Universidade de Brasília
Departamento de Estatística

Análise de Perfil de Metilação de Genoma Completo em Câncer Colorretal

Tamara Talita Rodrigues Dias

Orientadora: Prof. Dra. Joanlise Marco de Leon Andrade

Brasília
2021

Tamara Talita Rodrigues Dias

Análise de Perfil de Metilação de Genoma Completo em Câncer Colorretal

Orientadora:
Prof. Dra. Joanlise Marco de Leon Andrade

Relatório apresentado para o Departamento de Estatística, Instituto de Ciências Exatas, Universidade de Brasília, como parte dos requisitos necessários para o grau de Bacharel em Estatística.

Brasília
2021

Agradecimentos

Aos meus pais Sérgio e Fátima, por todo o apoio em toda a minha caminhada. Por serem zelosos e por sempre darem o máximo para a boa criação dos filhos, são impecáveis. Tenho muita sorte. Ao meu irmão Pedro, que mesmo mais novo tem me ensinado muita coisa.

Ao Gabriel Reis, meu namorado, que para além de muitos outros aspectos contribuiu muito na minha vida acadêmica, me ensinou a ter mais disciplina e ser mais crítica. Meu maior exemplo.

À minha família do Ceará, meus avós, tias, tios, primos. Exemplos de pessoas humildes e resilientes, tenho muito orgulho.

Ao Matheus Cavalcanti, Renan Almeida e todos os Illuminati, meus amigos da vida, por terem mantido minha sanidade durante os anos que estudamos e por toda amizade.

Aos amigos que a ESTAT me proporcionou, principalmente aos amigos do Social que levo no coração. À Isabela Harumi O. Yamaguchi e Leslie Miho, minhas parceiras no curso. À Isabela Harumi L. Motoki, por ter se tornado uma grande amiga.

À todas as professoras e professores que contribuíram direta ou indiretamente para que eu pudesse chegar até aqui, especialmente às professoras Juliana Betini e Maria Teresa Leão, por serem muito humanas e por demonstrarem tanto carinho com a profissão. Ao professor George Freitas von Borries pelos ensinamentos e conselhos.

À professora Joanlise Marco de Leon Andrade, a melhor orientadora que eu poderia ter. Extremamente atenciosa e compreensível. Uma inspiração como mulher e como profissional. Agradeço enormemente por ter contribuído com tantos ensinamentos.

Ao Departamento de Estatística, à Universidade de Brasília e todo o seu corpo docente que sempre proporcionaram um ensino de qualidade.

Por fim, à todas as mulheres que lutaram e lutam para que todas nós tenhamos direito à educação.

Resumo

Introdução: O câncer colorretal (CCR) é um dos tipos de câncer que mais afeta homens e mulheres no Brasil. Estudos prévios já identificaram diversos genes associados ao CCR. O processo de metilação envolve modificações químicas que podem causar alteração na expressão gênica. Este trabalho teve por objetivo a análise de perfil de metilação de genoma completo para a identificação de posições (DMPs) e regiões diferencialmente metiladas (DMRs) entre indivíduos com e sem câncer colorretal.

Metodologia: Foram comparados dados de metilação de indivíduos com câncer colorretal (n = 4) e controles saudáveis (n = 4) utilizando-se o chip *Infinium MethylationEPIC BeadChip* (EPIC), que mede níveis de metilação em mais de 886 mil *probes*. Procedimentos de controle de qualidade e de normalização pelos métodos `Funnorm` e `Relic` foram realizados.

Resultados: Foram identificadas 111 DMRs, contendo 58 genes distintos, dos quais *GPSM2*, *LOC647121*, *RUNX3* e *IGFBP3* foram previamente identificados como prognósticos de câncer colorretal, além de outros genes biomarcadores de CCR e de genes que possuem relação com supressão de tumor.

Conclusão: Mesmo com amostras pequenas, foi possível se identificar muitos genes em regiões diferencialmente metiladas. Mais estudos são necessários para confirmação de tais resultados e melhor entendimento das funções desses genes no desenvolvimento e progressão de CCR.

Palavras-chave: câncer colorretal, metilação, metiloma, epigenética, epigenoma, genoma completo, regiões diferencialmente metiladas, DMR, minfi, bumphuter.

Abstract

Introduction: Colorectal cancer (CRC) is one of the types of cancer that most affects men and women in Brazil. Previous studies have already identified several genes associated with CRC. Methylation processes involves chemical modifications that can cause changes in gene expression. This paper aimed to analyze genome-wide DNA methylation profiling for the identification of positions (DMPs) and differentially methylated regions (DMRs) between individuals with and without colorectal cancer.

Methods: Methylation data from individuals with colorectal cancer (n = 4) and healthy controls (n = 4) were compared using the *Infinium MethylationEPIC BeadChip* (EPIC), which measures methylation levels in more than 886 thousand probes. Quality control and normalization procedures using the `Funnorm` and `Relic` methods were performed.

Results: 111 DMRs were identified, with 58 distinct genes, of which *GPSM2*, *LOC647121*, *RUNX3* and *IGFBP3* were previously identified as colorectal cancer prognosis, in addition to other biomarker genes of CRC and genes related to tumor suppression.

Conclusions: Even with small samples, it was possible to identify many genes in differentially methylated regions. Further studies are needed to confirm these results and to better understand the functions of these genes in the development and progression of CRC.

Keywords: colorectal cancer, methylation, methylome, epigenetics, epigenome, genome-wide, differentially methylated regions, DMR, minfi, bumphunter.

Lista de Ilustrações

Lista de Tabelas e Quadros

Lista de Tabelas

Lista de Quadros

Lista de abreviaturas e siglas

CCR	Câncer Colorretal
CpG	Citosina que precede Guanina
DNA	Ácido Desoxirribonucleico
FDR	*False Discovery Rate*
Funnorm	*Functional Normalization*
FWER	*Family-wise Error Rate*
DMP	Posição Diferencialmente Metilada
DMR	Região Diferencialmente Metilada
IDAT	Dados de Intensidade
NOOB	*Normal-exponential out-of-band*
Relic	*Regression on Logarithm of Internal Control Probes*
SNP	Polimorfismo de Nucleotídeo Único
SWAN	*Subset-quantile Within Array Normalization*

Sumário

1 Introdução

Muitos estudos têm permitido a avaliação e identificação de diferentes fatores genéticos associados ao câncer. Um exemplo são as pesquisas feitas no ramo da epigenética. A epigenética estuda alterações hereditárias que ocorrem no código genético, sem modificação direta na sequência do DNA, ou seja, são alterações químicas que causam alterações na expressão dos genes, das proteínas, que podem levar a eventuais alterações em funções celulares (EGGER et al., 2004).

O câncer é o resultado de mutações genéticas que ocorrem no DNA, causando um crescimento desordenado das células que formam tumores (INCA, 2019). Estudos mostram que as alterações epigenéticas afetam quase todas as etapas de desenvolvimento de um tumor cancerígeno. A metilação é um tipo de modificação epigenética que ocorre somente na molécula de DNA (JONES; BAYLIN, 2002; OLIVEIRA et al., 2010).

No genoma de mamíferos a metilação ocorre em partes específicas dos genes e envolve alterações químicas na citosina que geralmente precede uma guanina (dinucleotídeo CpG) (OLIVEIRA et al., 2010). Nucleotídeos são os componentes básicos do DNA que consistem em quatro elementos: citosina (C), guanina (G), adenina (A) e timina (T) (BRYSON, 2004). Um dinucleotídeo contém dois nucleotídeos, sendo os nucleotídeos C e G no caso do CpG. Não há muitos dinucleotídeos CpG espalhados pelo DNA, mas no gene existem as chamadas ilhas CpG que contêm densidades mais altas de C e G. Há algumas ilhas CpG espalhadas pelo código genético (BIRD, 2002; OLIVEIRA et al., 2010).

As regiões no DNA codificantes de muitos genes contém CpGs que podem ser metilados e servem como *hotspots* em doenças genéticas humanas. *Hotspots* são os locais em que as mutações ocorrem a uma taxa até 100 vezes mais que o normal (JONES et al., 1992).

Atualmente a metilação é reconhecida como um processo epigenético muito importante, que influencia as atividades dos genes. Ter o padrão correto de metilação é primordial para manter células e órgãos saudáveis. Quando a metilação é desregulada, vários tipos de distúrbios podem ocorrer, levando ao surgimento de doenças (MOORE; LE; FAN, 2013; COSTELLO; PLASS, 2001). No caso de câncer, alguns genes que deveriam estar reprimidos são expressos em função de metilação desregulada e acabam favorecendo, por exemplo, a proliferação celular. Da mesma forma, há genes supressores de tumores que deviam estar ativos mas são silenciados, contribuindo para o desenvolvimento e progressão do câncer. Os tumores possuem a capacidade de reprimir tais genes pela metilação de DNA. (OLIVEIRA et al., 2010).

Estudos realizados pelo do INCA de 2020 mostram que, no Brasil, o câncer colorretal é o segundo tipo de câncer que mais afetou mulheres e homens. Câncer colorretal

é um tumor maligno que se desenvolve no intestino grosso, ou seja, no cólon ou no reto, que é a parte final do intestino. Em 2018 o câncer colorretal foi o terceiro tipo da doença que mais levou mulheres e homens a óbito no Brasil (INCA, 2020).

Nesse contexto, o objetivo do presente trabalho é avaliar dados de metilação em todo o genoma entre indivíduos com e sem câncer colorretal. Para isso serão realizados procedimentos de controle de qualidade e filtragem a fim de identificar posições e regiões diferencialmente metiladas, e assim detectar os genes localizados nessas regiões e a relação desses com o câncer colorretal.

2 Metodologia

A análise de perfil de metilação de genoma completo envolve algumas etapas. Inicialmente foram realizados procedimentos de controle de qualidade para filtrar amostras de tecido e *probes* de baixa qualidade. Valores foram normalizados e então utilizados para identificar posições e regiões diferencialmente metiladas.

Probes são sequências de DNA usadas para pesquisar sua sequência complementar em uma amostra do genoma. Maiores detalhes serão dados no decorrer da descrição metodológica.

A Figura 1 apresenta um fluxograma com as etapas das análises realizadas neste trabalho. Todas as siglas e etapas estão descritas no texto. Abaixo de algumas etapas estão os pacotes e as funções utilizadas, respectivamente, no *software* `R`.

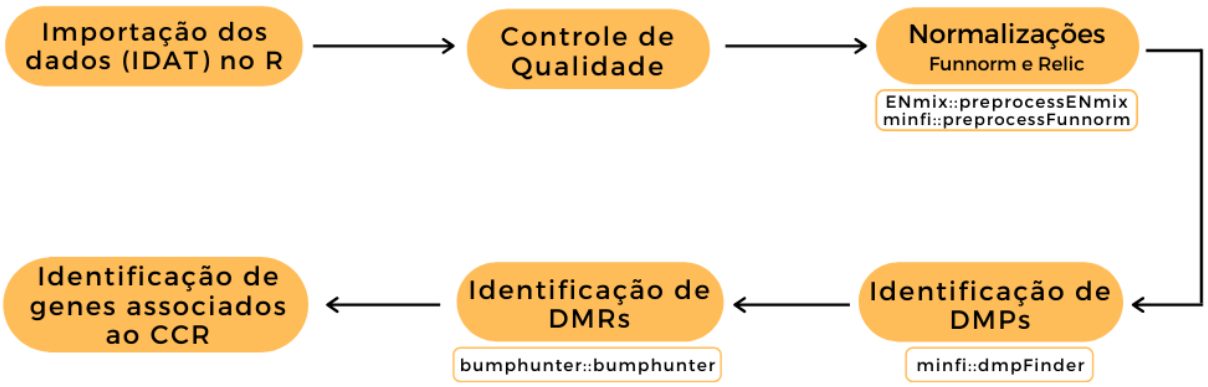

Figura 1: Fluxograma das etapas e análises realizadas.

Fonte: Elaboração própria.

2.1 Conjunto de Dados

Os dados são refentes à amostras de tecido de 12 indivíduos: 4 com câncer colorretal, 4 com pólipos benignos e 4 controles saudáveis (não apresentavam pólipos ou câncer). A amostra analisada tem relação com o estudo de PEREIRA, 2017, em que 106 indivíduos foram submetidos ao rastreamento do CCR.

As observações são comumente denominadas amostras (de tecidos ou de sangue) no contexto de estudos genéticos. Para evitar equívoco em relação ao termo estatístico, será empregado o termo “amostra(s) de tecido”.

As amostras de tecido foram coletadas em biópsia realizada durante o exame de

colonoscopia a qual os participantes foram submetidos. Nos casos com suspeita de CCR foram coletadas até três amostras em diferentes locais da lesão. O material coletado em pólipos foi dividido em duas partes: para análises histopatológicas e moleculares. Nos grupos controle as amostras foram coletadas a 15 cm da borda anal.

A amostragem foi de conveniência, com captação de participantes na Clínica do Aparelho Digestivo, especializada em colonoscopia, e no Hospital Dom Orione, ambas instituições privadas, e no Hospital Regional de Araguaína, uma instituição pública, todos localizados na cidade de Araguaína, no estado do Tocantins. Foram convidados a participar indivíduos submetidos a colonoscopia, sendo incluídos aqueles que aceitaram participar desta pesquisa e que atendiam aos critérios de inclusão. A captação desses indivíduos ocorreu no período de julho de 2014 a julho de 2015, perfazendo um total de 13 meses.

A Tabela 1 contém informações de cada grupo. Casos e pólipos tinham idades próximas em média (54,75 $\pm$ 3,4 anos para casos e 56,25 $\pm$ 5,38 para pólipos). O grupo controle saudável tinha média e desvio padrão de idade um pouco superiores (60,3 $\pm$ 7,77 anos), mas não continha a informação de uma pessoa.

O ideal seria que todos os grupos tivessem o mesmo número de mulheres e homens, mas o grupo de pessoas com pólipo foi composto apenas por mulheres enquanto que nos outros grupos metade dos indivíduos eram do sexo feminino.

Tabela 1: Análise descritiva por grupo de indivíduos

Grupo	**Média da idade**	**Desvio padrão da idade**	**Proporção de mulheres**
Câncer	54,75	3,40	50%
Controle*	60,30	7,77	50%
Pólipo	56,25	5,38	100%

*o grupo de indivíduos controle saudável continha informação de apenas 3 idades.

O Quadro 1 mostra o resultado do teste de Kruskall-Wallis, realizado a fim de comparar as médias dos grupos. Considerando um nível de significância de 5%, nota-se que o p-valor é superior, ou seja, as médias dos grupos são iguais.

Quadro 1: Teste de comparação dos grupos quanto à idade.

Estatística	G.L	P-valor
1,15	2	0,56

*G.L = graus de liberdade.

2.2 Dados de metilação

Os dados de metilação foram obtidos com base no chip do tipo *Infinium MethylationEPIC BeadChip* (EPIC), desenvolvido pela empresa `Illumina` para fornecer mapas de áreas de metilação do DNA, utilizando amostras de tecido das regiões de interesse. Esse chip permite a medição de 866.836 sequências (ou *probes*) de diferentes locais (ou *sites*) de metilação, sendo um produto mais avançado que o chip do tipo *Infinium HumanMethylation450 BeadChip* (HM450), líder na indústria. O chip EPIC conta com mais de 90% dos CpGs do chip HM450 e 350,000 CpGs adicionais em regiões melhoradoras (ILLUMINA, 2015).

Probe é uma sequência do trecho do DNA usado para buscar sua sequência complementar em uma amostra do genoma. Os chips possuem vários *probes* que são colocados em contato com o código genético do tecido selecionado. Cada *probe* se liga a sua sequência complementar do trecho de DNA da amostra em um processo denominado hibridização (NIH,). Para que o trecho do DNA hibridize ao *probe*, o trecho primeiro passa pela conversão bissulfito de sódio, que converte citosinas (C) não metiladas em timina (T) e deixando a citosina metilada intacta (PIDSLEY et al., 2016).

