Master of Science in Omics Data Analysis

Master Thesis

Methylation Data Analysis Associated with Alzheimer's Disease

by

Jose Luis Mosquera

Supervisor : Professor Malu Calle Rosingana

Department of Systems Biology, University of Vic

Co-supervisor : Dr. Marta Barrachina Castillo

Bellvitge Biomedical Research Institute

Department of Systems Biology

University of Vic - Central University of Catalonia

September 2015

Preface

This Master's thesis was carried out at the lab of the Institute of Neuropathology–Bellvitge University Hospital (IDIBELL), and in collaboration with the Department of Systems Biology at the University of Vic–Central University of Catalonia, as a part of the program Master of Science in Omics Data Analysis.

The main idea of the project was brought up based on results obtained after a previous collaboration between the co-supervisor Dr. Marta Barrachina Castillo and the MSc candidate.

Readers of this report should be familiarized with DNA sequencing technologies used in omics experiments, as well as with statical concepts associated with Fisher's Exact Test and Odds Ratio, and programming concepts based on the R scripting language.

Vic, 2015-09-21

Jose Luis Mosquera

Acknowledgment

First and foremost, I would like to mention a special debt of gratitude to **Dr Marta Barrachina**, as the co–supervisor of this master's thesis, for having guided and supported me, correcting my scientific work with an interest and dedication that have greatly exceeded all the expectations that I could have had. Likewise, I would like to thank **Prof. Malu Calle** for her work as a supervisor and as a chair of the Master's degree in Omics Data Analysis.

It is also my deep personal desire to thank all the professors and lecturers of the master for everything they have taught me. I would not wish to neglect a big thank you to the collaborators of Dr Marta Barrachina, specially **Dr Marta Blanch**, at the IDIBELL.

Finally, I would like to express my eternal gratitude to **Antonio**, **Carmen**, **Chus**, my parents and my brother; they have always been with me, through thick and thin.

J.L.M.

Abstract

Alzheimer disease (AD) is associated with distinctive changes in DNA methylation patterns. DNA methylation is a chemical modification that is involved in gene silencing. These modifications give rise to heritable changes in gene expression without altering the DNA sequence. In a first study, researchers observed differential methylated regions (DMRs) in a specific DNA in postmortem brains from AD patients. Due to this fact we decided to focus on two target genes from AD in order to compare differences in methylation between different experimental situations. In this project, we have developed a pipeline for finding DMR between different experimental conditions based on methylation data generated with the 454 GS FLX system from Roche, and DNA bisulfite-treated samples.

Contents

List of Tables

List of Figures

Glossary

AD Alzheimer's Disease. 2, 7, 10

API Application Programming Interface. 38

CI Confidence Interval. 34, 35

DMR Differentially Methylated Region. 10, 11, 13, 17, 18, 29, 33, 37

DNA Deoxyribonucleic Acid. vii, 3–10, 14, 28

FDR False Discovery Rate. 35

MID Multiple Identifiers. 14, 22, 24, 25, 27, 29, 30

NGS Next-Generation Sequencing. 7, 8

OR Odds Ratio. 33–35

RNA Ribonucleic Acid. 3

TSS Transcription Start Sites. 7

Chapter 1

Introduction: Scope of the Thesis

Alzheimer's Disease (AD) is the most common cause of dementia in humans [Knopman (2011)]. It progress gradually and causes a irreversible neurodegenerative disorder, that ends with dementia [Mastroeni et al. (2011)]. Dementia diagnosis due to AD is established using clinical criteria [McKhann et al. (1984)]. Two of the most important risk factors considered for diagnosis are advancing age and other cases in a family history. However, etiology and pathogenesis of AD remains up in the air because they are complex. What has been proved is that large amounts of neurofibrillary tangles and extracellular accumulation of β-amyloid peptide in the form of extracellular senile plaques and blood vessel deposits, characterize and confirm the diagnosis of AD, as evidenced by autopsies of affected brains [Bertram and Tanzi (2011)]. However, currently, there are not yet available biomarkers for AD being used as a routine diagnostic approach [Knopman (2011)]. Even so, while etiology and pathogenesis of AD still remains up in the air, actually, they involve many environmental risk factors, genetic mutations, direct regulation of gene expression or epigenetic modifications that are able to stimulate changes in phenotype by altering transcriptional activity in multiple genes encompassing several biological pathways [Mastroeni et al. (2011)]. For all that during the last decade scientific community have focused on finding and incorporating biomarkers for AD into diagnosis [Knopman (2011)].

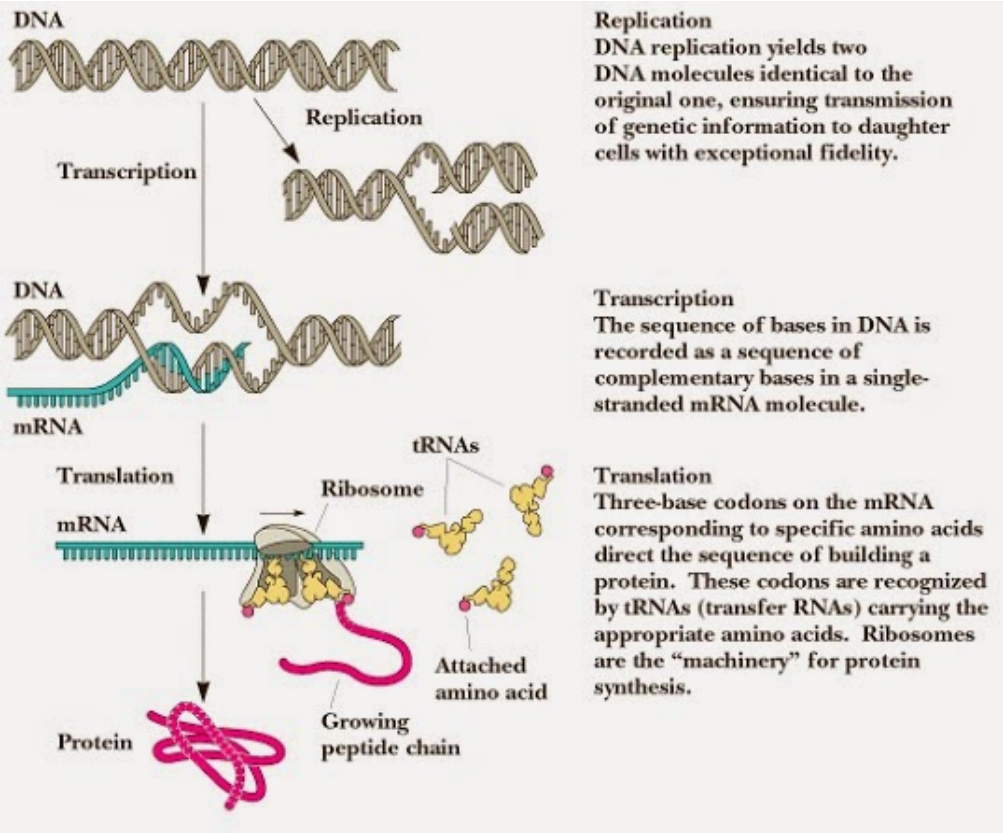

Figure 1.1: Central Dogma of Molecular Biology. Courtesy of Bio-Resource.

1.1 Principles of Epigenetic Mechanisms

Often the study of a human disease is focused on genome mechanisms[1]. However, it is unlikely to understand the cellular mechanisms involved in complex diseases, such as cancer or Alzheimer, just by studying only the DNA sequences on a linear genome.

The Central Dogma of Molecular Biology provides a basic explanation of the flow of the information within a biological system [Crick (1970)]. It states that DNA of a cell is transcribed to RNA, and then it is translated to proteins, that perform specific biological processes and molecular functions (Figure 1.1). However, in multicellular organisms, like in humans, cells sharing identical DNA sequences can have different functions in response to alterations in the environment (e.g. age, gender, lifestyle, or a disease state). Usually, these variations are called epigenetic changes.

[1]The human genome is the complete set of genetic information of Homo sapiens [Brown (2002)]. It is encoded as DNA in 23 chromosome pairs found in cell nuclei and a small DNA molecule within a individual mitochondria. It consists of approximately 3.2 billion DNA base pairs.

1.1.1 Epigenetics

Epigenetics is the study of heritable changes in gene expression that are not due to mutations in the DNA sequence [Eccleston et al. (2007), Bird (2007) and Goldberg et al. (2007)]. In other words, they are modifications in the phenotype with no changes in the genotype. These changes can be assumed and accumulated by an person, and then they can be passed from generation to generation.

Epigenetic mechanisms regulate a large number of essential biological processes, such as those responsible for cell division or cell differentiation [Bonetta (2008)]. For instance, two of the major biological processes that result in epigenetic modifications to the genome are: DNA methylation and histone modification (section 1.1.3) [Goldberg et al. (2007), Bonetta (2008) and Portela and Esteller (2010)]. These mechanisms are also involved in some complex diseases like cancer [Jones and Baylin (2007), Agrawal et al. (2007), and Colnot et al. (2004)], diabetes [Ling and Groop (2009) and Hanson and Godfrey (2015)], Parkinson [Feng et al. (2015) and Kaidery et al. (2013)] or, of course, Alzheimer [Bennett et al. (2015), Marques and Outeiro (2013) and Portela and Esteller (2010)]. The essential idea is that epigenetic mechanisms can be influenced by different factors and/or processes (Figure 1.2). That is, different exposures to environmental factors, drugs, diets, aging, etc. during the development and throughout the whole life of a person can modify DNA and the proteins bound to the genome [Bonetta (2008)].

1.1.2 Epigenomics

With the advent of the omic era huge quantities of information have been generated, and it has become a major breakthrough for molecular biology. This revolution began with the deciphering of the whole genome sequences of several organisms –among them the human genome–, and rapidly, similar ideas were applied to the study of the transcriptome, proteome and metabolome. This resulted in the emergence of omic studies: genomics, transcriptomics, proteomics and metabolomics (Figure 1.3) [Mosquera (2014)]. But, this revolution has been possible thanks to a new generation of high-technologies known as *high-throughput technologies*. These technologies allow the performance, in a routine way, of new types of experiments to analyze simultaneously the behavior of thousands of features under different conditions. There

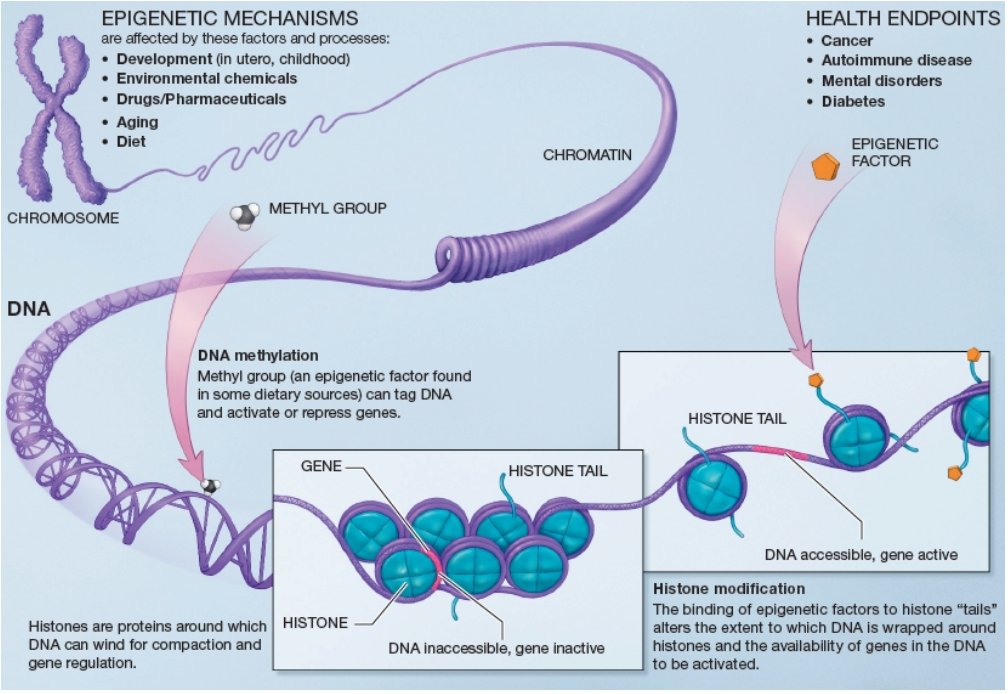

Figure 1.2: Illustration of how epigenetic mechanisms and modifications can influence DNA coiling around histones causing heritable changes. Courtesy of Wiki Commons.

are different types of high-throughput technologies (e.g. microarrays, next generation sequencing or mass spectrometry) that allow the performance of a broad range of omic experiments. In the case of epigenetics the irruption of such technologies and different bioinformatic resources, storing and mining an overwhelming quantity of data, made way for the epigenomics.

