4.1.4. Post-processing of the reads

After the genome assembly step is complete, the post-processing of the reads occurs and a profiling of the methylation states is made from the alignments where the absolute DNA methylation levels are inferred from the frequency of cytosines and thymines that align to each cytosine in the genomic DNA sequence.(Bock 2012) In order to do so, if the alignment uses the three-letter approach, the reference genome would need to be converted to a four-letter sequence. Then, a thymine that turns into a cytosine after the four-letter conversion and makes a C/T mismatch indicates an unmethylated cytosine while a cytosine in both the read and the reference genome indicates methylation. The same happens to the reverse complement reads.(Hackenberg et al. 2012) (Figure 17)

The methylation level is designated between 1 and 0 for completely methylated and completely unmethylated, respectively. This concept is determined by a relation between

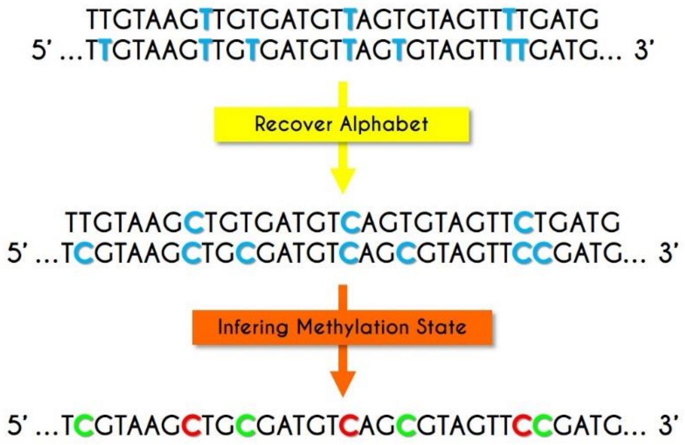

TTGTAAGTTGTGATGTTAGTGTAGTTTTGATG
5' ...TTGTAAGTTGTGATGTTAGTGTAGTTTTGATG... 3'
Recover Alphabet
TTGTAAGCTGTGATGTCAGTGTAGTTCTGATG
5' ...TCGTAAGCTGCGATGTCAGCGTAGTTCCGATG... 3'
Infering Methylation State
5' ...TCGTAAGCTGCGATGTCAGCGTAGTTCCGATG... 3'

Figure 17 – Representation of the inference of methylation state. The reads and the reference genome are converted again into the four-letter sequence and this allows the inference of methylation state. The C/T mismatches indicate an unmethylated cytosine (green) while a cytosine in both the read and the reference genome indicated methylation (red). Adapted from (Hackenberg et al. 2012)

the total number of reads and the number of methylcytosines that map to a certain position.(Hackenberg et al. 2012) However, certain technologies can improve the accuracy of methylation calls by local realignment, analysis of sequence quality scores and statistical modelling of allele distributions.(Bock 2012)

4.1.5. Quality Control

The quality control of mapped data is essential to guarantee high-quality information about DNA methylation. Therefore, it is essential to determine common sources of errors and possible solutions for that, as well as to report the quality of data during the processing (Figure 18).

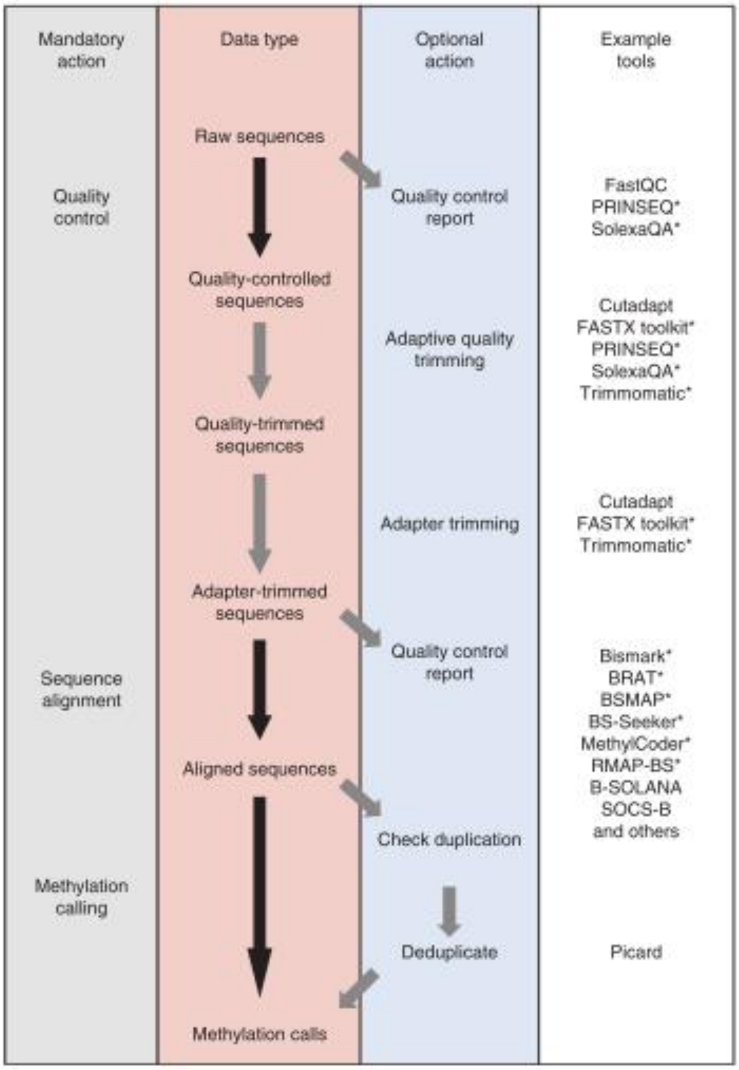


Figure 18 - Representation of the recommended workflow to analyse bisulfite-sequencing data. Each quality control step presented involves the use of several appropriate bioinformatics tools. The grey narrows represent optional steps and the black ones mandatory steps. Adapted from (Krueger et al. 2012)

Firstly, it is essential to guarantee that the methylation calling is made correctly by measuring the base quality in PHRED.(Hackenberg et al. 2012) This parameter is often visualized through a PHRED score versus cycle plot and it is common to observe that the quality score tend to decrease at the end of the reads.(Guo et al. 2013) To avoid incorrect methylation calls it is recommended to only use the good quality portion of the data, trimming the end of the reads. (Krueger & Andrews 2012) This step should be taken in the pre-processing of the raw sequencing data. Selecting restricted alignment parameters would also increase the mapping stringency preventing sequences with several mismatches from aligning, thus reducing the number of erroneously inferred methylation states but at the cost of reducing mapping efficiency.(Krueger et al. 2012)

Secondly, the incomplete bisulfite conversion can cause an overestimation or underestimation of DNA methylation levels in some difficult samples that require extensive optimization.(Bock 2012) It was already proposed to use non-CpG contexts to detect reads that are likely not bisulfite-converted, discarding reads with more than 3 methylated cytosines in a non-CpG context but this will affect the studies that involve non-CpG methylation analysis.(Krueger & Andrews 2012) Another possible solution could be an extended or repeated bisulfite treatment that could raise the conversion of unmethylated cytosines but this would also affect the quality of DNA of the accuracy of bisulfite conversion.(Bock 2012) Thirdly, it is possible to use spike-in controls of non-native DNA with a known methylation state but it can’t be forgot that such controls might not have the same conversion properties that the DNA of interest.(Krueger et al. 2012)

Lastly, if the 3’ adaptor sequences are not removed correctly, they will remain in the reads, decreasing dramatically the mapping efficiency of the read and causing random methylation calls. It was already shown that mapping efficiency decreases with adaptor contamination and that each addition of cytosine in the adaptor spikes the level of methylation. For these reasons, it is essential to trim the 3’ adaptor sequences in the pre-processing of the reads step.(Krueger et al. 2012) The monitorization of GC content distribution and cytosine distribution can give some insights about a possible adapter contamination. Although this analysis yields variable results across species, it is known that the GC content distribution and the cytosine distribution can vary between 20-30% and 1-2% in mammalians, respectively. Then, if the occurrence of GC content and cytosine content rises up to 40-60% and 20%, respectively, this might indicate an adapter contamination.(Krueger & Andrews 2012)

Another common source of errors in methylation sequencing are the Single Nucleotide Variants (SNVs) (Hackenberg, Barturen & J.L. Oliver 2012). SNVs are variations in just one nucleotide between the reference and the sequenced genome(Hackenberg et al. 2012) . In a CpG sequence context there are usually two alleles, C and T, corresponding to the reference genome and sequence genome respectively. (Hackenberg et al. 2012) If the presence of this sequence variant is unknown or ignored, the inference would be that the cytosine annotated in the reference genome is unmethylated, while the correct conclusion should be that no cytosine exists in the genome and therefore no methylation state can be detected.(Krueger et al. 2012)

Lastly, given the size of a mammalian genome, the appearance of several independent fragments which align to the same genomic position is unlikely. However, a look at sequence distribution levels can quickly tell whether it is expected a lot of duplicate alignments, whether the library is diverse or whether the sample suffered from PCR amplification errors. In mammals, a sequence duplication level of 10% is indicative of a diverse library, but an 80% level indicates that the sample suffered from PCR duplication. Then, for large genomes removing duplicate reads that have the same orientation, start and end positions is essential.(Krueger & Andrews 2012)

4.1.6. Data visualization, statistical analysis and validation of results

After the data quality control and mapping steps, it is necessary to proceed to data visualization, statistical analysis and validation of results. Firstly, a genome browser is used to visualize and inspect a selection of genomic regions. It is necessary to use web available genome annotations to compare with the sequencing data. Then, it is important to identify differential methylated regions that exhibit consistently different DNA methylation levels between sample groups.(Bock 2012)

Lastly, the biological interpretation of data is done by comparison of the obtained list of DMRs. This list might be validated in accuracy and reproducibility by a manual comparison of the strongest DMR in a genome browser, visualization of global properties of the DMR list by quality-control plots and confirmation of the biological reproducibility of a DMR. Then, the data is interpreted by the biologist with help of additional computational tools.(Bock 2012)

4.1.7. Bioinformatics Tools

The evolution of the NGS technologies and the study of its data has demanded the development of several software applications in the last years.(Hackenberg et al. 2012) In the case of methylome analysis there has been an effort to develop bioinformatics tools that integrate several of the above mentioned steps. Currently, it is possible to distinguish two types of tools – alignment tools that perform the pre-processing of the reads and the alignment step but do not report methylation levels and the full pipeline tools that perform all necessary steps from the pre-processing to the methylation profiling, error control and statistical analysis.(Hackenberg et al. 2012) Hereafter, the three bioinformatics tools that were used in this project will be presented: Bismark Bisulfite Mapper(Babraham Bioinformatics 2016), Integrative Genomics Viewer(Robison T. et al. 2012) and Methy-Pipe(Jiang et al. 2014).

4.1.7.1. Bismark Bisulfite Mapper

Bismark Bisulfite Mapper is a set of tools for a time-efficient analysis of bisulfite-sequencing data written on Perl(Babraham Bioinformatics 2016) that has as main features: bisulfite mapping and methylation calling in one single step, support of single-end or paired-end read alignments, possibility to adjust the seed length and number of mismatches and the possibility of a discriminated output between cytosine methylation in a CpG, CHH or CHG context.(Babraham Bioinformatics 2016)

Bismark uses a three-letter aligner named Bowtie that aims to find a unique alignment (i.e. the genome position to which the read aligns under a given set of parameters(Hackenberg et al. 2012)) by running four alignment processes

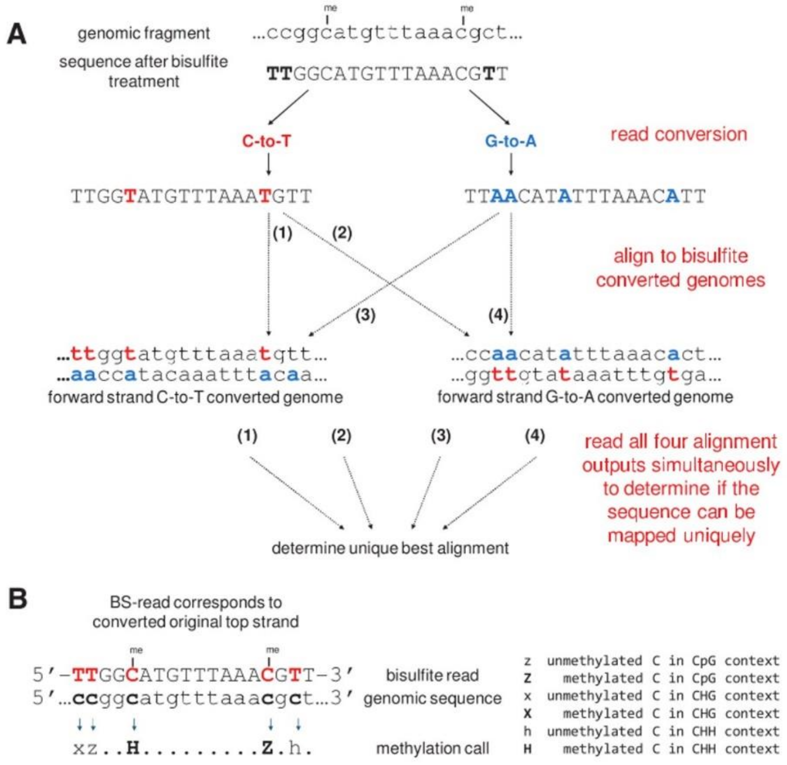

A
genomic fragment
me
me
...ccggcatgtttaaacgct...
sequence after bisulfite treatment
TTGGCATGTTTAAACGTT
C-to-T
G-to-A
read conversion
TTGGTATGTTTAAATGTT
TTAACATATTTAAACATT
(1)
(2)
(3)
(4)
align to bisulfite converted genomes
...ttggtatgtttaaatgtt...
...aaccatacaaatttacaa...
forward strand C-to-T converted genome
...ccaacatatttaaacact...
...ggttgtataaatttgtga...
forward strand G-to-A converted genome
(1)
(2)
(3)
(4)
read all four alignment outputs simultaneously to determine if the sequence can be mapped uniquely
determine unique best alignment
B
BS-read corresponds to converted original top strand
me
me
5′-TTGGCATGTTTAAACGTT-3′
bisulfite read
5′...ccggcatgtttaaacgct...3′
genomic sequence
xz..H.........Z.h.
methylation call
z unmethylated C in CpG context
Z methylated C in CpG context
x unmethylated C in CHG context
X methylated C in CHG context
h unmethylated C in CHH context
H methylated C in CHH context

Figure 19 – Overview of Bismark bisulfite mapper. a) After the bisulfite treatment, the reads are converted (C to T and G to A) and each of them is aligned to a reference genome that suffers the same type of conversion in order to determine the unique best alignment. B) The sequence reads with a unique alignment are compared to the original strand that allows the software to determine the cytosine methylation states in all methylation contexts.

simultaneously.(Krueger & Andrews 2011) The number of alignment processes of a mapper should be adaptable to all types of library preparation methods. For example, since in the non-directional library four alignments are built, Bismark should handle them (Krueger & Andrews 2011) (Figure 19). Sequence reads are firstly transformed into bisulfite-converted forward and reverse reads(Babraham Bioinformatics 2016) where the cytosines are converted to thymines and guanines to adenines, respectively. Each of them is aligned in parallel to reference genomes that suffer the same type of transformation(Krueger & Andrews 2011) and then the sequence reads with an unique alignment are compared to the normal genomic sequence that allows the inference of all cytosine methylation states.(Babraham Bioinformatics 2016)

Unlike other bisulfite mappers, Bismark contains an extraction process that determines the methylation state of each cytosine position in the read(Krueger & Andrews 2011), producing a report containing several useful information: 1) summary of alignment parameters used; 2) number of sequences analysed; 3) number of sequences with a unique best alignment; 4) statistics summarising the bisulfite strand where the best unique alignment came from; 5) number of cytosines analysed; 6) number of methylated and unmethylated cytosines and 7) percentage cytosines methylation in CpG, CHH or CHG context.(Babraham Bioinformatics 2016)

4.1.7.2. Integrative Genomics Viewer

The Integrative Genomics Viewer (IGV) is a high-performance desktop tool for interactive visual exploration of diverse and large-scale genomic data written in Java that appeared in 2007 but only in 2009 was adapted to short-read sequence alignments.(Thorvaldsdóttir et al. 2013) IGV supports integration of aligned sequence reads, mutations and copy number data, RNAi screens, gene expression, methylation and genome annotations.(Robison T. et al. 2012) This software also allows investigators to flexibly visualize different types of data together, supporting the view and the manipulation of multiple genomic regions side by side. Additionally, it integrates data with the display of sample attribute information, supports direct manipulation, navigation and real-time interaction at all scales of genome resolution, from whole-genome to single base pairs.(Thorvaldsdóttir et al. 2013)

In the IGV, the reference genome must be selected from several reference genomes available from public sources or user incorporated. About loading and viewing data, IGV supports a wide variety of file formats for genome annotation, sequence alignment, variant call and microarray data as also imported metadata information. IGV allows a simultaneous viewing of multiple data sets with default appearance and has available view options depending on the data type.(Thorvaldsdóttir et al. 2013)

4.1.7.3. Methy-Pipe

Methy-Pipe was implemented in 2014 by Jiang, P. *et al.*(Jiang et al. 2014) and is a full pipeline tool that not only meets the core methylation data analysis but also provides tools to facilitate the downstream analysis in an efficient and integrative manner. It is implemented using Perl, R and C++ and can be run in a Linux operating system. Methy-Pipe analyzes high-throughput bisulfite sequencing reads on FASTQ format from either single or paired-end libraries using two consecutive software modules.(Jiang et al. 2014)

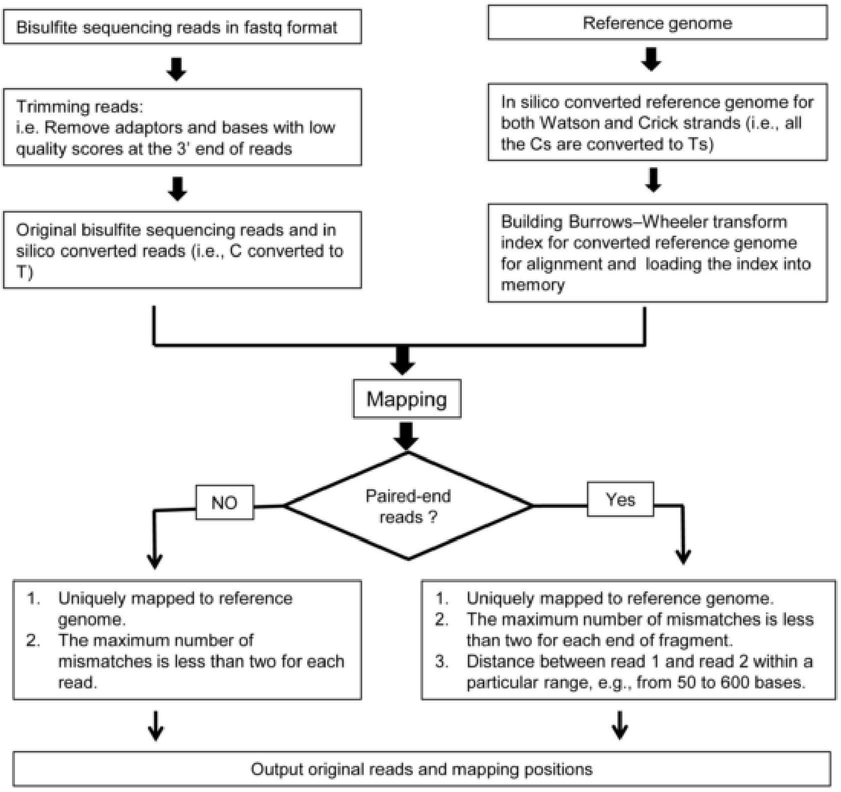

Bisulfite sequencing reads in fastq format
Reference genome
Trimming reads:
i.e. Remove adaptors and bases with low quality scores at the 3' end of reads
In silico converted reference genome for both Watson and Crick strands (i.e., all the Cs are converted to Ts)
Original bisulfite sequencing reads and in silico converted reads (i.e., C converted to T)
Building Burrows–Wheeler transform index for converted reference genome for alignment and loading the index into memory
Mapping
Paired-end reads ?
NO
Yes
1. Uniquely mapped to reference genome.
2. The maximum number of mismatches is less than two for each read.
1. Uniquely mapped to reference genome.
2. The maximum number of mismatches is less than two for each end of fragment.
3. Distance between read 1 and read 2 within a particular range, e.g., from 50 to 600 bases.
Output original reads and mapping positions

Figure 20 – Representation of bisulfite sequence read alignment in BS Aligner. Its process starts with a conversion of reference genomes and bisulfite sequencing reads in FASTQ format. A BWT (Burrows-Wheeler Transform) algorithm is used to create whole genome sequence indices that are firstly loaded into the computer memory. In the alignment the paired-end or single-end reads have a different process where in the paired-end reads the insert size is also taken in account addition to considering the number of mismatches. After this, it is obtained a text file with output information and mapping positions.

The first module, named BSAligner, is designed bisulfite-treated read alignment, but includes data pre-processing. In the pre-processing step, the adaptors and the bases with quality score below five are removed. After the preparation of reference genome and reads through the three-letter approach, where all Cs are converted to Ts, the pre-processed and converted reads are aligned to the pre-converted reference genomes and all the reads that can be aligned back to the Watson and Crick strands are discarded. The remaining are replaced by the original bisulfite sequencing reads and used for downstream analysis (Figure 20). (Jiang et al. 2014)

The second module is a data analysis tool named BS Analyzer that reports the basic statistics and sequencing quality of the data, profiles the regional and global methylation level, identifies DMRs for paired samples, annotates and visualizes genome-wide methylation data.(Jiang et al. 2014)

Jian, P. *et al.* (2014) demonstrated that Methy-Pipe can efficiently and accurately analyze the whole genome bisulfite sequencing data and in comparison with other software packages, it has more functionality and greater usability since it integrates the core and the downstream data analysis into one single package and uses high-performance computing clusters to parallelize data analysis, speeding up the analysis of bisulfite sequencing.(Jiang et al. 2014) Particularly, the BSAligner demonstrated on outperformance in comparison to Bismark Bisulfite Mapper in terms of computation time and with a comparable alignment accuracy.(Jiang et al. 2014) Since Bismark Bisulfite Mapper has shown to outperform many previously reported mapping programs, like BSMAP, BS Seeker and MAQ in terms of the ability for paired-end read alignment and running time, Methy-Pipe became quite relevant in the epigenomics research community.(Jiang et al. 2014)

4.2. Bioinformatics in microarray-based methods for methylome analysis

The accurate interpretation and analysis of microarray data requires the application of several bioinformatics methodologies that can be structured into several steps - file extraction, quality control, normalization, data analysis and biological interpretation (Figure 21).

The Illumina BeadChip 450k array originates an Illumina Intensive Data (idat) file format as raw data and each sample has a Red and a Green idat file that represent the intensities of the methylated and unmethylated probes. They are used to determine the β-value (Equation 1), obtained through the methylation scores for each CpG, ranging from 0 (unmethylated) to 1 (fully methylated) on a continuous scale. It is calculated from the intensity of the M and U alleles as the ratio of fluorescent signals and the 100 constant exists to stabilize the β-values when the intensities are low.(Wu & Kuan 2018) The M-value can also be used to measure methylation where a normalized M-value near 0 signifies a semi-methylated locus, a positive M-value indicates that more molecules are methylated than unmethylated and a negative M-value have the opposite interpretation (Equation 2).(Wright et al. 2016) Although an M-value is attractive in that it can be used in many statistical models derived for expression arrays that assume normality, β-values are much more biologically interpretable.(Wright et al. 2016)

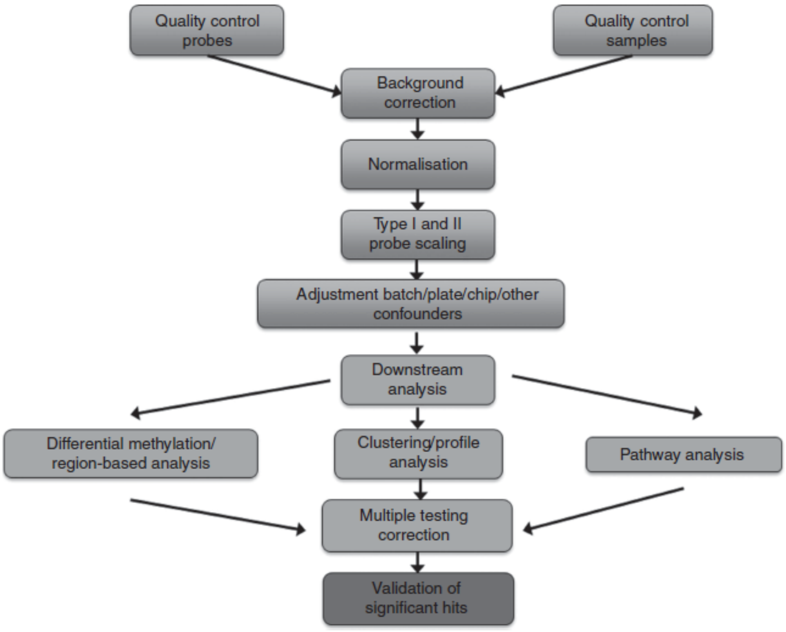

Quality control probes
Quality control samples
Background correction
Normalisation
Type I and II probe scaling
Adjustment batch/plate/chip/other confounders
Downstream analysis
Differential methylation/ region-based analysis
Clustering/profile analysis
Pathway analysis
Multiple testing correction
Validation of significant hits

Figure 21 – Typical procedure of a bioinformatics microarray protocol. Starts with the quality control steps and is followed by normalization, statistical analysis, biological interpretation and validation of data.

$$\beta = \frac{\mathrm{M}}{\mathrm{M} + \mathrm{U} + 100} \qquad \text{Equation 1}$$

$$\mathrm{M} = \log_2 \frac{\mathrm{Max}(\mathrm{M}, 0)}{\mathrm{Max}(\mathrm{U}, 0)} = \log_2 \frac{\beta}{1 - \beta} \qquad \text{Equation 2}$$

Some examples of software tools for the analysis and interpretation of DNA methylation microarray data are *methylumi*, *minfi*, *wateRmelon*, *ChAMP* and *RnBeads*. All of them are Bioconductor R packages (Figure 22) that will be explained bellow.

methylumi enables the user to perform quality control interrogation, three methods of background correction and normalization and also works with GoldenGate and 27k array. *minfi* does not provide a single function to run the entire pipeline but it is frequently updated to offer methods for the newest analysis options available to 450k users. It has DMR calling, block finding modules and a new between-array normalization algorithm, termed functional normalization. It has quality control reports with a HTML option that included visualization of the array's internal controls(Morris & Beck 2015). *watermelon* provides access to 15 normalization methods and 3 performance metrics based on three natural controls. *ChAMP* automates some of the *minfi* functions for a more inexperienced R user and offers eight functions that can be manipulated to set parameters. *RnBeads* has four normalization methods and a detailed HTML report that describes the analyses done along with results and images. It includes functions for annotation inference and data visualization.

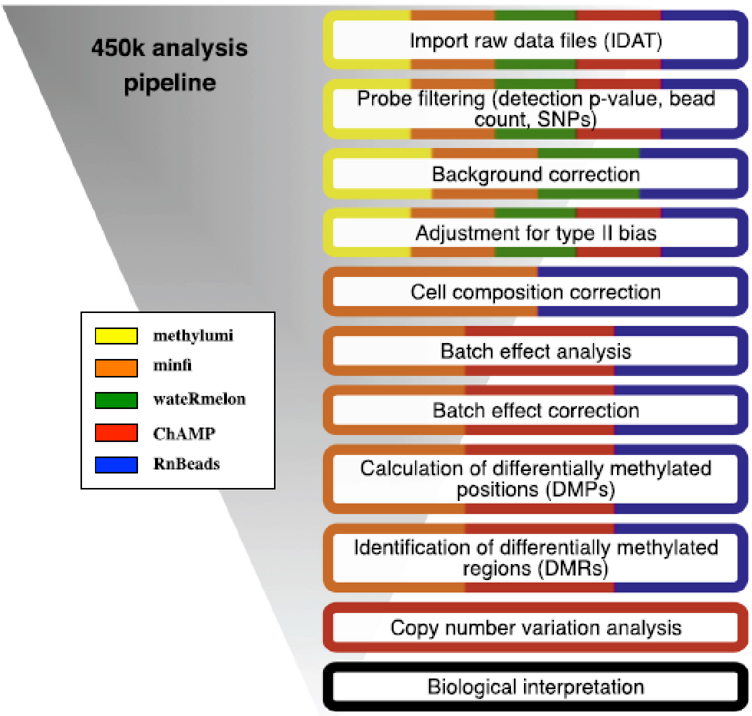

450k analysis pipeline
methylumi
minfi
wateRmelon
ChAMP
RnBeads
Import raw data files (IDAT)
Probe filtering (detection p-value, bead count, SNPs)
Background correction
Adjustment for type II bias
Cell composition correction
Batch effect analysis
Batch effect correction
Calculation of differentially methylated positions (DMPs)
Identification of differentially methylated regions (DMRs)
Copy number variation analysis
Biological interpretation

Figure 22 – Pipeline steps offered by several softwares used in 450k methylation array. With a broadest offer proposed by ChAMP, RnBeads and *minfi*. Adapted from (Morris & Beck 2015)

4.2.1. Quality control

The quality control step is the first step in any pipeline for microarrays, consisting on the estimation of the quality of a dataset and selection of reliable probes and samples.(Touleimat & Tost 2012) The quality control or sample filtering is made because the Infinium arrays include several control probes for determining the data quality including sample-independent and dependent controls.(Wilhelm-Benartzi et al. 2013) Some probes can assess the bisulfite conversion efficiency or background fluorescence levels.(Wright et al. 2016) The process can be made through the detection of poorly performing samples using diagnostic plots of control probes or using the raw signal intensities of the control probes and determining whether they are beyond the expected range of the signal intensities across all samples. (Wilhelm-Benartzi et al. 2013) If the control probe intensity values fall outside the clustering values for other samples, this could indicate a compromised sample.(Wright et al. 2016) Usually, principal component analysis is performed to detect potential batch effects when samples are processed on more than one array.

On the other hand, the filtering of probes is made if a certain proportion of samples have a detection P-value below a certain specified threshold.(Wright et al. 2016) This method can eliminate probes with intensity levels at or near background intensity, determined by several negative probe controls included in 450k array, poorly represented CpG sites and variable target sequences.(Wright et al. 2016) The quality control of probes can be made removing probes that fail to measure DNA methylation in a certain proportion of the total samples, through the identification of probes that failed to hybridize to a minimum of beads and cannot be detected by array.(Wright et al. 2016) Additionally, the probes located between single nucleotide polymorphisms (SNPs) should also be excluded, since this features can disrupt probe binding at that site, representing false low intensity signals, affecting the DNA methylation measurement.(Wright et al. 2016) The probes associated with sex chromosomes (X and Y) can also be removed since they account for the larger gender effects that researchers have found, or even the CpG probes located near short insertions or deletions or the ones that map to multiple locations on the genome, since they can produce difficult results to interpret.(Wright et al. 2016),(Wilhelm-Benartzi et al. 2013)

4.2.2. Preprocessing and normalization

The measurement of methylation levels can be affected by several sources of systematic variation in microarray experiments and so the data generated need to be normalized before the application of any mathematical methods.(Khademhosseini et al. 2013) Normalization removes the impact of nonbiological influences on the data, requiring the adjustment in three technical artifacts: non-specific background fluorescence, red/green dye bias and rescaling for probe type differences.(Wright et al. 2016) The normalization is made through between-array normalization, removing technical artefacts between samples on different arrays or through within-array normalization, correcting for intensity-related dye biases.(Wilhelm-Benartzi et al. 2013)

The background correction methods help to remove nonspecific signal from total signal and corrects for between-array artefacts.(Wilhelm-Benartzi et al. 2013) More advanced model-based background correction methods take advantage of the 450k array technology to measure the intensity level of type I probes outside of their specified color band and have been shown to be superior to subtractive methods that rely exclusively on the negative probes.(Wright et al. 2016) Additionally, owing to the difference in labeling efficiency and scanning properties of the two color channels, the intensities measured in the two color channels might be imbalanced.(Touleimat & Tost 2012) Therefore, it is necessary to make a color balance adjustment if the color effect is inconsistent across samples.(Touleimat & Tost 2012)

Furthermore, as referred before, the 450k array technology uses two different types of probes that need to be rescaling to make its distributions comparable.(Wilhelm-Benartzi et al. 2013) The first method proposed to correct this divergence was peak-based correction where the Infinium II data is rescaled on the basis of the Infinium I data, assuming a bimodal shape of the methylation density profiles.(Wilhelm-Benartzi et al. 2013) However, it is known that the density distribution does curves and does not work well when the density distribution does not exhibit well-defined peaks or nodes.(Wilhelm-Benartzi et al. 2013) Therefore, currently there are four alternative approaches to this method – subset-quantile within-array normalization (used in *minfi*), subset quantile normalization, β-mixture quantile dilation (BMIQ) normalization method (used in wateRmelon) and funnorm normalization method (also available in *minfi*).(Wilhelm-Benartzi et al. 2013) The first determines an average quantile distribution using a subset of probes defined to be biologically similar on the basis of CpG content and allows the Infinium I and II probes to be normalized together.(Wilhelm-Benartzi et al. 2013) The second uses the genomic location of CpGs to create probe groups through which they apply subset quantile normalization.(Wilhelm-Benartzi et al. 2013) The reference quantiles used in this approach are based on type I probes with significant detection P-values.(Wilhelm-Benartzi et al. 2013) The third uses quantiles to normalize the type II probes using a β-mixture model fit to the type I and II probes separately and then transforms the probabilities of class membership of the type II probes into quantiles using the parameters of the β-distributions of the type I distribution.(Wilhelm-Benartzi et al. 2013) Finally, the fourth is used when a global DNA methylation shift is expected, using internal control probes present on the array to infer between-array technical variations.(Hansen 2018)

4.2.3. Batch correction and cell composition

DNA methylation arrays are susceptible to batch effects, effects caused by a group of samples that undergo an experimental processing step in tandem, potentially introducing DNA methylation differences that reflect differences between batches and not in experimental factors of interest.(Wright et al. 2016) The most common type of batch is observed when experimental procedures necessitate the processing of samples in separate groups or in different days.(Wright et al. 2016) Although normalization has already been shown to reduce some component of batch effects, not all are adjusted, being necessary to use methods to correct batch effects.(Wilhelm-Benartzi et al. 2013)

However, array position effects may also exist and thus new batch correction techniques may be needed to take those into account.(Wilhelm-Benartzi et al. 2013) In the cases where the true sources of batch effects are unknown or cannot be correctly modelled, is necessary to use a method that estimates the source of batch effects, like Surrogate Variable Analysis (SVA), or a method that identifies features that correlate the phenotype of interest in the presence of potential confounding factors, like Independent Surrogate Variable Analysis (ISVA).(Wilhelm-Benartzi et al. 2013)

Additionally, it is known that the DNA methylation can vary by cell type and when we compare a group of samples that contain different cell types if its amount changes, it can affect the results.(Wright et al. 2016) Therefore, if a sample include abnormal cell-type proportions, the identification of significant DNA methylation differences due to this is essential to avoid associations of this variations to the health condition being evaluated in the research.(Wright et al. 2016) So, the evaluation of DNA methylation in mixed cell tissues should use statistical corrections to estimate heterogeneity of cell types found among samples.(Wright et al. 2016)

4.2.4. Data analysis

After all the quality control and normalization steps, methylated positions need to be properly compared between groups through several statistical methods, mainly through the identification of differential methylated individual CpG positions (DMPs) and DMRs and clustering analysis. The former is an essential step in the analysis of array-based DNA methylation data since it consists on the grouping of objects into clusters according to their similarity.(Wilhelm-Benartzi et al. 2013)

The identification of DMRs, composed of multiple near DMPs, is essential to identify methylation differences, like probe-wise or locus-specific methylation differences, between specific groups such as cases and controls. (Wright et al. 2016)'(Wilhelm-Benartzi et al. 2013) Since the probes are placed in a sparse and non-uniform way in 450k, the identification of DMRs remains a challenge.(Wright et al. 2016) Therefore, it is recommended that both DMPs and DMRs detection be run in tandem.(Wright et al. 2016) Additionally, it is essential to take into account CpG proximity, since nearby CpG loci tend to have methylation levels highly correlated, and the tissue being sampled because the extent of methylation can reflect true changes to the methylome but can also represent heterogeneity in underlying cell-type distributions.(Wilhelm-Benartzi et al. 2013) Some tool examples to determine DMPs and DMRs are MethVisual, *minfi*, limma, IMA, CHARM and EVORA, each one with specific features.(Wilhelm-Benartzi et al. 2013) Once the analysis has identified top hits, it is necessary to make a multiple testing correction to reduce the likelihood of false-positive loci by adjusting statistical confidence measures by the number of tests performed.(Wright et al. 2016)

4.2.5. Biological interpretation

The interpretation of biological data is the most important step once the bioinformatics analysis of genomic data is concluded(Wright et al. 2016). There are several interpretation-oriented approaches aimed at understanding the biological and clinical significance of DNA methylation data.(Wright et al. 2016)

Many researchers use functional enrichment analysis to reveal biological roles of differentially expressed DMRs through mapping them to the nearby genes in a process named Gene Ontology (GO).(Wright et al. 2016)˒(Wilhelm-Benartzi et al. 2013) Although the mapping process is made associating the probe ID of each DMR with gene names, if a CpG site maps to several nearby genes, one may elect to use all these genes.(Wright et al. 2016) It is recommended in the DNA methylation analysis to stratify the data by gene region to decrease the potential for bias originated by the different number of probes in specific regions.(Wright et al. 2016) After this, the functional enrichment can be performed using Gene Set Enrichment Analysis (GSEA), Database for Annotation, Visualization and Integrated Discovery (DAVID) or ToppGene.(Wright et al. 2016) However, if a CpG site maps to multiple nearby genes, the regulatory context of DMRs should be evaluated, in a process named regulatory enrichment analysis, since it may be difficult to know which gene is truly regulated by the methylation differences in that CpG site.(Wright et al. 2016)

CHAPTER II OBTAINING DATA

Using data from public databases

1. INTRODUCTION

Employing informatics to study biological data has become a routine in recent epigenomic studies, rising the concept of "*in silico* experiments", i.e. the use of several bioinformatics tools to extensively study biological systems, saving time, expenses and human resources. Therefore, retrieving information from databases of genomic, proteomic and transcriptomic data, even if already studied, is essential so that researches can use them differentially or to validate different approaches. Under this goal data mining becomes an important step for successful retrieving of required contents from any database.

The International Nucleotide Sequence Database (INSD) is one of the major initiatives in public domain data sharing and consists of three collaborators: DNA Data Bank of Japan (DDBJ), European Nucleotide Archive (ENA) and National Center for Biotechnology Information (NCBI).(Karsch-Mizrachi et al. 2018) These partners work together to preserve all public domain nucleotide sequence data, turning it accessible in standardized formats across the three sites through daily data exchange. INSD data is free to users leading to future new important discovers since INSD databases are data hosts but not owners.(Karsch-Mizrachi et al. 2018) In 2017 the assembled and annotated data consisted of a 2,650 total trillion bases, i.e. about 3,2 Petabytes.(Karsch-Mizrachi et al. 2018)

The National Center for Biotechnology Information (NCBI) was created in 1988 as a division of the United States National Library of Medicine (NLM) at the National Institute of Health (NIH). Currently, it provides biological data and resources focused on literature, health, genomes, genes, protein and chemicals. (Agarwala et al. 2016)

GEO, Gene Expression Omnibus, is a NCBI's data repository that aims to provide high-throughput functional genomic data through a user-friendly database with an open and flexible design that facilitates submission and can hold raw and processed data for further study.(Edgar 2002) This platform contains data from gene expression, gene copy number, gene-protein interactions and methylation profiling generated by microarray and NGS technology. (Agarwala et al. 2016),(Edgar 2002)

GEO is organized into two Entrez databases: GEO datasets where the data is organized in a study-level format so that users can search for studies relevant to their interests and GEO profiles where the data is organized in a gene-level manner that users can search for gene expression profiles.(NCBI n.d.) In both cases, the information is organized in platforms, series and experiments. A platform, represented as GPL prefix, defines the technology used in an experiment and is associated with a list of probes that may detect a certain set of molecules. A sample, represented as a GSM prefix, defines an individual sample, its handling conditions and characteristics. A series, represented with a GSE prefix, defines an organization of samples that belong to the same experiment.(Edgar 2002)

A particular advantage of GEO is its usage through Bioconductor, a software project that provides tools for the analysis and comprehension of high-throughput genomic data, using the R programming language.(Bioconductor n.d.) GEOmetadb, GEOquery, GEOsearch and GEOsubmission are examples of this applicability. GEOmetadb makes querying the GEO metadata easier and more effective through the use of GEOmetadb.SQLite, a locally database that is regularly updated. This allows the GEO search to be more detailed than using the online search NCBI tool.(Zhu et al. 2008) On the other hand, GEOquery is a package that downloads the Simple Omnibus Format in Text (SOFT) files, that are designed for rapid batch submission and includes information about the experiment, from GEO.(Sean & Meltzer 2007)

Since one of the objectives of this thesis was to process and analyze raw data, obtained from public databases, as a way to gather some genome-wide tissue-specific epigenetic information related to aging and to gain experience with the bioinformatical tools needed to do so, the first step was to obtain the files before its analysis. The Sequence Read Archive (SRA) is an international public archive for next-generation sequencing data that is also under the guidance of INSDC and its usage is also facilitated by Bioconductor, specifically SRAdb. SRAdb.SQLite is also regularly updated and therefore this package was used to download the necessary raw data for our analysis. (Leinonen et al. 2011), (Zhu et al. 2013)

2. METHODOLOGIES ADOPTED

Due to the reduced costs associated with *in silico* approach to study epigenomic variances, this Master's Thesis was early based on the study and data mining of metadata available on public databases. Firstly, *Mus musculus* was used as model organism of aging, since the institute aimed to study mice using bisulfite targeted sequencing in the near future. However, due to the limitations verified in this process and described on the discussion of this chapter, we turned our attention *Homo sapiens*. Still, the main goal was to construct tissue-specific genomic maps with age-related DMPs in human and use it in further studies of iBiMED.

2.1. Obtaining NGS data of *Mus musculus*

The process of obtaining NGS data is made in several steps, starting with the selection of datasets from GEO, extracting additional information present in the files that are relevant for the analysis, and finally downloading the raw NGS data. All of these processes were made using several packages of Bioconductor in RStudio 64 bits version 3.4.3.

Firstly, datasets were obtained using GEOmetadb package (Zhu et al. 2008). In this process, a script was written to select datasets by organism, methylation profiling method and technology. (Box 1)

Box 1 – Script for obtainment *Mus musculus* dataset after filtration using GEOmetadb package

```
#install Geometadb
source(https://bioconductor.org/biocLite.R)
biocLite(“GEOmetadb)

#Download the data and library Geometadb
library(GEOmetadb)
if(!file.exists('GEOmetadb.sqlite')) getSQLiteFile()
con <- dbConnect(SQLite(),'GEOmetadb.sqlite')

#Filtration by features of our interest
sql<-paste("SELECT gse.gse, gse.type, gse.title, gse.summary, gse.overall_design,
gse.status, gse.pubmed_id, gpl.gpl, gpl.title, gpl.technology, gpl.organism,
gsm.gsm, gsm.type, gsm.organism_ch1, gsm.source_name_ch1, gsm.characteristics_ch1,
gsm.supplementary_file, gsm.characteristics_ch2, gsm.status",
    "FROM",
    " gsm JOIN gse_gsm ON gsm.gsm=gse_gsm.gsm",
    " JOIN gse ON gse_gsm.gse=gse.gse",
    " JOIN gse_gpl ON gse_gpl.gse=gse.gse",
    " JOIN gpl ON gse_gpl.gpl=gpl.gpl",
    "WHERE",
    "gse.type LIKE '%Methylation profiling by high throughput sequencing%' AND
    gpl.technology LIKE '%high-throughput sequencing%' AND
    gsm.organism_ch1 LIKE 'Mus%' AND
    gpl.organism LIKE 'Mus%' AND
    gsm.type LIKE '%SRA%'", sep=" ")
data <- dbGetQuery(con,sql)

library(xlsx)
write.xlsx(data, file="data.xlsx")
```

Then, the dataset was manually refined to assume that only healthy, adult and bisulfite-seq analysed mice were included. The GEOquery package, downloaded using `getGEO` command, was used to obtain the metadata in SOFT files of the list of experiments obtained before.Finally, the FASTQ files were downloaded using SRAdb package(Zhu et al. 2013) (Box 2).

Box 2 – Script for downloading SOFT and FASTQ files from samples of *Mus musculus*

```
#install GEOquery
source("https://bioconductor.org/biocLite.R")
biocLite("GEOquery")

library(GEOquery)
geo <- c('GSM0000000') # where geo is a vector of multiple GSMs
for (i in 1:length(geo)){
       getGEO(geo[i])
}

#install SRAdb
source(“https://bioconductor.org/biocLite.R”)
biocLite(“SRAdb”)
library(SRAdb)
sra_dbname <- 'SRAmetadb.sqlite'
sra_con <- dbConnect(dbDriver("SQLite"), sra_dbname)
list <- c('SRX000000') # where list is a vector of multiple SRXs
getSRAfile(in_acc=(list),sra_con=sra_con,destDir=getwd(),fileType='fastq')
```

2.2. Filtering and quality control of *Mus musculus* NGS data

Mus musculus samples obtained as explained above went through the first step of the bioinformatics pipeline described in the previous chapter in section 4. Samples were filtered using our homemade protocol (BIOVIA PipelinePilot 2017) with filters that trim reads by quality with a quality cutoff of 20, length with a minimum length of 50, ambiguity with a threshold of 5 and average quality with a quality score lower than 20 (Figure 23). The choosen parameter values have already been tested and were used routinely at iBiMED.

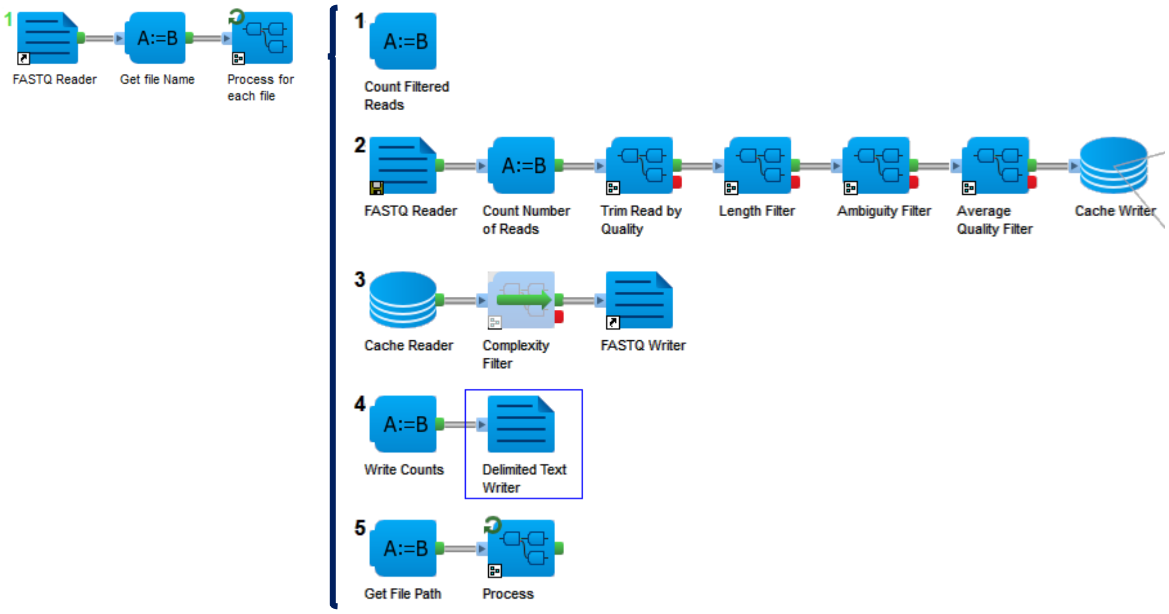


Figure 23 - Homemade pipeline for the preprocessing and filtering of reads in Pipeline Pilot. Filters for trimming by quality, length, ambiguity and average quality were used with standard parameters.

