UNIVERSIDADE ESTADUAL PAULISTA "JÚLIO DE MESQUITA FILHO"
INSTITUTO DE BIOCIÊNCIAS
PROGRAMA DE PÓS-GRADUAÇÃO CIÊNCIAS BIOLÓGICAS - GENÉTICA

Análise Modular de Dados de Expressão para Avaliação de Múltiplos Tumores com Auxílio de Métodos Heurísticos.

Giordano Bruno Sanches Seco

Tese apresentada ao Instituto de Biociências, Campus de Botucatu, UNESP, em preenchimento parcial dos requisitos para a obtenção do título de Doutor no Programa de Pós-Graduação em Ciências Biológicas - Genética.

Orientador: Prof. Dr. José Luiz Rybarczyk Filho

Botucatu, Agosto de 2022.

UNIVERSIDADE ESTADUAL PAULISTA - "JÚLIO DE MESQUITA FILHO"
INSTITUTO DE BIOCIÊNCIAS DE BOTUCATU

Análise Modular de Dados de Expressão para Avaliação de Múltiplos Tumores com Auxílio de Métodos Heurísticos.

Aluno: Giordano Bruno Sanches Seco
Orientador: Prof. Dr. José Luiz Rybarczyk Filho

Botucatu, Agosto de 2022.

FICHA CATALOGRÁFICA ELABORADA PELA SEÇÃO TÉC. AQUIS. TRATAMENTO DA INFORM.
DIVISÃO TÉCNICA DE BIBLIOTECA E DOCUMENTAÇÃO - CÂMPUS DE BOTUCATU - UNESP
BIBLIOTECÁRIA RESPONSÁVEL: ROSANGELA APARECIDA LOBO-CRB 8/7500

Seco, Giordano Bruno Sanches.
Análise modular de dados de expressão para avaliação de múltiplos tumores com auxílio de métodos heurísticos / Giordano Bruno Sanches Seco. - Botucatu, 2022

Tese (doutorado) - Universidade Estadual Paulista "Júlio de Mesquita Filho", Instituto de Biociências de Botucatu
Orientador: José Luiz Rybarczyk-Filho
Capes: 20200005

1. Aprendizado do computador. 2. Bioinformática. 3. Câncer.

Palavras-chave: Aprendizado de máquina; Bioinformática; Câncer.

Agradecimentos

O doutorado foi um marco importante tanto em minha carreira quanto em minha vida. Se não fosse a disposição e colaboração de diversos indivíduos e instituições, teria sido quase impossível, para mim, concluir esse laborioso processo. Felizmente, este processo é tão laborioso quanto é recompensador, pois através dos desafios encontrados fui capaz de refinar a minha capacidade como pesquisador científico. De alguma forma espero ter contribuído com construção de conhecimento com esta tese e espero que, no futuro, toda essa experiência adquirida possa ser traduzida em contribuições ainda maiores para a ciência brasileira. Dito isso, gostaria de realizar alguns agradecimentos a indivíduos e entidades que considero especialmente importantes para a conclusão desta tese:

- Primeiramente quero agradecer a CAPES, cujo auxílio financeiro mediado pela bolsa registrada no processo GEN180122, tornou possível a execução desta pesquisa;
- Agradeço ainda ao Programa de Pós-Graduação em Ciências Biológicas (genética), que não apenas me aceitou como aluno, como também possibilitou a aquisição da bolsa mencionada acima;
- Agradeço a todos docentes e funcionários tanto na Seção Técnica de Pós-Graduação quanto no programa de Pós-Graduação em Ciências Biológicas (Genética), cujo suporte foi de extrema importância durante a execução da pesquisa;
- Agradeço a instituição UNESP como um todo, que me deu suporte desde o ano de 2012, como graduando, até a finalização desta tese no ano de 2022. Em especial agradeço ao Instituto de Biociências (IBB) no campus de Botucatu-SP, cuja infraestrutura possibilitou a execução da pesquisa;
- Gostaria de agradecer imensamente ao núcleo de computação científica da UNESP, que forneceu a infraestrutura computacional para execução das partes mais computacionalmente custosas desta pesquisa. Agradeço especialmente à equipe GridUnesp, que me deu suporte para execução dos procedimentos necessários;
- Agradeço a todos os docentes e funcionários do departamento de Biofísica e Farmacologia, que deram assistência e suporte para a execução desse projeto;

- Agradeço imensamente ao meu orientador Prof. Dr. José Luiz Rybarczyk-Filho, que me deu orientação e suporte para execução deste projeto de pesquisa;
- Agradeço imensamente à Dra. Agnes Alessandra Sekijima Takeda, que também forneceu orientação e *insights* valiosos durante a execução deste projeto de pesquisa;
- Agradeço imensamente aos membros do laboratório de estudos em Biocomplexidade, pelo companheirismo e suporte durante a execução deste projeto. Em especial, quero agradecer aos indivíduos André Luiz Molan e José Rafael Pilan pela amizade e suporte;
- Por fim agradeço à minha família, cujo suporte financeiro e especialmente o emocional tornaram minha estadia na cidade de Botucatu possível.

Lista de Figuras

Lista de Tabelas

Sumário

Resumo

Os grandes avanços nas ciências biológicas, especialmente na área das ciências ômicas, tem permitido aos pesquisadores realizar a coleta de grandes quantidades de dados biológicos. O TCGA (*The Cancer Genome Atlas*) é uma base de dados *online* que contém dados de diversos tipos de tumores e níveis de informação biológica, como por exemplo: expressão de mRNA (RNA-seq), expressão de microRNA (miRNA; miRseq) e metilação. Apesar das possibilidades que essa abundância de dados traz para os pesquisadores, tal área ainda carece de métodos de integração desses dados. O objetivo desta tese é o desenvolvimento de uma metodologia integrativa para avaliação de tecidos tumorais com o uso de três níveis de informação biológica (dados de RNA-seq, metilação e miRNA). Essa metodologia é baseada em *machine learning* não supervisionado, para obter um mapa modular que é composto por tumores e seus processos biológicos. Os processos biológicos foram prospectados do *Gene Ontology* e avaliados os níveis de profundidade de cada processo biológico para os módulos, bem como a frequência média de sondas induzidas ou reprimidas dos processos. Por meio de análise estatística, foi possível elaborar os mapas modulares que foram capazes de determinar vias biológicas de importância geral em 32 tumores. Tais achados envolvem vias como: síntese e transporte de lipídios, regulação da metilação, fibrinólise, regulação da atividade transcricional, transição epitélio-mesênquima, homeostase iônica, regulação de receptores de células do sistema imune e regulação de vias de sinalização mediadas por proteínas tirosina quinase. Foi realizada também uma análise mais específica do tumor mamário (BRCA),com auxílio de redes mRNA-mRNA (RNAseq e metilação) e miRNA-mRNA (miRseq). Esta última análise, conseguiu demonstrar que o tumor mamário tinha como alteradas, principalmente, as vias de diferenciação e desenvolvimento tecidual, efluxo de colesterol, angiogênese e vascularização. Em especial, a metodologia foi capaz de determinar genes e miRNAs cuja expressão ou metilação apresentam-se significantemente induzidos ou reprimidos nos dados como: o AR, ARHGDIG, PPARG, SUMO1 e o STAT3. Saliento que essa é uma metodologia de análise exploratória alternativa desenvolvida para reaproveitamento de grandes quantidades de dados biológicos. Especialmente quando esses dados não apresentem paridade caso/controle, o que dificulta a realização da análise de expressão diferencial, considerada como padrão ouro.

Abstract

Recent advances in biological sciences, especially in the area of omics sciences, have allowed researchers to collect large amounts of biological data. TCGA (*The Cancer Genome Atlas*) is an *online* database that contains data on different types of tumors and levels of biological information, such as: mRNA expression (RNA-seq), microRNA expression (miRNA; miRseq) and methylation. Despite the possibilities that this abundance of data brings to researchers, the area still lacks methods of integrating these data. The objective of this thesis is the development of an integrative methodology for the evaluation of tumor tissues using three levels of biological information (RNA-seq, methylation and miRNA data). This methodology is based on unsupervised *machine learning*, to obtain a modular map that is composed of tumors and their biological processes. The biological processes were prospected from Gene Ontology and the depth levels of each biological process for the modules were evaluated, as well as the average frequency of induced or suppressed probes of the processes. Through statistical analysis, it was possible to elaborate modular maps that were able to determine biological pathways of general importance in 32 tumors. Our findings involve pathways such as: lipid synthesis and transport, methylation regulation, fibrinolysis, transcriptional activity regulation, epithelial-mesenchymal transition, ionic homeostasis, regulation of immune system cell receptors and regulation of signaling pathways mediated by tyrosine kinase proteins. A more specific analysis of the mammary tumor (BRCA) was also performed, with the aid of mRNA-mRNA (RNAseq and methylation) and miRNA-mRNA (miRseq) networks. This last analysis was able to demonstrate that the mammary tumor had mainly altered pathways of tissue differentiation and development, cholesterol efflux, angiogenesis and vascularization. In particular, the methodology was able to determine genes and miRNAs whose expression or methylation are significantly induced or repressed in data such as: AR, ARHGDIG, PPARG, SUMO1 and STAT3. I emphasize that this is an alternative exploratory analysis methodology developed to reuse large amounts of biological data. Especially when these data do not present case/control parity, which makes it difficult to perform differential expression analysis, considered the gold standard.

1 *Introdução*

1.1 Ciências Ômicas e *Big Data*

Nas últimas duas décadas, houve uma grande evolução das tecnologias experimentais para obtenção de dados referentes a diversos aspectos de um sistema biológico, em especial, as ciências ômicas e suas tecnologias associadas, que são capazes de gerar *terabytes* de informações em um único estudo. As ciências ômicas têm o objetivo de utilizar tecnologias de alto desempenho para estudar o que é chamado de "**Dogma Central da Biologia Molecular**" explanado na Figura 1.1.

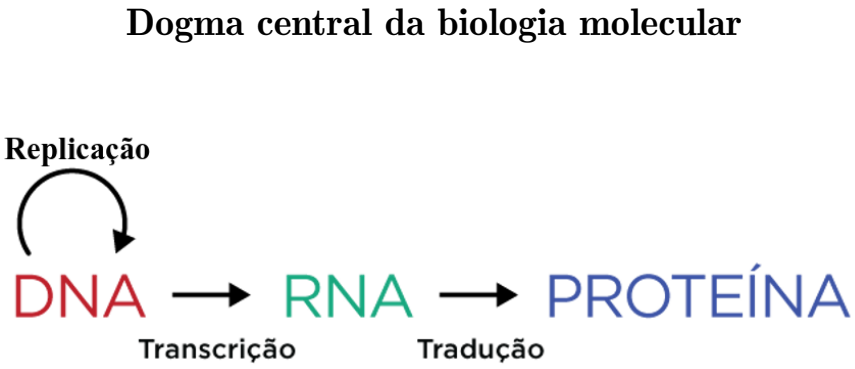


Figura 1.1: Segundo o dogma central da biologia molecular, o material genético (genoma) armazena em si toda a informação biológica do organismo. O genoma pode então passar sobre o processo de transcrição, dando origem aos RNA mensageiros (mRNA) e estes, mediante o processo de tradução dão origem as proteínas que são as principais responsáveis pelos fenótipos observados. Apesar da denominação de dogma, muitos conhecimentos novos foram adicionadas a este modelo, como por exemplo, hoje, sabe-se que nem todo o genoma tem função codificante assim como nem todo o RNA é mensageiro. Apesar das alterações que vão sendo realizadas conforme novos conhecimentos são produzidos, o teorema de que a informação flui do genoma para fenótipo observado mediante diversas etapas ainda permanece inalterado. As ciências ômicas atuam utilizando tecnologia de ponta para estudar essas etapas de maneira específica.

O fluxo de informação do genoma para os fenótipos, mediado por diversas etapas, é o objeto de estudo das ciências ômicas. Cada uma dessas etapas representa um aspecto específico do sistema biológico, as mais estudas estão ilustradas na Figura 1.2.

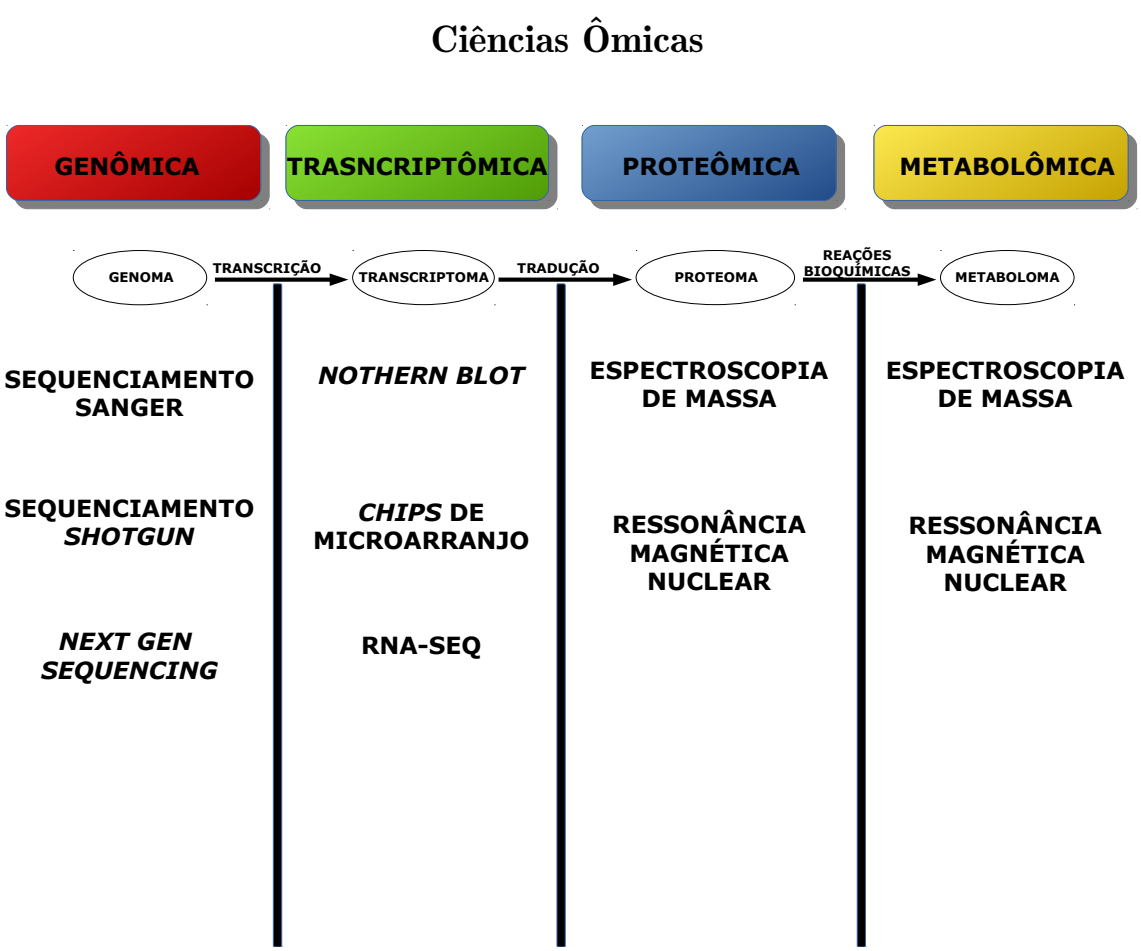


Figura 1.2: Na figura, está ilustrado o fluxo de informação biológica do genoma para o metaboloma, identicando o nome do processo responsável por catalisar a transição de um nível de informação para o outro. Esses níveis de informação, genoma, transcriptoma, proteoma e metaboloma estão separados por linhas grossas na figura acima, criando seus respectivos espaços. Na parte superior de cada um desses espaços, existem retângulos coloridos com o nome ciência ômica que estuda o respectivo nível de informação biológica, cercado por uma elipse. Na parte inferior de cada espaço de nível de informação, estão listados os nomes de alguns experimentos tipicamente utilizados pela respectiva ciência ômica.

O volume dos conjuntos de dados gerados por estudos das ciências ômicas faz com que tal área apresente muitas vezes os mesmos desafios encontrados para o *Big Data* em geral. De fato, a grande quantidade de dados gerados pelas ciências ômicas é em sí um problema de *Big Data*. Por essa razão muitas vezes as mesmas técnicas utilizadas para solução de desafios no *Big Data* podem também ser aplicadas para solucionar problemas em dados biológicos (MARX, 2013).

O termo *Big Data* se refere a grande quantidade de dados que as atuais tecnologias são capazes de gerar. Esses conjunto de dados são demasiadamente grandes para avaliação

humana ou mesmo com o auxilio de *software* e metodologias tradicionais. O *Big Data* é uma área que lida com 4 principais conceitos (os 4 **V**s) (HILBERT, 2016; GOLDSTEIN; SPATT; YE, 2021) envolvendo o uso de grandes conjuntos de dados: o **volume**, a **variedade**, a **veracidade** e a **velocidade**.

O **volume**, obviamente, se refere a quantidade de dados gerados, armazenados e a serem processados. A definição de quantidade de dados para o que pode ser considerado como *Big Data* é algo que está constantemente variando de acordo com as tecnologias, um único experimento de RNA-seq, por exemplo, pode gerar *gigabytes* de dados (ou até mesmo *terabytes*) (MARX, 2013). Trabalhar com uma quantidade tão grande de dados não é algo se pode fazer sem auxílio de ferramentas adequadas (HILBERT, 2016; GOLDSTEIN; SPATT; YE, 2021).

A **variedade** corresponde à natureza e classe dos dados, como por exemplo, se os dados são estruturados ou não estruturados. Dados estruturados são aqueles que estão previamente organizados e indexados, uma tabela que informa o valor de expressão de genes em cada uma das amostras de um experimento é um bom exemplo de dado estruturado (MARX, 2013). Dados não estruturados, analogamente, são aqueles que não possuem organização prévia, um texto, uma imagem e um arquivo de sequenciamento sem a devida indexação são exemplos de dados não estruturados (HILBERT, 2016; GOLDSTEIN; SPATT; YE, 2021).

A **veracidade** é a qualidade dos dados obtidos, ou seja, o quão confiáveis são os dados. Na biologia tal conceito está intrinsecamente ligado aos protocolos de realização do experimento em questão e a etapa de pré-processamento dos dados gerados pelo mesmo. A Figura 1.3 mostra um diagrama ilustrando o conceito de *Big Data*, os 4 **V**s e as etapas seguintes envolvendo o uso dos dados (HILBERT, 2016; GOLDSTEIN; SPATT; YE, 2021).

A **velocidade** é a frequência com a qual os dados são gerados e processados. Este constitui um desafio constante no *Big Data*, talvez, nem tanto na área de pesquisa biológica, mas, no mercado, existem empresas que têm a necessidade de trabalhar com dados gerados em tempo real (HILBERT, 2016; GOLDSTEIN; SPATT; YE, 2021).

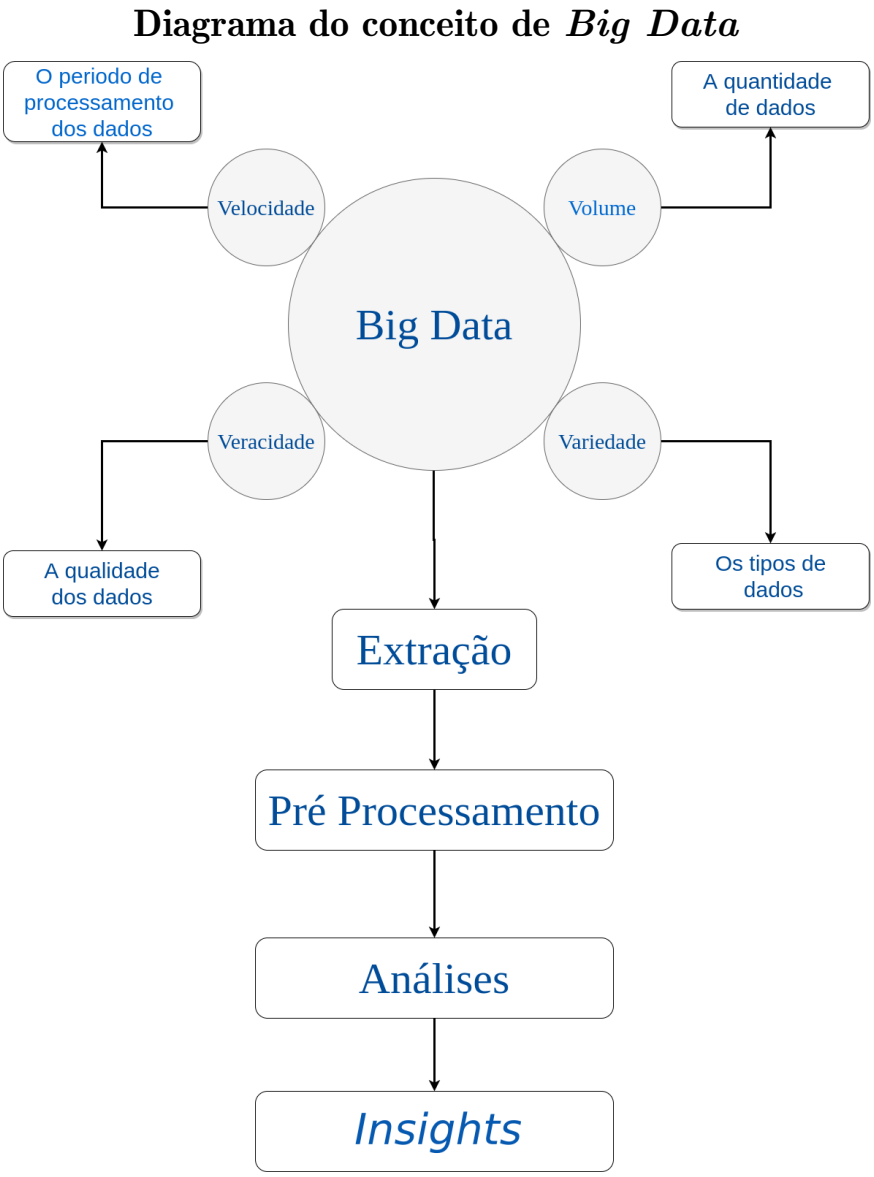


Figura 1.3: O diagrama ilustra o conceito de *Big Data* segundo suas 4 principais características (volume, variedade, velocidade e veracidade) e os procedimentos necessários para derivação de *insights*, a principal motivação desta área.

