Desenvolvimento de um pipeline para análise genômica de metilação do DNA

Henrique Cursino Vieira

Dissertação apresentada
ao
Programa Interunidades em Bioinformática
da
Universidade de São Paulo
para
obtenção do título
de
Mestre em Ciências

Programa: Interunidades em Bioinformática

Orientador: Prof[a]. Dr[a]. Helena Brentani

Durante o desenvolvimento deste trabalho o autor recebeu auxílio financeiro da CAPES

São Paulo, fevereiro de 2016

Desenvolvimento de um pipeline para análise genômica de metilação do DNA

Esta é a versão original da dissertação elaborada pelo candidato Henrique Cursino Vieira, tal como submetida à Comissão Julgadora.

Agradecimentos

Foi um longa jornada de conhecimento, do qual tive o prazer de ser acompanhado por pessoas maravilhosas que me auxiliaram com meu crescimento pessoal e intelectual. Quero agradecer a Prof^{a} Dr^{a}. Helena Brentani por toda a paciência, conhecimento e amizade que teve por mim nestes 5 anos. Agradeço o apoio, carinho e amizade da equipe do laboratório de Patologia Clínica, Thais, Gisele, Jean e a Elza e a equipe de pesquisa orientada pela Prof^{a} Dr^{a}. Helena Brentani, Luciana, Cecília, Aderbal, Bianca, Ana Tahira, Caroline, Kátia, Daniel Mariani, Leandro e Laura. Aos amigos que conheci, Débora Zambori, Mariana Maschietto, Ana Krepischi, Tatiane Rodrigues, Érica Sara, Renato Puga, Prof^{o}. Dr^{o}. Carlos Pereira, Amanda Rusiska e Gesiele Barros. A Walkiria Karla Resende, minha noiva, que conheci a 5 anos atrás, através do nosso mesmo interesse em bioinformática. Aos meus pais e minha irmã, Maria, Manoel e Marina que me deram toda a força para sempre seguir em frente e poder atingir meus objetivos. Agradeço de coração a todos.

Resumo

Vieira, H. C. **Pipeline de análise genômica de Metilação do DNA**.

Background A metilação do DNA é uma das principais modificações epigenéticas estudadas. A plataforma Illumina Infinium HumanMethylation450 oferece o menor custo para interrogar o status de metilação de 480.000 dinucleotídeos CpG distribuídos pelo genoma. Para o estudo do padrão de metilação, os dados são extraídos pelo software GenomeStudio e podem ser analisados neste software ou exportados para análises estatísticas, em geral realizadas em uma das inúmeras pipelines descritas na literatura. Neste nosso trabalho, apresentamos o uso uma análise de metilação diferencial que considere uma relação de interdependência entre as CpGs, sem a necessidade de assumir a normalidade dos dados e sem a necessidade da aplicação da correção para múltiplas testagens. Em nosso trabalho comparamos quatro testes estatísticos, o teste t Student, o teste Wilcoxon e o teste Empírico de Bayes, utilizados nas pipelines de analise de metilação ChAMP, RNBeads e IMA, e o método QUOR, através de um único fluxo de análise. Relacionamos as posições diferencialmente metiladas (DMPs) detectadas entre os 4 testes para determinar quais são as diferenças entre os testes. Analisamos também o uso do valor Beta e do valor M, que são medidas para mensurar o valor de metilação, entre os 4 testes.

;Resultados O número de DMPs detectadas entre os três testes, o teste t Student, o teste Wilcoxon e o teste Empírico de Bayes, foi muito próximos. Conforme aumentamos o corte de confiança para o método QUOR, mais as DMPs detectadas estão em intersecção as DMPs detectadas pelos três testes e ao utilizar um valor de corte de confiança 0.9999, conseguimos encontrar as DMPs que são realmente diferencialmente metiladas. A relação das DMPs detectadas pelos testes t Student, o teste Wilcoxon e o teste Empírico de Bayes, apresentam baixo valor de confiança, demonstrando que não há uma diferença de metilação significativa nestas DMPs detectadas.

;Discussão O valor de Beta se mostra menos confiável em comparação do valor M. Os testes t Student, o teste Wilcoxon e o teste Empírico de Bayes não demonstraram muita diferença na detecção das DMPs entre eles. O QUOR necessita utilizar um valor de confiança muito alto para determinar as DMPs com diferença de metilação significativa.

;Conclusão Atualmente, os microarrays para análise de metilação de DNA são plataformas de baixo custo e grande abrangência. O desafio encontra-se na análise da imensa quantidade de dados gerados. É preferível o uso do valor M ao invés do valor Beta, cada conjunto de dados deve ser investigado e aplicar diferentes tipos de filtragens, normalizações e correções. O fluxo que desenvolvemos pode ser aplicado primariamente para detecção das DMPs em casos onde não está muito bem definido a hipótese. Outras abordagens poderão ser tomadas para o melhoramento do resultado.

Palavras-chave: metilação do DNA, lâmina Illumina HumanMethylation450, diferença de metilação entre grupos, posições diferencialmente metiladas.

Abstract

Vieira, H. C. **Pipeline of genomic analysis of DNA methylation**.

Background The DNA methylation is a major epigenetic changes studied. The Illumina Infinium HumanMethylation450 platform provides the lowest cost to interrogate the methylation status of 480,000 CpG dinucleotide distributed throughout the genome. For the study of methylation pattern, the data is extracted by GenomeStudio software and can be analyzed in software or exported for statistical analysis in general carried out in one of the numerous pipelines described in the literature. In our work, we present the use of differential methylation analysis that considers an interdependent relationship between CpG without the need to assume the normality of the data and without the need to apply the correction for multiple testings.

In our study we compare four statistical tests, Student's t test, Wilcoxon test and the empirical Bayes test, used in methylation analysis pipelines ChAMP, RNBeads and IMA, and Quor method, through a single analysis flow. We relate the differentially methylated positions (DMPs) detected among the 4 tests to determine what are the differences between tests. We also analyzed the use of beta value and the M value, which is measured to measure the methylation value among the four tests.

;Results The number of DMPs detected between the three tests, the Student t test, Wilcoxon test and the empirical Bayes test, was very close. As we increase the confidence of court for Quor method detected more DMPs are in the intersection DMPs detected by three tests and using a reliable cutoff 0.9999, we find the DMPs that are truly differentially methylated. The list of DMPs detected by Student t test, Wilcoxon test and the empirical Bayes test, have a low amount of confidence, demonstrating that there is no significant difference in these methylation detected DMPs.

;Discuss The value of Beta proves less reliable in comparison of the value M. The test t Student, the Wilcoxon test and the empirical Bayes test did not show much difference in the detection of DMPs between them. The Quor need to use a very high confidence value to determine the DMPs with significant methylation difference.

;Conclusion Currently, the microarray for DNA methylation analysis are platforms low cost and wide coverage. The challenge lies in the analysis of the vast amount of generated data. It is preferable to use the value M rather than the beta value, each set of data should be investigated and apply different types of filtering, normalization and corrections. The flow we developed can be applied primarily for the detection of DMPs in cases where it is not very well defined hypothesis. Other approaches can be taken to improve the result.

Keywords: DNA methylation, Illumina HumanMethylation450 BeadChip, methylation difference

between groups, differentially methylated positions.

Sumário

Lista de Abreviaturas ix

Lista de Símbolos xi

Lista de Figuras xiii

Lista de Tabelas xv

1 Introdução 1

1.1 Epigenética . . . 1
1.2 Metilação do DNA . . . 1
1.3 Tipos de ensaio . . . 2
1.4 Tecnologia do microarranjo Illumina Infinium HumanMethylation450 . . . 4
1.5 Análise e interpretação dos dados de metilação do DNA . . . 6
1.6 Relação entre valor β e valor M . . . 8
1.7 Fluxo de análise . . . 9
1.8 Importação, qualidade e pré-processamento . . . 9
1.9 Correção . . . 10
1.10 Normalização . . . 11
1.11 Detecção de diferenças no padrão de metilação . . . 12
1.12 Múltiplas testagens . . . 13

2 Justificativa e objetivos 15

2.1 Justificativa . . . 15
2.2 Objetivos . . . 16
2.2.1 Objetivo geral . . . 16
2.2.2 Objetivos específicos . . . 16

3 Materiais e métodos 17

3.1 Obtenção dos dados . . . 17
3.1.1 Amostra 1 COCAINA/CRACK . . . 17
3.1.2 Amostra 2 GRAVIDAS – estudo do impacto do estresse mãe-bebê . . . 17
3.2 Pré-processamento . . . 17
3.3 Normalização . . . 18
3.4 Análise e correções de efeitos de lote . . . 18
3.5 Identificação de posições diferencialmente metiladas . . . 18

4 Resultados **19**
4.1 Pré-processamento, normalização e correção do dado 19
4.2 Identificação das DMPs 22
4.3 Intersecção dos resultados entre o QUOR e os testes 22
4.4 Intersecção dos resultados entre todos os testes 22
4.5 Comparação entre os cortes de Confiança 24

5 Discussão e conclusão **37**
5.1 Discussão 37
5.2 Conclusão 38

Referências Bibliográficas **39**

Glossário **47**

Lista de Abreviaturas

BMIQ - Beta-mixture quantile normalization
CNA - Copy Number Aberrations / Aberracões do número de cópia
CNV - Copy Number Variation / Variações no número de cópia
CpG - 5'—C—phosphate—G—3' / 5'—C—fosfato de ligação—G—3'
DMP - Differentially Methylated Positions / Posições diferencialmente metiladas
DMR - Differentially Methylated Regions / Regiões diferencialmente metiladas
DNA - Deoxyribonucleic acid / Ácido Desoxirribonucléico
FDR - False discovery rate / Taxa de falsos descobertos
IMA - Illumina Methylation Analyzer
ISVA - Independent Surrogate Variable Analysis / Análise de variável substituta independente
MDS - Multidimensional Scaling
MVP - Methylation variable position / Metilação de posição variável
PBC - Peak Based Correction / Correção por pico
PCR - Polymerase Chain Reaction / Reação em cadeia da polimerase
QN - Quantile Normalization / Normalização quantílica
SNP - Single Nucleotide Polymorphism / Polimorfismo de nucleotídeo único
SQN - Subset quantile normalization
SVA - Surrogate Variable Analysis / Análise de variável substituta
SWAN - Subset-quantile Within Array Normalization
TSS - Transcription Start Site
UCSC - University of California, Santa Cruz
UTR - Untranslated Region

Lista de Símbolos

β - valor Beta
M - valor M

Lista de Figuras

1.1 Metilação do DNA 2
1.2 Padrões de metilação do DNA 3
1.3 Visão geral da cobertura e design da lâmina Infinium HumanMethylation450 array 5
1.4 Contexto genômico da metilação da CpG 6
1.5 Visão geral dos ensaios Infinium I e Infinium II 7
1.6 Gráfico da média e o desvio-padrão 9
1.7 Visão geral do fluxo de analise 9

4.1 Agrupamento através do escalonamento multidimensional e distribuição dos valores de β sem normalização para cada sonda 20
4.2 Agrupamento através do escalonamento multidimensional e distribuição dos valores de β com normalização para cada sonda 20
4.3 Homoscedaticidade e heteroscedasticidade entre valor M e valor β 21
4.4 Intersecção das DMPs detectadas com exceção das DMPs detectadas pelo QUOR - COCAÍNA/CRACK - M 26
4.5 Intersecção das DMPs detectadas com exceção das DMPs detectadas pelo QUOR - COCAÍNA/CRACK - β 27
4.6 Intersecção das DMPs detectadas com exceção das DMPs detectadas pelo QUOR - GRAVIDAS - M 28
4.7 Intersecção das DMPs detectadas com exceção das DMPs detectadas pelo QUOR - GRAVIDAS - β 29
4.8 CpG detectada com valor de confiança maior que 0.99 - COCAÍNA/CRACK - M . 30
4.9 CpG detectada com valor de confiança maior que 0.99 - COCAÍNA/CRACK - β . 32
4.10 CpG detectada com valor de confiança maior que 0.9999 - COCAÍNA/CRACK - M 33
4.11 CpG detectada com valor de confiança maior que 0.9999 - COCAÍNA/CRACK - β 34
4.12 CpG detectada com valor de confiança maior que 0.99 - GRAVIDAS - M 34
4.13 CpG detectada com valor de confiança maior que 0.99 - GRAVIDAS - β 35

Lista de Tabelas

4.1 Quantidade de sondas removidas por conjunto de dados 19
4.2 Relação das DMPs detectadas com o conjunto de dados COCAÍNA/CRACK . . . 22
4.3 Relação das DMPs detectadas com o conjunto de dados GRAVIDAS 23
4.4 Intersecção das DMPs detectadas entre cada teste com as DMPs detectadas com QUOR do conjunto de dados COCAÍNA/CRACK sem correção 24
4.5 Intersecção das DMPs detectadas entre cada teste com as DMPs detectadas com QUOR do conjunto de dados COCAÍNA/CRACK com correção 25
4.6 Intersecção das DMPs detectadas entre cada teste com as DMPs detectadas com QUOR do conjunto de dados GRAVIDAS sem correção 30
4.7 Intersecção das DMPs detectadas entre cada teste com as DMPs detectadas com QUOR do conjunto de dados GRAVIDAS com correção 31
4.8 Intersecção entre todas as DMPs detectadas entre os 4 testes, DMPs detectadas somente pelo QUOR e DMPS detectadas pelos testes **"Paramétrico"**, **"Não paramétrico"** e **"Bayesiano"** no conjunto de dados COCAÍNA/CRACK 31
4.9 Intersecção entre todas as DMPs detectadas entre os 4 testes, DMPs detectadas somente pelo QUOR e DMPS detectadas pelos testes **"Paramétrico"**, **"Não paramétrico"** e **"Bayesiano"** no conjunto de dados GRAVIDAS 32

Capítulo 1

Introdução

1.1 Epigenética

O termo "epigenética" foi introduzido por Conrad Waddington em 1946 e definido como um conjunto de mecanismos moleculares que convertem a informação genética em traços e fenótipos observáveis que controlam a expressão gênica de um organismo sem alterar a sequência de DNA (Deoxyribonucleic Acid/Ácido desoxirribonucleico) (Portela e Esteller [2010]). Os mecanismos epigenéticos têm participação fundamental em processos celulares, como a diferenciação celular e a inativação do cromossomo X (Portela e Esteller [2010]; Suzuki e Bird [2008]), e, de uma forma potencialmente hereditária (Portela e Esteller [2010], Bird [2002]), são mantidos e passados de uma geração a outra, persistindo por meio da mitose ou até mesmo da meiose (Portela e Esteller [2010]). O genoma pode apresentar características e comportamentos muito distintos devido às diferenças na expressão dos genes, que pode ser modificada ao longo da vida conforme são "ativados" e "desativados" por agentes externos aos próprios genes (Esteller [2002]). Gêmeos monozigóticos, por exemplo, que são idênticos em termos de DNA, diferem entre si quanto ao padrão de metilação do DNA e às modificações no perfil das histonas, levando ao surgimento de diversas doenças como câncer ou desordens autoimunes não concordantes em pares de gêmeos monozigóticos (Portela e Esteller [2010]).

As modificações epigenéticas podem ser agrupadas em três principais categorias: metilação do DNA, modificação de histonas e remodelagem da cromatina (Portela e Esteller [2010]). A metilação do DNA, foco deste estudo, constitui a modificação epigenética mais amplamente pesquisada (Portela e Esteller [2010]; Marabita *et al.* [2013]; Smith e Meissner [2013]; Sandoval *et al.* [2011]).

1.2 Metilação do DNA

A metilação do DNA consiste na adição de um agrupamento metil (CH3) ao carbono da posição cinco da citosina (C), catalizada pela enzima DNA metil-transferase em dinucleotídeos CpG, que são regiões em que uma citosina (C) precede uma guanina (G) na sequência de DNA ('C-fosfato de ligação-G'). Trata-se de um processo comum em todo o genoma, conforme apresentado na Figura 1.2 (Portela e Esteller [2010]).

O processo de metilação do DNA é a modificação epigenética mais bem caracterizada até o momento, exercendo grande importância no silenciamento e na regulação gênica, em particular no imprinting genômico, na inativação do cromossomo X e no silenciamento de transpósons. Esses dinucleotídeos se agrupam no genoma em regiões conhecidas como ilhas CpG ("CpG Island"), definidas como regiões com mais de 200 pares de bases com um conteúdo $G + C$ maior que 50% e uma razão das frequências de CpG observada versus esperada de no mínimo 60% (Gardiner-Garden e Frommer [1987]).

Cerca de 60% dos promotores dos genes estão associados com ilhas CpG, que são normalmente não metiladas em células normais diferenciadas, e cerca de 6% são encontrados metilados de maneira tecido-específica durante o desenvolvimento inicial ou em tecidos diferenciados (Straussman *et al.*

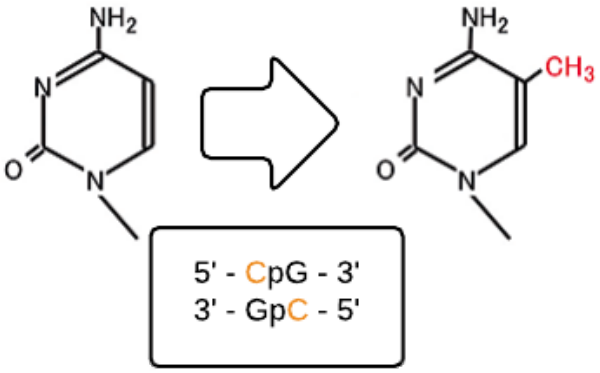


Figura 1.1: *A metilação da Citosina (C) ocorre através da adição de um grupo metil (CH3) a posição cinco da Citosina (C). A reação acontece na presença da enzima metil-transferase.*

[2009]). A metilação em ilhas CpG pode ser associada ao silenciamento gênico, conforme indicado na Figura 1.2.a (Portela e Esteller [2010]), podendo, contudo, ocorrer a metilação fora dessas ilhas CpG, em regiões com menor densidade de CpGs localizadas a aproximadamente 2 kb de distância das ilhas CpG que foram denominadas de "margens da ilha CpG" ("CpG Island shores") (Figura 1.2.b). A metilação destas regiões também foi associada ao controle transcricional (Irizarry [2009]; Doi *et al.* [2009]) e, de forma menos frequente, à ativação transcricional, particularmente quando ocorre no corpo dos genes (Figura 1.2.c). A metilação nesta região está relacionada à eficiência de elongação e prevenção de iniciações espúrias da transcrição (Zilberman e Henikoff [2007]), relacionando-se ao controle de expressão de isoformas tempo/local específicas. Uma fração significativa de CpGs estão profundamente metiladas é encontrada em elementos repetitivos (Figura 1.2.d). A metilação destas áreas se faz necessária para proteger a integridade genômica por meio da prevenção da reativação de sequências que causam instabilidade cromossômica, translocações e quebras gênicas (Esteller (2007)).

O equilíbrio entre a estabilidade e a flexibilidade no padrão de metilação do DNA pode ser observado nas interações entre o meio externo e o indivíduo, permitindo determinar riscos e diagnósticos de doenças (Bock [2012]). Muitos estudos apontam a relação entre a metilação do DNA e diversos processos biológicos. Além disso, alterações nos padrões de metilação foram relatadas em casos de câncer e de diversas outras doenças (Bock [2012]; Reik *et al.* [2001]; Irizarry *et al.* [2008]; Dyson *et al.* [2014]; Sandoval *et al.* [2011]).

A região diferencialmente metilada (DMR - Differentially Methylated Region) é uma região genômica com múltiplos locais de CpG adjacentes que exibem diferentes estados de metilação em múltiplas amostras, capazes de distinguir um fenótipo de outro, e fornece o exemplo mais bem analisado de variação de metilação. Sua identificação e utilidade têm implicações de longo alcance para aplicações clínicas, reduzindo a escala do genoma para um grupo menor de regiões. Nesse sentido, os objetivos dos estudos de DMR podem ser divididos em dois tipos: **(i)** identificação de DMRs em todas as populações e **(ii)** identificação de DMRs dentro de uma população (Hsiao *et al.* [2014]; Butcher e Beck [2014b]). As associações das DMRs foram listadas em Hsiao *et al.* [2014].

1.3 Tipos de ensaio

A expansão dos estudos sobre a metilação do DNA nos últimos anos permitiu o avanço das análises de metilação (Bibikova *et al.* [2011]). O desenvolvimento de tecnologias de microarranjo e de sequenciamento permitiu o surgimento de diversas metodologias para o mapeamento da metilação do DNA, possibilitando o mapeamento de sítios de metilação com alta resolução e em grande número de amostras e apresentando vantagens e desafios para o processamento dos dados, para a análise estatística e para a interpretação biológica (Bibikova *et al.* [2011]; Bock [2012]; Sandoval *et al.* [2011]).

