Universidade de Aveiro

Departamento de Química

Ano 2016

VASCO SILVA OLIVEIRA CLUNY

ESTUDO EXPLORATÓRIO DOS PADRÕES EPIGENÓMICOS ASSOCIADOS AO ENVELHECIMENTO

EXPLORATORY STUDY OF AGE RELATED EPIGENOMIC PATTERNS

Universidade de Aveiro

Ano 2016

Departamento de Química

VASCO SILVA OLIVEIRA CLUNY

ESTUDO EXPLORATÓRIO DOS PADRÕES EPIGENÓMICOS ASSOCIADOS AO ENVELHECIMENTO

EXPLORATORY STUDY OF AGE RELATED EPIGENOMIC PATTERNS

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

o júri

Presidente	Professor Doutor João Filipe Colardelle da Luz Mano Professor Catedrático, Universidade de Aveiro
Arguente	Doutora Patrícia Joana Morais Ferreira Oliveira Bolseira de Pós-Doutoramento, Ipatimup – Instituto de Patologia e Imunologia Molecular da Universidade de Porto
Orientador científico	Professora Doutora Gabriela Maria Ferreira Ribeiro de Moura Professora Auxiliar, Universidade de Aveiro

agradecimentos

Quero agradecer ao Professor Doutor Manuel Santos a oportunidade que me concedeu em escolher este fascinante tema para dissertação desta tese de mestrado e em facultar as melhores condições para o desenvolvimento da mesma. Quero agradecer também à Professora Doutora Gabriela Moura a dedicada orientação e disponibilidade constante para prestar qualquer esclarecimento.

Quero agradecer ao curador do NCBI Richard Lapoint, ao professor José Luís Oliveira e ao meu amigo André Leitão pela ajuda preciosa que me deram na utilização de bases de dados e procura de ficheiros das bases de dados. Quero a agradecer á Carolina Conceição pela troca de ideias, informação e amizade no início da Tese. Quero agradecer ao Dr Kevin Bowling do Hudson Alpha Institute for Biotechnology por me ter facultado *metadata* em falta de amostras nas bases de dados, usadas no nosso *dataset.*

Quero agradecer à minha família, aos meus amigos e amigas por todo o apoio e compreensão que manifestaram desde o ínicio em relação aos momentos importantes que não pude partilhar presencialmente com eles e com elas. Quero agradecer em particular à Andreia Reis, Rita Coimbra e Manuela Oliveira toda a ajuda, apoio técnico e (algumas!) críticas prestadas ao longo de toda a experiência e cruciais para a concretização desta Tese.

palavras-chave ADN, bases de dados, *bisulfite-seq*, cérebro, envelhecimento, epigenoma, ilhas CpG, NGS, metilação, metilcitosinas, PCR, sequenciação, *reads*, sangue.

resumo

Sabe-se hoje que o genoma humano, para além da sua sequencia nucleotídica, revela várias alterações químicas no DNA, nomeadamente metilações das citosinas. Estas modificações estabelecem padrões específicos que podem ser transmitidos de uma geração para a seguinte e exercem controlo sobre os genes que são expressos a cada momento nas células, tecidos ou orgãos. Esta tese teve como objectivos: explorar as principais bases de dados que contêm dados epigenómicos relevantes; obter ficheiros *fastq* de bibliotecas bisulfite-seq aplicando métodos de *data mining* a dados reais de bases de dados públicas de sequenciação de segunda geração; alinhar e mapear estes ficheiros usando software adequado (Methy-Pipe); fazer uma análise comparative por forma obter características associadas ao envelhecimento saudável de indivíduos e á evolução do epigenoma ao longo da vida; finalmente é esperado que, após atingidos os objectivos anteriores, se perceba o contributo do epienoma no envelhecimento saudável das populações .

keywords Aging, epigenome, bisulfite-seq, blood, brain,CpG islands, DNA, databases, epigenetics, NGS, methylation, methylcytosines, PCR, sequencing, reads

abstract It is already known today, that the human genome, in addition to its nucleotide sequence, shows multiple chemical modifications at the DNA level, namely cytosine methylations. These modifications changes establish specific patterns that can be transmitted from generation to generation and exercise control over the genes that are expressed at every moment in the life of the cells / tissues / organs. This thesis aimed to: understand the contribution of the epigenome to a healthy lifestyle; to explore the main databases containing relevant epigenomic data; to obtain fastq files of bisulfite-seq libraries by applying data mining methods to real data from next generation public databases; to align and map these files using adequate software (Methy Pipe); to do a comparative analysis in order to identify features associable to a healthy aging of individuals and the evolution of the epigenome in humans throughout life.In doing so, it is expected that this work will contribute to the understanding of the contribution of the epigenome to a healthy lifestyle.

List of Abbreviations and Acronyms

ATP-Adenosine Tri Phosphate
BS-seq-Bisulphite Sequencing
BER-Base Excision Repair
CCD-Charged Coupled Device
CGI-CpG Islands
DAS-Distributed Annotation System
DDBJ-DNA Database of Japan
DMP-Differentially Methylated Position
DMR-Differentially Methylated Regions
DNA-Deoxyribo Nucleic Acid
DNTPs-DeoxyNucleotides Triphosphate
DNMT-DNA Methyl Transferase
EBI-European Bioinformatic Institute
ENA-European Nucleotide Archive
ESC-Embryonic Stem Cells
FRET-Fluorescence Resonance Energy Transfer
GEO-Gene Expression Omnibus
HGP-Human Genome Project
HPLC-High Performance Liquid Chromatography
HTS-High throughput sequencing
INSDC-International Nucleotide Sequence Database Collaboration
IPSC-Induced Pluripotent Stem Cells
5cC5carboxylcytosine
5fC-5formylcytosine
5mC-5methylcytosine
5hmC-hidroxymethylcytosines
5hmU-hydroxymethyluracil LINE-Long Interspaced elements
MBD-Methyl-CpG Binding Domain
MD-Methylation Density
MeDIP-seq-Methylated DNA Immuno-precipitation sequencing
Methyl-seq-Methylation Sequencing
mQTLs-Methylation Quantitative Trait oci
mRNA-messenger RNA
MRE-seq-Methyl-sensitive Restriction Enzyme Sequencing
NCBI-National Center for Biotechnology Information
NHGRI-National Human Genome Research Institute
NIH-National Institutes of Health

Table of Contents

List of Figures and Tables

1.INTRODUCTION

1.1 EPIGENETICS, EPIGENOMICS AND EPIGENOMIC MECHANISMS

1.1.1 Basic Concepts and Definitions

Epigenetic is a recent scientific field that refers to the study of every process that regulates the expression of certain genes without causing any change in the primary sequence of the genome. Epigenomics, on the other hand, refers to the global analysis of epigenetic changes that occur genome-wide (Fraga, 2009).

The renewed interest in epigenetics has led to new findings about the relationship between epigenetic changes and a host of disorders including age related disorders. The increased knowledge and improved technologies in epigenetics made recently, allow us to better understand the interplay between genetics and epigenetics; and hopefully will lead to the development of new targeted and personalized treatments across the clinical spectrum.

There are three main epigenetic mechanisms: including non-coding RNA (ncRNA), histone modification and DNA methylation; that regulate various biological processes (see figure 1).

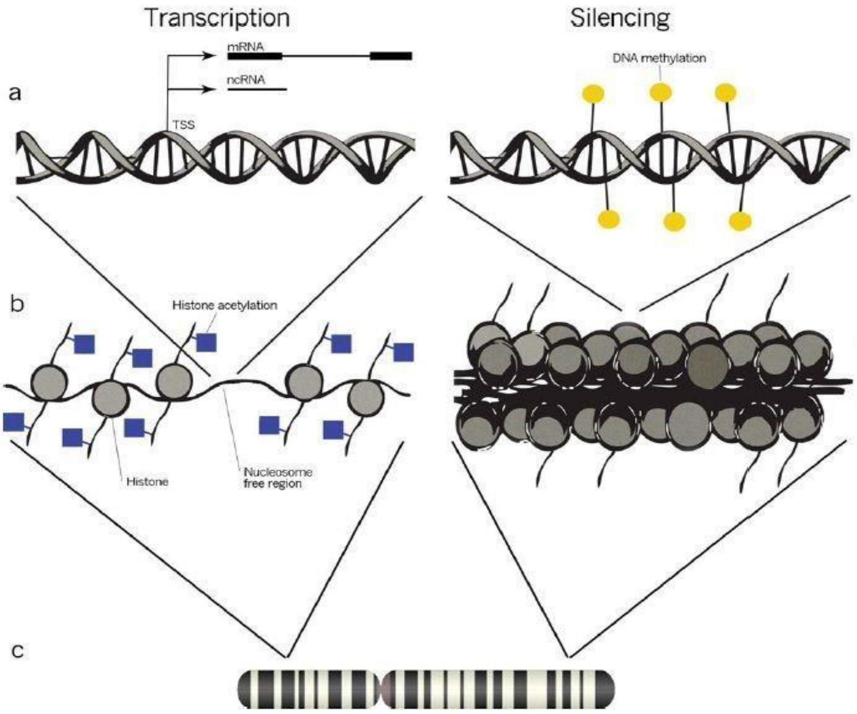


Figure 1 - Overview of epigenetic features. "Each chromosome (panel c) consists of both condensed and open chromatin regions (panel b), with different histone modifications present. Loose regions are, for example, characterized by histone lysine acetylation and the possibility of gene expression. Nucleosome (re)positioning results in nucleosome free regions, for example, at the transcription start site (TSS) (panel b), which is required for gene transcription (panel a). The resulting transcriptome not only consists of coding mRNAs, but also of noncoding RNAs (ncRNA). Promoter regions of transcriptionally silenced genes are typically densely packed without nucleosome free regions, lack histone lysine acetylation (panel b), and are often featured by DNA methylation (panel a)." Adapted from Klaas Mensaert, (2014).

1.1.2 DNA methylation and demethylation effects

One of the best understood molecular epigenetic mechanisms, and on which rests this study is the methylation of cytosine residues in DNA specific position. Cytosine methylation group (5-methylcytosine or 5mC), is caused by the covalent addition of a methyl group from S-adenosyl methionine to carbon 5 of cytosines, by a family of DNA methyltransferases (He, Chen, & Zhu, 2011) (see figure 2).

This mechanism is essential for normal development and cell/tissue differentiation, but is also associated with a number of key processes including genomic imprinting (an epigenetic mechanism that consists in silencing an allele depending on its parent of origin), X chromosome inactivation, suppression of repetitive elements (such as retrotransposons) and oncogenes; and to maintain chromosome stability (due to the hypermethylation of telomeres and centromeres).

(Ester Lara V. C., 2011). Once established, global cytosine methylation patterns must be conserved in order to keep transposons and oncogenes repressed, thus preserving cell integrity (Kim, Samaranayake, & Pradhan, 2009;jacobsen, 2010). The maintenance of methylation patterns through the genome is a complex but very important process involving many factors, but mainly catalyzed by DNMT1 methyltransferase during DNA replication and regulated by cofactors such as Np95 (Winnefeld & Lyko, 2012; Igor.P.Pogribny, 2009). It has been speculated that decreased DNMT1 expression might contribute to global hypomethylation upon aging (Weidner CI, 2014). In addition, the *de novo* methyltransferases DNMT3A and DNMT3B, do not only contribute to methylation maintenance, but also have the capacity to establish new methylation marks (Marc Winnefeld, 2012). Another member of the DNMT3 family is DNMT3L, which has no catalytic activity, but interacts with DNMT3A and DNMT3B and stimulates their enzymatic activity. The mechanism by which methyl groups are removed from methylated DNA is thought to be initiated by the oxidation of 5′-methylcytosine into 5-hyroxymethylcytosine, which is catalyzed by the ten-eleven translocation (TET) family enzymes. The oxidized form of 5mC by the ten- eleven translocation 1,2, or 3 (TET1,2,3), 5hmC, can be demethylated, also by TET1,2,3, into 5-formylcytosine(5fC) and finally into 5-carboxylcytosine(5caC)(figure 2) also plays an important functional role.

Activation-induced cytidine deaminase/apolipoproteinB mRNA-editing, enzyme-catalytic, polypeptide (AID/APOBEC) family of deaminases can also deamination 5hmC, forming 5-hydroxymethyluracil (5hmU). This led to the hypothesis that 5hmC is an active demethylation mark of 5mC.

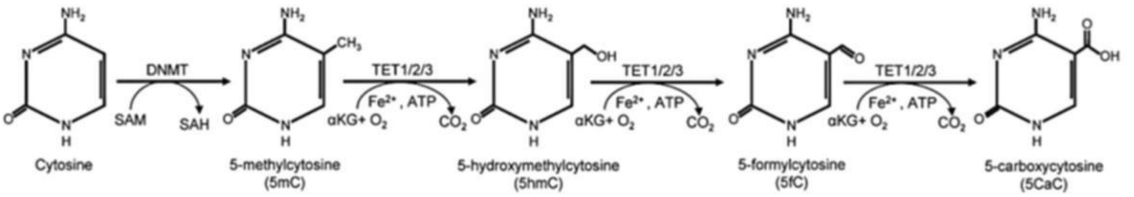


Figure 2 – "DNA cytosine methylation reaction catalyzed by DNMTs and DNA 5-hydroxymethylcytosine oxidative reactions catalyzed by the TET family enzymes." Adapted from Vichithra R. B. Liyanage 1, 2014.

After 5fC and 5caC are formed, the N-glycosidic bond is destabilized and subsequently thymine DNA glycosylase (TDG) and methylated DNA binding domain- containing protein 4 (MBD4) glycosylases initiate Base Excision Repair (BER) by removing the modified base, leading to an apurine/pyramidine site (AP site) (Vichithra R. B. Liyanage 1, 2014). These sites are toxic and have to be replaced with a base. AP endonuclease 1 (APEX1) then cleaves the AP site, allowing DNA polymerase to re-insert the appropriate base, in this case cytosine.

In Embrionic Stem Cells (ESCs), 5hmC levels dominate that of 5fC and 5caC levels, implying that TET expression is strictly controlled. (Vichithra R. B. Liyanage 1, 2014).

In somatic cells, 5-mC occurs mostly in the CpG context, where a cytosine nucleotide is located next to a guanidine nucleotide. An exception to this rule can be seen in embryonic stem (ES) cells, where a considerable amount of 5-mC can also be seen in non-CpG contexts. In the human genome, methylated CpGs cover approximately 1.5% of genomic DNA and affect 70–80% of all CpG dinucleotides in the genome, being irregularly distributed into CpG-poor regions and CpG-rich regions named "CpG islands" (CGI) (Bestor, Edwards, & Boulard, 2014) (figure 3). In normal cells, the CpG islands are usually unmethylated, unlike the rest of the CpG which are usually methylated. These are found mainly in the 5' end of the regulatory region, particularly enhancers and transcription factor-binding sites, of approximately half of all genes allowing for cell/tissue differentiation. Repetitive sequences also contain a high percentage of CpG dinucleotides that are normally hypermethylated (Lister *et al.*, 2009). The rest of CGI are either within or between characterized transcription units and have been termed ''orphan'' CGIs to reflect uncertainty over their significance (Illingworth *et al.*, 2010).

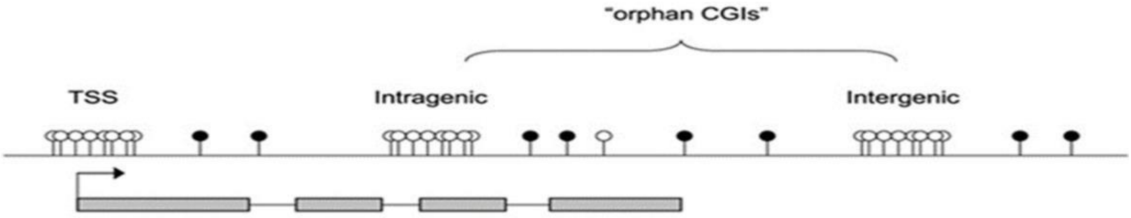


Figure 3 - The genomic distribution of CGIs. "(A) CGIs can be located at annotated TSSs, within gene bodies (Intragenic), or between annotated genes (Intergenic). Intragenic and intergenic CGIs of unknown function are classed as ''orphan'' CGIs. (Empty circles) Unmethylated CpG residues. (Filled circles) Methylated CpG residues." Adapted from Bird, 2011.

What distinguishes DNA methylation in vertebrate genomes is the fact that only a portion of CpGs are methylated in a given cell type, thus constituting differential methylated regions (DMR), which are stretches of DNA in an organism's genome that have different DNA methylation patterns between samples (Szyf, 2010). The difference between these methylated regions is mainly due to factors, such as gender, tissue/cell type and also age, which is the one we are most interested in.

Moreover, in the DNA there are sequence variants that are associated with DNA methylation patterns dispersed throughout the genome across different tissues, known as methylation quantitative trait loci (mQTLs), (Zhang *et al.*, 2014).These have been demonstrated to act mostly at cis level (such as promoters and enhancers) while the majority of estimated genetic variation that influences methylation levels is acting at trans level (such as transcription factors), increasing the level of regulation complexity between methylation and the DNA sequence(Gaunt *et al.*, 2016).

Levels of DNA methylation at a promoter-associated CpG island are generally associated with gene repression, although the opposite effect has been noted in particular genes (Meaghan, 2015). In general, CpG-rich promoters are largely unmethylated, regardless the state of expression; whereas CpG-poor promoters drift toward partially methylated states during prolonged inactivityand begin demethylation when transcription is initiated (Bestor *et al.*, 2014). It is believed that methylated promoter CGIs are usually restricted to genes at which there is long- term stabilization of repressed states, such as imprinted genes and genes located in the inactive X chromosome (P. A. Jones, 2012). However, many CGI free promoters are active in a tissue specific manner suggesting that they can be tightly regulated, as well (Deaton & Bird, 2011).

Conversely, DNA methylation in the gene body is often positively associated with levels of gene expression (Lister *et al.*, 2009; Gutierrez Arcelus *et al.*, 2013), which is thought to be due to the tissue-restricted use of CpG islands as alternative transcription start sites. Recent findings about intragenic (or gene body methylation) and its role in alternative splicing have changed the conventional view of the role of DNA methylation in transcription by proving to be an enrichment of DNA methyl marks within exons in contrast to the nearby intronic regions.

Moreover, there are some differences in the CpG and methylation density between splice donor and acceptor sites. The involvement of DNA methylation in splicing was further observed in alternate exons and spliced exons (Liyanage *et al.*, 2014). Interestingly, this negative correlation does not hold true when comparing expression and DNA methylation for a specific gene across individuals (Bestor *et al.*, 2014; M. J. Jones, Goodman, & Kobor, 2015).

Non-CpG methylation has been detected in mammals recently. Studies have found CpH (H = A/C/T) methylation to be present in cultured pluripotent stem cells, namely embryonic stem cells (ESCs), induced pluripotent stem cells (IPSC), and adult stem cells, where it may help repressing genes as cells transition into their differentiated state. Low levels of CpA methylation have been observed in early mouse embryos and ESCs, but are significantly decreased in somatic cells (Ursula Munoz Najar, 2011). Several recent profiling studies have shown the presence of CpH methylation in the adult mouse cortex and human brains which consist of mixtures of many neural subtypes. Thus, it indicates a tissue-specific distribution that is different from those genes that were previously identified in embryonic stem cells and the brain (Schultz *et al.*, 2015). Actually, as observed in studies done in brain tissue, the preferential CpH methylation in CpG- depleted regions suggests that CpH methylation might compensate for the lack of CpGs and increase the local mC density in neurons without adding constitutively methylated new CpG dinucleotides to the genome (Junjie U Guo, 2014).

DNA methylation offers significant advantages as a biomarker over expression-based and proteomic based markers, namely: the high stability of the DNA , which can survive routine processing for histopathology; the possibility to compare DNA methylation levels with absolute reference points (completely methylated or completely unmethylated DNA); the ability to amplify and identify by by polymerase chain reaction (PCR)-based approaches even when alterations are present only in a few cells (Olkhov-Mitsel & Bapat, 2012).However it is also difficult to establish a suitable reference for comparison due precisely to the susceptibility of DMRs to change according to the aforementioned factors.

1.2 THE EFFECT OF DNA METHYLATION ON AGING AND HUMAN AGE-RELATED DISEASES

Aging is a complex process that results from an advanced state of a series of degenerative processes such as somatic mutations, telomere attrition, activation of transposable elements and oxidative stress which ultimately leads to a loss of physiological integrity and increased susceptibility to diseases over time (Weidner & Wagner, 2014). Although age can be measured chronologically from the date of birth (chronological age), it can also be measured by a set of health-related biomarkers (biological age). In this regard, epigenomic studies on phenotypes associated with aging can help detect molecular changes related to the biological aging process. The retrotransposon theory of aging, for example, which hypothesizes that epigenetically silenced transposable element become deleteriously activated as cellular defense and surveillance mechanisms break down with age, has been supported recently (Wood *et al.*, 2016).

It is clear that the genetic component of methylation variation across the genome is relevant, but, in what concerns to variation between individuals, environmental or stochastic influences are a more important determinant of sex-specific and age-specific methylation than genetic influences, which is estimated to be around 25–30 % (Gaunt *et al.*, 2016; van Dongen *et al.*, 2016). In what regards to the gender effect on DNA methylation, Tsong *et al.* (2005) suggests that gender is at least as strong a predictor of methylation level in the genes under study as age and Allison M. Cotton (2011) proved that X-linked promoters show differences in methylation dependent on sex and CpG density in four autosomal genes (ESR1, MTHFR, CALCA and MGMT) (see table 13 in appendix).

Some aspects of mammalian aging result from an age associated decrease in number and decline in the replicative function of adult or somatic stem cells. One of the major mechanisms known to interfere with somatic stem cell function during aging is the accumulation of unrepaired DNA and chromosomal damage which, consequently, prevents the right production of differentiated cells for proper tissular function (Huidobro, Fernandez, & Fraga, 2013)

Once aging is the main risk factor of several human disorders, it is very likely that age related processes including epigenetic alterations and oxidative stress, promote the onset and development of these illnesses (Lardenoije & Iatrou, 2015). Complex disorders, such as neurodegenerative, are caused by the contribution of genetic and environmental factors and not one isolated (Sanchez-Mut *et al.*, 2016). Early life experiences modify the neurobiology of development and such influences continue to affect biological patterns and psychological outcomes in adulthood (Kanherkar, Bhatia- Dey, & Csoka, 2014).

In this context, epigenetics, acting as a mediator between genome and environment, is a key modulator of adult neurogenesis, affecting extracellular signaling molecules and patterns of neural circuit activity (figure 4).

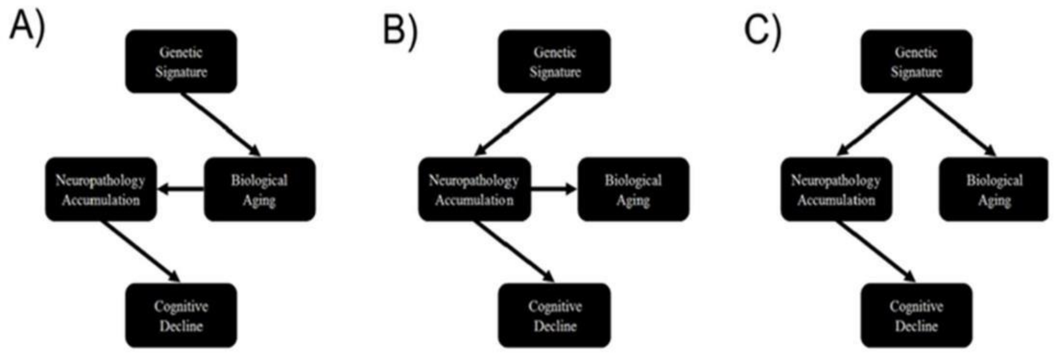


Figure 4 - Causal scenarios that can explain the significant genetic correlations between epigenetic age, neuropathology and cognitive decline. Adapted from (Levine, Lu, Bennett, & Horvath, 2015).

Despite a global DNA methylation in early life and gradual demethylation in later life across the genome, these changes are not symmetrical. Actually, these changes can be explained considering the periods and locations where they occur: in the early life when the rate of change is much higher and DNA methylation is gained globally, mainly at island shores and intergenic regions; and later life when DNA the rate of change is slower and DNA is lost globally but still gained at islands and shores (Gronniger *et al.*, 2010; Meaghan, 2015). So the CpG sites inside CGIs are more likely to gain rather than loose DNA methylation with aging (Numata *et al.*, 2012).

Because most CpGs are located outside of CpG islands and are highly methylated, this leads to a global loss of DNA methylation in later life as well as a tendency for DNA methylation levels to shift toward the mean with increased age. This changes make what is called the epigenome erosion and are responsible for what is called broadly the epigenetic drift. This drift was noticed for the first time in studies done with identical twins where it was proven that the difference of DNA methylation patterns between the twins increases throughout aging (Fraga *et al.*, 2005). A component of this drift is tissue-specific, but another component is tissue-independent, aiming for stem cell differentiation pathways which may explain the increased dysfunction of stem cell with age (Teschendorff, West, & Beck, n.d.). Another interesting concept, other than the epigenetic drift, is the epigenetic clock which regards specific genomic sites that are more likely to suffer methylation changes in throughout aging (M. J. Jones, Goodman, & Kobor, n.d.).

Previous studies noticed that CpG sites in genes such as those involved in cancer and tumor suppression, DNA repair, and telomere maintenance have mostly increased methylation with aging. These include MGMT, ESR1, RASSF, RAD50, GSTP1/GTS3, RARB, MYOD1, LAMB1, and the Werner gene WRN, the latter gene associated with a premature aging syndrome (for more details about these genes see table 14 in Appendix). A more recent study found genes with age-related methylation changes throughout life, and particularly in the transition from fetal to postnatal life period, for genes, such as DLG4, DRD2, NOS1, NRXN1, and SOX10, that have been implicated in schizophrenia and autism (for more details about these genes see table 14 in Appendix) (Numata *et al.*, 2012).

1.3 THE USE OF BLOOD AS A SAMPLE TISSUE FOR DETECTION OF HUMAN AGE-RELATED DISEASES

In one hand, some investigators believe that the correlations between blood and brain versus two brain tissues are very similar, and so blood could be used to predict the methylation patterns in a specific brain tissue in a similar degree as another brain tissue (K. A. Aberg *et al.*, 2014). Some even found that epigenetic age acceleration in dorsolateral prefrontal cortex (DLPFC) is highly heritable in a similar way reported for blood (Levine *et al.*, 2015).

Actually, the use of surrogate tissues seems supported by studies suggesting that tissue-specific DMRs, (Brock C. Christensen, 2009) constitute only a limited proportion of all methylated sites. Blood, in particular, is a multicellular tissue and this heterogeneity can reduce inter individual differences as the cell type differences may average out across subjects (Konstantin Shakhbazov, 2016).

This can be explained by three possible reasons. First, peripheral tissues may reveal methylation marks resulting from the epigenetic *de novo* reactions affecting germline and embryogenesis (K. A. Aberg *et al.*, 2013) (Monk M, 1987*;* Efstratiadis, 1994). As the epigenetic profile is inherited, these epigenetic mutations can also be detected in multiple tissues. Second, blood contains cells that may be modified while they travel through unhealthy tissues, includin cell- free DNA from those tissues (serum DNA can define tumor-specific genetic and epigenetic markers in gliomas of various grades). As such, traces of the aberrant methylation in disease- targeted regions may be detectable in blood. Finally, and perhaps most importantly, environmental factors such as diet, drugs and lifestyle factors, as well as genetic polymorphisms can affect methylation levels. Although these changes may only affect some tissues, it is very likely that the changes themselves are more global and cause similarities in methylation profiles across tissues. (McGowan, Meaney, & Szyf, 2008*;* Pilsner, 2007).

On the other hand, there is also data supporting the opposite idea, that is, tissues have specific DMR and so, they can’t be replaced by any other tissue. It has been even shown that tissue specific variation in DNA methylation greatly exceeds interindividual differences within any one tissue (Davies *et al.*, 2012). As an example of differentially methylated genes were reported the following genes (in brain, pleura, lung, blood and solid tissues): ESR1, GSTP1, IGF2, MGMT, MYOD1, MYOD1, RARB, RASSF1, RASSF1, DNMT1, DNMT3B, HDAC1, HDAC5, HDAC7A, HDAC11, LAMB1, RAD50 , TERT and WRN.

This is particularly problematic for neurodegenerative diseases because a typical brain biopsy contains multiple cell types, including neurons, astrocytes and other glia cells, all with different methylation patterns (Lord & Cruchaga, 2014).

Particularly in Alzheimer’s Disease’s (AD) case, the data supports that the top-ranked Differentially Methylated Positions (DMPs) in blood are distinct to those identified in the brain and there is no significant overlap with either cortex or cerebellum suggesting that AD-associated DMPs in blood are unlikely to be directly related to the actual neurodegenerative process itself.

Although distinct from Alzheimer Disease's associated changes occurring in the brain, many of the Alzheimer´s Disease-associated DMPs (ANK1, RPL13, RHBDF2, CDH23, ABCA7 and BIN1) (see table 15 in Appendix) identified in blood before death may be used as detectable transcriptomic changes and, given the "relative stability and ease of profiling DNA modifications compared to RNA, have potential utility as diagnostic biomarkers of the disorder" (Davies *et al.*, 2012) (see tables in appendix).

1.4 EPIGENOMIC GENOME-WIDE PROJECTS

Believing that our knowledge of the Epigenome will extend our understanding of the genome regulation, development, disease etiology, and even define polymorphic variation in populations susceptibility to diseases, some Epigenomic Genome-Wide projects were initiated in order to explore and understand epigenetic mechanisms and elements.

1.4.1 Roadmap Epigenomic Project

"The Roadmap Epigenomic Program (also known as Epigenomic Roadmap initiative), launched by NIH (2008), seeks to create a series of epigenome maps to study epigenetic mechanisms, develop new epigenetic analytics, generate a repository and long-term data archive, standardize procedures and practices in epigenomics and support new technologies for these" (Shakya, O'Connell, & Ruskin, 2012). As part of the $190 million, five-year initiative, the Roadmap Epigenomics Mapping Consortium44 was formed to provide a public database for human epigenomic data, the Human Epigenome Atlas. (http://www.roadmapepigenomics.org/overview/epigenomics-human-health).

To attain substantial coverage of the human epigenome, International Human Epigenome Consortium (IHEC) aims to decipher at least 1,000 epigenomes in the next years. Officially launched in Paris (Bae, 2013), "with an initial (first phase) budget target of $130 million, IHEC aims to coordinate the mapping of epigenomes from the NIH's Epigenomics Mapping Consortium and from the European Epigenome Networkof Excellence, the Danish National Research Foundation Centre for Epigenetics, and the Australian Epigenetic Alliance. The IHEC web portal provides links to databases, such as GEO, ARRAYEXPRESS and DDBJ, where epigenetic sequencing data will be made available "(Shakya, O'Connell, *et al.*, 2012).

Several papers published by the Roadmap Epigenomic Project up to now investigate histone modifications and DNA methylation providing insights into the relationship between histone signatures and gene expression throughout development and adult life.

1.4.2 Encode

Another important program including epigenetic data, is the Encyclopedia of DNA Elements (ENCODE). "This is supported by the ENCODE Consortium, an international collaboration of research groups funded by the National Human Genome Research Institute (NHGRI). The NHGRI began as the National Center for Human Genome Research (NCHGR), which was established in 1989 to carry out the role of the National Institutes of Health (NIH) in the International Human Genome Project (HGP) and is today a component of the National Institutes of Health (NIH) and the Department of Health and Human Services (DHHS)" (https://www.genome.gov/10001763/about-nhgri-a-brief-history-and-timeline/).

The purpose of this project is to identify every functional element in the human genome sequence, both at the protein as well as RNA levels, and regulatory elements that control cells and circumstances in which a gene is active (Shakya, O'Connell, *et al.*, 2012). A remarkable feature of this project was the discovery of hundreds of thousands of enhancer-like regions in the mammalian genome that regulate gene expression at long range, through epigenomic signatures. From this vast set, each cell type is regulated by a subset of perhaps 20,000–40,000 enhancers, which determine its particular gene-expression profile (Romanoski, Glass, Stunnenberg, Wilson, & Almouzni, 2015).

1.4.3 MethDB

The concept of methylomes was first introduced by Andrew Feinberg, defined as "the complete set of DNA methylation modifications of a cell" (Novik *et al.*, 2002). MethDB is a source for experimentally confirmed methylome data designed to store and annotate information on the occurrence of methylated cytosines in DNA. Until recently, it contained 219,905 methylation data items and 5,382 methylation patterns or profiles for 48 species, 1,511 individuals, 198 tissues and cell lines and 79 phenotypes. MethDB "also has a public online submission system available."(Shakya, O'Connell, & Ruskin, 2012)

The resource “forms part of an integrated network of biological databases through DAS (Distributed Annotation System), enabling the epigenetic data to be viewed as a layer in the human genome, and is also connected to Ensemble” (for DNA sequences with available MethDB data aligned to NCBI Refseq) (Shakya, Connell, & Ruskin, 2012).

1.5 PUBLIC DATABASES, REPOSITORIES AND LIBRARIES WITH EPIGENETIC INFORMATION

The rise of massively parallel sequencing technologies has enabled innumerous research possibilities, such as: elucidation of the human microbiome, discovery of polymorphisms and mutations, mapping of protein–DNA interactions, and positioning of nucleosomes;among others (http://www.ncbi.nlm.nih.gov/books/NBK47539/). In order to reach these goals, researchers must be able to store, access and use the big volume of short read data generated from massively parallel sequencing experiments into runs. In an attempt to deal with this situation, there have been created public databases repositories and libraries to store all the data.

1.5.1 INSDC

The International Nucleotide Sequence Database Collaboration (INSDC) is an initiative that operates between DDBJ, EMBL-EBI and NCBI (figure 5). It contains the spectrum of raw data reads, alignments and assemblies, annotated with metadata and setup configurations. (http://www.insdc.org/) The INSDC advisory board, is made up of members of each of the databases' advisory bodies. who endorsed and reaffirmed the existing data-sharing policy of the three databases that make up the INSDC (ENA, NCBI and DDBJ). These must be followed by anyone who wants to submit their data (figure 5 (http://www.insdc.org/about).

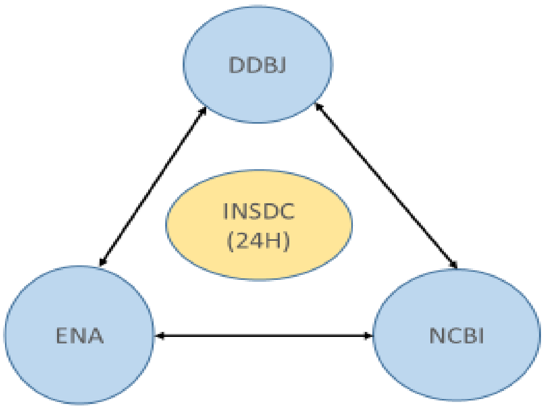


Figure 5 – There is permanent data exchange between the different international databases: NCBI, EBI DDBJ. Picture done by the author

1.5.2 NCBI: Genbank and GEO

As the US national resource for molecular biology information, “NCBI is responsible for: creating automatic systems capable of storing and analyzing knowledge about molecular biology, biochemistry, and genetics; facilitating the use of such databases and software by the research and medical community; coordinating efforts to gather biotechnology information both nationally and internationally; and performing research into advanced methods of computer-based information processing for analyzing the structure and function of biologically important molecules” (http://www.ncbi.nlm.nih.gov/home/about/).

GenBank is the NCBI’s genetic sequence database, an annotated library of every available DNA sequence. The complete release notes for the current version of GenBank are available on the NCBI ftp site. These databases were conceived to provide and promote access within the scientific community to the most up to date and comprehensive DNA sequence information. Therefore, NCBI does not restrict the use or distribution of the GenBank data. However, some submitters may claim patent, intellectual property rights in the data they have submitted (http://www.ncbi.nlm.nih.gov/genbank/).

GEO is NCBI’S international public repository that archives and freely distributes microarray, next-generation sequencing, and other types of high-throughput data. The three main goals of GEO are to: store high-throughput functional genomic data in a versatile database; facilitate submission procedures and formats that support metadata; provide user-friendly mechanisms that allow users to search and download studies and gene expression profiles of interest (http://www.ncbi.nlm.nih.gov/geo/info/overview.html). The main data records found in GEO are: Platform (GPL), or the technology used and the features detected; Sample (GSM) or preparation and description of the sample, with their own set of runs (SRR); Series (GSE) defines a set of samples and how they are related; and DataSets (GDS) which is the sample data collections assembled by GEO staff. Both samples and Series are submitted by experimentalists, the platforms are submitted by the manufacturers and the datasets are curated by NCBI experts.

1.5.3 EMBL/EBI and ArrayExpress

The European Nucleotide Archive (ENA) stores experimental data that are based in nucleotide sequencing workflows. The workflows typically include the isolation and preparation of material for sequencing, a run of a sequencing machine in which sequencing data are produced and a subsequent bioinformatics analysis pipeline. “ENA records this information in a data model that covers input information (sample, experimental setup, machine configuration), output machine data (sequence traces, reads and quality scores) and interpreted information (assembly, mapping, functional annotation) “(http://www.ebi.ac.uk/ena/about). There are three ENA data types: reads, assemblies and annotations (http://www.ebi.ac.uk/ena/submit/data-formats) (figure6).

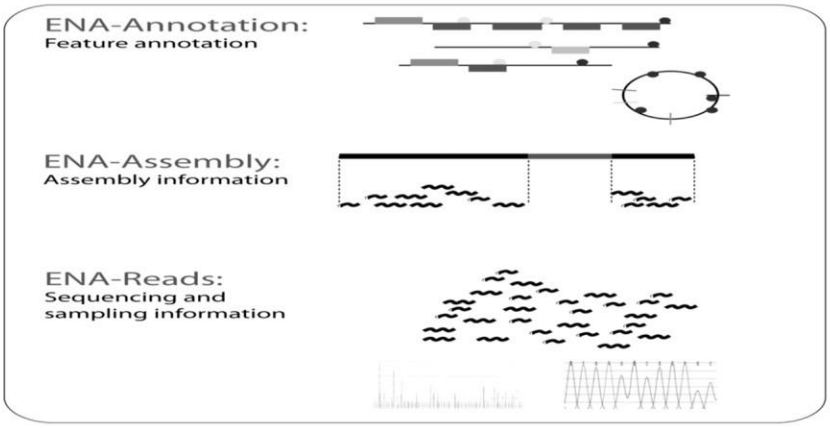


Figure 6 - The three ENA different tiers: annotations, assemblies and reads. Adapted from (https://www.ebi.ac.uk/training/online/course/nucleotide-sequence-data-resources-ebi/what-ena)

ArrayExpress is a database that stores microarray and high through-put sequencing experiments described and archived according to the community guidelines for microarray (MIAME) and High Throughput Sequencing (MINSEQE) data. The functional genomics data collected in ArrayExpress is organized into experiments, which are defined as collections of assays often related to a scientific publication. ArrayExpress can be searched to yield information about functional genomic experiments. For example, one can access data files and sample annotations relating to an experiment of interest (figure 7). Submitting functional genomics data to ArrayExpress is required by some major publishers and encouraged by others. This benefits the whole scientific community, enabling data available as part of the public record of science and facilitating meta-analysis.(https://www.ebi.ac.uk/training/online/course/arrayexpress-discover-functional-genomics- data-qui/what-arrayexpress).

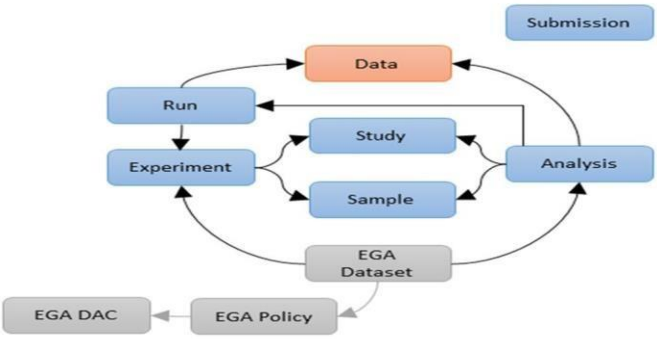


Figure 7 - The relation between study's, samples, runs and experiments in ENA database. Adapted from (http://www.ebi.ac.uk/ena/submit/metadata-model).

1.5.4 DDBJ

As the two other databases, DDBJ Center also collects nucleotide sequences from researchers, mainly from Japan, and their accession numbers. The accession number assigned to each sequence data is unique and internationally recognized to guarantee the submitter the property of the submitted and published data. Due to its origin, DDBJ submits the largest majority of Japanese data to INSDC, but also accepts data and accession numbers to researchers from any other countries. (http://www.ddbj.nig.ac.jp/intro-e.html).

1.5.5 Other Useful Databases with DNA Methylation data

MethylomeDB- "The Brain Methylome Database includes genome-wide DNA methylation profiles for human and mouse brains". (http://ww w.neuroepigenomics.org/methylomedb/)
DiseaseMeth "A web based resource focused on the aberrant methylomes of human diseases". (http://www.bio-bigdata.com/diseasemeth/).
NGSmethDB "A dedicated database for the storage, browsing and data mining of whole-genome, single-base-pair resolution methylomes. It collects NGS data from high-throughput sequencing together with bisulfite conversion of DNA from literature and public repositories, then generating high-quality chromosome methylation maps for many different tissues, pathological conditions and species". (http://bioinfo2.ugr.es:8080/NGSmethDB/).
Deepblue Epigenomics – "provides a central data access hub for large collections of epigenomic data. It organizes the data from different sources using controlled vocabularies and ontologies. The data is stored in the server, where the users can access the data programmatically or by web interface" (http://deepblue.mpi-inf.mpg.de/)

Table 1-some databases with useful DNA methylation data for epigenomic studies

1.6 SRA and SRA Toolkit

Due to the community need, the International Nucleotide Sequence Database Collaboration (INSDC), have developed the Sequence Read Archive (SRA) data storage and retrieval system (http://www.ncbi.nlm.nih.gov/books/NBK47539/). SRA is the database that stores sequence data obtained from next generation sequence (NGS) technology.

Through this database, one can query metadata to retrieve the sequence read files for download and further analyses. Specifically, SRA archives raw files of NGS data for various organisms from several platforms and requires per-base quality scores for all submitted data. Thus, unlike GenBank and some other NCBI repositories, FASTA and other formats are not sufficient for submission, although FASTA can be submitted.

The SRA Toolkit and System Development Kit from "NCBI is a collection of tools and libraries for using data in the INSDC Sequence Read Archives. Much of the data submitted these days contain alignment information. The process to restore original data, for example as FASTQ, requires fast access to the reference sequences to which the original data was aligned. NCBI recommends SRA users to dedicate local disk space to store reference sequences that need to be downloaded from the NCBI SRA site". (https://ncbi.github.io/sra-tools/)

1.7 CHRONOLOGICAL VIEW OF SEQUENCING TECHNOLOGIES

The Human Genome Project, initiated in 1990 and finished in 2003, in which researchers sequenced the entire Human genome, arose from two key ideas that emerged in the early 1980s: that new genomic discoveries could greatly improve biomedical research, by allowing researchers to attack problems in a comprehensive and unbiased fashion; and that the creation of such global views would require an effort in infrastructure building, as never seen before (Lander *et al.*, 2001). The technology used in this project was based in Sanger sequencing principles. A few years later, in 2008, it was initiated the 1000 Genome Project which aimed to catalogue human genetic variations from 1092 genomes of individuals from different ethnics. This project used next generation sequencing technologies (Auton *et al.,* 2015). Both these projects provided invaluable data about the Human Genome, but also highlighted the value of sequencing technologies and the urge for faster and cheaper ones.

1.7.1 First Generation Sequencing

First-generation sequencing was first developed by Sanger in 1975 (the chain- termination method) and, a few years later by Maxam and Gilbert (a chemical sequencing method).However Sanger sequencing prevailed given it was more user friendly and more amenable to being scaled up.

Sanger sequencing is a method that mixes dye-labelled normal deoxynucleotides (dNTPs) and dideoxy-modified and chromatophore labelled dNTPs (figure 8). A standard single primer PCR reaction is carried out and, as elongation occurs, some strands incorporate a dideoxy-dNTP, and so, ending elongation. The strands are then split up on a gel and the terminal base label of each strand is identified by laser excitation and spectral emission analysis (Sara Goodwin, 2016). The outcome is read with a mean length of 800 bases, although may be extended until 1000 bases. While fully automated implementations of this approach were the mainstay for the original sequencing of the human genome, their main limitation was the small amounts of DNA that could be processed per unit time, referred to as throughput, as well as high cost, taking roughly 10 years and three billion dollars to sequence the first human genome (Schadt, Turner, & Kasarskis, 2010).

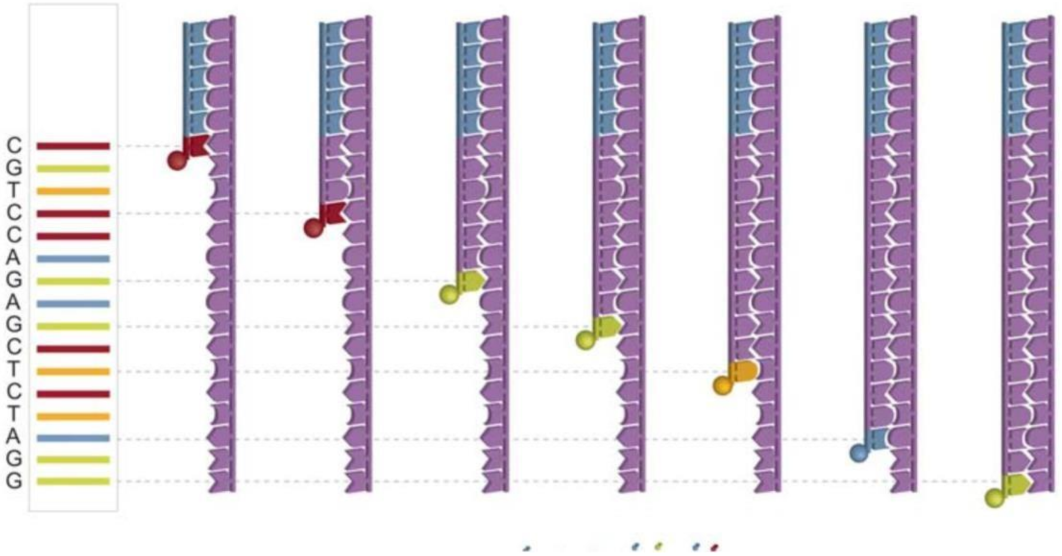


Figure 8 - A modern implementation of Sanger sequencing is shown to illustrate differential labeling and use of terminator chemistry followed by size separation to resolve the sequence. Adapted from (Schadt et al., 2010).

1.7.2 Massively Parallel Sequencing or Second Generation Sequencing Platforms

Next-generation sequencing (NGS), or high-throughput sequencing, is the term used to describe a number of different sequencing technologies including, Roche 454 sequencing, Ion torrent: Proton / PGM sequencing, SOLiD sequencing, and Illumina (Solexa) sequencing (José L. Oliver, 2012). A commonality of Next-Generation Sequencing methods is the simplified workflow used to prepare genes for sequencing. Library preparation includes: fragmenting the DNA (through sonification, enzymatic cleavage, or any other method); ligation of an adapter sequence, barcode and primer and size selection of the fragments. Previous methods relied on capillary electrophoresis, which could only read up to 96 wells at a time. NGS's massively parallel sequencing allowed for innumerous of reads to run simultaneously, although most reads come out as short, unless additional techniques such as mate-pair sequencing are used.

There are currently numerous NGS platforms available (Metzker, 2010). The most well-known platforms include an array-based pyrosequencing approach, such as 454 sequencing (Balzer, Malde, Lanzn, Sharma, & Jonassen, 2011), a sequencing-by synthesis method called Illumina sequencing (Bentley, 2006), and sequencing-by-ligation method named SOLiD sequencing (Valouev *et al.*, 2008). New sequencing technologies are being developed continuously.

With the advent of PCR and its variations, no longer DNA fragments were transformed into bacterial cells to replicate DNA. Instead, NGS techniques use two different types of PCR: emulsion PCR (or emPCR) and bridge/cluster PCR (Teng & Xiao, 2009) (figure 9).

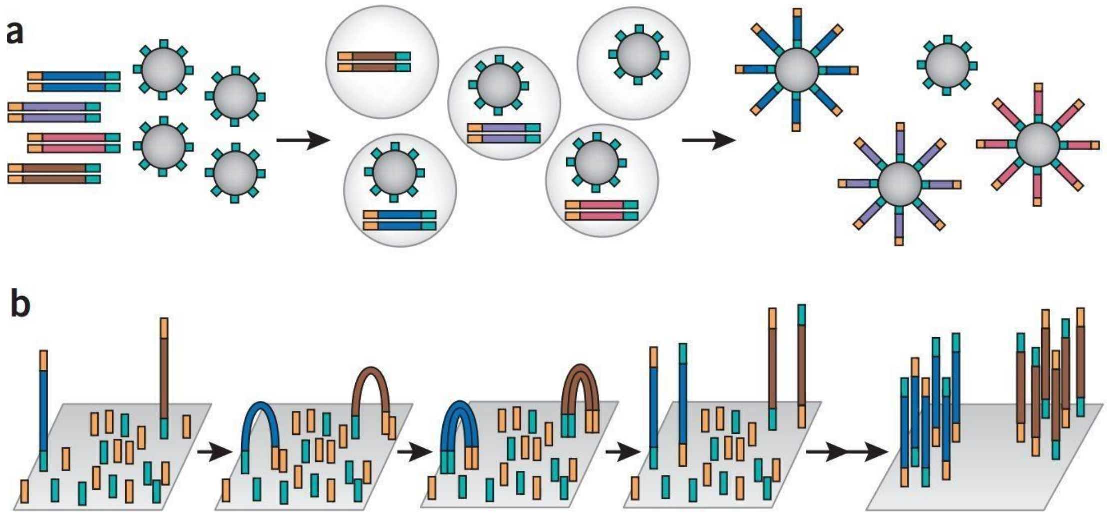
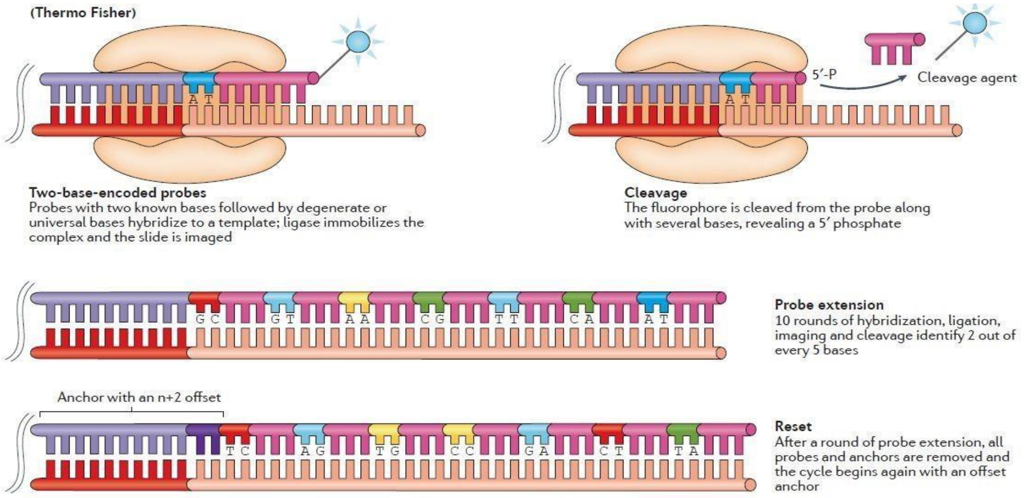


Figure 9 – "The 454, and SOLiD platforms rely on emulsion PCR to amplify clonal sequencing features. In brief, an in vitro–constructed adaptor flanked shotgun library (shown as gold and turquoise adaptors flanking unique inserts) is PCR amplified (that is, multi-template PCR, not multiplex PCR, as only a single primer pair is used, corresponding to the gold and turquoise adaptors) in the context of a water-in-oil emulsion. One of the PCR primers is tethered to the surface (5′- attached) of micron-scale beads that are also included in the reaction. A low template concentration results in most bead-containing compartments having either zero or one template molecule present. In productive emulsion compartments (where both a bead and template molecule is present), PCR amplicons are captured to the surface of the bead. After breaking the emulsion, beads bearing amplification products can be selectively enriched. Each clonally amplified bead will bear on its surface PCR products corresponding to amplification of a single molecule from the template library. The Solexa technology relies on bridge PCR (aka 'cluster PCR') to amplify clonal sequencing features. In brief, an in vitro–constructed adaptor-flanked shotgun library is PCR amplified, but both primers densely coat the surface of a solid substrate, attached at their 5′ ends by a flexible linker. As a consequence, amplification products originating from any given member of the template library remain locally tethered near the point of origin. At the conclusion of the PCR, each clonal cluster contains ~1,000 copies of a single member of the template library. Accurate measurement of the concentration of the template library is critical to maximize the cluster density while simultaneously avoiding overcrowding." Adapted from Next-generation DNA sequencing, Jay Shendure1 & Hanlee Ji2).

SOLiD Helicos

The SOLiD platform uses a sequence by ligation approach, in which libraries of fragments are clonally amplified on the surface of a 1micron bead and an oligo complementary to one of the two adapters used in the library construction is covalently bound.

“Clonal amplification is achieved by limiting dilution of the fragment library during emulsion PCR, which is performed as an emulsion generated by mechanical whipping of an aqueous solution containing PCR reagents, amplification beads, the library and oil. Following emulsion PCR ‘loaded’ beads are enriched by hybridization of the alternate adapter to complementary oligos covalently attached to a polystyrene bead” (Hirst & Marra, 2010) (figure 10).



Two-base-encoded probes
Probes with two known bases followed by degenerate or universal bases hybridize to a template; ligase immobilizes the complex and the slide is imaged

Cleavage
The fluorophore is cleaved from the probe along with several bases, revealing a 5′ phosphate

Probe extension
10 rounds of hybridization, ligation, imaging and cleavage identify 2 out of every 5 bases

Reset
After a round of probe extension, all probes and anchors are removed and the cycle begins again with an offset anchor

Figure 10 - SOLiD sequencing. “Following cluster generation or bead deposition onto a slide, fragments are sequenced by ligation, in which a fluorophore-labelled two-base-encoded probe, which is composed of known nucleotides in the first and second positions (dark blue), followed by degenerate or universal bases (pink), is added to the DNA library. The two-base probe is ligated onto an anchor (light purple) that is complementary to an adapter (red), and the slide is imaged to identify the first two bases in each fragment. Unextended strands are capped by unlabeled probes or phosphatase to maintain cycle synchronization. Finally, the terminal degenerate bases and the fluorophore are cleaved off the probe, leaving a 5 bp extended fragment. The process is repeated ten times until two out of every five bases are identified. At this point, the entire strand is reset by removing all of the ligated probes and the process of probe binding, ligation, imaging and cleavage is repeated four times, each with an n + 1, n + 2, n + 3 or n + 4 offset anchor.” Adapted from (Goodwin, McPherson, & McCombie, 2016)

A related system to the SOLiD is the Polonator, based in part on the system developed by J.S. and the Church group 13 at Harvard. This platform also uses sequencing features generated by emulsion PCR and sequencing by ligation. However, this instrument is substantially cheaper than that of other second-generation sequencing instruments. Additionally, “the instrument is open source and programmable, potentially enabling user innovation (e.g., the use of alternative biochemistries).” The current read-lengths, however, may be significantly limiting (Ji, 2008).

Roche 454 Sequencing

Roche 454 uses a pyrosequencing approach. “Pyrosequencing is a non-electrophoretic, bioluminescence method that measures the release of inorganic pyrophosphate by proportionally converting it into visible light using a series of enzymatic reactions. Unlike other sequencing approaches that use modified nucleotides to terminate DNA synthesis, the pyrosequencing method manipulates DNA polymerase by the single addition of a dNTP in limiting amounts. Upon incorporation of the complementary dNTP, DNA polymerase extends the primer and pauses. DNA synthesis is reinitiated following the addition of the next complementary dNTP in the dispensing cycle. The order and intensity of the light peaks are recorded as flowgrams, which reveal the underlying DNA” (Metzker, 2009). As in Illumina, the DNA or RNA is fragmented into shorter reads, in this case up to 1kb. Generic adaptors are added to the ends and these are annealed to beads, one DNA fragment per bead.

“The fragments are then amplified by PCR using adaptor-specific primers. Each bead is then placed in a single well of a slide. So, each well will contain a single bead, covered in many PCR copies of a single sequence. The wells also contain DNA polymerase and sequencing buffers. The slide is flooded with one of the four NTP species. Where this nucleotide is next in the sequence, it is added to the sequence read. If that single base repeats, then more will be added. So if we flood with Guanine bases, and the next in a sequence is G, one G will be added, however if the next part of the sequence is GGGG, then four Gs will be added. The addition of each nucleotide releases a light signal. These locations of signals are detected and used to determine which beads the nucleotides are added to. This NTP mix is washed away.” (http://www.ebi.ac.uk/training/online/course/ebi-next-generation-sequencing-practical-course/what-next-generation-dna-sequencing/454-seque)

The next NTP mix is now added and the process repeated, cycling through the four NTPs (see figure 11). This kind of sequencing generates graphs for each sequence read, showing the signal density for each nucleotide wash.

The sequence can then be determined computationally from the signal density in each wash. All of the sequence reads we get from 454 will have different lengths, because different numbers of bases will be added with each cycle (https://www.ebi.ac.uk/training/online/course/ebi-next-generation-sequencing-practical-course/what-next-generation-dna-sequencing/454-seque).

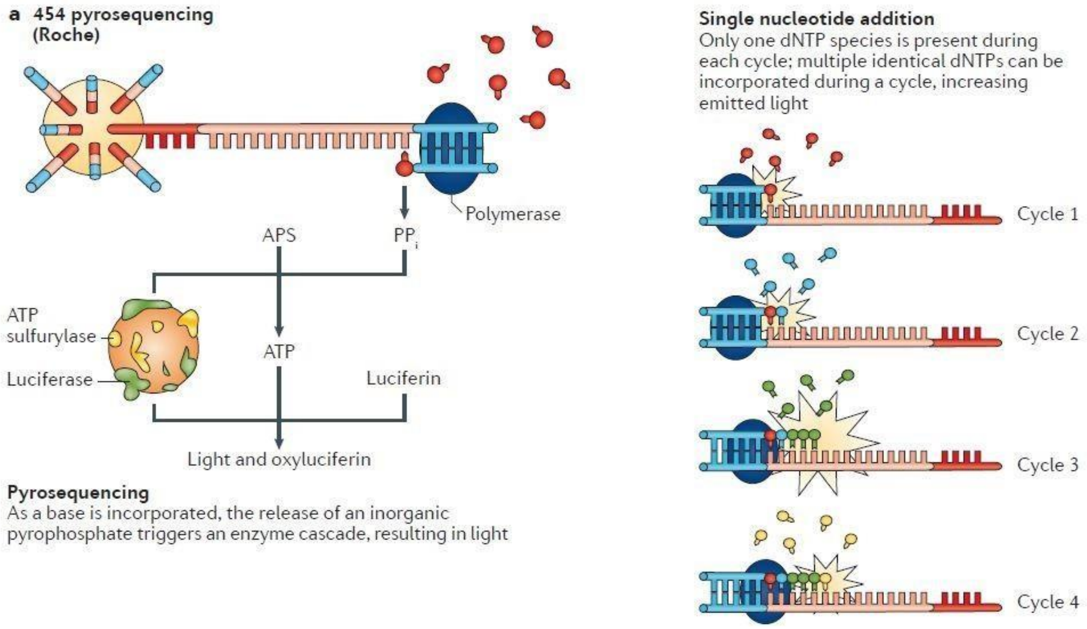

a 454 pyrosequencing (Roche)
Polymerase
APS
PPi
ATP sulfurylase
ATP
Luciferase
Luciferin
Light and oxyluciferin
Pyrosequencing
As a base is incorporated, the release of an inorganic pyrophosphate triggers an enzyme cascade, resulting in light
Single nucleotide addition
Only one dNTP species is present during each cycle; multiple identical dNTPs can be incorporated during a cycle, increasing emitted light
Cycle 1
Cycle 2
Cycle 3
Cycle 4

Figure 11 - Roche 454 (pyro) sequencing. "After bead-based template enrichment, the beads are arrayed onto a microtiter plate along with primers and different beads that contain an enzyme cocktail. During the first cycle, a single nucleotide species is added to the plate and each complementary base is incorporated into a newly synthesized strand by a DNA polymerase. The by-product of this reaction is a pyrophosphate molecule (PPi). The PPi molecule, along with ATP sulfurylase, transforms adenosine 5′ phosphosulfate (APS) into ATP. ATP, in turn, is a cofactor for the conversion of luciferin to oxyluciferin by luciferase, for which the by-product is light. Finally, a pyrase is used to degrade any unincorporated bases and the next base is added to the wells. Each burst of light, detected by a charge-coupled device (CCD) camera, can be attributed to the incorporation of one or more bases at a particular bead". Adapted from (Goodwin, McPherson, & McCombie, 2016)

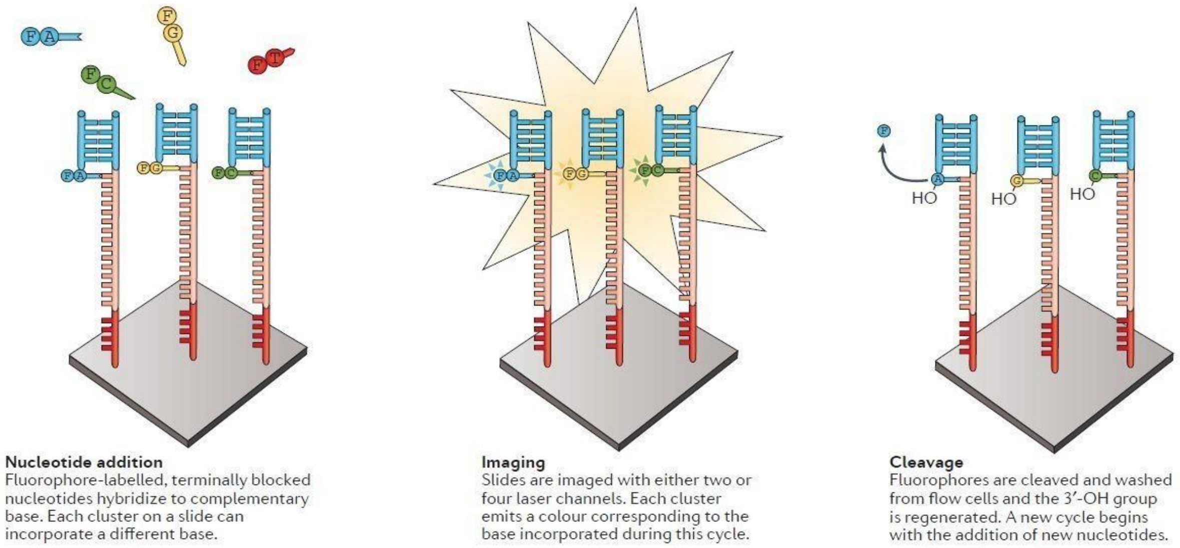
Illumina Genome Analyzer

Illumina sequencing uses a sequence-by-synthesis (SBS) approach, a cyclic method that comprises nucleotide incorporation, fluorescence imaging and cleavage (figure 12). Each read is then cluster-PCR amplified, creating a spot with many copies of the same read. They are then separated into single strands to be sequenced. The slide is flooded with nucleotides and DNA polymerase (Metzker, 2009). In the first step, the DNA polymerase, bound to the primed template, adds or incorporates just one fluorescently modified nucleotide, which represents the complement of the template base. Following incorporation, the remaining unincorporated nucleotides are washed away. Imaging is then performed to determine the identity of the incorporated nucleotide. This is followed by a cleavage step where the terminating/inhibiting group and fluorescent dye are removed.100-150bp reads are used.

Somewhat longer fragments are ligated to generic adaptors and annealed to a slide using the adaptors. Current sequencing on the Illumina platform often produces data whose quality deteriorates towards later cycles.

Up to a read length of nearly 60-70bp, the quality of these reads is considered excellent (> Phred 30). After that, however, Phred scores tend to drop dramatically in a fairly large number of sequences, probably due to wrong nucleotide addition (or “background noise”), which means the rates at which bases are called erroneously increase and need to be processed before alignments (Andrews, 2013).Base call errors in reads can result in alignment mismatch (reduced mapping efficiency), incorrect methylation calls or, in the worst case, misalignments (which will most likely also generate incorrect methylation calls) (https://www.ebi.ac.uk/training/online/course/ebi-next-generation-sequencing-practical- course/what-next-generation-dna-sequencing/illumina-).

With the Illumina genome analyzer, nearly a terabyte of image files is generated during a single run where image files are analyzed and converted into sequence reads. Before any high throughput sequencing experiment, it is recommended to design and test a data analysis pipeline (Kyle R. Pomraning K. M, 2009).



Nucleotide addition
Fluorophore-labelled, terminally blocked nucleotides hybridize to complementary base. Each cluster on a slide can incorporate a different base.

Imaging
Slides are imaged with either two or four laser channels. Each cluster emits a colour corresponding to the base incorporated during this cycle.

Cleavage
Fluorophores are cleaved and washed from flow cells and the 3′-OH group is regenerated. A new cycle begins with the addition of new nucleotides.

Figure 12 - Sequencing by synthesis: cyclic reversible termination approaches. Illumina. “After solid-phase template enrichment, a mixture of primers, DNA polymerase and modified nucleotides are added to the flow cell. Each nucleotide is blocked by a 3′-O-azidomethyl group and is labelled with a base-specific, cleavable fluorophore (F). During each cycle, fragments in each cluster will incorporate just one nucleotide as the blocked 3′ group prevents additional incorporations. After base incorporation, unincorporated bases are washed away and the slide is imaged by total internal reflection fluorescence (TIRF) microscopy using either two or four laser channels; the colour (or the lack or mixing of colours in the two-channel system used by NextSeq) identifies which base was incorporated in each cluster. The dye is then cleaved and the 3′-OH is regenerated with the reducing agent tris(2-carboxyethyl)phosphine (TCEP). The cycle of nucleotide addition, elongation and cleavage can then begin again”. Adapted from (Sara Goodwin, 2016).

Ion Torrent

In contrast to other platforms such as Illumina and 454, Ion torrent and Ion proton sequencing is based in the release of an H+ ion due to the addition of a dNTP to a DNA polymer, although the input DNA or RNA is also fragmented, this time ~200bp. Adaptors are added and one molecule is placed onto a bead. The molecules are amplified on the bead by emulsion PCR, where PCR reagents, primer-coated particles, and a low concentration of template fragments are mixed with oil and emulsified, forming micro reactions. Each bead is placed into an individual nano well of a slide, which is then flooded with a single species of dNTP, along with buffers and polymerase. Nucleotides are added cyclically one at a time, being registered by the unleash of an H+ ion (Salipante et al., 2014). The pH is detected in each of the wells, as each H+ ion released will decrease the pH. The changes in pH allow us to determine if that base, and how many thereof, was added to the sequence read. The dNTPs are washed away, and the process is repeats itself with different dNTP species.

(https://www.ebi.ac.uk/training/online/course/ebi-next-generation-sequencing-practical-course/what-next-generation-dna-sequencing/ion-torre) .

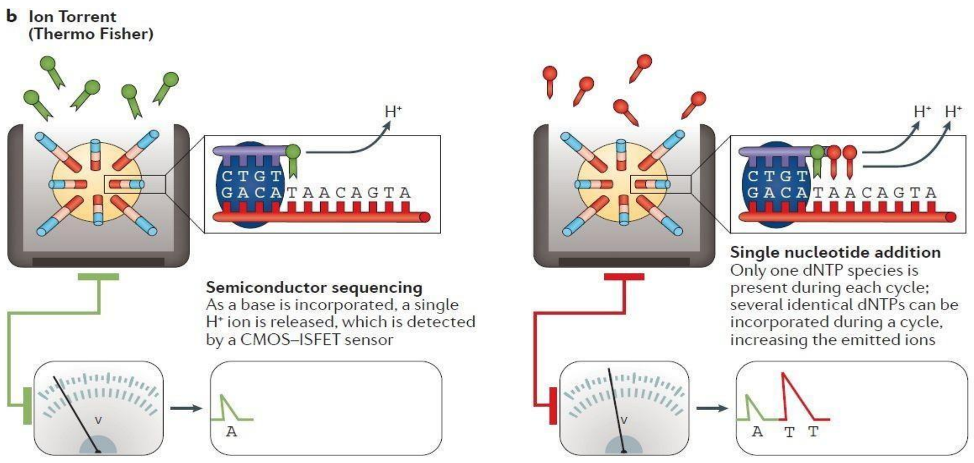

b Ion Torrent
(Thermo Fisher)
H⁺
C T G T
G A C A T A A C A G T A
Semiconductor sequencing
As a base is incorporated, a single H⁺ ion is released, which is detected by a CMOS–ISFET sensor
V
A
H⁺ H⁺
C T G T
G A C A T A A C A G T A
Single nucleotide addition
Only one dNTP species is present during each cycle; several identical dNTPs can be incorporated during a cycle, increasing the emitted ions
V
A T T

Figure 13 - Ion torrent sequencing. "After bead-based template enrichment, beads are carefully arrayed into a microtiter plate where one bead occupies a single reaction well. Nucleotide species are added to the wells one at a time and a standard elongation reaction is performed. As each base is incorporated, a single H+ ion is generated as a by-product. The H+ release results in a 0.02 unit change in pH, detected by an integrated complementary metal-oxide semiconductor (CMOS) and an ion-sensitive field-effect transistor (ISFET) device. After the introduction of a single nucleotide species, the unincorporated bases are washed away and the next is added." Adapted from (Goodwin *et al.*, 2016).

1.7.3 Third Generation of Sequencing Platforms

The third generation of sequencing platforms are distinct from their forbearers, in that they are designed to sequence DNA at the level of a single molecule. The advantages of such an approach include simple library preparation, massively parallel sequencing at long read lengths (which is very useful to detect repetitive elements) and, importantly, the lack of the repeated PCR amplifications before sequencing (Hirst & Marra, 2010). The emerging third-generation technologies are the PacBio, Life Technologies and Oxford Nanopore platforms.

1.8 FASTQ FILES FORMAT

When high-throughput sequencing instruments began generating millions of reads per run, there was a demand for a way to check the quality of each base call. In order to represent both the sequence and the probability of each nucleotide to be well sequenced, FASTQ format was invented. The "Q" suggests quality, as in the quality of the read. In this format, the quality scores are represented by ASCII characters, instead of a sequence of numbers. This type of coding is more efficient because it only requires 1 byte, instead of the generally 3 bytes required. It is shown below an example of a FASTQ file (http://binf.snipcademy.com/lessons/sequence- file-formats/fastq).

```
@SRR478995.2 HWI-ST565_0122:5:1101:1475:2117/1

TTTTAAGAAGTTTTTGAGTTTGTTTTTATTAGTATTTATTTTATAGAAAGATATTTTTTTGTGGTTTTGGGGTT
TTTTTTTTGTACTAAGGTTTTTTTAG

+

?@CFA;D=ADDHHIIICGEHIIHGIGIGHIIIIFHIIGHIGGFGCGDGHFDGC*BFHII''..(.5C(?/539?BB(39&)&&(
4(:(:4(:@ABDDD##
```

Figure 14 - The FASTQ file format. The first line gives the name of the read and the number of pair at the end (in this case pair one). The second and third lines show the nucleotide sequence, which is a, c, g, or t. Then there's a plus. Finally, the fifth and sixth lines give the quality (ASCII code) of each nucleotide from the second and third lines and this code can be looked up/searched to see for what these qualities are.

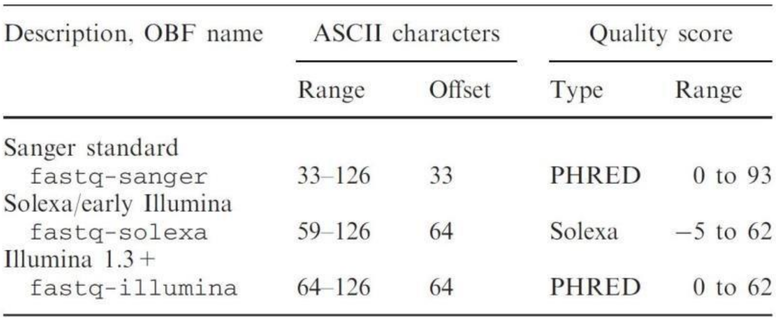
In order to be intelligible and easily edited, ASCII printable characters were restricted to 32–126 (decimal), and since ASCII 32 is the space character, Sanger FASTQ files use ASCII 33–126 characters to encode PHRED qualities from 0 to 93 (i.e. PHRED scores with an ASCII offset of 33. The quality score is the probability that a base is incorrectly identified, and is often parameterized by bioinformatics tools such that reads beyond a threshold of inaccuracy are discarded from the fastq file.

The quality score, or PHRED quality score, was named after the PHRED software which reads DNA sequencing traces, calls bases and assigns each a quality value. This gives a very broad range of error probabilities, from 1.0 (a wrong base) through to 109.3 (an extremely accurate read), so the Sanger FASTQ format is convenient both for raw sequencing reads and post-processed assemblies where higher qualities scores are read. ENA/EMBL ArrayExpress provides FASTQ files so that they don't require uncompressing but take longer to download as they are larger files. Unlike Arrayexpress, NCBI GEO provides raw data as SRA files, which are compressed versions of the FASTQ files which means that paired reads come in two the two FASTQ files for the two paired end reads. So there is a trade off when choosing between these options

In 2004, Solexa, Inc. introduced their own version of the FASTQ format. Although the FASTQ format only stores a single quality score per letter, Solexa also produced quality scores for all four bases, and in order to represent low-quality information more accurately, an alternative logarithmic scale was used (Cock, Fields, Goto, Heuer, & Rice, 2009).

Although Illumina initially continued to use the Solexa FASTQ variant, from Genome Analyzer Pipeline version 1.3 onwards, PHRED quality scores rather than Solexa scores were used. "However, rather than adopt the original Sanger format, Illumina introduced a third incompatible FASTQ variant designed to be interchangeable with their earlier 'Solexa FASTQ' files for good quality reads. The Illumina 1.3+ FASTQ variant encodes PHRED scores with an ASCII offset of 64, and so can hold PHRED scores from 0 to 62 (ASCII 64–126), although currently raw Illumina data quality scores are only expected in the range 0–40."(Cock, Fields, Goto, Heuer, & Rice, 2009)

Illumina could have adopted the original Sanger format and, consequently, could have unified the FASTQ format. Although Illumina 1.3+ FASTQ variant is interchangeable with the earlier Solexa FASTQ version for high quality reads, there are currently three incompatible FASTQ variants (figure 17).



Description, OBF name	ASCII characters		Quality score	
	Range	Offset	Type	Range
Sanger standard `fastq-sanger`	33–126	33	PHRED	0 to 93
Solexa/early Illumina `fastq-solexa`	59–126	64	Solexa	−5 to 62
Illumina 1.3+ `fastq-illumina`	64–126	64	PHRED	0 to 62

Figure 15 - The three described FASTQ variants. In columns there is the description, format name used in Open Bioinformatics foundation (which is a non-profit, volunteer-run group dedicated to promoting the practice and philosophy of Open Source software development and Open Science within the biological research community. projects, range of ASCII characters permitted in the quality string (in decimal notation), ASCII encoding offset, type of quality score encoded and the possible range of scores. Adapted from (Cock et al., 2009).

2.OBJECTIVES

The primary goals of this thesis are:

1-To know and make use of the main biological databases with relevant data for epigenomics;

2-To obtain fastq files from methyl-seq and/or bisulfite-seq next generation sequencing from public databases (ENA, GEO or DDBJ), of healthy individuals with their respective metadata attached giving information about the age, gender, tissue/organ, health status

3-To use the current bioinformatic protocol from the laboratory to map methylations, using the files obtained previously. Adapt, if necessary, the protocol for any other possible modifications

4-To do a comparative analysis with the results obtained previously in order to identify any possible features related to healthy human ageing

5-To do a comparative analysis using brain and blood samples from healthy individuals with samples from age matched individuals in order to establish possible markers for human healthy aging

6-To further understand the of the Epigenome in a healthylifestyle.

3. METHODOLOGIES

3.1 PRIMITIVE METHODS AND TECHNOLOGY OF DNA MEHTYLATION QUANTIFICATION

DNA methylation research can be approached in several ways because there is a wide range of techniques available for the study of the occurrence and localization of methyl cytosine in the genome.

The focus of a study might be the methylation status of a gene of interest (locus-specific study) or of a large number of genes (genome-wide study), or the total DNA methylation content in a cell or tissue under normal or pathological conditions (global study). The techniques used for qualitative analysis of DNA methylation yield information about the methylation status of a gene or comparative information in paired samples. Quantitative methylation provides information about particular genes or, a particular CpG in the region of interest. (Ester Lara V. C., 2011).

The first techniques used to explore the epigenetic patterns were based on the separation of methylated and unmethylated deoxynucleotides. Quantification of methylcytosine in the genomic DNA can be done by high-performance separation techniques or by enzymatic/chemical means. The most significant technique at the time was the separation of purines and pyrimidines through paper chromatography, based on the quantitative hydrolysis of DNA using DNase I and nuclease P1, followed by treatment with alkaline phosphatase. The individual bases can then be monitored based on their UV absorbance at 254 and 280 nm. The Reverse Phase-High Performance Liquid Cromatography (RP-HPLC) method was further improved throughout the 1980s with incorporation of mass spectrometry with standard High Performance Liquid Chromatography (HPLC) (Alan Harrison, 2011).

Because, HPLC based methods demand specialized machinery, alternative separation techniques were developed. Bestor et al (1984) used two restriction endonucleases, Msp1 and Taq1 to distinuish between methylated and unmethylated-CpG residues in their restriction sites, CCGG and TCGA respectively. Digested DNA is first labeled with a 32P isotope in the 5′ end- and then hydrolyzed to deoxyribonucleotide monophosphate followed by separation in two dimensions via thin-layer chromatography (TLC).

Quantitative analysis of DNA methylation is measured by the ratio between C and 5mC fractions after separation. In the former, DNA can be digested into single nucleotides and total genomic 5-methylcytosine, and can be quantified by high- performance liquid chromatography thin-layer chromatography, liquid chromatography/mass spectroscopy or high performance capillary electrophoresis, which is the best choice because is faster, cheaper, and more sensitive than chromatography based techniques (Fraga M. F., 2002).

The enzymatic/means are never as accurate as the former, and sometimes their resolution is limited by the endonuclease cleavage sites. However, enzymatic/chemical approaches were still able to replace the earlier methods because, unlike separation techniques, they do not require expensive and complex equipment that is not always available, require minor amounts of DNA and, so, are much more convenient.

An important impulse for epigenetic research was the adoption of DNA microarrays to methylation profiling (Estécio, Yan, Huang, & Issa, 2008) but the generation of whole genome, single-base-pair resolution methylation maps became possible just recently with the upcoming Next-Generation Sequencing (NGS) or High-Throughput Sequencing (HTS) (José L.Oliver, 2012).

Microarray-Based Methylome-Wide Analysis and Platforms

Throughout the 1990s, the development of DNA microarray technology caused a revolution in functional genomics, allowing high-throughput analysis of single nucleotide genomic variants (Southern *et al.*, 1999). Microarrays consist of an a series of packed microscopic spots of DNA or RNA fragments, called features. Three main classes of microarray-based methods have been developed to map 5mC patterns in genomes: methods enriching for highly methylated regions using an antibody specific for 5mC or methyl- binding proteins; methods based upon bisulfite modification; and methods utilizing restriction enzymes for methylated sites (Yu-I Weng, 2009) Microarrays allow to obtain information of the "mean" methylation values of a given region, however the methylation pattern is not revealed at a single base pair resolution (José L. Oliver, 2012).There are some exceptions of arrays for individual CpGs which cover several thousand sites

The regions covered by the arrays can be short sections of DNA used as probes to hybridize to RNA or DNA from sample (called target). “The underlying principle across microarray-based methods is the same: methylated and unmethylated fragments of the genome are split and analyzed. When previously known probes hybridize to a microarray, areas of the genome that are methylated or unmethylated are quantified and identified. This is achieved by fluorescence-based detection of fluorophore-labeled targets thus giving the level of abundance of nucleic acid sequences in the target.” (Parle-McDermott, 2011). In standard microarrays, the probes are attached to a solid surface, which can be a solid surface or a silicon chip, by a covalent bond to a chemical matrix. (Teng & Xiao, 2009).

While powerful, microarrays do have some drawbacks such as the need for a priori knowledge of the genome or genomic features. This directly harms genome annotations when they are incomplete, incorrect, or outdated. Furthermore, metagenomics approaches (where the genetic content from undefined mixtures of organisms in an environment is sampled *en masse*) are difficult due to this restriction of microarrays. Another major obstacle in microarray analysis is cross hybridization between similar sequences. This restricts microarray analysis to the non-repetitive regions of genomes and hampers the analysis of related genes (or features), alternatively spliced transcripts, allelic gene variants, and SNPs (Nelson, 2009). Finally, particularly in those cases based on enrichment of methylated DNA, it can survey for the presence or absence of methylated DNA, but it does not inform about the extent and pattern of CpG methylation in a given region. Often, more studies must be conducted on a single gene basis to confirm the results of such microarray experiments. (Manuscript, 2014).

Illumina® technologies, such as the Illlumina Infinium and GoldenGate beads array, which analyzes bisulfite-converted DNA, have designed bead arrays to analyze the methylation content in different samples simultaneously (Ester Lara V. C., 2011).

3.1.1 High Throughput Sequencing Methods for Cytosine Methylation Mapping

Enrichment-based Methods

MeDIP-Seq

Methylated DNA Immuno-precipitation sequencing (MeDIP-Seq) is, as the name of the technique suggests, based in immunoprecipitation. Similarly to the microarray, the fragmented DNA is enriched based on its methylation content. Antibodies are raised against a single stranded methyl-cytosine and so the immuno-precipitation occurs in a denatured state. To prevent over repetitive content in the subsequent library through preferential annealing of very methylated genomic repeats, library construction is performed before the immuno-precipitation and amplified following enrichment by PCR. Highly specific isolation and enrichment of methylated DNA provides an advantage for the convenient and comprehensive identification of methylation status of normal and diseased cells. The methylated DNA immunoprecipitation protocol uses an antibody specific to methylcytosine in order to capture methylated DNA. An ideal MeDIP assay should be very sensistive and specific, have minimal background and fast high-throughput capability (Kelsie L. Thu, 2009).

MBD-Seq

Methylated DNA Binding Domain sequencing (MBD-seq) shares the same concept of MeDIP- seq, because in both techniques, genomic fragments are enriched based on their methylation content. However, MBD-seq uses recombinant methylated-CpG binding proteins MECP2 or MBD2 to enrich for methylated DNA fragments from a collection of fragmented genomic DNA with 100–300 bp in length. Following enrichment of methylated double stranded DNA fragments, standard techniques are utilized to create a representative library of the methylated fraction of the genome (Marra, 2011).

MRE-Seq

This technique involves parallel digestion with methylation sensitive restriction enzymes (HpaII, AciI, and Hin6I), selection of cutted fragments of approximately 50bp–300bp, pooling the digests, library construction, and sequencing. This interrogates higher CpG density regions because they have many unmethylated recognition sites for these enzymes. Therefore, the coverage of MRE-seq and enrichment methods is notably complementary. (Harris et al.2010*)*.

OVERALL COMPARISON BETWEEN METHODS

MeDIP and bisulphite-based techniques both use one DNA strand, so both are therefore compatible with previously denatured DNA samples.

MeDIP sequencing as well as MBD- isolated DNA sequencing (MBD-seq) can detect differentially methylated regions (DMRs) and capture nearly the same fraction of the methylome. However, “while the proteins used for MBD-based capture strictly bind to methylated CpGs, the antibody used in MeDIP does not discriminate methylated C in the DNA fragments” (Klaas Mensaert, 2014). MeDIP enrichment profiles are preferentially used to distinguish between highly or low methylated CpG dense regions, but not recommended to decipher a methylome on the basis of enrichment data only.

These techniques enrich mainly low CpG density regions, as well as few methylated CpG islands. On the other hand, MRE-seq interrogates more CpG density regions because they have an abundance of unmethylated recognition sites for these enzymes (R. Alan Harris, 2010). Purification-based methods reduce overall cost by limiting the amount of DNA to be sequenced while maintaining a genome-wide approach, however there are also several drawbacks .

First, neither MBD-seq or MeDIP have the base pair resolution of bisulfite sequencing. Second, CpG-density and GC-content affect, the efficiency of affinity purification particularly for Methyl- seq (Bock *et al.*, 2010; Robinson *et al.*, 2011) and subsequent sequencing (Robinson *et al.*, 2011; Benjamini and Speed, 2012). As a consequence, several methylated regions cannot be captured and sequenced by MBD-seq and/or MeDIP-seq, sugesting that the genome-wide character of both methods is limited.

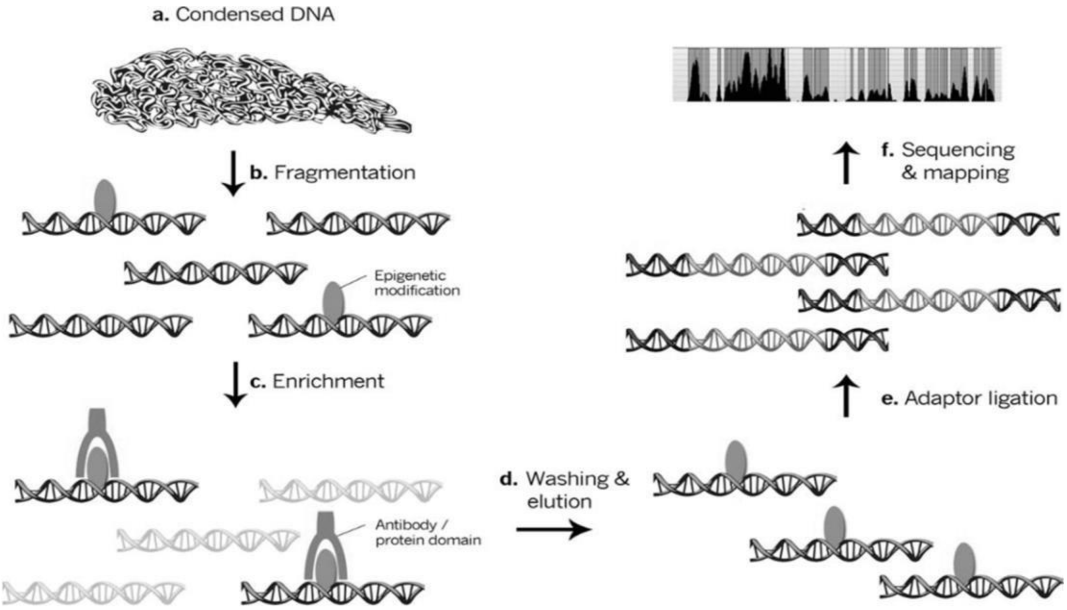


Figure 16 - Enrichment-based sequencing. "After DNA fragmentation (a, b), DNA fragments bearing the specific epigenetic modification of interest are captured using antibodies or specific protein domains. Unbound fragments can be washed away, whereas an elution step is required to obtain the DNA fragments of interest (d). After adaptor ligation and sequencing, sequence reads are aligned to a reference genome to identify the epigenetically modified loci (e, f)". Adapted from (Hirst & Marra, 2010).

Another benefit when using enrichment methods, is the ability to retain every nucleotide, which increases the rate of uniquely mappable sequence reads and allows more genotype-epigenotype correlations. However, enrichment methods do not yield precise quantification of methylation levels. The inability of enrichment methods to quantify methylation was addressed by integrating MeDIP-seq to map methylated regions with MRE-seq to map unmethylated CpG sites (R. Alan Harris, 2010).

3.1.2.1 Bisulfite Based Sequencing Methods

These methods are based on the discovery that, after long incubation with sodium bisulfite, cytosines in single-stranded DNA are deaminated to give uracil. 5-methyl cytosines are immune to this transformation and, therefore, any cytosines presented in bisulfite-treated DNA must have been methylated (Yoshihisa Watanabe, 2010).

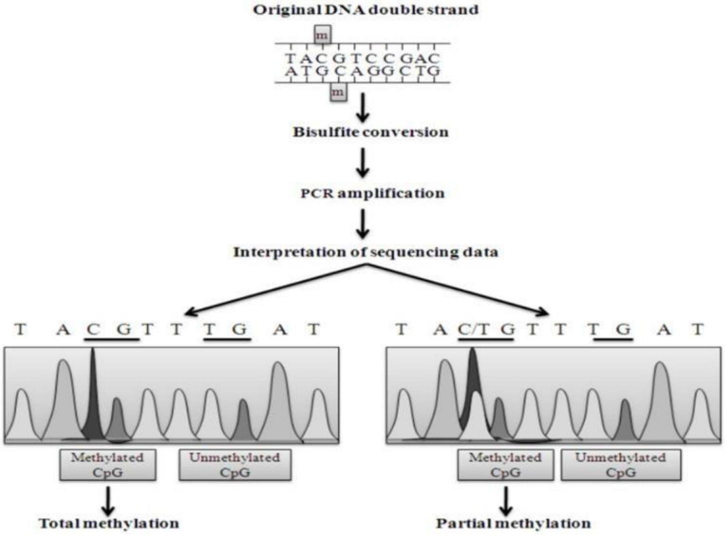


Figure 17 - Principles of bisulfite methods and interpretation of methylation sequencing results. "After fragmentation, bisulfite treatment and PCR amplification, all unmethylated cytosines (C) convert to thymine (T) and the presence of a C-peak indicates the presence of 5mC in the genome. Total methylation or complete conversion of a single residue shows a single peak. The presence of both C- and T-peaks indicates partial methylation or potentially incomplete bisulfite conversion." Adapted from Yuanyuan Li, 2011.

Following the bisulfite treatment, which is performed under denaturing conditions, the library is PCR amplified using PCR primers that extend the adapter sequencing and allowing for clonal amplification and sequencing. The main difference between these methods is the nature of the library, that is, the way how fragments and reads are generated and amplified by PCR, respectively. The majority of bisulfite libraries are directional, which preserves strand specificity; or non directional libraries, which do not preserve strand specificity, so strand identity of a bisulfite read is a priori unknown (Krueger, Kreck, Franke, & Andrews, 2012; Pomraning, Smith, & Freitag, 2009).

Bisulfite-Seq

In this protocol, the fragments are first generated by sonication, then go through end-repair, adapter ligation, and treated with sodium bisulfite (Cheng & Zhu, 2013).

At the end, after two consecutive PCR amplifications, forward and reverse reads, as well as their reverse complementary strands are yielded. These type of library is called non- directional, and so, it does not preserve strand identity. When performed genome-wide this protocol is also called Whole Genome Bisulfite Sequencing (WGBS)

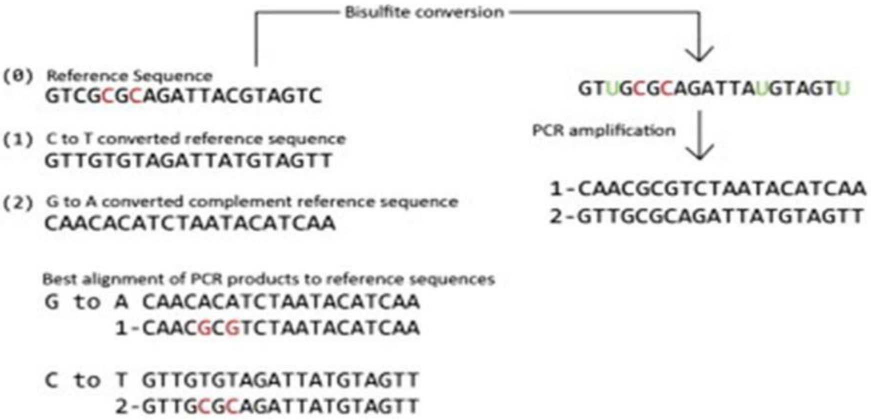


Figure 18 - Analysis of reads from high throughput Bisulfite sequencing. **"To map sequence reads derived from BS-Seq, two additional reference genomes are prepared from a current reference genome (0). The first (1) is the reference genome with all cytosines changed to thymines. The second(2) is the complement sequence to the genome with all guanines changed to adenines. After bisulfite conversion the DNA is subjected to PCR amplification resulting intwo main products from any given sequence. The products are sequenced and aligned to their best hits in either of the two converted reference genomes. C/T mismatches (in the C to T converted reference sequence) and G/A mismatches (in the G to A converted complement reference sequence) indicate the position of a methylated cytosine. Methylated cytosines are red while uracils derived from converted unmethylated cytosines are shown in green". Adapted from (Kyle R. Pomraning K. M., 2009).**

MethylC-seq

As in Bisulfite-seq protocols, fragments are also generated by sonication but in this case the reads come from the forward and reverse bisulfite treated DNA fragments, which means that the library is directionsl. Therefore, all cytosines of the input reads will be converted to thymines and will be tried to align to the C to T converted reference.

RRBS

Reduced representation bisulphite sequencing (RRBs) was introduced to select naturally CpGs enriched regions by size-fractionation of DNA fragments after BglII digestion121 or after MspI digestion (Lee, 2014). Directional libraries, the most wide-spread way of (RR)BS, only ever sequences reads originating from the original top (OT) or original bottom (OB) strands. Often, the library is also size-selected for fragments with length range within 40 and 220bp .This fragment size has been shown to be represented in the sample and yield information on the vast majority of CpG islands (CGIs) in the human or mouse genome. This is followed by end- repair, A-tailing, adapter ligation and bisulfite conversion. Thus, depending on their methylation status, the first three bases of almost all RRBS reads are either CGG or TGG. This applies to reads from both the OT and OB strand, and as nearly all reads in a directional RRBS experiment start with one of these two ways, every read suplies information on at least one CpG right in the start. Finally, amplification by PCR converts every unmethylated cytosine to a thymidine while leaving methylated cytosines intact. The fairly small fragment size of RRBS fragments can become a potential problem especially for sequencing reads with high read length (e.g. > 75bp or >100bp). RRBS provides substantial coverage of CpGs in CGIs, but low CpG coverage genome-wide. (Harris *et al.*, 2011).

Non-directional bisulfite sequencing is less common, but has been performed in several studies (Cokus *et al.*, 2008; Popp *et al.*, 2010; Smallwood *et al.*, 2011; Hansen *et al.*, 2011; Kobayashi *et al.*, 2012).

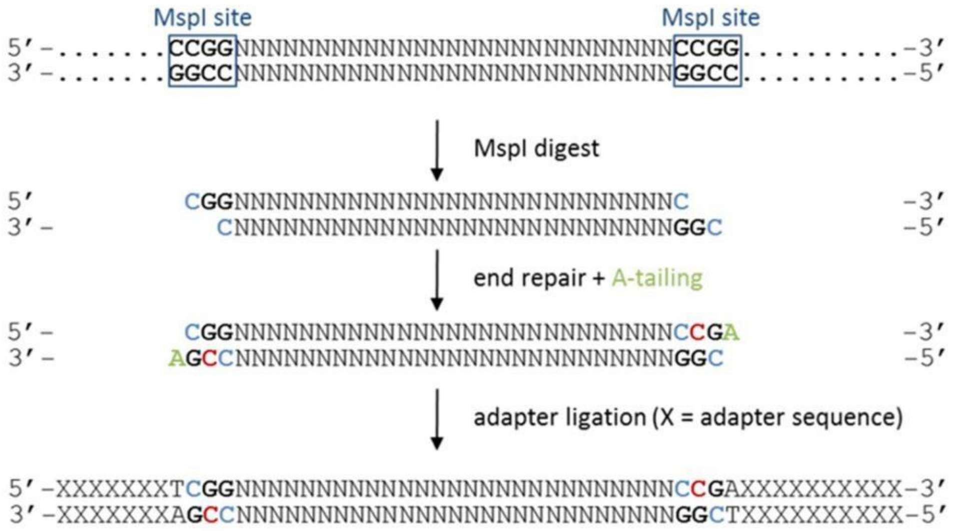


Figure 19 - Preparation of a directional RRBS library. “Cytosines in blue retain the original genomic methylation state, whereas cytosines in red are introduced experimentally during the fragment end-repair reaction (this can be accomplished with either unmethylated or methylated cytosines but the trend seems to be that unmethylated cytosines are being used primarily now)”. Adapted from (Reduced Representation Bisulfite-Seq –A Brief Guide to RRBS, 2013).

In the paired-end library, each end of the DNA fragment is sequenced, resulting in two reads: one coming from one of the original strand and the other coming from the complementary strand, respectively. In these type of libraries, a considerable number of the wrong mappings can be avoided.

Normally, the approximate fragment length distribution is known and therefore a narrow window on the genome can be established to which both reads must map. So, if the two mate reads are independently mapped to the genome and the result of the best alignment yields reads far away from each other, these mappings can be eliminated as at least one of the two alignments will be incorrect (Hackenberg, Barturen, Oliver, Genética, & Ciencias, 2010).

Other than reading into the adapter on the other side, paired-end reads may also generate potentially redundant methylation calls. These need to be discarded if positions are filtered for a certain coverage by independent reads, since overlapping regions are overrepresented, although not twice as much.

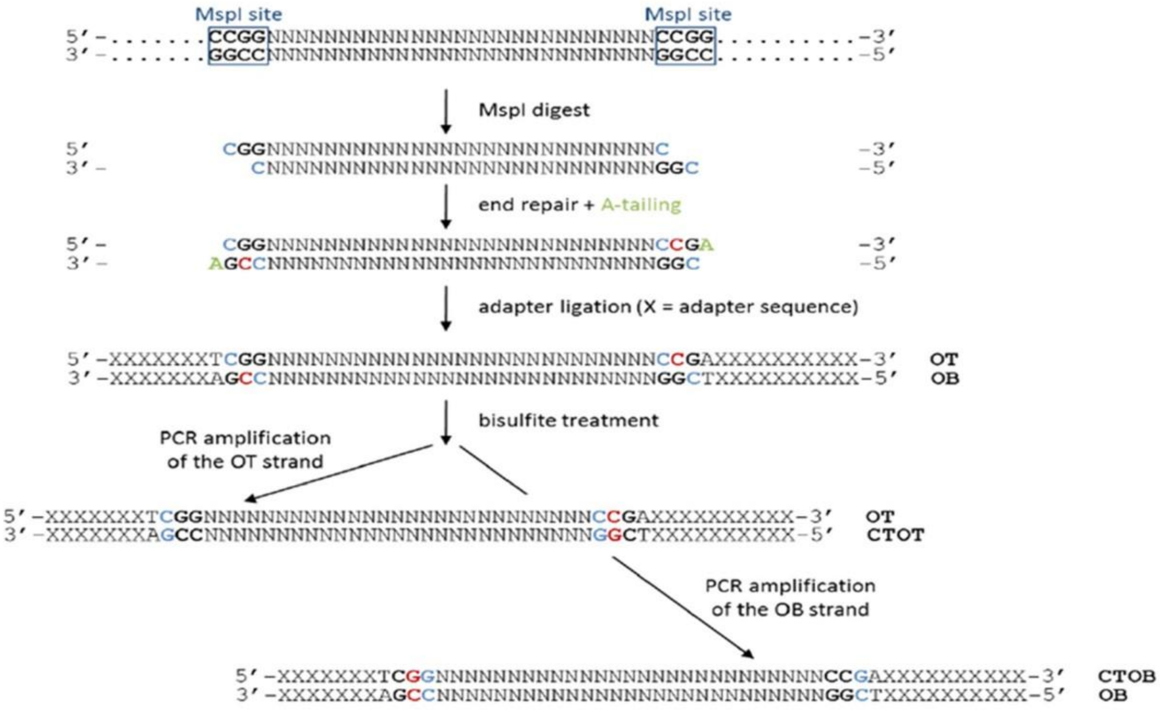

MspI site
MspI site
MspI digest
end repair + A-tailing
adapter ligation (X = adapter sequence)
OT
OB
bisulfite treatment
PCR amplification of the OT strand
OT
CTOT
PCR amplification of the OB strand
CTOB
OB

Figure 20 - Preparation of a non directional paired-end RRBS library. Again, cytosines in blue retain the original genomic methylation state, whereas cytosines in red are introduced experimentally during the fragment end-repair reaction. Although this can be accomplished with either unmethylated or methylated cytosines, the trend seems to be that unmethylated cytosines are being used primarily now. Adapted from (Andrews, 2013).

Many sequencing platforms cannot sequence the DNA directly, so the addition of platform specific adapter sequences to the ends is needed to support the sequencing chemistry reactions. These reactions are designed so that the primer anneals right upstream of the insert to be sequenced and the first sequence to be generated is the desired insert sequence, and not the adapter.

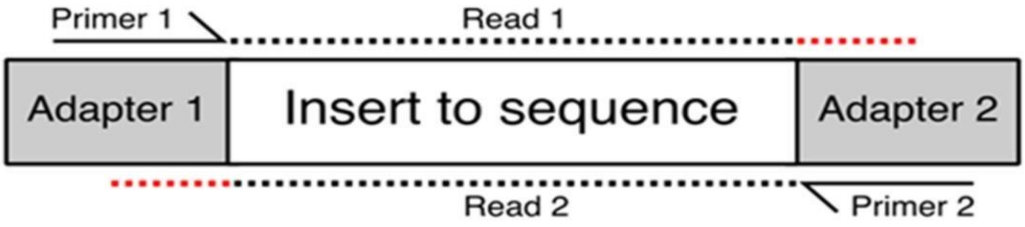


Figure 21 - Paired reads." A linker ligation step coupled to PCR after bisulfite treatment selects only the sequences that have undergone complete cytosine to uracil conversion and, because only 18 PCR cycles are used, this allows for unbiased amplification." Adapted from (Pomraning et al., 2009).

> “After successful bisulfite PCR amplification or sub-cloning procedures, DNA methylation status can be interpreted by further sequencing analysis. Direct sequencing of PCR products may be easily accessible; however, a series of problems limit its application such as failing to read the entire target region and high background interference. Cloning sequencing can provide useful methylation information on a molecular basis. To obtain high confidence in the results, a large number of clones (minimum 5, ideally 10) need to be sequenced, which can be time- and labor-intensive. DNA methylation status can be interpreted by comparing the sequencing results and the original DNA sequence. During bisulfite sequencing the treatment of DNA with sodium bisulfite converts cytosines into uracils, whereas methylcytosines remain unmodified. Uracils are read as thymines by DNA polymerase, so, after PCR amplification, unmethylated cytosines appear as thymines. By comparing the modified DNA with the original sequence, the methylation state of the original DNA can therefore be deduced. Incomplete bisulfite conversion is indicated if both C-and T-peaks appear. The ratio of 5mC to C can be interpreted by analyzing the relative square area of these two bands” (Tollefsbol, 2004) (figure21).

Overview and Comparison between bisulfite based methods

The DNA methylation is erased by PCR and not detected by hybridization, therefore, most techniques rely on a methylation pretreatment of the DNA before hybridization, amplification or sequencing.

The biggest advantage of bisulfite-based methods is to allow quantitative comparisons of methylation levels at single base resolution. Bisulfite-based methods also detect hydroxylmethylation, but they do not differ from methylation. In the original methodology, bisulfite treated genomic regions were amplified by site specific PCR, cloned and sequenced by Sanger’s method. Sequence are assessed one by one and visualized as a matrix with the CpG content of each clone represented as a row.

The most critical issue in BS-Seq experiments in general is to be able to obtain a complete bisulfite conversion of the sample DNA material, while avoiding erroneous conversion of methylated cytosine to uracil is also important. To achieve complete conversion, the two most critical parameters are incubation time and incubation temperature. While maximum bisulfite conversion occurs at either 95°C incubation temperature for a short incubation time of 1 hour, or 55°C for longer incubation of 4-18 hours, these conditions affect DNA stability, and lead to a degradation of 84-96% of available DNA. Other quality issues of Bisulfite-Seq are sensitivity and reproducibility of the method, which are important factors due to the often small amounts of initial original DNA material available for analysis. The amount of unmethylated cytosines which are not converted by bisulfite treatment is known to as ’non-conversion-’ or ’false methylation’ rate, which depends on the completeness of bisulfite conversion. This has been expected to be nearly complete in existing studies, but may be significantly lower for gentler bisulfite conversion treatment.

RRBS revealed to be very useful in the determining the methylation status of discrete genomic regions but it cannot be applied to whole genome studies (K. a Aberg *et al.*, 2014). Notably, RRBS can be applied to a minute amount of input DNA. However, it cannot be used to examine particular regions of interest unless they are adequately flanked by the restriction enzyme sites (Miura & Ito, 2015). In all of these bisulfite PCR- based methods, primer design is the key for successful amplification. Ideally, these shouldn´t contain any CpGs due to CpG density, but if this is not possible, one CpG site can be included at the 5’ end. Such primers must be synthesized as Y (C/T) in the forward strand and R (G/A) in the reverse stand and should incorporate enough cytosines in the original sequence to avoid amplification of unchaned DNA.

The main concern for bisulfite-PCR occurs when methylated and unmethylated DNA molecules sometimes amplify with significantly different efficiencies which can bias the final amplification result (Shen & Waterland, 2007).

Bisulfite treatment of 5-hydroxymethylcytosine (5hmC) provides a similar result as 5-methylcytosine, meaning that bisulfite based methods can be used to detect whether a position is (hydroxy-) methylated but not to determine the exact type of modification, thus indicating that previous studies using bisulfite methods may have been simultaneously examining 5- mC and 5-hmC (Krueger *et al.*, 2012). Based on the rapidly increasing interest in the epigenomic role of 5- hmC, and its chemical similarity to 5-mC, new and specific 5-hmC analysis schemes have been developed.