Os desafios associados aos 4 **V**s, se manifestam durante todo o processo de uso dos dados, exemplificados pelas outras etapas ilustradas na Figura 1.3: a **extração**, o **pré-processamento**, a **análise** e a derivação de *insights*. Embora grandes quantidades de dados tenham em sí uma abundância de informações, derivar *insights* valiosos ainda é um desafio grande. Este, no entanto, é a razão pela qual uma outra área, o **aprendizado de máquina** (**AM**/*Machine Learning*), se desenvolveu junto ao *Big Data*, agindo principalmente nas etapas de pré-processamento e análise, para otimizar o processo de derivação de *insights* (HILBERT, 2016; GOLDSTEIN; SPATT; YE, 2021). Gerar informações úteis partir de grandes conjuntos de dados é em sí um procedimento de agregação de **valor**. Em empresas que lidam diretamente com serviços associadas a tecnologia de informação, esse procedimento de agregação de valor por um profissional em *Big Data* se dá na forma de tomadas de **decisões inteligentes**. Algo análogo ocorre ná área da biologia, certamente algumas das técnicas experimentais previamente mencionadas, como o RNA-seq, necessitam de protocolos de análises rigorosos para derivação de *insights* que possibilitam chegar a conclusões válidas sobre o experimento em questão (HILBERT, 2016; GOLDSTEIN; SPATT; YE, 2021).

1.2 Câncer e Bases de Dados (GDC e TCGA)

De acordo com a organização mundial da saúde (*World Health Organization*) o câncer é a segunda maior causa de morte no mundo. Em 2020, foram 10 milhões de mortes registradas causadas por câncer (aproximadamente 1 em cada 6 **mortes registradas** era causada por câncer). Os cânceres de pulmão, fígado, colorectal, estômago e mamário são os maiores responsáveis pelas mortes de câncer, como pode ser observado no endereço eletrônico `https://www.who.int/news-room/fact-sheets/detail/cancer#:~:text=Cancer%20is%20a%20leading%20cause,and%20rectum%20and%20prostate%20cancers.`. Na última década, métodos de aprendizado de máquina têm sido utilizados para o estudo dessa doença, principalmente com objetivo de facilitar o manejamento clínico de pacientes (prognóstico, predição de sobrevivência, recorrência e susceptibilidade) e estudo de mecanismos moleculares da doença como é apresentado por Hanahan e colaboradores (HANAHAN; WEINBERG, 2011).

A aplicação de métodos de aprendizado de máquina em dados biológicos para análise não é algo recente, mas apesar disso não há um protocolo considerado padrão ouro para esse tipo de análise (CRUZ; WISHART, 2006). Métodos clássicos de aprendizado de máquina (AM) como regressão logística, regressão linear, K-*means* e clusterização[1]hierárquica são tipicamente aplicados em dados biológicos (genômicos, epigenômicos, transcriptômicos, etc...) para realizar reconhecimento de padrões, com a finalidade de melhor conhecer o comportamento da doença. Atualmente, uma nova tendência tem acompanhado as análises de AM, que é integrar diversos tipos de informações biológicas para enriquecer o processo de análise (CRUZ; WISHART, 2006). Ao invés de utilizar apenas, por exemplo, dados transcriptômicos de um experimento de RNA-seq para derivar *insights* sobre a doença, seria mais interessante criar um protocolo de análise que integrasse dados transcriptômicos juntamente com dados de sequenciamento, mutações e metilação. Dessa maneira, as conclusões irão contemplar diversos aspectos do sistema biológico e estudos mostram que essa integração de diversos tipos de informações biológicas pode trazer melhores resultados do que a utilização de um único tipo de informação (ZHANG et al., 2016). Isto tem ocorrido principalmente devido às tecnologias de alto rendimento das ciências ômicas, que causaram um grande influxo de dados sobre diversos aspectos do sistema biológico, incluindo genômica, proteômica, metabolômica e até mesmo imagens (EXARCHOS; GOLETSIS; FOTIADIS, 2012). Devido a grande quantidade de informações geradas, existem iniciativas que tentam organizá-las em bases de dados públicas e no caso dos dados sobre o câncer pode-se citar o *Genomic Data Commons* (GDC) como exemplo.

1.2.1 *Genomic Data Commons* (GDC)

O GDC é um sistema de informação para armazenar, analisar e compartilhar dados biológicos (genômicos, transcriptômicos) e clínicos de pacientes com câncer. As tecnologias de alto rendimento, utilizado pelas ciências ômicas no estudo de genomas e transcriptomas de câncer, deu origem a um problema de *Big Data* que impede muitos estudiosos não familiarizados com problemas dessa área, como biólogos e oncologistas, de extrair conhecimento desses dados. Portanto, o GDC visa democratizar o acesso e compartilhamento aos dados

[1]O termo "clusterização" é um neologismo da palavra "*clustering*" em inglês que significa "agrupamento" em português. Esse neologismo é tipicamente utilizado para nomear metodologias de análise que envolvem agrupamento de dados. No contexto desta tese, o neologismo "clusterização" foi utilizada para representar exatamente o mesmo conceito de análise de agrupamento que seu análogo em inglês, "*clustering*" representa.

biológicos e clínicos do câncer. O GDC organiza seus dados segundo uma arquitetura que pode ser observada na Figura 1.4.

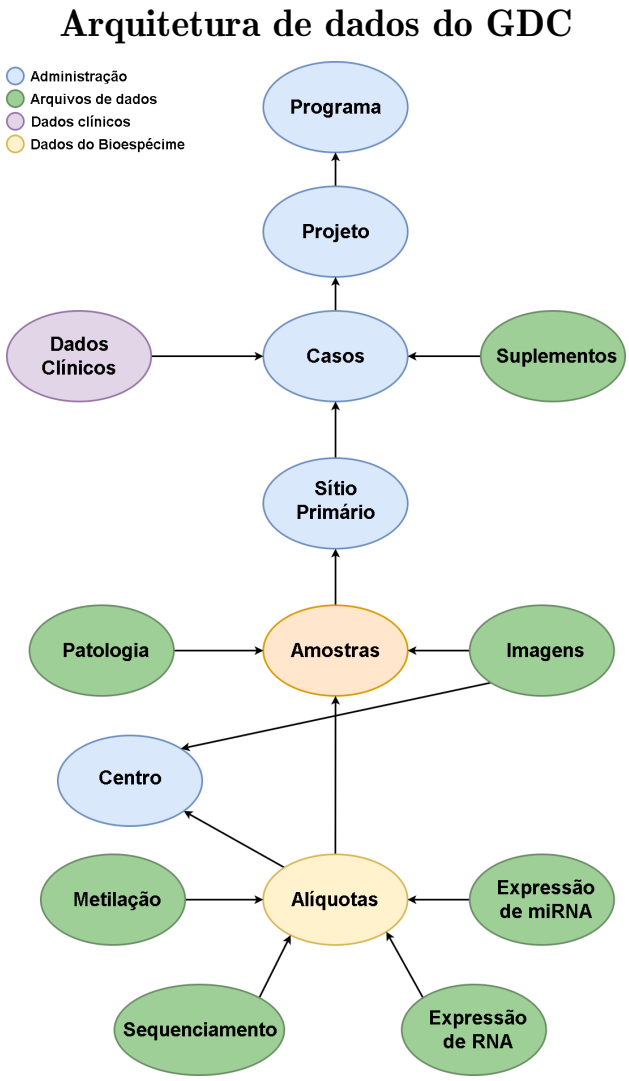


Figura 1.4: A figura ilustra a arquitetura de dados utilizada pelo GDC (*Genomic Data Commons*). As setas indicam a direção de um elemento de menor hierarquia da estrutura para um de maior hierarquia. Note que as cores tem significados: os elementos azuis são os de maior hierarquia da estrutura e representam etapas ou classes de administração dos dados, por exemplo de que estudo ou projeto (Projeto) proveniente ou de qual instituição (Centro) foi obtido. As hierarquias das outras três cores podem ser difíceis de se determinar entre sí, mas estarão sempre abaixo da cor azul (algum tipo de administração). Os dados clínicos referentes a cada caso estão marcados em roxo. Em laranja, metadados do bio espécime, esses geralmente se resumem a algum tipo de nomenclatura padronizada para as amostras e alíquotas. Por fim, os elementos verdes, que em geral são os de mais baixa hierarquia da estrutura e correspondem aos arquivos de dados experimentais (metilação, sequenciamento e expressão) ou metadados importantes como dados de patologia, suplementos e imagens. Esta figura é simultaneamente uma adaptação e redução da figura encontrada no endereço eletrônico `https://gdc.cancer.gov/about-gdc/scientific-reports/tester/gdc-data-model-components`, onde uma estrutura muito mais detalhada pode ser encontrada.

Na Figura 1.4, pode se observar a estrutura de dados utilizada pelo GDC. É importante ter em mente que as setas representam a relação de hierarquia entre os elementos na árvore. No caso, as setas sempre partem de um elemento de menor hierarquia para um de maior

hierarquia. Sendo assim, o usuário pode agrupar/selecionar dados tanto segundo aspectos administrativos como programa, projeto e sítio primário, quanto aspectos específicos dos dados em sí, como tipo de experimento ou patologia.

Os dados no GDC podem ser acessados, de diversas maneiras incluindo: endereço eletrônico, interface gráfica *user friendly* (*GDC Data Portal*), a ferramenta de linha de comando (*Data Transfer Tool*) e pelo API (*Application Programming Interface*) do endereço eletrônico. No *GDC Data Portal* `https://portal.gdc.cancer.gov/`, pode se consultar detalhadamente cada um desses métodos de acesso aos dados do GDC, como mostra a Figura 1.5 abaixo:

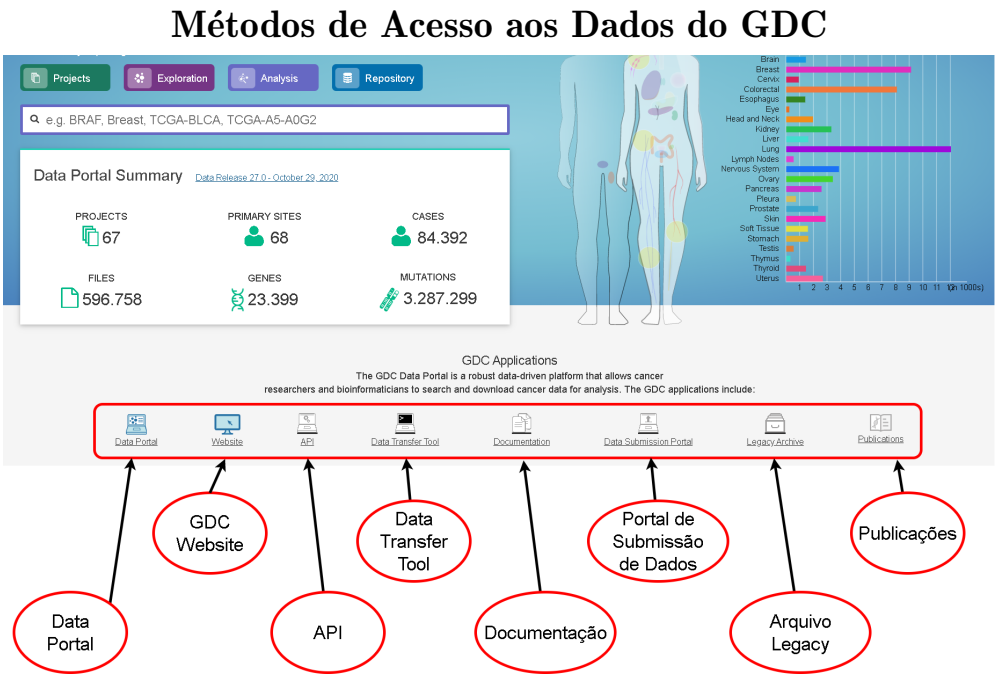


Figura 1.5: Na figura acima, temos a página inicial ao se acessar o endereço eletrônico *https://portal.gdc.cancer.gov/*, na qual pode se observar uma sumarização dos dados contidos no domínio. A parte inferior da página contém um menu com ícones, circulados em vermelho. Esses ícones representam: maneiras de acessar os dados ("*Data Portal*","*GDC Website*", "*API*","*Data Transfer Tool*" e "Arquivo *Legacy*"), maneira de submeter dados ("Portal de Submissão de Dados"), documentações da base de dados ("Documentaçao") e suas publicações ("Publicações").

O **GDC Data Portal** é um *webservice*, que permite navegar, consultar e baixar dados e metadados armazenados no endereço eletrônico do GDC com interface gráfica amigável para usuários. A Figura 1.6 abaixo mostra a página inicial do **GDC Data Portal**

Página inicial do GDC

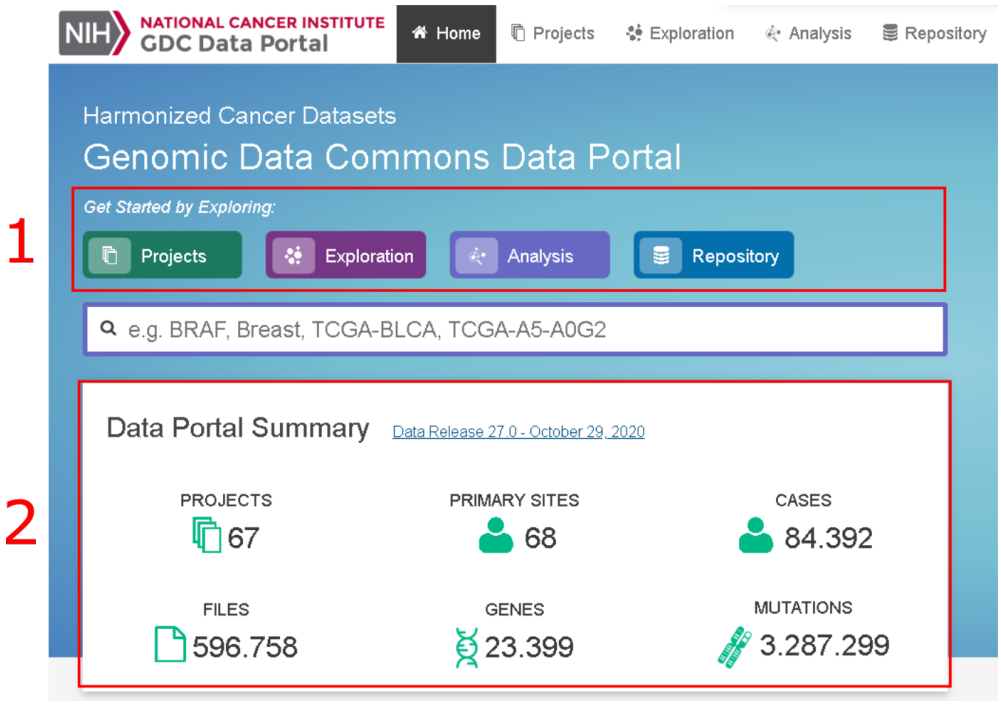


Figura 1.6: Em 1 pode-se observar as 4 principais formas de acesso aos dados armazenados no GDC. No caso, esse se trata do acesso via endereço eletrônico, mas existem outras maneiras de trabalhar com o GDC. Em *Projects*, os dados podem ser acessados como parte de um dos muitos projetos armazenados no GDC. Esses projetos podem ser subdivididos de acordo com outros critérios como: sítio primário, tipo de doença, categoria dos dados, estratégia experimental e o programa ao qual os projetos pertencem. Em *Exploration*, o usuário tem acesso a uma interface em que este pode explorar os dados segundo outros critérios que incluem: programa, projeto, sítio primário, tipo de doença e tipo de amostra. Em *Analysis*, o usuário tem acesso a uma interface gráfica que lhe permite realizar algumas análises, cujo *pipeline* já foi implementado à plataforma, cabendo ao usuário selecionar as amostras/dados a serem analisados. Em 2, pode se observar um sumário dos dados contemplados pelo GDC onde pode se observar: o número total de projetos, o número de sítios primários, o número de casos, o número de arquivos, o número de genes e o número de mutações.

Um dos critérios mais amplos para seleção dos dados é o projeto a que eles pertencem. O projeto é uma combinação do **programa** ao qual o projeto pertence e do **sítio primário** do tumor. Um exemplo seria o projeto **TCGA-BLCA**, basicamente os dados do projeto pertencem ao programa TCGA e são referentes ao tumor de bexiga (**BL**adder **CA**ncer). Abaixo segue uma lista de programas que estão armazenados no GCG e o número de projetos que contemplam:

Programas do GDC

Tabela 1.1: Programas que existem dentro do GDC até o momento (12/2020), bem como o número de projetos que existe dentro de cada um desses programas. Se observarmos o sumário do GDC na Figura 1.6, nota-se que há 67 projetos dentro do GDC, esse é justamente o total da soma da coluna "Número de Projetos".

Programa	Sigla	Número de Projetos
The Cancer Genome Atlas	TCGA	33
Therapeutically Applicable Research to Generate Effective Treatments	TARGET	9
Genomics Evidence Neoplasia Information Exchange	GENIE	8
BETAML1.0	BETAML	2
Cancer Genome Characterization Initiatives	CGCI	2
Count Me In	CMI	2
Clinical Proteomic Tumor Analysis Consortium	CPTAC	2
Clinical Trial Sequencing Project	CTSP	1
FM-AD	FM	1
Human Cancer Models Initiative	HCMI	1
The Multiple Myeloma Research Foundation	MMRF	1
National Cancer Institute Center for Cancer Research	NCICCR	1
Philadelphia-Negative Neutrophilic Leukemias (CNL/aCML/MDS/MPNu)	OHSU	1
Pancreatic Cancer Organoid Profiling	ORGANOID	1
VA Research Precision Oncology Program	VAREPOP	1
Genomic Characterization of Metastatic Castration Resistant Prostate Cancer	WCDT	1

O TCGA (*The Cancer Genome Atlas*), é uma das bases de dados contempladas pelo GDC, que produziu mais de 2.5 petabytes de dados que estão disponíveis no sítio `https://portal.gdc.cancer.gov/` (TCGA, 2018).

A Figura 1.7 mostra informações sobre a quantidade de dados que o TCGA atualmente contempla em seu domínio.

Dados armazenados no TCGA

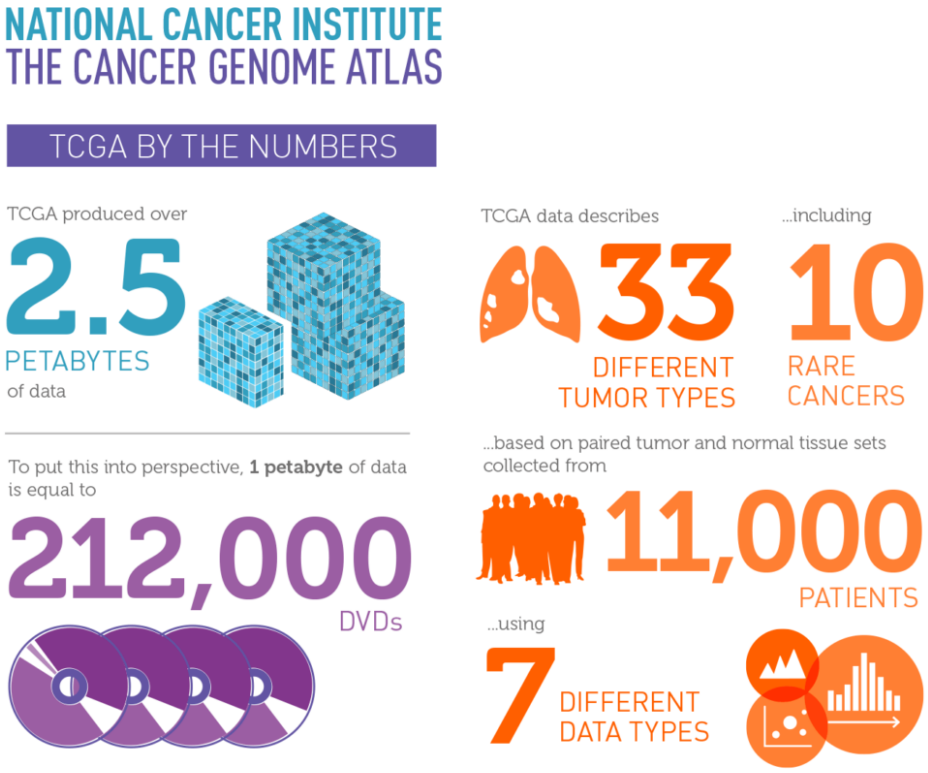


Figura 1.7: Dados armazenados no TCGA: A coluna esquerda, em azul e roxo, contém informações sobre a quantidade bruta de dados que o TCGA armazena no sítio `https://portal.gdc.cancer.gov/`. Na coluna direita, em laranja, pode se observa as classificação desses dados quanto aos tipos de tumor (33), experimentos (7) e número de pacientes (11000). Adaptado de `https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga/history`

Os dados podem ser categorizados e filtrados de diversas maneiras: tipo de tumor, tipo de experimento (rna-seq, metilação, miRNA, sequenciamento), tipo de plataforma e tipo de acesso (TCGA, 2018). Isso torna essa base de dados ideal para diversos estudos exploratórios, pois resultados de diferentes estudos podem ser comparados e analisados. Os 33 tipos de tumores/tecidos contemplados pelo TCGA possuem uma abreviação de letras associadas a seu nome, tal relação pode ser encontrada na Tabela 1.2. Note que um mesmo tecido pode ter mais de um tipo de tumor como é caso de 16 e 17 (TCGA, 2018).

Os 33 tumores do TCGA

Tabela 1.2: Relação entre abreviações e nomes de conjuntos de dados tumorais. Retirado de `https://gdc.cancer.gov/resources-tcga-users/tcga-code-tables/tcga-study-abbreviations`

	Abreviação	Nome
1	AML	*Acute Myeloid Leukemia*
2	ACC	*Adrenocortical carcinoma*
3	BLCA	*Bladder Urothelial Carcinoma*
4	LGG	*Brain Lower Grade Glioma*
5	BRCA	*Breast invasive carcinoma*
6	CESC	*Cervical squamous cell carcinoma and endocervical adenocarcinoma*
7	CHOL	*Cholangiocarcinoma*
8	LCML	*Chronic Myelogenous Leukemia*
9	COAD	*Colon adenocarcinoma*
10	CNTL	*Controls*
11	ESCA	*Esophageal carcinoma*
12	FPPP	*FFPE Pilot Phase II*
13	GBM	*Glioblastoma multiforme*
14	HNSC	*Head and Neck squamous cell carcinoma*
15	KICH	*Kidney Chromophobe*
16	KIRC	*Kidney renal clear cell carcinoma*
17	KIRP	*Kidney renal papillary cell carcinoma*
18	LIHC	*Liver hepatocellular carcinoma*
19	LUAD	*Lung adenocarcinoma*
20	LUSC	*Lung squamous cell carcinoma*
21	DLBC	*Lymphoid Neoplasm Diffuse Large B-cell Lymphoma*
22	MESO	*Mesothelioma*
23	MISC	*Miscellaneous*
24	OV	*Ovarian serous cystadenocarcinoma*
25	PAAD	*Pancreatic adenocarcinoma*
26	PCPG	*Pheochromocytoma and Paraganglioma*
27	PRAD	*Prostate adenocarcinoma*
28	READ	*Rectum adenocarcinoma*
29	SARC	*Sarcoma*
30	SKCM	*Skin Cutaneous Melanoma*
31	STAD	*Stomach adenocarcinoma*
32	TGCT	*Testicular Germ Cell Tumors*
33	THYM	*Thymoma*
34	THCA	*Thyroid carcinoma*
35	UCS	*Uterine Carcinosarcoma*
36	UCEC	*Uterine Corpus Endometrial Carcinoma*
37	UVM	*Uveal Melanoma*

A quantidade de dados disponíveis para cada tipo de tumor não tem uma distribuição uniforme. Alguns tumores como o BRCA (*Breast Invasive Carcinoma*) são mais amplamente estudados e consequentemente possuem mais dados disponibilizados para acesso público (TCGA, 2018). A Figura 1.8 mostra uma distribuição de quantidade de amostras por tipo de tumor.