Epigenomics is defined as, "the study of epigenetic modifications across the whole genome" [Bonetta (2008)], which is referred as the *epigenome*. To know the epigenome is important because it provides information for when some of the proteins encoded by the 25,000 genes of the human genome are produced, and where this production takes place (i.e in which cells or tissues).

1.1.3 Epigenetic Modifications

In eukaryote cells, genomic DNA is packaged with histone proteins into chromatin to form structures that make up chromosomes (Figure 1.2). This composition is the key for regulating the accessibility to gene and the corresponding function, which is controlled by epigenetic modifications. Probably the most important epigenetic modifications are DNA methylation and

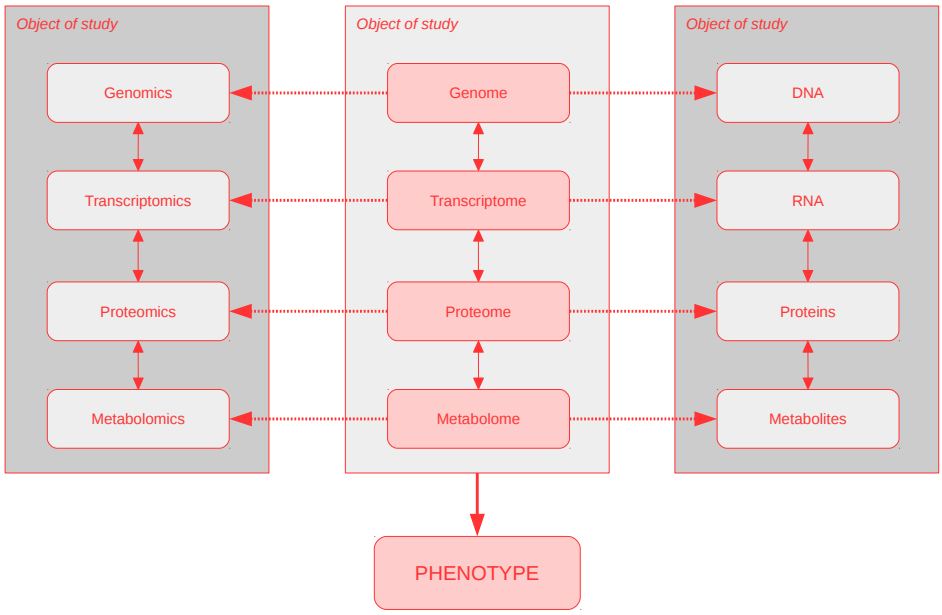

Figure 1.3: Illustration of the relationships between genome, transcriptome, proteome and metabolome in omic studies. Courtesy of Mosquera (2014).

histone modifications. Histones are proteins that DNA winds around to compact into chromosomes and contain certain chemical tags [Bonetta (2008)]. Different chemical modifications on the histones have an effect on the genic control. For example, a histone modification can occurs when a chemical modification in histone tails do not let the DNA coil as tightly, causing that the availability of genes in the DNA is activated, and in consequence the mRNA is transcribed. Now, on the DNA methylation what happens is the opposite effect. That is, when a methyl group is added to a cytosine residue, the DNA is coiled more tightly, due to the binding of DNA–methyl–binding proteins that recruit cofactors related to gene repression, and as a result the mRNA cannot be transcribed.

1.2 DNA Methylation Analysis

DNA methylation is the most well–characterized epigenetic modification [Laird (2010), Bonetta (2008) and Feinberg and Tycko (2004)]. It is involved in gene silencing [Wood (2014), Feinberg

and Tycko (2004) and Laird (2010)], and it is often found in the contexts CpG or CpHpG, where H can be A, T or C [Wood (2014) and Laird (2010)]. Moreover, it has been also observed that in embryonic stem cells and in neural development, non-CpG methylation is prevalent [Lister et al. (2013), Lister et al. (2009), and Dodge et al. (2002)].

DNA methylation analysis experimented a qualitative leap forward with the advent of high-throughput technologies [Jones (2012) and Feinberg and Tycko (2004)]. This fact allowed to introduce improvements on the genome-wide mapping of methylation [Jones (2012)]. Current methods are able to measure DNA methylation at a different genomics contexts (e.g. transcriptional start sites (TSS) with or without CpG islands, in gene bodies, at regulatory elements and at repeat sequences. This fact has helped to prove that methylation in cancer is perturbed. Hence, researchers moved the focus on other diseases like AD. It is reasonable that mapping CpG methylation can highlight genetic/epigenetic interactions and in consequence to help finding disease risk loci.

There is a large list of methods for detecting such changes in methylation across the entire genome (Table 1.4) [Laird (2010)]. In this report, we only focus on the bisulfite conversion combined with the Next Generation Sequencing (NGS) technologies [Korshunova et al. (2008) and Cokus et al. (2008)].

Pretreatment	Analytical step			
	Locus-specific analysis	Gel-based analysis	Array-based analysis	NGS-based analysis
Enzyme digestion	• *Hpa*II-PCR	• Southern blot • RLGS • MS-AP-PCR • AIMS	• DMH • MCAM • HELP • MethylScope • CHARM • MMASS	• Methyl–seq • MCA–seq • HELP–seq • MSCC
Affinity enrichment	• MeDIP-PCR		• MeDIP • mDIP • mCIP • MIRA	• MeDIP–seq • MIRA–seq
Sodium bisulphite	• MethyLight • EpiTYPER • Pyrosequencing	• Sanger BS • MSP • MS-SNuPE • COBRA	• BiMP • GoldenGate • Infinium	• RRBS • BC–seq • BSPP • WGSBS

AIMS, amplification of inter-methylated sites; BC–seq, bisulphite conversion followed by capture and sequencing; BiMP, bisulphite methylation profiling; BS, bisulphite sequencing; BSPP, bisulphite padlock probes; CHARM, comprehensive high-throughput arrays for relative methylation; COBRA, combined bisulphite restriction analysis; DMH, differential methylation hybridization; HELP, *Hpa*II tiny fragment enrichment by ligation-mediated PCR; MCA, methylated CpG island amplification; MCAM, MCA with microarray hybridization; MeDIP, mDIP and mCIP, methylated DNA immunoprecipitation; MIRA, methylated CpG island recovery assay; MMASS, microarray-based methylation assessment of single samples; MS-AP-PCR, methylation-sensitive arbitrarily primed PCR; MSCC, methylation-sensitive cut counting; MSP, methylation-specific PCR; MS-SNuPE, methylation-sensitive single nucleotide primer extension; NGS, next-generation sequencing; RLGS, restriction landmark genome scanning; RRBS, reduced representation bisulphite sequencing; –seq, followed by sequencing; WGSBS, whole-genome shotgun bisulphite sequencing.

Figure 1.4: Table describing main principles of DNA analysis. Courtesy of Laird (2010).

1.2.1 Bisulfite Sequencing

NGS in combination with bisulfite treatment is probably the most widely used technique for studying DNA methylation modifications [Chatterjee et al. (2012)]. Basically, bisulfite treatment (*aka* sodium bisulfite treatment or simply bisulfite conversion) of DNA converts non-methylated cytosines (C) to uracils (U), but leaves 5-methylcytosine (5mC) residues unaffected. In oder words, bisulfite conversion introduces modifications in the DNA sequence that depends on the methylation status of specific cytosines residues. Then by applying sequencing methods on the bisulfite-treated DNA, one can determine methylation status of CpG sites.

Figure 1.5 shows the outline of bisulfite treatment of DNA, where nucleotides colored in red indicate 5-methylcytosines resistant to conversion and nucleotides colored in blue are unmethylated cytosines converted to uracils by the bisulfite treatment.

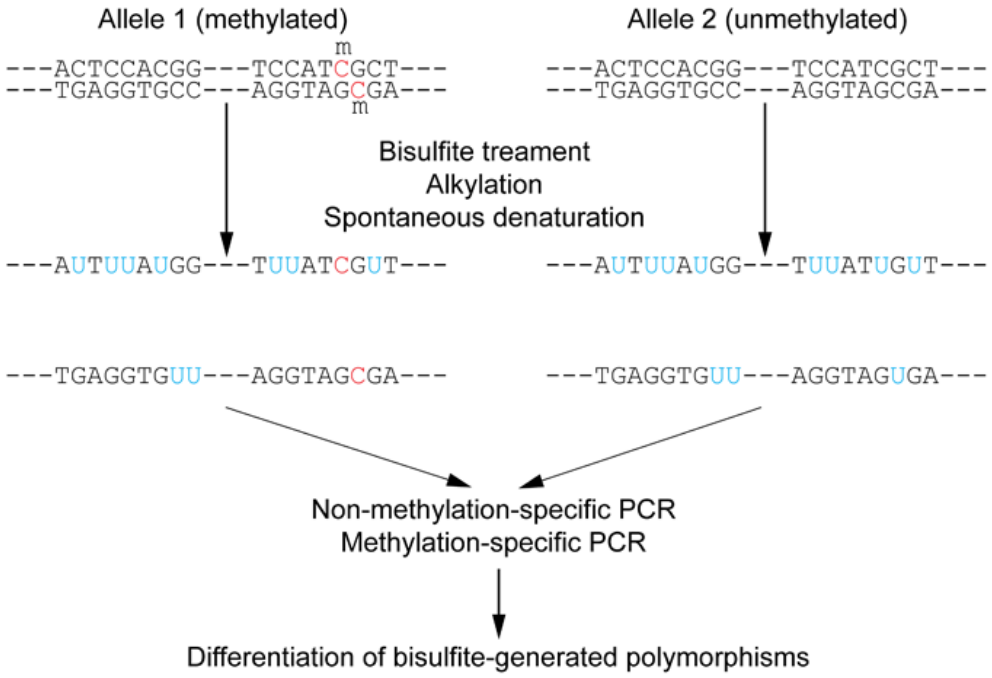

Figure 1.5: Illustration of the outline of bisulfite treatment of genomic DNA. Courtesy of Wiki Commons.

1.2.2 Downstream Analysis

Methylation data analysis can be performed using different types of tools and configurations [Laird (2010), Henry et al. (2014), and Chatterjee et al. (2012)]. Such analyses depends on the objectives of the study, the target sequences, the type of methylation protocol, and the high-throughput technology for sequencing the DNA (Figure 1.6). The overall steps for processing these kinds of data have been quite well established during the last years, and the main differences between them are the downstream analyses [Laird (2010)].