Os *probes* do chip EPIC foram desenvolvidos para permitir medições nas ilhas CpG, cujas regiões promotoras podem envolver processos de metilação do DNA, que induzem à expressão do gene (PIDSLEY et al., 2016). Os dados metilação da empresa `Illumina` geralmente são obtidos na forma de arquivos de Dados de Intensidade (IDAT). Este é um formato desenvolvido pela empresa, em que os dados são gerado pelo *scanner* que armazena as intensidades de metilação para cada matriz de chip (MAKSIMOVIC; PHIPSON; OSHLACK, 2016).

O nível de metilação de cada *probe* é determinando pelo cálculo da proporção de fluorescência da coloração dos canais metilados e não metilados (ILLUMINA,).

A Figura 2 mostra a estrutura do chip EPIC que comporta 8 amostras.

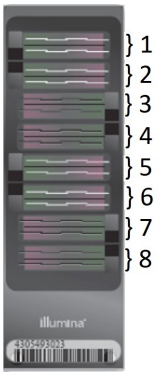

Figura 2: O *Infinium Methylation EPIC BeadChip*.

Fonte: Adaptado de Illumina (2015).

As amostras de material genético dos participantes foram alocadas em 2 chips do tipo EPIC, sendo 4 *arrays* para os indivíduos com pólipo, 4 para o controle saudável em um chip e 4 *arrays* para o grupo de pessoas com câncer em outro chip. Portanto, a disposição do material nos chips traz à tona um problema de delineamento, pois cada chip pode ser considerado um bloco, cujo efeito não pode ser avaliado. O ideal seria alocar aleatoriamente 2 *arrays* de cada grupo em cada chip.

Cada amostra de tecido é medida em somente um *array* e o conjunto de amostras medidas em cada chip é denominado matriz de chip. Cada matriz de chip representa uma matriz com valores de todas as variáveis de metilação para um indivíduo, medidos em dois canais de cores diferentes (vermelho e verde). Para cada CpG há dois tipos de medidas: intensidade metilada (*methylated*, representada por M) e intensidade não metilada (*unmethylated*, representada por U). Por conta da configuração (*design*) dos *probes* utilizados, os sinais são exibidos de duas formas:

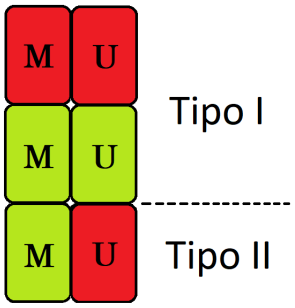

Figura 3: Representação dos tipos de *design* através das cores dos canais e tipos de metilação.

Fonte: Baseado em Fortin; Hansen (2015).

No *design* tipo I há uma cor para cada CpG com um *probe* para o sinal metilado e um *probe* para o sinal não metilado, ou seja, são quatro probes para um CpG. No *design* tipo II apenas um *probe* é utilizado para cada CpG e a cor verde representa o sinal metilado eo vermelho mede o sinal não metilado (PIDSLEY et al., 2016).

O *design* do tipo II ocupa a metade do espaço que o *design* do tipo I ocupa no chip. Devido à características particulares do *design* do tipo I, uma certa porcentagem dos *probes* desse *design* deve ser mantida no chip. Os dois tipos de *design* possuem um intervalo dinâmico de coloração ligeiramente diferente, podendo levar a um erro do tipo II durante a análise (deixar de rejeitar a hipótese nula quando ela é falsa). Esse é um dos motivos pelos quais deve-se realizar normalizações das matrizes de chip (MORRIS; BECK, 2015).

O nível de metilação pode ser expresso por duas medidas, quais sejam o `valor-Beta` e o `valor-M`, calculados como a seguir:

$$Valor\ -Beta = \frac{M}{M + U + 100} \tag{2.2.1}$$

e

$$Valor\ -M = log_2(\frac{M}{U}), \tag{2.2.2}$$

em que M representa o sinal metilado e U, o sinal não metilado. O `valor-Beta` varia entre 0 e 1, sendo próximo de 1 quando houve metilação no gene 0 caso contrário (FORTIN; HANSEN, 2015; HANSEN MARTIN ARYEE, 2020; MAKSIMOVIC; PHIPSON; OSHLACK, 2016). O `valor-M` pode incluir valores negativos e positivos. Além desses valores, há também o Número de Cópias, do inglês *Copy Number* (CN), obtido através da soma dos sinais metilados e não metilados. Em termos biológicos o CN é o número de cópias de um trecho de DNA no genoma de um organismo (BIOLOGY, 2005).

O trabalho de DU et al., 2010 concluiu que o `valor-Beta` tem interpretação biológica mais intuitiva, enquanto que o `valor-Beta` é mais apropriado para a análises estatísticas de diferença nos níveis de metilação. Estudos como os de JAFFE et al., 2012, FORTIN et al., 2014, SLIEKER et al., 2013 e muitos outros utilizaram o `valor-Beta` para esse tipo de análise. Com isso, os dois valores foram empregados na identificação de posições e regiões diferencialmente metiladas neste trabalho.

2.3 Ferramentas Computacionais

As análises foram implementadas utilizando-se os programas `R` (versão 4.0.0) e `RStudio` (versão 1.4.1106). Para analisar e visualizar matrizes de metilação de DNA do chip *Infinium Methylation EPIC BeadChip* empregou-se o pacote `minfi`, que faz parte do projeto `Bioconductor`, projeto de *software* livre, de código aberto e de desenvolvimento aberto para análise e compreensão de dados genômicos (FORTIN; TRICHE; HANSEN, 2017).

2.4 Controle de Qualidade

Alguns procedimentos de controle de qualidade, descritos a seguir, foram utilizados para a eventual exclusão de amostras de tecido ou de *probes* de baixa qualidade.

Inicialmente, calculou-se "p-valores de detecção" de sinal para cada CpG em cada amostra. O método utilizado pelo pacote `minfi` para calcular tais p-valores compara o sinal total (M+U) para cada *probe* com o nível de sinal de fundo (*background signal*), que é estimado a partir dos *probes* de controle negativo, assumindo-se distribuição normal

(HANSEN et al., 2014). O chip EPIC contém 635 *probes*-controle. P-valores muito pequenos são indicativos de um sinal confiável, mas p-valores $> 0{,}01$ indicam um sinal de baixa qualidade.

O cálculo da média dos p-valores de detecção pode ser utilizado para classificar a qualidade geral de cada amostra (indivíduo) com respeito à reprodutibilidade dos sinais. Amostras com muitos *probes* de má qualidade apresentam altas médias de p-valores e devem ser excluídas (MAKSIMOVIC; PHIPSON; OSHLACK, 2016).

Recomenda-se ainda que *probes* afetados por marcadores do tipo SNP (*single nucleotide polymorfism*) sejam excluídos. SNPs envolvem mutações de um par do nucleotídeo na sequência de DNA em pelo menos 1% da população e ocorrem em abundância no genoma (BROOKES, 1999). Os SNPs podem ocorrer por um erro de replicação do genoma, mas também podem ser responsáveis por características diferentes entre indivíduos.

Adicionalmente, foram excluídos *probes* com reação cruzada, que mapeiam mais de uma posição no genoma (MAKSIMOVIC; PHIPSON; OSHLACK, 2016).

Finalmente, excluiu-se *probes* em cromossomos X e Y pois a proporção de indivíduos por sexo era diferente entre os grupos, o que poderia levar à falsa detecção de diferenças de metilação nos cromossomos sexuais entre os grupos.

2.5 Métodos de Normalização

Métodos de normalização são utilizados a fim de tornar as amostras de tecido mais comparáveis, minimizando-se a variação interna (*within*) indesejada, que ocorre devido à uma diferença entre os *designs* tipo I e o tipo II. O tipo II tem média de deslocamento para cima do `valor-Beta` para a intensidade não metilada e deslocamento para baixo para a intensidade metilada. Além disso, os `valores-Beta` obtidos dos *probes* do tipo II têm distribuição mais estreita e são menos reprodutíveis do que aqueles obtidos a partir do tipo I, levando a um viés de *design*. Portanto, a etapa de normalização é essencial (WANG; WU; WANG, 2018).

Exemplos de variação indesejada têm relação com as diferenças em preparação das amostras, com a quantidade de reagente colocado nos chips, diferenças no escaneamento dos chips, entre outros.

Para fins de comparação, valores brutos de metilação foram tratados por 6 métodos de normalização distintos: `Quantílica`, `Funnorm`, `Relic`, `Illumina`, `NOOB` e `SWAN`.

O trabalho de FORTIN et al., 2014 mostrou que a normalização `Funnorm` apresenta melhor desempenho para diminuir a variação técnica entre grupos quando comparada às normalizações `SWAN`, `Quantile`, `NOOB`, dentre outras não utilizadas neste estudo.

FORTIN; HANSEN, 2015, sugeriu o uso da normalização `Funnorm` quando o objetivo envolve análises de comparação global, como em delineamentos do tipo caso-controle, em que o uso da normalização `Quantile` não é recomendado. Os autores também concluíram que as normalizações `SWAN` e `Illumina` não apresentam um desempenho tão bom quanto as outras.

Em paralelo, XU et al., 2017, mostraram que a normalização `Relic`, por eles desenvolvida, é superior à normalização `Illumina`, dentre outras também não utilizadas. Este método de normalização está implementado no pacote `ENmix`, diferente das outras normalizações, implementadas no pacote `minfi`.

Por serem consideradas mais adequadas pela literatura, optou-se pela utilização de dados normalizados pelos métodos `Funnorm` e `Relic` nas análises subsequentes de maior interesse. A seguir os 2 métodos selecionados são apresentados com maior detalhamento, enquanto os demais são descritos brevemente.

2.5.1 Normalização Quantílica

O objetivo da normalização quantílica, do inglês *Quantile*, é fazer com que a distribuição das intensidades dos *probes*, metilada e não metilada, seja a mesma para cada matriz de chip. Além dos sinais metilados e não metilados, também é realizada uma estratificação por *probes* do tipo I e tipo II (HANSEN MARTIN ARYEE, 2020).

A normalização quantílica, desenvolvida por BOLSTAD et al., 2003, é motivada pela ideia de que um gráfico Q-Q mostra que a distribuição de dois vetores de dados é a mesma se o gráfico for uma linha reta diagonal, e não a mesma se o gráfico não apresentar uma linha diagonal.

Esse conceito é estendido para n dimensões, ou seja, o número de vetores (do chip), um para cada indivíduo. Se todos os n vetores tiverem a mesma distribuição, os quantis nas n dimensões serão posicionados em uma linha reta ao longo da linha dada pelo vetor unitário $(\frac{1}{\sqrt{n}}, ..., \frac{1}{\sqrt{n}})$. Desse modo, um conjunto de dados pode ter a mesma distribuição do quantil médio se os pontos dos quantis n dimensionais forem projetados na diagonal.

Definindo $\boldsymbol{q}_k = (q_{k1}, ..., q_{kn})$ para $k = 1, ..., p$ como o vetor de k-ésimos quantis para os n indivíduos; $\boldsymbol{q}_k = (q_{k1}, ..., q_{kn})$ e $\boldsymbol{d} = (\frac{1}{\sqrt{n}}, ..., \frac{1}{\sqrt{n}})$ como a diagonal unitária.

Para transformar os quantis de modo que todos eles fiquem ao longo da diagonal, considera-se a projeção de $\boldsymbol{q}$ em $\boldsymbol{d}$.

$$proj_{\boldsymbol{d}}\boldsymbol{q}_k = (\frac{1}{n}\sum_{j=1}^{n} q_{kj}, ..., \frac{1}{n}\sum_{j=1}^{n} q_{kj}). \tag{2.5.1}$$

Com isso cada vetor terá a mesma distribuição tomando-se o quantil médio e o substituindo como o valor da observação no conjunto de dados original. O seguinte algoritmo, ilustrado na Figura 4, pode ser utilizado para normalizar um conjunto de vetores de dados para que cada amostra de tecido possua a mesma distribuição:

1. Dados n vetores de chip de tamanho p, em que p é o número de *probes*, forma-se uma matriz X de dimensão p x n onde cada vetor de chip é uma coluna;
2. Ordena-se cada coluna de X para gerar o $X_{ordenado}$;
3. Toma-se as médias das linhas de $X_{ordenado}$ e atribui-se a média para cada elemento na linha para obter $X'_{ordenado}$;
4. O $X_{normalizado}$ é obtido organizando cada coluna de $X'_{ordenado}$ para ter a mesma ordenação do X original.

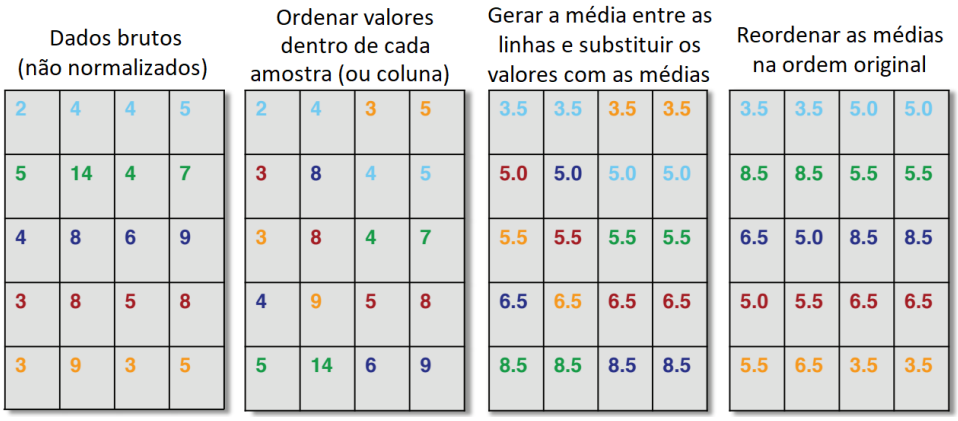

Figura 4: Representação de etapas da normalização quantílica. As linhas representam os *probes* e as colunas representam as amostras, que neste caso são as de tecido.

Fonte: Adaptado de Hicks; Irizarry (2014).

Ressalta-se que este método força os valores dos quantis a serem iguais, o que representa uma limitação do método, principalmente quando há real diferença biológica nas distribuições entre grupos ou valores discrepantes. Nessas situações, a normalização minimizaria as diferenças entre grupos. Por essa razão, tal método de normalização não é recomendado em casos em que diferenças globais podem ser observadas, como por exemplo em estudos com delineamento do tipo caso-controle (HANSEN MARTIN ARYEE, 2020; TOULEIMAT; TOST, 2012; BOLSTAD et al., 2003; HICKS; IRIZARRY, 2014).

2.5.2 Normalização Funnorm

A chamada normalização funcional, do inglês *Functional Normalization* (`Funnorm`), estende a ideia da normalização quantílica ajustando, para cada covariável conhecida, a variação indesejada. A normalização `Funnorm` controla a variabilidade indesejada com

base em "*probes*-controle" presentes na matriz de chip. Seu algoritmo não envolve suposições sobre distribuições, o que torna a normalização aplicável em casos em que mudanças globais são esperadas, como em comparações entre um grupo controle e um grupo de pessoas com câncer, por exemplo (FORTIN; HANSEN, 2015).