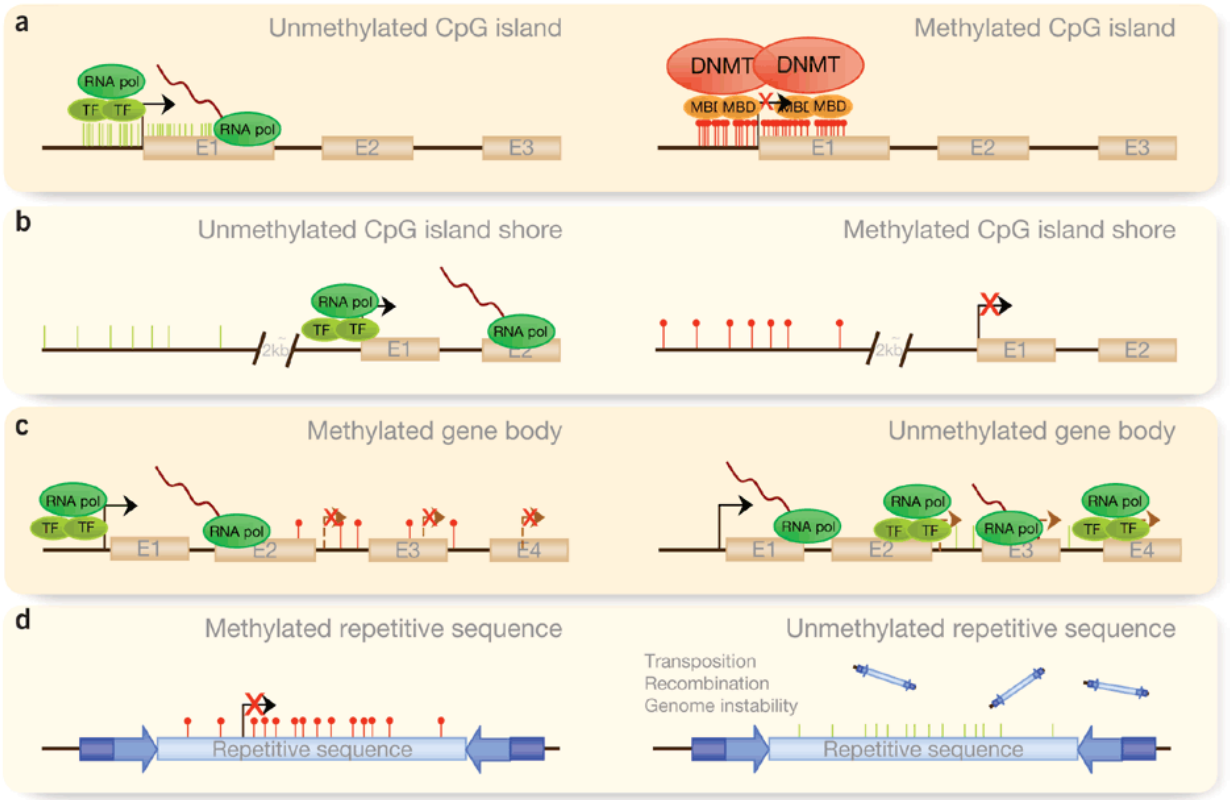


Figura 1.2: *Padrões de metilação do DNA. A metilação do DNA pode ocorrer em diferentes regiões do genoma. O funcionamento normal da metilação está representado na coluna da esquerda e alterações desse padrão são mostradas à direita. (A) As ilhas CpG estão localizadas nos promotores dos genes e são normalmente não metiladas, permitindo a transcrição. A hipermetilação nestas ilhas CpGs localizadas no promotor podem levar à inativação transcricional deste gene. (B) Um número menor de CpGs estão localizadas nas margens da ilha CPG (island shores) que estão localizadas a 2 mil pares de base de distância da ilha CpG, podendo também influenciar na inativação transcricional deste gene. (C) Quando a metilação ocorre no corpo do gene, a transcrição é facilitada, porque a metilação impede que inícios transcricionais espúrios sejam utilizados. Em caso de doença, o corpo tende a desmetilar o gene, permitindo que a transcrição seja iniciada em vários locais errados. (D) As sequências repetitivas tendem a estar hipermetiladas, evitando a instabilidade cromossômica, translocações e ruptura do gene por reativação de seqüências de endoparasitas. Este padrão pode ser alterado por doenças. Figura retirada de "Epigenetic modifications and human disease", Portela e Esteller [2010].*

O uso de técnicas como o pirossequenciamento é útil quando há poucas CpGs para estudo e muitas amostras (Sandoval *et al.* [2011]). O pirossequenciamento é um método de sequenciação por síntese, que monitora a incorporação quantitativa em tempo real de nucleótidos por meio da conversão enzimática do pirofosfato para um sinal de luz proporcional. As medidas quantitativas são de especial importância para a análise de metilação do DNA em várias situações desenvolvimentais e patológicas. Dessa forma, a análise de padrões de metilação do DNA por pirossequenciamento combina um protocolo de reação simples com medidas reprodutíveis e precisas do grau de metilação em vários CpGs em estreita proximidade com alta resolução quantitativa. Após o tratamento com bissulfito e PCR (Polymerase Chain Reaction/Reação em cadeia da polimerase), o grau de cada metilação de CpG em cada posição em uma sequência é determinada a partir da proporção de T e C. O processo de purificação e de sequenciação pode ser repetido para o mesmo modelo a fim de analisar outros CpG no mesmo produto de amplificação (Tost *et al.* [2003]).

Dentre as metodologias desenvolvidas, o microarranjo representa uma poderosa ferramenta para os estudos de metilação em larga escala, permitindo a análise de um maior número de amostras a um custo relativamente acessível (Bibikova *et al.* [2011]). Em muitos casos, as alterações na metilação do DNA são sutis, e a variabilidade biológica pode ser elevada, tornando particularmente relevante o tamanho da amostra a ser estudada (Marabita *et al.* [2013]).

1.4 Tecnologia do microarranjo Illumina Infinium HumanMethylation450

Em 2007, foi desenvolvido pela empresa Illumina® um microarranjo de metilação de DNA, conhecido como GoldenGate, com aproximadamente 1500 CpGs associadas a mais de 800 genes relacionados ao câncer (Bibikova *et al.* [2006]). Um ano depois, Bibikova et al. aprimoraram essa tecnologia, desenvolvendo o microarranjo Infinium HumanMethylation27 BeadChip, com cerca de 27 mil CpGs associadas a aproximadamente 14 mil genes (Bibikova *et al.* [2009]). Em 2011, foi desenvolvido um novo microarranjo pela empresa Illumina®, a lâmina Infinium HumanMethylation450 BeadChip, que possui 485.577 sítios (482.421 CpGs, 3.091 não CpGs e 65 SNPs variados), com uma cobertura de 99% de genes RefSeq com múltiplas sondas por gene e 96% de regiões de ilhas CpG e "shelf" (Bibikova *et al.* [2011]; Sandoval *et al.* [2011]). Quanto à distribuição genômica funcional deste último modelo, 200.339 CpGs estão localizadas em regiões promotoras, 150.212 no corpo de genes, 15.383 em 3'UTRs e 119.830 em regiões intergênicas (Sandoval *et al.* [2011]). Em relação ao mapeamento genômico das ilhas, 150.254 sítios encontram-se em ilhas, 112.072 sítios em "shores", 47.161 sítios em "shelves" e 176.277 sítios em "open seas" (regiões intergênicas) (Sandoval *et al.* [2011]), Figura 1.3.

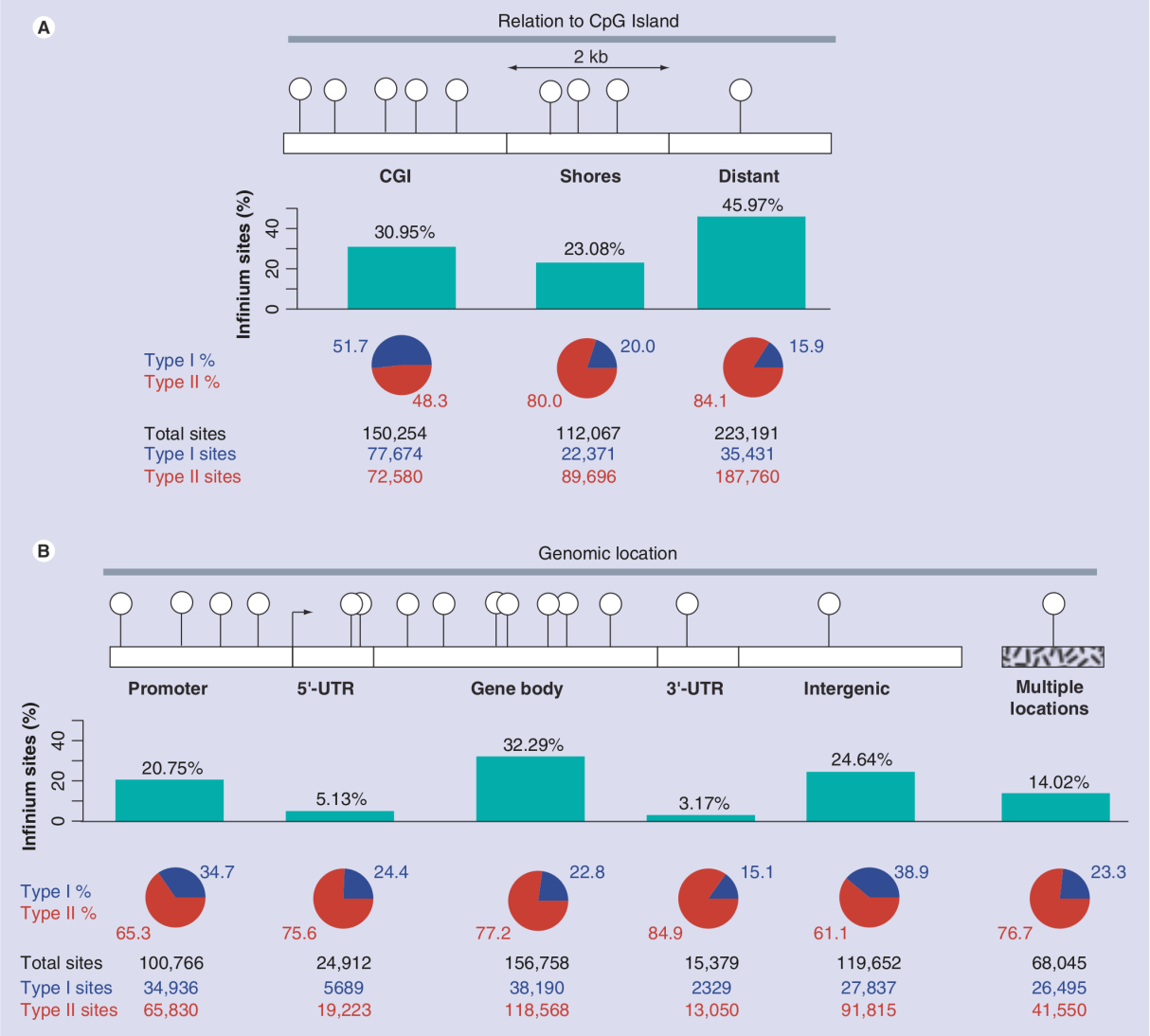


Figura 1.3: *Visão geral da cobertura e design da lâmina Infinium HumanMethylation450. Figura retirada de "Evaluation of the Infinium Methylation 450K technology", Dedeurwaerder et al. [2011].*

A maior parte das sondas na lâmina Infinium HumanMethylation450 BeadChip é focada em genes e ilhas (Dyson *et al.* [2014]), seleção que foi definida pelo Consórcio de Pesquisadores de Epigenética. A subdivisão das regiões de ilhas CpG e dos genes foi definida de acordo com as classificações do banco de dados "UCSC genome browser" (Bibikova *et al.* [2011]). Para a definição das anotações para cada sítio, foi utilizado o contexto de posicionamento relativo tanto às regiões gênicas quanto às ilhas CpGs mais próximas (Figura 1.4).

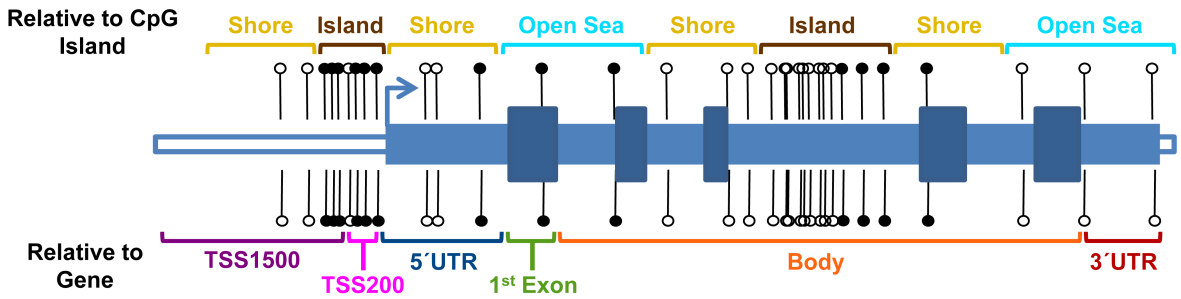


Figura 1.4: *Contexto genômico da metilação da CpG: Diagrama que descreve o contexto genômico relativo à ilha CpG mais próxima (esquema superior) ou do gene (esquema inferior). No contexto relativo a ilha CPG, foi definido como sendo a 'ilha' (em marrom), a 'margem' de 4 kb flanqueando a ilha (em amarelo), ou 'mar aberto' após os 4 kb da 'margem' da 'ilha'. O contexto relativo ao gene foi definido em relação a ORF (open reading frame) mais próxima: local dentro de 1500 pares de base de distância do sítio de início da transcrição (transcription start site - TSS) (TSS1500; em roxo) ou local dentro de 200 pares de base de distância do sítio de início da transcrição (TSS200; em rosa); na região 5'UTR (Untranslated Region - UTR) (em azul), no primeiro exon de um transcrito (em verde); no corpo de gene (em laranja) ou 3' UTR (em vermelho). Figura retirada de "Genome-Wide DNA Methylation Analysis Predicts an Epigenetic Switch for GATA Factor Expression in Endometriosis", Dyson et al. [2014]*

Dessa forma, 80% das 470.540 CpGs examinadas no microarranjo estão ligadas a um gene mapeado nas regiões promotoras ou localizado próximo a elas (área que abrange os promotores 5' UTRs e o primeiro éxon), dentro do corpo do gene (tipicamente intrônico) ou na região 3' UTR do gene (Dyson *et al.* [2014]). A anotação das sondas baseada na proximidade das ilhas CpG mapeia, aproximadamente, um terço das CpGs dentro das ilhas, um terço nas regiões próximas às ilhas ("shores" e "shelves") e um terço em "open sea" (Bibikova *et al.* [2011]). Esta lâmina possui dois ensaios experimentais com características distintas, chamados de Infinium I (135.501 sondas) e Infinium II (350.076 sondas). O ensaio Infinium I segue a tecnologia do seu precursor, o Infinium HumanMethylation27 BeadChip, que é composto de duas sondas, uma sonda para o locus "metilado" e uma sonda para o locus "não metilado". Para cada sítio, a terminação 3' da sonda é desenhada para parear com a citosina (metilado) ou com a timina (não metilado), como pode ser visto na Figura 1.5.a. O ensaio experimental Infinium II, por sua vez, utiliza somente uma sonda que complementa a base anterior do sítio em questão na terminação 3' da sonda, enquanto uma extensão de base única resulta na adição de uma base G (guanina) ou A (adenina) marcada com fluoróforo, complementar à C (citosina) "metilada" ou à T (timina) "não metilada" (Figura 1.5.b). Para ambos os experimentos, a medida de metilação é calculada com base na intensidade da fluorescência dos corantes verde (Cy3) e vermelho (Cy5) (Bibikova *et al.* [2011]; Pidsley *et al.* [2013]).

Algumas características relacionadas às sondas, tais como **(a)** o espaçamento não regular das sondas, que não estão em posições equidistantes ao longo do genoma; **(b)** a não uniformidade das sondas, pois não existe definição para o número de sondas que representam promotores ou quaisquer outras regiões definidas; e **(c)** a heterogeneidade entre o número e as representações genômicas usando Infinium I e II, são consideradas nas pré-análises de experimentos com a lâmina Illumina Infinium HumanMethylation450, pois podem interferir nas análises e devem ser corrigidas por meio do uso de um ou mais pipelines existentes, resultando em uma matriz com o nível de metilação para cada sítio em cada amostra, que servirá como ponto de partida para análises subsequentes (Bock [2012]).

1.5 Análise e interpretação dos dados de metilação do DNA

As tecnologias para o mapeamento de metilação de DNA estão se tornando cada vez mais disponíveis em laboratórios não especialistas, devido a uma necessidade crescente de ferramentas computacionais de fácil utilização que permitam a manipulação e a análise de grandes conjuntos de

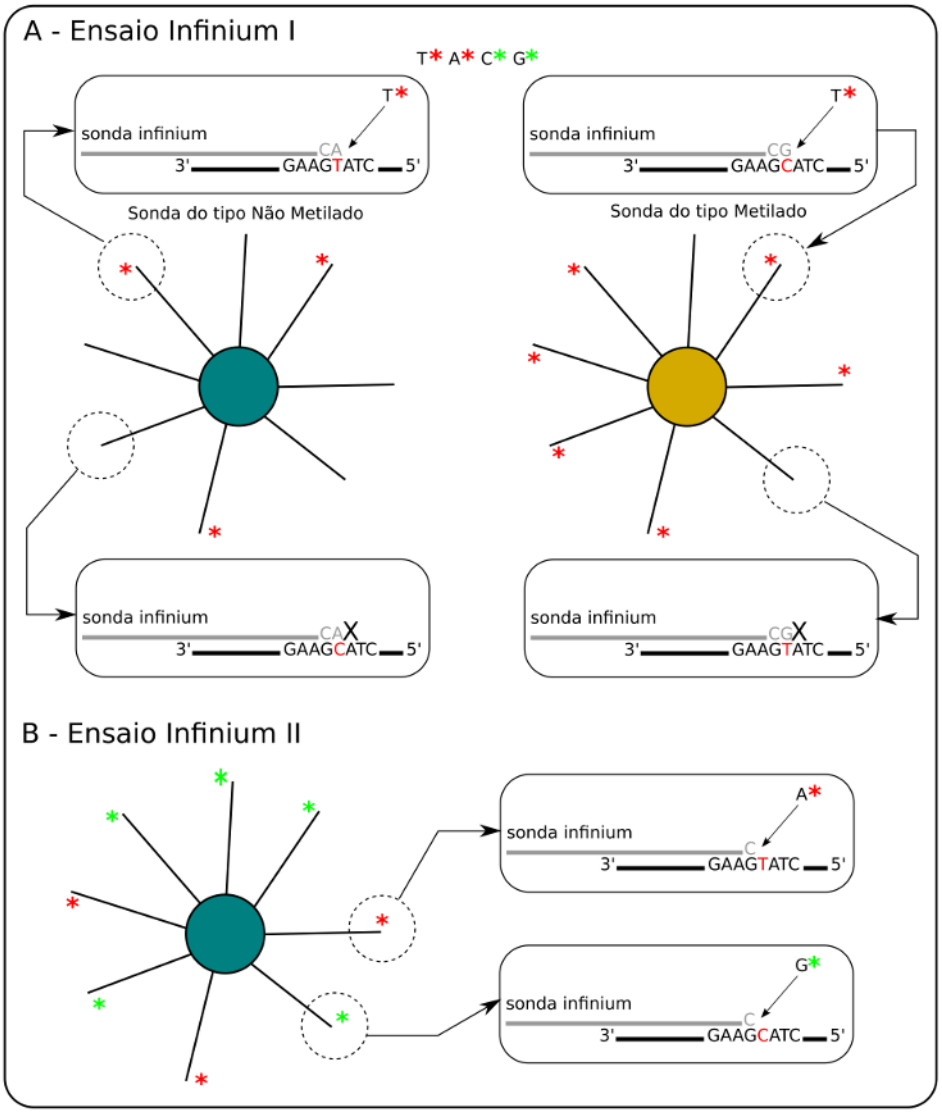


Figura 1.5: *Visão geral dos ensaios Infinium I e Infinium II. (A) Infinium I e (B) Infinium II, presente no Infinium Methylation 450K array. M: Metilado; U: Não-metilado. Adaptada de "Evaluation of the Infinium Methylation 450K technology", Dedeurwaerder et al. [2011].*

dados de metilação do DNA (Bock [2012]). Em função das diversas singularidades apresentadas pela lâmina Illumina Infinium HumanMethylation450, diversas metodologias e vários fluxos de análise foram publicados e revisados por Bock [2012], Dedeurwaerder *et al.* [2014], Pidsley *et al.* [2013] e Morris e Beck [2014]. Como em outros tipos de microarranjos (por exemplo, para a genotipagem e os perfis de transcrição), o processamento de dados da Infinium HumanMethylation450 compreende o processamento de imagem e a normalização de dados como os seus principais passos (Bock [2012]). O processamento de imagem é realizado normalmente por meio do software Illumina BeadScan® fornecido pela empresa Illumina. Já para normalizar os valores de intensidade de cada sonda e para inferir os níveis absolutos de metilação do DNA existem diversas opções (Bock [2012]). Uma destas opções consiste na utilização do software comercial Illumina GenomeStudio (Bock [2012]), assim como no uso das funções de extração de dados contidos nos pacotes MINFI (Morris e Beck [2014]) e Methylumi Davis e Bilke [2012], caso em que os dados de fluorescências são convertidos em uma medida denominada β.

