XXXI Congresso de Iniciação Científica Unicamp 2023

MÉTODOS ESTATÍSTICOS APLICADOS AO ESTUDO DE PADRÕES DIFERENCIAIS DE METILAÇÃO DE DNA

Palavras-chave: bioinformática, metilação de DNA, EWAS

Autores:
Victor Vinícius Gomes [IMECC, Unicamp]
Profa. Samara Flamini Kiihl (orientadora) [DE - IMECC, Unicamp]

1 Introdução

A epigenética é a área da biologia que estuda as mudanças no fenótipo do seres e que, ao mesmo tempo, não envolva alterações no código genético. Um tipo de mudança muito estudada é a metilação do DNA. Trata-se da interação de uma molécula de metil com as bases nitrogenadas Citosina (C) e Guanina (G), unificadas por um Fósforo (p). Nos mamíferos, cerca de 70 a 80% de todos os CpGs estão metilados [10] e, além disso, estão concentrados, em sua maioria, em locais chamados de ilhas de CpG (ou CGIs). Estas ilhas são mais comuns em regiões de genes promotores (cerca de 70% do total [5]), e não apresentam metilação do DNA. Mudanças nas taxas de metilação, no entanto, podem ser indicativos de alguma desregulação - possivelmente causada por fatores ambientais - e são chamadas de hiper ou hipometilação, quando são maiores ou menores que as taxas normais, respectivamente.

A análise diferencial da metilação é usada tanto para saber se as alterações estão sendo causadas por mudanças no perfil de metilação dos indivíduos quanto para identificar especificamente quais os CpGs responsáveis por causar essas alterações fenotípicas. Considerando a natureza dos estudos, estes são chamados de associativos. Neste projeto de iniciação científica, uma análise de metilação diferencial do DNA completa é desenvolvida, desde a escolha de uma base de dados adequada, a definição do pré-processamento a ser aplicado (filtragens, normalização, etc) e, posteriormente, a análise em si, juntamente com a disponibilização dos códigos desenvolvidos.

2 Metodologia

2.1 Metilação do DNA

A metilação do DNA é observada naturalmente nos organismos e está relacionada à ação de repressão da transcrição genética. É essencial para o desenvolvimento saudável de mamíferos e está associada a diversos eventos, como a formação da estrutura de cromatina. Além de sua importância natural, recentemente, estudos estão sendo conduzidos para analisar sua associação com doenças autoimunes e mesmo com diversos tipos de câncer. Além de ajudar a entender tais doenças melhor, a metilação do DNA é um evento que não altera o DNA, portanto pode ser considerada como estratégia de tratamento para tais doenças, futuramente [11].

Atualmente, no entanto, a investigação de quais doenças apresentam associação com a metilação e onde no DNA essa associação é encontrada é de grande interesse. Muito deste estudo necessita de análises estatísticas que auxiliam no fortalecimento das evidências de tais associações. Estes estudos apenas são possíveis com a rápida evolução tecnológica e científica envolvendo a coleta de DNA e de métodos de identificação de locais onde existe, ou não, metilação do mesmo. Neste projeto, o método usado e mencionado a todo momento é o de arranjos (ou *Illumina Methylation assay*), por suas diversas vantagens.

2.2 Arranjos

Os arranjos da marca *Illumina* usam da tecnologia chamada *beadChip* para gerar o perfil de metilação do DNA em todo o genoma (detalhes em *Bead-Based Microarray Technology*). A tecnologia 27K foi a primeira a ser lançada, atingindo aproximadamente 27 mil dinucleotídeos de CpG em 14 mil genes [19]. Depois, o arranjo 450K (*450k Array*) foi lançado, estendendo o alcance para mais de 450 mil CpGs. Mais recentemente, em 2016, o arranjo do tipo EPIC (*EPIC Array*) foi desenvolvido, com quase o dobro do alcance de seu predecessor, chegando a mais que 850 mil sítios de CpG. Cada CpG interrogado é chamado de sonda, e cada indivíduo de um estudo é chamado de amostra. Com relação ao funcionamento dos arranjos, é importante considerar a interpretação do que é capturado pelos mesmos. Ao interrogar um sítio de CpG, a tecnologia