3.1.2.2 High Throughput Methods That Can Detect 5hmC

Due to the inability of bisulfite sequencing based methods to distinguish between methyl and hydroxymethylcytosines, new techniques have been developed to detect only the latter. Some of these techniques have similar approaches to the ones mentioned previously for microarray based methods, such as enzymatic digestion with (hydroxy) methylation sensitive enzymes (as is the case for RRHP and ABA-seq) (Petterson, Chung, Tan, Sun, & Jia, 2014; Horton *et al.*, 2014) imunoprecepitation detection assays with antibodys (such as HMEDIP) (John P. Thomson, 2013) and binding proteins (JBP) (Skinner *et al.*, 2015). Other methods, such as OXB-seq or TAB-seq, have an oxidative demethylation detection approach, in which hydroxymethylcytosines are oxidized until 5-formylcytosine and 5- carboxylcytosine, respectively, before detection (Yu *et al.*, 2012).

3.1.3 Bisulfite-Seq Dataset Analysis and Workflow

The general process of converting the sequencing data into methylation maps can be divided into 3 steps: pre-processing of the reads, alignment and the profiling of the methylation states from the alignments. Some of these steps are shared by all of the tools, but others, however, are unique to a single or few applications. (Hackenberg *et al.*, 2010). The analysis of methylation from BS-Seq is usually direct, but one should be careful with initial quality control, trimming and suitable alignment of BS-Seq libraries since these are keen to a variety of errors or biases that can be missed with other sequencing applications (Krueger & Andrews, 2012).

3.1.3.1 Pre-Processing of the Reads

The pre-processing of the reads can be separated in two steps: elimination of low quality reads, and preparation of the reads for the alignment step. Because Illumina sequence reads loose quality towards the 3' end and, Lister *et al.* (2009) proposed to trim the read before the first occurrence of a low quality base call (PHRED score <= 2) in order to use the high quality part. Another step which might increase the alignment accuracy is the removal of the artificial adapter sequences because the addition of these adapter sequences on the end of reads in a library will probably cause those reads to fall out of any downstream analysis fairly quickly and introduce large numbers of mismatches into any genomic alignments (José L.Oliver, 2012;*https://sequencing.qcfail.com/articles/read- through-adapters-can-appear-at-the-ends-of-sequencing-reads/*).

When some softwares do a 3-letter alignment, reads must before be manipulated in order to replace the unconverted cytosines by thymines. Since the maximum GC content of (mammalian) BS-Seq libraries ranges from 20 to 30%, the per-base GC-content plot can be another way of spotting contaminating sequences. The GC profile can be increased up to 40-60% due to adaptor contamination, but this can usually be fixed bytrimming the sequence file. It is also recommended to remove: base calls with a Phred score of 20 or lower (assuming Sanger encoding), any signs of the Illumina adapter sequence from the 3' end (AGATCGGAAGAGC), and any sequences that got shorter than 20 bp(Krueger & Andrews, 2012).

3.1.3.2 Alignment

After the treatment of the genomic DNA with sodium bisulfite, the DNA is subjected to PCR amplification and the sequence complexity is reduced as unmethylated cytosines, are converted into thymine. As the methylation state of bisulfite-treated DNA must be inferred by comparison to an unchanged reference sequence, it is crucial to obtain a correct alignment. This is difficult because the aligned sequences do not exactly match the reference due to the conversion, so, usually, only unique alignments are accepted by programs and softwares. In some cases, one and the same read has a unique alignment in a 4-letter presentation but maps to several positions in a 3-letter alphabet, without loosing quality. In these cases, the information carried by the read islost.

Also, as cytosine methylation is not symmetrical, each strand of DNA in the reference genome must be considered separately. A single site can have a different methylation state in different cells, so the percentage of methylation at each site needs to be pre-determined when handling with mixtures of cells or tissues (Pomraning *et al.,* 2009).

Another challenge in bisulfite read mapping is the enlarged search space. This is because the forward and reverse strands of bisulfite treated DNA sequences are not complementary to each other as the bisulfite just acts on cytosines. As a consequence, both bisulfite forward and reverse strands have their own reverse complementary strands (José L. Oliver, 2012).As in the bisulfite-seq, three reference genomes must be used for alignment of bisulfite-converted reads (as shown in figure 20). Methylated cytosines can be identified by a mismatch when a cytosine is aligning to a thymine or a guanine aligning to an adenine and if the position was originally a cytosine or guanine in the original sequence (Pomraning *et al.*, 2009).

“Considering that mapping efficiencies of BS-seq and standard genomic reads converge quickly for read lengths greater than 40 base pairs, single-end reads of 50–75 base pairs seem to offer a reasonable compromise between high quality read sequencing and good mapping efficiency”, at least using Ilumina Genome analyzers (Krueger, 2015).

3.1.3.3 Post-Processing and Output

After the reads have been aligned to the reference genome, each cytosines can be analyzed to assess its methylation status. First, both the reads and the reference sequences need to be converted back to a 4-letter alphabet. Methylated cytosines are then indicated by C/C matches while cytosines are given by a T/C mismatch in the alignment (also G/G and G/A in the case of BS-Seq). The methylation level of a given cytosine position in the genome is given by the number of methylcytosines divided by the total number of reads that map to the position. In this way, the methylation level lies within 0 (completely unmethylated) and 1 (completely methylated). Intermediate methylation levels may be caused when a cell population is used to extract DNA and these values can indicate fluctuations at a given position between the individual cells and allele specific methylation.

Because bisulfite sequencing can assess the methylation level of each individual cytosine can be assessed, not only the methylation levels of CpGs can be determined but also other sequence contexts like CHG or CHH (Lister *et al.*, 2009).

3.1.3.4 Quality Control and Common Sources of Errors

While not completely preventable, a poor sequencing run will have several misinforming sequencing reads such as un-mappable reads, PCR duplicates, low quality reads, adapter dimer or sequencing adapter reads. Several factors are crucial to determine the methylation state of a read from a BS-seq experiment:
First, the sequence of the read must be correct and totally derive from a bisulfite-treated sequence in the original genome.

Second, the read must be properly mapped to the match position of the targeted genome. If any of these criteria fails, the result is the generation of incorrect methylation calls which, in turn, can deviate the conclusions drawn from the whole experiment.

If a base is misaligned or miscalled, the methylation rate will be around 50% because both cytosine and thymine are equally likely to be misplaced against a genomic cytosine (Krueger *et al.,* 2012). Third, both BS-seq alignments and methylation calls assume that the genomic reference sequence that reads are compared to remains unchanged. Thus, if no SNP information is available, extensive systematic errors may occur. Such effects could be minimized considering available genomic-variation data, for example, “from SNP databases into the reference sequence before bisulfite alignments are carried out or by using nucleotide information of the opposing genomic strand” (Krueger *et al.*, 2012).

In real data, the quality of base calls tends to fall as the length of the reads increases. Until 60-70bp, the quality is excellent (> Phred 30). From then on, however, Phred scores tend to decrease significantly in a fairly large number of sequences, which means the rates at which bases are miscalled increase significantly. Base call errors in reads can result in misalignments which will most likely also generate incorrect methylation calls (Andrews, 2013). Particularly in RRBS libraries, artificial methylation calls that arise during the end repair step are a common source of errors. “ As base-call errors are random, the frequency for each base will tend toward 25% each at positions with high error rate. Also, if the read length is longer than the MspI-MspI fragment itself, the sequencing read may continue to read into the adapter sequence on the 3’ end” (Andrews, 2013). Removal of 3’-MspI-sites, as well as low-quality bases of the reads, in RRBS data analysis, is crucial since its methylation state is determined by the cytosine nucleotide used for library preparation (Krueger & Andrews, 2012).

Paired-end RRBS libraries “may contain reads originating from filled-in MspI sites at the beginning of the reads, which consequently need to be excluded from downstream analysis” (Felix Krueger, 2012). Another frequent problem of paired-end alignments seems to be a low mapping efficiency which may result from setting the lower and upper fragment lengths too narrowly. Such stringent settings are not recommended because, quite often, the size-selected fragments turn out to be much smaller or larger than intended (Andrews, 2013).

3.2 USED METHODS AND TECHNOLOGY

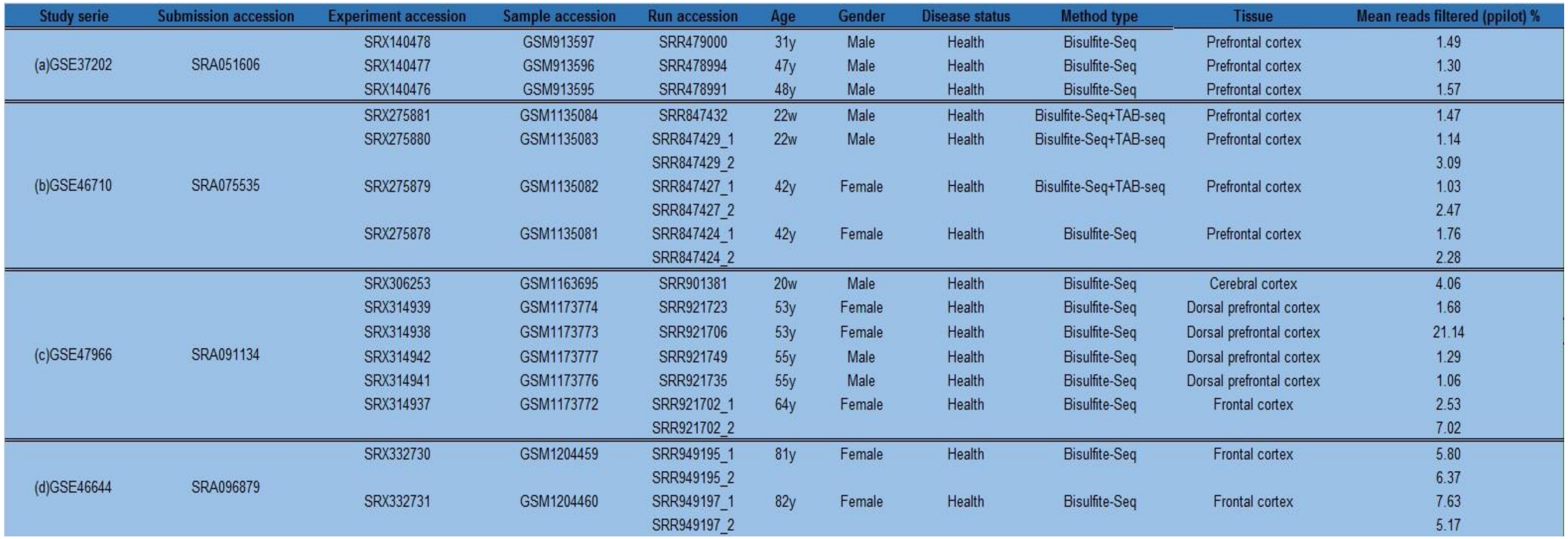
3.2.1 Data Mining: Finding Bisulfite-Seq Data and Metadata

First I searched for SRA files of samples from tissues/organs most abundantly present in the databases and found them to be the brain (Pre Frontal Cortex) and (peripheral) blood which was apparently a good idea because both of them are involved, particularly the brain, in age related diseases. We did so through searching in NCBI and EMBL/ENA and then confirmed the statistic results in DeepBlue Epigenomics.

Through NCBI I searched for epigenetic studies submitted to the SRA archive that were done with humans using "bisulfite-seq" in the first search of Entrez: "strategy bisulfite seq"[Properties] AND "study type epigenetics"[Properties] AND "Homo sapiens [Organism]". Then through ArrayExpress I first searched with the query "Human AND bisulfite AND age AND sex "and then narrowed down the search to "Human AND peripheral blood AND DNA methylation AND bisulfite".

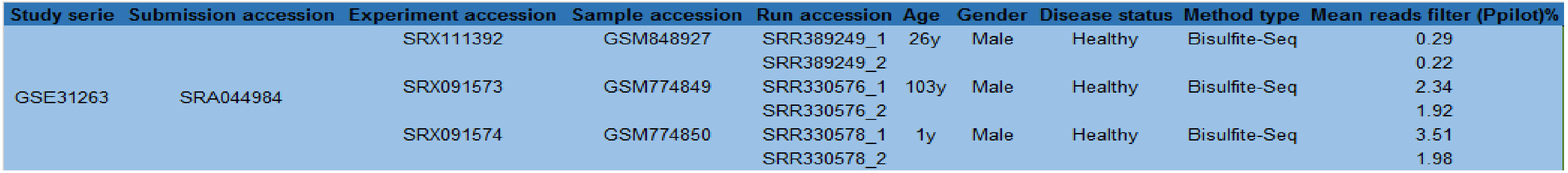
Because most of the samples resulting from the initial search didn't completely match with our query or didn't completely fulfilled the metadata required, we refined our search using Rstudio and the Bioconductor GEO metadb package, which led to the following results (see table 2 and 3).

Table 2 - Brain samples found and downloaded with its respective metadata annotation (age, gender, tissue), accession numbers (study serie, submission accession, experiment accession, sample accession and run accession) and additional information about sequencing method (method type), pre processing parameter (mean filter reads).Paired reads have the same accession numbers, differing only the run accession. Each gse serie number has its own citation number: a) 22922032; b) 24594098; c) 26030523; d) 23925113.



Study serie	Submission accession	Experiment accession	Sample accession	Run accession	Age	Gender	Disease status	Method type	Tissue	Mean reads filtered (ppilot) %
(a)GSE37202	SRA051606	SRX140478	GSM913597	SRR479000	31y	Male	Health	Bisulfite-Seq	Prefrontal cortex	1.49
		SRX140477	GSM913596	SRR478994	47y	Male	Health	Bisulfite-Seq	Prefrontal cortex	1.30
		SRX140476	GSM913595	SRR478991	48y	Male	Health	Bisulfite-Seq	Prefrontal cortex	1.57
(b)GSE46710	SRA075535	SRX275881	GSM1135084	SRR847432	22w	Male	Health	Bisulfite-Seq+TAB-seq	Prefrontal cortex	1.47
		SRX275880	GSM1135083	SRR847429_1	22w	Male	Health	Bisulfite-Seq+TAB-seq	Prefrontal cortex	1.14
				SRR847429_2						3.09
		SRX275879	GSM1135082	SRR847427_1	42y	Female	Health	Bisulfite-Seq+TAB-seq	Prefrontal cortex	1.03
				SRR847427_2						2.47
		SRX275878	GSM1135081	SRR847424_1	42y	Female	Health	Bisulfite-Seq	Prefrontal cortex	1.76
				SRR847424_2						2.28
(c)GSE47966	SRA091134	SRX306253	GSM1163695	SRR901381	20w	Male	Health	Bisulfite-Seq	Cerebral cortex	4.06
		SRX314939	GSM1173774	SRR921723	53y	Female	Health	Bisulfite-Seq	Dorsal prefrontal cortex	1.68
		SRX314938	GSM1173773	SRR921706	53y	Female	Health	Bisulfite-Seq	Dorsal prefrontal cortex	21.14
		SRX314942	GSM1173777	SRR921749	55y	Male	Health	Bisulfite-Seq	Dorsal prefrontal cortex	1.29
		SRX314941	GSM1173776	SRR921735	55y	Male	Health	Bisulfite-Seq	Dorsal prefrontal cortex	1.06
		SRX314937	GSM1173772	SRR921702_1	64y	Female	Health	Bisulfite-Seq	Frontal cortex	2.53
				SRR921702_2						7.02
(d)GSE46644	SRA096879	SRX332730	GSM1204459	SRR949195_1	81y	Female	Health	Bisulfite-Seq	Frontal cortex	5.80
				SRR949195_2						6.37
		SRX332731	GSM1204460	SRR949197_1	82y	Female	Health	Bisulfite-Seq	Frontal cortex	7.63
				SRR949197_2						5.17

Table 3 - peripheral blood samples found and downloaded with its respective metadata annotation (age, gender, tissue), accession numbers (study serie, submission accession, experiment accession, sample accession and run accession) and additional information about sequencing method (method type), pre processing parameter (mean filter reads).Paired reads have the same accession numbers, differing only in the run accession.



Study serie	Submission accession	Experiment accession	Sample accession	Run accession	Age	Gender	Disease status	Method type	Mean reads filter (Ppilot)%
GSE31263	SRA044984	SRX111392	GSM848927	SRR389249_1	26y	Male	Healthy	Bisulfite-Seq	0.29
				SRR389249_2					0.22
		SRX091573	GSM774849	SRR330576_1	103y	Male	Healthy	Bisulfite-Seq	2.34
				SRR330576_2					1.92
		SRX091574	GSM774850	SRR330578_1	1y	Male	Healthy	Bisulfite-Seq	3.51
				SRR330578_2					1.98

3.2.2 PRE PROCESSING READS WITH PIPELINE PILOT

After having found and downloaded every file and its corresponding metadata, we used the pipeline pilot to do the first quality control. “Pipeline Pilot presents a visual working environment for viewing and editing protocols and for running them on a server. Each protocol is comprised by a set of components that perform operations such as data reading, calculation, merging, and filtering. The connections between each components define the sequence in which data is processed. Data from files, databases, and the web is merged, compared, and processed, according to the logic of the protocol. Multiple components can be incorporated into a single component exposed at the outer level of a protocol. The component that comprises these components is known as a "subprotocol". A protocol can include any number of subprotocols. Each one is a complete protocol with its own set of inner components” (Biovia Help Center). The reads were filtered using the homemade script.

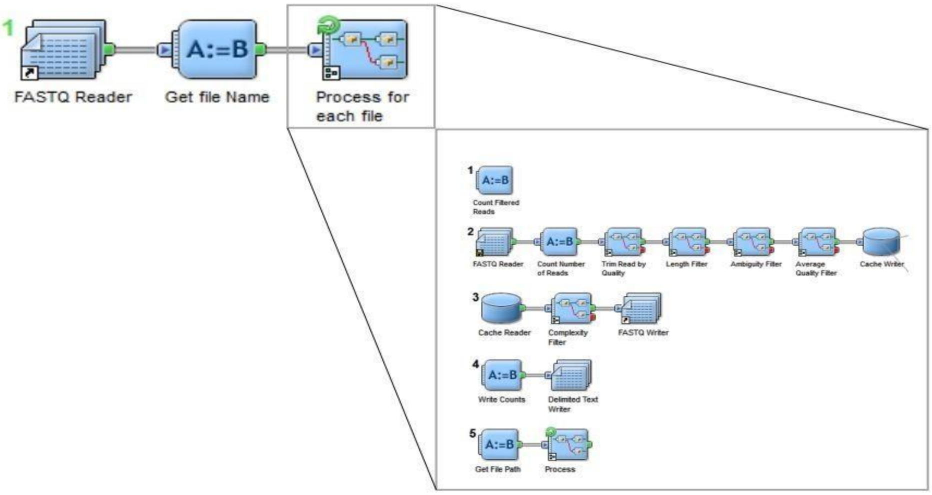


Figure 22 - Pipeline Pilot example of a protocol for pre processing reads. Protocols are made of components (blue boxes) linned among them to indicate the flow of the data and ordered accordingly, each component can integrate subprotocols as exemplified for the last component of the figure.

The protocol used for read filtering includes the following subprotocols:

Trim Read by Quality - Remove fragments with bad quality from the processed read (default threshold = 20 and minimum remaining length =50).

Length Filter - Remove reads too small (default threshold = 50).

Ambiguity Filter - Remove reads with a high percentage of N occurrence (default threshold = 5%).

Average Quality Filter - Remove reads with a low average quality (default quality cutoff threshold = 20).

Filter unpaired reads-Removes those runs that don’t come with both paired reads files

Trim read by length-Removes bp from either 5’end and/or 3’end of the reads. This component was only applied to reads that presented abnormal sequence base content and low quality either at the beginning or the end of the read, respectively.

In the case of paired reads runs, another protocol was necessary (**Manipulate FASTQ paired read files**) to merge both paired reads in a single file so it could be used more accurately as input for Methy Pipe.

3.2.3 CHECKING QUALITY CONTROL WITH FASTQC

After filtering all the reads in Pipeline Pilot, we checked the output reads using FASTQC. FASTQC provides a modular set of analyses, which one can use to quickly tell if whether the data has any problems of which one should be aware before doing any further analysis. Allthouh some reads from our dataset had higher quality than others, all of them fulfilled the quality criteria. As a way to exemplify how main quality issues that can be detected with FASTQC, I will present a series of FASTQC plots from a sample chosen as representative of a globaly good quality of reads.

Quality Control of a filtered single end reads run: SRR921706

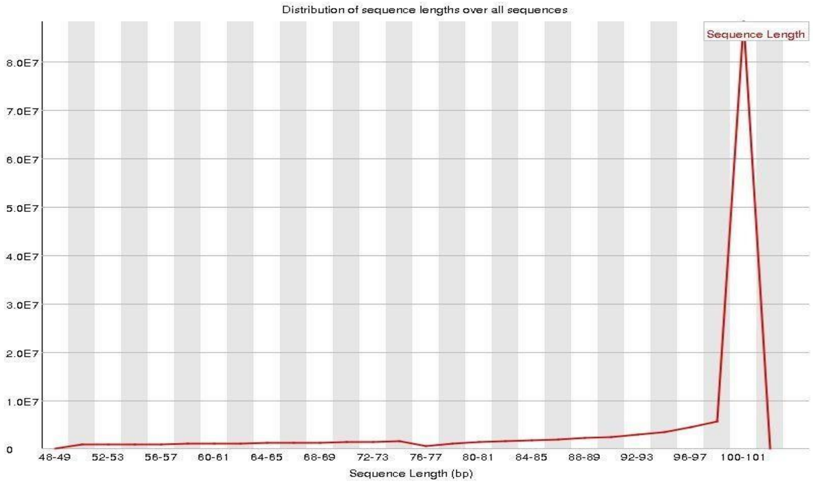


Figure 23 - Distribution of sequence length (in pbs) over all sequences.

Idealy, every sequence should be 100pb long, although in this case some sequences have length lower than 100 bp and the majority have between 99-101 pbs.

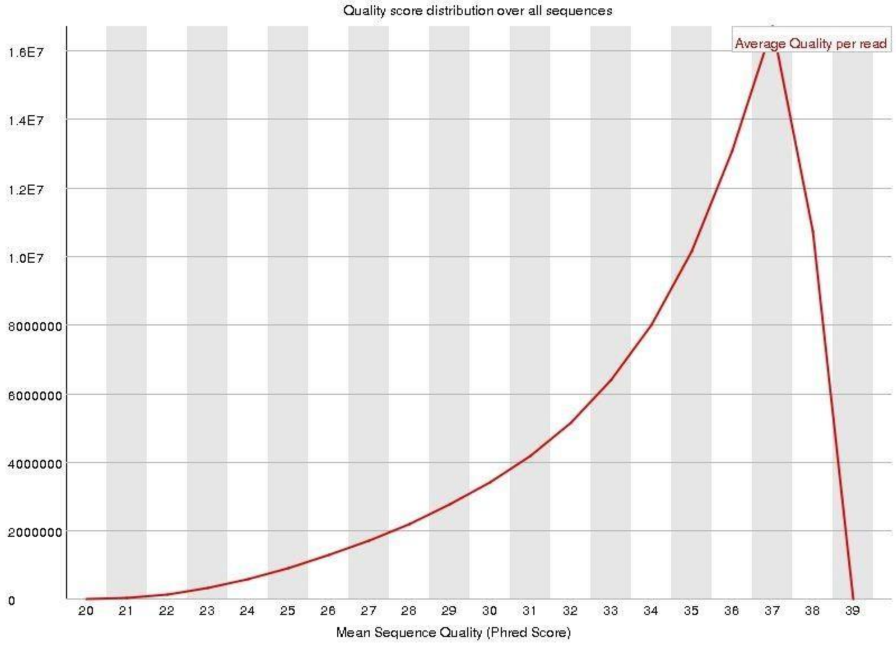


Figure 24 - Mean sequence quality over all sequences.

Most reads have a mean sequence quality Phred score near 37, which is very good.

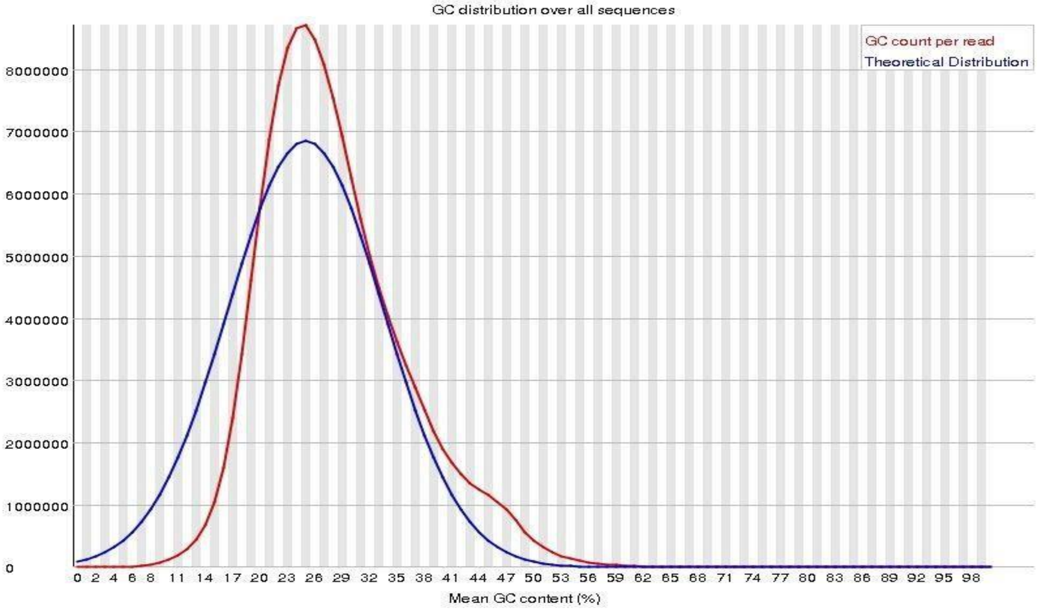


Figure 25 - Mean GC content distribution over all sequences.

The typical GC content of (mammalian) BS-Seq libraries peaks between 20 and 30%. This plot represents the normal distribution on GC content across all sequences, which is near 23%, as expected.

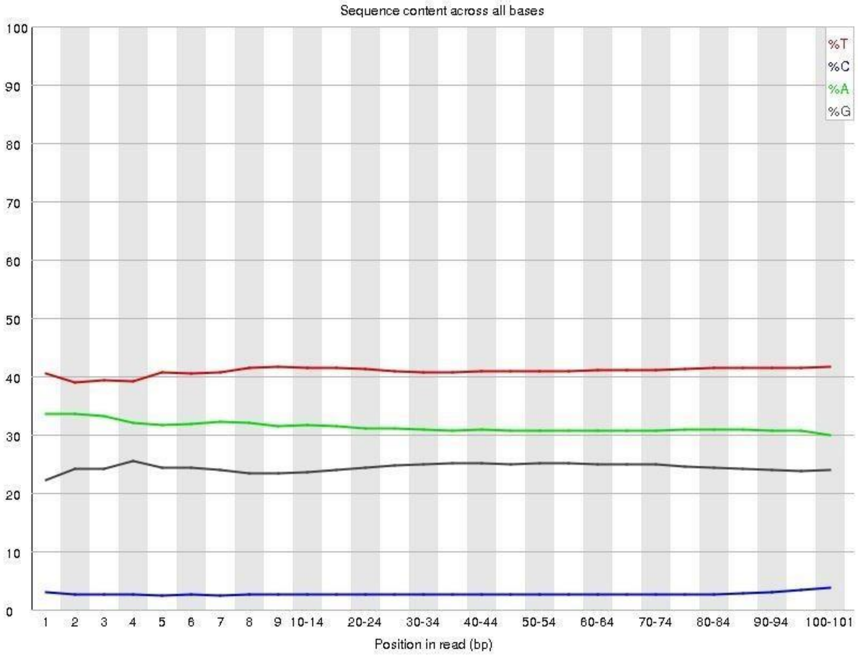


Figure 26 - Mean GC content distribution over all sequences.

Typical BS -seq experiments in mammals tend to have an average cytosine content of ~1-2% throughout the entire sequence length. This may certainly be different for cell types or organisms with different methylation rates, especially in non-CG context.

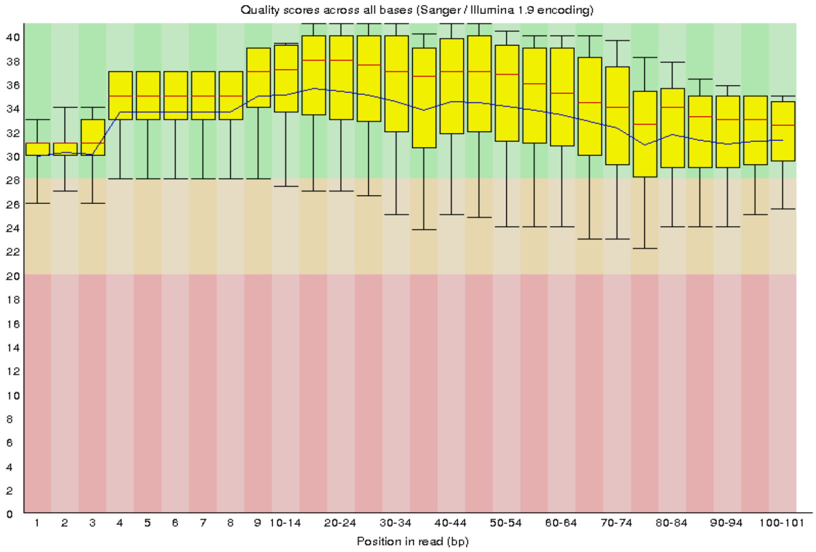


Figure 27 - Quality scores across all bases in every position of the sequence.

If the occurrence of C seems to increase to more than 20% towards later cycles one is almost certainly looking at adapter contamination of variable length. In contrast to other sequencing techniques, the low overall C content of BS-Seq libraries makes adapter contamination easy to spot and remove. Good mean quality scores, as represented here with boxplots, must be over 30.

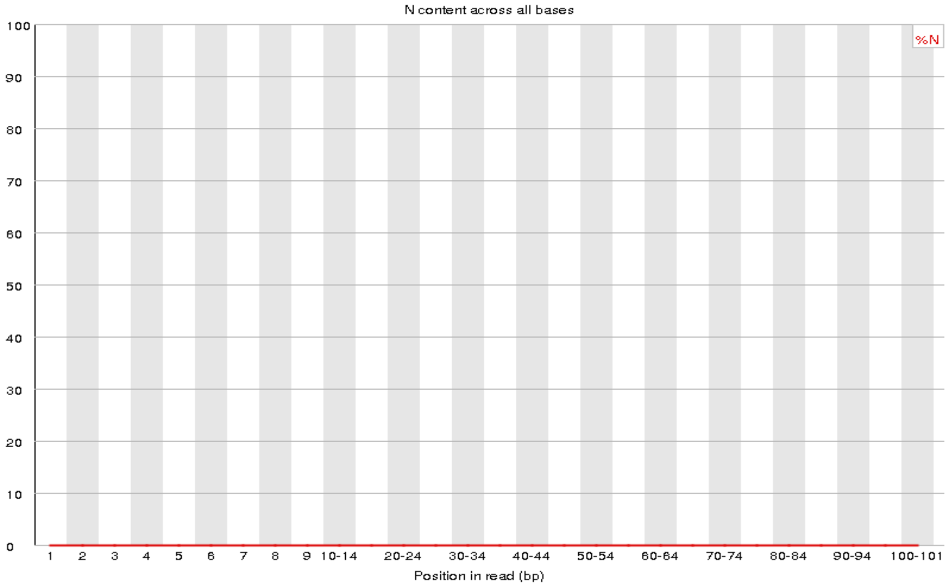


Figure 28 - N content across all bases in every position of read, where N means an undetermined base.

N content plot shows the ideal situation where N content across every base is zero, which means that the base composition in every position was determined.

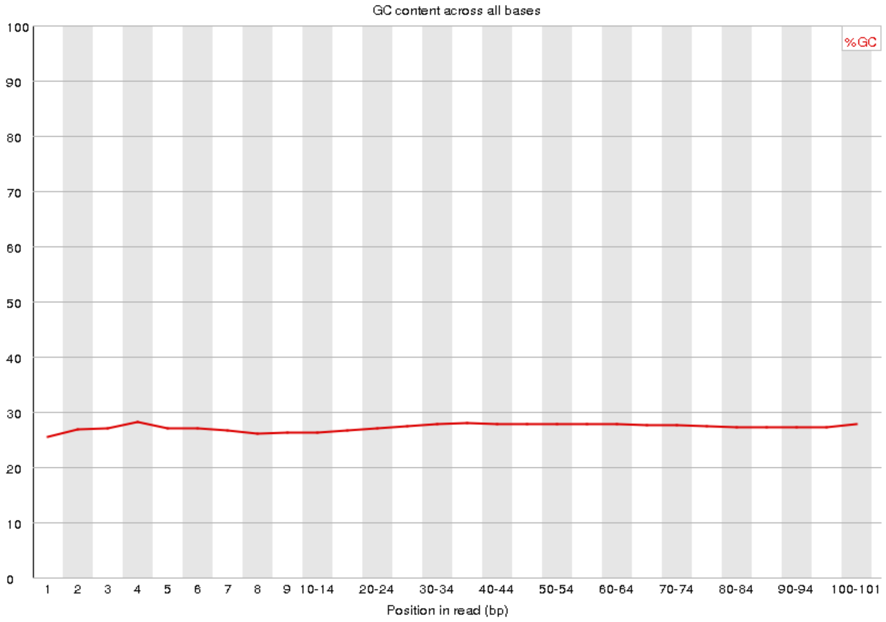


Figure 29 - GC content across all bases in every position of the sequence.

The per-base GC-content plot can be another way of spotting contaminating sequences. Adapter contamination can shift the GC profile to 40-60%, but this can usually be fixed by adapter trimming the sequence file.

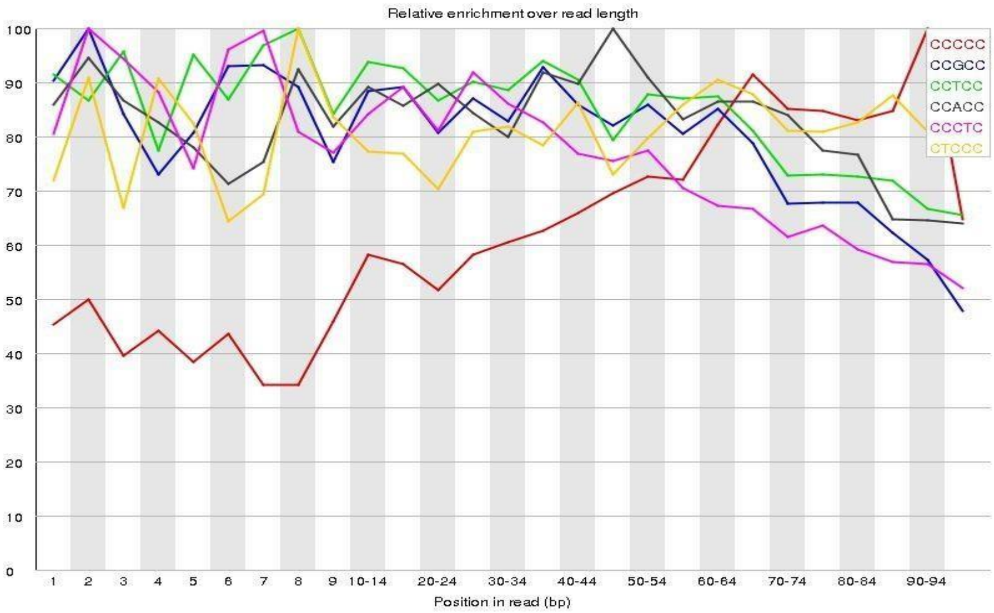


Figure 30 -Kmer plot showing relative enrichment over read length in every position of sequences.

FastQC estimates over expected ratio of individual k-mers by considering the overall frequency of all bases in the library. Once C is usually undererepresented in the sequence file (~1-2%, see), the probability of encountering C containing k-mers is so low that one can easily get high observed/expected ratios from just a few occurrences of C- containing k-mers.

Whereas a prediction of sequence duplication level is 10% for a mammalian shotgun BS-Seq experiment is exected, a level of 80% leads to believe that the sample is highly suffering from PCR duplication which should probably be removed before starting with further analysis. If the over-represented sequence plot contains any sequences it may provide more clues of the potential source of contamination, usually Illumina adapters or primer sequences as a result of primer-dimers.

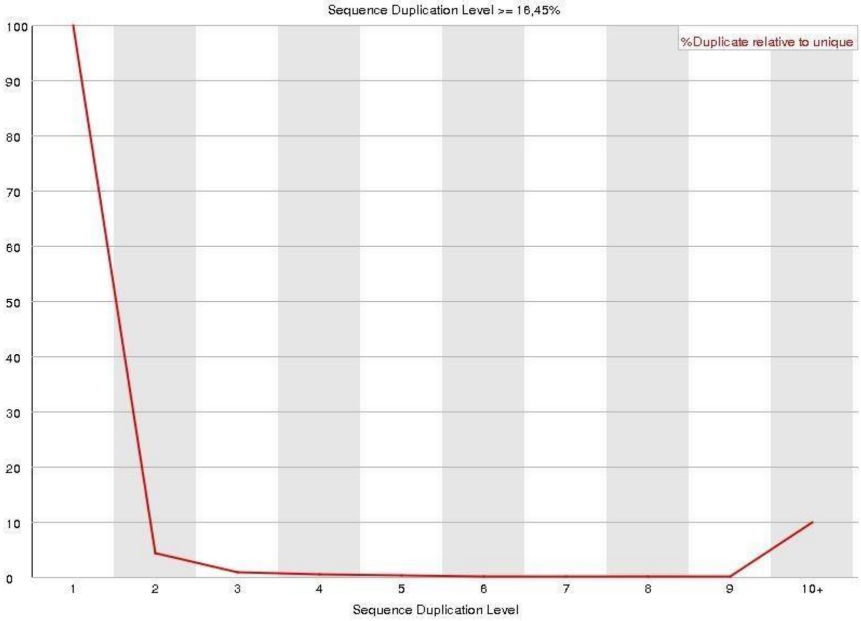


Figure 31 - Sequence duplication levels of read.

3.2.3 ALIGNMENT WITH METHY PIPE

Finaly, for the alignment process and statistic results, we used the software Methypipe. Methypipe is directed towards the analysis of high- or low-resolution DNA methylomes in multiple species, handling (hydroxy-)methyl-cytosines in both CpG and non-CpG sequence context. Methy Pipe is able to perform multiple whole-genome bisulfite sequencing experiments, while maintaining the ability of integrating targeted genomic data. The input data consists of high-throughput bisulfite sequencing reads sequenced from either single or paired end libraries prepared according to the MethylC/Bisulfite-Seq protocol in FASTQ format.

Methypipe Modules and Functions

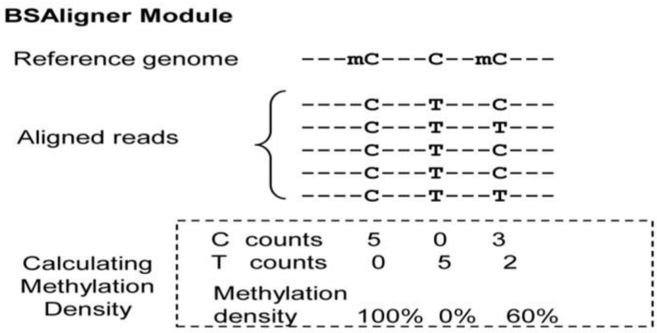


Figure 32 - Scheme of the functions of BSAligner Module.Adapted from (Jiang *et al.*, 2014).

Before the alignment, the reads may need to be pre-processed (if not completely pre processed already with other softwares such as pipeline pilot). First, the low-quality bases (that is, bases with quality score lower than 5) and sequenced adaptors at the 3’ ends of the reads are removed. The preprocessed reads are then mapped to C-to-T converted reference genomes before executing alignment.

Paired-end and single-end reads use different alignment approaches. Single-end reads are mapped to reference genome by allowing at most 2 mismatches and only uniquely mapped reads are kept for further analysis; whereas paired-end reads not only consider the number of mismatches and aligned hits, but also take into account the insert size between the paired-end reads (e.g., from 50 to 600 bases). The ambiguous reads mapped to both forward and reverse strands are removed. Finally, the alignments yielded in a text file that records the aligned chromosomes, positions, mismatches as well as sequencing qualities. To present clear insights of the methylation status of each sample, Methy-Pipe generates genome-wide methylation profiles using Methylation Densitys (MDs, calculated as shown in equation 1) of fixed windows across the whole genome to visualize the MDs in a scatter plot. In these plots, each dot represents a genomic region with a fixed length of 100 kb. The MDs of these fixed windows are plotted against their genomic locations in the reference genome. (see the plots of the section Methylation Density across the chromosomes, in the Appendix).

$$MD = \frac{\sum_1^n C_{(i)}}{\sum_1^n (C_{(i)} + T_{(i)})} * 100\%$$

Figure 33 - Calculation of methylation density formula, where, in a given genomic region” Ci is the number of cytosines and Ti is the number of thymines, in the ith position.n is the total number of cytosines, C(i) is the total number of sequenced cytosines at the ith cytosine position in the reference genome, suggesting the methylated event, and T(i) is the total number of sequenced thymines at the ith position which is suggestive of unmethylated event. When n equals to 1, MD at a single-base resolution could be calculated.”Adapted from Peiyong Jiang, 2014.

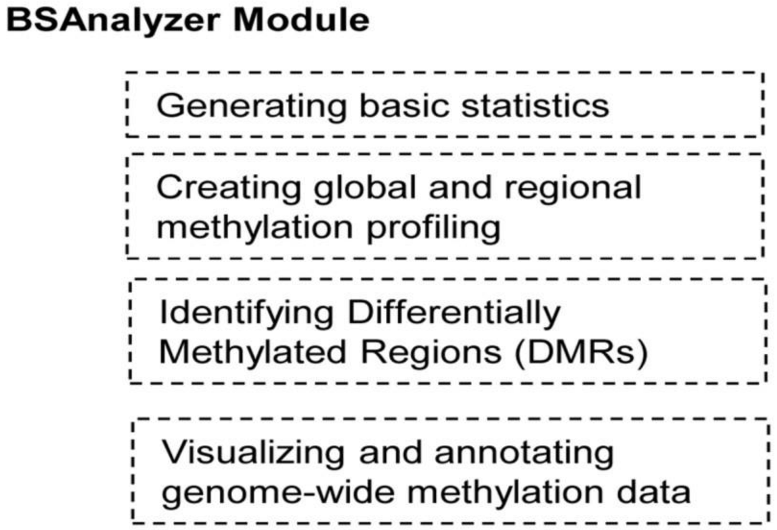

BSAnalyzer Module
Generating basic statistics
Creating global and regional methylation profiling
Identifying Differentially Methylated Regions (DMRs)
Visualizing and annotating genome-wide methylation data

Figure 34 - Scheme of the functions of the BSAnalyzer Module.Adapted from Peiyong Jiang, 2014.

MethtyPipe identifies Differentially methylated regions (DMRs) that have been widely identified among tissues, developmental cells and cancer types as being involved in tissue-, cell- or cancer-specific gene expression. Therefore, the identification and analysis of DMRs for paired or multiple samples is of wide interest (Su *et al.,* 2013).

The identification of genome-wide DMRs between two compared samples is achieved through 4 steps: determination of the seed regions (with a 500 bps extension window from the 5’ part of the read of the two samples); identification of differentially methylated seed regions; extension of differentially methylated seed regions; and merging of adjacent differentially methylated seed regions. (Peiyong Jiang, 2014) (figure 38). As cutoff, we admitted DMRs with p-values (which indicate statistical significance) <0,01.

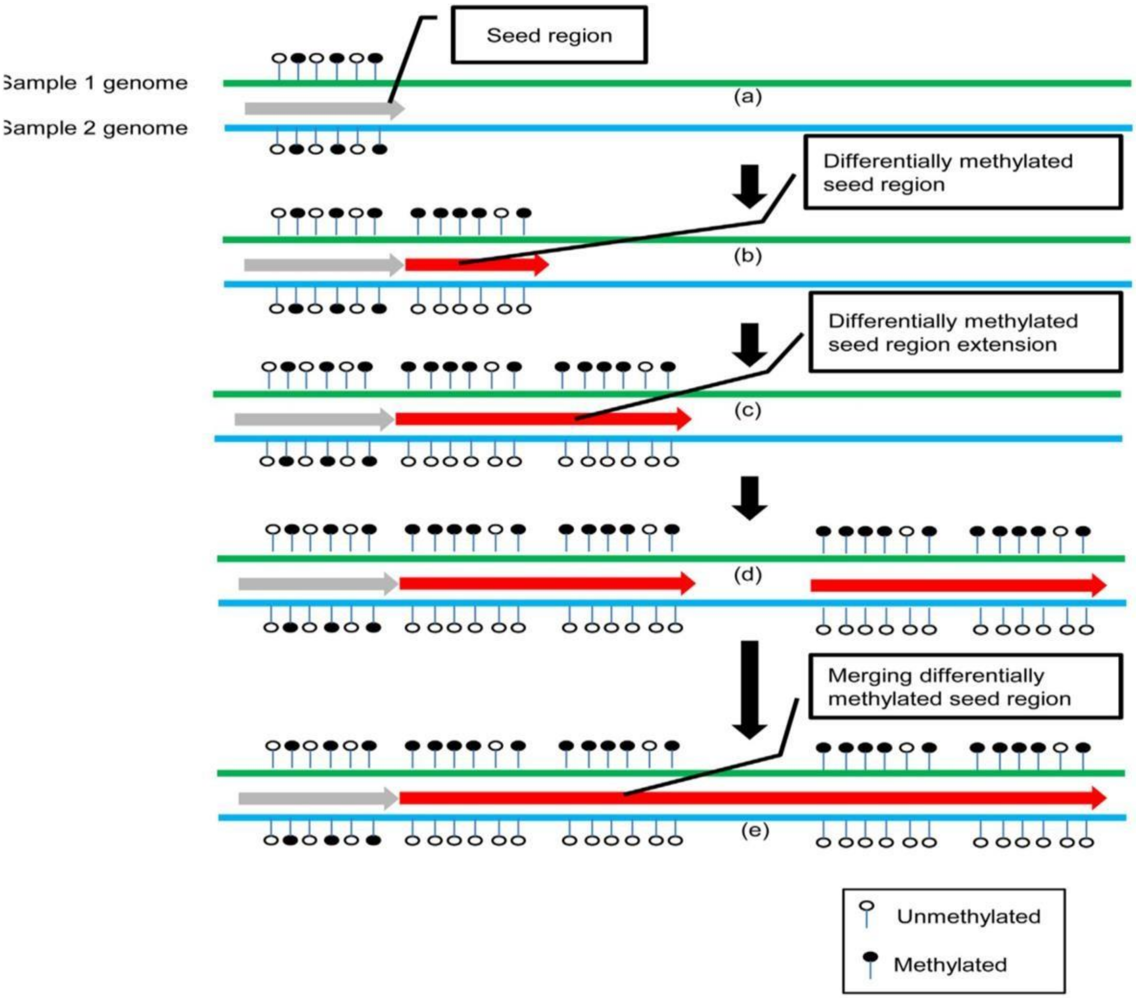

Seed region
Sample 1 genome
Sample 2 genome
(a)
Differentially methylated seed region
(b)
Differentially methylated seed region extension
(c)
(d)
Merging differentially methylated seed region
(e)
Unmethylated
Methylated

Figure 35 - Principle of DMR detection by BSAnalyzer. "(A) Firstly, starting from one end of the genome to search for a seed region (i.e., 500 bps) using a sliding window. (B) If the seed region is located, Mann-Whitney test will be used to test if the seed region is a differentially methylated seed region. (C) Two adjacent differentially methylated seed regions are merged into one extended seed region (seed region extension). (D) Two discontinued differentially methylated regions are further merged together if they are within a certain distance (e.g. less than 1000 bps) for further differential methylation test." Adapted from Peiyong Jiang, 2014.

3.2.4 PRELIMINARY ANALYSIS OF DMRS WITH G:PROFILER

Among all the outputs that Methy-Pipe retrieved, we were mainly interested in obtaining DMRs. Since we had extensive DMRs annotation tables comparing several pairs, a deep and thorough analysis of these genes and their functions would be unfeasible within the scope of this thesis (the most extensive tables were those comparing DMRs between the test file SRR949197 from 82-year-old female brain and control SRR847424 from 42-year-old female brain, which retrieved 810 hypermethylated genes and 2729 hypermehtylated genes; and those comparing the test file SRR330576 from 103 year old male blood with the control file SRR330578 from 1 year old male blood which retrieved 2319 hyper methylated genes an 1212 hypomethylated genes).

Therefore, we began a preliminary analysis of the main functions affected by these genes with an online tool called G:PROFILER. (g:Profiler).This is a public web server for characterizing and manipulating gene lists of high-throughput genomics. It has a simple user- friendly web interface with powerful visualization and is currently available for 80+ species, including mammals, plants, fungi, insects, etc from Ensembl and Ensembl Genomes. The core of the g:Profiler, performs statistical enrichment analysis to provide interpretation to user-provided gene lists, ordered gene lists and chromosomal regions. It studies many sources of functional evidence, such as Gene Ontology terms, biological pathways, regulatory motifs of transcription factors, microRNAs, human disease annotations and protein-protein interactions.

The tool we used, GOST, performs functional profiling of gene lists using various kinds of biological evidence. It also performs statistical enrichment analysis to find over-representation of information such as Gene Ontology terms, regulatory DNA elements, human disease gene annotations, and protein-protein interaction networks. The basic input of g:GOSt is a list of genes. (http://biit.cs.ut.ee/gprofiler/page.cgi?welcome). The output is a tabular graphic where genes are shown in columns, functions in rows, and colored table cells showing functional associations. (http://biit.cs.ut.ee/gprofiler/).

g:GOSt uses multiple testing correction algorithms for distinguishing significant results from random matches. We used the Bonferroni Correction to show more clearly the difference between approximate values.

3.2.5 PRELIMINARY ENRICHMENT ANALYSIS OF METHYLATION DENSITY PLOTS USING SPSS

Given the enormous amount of entries of the Regional Methylation Density Calculation tables generated by Methy-Pipe, it would be unfeasible to display them, so we used SPSS to do a preliminary enrichment analysis with these because of its capacity to manage large amount of data.

“SPSS is a widely used program for statistical analysis in social science. It is also used by market researchers, health researchers, survey companies, government, education researchers, marketing organizations, data miners, and others.

In addition to statistical analysis, data management (case selection, file reshaping, creating derived data) and data documentation (a metadata dictionary was stored in the datafile) are features of the base software.” (http://www.ibm.com/analytics/us/en/technology/spss/)

The main output elements were retrotransposons (LINE and SINE), given the frequency and relevant role on aging.

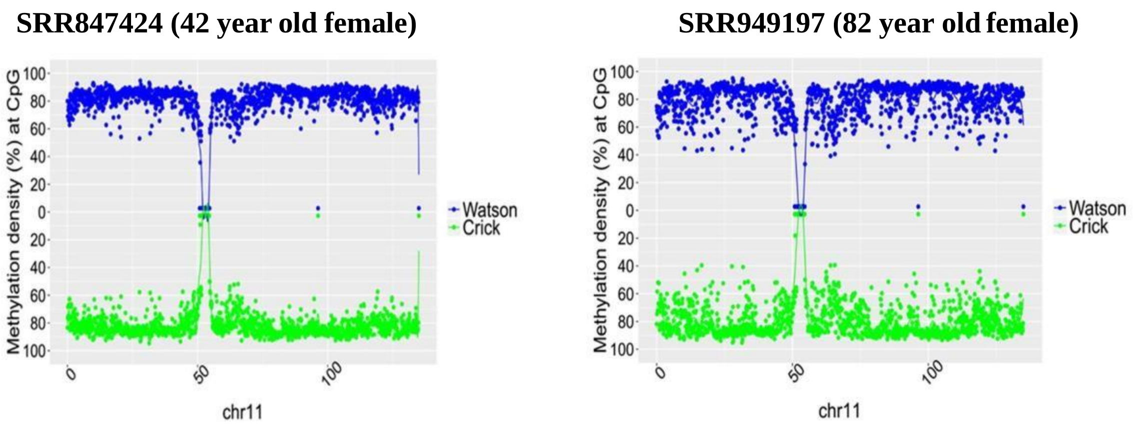
4 RESULTS AND DISCUSSION

We summarized the main conclusions we got from our samples from a biological standpoint according to the output sections:

4.1METHYPIPE OUTPUTS

4.1.1 Methylation Density across thechromosomes

In this section, we performed whole chromosome methylation profiling within a fixed window approach for brain and blood samples. In general, we can´t notice much difference among the plots, except in some cases (such as the chromosome 11) where those from older donors (64, 81 and 82) show a more dissipated methylation density cloud, which indicates they have more CpG sites less methylated (methylation density ranges from nearly 50-80% as opposed to the younger donors whose methylation density ranges from 70-80%). This could be explained by the epigenome erosion that states there is a global hypomethylation throughout aging which decreases transcription regulation and increases genome instability and susceptibility to develop diseases but it should also be taken into consideration that the tissues that the samples came from are not exactly the same: the sample from the younger donor derived from pre frontal cortex whereas the sample from the older donor originated from frontal cortex.


SRR847424 (42 year old female)

SRR949197 (82 year old female)

Figure 36 - Methylation Density in chromosomes 11 of both sample files SRR847424 and SRR949197 from brain. Here is evident the MD profile of CpGs to be clustered around 80% in the first file as opposed to the second file where the MD profile of Cps is more dissipated, ranging 50- 80%.

4.1.2 Base Content Percentage throughout the sequence cycles

We observed, for most samples, a very small percentage of cytosines C (1~2%), a small percentage of Guanines and Adenines (25% and 30%, respectively), and a high percentage of thymines (~45%) along all the sequence cycle (which roughly corresponds to the length of the sequences). This confirms, that most cytosines are unmethylated because, again, after bisulfite treatment and PCR amplification, unmethylated cytosines are converted to thymines and so, their naturally occurrence percentage content changes (decrease of cytosines from ~25% to ~1/2% and an increase of thymines from ~25% to ~43%).

The peaks we observe (mainly from the files SRR 47900, SRR 478994 and SRR478991, which belong to the same dataset) might be probably due to incomplete conversion of cytosines into uracils during the bisulfite treatment.

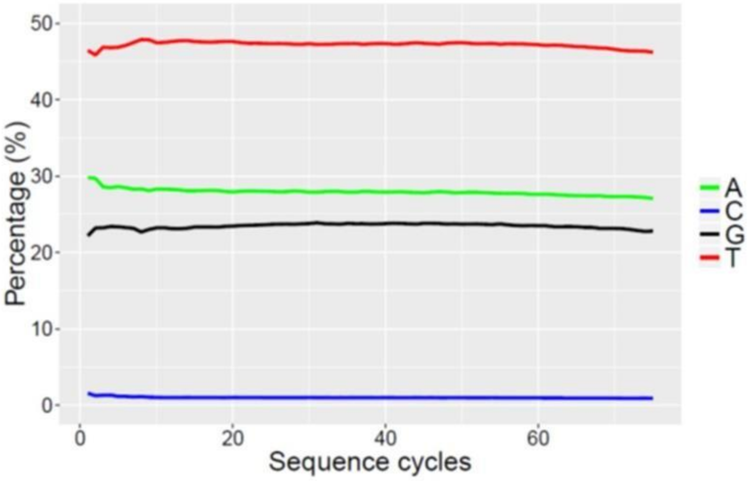


Figure 37 - Base content percentage across sequence cycles from fil SRR949197 (82 year old female). The plots present the base frequency at each sequencing cycle, where X-axis indicates the sequencing cycle and Y-axis indicates the base frequency. Similarly, to the plot from FASTQC, each cycle here is equivalent to the nucleotide position within the read.

4.1.3 Methylation density from CpG sites around TSS regions with 200 bp bins in Watson and crick strands

Here CpG sites are considered to have a length of 200 pb (bin). Methy-Pipe performs the profile with the mean methylation density for each CpG within 5000 pbs, upstream and downstream from the transcription start sites.

MDs around (TSSs) (Peiyong Jiang, 2014), are usually correlated with low levels of expression. From a first insight of those plots we can see that each chromosome has an overall similar methylation pattern among the samples, presenting very low methylation density near the TSSs. This is to be expected, because it follows the methylation-induced expression theory that states that CpG-rich promoters remain largely unmethylated regardless the state of expression. We were able to confirm this hypothesis, with our plots from the brain samples showing a very similar profile among them and in comparison, with the output plots of the blood samples, for every life stage. This is supported by the literature since age related changes in MD near TSS are local and can either increase or decrease giving an overall constant result.

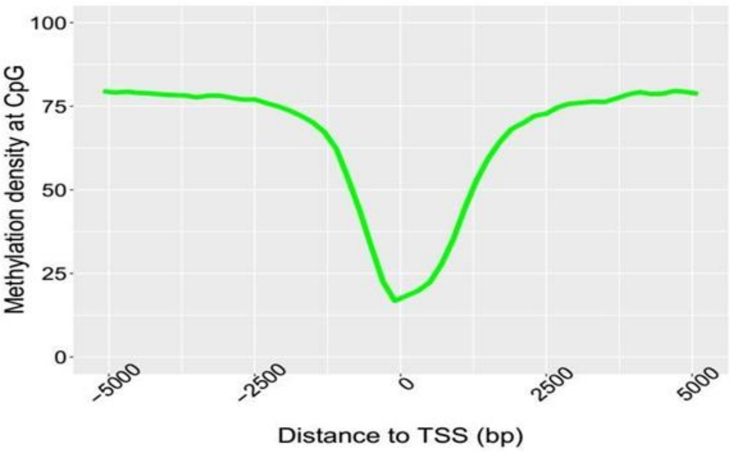


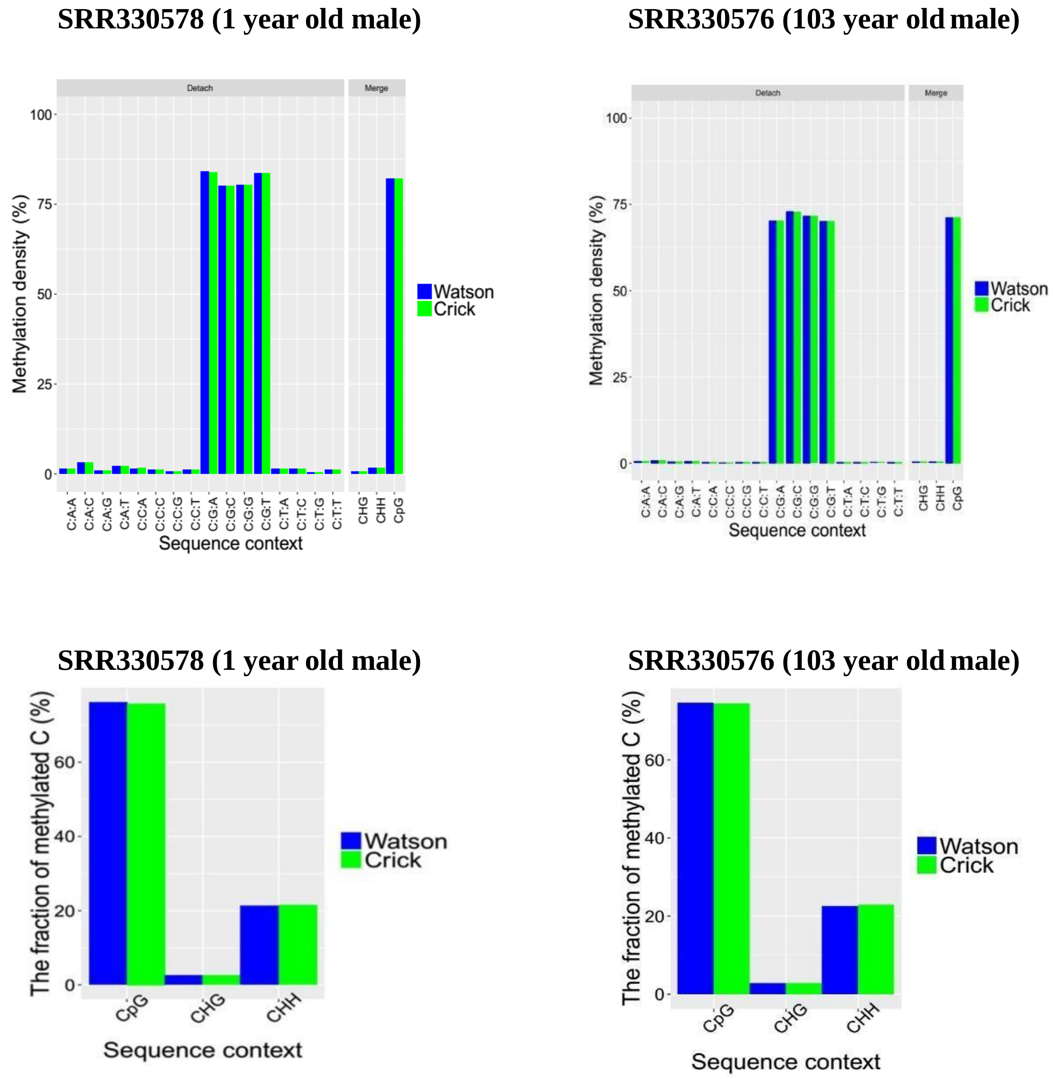
Figure 38 - MD from CpGs around TSS in the watson strand from the file SRR921723 (53 year old female). Methy- Pipe builds the profile with the mean methylation density for each CpG within 5000 pbs, upstream and downstream from the transcription start sites, computing the average among 200pb bin.

4.1.4 Methylation Density According to the Sequence Context

“Fractional methylated C is calculated as the proportion of the methylated cytosines at a particular sequence context over total methylated C sequenced”(Jiang et al., 2014).To determine the methylation density according to the sequence context, Methy-Pipe performs whole genome methylation profiling within different sequence contexts. MDs at different sequence contexts, namely (CAC, CAA, CAG, CAT, CCA, CCC, CCG, CCT, CGA, CGC, CGG,CGT, CTA, CTC, CTG, CTT), are calculated.

Although only a minor portion of all cytosines throughout the genome are methylated “mCH and hmC constitute major, and nonoverlapping, components of the methylated fraction of the genome in adult frontal cortex (mCG = 57.2%,mCH = 25.6%, hmC = 17.2%), with neurons contributing with a major percentage account of mCH” (nearly 53% in 50 year old adults) (Lister *et al.*, 2013).

At a genome-wide level, “it has been noticed that neuronal mCH is depleted in expressed genes, with mCH levels throughout the 5′-upstream, gene-body, and 3′-downstream regions inversely correlated with the abundance of the associated transcript” (Kinde, Gabel, Gilbert, Griffith, & Greenberg, 2015). Research using genome wide single-base resolution analysis in fetus and adults has found that “mCH is absent in the fetal cortex but accumulates in the early post-natal life (in the first 2 years in humans, characterized by a burst in synaptogenesis followed by activity-dependent pruning of excess synapses meaning that neurons could use this modification to sculpt their gene expression during critical periods" (Tognini, Napoli, & Pizzorusso, 2015); followed by slower accumulation of mCH during later adolescence which may eventualy become the predominant form of DNA methylation in mature neurons (Lister *et al.*, 2013).Our results do support this theory, given the trend towards an increase in CpH and decrease in CpG methylation, when comparing the cytosine methylated fraction of samples from younger male donors to their older donors (figure 39).



SRR330578 (1 year old male)

SRR330576 (103 year old male)

SRR330578 (1 year old male)

SRR330576 (103 year old male)

Figure 39 - Methylation Density and Fraction of Methylated Cytosines according to the sequence context of sample file SRR330578 (1 year old male) and SRR330576 (103 year old male). The fraction of Methylated Cytosines in CpG context is relatively higher in case of the first sample than in the second sample, and conversely, the fraction of Methylated Cytosines in CpH context is higher in the second sample.

When comparing middle age males (55 year olds) with their age matched females (53 year olds) from brain samples we can observe two different situations: one where the dominant methylated sequence context is CpG (which is the case for the sample files SRR921723, SRR921749, which are from glia cell samples) and the other where it is CpH (as is the case for sample files SRR921735 and SRR921706 which are from neuron cells). As reported by (Lister, 2013), neurons are globally more enriched for mCH compared with glia, which is consistent with our results. This is an important result since it proves that our methodology is sensitive enough to detect such relevant differences between methylomes representative of similar ages. Additionally, it stresses the high importance of quantifying methylcytosines within their nucleotide context, due to its functional relevance as well (figure 40).

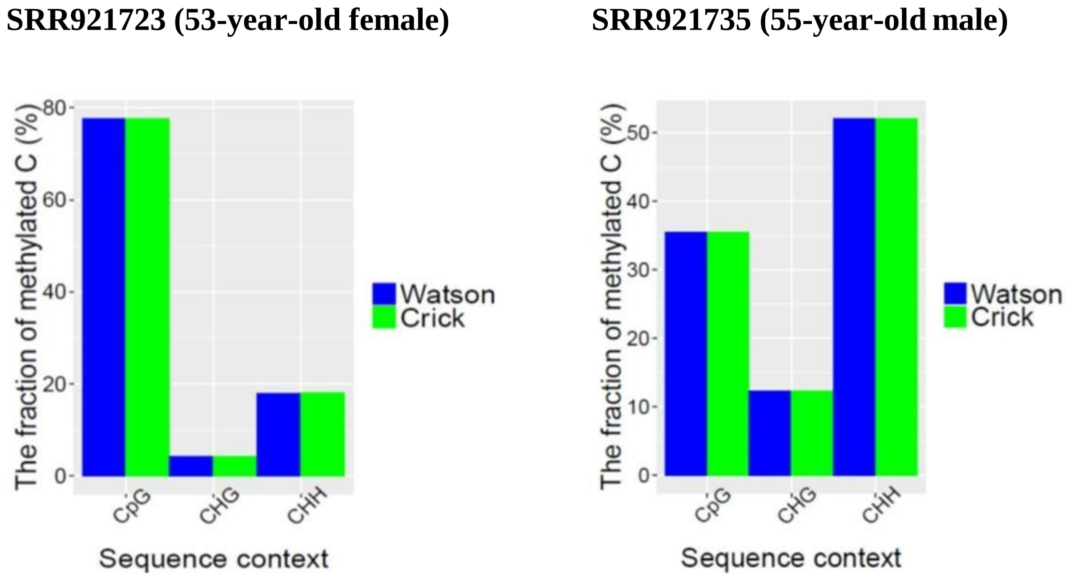

SRR921723 (53-year-old female) **SRR921735 (55-year-old male)**

Figure 40 - The fraction of methylated C according to the sequence context for files SRR921723 and SRR921735. Since the first sample file is from glia cells, we can see CpG as the major methylated context, whereas the second sample, that is from neurons the major methylated context is CpH (H=A,T,G) which is according to the previous results.

Some output plots (from files SRR847432, and SRR847427_1/SRR847427_2) had to be disconsidered due to the fact that these samples were not only bisulfite sequenced, but also tab sequenced. This, in turn, retrieved only hydroxymethylations, which reduced the overall total methylation (methylation and hydroxymethylations). We could also immediately notice a very similar methylation density between both strands (Watson and Crick) that is to be expected, because in general each nucleotide has the same genome-wide frequency in both strands (at least in CG context, due to base pair complementarity). We also confirmed previous observations about the preferred non CpG sequences in neurons as being CAC (Lister *et al.*, 2013).

4.1.5 Enrichment Analysis of Regional Methylation Density using SPSS

The results we obtained clearly highlighted particular methylated regions: the transposons and retrotransposons, which shouldn't be surprising at all. Indeed, nearly 40% of the genome is composed of sequences derived from these mobile genetic sequences

A major fraction of methylation sites inside the genome are in repeat sequences and transposable elements, like SINE and long-interspersed nuclear element (LINE-1) which are among the most common and best characterized repetitive elements. Alu is the most abundant of the short-interspersed nuclear elements (SINE) with more than a million copies per genome, composing approximately 11% of the mass of human genome and contain 30% of its methylation sites. They are nonautonomous non-long-terminal-repeat retrotransposons derived from human gene 7SL (which codes for the RNA component of the signal recognition particle ribonucleoprotein complex); and are the most successful human SINEs (Angela Macia, 2011).

Despite the high prevalence of transposable elements in the human genome and the abundance of several LINE and SINE subfamilies in this genome, apparently at present only certain members of each class are active (Macia *et al.*, 2011). Previous research established that heterochromatin is enriched for transposable elements, which are known to have harmful effects on the genome when transposed, leading some researchers to postulate a role for TEs in ageing. Moreover, it has been suggested that an age-related expression of transposable elements may contribute to ageing, consistent with the previously proposed retrotransposon theory of ageing (Wood *et al.*, 2016). When comparing the methylation density between younger and older donors, that of older donors have, overall, a slightly lower methylation density in LINE and SINE elements, as predicted by the retrotransposon theory of age. Therefore, our results also suggest that our workflow was able, once more to lead to conclusions consistent with current theories about aging.

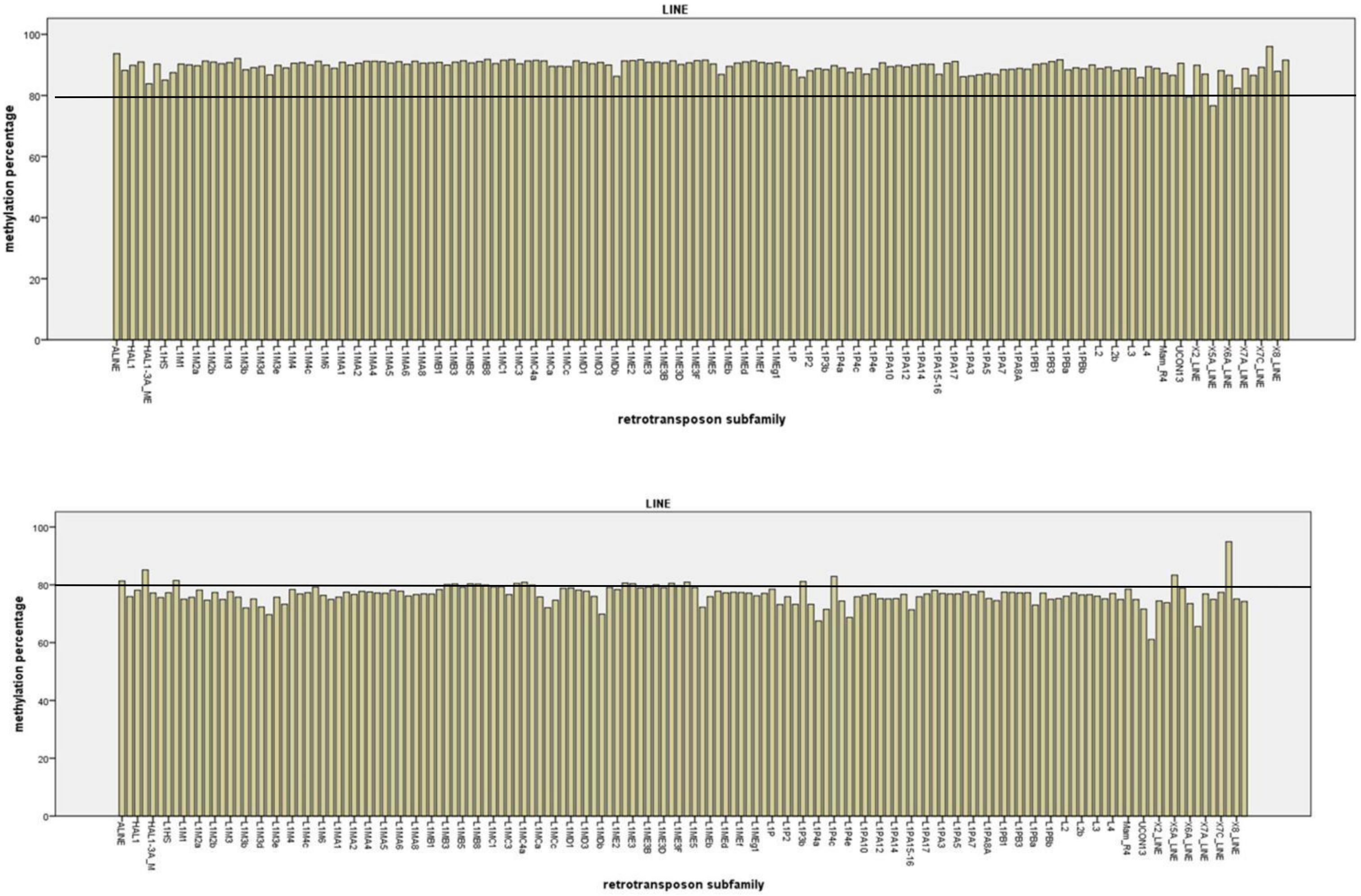


Figure 41 - Enrichment analysis of files SRR330578 (shown on top) and SRR330576 (shown on bottom) for LINE elements. It is clear that in case of file SRR330578 (1 year old male from blood) most LINE subfamilies have methylation density over 80%, unlike file SRR330576 (103 year old male from blood) most LINE subfamilies have a methylation percentage under 80%.

4.2 PRELIMINARY ENRICHMENT ANALYSIS OF DIFFERENTALLY METHYLATED REGIONS WITH G:PROFILER

Our analysis of differential Methylated Regions was based on three main different type of comparisons: Older individual (82 year old female and 103 year old male) with younger individual (42 year old female, 1 year old male) within the same tissue sample (brain and blood, respectively) so we could discuss the effect of aging within the same gender; middle age male individuals with middle age female individuals from the same tissue sample (brain) to ascertain the effect of gender; and young adults from both tissue samples (31 year old from brain samples and 26 year old from blood samples) to evaluate the effect of tissue difference. For the first comparisons, we will briefly discuss the main biological processes (those with the highest statistical significance, given by the lowest -log(pvalue)) affected by the genes found in the DMRs. The complete DMR data obtained for every comparison described below are presented in tables 4 to 16, together the genes found to be associated with those regions.

This shall be, however, an over simplified analysis since it is only based on the transcription-induced demethylation theory mentioned previously in the introduction (mostly applied to promoter regions) that may not always be accurate, which states: biological processes assigned to hypo methylated regions will be considered upregulated and biological processes assigned to hypermethylated regions will be considered downregulated. Unfortunately, we were not able to assign genes to the DMRs detected for the two last types of comparisons due to a relatively smaller amount of DMRs to start with or to their location within the genome that could not be associated with any nearby genes.

4.2.1 Assessing the effect of Aging Upon Methylation

From the first comparison between the brain samples (82-year-old with 42-year-old females) we can see up regulated biological process involved in carcinogenesis found in the hypo methylated genes. Since DNA hypomethylation can activate oncogenes and initiate chromosome instability, whereas DNA hypermethylation initiates silencing of tumor suppressor genes, these results are in line with the hypothesis of increasing susceptibility to cancer development with age.

From the list of genes mentioned previously in the Introduction chapter as being hypermethylated with aging, we could find MGMT and WRN and from those genes related with Alzheimer´s disease, we could also find ANK1 and RHBDF2 included in the table of hypo methylated regions. Again, this shows that our methodology is sensitive and generates results as expected, suggesting the onset of age related diseases (figure 42).

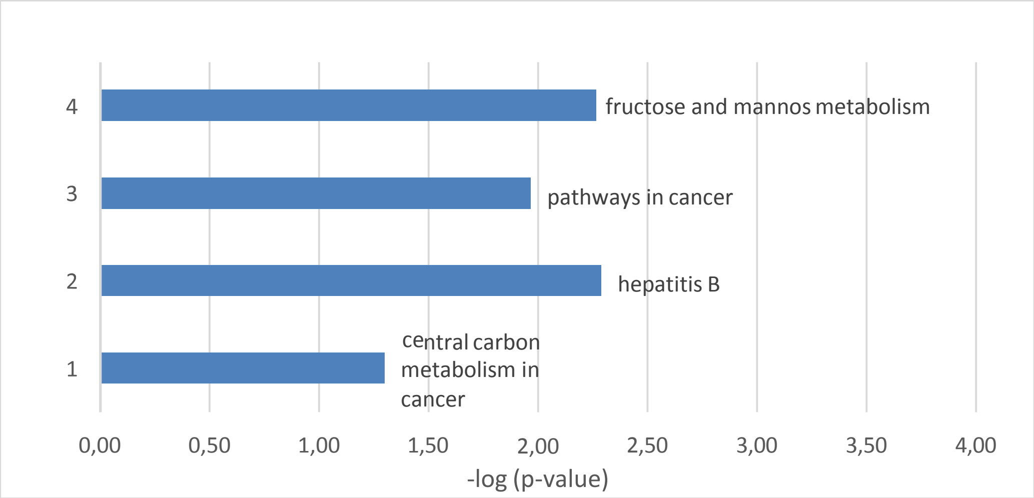


Figure 42 - Putative up regulated biological processes assigned to the hypo methylated genes found between the test sample SRR949197 (82-year-old female sample) and the control sample SRR847427 (42 year old female) from brain.

As downregulated processes, we had highlighted the regulation of actin cytoskeleton. Actin cytoskeleton dynamics plays a crucial part in processes such as: embryonic morphogenesis, immune surveillance, angiogenesis and tissue repair and regeneration, mediating the formation of cellular structures such as lamellipodia, filopodia, stress fibers and focal adhesions. (Lee & Dominguez, 2010). In this regard, it is interesting to confirm the conclusions from a recent study done with samples from Sri Lanka which states that "aging cytoskeletal pathologies are comparatively higher in elderly Sri Lankans and this might be due to their genetic, dietary and/ or environmental variations" (Wijesinghe P, 2016) (figure 43).

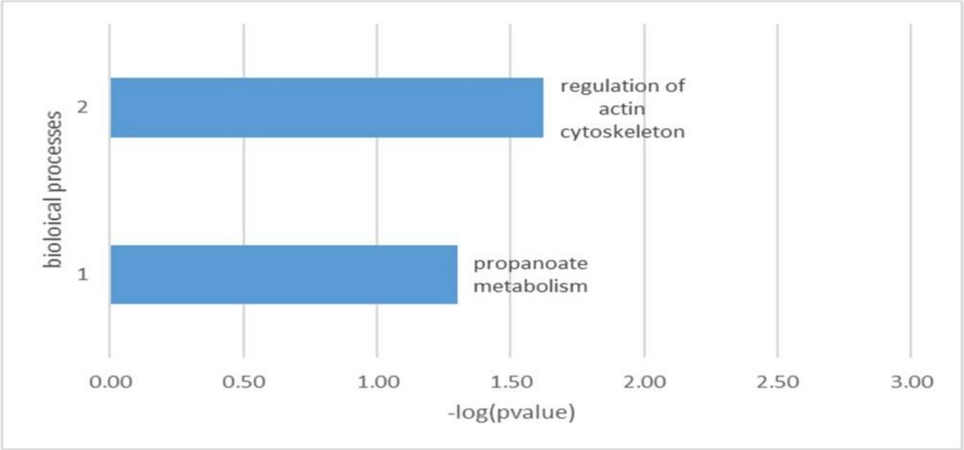


Figure 43 – Putative down regulated biological processes assigned to the hyper methylated genes found between the test sample of SRR949197 (82-year-old female sample) and the control sample of SRR847427 (42-year-old female) from brain.

For the second comparison between blood samples of the centenarian and the newborn, we can also highlight some upregulated biological functions: type II diabetes mellitus t and natural killer cells mediated cytotoxicity (figure 44). The incidence and prevalence of Type 2 diabetes increases with age although the underlying mechanisms behind why diabetes is increasing in the elderly is still not clearly understood. It has been proposed that insulin resistance increases with age due to increased adiposity, decreased lean muscle mass, deficient nutrition, and reduced physical activity (Gunasekaran & Gannon, 2011). On the other hand, natural killer cells mediated cytotoxicity is a process by which mature natural killer (NK) cells induce target cell death (Zamai et al., 1998), which in turn debilitates the immune system. In conclusion, both these processes are more likely to occur in older people, as our results corroborate. We could detect some Alzheimer related genes, as well, in the samples: ANK1 and CDH23

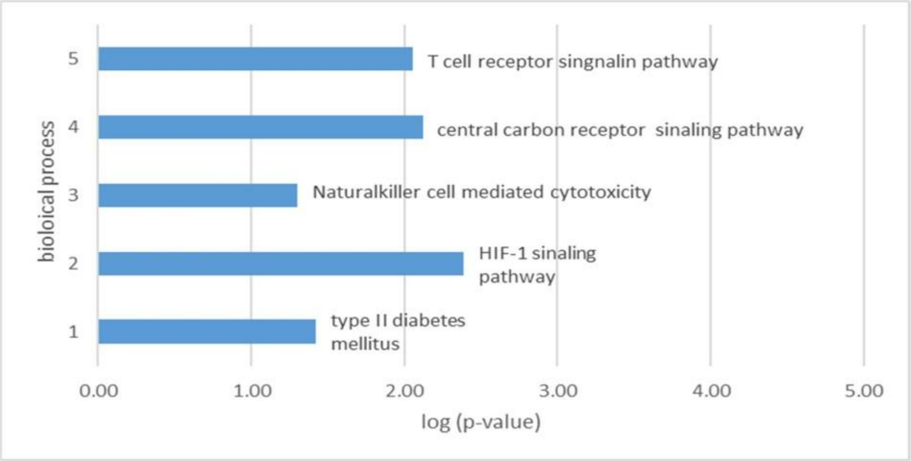


Figure 44 - Putative up regulated biological processes assigned to the hypo methylated genes found between the test sample of SRR330576 (103year old male) and the control sample of SRR330578 (1-year-old male) from blood.

As the main downregulated biological processes we can clearly highlight age related processes: the mTOR pathway, involved in the regulation of the cell cycle; the Wnt signaling pathway, involved in the cellular proliferation and embryonic development; phospholipase D signaling pathway, involved in functions, such as: growth/proliferation, vesicle trafficking, cytoskeleton modulation, development, and morphogenesis (Jang, Lee, Hwang, & Ryu, 2012); and signaling pathways regulating pluripotency of stem cells, all cellular mechanisms known to be modulated by age.

It is also interesting to notice as one of the main downregulated biological processes the retrograde endocannabinoide system (figure 45). This system has many interactions with other signaling and neuromodulatory systems and is the principal mode by which endocannabinoids mediate short- and long-term forms of plasticity at both excitatory and inhibitory synapses. It is believed that by modulating synaptic strength, endocannabinoids can regulate a wide range of neural functions, including cognition, motor control, feeding behaviors and pain. (Purpura & Einstein, 2013) Due to this crucial role in the central nervous system function, and, particularly in the brain, this downregulation is very consistent with our hypothesis of increased risk of neurodegenerative diseases with aging. We could also detect some of those genes mentioned in the introduction chapter as being differentially methylated with aging, namely MGMT and NOS1 in the samples.

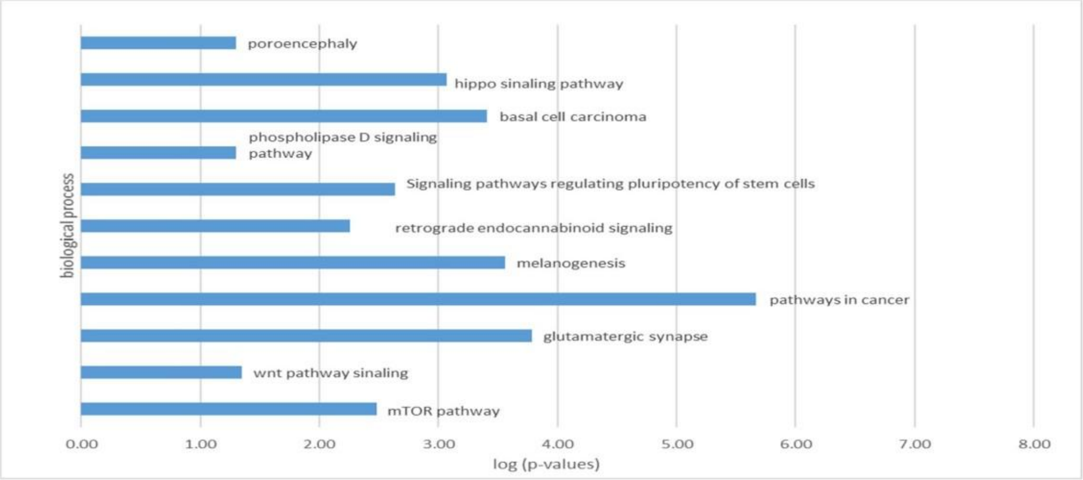


Figure 45 - Putative down regulated biological processes assigned to the hyper methylated genes found between the test sample SRR330576 (103year old male) with the control sample SRR330578 (1 year old male) from blood.

4.2.2 Assessing the effect of gender upon methylation

In this case, we made four comparisons: test sample files of SRR9211723 (53 year old female) with controls SRR9211749 and 921735 (55 year old males); and test files SRR921706 (53 year old female) also with control files SRR921749 and SRR9211735. Although we were not able to assign any gene to the DMRs detected, we were able to notice that each comparison revealed a Hypo Methylated Region on the X chromosome which confirms the initial expectations due to the chromosome X inactivation.

4.2.3 Assessing the effects of differential tissues (brain and blood) samples upon methylation

From the comparison between the samples of both tissues (31 year old male sample brain with 26 year old male blood sample) we could find some DMRs and, although we couldn´t assign them any gene, we cannot exclude the hypothesis of these DMRs regions of having any functional role (because they could be affecting regulatory regions, at *trans* level).

Table 4-Hypo methylated regions found between the test file SRR949197 (82-year-old female) and control sample of SRR84742 (42-year-old female) from brain tissue. Column CHR indicates the chromosome where the DMR is located, columns START and END indicate the beginning and end of the DMR position in the aforementioned chromosome, column TYPE indicates whether the DMRs are of the type hypo or hyper; the columns TEST (C/T) and CONTROL (C/T) indicate the number of cytosines/thymines in DMR of the test and control sample, respectively: the columns METH%(TEST/CONTROL) indicate the methylation percentage of the DMR in test and control samples, respectively; the column P value indicates the statistical significance of each DMR given by its Pvalue; the CpG (TEST/CONTROL)columns indicate the number of CpGs in eacha DMR for TEST and CONTROL samples; and AFECTED REGION indicates the gene and region affected by the DMR detected. This is only a small portion of the original table which can be supplied if requested

CHR	START	END	TYPE	TEST		CONTROL		%METH		P-VALUE	CpG		AFFECTED REGION
				C	T	C	T	TEST	CONTROL		TEST	CONTROL	
chrX	152127500	152128000	hypo	42	215	27	18	16.34	60	5.30E-63	18	6	ZNF185:intron
chrX	153361600	153362600	hypo	88	209	81	53	29.63	60.45	5.30E-63	29	21	MECP2:intron
chrX	137794300	137794800	hypo	51	167	30	26	23.39	53.57	1.44E-57	12	6	FGF13:promoter
chrX	130926900	130927400	hypo	46	84	46	11	35.38	80.7	7.58E-51	11	8	LOC286467:intron
chrX	135848600	135849100	hypo	23	71	23	13	24.47	63.89	7.58E-51	11	5	ARHGEF6:intron
chrX	117957200	117957700	hypo	43	124	22	12	25.75	64.71	3.10E-50	15	6	ZCCHC12:promoter
chrX	119377900	119378400	hypo	16	71	36	30	18.39	54.55	3.10E-50	9	8	NKAPP1:intron
chrX	110188000	110188500	hypo	17	78	19	18	17.89	51.35	1.10E-36	12	6	PAK3:intron
chrX	117957200	117957700	hypo	43	124	22	12	25.75	64.71	1.10E-36	15	6	ZCCHC12:promoter
chrX	106871400	106871900	hypo	68	213	37	30	24.2	55.22	1.99E-33	34	11	PRPS1:promoter, PRPS1:5UTR
chrX	106449600	106450100	hypo	118	225	40	14	34.4	74.07	1.92E-28	22	7	NUP62CL:promoter, CXorf41:promoter, NUP62CL:5UTR, CXorf41:5UTR
chrX	106362000	106362500	hypo	24	78	23	17	23.53	57.5	2.48E-28	10	5	RBM41:promoter RBM41:5UTR

Table 5-Hyper methylated regions found between the test file SRR949197 (82-year-old female) and control sample file of SRR84742 (42-year-old female), from brain sample. Column CHR indicates the chromosome where the DMR is located, columns START and END indicate the beginning and end of the DMR position in pbs; column TYPE indicates whether the DMRs are of the type hypo or hyper; the columns TEST (C/T) and CONTROL (C/T) indicate the number of cytosines/thymines in DMR of the test and control sample, respectively: the columns METH%(TEST/CONTROL) indicate the methylation percentage of the DMR in test and control samples, respectively; the column P value indicates the statistical significance of each DMR given by its Pvalue; the CpG (TEST/CONTROL)columns indicate the number of CpGs in each DMR for TEST and CONTROL samples; and AFECTED REGION indicates the gene and region affected by the DMR detected. This is only a small portion of the original table which can be supplied if requested

CHR	START	END	TYPE	TEST		CONTROL		%METH		P-VALUE	CpG		AFFECTED REGION
				C	T	C	T	TEST	CONTROL		TEST	CONTROL	
chr12	49297400	49297900	hyper	38	17	23	52	69.09	30.67	9.98E-03	8	12	CCDC65:promoter, CCDC65:5UTR
chr11	118498900	118499400	hyper	116	68	13	35	63.04	27.08	9.92E-03	20	7	PHLDB1:CDS:5
chr5	135415300	135416300	hyper	338	233	21	177	59.19	10.61	3.55E-15	35	28	VTRNA2-1:ncexon:1
chr4	187125100	187126100	hyper	307	47	14	71	86.72	16.47	1.29E-09	36	14	CYP4V2:intron
chr4	57252800	57253300	hyper	35	15	0	65	70	0	3.45E-09	10	6	AASDH:intron
chr16	55362100	55365600	hyper	646	1379	61	376	31.9	13.96	4.25E-09	133	68	IRX6:CDS:5 IRX6:CDS:6
chr10	42862500	42864000	hyper	183	424	2	132	30.15	1.49	5.09E-09	59	21	LOC441666:ncexon:1
chr5	134259200	134259700	hyper	124	369	34	393	25.15	7.96	5.13E-09	20	14	PCBD2:intron
chr16	55794100	55795100	hyper	378	121	16	77	75.75	17.2	1.11E-08	40	16	CES1P1:ncexon:1
chr6	163680800	163682300	hyper	92	288	1	136	24.21	0.73	6.86E-08	34	22	PACRG:intron
chr20	21491400	21492400	hyper	36	87	13	207	29.27	5.91	5.29E-07	18	29	NKX2-2:3UTR
chr8	145670400	145670900	hyper	122	131	0	53	48.22	0	9.06E-07	24	7	TONSL:promoter
chr18	21167100	21167600	hyper	44	23	5	60	65.67	7.69	1.65E-06	5	6	NPC1:promoter
chr18	21167100	21167600	hyper	44	23	5	60	65.67	7.69	1.65E-06	5	6	NPC1:promoter
chr19	14639300	14639800	hyper	45	49	4	75	47.87	5.06	1.67E-06	13	8	TECR:promoter
chr19	14639300	14639800	hyper	45	49	4	75	47.87	5.06	1.67E-06	13	8	TECR:promoter
chr8	101169600	101170100	hyper	28	15	3	57	65.12	5	2.03E-06	7	8	SPAG1:promoter

chr8	101169600	101170100	hyper	28	15	3	57	65.12	5	2.03E-06	7	8	SPAG1:promoter
chr19	55591100	55591600	hyper	88	92	1	54	48.89	1.82	2.20E-06	25	8	EPS8L1:promoter
chr19	55591100	55591600	hyper	88	92	1	54	48.89	1.82	2.20E-06	25	8	EPS8L1:promoter
chr7	111396300	111397800	hyper	143	131	27	123	52.19	18	2.66E-06	25	21	DOCK4:intron
chr1	3633300	3633800	hyper	43	9	0	32	82.69	0	2.87E-06	7	5	TP73:intron
chr20	44001200	44003100	hyper	171	276	15	124	38.26	10.79	3.24E-06	54	26	TP53TG5:CDS:5
chr15	78422900	78423400	hyper	87	32	4	43	73.11	8.51	4.27E-06	14	7	CIB2:intron

Table 6-Hypo methylated Regions found between the test file SRR330676 (103 year old male) with the control file SRR330578 (1 year old male) from blood sample. Column CHR indicates the chromosome where the DMR is located, columns START and END indicate the beginning and end of the DMR position in the aforementioned chromosome, column TYPE indicates whether the DMRs are of the type hypo or hyper; the columns TEST (C/T) and CONTROL (C/T) indicate the number of cytosines/thymines in DMR of the test and control sample, respectively: the columns METH%(TEST/CONTROL) indicate the methylation percentage of the DMR in test and control samples, respectively; the column P value indicates the statistical significance of each DMR given by its Pvalue; the CpG (TEST/CONTROL)columns indicate the number of CpGs in each DMR for TEST and CONTROL samples; and AFFECTED REGION indicates the gene and region affected by the DMR detected. This is only a small portion of the original table which can be supplied if requested

CHR	START	END	TYPE	TEST		CONTROL		%METH		P-VALUE	CpG		AFFECTED REGION
				C	T	C	T	TEST	CONTROL		TEST	CONTROL	
chr1	111215900	111216900	hypo	7	329	80	125	2.08	39.02	5.43E-21	46	32	KCNA3:CDS:1
chr19	1905500	1906000	hypo	1	244	19	17	0.41	52.78	3.93E-20	37	5	ADAT3:promoter
chr19	2150300	2152300	hypo	2	422	33	95	0.47	25.78	6.11E-20	58	22	AP3D1:promoter, AP3D1:5UTR
chr12	1701700	1702700	hypo	16	234	86	69	6.4	55.48	8.64E-17	37	21	FBXL14:CDS:1
chr6	501900	502400	hypo	0	108	33	3	0	91.67	1.61E-15	12	6	EXOC2:intron
chr1	52498000	52499000	hypo	4	193	45	53	2.03	45.92	1.26E-14	30	17	KTI12:CDS:1
chr9	92033200	92035600	hypo	7	164	86	74	4.09	53.75	2.54E-14	27	24	SEMA4D:promoter, SEMA4D:5UTR
chr5	41870700	41871200	hypo	0	91	40	10	0	80	5.79E-13	13	5	OXCT1:promoter, OXCT1:5UTR

chr3	71275900	71276900	hypo	1	145	41	51	0.68	44.57	1.73E-12	23	14	FOXP1:intron
chr10	124638500	124640100	hypo	38	319	101	135	10.64	42.8	2.62E-12	54	32	FAM24B:promoter, FAM24B:5UTR
chr2	231788900	231789900	hypo	20	154	60	22	11.49	73.17	1.11E-11	19	14	GPR55:promoter, GPR55:5UTR
chr17	36283000	36294100	hypo	11353	3024	9208	1067	78.97	89.62	3.22E-11	307	306	TBC1D3F:promoter, TBC1D3:5UTR
chr1	17053400	17055000	hypo	323	346	490	65	48.28	88.29	4.09E-11	42	42	MIR3675:intron
chr16	30077700	30078200	hypo	5	143	32	30	3.38	51.61	4.70E-11	20	7	ALDOA:promoter
chr16	2390800	2391300	hypo	0	154	34	75	0	31.19	1.27E-10	26	20	ABCA3:promoter
chr3	52279900	52280900	hypo	8	246	36	85	3.15	29.75	1.33E-10	38	18	PPM1M:promoter, PPM1M:5UTR
chr9	130829100	130829600	hypo	0	71	54	20	0	72.97	1.44E-10	10	12	NAIF1:promoter, NAIF1:5UTR
chr19	2945500	2946000	hypo	7	123	25	11	5.38	69.44	2.43E-10	15	7	ZNF77:promoter
chr16	3116000	3117800	hypo	15	117	84	38	11.36	68.85	2.91E-10	18	18	IL32:CDS:2,IL32:CDS:3
chr22	39495900	39496400	hypo	5	75	61	12	6.25	83.56	4.66E-10	9	11	APOBEC3H:promoter, APOBEC3H:5UTR

Table 7-Hyper Methylated Regions found between the test file SRR330676 (103-year-old male blood) sample and the control file of SRR330578 (1 year old male), from blood sample. Column CHR indicates the chromosome where the DMR is located, columns START and END indicate the beginning and end of the DMR position in the pbs, column TYPE indicates whether the DMRs are of the type hypo or hyper; the columns TEST (C/T) and CONTROL (C/T) indicate the number of cytosines/thymines in DMR of the test and control sample, respectively: the columns METH%(TEST/CONTROL) indicate the methylation percentage of the DMR in test and control samples, respectively; the column P value indicates the statistical significance of each DMR given by its Pvalue; the CpG (TEST/CONTROL)columns indicate the number of CpGs in each DMR for TEST and CONTROL samples; and AFECTED REGION indicates the gene and region affected by the DMR detected. This is only a small portion of the original table which can be supplied if requested

CHR	START	END	TYPE	TEST		CONTROL		%METH		P-VALUE	CpG		AFFECTED REGION
				C	T	C	T	TEST	CONTROL		TEST	CONTROL	
chr1	47897900	47900400	hyper	143	181	30	416	44.14	6.73	5.07E-22	49	66	MGC12982:ncexon:1
chr10	50818400	50822800	hyper	327	811	86	876	28.73	8.94	4.77E-21	167	145	CHAT:promoter, SLC18A3:promoter, CHAT:5UTR, CHAT:5UTR, SLC18A3:5UTR
chr9	124981900	124983400	hyper	138	301	12	364	31.44	3.19	2.06E-18	63	58	LHX6:5UTR
chr14	65007200	65009200	hyper	373	527	78	508	41.44	13.31	4.09E-18	132	96	HSPA2:promoter, HSPA2:5UTR
chr13	58207100	58208700	hyper	272	611	58	609	30.8	8.7	6.32E-18	110	98	PCDH17:CDS:1
chr2	157185200	157187200	hyper	140	56	38	225	71.43	14.45	3.49E-16	30	44	NR4A2:promoter NR4A2:5UTR
chr1	200842200	200843700	hyper	168	89	23	190	65.37	10.8	5.77E-16	42	38	GPR25:CDS:1
chr1	119525700	119529300	hyper	171	476	17	405	26.43	4.03	6.92E-16	99	72	TBX15:intron
chr9	140056300	140057800	hyper	97	105	17	244	48.02	6.51	2.09E-15	34	42	GRIN1:CDS:12 GRIN1:CDS:13, GRIN1:CDS:14, GRIN1:CDS:15,

													GRIN1:CDS:16, GRIN1:CDS:17
chr20	44686600	44687600	hyper	95	111	5	185	46.12	2.63	3.77E-15	35	31	SLC12A5:3UTR
chr17	41465900	41466400	hyper	139	513	17	478	21.32	3.43	9.94E-15	27	22	LOC100130581:ncexon:1
chr3	73672700	73674700	hyper	95	219	18	365	30.25	4.7	1.41E-14	50	63	PDZRN3:promoter PDZRN3:5UTR
chr22	19710000	19711000	hyper	97	38	10	134	71.85	6.94	3.18E-14	20	25	GP1BB:promoter
chr1	201617800	201618800	hyper	172	55	28	158	75.77	15.05	4.10E-14	35	31	NAV1:CDS:1
chr3	62304100	62305100	hyper	74	135	18	315	35.41	5.41	8.00E-14	31	44	C3orf14:promoter
chr2	25383800	25384800	hyper	51	26	18	198	66.23	8.33	9.64E-14	14	32	POMC:CDS:2
chr1	149286100	149288100	hyper	200	415	43	420	32.52	9.29	2.56E-13	62	52	LOC388692:ncexon:1
chr16	66612300	66613300	hyper	57	122	12	303	31.84	3.81	2.62E-13	26	39	CMTM2:promoter
chr2	20869100	20871700	hyper	130	174	25	262	42.76	8.71	3.19E-13	52	49	GDF7:CDS:2
chr2	172944900	172949300	hyper	124	394	18	397	23.94	4.34	6.07E-13	82	66	DLX1:promoter
chr10	112838500	112839000	hyper	48	15	10	136	76.19	6.85	6.32E-13	11	19	ADRA2A:CDS:1

Table 8-Hyper Methylated regions found between the test file SRR479000 (31-year-old male) from brain and the control file SRR389249 (26 year old male) from blood sample. There couldn't be assigned any gene to the DMRs. Column CHR indicates the chromosome where the DMR is located, columns START and END indicate the beinning and end of the DMR position pbs, column TYPE indicates whether the DMRs are of the type hypo or hyper; the columns TEST (C/T) and CONTROL (C/T) indicate the number of cytosines/thymines in DMR of the test and control sample, respectively: the columns METH%(TEST/CONTROL) indicate the methylation percentage of the DMR in test and control samples, respectively; the column P value indicates the statistical significance of each DMR given by its Pvalue; the CpG (TEST/CONTROL)columns indicate the number of CpGs in eacha DMR for TEST and CONTROL samples; and AFECTED REGION indicates the gene and region affected by the DMR detected.

CHR	START	END	TYPE	TEST		CONTROL		%METH		P-VALUE	CPG	
				C	T	C	T	TEST	CONTROL		TEST	CONTROL
chr16	33962800	33964000	hyper	723	1860	85	391	27.99	17.86	2.96E-04	54	25
chr4	49099300	49099800	hyper	106	234	14	113	31.18	11.02	3.83E-04	12	6
chr4	49109900	49110400	hyper	196	262	19	100	42.79	15.97	1.01E-04	12	5
chr4	49142100	49142600	hyper	84	114	6	41	42.42	12.77	5.39E-03	7	6
chr4	49653100	49653600	hyper	159	336	15	93	32.12	13.89	3.10E-03	10	6
chr5	99387200	99390500	hyper	402	916	18	1330	30.5	1.34	1.81E-70	66	41

Table 9-Hypo Methylated regions found between the test file SRR479000 (31 year old male) from blood and the control file SRR389249 (26 year old male) from blood sample. There couldn’t be assigned any gene. Column CHR indicates the chromosome where the DMR is located, columns START and END indicate the beginning and end of the DMR position in pbs, column TYPE indicates whether the DMRs are of the type hypo or hyper; the columns TEST (C/T) and CONTROL (C/T) indicate the number of cytosines/thymines in DMR of the test and control sample, respectively: the columns METH%(TEST/CONTROL) indicate the methylation percentage of the DMR in test and control samples, respectively; the column P value indicates the statistical significance of each DMR given by its Pvalue; the CpG (TEST/CONTROL)columns indicate the number of CpGs in each DMR for TEST and CONTROL samples. There couldn´t be assigned any gene to the DMRs.

CHR	START	END	TYPE	TEST		CONTROL		%METH		P-VALUE	CpG	
				C	T	C	T	TEST	CONTROL		TEST	CONTROL
chr10	42392300	42396600	hypo	5459	2245	797	138	70.86	85.24	3.16E-04	170	67
chr16	46386200	46390400	hypo	5058	2496	799	196	66.96	80.3	3.48E-04	131	64
chr16	46427400	46429900	hypo	6847	2459	2615	597	73.58	81.41	1.00E-03	102	86

Table 10-Hypo Methylated Regions between test sample of SRR921723 (53 year old female) and the control sample SRR921735 (55 year old male), from brain. Column CHR indicates the chromosome where the DMR is located, columns START and END indicate the beginning and end of theDMR position in pbs, column TYPE indicates whether the DMRs are of the type hypo or hyper; the columns TEST (C/T) and CONTROL (C/T) indicate the number of cytosines/thymines in DMR of the test and control sample, respectively: the columns METH%(TEST/CONTROL) indicate the methylation percentage of the DMR in test and control samples, respectively; the column P value indicates the statistical significance of each DMR given by its Pvalue; the CpG (TEST/CONTROL)columns indicate the number of CpGs in each DMR for TEST and CONTROL samples. There couldn’t be assigned any gene.

CHR	START	END	TYPE	TEST		CONTROL		%METH		P-VALUE	CpG	
				C	T	C	T	TEST	CONTROL		TEST	CONTROL
chr15	28817600	28818100	hypo	9	30	31	7	23.08	81.58	3.32E-03	6	5
chr17	31149200	31149700	hypo	37	180	145	293	17.05	33.11	8.92E-04	17	17
chr1	91852400	91853400	hypo	273	1321	528	1237	17.13	29.92	8.06E-12	19	19
chr21	9825200	9827700	hypo	1765	7661	3754	9753	18.72	27.79	1.32E-35	334	335
chr21	10732900	10733400	hypo	17	42	29	9	28.81	76.32	7.54E-03	5	5
chr7	74631100	74631600	hypo	11	40	35	1	21.57	97.22	1.35E-04	6	6
chr8	70602000	70602500	hypo	112	451	211	407	19.89	34.14	3.16E-05	10	10
chrX	108297000	108297500	hypo	24	139	36	54	14.72	40	5.23E-04	9	5

Table 11-Hyper Methylated Regions between the test sample of SRR921723 (53-year-old female) and the control sample SRR921735 (55-year-old male) from brain. Column CHR indicates the chromosome where the DMR is located, columns START and END indicate the beginning and end of the DMR position in pbs, column TYPE indicates whether the DMRs are of the type hypo or hyper; the columns TEST (C/T) and CONTROL (C/T) indicate the number of cytosines/thymines in DMR of the test and control sampe, respectively: the columns METH%(TEST/CONTROL) indicate the methylation percentage of the DMR in test and control samples, respectively; the column P value indicates the statistical significance of each DMR given by its Pvalue; the CpG (TEST/CONTROL)columns indicate the number of CpGs in each DMR for TEST and CONTROL samples, respectively.There couldn't be assigned any gene.

CHR	START	END	TYPE	TEST		CONTROL		%METH		P-VALUE	CpG	
				C	T	C	T	TEST	CONTROL		TEST	CONTROL
chr16	22482600	22484700	hyper	173	43	149	133	80.09	52.84	3.74E-03	21	26
chr4	190611000	190611500	hyper	19	7	4	25	73.08	13.79	3.97E-03	5	5
chr9	67791800	67792300	hyper	57	39	10	41	59.38	19.61	2.97E-03	15	9

Table 12-Hyper Methylated Regions between the test sample file of SRR921706 (53 year old female) and the control sample file SRR921735 (55 year old male) from brain. Column CHR indicates the chromosome where the DMR is located, columns START and END indicate the beginning and end of the DMR position in pbs, column TYPE indicates whether the DMRs are of the type hypo or hyper; the columns TEST (C/T) and CONTROL (C/T) indicate the number of cytosines/thymines in DMR of the test and control sample, respectively: the columns METH%(TEST/CONTROL) indicate the methylation percentage of the DMR in test and control samples, respectively; the column P value indicates the statistical significance of each DMR given by its Pvalue; the CpG (TEST/CONTROL)columns indicate the number of CpGs in each DMR for TEST and CONTROL samples; There couldn't be assigned any gene.