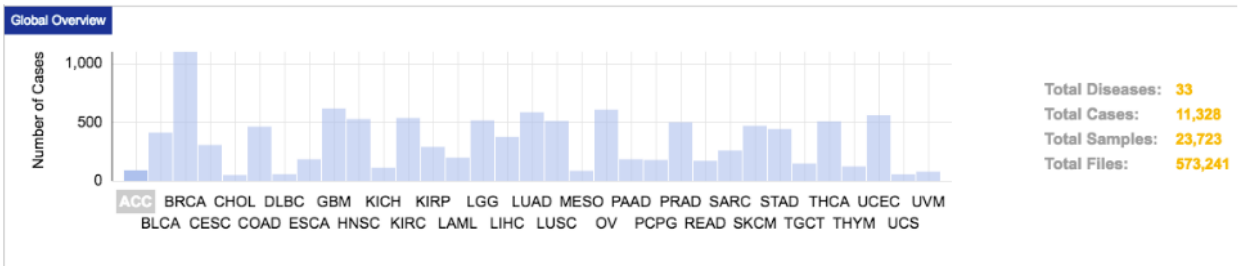


Figura 1.8: A distribuição de dados armazenados no sítio do TCGA. Pode se observar como existe uma grande discrepância na distribuição dos casos entre os tipos de tumores. O BRCA possui mais de 1000 casos disponíveis enquanto o CHOL possui menos de 100. Obtido de `https://www.sevenbridges.com/tcga/`

A Figura 1.7, menciona a existência de 7 tipos de dados. A Tabela 1.3 contém uma relação que mostra tipos de dados, os centros, a nomenclatura do tipo de dado no FTP (*File Transfer Protocol*) e se este tipo de dado tem acesso livre no endereço eletrônico `https://portal.gdc.cancer.gov/`. Note que esses dados proveem de diversas estratégias experimentais como:

- *Whole Genome Sequencing* (*WGS*)
- *Whole Exome Sequencing* (*WXS*)
- *Genotyping Array*
- *Methylation Array*
- *RNA-Seq*
- *miRNA-Seq*
- *Tissue Slides Diagnostic Slides* (Imagens)

Quantos aos tipos de centros, mais informações sobre cada um deles pode ser encontrado no domínio do CGC em `https://gdc.cancer.gov/resources-tcga-users/tcga-code-tables/center-codes`.

Tipos de dados contemplados pelo TCGA

Tabela 1.3: Tipos de dados contemplados pelo TCGA. Retirado de `https://gdc.cancer.gov/resources-tcga-users/tcga-code-tables/data-types`

	Tipo de Centro	Tipo de Dado	FTP	Acesso Livre
1	CGCC	*Expression-Genes*	*transcriptome*	Sim
2	CGCC	*Expression-Exon*	*exon*	Sim
3	CGCC	*Expression-miRNA*	*mirna*	Sim
4	CGCC	*CNV (CN Array)*	*cna*	Sim
5	CGCC	*DNA Methylation*	*methylation*	Sim
6	CGCC	*CNV (SNP Array)*	*snp*	Sim
7	CGCC	*SNP Frequencies*	*snp*	Não
8	CGCC	*SNP Copy Number Results*	*cna*	Não
9	GSC	*Trace-Sample Relationship*	*tracerel*	Sim
10	GSC	*Somatic Mutations*	*mutations*	Sim
11	BCR	*Complete Clinical Set*	*clin*	Sim
12	BCR	*Minimal Clinical Set*	*clin*	Sim
13	GSC	*Sequencing Trace*	*seq*	Não
14	BCR	*Tissue Slide Images*	*slide_images*	Sim
15	GSC	*Short-Read Relationship*	*sr*	Sim
16	GSC	*BAM-file Relationship*	*bam*	Sim
17	CGCC	*Quantification-miRNA*	*mirnaseq*	Sim
18	CGCC	*Quantification-miRNA Isoform*	*mirnaseq*	Sim
19	CGCC	*Annotations*	*annotations*	Sim
20	CGCC	*miRNA Sequence-Variants*	*mirnaseq*	Sim
21	CGCC	*Quantification-Junction*	*rnaseq*	Sim
22	CGCC	*Quantification-Gene*	*rnaseq*	Sim
23	CGCC	*Quantification-Exon*	*rnaseq*	Sim
24	BCR	*Reports*	*reports*	Sim
25	CGCC	*RNASeq*	*rnaseq*	Sim
26	CGCC	*miRNASeq*	*mirnaseq*	Sim
27	CGCC	*Expression-Protein*	*protein_exp*	Sim
28	CGCC	*Raw Sequence*	*raw_sequence*	Sim
29	CGCC	*Fragment Analysis Results*	*fragment_analysis*	Sim
30	CGCC	*GCC Reports*	*gcc_reports*	Sim
31	CGCC	*RNA Sequence-Variants*	*rnaseq*	Sim
32	CGCC	*RNA Structural-Variants*	*rnaseq*	Sim
33	CGCC	*RNASeqV2*	*rnaseqV2*	Sim
34	CGCC	*CNV (Low Pass DNASeq)*	*cna*	Sim
35	GSC	*Protected Mutations*	*mutations_protected*	Sim
36	CGCC	*Methylation - Bisulfite Sequencing*	*bisulfiteseq*	Sim
37	CGCC	*TotalRNASeqV2*	*totalrnaseqV2*	Sim
38	GDAC	*analyses*	*analyses*	Sim
39	GDAC	*stddata*	*stddata*	Sim
40	GDAC	*Firehose Reports*	*reports*	Sim

Como previamente mencionado, embora todos os dados estejam publicamente disponíveis, nem todos têm acesso livre (TCGA, 2018). Como pode se observar na última coluna da Tabela 1.3, alguns tipos de dados tem o que se chama de "acesso controlado", onde um cadastro e permissão são necessários antes de ter acesso a tais dados. Geralmente dados de sequenciamento e de SNPs (*Single nucleotide polymorphisms*) têm acesso restrito por razões que envolvem a privacidade do paciente (TCGA, 2018). Outro ponto interessante de se observar é que existem dados estruturados (expressão) e não estruturados (imagens) disponíveis para análise. Os dados estruturados são, tradicionalmente, os mais utilizados como mostrado em (ZHANG et al., 2016), entretanto, a incorporação de mais tipos dados a protocolos de análise de grandes quantidades de dados é algo que pode torna-las mais robustas, e dados não estruturados como as imagens de diagnósticos não são exceção. Inclusive, existe um artigo publicado por Mckinney e colaboradores (2020), na revista *Nature*, que mostra o desenvolvimento de sistema de AI (*Artificial Intelligence*) que conseguiu ter uma acurácia média maior que a de médicos para detecção de câncer de mama a partir de imagens de diagnóstico raio-x (MCKINNEY et al., 2020).

1.3 RNA-seq, miRseq e Metilação: *Design* Experimental

1.3.1 RNA-Seq

O RNA-seq (RNA *Sequencing*), é uma tecnologia de sequenciamento do tipo *Next Generation Sequencing* (NGS), cujo objetivo é mensurar a presença e a quantidade de ácido ribonucleico (RNA) numa amostra biológica (WANG; GERSTEIN; SNYDER, 2009). O primeiro método de sequenciamento (primeira geração) foi o de *Sanger*, trazendo consigo o conceito de sequenciamento capilar das amostras por meio da automatização da técnica de incorporação seletiva de dideoxinucleotídeos à cadeia de nucleotídeos de uma molécula de DNA sendo replicada *in vitro* por ação da enzima DNA polimerase (WANG; GERSTEIN; SNYDER, 2009).

A segunda geração de métodos de sequenciamento (ou NGS) trouxe o paradigma do sequenciamento massivo e paralelo (*Shotgun Sequencing*) (SCHUSTER, 2008), baseada na leitura simultânea de milhões de pedaços de *templates* de cDNA (DNA complementar) chamados de *short-reads* que estão dispostos sob a superfície de um dispositivo chamado *Flow Cell* (ZHONG et al., 2011). Alguns detalhes desse procedimento, como a geração da biblioteca de cDNA e o próprio processo de sequenciamento, variam entre as tecnologias fornecidas por diferentes empresas como a Illumina com a amplificação por ponte e sequenciamento por *reversible dye terminator* e a Roche com amplificação por PCR de emulsão e sequenciamento com a técnica de pirosequenciamento (ZHONG et al., 2011). A Figura 1.9 mostra o *design* experimental do experimento de RNA-Seq.

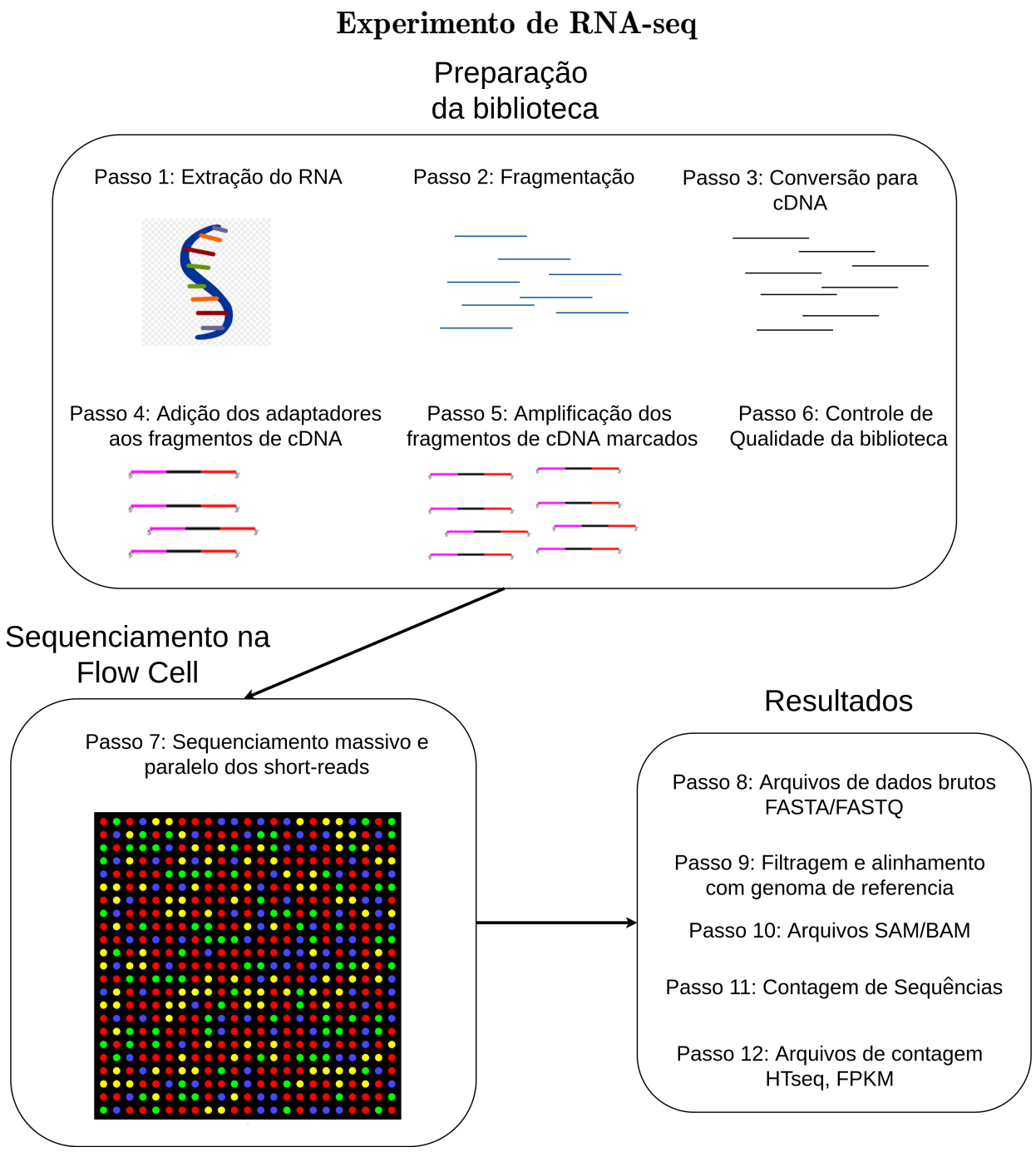


Figura 1.9: *Design* experimental do RNA-Seq: Os primeiros 6 passos compõe a primeira fase do experimento que é a da construção da biblioteca de material genético a ser sequenciado. Um experimento de RNA-Seq tradicional usa uma biblioteca de cDNA (DNA complementar) para mensurar a expressão de genes codificantes, mas com auxílio de tecnologias atuais, essa biblioteca pode conter outros tipos de material genético e até mesmo o genoma todo. A primeira fase é composta dos seguintes passos: o RNA é extraído da amostra (1), fragmentado em pedaços menores (2), convertido em dupla fita ou cDNA com auxilio da enzima transcriptase reversa (3), marcado com sequências adaptadoras (4), as sequências marcadas são então amplificadas por PCR (5) e por fim é realizado o controle de qualidade dos *short-reads* da biblioteca (6). A próxima fase é o sequenciamento (7) em sí, que, como mencionado segue o paradigma de sequenciamento massivo e paralelo. A técnica de sequenciamento pode variar entre as empresas, a Illumina, por exemplo, utiliza a técnica de *reverse dye terminator*. Nesta técnic, a biblioteca é aplicada sobre a superfície de uma *Flow Cell* e os *short-reads* são, então, sequenciados através de nucleotídeos marcados com cores. A partir do sinal luminoso emitido pela *Flow Cell* é possível inferir a sequencia dos *short-reads*. Plataformas mais recentes oferecem opções de realizar a leitura dos *short-reads* duas vezes em sentidos contrários, esse tipo de RNA-Seq é chamado de *paired-end* e costuma oferecer vantagens em relação ao RNA-Seq *single-end* como a detecção apurada de SNPs e deleções. A terceira fase, dos passos 7 a 12, é composta pela sumarização e filtragem dos dados brutos. Experimentos de RNA-Seq geram arquivos do FASTA (8) para armazenar os nomes dos *short-reads* e a sequência determinada. Plataformas mais modernas oferecem seus dados brutos no formato FASTAQ, que contém as mesmas informações do formato FASTA associadas a um *score* de qualidade. Em seguida, algoritmos são utilizados para realizar a filtragem e alinhamento dos *short-reads* (9) com auxilio de algorítimos e armazenar essas sequências mapeadas em arquivo do formato SAM ou BAM (10), este último é simplesmente um arquivo binário do primeiro. Por fim, algoritmos são utilizados para realizar a contagem de genes (11), essa matriz de *gene counts* pode ser fornecida em seu formato bruto, HTseq, ou normalizada (FPKM) (12) (ZHONG et al., 2011).

As tecnologias de sequenciamento, inicialmente, tinham como principal objetivo conhecer as sequências de nucleotídeos que compõe o genoma de organismos (WANG; GERSTEIN; SNYDER, 2009). A tendência de utilizá-las para mensurar o que é chamado de expressão gênica é algo que surgiu nas últimas décadas mediante desafios que surgiram com a principal técnica que era utilizada para esse fim, os *chips* de microarranjo (WANG; GERSTEIN; SNYDER, 2009). Em resumo, os *chips* de microarranjo determinavam a expressão gênica a partir da leitura de sinais luminosos na superfície de um *chip* cheio de sondas com pedaços do genoma de um organismo qualquer. Essas sondas realizam hibridização com uma biblioteca de cDNA marcada com fluoróforos com base na complementaridade entre as sequencias da sonda e da molécula marcada. Com a leitura dos sinais luminosos na superfície do *chip*, é possível saber o quanto o material genético de uma sonda em uma determinada amostra (Caso) está expresso relativamente a outra amostra (Controle) (WOO et al., 2004). A Figura 1.10 contém um diagrama ilustrando o design experimental dos microarranjos junto com uma extensa explanação.

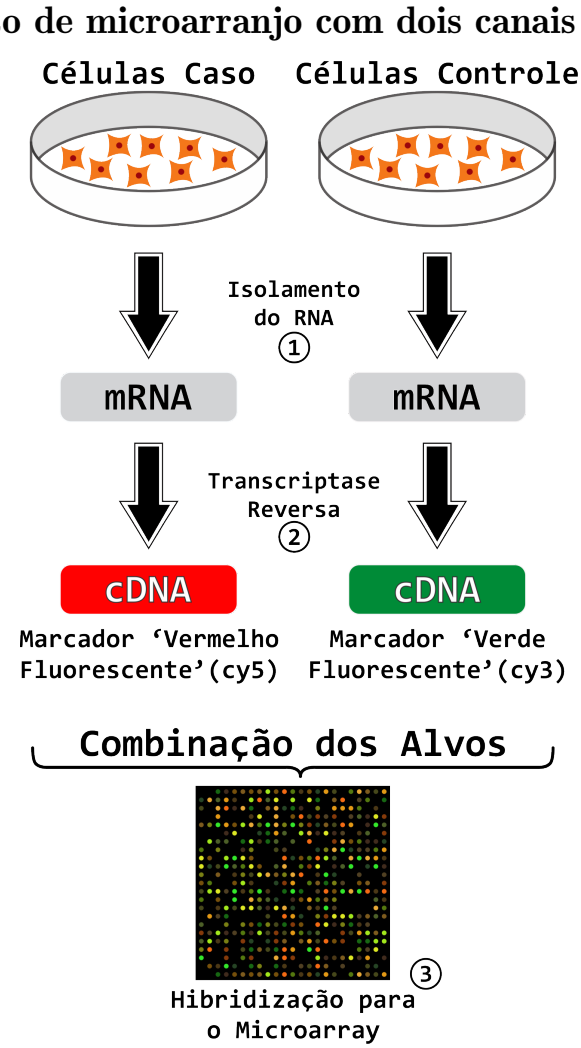


Figura 1.10: Ilustração das etapas de um experimento de microarranjo de dois canais (duas cores). Aqui duas amostras, uma caso e a outra o controle, tem seu mRNA extraído, isolado e utilizado para construção de uma biblioteca de cDNA (DNA sem íntrons) marcada. Geralmente, essa marcação é realizada via uso de fluoróforos, moléculas que emitem ondas eletromagnéticas no comprimento da luz visível (390 nm a 700 nm). Os fluoróforos mais utilizados são o Cy3 emitindo luz verde (570 nm) e o Cy5 que emite luz vermelha (670 nm) (HOEN et al., 2003). O cDNA então é hibridizado com as sondas no *chip* de microarranjo. O *chip* então é lavado e colocado para leitura em um *scanner* para registrar as intensidades luminosas de cada sonda. Analisando as intensidades luminosas de cada cor é possível determinar o quão sub ou superexpresso um gene está no caso em relação ao seu controle, já que a intensidade luminosa está diretamente associada a abundância de mRNA em cada situação (caso e controle) (NRC et al., 2007). Existe também o microarranjo de um canal só, suas etapas são idênticas ao microarranjo de dois canais, a diferença é que para comparar um caso e um controle são necessários dois chips de microarranjo. Essa é uma estratégia interessante, adotada pela linha GeneChip®da Affymetryx, já que elimina o problema das escalas de intensidades de cores diferentes que existe no microarranjo de duas cores. Outros problemas atenuados nesta linha incluem hibridização não específica, o chip da Affymetryx por exemplo tem para cada sonda milhares de cópias complementares à um tipo de oligonucleotídio de 25 pb (pares de bases) (essa sonda é chamada de *Perfect Match*), cada sonda PM tem uma sonda *mismatch* (MM) correspondente que contém milhares de cópias do mesmo oligonucleotídio da sonda PM, porém com a alteração de um nucleotídeo (o décimo terceiro). Assim as sondas MM podem ser utilizadas para avaliar a hibridização não específica de cDNA e servir como uma maneira de quantificar o ruído (DALMA-WEISZHAUSZ et al., 2006). Adaptado de (BIAGI-JUNIOR,).

Pela breve descrição, é possível notar a primeira grande vantagem do RNA-Seq, a sua plataforma (*Flow Cell*) não é organismo específica como o *chip* de microarranjo. Existem ainda mais vantagens para o RNA-Seq, como por exemplo, com esta técnica é possível analisar fenômenos que não podiam ser analisados com os *chips* de microarranjo como *splicing* alternativo, modificações pós-transcricionais, fusão de genes e mutações/SNPs (*Single Nucleotide Polymorphisms*). A depender de como a biblioteca de DNA é preparada, é possível analisar também outros tipos de transcritos além do mRNA (RNA mensageiro, principal componentes das bibliotecas de cDNA) como: micro RNA (miRNA), RNA transportador (tRNA), RNA ribossômico, lncRNA (*long non coding RNA*) e RNA pequenos (*small RNA*) (WANG; GERSTEIN; SNYDER, 2009). A Figura 1.11 mostra o fluxograma seguido pelo TCGA para disponibilização dos dados de RNA-seq.