O modelo apresentado a seguir é geral, não sendo específico para dados de metilação. Considera-se $\mathbf{Y}_1, ..., \mathbf{Y}_n$ vetores de alta dimensão, cada um associado a um conjunto de covariáveis escalares $Z_{i,j}$ com $i = 1, ...n$ amostras indexadas e $j = 1, ..., m$ covariáveis indexadas. Idealmente, essas covariáveis estão relacionadas com variações indesejadas não associadas à variações biológicas. A normalização `Funnorm` busca remover tal influência. Para cada vetor $\mathbf{Y}_i$, a função quantílica empírica é formada para a sua distribuição marginal, denotada por q_i^{emp}. Assim, funções quantílicas são definidas no intervalo unitário utilizando a variável $r \in [0, 1]$ para avaliá-los no ponto, como $q_i^{emp}(r)$. O modelo na forma pontual se dá como a seguir:

$$q_i^{emp}(r) = \alpha(r) + \sum_{j=1}^{m} Z_{i,j}\theta_j(r) + \epsilon_i(r), \tag{2.5.2}$$

que possui a forma funcional:

$$q_i^{emp} = \alpha + \sum_{j=1}^{m} Z_{i,j}\theta_j + \epsilon_i, \tag{2.5.3}$$

em que

- α é a média das funções de quantis em todas as amostras;
- θ_j são funções de coeficientes associadas às covariáveis;
- ϵ_i são funções de erro, consideradas independentes e centralizadas em torno de 0.

Neste modelo, o termo $\sum_{j=1}^{m} Z_{i,j}\theta_j$ representa a variação nas funções quantílicas explicadas pelas covariáveis. Ao se especificar covariáveis conhecidas, que medem a variação indesejada e que não estão associadas a um sinal biológico, a normalização `Funnorm` remove a variação indesejada reduzindo o valor do erro aleatório (o último termo).

As estimativas de $\hat{\theta}_j$ para $j = 1, ..., m$ são obtidas através de interpolação linear. Com isso formam-se os quantis normalizados funcionais por

$$q_i^{Funnorm}(r) = q_i^{emp}(r) - \sum_{j=1}^{m} Z_{i,j}\hat{\theta}_j(r). \tag{2.5.4}$$

Então, $\mathbf{Y}_i$ são transformados para a quantidade funcional normalizada $\tilde{\mathbf{Y}}_i$ através da equação

$$\tilde{\mathbf{Y}}_i = q_i^{Funnorm}(r)((q_i^{emp}(r))^{-1}(\mathbf{Y}_i)). \tag{2.5.5}$$

Isso garante que a distribuição marginal de $\tilde{\mathbf{Y}}_i$ tenha $q_i^{Funnorm}$ como função quantílica.

Para matrizes de chip do tipo EPIC, a normalização `Funnorm` é aplicada nas intensidades metiladas e não metiladas separadamente. Como é esperado que a relação entre os valores de metilação e os *probes*-controle se diferenciem entre *probes* do tipo I e do tipo II, a normalização também é aplicada separadamente por tipo de *probe* para a obtenção de distribuições quantílicas mais representativas. Isso resulta em quatro aplicações separadas da normalização `Funnorm`, usando exatamente a mesma matriz de covariáveis.

Para definir as covariáveis são utilizados os dois primeiros componentes em uma análise de componentes principais. A definição do número de componentes ($m = 2$) foi obtida de forma empírica pelos autores da normalização.

A normalização `Funnorm` foi proposta a remover apenas a variação nas distribuições marginais dos dois canais de metilação associados aos *probes*-controle. Desse modo, qualquer diferença de metilação biológica global (e real) entre as amostras seria preservada (HANSEN MARTIN ARYEE, 2020; FORTIN et al., 2014).

2.5.3 Normalização Relic

A normalização por "regressão no logaritmo de controle interno", do inglês *Regression on Logarithm of Internal Control Probes* (`Relic`), se propõe a corrigir o viés de coloração (*dye-bias*) em toda a matriz de chip, utilizando a intensidade dos valores de *probes*-controle internos emparelhados, que monitoram os dois canais de cores (verde e vermelho). Esse viés ocorre pois os canais de cores funcionam de forma diferente, com valores de intensidade geralmente mais altos no canal vermelho. A não realização deste ajuste pode causar um impacto direto no cálculo dos `valores-Beta`.

O chip EPIC possui 85 pares de *probes*-controle de normalização interna. Em cada par, os dois *probes* foram projetados para atingir a mesma região de DNA: um *probe* irá incorporar uma base A ou T (canal vermelho), e o outro *probe* irá incorporar uma base G ou C (canal verde). Desse modo, o monitoramento do desempenho da matriz de chip em diferentes canais de cores é realizado para a correção do *dye-bias*. Se não houvesse *dye-bias*, os valores de intensidade dos dois *probes* de cada par seriam os mesmos, com uma razão próxima a 1.

A normalização `Relic` envolve inicialmente uma regressão nos logaritmos dos valores de intensidade dos *probes*-controle para derivar uma relação quantitativa entre os canais vermelhos e verdes e, em seguida, utiliza-se tal relação para corrigir o *dye-bias* nos valores de intensidade para toda a matriz de chip. Especificamente, para cada amostra de tecido a normalização `Relic` ajusta todos os valores de intensidade do canal verde como:

$$I_{i,adj} = e^{\hat{\beta}_1 log(I_i)+\hat{\beta}_0}, \tag{2.5.6}$$

em que

- i é a localização (*index*) do *probe*;
- I_i indica o valor da intensidade do *probe* no canal verde;
- $I_{i,adj}$ representa o valor ajustado da intensidade;
- $\hat{\beta}_0$ e $\hat{\beta}_1$ são os coeficientes de regressão linear das intensidades (em escala logaritmica) entre os *probes* controle, emparelhados para a mesma amostra de tecido [1].

Os valores de intensidade do canal vermelho permanecem inalterados. Ao se originar a relação entre os canais vermelho e verde usando valores de intensidade transformados por logarítmo de *probes* controle, são observados apenas valores não negativos após o ajuste (XU et al., 2017).

2.5.4 Normalização Illumina

A normalização denominada `Illumina` foi desenvolvida pela empresa homônima e implementada no *Software* `Genome Studio`. A implementação da normalização de controle do pacote `minfi` é equivalente à disponível no `Genome Studio`. Porém, tal método exige a seleção de uma matriz de chip de referência e não está claro como o *software* `Genome Studio` seleciona a matriz de chip de referência, mas é possível a especificação manual desse parâmetro. A correção de fundo (*background*) do pacote `minfi` é aproximadamente igual ao do *Software* `Genome Studio` (HANSEN MARTIN ARYEE, 2020).

2.5.5 Normalização NOOB

O método de normalização "normal-exponencial fora de banda", do inglês *Normal-exponential out-of-band* (`NOOB`), é uma normalização que utiliza um método de correção de *background* que controla o viés de coloração ou tingimento (*dye-bias*), caracterizado

[1] Os logaritmos dos valores de intensidade do canal verde, ou seja, em G e C, são modelados como variáveis independentes

por uma diferença de intensidade entre amostras de tecido marcadas com corantes diferentes. O viés de coloração é utilizado no lugar da expressão do gene (HANSEN MARTIN ARYEE, 2020; JR et al., 2013).

2.5.6 Normalização SWAN

O método de normalização quantílica por subgrupo dentro da matriz do chip, do inglês *Subset-quantile Within Array Normalization* (`SWAN`) é um método de normalização em que *probes* do tipo I e tipo II são normalizados juntos em uma única matriz de chip.

O algoritmo combina as distribuições dos valores beta dos *probes* de tipo I e tipo II, em que a distribuição quantílica média é determinada usando um subconjunto de *probes* definidos como biologicamente semelhantes com base no conteúdo CpG (HANSEN MARTIN ARYEE, 2020; MAKSIMOVIC; GORDON; OSHLACK, 2012).

2.6 Identificação de Posições e Regiões Diferencialmente Metiladas

Posições diferencialmente metiladas, do inglês *Differentially Methylated Positions* (DMP), são posições CpGs em que a metilação está associada com um fenótipo contínuo ou categorizado. Para se identificar as DMPs, testa-se a associação entre o nível de metilação e a resposta fenotípica para cada posição (representada por *probes*) do genoma. Fenótipos contínuos são testados por regressão linear. O teste F pode ser utilizado quando o fenótipo de interesse é categorizado, como é o caso neste trabalho. O teste pode ser realizado com base em `valores-Beta` ou `valores-M` (HANSEN et al., 2014).

O modelo ajustado a partir de mínimos quadrados para cada *probe* é expresso por:

$$Y_{ij} = \beta_0(l_j) + \beta_1(l_j)X_j + \beta_2(l_j)X_j + \epsilon_{ij}, \tag{2.6.1}$$

em que:

- Y_{ij} assume `valor-Beta` ou `valor-M` para o CpG no j-ésimo *probe* do i-ésimo indivíduo;
- i representa o i-ésimo indivíduo;
- l_j indica a localização no genoma do j-ésimo *probe*;
- X_j representa os níveis do fenótipo;
- $\beta_0(l)$ é o intercepto;

- $\beta_j(l)$ ($\beta_1(l)$ ou $\beta_2(l)$) representa o parâmetro de interesse, que é o efeito do fenótipo X_j em Y_{ij} e
- ϵ_i indica o termo do erro, que inclui a variabilidade associada com erros de medição e a variabilidade biológica natural.

Para cada *probe*, o modelo ajustado fornece os coeficientes e seus erros-padrões estimados e outros valores necessários para o cálculo do teste F, utilizado para verificar se a posição pode ser considerada como diferencialmente metilada. Quando o estudo envolve amostras inferiores a 10 unidades por grupo um procedimento de redução de variância (*variance shrinkage*) é recomendando. O método implementado no pacote `minfi` envolve as médias das distribuições *a posteriori* de Bayes, descrito em MCCARTHY; SMYTH, 2009. Tal método assume uma distribuição *a priori* para a variância *a posteriori*, de modo que a variância reduzida pode ser expressa por:

$$\tilde{\sigma}_j^2 = \frac{\sigma_j^2 d_j + \sigma_0^2 d_0}{d_j + d_0}, \tag{2.6.2}$$

em que:

- σ_j^2 representa a variância observada do j-ésimo *probe*;
- σ_0^2 indica a variância esperada (da população dos genes) e
- d são os graus de liberdade dos resíduos (observados e *a priori*) para o modelo linear do *probe* j.

Após o procedimento de redução da variância, o teste F é calculado da seguinte forma:

$$F = \left(\frac{\hat{\beta}_j(l)}{\tilde{\sigma}_j \sqrt{v_g}} \right)^2, \tag{2.6.3}$$

em que $v_g = c^T V_g c$; c é um vetor de constantes. Por exemplo, na comparação de dois grupos, pode-se obter $c^T = (0, 1)$ para escolher o coeficiente relacionado à diferença entre os dois grupos. V_g é uma matriz definida não positiva e que não depende de σ_j^2.

Por fim é realizada uma correção para testes múltiplos, que transforma p-valores em q-valores. Os q-valores são obtidos através de uma estimação do FDR (*False Discovery Rate*) da seguinte maneira (STOREY; TIBSHIRANI, 2003):

1. Os p-valores são ordenados em ordem crescente $p_{(1)}, ..., p_{(m)}$.
2. Para um intervalo de λ, como $\lambda = 0, 0.05, 0.10, ..., 0.95$, calcula-se

$$\hat{\pi}_0(\lambda) = \frac{\sum_{j=1}^{m} I_{\{p_j > \lambda\}}}{m(1-\lambda)}, \tag{2.6.4}$$

em que m é o número de p-valores e $j = 1, ..., m$. I é uma variável indicadora. Se $\{p_j > \lambda\}$, então $I_{\{p_j > \lambda\}}$ é igual a 1, e 0 caso contrário.

3. Seja $\hat{f}$ o *spline* cúbico natural (um tipo de interpolação), com 3 graus de liberdade de $\hat{\pi}_0(\lambda)$ em λ.
4. A estimativa de $\hat{\pi}_0(\lambda)$ será

$$\hat{\pi}_0 = \hat{f}(1). \tag{2.6.5}$$

5. Calcula-se

$$\hat{q}(p_{(m)}) = \min_{t \geq p_{(m)}} \frac{\hat{\pi}_0 mt}{\sum_{j=1}^{m} I_{\{p_j \leq \lambda\}}}. \tag{2.6.6}$$

em que t é um valor entre $0 < t \leq 1$. I também é uma variável indicadora. Se $\{p_j \leq \lambda\}$, então $I_{\{p_j \leq \lambda\}}$ é igual a 1, e 0 caso contrário.

6. Para $i = m-1, m-2, ..., 1$, calcula-se

$$\hat{q}(p_{(i)}) = \min_{t \geq p_{(i)}} \frac{\hat{\pi}_0 mt}{\sum_{j=1}^{m} I_{\{p_j \leq \lambda\}}} = \min\left(\frac{\hat{\pi}_0 m p_{(i)}}{i}, \hat{q}(p_{(i+j)})\right). \tag{2.6.7}$$

Os q-valores serão então dados por $\hat{q}(p_{(i)})$. Quando o q-valor é significativo, ou seja, menor que o nível de significância estabelecido, o *probe* é considerado uma posição diferencialmente metilada.

A busca e identificação de Regiões Diferencialmente Metiladas, do inglês *Differentially Mehylated Regions* (DMR), é realizada por meio do comando `bumphunter`, que envolve a busca por "picos" (*bumps*). A identificação de DMRs envolve a busca por regiões do genoma que sejam diferencialmente metiladas entre apenas dois grupos (FORTIN; HANSEN, 2015). No presente estudo comparou-se os grupos câncer e controle saudável.

Para se identificar as DMRs, inicialmente é necessário encontrar *clusters*, que são grupos de *probes* de modo que duas localizações consecutivas de *probes* no *cluster* não sejam separadas por mais do que alguma distância. No caso de *microarrays* os *clusters* são fornecidos pelo fabricante do chip (HANSEN et al., 2014).

O modelo estatístico para calcular os coeficientes das DRMs é similar ao modelo da equação 2.6.1 para identificar as DMPs:

$$Y_{ij} = \beta_0(l_j) + \beta_1(l_j)X_j + \epsilon_{ij}, \tag{2.6.8}$$

em que

- Y_{ij} assume `valor-Beta` ou `valor-M` para o CpG no j-ésimo *probe* do i-ésimo indivíduo;
- i representa o i-ésimo indivíduo;
- l_j indica localização no genoma do j-ésimo *probe*;
- X_j representa os fenótipos (1 para caso e 0 para controle);
- $\beta_0(l)$ é o intercepto;
- $\beta_1(l)$ é o parâmetro de interesse, efeito do fenótipo X_j em Y_{ij};
- ϵ_i indica o termo do erro, que inclui a variabilidade associada com erros de medição e a variabilidade biológica natural.

Assume-se que $\beta_1(l)$ será igual a zero na maior parte do genoma e localizações com $\beta_1(l) \neq 0$ serão considerados como picos (HANSEN; IRIZARRY, 2013).

Calcula-se então a estatística de teste t para cada *probe*, dada por:

$$t = \frac{\hat{\beta}_1(l)}{\sqrt{v\hat{a}r(\hat{\beta}_1(l))}}. \tag{2.6.9}$$

A busca por DMRs candidatas ocorre dentro de cada *cluster*, em que mais de uma DMR pode ser identificadas.

Para testar a significância da região, o algoritmo realiza permutações através de procedimentos de *bootstrap*, com um valor de corte (*cutoff*) que corresponde ao percentual mínimo de diferença nos `valores-Beta` ou `valores-M` entre grupos. Neste estudo, adotou-se o valor de corte de 0.2, ou mínimo de 20% de diferença nos valores entre casos e controles.

Cada uma das reamostras de *bootstraps* irá produzir um "perfil nulo" estimado, para os quais pode-se definir as "regiões candidatas nulas". Para cada região candidata observada, determina-se quantas regiões nulas são "mais extremas" (mais longas e com maior valor médio) (IRIZARRY et al., 2013).

O p-valor é a porcentagem de regiões candidatas obtidas por `bootstraps` que são tão ou mais extremas que a região observada. Deve ser interpretado com cuidado,

devido às proporções teóricas que não são bem compreendidas. Também é realizada uma avaliação de incerteza mais conservadora com base no FWER (*family-wise error rate*), que calcula, para cada região observada, a proporção de reamostras de *bootstraps* pelo menos uma região tão extrema quanto a região observada. (JAFFE et al., 2012).