CHR	START	END	TYPE	TEST		CONTROL		%METH		P-VALUE	CpG	
				C	T	C	T	TEST	CONTROL		TEST	CONTROL
chr10	39139900	39140400	hyper	132	48	10	26	73.33	27.78	7.76E-03	11	5
chr5	99387400	99387900	hyper	25	82	5	81	23.36	5.81	3.83E-03	8	6
chr5	99389300	99389800	hyper	60	15	12	99	80	10.81	5.50E-10	5	5
chr9	67791800	67792300	hyper	92	60	10	41	60.53	19.61	1.60E-03	15	9

Table 13-Hypo Methylated Regions between the test file of SRR921706 (53-year-old female) and the control file SRR921735 (55 year old male), from brain. Column CHR indicates the chromosome where the DMR is located, columns START and END indicate the beginning and end of the DMR position in pbs; column TYPE indicates whether the DMRs are of the type hypo or hyper; the columns TEST (C/T) and CONTROL (C/T) indicate the number of cytosines/thymines in DMR of the test and control sample, respectively: the columns METH%(TEST/CONTROL) indicate the methylation percentage of the DMR in test and control samples, respectively; the column P value indicates the statistical significance of each DMR given by its Pvalue; the CpG (TEST/CONTROL)columns indicate the number of CpGs in each DMR for TEST and CONTROL samples.There couldn't be assigned any gene.

CHR	START	END	TYPE	TEST		CONTROL		%METH		P-VALUE	CpG	
				C	T	C	T	TEST	CONTROL		TEST	CONTROL
chr17	31149200	31150200	hypo	190	1172	344	629	13.95	35.35	1.78E-21	54	45
chr17	33477700	33478700	hypo	155	836	246	529	15.64	31.74	2.55E-10	26	24
chr1	91852500	91853500	hypo	306	1746	528	1237	14.91	29.92	4.18E-19	19	19
chr1	156185900	156186900	hypo	434	2555	794	1411	14.52	36.01	1.07E-44	56	53
chr20	26188400	26190900	hypo	425	2678	739	1565	13.7	32.07	2.14E-38	116	115
chr2	133025500	133026000	hypo	87	466	101	301	15.73	25.12	3.32E-03	18	18
chr4	49142100	49142600	hypo	24	131	52	79	15.48	39.69	4.56E-04	7	7
chr5	71146300	71147300	hypo	108	729	157	408	12.9	27.79	1.14E-08	18	18
chrX	108297300	108297800	hypo	88	446	180	419	16.48	30.05	2.28E-05	18	20

5 FUTURE WORK

In the adult mammalian brain, while the DNA epigenome is globally stable at the genome-wide level (Ma *et al.*, 2009), “evidence suggests the presence of active DNA modifications at specific genomic loci and these modifications are critical for certain types of brain plasticity” (Day and Sweatt, 2010; Ma *et al.*, 2010). Because it is becoming evident that DNA methylation are a powerful epigenetic tool to modulate neuronal functions and synaptic plasticity, methylation is one of the most broadly studied and well- characterized epigenetic modifications in studies done by Griffith and Mahler who suggested that DNA methylation may be important in long term memory function (Bestor *et al.,* 2014). This evidence has been reinforced with the discovery of the Rett Syndrome, a neurological disorder linked to mutations in methyl-CpG-binding protein 2 (MeCP2) which implies that alterations to the methylome landscape might interfere also with the normal DNA binding of MeCP2 or other functions, leading to neurological symptoms. In this regard, one important area of research is to study samples of brain from peoples suffering from different neurological disorders in order to evaluate the methylome contribution for the condition, especially in an interactive approach where DNA variants, epigenetics and gene expression data could be studied together.

Current data indicates that hmC is substantially enriched in neurons compared with different cells and accounting for nearly 25% modified CG dinucleotides in the frontal cortex(Kinde *et al.*, 2015). Studies that examined the first two years of human development have reported that DNA methylation levels increase rapidly and then stabilize by adulthood in the brain. (Alisch *et al.*, 2012; Lister *et al.*, 2013). Also, the adult mammalian brain contains the highest levels of hmC that have been observed among tissues (Lister *et al.*, 2013) and is generally found near transcription start sites, making it essential for proper development. There is, so far, two roles assigned to this base: it works as an intermediate mark and/or plays a role as modulator of genomic function (Kinde *et al.*, 2015). Considering this, it would also be interesting, for future directions, to compare more samples from brain and blood that went through TAB seq or any other method capable to detect 5hmc to see the relative methylation fraction variation throughout aging and see how they differ between these tissues. Through this approach, we could determine that 5hmC also plays a functional role, particularly in brain.

Overall, in order to be more statistically relevant, biologically meaningful and avoid some caveats inherent to cross sectional studies, the ideal study on this subject would have to include more samples (perhaps through searching data from microarray sequencing).It would also be interesting, if possible, to use data from longitudinal studies in order to assess intraindividual changes in the methylation; although none of these kind have been found up to this study. Finally, to validate these results, I would recommend to run expression experiments such as mRNA-seq, especially in the case of the comparison between samples from both tissue samples (brain from 31-year-old from and blood from 26 year, given the mismatch of DMR between these tissues) to evaluate the difference of the transcriptome of these two tissues and to do a comparative analysis with their respective methylome.

6 FINAL REMARKS

Although we had some limitations in the beginning of our work, we were able to accomplish our goals.

Our main limitation was the difficulty in finding a good amount of well annotated samples. This was due to two main reasons: the first one had to do with the high throughput sequencing method used as query input, and the second had to do with the misinformation in metadata. Indeed, most of the studies done so far on DNA methylation used microarrays as the main sequencing method and not bisulfite-based NGS methods; and most of the samples missed the metadata regarding age, gender or health status, which were the main contributors to the DMRs, and the ones we were trying to assess. Because of this, we started our data set with 18 sample files from blood and 65 sample files from brain, and in the end we narrowed to only 3 (peripheral) blood sample files and 16 effectual brain sample files (shown in tables 2 and 3, respectively).

However, our goals have been accomplished and we could validate most of our initial hypothesis: we were able explore and learn how to use the main databases in order to obtain fastq files from methyl-seq and bisulphite-seq next generation sequencing reads; we validated our homemade protocol, and script to do both quality control and alignment of the reads, respectively with results that replicate what has been described in the literature. Finally, although we had a small dataset, we could still verify that the epigenome is modulated by age, both in brain and blood, regardless the gender (at least in healthy subjects).Furthermore the functions that could be affected are also relatable to aging, suggesting that our methodology can be used in the future to confirm general and particular features features associated to a healthy and pathological status, and thus to contribute to expand our knowledge about the role of the epigenome as a modulator of human health in every stage of life.

References

Aberg, K. A., Xie, L. Y., McClay, J. L., Nerella, S., Vunck, S., Snider, S., … van den Oord, E. J. C. G. (2013). Testing two models describing how methylome-wide studies in blood are informative for psychiatric conditions. Epigenomics, 5(4), 367–77. https://doi.org/10.2217/epi.13.36

Aberg, K. a, Mcclay, J. L., Nerella, S., Xie, L. Y., Clark, S. L., Hudson, A. D., … Magnusson, P. K. E. (2014). NIH Public Access, *4*(6), 605–621. https://doi.org/10.2217/epi.12.59.MBD-seq

Andrews, S. (2013). Reduced Representation Bisulfite-Seq – A Brief Guide to RRBS. *Babraham Bioinformatics*, 1–12. Retrieved from papers3://publication/uuid/A5B88C00-5F40-41F6-8A67-57BD85491A01

Auton, A., Abecasis, G. R., Altshuler, D. M., Durbin, R. M., Abecasis, G. R., Bentley, D. R., … Abecasis, G. R. (2015). A global reference for human genetic variation. *Nature*, *526*(7571), 68–74. https://doi.org/10.1038/nature15393

Bae, J.-B. (2013). Perspectives of international human epigenome consortium. *Genomics & Informatics*, *11*(1), 7–14. https://doi.org/10.5808/GI.2013.11.1.7

Balzer, S., Malde, K., Lanz??n, A., Sharma, A., & Jonassen, I. (2011). Characteristics of 454 pyrosequencing data-enabling realistic simulation with flowsim. *Bioinformatics*, *27*(13), i420–i425. https://doi.org/10.1093/bioinformatics/btq365

Bestor, T. H., Edwards, J. R., & Boulard, M. (2014). Notes on the role of dynamic DNA methylation in mammalian development. *Proceedings of the National Academy of Sciences of the United States of America*, *112*(22), 6796–6799. https://doi.org/10.1073/pnas.1415301111

Cheng, L., & Zhu, Y. (2013). Genome analysis A classification approach for DNA methylation profiling with bisulfite next-generation sequencing data, 1–8. https://doi.org/10.1093/bioinformatics/btt674

Cock, P. J. A., Fields, C. J., Goto, N., Heuer, M. L., & Rice, P. M. (2009). The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina

FASTQ variants. *Nucleic Acids Research*, *38*(6), 1767–1771. https://doi.org/10.1093/nar/gkp1137

Davies, M. N., Volta, M., Pidsley, R., Lunnon, K., Dixit, A., Lovestone, S., … Mill, J. (2012). Functional annotation of the human brain methylome identifies tissue-specific epigenetic variation across brain and blood. *Genome Biology*, *13*(6), R43. https://doi.org/10.1186/gb-2012-13-6-r43

Deaton, A., & Bird, A. (2011). CpG islands and the regulation of transcription. *Genes & Development*, *25*(10), 1010–1022. https://doi.org/10.1101/gad.2037511.1010

Estécio, M. R. H., Yan, P. S., Huang, T. H. M., & Issa, J. P. J. (2008). Methylated CpG island amplification and microarray (MCAM) for high-throughput analysis of DNA methylation. *Cold Spring Harbor Protocols*, *3*(3), 1–8. https://doi.org/10.1101/pdb.prot4974

Fraga, M. F. (2009). Genetic and epigenetic regulation of aging. *Current Opinion in Immunology*, *21*(4), 446–453. https://doi.org/10.1016/j.coi.2009.04.003

Fraga, M. F., Ballestar, E., Paz, M. F., Ropero, S., Setien, F., Ballestar, M. L., … Esteller, L. (2005). Epigenetic differences arise during the lifetime of monozygotic twins. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(30), 10604–9. https://doi.org/10.1073/pnas.0500398102

Gaunt, T. R., Shihab, H. A., Hemani, G., Min, J. L., Woodward, G., Lyttleton, O., … Relton, C. L. (2016). Systematic identification of genetic influences on methylation across the human life course. *Genome Biology*, *17*(1), 61. https://doi.org/10.1186/s13059-016-0926-z

Goodwin, S., McPherson, J. D., & McCombie, W. R. (2016). Coming of age: ten years of next-generation sequencing technologies. *Nat Rev Genet*, *17*(6), 333–351. https://doi.org/10.1038/nrg.2016.49

Hackenberg, M., Barturen, G., Oliver, J. L., Genética, D. De, & Ciencias, F. De. (2010). DNA Methylation Profiling from High-Throughput Sequencing Data.

Harris, R. A., Wang, T., Coarfa, C., Nagarajan, R. P., Hong, C., Sara, L., … Hirst, M.

(2011). NIH Public Access. *October*, *28*(10), 1097–1105. https://doi.org/10.1038/nbt.1682.Comparison

He, X.-J., Chen, T., & Zhu, J.-K. (2011). Regulation and function of DNA methylation in plants and animals. *Cell Research*, *54*(3), 442–465. https://doi.org/10.1038/cr.2011.23

Hirst, M., & Marra, M. A. (2010). Next generation sequencing based approaches to epigenomics. *Briefings in Functional Genomics*, *9*(5–6), 455–465. https://doi.org/10.1093/bfgp/elq035

Horton, J. R., Borgaro, J. G., Griggs, R. M., Quimby, A., Guan, S., Zhang, X., … Cheng, X. (2014). Structure of 5-hydroxymethylcytosine-specific restriction enzyme, AbaSI, in complex with DNA. *Nucleic Acids Research*, *42*(12), 7947–7959. https://doi.org/10.1093/nar/gku497

Huidobro, C., Fernandez, A. F., & Fraga, M. F. (2013). Aging epigenetics: Causes and consequences. *Molecular Aspects of Medicine*, *34*(4), 765–781. https://doi.org/10.1016/j.mam.2012.06.006

Jiang, P., Sun, K., Lun, F. M. F., Guo, A. M., Wang, H., Chan, K. C. A., … Sun, H. (2014). Methy-Pipe: An integrated bioinformatics pipeline for whole genome bisulfite sequencing data analysis. *PLoS ONE*, *9*(6). https://doi.org/10.1371/journal.pone.0100360

Jones, M. J., Goodman, S. J., & Kobor, M. S. (n.d.). DNA methylation and healthy human aging. https://doi.org/10.1111/acel.12349

Jones, M. J., Goodman, S. J., & Kobor, M. S. (2015). DNA methylation and healthy human aging. *Aging Cell*, *14*(6), 924–932. https://doi.org/10.1111/acel.12349

Jones, P. A. (2012). Functions of DNA methylation: islands, start sites, gene bodies and beyond. *Nature Reviews Genetics*, *13*(7), 484–492. https://doi.org/10.1038/nrg3230

Kanherkar, R. R., Bhatia-Dey, N., & Csoka, A. B. (2014). Epigenetics across the human lifespan. *Frontiers in Cell and Developmental Biology*, *2*(September), 49. https://doi.org/10.3389/fcell.2014.00049; 10.3389/fcell.2014.00049

Kim, J. K., Samaranayake, M., & Pradhan, S. (2009). Epigenetic mechanisms in mammals. *Cellular and Molecular Life Sciences*, *66*(4), 596–612. https://doi.org/10.1007/s00018-008-8432-4

Kinde, B., Gabel, H. W., Gilbert, C. S., Griffith, E. C., & Greenberg, M. E. (2015). Reading the unique DNA methylation landscape of the brain: Non-CpG methylation, hydroxymethylation, and MeCP2. *Proceedings of the National Academy of Sciences of the United States of America*, *112*(22), 6800–6. https://doi.org/10.1073/pnas.1411269112

Krueger, F. (2015). 1 Bisulfite-Sequencing theory and Quality Control, (January).

Krueger, F., & Andrews, S. R. (2012). Quality Control , trimming and alignment of Bisulfite-Seq data (Prot 57). *Practical Info NICE!!*, (July), 1–13.

Krueger, F., Kreck, B., Franke, A., & Andrews, S. R. (2012). DNA methylome analysis using short bisulfite sequencing data. *Bioinformatics*, *9*(2), 145–51. https://doi.org/10.1038/nmeth.1828

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., … International Human Genome Sequencing, C. (2001). Initial sequencing and analysis of the human genome. *Nature*, *409*(6822), 860–921. https://doi.org/10.1038/35057062

Lange, N. E., Sordillo, J., Tarantini, L., Bollati, V., Sparrow, D., Vokonas, P., … Demeo, D. L. (2012). Alu and LINE-1 methylation and lung function in the normative ageing study. *BMJ Open*, *2*(5), e001231-. https://doi.org/10.1136/bmjopen-2012-001231

Lee, S. H., & Dominguez, R. (2010). Regulation of actin cytoskeleton dynamics in cells. *Molecules and Cells*, *29*(4), 311–325. https://doi.org/10.1007/s10059-010-0053-8

Levine, M. E., Lu, A. T., Bennett, D. A., & Horvath, S. (2015). Epigenetic age of the pre-frontal cortex is associated with neuritic plaques, amyloid load, and Alzheimer's disease related cognitive functioning. *Aging*, *7*(12), 1198–1211. https://doi.org/10.18632/aging.100864

Lister, R., Mukamel, E. a, Nery, J. R., Urich, M., Puddifoot, C. a, Nicholas, D., … He, C. (2013). Global Epigenomic Reconfiguration During Mammalian Brain Development.

Science (New York, N.Y.), *341*(mC), 629–643. https://doi.org/10.1126/science.1237905

Lister, R., Pelizzola, M., Dowen, R. H., Hawkins, R. D., Hon, G., Tonti-Filippini, J., … Ecker, J. R. (2009). Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature*, *462*(7271), 315–22. https://doi.org/10.1038/nature08514

Liyanage, V. R., Jarmasz, J. S., Murugeshan, N., Del Bigio, M. R., Rastegar, M., & Davie, J. R. (2014). DNA modifications: function and applications in normal anddisease States. *Biology (Basel)*, *3*(4), 670–723.https://doi.org/10.3390/biology3040670

Lord, J., & Cruchaga, C. (2014). The epigenetic landscape of Alzheimer's disease. *Nature Neuroscience*, *17*(9), 1138–40. https://doi.org/10.1038/nn.3792

Macia, A., Munoz-Lopez, M., Cortes, J. L., Hastings, R. K., Morell, S., Lucena-Aguilar, G., … Garcia-Perez, J. L. (2011). Epigenetic control of retrotransposon expression in human embryonic stem cells. *Mol Cell Biol*, *31*(2), 300–316. https://doi.org/10.1128/MCB.00561-10

Manuscript, A. (2014). Sequencing, *340*(2), 1–16. https://doi.org/10.1016/j.canlet.2012.10.040.Analyzing

McGowan, P., Meaney, M., & Szyf, M. (2008). Diet and the epigenetic (re) programming of phenotypic differences in behavior. *Brain Research*, *1237*, 12–24. https://doi.org/10.1016/j.brainres.2008.07.074.Diet

Metzker, M. L. (2010). Sequencing technologies - the next generation. *Nature Reviews. Genetics*, *11*(1), 31–46. https://doi.org/10.1038/nrg2626

Miura, F., & Ito, T. (2015). Highly sensitive targeted methylome sequencing by post-bisulfite adaptor tagging. *DNA Research*, *22*(1), 13–18. https://doi.org/10.1093/dnares/dsu034

Munroe, D. J., & Harris, T. J. R. (2010). Third-generation sequencing fireworks at Marco Island. *Nature Biotechnology*, *28*(5), 426–428. https://doi.org/10.1038/nbt0510-426

Novik, K. L., Nimmrich, I., Genc, B., Maier, S., Piepenbrock, C., Olek, A., & Beck, S. (2002). Epigenomics: Genome-wide study of methylation phenomena. *Current Issues in Molecular Biology*, *4*(4), 111–128.

Numata, S., Ye, T., Hyde, T. M., Guitart-Navarro, X., Tao, R., Wininger, M., … Lipska, B. K. (2012). DNA methylation signatures in development and aging of the human prefrontal cortex. *American Journal of Human Genetics*, *90*(2), 260–272. https://doi.org/10.1016/j.ajhg.2011.12.020

Olkhov-Mitsel, E., & Bapat, B. (2012). Strategies for discovery and validation of methylated and hydroxymethylated DNA biomarkers. *Cancer Medicine*, *1*(2), 237–60. https://doi.org/10.1002/cam4.22

Petterson, A., Chung, T., Tan, D., Sun, X., & Jia, X.-Y. (2014). RRHP: A tag-based approach for 5-hydroxymethylcytosine mapping at single-site resolution. *Genome Biology*, *15*(9), 456. https://doi.org/10.1186/s13059-014-0456-5

Pomraning, K. R., Smith, K. M., & Freitag, M. (2009). Genome-wide high throughput analysis of DNA methylation in eukaryotes. *Methods*, *47*(3), 142–150. https://doi.org/10.1016/j.ymeth.2008.09.022

Romanoski, C. E., Glass, C. K., Stunnenberg, H. G., Wilson, L., & Almouzni, G. (2015). Epigenomics: Roadmap for regulation. *Nature*, *518*(7539), 314–316. https://doi.org/10.1038/518314a

Sanchez-Mut, J. V, Heyn, H., Vidal, E., Moran, S., Sayols, S., Delgado-Morales, R., … Esteller, M. (2016). Human DNA methylomes of neurodegenerative diseases show common epigenomic patterns. *Translational Psychiatry*, *6*(1), e718. https://doi.org/10.1038/tp.2015.214

Schadt, E. E., Turner, S., & Kasarskis, A. (2010). A window into third-generation sequencing. *Human Molecular Genetics*, *19*(R2), 227–240. https://doi.org/10.1093/hmg/ddq416

Schultz, M. D., He, Y., Whitaker, J. W., Hariharan, M., Mukamel, E. A., Leung, D., … Ecker, J. R. (2015). Human body epigenome maps reveal noncanonical DNA methylation variation. *Nature*, *523*(7559), 212–216. https://doi.org/10.1038/nature14465

Shakya, K., Connell, M. J. O., & Ruskin, H. J. (2012). The landscape for epigenetic / epigenomic biomedical resources. *Landes Bioscience*, *7*(9), 982–986.

Shakya, K., O'Connell, M. J., & Ruskin, H. J. (2012). The landscape for epigenetic/epigenomic biomedical resources. *Epigenetics*, *7*(9), 982–986. https://doi.org/10.4161/epi.21493

Shen, L., & Waterland, R. A. (2007). Methods of DNA methylation analysis. *Curr Opin Clin Nutr Metab Care*, *10*(5), 576–81. https://doi.org/10.1097/MCO.0b013e3282bf6f43

Su, J., Yan, H., Wei, Y., Liu, H., Liu, H., Wang, F., … Zhang, Y. (2013). CpG-MPs: Identification of CpG methylation patterns of genomic regions from high-throughput bisulfite sequencing data. *Nucleic Acids Research*, *41*(1), 1–15. https://doi.org/10.1093/nar/gks829

Teng, X., & Xiao, H. (2009). Perspectives of DNA microarray and next-generation DNA sequencing technologies. *Science in China, Series C: Life Sciences*, *52*(1), 7–16. https://doi.org/10.1007/s11427-009-0012-9

Teschendorff, A. E., West, J., & Beck, S. (n.d.). Age-associated epigenetic drift: implications, and a case of epigenetic thrift? https://doi.org/10.1093/hmg/ddt375

The fraction of methylated C (%) Sequence context Watson Crick. (n.d.), 40.

Tognini, P., Napoli, D., & Pizzorusso, T. (2015). Dynamic DNA methylation in the brain: a new epigenetic mark for experience-dependent plasticity. *Frontiers in Cellular Neuroscience*, *9*(August), 331. https://doi.org/10.3389/fncel.2015.00331

Tollefsbol, T. O. (2004). Epigenetics Protocols. *Platelets*, *287*(18), 316. https://doi.org/10.4081/ejh.2012.br8

van Dongen, J., Nivard, M. G., Willemsen, G., Hottenga, J.-J., Helmer, Q., Dolan, C. V, … Boomsma, D. I. (2016). Genetic and environmental influences interact with age and sex in shaping the human methylome. *Nature Communications*, *7*, 11115. https://doi.org/10.1038/ncomms11115

Weidner, C. I., & Wagner, W. (2014). The epigenetic tracks of aging. *Biological Chemistry*, *395*(11), 1307–1314. https://doi.org/10.1515/hsz-2014-0180

Winnefeld, M., & Lyko, F. (2012). The aging epigenome: DNA methylation from the cradle to the grave. *Genome Biology*, *13*(7), 165. https://doi.org/10.1186/gb4033

Wood, J. G., Jones, B. C., Jiang, N., Chang, C., Hosier, S., Wickremesinghe, P., … Helfand, S. L. (2016). Chromatin-modifying genetic interventions suppress age-associated transposable element activation and extend life span in *Drosophila*. *Proceedings of the National Academy of Sciences*, 201604621. https://doi.org/10.1073/pnas.1604621113

Yu, M., Hon, G. C., Szulwach, K. E., Song, C. X., Zhang, L., Kim, A., … He, C. (2012). Base-resolution analysis of 5-hydroxymethylcytosine in the mammalian genome. *Cell*, *149*(6), 1368–1380. https://doi.org/10.1016/j.cell.2012.04.027

Zhang, H., Wang, F., Kranzler, H. R., Yang, C., Xu, H., Wang, Z., … Gelernter, J. (2014). Identification of methylation quantitative trait loci (mQTLs) influencing promoter DNA methylation of alcohol dependence risk genes. *Human Genetics*, *133*(9), 1093–1104. https://doi.org/10.1007/s00439-014-1452-2

APPENDIX

R code to search for brain (Pre Frontal Cortex) sample files

```
library(GEOmetadb)
if(!file.exists('GEOmetadb.sqlite')) getSQLiteFile()
con <- dbConnect(SQLite().'GEOmetadb.sqlite')

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 'Homo%' AND
             gpl.organism LIKE 'Homo%' AND
             gsm.type LIKE '%SRA%'AND
             gsm.source_name_ch1 LIKE '%cerebral cortex%'". sep="
")


data1 <- dbGetQuery(con.sql)
```

```
library(xlsx)
write.xlsx(data1. file="datasource.xlsx")
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 'Homo%' AND
             gpl.organism LIKE 'Homo%' AND
             gsm.type LIKE '%SRA%'AND
             gsm.characteristics_ch1 LIKE '%cerebral cortex%'".
sep=" ")

data2 <- dbGetQuery(con.sql)

library(xlsx)
write.xlsx(data2. file="datachara.xlsx")
-----------------------------------------------------------
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 'Homo%' AND
              gpl.organism LIKE 'Homo%' AND
              gsm.type LIKE '%SRA%'AND
              gsm.characteristics_ch1 LIKE '%frontal cortex%'".
sep=" ")
```

```
data <- dbGetQuery(con.sql)
library(xlsx)
write.xlsx (data. file="chara.xlsx")
```

R CODE USED TO SERCH FOR PERIPHERAL BLOOD SAMPLES

```
library(GEOmetadb)
if(file.exists('GEOmetadb.sqlite')) getSQLiteFile()
con <- dbConnect(SQLite().'GEOmetadb.sqlite')

####DATA WHOLE BLOOD####
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%'

gsm.organism_ch1 LIKE 'Homo%' AND gpl.organism
LIKE 'Homo%' AND gsm.type LIKE '%SRA%'AND



              gsm.characteristics_ch1 LIKE '%whole blood%'". sep="
")
data <- dbGetQuery(con.sql)
```

R code for peripheralblood

```
library(GEOmetadb) #acess geometadb functions#
if(!file.exists('GEOmetadb.sqlite')) getSQLiteFile()
con <- dbConnect(SQLite().'GEOmetadb.sqlite') #connect to the
database#
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". #joins the tables gse. gsm e gpl keeping the
common fields#
             "  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". #filters the fields with the respective#
             "gse.type LIKE '%Methylation profiling by high
throughput sequencing%' AND
              gpl.technology LIKE '%high-throughput sequencing%'
AND
              gsm.organism_ch1 LIKE 'Homo%' AND
              gpl.organism LIKE 'Homo%' AND
              gsm.type LIKE '%SRA%'AND
              gsm.characteristics_ch1 LIKE '%peripheral blood%'".
sep=" ")
```

```
data <- dbGetQuery(con.sql) #devolve os resultados da filtragem na
BD#
library(xlsx)
write.xlsx(data. file="datachara.xlsx")
#this will writ the outcome in an excel file#


library(GEOquery)
#peripheral blood#
geo <- c('GSM…'.'GSM…'.…)
 for (i in 1:length(geo)){
    getGEO(geo[i])
 }
#this code will download metadata of the FASTQ files using GSM
id’s#
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 'Homo%' AND
              gpl.organism LIKE 'Homo%' AND
              gsm.type LIKE '%SRA%'AND
              gsm.source_name_ch1 LIKE '%peripheral blood%'".
sep=" ")
```

```
data <- dbGetQuery(con.sql)
library (SRAdb)
sra_dbname <- 'SRAmetadb.sqlite'
sra_con <- dbConnect (dbDriver("SQLite"). sra_dbname )
res <- dbGetQuery(sra_con. "select run_accession.
experiment_accession. sample_alias. submission_accession.


study_name from sra_ft where experiment_accession in
('SRX….'SRX….'.….)")
#The first three lines of the code give access to the SRA database
and the last one retrieves  a table from sra_ft with the mentioned
fields from dbgetuery function with the corresponding experiment
accessions#
library (xlsx)
write.xlsx (res. file="ids.xlsx")

getFASTQfile( in_acc = c('SRR…'.'SRR…'.…). sra_con. destDir =
getwd(). srcType = 'ftp')
----------------------------------------------------------------
   library(GEOmetadb)
if(!file.exists('GEOmetadb.sqlite')) getSQLiteFile()
con <- dbConnect(SQLite().'GEOmetadb.sqlite')
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           ")
```

```
              m.organism_ch1 LIKE 'Homo%' AND
              gpl.organism LIKE 'Homo%' AND
g
              gsm.type LIKE '%SRA%'AND
s
              gsm.source_name_ch1 LIKE '%frontal cortex%'". sep="
data <- dbGetQuery(con.sql)
```

```
Library (xlsx)
 write.xlsx(data. file="datasource.xlsx")
```

R code to search for brain (Pre Frontal Cortex) sample files

```
library(GEOmetadb)

if(!file.exists('GEOmetadb.sqlite')) getSQLiteFile()

con <- dbConnect(SQLite().'GEOmetadb.sqlite')


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 'Homo%' AND

             gpl.organism LIKE 'Homo%' AND

             gsm.type LIKE '%SRA%'AND
```

```
            gsm.source_name_ch1 LIKE '%cerebral cortex%'".  sep="
")

data1 <- dbGetQuery(con.sql)


library(xlsx)

write.xlsx(data1. file="datasource.xlsx")

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 'Homo%' AND

            gpl.organism LIKE 'Homo%' AND

            gsm.type LIKE '%SRA%'AND

            gsm.characteristics_ch1  LIKE  '%cerebral  cortex%'".
sep=" ")

data2 <- dbGetQuery(con.sql)
```

R code to search for brain (Pre Frontal Cortex) sample files

```
library(GEOmetadb)
if(!file.exists('GEOmetadb.sqlite')) getSQLiteFile()
con <- dbConnect(SQLite().'GEOmetadb.sqlite')

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 'Homo%' AND
             gpl.organism LIKE 'Homo%' AND
             gsm.type LIKE '%SRA%'AND
             gsm.source_name_ch1 LIKE '%cerebral cortex%'". sep="
")

data1 <- dbGetQuery(con.sql)
```

```
library(xlsx)
write.xlsx(data1. file="datasource.xlsx")
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 'Homo%' AND
             gpl.organism LIKE 'Homo%' AND
             gsm.type LIKE '%SRA%'AND
             gsm.characteristics_ch1 LIKE '%cerebral cortex%'".
sep=" ")

data2 <- dbGetQuery(con.sql)

library(xlsx)
write.xlsx(data2. file="datachara.xlsx")
------------------------------------------------------------
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 'Homo%' AND
              gpl.organism LIKE 'Homo%' AND
              gsm.type LIKE '%SRA%'AND
              gsm.characteristics_ch1 LIKE '%frontal cortex%'".
sep=" ")



data <- dbGetQuery(con.sql)
library(xlsx)
write.xlsx (data. file="chara.xlsx")
```

R CODE USED TO SERCH PERIPHERAL BLOOD SAMPLES

```
library(GEOmetadb)
if(file.exists('GEOmetadb.sqlite')) getSQLiteFile()
con <- dbConnect(SQLite().'GEOmetadb.sqlite')

####DATA WHOLE BLOOD####
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 'Homo%' AND
              gpl.organism LIKE 'Homo%' AND
              gsm.type LIKE '%SRA%'AND
              gsm.characteristics_ch1 LIKE '%whole blood%'". sep="
")
```

```
data <- dbGetQuery(con.sql)
```

R code for peripheralblood

```
library(GEOmetadb) #acess geometadb functions#
if(!file.exists('GEOmetadb.sqlite')) getSQLiteFile()
con <- dbConnect(SQLite().'GEOmetadb.sqlite') #connect to the
database#
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". #joins the tables gse. gsm e gpl keeping the
common fields#
             "  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". #filters the fields with the respective#
             "gse.type LIKE '%Methylation profiling by high
throughput sequencing%' AND
              gpl.technology LIKE '%high-throughput sequencing%'
AND
              gsm.organism_ch1 LIKE 'Homo%' AND
              gpl.organism LIKE 'Homo%' AND
              gsm.type LIKE '%SRA%'AND
              gsm.characteristics_ch1 LIKE '%peripheral blood%'".
sep=" ")


data <- dbGetQuery(con.sql) #devolve os resultados da filtragem na
BD#
library(xlsx)
write.xlsx(data. file="datachara.xlsx")
#escreve os resultados num ficheiro excel#
library(GEOquery)
#peripheral blood#
geo <- c('GSM…'.'GSM…'.…)
 for (i in 1:length(geo)){
    getGEO(geo[i])
 }
#this code will download metadata of the FASTQ files using GSM
id’s#
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 'Homo%' AND
              gpl.organism LIKE 'Homo%' AND
              gsm.type LIKE '%SRA%'AND
              gsm.source_name_ch1 LIKE '%peripheral blood%'".
sep=" ")


----------------------------------------------------------------
data <- dbGetQuery(con.sql)
library (SRAdb)
sra_dbname <- 'SRAmetadb.sqlite'
sra_con <- dbConnect (dbDriver("SQLite"). sra_dbname )
res <- dbGetQuery(sra_con. "select run_accession.
experiment_accession. sample_alias. submission_accession.
study_name from sra_ft where experiment_accession in
('SRX….'SRX….'.….)")
#The first three lines of the code give acess to the SRA database
and the last one retrieves a table from sra_ft with the mentioned
```

```
fields from dbgetuery function with the corresponding experiment
accessions#
library (xlsx)
write.xlsx (res. file="ids.xlsx")

getFASTQfile( in_acc = c('SRR…'.'SRR…'.…). sra_con. destDir =
getwd(). srcType = 'ftp')
----------------------------------------------------------------
   library(GEOmetadb)
if(!file.exists('GEOmetadb.sqlite')) getSQLiteFile()
con <- dbConnect(SQLite().'GEOmetadb.sqlite')
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 'Homo%' AND
              gpl.organism LIKE 'Homo%' AND
              gsm.type LIKE '%SRA%'AND
              gsm.source_name_ch1 LIKE '%frontal cortex%'". sep="
")


data <- dbGetQuery(con.sql)


----------------------------------------------------------------
Library (xlsx)
write.xlsx(data. file="datasource.xlsx")
```

R CODE TO DOWNLOAD FASTQ FILES FROM THE DATABASES

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

biocLite("SRAdb")

browseVignettes("SRAdb")


library(SRAdb)

sqlfile <- 'SRAmetadb.sqlite'

if(!file.exists('SRAmetadb.sqlite')) sqlfile <<- getSRAdbFile()

sra_con <- dbConnect(SQLite().sqlfile)

list<-c('SRX’.…')