BeadChip simplificadamente recebe uma resposta de luz (verde ou vermelha) que indica o estado de tal sítio. A intensidade dessa luz, juntamente com a cor e o tipo de chip usado são analisados para quantificar a possibilidade de metilação daquele sítio. Em geral, tem-se a intensidade metilada (M) e não metilada (U) e calculam-se dois valores: o valor de beta (β) e o valor M; suas fórmulas são indicadas em 1 e 2, respectivamente. O valor de beta se refere a proporção de intensidade metilada, enquanto o valor de M indica o log2 da razão entre a intensidade metilada e a intensidade não metilada.

$$\beta = \frac{M}{M+U} \tag{1}$$

$$Mval = log_2\left(\frac{M}{U}\right) \tag{2}$$

Conforme discutido na literatura, durante o desenvolvimento de análises estatísticas, o valor de M será priorizado; para a apresentação de resultados gráficos, o valor de beta será usado. A decisão leva em conta as melhores propriedades estatísticas do valor de M, e a maior intuitividade por trás do valor de beta [6].

3 Resultados e discussão

3.1 Leitura e Processamento dos Dados

O projeto foi desenvolvido com o auxílio do *software R* [13]. Para as análises, os pacotes *minfi* [12] e *limma* [14] foram empregados, fornecendo o ferramental essencial para o estudo adequado e completo dos dados analisados.

A base de dados usada no projeto foi obtida no *website* `https://www.ncbi.nlm.nih.gov/geo/`. Cada indivíduo analisado tem 2 arquivos associados, um contendo as intensidades verdes e outro contendo as intensidades vermelhas, as quais fornecem a base para encontrar os valores de β e M. É possível realizar o *download* dos dados pelo *link* `https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE147430`, no entanto, recomenda-se a utilização de um *software* intermediário, considerando o tamanho e organização dos arquivos. A base fala sobre os efeitos do fumo na metilação do DNA humano em células CD8 T do sangue [16], e contém amostras de 132 indivíduos, sendo 121 com arranjos do tipo 450K e 11 com arranjos do tipo EPIC.

3.2 Pré-Processamento

O pré-processamento dos dados é essencial para mitigar os efeitos de ruídos aleatórios, assim como efeitos de lote que podem acarretar em vieses de grande impacto nas análises posteriores. Algumas formas de processamento diferentes foram exploradas, buscando reduzir as inconsistências encontradas entre os diferentes métodos. O *download* e leitura dos dados é feita usando a função *getGEO()* do pacote *GEOquery* [4]. Uma vez baixados, os arquivos são organizados em pastas específicas (uma para os arranjos *450K* e outra para os arranjos *EPIC*), onde depois são lidos como experimentos separados, juntamente com as informações fenotípicas (as características importantes para a análise diferencial, além de possíveis características de confundimento, como o tipo de arranjo) referentes aos experimentos e posteriormente unificados, como mencionado acima. Ao final do processo de leitura dos dados, tem-se um único conjunto de dados, contendo as informações referentes às intensidades (brutas, não-processadas) de luz vermelha e verde dos experimentos.

3.2.1 Controle de Qualidade

Após os passos de leitura, inicia-se a filtragem de amostras (sejam elas de sondas ou indivíduos) ruins. O pacote *Minfi* vem carregado de funções especificamente designadas para tratar de amostras que apresentam comportamento anômalo. O controle de qualidade deve ser feito antes da normalização, para que a influência de tais amostras nas análises seja minimizada, evitando vieses.

Para encontrar locais com intensidade total (metilados + não-metilados) pequena, usa-se a função *detectionP()* do pacote *Minfi*, que compara essa intensidade total com o nível da intensidade do sinal de fundo. O resultado, nos dados estudados, pode ser observado na Figura 1, e indica que apenas uma amostra será descartada, devido ao seu comportamento irregular.

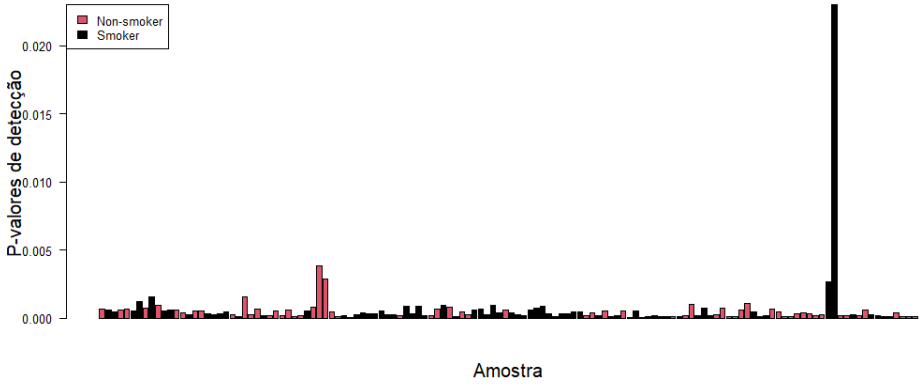

Figura 1: P-valores de detecção para as amostras, agrupando por fumantes e não-fumantes.

