# Aplicação de métodos de aprendizado de máquina não-supervisionado em dados de metilação de DNA de pacientes de COVID-19

Orientadora: Profa. Samara Kiihl
Aluno: Guilherme Pereira de Freitas

**Resumo**

Estudos recentes vêm demonstrando a importância da metilação de DNA, um marcador epigenético importante, na regulação da expressão gênica. Métodos de agrupamento ou de aprendizado de máquina não-supervisionado são utilizados para extrair informações para diagnóstico precoce e tratamentos a partir de dados de alta dimensão dos estudos epigenômicos. Neste projeto, iremos aplicar métodos de agrupamento em dados de metilação de DNA de 407 pacientes de COVID-19, onde 194 (47.7%) estavam com sintomas leves e 213 (52.3%), com o objetivo de encontrar grupos que podem ter sido formatos por influência da doença.

**Palavras-chave:** bioinformática, metilação de DNA, aprendizado de máquina não supervisionado, métodos de agrupamento, COVID-19, EWAS

## 1 Introdução

Embora todas as células de um organismo apresentem, essencialmente, o mesmo conteúdo genético, suas funções e particularidades se dão por meio do regulamento da expressão gênica. Tal regulamento ocorre por meio de mecanismos epigenéticos, como a metilação do DNA, modificação de histonas e outros processos mediados por RNA, que influenciam principalmente a expressão gênica a nível de transcrição (Gibney e Nolan 2010).

O mecanismo estudado foi a metilação da citosina (5mC), que acontece em áreas específicas de regulação, como regiões promotoras ou de heterocromatina. Esse fenômeno pode modificar, significativamente, a expressão temporal e espacial dos genes e a remodelação da cromatina (Heyn e Esteller 2012). Em mamíferos, as citosinas metiladas estão restritas às CpGs (citosina-fosfato-guanina), onde elas antecedem uma guanina (G) na direção de 5'. Vale lembrar que o DNA é formado por 4 nucleotídeos e

[redacted]

portanto existem 16 p[redacted]uência, o que ajuda a identificar as ilhas CpGs, pois estas apresentam uma frequência maior desse par (CG) quando comparadas com outras regiões arbitrárias (Gibney e Nolan 2010).

# 2 Banco de dados

O banco de dados utilizado foi retirado do estudo de associação epigenômica ampla (EWS - *epigenome-wide association study*) de COVID-19 realizado por (Moura et al. 2021). Os dados estão disponíveis no repositório público *The Gene Expression Omnibus* (GEO), sob código de acesso GSE168739. Trata-se de 407 pacientes de COVID-19 sem comorbidades e com idade máxima de 61 anos, onde 194 (47.7%) estavam com sintomas leves e 213 (52.3%) estavam com sintomas graves. Os dados foram coletados através da plataforma *Infinium MethylationEPIC BeadChip*, totalizando em aproximadamente 850 mil ilhas CpGs para cada indivíduo.

# 3 Pré-processamento

O fluxo de pré-processamento foi feito seguindo o passo a passo descrito no artigo "A cross-package Biodonductor workflow for analysing methylation array data" (Maksimovic, Phipson, e Oshlack 2016), por meio das ferramentas dispostas no pacote Bioconductor (Huber et al. 2015), disponíveis para a linguagem R (R Core Team 2020). Os algoritmos são aplicados na matriz de p-valores.

A matriz de p-valores é obtida comparando-se a distribuição das intensidades, para cada par de indivíduos e ilhas, com a distribuição do ruído de fundo (que por sua vez, foi calculado a partir das sondas de controle). Cada um dos ensaios (tipo I e tipo II) apresenta sua distribuição própria do ruído de fundo, bem como a intensidade de metilação dos indivíduos.

Sem entrar em muitos detalhes, as etapas de pré-processamento consistiram em remover amostras de baixa qualidade, aplicar a normalização quantílica (Touleimat e Tost 2012) e remover sondas de CpGs referentes ao sexo do paciente e/ou com reatividade cruzada (Chen et al. 2013).

Devido a limitações computacionais para realizar as etapas de pré-processamento, esse trabalho usou recursos do Centro Nacional de Processamento de Alto Desempenho em São Paulo (CENAPAD-SP).