getSRAfile( in_acc = (list). sra_con = sra_con. destDir = getwd().
fileType = 'fastq')
```

Table 13 – Differentially methylated autosomal genes with gender.

gene	location	FUNCTION
ESR1	Start (pbs):151,656,691 End (pbs):152,129,619 Chromosome:6	This gene encodes an estrogen receptor, a ligand-activated transcription factor composed of several domains important for hormone binding, DNA binding, and activation of transcription. The protein localizes to the nucleus where it may form a homodimer or a heterodimer with estrogen receptor 2. Estrogen and its receptors are essential for sexual development and reproductive function, but also play a role in other tissues such as bone. Estrogen receptors are also involved in pathological processes including breast cancer, endometrial cancer, and osteoporosis. Alternative promoter usage and alternative splicing result in dozens of transcript variants, but the full-length nature of many of these variants has not been determined. [provided by RefSeq, Mar 2014]
MTHFR	Start:11,785,723 bp End:11,806,920 bp Chromosome:1	MTHFR (Methylenetetrahydrofolate Reductase) is a Protein Coding gene. The protein encoded by this gene catalyzes the conversion of 5,10-methylenetetrahydrofolate to 5-methyltetrahydrofolate, a co-substrate for homocysteine remethylation to methionine. Genetic variation in this gene influences susceptibility to occlusive vascular disease, neural tube defects, colon cancer and acute leukemia, and mutations in this

		gene are associated with methylenetetrahydrofolate reductase deficiency.[provided by RefSeq, Oct 2009]
CALCA	calcitonin related polypeptide	This gene encodes the peptide hormones calcitonin, calcitonin gene-related peptide and katacalcin by tissue-specific alternative RNA splicing of the gene transcripts and cleavage of inactive precursor proteins. Calcitonin is involved in calcium regulation and acts to regulate phosphorus metabolism. Calcitonin gene-related peptide functions as a vasodilator and as an antimicrobial peptide while katacalcin is a calcium-lowering peptide. Multiple transcript variants encoding different isoforms have been found for this gene.[provided by RefSeq, Aug 2014]
MGMT	Start(pbs):129,467,184 End(pbs):129,768,042 Chromosome:10	MGMT (O-6-Methylguanine-DNA Methyltransferase) is a Protein Coding gene. Diseases associated with MGMT include Spinal Cord Astrocytoma and Glioblastoma. Among its related pathways are p53 Pathway (RnD) and DNA Damage Reversal. The protein encoded by this gene is a DNA repair protein that is involved in cellular defense against mutagenesis and toxicity from alkylating agents. The protein catalyzes transfer of methyl groups from O(6)-alkylguanine and other methylated moieties of the DNA to its own molecule, which repairs the toxic lesions. Methylation of the genes promoter has been associated with several cancer types, including colorectal cancer, lung cancer, lymphoma and glioblastoma. [provided by RefSeq, Sep 2015]

Table 14 - Differentially methylated related to healthy and pathologic aging.

GENE	LOCATION	FUNCTION
MGMT	Start(pbs):129,467,184 End(pbs):129,768,042 Chromosome:10	MGMT (O-6-Methylguanine-DNA Methyltransferase) is a Protein Coding gene. Diseases associated with MGMT include Spinal Cord Astrocytoma and Glioblastoma. Among its related pathways are p53 Pathway (RnD) and DNA Damage Reversal. The protein encoded by this gene is a DNA repair protein that is involved in cellular defense against mutagenesis and toxicity from alkylating agents. The protein catalyzes transfer of methyl groups from O(6)-alkylguanine and other methylated moieties of the DNA to its own molecule, which repairs the toxic lesions. Methylation of the genes promoter has been associated with several cancer types, including colorectal cancer, lung cancer, lymphoma and glioblastoma. [provided by RefSeq, Sep 2015]
ESR1	Start (pbs):151,656,691 End (pbs):152,129,619 Chromossome:6	This gene encodes an estrogen receptor, a ligand-activated transcription factor composed of several domains important for hormone binding, DNA binding, and activation of transcription. The protein localizes to the nucleus where it may form a homodimer or a heterodimer with estrogen receptor 2. Estrogen and its receptors are essential for sexual development and reproductive function, but also play a role in other tissues such as bone. Estrogen receptors are also involved in pathological processes including breast cancer, endometrial cancer, and osteoporosis. Alternative promoter usage and alternative splicing result in dozens of transcript variants, but the full-length nature of many of these variants has not been determined. [provided by RefSeq, Mar 2014]

RASSF1	Start (pbs):50,329,782 End(pbs):50,340,980 Chromosome:3	This gene encodes a protein similar to the RAS effector proteins. Loss or altered expression of this gene has been associated with the pathogenesis of a variety of cancers, which suggests the tumor suppressor function of this gene. The inactivation of this gene was found to be correlated with the hypermethylation of its CpG-island promoter region. The encoded protein was found to interact with DNA repair protein XPA. The protein was also shown to inhibit the accumulation of cyclin D1, and thus induce cell cycle arrest. Several alternatively spliced transcript variants of this gene encoding distinct isoforms have been reported. [provided by RefSeq, May 2011]
RAD50	Start (pbs):132,556,019 End (pbs):132,646,344 Chromosome:5	The protein encoded by this gene is highly similar to Saccharomyces cerevisiae Rad50, a protein involved in DNA double-strand break repair. This protein forms a complex with MRE11 and NBS1. The protein complex binds to DNA and displays numerous enzymatic activities that are required for nonhomologous joining of DNA ends. This protein, cooperating with its partners, is important for DNA double-strand break repair, cell cycle checkpoint activation, telomere maintenance, and meiotic recombination. Knockout studies of the mouse homolog suggest this gene is essential for cell growth and viability. Mutations in this gene are the cause of Nijmegen breakage syndrome-like disorder.[provided by RefSeq, Apr 2010]
GSTP1	Start (pbs):67,583,595 End (pbs):67,586,660 Chromosome: 11	Glutathione S-Transferase Pi 1 is a Protein Coding gene. Glutathione S-transferases (GSTs) are a family of enzymes that play an important role in detoxification by catalyzing the conjugation of many hydrophobic and electrophilic compounds with reduced glutathione. Based on their biochemical, immunologic, and structural properties, the soluble GSTs are categorized into 4 main classes: alpha, mu, pi, and theta. This GST family member is a polymorphic gene encoding active, functionally different GSTP1 variant proteins that are thought to function in xenobiotic metabolism and play a role in susceptibility to cancer, and other diseases.
RARB	Start (pbs):25,174,332 End (pbs):25,597,932 Chromosome: 3	This gene encodes retinoic acid receptor beta, a member of the thyroid-steroid hormone receptor superfamily of nuclear transcriptional regulators. This receptor localizes to the cytoplasm and to subnuclear compartments. It binds retinoic acid, the biologically active form of vitamin A which mediates cellular signalling in embryonic morphogenesis, cell growth and differentiation. It is thought that this protein limits growth of many cell types by regulating gene expression. The gene was first identified in a hepatocellular carcinoma where it flanks a hepatitis B virus integration site. Alternate promoter usage and differential splicing result in multiple transcript variants. [provided by RefSeq, Mar 2014]
MYOD1	Start (pbs):17,719,563 End (pbs):17,722,131 Chromosome: 11	This gene encodes a nuclear protein that belongs to the basic helix-loop-helix family of transcription factors and the myogenic factors subfamily. It regulates muscle cell differentiation by inducing cell cycle arrest, a prerequisite for myogenic initiation. The protein is also involved in muscle

		regeneration. It activates its own transcription which may stabilize commitment to myogenesis. [provided by RefSeq, Jul 2008]
LAMB1	Start (pbs):107,923,799 End (pbs):108,003,359 Chromosome :7	Laminin Subunit Beta 1 is a Protein Coding gene. This gene encodes the beta chain isoform laminin, beta 1. The beta 1 chain has 7 structurally distinct domains which it shares with other beta chain isomers. The C- terminal helical region containing domains I and II are separated by domain alpha, domains III and V contain several EGF-like repeats, and domains IV and VI have a globular conformation. Laminin, beta 1 is expressed in most tissues that produce basement membranes, and is one of the 3 chains constituting laminin 1, the first laminin isolated from Engelbreth-Holm- Swarm (EHS) tumor. A sequence in the beta 1 chain that is involved in cell attachment, chemotaxis, and binding to the laminin receptor was identified and shown to have the capacity to inhibit metastasis. [provided by RefSeq, Aug 2011]
WRN	Start (pbs):31,033,262 End (pbs): 31,173,769 Chromosome:8	This gene encodes a member of the RecQ subfamily and the DEAH (Asp- Glu-Ala-His) subfamily of DNA and RNA helicases. DNA helicases are involved in many aspects of DNA metabolism, including transcription, replication, recombination, and repair. This protein contains a nuclear localization signal in the C-terminus and shows a predominant nucleolar localization. It possesses an intrinsic 3' to 5' DNA helicase activity, and is also a 3' to 5' exonuclease. Based on interactions between this protein and Ku70/80 heterodimer in DNA end processing, this protein may be involved in the repair of double strand DNA breaks. Defects in this gene are the cause of Werner syndrome, an autosomal recessive disorder characterized by premature aging. [provided by RefSeq, Jul 2008]
DLG4	Start (pbs):7,189,890 End (pbs):7,220,050 Chromosome:17	Discs Large MAGUK Scaffold Protein 4) is a Protein Coding gene. This gene encodes a member of the membrane-associated guanylate kinase (MAGUK) family. It heteromultimerizes with another MAGUK protein, DLG2, and is recruited into NMDA receptor and potassium channel clusters. These two MAGUK proteins may interact at postsynaptic sites to form a multimeric scaffold for the clustering of receptors, ion channels, and associated signaling proteins. Multiple transcript variants encoding different isoforms have been found for this gene. [provided by RefSeq, Jul 2008]