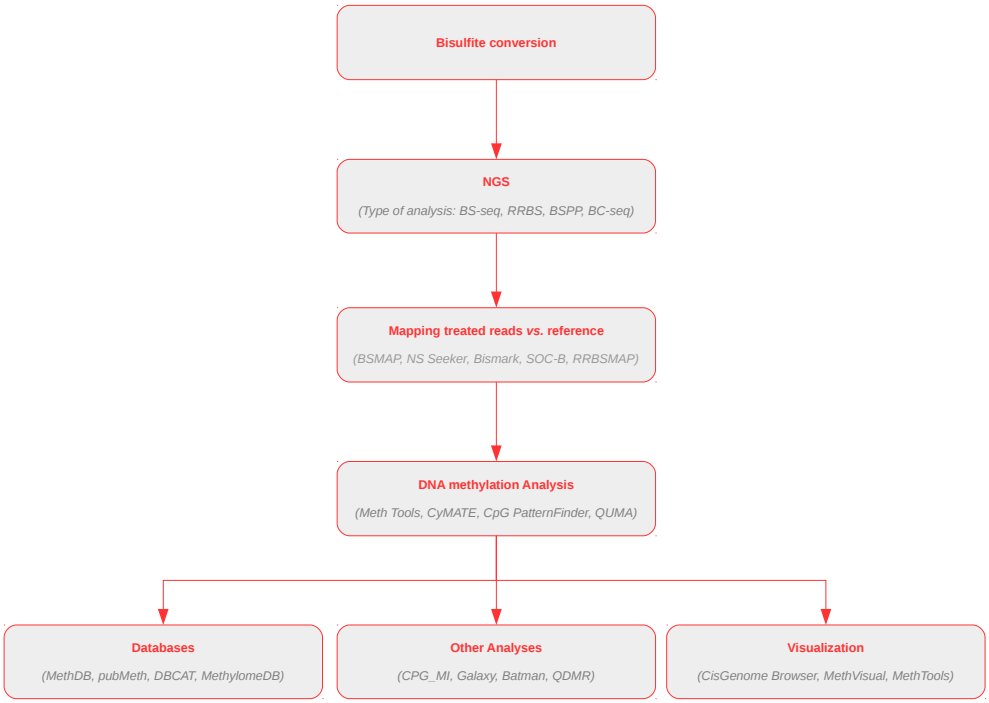

Figure 1.6: Different existing bioinformatic resources for a methylation data analysis based on a bisulphite conversion.

Chapter 2

Hypothesis, Objectives and Outline of the Thesis

Alzheimer disease (AD) is associated with distinctive changes in DNA methylation patterns 1. DNA methylation is a chemical modification that is involved in gene silencing. These modifications give rise to heritable changes in gene expression without altering the DNA sequence.

2.1 Hypothesis: Knowledge Gap

In a previous study, researchers observed differential methylated regions (DMR) in a specific DNA in postmortem brains from AD patients. Considering these results, they decided to focus on two target genes from AD in order to compare differences in methylation between different experimental situations. Due to this fact, researchers would need to have a pipeline for finding DMR. Due to this fact, different tools and approaches for performing this kind of analysis were reviewed from the literature [Laird (2010), Henry et al. (2014), Gentleman et al. (2004), Lutsik et al. (2011), Krueger and Andrews (2011), Xi and Li (2009), and Chen et al. (2010), among others]. However, none of the strategies reviewed were well adaptable to the purposes of the researchers. Some of them are devoted to look for whole genome differences, other are intended to work with sequencing system from other companies (e.g. illumina sequencing systems, or other are focused on different downstream analyses 1.2.2. Therefore, a pipeline for finding DMR based on bisulfite sequencing using a 454 GS FLX system from Roche for target DNA genomic regions

was required. Thus, in order to shed light on this issue, the main and specific objectives that this master's thesis will address are presented in the following sections.

2.2 Objectives

The context of this master thesis is focused on the methods and tools that have been developed, implemented and used to build a specific pipeline for finding a targeted DNA methylation Data Analysis, mainly associated with Alzheimer's Disease.

2.2.1 Main Objective

The main objective of this master's thesis is to develop and implement a pipeline for finding Differentially Methylated Regions (DMR) between different experimental conditions of a specific region of a DNA sequence, whose samples have been treated with bisulfite and sequenced with a 454 GS FLX system from Roche.

2.2.2 Specific Objectives

In order to accomplish the main objective of the project, specific questions were proposed by researchers. Specifically, the specific objectives of this master's thesis are:

1. To find DMR in a target gene 1 between:
   (a) a experimental condition A and a control group.
2. To find DMR in a target gene 2 between:
   (a) a experimental condition B and a control group.
   (b) a experimental condition D and a control group.
   (c) a experimental condition D and a experimental condition B.

In all the cases, comparisons performed have to take into account the contexts. That is, they are performed on CG sites, CHG sites, and CHH sites, where H = A, T, or C.

2.2.3 Limitations

Due to the fact that this research project is still ongoing, we agreed with Dr Marta Barrachina and their collaborators , inter alia, to impose special conditions for the handling of these sensitive data. Therefore, neither raw data nor processed data as well as biological results will be reported in this thesis.

2.3 Outline

This master's thesis is organized in five chapters. Chapter one discusses the background of the project and presents main ideas of the scope of the thesis. Chapter two focuses on the hypothesis formulation, states both the main objective and then specific questions to answer, and exposes the limitations of the research. In chapter three, materials are presented. Chapter four is devoted describe tools and methods as well as to present the results of the implementation of the pipeline. Finally, chapter five summarizes the contributions and proposes some recommendations for further work.

Chapter 3

Material

Due to the necessities of the researchers (section 2.1), the main goal of this project is the development of a pipeline that allows the identification of DMR between different experimental conditions (section 2.2.1). In order to achieve this objective, two specific comparisons between two experimental conditions from a experimental design, agreed with researchers, were formulated (section 2.2.2). In this chapter, we first expound the experimental conditions and design, the sample size and raw data used.

3.1 Experimental Conditions

The experimental conditions considered in the data analysis were:

1. Phenotype group (*Group*)
    (a) Group A (*A*)
    (b) Group B (*B*)
    (c) Group C (*C*)
    (d) Group D (*D*)
2. Region of interest (*Amplicon* or *Gene*)
    (a) Region associated with gene 1 (*Gene1*)
    (b) Region associated with gene 2 (*Gene2*)

Figure 3.1: Illustration of the FLX primers design for the gene of interest.

3.2 Experimental Design and Sample Size

Data analysis was based on a total of 44 samples; 8 from group A, 8 from group B, 8 from group C associated with gene1, 8 from group C associated with gene2, and 8 from group D. Table 3.1 shows the allocation of each sample to each experimental condition.

3.3 Raw Data

DNA was isolated from the mouse neocortex and all samples were bisulfite-treated. In order to avoid technical batch effects, bisulfite treatment was carried out in parallel and using the same stock. In order to identify the DNA sequence region of interest from each sample after, two primer adapters were designed, according to the guidelines for the 454 GS FLX System from Roche [Roche (2009)]. The primers used consist of four components, namely, a directional primer (Primer A or Primer B), a four-base library "key", a Multipex Identifier (MID) [454 Life Sciences Corporation (2012)] and a template specific sequence (*aka* consensus primer) (Figure 3.1). More specifically:

- Forward primer (Primer A-Key-MID-consensus primer):

  5'-CGTATCGCCTCCCTCGCGCCA-TCAG-MID-template specific sequence-3'

- Reverse primer (Primer B-Key-MID-consensus primer):

  5'-CTATGCGCCTTGCCAGCCCGC-TCAG-MID-template specific sequence-3'

454 GS FLX Titanium pyrosequencer was used for sequencing the DNA bisulfite sequences. This process generated two types of plain text files; `reads.fasta` is a fasta file containing the nucleotides sequence for each read, and `reads.qual` containing the corresponding Phred scores for each fasta sequence [Ewing and Green (1998)]. These two files were used as the basis for the data analysis.

Table 3.1: Allocation of each sample to each experimental condition.

sampleID	sample	sampleName	Lane	MID	Gene	Group
S01	sample.01	Gene1.MID.1	1	MID.1	Gene1	C
S02	sample.02	Gene1.MID.2	1	MID.2	Gene1	C
S03	sample.03	Gene1.MID.3	1	MID.3	Gene1	C
S04	sample.04	Gene1.MID.4	1	MID.4	Gene1	C
S05	sample.05	Gene1.MID.5	1	MID.5	Gene1	C
S06	sample.06	Gene1.MID.6	1	MID.6	Gene1	C
S07	sample.07	Gene1.MID.7	1	MID.7	Gene1	C
S08	sample.08	Gene1.MID.8	1	MID.8	Gene1	C
S09	sample.09	Gene1.MID.9	1	MID.9	Gene1	C
S10	sample.10	Gene1.MID.10	1	MID.10	Gene1	C
S11	sample.11	Gene1.MID.11	1	MID.11	Gene1	A
S12	sample.12	Gene1.MID.12	1	MID.12	Gene1	A
S13	sample.13	Gene1.MID.13	1	MID.13	Gene1	A
S14	sample.14	Gene1.MID.14	1	MID.14	Gene1	A
S15	sample.15	Gene1.MID.15	1	MID.15	Gene1	A
S16	sample.16	Gene1.MID.16	1	MID.16	Gene1	A
S17	sample.17	Gene1.MID.17	1	MID.17	Gene1	A
S18	sample.18	Gene1.MID.18	1	MID.18	Gene1	A
S19	sample.19	Gene1.MID.19	1	MID.19	Gene1	A
S20	sample.20	Gene1.MID.20	1	MID.20	Gene1	A
S21	sample.21	Gene2.MID.1	1	MID.1	Gene2	C
S22	sample.22	Gene2.MID.2	1	MID.2	Gene2	C
S23	sample.23	Gene2.MID.3	1	MID.3	Gene2	C
S24	sample.24	Gene2.MID.4	1	MID.4	Gene2	C
S25	sample.25	Gene2.MID.5	1	MID.5	Gene2	C
S26	sample.26	Gene2.MID.6	1	MID.6	Gene2	C
S27	sample.27	Gene2.MID.7	1	MID.7	Gene2	C
S28	sample.28	Gene2.MID.8	1	MID.8	Gene2	C
S29	sample.29	Gene2.MID.9	1	MID.9	Gene2	B
S30	sample.30	Gene2.MID.10	1	MID.10	Gene2	B
S31	sample.31	Gene2.MID.11	1	MID.11	Gene2	B
S32	sample.32	Gene2.MID.12	1	MID.12	Gene2	B
S33	sample.33	Gene2.MID.13	1	MID.13	Gene2	B
S34	sample.34	Gene2.MID.14	1	MID.14	Gene2	B
S35	sample.35	Gene2.MID.15	1	MID.15	Gene2	B
S36	sample.36	Gene2.MID.16	1	MID.16	Gene2	B
S37	sample.37	Gene2.MID.17	1	MID.17	Gene2	D
S38	sample.38	Gene2.MID.18	1	MID.18	Gene2	D
S39	sample.39	Gene2.MID.19	1	MID.19	Gene2	D
S40	sample.40	Gene2.MID.20	1	MID.20	Gene2	D
S41	sample.41	Gene2.MID.21	1	MID.21	Gene2	D
S42	sample.42	Gene2.MID.22	1	MID.22	Gene2	D
S43	sample.43	Gene2.MID.23	1	MID.23	Gene2	D
S44	sample.44	Gene2.MID.24	1	MID.24	Gene2	D

Chapter 4

Methods and Results: The Pipeline

In this chapter, the outline of the pipeline is introduced. Then, availability and requirements of the program are indicated, and finally, the flow of the execution and main results generated are described step by step, indicating which methods or software tools have been implemented or integrated.

4.1 Outline

As mentioned in section 1.2.2, there are different types of tools and configurations for finding DMR. However, based on our review, none of the pipelines proposed were well-adapted for the interests of the researchers (section 2.1). For this reason, and based on the general outline of a pipeline (Figure 1.6), the pipeline that we have developed and implemented consists of the following processes:

1. *Quality control of raw data*: quality assessment and checks on raw reads generated by the sequencer.

2. *Preprocessing*: in order to prepare raw data for the alignment process, sequence reads are filtered, splitted, and trimmed according to different criteria.

3. *Alignment process*: mapping preprocessed reads against the reference sequence and identification of methylated sites.