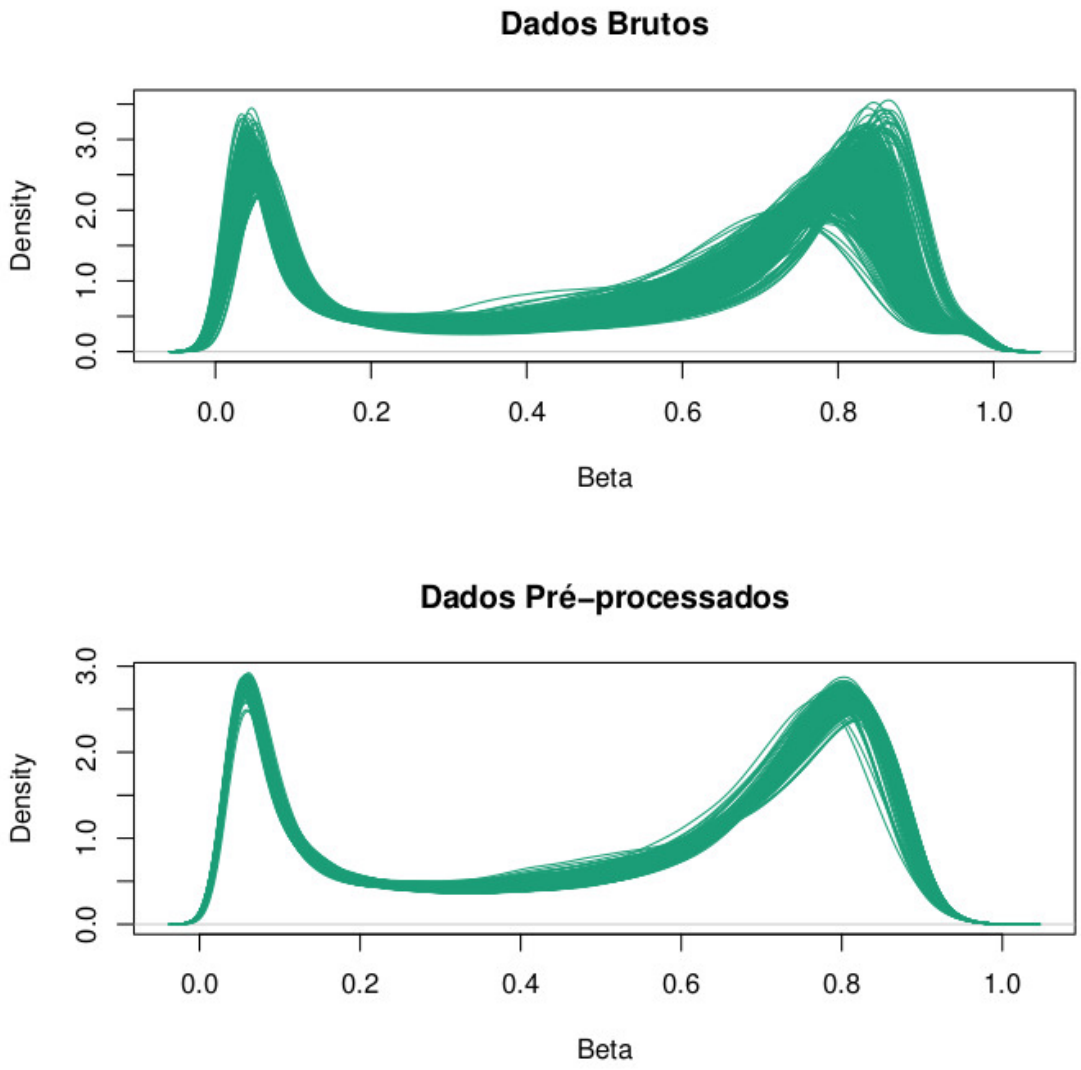



Figura 1: Densidade das taxas de metilação antes e depois do pré-proprocessamento