Fluxograma de submissão de dados de RNA-seq no TCGA

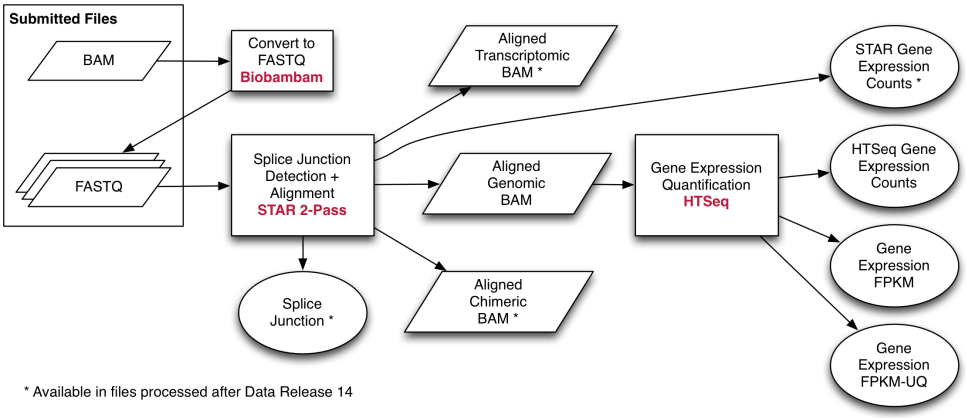


Figura 1.11: Fluxograma do TCGA para RNA-Seq: Os dados de saída do experimento de RNA-Seq (FASTQ ou BAM), sendo que os arquivos do formato BAM são primeiramente convertidos em formato FASTQ pelo protocolo *biobambam*. A partir desses arquivos, ocorrem as etapas de filtragem, alinhamento e detecção de junções de *splicing* alternativo com o protocolo *STAR 2-Pass*. Finalmente, o TCGA gera e disponibiliza vários formatos de arquivos como o de contagem bruta (HTseq) e contagem normalizada (FPKM e FPKM-UQ). Obtido de `https://docs.gdc.cancer.gov/Data/Bioinformatics_Pipelines/Expression_mRNA_Pipeline/`

1.3.2 miRseq

O protocolo de um experimento de **miRseq** é idêntico ao do RNA-Seq, o que é alterado entre os dois tipos de experimento, obviamente, é o material de entrada (a biblioteca de material genético). No RNA-Seq, a biblioteca é composta primordialmente de cDNA, ou a porção codificante do genoma. No miRseq, a biblioteca é enriquecida para captura de RNAs pequenos por meio do isolamento do RNA total em uma amostra seguido de filtragem por

tamanho em eletroforese (TAM; TSAO; MCPHERSON, 2015).

Hoje em dia, pesquisadores dão muita importância ao estudo da porção não codificante do genoma, pois esta porção pode ter outras funções importantes, como modulação da expressão do genoma codificante. Os miRNAs em especial, são úteis na determinação de padrões de expressão que são altamente tecido específica e doença-específica. Existem estudos que apontam como a expressão desregulada de miRNAs pode ter função importante no desenvolvimento de patologias como o câncer(TAM; TSAO; MCPHERSON, 2015).

1.3.3 Metilação

Metilação é o nome de uma transformação química onde um grupo metila (CH_3) é a adicionado a um outro substrato. Em termos mais abrangentes, a metilação é uma forma específica de alquilação, onde um grupo metila substitui um átomo de hidrogênio (BIRD, 1992). O processo de metilação do DNA é algo que tem recebido grande atenção por parte dos pesquisadores atualmente. Isto porque esse é o principal mecanismo de regulação epigenética que se tem conhecimento. Epigenética é conceito de mudanças fenotípicas herdáveis em seres vivos que não envolve alteração direta do genoma (GOLDBERG; ALLIS; BERNSTEIN, 2007). A metilação é um mecanismo de regulação da atividade/expressão do genoma. Em mamíferos, a metilação age tipicamente para a inibição do processo de transcrição gênica e foi mostrado como sendo mecanismo essencial para o desenvolvimento embrionário normal e também como estando diretamente associado a outros fenômenos como *genomic imprinting*, repressão de elementos transponíveis, inativação do cromossomo X, envelhecimento e até mesmo carcinogênese (DAS; SINGAL, 2004). A transformação da citosina em 5-metil-citosina é a forma de metilação mais estuda em seres humanos, esta transformação pode ser observada na Figura 1.12.

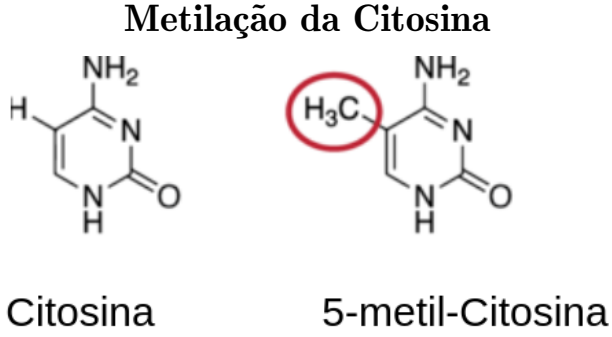


Figura 1.12: Diferença da molécula de citosina não metilada para a metilada, na qual um de seus átomos de hidrogênio é substituído por um grupo metila. Adaptado de `https://commons.wikimedia.org/wiki/File:DNA_methylation.png`

Para avaliação de padrões de metilação do genoma, utiliza-se a técnica de sequenciamento de bissulfito. Em resumo, após a extração do material genético de uma amostra este é tratado com bissulfito (HSO_3) que converte resíduos de citosina em uracila, mas deixa resíduos de 5-metil-citosina inalterados. Assim o tratamento com bissulfito introduz mudanças nucleotídeo-específicas no DNA, que dependem do estado de metilação ao não do nucleotídeo (HUANG et al., 2010). Após o tratamento, o sequenciamento de bissulfito segue o mesmo *workflow* do RNA-seq incluindo: preparação da biblioteca, amplificação e sequenciamento (HUANG et al., 2010). Como no RNA-Seq, existem diversas tecnologias diferentes para realização desses passos, sendo que o paradigma do sequenciamento paralelo e massivo também é mantido. A tecnologia de sequenciamento de genoma inteiro pode ser aplicada a uma sequência de DNA previamente tratada e amplificada para determinação dos níveis de metilação, com auxílio, por exemplo, do protocolo Illumina *Hiseq*®ilustrado na figura 1.9. Entretanto, a própria Illumina desenvolveu uma plataforma específica para analisar níveis de metilação numa amostra, a tecnologia *Infiniun*®, baseada no uso *chips* parecidos com o de um microarranjo chamados *BeadChip*®. A Figura 1.13 mostra o *design* experimental do Illumina *BeadChip* 450k®(MAKSIMOVIC; GORDON; OSHLACK, 2012).

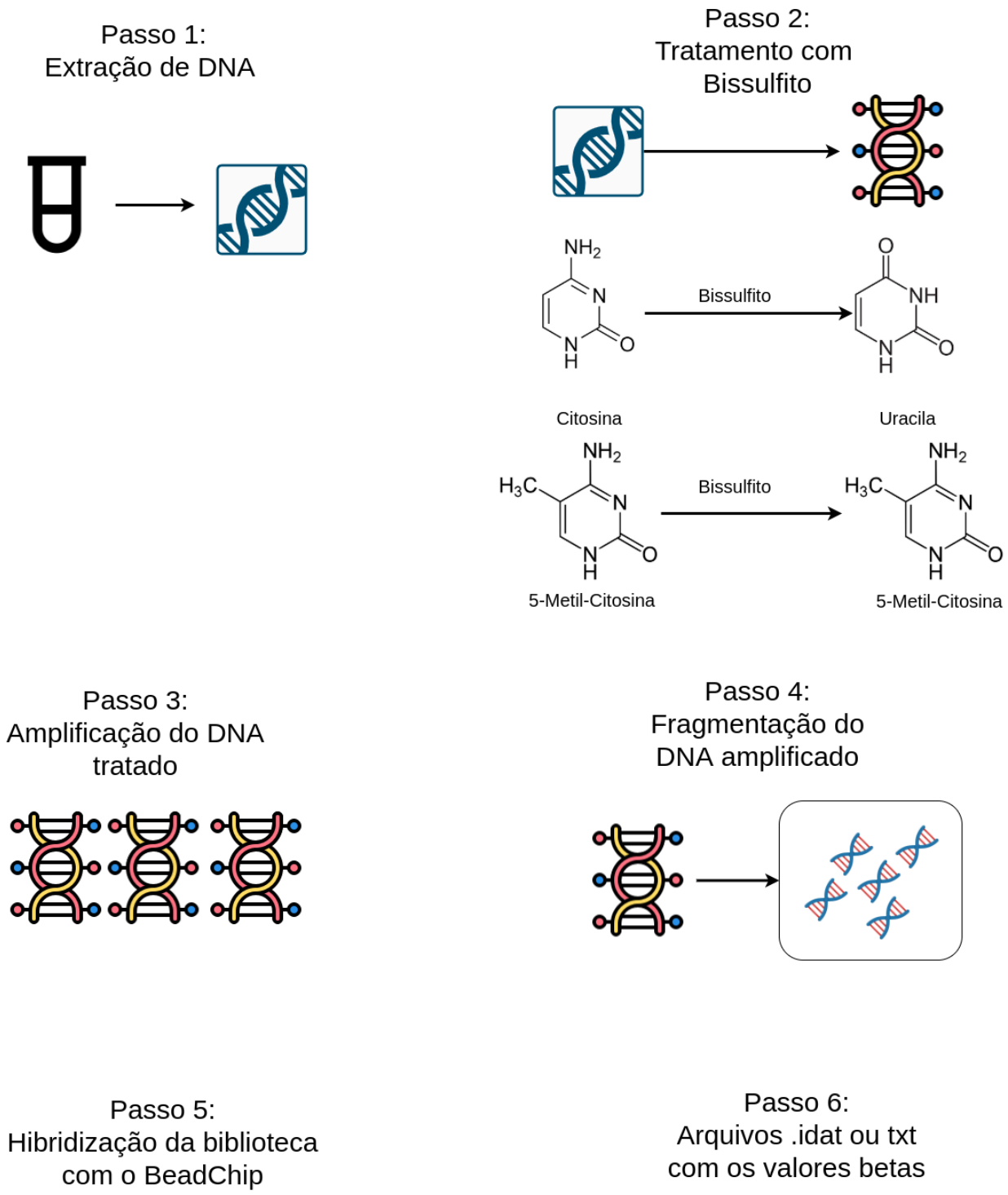


Figura 1.13: *Workflow* do *design* experimental da tecnologia Illumina *BeadChip*®. O *BeadChip*®da Illumina tem a vantagem de oferecer cobertura de todo o genoma humano. Existem duas versões o *Infinium HumanMethylation27 BeadChip*®e o *Infinium HumanMethylation450 BeadChip*®. O primeiro possui 27000 sondas com a tecnologia *Infinium I*, já o segundo 450000 sondas com tecnologia *Infiniun II*. Fora isso, o *workflow* de uso deles é idêntico: o primeiro passo é a extração de DNA (1) seguido do tratamento com bissulfito (2). Nesta etapa (2), o bissulfito converte citosinas em uracila por reação de deaminação enquanto as 5-metil-citosinas permanecem inalteradas. Em seguida, o DNA tratado com bissulfito é amplificado por PCR (3) e enzimaticamente fragmentado (4). Com a biblioteca pronta, ocorre sua aplicação sobre o Illumina BeadChip®(5). Finalmente, os valores beta são determinados e fornecidos no formato de arquivos *.idat, embora seja comum encontra-los também no formato .txt (MAKSIMOVIC; GORDON; OSHLACK, 2012).

1.4 Aprendizado de Máquina (AM)

O aprendizado de máquina é algo que tem sido amplamente utilizado para derivar *insights* de conjuntos de dados de grande porte. Essa área do conhecimento é basicamente um conjunto de métodos estatísticos que visa dar ao computador a habilidade de "aprender" de maneira gradual a partir do reconhecimento de padrões em um conjunto de dados de entrada (TARCA et al., 2007). A Figura 1.14 contém um diagrama ilustrando a subdivisão dos métodos de aprendizado de máquina.

Os métodos de aprendizado de máquina geralmente são divididos em dois tipos: supervisionados e não supervisionados (ANG et al., 2015). Como pode se observar na Figura 1.14 no aprendizado de máquina supervisionado as principais abordagens utilizadas são a classificação e a regressão, ambas abordagens têm um viés preditivo. Isso se dá pela própria natureza dos dados de entrada, uma vez que estes já estão previamente categorizados, ou seja, existe o que se chama de uma variável alvo. Partindo desse ponto, é natural imaginar que o pesquisador esteja interessado em saber, por exemplo, qual configuração das outras variáveis presentes no conjunto de dados esteja mais associada com a variável alvo (ANG et al., 2015).

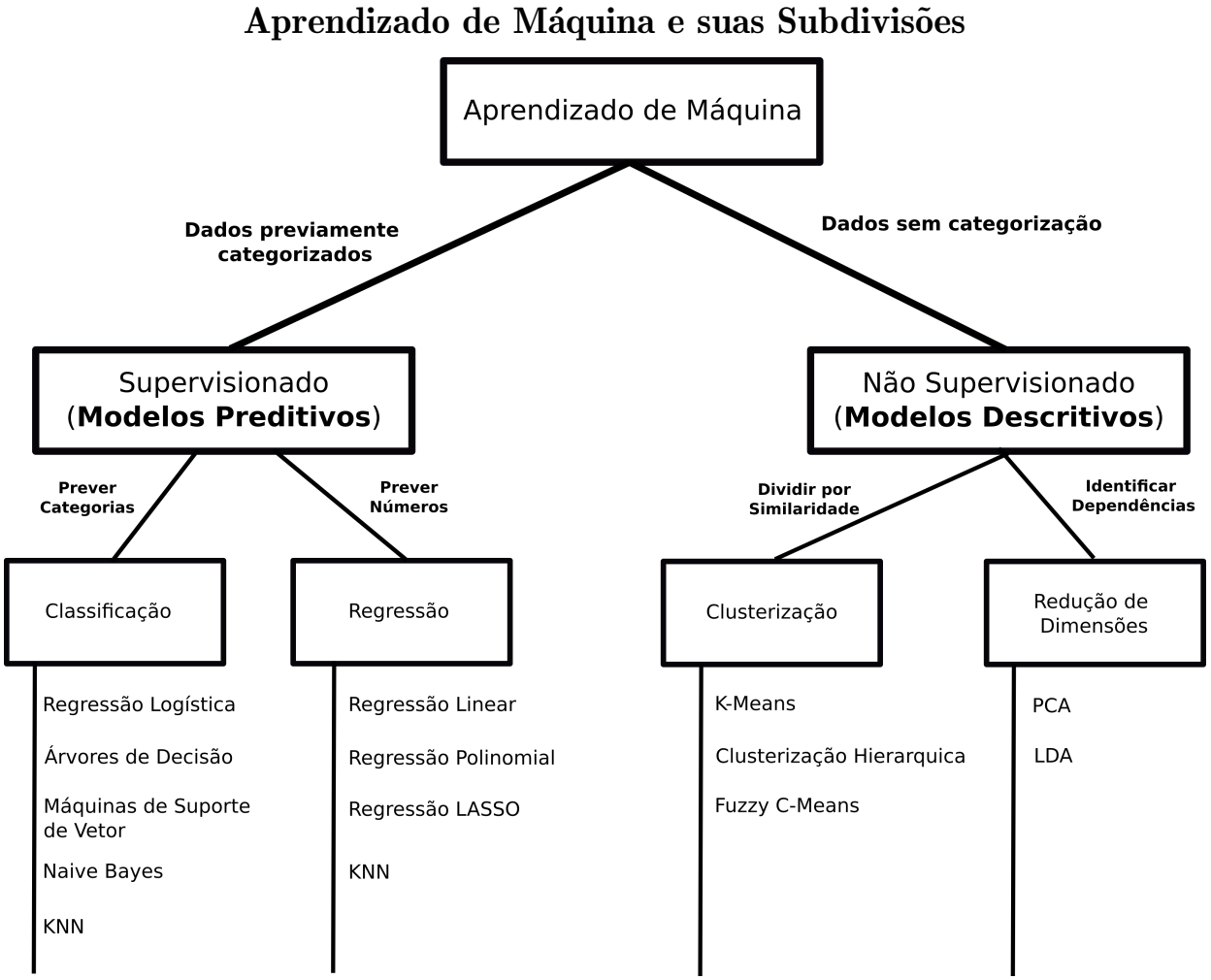


Figura 1.14: O aprendizado de máquina é, de maneira geral, divido em dois conjuntos de metodologias: supervisionados e não supervisionados. Isso está diretamente associado ao conjunto de dados de entrada e ao objetivo final da análise. Métodos supervisionados como regressão logística, K *nearest neighbours* (KNN) e máquinas de suporte de vetores (SVM) são usualmente utilizados para construir modelos preditivos. Esses modelos podem ter como saída classes (Classificação) ou números (Regressão). Os modelos não supervisionados são usados para fazer uma descrição dos dados de entrada e tipicamente abragem os algoritmos de clusterização como KNN, K-*Means*, *Fuzzy* C-*Means* e clusterização hierárquica onde o principal objetivo é agrupar os dados de entrada por alguma medida de similaridade. Os de redução de dimensão como *Principal Component Analisys* (PCA) e *Linear Discriminant Analysis* (LDA) têm o objetivo de encontrar dependências ocultas entre as variáveis do conjunto de dados, afim de reduzir a redundância dos dados. Note que um mesmo algoritmo pode ser utilizado para diversas finalidades, como o KNN, novamente isso depende inteiramente dos dados de entrada e do objetivo da análise (ANG et al., 2015).

Metodologias de regressão realizam previsões através de resultados numéricos, para entender esse conceito basta entender uma *regressão linear*, que é uma das mais simples. Nessa metodologia, imagine um conjunto de dados que mede a expressão do gene G em função da concentração de um fármaco. A regressão linear, em sua forma mais simples, consiste em determinar a equação de reta $y = ax + b$ que melhor se ajuste aos dados de treinamento, onde y é o valor da expressão, a o coeficiente angular da reta e c o coeficiente linear da reta (BEWICK; CHEEK; BALL, 2003). Na Figura 1.15 abaixo, pode se observar um exemplo

de regressão linear com dados artificiais para nosso suposto conjunto de dados.

	Expressão	mg\l
1	400.00	10.00
2	398.54	20.00
3	396.92	30.00
4	396.43	40.00
5	391.76	50.00
6	391.49	60.00
7	387.44	70.00
8	386.61	80.00
9	383.32	90.00
10	381.55	100.00

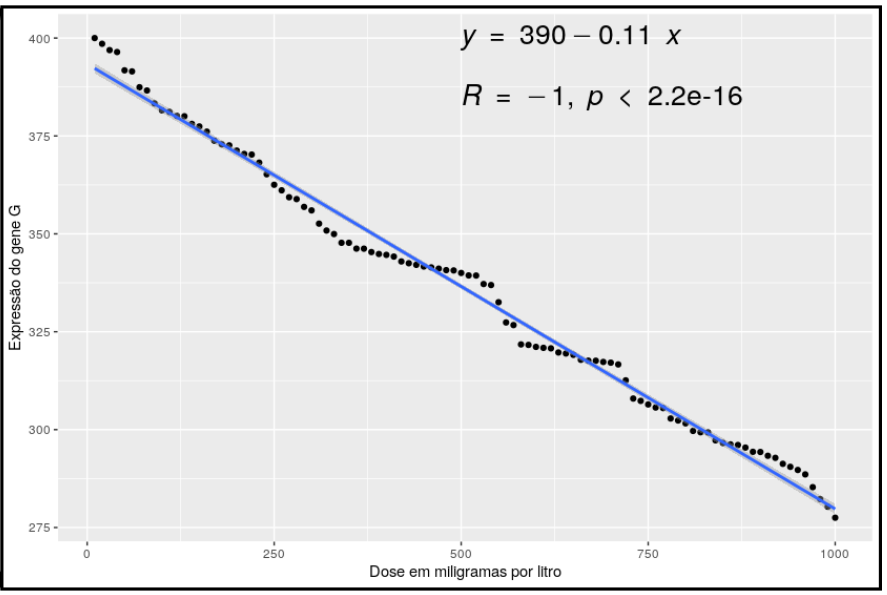


Figura 1.15: Um exemplo de regressão linear. À esquerda, há uma tabela com as 10 primeiras observações do conjunto de dados artificial que registra a expressão do Gene G fictício em função da concentração em microgramas por litro (mg\l) de um fármaco arbitrário. À direita, há um gráfico feito com o pacote "*ggplot2*" onde os pontos pretos são os valores reais registrados na tabela, enquanto a reta azul é o modelo construído pela regressão linear. Pode se observar, o modelo resultante na forma da equação $y = 390 - 0.11x$ juntamente com o coeficiente de correlação $R = -1$ com um p-valor de praticamente 0. Isso mostra que a regressão linear é capaz de explicar praticamente 100% do comportamento dos dados desse conjunto de dados fictícios.

Ainda na classe dos métodos supervisionados existem os métodos de classificação como a regressão logística, cujo objetivo é realizar a predição de classes e não de números como na regressão linear. Variáveis não numéricas são chamadas de variáveis categóricas (CHENG; HÜLLERMEIER, 2009):

- Binárias: Possuem apenas duas configurações possíveis, 0 ou 1, sim ou não, morto ou vivo, etc...
- Ordenadas: Possuem várias configurações, mas existe uma ordem sequencial intrínseca a elas como: nivéis de expressão gênica, faixa de idade, etc...
- Não Ordenadas: Tem várias configurações possíveis que não possuem nenhuma sequencia intrínseca associada: nome do gene, cor dos olhos, etc...

A matemática por trás da regressão logística mais simples é, a princípio, parecida com a da regressão linear. É assumido que os valores da variável alvo y podem ser representados

como a equação da reta $y = ax + b$. Entretanto existem algumas diferenças importantes, causadas pela natureza da variável alvo. Imagine que o conjunto de dados artificial do gene G agora representa um ensaio sobre uma droga que é aplicada em uma determinada dose para cada paciente com uma mesma doença arbitrária. Cada ponto do conjunto de dados agora representa o resultado do tratamento em um dos pacientes e existe também uma terceira coluna informando se o paciente sobreviveu ou não.

O objetivo é criar um modelo capaz de predizer, a partir da expressão do gene G, se o paciente sobreviveu ou não ao tratamento. A diferença notável que existe entre esse caso e o exemplo anterior, onde foi utilizada a regressão linear para prever a expressão do gene a partir da dose, é que a variável alvo não é numérica e continua, mas uma variável binária. É de se imaginar que a equação $y = ax + b$ não irá, por si só, retornar valores binários. Algumas mudanças são necessárias nesse cenário, pois aqui o espaço representa a **probabilidade** da variável alvo ser de uma classe ou não (CHENG; HÜLLERMEIER, 2009). Essa probabilidade, obviamente, nunca pode ser negativa e nem maior que 1. Portanto, é aplicada uma função exponencial a equação da reta (BEWICK; CHEEK; BALL, 2005):

$$\begin{aligned} exp(y) &= exp(ax+b) \\ P &= exp(ax+b) \end{aligned} \tag{1.1}$$

A probabilidade P nunca pode ser maior que 1, portanto o valor é dividido por ele mesmo acrescido de 1 (BEWICK; CHEEK; BALL, 2005).

$$P = \frac{exp(ax+b)}{exp(ax+b)+1} \tag{1.2}$$

Essa equação pode ser escrita em termos de *log*, assim a probabilidade em *log* (origem do nome "logística") fica (BEWICK; CHEEK; BALL, 2005):

$$log(\frac{P}{1-P}) = ax+b \tag{1.3}$$

A Figura 1.16 ilustra a diferença gráfica dos dois modelos de regressão (linear e logístico). Logo em seguida, a Figura 1.17 mostra o novo conjunto de dados artificial do gene G, e um gráfico com o modelo de regressão logística derivado.