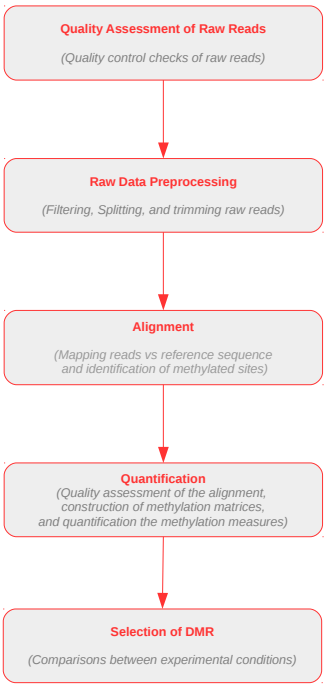

Figure 4.1: Illustration depicting the flow of the pipeline.

4. *Quantification process*: quality assessment of the alignment process, construction of methylation matrices associated with each context (i.e. CG, CHG or CHH), and quantification of the methylation measures for each sample and target region (i.e. gene).

5. *Selection of DMR*: finding the differentially methylated regions (i.e. sites) between the experimental conditions.

Figure 4.1 shows the flow of the pipeline throughout each process of execution implemented.

4.2 Availability

The pipeline, that we called `pmda`, is freely available under a License GPL-2 (`http://www.r-project.org/Licenses/GPL-2`). It can be downloaded from the GitHub repository `https://github.com/jlmosquera/pmda`.

4.3 Requirements

To use the pipeline one must have `R` 3.2 (or greater) [R Core Team (2015)] installed, as well as `Bioconductor` 3.1 [Gentleman et al. (2004)] (or greater). But also, some extra packages from CRAN (`http://CRAN.R-project.org/`) [R Core Team (2015)] are required. Specifically, `ShortRead` [Morgan et al. (2009)], `tools` [R Core Team (2015)], `reshape2` [Wickham (2007)], `parallel` [R Core Team (2015)], and `gplots` [Warnes et al. (2015)]. It is also essential to have `Perl` 5 18.2 [Team (2015)], `FastQC` [Andrews (2015)] and `BiQ Analyzer HT` [Lutsik et al. (2011)] well installed.

4.4 Execution

To run the pipeline, the user must to:

1. Set up the required parameters in file `parameter.R`.
2. Open a terminal console, and execute the file `run_pmda.R` as follows

```
$ ./run_pmda.R
```

During the execution, the program calls sequentially the `R` script sources corresponding to each process of the pipeline. Results are stored in a new path (or folder) called `./results/`. In turn, this folder contains five new folders named `qc.raw`, `preprocessing`, `alignment`, `quantification`, and `comparison`, where the files yielded by the program at each step are respectively stored.

The following subsections describe the methods implemented and the software tools involved at each step of the pipeline, and also presents the files provided by the program at the end of each process.

4.4.1 Quality Control of Raw Data

The goal of this process is assessing the quality of read sequences generated by the 454 GS FLX System from Roche. It consists of two main steps:

1. `FastQC` on raw reads.

2. Descriptive statistics of raw reads.

4.4.1.1 FastQC on Raw Reads

First step can be only executed when the user provides either files `.fasta` and `.qual` or a file `.fastq`[1]. In such case, the program runs the software `FastQC` and provides the user with different types of quality checks, which are organized in an interactive `.html` file. The information provided is listed as:

- *Basic Statistics*: generates some simple composition statistics for the file analysed.
- *Per base sequence quality*: an overview of the range of quality values across all bases at each position in the FastQ file.
- *Per sequence quality scores*: allows to see if a subset of the sequences have universally low quality values.
- *Per base sequence content*: plots out the proportion of each base position in a file for which each of the four normal DNA bases has been called.
- *Per base GC content*: shows graphically the GC content of each base position in a file.
- *Per 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.
- *Per base N content*: represents in a plot the percentage of base calls at each position for which an N was called.
- *Sequence Length Distribution*: a graph showing the distribution of fragment sizes in the file which was analysed.
- *Sequence Duplication Levels*: counts the degree of duplication for every sequence in the set and creates a plot showing the relative number of sequences with different degrees of duplication.

[1] A `.fastq` is a plain text file format containing both a sequence (usually nucleotide sequence) and its corresponding quality scores (usually Phred scores).

- *Overrepresented sequences*: lists all of the sequence which make up more than 0.1% of the total.

- *Kmer Content*: counts the enrichment of every 5-mer within the sequence library.

4.4.1.2 Descriptive Statistics on Raw Reads

Second step runs an `R` scripting code that yields the following files devoted to provide control checks that allows to assess the quality of the raw reads. This step shall be run regardless of the execution of `FastQC` , which in such a case it can be interpreted as a complementary information. Descriptive statistic results are provided in the following files:

- `descriptive.reads.csv`: summary table showing the number of reads, and the minimum, first quartile, median, mean, third quartile and maximum read length per lane.

- `distribution.reads.csv`: summary table showing the absolute frequency, cumulative absolute frequency, relative frequency and cumulative relative frequency for intervals of read lengths of 50 base pairs.

- `histogram.read.lengths.read.pdf`: histogram describing the number of reads per each read length.

- `density.overall.read.quality.pdf`: density of overall average read quality per each lane.

- `base.call.frequency.table.csv`: summary table showing the number and the percentage of each base per each lane.

- `base.call.frequency.plot.pdf`: bar plot of base call frequency over all reads per lane.

4.4.2 Raw Data Preprocessing

Raw data preprocessing is intended to prepare files for the alignment process. It consists of three main steps:

1. Filtering raw reads.

2. Splitting and trimming reads by MID primers.

3. Splitting reads by consensus primers.

4.4.2.1 Filtering Raw reads

The high-throughput sequencing technologies have been represented a revolution (section 1.1.2). However, these technologies have some limitations [Kircher and Kelso (2010) and Tucker et al. (2009)]. This step was implemented in order to filter reads containing some technical errors. Currently, there are two types of filters implemented:

1. Removing reads that are less than minimum length of nucleotides.

2. Trimming and filtering reads containing bases `N` [454 Life Sciences Corporation (2012)].

As a result of applying these filters, the program yields two files named `reads.filtered.fastq` and `reads.filtered.fasta`. These files contain the the sequences in `fastq` and `fasta` text format respectively. Moreover, the program also provides a brief summary describing the number and lengths of the reads filtered stored in a file called `summary.filtering.raw.reads.csv`.

Table 4.1 shows the summary table describing the remaining number and lengths of the reads after performing each filter during the execution of the data analyzed.

Table 4.1: Summary table describing the number an length of the reads after performing the filtering step.

Filtering.Criteria	N	Min	Q1	Median	Mean	Q3	Max
(none)	676498	41	415	458	414	468	1338
reads with length < 150	636641	151	417	462	433,4	468	1338
remove and trim reads with N's	613425	151	417	463	434,5	468	967

In file `parameters.R`, a user can also order to run a quality control of filtered reads performed by using `FastQC`.

4.4.2.2 Splitting and Trimming Reads by MID Primers

Primers are useful for the task of identifying amplicons and samples. However, removal of primers from reads improves the quality of results from downstream analyses [Criscuolo and

Brisse (2013) an Laird (2010)]. This step of the pipeline is devoted to split reads by MID primers and then trimming such adapters from the reads.

Leaving aside the input parameters provided in file `parameters.R`, in order to prepare the reads according to each sample and amplicon, this step also requires that the user provides some extra files associated with information about the lanes, samples, and primers. These files must be placed in the folder `./data/` and, specifically, they contain the following information (Appendix A):

- `lanes.csv`: a tabulated text file with two columns indicating for each lane:
  - *Lane.ID*: number of lane.
  - *FileName*: name of the `.fasta` file storing the read sequences associated with *Lane.ID*.
- *primers.csv*: a tabulated text file provinding information about the consensus primer, specifically:
  - *Primer.ID*: number of consensus primer.
  - *Ampl.Name*: name of the amplicon (or gene of interest).
  - *Sense*: Sense of the sequence. *fw* indicates a forward sequence and *rv* indicates a reverse sequence.
  - *Primer.Seq*: nucleotide sequence associated with *Primer.ID*, *Ampl.Name*, and *Sense*.
  - *Start*: starting position of the consensus primer in the reference sequence.
  - *Num*: number of nucleotides o the primer.
- *mids.csv*: a tabulated text file indicating for each MID primer the following information:
  - *MID.ID*: number of MID.
  - *MID.Seq*: nucleotide sequence associated with *MID.ID*.
- *samples.csv*: a tabulated text file relating samples, primers, amplicons and reference sequences
  - *Sample.ID*: sample identification.

- *Ampl.Name*: name of the amplicon provided in file *primers.csv*.
- *Patient.ID*: sample name.
- *Lane.ID*: number of lane provided in file *lanes.csv*.
- *Primer.ID*: number of consensus primer provided in file *primers.csv*.
- *RefSeq.ID*: number of reference sequence provided in file *RefSeqs.csv*.

After performing this step, the program generates a large list of `fastq` and `fasta` text format files. Each couple of files are associated with a unique MID, they are named, respectively, as `MID.nn.fastq` and `MID.nn.fasta` where `nn` indicates the number of MID primer, being 0 those reads that have not been identified. These list of files are stored in a new folder called `./data/splitted/mids/`.

The program also yields a bar plot showing the number of reads splitted by MID primer, which is stored in file `reads.by.MID.pdf`. Figure 4.2-A shows an example of this mentioned plot. In dark gray color is highlighted the number of reads non-identified by MID primerand in red color the MID with the lowest number of reads (less than 1% of the total number of reads).

4.4.2.3 Splitting Reads by Consensus Primers

Third step of raw data preprocessing splits reads by the consensus primers. This allows to identify the reads by amplicon. However, consensus primers are not removed from the reads. `BiQ Analyzer HT` is the software tool that has been integrated in the pipeline for the alignment process [Lutsik et al. (2011)]. The developers of this tool mention in the documentation of the software that trimming primers is not required because the algorithms implemented in the program already takes this fact into account. Hence, after splitting reads by consensus primers, adapters are not removed from the read sequences.

As a result, this step yields also a large list of `fastq` and `fasta` text format files such that, for each pair of files `MID.nn.fastq` and `MID.nn.fasta` generates two files for (i.e. four new files) storing the reads already splitted. These files are named `xxx_MID.nn.fw.fastq` and `xxx_MID.nn.rv.fastq`, `xxx_MID.nn.fw.fasta` and `xxx_MID.nn.rv.fasta` where `xxx` indicates the amplicon, textttnn the number of MID primer and `fw` or `rv` if the file contains forward or reverse read sequences respectively. These large list of files are stored in a new folder called