DRD2	Start (pbs):113,409,595 End (pbs):113,475,691 Chromosome 11	Dopamine Receptor D2 gene encodes the D2 subtype of the dopamine receptor. This G-protein coupled receptor inhibits adenylyl cyclase activity. A missense mutation in this gene causes myoclonus dystonia; other mutations have been associated with schizophrenia. Alternative splicing of this gene results in two transcript variants encoding different isoforms. A third variant has been described, but it has not been determined whether this form is normal or due to aberrant splicing. [provided by RefSeq, Jul 2008]
NOS1	Start(pbs):17,208,142 End (pbs):117,452,170 Chromosome 12	The protein encoded by this gene belongs to the family of nitric oxide synthases, which synthesize nitric oxide from L-arginine. Nitric oxide is a reactive free radical, which acts as a biologic mediator in several processes, including neurotransmission, and antimicrobial and antitumoral activities. In the brain and peripheral nervous system, nitric oxide displays many properties of a neurotransmitter, and has been implicated in neurotoxicity associated with stroke and neurodegenerative diseases, neural regulation of smooth muscle, including peristalsis, and penile erection. This protein is

		ubiquitously expressed, with high level of expression in skeletal muscle. Multiple transcript variants that differ in the 5' UTR have been described for this gene but the full-length nature of these transcripts is not known. Additionally, alternatively spliced transcript variants encoding different isoforms (some testis-specific) have been found for this gene.[provided by RefSeq, Feb 2011]
NRXN1	Start (pbs):49,918,505 End (pbs):51,225,575 Chromosome:2	This gene encodes a single-pass type I membrane protein that belongs to the neurexin family. Neurexins are cell-surface receptors that bind neuroligins to form Ca(2+)-dependent neurexin/neuroligin complexes at synapses in the central nervous system. This complex is required for efficient neurotransmission and is involved in the formation of synaptic contacts. Three members of this gene family have been studied in detail and are estimated to generate over 3,000 variants through the use of two alternative promoters (alpha and beta) and extensive alternative splicing in each family member. Recently, a third promoter (gamma) was identified for this gene in the 3' region. Mutations in this gene are associated with Pitt- Hopkins-like syndrome-2 and may contribute to susceptibility to schizophrenia. [provided by RefSeq, Aug 2016]
SOX10	Start (pbs):37,970,686 End (pbs):37,987,422 Chromosome:22	This gene encodes a member of the SOX (SRY-related HMG-box) family of transcription factors involved in the regulation of embryonic development and in the determination of the cell fate. The encoded protein may act as a transcriptional activator after forming a protein complex with other proteins. This protein acts as a nucleocytoplasmic shuttle protein and is important for neural crest and peripheral nervous system development. Mutations in this gene are associated with Waardenburg-Shah and Waardenburg-Hirschsprung disease. [provided by RefSeq, Jul 2008]

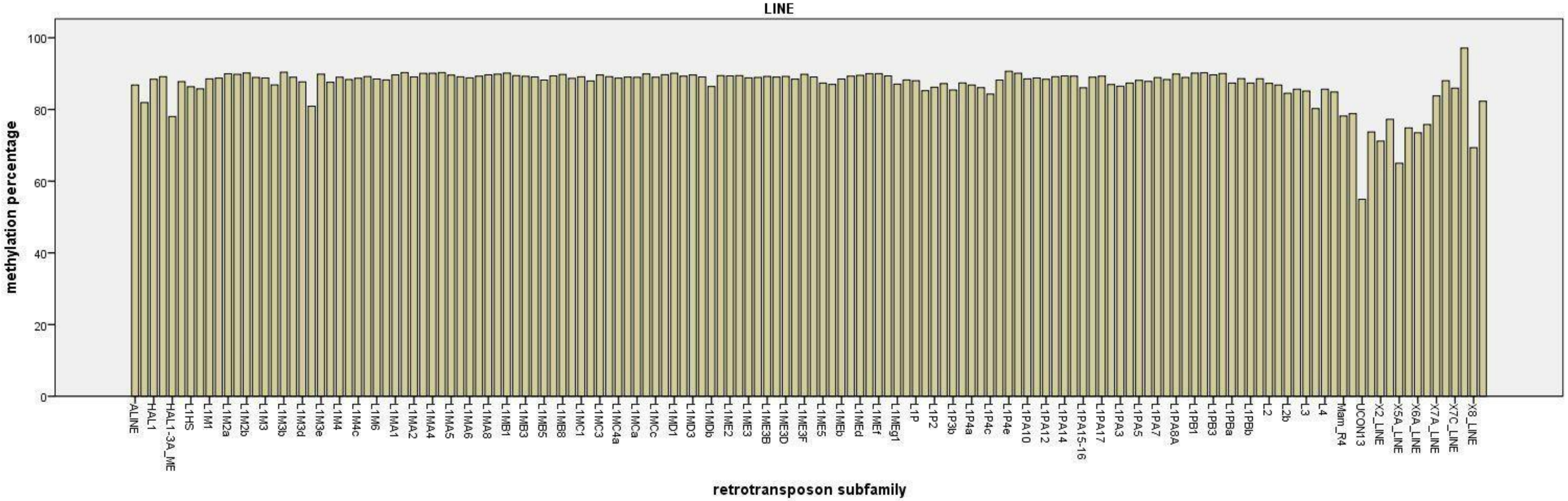
Table 15- Differentially methylated genes with Alzheimer's Disease

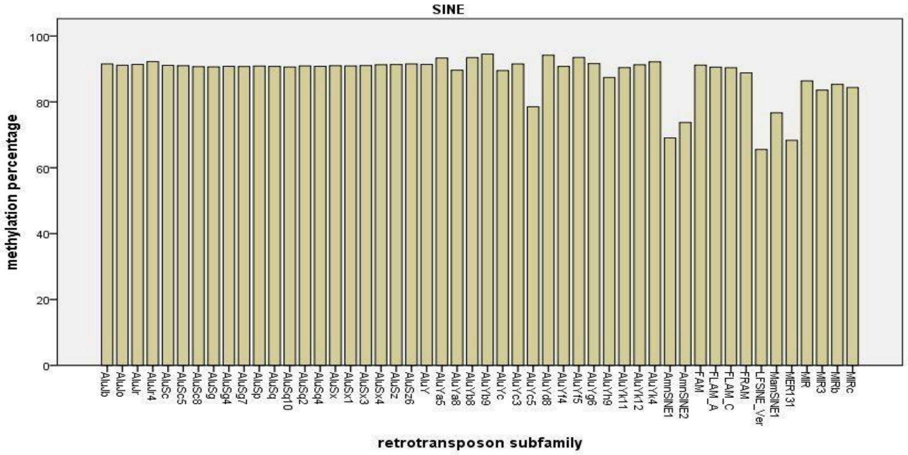
gene	location	function
ANK1	Start (pbs):41,653,220 End (pbs):41,896,762 Chromosome:8	ANK1 (Ankyrin 1) is a Protein Coding gene. Attaches integral membrane proteins to cytoskeletal elements; binds to the erythrocyte membrane protein band 4.2, to Na-K ATPase, to the lymphocyte membrane protein GP85, and to the cytoskeletal proteins fodrin, tubulin, vimentin and desmin. Erythrocyte ankyrins also link spectrin (beta chain) to the cytoplasmic domain of the erythrocytes anion exchange protein; they retain most or all of these binding functions. Isoform Mu17 together with obscurin in skeletal muscle may provide a molecular link between the sarcoplasmic reticulum and myofibrils
RPL13	Start (pbs):89,560,657 End:89,566,829 Chromosome 16	RPL13 (Ribosomal Protein L13) is a Protein Coding gene. The protein belongs to the L13E family of ribosomal proteins. It is located in the cytoplasm. This gene is

		expressed at significantly higher levels in benign breast lesions than in breast carcinomas. Alternatively spliced transcript variants encoding distinct isoforms have been found for this gene. As is typical for genes encoding ribosomal proteins, there are multiple processed pseudogenes of this gene dispersed through the genome.
RHBDF2	Start (pbs):76,470,891 End (pbs):76,501,790 Chromosome:17	RHBDF2 (Rhomboid 5 Homolog 2) is a Protein Coding gene. Rhomboid protease-like protein which has no protease activity but regulates the secretion of several ligands of the epidermal growth factor receptor. Indirectly activates the epidermal growth factor receptor signaling pathway and may thereby regulate sleep, cell survival, proliferation and migration (By similarity).
CDH23	Start (pbs):71,396,934 End (pbs):71,815,947 Chromosome: 10	This gene is a member of the cadherin superfamily, whose genes encode calcium dependent cell-cell adhesion glycoproteins. The encoded protein is thought to be involved in stereocilia organization and hair bundle formation. The gene is located in a region containing the human deafness loci DFNB12 and USH1D. Usher syndrome 1D and nonsyndromic autosomal recessive deafness DFNB12 are caused by allelic mutations of this cadherin-like gene. Upregulation of this gene may also be associated with breast cancer. Alternative splice variants encoding different isoforms have been described. [provided by RefSeq, May 2013]
ABCA7	Start (pbs):1,040,101 End (pbs):1,065,572 Chromosome: 19	ABCA7 (ATP Binding Cassette Subfamily A Member 7) is a Protein Coding gene. Plays a role in phagocytosis by macrophages of apoptotic cells. Binds APOA1 and may function in apolipoprotein-mediated phospholipid efflux from cells. May also mediate cholesterol efflux. May regulate cellular ceramide homeostasis during keratinocytes differentiation.
BIN1	Start (pbs):127,048,023 End (pbs) :127,107,400 Chromosome:2	BIN1 (Bridging Integrator 1) is a Protein Coding gene. This gene encodes several isoforms of a nucleocytoplasmic adaptor protein, one of which was initially identified as a MYC-interacting protein with features of a tumor suppressor. Isoforms that are expressed in the central nervous system may be involved in synaptic vesicle endocytosis and may interact with dynamin, synaptojanin, endophilin, and clathrin. May act as a tumor suppressor and inhibits malignant cell transformation.

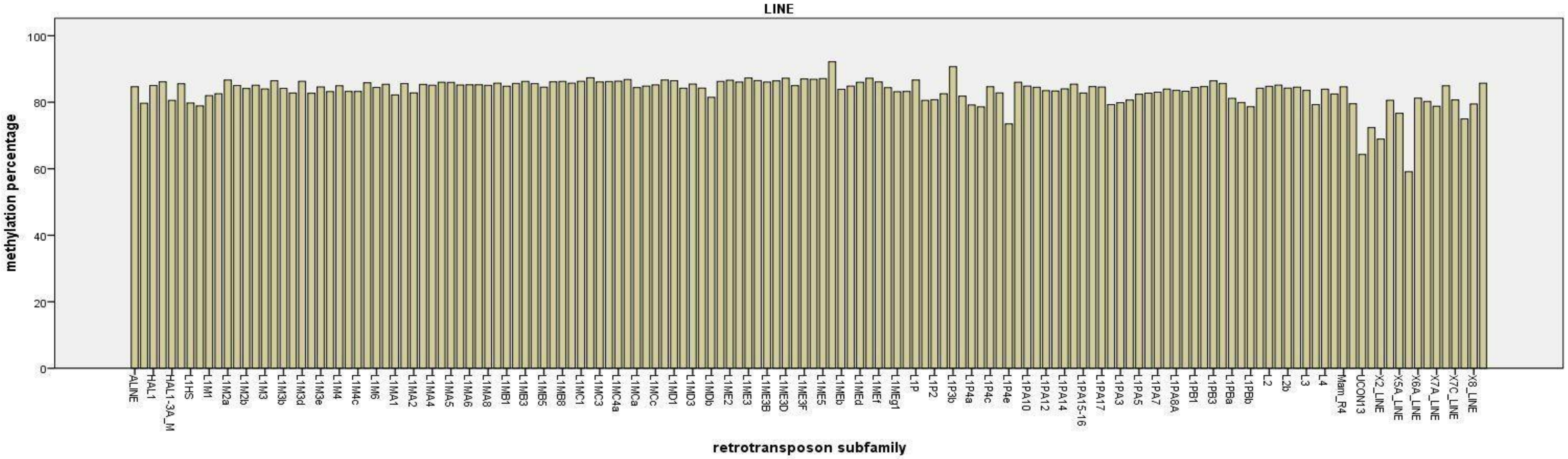
Plots generated from Regional Methylation Density Calculation Tables for Brain samples (male and female)

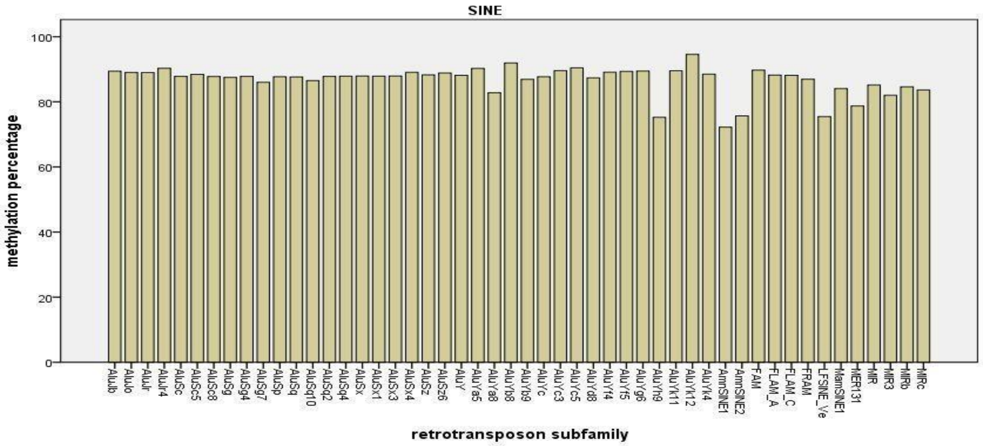
SRR901381 (20 weeks)



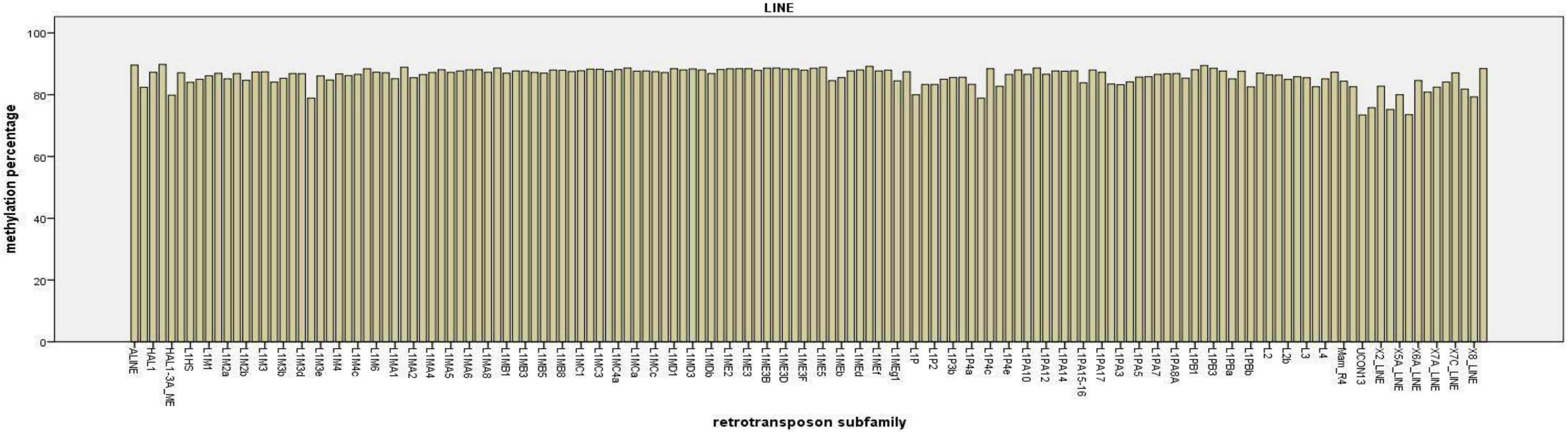


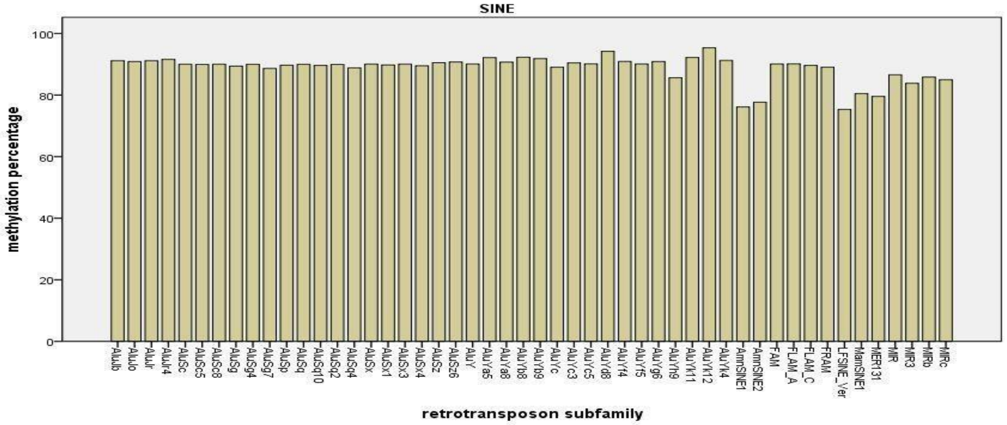
SRR479000 (31 years)



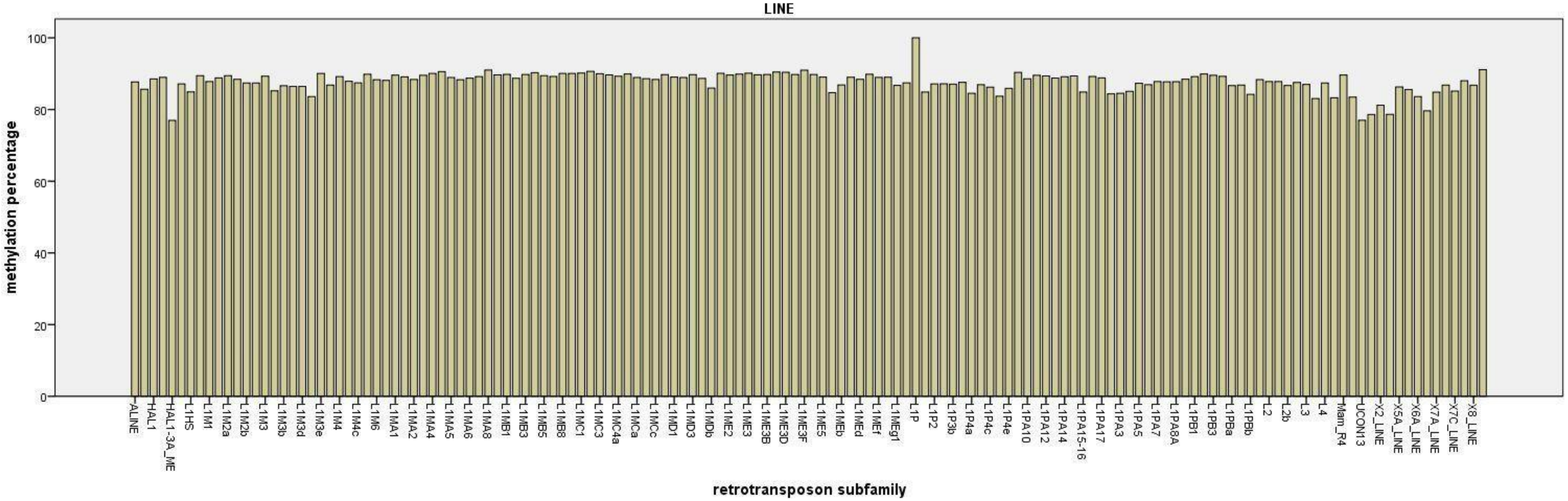


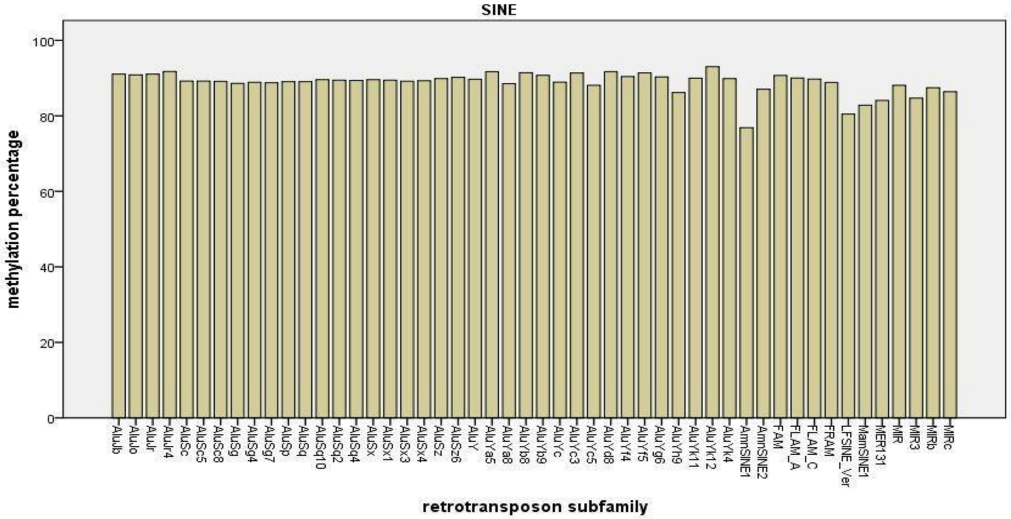
SRR478994 (47 year old male)



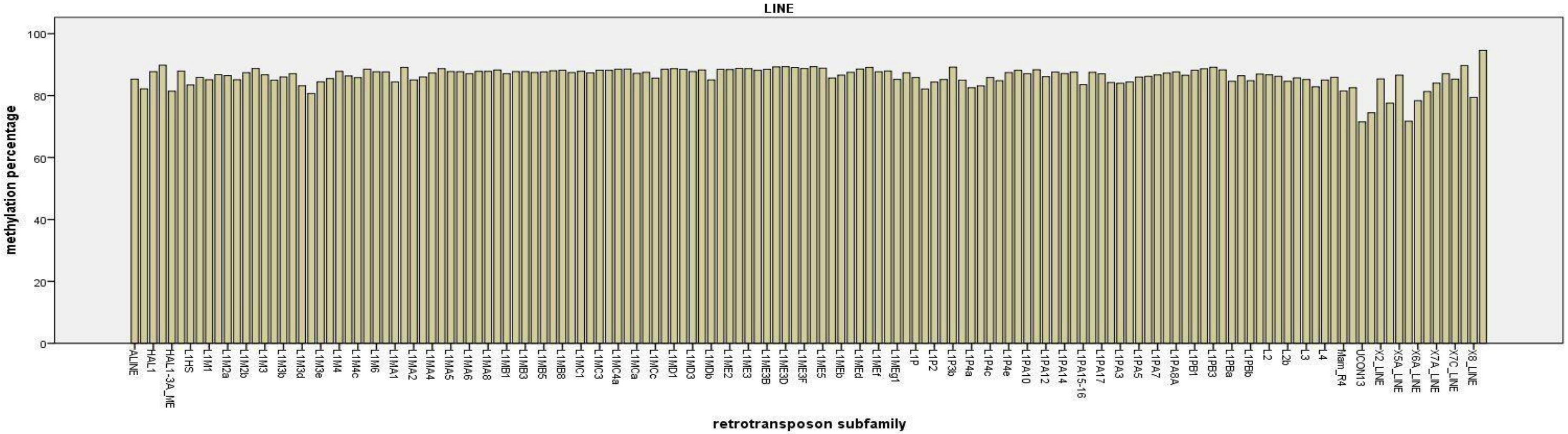


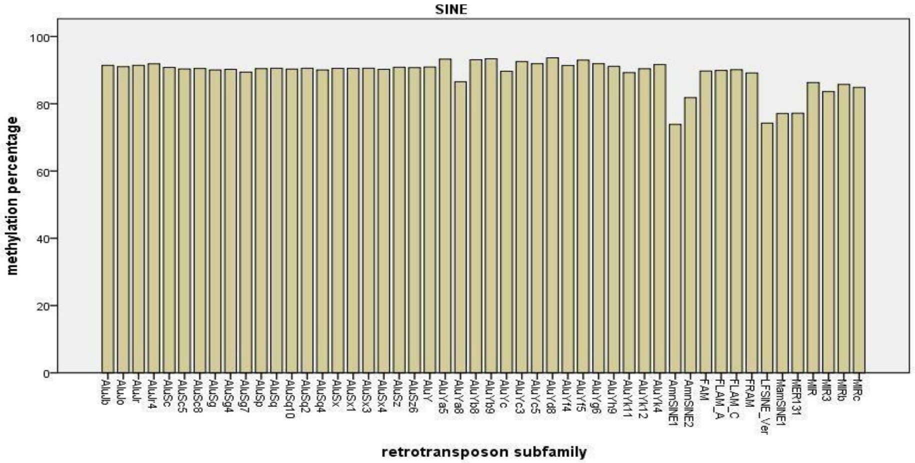
SRR478991 (48 year old male)

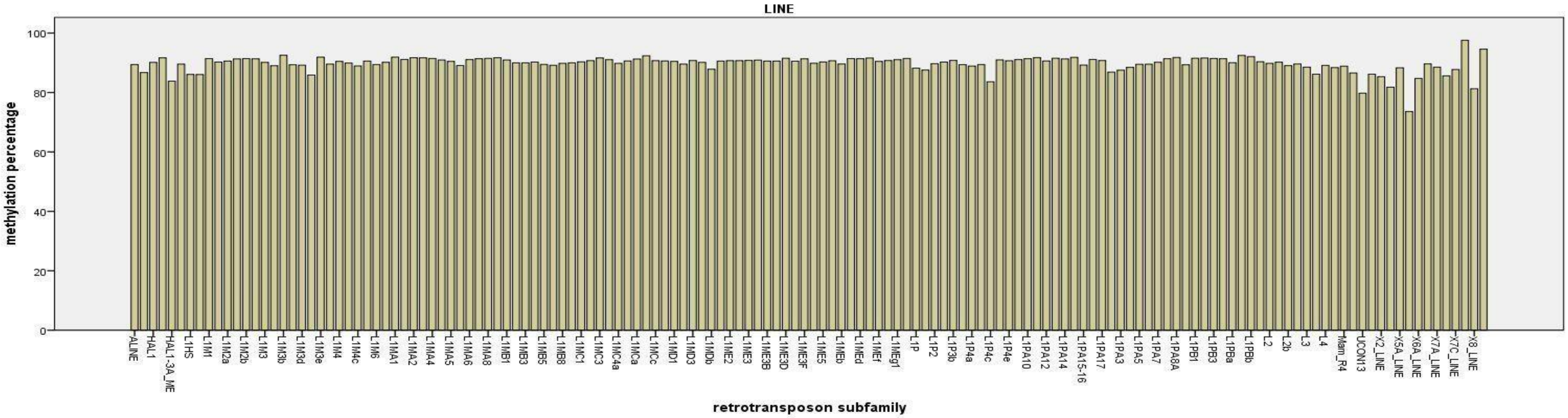


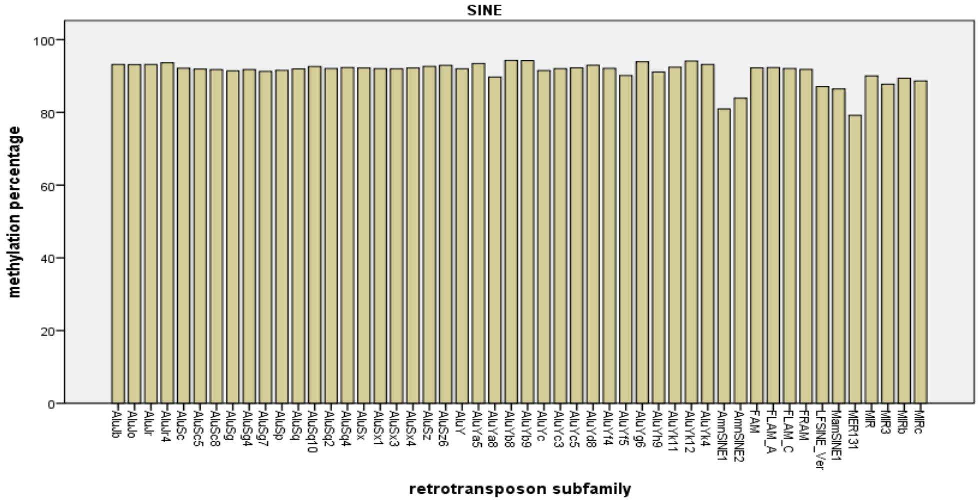


SRR921749 (55 year old male)

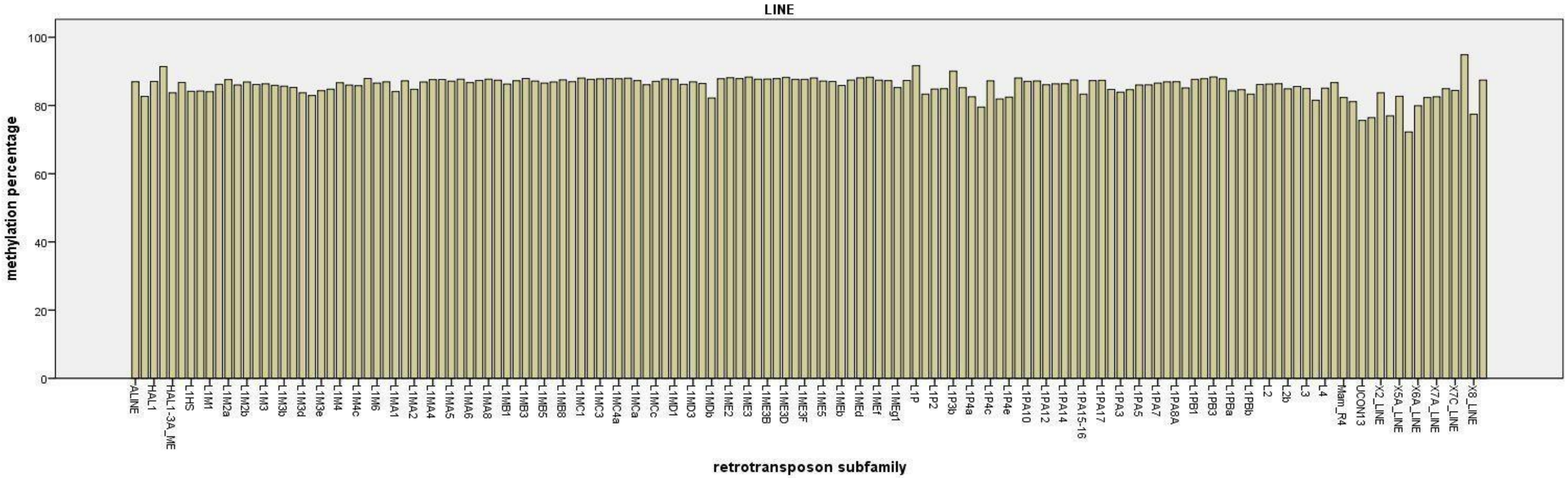


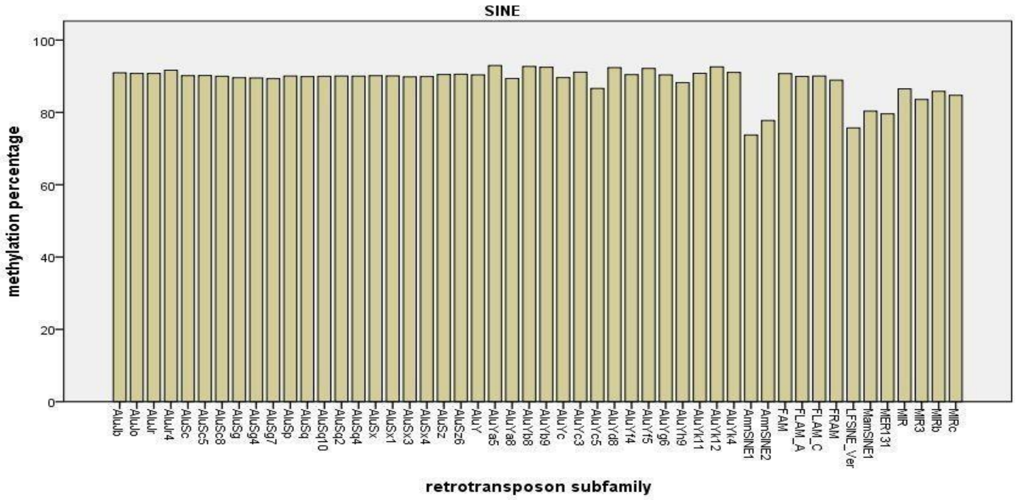



LINE
methylation percentage
0
20
40
60
80
100
ALINE
HAL1
HAL1-3A_ME
L1HS
L1M1
L1M2a
L1M2b
L1M3
L1M3b
L1M3d
L1M3e
L1M4
L1M4c
L1M6
L1MA1
L1MA2
L1MA4
L1MA5
L1MA6
L1MA8
L1MB1
L1MB3
L1MB5
L1MB8
L1MC1
L1MC3
L1MC4a
L1MCa
L1MCc
L1MD1
L1MD3
L1MDb
L1ME2
L1ME3
L1ME3B
L1ME3D
L1ME3F
L1ME5
L1MEb
L1MEd
L1MEf
L1MEg1
L1P
L1P2
L1P3b
L1P4a
L1P4c
L1P4e
L1PA10
L1PA12
L1PA14
L1PA15-16
L1PA17
L1PA3
L1PA5
L1PA7
L1PA8A
L1PB1
L1PB3
L1PBa
L1PBb
L2
L2b
L3
L4
Mam_R4
UCON13
X2_LINE
X5A_LINE
X6A_LINE
X7A_LINE
X7C_LINE
X8_LINE
retrotransposon subfamily

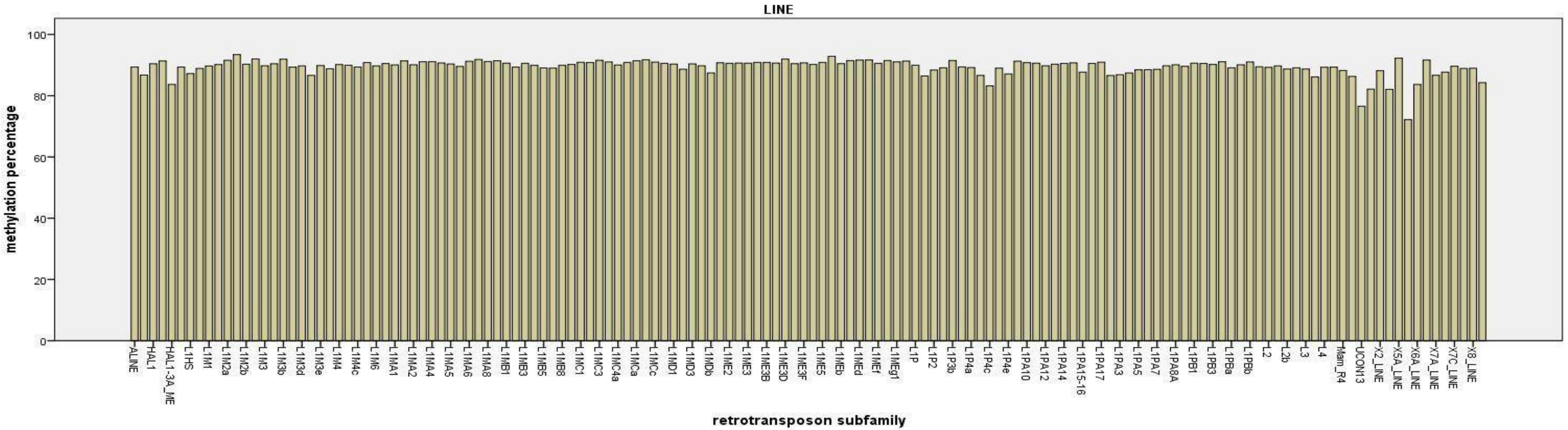

SINE
methylation percentage
0
20
40
60
80
100
AluJb
AluJo
AluJr
AluJr4
AluSc
AluSc5
AluSc8
AluSg
AluSg4
AluSg7
AluSp
AluSq
AluSq10
AluSq2
AluSq4
AluSx
AluSx1
AluSx3
AluSx4
AluSz
AluSz6
AluY
AluYa5
AluYa8
AluYb8
AluYb9
AluYc
AluYc3
AluYc5
AluYd8
AluYf4
AluYf5
AluYg6
AluYh9
AluYk11
AluYk12
AluYk4
AmnSINE1
AmnSINE2
FAM
FLAM_A
FLAM_C
FRAM
LFSINE_Ver
MamSINE1
MER131
MIR
MIR3
MIRb
MIRc
retrotransposon subfamily

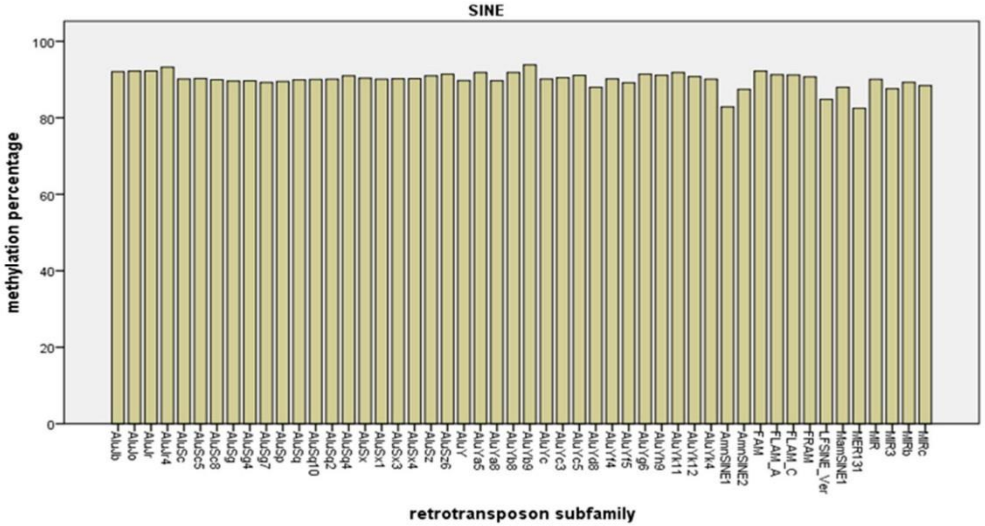
SRR847424 (42 year old female)



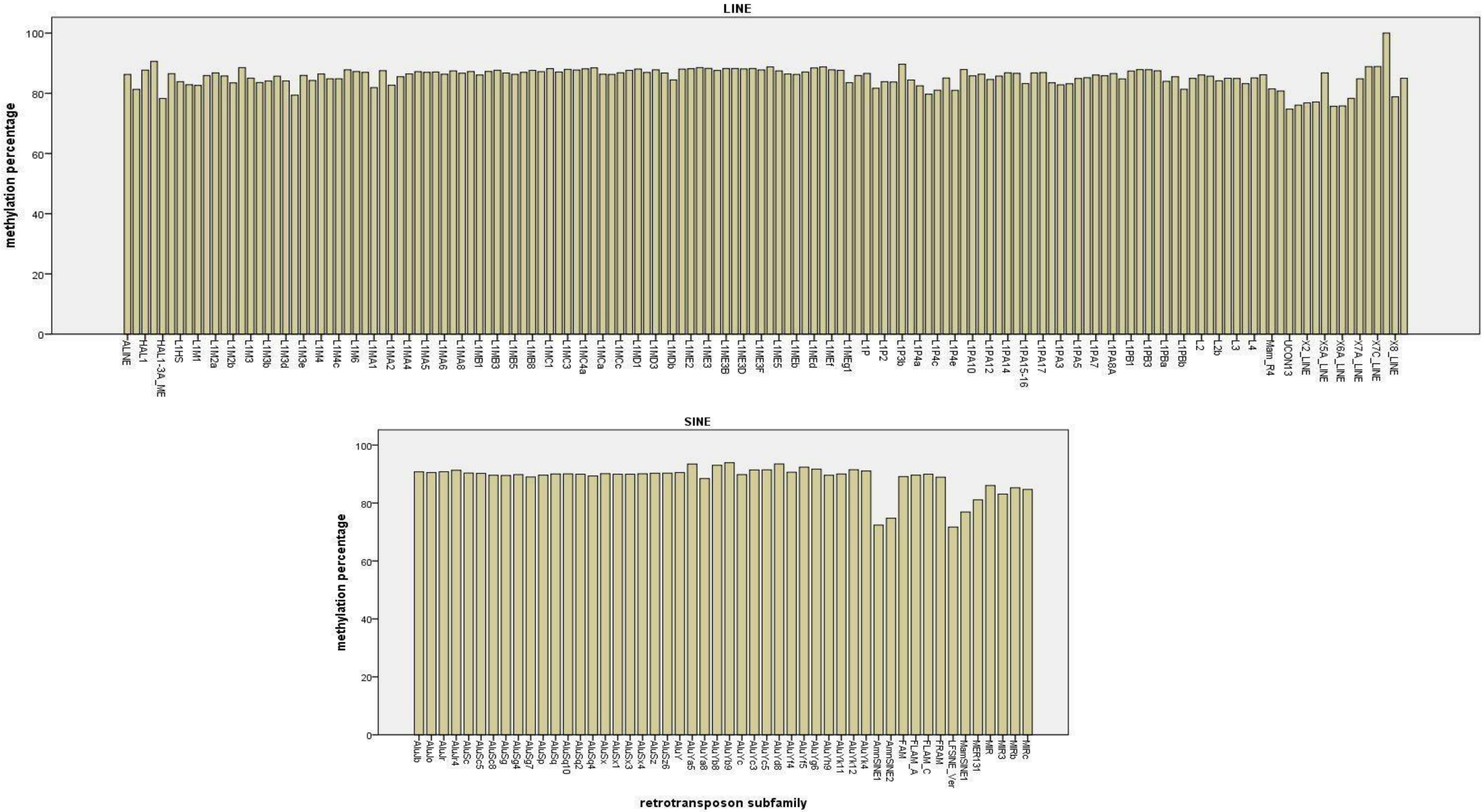


SRR921706 (53 year old female)

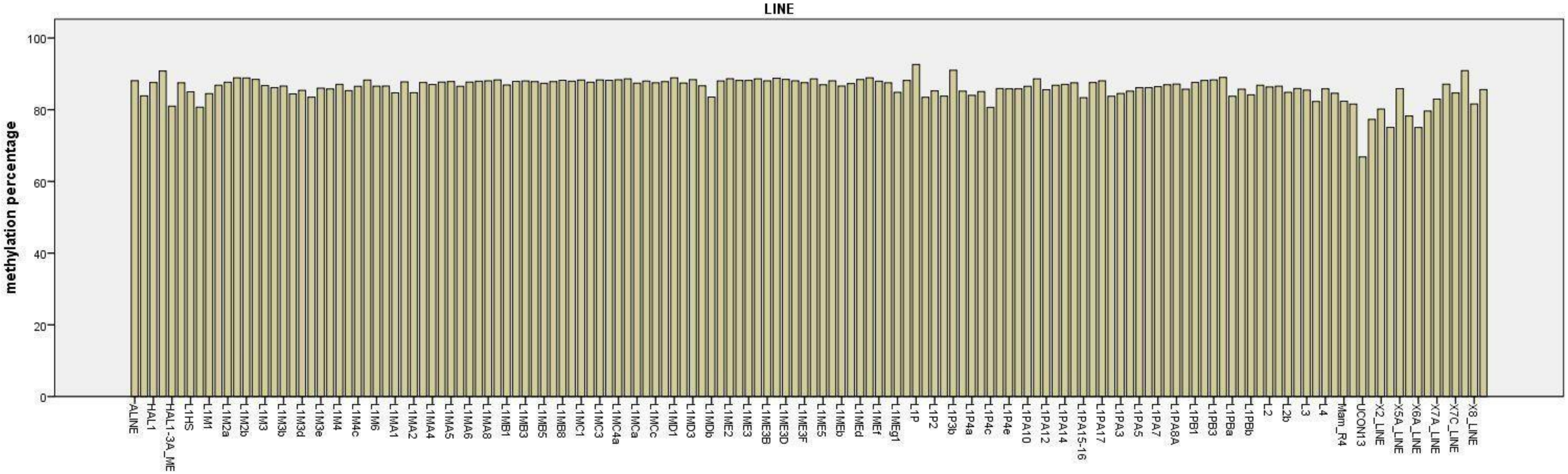


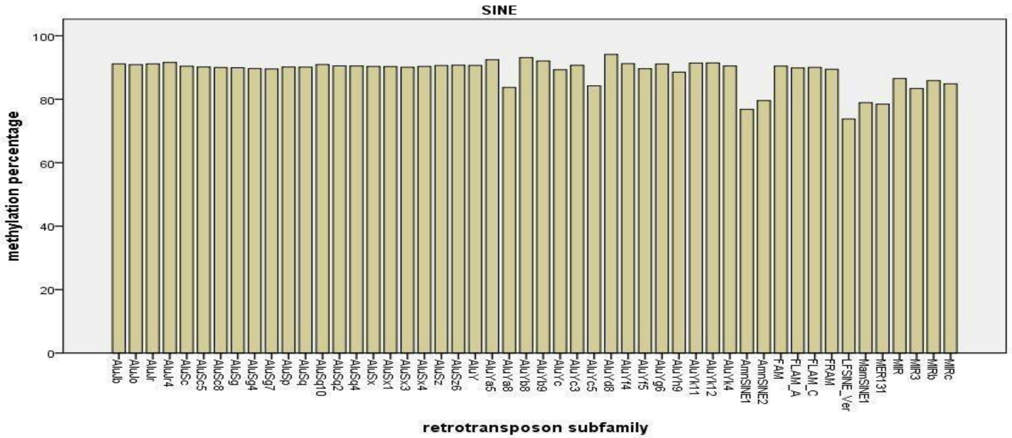


SRR921723 (53 year old female)

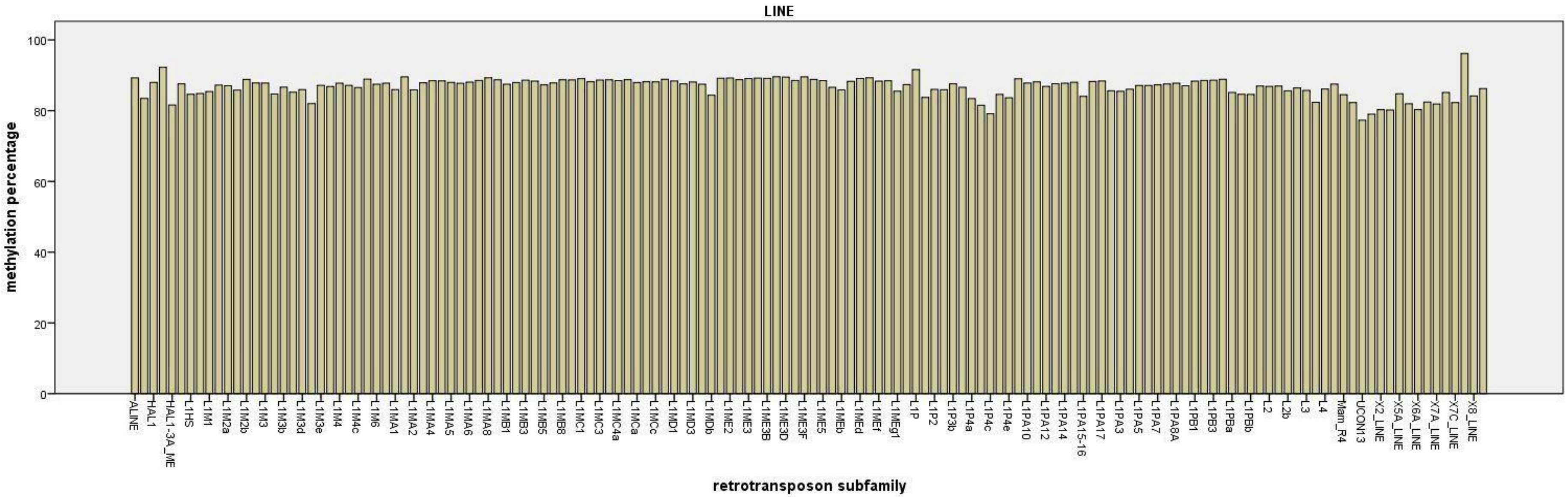


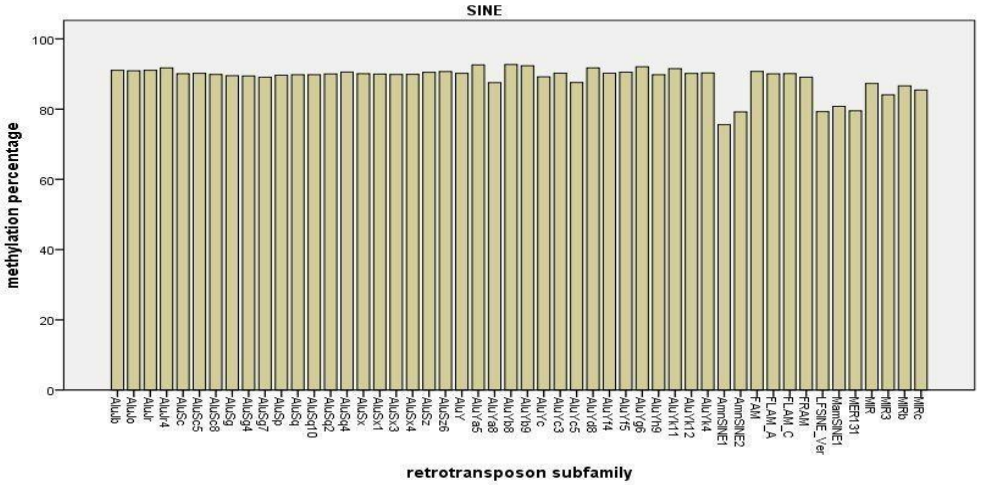
SRR921702 (64 year old female)



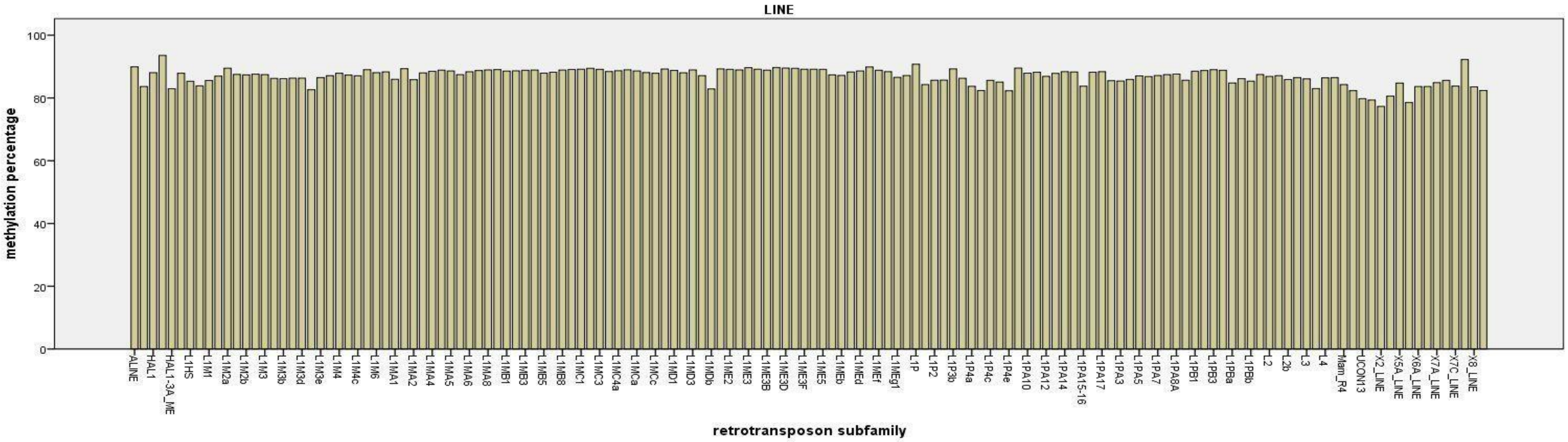


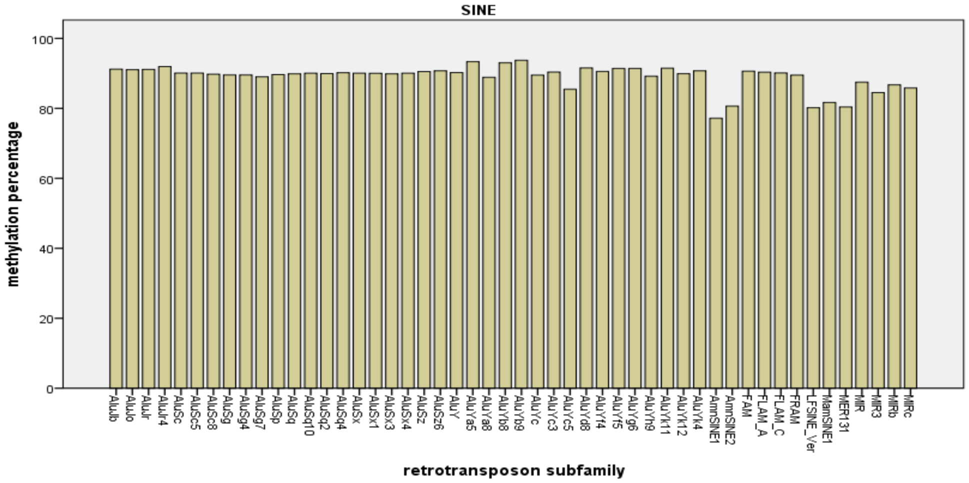
SRR949195 (81 year old female)





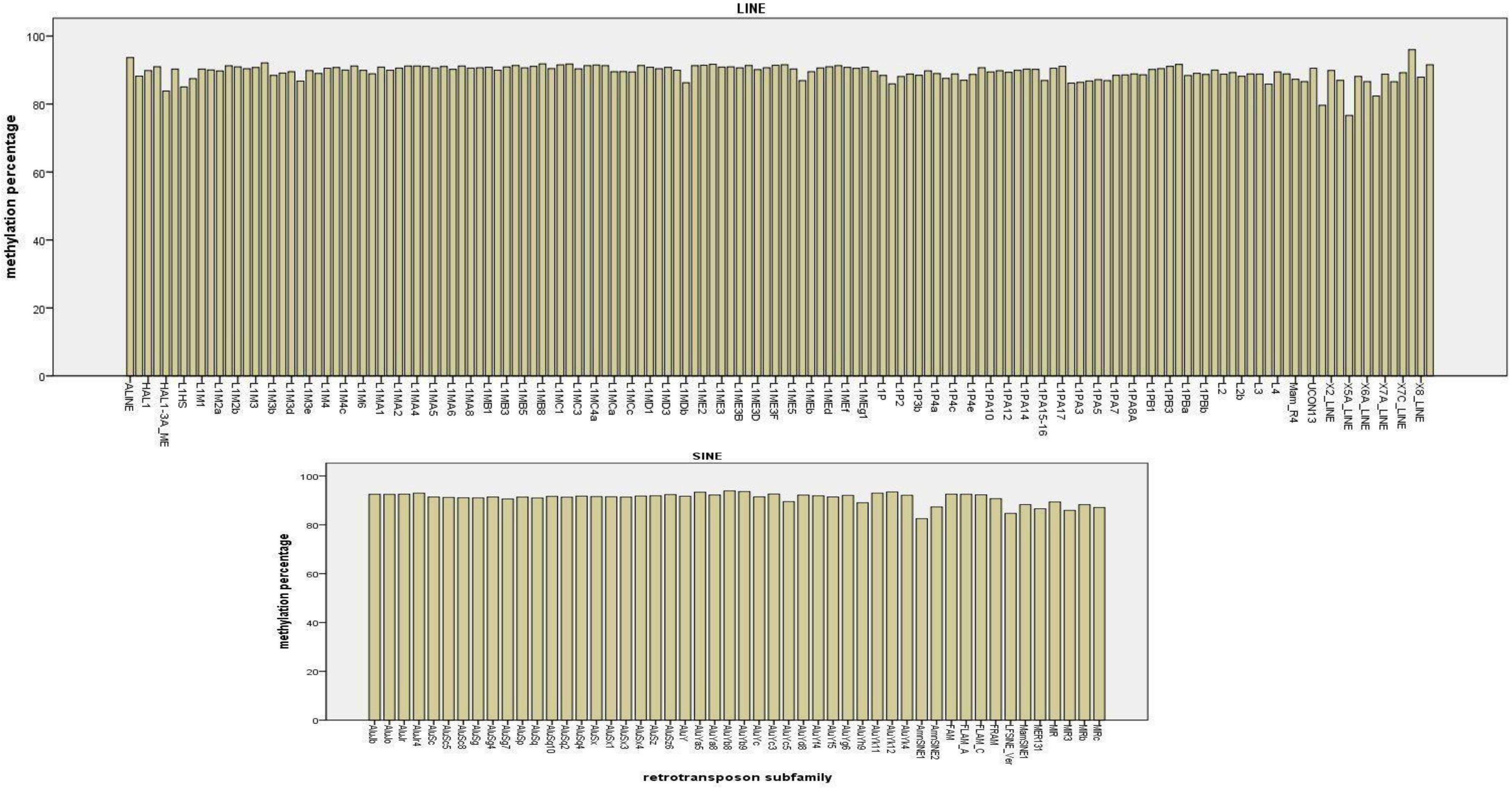
SRR949197 (82 year old female)



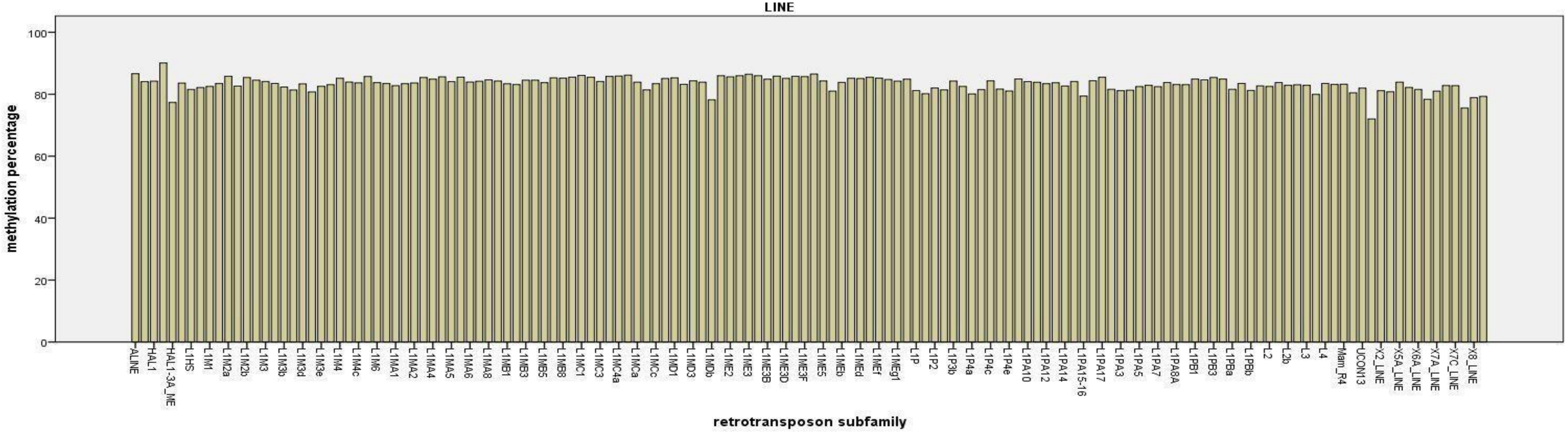


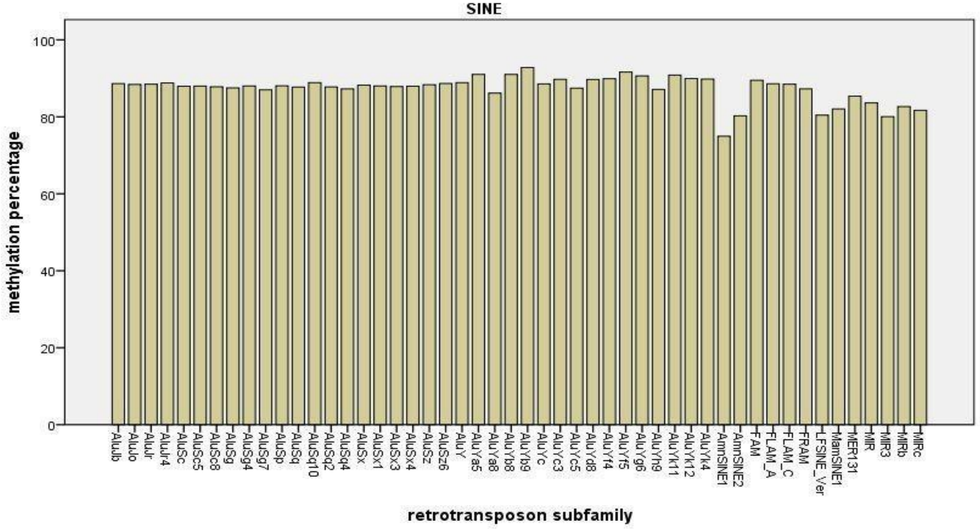
Plots generated from Regional Methylation Density Calculation Plots for Blood samples

SSR330578 (1 year old)

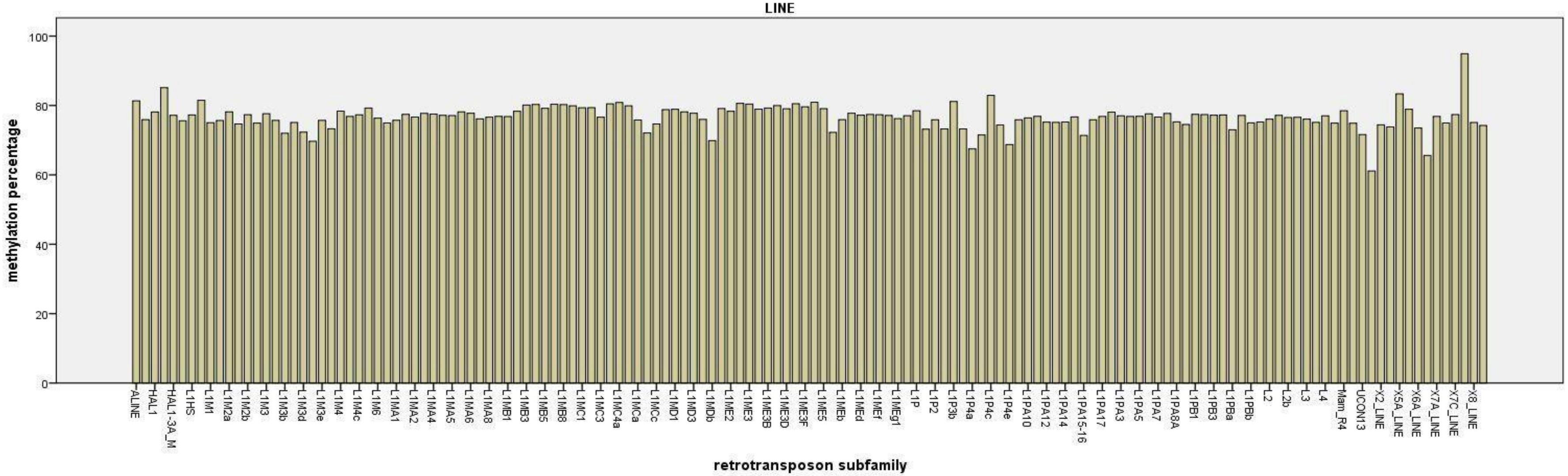


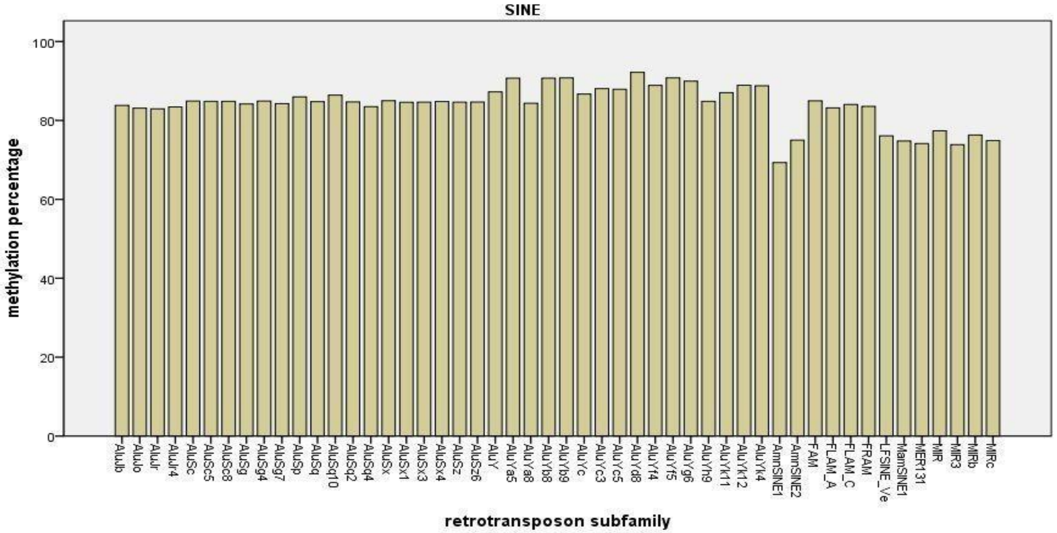
SSR389249 (26 year old male)





SSR330576 (103 year old male)

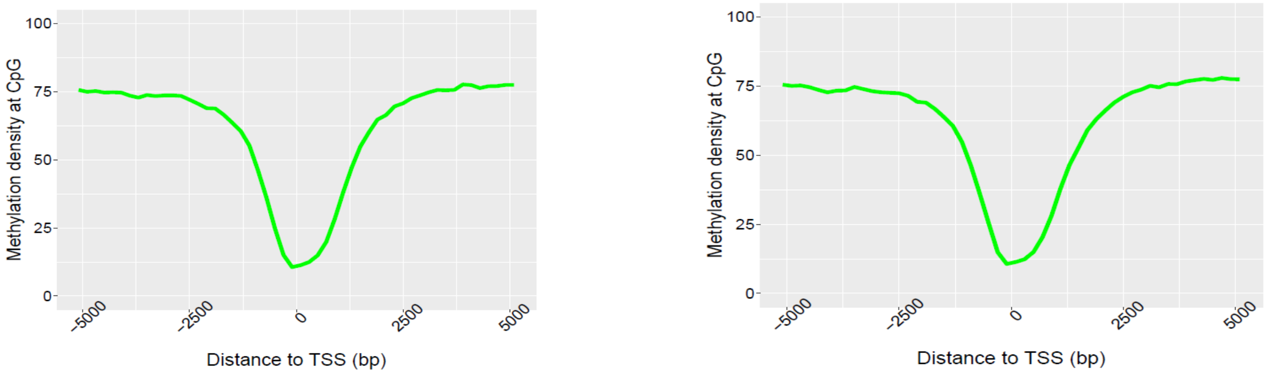




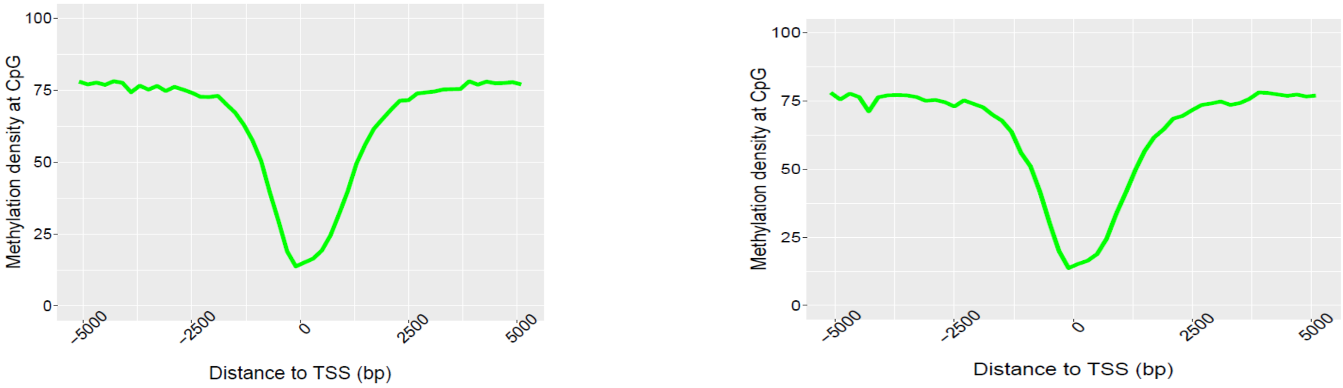
METHY PIPE BRAIN SAMPLES OUTPUTS

Methylation density from CpG sites around TSS regions with 200 bp bins in Watson (left) and Crick (right) strands

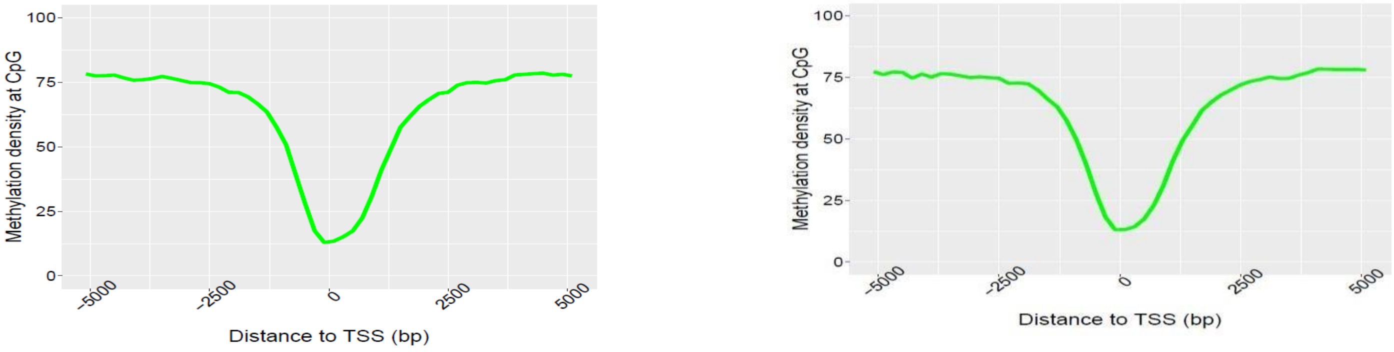
SRR901381 (20 weeks old male)



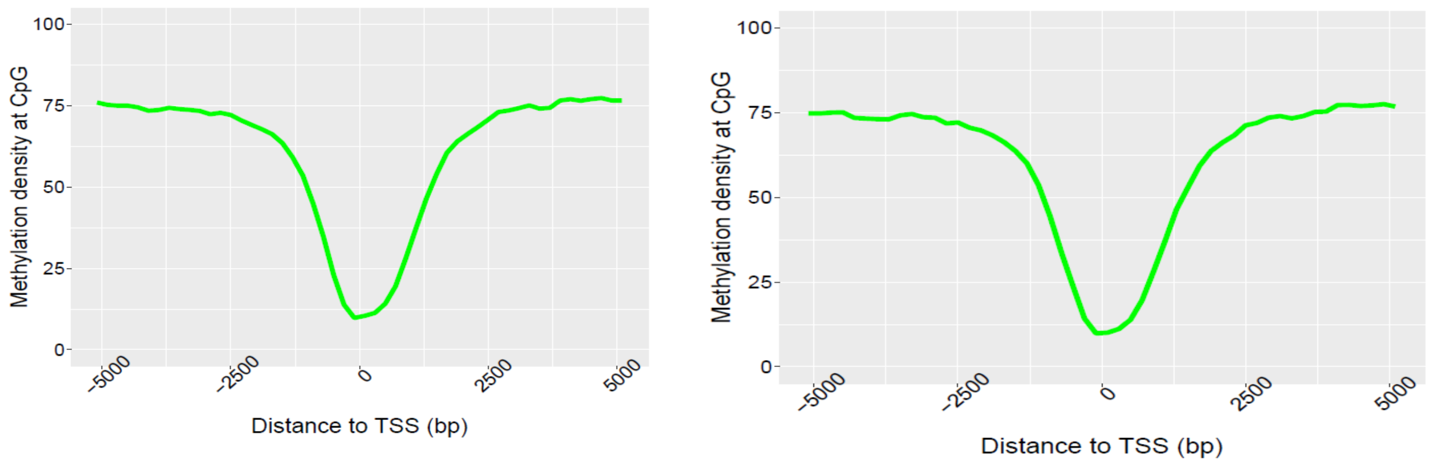
SRR 47900 (31 year old male)



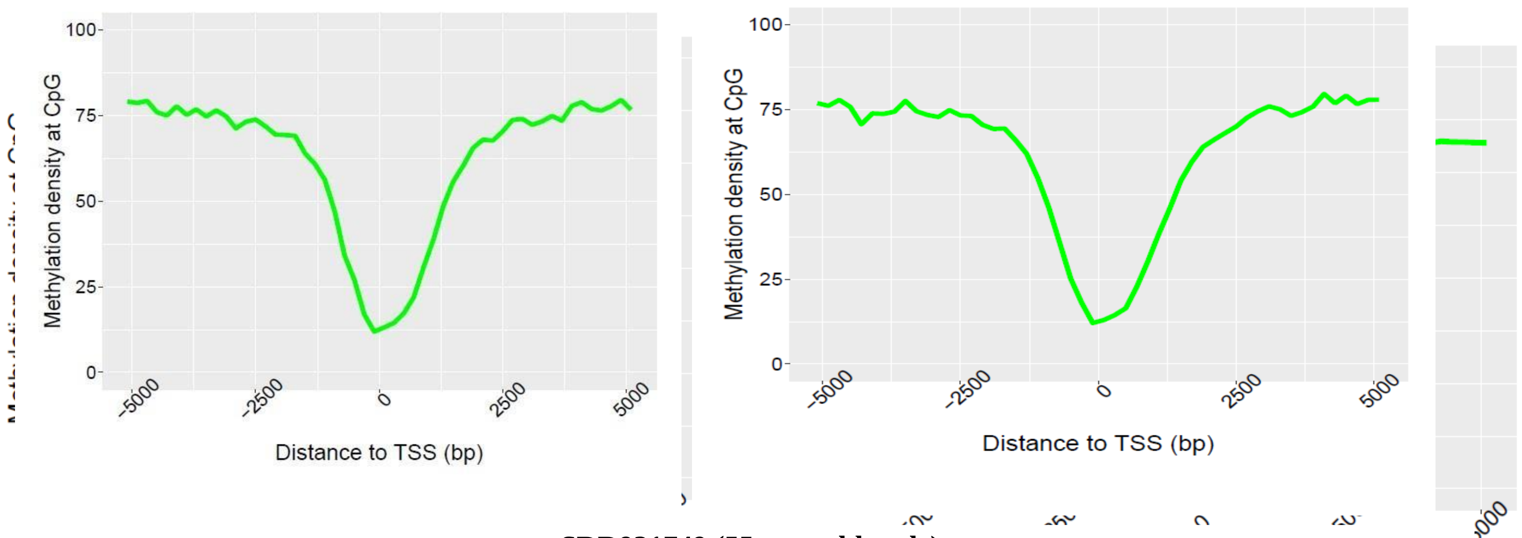
SRR 478994 (47 year old male)



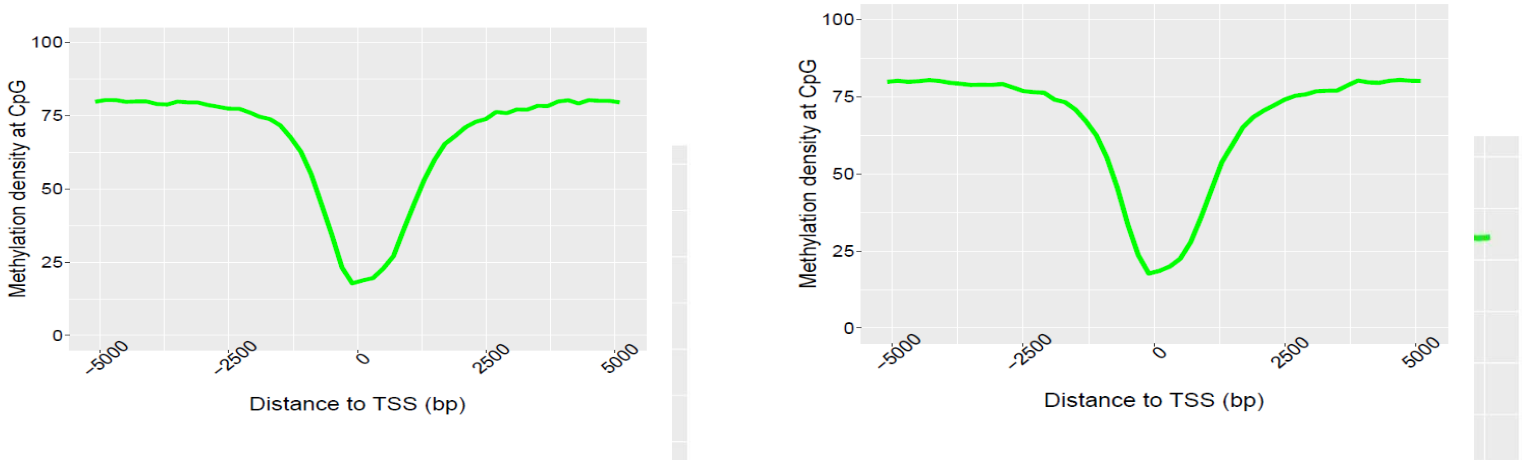
SRR478991 (48 year old male)



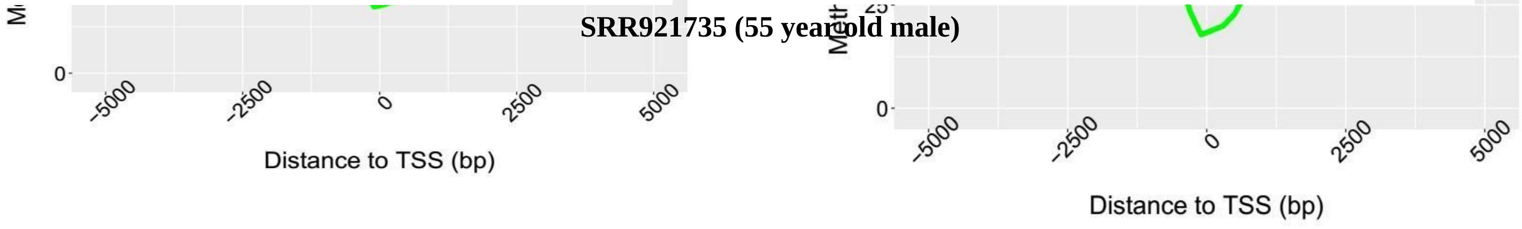
SRR921729 (55 year old male)



SRR921749 (55 year old male)

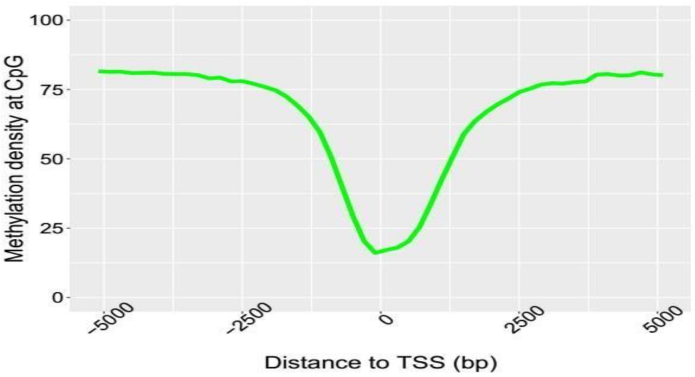


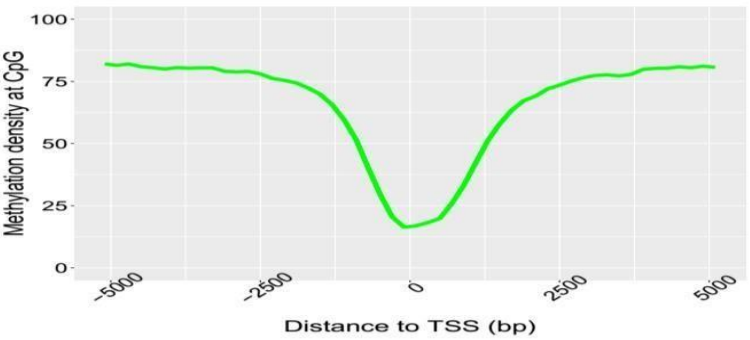
SRR921735 (55 year old male)



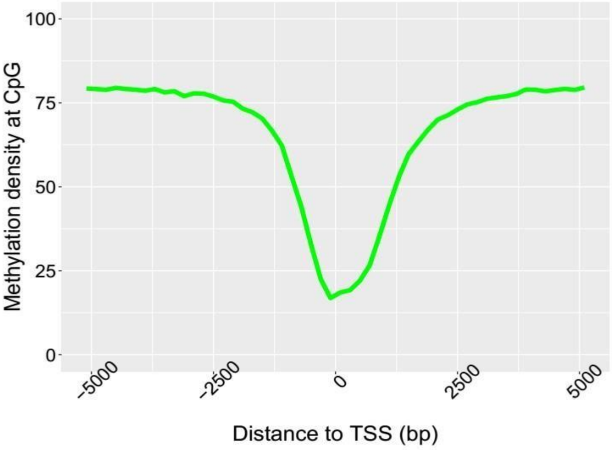
SRR847424 (42 year old female)

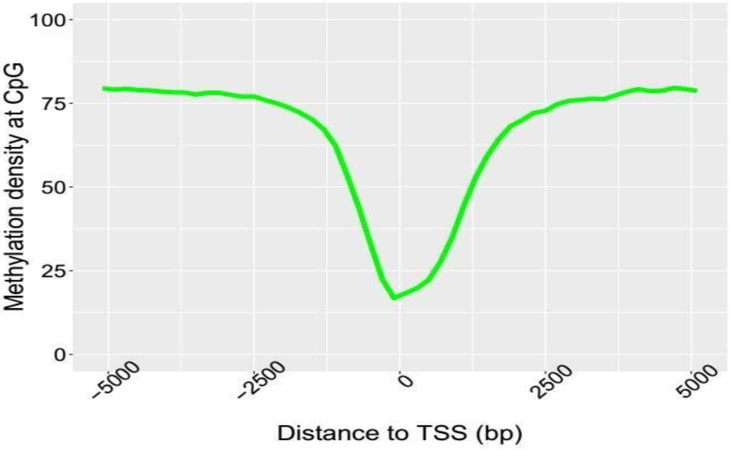
SRR921706 (53 yeal old female)



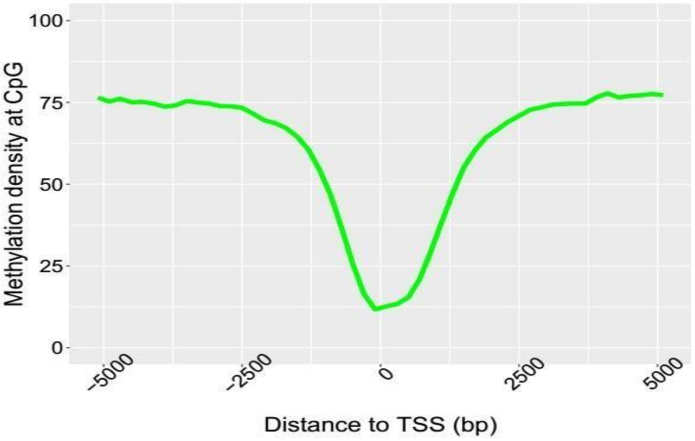


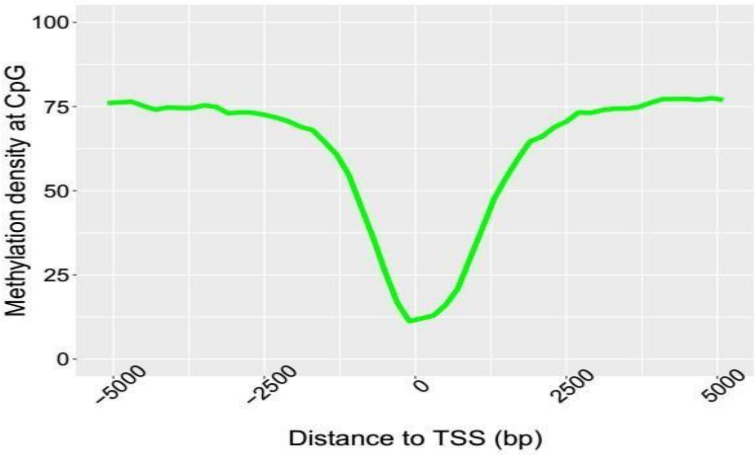
SRR921723 (53 year old female)



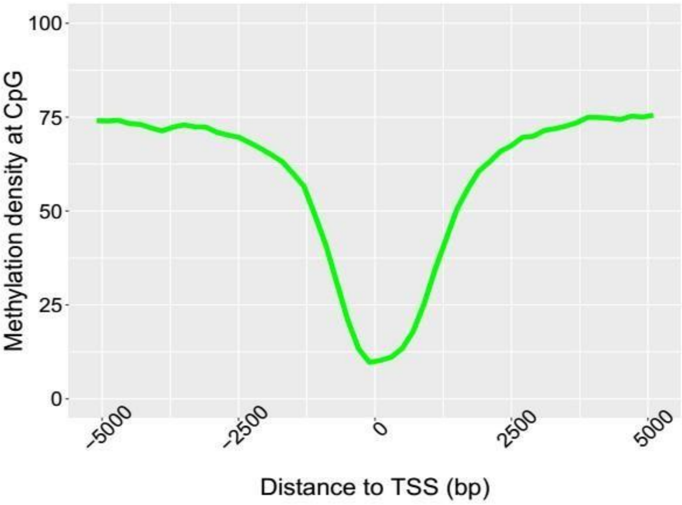


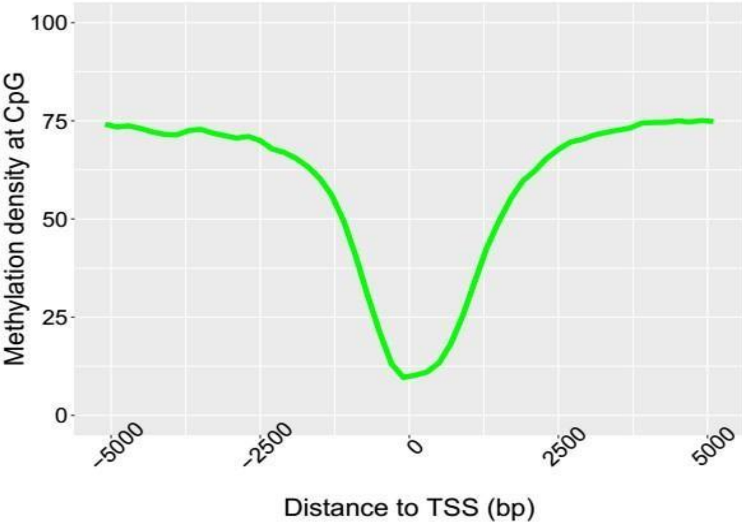
SRR921702 (64 year old female)



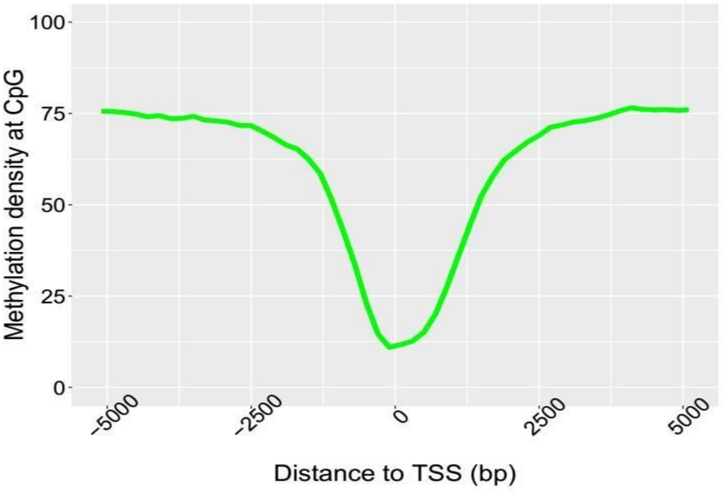


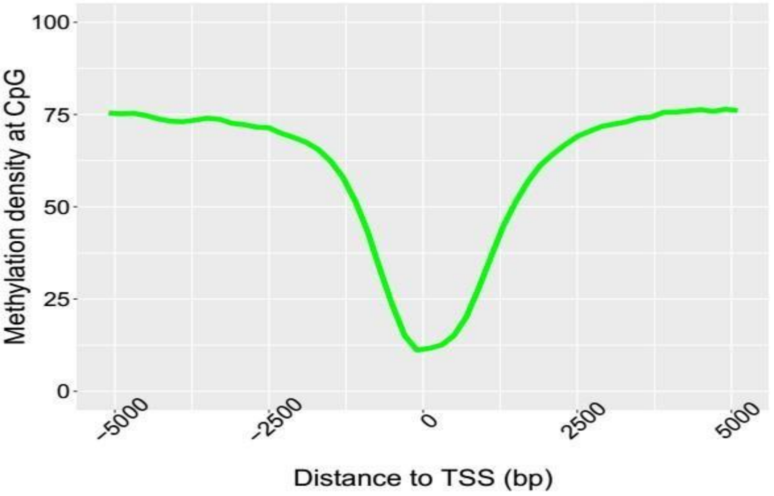
SRR949195 (81 year old female)





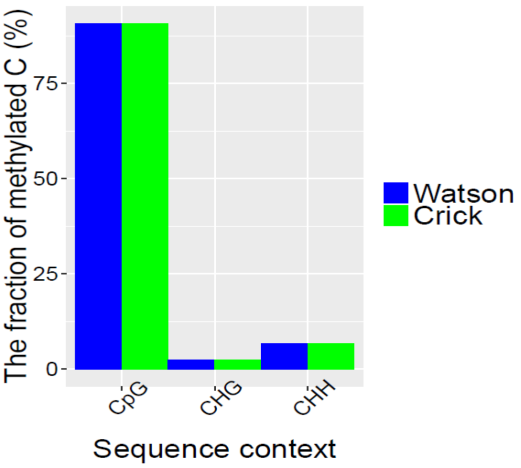
SRR949197 (82 year old female)



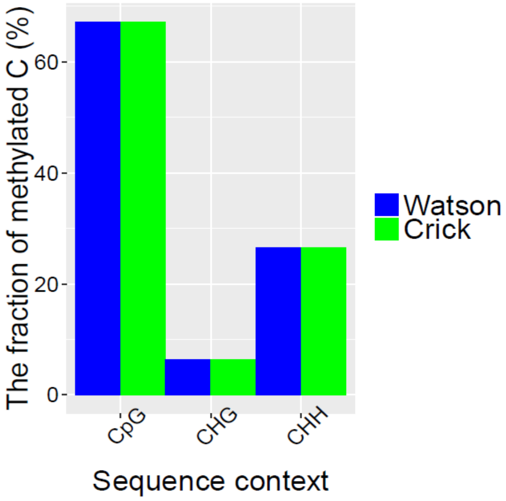


The fraction of methylated C according to the sequence context

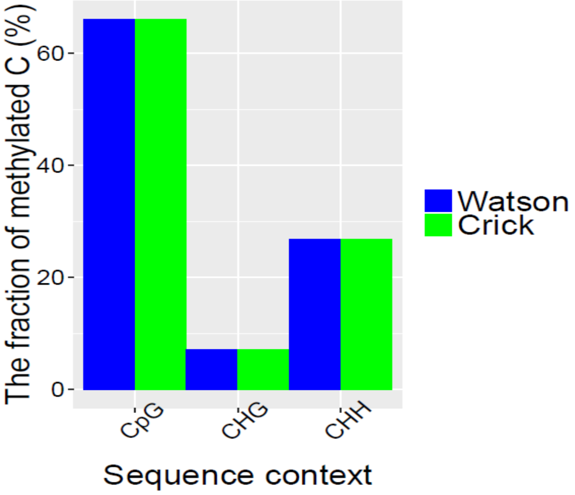
SRR901381 (20 weeks old male)



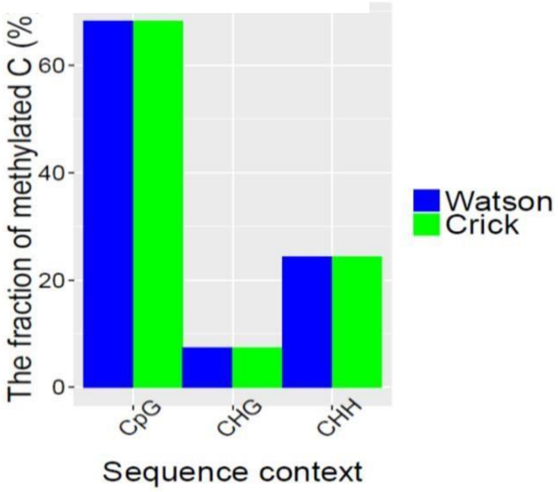
SRR 47900 (31 year old male)



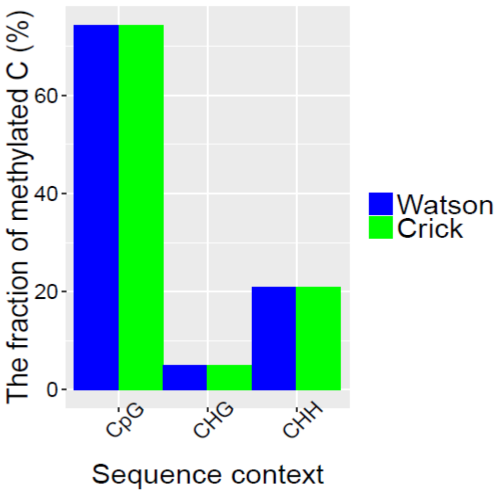
SRR 478994 (47 year old male)



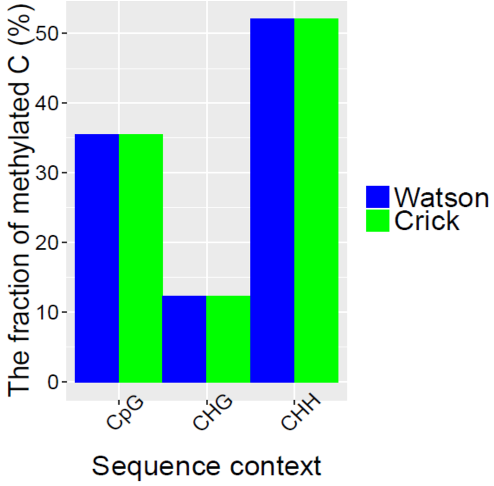
SRR478991 (48 year old male)