Uma forma eficiente de detectar amostras anômalas é o uso do log2 da mediana dos valores de intensidades metiladas e não-metiladas por indivíduo. Empiricamente, os autores do pacote *Minfi* definem um limiar que separa, com grande certeza, amostras bem-comportadas de amostras anômalas. Observa-se, na Figura 2, quais amostras foram classificadas como anômalas, juntamente com a linha separadora, definida pelos limiares das intensidades log2.

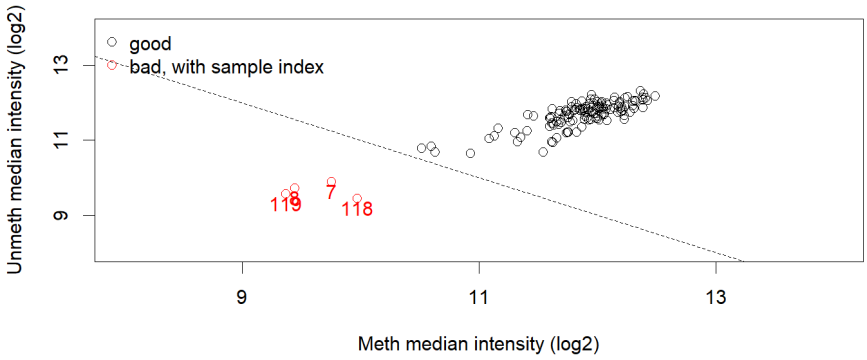

Figura 2: Gráfico de dispersão de medianas de valores metilados e não-metilados por indivíduo.

Outro gráfico que acaba contribuindo com a análise de amostras anômalas é o gráfico de densidade para os valores de beta de cada indivíduo. A Figura 3 indica como essa densidade se comporta nos dados antes do pré-processamento; cada linha (densidade) representa um indivíduo. A bimodalidade observada é esperada, ao mesmo tempo que a variação entre indivíduos deve ser razoavelmente pequena, pois pode ser efeito de fatores técnicos e não representativa das diferenças fenotípicas estudadas.

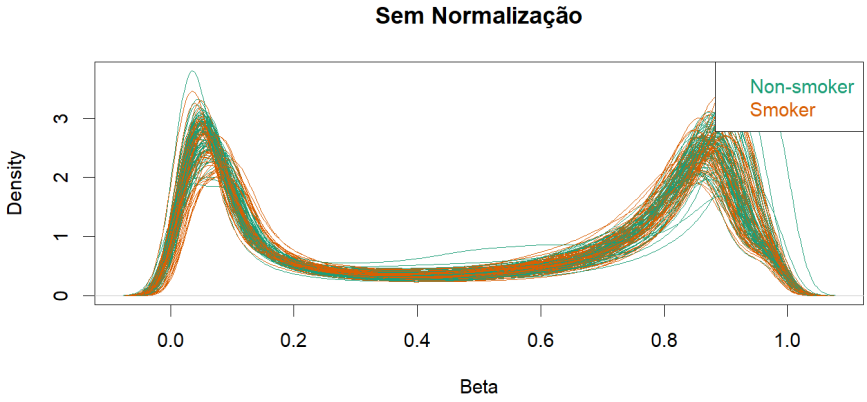

Figura 3: Densidade dos valores de beta antes da normalização, agrupando por fumantes e não-fumantes.

3.2.2 Normalização

A normalização foi feita comparando os resultados entre duas escolhidas: a normalização *ssNoob* [18] e a Quantílica [17]. O gráfico de densidade dos valores de Beta na Figura 4 mostra a comparação entre os dois métodos nos dados. Como a variabilidade entre indivíduos foi menor no caso da normalização quantílica - ao mesmo tempo que não indicou amostras anômalas pós normalização - esta foi escolhida para os procedimentos posteriores de análise.