# 4 Aprendizado de máquina não-supervisionado

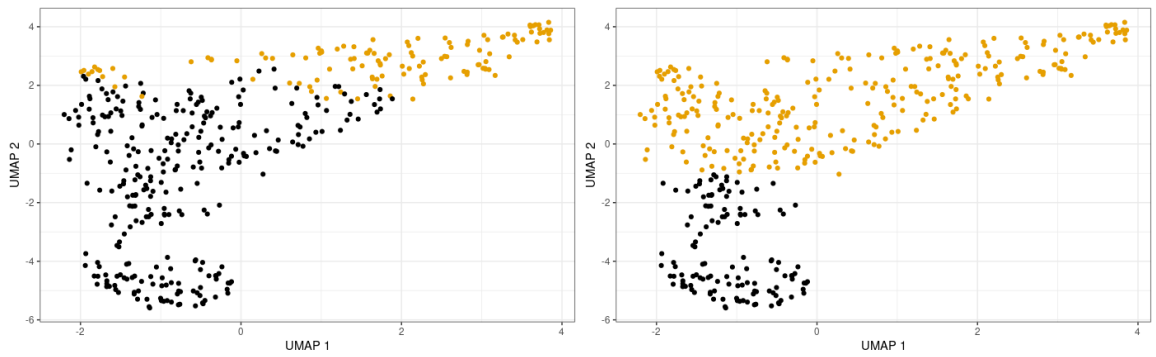
O aprendizado de máquina não-supervisionado também é conhecido como análise de cluster, ou análise de agrupamento. Uma das maiores diferenças entre aprendizado de máquina supervisionado e não-supervisionado está na falta de dados de treinamento para a última, bem como a falta de um target para tal. Os pré-requisitos para aplicar as técnicas de agrupamento se dão na escolha das variáveis, hiperparâmetros e tipo de distância adotada (Gentleman e Carey 2008).

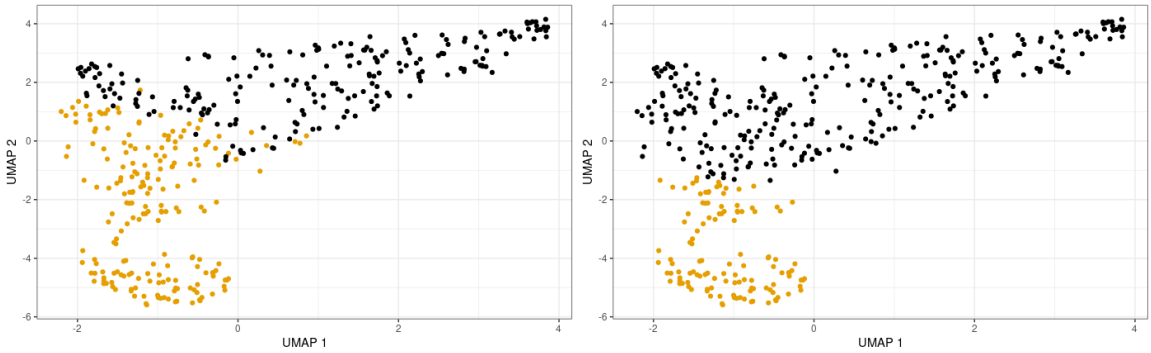
Com a matriz final dos betas pré-processados, podemos calcular a matriz de dissimilaridade entre os indivíduos por meio da distância euclidiana, dada pela fórmula $D(X,Y) = \sqrt{\sum_i^n (x_i - y_i)^2}$, onde $n$ é o número total de CpGs e X e Y são os vetores de betas de dois indivíduos. Cada um dos métodos foi aplicado na matriz de distâncias completa e em duas compontentes reduzidas pelo método Uniform Manifold Approximation and Projection (UMAP) (McInnes, Healy, e Melville 2020).

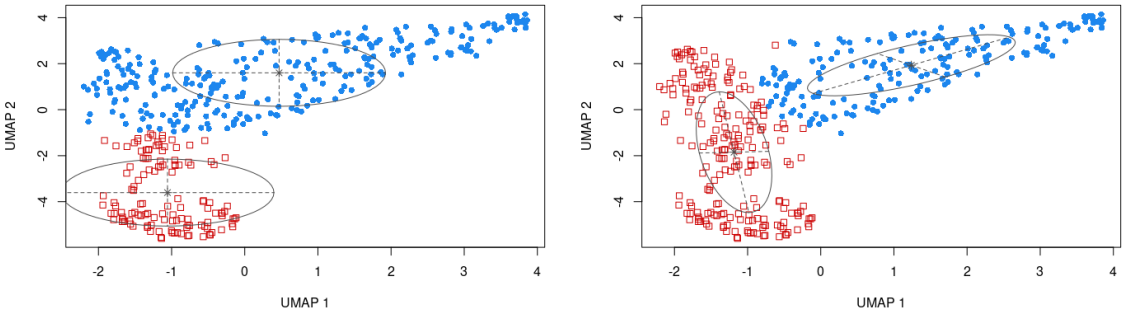
Para avaliar a formação de clusteres, em cada um dos métodos, utilizou-se as técnicas de Silhueta(Rousseeuw 1986) e Gráfico de Elbow (Hayasaka 2022).