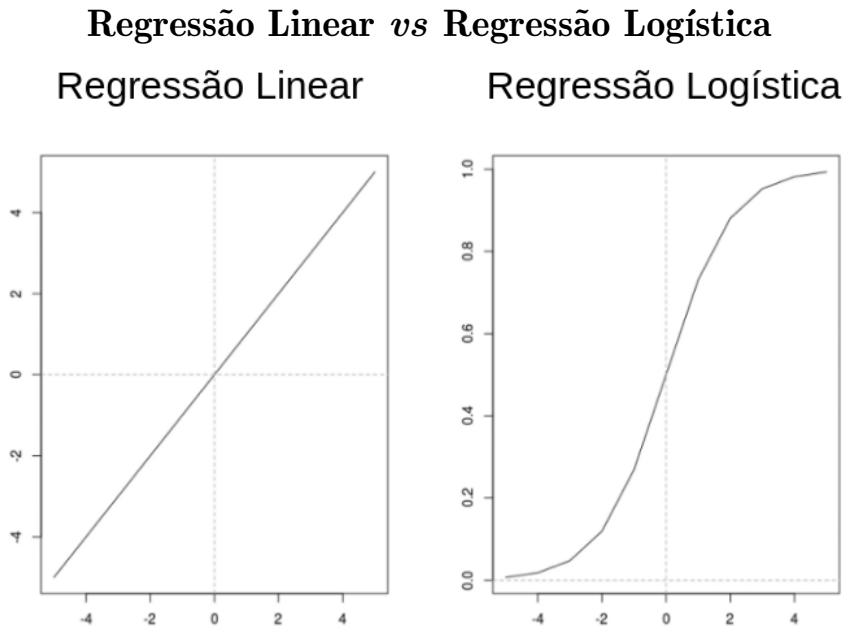
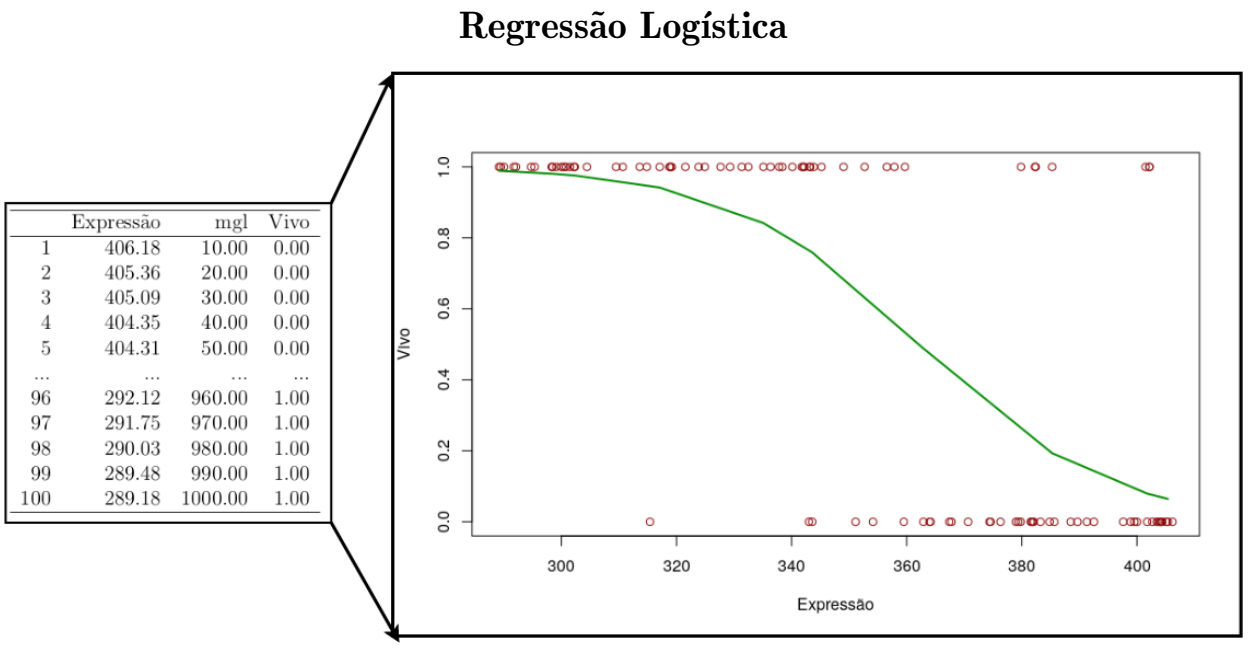


Figura 1.16: Uma ilustração das regressões linear e logística.



	Expressão	mgl	Vivo
1	406.18	10.00	0.00
2	405.36	20.00	0.00
3	405.09	30.00	0.00
4	404.35	40.00	0.00
5	404.31	50.00	0.00
...	...	...	...
96	292.12	960.00	1.00
97	291.75	970.00	1.00
98	290.03	980.00	1.00
99	289.48	990.00	1.00
100	289.18	1000.00	1.00

Figura 1.17: Regressão logística com o conjunto de dados do gene *G*. À esquerda há uma tabela com os 5 primeiros e os 5 últimos registros da tabela. À direita, há um gráfico onde os pontos vermelhos representam os valores da coluna "Vivo" da tabela e também a curva verde correspondente ao modelo de regressão logística derivado. Note que a curva determina a probabilidade do individuo ter sobrevivido a partir da expressão do gene, cabe ao pesquisador determinar qual o valor crítico que separa uma configuração da outra. Exemplo: toda expressão com de $y \geq 0.4$ é classificada como "morto".

A clusterização KNN (K *Nearest Neighbors*) é um método não paramétrico para aprendizado supervisionado que pode ser utilizado tanto para classificação como para regressão

(ZHANG, 2016). A ideia por trás do algoritmo é classificar um conjunto de valores não categorizados a partir de um conjunto de dados previamente categorizado (ZHANG, 2016). Um ponto é classificado dependendo da classe dos seus K vizinhos mais próximos (portanto, o nome K *Nearest Neighbors*). Por exemplo, se o K é determinado como 3 e em um ponto arbitrário dois dos seus vizinhos mais próximos pertencem a classe "A" e um a classe "B", o algoritmo iria determinar que esse ponto pertence a classe "A". Assim esse tipo de algoritmo pode ser resumido em alguns passos(ZHANG, 2016):

1. Determinar o valor de K;
2. Determinar a métrica de distância a ser utilizada;
3. Calcular a métrica de distância de um ponto não classificado a todos os outros classificados e determinar os k vizinhos mais próximos;
4. Determinar as classes dos K vizinhos mais próximos;
5. Retornar a classe predita.

É um algoritmo relativamente simples, como pode se observar, no entanto existem dois problemas comuns ao utilizá-lo: qual a melhor métrica de distância a se utilizar e qual o melhor número de vizinhos K. Valores de K muito baixos, em relação ao conjunto de dados em questão, desfrutam da vantagem de ter um custo computacional mais baixo e também de terem um viés menor, mas uma variância elevada devido ao fato do modelo estar mais exposto ao ruído dos dados. Analogamente um K muito elevado terá o problema de um custo computacional maior e a vantagem de te um especificidade maior no modelo criado, mas com risco de introduzir viés ao modelo (ZHANG, 2016).

Existem muitas métricas de distância para algoritmos de clusterização, as mais comuns são a distância **Euclidiana** e a distância de ***Manhattan*** (SINWAR; KAUSHIK, 2014). A distância **Euclidiana** é a mais simples e é obtida com a aplicação do teorema de pitágoras a dois ponto em um espaço euclidiano, sendo assim a distância Euclidiana entre dois pontos é definida como a raiz quadrada da soma dos quadrados dos catetos. A distância de *Manhattan* é definida como a soma das distâncias absolutas entre dois pontos (SINWAR; KAUSHIK, 2014). A Figura 1.18 mostra a definição de cada uma delas:

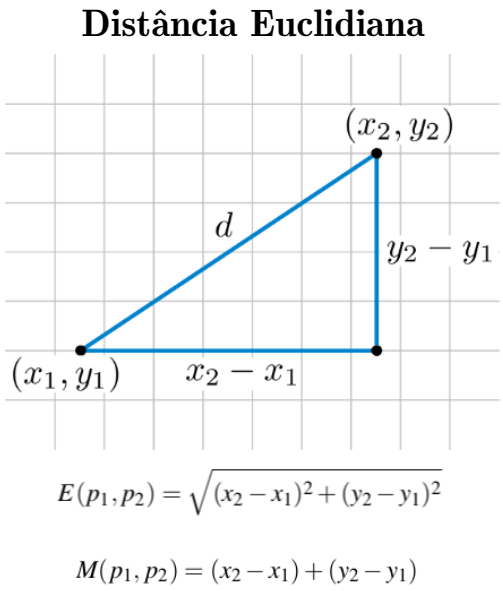


Figura 1.18: A definição de distância Euclidiana (E) e de *Manhattan* (M) entre dois pontos p_1 (x_1, y_1) e p_2 (x_2, y_2). Adaptado de `http://rosalind.info/glossary/euclidean-distance/`

Obviamente não existe uma resposta universal para determinação do K e métrica "ideais", isso é algo que faz parte do trabalho de **meta análise** do conjunto de dados. Para isso é necessário que o analista tenha em mente qual é o objetivo de sua análise, qual a pergunta que deve ser respondida, e a partir disso, delinear o processo de análise e meta análise dos dados. Na prática, é um processo de tentativa e erro, onde o analista deve avaliar o modelo para as configurações de parâmetros abordadas. Esse procedimento é tipicamente também chamado de hiper parametrização do modelo de aprendizado de máquina, e muitas vezes pode ser mais demorado e trabalhoso do que o próprio processo de análise (ZHANG, 2016).

Existem também a classe de **modelos não supervisionados**, a diferença primordial entre os supervisionados e os não supervisionados é que no primeiro observa-se a existência de uma variável alvo no conjunto de dados e no segundo não (ANG et al., 2015). Por esta razão, os métodos não supervisionados possuem um viés mais descritivo do que preditivo, como pode se observar na Figura 1.14. Assim, deste modo, esses métodos são utilizados principalmente em **análises exploratórias**. Onde o objetivo principal não é realizar previsões, mas delinear o comportamento do dados e determinar a dependência ou não das variáveis de um conjunto de dados. Esses métodos são de grande auxílio para cientistas que queiram conhecer melhor os dados com os quais estão trabalhando. Hoje em dia, até existem métodos de AM chamados de semi-supervisionados, que combinam os dois tipos (supervisionados e não supervisionados) (ANG et al., 2015).

Duas classes de métodos não supervisionados muito utilizadas são as clusterizações (agrupamentos) e a redução de dimensões. Na primeira classe, clusterização, o objetivo é descrever os dados a partir de uma medida de similaridade, para realizar o agrupamento dos mesmos (TARCA et al., 2007). Existem muitos tipos de algoritmos de clusterização, sendo que o KNN, K-*means* a clusterização hierárquica são alguns dos mais tipicamente utilizados para dados biológicos (TARCA et al., 2007).

K-*means* é um algoritmo não supervisionado e iterativo de clusterização que visa encontrar o máximo local em cada iteração considerando que, no conjunto de dados fornecido, existem k *clusters* (KODINARIYA; MAKWANA, 2013). Este algoritmo funciona nestes 5 passos:

1. Especificar o número desejado de *clusters* K;
2. Atribuir aleatoriamente cada ponto de dados a um *cluster*;
3. Determinar o centróide de cada *cluster*;
4. Atribuir cada ponto a um *cluster* cujo centroide esteja mais próximo;
5. Recomputar os centroides nos novos *cluster*s formados e repetir o processo de reagrupamento;
6. Repetir os passos 4-5 até atingir uma configuração optimal.

Na Figura 1.19, é ilustrado o processo da clusterização K-*means* com $k = 2$ para um conjunto de dados artificial onde, novamente, a expressão do gene G é medida em função da concentração de um fármaco arbitrário. Ambas as colunas tiveram seus valores normalizados para uma escala ente 0 e 1.

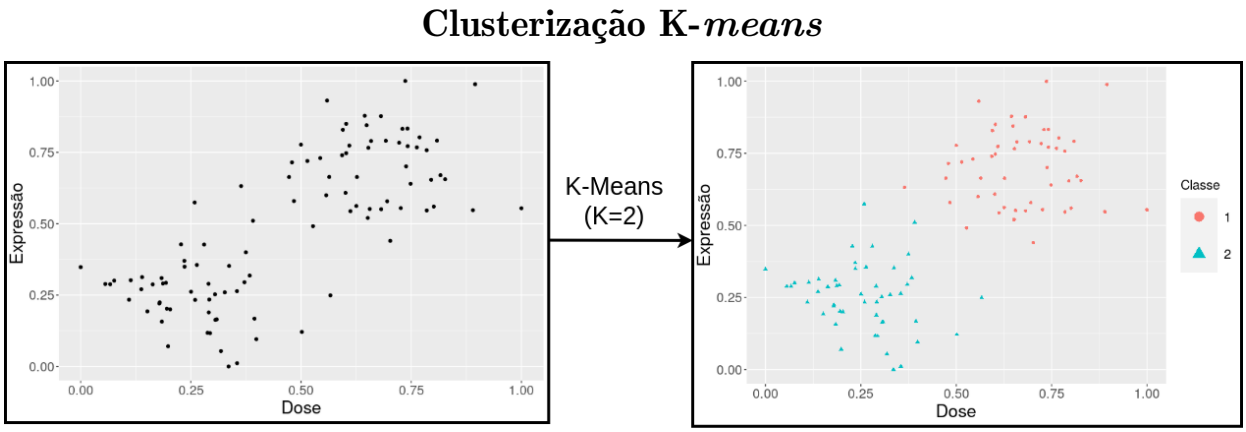


Figura 1.19: Realizando a clusterização K-*means* de um conjunto de dados artificial que registra a expressão do gene G em função concentração de um fármaco arbitrário. À esquerda, existe um *scatter plot* do conjunto de dados, é visível que este não tem o comportamento linear que se observa na Figura 1.15. À direita, o mesmo gráfico com os pontos coloridos segundo a classe (*cluster*) que foi dada a cada um deles segundo o algoritmo k-*means* com k=2 que foi aplicado ao conjunto de dados.

A vantagem deste método não supervisionado é que mesmo não tendo uma variável alvo, ele permite facilitar a visualização de comportamentos e tendências no conjunto de dados(KODINARIYA; MAKWANA, 2013). No exemplo da Figura 1.19, fica óbvio que a partir da dose 0.5, o fármaco induz uma expressão gênica média consideravelmente maior.

A clusterização hierárquica é um método de aprendizado de máquina não supervisionado cujo objetivo é dividir os dados em um hierarquia de *clusters* determinada por uma medida de similaridade entre os pontos (MURTAGH; CONTRERAS, 2012). Existem, tipicamente, duas abordagens para isso: aglomerativa e divisiva. A aglomerativa, também conhecida com ***botton-up***, é uma abordagem onde um conjunto de dados com n observações inicia o processo com n *clusters* (cada ponto é um cluster). De maneira sucessiva, o algoritmo aglomera os dois pontos mais similares, de maneira que no final existe um único grande *cluster*, composto por todos os outros que foram determinados. No final, obtém-se um dendrograma, que é um diagrama que ilustra os processos sucessivos de aglomeração dos elementos (ponto ou agrupamento de pontos) (MURTAGH; CONTRERAS, 2012).

A abordagem divisiva é justamente o oposto da aglomerativa, o conjunto de dados todo é considerado um grande *cluster* e este é sucessivamente dividido em *clusters* menores cujos componentes tenham a maior similaridade possível. Isso é feito até que sejam determinados n *clusters*, onde n é o número total de observações no conjunto de dados. Ou seja, no final, cada observação do conjunto de dados é em si um *cluster*, esta abordagem também gera um dendrograma como resultado (MURTAGH; CONTRERAS, 2012).

É mencionado que tanto a abordagem aglomerativa quanto a divisiva se utilizam do conceito de "similaridade" para determinação dos *clusters*. A determinação desse conceito tem duas frentes: a métrica de similaridade e o método de medida de similaridade (*Linkage criteria*) (SASIREKHA; BABY, 2013). A métrica de similaridade é tipicamente considerada como uma das muitas métricas de distância como a euclidiana e *Manhattan*. Métodos de determinação de similaridade (*Linkage Criteria*) são operadores que são aplicados a métrica de distância definida que têm a função de quantificar e determinar, de fato, o conceito de similaridade entre dois pontos quaisquer no conjunto de dados (SASIREKHA; BABY, 2013). A Tabela 1.4 mostra alguns desses métodos e seus critérios para determinação da similaridade máxima que define a formação sucessiva de *clusters* no dendrograma da clusterização hierárquica.

Clusterização K-*means*

Tabela 1.4: Métodos de determinação de similaridade na clusterização hierárquica e seus respectivos critétios (SASIREKHA; BABY, 2013)

Nome	Critério
Maximun Linkage/ Complete Linkage	Similaridade máxima é definida como o valor máximo da métrica das distâncias entre os elementos do conjunto
Minimum Linkage/ Single Linkage	Similaridade máxima é definida como o valor mínimo da métrica das distâncias entre os elementos do conjunto
Unweighted average linkage clustering (or UPGMA)	Similaridade máxima é definida como o valor da média aritmética da métrica das distâncias entre os elementos do conjunto
Weighted average linkage clustering (or WPGMA)	Similaridade máxima é definida como o valor da média ponderada das métricas de distância entre os elementos do conjuntos
Centroid linkage clustering or (UPGMC)	Similaridade máxima é definida como o valor do centróide das métrica de distância entre os elementos do conjunto
Ward	Similaridade máxima é definida como a fusão de elementos que minimiza a variância interna do novo cluster.

Para exemplificar melhor como o método funciona considere ainda o mesmo conjunto de dados de expressão do gene G fictício utilizado na Figura 1.19. Ao utilizar a função *hclust* da linguagem R para aplicar a clusterização hierárquica, com método de *Ward* e métrica de distância Euclidiana, foram gerados os gráficos presentes na Figura 1.20.

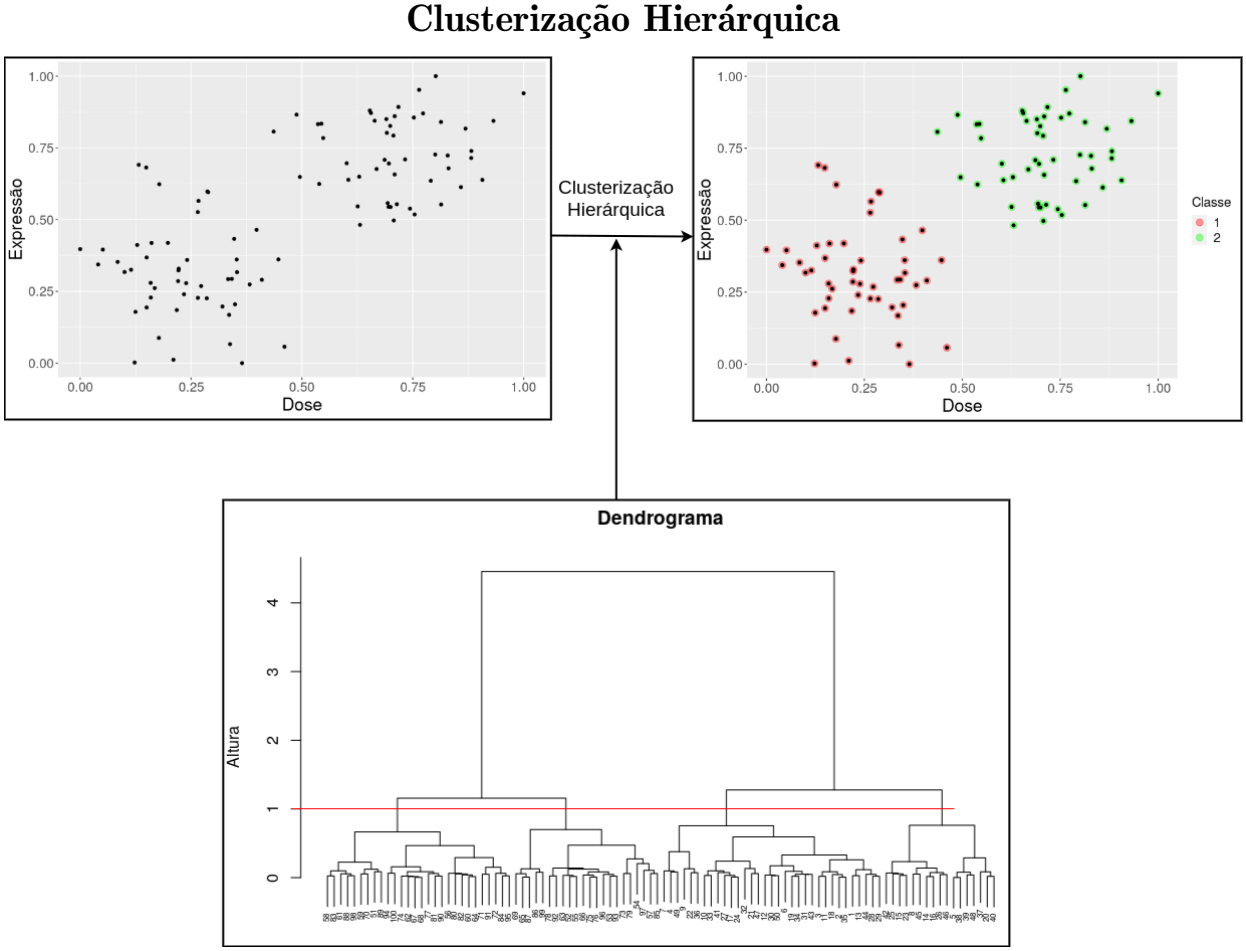


Figura 1.20: Realizando a clusterização hierárquica de um conjunto de dados artificial que registra a expressão do gene G em função concentração de um fármaco arbitrário. À esquerda, existe um *scatter plot* dos pontos do conjunto de dados onde o eixo x representa a dose e o eixo y representa a expressão do gene G fictício. A clusterização hierárquica é aplicada com auxílio da função *hclust* na linguagem R, considerando o método de *Ward* com métrica de distância Euclidiana. Essa função (*hclust*) acaba por gerar o dendrograma na parte inferior da figura. Note a linha vermelha na altura 1, tudo numa altura igual ou maior que ela é considerada um *cluster* válido. Sendo assim, fica fácil visualizar no próprio dendrograma dois grandes *clusters*. O gráfico à direita é idêntico ao da esquerda, mas agora os pontos estão classificados (1 ou 2) segundo o modelo gerado pelo dendrograma. Apesar da clusterização hierárquica ser um método bem diferente do K-*means* mostrado na Figura 1.19, ambos permitem a visualização de um mesmo comportamento, a de que a dose superior a 0.5 parece causar um aumento considerável na expressão média do gene G.