The reads were then quality controlled using FASTQC (version 0.11.5) (Babraham Bioinformatics 2010). This is a Babrahams Bioinformatics software that provides a graphical environment for quality control checks on raw sequence data coming from high throughput sequencing pipelines. Since in paired read situations, files with paired reads should have the same number of reads, it becomes necessary a last filtering step to discard the reads that have no pair in the opposite file (Figure 24).

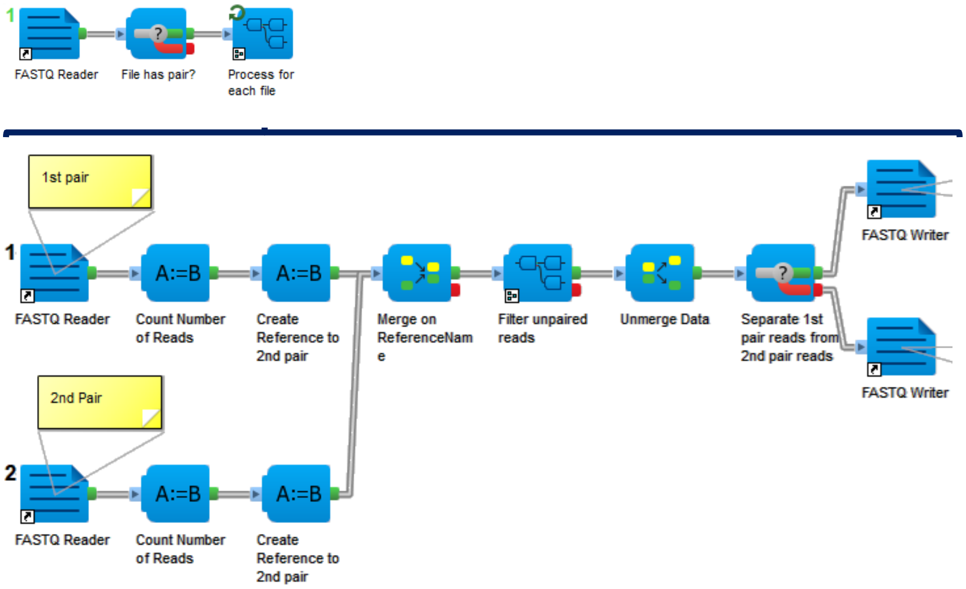


Figure 24 - Homemade pipeline for filtering unpaired reads in Pipeline Pilot. The process is runned after the first preprocessing step and further quality control using FASTQC.

2.3. Alignment of *Mus musculus* samples

Following the preprocessing and quality control steps, the samples wen't to the alignment step. At this moment, and since the software used in our pipeline (MethyPipe) was not prepared for the alignment of samples other than human, we decided to run them on Bismark. For this, Bismark Bisulfite Mapper version 0.14.4 (Babraham Bioinformatics 2016) was used. The specifications used were the default ones and the reference genome was mm10 from C57BL/6J mouse strain.

2.4. Obtaining microarray data from *Homo sapiens*

Human samples were obtained on GEO DataSets(NCBI n.d.) query and browsed using the following parameters: *"Homo sapiens [Organism] AND ("methylation profiling by array" [DataSet Type] OR "methylation profiling by genome tiling array" [DataSet Type] OR "methylation profiling by high throughput sequencing" [DataSet Type] OR "methylation profiling by SNP array" [DataSet Type])"*. This procedure reduced the 96453 available series of samples in GEO to 2513 that were filtered with "age" as the "Attribute Name". This resulted in 471 series of samples.

Afterwards, the series were manually analyzed to assure that samples were from tissues of perfectly healthy adult individuals. Individuals with reported bad life styles (e.g. smoking, drinking) and known diseases were excluded. Experiences that manipulated individuals were also discarded. The ethnic differences were ignored. Then, the GEOquery package, downloaded using `getGEO` command, was used to obtain the metadata in SOFT files and idat files from these experiments, using RStudio 64 bits version 3.4.3 as below (Box 3).

Box 3 – Script for downloading SOFT and idat files using GEOquery

```
library(GEOquery)

getGEOSuppFiles("GSE105123")
untar("GSE105123_RAW.tar", exdir = "GSE105123/idat")
```

3. RESULTS

3.1. Dataset from *Mus musculus*

13302 samples, 574 series and 23 platforms were obtained using GEOmetadb package. These were refined, through a manual analysis, to check for healthy, non-embryo, non-embryonic stem cells, non-cell lines, bisulfite-seq samples and with age information, which resulted in 64 samples and 10 series (Supplementary Table 1). However, to turn this set comparable, a lot of similarities were necessary - mouse strain, library preparation strategy and type of reads used in library preparation. Since we wanted to compare the methylation status through age in several tissues, it was necessary to have at least two samples with different ages from each tissue. In the case of GEO samples with the same tissue and age specificity, they were selected according to the higher number of final reads, after the filtration step described below. The resultant Dataset A is represented on Table 5.

Table 5 - Dataset A obtained for *Mus musculus* with information about GEO experiment, GEO sample, strain, age, tissue, technology, its reference genome and library strategy.

GEO Series	*GEO Sample*	*Strain*	*Age*	*Tissue*	*Reference Genome*	*Platform*	*Library strategy*
GSE68618	GSM1677165	C57B|6	16-18 mo	Pancreas	mm9	HiSeq2000	Bisulfite-seq
GSE68618	GSM1677166	C57B|6	4-6 w	Pancreas	mm9	HiSeq2000	Bisulfite-seq
GSE70317	GSM1723692	C57BL/6N	7 w	Liver	mm9	MiSeq	Bisulfite-seq
GSE72177	GSM1857045	C57/BL6	22 w	Liver	mm9	HiSeq2000	Bisulfite-seq
GSE92486	GSM2430564	C3B6F1	5m	Liver	mm10	HiSeq2500	Bisulfite-seq
GSE92486	GSM2430570	C3B6F1	26m	Liver	mm10	HiSeq2500	Bisulfite-seq

3.1.1. Preprocessing steps

After sample selection quality and trimming, the retrieving fastq files were selected according to the higher number of final reads. The results of the selected fastq files are summarized in Table 6. The complete list of results before selection can be seen in Supplementary Table 2.

Table 6 – First filtering step of Dataset A and the distribution of filtered reads among the several stages of the process of the already selected samples. The selection was made according do the higher mean final reads of each sample.

GEO Series	*GEO Sample*	*Acession fastq file*	*Initial reads*	*Trim Filter*	*Ambiguity Filter*	*Quality Filter*	*Final reads*	*Mean reads filtered*
GSE68618	GSM1677165	SRR2034989_1	240688878	16247555	94637	635	224346051	7%
		SRR2034989_2	240688878	16301684	104652	66285	224216257	
		SRR2034993_1	213105508	2564085	52936	34	210488453	2%
		SRR2034993_2	213105508	6646782	48550	13029	206397147	
GSE70317	GSM1723692	SRR2079727_1	199775	0	29	0	199746	0,05%
		SRR2079727_2	199775	1	168	3	199603	
GSE72177	GSM1857045	SRR2173864_1	37550718	260253	3477	0	37286988	1%
		SRR2173864_2	37550718	603998	5235	766	36940719	
	GSM1857046	SRR5115679_1	70538476	1414173	4174	18413	69101716	2%
		SRR5115679_2	70538476	1414173	4174	18413	69101716	
GSE92486	GSM2430570	SRR5115685_1	229564398	500653	1784	1777	229060184	0,4%
		SRR5115685_1	229564398	1090046	114848	7429	228352075	

3.1.2. Quality control

Samples were quality-controlled using FASTQC. This tool produces reports about the base sequence quality, sequence quality scores, base sequence content, base GC content, sequence GC content, base N content, sequence length distritution, sequence duplication levels, overrepresented sequences and Kmer content.(Babraham Bioinformatics 2010) This information was analysed in our samples and decisions were based in the official documentation of FASTQC. (Babraham Bioinformatics 2010)

In general, all the samples presented a bad quality control in kmer plots and base sequence content and several warnings in sequence GC content. However, there is empirical evidence of overrepresentation of reads in methylated DNA that can arise in the construction of sequencing libraries using bisulfite.(Ji et al. 2014) In any case, the other reports generally presented a good quality and we decided to continue the procedure to the alignment step. (Table 7)

Table 7 – FASTQC outcome across Dataset A in both paired reads demonstrated a bad quality in kmer content and base sequence content and several warnings in sequence GC content.

FASTQC Steps	*SRR2034989*		*SRR2034993*		*SRR2079727*		*SRR2173864*		*SRR5115679*		*SRR5115685*	
	R1	*R2*	*R1*	*R2*	*R1*	*R2*	*R1*	*R2*	*R1*	*R2*	*R1*	*R2*
Base sequence quality	G	G	G	G	G	G	G	G	G	G	G	G
Tile sequence quality	W	G	G	W	N/A	N/A	G	G	G	G	G	G
Sequence quality scores	G	G	G	G	G	G	G	G	G	G	G	G
Base sequence content	B	B	B	B	B	B	B	B	B	B	B	B
Sequence GC content	B	G	B	B	B	B	W	W	G	G	G	G
Base N content	G	G	G	G	G	G	G	G	G	G	G	G
Sequence Length Distribution	W	W	W	W	W	W	W	W	W	W	W	W
Sequence Duplication Levels	W	G	B	B	B	B	G	G	G	G	W	W
Overrepresented sequences	B	W	W	W	B	B	G	G	G	W	G	G
Adapter Content	B	G	B	B	G	G	G	G	G	G	G	G
Kmer Content	B	B	B	B	B	W	B	W	W	B	B	B

G – Good quality control | W – Warning quality control | B – Bad quality control | N/A – non-available

Bellow we exemplify the plots available and the rest of the data is presented in Supplementary Figures 1-10. The plot of base sequence quality score (Figure 25, left) shows that the sample presented a good quality although it is possible to see a falling of quality in the end of the run progress which is normal since the sequencing chemistry tends to degrade with increasing read length especially for long runs. However, in this case the trimming was not necessary since that was not significantly lower in quality. In the case of per sequence quality scores (Figure 25, right), we also had a good quality of samples, since they were all above 27 of quality score.

On the other hand, we compared of two samples with a better and worst scenario using per tile sequence quality as can be assessed in Figure 26. The worst samples in this case represent individual specific events confined to a specific area or range of cycles since they do not appear all over the run and cannot be considered bad quality samples.

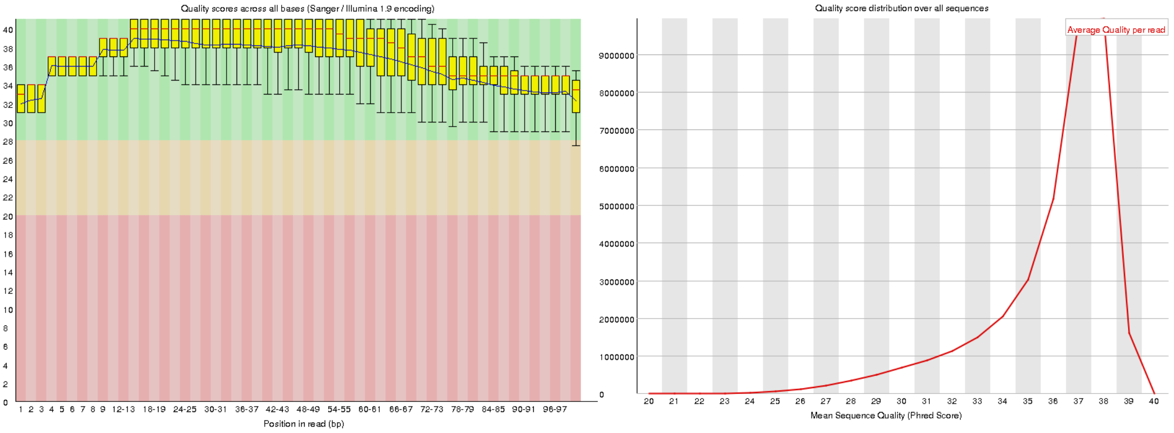


Figure 25 - Base sequence quality (left) and sequence quality score (right) of the SRR2173864 sample. The sample reveals a good quality in both categories of FASTQC.

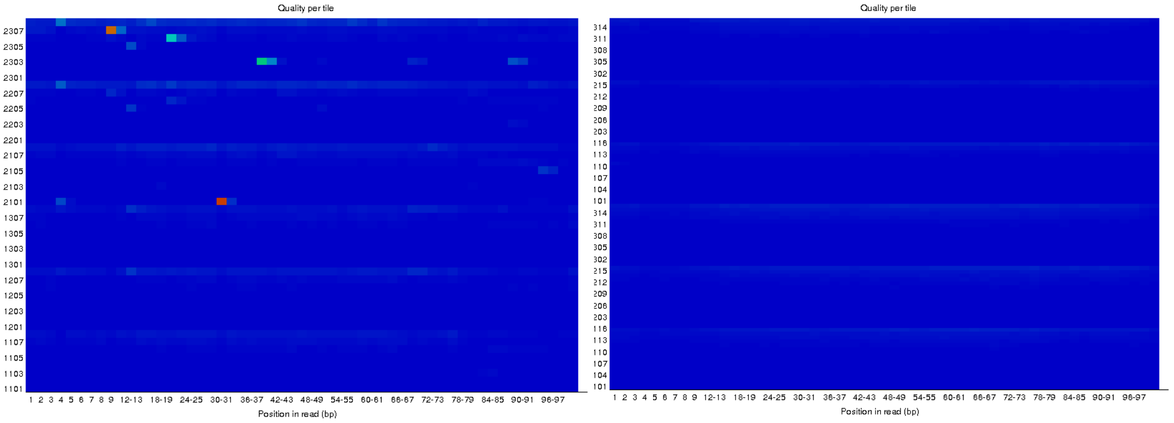


Figure 26 - Comparison of per tile sequence quality FASTQC step of a bad sample (SRR2034989, left) and a good sample (SRR2173864, right).

In the case of base sequence content, the proportion of each base, for which each of the normal DNA bases has been called, is presented in a plot. It would be expected that the lines in this plot run parallel with each other reflecting no major difference between the different bases of a sequence run. However, some types of library will always produce biased sequence composition as bisulfite converted sequences, since most of cytosines were converted to thymines yielding an excel of thymines in the plot. In the case of Figure 27, FASTQC rejects the sample based on the 20% greater amount of thymines and adenines. The warnings presented in the per sequence length distribution (Figure 27, right) are related to sequences that possibly are not of the same length.

In Figure 28 we present a normal distribution of GC content (Figure 28, right) and a bad one (Figure 28, left). The sequence GC content measures the GC content across the whole length of each sequence in a file and compares it to a modelled normal distribution of GC content made from the observed data and used to build a reference distribution. The second peak observed in the graphs corresponds to the overall GC content of the underlying sample. Any unusually shaped distribution could indicate a contaminated library since the sum of the deviations from the normal distribution represents more than 30% of the reads.

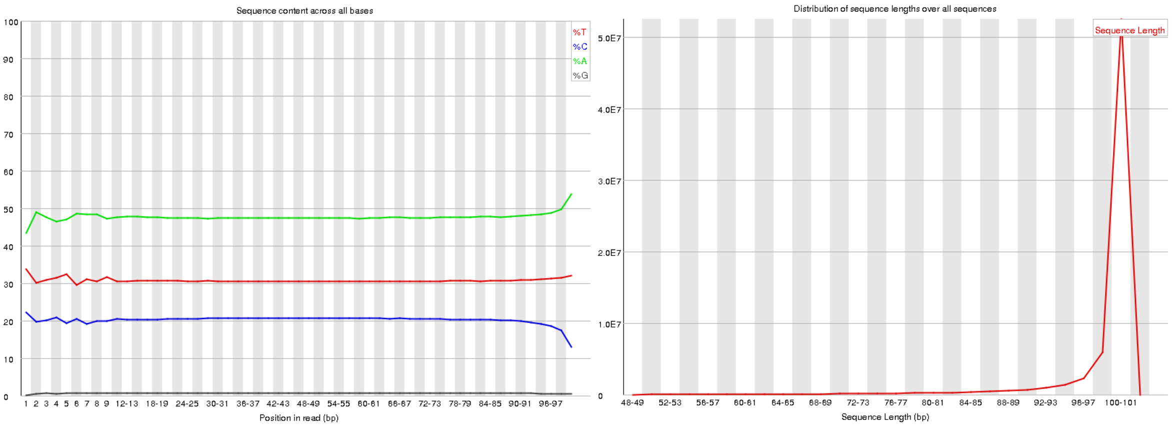


Figure 27 – Quality control per base sequence content and sequence length distribution of SRR5115679 sample. The sample revealed a bad quality in per base sequence content and several warnings in sequence length distributions.

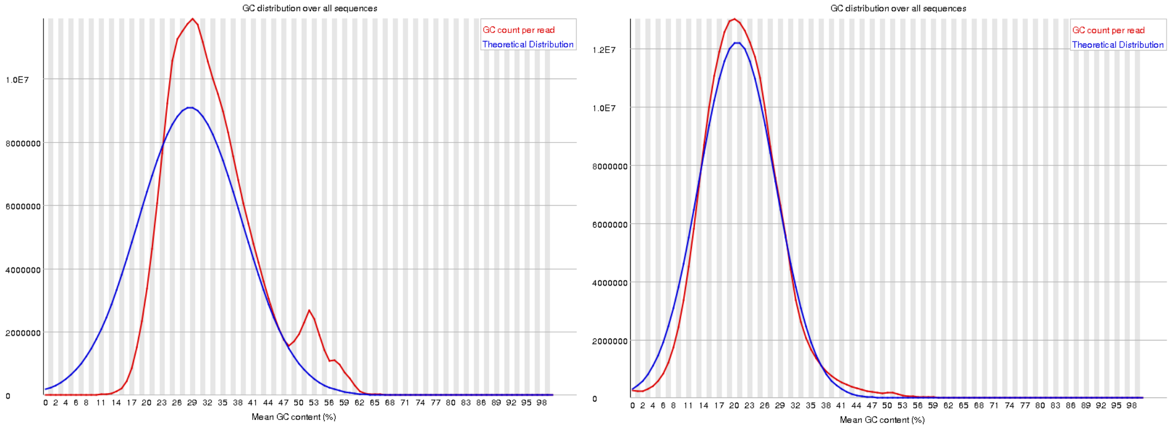


Figure 28 – Comparison of per sequence GC content FASTQC step of a bad sample (SRR2034989, left) and a good sample (SRR5115685, right).

Additionally, the sequence duplication levels represents the degree of duplication for every sequence in a library. While a low level of duplication may indicate a very high level of coverage of the target sequence, a high level is more likely to indicate some kind of enrichment bias. According to the literature, contaminants will tend to produce spikes towards the right of the plot and if peaks appear in the blue trace there should be a large number of different highly duplicated sequences which might indicate either a contaminant set or a severe technical duplication, as we can see in Figure 29 with the failure of SRR2034993, caused by more than 50% of non-unique sequences.

To evaluate call quality of bases, the percentage of FASTQC plots positions for which an N was called (Figure 30). This happens because when a sequencer is unable to make a base call with sufficient confidence, then it will call it an N. FASTQC also builts Kmer profiles, which report possible overrepresentated sequences in samples. In the figure, we show a bad sample with this respect, since six most biased Kmers are represented by sharp spikes of enrichment at a single point in the sequence, rather than a progressive or broad enrichment.

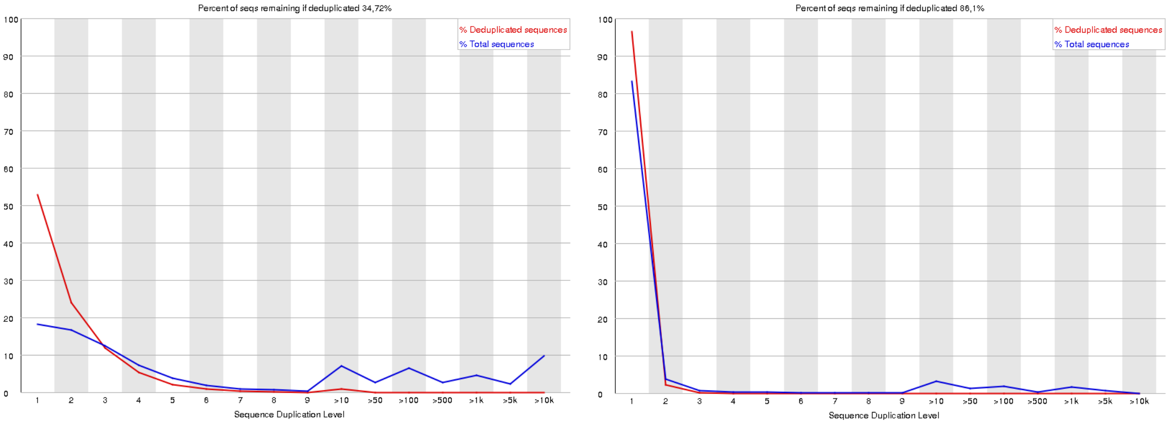


Figure 29 – Comparison of sequence duplication levels FASTQC step of a bad sample (SRR2034993, left) and a good sample (SRR5115679, right)

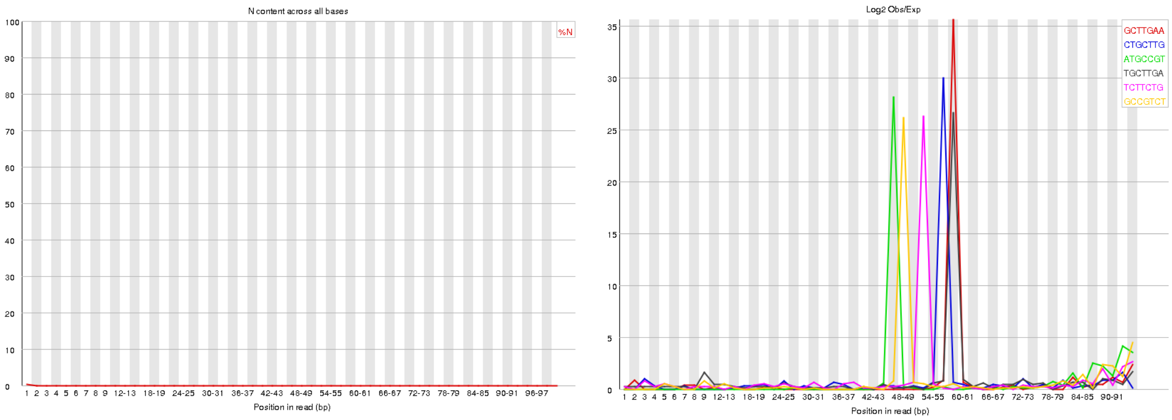


Figure 30 - Quality control per base N content (left) and Kmer content (right) of SRR2173864 sample. The sample reveals a good quality in per N content and bad quality in Kmer content.

3.1.3. Alignment step

After the alignment step using Bismark Bisulfite Mapper, we obtained the results presented in Table 8 with a really low mapping efficiency. Therefore, through this results we concluded that the alignment of reads to the reference genome had not a good quality, which could probably be caused by the several problems detected in the FASTQC step. For all those reasons, we could not use these samples to study the T-DMRs across age, since differences reported between the samples could be caused by a bad quality of samples or alignment instead of methylation differences.

Table 8 – Results obtained from the Bismark bisulfite mapper with mapping efficiency and the methylated cytosines detected in the several contexts.

	Mapping efficiency (%)	*C methylated in CpG context (%)*	*C methylated in CHG context (%)*	*C methylated in CHH context (%)*	*C methylated in unknown context (CN or CHN) (%)*
SRR2034989	40,30	72,60%	8,10%	11,50%	5,00%
SRR2034993	43,4	76,4	13,9	19,8	6,5
SRR2079727	0	0	0	0	0
SRR5115679	75	75,7	0,4	0,5	1,1
SRR5115685	76,4	75,9	0,3	0,3	0,8

3.2. Dataset from *Homo sapiens*

After the first filtering step using GEOmetadb we obtained 2103 samples from healthy individuals, together with age and gender information. However, due to the need for raw data to proceed to the bioinformatics protocol, the previous number of samples was reduced to 1703, as summarized in Table 9.

Since, data from Illumina BeadChip 450k was prevalent, representing 97% of the already filtered total, the following plots only focused on the GPL13534 platform and this was the only one being used in our analysis. This resulted in 27 series and 1650 samples from 12 different tissues as distributed in Figure 31. Visibly, blood was the overrepresented sample type with 1334 samples, followed by buccal, lung and brain cells.

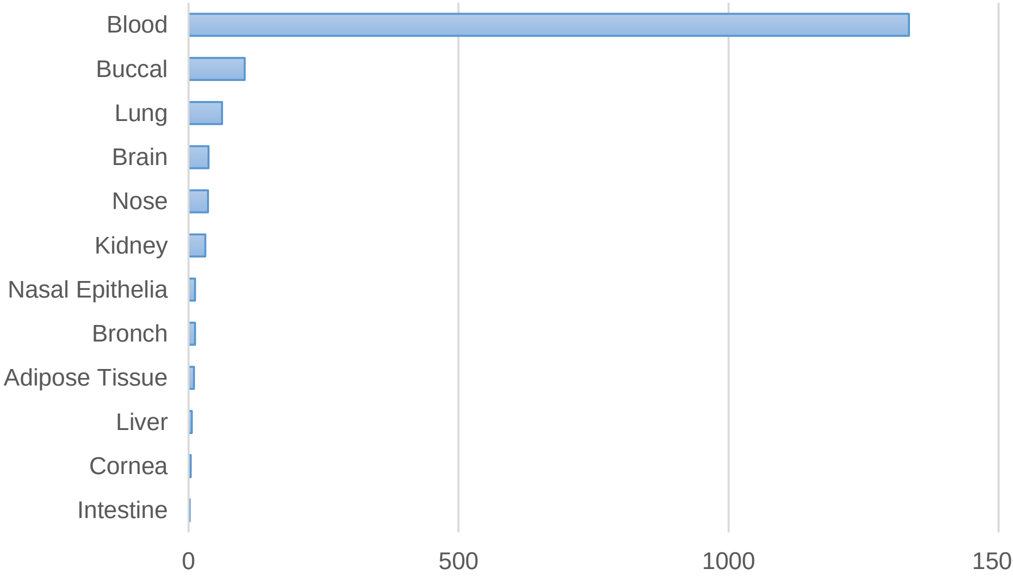


Figure 31 – Distribution of tissues among Dataset B.2 with a prevalent advantage of blood with 1333 samples, followed by buccal cells, lung and brain.

Table 9 – Dataset B.1 obtained for *Homo sapiens* and microarray search with information about GEO experiment, superseries, article PMID, number of samples across gender, age interval, tissue type and platform used

Serie	*Superserie*	*PMID*	*No Samples*	*No. Males*	*No. Females*	*Age*	*Tissue*	*Tissue-specificity*	*Cell-specificity*	*Technology*	*Platform*
GSE99755	N/A	29362489	37	18	19	24-58	Blood	Whole blood	N/A	Illumina 450k	GPL13534
GSE98876	N/A	28747766	5	5	0	31-64	Blood	Peripheral blood	PBMC (CD3 T cell)	Illumina 450k	GPL13534
GSE98203	N/A	28556790	9	7	2	24-53	Brain	Neurons	N/A	Illumina 450k	GPL13534
GSE94462	N/A	N/A	4	2	2	46-75	Cornea	Endothelium tissue	N/A	Illumina 450k	GPL13534
GSE88940	N/A	N/A	10	0	10	11-19	Adipos	Visceral	N/A	Illumina 450k	GPL13534
GSE87648	GSE87650	27886173	92	44	48	18-69	Blood	Whole blood	N/A	Illumina 450k	GPL13534
GSE87640	GSE87650	27886173	18	12	6	24-58	Blood	Whole blood	N/A	Illumina 450k	GPL13534
		N/A	16	0	0	24-58	Blood	Peripheral blood	PBMC (CD4)	Illumina 450k	GPL13534
		N/A	15	0	0	24-58	Blood	Peripheral blood	PBMC (CD8)	Illumina 450k	GPL13534
		N/A	16	0	0	24-58	Blood	Peripheral blood	PBMC (CD14)	Illumina 450k	GPL13534
GSE87571	N/A	23826282	732	341	389	14-94	Blood	Whole blood	N/A	Illumina 450k	GPL13534
GSE85647	GSE85649	28549776	6	0	6	23-52	Blood	Peripheral blood	PBMC (CD14)	Illumina 450k	GPL13534
GSE85566	GSE85568	27942592	42	15	27	19-59	Lung	Airway epithelial	N/A	Illumina 450k	GPL13534
GSE85506	N/A	N/A	21	0	21	19-80	Blood	Peripheral blood	N/A	Illumina 450k	GPL13534
GSE84395	N/A	N/A	16	12	4	28-79	Lung	Pulmonary cells	N/A	Illumina 450k	GPL13534
GSE80261	N/A	27358653	104	44	60	5-18	Buccal	Epithelial cells	N/A	Illumina 450k	GPL13534
GSE79100	N/A	N/A	31	15	16	18-78	Kidney	N/A	N/A	Illumina 450k	GPL13534
GSE71955	GSE71957	26459776	31	0	0	35-79	Blood	Peripheral blood	PBMC (CD4)	Illumina 450k	GPL13534
		N/A	31	0	0	35-79	Blood	Peripheral blood	PBMC (CD8)	Illumina 450k	GPL13534
GSE67484	GSE67485	N/A	2	2	0	64	Liver	N/A	N/A	Illumina 450k	GPL13534
		N/A	2	2	0	25	Intstine	Small Intestine	N/A	Illumina 450k	GPL13534
GSE65163	GSE65205	N/A	36	17	19	9-12	Nose	Nasal Epithelium	N/A	Illumina 450k	GPL13534
GSE63179	N/A	25706862	4	4	0	25	Brain	Cerebellum	N/A	Illumina 450k	GPL13534
GSE61107	N/A	24399042	24	19	5	53-90	Brain	Frontal cortex	N/A	Illumina 450k	GPL13534
GSE51057	N/A	24278132	177	0	177	34-65	Blood	Peripheral blood	Leukocytes	Illumina 450k	GPL13534
GSE43091	GSE43273	24735922	4	0	4	37-71	Liver	N/A	N/A	Illumina 450k	GPL13534
GSE42861	N/A	23334450	76	16	60	24-70	Blood	Peripheral blood	Leukocytes	Illumina 450k	GPL13534
GSE32396	N/A	22346766	15	0	15	50-80	Blood	Whole blood	White cells	Illumina 27k	GPL8490
GSE32393	N/A	22346766	23	0	23	19-75	Breast	N/A	N/A	Illumina 27k	GPL8490
GSE30759	GSE30760	22453030	15	0	15	33-69	Uterus	Cervic	N/A	Illumina 27k	GPL8490
GSE107737	N/A	N/A	12	12	0	18-29	Blood	Whole blood	N/A	Illumina 450k	GPL13534
GSE107226	N/A	N/A	4	0	0	47-59	Lung	Fibroblast	N/A	Illumina 450k	GPL13534
GSE105123	GSE105124	PMC39623	19	11	8	19-23	Blood	Peripheral blood	PBMC	Illumina 450k	GPL13534
GSE104471	GSE104472	28294656	12	6	6	24-45	Bronch	Bronchial epithelia	N/A	Illumina 450k	GPL13534
		N/A	12	6	6	24-45	Nasal		N/A	Illumina 450k	GPL13534
		N/A	12	6	6	24-45	Blood	Peripheral blood	PBMC	Illumina 450k	GPL13534
GSE102177	N/A	N/A	18	10	8	4-14	Blood	Peripheral blood	N/A	Illumina 450k	GPL13534

The blood samples consisted of two types of blood cell populations – whole blood and peripheral blood. Between the samples from peripheral blood we got peripheral blood mononuclear cells (PBMCs), undefined leukocytes or undefined peripheral blood samples, with a visible advantage of PBMCs with 17% of the samples (Figure 32).

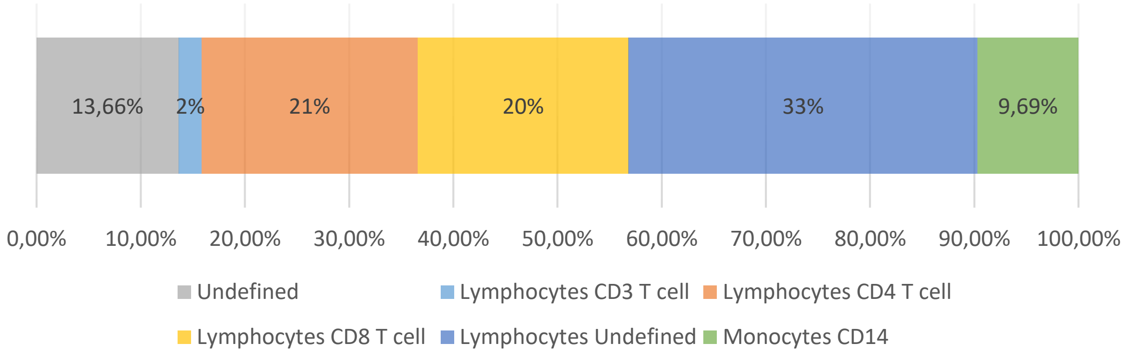


Figure 32 – Representation of different types of PBMCs present in Dataset B.2 with a biggest prevalence of lymphocytes.

Additionally, it is known that PBMCs include several types of mononuclear cells of peripheral blood like lymphocytes and monocytes. In our dataset we obtained both types of PBMCs and a small contribution of undefined PBMCs. Between lymphocytes our dataset had a prevalent presence of undefined lymphocytes, followed by lymphocytes CD4 T cells, also known ad T helper cells (Th), lymphocytes CD8 T cells, also known as cytotoxic T cells (Tc), and CD3 T cells. Lastly, we also had 10% of monocytes CD14 (Figure 33).

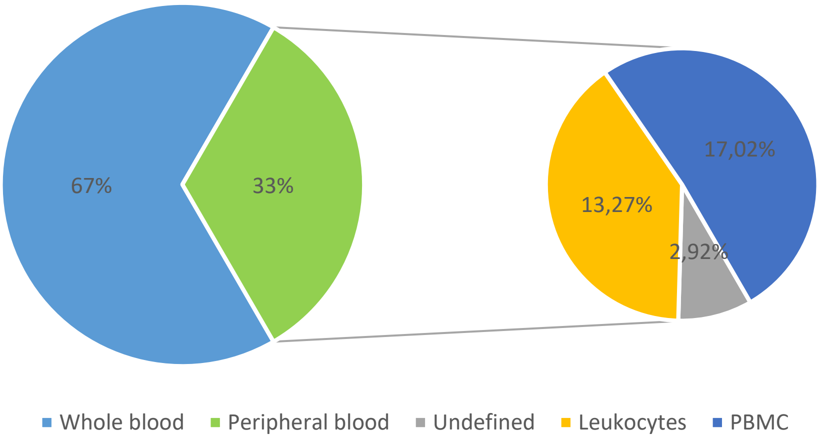


Figure 33 – Blood specificity across Dataset B.2 with a prevalence of whole blood samples (67%). The peripheral blood samples included undefined samples, leukocytes and PBMCs.

The collected data had also a wide variety of individual ages with a prevalence of individuals with ages between 41 and 60 years as shown in Figure 34. However, as referred before it is known that the methylome has a big importance in the early development of an individual and several appreciations have been made to methylome in childhood and adolescence that are not related with aging. Therefore, the samples with ages below 18 were eliminated from our dataset. In the end, our dataset was ready for further studies and its full details are presented in Supplementary Table 3 (Dataset B.2).

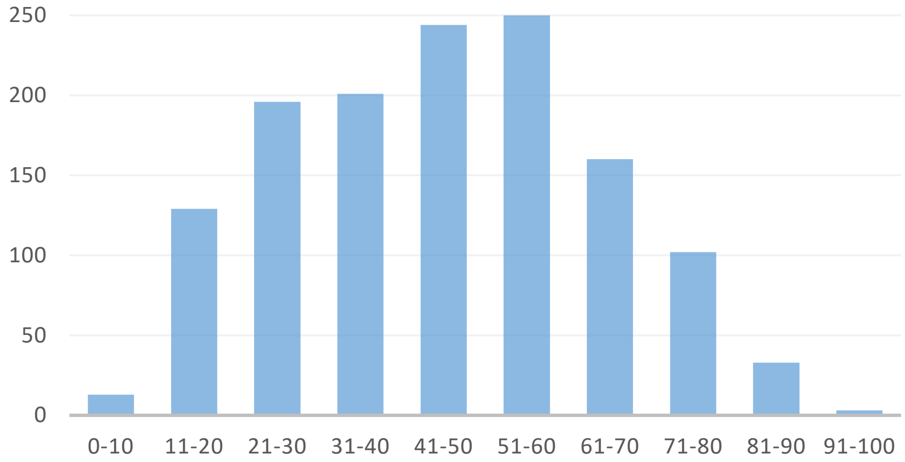


Figure 34 – Age distribution on dataset with a prevalence of individuals between 41 and 60 years old and low amount of individuals with 81-100 and 0-10 years old.

4. Discussion

In silico experiments can be used in genomics to improve data throughput, especially in experiments that usually would be expensive and prolongated. The use of public databases and data mining to recicle datasets already used by biological scientists for the same or different purposes is essential in this process. Additionally, wanted to use age-specific *Mus musculus* samples as a biological model, since it can be taken as representative of human methylome has already been reported. Therefore, our research started looking for a dataset that would be explored and used as a basis to do targeted bisulfite sequencing studies with aged murines at iBiMED.

The murine dataset initially consisted of 64 samples from healthy mice. These excluded those samples from embryo, embryonic stem cells, cell lines and were only from bisulfite-seq. However, due to the need of an age time course inside the selected tissue, the research focused only on 6 samples from pancreas and liver those with the highest number of final reads after filtration with Pipeline Pilot, although with variances in the used reference genomes and technologies. The preprocessing steps associated with NGS technologies were developed, starting with a filtering step followed by a quality control and a last filtering step that filtered the unpaired reads. Since the samples had a general good quality, they were ready to proceed to the alignment step, which started with MethyPipe but later changed to Bismark, due to the above mentioned limitations of the previous software. In this last step, it was concluded that MethyPipe could not be used with samples that were human, since the software has an incorporation of the alignment genome and this feature cannot be changed. For this reason, Bismark Bisulfite Mapper was used to map the reads and a low efficiency (<50%) was obtained for three samples and only two samples had a better alignment efficiency (≈70%).

Additionally, our Dataset A didn't include samples from the same sequencing platform and the number of available samples could not provide a relevant statistical analysis of methylation across tissues so, it was useless to continue with this approach. Previous work in the Institute also recommended the usage of a significant amount of human samples that would improve the statistical relevance of the study, suggesting microarray technologies as a potential escape from NGS expenses and lack of available data.(Cluny 2016) Through a database search it was concluded that the microarray approach had a more prevalent contribution than NGS for past methylation studies and that NGS data was distributed by a lot of library preparation protocols that could not be compared between each other.

Therefore, in view of the advantages and disadvantages stated above, we concluded that the usage of *Homo sapiens* and microarray as keywords for our search would be the best approach to study DNA methylation variation across age and tissue. Furthermore, this stratedy is also supported by the fact that in the future the institute aims to study the methylome across age through young and old people using microarrays. In the new dataset, the diversity of tissues and age ranges were appreciable and the usage of the same platform and DNA treatment techniques was also a remark. Globally, the aim of our search could be reached and a wide methylation map could be constructed.

In the Dataset B.2 from *Homo sapiens*, we obtained 1650 samples from the most widely used microarray chip – Illumina Methylation 450k - to study human methylation with a huge diversity in tissue types. However, given the general limitation in obtaining large number of samples from all tissues, blood has already been reported as an attractive, easy and available source of DNA.(Reinius et al. 2012) According to several studies, it is known that alteration in DNA methylation patterns can be detected in the blood of patients with diseases or even solid tumours samples.(Reinius et al. 2012) Therefore, our study was focused on blood as the primary tissue for methylation variations to be used routinely at iBiMED. However, the tissue-specific cell variation in our dataset was visible since it integrated 887 whole blood samples and 443 peripheral blood ones. Between them 17% were PBMCs distributed in lymphocytes of several types (76%), monocytes (10%) and cells without any specification (14%).

According to literature, cell heterogeneity among blood samples may act as a confounder when measuring DNA methylation in whole blood. It was already reported that there are differentially methylated regions between the several purified cell populations of blood and caution should be taken in the interpretation of whole blood results particularly for immune-related genes. The comparison between granulocytes and monocytes has already revealed some cell-specific differentially methylated sites. To diminuish this problem, alghorithms have been developed to use cell-type-specific positions to determine the relative amount of each cell type per sample.

Elsewhere, it is also known that CpG islands and 5'UTR – CpG sites in regions of high density - are more often unmethylated, while CpG sites located in introns, 3'UTRs and repetitive elements were methylated. Lastly, the existence of sex-specific methylation patterns have also been studied. This patterns are present not only in sex chromosomes but also in autosomal ones – it is known that global and autosomal CpG methylation has a tendency for higher methylation in males and sex-specific differences at varying numbers of CpG probes, across different chromosomes. The X-chromossome inactivation is accompanied with widespread CpG hypermethylation but sex-specific methylation has also been shown to be modified by sex hormones and there are already several DMRs that are known to exist in autosomal chromosomes, related with sex variances. Although the cross-reacting probes are excluded in the bioinformatics protocols, there is evidence that the number of sex-specific DMRs is still high.

Therefore, reviewing our dataset, we can expect a biased distribution of samples according to cell types, genders or genome regions effects of methylation. For all of these reasons, it is essential to meticulously compare our samples: to remove probes theoretically related with these features and to analyse our clusters carefully in order to see tendencies that could introduce non-age related biases.

We should also stress that, the lack of information or the incorrectly insert of information by the researchers when submitting their own data was detrimental and clearly impacted the timeline of our work. This forced us to perform a manual study for each sample which was even harder due to the lack of a userfriendly platform displayed by NCBI. Therefore, in the near future it is essential to create an in-house database, to store this data, now that is was completely verified, to improve the reuse it in other studies or even to improve the pipeline used to explore the NCBI dataset. Through this progress, the sustentability of *in silico* experiments at iBiMED should be highly improved.

CHAPTER III DATA ANALYSIS

Methylomics of aging using public datasets

1. INTRODUCTION

The main goal of this chapter was to use public data to find differential methylated regions that could be associated with the healthy aging of individuals. For that purpose we adopted an in silico concept using minfi. *minfi* is a software developed in 2014 for the analysis of Illumina Infinium methylation arrays, particularly the Illumina Infinium HumanMethylation 450 BeadArray(Aryee et al. 2014). This software has already been used in human methylomic studies and has been referred as a good choice for the several steps comprised in the microarray bioinformatics pipeline(Wright et al. 2016)'(Morris & Beck 2015). *minfi* separates the annotation step – genomic location of methylation loci and nearby features – from array design interpretation - how probes are matched with relevant color channels to produce Meth or Unmeth signals(Aryee et al. 2014). With this particular feature, the annotation process can be updated using, for example, later human genome builts(Aryee et al. 2014).

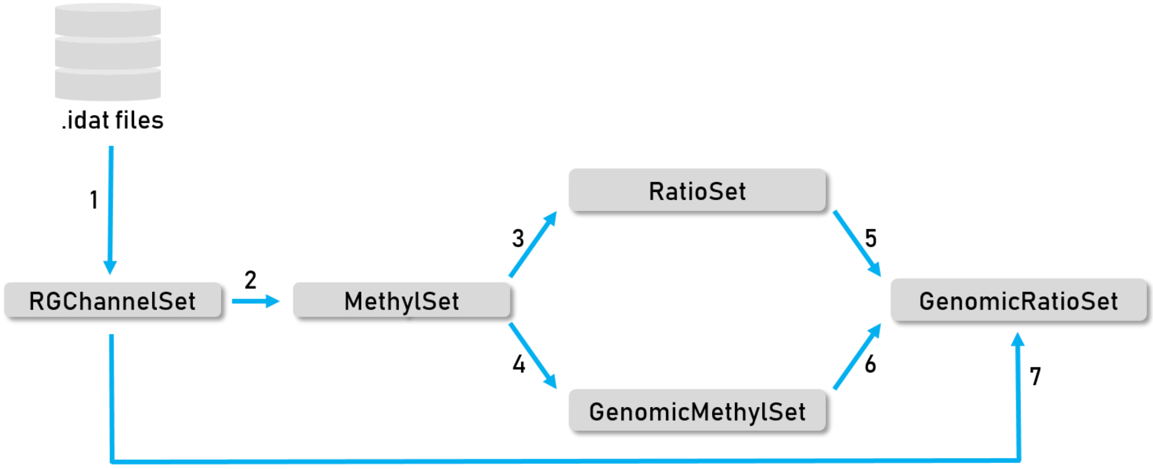


Figure 35 – Organization of *minfi* in several computer classes as used in the several steps of the pipeline. The .idat files are the starting point of the analysis and are followed by the usage of several functions. (1) read.450k.exp(); (2) preprocessRaw(), preprocessSWAN() or preprocessIllumina(); (3) ratioConvert(); (4) mapToGenome(); (5) mapToGenome (); (6) ratioConvert(); (7) preprocessQuantile() or preprocessFunnorm().

The software is organized in several computer classes for the several steps of the microarray pipeline (Aryee et al. 2014). The process starts with .idat files that are read into an RGChannelSet, a class that contains the raw intensities as two matrices, with the red and the green channel, the intensities of the internal control probes and a manifest object with the probe design information of the array(Fortin & Hansen 2016). Once these data are processed into methylation measurements, they can be stored in four additional classes representing several stages of preprocessed data: MethylSet, GenomicMethylSet, RatioSet and GenomicRatioSet (Figure 35)(Aryee et al. 2014).

`MethylSet` is the result of the preprocessing step and contains normalized data and a matrice with the methylated and unmethylated signals.(Hansen 2018) Using the accessors `getMeth` and `getUnmeth`, both methylated and unmethylated matrices can be obtained(Hansen 2018). On the other hand, `RatioSet` is a class designed to store β-values and/or M-values, concepts defined in chapter 1 section 4, and is irreversible, which means it is not possible to retrieve the methylated and unmethylated signals from a `RatioSet`(Fortin & Hansen 2016). The `RatioSet` is obtained from `MethylSet` using the `ratioConvert` function(Aryee et al. 2014). In both cases, the genomic prefix in the class name indicates that methylation loci have been associated with a genomic location which is a nonreversible transformation, as it entails choosing a reference genome and discarding unmapped probes(Aryee et al. 2014). This operation is made using the function `mapToGenome`, which allows the user to choose a human genome build(Aryee et al. 2014).