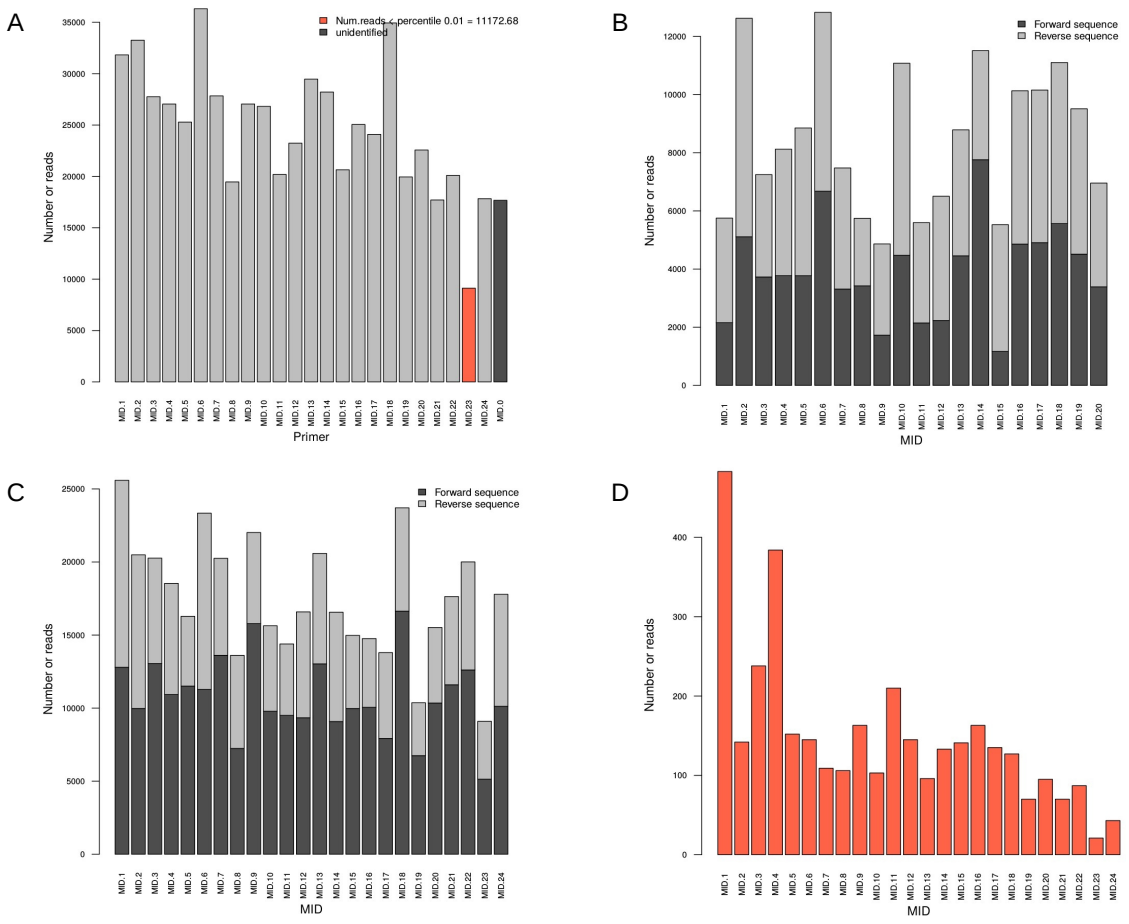

Figure 4.2: Bar plots associated with the steps of splitting and trimming during the execution of the raw data preprocessing. Bar plot shows the number of reads by MID primer. Bar plots B and C shows the number of reads by MID primer and amplicon (e.g gene1 and gene2). Bar plot D shows the number of by MID primer that have not been splitted by amplicon.

`./data/splitted/consensus/`. Moreover, due to the fact that `BiQ Analyzer HT` requires for each sample and target DNA sequence a `.fasta` file with the appropriate read sequences, the pipeline concatenates the sequences from each pair of files `xxx_MID.nn.fw.fasta` and `xxx_MID.nn.rv.fasta` and yield a single `.fasta` file for each amplicon and MID, named `xxx.MID.nn.fasta`. These new files are stores in new folder named `./data/alignment/xxx/`.

In order to summarize the results of this process, the program yields a `.pdf` file called `reads.by.amplicon.and.MID.pdf` that contains a bar plot for each amplicon profiled plus a bar plot with the number of unidentified reads per MID primer (plots B–D in Figure 4.2), and a tabulated text file with the number of reads per amplicon, MID and sense (Table 4.2)

In file `parameters.R`, a user can also order to run a quality control of splitted reads which is

Table 4.2: Summary table describing the number of reads identified by amplicon, MID and sense of the sequence.

Ampl.Name	MID.ID	Sense	Num.Reads
gene1	MID.1	fw	2156
gene1	MID.1	rv	3597
gene1	MID.2	fw	5110
gene1	MID.2	rv	7513
...	...	...	...
gene1	MID.20	fw	3389
gene1	MID.20	rv	3569
gene2	MID.1	fw	12795
gene2	MID.1	rv	12796
gene2	MID.2	fw	9976
gene2	MID.2	rv	10515
...	...	...	...
gene2	MID.23	rv	3961
gene2	MID.24	fw	10124
gene2	MID.24	rv	7665
unidentified	MID.1	NA	483
unidentified	MID.2	NA	142
...	...	...	...
unidentified	MID.24	NA	43

performed again by `FastQC`.

4.4.3 Alignment Process

Alignment process is devoted to map read sequences treated with bisulfite from each sample onto a target genomic reference sequence. There are different method and tools for mapping bisulfite-treated reads to a reference sequence (*aka* methylation aligners) [Henry et al. (2014), Krueger and Andrews (2011), Xi and Li (2009), and Chen et al. (2010)]. In this pipeline, we have selected the `BiQ Analyzer HT` from the Max Plank Institut Informatik (`http://biq-analyzer-ht.bioinf.mpi-inf.mpg.de/`) [Lutsik et al. (2011)].

`BiQ Analyzer HT` takes as input files `xxx.MID.nn.fasta` stored in path `./data/alignment/xxx/` (paragrah 4.4.2.3) and a reference DNA genomic sequence provided in `fasta` text format stored in a file at the folder `./data/`. Then, the software maps, one by one, the read sequences in each file to the reference sequence. As a result, `BiQ Analyzer HT` outputs the DNA methylation information in different files. Briefly, main files reported for each

`.fasta` file aligned are (Appendix A):

- `summary.dat`: a short summary of the analysis run.
- `alignment.mfa`: `.fasta` file containing multiple alignment of the bisulfite reads to the genomic reference sequence.
- `heatmap.png`: a methylation heatmap, containing methylation patterns per read. Methylation heatmap represents the extracted methylation patterns of the bisulfite reads graphically. Columns of the heatmap are formed by methylation sites found in the reference sequence by matching the analyzed methylation context (i.e. CG, CHG or CHH), while rows correspond to read sequences.
- `pearlNecklace.png`: a pearl–necklace diagram, summarizing methylation information for each CpG. Pearl-Necklace diagram summarizes methylation information for the whole set of filtered reads, by identified methylation sites. For each site the diagram has a colored rectangle plotting a distribution of the methylated, unmethylated and unrecognized states of this site in the given set of bisulfite reads. In other words, the diagram gives a "mean" methylation profile of the read population.
- `results.tsv` is a tabulated text file containing methylation information for each analyzed read.

All these files are generated for each context, amplicon and MID and they are stored in a new paths named `./results/alignment/cpg/xxx/MID.nn/` where `cpg` indicates the context, `xxx` the amplicon (or target region) and `nn` the number of MID.

4.4.4 Quantification Process

In order to perform a downstream analyses such as the selection of DMR between two experimental conditions, it is necessary to quantify the number of reads (un)methylated per each site and sample in order to compute and compare statistical measures.

Mentioned in section 4.4.3, alignment process performed with `BiQ Analyzer HT` , generates a large list of files. Among the files provided for just one single mapping, the most important

file is `results.tsv`. This file consists of a enormous table describing the methylation information for each read, specifically, it contains the following data (Appendix B):

- *ID*: the identification of the read aligned.
- *Alignment score*: "matching score" used in the definition of the sequence identity.
- *Sequence identity*: percentage of identity for the read aligned.
- *Methylation pattern*: pattern of all the sites identified for the read (0:unmethyled, 1:methyled, x:missing site).
- *Mean methylation level*: percentage of de number of methylated sites.
- *Missing sites*: number of not present sites after alignment.
- *Conversion rate*: percentage of the bisulphited conversion rate of the read.
- *Reference sequence*: name of the reference sequence.

This file is generated for each context (i.e. CG, CHG or CHH), region of interest (i.e. gene or amplicon), and MID.

Quantification is the largest the process implemented in the pipeline. It consists of four main steps:

1. Pile up tables from all "results.tsv" files into one single table associated with a unique context and amplicon.
2. Quality assessment of the aligment process.
3. Quantification.
4. Heatmap plot of quantification measures (i.e. beta values).

4.4.4.1 Pile Up Methylation Alignment Results

In order to asses the quality of the alignments and then quantify the methylation of each site and sample, the pipeline executes a functions that takes all the files called "results.tsv" from all the MID in each amplicon and context, and concatenates all the information into a one single table. This table identifies the origin of each read binding a column that indicate the MID (i.e. the patient ID). This concatenation is done for each context and amplicon, and the resulting files are stored in a new folder called `./results/quantification/`. These files are named `cpg.xxx.csv` where `cpg` indeicates the context (i.e. CG, CHG or CHH) and `xxx` the name of the amplicon.

4.4.4.2 Quality Assessment of Alignment Process

Quality assessment of alignment is based on files containing the methylation information piled up from the original files `results.tsv` generated by `BiQ Analyzer HT` (section 4.4.4.1). This step yields three files containing descriptive plots for the corresponding measures reported in files `cpg.xxx.csv`. Specifically,

- `cpg.Aligment.Score.tiff`: density and box plots of alignment score per each amplicon in a specific context.
- `cpg.Mean.Methylation.tiff`: density and box plots of the % of (detected) methylation sites per each amplicon in a specific context.
- `cpg.Missing.Sites.tiff`:density and box plots of missing methylation sites per each amplicon in a specific context.

Figures 4.3, 4.4, and 4.5 show and example of the densities plots and box plots associated with the alignment scores, the mean methylation and the missing sites measures provided by files `cpg.xxx.csv`, respectively. In the three cases box plots show two horizontal lines. Red color line show the mean of the measure, while orange line shows the media.

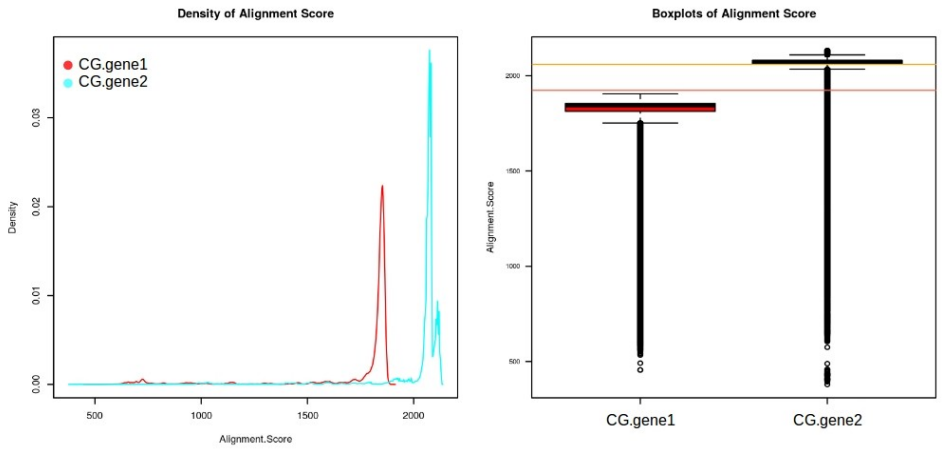

Figure 4.3: Density plot (left side)and box plots (right side) of alignment score per each amplicon in a specific context. Red color line show the mean of the measure, while orange line shows the media.

4.4.4.3 Quantification of Methylation Levels

For performing comparisons between the different experimental conditions it is necessary to summarize the methylation status of each site, in each context and amplicon. This process is called quantification. Probably the most widely use method for quantifying methylation levels is the β-value [Du et al. (2010)]. The β-value is the ratio of methylated reads per site and the overall the sum of methylated and unmethylated reads per site, that is

$$\beta_{i,j} = \frac{M}{M+U}$$

where M is the number of methylated reads in site i and MID j, and U is the number of unmethylated reads in site i and MID j. The $beta$-value statistic results in a number between 0 and 1, such that under ideal conditions, a value of 0 indicates that all copies of the CpG site in the sample were completely unmethylated and a value of 1 indicates that every copy of the site was methylated. The pipeline computes β-values and provides M-values that can be interpreted as a "inverse" of β-values [Du et al. (2010)].