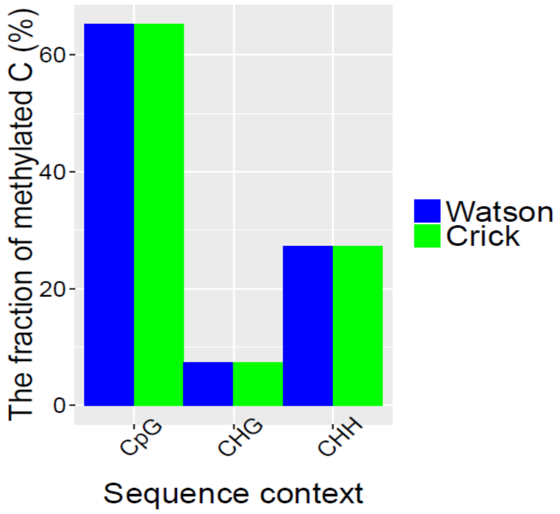
SRR921749 (55 year old male)



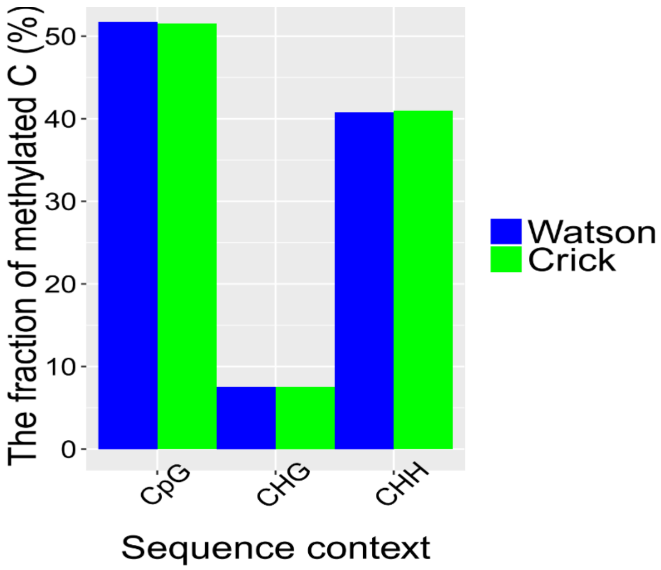
SRR921735 (55 year old male)



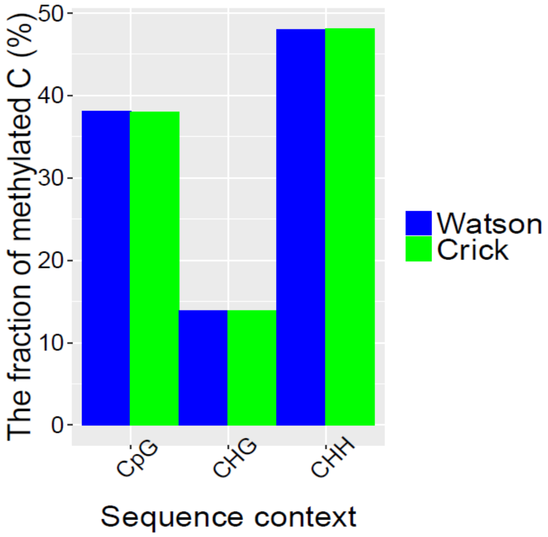
SRR921729 (55 year old male)



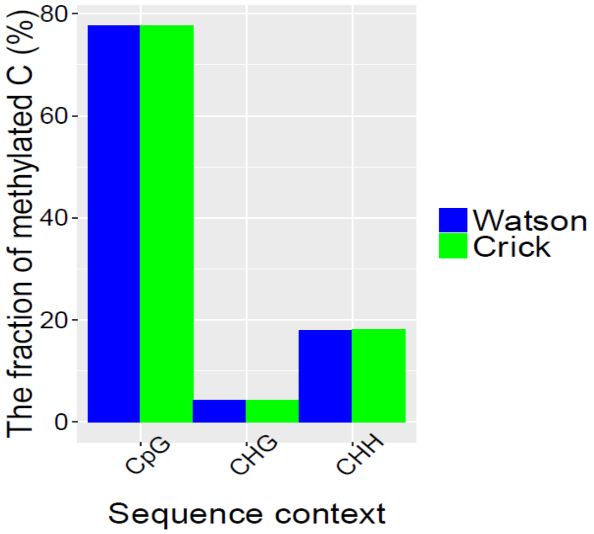
SRR847424 (42 year old female)



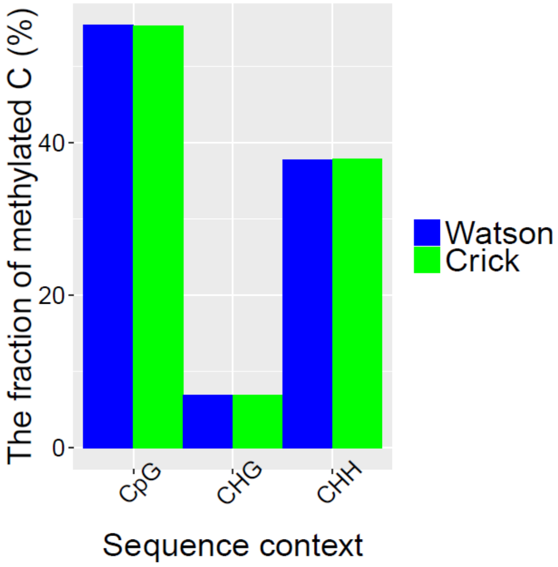
SRR921706 (53 yeal old female)



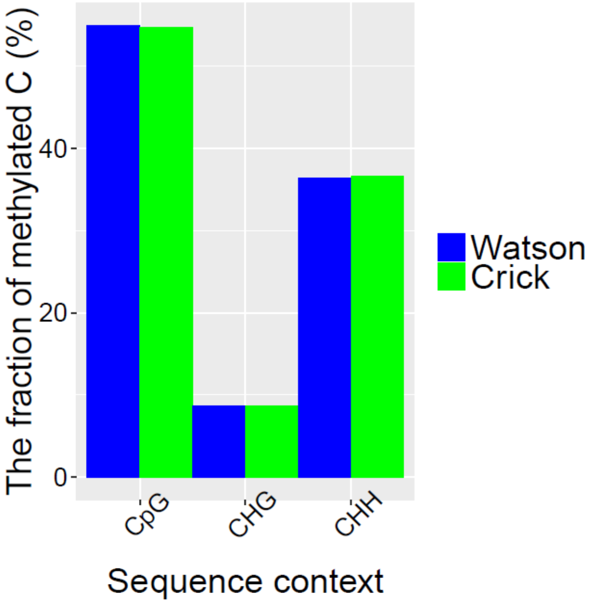
SRR921723 (53 year old female)



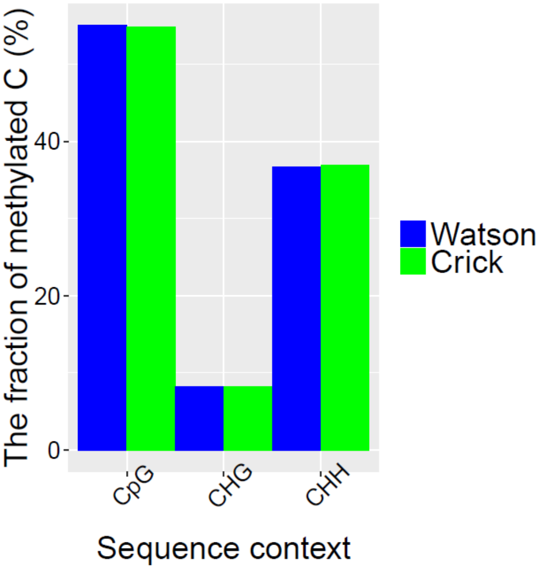
SRR921702 (64 year old female)



SRR949195 (81 year old female)

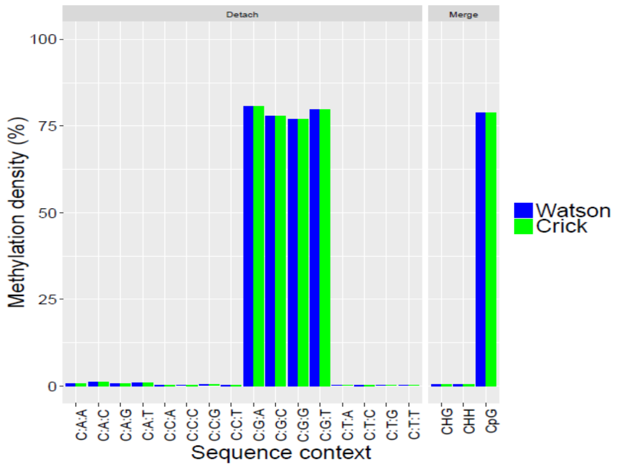


SRR949197 (82 year old female)

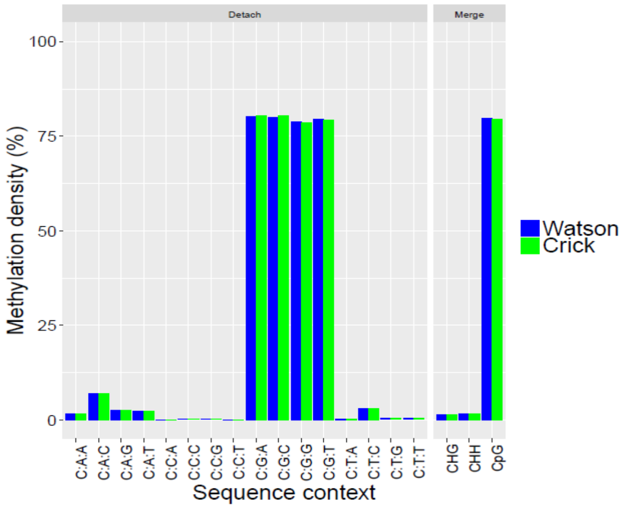


Methylation Density according to the Sequence Context

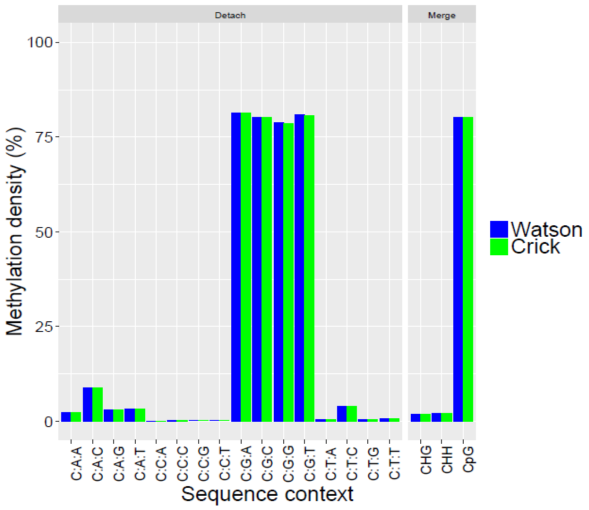
SRR901381 (20 weeks old male)



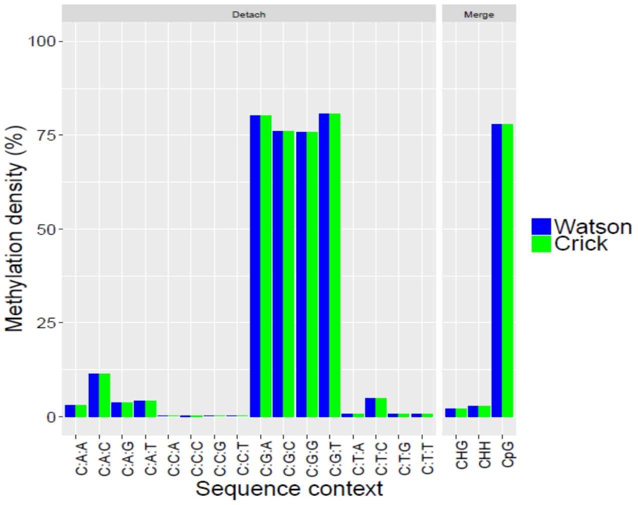
SRR 47900 (31 year old male)



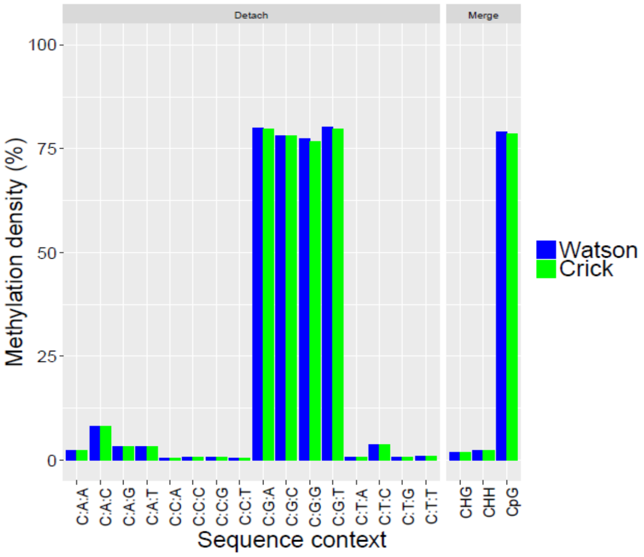
SRR 478994 (47 year old male)



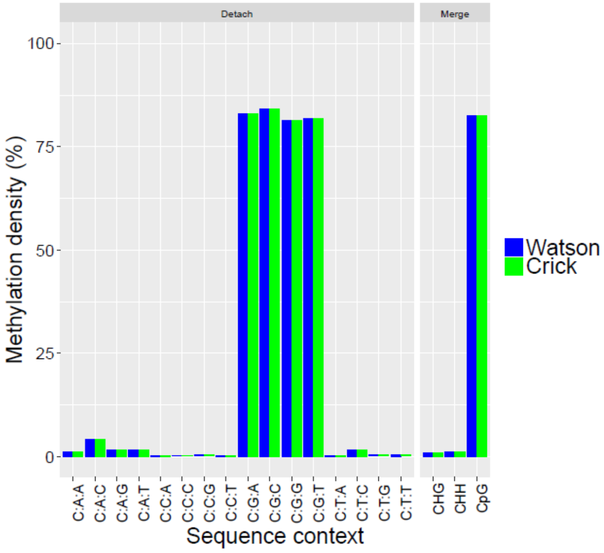
SRR478991 (48 year old male)



SRR921729 (55 year old male)

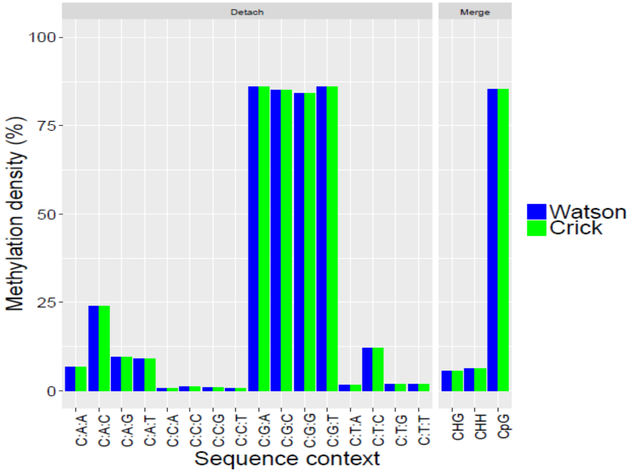


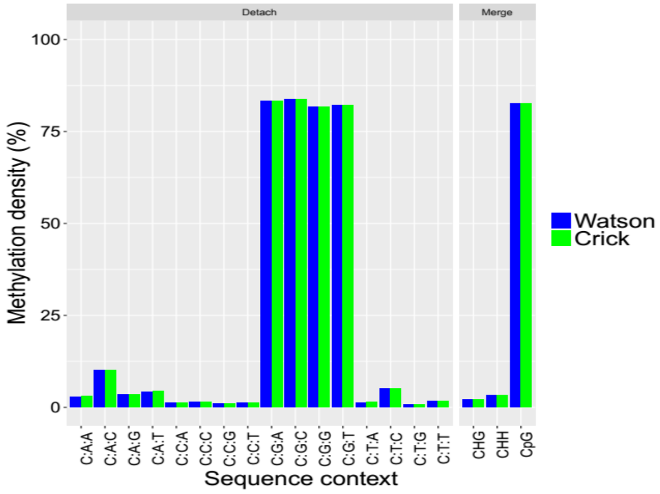
SRR921749 (55 year old male)



SRR921735 (55 year old male)

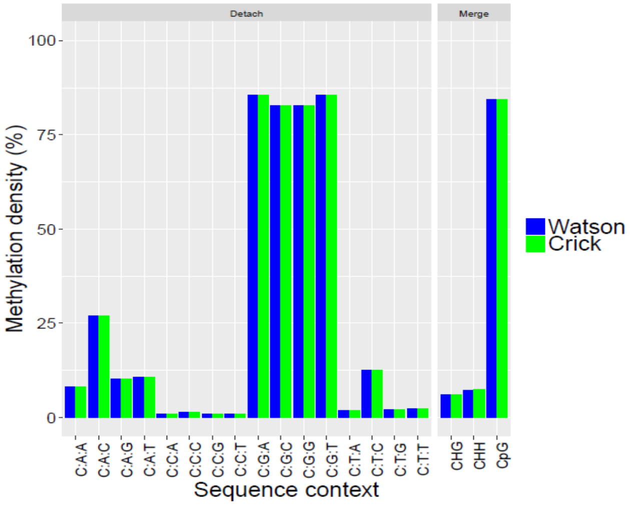
SRR847424 (42 year old female)

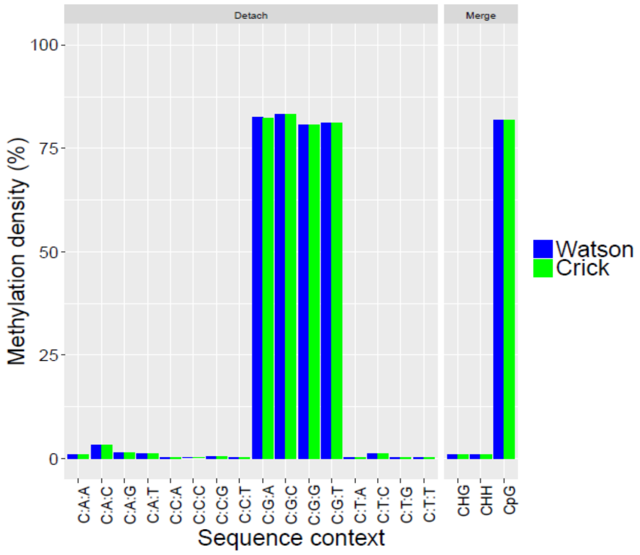




SRR921706 (53 year old female)

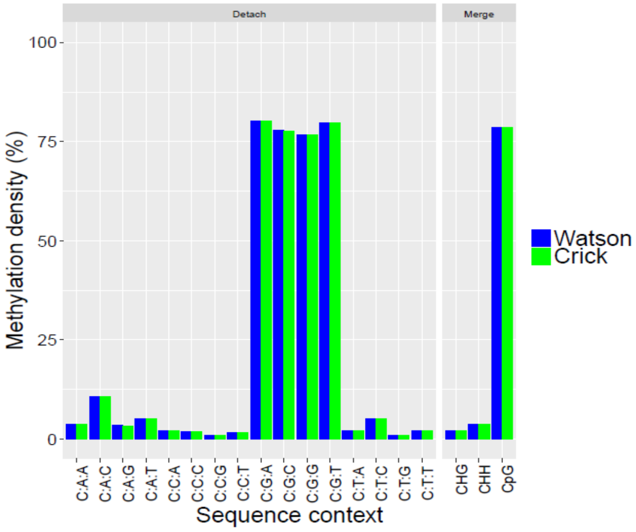
SRR921723 (53 year old female)

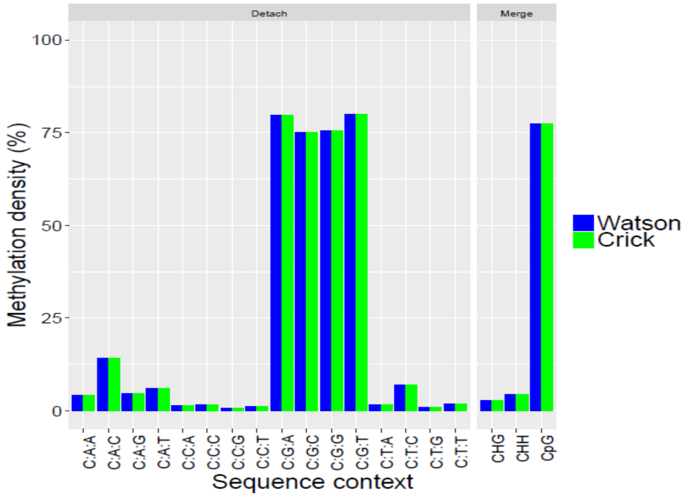




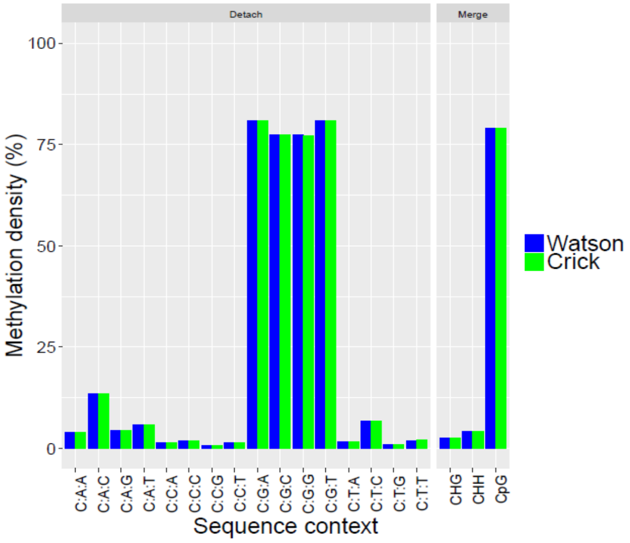
SRR921702 (64 year old female)

SRR949195 (81 year old female)


Detach
Merge
100
75
50
25
0
Methylation density (%)
C:A:A
C:A:C
C:A:G
C:A:T
C:C:A
C:C:C
C:C:G
C:C:T
C:G:A
C:G:C
C:G:G
C:G:T
C:T:A
C:T:C
C:T:G
C:T:T
CHG
CHH
CpG
Sequence context
🟦 Watson
🟩 Crick

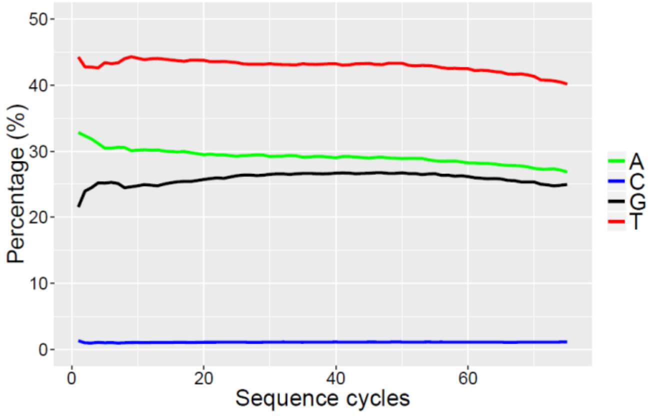

Detach
Merge
100
75
50
25
0
Methylation density (%)
C:A:A
C:A:C
C:A:G
C:A:T
C:C:A
C:C:C
C:C:G
C:C:T
C:G:A
C:G:C
C:G:G
C:G:T
C:T:A
C:T:C
C:T:G
C:T:T
CHG
CHH
CpG
Sequence context
🟦 Watson
🟩 Crick

SRR949197 (82 year old female)

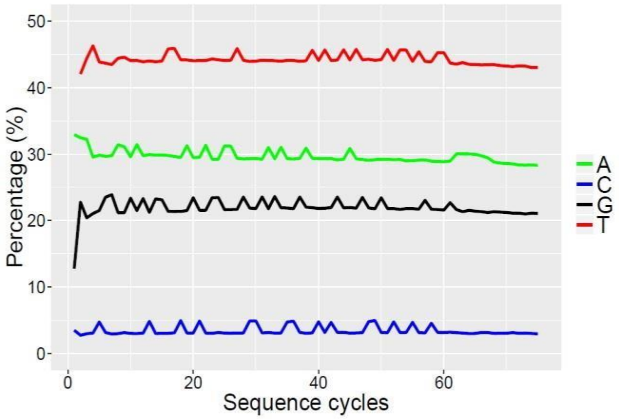


Base content percentage across sequence cycles

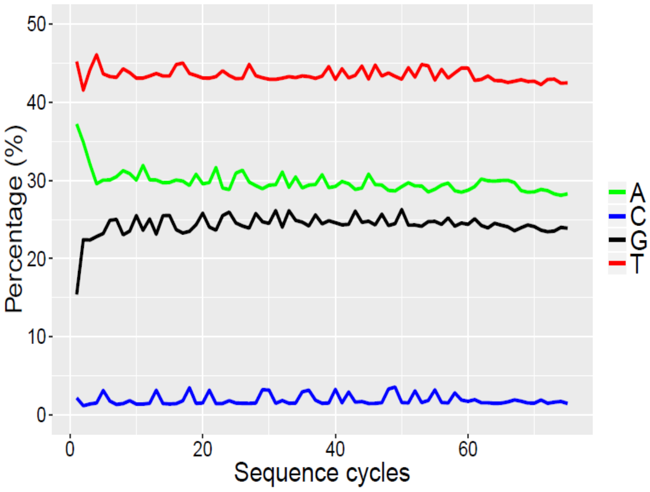
SRR901381 (20 weeks old male)



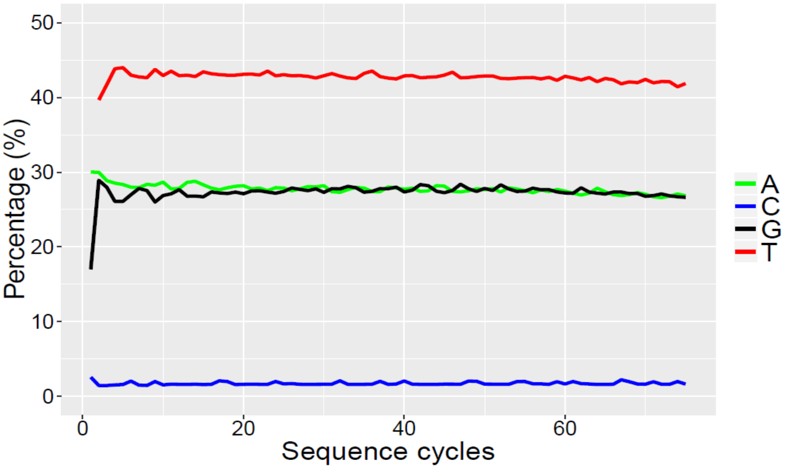
SRR47900 (31 year old male)



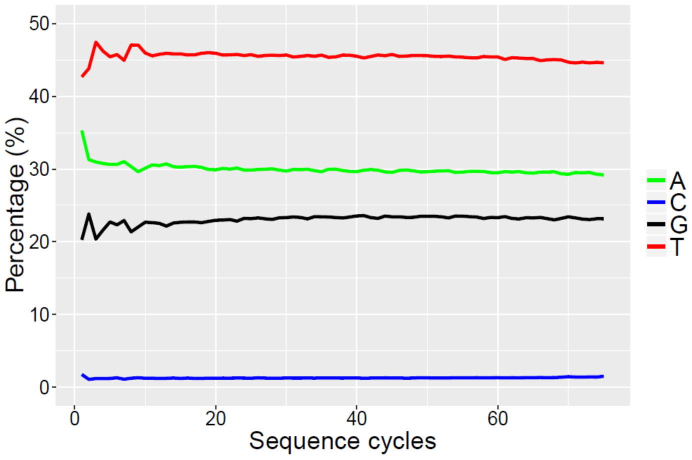
SRR 478994 (47 year old male)



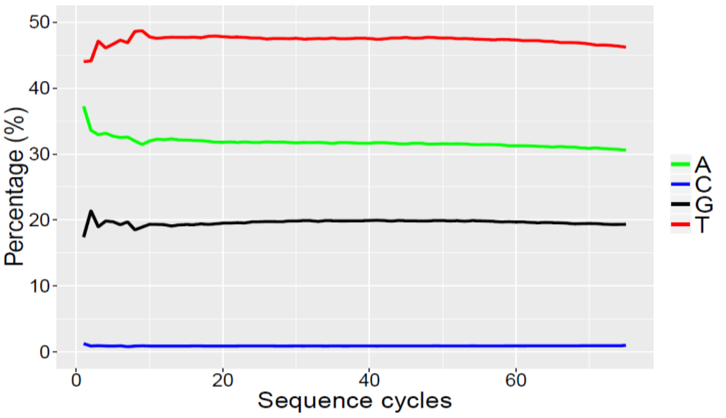
SRR478991 (48 year old male)



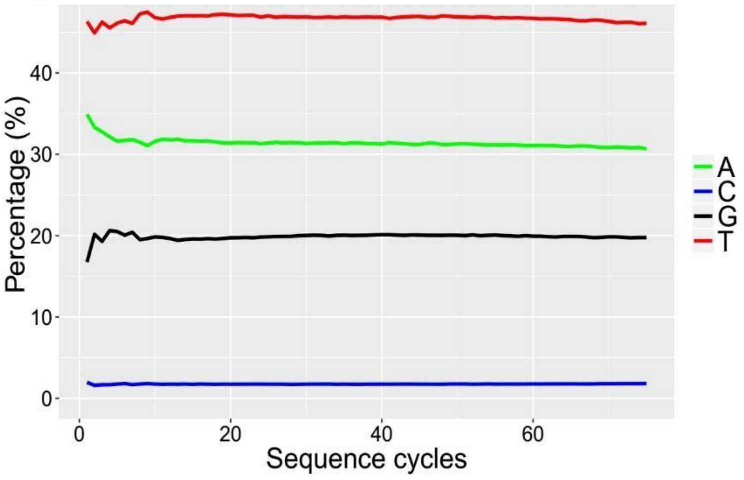
SRR921729 (55 year old male)



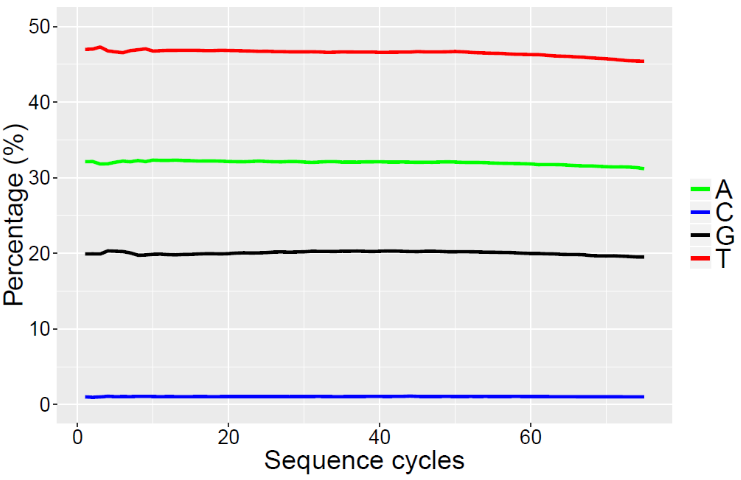
SRR921749 (55 year old male)



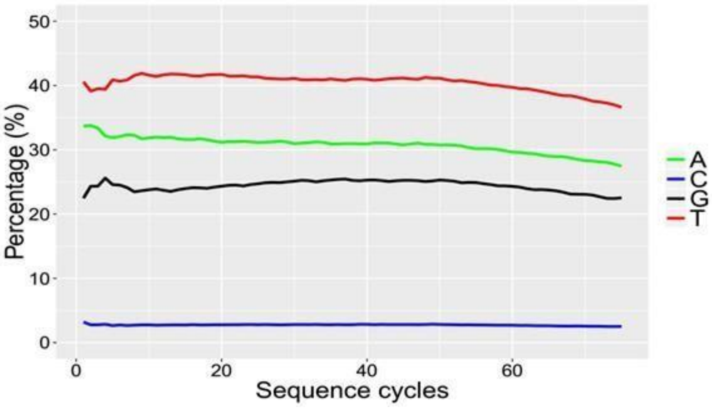
SRR921735 (55 year old male)



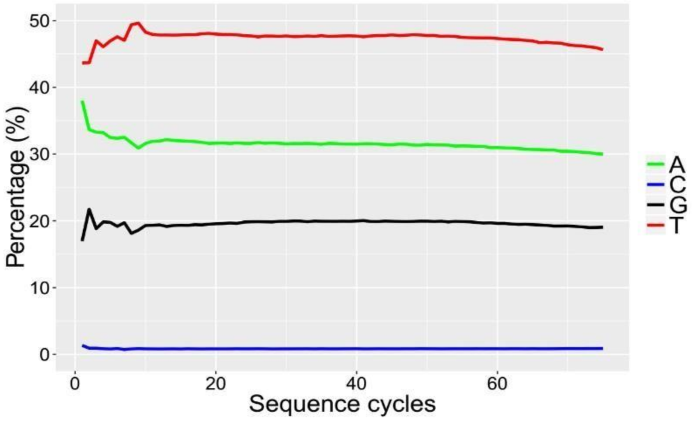
SRR847424 (42 year old female)



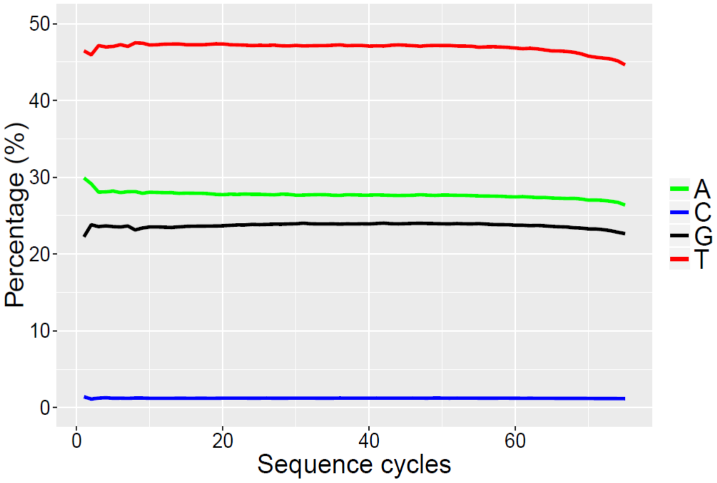
SRR921706 (53 year old female)



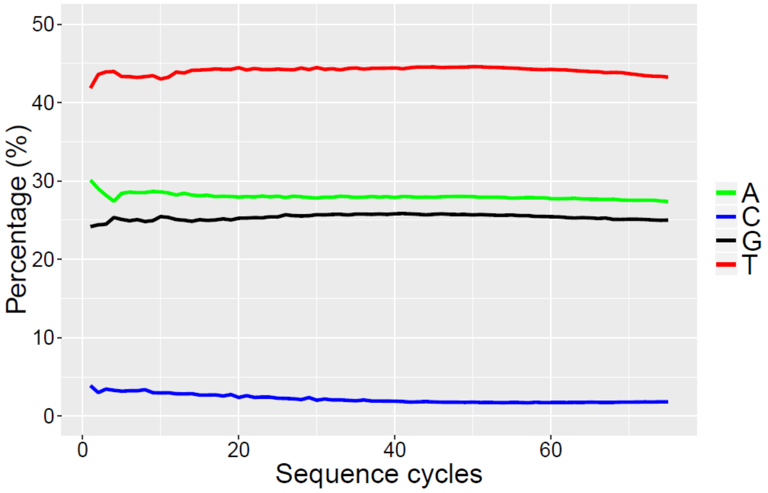
SRR921723 (53 year old female)



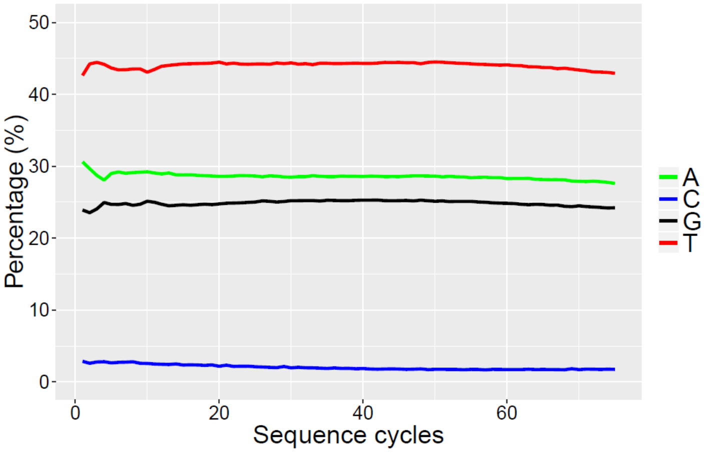
SRR921702 (64 year old female)



SRR949195 (81 year old female)

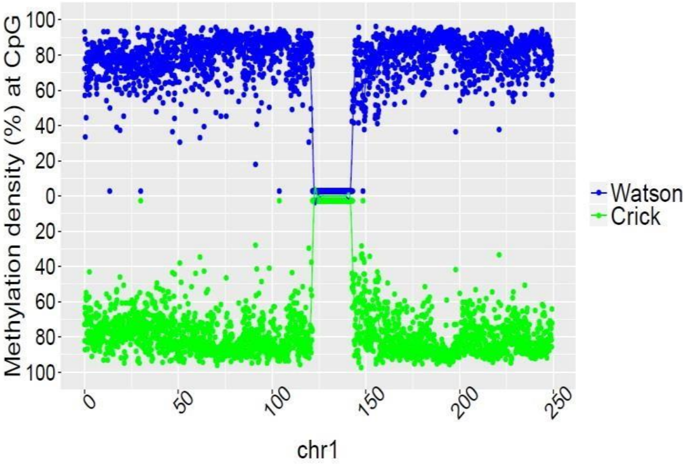


SRR949197 (82 year old female)

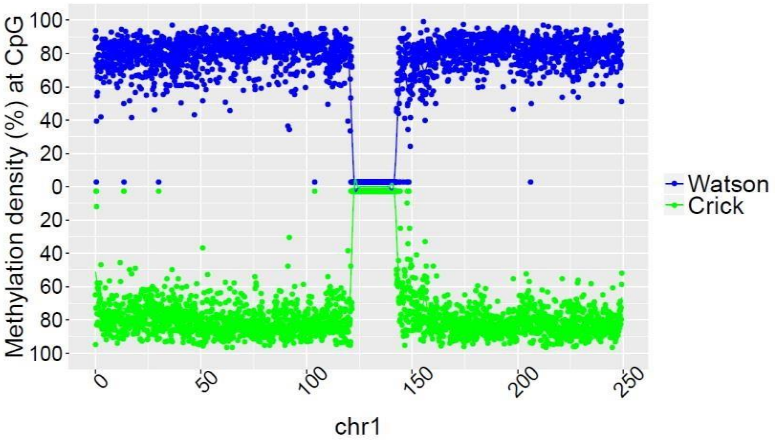


methylation density across cromossomes (distance in mega pb)

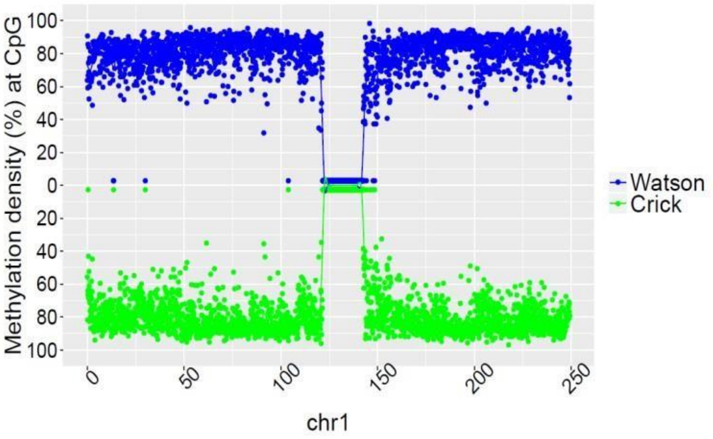
SRR901381 (20 weeks old male)



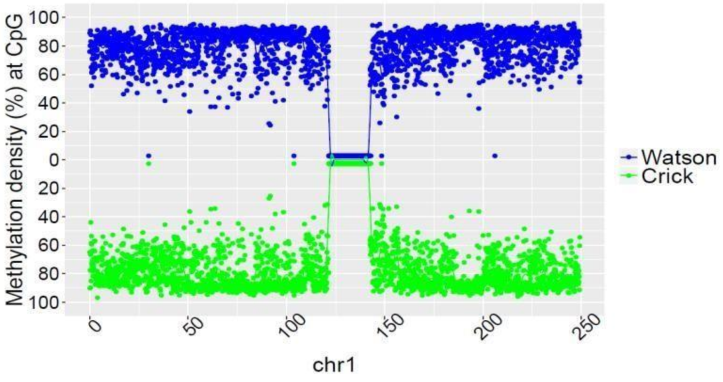
SRR47900 (31 year old male)



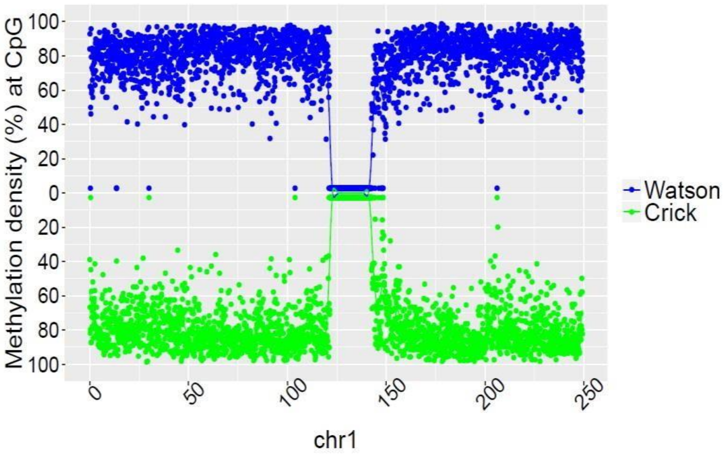
SRR 478994 (47 year old male)



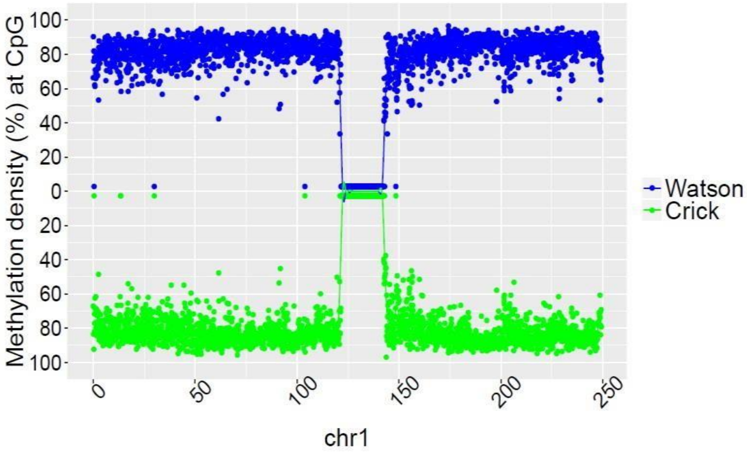
SRR478991 (48 year old male)



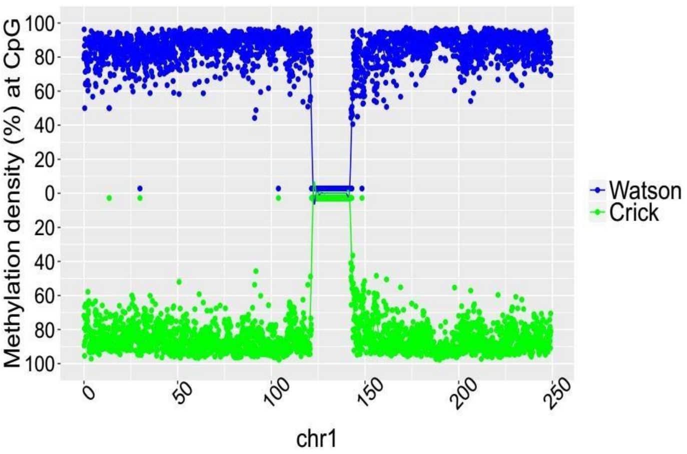
SRR921729 (55 year old male)



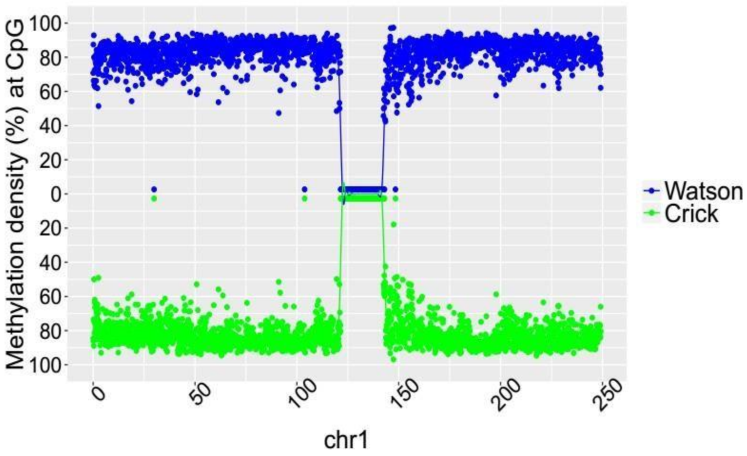
SRR921749 (55 year old male)



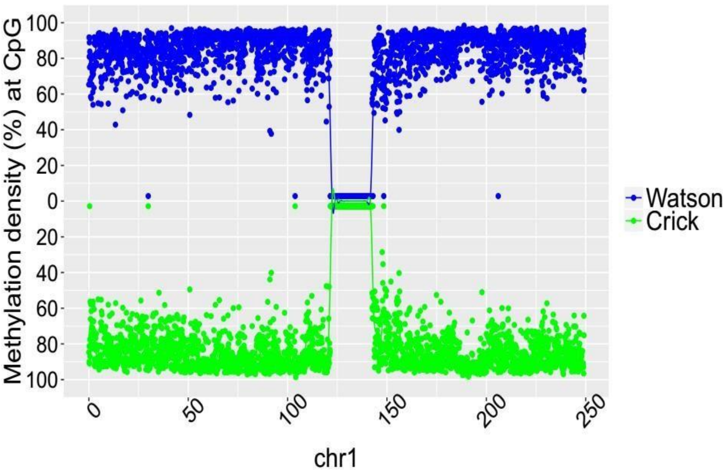
SRR921735 (55 year old male)



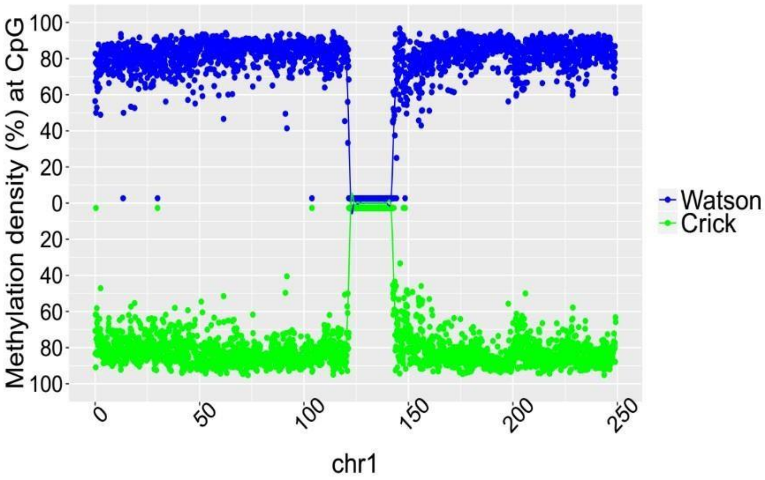
SRR847424 (42 year old female)



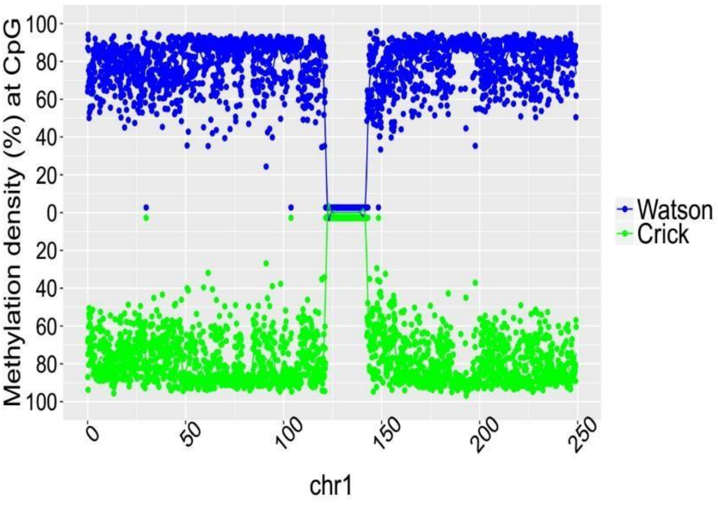
SRR921706 (53 year old female)



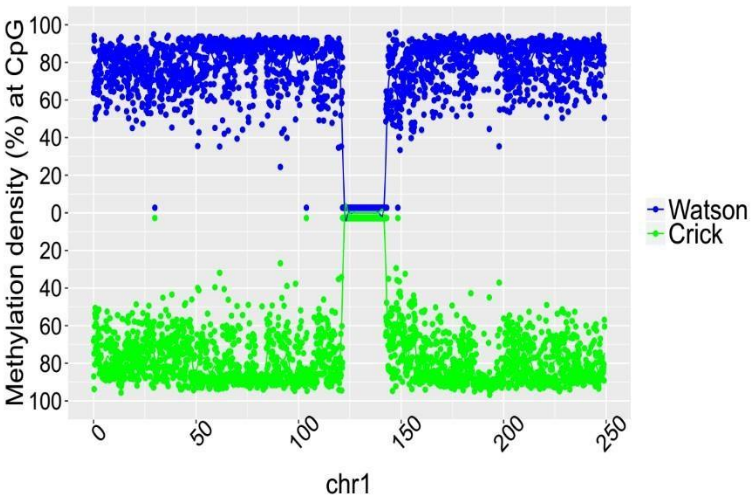
SRR921723 (53 year old female)



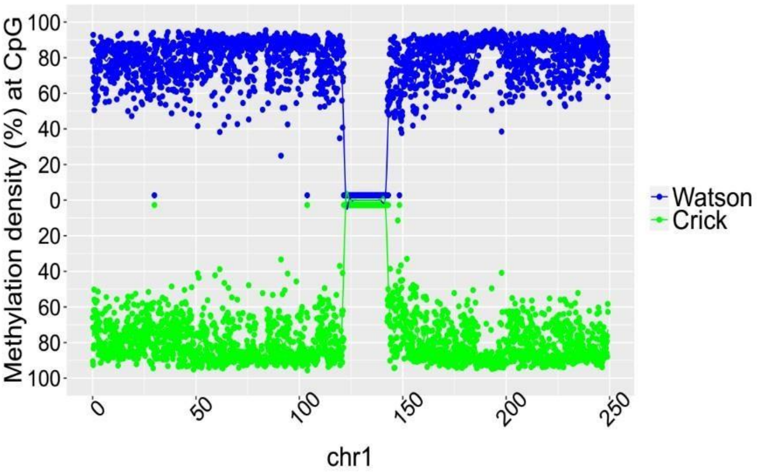
SRR921702 (64 year old female)



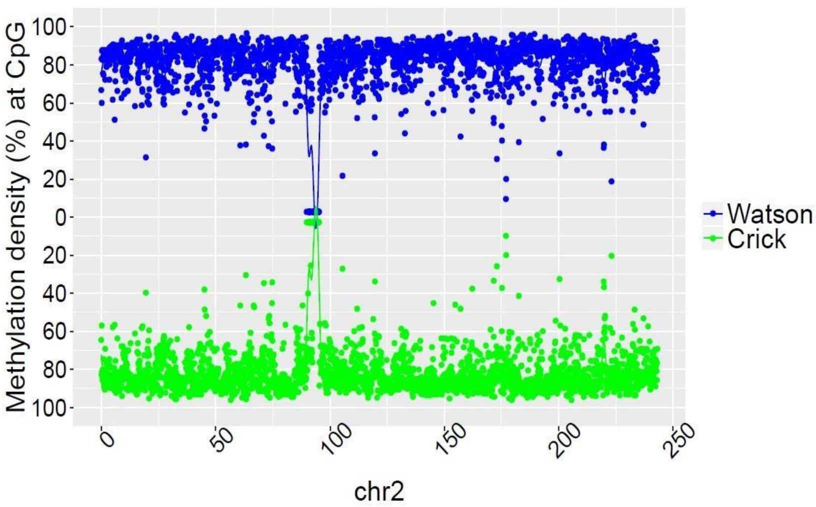
SRR949195 (81 year old female)



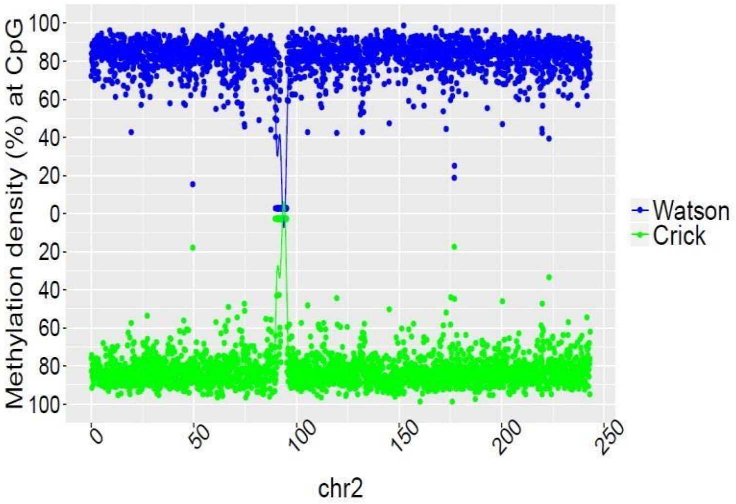
SRR949197 (82 year old female)



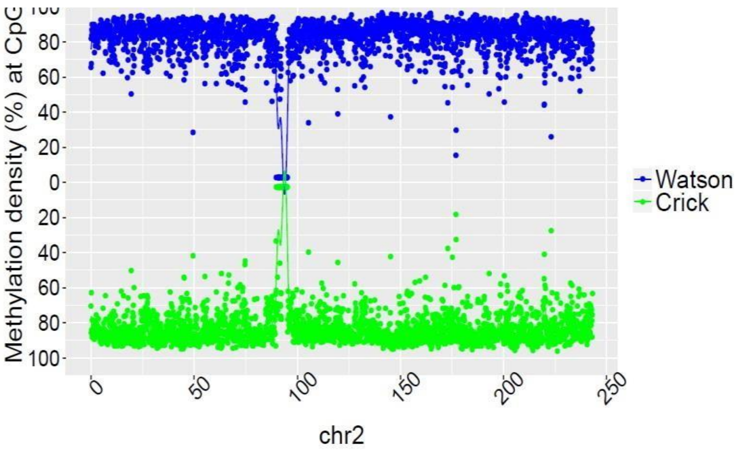
SRR901381 (20 weeks old male)



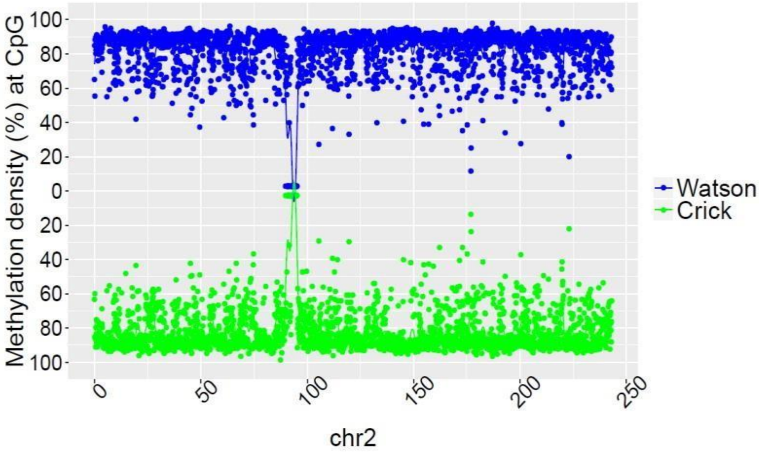
SRR47900 (31 year old male)



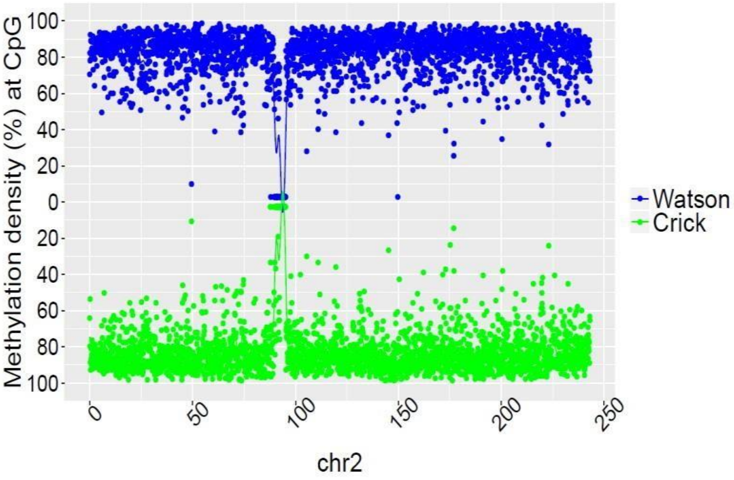
SRR 478994 (47 year old male)



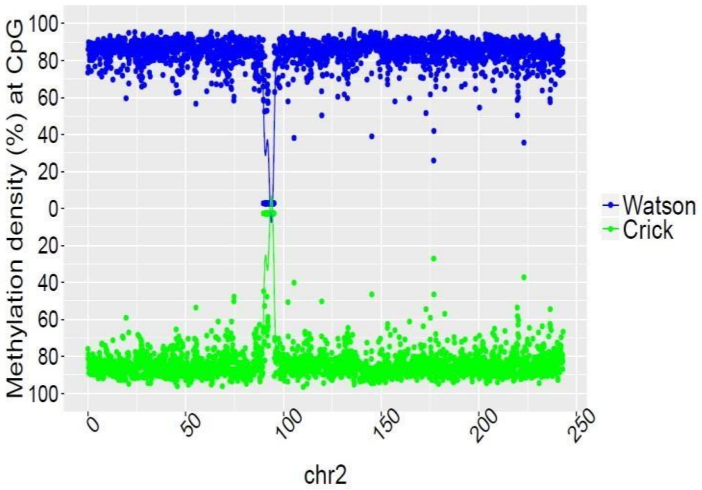
SRR478991 (48 year old male)



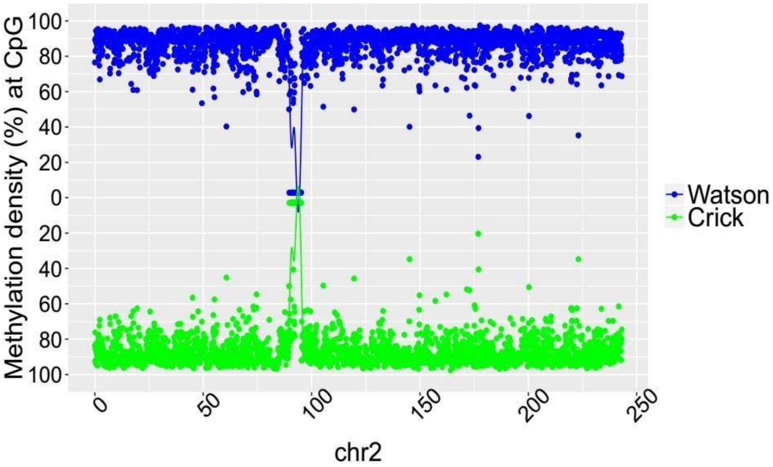
SRR921729 (55 year old male)



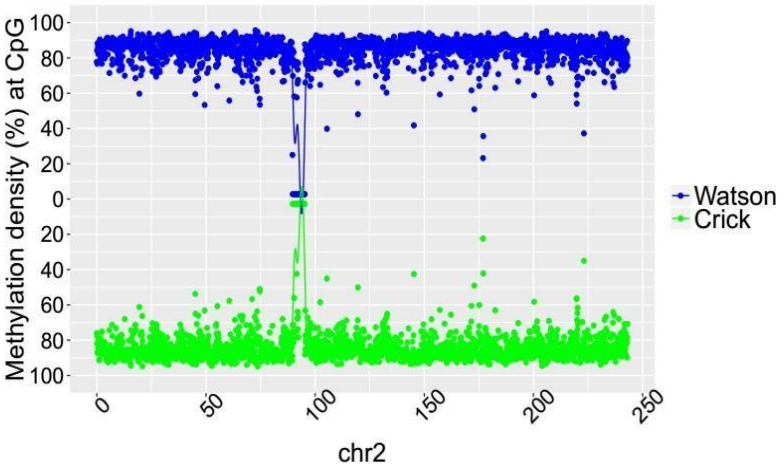
SRR921749 (55 year old male)



SRR921735 (55 year old male)

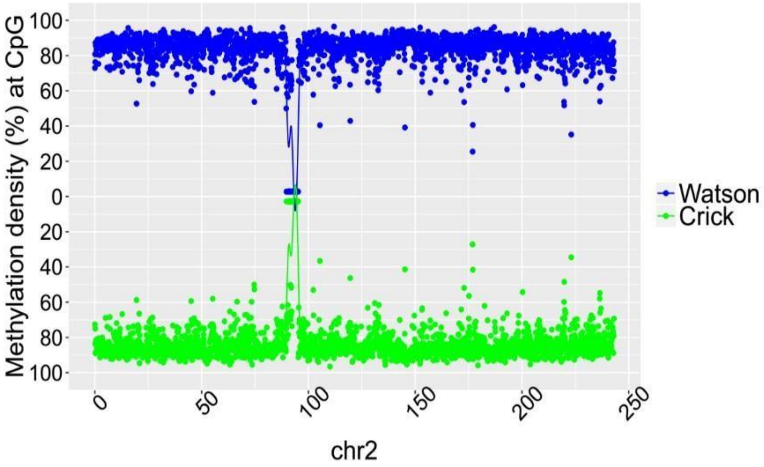


SRR847424 (42 year old female)



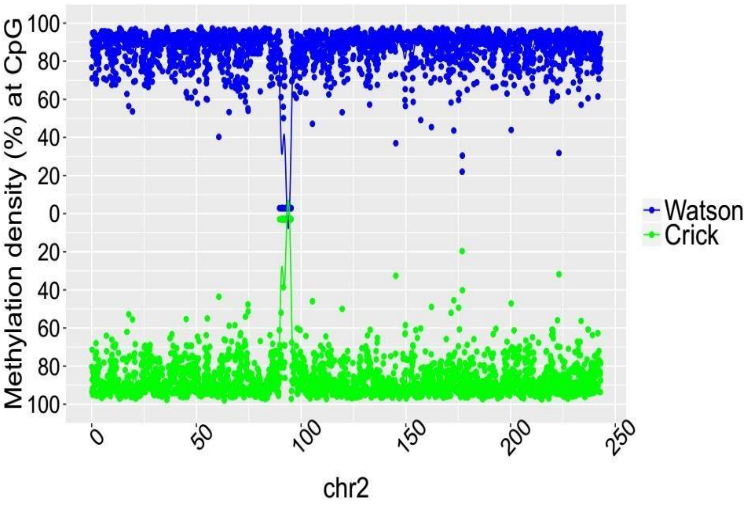
SRR921706 (53 year old female)

SRR921723 (53 year old female)

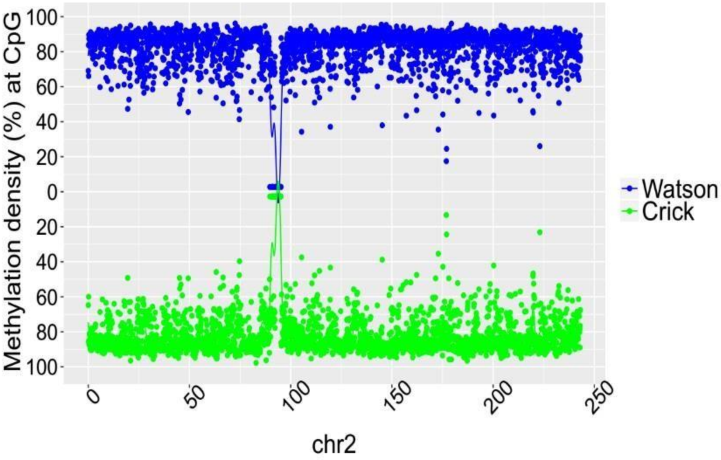


SRR921729 (55 year old male)

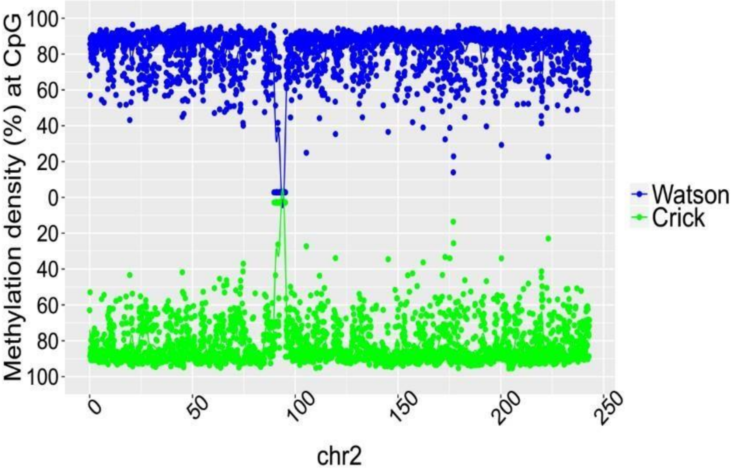
SRR921749 (55 year old male)



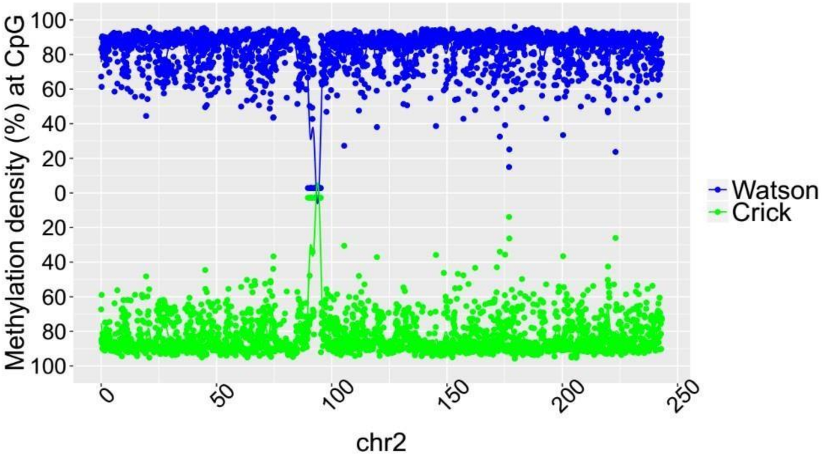
SRR921702 (64 year old female)



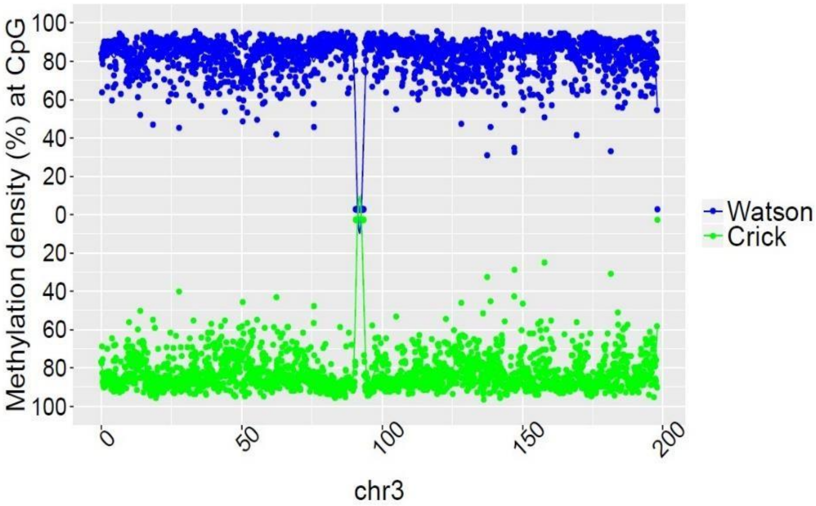
SRR949195 (81 year old female)



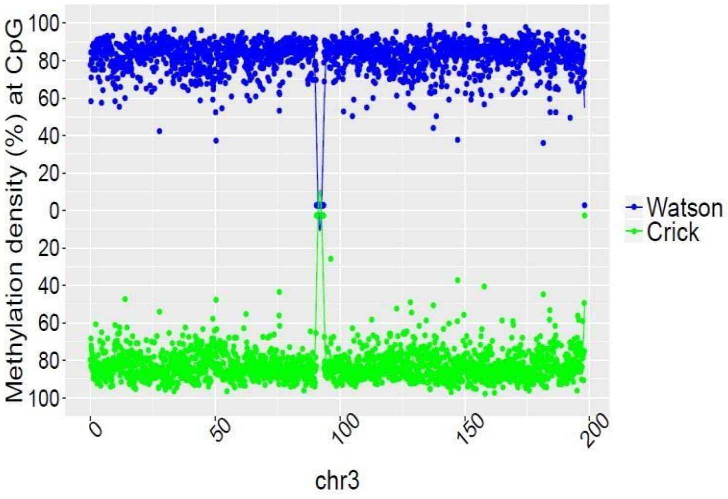
SRR949197 (82 year old female)



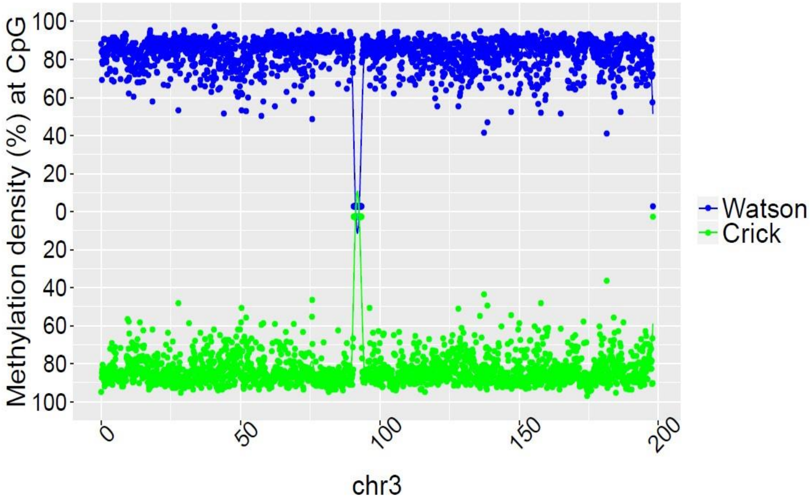
SRR901381 (20 weeks old male)



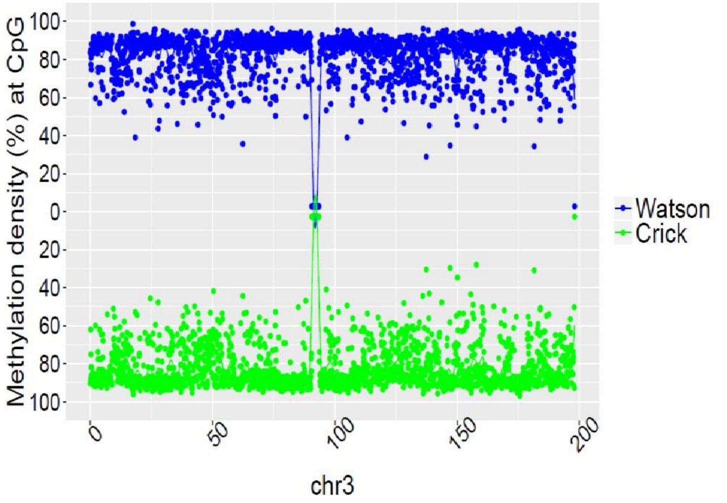
SRR47900 (31 year old male)



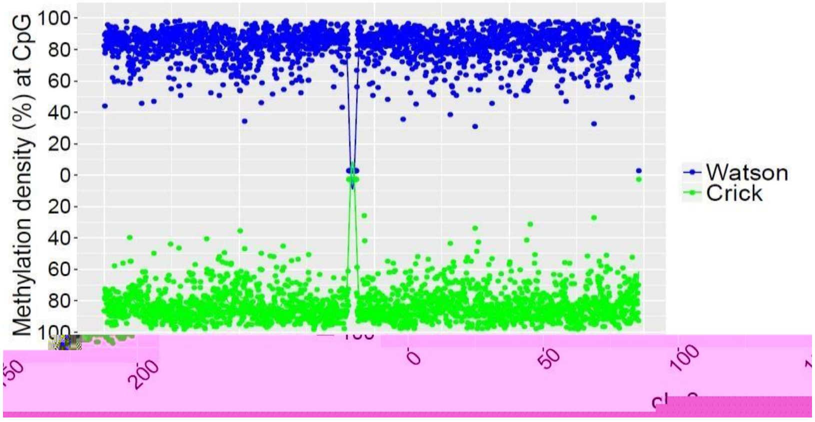
SRR 478994 (47 year old male)



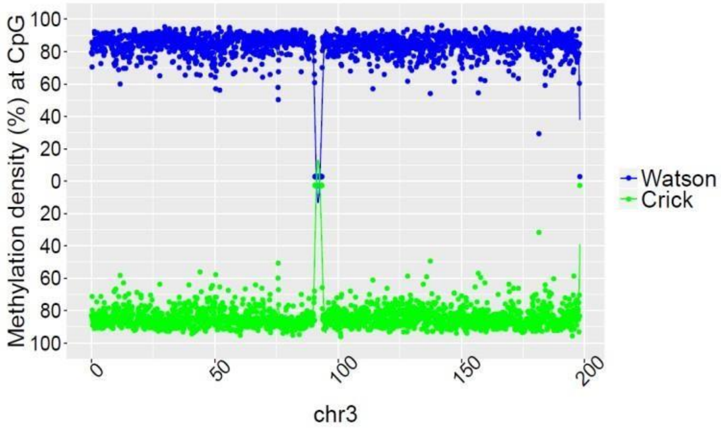
SRR478991 (48 year old male)



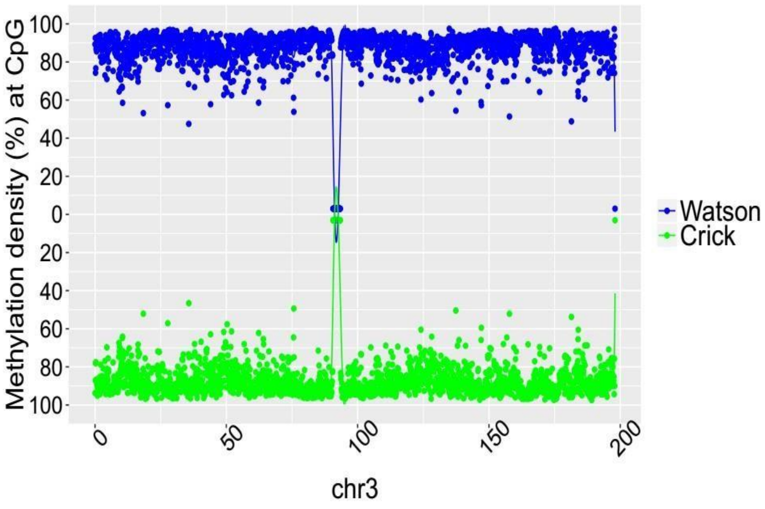
SRR921729 (55 year old male)



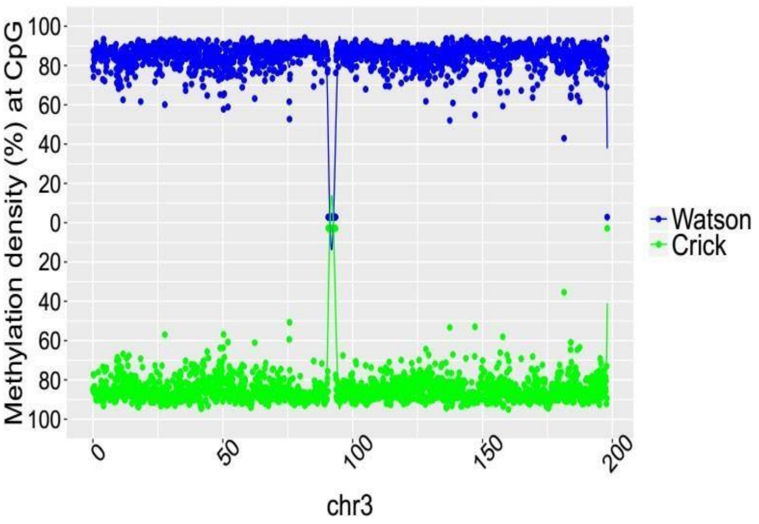
SRR921749 (55 year old male)



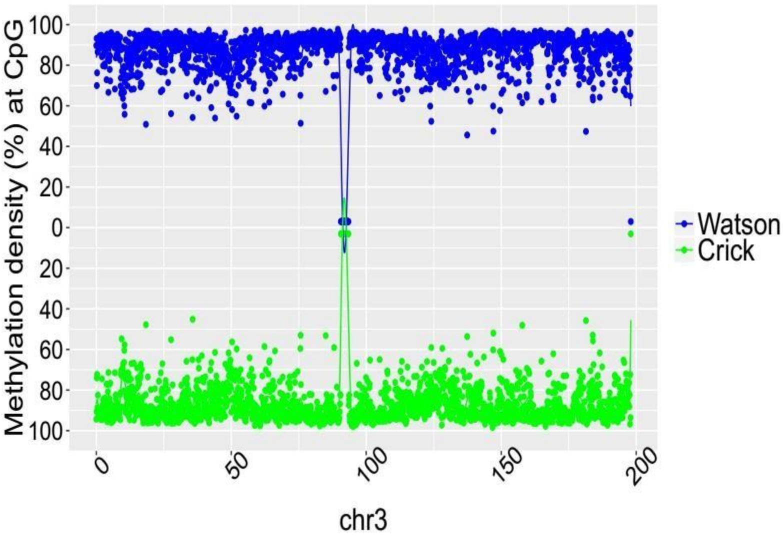
SRR921735 (55 year old male)



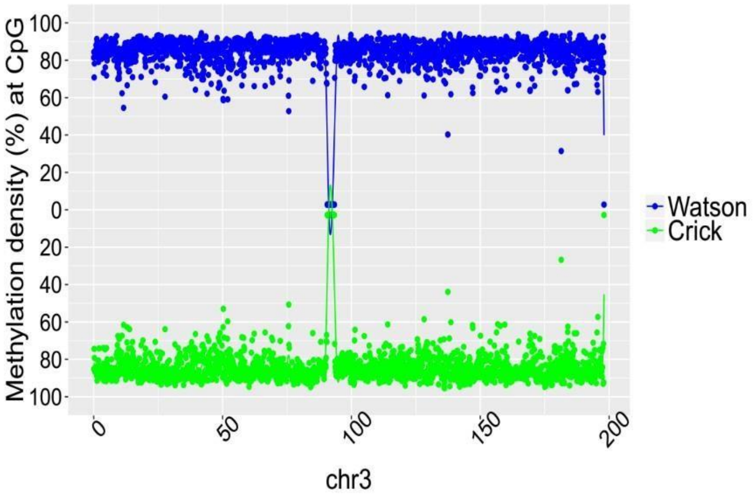
SRR847424 (42 year old female)



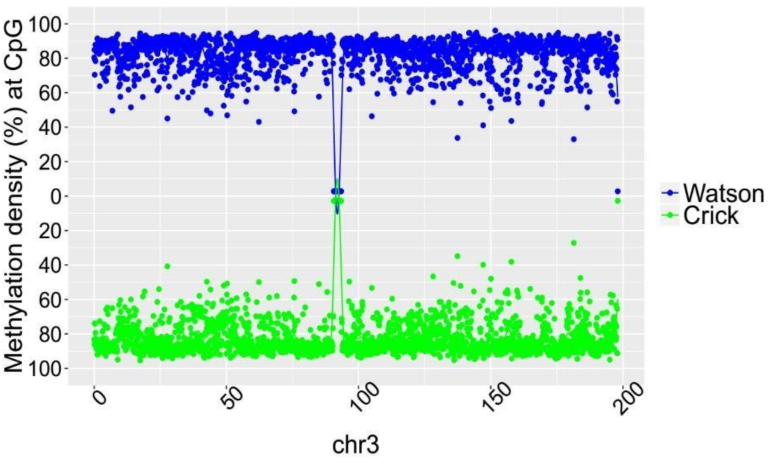
SRR921706 (53 year old female)



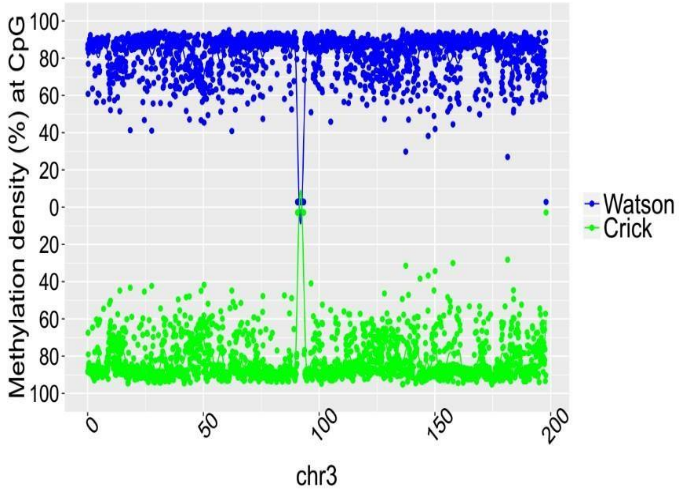
SRR921723 (53 year old female)



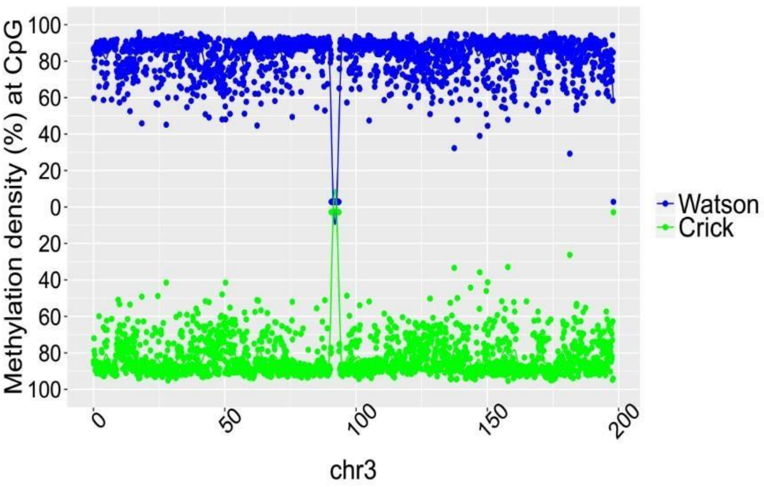
SRR921702 (64 year old female)



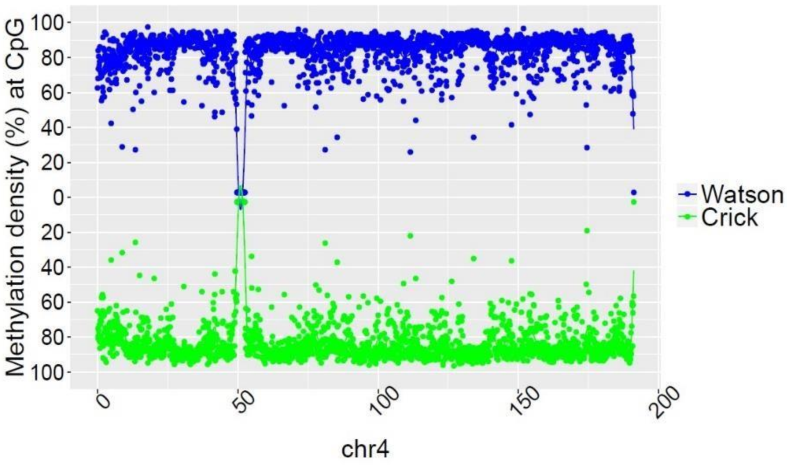
SRR949195 (81 year old female)



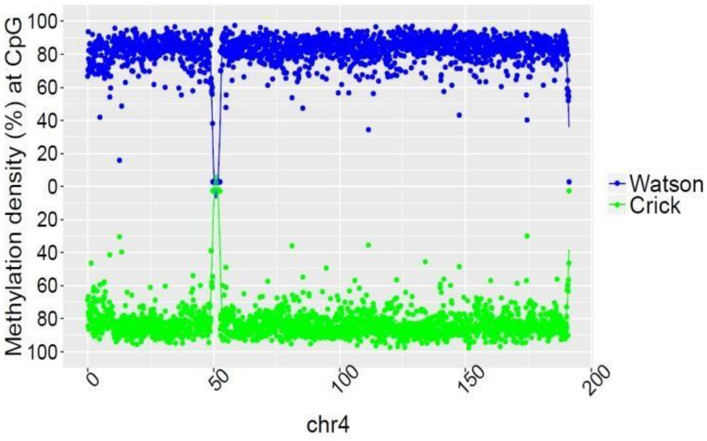
SRR949197 (82 year old female)



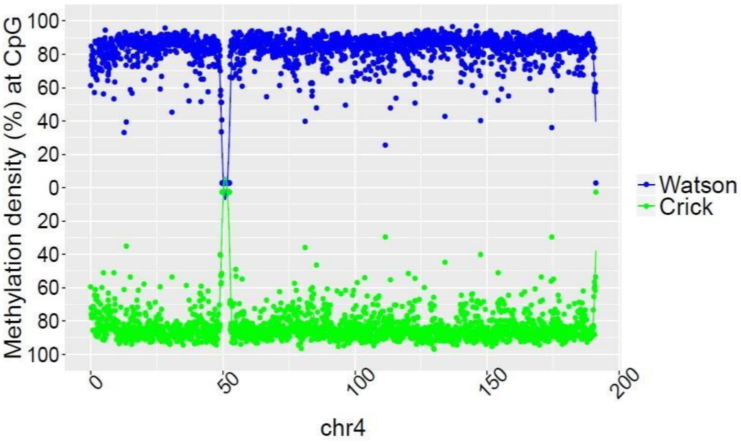
SRR901381 (20 week male)



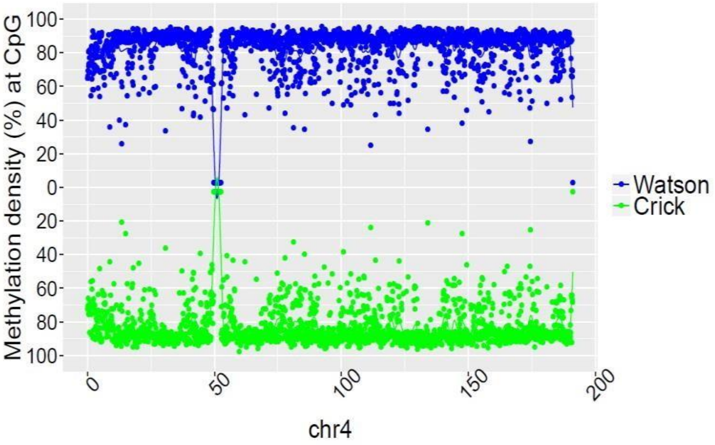
SRR47900 (31 year old male)



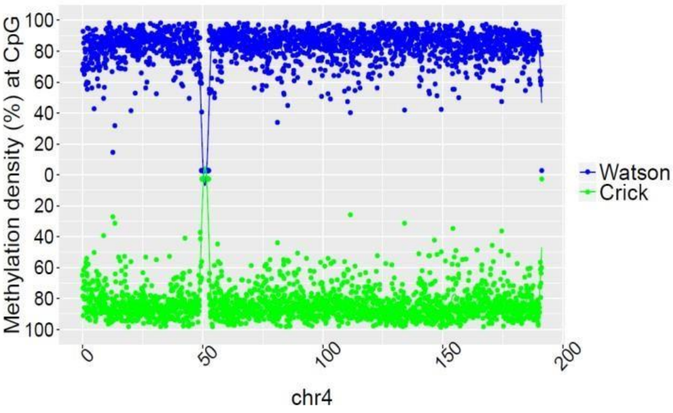
SRR 478994 (47 year old male)



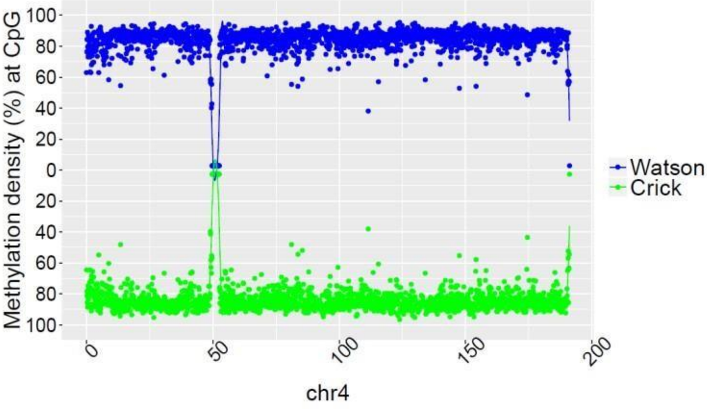
SRR478991 (48 year old male)



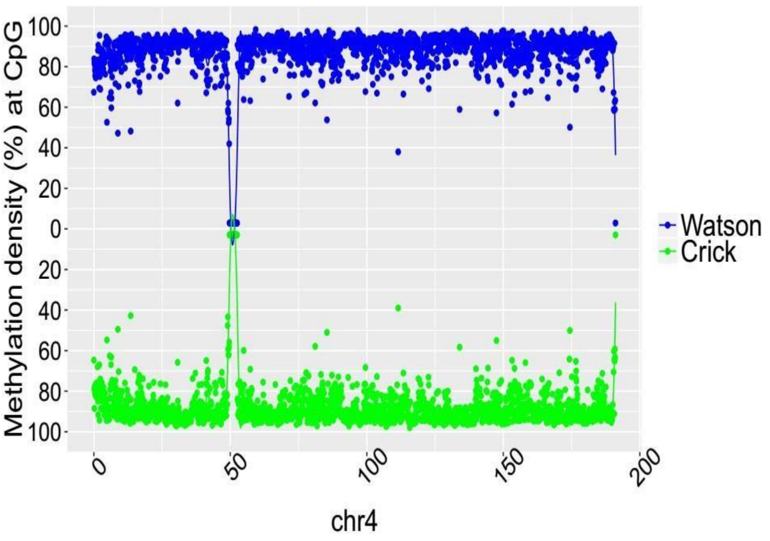
SRR921729 (55 year old male)



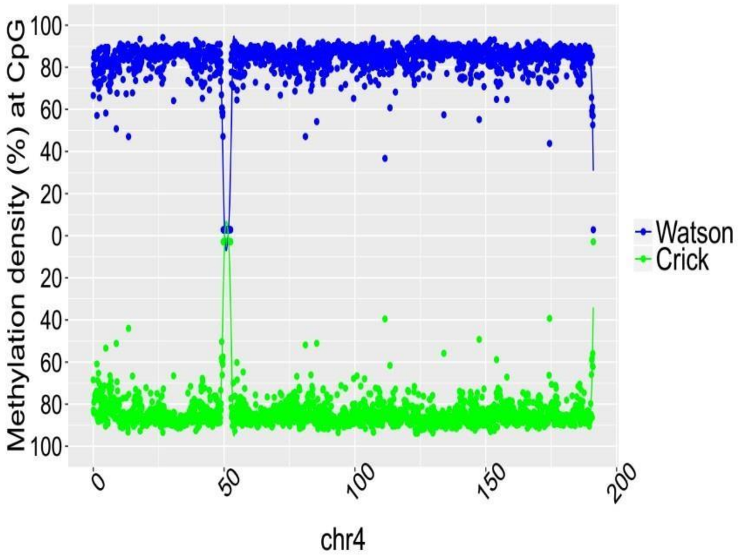
SRR921749 (55 year old male)



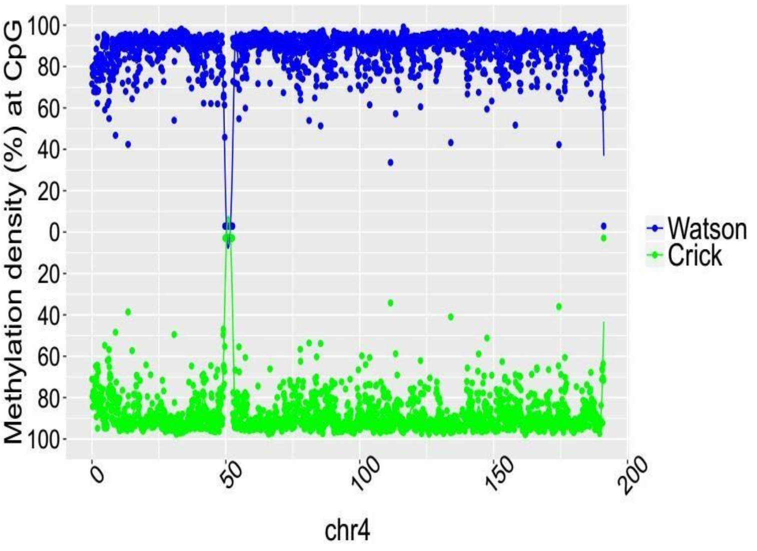
SRR921735 (55 year old male)



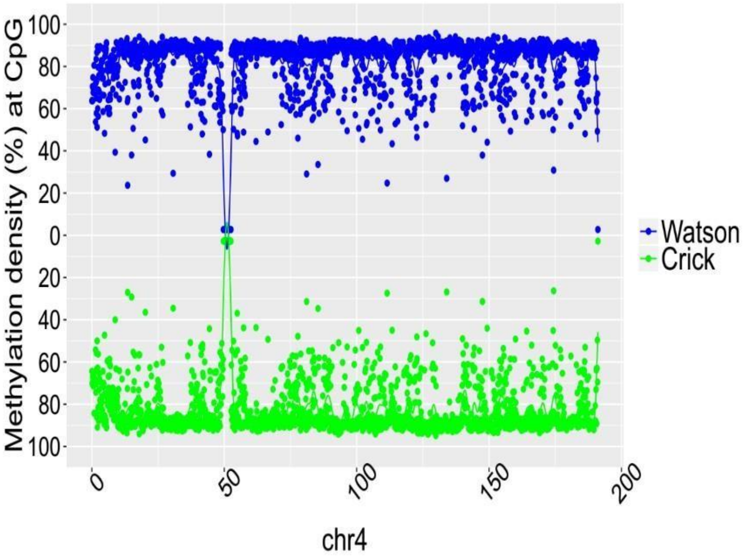
SRR847424 (42 year old female)



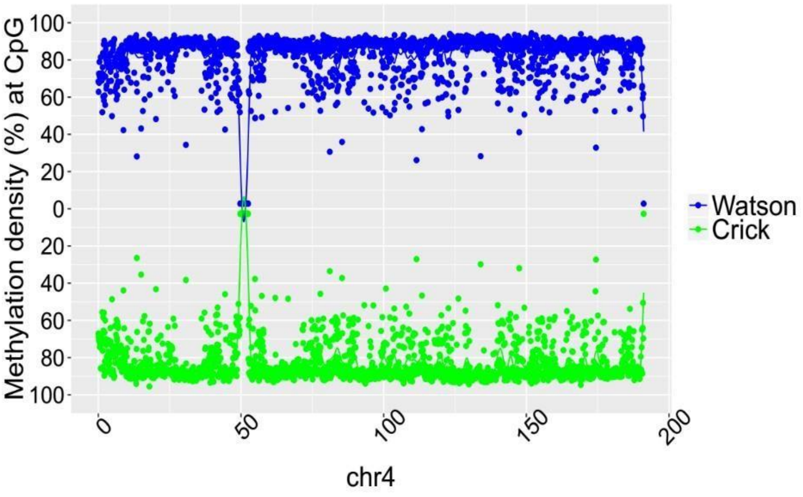
SRR921702 (64 year old female)



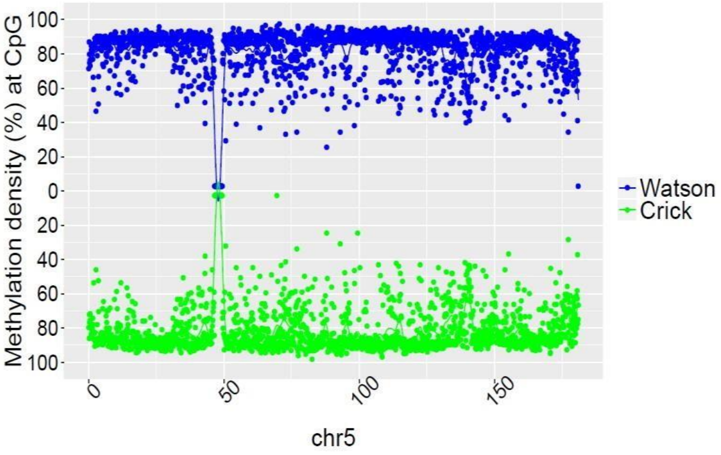
SRR949195 (81 year old female)



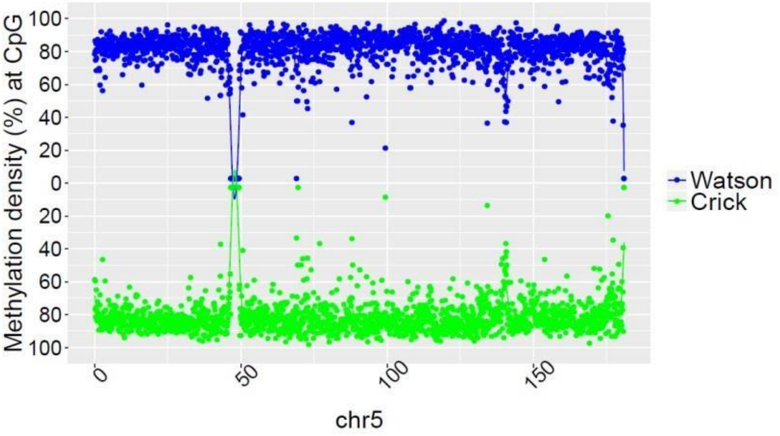
SRR949197 (82 year old female)



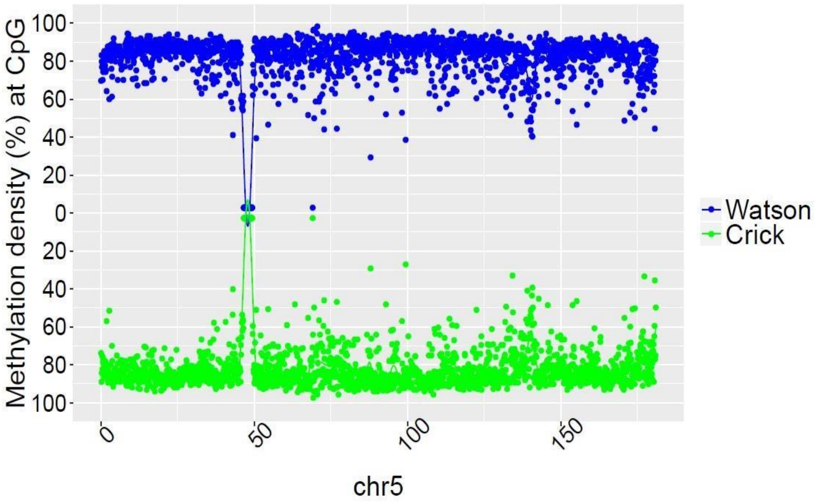
SRR901381 (20 weeks old male)



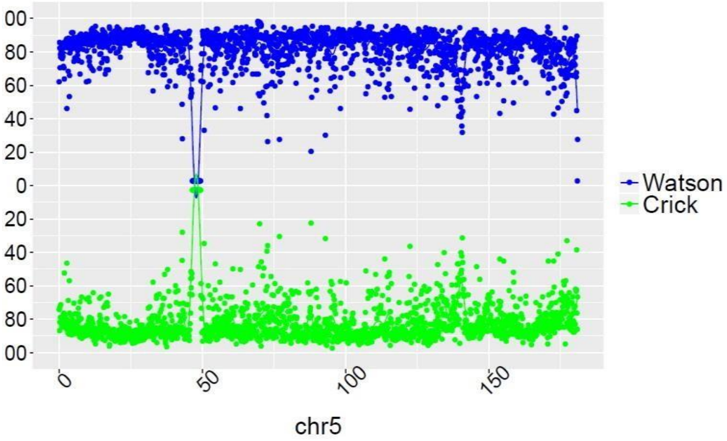
SRR47900 (31 year old male)



SRR 478994 (47 year old male)

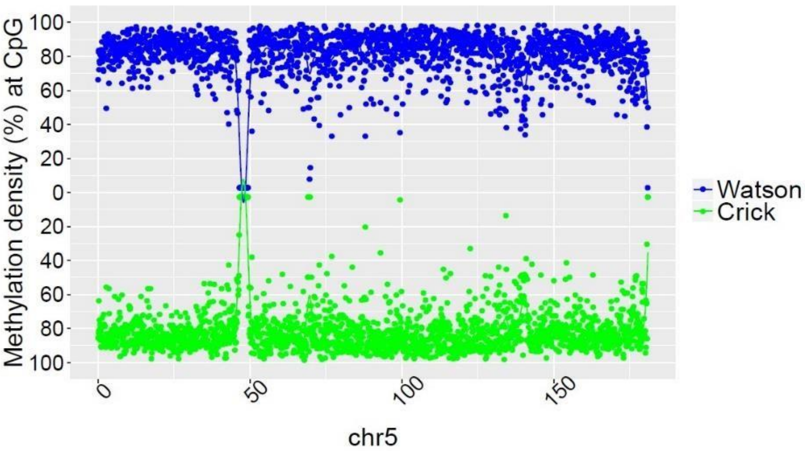


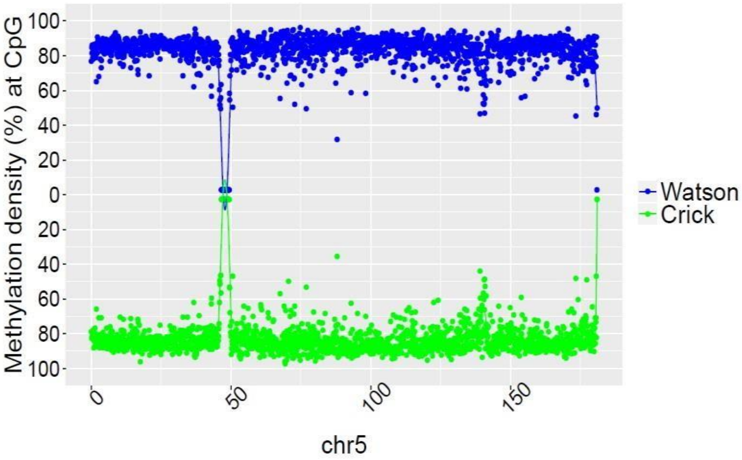
SRR478991 (48 year old male)



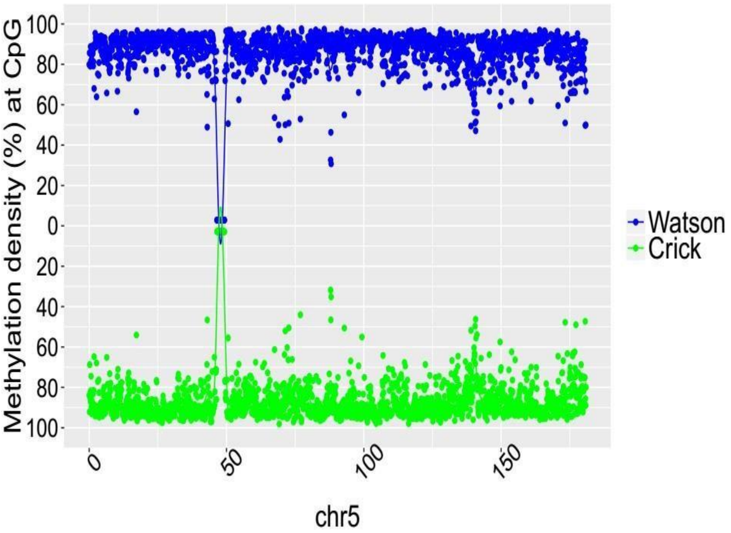
SRR921729 (55 year old male)

SRR921749 (55 year old male)

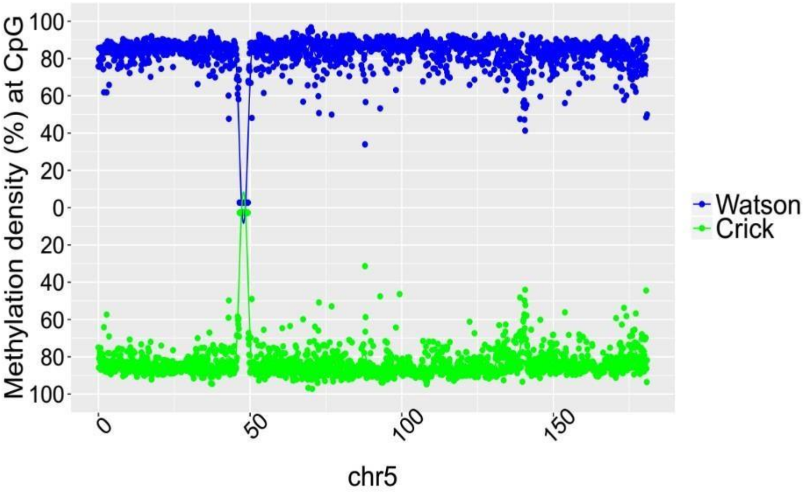

Methylation density (%) at CpG
100
80
60
40
20
0
20
40
60
80
100
0
50
100
150
chr5
Watson
Crick


Methylation density (%) at CpG
100
80
60
40
20
0
20
40
60
80
100
0
50
100
150
chr5
Watson
Crick

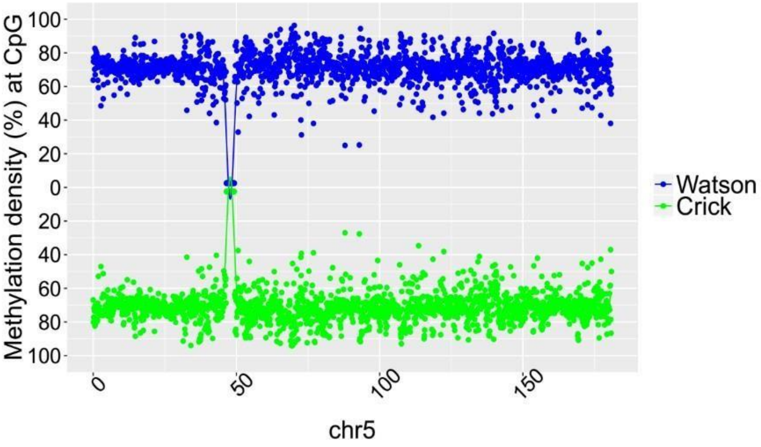
SRR921735 (55 year old male)



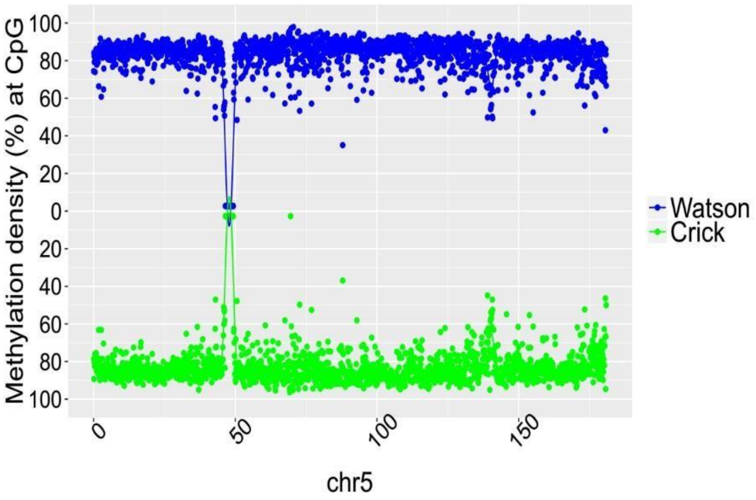
SRR847424 (42 year old female)



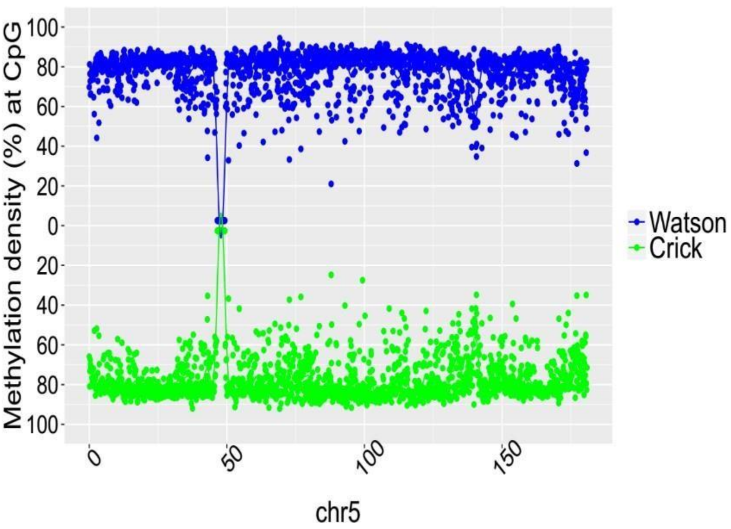
SRR921706 (53 year old female)



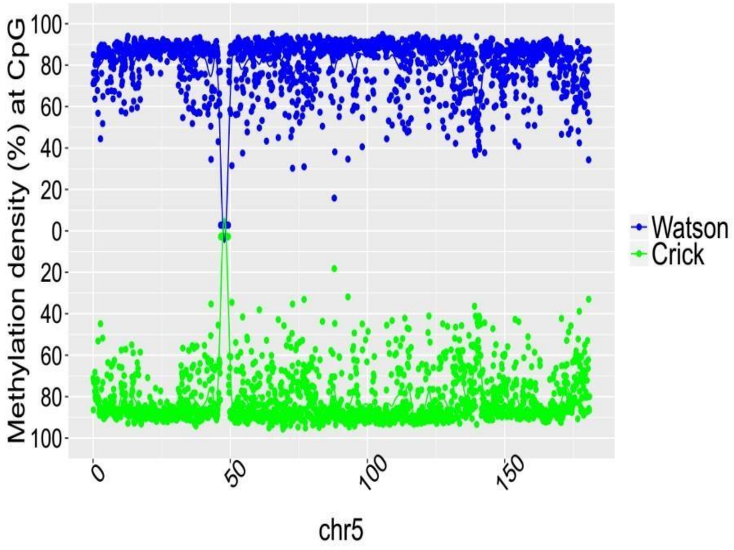
SRR921723 (53 year old female)



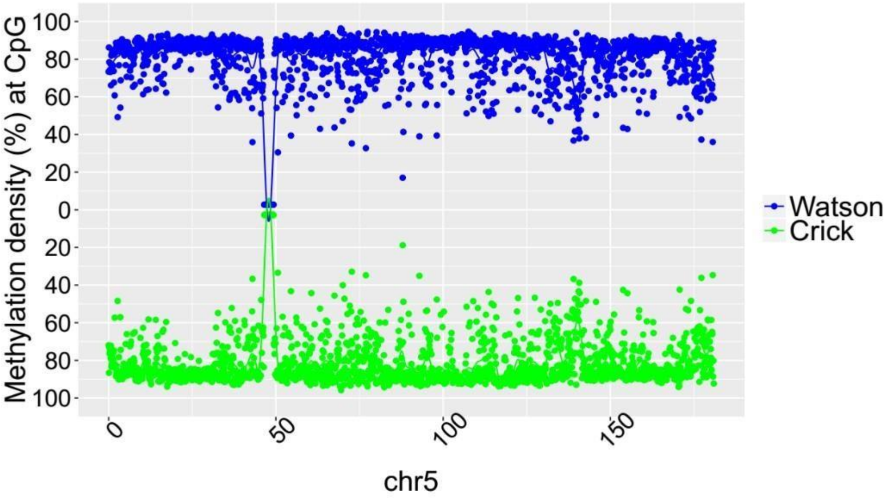
SRR921702 (64 year old female)



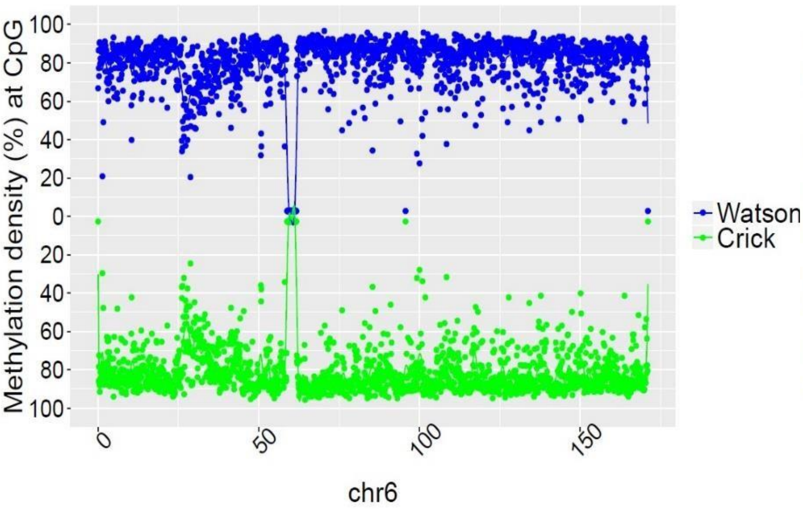
SRR949195 (81 year old female)



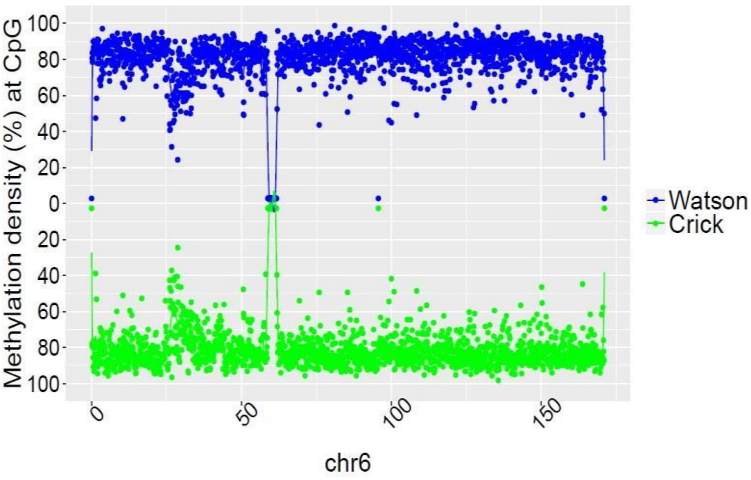
SRR949197 (82 year old female)



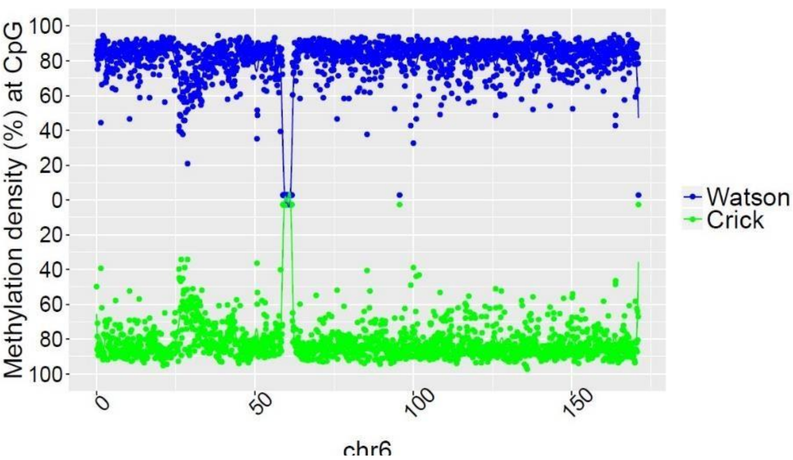
SRR901381 (20 weeks old male)



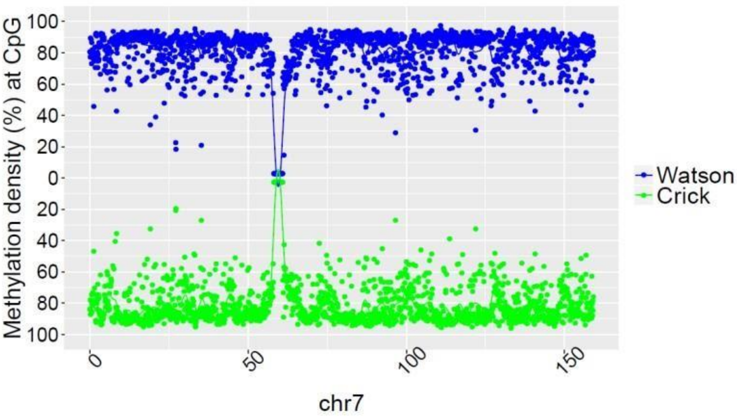
SRR47900 (31 year old male)



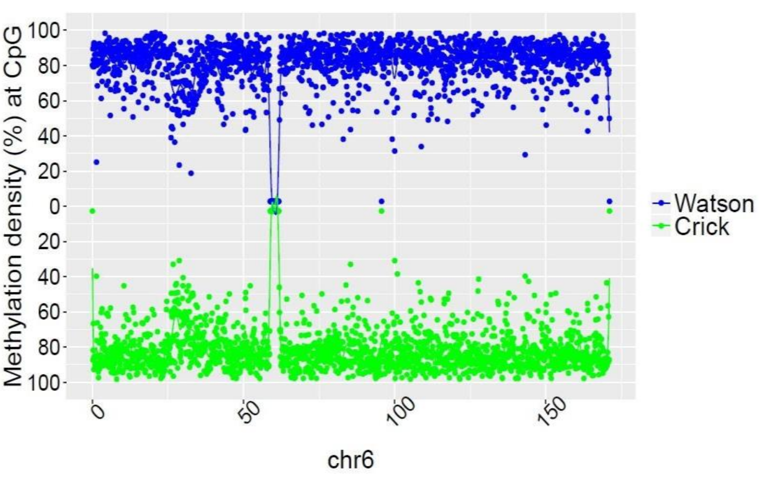
SRR 478994 (47 year old male)



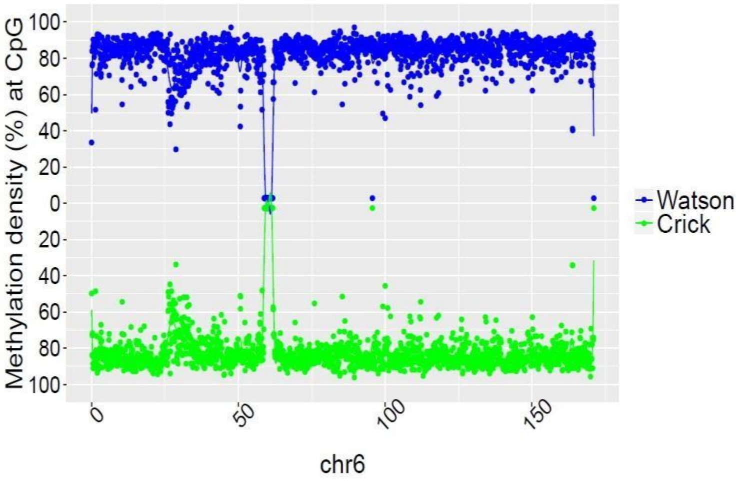
SRR478991 (48 year old male)



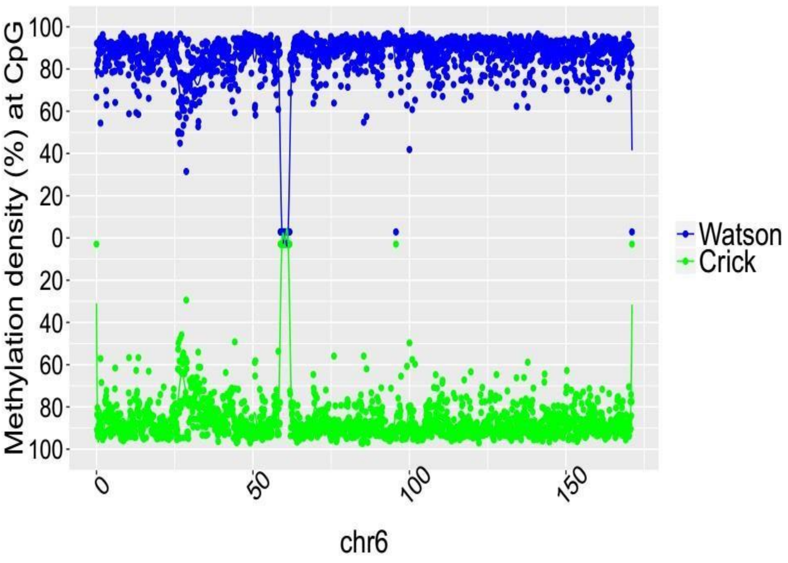
SRR921729 (55 year old male)



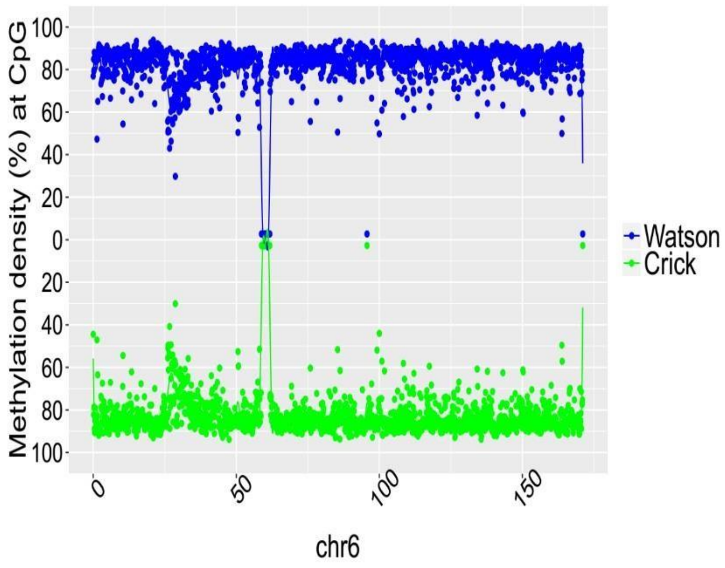
SRR921749 (55 year old male)



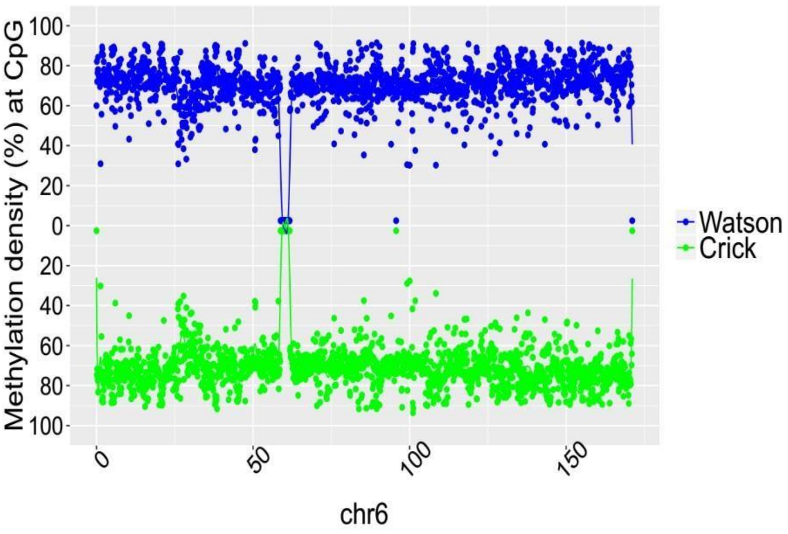
SRR921735 (55 year old male)



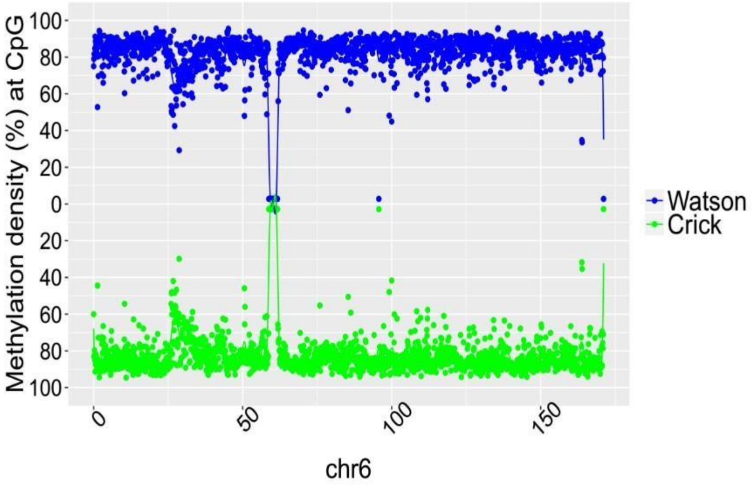
SRR847424 (42 year old female)



SRR921706 (53 year old female)

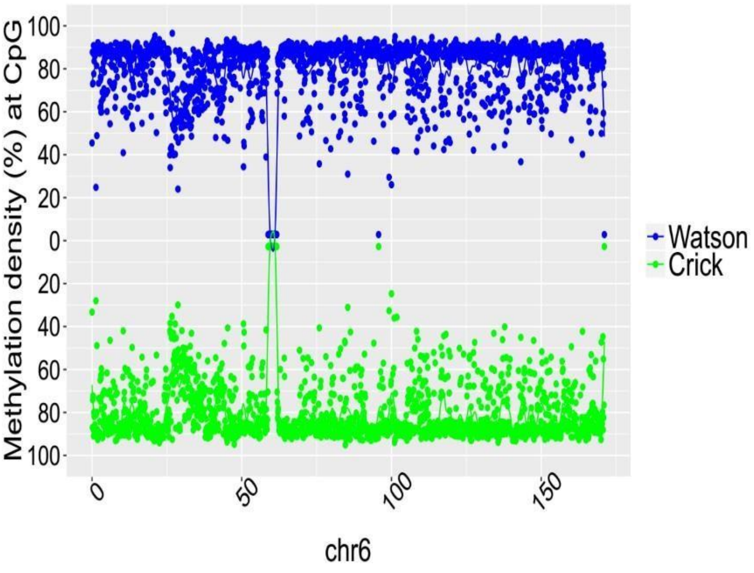


SRR921723 (53 year old female)



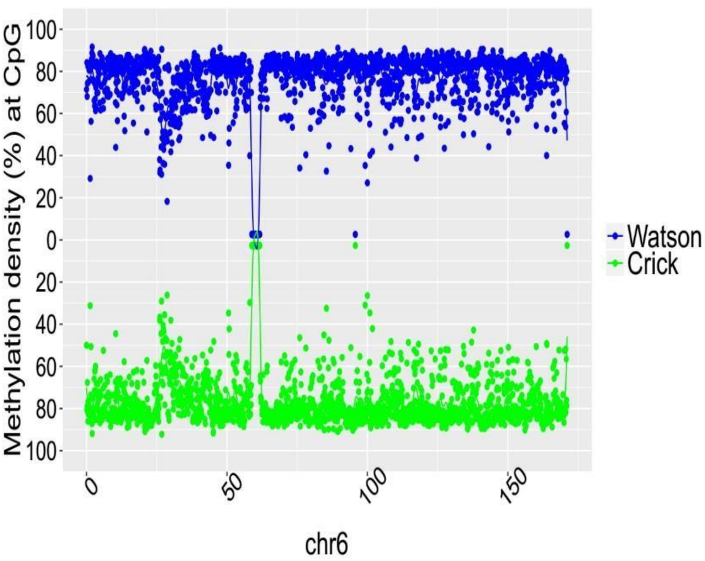
SRR921702 (64 year old female)

SRR949195 (81 year old female)

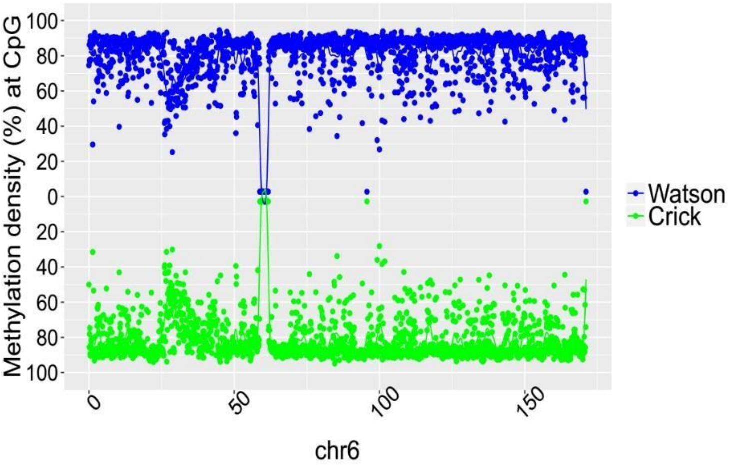


SRR921735 (55 year old male)

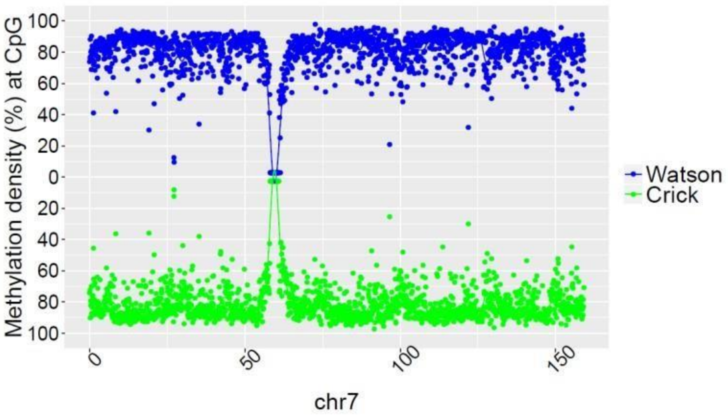
SRR847424 (42 year old female)



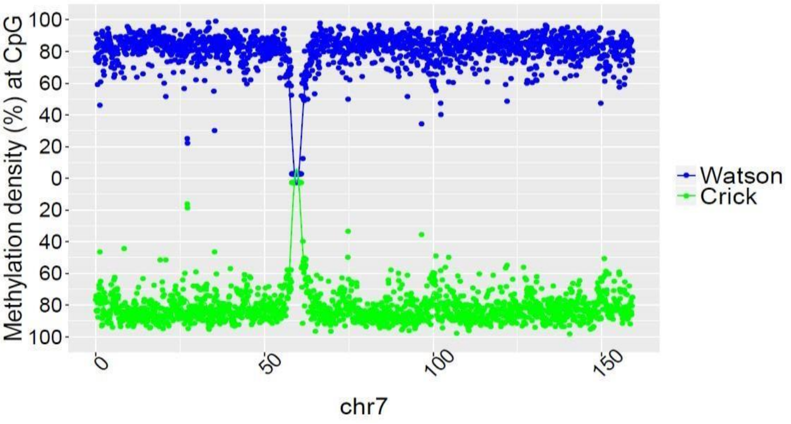
SRR949197 (82 year old female)



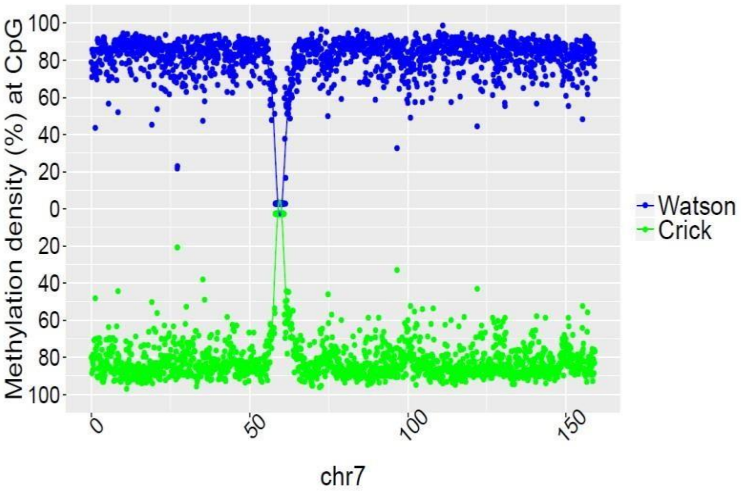
SRR901381 (20 weeks old male)



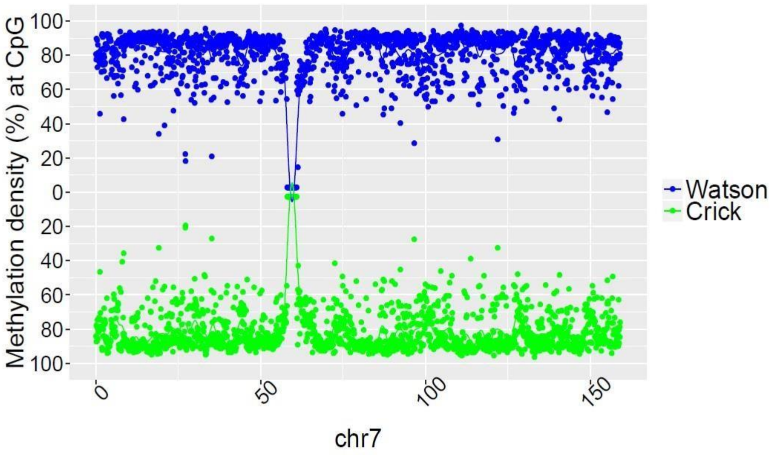
SRR47900 (31 year old male)



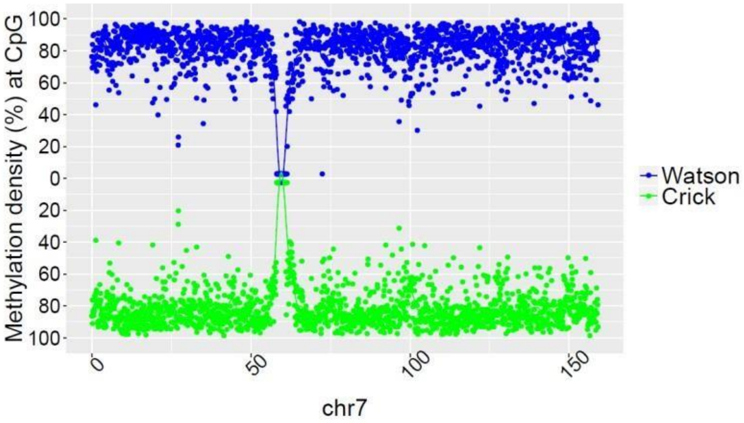
SRR 478994 (47 year old male)



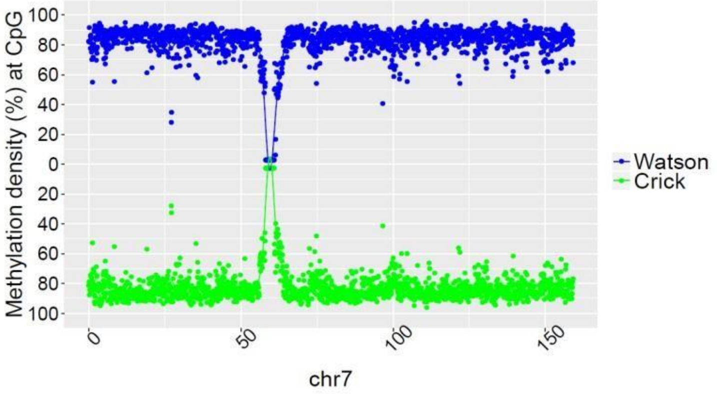
SRR478991 (48 year old male)



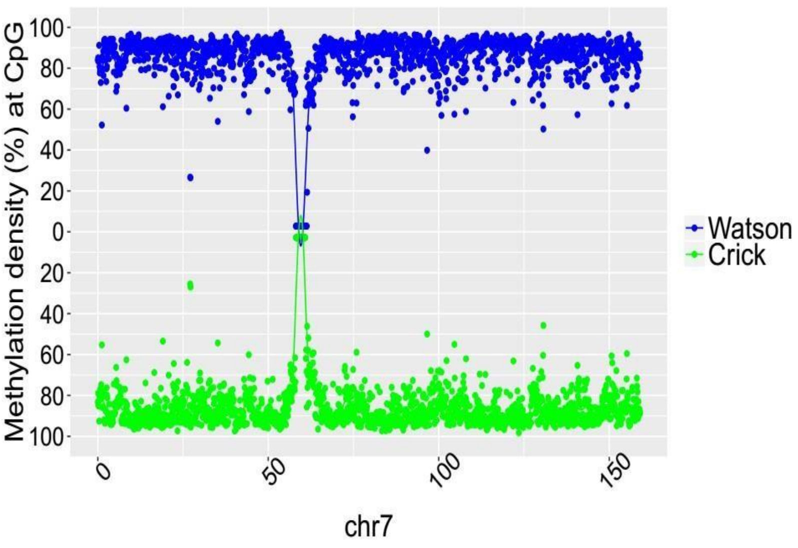
SRR921729 (55 year old male)



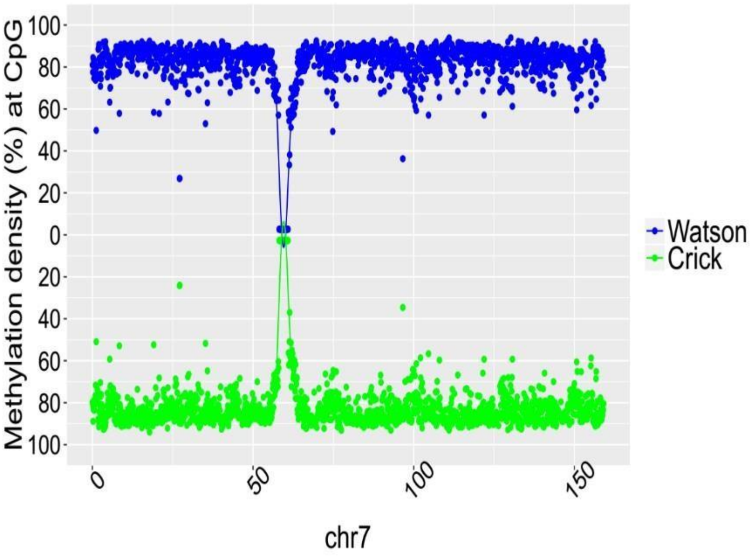
SRR921749 (55 year old male)



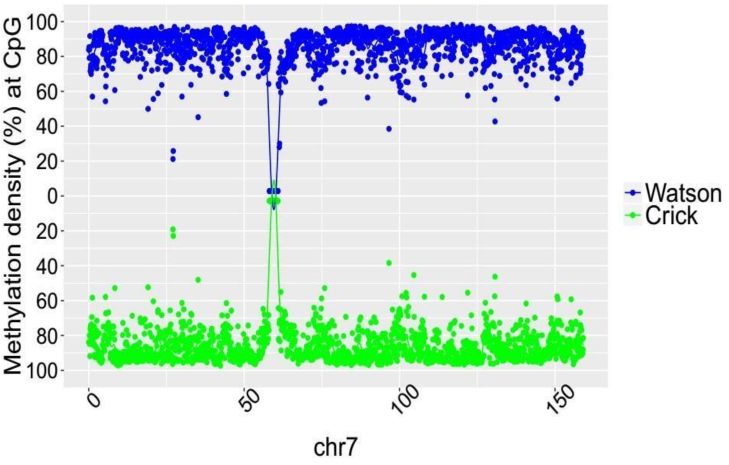
SRR921735 (55 year old male)



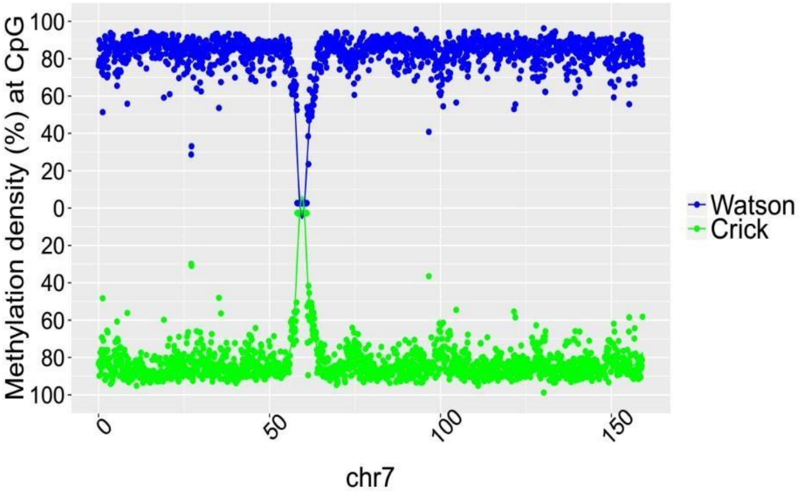
SRR847424 (42 year old female)



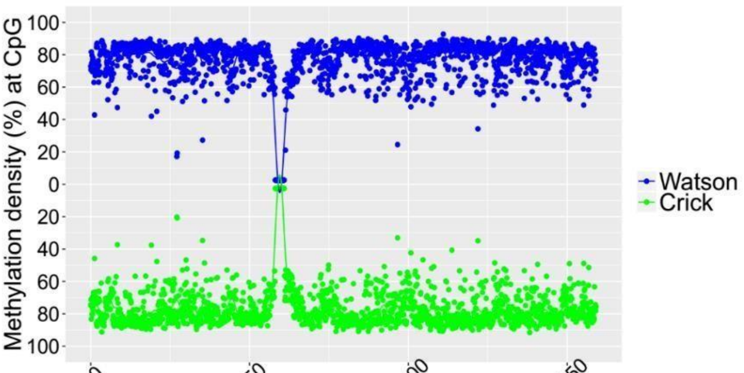
SRR921706 (53 year old female)



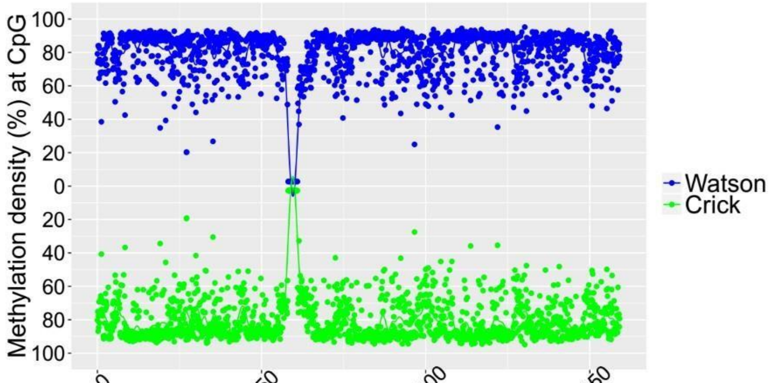
SRR921723 (53 year old female)



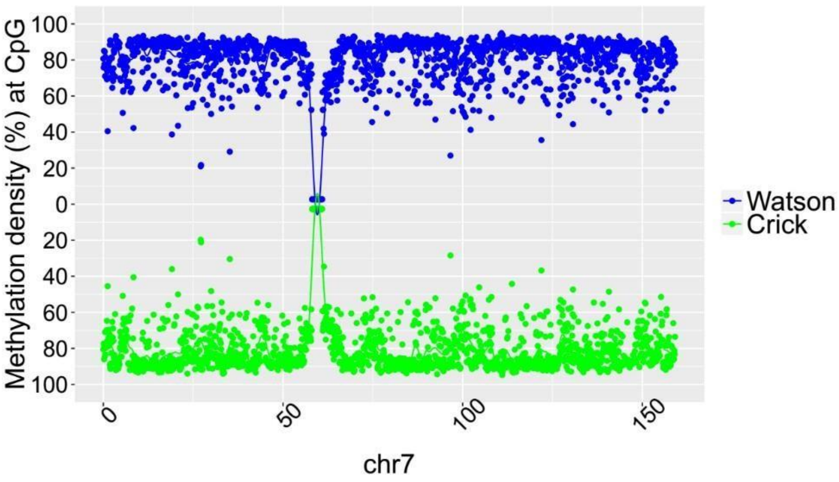
SRR921702 (64 year old female)



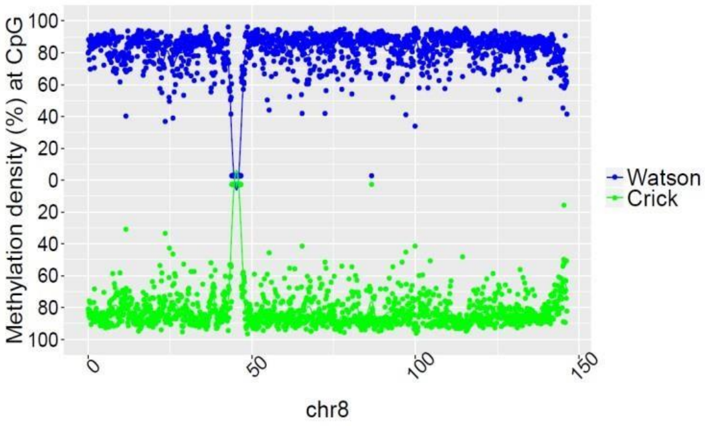
SRR949195 (81 year old female)



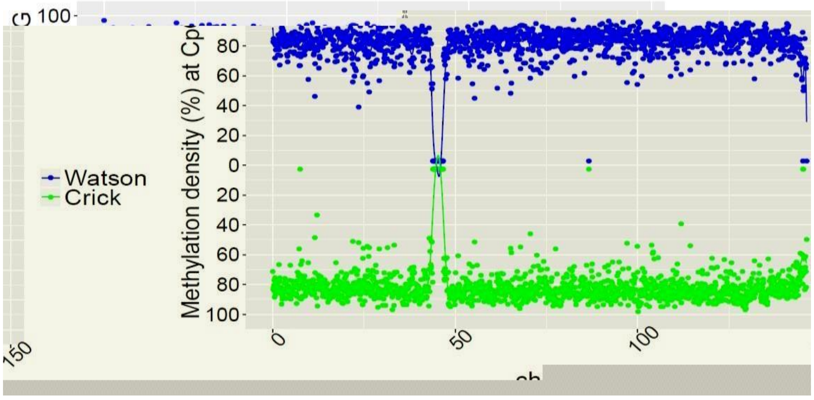
SRR949197 (82 year old female)



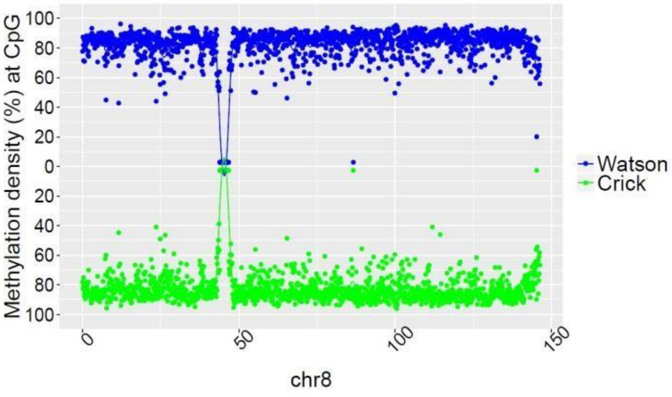
SRR901381 (20 weeks old male)



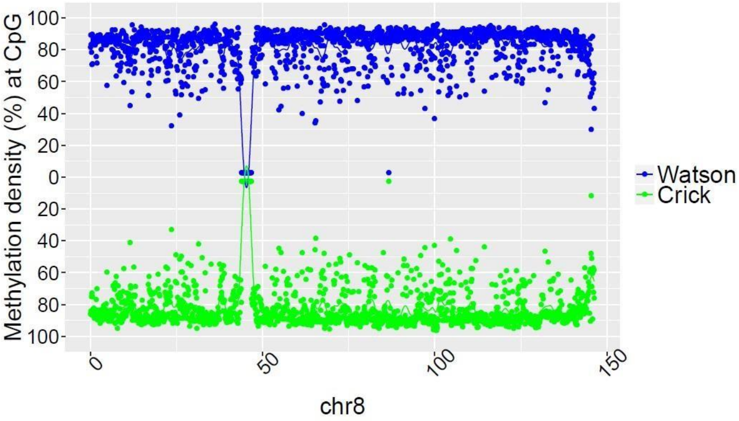
SRR47900 (31 year old male)



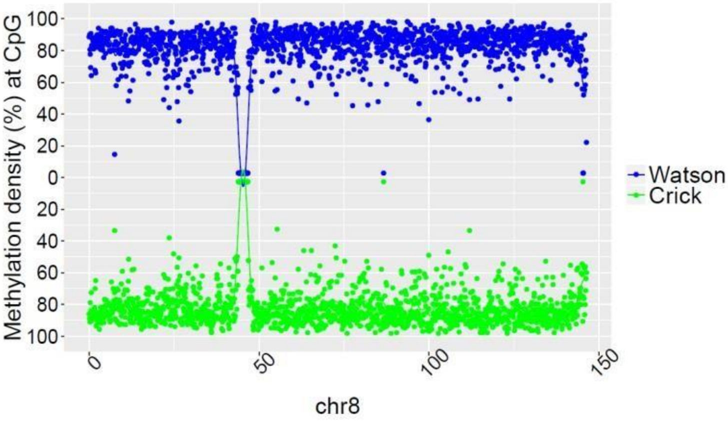
SRR 478994 (47 year old male)



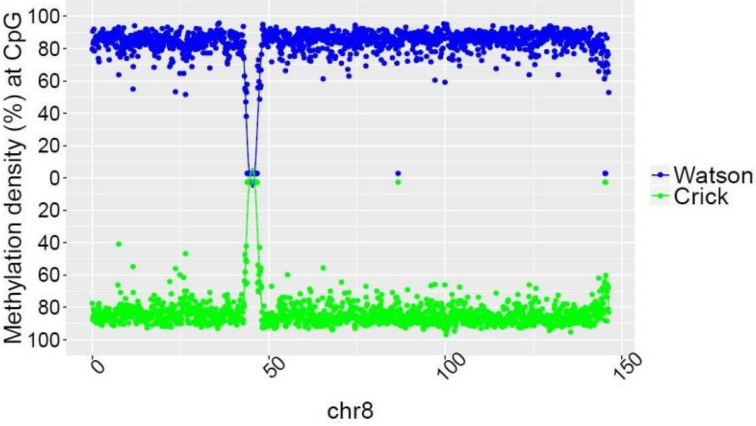
SRR478991 (48 year old male)



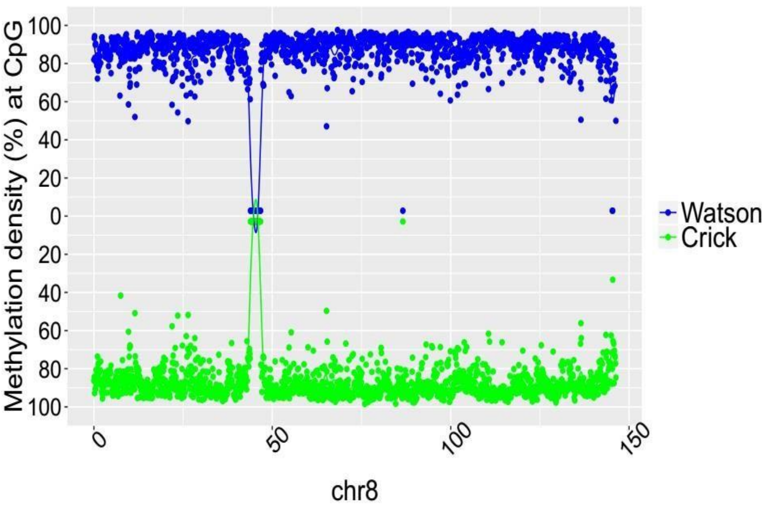
SRR921729 (55 year old male)



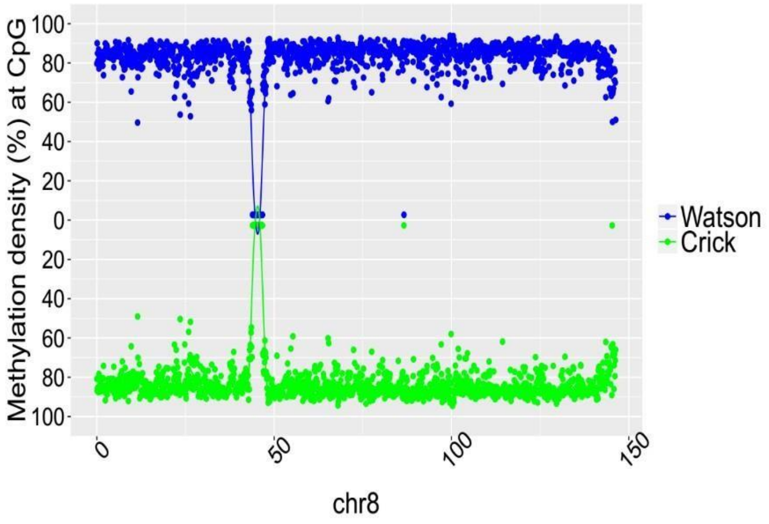
SRR921749 (55 year old male)



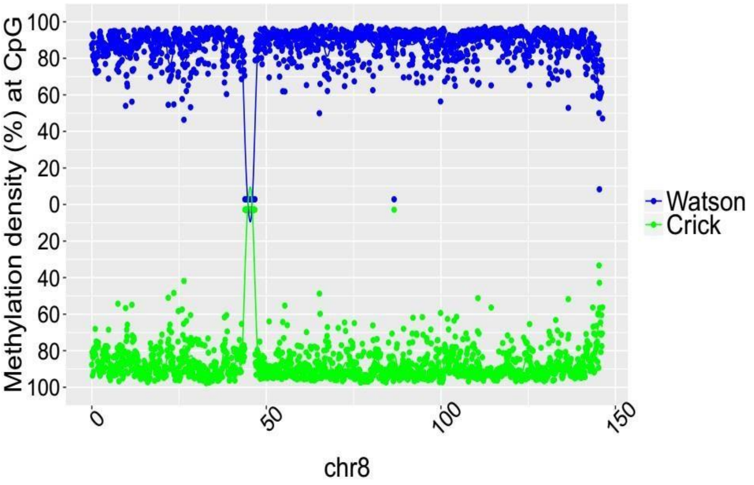
SRR921735 (55 year old male)



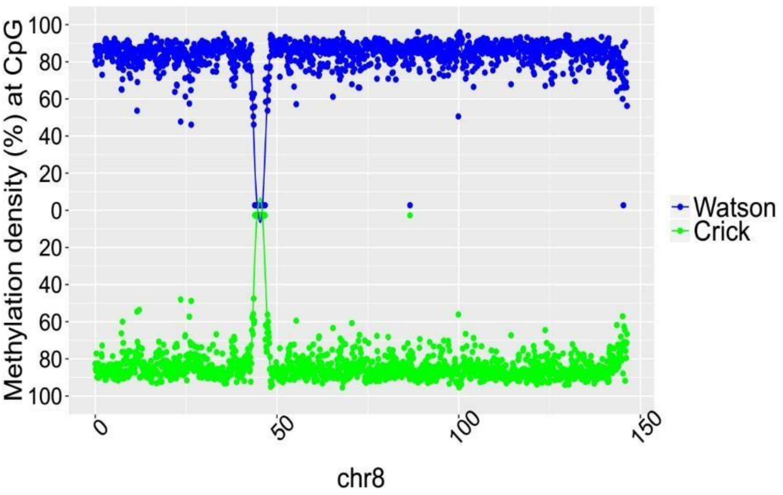
SRR847424 (42 year old female)



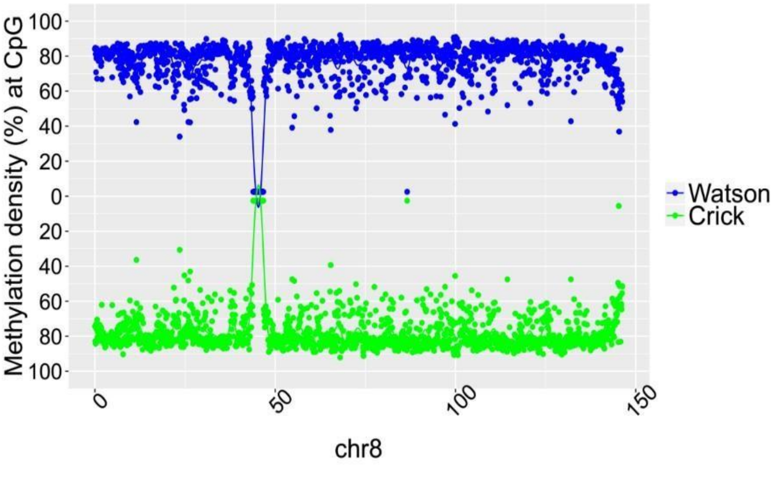
SRR921706 (53 year old female)



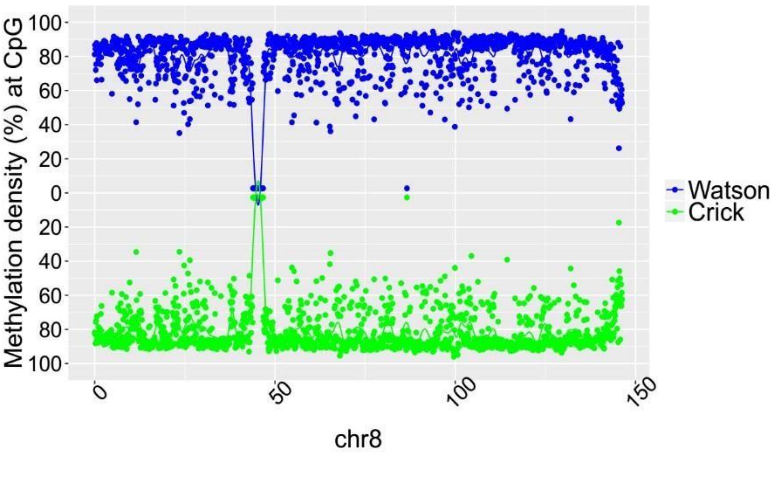
SRR921723 (53 year old female)



