



UNIVERSIDADE FEDERAL DE PERNAMBUCO
CENTRO DE INFORMÁTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

GEORGE DE VASCONCELOS CARVALHO NETO

Mineração de subgrupos aplicada a dados de metilação de pacientes com câncer de mama

Recife

2022

GEORGE DE VASCONCELOS CARVALHO NETO

Mineração de subgrupos aplicada a dados de metilação de pacientes com câncer de mama

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Ciência da Computação da Universidade Federal de Pernambuco, como requisito parcial para a obtenção do título de Mestre em Ciência da Computação.

Área de Concentração: Inteligência Computacional

Orientador (a): Prof. Dr. Ricardo Bastos Cavalcante Prudêncio

Coorientador (a): Prof. Dr. André Câmara Alves do Nascimento

Recife

2022

Catálogo na fonte
Bibliotecária Monick Raquel Silvestre da S. Portes, CRB4-1217

C331m Carvalho Neto, George de Vasconcelos
Mineração de subgrupos aplicada a dados de metilação de pacientes com
câncer de mama / George de Vasconcelos Carvalho Neto. – 2022.
74 f.: il., fig.

Orientador: Ricardo Bastos Cavalcante Prudêncio.
Dissertação (Mestrado) – Universidade Federal de Pernambuco. CIn,
Ciência da Computação, Recife, 2022.

Inclui referências.

1. Inteligência computacional. 2. Mineração. I. Prudêncio, Ricardo Bastos
Cavalcante (orientador). II. Título.

006.31

CDD (23. ed.)

UFPE - CCEN 2022-113

George de Vasconcelos Carvalho Neto

**“Mineração de subgrupos aplicada a dados de metilação de pacientes com
câncer de mama”**

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Ciência da Computação da Universidade Federal de Pernambuco, como requisito parcial para a obtenção do título de Mestre em Ciência da Computação. Área de Concentração: Inteligência Computacional.

Aprovado em: 25 de março de 2022.

BANCA EXAMINADORA

Prof. Dr. Ricardo Bastos Cavalcante Prudêncio
Centro de Informática/UFPE
(**Orientador**)

Profa. Dra. Danyelly Bruneska Gondim Martins
Departamento de Bioquímica / UFPE

Prof. Dr. Marcelo Rodrigo Portela Ferreira
Departamento de Estatística / UFPB

Dedico essa dissertação a toda minha família, amigos e professores que me deram todo suporte necessário até aqui.

AGRADECIMENTOS

A família, especialmente meus pais Jorgerson e Lindalva pelo amor incondicional. A minha avó Oleci, tia avó Otávia e minhas tias Geresa e Jaci pelo investimento em minha educação e qualidade de vida. A minha irmã Danielle e minha sobrinha Yale por trazer leveza aos momentos em família. A minha companheira Ana por sempre estar ao meu lado sem pestanejar.

Ao meu orientador, Prof. Dr. Ricardo Prudêncio pela oportunidade e conhecimento compartilhado. Ao meu coorientador Prof. Dr. André Câmara, Dr. Matthew J. Suderman e Dr. Wilson Júnior pela disponibilidade e pelos conhecimentos passados. Aos meus antigos chefes Dr. Bosco e Marcel Caraciolo pelos ensinamentos e investimentos.

Aos amigos que durante todo esse processo fizeram parte dos momentos bons e mais difíceis do curso, em especial Wilder, Lucas, Dyego, Marcelo, Leonardo, Tiago, Rafael, Tony, Geórgia, Rodrigo, vocês foram minha terapia.

A Universidade Federal de Pernambuco e aos membros do corpo docente do centro de Informática pelos ensinamentos durante o curso. A banca examinadora, que se dispôs a colaborar com a melhoria do presente trabalho.

A Lyra, Pingo e Lilly, por serem mais que apenas cães.

RESUMO

O câncer de mama é um dos tipos de câncer mais prevalentes em mulheres. Com uma ênfase crescente na pesquisa relacionada ao câncer muitos trabalhos foram realizados para classificar o diagnóstico, a sobrevivência e a recorrência desse tipo de câncer. Como um dos biomarcadores epigenéticos mais bem descritos em cânceres humanos, a metilação do DNA desempenha um papel essencial na regulação da expressão gênica, além de estar implicada no prognóstico e na terapêutica de muitos tipos de câncer. Nesse estudo exploramos os perfis de metilação de DNA capturando grupos de pacientes com câncer de mama que possuem diferentes marcadores e possível diversidade fenotípica visando melhorar o prognóstico desses pacientes no nível epigenético. Perfis de metilação do DNA de um estudo caso-controle com 235 indivíduos com câncer de mama e 424 indivíduos com diagnóstico negativo para câncer foram obtidos do banco de dados *Gene Expression Omnibus* (GEO). Os marcadores diferencialmente metilados foram identificados por teste T moderado. Novas características foram adicionadas aos dados usando ferramentas de predição de idade biológica, tipos celular e condição de fumante. Subgrupos de pacientes foram encontrados usando as ferramentas de mineração de subgrupos *Beam search* e SSDP+. Um total de 20 subgrupos foram selecionados, onde 10 subgrupos foram encontrados pelo algoritmo *Beam search* e os outros 10 subgrupos pelo algoritmo SSDP+. O algoritmo *Beam search* encontrou subgrupos com mais elementos e com alto grau de redundância englobando diferentes genes relacionados previamente ao câncer de mama. Já o algoritmo SSDP+ encontrou subgrupos com menos elementos e alto grau de diversidade, mas que por englobar menos regiões não tiveram genes previamente relacionados a câncer de mama. Nosso estudo identificou múltiplos painéis de metilação de DNA associados a grupos específicos de pacientes com câncer de mama que podem, se validados clinicamente, ser úteis para a previsão do risco de câncer de mama no futuro.

Palavras-chaves: câncer de mama; metilação de DNA; sangue periférico; mineração de subgrupos; painel de risco.

ABSTRACT

Breast cancer is one of the most prevalent types of cancer in women, with an increasing emphasis on research related to cancer, much work has been done to classify the diagnosis, survival or recurrence of this type of cancer. As one of the most described epigenetic biomarkers in human cancers, DNA methylation plays an essential role in the regulation of gene expression and has been implicated in the prognosis and therapeutics of many cancers. In this study we explored DNA methylation profiles capturing heterogeneous groups of breast cancer patients to improve the prognosis of these patients at the epigenetic level. DNA methylation profiles from a case-control study with 235 breast cancer subjects and 424 healthy subjects negative for cancer were obtained from the Gene Expression Omnibus (GEO) database. Differently methylated markers were identified by moderate T test. New features were added to the data using predictive tools for biological age, cell types and smoking status. Patient subgroups were found using the Beam search and SSDP+ subgroup mining tools. A total of 20 subgroups were selected, 10 of which were found by Beam search and 10 by SSDP+. Where the Beam search algorithm found larger subgroups with a high degree of redundancy encompassing different genes previously related to breast cancer, while SSDP+ found smaller subgroups with a high degree of diversity, but which, by encompassing fewer regions, did not have genes previously related to breast cancer. Our study has identified multiple DNA methylation panels associated with specific groups of breast cancer patients that could, if clinically validated, be useful for predicting future breast cancer risk.

Keywords: breast cancer; DNA methylation; peripheral blood; subgroup discovery; risk panel.

LISTA DE FIGURAS

- Figura 1 – Dogma central da biologia molecular, onde o DNA é transcrito para RNA e o RNA é traduzido para proteínas. 19
- Figura 2 – Mecanismos epigenéticos: A) O nucleossomo é composto pela hélice de DNA enrolado a histona, onde na cauda da histona pode haver modificações como metilação, acetilação ou fosforilação; B) A metilação de DNA é a adição de um grupo metil (metilado) ou não (não metilado) a um nucleotídeo citosina da sequência de DNA; C) Moléculas de RNA podem estar na forma de microRNA e afetar a expressão gênica se ligando ao DNA ou afetando o RNA mensageiro de várias formas. 19
- Figura 3 – Fluxo dos processos feitos: A) Download dos dados GSE1032 no GEO do NCBI e posteriores processos de controle de qualidade, filtragem e análise diferencial das metilações; B) Adição de novas anotações (predição da condição de fumante, predição das composição celular, predição da idade biológica); C) Processo de categorização dos dados para servir de entrada para os algoritmos *Beam Search* e SSDP+; D) Testes para descoberta dos GO termos e vias metabólicas do KEGG relacionadas aos subgrupos encontrados. 30
- Figura 4 – Fluxo dos processos antes da análise diferencial das metilações: pre-filtro de qualidade das amostras, normalização dos dados e pós-filtros de qualidade das sondas. 33
- Figura 5 – Distribuição dos valores encontrados em todas as colunas disponíveis. Onde A temos apenas os pacientes com diagnóstico negativo para câncer, enquanto em B apenas os pacientes com câncer. 37
- Figura 6 – Gráfico de barras da predição do status de fumante para todos os pacientes. Onde A temos apenas os pacientes com diagnóstico negativo para câncer, enquanto em B apenas os pacientes com câncer. 38
- Figura 7 – Correlação entre as colunas de características disponíveis. É possível perceber uma alta correlação negativa entre a coluna "time of diagnosis" e a classe alvo ("cancer type classes"). 39

Figura 8 – Diagrama de Venn com as CpGs agrupadas nos subgrupos encontrados pelo algoritmo Beam Search.	43
Figura 9 – Diagrama de Venn com as CpGs agrupadas nos subgrupos encontrados pelo algoritmo SSDP+ usando a métrica de qualidade Qg.	46
Figura 10 – Diagrama de Venn com as CpGs agrupadas nos subgrupos encontrados pelo algoritmo SSDP+ usando a métrica de qualidade WRAcc.	46
Figura 11 – Distribuição das características para todos os pacientes englobados nos subgrupos encontrados pelo Beam search.	47
Figura 12 – Gráfico de barras da predição do status de fumante para todos os pacientes englobados nos subgrupos encontrados pelo Beam search.	47
Figura 13 – Distribuição das anotações relacionadas a composição celular para os subgrupos encontrados usando o SSDP+.	48
Figura 14 – Distribuição das anotações relacionadas a idade e tempo para diagnóstico para os subgrupos encontrados usando o SSDP+ com a métrica de qualidade Qg.	49
Figura 15 – Gráfico de barras da predição do status de fumante para os subgrupos encontrados usando o SSDP+ com a métrica de qualidade Qg.	49

LISTA DE TABELAS

Tabela 1 – Estatística descritiva das colunas com valores numéricos.	36
Tabela 2 – Número de classes para cada coluna dos dados que serviram de entrada para os algoritmos de descoberta de subgrupos.	41
Tabela 3 – Tabela com os 4 primeiros subgrupos encontrados pelo algoritmo Beam Search e suas respectivas classes.	44
Tabela 4 – Tabela com os 10 subgrupos encontrados pelo algoritmo SSDP+ para a métrica de qualidade Qg.	45
Tabela 5 – Tabela com as anotações das CpGs englobadas nos subgrupos encontrados tanto pelo Beam search quanto pelo SSDP+ para a métrica Qg.	50
Tabela 6 – Tabela com os 10 primeiros termos de ontologia genética para os 10 subgrupos encontrados pelo algoritmo Beam search.	51
Tabela 7 – Tabela com os 10 primeiros termos de ontologia genética para os 10 subgrupos encontrados pelo algoritmo SSDP+ para a métrica de qualidade Qg.	52
Tabela 8 – Tabela com as 10 primeiras vias metabólicas para os 10 subgrupos encontrados pelo algoritmo Beam search.	53
Tabela 9 – Tabela com as 10 primeiras vias metabólicas para os 10 subgrupos encontrados pelo algoritmo SSDP+ para a métrica de qualidade Qg.	54
Tabela 10 – Tabela de genes hipermetilados presentes nas regiões de metilação dos subgrupos encontrados pelo algoritmo <i>Beam search</i>	56
Tabela 11 – Tabela de genes hipometilados presentes nas regiões de metilação dos subgrupos encontrados pelo algoritmo <i>Beam search</i>	57
Tabela 12 – Tabela de genes hiper-metilados e hipo-metilados presentes nas regiões de metilação dos subgrupos encontrados pelo algoritmo <i>SSDP+</i>	60

LISTA DE ABREVIATURAS E SIGLAS

CD4	Cluster of Differentiation 4
CD8	Cluster of Differentiation 8
CGI	CpG Islands
CpG	Sítio citosina-fosfato-guanina
DMRs	Differentially Methylated Regions
DNA	Deoxyribonucleic Acid
DPs	Descriminative patterns
EPIC	European Prospective Investigation into Cancer and Nutrition
FDR	False Discovery Rate
GEO	Gene Expression Omnibus
GO	Gene Ontology
IDAT	Intensity Data files
KEGG	Kyoto Encyclopedia of Genes and Genomes
MBD	Mamografic Breast Density
NK	Natural Killer
RefSeq	Reference Sequence Database NCBI
RNA	Ribonucleic acid
SNPs	Single Nucleotide Polymorphisms
SSDP+	Simple Search Discriminative Patterns Plus
T2D	Type 2 Diabetes
TCGA	The Cancer Genome Atlas

LISTA DE SÍMBOLOS

$\in$ Pertence

$\rightarrow$ Implica

SUMÁRIO

1	INTRODUÇÃO	14
1.1	OBJETIVOS	16
1.2	ORGANIZAÇÃO	16
2	REVISÃO DA LITERATURA	18
2.1	METILAÇÃO DE DNA	18
2.2	CÂNCER DE MAMA	22
2.3	DESCOBERTA DE SUBGRUPOS	24
2.3.1	Beam search	26
2.3.2	SSDP+	27
3	MATERIAL E MÉTODOS	29
3.1	DADOS	30
3.2	ANÁLISE DIFERENCIAL DAS METILAÇÕES	31
3.3	ENRIQUECIMENTO DE DADOS FENOTÍPICOS	33
3.3.1	Metadados dos pacientes	34
3.3.2	Predição da condição de fumante	34
3.3.3	Composição celular	34
3.3.4	Estimando idade biológica	34
3.3.5	Organizando os dados	35
3.4	CATEGORIZAÇÃO DOS DADOS	40
3.5	BEAM SEARCH	41
3.6	SSDP+	41
3.7	ANOTAÇÃO DOS SUBGRUPOS	42
4	DESCOBERTA DOS SUBGRUPOS	43
5	DISCUSSÃO	55
5.1	SUBGRUPOS ENCONTRADOS PELO BEAM SEARCH	55
5.2	SUBGRUPOS ENCONTRADOS PELO SSDP+	59
6	CONCLUSÃO	62
6.1	CONTRIBUIÇÕES	63
6.2	LIMITAÇÕES E TRABALHOS FUTUROS	63
	REFERÊNCIAS	64

1 INTRODUÇÃO

Considerado um dos marcadores epigenéticos mais bem descritos para câncer em humanos, a metilação de Deoxyribonucleic Acid (DNA) tem um papel essencial na regulação da expressão de genes com implicação no prognóstico e terapêutica de vários tipos de doenças complexas como diabetes, esquizofrenia, doenças cardiovasculares e vários tipos de câncer. Para os mamíferos, a metilação do DNA é um mecanismo epigenético que envolve a transferência de um grupo metil para a posição C5 da citosina para formar 5-metil citosina. Esse mecanismo regula a expressão gênica recrutando proteínas envolvidas na repressão gênica ou inibindo a ligação do(s) fator(es) de transcrição ao DNA. A metilação de DNA é conhecida por variar entre tipos de células em certos tecidos, essas variações de metilação têm sido estudadas por ter um papel importante na gênese de tumores e na progressão de tipos de câncer. Entre as plataformas de perfil de metilação de DNA em humanos, o “Illumina Infinium HumanMethylation450 Bead-Chip” (450K) tem sido aceito como uma das tecnologias mais eficientes cobrindo 480K sítios de nucleotídeo único Sítio citosina-fosfato-guanina (CpG) (metilação), incluindo 98% das ilhas de CpG no genoma humano.

Existem diversos trabalhos em ciências biomédicas usando populações para encontrar grupos de pacientes que estejam relacionados a determinada doença. Esses estudos são normalmente organizados por uma hipótese prévia de que determinado grupo com certas características podem estar relacionados a doença sob investigação ou a um tratamento. Muitas vezes essas hipóteses não são estatisticamente significativas, sendo necessário recomeçar um novo estudo a partir de novos dados referentes a outras populações com diferentes características mas ainda com a incerteza de haver uma relação significativa com a hipótese. Nos estudos com câncer é possível ver frequentemente uma distribuição heterogênea dos pacientes para uma mesma doença em relação aos sintomas ou mesmo em relação ao tratamento, o que aumenta a quantidade de hipóteses a serem estudadas. A maior parte literatura sobre metilações de DNA relacionados a tipos de câncer usam algoritmos de clusterização para agrupar pacientes e assim identificar grupos heterogêneos seguidos de um modelo de classificação para identificar esses grupos (CAPPER et al., 2018; ONWUKA et al., 2020; HUANG et al., 2020; WOCKNER et al., 2014; ZHANG et al., 2018a; CHEN; ZHANG; DAI, 2019). Regras de associação em mineração de dados são usadas para descobrir regras interessantes que se associam a determinados itens dentro de um grupo de dados. Enquanto que mineração de subgrupos engloba um conjunto

de modelos e algoritmos situados entre os modelos de predição e as regras de associação, não sendo usada para predição. Dessa forma, tem por objetivo trazer resultados em forma de regras que acabam sendo mais fáceis de serem interpretadas pelo usuário e ter uma descrição compacta dos padrões mais interessantes dos dados.

Seguindo uma tendência mundial, o câncer de mama é um dos tipos de câncer mais prevalentes em mulheres. Com o aumento da prevalência desse tipo de câncer entre as doenças não transmissíveis, ele acabou virando um problema de saúde pública para o sistema de saúde brasileiro e isso exige intervenções vinculadas a tecnologias leves, leve-duras e duras, junto com mudanças no estilo de vida dentro de um projeto terapêutico de cuidado contínuo. A possibilidade de cura existe quando o diagnóstico dessa doença é feito precocemente, porém o estudo de Brito-Silva et al. aponta que no Brasil os dados de mortalidade são elevados graças a um diagnóstico tardio da doença por vários motivos, sendo alguns deles a dificuldade de acesso a serviços de saúde, carência de serviços oncológicos, falha no processo de capacitação oncológica para profissionais e outros (BRITO-SILVA et al., 2014). Por isso, os dados escolhidos para esse trabalho são dados públicos de pacientes com câncer de mama, dado que o sequenciamento foi feito antes dos pacientes serem diagnosticados com a doença, possibilitando a descoberta de biomarcadores que indicam o câncer antes mesmo do paciente ser diagnosticado.

Já foram propostos diversos biomarcadores para câncer de mama, mas um diagnóstico precoce desses pacientes ainda é um desafio (VAINIO; BIANCHINI et al., 2002; SAWYERS, 2008). Além disso, os métodos de imagem são restritos ao tamanho e volume do tecido tumoral para a análise (BONINI; MONDINO, 2015). A detecção de câncer de mama hoje depende de métodos invasivos como a biópsia de tecido tumoral (SAGIV-BARFI et al., 2015). Assim, a detecção desse tipo de câncer nos estágios iniciais em que o tumor não está presente poderia beneficiar o diagnóstico e prognóstico desses pacientes. Um recente estudo de Caini et al. que investigou padrões de metilações de DNA entre mulheres com alta e baixa densidade mamográfica (Mamografic Breast Density (MBD)) que desenvolveram câncer de mama usando amostras de sangue antes do diagnóstico de câncer, não foi encontrada nenhuma diferença significativa para CpGs únicas, mas encontraram regiões diferencialmente metiladas (Differentially Methylated Regions (DMRs)) em ambos os grupos de MBD com câncer de mama com genes envolvidos na regulação da transcrição do DNA e na apoptose celular, o que sugere que os dois grupos diferem de alguma forma quanto ao mecanismo de carcinogêneses (CAINI et al., 2021). Existem poucos trabalhos usando algoritmos de mineração de subgrupos no contexto de dados de alta dimensionalidade e até o momento nenhum estudo aplicado a

dados de metilação que possuem uma alta dimensionalidade. Assim, a aplicação de mineração de subgrupos em dados de metilação parece ser uma abordagem interessante, facilitando a descoberta de populações para estudos de associação baseados em hipóteses prévias. No nosso trabalho foram utilizados dados públicos de metilação de pacientes com câncer de mama com o intuito de encontrar grupos de pacientes com características específicas que possam facilitar o diagnóstico e prognóstico precoce de pacientes no futuro.

1.1 OBJETIVOS

Como objetivos do nosso trabalho temos:

- Utilizar dados de metilação de DNA de alta dimensionalidade para pacientes com câncer de mama e pacientes com diagnóstico negativo para câncer.
- Automatizar esse processo de a partir de um grupo de pacientes encontrar subgrupos com características específicas que estejam intrinsecamente relacionados a uma doença, no nosso caso o câncer de mama.
- Explorar algoritmos de descoberta de subgrupos em dados de alta dimensionalidade no domínio biológico.
- Validar a metodologia proposta para detectar subgrupos estatisticamente significantes relacionados ao câncer de mama.

1.2 ORGANIZAÇÃO

As próximas sessões estão organizadas da seguinte maneira.

- No Capítulo 2, é feita uma revisão da literatura falando sobre metilação de DNA, câncer de mama e descoberta de subgrupos.
- No capítulo 3 são apresentados os dados e as metodologias aplicadas desde a análise das metilações, análise dos dados de metilações, anotação de novas informações até a utilização dos algoritmos de descoberta de subgrupos.
- No capítulo 4 são apresentados os resultados obtidos no processo de descoberta dos subgrupos.

- No capítulo 5 é discutida a importância dos subgrupos encontrados e os pontos fortes do trabalho.
- No capítulo 6 o trabalho é finalizado com um resumo dos objetivos alcançados e discutindo as limitações e trabalhos futuros.

2 REVISÃO DA LITERATURA

Nesse capítulo é feita uma revisão sobre os assuntos necessários para o entendimento do projeto, dividindo-se da seguinte forma:

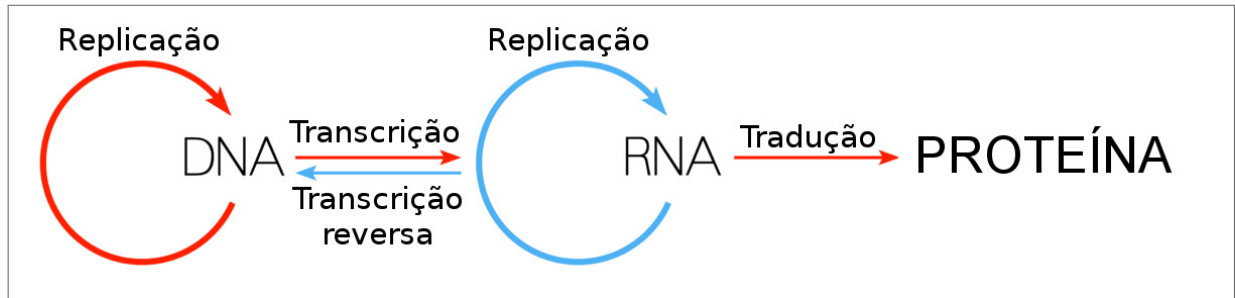
- No subcapítulo 2.1 são introduzidos os conceitos relacionados ao dogma central da biologia molecular, epigenética, metilação de DNA e feita uma revisão sobre os estudos relacionados a metilação de DNA.
- O subcapítulo 2.2 introduz a ideia de câncer, descreve estatísticas dessa doença no mundo e no Brasil, descreve o câncer de mama e suas estatísticas, especifica a importância de se trabalhar com câncer de mama e cita alguns estudos com metilação em pacientes com câncer de mama.
- No subcapítulo 2.3 introduz o conceito de descoberta ou mineração de subgrupos, descreve os dois algoritmos utilizados no projeto e suas métricas de avaliação.

2.1 METILAÇÃO DE DNA

Todas as células do nosso corpo possuem o mesmo código genético em forma de ácido desoxirribonucleico (DNA). Essa informação genética é formada por sequências de nucleotídeos: Adenina (A), Citosina (C), Timina (T) e Guanina (G). Os nossos genes são resultado de várias combinações de nucleotídeos que podem dar origem a grupos de proteínas que são específicas para um tipo de célula e seu funcionamento. De acordo com o dogma central da biologia molecular (Figura 1), genes que são expressos dando origem a proteínas são primeiramente transcritos em ácido ribonucleico (Ribonucleic acid (RNA)) para então serem traduzidos em proteínas (FARKAS, 2014).