After performing this step, the program yields as a result for each context and amplicon the following files:

- `cpg.xxx.num.Meth.Sites.csv`: number of (un)methylated sites per MID for each con-

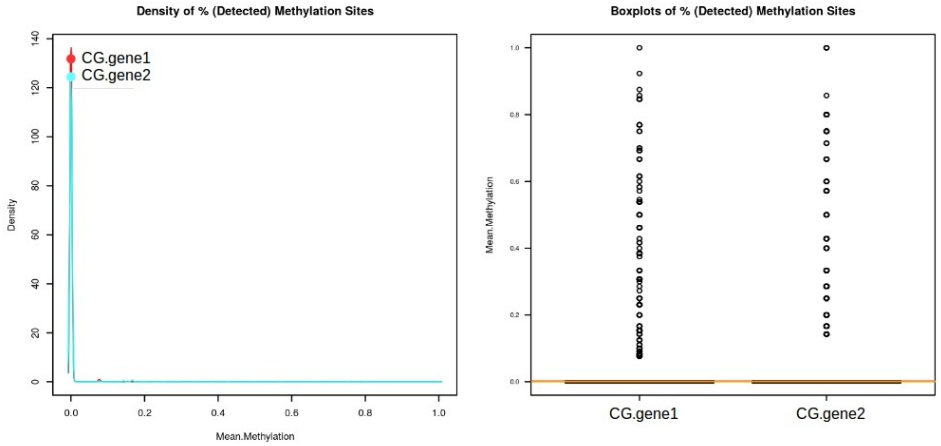

Figure 4.4: Density plot (left side)and box plots (right side) of the % of (detected) methylation sites per each amplicon in a specific context. Red color line show the mean of the measure, while orange line shows the media.

text (`cpg`) and amplicon (`xxx`).

- `cpg.xxx.beta.values.csv`: β-values of sites per MIDs for each context (`cpg`) and amplicon (`xxx`).
- `cpg.xxx.M.values.csv`: M-values of sites per MIDs for each context (`cpg`) and amplicon (`xxx`).
- `cpg.xxx.Meth.Pattern.csv`: methylation pattern for each read by context (`cpg`) and amplicon (`xxx`).
- `cpg.xxx.mid.freq.csv`: frequencies for each context and MID excluding missing sites by context (`cpg`) and amplicon (`xxx`).
- `cpg.xxx.Met.summary`: summary of frequencies for each site in each context

4.4.4.4 Heatmap of Methylation Levels

Last step of quantification process is a devoted to plot a heatmap for each context and amplicon, `cpg.xxx.heatmap.pdf` (Figure 4.6). This figure facilitates the review of comparisons associated with an specific situation. It reports a z-score of β-values for each site and MID in a specific context and amplicon.

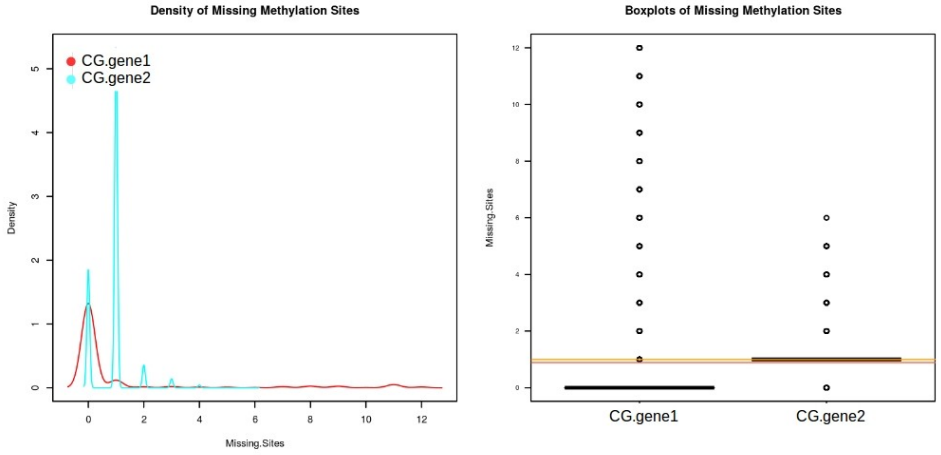

Figure 4.5: Density plot (left side)and box plots (right side) of missing methylation sites per each amplicon in a specific context. Red color line show the mean of the measure, while orange line shows the media.

4.4.5 Selection of Differentially Methylated Regions (DMR)

Finally, last step is the process devoted to find the differentially methylated regions (DMR) between each pair of experimental conditions. This analysis is based on the Fisher's Exact Test [Fisher (1922) and Agresti (2002)]. Briefly, for each site the program tests if (un)methylation and phenotypes are independent, or equivalently if the odds ratio (OR) is equal to 1. Associated P–values are adjusted for correcting the problem of multiple testing by the method of Benjamini and Hochberg [Benjamini and Hochberg (1995)].

This process requires that the user provides an extra file associated with the comparison to be performed. This file must be placed in the folder `./data/` and it must to contain the following information (Appendix A):

- *Gene*: number of amplicon.
- *Comp.Label*: label to be shown in resulting files for identifying the comparison.
- <*labels*>: labels associated associated with each experimental condition.

After performing this step, the program yields the following two files for each context and amplicon:

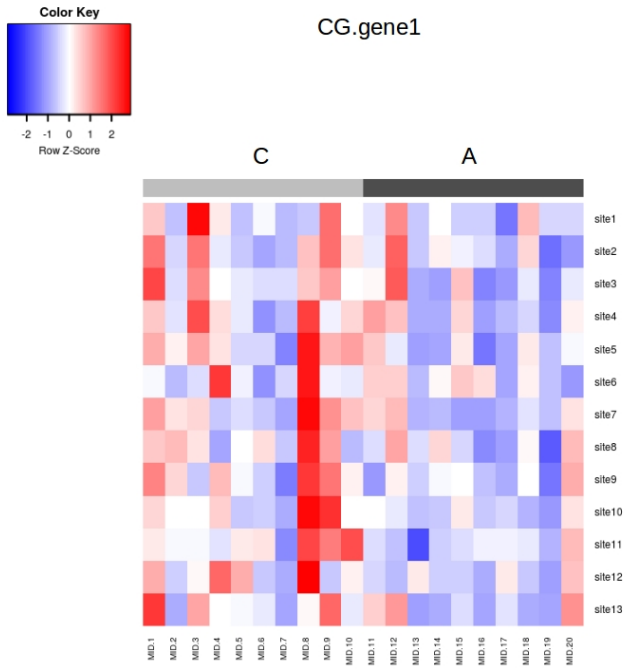

Figure 4.6: Heatmap plot of beta values associated with experimental conditions A and C in the context of sites CG for the gene1.

- `cpg.xxx.Comparisons.csv`: table with the OR, the confidence interval (CI) for the OR, the raw p–value and the adjusted pvalue for each condition and site in a specific context and amplicon.

- `cpg.xxx.plotOR_x_vsy.pdf`: plot of the OR, whit their CI for all the sites of a specific context in an amplicon and comparison `x_vs_y`.

Moreover, in order to summarize help the user with the summarization of the results, the pipeline yields a file called `numDMR.csv`, where for each context, gene and comparison indicates the number of DMR according to show a False Discovery Rate (FDR) lower than 0.01 and 0.05.

Table 4.3 shows the results provided by the pipeline for the comparison A *vs* B in the context of site CG for the gene1.

Table 4.4 shows the results the number of DMR for the data analysis performed.

Figure 4.7 shows al

Table 4.3: Summary table describing the results of comparison in CG for gene1 betwen group A and group C.

GG.Site	C_vs_A.OR	C_vs_A.Lower.OR	C_vs_A.Upper.OR	C_vs_A.Pvalue	C_vs_A.Adj.Pvalue
site1	1.2206178997	0.8940349751	1.6734518705	0.1971028932	0.3844981015
site2	1.1883892672	0.9077696805	1.5601554391	0.2087895516	0.3844981015
site3	1.39272528	1.0615775769	1.8350945525	0.0146559878	0.1905278413
site4	1.369766737	0.9976846646	1.891587833	0.0461108279	0.1998135875
site5	1.5336674469	1.0349860495	2.2991117058	0.0310774655	0.1998135875
site6	0.8956542199	0.6223124368	1.2893874015	0.5350219357	0.5600505906
site7	1.3101379355	0.9488629274	1.8188091895	0.099794099	0.2594646575
site8	1.116276708	0.7940933758	1.5749295255	0.5600505906	0.5600505906
site9	1.4036077779	0.9489051583	2.0962737572	0.0905287885	0.2594646575
site10	1.2741010287	0.8169658873	2.0067883854	0.2859961928	0.4131056118
site11	1.2307982827	0.8180214014	1.8659252142	0.3237930795	0.4209310033
site12	1.2526392212	0.8598729488	1.8370354238	0.2366142163	0.3844981015
site13	1.109583105	0.7844931373	1.5752392132	0.5541254153	0.5600505906

Table 4.4: Summary table of the number of DMR found for each context, gene and comparison, and according to a FDR<0.01 or a FDR<0.05

Context	Gene	Comparison	Num.Sites	Adj.Pvalue < 0.01	Adj.Pvalue < 0.05
CG	gene1	C_vs_A	13	0	0
CG	gene2	C_vs_B	7	4	4
CG	gene2	C_vs_D	7	3	6
CG	gene2	B_vs_D	7	1	1
CHG	gene1	C_vs_A	9	0	0
CHG	gene2	C_vs_B	8	0	0
CHG	gene2	C_vs_D	8	0	0
CHG	gene2	B_vs_D	8	0	1
CHH	gene1	C_vs_A	72	0	0
CHH	gene2	C_vs_B	14	1	1
CHH	gene2	C_vs_D	14	1	2
CHH	gene2	B_vs_D	14	1	2

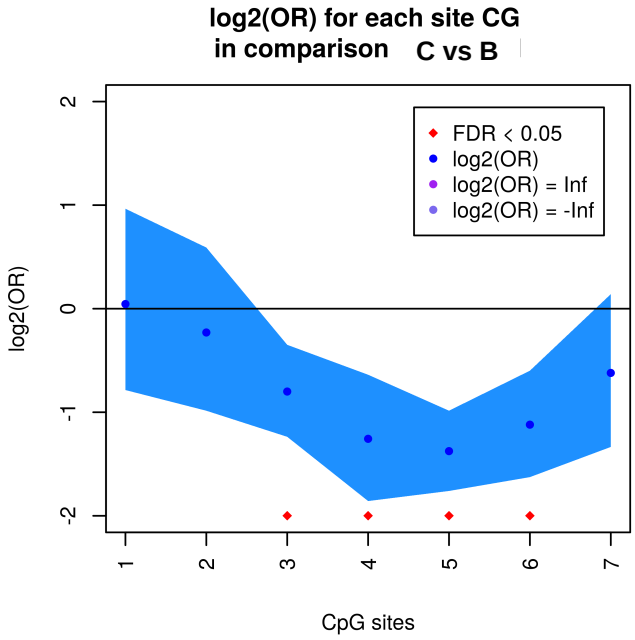

Figure 4.7: OR plot describing the $log_2(OR)$ values for each CG site when comparing group C *vs* group B in gene2. Red dots indicate DMR according to a FDR lower than 0.05. Light blue shadows is the 95% confidence band.

Chapter 5

Conclusions and Recommendations for Further Work

5.1 Conclusions

This master's thesis which focused on the "Methylation Data Analysis Associated with Alzheimer's Disease" is an attempt help researchers at the lab of Dr Marta Barrachina providing them a tool for finding differentially methylated regions based on DNA bisulfite sequencing of target regions by using the 454 GS FLX system from Roche.

The project has focused on main aspect:

1. The development and implementation of a pipeline for finding Differentially Methylated Regions (DMR) between different experimental conditions of a specific region of a DNA sequence, whose samples have been treated with bisulfite and sequenced with a 454 GS FLX system.

With regard to this issue, it has have been taken advantage of a experimental design and specific objectives in which researchers were interested. This fact has allowed the implementation of a pipeline that

- It is completely coded in `R` language.
- Performs a whole data analysis for finding DMR.

- Integrates properly external tools (`Perl` script, `FastQC` and `BiQ Analyzer HT`) for taking advantage of their results, and based on them performing new computations.
- Provides the user appropriate measures, summaries and plots at each step of the pipeline.
- The pipeline is called `pmd` and it is freely available at the GitHub repository https://github.com/jlmosquera/pmda under a License GPL-2 (http://www.r-project.org/Licenses/GPL-2).