O valor β, Fórmula 1.1, é calculado pela razão do sinal de fluorescência do alelo metilado sobre a soma do sinal de fluorescência do alelo metilado com o sinal de fluorescência do alelo não metilado, podendo variar entre 0 e 1 (0 para não metilado e 1 para totalmente metilado). Trata-se de uma medida absoluta do nível de metilação do DNA facilmente interpretada, embora imponha sérios desafios quando aplicada a muitos modelos estatísticos (Bibikova *et al.* [2006]; Siegmund *et al.* [2012]; Du *et al.* [2010]; Wilhelm-Benartzi *et al.* [2013]).

$$\beta = \frac{metilado}{metilado \; + \; não\, metilado} \tag{1.1}$$

Uma proposta alternativa é utilizar o valor M (Irizarry *et al.* [2008]; Du *et al.* [2010]; Zhuang *et al.* [2012]), Fórmula 1.2, adaptado dos métodos de análise de expressão de mRNA em microarranjo, sendo calculado a partir do log_2 da razão entre o metilado sobre o não metilado. O valor de M é considerado estatisticamente mais válido em testes estatísticos utilizados em estudos de expressão de genes e de metilação devido à sua natureza mais homocedástica. Apesar disso, o valor M não é diretamente interpretável em termos de uma porcentagem de metilação, tal como o valor β (Du *et al.* [2010]; Zhuang *et al.* [2012]).

$$M = log_2 \frac{metilado}{não\, metilado} \tag{1.2}$$

1.6 Relação entre valor β e valor M

Ao verificar o valor das intensidades dos sítios CpGs (metilado + não metilado) da lâmina Illumina Infinium HumanMethylation27, Du *et al.* [2010] constataram que aproximadamente 95% dos sítios CpGs possuem valores de intensidade acima de 1000. A relação entre os valores β e M pode ser deduzida pelo uso das Fórmulas 1.3. Além disso, Du *et al.* [2010] observaram a média e o desvio padrão em relação ao valor β e M da lâmina Illumina Infinium HumanMethylation27 (Figura 1.6), que foram calculados utilizando amostras de replicatas técnicas. O desvio padrão do valor β é mais comprimido quando está entre os intervalos 0 a 0.2 e 0.8 a 1. Isso significa que o valor β possui significante heterocedasticidade em intervalos baixos (0 a 0.2) e altos (0.8 a 1) de metilação.

A heterocedasticidade pode ser resolvida após a conversão do valor β em valor M utilizando a Fórmula 1.3. O valor M apresenta homocedasticidade por manter o desvio padrão aproximadamente constante ao longo do intervalo dos valores M. Por esse motivo, o valor M é mais válido para abordagens estatísticas em que se assume a homocedasticidade dos dados.

$$\beta = \frac{2^M}{2^M + 1} \quad M = log_2 \frac{\beta}{1 - \beta} \tag{1.3}$$

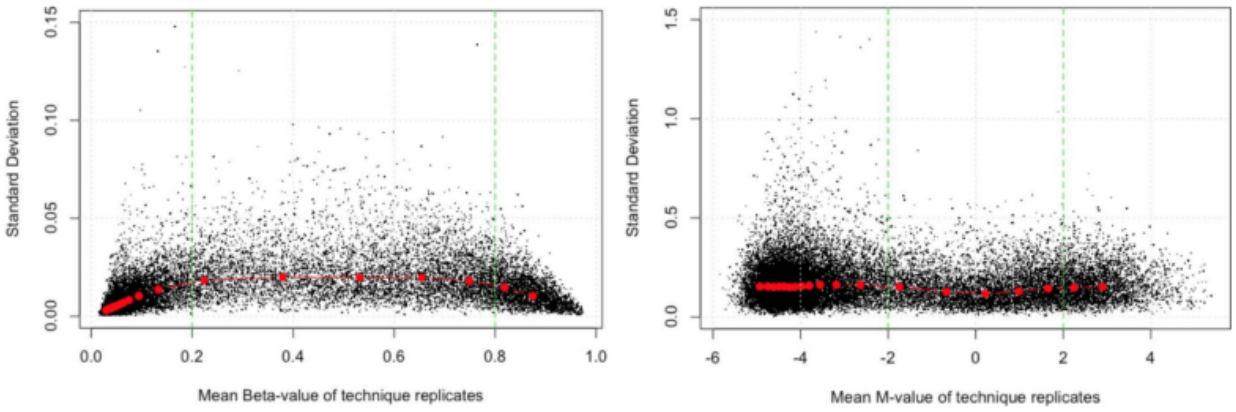


Figura 1.6: *Gráfico da média e o desvio-padrão das relações de replicatas técnicas. Valor β (à esquerda) e valor M (à direita). Figura retirada de "Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis.", Du et al. [2010].*

1.7 Fluxo de análise

Etapas importantes devem ser incluídas na pipeline de análise da Infinium HumanMethylation450, como normalização, análise de efeito lote, polimorfismo de nucleotídeo único (SNP) de sinalização, detecção de aberrações do número de cópia (CNAs) e segmentação de posições variáveis de metilação (MVPs) em DMRs biologicamente relevantes (Morris *et al.* [2014]). Como destaca Bock [2012], o fluxo de análise apresenta as seguintes etapas observadas resumidamente na Figura 1.7

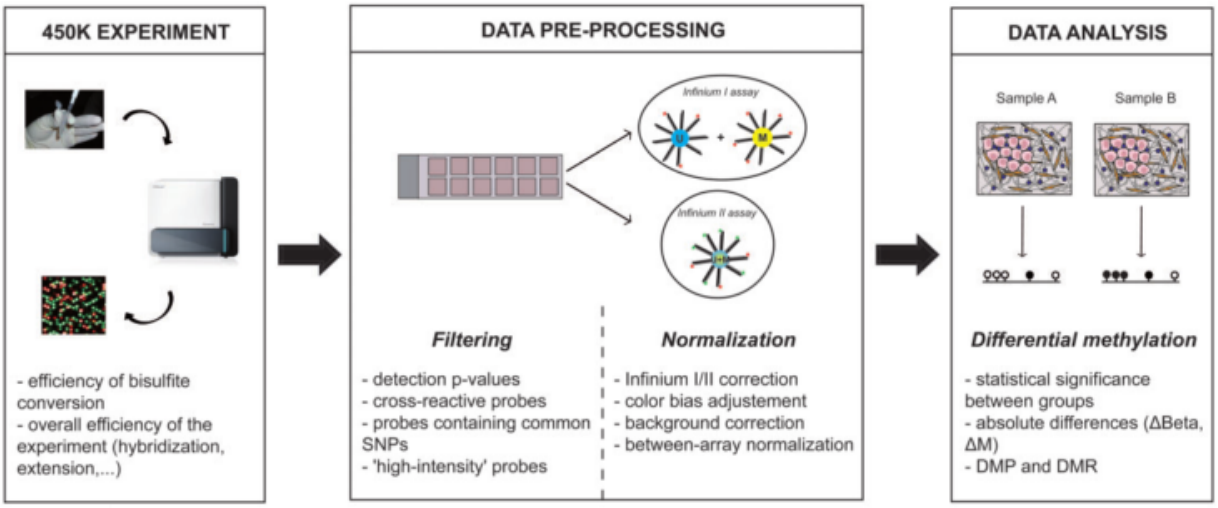


Figura 1.7: *Visão geral do fluxo de analise realizado do pré processamento a interpretação do resultado obtido. Figura retirada de "A comprehensive overview of Infinium HumanMethylation450 data processing", Dedeurwaerder et al. [2014].*

1.8 Importação, qualidade e pré-processamento

Os dados brutos de metilação do DNA (contidos em arquivos IDAT são importados usando o illuminaio (Smith (2013)), que é uma ferramenta para extração do sinal de intensidade de fluorescência implementada no pacote MINFI (Aryee *et al.* [2014]). Diversos pipelines realizam a etapa de pré-processamento, sendo necessários alguns estágios para preparar os dados antes da realização das análises estatísticas, como a filtragem de sondas que não hibridizaram ou que hibridizaram

com baixa qualidade, a correção de viés biológico e técnico e a normalização para eliminar potenciais efeitos de lote (Dedeurwaerder *et al.* [2014], Morris e Beck [2014]). O processo de controle de qualidade envolve uma série de métricas para determinar o sucesso da conversão de bissulfito e a subsequente hibridação do microarranjo (Morris e Beck [2014]). A filtragem de sonda é realizada para remover as sondas **(i)** que não conseguiram hibridar, **(ii)** que não são representadas por um mínimo de três beads no microarranjo, **(iii)** que se sobrepõem a polimorfismos de nucleotídeo único (Single Nucleotide Polymorphism - SNPs) e **(iv)** que apresentam mapeamento em mais de uma localização genômica (Morris e Beck [2014]).

Há 65 sondas na lâmina Infinium HumanMethylation450 BeadChip que são SNPs altamente polimórficos (Pidsley *et al.* [2013]). Para interpretar adequadamente os dados de metilação destes CpGs polimórficos, faz-se necessário um conhecimento a priori de cada genótipo individual. Eles poderiam ser afetados por polimorfismo genético, mas a maioria desses SNPs é rara e apresenta frequências muito baixas de alelos alternativos. Desse modo, não é esperado ter um efeito importante sobre os dados de metilação quando a população em estudo não demonstra uma frequência significativa do alelo raro (Chen *et al.* [2012]).

O viés técnico pode ocorrer por diferentes razões, tais como por um escaner mal calibrado, por problemas de leitura do sinal de algumas sondas com baixa intensidade ou por algum distúrbio com o microarranjo. Este problema é traduzido por uma alta detecção de p-valor, sendo altamente recomendada a filtragem destas sondas (Dedeurwaerder *et al.* [2014]). Para aferir a qualidade das amostras, gráficos de diagnóstico de sondas podem ser gerados no GenomeStudio® (Bibikova *et al.* [2011]) ou no pacote HumMethQCReport (Mancuso *et al.* [2011]) do software R, que oferece um conjunto de opções gráficas para mensurar a qualidade das amostras (Wilhelm-Benartzi *et al.* [2013]). Outras opções para o controle de qualidade, assim como o pipeline de pré-processamento e análise, estão disponíveis em pacotes do Bioconductor (Gentleman [2004]) para o software R, como o Complete pipeline for Infinium® Human Methylation 450K de Touleimat e Tost [2012], o IMA de Wang *et al.* [2012], o MINFI de Aryee *et al.* [2014], o MethyLumi de Davis e Bilke [2012], o wateRmelon de Pidsley *et al.* [2013], o RnBeads de Assenov *et al.* [2009] e o ChAMP de Morris *et al.* [2014] (Bock [2012]). Foi observado, por meio dos testes entre as formas de pré-processamento divulgados, que a seleção cuidadosa de passos de pré-processamento pode minimizar a variância e, assim, melhorar o poder estatístico, especialmente para a detecção de pequenas variações de dados de metilação do DNA (Pidsley *et al.* [2013]).

1.9 Correção

Após os passos que garantem a qualidade dos dados, é necessário corrigi-los. Diversos métodos realizam a correção de fundo, o ajuste de bias do corante e o ajuste da Infinium II (Bock [2012]; Morris e Beck [2014]). Apesar do uso de algoritmos de correção para reduzir artefatos técnicos, várias fontes de viés tendem a persistir. É necessário, desse modo, analisar os dados para potenciais efeitos de lote, que são uma fonte comum de variação em experimentos de alto rendimento. Tais efeitos representam medições relacionadas com condições que não são variáveis biológicas ou científicas do estudo (ou seja, data de experimento, chip ou instrumento utilizado, lote de reagentes empregado, técnico responsável pelas análises etc.) (Morris e Beck [2014]). Estas variações podem afetar as etapas de análise subsequentes se não forem tomadas medidas adequadas (Leek *et al.* [2010]; Teschendorff *et al.* [2011]; Bock [2012]; Morris e Beck [2014]). Esse problema faz com que seja difícil distinguir entre variação técnica indesejável e diferenças biológicas significativas, sendo aconselhável, portanto, processar as amostras para minimizar a variação entre as fontes potenciais de efeitos em lote por meio, por exemplo, do uso de um design cuidadoso do estudo (Bock [2012]; Morris e Beck [2014]). Os métodos mais comuns para correção do efeito de lote são o método supervisionado de correção ComBat (Johnson *et al.* [2007]), um método Bayesiano empírico que estima os parâmetros para localização e ajuste de escala de cada lote para cada gene de forma independente (Chen *et al.* [2011]); a análise de variável substituta (Surrogate Variable Analysis - SVA) Leek *et al.* [2012]; e a análise de variável substituta independente (Independent Surrogate

Variable Analysis - ISVA) (Teschendorff *et al.* [2011]) (Morris e Beck [2014]).

Outras possíveis correções são necessárias dependendo da origem da amostra e do que o estudo se propõe a responder. A variação do número de cópias (Copy Number Variation - CNV) é um tipo de variação genética amplamente encontrada em genomas de mamíferos. Sabe-se que um número substancial de CNV tem impacto significativo sobre as doenças humanas (Zhang *et al.* [2014]). O CNV possui influência sobre a metilação em quaisquer regiões com perdas de heterozigotos ou com ganhos de cópia única quando comparadas com as regiões de número de cópias normais (Feber *et al.* [2014]; Houseman *et al.* [2009]). A correção estatística para aberrações do número de cópias visa corrigir possível viés sobre o valor de metilação que pode influenciar os resultados de detecção de metilação diferencial (Robinson *et al.* [2012]).

Deve-se, também, considerar a correção para a heterogeneidade celular (Irizarry, 2014; Jaffe e Irizarry [2014]; Morris e Beck [2014]). O sangue, tal como muitos outros tecidos, é constituído de diferentes tipos de células, cada uma com distintos perfis de metilação que podem variar em proporção com a idade ou o estado da doença (Morris e Beck [2014]; Reinius *et al.* [2012]; Houseman *et al.* [2012]; Houseman *et al.* [2015]) desenvolveu um processo utilizando dados de referência da Infinium HumanMethylation450 para estimar as proporções relativas dos diferentes tipos de células no sangue e corrigir o dado de metilação. Este processo foi incorporado nos pacotes MINFI e ao RnBeads (Morris e Beck [2014]).

1.10 Normalização

A normalização é uma etapa especialmente importante na plataforma Infinium HumanMethylation450, por combinar dois ensaios diferentes: a Infinium I e a Infinium II na mesma lâmina (Bibikova *et al.* [2011]; Sandoval *et al.* [2011]). Uma série de métodos de normalização estão disponíveis para lidar com esta questão de forma ligeiramente diferente (Marabita *et al.* [2013]; Yousefi *et al.* [2013]; Bock [2012]; Siegmund [2011]). Em ordem cronológica de desenvolvimento, há: correção baseada em pico (Peak Based Correction - PBC) (Dedeurwaerder *et al.* [2011]), SQN (Touleimat e Tost [2012]), subconjunto-quantile interno de vetor normalização (SWAN) (Makismovic *et al.* [2012]) e beta-mistura quantile normalização (BMIQ) (Teschendorff *et al.* [2013]) (Morris *et al.* [2014]). O software comercial Illumina GenomeStudio fornece um algoritmo básico para a normalização do sinal e a subtração de fundo usando sondas de controle positivas e negativas. Um algoritmo semelhante foi implementado no R / Bioconductor como parte dos pacotes open-source MINFI e methylumi (Bock [2012]). Pidsley *et al.* [2013] alerta para o fato de que, para investigações de diferenças sutis, como as observadas em doenças complexas comuns, tais como a esquizofrenia (Dempster *et al.* [2011]; Kinoshita *et al.* [2013]) e a diabetes (Rakyan *et al.* [2011]), existe a necessidade de assegurar a máxima sensibilidade para detectar a metilação de DNA diferencial Pidsley *et al.* [2013]. Estudos recentes demonstram que os passos de normalização mais sofisticados podem melhorar a qualidade dos dados e reduzir a variação técnica (Dedeurwaerder *et al.* [2011]; Touleimat e Tost [2012]; Wang *et al.* [2015]; Bock [2012]). Uma abordagem pragmática para as limitações de simples métodos baseados na relação para calcular os valores de metilação do DNA, comuns na literatura (Sun *et al.* [2011]), tem sido a de normalização quantílica (Quantile Normalization - QN) do valor β. A QN é uma técnica bem estabelecida na análise da expressão de genes, devido ao seu bom desempenho (Irizarry *et al.* [2003]). Para dados de microarranjos de múltiplas amostras formatados como uma matriz com uma coluna por amostra e uma linha por característica, a QN é uma transformação não linear que substitui cada sinal de intensidade pela média das características com o mesmo valor de cada matriz. Um ponto fraco potencial da QN é que, em partes da distribuição com alguns valores (havendo, portanto, relativamente grandes diferenças interquartil), pode introduzir alterações consideráveis, de modo que essas grandes mudanças poderiam aumentar a variância por meio amostras de características individuais, ao invés de reduzi-la conforme desejado (Pidsley *et al.* [2013]). O resultado da normalização de dados é uma tabela de valores β e, opcionalmente, de valores M, que serve como o ponto de partida para análises posteriores (Bock [2012]).

1.11 Detecção de diferenças no padrão de metilação

Após as etapas de pré-processamento, o passo seguinte é a detecção das posições diferencialmente metiladas (differentially methylated positions - DMPs), que são CpGs individuais diferencialmente metiladas em tecidos específicos, e a detecção das regiões diferencialmente metiladas (differentially methylated regions - DMRs), que são regiões que contêm ao menos três DMPs, com distância <= 1kb, interrompidas por no máximo três não DMPs (Slieker *et al.* [2013]).

A forma mais básica de detecção de uma DMP consiste na utilização do teste t ou do teste de Wilcoxon, que comparam os níveis de metilação do DNA de cada citosina com dados suficientes entre dois grupos amostrais (Wang *et al.* [2012]). Vários métodos mais avançados têm sido descritos que visam melhorar a detecção de DMPs, como modelos de misturas (Wang [2012]), Shannon entropia (Zhang *et al.* [2011]), valor logístico M (Du *et al.* [2011]), seleção de características (Zhuang *et al.* [2012]), estratificação do teste t (Chen *et al.* [2012]), agregação de regiões genômicas por tipo (Poage *et al.* [2013]) ou regressão linear em combinação com a remoção do efeito de lotes e detecção de pico (Jaffe *et al.* [2012]).

Além disso, foram propostos métodos alternativos para aumentar o poder estatístico, modelando a dependência entre os testes estatísticos realizados para CpGs vizinhos (Kuan e Chiang [2012]; Bock [2012]). As comparações estatísticas podem ser direcionadas para regiões genômicas maiores, em vez de CpGs individuais (DMP), de tal modo que CpGs vizinhas com diferenças na metilação do DNA semelhantes reforcem umas às outras e deem origem a resultados mais significativos. Tais regiões são denominados de regiões diferencialmente metiladas (DMR) e são sequências genômicas discretas que possuem uma assinatura distinta através de um número de CpGs (e/ou não CpGs), capazes de distinguir um fenótipo de outro. Sua identificação e utilidade têm sido de longo alcance nas aplicações clínicas, porque o uso das DMRs reduz a escala do genoma para um número menor de regiões. Além disso, uma vez que as DMRs forem validadas e replicadas, abre-se o caminho para a redução de tempo, de custos e de esforço, possibilitando a realização de ensaios experimentais mais eficazes que irão melhorar os estudos funcionais posteriores e fornecer ferramentas de diagnóstico médico (Butcher e Beck [2014a]).

Embora a identificação de DMRs seja mais facilmente feita entre dois grupos, análises mais complexas podem ser projetadas (Morris e Beck [2014]). As DMRs podem ser pequenas, compostas de poucas CpGs, ou tão grandes quanto um locus de gene inteiro, dependendo do interesse biológico envolvido e dos métodos de bioinformática utilizados para a sua identificação. Embora um único CpG metilado possa, ocasionalmente, ser ligado à regulação da expressão do gene (Xu *et al.* [2007]) e possa afetar o risco à doença (Raval *et al.* [2007]; Moser *et al.* [2009]), a grande maioria das DMRs relatadas na literatura está dentro de uma gama de tamanhos de algumas centenas a algumas milhares de bases. Esta gama coincide com os tamanhos típicos de regiões reguladoras de gene, e acredita-se que DMRs podem controlar a repressão transcricional de um tipo específico de célula de um gene-relacionado (Bird [2002]; Jones [2012]; Deaton e Bird [2011]; Bock [2012]).

Este método pode ser aplicado ao genoma ou pode ser focado sobre um conjunto previamente selecionado de regiões genômicas candidatas (Bock *et al.* [2010]) – a última abordagem aumenta substancialmente o poder estatístico para detectar essas DMRs que estão localizados entre as regiões candidatas, mas à custa de perder DMRs em outros lugares (Bock [2012]). Uma abordagem simples seria contar as CpGs significativas em uma janela deslizante de tamanho fixo. Dessa forma, uma DMR poderia ser definida se uma janela (ou janelas contíguas) de determinado tamanho capturasse um número determinado de sondas associadas significativamente. No entanto, esta abordagem é controversa, devido à distribuição das CpGs e ao risco de restringir as DMRs para somente regiões com maior número de sondas (Butcher e Beck [2014a]). Há uma série de métodos que utilizam esta abordagem e que podem ser aplicados à lâmina 450K. Estes métodos são: LASSO (Butcher e Beck [2014a]), Bump Hunting (Jaffe *et al.* [2012]), Block Finding (Hansen *et al.* [2011]), AClust (Sofer *et al.* [2013]) e DMRcate (Peters *et al.* [2014]).