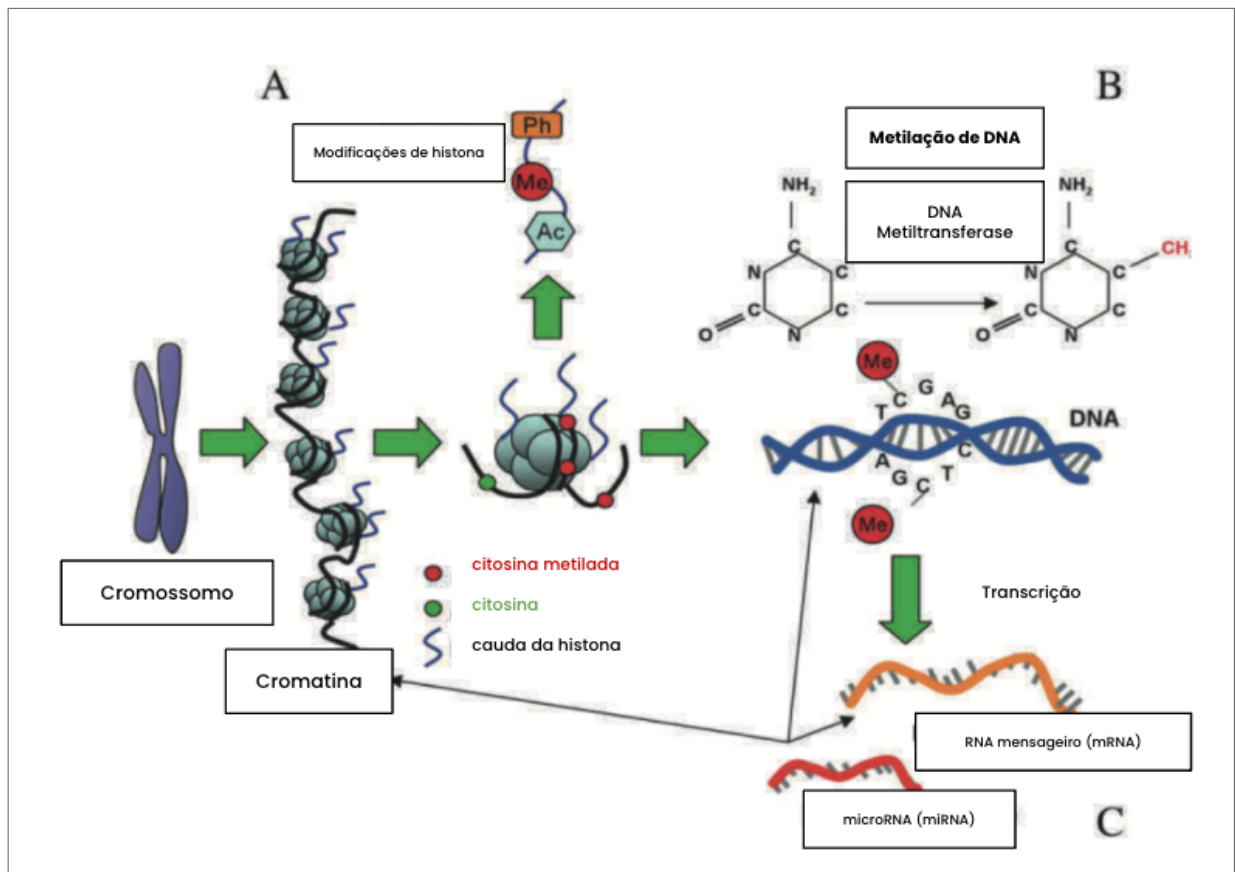
A epigenética é o estudo de como o nosso comportamento e o ambiente em que vivemos podem causar mudanças que afetam o comportamento dos nossos genes. Diferente das alterações genéticas, as alterações epigenéticas podem ser reversíveis e não alteram a sequência de DNA, porém elas podem mudar a forma como o nosso corpo lê essa sequência ((OS); HEALTH, 2020). Os mecanismos epigenéticos podem ser divididos em três grandes grupos: metilação de DNA, modificações de histona e RNAs não-codificantes (Figura 2) (FARKAS, 2014).

Figura 1 – Dogma central da biologia molecular, onde o DNA é transcrito para RNA e o RNA é traduzido para proteínas.



Fonte: (PHILIP, 2006)

Figura 2 – Mecanismos epigenéticos: A) O nucleossomo é composto pela hélice de DNA enrolado a histona, onde na cauda da histona pode haver modificações como metilação, acetilação ou fosforilação; B) A metilação de DNA é a adição de um grupo metil (metilado) ou não (não metilado) a um nucleotídeo citosina da sequência de DNA; C) Moléculas de RNA podem estar na forma de microRNA e afetar a expressão gênica se ligando ao DNA ou afetando o RNA mensageiro de várias formas.



Fonte: (FARKAS, 2014)

A metilação de DNA (Figura 2A) é a adição de um radical metil (CH_3) por meio de uma ligação covalente a um nucleotídeo que tipicamente é uma citosina seguida por uma guanina, por isso conhecidas como sítios CpG (citosina-fosfato-guanina). Regiões metiladas do DNA (hiper metilação) estão associadas a uma cromatina condensada que quando perto de regiões

promotoras causa a repressão da transcrição. Já as regiões não metiladas (hipo metilação) estão associadas com regiões de cromatina aberta que permitem a transcrição do gene (LEVY et al., 2020). Regiões no DNA que são ricas em sítios de CpG são chamadas de ilhas de CpG (do inglês *CpG islands*, CGI). No genoma humano temos aproximadamente 25.000 CpG Islands (CGI)s onde cerca de 60% delas estão associadas a regiões promotoras de vários genes. As regiões promotoras são sequências de DNA que podem se ligar a proteínas (conhecidas como fatores de transcrição) que podem ativar ou desativar a expressão de um determinado gene (FARKAS, 2014).

Desde o momento do nascimento, os níveis de metilação do genoma de mamíferos podem ser modificados pela exposição ao ambiente, alterando a expressão de genes relacionados a diferentes funções biológicas do corpo (STARNAWSKA; DEMONTIS, 2021). Padrões de metilação de DNA estão associados à expressão de tipos de células específicas. Alterações nesses padrões de metilação têm sido associadas ao envelhecimento e a exposição ao ambiente. Assim, já está bem estabelecido que alterações do padrão de metilação do DNA podem contribuir para progressão de vários tipos de câncer e doenças complexas como diabetes, esquizofrenia e doenças cardiovasculares (LEVY et al., 2020). No trabalho de Zhang et al. estudando pacientes com carcinoma hepatocelular, foi percebido que esses pacientes possuem assinaturas de metilação de DNA que intensificam a progressão do câncer conseguindo distinguir os tipos de câncer. Também foi percebido que essas diferenças ocorrem nos primeiros estágios do câncer, podendo ser utilizado como uma abordagem não invasiva na detecção da doença (ZHANG et al., 2018b). Em um estudo feito por Joo et al., foi possível identificar 24 marcadores de metilação de DNA que foram herdados dentro das famílias com câncer de mama estudadas, o que pode reduzir a quantidade de biomarcadores estudados para o risco de câncer (JOO et al., 2018). A heterogeneidade tumoral tem sido um problema para o diagnóstico e terapia de pacientes com câncer de mama. Zhang et al. em seu trabalho identificaram nove diferentes subtipos de câncer de mama usando dados abertos do The Cancer Genome Atlas (TCGA) (The Cancer Genome Atlas) que foram considerados distintos ao nível molecular e epidemiológico. As metilações de DNA que diferenciam esses subtipos podem ser usadas como biomarcadores para diagnóstico e tratamentos personalizados de pacientes com esses subtipos de câncer de mama (ZHANG et al., 2018a).

Mudanças no perfil de metilação também têm sido estudadas em populações com transtornos mentais. Em um estudo com pacientes diagnosticados com esquizofrenia, Jaffe et al. identificaram cerca de 2.000 CpGs diferencialmente expressas entre seus pacientes e pacientes

controle e perceberam que a maioria dos genes estavam relacionados ao desenvolvimento e diferenciação neural, onde houve uma alta correlação entre essas CpGs e a transição do período pré-natal e pós-natal (JAFJE et al., 2016). Em outro estudo com pacientes com esquizofrenia, Wockner et al. identificaram CpGs diferencialmente metiladas que serviram como diferenciação entre pacientes com diagnóstico negativo para esquizofrenia e com a doença. Apesar das limitações do trabalho, como os pacientes foram submetidos a drogas que podem influenciar os níveis de metilação do DNA e não terem levado em consideração os tipos celulares, os resultados demonstram o potencial uso dos perfis de metilação como indicadores de prognóstico para pacientes com esquizofrenia (WOCKNER et al., 2014).

Examinando o impacto de mães com depressão em seus filhos, Nemoda et al. exploraram se esse impacto influencia a metilação de DNA dos filhos. Apesar das limitações do estudo, como o tamanho da amostragem, variáveis geográficas como uso de tabaco durante a gestação que afetam os sítios de metilação diferenciando grupos de amostra, foi possível identificar quase 2.000 genes diferencialmente metilados, mas apenas 139 foram associados a depressão maternal. Com os dados de suporte que eles tinham de amostras de cérebro dos pacientes já adultos, eles viram que a depressão materna tem um efeito no perfil de metilação de seus filhos, podendo então persistir na vida adulta (NEMODA et al., 2015). Utilizando um grupo de pacientes com diferentes tipos de câncer de cérebro, Capper et al. criaram um modelo para classificar os diferentes tipos de câncer e para isso utilizaram dados de metilação de DNA. Com a criação de um banco de dados interno com amostras referentes aos tipos de tumores encontrados, o grupo espera detectar com maior facilidade tipos de tumores raros e refinar esse classificador. Assim, usar perfis de metilação como parte da classificação histo molecular poderá aumentar a acurácia dos diagnósticos e servirá como referência para outros tipos de tumores (CAPPER et al., 2018).

Uma revisão feita por Willmer et al. usando estudos com diferentes grupos de populações de pacientes com diabetes tipo 2 (Type 2 Diabetes (T2D)) identificou 8 sítios de CpG diferencialmente expressos que estão potencialmente associados com T2D. Porém eles indicaram limitações como os dados serem de coleta de sangue que não é um tecido de resposta à insulina e a necessidade de estudos longitudinais para comprovar essa relação (WILLMER et al., 2018). Huang et al. identificaram 13 novos sítios de CpG associados com pressão sanguínea, replicando os resultados em outros dados. Apesar da limitação de não ter sido feito nenhum estudo *in vitro* ou *in vivo* para confirmar o impacto desses novos sítios na expressão de genes e subsequentemente na pressão sanguínea de pacientes, o estudo serve como embasamento para

novos estudos de correlação futuros e dá uma visão sobre as origens genéticas e ambientais da relação entre metilação de DNA e pressão sanguínea (HUANG et al., 2020).

O sequenciamento de genoma completo por dissulfeto é o padrão ouro para a quantificação de metilações e não-metilações de sítios CpG. O repositório público Gene Expression Omnibus (GEO) (EDGAR; DOMRACHEV; LASH, 2002) possui mais de 100 mil amostras de metilação de DNA humano. A maioria desses dados foi mensurado usando o kit da Illumina Infinium HumanMethylation450 BeadChip (BIBIKOVA et al., 2011). Esse kit possui 485.577 probes contemplando 99% dos genes presentes no Reference Sequence Database NCBI (RefSeq), além de outras áreas do genoma (como CpG Open Seas) e sítios de CpG em regiões de densidade variável (como ilhas de CpG, Shelves e Shores), contemplando 96% das ilhas de CpG (SALA et al., 2020). Esse kit usa duas metodologias químicas (infinium 1 e infinium 2). As duas são baseadas na quantificação da genotipagem do polimorfismo C/T gerado pela conversão do dissulfeto no DNA (DEDEURWAERDER et al., 2014). Sendo uma tecnologia popular para a exploração de metilação de DNA em pessoas saudáveis e pacientes com doenças, existem diversos passos para o processamento desses dados, incluindo controle de qualidade, filtragem, normalização e testes estatísticos para descoberta de probes diferencialmente metiladas (MAKSIMOVIC; PHIPSON; OSHLACK, 2016).

2.2 CÂNCER DE MAMA

O câncer de mama é uma neoplasia maligna causada por alterações genéticas hereditárias ou por influências ambientais causando um crescimento anormal das células mamárias, sendo o câncer mais comum e o quinto câncer mais mortal mundialmente, enquanto que no Brasil é o que possui mais novos casos e o que mais mata mulheres (MACHADO; SOARES; OLIVEIRA, 2017; COSTA et al., 2020). De acordo com o Instituto do Câncer, estima-se cerca de 66 mil novos casos desse tipo de câncer por ano (BRASIL, 2020). Os fatores de riscos para o desencadeamento da doença são o histórico familiar, genético, idade elevada, menarca precoce, menopausa tardia, nuliparidade, sedentarismo, obesidade, presença de vícios, influências ambientais e outros (SILVA; RIUL, 2011). A possibilidade de cura aumenta com o diagnóstico precoce da doença, e para isso existem dois grupos de medidas preventivas ao câncer de mama, as medidas primárias e secundárias. Entre as primárias podemos citar um estilo de vida saudável, isso inclui uma alimentação balanceada, prática de exercícios físicos regulares, evitar o uso de substâncias como álcool e tabaco. Dentro das secundárias, temos fatores mais específicos com

autoexame, avaliações periódicas com especialistas e realização de exames de rotina, visando um diagnóstico precoce para uma maior chance de cura (PINHO; COUTINHO, 2007).

O estudo de Machado et al. aponta que as taxas de mortalidades de pacientes com câncer de mama no Brasil são motivadas principalmente pelo diagnóstico tardio. Alguns fatores que agravam esse processo são as dificuldades ao acesso aos serviços de saúde, carência de serviços oncológicos fora das metrópoles, falhas na capacitação oncológica de profissionais de saúde, descontinuidade do acesso desde os serviços de atenção básica até serviços mais especializados e a dificuldade dos gestores organizarem os fluxos assistenciais. Somado a isso, podem ser considerados fatores macro contextuais (elementos demográficos, econômicos e sociais) e fatores micro contextuais (crenças, relações familiares e interpessoais) (MACHADO; SOARES; OLIVEIRA, 2017). Os programas de rastreamentos e detecção precoce do câncer de mama normalmente se apoiam no auto exame das mamas, no exame clínico das mamas e na mamografia (SOUSA et al., 2020). Existe uma discussão sobre o estresse causado em pacientes pelo acompanhamento por mamografia (SCHÜNEMANN et al., 2020; LOVING; AMINOLOLAMA-SHAKERI; LEUNG, 2021). Com o objetivo de diminuir as taxas de mortalidade dessa doença, tem-se usado exames genéticos para quantificar os riscos, discutir opções terapêuticas, propor manejo clínico ou cirúrgico de forma preventiva para pacientes e seus parentes em casos de histórico familiar (SOUSA et al., 2020). Alterações epigenéticas como a metilação do DNA têm sido associadas a diversas doenças, incluindo o câncer de mama. Já foi demonstrado que, assim como variantes genéticas, alterações na metilação de DNA parecem marcar células normais pré-neoplásicas que posteriormente se transformam e podem dar origem a um câncer, sugerindo que as alterações de metilação do DNA são anteriores ao surgimento do câncer de mama (TESCHENDORFF et al., 2016).

Os padrões de metilação de DNA atuam como uma conexão entre as exposições ambientais e a expressão de genes, sendo potenciais biomarcadores para avaliar o risco do câncer de mama e permitindo assim uma abordagem personalizada com estratégias visando a redução de complicações. Um estudo com mulheres que desenvolveram câncer de mama durante um acompanhamento de aproximadamente nove anos tentou evidenciar diferenças nos padrões de metilação de DNA de acordo com índices baixos e altos da densidade da mamografia. Esse estudo não encontrou CpGs diferencialmente metiladas para índices altos ou baixos de mamografia, mas descobriu 140 regiões diferencialmente metiladas englobando 35 genes codificantes de proteína implicados na regulação da transcrição do DNA, sinalização da apoptose e ligação ao DNA. Isso mostra que padrões de metilação de DNA antes do diagnóstico diferem entre

índices alto e baixo de densidade na mamografia (CAINI et al., 2021).

O estudo de Ennour-Idrissi et al. identificou padrões de metilação em tecidos mamários de pacientes com diagnóstico negativo para câncer que pudessem predispor o desenvolvimento de câncer. Para isso eles compararam amostras com diagnóstico negativo antes do paciente ter câncer com amostras dos pacientes que não tiveram câncer. Assim, foram identificadas 7314 CpGs diferencialmente metiladas que apenas 53 CpGs passaram pela correção de *Bonferroni* ($p < 0.01$), dessas 43 compreendem genes envolvidos em doenças metabólicas. Além disso, quatro genes estavam diferencialmente expressos tanto na análise de CpGs diferencialmente expressas quanto na análise de regiões diferencialmente expressas, enquanto três outros genes estavam diferencialmente expressos para a comparação entre tecidos mamários e sangue. Esses resultados suportam a hipótese de que metilações diferencialmente expressas em genes relacionados a câncer em tecidos mamários sem alterações morfológicas acontecem antes da detecção do câncer de mama. Algumas dessas metilações detectadas no sangue podem ter sido causadas no desenvolvimento e propagadas de alguma forma, o que pode ser usado como biomarcadores para um acompanhamento não invasivo para a identificação de mulheres com um maior risco de desenvolverem câncer de mama (ENNOUR-IDRISSI et al., 2020).

2.3 DESCOBERTA DE SUBGRUPOS

A mineração de subgrupos (*subgroup mining*) foi formulada primeiramente por Klosgen e Wrobel (KLÖSGEN, 1996; WROBEL, 1997). Em seus trabalhos, o problema da identificação de subgrupos foi definido da seguinte forma: Dada uma população de indivíduos e a propriedade desses indivíduos, estamos interessados em encontrar subgrupos populacionais que são estatisticamente "mais interessantes" (ex.: são os maiores possíveis e possuem a métrica de avaliação mais significativa em respeito a característica de interesse. Diferente dos métodos de classificação e regressão o objetivo da descoberta de subgrupos (*subgroup discovery*) não é primeiramente ter um modelo preditivo, mas em vez disso ter uma representação que seja de fácil interpretação pelo usuário e disponibilizar uma descrição compacta dos padrões mais interessantes para os dados (HERRERA et al., 2011; LEMMERICH; BECKER, 2018). A mineração de subgrupos tem diversas aplicações incluindo análises espaciais e planos de campanhas de marketing (KLÖSGEN; MAY, 2002; LAVRAČ et al., 2004). Nos domínios clínicos e médicos, a descoberta de subgrupos pode ser útil no prognóstico médico, resposta a tratamentos e medicina personalizada (BAUMEISTER; ATZMUELLER; PUPPE, 2006). A mineração de subgrupos se pre-

ocupa em achar relações entre diferentes variáveis de um grupo de dados em relação a uma característica alvo. Dado que essa metodologia foca na extração de relações com características interessantes acaba não sendo necessário obter relações completas, sendo as relações parciais suficientes. Essas relações são descritas em forma de regras (R), que consistem em uma descrição do subgrupo, que formalizando seria:

$$R : Cond \rightarrow Target = Value$$

Onde $Target = Value$ é um valor da variável de interesse (também conhecido como Classe) e $Cond$ é um conjunto de atributos que são capazes de descrever uma distribuição estatisticamente não usual em relação ao $Target = Value$. Como exemplo, vejamos um grupo de dados (D) com quatro variáveis:

- $Idade \in \{Menos\ de\ 25, 25\ a\ 60, Mais\ de\ 60\}$
- $Sexo \in \{H, M\}$
- $País \in \{Brasil, Emirados\ Árabes\ Unidos, Canadá, Argentina, Portugal, Espanha\}$
- $Dinheiro \in \{Pobre, Normal, Rico\}$

Algumas das possíveis regras contendo subgrupos são:

- $R_1 : (Idade = Mais\ de\ 60\ AND\ País = Emirados\ Árabes\ Unidos) \rightarrow Dinheiro = Rico$
- $R_2 : (Idade = 25\ a\ 60\ AND\ Sexo = M) \rightarrow Dinheiro = Normal$

A regra R_1 representa um subgrupo de pessoas dos Emirados Árabes Unidos com mais de 60 que possuem uma probabilidade incomum de serem ricos quando comparados ao resto da população. Enquanto na regra R_2 representa um subgrupo de mulheres com idade entre 25 e 60 anos que são mais prováveis de ter uma economia normal do que o resto da população.

A descoberta de subgrupos acaba se diferenciando das técnicas de classificação basicamente porque tem o objetivo de descrever o conhecimento dos dados de forma simples e interpretável, enquanto um classificador tenta prever esse conhecimento com precisão e complexidade (HERRERA et al., 2011). Analisar manualmente todos os subgrupos para um problema acaba sendo inviável, assim já foram propostos algoritmos para extrair os subgrupos considerados

mais relevantes para o problema e posteriormente serem selecionados por um especialista no domínio do problema. Entre as heurísticas de mineração de subgrupos tem se destacado as baseadas em pesquisa de feixe (*Beam search*) e as de computação evolucionária (GAMBERGER; LAVRAC, 2002; LEEUWEN; KNOBBE, 2012; MARTÍN et al., 2016; LUCAS; VIMIEIRO; LUDERMIR, 2018).

2.3.1 Beam search

Beam search é um método simples muitas vezes usado para evitar pesquisas exaustivas, considerada uma boa abordagem para exploração de um espaço de hipóteses. Supondo que temos um grafo (G) que queremos percorrer até chegar em um nó específico, começamos com o nó raiz. O primeiro passo é expandir o nó inicial com todos os nós possíveis. Então, a cada passo do algoritmo escolhemos um número específico de nós dentro dos possíveis para expandir, esse número de nós é a largura do feixe (beam, w) que é um parâmetro de entrada dado pelo usuário. Os nós são selecionados de acordo com um custo heurístico e a expansão dos nós continua até atingir o nó objetivo. Ao selecionar um pequeno valor para w o algoritmo acaba ignorando alguns caminhos que poderiam ter sido mais promissores, assim valores altos de w podem ajudar na pesquisa de soluções melhores, porém terá um custo computacional alto (IMPARATO, 2012; KARMA, 2021).

A definição de relevância de um subgrupo é dada por métricas de avaliação, sendo as uma das métricas disponíveis a *WRAcc* (*weighted relative accuracy*).

$$WRAcc(s) = \frac{TP + FP}{|D|} \left(\frac{TP}{TP + FP} - \frac{|D^+|}{|D|} \right)$$

Onde TP é o suporte positivo ($TP = |c^+(s)|$), FP o suporte negativo ($FP = |c^-(s)|$), D todo o dado de entrada. Essa métrica acaba sendo um balanceamento entre cobertura ($\frac{TP+FP}{|D|}$) e acurácia relativa ($\frac{TP}{TP+FP} - \frac{|D^+|}{|D|}$). O valor resultante varia entre $-0,25$ e $+0,25$, sendo $+0,25$ subgrupos totalmente puros. Outra métrica disponível é o *lift* que é utilizado para definir o grau de interesse de uma regra de associação. Sendo a razão entre o suporte observado e o esperado se X e Y fossem independentes.

$$lift(X \Rightarrow Y) = \frac{supp(X \cup Y)}{supp(X) * supp(Y)}$$

Onde um valor de lift igual a 1 indica que a probabilidade da ocorrência do antecedente e a do consequente são independentes uma da outra, não podendo ser traçada nenhuma regra envolvendo os dois eventos. Enquanto maior que 1 indica que as duas ocorrências são dependentes uma da outra, tornando essas regras potencialmente úteis para prever o consequente (Y) em conjuntos de dados futuros. Já valores de lift menor que 1 significa que a presença de um item tem efeito negativo na presença do outro item e vice-versa.

2.3.2 SSDP+

A ferramenta Simple Search Discriminative Patterns Plus (SSDP+) usa métodos evolucionários e *Beam search* para achar top-k padrões discriminativos (*Discriminative patterns*, s) de forma adaptada para dados de alta dimensionalidade, onde k é escolhido pelo usuário. O algoritmo genético usado pesquisa padrões "frequentes-diversos" (*Diverse-frequent patterns*) em grandes dados, de forma que inclui para a próxima geração os melhores indivíduos de uma população antiga P_{old} e outros criados por operadores genéticos ($P_c \leftarrow crossOver(P_{old})$ e $P_m \leftarrow mutation(P_{old})$), sendo o tamanho das populações iguais ($|P_{old}| = |P_c| = |P_m|$). Além disso, o algoritmo usa dois conceitos advindos do algoritmo *Beam search*. O primeiro é inicializar a pesquisa usando todos os s de uma dimensão, isso garante todos os itens na pesquisa. O segundo é que as pesquisas na dimensão d são feitas pelos melhores Discriminative patterns (DPs) menores que d , então no SSDP+ a população inicial é formada por todos os DPs de uma dimensão e os operadores genéticos são responsáveis por expandir para mais dimensões. SSDP+ salva os melhores top-k subgrupos avaliados que são distintos e não-dominantes, sendo um subgrupo s considerado dominado por outro subgrupo s' se $c^+(s) \subseteq c^+(s')$ e $c^-(s) \subseteq c^-(s')$, onde c^+ e c^- são os exemplos positivos e negativos no subgrupo respectivamente. Parte dos subgrupos considerados redundantes são salvos no cache para minimizar o risco de descartar informações relevantes (PONTES; VIMIEIRO; LUDERMIR, 2016; LUCAS; VIMIEIRO; LUDERMIR, 2018).