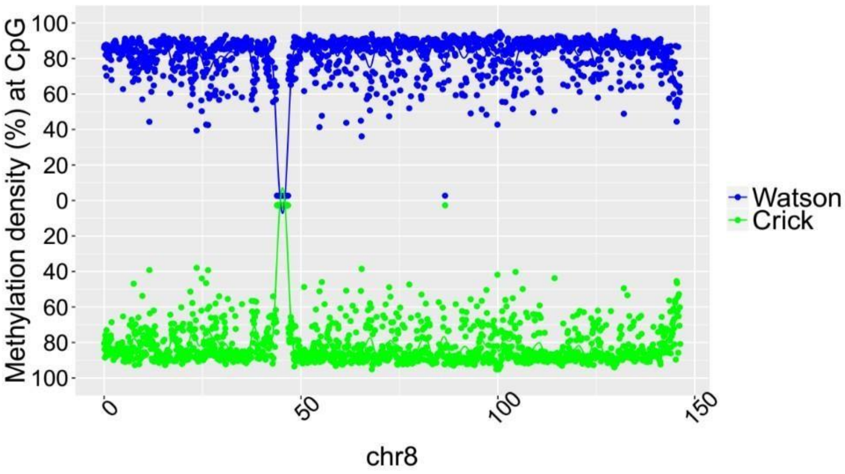
SRR921702 (64 year old female)



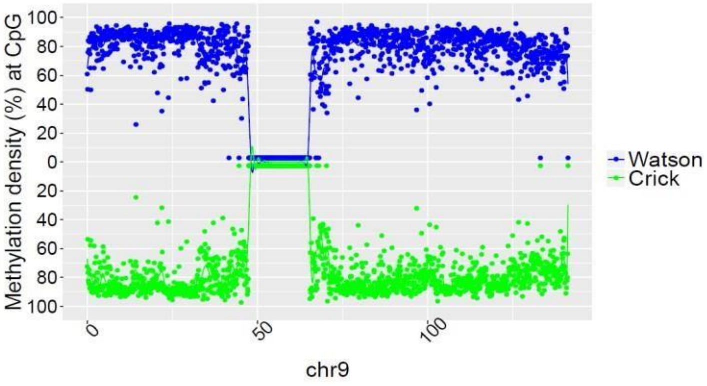
SRR949195 (81 year old female)



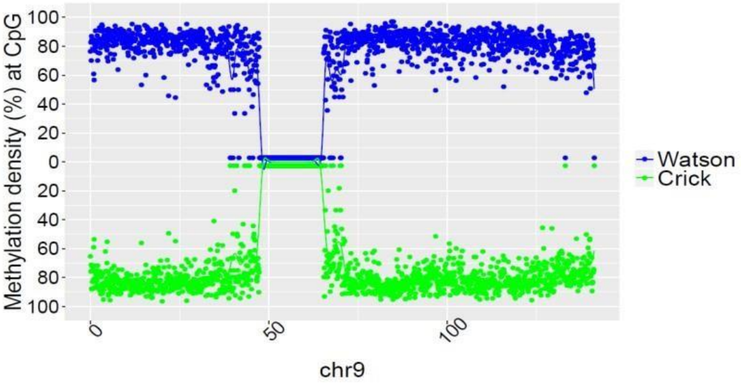
SRR949197 (82 year old female)



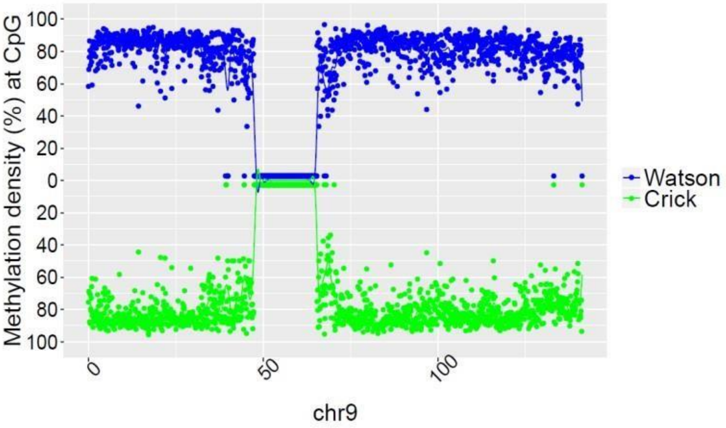
SRR901381 (20 weeks old male)



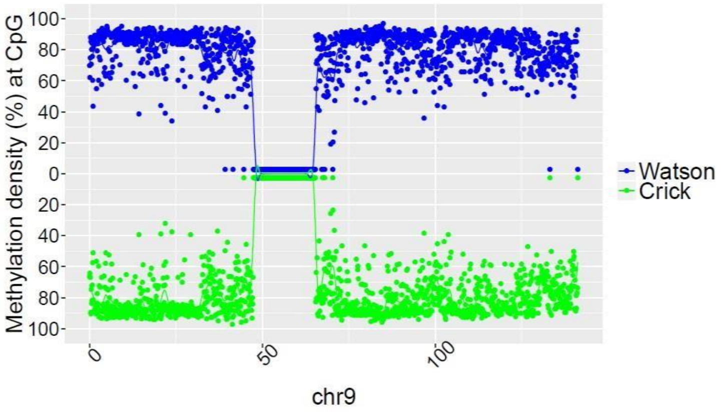
SRR47900 (31 year old male)



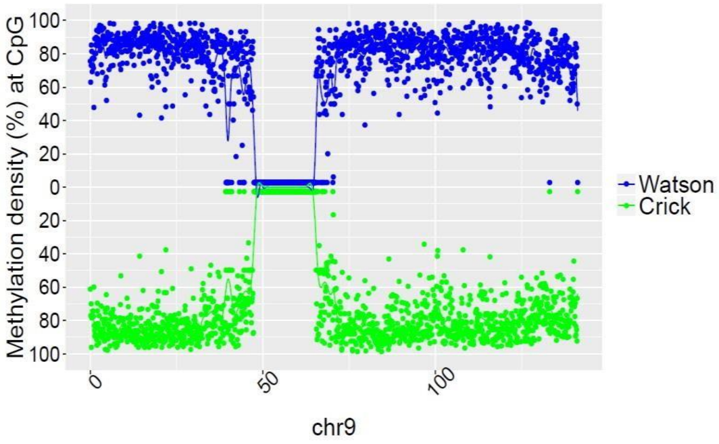
SRR 478994 (47 year old male)



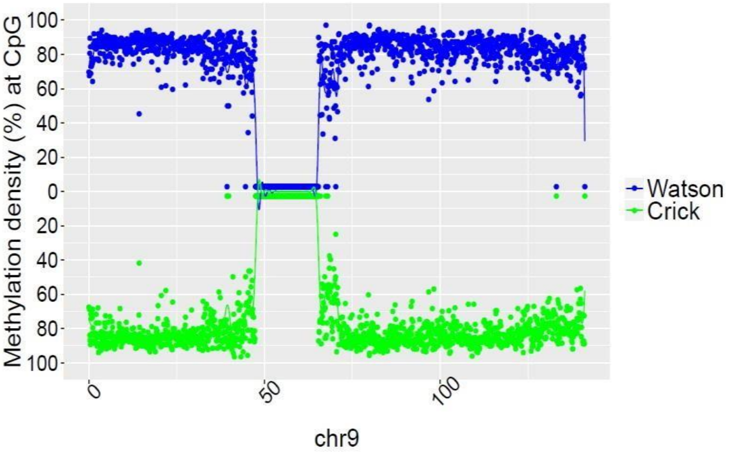
SRR478991 (48 year old male)



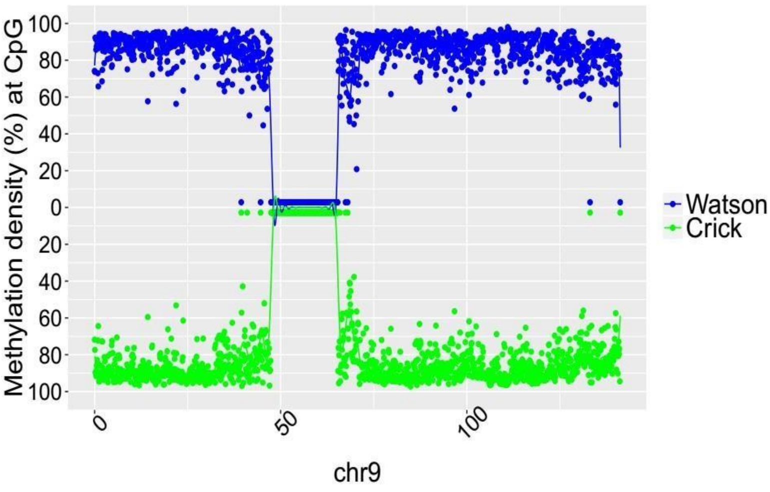
SRR921729 (55 year old male)



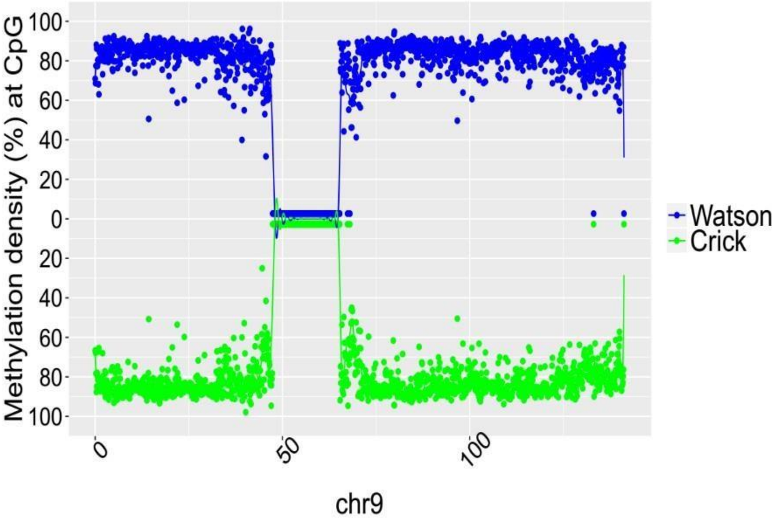
SRR921749 (55 year old male)



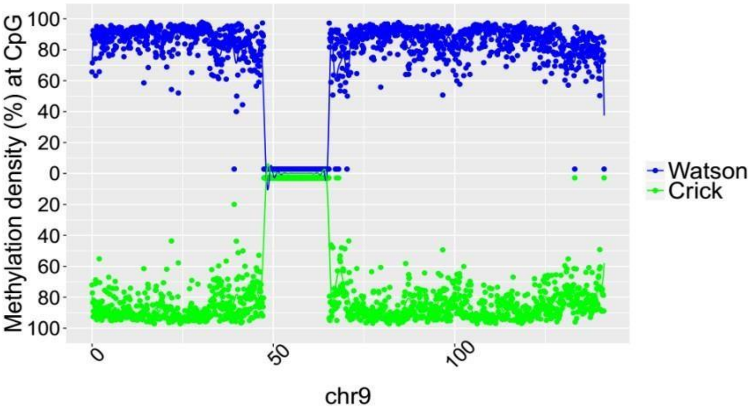
SRR921735 (55 year old male)



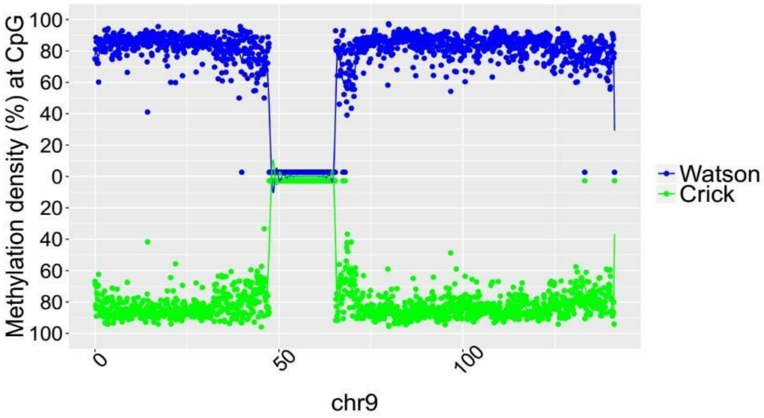
SRR847424 (42 year old female)



SRR921706 (53 year old female)

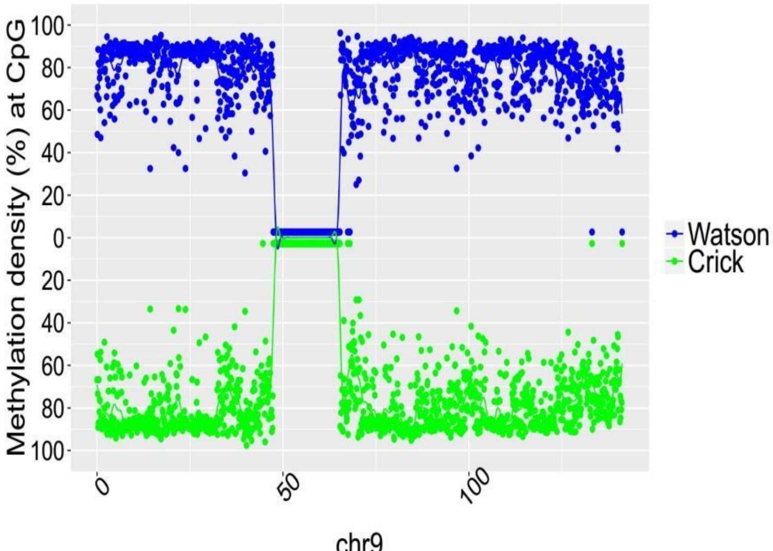


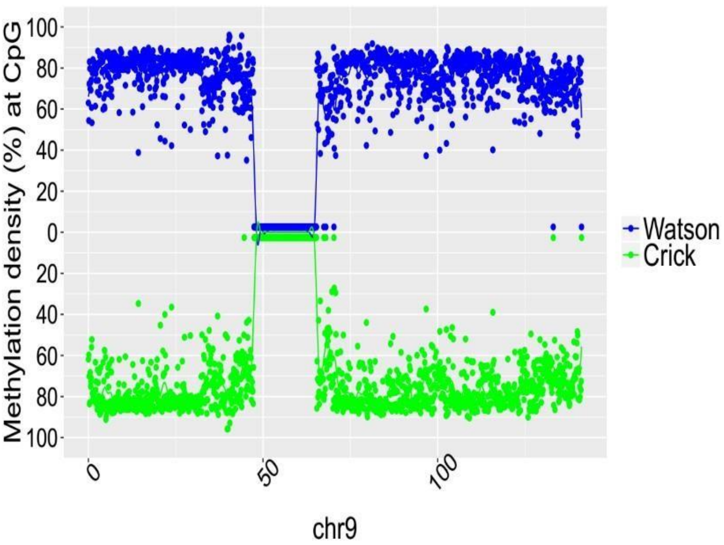
SRR921723 (53 year old female)



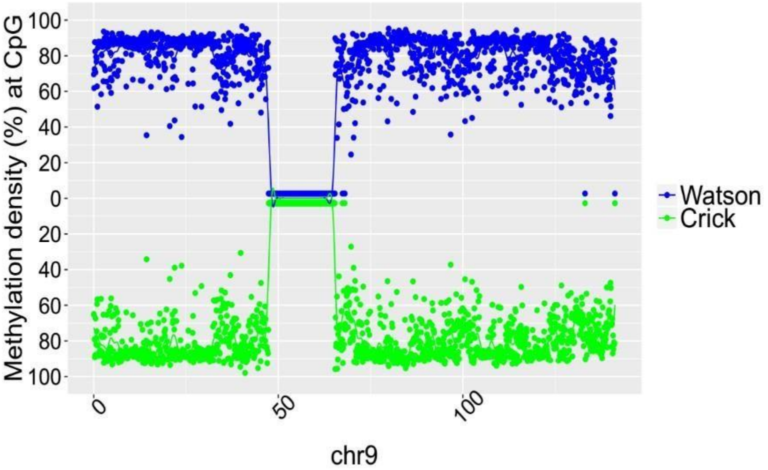
SRR921702 (64 year old female)

SRR949195 (81 year old female)

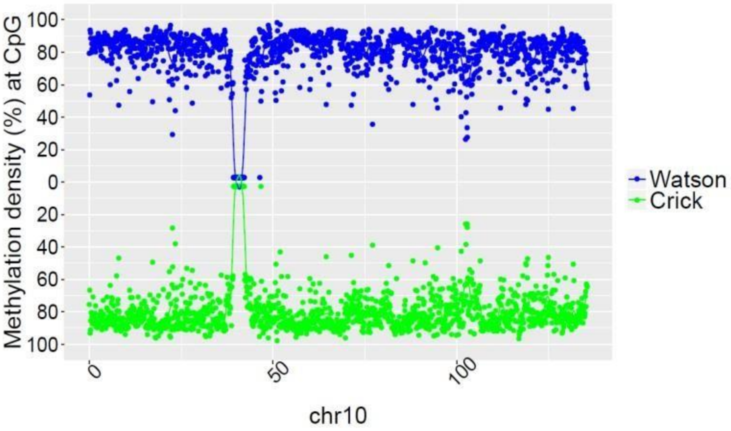



Methylation density (%) at CpG
100
80
60
40
20
0
20
40
60
80
100
0
50
100
chr9
Watson
Crick

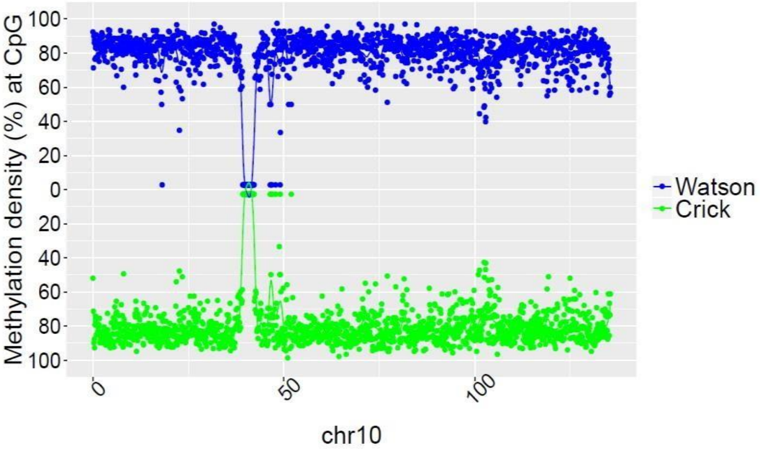
SRR949197 (82 year old female)



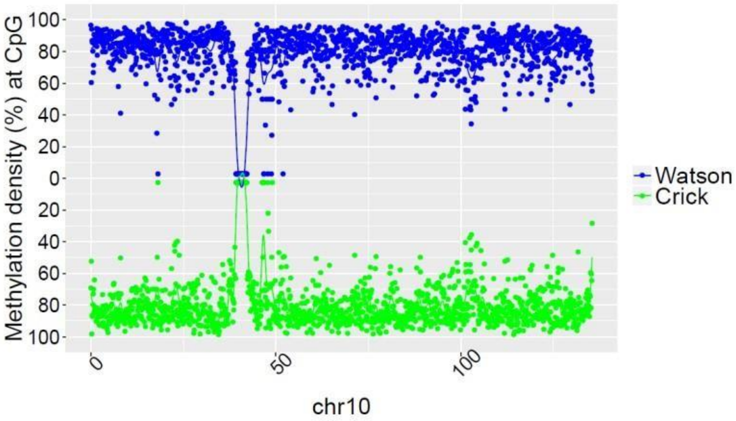
SRR901381 (20 weeks old male)



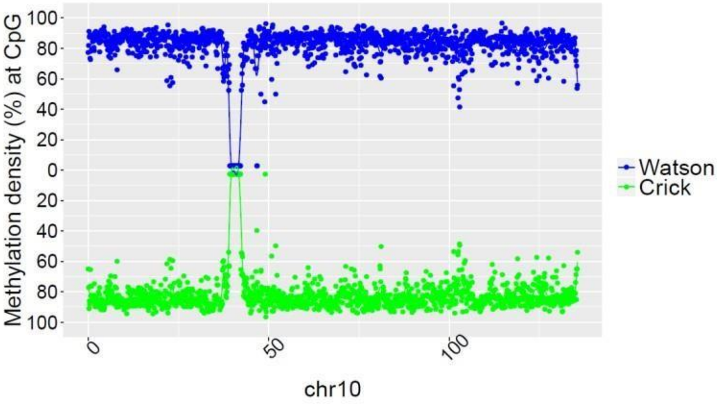
SRR47900 (31 year old male)



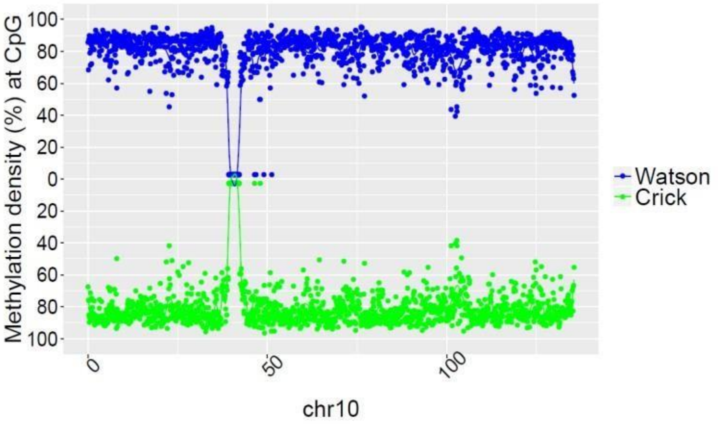
SRR 478994 (47 year old male)



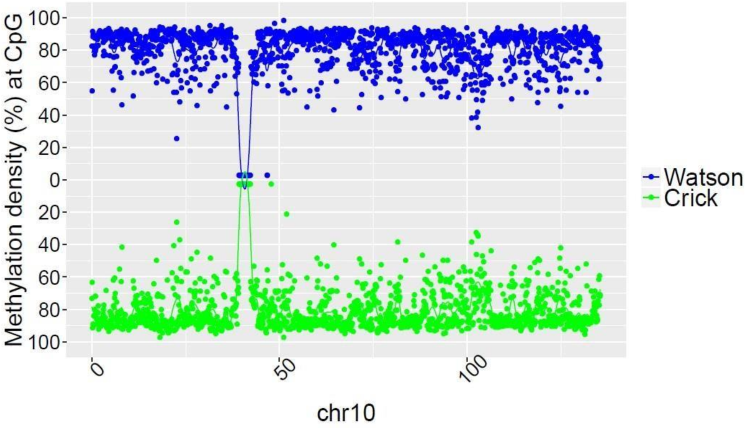
SRR478991 (48 year old male)



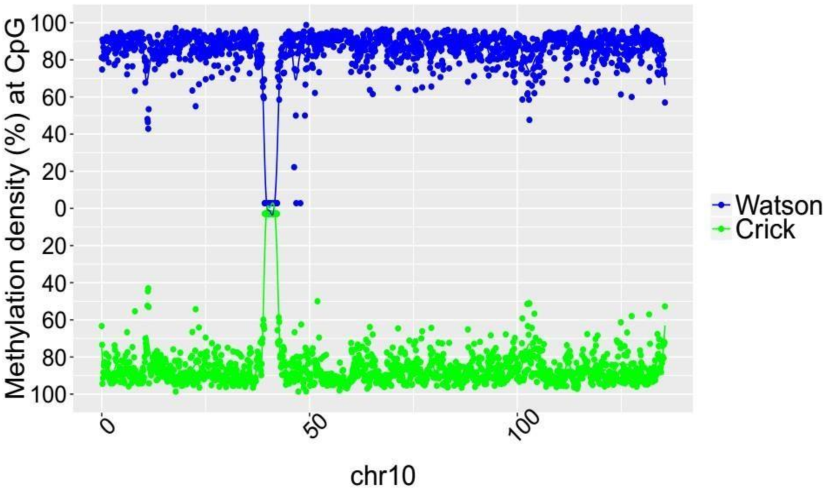
SRR921749 (55 year old male)



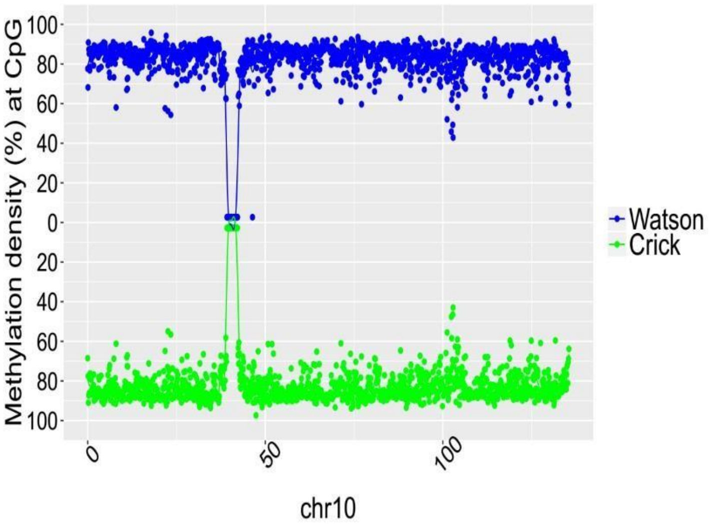
SRR921729 (55 year old male)



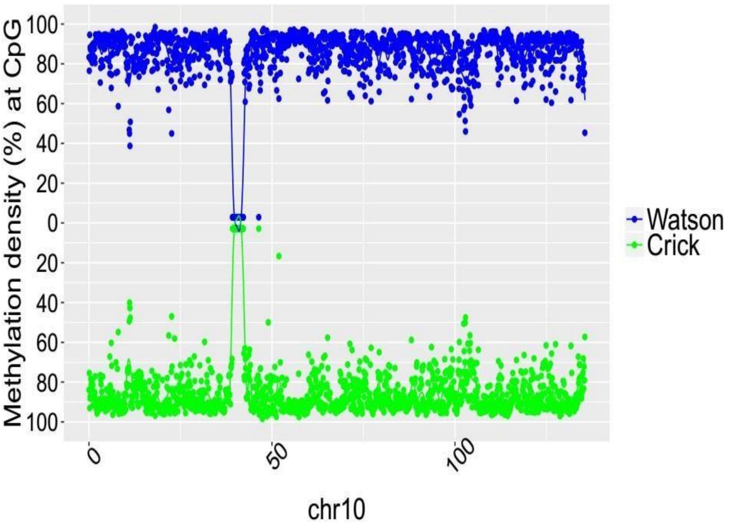
SRR921735 (55 year old male)



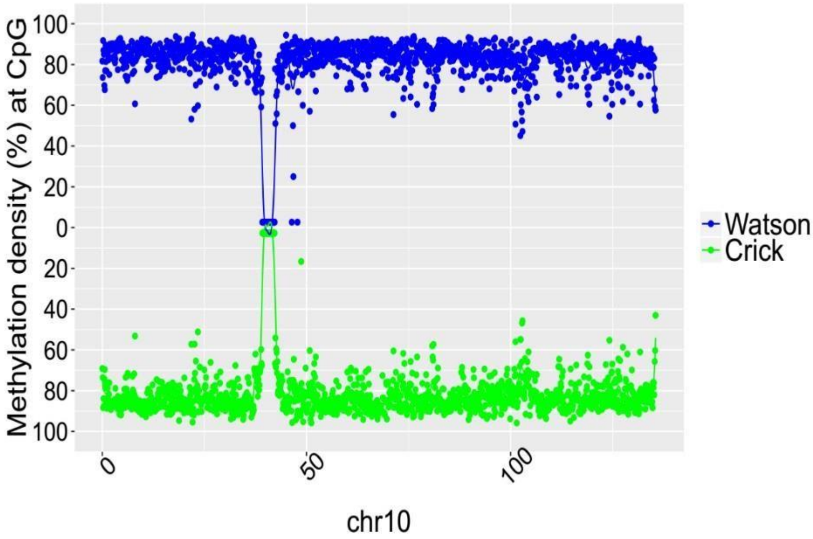
SRR847424 (42 year old female)



SRR921706 (53 year old female)

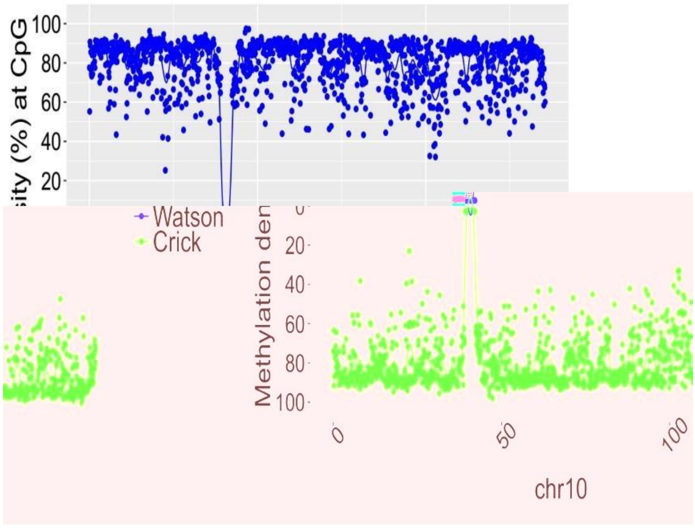


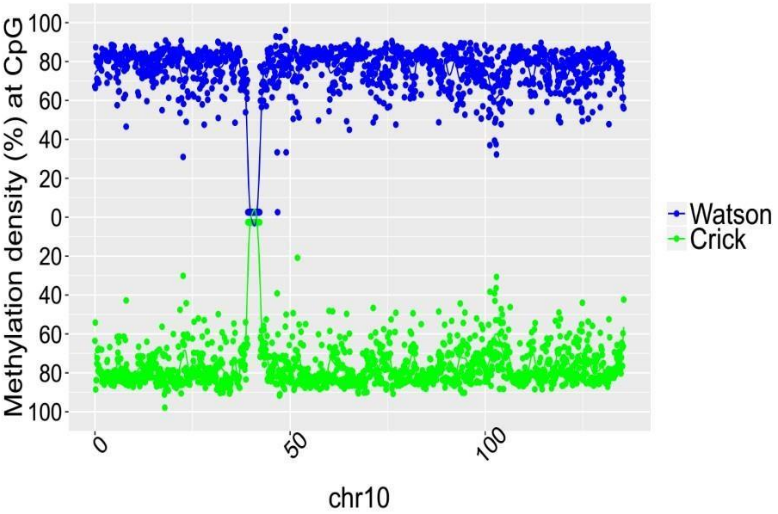
SRR921723 (53 year old female)



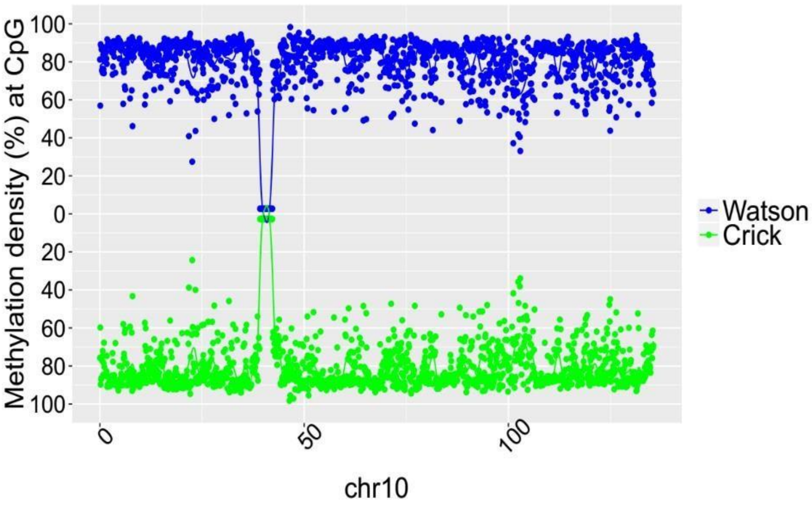
SRR921702 (64 year old female)

SRR949195 (81 year old female)

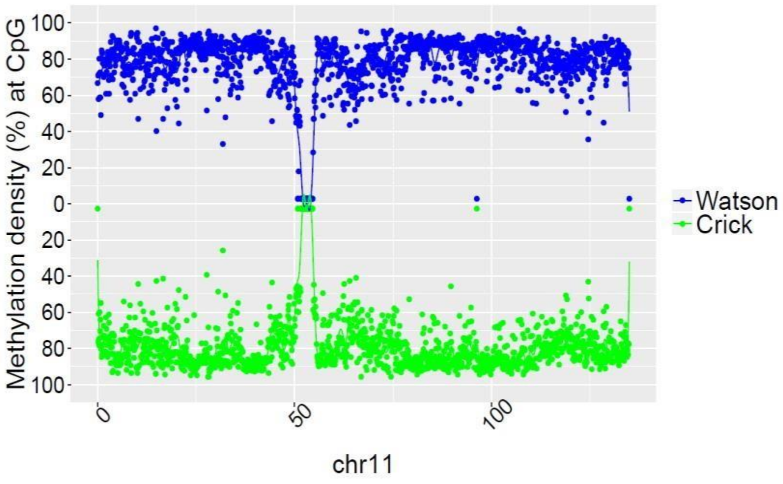



Methylation density (%) at CpG
100
80
60
40
20
0
20
40
60
80
100
0
50
100
chr10
Watson
Crick

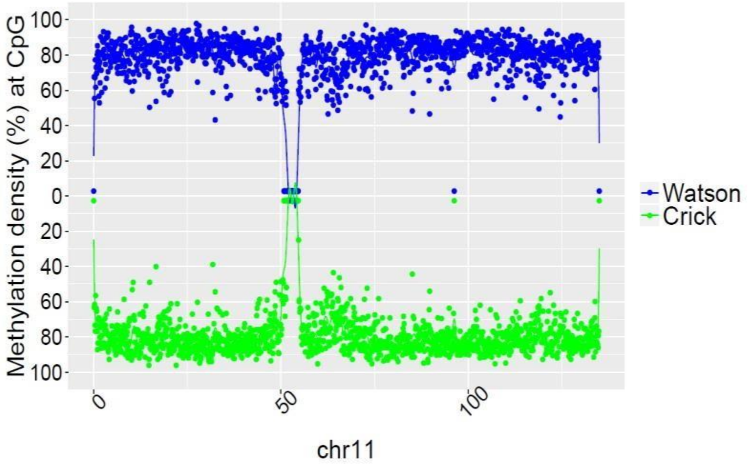
SRR949197 (82 year old female)



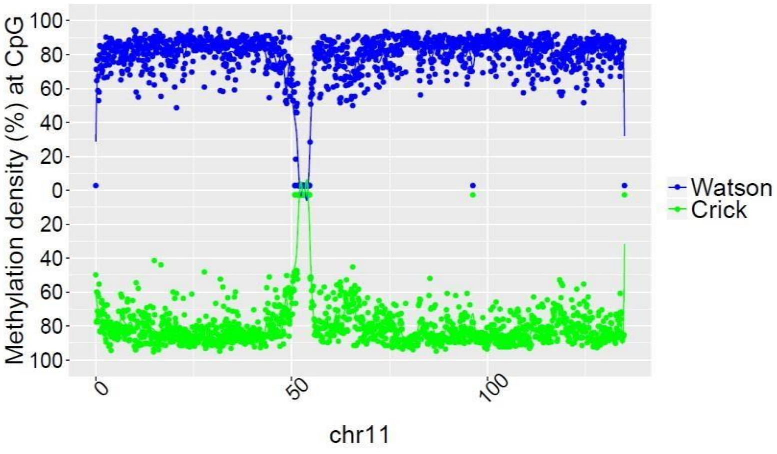
SRR901381 (20 weeks old male)



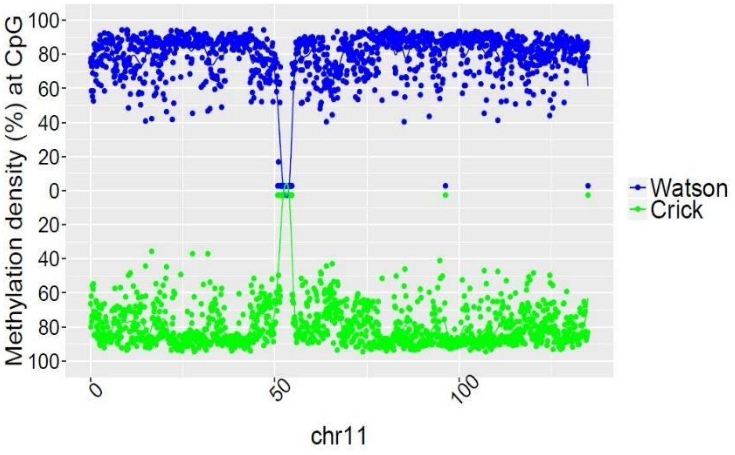
SRR47900 (31 year old male)



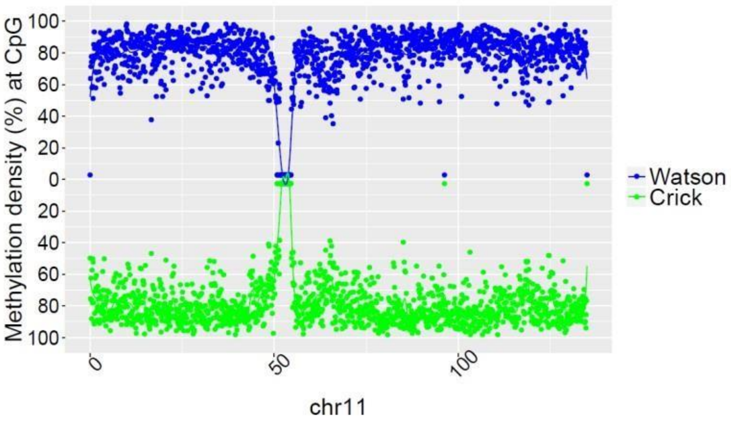
SRR 478994 (47 year old male)



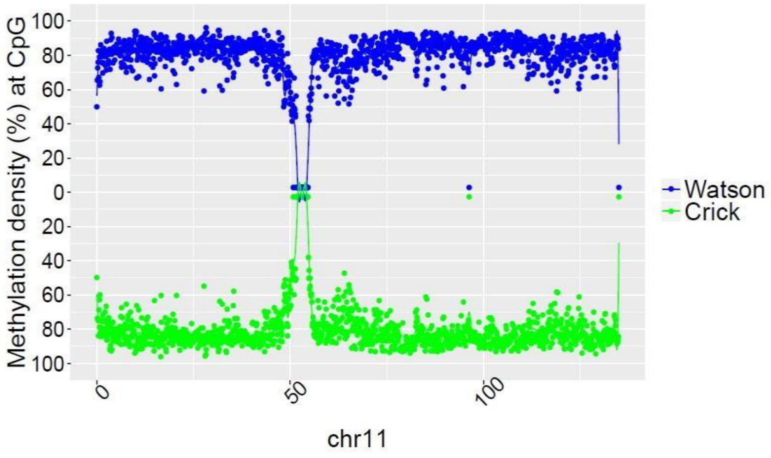
SRR478991 (48 year old male)



SRR921729 (55 year old male)

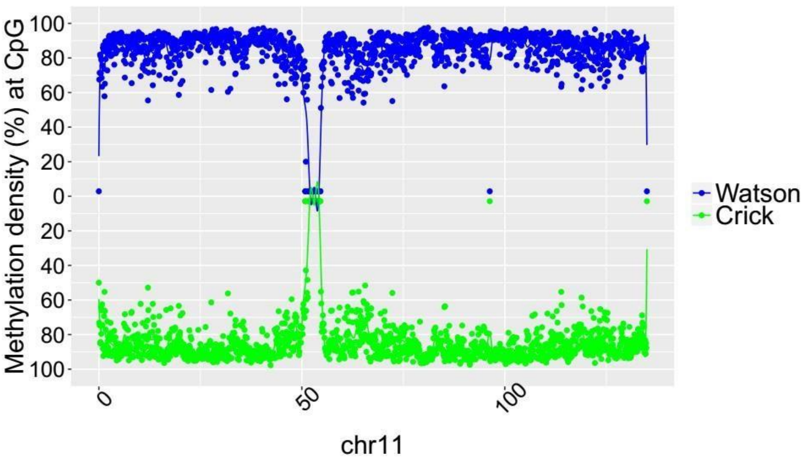


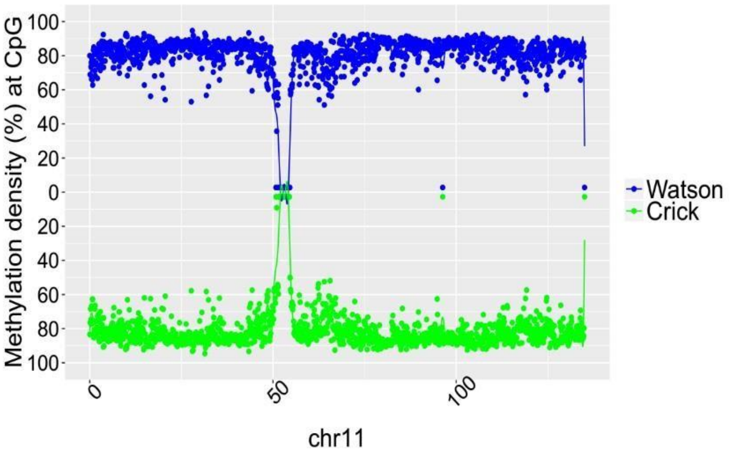
SRR921749 (55 year old male)



SRR847424 (42 year old female)

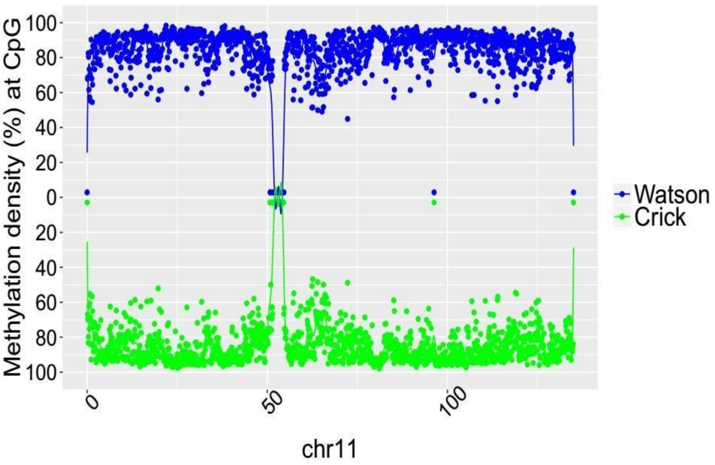
SRR921735 (55 year old male)

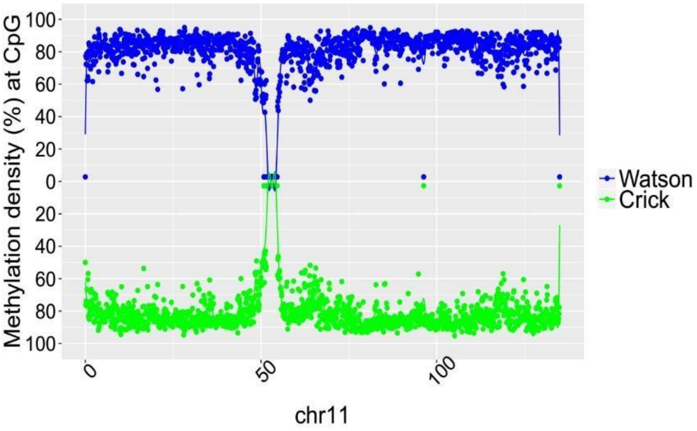




SRR921706 (53 year old female)

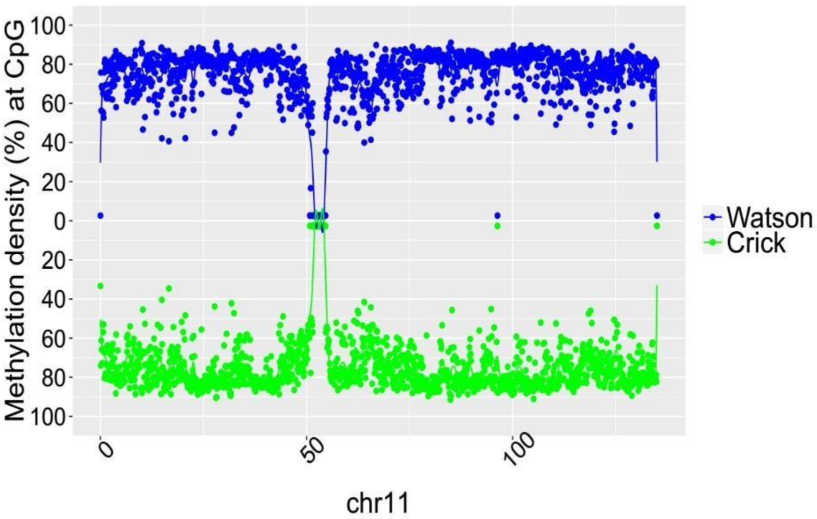
SRR921723 (53 year old female)

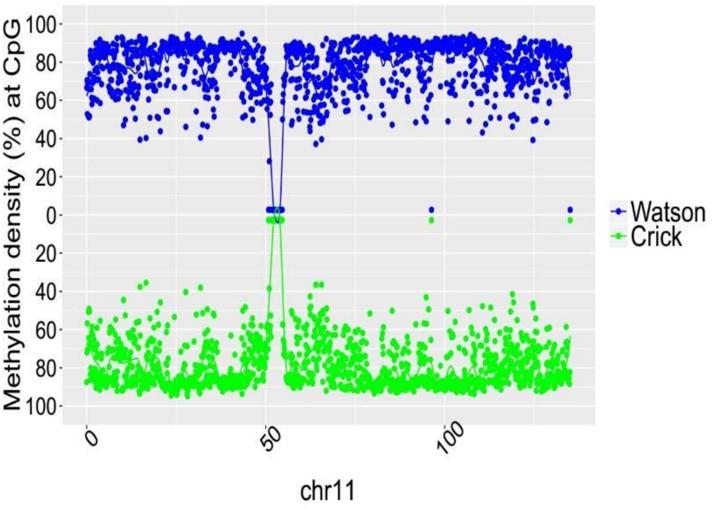




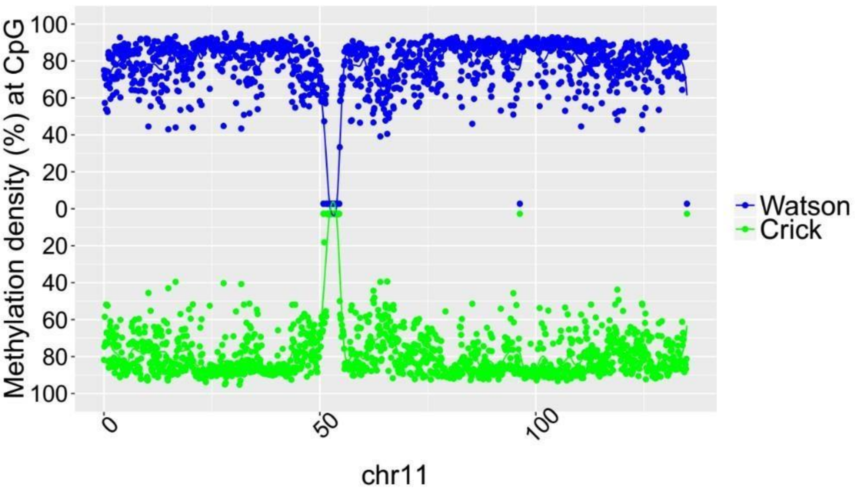
SRR921702 (64 year old female)

SRR949195 (81 year old female)

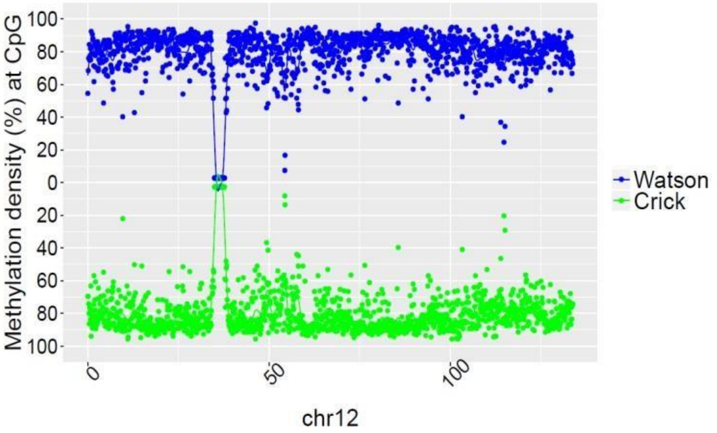




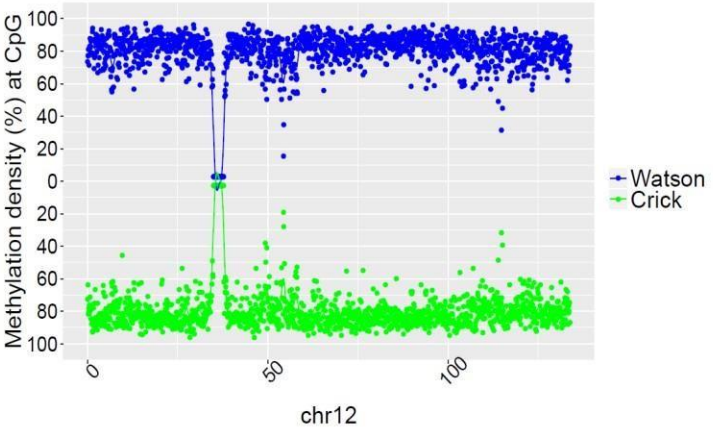
SRR949197 (82 year old female)



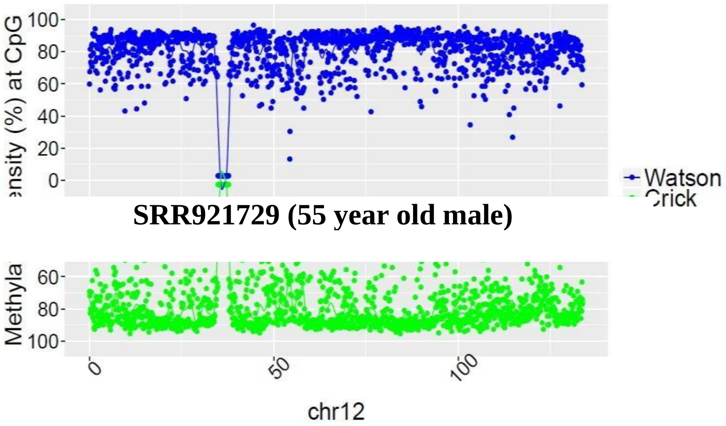
SRR901381 (20 weeks old male)



SRR47900 (31 year old male)

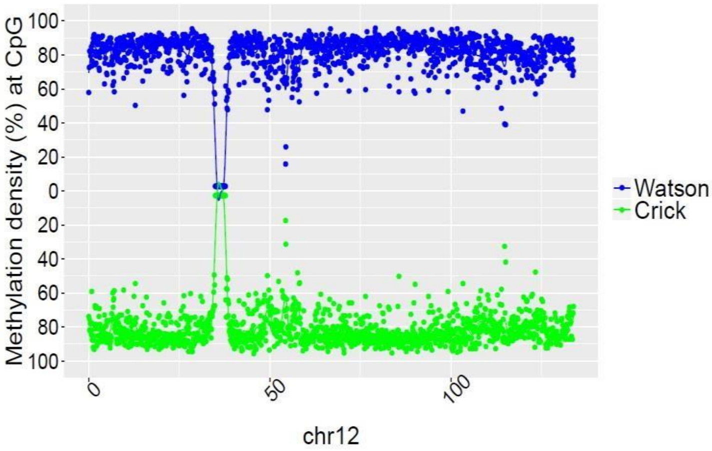


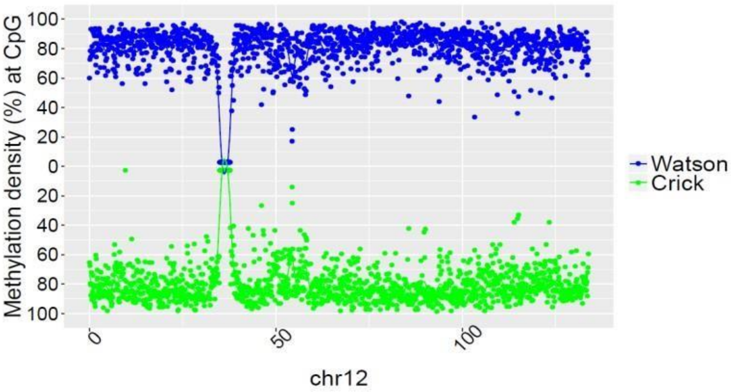
SRR 478994 (47 year old male)



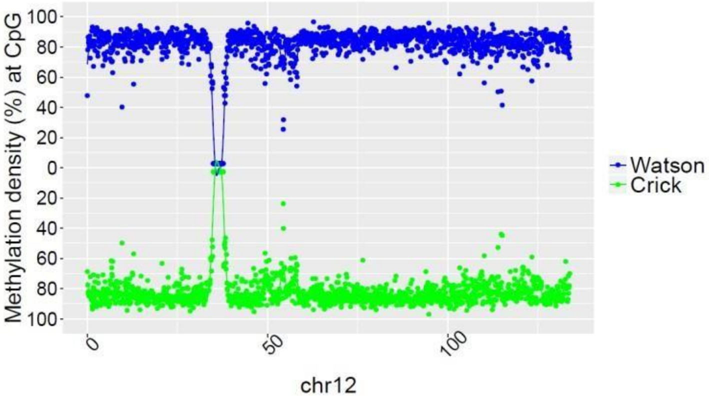
SRR921729 (55 year old male)

SRR478991 (48 year old male)

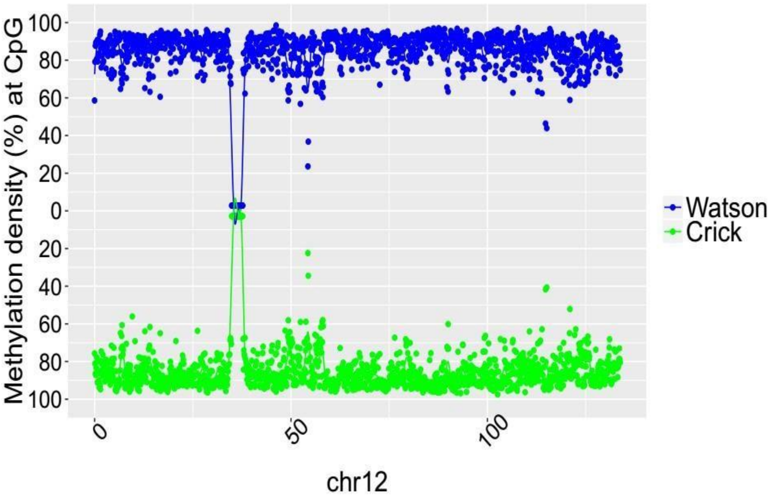




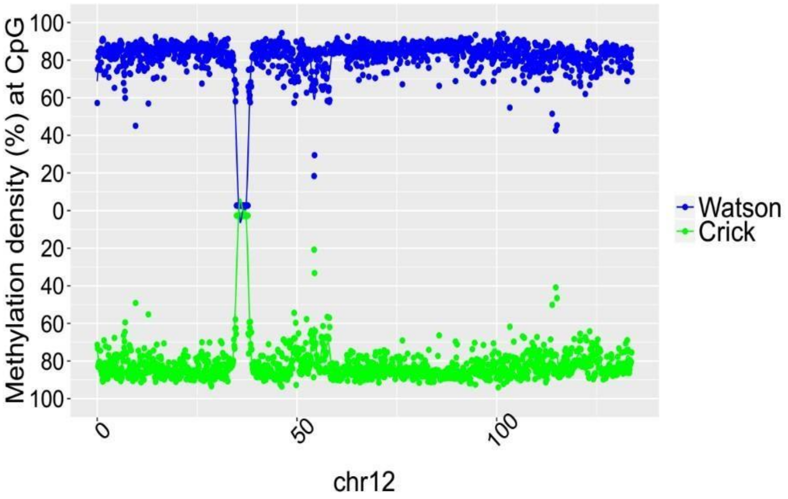
SRR921749 (55 year old male)



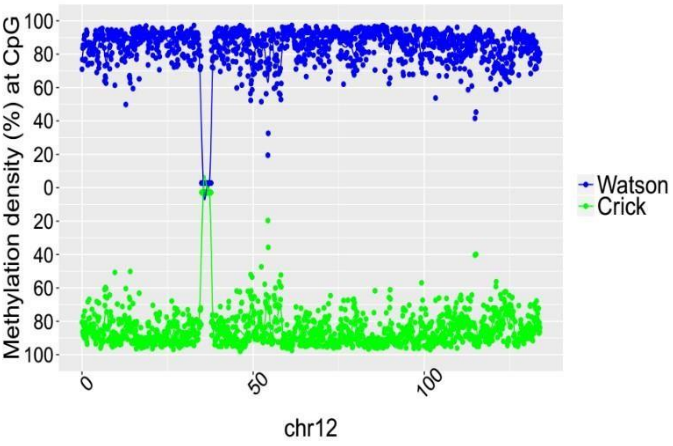
SRR921735 (55 year old male)



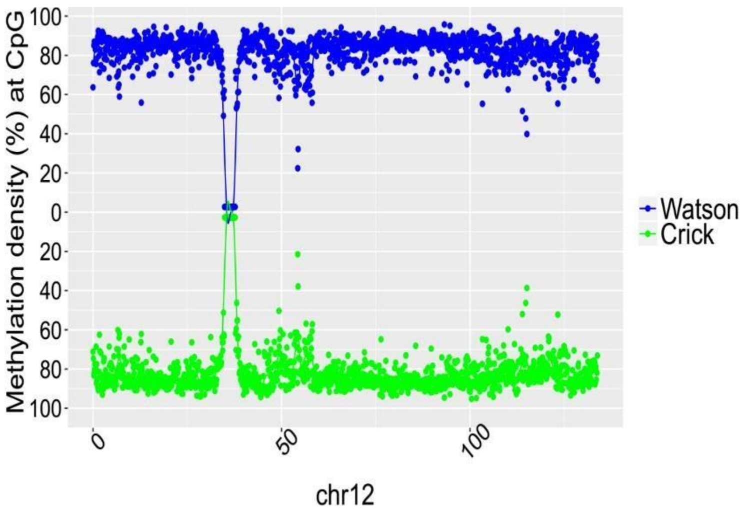
SRR847424 (42 year old female)



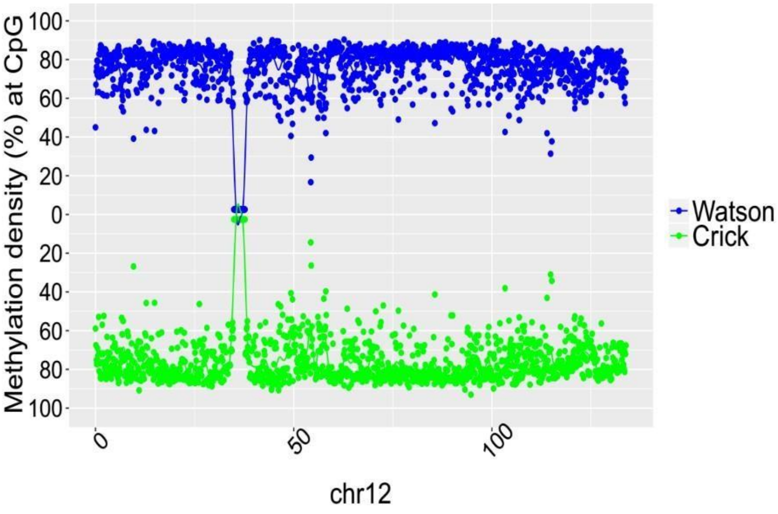
SRR921706 (53 year old female)



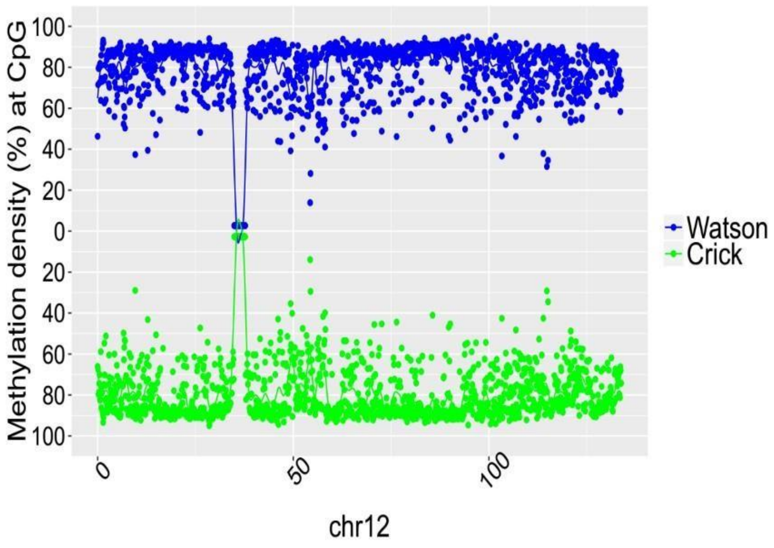
SRR921723 (53 year old female)



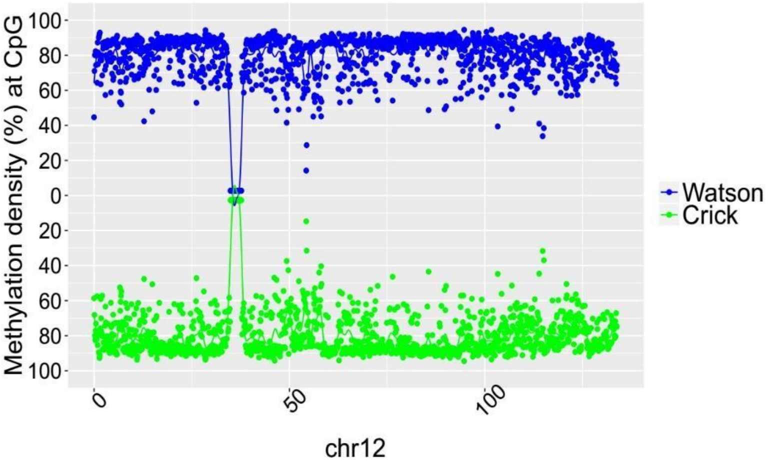
SRR921702 (64 year old female)



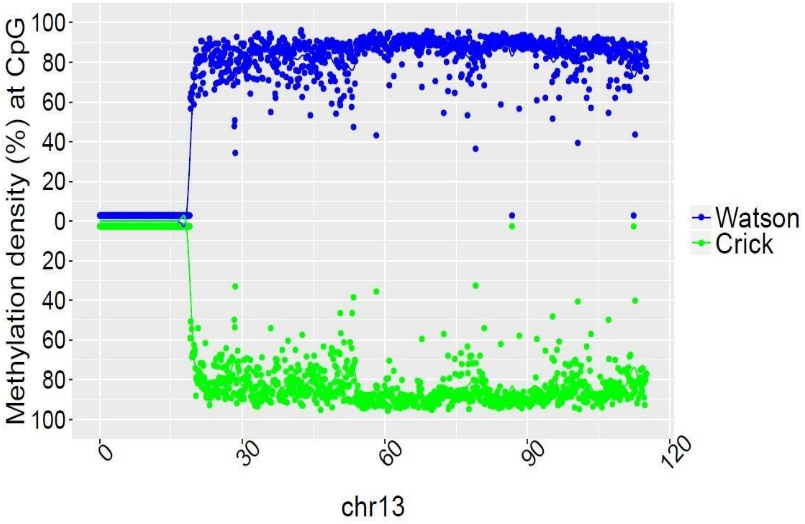
SRR949195 (81 year old female)



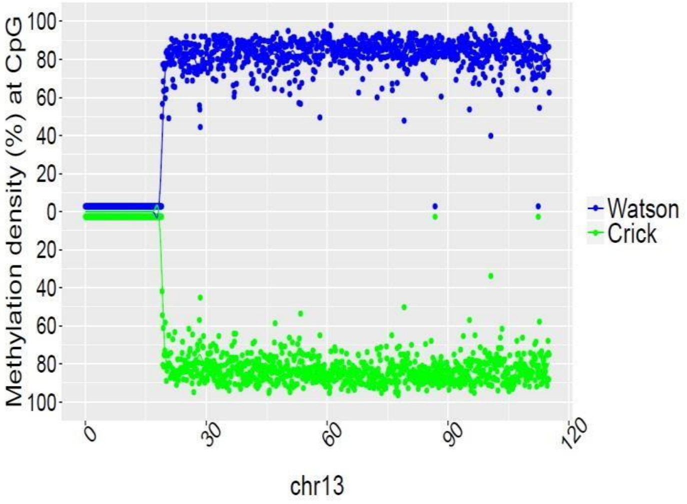
SRR949197 (82 year old female)



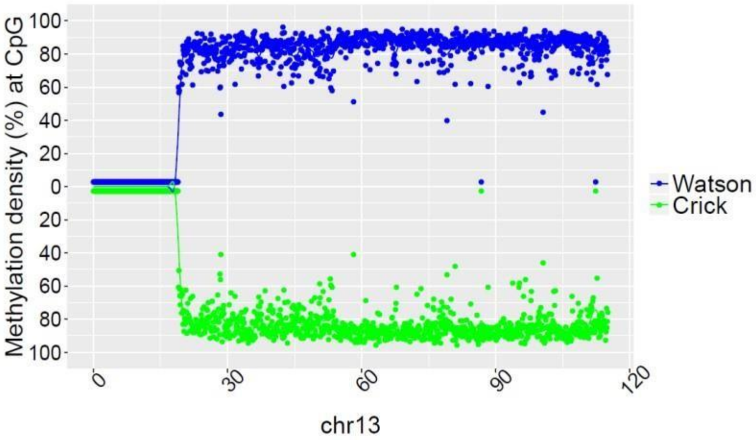
SRR901381 (20 weeks old male)



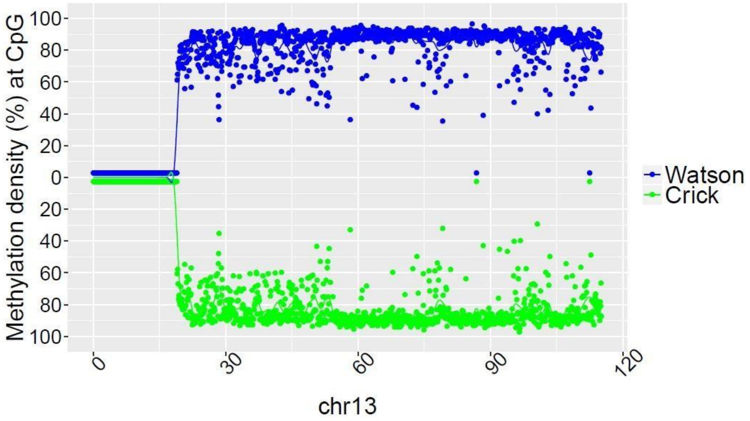
SRR47900 (31 year old male)



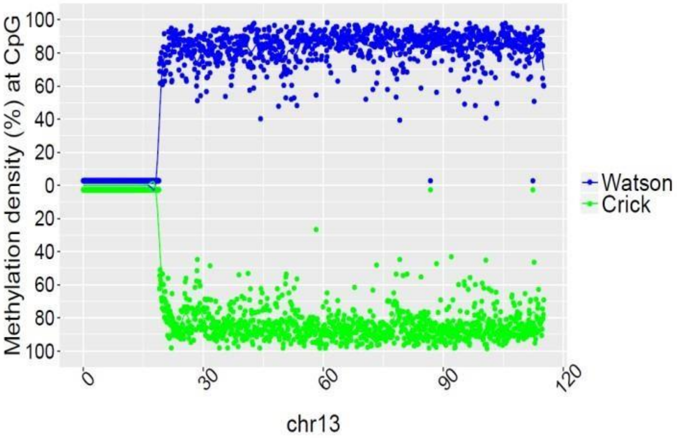
SRR 478994 (47 year old male)



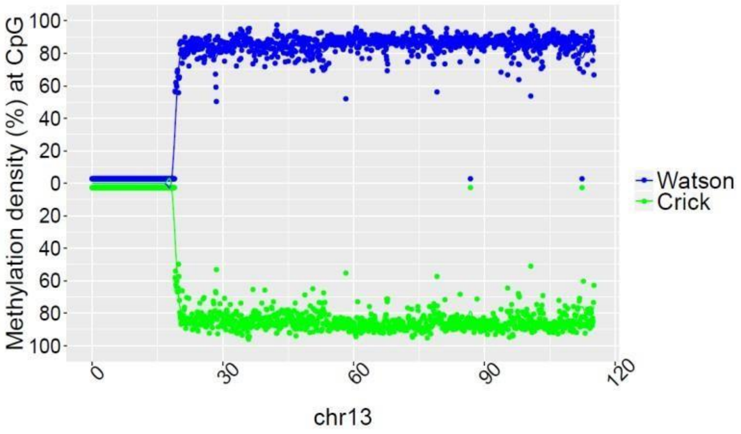
SRR478991 (48 year old male)



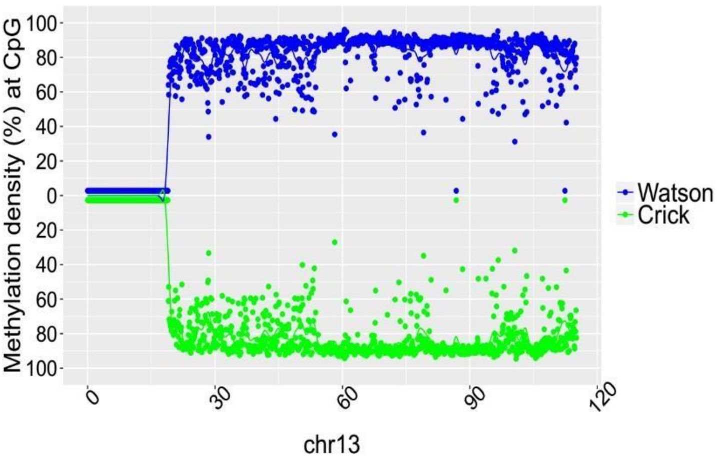
SRR921729 (55 year old male)



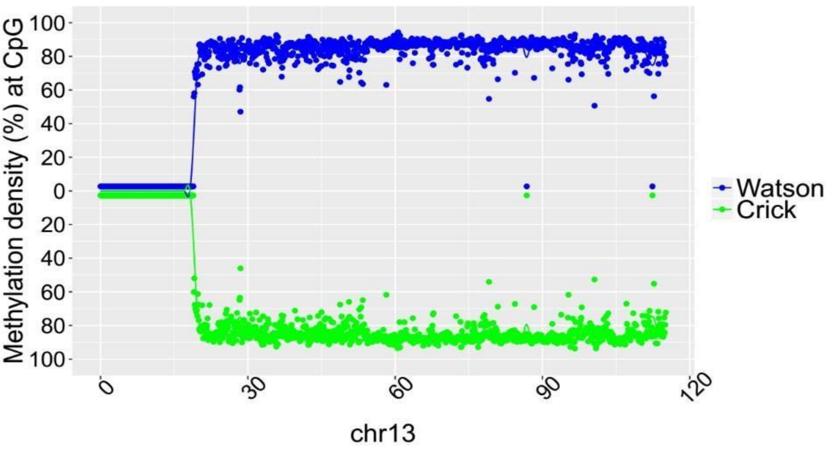
SRR921749 (55 year old male)



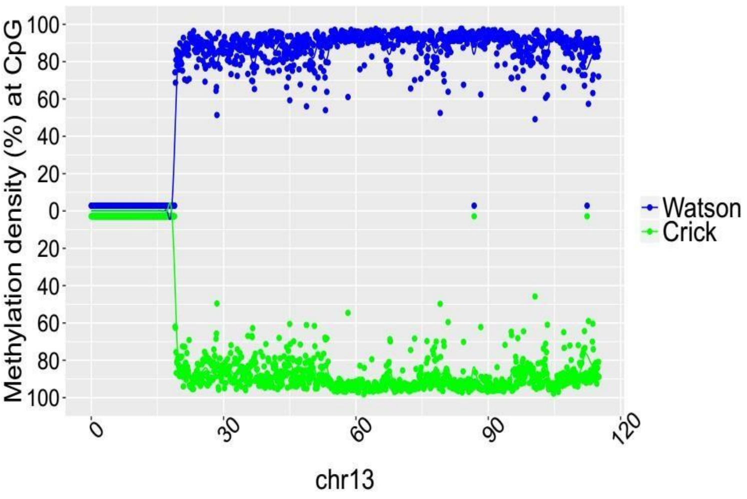
SRR921735 (55 year old male)



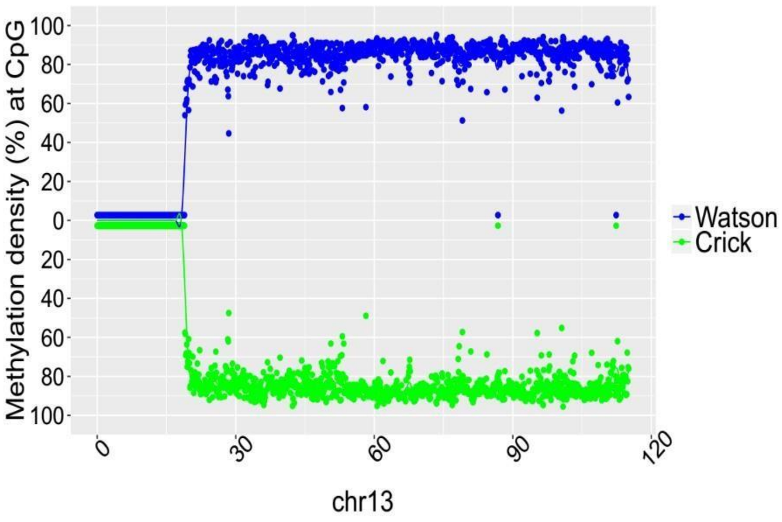
SRR847424 (42 year old female)



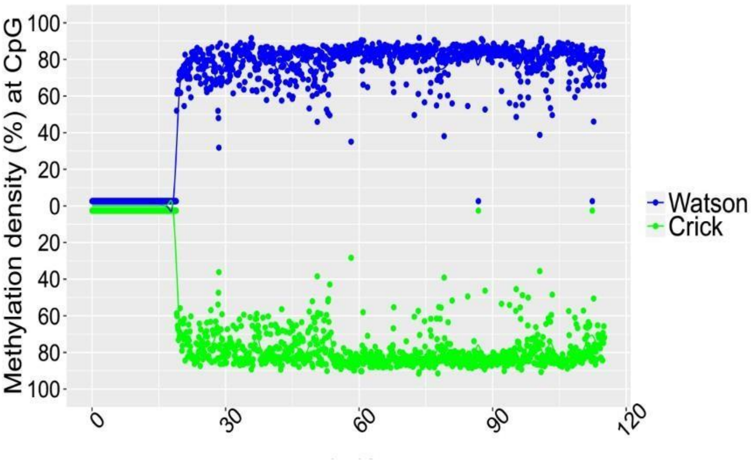
SRR921706 (53 year old female)



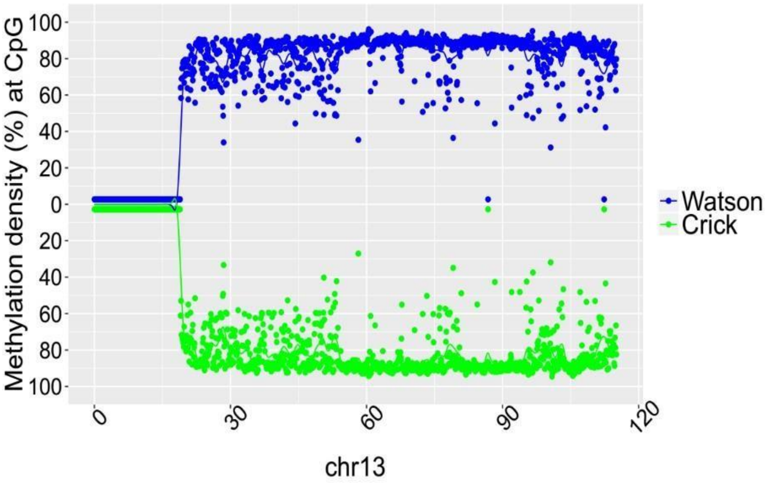
SRR921723 (53 year old female)



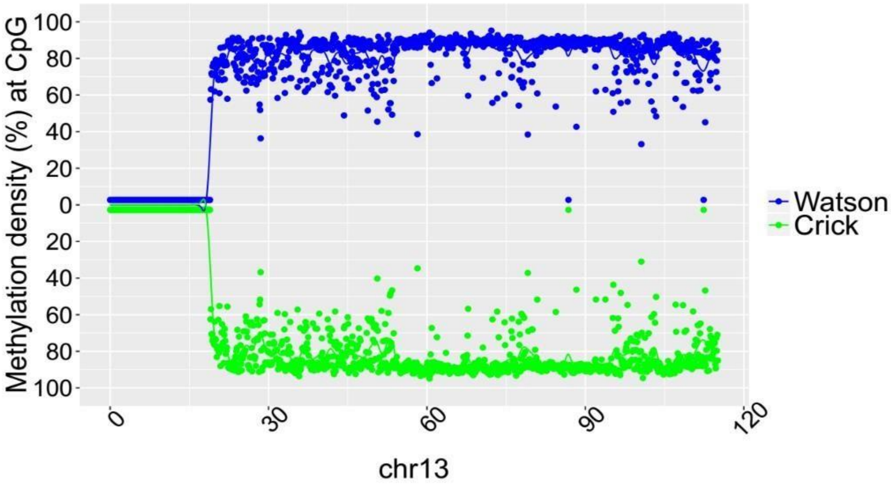
SRR921702 (64 year old female)



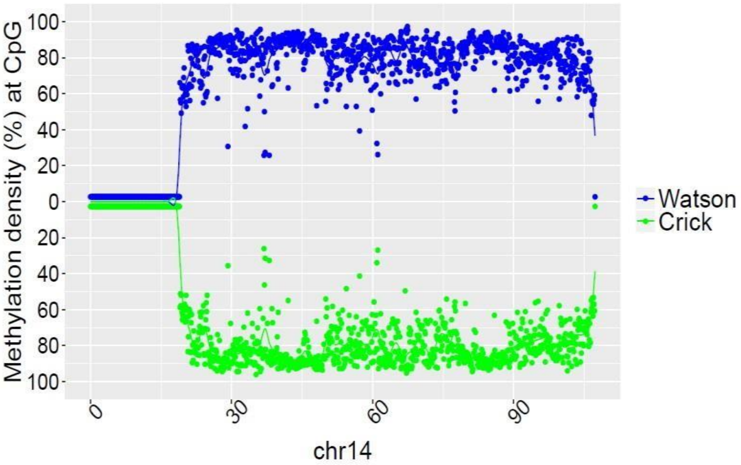
SRR949195 (81 year old female)



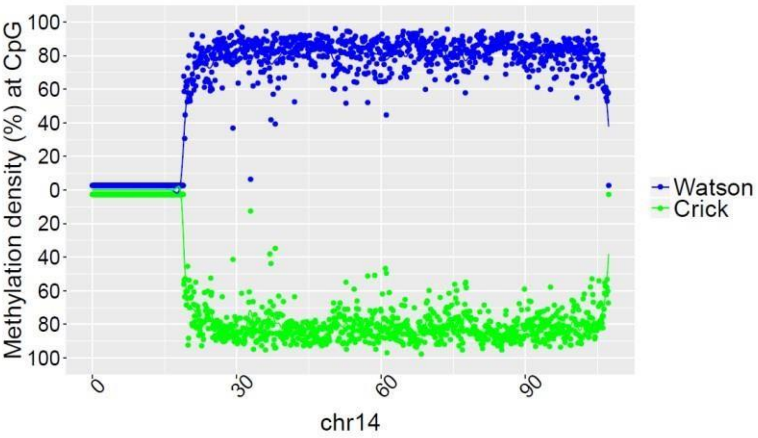
SRR949197 (82 year old female)



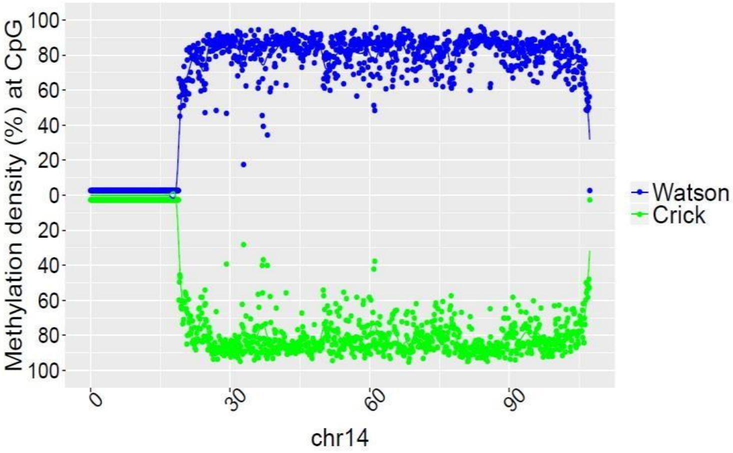
SRR901381 (20 weeks old male)



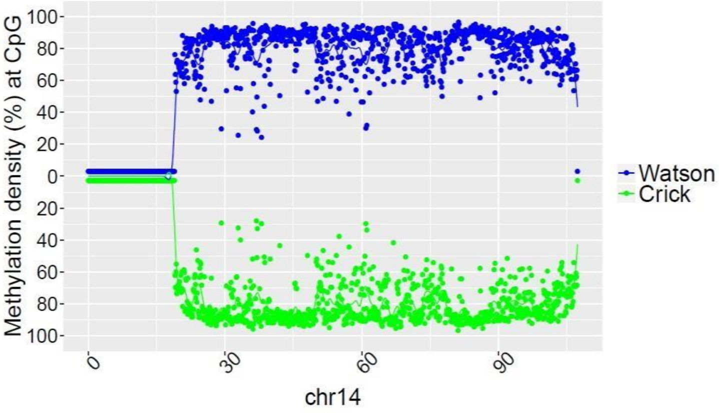
SRR47900 (31 year old male)



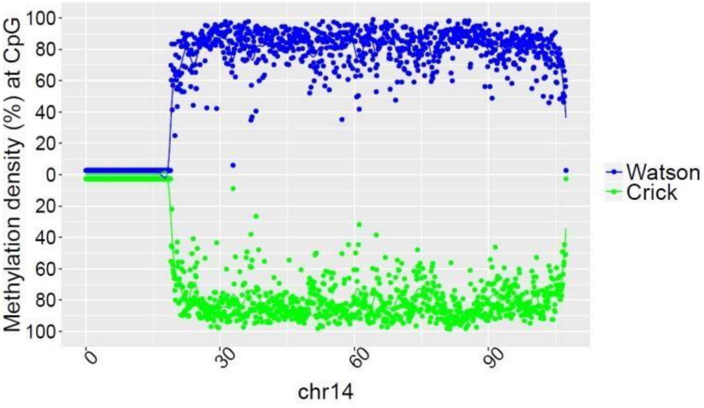
SRR 478994 (47 year old male)



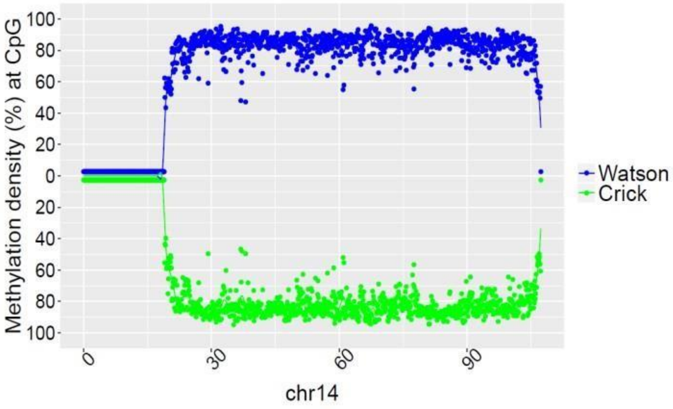
SRR478991 (48 year old male)



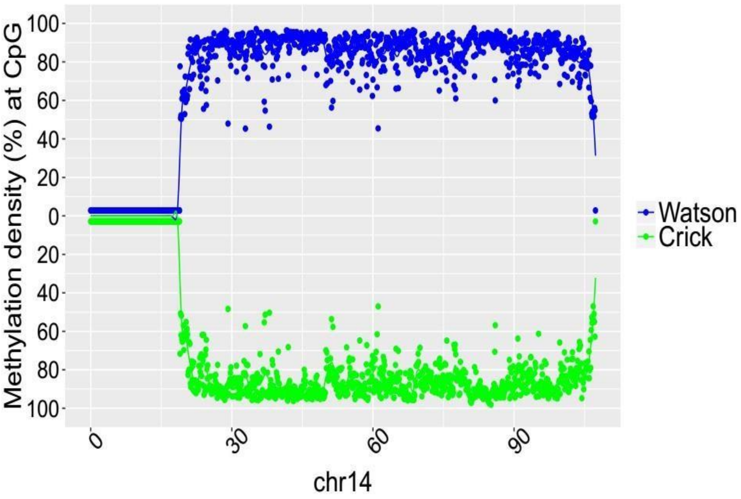
SRR921729 (55 year old male)



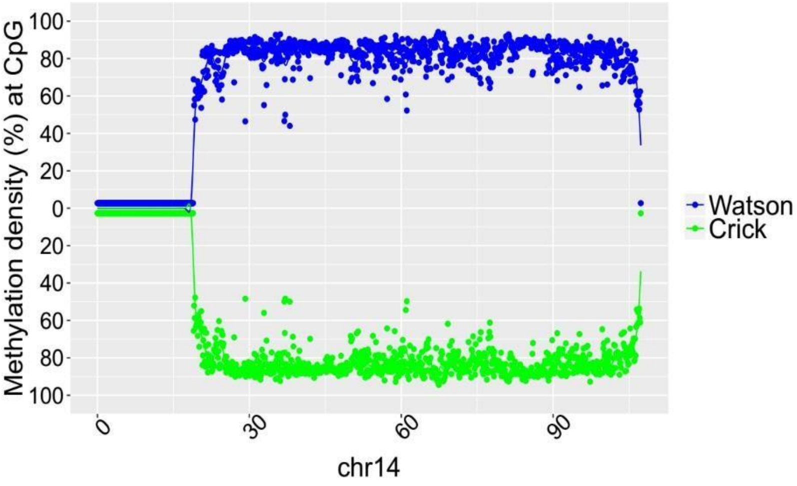
SRR921749 (55 year old male)



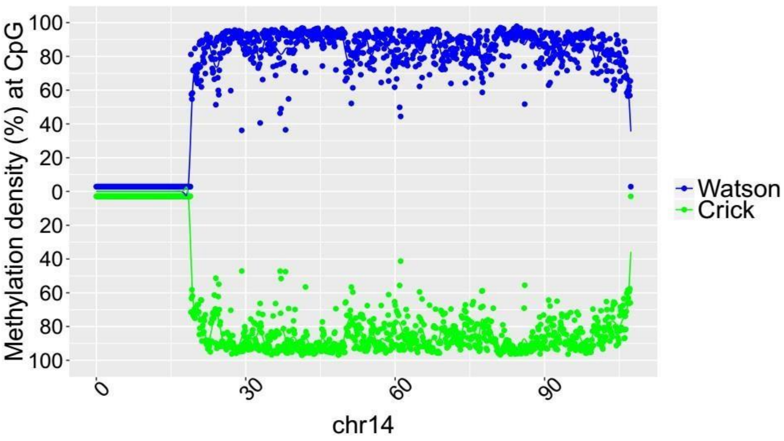
SRR921735 (55 year old male)



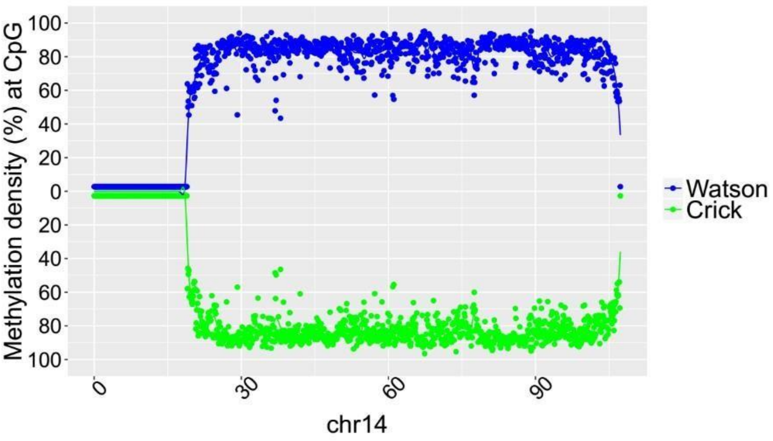
SRR847424 (42 year old female)



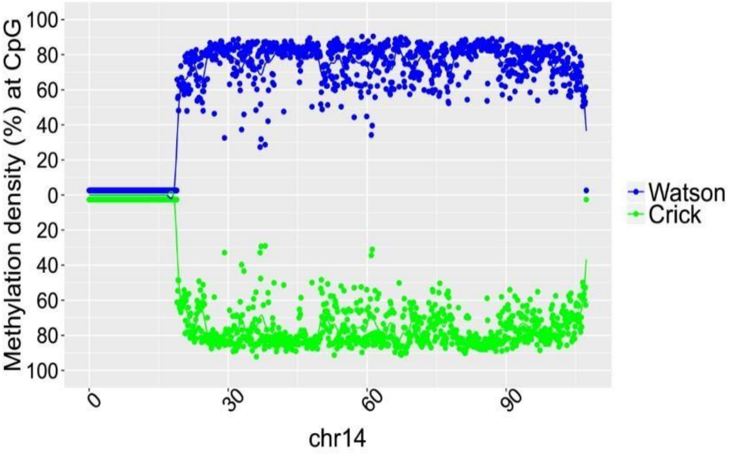
SRR921706 (53 year old female)



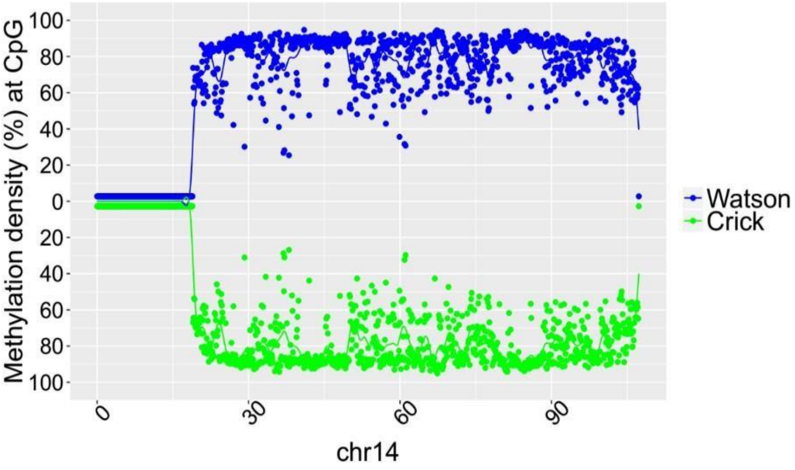
SRR921723 (53 year old female)



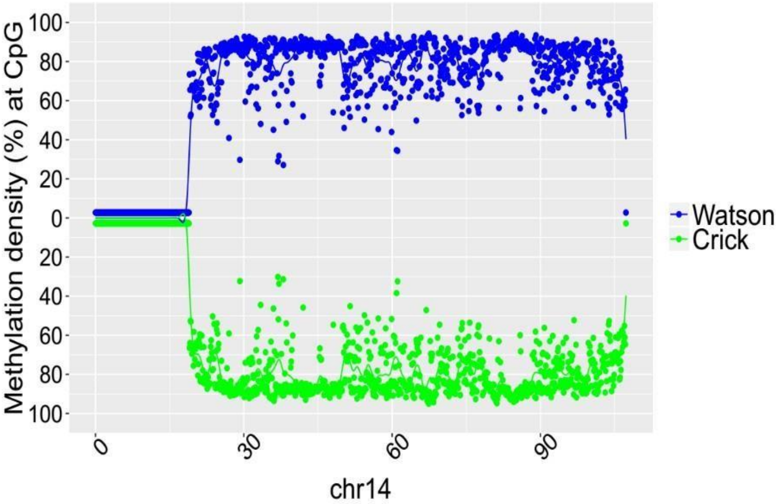
SRR921702 (64 year old female)



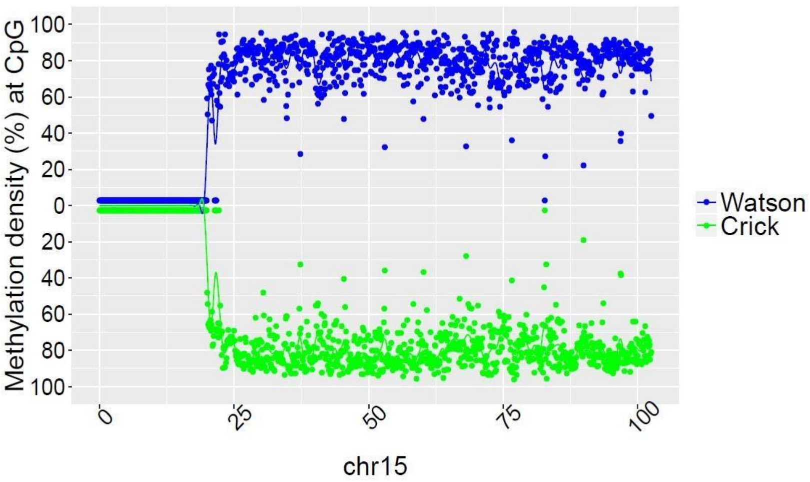
SRR949195 (81 year old female)



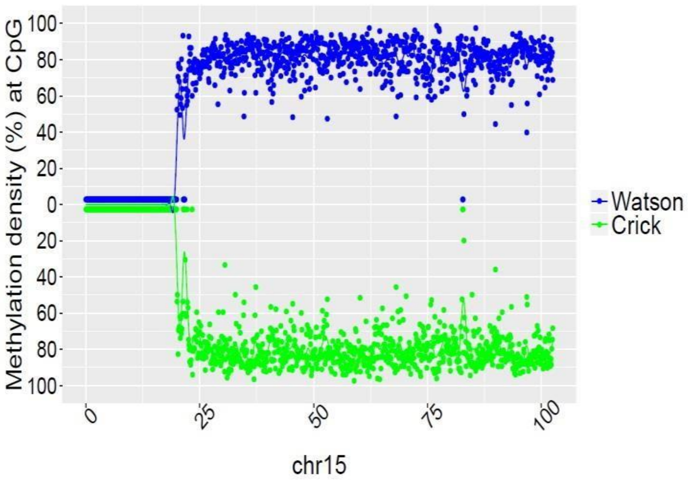
SRR949197 (82 year old female)



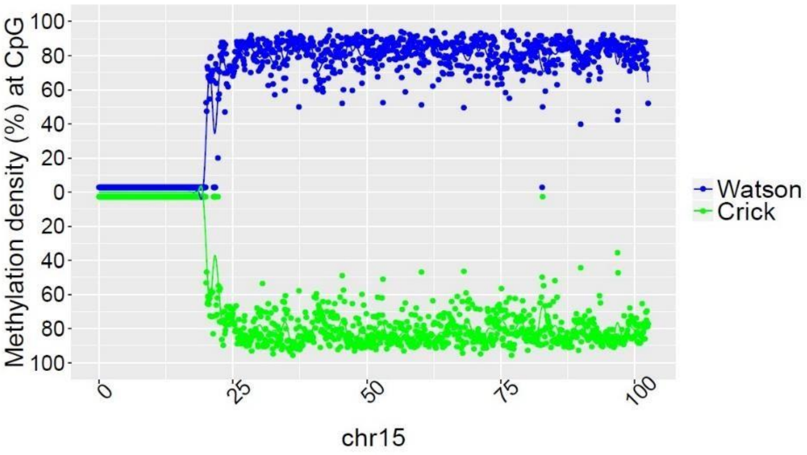
SRR901381 (20 weeks old male)



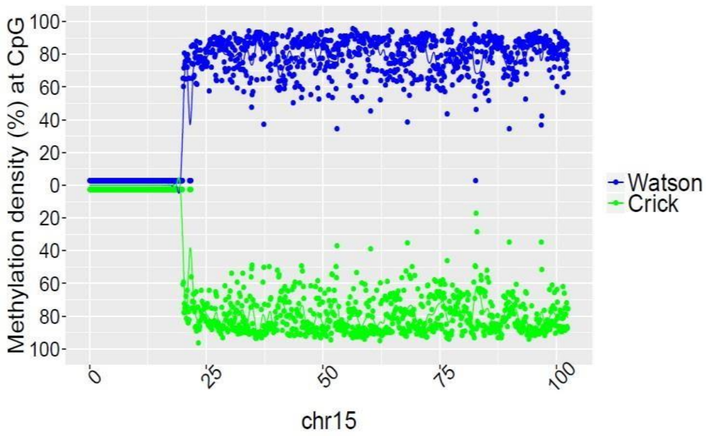
SRR47900 (31 year old male)



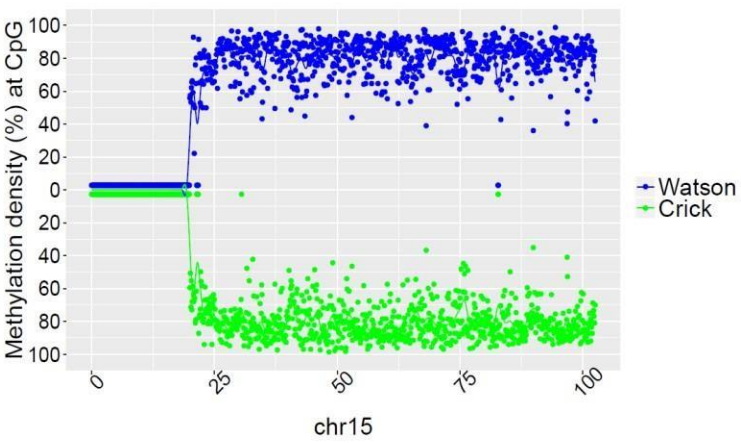
SRR 478994 (47 year old male)



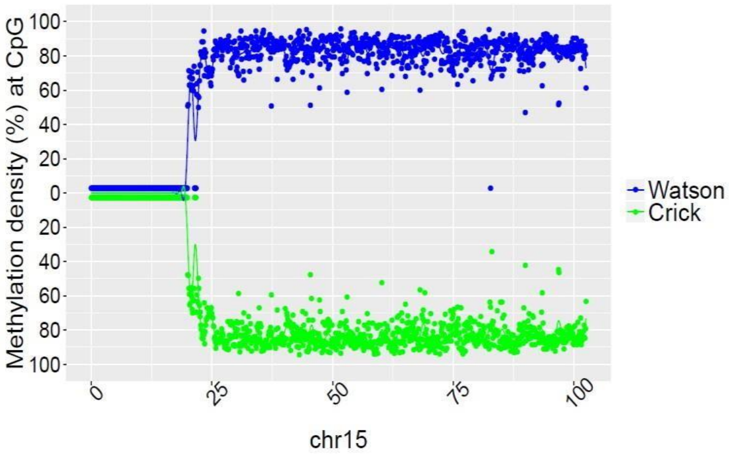
SRR478991 (48 year old male)



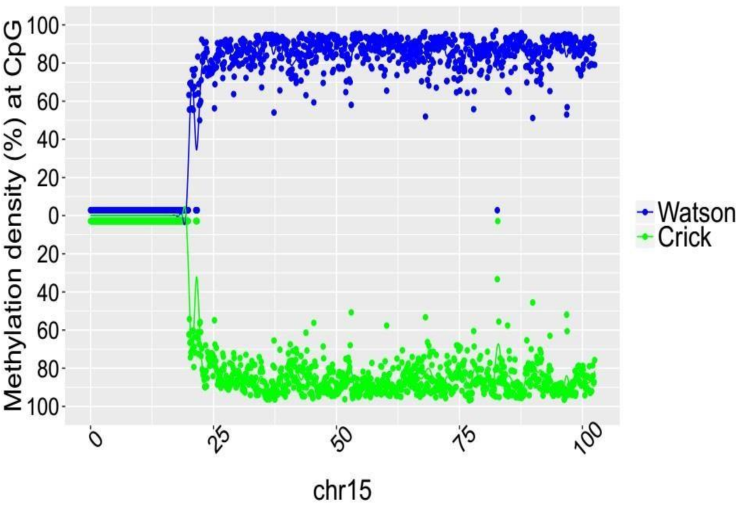
SRR921729 (55 year old male)



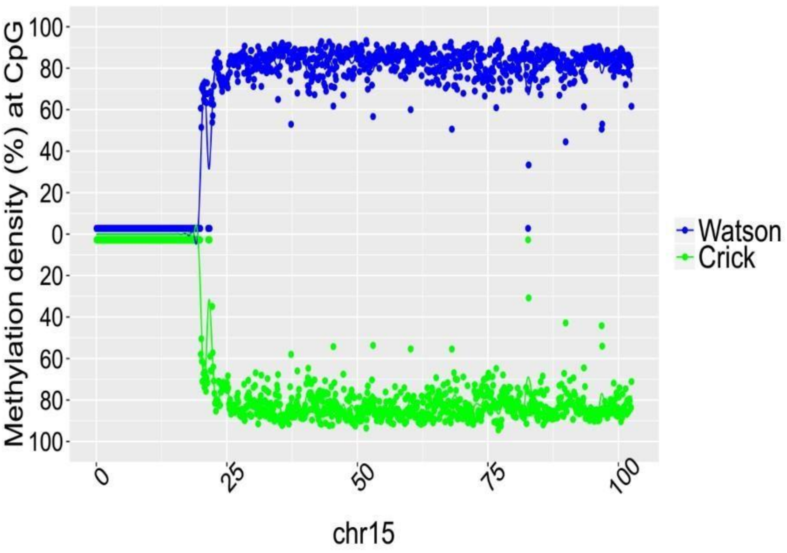
SRR921749 (55 year old male)



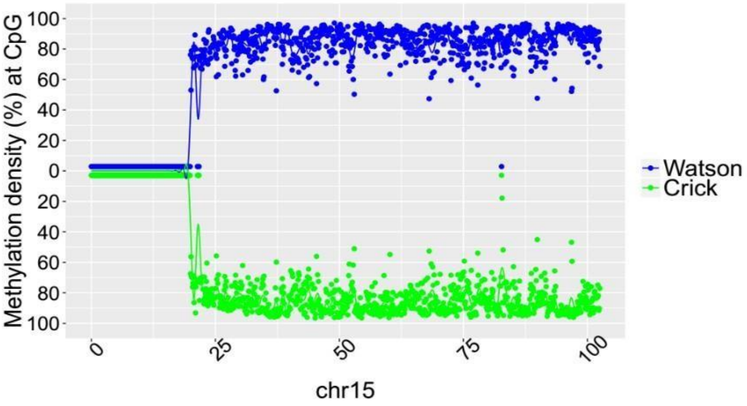
SRR921735 (55 year old male)



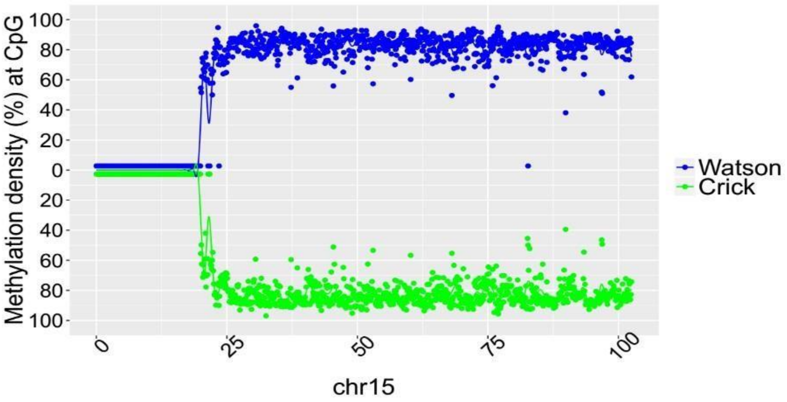
SRR847424 (42 year old female)



SRR921706 (53 year old female)

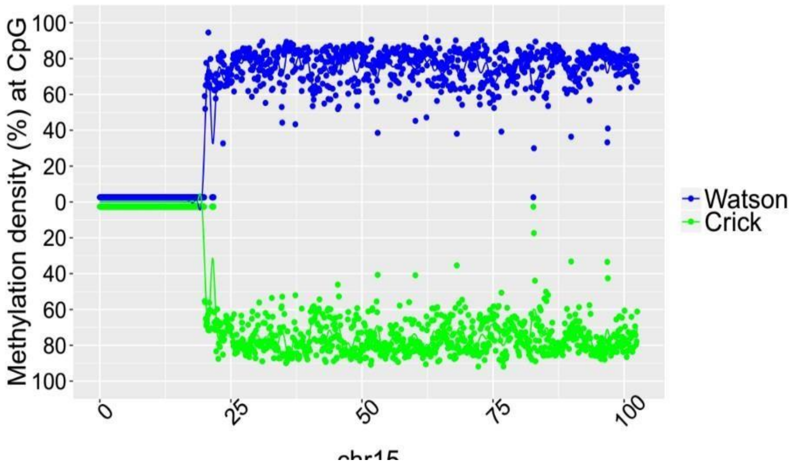


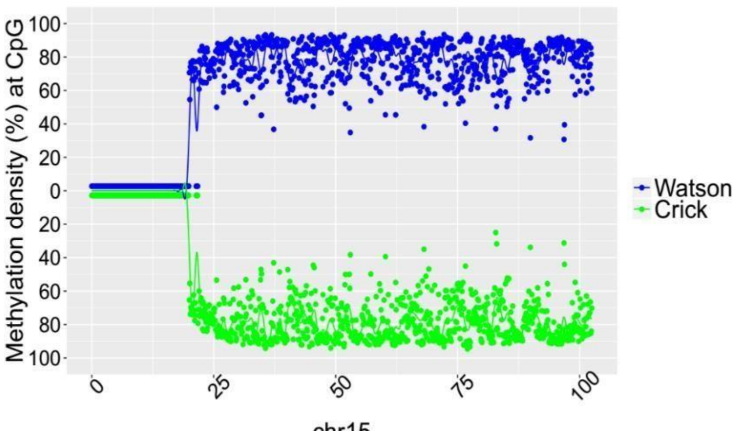
SRR921723 (53 year old female)



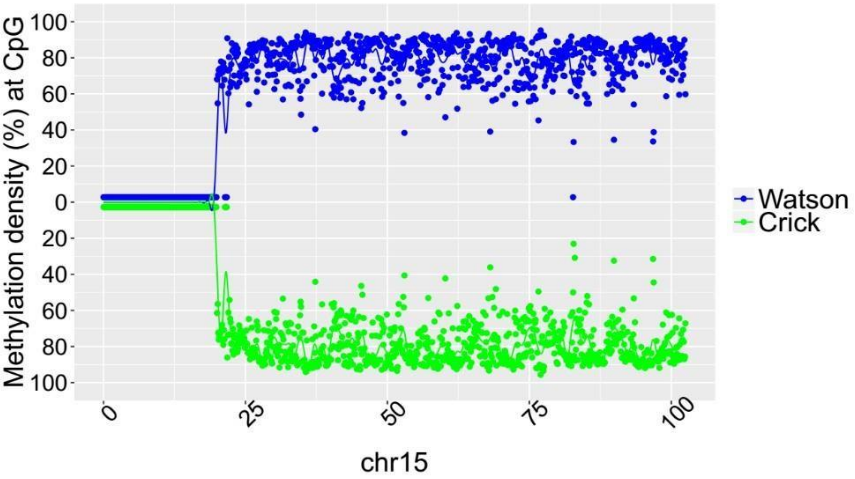
SRR949195 (81 year old female)

SRR921702 (64 year old female)

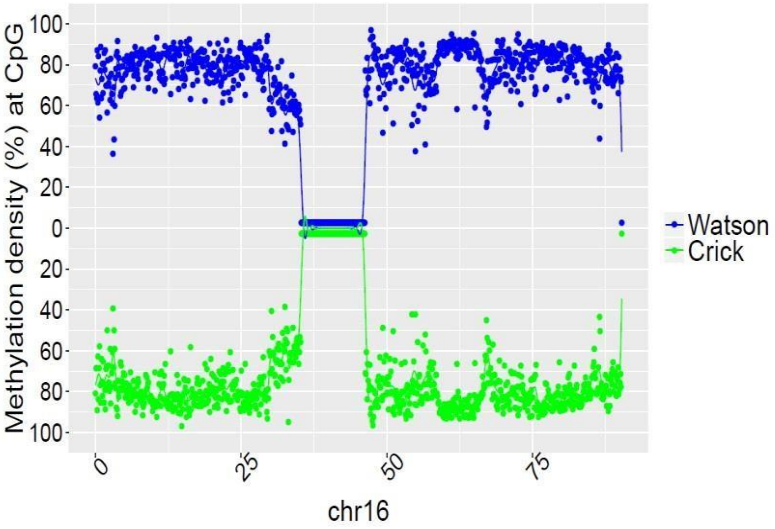




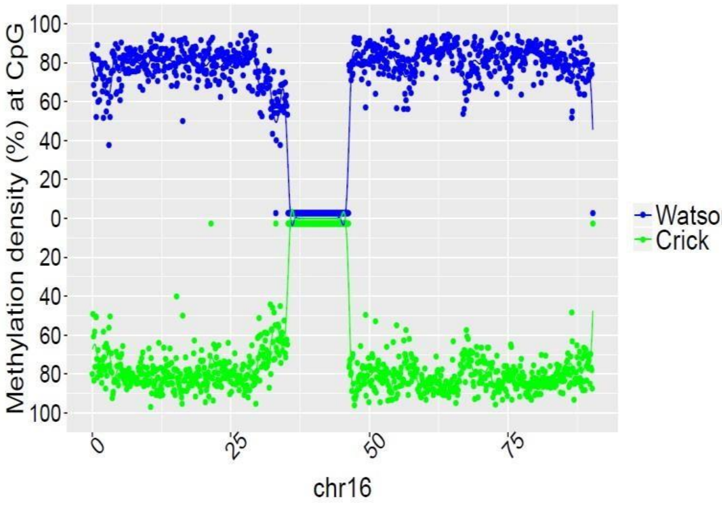
SRR949197 (82 year old female)



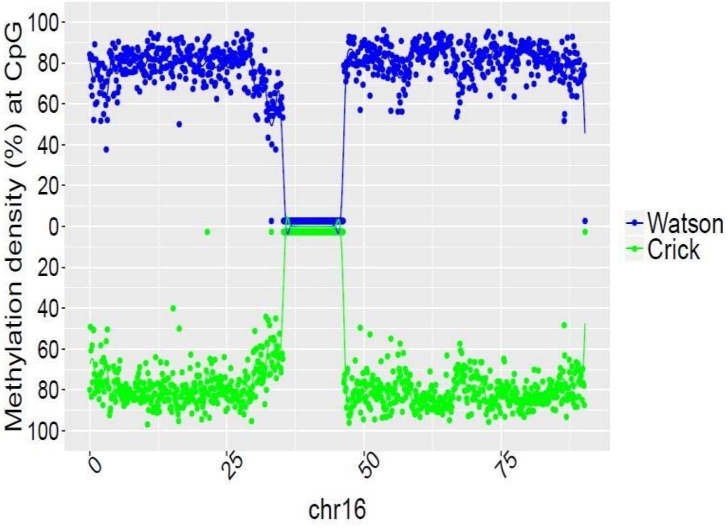
SRR901381 (20 weeks old male)



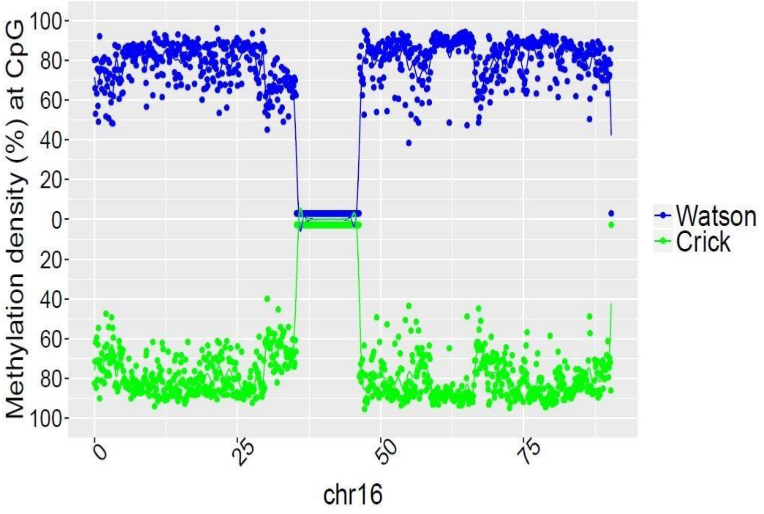
SRR47900 (31 year old male)



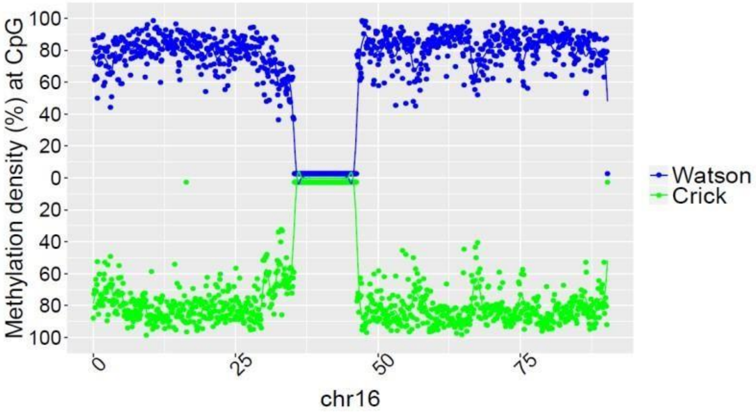
SRR 478994 (47 year old male)



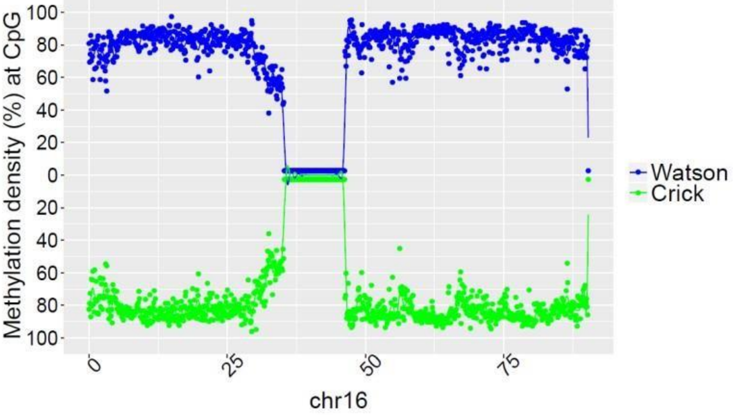
SRR478991 (48 year old male)



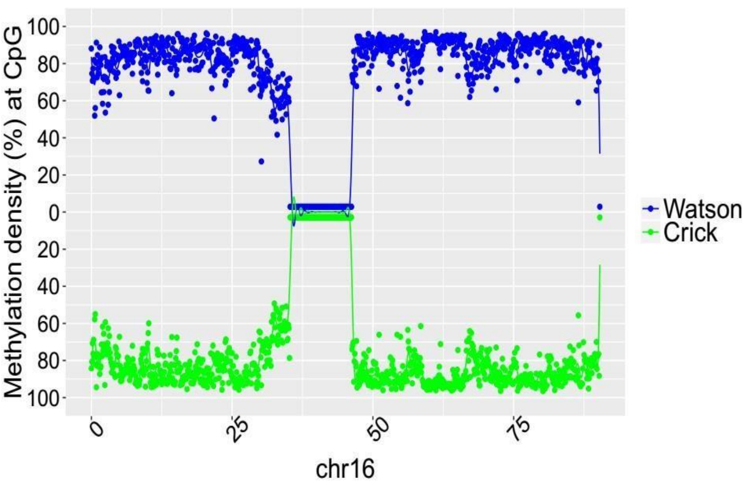
SRR921729 (55 year old male)



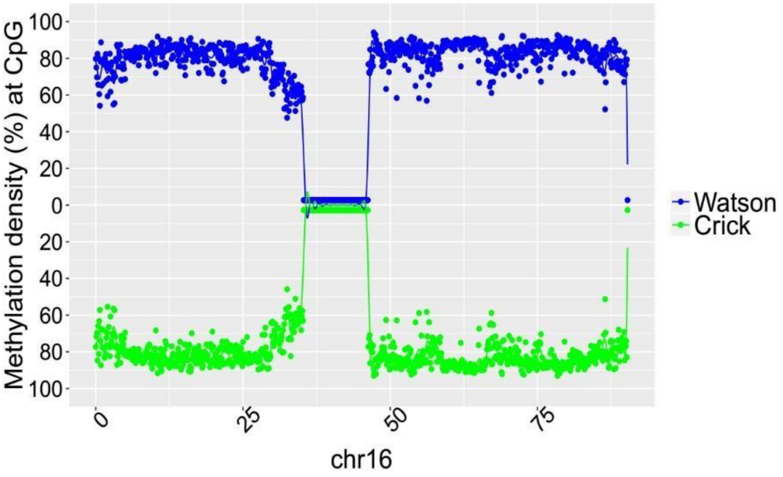
SRR921749 (55 year old male)



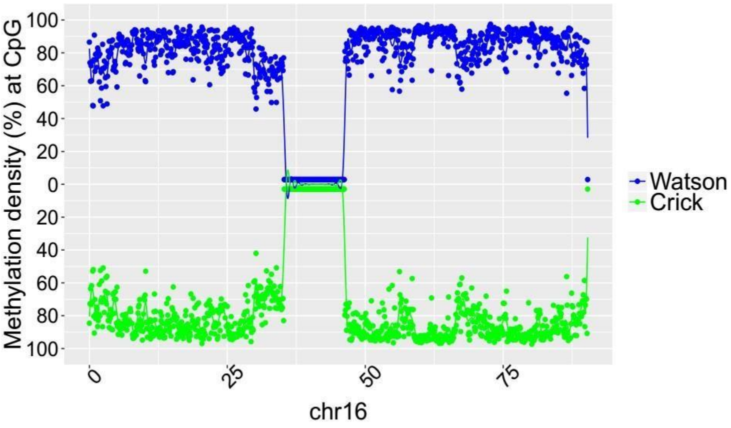
SRR921735 (55 year old male)



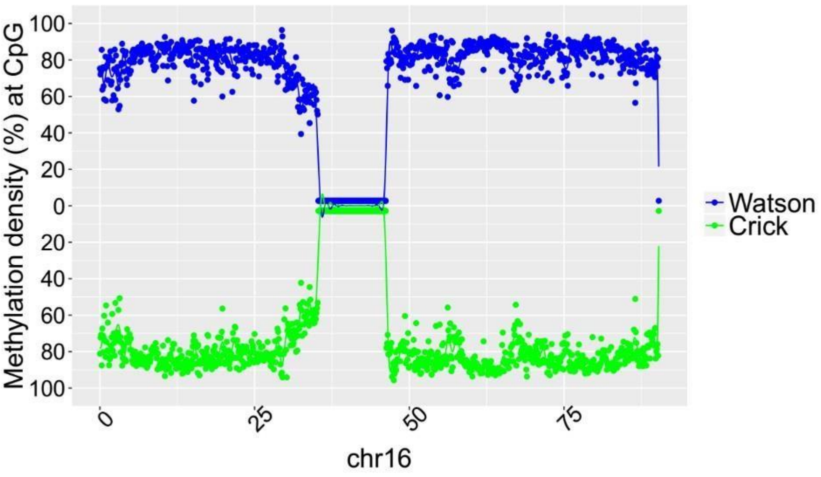
SRR847424 (42 year old female)



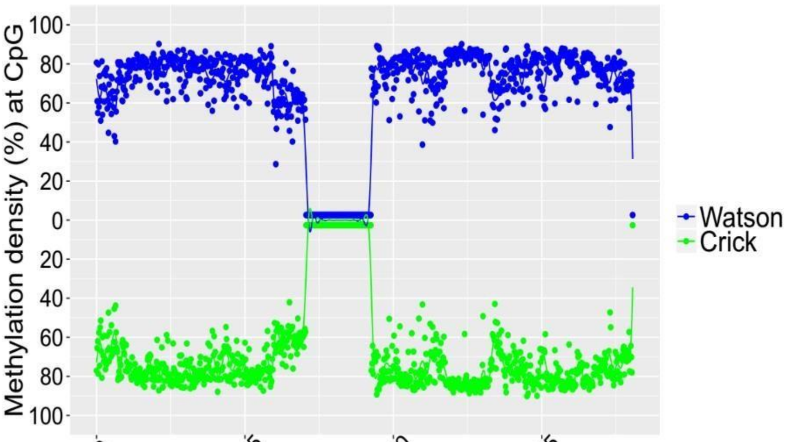
SRR921706 (53 year old female)



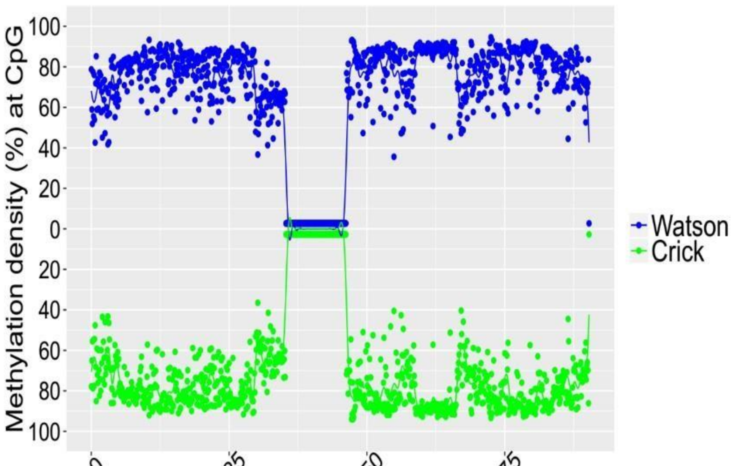
SRR921723 (53 year old female)



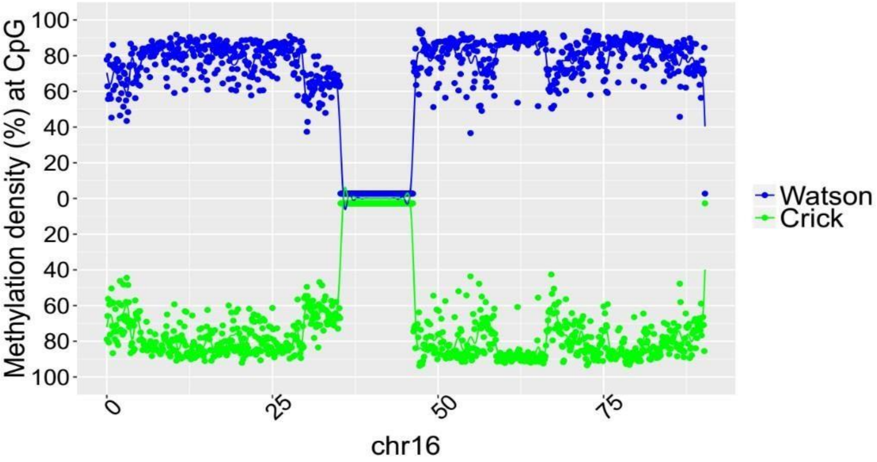
SRR921702 (64 year old female)



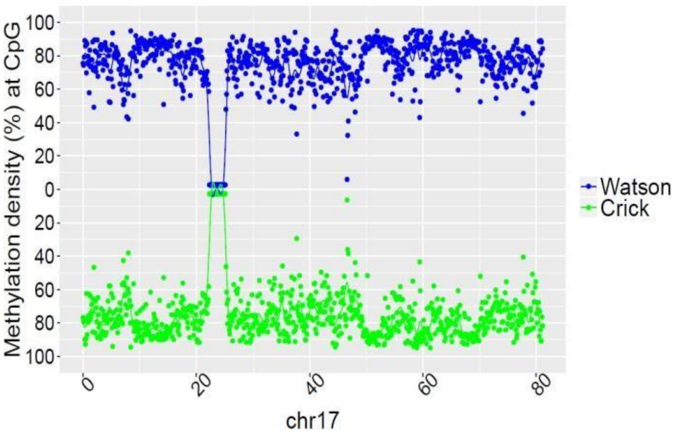
SRR949195 (81 year old female)



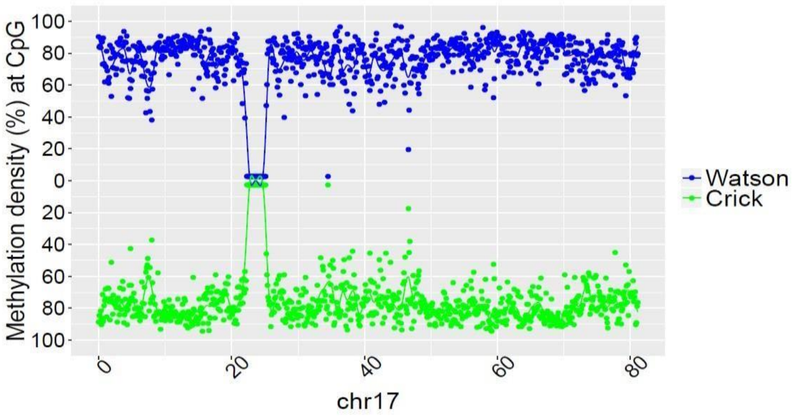
SRR949197 (82 year old female)



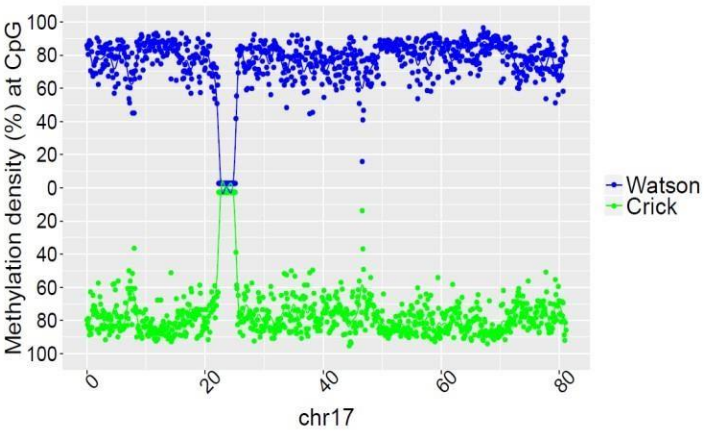
SRR901381 (20 weeks old male)



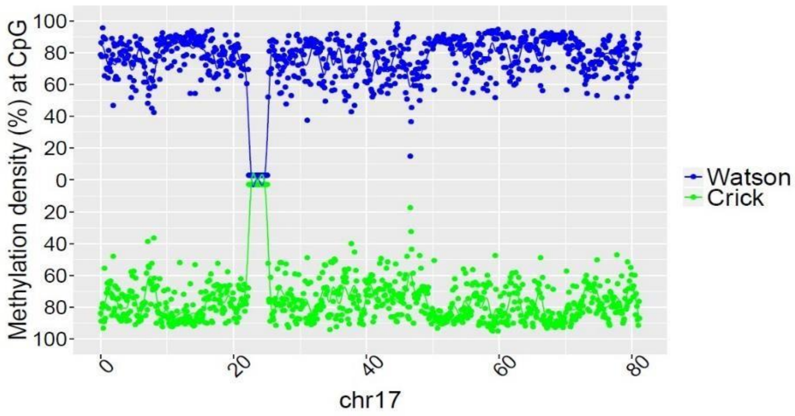
SRR47900 (31 year old male)



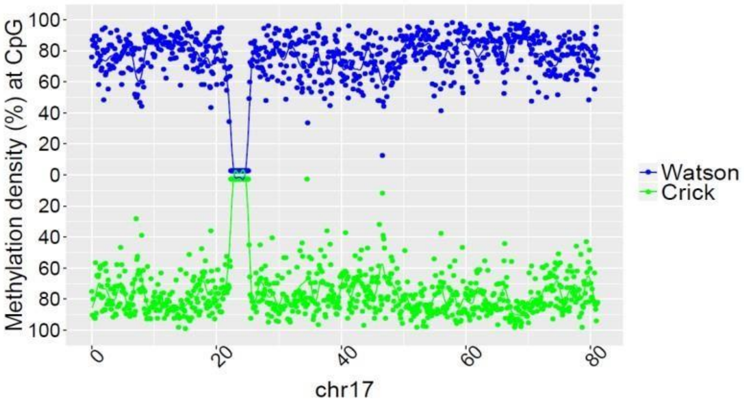
SRR 478994 (47 year old male)



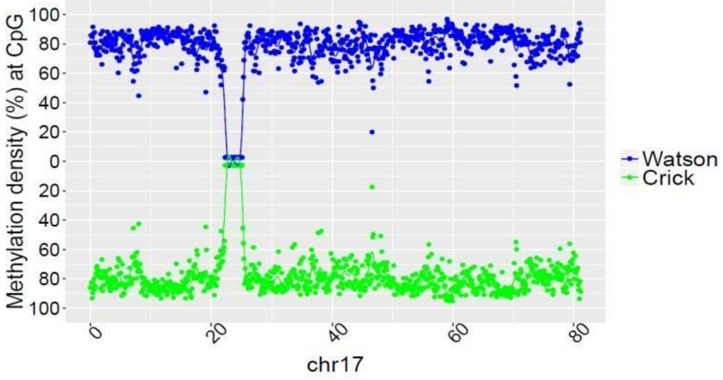
SRR478991 (48 year old male)



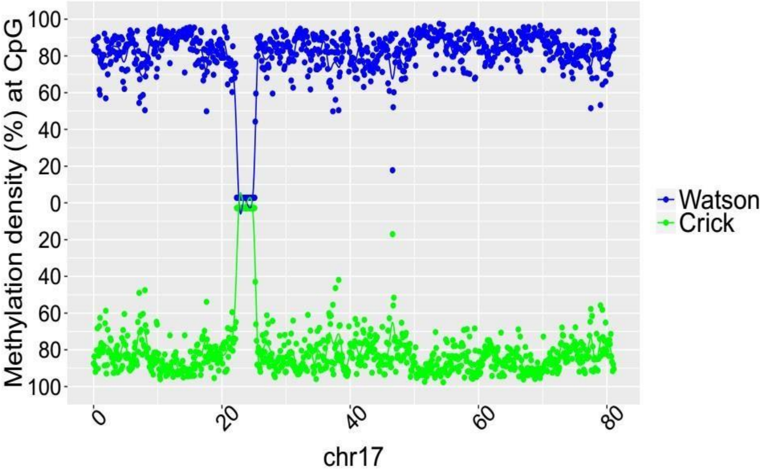
SRR921729 (55 year old male)



SRR921749 (55 year old male)



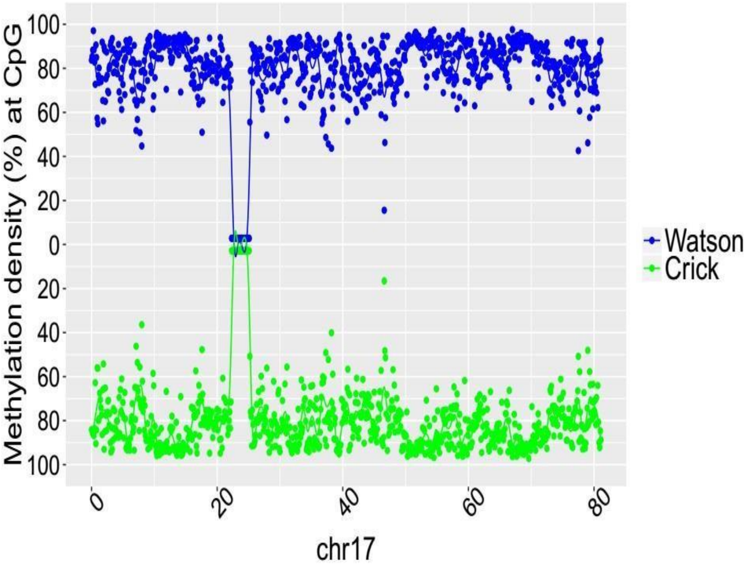
SRR921735 (55 year old male)



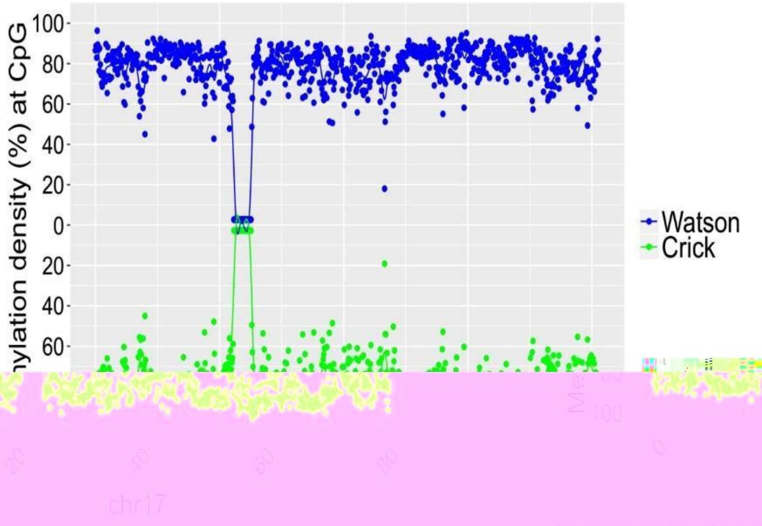
SRR847424 (42 year old female)

Não é possível apresentar a imagem.

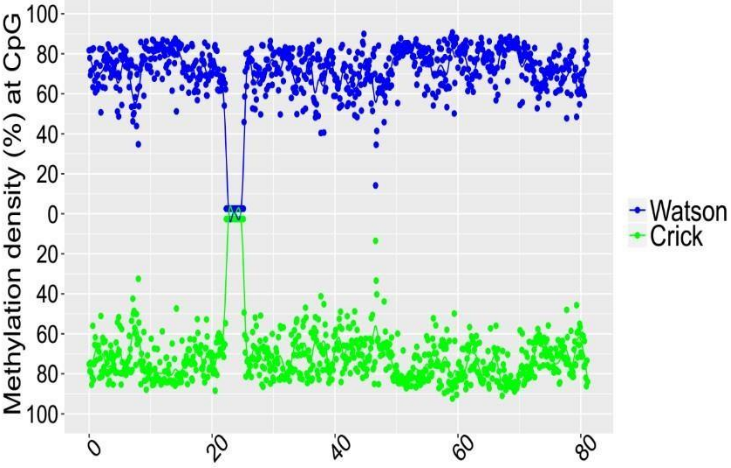
SRR921706 (53 year old female)



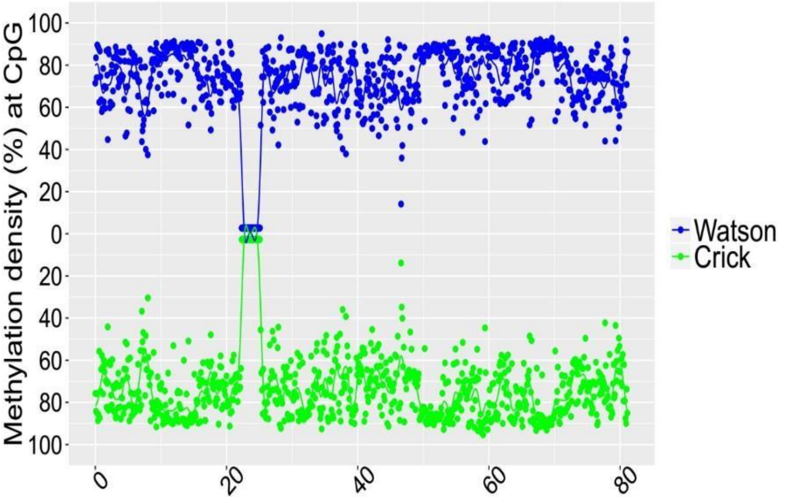
SRR921723 (53 year old female)



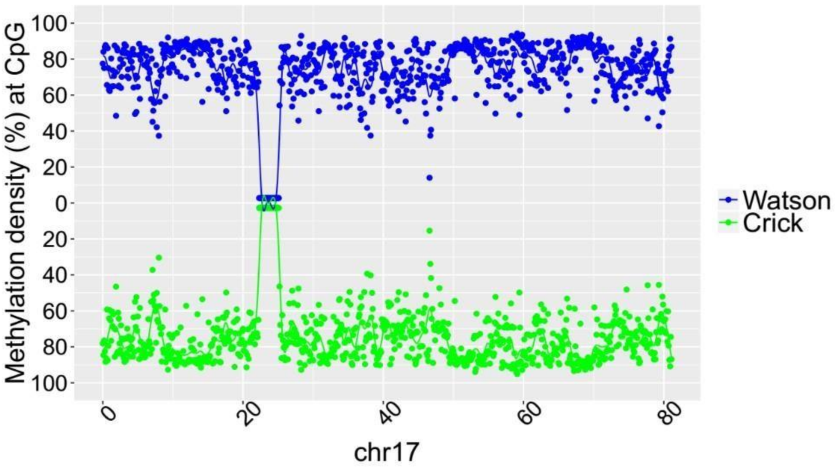
SRR921702 (64 year old female)



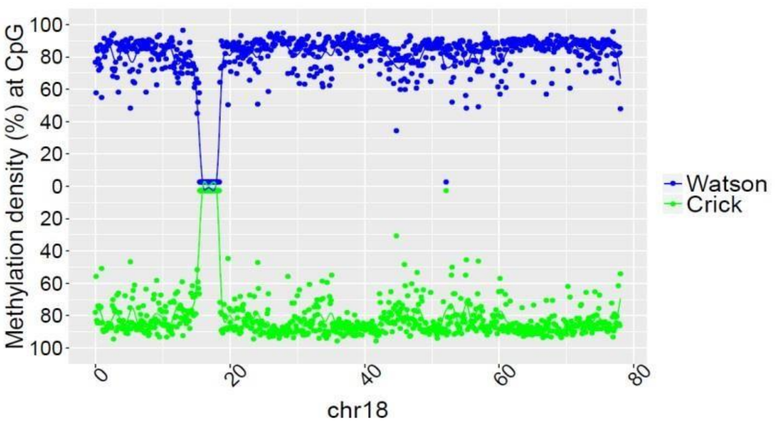
SRR949195 (81 year old female)



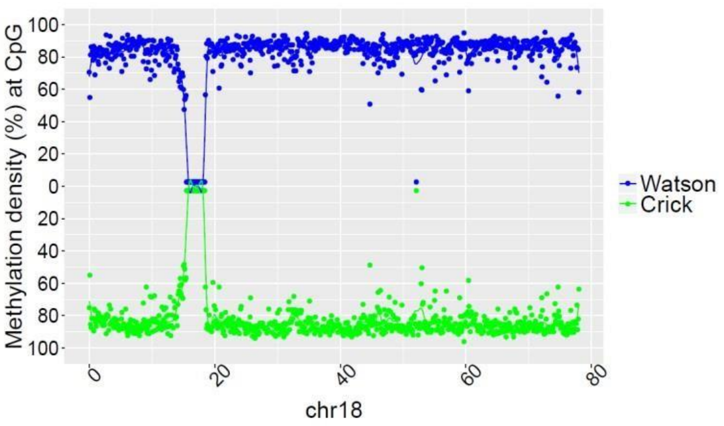
SRR949197 (82 year old female)



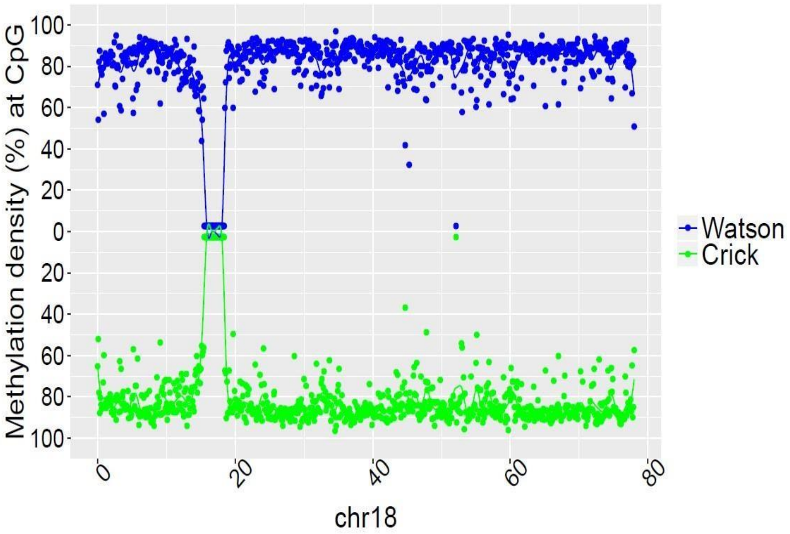
SRR901381 (20 weeks old male)



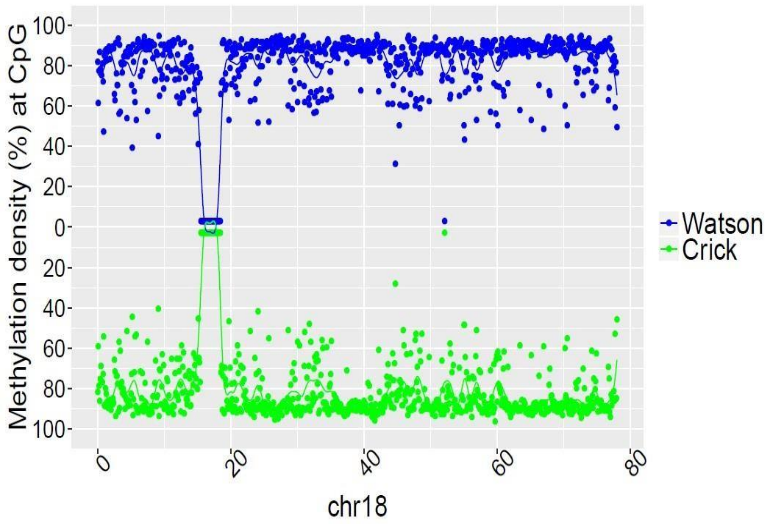
SRR47900 (31 year old male)



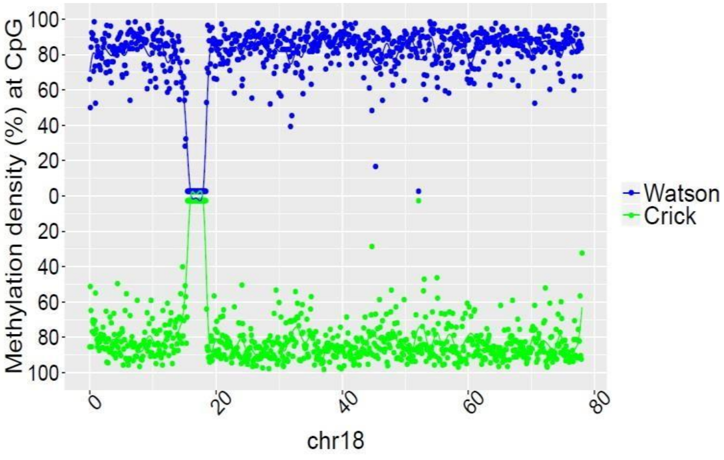
SRR 478994 (47 year old male)



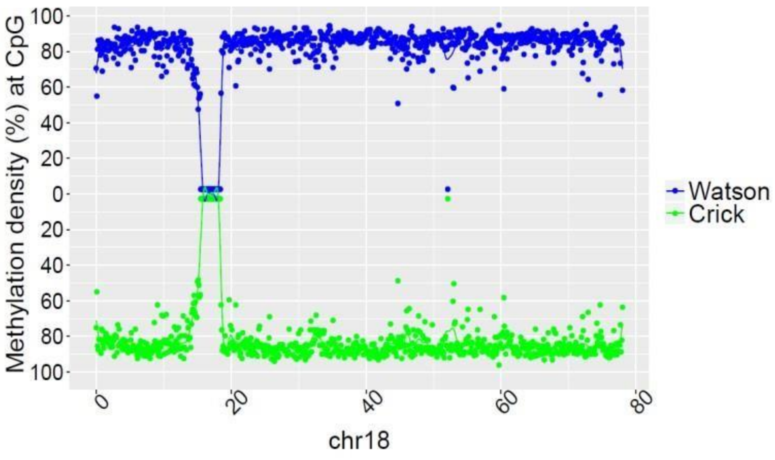
SRR478991 (48 year old male)



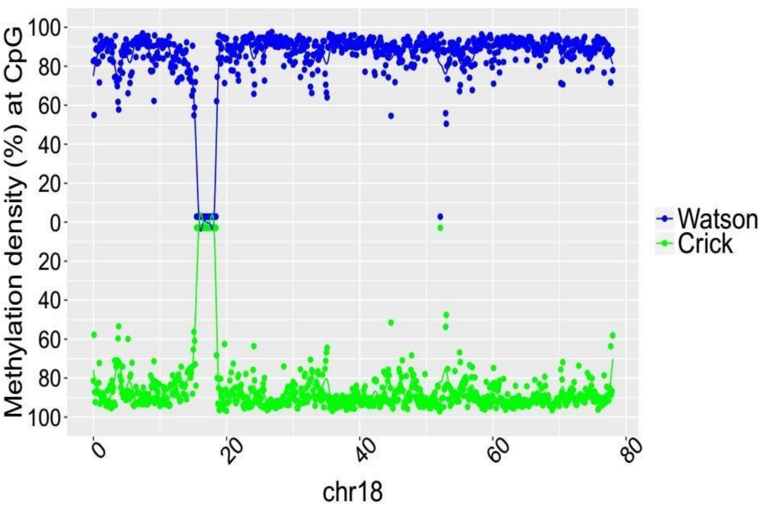
SRR921729 (55 year old male)



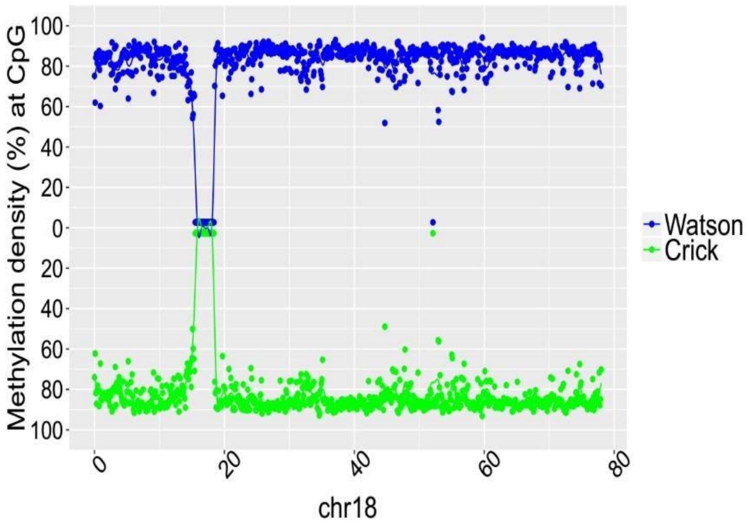
SRR921749 (55 year old male)



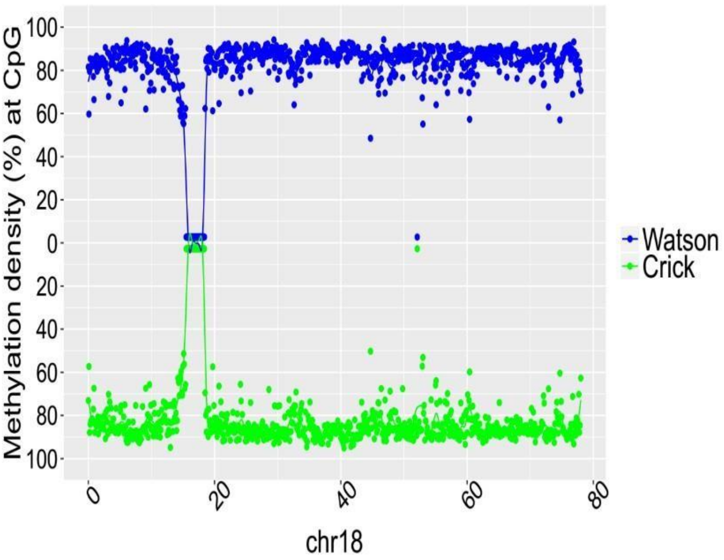
SRR921735 (55 year old male)



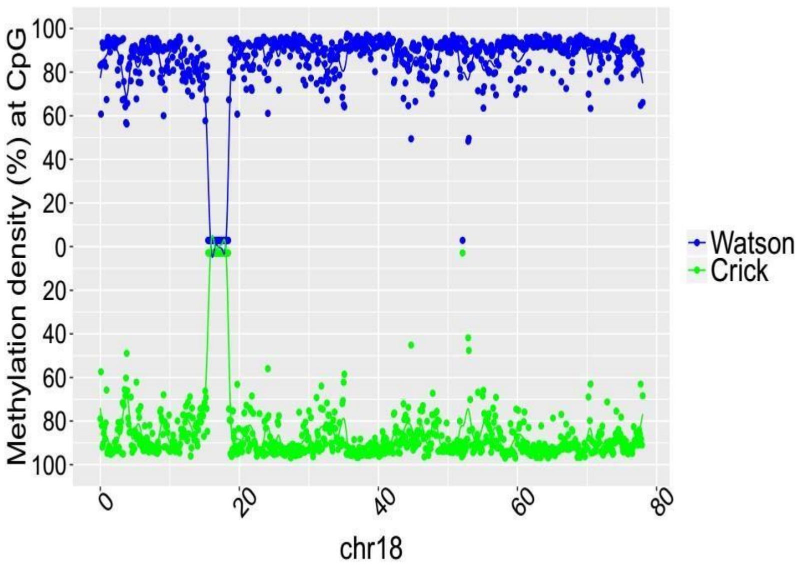
SRR847424 (42 year old female)



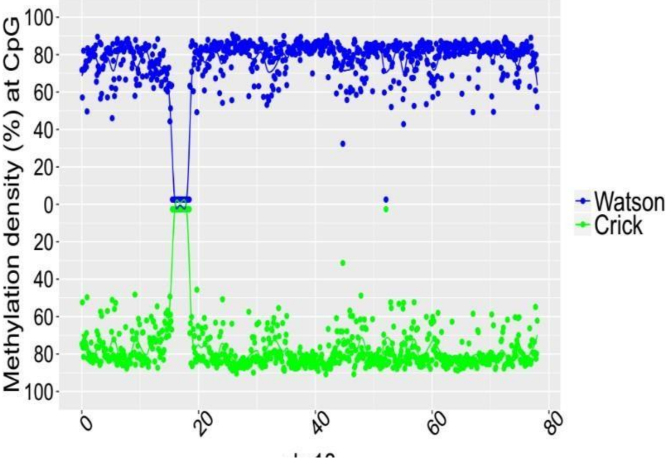
SRR921706 (53 year old female)



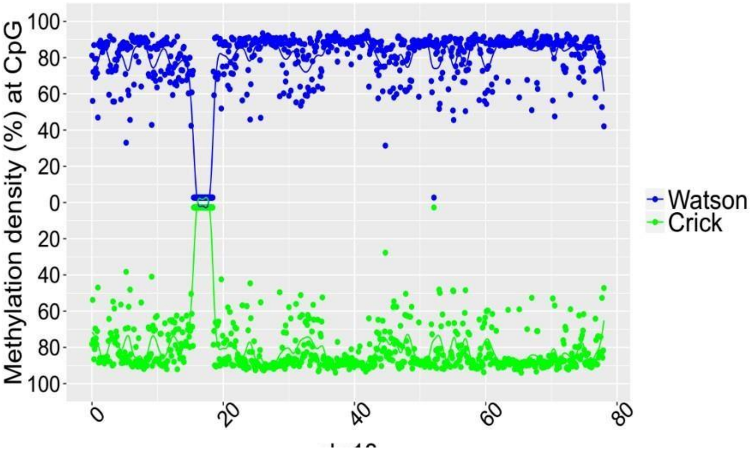
SRR921723 (53 year old female)



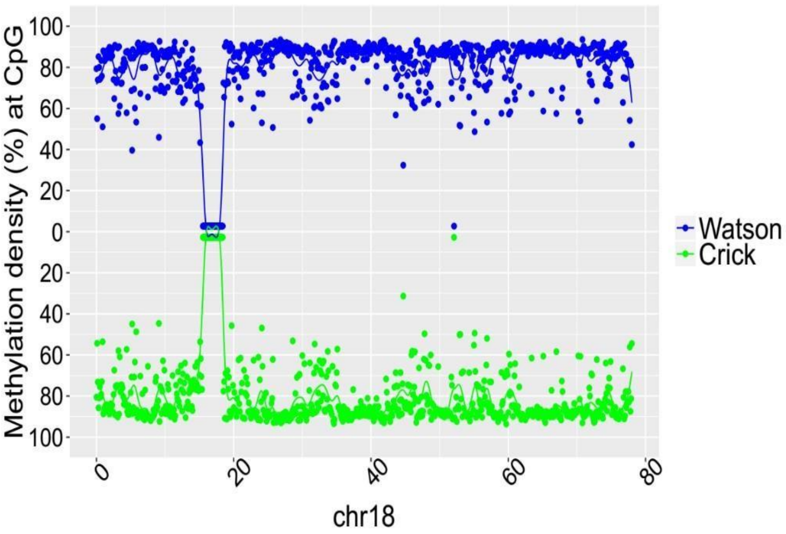
SRR921702 (64 year old female)



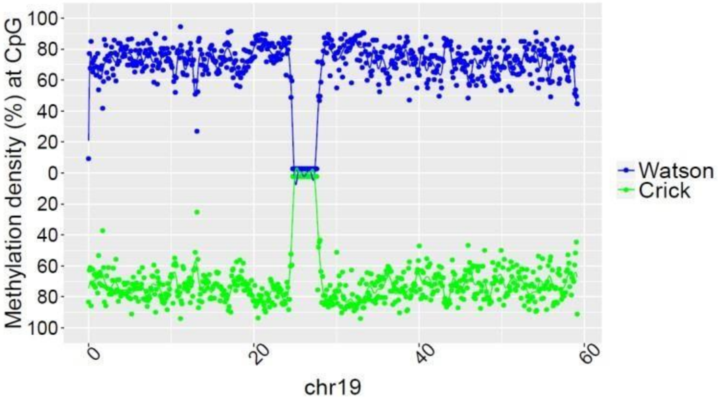
SRR949195 (81 year old female)



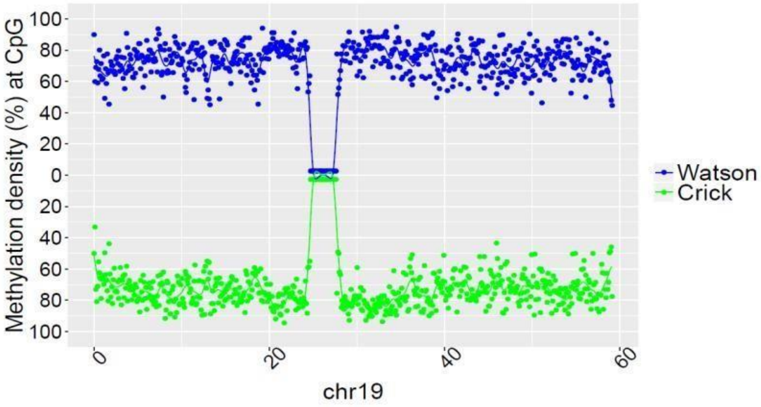
SRR949197 (82 year old female)



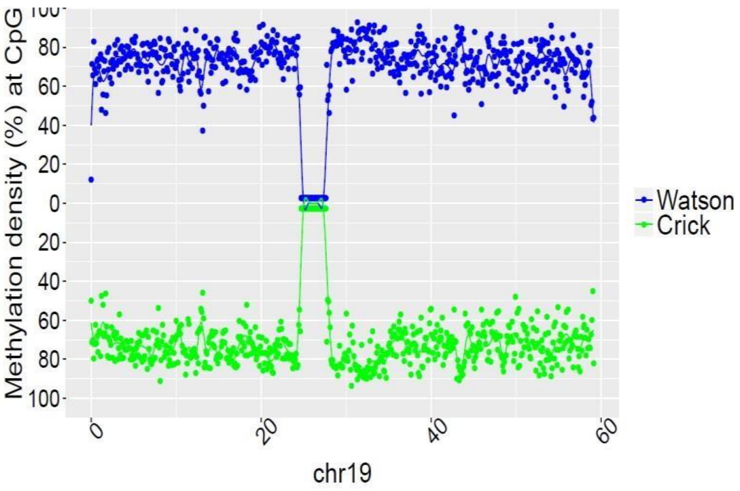
SRR901381 (20 weeks old male)



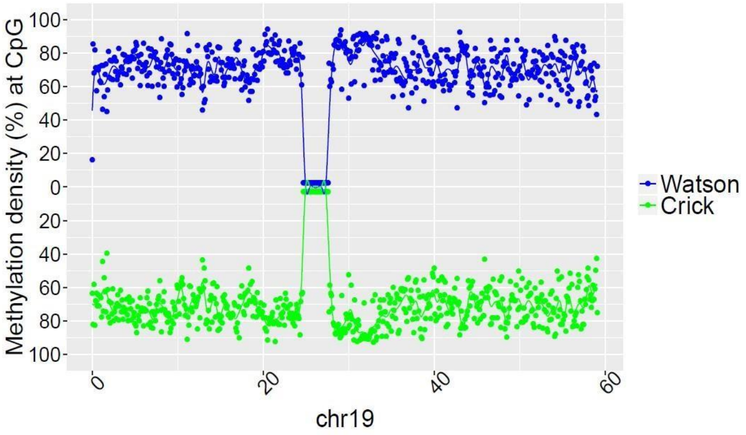
SRR47900 (31 year old male)



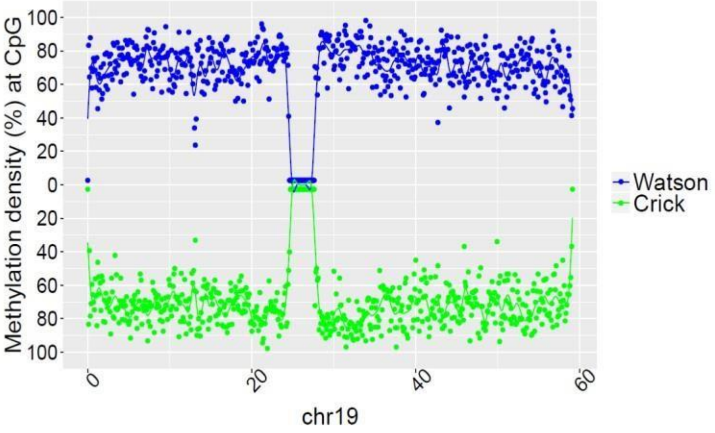
SRR478994 (47 year old male)



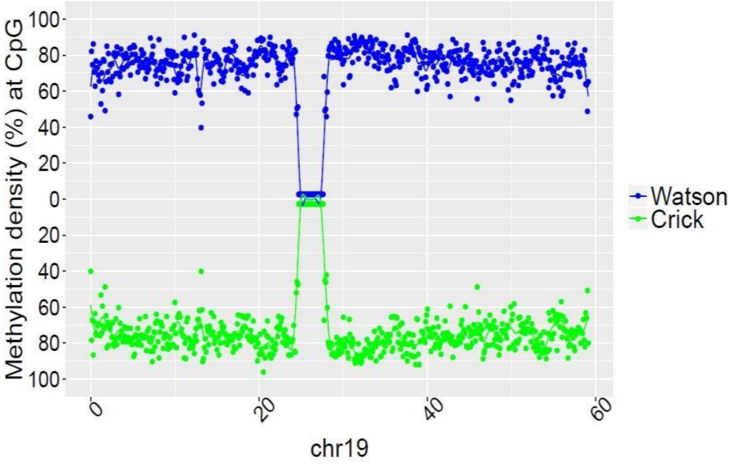
SRR478991 (48 year old male)



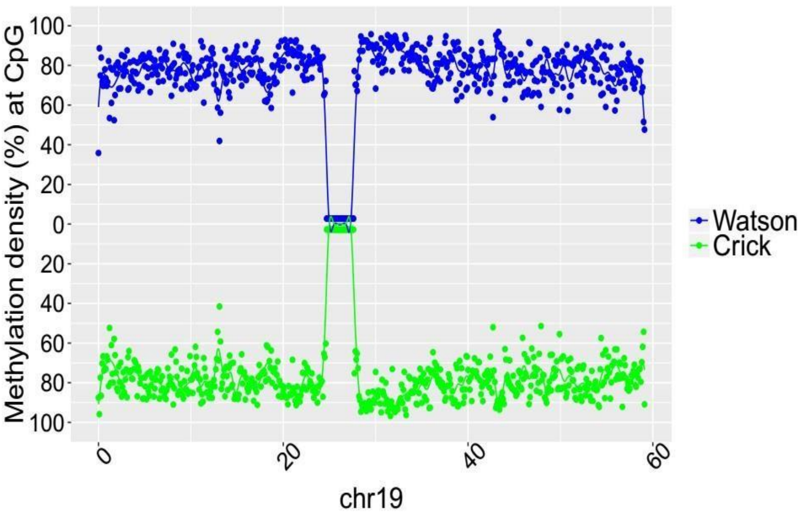
SRR921729 (55 year old male)



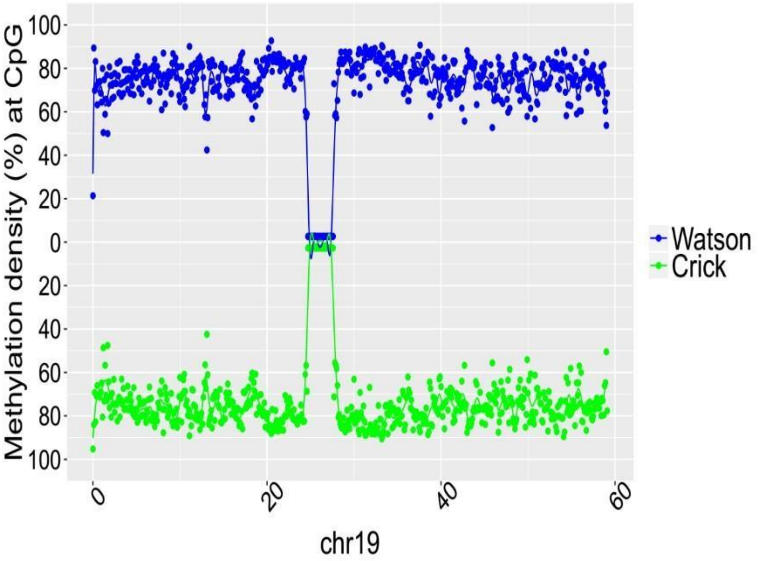
SRR921749 (55 year old male)



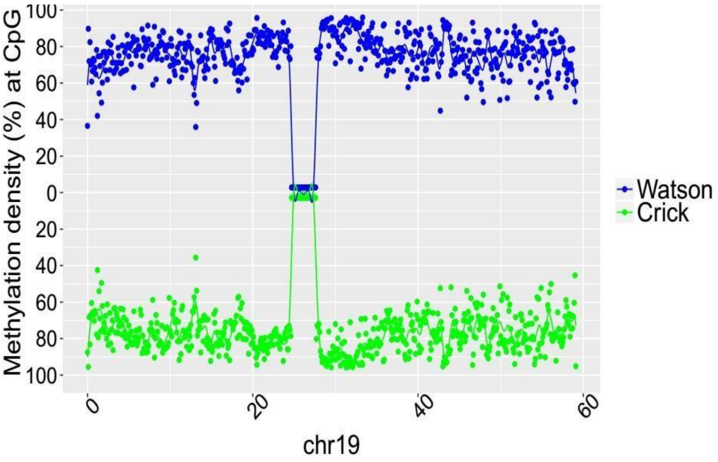
SRR921735 (55 year old male)



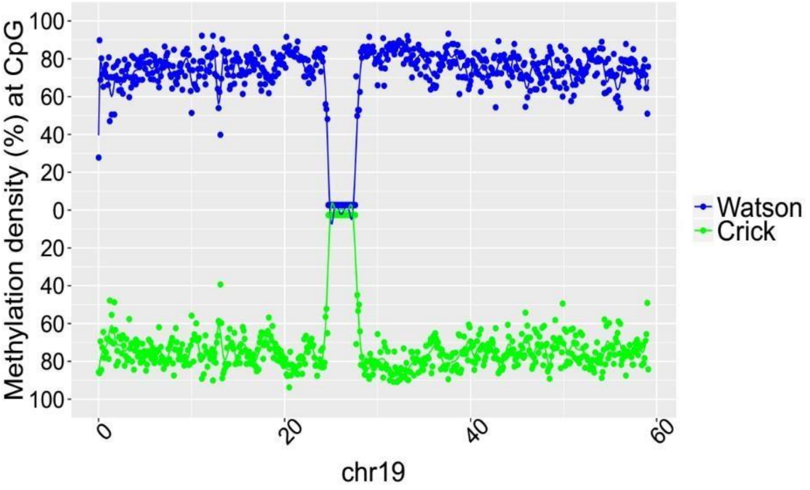
SRR847424 (42 year old female)



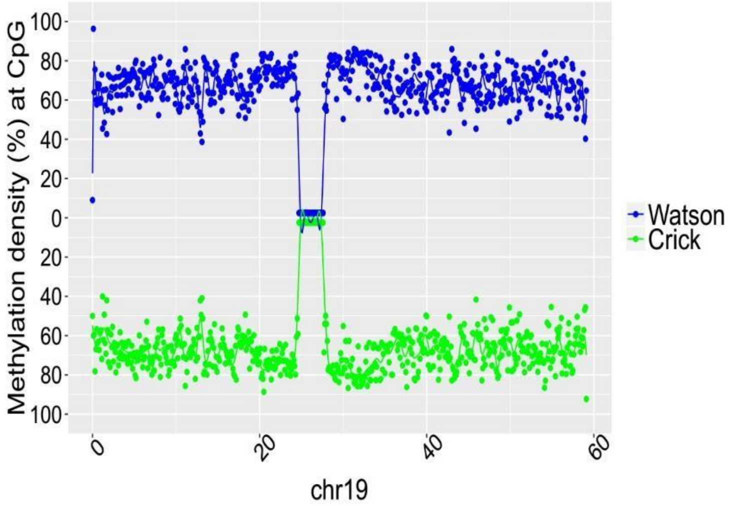
SRR921706 (53 year old female)



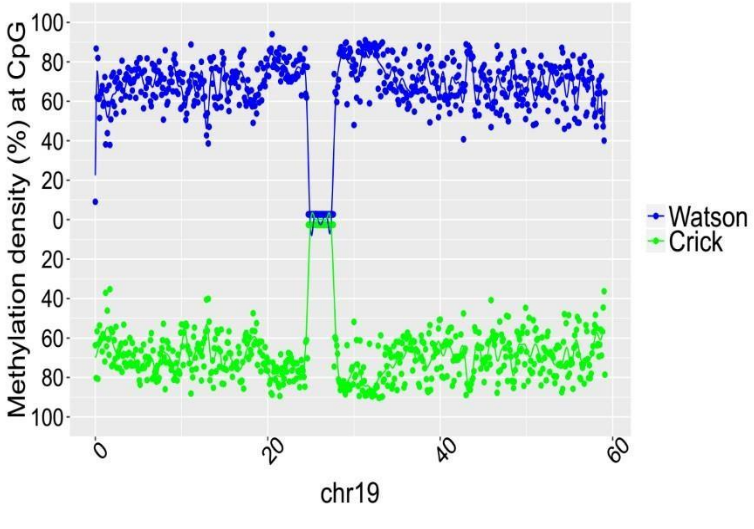
SRR921723 (53 year old female)