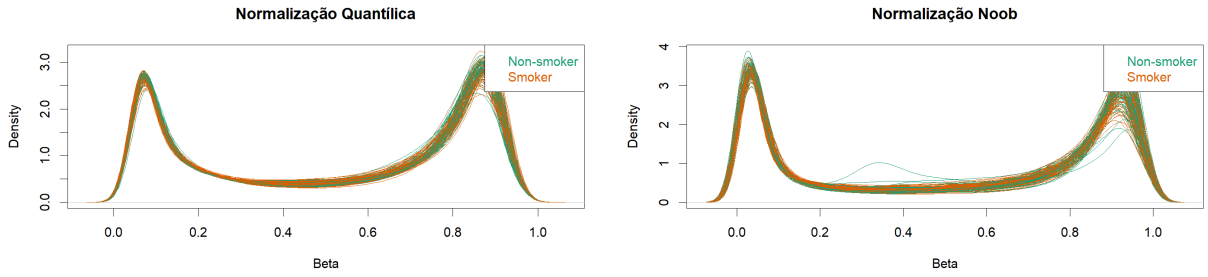

Figura 4: Densidade dos valores de beta usando normalização Quantílica (Esquerda) e ssNoob (Direita), agrupando por fumantes e não-fumantes.

3.2.3 Filtragens Adicionais

As filtragens adicionais incluem a remoção de sondas específicas que podem adicionar vieses indesejados nas análises. Sondas relativas ao sexo dos indivíduos (cromossomos X e Y) por serem de difícil incorporação nas análises [9], *cross-hibridized probes* [1] e *Single Nucleotide Polymorphism* (SNP) [3] devem ser removidos, no geral, como boa prática, quando não são do interesse geral da análise.

Além das filtragens mencionadas, observou-se um agrupamento perfeito (em gráficos feitos usando métodos de redução de dimensionalidade) de amostras vindas de arranjos *450K* e *EPIC*, como é visível na Figura 5, onde foi aplicado método *Multidimensional Scaling* (ou MDS) [8] para visualizar a distribuição espacial dos indivíduos pelos valores M por CpG de cada um, agrupando por *design* do arranjo. Assim, decidiu-se remover as amostras do tipo *EPIC*, já que tratam-se de apenas 11 indivíduos. Vale notar, no entanto, que esta decisão não é a ideal, já que existem métodos de correção de vieses deste tipo, mas por se tratar de uma quantidade pequena de indivíduos com este tipo de arranjo, a análise não seria fortemente afetada pela remoção dos mesmos.

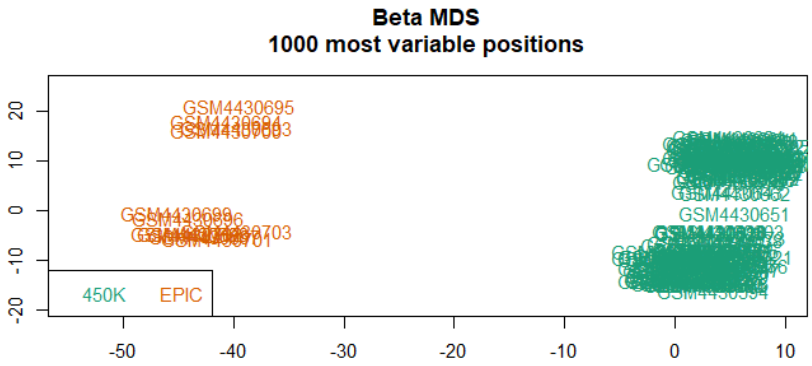

Figura 5: Gráfico de Escalonamento Multidimensional agrupando por design do arranjo.

3.3 Análise Diferencial

A análise diferencial é feita para encontrar quais CpGs apresentam evidências estatísticas de diferenciação por meio da variável de interesse: neste caso, o status de fumante dos indivíduos. Para cada CpG, é ajustado um modelo de regressão linear que usa como covariável o status de fumante e, como variável dependente, o valor M (Equação 2), que é o logito da metilação do CpG. O modelo considera cada indivíduo como uma observação única para dado CpG. Alguns procedimentos de correção de P-valor são utilizados: são ajustados cerca de 450 mil modelos, o que aumenta muito a chance de descobertas falsas (também conhecido como o erro do tipo I, rejeitar a hipótese nula quando na verdade esta é verdadeira). Os procedimentos usados foram o *Empirical Bayes* [15], para computar as estatísticas moderadas dos testes de hipótese de interesse, usando uma distribuição melhor comportada para os erros padrão dos parâmetros dos modelos, e o método BH (método de Benjamini-Hochberg [2]) para correção de p-valores. Por fim, tem-se uma lista com os CpGs significativamente diferenciais encontrados.