To start data processing, *minfi* provides several plots for quality control check of the data, such as density plots or control probe plots(Hansen & Aryee 2012). The simplest quality control plot uses the log median intensity in both methylated and unmethylated intensities and when plotting the two medians against each other, good samples will cluster together while failed ones will tend to separate and have lower median intensities(Fortin & Hansen 2016). This data can be explored in order to look at the β-value densities of the samples and even to compare all plots at the same time in an interactive manner. On the other hand, the control probes plot allows plotting individual control probe types. There are at least 9 types of control probes (such as: staining, hybridization, extension, target removals, bisulfite conversion, specificity, non-polymorphic and norm probes) integrated in 450k array. Through the control probes plot, the researcher can evaluate and quality control the several steps in sample preparation microarray process. (Fortin & Hansen 2016)'(Hansen & Aryee 2012).

Then, a preprocessing or normalization step is carried out according to the chosen method(Fortin & Hansen 2016). The `preprocessRaw` option does not perform any normalization, i.e. it uses `RGChannelSet` as input and `MethylSet` as output(Fortin & Hansen 2016). The `preprocessIllumina` implements the preprocessing choices as available in Genome Studio, making a background substraction, a control normalization and using `RGChannelSet` as input and `MethylSet` as output(Fortin & Hansen 2016). On the other hand, the `preprocessSWAN` performs subset-quantile within array normalization

(SWAN) that corrects the differences between the Type I and Type II probes by applying a within-array quantile normalization separately for different subsets of probes(Fortin & Hansen 2016). The input used in SWAN is `RGChannelSet` or `MethylSet` and the output is `MethylSet`(Fortin & Hansen 2016). The `preprocessQuantile`, with `RGChannelSet` as input and `GenomicRatioSet` as output, implements the stratified quantile normalization preprocessing explained before, but it is not recommended for datasets where global changes are expected(Fortin & Hansen 2016). Another option is `preprocessNoob`, with `RGChannelSet` as input and `MethySet` as output, which implements the noob background subtraction method with dye-bias normalization(Fortin & Hansen 2016). The background noise is estimated from the out-of-band probes and removed for each sample while the dye-bias normalization uses a subset of control probes to estimate the dye bias(Fortin & Hansen 2016). Lastly, the already described `preprocessFunnorm` that has as input `RGChannelSet` and as output `GenomicRatioSet` and that is particularly useful for studies that compare conditions with known large-scale differences, like between-tissue studies(Fortin & Hansen 2016). The function applies the `preprocessNoob` function as a first step for background subtraction and then uses the first two principal components of the control probes to explore any unwanted variation(Fortin & Hansen 2016). After the normalization process, *minfi* still offers a way to correct batch effects correction SVA, removing probes with known SNPs associated with the same CpG site and cell- estimation in case of complex samples.

Searching DMPs and DMRs is the next step(Fortin & Hansen 2016). The identification of DMPs by `dmpFinder` function reveals differentially methylated positions between two or more sample groups using an F-test(Hansen & Aryee 2012). Then, to find DMRs the `bumphunter` function is used, samples are firstly clustered of probes. To make clusters, candidate regions need to be tested for significance and for that the algorithm uses permutations(Fortin & Hansen 2016). However, since bump hunting focuses on methylation changes around gene promoters, it is necessary to use a block finding step to find long-range alterations(Aryee et al. 2014). The function `blockFinder` groups the average methylation values in open-sea probe cluster into larger regions and the bump hunting process can then be applied with a large smoothing window(Aryee et al. 2014).

Finally, after data processing using *minfi*, the creation of interactive visualizations for genomic scale data is essential. This can be offered by several visualizers that have an integration with *minfi* through R like ggplot2, Gviz and epiviz or even by importation of data into UCSC Genome Browser or IGV (Morris & Beck 2015).

2. Methodologies Adopted

Dataset B.2 was analysed with *minfi* Bioconductor package version 1.26.0 (Aryee et al. 2014) in RStudio 64 bits version 3.4.3. The relevant functions used in our pipeline that were not included in *minfi* package are described on Table 10. The pipeline followed was adapted from (Hansen 2018; Fortin & Hansen 2016),(Maksimovic & Phipson 2015) and was divided into two steps: (1) an analysis of each individual experiment of the dataset; and (2) a global analysis joining all experiments.

Table 10 – List of used functions, its description and respective packages, excluding the *minfi* or R incorporated functions.

Library	*Useful functions*	*Function description*	*References*
fastcluster	hclust	Hierarchical agglomerative clustering	(Müllner 2013)
DMRcate	rmSNPandCH	Filters a matrix of M/-values by distance to SNP	(Peters et al. 2015)
	cpg.annotate	Annotates a matrix of M-values with probe weights and chromosomal position	
	extractRanges	Takes a dmrcate.output object and produces the corresponding GRanges object	
	DMR.plot	Plots an individual DMR as found by dmrcate	
qqman	manhattan	Creates a manhattan plot from PLINK assoc output	(Turner 2014)

After obtaining data from `GEOquery` library, as described in Chapter II, the idat files needed to be decompressed since *minfi* does not support reading compressed idat files. The idat files were then read, using the function `read.metharray.exp`, into the class `RGChannelSet` that will be used in the future when we want to refer to the raw data, and the data is accessed from a data sheet experiment using the command `pData`. Since the phenotype data comprises much information, we needed to simplify and reduce it by focusing on the most relevant information for our analysis. For this, the table was filtered and only the data was integrated into the methylation data. The `RGChannelSet` also stores a manifest object that contains the probe design information of the array, obtained through the function `getManifest`. (Box 4)

Box 4 – Script of the manipulation of input data of Dataset B.2 starting with the import of all the libraries used in the entire script

```
library(minfi)
library(fastcluster)
library(IlluminaHumanMethylation450kmanifest)
library(IlluminaHumanMethylation450kanno.ilmn12.hg19)
library(RColorBrewer)
library(limma)
library (qqman)
ann450k = getAnnotation(IlluminaHumanMethylation450kanno.ilmn12.hg19)


#Input data and manipulation
idatFiles <- list.files("GSE105123/idat", pattern = "idat.gz$", full = TRUE)
```

```
sapply(idatFiles, gunzip, overwrite = TRUE)
rgSet <- read.metharray.exp("GSE105123/idat")
source("geo_data.R")
geoMat <- getGEO("GSE105123")
pD.all <- pData(geoMat[[1]])
pD <- pD.all[, c("title", "geo_accession", "characteristics_ch1",
"characteristics_ch1.1", "characteristics_ch1.2", "characteristics_ch1.3")]
head(pD)
names(pD)[c(3,4)] <- c("sex", "age")
pD$sex <- sub("gender: ", "", pD$sex)
pD$age <- sub("age: ", "", pD$age)
sampleNames(rgSet) <-  sub("(GSM\\d+)_.*", "\\1", sampleNames(rgSet))
pD <- pD[sampleNames(rgSet),]
pD<-as(pD,"DataFrame")
pData(rgSet) <- pD
phenodata<-pData(rgSet) #info
manifest <- getManifest(rgSet) #info
head(getProbeInfo(manifest)) #info

#save the manipulation data into R object
saveRDS(rgSet, file = "rgSet_GSE105123.rds")
saveRDS(phenodata, file = "pheno_GSE105123.rds")
```

2.1. Quality control

In the quality control step (Box 5) we performed by a clustering using `hclust` function that worked through the calculation of raw data β-values (`getBeta` function). The P-value was calculated using the function `detectionP` and the mean of p-values inside the same sample was represented in a boxplot. Then, we carried out a quality control report that plotted the most common analysis available at *minfi* studying samples by age and gender and building a control probes plot. Finally, we checked the global quality of our samples making a QC plot using the functions `getQC` and `plotQC` and a boxplot of the difference between unmethylated and methylated channels.

Box 5 – Script of the quality control step in Dataset B.2 including a clustering, calculation of mean p-values and quality control reports with a wide variety of integrated plots.

```
#calculate the beta values with raw data and clustering the samples
beta <- getBeta(rgSet, type="Illumina")
d <- dist(t(beta),method="euclidean")
fit <- hclust(d, method="complete")
pdf("clustering.pdf", onefile=T, paper="a4r")
plot(fit, cex = 0.8)
dev.off()

#calculate the mean P-values across all samples to identify any failed samples
pal = colorRampPalette(c('cadetblue3'))(50)
detP = detectionP(rgSet)
head(detP)
pdf("p_values.pdf")
barplot(colMeans(detP),col=pal[factor(phenodata$geo_accession)],las=2,cex.names=0.8
, main="Mean detection p-values", ylim=c(0,0.0009))
dev.off()


#qc_report by group before filtering and normalization
```

```
qcReport(rgSet,sampNames=phenodata$geo_accession,sampGroups=phenodata$age,pdf="qcRe
port_Age.pdf")
qcReport(rgSet,sampNames=phenodata$geo_accession,sampGroups=phenodata$sex,pdf="qcRe
port_Gender.pdf")
raw <- preprocessRaw(rgSet)
meth <- minfi::getMeth(raw)
dim(meth)
qc <- getQC(raw)
pdf("raw.pdf")
plotQC(qc)
dev.off()
```

2.2. Preprocessing and normalization

The preprocessing and normalization step (Box 6) started with a comparison of the four methods available in *minfi* using the `preprocessRaw`, `preprocessSWAN`, `preprocessQuantile` and `preprocessFunnorm` functions and consequent density plotting with variations of age or gender. Due to the origin of our samples, the data is also evaluated using a normalization by cell type. After the comparison of all methods and selection of the Quantile normalization method, the data was compared before and after normalization in the β-value variations using density plots.

Box 6 – Script of the normalization step starting by a comparison of all available normalization methods, followed by a comparison of the data before and after normalization.

```
#make different normalization and see the differences between them
mSetRaw = preprocessRaw(rgSet) #raw method
mSetSw = preprocessSWAN(rgSet = rgSet, mSet = mSetRaw, verbose=TRUE) #swan method
mSetSq = preprocessQuantile(rgSet) #quantile method
funSq = preprocessFunnorm(rgSet, bgCorr = TRUE, dyeCorr = TRUE) #funnorm method
cells<-estimateCellCount(rgSet, compositionCells="Blood", returnAll=TRUE)

#differences between normal quantile and quantile with type cells normalization
corr.test(getBeta(mSetSq), getBeta(cells$normalized))
plot(getBeta(mSetSq), getBeta(cells$normalized))

#plotting the differences of normalization method - gender variable
pdf("densityplots_sex.pdf")
par(mfrow=c(1,3))
densityPlot(rgSet, sampGroups = phenodata$Gender, main="Raw")
densityPlot(getBeta(mSetSw), sampGroups = phenodata$sex,main="SWAN")
densityPlot(getBeta(mSetSq), sampGroups = phenodata$sex,main="Quantile")
densityPlot(getBeta(funSq), sampGroups = phenodata$sex,main="FunNorm")
dev.off()

#plotting the differences of normalization method - age variable
pdf("densityplots_age.pdf")
par(mfrow=c(1,3))
densityPlot(rgSet, sampGroups = phenodata$age, main="Raw", legend=FALSE)
densityPlot(getBeta(mSetSw), sampGroups = phenodata$age,main="SWAN", legend=FALSE)
densityPlot(getBeta(mSetSq), sampGroups = phenodata$age,main="Quantile",
legend=FALSE)
densityPlot(getBeta(funSq), sampGroups = phenodata$age,main="FunNorm",
legend=FALSE)
dev.off()
```

```
#visualize what the data looks like before and after normalization - age variable
pdf("densityplots_before_after_norm_rawvsquantile.pdf")
par(mfrow=c(1,2))
densityPlot(rgSet, sampGroups=phenodata$age,main="Raw", legend=FALSE)
densityPlot(getBeta(mSetSq), sampGroups = phenodata$age,main="Quantile",
legend=FALSE)
dev.off()
```

2.3. Filtering

Before the beginning of the filtering step, a multidimensional scaling plot was made to look at the level of similarity of individual samples across the dataset and to identify the main sources of variation inside our dataset across gender and age. (Box 7) Then, probes with P-value above 0.01, probes from the X or Y chromosomes and the ones with common SNPs at CpG sites were removed. Finally the probes that were shown to be mapped to multiple places in the genome were also removed. The filters used were based on those recommended in literature.(Pidsley et al. 2016) After the filtering steps, the dendograms and MDS plots were repeated and also the clustering of samples in order to reanalyze the main sources of variation inside our dataset and proceed to the next step.

Box 7 – Script of the filtering step starting and ending with a MDS plot in order to evaluate the major sources of variation in our dataset.

```
#MDS plots to look at largets sources of variation - age
pdf("MDSplot_quantile.pdf")
par(mfrow=c(1,2))
colfunc <- colorRampPalette(c("royalblue", "white"))
colorgrad<-colfunc(length(sort(unique(pData(mSetSq)$age))))
zzz<-as.factor(pData(mSetSq)$age)
plotMDS(getM(mSetSq), top=1000, gene.selection = "common", col=colorgrad[zzz],
pch=16)

#MDS plots to look at largets sources of variation - gender
pal = colorRampPalette(c('darksalmon', "cadetblue3"))(2)
plotMDS(getM(mSetSq), top=1000, gene.selection = "common", col =
pal[factor(phenodata$sex)], pch = 16)
dev.off()

#Filtering step
#remove any probes that have failed in one or more samples
detP = detP[match(featureNames(mSetSq),rownames(detP)),] #ensure probes are in the
same order in the mSetSq and detP objects
keep = rowSums(detP < 0.01) == ncol(mSetSq) #By default detection P-values with a
value >0.01 are set to NA
table(keep)
mSetSqFlt = mSetSq[keep,]
mSetSqFlt

#if your data includes males and females, remove the sex chromosomes
ann450k = getAnnotation(IlluminaHumanMethylation450kanno.ilmn12.hg19)
keep = !(featureNames(mSetSqFlt) %in% ann450k$Name[ann450k$chr %in%
c("chrX","chrY")])
table(keep)
mSetSqFlt = mSetSqFlt[keep,]
```

```
#remove probes with SNPs at CpG site
mSetSqFlt = dropLociWithSnps(mSetSqFlt)

#exclude cross reactive probes
xReactiveProbes = read.csv(file="48639-non-specific-probes-Illumina450k.csv",
stringsAsFactors=FALSE)
keep = !(featureNames(mSetSqFlt) %in% xReactiveProbes$TargetID)
table(keep)
mSetSqFlt = mSetSqFlt[keep,]

#dendogram after normalization and filter
beta <- getBeta(mSetSqFlt)
d <- dist(t(beta),method="euclidean")
fit<- hclust(d, method="complete")
pdf("clustering_afternormalization_filtering.pdf", onefile=T, paper="a4r")
plot(fit, cex = 0.8)
dev.off()

#MSD plots after filtering and normalization
pdf("MDSplots_afterfiltering_quantile.pdf")
par(mfrow=c(1,2))
colfunc <- colorRampPalette(c("royalblue", "white"))
colorgrad<-colfunc(length(sort(unique(pData(mSetSqFlt)$age))))
zzz<-as.factor(pData(mSetSqFlt)$age)
plotMDS(getM(mSetSqFlt), top=1000, gene.selection = "common", col=colorgrad[zzz],
pch=16)
pal = colorRampPalette(c('darksalmon', "cadetblue3"))(2)
plotMDS(getM(mSetSqFlt), top=1000, gene.selection = "common", col =
pal[factor(phenodata$sex)], pch=16)
dev.off()

#save the R object after filtering and normalization
saveRDS(mSetSqFlt, file = "quantile_filterNorm.rds")
```

2.4. Joining experiments

After the individual analysis of experiments, the protocol joins the samples from Dataset B.2 using the function `combinearrays` from *minfi* that creates a virtual global array (Box 8). The experiments were joined in pairs, since the function does not support a global simultaneous joining process. After that, a clustering and a MDS plot was made in order to see the potential source of bias of our global dataset and exclude the major confounders such as cell types.

Box 8 – Script of merging of all samples from Dataset B.2, in order to obtain Dataset B.3 after normalization and filtering followed by clustering and MDS plots.

```
#Join GEO datasets already normalized and filtered
gse104471<-readRDS("GSE104471/quantile_filterNorm.rds")
gse105123<-readRDS("GSE105123/quantile_filterNorm.rds")
gse107737<-readRDS("GSE107737/quantile_filterNorm.rds")
gse87571<-readRDS("GSE87571/quantile_filterNorm.rds")
gse42861<-readRDS("GSE42861/quantile_filterNorm.rds")
gse51057<-readRDS("GSE51057/quantile_filterNorm.rds")
gse71955<-readRDS("GSE71955/quantile_filterNorm.rds")
gse85506<-readRDS("GSE85506/quantile_filterNorm.rds")
gse85647<-readRDS("GSE85647/quantile_filterNorm.rds")
gse87640<-readRDS("GSE87640/quantile_filterNorm.rds")
```

```
gse98876<-readRDS("GSE98876/quantile_filterNorm.rds")
gse99755<-readRDS("GSE99755/quantile_filterNorm.rds")
gse104471_105123<-combineArrays(gse105123, gse104471,
outType="IlluminaHumanMethylationEPIC", verbose=TRUE)
gse107737_87571<-combineArrays(gse107737, gse87571,
outType="IlluminaHumanMethylationEPIC", verbose=TRUE)
gse42861_51057<-combineArrays(gse42861, gse51057,
outType="IlluminaHumanMethylationEPIC", verbose=TRUE)
gse71955_85506<-combineArrays(gse71955, gse85506,
outType="IlluminaHumanMethylationEPIC", verbose=TRUE)
gse85647_87640<-combineArrays(gse85647, gse87640,
outType="IlluminaHumanMethylationEPIC", verbose=TRUE)
gse98876_99755<-combineArrays(gse98876, gse99755,
outType="IlluminaHumanMethylationEPIC", verbose=TRUE)
gse104471_105123__gse107737_87571<-combineArrays(gse104471_105123, gse107737_87571,
outType="IlluminaHumanMethylationEPIC", verbose=TRUE)
gse42861_51057__gse71955_85506<-combineArrays(gse42861_51057, gse71955_85506,
outType="IlluminaHumanMethylationEPIC", verbose=TRUE)
gse85647_87640__gse98876_99755<-combineArrays(gse85647_87640, gse98876_99755,
outType="IlluminaHumanMethylationEPIC", verbose=TRUE)
gse104471_105123__gse107737_87571___gse42861_51057__gse71955_85506<-
combineArrays(gse42861_51057__gse71955_85506, gse104471_105123__gse107737_87571,
outType="IlluminaHumanMethylationEPIC", verbose=TRUE)
all_geo<-
combineArrays(gse104471_105123__gse107737_87571___gse42861_51057__gse71955_85506,
gse85647_87640__gse98876_99755, outType="IlluminaHumanMethylationEPIC",
verbose=TRUE)
##all_geo<-all_geo[,-c(1:18)] ###remove the experiment that has samples with age
below 18
##remove unnecessary columns
pData(all_geo)$characteristics_ch1<-NULL
pData(all_geo)$characteristics_ch1.1<-NULL
pData(all_geo)$characteristics_ch1.2<-NULL
pData(all_geo)$characteristics_ch1.4<-NULL
pData(all_geo)$characteristics_ch1.3<-NULL

#manipulate data to get all uniform
pData(all_geo)$sex <- sub("Female", "F", pData(all_geo)$sex)
pData(all_geo)$sex <- sub("Male", "M", pData(all_geo)$sex)
pData(all_geo)$sex <- sub("Gender: ", "", pData(all_geo)$sex)
pData(all_geo)$sex <- sub("gender: ", "", pData(all_geo)$sex)
pData(all_geo)$sex <- sub("Sex: ", "", pData(all_geo)$sex)
pData(all_geo)$sex <- sub("female", "F", pData(all_geo)$sex)
pData(all_geo)$sex <- sub("male", "M", pData(all_geo)$sex)
pData(all_geo)$sex <- sub("m", "M", pData(all_geo)$sex)
pData(all_geo)$sex <- sub("f", "F", pData(all_geo)$sex)
pData(all_geo)$age <- sub("\\.\\d+", "", pData(all_geo)$age)
saveRDS(all_geo, file = "all_geo.rds")

#clustering and density plots of all the datasets
pdf("densityplots_allgeo.pdf")
par(xpd=NA,oma=c(3,0,0,0))
densityPlot(getBeta(all_geo), sampGroups = pData(all_geo)$sex)
densityPlot(getBeta(all_geo), sampGroups = pData(all_geo)$age, legend=FALSE)
dev.off()
beta <- getBeta(all_geo) #[,1:19] number of samples
d <- dist(t(beta),method="euclidean")
fit<- hclust(d, method="complete")
plot(fit)
clusterCols<-c("lightgreen", "cadetblue3")
color<-clusterCols[as.factor(pData(all_geo)$sex)]
as.dendrogram(fit) %>% set("labels_col", color) %>% plot()
labels<-fit$labels
```

2.5. Statistical analysis

After the elimination of the biased samples, samples were selected to integrate the “old” (above 70 years old) and “young” (between 18 and 33 years old) groups and global comparison between methylated states of both groups was made using a density plot (Box 9). The data goes through a Pearson correlation analysis using the three methylation quartiles in order to analyse the global tendency of our dataset.

Box 9 – Script of the statistical analysis step in which the correlation coefficients were calculated and the global density plots were made

```
#separate dataset into young and old individuals
young<-pData(mSetSqFlt)[which(pData(mSetSqFlt)$age>=18 & pData(mSetSqFlt)$age<33),]
old<-pData(mSetSqFlt)[which(pData(mSetSqFlt)$age>=70),] #mSetSqFlt is refered only
to GSE87571
young["years"] <- NA
young$years <- "y"
old["years"] <- NA
old$years <- "o"
newtable<-rbind(young, old)
r1<-(pData(mSetSqFlt)$geo_accession %in% young$geo_accession)
gse_y=mSetSqFlt[, which(r1)]
beta_y<-getBeta(gse_y)
r2<-(pData(mSetSqFlt)$geo_geo_accession %in% old$geo_accession)
gse_o=mSetSqFlt[, which(r2)]
beta_o<-getBeta(gse_o)
toremove<-(pData(geo)$geo_accession %in% newtable$geo_accession)
gse_new=geo[, which(toremove)]
colData(gse_new)$years<-newtable$years
beta_new<-getBeta(gse_new)

#correlation between methylation and young/old individuals
age <- as.numeric(pData(gse_new)$age)
q1=apply(beta_new,2,function(x){quantile(x,0.25)})
q2=apply(beta_new,2,function(x){quantile(x,0.5)})
q3=apply(beta_new,2,function(x){quantile(x,0.75)})
cor.test(q1, age)
cor.test(q2, age)
cor.test(q3, age)

#density plot methylation vs age across chromosomes
meany<-rowMeans(beta_y)
meano<-rowMeans(beta_o)
annotation_o<-merge(meano, ann450k, by="row.names", all.x=TRUE)
annotation_o$chr <- sub("chr", "", annotation_o$chr)
annotation_y<-merge(meany, ann450k, by="row.names", all.x=TRUE)
annotation_y$chr <- sub("chr", "", annotation_y$chr)
pdf("meth_chrFINAL.pdf")
plot(c(1,22), c(0,1))
axis(1, c(1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20,21,22))
lines(as.numeric(annotation_y$x), col="blue", type="o", pch=22, lty=2)
lines(as.numeric(annotation_o$x), type="o", pch=22, lty=2, col="purple")
dev.off()
#density plot methylation vs age global
meany<-rowMeans(beta_y)
meano<-rowMeans(beta_o)
a<-density(meany)
b<-density(meano)
```

```
pdf("youngvsold(bluevspurple)_FINAL.pdf", onefile=T, paper="a4r")
plot(a, col="blue") #young
lines(b, col="purple") #old
dev.off()
```

2.6. Finding DMPs

Finding DMPs was the next step that gave us some insights about methylation across aging (Box 10). For DMPs finding, the `dmpFinder` was used, since it is the recommended function of *minfi*, and starts by the definition of our phenotype of interest – age as a continuous phenotype. According to the literature(Kuo 2017), the local significance level of a probe cannot be determined as a single event but a multiple testing correction should be used. For that reason, the p-value was adjusted using Bonferroni correction method and the DMPs with adjusted P-values above 0.05 were considered not significant. A manhattan and a volcano plot were made in order to display the annotated DMPs across the chromosomes and the variations in Δβ-values and the negative logarithm of its p-value, respectively. These plots allow us to determine the most significant probes according to p-value and differential methylation quantification (the amount of effect).

Box 10 – Script of the DMP finding step for Dataset B.3 followed by its representation in volcano and manhattan plots.

```
#find DMPs
beta <- getBeta(mSetSqFlt)
age <- pData(mSetSqFlt)$age
dmp_age <- dmpFinder(beta, pheno = age  , type = "continuous")
dmp<-dmp[dmp$pval<0.05,]
young<-rownames(pData(mSetSqFlt))[ pData(mSetSqFlt)$SW_Age >= 18 &
pData(mSetSqFlt)$SW_Age < 33]
old<-rownames(pData(mSetSqFlt))[ pData(mSetSqFlt)$SW_Age >= 70]
dmpCpgs = rownames(dmpfinal_age)
dmpfinal_age$young = rowMeans(beta[dmpCpgs, young, drop=F])
dmpfinal_age$old = rowMeans(beta[dmpCpgs, old, drop=F])
dmpfinal_age$deltaBeta = dmpfinal_age$old - dmpfinal_age$young
p_adjusted<-p.adjust(dmp_age$pval, method="bonferroni") #p value adjusted with
bonferroni method
dmpfinal<-cbind(dmp_age, p_adjusted)
annotation_age<-merge(dmpfinal_age, ann450k, by="row.names", all.x=TRUE)
annotation_age$log10<-NULL
annotation_age$log10<--(log10(annotation$pval))
annotation_age$chr <- sub("chr", "", annotation$chr)
annotation_age$chr <- as.numeric(annotation$chr)
y_o<-annotation_age
write.table(y_o, file="annotation_dmps_age.txt")

#volcano plot - age
pdf("volcanoplot_age.pdf")
with(y_o, plot(y_o$deltaBeta, y_o$log10, pch=20, main=""))
abline(h = 5.0, col = "blue", lty = 2, lwd = 1)
abline(v = c(-0.1,0.1), col = "blue", lty = 2, lwd = 1)
with(subset(y_o, y_o$log10<5.0), points(deltaBeta, log10, pch=20, col="gray"))
with(subset(y_o, y_o$deltaBeta< -0 & y_o$log10>5.0), points(deltaBeta, log10,
pch=20, col="red"))
```

```
with(subset(y_o, y_o$deltaBeta> 0 & y_o$log10>5.0), points(deltaBeta, log10,
pch=20, col="green"))
dev.off()
#manhattan plot-age
pdf("manhattan-age.pdf")
manhattan(annotation_age, chr="chr", bp="pos", p="pval", snp="Row.names",
col=c("grey","skyblue")) ##check the borderline
dev.off()

#most significative probe linear regression
theone<-as.numeric(beta[rownames(beta)=="cg16867657", ])
pdf("dmp_age_cg16867657.pdf", onefile=T, paper="a4r")
plot(age, theone, xlim=c(1, 100), ylim=c(0, 0.9))
abline(lm(theone ~ age), col="blue")
dev.off()
summary(lm(theone ~ age))$r.squared
```

3. Results

The manifest object obtained for this chapter included some important information about the array, mainly the quantity of probes used and its type – Type I, II, Control, SNP Type I and SNP Type II. This manifest was the same for all the samples studied, since they were all analysed using 450k technology (Box 11).

Box 11 – Manifest object general information about Dataset B.2 and B.3

```
##IlluminaMethylationManifest object
##Annotation
    array: IlluminaHumanMethylation450k
##Number of type I probes: 135476
##Number of type II probes: 350036
##Number of control probes: 850
##Number of SNP type I probes: 25
##Number of SNP type II probes: 40
```

Next, the results related to individual experiments (quality control, normalization and filtering) will be exemplified using only the 19 samples from experiment GSE105123 of Dataset B.2. The data from the rest of datasets is presented in Supplementary Figures 11-15.

3.1. Quality control, normalization and filtering

Once the data was imported into R, the quality control step took place. The p-value was determined and represented in a plot (Figure 36). Small p-values are indicative of a reliable signal. Usually a significance level lower than 0.01 indicates a good quality signal, we could conclude about the general quality of our samples in terms of the overall signal reliability, according to the literature (Maksimovic et al. 2016). Additionally, it is known that when plotting both the log median intensity of methylated and unmethylated channels against each other the good samples will cluster together while the failed samples will tend to separate and have lower median intensities. The control probes were also quality checked since they are normally used to assess the overall quality of sample preparation protocol.

Next, and as a way to deeper our quality control, we decided to do the analysis of the β-value distributions in the density plots (Figure 37). From this we could conclude that the data needed normalization since there were several deviations of the characteristic shape of the plot that should have one node close to 0% methylation and a second close to 100%. Control probes were also checked (data not shown). In Figure 37 the density plot obtained through `qcreport` for the age variable is presented. In the case of gender, the shape of the plot was similar (data not shown).

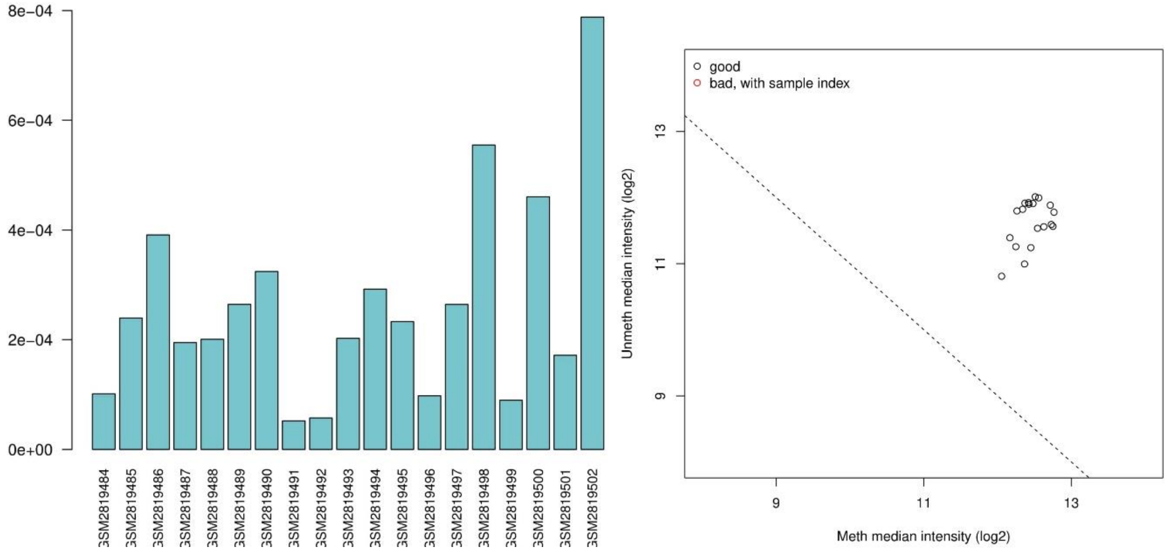


Figure 36 – Quality control plot with the mean detection of p-values (y axis) in each sample (x axis) (left) and the representation of both the log median intensity of methylated and unmethylated channels against each other (right). The left plot reveals a general quality of samples in terms of the overall signal reliability while in the right the good samples clustering together.

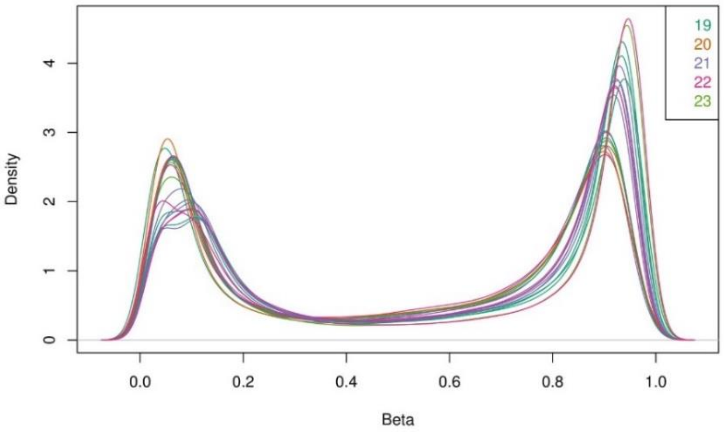


Figure 37 – Density plot for samples quality control. The plot represents the distribution of β-values across several sample groups, with age as phenotype of interest, before the normalization procedure.

The normalization step started with a test to the three normalization methods available in *minfi* – SWAN, Quantile and Funnorm. A direct comparison of these plots was performed to the density plot using raw data, in order to choose the best one. Through the analysis of Figure 38 we concluded that the Quantile normalization was the best method for our study. Although the funnorm method is particularly useful for studies comparing conditions with known large-scale differences, it is also known that the quantile function is better to study single tissue variations. However, the quantile normalization presented better results in the density plots from all variables – age and gender – so we decided to use this method.

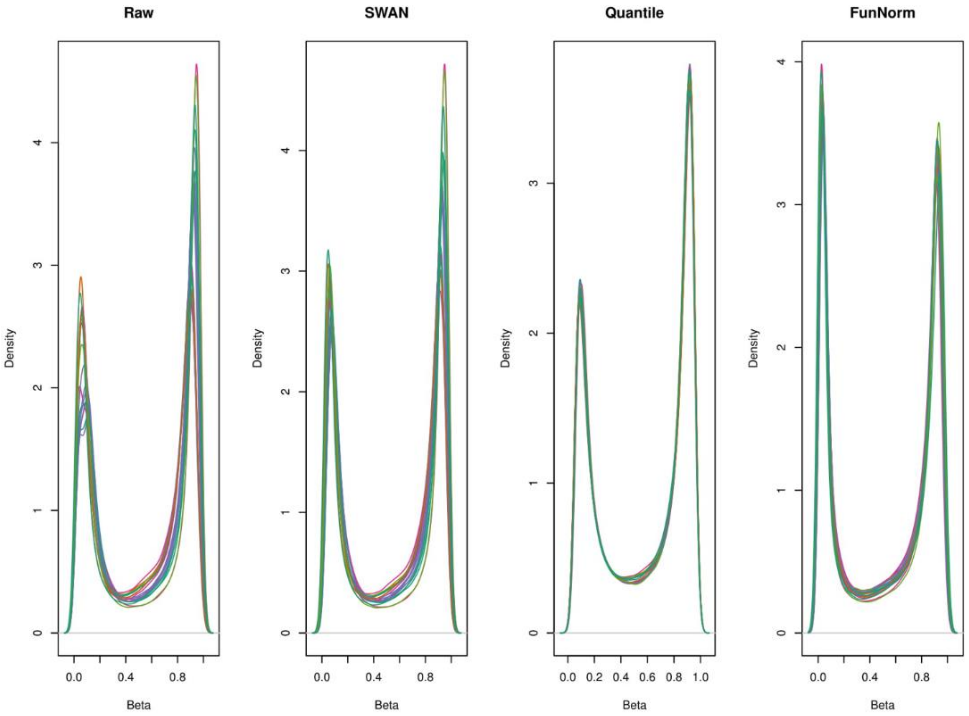


Figure 38 – Density plots of the three available normalization processes for the age phenotype in Dataset B.2. The best normalization was achieved with Quantile normalization which shows a most uniform plot.

To acess the effect of the normalization process, we compared the several statistical analysis made before and after the normalization methods, in order to determine the kind of variables that were globally affecting our dataset. Through the clustering of samples on a condensed dissimilarity matrix, we concluded that our methylation data was clearly influenced by the gender of the samples, turning unfeasible the characterization of other features, like age. These findings were confirmed by the MDS plots in which the greatest source of variation, captured by dimension one (or principal component 1) of the plot, was also gender (Figure 39). This is according to the literature since it is known that gender accounts for the larger methylation effects even at autosomal chromosomes (Wright et al. 2016).

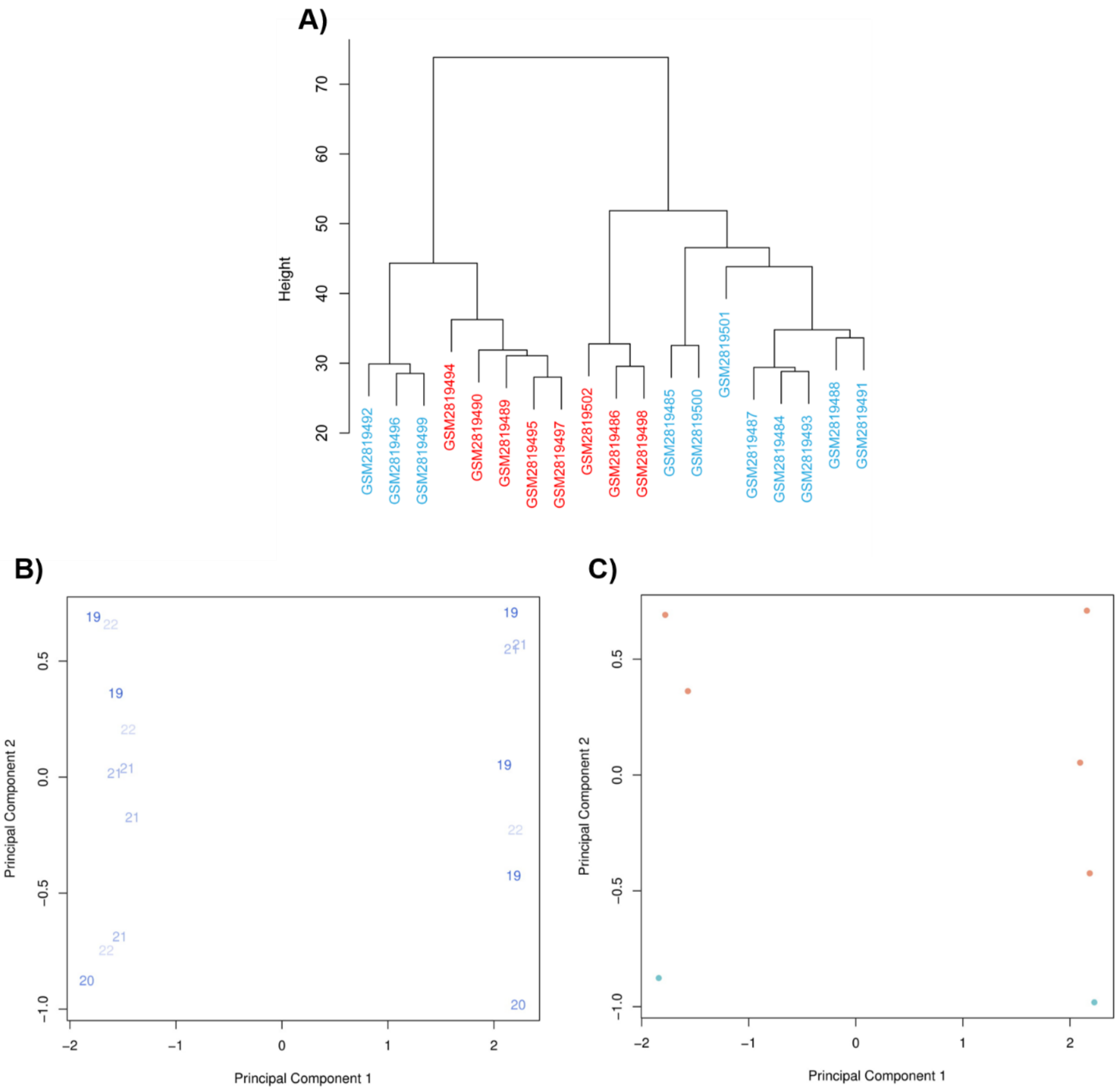


Figure 39 – Clustering (A) and MDS plot of age (B) and gender (C) before the normalization and filtration procedures A gender tendency is observed, which affects the global methylation in our dataset.

Therefore, it was essential to carry out a filtering process starting by the removal of probes with p-value above 0.01 which means they failed in one or more samples, the remotion of sex chromosome probes and the removal of probes with SNPs at CpG sites. Finally, the cross-reactive probes were excluded and we obtained as a final product a total of 455 400 valid probes. The probe removal for all the experiments of Dataset B.2 was summarized in Table 11. In this table we can also detect the filtration of a great amount of probes in GSE87648 which made us remove this experiment from our dataset. The MDS and the clustering plot were repeated and there was no long a variable affecting our results (Figure 40). For these reasons, the protocol proceeded.

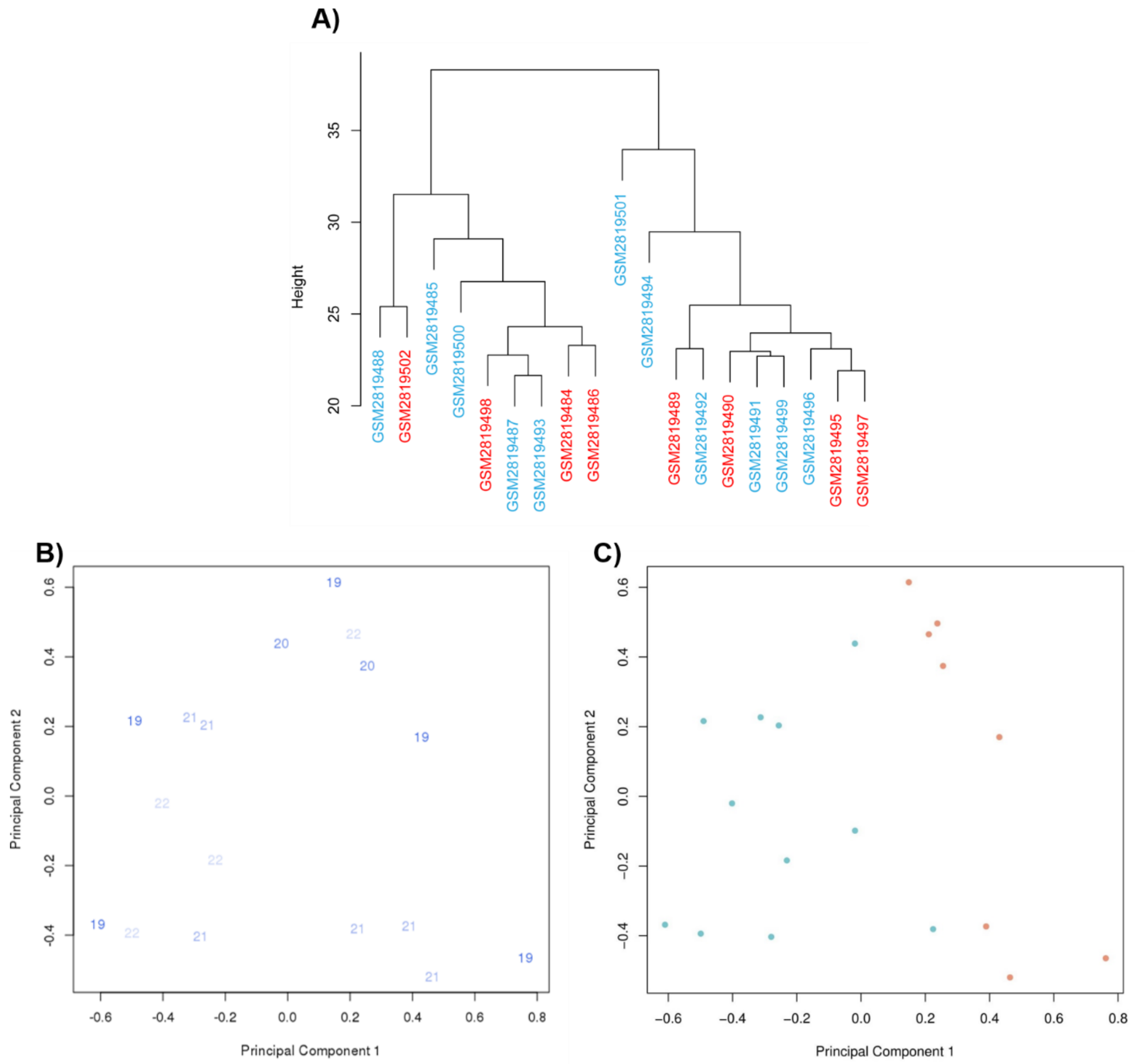


Figure 40 - Clustering (A) and MDS plot of age (B) and gender (C) of Dataset B.2 after the normalization and filtration procedures. The gender effect is no longer visible, since samples are mixed, which means that the dataset is ready to be proceed with the analysis.

Table 11 – Variation of initial and maintained probes after filtering

GEO Serie	*Initial probes*	*Final probes*
GSE102177	622399	456041
GSE104471	622399	456226
GSE105123	622399	455400
GSE107737	622399	456077
GSE42861	622399	451244
GSE51057	622399	453882
GSE71955	622399	454341
GSE85506	622399	455830
GSE85647	622399	450132
GSE87571	622399	446785
GSE87640	622399	456041
GSE87648	622399	277867
GSE98876	622399	456256
GSE99755	622399	455215

After joining all experiments, the clusters and the MDS plots of a virtual global array were repeated (Supplementary Figure 16-17) in order to see if there were differences between experiments as reported in Chapter II that could introduce biases in our study. Through its analysis we concluded that there was a general clustering of samples according to the experiment and type of blood category. The different types of blood cells used, different laboratory conditions or methodologies adopted or even sample information, which was not published by the original owners of the data, can constitute possible reasons for these findings. Therefore, further analysis only used the GSE87571 experiment, from now on called DataSet B.3, since this was the one with the highest number of samples and ages range (Figure 41) and did not include different blood cell types. In this final dataset, 72% of probes (446785) passed the quality control step, it included 664 individuals, 356 females and 308 males, the age ranged between 18 and 94 years old and the mean of individual age was 50 years old.

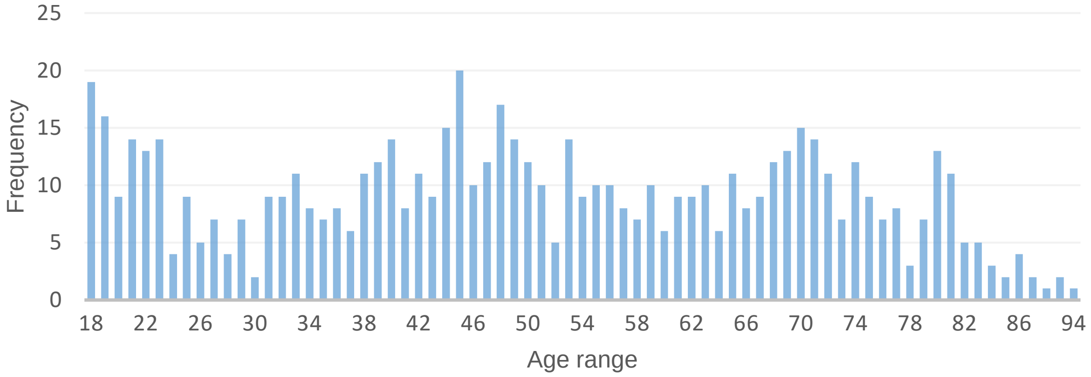


Figure 41 – Age variation in GSE87571 dataset with a range of ages between 18 and 94 years old.