Uma outra classe de métodos de aprendizado de máquina não supervisionado que pode se observar na Figura 1.14 é a **redução de dimensão**. A ideia central dessa classe de métodos é reduzir o número de dimensões ou variáveis do conjunto de dados com o qual se está trabalhando (JOLLIFFE; CADIMA, 2016), de maneira a se facilitar a visualização de quais

variáveis são responsáveis pela maior parte das variações observadas no conjunto de dados. A análise de componentes principais (PCA), é o método mais utilizado nessa classe. Apesar de ser considerado um método de redução de dimensão, ele não vai fazer isso diretamente, ou vai garantir que isso possa ser feito para o conjunto de dados em questão. Entretanto, o PCA permite que o pesquisador consiga visualizar as variáveis que mais contribuem para a variância intrínseca do conjunto de dados. Dessa maneira, o analista poderá, por iniciativa própria, determinar que, por exemplo, duas entre mil variáveis do conjunto de dados são suficientes (JOLLIFFE; CADIMA, 2016).

Em resumo, a análise de componentes principais é um método de transformação linear ortogonal. O seu objetivo é determinar as autointituladas componentes principais, a partir da variância do conjunto de dados. Essa componente se tornara um novo eixo de um gráfico para ilustrar os dados e corresponde a variância de uma das variáveis do conjunto de dados. O eixo x do novo gráfico será a componente principal 1 (PC1) e corresponde a maior variância calculada, o eixo y seria a PC2 e corresponde a segunda maior variância calculada (JOLLIFFE; CADIMA, 2016). A Figura 1.21 mostra as componentes principais 1 e 2 calculadas para um conjunto de dados aleatório.

Cálculo das componentes principais

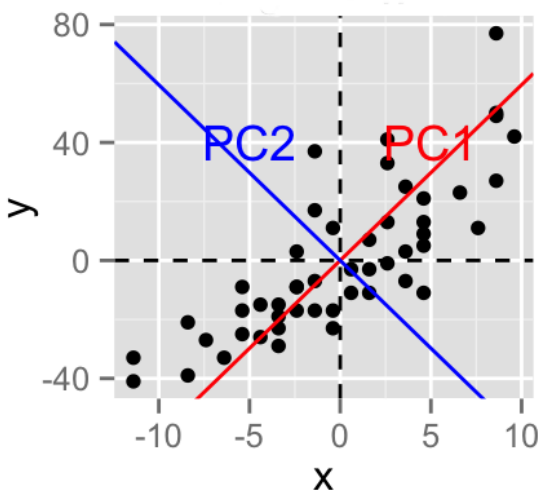


Figura 1.21: Cálculo das componentes principais 1 e 2 para um conjunto de dados aleatório. O próximo passo seria representar o conjunto de dados em um novo espaço bidimensional onde PC1 seria o eixo x e PC2 o eixo y. Note que a ordem da componente indica sua importância, ou seja, a componente PC1 tem uma contribuição (variância) maior que PC2, PC2 maior que PC3, etc...

Por essa descrição, é possível perceber que PCA permite ao analista ilustrar um conjunto de dados de 3 variáveis em duas dimensões, naturalmente isso não teria muita utilidade para conjuntos de dados com 5 ou mais variáveis, a menos que, por exemplo, 2 variáveis entre outras 998 sejam responsáveis por 70% da variância observada no conjunto de dados. Este seria um caso no qual o analista poderia descartar algumas variáveis para observar o comportamento dos dados segundo PC1 e PC2. Mas, mesmo que esse não fosse o caso, a PCA ainda teria utilidade, pois, ela pode ser aplicada a qualquer número de dimensões maior ou igual a 2. Assim seria possível determinar a importância que cada uma das variáveis tem para a variância intrínseca do conjunto de dados, oque por sí só é uma informação muito útil.

Na PCA, a determinação das componentes principais é dada com auxílio das seguintes fórmulas (JOLLIFFE; CADIMA, 2016):

$$V_i = \frac{1}{n-1} \sum_{m=1}^{n} (X_{im} - \overline{X_i})^2 \tag{1.4}$$

Onde V_i é a variância da coluna/atributo i do conjunto de dados (JOLLIFFE; CADIMA, 2016):

$$C_{ij} = \frac{1}{n-1} \sum_{m=1}^{n} (X_{im} - \overline{X_i})(X_{jm} - \overline{X_j}) \tag{1.5}$$

Onde, C_{ij} é a covariância entre as variáveis i e j; X_{im} é valor da variável i no ponto m e analogamente X_{jm} é valor da variável j no ponto m; $\overline{X_i}$ é a média da variável i e analogamente $\overline{X_j}$ é a média da variável j. Assim sendo, o termo $\frac{1}{n-1}\sum_{m=1}^{n}(X_{im} - \overline{X_i})(X_{jm} - \overline{X_j})$ representa o somatório da covariância de todos os m pontos. O raciocínio geométrico da PCA é rotacionar os n eixos de um espaço n-dimensional de forma tal que seja determinado um novo eixo que maximize a variância interna da variável i, que seja PC1 (componente pricipal 1). Em seguida, o PCA determina o eixo que tem a segunda maior variância PC2 e que tenha uma covariância nula com PC1. Desta maneira, a ortogonalidade dos eixos é garantida (JOLLIFFE; CADIMA, 2016). Depois disso, o eixo com a terceira maior variância PC3 e com covariância nula a TODOS os eixos previamente encontrados é determinado e assim sucessivamente.

Para exemplificar a aplicação da PCA, o conjunto de dados *wbdc.data* que pode ser obtido no sítio `http://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsi`

n+%28Diagnostic%29 será utilizado. Esse conjunto de dados é constituído por 32 *features* de um diagnóstico de tumor de mama. As primeiras 5 observações das primeiras 10 colunas são mostradas na Tabela 1.5:

Dataset wdbc.data

Tabela 1.5: Conjunto de dados *wdbc.data*: As primeiras 5 observações das primeiras 6 colunas. Este conjunto de dados associa o diagnóstico benigno (B) ou maligno (M) junto com 30 variáveis de resultados do exame realizado.

	id	diagnosis	radius_mean	texture_mean	perimeter_mean	area_mean
1	842302	M	17.99	10.38	122.80	1001.00
2	842517	M	20.57	17.77	132.90	1326.00
3	84300903	M	19.69	21.25	130.00	1203.00
4	84348301	M	11.42	20.38	77.58	386.10
5	84358402	M	20.29	14.34	135.10	1297.00

Foi utilizada a função *prcomp* do R para aplicar o PCA às colunas 3 a 32 do conjunto de dados *wdbc.data*, com os parâmetros "*center=TRUE*" para centralização da matriz fornecida e "*scale=TRUE*" para normalização dos valores centralizados. Esse procedimento, essencialmente, transforma a matriz numérica de entrada numa matriz de *z-scores*, ou seja, uma matriz onde o valor de cada célula mostra apenas quantos desvios padrões o valor de expressão original está acima (valores positivos) ou abaixo (valores negativos) da expressão gênica média da coluna. A Tabela 1.6 mostra o sumário da PCA para os 6 primeiros componentes determinados (PC).

Sumário de Componentes Principais

Tabela 1.6: Sumário da PCA: Pode-se observar o desvio padrão, porcentagem da variância e porcentagem cumulativa das componentes principais nas linhas da tabela, enquanto nas colunas observa-se os componentes (PC).

	PC1	PC2	PC3	PC4	PC5	PC6
Desvio Padrão	3.64	2.39	1.68	1.41	1.28	1.10
Porcentagem de Variância	0.44	0.19	0.09	0.07	0.05	0.04
Porcentagem Cumulativa	0.44	0.63	0.73	0.79	0.85	0.89

É notável o fato de que somente 6 das 30 variáveis são responsáveis por quase 90% da variância intrínseca do conjunto de dados. Agora são ilustrados todos os pontos em função das componentes PC1 e PC2, os quais foram coloridos de acordo com seu diagnóstico benigno (B) ou maligno (M):

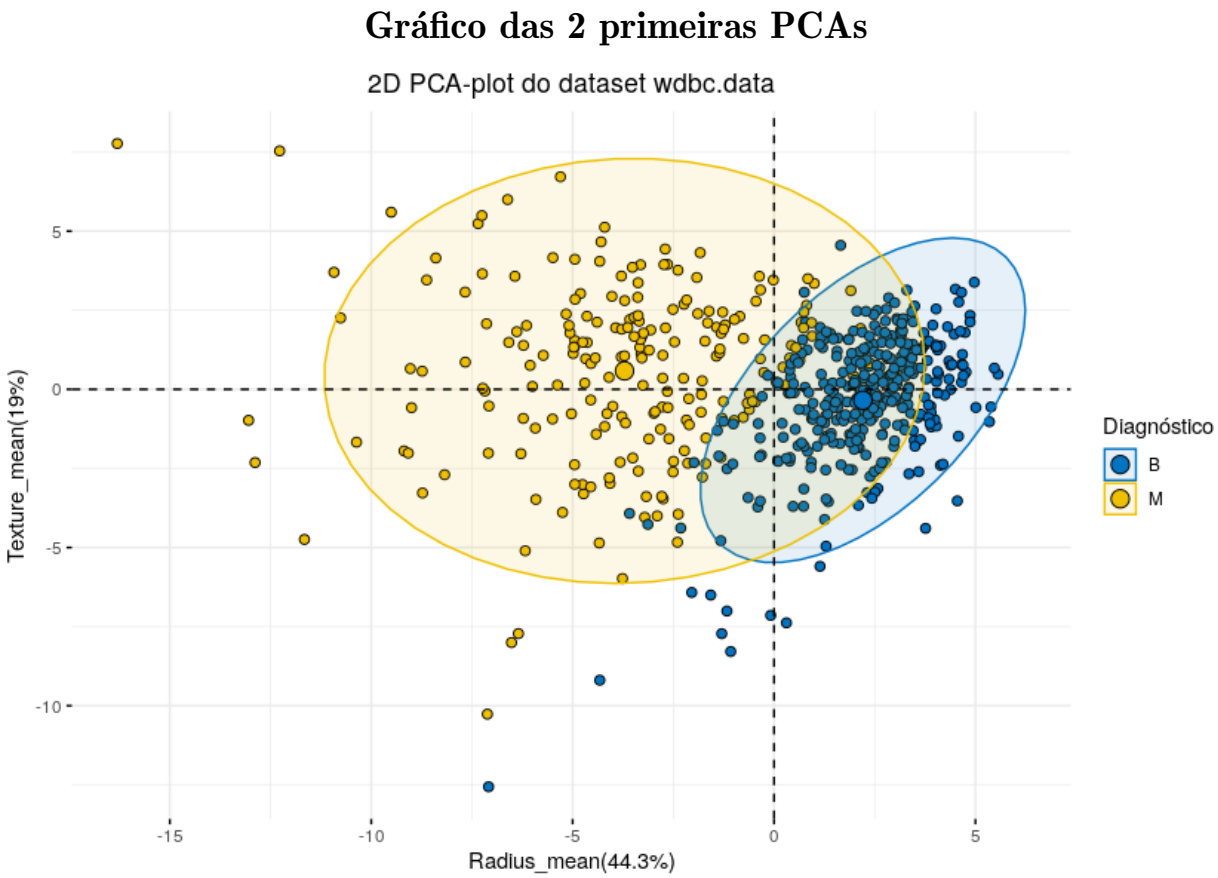


Figura 1.22: Gráfico do PCA do conjunto de dados *wdbc.data*: O gráfico das componentes PC1 e PC2, colorido segundo o diagnóstico benigno (B) ou maligno (M) do tumor. Pode se observar que o modelo facilita a visualização do comportamento de clusterização/agrupamento dos dois diagnósticos.

Esse simples estudo de caso mostra como o PCA pode ser utilizado para visualização de grandes quantidades de dados. Mesmo quando o conjunto de dados em questão possui um número elevado de variáveis.

1.5 Bioinformática: Aplicações de AM e Tendências

As aplicações do aprendizado de máquina são amplas e contemplam desde a solução de problemas do cotidiano como controle de fraude de cartões de crédito à solução de problemas em área de pesquisa como a bioinformática (NASRABADI, 2007).

De fato, a bioinformática é uma área em grande estudo atualmente, devido principalmente ao grande influxo de dados produzidos pelas ciências ômicas. Existem um grande número de aplicações do aprendizado de máquina nos dados das ciências ômicas: descoberta

de novas classes nos dados de entrada por meio de algorítimos não supervisionados como é feito por Wang e colaboradores, onde clusterização não supervisionada é utilizada para desenvolver um modelo de classificação de tumores gástricos e aprimorar prognóstico da doença (WANG et al., 2018). Outra aplicação seria a criação de modelos preditivos, isso é realizado por Gilmore e colaboradores (2010), no qual, *suport vector machines* são utilizadas para reconhecimento de tecidos com melanoma ou também como Ayer e colaboradores (2010) utilizaram *artificial neural networks* para determinação de um modelo de predição de risco de câncer mamário por meio da integração de dados clínicos, moleculares, genômicos e demográficos (GILMORE; HOFMANN-WELLENHOF; SOYER, 2010; AYER et al., 2010).

Uma tendência na bioinformática é que os métodos do aprendizado de máquina têm sido utilizados em conjunto com as redes de interação para modelagem e estudos dos elementos que compõe um sistema biológico. Isso é importante porque o uso de redes pode não apenas facilitar a visualização de resultados no contexto biológico como também pode fornecer mais *features* para um método de aprendizado de máquina, como por exemplo tipos de interação entre os nodos, quantidade de interações e centralidade. As *features* são propriedades mensuráveis do fenômeno de interesse, a determinação desses parâmetros influencia diretamente a saída de um algoritmo (NASRABADI, 2007). Dessa maneira, além de *features* normalmente utilizadas para estudo dos dados (como expressão gênica) é possível utilizar novas *features* que levam em consideração parâmetros associados com as interações dos elementos (genes, transcritos e proteínas) de um sistema biológico sob determinada condição (LI; WU; NGOM, 2016). No trabalho de Xing e colaboradores (2018), por exemplo, dados de câncer renal são utilizados para avaliação da expressão de *clusters* de lncRNAs em uma rede integrada com miRNAs através de um modelo de regressão de cox (XING et al., 2018).

Tal abordagem é muito útil para estudar vias biológicas alteradas em diferentes condições, por exemplo, estudos toxicogenômicos com o auxílio de redes e de algoritmos de aprendizado de máquina podem determinar genes ou vias metabólicas significantemente alteradas que podem explicar a toxicidade ou efeitos adversos de uma droga como é explicado em (BAREL; HERWIG, 2018). A mesma abordagem pode ser utilizada para estudo de doenças como o câncer, como é apresentada por Kourou e colaboradores (2015), onde é feita uma revisão de métodos de aprendizado de máquina utilizados para predição de prognósticos da doença (KOUROU et al., 2015).

1.6 Câncer: Integrando Dados Biológicos com Anotações Funcionais e Clínicas

Estudos que usam dados biológicos e aprendizado de máquina são utilizados para reconhecimento de padrões nos dados de entrada com finalidade de determinar associações desses padrões com diversas variáveis e então realizar avaliação de prognósticos e predições. Essas variáveis são as anotações associados aos dados de entrada, no caso de um estudo que pretenda utilizar aprendizado de máquina para derivar um modelo que consiga predizer chances de sobrevivência de um paciente seria necessário associar os dados de entrada a anotações clínicas sobre a sobrevivência, um desse tipo é realizado por Zhang e colaboradores (ZHANG et al., 2016). No caso de um estudo que queira estudar vias biológicas do câncer, seria necessário o uso de anotações funcionais associadas aos dados de entrada. Exemplos de anotações clínicas que modelam o sistema biológico e as associam a genes são o *Gene Ontology* e o *KEGG* (*Kyoto Encyclopedia of Genes and Genomes*).

Em 2004, Segal e colaboradores desenvolveram uma maneira interessante de se estudar grandes conjuntos de dados sobre o câncer. A metodologia é chamada de mapa modular (*modular map*) e consiste-se em utilizar dados de expressão gênica de microarranjo em conjunto com anotações clínicas dos experimentos e anotações funcionais para a criação de um mapa compreensivo dos dados. No fim, esse método gera um grande *heatmap* que define módulos funcionais para os genes (com auxílio das anotações funcionais) e associa tais módulos a condições clínicas. Dessa maneira esse "mapa" permite analisar quais módulos funcionais estão super ou subexpressos para determinada condição clínica. Esse método é uma maneira excelente de se realizar estudos exploratórios de grande quantidades de dados biológicos e pode ser facilmente adaptado para outras situações (SEGAL et al., 2004).

No estudo mencionado, Segal e colaboradores utilizam o método para determinar *clusters* de genes associados a funções biológicas e condições clínicas para um conjunto de dados referente a 11 tipos de tumores. O mapa modular apresenta observações interessantes como módulos do ciclo celular induzidos na presença de câncer ou como o módulo da inibição de crescimento está inibido na presença da doença (câncer de fígado e leucemia) (SEGAL et al., 2004).

1.7 Justificativa

Nos tempos atuais, métodos que consigam integrar grandes quantidades de diferentes tipos de dados biológicos para derivar *insights* são muito necessários aos pesquisadores. O uso de métodos de aprendizado de máquina é uma das soluções mais utilizadas no *design* de tais metodologias. O mapa modular criado por Segal e colaboradores (2004) utiliza dados biológicos e anotações (clínicas e funcionais) junto com algorítimos de clusterização hierárquica (aprendizado de máquina supervisionado) e é um bom exemplo dessas metodologias (SEGAL et al., 2004). Entretanto, existe uma limitação desse método que nos dias de hoje pode ser resolvida com auxílio da base de dados TCGA: o método de Segal utiliza apenas dados de expressão gênica de microarranjo.

Atualizar essa metodologia para que possa integrar outros tipos de informações biológicas às anotações funcionais pode fornecer perspectivas diferentes sobre o câncer que possuem o potencial de desvendar novas informações sobre a doença. Está é uma hipótese com respaldo na literatura científica, pois, sabe-se que o câncer é uma doença multifatorial. Perfís metabólicos e transcriptômicos são tradicionalmente mais utilizados para caracterização de tumores, entretanto o TCGA permite a utilização de outras informações associadas a epigenética (metilação) e regulação pós-transcricional (miRNAs). Atualmente, o TCGA possui esses dados mencionados anteriormente para mais de 30 tipos de tumores.

2 Objetivos

2.1 Objetivos Gerais

Com todas essas considerações feitas, essa tese propõe uma atualização da metodologia de Segal onde o mapa modular será adaptado para uso de dados de expressão de *RNA-Seq* de mRMA e miRNA e de dados de metilação. Também é considerada a utilização de um maior número de tipos de tumores para a construção dos mapas e alterações no algoritmo de clusterização (aprendizado de máquina). Com o intuito de facilitar a visualização de eventuais observações realizadas com este mapa, será criada uma rede que integre as interações: mRNA-mRNA (RNA-seq e metilação) e miRNA-mRNA (miRseq). Assim sendo, o objetivo geral desta tese, é o desenvolvimento de uma metodologia integrativa para avaliação de tecidos tumorais com o uso de três níveis de informação biológica (dados de RNA-seq, metilação e miRNA). Essa metodologia é baseada em *machine learning* não supervisionado, para obter um mapa modular que é composto por tumores e seus processos biológicos.

2.2 Objetivos Específicos

- Análise dos módulos reprimidos/induzidos dos dados de RNA-seq;
- Análise dos módulos reprimidos/induzidos dos dados de metilação;
- Análise dos módulos reprimidos/induzidos dos dados de miRNA;
- Uso de redes biológicas no mapa modular;
- Avaliar o comportamento dos módulos para os tipos de tumores;
- Avaliação do BRCA para o tumor de mama.

3 Material e Métodos

3.1 O Mapa Modular

Esta é uma metodologia criada por *Eran Segal* e colaboradores em (SEGAL et al., 2004), cuja finalidade é caracterizar condições experimentais (doença, por exemplo) por meio da definição de módulos induzidos e inibidos, os quais são compostos por conjuntos de genes. A partir de dados de expressão gênica e de conjuntos de genes previamente definidos, que funcionam como o entrada do método, determinam-se genes induzidos e/ou reprimidos dentro dos conjuntos de genes originais de maneira que é possível atribuir significado biológico para esses os módulos. Assim, as condições associadas aos dados de entradas podem ser caracterizadas e estudadas de maneira a elaborar hipóteses sobre o comportamento do organismo perante a doença ou condição em estudo. A metodologia original desenvolvida em (SEGAL et al., 2004), dependia do uso de dados de expressão gênica de 1975 *chips* de microarranjo considerando 11 tumores.

Para este trabalho, a metodologia foi adaptada para uso de dados provindos de experimentos de RNA-seq e mais 2 tipos de dados: micro RNA (miRseq) e metilação. O TCGA disponível no *Genomic Data Commons* (GDC), descrito na Subseção 1.2.1, atualmente, possui dados referentes a 33 tipos de tumores. Foram prospectados os 3 tipos de dados no GDC, para todos os casos, em todos os tumores disponíveis apenas no projeto *TCGA*. Como condição de seleção, os casos obviamente deveriam ter os 3 tipos de dados, a sua anotação clínica e ter acesso livre (TCGA, 2018).

Nossa adaptação do mapa modular segue o *workflow* ilustrado na Figura 3.1. Para a adaptação do mapa modular, foram utilizados pouco menos de 10000 casos, pois esse é o número de situações em que todas as condições para seleção são satisfeitas. O número exato varia com o tipo de dado (RNA-seq, metilação e miRseq) e serão mencionados nas subseções

abaixo. Todos os processos descritos nas seções abaixo foram realizados no ambiente *Rstudio* da linguagem R.