Além da variável referente ao status de fumo, considerou-se a utilização do sexo dos indivíduos (não disponível na base, mas produto adicional da normalização quantílica). A integração desta nos modelos se mostrou não significativa, e foi usada a biblioteca VennDetail [7], para análise da intersecção entre os CpGs significativos entre ambos os modelos, avaliando a concordância de ambos. Como resultado da análise, e considerando a não significância do sexo nos modelos, as análises posteriores seguiram apenas com a variável do status de fumante.

Considerando as configurações finais de modelos, 86 CpGs apresentaram significância estatística através de todos os indivíduos, podendo ser, posteriormente, estudados a fundo por profissionais adequados. Esta lista será disponibilizada, juntamente com o código, no GitHub do projeto. Por fim, algumas análises mais simples de *Gene Ontology* (ou *GO*) foram aplicadas aos CpGs significantes, buscando padrões nas funções dos genes afetados por estes CpGs, mas não foram encontrados resultados relevantes.

Referências

[1] Yi an Chen, Mathieu Lemire, Sanaa Choufani, Darci T. Butcher, Daria Grafodatskaya, Brent W. Zanke, Steven Gallinger, Thomas J. Hudson, and Rosanna Weksberg. Discovery of cross-reactive probes and polymorphic cpgs in the illumina infinium humanmethylation450 microarray. *Epigenetics*, 8(2):203–209, 2013. PMID: 23314698.

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

[3] Zebracka-Gala J Daca-Roszak P, Pfeifer A, Rusinek D, Szybinska A, Jarzab B, Witt M, and Zietkiewicz E. Impact of snps on methylation readouts by illumina infinium humanmethylation450 beadchip array: implications for comparative population studies. *BMC Genomics*, 2015.

[4] Sean Davis and Paul Meltzer. Geoquery: a bridge between the gene expression omnibus (geo) and bioconductor. *Bioinformatics*, 14:1846–1847, 2007.

[5] Aimée M. Deaton and Adrian Bird. Cpg islands and the regulation of transcription. *Genes & Development*, 2011.

[6] Zhang Du, P., X., Huang, and CC. et al. Comparison of beta-value and m-value methods for quantifying methylation levels by microarray analysis. *BMC Bioinformatics*, 11, 2010.

[7] Kai Guo and Brett McGregor. *VennDetail: A package for visualization and extract details*, 2022. R package version 1.14.0.

[8] P Groenen I Borg. *Modern Multidimensional Scaling: theory and applications.*, volume 2. Springer-Verlag, New York, 2005.

[9] Wong Inkster, A.M., M.T., Matthews, and A.M. et al. Who's afraid of the x? incorporating the x and y chromosomes into the analysis of dna methylation array data. *Epigenetics & Chromatin*, 2023.

[10] Andrew P. Feinberg et al. Liora Z. Strichman-Almashanu, Richard S. Lee. A genome-wide screen for normally methylated human cpg islands that can identify novel imprinted genes. *Genome Research*, 2002.

[11] Leszek Roszkowski Marzena Ciechomska and Wlodzimierz Maslinski. Dna methylation as a future therapeutic and diagnostic target in rheumatoid arthritis. *Cells*, 2019.

[12] Aryee MJ, Jaffe AE, Corrada-Bravo H, Ladd-Acosta C, Feinberg AP, Hansen KD, and Irizarry RA. Minfi: A flexible and comprehensive bioconductor package for the analysis of infinium dna methylation microarrays, 2014.

[13] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2022.

[14] Matthew E Ritchie, Belinda Phipson, Di Wu, Yifang Hu, Charity W Law, Wei Shi, and Gordon K Smyth. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research*, 43(7):e47, 2015.

[15] Gordon Smyth. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Statistical applications in genetics and molecular biology*, 3:Article3, 02 2004.

[16] Martos SN, Campbell MR, Lozoya OA, and Wang X et al. Single-cell analyses identify dysfunctional cd16+ cd8 t cells in smokers. *Cell Rep Med*, 2020.

[17] N Touleimat and J Tost. Complete pipeline for infinium human methylation 450k beadchip data processing using subset quantile normalization for accurate dna methylation estimation. *Epigenomics*, 4, 2012.

[18] Van Den Berg D Triche TJ, Weisenberger DJ, Laird PW, and Siegmund KD. Low-level processing of illumina infinium dna methylation beadarrays, 2013.

[19] DJ. et al. Weisenberger. Comprehensive dna methylation analysis on the illumina infinium assay platform. 2008.