However, since the type of sample of this dataset was whole blood, it could be expected that our differential methylation analysis become biased if there was cell composition major fluctuations. To test this hypothesis, we normalized the samples according to the cell type and compared the global patterns of methylation of samples normalized as such and as before (Quantile method). We made a Pearson correlation test and obtained a t-test of 5363300 with 322380000 degrees of freedom, a p-value under $2.2x10^{-16}$ and a correlation coefficient of 0.9999. A plot was also made in order to validate our expectations (Figure 42), from which we can conclude that the normalization using quantile only or taking also cells into consideration would not influence the results. From this results we conclude that the normalization by cell type would not influence the differential methylation analysis and for that reason it was not considered, reason why we only performed a quantile normalization method.

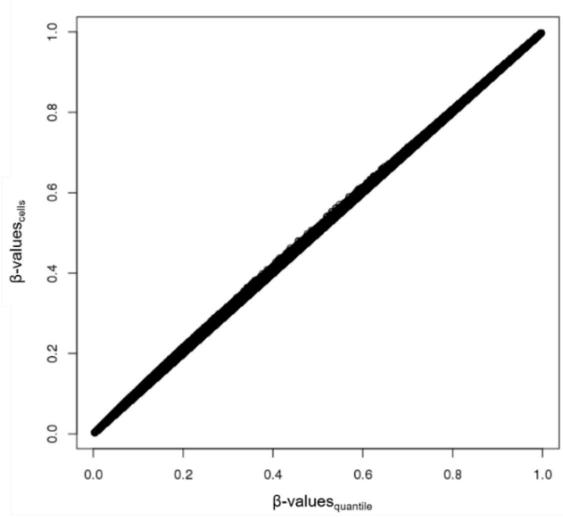


Figure 42 – Comparison of β-values using quantile normalization or using quantile and cell normalization. Values are highly correlated and therefore the normalization should not influence the results.

3.2. Differential methylation analysis

In order to evaluate the global tendency of the methylation values with our phenotype, we calculated Pearson correlation in our final dataset. In young individuals, we demonstrated an increasing of the first quartile methylation with age (R = 0.2720, p-value = 0.00110) and a decreasing in second (R = -0.2112, p-value = 0.01195) and third quartile methylation (R = -0.1518, p-value = 0.07231). On the other hand, in old individuals we demonstrated an increasing of the first quartile methylation with age (R = 0.01159, p-value = 0.8911), a decreasing of the median quartile (R=-0.1556, p-value = 0.0645) and an increasing in the third quartile (R = 0.1138, p-value = 0.1776). This data was according to the expected and already published (Johansson et al. 2013).

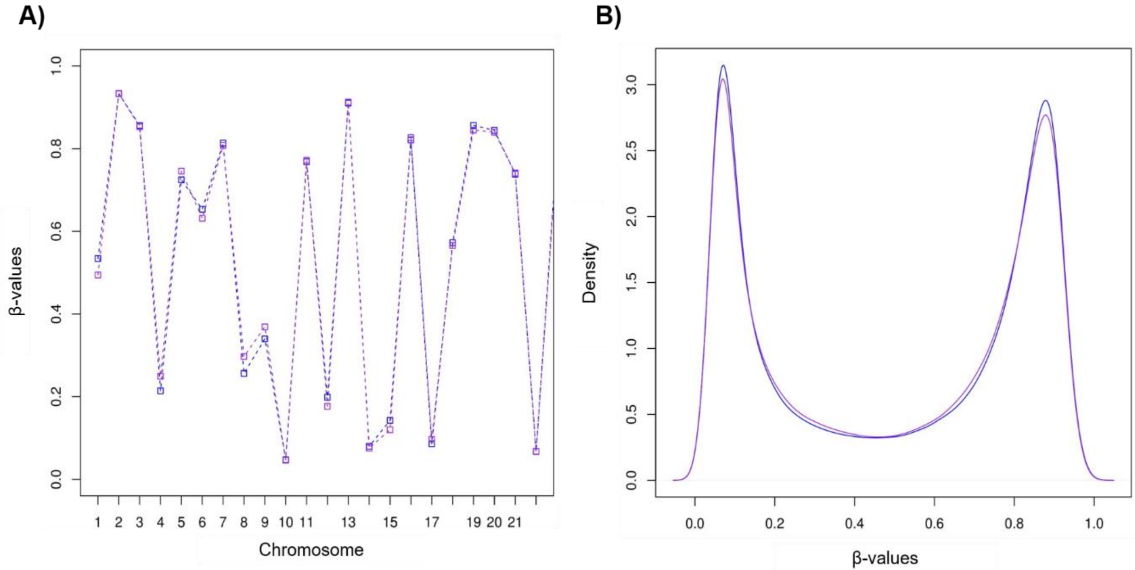


Figure 43 – Global comparison of methylation between young (blue) and old (purple) individuals of Dataset B.3 (A) Methylation across chromosomes; (B) Density plot. The highest differences are located in chromosomes 1, 4, 5, 6, 8, 9, 12, 15, 19.

In order to make a differential methylation analysis across age, the dataset was separated in young and old individuals which resulted in 141 young and 142 old individuals.We performed a density plot between old (purple) and young (blue) individuals that shows a global variation between the autosomal methylation levels in both phenotypes although the methylated and unmethylated sites appeared with a similar density in our samples (Figure 43B). The variation of global methylation in both phenotypes was refined and analysed according to its distribution across chromosomes. The highest differences between both groups were located in chromosomes 1, 4, 5, 6, 8, 9, 12, 15 and 19 (Figure 43A).

We identified 95978 significant a-DMPs and represented them according to the negative logarithm of detection p-value and chromosome positions (manhattan plot) or variation of β-values (volcano plot) (Figure 44). According to the literature(Kuo 2017), the local significance level of a probe cannot be determined as a single event but should use a multiple testing correction. In this case, we have 450k probes and wanted to consider a significant adjusted p-value of 0.05. For that reason, we considered a significant raw p-value of $1.11x10^{-7}$.Of the total of significant probes, 76% with gene associations, 23% enhancer associated and 14% DMR associated. In the analysis of the volcano plots we identified the positive or negative β-value variations across DMPs in the case of young/old individuals. In Table 12 the top-significative DMPs are presented according to its position in the genome, associated genes and its functions and also with statistical information about DMP finding – Bonferroni adjusted p-value, -log10 (raw p-value), Q-value, Δβ and the slope of a linear regression of the methylation rate of the probe across samples and its age.

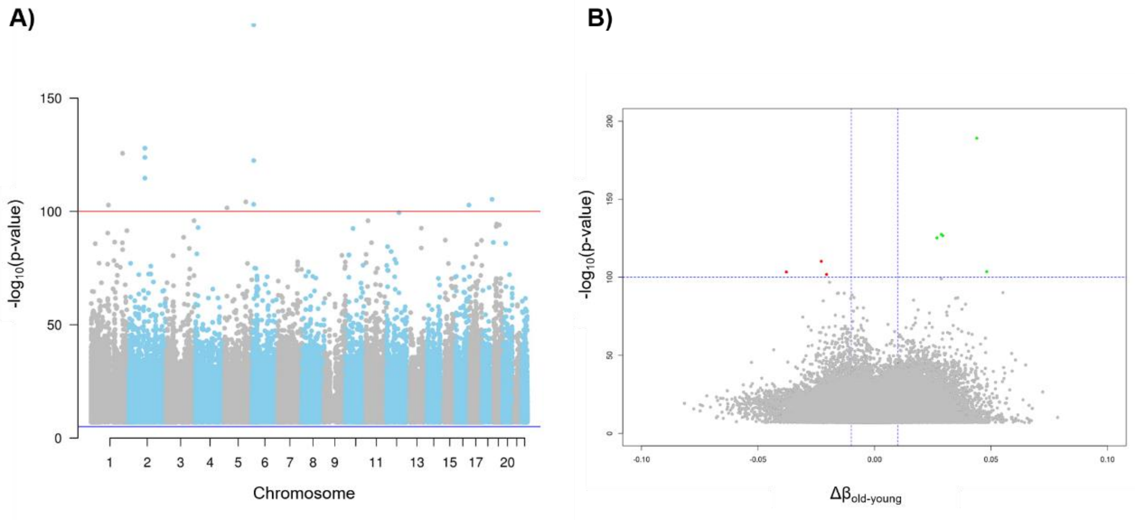


Figure 44 –Representation of all DMPs found with age in Dataset B.3 in a (A) manhattan plot with a selected threshold of p-value = $1x\ 10^{-100}$; and in a (B) volcano plot using a selected threshold of p-value = $1x10^{-100}$ and $|\Delta\beta|>0.2$.

Table 12 – Summary of the top-significant DMPs with age phenotype in Dataset B.3. Probes are orders by statistical significance – 8 are methylation gains while 4 are losers.

Probes	*slope*	*P_{adj}-value*	*Q-value*	*Δβ*	*Chr*	*Position*	*Islands Name*	*Relation to Island*	*Gene Group*	*Gene Name*	*Gene Function*	*Log_{10}(of p-val*
cg16867657	0.00431	1,85x10^{-177}	5,02Ex10^{-178}	0,0349	6	11044877	chr6:11043913 - 11045206	Island	TSS1500	ELOVL2	Fatty acid elongase activity	182,382
cg22454769	0.00348	5,44x10^{-123}	7,37x10^{-124}	0,0286	2	106015767	chr2:106014878 - 106015884	Island	TSS200; 5'UTR	FHL2	Protein binding	127,914
cg10501210	-0.00505	1,06x10^{-120}	9,60x10^{-122}	-0,0460	1	207997020	N/A	Open Sea	N/A	N/A	N/A	125,623
cg06639320	0.00330	6,81x10^{-119}	4,61x10^{-120}	0,0252	2	106015739	chr2:106014878 - 106015884	Island	TSS200; 5'UTR	FHL2	Protein binding	123,817
cg21572722	0.00217	1,54x10^{-117}	8,32x10^{-119}	0,0210	6	11044894	chr6:11043913 - 11045206	Island	TSS1500	ELOVL2	Fatty acid elongase activity	122,464
cg24079702	0.00335	8,90x10^{-110}	4,02x10-^{111}	0,0265	2	106015771	chr2:106014878 - 106015884	Island	TSS200; 5'UTR	FHL2	Protein binding	114,701
cg19283806	-0.00256	1,95x10^{-100}	7,55x10^{-102}	-0,0161	18	66389420	N/A	Open Sea	5'UTR	CCDC102B	Protein binding	105,359
cg23500537	0.00243	2,63x10^{-99}	8,92x10^{-101}	0,0217	5	140419819	N/A	Open Sea	N/A	N/A	N/A	104,229
cg07082267	-0.00019	6,83x10^{-98}	1,69x10^{-99}	-0,0067	16	85429035	N/A	Open Sea	N/A	N/A	N/A	102,816
cg16008966	-0.00235	6,88x10^{-98}	1,69x10^{-99}	-0,0136	1	114761794	N/A	Open Sea	N/A	N/A	N/A	102,812
cg24724428	0.00393	3,34x10^{-98}	1,01x10^{-99}	0,0352	6	11044888	chr6:11043913 - 11045206	Island	TSS1500	ELOVL2	Fatty acid elongase activity	103,126
cg06782035	0.00330	1,25x10^{-96}	2,83x10^{-98}	0,0345	5	16179135	chr5:16179064 - 16180420	Island	Body	MARCH11	Transferase activity	101,552

In the top-significant DMPs, the most significant DMP was cg16867657 located in chromosome 6, with an adjusted p-value of 1.85x10^{-177} and with an increase of methylation across age of about 4%. The representation of the DNA methylation level across all individuals from our dataset shows a positive correlation between age and the methylation state for this probe (Figure 45). cg16867657 is located in the promoter of ELOVL2 gene, particularly 1500 nucleotides distant from the transcription start site (TSS). ELOVL2 is a gene associated with the elongase activity of fatty acids. This gene was associated to other 3 significant probes, all of them located in the promoter, at a distance of 1500 nt (nucleotides) from the TSS, and all of them with an increase in methylation across age. The second most significant DMP was cg22454769 located in chromosome 2, with an adjusted p-value of 5.44x10^{-123} and with an increase of methylation across age of about 3%. This probe was located in the 5'UTR and promoter region, at a distance of 2000 nt from the TSS, of the FHL2 gene, a gene associated with protein binding that was also found in another two significant probes, all showing an increase of methylation level across age.

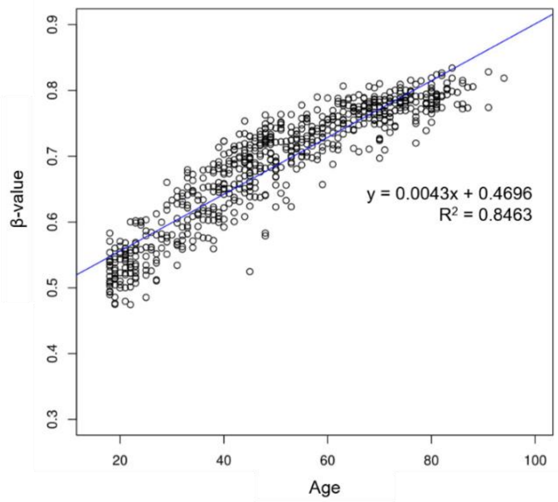


Figure 45 – Representation of the DNA methylation level of cg16867657 across age in Dataset B.3 and corresponding regression line.

As a way to explore global trends relative to methylation degree in different genomic regions, we performed a comparison between the total hypo and hypermethylated sites and its relation to island and gene positions (Figure 46). Through this analysis we concluded that the global variation of the number of hyper and hypomethylated sites in genomic region is minimal (0.1 – 2.8% of variation). The probes inside CGI represent the majority of our analysis, with about 41% of total probes and an average of 2.23 markers per island ranging from 1 to 36. The open sea positions were also relevant since they constituted about 13% and 15% of the hypomethylated and hypermethylated probes in our analysis, respectively. On the other hand, in the case of the gene affected part, we observed that about 22% and

18% of the sites were hypermethylated or hypomethylated and that both are mainly located in the body of a gene. This represents about 41% of the hypermethylated sites associated with genes and 37% of the hypomethylated sites respectively, associated with genes.

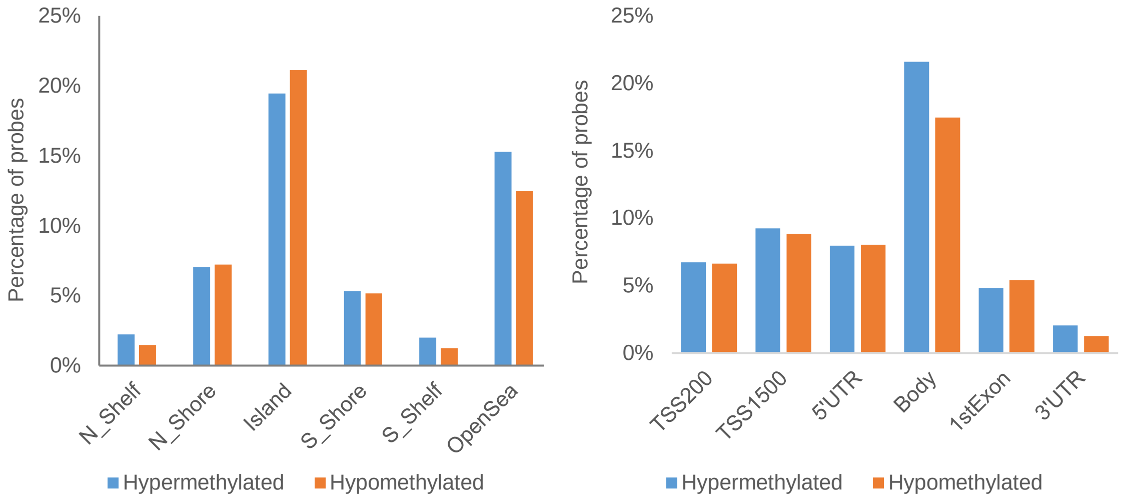


Figure 46 – Distribution of DMPs of Dataset B.3 according to its relation with the genomic CpG region (left) and gene part (right). There was a prevalence of DMPs located in CGI and body of genes.

3.3. Gene ontology analysis

Since the total of significant DMPs was too high (N = 95978) to use directly on gene ontology analysis, we performed a selection according to the procedure suggested by (Johansson et al. 2013). The island locations with more than three significant probes were separated between hypo and hypermethylated, the top 500 genes for each condition were included and only one gene per CGI was included, in order to avoid introducing bias downstream. The gene ontology analysis was then performed using PANTHER overrepresentation SLIM test either in biological process, molecular function and cellular components.

Table 13 shows gene functions associated with the hyper and hypomethylated phenotypes. We observed that the biological processes associated with genes with hypermethylated CGIs were overrepresented mainly in functions associated with development and differentiation while the molecular functions are associated with transcription regulation or cell signalling. In the case of the hypomethylated CGIs, we reveal a biggest influence in the regulation of cellular cycle, signal transduction and protein binding.

Table 13 – Gene Ontology enrichment analysis for biological process, molecular function and cellular components in hypo and hypermethylated genes of Dataset B.3. GO terms are ordered by hierarchy and statistical significance.

Category		*Function*	*UR/OR*	*FE*	*Raw p-value*
Hypermethylated	Biological process	muscle organ development (GO:0007517)	OR	12.32	2.15E-05
		digestive tract mesoderm development (GO:0007502)	OR	12.10	5.75E-04
		segment specification (GO:0007379)	OR	10.32	5.37E-10
		heart development (GO:0007507)	OR	10.00	1.08E-03
		pattern specification process (GO:0007389)	OR	8.81	6.90E-11
		ectoderm development (GO:0007398)	OR	8.14	2.12E-17
		embryo development (GO:0009790)	OR	7.59	1.85E-08
		anatomical structure morphogenesis (GO:0009653)	OR	6.12	7.95E-08
		cyclic nucleotide metabolic process (GO:0009187)	OR	5.64	2.28E-05
		mesoderm development (GO:0007498)	OR	5.13	2.76E-10
		neuron-neuron synaptic transmission (GO:0007270)	OR	4.73	1.02E-03
		behavior (GO:0007610)	OR	4.71	5.41E-03
		cell-cell adhesion (GO:0016337)	OR	3.26	4.18E-03
		developmental process (GO:0032502)	OR	3.22	3.78E-21
		system development (GO:0048731)	OR	2.93	8.38E-06
		regulation of transcription from RNA polymerase II promoter (GO:0006357)	OR	2.79	1.38E-06
		cell differentiation (GO:0030154)	OR	2.62	2.01E-05
		synaptic transmission (GO:0007268)	OR	2.56	5.57E-04
		cell-cell signaling (GO:0007267)	OR	2.42	1.27E-04
		nervous system development (GO:0007399)	OR	2.38	4.35E-03
		transcription from RNA polymerase II promoter (GO:0006366)	OR	2.15	2.24E-04
		regulation of phosphate metabolic process (GO:0019220)	OR	2.14	2.22E-03
		transcription, DNA-dependent (GO:0006351)	OR	1.97	2.02E-04
		single-multicellular organism process (GO:0044707)	OR	1.93	3.66E-06
		system process (GO:0003008)	OR	1.92	3.62E-04
		multicellular organismal process (GO:0032501)	OR	1.91	4.25E-06
		neurological system process (GO:0050877)	OR	1.80	2.91E-03
		intracellular signal transduction (GO:0035556)	OR	1.72	3.82E-03
		cell communication (GO:0007154)	OR	1.43	2.67E-03
	Molecular function	glutamate receptor activity (GO:0008066)	OR	7.37	8.93E-04
		adenylate cyclase activity (GO:0004016)	OR	5.45	7.45E-05
		sequence-specific DNA binding RNA polymerase II transcription factor activity (GO:0000981)	OR	4.48	3.20E-07

	Category	*Function*	*UR/OR*	*FE*	*Raw p-value*
Hypermethylated	Molecular function	sequence-specific DNA binding transcription factor activity (GO:0003700)	OR	3.21	1.78E-11
		G-protein coupled receptor activity (GO:0004930)	OR	2.98	1.64E-04
		ion channel activity (GO:0005216)	OR	2.57	1.13E-03
		DNA binding (GO:0003677)	OR	2.55	8.13E-09
		nucleic acid binding (GO:0003676)	OR	1.91	7.46E-06
		binding (GO:0005488)	OR	1.44	9.41E-06
Hypomethylated	Biological process	regulation of cell cycle (GO:0051726)	OR	3.85	4.08E-04
		intracellular signal transduction (GO:0035556)	OR	2.07	1.60E-04
	Molecular function	protein binding (GO:0005515)	OR	1.66	5.78E-05

UR - Underrepresentation | OR - Overrepresentation | FE - Fold Enrichment

4. Discussion

Currently, the influence of DNA methylation as an epigenetic mechanism that is significantly changed with aging in human is already widely accepted by the scientific community. For that reason, in this chapter we used in-silico experiments in order to validate our bioinformatics pipeline and to try to determine new epigenetic markers of age. According to our data, both young and old individuals demonstrated to have a positive and negative correlation with the lower and median levels of methylation across age, respectively. On the other hand, the higher levels of methylation are positive correlated with age in old individuals and negative correlated with age in young individuals.

We should focus into DMPs, since they represent the most important regions to study when looking for methylation patterns. According to the literature, several lists of the most methylation-influenced genes during aging have already been published. From those, ELOVL2, FHL2, CCDC102B, ZNF423, ASPA, PDE4C and C1orf132 should be taken into consideration into a methylation analysis. In our study, we identified 12 significant positions associated with aging. From these, 8 presented a hypermethylated pattern and 7 were located inside CGIs associated with specific genes – ELOVL2, FHL2, CCDC102B and MARCH11. The gene associated with the most significative probe, ELOVL2, encodes for a transmembrane protein involved in the synthesis of long polyunsaturated fatty acids, molecules involved in functions like energy production, modulation of inflammation and maintenance of cell membrane integraty, that is mainly expressed in liver but that was also determined to be hypermethylated in blood (Garagnani et al. 2012; Bacalini et al. 2017).

This gene has been extensively reported as an epigenetic marker of age that variates its methylation level across age from 7% to 91%. For this reason can be used to calculate the human age or even the number of cell divisions in a cell culture(Garagnani et al. 2012; Bacalini et al. 2017). On the other hand, FHL2, the gene identified in the second most significative probe, encodes a transcriptional cofactor (called FHL2) that can interact with many different proteins and is involved in organ differentiation, development, cell apoptosis and carcinogenesis(Wang et al. 2016). This gene was also already reported as an epigenetic marker of age, although its hypermethylation degree with age is more restricted than ELOVL2, consisting of 12% to 53%(Garagnani et al. 2012; Bacalini et al. 2017). In the case of hypomethylation markers, CCDC102B seems to play an important role. This gene is a 297-aminoacid protein coding gene but with an unknown function(Park et al. 2016). In the case of MARCH11, that codes for a family of membrane-bound E3 ubiquitin ligases which add ubiquitin to target lysines in substrate proteins signaling their intracellular transport, it has not been yet reported in methylation across age, but according to our results it might be promising as a methylation aging marker. In Figure 47 there is made a representation of the behavior in Dataset B.3 of the most methylation-influenced genes in aging together with the probe of MARCH11.

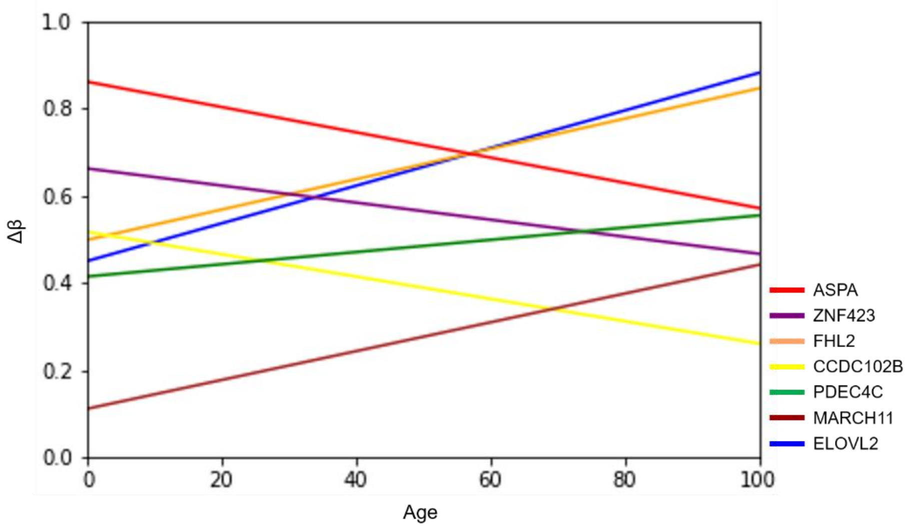


Figure 47 – Comparison of the linear regressions obtained for ASPA (cg02228185), ZNF423 (cg04208403), FHL2 (cg 22454769), CCDC102B (cg19283806), PDEC4C (cg17861230), MARCH11 (cg06782035), ELOVL2 (cg16867657) in Dataset B.3.

Exploring our Gene Ontology analysis, the processes associated with the majority of hypermethylated islands are associated with development, differentiation, tissue specifications and morphogenesis, all of these processes are known to be highly influenced by methylation across age due to the differential gene expression associated with it. The presence of processes that affect regulation of transcription was also overrepresented in hypermethylated genes, as it would be expected, since this is the most reported effect of hypermethylation across age. However, the number of enrichment events obtained for hypomethylated probes was residual although we did the Gene Ontology analysis using an equal number of genes in the hyper and hypomethylated conditions. This is also reported in literature, since it is expected that the CpG sites hypermethylated during aging are enriched to common processes and exhibit shared features, while hypomethylated sites are not homogenous and may ocurr sporadically at sites with a less central role.

Therefore, we can conclude that our results agree with the literature and that the performed procedure is promising in the determination of methylation patterns across age in the Portuguese population. However, in further explorations of DNA methylation across age it is essential to guarantee a larger cohort with homogenous tissue specificities. The exploration of other tissues for differentiated methylation pattern discovery mainly brain and liver, and its comparison with whole-blood methylomics markers would also be a promising goal, in order to improve the usage of blood as an accessible source of information for this type of studies.

CHAPTER IV DATA VALIDATION

Methylome analysis of samples from iBiMED

1. INTRODUCTION

The main goal of this chapter was to validate the methodology and the results about age-differential methylation patterns obtained in the previous chapter. For that purpose, the project POCI-01-0145-FEDER-016428-PAC-MEDPERSYST, developed between iBiMED and Life and Health Sciences Research Institute (ICVS), provided samples from healthy individuals evaluated in good and bad cognitive performance to methylome analysis (Serre-Miranda et al. 2015).

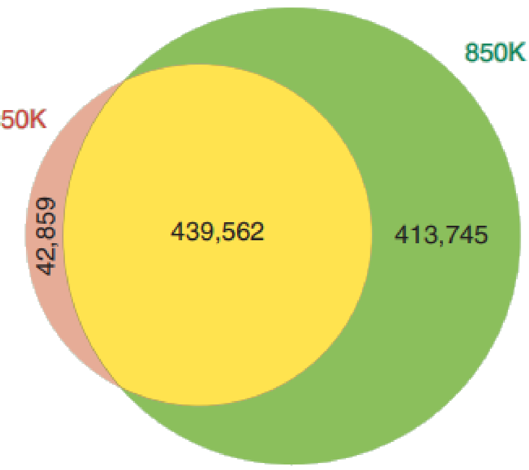


Figure 48 – Differences in the 450k and 850k CpG sites coverage. The 850k technology covers 91% of the CpGs included in the 450k array, making a difference between the two techniques of 42 859 CpG sites included in 450k and not in 850k and 413 745 CpG sites included in 850k and not in 450k.

As refered before, the development of tools for genome wide scale analyses of epigenetics influences on transcription is in constant development (Moran et al. 2016). Although the importance and impact that the 450k array had in the genomic research, the release of the Illumina Infinium MethylationEPIC BeadChip technology enabled the study of methylation also in enhancer regions (covered in 333 265 CpG sites) (Moran et al. 2016). These regions affect the transcription process through the looping and contact of DNA elements interspersed at great genomic distance (Moran et al. 2016). Therefore, given that methylation status can affect the binding of cognate transcription factors, it is probable that DNA methylation differences in enhancer sequences exert a major role in cell and tissue functionality.(Moran et al. 2016) The EPIC array, also known as the 850k array, can cover a total of 853 307 CpG sites, including those present in and the 450k array. In fact, this technology interrogated the methylation status of 91% of the CpGs included in the 450k array (Figure 48).

Although the *minfi* package described before software package designed for the Illumina HumanMethylation450 array, it was already described as a useful package for handling DNA methylation data from other arrays, like HumanMethylationEPIC. (Fortin et al. 2017) This adaptation requires a convertion of EPIC array to a virtual 450k array by joint normalization and processing of data from both platforms and an estimation of cell type proportions for EPIC samples using external reference data from 450k. (Fortin et al. 2017)

2. Methodologies adopted

The dataset used for validation of this work was constituted by samples from healthy individuals with ages between 52 and 77 years old. In the beginning of the analysis, our dataset was composed of 48 samples but it was reduced to 41 due to inconsistencies between the gender reported and the observed one. The dataset used, from now on called Dataset C, is presented in Table 12. Since these samples have been studied in other project, donnors have been evaluated to cognitive performance as good (1) or bad (4).

Table 14 – Dataset C used to validate our data. Samples are from 41 male and female healthy individuals with ages between 52 and 77 years old evaluated as good (1) or bad (4) as cognitive performance.

Sample Code	*Sample Name*	*Age*	*Gender*	*Cognitive Performance*	*Basename*
B06_M_60_4	sw0033C_A	60	Male	4	202060330094_R02C01
E06_M_77_1	sw0242C_A	77	Male	1	202060330094_R05C01
C03_F_64_4	sw0247C_A	64	Female	4	202053820063_R03C01
B02_F_68_1	sw0269C_A	68	Female	1	202053820039_R02C01
E02_F_58_1	sw0291C_A	58	Female	1	202053820039_R05C01
H02_F_68_1	sw0318C_A	68	Female	1	202053820039_R08C01
D01_F_62_4	sw0397C_A	62	Female	4	202053820031_R04C01
F05_M_57_4	sw0410C_A	57	Male	4	202060330086_R06C01
G06_M_68_1	sw0457C_A	68	Male	1	202060330094_R07C01
H01_F_70_4	sw0544C_A	70	Female	4	202053820031_R08C01
F06_M_77_1	sw0598C_A	77	Male	1	202060330094_R06C01
F01_F_61_1	sw0668C_A	61	Female	1	202053820031_R06C01
A06_M_72_1	sw0753C_A	72	Male	1	202060330094_R01C01
E01_F_68_4	sw0879C_A	68	Female	4	202053820031_R05C01
C01_F_70_1	sw0930C_A	70	Female	1	202053820031_R03C01
E03_F_58_4	sw1055C_A	58	Female	4	202053820063_R05C01
A03_F_71_4	sw1133C_A	71	Female	4	202053820063_R01C01

Sample Code	*Sample Name*	*Age*	*Gender*	*Cognitive Performance*	*Basename*
B03_F_63_1	sw1536C_A	63	Female	1	202053820063_R02C01
B01_F_66_4	sw1636C_A	66	Female	4	202053820031_R02C01
H03_F_69_4	sw1647C_A	69	Female	4	202053820063_R08C01
G02_F_69_4	sw1717C_A	69	Female	4	202053820039_R07C01
B04_M_64_1	sw1938C_A	64	Male	1	202053820069_R02C01
A02_F_69_1	sw2100C_A	69	Female	1	202053820039_R01C01
D03_F_52_1	sw2154C_A	52	Female	1	202053820063_R04C01
H06_M_68_1	sw2183C_A	68	Male	1	202060330094_R08C01
F03_F_57_1	sw2269C_A	57	Female	1	202053820063_R06C01
E05_M_58_1	sw2416C_A	58	Male	1	202060330086_R05C01
F02_F_71_1	sw2503C_A	71	Female	1	202053820039_R06C01
A01_F_60_4	sw2544C_A	60	Female	4	202053820031_R01C01
G05_M_69_4	sw2581C_A	69	Male	4	202060330086_R07C01
D05_M_76_4	sw2714C_A	76	Male	4	202060330086_R04C01
B05_M_58_1	sw2750C_A	58	Male	1	202060330086_R02C01
D02_F_72_4	sw2850C_A	72	Female	4	202053820039_R04C01
H05_M_63_4	sw2888C_A	63	Male	4	202060330086_R08C01
C02_F_68_4	sw2906C_A	68	Female	4	202053820039_R03C01
C05_M_74_4	sw2915C_A	74	Male	4	202060330086_R03C01
G01_F_75_1	sw2957C_A	75	Female	1	202053820031_R07C01
C06_M_74_1	sw3075C_A	74	Male	1	202060330094_R03C01
A04_M_57_1	sw3231C_A	57	Male	1	202053820069_R01C01
G03_F_54_4	sw3747C_A	54	Female	4	202053820063_R07C01
D06_M_52_4	sw3833C_A	52	Male	4	202060330094_R04C01

The microarray technology used with those samples is different from the one described in the last chapter. However, it has already been reported that *minfi* can be used with 850k arrays. For that reason, the pipeline used was similar to the one presented in the last chapter with only small changes that will be presented bellow.

Firstly, we decided to use the normalization by cell type since it influenced our results, when compared to quantile normalization. However, the strongest difference in the methodology adopted was on DMP finding step. In this case, both the cognitive performance and the age were taken as phenotypes of interest and were categorized into categorical and continuous phenotypes, respectively. In the case of the age as a phenotype of interest, the samples were separated in younger (51-70 years old) and older (>70 years old) individuals so that we could analyze the impact of age variation in methylation. After DMP calling, as usually, it is recommended to ajust the p-value of probes using the Bonferroni method. This, however, didn’t allow us to obtain significant DMPs for cognitive performance phenotype. For that reason this method was only used for the age phenotype.

In the case of cognitive performance phenotype, we only excluded the DMPs with a P-value above 0.05 and had the rest of them into consideration in the later analysis as suggestive, non-significant ones. We also did manhattan and volcano plots. (Box 12)

Box 12 - Script for the DMP finding step of Dataset C followed by its representation in volcano and manhattan plots. “Clusters” is the term used for designate cognitive performance phenotype

```
#find DMPs
beta <- getBeta(mSetSqFlt)
age <- pData(mSetSqFlt)$SW_Age
clusters <- pData(mSetSqFlt)$clusters
dmp_age <- dmpFinder(beta, pheno = age  , type = "continuous")
dmp_clusters <- dmpFinder(beta, pheno = clusters  , type = "categorical")

#deltabeta - clusters
good<-rownames(pData(mSetSqFlt))[ pData(mSetSqFlt)$clusters == 1]
bad<-rownames(pData(mSetSqFlt))[ pData(mSetSqFlt)$clusters == 4]
dmpCpgs = rownames(dmpfinal_clusters)
dmpfinal_clusters$good = rowMeans(beta[dmpCpgs, good, drop=F])
dmpfinal_clusters$bad = rowMeans(beta[dmpCpgs, bad, drop=F])
dmpfinal_clusters$deltaBeta = dmpfinal_clusters$bad - dmpfinal_clusters$good
ann850k = getAnnotation(IlluminaHumanMethylationEPICanno.ilm10b2.hg19)
annotation<-merge(dmpfinal_clusters, ann850k, by="row.names", all.x=TRUE)
annotation$log10<-NULL
annotation$log10<--(log10(annotation$pval))
annotation$chr <- sub("chr", "", annotation$chr)
annotation$chr <- as.numeric(annotation$chr)
bad_good<-annotation
write.table(bad_good, file="annotation_dmps_cluster.txt")

#volcano plot - clusters
pdf("volcanoplot_clusters.pdf")
with(bad_good, plot(bad_good$deltaBeta, bad_good$log10, pch=20, main=""))
abline(h = 5.0, col = "blue", lty = 2, lwd = 1)
abline(v = c(-0.1,0.1), col = "blue", lty = 2, lwd = 1)
with(subset(bad_good, bad_good$log10<5.0), points(deltaBeta, log10, pch=20,
col="gray"))
with(subset(bad_good, bad_good$deltaBeta< -0 & bad_good$log10>5.0),
points(deltaBeta, log10, pch=20, col="red"))
with(subset(bad_good, bad_good$deltaBeta> 0 & bad_good$log10>5.0),
points(deltaBeta, log10, pch=20, col="green"))
dev.off()

#manhattan plot-clusters
pdf("manhattan-clusters.pdf")
manhattan(annotation, chr="chr", bp="pos", p="pval", snp="Row.names",
col=c("grey","skyblue")) ##check the borderline
dev.off()
cluster <- as.numeric(pData(mSetSqFlt)$clusters)
one<-as.numeric(beta[rownames(beta)=="cg09592155", ]) ## the probe with most
significance in the clusters feature (checked corrected!)
pdf("dmp_clusters_cg09592155.pdf", onefile=T, paper="a4r")
boxplot(one~cluster)
dev.off()

#deltabeta - age
young<-rownames(pData(mSetSqFlt))[ pData(mSetSqFlt)$SW_Age >= 52 &
pData(mSetSqFlt)$SW_Age < 70]
old<-rownames(pData(mSetSqFlt))[ pData(mSetSqFlt)$SW_Age >= 70]
dmpCpgs = rownames(dmpfinal_age)
dmpfinal_age$young = rowMeans(beta[dmpCpgs, young, drop=F])
dmpfinal_age$old = rowMeans(beta[dmpCpgs, old, drop=F])
```

```
dmpfinal_age$deltaBeta = dmpfinal_age$old - dmpfinal_age$young
p_adjusted<-p.adjust(dmpfinal_age$pval, method="fdr") ##p value adjusted with
fdrmethod
dmpfinal_age<-cbind(dmpfinal_age, p_adjusted)
ann850k = getAnnotation(IlluminaHumanMethylationEPICanno.ilm10b2.hg19)
annotation_age<-merge(dmpfinal_age, ann850k, by="row.names", all.x=TRUE)
annotation_age$log10<-NULL
annotation_age$log10<--(log10(annotation$pval))
annotation_age$chr <- sub("chr", "", annotation$chr)
annotation_age$chr <- as.numeric(annotation$chr)
y_o<-annotation_age
write.table(y_o, file="annotation_dmps_age.txt")

#volcano plot - age
pdf("volcanoplot_age.pdf")
with(y_o, plot(y_o$deltaBeta, y_o$log10, pch=20, main=""))
abline(h = 5.0, col = "blue", lty = 2, lwd = 1)
abline(v = c(-0.1,0.1), col = "blue", lty = 2, lwd = 1)
with(subset(y_o, y_o$log10<5.0), points(deltaBeta, log10, pch=20, col="gray"))
with(subset(y_o, y_o$deltaBeta< -0 & y_o$log10>5.0), points(deltaBeta, log10,
pch=20, col="red"))
with(subset(y_o, y_o$deltaBeta> 0 & y_o$log10>5.0), points(deltaBeta, log10,
pch=20, col="green"))
dev.off()

#manhattan plot-age
pdf("manhattan-age.pdf")
manhattan(annotation_age, chr="chr", bp="pos", p="pval", snp="Row.names",
col=c("grey","skyblue")) ##check the borderline
dev.off()
theone<-as.numeric(beta[rownames(beta)=="cg16867657", ]) ##the probe with most
significance in the age feature (checked corrected!)
pdf("dmp_age_cg16867657.pdf", onefile=T, paper="a4r")
plot(age, theone, xlim=c(1, 100), ylim=c(0, 0.9))
abline(lm(theone ~ age), col="blue")
dev.off()
summary(lm(theone ~ age))$r.squared
```

In order to proceed to the validation of results, which was the main goal of this chapter, we merged Datasets B.3 and C, obtaining Dataset D (Box 13). These datasets were joined before quality control, normalization and filtering steps for each of the individual datasets and for that reason all the procedure was repeated for Dataset D. Cell-type normalization was used and in this case for DMP calling, the correction method of Bonferroni was used in order to select the significant probes affected by aging. Additionally, as before probes with a p-value > 0.05 were removed.

Box 13 – Script of merging of Dataset C and B.3 in order to obtain Dataset D, before normalization and filtering

```
#Joining Dataset C with Dataset B.3 before normalization and filtering
minho<-readRDS("minho/rgSet_minho.rds")
pData(minho)$age<-pData(minho)$SW_Age
pData(minho)$sex<-pData(minho)$Gender
pData(minho)$type<-c("MINHO")
pData(minho)$SW_Age<-NULL
pData(minho)$Gender<-NULL
pData(minho)$geo_accession<-pData(minho)$title
pData(minho)$title<-NULL
gse87571<-readRDS("GSE87571/rgSet_gse87571.rds")
```

```
pData(gse87571)$type<-c("GSE")
gse_minho<-combineArrays(gse87571, minho, outType="IlluminaHumanMethylationEPIC",
verbose=TRUE)

#remove unnecessary columns
pData(gse_minho)$characteristics_ch1<-NULL
pData(gse_minho)$characteristics_ch1.1<-NULL
pData(gse_minho)$characteristics_ch1.2<-NULL
pData(gse_minho)$characteristics_ch1.4<-NULL
pData(gse_minho)$characteristics_ch1.3<-NULL

#manipulate data to get all uniform
pData(gse_minho)$sex <- sub("Female", "F", pData(all_geo)$sex)
pData(gse_minho)$sex <- sub("Male", "M", pData(all_geo)$sex)
pData(gse_minho)$sex <- sub("Gender: ", "", pData(all_geo)$sex)
pData(gse_minho)$sex <- sub("gender: ", "", pData(all_geo)$sex)
pData(gse_minho)$sex <- sub("Sex: ", "", pData(all_geo)$sex)
pData(gse_minho)$sex <- sub("female", "F", pData(all_geo)$sex)
pData(gse_minho)$sex <- sub("male", "M", pData(all_geo)$sex)
pData(gse_minho)$sex <- sub("m", "M", pData(all_geo)$sex)
pData(gse_minho)$sex <- sub("f", "F", pData(all_geo)$sex)
pData(gse_minho)$age <- sub("\\.\\d+", "", pData(all_geo)$age)
saveRDS(gse_minho, file = "rgSet_gseMinho.rds")
pdf("densityplots_gse87571_minho.pdf")
par(xpd=NA,oma=c(3,0,0,0))
densityPlot(getBeta(gse_minho), sampGroups = pData(gse_minho)$sex)
densityPlot(getBeta(gse_minho), sampGroups = pData(gse_minho)$age, legend=FALSE)
dev.off()
```

3. RESULTS

In order to analyse this dataset, we obtained the manifest object that included information about the array, such as the probes used – Type I, II, Control, SNP Type I and SNP Type II. This manifest was the same for all the samples studied, since they were all analysed using EPIC technology.

Box 14 – Manifest object general information about Dataset C

```
## IlluminaMethylationManifest object
##Annotation
    array: IlluminaHumanMethylationEPIC
##Number of type I probes: 142262
##Number of type II probes: 724574
##Number of control probes: 635
##Number of SNP type I probes: 21
##Number of SNP type II probes: 38
```

3.1. Quality control, normalization and filtering

The results obtained for the dataset quality control were similar to the ones presented before. As shown in Figure 49, our dataset revealed a mean reliable signal across samples and we see a clustering of the good samples, as expected. As to the preprocessing and normalization step, we also started through the comparison of the normalization methods availale on *minfi* in order to choose the best one for the dataset. As said before, the funnorm method is useful for studies with large methylation differences across the phenotypes. However, the quantile normalization method demonstrated better results in the density plots (Figure 50) and so as before, together with cell-type normalization, from all variables and we decided to use this method, we decided to use this method.

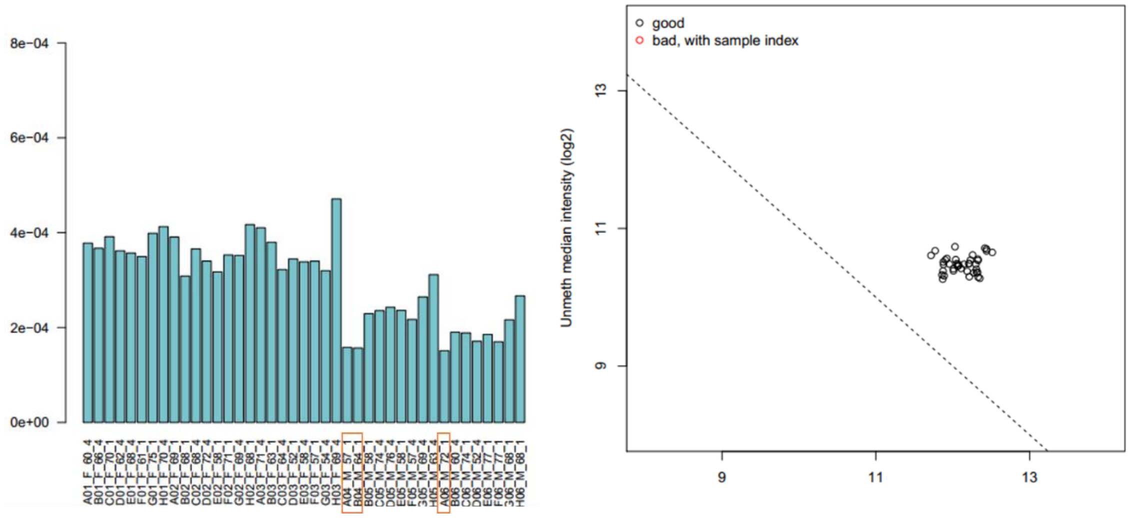


Figure 49 – Quality control plot with mean detection of p-values (y axis) per sample(x axis) (left) and the representation of both the log median intensity of methylated and unmethylated channels against each other (right). The left plot reveals a general quality of samples in terms of the overall signal reliability. The right plot presents the good samples clustering together.

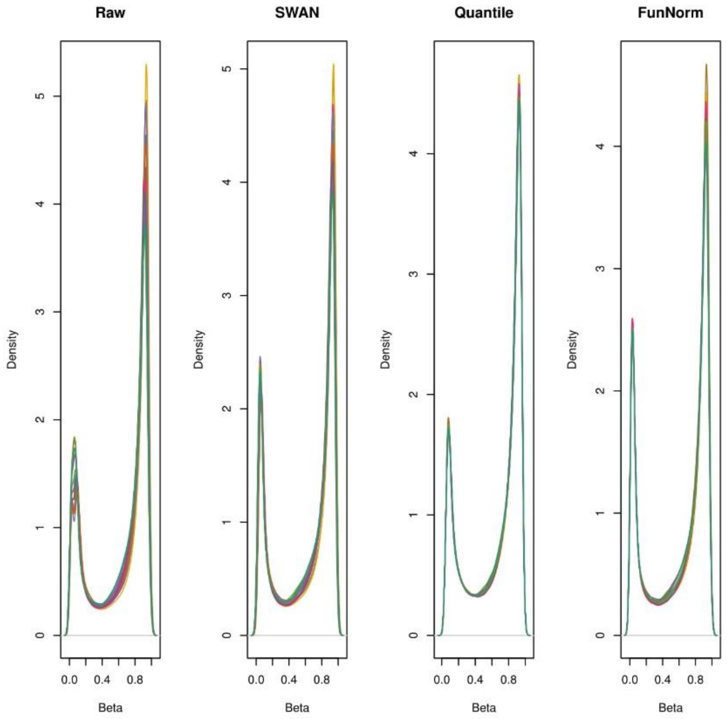


Figure 50 – Density plots of the three available normalization processes for the age phenotype. The best normalization was achieved with Quantile normalization which shows a most uniform plot.