1.12 Múltiplas testagens

É importante, para qualquer método estatístico que teste as diferenças na metilação do DNA em um grande número de loci genômicos, aplicar uma correção para múltiplos testes de hipóteses. Esta correção é feita quase exclusivamente por meio do controle da taxa de falsa descoberta (false discovery rate - FDR). Para isso, a distribuição de p-valores não corrigidos é analisada, e uma FDR é inferida de cada DMR. A inferência é, muitas vezes, feita usando o método do valor Q disponível no R / Bioconductor (Storey e Tibshirani [2003]). Por causa do grande número de CpGs no genoma, somente as fortes diferenças de único CpG tendem a permanecer significativas após várias correções de teste. O resultado é muitas vezes uma taxa de falso-negativo elevado, especialmente quando o número de amostras e tamanhos de efeito é pequeno (Bock [2012]).

Capítulo 2

Justificativa e objetivos

2.1 Justificativa

O estudo da epigenética está em rápida expansão. Nesse contexto, a metilação do DNA é uma das áreas que mais avançam em virtude dos avanços tecnológicos que tornaram possível a realização de estudos em grande escala, como a aprimoração do uso de microarranjos e o seu baixo custo em relação a outros métodos de detecção de metilação. No entanto, a determinação do valor absoluto da metilação por meio do uso de microarranjo tem sido problemática, devido a erros sistemáticos e a variabilidades indesejadas. Os procedimentos para pré-processamento estão em melhoramento contínuo, e as correções do dado são incorporadas à medida que se descobrem novas variáveis a partir do tipo de amostra que influenciam o valor da metilação. Cada pipeline possui metodologias e técnicas diferentes, que envolvem muitos passos a serem seguidos, gerando resultados diferentes e não comparáveis entre si (Morris e Beck [2014]; Li *et al.* [2015]; Bock [2012]). Os pipelines desenvolvidos para determinar sítios diferencialmente metilados (DMP) em grupos de interesse contam com diferentes métodos para realizar esta tarefa, mas esbarram em problemas relacionados ao valor de metilação e ao design do microarranjo empregado (Morris e Beck [2014]; Li *et al.* [2015]; Bock [2012]). Muitos dos métodos utilizados para a detecção de DMP confiam em hipóteses irrealistas, assumindo, eventualmente, a normalidade das amostras, apresentando comportamento assintótico de algumas estatísticas, analisando razoavelmente as amostras de grandes dimensões, realizando a aproximação dos cálculos, possuindo a limitação de comparações de somente duas populações e equivalência no número de amostras nos grupos, necessitando de correção para múltiplos testes (para evitar a ocorrência de resultados significativos por acaso) e não tendo nenhuma evidência para a hipótese nula, estes problemas podem ser agravados se os conjuntos de dados forem relativos a somente alguns pacientes.

O método QUOR é uma alternativa aos outros métodos, que foi elaborada para o uso em estudos de expressão gênica e que pode ser também aplicada para a análise de metilação. Permite o uso de um número diferente de amostras e de dados faltantes e ainda possibilita comparar corretamente estas variáveis. Ademais, todos os cálculos são realizados de forma exata, sem qualquer suposição assintótica (Pereira *et al.* [2012]). Assim, o uso do QUOR para a análise de metilação não terá os problemas que análises estatísticas para diferenças entre grupos possuem. O QUOR é um método não paramétrico para amostras independentes (Pereira *et al.* [2012]), tendo como objetivo comparar a mediana entre dois ou mais grupos e classificá-los. A hipótese estatística a ser testada é de que a mediana do grupo X é menor que a mediana do grupo Y, isto é, trata-se de averiguar qual é a probabilidade de que X seja menor ou maior que Y, admitindo que X e Y sejam variáveis aleatórias. Exemplificando: é possível ter duas populações e avaliar os valores medianos de uma variável que representa o nível de expressão de um gene particular. O objetivo consiste, assim, em obter a confiança de que "a mediana da expressão desse gene na primeira população é estritamente menor (ou maior) do que a mediana da expressão do mesmo gene na segunda população" (Pereira *et al.* [2012]).

Da mesma forma, usado para a análise de expressão gênica, o QUOR pode sugerir que uma

CpG está hipometilada, hipermetilada ou que não há nenhuma diferença significativa para esta CpG entre as populações (Pereira *et al.* [2012]). O método proposto utiliza intervalos de confiança não paramétricos para quantis com base na distribuição binomial. Seu objetivo é calcular um valor de confiança que indica o quanto se acredita que os parâmetros quantílicos de diferentes populações/grupos são ordenados entre si (Pereira *et al.* [2012]). Entre as propriedades do QUOR, destaca-se sua natureza não paramétrica, a possibilidade de processar conjuntos de dados com valores faltantes e com variação do número de amostras de cada população (entre vários testes) e o fato de que o corte e a interpretação da confiança podem evitar a necessidade de correção para múltiplos testes (Pereira *et al.* [2012]).

2.2 Objetivos

2.2.1 Objetivo geral

Comparar métodos de análise de metilação diferencial usando a lâmina Infinium HumanMethylation450 BeadChip com método QUOR.

2.2.2 Objetivos específicos

Comparar as diferentes formas de análise sítio-específico de metilação diferencial e correção para múltiplos testes, ou seja, empregar diferentes técnicas estatísticas a partir da distribuição dos dados amostrais e das hipóteses do trabalho.

Usar diferentes conjuntos de dados com números amostrais diferentes para realizar as comparações entre os diferentes métodos de análise.

Capítulo 3

Materiais e métodos

3.1 Obtenção dos dados

Foram utilizados dois conjuntos de dados para os testes realizados neste trabalho. A seguir, é descrito cada um destes conjuntos.

3.1.1 Amostra 1 COCAINA/CRACK

A amostra 1 consiste de um conjunto de dados obtido a partir de amostras de sangue periférico de 24 casos de abuso/dependência de cocaína e crack e de 24 casos de controle, provenientes do banco de DNA do Programa de Genética e Farmacogenética (ProGene) do Instituto de Psiquiatria do Hospital das Clínicas da Faculdade de Medicina da Universidade de São Paulo (Ipq-HCFMUSP). O conjunto de dados Infinium HumanMethylation450 BeadChip deste estudo foi submetido ao NCBI Gene Expression Omnibus sob o acesso GSE77056.

3.1.2 Amostra 2 GRAVIDAS – estudo do impacto do estresse mãe-bebê

A amostra 2 consiste em um conjunto de dados de sangue de cordão umbilical, oriundos do projeto INPD realizado pela Faculdade de Medicina da Universidade de São Paulo (FMUSP) em parceria com o Hospital Universitário da Universidade de São Paulo (HU-USP). Este conjunto de dados possui 70 amostras de sangue de cordão, analisadas por meio da lâmina Illumina HumanMethylation450 Beadchip, contendo 30 casos e 40 controles.

3.2 Pré-processamento

Importação dos dados, filtragem de sondas, conversão em β ou M e correções O pré-processamento engloba os seguintes passos: importação dos dados dos arquivos IDAT; filtragem de sondas (remoção de sondas que falham ao hibridizar por apresentar p-valor alto de detecção, de sondas que não possuem no mínimo três beads na lâmina, de sondas relacionadas a SNP e de sondas que hibridizam em múltiplas localizações do genoma); remoção de sondas relacionadas ao cromossomo XY, em dependência do objetivo do estudo; e conversão dos valores de intensidade do sinal em valores β ou M.

As correções aplicadas correspondem ao ajuste entre os ensaios Infinium I e Infinium II.

Para dados de origem sanguínea, empregou-se a correção por composição celular. Neste caso, foi utilizado o método EstimateCellCounts contido no pacote Minfi (Aryee *et al.* [2014]) para calcular a composição celular, conforme definido por Houseman *et al.* [2012]. Utilizou-se, então, o valor padrão para cada parâmetro deste método.

3.3 Normalização

Para a normalização do dado, foi usado o método BMIQ, implementado no pacote ChAMP (Morris *et al.* [2014]) e encontrado, também, nos pacotes RnBeads (Assenov *et al.* [2009]) e wateRmelon (Pidsley *et al.* [2013]).

3.4 Análise e correções de efeitos de lote

Para a análise e correção de possíveis efeitos de lote, foram empregados, respectivamente, o método SVD e o método ComBat.

3.5 Identificação de posições diferencialmente metiladas

Com o objetivo de comparar as diferentes abordagens para a análise de metilação entre as pipelines, foram usados os seguintes métodos e pacotes:

- teste empírico de Bayes, implementado no pacote Limma e contido nas pipelines ChAMP, RnBeads e IMA;
- teste t Student, contido nas pipelines RnBeads e IMA;
- teste Wilcoxon, contido na pipeline IMA.

Outro interesse deste estudo consiste em comparar o método QUOR aos métodos supracitados.

Capítulo 4

Resultados

O estudo foi conduzido para a comparação entre o número de CpGs detectadas com o uso de quatro tipos de análises estatísticas para avaliar a diferença entre os grupos no dois conjuntos de dados. Os testes utilizados foram: teste t Student, teste Wilcoxon, teste Empírico de Bayes e método QUOR. Além de comparar o número de DMPs detectadas da intersecção dos três testes com o QUOR, também foram comparadas as diferenças entre os resultados obtidos, utilizando o valor β e o valor M. Para cada estudo, os dados foram submetidos ao mesmo fluxo de análise.

4.1 Pré-processamento, normalização e correção do dado

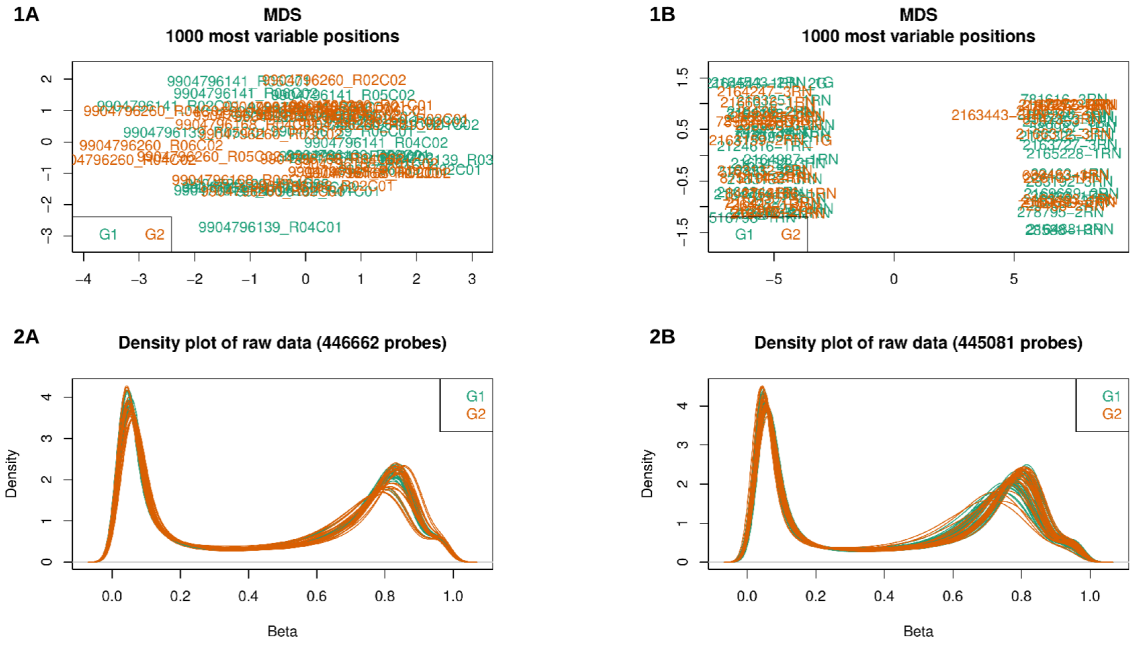
No processo de importação e filtragem de sondas, das 485.577 sondas iniciais, foram mantidas 433.975 sondas e 48 amostras do conjunto de dados COCAÍNA/CRACK. E foram mantidas 445.801 sondas e 70 amostras do conjunto de dados GRAVIDAS, a Tabela 4.1 apresenta a quantidade de sondas removidas por tipo de filtragem aplicada durante o pré-processamento. Utilizandos os valores de β não normalizados não é possível separar em dois grupos distintos, caso e controle, dos conjuntos de dados COCAÍNA/CRACK e GRAVIDAS, através do escalonamento multidimensional (Multidimensional scaling - MDS), visto nas Figuras 4.1.1A e Figura 4.1.1B. É observável que os valores β ainda não normalizados, não atingem o valor máximo, estando distribuídos próximos do valor 0.8, este é um problema na lâmina Illumina Infinium HumanMethylation450, por haver dois experimentos diferentes para detecção da metilação (Dedeurwaerder *et al.* [2011]; Touleimat e Tost [2012]; Morris e Beck [2014]; Teschendorff *et al.* [2013]), Figura 4.1.2A e Figura 4.1.2B.

Conjunto de dados	p valor de detecção > 0.01	número de beads < 3	relacionado a SNP	relacionado a alinhamentos múltiplos	relacionado ao cromossomo X e Y
Cocaína/Crack	1272	284	28792	8502	11504
Gravidas	3133	297	28508	8486	11108

Tabela 4.1: *Quantidade de sondas removidas por conjunto de dados, cada coluna é um filtro aplicado ao conjunto de dados.*

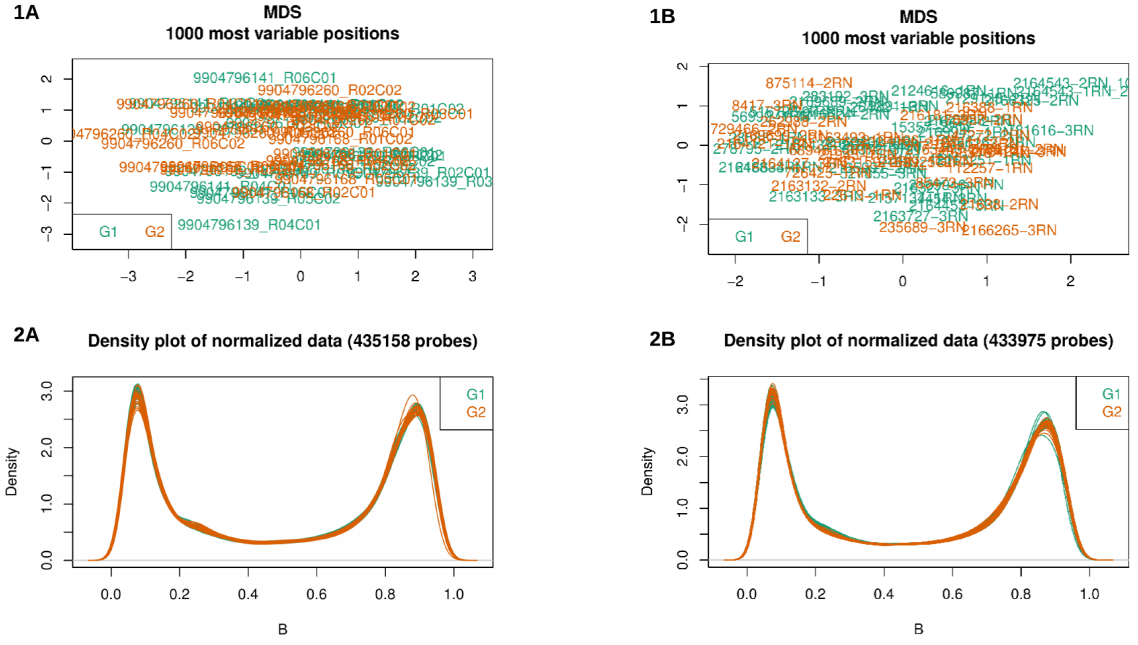
Após a normalização, os dois conjuntos de dados ainda não definem dois grupos distintos, Figura 4.2.1A e 4.2.1B. A normalização corrigiu o problema, fazendo com que o valor β possa alcançar valores mais altos, como também aumentou sua concentração em um só valor, não mais distribuídos como visto anteriormente, Figura 4.2.2A e 4.2.2B.

Na Figura 4.3, é observado o valor β é mais heterocedástico do que o valor M, apresentando baixa variabilidade nos valores mais baixos e altos de β, enquanto o valor M, sendo este mais homocedástico, possui um certo equilíbrio entre a variabilidade. Esta observação também foi demonstrada e explicada por Du *et al.* [2010] e Zhuang *et al.* [2012].



(a) Conjunto de dados COCAÍNA/CRACK (b) Conjunto de dados GRAVIDAS

Figura 4.1: *Agrupamento através do escalonamento multidimensional e distribuição dos valores de β sem normalização para cada sonda*



(a) Conjunto de dados COCAÍNA/CRACK (b) Conjunto de dados GRAVIDAS

Figura 4.2: *Agrupamento através do escalonamento multidimensional e distribuição dos valores de β com normalização para cada sonda*

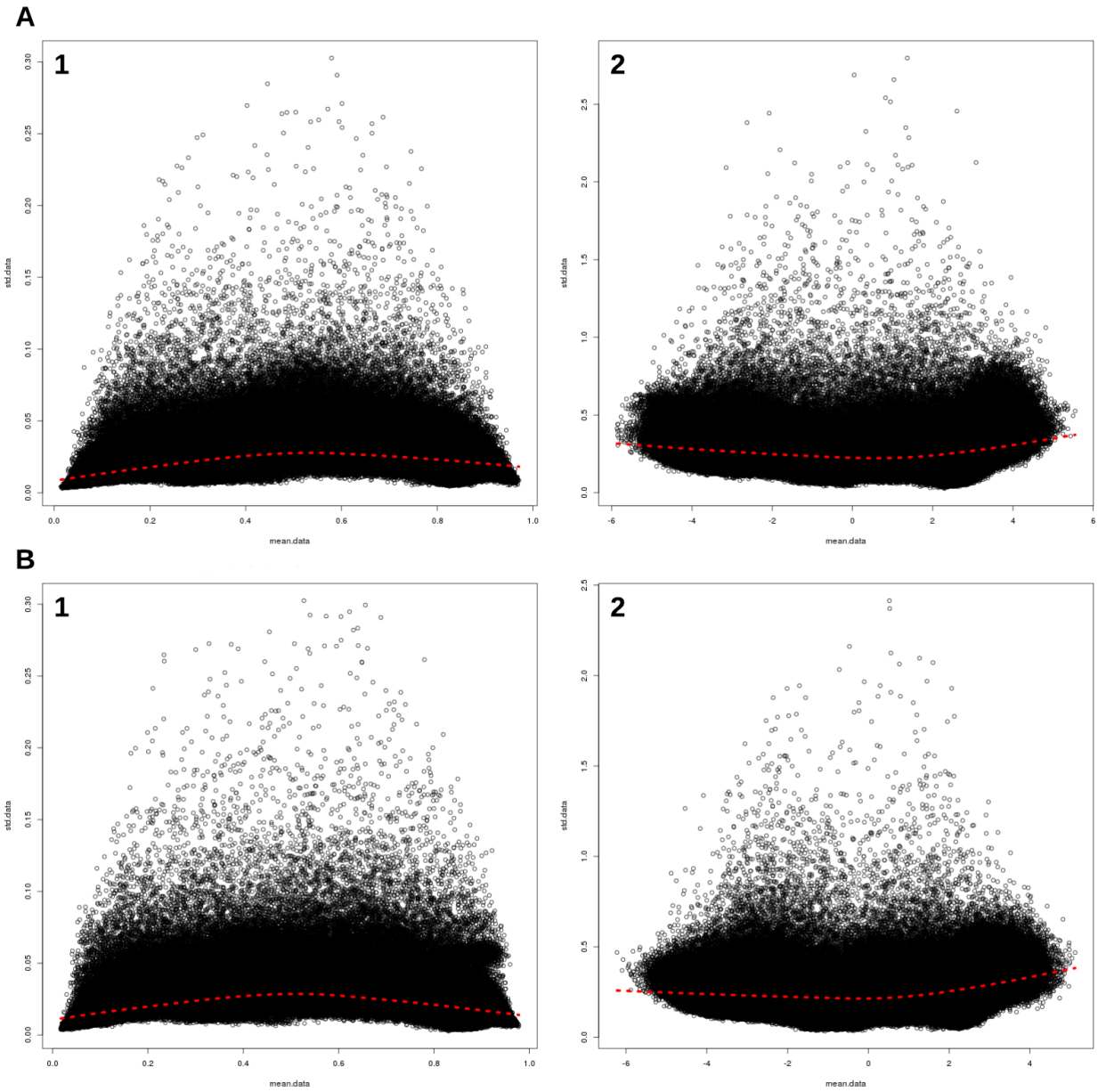


Figura 4.3: *Homoscedaticidade e heteroscedasticidade entre valor M e valor β.*

4.2 Identificação das DMPs

Denominamos como **"Paramétrico"** o teste utilizando t Student, como **"Não paramétrico"** o teste realizado com Wilcoxon e como **"Bayesiano"** o teste Empirico de Bayes, estes testes identificaram um número aproximado de DMPs entre elas e uma alta intersecção entre os seus valores β e M do conjunto de dados COCAÍNA/CRACK, Tabela 4.2. Para o conjunto de dados maior GRAVIDAS, a quantidade de DMPs detectadas usando o valor M foi menor e aparentam estar contidas nas DMPs detectadas usando o valor β, Tabela 4.3. O número de DMPs detectadas pelo teste QUOR para ambos os conjuntos de dados observados, se mantiveram aproximadas, demonstrando menor sensibilidade do QUOR em relação do uso do valor β e do valor M.