Tais procedimentos de `bootstrap` envolvem bastante tempo computacional, pois recomenda-se utilizar pelo menos 1000 reamostragens em tal contexto. Uma das formas de se reduzir o tempo computacional seria aumentar o valor de corte, o que diminuiria o número de *bumps* a serem identificados (FORTIN; HANSEN, 2015). As análises para a identificação de DMRs pode também ser realizada por cromossomo, estratégia adotada neste trabalho.

3 Resultados

Dados de metilação em amostras de tecido de 12 indivíduos, em 3 grupos de 4 indivíduos com câncer colorretal, com pólipos benignos ou controles saudáveis, foram coletados e analisados. As etapas de análise dos dados envolveram procedimentos de controle de qualidade, aplicação de métodos de normalização e identificação de posições e regiões diferencialmente metiladas. Os resultados obtidos são apresentados a seguir.

3.1 Procedimentos de Controle de Qualidade

A Figura 5 apresenta as médias de p-valores de detecção obtidos para cada indivíduo, considerando-se todos os *probes*. A média dos p-valores de nenhum indivíduo sequer chegou próximo do valor de corte 0,01 (no gráfico equivale a 100 = 0,01 x 10.000). Sendo assim, todos os indivíduos foram mantidos no estudo.

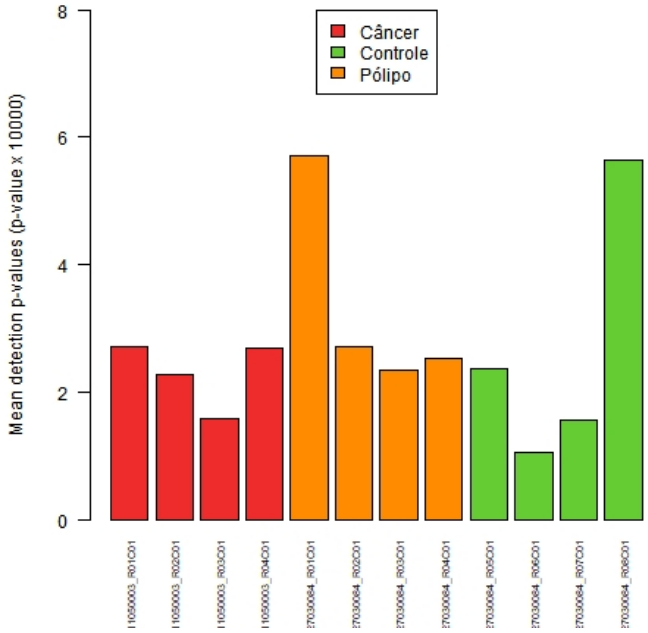

Figura 5: Média dos p-valores de detecção (p-valor x 10.000) para cada indivíduo.

A densidade do `valor-Beta` para cada indivíduo é apresentada na Figura 6. `Valores-Beta` próximos de 1 indicam a presença de genes metilados, e valores próximos de 0 indicam a presença de genes não metilados. Portanto, como seria esperado, observa-se um pico próximo de cada extremo da distribuição dos valores. Para os três grupos de indivíduos, o comportamento da distribuição é semelhante. Porém, para indivíduos do

grupo controle, os picos próximos de 0 são ligeiramente inferiores, assim como os próximos de 1, superiores aos valores de indivíduos dos outros grupos.

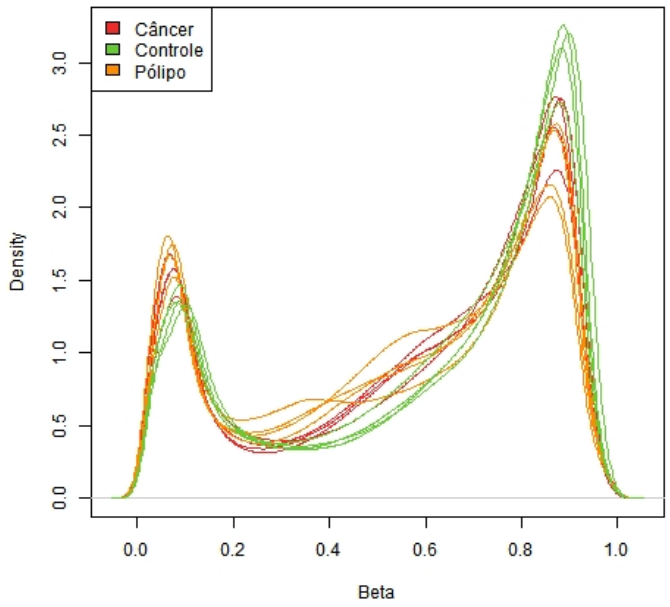

Figura 6: Densidade do `valor-Beta` para cada indivíduo.

A Tabela 2 mostra o número de *probes* excluídos por cada critério de controle de qualidade (CQ). Destaca-se que tais exclusões foram feitas de forma sequencial. O número total de *probes* excluídos representa 10,38% dos 866.836 *probes* incluídos no chip.

Tabela 2: Número de *probes* excluídos em cada processo do Controle de Qualidade.

Etapa do CQ	**Probes removidos**
P-valor de detecção	5.218
SNPs	29.206
Reação Cruzada	40.133
Cromossomos X e Y	17.476
Total	92.033

3.2 Análise exploratória

A Figura 7 exibe a distribuição dos valores brutos (não normalizados) `Beta`, `M` e `CN` antes e depois dos processos de CQ. Em geral, não há muita diferença nas distribuições entre grupos para os `valores-Beta` e `valores-M`, mas o grupo controle apresentou

medianas ligeiramente menores.

A distribuição do `valor-M` varia entre -10 a 10 antes ou após procedimentos de CQ. Já o valor CN se distribui de 4 a 16 antes e, de 10 e 16 após tais procedimentos. O valor-CN é que mais sofreu alteração na distribuição dos valores após o CQ.

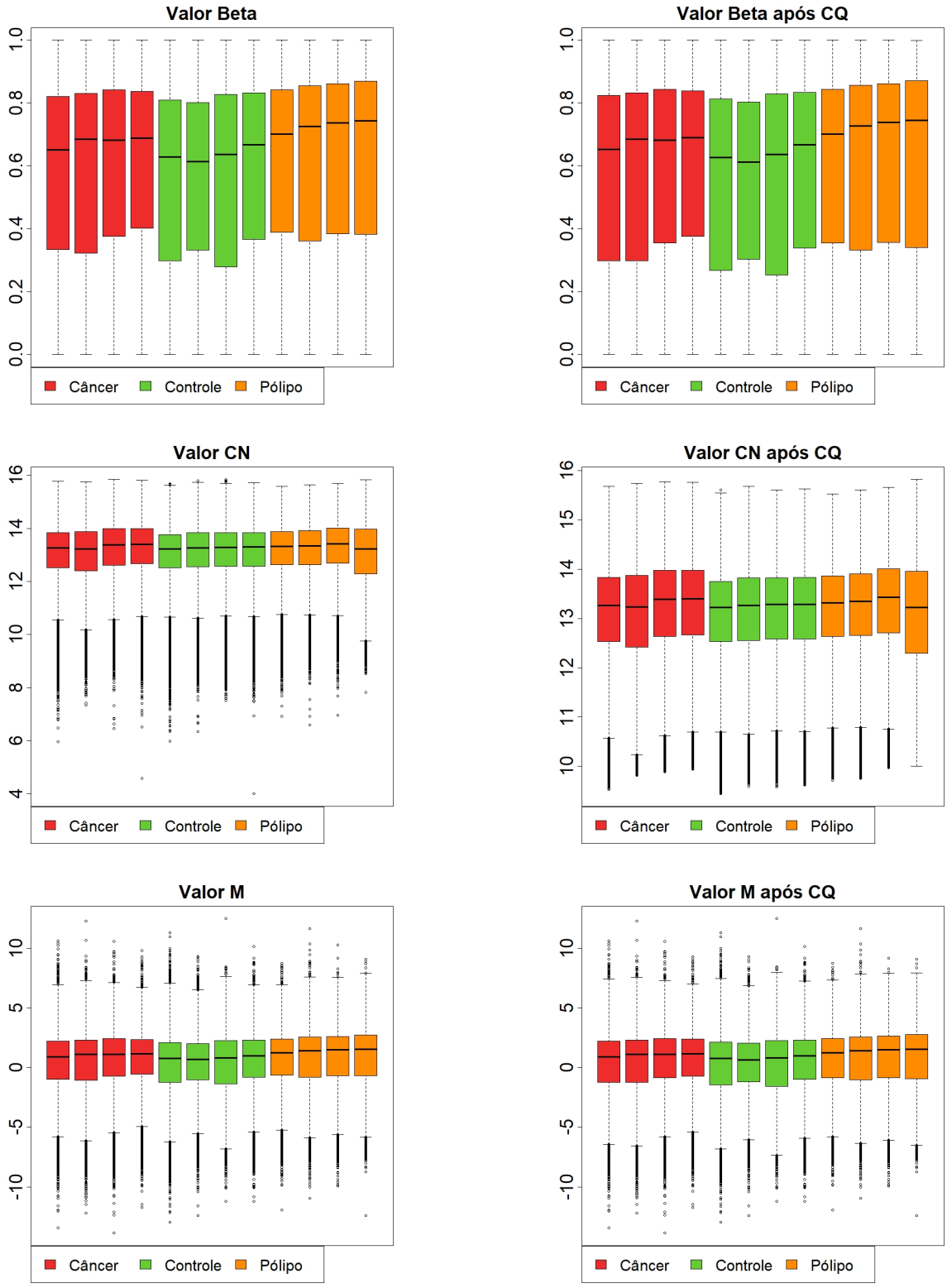

Figura 7: Boxplots dos valores brutos de Beta, M e CN antes e após procedimentos de controle de qualidade (CQ).

3.3 Normalizações

Por simplicidade, `valores-Beta` e `valores-M` normalizados pelos métodos `Funnorm` e `Relic` passarão a ser tratados como vBeta.Funnorm, vM.Funnorm, vBeta.Relic e vM.Relic, repectivamente. Observa-se que as distribuições de vBeta.Funnorm e vBeta.Relic são bastante semelhantes (Figura 8) e sem a presença de valores discrepantes. Novamente as medianas dos indivíduos do grupo controle parecem inferiores às outras.

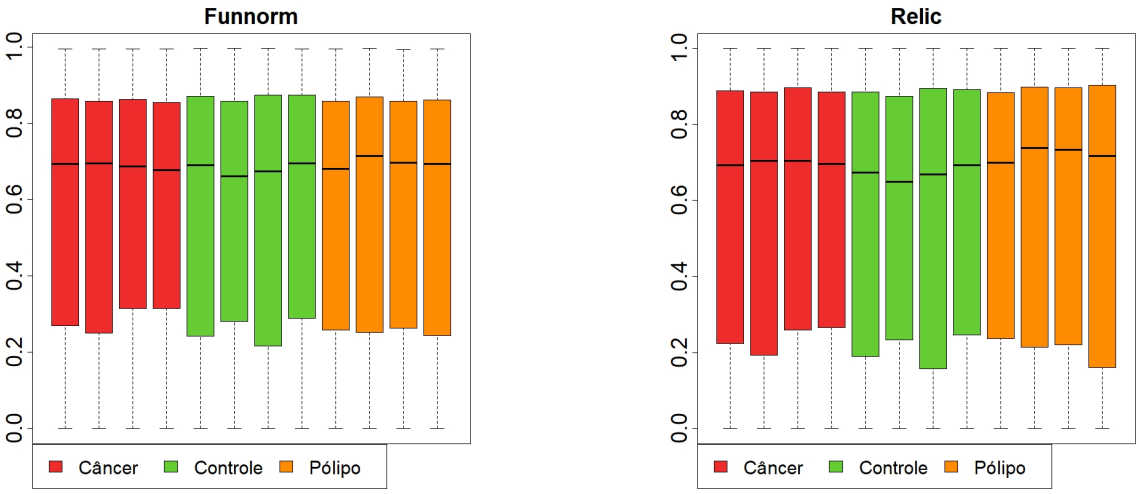

Figura 8: Boxplot do `valor-Beta` para as normalizações `Funnorm` e `Relic`.

As distribuições do `valor-M` (Figura 9) apresentam maior disparidade entre os métodos de normalização `Funnorm` e `Relic`. A primeira não possui *outliers* superiores e varia entre -15 e 5, enquanto a segunda, entre -20 e 20.

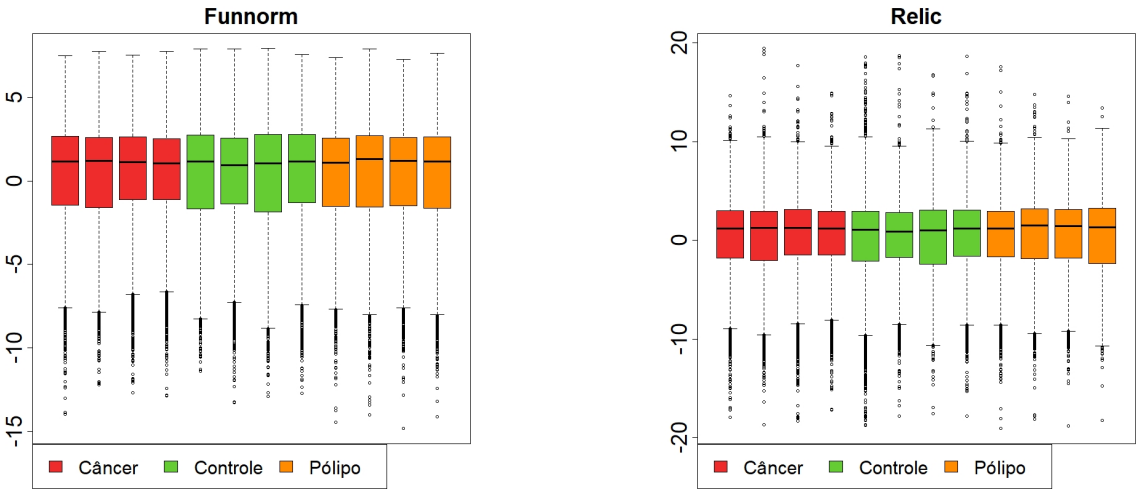

Figura 9: Boxplot do `valor-M` para as normalizações `Funnorm` e `Relic`.

Para o valor CN parece existir uma ligeira diferença nos valores resultantes das 2 normalizações. Ambas apresentam pontos discrepantes inferiores. A amplitude do valor CN para a normalização `Relic` é maior que a da normalização `Funnorm`. Além disso, para a normalização `Relic`, a mediana do grupo controle parece estar mais equiparada aos outros grupos quando comparada à `Funnorm`.

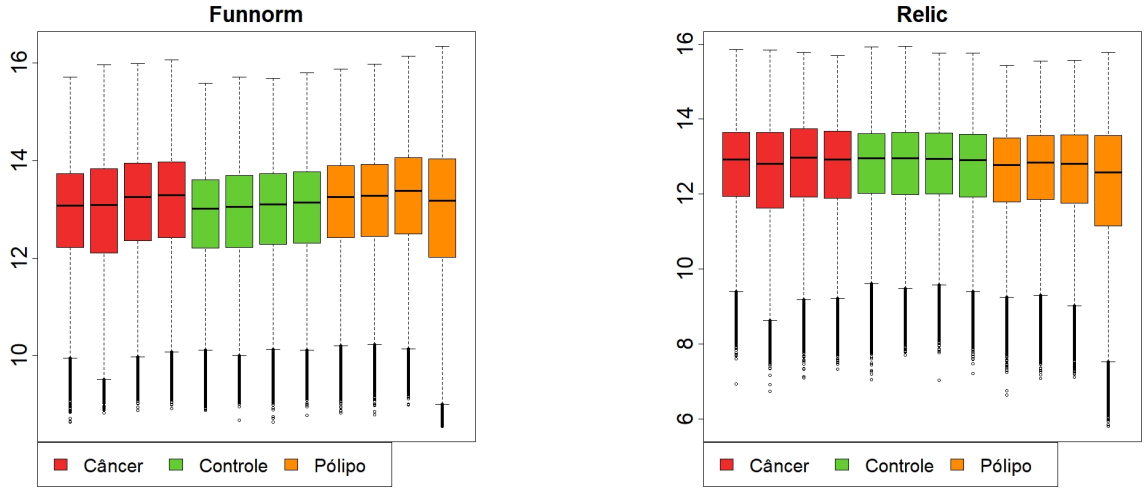

Figura 10: Boxplot do valor CN para as normalizações `Funnorm` e `Relic`.