Além de usar a métrica de avaliação $WRAcc$ existe a opção de usar a outra métrica disponível Qg .

$$Qg(s) = \frac{TP}{FP + g}$$

Onde g é um parâmetro de generalização, sendo que valores altos de g retornam normal-

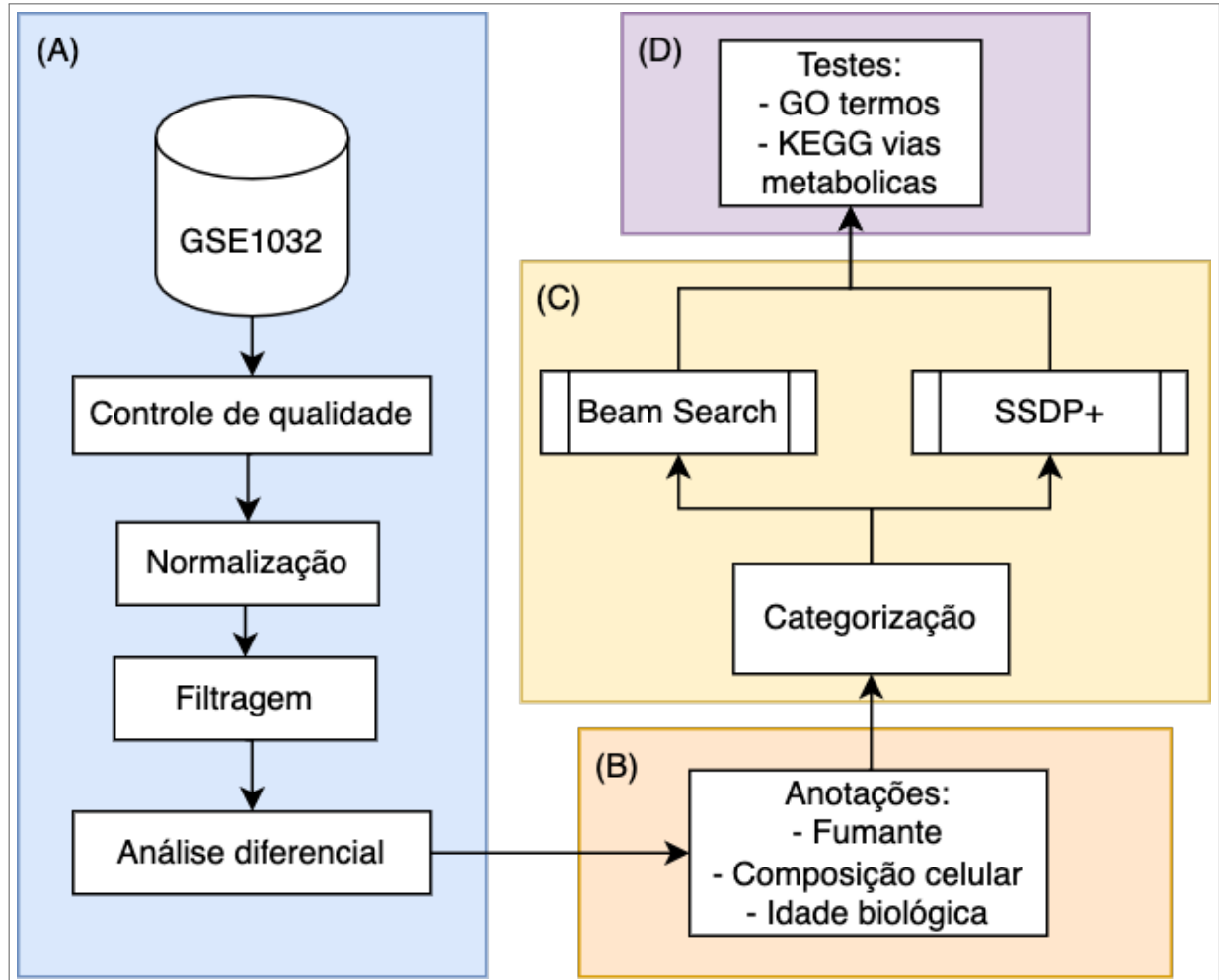
mente subgrupos mais gerais com menos precisão e baixos valores de g retornam subgrupos específicos com alta precisão. O valor padrão de g igual a 1.

3 MATERIAL E MÉTODOS

Nesse capítulo são apresentados o processamento dos dados e os resultados obtidos pela exploração desses dados (Figura 3), dividindo-se em:

- No subcapítulo 3.1 os dados utilizados são referenciados e o processo de download é descrito.
- No subcapítulo 3.2 são descritas todas as etapas do processamento dos dados até a análise diferencial das metilações.
- No subcapítulo 3.3 são detalhadas a adição de anotações fenotípicas aos resultados obtidos da análise diferencial das metilações, além da distribuição dessas características nos dados.
- No subcapítulo 3.4 é explicado como foi feita a categorização dos dados que serviram de entrada para os algoritmos de mineração de subgrupos, assim como a distribuição dos números de classes obtidas após a aplicação da categorização.
- Nos subcapítulos 5.1 e 5.2 são descritas as ferramentas e os parâmetros usados.
- No subcapítulo 3.3 são especificados os bancos de dados utilizados para anotar os dados finais afim de obter mais características para os subgrupos encontrados.

Figura 3 – Fluxo dos processos feitos: A) Download dos dados GSE1032 no GEO do NCBI e posteriores processos de controle de qualidade, filtragem e análise diferencial das metilações; B) Adição de novas anotações (predição da condição de fumante, predição das composição celular, predição da idade biológica); C) Processo de categorização dos dados para servir de entrada para os algoritmos *Beam Search* e SSDP+; D) Testes para descoberta dos GO termos e vias metabólicas do KEGG relacionadas aos subgrupos encontrados.



Fonte: Elaborada pelo autor (2022)

3.1 DADOS

Os dados contendo o status de metilação de DNA (GSE51032) são de um estudo multicêntrico com o objetivo de investigar as complexas relações entre nutrição e vários fatores do estilo de vida, além da etiologia do câncer e de outras doenças crônicas (RIBOLI et al., 2002). O cohort European Prospective Investigation into Cancer and Nutrition (EPIC) Italy é um sub-cohort de 46.857 voluntários com um estilo de vida e questionário com histórico pessoal padrão, dados antropométricos e amostras de sangue para extração de DNA. No último acompanhamento em 2010, 424 pacientes se mantiveram sem nenhum tipo de câncer, enquanto 235 desenvolveram câncer de mama (ONWUKA et al., 2020). Os dados foram obtidos através da plataforma GEO

com número de acesso GSE51032 (www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE51032). Essas dados são advindos do kit *Illumina Human Methylation 450 Beadchip*. Utilizamos a biblioteca GEOquery do R para ter acesso ao link para download desses dados brutos e da tabela de metadados, os dados brutos foram salvos usando *wget*.

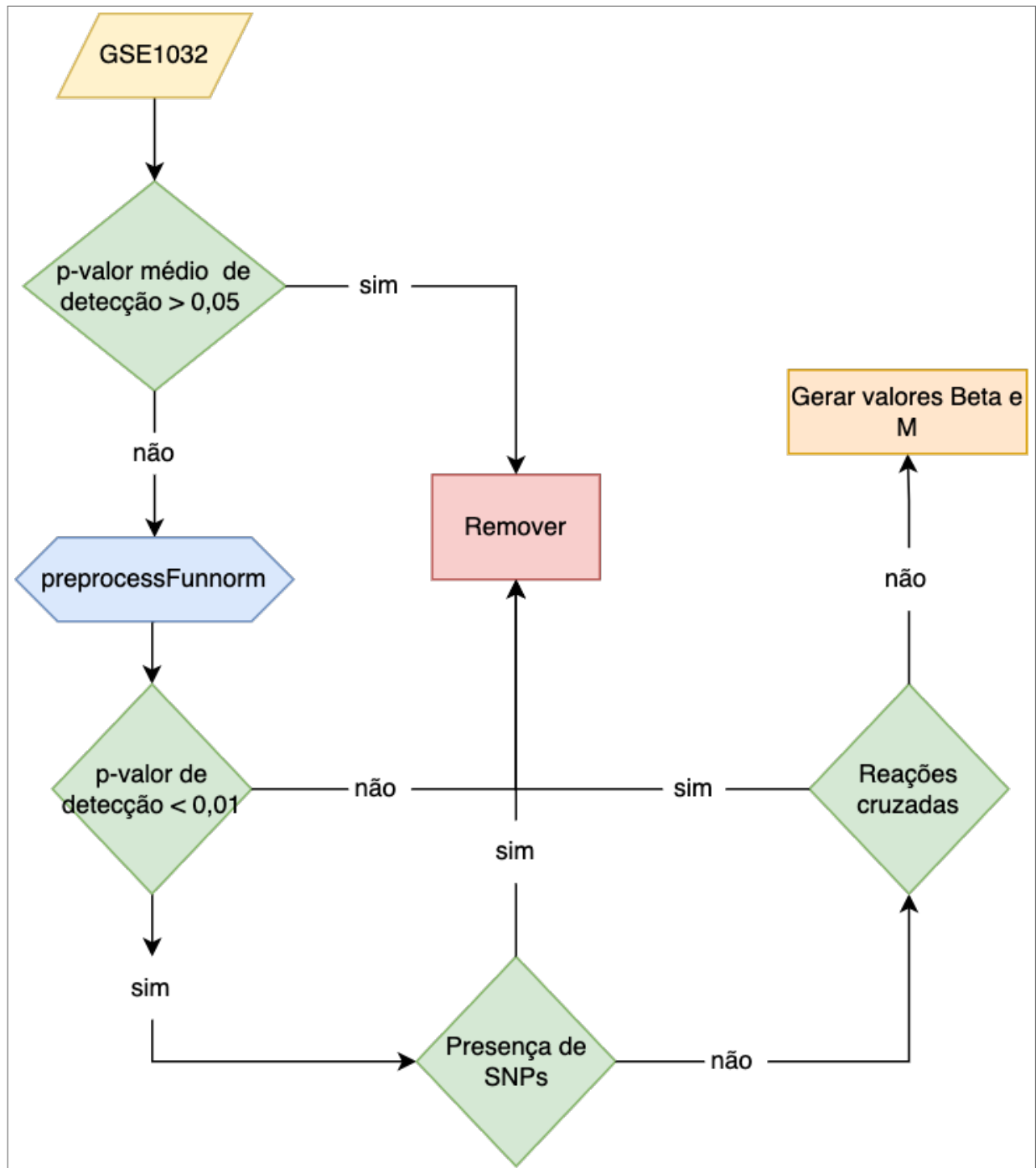
3.2 ANÁLISE DIFERENCIAL DAS METILAÇÕES

Todo o pré-processamento, normalização dos dados, filtragens, análise diferencial das metilações foram baseados em um pipeline feito por Maksimovic et al. (MAKSIMOVIC; PHIPSON; OSHLACK, 2016) usando as bibliotecas da linguagem R chamada Minfi (ARYEE et al., 2014) e a biblioteca Limma (RITCHIE et al., 2015) (Figura 3A). Com a biblioteca Minfi fizemos a leitura dos dados brutos do tipo Intensity Data files (IDAT), Foram extraídos para 485.512 sítios de CpGs para os pacientes sem câncer (n=424) e com câncer de mama (n=235). Os dados do sexo masculino foram excluídos por ser uma amostragem pequena (84 sem câncer e 2 com câncer), sobrando 233 pacientes com câncer de mama e 340 pacientes negativos para câncer do sexo feminino. Em seguida, para o controle de qualidade as amostras com baixa qualidade (p-valor < 0,05) foram removidas, para o nosso caso nenhuma amostra foi removida. Para minimizar a variação entre as amostras foi feita a normalização dos dados utilizando a função *Functional Normalization* (FunNorm) por ser considerada mais apropriada para dados com diferenças de metilação entre pacientes com câncer e sem (FORTIN et al., 2014).

As sondas com baixo desempenho são filtradas antes da análise de metilação diferencial por não serem confiáveis, assim essas sondas que falharam em pelo menos um paciente foram removidas usando um p-valor de detecção de 0,01. Além disso, todas as sondas com variantes de único polimorfismo (Single Nucleotide Polymorphisms (SNPs)) foram removidas para evitar influências desconhecidas causadas pelo ambiente, porém as probes mapeadas nos cromossomos sexuais foram mantidas, já que os voluntários do sexo masculino foram removidos e assim haver a possibilidade de encontrar relações interessantes dado que o câncer de mama é mais comum em mulheres. Probes com reações cruzadas, que se referem a probes que tiveram mapeamento em diferentes posições do genoma, também foram removidas (CHEN et al., 2013) (Figura 4). Para a análise diferencial das metilações foram feitas comparações em pares entre pacientes com câncer e sem câncer levando em consideração as diferenças individuais usando a biblioteca Limma (RITCHIE et al., 2015) da linguagem R. Essa análise foi feita com uma matriz de valores-M ($\log_2$ dos valores Beta que são a proporção de metilação para cada CpG)

obtendo valores de teste t de Student moderado e p-valores associados para cada sítio de CpG. Como são feitos milhares de testes dada a quantidade de CpGs disponíveis para saber o quanto significativo é uma mudança nos níveis de metilação entre dois grupos (negativo e positivo para câncer) usamos um limiar de 5% de taxa de descobertas falsas (False Discovery Rate (FDR)), isso significa que na nossa lista de metilações diferencialmente expressas, 5% serão descobertas falsas (BENJAMINI; HOCHBERG, 1995). Após o fluxo de análise diferencial das metilações foram encontradas 152.894 CpGs diferencialmente expressas entre pacientes sem e com câncer de mama, onde 82.787 estavam altamente expressas e 70.107 com baixa expressão. Ao final da análise diferencial temos uma lista de 152.894 CpGs e a tabela com os valores Beta das expressões de todas as CpGs do kit (485 mil no total). Anotamos todas as CpGs diferencialmente expressas usando a biblioteca *IlluminaHumanMethylation450kanno.ilmn12.hg19* para obter os possíveis genes pertencentes aos intervalos dessas regiões. A tabela com valores Beta para todas as CpGs iniciais serviram de entrada para obter novas anotações, como idade biológica, composição celular e condição de fumante (Figura 3B).

Figura 4 – Fluxo dos processos antes da análise diferencial das metilações: pre-filtro de qualidade das amostras, normalização dos dados e pós-filtros de qualidade das sondas.



Fonte: Elaborada pelo autor (2022)

3.3 ENRIQUECIMENTO DE DADOS FENOTÍPICOS

Os valores beta das metilações foram enriquecidos com os metadados já disponibilizados com os dados brutos, mas além disso foram obtidas predições de dados fenotípicos como a predição da condição de fumante, composição celular e idade biológica. Esse enriquecimento

foi feito visando encontrar subgrupos de pacientes que fossem caracterizados por alguma(s) dessas informações fenotípicas.

3.3.1 Metadados dos pacientes

Junto aos dados de metilação (GSE51032) tivemos acesso aos dados fenotípicos disponíveis para cada paciente, como: gênero, idade, idade quando teve a menarca e tempo para o diagnóstico. Além disso, usamos apenas os pacientes com diagnóstico positivo e negativo para câncer de mama para esse estudo, já que esses dados possuem pacientes com outros tipos de câncer.

3.3.2 Predição da condição de fumante

A biblioteca EpiSmokEr (*Epigenetic Smoking status Estimator*) é um classificador implementado em R para estimar a condição de fumante de uma pessoa baseado no perfil de metilação para o kit *Infinium HumanMethylation450* (BOLLEPALLI et al., 2019). Usamos como dados de entrada foram usados os valores Beta dos pacientes e uma planilha identificando o sexo de cada paciente. O classificador estima a probabilidade do indivíduo ser atualmente fumante (*Current smoker*), ex-fumante (*Former smoker*) e nunca ter fumado (*Never smoker*).

3.3.3 Composição celular

Como metilações são específicas para cada tipo de célula e dados de *Infinium HumanMethylation450* fornecem valores de metilação para populações de células, essa composição celular pode trazer uma variabilidade biológica que pode se relacionar a idade (JAFPE; IRIZARRY, 2014). Da biblioteca Minfi do R usamos a função *estimateCellCounts* que estima a proporção relativa dos tipos de células em uma amostra, dado valores Beta para amostras de sangue a função retornou a proporção de linfócitos, monócitos, células B e neutrófilos.

3.3.4 Estimando idade biológica

Através do perfil de metilações de DNA é possível estimar a idade biológica de um indivíduo usando modelos de predição. Dentre os modelos mais usados, citamos o modelo proposto por

Horvath baseado em 353 CpGs, um modelo proposto por Hannum baseado em 71 CpGs e um modelo chamado DNAm PhenoAge, proposto por Levine baseado em 513 CpGs (HORVATH, 2013; HANNUM et al., 2013; LEVINE et al., 2018). Conhecidos como relógios epigenéticos, esses modelos são considerados biomarcadores de idade biológica precisos, onde o desvio entre a idade cronológica e biológica por metilações é referenciado como aceleração da idade epigenética, podendo ser positiva ou negativa. Essa aceleração da idade epigenética tem sido correlacionada com condições relacionadas a idade, como doença de Parkinson, tempo até a morte, fragilidade, declínio cognitivo e físico (BREITLING et al., 2016; HORVATH; RITZ, 2015; MARIONI et al., 2015a; MARIONI et al., 2015b). ENmix é uma biblioteca em R para controle de qualidade e análise de dados de metilação, nela usamos a função *methyAge* dando os valores Beta como dados de entrada onde ela retornou as estimativas de idade para cada paciente de acordo com os modelos de Horvath, Hannum e PhenoAge.

3.3.5 Organizando os dados

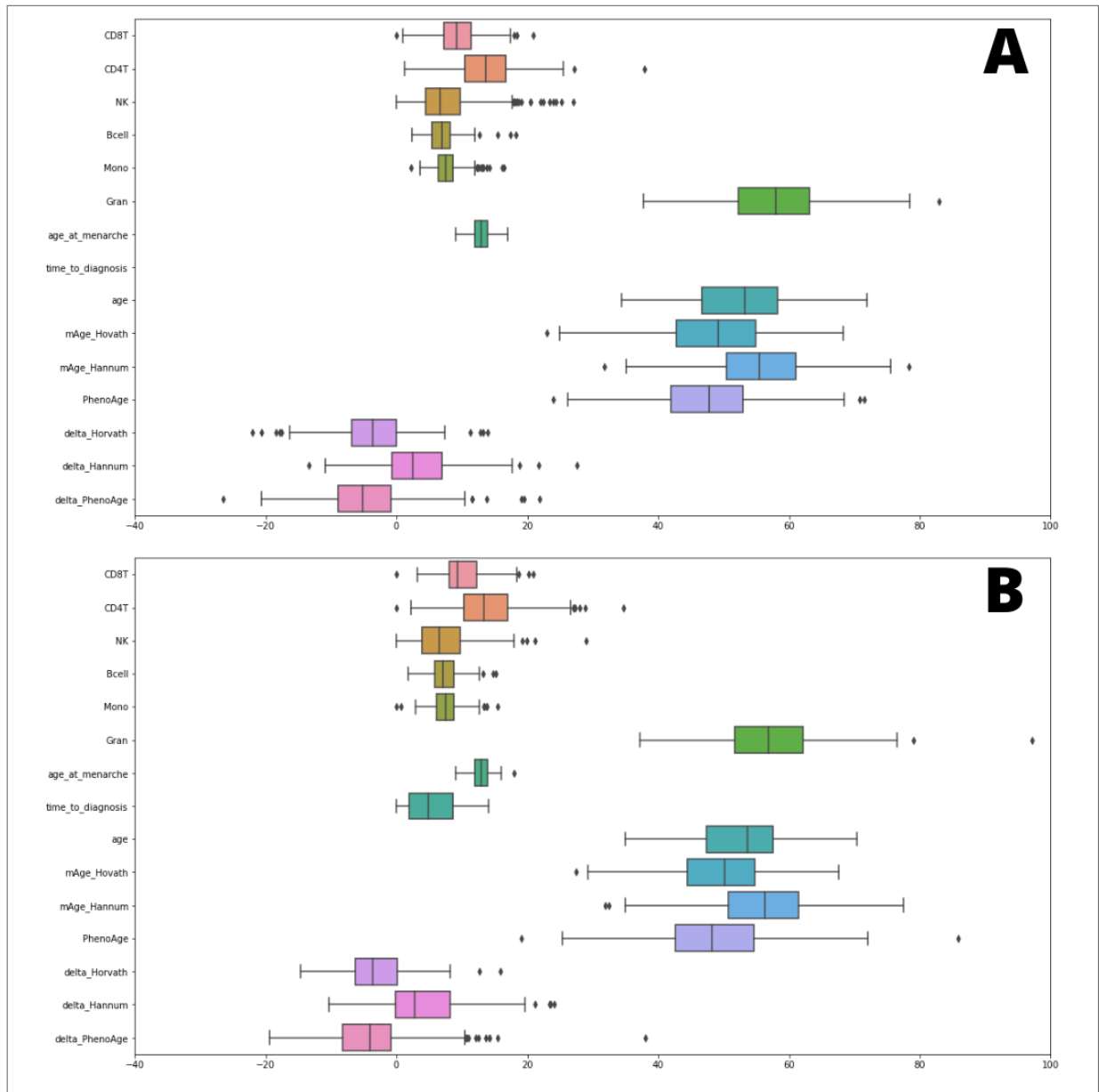
Os dados fenotípicos originais como idade, idade quando teve a menarca e tempo para diagnóstico; mais as anotações que foram adicionadas como condição de fumante (nunca: 425, fumante: 137, desistiu: 11), idade biológica por Horvath, Hannum, PhenoAge e composição celular para células T Cluster of Differentiation 8 (CD8), células T Cluster of Differentiation 4 (CD4), linfócitos Natural Killer (NK), células B, monócitos e granulócitos (Tabela 1) foram combinados em uma única tabela (Figura 5, 6). Apenas a coluna do tempo para diagnóstico foi removida por ter uma alta correlação negativa com a classe de interesse (Figura 7), isso ocorreu porque os valores dessa coluna só estão presentes em pacientes com câncer. Em uma primeira análise usando essa característica nos dados de entrada para os algoritmos de descoberta de subgrupos todos os subgrupos obtidos tiveram apenas "tempo para diagnóstico" como característica nos subgrupos e assim a necessidade de excluir essa informação dos dados de entrada para os algoritmos de descoberta de subgrupos. Essa tabela com todos os fenótipos e valores Beta foi filtrada para ter apenas as 152 mil CpGs diferencialmente expressas.

Tabela 1 – Estatística descritiva das colunas com valores numéricos.

Colunas	Mínimo	Media	Mediana	Desvio padrão	Máximo
age	34,37	52,49	53,39	7,4	71,94
age at menarch	9	12,71	13	1,53	18
time of diagnosis	0,04	5,36	4,87	3,69	14,13
Age Horvath	22,93	49,14	49,51	7,89	68,29
Age Hannum	31,74	56,1	55,84	8,05	78,28
Pheno Age	19,14	48,05	48,12	8,46	85,8
CD8T	0	9,66	9,26	3,32	20,87
CD4T	0	13,69	13,49	5,26	37,87
NK	0	7,42	6,69	4,6	29,02
Bcell	1,81	7,14	7,02	2,18	18,3
Mono	0	7,64	7,47	2,1	16,5
Gran	37,27	57,52	57,57	8,05	97,21

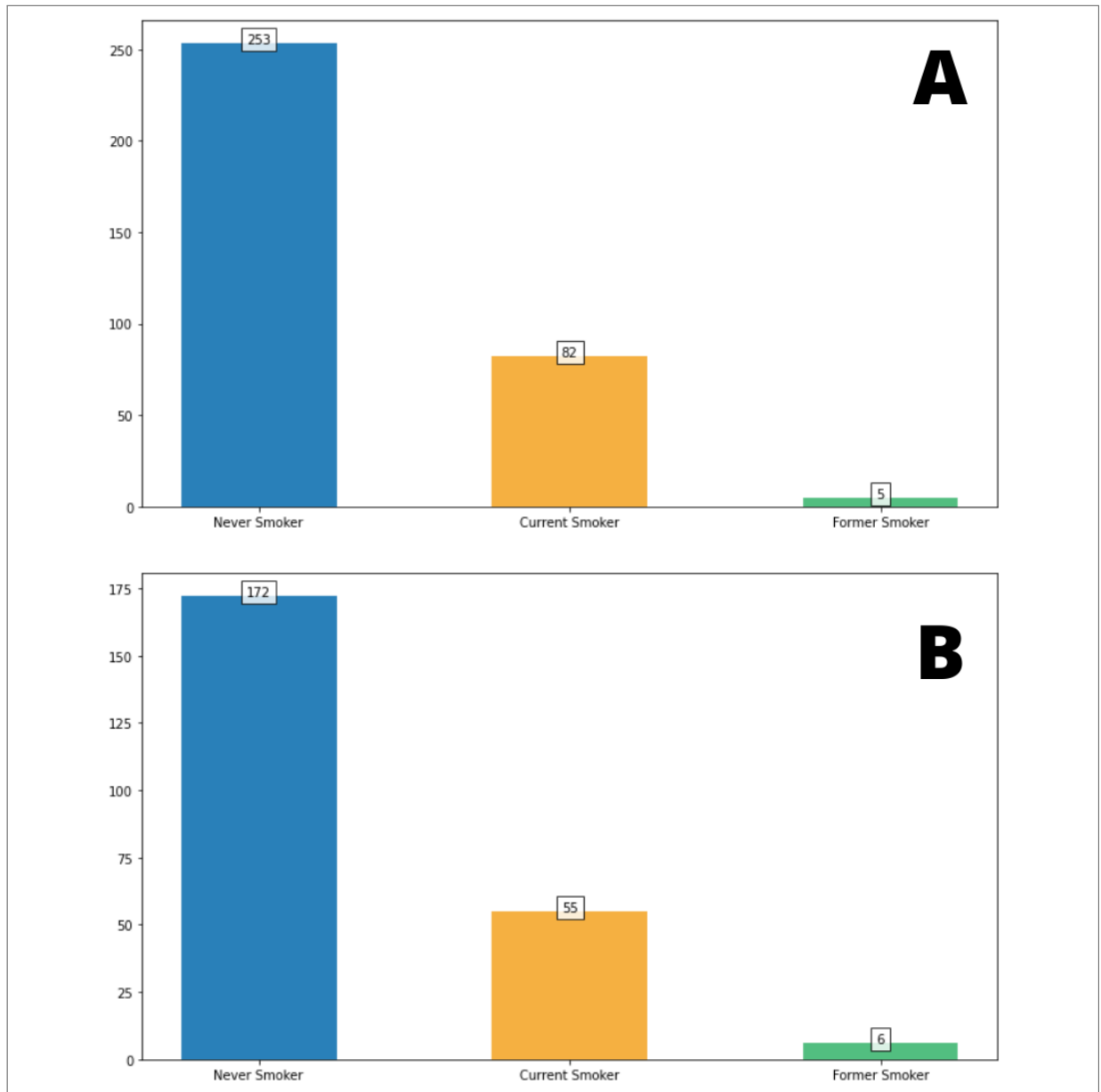
Fonte: Elaborada pelo autor (2022)

Figura 5 – Distribuição dos valores encontrados em todas as colunas disponíveis. Onde A temos apenas os pacientes com diagnóstico negativo para câncer, enquanto em B apenas os pacientes com câncer.