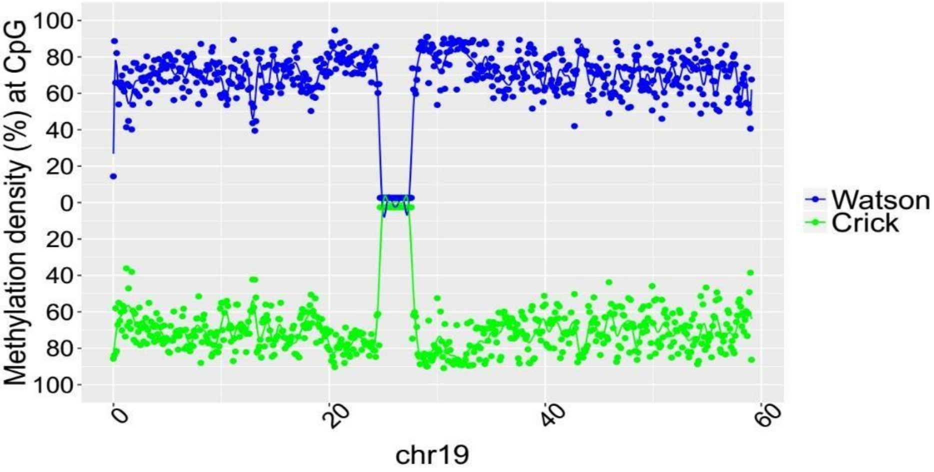
SRR921702 (64 year old female)



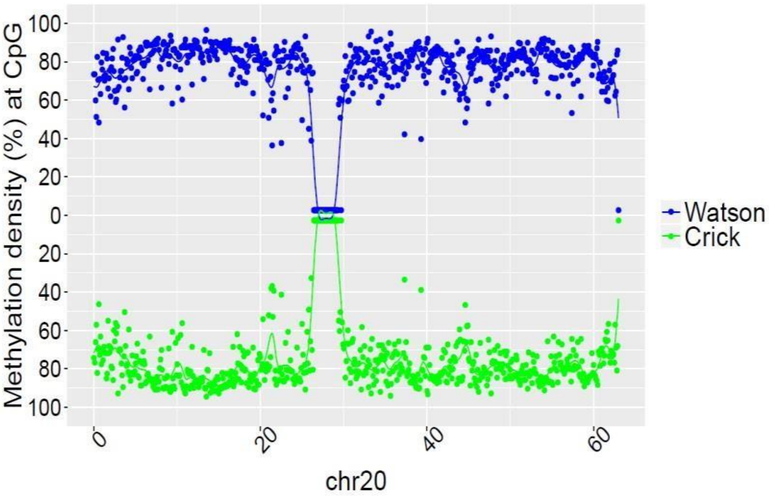
SRR949195 (81 year old female)



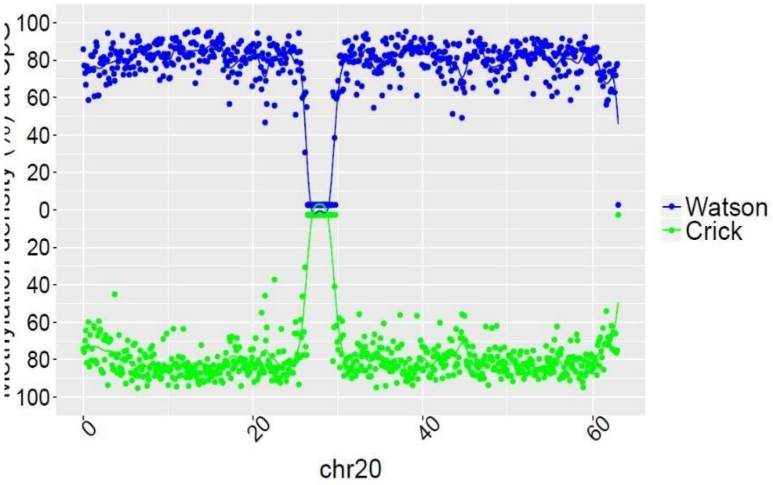
SRR949197 (82 year old female)



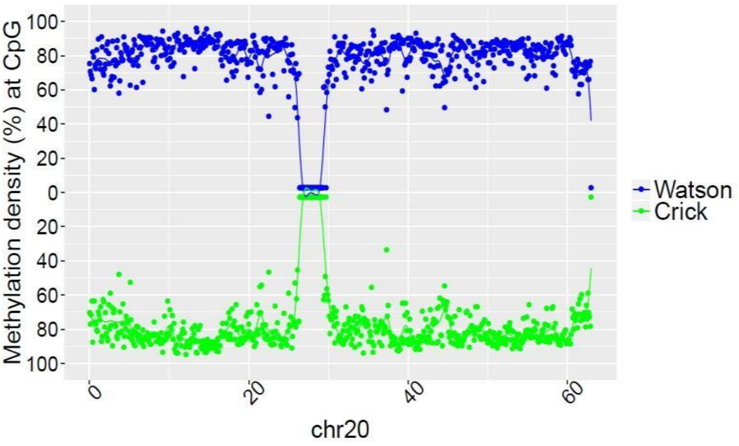
SRR901381 (20 weeks old male)



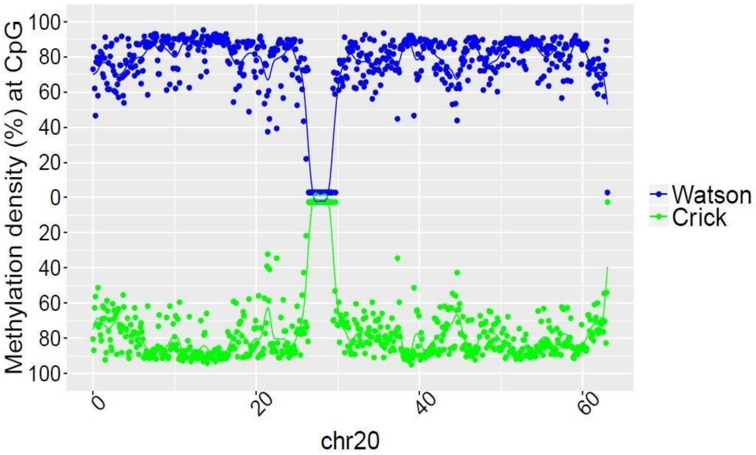
SRR47900 (31 year old male)



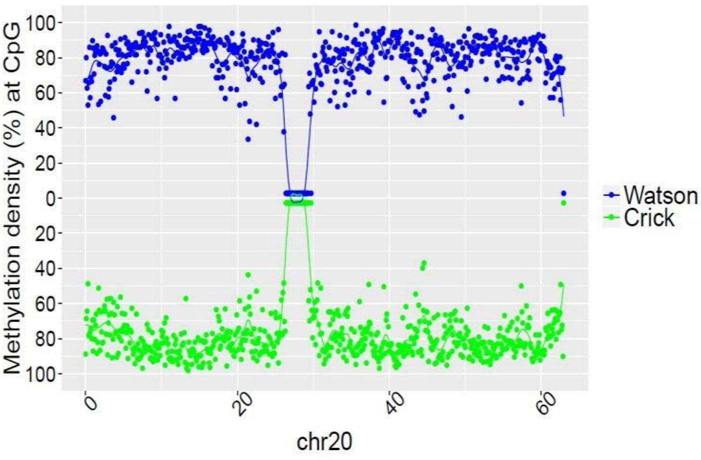
SRR 478994 (47 year old male)



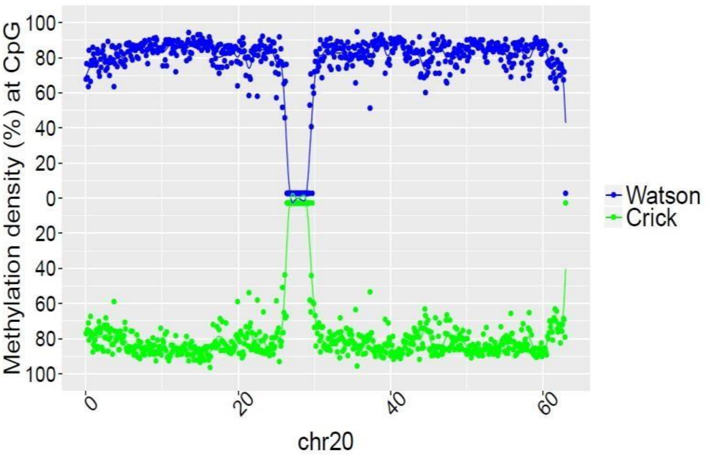
SRR478991 (48 year old male)



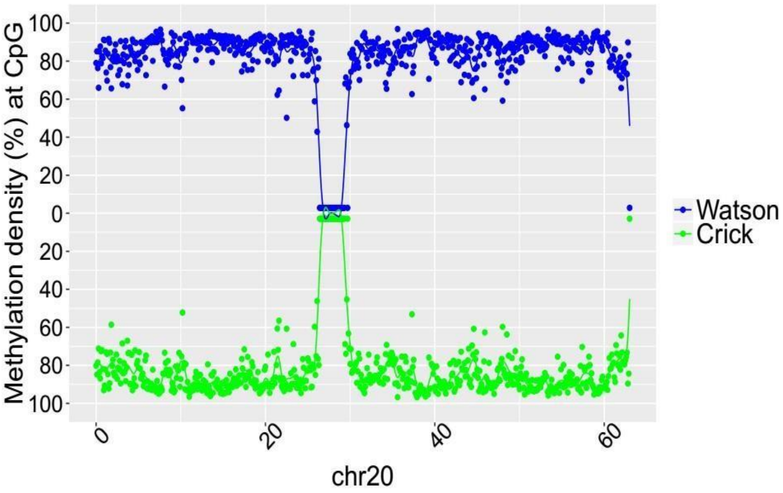
SRR921729 (55 year old male)



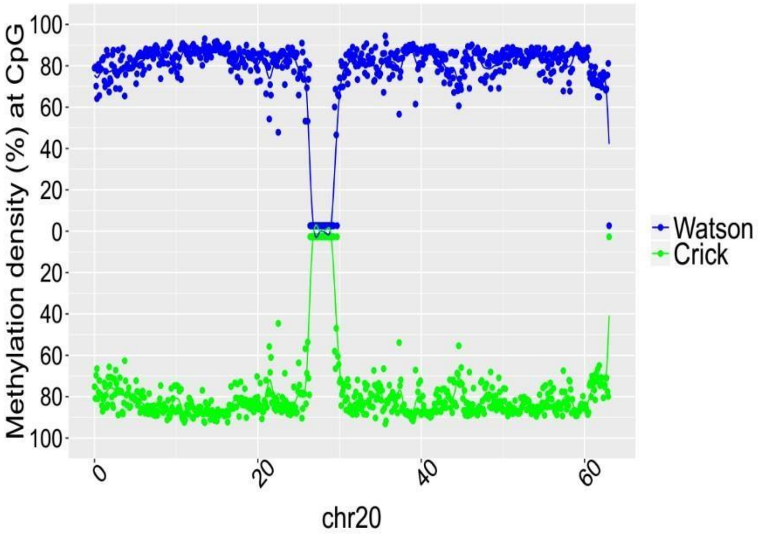
SRR921749 (55 year old male)



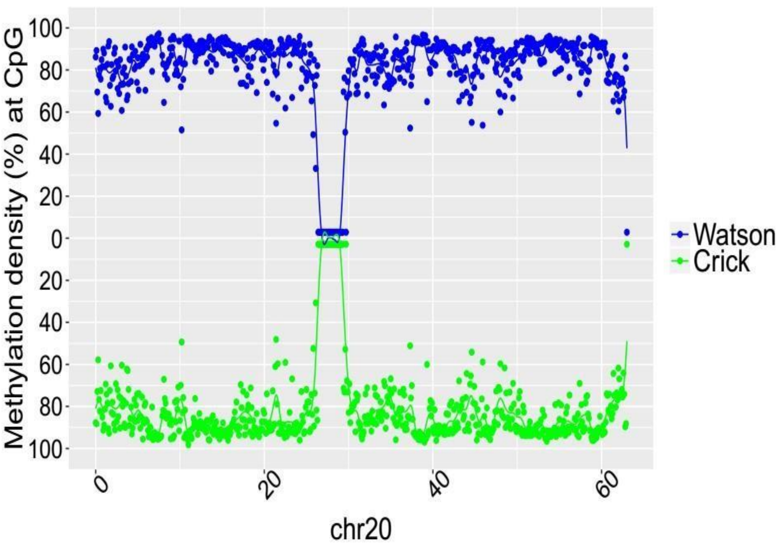
SRR921735 (55 year old male)



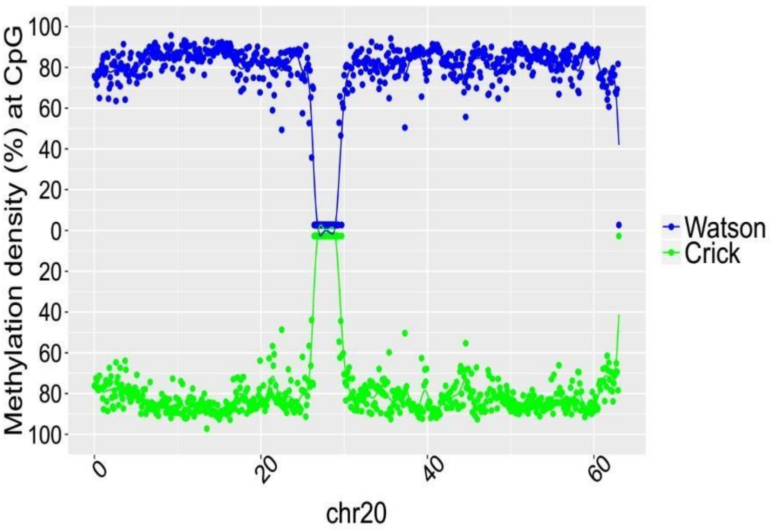
SRR847424 (42 year old female)



SRR921706 (53 year old female)

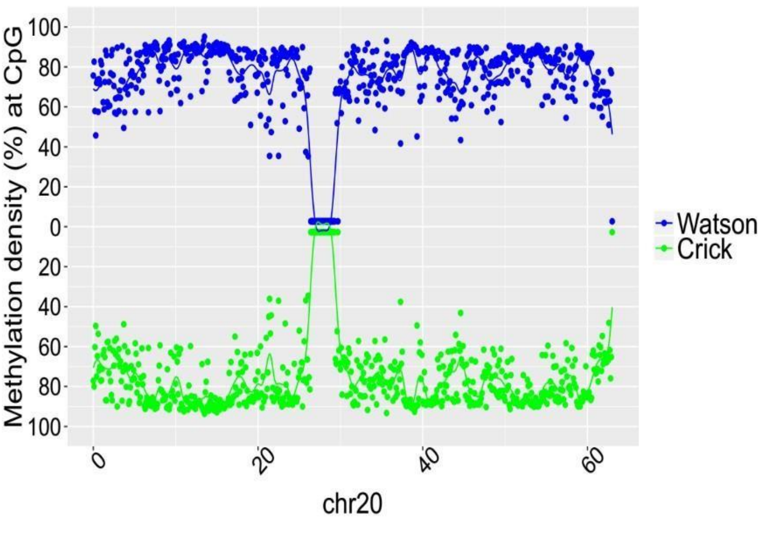


SRR921723 (53 year old female)



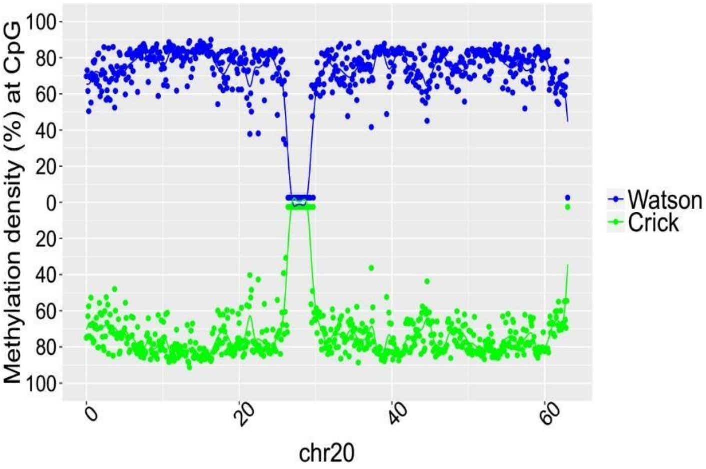
SRR921702 (64 year old female)

SRR949195 (81 year old female)

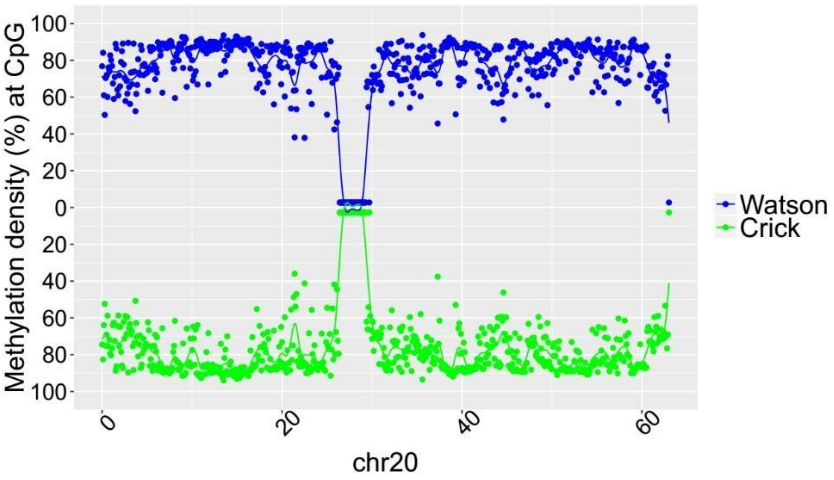


SRR921735 (55 year old male)

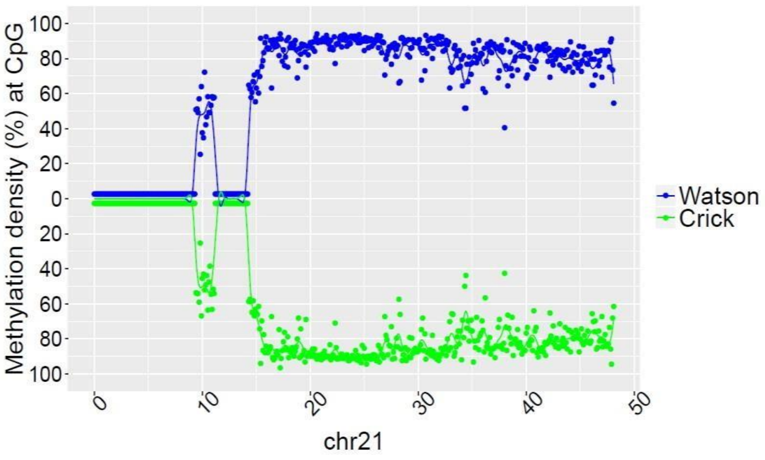
SRR847424 (42 year old female)



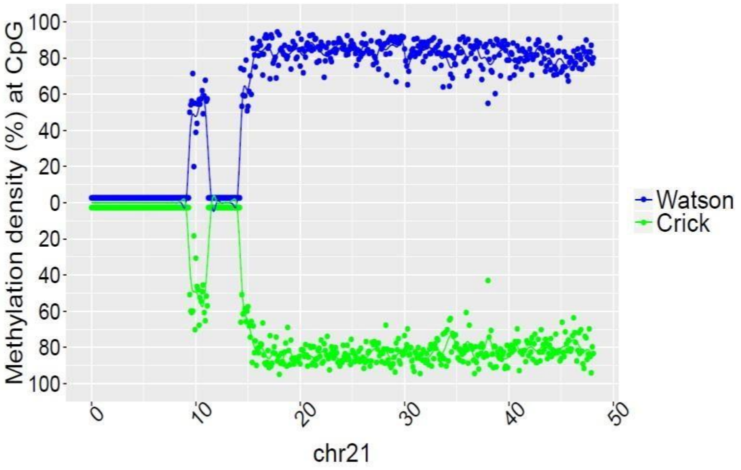
SRR949197 (82 year old female)



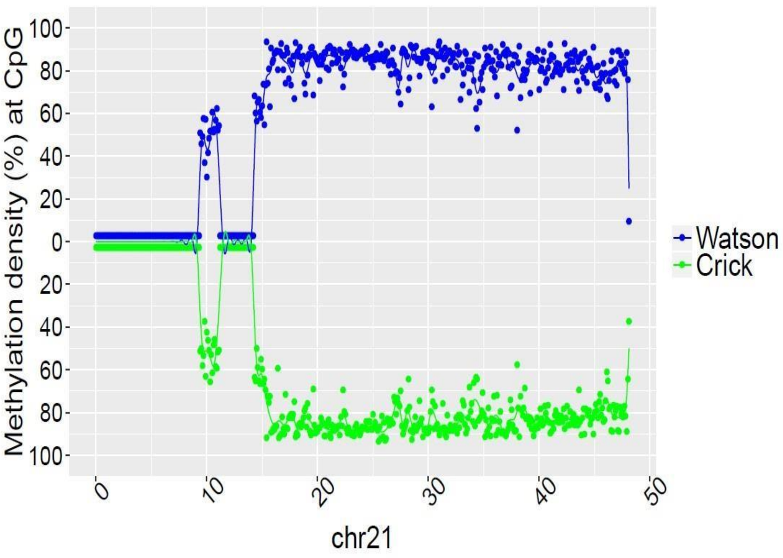
SRR901381 (20 weeks old male)



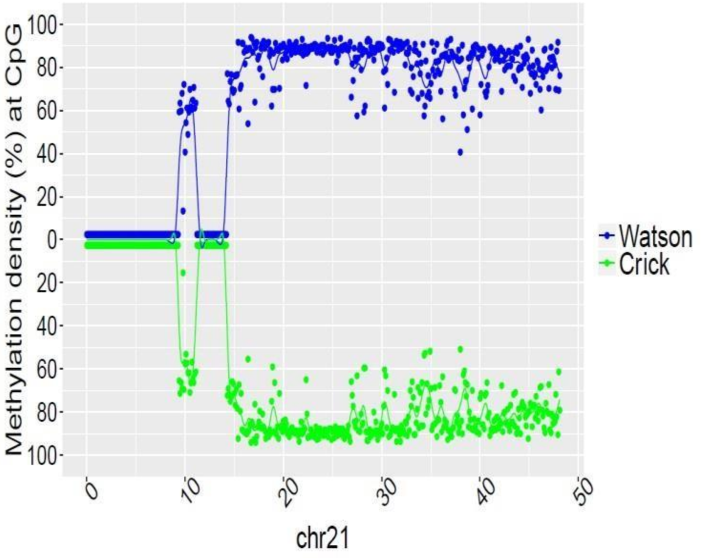
SRR47900 (31 year old male)



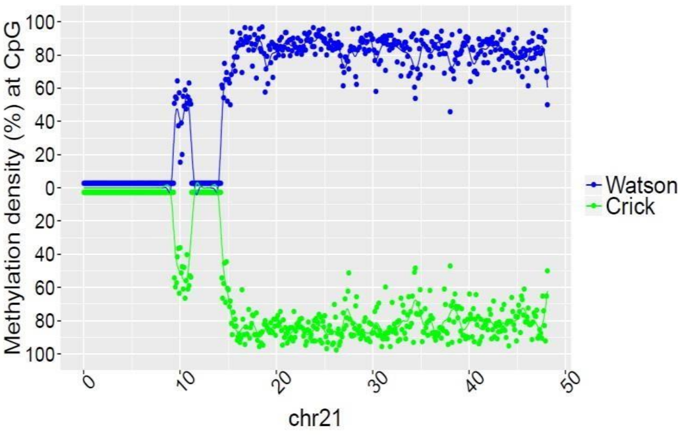
SRR478994 (47 year old male)



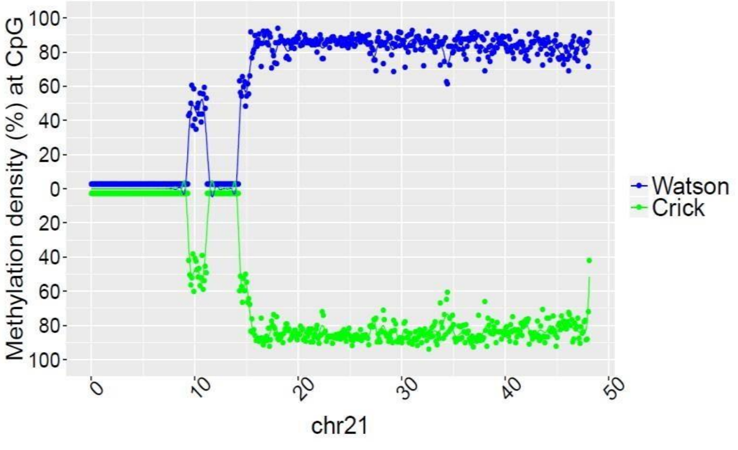
SRR478991 (48 year old male)



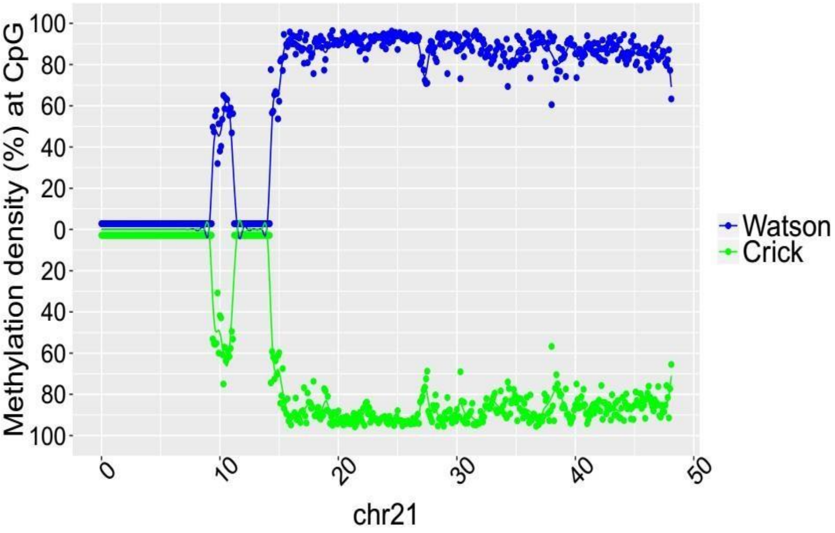
SRR921729 (55 year old male)



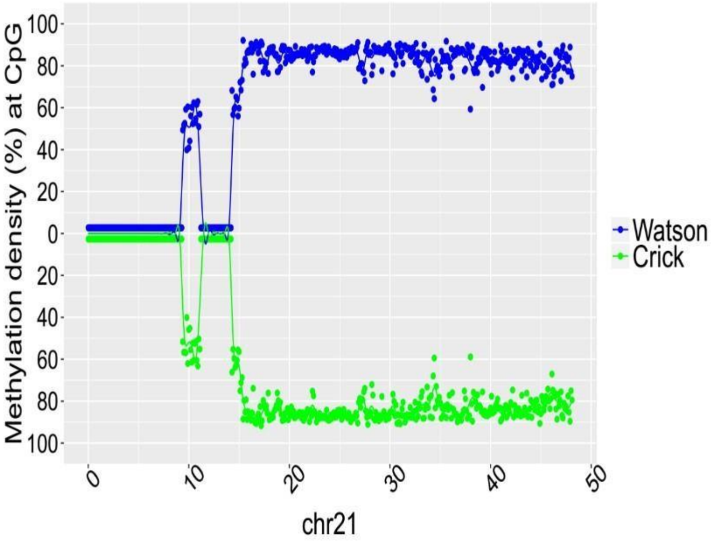
SRR921749 (55 year old male)



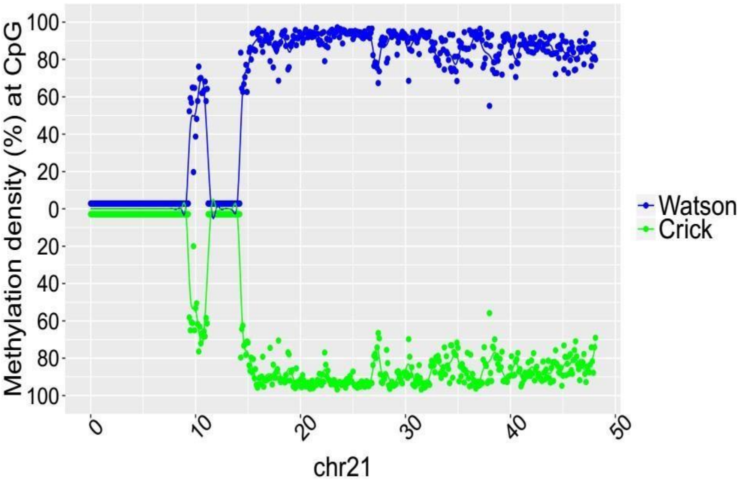
SRR921735 (55 year old male)



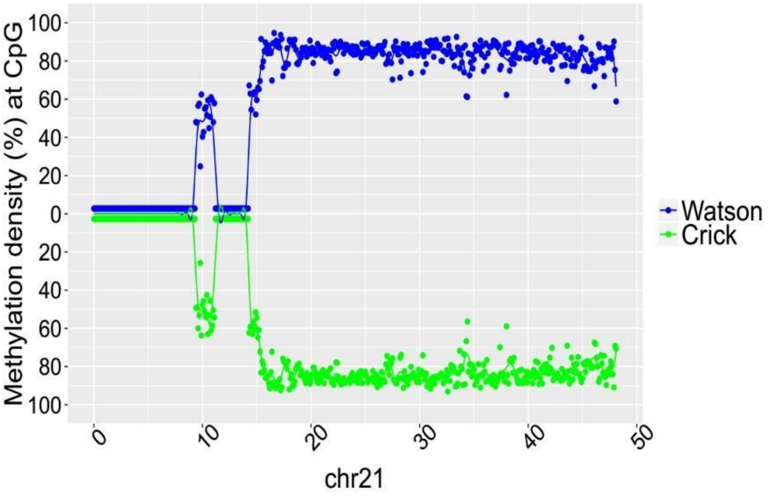
SRR847424 (42 year old female)



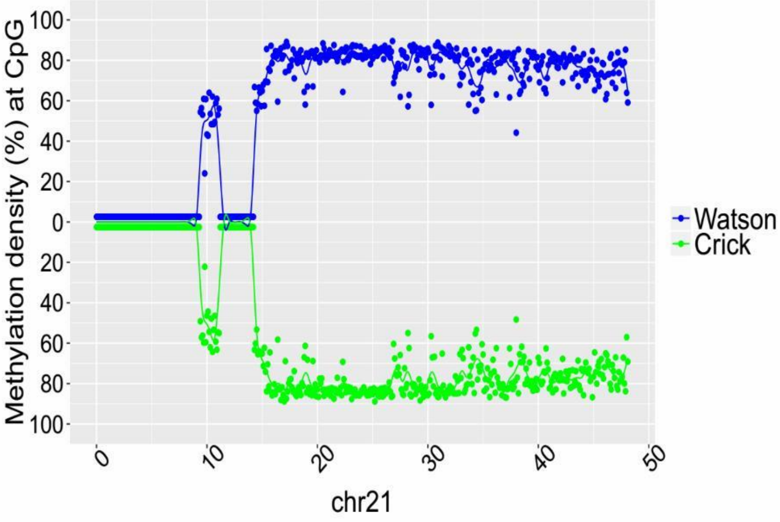
SRR921706 (53 year old female)



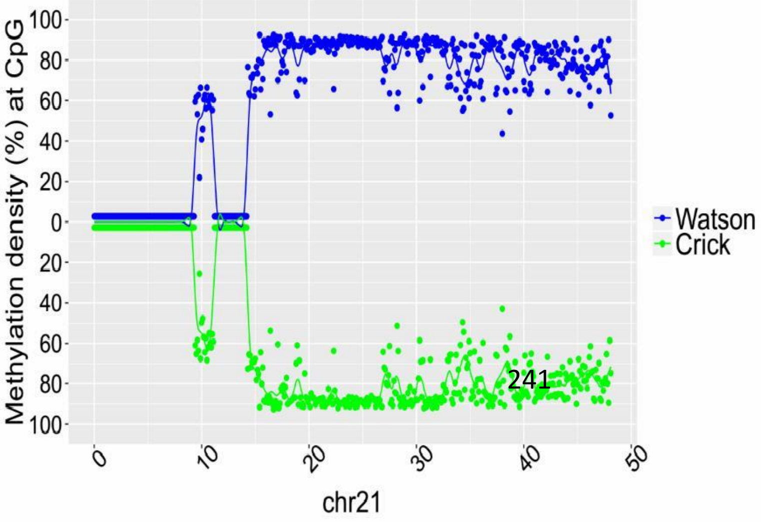
SRR921723 (53 year old female)



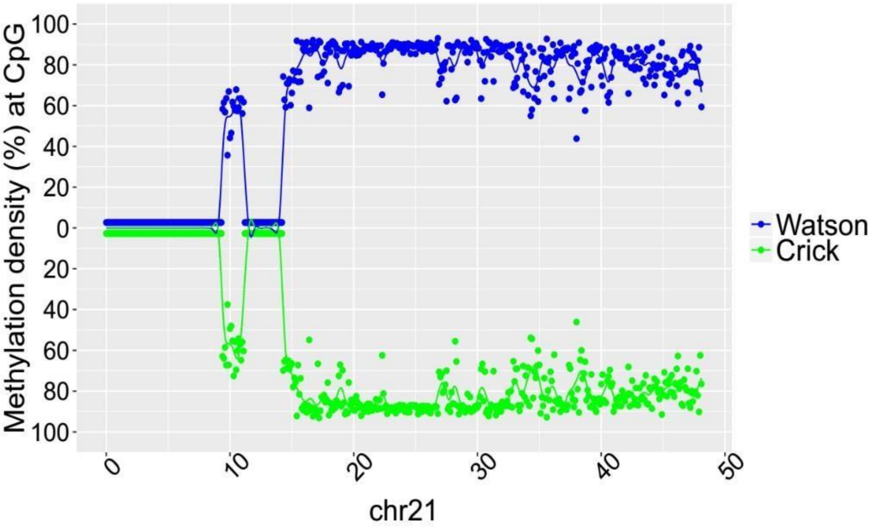
SRR921702 (64 year old female)



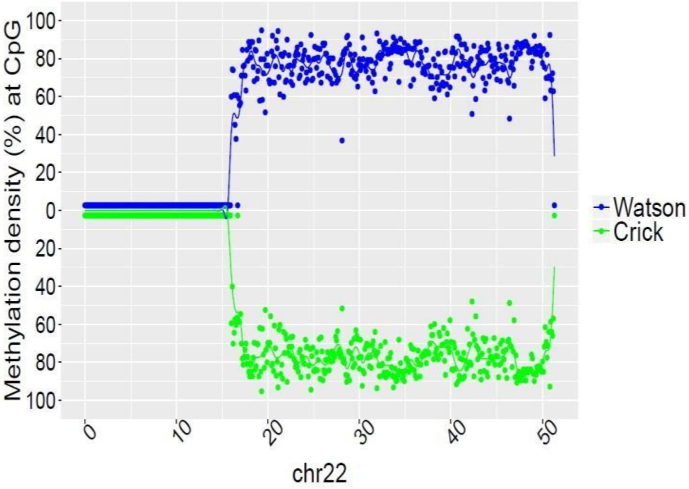
SRR949195 (81 year old female)



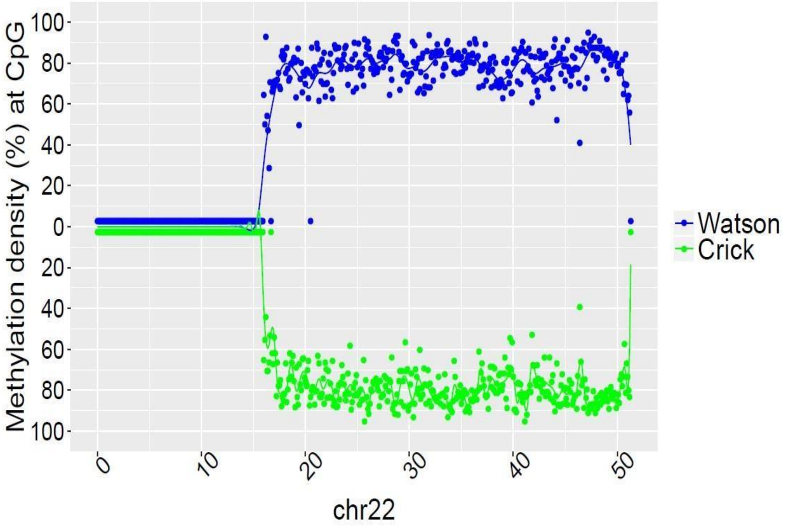
SRR949197 (82 year old female)



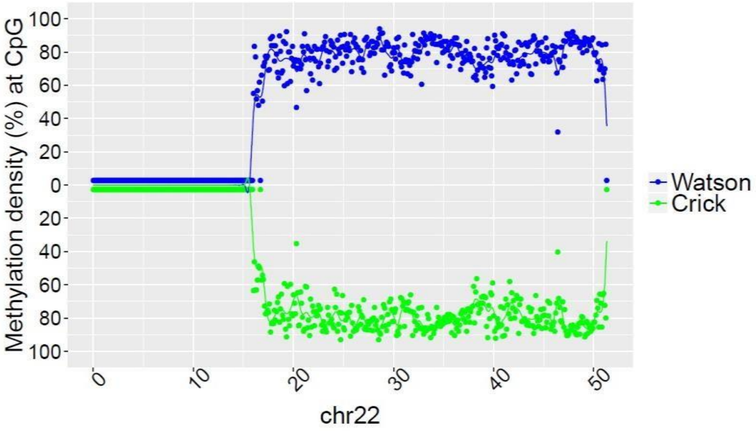
SRR901381 (20 weeks old male)



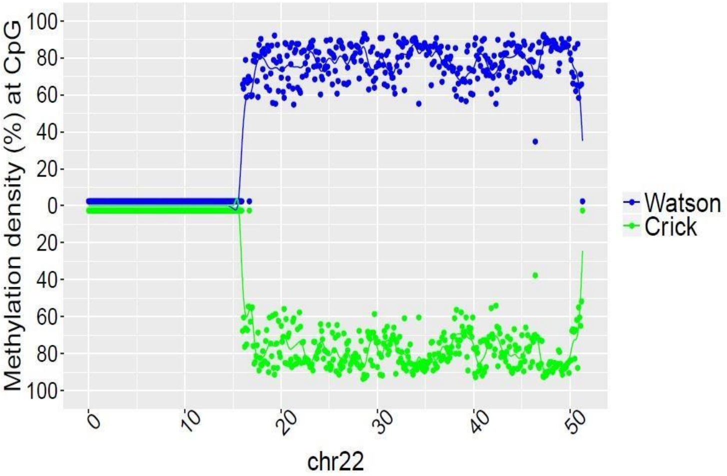
SRR47900 (31 year old male)



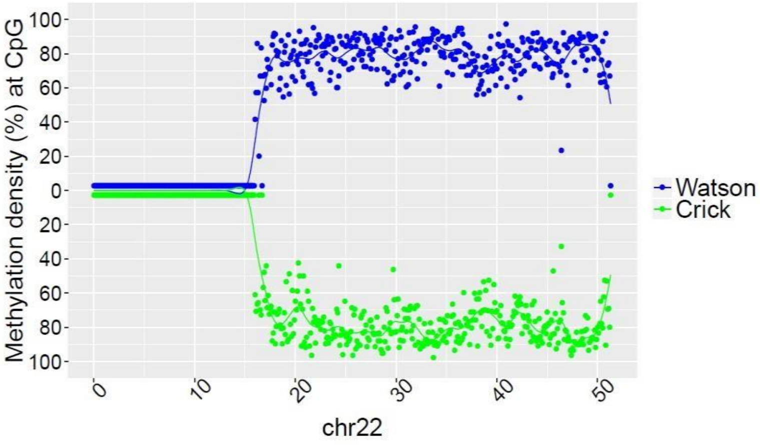
SRR 478994 (47 year old male)



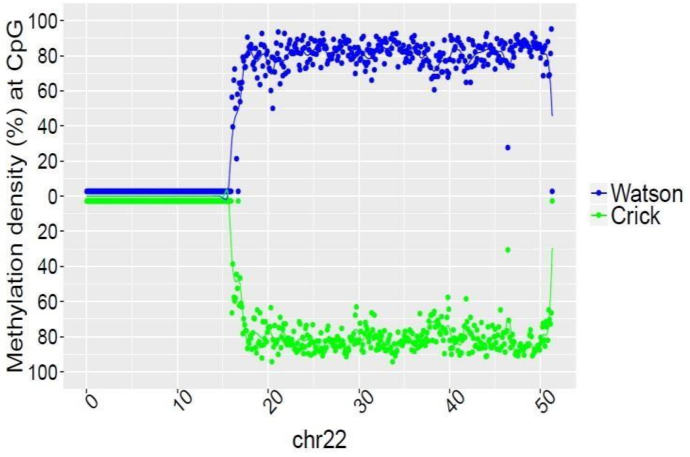
SRR478991 (48 year old male)



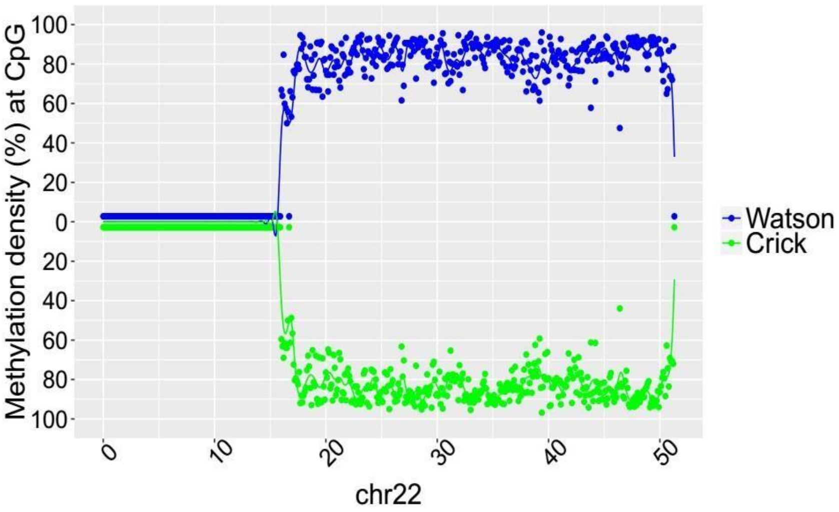
SRR921729 (55 year old male)



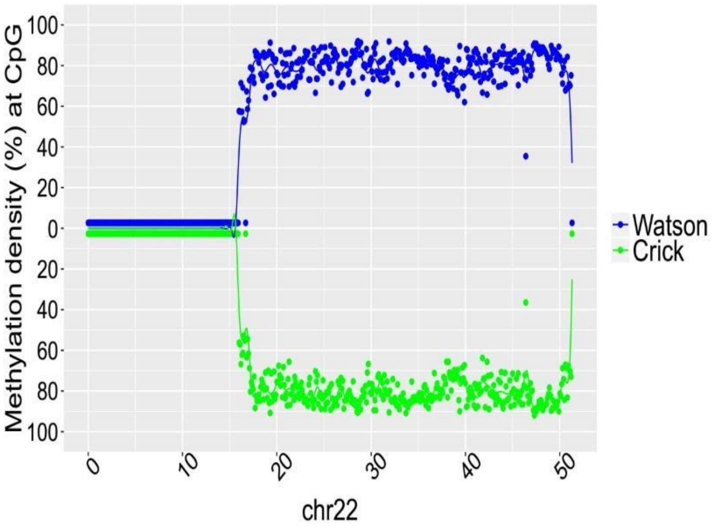
SRR921749 (55 year old male)



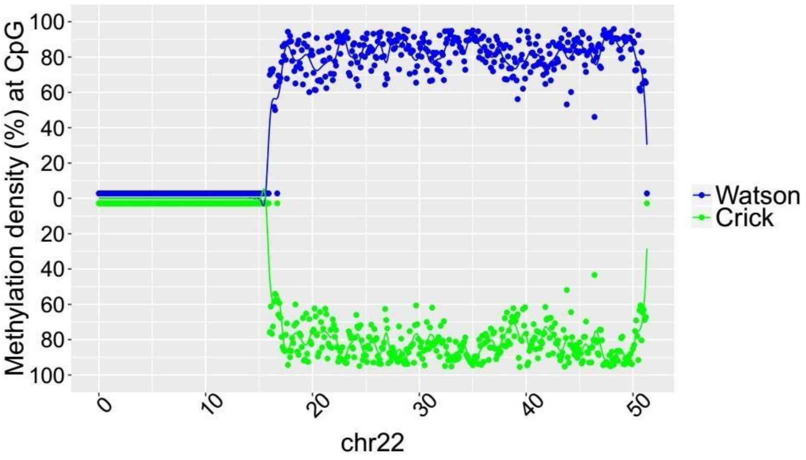
SRR921735 (55 year old male)



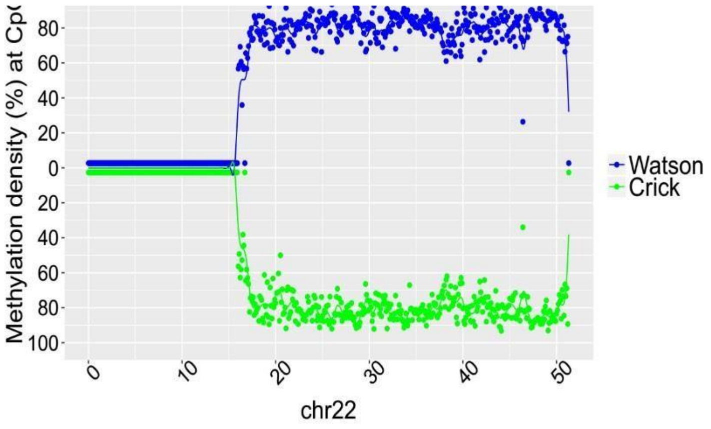
SRR847424 (42 year old female)



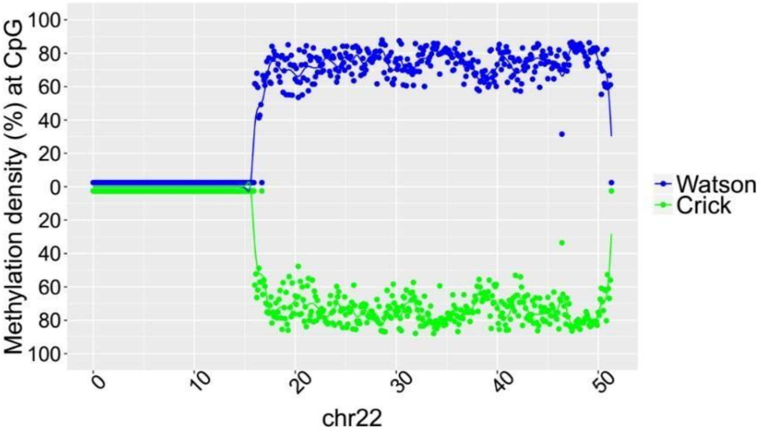
SRR921706 (53 year old female)



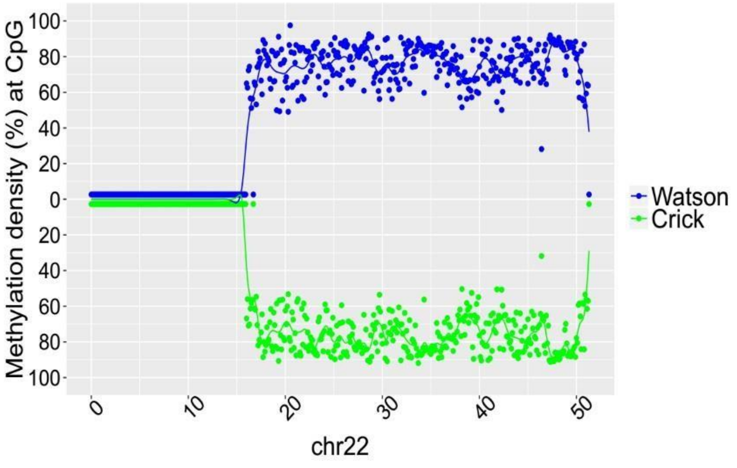
SRR921723 (53 year old female)



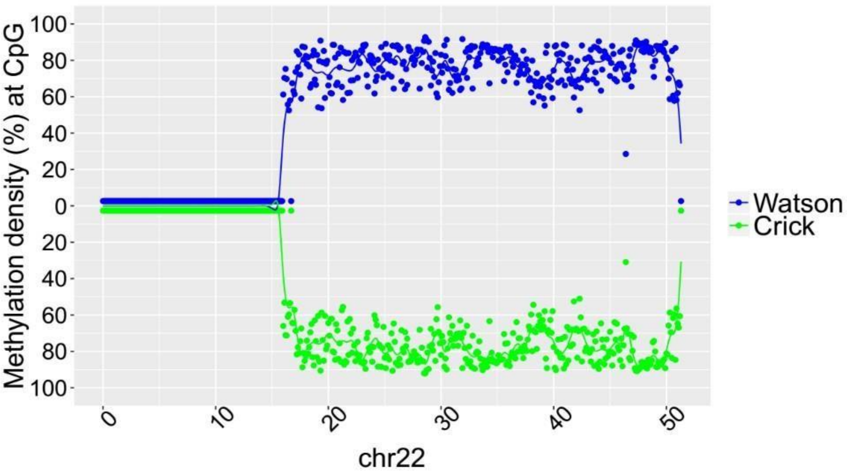
SRR921702 (64 year old female)



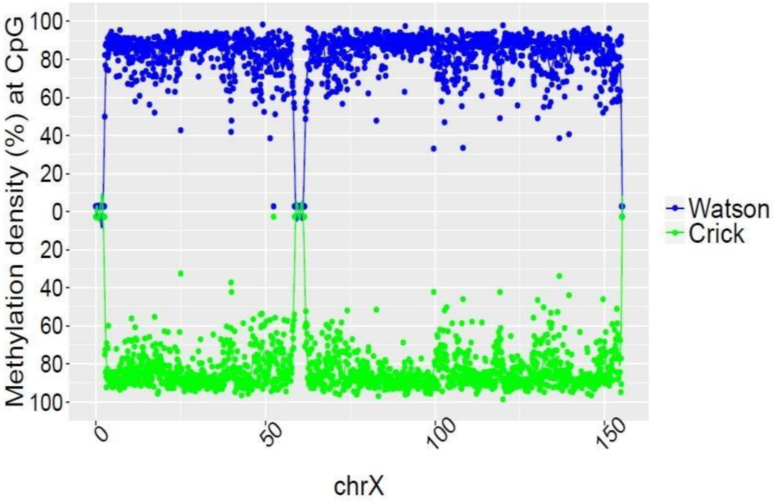
SRR949195 (81 year old female)



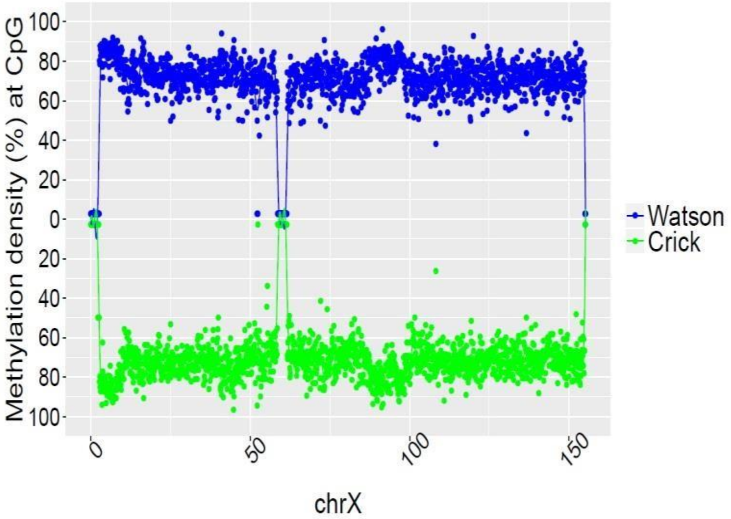
SRR949197 (82 year old female)



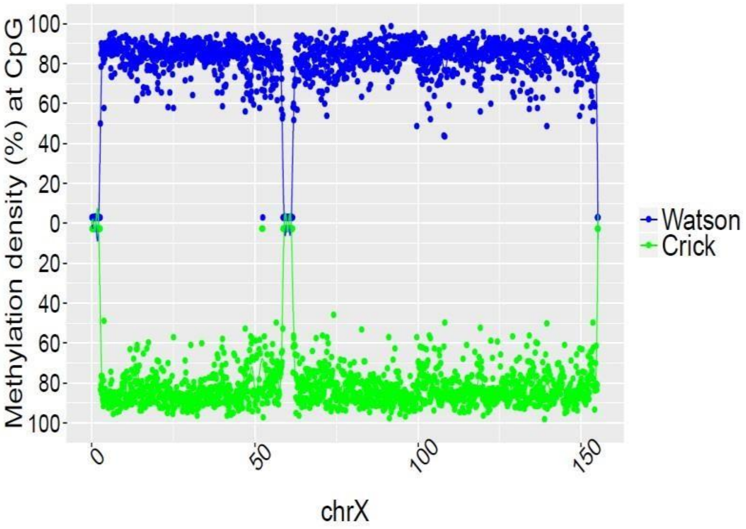
SRR901381 (20 weeks old male)



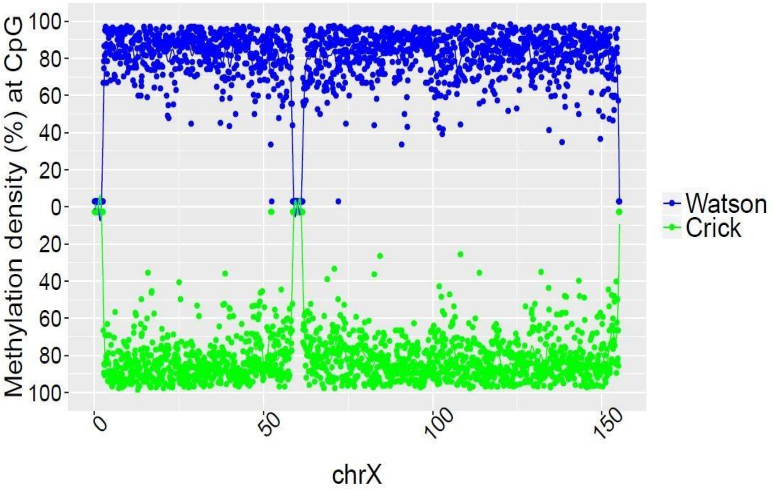
SRR47900 (31 year old male)



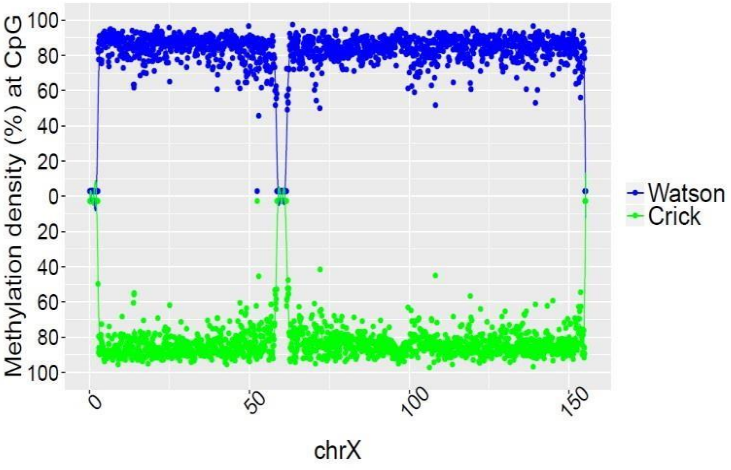
SRR 478994 (47 year old male)



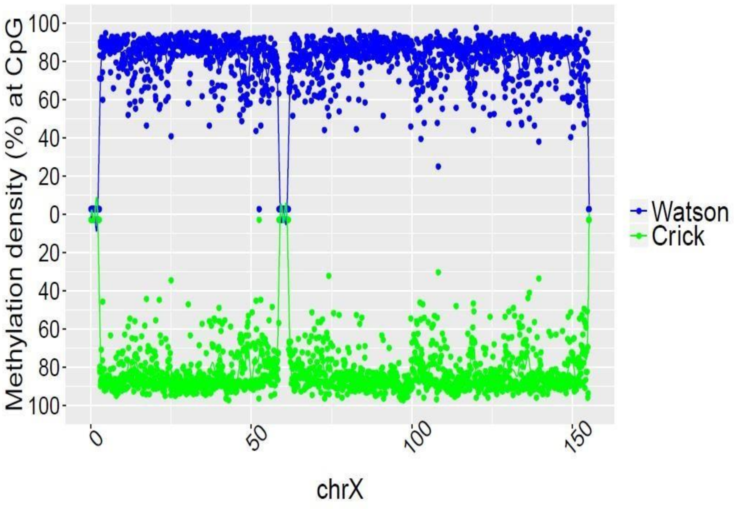
SRR478991 (48 year old male)



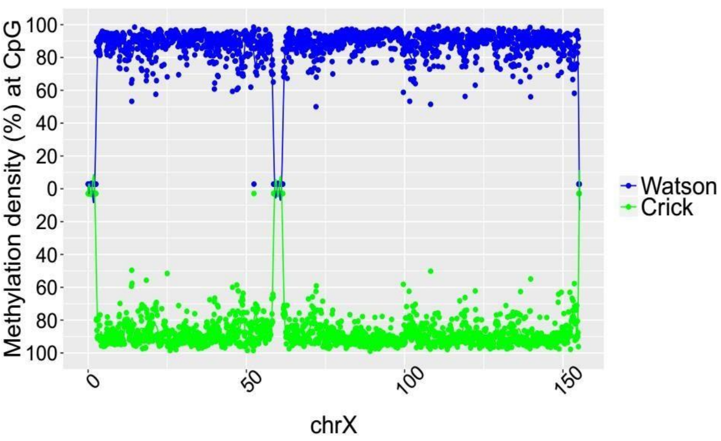
SRR921729 (55 year old male)



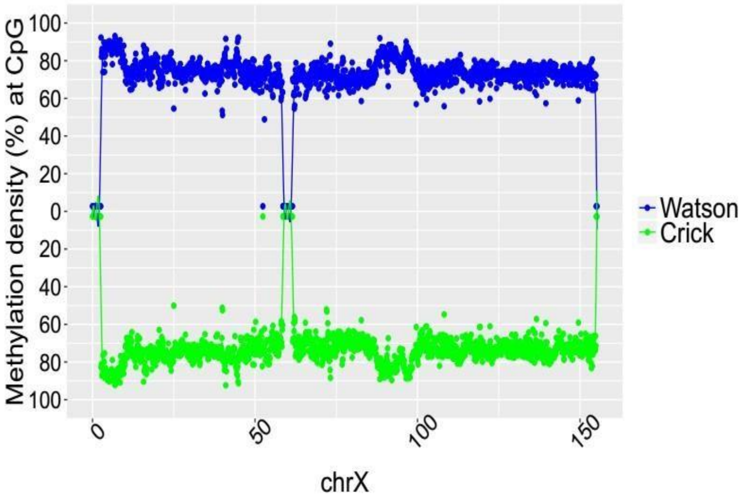
SRR921749 (55 year old male)



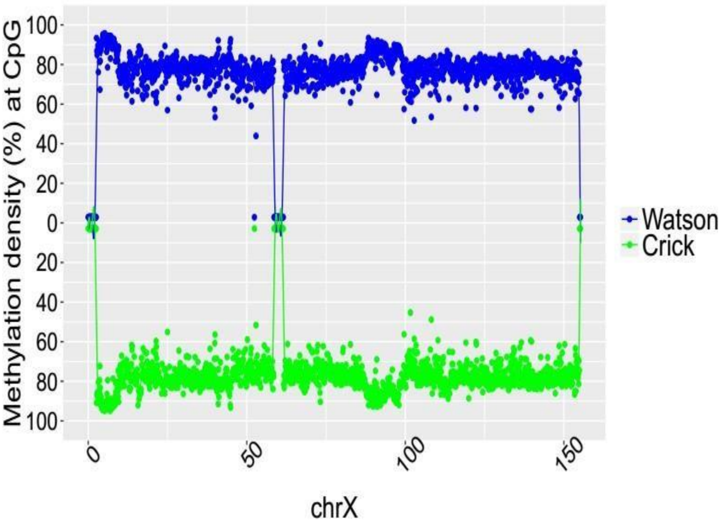
SRR921735 (55 year old male)



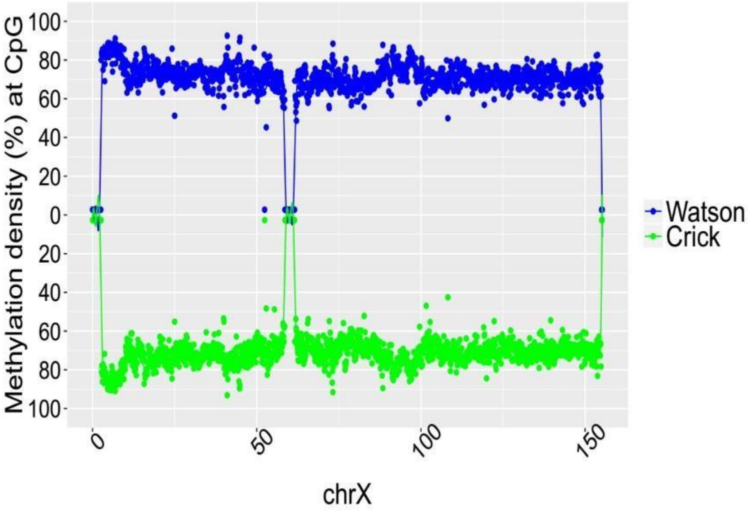
SRR847424 (42 year old female)



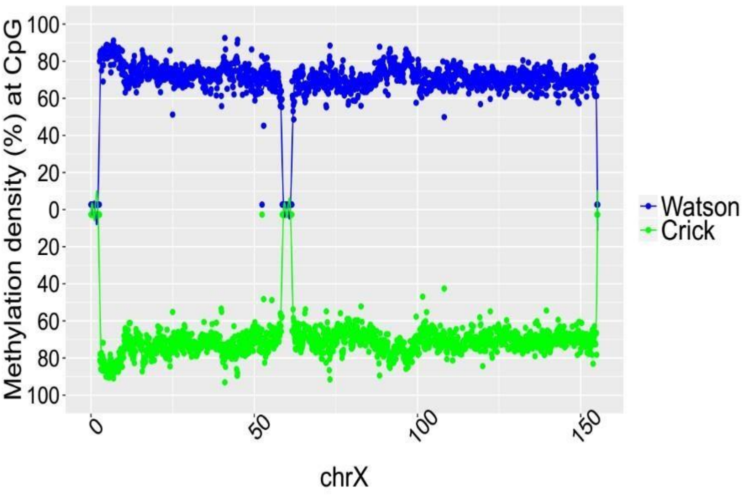
SRR921706 (53 year old female)



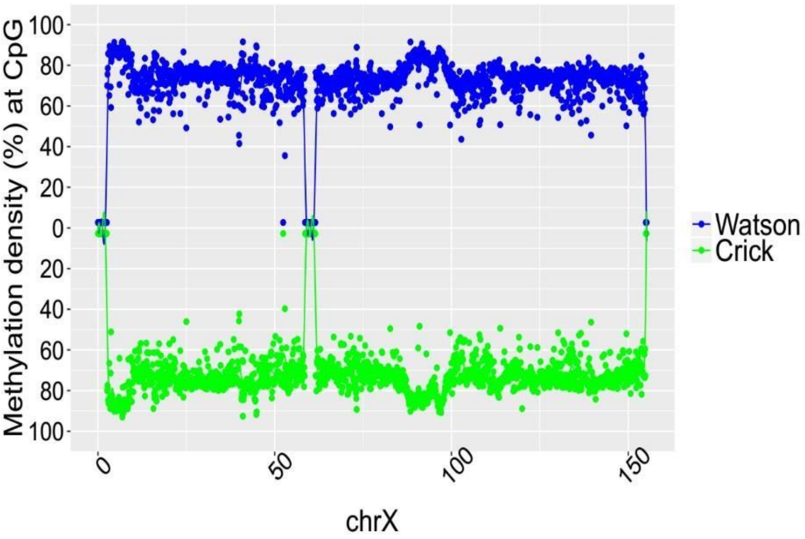
SRR921723 (53 year old female)



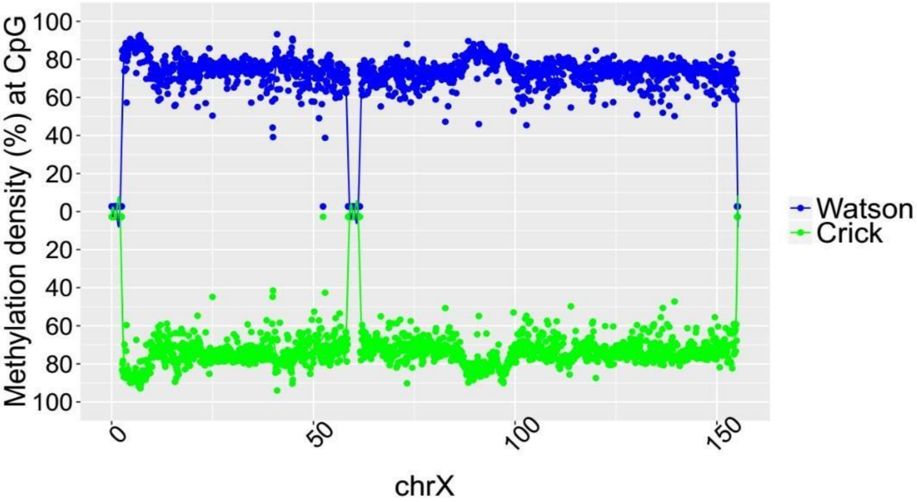
SRR921702 (64 year old female)



SRR949195 (81 year old female)



SRR949197 (82 year old female)

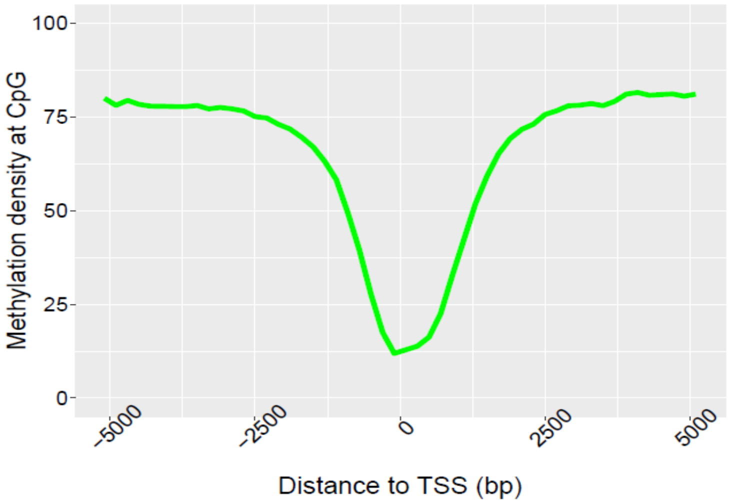


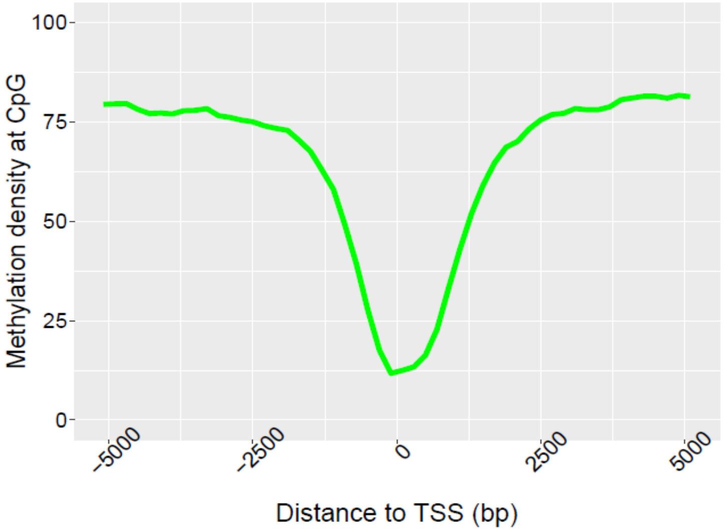
Methylation density (%) at CpG
100
80
60
40
20
0
20
40
60
80
100
0
50
100
150
chrX
Watson
Crick

METHY PIPE OUTPUT OF BLOOD SAMPLES

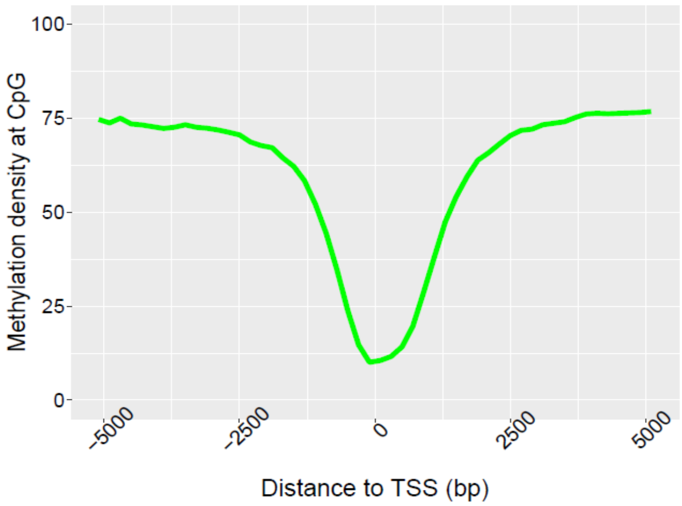
Methylation density from CpG sites around TSS regions with 200 bp bins in Watson (left) and crick (right) strands

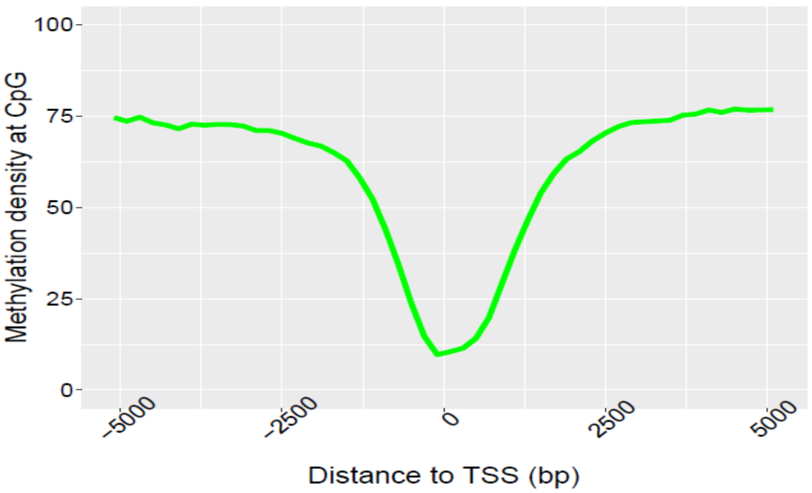
SRR330578 (1 year old male)



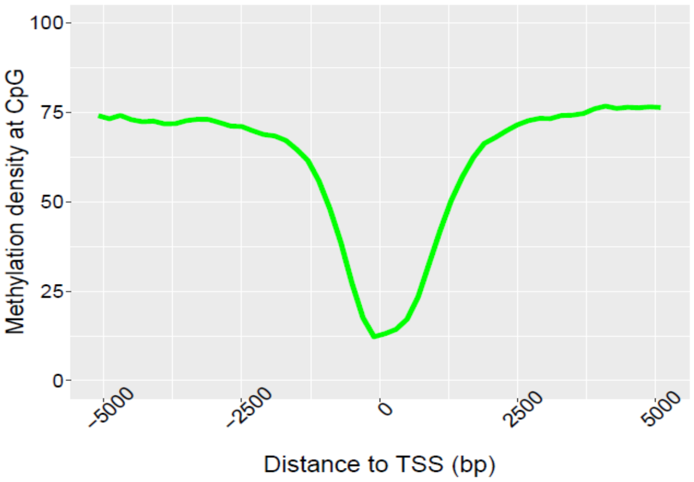


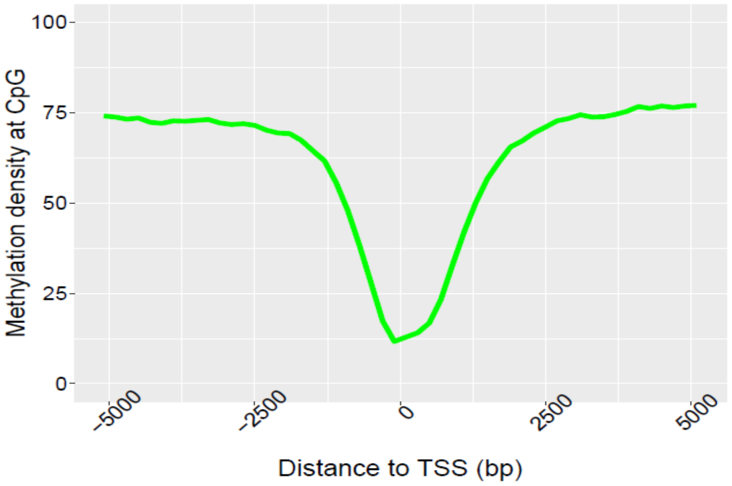
SRR389249 (26 year old male)





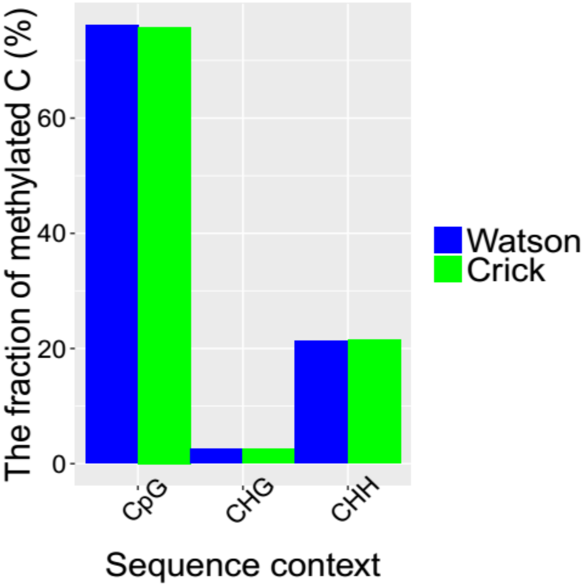
SRR330576 (103 year old male)



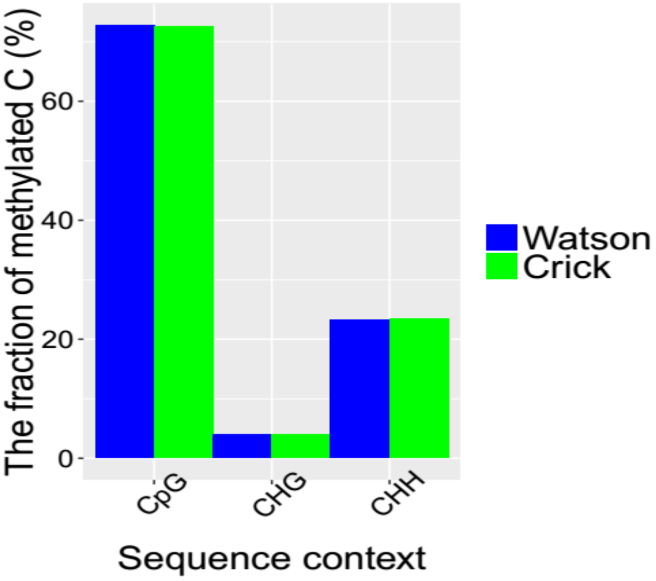


The fraction of methylated C according to the sequence context

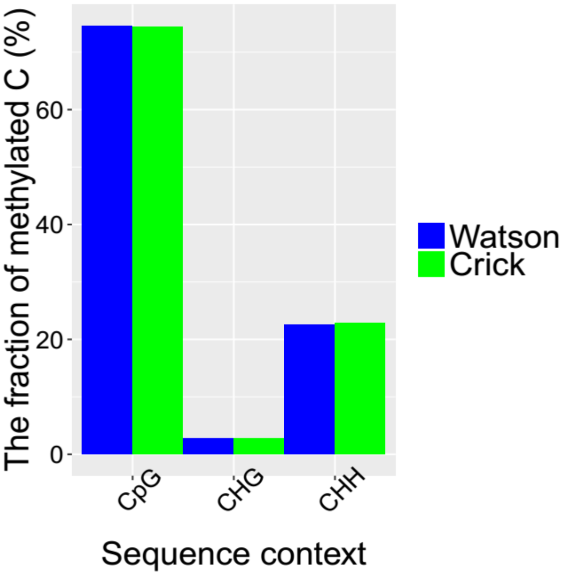
SRR330578 (1 year old male)



SRR389249 (26 year old male)

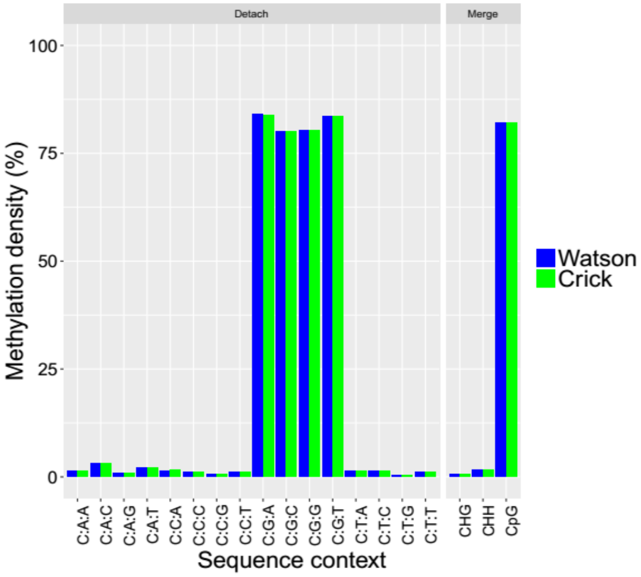


SRR330576 (103 year old male)

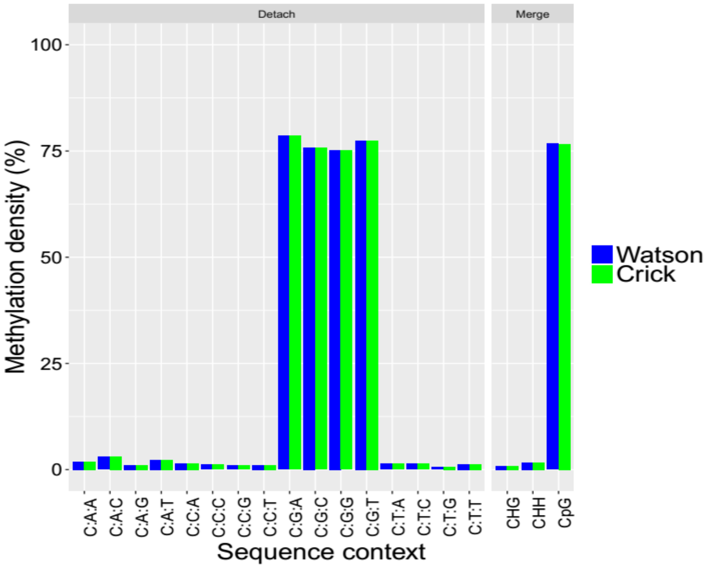


Methylation Density According to the Sequence Context

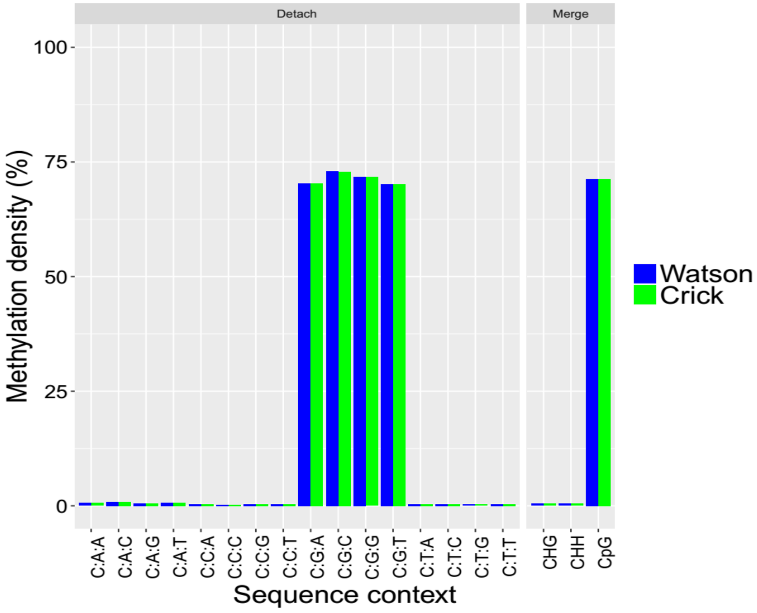
SRR330578 (1 year old male)



SRR389249 (26 year old male)

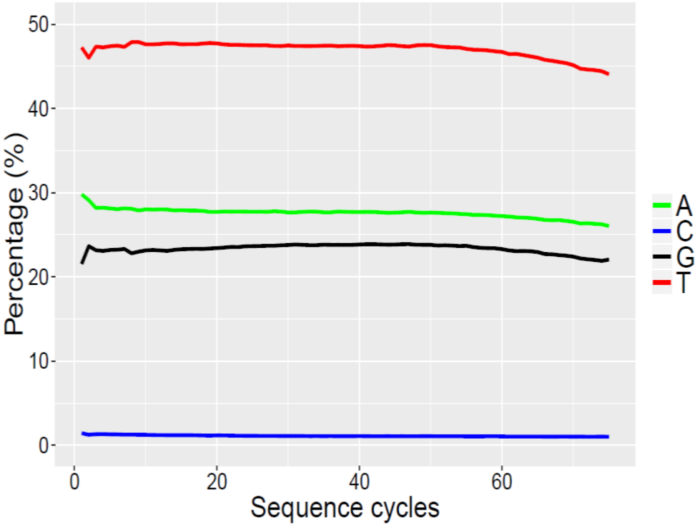


SRR330576 (103 year old male)

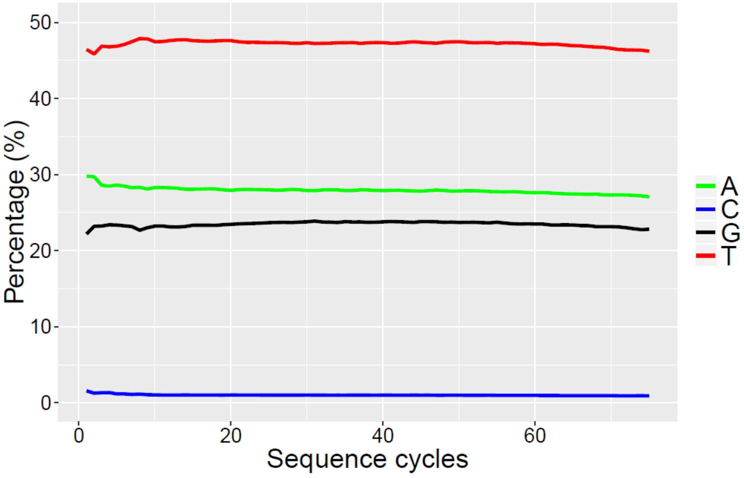


Base content percentage across sequence cycles

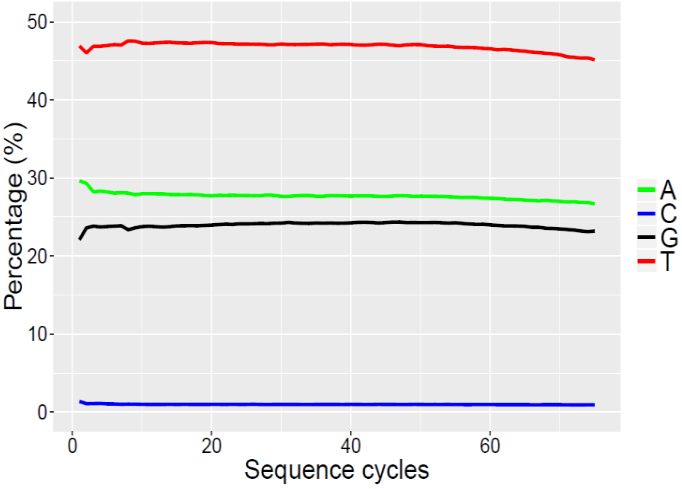
SRR330578 (1 year old male)



SRR389249 (26 year old male)

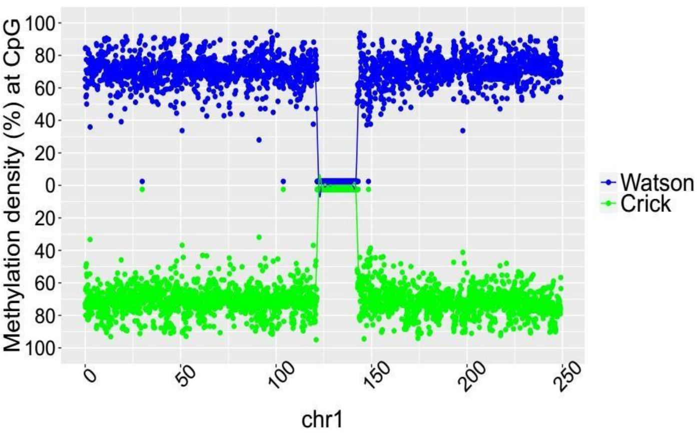


SRR330576 (103 year old male)

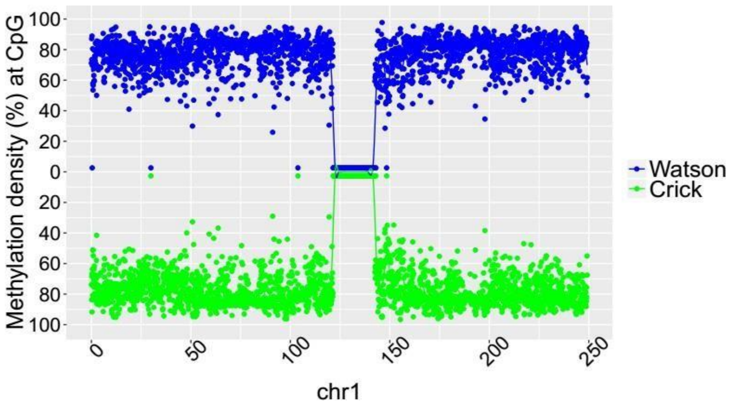


METHYLATION DENSITY ACROSS CHROMOSSOMES (distance in mega pbs)

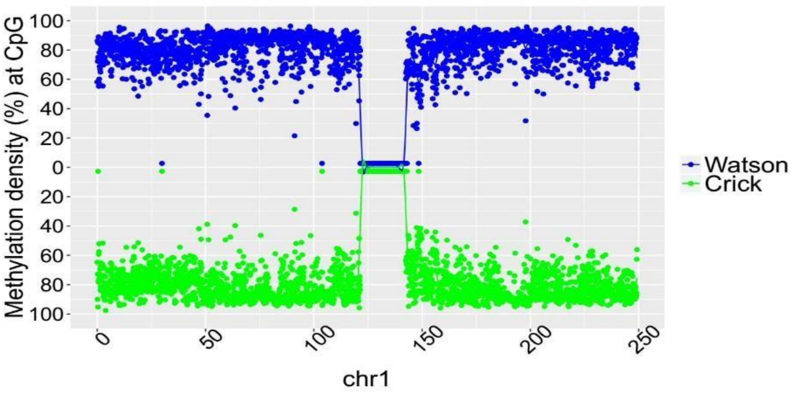
SRR330578 (1 year old male)



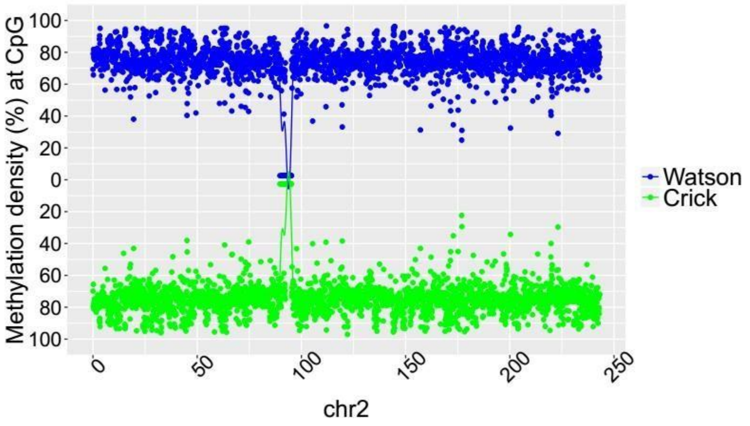
SRR389249 (26 year old male)



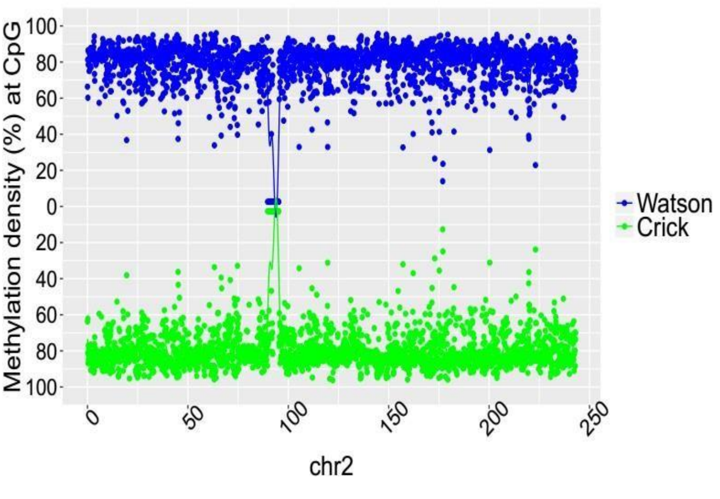
SRR330576 (103 year old male)



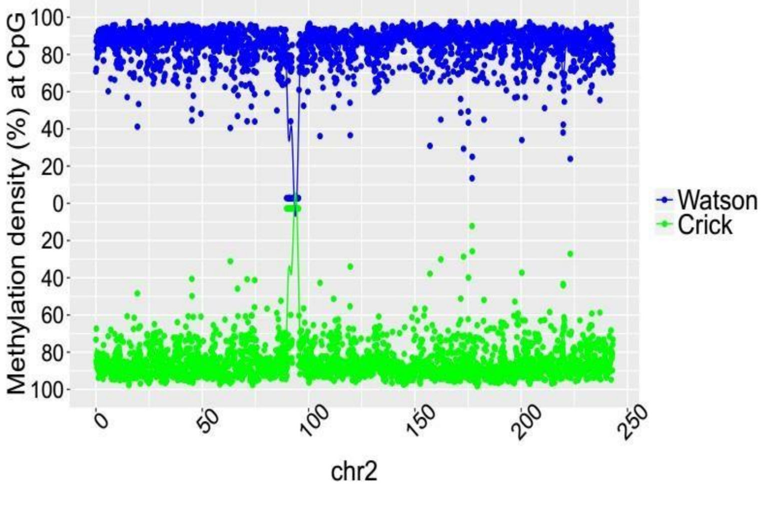
SRR330578 (1 year old male)



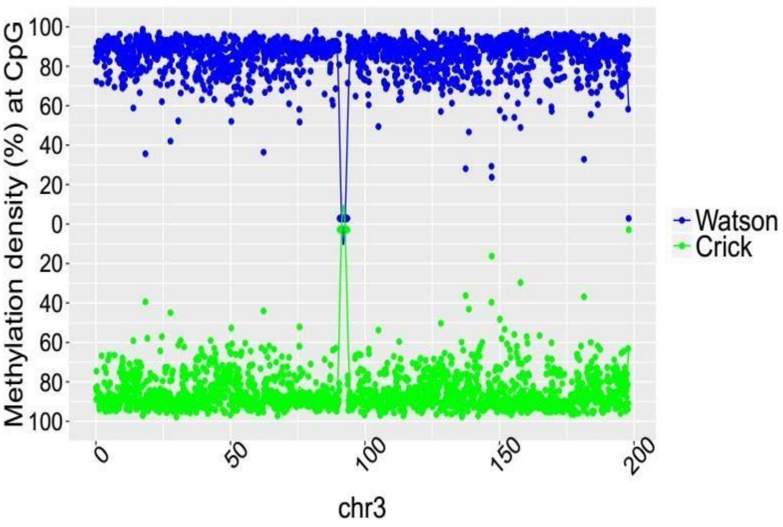
SRR389249 (26 year old male)



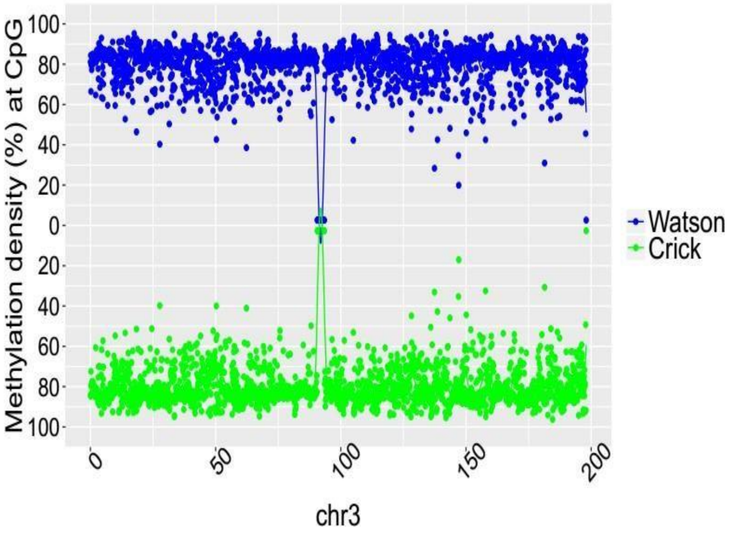
SRR330576 (103 year old male)



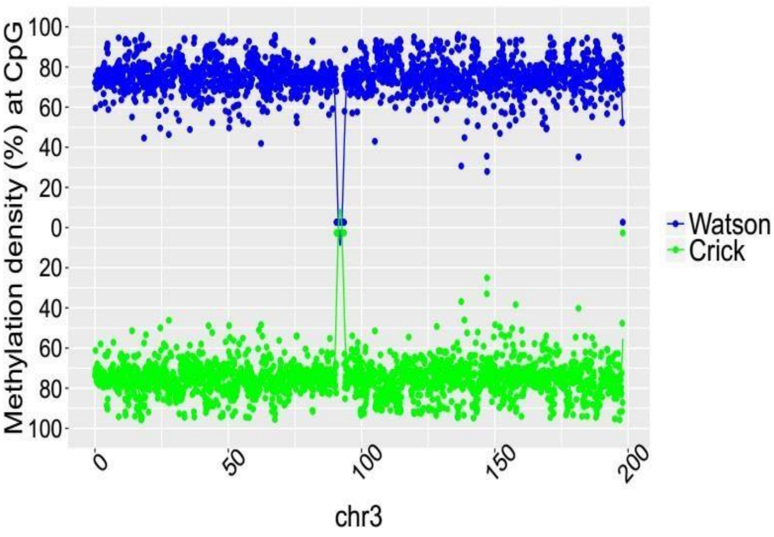
SRR330578 (1 year old male)



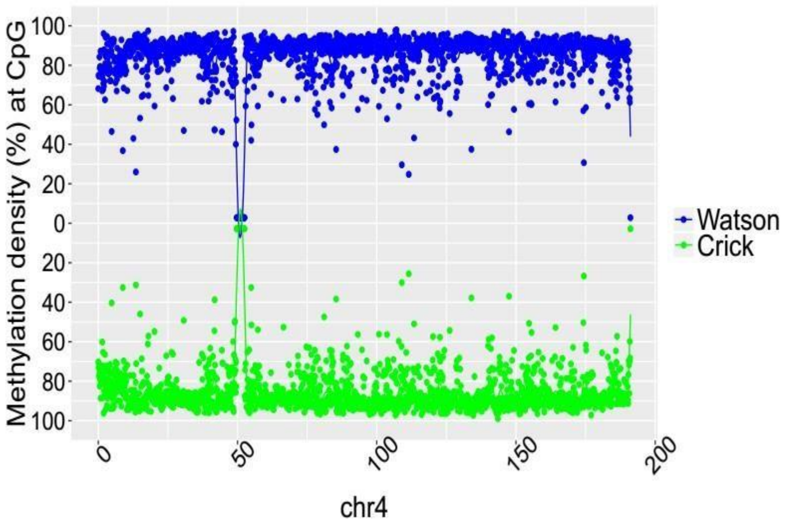
SRR389249 (26 year old male)



SRR330576 (103 year old male)

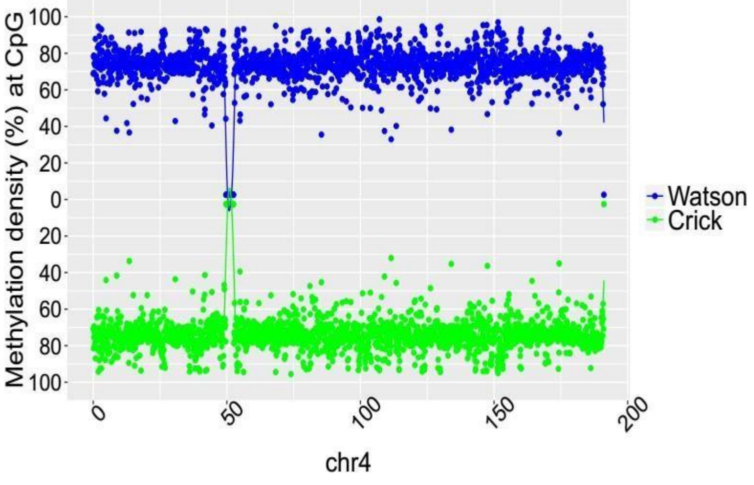


SRR330578 (1 year old male)



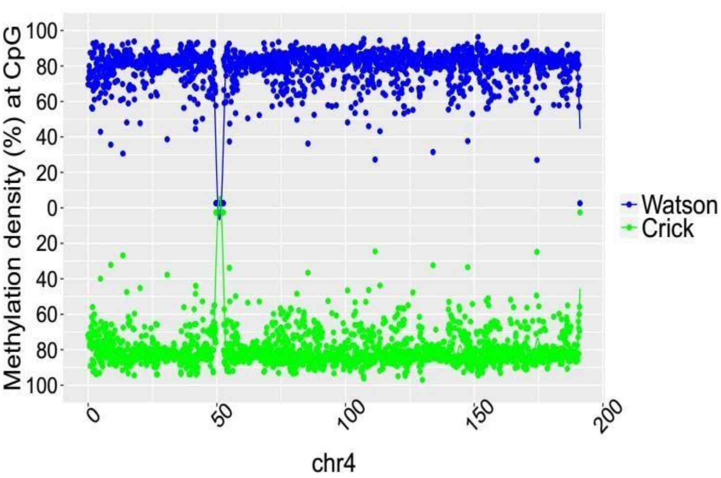
SRR389249 (26 year old male)

SRR330576 (103 year old male)

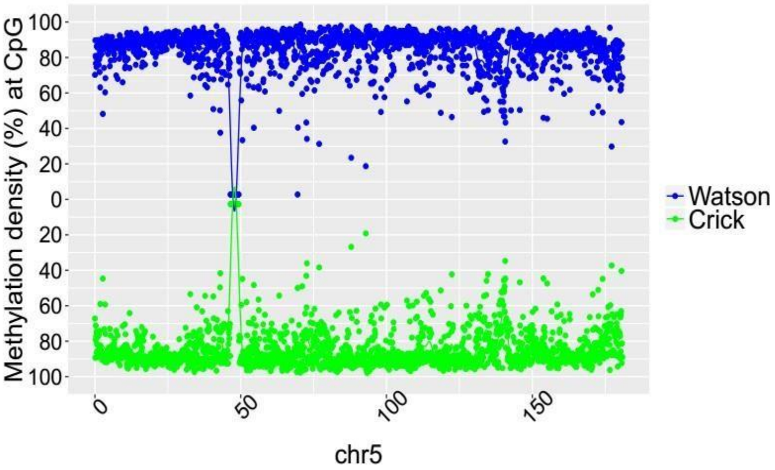


SRR330578 (1 year old male)

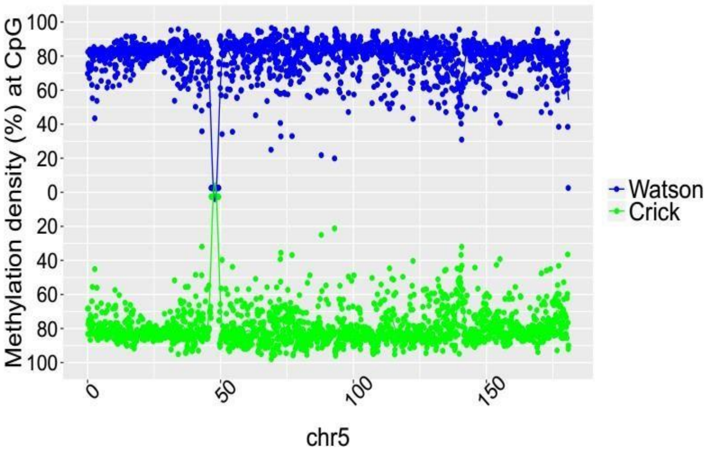
SRR389249 (26 year old male)



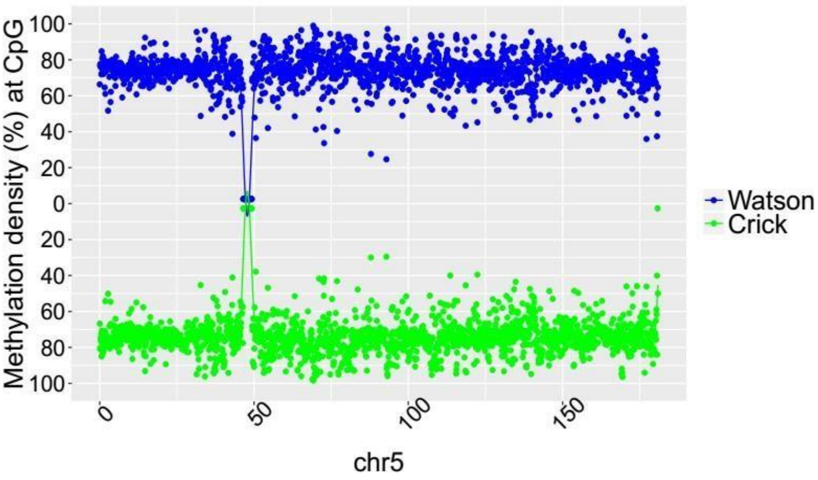
SRR330578 (1 year old male)



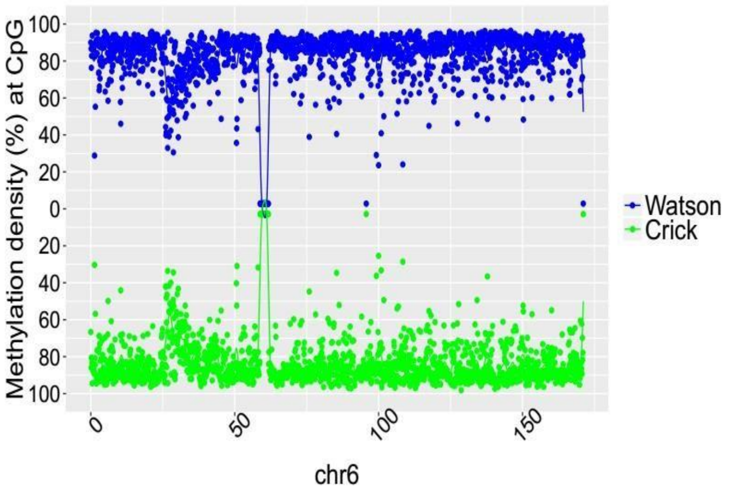
SRR389249 (26 year old male)



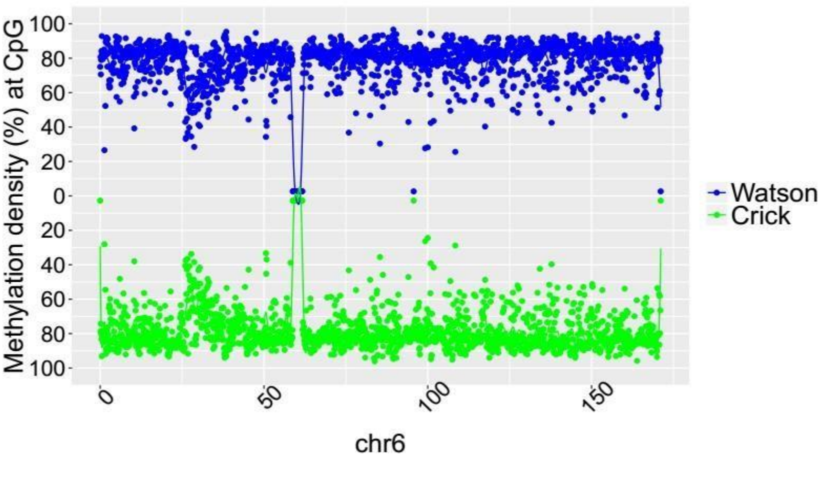
SRR330576 (103 year old male)



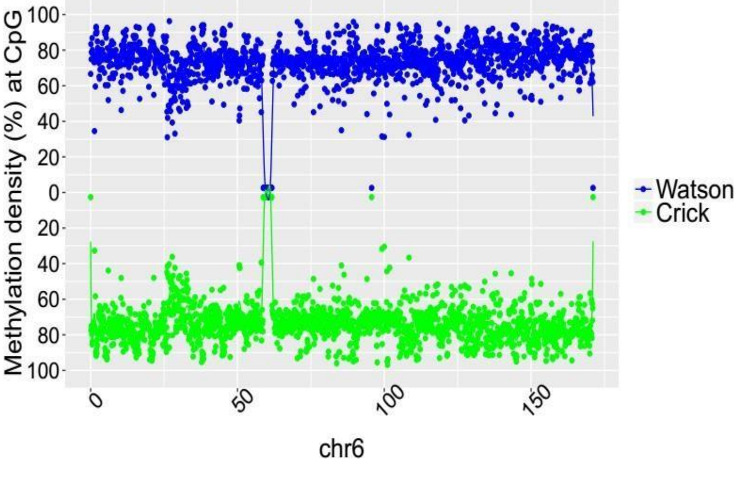
SRR330578 (1 year old male)



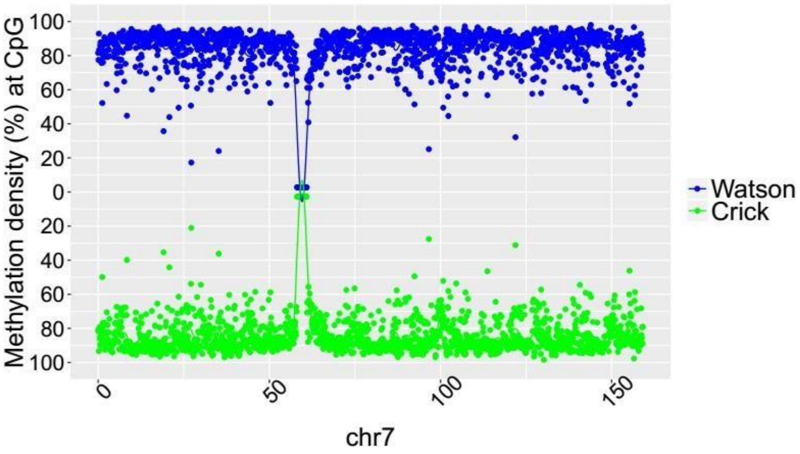
SRR389249 (26 year old male)



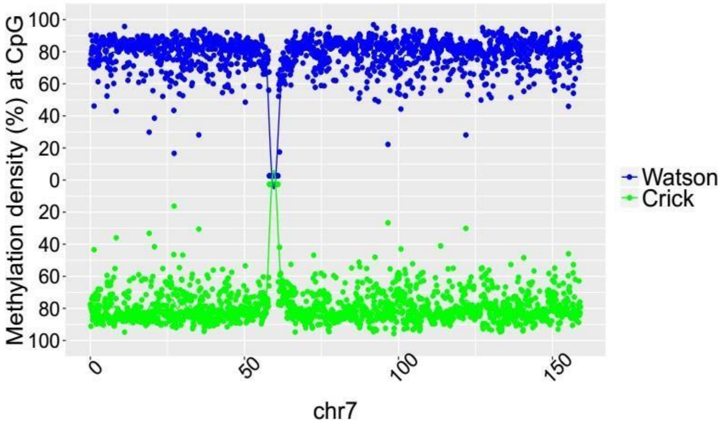
SRR330576 (103 year old male)



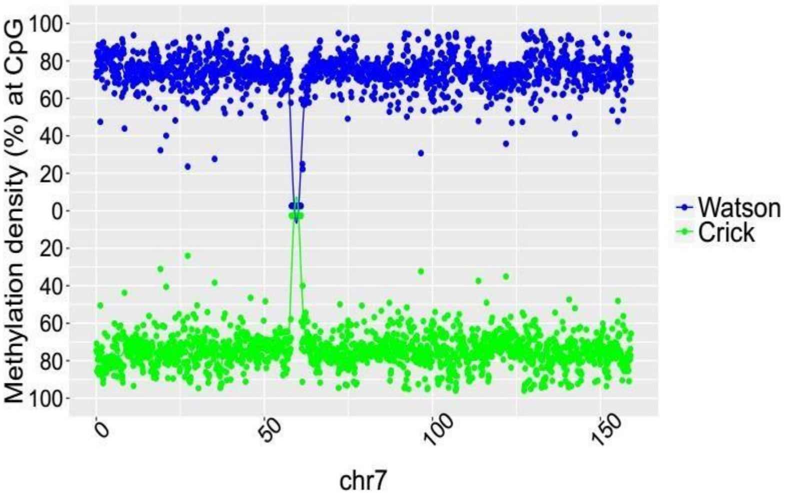
SRR330578 (1 year old male)



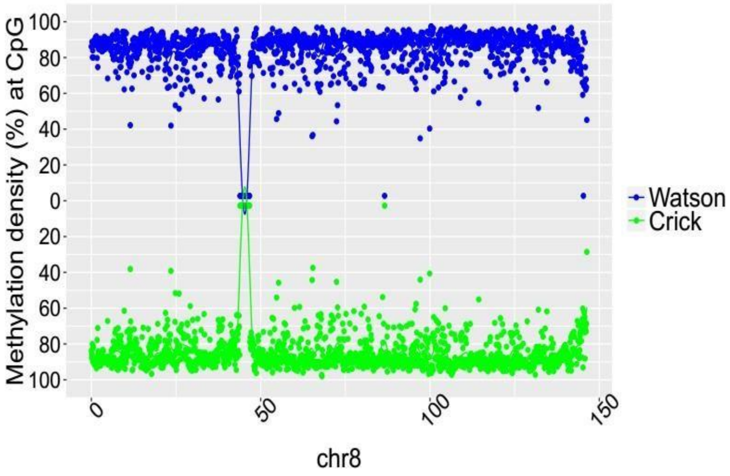
SRR389249 (26 year old male)



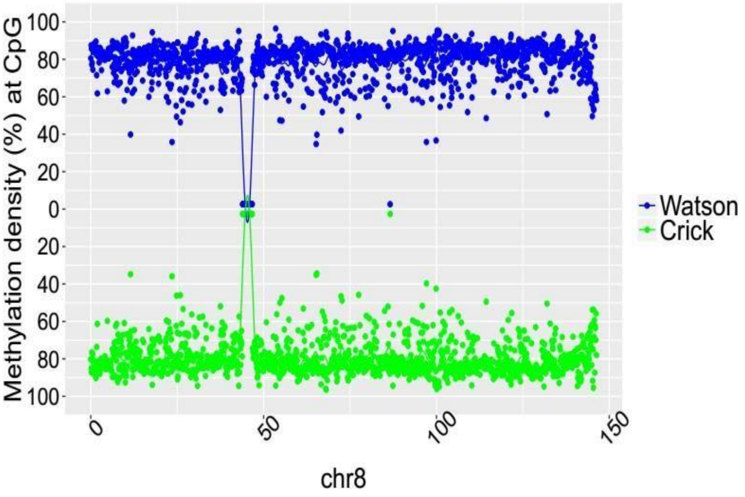
SRR330576 (103 year old male)



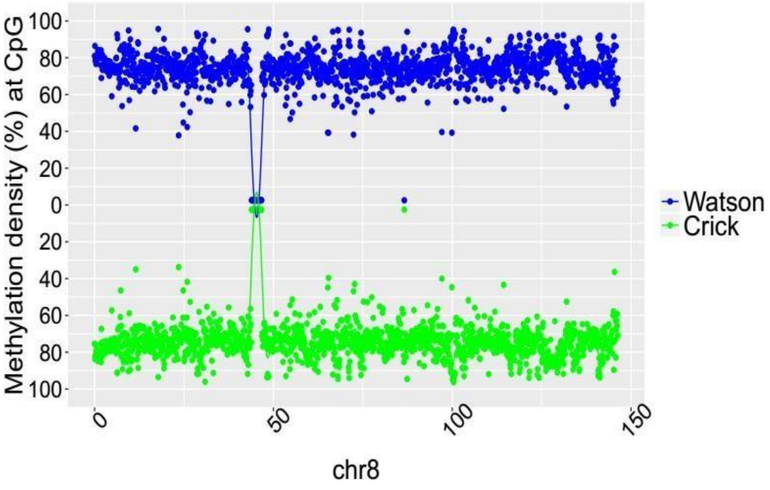
SRR330578 (1 year old male)



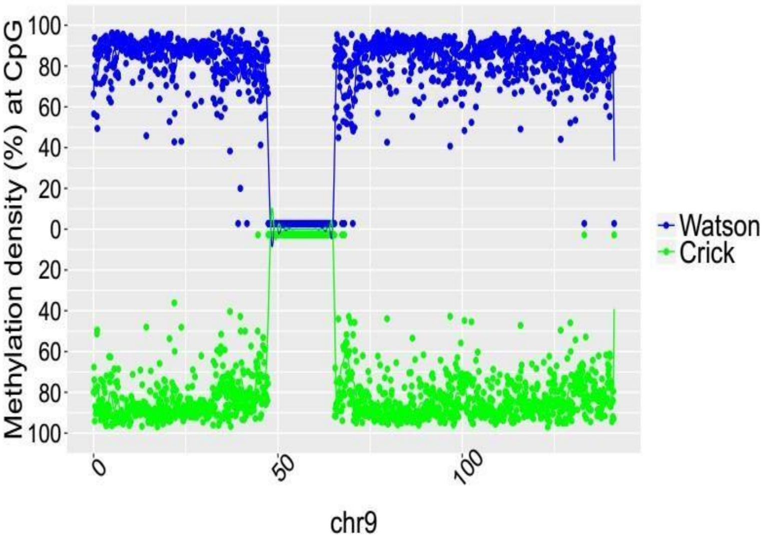
SRR389249 (26 year old male)



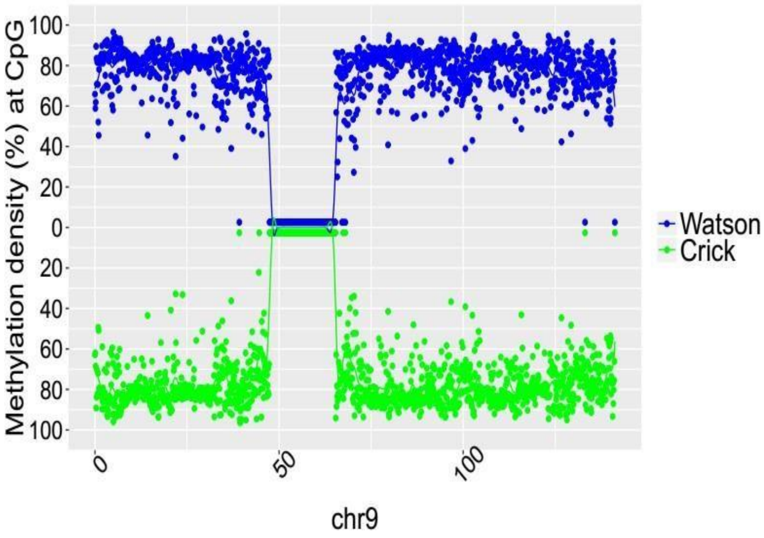
SRR330576 (103 year old male)



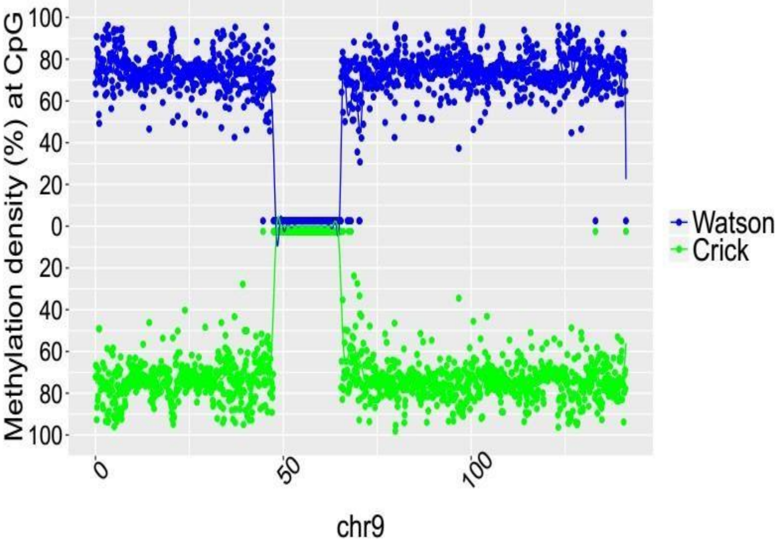
SRR330578 (1 year old male)



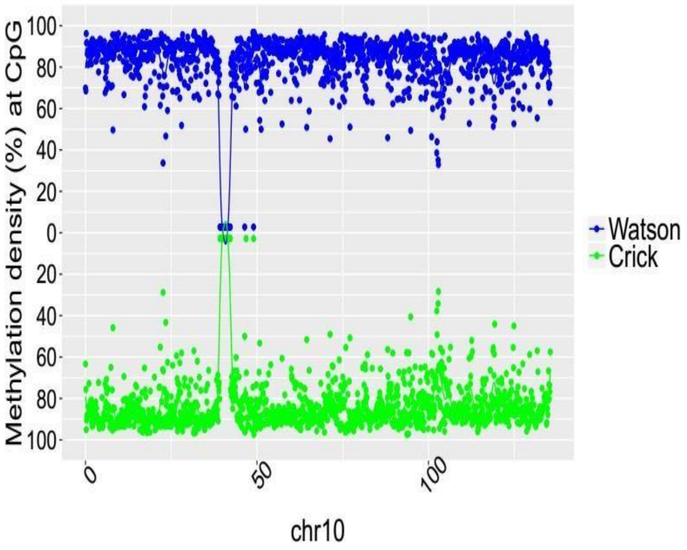
SRR389249 (26 year old male)



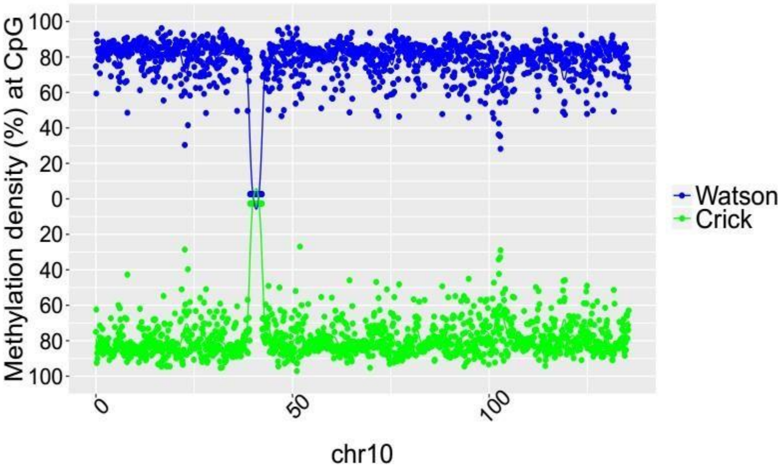
SRR330576 (103 year old male)



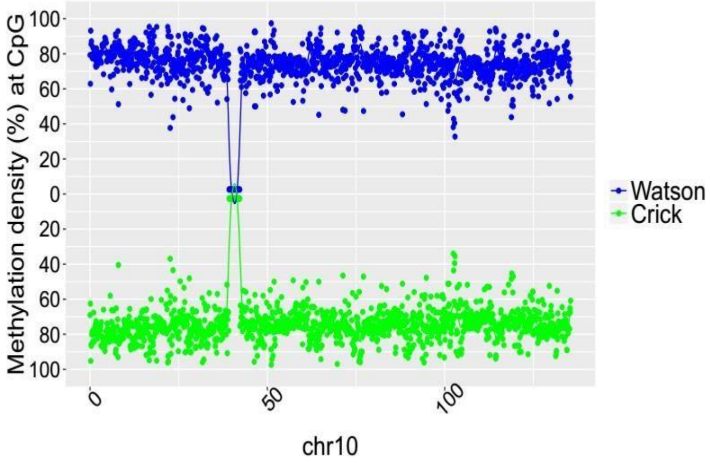
SRR330578 (1 year old male)



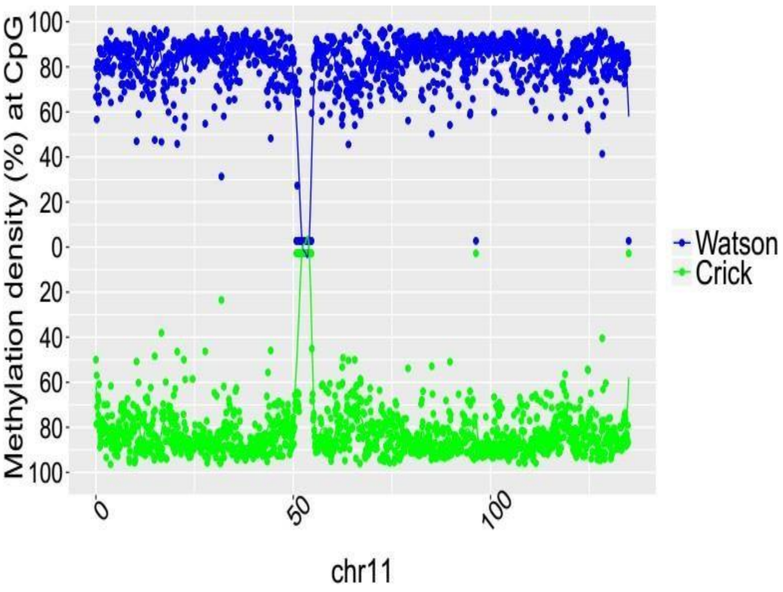
SRR389249 (26 year old male)



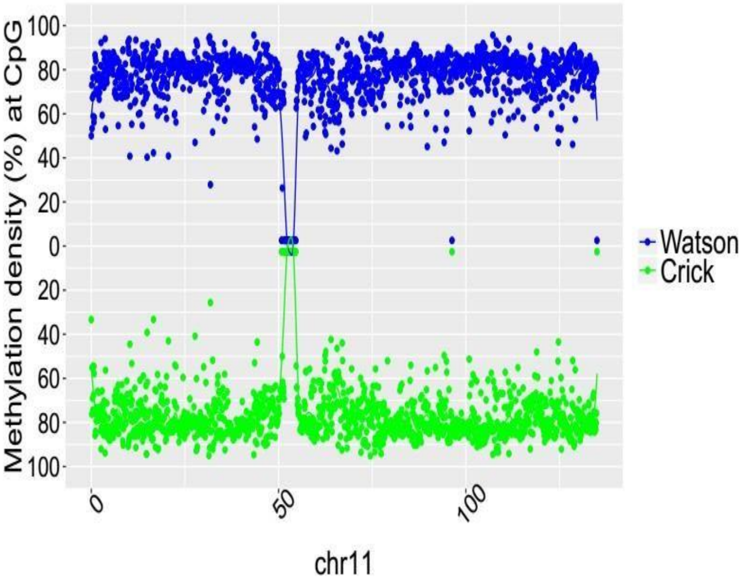
SRR330576 (103 year old male)



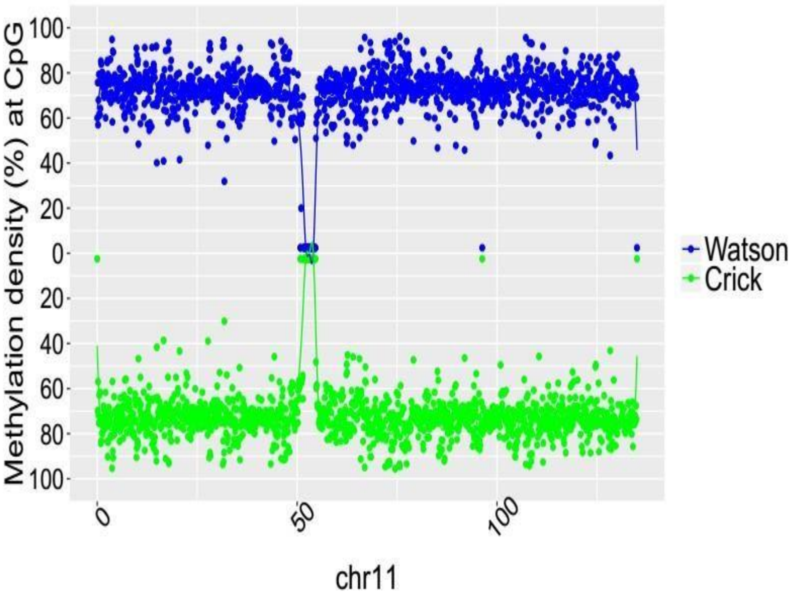
SRR330578 (1 year old male)



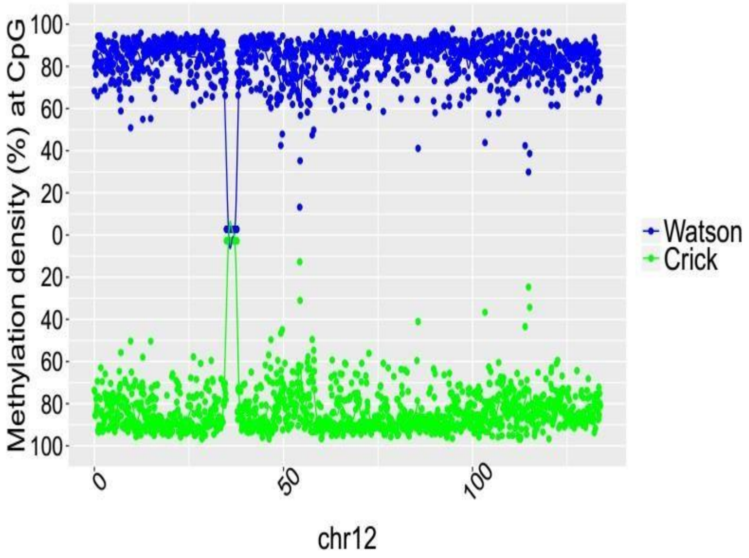
SRR389249 (26 year old male)



SRR330576 (103 year old male)

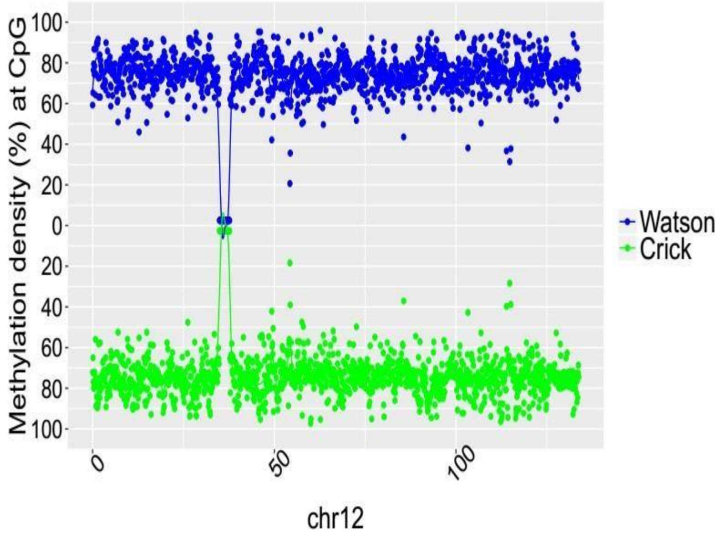


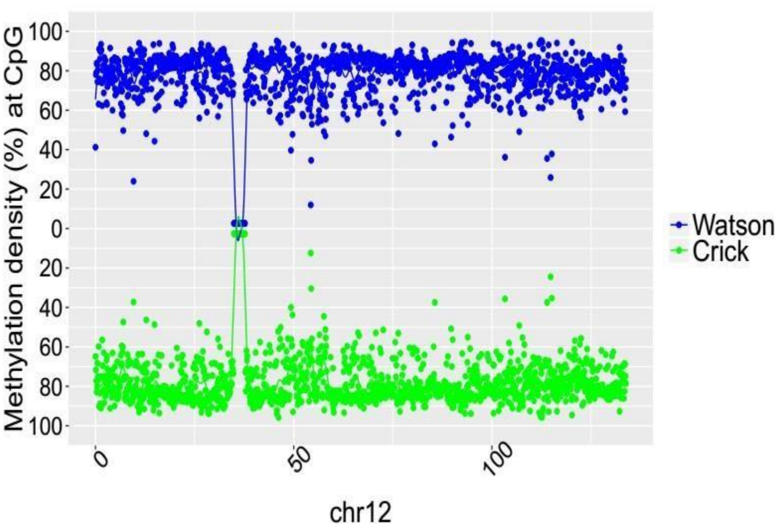
SRR330578 (1 year old male)



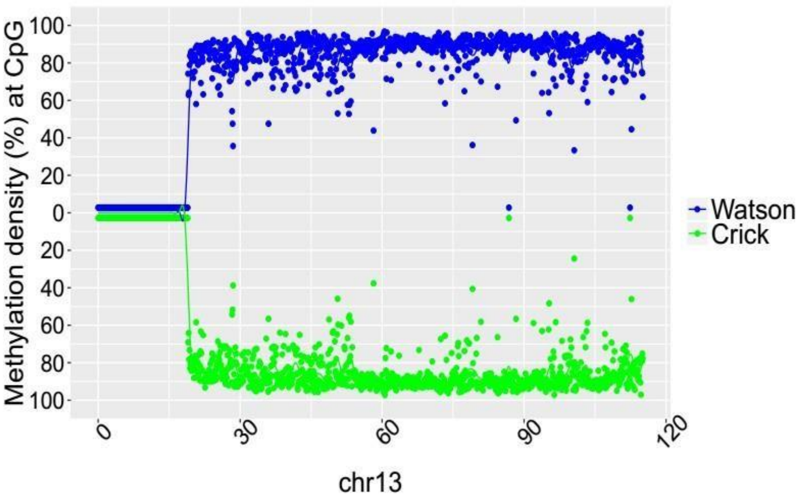
SRR389249 (26 year old male)

SRR330576 (103 year old male)

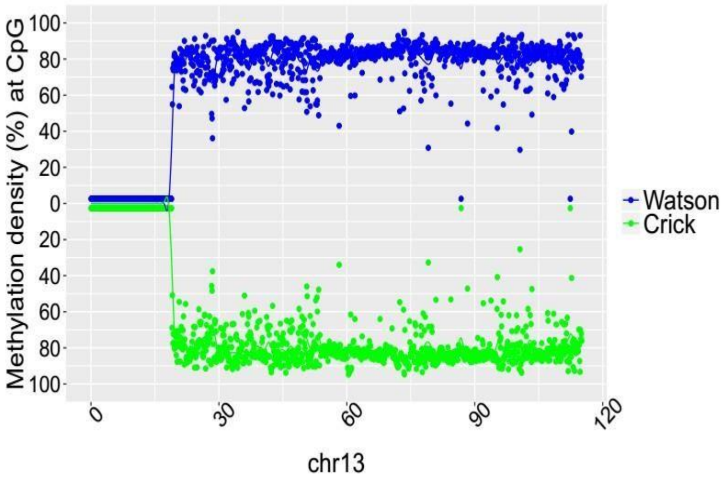



Methylation density (%) at CpG
100
80
60
40
20
0
20
40
60
80
100
0
50
100
chr12
Watson
Crick

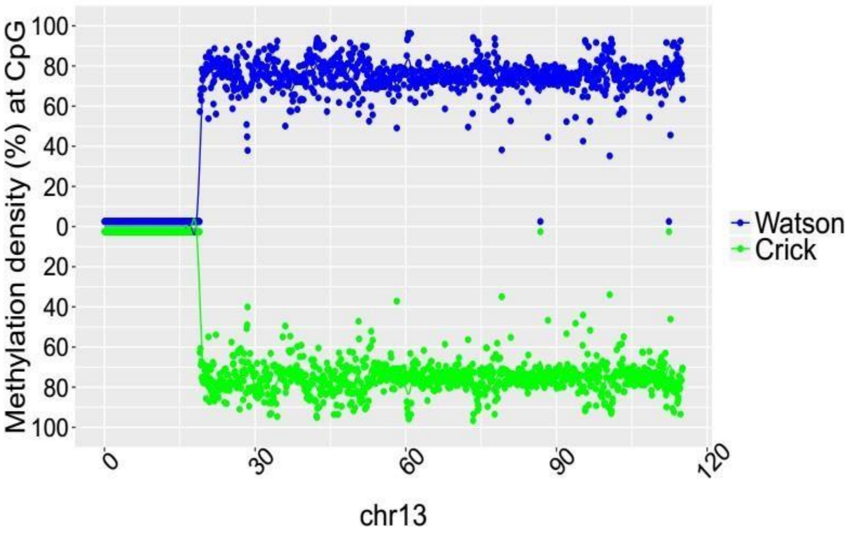
SRR330578 (1 year old male)



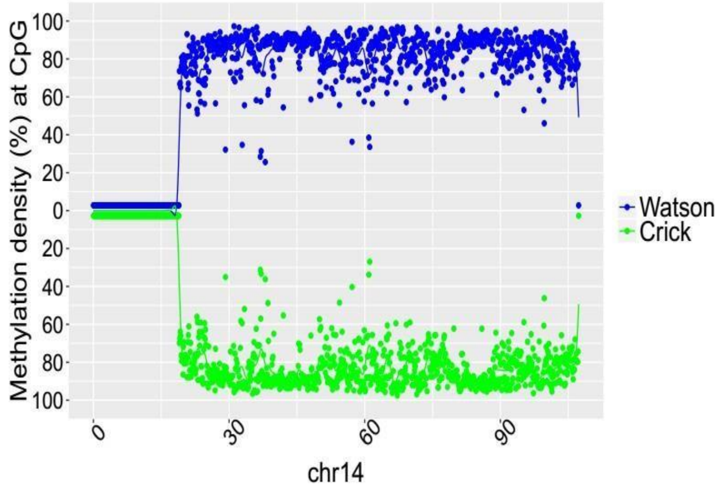
SRR389249 (26 year old male)



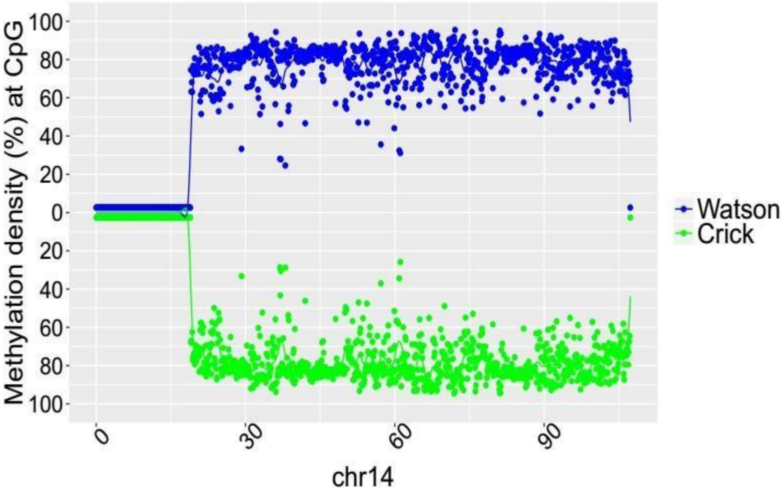
SRR330576 (103 year old male)



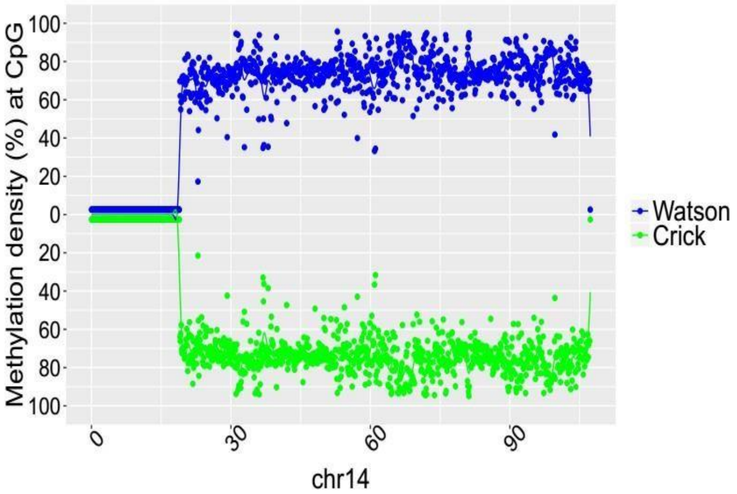
SRR330578 (1 year old male)



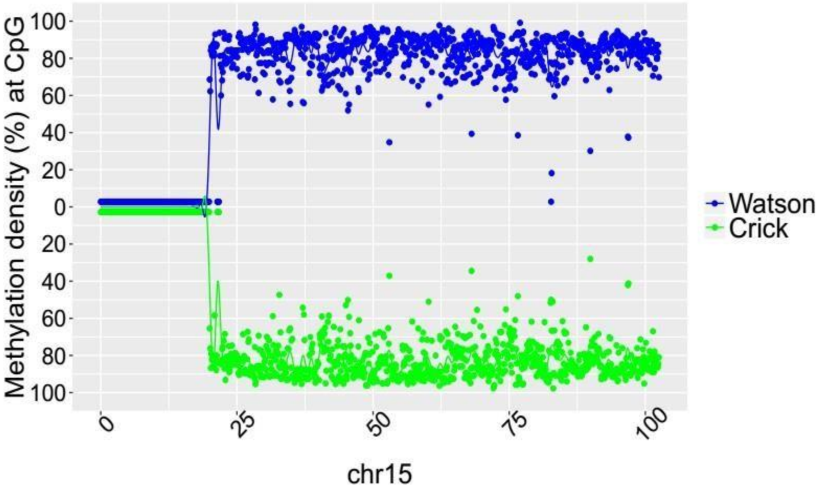
SRR389249 (26 year old male)



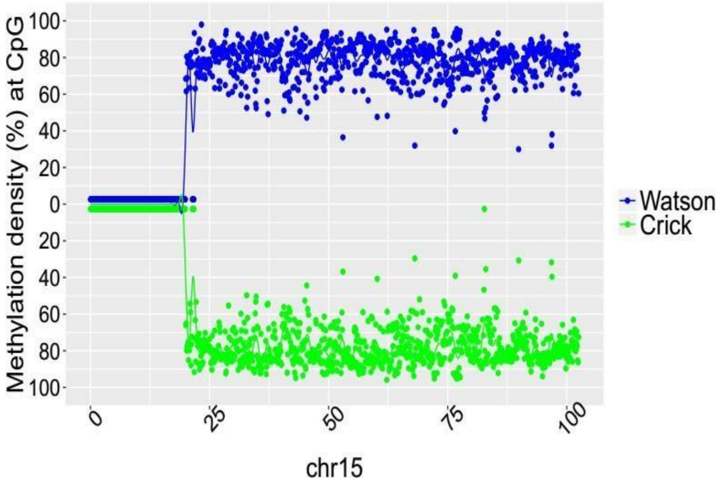
SRR330576 (103 year old male)



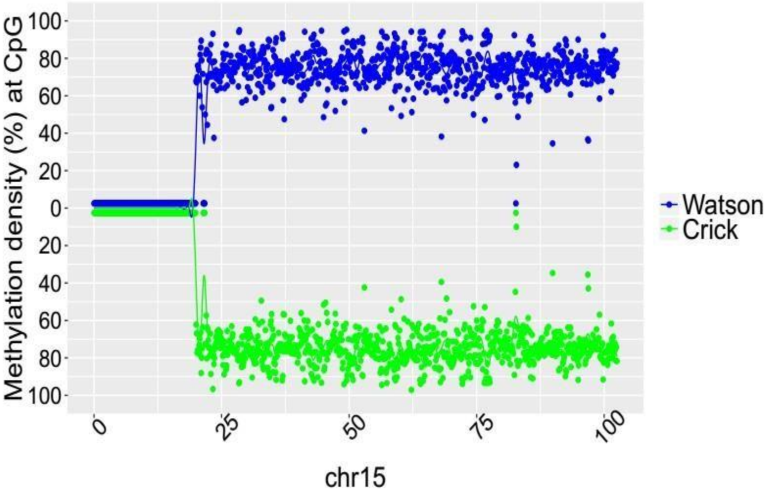
SRR330578 (1 year old male)



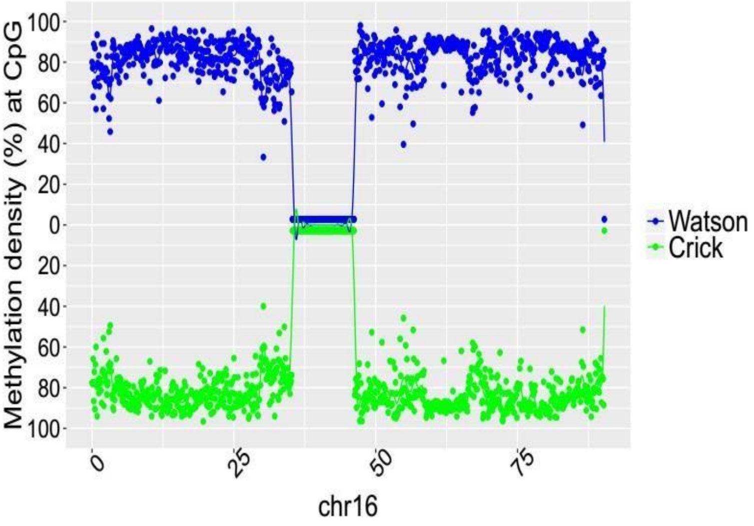
SRR389249 (26 year old male)



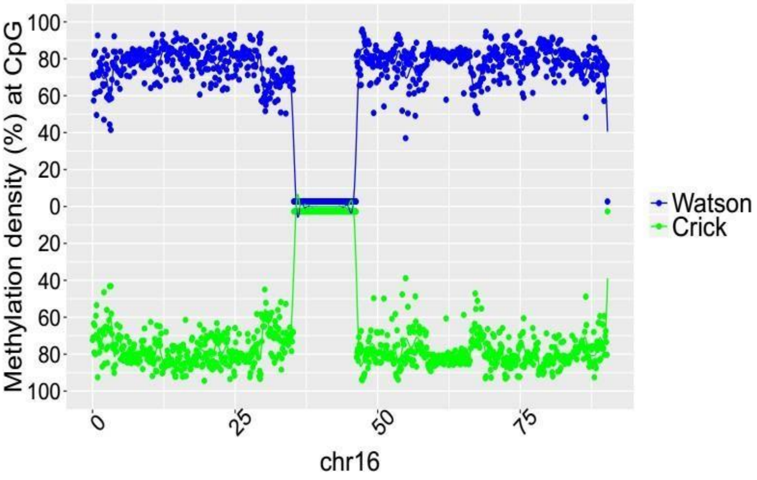
SRR330576 (103 year old male)



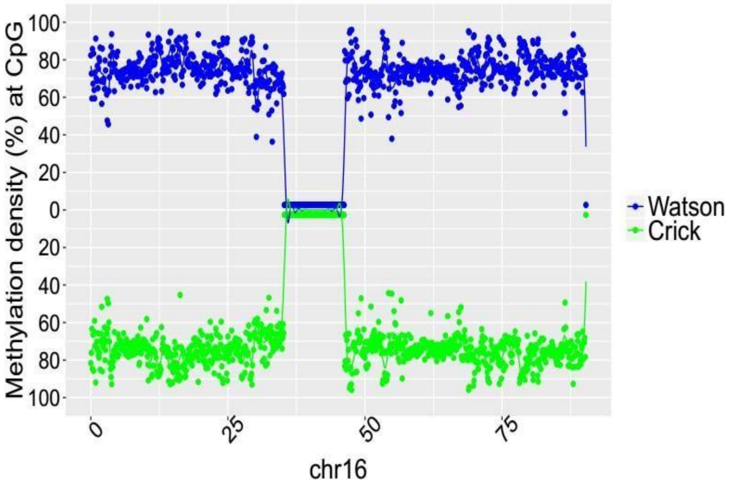
SRR330578 (1 year old male)



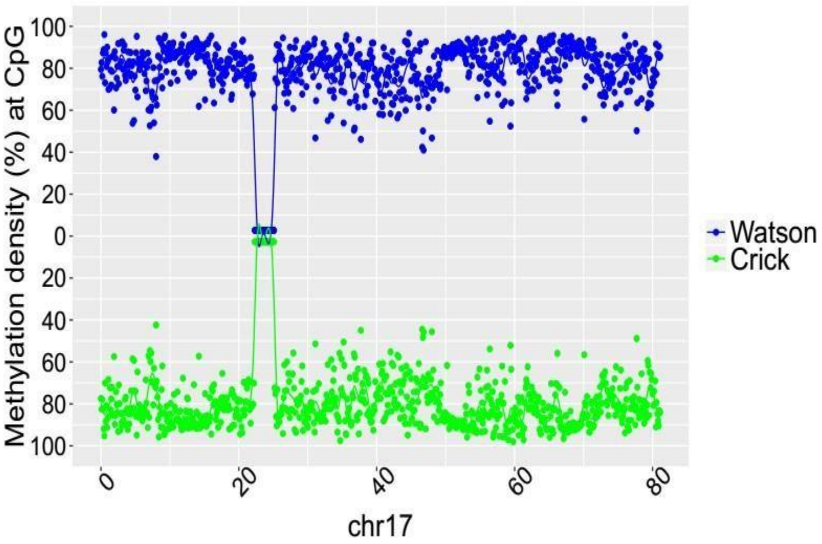
SRR389249 (26 year old male)



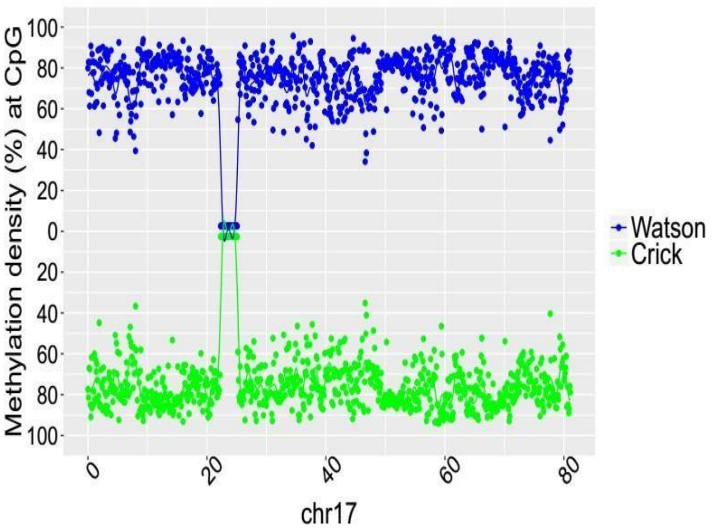
SRR330576 (103 year old male)



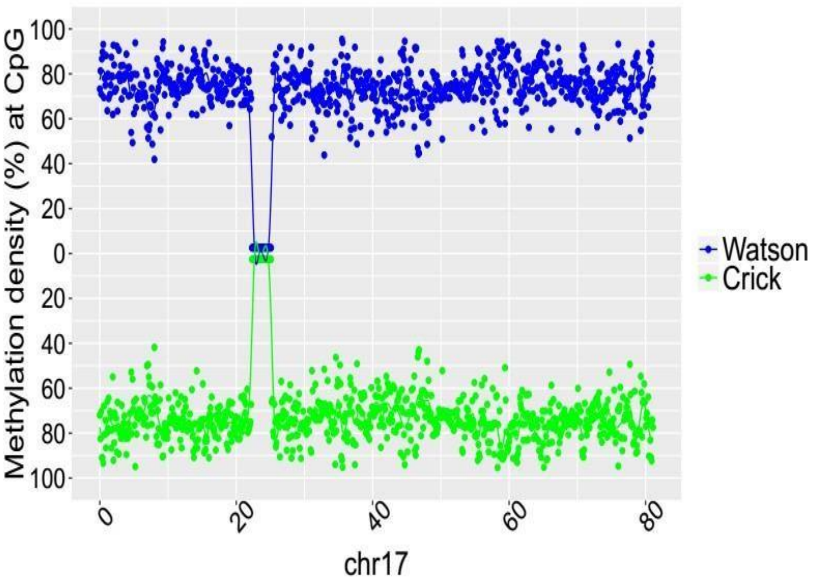
SRR330578 (1 year old male)



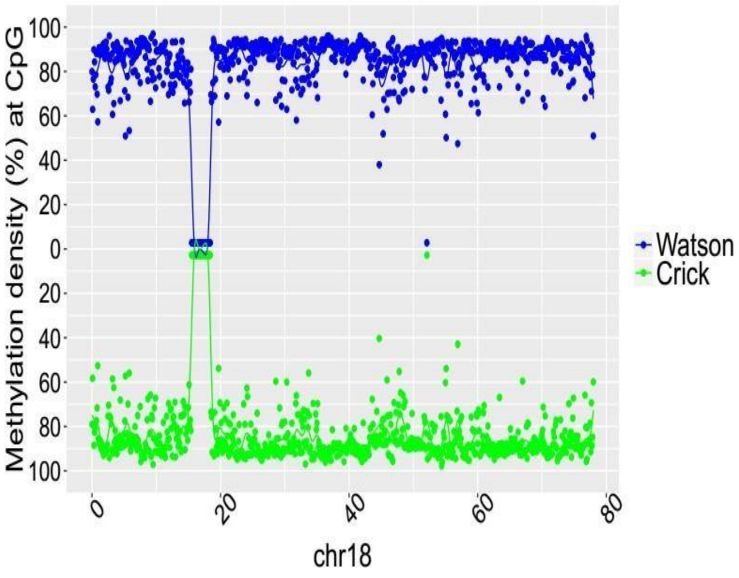
SRR389249 (26 year old male)



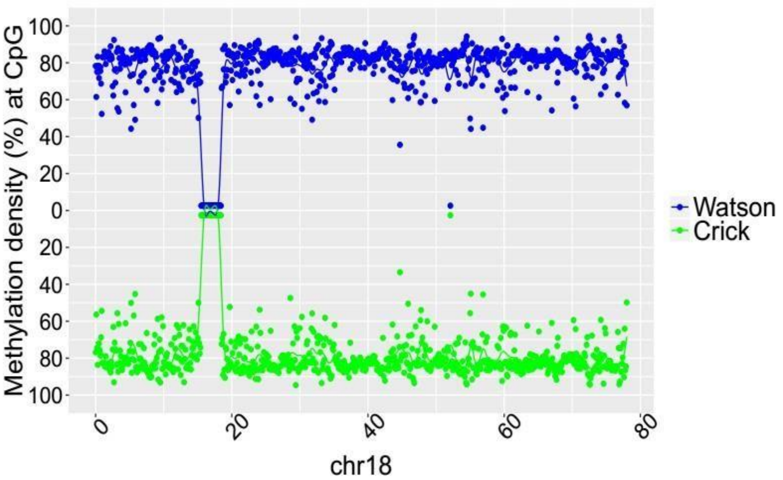
SRR330576 (103 year old male)



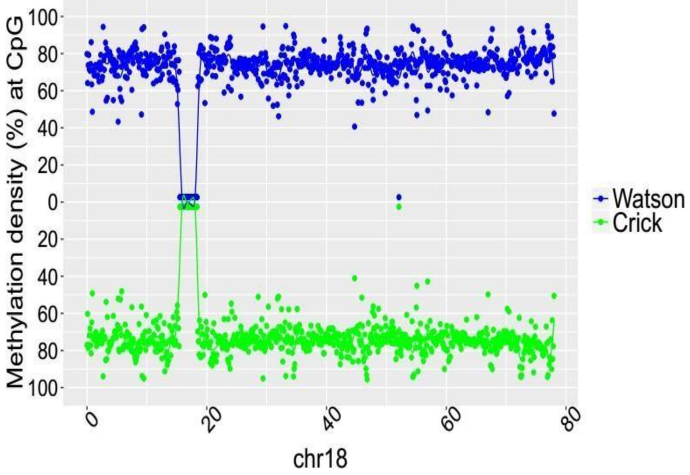
SRR330578 (1 year old male)



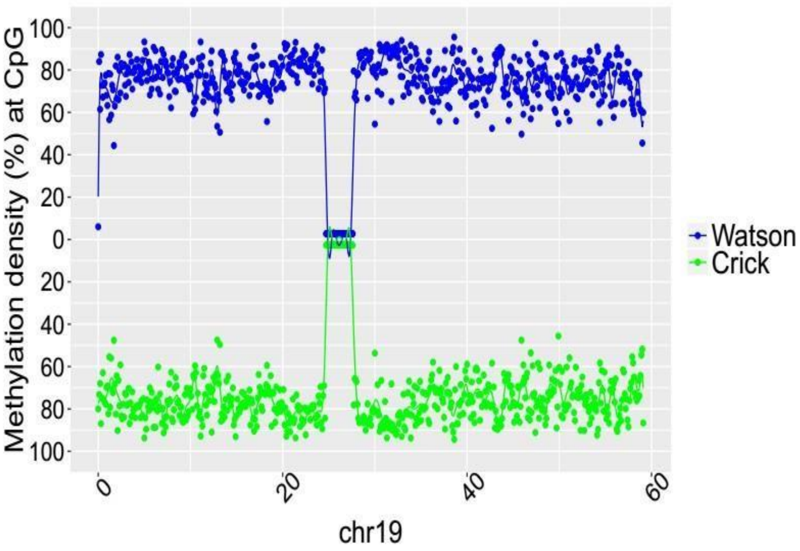
SRR389249 (26 year old male)



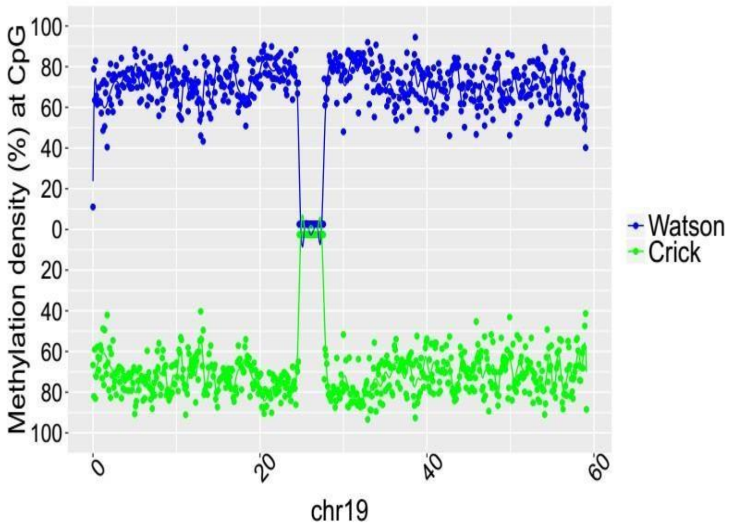
SRR330576 (103 year old male)



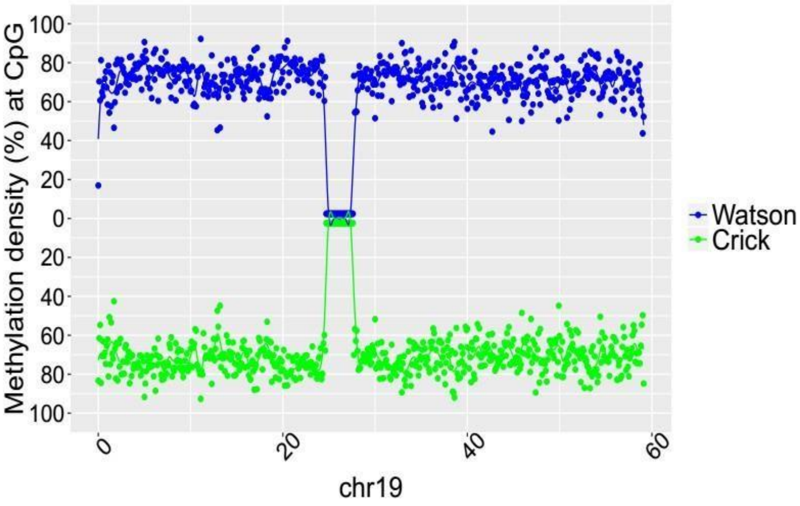
SRR330578 (1 year old male)



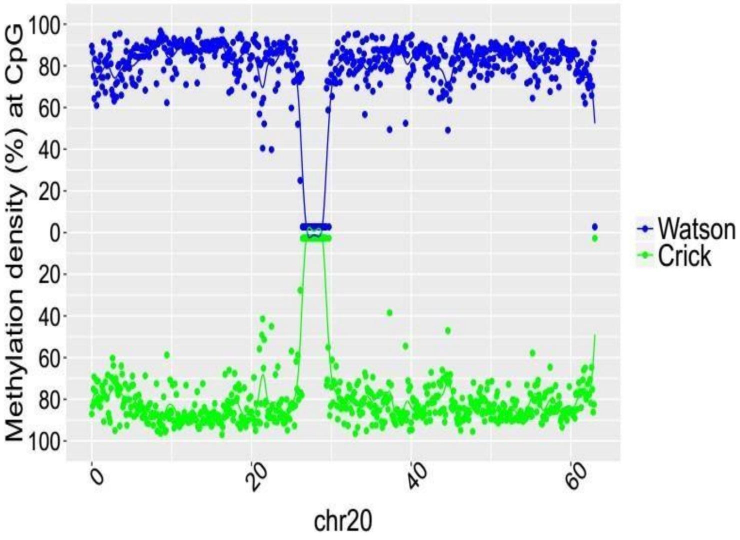
SRR389249 (26 year old male)



SRR330576 (103 year old male)

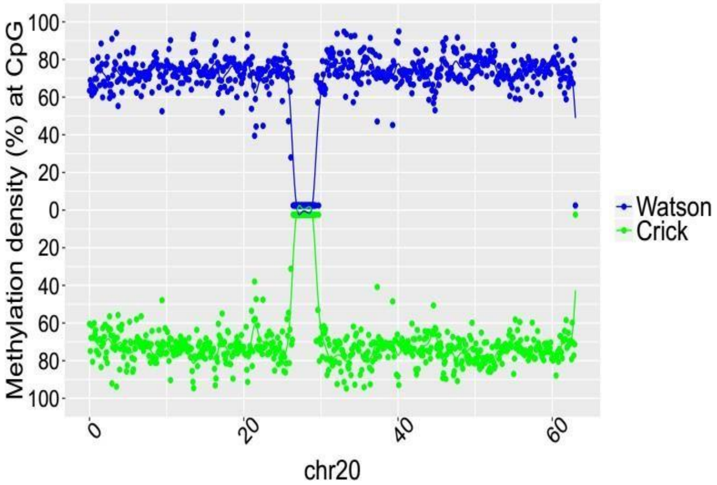


SRR330578 (1 year old male)



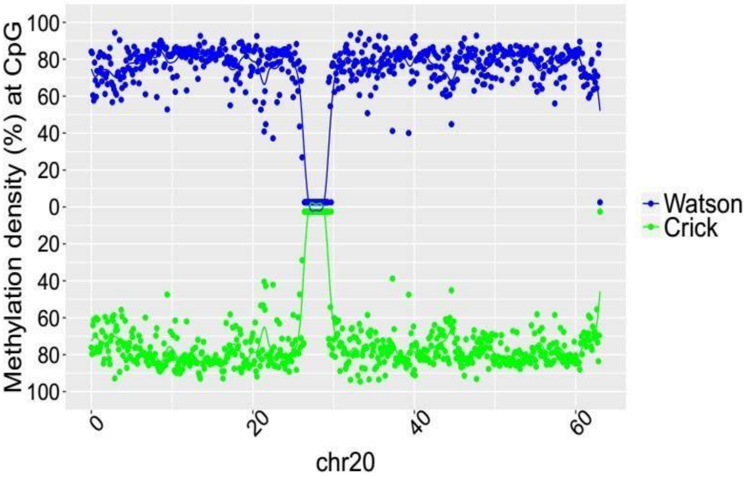
SRR389249 (26 year old male)

SRR330576 (103 year old male)

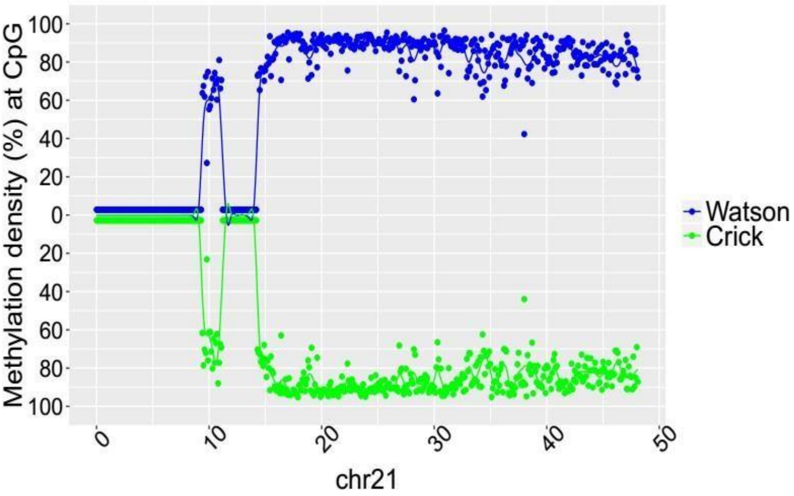


SRR330578 (1 year old male)

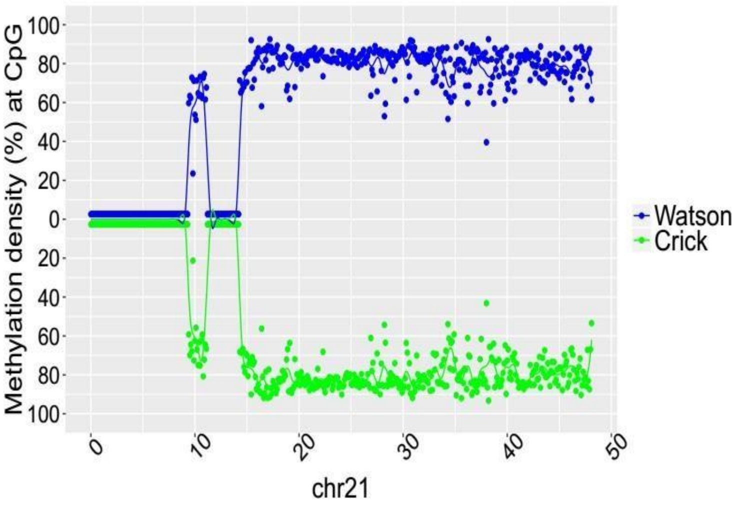
SRR389249 (26 year old male)



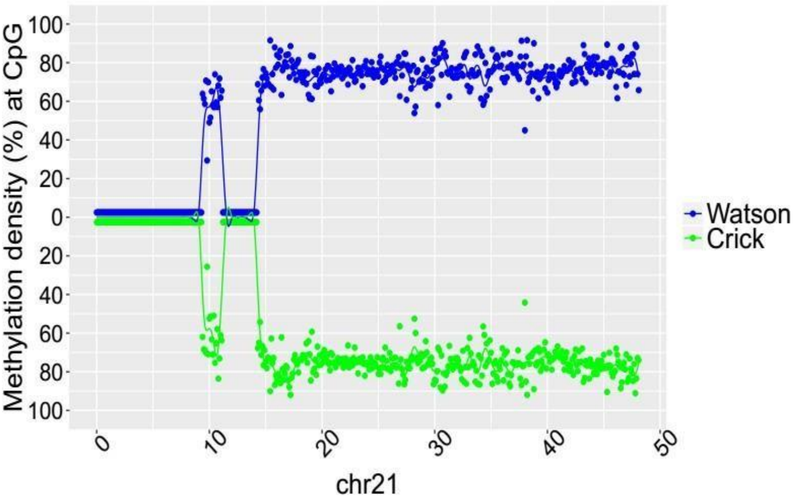
SRR330578 (1 year old male)



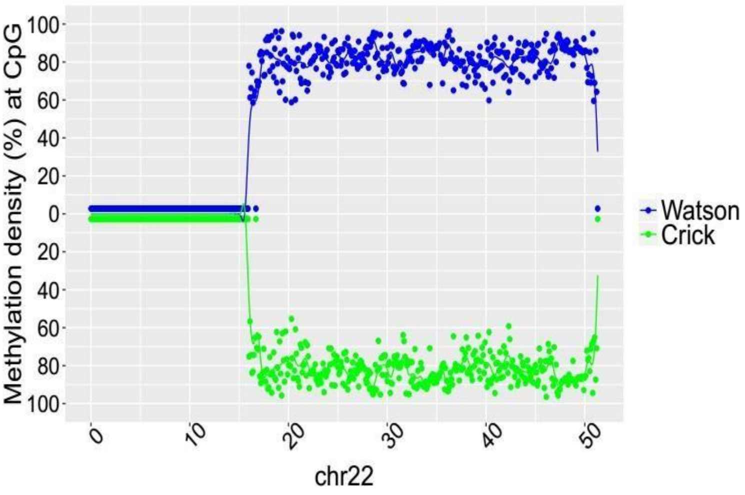
SRR389249 (26 year old male)



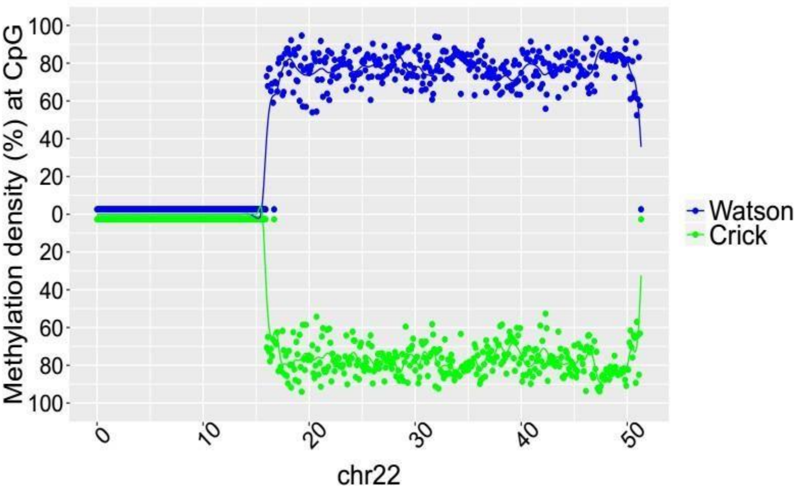
SRR330576 (103 year old male)



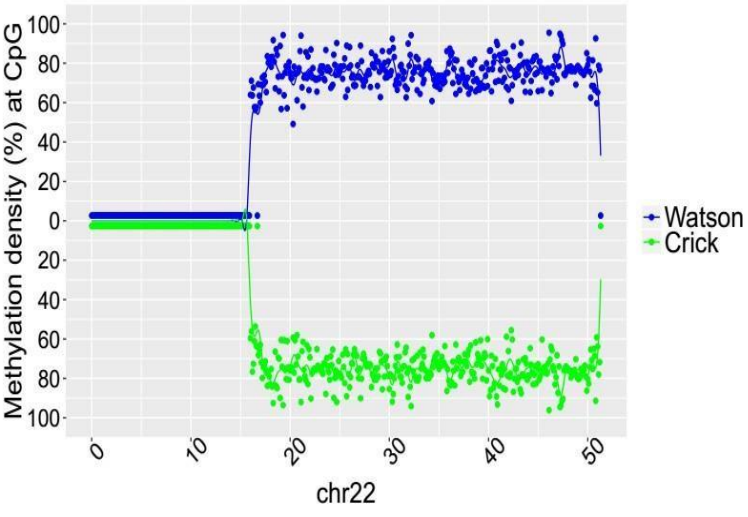
SRR330578 (1 year old male)

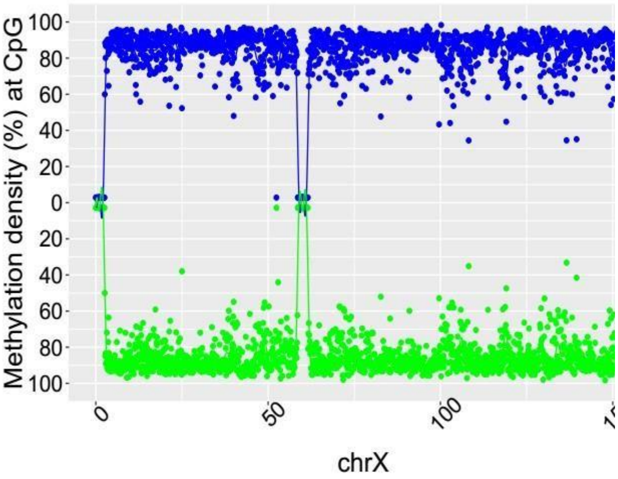


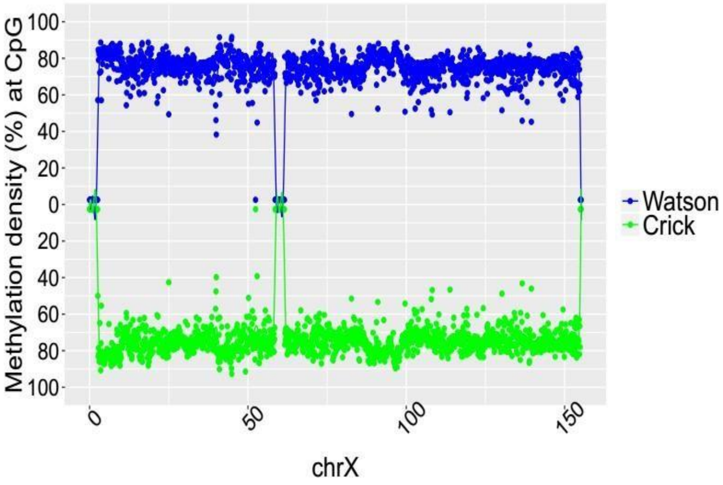
SRR389249 (26 year old male)



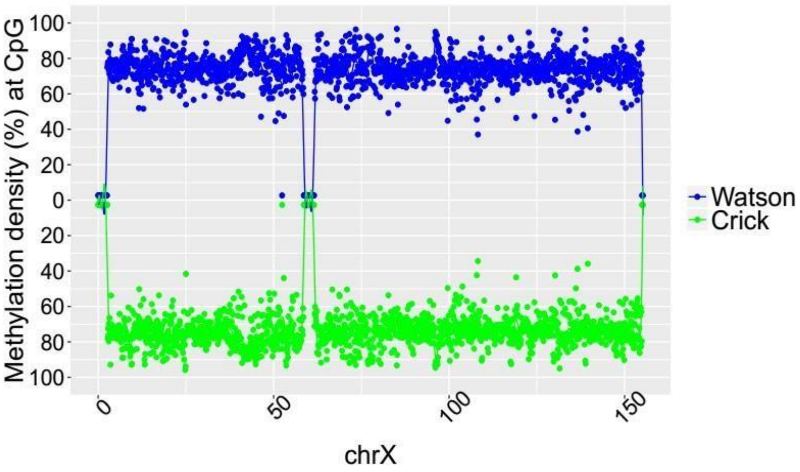
SRR330576 (103 year old male)







SRR330576 (103 year old male)


Methylation density (%) at CpG
chrX
Watson
Crick