As distribuições de valores Beta, M e CN para as demais normalizações são apresentadas nas Figuras 21, 22 e 23 do Apêndice.

3.4 Posições Diferencialmente Metiladas

A Tabela 3 apresenta o número de DMPs para os valores Beta e M ajustados pelas normalizações escolhidas. Observa-se que os números de DMPs identificadas para valores normalizados pelo método `Relic` foram muito superiores aos obtidos pelo método `Funnorm`. Ao todo foram identificadas 3.186 DMPs.

Tabela 3: Número de posições diferencialmente metiladas para os valores `Beta` e `M` normalizados pelos métodos `Funnorm` e `Relic`.

Normalização	**Valor-Beta**	**Valor-M**
Funnorm	149	12
Relic	1.136	1.889

As Figuras 11 e 12 apresentam diagramas de Venn[2] com a distribuição do número de DMPs identificadas com valores `Beta` e `M`, respectivamente (normalizados pelos métodos `Funnorm` e `Relic`). A maior parte (n = 113) das DMPs identificadas com base em vBeta.Funnorm foi também identificada com vBeta.Relic. Nenhuma DMP foi identificada exclusivamente com base em vM.Funnorm.

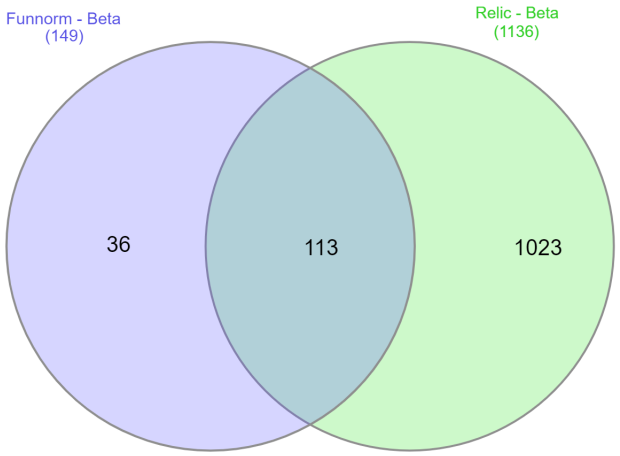

Figura 11: DMPs identificadas com as normalizações `Funnorm` e `Relic` do `valor-Beta`.

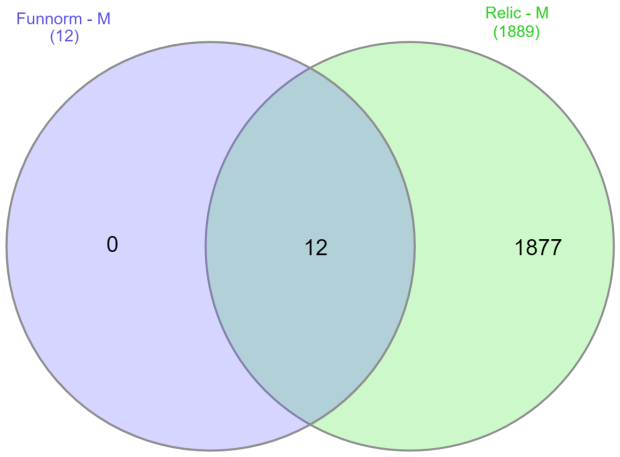

Figura 12: DMPs identificadas com as normalizações `Funnorm` e `Relic` do `valor-M`.

[2]Todos os diagramas de Venn foram elaborados no site <http://www.interactivenn.net/>

O Quadro 2 apresenta um exemplo de saída do *software R* para algumas DMPs da normalização `Funnorm`. Para cada *probe* são disponibilizados o valor do intercepto, o valor da estatística F, o p-valor e o q-valor (corrigido por testes múltiplos). Uma posição é considerada DMP quando o q-valor é inferior a 0,05.

Quadro 2: Exemplo de saída do *software R* de DMPs para a normalização `Funnorm` do `valor-Beta`

Probe	**Intercepto**	**F**	**p-valor**	**q-valor**
cg09103061	0,1812	132,75	$5x10^{-8}$	0,02
cg04043710	0,1180	112,31	$1x10^{-7}$	0,03
cg17499212	0,0313	95,66	$2x10^{-7}$	0,03
cg17029731	-0,0154	89,31	$3x10^{-7}$	0,03
cg05751616	0,0756	85,64	$4x10^{-7}$	0,03
cg06263372	-0,0008	79,85	$6x10^{-7}$	0,03
cg21183638	0,1289	77,74	$7x10^{-7}$	0,03
cg21183638	0,0040	77,05	$7x10^{-7}$	0,03
cg02021127	-0,1928	76,48	$8x10^{-7}$	0,03
cg08164795	0,0877	75,70	$8x10^{-7}$	0,03

3.5 Regiões Diferencialmente Metiladas

A Tabela 4 apresenta o número de DMRs identificadas para as normalizações escolhidas, com valor de corte (*cutoff*) para a diferença percentual mínima de 0,2 e 1000 reamostragens por *bootstraps*. Ao todo foram identificadas 111 DMRs. Observa-se que o número de DMRs encontradas, é inferior ao número de DMPs, com exceção de vM.Funnorm (comparação entre Tabelas 3 e 4). Além disso, o `valor-M` identificou mais DMRs que o `valor-Beta`.

Tabela 4: Número de regiões diferencialmente metiladas para os valores `Beta` e `M` das normalizações `Funnorm` e `Relic`.

Normalização	**Valor-Beta**	**Valor-M**
Funnorm	19	43
Relic	15	34

O Quadro 3 ilustra um exemplo da saída do `R` para as DMRs. A definição de cada coluna se dá da seguinte forma:

- CR: cromossomo;
- Início e fim: começo e fim da região no cromossomo;
- Valor: média do coeficiente estimado na região (geralmente representa a diferença entre dois grupos);
- Área: é o valor absoluto da soma dos coeficientes estimados na região;

- Cluster: ID do cluster;
- I.Início e I.Fim: índice do *probe* na matriz de metilação original onde a região começa e termina;
- L e C.L: número de *probes* na região e no cluster;
- Pvalor e P.Área: p-valores de testes referentes à região e à área, respectivamente;
- FWER e F.Área: FWER de testes referentes à região e à área, respectivamente.

A região considerada uma DMR é a que apresenta FWER inferior a 0,05.

Quadro 3: Exemplo de saída do *software R* de DMRs para a normalização `Funnorm` do `valor-Beta`.

CR	Início	Fim	Valor	Área	Cluster	I.Início	I.Fim	L	C.L	Pvalor	FWER	P.Área	F.Área
1	158150581	158151363	-3.35	20.11	25157	52388	52393	6	6	$3x10^{-6}$	0.03	$1x10^{-4}$	0.59
1	25256369	25259034	-0.90	26.86	7703	16695	16724	30	30	$1x10^{-6}$	0.04	$5x10^{-5}$	0.34
1	186649153	186650479	-2.13	36.24	29599	60657	60673	17	17	$2x10^{-6}$	0.06	$1x10^{-5}$	0.13
1	153234037	153234809	-3.14	15.68	23734	48773	48777	5	5	$8x10^{-6}$	0.06	$4x10^{-4}$	0.78
1	159842863	159842929	3.15	9.45	25386	52819	52821	3	4	$1x10^{-5}$	0.07	$1x10^{-3}$	0.96
1	49242359	49242934	-2.52	25.22	13643	29179	29188	10	10	$1x10^{-5}$	0.13	$7x10^{-5}$	0.39

3.6 Identificação de Genes

Ocorreram casos em que a mesma região (contendo um gene) foi identificada por diferentes métodos de normalização e/ou de valores (`valor-Beta` ou `valor-M`), o que indica boa consistência nos resultados. Além disso, houve casos em que o mesmo gene foi identificado em regiões ligeiramente diferentes de modo que, ao todo foram identificados 58 genes distintos através dos valores `Beta` e `M` normalizados pela `Funnorm` e `Relic`. A Figura 13 mostra que há 6 genes identificados em comum com valores `Beta` e `M` e normalizações. Dezoito genes foram identificados somente por valores normalizados pelo método `Funnorm`.

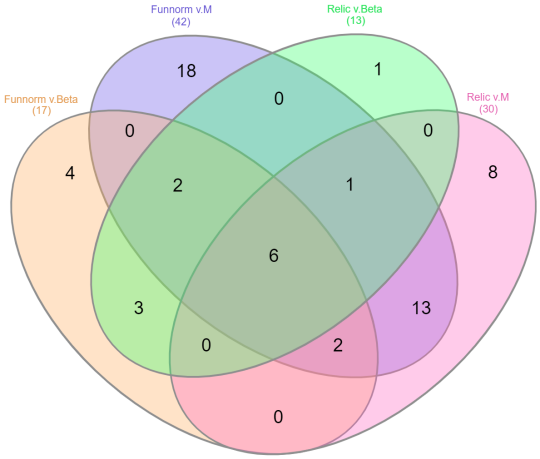

Figura 13: Genes em comum entre as normalizações e valores `Beta` e `M`.

Quando se avalia apenas o `valor-M`, que possibilitou a identificação de mais DMRs nas duas normalizações, verifica-se que há 22 DMRs em comum (Figura 14).

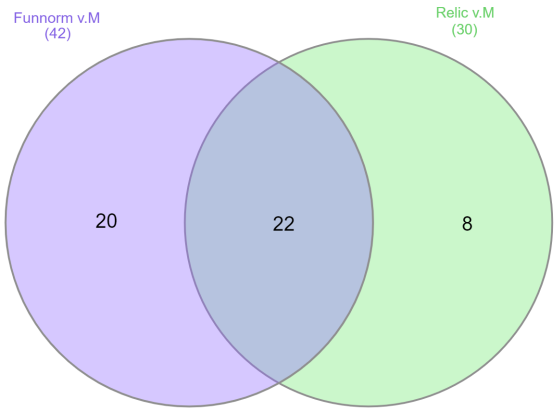

Figura 14: Genes em comum das normalizações `Funnorm` e `Relic` do `valor-M`.

A Tabela 5 apresenta os genes diferencialmente metilados entre os grupos câncer e controle saudável. A informação sobre o gene ser prognóstico de CCR foi obtida no Atlas de Proteína Humana (*The Human Protein Atlas*) [3].

Tabela 5: Lista de genes encontrados na identificação de DMRs.

CR	P. inicial	P. final	Norm.	Valor	Gene	Prog.
1	183441292	183441331	F	M	*SMG7*	Não
1	109419585	109419685	F	M	*GPSM2*	Sim
1	121260690	121261404	F	M	*LOC647121*	Sim
1	186649153	186650479	F	M	*PTGS2*	Não
1	158150581	158151363	F	M	*CD1D*	Não
1	25256369	25259034	R	M	*RUNX3*	Sim
2	128768557	128768557	F	M	*SAP130*	Não
2	154333914	154335766	F/R	M	*RPRM*	Não
2	70994758	70995607	F/R	M	*ADD2*	Não
2	182321783	182322501	F	B	*ITGA4*	Não
2	68546467	68547088	F	B	*CNRIP1*	Não
3	93781653	93781653	F	M	*NSUN3*	Não
4	96468962	96471143	F/R	M	*UNC5C*	Não

Continua

[3] *The Human Protein Atlas*: <https://www.proteinatlas.org/>

Conclusão

CR	P. inicial	P. final	Norm.	Valor	Gene	Prog.
4	176986621	176987577	F	M	*WDR17*	Não
4	81186592	81188471	R	M	*FGF5*	Não
6	10837744	10837744	F	M	*MAK*	Não
6	133561614	133562776	F/F/R	M/B/B	*EYA4*	Não
7	102985355	102985355	F	M	*DNAJC2*	Não
7	93519220	93520566	F/R/F/R	M/B/M/B	*TFPI2*	Não
7	99155595	99157059	F/R	M	*ZNF655*	Não
7	142494204	142495846	F/R/R	M/M/B	*TRBJ2-3*	Não
7	45960806	45963523	F/R	M	*IGFBP3*	Sim
7	27182435	27185732	F	M	*HOXA5*	Não
7	27140797	27144595	R	M	*HOXA2*	Não
8	42751762	42751762	F	M	*RNF170*	Não
8	67344190	67345006	F/R/F	M/B/B	*ADHFE1*	Não
8	72753888	72757004	F	M	*MSC*	Não
8	145105503	145107199	F/R	B	*OPLAH*	Não
9	132382433	132382463	F/R	M	*C9orf50*	Não
9	100614879	100615357	F/R	B	*FOXE1*	Não
9	91606128	91606450	R	B	*C9orf47*	Não
10	7450112	7454759	F/R/F/R	M/B/M/B	*SFMBT2*	Não
11	2160882	2162582	F/R	M	*IGF2*	Não
11	32454718	32457878	F/R	M	*WT1*	Não
11	123300839	123301758	F/R/F/R	M/B/M/B	*AP000783.1*	Não
11	79148183	79152112	F	M	*TENM4*	Não
11	18813466	18813484	F	B	*PTPN5*	Não
11	2889602	2891360	R	M	*KCNQ1DN*	Não
11	110581434	110584091	R	M	*ARHGAP20*	Não
12	95941571	95943131	F/R/F/R	M/B/M/B	*USP44*	Não
12	45268163	45270997	F	M	*NELL2*	Não
13	78492916	78494067	F	M	*EDNRB*	Não
13	37004536	37006265	F/R	M	*CCNA1*	Não
13	28498091	28498384	F/R	B	*PDX1*	Não
13	112759893	112760228	F	B	*LINC00403*	Não
13	103046499	103047287	R	M	*FGF14*	Não
14	102247610	102248073	F/F/R/R	M/B/M/B	*PPP2R5C*	Não
14	70653919	70656116	F/R	M	*SLC8A3*	Não
15	48936953	48937213	F/R	M	*FBN1*	Não

Continua

Conclusão

CR	P. inicial	P. final	Norm.	Valor	Gene	Prog.
15	74420307	74422572	F/R	M	*LOC283731*	Não
15	79382548	79383980	F/R	M	*RASGRF1*	Não
16	70472678	70472993	F	M	*ST3GAL2*	Não
16	58497230	58499074	R	M	*NDRG4*	Não
17	46655561	46656892	F/F/R/R	M/B/M/B	*HOXB4*	Não
17	75368750	75370611	F/R/F	M/M/B	*SEPT9*	Não
19	58739734	58740747	F/R/F	M/M/B	*ZNF544*	Não
20	57427010	57427977	F	M	*GNAS*	Não
20	24449668	24452131	R	M	*SYNDIG1*	Não

Abreviações: CR = Cromossomo; P. = Posição; Norm. = Normalização; F = `Funnorm`; R = `Relic`; B = Beta; Prog. = Prognóstico de CCR; Exp. = Expressão da proteína em CCR; Pend. = Pendente.

A Tabela 8 do Apêndice contém os nomes completos dos genes. Os 4 genes da Tabela 6 são prognósticos de CCR, segundo o atlas, todos identificados com o `valor-M`, e um gene em comum entre as normalizações `Funnorm` e `Relic`.

Tabela 6: Genes prognósticos de CCR.