Fonte: Elaborada pelo autor (2022)

Figura 6 – Gráfico de barras da predição do status de fumante para todos os pacientes. Onde A temos apenas os pacientes com diagnóstico negativo para câncer, enquanto em B apenas os pacientes com câncer.



Fonte: Elaborada pelo autor (2022)

Figura 7 – Correlação entre as colunas de características disponíveis. É possível perceber uma alta correlação negativa entre a coluna "time of diagnosis" e a classe alvo ("cancer type classes").

	age_at_menarche	age	cancer_type_classes	time_to_diagnosis	CD8T	CD4T	NK	Bcell	Mono	Gran	mAge_Hovath	mAge_Hannum	PhenoAge
age_at_menarche	1.00	0.17	0.08	-0.07	-0.00	0.02	0.10	0.01	0.00	-0.07	0.15	0.11	0.12
age	0.17	1.00	0.01	-0.07	-0.09	0.17	0.20	0.16	-0.14	-0.18	0.78	0.67	0.64
cancer_type_classes	0.08	0.01	1.00	-0.75	-0.11	-0.01	0.03	-0.08	0.02	0.03	-0.02	-0.05	-0.05
time_to_diagnosis	-0.07	-0.07	-0.75	1.00	0.10	-0.03	-0.08	0.03	-0.04	0.04	-0.01	0.08	0.07
CD8T	-0.00	-0.09	-0.11	0.10	1.00	0.03	-0.01	0.26	-0.05	-0.46	0.01	-0.07	-0.07
CD4T	0.02	0.17	-0.01	-0.03	0.03	1.00	-0.07	0.29	-0.26	-0.63	0.06	-0.11	-0.10
NK	0.10	0.20	0.03	-0.08	-0.01	-0.07	1.00	0.02	-0.01	-0.50	0.23	0.18	0.12
Bcell	0.01	0.16	-0.08	0.03	0.26	0.29	0.02	1.00	-0.08	-0.51	0.08	0.00	0.02
Mono	0.00	-0.14	0.02	-0.04	-0.05	-0.26	-0.01	-0.08	1.00	-0.05	-0.10	-0.10	0.05
Gran	-0.07	-0.18	0.03	0.04	-0.46	-0.63	-0.50	-0.51	-0.05	1.00	-0.16	0.03	0.02
mAge_Hovath	0.15	0.78	-0.02	-0.01	0.01	0.06	0.23	0.08	-0.10	-0.16	1.00	0.78	0.75
mAge_Hannum	0.11	0.67	-0.05	0.08	-0.07	-0.11	0.18	0.00	-0.10	0.03	0.78	1.00	0.79
PhenoAge	0.12	0.64	-0.05	0.07	-0.07	-0.10	0.12	0.02	0.05	0.02	0.75	0.79	1.00

Fonte: Elaborada pelo autor (2022)

3.4 CATEGORIZAÇÃO DOS DADOS

Depois de feito o enriquecimento de dados fenotípicos aos dados das metilações consideradas diferencialmente expressas entre os voluntários com câncer e sem câncer, além dos níveis de metilação a maioria das características fenotípicas possuíam valores contínuos que precisaram ser transformados em classes para a análise seguinte de mineração de subgrupos. Para distribuir as características em classes, usamos a seguinte fórmula para descobrir a quantidade de classes para cada característica:

$$2^k \geq n \quad (3.1)$$

Onde k é número de classes e n o número total de valores para a característica em análise. Sabendo a quantidade de classes calculamos a largura dos intervalos das classes, para isso a fórmula:

$$\text{largura da classe} = \frac{\text{máximo} - \text{mínimo}}{\text{número de classes}} \quad (3.2)$$

O resultado deve ser arredondado pra cima mesmo em casos de valores inteiros, então calculamos os limites inferiores do intervalo da classe a partir do valor mínimo presente para a característica somando a largura calculada anteriormente. Para obter o limite superior do intervalo da classe subtraímos um do próximo limite inferior. O número de classes obtidas para cada coluna podem ser vistos na Tabela 2, no caso das CpGs algumas tiveram 17 classes enquanto outras 18 classes graças a distribuição dos valores de expressão serem mais variáveis. A tabela resultante da categorização dos dados serviu de entrada para os algoritmos *Beam Search* e SSDP+ (Figura 3C).

Tabela 2 – Número de classes para cada coluna dos dados que serviram de entrada para os algoritmos de descoberta de subgrupos.

Colunas	Número de classes
time of diagnosis	9
age at menarch	11
age	10
CD8T	7
CD4T	10
NK	10
Bcell	9
Mono	9
Gran	9
Predicted smoking status	3
Age Hannum	10
Age Horvath	10
Pheno Age	9
CpGs	17,18

Fonte: Elaborada pelo autor (2022)

3.5 BEAM SEARCH

A tabela com os valores Beta das metilações e as características fenotípicas foi dada como dado de entrada para a biblioteca *pysubgroup* que disponibiliza a implementação de funções de mineração de subgrupos como o algoritmo Beam search em Python (LEMMERICH; BECKER, 2018). O *Beam search* recebe dois parâmetros além dos dados, o *depth* (10) e o *result set size* (100) que são o tamanho da lista de subgrupos retornados e o tamanho máximo para cada subgrupo respectivamente. Além desses parâmetros, foi utilizado o *beam width adaptive* para que o tamanho do subgrupo seja ajustado a medida que são encontrados.

3.6 SSDP+

Foi feita a cópia do repositório original do SSDP+ para se criar uma cópia do código com parâmetros e facilitar testes. Para executar o algoritmo, é preciso além dos dados, passar quantos subgrupos vão ser agrupados no arquivo final (10), o tamanho do cache que será criado com subgrupos parecidos com o encontrado originalmente (5), o valor de similaridade (0.5) e a classe alvo na coluna de interesse. Para o valor de similaridade, foi decidido um

valor intermediário já que valores próximos de 1 indicam alta similaridade entre os subgrupos e valores próximo de 0 alta diversidade entre os subgrupos.

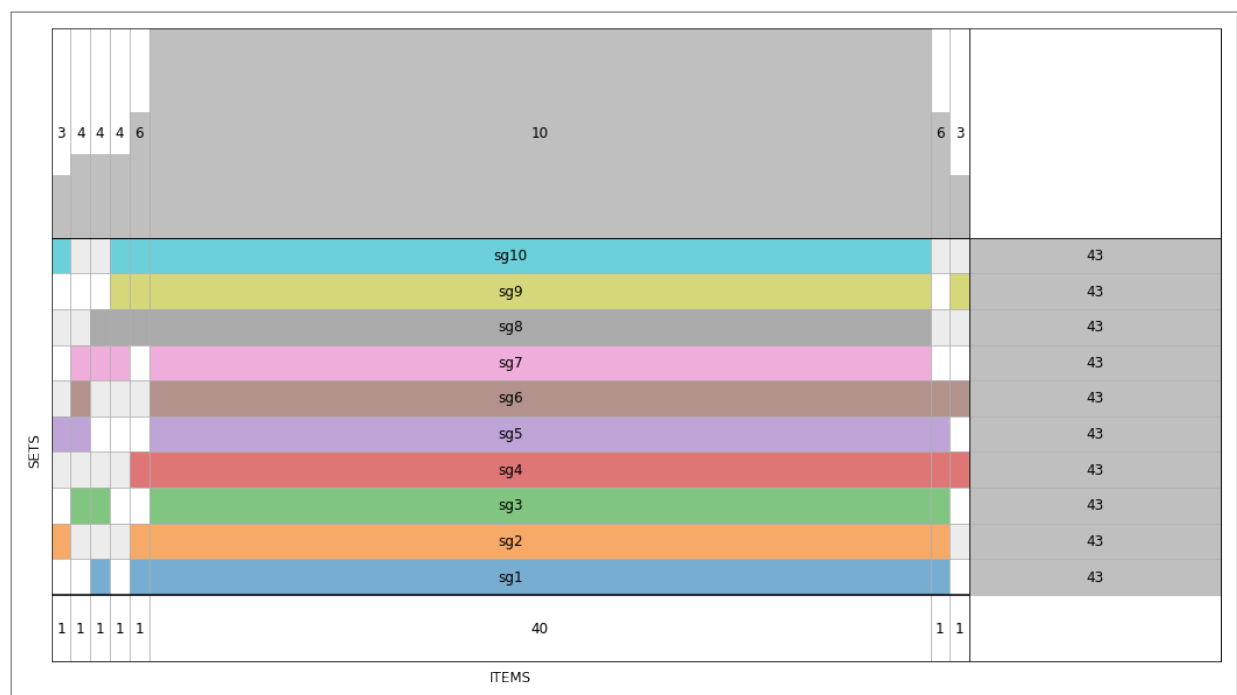
3.7 ANOTAÇÃO DOS SUBGRUPOS

A fim de obter uma compreensão dos processos biológicos em que os subgrupos de CpGs diferencialmente metilados podiam estar envolvidos, realizamos a análise de ontologia genética termos Gene Ontology (GO)) e análise de possíveis vias metabólicas (Kyoto Encyclopedia of Genes and Genomes (KEGG)) utilizando a função *gometh* do pacote *missMethyl* (PHIPSON; MAKSIMOVIC; OSHLACK, 2016). Essa função recebe como dado de entrada um vetor de caracteres (por exemplo, cg20832020) das CpGs consideradas significativas. Para o teste de ontologia genética (padrão), foi especificado *collection="GO"*, enquanto que para a análise de vias metabólicas foi especificado *collection = "KEGG"*.

4 DESCOBERTA DOS SUBGRUPOS

Os subgrupos são formado por regras, onde essas regras são caracterizadas por grupos de CpGs contendo determinado categoria de expressão (Tabelas 3 e 4). Para o algoritmo *Beam search* os subgrupos encontrados tiveram o mesmo tamanho (43 CpGs) englobando 151 pacientes do total de 233 pacientes com câncer, ou seja, subgrupos puros. Os 151 pacientes englobados por todos os 10 subgrupos encontrados foram basicamente os mesmos, variando entre 1 ou 2 pacientes para cada subgrupo. Todos os subgrupos tiveram também o mesmo valor da métrica de qualidade WR_{Acc} (0,1564) e $lift$ (2,4592), indicando o quão parecidos são esses subgrupos (Tabela 3). Para comparar esses subgrupos foi feito um diagrama de Venn (Figura 8), nesse diagrama é possível identificar que das 43 CpGs pertencentes aos subgrupos encontrados, 40 se repetiam e apenas 3 posições variaram entre 7 CpGs nos subgrupos.

Figura 8 – Diagrama de Venn com as CpGs agrupadas nos subgrupos encontrados pelo algoritmo Beam Search.



Fonte: Elaborada pelo autor (2022)

O SSDP+ foi utilizado duas vezes para cada métrica de qualidade disponível, WR_{Acc} e Qg . Os resultados obtidos com a métrica Qg são menores em tamanho e diversos (Figura 9), englobando menos pacientes mas sendo todos subgrupos puros (Tabela 4). Já os subgrupos resultantes da análise com a métrica WR_{Acc} foram subgrupos diversos (Figura 10), porém nenhum dos subgrupos encontrados eram puros, ou seja, todos os subgrupos tinham pacientes

Tabela 3 – Tabela com os 4 primeiros subgrupos encontrados pelo algoritmo Beam Search e suas respectivas classes.

subgrupo	metilações	amostragem	lift	wracc
1	cg00015930=16, cg00262415=0, cg00505073=15, cg01016662=0, cg01031032=0, cg01622006=0, cg01892567=16, cg01962937=0, cg02260340=0, cg02527528=0, cg02647825=15, cg03161767=0, cg03255749=15, cg03397332=0, cg03801179=0, cg04104256=0, cg04285418=0, cg04551440=0, cg05006947=9, cg05046026=0, cg05664296=15, cg06097659=0, cg06668922=15, cg06770554=0, cg06775930=0, cg07062711=0, cg07610406=0, cg09449449=15, cg10318121=0, cg10437571=0 , cg12071806=0, cg12510286=0 , cg14651446=0, cg15717853=0, cg15859496=14, cg17386812=15 , cg20972117=0, cg21279459=9, cg24154336=16, cg25373297=0, cg26821539=0, cg27544799=0, ch. 9.2042397F=0	43	2,4592	0,1564
2	cg00015930=16, cg00262415=0, cg00505073=15, cg01016662=0, cg01031032=0, cg01622006=0, cg01892567=16, cg01962937=0, cg02260340=0, cg02527528=0, cg02647825=15, cg03161767=0, cg03255749=15, cg03397332=0, cg03801179=0, cg04104256=0, cg04285418=0, cg04551440=0, cg05006947=9, cg05046026=0, cg05664296=15, cg06097659=0, cg06668922=15, cg06770554=0, cg06775930=0, cg07062711=0, cg07610406=0, cg09449449=15, cg10318121=0, cg10437571=0 , cg12071806=0, cg13613388=15 , cg14651446=0, cg15717853=0, cg15859496=14, cg17386812=15 , cg20972117=0, cg21279459=9, cg24154336=16, cg25373297=0, cg26821539=0, cg27544799=0, ch. 9.2042397F=0	43	2,4592	0,1564
3	cg00015930=16, cg00262415=0, cg00505073=15, cg01016662=0, cg01031032=0, cg01622006=0, cg01892567=16, cg01962937=0, cg02260340=0, cg02527528=0, cg02647825=15, cg03161767=0, cg03255749=15, cg03397332=0, cg03801179=0, cg04104256=0, cg04285418=0, cg04551440=0, cg05006947=9, cg05046026=0, cg05664296=15, cg06097659=0, cg06668922=15, cg06770554=0, cg06775930=0, cg07062711=0, cg07610406=0, cg09449449=15, cg10318121=0, cg10437571=0 , cg12071806=0, cg12510286=0 , cg14651446=0, cg15717853=0, cg15859496=14, cg20972117=0, cg21279459=9, cg21881074=0 , cg24154336=16, cg25373297=0, cg26821539=0, cg27544799=0, ch. 9.2042397F=0	43	2,4592	0,1564
4	cg00015930=16, cg00262415=0, cg00505073=15, cg01016662=0, cg01031032=0, cg01622006=0, cg01892567=16, cg01962937=0, cg02260340=0, cg02527528=0, cg02647825=15, cg03161767=0, cg03255749=15, cg03397332=0, cg03801179=0, cg04104256=0, cg04285418=0, cg04551440=0, cg05006947=9, cg05046026=0, cg05664296=15, cg06097659=0, cg06668922=15, cg06770554=0, cg06775930=0, cg07062711=0, cg07610406=0, cg09449449=15, cg10318121=0, cg10437571=0 , cg12071806=0, cg14651446=0, cg15717853=0, cg15859496=14, cg17386812=15 , cg20972117=0, cg21279459=9, cg21327609=0 , cg24154336=16, cg25373297=0, cg26821539=0, cg27544799=0, ch. 9.2042397F=0	43	2,4592	0,1564

Fonte: Elaborada pelo autor (2022)

negativos para câncer além dos pacientes positivos para câncer de mama, incluindo o cache com subgrupos similares.

Os subgrupos resultantes do SSDP+ usando a métrica WR_{Acc} não foram trabalhados, uma vez que todos eles tiveram pacientes negativos para câncer, não sendo de interesse pensando na possibilidade de ter um painel de CpGs que diferenciem subgrupos de pacientes com câncer de mama essa métrica acabou não se demonstrando eficaz para essa proposta. Como todos os subgrupos encontrados usando *Beam search* continham os mesmos pacientes, a distribuição das anotações nesses subgrupos foi a mesma (Figura 11, 12). Já os subgrupos obtidos usando a métrica Qg no algoritmo SSDP+ foram mais diversos contendo diferentes pacientes dentro desses subgrupos, assim a distribuição das anotações também acabou variando (Figura 13, 14, 15).

Com a descoberta de subgrupos foram separadas as listas de CpGs em cada subgrupo puro resultante do *Beam search* e do SSDP+, esse último apenas para a métrica de qualidade Qg . De um arquivo maior contendo todas as CpGs diferencialmente expressas anotadas com os

Tabela 4 – Tabela com os 10 subgrupos encontrados pelo algoritmo SSDP+ para a métrica de qualidade Qg.

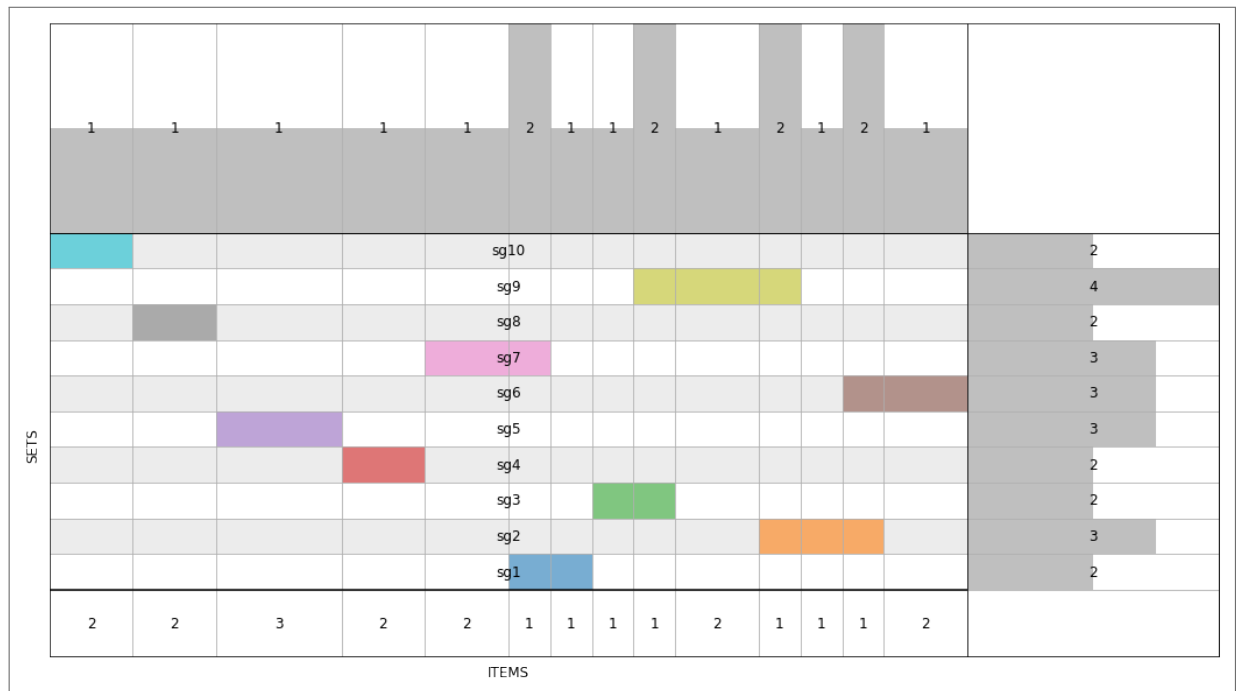
subgrupo	metilações	amostragem	qg
1	cg03610970=13,cg22793136=9	25	1
2	cg08879482=14,cg24434368=9,cg09706833=3	23	1
3	cg14154819=16,cg10668569=14	22	1
4	cg22947748=1,cg03534326=11	22	1
5	cg24367316=9,cg16995299=14,cg06532611=13	21	1
6	cg02674789=14,cg26684511=9,cg24434368=9	20	1
7	cg12785643=9,cg10110581=0,cg03610970=13	20	1
8	cg17820247=0,cg10872521=9	20	1
9	cg24463471=15,cg14154819=16,cg25741023=9,cg09706833=3	20	1
10	cg09444358=13,cg13551227=1	19	1

Fonte: Elaborada pelo autor (2022)

genes e regiões dessas metilações foram filtradas apenas as CpGs encontradas nessa lista de subgrupos puros na tentativa de descobrir os genes que estavam nessas regiões e assim facilitar a discussão sobre a importância desses subgrupos encontrados levando em consideração a literatura sobre esses genes e o câncer de mama, porém nem todas as CpGs possuem genes em sua região (Tabela 5). A fim de obter uma compreensão dos processos biológicos em que essas listas de subgrupos de CpGs diferencialmente metiladas podem estar envolvidos, foi feito o teste para descoberta de termos de ontologia genética (termos GO) e para vias metabólicas (KEGG). Como resultado para os 10 primeiros subgrupos encontrados pelos SSDP+ e *Beam search* temos tabelas com os 10 primeiros GO termos (Tabelas 6 e 7) e vias metabólicas (Tabelas 8 e 9) mais significantes encontrados para os genes presentes nas regiões de metilações diferencialmente expressas.

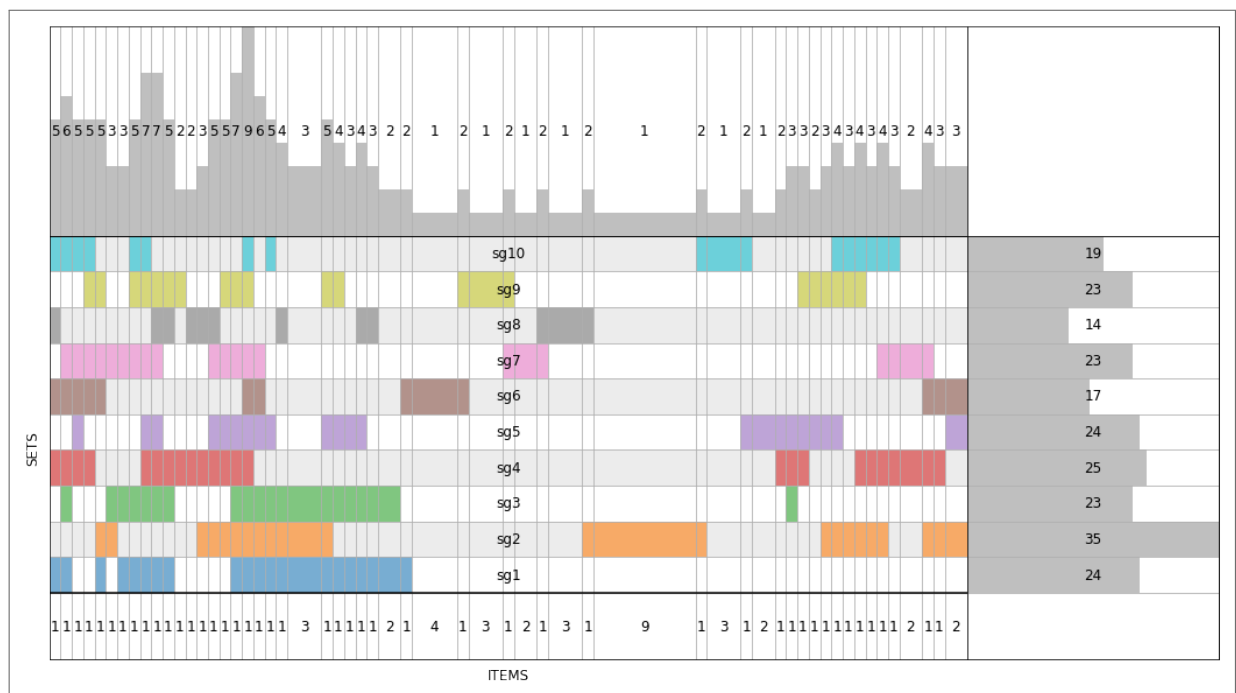
Para os subgrupos encontrados usando *Beam search* tivemos subgrupos muito parecidos, onde variaram basicamente 3 CpGs do total de 43 encontradas para cada um, ou seja, todos eles tiveram 40 CpGs em comum. Por terem todas as CpGs em comum, os termos GO encontrados se repetiram também. Por exemplo, os termos GO:0150070, GO:0150072, GO:0015191,

Figura 9 – Diagrama de Venn com as CpGs agrupadas nos subgrupos encontrados pelo algoritmo SSDP+ usando a métrica de qualidade Qg.