(a) Quantidade de DMPs detectadas entre os testes **"Paramétrico"**, **"Não paramétrico"** e **"Bayesiano"** sem a correção por múltipla testagem

	COCAÍNA/CRACK (p valor = 0.01)		
	sem correção para múltipla testagem		
	M	β	Intersecção
paramétrico	26261	26532	23106
não paramétrico	26524	26367	23257
Bayesiano	25830	26426	22424

(b) Quantidade de DMPs detectadas entre os testes **"Paramétrico"**, **"Não paramétrico"** e **"Bayesiano"** com a correção por múltipla testagem

	COCAÍNA/CRACK (p valor ajustado = 0.01)		
	com correção para múltipla testagem		
	M	β	Intersecção
paramétrico	3014	3118	2709
não paramétrico	3109	3229	2843
Bayesiano	2669	3218	2434

(c) Quantidade de DMPs detectadas com QUOR

	COCAÍNA/CRACK		
	M	β	Intersecção
QUOR (conf = 90)	26073	24313	19892
QUOR (conf = 95)	13373	12449	10229
QUOR (conf = 99)	2182	2041	1759
QUOR (conf = 99.99)	160	159	143

Tabela 4.2: *Relação das DMPs detectadas com o conjunto de dados COCAÍNA/CRACK*

4.3 Intersecção dos resultados entre o QUOR e os testes

Comparamos então a sobreposição do resultado obtido pelo QUOR com os testes **"Paramétrico"**, **"Não paramétrico"** e **"Bayesiano"**, sem e com múltipla testagem. Conforme aumentamos o valor de corte para as DMPs detectadas pelo QUOR para o conjunto de dados COCAÍNA/CRACK, Tabela 4.4, Tabela 4.5 e para o conjunto de dados GRAVIDAS, Tabela 4.6 e Tabela 4.7, verificamos que o QUOR possui uma alta intersecção ao que foi detectado pelos três testes.

4.4 Intersecção dos resultados entre todos os testes

Realizamos a intersecção entre todos os testes e selecionamos as DMPs detectadas por todos os 4 testes, as DMPs únicas do método QUOR e as DMPs únicas detectadas pelos testes

(a) Quantidade de DMPs detectadas entre os testes **"Paramétrico"**, **"Não paramétrico"** e **"Bayesiano"** sem a correção por múltipla testagem

	GRAVIDAS (p valor = 0.01)		
	sem correção para múltipla testagem		
	M	β	Intersecção
paramétrico	5890	14242	4810
não paramétrico	5108	12503	4289
Bayesiano	6175	13677	4939

(b) Quantidade de DMPs detectadas entre os testes **"Paramétrico"**, **"Não paramétrico"** e **"Bayesiano"** com a correção por múltipla testagem

	GRAVIDAS (p valor ajustado = 0.01)		
	com correção para múltipla testagem		
	M	β	Intersecção
paramétrico	0	0	0
não paramétrico	0	2	0
Bayesiano	0	0	0

(c) Quantidade de DMPs detectadas com QUOR

	GRAVIDAS		
	M	β	Intersecção
QUOR (conf = 90)	8590	8024	5030
QUOR (conf = 95)	2134	1832	1104
QUOR (conf = 99)	100	63	40
QUOR (conf = 99.99)	0	0	0

Tabela 4.3: *Relação das DMPs detectadas com o conjunto de dados GRAVIDAS*

(a) Utilizando o valor M sem correção

		QUOR 90	QUOR 95	QUOR 99	QUOR 99.99
Paramétrico	em comum	16160	10465	2157	160
	quor	9913	2908	25	0
	paramétrico	10101	15796	24104	26101
Não paramétrico	em comum	17771	11473	2178	160
	quor	8302	1900	4	0
	não paramétrico	8753	15051	24346	26364
Bayesiano	em comum	15958	10395	2150	160
	quor	10115	2978	32	0
	Bayesiano	9872	15435	23680	25670

(b) Utilizando o valor β sem correção

		QUOR 90	QUOR 95	QUOR 99	QUOR 99.99
Paramétrico	em comum	15037	9682	2009	159
	quor	9276	2767	32	0
	paramétrico	11495	16850	24523	26373
Não paramétrico	em comum	16478	10543	2035	159
	quor	7835	1906	6	0
	não paramétrico	9889	15824	24332	26208
Bayesiano	em comum	14985	9652	2007	159
	quor	9328	2797	34	0
	Bayesiano	11441	16774	24419	26267

Tabela 4.4: *Intersecção das DMPs detectadas entre cada teste com as DMPs detectadas com QUOR do conjunto de dados COCAÍNA/CRACK sem correção*

“Paramétrico”, **“Não paramétrico”** e **“Bayesiano”**, Tabela 4.8 e Tabela 4.9. Como visto anteriormente, com o aumento do corte de confiança, as DMPs detectadas pelo QUOR estão contidas nas DMPS detectadas pelos 3 testes, aparentemente demonstrando que o p valor mesmo ajustado por múltipla testagem ainda possui falsos positivos.

4.5 Comparação entre os cortes de Confiança

Os valores de corte de confiança 0.90, 0.95 e 0.99 apesar de serem altos valores de confiança, mostraram-se com baixa eficácia como critério de corte para o que estava realmente diferencialmente metilado, a grande variabilidade do dado de metilação possue influência no valor de corte de confiança. As DMPs detectadas pelos testes **“Paramétrico”**, **“Não paramétrico”** e **“Bayesiano”** e que não foram detectadas pelo QUOR, possuem o valor de confiança dentro o intervalor 0.95 a 0.99, tanto com uso do valor β como com valor M com o conjunto de dados COCAÍNA/CRACK, Figura 4.4 e Figura 4.5.

No conjunto de dados GRAVIDAS, as DMPs detectadas pelos testes **“Paramétrico”**, **“Não paramétrico”** e **“Bayesiano”** e que não foram detectadas pelo QUOR, apresenta uma quantidade maior de DMPs com valores de confiança abaixo de 0.90, principalmente com o uso do valor β, Figura 4.6 e Figura 4.7.

O p valor corrigido por múltipla testagem e abaixo do corte 0.01 e com valor de confiança alto como 99, demonstra detectar também sondas que não são diferencialmente metiladas, ou seja, falsos positivos, podemos observar a sobreposição das distribuições, como visto na Figura 4.8 e Figura 4.9.

Ao aumentarmos o valor de corte de confiança para um valor mais restritivo, como 0.9999 para o conjunto de dados COCAINA/CRACK conseguimos então detectar as DMPs reais, que são

(a) Utilizando o valor M com correção

		M com correção			
		QUOR 90	QUOR 95	QUOR 99	QUOR 99.99
Paramétrico	em comum	2994	2862	1578	160
	quor	23079	10511	604	0
	paramétrico	20	152	1436	2854
Não paramétrico	em comum	3090	3006	1696	160
	quor	22983	10367	486	0
	não paramétrico	19	103	1413	2949
Bayesiano	em comum	2658	2557	1489	160
	quor	23415	10816	693	0
	Bayesiano	11	112	1180	2509

(b) Utilizando o valor β com correção

		β com correção			
		QUOR 90	QUOR 95	QUOR 99	QUOR 99.99
Paramétrico	em comum	3052	2863	1520	159
	quor	21261	9586	521	0
	paramétrico	66	255	1598	2959
Não paramétrico	em comum	3165	3026	1648	159
	quor	21148	9423	393	0
	não paramétrico	64	203	1581	3070
Bayesiano	em comum	3144	2923	1535	159
	quor	21169	9526	506	0
	Bayesiano	74	295	1683	3059

Tabela 4.5: *Intersecção das DMPs detectadas entre cada teste com as DMPs detectadas com QUOR do conjunto de dados COCAÍNA/CRACK com correção*

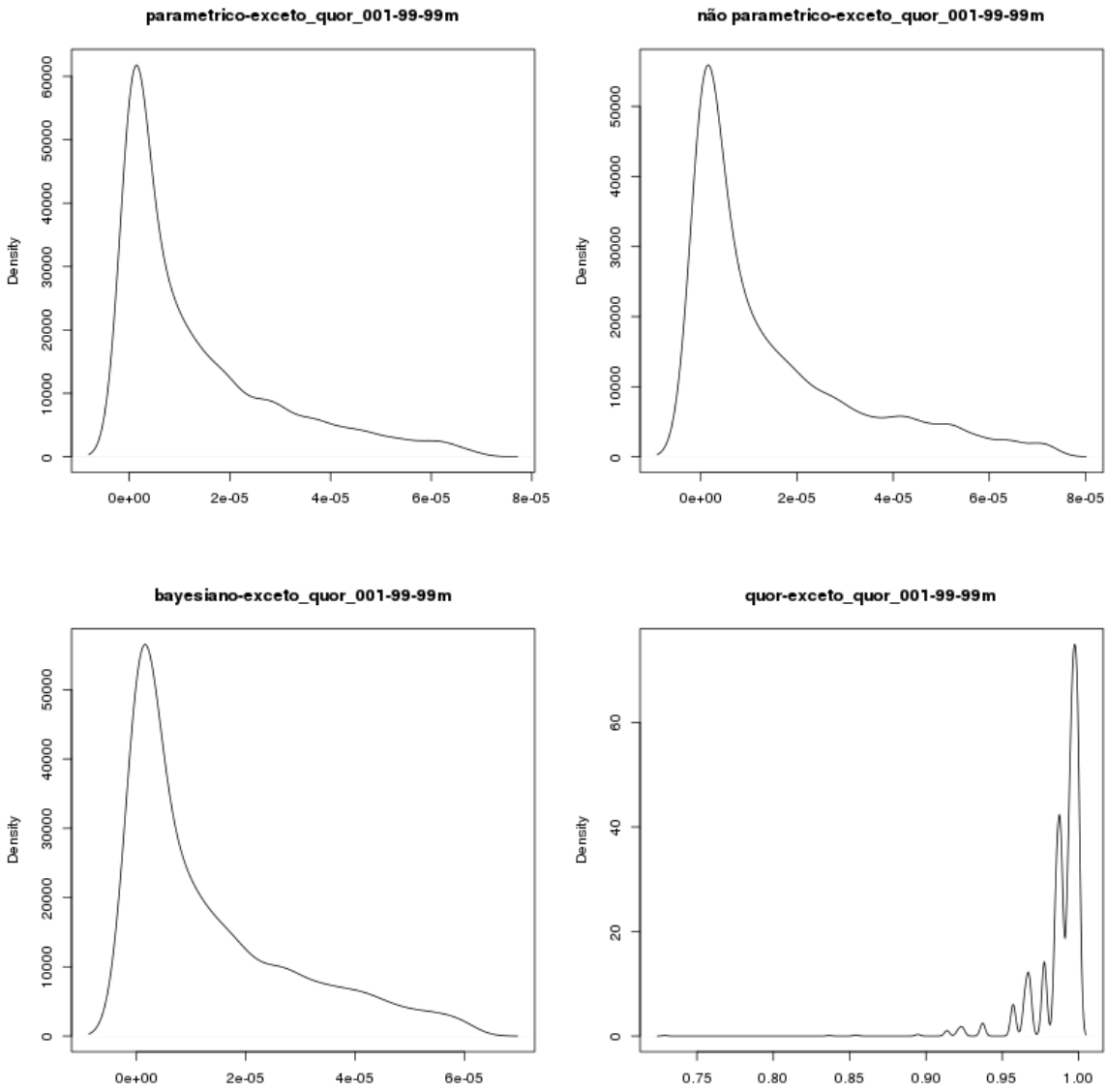


Figura 4.4: *Intersecção das DMPs detectadas com o uso dos três testes utilizando o valor M com correção para múltipla testagem com conjunto de dados COCAÍNA/CRACK.*

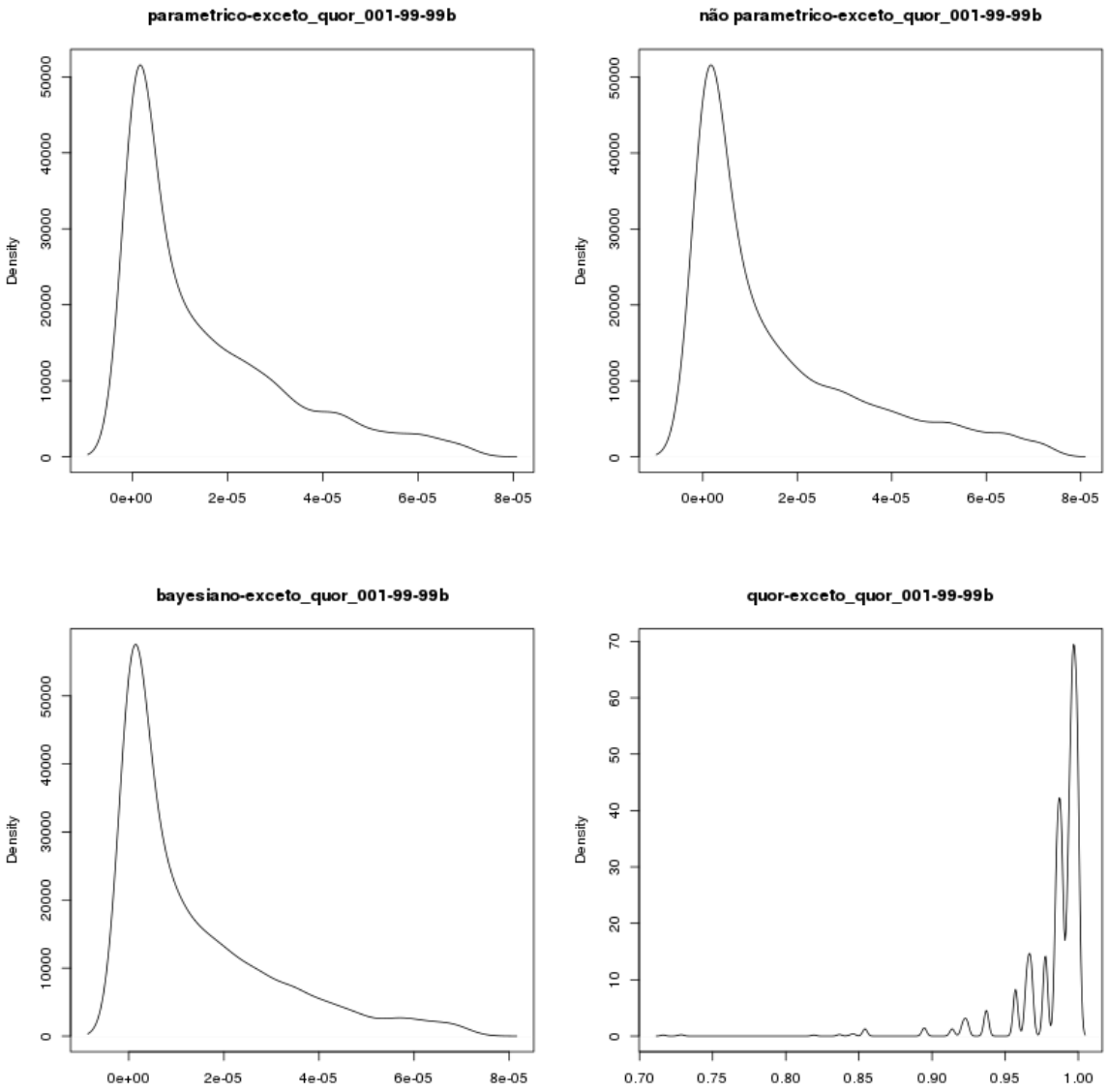


Figura 4.5: *Intersecção das DMPs detectadas com o uso dos três testes utilizando o valor β com correção para múltipla testagem com conjunto de dados COCAÍNA/CRACK.*

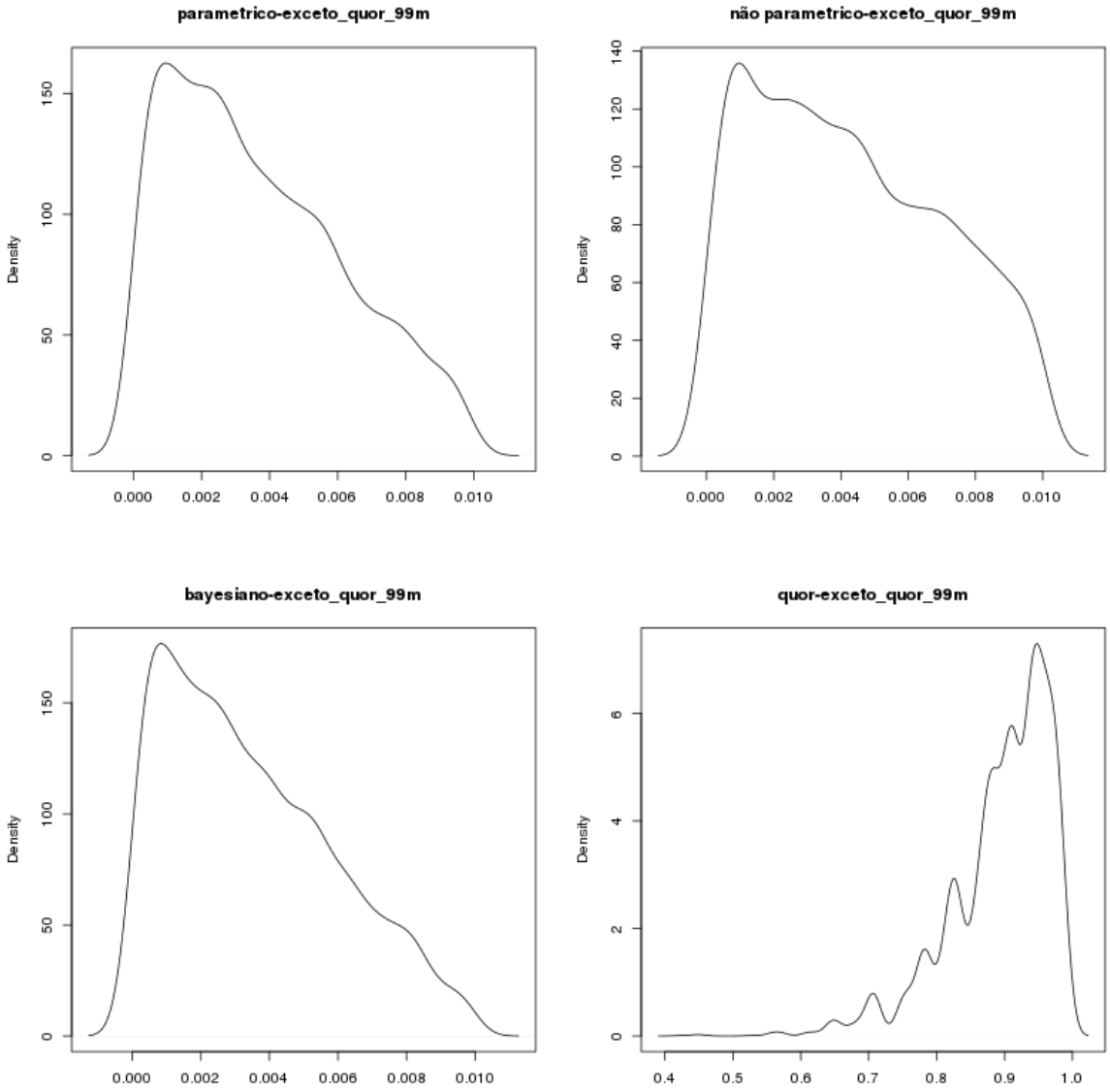


Figura 4.6: *Intersecção das DMPs detectadas com o uso dos três testes utilizando o valor M com correção para múltipla testagem com conjunto de dados GRAVIDAS.*