Cromossomo	Normalização	Valor	Gene
1	Funnorm	M	*GPSM2*
1	Funnorm	M	*LOC647121*
2	Relic	M	*RUNX3*
7	Funnorm/Relic	M	*IGFBP3*

As Figuras 15 a 20 mostram o comportamento de valores `Beta` e `M` em 3 DMRs que incluem respectivamente os genes *LOC647121*, *SEPT9* e *TFPI2*.

Os pontos representam os valores para cada indivíduo dos grupos de câncer (em vermelho) e controle (em azul). As linhas representam as medianas de cada grupo mantendo-se as respectivas cores por grupo. A marcação em verde delimita a DMR. Quando comparados os valores `Beta` e `M`, há pouca diferença no comportamento das observações entre os gráficos, apesar da amplitude entre eles ser distinta.

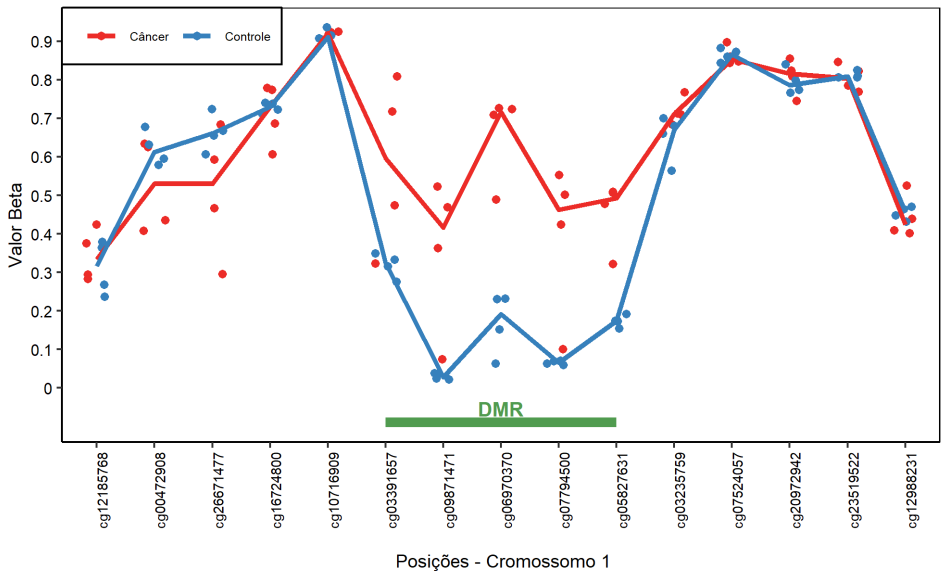

Figura 15: DMR no cromossomo 1 em que o gene *LOC647121* foi identificado com `valor-Beta`.

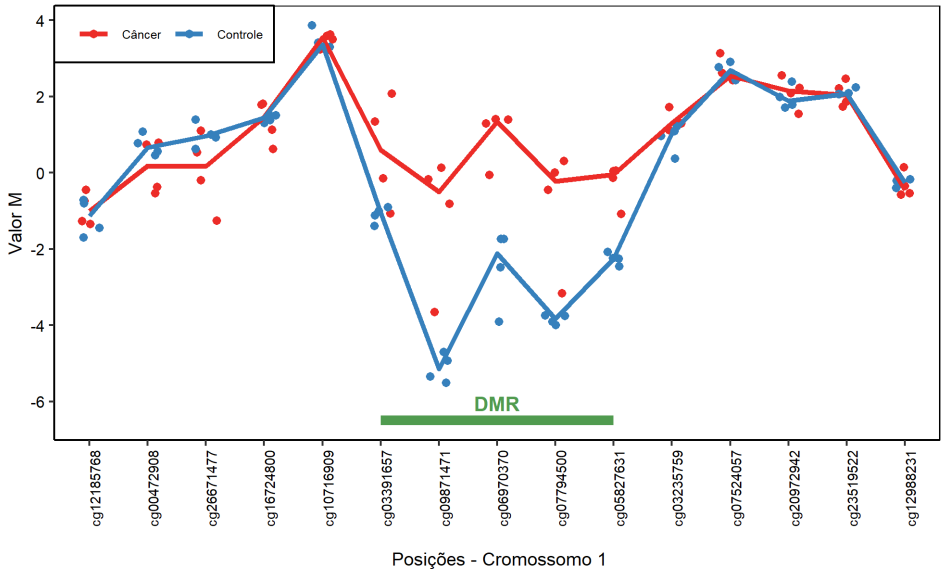

Figura 16: DMR no cromossomo 1 em que o gene *LOC647121* foi identificado com `valor-M`.

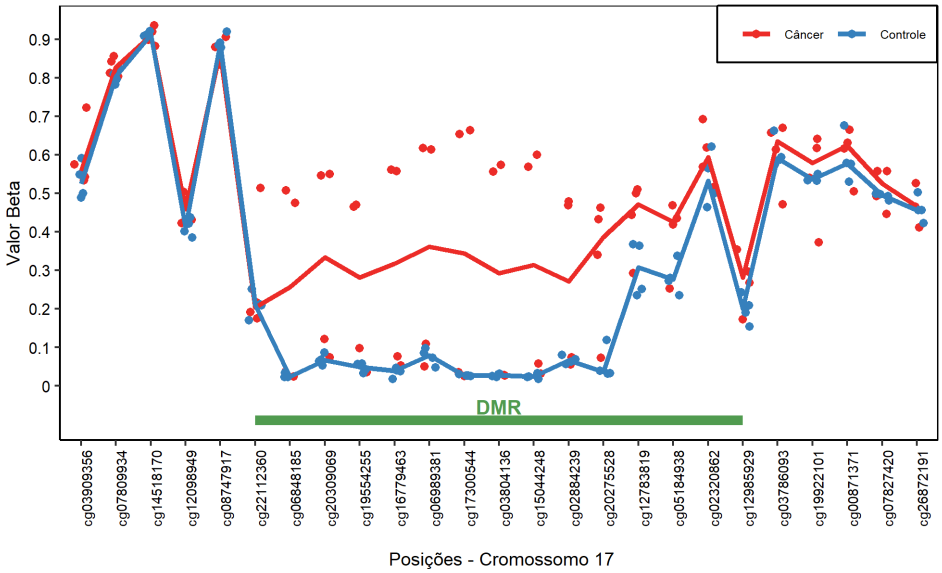

Figura 17: DMR no cromossomo 17 em que o gene *SEPT9* foi identificado com `valor-Beta`.

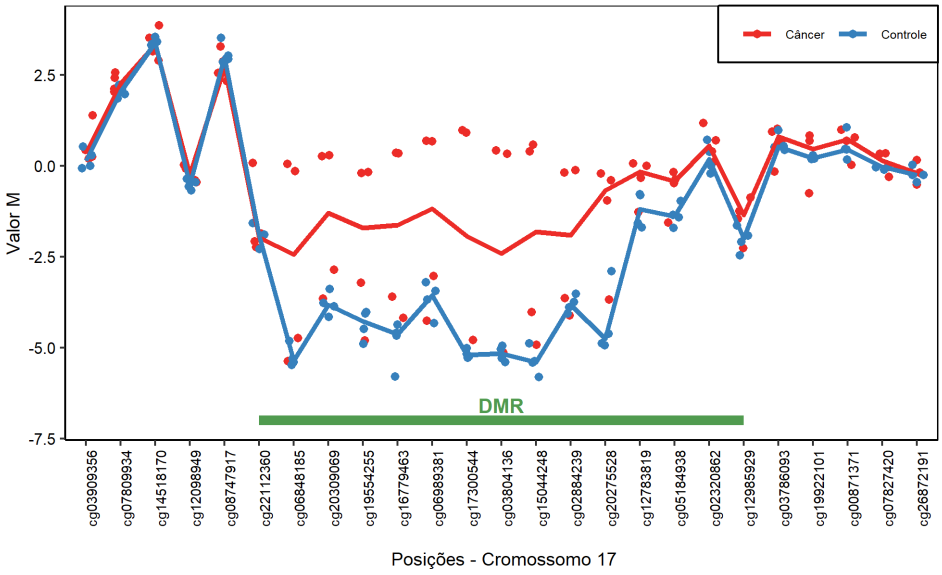

Figura 18: DMR no cromossomo 17 em que o gene *SEPT9* foi identificado com `valor-M`.

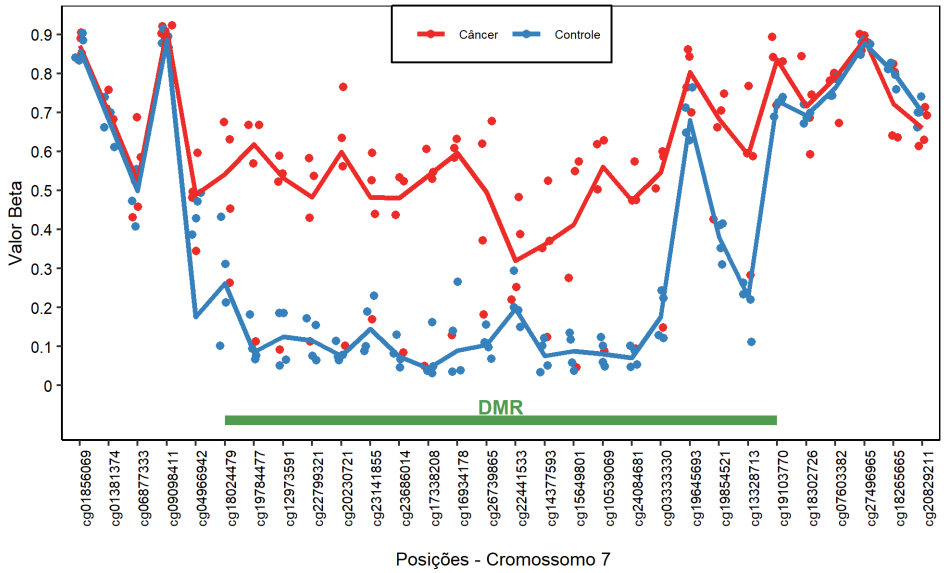

Figura 19: DMR no cromossomo 7 em que o gene *TFPI2* foi identificado com `valor-Beta`.

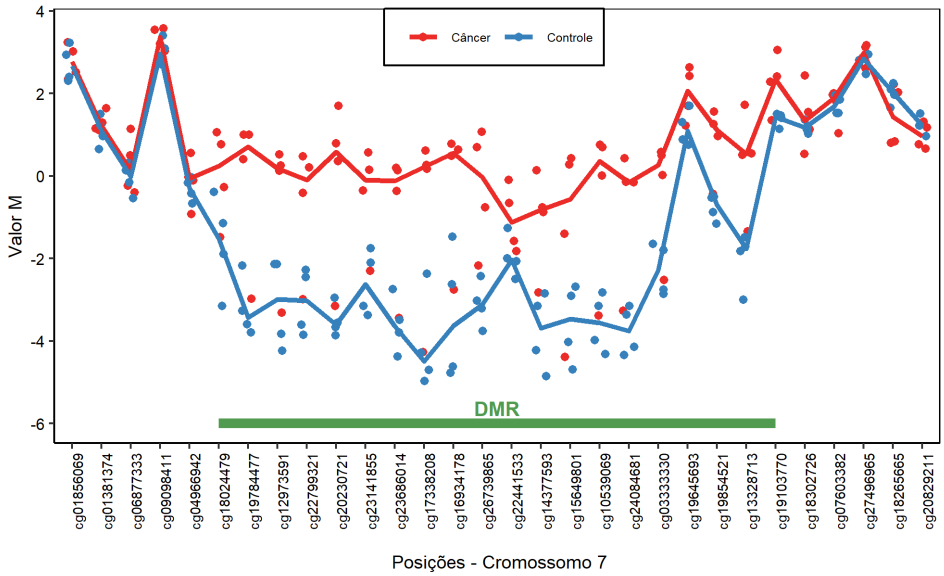

Figura 20: DMR no cromossomo 7 em que o gene *TFPI2* foi identificado com `valor-M`.

4 Discussão e conclusão

Neste trabalho foram empregadas técnicas para analisar dados de metilação em todo o genoma de 12 indivíduos (4 com câncer colorretal, 4 com pólipos e 4 controles saudáveis). Métodos de controle de qualidade da conversão bisulfito e do processo de hibridização nos chips de metilação foram realizados. A exclusão de *probes* com altos p-valores de detecção removeu *probes* que não hibridizaram. *Probes* afetados por marcadores SNP e nos cromossomos sexuais X e Y também foram removidos. Tais processos resultaram em uma diminuição de aproximadamente 10% do total de *probes*. Os valores `Beta, M e CN`, mediram o nível de metilação de cada *probe*, para cada amostra. O `valor-Beta` varia de 0 a 1 e é apropriado para a análise exploratória dos níveis de metilação. Já o `valor-M` apresentou a maior variação de valores e muitos valores discrepantes.

A escolha dos métodos de normalização `Funnorm` e `Relic` foi baseada na literatura por serem adequadas para análises de comparação global, o que não era comportado por algumas das outras normalizações disponíveis. A identificação de posições diferencialmente metiladas, para os 3 grupos de amostras de tecido, se mostrou bem diferente entre as normalizações dos valores `Beta` e `M`. Para a normalização `Funnorm` o número de DMPs foi maior para o `valor-Beta`, contrário ao que aconteceu com a normalização `Relic`. De toda forma, o número de DMPs encontradas com a normalização `Relic` foi muito maior (quase 8 vezes) para o `valor-Beta` quando comparado à normalização `Funnorm`. É possível que haja mais resultados falso positivos ou que o método Relic forneça valores que evidenciem maiores diferenças entre os grupos. Além disso, todas as DMPs identificadas com vM.Funnorm foram também detectadas com vM.Relic. Isso não aconteceu em relação ao `valor-Beta`, mas cerca de 76% das DMPs da normalização `Funnorm` também foram encontradas pela normalização `Relic`.

Devido ao grande número de DMPs identificadas, os genes relacionados à essas posições não foram detectados neste estudo. Mas é válido apontar que nem todas as posições possuem genes. Portanto, o número de genes seria menor que o número de DMPs.

Os números de regiões identificadas como diferencialmente metiladas foram menores que os de DMPs. Para as duas normalizações com `valor-M`, foram identificadas mais regiões que com o `valor-Beta`. No total foram encontradas 111 DMRs, que implicam em 58 genes. A normalização `Funnorm` identificou mais genes, mas na comparação com o `valor-M` foram observados 22 genes em comum.

Quatro dos genes identificados neste trabalho são prognósticos de câncer colorretal de acordo com as pesquisas feitas no Atlas de Proteína Humana. São eles: *GPSM2, LOC647121, RUNX3 e IGFBP3*. Todos foram detectados através do `valor-M`. O gene

LOC647121, também chamado de *EMBP1*, foi identificado como um biomarcador de CCR por LI et al., 2019. Ou seja, é um gene indicativo da presença de CCR no corpo. O estudo de KU et al., 2004 concluiu que a baixa expressão do gene *RUNX3* ocorre devido a hipermetilação de sua ilha CpG em câncer colorretal, o que corrobora para o gene ser diferencialmente metilado em CCR. Alterações no gene *IGFBP3* estão associadas ao risco de desenvolvimento de câncer colorretal, como mostra o estudo de MURPHY et al., 2020.

Um dos genes identificados, o *SEPT9*, é um marcador hipermetilado específico de câncer colorretal (ARELLANO et al., 2020). Esse gene é clinicamente usado como biomarcador na triagem para CCR. Exames de triagem geralmente são realizados para detecção precoce de doenças. O gene *ADD2*, também detectado, é um potencial biomarcador de CCR, como conclui o estudo de WEI et al., 2016.