Fonte: Elaborada pelo autor (2022)

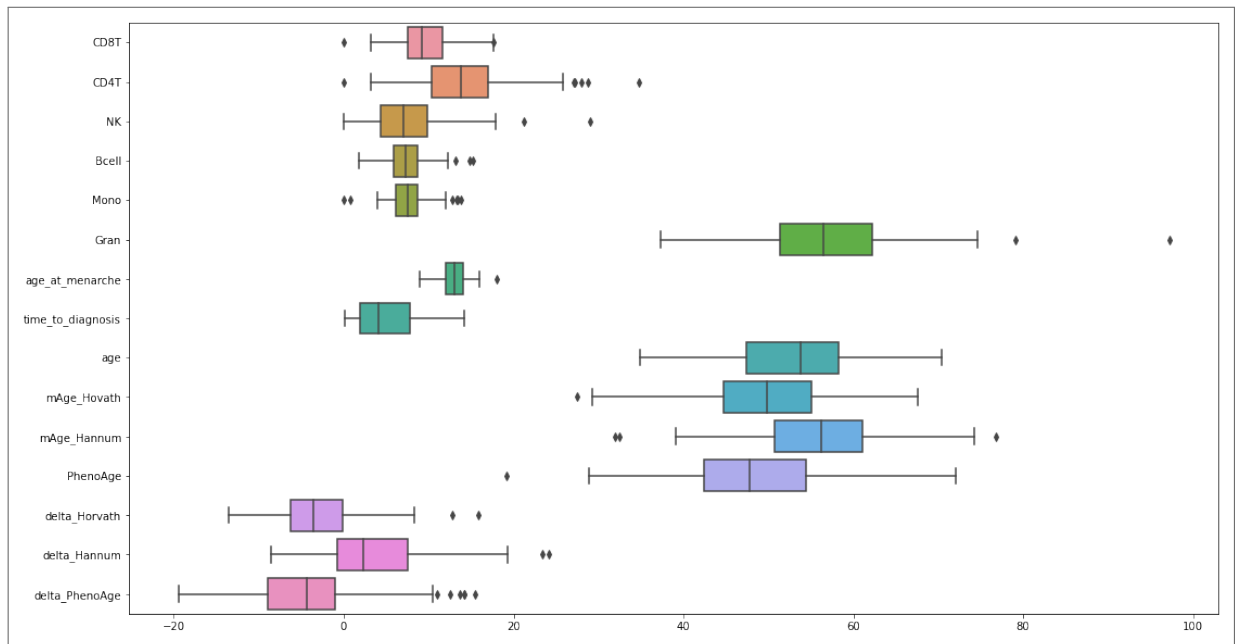
Figura 10 – Diagrama de Venn com as CpGs agrupadas nos subgrupos encontrados pelo algoritmo SSDP+ usando a métrica de qualidade WRAcc.



Fonte: Elaborada pelo autor (2022)

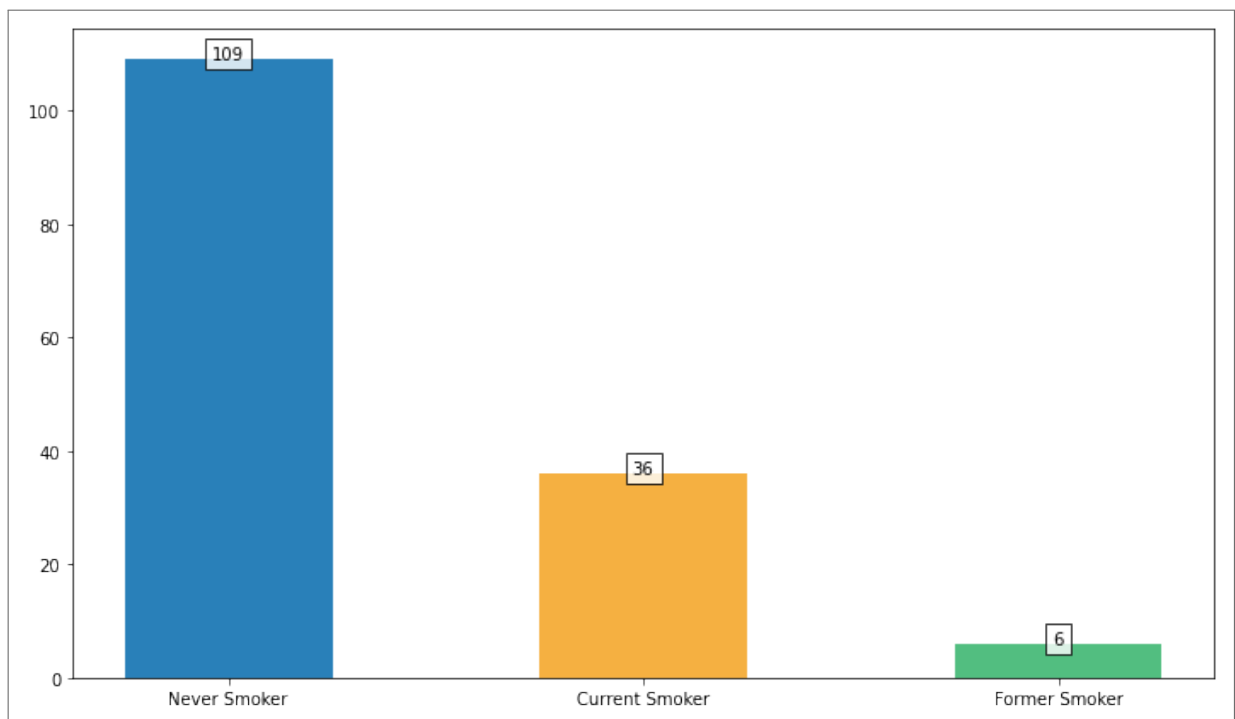
GO:0015821, GO:0043865, GO:0004454, GO:0150077 foram comuns a todos os subgrupos, provavelmente porque esses termos só possuem um gene relacionado e que esta diferenci-

Figura 11 – Distribuição das características para todos os pacientes englobados nos subgrupos encontrados pelo Beam search.



Fonte: Elaborada pelo autor (2022)

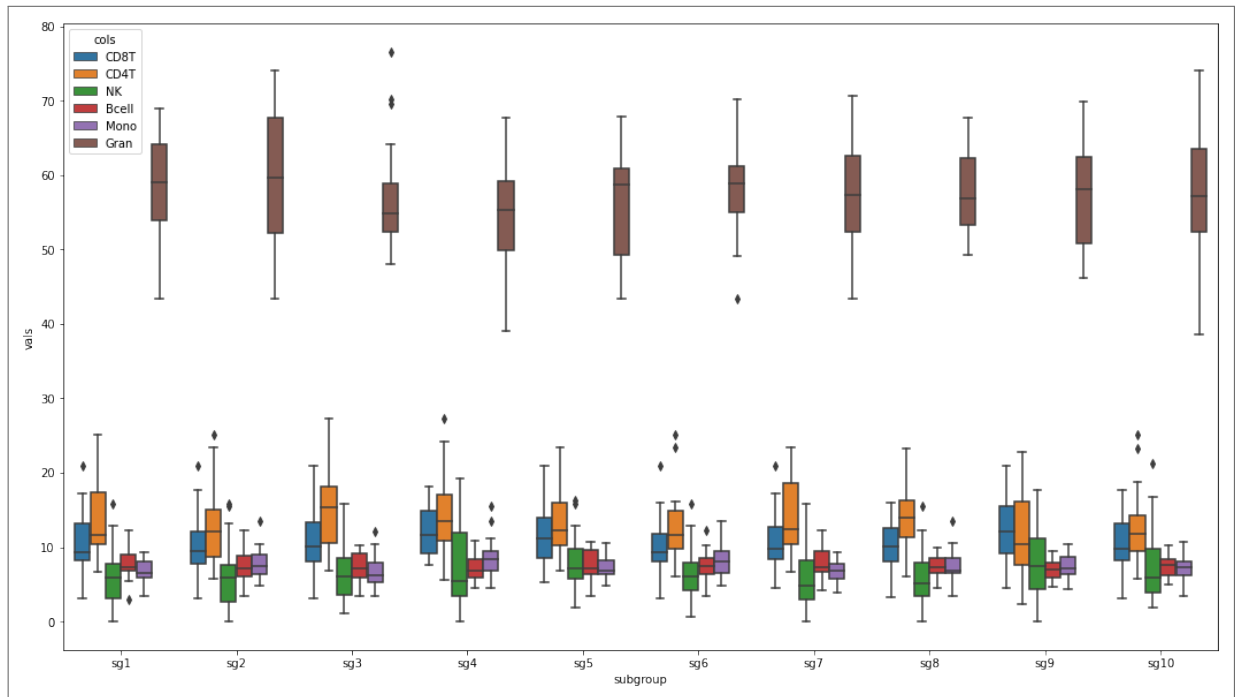
Figura 12 – Gráfico de barras da predição do status de fumante para todos os pacientes englobados nos subgrupos encontrados pelo Beam search.



Fonte: Elaborada pelo autor (2022)

almente expresso nos subgrupos (Tabela 6). Sobre as vias metabólicas para esses mesmos subgrupos, não tivemos nenhuma via estatisticamente significativa para um p-valor de 0,01,

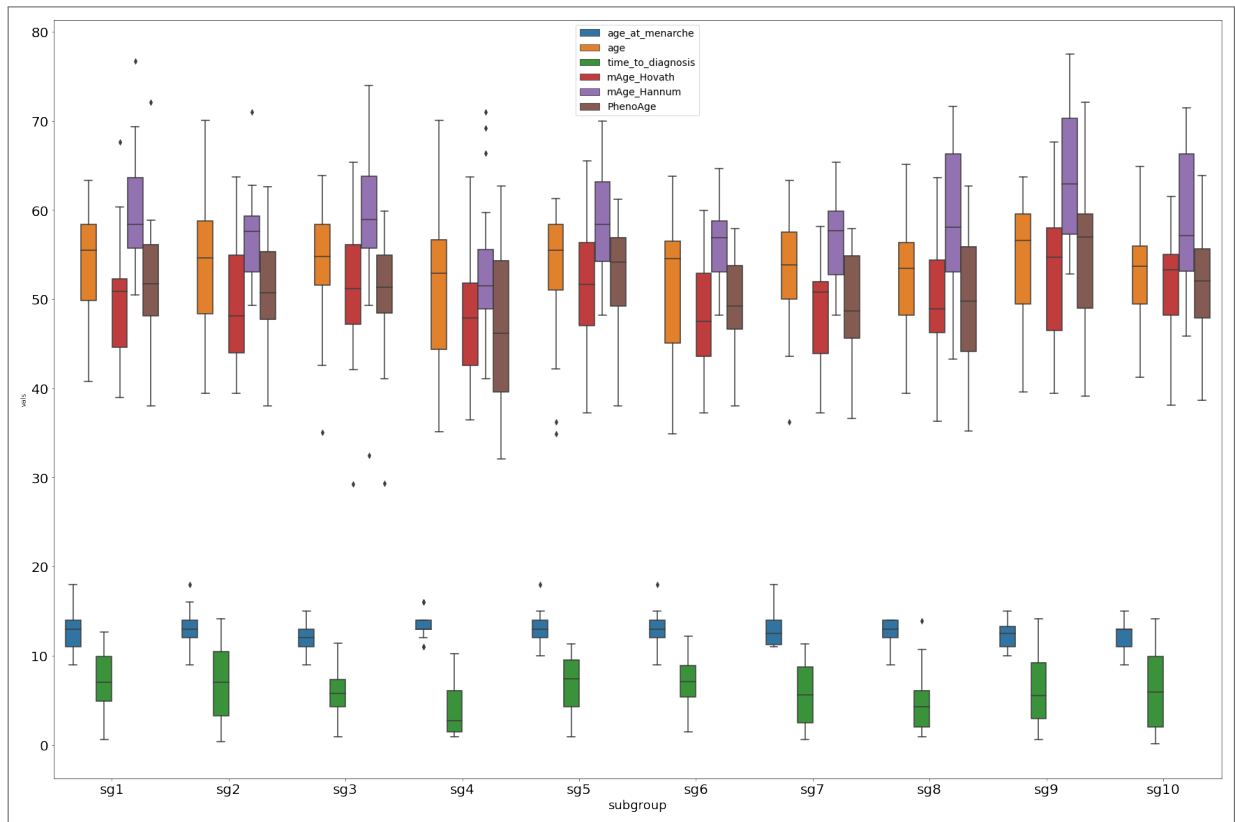
Figura 13 – Distribuição das anotações relacionadas a composição celular para os subgrupos encontrados usando o SSDP+.



Fonte: Elaborada pelo autor (2022)

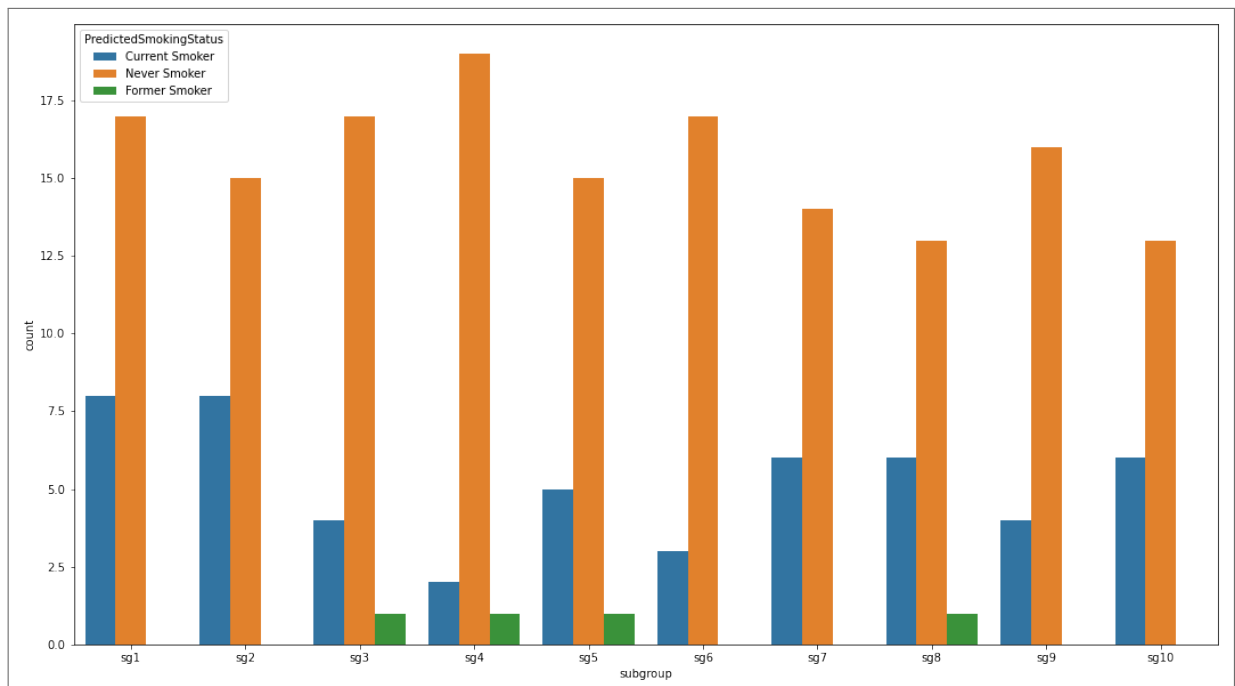
isso provavelmente ocorreu porque não houve genes diferencialmente expressos suficientes nos subgrupos para caracterizar as vias metabólicas (Tabela 8). Já os subgrupos encontrados usando SSDP+ com a métrica Qg temos uma alta diversidade de termos GO, mostrando o quanto esses subgrupos são diferentes uns dos outros, apenas o subgrupo 7 teve menos termos GO significativos para um p-valor de 0,01 (Tabela 7). Sobre as vias metabólicas relacionadas a esses subgrupos, apenas duas vias metabólicas foram significantes para um p-valor de 0,01 para o subgrupo 8, "SNARE interactions in vesicular transport" e "Vibrio cholerae infection" (Tabela 9). Outro detalhe é que não foi obtido nenhum termo GO ou vias metabólicas para o subgrupo 4, já que não foi encontrado nenhum gene presente nas regiões de metilação englobadas por esse subgrupo.

Figura 14 – Distribuição das anotações relacionadas a idade e tempo para diagnóstico para os subgrupos encontrados usando o SSDP+ com a métrica de qualidade Qg.



Fonte: Elaborada pelo autor (2022)

Figura 15 – Gráfico de barras da predição do status de fumante para os subgrupos encontrados usando o SSDP+ com a métrica de qualidade Qg.



Fonte: Elaborada pelo autor (2022)

Tabela 5 – Tabela com as anotações das CpGs englobadas nos subgrupos encontrados tanto pelo Beam search quanto pelo SSDP+ para a métrica Qg.

Name	UCSC_RefGene_Name	UCSC_RefGene_Accession	chr	pos	Probe_rs	Islands_Name	Relation_to_Island	beam_search	ssdp
cg00015930	NA	NA	chr1	46921746	NA	NA	OpenSea	all	NA
cg00262415	PUSL1;ACAP3	NM_153339;NM_030649	chr1	1242732	NA	chr1:1242400-1245185	Island	all	NA
cg00505073	C20orf165	NM_080608	chr20	44516626	NA	chr20:44518897-44520361	N_Shelf	all	NA
cg01016662	NA	NA	chr1	110673082	NA	chr1:110672810-110673339	Island	all	NA
cg01031032	CD200	NM_001004196;NM_005944	chr3	112051894	NA	chr3:112051893-112052406	Island	all	NA
cg01622006	NA	NA	chr6	170463397	rs3104183	NA	OpenSea	all	NA
cg01892567	NOM1	NM_138400	chr7	156760930	rs6948221	chr7:156760878-156761236	Island	all	NA
cg01962937	HTRA1	NM_002775;NM_002775	chr10	124221066	NA	chr10:124220338-124222240	Island	all	NA
cg02260340	RNF5P1;RNF5;AGPAT1	NR_003129;NM_006913;NM_032741	chr6	32145753	NA	NA	OpenSea	all	NA
cg02527528	NA	NA	chr16	66907095	NA	chr16:66907017-66907272	Island	all	NA
cg02647825	TSEN2	NM_001145395;NM_001145392;NM_025265;NM_001145393;NM_001145394	chr3	12527333	NA	chr3:12525749-12526346	S_Shore	all	NA
cg03161767	NOC2L	NM_015658	chr1	894207	NA	chr1:894313-902654	N_Shore	all	NA
cg03255749	NXN	NM_022463	chr17	739405	NA	chr17:738868-739377	S_Shore	all	NA
cg03397332	NA	NA	chr5	1385979	NA	chr5:1385978-1386862	Island	all	NA
cg03801179	MATN3	NM_002381;NM_002381	chr2	20212423	NA	chr2:20212067-20212783	Island	all	NA
cg04104256	CKAP5	NM_014756;NM_001008938	chr11	46868062	NA	chr11:46867194-46868077	Island	all	NA
cg04285418	GTDC1	NM_001164629	chr2	145089952	NA	chr2:145089453-145090397	Island	all	NA
cg04551440	KATNAL1;KATNAL1	NM_001014380;NM_032116	chr13	30881194	NA	chr13:30880899-30881939	Island	all	NA
cg05006947	SLC38A7	NM_018231	chr16	58718479	NA	chr16:58718309-58718703	Island	all	NA
cg05046026	C11orf10;FEN1;MIR611	NM_014206;NM_004111;NR_030342	chr11	61559981	NA	chr11:61559907-61560549	Island	all	NA
cg05664296	FPR1	NM_002029	chr19	52249599	NA	NA	OpenSea	all	NA
cg06097659	TNFRSF18	NM_148901;NM_148901;NM_004195;NM_004195;NM_148902;NM_148902	chr1	1141991	NA	chr1:1141670-1142150	Island	all	NA
cg06668922	LPCAT1	NM_024830	chr5	1467884	rs58459386	chr5:1467831-1468263	Island	all	NA
cg06770554	C1orf56	NM_017860	chr1	151020549	NA	chr1:151020162-151020627	Island	all	NA
cg06775930	NA	NA	chr6	126069089	NA	chr6:126068722-126071320	Island	all	NA

Fonte: Elaborada pelo autor (2022)

Tabela 6 – Tabela com os 10 primeiros termos de ontologia genética para os 10 subgrupos encontrados pelo algoritmo Beam search.

ID	ontologia	termo	número de genes	genes diferencialmente metilados	p-valor	FDR	subgrupo
GO:0150070	BP	regulation of arginase activity	1	1	0,0019	1	1
GO:0150072	BP	positive regulation of arginase activity	1	1	0,0019	1	1
GO:0015191	MF	L-methionine transmembrane transporter activity	1	1	0,0019	1	1
GO:0015821	BP	methionine transport	1	1	0,0019	1	1
GO:0043865	MF	methionine transmembrane transporter activity	1	1	0,0019	1	1
GO:0004454	MF	ketoheokinase activity	1	1	0,0021	1	1
GO:0150077	BP	regulation of neuroinflammatory response	37	2	0,0028	1	1
GO:0150073	BP	regulation of protein-glutamine gamma-glutamyltransferase activity	2	1	0,0030	1	1
GO:0150074	BP	positive regulation of protein-glutamine gamma-glutamyltransferase activity	2	1	0,0030	1	1
GO:0030690	CC	Noc1p-Noc2p complex	1	1	0,0031	1	1
GO:0150070	BP	regulation of arginase activity	1	1	0,0018	1	2
GO:0150072	BP	positive regulation of arginase activity	1	1	0,0018	1	2
GO:0015191	MF	L-methionine transmembrane transporter activity	1	1	0,0018	1	2
GO:0015821	BP	methionine transport	1	1	0,0018	1	2
GO:0043865	MF	methionine transmembrane transporter activity	1	1	0,0018	1	2
GO:0150077	BP	regulation of neuroinflammatory response	37	2	0,0026	1	2
GO:0150073	BP	regulation of protein-glutamine gamma-glutamyltransferase activity	2	1	0,0029	1	2
GO:0150074	BP	positive regulation of protein-glutamine gamma-glutamyltransferase activity	2	1	0,0029	1	2
GO:0150076	BP	neuroinflammatory response	41	2	0,0031	1	2
GO:0030690	CC	Noc1p-Noc2p complex	1	1	0,0032	1	2
GO:0150070	BP	regulation of arginase activity	1	1	0,0020	1	3
GO:0150072	BP	positive regulation of arginase activity	1	1	0,0020	1	3
GO:0015191	MF	L-methionine transmembrane transporter activity	1	1	0,0020	1	3
GO:0015821	BP	methionine transport	1	1	0,0020	1	3
GO:0043865	MF	methionine transmembrane transporter activity	1	1	0,0020	1	3

Fonte: Elaborada pelo autor (2022)

Tabela 7 – Tabela com os 10 primeiros termos de ontologia genética para os 10 subgrupos encontrados pelo algoritmo SSDP+ para a métrica de qualidade Qg.