Afterwards, we carried out a statistical analysis that evidenced the need of a filtration step, as in the last chapter. Next we carried out a filtering step for removal of sex chromosomes probes, probes with p-values above 0.01, probes that failed in one or more samples, probes with SNPs in CpG sites and cross-reactive probes. This resulted in 21584 probes remained. The MDS and the clustering plot using the filtered data confirmed that gender was no long a variable affecting our results (Figure 51), and so we proceed with the protocol.

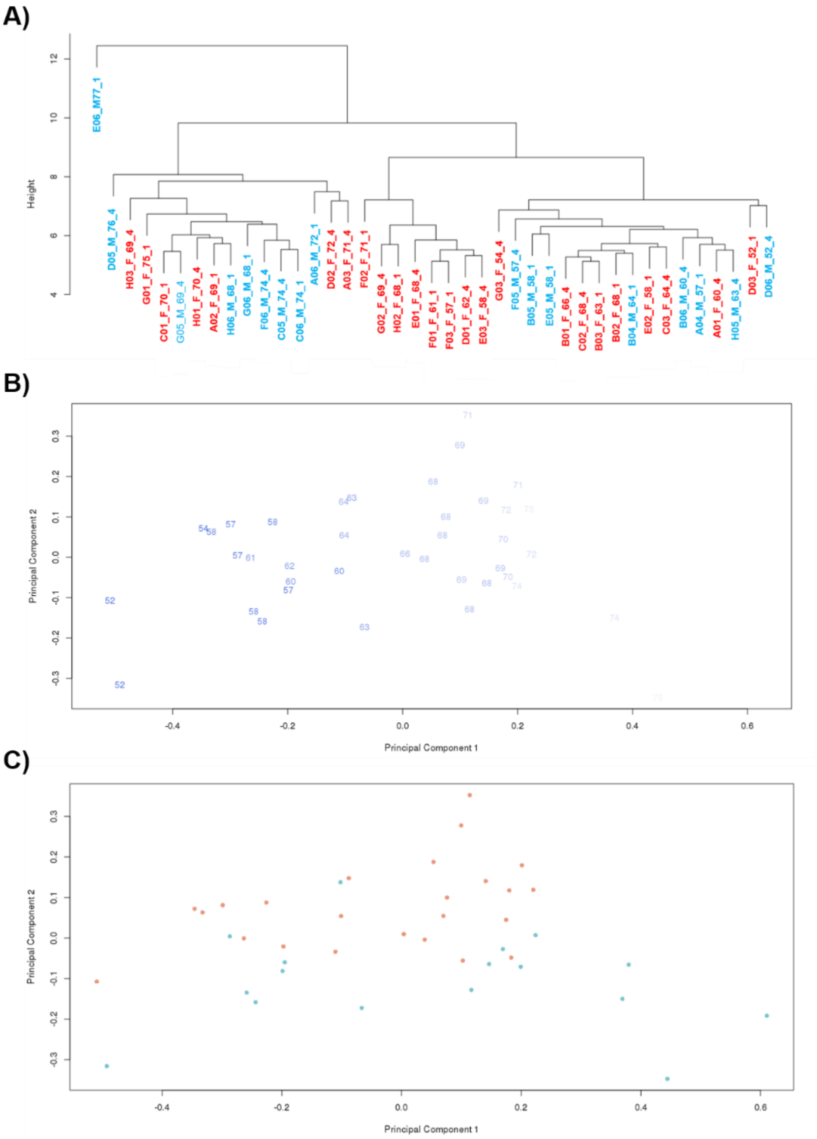


Figure 51 – Clustering (A) and MDS plots of age (B) and gender (C) of Dataset C after the normalization and filtration procedures. The gender effect is no longer visible, since samples are mixed, which means the dataset is ready to proceed with the analysis.

3.2. Differential methylation analysis

The dataset was separated in young (52-70 years old) and old (>70 years old) individuals in order to make a differential methylation analysis comparing both phenotypes. This resulted in 31 young and 10 old individuals. We compared the variation of global methylation in old (purple) and young (blue) phenotypes across chromosomes and infered that the most significant variations could be associated with chromosomes 1,5,10, 11 and15 (Figure 52A). A density plot across both phenotypes was also performed and that didn't showed a global variation between the autosomal methylation levels in both phenotypes although the methylated sites constituted about twice of the unmethylated ones (Figure 52B).

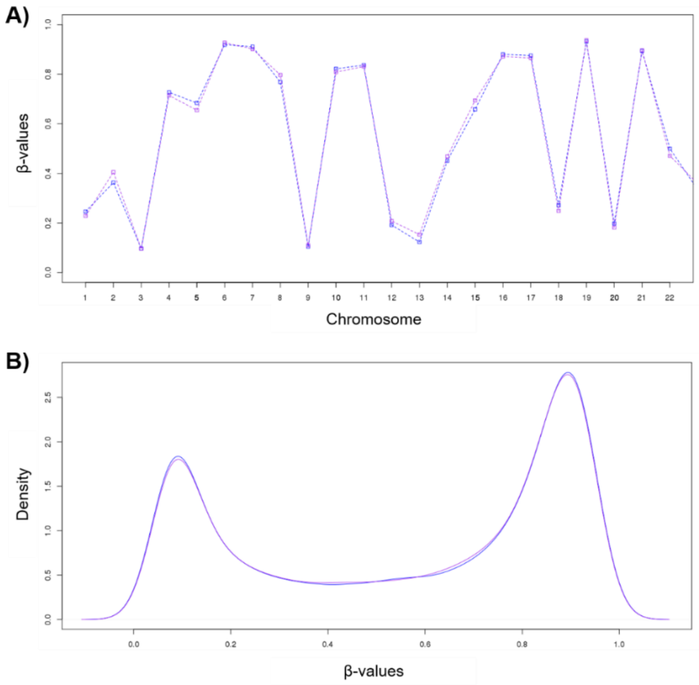


Figure 52 – Global comparison of methylation between young and old individuals (A) Methylation across chromosomes; (B) Density plot. The highest differences are located in chromosomes 1, 5, 10, 11 and 15 and the methylated sites constitute about twice of the unmethylated ones.

We identified a total of 21584 a-DMPs without p-value cutoff and represented them according to the negative logarithm of detection p-value and chromosome positions (manhattan plot) or variation of β-values (volcano plot) (Figure 53). Of the total of probes, 74% of our probes were gene associated, 21% were enhancer associated and 8% were DMR associated. Considering that we have 850k probes in our array and that a significant adjusted p-value should be inferior to 0.05, we conclude that our significant probes should have a raw p-value inferior to 5.88 x 10^{-8} and suggestive probes should have a raw p-value inferior to 1x10^{-5}, that corresponds to a -$\log_{10}$(raw p-value) of 5. In view of this, we conclude that there weren't identified significant probes in aging, but only suggestive ones.

In the analysis of the volcano plots we identified the positive or negative β-value variations across DMPs in the case of young or old individuals. From these, we selected 6 suggestive DMPs relative to age, 4 with an increase of methylation and 2 with a decrease of methylation. All of these markers had a $|\Delta\beta|<10\%$ which corresponds to a reduced overall effect. The information for all of the suggestive probes is presented in Supplementary Table 4. In the case of the age phenotype, the most suggestive DMP was cg06639320 located in chromosome 2 (adjusted p-value = 0.00418) with an increase of 2.5% in methylation across age. It was located in the promoter of FHL2 gene, particularly 200 nucleotides distant from the TSS, but also with annotations relating it to the 5' UTR. cg16867657 (adjusted p-value = 0.00747) was the second most suggestive probe of our dataset, located in chromosome 6, with an increase of 0.9% of methylation rate across age and located in TSS1500 of the ELOVL2 gene.

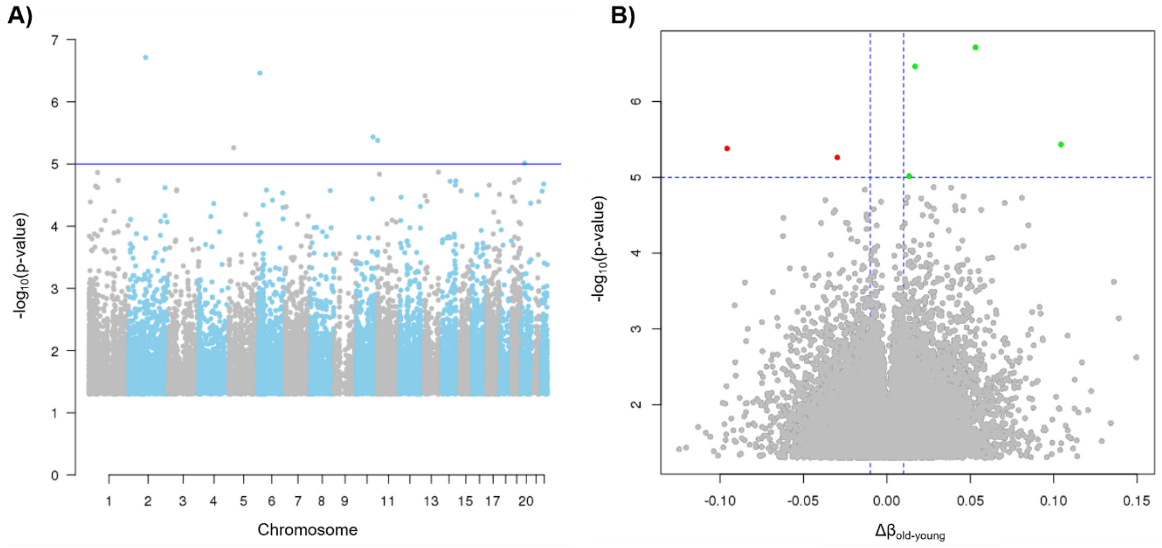


Figure 53 – Representation of all DMPs found with age in Dataset C in a (A) manhattan plot with a selected threshold of p-value = 1x 10^{-5}; and in a (B) volcano plot using a selected threshold of p-value = 1x10-5 and |Δβ|>0.2.

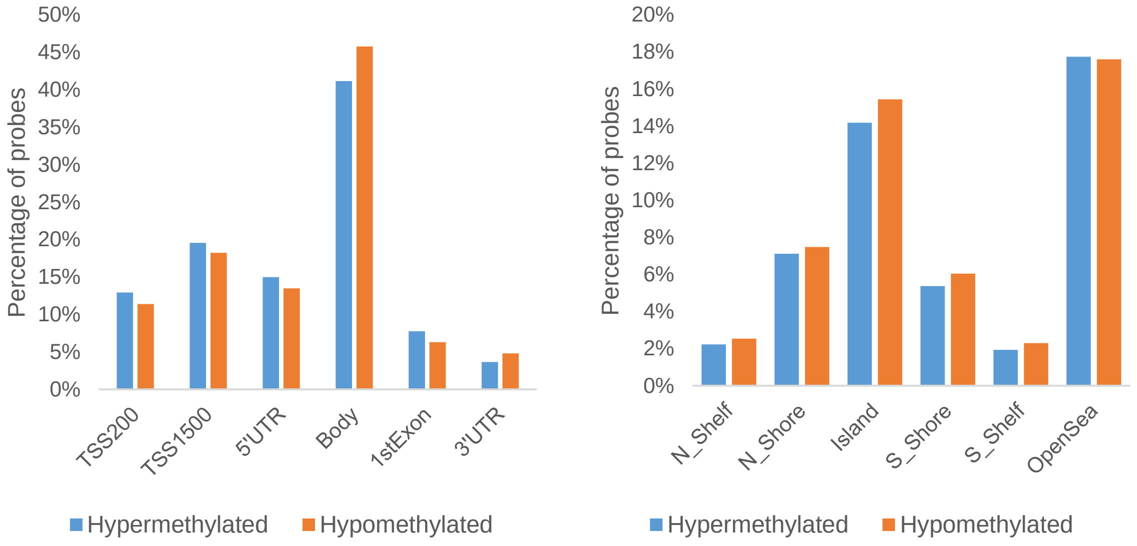


Figure 54 – Distribution of DMPs of Dataset C according to its relation with genomic CpG location (left) and gene part (right). There was a preference for DMPs located in body of genes and open sea.

The marker genomic location was analysed (Figure 54) and, similarly to the described in the last chapter, the markers for both phenotypes had a similar behavior either if they were hypo or hypermethylated. However, in Dataset C the probes in open sea represent the majority of the cases, with 35% of markers, followed by the island location, with 30% of markers and a mean of 1.23 markers per island. In the case of gene position, the body location is still the most preferred localization on 44% of probes, with 46% and 41% hypermethylated and hypomethylated.

3.3. Validation of data

3.3.1. The effect of blood cell population

When Datasets B.3 and C were merged, resulting in Dataset D, the respective dendogram (Supplementary Figure 18) revealed a strong clustering of samples according to its dataset origin. These findings can be justified by different tissue-specificities of samples, the different laboratory conditions or procedures used or even different ancestry. However, as said before, the cell-specificity in blood was already reported as a major source of biases in methylation among individuals and could also be, in our dataset, the major source of bias. For that reason, we estimated the cell type composition for all blood samples using the function `estimatecellcounts` of *minfi* (Table 15 and Supplementary Figure 19) and concluded that the Dataset B.3 had a huge percentage of granulocytes followed by CD4T lymphocyte, while the Dataset C had mostly CD4T and NK lymphocytes. Since there was a dissimilarity between both datasets that influenced our results, another normalization step, apart from quantile, was performed, based on cell-type composition.

Table 15 – Estimation of the average percentage of each type of blood cell for both datasets

Type of cells	*Dataset B.3*	*Dataset C*
CD8 T Lymphocyte	7,66%	9.65%
CD4 T Lymphocyte	14,05%	36.41%
NK Lymphocyte	9,47%	27.14%
B Lymphocyte	5.00%	9.06%
Leukocyte monocyte	8.10%	17.09%
Leukocyte granulocyte	56.52%	1.12%

Additionally and since our validation samples had information about cell composition as measured by flow cytometry, we estimated the cell composition of this dataset and compared the results with the provided results. According to Figure 55, it is visible that the number of cells estimated by us is very well correlated with provided data and therefore represents the real cell composition of samples.

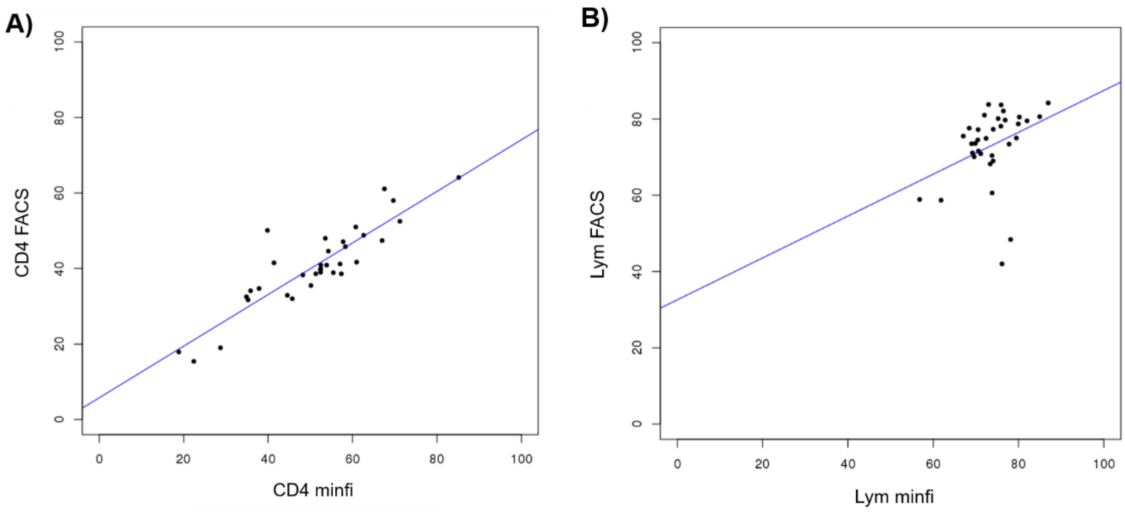


Figure 55 – Comparison of cell counts using *minfi* and flow cytometry (A) for CD4 and for (B) lymphocytes.

3.3.2. Differential methylation analysis

After normalization and filtering, the analysis of MDS plots allowed us to infer that normalization removed all biases, and so we initiated the differential methylation analysis. For this, we started by a Pearson correlation analysis, similarly to before, in which we evidenced an increasing of methylation with age in the median quantile (R = 0.4456, p-value $< 2.2x10^{-16}$) representing the most significant difference, followed by a decrease of methylation in the third quantile (R = -0.2024, p-value = $5.918x10^{-08}$) and an increase of methylation in the first one (R = 0.1808, p-value = $1.357x10^{-06}$).

Similarly to the processes adopted before, the dataset was separated in young (18 – 32 years old) 141 and old (> 70 years old) 152 individuals with which we performed a global comparison using density plots and chromosome plots (Figure 56). Here, the number of methylated and unmethylated sites was similar but there was a reduction in global methylation of old individuals in comparison to young individuals. Additionally, chromosomes 1, 3, 4, 7, 8, 11, 14, 15, 18 and 21 evidenced higher differences in global methylation across both phenotypes.

We identified 70422 significant a-DMPs that were represented in manhattan and volcano plots (Figure 57). From them, 75% were gene associated, 22% were enhancer associated and 13% were DMR associated. Through this analysis we 10 top-significative DMPs, 6 of them positive correlated with age (> 10% variance) and 4 negative correlated with age (3 with > 10% variance and 1 with <10% variance). (Table 16)

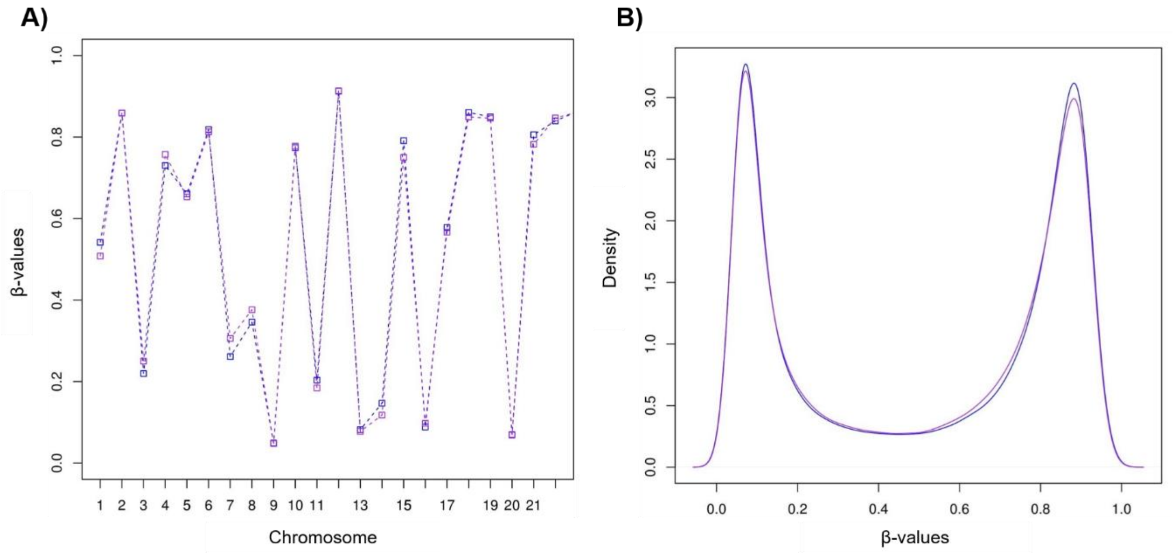


Figure 56 – Global comparison of methylation between young (blue) and old (purple) individuals of Dataset D (A) Methylation across chromosomes; (B) Density plot. The highest differences were located in chromosomes 1, 3, 4, 7, 8, 11, 14, 15, 18 and 21 and the number of methylated probes in the old individuals was smaller.

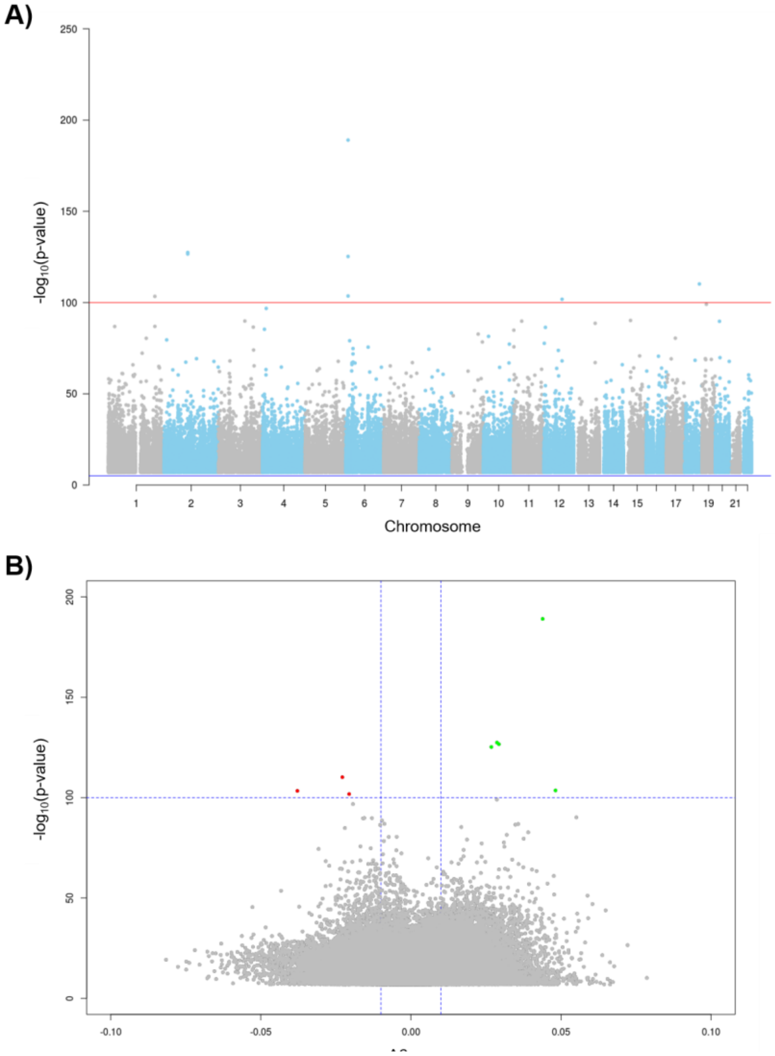


Figure 57 – Representation of all DMPs found with age in Dataset D in a (A) manhattan plot a selected threshold of p-value = 1x10-150; and in a (B) volcano plot using a selected threshold of p-value = 1x10-50 and |Δβ|>0.2.

Table 16 – Summary of the top-signiticant DMPs with age phenotype in Dataset D. Probes are ordered by statistical significance – 6 are methylation gains while 4 are losers.

Probes	*slope*	*P_{adj}-value*	*Q-value*	*Δβ*	*Chr*	*Position*	*Islands Name*	*Relation to Island*	*Gene Group*	*Gene Name*	*Gene Function*	*Log10*
cg16867657	0.00435	3.19E-184	9.25E-185	0.0439	6	11044877	chr6:11043913 -11045206	Island	TSS1500	ELOVL2	Fatty acid elongase activity	189.042
cg06639320	0.00328	1.29E-122	1.87E-123	0.0287	2	106015739	chr2:106014878 – 106015884	Island	TSS200; 5'UTR	FHL2	Protein binding	127.436
cg22454769	0.00341	8.75E-122	8.46E-123	0.0293	2	106015767	chr2:106014878 – 106015884	Island	TSS200 5'UTR	FHL2	Protein binding	126.603
cg21572722	0.00222	2.02E-120	1.47E-121	0.0268	6	11044894	chr6:11043913 - 11045206	Island	TSS1500	ELOVL2	Fatty acid elongase activity	125.239
cg19283806	- 0.00262	2.32E-105	1.34E-106	-0.0229	18	66389420		OpenSea	5'UTR	CCDC102B	Protein binding	110.181
cg24724428	0.00404	9.58E-99	4.63E-100	0.0482	6	11044888	chr6:11043913 - 11045206	Island	TSS1500	ELOVL2	Fatty acid elongase activity	103.564
cg10501210	- 0.00477	1.55E-98	6.41E-100	-0.0378	1	207997020		OpenSea				103.356
cg11649376	- 0.00252	5.82E-97	2.11E-98	-0.0206	12	81473234	chr12:81471569 – 81472119	S_Shore	Body	ACSS3		101.780
cg07544187	0.00308	3.05E-94	9.84E-96	0.0286	19	19651235	chr19:19650683 - 19651274	Island	Body	CILP2		99.061
cg08262002	- 0.00258	5.89E-92	1.71E-93	-0.0193	4	16575323		OpenSea	Body	LDB2		96.775

As expected, the most significant probe was the same detected in Dataset B.3, cg16867657 (Δβ = 4%, adjusted p-value = 3.1865x10^{-184}), in promoter of ELOVL2, followed by cg06639320 (Δβ = 3%, adjusted p-value = 1.2855x10^{-122}) and cg22454769 (Δβ = 3%, adjusted p-value = 2.0221x10^{-120}), in promoter of FHL2 gene. The two most significant probes were represented in a linear regression confirming its hypermethylation across age, although the probe associated with ELOVL2 demonstrates a higher methylation rate than FHL2 (Figure 58). Through the analysis of Table 17, that shows a Pearson correlation for each of the most relevant probes, we observe that in both probes the first quartile of methylation level has a negative correlation with age but the third quantile of methylation level reveals a positive correlation with age.

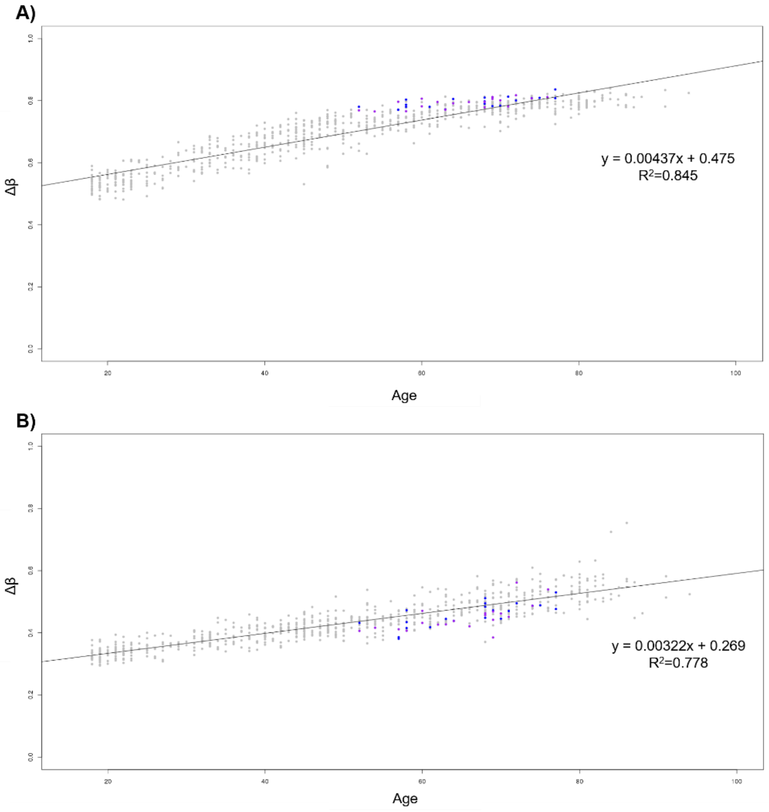


Figure 58 – Representation of the DNA methylation level of the two most significant probes of Dataset D. (A) cg16867857; (B) cg06639320

Table 17 - Pearson correlation to the most significant probes identified in Dataset D

	cg16867657	*cg06639320*
1Q	-0.024503	-0.021897
Median	0.000551	-0.002873
3Q	0.025945	0.018746
Max	0.092927	0.207215
p-value	< 2.2e-16	< 2.2e-16

A comparison between hypo and hypermethylated sites and genomic CpG region and gene part localization was then made (Figure 59). Differently from the results from previous chapter, where no major variance between hypo and hypermethylated sites for the same location was detected, in this case we can observe that in CGI and open sea positions the difference between hypo and hypermethylated sites is 4% and 8%, respectively. In the case of gene part localization, there was a difference between the hypo and hypermethylated sites in body of genes of about 10%. However, the top positions remained the same, with a prevalence of 42% of the probes identified inside CGI and 38% inside gene bodies. Also, 34% and 40% of the hyper and hypomethylated probes, respectively, in our study, were located in gene body and that there were an average of 1.51 markers per island.

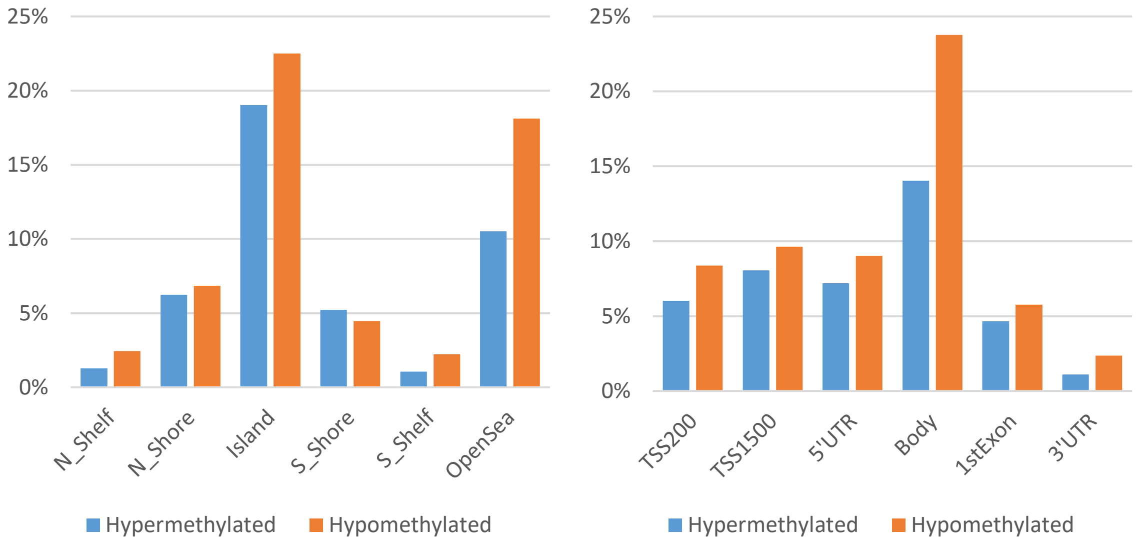


Figure 59 – Distribution of DMPs of dataset D according to its relation with the genomic CpG region (left) and gene part (right). There was a prevalence of DMPs located in CGI antibody of genes.

3.3.3. Gene ontology analysis

The protocol used for gene ontology analysis was the same used for Dataset B.3 and the test used was PANTHER Overrepresentation SLIM test either for biological process, molecular function or cellular components. Table 18 shows gene functions associated with the hyper and hypomethylated phenotypes and reveals a good similarity with the data from Dataset B.3.

Table 18 – Gene Ontology enrichment analysis for biological process, molecular function and cellular components in hypo and hypermethylated genes of Dataset B.3. GO terms are ordered by hierarchy and statistical significance.

	Category	*Function*	*UR / OR*	*FE*	*Raw p-value*
Hypermethylation	Biological function	Digestive tract mesoderm development (GO:0007502)	OR	18.64	3.83E-04
		Segment specification (GO:0007379)	OR	9.08	1.98E-05
		Ectoderm development (GO:0007398)	OR	8.02	4.10E-12
		Embryo development (GO:0009790)	OR	7.35	3.01E-05
		Synaptic vesicle exocytosis (GO:0016079)	OR	7.33	3.78E-03
		Pattern specification process (GO:0007389)	OR	7.02	3.93E-05
		Anatomical structure morphogenesis (GO:0009653)	OR	6.53	1.57E-05
		Mesoderm development (GO:0007498)	OR	6.06	1.95E-09
		Neurotransmitter secretion (GO:0007269)	OR	4.47	2.83E-02
		Cell-cell adhesion (GO:0016337)	OR	4.02	1.30E-02
		Developmental process (GO:0032502)	OR	3.49	1.12E-18
		Nervous system development (GO:0007399)	OR	3.38	1.11E-03
		System development (GO:0048731)	OR	3.13	2.78E-04
		Synaptic transmission (GO:0007268)	OR	2.97	2.57E-03
		Regulation of transcription from RNA polymerase II promoter (GO:0006357)	OR	2.84	2.25E-04
		Cell differentiation (GO:0030154)	OR	2.71	9.90E-04
		Celular component morphogenesis (GO:0032989)	OR	2.68	6.37E-03
		Cell-cell signaling (GO:0007267)	OR	2.37	1.05E-02
		Transcription from RNA polymerase II promoter (GO:0006366)	OR	2.29	2.42E-03
		Transcription, DNA-dependent (GO:0006351)	OR	2.01	7.35E-03
		Neurological system process (GO:0050877)	OR	1.99	1.13E-02
		Multicellular organismal process (GO:0032501)	OR	1.98	2.36E-04
		Single-multicellular organism process (GO:0044707)	OR	1.96	4.01E-04
		System process (GO:0003008)	OR	1.88	2.65E-02
		Catabolic process (GO:0009056)	UR	.36	4.29E-02
	Cellular component	Neuron projection (GO:0043005)	OR	5.04	3.39E-06
		Dendrite (GO:0030425)	OR	4.77	1.49E-02
		Extracelular matrix (GO:0031012)	OR	3.57	4.71E-02
		Cell projection (GO:0042995)	OR	3.44	1.72E-04
		Intracellular (GO:0005622)	UR	.70	3.67E-02
		Macromolecular complex (GO:0032991)	UR	.46	1.98E-02
	Molecular function	G-protein coupled receptor activity (GO:0004930)	OR	3.44	2.57E-03
		Sequence-specific DNA binding transcription factor activity (GO:0003700)	OR	3.08	1.49E-06
		DNA binding (GO:0003677)	OR	2.29	5.43E-04

	Category	Function	UR/OR	FE	Raw p-value
	Molecular function	Nucleic acid binding (GO:0003676)	OR	1.83	6.85E-03
		Binding (GO:0005488)	OR	1.53	2.23E-04
Hypomethylated	Cellular component	Early endosome membrane (GO:0031901)	OR	6.40	5.32E-03
		Early endosome (GO:0005769)	OR	4.14	7.91E-03
		Endosome (GO:0005768)	OR	2.39	4.64E-02
		Cytosol (GO:0005829)	OR	1.48	4.04E-02
		Nucleus (GO:0005634)	OR	1.38	2.78E-02
		Intracelular membrane-bounded organelle (GO:0043231)	OR	1.29	7.01E-03
		Cytoplasmic part (GO:0044444)	OR	1.29	3.79E-02
		Membrane-bounded organelle (GO:0043227)	OR	1.25	5.92E-03
		Intracelular organelle (GO:0043229)	OR	1.25	1.15E-02
		Organelle (GO:0043226)	OR	1.21	1.03E-02
		Organelle (GO:0044424)	OR	1.21	1.08E-02
		Intracellular (GO:0005622)	OR	1.19	8.60E-03
		Cell (GO:0005623)	OR	1.12	4.14E-02
		Celular_component (GO:0005575)	OR	1.08	4.85E-02

UR – Underrepresentation | OR – Overrepresentation | FE – Fold Enrichment

3.4. Cognitive performance phenotype

As refered before, the samples from in Dataset C included information about the good and bad cognitive performance of individuals. For that reason, that additional information was used in our research in order to evaluate the differential methylation in that phenotype. The procedure adopted was the same described for aging.

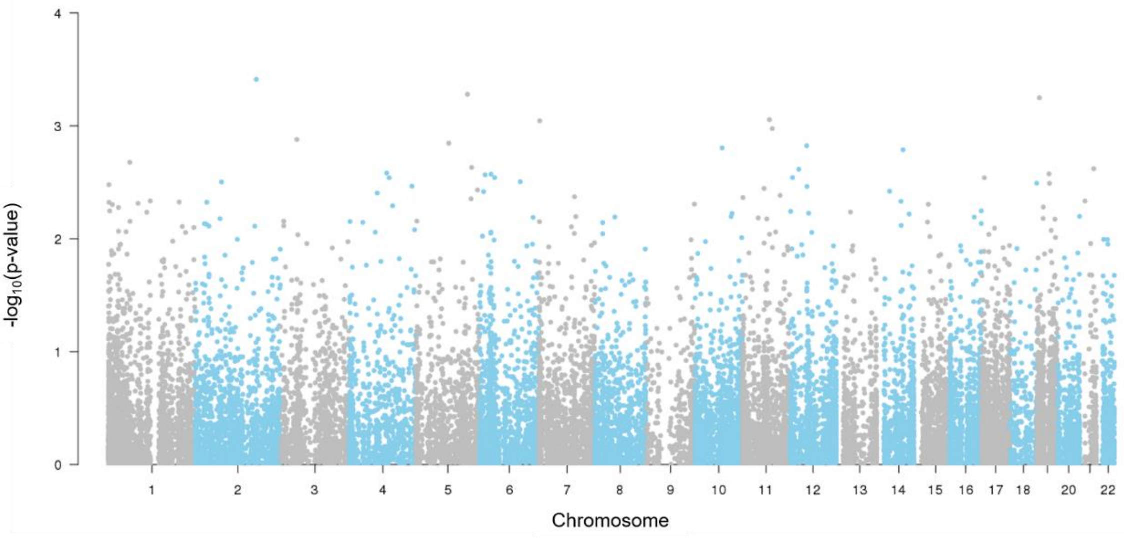


Figure 60 – Manhattan plot of DMPs found for cognitive performance phenotype in Dataset C. There were not significative neither suggestive probes identified.

In the DMP finding step, 21584 probes were identified (Figure 60). According to the explained before, the significance and suggestive raw p-values for Dataset C needed to be lower than 5.88x10^{-8} and 1x10^{-5}, respectively. In view of this, and since the probe with the lower p-value was cg05756320 with a raw p-value of 3.87x10^{-4}, we conclude that there weren't identified suggestive neither significant probes in the cognitive performance phenotype and for that reason the cognitive performance phenotype was not used in further data exploration.

4. DISCUSSION

The usage of an independent dataset, treated with the same experimental protocol is essential to validate the methodology as well as the results and to reach effective conclusions. For that reason, in this chapter we used a Portuguese cohort with 41 healthy individuals with ages between 52 and 77 years that have previously been characterized in order to validate the results from last chapter.

In the case of aging, in Dataset C, since the age range was more reduced, we only expected to observe the probes that showed differences even between close ages across age and that demonstrated differences between middle-age and elderly people. In comparison to Dataset B.3, we also identified cg06639320, associated to the FHL2 gene and cg16867657, associated to the ELOVL2. cg12662887, located in S_Shore of chr10:105344173-105345039 island affecting NEURL gene, was also identified. NEURL gene, also called NEURL1, encodes for neutralized E3 ubiquitin protein ligase 1 that plays a role in hippocampal-dependent synaptic plasticity, learning and memory. This gene has also been demonstrated in causing apoptosis and downregulating Notch target genes in medulloblastoma (Teider et al. 2010) and together with other three genes was shown to predict a prognosis of non-metastatic renal cell carcinoma(Van Vlodrop et al. 2017). However, it was not reported as a methylation marker yet although it has been identified in one study as a cancer-related gene that showed methylation differences between children and >10 years old individuals (Numata et al. 2012). cg12662887 was also identified in the Dataset B.3 with an adjusted p-value = 2.0714x10^{-9} as well as NEURL gene isoforms, like NEURL1B, NEURL2, NEURL3 and NEURL4. Looking into the most influenced and reported genes during aging – ELOVL2, FHL2, CCDC102B, ZNF423, ASPA, PDE4C and C1orf132 – we observed that Dataset C included all of them except C1orf132, but all with a p-value extremely high with the exception of FHL2 and ELOVL2 that remained in the top of the table as refered before. MARCH11, one of the significant genes identified in Dataset B.3, was absent from this validation.

In Dataset D, the significant probes were similar to the ones found in Dataset B.3, which was expected due to the merging of both datasets, however cg12662887 (NEURL) and cg06782035 (MARCH11) were not found in Dataset D. Through this we concluded that the most conservative positions among aging of healthy individuals may be cg16867657 and cg06639320 that were both positively correlated with age. The presence of these probes in all datasets made us conclude that they seem to be good markers of age since they were present in a larger cohort but also in a cohort limited in the range of ages and in the number of individuals. Figure 61 makes a comparison of the most important probes of our study and its behavior in all datasets were they were identified. Indeed, although with variable slopes, it is relevant that all show similar trends.

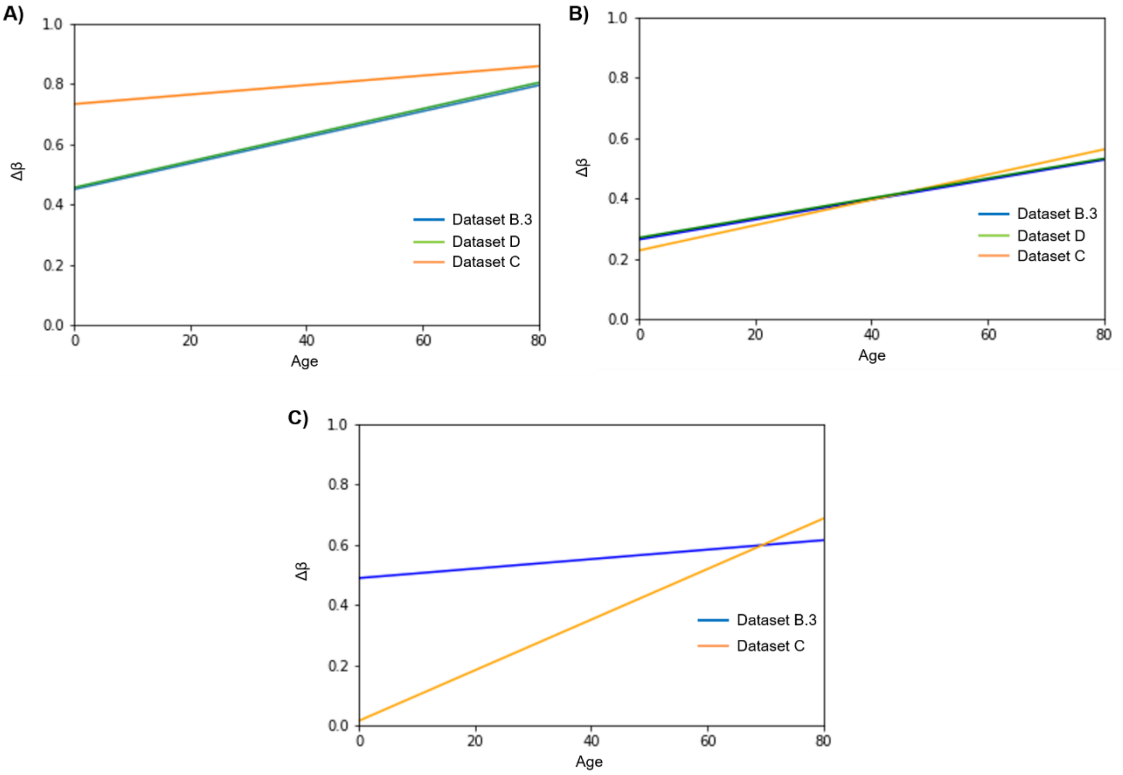


Figure 61 – Comparison of the linear regressions obtained for cg16867657 (ELOVL2) (A), cg06639320 (FHL2) (B) and cg12662887 (NEURL) (C) in the several datasets.

About marker genomic locations, the main difference between Dataset C, analysed with 850k array and Dataset B.3, analysed with 450k array, was the most frequent genomic position of markers – open sea for Dataset C and CGI for Dataset B.3. Even though, the percentage of probes in body of genes in both datasets was same and the ones CGI and gene body were similar.

As to Gene Ontology analysis, the majority of hypermethylated islands were associated with similar functions that were reported for Dataset B.3. The absence of enrichment for hypomethylated regions remain an issue. As mentioned before, these results were expected since Dataset B.3 was included on Dataset D. However, in an ideal situation the number of individuals in each of the cohorts would be closer and its age distribution and the number of identified probes would have the same ranges so that each could be studied separately and compared. This would only be possible if the datasets were similar in all the categories – ancestry, healthy stage, number of samples, type of tissue, laboratory conditions and methodologies used – which was not the case. For all of these reasons, our work is according to the described in the literature for methylation in aging. However, in order to improve the detection of NEURL and MARCH11 as markers of age it would be necessary to repeat the protocol using the conditions refered above.

Although the selected individuals were healthy, the differences between cognitive performance of both groups were expected to show corresponding variations in their methylome, which justified the differential methylation analysis, with a big focus on DMP finding. Since the statistical power of our test only resulted in adjusted p-values above 0.05, there weren't found significantly neither suggestive DMPs in cognitive performance.

Altogether, we can conclude that the bioinformatical pipeline allowed us to explore the methylome of both datasets with superimposition of results, in spite of the identified drawbacks. Therefore, the main goal of the project was accomplished and a new methodology for epigenetical studies become available for future research at iBiMED.

FINAL REMARKS

Through the usage of *in silico* experiments based on microarray data from public databases, we were able to take valid conclusions about the variations of methylome across the aging process in healthy individuals. The most common and cited markers of age were identified in the scope of this thesis and ELOVL2 and FHL2 have still shown its activity even in a dataset with a lower range of ages. Additionally, genes like MARCH11 and NEURL demonstrated to be differential methylated with age, plus the last one was present in both datasets.

However, the small-size of our validation dataset was the biggest challenge in this project. For that reason, the validation of our conclusions was not performed as the expected and it wasn't possible to take conclusions about the cognitive performance of individuals. On the other hand, the initial curation process was hampered by the lack of information or the incorrect annotations published by researchers or even the non-user friendly platform of NCBI. This forced us to perform a manual study for each sample which was very time-consuming and in the case of Next-Generation Sequencing with no payback.

For all of these reasons, if it is pretended to continue the *in-silico* experiments in iBiMED it is essential to create an in-house database to store the data and improve the research of interesting samples. On the other hand, the number of needed samples in order to have sufficient power to detect a meaningful difference in DNA methylation patterns needs to be estimated. For that purpose, there are several available softwares like G*Power and Quanto that should be used before the definition of the number of need samples in a test. Through this determination, biological samples can be used with microarray technologies and the differential methylation patterns will be more conclusive. The study of blood as a most accessible tissue and its comparison with other tissues methylome, like brain and liver, is also an interesting goal in order to determine their similarities and to improve the usage of blood as an accessible source of information for this type of studies.

Finally, the study of methylome is promising and since it is influenced by the environment of individuals, it is essential to explore this topic in the iBiMED. Through this work, the institute developed a methylation microarray pipeline that can be used in the near future to deepen methylome studies even in healthy or diseased individuals.

BIBLIOGRAPHIC REFERENCES

A.D. Riggs, 1975. X inactivation, differentiation, and DNA methylation. *Cytogenetics and cell genetics*, 14, pp.9–25.

Agarwala, R. et al., 2016. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research*, 44(D1), pp.D7–D19.

Ansorge, W.J., 2009. Next-generation DNA sequencing techniques. *New Biotechnology*, 25(4), pp.195–203.

Antequera, F. & Bird, A., 1993. Number of CpG islands and genes in human and mouse. *Proceedings of the National Academy of Sciences*, 90, pp.11995–11999.

Aryee, M.J. et al., 2014. Minfi: A flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. *Bioinformatics*, 30(10), pp.1363–1369.