5.2 Recommendations for Further Work

In order to facilitate the data analysis for potential user, it would be interesting to build an `R` package with all the functions and sources developed as well as to develop an Application Programming Interface (API) for those user less familiarized with this kind of data analysis. Current pipeline only is thought to perform simple comparisons, however, it is possible that more complex experimental designs require the use of considering linear regression. Thus integrating approches such as the `limma` package from `Bioconductor` [Gentleman et al. (2004)], the user count take into account other interactions between different effects.

Appendix A

Input Files Required by the Pipeline

A.1 `lanes.csv`

Table A.1 shows an example of the information provided in file `lanes.csv` to the pipeline.

Table A.1: `lanes.csv`.

Lane.ID	FileName
1	reads.fasta

A.2 `mids.csv`

Table A.2 shows an example of the information provided in file `mids.csv` to the pipeline.

Table A.2: `mids.csv`.

MID.ID	MID.Seq
1	ATAGAGTACT
2	CACGCTACGT
3	CAGTAGACGT
4	CGACGTGACT
5	TACACACACT
6	TACACGTGAT
7	TACAGATCGT
8	TACGCTGTCT

A.3 primers.csv

Table A.3 shows an example of the information provided in file primers.csv to the pipeline. In order to preserve the data analyzed, nucleotide sequences provided in this table are not the ones used for the data analysis performed. This table shows information associated with two hypothetical amplicons analyzed.

Table A.3: mids.csv.

Primer.ID	Ampl.Name	Sense	Primer.Seq	Start	Num
1	gene1	fw	ATAGTAAATGTTTTTATTTATTCTTGTATTG	1	30
1	gene1	rv	GTTTATAAATCCAATACTCATCCTAATCAT	370	30
2	gene2	fw	TGTAAGTTTGATTGTATTTATGGGGGATTA	1	30
2	gene2	rv	ACTACAAAATTCAATCATTAAATA	475	25

A.4 samples.csv

Table A.4 shows an example of the information provided in file samples.csv to the pipeline. Data provided here are not the ones used for the data analysis performed. This table shows information associated with two hypothetical amplicons profiled in tha same lane.

Table A.4: mids.csv.

Sample.ID	Ampl.Name	Patient.ID	Lane.ID	MID.ID	Primer.ID	RefSeq.ID
s01	gene1	gene1.01	1	1	1	1
s02	gene1	gene1.02	1	2	1	1
s04	gene1	gene1.04	1	4	1	1
s05	gene1	gene1.05	1	5	1	1
s06	gene1	gene1.06	1	6	1	1
s07	gene1	gene1.07	1	7	1	1
s08	gene1	gene1.08	1	8	1	1
s09	gene2	gene2.09	1	1	2	2
s10	gene2	gene2.10	1	2	2	2
s11	gene2	gene2.11	1	3	2	2
s12	gene2	gene2.12	1	4	2	2
s13	gene2	gene2.13	1	5	2	2
s14	gene2	gene2.14	1	6	2	2

A.5 comparisons.csv

Table A.5 shows an example of the information provided in file `comparisons.csv` to the pipeline. Actually, it is the contrasts matrix, where a user must take into account that values can be 1, 0 or -1. Currently, the pipeline only supports simple contrasts. Thus, a cell with a value 1 indicates that the the experimental condition is the first in the comparison and a value -1 is intended for the second experimental condition. A value 0 indicates that a label is not included in the comparison.

Table A.5: `comparisons.csv`.

Gene	Comp.Label	C	A	B	D
gene1	A_vs_C	-1	1	0	0
gene2	B_vs_C	-1	0	1	0
gene2	D_vs_C	-1	0	0	1
gene2	D_vs_B	0	0	-1	1

Appendix B

Output Files Provided by BiQ Analyzer HT

B.1 heatmap.png

Figure B.1 shows the type of methylation heatmap that BiQ Analyzer HT provides in filee heatmap.png.

B.2 perlNecklace.png

Figure B.2 shows a pearl–neclace diagram yielded by BiQ Analyzer HT.

B.3 results.tsv

Table B.1 shows an example of the methylation information privide by BiQ Analyzer HT , for each read sequence mapped to a reference genomic sequence of "Gene1".

Table B.1: results.tsv.

ID	Alignment score	Sequence identity	Methylation pattern	Mean methylation level	Missing sites	Conversion rate	Reference sequence
read.1	1444	0.7826	xxx1111000111	0.7	3	0.5349	Gene1
read.2	1868	0.9925	0000001111111	0.5385	0	0.5116	Gene1
read.3	1868	0.9926	0101100011110	0.5385	0	0.6124	Gene1
read.4	1861	0.9928	1000111110000	0.4615	0	0.7984	Gene1
...	...	...	...	...	...	...	...
read.3000	1859	0.9957	0000000000000	0	0	0.9845	Gene1

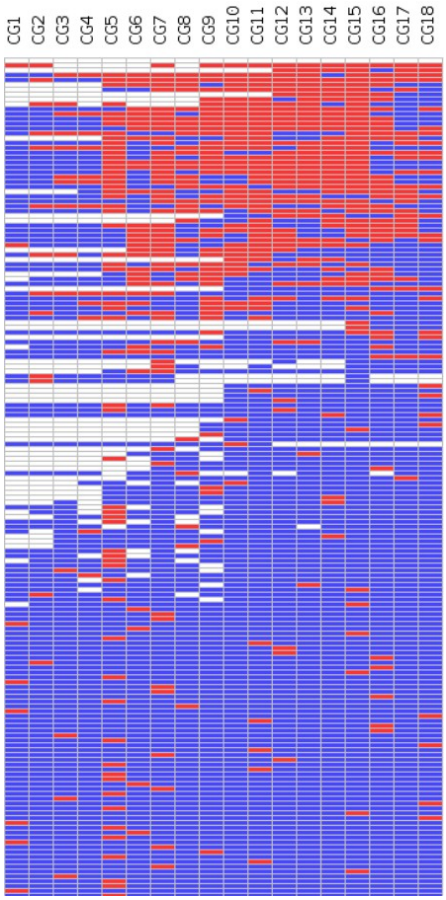

Figure B.1: Illustration of a piece of a methylation heatmap provided by `BiQ Analyzer HT` showing methylation patterns per each read and site.

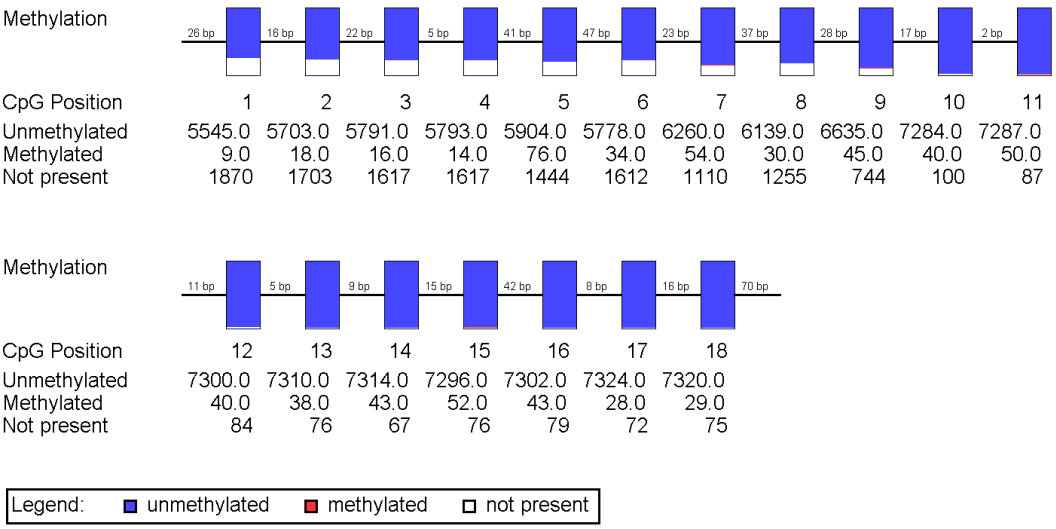

Figure B.2: Illustration of a perl–necklace diagram generated by `BiQ Analyzer`.

Appendix C

File `parameters.R`

The following `R` scripting code are the parameters required by the pipeline. That is, a potential user of the of the program must select and set up the appropriate parameters for his/her analysis, before running the `R` script `run_pmda.R`.

```
#####################################################
## chunk 1: R Output Options
#####################################################

Sys.setlocale('LC_ALL','C')          # To avoid warnings due to language
options(width = 170)

mc.cores <- 3                        # Number of cores

#####################################################
## chunk 1: Project Identification
#####################################################

.project    <- "<project_name>"                      # Project name
.title      <- "<title_of_the_project>"              # Project title
.abstract   <- "<brief_description_of_the_project>"  # Project description
```

```
.researcher <- "<name_of_the_researcher>"            # Researcher name
.lab        <- "<name_of_the_lab-_organization>"     # Lab (Organization)
.email      <- "<contact_@_e-mail>"                  # Contact e-mail
.analysts   <- "<name_of_the_data_analyst>"          # Data Analyst


##################################################
## chunk 2: Paths
##################################################

## Main path

.main.path <- "<path_to_the_project>"    # e.g. "/home/Documents/myproject"

## Basic structure

.wd <- file.path(.main.path, .project)  # Working path
setwd(.wd)

.dat <- file.path(.wd, "data")           # Data
.fun <- file.path(.wd, "functions")      # R functions
.res <- file.path(.wd, "results")        # Results

## Paths to software tools

.r      <- file.path(.fun, "R")          # path name to R functions
.perl   <- file.path(.fun, "perl")       # path name to perl scripts
.fastqc <- "<path_to_FastQC>"            # e.g. "/home/Software/FastQC"
.biq    <- "<path_to_BiQAnalyzerHT"      # e.g. "/home/Software/'BiQAnalyzerHT"
```

```
#####################################################
## chunk 4: Quality Control of Raw Data
#####################################################
##
## Parameters associated with file: "process01.qc_raw_reads.R"
##

## 1. FastQC of raw reads

run.FastQC.raw    <- TRUE          # TRUE runs the software FastQC on raw reads
convert.to.fastq <- TRUE           # TRUE if does not exist a .fastq file, but
                                   # fasta.fn and qual.fn exist.
                                   # FALSE is any other case
.run <- file.path(.dat, "run")     # path name to .fasta and .qual files
fasta.fn <- "reads.fasta"          # file name of the file with fasta sequences.
                                   # NULL, otherwise
qual.fn  <- "reads.qual"           # file name of the file with qual sequences.
                                   # NULL, otherwise

fastq.fn <- "reads.fastq"          # file name of the file with fastq sequences.
                                   # NULL, otherwise

## 2. Load Data Files

lanes.fn <- "lanes.csv"            # file name of the file with lanes

save.reads <- TRUE                 # TRUE saves in a file called 'reads.Rda' an
                                   # object with raw reads, and in case of having a
                                   # fastq file, it also save an object with fastq
```

```
                                # sequences. FALSE loads the object(s) mentioned
                                # above

## 3. QC of raw reads

run.QC.raw <- TRUE              # TRUE builds statistics summaries and plots


#####################################################
## chunk 5: Pre-processing: Filtering Raw Reads
#####################################################
##
## Parameters associated with file:
##                              'process02.preprocessing_1.filetering_raw_reads.R'
##

## 1. Filtering raw reads

filtering <- TRUE               # TRUE applies different filtering criteria to
                                # raw reads
min.len <- 150                  # minimum read length accepted for data analysis

## 2. FastQC of filtered reads

run.FastQC.filtered <- TRUE     # TRUE runs software FastQC on filtered reads


#####################################################
## chunk 6: Pre-processing: Splitting and Trimming
#####################################################
```

```
##
## Parameters associated with file:
##                        'process02-preprocessing_2.splitting_and_trimming_reads.R'
##

## 1. Load Data Files

samples.fn <- "samples.csv"      # file name of the file with sample targets info
primers.fn <- "primers.csv"      # file name of the file with consensus primer
                                 # sequences
mids.fn    <- "mids.csv"         # file name of the file with MID primer sequences

## 2. Splitting and trimming reads by MID primers

run.split.MIDs <- TRUE           # TRUE runs steps splitting and trimming reads by
                                 # MID primers. FALSE loads reads already splitted
                                 # and trimmed

max.start.pos    <- 5            # Maximum starting position of primers
max.mm.mid       <- 0            # Maximum number of mismatches in each MID primer
with.indels.mid <- FALSE         # Indels are not allowed in MID primer

## 3. Splitting reads by consensus primers

run.split.cons <- TRUE           # TRUE runs step splitting by consensus primers

max.mm.cons      <- 3            # Maximum number of mismatches in each consensus
                                 # primer
with.indels.cons <- TRUE         # Indels are not allowed in consensus primers
```

```
## 4. FastQC of reads splitted by consensus primers

run.FastQC.consensus <- TRUE     # TRUE runs the software FastQC on reads splitted



#####################################################
## chunk 7: Aligment process
#####################################################
##
## Parameters associated with file: 'process03_alignment.R'
##

cpg <- c("CG", "CHG", "CHH")     # Contexts where methylation sites in the
                                 # reference sequence must be found



#####################################################
## chunk 8: Quantification
#####################################################
##
## Parameters associated with file: 'process04_quantification.R'
##

## 1. Load Data Files

targets.fn <- "targets.txt"      # file name of the file with targets info of
                                 # each sample
pile.up <- TRUE                  # TRUE pile up all 'result.tsv' generated by
                                 # alignment. FALSE loads concatenated tables
                                 # resulting from the pile up step
```

```
## Plot heatmaps with beta values

target.samples.order <- NULL     # Numeric vector indicating the order in which
                                 # samples should be shown NULL, which is the
                                 # parametrization by default, takes the order of
                                 # samples in targets.fn file

experimental.condition <- "Group"


#####################################################
## chunk 10: Comparisons
#####################################################
##
## Parameters associated with file: 'process05_comparissons.R'
##

comparisons.fn <- "comparisons.txt" # file name of the file with the comparisons
```