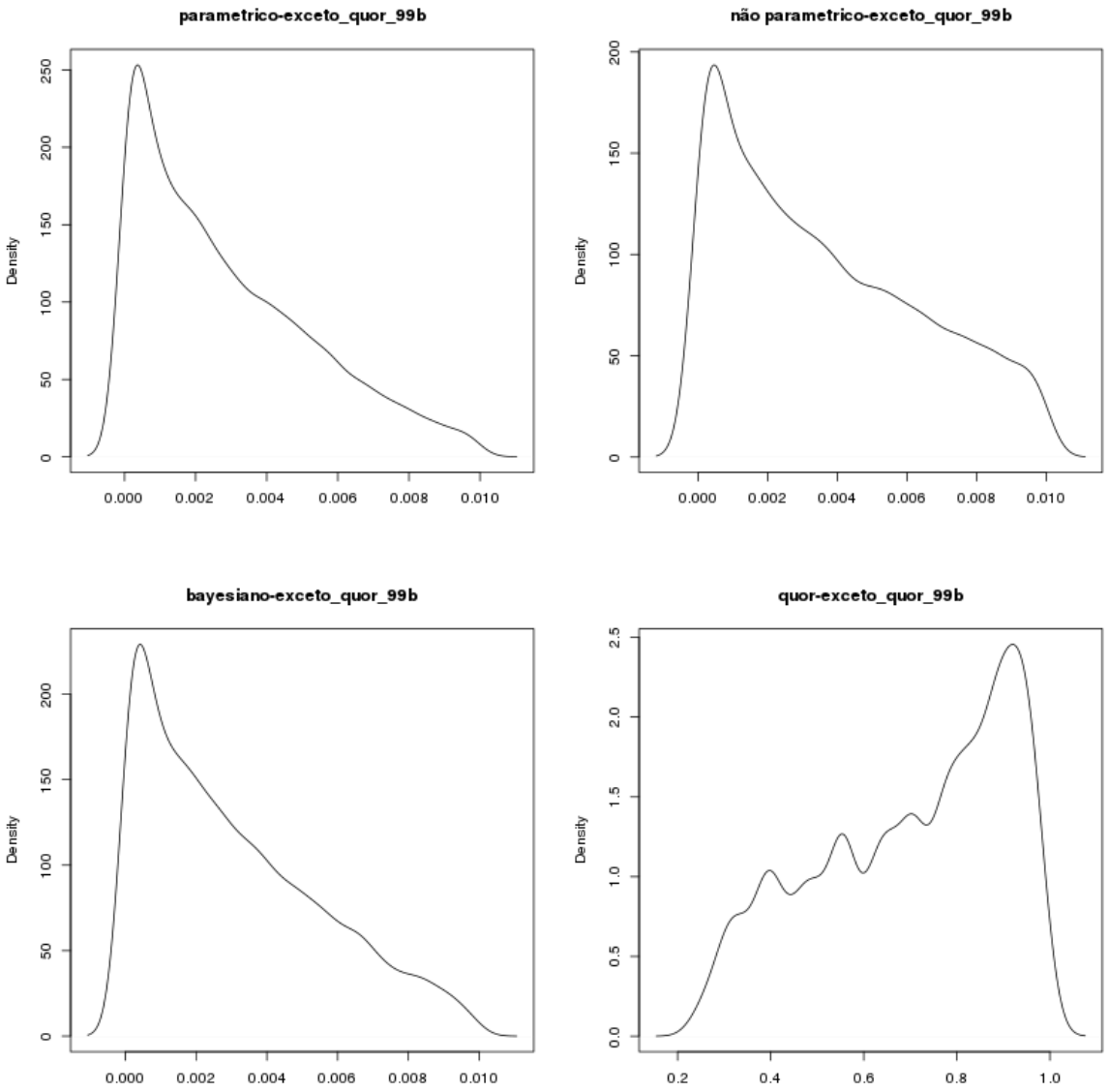


Figura 4.7: *Intersecção das DMPs detectadas com o uso dos três testes utilizando o valor β com correção para múltipla testagem com conjunto de dados GRAVIDAS.*

(a) Utilizando o valor M sem correção

		QUOR 90	QUOR 95	QUOR 99
Paramétrico	em comum	2731	1197	88
	quor	5859	937	12
	paramétrico	3159	4693	5802
Não paramétrico	em comum	2962	1349	97
	quor	5628	785	3
	não paramétrico	2146	3759	5011
Bayesiano	em comum	2801	1205	88
	quor	5789	929	12
	Bayesiano	3374	4970	6087

(b) Utilizando o valor β sem correção

		QUOR 90	QUOR 95	QUOR 99
Paramétrico	em comum	2605	939	55
	quor	5419	893	8
	paramétrico	11637	13303	14187
Não paramétrico	em comum	2849	1077	59
	quor	5175	755	4
	não paramétrico	9654	11426	12444
Bayesiano	em comum	2585	930	54
	quor	5439	902	9
	Bayesiano	11092	12747	13623

Tabela 4.6: *Intersecção das DMPs detectadas entre cada teste com as DMPs detectadas com QUOR do conjunto de dados GRAVIDAS sem correção*

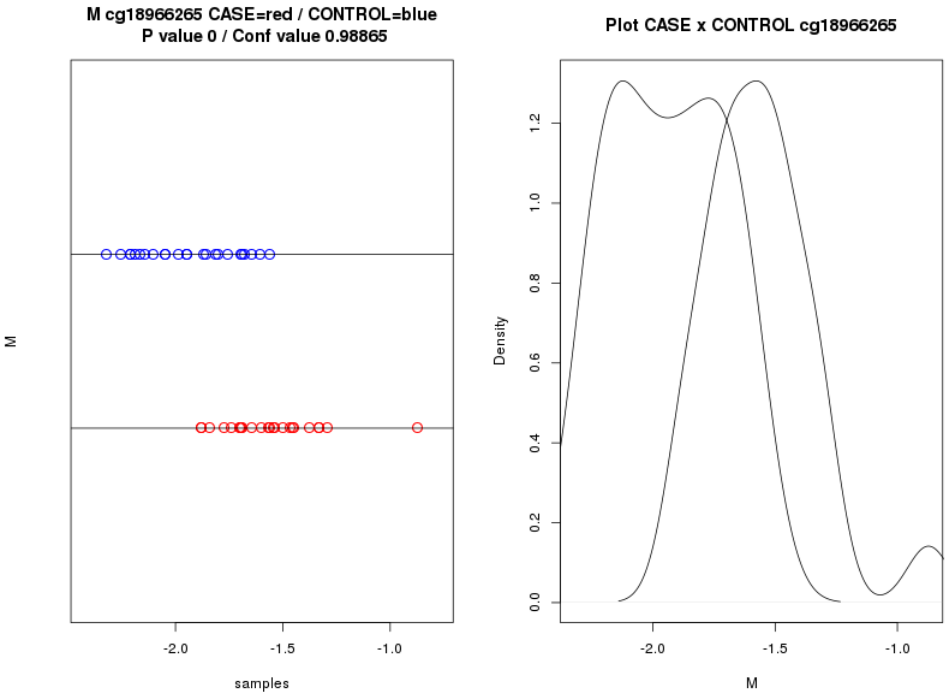


Figura 4.8: *Apesar do valor de confiança muito alto utilizando o valor Beta e M demostrados nos graficos A e B, este valor de confiança não representa uma diferença de metilação entre os grupos, como observados pelos graficos de distribuição sobrepostos.*

(a) Utilizando o valor M com correção

		QUOR 90	QUOR 95	QUOR 99
Paramétrico	em comum	0	0	0
	quor	8590	2134	100
	paramétrico	0	0	0
Não paramétrico	em comum	0	0	0
	quor	8590	2134	100
	não paramétrico	0	0	0
Bayesiano	em comum	0	0	0
	quor	8590	2134	100
	Bayesiano	0	0	0

(b) Utilizando o valor β com correção

		QUOR 90	QUOR 95	QUOR 99
Paramétrico	em comum	1	0	0
	quor	8023	1832	63
	paramétrico	1	2	2
Não paramétrico	em comum	1	0	0
	quor	8023	1832	63
	não paramétrico	1	2	2
Bayesiano	em comum	0	0	0
	quor	8024	1832	63
	Bayesiano	0	0	0

Tabela 4.7: *Intersecção das DMPs detectadas entre cada teste com as DMPs detectadas com QUOR do conjunto de dados GRAVIDAS com correção*

valor de corte de confiança	Teste	Sem correção		Com correção	
		M	Beta	M	Beta
90	Todos	15144	14294	2394	2662
	Somente QUOR	7605	7253	22634	20697
	Exceto QUOR	6667	8043	7	43
95	Todos	10143	9453	2338	2557
	Somente QUOR	1750	1778	10091	9091
	Exceto QUOR	11668	12884	63	148
99	Todos	2150	2006	1436	1478
	Somente QUOR	4	6	451	358
	Exceto QUOR	19661	20331	965	1227
99.99	Todos	160	159	160	159
	Somente QUOR	0	0	0	0
	Exceto QUOR	21651	22178	2241	2546

Tabela 4.8: *Intersecção entre todas as DMPs detectadas entre os 4 testes, DMPs detectadas somente pelo QUOR e DMPS detectadas pelos testes* ***"Paramétrico"****,* ***"Não paramétrico"*** *e* ***"Bayesiano"*** *no conjunto de dados COCAÍNA/CRACK*

valor de corte de confiança	Teste	Sem correção		Com correção	
		M	Beta	M	Beta
90	Todos	2377	2314	0	0
	Somente QUOR	5241	4910	8590	8023
	Exceto QUOR	1753	8325	0	0
95	Todos	1114	889	0	0
	Somente QUOR	721	721	2134	1832
	Exceto QUOR	3016	9750	0	0
99	Todos	86	54	0	0
	Somente QUOR	3	4	100	63
	Exceto QUOR	4044	10585	0	0

Tabela 4.9: *Intersecção entre todas as DMPs detectadas entre os 4 testes, DMPs detectadas somente pelo QUOR e DMPS detectadas pelos testes* ***"Paramétrico"****,* ***"Não paramétrico"*** *e* ***"Bayesiano"*** *no conjunto de dados GRAVIDAS*

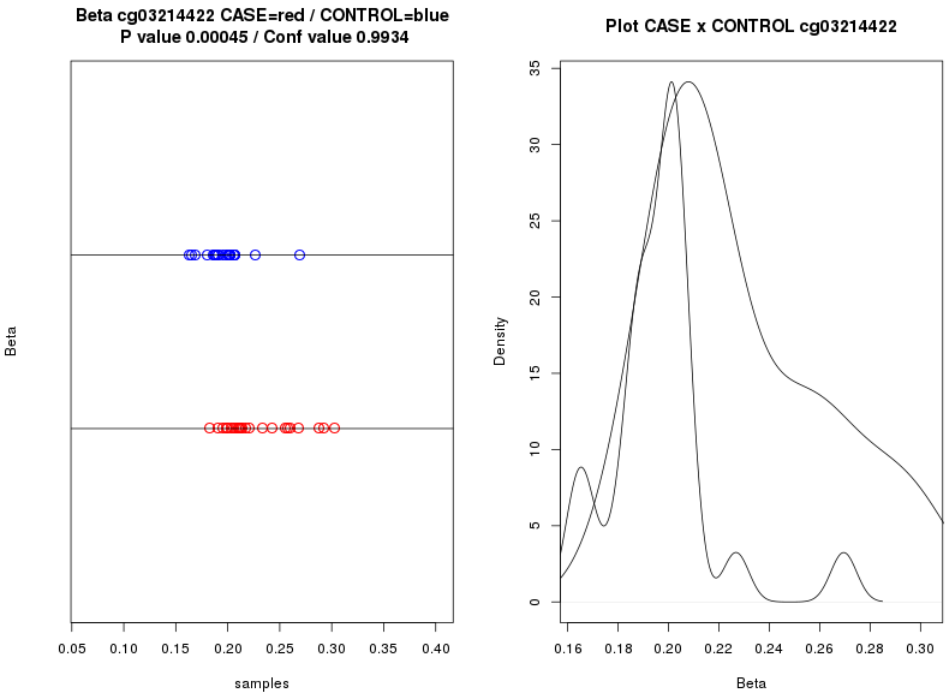


Figura 4.9: *Apesar do valor de confiança muito alto utilizando o valor Beta e M demostrados nos graficos A e B, este valor de confiança não representa uma diferença de metilação entre os grupos, como observados pelos graficos de distribuição sobrepostos.*

realmente diferencialmente metiladas entre o grupo caso e controle, Figura 4.10 e Figura 4.11.

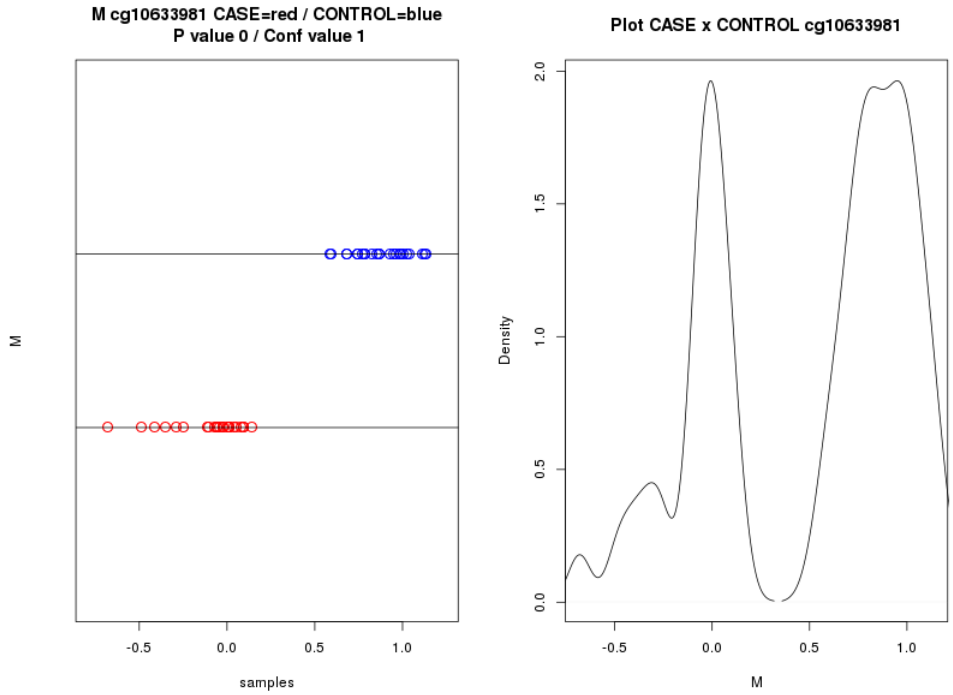


Figura 4.10: *Distribuição dos valores M de caso e controle da DMP cg10633981 com valor de confiança maior que 0.9999*

Para o conjunto de dados GRAVIDAS, o uso de um corte mais restritivo como 0.9999 ou menor, obteve zero DMPs detectadas. Admitimos o corte de 0.99 de confiança, mas não detectou nenhuma DMP com real diferença entre caso e controle, utilizando valor M, Figura 4.12 e Figura 4.13.

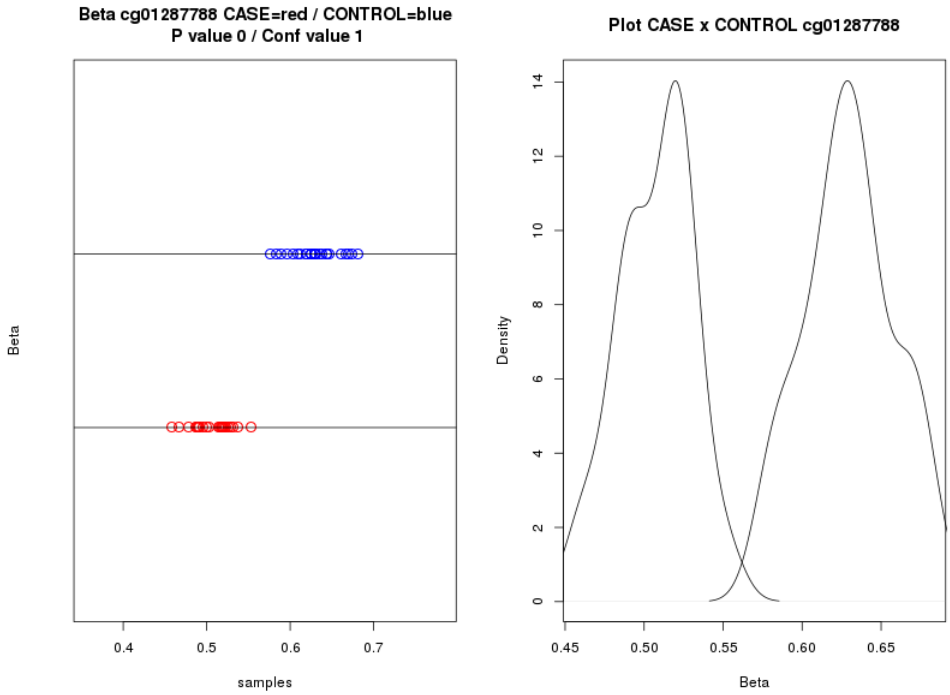


Figura 4.11: *Distribuição dos valores β de caso e controle da DMP cg01287788 com valor de confiança maior que 0.9999*

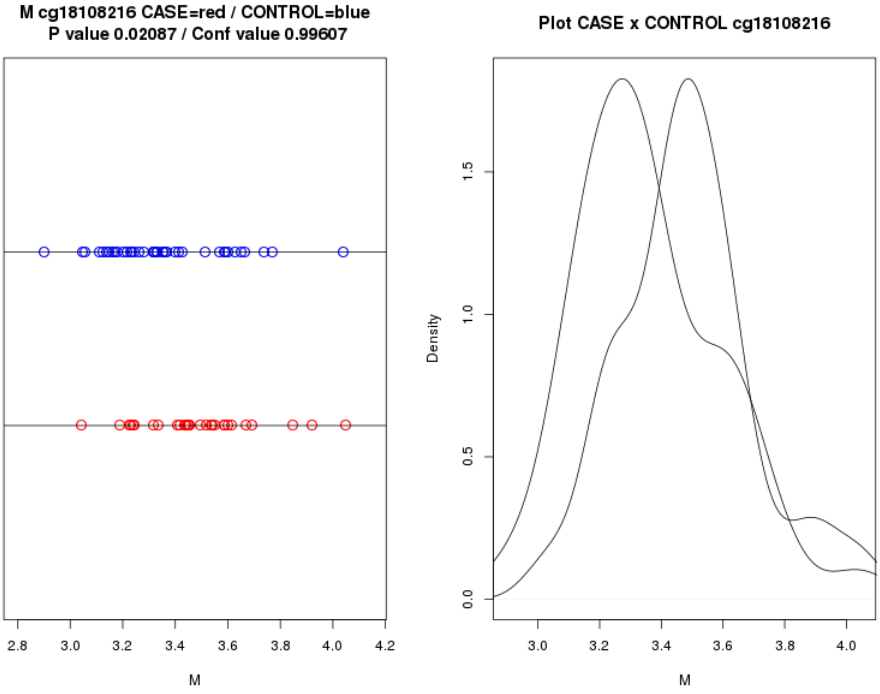


Figura 4.12: *Gráfico da distribuição das DMPs reais detectadas.*

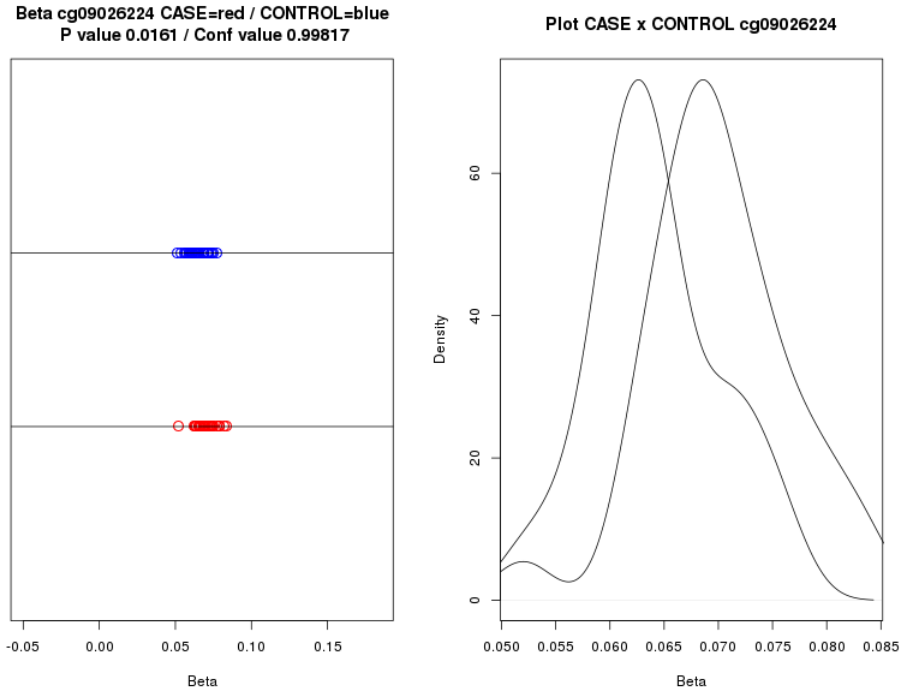


Figura 4.13: *Gráfico da distribuição das DMPs reais detectadas.*

Capítulo 5

Discussão e conclusão

5.1 Discussão

Ao longo do nosso estudo comparamos o uso do valor β e o valor M na analise de metilação de DNA com uso da lâmina Illumina HumanMethylation450. No conjunto de dados COCAÍNA/CRACK o número de DMPs detectadas com os quatro testes utilizando o valor M e o valor β foram aproximados entre os dois valores. Para o caso do conjunto de dados GRAVIDAS, podemos observar que o numero de DMPs detectadas sem a correção do p valor com os testes **"Paramétrico"**, **"Não paramétrico"** e **"Bayesiano"** utilizando o valor M foi muito menor do que o número de DMPs detectadas sem a correção do p valor utilizando o valor β, além de demonstrar que as DMPs detectadas com valor M estão em intersecção com as DMPs detectadas com valor β. O número de DMPs detectadas com QUOR utilizando o valor M e o valor β, mantiveram-se muito aproximados, indicando que o QUOR pode não sofrer influência da heteroscedasticidade do valor β. Isto aponta que o valor β pode levar aos testes **"Paramétrico"**, **"Não paramétrico"** e **"Bayesiano"** a detectarem um número maior de falsos positivos quando não realizada a correção por múltipla testagem, e o valor M, com seu perfil menos heteroscedástico, permite uma melhor detecção de verdadeiros positivos Zhuang *et al.* [2012].