Babraham Bioinformatics, 2016. Bismark Bisulfite Mapper – User Guide -v0.15.0.

Babraham Bioinformatics, 2010. FASTQC. *Documentation*.

Babraham Bioinformatics, 2013. Reduced Representation Bisulfite-Seq – A Brief Guide to RRBS. Available at: papers3://publication/uuid/A5B88C00-5F40-41F6-8A67-57BD85491A01.

Bacalini, M.G. et al., 2017. Systemic Age-Associated DNA Hypermethylation of ELOVL2 Gene: In Vivo and in Vitro Evidences of a Cell Replication Process. *Journals of Gerontology - Series A Biological Sciences and Medical Sciences*, 72(8), pp.1015–1023.

Barski, A. et al., 2007. High-Resolution Profiling of Histone Methylations in the Human Genome. *Cell*, 129, pp.823–837.

Bell, J.T. et al., 2012. Epigenome-Wide Scans Identify Differentially Methylated Regions for Age and Age-Related Phenotypes in a Healthy Ageing Population. *PLoS Genetics*, 8(4).

Bewerunge-Hudler, M. et al., 2014. Cross-sectional and longitudinal changes in DNA methylation with age : an epigenome-wide analysis revealing over 60 novel age-associated CpG sites. *Human Molecular Genetics*, 23(5), pp.1186–1201.

Bibikova, M. et al., 2011. High density DNA methylation array with single CpG site resolution. *Genomics*, 98(4), pp.288–295.

Bibikova, M. & Fan, J.-B., 2010. Genome-wide DNA methylation profiling. *Wiley Interdisciplinary Reviews: Systems Biology and Medicine*, 2(2), pp.210–223.

Bioconductor, Bioconductor. Available at: https://www.bioconductor.org/ [Accessed April 30, 2018].

Bird, A., 2002. DNA methylation patterns and epigenetic memory. *Genes & Development*, 16, pp.6–21.

Bock, C., 2012. Analysing and interpreting DNA methylation data. *Nature Reviews Genetics*, 13, pp.705–719.

Bollati, V. et al., 2010. Decline in Genomic DNA Methylation through Aging in a Cohort of Elderly Subjects. *Mechanisms of Ageing and Development*, 130(4), pp.234–239.

Borevitz, J.O. et al., 2015. Genomic variation across landscapes: insights and applications. *New Phytologist*, 207, pp.953–967.

Bumgarner, R., 2013. Overview of dna microarrays: Types, applications, and their future. *Current Protocols in Molecular Biology*, (SUPPL.101), pp.1–11.

Callinan, P.A. & Feinberg, A.P., 2006. The emerging science of epigenomics. *Human Molecular Genetics*, 15(1), pp.95–101.

Chen, Y.A. et al., 2013. Discovery of cross-reactive probes and polymorphic CpGs in the Illumina Infinium HumanMethylation450 microarray. *Epigenetics*, 8(2), pp.203–209.

Clark, S.J. et al., 1994. High sensitivity mapping of methylated cytosines. *Nucleic Acids Research*, 22(15), pp.2990–2997.

Cluny, V., 2016. *Exploratory study of age related to epigenomic patterns*.

Cock, P.J.A. et al., 2009. The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. *Nucleic Acids Research*, 38(6), pp.1767–1771.

Daniel, C., Lagergren, J. & Öhman, M., 2015. RNA editing of non-coding RNA and its role in gene regulation. *Biochemie*, 117, pp.22–27.

Dedeurwaerder, S. et al., 2011. Evaluation of the Infinium Methylation 450K technology. *Epigenomics*, 3(6), pp.771–784.

Dijk, E.L. van, Jaszczyszyn, Y. & Thermes, C., 2014. Library preparation methods for next-generation sequencing: Tone down the bias. *Experimental Cell Research*, 322(1), pp.12–20.

Djebali, S. et al., 2012. Landscape of transcription in human cells. *Nature*, 489, pp.101–108.

Edgar, R., 2002. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Research*, 30(1), pp.207–210.

El-Metwally, S. et al., 2013. Next-Generation Sequence Assembly: Four Stages of Data Processing and Computational Challenges. *PLoS Computational Biology*, 9(12).

Escalante, A.E. et al., 2014. The study of biodiversity in the era of massive sequencing. *Revista Mexicana de Biodiversidad*, 85(4), pp.1249–1264.

Esteller, M., 2002. CpG island hypermethylation and tumor suppressor genes : a booming present, a brighter future. *Oncogene*, 21, pp.5427–5440.

Falahi, F., Sgro, A. & Blancafort, P., 2015. Epigenome engineering in cancer: fairytale or a realistic path to the clinic? *Frontiers in oncology*, 5(22).

Farlik, M. et al., 2015. Single-Cell DNA Methylome Sequencing and Bioinformatic Inference of Epigenomic Cell-State Dynamics. *Cell Reports*, 10(8), pp.1386–1397.

Flusberg, B.A. et al., 2010. Direct detection of DNA methylation during single-molecule, real- time sequencing. *Nature Methods*, 7(6), pp.461–465.

Fortin, J.P. & Hansen, K.D., 2016. Analysis of 450k data using minfi. , pp.1–42.

Fortin, J.P., Triche, T.J. & Hansen, K.D., 2017. Preprocessing, normalization and integration of the Illumina HumanMethylationEPIC array with minfi. *Bioinformatics*, 33(4), pp.558–560.

Friedman, N. & Rando, O.J., 2015. Epigenomics and the structure of the living genome. *Genome Research*, 25(10), pp.1482–1490.

Garagnani, P. et al., 2012. Methylation of ELOVL2 gene as a new epigenetic marker of age. *Aging Cell*, 11(6), pp.1132–1134.

GenomeWeb, Roche Shutting Down 454 Sequencing Business.

Goodwin, S., McPherson, J.D. & McCombie, W.R., 2016. Coming of age: ten years of next-generation sequencing technologies. *Nature Reviews Genetics*, 17(6), pp.333–351.

Guo, Y. et al., 2013. Three-stage quality control strategies for DNA re-sequencing data. *Briefings in Bioinformatics*, 15(6), pp.879–889.

Hackenberg, M., Barturen, G. & Oliver, J.L., 2012. DNA Methylation Profiling from High-Throughput Sequencing Data. *DNA Methylation - From Genomics to Technology*, pp.29–54.

Hannum, G. et al., 2013. Genome-wide Methylation Profiles Reveal Quantitative Views of Human Aging Rates. *Molecular Cell*, 49(2), pp.359–367. Available at: http://dx.doi.org/10.1016/j.molcel.2012.10.016.

Hansen, K. & Aryee, M., 2012. The minfi User's Guide Analyzing Illumina 450k Methylation Arrays. *Kasper D. Hansen*, pp.1–21.

Hansen, K.D., 2018. *Analysis of 450k DNA methylation data with minfi*, Available at: https://kasperdanielhansen.github.io/genbioconductor/html/minfi.html.

Harrison, A. & Parle-McDermott, A., 2011. DNA methylation: A timeline of methods and applications. *BioTechniques*, 2(74).

Hirst, M. & Marra, M.A., 2011. Next generation sequencing based approaches to epigenomics. *Briefings in Functional Genomics*, 9(6), pp.455–465.

Holliday, R. & Pugh, J.E., 1975. DNA Modification Mechanisms and Gene Activity during Development. *Science*, 187(4173), pp.226–232.

Hon, G.C., Hawkins, R.D. & Ren, B., 2009. Predictive chromatin signatures in the mammalian genome. *Human Molecular Genetics*, 18(2), pp.195–201.

Huang, Y.W., Huang, T.H.M. & Wang, L.S., 2010. Profiling DNA methylomes from microarray to genome-scale sequencing. *Technology in Cancer Research and Treatment*, 9(2), pp.139–147.

Illumina, I., 2010. Illumina Sequencing Technology.

International Human Genome Sequencing Consortium, 2004. Finishing the euchromatic sequence of the human genome. *Nature*, 431, pp.931–945.

Jeong, H.M. et al., 2016. Efficiency of methylated DNA immunoprecipitation bisulphite sequencing for whole-genome DNA methylation analysis. *Epigenomics*, 8(8), pp.1061–1077.

Ji, L. et al., 2014. Methylated DNA is over-represented in whole-genome bisulfite sequencing data. *Frontiers in Genetics*, 5(SEP), pp.1–10.

Jiang, P. et al., 2014. Methy-Pipe: An integrated bioinformatics pipeline for whole genome bisulfite sequencing data analysis. *PLoS ONE*, 9(6).

Johansson, Å., Enroth, S. & Gyllensten, U., 2013. Continuous Aging of the Human DNA Methylome Throughout the Human Lifespan. *PLoS ONE*, 8(6).

Johnson, A.A. et al., 2012. The Role of DNA Methylation in Aging, Rejuvenation, and Age-Related Disease. *Rejuvenation research*, 15(5), pp.483–494.

Jones, M.J., Goodman, S.J. & Kobor, M.S., 2015. DNA methylation and healthy human aging. *Aging Cell*, 14(6), pp.924–932.

Jones, P.A. & Baylin, S.B., 2007. The Epigenomics of Cancer. *Cell*, 128, pp.683–692.

Jones, P.L. et al., 1998. Methylated DNA and MeCP2 recruit histone deacetylase to repress Methylated DNA and MeCP2 recruit histone deacetylase to repress transcription. *Nature Genetics*, 19, pp.187–191.

Kaikkonen, M.U., Lam, M.T.Y. & Glass, C.K., 2011. Non-coding RNAs as regulators of gene expression and epigenetics. *Cardiovascular Research*, 90, pp.430–440.

Karsch-Mizrachi, I., Takagi, T. & Cochrane, G., 2018. The international nucleotide sequence database collaboration. *Nucleic Acids Research*, 46(D1), pp.D48–D51.

Ke, R. et al., 2016. Fourth Generation of Next-Generation Sequencing Technologies: Promise and Consequences. *Human Mutation*, 37(12), pp.1363–1367.

Khademhosseini, A., Suh, K.-Y. & Zourob, M., 2013. *Biological Microarrays - Methods and protocols*,

Kohli, R.M. & Zhang, Y., 2013. TET enzymes, TDG and the dynamics of DNA demethylation. *Nature*, 502(7472), pp.472–479.

Kornberg, R.D. & Lorch, Y., 1999. Twenty-Five Years of the Nucleosome , Fundamental Particle of the Eukaryote Chromosome. *Cell*, 98, pp.285–294.

Krueger, F. et al., 2012. DNA methylome analysis using short bisulfite sequencing data. *Nature Methods*, 9(2), pp.145–151. Available at: http://www.nature.com/doifinder/10.1038/nmeth.1828.

Krueger, F. & Andrews, S.R., 2011. Bismark: A flexible aligner and methylation caller for Bisulfite-Seq applications. *Bioinformatics*, 27(11), pp.1571–1572.

Krueger, F. & Andrews, S.R., 2012. Quality Control , trimming and alignment of Bisulfite-Seq data. *Epigenesys*, (July), pp.1–13.

Kuo, K.H.M., 2017. Multiple Testing in the Context of Gene Discovery in Sickle Cell Disease Using GWAS. *Genomics Insights*, 10, pp.1–10.

Laird, P.W., 2010. Principles and challenges of genome- wide DNA methylation analysis. *Nature Reviews Genetics*, 11, pp.191–203.

Laird, P.W., 2003. The power and the promise of DNA methylation markers. *Nature reviews*, 3, pp.253–266.

Lehninger, A.L., Nelson, D.L. & Cox, M.M., 2005. *Principles of Biochemistry* Fourth Edi.,

Leinonen, R., Sugawara, H. & Shumway, M., 2011. The sequence read archive. *Nucleic Acids Research*, 39(SUPPL. 1), pp.2010–2012.

Li, D. et al., 2015. Combining MeDIP-seq and MRE-seq to investigate genome-wide CpG methylation. *Methods*, 72, pp.29–40.

Li, H., Ruan, J. & Durbin, R., 2008. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Research*, 18, pp.1851–1858.

Lister, R. et al., 2009. Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature*, 462(7271), pp.315–322.

Lister, R. & Ecker, J.R., 2009. Finding the fifth base: Genome-wide sequencing of cytosine methylation. *Genome Research*, 19(6), pp.959–966.

Liu, L. et al., 2012. Comparison of Next-Generation Sequencing Systems. *Journal of Biomedicine and Biotechnology*, 2012.

Luger, K. et al., 1997. Crystal Structure of the nucleosome core particle at 2.8A resolution. *Nature*, 389(6648), pp.251–260.

Maksimovic, J. & Phipson, B., 2015. RPubs. *BioinfoSummer2015: 450k Analysis Workshop*. Available at: https://rpubs.com/anavoj/133334.

Maksimovic, J., Phipson, B. & Oshlack, A., 2016. A cross-package Bioconductor workflow for analysing methylation array data. *F1000Research*, 5, p.1281.

Mardis, E.R., 2008. Next-Generation DNA Sequencing Methods. *The Annual Review of Genomics and Human Genetics*, 9, pp.387–402.

Mattick, J.S. & Makunin, I. V., 2006. Non-coding RNA. *Human Molecular Genetics*, 15(1), pp.17–29.

Maxam, A.M. & Gilbert, W., 1977. A new method for sequencing DNA. *Proceedings of the National Academy of Sciences*, 74(2), pp.560–564.

McGowan, P.O. & Szyf, M., 2010. Environmental epigenomics: understanding the effects of parental care on the epigenome. *Essays in biochemistry*, 48(1), pp.275–287.

Mensaert, K. et al., 2014. Next-Generation Technologies and Data Analytical Approaches for Epigenomics. *Environmental and Molecular Mutagenesis*, 55(3), pp.155–170.

Metzker, M.L., 2010. Sequencing technologies — the next generation. *Nature Reviews Genetics*, 11(1), pp.31–46.

Metzker, M.L., 2010. Sequencing technologies — the next generation. *Nature Reviews Genetics*, 11(1), pp.31–46. Available at: http://dx.doi.org/10.1038/nrg2626.

Miller, M.B. & Tang, Y.W., 2009. Basic concepts of microarrays and potential applications in clinical microbiology. *Clinical Microbiology Reviews*, 22(4), pp.611–633.

Moran, S., Arribas, C. & Esteller, M., 2016. Validation of a DNA methylation microarray for 850,000 CpG sites of the human genome enriched in enhancer sequences. *Epigenomics*, 8(3), pp.389–399.

Morris, T.J. & Beck, S., 2015. Analysis pipelines and packages for Infinium HumanMethylation450 BeadChip (450k) data. *Methods*, 72(C), pp.3–8.

Müllner, D., 2013. fastcluster : Fast Hierarchical, Agglomerative Clustering Routines for R and Python. *Journal of Statistical Software*, 53(9).

NCBI, Gene Expression Omnibus. Available at: https://www.ncbi.nlm.nih.gov/geo/ [Accessed April 28, 2018].

Numata, S. et al., 2012. DNA methylation signatures in development and aging of the human prefrontal cortex. *American Journal of Human Genetics*, 90(2), pp.260–272.

Okano, M. et al., 1999. DNA Methyltransferases Dnmt3a and Dnmt3b Are Essential for De Novo Methylation and Mammalian Development. *Cell*, 99, pp.247–257.

Ozanne, S.E. & Constância, M., 2007. Mechanisms of disease: the developmental origins of disease and the role of the epigenotype. *Nature*, 3(7), pp.539–546.

Park, J.L. et al., 2016. Identification and evaluation of age-correlated DNA methylation markers for forensic use. *Forensic Science International: Genetics*, 23, pp.64–70.

Parle-McDermott, A. & Ozaki, M., 2011. The Impact of Nutrition on Differential Methylated Regions of the Genome. *Advances in Nutrition*, 2(6), pp.463–471.

Pavlopoulos, G.A. et al., 2013. Unraveling genomic variation from next generation sequencing data. *BioData Mining*, 6(1), p.13.

Peters, T.J. et al., 2015. De novo identification of differentially methylated regions in the human genome. *Epigenetics & Chromatin*, 8(1), p.6.

Pidsley, R. et al., 2016. Critical evaluation of the Illumina MethylationEPIC BeadChip microarray for whole-genome DNA methylation profiling. *Genome Biology*, 17(1), pp.1–17.

Pomraning, K.R., Smith, K.M. & Freitag, M., 2009. Genome-wide high throughput analysis of DNA methylation in eukaryotes. *Methods*, 47(3), pp.142–150.

Quail, M. et al., 2012. A tale of three next generation sequencing platforms: comparison of Ion torrent, pacific biosciences and illumina MiSeq sequencers. *BMC Genomics*, 13(1), p.341.

Quinodoz, S. & Guttman, M., 2014. Long noncoding RNAs: an emerging link between gene regulation and nuclear organization. *Trends in Cell Biology*, 24(11), pp.651–663.

Reinius, L.E. et al., 2012. Differential DNA Methylation in Purified Human Blood Cells : Implications for Cell Lineage and Studies on Disease Susceptibility. *PloS ONE*, 7(7).

Robison T., J. et al., 2012. Integrative Genomics Viewer. *Nature Biotechnology*, 29(1), pp.24–26.

Romanoski, C.E. et al., 2015. Roadmap for regulation. *Nature*, 518, pp.314–316.

Sander, F., Goulson, A.R. & Road, H., 1975. A Rapid Method for Determining Sequences in DNA by Primed Synthesis with DNA Polymerase. , 94(3), pp.441–448.

Sanger, F. & Nicklen, S., 1977. DNA sequencing with chain-terminating. *Proceedings of the National Academy of Sciences*, 74(12), pp.5463–5467.

Schroeder, D.I. et al., 2011. Large-scale methylation domains mark a functional subset of neuronally expressed genes. *Genome Research*, 21(10), pp.1583–1591.

Schultz, M.D. et al., 2015. Human body epigenome maps reveal noncanonical DNA methylation variation. *Nature*, 523(7559), pp.212–216.

Sean, D. & Meltzer, P.S., 2007. GEOquery: A bridge between the Gene Expression Omnibus (GEO) and BioConductor. *Bioinformatics*, 23(14), pp.1846–1847.

Serre-Miranda, C. et al., 2015. Effector memory CD4 $^+$ T cells are associated with cognitive performance in a senior population. *Neurology - Neuroimmunology Neuroinflammation*, 2(1), p.e54. Available at: http://nn.neurology.org/lookup/doi/10.1212/NXI.0000000000000054.

Shendure, J. & Ji, H., 2008. Next-generation DNA sequencing. *Nature Biotechnology*, 26(10), pp.1135–1145.

Shokralla, S. et al., 2012. Next-generation sequencing technologies for environmental DNA research. *Molecular Ecology*, 21, pp.1794–1805.

Simpson, J.T. et al., 2017. Detecting DNA cytosine methylation using nanopore sequencing. *Nature Methods*, 14(4), pp.407–410. Available at: http://dx.doi.org/10.1038/nmeth.4184.

Suganuma, T. & Workman, J.L., 2011. Signals and Combinatorial Functions of Histone Modifications. *Annual Review of Biochemistry*, 80, pp.473–499.

Swerdlow, H. et al., 1990. Capillary gel electrophoresis for DNA sequencing. Laser-induced fluorescence detection with the sheath flow cuvette. *Journal of Chromatography*, 516, pp.61–67.

Takai, D. & Jones, P.A., 2002. Comprehensive analysis of CpG islands in human chromosomes 21 and 22. *Proceedings of the National Academy of Sciences*, 99(6), pp.3740–3745.

Teider, N. et al., 2010. Neuralized1 causes apoptosis and downregulates Notch target genes in medulloblastoma. *Neuro-Oncology*, 12(12), pp.1244–1256.

Thermo Fisher Scientific, 5500 Series Genetic Analyzers Discontinuance Letter.

Thorvaldsdóttir, H., Robinson, J.T. & Mesirov, J.P., 2013. Integrative Genomics Viewer (IGV): High-performance genomics data visualization and exploration. *Briefings in Bioinformatics*, 14(2), pp.178–192.

Touleimat, N. & Tost, J., 2012. Complete pipeline for Infinium ® Human Methylation 450K BeadChip data processing using subset quantile normalization for accurate DNA methylation estimation. *Epigenomics*, 4(3), pp.325–341.

Toyota, M. et al., 2009. Cancer epigenomics: Implications of DNA methylation in personalized cancer therapy. *Cancer Science*, 100(5), pp.787–791.

Tsai, P.-C., Spector, T.D. & Bell, J.T., 2012. Using epigenome-wide association scans of DNA methylation in age-related complex human traits. *Epigenomics*, 4(5), pp.511–526.

Turner, S.D., 2014. qqman: an R package for visualizing GWAS results using Q-Q and manhattan plots. *The Journal of Open Source Software*.

Van Vlodrop, I.J.H. et al., 2017. A four-gene promoter methylation marker panel consisting of GREM1, NEURL, LAD1, and NEFH predicts survival of clear cell renal cell cancer patients. *Clinical Cancer Research*, 23(8), pp.2006–2018.

Wan, J. et al., 2015. Characterization of tissue-specific differential DNA methylation suggests distinct modes of positive and negative gene expression regulation. *BMC Genomics*, 16, pp.1–11.

Wang, Q. et al., 2016. Four and a half LIM domains 2 contributes to the development of human tongue squamous cell carcinoma. *Journal of Molecular Histology*, 47(2), pp.105–116.

Wilhelm-Benartzi, C.S. et al., 2013. Review of processing and analysis methods for DNA methylation array data. *British Journal of Cancer*, 109(6), pp.1394–1402.

Wilson, B.J. & Nicholls, S.G., 2015. The Human Genome Project, and recent advances in personalized genomics. *Risk management and healthcare policy*, 8, pp.9–20.

Winnefeld, M. & Lyko, F., 2012. The aging epigenome: DNA methylation from the cradle to the grave. *Genome Biology*, 13(7).

Wright, M.L. et al., 2016. Establishing an analytic pipeline for genome-wide DNA methylation. *Clinical Epigenetics*, 8(1), pp.1–10.

Wu, H. et al., 2005. Hypomethylation-linked activation of PAX2 mediates tamoxifen-stimulated endometrial carcinogenesis. *Nature*, 438, pp.981–987.

Wu, M.C. & Kuan, P.F., 2018. A guide to illumina beadchip data analysis. *Methods in Molecular Biology*, 1708, pp.303–330.

Yong, W.-S., Hsu, F.-M. & Chen, P.-Y., 2016. Profiling genome-wide DNA methylation. *Epigenetics & Chromatin*, 9(1), p.26.

Zhang, H., 2016. Overview of Sequence Data Formats. In E. Mathé & S. Davis, eds. *Statistical Genomics: Methods and Protocols*. Humana Press, pp. 3–17.

Zhang, Y. et al., 1999. Analysis of the NuRD subunits reveals a histone deacetylase core complex and a connection with DNA methylation. *Genes & Development*, 13(15), pp.1924–1935.

Zhu, Y. et al., 2008. GEOmetadb: Powerful alternative search engine for the Gene Expression Omnibus. *Bioinformatics*, 24(23), pp.2798–2800.

Zhu, Y. et al., 2013. SRAdb: query and use public next-generation sequencing data from within R. *BMC Bioinformatics*, 14, p.19.

Supplementary Files

Supplementary Table 1 – First *Mus musculus* dataset before the selection of data

GEO Series	***GEO Platform***	***GEO Samples***	***Strain***	***Age***	***Tissue***	***Acession fastq***	***Reads type***	***Genome ref***	***Technology***
GSE52330	GPL13112	GSM1263221	C57Black6	8-10w	Dentate gyri	SRR1029055	paired	mm9	HiSeq2000
			C57Black6	8-10w	Dentate gyri	SRR1029056	paired	mm9	HiSeq2000
			C57Black6	8-10w	Dentate gyri	SRR1029057	paired	mm9	HiSeq2000
			C57Black6	8-10w	Dentate gyri	SRR1029058	paired	mm9	HiSeq2000
		GSM1263222	C57Black6	8-10w	Dentate gyri	SRR1029059	paired	mm9	HiSeq2000
			C57Black6	8-10w	Dentate gyri	SRR1029060	paired	mm9	HiSeq2000
			C57Black6	8-10w	Dentate gyri	SRR1029061	paired	mm9	HiSeq2000
			C57Black6	8-10w	Dentate gyri	SRR1029062	paired	mm9	HiSeq2000
			C57Black6	8-10w	Dentate gyri	SRR1029063	paired	mm9	HiSeq2000
			C57Black6	8-10w	Dentate gyri	SRR1029064	paired	mm9	HiSeq2000
			C57Black6	8-10w	Dentate gyri	SRR1029065	paired	mm9	HiSeq2000
			C57Black6	8-10w	Dentate gyri	SRR1029066	paired	mm9	HiSeq2000
			C57Black6	8-10w	Dentate gyri	SRR1029067	paired	mm9	HiSeq2000
			C57Black6	8-10w	Dentate gyri	SRR1029068	paired	mm9	HiSeq2000
			C57Black6	8-10w	Dentate gyri	SRR1029069	paired	mm9	HiSeq2000
			C57Black6	8-10w	Dentate gyri	SRR1029070	paired	mm9	HiSeq2000
GSE60062	GPL13112	GSM1464464	C57Black6	8w	Cerebellum	SRR1536120	single	mm9	HiSeq2000
			C57Black6	8w	Cerebellum	SRR1536121	single	mm9	HiSeq2000
			C57Black6	8w	Cerebellum	SRR1536122	single	mm9	HiSeq2000
			C57Black6	8w	Cerebellum	SRR1536123	single	mm9	HiSeq2000
		GSM1464465	C57Black6	8w	Cortex	SRR1536124	single	mm9	HiSeq2000
			C57Black6	8w	Cortex	SRR1536125	single	mm9	HiSeq2000
			C57Black6	8w	Cortex	SRR1536126	single	mm9	HiSeq2000
			C57Black6	8w	Cortex	SRR1536127	single	mm9	HiSeq2000
GSE63137	GPL13112	GSM1541958	C57BL6J/129	8-11w	Neocortex	SRR1647862	single	mm10	HiSeq2000
			C57BL6J/129	8-11w	Neocortex	SRR1647863	single	mm10	HiSeq2000
			C57BL6J/129	8-11w	Neocortex	SRR1647864	single	mm10	HiSeq2000
		GSM1541959	C57BL6J/129	8-11w	Neocortex	SRR1647865	single	mm10	HiSeq2000
			C57BL6J/129	8-11w	Neocortex	SRR1647866	single	mm10	HiSeq2000
			C57BL6J/129	8-11w	Neocortex	SRR1647867	single	mm10	HiSeq2000
		GSM1541960	C57BL6J/129	8-11w	Neocortex	SRR1647868	single	mm10	HiSeq2000

			C57BL6J/129	8-11w	Neocortex	SRR1647869	single	mm10	HiSeq2000
			C57BL6J/129	8-11w	Neocortex	SRR1647870	single	mm10	HiSeq2000
		GSM1541961	C57BL6J/129	8-11w	Neocortex	SRR1647875	single	mm10	HiSeq2000
			C57BL6J/129	8-11w	Neocortex	SRR1647876	single	mm10	HiSeq2000
			C57BL6J/129	8-11w	Neocortex	SRR1647877	single	mm10	HiSeq2000
			C57BL6J/129	8-11w	Neocortex	SRR1647878	single	mm10	HiSeq2000
		GSM1541962	C57BL6J/129	8-11w	Neocortex	SRR1647871	single	mm10	HiSeq2000
			C57BL6J/129	8-11w	Neocortex	SRR1647872	single	mm10	HiSeq2000
		GSM1541963	C57BL6J/129	8-11w	Neocortex	SRR1647873	single	mm10	HiSeq2000
			C57BL6J/129	8-11w	Neocortex	SRR1647874	single	mm10	HiSeq2000
			C57BL6J/129	8-11w	Neocortex	SRR1647879	single	mm10	HiSeq2000
GSE67292	GPL16417	GSM1643930	C57BL/6	10-11w	Cerebellum	SRR1930024	paired	mm9	MiSeq
		GSM1643931	C57BL/6	10-11w	Cerebellum	SRR1930025	paired	mm9	MiSeq
GSE68618	GPL13112	GSM1677165	C57B|6	16-18m	Pancreatic Beta Cells	SRR2034988	paired	mm9	HiSeq2000
			C57B|6	16-18m	Pancreatic Beta Cells	SRR2034989	paired	mm9	HiSeq2000
			C57B|6	16-18m	Pancreatic Beta Cells	SRR2034990	paired	mm9	HiSeq2000
		GSM1677166	C57B|6	4-6w	Pancreatic Beta Cells	SRR2034991	paired	mm9	HiSeq2000
			C57B|6	4-6w	Pancreatic Beta Cells	SRR2034992	paired	mm9	HiSeq2000
			C57B|6	4-6w	Pancreatic Beta Cells	SRR2034993	paired	mm9	HiSeq2000
GSE70317	GPL16417	GSM1723681	C57BL/6N	7w	Brain	SRR2079716	paired	mm9	MiSeq
		GSM1723682	C57BL/6N	7w	Heart	SRR2079717	paired	mm9	MiSeq
		GSM1723683	C57BL/6N	7w	Heart	SRR2079718	paired	mm9	MiSeq
		GSM1723684	C57BL/6N	7w	Heart	SRR2079719	paired	mm9	MiSeq
		GSM1723685	C57BL/6N	7w	Kidney	SRR2079720	paired	mm9	MiSeq
		GSM1723686	C57BL/6N	7w	Spleen	SRR2079721	paired	mm9	MiSeq
		GSM1723687	C57BL/6N	7w	Kidney	SRR2079722	paired	mm9	MiSeq
		GSM1723688	C57BL/6N	7w	Kidney	SRR2079723	paired	mm9	MiSeq
		GSM1723689	C57BL/6N	7w	Brain	SRR2079724	paired	mm9	MiSeq
		GSM1723690	C57BL/6N	7w	Kidney	SRR2079725	paired	mm9	MiSeq
		GSM1723691	C57BL/6N	7w	Liver	SRR2079726	paired	mm9	MiSeq
		GSM1723692	C57BL/6N	7w	Liver	SRR2079727	paired	mm9	MiSeq
		GSM1723693	C57BL/6N	7w	Spleen	SRR2079728	paired	mm9	MiSeq
		GSM1723694	C57BL/6N	7w	Liver	SRR2079729	paired	mm9	MiSeq
		GSM1723695	C57BL/6N	7w	Liver	SRR2079730	paired	mm9	MiSeq
		GSM1723696	C57BL/6N	7w	Spleen	SRR2079731	paired	mm9	MiSeq
		GSM1723697	C57BL/6N	7w	Liver	SRR2079732	paired	mm9	MiSeq

		GSM1723698	C57BL/6N	7w	Liver	SRR2079733	paired	mm9	MiSeq
		GSM1723699	C57BL/6N	7w	Liver	SRR2079734	paired	mm9	MiSeq
	GPL13112	GSM1723710	C57BL/6N	7w	Brain	SRR2079745	single	mm9	HiSeq 2000
		GSM1723711	C57BL/6N	7w	Liver	SRR2079746	single	mm9	HiSeq 2000
		GSM1723712	C57BL/6N	7w	Liver	SRR2079747	single	mm9	HiSeq 2000
		GSM1723713	C57BL/6N	7w	Liver	SRR2079748	single	mm9	HiSeq 2000
		GSM1723714	C57BL/6N	7w	Liver	SRR2079749	single	mm9	HiSeq 2000
GSE72177	GPL13112	GSM1857044	C57/BL6	22 w	Liver	SRR2173835	paired	mm9	HiSeq 2000
			C57/BL6	22 w	Liver	SRR2173836	paired	mm9	HiSeq 2000
			C57/BL6	22 w	Liver	SRR2173837	paired	mm9	HiSeq 2000
			C57/BL6	22 w	Liver	SRR2173838	paired	mm9	HiSeq 2000
			C57/BL6	22 w	Liver	SRR2173839	paired	mm9	HiSeq 2000
			C57/BL6	22 w	Liver	SRR2173840	paired	mm9	HiSeq 2000
			C57/BL6	22 w	Liver	SRR2173841	paired	mm9	HiSeq 2000
			C57/BL6	22 w	Liver	SRR2173842	paired	mm9	HiSeq 2000
			C57/BL6	22 w	Liver	SRR2173843	paired	mm9	HiSeq 2000
			C57/BL6	22 w	Liver	SRR2173844	paired	mm9	HiSeq 2000
			C57/BL6	22 w	Liver	SRR2173845	paired	mm9	HiSeq 2000
			C57/BL6	22 w	Liver	SRR2173846	paired	mm9	HiSeq 2000
			C57/BL6	22 w	Liver	SRR2173847	paired	mm9	HiSeq 2000
			C57/BL6	22 w	Liver	SRR2173848	paired	mm9	HiSeq 2000
			C57/BL6	22 w	Liver	SRR2173849	paired	mm9	HiSeq 2000
			C57/BL6	22 w	Liver	SRR2173850	paired	mm9	HiSeq 2000
		GSM1857045	C57/BL6	22 w	Liver	SRR2173851	paired	mm9	HiSeq 2000
			C57/BL6	22 w	Liver	SRR2173852	paired	mm9	HiSeq 2000
			C57/BL6	22 w	Liver	SRR2173853	paired	mm9	HiSeq 2000
			C57/BL6	22 w	Liver	SRR2173854	paired	mm9	HiSeq 2000
			C57/BL6	22 w	Liver	SRR2173855	paired	mm9	HiSeq 2000
			C57/BL6	22 w	Liver	SRR2173856	paired	mm9	HiSeq 2000
			C57/BL6	22 w	Liver	SRR2173857	paired	mm9	HiSeq 2000
			C57/BL6	22 w	Liver	SRR2173858	paired	mm9	HiSeq 2000
			C57/BL6	22 w	Liver	SRR2173859	paired	mm9	HiSeq 2000
			C57/BL6	22 w	Liver	SRR2173860	paired	mm9	HiSeq 2000
			C57/BL6	22 w	Liver	SRR2173861	paired	mm9	HiSeq 2000
			C57/BL6	22 w	Liver	SRR2173862	paired	mm9	HiSeq 2000
			C57/BL6	22 w	Liver	SRR2173863	paired	mm9	HiSeq 2000
			C57/BL6	22 w	Liver	SRR2173864	paired	mm9	HiSeq 2000
			C57/BL6	22 w	Liver	SRR2173865	paired	mm9	HiSeq 2000
			C57/BL6	22 w	Liver	SRR2173866	paired	mm9	HiSeq 2000
		GSM1857046	C57/BL6	22 w	Liver	SRR2173867	paired	mm9	HiSeq 2000
			C57/BL6	22 w	Liver	SRR2173868	paired	mm9	HiSeq 2000
			C57/BL6	22 w	Liver	SRR2173869	paired	mm9	HiSeq 2000
			C57/BL6	22 w	Liver	SRR2173870	paired	mm9	HiSeq 2000
			C57/BL6	22 w	Liver	SRR2173871	paired	mm9	HiSeq 2000

			C57/BL6	22 w	Liver	SRR2173872	paired	mm9	HiSeq 2000
			C57/BL6	22 w	Liver	SRR2173873	paired	mm9	HiSeq 2000
			C57/BL6	22 w	Liver	SRR2173874	paired	mm9	HiSeq 2000
			C57/BL6	22 w	Liver	SRR2173875	paired	mm9	HiSeq 2000
			C57/BL6	22 w	Liver	SRR2173876	paired	mm9	HiSeq 2000
			C57/BL6	22 w	Liver	SRR2173877	paired	mm9	HiSeq 2000
			C57/BL6	22 w	liver	SRR2173878	paired	mm9	HiSeq 2000
			C57/BL6	22 w	liver	SRR2173879	paired	mm9	HiSeq 2000
			C57/BL6	22 w	liver	SRR2173880	paired	mm9	HiSeq 2000
			C57/BL6	22 w	liver	SRR2173881	paired	mm9	HiSeq 2000
			C57/BL6	22 w	liver	SRR2173882	paired	mm9	HiSeq 2000
GSE78955	GPL17021	GSM2082053	C57BL/6	7w	erthroblast	SRR3208877	single	mm10	HiSeq2500
GSE92486	GPL17021	GSM2430564	C3B6F1	5m	liver	SRR5115679	paired	GRCm38	HiSeq2500
		GSM2430565	C3B6F1	5m	liver	SRR5115680	paired	GRCm38	HiSeq2500
		GSM2430566	C3B6F1	5m	liver	SRR5115681	paired	GRCm38	HiSeq2500
		GSM2430570	C3B6F1	26m	liver	SRR5115685	paired	GRCm38	HiSeq2500
		GSM2430571	C3B6F1	26m	liver	SRR5115686	paired	GRCm38	HiSeq2500
		GSM2430572	C3B6F1	26m	liver	SRR5115687	paired	GRCm38	HiSeq2500
GSE49191	GPL13112	GSM1206262	C57BL/6	12m	Hematopoietic stem cells	SRR950173	paired	mm9	HiSeq2000
		GSM1206263	C57BL/6	12m	Hematopoietic stem cells	SRR950174	paired	mm9	HiSeq2000
		GSM1206264	C57BL/6	12m	Hematopoietic stem cells	SRR950175	paired	mm9	HiSeq2000
		GSM1206265	C57BL/6	12m	Hematopoietic stem cells	SRR950176	paired	mm9	HiSeq2000
		GSM1206266	C57BL/6	12m	Hematopoietic stem cells	SRR950177	paired	mm9	HiSeq2000
		GSM1206267	C57BL/6	12m	Hematopoietic stem cells	SRR950178	paired	mm9	HiSeq2000
		GSM1206268	C57BL/6	12m	Hematopoietic stem cells	SRR950179	paired	mm9	HiSeq2000
		GSM1206268	C57BL/6	12m	Hematopoietic stem cells	SRR950180	paired	mm9	HiSeq2000
GSE49623	GPL13112	GSM1202738	C57BL/6	8w	Testis	SRR948779	paired	mm9	HiSeq2000
	GPL13112	GSM1202739	C57BL/6	8w	Testis	SRR948780	paired	mm9	HiSeq2000
	GPL13112	GSM1202740	C57BL/6	8w	Testis	SRR948781	paired	mm9	HiSeq2000
	GPL13112	GSM1202741	C57BL/6	8w	Testis	SRR948782	paired	mm9	HiSeq2000
	GPL13112	GSM1202742	C57BL/6	8w	Testis	SRR948783	paired	mm9	HiSeq2000
	GPL13112	GSM1202743	C57BL/6	8w	Testis	SRR948784	paired	mm9	HiSeq2000
	GPL13112	GSM1202744	C57BL/6	8w	Testis	SRR948785	paired	mm9	HiSeq2000
	GPL13112	GSM1202745	C57BL/6	8w	Testis	SRR948786	paired	mm9	HiSeq2000
	GPL13112	GSM1202746	C57BL/6	8w	Testis	SRR948787	paired	mm9	HiSeq2000
	GPL13112	GSM1202747	C57BL/6	8w	Testis	SRR948788	paired	mm9	HiSeq2000
	GPL13112	GSM1202748	C57BL/6	8w	Testis	SRR948789	paired	mm9	HiSeq2000
	GPL13112	GSM1202749	C57BL/6	8w	Testis	SRR948790	paired	mm9	HiSeq2000
	GPL13112	GSM1202750	C57BL/6	8w	Testis	SRR948791	paired	mm9	HiSeq2000
	GPL13112	GSM1202751	C57BL/6	8w	Testis	SRR948792	paired	mm9	HiSeq2000

	GPL13112	GSM1202752	C57BL/6	8w	Testis	SRR948793	paired	mm9	HiSeq2000
	GPL13112	GSM1202753	C57BL/6	8w	Testis	SRR948794	paired	mm9	HiSeq2000

Supplementary Table 2 – Filtration step across fastq files from all the samples selected on our dataset in order to select the ones with more final reads

Experiment	*Initial Reads*	*Final Reads*	*Mean final reads*	*Trim Filter*	*Ambiguity Filter*	*Quality Filter*	*Reads filtered*	*Mean filtered reads (%)*	*Mean Filtered reads*
SRR2034988_1	97036309	85739726	88939173,5	4940712	6355774	97	11296583	8%	8097136
SRR2034988_2	97036309	92138621		4799395	78671	19622	4897688		
SRR2034989_1	240688878	224346051	224281154	16247555	94637	635	16342827	7%	16407724
SRR2034989_2	240688878	224216257		16301684	104652	66285	16472621		
SRR2034990_1	219562678	217648567	215517266,5	1868356	45685	70	1914111	2%	4045412
SRR2034990_2	219562678	213385966		6069893	79521	27298	6176712		
SRR2034991_1	197404164	186038285	186312251	11276956	88586	337	11365879	6%	11091913
SRR2034991_2	197404164	186586217		10692708	85879	39360	10817947		
SRR2034992_1	185532031	168397286	168599661,5	16511233	623321	191	17134745	9%	16932370
SRR2034992_2	185532031	168802037		16601601	74645	53748	16729994		
SRR2034993_1	213105508	210488453	208442800	2564085	52936	34	2617055	2,19%	4662708
SRR2034993_2	213105508	206397147		6646782	48550	13029	6708361		
SRR2079726_1	174390	174374	174302	0	16	0	16	0,05%	88
SRR2079726_2	174390	174230		1	153	6	160		
SRR2079727_1	199775	199746	199674,5	0	29	0	29	0,05%	100,5
SRR2079727_2	199775	199603		1	168	3	172		
SRR2079729_1	169085	169072	168996	0	13	0	13	0,05%	89
SRR2079729_2	169085	168920		3	156	6	165		
SRR2079730_1	164943	164920	164853,5	0	23	0	23	0,05%	89,5
SRR2079730_2	164943	164787		0	156	0	156		
SRR2079732_1	40269	40267	40250	0	2	0	2	0,05%	19
SRR2079732_2	40269	40233		0	36	0	36		
SRR2079733_1	45544	45540	45525	0	4	0	4	0,04%	19
SRR2079733_2	45544	45510		0	34	0	34		
SRR2079734_1	34105	34104	34089	0	1	0	1	0,05%	16
SRR2079734_2	34105	34074		0	31	0	31		
SRR2173835_1	35307975	35037366	34875491	264846	5761	2	270609	1,22%	432484
SRR2173835_2	35307975	34713616		589278	4286	795	594359		
SRR2173836_1	34226300	33958775	33806740	260168	7354	3	267525	1,23%	419560
SRR2173836_2	34226300	33654705		554648	15990	957	571595		
SRR2173837_1	34096758	33827611	33666545	264983	4163	1	269147	1,26%	430213
SRR2173837_2	34096758	33505479		567819	22443	1017	591279		
SRR2173838_1	34261936	33987515	33836245,5	268946	5475	0	274421	1,24%	425690,5
SRR2173838_2	34261936	33684976		559493	16513	954	576960		
SRR2173839_1	34596209	34326468	34171919,5	264738	5003	0	269741	1,23%	424289,5
SRR2173839_2	34596209	34017371		559142	18706	990	578838		

SRR2173840_1	34780215	34511631	34346269,5	263212	5372	0	268584	1,25%	433945,5
SRR2173840_2	34780215	34180908		564572	33682	1053	599307		
SRR2173841_1	32462856	32122870	31964537	335504	4482	0	339986	1,54%	498319
SRR2173841_2	32462856	31806204		645760	9509	1383	656652		
SRR2173842_1	37119474	36771303	36472536,5	347540	619	12	348171	1,74%	646937,5
SRR2173842_2	37119474	36173770		921621	18663	5420	945704		
SRR2173843_1	35591830	35314410	35156573,5	269715	7704	1	277420	1,22%	435256,5
SRR2173843_2	35591830	34998737		587536	4825	732	593093		
SRR2173844_1	35048690	34776184	34618471,5	265007	7497	2	272506	1,23%	430218,5
SRR2173844_2	35048690	34460759		580706	6488	737	587931		
SRR2173845_1	35457350	35177170	34998089,5	273504	6675	1	280180	1,30%	459260,5
SRR2173845_2	35457350	34819009		630031	7539	771	638341		
SRR2173846_1	35592596	35310589	35140894	275366	6639	2	282007	1,27%	451702
SRR2173846_2	35592596	34971199		613883	6809	705	621397		
SRR2173847_1	35582221	35313010	35154531,5	266467	2744	0	269211	1,20%	427689,5
SRR2173847_2	35582221	34996053		582268	3085	815	586168		
SRR2173848_1	35789090	35517598	35352952,5	267819	3671	2	271492	1,22%	436137,5
SRR2173848_2	35789090	35188307		594646	5320	817	600783		
SRR2173849_1	33757980	33487695	33333938	260710	9573	2	270285	1,26%	424042
SRR2173849_2	33757980	33180181		559360	17511	928	577799		
SRR2173850_1	34641232	34367159	34208084	265153	8918	2	274073	1,25%	433148
SRR2173850_2	34641232	34049009		574452	16790	981	592223		
SRR2173851_1	36966380	36700991	36532522	259497	5890	2	265389	1,17%	433858
SRR2173851_2	36966380	36364053		597282	4387	658	602327		
SRR2173852_1	35868592	35604794	35446049,5	256076	7721	1	263798	1,18%	422542,5
SRR2173852_2	35868592	35287305		563366	17009	912	581287		
SRR2173853_1	35765567	35500075	35331625	261168	4324	0	265492	1,21%	433942
SRR2173853_2	35765567	35163175		578112	23344	936	602392		
SRR2173854_1	35925705	35653350	35495605,5	266521	5834	0	272355	1,20%	430099,5
SRR2173854_2	35925705	35337861		569484	17446	914	587844		
SRR2173855_1	36257920	35990913	35830241	261610	5394	3	267007	1,18%	427679
SRR2173855_2	36257920	35669569		567663	19776	912	588351		
SRR2173856_1	36441355	36175716	36003099,5	259979	5660	0	265639	1,20%	438255,5
SRR2173856_2	36441355	35830483		574670	35223	979	610872		
SRR2173857_1	33218255	32897305	32727701	316513	4432	5	320950	1,48%	490554
SRR2173857_2	33218255	32558097		649341	9633	1184	660158		
SRR2173858_1	34920839	34617352	34332683,5	302807	671	9	303487	1,68%	588155,5
SRR2173858_2	34920839	34048015		850242	17359	5223	872824		
SRR2173859_1	37290561	37019258	36853905	263159	8142	2	271303	1,17%	436656
SRR2173859_2	37290561	36688552		596292	5045	672	602009		
SRR2173860_1	36781591	36516259	36351516	257711	7617	4	265332	1,17%	430075
SRR2173860_2	36781591	36186773		587553	6578	687	594818		
SRR2173861_1	37253651	36978884	36791357,5	267821	6945	1	274767	1,24%	462293,5
SRR2173861_2	37253651	36603831		641326	7782	712	649820		