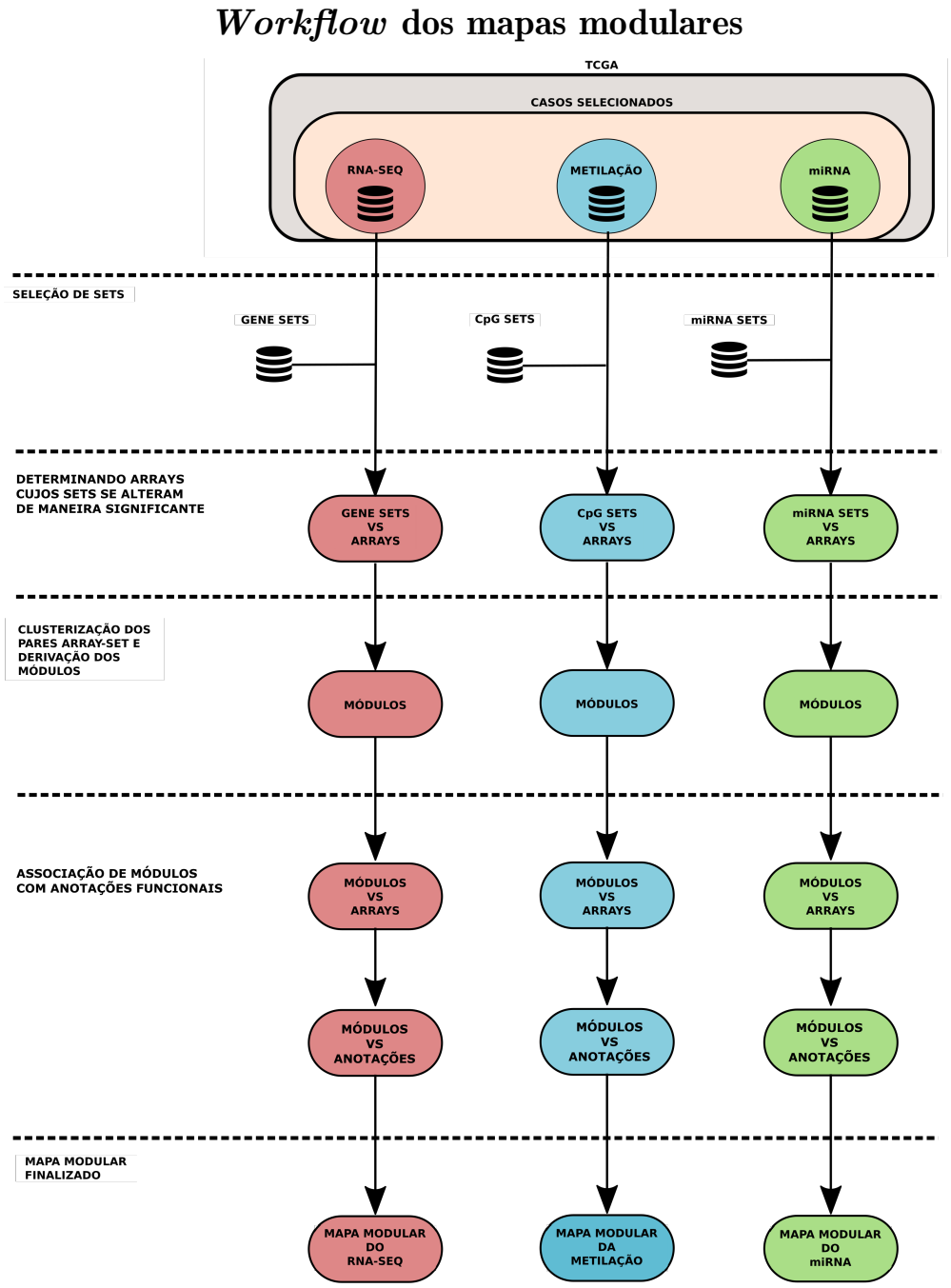


Figura 3.1: *Workflow* descrevendo os passos da metodologia utilizada para obtenção dos 3 tipos de mapas modulares. Cada tipo de dado é indicado por cor: rosa (RNA-seq), azul (metilação) e verde (miRNA). Note que o diagrama é unidirecional, demonstrando o fluxo de informação, bem como qual é a entrada e a saída do método. O *workflow* é iniciado representando a seleção de dados de interesse no TCGA e é finalizado com a construção de cada mapa modular. Note que, existem várias etapas que precisam ser realizadas entre a coleta de dados e a finalização do mapa modular. Tais etapas são separadas umas das outras pelas linhas pontilhadas. Abaixo de cada uma delas existe uma breve descrição do que é realizado na respectiva etapa. Cada uma dessas etapas é discutida com detalhe ainda nesta seção.

3.2 Coleta e Pré-processamento de dados

A coleta de dados brutos foi feita com auxílio da linguagem de programação R. Foi desenvolvido um *script* para *download* dos dados brutos dos 33 tumores a partir do repositório `http://gdac.broadinstitute.org/runs/` (*GDAC*). Nele, estão dispostos publicamente diversas versões dos dados brutos que também são encontrados no TCGA. Primeiramente, o *script* determinou qual a versão mais recente dos dados que estavam publicamente disponíveis, no caso, a versão da data *2016/01/18* era a mais atual.

É importante mencionar que o domínio *GDAC*, por padrão, realiza algumas filtragens antes de disponibilizar os dados para *download*. Essas filtragens podem ou não ser realizadas pelo TCGA e por isso pode haver discrepância entre o número de amostras disponíveis no TCGA e no domínio *GDAC*. Em resumo, existem 4 tipos de filtragens feitas pelo *GDAC*: de replicatas, remoção por submissor, remoção por *black list* e remoção por *FFPE*.

Algumas amostras possuem várias replicatas e por padrão o domínio *GDAC* dispõe apenas 1 replicata por amostra. Assim, o domínio *GDAC* segue um rigoroso protocolo para selecionar apenas uma das replicatas. Detalhes desse protocolo podem ser encontrados nos *reports* do domínio *GDAC*, como exemplo tem-se o link: `http://gdac.broadinstitute.org/runs/stddata__2015_02_04/samples_report/COAD_Replicate_Samples.html`. O TCGA não faz a filtragem descrita nesse paragrafo, disponibilizando todas as replicatas.

Outras maneiras de filtragem incluem remoções de amostras que são requisitadas pelos seus próprios submissores e aquelas que, por decisão do próprio do próprio domínio, são removidas (*black list*). Ambas essas filtragens são realizadas também pelo TCGA, devido a sua natureza. Por fim, existe a remoção de amostras FFPE (*Formalin Fixed Paraffin Embedded*), pois segundo o domínio *GDAC*, estas amostras não são apropriadas para análises de biologia molecular. Esta última filtragem não é realizada pelo TCGA, as amostras FFPE estão disponíveis para *download*.

Abaixo é detalhada a coleta e pré-processamento de cada um dos 3 tipos dados.

3.2.1 RNA-seq: Coleta e Pré-processamento

Os dados de RNA-seq de cada um dos 33 tumores são obtidos de seus respectivos repositórios em `http://gdac.broadinstitute.org/runs/`, por meio de arquivos compactados

do tipo **.tar.gz*. Os arquivos têm uma nomenclatura grande, que envolve a concatenação da abreviação do tumor no TCGA e o tipo de dado que este armazena. Por exemplo, "*BLCA.mRNAseq Preprocess.Level_3*" é uma parte do nome do arquivo que contém os dados de contagem bruta de transcritos de mRNA. O domínio *GDAC* e por extensão, o TCGA, usam o mesmo protocolo para traçar o perfil transcriptômico a partir dos dados de sequenciamento, independente de sua plataforma, esse protocolo é brevemente descrito na Figura 1.11. Assim, no final, ficam disponíveis, os arquivos de contagem bruta (*HTSeq Counts*) e de contagens normalizadas *HTseq FPKM/FPKM-UQ*. Neste estudo foram utilizados os arquivos de contagem bruta (*HTSeq Counts*), pois assim, é possível aplicar as metodologias de normalização que forem pertinentes. Cada coluna desse arquivo *HTSeq Count* armazena a contagem de transcritos de uma amostra de algum indivíduo e cada linha corresponde ao *Gene Symbol* de algum gene.

Os arquivos com o *Htseq Count* de cada tumor foram baixados e armazenados em diretórios, que depois são lidos pela linguagem R no *Rstudio* em forma de "*dataframe*". A seguir foram selecionadas apenas as amostras que estão simultaneamente disponíveis no TCGA e no arquivo prospectado. Isso foi feito por meio do *download* dos meta-dados do tumor em questão com auxílio da biblioteca em R *TCGAbiolinks*. Posteriormente foram selecionadas as amostras de interesse, no caso, de tumores primários, pois a princípio avaliar a questão da reincidência e da metástase nesse estudo é algo não desejado. O tumor de leucemia (LAML, *Acute Myeloid Leukemia*) é uma exceção nesse caso, uma vez que não há amostras de controle, apenas amostras de sangue periférico, que foram utilizadas como amostras tumorais. Outro caso a ser mencionado é o do tumor mamário (BRCA). No arquivo, é possível encontrar poucas amostras de indivíduos do sexo masculino, essas amostras são removidas de modo que o conjunto final tenha apenas indivíduos do sexo feminino.

A seguir, as amostras que compõe cada tumor são organizadas em *dataframes*, tumor-a-tumor, totalizando 33 *dataframes* para os dados de RNA-seq, uma para cada tumor. Estes *dataframes*, contêm valores brutos de expressão gênica para os dados de RNA-seq, as colunas são amostras enquanto linhas são sondas/genes. Em seguida, é realizada uma remoção de amostras que estar repetidas, como mencionado anteriormente nesta Seção 3.2, o GDAC dispõe apenas uma amostra por caso, sem replicatas. Portanto amostras repetidas precisam ser removidas. Toda a etapa de prospecção de dados bruto descrita até o momento, está ilustrada na Figura 3.2 na Seção 3.4.

Posteriormente, foi realizada a transformação dos dados destes *dataframes* para um escala $\log(2)$. Depois disso, as amostras de controle foram removidas de cada um dos 33 *dataframes*, quando estão presentes. A princípio não foram consideradas as amostras de controle, somente as de tumores.

Posteriormente, cada uma dos 33 *dataframes* passa por um processo normalização, remoção de *outliers* e de correção para efeitos de lote. Foi utilizada a função *normalizeBetweenArrays* do pacote *limma* para aplicar uma normalização do tipo *quantile* para os genes entre amostras (*arrays*) diferentes. É importante salientar que valores NA (*Not Available*) são desconsiderados no cálculo da normalização. Os valores NA serão tratados mais a frente.

A identificação e remoção de possíveis *outliers* foi feita utilizando a técnica do PCA (*Principal Component Analysis*), descrita na Seção 1.4, juntamente com o teste estatístico de Rosner. Com o PCA, foram determinados os componentes principais PC1 e PC2, que correspondem as 2 sondas responsáveis pelas duas maiores fontes de variância intrínseca, em cada um dos 33 *dataframes*. Com PC1 como eixo x e PC2 como eixo y, as amostras são transpostas para um novo espaço, onde os pontos/observações correspondem a variância ajustada da expressão/metilação de cada amostra no *dataframe* em questão. Neste nova configuração foi utilizado o teste de Rosner para identificação de *outliers*. Existem 2 paramêtros necessários este teste, o p-valor crítico "α" e número estimado de *outliers* "K". Empiricamente determinamos que "K" seria igual a um centésimo do número de amostras (colunas) do *dataframe* em questão, arredondado para cima. Sendo assim, o valor mínimo possível de "K" é 1. Quanto ao p-valor crítico "α", foi determinado, também empiricamente, que seu valor seria de 0.5. Após a aplicação do processo descrito acima, os *outiliers* detectados foram removidos dos *dataframes*.

As quatro principais causas de introdução de efeitos de lote nos dados do TCGA são: os lotes de processamento de amostras (*batch*), o centro onde as amostras são processadas (*center*), a local de coleta das amostras (*TSS*) e o sexo dos indivíduos (*sex*). Para mensurar a existência de efeitos de lote, foi utilizado o *Dispersion Separability Criterion* (DSC), criado pelo Instituto de Bioinformática *MDAnderson* e definido como:

$$DSC = \frac{D_b}{D_w} \tag{3.1}$$

Onde, D_b é a dispersão existente entre amostras de lotes diferentes e D_w é a dispersão

entre amostras dentro de um mesmo lote. Valores altos de DSC significam que existe uma dispersão maior entre amostras de lotes diferentes. As amostras que pertencem a um mesmo lote têm maior probabilidade de serem mais parecidas entre sí do que amostras de lotes diferentes. De acordo com o Instituto *MDAnderson*, valores entre 0 e 0.3 não indicam presença de efeitos de lote nos dados. Entre 0.31 e 0.5 existe uma baixa presença de efeitos de lote, se estes devem ser corrigidos ou não é algo que deve passar pelo crivo do pesquisador/analista. Por fim, valores mais altos que 0.5 indicam alta presença de efeitos de lote e o pesquisador é fortemente aconselhado a considerar metologias para corrigí-los.

Foram avaliados os valores de DSC para as quatro fontes de efeitos de lote em nossos dados RNA-seq, quando este superava o valor de 0.3, foi usado um método de correção de efeitos de lote implementado no pacote *limma*, na função *removeBatchEffect*. A correção de um tipo de efeito de lote (*batch*, por exemplo) pode afetar o DSC de outro (*center*). Por esta razão, foi repetido o processo constituído por medir o DSC, corrigir os efeitos, e medí-lo novamente, até garantir que este estivesse abaixo do valor 0.3 para todas as quatro fontes de efeitos de lote. É importante salientar que de maneira empírica, foi determinado que lotes com menos que 3 amostras seriam removidos do conjunto de dados, de maneira a reduzir o viés estatístico que é introduzido ao conjunto por um número muito pequeno de amostras. Dessa maneira a filtragem de dados em certas situações foi automatizada, como por exemplo a do mencionado tumor mamário (BLCA), onde temos mais de 98% dos indivíduos pertencentes ao sexo feminino e somente 3 amostras do sexo masculino. Todo esse procedimento de correção de efeitos de lote é aplicado separadamente aos 33 *dataframes* de dados de tumor. No final, foram obtidos 33 *dataframes*, totalizando 6.7 GB de dados de RNAseq contemplando 8904 casos. Essa etapa de pré-processamento esta ilustrada na Figura 3.3 na Seção 3.4.

3.2.2 Metilação: Coleta e Pré-processamento

O procedimento de *download* dos dados de metilação segue os mesmos passos descritos na Subseção 3.2.1 para dados de RNA-seq. O *script* identificou a última versão dos dados públicos no mesmo repositório (`http://gdac.broadinstitute.org/runs/`) e procurou, para cada um dos 33 tumores, arquivos *tar.gz* cujo nome continha a combinação do nome do tumor e a *string* "*Methylation_Preprocess.Level_3*". Esses arquivos contém amostras como colunas e ilhas CPG como linhas, cada célula dessa tabela contém valores beta de metilação.

As ilhas CPG e os valores beta de metilação são brevemente discutidos na introdução na Figura 1.13.

Em seguida, os dados de metilação de cada um dos 33 tumores são salvos em diretórios separados e são posteriormente lidos como "*dataframes*" no ambiente *Rstudio* da linguagem R. Para cada tumor são selecionadas as amostras de interesse: que estejam simultaneamente no TCGA e no domínio *GDAC* e que sejam tumores primários. Para cada tumor foi montado um *dataframe* com valores de metilação, amostras como colunas e linhas como sondas/genes. Foram obtidas 33 *dataframes*, uma para cada tumor. As amostras de controle foram removidas de cada uma dos 33 *dataframes*, quando existiam. Separadamente, a cada um destes 33 *dataframes*, foi aplicado o mesmo processo de normalização "*quantile*" do RNA-seq, remoção de *outliers* e de correção de efeitos de lote, todos estes estão descritos na Subseção 3.2.1. É importante salientar que os dados de metilação, chamados valores beta, existem em uma escala de 0 a 1, e por essa razão, não passaram pela fase de transformação de dados para escala $\log(2)$ que ocorre com os ddos de RNA-seq e de miRseq. Ao fim, foram obtidos 33 *dataframes*, totalizando de 11.3 GB de dados de metilação contemplando 8705 casos.

3.2.3 miRseq: Coleta e Pré-processamento

O procedimento de *download* dos dados de micro RNA segue os mesmos passos descritos na Seção 3.2.1 para dados de RNA-seq. O *script* identificou a última versão dos dados públicos no mesmo repositório (`http://gdac.broadinstitute.org/runs/`) e procurou, para cada um dos 33 tumores, arquivos **.tar.gz* cujo nome continha a combinação do nome do tumor e a *string* "*miRseq_Preprocess.Level_3*". Esses arquivos contêm amostras como colunas e miRNAs como linha. Cada célula dessa tabela contém valores de expressão gênica. Os experimentos para mensurar os niveis de expressão dos miRNAs são brevemente discutidos na Subseção 1.3.2.

Em seguida, os dados de miRNA de cada um dos 33 tumores são salvos em diretórios separados e são, posteriormente, lidos como *dataframes* no ambiente Rstudio da linguagem R. Durante essa etapa, foi descoberto que não existem dados de miRseq para o tumor GBM (*Glioblastoma Multiform*), por essa razão esse tumor foi excluído das análises, inclusive nos dados de RNA-seq e metilação.

Para cada um dos 32 *dataframes*, foram selecionadas as amostras de interesse: que estejam simultaneamente no TCGA e no domínio *GDAC* e que sejam tumores primários. Para cada *dataframe*, foram removidos amostras de controle, quando estas existiam. Depois, foram aplicadas, separadamente, a cada um destes 32 *dataframes*, o mesmo processo de normalização "*quantile*" do RNA-seq, remoção de *outliers* e de correção de efeitos de lote, todos estes estão descritos na Subseção 3.2.1. Ao fim, foi obtido um total de 165 MB de dados de miRseq contemplando 9161 casos.

3.2.4 Determinação de *Gene Sets*: *Gene Ontology* (GO)

A ferramenta *Gene Ontology* (GO) (ASHBURNER et al., 2000) está disponível no endereço eletrônico (`http://www.geneontology.org/`) e tem como objetivo classificar e organizar as descrições dos genes e seus respectivos produtos. O GO iniciou, em 1998, com somente três organismos: *Saccharomyces cerevisiae*, *Mus musculus* e *Drosophila melanogaster*, mas atualmente há mais espécies catalogadas, incluindo *Homo sapiens Rattus norvegicus* que são de interesse para esse trabalho. As informações foram organizadas em três classificações independentes: processos biológicos, componentes celulares e funções moleculares. Cada uma tem uma abordagem diferente em relação a associação do gene e seu produto.

- **Função Molecular (MF)**: Representa as atividades moleculares ao qual o gene está associado. Descreve, principalmente, ligações moleculares (ligação com enzimas, com oxigênio, ferro, ácidos graxos, etc) e atividade de enzimas específicas (aromatase, quinina 3-monoxigenase, testosterona 6-beta-hidroxilase, etc).
- **Compartimento Celular (CC)**: Representa as localizações celulares em que o produto do gene desempenha sua função. Descreve, principalmente, organelas (mitocôndria, nucleoplasma, citoplasma), membranas (membrana celular, nuclear, do retículo endoplasmático) e complexos moleculares (complexo MCM, ciclina B1-CDK1, Ndc80, etc).
- **Processo Biológico (BP)**: Representa os objetivos a serem atingidos pelas funções moleculares ao qual o gene está associado (replicação do DNA, processos metabólicos e catabólicos de substâncias, atividade de ontologias enzimáticas como epoxigenases, transporte de moléculas, etc).

As ontologias estão organizadas de maneira hierárquica, assim existem ontologias mãe e filhas sendo uma ontologia pode ser mãe de mais de uma ontologia filha e estas podem ter mais de uma mãe. Quanto maior for a hierarquia da ontologia mais geral e abrangente ela é, caso contrário a ontologia tende a ser cada vez mais específica. Graficamente, as ontologias são representadas em forma de árvores, esses gráficos também são chamados de grafos direcionados não-cíclicos.

Nesta tese, as ontologias do *Homo sapiens* foram utilizadas por meio do pacote "GO.db" para o R, disponível na plataforma *bioconductor*. As ontologias do tipo BP (*Biological Process*) foram a principal fonte de formação dos *sets* de genes que o método do mapa modular necessita como *input*. Atualmente (24/07/2020), existem 29699 ontologias do tipo BP, mas muitas delas são redundantes ou contém um número muito pequeno ou alto de genes anotados, por essa razão, foram aplicadas filtragens nesse conjunto com a finalidade de reduzir a quantidade de *Gene Sets* utilizados. A filtragem se baseou no número de ontologias filhas que cada uma possuía, foram selecionadas todas as ontologias BP que tinham anotados mais de 10 termos e menos de 50, totalizando 2299 delas.

3.2.5 *CpG Sets* e *miRNA Sets*

Os *CpG Sets* e os *miRNA Sets* são análogos aos *Gene Sets* e são utilizados como dados de entrada, respectivamente, nos mapas modulares da metilação e dos microRNAs. Estes dois últimos *Sets* foram derivados diretamente dos *Gene sets*, que, como mencionado na Subseção anterior, são as ontologias BP selecionadas após a filtragem. Para isso, foi necessário o uso de bases dados adicionais. Para montar os *CpG Sets*, foi utilizado o pacote "IlluminaHumanMethylation450kanno.ilmn12.hg19" do R para obter a relação CpG-Genes. Desta maneira, substitui-se cada gene dentro de um *Gene Set* pelo conjunto de CpG que o compõe, de forma que, no final, obtém-se o *GpG Set* correspondente (HANSEN, 2015). No final, foram prospectadas um total de 2299 GOs.

Os *miRNA Set* foram determinados de maneira análoga aos *CpG Sets*, as relações miRNA-GO foram obtidas no Rstudio com auxílio da API do endereço eletrônico `https://www.ebi.ac.uk/QuickGO`, que hospeda uma lista de relação entre miRNAs e GOs (*Gene Ontology*). Estas relações são agregadas a partir de outras bases de dados com o miRBase (KOZOMARA; GRIFFITHS-JONES, 2013) e o starBase (LI et al., 2013), as quais possuem referências para as relações derivadas no *QuickGO*. Os miRNAs tiveram uma atenção espe-

cial da comunidade científica na última década e por isso informações sobre os papéis que podem ter em processos biológicos ainda são limitadas. Isso resulta em uma quantidade de informações consideravelmente menor para o mapa modular de miRNAs em relação aos de RNA-seq e de metilação. No final, foram prospectadas apenas 426 GOs. Esta baixa quantidade se deve a pouca quantidade de informações disponíveis para os miRNAs.

3.2.6 Pré-processamento e Elementos Significantemente Alterados

Após a obtenção dos 33 (ou 32) *dataframes* de expressão/metilação pré-processados segundo as Subseções 3.2.1, 3.2.2 e 3.2.2, os dados foram concatenados em 3 objetos chamados DG (*Dataframe Geral*), um para cada tipo de dado. Essa etapa de concatenação de dados é ilustrada na Figura 3.4 na Seção 3.4.

No caso do mapa modular publicado por *Eran Segal*, os dados, caso e controle, são agrupados em uma matriz de sondas (linhas) por *Arrays* (colunas) e normalizados pela média de expressão na matriz, equivalente a realizar uma centralização da média dos valores da mesma. Dessa maneira, os dados de todos os *Arrays* na matriz são transformados para uma escala comum (SEGAL et al., 2004).

Na metodologia de *Segal*, a diferença da alteração de uma sonda/gene entre os *Arrays* (caso/controle) é considerada relevante se o *Log Fold Change*, cuja fórmula é definida abaixo:

$$log(\frac{caso}{controle}) \tag{3.2}$$

for maior que 2 (induzidos) ou menor que -2 (reprimidos) em relação a média dos *Arrays* e possuir um q-valor menor que 0.05, sendo que primeiramente o p-valor é calculado com um teste hipergeométrico e posteriormente corrigido por um FDR (*False Discovery Rate*) de 0.05 feito pelo teste de Benjamini-Hochberg (SEGAL et al., 2004).