A metilação do gene *TFPI2* é frequentemente observada em câncer colorretal. Esse gene pode agir como supressor de tumor em CCR e sua metilação pode apresentar um risco potencial de malignidade em CCR (HIBI et al., 2010). O gene *CCNA1* foi considerado como um gene que, quando erroneamente metilado, está relacionado com o desenvolvimento de tumores malignos (YANG et al., 2015). Outro gene importante é o *RPRM*, em que a desregulação na metilação desse gene pode contribuir para a patogênese de CCR (CORVALAN et al., 2012).

O *CD1D* é um exemplo de gene ligado a uma molécula que contribui para a resposta imune antitumoral. Polimorfismos em genes associados à essa molécula acabam modificando sua função (GOLMOGHADDAM et al., 2011). O gene supressor de tumor *USP44* é inativado epigeneticamente em adenomas colorretais (SLOANE et al., 2014).

Muito ainda pode ser explorado com as metodologias de identificação de posições e regiões diferencialmente metiladas. Com esse conjunto de dados, sugere-se comparações entre os grupos: câncer vs. pólipo e controle saudável vs. pólipo, para identificação de DMPs e DMRs. O uso do `valor-CN` como nível de metilação na busca de posições e regiões diferencialmente metiladas também seria interessante. Sugere-se ainda a utilização de análises mais aprofundadas das posições diferencialmente metiladas e o uso de outras metodologias voltadas para análise de dados de metilação a fim de comparar os resultados obtidos neste trabalho. Finalmente, seria de grande importância o desenvolvimento de funções que facilitem o processo de criação de gráficos que mostrem as DMRs, como os apresentados neste trabalho.

Em suma, apesar dos pequenos tamanhos de amostra (4 por grupo), a técnica de identificação de regiões diferencialmente metiladas detectou muitos genes relacionados ao CCR. Devido a isso, o método de busca `bumphunter` se mostrou adequado para identificar DMRs. Além disso, nem todos os genes haviam sido reportados como associados à CCR. Portanto, os genes encontrados neste trabalho são um incentivo para que novas análises

que relacionem o CCR e esses genes possam ser feitas. Por fim, o estudo de metilação de genes em câncer tem se mostrado muito importante na elucidação de componentes moleculares no desenvolvimento e progressão dessas doenças, bem como para avanços em terapêuticas, e deve continuar sendo incentivado e aprofundado.

Referências

ARELLANO, M. L. et al. A first step to a biomarker of curative surgery in colorectal cancer by liquid biopsy of methylated septin 9 gene. *Disease Markers*, Hindawi, v. 2020, 2020.

BIOLOGY, C. D. of. *Copy Number*. 2005. <https://medical-dictionary.thefreedictionary.com/copy+number>.

BIRD, A. Dna methylation patterns and epigenetic memory. *Genes & development*, Cold Spring Harbor Lab, v. 16, n. 1, p. 6–21, 2002.

BOLSTAD, B. M. et al. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, Oxford University Press, v. 19, n. 2, p. 185–193, 2003.

BROOKES, A. J. The essence of snps. *Gene*, Elsevier, v. 234, n. 2, p. 177–186, 1999.

BRYSON, B. *A short history of nearly everything*. [S.l.]: Broadway, 2004.

CORVALAN, A. H. et al. Rprm (reprimo, tp53 dependent g2 arrest mediator candidate). *Atlas of Genetics and Cytogenetics in Oncology and Haematology*, Jean-Loup Huret (Editor-in-Chief); INIST-CNRS (Publisher), 2012.

COSTELLO, J. F.; PLASS, C. Methylation matters. *Journal of medical genetics*, BMJ Publishing Group Ltd, v. 38, n. 5, p. 285–303, 2001.

DU, P. et al. Comparison of beta-value and m-value methods for quantifying methylation levels by microarray analysis. *BMC bioinformatics*, Springer, v. 11, n. 1, p. 1–9, 2010.

EGGER, G. et al. Epigenetics in human disease and prospects for epigenetic therapy. *Nature*, Nature Publishing Group, v. 429, n. 6990, p. 457–463, 2004.

FORTIN, J.-P.; HANSEN, K. D. *Analysis of 450k data using minfi*. 2015. <https://www.bioconductor.org/help/course-materials/2015/BioC2015/methylation450k.html>.

FORTIN, J.-P. et al. Functional normalization of 450k methylation array data improves replication in large cancer studies. *Genome biology*, Springer, v. 15, n. 11, p. 503, 2014.

FORTIN, J.-P.; TRICHE, T. J.; HANSEN, K. D. Preprocessing, normalization and integration of the illumina humanmethylationepic array with minfi. *Bioinformatics*, v. 33, n. 4, 2017.

GOLMOGHADDAM, H. et al. Cd1a and cd1d genes polymorphisms in breast, colorectal and lung cancers. *Pathology & Oncology Research*, Springer, v. 17, n. 3, p. 669–675, 2011.

HANSEN, K. D. et al. Package 'minfi'. 2014.

HANSEN, K. D.; IRIZARRY, R. A. The bumphunter user's guide. 2013.

HANSEN MARTIN ARYEE, R. A. I. K. D. *Analyze Illumina Infinium DNA methylation arrays*. 2020. <https://www.bioconductor.org/packages/release/bioc/manuals/minfi/man/minfi.pdf>.

HIBI, K. et al. Methylation of tfpi2 gene is frequently detected in advanced well-differentiated colorectal cancer. *Anticancer research*, International Institute of Anticancer Research, v. 30, n. 4, p. 1205–1207, 2010.

HICKS, S. C.; IRIZARRY, R. A. When to use quantile normalization? *BioRxiv*, Cold Spring Harbor Laboratory, p. 012203, 2014.

ILLUMINA. *Understanding epigenetic changes: quantitative measurement of methylation at individual CpG sites for high-resolution analysis*. <https://www.illumina.com/science/technology/microarray/infinium-methylation-assay.html>.

ILLUMINA. *Infinium MethylationEPIC BeadChip da Illumina*. 2015. <https://www.molmed.medsci.uu.se/digitalAssets/491/c_491080-l_1-k_epic-data-sheet-2015.pdf>.

INCA, I. N. de C. *O que é câncer?* 2019. <https://www.inca.gov.br/o-que-e-cancer>.

INCA, I. N. de C. *Estatísticas de câncer*. 2020. <https://www.inca.gov.br/numeros-de-cancer>.

IRIZARRY, R. A. et al. Package 'bumphunter'. 2013.

JAFFE, A. E. et al. Bump hunting to identify differentially methylated regions in epigenetic epidemiology studies. *International journal of epidemiology*, Oxford University Press, v. 41, n. 1, p. 200–209, 2012.

JONES, P. A.; BAYLIN, S. B. The fundamental role of epigenetic events in cancer. *Nature reviews genetics*, Nature Publishing Group, v. 3, n. 6, p. 415–428, 2002.

JONES, P. A. et al. Methylation, mutation and cancer. *Bioessays*, Wiley Online Library, v. 14, n. 1, p. 33–36, 1992.

JR, T. J. T. et al. Low-level processing of illumina infinium dna methylation beadarrays. *Nucleic acids research*, Oxford University Press, v. 41, n. 7, p. e90–e90, 2013.

KU, J.-L. et al. Promoter hypermethylation downregulates runx3 gene expression in colorectal cancer cell lines. *Oncogene*, Nature Publishing Group, v. 23, n. 40, p. 6736–6742, 2004.

LI, J. et al. Detection of colorectal cancer in circulating cell-free dna by methylated cpg tandem amplification and sequencing. *Clinical chemistry*, Oxford University Press, v. 65, n. 7, p. 916–926, 2019.

MAKSIMOVIC, J.; GORDON, L.; OSHLACK, A. Swan: Subset-quantile within array normalization for illumina infinium humanmethylation450 beadchips. *Genome biology*, Springer, v. 13, n. 6, p. R44, 2012.

MAKSIMOVIC, J.; PHIPSON, B.; OSHLACK, A. A cross-package bioconductor workflow for analysing methylation array data. *F1000Research*, Faculty of 1000 Ltd, v. 5, 2016.

MCCARTHY, D. J.; SMYTH, G. K. Testing significance relative to a fold-change threshold is a treat. *Bioinformatics*, Oxford University Press, v. 25, n. 6, p. 765–771, 2009.

MOORE, L. D.; LE, T.; FAN, G. Dna methylation and its basic function. *Neuropsychopharmacology*, Nature Publishing Group, v. 38, n. 1, p. 23–38, 2013.

MORRIS, T. J.; BECK, S. Analysis pipelines and packages for infinium humanmethylation450 beadchip (450k) data. *Methods*, Elsevier, v. 72, p. 3–8, 2015.

MURPHY, N. et al. Circulating levels of insulin-like growth factor 1 and insulin-like growth factor binding protein 3 associate with risk of colorectal cancer based on serologic and mendelian randomization analyses. *Gastroenterology*, Elsevier, v. 158, n. 5, p. 1300–1312, 2020.

NIH, N. H. G. R. I. *Probe*. <https://www.genome.gov/genetics-glossary/Probe>.

OLIVEIRA, N. F. P. de et al. Metilação de dna e câncer. *Revista Brasileira de Cancerologia*, v. 56, n. 4, p. 493–499, 2010.

PEREIRA, A. D. Associação entre o consumo alimentar e a metilação dos genes rassf1a e hic1 em indivíduos em rastreamento de câncer colorretal. 2017.

PIDSLEY, R. et al. Critical evaluation of the illumina methylationepic beadchip microarray for whole-genome dna methylation profiling. *Genome biology*, BioMed Central, v. 17, n. 1, p. 1–17, 2016.

SLIEKER, R. C. et al. Identification and systematic annotation of tissue-specific differentially methylated regions using the illumina 450k array. *Epigenetics & chromatin*, BioMed Central, v. 6, n. 1, p. 1–12, 2013.

SLOANE, M. A. et al. Epigenetic inactivation of the candidate tumor suppressor usp44 is a frequent and early event in colorectal neoplasia: Epigenetic inactivation of the candidate tumor suppressor usp44 is a frequent and early event in colorectal neoplasia. *Epigenetics*, Taylor & Francis, v. 9, n. 8, p. 1092–1100, 2014.

STOREY, J. D.; TIBSHIRANI, R. Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences*, National Acad Sciences, v. 100, n. 16, p. 9440–9445, 2003.

TOULEIMAT, N.; TOST, J. Complete pipeline for infinium® human methylation 450k beadchip data processing using subset quantile normalization for accurate dna methylation estimation. *Epigenomics*, Future Medicine, v. 4, n. 3, p. 325–341, 2012.

WANG, Z.; WU, X.; WANG, Y. A framework for analyzing dna methylation data from illumina infinium humanmethylation450 beadchip. *BMC bioinformatics*, Springer, v. 19, n. 5, p. 115, 2018.

WEI, J. et al. Discovery and validation of hypermethylated markers for colorectal cancer. *Disease markers*, Hindawi, v. 2016, 2016.

XU, Z. et al. Relic: a novel dye-bias correction method for illumina methylation beadchip. *BMC genomics*, Springer, v. 18, n. 1, p. 4, 2017.

YANG, B. et al. Correlation of ccna1 promoter methylation with malignant tumors: A meta-analysis introduction. *BioMed research international*, Hindawi, v. 2015, 2015.

5 Apêndice

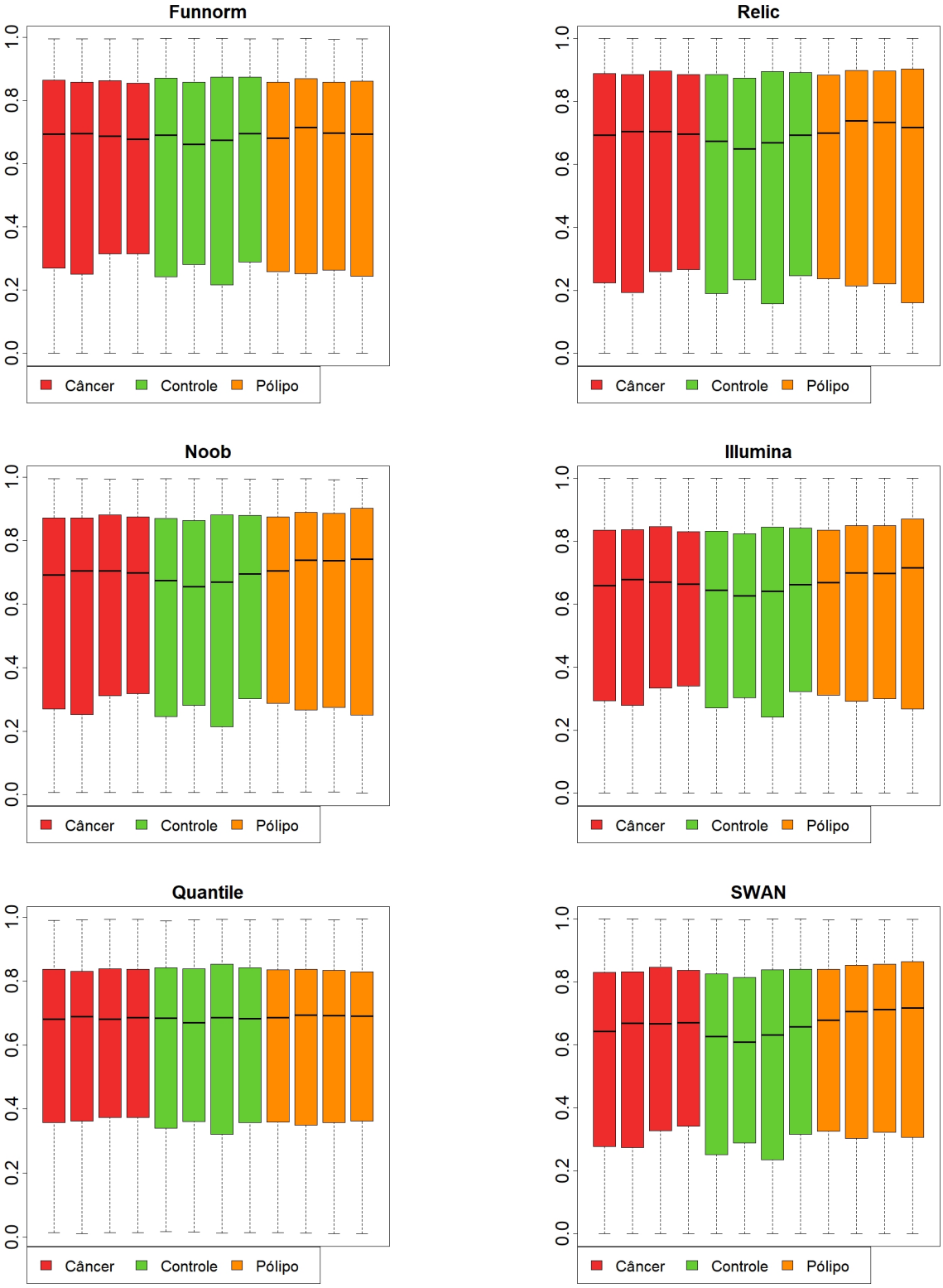

Figura 21: Boxplots de todas as normalizações do valor-Beta.