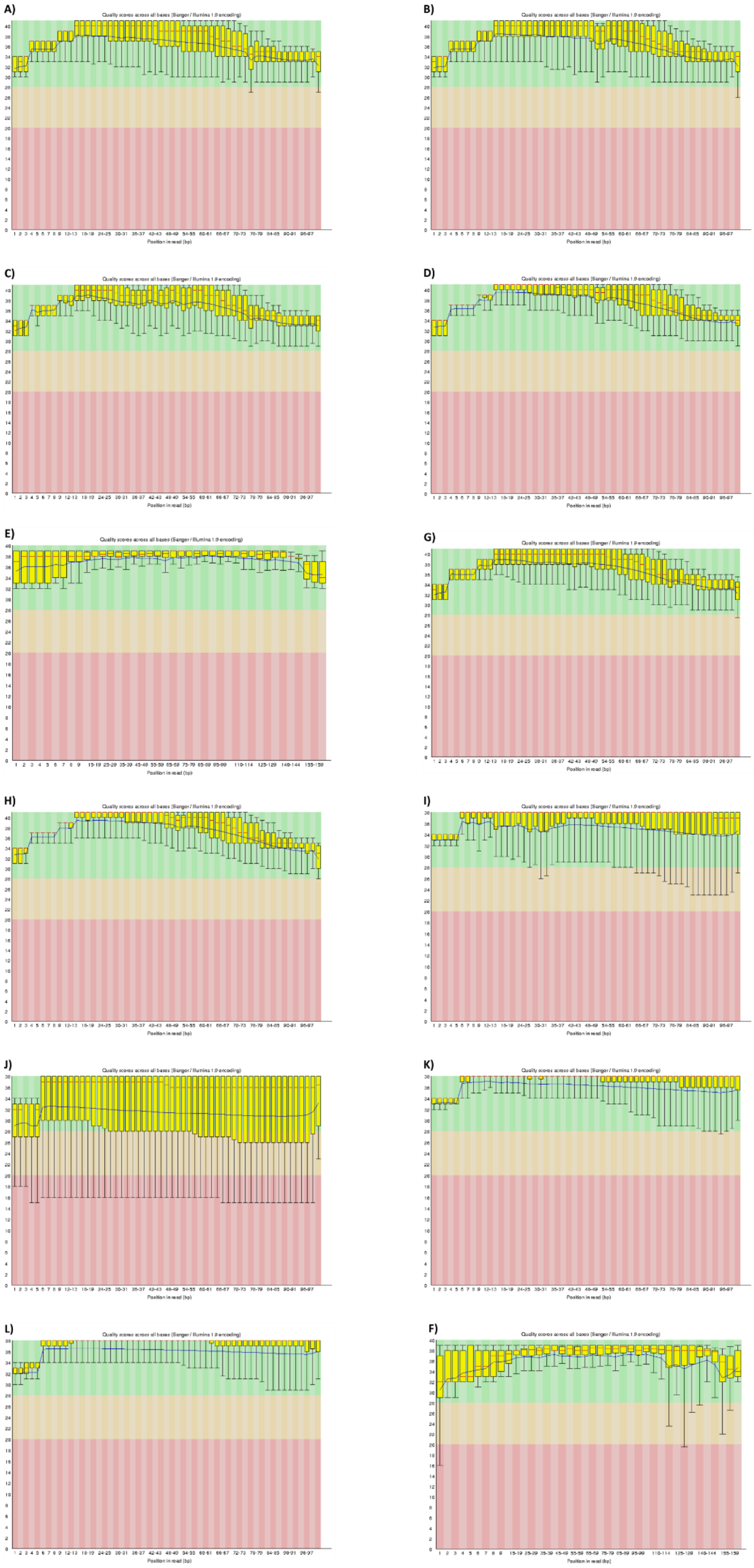
SRR2173862_1	37310731	37035523	36857103	268564	6643	1	275208	1,22%	453628
SRR2173862_2	37310731	36678683		624441	6913	694	632048		
SRR2173863_1	37309664	37047433	36882305,5	259470	2761	0	262231	1,15%	427358,5
SRR2173863_2	37309664	36717178		588679	3054	753	592486		
SRR2173864_1	37550718	37286988	37113853,5	260253	3477	0	263730	1,16%	436864,5
SRR2173864_2	37550718	36940719		603998	5235	766	609999		
SRR2173865_1	35422931	35152553	34992174	260325	10052	1	270378	1,22%	430757
SRR2173865_2	35422931	34831795		571870	18377	889	591136		
SRR2173866_1	36323676	36051425	35885291,5	262829	9422	0	272251	1,21%	438384,5
SRR2173866_2	36323676	35719158		586179	17415	924	604518		
SRR2173867_1	35465955	35206703	35050242,5	253630	5621	1	259252	1,17%	415712,5
SRR2173867_2	35465955	34893782		567320	4227	626	572173		
SRR2173868_1	34577434	34316816	34168847	253250	7368	0	260618	1,18%	408587
SRR2173868_2	34577434	34020878		539540	16213	803	556556		
SRR2173869_1	34508430	34245514	34086720	258720	4195	1	262916	1,22%	421710
SRR2173869_2	34508430	33927926		556870	22773	861	580504		
SRR2173870_1	34667367	34398880	32709573	262976	5510	1	268487	1,12%	387439
SRR2173870_2	31526657	31020266		490393	15269	729	506391		
SRR2173871_1	34954414	34688720	34539329,5	260495	5199	0	265694	1,19%	415084,5
SRR2173871_2	34954414	34389939		544687	18958	830	564475		
SRR2173872_1	35148316	34885636	34723328,5	257231	5448	1	262680	1,21%	424987,5
SRR2173872_2	35148316	34561021		551881	34547	867	587295		
SRR2173873_1	31666460	31352437	31196429	309704	4317	2	314023	1,48%	470031
SRR2173873_2	31666460	31040421		615665	9238	1136	626039		
SRR2173874_1	33672268	33372935	33107535	298718	602	13	299333	1,68%	564733
SRR2173874_2	33672268	32842135		808226	17002	4905	830133		
SRR2173875_1	35769411	35503458	35348734,5	258024	7927	2	265953	1,18%	420676,5
SRR2173875_2	35769411	35194011		569798	4980	622	575400		
SRR2173876_1	35144941	34884002	34730722	253662	7276	1	260939	1,18%	414219
SRR2173876_2	35144941	34577442		560567	6291	641	567499		
SRR2173876_2	35144941	34577442	34959995,5	560567	6291	641	567499	1,19%	418201,5
SRR2173877_1	35611453	35342549		262063	6841	0	268904		
SRR2173877_2	35611453	34993349	34993349	609788	7727	589	618104	1,74%	618104
SRR2173877_2	35611453	34993349		609788	7727	589	618104		
SRR2173878_1	35764081	35492816	35328928,5	264865	6399	1	271265	1,22%	435152,5
SRR2173878_2	35764081	35165041		591829	6558	653	599040		
SRR2173879_1	35746441	35488895	35334599,5	254927	2618	1	257546	1,15%	411841,5
SRR2173879_2	35746441	35180304		562426	3022	689	566137		
SRR2173880_1	35981657	35722696	35561990,5	255562	3398	1	258961	1,17%	419666,5
SRR2173880_2	35981657	35401285		574588	5114	670	580372		
SRR2173881_1	34201029	33935058	33784425,5	256314	9656	1	265971	1,22%	416603,5
SRR2173881_2	34201029	33633793		548788	17631	817	567236		
SRR2173882_1	35020926	34752302	34596804	259499	9124	1	268624	1,21%	424122
SRR2173882_2	35020926	34441306		561747	16978	895	579620		

SRR5115679_1	70538476	70247072	69674394	219379	71823	202	291404	1,22%	864082
SRR5115679_2	70538476	69101716		1414173	4174	18413	1436760		
SRR5115680_1	44615639	44435851	43925276	157758	21871	159	179788	1,55%	690363
SRR5115680_2	44615639	43414701		1181304	3153	16481	1200938		
SRR5115681_1	49504536	49316335	48712325,5	169030	19004	167	188201	1,60%	792210,5
SRR5115681_2	49504536	48108316		1373083	3701	19436	1396220		
SRR5115685_1	229564398	229060184	228706129,5	500653	1784	1777	504214	0,37%	858268,5
SRR5115685_2	229564398	228352075		1090046	114848	7429	1212323		
SRR5115686_1	175947615	175607408	175263755,5	337445	2059	703	340207	0,39%	683859,5
SRR5115686_2	175947615	174920103		970730	52264	4518	1027512		
SRR5115687_1	163041401	162667009	162303303	373656	134	602	374392	0,45%	738098
SRR5115687_2	163041401	161939597		1012717	82821	6266	1101804		

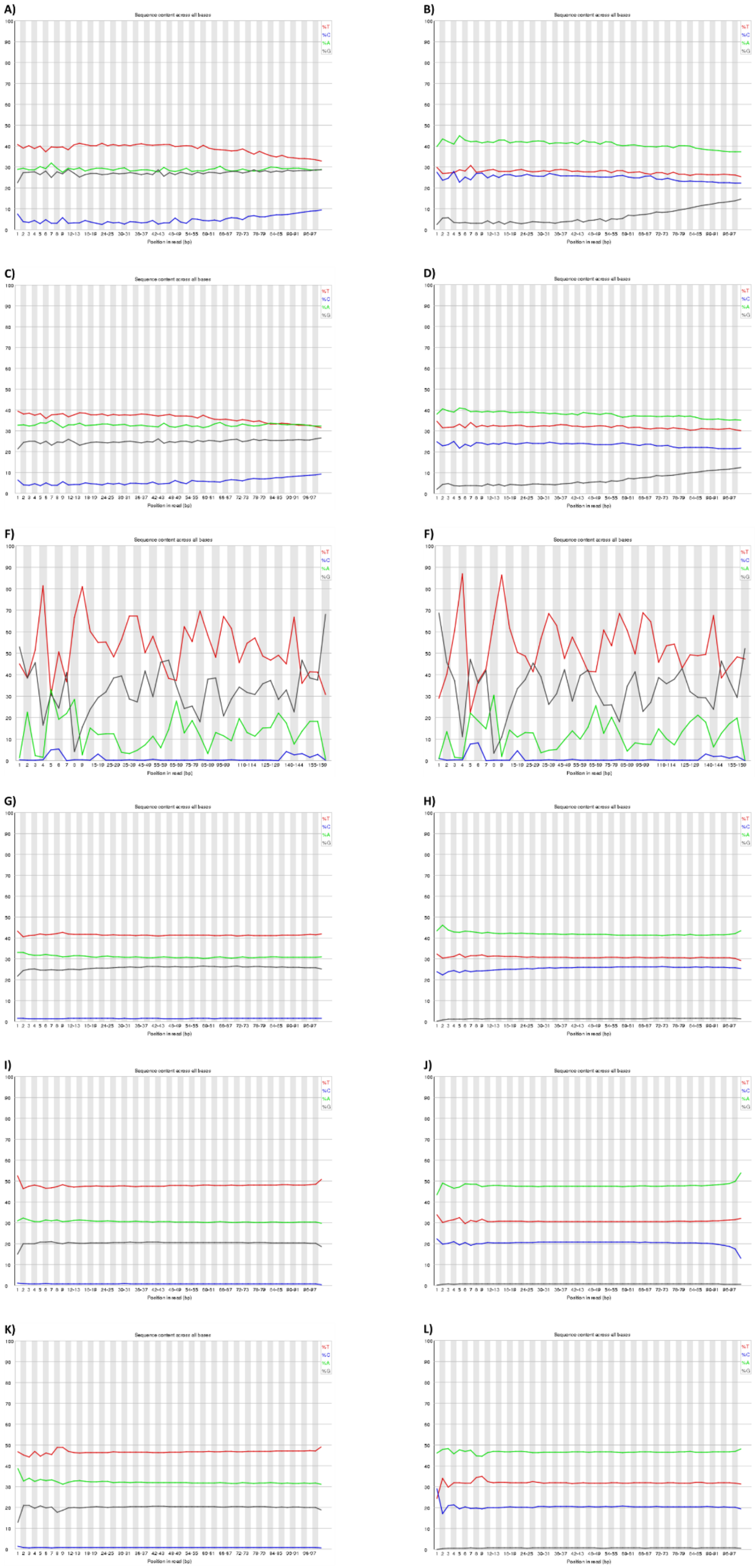
Supplementary Table 3 – Final dataset of blood samples. Specifications about GEO serie, superserie, PMID, number of samples, tissue and cell specificity, array reference and platform used, are presented.

GEO Serie	*GEO Superserie*	*PMID*	*Samples*	*Female*	*Male*	*Age*	*Tissue*	*Tissue specificity*	*Cell specificity*	*Array ref*	*Platform*
GSE99755	-	29362489	37	19	18	24-58	Blood	Whole blood		450k	GPL13534
GSE98876	-	28747766	5	0	5	31-64	Blood	Peripheral blood	PBMC (CD3 T cell)	450k	GPL13534
GSE87648	GSE87650	27886173	92	48	44	18-69	Blood	Whole Blood		450k	GPL13534
GSE87640	GSE87650	27886173	18	6	12	24-58	Blood	Whole blood		450k	GPL13534
GSE87640	GSE87650	27886173	16	7	9	24-58	Blood	Peripheral blood	PBMC (CD4)	450k	GPL13534
GSE87640	GSE87650	27886173	15	6	9	24-58	Blood	Peripheral blood	PBMC (CD8)	450k	GPL13534
GSE87640	GSE87650	27886173	16	7	9	24-58	Blood	Peripheral blood	PBMC (CD14)	450k	GPL13534
GSE87571	-	23826282	664	356	308	18-94	Blood	Whole Blood		450k	GPL13534
GSE85647	GSE85649	28549776	6	6	0	23-52	Blood	Peripheral Blood	PBMC (CD14)	450k	GPL13534
GSE85506	-	28621701	21	21	0	19-80	Blood	Peripheral blood		450k	GPL13534
GSE71955	GSE71957	26459776	31	28	3	35-79	Blood	Peripheral blood	PBMC (CD4)	450k	GPL13534
GSE71955	GSE71957	26459776	31	28	3	35-79	Blood	Peripheral blood	PBMC (CD8)	450k	GPL13534
GSE51057	-	24278132	177	177	0	34-65	Blood	Peripheral blood	Leukocytes	450k	GPL13534
GSE42861	-	23334450	76	60	16	24-70	Blood	Peripheral blood	Leukocytes	450k	GPL13534
GSE107737	-	-	12	0	12	18-29	Blood	Whole Blood		450k	GPL13534
GSE105123	GSE105124	24658407	19	8	11	19-23	Blood	Peripheral Blood	PBMC	450k	GPL13534
GSE104471	GSE104472	28294656	12	6	6	24-45	Blood	Peripheral Blood	PBMC	450k	GPL13534

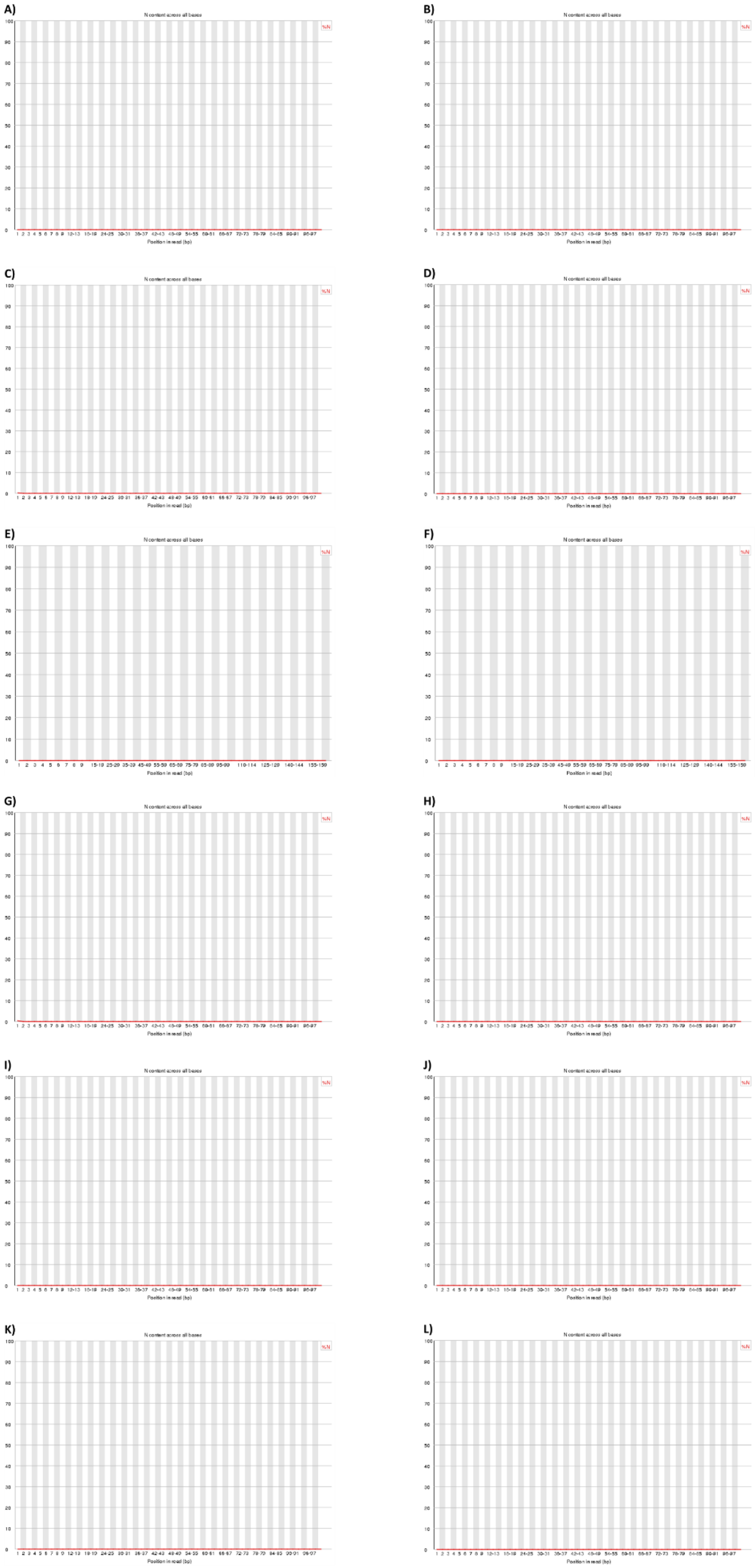
Supplementary Figure 1 – FASTQC reports of Dataset A, per base sequence quality of SRR2034989 (A/B); SRR2034993 (C/D); SRR2079727 (E/F); SRR2173864 (G/H); SRR5115679(I/J); SRR5115685 (K/L)



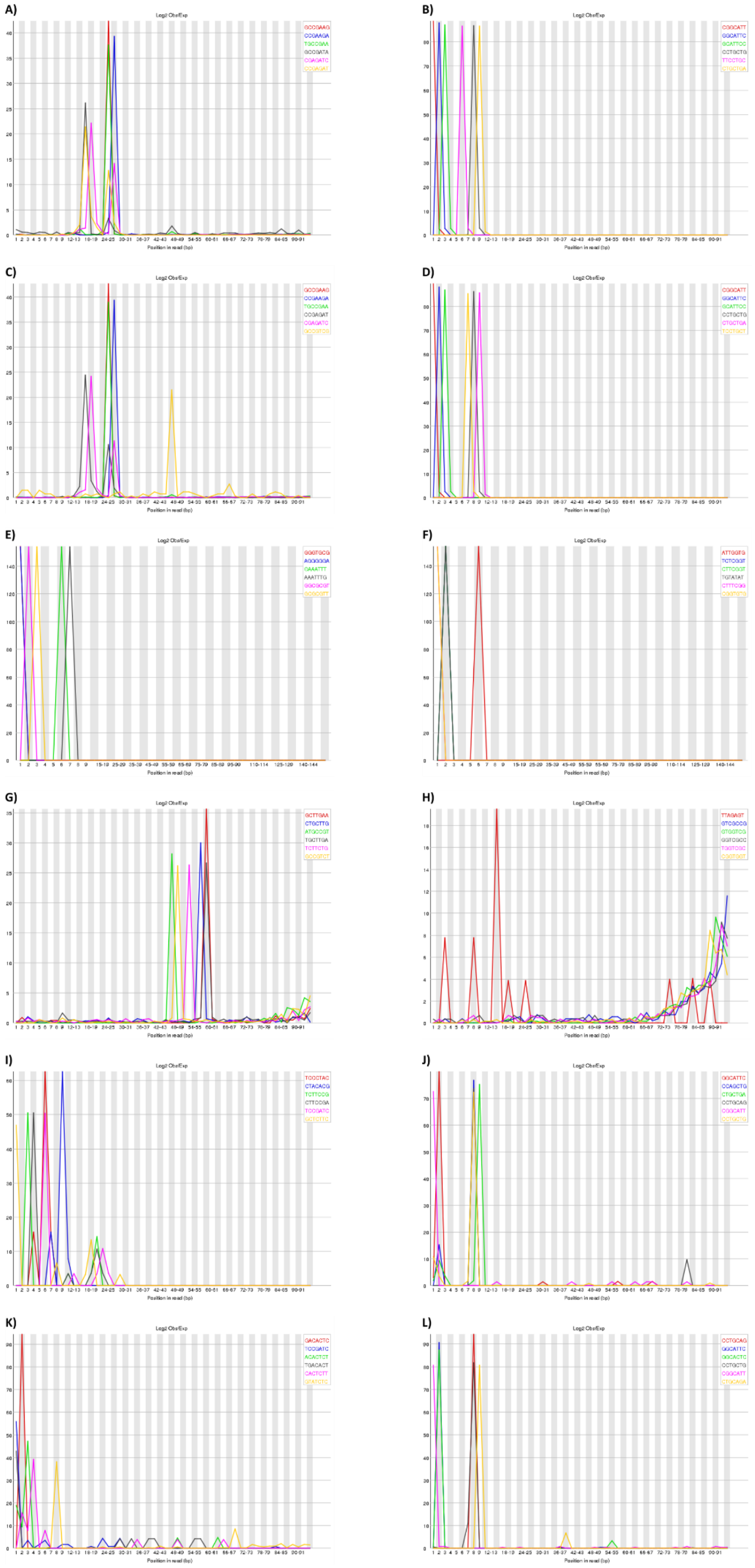
Supplementary Figure 2 - FASTQC reports of Dataset A, per base sequence content of SRR2034989 (A/B); SRR2034993 (C/D); SRR2079727 (E/F); SRR2173864 (G/H); SRR5115679(I/J); SRR5115685 (K/L)



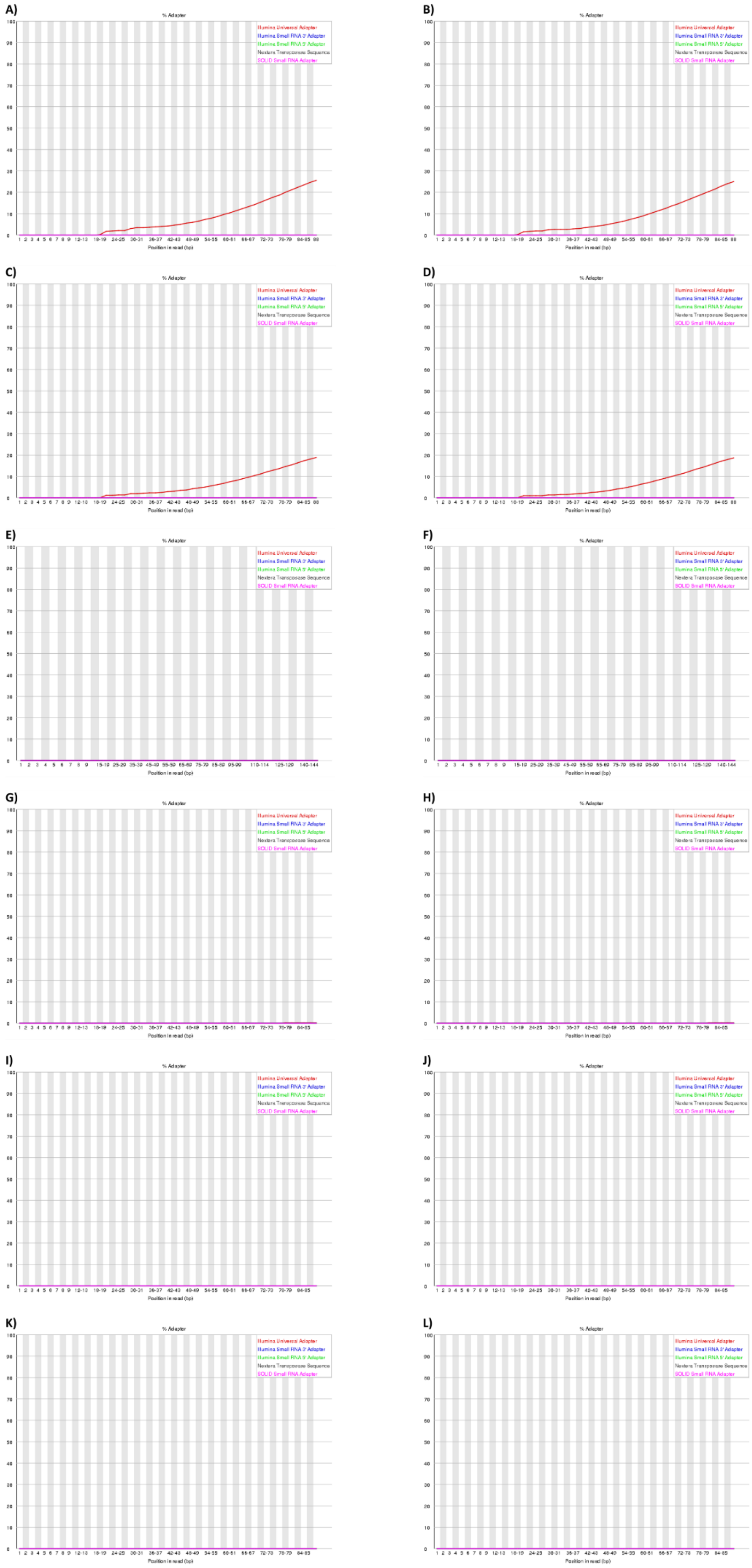
Supplementary Figure 3 - FASTQC reports of Dataset A, per base N content of SRR2034989 (A/B); SRR2034993 (C/D); SRR2079727 (E/F); SRR2173864 (G/H); SRR5115679(I/J); SRR5115685 (K/L)



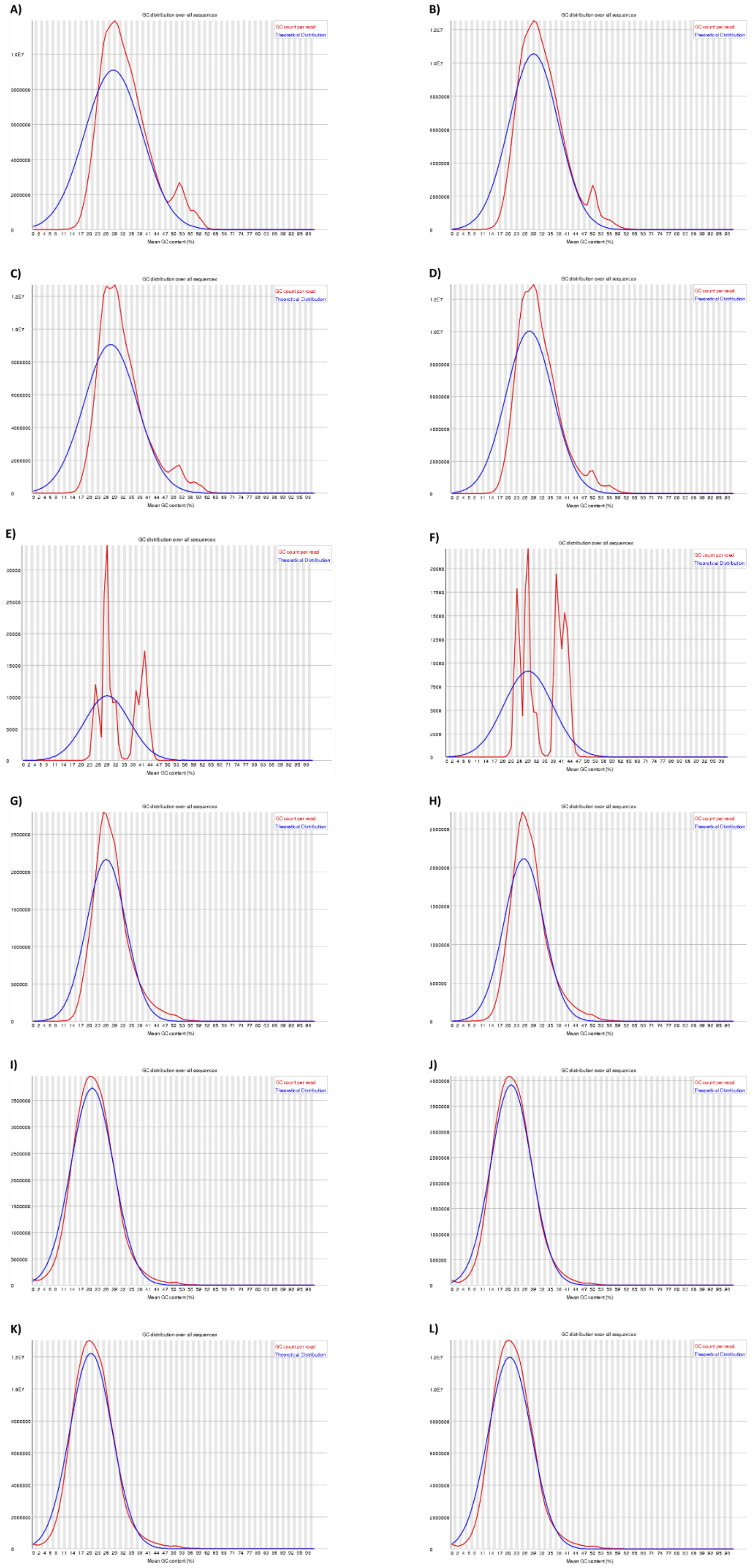
Supplementary Figure 4 - FASTQC reports of Dataset A, Kmer content of SRR2034989 (A/B); SRR2034993 (C/D); SRR2079727 (E/F); SRR2173864 (G/H); SRR5115679(I/J); SRR5115685 (K/L)



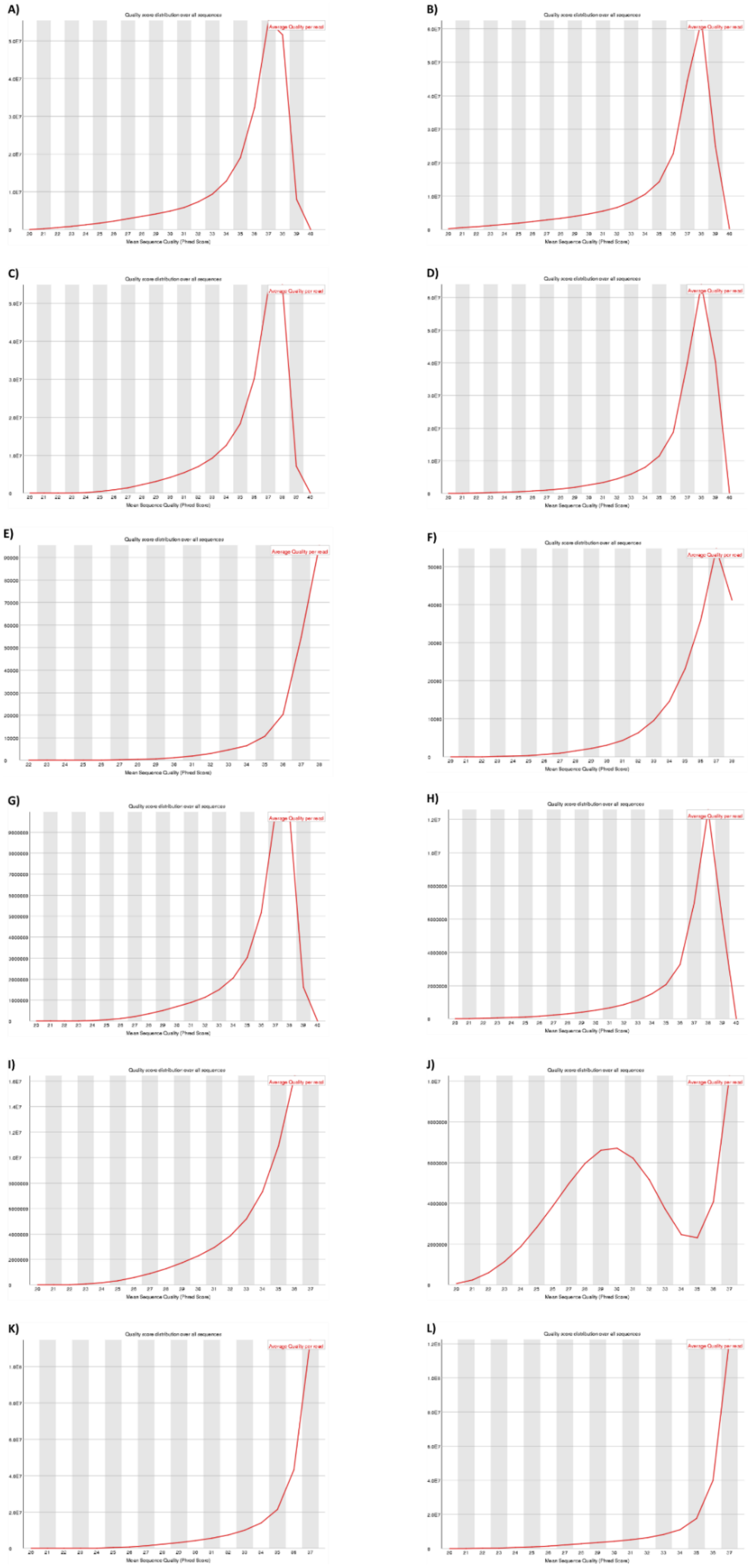
Supplementary Figure 5 - FASTQC reports of Dataset A, adapter content of SRR2034989 (A/B); SRR2034993 (C/D); SRR2079727 (E/F); SRR2173864 (G/H); SRR5115679(I/J); SRR5115685 (K/L)



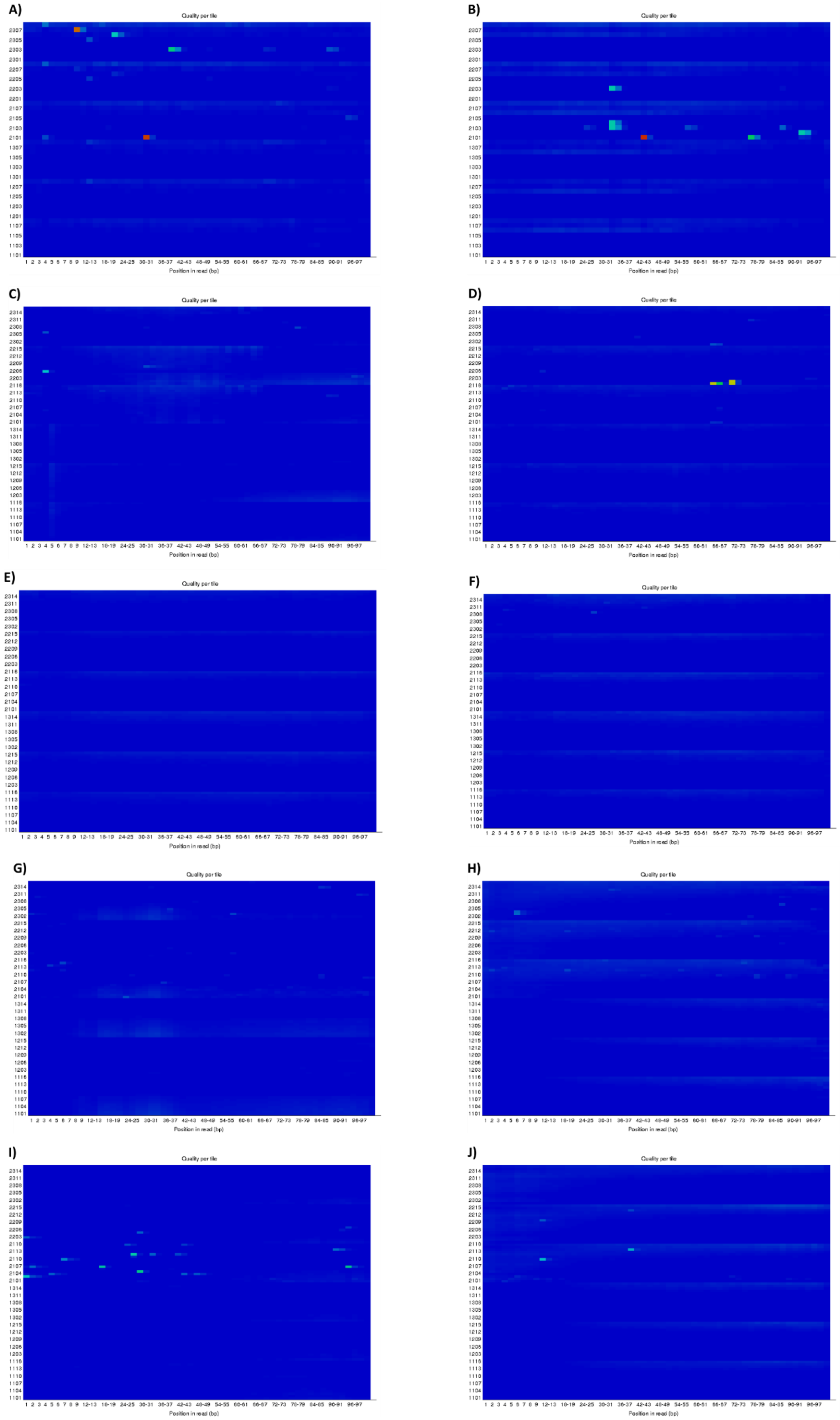
Supplementary Figure 6 - FASTQC reports of Dataset A, per sequence GC content of SRR2034989 (A/B); SRR2034993 (C/D); SRR2079727 (E/F); SRR2173864 (G/H); SRR5115679(I/J); SRR5115685 (K/L)



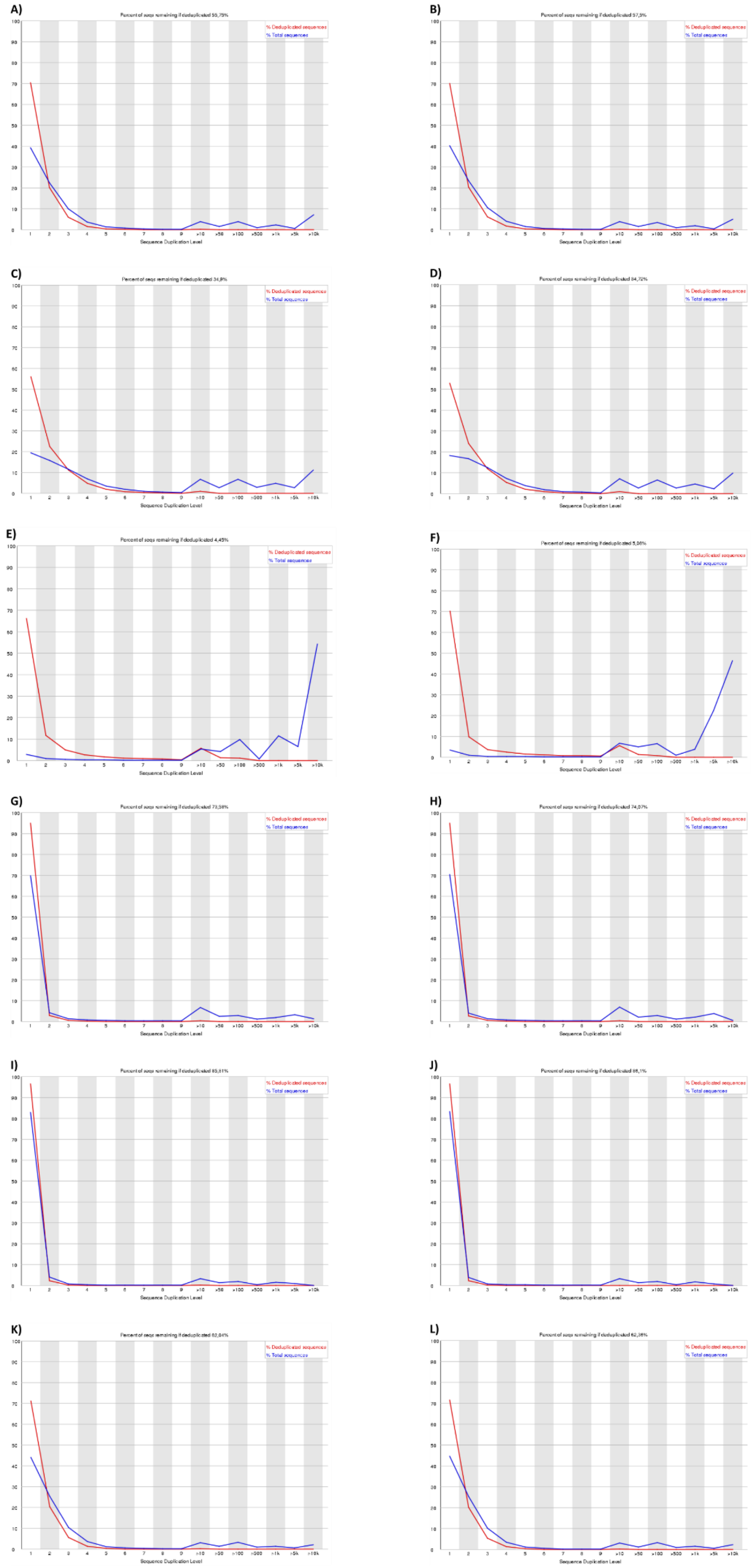
Supplementary Figure 7 - FASTQC reports of Dataset A, per sequence quality scores of SRR2034989 (A/B); SRR2034993 (C/D); SRR2079727 (E/F); SRR2173864 (G/H); SRR5115679(I/J); SRR5115685 (K/L)



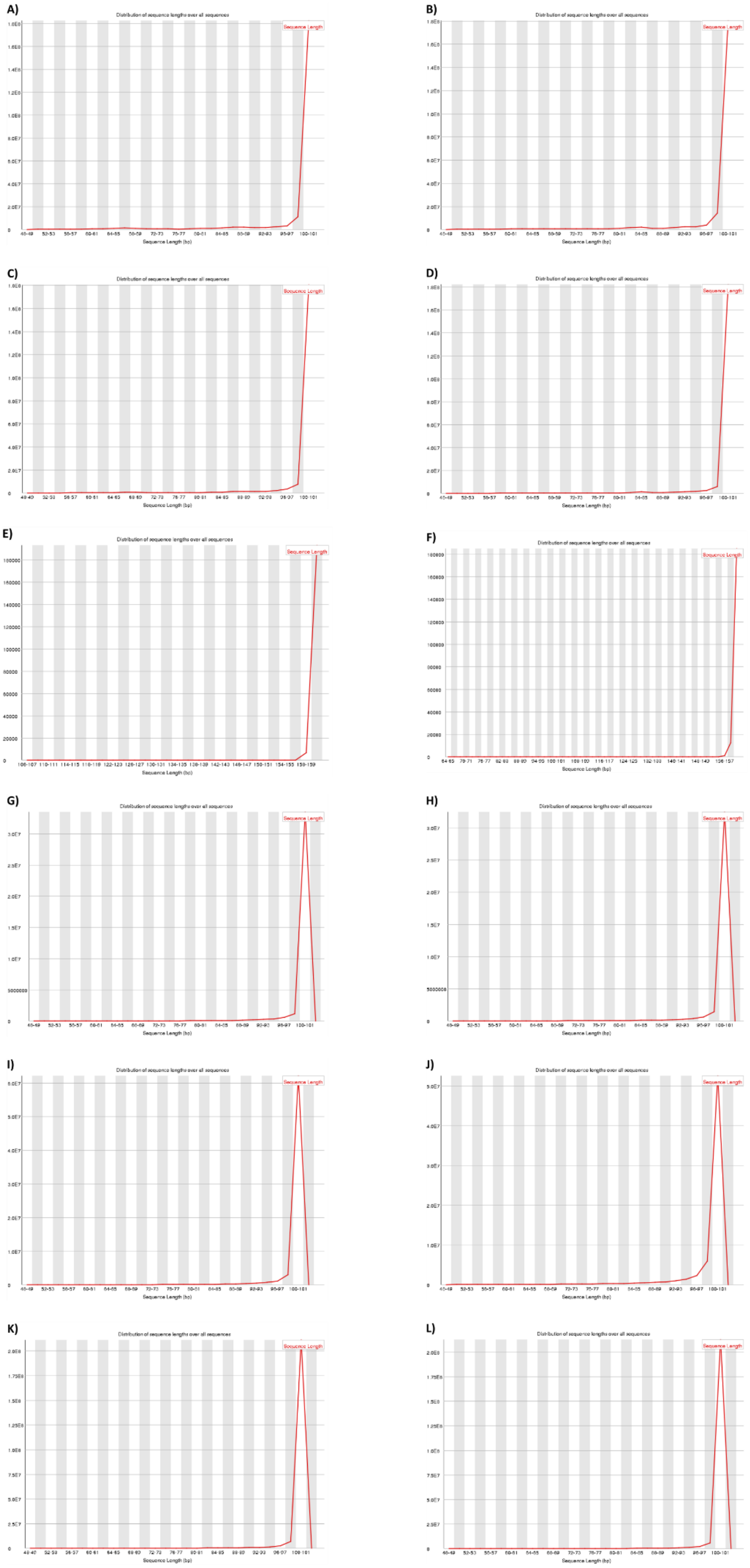
Supplementary Figure 8 - FASTQC reports of Dataset A, per tile sequence quality of SRR2034989 (A/B); SRR2034993 (C/D); SRR2173864 (E/F); SRR5115679(G/H); SRR5115685 (I/J)



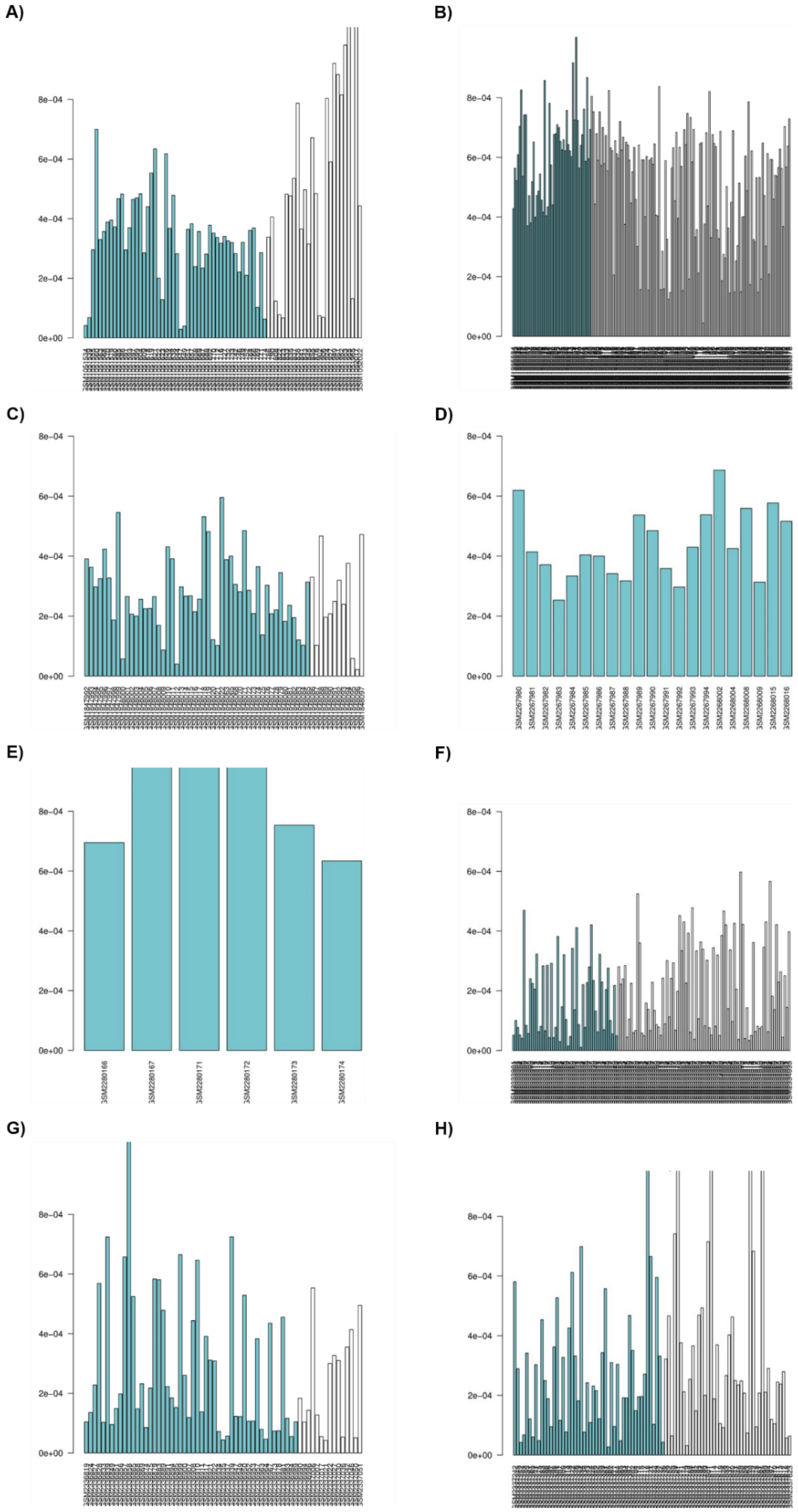
Supplementary Figure 9 - FASTQC reports of Dataset A, sequence duplication levels of SRR2034989 (A/B); SRR2034993 (C/D); SRR2079727 (E/F); SRR2173864 (G/H); SRR5115679(I/J); SRR5115685 (K/L)