Os métodos de agrupamento utilizados foram K-Médias (Lloyd 1982), PAM - Partition Around Medoids (Kaufman Rousseeuw 2009), Agrupamento Hierárquico (Mardia, Bibby, e Kent 1979) e Mistura de Modelos.

# 5 Resultados




UMAP 2
UMAP 1
UMAP 2
UMAP 1


UMAP 2
UMAP 1
UMAP 2
UMAP 1

# Referências

Chen, Yi-an, Mathieu Lemire, Sanaa Choufani, Darci T. Butcher, Daria Grafodatskaya, Brent W. Zanke, Steven Gallinger, Thomas J. Hudson, e Rosanna Weksberg. 2013. "Discovery of cross-reactive probes and polymorphic CpGs in the Illumina Infinium HumanMethylation450 microarray". *Epigenetics* 8 (2): 203–9. https://doi.org/10.4161/epi.23470.

Gentleman, R., e V. J. Carey. 2008. "Unsupervised Machine Learning". In *Bioconductor Case Studies*, 137–57. Springer New York. https://doi.org/10.1007/978-0-387-77240-0_10.

Gibney, E R, e C M Nolan. 2010. "Epigenetics and gene expression". *Heredity* 105 (1): 4–13. https://doi.org/10.1038/hdy.2010.54.

Hayasaka, Satoru. 2022. "How many clusters?". *Medium*. Towards Data Science. https://towardsdatascience.com/how-many-clusters-6b3f220f0ef5.

Heyn, Holger, e Manel Esteller. 2012. "DNA methylation profiling in the clinic: applications and challenges". *Nature Reviews Genetics* 13 (10): 679–92. https://doi.org/10.1038/nrg3270.

Huber, W., V. J. Carey, R. Gentleman, S. Anders, M. Carlson, B. S. Carvalho, H. C. Bravo, et al. 2015. "Orchestrating high-throughput genomic analysis with Bioconductor". *Nature Methods* 12 (2): 115–21. http://www.nature.com/nmeth/journal/v12/n2/full/nmeth.3252.html.

Kaufman Rousseeuw, Leonard Peter J -. 2009. *Finding Groups in Data*. John Wiley & Sons, Inc.

Lloyd, S. 1982. "Least squares quantization in PCM". *IEEE Transactions on Information Theory* 28 (2): 129–37. https://doi.org/10.1109/TIT.1982.1056489.

Maksimovic, Jovana, Belinda Phipson, e Alicia Oshlack. 2016. "A cross-package Bioconductor workflow for analysing methylation array data". *F1000Research* 5 (junho): 1281. https://doi.org/10.12688/f1000research.8839.1.

Mardia, K V, J M Bibby, e J T Kent. 1979. *Multivariate analysis*. Livro. http://www.loc.gov/catdir/toc/els031/79040922.html.

McInnes, Leland, John Healy, e James Melville. 2020. "UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction". http://arxiv.org/abs/1802.03426.

Moura, Manuel Castro de, Veronica Davalos, Laura Planas-Serra, Damiana Alvarez-Errico, Carles Arribas, Montserrat Ruiz, Sergio Aguilera-Albesa, et al. 2021. "Epigenome-wide association study of COVID-19 severity with respiratory failure". *EBioMedicine* 66 (abril): 103339. https://doi.org/10.1016/j.ebiom.2021.103339.

R Core Team. 2020. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Rousseeuw, Peter J. 1986. "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis".

Touleimat, Nizar, e Jörg Tost. 2012. "Complete pipeline for Infinium Human Methylation 450K BeadChip data processing using subset quantile normalization for accurate DNA methylation estimation". *Epigenomics* 4 (3): 325–41. https://doi.org/10.2217/epi.12.21.