O valor de confiança usado como corte para determinar quais sondas sejam DMPs, deve ser muito restritivo, pois mesmo um alto valor de confiança como 0.99 ainda não é suficiente para determinar uma DMP em conjuntos de dados que possuem pequenos alterações no padrão de metilação.

O uso de diferentes abordagens estatísticas detectaram quantidades de DMPs muito próximos apesar da diferença das premissas de cada teste, podendo então apontar que, o que diferencia os resultados obtidos por cada pipeline (Morris e Beck [2014]; Li *et al.* [2015]; Bock [2012]), são os métodos empregados para filtragem, normalização e correção do dado e não sobre o método estatístico empregado para detecção de diferenças entre grupos. Apesar do conjunto de dados GRAVIDAS ter passado pelo mesmo fluxo aplicado ao conjunto de dados COCAÍNA/CRACK, não obtivemos nenhuma DMP com os testes **"Paramétrico"** e **"Bayesiano"**, o teste **"Não paramétrico"** detectou duas DMPs com valor β e o QUOR detectou 100 DMPs com valor M e 63 DMPs com valor β. Estas DMPs detectadas apesar de serem significativas, ao observar a distribuição destas DMPs entre caso e controle, percebemos que não há uma diferença real entre elas.

Cada dado necessita de filtragens, normalizações, correções e métodos para detecção das DMPs diferentes, não havendo uma pipeline que funcionará em todas as vezes. É necessário, estudar o dado e testar quais opções deverá aplicar sobre o conjunto de dados para detectar as DMPs com qualidade, quantidade e significância biológica (Zhuang *et al.* [2012]; Johnstone *et al.* [2013]). Em casos onde as hipóteses não estão bem definidas, o QUOR pode ser aplicado como primeira análise, utilizando o valor M e um valor alto de confiança como corte.

5.2 Conclusão

Atualmente, o microarranjo para analise de metilação de DNA Illumina HumanMethylation450 ainda continua sendo, desde seu lançamento uma plataforma de baixo custo e de grande abrangência comparativamente com outras técnicas no mercado. O desafio que ainda permanece está em realizar a analise desta grande quantidade de dados gerados e obter resultados com grande significância biológica. Sem dúvida mais processos de controle de qualidade e correção do dado terão de serem desenvolvidos para remoção de eventuais viéses que podem impactar no resultado final e na sua interpretação, algo que poderá melhorar com o uso mais intensivo de delineamento de experimento. Mas como podemos observar, cada conjunto de dados possui suas particularidades, sendo necessário investigar e testar as diferentes opções que as pipelines desenvolvidas podem oferecer.

Demonstramos ao realizar um unico fluxo para os passos de preprocessamento, normalização e correção, comparando os métodos utilizados para detecção de diferença entre grupos que as pipelines mais conhecidas utilizam, como ChAMP, RNBeads e IMA, resultam na identificação de um número muito próximo de CpGs diferencialmente metilados. Estes métodos tem como premissa a interdependência entre os sítios CpGs e assume-se ou não uma distribuição, como a distribuição normal, detectando muitas DMPs significativas que deverão ter seus p valores corrigidos por algum método de múltipla testagem. Dentre essas DMPs detectadas ainda haverá falsos positivos.

Ao comparar com um método que determine a diferença de metilação entre populações, sem a necessidade de assumir a normalidade dos dados, sem a necessidade de correção para múltiplas testagens e que pode ser aplicado sobre um número de amostras relativamente pequeno, dada a alta variabilidade do valor de metilação entre indivíduos, observamos que o nosso fluxo de analise pode ser utilizado como uma forma primária para a detecção de DMPs, antes da aplicação de outros métodos, com o uso de um valor restritivo de corte e com o uso do valor M. Há a necessidade de verificar a significância biológica destes resultados obtidos com o QUOR.

Referências Bibliográficas

Aryee *et al.***(2014)** Martin J Aryee, Andrew E Jaffe, Hector Corrada-Bravo, Christine Ladd-Acosta, Andrew P Feinberg, Kasper D Hansen e Rafael A Irizarry. Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. *Bioinformatics (Oxford, England)*, 30(10):1363–9. ISSN 1367-4811. doi: 10.1093/bioinformatics/btu049. URL http://bioinformatics.oxfordjournals.org/content/30/10/1363. Citado na pág. 9, 10, 17

Assenov *et al.***(2009)** Yassen Assenov, Fabian Müller, Pavlo Lutsik, Jörn Walter, Thomas Lengauer e Christoph Bock. Comprehensive Analysis of DNA Methylation Data with RnBeads. 40 (2):189–197. doi: 10.1038/ng.75.Six. Citado na pág. 10, 18

Bibikova *et al.***(2006)** Marina Bibikova, Zhenwu Lin, Lixin Zhou, Eugene Chudin, Eliza Wickham Garcia, Bonnie Wu, Dennis Doucet, Neal J Thomas, Yunhua Wang, Ekkehard Vollmer, Torsten Goldmann, Carola Seifart, Wei Jiang, David L Barker, Mark S Chee, Joanna Floros e Jian-Bing Fan. High-throughput DNA methylation profiling using universal bead arrays. *Genome research*, 16(3):383–93. ISSN 1088-9051. doi: 10.1101/gr.4410706. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1415217{&}tool=pmcentrez{&}rendertype=abstract. Citado na pág. 4, 8

Bibikova *et al.***(2009)** Marina Bibikova, Jennie Le, Bret Barnes, Shadi Saedinia-Melnyk, Lixin Zhou, Richard Shen e Kevin L. *Genome-wide DNA methylation profiling using Infinium assay*. Epigenetics. Citado na pág. 4

Bibikova *et al.***(2011)** Marina Bibikova, Bret Barnes, Chan Tsan, Vincent Ho, Brandy Klotzle, Jennie M Le, David Delano, Lu Zhang, Gary P Schroth, Kevin L Gunderson, Jian-Bing Fan e Richard Shen. High density DNA methylation array with single CpG site resolution. *Genomics*, 98(4):288–95. ISSN 1089-8646. doi: 10.1016/j.ygeno.2011.07.007. URL http://www.ncbi.nlm.nih.gov/pubmed/21839163. Citado na pág. 2, 4, 5, 6, 10, 11

Bird(2002) Adrian Bird. DNA methylation patterns and epigenetic memory. *Genes & development*, 16(1):6–21. ISSN 0890-9369. doi: 10.1101/gad.947102. URL http://www.ncbi.nlm.nih.gov/pubmed/11782440. Citado na pág. 1, 12

Bock(2012) Christoph Bock. Analysing and interpreting DNA methylation data. *Nature Reviews Genetics*, 13(10):705–719. ISSN 1471-0056. doi: 10.1038/nrg3273. URL http://www.ncbi.nlm.nih.gov/pubmed/22986265. Citado na pág. 2, 6, 8, 9, 10, 11, 12, 13, 15, 37, 47

Bock *et al.***(2010)** Christoph Bock, Eleni M Tomazou, Arie B Brinkman, Fabian Müller, Femke Simmer, Hongcang Gu, Natalie Jäger, Andreas Gnirke, Hendrik G Stunnenberg e Alexander Meissner. Quantitative comparison of genome-wide DNA methylation mapping technologies. *Nature biotechnology*, 28(10):1106–14. ISSN 1546-1696. doi: 10.1038/nbt.1681. URL http://www.scopus.com/inward/record.url?eid=2-s2.0-77957937011{&}partnerID=tZOtx3y1. Citado na pág. 12

Butcher e Beck(2014a) Lee M. Butcher e Stephan Beck. Probe Lasso: a novel method to rope in differentially methylated regions with 450k DNA methylation data. *Methods*, 72:21–28. ISSN

10462023. doi: 10.1016/j.ymeth.2014.10.036. URL http://dx.doi.org/10.1016/j.ymeth.2014.10.036. Citado na pág. 12

Butcher e Beck(2014b) Lee M. Butcher e Stephan Beck. Probe Lasso: a novel method to rope in differentially methylated regions with 450k DNA methylation data. *Methods*, 72:21–28. ISSN 10462023. doi: 10.1016/j.ymeth.2014.10.036. URL http://dx.doi.org/10.1016/j.ymeth.2014.10.036. Citado na pág. 2

Chen *et al.***(2011)** Chao Chen, Kay Grennan, Judith Badner, Dandan Zhang, Elliot Gershon, Li Jin e Chunyu Liu. Removing batch effects in analysis of expression microarray data: An evaluation of six batch adjustment methods. *PLoS ONE*, 6(2). ISSN 19326203. doi: 10.1371/journal.pone.0017238. Citado na pág. 10

Chen *et al.***(2012)** Zhongxue Chen, Qingzhong Liu e Saralees Nadarajah. A new statistical approach to detecting differentially methylated loci for case control Illumina array methylation data. páginas 1–5. Citado na pág. 10, 12

Davis e Bilke(2012) Sean Davis e Sven Bilke. An Introduction to the methylumi package. páginas 1–10. Citado na pág. 8, 10

Deaton e Bird(2011) Am Deaton e Adrian Bird. CpG islands and the regulation of transcription. *Genes & development*, 25(10):1010–1022. doi: 10.1101/gad.2037511.1010. URL http://genesdev.cshlp.org/content/25/10/1010.short. Citado na pág. 12

Dedeurwaerder *et al.***(2011)** Sarah Dedeurwaerder, Matthieu Defrance, Emilie Calonne, Hélène Denis, Christos Sotiriou e François Fuks. Evaluation of the Infinium Methylation 450K technology. *Epigenomics*, 3(6):771–784. ISSN 1750-1911. doi: 10.2217/epi.11.105. URL http://www.ncbi.nlm.nih.gov/pubmed/22126295. Citado na pág. 5, 7, 11, 19

Dedeurwaerder *et al.***(2014)** Sarah Dedeurwaerder, Matthieu Defrance, Martin Bizet, Emilie Calonne, Gianluca Bontempi e Fran??ois Fuks. A comprehensive overview of Infinium HumanMethylation450 data processing. *Briefings in bioinformatics*, 15(6):929–941. ISSN 14774054. doi: 10.1093/bib/bbt054. URL http://bib.oxfordjournals.org/content/early/2013/08/27/bib.bbt054.abstract. Citado na pág. 8, 9, 10

Dempster *et al.***(2011)** Emma L. Dempster, Ruth Pidsley, Leonard C. Schalkwyk, Sheena Owens, Anna Georgiades, Fergus Kane, Sridevi Kalidindi, Marco Picchioni, Eugenia Kravariti, Timothea Toulopoulou, Robin M. Murray e Jonathan Mill. Disease-associated epigenetic changes in monozygotic twins discordant for schizophrenia and bipolar disorder. *Human Molecular Genetics*, 20(24):4786–4796. ISSN 09646906. doi: 10.1093/hmg/ddr416. Citado na pág. 11

Doi *et al.***(2009)** Akiko Doi, In-Hyun Park, Bo Wen, Peter Murakami, Martin J Aryee, Rafael Irizarry, Brian Herb, Christine Ladd-Acosta, Junsung Rho, Sabine Loewer, Justine Miller, Thorsten Schlaeger, George Q Daley e Andrew P Feinberg. Differential methylation of tissue- and cancer-specific CpG island shores distinguishes human induced pluripotent stem cells, embryonic stem cells and fibroblasts. *Nature genetics*, 41(12):1350–3. ISSN 1546-1718. doi: 10.1038/ng.471. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2958040{&}tool=pmcentrez{&}rendertype=abstract. Citado na pág. 2

Du *et al.***(2010)** Pan Du, Xiao Zhang, Chiang-Ching Huang, Nadereh Jafari, Warren A Kibbe, Lifang Hou e Simon M Lin. Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis. *BMC bioinformatics*, 11(1):587. ISSN 1471-2105. doi: 10.1186/1471-2105-11-587. URL http://www.biomedcentral.com/1471-2105/11/587. Citado na pág. 8, 9, 19

Du *et al.***(2011)** Pan Du, Gang Feng, Spencer Huang, Warren A Kibbe e Simon Lin. Analyze Illumina Infinium methylation microarray data. 2. Citado na pág. 12

Dyson *et al.***(2014)** Matthew T. Dyson, Damian Roqueiro, Diana Monsivais, C. Mutlu Ercan, Mary Ellen Pavone, David C. Brooks, Toshiyuki Kakinuma, Masanori Ono, Nadereh Jafari, Yang Dai e Serdar E. Bulun. Genome-Wide DNA Methylation Analysis Predicts an Epigenetic Switch for GATA Factor Expression in Endometriosis. *PLoS Genetics*, 10(3):e1004158. ISSN 15537404. doi: 10.1371/journal.pgen.1004158. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3945170{&}tool=pmcentrez{&}rendertype=abstract. Citado na pág. 2, 5, 6

Esteller(2002) Manel Esteller. CpG island hypermethylation and tumor suppressor genes: a booming present, a brighter future. *Oncogene*, 21:5427–5440. ISSN 0950-9232. doi: 10.1038/sj.onc.1205600. Citado na pág. 1

Feber *et al.***(2014)** Andrew Feber, Paul Guilhamon, Matthias Lechner, Tim Fenton, Gareth a Wilson, Christina Thirlwell, Tiffany J Morris, Adrienne M Flanagan, Andrew E Teschendorff, John D Kelly e Stephan Beck. Using high-density DNA methylation arrays to profile copy number alterations. *Genome biology*, 15(2):R30. ISSN 1465-6914. doi: 10.1186/gb-2014-15-2-r30. URL http://www.ncbi.nlm.nih.gov/pubmed/24490765. Citado na pág. 11

Gardiner-Garden e Frommer(1987) M Gardiner-Garden e M Frommer. CpG islands in vertebrate genomes. *Journal of molecular biology*, 196:261–282. ISSN 00222836. doi: 10.1016/0022-2836(87)90689-9. Citado na pág. 1

Gentleman(2004) Robert C Gentleman. Bioconductor: open software development for computational biology and bioinformatics. (10). Citado na pág. 10

Hansen *et al.***(2011)** Kasper Daniel Hansen, Winston Timp, Héctor Corrada Bravo, Sarven Sabunciyan, Benjamin Langmead, Oliver G McDonald, Bo Wen, Hao Wu, Yun Liu, Dinh Diep *et al.* Increased methylation variation in epigenetic domains across cancer types. *Nature genetics*, 43(8):768–775. Citado na pág. 12

Houseman *et al.***(2009)** E Andrés Houseman, Brock C Christensen, Margaret R Karagas, Margaret R Wrensch, Heather H Nelson, Joseph L Wiemels, Shichun Zheng, John K Wiencke, Karl T Kelsey e Carmen J Marsit. Copy number variation has little impact on bead-array-based measures of DNA methylation. *Bioinformatics (Oxford, England)*, 25(16):1999–2005. ISSN 1367-4811. doi: 10.1093/bioinformatics/btp364. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2723008{&}tool=pmcentrez{&}rendertype=abstract. Citado na pág. 11

Houseman *et al.***(2012)** E Andres Houseman, William P Accomando, Devin C Koestler, Brock C Christensen, Carmen J Marsit, Heather H Nelson, John K Wiencke e Karl T Kelsey. DNA methylation arrays as surrogate measures of cell mixture distribution. *BMC Bioinformatics*, 13(1):86. ISSN 1471-2105. doi: 10.1186/1471-2105-13-86. Citado na pág. 11, 17

Houseman *et al.***(2015)** E Andres Houseman, Karl T Kelsey, John K Wiencke e Carmen J Marsit. Cell-composition effects in the analysis of DNA methylation array data: a mathematical perspective. *BMC Bioinformatics*, 16(1):95. ISSN 1471-2105. doi: 10.1186/s12859-015-0527-y. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4392865{&}tool=pmcentrez{&}rendertype=abstract. Citado na pág. 11

Hsiao *et al.***(2014)** Ching-Lin Hsiao, Ai-Ru Hsieh, Ie-Bin Lian, Ying-Chao Lin, Hui-Min Wang e Cathy S J Fann. A novel method for identification and quantification of consistently differentially methylated regions. *PloS one*, 9(5):e97513. ISSN 1932-6203. doi: 10.1371/journal.pone.0097513. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4018258{&}tool=pmcentrez{&}rendertype=abstract. Citado na pág. 2

Irizarry(2009) Rafael A Irizarry. *Genome-wide methylation analysis of human colon cancer reveals similar hypo- and hypermethylation at conserved tissue-specific CpG island shores*. NIH Public Access. Citado na pág. 2

Irizarry *et al.***(2003)** Rafael a Irizarry, Bridget Hobbs, Francois Collin, Yasmin D Beazer-Barclay, Kristen J Antonellis, Uwe Scherf e Terence P Speed. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics (Oxford, England)*, 4(2): 249–264. ISSN 1465-4644. doi: 10.1093/biostatistics/4.2.249. Citado na pág. 11

Irizarry *et al.***(2008)** Rafael a Irizarry, Christine Ladd-Acosta, Benilton Carvalho, Hao Wu, Sheri a Brandenburg, Jeffrey a Jeddeloh, Bo Wen e Andrew P Feinberg. Comprehensive high-throughput arrays for relative methylation (CHARM). *Genome research*, 18(5):780–90. ISSN 1088-9051. doi: 10.1101/gr.7301508. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2336799{&}tool=pmcentrez{&}rendertype=abstract. Citado na pág. 2, 8

Jaffe e Irizarry(2014) Andrew E Jaffe e Rafael a Irizarry. Accounting for cellular heterogeneity is critical in epigenome-wide association studies. *Genome biology*, 15(2):R31. ISSN 1465-6914. doi: 10.1186/gb-2014-15-2-r31. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4053810{&}tool=pmcentrez{&}rendertype=abstract. Citado na pág. 11

Jaffe *et al.***(2012)** Andrew E Jaffe, Peter Murakami, Hwajin Lee, Jeffrey T Leek, M Daniele Fallin, Andrew P Feinberg e Rafael a Irizarry. Bump hunting to identify differentially methylated regions in epigenetic epidemiology studies. *International journal of epidemiology*, 41(1):200–9. ISSN 1464-3685. doi: 10.1093/ije/dyr238. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3304533{&}tool=pmcentrez{&}rendertype=abstract. Citado na pág. 12

Johnson *et al.***(2007)** W. Evan Johnson, Cheng Li e Ariel Rabinovic. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics*, 8(1):118–127. ISSN 14654644. doi: 10.1093/biostatistics/kxj037. Citado na pág. 10

Johnstone *et al.***(2013)** Daniel Johnstone, Carlos Riveros, Moones Heidari, Ross Graham, Debbie Trinder, Regina Berretta, John Olynyk, Rodney Scott, Pablo Moscato e Elizabeth Milward. Evaluation of Different Normalization and Analysis Procedures for Illumina Gene Expression Microarray Data Involving Small Changes. *Microarrays*, 2(2):131–152. ISSN 2076-3905. doi: 10.3390/microarrays2020131. URL http://www.mdpi.com/2076-3905/2/2/131/. Citado na pág. 37

Jones(2012) Peter a Jones. Functions of DNA methylation: islands, start sites, gene bodies and beyond. *Nature reviews. Genetics*, 13(7):484–92. ISSN 1471-0064. doi: 10.1038/nrg3230. URL http://www.ncbi.nlm.nih.gov/pubmed/22641018. Citado na pág. 12

Kinoshita *et al.***(2013)** Makoto Kinoshita, Shusuke Numata, Atsushi Tajima, Shinji Shimodera, Shinji Ono, Akira Imamura, Jun-ichi Iga, Shinya Watanabe, Kumiko Kikuchi, Hiroko Kubo *et al.* Dna methylation signatures of peripheral leukocytes in schizophrenia. *Neuromolecular medicine*, 15(1):95–101. Citado na pág. 11

Kuan e Chiang(2012) Pei Fen Kuan e Derek Y Chiang. Integrating prior knowledge in multiple testing under dependence with applications to detecting differential dna methylation. *Biometrics*, 68(3):774–783. Citado na pág. 12

Leek *et al.***(2010)** Jeffrey T Leek, Robert B Scharpf, Héctor Corrada Bravo, David Simcha, Benjamin Langmead, W Evan Johnson, Donald Geman, Keith Baggerly e Rafael a Irizarry. Tackling the widespread and critical impact of batch effects in high-throughput data. *Nature reviews. Genetics*, 11(10):733–739. ISSN 1471-0056. doi: 10.1038/nrg2825. URL http://dx.doi.org/10.1038/nrg2825. Citado na pág. 10

Leek *et al.***(2012)** Jeffrey T. Leek, W. Evan Johnson, Hilary S. Parker, Andrew E. Jaffe e John D. Storey. The SVA package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics*, 28(6):882–883. ISSN 13674803. doi: 10.1093/bioinformatics/bts034. Citado na pág. 10