Bibliography

454 Life Sciences Corporation (2012). *GS Reference Mapper*.

Agrawal, A., Murphy, R. F., and Agrawal, D. K. (2007). DNA methylation in breast and colorectal cancers. *Modern Pathology*, 20:711–721.

Agresti, A. (2002). *Categorical Data Analysis*. John Wiley and Sons, 2nd edition edition.

Andrews, S. (2015). *FastQC: A Quality Control Tool for High Throughput Sequence Data*.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society B*, 57(1):289–300.

Bennett, D. A., Yu, L., Yang, J., Srivastava, G. P., Aubin, C., and De Jagerb, P. L. (2015). Epigenomics of Alzeimer's Disease. *Translational Research*, 165(1):200–220.

Bertram, L. and Tanzi, R. E. (2011). *Genetics of Alzheimer's Disease*, pages 51–61. Wiley-Blackwell, 2nd edition edition.

Bird, A. (2007). Perceptions of Epigenetics. *Nature*, 447:396–398.

Bonetta, L. (2008). Epigennomics: The New Tool in Studying Complex Diseases. *Nature Education*, 1(1):178.

Brown, T. A. (2002). *Genomes*. Oxford: Wiley-Liss, 2nd edition.

Chatterjee, A., Stockwell, P. A., Rodger, E. J., and Morison, I. M. (2012). Comparison of alignment software for genome-wide bisulphite sequence data. *Nucleic Acids Research*, pages 1–8.

Chen, P.-Y., Cokus, S. J., and Pellegrini, M. (2010). Bs seeker: precise mapping for bisulfite sequencing. *BMC Bioinformatics*, 11(203):1–6.

Cokus, S. J., Feng, S., Zhang, X., Chen, Z., Merriman, B., Haudenschild, C. D., Pradhan, S., Nelson, S. F., Pellegrini, M., and Jacobsen, S. E. (2008). Shotgun bisulphite sequencing of the arabidopsis genome reveals dna methylation patterning. *Nature*, 452:215–219.

Colnot, S., Niwa-Kawakita, M., Hamard, G., Godard, C., Plenier, S. L., Houbron, C., Romagnolo, B., Berrebi, D., Giovannini, M., and Perret, C. (2004). Colorectal cancers in a new mouse model of familial adenomatous polyposis: influence of genetic and environmental modifiers. *Laboratory Investigation*, 84:1619–1630.

Crick, F. (1970). Central Dogma of Molecular Biology. *Nature*, 227(5258):561–563.

Criscuolo, A. and Brisse, S. (2013). AlienTrimmer: A tool to quickly and accurately trim off multiple short contaminant sequences from high-throughput sequencing reads. *Genomics*, 102(5–6):500–506.

Dodge, J. E., Ramsahoye, B. H., Wo, Z., Okano, M., and Li, E. (2002). De novo methylation of mmlv provirus in embryonic stem cells: Cpg versus non-cpg methylation. *Gene*, 289(1–2):41–48.

Du, p., Zhang, X., Huang, C.-C., Jafari, N., Kibbe, W. A., Hou, L., and Lin, S. M. (2010). Comparison of Beta-value and M–value methods for quantifying methylation levels by microarray analysis. *BMC Bioinformatics*, 11(587):1–9.

Eccleston, A., DeWitt, N., Gunter, C., Marte, B., and Nath, D. (2007). Epigenetics. *Nature*, 447:395.

Ewing, B. and Green, P. (1998). Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Research*, 8(3):186–194.

Feinberg, A. P. and Tycko, B. (2004). The history of Cancer Epigenetics. *Nature Reviews Cancer*, 4:143–153.

Feng, Y., Jankovic, J., and Wu, Y.-C. (2015). Epigenetic Mechanisms in Parkinson's Disease. *Journal of Neurological Sciences*, 349(1–2):3–9.

Fisher, R. A. (1922). On the interpretation of χ^2 from contingency tables, and the calculation of p. *Journal of the Royal Statistical Society*, 85(1):87–94.

Gentleman, R., Carey, V., Bates, D., and others (2004). Bioconductor: Open software development for computational biology and bioinformatics. *Genome Biology*, 5:R80.

Goldberg, A. D., Allis, C. D., and Bernstein, E. (2007). Epigenetics: A Landscape Takes Shape. *Cell*, 128(4):635–638.

Hanson, M. A. and Godfrey, K. M. (2015). Genetics: Epigenetic Mechanisms Underlying Type 2 Diabetes mellitus. *Nature*, 11:261–263.

Henry, V. J., Bandrowski, A. E., Pepin, A.-S., Gonzalez, B. J., and Desfeux, A. (2014). Omictools: an informative directory for multi–omic data analysis. *Database*, 2014:1–5.

Jones, P. A. (2012). Functions of DNA Methylation: Islands, Start Sites, Gene Bodies and Beyond. *Nature Reviews Genetics*, 13:484–492.

Jones, P. A. and Baylin, S. B. (2007). The Epigenomics of Cancer. *Cell*, 128(4):683–692.

Kaidery, N. A., Tarannum, S., and Thomas, B. (2013). Epigenetic Landscape of Parkinson's Disease: Emerging Role in Disease Mechanisms and Therapeutic Modalities. *Neurotherapeutics*, 10:698–708.

Kircher, M. and Kelso, J. (2010). High-throughput dna sequencing – concepts and limitations. *BioEssays*, 32(6):524–536.

Knopman, D. (2011). *Clinical Aspects of Alzheimer's Disease*, pages 37–50. Wiley-Blackwell, 2nd edition edition.

Korshunova, Y., Maloney, R. K., Lakey, N., Citek, R. W., Bacher, B., Budiman, A., Ordway, J. M., McCombie, W. R., Leon, J., Jeddeloh, J. A., and McPherson, J. D. (2008). Massively parallel bisulphite pyrosequencing reveals the molecular complexity of breast cancer-associated cytosine-methylation patterns obtained from tissue and serum DNA. *Genome Research*, 18(1):19–29.

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

Laird, P. W. (2010). Principles and Challenges of Genome-Wide DNA Methylation Analysis. *Nature Reviews Genetics*, 11(12):191–203.

Ling, C. and Groop, L. (2009). Epigenetics: A Molecular Link Between Environmental Factors and Type 2 Diabetes. *Diabetes*, 58(12):2718–2725.

Lister, R., Mukamel, E. A., Nery, J. R., Urich, M., Puddifoot, C. A., Johnson, N. D., Lucero, J., Huang, Y., Dwork, A. J., Schultz, M. D., Yu, M., Tonti-Filippini, J., Heyn, H., Hu, S., Wu, J. C., Rao, A., Esteller, M., He, C., Haghighi, F. G., Sejnowski, T. J., Behrens, M. M., and Ecker, J. R. (2013). Global epigenomic reconfiguration during mammalian brain development. *Science*, 341(6146).

Lister, R., Pelizzola, M., Dowen, R. H., Hawkins, R. D., Hon, G., Tonti-Filippini, J., Nery, J. R., Lee, L., Ye, Z., Ngo, Q.-M., Edsall, L., Antosiewicz-Bourget, J., Stewart, R., Ruotti, V., Millar, A. H., Thomson, J. A., Ren, B., and Ecker, J. R. (2009). Global epigenomic reconfiguration during mammalian brain development. *Nature*, 462(7271):315–322.

Lutsik, P., Feuerbach, L., Arand, J., Lengauer, T., Walter, J., and Bock, C. (2011). BiQ Analyzer HT: locus-specific analysis of DNA methylation by high-throughput bisulfite sequencing. *Nucleic Acids Research*, 39(suppl 2):W551–W556.

Marques, S. and Outeiro, T. F. (2013). Epigenetics in Parkinson's and Alzheimer's Diseases. In Kundu, T. K., editor, *Epigenetics: Development and Disease*, volume 61 of *Subcellular Biochemistry*, pages 507–525. Springer Netherlands.

Mastroeni, D., Grover, A., Delvaux, E., Whiteside, C., Coleman, P. D., and Rogers, J. (2011). Epigenetics Mechanisms in Alzheimer's disease. *Neurobiol Aging*, 32(1):1161–1180.

McKhann, G., Drachman, D., Folstein, M., Katzman, R., Price, D., and Stadlan, E. M. (1984). Clinical diagnosis of Alzheimer's disease: report of the NINCDS-ADRDA Work Group under the auspices of Department of Health and Human Services Task Force on Alzheimer's Disease. *Neurology*, 34(7):939–944.

Morgan, M., Anders, S., Lawrence, M., Aboyoun, P., Pagès, H., and Gentleman, R. (2009).

ShortRead: a Bioconductor package for input, quality assessment and exploration of high-throughput sequence data. *Bioinformatics*, 25:2607–2608.

Mosquera, J. L. (2014). *Methods and Models for the Analysis of Biological Significance Based on High Throughput Data*. PhD thesis, University of Barcelona, Department of Statistics.

Portela, A. and Esteller, M. (2010). Epigenetic Modifications and Human Disease. *Nature Biotechnology*, 28:1057–1068.

R Core Team (2015). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Roche (2009). *Technical Bulletin: Amplicon Fusion Primer Design Guidelines for GS FLX Titanium Series Lib-A Chemistry*.

Team, P. C. (2002–2015). *Perl 5*.

Tucker, T., Marra, M., and Friedman, J. M. (2009). Massively Parallel Sequencing: The Next Big Thing in Genetic Medicine. *American Journal of Human Genetics*, 85(2):142–154.

Warnes, G. R., Bolker, B., Bonebakker, L., Gentleman, R., Huber, W., Liaw, A., Lumley, T., Maechler, M., Magnusson, A., Moeller, S., Schwartz, M., and Venables, B. (2015). *gplots: Various R Programming Tools for Plotting Data*. R package version 2.17.0.

Wickham, H. (2007). Reshaping data with the reshape package. *Journal of Statistical Software*, 21(12):1–20.

Wood, H. (2014). AD-susceptible brain regions exhibit altered DNA methylation. *Nature Reviews Neurology*, 10(548).

Xi, Y. and Li, W. (2009). BSMAP: whole genome Bisulfite Sequence MAPping program. *BMC Bioinformatics*, 10(232):1–9.