ID	ontologia	termo	número de genes	genes diferencialmente metilados	p-valor	FDR	subgrupo
GO:0048266	BP	behavioral response to pain	12	1	0,0012	1	1
GO:1902287	BP	semaphorin-plexin signaling pathway involved in axon guidance	11	1	0,0016	1	1
GO:0002116	CC	semaphorin receptor complex	11	1	0,0018	1	1
GO:0017154	MF	semaphorin receptor activity	12	1	0,0020	1	1
GO:1902285	BP	semaphorin-plexin signaling pathway involved in neuron projection guidance	13	1	0,0021	1	1
GO:0032793	BP	positive regulation of CREB transcription factor activity	19	1	0,0021	1	1
GO:0004888	MF	transmembrane signaling receptor activity	970	2	0,0025	1	1
GO:0048265	BP	response to pain	27	1	0,0035	1	1
GO:0038023	MF	signaling receptor activity	1144	2	0,0036	1	1
GO:0060089	MF	molecular transducer activity	1144	2	0,0036	1	1
GO:1902435	BP	regulation of male mating behavior	2	1	0,0014	1	2
GO:1902437	BP	positive regulation of male mating behavior	2	1	0,0014	1	2
GO:0034983	BP	peptidyl-lysine deacetylation	8	1	0,0021	1	2
GO:0051599	BP	response to hydrostatic pressure	6	1	0,0025	1	2
GO:0031078	MF	histone deacetylase activity (H3-K14 specific)	10	1	0,0025	1	2
GO:0032041	MF	NAD-dependent histone deacetylase activity (H3-K14 specific)	10	1	0,0025	1	2
GO:0060179	BP	male mating behavior	8	1	0,0028	1	2
GO:0030955	MF	potassium ion binding	11	1	0,0030	1	2
GO:0070933	BP	histone H4 deacetylation	11	1	0,0034	1	2
GO:0017136	MF	NAD-dependent histone deacetylase activity	15	1	0,0036	1	2
GO:0019894	MF	kinesin binding	39	1	0,0027	1	3
GO:0022011	BP	myelination in peripheral nervous system	25	1	0,0029	1	3
GO:0032292	BP	peripheral nervous system axon ensheathment	25	1	0,0029	1	3
GO:0014044	BP	Schwann cell development	27	1	0,0030	1	3
GO:0090307	BP	mitotic spindle assembly	62	1	0,0038	1	3
GO:0014037	BP	Schwann cell differentiation	35	1	0,0040	1	3
GO:0051496	BP	positive regulation of stress fiber assembly	52	1	0,0049	1	3
GO:0051298	BP	centrosome duplication	69	1	0,0049	1	3

Fonte: Elaborada pelo autor (2022)

Tabela 8 – Tabela com as 10 primeiras vias metabólicas para os 10 subgrupos encontrados pelo algoritmo Beam search.

ID	descrição	número de genes	genes diferencialmente metilados	p-valor	FDR	subgrupo
path:hsa05332	Graft-versus-host disease	27	1	0,0559	1	1
path:hsa05330	Allograft rejection	30	1	0,0609	1	1
path:hsa05320	Autoimmune thyroid disease	34	1	0,0702	1	1
path:hsa00051	Fructose and mannose metabolism	32	1	0,0742	1	1
path:hsa04940	Type I diabetes mellitus	34	1	0,0751	1	1
path:hsa00565	Ether lipid metabolism	46	1	0,0927	1	1
path:hsa00600	Sphingolipid metabolism	47	1	0,0968	1	1
path:hsa05150	Staphylococcus aureus infection	63	1	0,1059	1	1
path:hsa04612	Antigen processing and presentation	54	1	0,1219	1	1
path:hsa05416	Viral myocarditis	52	1	0,1272	1	1
path:hsa05332	Graft-versus-host disease	27	1	0,0543	1	2
path:hsa05330	Allograft rejection	30	1	0,0593	1	2
path:hsa05320	Autoimmune thyroid disease	34	1	0,0685	1	2
path:hsa04940	Type I diabetes mellitus	34	1	0,0740	1	2
path:hsa00565	Ether lipid metabolism	46	1	0,0899	1	2
path:hsa00600	Sphingolipid metabolism	47	1	0,0942	1	2
path:hsa04962	Vasopressin-regulated water reabsorption	44	1	0,0987	1	2
path:hsa05150	Staphylococcus aureus infection	63	1	0,1020	1	2
path:hsa04612	Antigen processing and presentation	54	1	0,1198	1	2
path:hsa05416	Viral myocarditis	52	1	0,1245	1	2
path:hsa05332	Graft-versus-host disease	27	1	0,0551	1	3
path:hsa05330	Allograft rejection	30	1	0,0598	1	3
path:hsa05320	Autoimmune thyroid disease	34	1	0,0686	1	3
path:hsa04940	Type I diabetes mellitus	34	1	0,0739	1	3
path:hsa00565	Ether lipid metabolism	46	1	0,0908	1	3

Fonte: Elaborada pelo autor (2022)

Tabela 9 – Tabela com as 10 primeiras vias metabólicas para os 10 subgrupos encontrados pelo algoritmo SSDP+ para a métrica de qualidade Qg.

ID	descrição	número de genes	genes diferencialmente	p-valor	FDR	subgrupo
path:hsa04072	Phospholipase D signaling pathway	144	1	0,0200	1	1
path:hsa05130	Pathogenic Escherichia coli infection	181	1	0,0215	1	1
path:hsa04360	Axon guidance	179	1	0,0270	1	1
path:hsa04015	Rap1 signaling pathway	208	1	0,0277	1	1
path:hsa04810	Regulation of actin cytoskeleton	209	1	0,0283	1	1
path:hsa04151	PI3K-Akt signaling pathway	324	1	0,0440	1	1
path:hsa05200	Pathways in cancer	501	1	0,0654	1	1
path:hsa00010	Glycolysis / Gluconeogenesis	60	0	1	1	1
path:hsa00020	Citrate cycle (TCA cycle)	28	0	1	1	1
path:hsa00030	Pentose phosphate pathway	27	0	1	1	1
path:hsa04974	Protein digestion and absorption	91	1	0,0301	1	2
path:hsa04613	Neutrophil extracellular trap formation	158	1	0,0407	1	2
path:hsa04371	Apelin signaling pathway	134	1	0,0446	1	2
path:hsa05034	Alcoholism	169	1	0,0466	1	2
path:hsa05203	Viral carcinogenesis	188	1	0,0627	1	2
path:hsa05206	MicroRNAs in cancer	265	1	0,0822	1	2
path:hsa05210	Colorectal cancer	84	0	1	1	2
path:hsa00010	Glycolysis / Gluconeogenesis	60	0	1	1	2
path:hsa00020	Citrate cycle (TCA cycle)	28	0	1	1	2
path:hsa00030	Pentose phosphate pathway	27	0	1	1	2
path:hsa00010	Glycolysis / Gluconeogenesis	60	0	1	1	3
path:hsa00020	Citrate cycle (TCA cycle)	28	0	1	1	3
path:hsa00030	Pentose phosphate pathway	27	0	1	1	3
path:hsa00040	Pentose and glucuronate interconversions	27	0	1	1	3
path:hsa00051	Fructose and mannose metabolism	32	0	1	1	3
path:hsa00052	Galactose metabolism	28	0	1	1	3

Fonte: Elaborada pelo autor (2022)

5 DISCUSSÃO

Nesse estudo se trabalhou a hipótese de encontrar grupos de pacientes que através da análises de suas metilações e características seria possível detectar o câncer de mama em seus estágios iniciais. Para isso foram reutilizados dados de metilação de DNA obtidos de amostras de sangue periférico para pacientes saudáveis e com câncer de mama através do banco de dados público GEO. Na análise desses dados foram identificadas CpGs diferencialmente metiladas entre pacientes negativos para câncer e positivos para câncer de mama para identificação de subgrupos de pacientes. As probes diferencialmente expressas foram submetidas aos algoritmos de descoberta de subgrupo SSDP+ e *Beam search*, observou-se que dentro dos subgrupos de CpGs encontrados é importante notar que alguns dos genes relacionados a essas CpGs já foram reportados por ter alguma relação com câncer de mama em estudos prévios.

5.1 SUBGRUPOS ENCONTRADOS PELO BEAM SEARCH

Para exemplificar o potencial dos genes englobados nos subgrupos encontrados pelo algoritmo *Beam search* das CpGs comuns a todos os subgrupos os genes com maior histórico bibliográfico relacionado ao câncer de mama foram CD200, RNF5, FEN1, AF1Q, SPHK1, HTRA1, FPR1, VANGL2, MATN3, TNFRSF18, LPCAT1, CKAP5 (Tabela 10 e 11). Existem diversos artigos indicando a diferença na expressão do gene CD200 em camundongos associado a metástase sendo descrito como um possível biomarcador imunoterápico (PODNOS et al., 2012; ERIN et al., 2015; CURRY et al., 2017). Enquanto que para pacientes com câncer de mama o gene CD200 aparece super expresso em pacientes com metástase quando comparado a pacientes sem metástase e em pacientes com recidiva em comparação aos pacientes sem progressão (VEER et al., 2002; WANG et al., 2005; MOREAUX et al., 2008). O gene CD200 está relacionado aos termos GO que se refere a regulação da resposta neuro inflamatória (GO:0150077), regulação da atividade da glutamina gama glutamil transferase (GO:0150073) e a regulação da atividade da enzima arginase (GO:0150072). Onde um alto nível dessa enzima já foi observado no soro de pacientes com câncer de mama, onde quanto mais avançado o câncer de mama, maior o nível dessa enzima no soro o que pode servir como um biomarcador biológico para indicar progressão do câncer de mama (POLAT et al., 2003; POREMBSKA et al., 2003). Enquanto que pacientes com câncer de mama com alta concentração de gama glutamil

transferase foram associados a uma baixa sobrevivência (STAUDIGL et al., 2015).

Tabela 10 – Tabela de genes hipermetilados presentes nas regiões de metilação dos subgrupos encontrados pelo algoritmo *Beam search*.

Subgrupo	Genes hipermetilados
1	C20orf165,NOM1,TSEN2,NXN,SLC38A7,FPR1,LPCAT1,MKL2,WSB2,HLA-B
2	C20orf165,NOM1,TSEN2,NXN,FPR1,LPCAT1,MKL2,NSF,WSB2,SLC38A7,HLA-B
3	C20orf165,NOM1,TSEN2,NXN,SLC38A7,FPR1,LPCAT1,MKL2,WSB2,HLA-B
4	C20orf165,NOM1,TSEN2,NXN,SLC38A7,FPR1,LPCAT1,MKL2,WSB2,HLA-B
5	C20orf165,NOM1,TSEN2,NXN,SLC38A7,FPR1,LPCAT1,MKL2,NSF,WSB2,HLA-B
6	C20orf165,NOM1,TSEN2,NXN,SLC38A7,FPR1,LPCAT1,MKL2,WSB2,HLA-B
7	C20orf165,NOM1,TSEN2,NXN,SLC38A7,FPR1,LPCAT1,MKL2,WSB2,HLA-B
8	C20orf165,NOM1,TSEN2,NXN,SLC38A7,FPR1,LPCAT1,MKL2,WSB2,HLA-B
9	C20orf165,NOM1,TSEN2,NXN,SLC38A7,FPR1,LPCAT1,MKL2,WSB2,HLA-B
10	C20orf165,NOM1,TSEN2,NXN,SLC38A7,FPR1,LPCAT1,MKL2,NSF,WSB2,HLA-B

Fonte: Elaborada pelo autor (2022)

O gene RNF5 tem como função de regular o controle da diferenciação celular, crescimento e sua desregulação já foi associada a tumorigênese (FANG et al., 2003). Em um estudo de Bromberg et al. foi observado uma alta expressão do gene RNF5 em células tumorais de câncer de mama e linhagens celulares relacionadas, seu alto nível de expressão foi associado a uma menor sobrevida para espécimes celulares de câncer de mama humano. Também foi observado níveis altos de RNF5 em espécimes celulares de melanoma metastático, leucemia e tumores de ovário e renal, sugerindo ser um evento comum durante o processo de progressão tumoral (BROMBERG et al., 2007). Um estudo mais recente de Wang et al. foi observada a importância do gene RNF5 na degradação da proteína PHGDH prevenindo a proliferação de células cancerígenas, onde o gene PHGDH já foi descrito como significativamente alto expresso em amostras de câncer de mama e associado a uma baixa expressão de RNF5 (WANG et al., 2020). A inibição de RNF5 em ratos com câncer de mama promoveu um crescimento tumoral e resistência ao tratamento por paclitaxel, além disso foi demonstrado a função do RNF5 na regulação das proteínas SLC1A5 e SLC38A2 para o metabolismo da glutamina que é um elemento chave no desenvolvimento tumoral e resposta a terapias (JEON et al., 2015).

Em um trabalho de Wang et al. foi apresentado dados que suportam que a alta expressão de FEN1 tem uma relação inversa com a sobrevivência de pacientes com câncer de mama e que em tratamentos a sua expressão é induzida na célula cancerígena resultando em uma maior resistência a drogas, tendo assim pacientes com baixa expressão de FEN1 uma alta taxa de sobrevivência (WANG et al., 2015). Além disso, o gene FEN1 já tiveram polimorfismos que aumentam a susceptibilidade ao câncer de mama (LV et al., 2014) e também já foi alvo de estudos com cúrcuma para a inibição da proliferação do câncer de mama, sendo assim

Tabela 11 – Tabela de genes hipometilados presentes nas regiões de metilação dos subgrupos encontrados pelo algoritmo *Beam search*.

Subgrupo	Genes hipometilados
1	PUSL1,ACAP3,CD200,HTRA1,RNF5P1,RNF5,AGPAT1,NOC2L,MATN3,CKAP5,GTDC1,KATNAL1,C11orf10,FEN1,MIR611,TNFRSF18,C1orf56,MEIS2,IKZF2,NDUFS2,SPHK1,KHK,RGS16,ZNF193,FBXL12,VANGL2,FKBP10,SC65,GAPVD1
2	PUSL1,ACAP3,CD200,HTRA1,RNF5P1,RNF5,AGPAT1,NOC2L,MATN3,CKAP5,GTDC1,KATNAL1,C11orf10,FEN1,MIR611,TNFRSF18,C1orf56,MEIS2,IKZF2,NDUFS2,SPHK1,RGS16,ZNF193,FBXL12,VANGL2,FKBP10,SC65,GAPVD1
3	PUSL1,ACAP3,CD200,HTRA1,RNF5P1,RNF5,AGPAT1,NOC2L,MATN3,CKAP5,GTDC1,KATNAL1,C11orf10,FEN1,MIR611,TNFRSF18,C1orf56,MEIS2,IKZF2,NDUFS2,SPHK1,KHK,RGS16,ZNF193,NAP1L4,FBXL12,VANGL2,FKBP10,SC65,GAPVD1
4	PUSL1,ACAP3,CD200,HTRA1,RNF5P1,RNF5,AGPAT1,NOC2L,MATN3,CKAP5,GTDC1,KATNAL1,C11orf10,FEN1,MIR611,TNFRSF18,C1orf56,MEIS2,IKZF2,NDUFS2,SPHK1,RGS16,ZNF193,LRRC23,FBXL12,VANGL2,FKBP10,SC65,GAPVD1
5	PUSL1,ACAP3,CD200,HTRA1,RNF5P1,RNF5,AGPAT1,NOC2L,MATN3,CKAP5,GTDC1,KATNAL1,C11orf10,FEN1,MIR611,TNFRSF18,C1orf56,MEIS2,IKZF2,NDUFS2,SPHK1,RGS16,ZNF193,NAP1L4,FBXL12,VANGL2,FKBP10,SC65,GAPVD1
6	PUSL1,ACAP3,CD200,HTRA1,RNF5P1,RNF5,AGPAT1,NOC2L,MATN3,CKAP5,GTDC1,KATNAL1,C11orf10,FEN1,MIR611,TNFRSF18,C1orf56,MEIS2,IKZF2,NDUFS2,SPHK1,RGS16,ZNF193,LRRC23,NAP1L4,FBXL12,VANGL2,FKBP10,SC65,GAPVD1
7	PUSL1,ACAP3,CD200,HTRA1,RNF5P1,RNF5,AGPAT1,NOC2L,MATN3,CKAP5,GTDC1,KATNAL1,C11orf10,FEN1,MIR611,TNFRSF18,C1orf56,MEIS2,IKZF2,NDUFS2,ECHDC1,SPHK1,KHK,RGS16,ZNF193,NAP1L4,FBXL12,VANGL2,FKBP10,SC65,GAPVD1
8	PUSL1,ACAP3,CD200,HTRA1,RNF5P1,RNF5,AGPAT1,NOC2L,MATN3,CKAP5,GTDC1,KATNAL1,C11orf10,FEN1,MIR611,TNFRSF18,C1orf56,MEIS2,IKZF2,NDUFS2,ECHDC1,SPHK1,KHK,RGS16,ZNF193,FBXL12,VANGL2,FKBP10,SC65,GAPVD1
9	PUSL1,ACAP3,CD200,HTRA1,RNF5P1,RNF5,AGPAT1,NOC2L,MATN3,CKAP5,GTDC1,KATNAL1,C11orf10,FEN1,MIR611,TNFRSF18,C1orf56,MEIS2,IKZF2,NDUFS2,ECHDC1,SPHK1,RGS16,ZNF193,LRRC23,FBXL12,VANGL2,FKBP10,SC65,GAPVD1
10	PUSL1,ACAP3,CD200,HTRA1,RNF5P1,RNF5,AGPAT1,NOC2L,MATN3,CKAP5,GTDC1,KATNAL1,C11orf10,FEN1,MIR611,TNFRSF18,C1orf56,MEIS2,IKZF2,NDUFS2,ECHDC1,SPHK1,RGS16,ZNF193,FBXL12,VANGL2,FKBP10,SC65,GAPVD1

Fonte: Elaborada pelo autor (2022)

considerado um potencial biomarcador para câncer de mama (CHEN et al., 2014; ZOU et al., 2018). O gene C1orf56 também conhecido como AF1Q quando em alta expressão em pacientes com câncer de mama caracterizou um pobre prognóstico ao promover metástase (LI et al., 2006; CHANG et al., 2008; PARK et al., 2015; PARK et al., 2019). O AF1Q também já foi caracterizado como biomarcador de pobre prognóstico em pacientes com leucemia (TSE et al., 2004), síndrome mielodisplásica (TSE et al., 2005) e câncer de ovário (TIBERIO et al., 2012).

O gene SPHK1 é comumente visto altamente regulado em células tumorais de pacientes com câncer de mama, já descrito em maior concentração em pacientes do tipo triplo negativo (DATTA et al., 2014; WANG et al., 2018). SPHK1 já foi vinculado a um prognóstico pobre e progressão da doença, possivelmente causando resistência a terapia, sendo assim um potencial biomarcador para redução do crescimento tumoral e metástase (GEFFKEN; SPIEGEL, 2018; NAGAHASHI et al., 2018). Além disso, o gene SPHK1 também está relacionado a regulação da

atividade da enzima arginase (GO:0150072), assim como o gene CD200.

O gene HTRA1 já foi descrito ter associação estatisticamente significativa quando em baixa expressão em pacientes com câncer de mama em um estágio tumoral mais avançado, além disso, a baixa expressão de HTRA1 parece estar associada a características clínicas mais agressivas como alto nível de linfonodos afetados, também já descrito para outros tipos de câncer como câncer de pulmão e melanoma maligno (BALDI et al., 2002; ESPOSITO et al., 2006; LEHNER et al., 2013). No estudo de Vecchi et al. foi observada que a inibição do FPR1 que é encontrado com alta expressão em pacientes com câncer de mama diminui o crescimento e metástases tumorais em camundongos (VECCHI et al., 2018). Para o gene FPR1 já foi descrito um polimorfismo que causa perda de função que afeta negativamente a sobrevida de pacientes com câncer de mama e também câncer de cólon e reto tratados por quimioterapia (VACCHELLI et al., 2016; SZTUPINSZKI et al., 2021). O gene FPR1 já foi descrito ter níveis significantes em células tumorais de diferentes tipos de câncer como glioblastoma, neuroblastoma e câncer de bexiga (AHMET et al., 2020). O gene VANGL2 é necessário para o desenvolvimento normal das glândulas mamárias e quando estudado usando cultura de células de câncer do tipo triplo negativo foi encontrado altamente expresso estando implicado na tumorigênese e migração celular, foi posteriormente descrito altamente expresso em pacientes com câncer de mama, estando associado a um pobre diagnóstico e envolvido no crescimento tumoral (BORG et al., 2013; HATAKEYAMA et al., 2014; PUVIRAJESINGHE et al., 2016; SMITH et al., 2019).

Existem poucos trabalhos discutindo a influência do gene MATN3 no câncer de mama, o trabalho de Post et al. associou os altos níveis de sua expressão com um melhor prognóstico dos pacientes, enquanto que Gu et al. estudando as influências da idade na promoção da progressão da doença viu que o silenciamento do MATN3 e outros genes inibiam a proliferação das células cancerígenas (POST et al., 2019; GU et al., 2020). A alta expressão do gene MATN3 também já foi associado a um pobre prognóstico em pacientes com câncer gástrico estando relacionado ao crescimento e progressão tumoral (LI et al., 2021; CHEN et al., 2021). No estudo de Roberti et al. com camundongos analisando mudanças durante a metástase em modelos com câncer de mama do tipo triplo negativo, a proteína TNFRSF18 foi encontrada com baixa regulação em células metastáticas (ROBERTI et al., 2012). Os trabalhos de Peng et al. e Zhu et al. desenvolveram uma lista de CpGs para a predição do prognóstico de pacientes com câncer de mama do tipo triplo negativo e outra lista de genes relacionados a imunidade de pacientes com câncer de mama respectivamente, onde o gene TNFRSF18 esteve presente diferencialmente expresso nas listas dos dois grupos (PENG et al., 2020; ZHU et al., 2020). O gene TNFRSF18

também já foi discutido em diferentes tipos de câncer como o câncer endometrial (WANG et al., 2018) e mieloma (LIU et al., 2011).

Ao estudar o gene LPCAT1, Abdelzaher et al. identificou que a super expressão de LPCAT1 foi significativamente associada ao aumento da atividade proliferativa para pacientes com status negativo de receptor de estrogênio (ER) e receptor de progesterona (PR), status positivo do receptor 2 do fator de crescimento epidérmico humano (HER2), bem como para os subtipos moleculares triplo-negativo e HER2, concluindo que o LPCAT1 está implicado na patogênese, evolução e progressão do câncer de mama e estar envolvido na invasão local e metástase do câncer de mama (ABDELZAHER; MOSTAFA, 2015). Recentemente, Lebok et al. concluiu que a super expressão de LPCAT1 está ligado a características tumorais adversas e prognóstico ruim em pacientes com câncer de mama (LEBOK et al., 2019). Além disso, o LPCAT1 já foi descrito para câncer de cólon e reto (MANSILLA et al., 2009), câncer hepático (ZHANG et al., 2020a), câncer de próstata (HAN et al., 2020) e leucemia (WANG et al., 2021). Analisando os efeitos adversos do tratamento usando Ado-trastuzumab emtansine (T-DM1) em pacientes com câncer de mama HER2-positivo, os estudos de Endo et al. determinou que a interação entre a porção DM1 e o gene CKAP5 tem um grande potencial para ser um fator que contribui na dose limite para toxicidade em pacientes tratados com T-DM1 (ENDO et al., 2018; ENDO et al., 2021). Além disso, CKAP5 já foi descrito ter importantes interações na migração e proliferação de glioma e na regulação da dinâmica dos microtúbulos influenciando a tumorigênese (HUANG et al., 2018; ROSSI et al., 2021).

5.2 SUBGRUPOS ENCONTRADOS PELO SSDP+

Enquanto que para o grupo mais significativo encontrado pelo SSDP+ temos apenas dois genes LPAR5 e PLXNC1 (Tabela 12). Existem poucos estudos trabalhando o gene LPAR5 e o câncer de mama, mas sua alta expressão foi associada a uma alta taxa de metástase linfática, progressão tumoral e uma baixa taxa de sobrevivência (ZHENG et al., 2020). Para o gene PLXNC1 não foi encontrado nenhum artigo estudando sua relação com câncer de mama, mas ele foi descrito recentemente em outros tipos como câncer de tireoide, câncer gástrico e câncer de estômago (LI et al., 2019; CHEN et al., 2020; NI et al., 2021). O subgrupo 8 encontrado pelo SSDP+ foi o único que teve alguma via metabólica significativa, os genes presentes nesse grupo foram VAMP1 e ERO1L. O gene ERO1L não possui muitos artigos estudando sua relação com câncer de mama, porém em um estudo sobre a desregulação da

autofagia na progressão do câncer de mama, esse gene foi um dos genes significativamente associados a sobrevida dos pacientes (GU et al., 2016). O gene ERO1L também já foi associado a câncer de pâncreas e de pulmão (ZHANG et al., 2020b; LEI et al., 2020). A via metabólica de interações SNARE no transporte vesicular (path:hsa04130) foi significativa ($P < 0,01$) para o subgrupo 8 encontrado pelo algoritmo SSDP+ e apesar de não termos encontrado nenhum artigo discutindo sua relação com o câncer de mama essa via metabólica é estudada por ser um alvo em potencial para o desenvolvimento de terapias anti-metastáticas (MENG; WANG, 2015; GORSHTAIN; GRAFINGER; COPPOLINO, 2021). A outra via metabólica que foi significativa para o subgrupo 8 foi a relacionada a infecção por *Vibrio cholerae* (path:hsa05110), onde o único estudo encontrado relacionando essa bactéria ao câncer de mama fala sobre o uso da enzima secretado por ela para induzir apoptose das células cancerígenas e regredir o crescimento tumoral usando modelos de camundongos (RAY; PAL, 2016).

Tabela 12 – Tabela de genes hiper-metilados e hipo-metilados presentes nas regiões de metilação dos subgrupos encontrados pelo algoritmo SSDP+.