Li *et al.***(2015)** Dongmei Li, Zidian Xie, Marc Le Pape e Timothy Dye. An evaluation of statistical methods for DNA methylation microarray data analysis. *BMC Bioinformatics*, 16(1):217. ISSN 1471-2105. doi: 10.1186/s12859-015-0641-x. URL http://www.ncbi.nlm.nih.gov/pubmed/26156501. Citado na pág. 15, 37

Makismovic *et al.***(2012)** Jovana Makismovic, Lavinia Gordon e Alicia Oshlack. SWAN: Subset-quantile Within Array Normalization for Illumina Infinium HumanMethylation450 BeadChips. *Genome Biology*, 13(6):R44. ISSN 1465-6906. doi: 10.1186/gb-2012-13-6-r44. URL http://genomebiology.com/2012/13/6/R44. Citado na pág. 11

Mancuso *et al.***(2011)** Francesco M Mancuso, Magda Montfort, Anna Carreras, Andreu Alibés e Guglielmo Roma. HumMeth27QCReport: an R package for quality control and primary analysis of Illumina Infinium methylation data. *BMC research notes*, 4:546. ISSN 1756-0500. doi: 10.1186/1756-0500-4-546. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3285701{&}tool=pmcentrez{&}rendertype=abstract. Citado na pág. 10

Marabita *et al.***(2013)** Francesco Marabita, Malin Almgren, Maléne E Lindholm, Sabrina Ruhrmann, Fredrik Fagerström-Billai, Maja Jagodic, Carl J Sundberg, Tomas J Ekström, Andrew E Teschendorff, Jesper Tegnér e David Gomez-Cabrero. An evaluation of analysis pipelines for DNA methylation profiling using the Illumina HumanMethylation450 BeadChip platform. *Epigenetics*, 8(3):333–46. ISSN 1559-2308. doi: 10.4161/epi.24008. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3669124{&}tool=pmcentrez{&}rendertype=abstract. Citado na pág. 1, 4, 11

Morris e Beck(2014) Tiffany J Morris e Stephan Beck. Analysis pipelines and packages for Infinium HumanMethylation450 BeadChip (450k) data. *Methods (San Diego, Calif.)*, (September): 6–11. ISSN 1095-9130. doi: 10.1016/j.ymeth.2014.08.011. URL http://www.ncbi.nlm.nih.gov/pubmed/25233806. Citado na pág. 8, 10, 11, 12, 15, 19, 37

Morris *et al.***(2014)** Tiffany J. Morris, Lee M. Butcher, Andrew Feber, Andrew E. Teschendorff, Ankur R. Chakravarthy, Tomasz K. Wojdacz e Stephan Beck. ChAMP: 450k Chip Analysis Methylation Pipeline. *Bioinformatics*, 30(3):428–430. ISSN 13674803. doi: 10.1093/bioinformatics/btt684. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3904520{&}tool=pmcentrez{&}rendertype=abstract. Citado na pág. 9, 10, 11, 18

Moser *et al.***(2009)** Dirk Moser, Savira Ekawardhani, Robert Kumsta, Haukur Palmason, Christoph Bock, Zoi Athanassiadou, Klaus-Peter Lesch e Jobst Meyer. Functional analysis of a potassium-chloride co-transporter 3 (SLC12A6) promoter polymorphism leading to an additional DNA methylation site. *Neuropsychopharmacology : official publication of the American College of Neuropsychopharmacology*, 34(2):458–467. ISSN 0893-133X. doi: 10.1038/npp.2008.77. Citado na pág. 12

Pereira *et al.***(2012)** Carlos Pereira, Carlos A de B. Pereira e Adriano Polpo. MedOr: Order of Medians Based on Confidence Statements. *arXiv.org*. URL http://arxiv.org/abs/1212.5405. Citado na pág. 15, 16

Peters *et al.***(2014)** TJ Peters, M Buckley, AL Statham, R Pidsley, SJ Clark e PL Molloy. Dmrcate: Illumina 450 k methylation array apatial analysis methods. *R package version*, 1(0). Citado na pág. 12

Pidsley *et al.***(2013)** Ruth Pidsley, Chloe C Y Wong, Manuela Volta, Katie Lunnon, Jonathan Mill e Leonard C Schalkwyk. A data-driven approach to preprocessing Illumina 450K methylation array data. *BMC genomics*, 14:293. ISSN 1471-2164. doi: 10.1186/1471-2164-14-293. URL http://www.biomedcentral.com/1471-2164/14/293$\delimiter"026E30F$nhttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3769145{&}tool=pmcentrez{&}rendertype=abstract. Citado na pág. 6, 8, 10, 11, 18

Poage *et al.***(2013)** Graham M. Poage, Rondi A. Butler, E. Andrés Houseman, Michael D. McClean, Heather H. Nelson, Brock C. Christensen, Carmen J. Marsit e Karl T. Kelsey. Identification of an Epigenetic Profile Classifier That Is Associated with Survival in Head and Neck Cancer. 72(11):2728–2737. doi: 10.1158/0008-5472.CAN-11-4121-T.Identification. Citado na pág. 12

Portela e Esteller(2010) Anna Portela e Manel Esteller. Epigenetic modifications and human disease. *Nature biotechnology*, 28(10):1057–68. ISSN 1546-1696. doi: 10.1038/nbt.1685. URL http://www.ncbi.nlm.nih.gov/pubmed/20944598. Citado na pág. 1, 2, 3

Rakyan *et al.***(2011)** Vardhman K Rakyan, Thomas a Down, David J Balding e Stephan Beck. Epigenome-wide association studies for common human diseases. *Nature reviews. Genetics*, 12 (8):529–541. ISSN 1471-0056. doi: 10.1038/nrg3000. URL http://dx.doi.org/10.1038/nrg3000. Citado na pág. 11

Raval *et al.***(2007)** Aparna Raval, Stephan M Tanner, John C Byrd, Elizabeth B Angerman, James D Perko, Shih-Shih Chen, Björn Hackanson, Michael R Grever, David M Lucas, Jennifer J Matkovic *et al.* Downregulation of death-associated protein kinase 1 (dapk1) in chronic lymphocytic leukemia. *Cell*, 129(5):879–890. Citado na pág.

Reik *et al.***(2001)** W Reik, W Dean e J Walter. Epigenetic reprogramming in mammalian development. *Science (New York, N.Y.)*, 293(5532):1089–93. ISSN 0036-8075. doi: 10.1126/science.1063443. URL http://www.ncbi.nlm.nih.gov/pubmed/11498579. Citado na pág. 2

Reinius *et al.***(2012)** Lovisa E Reinius, Nathalie Acevedo, Maaike Joerink, Göran Pershagen, Sven-Erik Dahlén, Dario Greco, Cilla Söderhäll, Annika Scheynius e Juha Kere. Differential DNA methylation in purified human blood cells: implications for cell lineage and studies on disease susceptibility. *PloS one*, 7(7):e41361. ISSN 1932-6203. doi: 10.1371/journal.pone.0041361. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3405143{&}tool=pmcentrez{&}rendertype=abstract. Citado na pág. 11

Robinson *et al.***(2012)** Mark D Robinson, Dario Strbenac, Clare Stirzaker, Aaron L Statham, Jenny Song, Terence P Speed e Susan J Clark. Copy-number-aware differential analysis of quantitative DNA sequencing data. páginas 1–8. doi: 10.1101/gr.139055.112.Freely. Citado na pág. 11

Sandoval *et al.***(2011)** Juan Sandoval, Holger a. Heyn, Sebastian Moran, Jordi Serra-Musach, Miguel Angel Pujana, Marina Bibikova e Manel Esteller. Validation of a DNA methylation microarray for 450,000 CpG sites in the human genome. *Epigenetics*, 6(6):692–702. ISSN 1559-2294. doi: 10.4161/epi.6.6.16196. URL http://www.landesbioscience.com/journals/epigenetics/article/16196/. Citado na pág. 1, 2, 4, 11

Siegmund(2011) Kimberly D Siegmund. Statistical approaches for the analysis of DNA methylation microarray data, jun 2011. ISSN 03406717. URL http://www.ncbi.nlm.nih.gov/pubmed/21519831. Citado na pág. 11

Siegmund *et al.***(2012)** Kimberly D Siegmund, Jeremy G Stone, Sandra L Siedlak, Massimo Tabaton, Asao Hirano, Rudy J Castellani, Corrado Santocanale, George Perry, Mark a Smith, Xiongwei Zhu, Hyoung-gon Lee, Mark H Chin, Mike J Mason, Wei Xie, Stefano Volinia, Mike Singer, Gayane Ambartsumyan, Otaren Aimiuwu, Laura Richter, Jin Zhang, Vanessa Ott, Michael Grunstein, Neta Lavon, Nissim Benvenisty, M Croce, Amander T Clark, Tim Baxter, April D Pyle e Mike a Teitell. Statistical Approaches for the Analysis of DNA Methylation Microarray Data. 22(6):1338–1344. doi: 10.1007/s00439-011-0993-x.Statistical. Citado na pág. 8

Slieker *et al.***(2013)** Roderick C Slieker, Steffan D Bos, Jelle J Goeman, Judith Vmg Bovée, Rudolf P Talens, Ruud van der Breggen, H Eka D Suchiman, Eric-Wubbo Lameijer, Hein Putter,

Erik B van den Akker, Yanju Zhang, J Wouter Jukema, P Eline Slagboom, Ingrid Meulenbelt e Bastiaan T Heijmans. Identification and systematic annotation of tissue-specific differentially methylated regions using the Illumina 450k array. *Epigenetics & chromatin*, 6(1):26. ISSN 1756-8935. doi: 10.1186/1756-8935-6-26. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3750594{&}tool=pmcentrez{&}rendertype=abstract. Citado na pág. 12

Smith e Meissner(2013) Zachary D Smith e Alexander Meissner. DNA methylation: roles in mammalian development. *Nature reviews. Genetics*, 14(3):204–20. ISSN 1471-0064. doi: 10.1038/nrg3354. URL http://www.ncbi.nlm.nih.gov/pubmed/23400093. Citado na pág. 1

Sofer *et al.*(2013) Tamar Sofer, Elizabeth D Schifano, Jane A Hoppin, Lifang Hou e Andrea A Baccarelli. A-clustering: a novel method for the detection of co-regulated methylation regions, and regions associated with exposure. *Bioinformatics*, página btt498. Citado na pág. 12

Storey e Tibshirani(2003) John D Storey e Robert Tibshirani. Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences of the United States of America*, 100(16):9440–9445. ISSN 0027-8424. doi: 10.1073/pnas.1530509100. Citado na pág. 13

Straussman *et al.*(2009) Ravid Straussman, Deborah Nejman, Douglas Roberts, Israel Steinfeld, Barak Blum, Nissim Benvenisty, Itamar Simon, Zohar Yakhini e Howard Cedar. Developmental programming of CpG island methylation profiles in the human genome. *Nature structural & molecular biology*, 16(5):564–71. ISSN 1545-9985. doi: 10.1038/nsmb.1594. URL http://www.ncbi.nlm.nih.gov/pubmed/19377480. Citado na pág. 1

Sun *et al.*(2011) Zhifu Sun, High Chai, Yanhong Wu, Wendy M White, Krishna V Donkena, Christopher J Klein, Vesna D Garovic, Terry M Therneau e Jean-Pierre a Kocher. Batch effect correction for genome-wide methylation data with Illumina Infinium platform. *BMC Medical Genomics*, 4(1):84. ISSN 1755-8794. doi: 10.1186/1755-8794-4-84. URL http://www.biomedcentral.com/1755-8794/4/84. Citado na pág. 11

Suzuki e Bird(2008) Miho M Suzuki e Adrian Bird. DNA methylation landscapes: provocative insights from epigenomics. *Nature reviews. Genetics*, 9(6):465–76. ISSN 1471-0064. doi: 10.1038/nrg2341. URL http://www.ncbi.nlm.nih.gov/pubmed/18463664. Citado na pág. 1

Teschendorff *et al.*(2011) Andrew E. Teschendorff, Joanna Zhuang e Martin Widschwendter. Independent surrogate variable analysis to deconvolve confounding factors in large-scale microarray profiling studies. *Bioinformatics*, 27(11):1496–1505. ISSN 13674803. doi: 10.1093/bioinformatics/btr171. Citado na pág. 10, 11

Teschendorff *et al.*(2013) Andrew E Teschendorff, Francesco Marabita, Matthias Lechner, Thomas Bartlett, Jesper Tegner, David Gomez-Cabrero e Stephan Beck. A beta-mixture quantile normalization method for correcting probe design bias in Illumina Infinium 450 k DNA methylation data. *Bioinformatics (Oxford, England)*, 29(2):189–96. ISSN 1367-4811. doi: 10.1093/bioinformatics/bts680. URL http://bioinformatics.oxfordjournals.org/content/29/2/189.full. Citado na pág. 11, 19

Tost *et al.*(2003) Jörg Tost, Jenny Dunker e Ivo Glynne Gut. Analysis and quantification of multiple methylation variable positions in CpG islands by Pyrosequencing. *BioTechniques*, 35(1):152–6. ISSN 0736-6205. URL http://www.ncbi.nlm.nih.gov/pubmed/12866415. Citado na pág. 4

Touleimat e Tost(2012) Nizar Touleimat e Jörg Tost. Complete pipeline for Infinium(®) Human Methylation 450K BeadChip data processing using subset quantile normalization for accurate DNA methylation estimation. *Epigenomics*, 4(3):325–41. ISSN 1750-192X. doi: 10.2217/epi.12.21. URL http://www.ncbi.nlm.nih.gov/pubmed/22690668. Citado na pág. 10, 11, 19

Wang *et al.***(2012)** Dan Wang, Li Yan, Qiang Hu, Lara E Sucheston, Michael J Higgins, Christine B Ambrosone, Candace S Johnson, Dominic J Smiraglia e Song Liu. IMA: an R package for high-throughput analysis of Illumina's 450K Infinium methylation data. *Bioinformatics (Oxford, England)*, 28(5):729–30. ISSN 1367-4811. doi: 10.1093/bioinformatics/bts013. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3289916{&}tool=pmcentrez{&}rendertype=abstract. Citado na pág. 10, 12

Wang(2012) Shuang Wang. Method to Detect Differentially Methylated Loci with Case- Control Designs using Illumina Arrays. *Genet Epidemiol*, 35(7):686–694. doi: 10.1002/gepi.20619. Method. Citado na pág. 12

Wang *et al.***(2015)** Ting Wang, Weihua Guan, Jerome Lin, Nadia Boutaoui, Glorisa Canino, Jianhua Luo, Juan Carlos Celedón e Wei Chen. A systematic study of normalization methods for Infinium 450K methylation data using whole-genome bisulfite sequencing data. *Epigenetics : official journal of the DNA Methylation Society*, 2294(June):37–41. ISSN 1559-2308. doi: 10.1080/15592294.2015.1057384. URL http://www.ncbi.nlm.nih.gov/pubmed/26036609. Citado na pág. 11

Wilhelm-Benartzi *et al.***(2013)** C S Wilhelm-Benartzi, D C Koestler, M R Karagas, J M Flanagan, B C Christensen, K T Kelsey, C J Marsit, E a Houseman e R Brown. Review of processing and analysis methods for DNA methylation array data. *Br J Cancer*, 109(6):1394–402. ISSN 1532-1827. doi: 10.1038/bjc.2013.496. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3777004{&}tool=pmcentrez{&}rendertype=abstract. Citado na pág. 8, 10

Xu *et al.***(2007)** Jian Xu, Scott D Pope, Ali R Jazirehi, Joanne L Attema, Peter Papathanasiou, Jason a Watts, Kenneth S Zaret, Irving L Weissman e Stephen T Smale. Pioneer factor interactions and unmethylated CpG dinucleotides mark silent tissue-specific enhancers in embryonic stem cells. *Proceedings of the National Academy of Sciences of the United States of America*, 104(30):12377–12382. ISSN 0027-8424. doi: 10.1073/pnas.0704579104. Citado na pág. 12

Yousefi *et al.***(2013)** P Yousefi, K Huen, R A Schall, A Decker, E Elboudwarej, H Quach, L Barcellos e N Holland. Considerations for normalization of DNA methylation data by Illumina 450K BeadChip assay in population studies. *Epigenetics*, 8(11):1141–1152. ISSN 15592294. doi: 10.4161/epi.26037. URL http://www.ncbi.nlm.nih.gov/pubmed/23959097. Citado na pág. 11

Zhang *et al.***(2014)** Xin Zhang, Renqian Du, Shilin Li, Feng Zhang, Li Jin e Hongyan Wang. Evaluation of copy number variation detection for a SNP array platform. *BMC bioinformatics*, 15(1):50. ISSN 1471-2105. doi: 10.1186/1471-2105-15-50. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4015297{&}tool=pmcentrez{&}rendertype=abstract. Citado na pág. 11

Zhang *et al.***(2011)** Yan Zhang, Hongbo Liu, Jie Lv, Xue Xiao, Jiang Zhu, Xiaojuan Liu, Jianzhong Su, Xia Li, Qiong Wu, Fang Wang e Ying Cui. QDMR: a quantitative method for identification of differentially methylated regions by entropy. *Nucleic acids research*, 39(9):e58. ISSN 1362-4962. doi: 10.1093/nar/gkr053. URL http://www.scopus.com/inward/record.url?eid=2-s2.0-79956006215{&}partnerID=tZOtx3y1http://nar.oxfordjournals.org/content/early/2011/02/08/nar.gkr053.full. Citado na pág. 12

Zhuang *et al.***(2012)** J Zhuang, M Widschwendter e Ae Teschendorff. A comparison of feature selection and classification methods in DNA methylation studies using the Illumina Infinium platform. *BMC Bioinformatics*, 13(1):59. ISSN 1471-2105. doi: 10.1186/1471-2105-13-59. URL http://discovery.ucl.ac.uk/1346600/. Citado na pág. 8, 12, 19, 37

Zilberman e Henikoff(2007) Daniel Zilberman e Steven Henikoff. Genome-wide analysis of DNA methylation patterns. *Development (Cambridge, England)*, 134:3959–3965. ISSN 0950-1991. doi: 10.1242/dev.001131. Citado na pág. 2

Glossário

Agrupamento Metil (CH3): Metil é um grupo funcional derivado do metano, contendo um átomo de carbono ligados a três átomos de hidrogênio.

R/Bioconductor: Uma ferramenta de linha de comando para processamento de dados, análises estatísticas e visualização de dados biológicos. (Bock [2012]).

valor β: Um termo para o nível de metilação do DNA, o qual se asemelha a uma distribuição Beta. (Bock [2012]).

valor M: são valores β transformados. A transformação atênua alguns problemas estatísticos do valor β (ou seja, intervalo de valor limitado e distribuição fortemente bimodal) ao custo de reduzir a interpretabilidade biológica. (Bock [2012]).

Efeitos de lote: desvios sistemáticos nos dados que não estão relacionados com a questão de pesquisa, mas que são provenientes de diferenças indesejáveis (e muitas vezes não reconhecidos) no manuseamento das amostras. (Bock [2012]).

Taxa de falsos descobertos (False discovery rate - FDR): proporção de erros devido à rejeição incorreta da hipótese H0.

Effect size: Uma medida para a força de associação entre duas variaveis que fornece informações complementares importantes para os p valores e para as taxas de descoberta de falsos positivos. (Bock [2012]).

Tiling map: Segmentação do genoma em janelas de tamanho fixo, tipicamente pequenas.

Regiões diferencialmente metiladas (Differentially methylated regions - DMRs): regiões genômicas que apresentam diferenças de metilação do DNA entre amostras, com significancia estatistica. (Bock [2012]).

RefSeq: Um conjunto de sequências integradas, não redundantes e bem anotadas de genomas, transcritos e proteinas.

Microarranjo: Microarranjos de DNA consistem num conjunto ordenado de milhares de moléculas de DNA cuja seqüência é conhecida. Dessa forma é criada uma matriz de fragmentos genéticos distintos e posicionados numa ordem pré-definida, a qual pode ter sua imagem capturada bem como digitalizada. Isso permite avaliar a expressão de milhares de genes simultaneamente através de métodos de processamento computacional de imagens.

Enzima DNA metiltransferase: transfere o grupo metil para o DNA (responsável pela metilação da citosina).

Dinucleotídeos CpG: citosina que geralmente precede a uma guanina.

Elementos repetitivos: trechos do DNA que são iguais a outros trechos do mesmo DNA.

Piro sequenciamento: é um método de sequenciamento de DNA (que estabelece a ordem de nucleotídeos no DNA) com base no princípio do "sequenciamento por síntese".

Sequenciamento: é uma série de métodos bioquímicos que têm como finalidade determinar a ordem das bases nitrogenadas adenina (A), guanina (G), citosina (C) e timina (T) da molécula de DNA.

Arquivos IDAT: formato usado para armazenar dados de microarranjo.

Imprinting: Normalmente tanto a copia materna quanto a copia paterna de cada gene tem o mesmo potencial para estarem ativos em alguma celula. No entanto o imprinting é um mecanismo epigenetico que modifica este potencial, restringindo a expressão de um gene para um dos dois cromossomos.

Transposons: são trechos de DNA que podem mudar seu número de cópias no genoma ou se mover nele, mudando de posição.