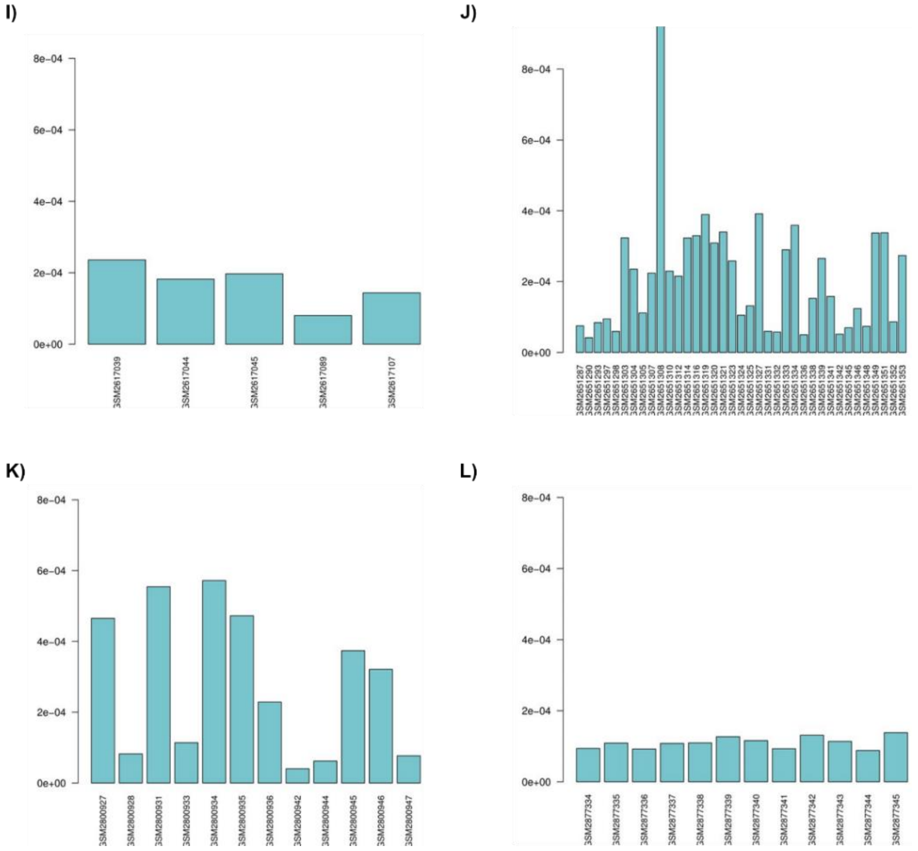


Supplementary Figure 10 - FASTQC reports of Dataset A, sequence length distribution of SRR2034989 (A/B); SRR2034993 (C/D); SRR2079727 (E/F); SRR2173864 (G/H); SRR5115679(I/J); SRR5115685 (K/L)

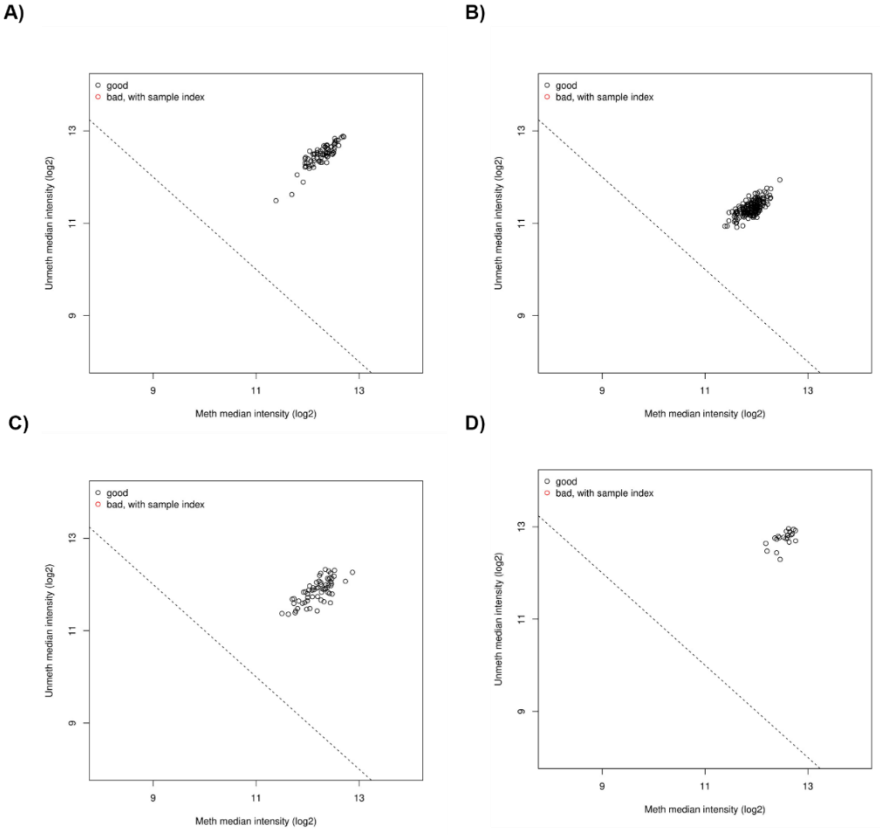


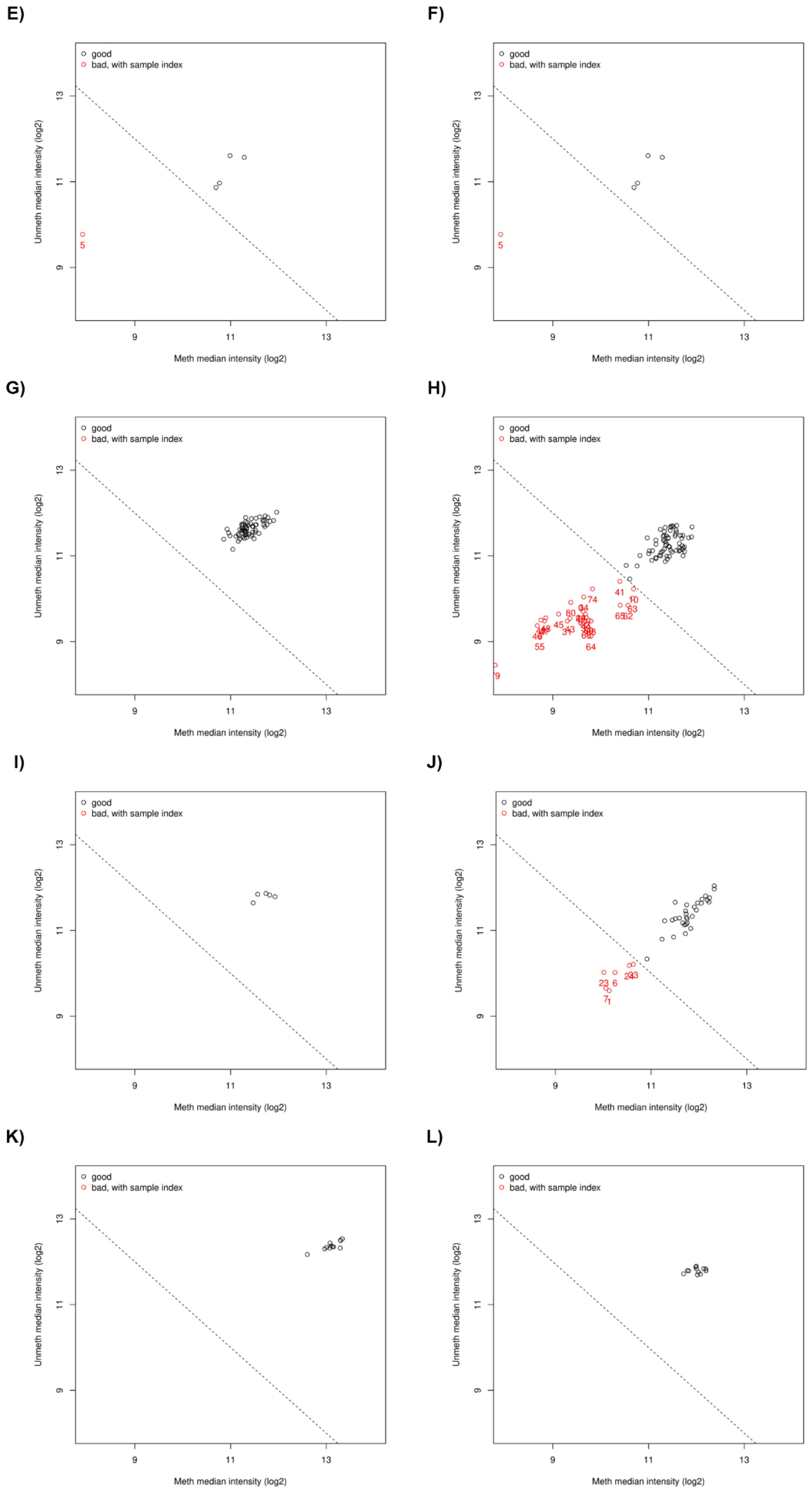
Supplementary Figure 11 – Mean detection p-values of (A) GSE42861, (B) GSE51057, (C) GSE71955, (D) GSE85506, (E) GSE85647, (D) GSE87571, (E)GSE87640, (F) GSE87648, (G) GSE98876, (H)GSE99766 (I) GSE104471, (J) GSE107737



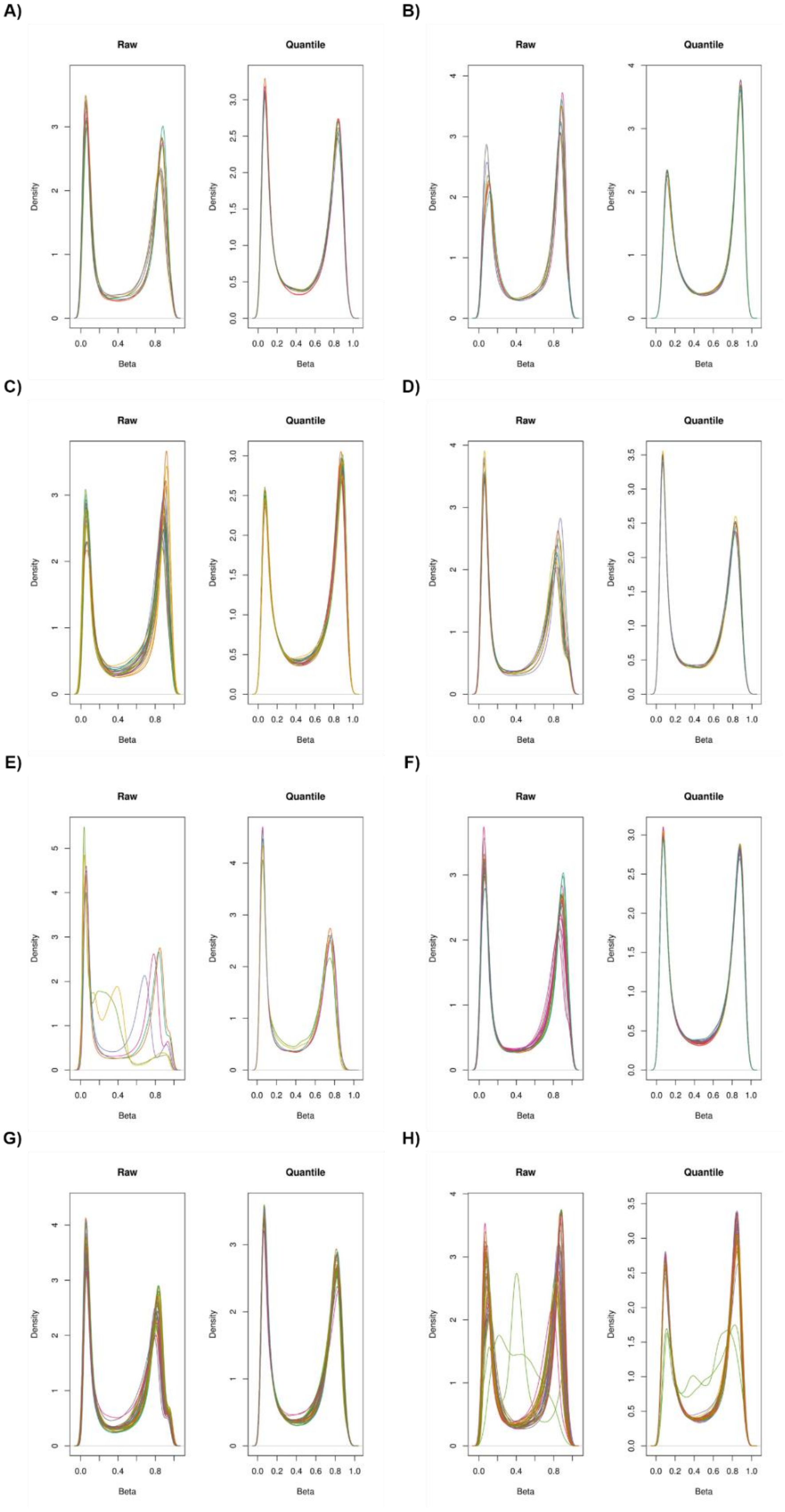


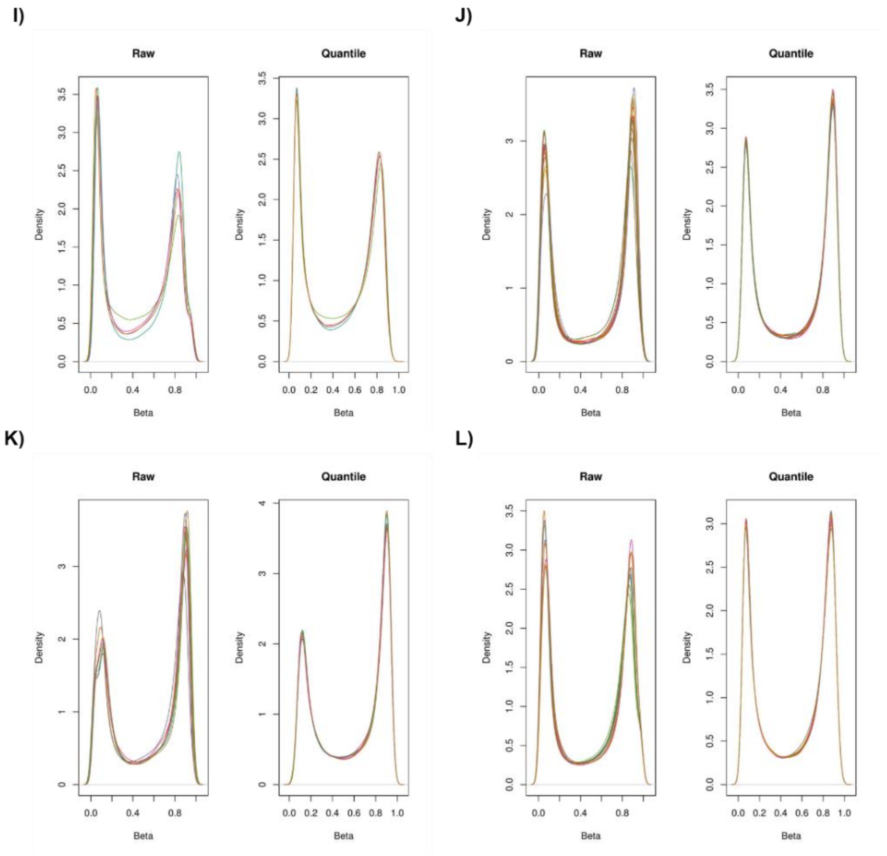
Supplementary Figure 12 – Quality control report of (A) GSE42861, (B) GSE51057, (C) GSE71955, (D) GSE85506, (E) GSE85647, (D) GSE87571, (E)GSE87640, (F) GSE87648, (G) GSE98876, (H)GSE99766 (I) GSE104471, (J) GSE107737



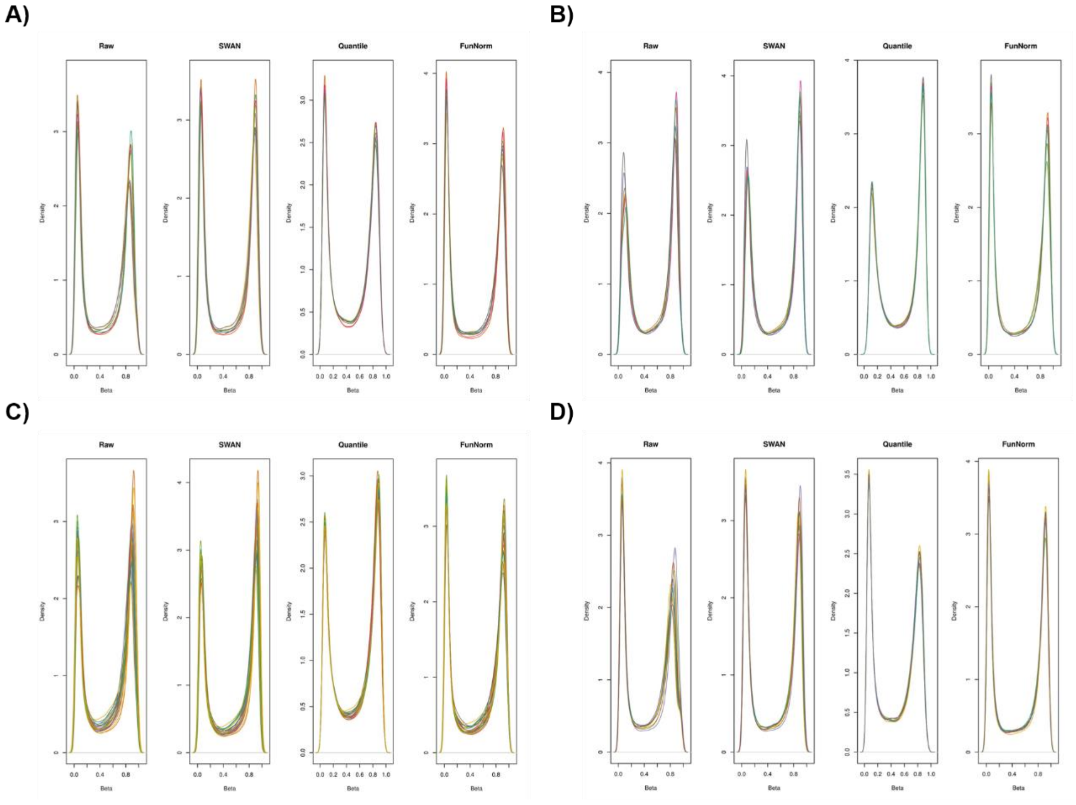

E)
F)
G)
H)
I)
J)
K)
L)
good
bad, with sample index
Unmeth median intensity (log2)
Meth median intensity (log2)
9
11
13
5
41
74
10
53
65
62
80
45
34
43
55
64
49
48
9
23
6
7
1
24
33

Supplementary Figure 13 – Comparison of data before and after normalization of (A) GSE42861, (B) GSE51057, (C) GSE71955, (D) GSE85506, (E) GSE85647, (D) GSE87571, (E)GSE87640, (F) GSE87648, (G) GSE98876, (H)GSE99766 (I) GSE104471, (J) GSE107737

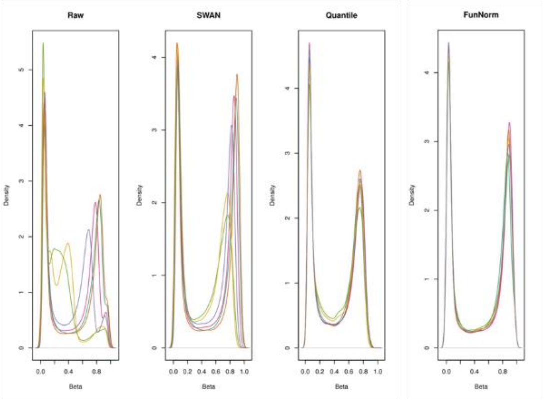




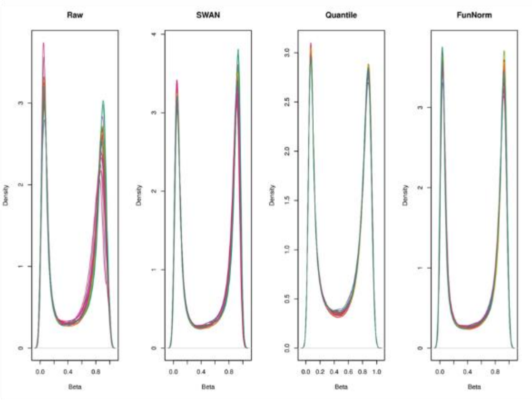
Supplementary Figure 14 – Comparison of the three available normalization methods of (A) GSE42861, (B) GSE51057, (C) GSE71955, (D) GSE85506, (E) GSE85647, (D) GSE87571, (E)GSE87640, (F) GSE87648, (G) GSE98876, (H)GSE99766 (I) GSE104471, (J) GSE107737



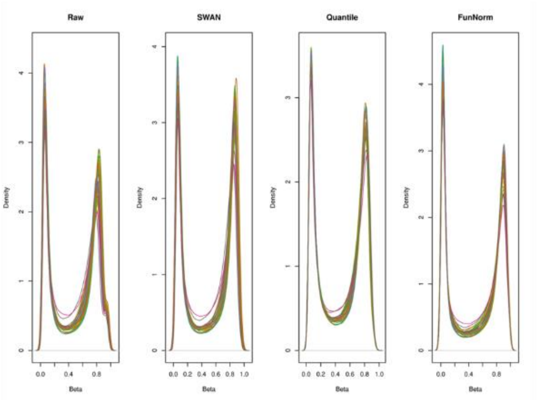
E)


Raw
SWAN
Quantile
FunNorm
Density
Beta

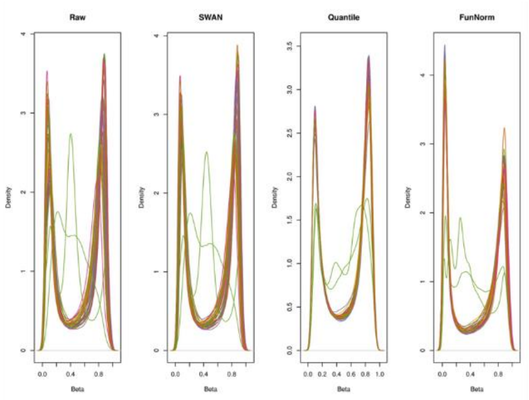
F)


Raw
SWAN
Quantile
FunNorm
Density
Beta

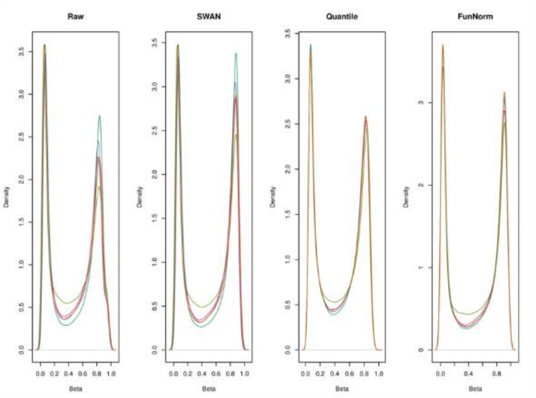
G)


Raw
SWAN
Quantile
FunNorm
Density
Beta

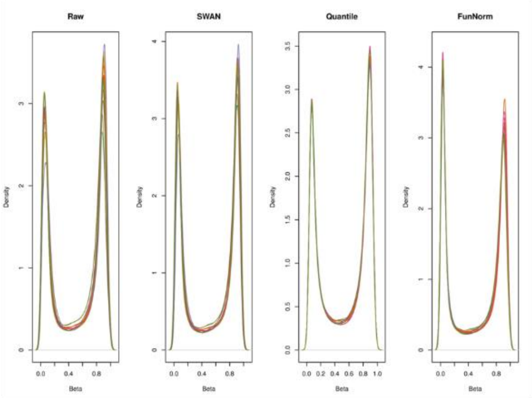
H)


Raw
SWAN
Quantile
FunNorm
Density
Beta

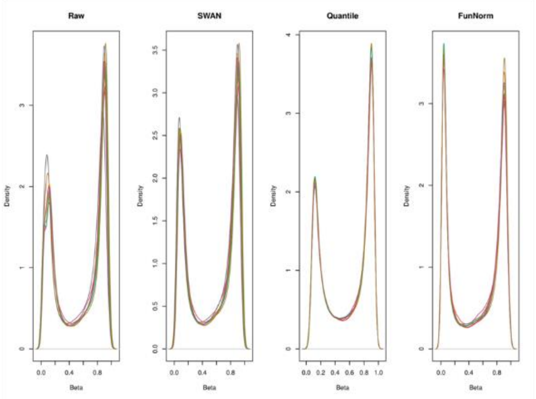
I)


Raw
SWAN
Quantile
FunNorm
Density
Beta

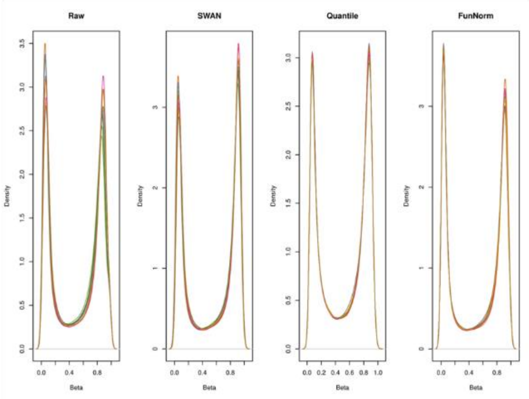
J)


Raw
SWAN
Quantile
FunNorm
Density
Beta

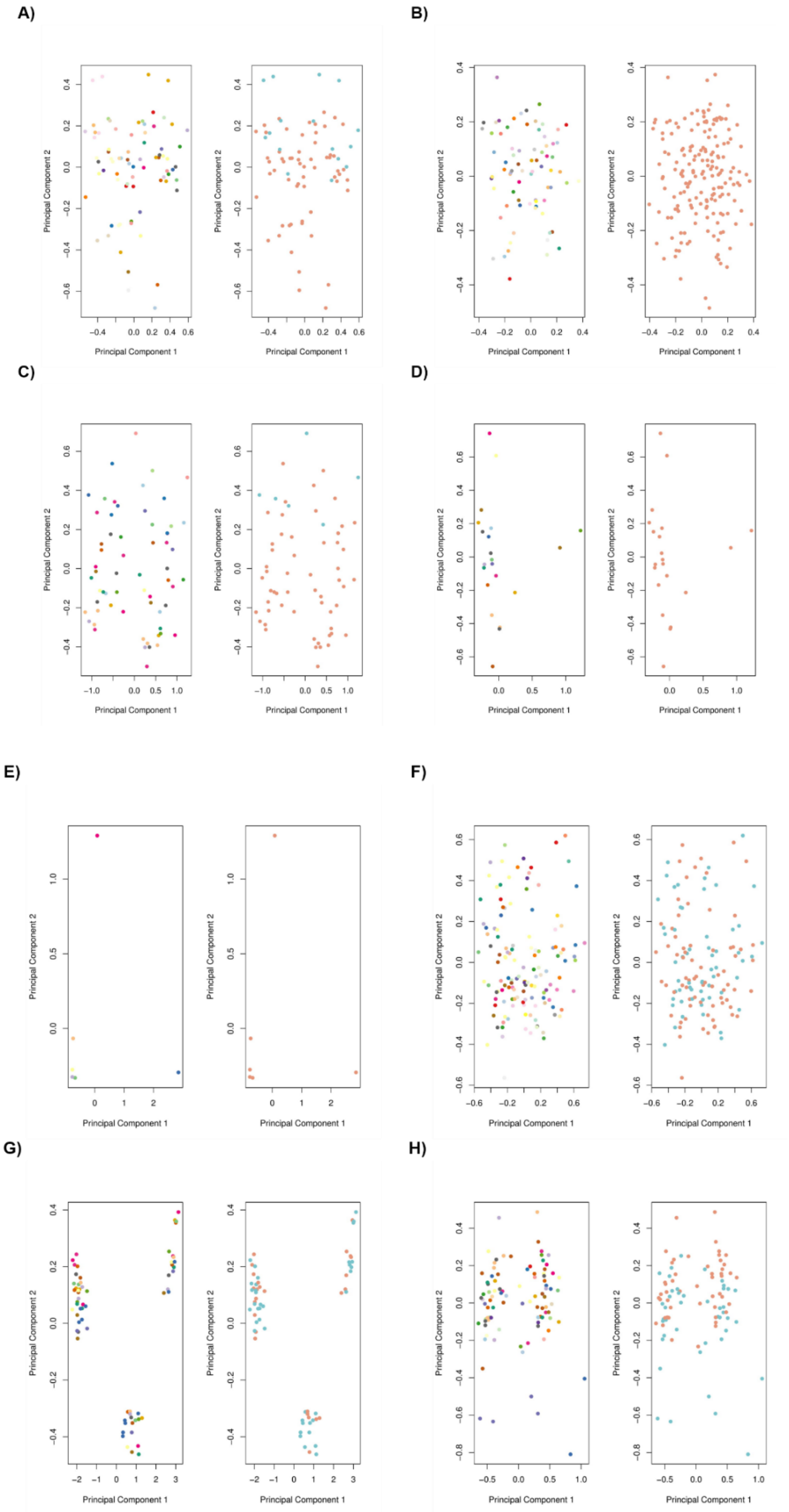
K)


Raw
SWAN
Quantile
FunNorm
Density
Beta

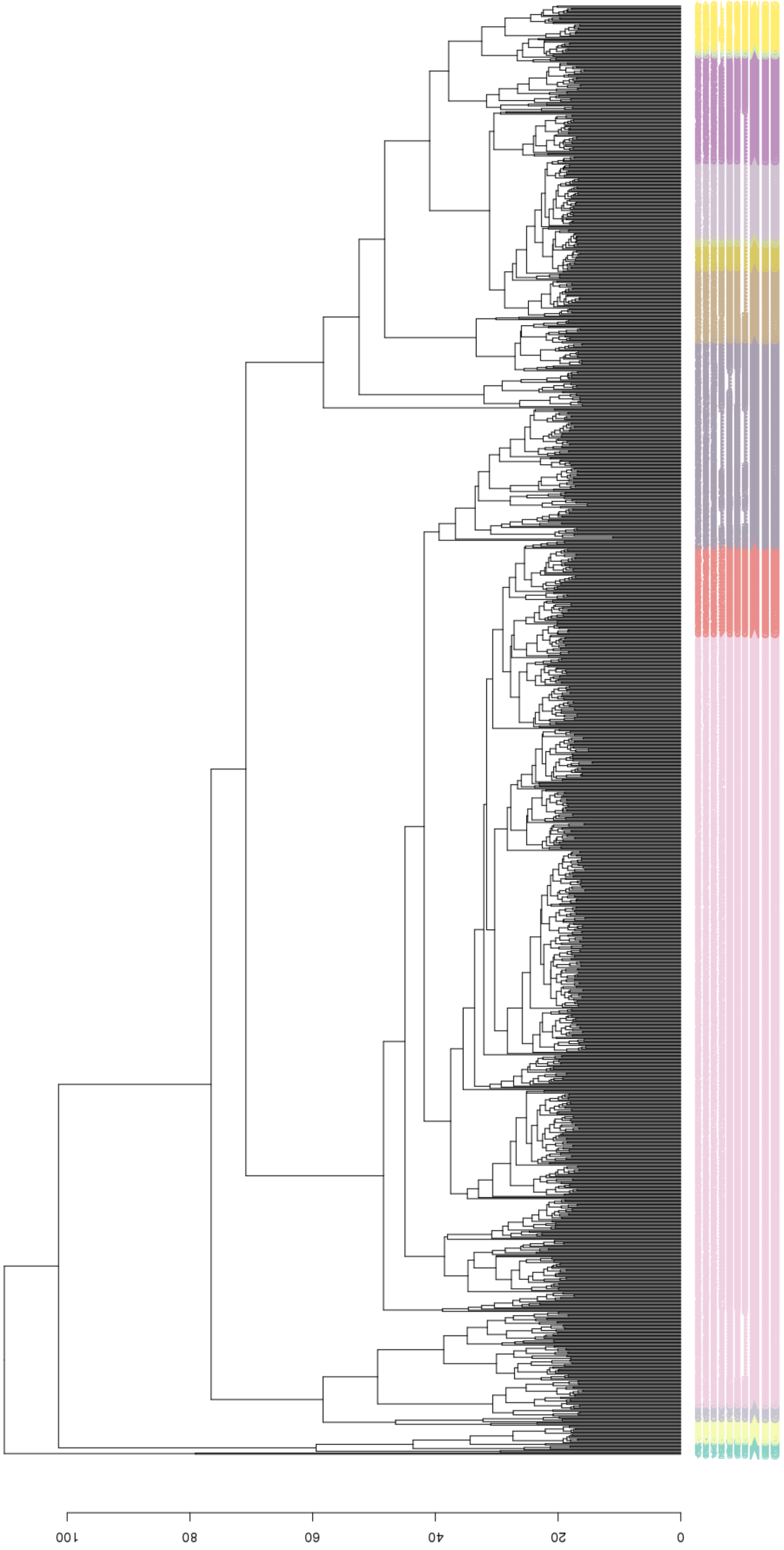
L)


Raw
SWAN
Quantile
FunNorm
Density
Beta

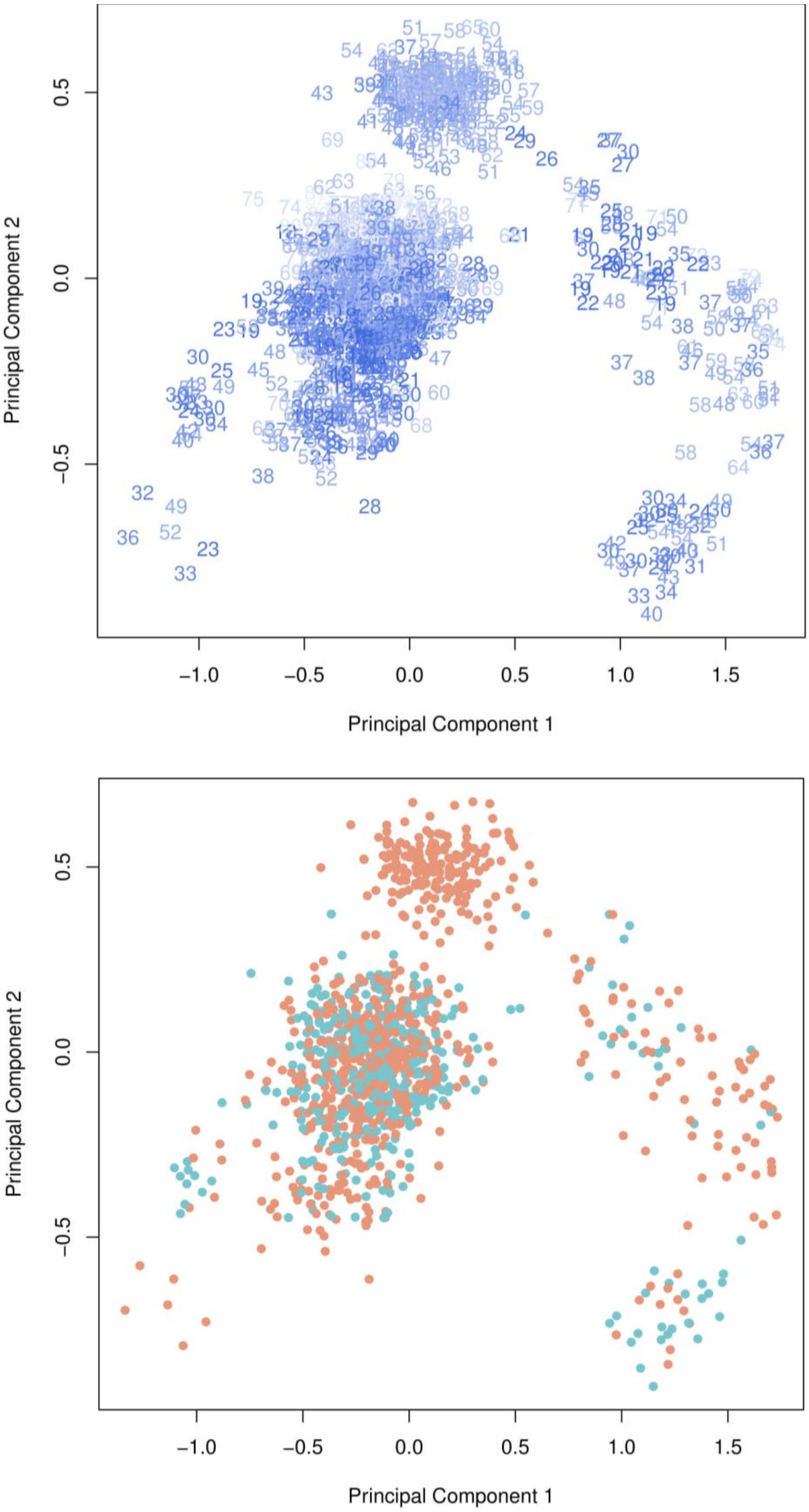
Supplementary Figure 15 – MDS plots after normalization and filtering of (A) GSE42861, (B) GSE51057, (C) GSE71955, (D) GSE85506, (E) GSE85647, (D) GSE87571, (E)GSE87640, (F) GSE87648, (G) GSE98876, (H)GSE99766 (I) GSE104471, (J) GSE107737



Supplementary Figure 16 – Clustering of a global virtual array used to test the junction of all datasets of Dataset B.2.



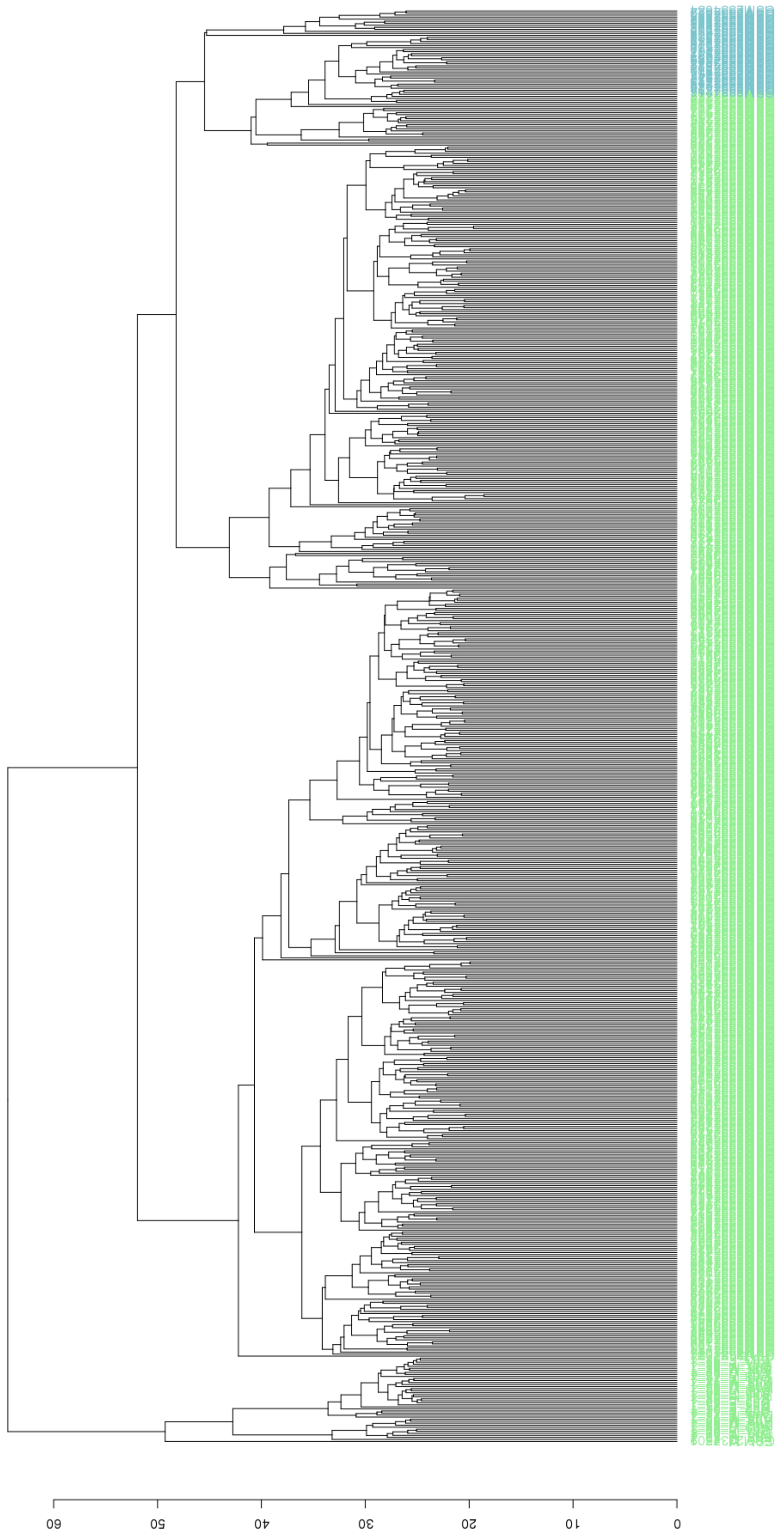
Supplementary Figure 17 – MDS plot of a global virtual array used to test the junction of all datasets of Dataset B.2.



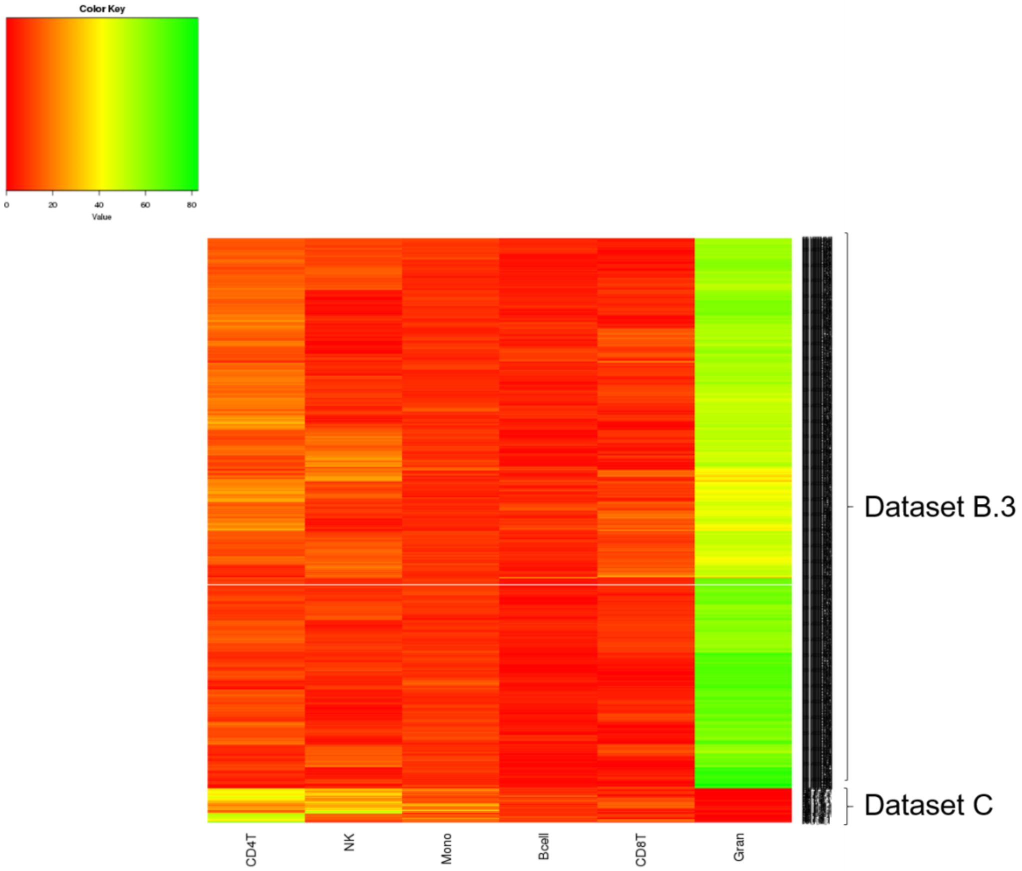
Supplementary Table 4 – Summary of the suggestive DMPs with age phenotype in Dataset C.

Probes	***slope***	***P_{adj}-value***	***Δβ***	***Chr***	***Position***	***Islands Name***	***Relation to Island***	***Gene Group***	***Gene Name***	***Gene Function***	***Log10***
cg06639320	0.00418	0.004183	0.0245	2	106015739	chr2:106014878-106015884	Island	1stExon, TSS200, TSS1500, 5’UTR	**FHL2**	Protein binding	6.71262
cg16867657	0.00157	0.007468	0.0087	6	11044877	chr6:11043913-11045206	Island	TSS200, TSS1500, 5’UTR	ELOVL2	Fatty acid elongase activity	6.46090
cg12662887	0.00838	0.07964	0,0181	10	105343920	chr10:105344173-105345039	N_Shore		NEURL	Transferase activity	5.43299
cg02314201	-0.01008	0.090036	-0.0453	10	134843775	chr10:134843465-134843776	Island				5.37971
cg24851859	-0.00293	0.118316	-0.0080	5	27532684		OpenSea				5.26108
cg06576732	0.00175	0.209114	0.0015	20	20344672	chr20:20344400-20350605	Island				5.01374

Supplementary Figure 18 – Clustering of Dataset B.3 (green) and Dataset C (blue), which make us obtain Dataset D



Supplementary Figure 19 - Heatmap of the cell composition of both datasets presented.


Color Key
Value
Dataset B.3
Dataset C
CD4T
NK
Mono
Bcell
CD8T
Gran