Subgrupo	Genes hipermetilados	Genes hipometilados
1	LPAR5, PLXNC1	-
2	HDAC4, ST7, ST7OT4, ST7OT1	COL18A1
3	ARHGEF10	-
4	-	-
5	SFRS11, LRRC40, SLC45A4, ZNF550	-
6	EIF2AK2, ETV5, ST7, ST7OT4, ST7OT1	-
7	GNL1, PRR3, LPAR5	AVL9, LSM5
8	ERO1L	VAMP1
9	RUNX3, ARHGEF10, TLX1, TLX1NB	COL18A1
10	-	KCNIP2

Fonte: Elaborada pelo autor (2022)

De acordo com nosso conhecimento até aqui, apenas um trabalho utilizou algoritmos de descoberta de subgrupos relacionados a uma longa ou curta sobrevivência em dados epidemiológicos de pacientes com câncer de mama (PARK; PARK; YOO, 2019). Nosso trabalho é o primeiro estudo a utilizar algoritmos de descoberta de subgrupos em dados de metilação de DNA para descoberta de grupos específicos de pacientes. Através dos grupos de genes pertencentes as regiões de metilações englobadas nos subgrupos encontrados é possível perceber a capacidade que esses algoritmos possuem de caracterizar grupos de pacientes e de junto a isso trazer associações entre possíveis biomarcadores. Para os dados que foram utilizados nesse trabalho o algoritmo *Beam search* demonstrou obter resultados mais interessantes quando comparado aos subgrupos obtidos pelo SSDP+, já que os subgrupos puros englobaram mais regiões e mais pacientes, obtendo assim genes que tinham um relevância prévia com cân-

cer de mama. Os principais pontos fortes desse estudo foram, primeiro usar dados coletados previamente a descoberta do diagnóstico de câncer de mama e assim tentar criar painéis de biomarcadores que possam indicar a susceptibilidade a essa doença, sendo esse método passível de ser utilizado para outros tipos de doenças. Segundo por mostrar como algoritmos de descoberta subgrupos podem facilitar e ajudar a encontrar aglomerados de pacientes com atributos similares e informativos sem a necessidade de hipóteses prévias como frequentemente é feito em grandes estudos populacionais.

6 CONCLUSÃO

A descoberta do câncer de mama hoje depende de métodos invasivos como a biópsia de tecido tumoral que normalmente é feito em um estágio mais avançado do câncer. Enquanto que os métodos de imagem são restritos ao tamanho e volume do tecido tumoral para a análise. Existem diversos biomarcadores já propostos para câncer de mama, mas um diagnóstico precoce desses pacientes ainda é um desafio. No nosso estudo foram encontrados grupos de pacientes que através da análises de suas metilações seria possível detectar o câncer de mama em seus estágios iniciais. Para isso foram reutilizados dados de metilação de DNA obtidos de amostras de sangue periférico para pacientes com diagnóstico negativo para câncer e com câncer de mama através do banco de dados público GEO.

Na análise desses dados foram identificadas CpGs diferencialmente metiladas entre pacientes com e sem câncer de mama auxiliando na identificação de subgrupos de pacientes. As sondas diferencialmente expressas foram submetidas aos algoritmos de descoberta de subgrupo SSDP+ e *Beam search*. Foram selecionados 20 subgrupos, sendo 10 subgrupos encontrados pelo *Beam search* e 10 pelo SSDP+. Onde o algoritmo *Beam search* encontrou subgrupos com mais elementos e alto grau de redundância englobando diferentes genes relacionados previamente ao câncer de mama, enquanto o SSDP+ encontrou subgrupos com menos elementos e alto grau de diversidade, mas que por englobar menos regiões não tiveram genes previamente relacionados a câncer de mama.

De acordo com nosso conhecimento até aqui, esse trabalho é o primeiro estudo a utilizar algoritmos de descoberta de subgrupos em dados de metilação de DNA para descoberta de grupos específicos de pacientes. Nossos achados mostram que é possível utilizar algoritmos de descoberta de subgrupos para detecção de subgrupos específicos de pacientes de câncer de mama usando dados de metilação de DNA obtidos do sangue periférico. Uma vez validados, nossos genes identificados podem ser usados para melhorar a estratificação de risco para a prevenção do câncer de mama e para o desenvolvimento de novas estratégias de prevenção e tratamento do câncer de mama. Além disso, a metodologia desse trabalho pode ser reaplicada em outros tipos de doenças ou de dados moleculares para descoberta de grupos de pacientes com diferentes sintomas dentro de uma mesma doença. Com isso, através desse estudo foi possível mostrar a capacidade que as ferramentas de descoberta de subgrupos podem trazer para os estudos clínicos baseados em testes moleculares.

6.1 CONTRIBUIÇÕES

Os códigos usados para analisar os dados estão disponíveis no Github (<<https://github.com/geocarvalho/master-degree>>), assim como os dados obtidos podem ser acessados no Google Drive (<<https://bit.ly/3tPYFjb>>).

6.2 LIMITAÇÕES E TRABALHOS FUTUROS

Existem alguns limitações no nosso trabalho. Esse trabalho é uma reutilização de dados obtidos no banco dados GEO, assim se faz necessário um estudo clínico para validar os subgrupos obtidos para validar a aplicação desse modelo, dado que nem todos os genes englobados nos subgrupos possuem uma bibliografia prévia relacionada ao câncer de mama. Outra limitação foi a quantidade de características fenotípicas disponíveis para esses dados, com mais atributos existe a possibilidade de encontrar subgrupos mais específicos englobando tanto dados moleculares quanto dados clínicos, demográficos e sociais. Assim, seria interessante testar a mesma abordagem em outras amostras de dados de metilação para pacientes com câncer de mama com o intuito de ver se os subgrupos encontrados nesse trabalho também estarão presentes em outras amostragens, além disso a metodologia aplicada no nosso estudo pode ser reutilizada em diferentes tipos de doenças. Com os resultados obtidos pelos subgrupos é possível utilizar biologia de sistemas para entender como os genes presentes nas CpGs encontradas interagem. Existem a possibilidade de trazer atualizações ao algoritmo SSDP+ para maximizar o tamanho dos subgrupos puros encontrados a nível de elementos e de amostras.

REFERÊNCIAS

- ABDELZAHER, E.; MOSTAFA, M. F. Lysophosphatidylcholine acyltransferase 1 (lpcat1) upregulation in breast carcinoma contributes to tumor progression and predicts early tumor recurrence. *Tumor Biology*, Springer, v. 36, n. 7, p. 5473–5483, 2015.
- AHMET, D. S.; BASHEER, H. A.; SALEM, A.; LU, D.; AGHAMOHAMMADI, A.; WEYERHÄUSER, P.; BORDIGA, A.; ALMENIAWI, J.; RASHID, S.; COOPER, P. A. et al. Application of small molecule fpr1 antagonists in the treatment of cancers. *Scientific reports*, Nature Publishing Group, v. 10, n. 1, p. 1–12, 2020.
- ARYEE, M. J.; JAFFE, A. E.; CORRADA-BRAVO, H.; LADD-ACOSTA, C.; FEINBERG, A. P.; HANSEN, K. D.; IRIZARRY, R. A. Minfi: a flexible and comprehensive bioconductor package for the analysis of infinium dna methylation microarrays. *Bioinformatics*, Oxford University Press, v. 30, n. 10, p. 1363–1369, 2014.
- BALDI, A.; LUCA, A. D.; MORINI, M.; BATTISTA, T.; FELSANI, A.; BALDI, F.; CATRICALA, C.; AMANTEA, A.; NOONAN, D. M.; ALBINI, A. et al. The htra1 serine protease is down-regulated during human melanoma progression and represses growth of metastatic melanoma cells. *Oncogene*, Nature Publishing Group, v. 21, n. 43, p. 6684–6688, 2002.
- BAUMEISTER, J.; ATZMUELLER, M.; PUPPE, F. Introspective subgroup analysis for interactive knowledge refinement. In: *International Florida Artificial Intelligence Research Society Conference (FLAIRS)*, Springer. [S.l.: s.n.], 2006. p. 402–407.
- BENJAMINI, Y.; HOCHBERG, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, Wiley Online Library, v. 57, n. 1, p. 289–300, 1995.
- BIBIKOVA, M.; BARNES, B.; TSAN, C.; HO, V.; KLOTZLE, B.; LE, J. M.; DELANO, D.; ZHANG, L.; SCHROTH, G. P.; GUNDERSON, K. L. et al. High density dna methylation array with single cpg site resolution. *Genomics*, Elsevier, v. 98, n. 4, p. 288–295, 2011.
- BOLLEPALLI, S.; KORHONEN, T.; KAPRIO, J.; ANDERS, S.; OLLIKAINEN, M. Epismoker: a robust classifier to determine smoking status from dna methylation data. *Epigenomics*, Future Medicine, v. 11, n. 13, p. 1469–1486, 2019.
- BONINI, C.; MONDINO, A. Adoptive t-cell therapy for cancer: The era of engineered t cells. *European journal of immunology*, Wiley Online Library, v. 45, n. 9, p. 2457–2469, 2015.
- BORG, J.-P.; BELOTTI, E.; DAULAT, A.; LEMBO, F.; BERTUCCI, F.; CHARAFE-JAUFFRET, E.; BIRNBAUM, D. *Deregulation of the non-canonical pathway in triple-negative breast cancer*. [S.l.]: Wiley Online Library, 2013.
- BRASIL, I. E. incidência de câncer no brasil| inca-instituto nacional de câncer. 2019. 2020.
- BREITLING, L. P.; SAUM, K.-U.; PERNA, L.; SCHÖTTKER, B.; HOLLECZEK, B.; BRENNER, H. Frailty is associated with the epigenetic clock but not with telomere length in a german cohort. *Clinical epigenetics*, BioMed Central, v. 8, n. 1, p. 1–8, 2016.

BRITO-SILVA, K.; BEZERRA, A. F. B.; CHAVES, L. D. P.; TANAKA, O. Y. Integralidade no cuidado ao câncer do colo do útero: avaliação do acesso. *Revista de Saúde Pública, SciELO Public Health*, v. 48, p. 240–248, 2014.

BROMBERG, K. D.; KLUGER, H. M.; DELAUNAY, A.; ABBAS, S.; DIVITO, K. A.; KRAJEWSKI, S.; RONAI, Z. Increased expression of the e3 ubiquitin ligase rnf5 is associated with decreased survival in breast cancer. *Cancer research, AACR*, v. 67, n. 17, p. 8172–8179, 2007.

CAINI, S.; FIORITO, G.; PALLI, D.; BENDINELLI, B.; POLIDORO, S.; SILVESTRI, V.; OTTINI, L.; AMBROGETTI, D.; ZANNA, I.; SAIEVA, C. et al. Pre-diagnostic dna methylation patterns differ according to mammographic breast density amongst women who subsequently develop breast cancer: a case-only study in the epic-florence cohort. *Breast Cancer Research and Treatment, Springer*, p. 1–10, 2021.

CAPPER, D.; JONES, D. T.; SILL, M.; HOVESTADT, V.; SCHRIMPF, D.; STURM, D.; KOELSCHE, C.; SAHM, F.; CHAVEZ, L.; REUSS, D. E. et al. Dna methylation-based classification of central nervous system tumours. *Nature, Nature Publishing Group*, v. 555, n. 7697, p. 469–474, 2018.

CHANG, X.-Z.; LI, D.-Q.; HOU, Y.-F.; WU, J.; LU, J.-S.; DI, G.-H.; JIN, W.; OU, Z.-L.; SHEN, Z.-Z.; SHAO, Z.-M. Identification of the functional role of af1q in the progression of breast cancer. *Breast cancer research and treatment, Springer*, v. 111, n. 1, p. 65–78, 2008.

CHEN, B.; ZHANG, Y.; WANG, Y.; RAO, J.; JIANG, X.; XU, Z. Curcumin inhibits proliferation of breast cancer cells through nrf2-mediated down-regulation of fen1 expression. *The Journal of steroid biochemistry and molecular biology, Elsevier*, v. 143, p. 11–18, 2014.

CHEN, J.; LIU, H.; CHEN, J.; SUN, B.; WU, J.; DU, C. Plxnc1 enhances carcinogenesis through transcriptional activation of il6st in gastric cancer. *Frontiers in oncology, Frontiers*, v. 10, p. 33, 2020.

CHEN, Q.; TAN, Y.; ZHANG, C.; ZHANG, Z.; PAN, S.; AN, W.; XU, H. *Frontiers in Oncology, Frontiers*, v. 11, p. 221, 2021.

CHEN, X.; ZHANG, J.; DAI, X. Dna methylation profiles capturing breast cancer heterogeneity. *BMC genomics, Springer*, v. 20, n. 1, p. 1–12, 2019.

CHEN, Y.-a.; LEMIRE, M.; CHOUFANI, S.; BUTCHER, D. T.; GRAFODATSKAYA, D.; ZANKE, B. W.; GALLINGER, S.; HUDSON, T. J.; WEKSBERG, R. Discovery of cross-reactive probes and polymorphic cpgs in the illumina infinium humanmethylation450 microarray. *Epigenetics, Taylor & Francis*, v. 8, n. 2, p. 203–209, 2013.

COSTA, P. V. D. P.; SILVA, J. M. L. da; SILVA, L. C. da; ROCHA, S. G.; COSTA, L. F. da; SOARES, M. K. M.; SANTOS, S. R. P. dos; LIMA, P. da C.; RODRIGUES, B. d. N. B.; ALMEIDA, M. da S. et al. A educação em saúde como ferramenta no combate ao câncer de mama: relato de experiência. *Research, Society and Development*, v. 9, n. 10, p. e6389108912–e6389108912, 2020.

CURRY, A.; KHATRI, I.; KOS, O.; ZHU, F.; GORCZYNSKI, R. Importance of cd200 expression by tumor or host cells to regulation of immunotherapy in a mouse breast cancer model. *PLoS One, Public Library of Science San Francisco, CA USA*, v. 12, n. 2, p. e0171586, 2017.

- DATTA, A.; LOO, S. Y.; HUANG, B.; WONG, L.; TAN, S. S.; TAN, T. Z.; LEE, S.-C.; THIERY, J. P.; LIM, Y. C.; YONG, W. P. et al. Sphk1 regulates proliferation and survival responses in triple-negative breast cancer. *Oncotarget*, Impact Journals, LLC, v. 5, n. 15, p. 5920, 2014.
- DEDEURWAERDER, S.; DEFRANCE, M.; BIZET, M.; CALONNE, E.; BONTEMPI, G.; FUKS, F. A comprehensive overview of infinium humanmethylation450 data processing. *Briefings in bioinformatics*, Oxford University Press, v. 15, n. 6, p. 929–941, 2014.
- EDGAR, R.; DOMRACHEV, M.; LASH, A. E. Gene expression omnibus: Ncbi gene expression and hybridization array data repository. *Nucleic acids research*, Oxford University Press, v. 30, n. 1, p. 207–210, 2002.
- ENDO, Y.; MOHAN, N.; DOKMANOVIC, M.; WU, W. J. Mechanisms contributing to ado-trastuzumab emtansine-induced toxicities: a gateway to better understanding of adc-associated toxicities. *Antibody therapeutics*, Oxford University Press, v. 4, n. 1, p. 55–59, 2021.
- ENDO, Y.; TAKEDA, K.; MOHAN, N.; SHEN, Y.; JIANG, J.; ROTSTEIN, D.; WU, W. J. Payload of t-dm1 binds to cell surface cytoskeleton-associated protein 5 to mediate cytotoxicity of hepatocytes. *Oncotarget*, Impact Journals, LLC, v. 9, n. 98, p. 37200, 2018.
- ENNOUR-IDRISSI, K.; DRAGIC, D.; ISSA, E.; MICHAUD, A.; CHANG, S.-L.; PROVENCHER, L.; DUROCHER, F.; DIORIO, C. Dna methylation and breast cancer risk: An epigenome-wide study of normal breast tissue and blood. *Cancers*, Multidisciplinary Digital Publishing Institute, v. 12, n. 11, p. 3088, 2020.
- ERIN, N.; PODNOS, A.; TANRIOVER, G.; DUYMUŞ, Ö.; COTE, E.; KHATRI, I.; GORCZYNSKI, R. Bidirectional effect of cd200 on breast cancer development and metastasis, with ultimate outcome determined by tumor aggressiveness and a cancer-induced inflammatory response. *Oncogene*, Nature Publishing Group, v. 34, n. 29, p. 3860–3870, 2015.
- ESPOSITO, V.; CAMPIONI, M.; LUCA, A. D.; SPUGNINI, E. P.; BALDI, F.; CASSANDRO, R.; MANCINI, A.; VINCENZI, B.; GROEGER, A.; CAPUTI, M. et al. Analysis of htra1 serine protease expression in human lung cancer. *Anticancer research*, International Institute of Anticancer Research, v. 26, n. 5A, p. 3455–3459, 2006.
- FANG, S.; LORICK, K. L.; JENSEN, J. P.; WEISSMAN, A. M. Ring finger ubiquitin protein ligases: implications for tumorigenesis, metastasis and for molecular targets in cancer. In: ELSEVIER. *Seminars in cancer biology*. [S.l.], 2003. v. 13, n. 1, p. 5–14.
- FARKAS, S. *DNA methylation in the placenta and in cancer with special reference to folate transporting genes*. Tese (Doutorado) — Örebro university, 2014.
- FORTIN, J.-P.; LABBE, A.; LEMIRE, M.; ZANKE, B. W.; HUDSON, T. J.; FERTIG, E. J.; GREENWOOD, C. M.; HANSEN, K. D. Functional normalization of 450k methylation array data improves replication in large cancer studies. *Genome biology*, Springer, v. 15, n. 11, p. 1–17, 2014.
- GAMBERGER, D.; LAVRAC, N. Expert-guided subgroup discovery: Methodology and application. *Journal of Artificial Intelligence Research*, v. 17, p. 501–527, 2002.

- GEFFKEN, K.; SPIEGEL, S. Sphingosine kinase 1 in breast cancer. *Advances in biological regulation*, Elsevier, v. 67, p. 59–65, 2018.
- GORSHTEN, G.; GRAFINGER, O.; COPPOLINO, M. G. Targeting snare-mediated vesicle transport to block invadopodium-based cancer cell invasion. *Frontiers in Oncology*, Frontiers, v. 11, p. 1847, 2021.
- GU, X.; WANG, B.; ZHU, H.; ZHOU, Y.; HORNING, A. M.; HUANG, T. H.; CHEN, Y.; HOUGHTON, P.; LAI, Z.; MICHALEK, J. E. et al. Age-associated genes in human mammary gland drive human breast cancer progression. *Breast Cancer Research*, BioMed Central, v. 22, n. 1, p. 1–15, 2020.
- GU, Y.; LI, P.; PENG, F.; ZHANG, M.; ZHANG, Y.; LIANG, H.; ZHAO, W.; QI, L.; WANG, H.; WANG, C. et al. Autophagy-related prognostic signature for breast cancer. *Molecular carcinogenesis*, Wiley Online Library, v. 55, n. 3, p. 292–299, 2016.
- HAN, C.; YU, G.; MAO, Y.; SONG, S.; LI, L.; ZHOU, L.; WANG, Z.; LIU, Y.; LI, M.; XU, B. Lpcat1 enhances castration resistant prostate cancer progression via increased mrna synthesis and paf production. *PloS one*, Public Library of Science San Francisco, CA USA, v. 15, n. 11, p. e0240801, 2020.
- HANNUM, G.; GUINNEY, J.; ZHAO, L.; ZHANG, L.; HUGHES, G.; SADDA, S.; KLOTZLE, B.; BIBIKOVA, M.; FAN, J.-B.; GAO, Y. et al. Genome-wide methylation profiles reveal quantitative views of human aging rates. *Molecular cell*, Elsevier, v. 49, n. 2, p. 359–367, 2013.
- HATAKEYAMA, J.; WALD, J. H.; PRINTSEV, I.; HO, H.-Y. H.; CARRAWAY, K. L. Vangl1 and vangl2: planar cell polarity components with a developing role in cancer. *Endocrine-related cancer*, Bioscientifica Ltd, v. 21, n. 5, p. R345–R356, 2014.
- HERRERA, F.; CARMONA, C. J.; GONZÁLEZ, P.; JESUS, M. J. D. An overview on subgroup discovery: foundations and applications. *Knowledge and information systems*, Springer, v. 29, n. 3, p. 495–525, 2011.
- HORVATH, S. Dna methylation age of human tissues and cell types. *Genome biology*, BioMed Central, v. 14, n. 10, p. 1–20, 2013.
- HORVATH, S.; RITZ, B. R. Increased epigenetic age and granulocyte counts in the blood of parkinson's disease patients. *Aging (Albany NY)*, Impact Journals, LLC, v. 7, n. 12, p. 1130, 2015.
- HUANG, D.; WANG, Y.; XU, L.; CHEN, L.; CHENG, M.; SHI, W.; XIONG, H.; ZALLI, D.; LUO, S. Gli2 promotes cell proliferation and migration through transcriptional activation of arhgef16 in human glioma cells. *Journal of Experimental & Clinical Cancer Research*, BioMed Central, v. 37, n. 1, p. 1–17, 2018.
- HUANG, Y.; OLLIKAINEN, M.; MUNIANDY, M.; ZHANG, T.; DONGEN, J. V.; HAO, G.; MOST, P. J. van D.; PAN, Y.; PERVJAKOVA, N.; SUN, Y. V. et al. Identification, heritability, and relation with gene expression of novel dna methylation loci for blood pressure. *Hypertension*, Am Heart Assoc, v. 76, n. 1, p. 195–205, 2020.
- IMPARATO, A. Interactive subgroup discovery. 2012.

JAFFE, A. E.; GAO, Y.; DEEP-SOBOSLAY, A.; TAO, R.; HYDE, T. M.; WEINBERGER, D. R.; KLEINMAN, J. E. Mapping dna methylation across development, genotype and schizophrenia in the human frontal cortex. *Nature neuroscience*, Nature Publishing Group, v. 19, n. 1, p. 40–47, 2016.

JAFFE, A. E.; IRIZARRY, R. A. Accounting for cellular heterogeneity is critical in epigenome-wide association studies. *Genome biology*, BioMed Central, v. 15, n. 2, p. 1–9, 2014.

JEON, Y. J.; KHELIFA, S.; RATNIKOV, B.; SCOTT, D. A.; FENG, Y.; PARISI, F.; RULLER, C.; LAU, E.; KIM, H.; BRILL, L. M. et al. Regulation of glutamine carrier proteins by rnf5 determines breast cancer response to er stress-inducing chemotherapies. *Cancer cell*, Elsevier, v. 27, n. 3, p. 354–369, 2015.

JOO, J. E.; DOWTY, J. G.; MILNE, R. L.; WONG, E. M.; DUGUÉ, P.-A.; ENGLISH, D.; HOPPER, J. L.; GOLDFAR, D. E.; GILES, G. G.; SOUTHEY, M. C. Heritable dna methylation marks associated with susceptibility to breast cancer. *Nature communications*, Nature Publishing Group, v. 9, n. 1, p. 1–12, 2018.

KARMA, A. *Beam search algorithm*. 2021. Disponível em: <<https://www.baeldung.com/cs/beam-search>>.

KLÖSGEN, W. Explora: A multipattern and multistrategy discovery assistant. In: *Advances in knowledge discovery and data mining*. [S.l.: s.n.], 1996. p. 249–271.

KLÖSGEN, W.; MAY, M. Spatial subgroup mining integrated in an object-relational spatial database. In: SPRINGER. *European Conference on Principles of Data Mining and Knowledge Discovery*. [S.l.], 2002. p. 275–286.

LAVRAČ, N.; CESTNIK, B.; GAMBERGER, D.; FLACH, P. Decision support through subgroup discovery: three case studies and the lessons learned. *Machine Learning*, Springer, v. 57, n. 1, p. 115–143, 2004.

LEBOK, P.; HASSEL, A. von; MEINERS, J.; HUBE-MAGG, C.; SIMON, R.; HÖFLMAYER, D.; HINSCH, A.; DUM, D.; FRAUNE, C.; GÖBEL, C. et al. Up-regulation of lysophosphatidylcholine acyltransferase 1 (lpcat1) is linked to poor prognosis in breast cancer. *Aging (Albany NY)*, Impact Journals, LLC, v. 11, n. 18, p. 7796, 2019.

LEEUWEN, M. V.; KNOBBE, A. Diverse subgroup set discovery. *Data Mining and Knowledge Discovery*, Springer, v. 25, n. 2, p. 208–242, 2012.

LEHNER, A.; MAGDOLEN, V.; SCHUSTER, T.; KOTZSCH, M.; KIECHLE, M.; MEINDL, A.; SWEEP, F. C.; SPAN, P. N.; GROSS, E. Downregulation of serine protease htra1 is associated with poor survival in breast cancer. *PloS one*, Public Library of Science San Francisco, USA, v. 8, n. 4, p. e60359, 2013.

LEI, Y.; ZANG, R.; LU, Z.; ZHANG, G.; HUANG, J.; LIU, C.; WANG, Z.; MAO, S.; CHE, Y.; WANG, X. et al. Ero1l promotes il6/sil6r signaling and regulates muc16 expression to promote ca125 secretion and the metastasis of lung cancer cells. *Cell death & disease*, Nature Publishing Group, v. 11, n. 10, p. 853–853, 2020.