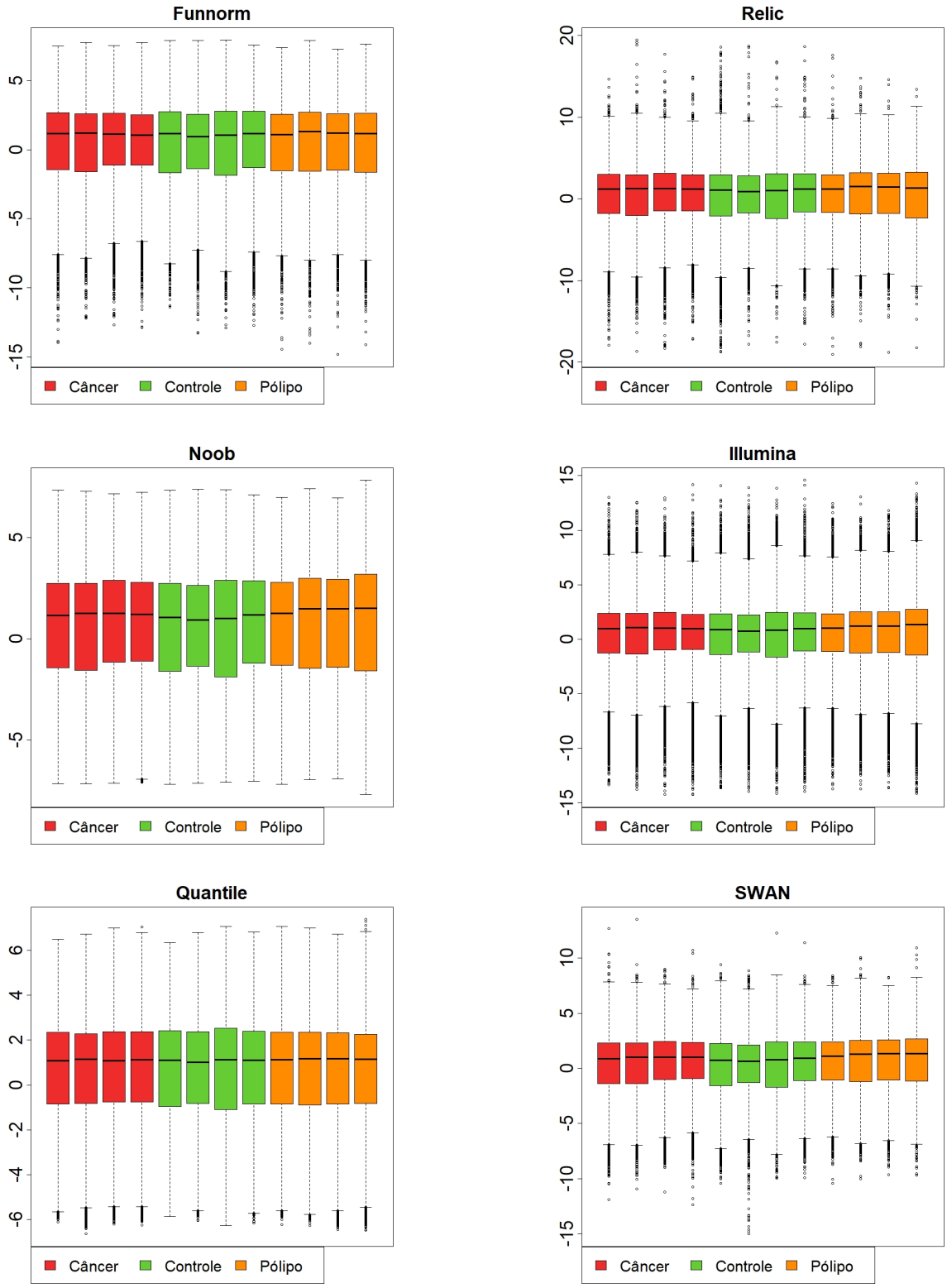

Figura 22: Boxplot de todas as normalizações do valor-M.

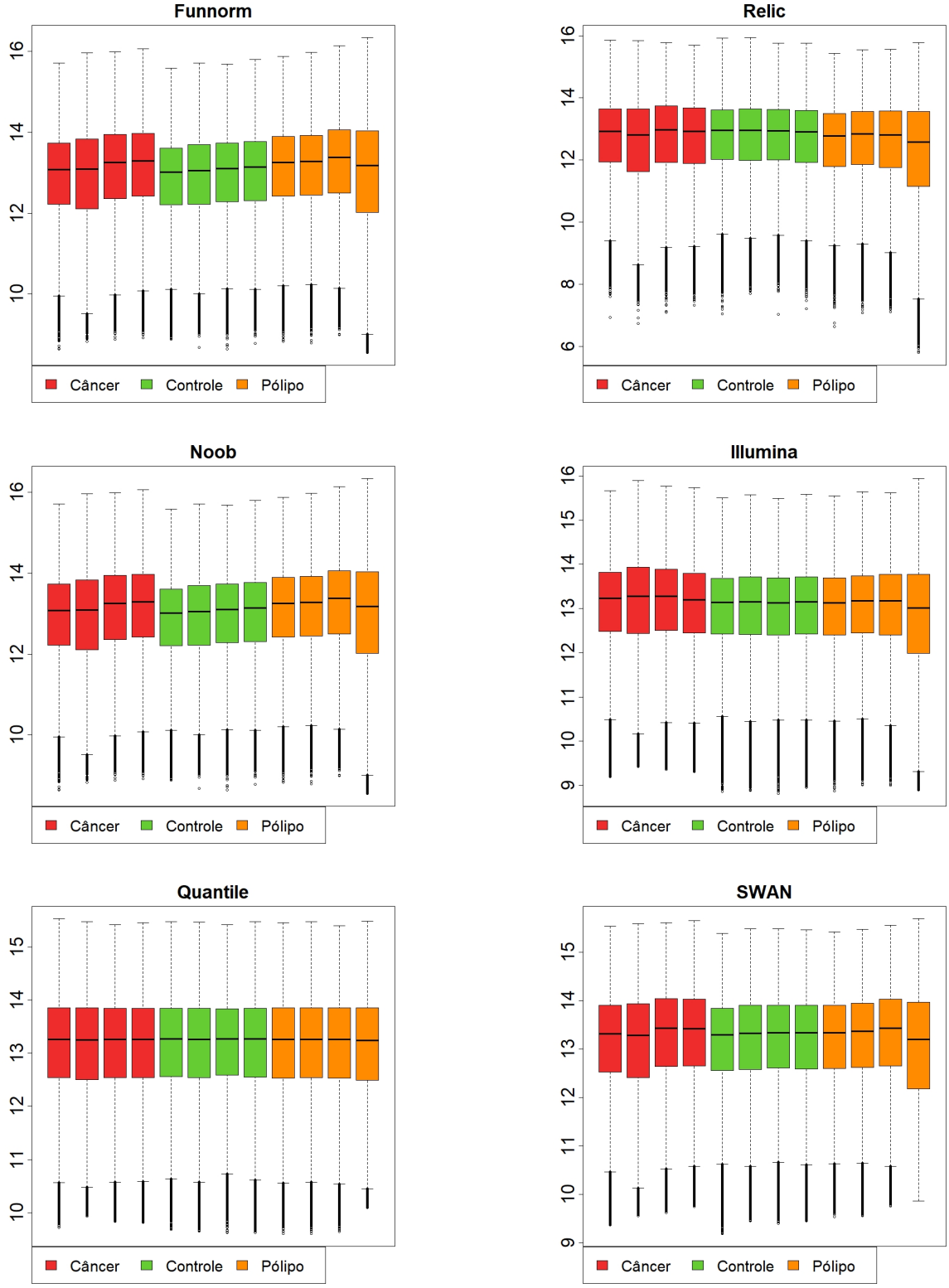

Figura 23: Boxplot de todas as normalizações do valor-CN.

Tabela 7: Número de posições diferencialmente metiladas para todas as normalizações.

Normalização	**Valor Beta**	**Valor M**
Funnorm	149	12
Relic	1.136	1.889
SWAN	120.447	147.387
Illumina	4.573	1
NOOB	5.089	21.011
Quantil	18	21

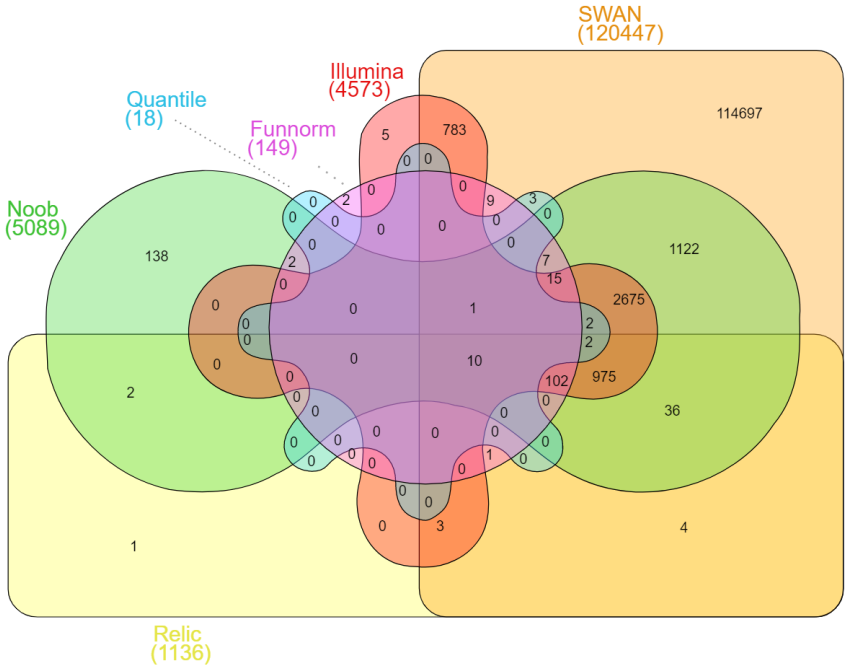

Figura 24: Diagrama de Venn das DMPs de cada normalização com Valor Beta.

Tabela 8: Lista de genes encontrados na identificação de DMRs.

CR	P. inicial	P. final	Norm.	Valor	Gene	Nome completo do gene	Prog.
1	183441292	183441331	F	M	SMG7	SMG7 Nonsense Mediated MRNA Decay Factor	Não
1	109419585	109419685	F	M	GPSM2	G Protein Signaling Modulator 2	Sim
1	121260690	121261404	F	M	LOC647121	Embigin Pseudogene 1 (EMBP1)	Sim
1	186649153	186650479	F	M	PTGS2	Prostaglandin-Endoperoxide Synthase 2	Não
1	158150581	158151363	F	M	CD1D	CD1d Molecule	Não
2	25256369	25259034	R	M	RUNX3	RUNX Family Transcription Factor 3	Sim
2	128768557	128768557	F	M	SAP130	Sin3A Associated Protein 130	Não
2	154333914	154335766	F/R	M	RPRM	Reprimo, TP53 Dependent G2 Arrest Mediator Homolog	Não
2	70994758	70995607	F/R	M	ADD2	Adducin 2	Não
2	182321783	182322501	F	Beta	ITGA4	Integrin Subunit Alpha 4	Não
2	68546467	68547088	F	Beta	CNRIP1	Cannabinoid Receptor Interacting Protein 1	Não
3	93781653	93781653	F	M	NSUN3	NOP2/Sun RNA Methyltransferase 3	Não
4	96468962	96471143	F/R	M	UNC5C	Unc-5 Netrin Receptor C	Não
4	176986621	176987577	F	M	WDR17	WD Repeat Domain 17	Não
4	81186592	81188471	R	M	FGF5	Fibroblast Growth Factor 5	Não
6	10837744	10837744	F	M	MAK	Male Germ Cell Associated Kinase	Não
6	133561614	133562776	F/F/R	M/Beta/Beta	EYA4	EYA Transcriptional Coactivator and Phosphatase 4	Não
7	102985355	102985355	F	M	DNAJC2	DnaJ Heat Shock Protein Family (Hsp40) Member C2v	Não
7	93519220	93520566	F/R/F/R	M/Beta/M/Beta	TFPI2	Tissue Factor Pathway Inhibitor 2	Não
7	99155595	99157059	F/R	M	ZNF655	Zinc Finger Protein 655	Não
7	142494204	142495846	F/R/R	M/M/Beta	TRBJ2-3	T Cell Receptor Beta Joining 2-3	Não
7	45960806	45963523	F/R	M	IGFBP3	Insulin Like Growth Factor Binding Protein 3	Sim
7	27182435	27185732	F	M	HOXA5	Homeobox A5	Não
7	27140797	27144595	R	M	HOXA2	Homeobox A2	Não
8	42751762	42751762	F	M	RNF170	Ring Finger Protein 170	Não
8	67344190	67345006	F/R/F	M/Beta/Beta	ADHFE1	Alcohol Dehydrogenase Iron Containing 1	Não
8	72753888	72757004	F	M	MSC	Musculin	Não

Continua

Conclusão

CR	P. inicial	P. final	Norm.	Valor	Gene	Nome completo do gene	Prog.
8	145105503	145107199	F/R	Beta	OPLAH	5-Oxoprolinase, ATP-Hydrolysing	Não
9	132382433	132382463	F/R	M	C9orf50	Chromosome 9 Open Reading Frame 50	Não
9	100614879	100615357	F/R	Beta	FOXE1	Forkhead Box E1	Não
9	91606128	91606450	R	Beta	C9orf47	Chromosome 9 Open Reading Frame 47	Não
10	7450112	7454759	F/R/F/R	M/Beta/M/Beta	SFMBT2	Scm Like With Four Mbt Domains 2	Não
11	2160882	2162582	F/R	M	IGF2	Insulin Like Growth Factor 2	Não
11	32454718	32457878	F/R	M	WT1	WT1 Transcription Factor	Não
11	123300839	123301758	F/R/F/R	M/Beta/M/Beta	AP000783.1		Não
11	79148183	79152112	F	M	TENM4	Teneurin Transmembrane Protein 4	Não
11	18813466	18813484	F	Beta	PTPN5	Protein Tyrosine Phosphatase Non-Receptor Type 5	Não
11	2889602	2891360	R	M	KCNQ1DN	KCNQ1 Downstream Neighbor	Não
11	110581434	110584091	R	M	ARHGAP20	Rho GTPase Activating Protein 20	Não
12	95941571	95943131	F/R/F/R	M/Beta/M/Beta	USP44	Ubiquitin Specific Peptidase 44	Não
12	45268163	45270997	F	M	NELL2	Neural EGFL Like 2	Não
13	78492916	78494067	F	M	EDNRB	Endothelin Receptor Type B	Não
13	37004536	37006265	F/R	M	CCNA1	Cyclin A1	Não
13	28498091	28498384	F/R	Beta	PDX1	Pancreatic And Duodenal Homeobox 1	Não
13	112759893	112760228	F	Beta	LINC00403	SOX1 Overlapping Transcript (SOX1-OT)	Não
13	103046499	103047287	R	M	FGF14	Fibroblast Growth Factor 14	Não
14	102247610	102248073	F/F/R/R	M/Beta/M/Beta	PPP2R5C	Protein Phosphatase 2 Regulatory Subunit B'Gamma	Não
14	70653919	70656116	F/R	M	SLC8A3	Solute Carrier Family 8 Member A3	Não
15	48936953	48937213	F/R	M	FBN1	Fibrillin 1	Não
15	74420307	74422572	F/R	M	LOC283731	Uncharacterized LOC283731	Não
15	79382548	79383980	F/R	M	RASGRF1v	Ras Protein Specific Guanine Nucleotide Releasing Factor 1	Não
16	70472678	70472993	F	M	ST3GAL2	ST3 Beta-Galactoside Alpha-2,3-Sialyltransferase 2	Não
16	58497230	58499074	R	M	NDRG4	NDRG Family Member 4	Não
17	46655561	46656892	F/F/R/R	M/Beta/M/Beta	HOXB4	Homeobox B4	Não
17	75368750	75370611	F/R/F	M/M/Beta	SEPT9	Septin 9	Não

Continua

Conclusão

CR	P. inicial	P. final	Norm.	Valor	Gene	Nome completo do gene	Prog.
19	58739734	58740747	F/R/F	M/M/Beta	ZNF544	Zinc Finger Protein 544	Não
20	57427010	57427977	F	M	GNAS	GNAS Complex Locus	Não
20	24449668	24452131	R	M	SYNDIG1	Synapse Differentiation Inducing 1	Não

Abreviações: CR = Cromossomo; P. = Posição; Norm. = Normalização; F = `Funnorm`; R = `Relic`; B = Beta; Prog. = Prognóstico de CCR; Exp. = Expressão da proteína em CCR; Pend. = Pendente.