LEMMERICH, F.; BECKER, M. pysubgroup: Easy-to-use subgroup discovery in python. In: SPRINGER. *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. [S.l.], 2018. p. 658–662.

- LEVINE, M. E.; LU, A. T.; QUACH, A.; CHEN, B. H.; ASSIMES, T. L.; BANDINELLI, S.; HOU, L.; BACCARELLI, A. A.; STEWART, J. D.; LI, Y. et al. An epigenetic biomarker of aging for lifespan and healthspan. *Aging (Albany NY)*, Impact Journals, LLC, v. 10, n. 4, p. 573, 2018.
- LEVY, J. J.; TITUS, A. J.; PETERSEN, C. L.; CHEN, Y.; SALAS, L. A.; CHRISTENSEN, B. C. Methylnet: an automated and modular deep learning approach for dna methylation analysis. *BMC bioinformatics*, BioMed Central, v. 21, n. 1, p. 1–15, 2020.
- LI, D.; XU, J.; DONG, X.; CHEN, W.; PAN, L.; JIANG, H.; PAN, J.; HUANG, Y. Diagnostic and prognostic value of matn3 expression in gastric carcinoma: Tcga database mining. *Journal of Gastrointestinal Oncology*, AME Publications, v. 12, n. 4, p. 1374, 2021.
- LI, D.-Q.; HOU, Y.-F.; WU, J.; CHEN, Y.; LU, J.-S.; DI, G.-H.; OU, Z.-L.; SHEN, Z.-Z.; DING, J.; SHAO, Z.-M. Gene expression profile analysis of an isogenic tumour metastasis model reveals a functional role for oncogene af1q in breast cancer metastasis. *European Journal of Cancer*, Elsevier, v. 42, n. 18, p. 3274–3286, 2006.
- LI, R.; TENG, X.; ZHU, H.; HAN, T.; LIU, Q. Mir-4500 regulates plxnc1 and inhibits papillary thyroid cancer progression. *Hormones and Cancer*, Springer, v. 10, n. 4, p. 150–160, 2019.
- LIU, Y.; ZHANG, Y.; QUANG, P.; NGO, H. T.; AZAB, F.; AZAB, A. K.; MORGAN, B.; MAISO, P.; REAGAN, M.; ROMBAOA, C. P. et al. Deregulation of tnfrsf18 (gitr) through promoter cpg island methylation induces tumor proliferation in multiple myeloma. *Blood*, Elsevier, v. 118, n. 21, p. 2424, 2011.
- LOVING, V. A.; AMINOLOLAMA-SHAKERI, S.; LEUNG, J. W. Anxiety and its association with screening mammography. *Journal of Breast Imaging*, Oxford University Press US, v. 3, n. 3, p. 266–272, 2021.
- LUCAS, T.; VIMIEIRO, R.; LUDERMIR, T. Ssd⁺: A diverse and more informative subgroup discovery approach for high dimensional data. In: IEEE. *2018 IEEE Congress on Evolutionary Computation (CEC)*. [S.l.], 2018. p. 1–8.
- LV, Z.; LIU, W.; LI, D.; LIU, L.; WEI, J.; ZHANG, J.; GE, Y.; WANG, Z.; CHEN, H.; ZHOU, C. et al. Association of functional fen1 genetic variants and haplotypes and breast cancer risk. *Gene*, Elsevier, v. 538, n. 1, p. 42–45, 2014.
- MACHADO, M. X.; SOARES, D. A.; OLIVEIRA, S. B. Significados do câncer de mama para mulheres no contexto do tratamento quimioterápico. *Physis: Revista de Saúde Coletiva*, SciELO Brasil, v. 27, p. 433–451, 2017.
- MAKSIMOVIC, J.; PHIPSON, B.; OSHLACK, A. A cross-package bioconductor workflow for analysing methylation array data. *F1000Research*, Faculty of 1000 Ltd, v. 5, 2016.
- MANSILLA, F.; COSTA, K.-A. da; WANG, S.; KRUGHØFFER, M.; LEWIN, T. M.; ØRNTTOFT, T. F.; COLEMAN, R. A.; BIRKENKAMP-DEMTRÖDER, K. Lysophosphatidylcholine acyltransferase 1 (lpcat1) overexpression in human colorectal cancer. *Journal of molecular medicine*, Springer, v. 87, n. 1, p. 85–97, 2009.
- MARIONI, R. E.; SHAH, S.; MCRAE, A. F.; CHEN, B. H.; COLICINO, E.; HARRIS, S. E.; GIBSON, J.; HENDERS, A. K.; REDMOND, P.; COX, S. R. et al. Dna methylation age of blood predicts all-cause mortality in later life. *Genome biology*, BioMed Central, v. 16, n. 1, p. 1–12, 2015.

- MARIONI, R. E.; SHAH, S.; MCRAE, A. F.; RITCHIE, S. J.; MUNIZ-TERRERA, G.; HARRIS, S. E.; GIBSON, J.; REDMOND, P.; COX, S. R.; PATTIE, A. et al. The epigenetic clock is correlated with physical and cognitive fitness in the lothian birth cohort 1936. *International journal of epidemiology*, Oxford University Press, v. 44, n. 4, p. 1388–1396, 2015.
- MARTÍN, D.; ALCALÁ-FDEZ, J.; ROSETE, A.; HERRERA, F. Nicgar: A niching genetic algorithm to mine a diverse set of interesting quantitative association rules. *Information Sciences*, Elsevier, v. 355, p. 208–228, 2016.
- MENG, J.; WANG, J. Role of snare proteins in tumourigenesis and their potential as targets for novel anti-cancer therapeutics. *Biochimica et Biophysica Acta (BBA)-Reviews on Cancer*, Elsevier, v. 1856, n. 1, p. 1–12, 2015.
- MOREAUX, J.; VEYRUNE, J. L.; REME, T.; VOS, J. D.; KLEIN, B. Cd200: a putative therapeutic target in cancer. *Biochemical and biophysical research communications*, Elsevier, v. 366, n. 1, p. 117–122, 2008.
- NAGAHASHI, M.; YAMADA, A.; KATSUTA, E.; AOYAGI, T.; HUANG, W.-C.; TERRACINA, K. P.; HAIT, N. C.; ALLEGOOD, J. C.; TSUCHIDA, J.; YUZA, K. et al. Targeting the sphk1/s1p/s1pr1 axis that links obesity, chronic inflammation, and breast cancer metastasis. *Cancer research*, AACR, v. 78, n. 7, p. 1713–1725, 2018.
- NEMODA, Z.; MASSART, R.; SUDERMAN, M.; HALLETT, M.; LI, T.; COOTE, M.; CODY, N.; SUN, Z.; SOARES, C.; TURECKI, G. et al. Maternal depression is associated with dna methylation changes in cord blood t lymphocytes and adult hippocampi. *Translational psychiatry*, Nature Publishing Group, v. 5, n. 4, p. e545–e545, 2015.
- NI, Z.; HUANG, C.; ZHAO, H.; ZHOU, J.; HU, M.; CHEN, Q.; GE, B.; HUANG, Q. Plxnc1: A novel potential immune-related target for stomach adenocarcinoma. *Frontiers in cell and developmental biology*, Frontiers Media SA, v. 9, 2021.
- ONWUKA, J. U.; LI, D.; LIU, Y.; HUANG, H.; XU, J.; LIU, Y.; ZHANG, Y.; ZHAO, Y. A panel of dna methylation signature from peripheral blood may predict colorectal cancer susceptibility. *BMC cancer*, BioMed Central, v. 20, n. 1, p. 1–11, 2020.
- (OS), O. o. G. Office of S.; HEALTH, P. P. *What is Epigenetics?* 2020. Disponível em: <<https://www.cdc.gov/genomics/disease/epigenetics.htm>>. Acesso em: 15 jul. 2021.
- PARK, J.; HWANG, J. Y.; THORE, A.; KIM, S.; TOGANO, T.; HAGIWARA, S.; PARK, J. W.; TSE, W. Af1q inhibited t cell attachment to breast cancer cell by attenuating intracellular adhesion molecule-1 expression. *Journal of cancer metastasis and treatment*, NIH Public Access, v. 5, 2019.
- PARK, J.; SCHLEDERER, M.; SCHREIBER, M.; ICE, R.; MERKEL, O.; BILBAN, M.; HOFBAUER, S.; KIM, S.; ADDISON, J.; ZOU, J. et al. Af1q is a novel tcf7 co-factor which activates cd44 and promotes breast cancer metastasis. *Oncotarget*, Impact Journals, LLC, v. 6, n. 24, p. 20697, 2015.
- PARK, J. V.; PARK, S. J.; YOO, J. S. Finding characteristics of exceptional breast cancer subpopulations using subgroup mining and statistical test. *Expert Systems with Applications*, Elsevier, v. 118, p. 553–562, 2019.

- PENG, Y.; SHUI, L.; XIE, J.; LIU, S. Development and validation of a novel 15-cpg-based signature for predicting prognosis in triple-negative breast cancer. *Journal of cellular and molecular medicine*, Wiley Online Library, v. 24, n. 16, p. 9378–9387, 2020.
- PHILIP. *Central dogma of molecular biology*. 2006. Disponível em: <<https://commons.wikimedia.org/wiki/File:CentralDogma.png>>. Acesso em: 10 jul. 2021.
- PHIPSON, B.; MAKSIMOVIC, J.; OSHLACK, A. missmethyl: an r package for analyzing data from illumina's humanmethylation450 platform. *Bioinformatics*, Oxford University Press, v. 32, n. 2, p. 286–288, 2016.
- PINHO, V. F. d. S.; COUTINHO, E. S. F. Variáveis associadas ao câncer de mama em usuárias de unidades básicas de saúde. *Cadernos de Saúde Pública*, SciELO Brasil, v. 23, n. 5, p. 1061–1069, 2007.
- PODNOS, A.; CLARK, D. A.; ERIN, N.; YU, K.; GORCZYNSKI, R. M. Further evidence for a role of tumor cd200 expression in breast cancer metastasis: decreased metastasis in cd200r1ko mice or using cd200-silenced emt6. *Breast cancer research and treatment*, Springer, v. 136, n. 1, p. 117–127, 2012.
- POLAT, M. F.; TAYSI, S.; POLAT, S.; BÖYÜK, A.; BAKAN, E. Elevated serum arginase activity levels in patients with breast cancer. *Surgery today*, Springer, v. 33, n. 9, p. 655–661, 2003.
- PONTES, T.; VIMIEIRO, R.; LUDERMIR, T. B. SsdP: A simple evolutionary approach for top-k discriminative patterns in high dimensional databases. In: IEEE. *2016 5th Brazilian Conference on Intelligent Systems (BRACIS)*. [S.l.], 2016. p. 361–366.
- POREMBSKA, Z.; LUBOŃSKI, G.; CHRZANOWSKA, A.; MIELCZAREK, M.; MAGNUSKA, J.; BARAŃCZYK-KUŹMA, A. Arginase in patients with breast cancer. *Clinica chimica acta*, Elsevier, v. 328, n. 1-2, p. 105–111, 2003.
- POST, A. E.; BUSSINK, J.; SMID, M.; SWEEP, F. C.; SPAN, P. N. Downregulation of matrix gla protein is a biomarker for tamoxifen-resistant and radioresistant breast cancer. *Biomarkers in medicine*, Future Medicine, v. 13, n. 10, p. 841–850, 2019.
- PUVIRAJESINGHE, T. M.; BERTUCCI, F.; JAIN, A.; SCERBO, P.; BELOTTI, E.; AUDEBERT, S.; SEBBAGH, M.; LOPEZ, M.; BRECH, A.; FINETTI, P. et al. Identification of p62/sqstm1 as a component of non-canonical wnt vangl2-jnk signalling in breast cancer. *Nature communications*, Nature Publishing Group, v. 7, n. 1, p. 1–15, 2016.
- RAY, T.; PAL, A. Par-1 mediated apoptosis of breast cancer cells by v. cholerae hemagglutinin protease. *Apoptosis*, Springer, v. 21, n. 5, p. 609–620, 2016.
- RIBOLI, E.; HUNT, K.; SLIMANI, N.; FERRARI, P.; NORAT, T.; FAHEY, M.; CHARRONDIERE, U.; HEMON, B.; CASAGRANDE, C.; VIGNAT, J. et al. European prospective investigation into cancer and nutrition (epic): study populations and data collection. *Public health nutrition*, Cambridge University Press, v. 5, n. 6b, p. 1113–1124, 2002.
- RITCHIE, M. E.; PHIPSON, B.; WU, D.; HU, Y.; LAW, C. W.; SHI, W.; SMYTH, G. K. limma powers differential expression analyses for rna-sequencing and microarray studies. *Nucleic acids research*, Oxford University Press, v. 43, n. 7, p. e47–e47, 2015.

- ROBERTI, M. P.; ARRIAGA, J. M.; BIANCHINI, M.; QUINTÁ, H. R.; BRAVO, A. I.; LEVY, E. M.; MORDOH, J.; BARRIO, M. M. Protein expression changes during human triple negative breast cancer cell line progression to lymph node metastasis in a xenografted model in nude mice. *Cancer biology & therapy*, Taylor & Francis, v. 13, n. 11, p. 1123–1140, 2012.
- ROSSI, F.; BELTRAN, M.; DAMIZIA, M.; GRELLONI, C.; COLANTONI, A.; SETTI, A.; TIMOTEO, G. D.; DATTILO, D.; CENTRÓN-BROCO, A.; NICOLETTI, C. et al. Circular rna znf609/ckap5 mrna interaction regulates microtubule dynamics and tumorigenicity. *Molecular cell*, Elsevier, 2021.
- SAGIV-BARFI, I.; KOHRT, H. E.; CZERWINSKI, D. K.; NG, P. P.; CHANG, B. Y.; LEVY, R. Therapeutic antitumor immunity by checkpoint blockade is enhanced by ibrutinib, an inhibitor of both btk and itk. *Proceedings of the National Academy of Sciences*, National Acad Sciences, v. 112, n. 9, p. E966–E972, 2015.
- SALA, C.; LENA, P. D.; DURSO, D. F.; PRODI, A.; CASTELLANI, G.; NARDINI, C. Evaluation of pre-processing on the meta-analysis of dna methylation data from the illumina humanmethylation450 beadchip platform. *Plos one*, Public Library of Science San Francisco, CA USA, v. 15, n. 3, p. e0229763, 2020.
- SAWYERS, C. L. The cancer biomarker problem. *Nature*, Nature Publishing Group, v. 452, n. 7187, p. 548–552, 2008.
- SCHÜNEMANN, H. J.; LERDA, D.; QUINN, C.; FOLLMANN, M.; ALONSO-COELLO, P.; ROSSI, P. G.; LEBEAU, A.; NYSTRÖM, L.; BROEDERS, M.; IOANNIDOU-MOUZAKA, L. et al. Breast cancer screening and diagnosis: a synopsis of the european breast guidelines. *Annals of internal medicine*, American College of Physicians, v. 172, n. 1, p. 46–56, 2020.
- SILVA, P. A. d.; RIUL, S. d. S. Câncer de mama: fatores de risco e detecção precoce. *Revista Brasileira de Enfermagem*, SciELO Brasil, v. 64, p. 1016–1021, 2011.
- SMITH, P.; GODDE, N.; RUBIO, S.; TEKESTE, M.; VLADAR, E. K.; AXELROD, J. D.; HENDERSON, D. J.; MILGROM-HOFFMAN, M.; HUMBERT, P. O.; HINCK, L. Vangl2 regulates luminal epithelial organization and cell turnover in the mammary gland. *Scientific reports*, Nature Publishing Group, v. 9, n. 1, p. 1–13, 2019.
- SOUSA, M. C. de; SANTOS, L. dos; SOUSA, C. C. de et al. Diagnóstico de câncer de mama por exames genéticos: uma revisão de literatura. *Brazilian Journal of Health Review*, v. 3, n. 2, p. 1786–1797, 2020.
- STARNAWSKA, A.; DEMONTIS, D. Role of dna methylation in mediating genetic risk of psychiatric disorders. *Frontiers in Psychiatry*, Frontiers Media SA, v. 12, 2021.
- STAUDIGL, C.; CONCIN, N.; GRIMM, C.; PFEILER, G.; NEHODA, R.; SINGER, C. F.; POLTERAUER, S. Prognostic relevance of pretherapeutic gamma-glutamyltransferase in patients with primary metastatic breast cancer. *PLoS One*, Public Library of Science San Francisco, CA USA, v. 10, n. 4, p. e0125317, 2015.
- SZTUPINSZKI, Z.; NAOUR, J. L.; VACCHELLI, E.; LAURENT-PUIG, P.; DELALOGUE, S.; SZALLASI, Z.; KROEMER, G. [S.l.]: Taylor & Francis, 2021. 1859064 p.

TESCHENDORFF, A. E.; GAO, Y.; JONES, A.; RUEBNER, M.; BECKMANN, M. W.; WACHTER, D. L.; FASCHING, P. A.; WIDSCHWENDTER, M. Dna methylation outliers in normal breast tissue identify field defects that are enriched in cancer. *Nature communications*, Nature Publishing Group, v. 7, n. 1, p. 1–12, 2016.

TIBERIO, P.; CAVADINI, E.; CALLARI, M.; DAIDONE, M. G.; APPIERTO, V. Af1q: a novel mediator of basal and 4-hpr-induced apoptosis in ovarian cancer cells. *PloS one*, Public Library of Science San Francisco, USA, v. 7, n. 6, p. e39968, 2012.

TSE, W.; DEEG, H. J.; STIREWALT, D.; APPELBAUM, F. R.; RADICH, J.; GOOLEY, T. Increased af1q gene expression in high-risk myelodysplastic syndrome. *British journal of haematology*, Wiley Online Library, v. 128, n. 2, p. 218–220, 2005.

TSE, W.; MESHINCHI, S.; ALONZO, T. A.; STIREWALT, D. L.; GERBING, R. B.; WOODS, W. G.; APPELBAUM, F. R.; RADICH, J. P. Elevated expression of the af1q gene, an mll fusion partner, is an independent adverse prognostic factor in pediatric acute myeloid leukemia. *Blood*, American Society of Hematology, v. 104, n. 10, p. 3058–3063, 2004.

VACCHELLI, E.; ENOT, D. P.; PIETROCOLA, F.; ZITVOGEL, L.; KROEMER, G. Impact of pattern recognition receptors on the prognosis of breast cancer patients undergoing adjuvant chemotherapy. *Cancer research*, AACR, v. 76, n. 11, p. 3122–3126, 2016.

VAINIO, H.; BIANCHINI, F. et al. Iarc handbooks of cancer prevention. *Breast cancer screening*, IARC Press: Lyon, France, v. 7, p. 40–49, 2002.

VECCHI, L.; ZÓIA, M. A. P.; SANTOS, T. G.; BESERRA, A. de O.; RAMOS, C. M. C.; COLOMBO, B. F. M.; MAIA, Y. C. P.; ANDRADE, V. P. de; MOTA, S. T. S.; ARAÚJO, T. G. de et al. Inhibition of the anxa1/fpr1 autocrine axis reduces mda-mb-231 breast cancer cell growth and aggressiveness in vitro and in vivo. *Biochimica et Biophysica Acta (BBA)-Molecular Cell Research*, Elsevier, v. 1865, n. 9, p. 1368–1382, 2018.

VEER, L. J. V.; DAI, H.; VIJVER, M. J. V. D.; HE, Y. D.; HART, A. A.; MAO, M.; PETERSE, H. L.; KOOY, K. V. D.; MARTON, M. J.; WITTEVEEN, A. T. et al. Gene expression profiling predicts clinical outcome of breast cancer. *nature*, Nature Publishing Group, v. 415, n. 6871, p. 530–536, 2002.

WANG, C.; WAN, X.; YU, T.; HUANG, Z.; SHEN, C.; QI, Q.; XIANG, S.; CHEN, X.; ARBELY, E.; LING, Z.-Q. et al. *Cell Reports*, Elsevier, v. 32, n. 6, p. 108021, 2020.

WANG, J.; ZHOU, L.; LI, Z.; ZHANG, T.; LIU, W.; LIU, Z.; YUAN, Y.-C.; SU, F.; XU, L.; WANG, Y. et al. Yy1 suppresses fen1 over-expression and drug resistance in breast cancer. *BMC cancer*, BioMed Central, v. 15, n. 1, p. 1–15, 2015.

WANG, K.; WU, Z.; SI, Y.; TANG, W.; XU, X.; CHENG, Y.; LIN, J. Identification of lpcat1 expression as a potential prognostic biomarker guiding treatment choice in acute myeloid leukemia. *Oncology Letters*, Spandidos Publications, v. 21, n. 2, p. 1–1, 2021.

WANG, S.; LIANG, Y.; CHANG, W.; HU, B.; ZHANG, Y. Triple negative breast cancer depends on sphingosine kinase 1 (sphk1)/sphingosine-1-phosphate (s1p)/sphingosine 1-phosphate receptor 3 (s1pr3)/notch signaling for metastasis. *Medical science monitor: international medical journal of experimental and clinical research*, International Scientific Information, Inc., v. 24, p. 1912, 2018.

- WANG, Y.; KLIJN, J. G.; ZHANG, Y.; SIEUWERTS, A. M.; LOOK, M. P.; YANG, F.; TALANTOV, D.; TIMMERMANS, M.; GELDER, M. E. Meijer-van; YU, J. et al. Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. *The Lancet*, Elsevier, v. 365, n. 9460, p. 671–679, 2005.
- WANG, Y.; REN, F.; CHEN, P.; LIU, S.; SONG, Z.; MA, X. Identification of a six-gene signature with prognostic value for patients with endometrial carcinoma. *Cancer medicine*, Wiley Online Library, v. 7, n. 11, p. 5632–5642, 2018.
- WILLMER, T.; JOHNSON, R.; LOUW, J.; PHEIFFER, C. Blood-based dna methylation biomarkers for type 2 diabetes: potential for clinical applications. *Frontiers in endocrinology*, Frontiers, v. 9, p. 744, 2018.
- WOCKNER, L. F.; NOBLE, E. P.; LAWFORD, B. R.; YOUNG, R. M.; MORRIS, C. P.; WHITEHALL, V. L.; VOISEY, J. Genome-wide dna methylation analysis of human brain tissue from schizophrenia patients. *Translational psychiatry*, Nature Publishing Group, v. 4, n. 1, p. e339–e339, 2014.
- WROBEL, S. An algorithm for multi-relational discovery of subgroups. In: SPRINGER. *European symposium on principles of data mining and knowledge discovery*. [S.l.], 1997. p. 78–87.
- ZHANG, H.; XU, K.; XIANG, Q.; ZHAO, L.; TAN, B.; JU, P.; LAN, X.; LIU, Y.; ZHANG, J.; FU, Z. et al. Lpcat1 functions as a novel prognostic molecular marker in hepatocellular carcinoma. *Genes & Diseases*, Elsevier, 2020.
- ZHANG, J.; YANG, J.; LIN, C.; LIU, W.; HUO, Y.; YANG, M.; JIANG, S.-H.; SUN, Y.; HUA, R. Endoplasmic reticulum stress-dependent expression of ero1l promotes aerobic glycolysis in pancreatic cancer. *Theranostics*, Ivyspring International Publisher, v. 10, n. 18, p. 8400, 2020.
- ZHANG, S.; WANG, Y.; GU, Y.; ZHU, J.; CI, C.; GUO, Z.; CHEN, C.; WEI, Y.; LV, W.; LIU, H. et al. Specific breast cancer prognosis-subtype distinctions based on dna methylation patterns. *Molecular oncology*, Wiley Online Library, v. 12, n. 7, p. 1047–1060, 2018.
- ZHANG, Y.; PETROPOULOS, S.; LIU, J.; CHEISHVILI, D.; ZHOU, R.; DYMOV, S.; LI, K.; LI, N.; SZYF, M. The signature of liver cancer in immune cells dna methylation. *Clinical epigenetics*, BioMed Central, v. 10, n. 1, p. 1–17, 2018.
- ZHENG, Y.; MIAO, X.; LI, J.; HU, M.; ZHU, Y.; LI, X.; ZHANG, Y. Trichostatin a alleviates the process of breast carcinoma by downregulating lpar5. *European review for medical and pharmacological sciences*, v. 24, n. 11, p. 6417–6425, 2020.
- ZHU, T.; ZHENG, J.; HU, S.; ZHANG, W.; ZHOU, H.; LI, X.; LIU, Z.-Q. Construction and validation of an immunity-related prognostic signature for breast cancer. *Aging (Albany NY)*, Impact Journals, LLC, v. 12, n. 21, p. 21597, 2020.
- ZOU, J.; ZHU, L.; JIANG, X.; WANG, Y.; WANG, Y.; WANG, X.; CHEN, B. Curcumin increases breast cancer cell sensitivity to cisplatin by decreasing fen1 expression. *Oncotarget*, Impact Journals, LLC, v. 9, n. 13, p. 11268, 2018.