No caso do mapa modular desenvolvido neste trabalho, foi encontrado um problema em seguir tal procedimento. No TCGA, os casos e os controles não estão devidamente pareados. Tal situação pode introduzir um grande viés estatístico em função do problema envolvendo número de casos e número de observações no conjunto de dados. Desta forma, foi determinado que seria utilizada uma outra abordagem para seleção de elementos que podem ser considerados como significantemente alterados, que neste trabalho serão referidos como "Sondas Induzidas ou Reprimidas" (SIRs). Para isso, foi criada uma matriz contendo os

dados brutos de cada sonda/gene para cada *Array*, a média da matriz foi centralizada com auxílio da função *scale* (linguagem R), de modo a reproduzir o procedimento realizado por *Segal* para colocar os valores de todos os *Arrays* em uma mesma escala. Tais valores recebem o nome de "Expressão Relativa" (ER) e serão de suma importância para o processo de clusterização e de análise. É importante salientar que, para fim de simplificação, a definição de "Expressão Relativa", se aplica aos 3 tipos de dados: RNA-seq, metilação e miRseq. Naturalmente, no caso dos dados de metilação, o que foi obtido é um valor de "Metilação Relativa" , pois os dados medem níveis de metilação, ao invés de "Expressão Relativa" como no caso do RNA-seq e do miRseq, cujos dados medem níveis de transcrição.

Foram consideradas como significativamente alterada, qualquer sonda com valor de "Expressão Relativa" maior ou igual a 3 vezes ao desvio padrão do *Array* em questão. Essas mesmas condições foram utilizadas para seleção de elementos significantemente alterados nas ilhas CpG e de miRNAs (SEGAL et al., 2004).

É muito importante salientar que isto não é a mesma coisa que uma análise de expressão gênica diferencial, onde existe caso e controle pareados. O que é sugerido é uma sumarização estatística da expressão das sondas, que só pode ser realizada em função do grande número de amostras (10000) disponíveis. Assim, nosso método viabiliza um estudo populacional do câncer e a modularidade que pretende-se determinar foi encontrada tendo como base as sondas cuja expressão esteja consistentemente induzida ou reprimida no conjunto de dados. A etapa de seleção de SIRs aqui descrita está ilustrada na Figura 3.5 na Seção 3.4.

3.2.7 Matriz de Porcentagem de Expressão: Determinação dos *arrays* em que a Alteração dos Elementos é Significante

O próximo passo foi a construção de uma matriz de *Arrays* por *Sets* (*Gene Set*, *CpG Set* e *miRNA Set*) em cada tipo de mapa modular, com a finalidade de possibilitar o cálculo de um p-valor através do teste exato de Fisher (equivalente a distribuição hipergeométrica) para identificação dos *Arrays* estatisticamente significantes no que se diz respeito a alteração (indução ou inibição) dos elementos que compõe o *Arrays* e o *Set* de determinado mapa modular (genes, CpGs e miRNAs). Essa matriz é construída a partir de frações de elementos alterados presentes e ausentes nos *Arrays* e no respectivo *Set* do mapa modular em questão. Após a montagem da matriz, os p-valores são determinados através da aplicação da fórmula do teste hipergeométrico. Foram considerados como significativos os pares *Arrays-Set* que

apresentarem p-valor menor que 0.05 e maior que 0 (SEGAL et al., 2004). Essa matriz de p-valores é então convertida para para uma matriz de porcentagem de expressão. Se o p-valor fosse nulo ou maior que 0.05, o valor de porcentagem de expressão assinalado é 0. Caso o p-valor fosse maior que 0 e menor que 0.05, foi calculado um valor de porcentagem de expressão (PE). Este valor corresponde à porcentagem de SIRs dentro do *Set* do par *Array-Set* em questão, portanto, o *Set* na verdade corresponde a um processo biológico (BP). Essa porcentagem é positiva para SIRs induzidas e negativa para SIRs reprimidas. Se um par *Array-Set* apresentar SIRs induzidas e reprimidas, a PE de maior valor em módulo é escolhida. Se as duas PEs forem iguais, é assinalado o valor para o par *Array-Set*. Assim, uma matriz de p-valores foi convertida para uma matriz de porcentagem de expressão, tal objeto também é chamado de DG (*Dataframe* Geral) de porcentagem de expressão. Esta etapa é ilustrada na Figura 3.6 na Seção 3.4.

Após essa etapa, os 3 DGs de porcentagem de expressão passam por algumas medidas de filtragem adicionais foram implementadas a fim de reduzir redundância: remover variáveis não informativas e diminuir o tempo de clusterização dos objetos criados. Primeiramente, foi calculado o número de valores não nulos para cada linha e coluna e então foi determinado a mediana de cada grupo (linhas e colunas). Em seguida, a remoção de todas as linhas e colunas nulas ou menores que a mediana calculada foi realizada. Assim, foram removidas variáveis com pouco ou nenhum valor informativo e o tempo de execução do *script* de clusterização foi reduzido. Ao término do processo, os 3 objetos tinham as seguintes dimensões:

- RNA-seq: 4483 amostras (colunas) e 1680 GOs (linhas);
- metilação: 4326 amostras (colunas) e 2263 GOs (linhas);
- miRseq: 4600 amostras (colunas) e 213 GOs (linhas).

3.2.8 A Heurística do Novo Mapa Modular: Identificação de *Clusters* de *Sets*

O processo de clusterização dos 3 objetos, foi realizado com auxílio da função "*pvclust*" do pacote homônimo em ambiente da linguagem R. A partir da matriz *Sets vs Arrays*, foi realizada a clusterização hierárquica dos *Arrays-Set*. Os *clusters*, na metodologia publicada

por *Eran Segal* em 2004, são formados de hierarquia menor para maior, utilizando a média como método de aglomeração e a correlação como a distância euclidiana para a construção da árvore de interações entre *clusters* a partir de uma distribuição de amostras. Cada arvore possui folhas, o menor nível hierárquico, de um único tipo de *Array-Set*, e cada folha está associada a um vetor que possui tamanho igual ao de *Arrays*. Os índices desse tal vetor são nulos nos *Arrays* em que o *Set* correspondente não apresentou alteração significativa, e no caso contrário esse índice é preenchido com a fração de elementos alterados do *Set* presentes no *Array*. Assim as folhas vão sendo agrupadas, em nodos interiores de hierarquia cada vez maior, os nodos também estão associados a vetores que são iguais a média dos vetores dos nodos filhos que compõe os nodos em questão (SHIMODAIRA, 2013).

Para a determinação da significância estatística dos *clusters*, foi utilizado o método *mutiscale bootstrap resampling*. O método é uma variação do *bootstrap resampling*, onde p-valores chamados *bootstrap probability* são assinados a um determinado *cluster* de acordo com uma razão de dendrogramas que apresentam a hipótese (o *cluster* em questão) pelo número total de dendrogramas gerados a partir de distribuições aleatórias (com reposição dos dados iniciais, essas são chamadas de *bootstrap samples*) derivadas dos dados iniciais (SHIMODAIRA, 2013). No caso do *bootstrap resampling*, são geradas *bootstrap samples* com o mesmo tamanho do conjunto original, por exemplo, 10 casos com 5 replicatas geraria distribuições com 10 linhas e 5 colunas, no fim apenas um p-valor (*bootstrap probability*) seria determinado. O *multiscale bootstrap resampling* se consiste em gerar diversos p-valores (*bootstrap probability*) para um conjunto de dados e depois ajustar esses p-valores BP a uma equação teórica, fornecendo então o p-valor AU (*aproximately unbiased*). Segundo os desenvolvedores do método, esse p-valor AU está corrigido para o viés introduzido pelo uso de distribuições aleatórias de um único tamanho, como é feito no método *bootstrap resampling*. No *multiscale bootstrap resampling*, os x p-valores BP utilizados para calcular o p-valor AU são gerados a partir de x conjuntos de distribuições cada uma contendo y distribuições aleatórias. Sendo que cada um dos x possui distribuições aleatórias com dimensões fixas, esse número de dimensões pode ser menor, igual ou maior aos do conjunto original (SHIMODAIRA, 2013). Foram considerados significantes os *clusters* que apresentaram um p-valor AU menor que 0.05 (ou confiança de 0.95), foram utilizados 5000 passos de *bootstrap*. Essa etapa de clusterização hierárquica está ilustrada na Figura 3.7 na Seção 3.4.

3.2.9 Definição dos Módulos

A definição dos módulos é feita a partir da identificação do processo biológico mais significante do módulo. Talvez essa não seja a melhor maneira de nomear ou associar um módulo a um processo biológico. Entretanto como este se trata de um problema em aberto e uma decisão devia ser tomada, foi determinado o uso desse método devido a sua simplicidade. Para realizar este procedimento, foi criada uma função que nomeava cada módulo que considerava os seguintes itens:

1. Porcentagem de sondas (genes, CPGs e miRNAs) induzidas e inibidas (SIRs) da ontologia;
2. Frequência média de sondas induzidas ou reprimidas da ontologia (*Score*);
3. Nível de profundidade da ontologia (*Grand_Score*).

O item "1" é evidente, corresponde a porcentagem das sondas de uma determinada ontologia que em algum momento são classificados como induzidas ou reprimidas no conjunto de dados total, incluindo todos os tumores. O item "2" , chamado de *Score*, corresponde a frequência média de sondas induzidas ou reprimidas da ontologia e é calculada como a soma da frequência de todos os genes diferencialmente expressos internos da ontologia no conjunto de dados inteiro dividida pelo número de SIRs únicos e diferentes existentes na ontologia. O terceiro item, "3", chamado de *Grand_score*, corresponde ao item "2" multiplicado pelo nível de profundidade da ontologia.

Os itens "1" e "2" foram criados para mensurar super-representação de SIRs dentro de uma ontologia segundo o conjunto de dados como um todo, de maneira que é possível quantificar e selecionar ontologias cujas SIRs tenham um nível mínimo de representatividade no conjunto de dados. O item "3" foi criado para considerar a especificidade da ontologia, utilizando sua profundidade na arvore de ontologias. Quanto mais profunda uma ontologia na hierarquia da arvore, mais especifica esta é. É natural imaginar que o método deva considerar um certo equilíbrio entre a especificidade da ontologia e a representatividade de suas SIRs no conjunto de dados, essa é a razão pela qual o item "3" foi criado, ja que este é a maneira mais eficiente de realizar a seleção da ontologia "*ideal*" dentro de um módulo formado.

Para nosso método foram considerados os itens da seguinte maneira:

- Qualquer porcentagem de SIRs maior que 0;
- Qualquer *Score* maior que 0;
- Qualquer profundidade maior que 2;
- Grand_score máximo.

Isso resume a maneira pela qual o processo biológico ideal é selecionado para cada módulo determinado. Esta etapa de seleção e de nomeação de módulos é ilustrada na Figura 3.7 na seção 3.4.

3.2.10 Finalização do Mapa Modular

A matriz obtida no item anterior juntamente com os q-valores determinados para suas células são utilizados para construção de um *heatmap* associando módulos e condições clínicas, o eixo paralelo destes dois contém gráficos para visualização da quantidade de elementos por módulos e *Arrays* por condições, respectivamente. Este construto gráfico representa a forma final do mapa modular. No final, foram obtidos 3 desses mapas modulares, um para os dados do RNA-seq, um para dados de metilação e um para miRNAs (SEGAL et al., 2004).

3.3 Análise Final dos Mapas Modulares

3.3.1 Avaliação da Clusterização Hierárquica

A primeira parte da análise final se consiste em avaliar a consistência dos resultados obtidos através da clusterização hierárquica. Sendo essa uma metodologia heurística, com algum grau de aleatoriedade envolvida, é natural que haja alguma medida de convergência de nossos resultados. Resultados sem convergência podem ser considerados aleatórios e portanto, de pouco interesse para a análise final. Maiores detalhes sobre essa avaliação são fornecidos na Seção 4.1.

3.3.2 Visualização dos Mapas Modulares

A segunda parte da análise final dos mapas modulares constitui uma avaliação visual, tendo como finalidade principal a determinação de módulos que se mostrem como positiva ou negativamente regulados em diversos tipos de tumores, que a princípio constitui o objetivo principal da tese. A identificação desses módulos consistentemente induzidos ou reprimidos se baseia na simples identificação visual de quais módulos apresentam majoritariamente um mesmo perfil de expressão em vários tumores. Foram determinados 3 perfis de expressão: positivo, negativo e nulo. O positivo é representado por valores entre [0,1] e visualmente representado pela cor vermelha, o negativo por valores entre [-1,0] e a cor verde, por fim o nulo pelo valor 0 e cor preta. Mais detalhes sobre essa etapa de análise são encontrados na Seção 4.1 a qual está subdividida em subseções para cada um dos 3 mapas modulares: RNA-seq 4.1.1, metilação 4.1.2 e miRseq 4.1.3.

3.3.3 Estudo de Caso com Auxílio de Redes

A última parte da análise final se consiste em realizar um estudo de caso de um dos tumores com auxílio de redes, para mostrar que os dados produzidos pela construção do mapa modular ainda podem ser reaproveitados para análises adicionais. Essa análise adicional tem o potencial de gerar mais *insights*, principalmente, no que tange à observação do perfil de expressão de certos alvos.

A primeira parte desse estudo de caso é determinar as sondas/SIRs de maior importância estatística segundo os dados do mapa modular, os nossos alvos. Para isso, foi utilizado um teste de reamostragem com reposição de 10000 passos para determinação de p-valores para as SIRs de cada tipo de mapa modular. Informações mais detalhadas são fornecidas na própria seção de análise, na Seção 4.2. Os alvos selecionados por esse teste são utilizados para montagem de uma rede interação. A natureza da rede depende do tipo de dado: proteína-proteína para os mapas modulares do RNA-seq/metilação e miRNA-proteína no caso do mapa modular do miRseq. Na subseção abaixo é descrito o protocolo para montagem das redes e as bases de dados utilizadas.

3.3.4 Construção de Redes de Integração para Análise de Expressão

Redes são a representação de elementos e suas respectivas interações entre sí, a natureza dos elementos (também chamados de nodos ou vértices) e das interações (também chamadas de arestas) são variáveis de rede para rede. Essas entidades são tipicamente representadas por matrizes de adjacência, que são matrizes binárias onde as linhas e colunas representam os elementos e os valores 1 e 0 são utilizados para representar a presença ou ausência, respectivamente, de interação entre os elementos. Redes biológicas têm sido cada vez mais utilizadas em análises de dados massivos produzidos pelos experimentos atuais e na visualização dos mesmos. Um exemplo de rede biológica são as que associam genes a seus produtos e elementos reguladores. Outro exemplo são redes de interação proteína-proteína (PPI), onde seus nodos são as proteínas e as arestas representam interações entre essas proteínas.

Para facilitar a interpretação dos mecanismos biológicos e visualização do perfil de expressão dos alvos durante o estudo de caso, foram utilizadas redes de interação. Para o mapa modular do RNA-seq foi construída uma rede PPI, as relações foram obtidas da base de dados *String*, discutida brevemente na Subseção 3.3.4.1. Para o mapa modular da metilação, também foi utilizada uma rede PPI, mas para construí-la foi necessário primeiro termos em mãos uma relação entre ilhas CPG e mRNAs. Essas relações podem ser retiradas de bases de dados da plataforma Illumina. Esse procedimento é descrito detalhadamente na Seção 3.2.5.

Por fim, existe o mapa modular do miRseq, no qual foi utilizado redes do tipo miRNA-mRNA. O *String* não trabalha com esse tipo de relação. Assim sendo, foi necessário utilizar a base de dados *Starbase ENCORI* para derivar as relações miRNA-mRNA, este protocolo é descrito detalhadamente na Subseção 3.3.4.2.

3.3.4.1 STRING: Um Banco de Dados de Interações Proteína-Proteína (PPI)

O STRING (*Search Tool for the Retrieval of Interacting Genes/Proteins*) é um banco de dados *online* de interações proteína-proteína. Esse banco de dados é mantido pelo Laboratório Europeu de Biologia Molecular (*European Laboratory for Molecular Biology* - EMBL), desde o ano 2000 (MERING et al., 2005; MERING et al., 2007; JENSEN et al., 2009; SZKLARCZYK et al., 2011; SZKLARCZYK et al., 2014) e pode ser encontrado no endereço

eletrônico "`https://string-db.org/`".

Hoje, o STRING encontra-se na versão 11.5 com dados referentes a 2031 espécies, aproximadamente 67 milhões de polipeptídeos registrados e mais de 20 bilhões de interações, essas informações são constantemente atualizadas no endereço eletrônico mencionado. A natureza das interações registradas varia. Atualmente, existem 7 tipos de interações com a qual esse banco de dados trabalha: *Conserved Neighborhood*, *Co-ocurrence*, *Fusion*, *Co-expression*, *Experiments*, *Databases* e *Textmining*.

As interações do tipo *Conserved Neighborhood* associam proteínas cujos genes codificantes (no genoma procarioto) estejam próximos um do outro (distância intergênica de no máximo 300 pares de bases). As interações do tipo *Co-ocurrence* servem para indicar proteínas cujos genes codificantes possuam homólogos em outra espécies. As interações do tipo *fusion* indicam proteínas cujos genes codificantes sofram o processo de fusão. As interações do tipo *Co-expression* indica proteínas cujos genes codificantes apresentam co-expressão. As interações do tipo *Experiments* são interações proteína-proteína retiradas de outros bancos de dados de interações entre proteínas (DIP,BIND,BioGRID). As interações do tipo *Databases* representam interações de proteínas considerando rotas metabólicas, são retiradas de bancos de dados como KEGG e Reactome. As interações do tipo *textmining* representam interações proteína-proteína encontradas nos resumos de artigos científicos (MERING et al., 2005; MERING et al., 2007; JENSEN et al., 2009; SZKLARCZYK et al., 2011; SZKLARCZYK et al., 2014).

Neste trabalho, as redes de interação proteína-proteína para *Homo sapiens* foram obtidas da base dados STRING, considerando apenas os tipos de interação *Experiments* e *Databases*. Essas interações são retiradas de bancos de dados curados manualmente, e portanto, possuem um certo grau de confiança. O score mínimo para que as interações sejam aceitas é de 0.7, o uso destes parâmetros restringe ao máximo o número de interações proteína-proteína que sejam falsos positivos. Essas redes foram utilizadas para visualização e interpretação de mecanismos biológicos associados aos módulos e condições que nos forem de interesse nos mapas modulares.

3.3.4.2 Starbase ENCORI: Um Banco de Dados de Interações miRNA-Alvo

O *Starbase ENCORI* é um endereço eletrônico encontrado em "`https://starbase.sysu.edu.cn/`" que se dedica a armazenar interações que miRNAs possam ter com outros transcritos. Isso é feito por meio da coleta de informações de diversas fontes que incluem: diferentes algoritmos de predição, dados de CLIP-seq e artigos publicados (LI et al., 2014). Existem diversas relações que podem ser encontradas com auxílio dessa base de dados, alguns exemplos incluem: miRNA-mRNA, miRNA-lncRNA (*long non coding RNA*), miRNA-circRNA (*circular RNA*), miRNA-ceRNA (*competing endogenous RNA*) e miRNA-pseudogene (LI et al., 2014). *Li* e colaboradores (2014), detalham em (LI et al., 2014) como funciona a base de dados, bem como a mesma foi criada. Essa ferramenta foi utilizada para prospectar relações do tipo miRNA-mRNA. Para entender o protocolo de seleção que foi utilizado, primeiro é explicado o formato dos resultados fornecidos pelo *ENCORI*. As relações são fornecidas em formato de tabela, onde são encontradas as seguintes colunas de interesse:

- *miRNAname*: o nome do miRNA;
- *geneID*: o nome ENSEMBL do mRNA alvo;
- *geneName*: o nome *Gene Symbol* do mRNA alvo;
- *PITA*: algoritmo de predição de interação miRNA-alvo por *kertesz* e colaboradores (KERTESZ et al., 2007);
- *RNA22*: algoritmo de predição de interações miRNA-alvo por *Loher* e colaboradores (LOHER; RIGOUTSOS, 2012);
- *miRmap*: algoritmo de predição de interações miRNA-alvo por *Vejnar* e colaboradores (VEJNAR; ZDOBNOV, 2012);
- *microT*: algoritmo de predição de interações miRNA-alvo por *Maragkakis* e colaboradores (MARAGKAKIS et al., 2009);
- *miRanda*: algoritmo de predição de interações miRNA-alvo por *Enright* e colaboradores (ENRIGHT et al., 2003);

- *PicTar*: algoritmo de predição de interações miRNA-alvo por *Krek* e colaboradores (KREK et al., 2005);
- *TargetScan*: algoritmo de predição de interações miRNA-alvo por *Agarwal* e colaboradores (AGARWAL et al., 2015);
- *clipExpNum*: número de experimentos CLIP-seq com evidencias da interação miRNA-alvo de acordo com *Li* e colaboradores (LI et al., 2014);
- *degraExpNum*: número de experimentos de sequenciamento de degradoma com evidências da interação miRNA-alvo, segundo (LI et al., 2014);
- *pancancer*: número de tumores cuja interação miRNA-alvo possui evidências segundo *Li* e colaboradores (LI et al., 2014).

As colunas "*miRNAname*", "*geneID*" e "*geneName*" representam identificadores para o miRNA e seu alvo. As colunas "*PITA*", "*RNA22*", "*miRmap*", "*microT*", "*miRanda*", "*PicTar*" e "*TargetScan*" representam seus respectivos algoritmos de predição. Sendo assim, estas colunas indicam se a relação miRNA-alvo possui alguma evidência, segundo o respectivo algoritmo de predição. As colunas "*clipExpNum*", "*degraExpNum*" e "*pancancer*" indicam o número de evidências que a relação miRNA-alvo tem de acordo com cada tipo de dado, de acordo com o que é descrito por *Li* e colaboradores em (LI et al., 2014).

Para realizar a seleção das relações miRNA-mRNA, foi seguido o seguinte protocolo: as interações devem estar presentes em pelo menos 5 dos 7 algoritmos considerados, devem apresentar pelo menos 5 experimentos CLIP-seq e devem ter evidências em pelo menos 3 tumores. Essas relações foram utilizadas para montar uma rede miRNA-mRNA para visualização do perfil de expressão dos alvos/nodos e de possíveis interações de interesse. Para realizar uma análise mais direcionada segundo os dados obtidos posteriormente, as relações miRNA-mRNA foram filtradas segundo os dados de RNA-seq ou metilação, selecionando apenas as relações miRNA-mRNA, cujo mRNA esteja presente em sua respectiva rede (RNA-seq ou metilação). Posteriormente, as relações miRNA-mRNA foram concatenadas às relações mRNA-mRNA da respectiva rede (RNA-seq ou metilação), cujo mRNA esteja presente nas relações miRNA-mRNA filtradas. No final, foi obtida uma rede com relações miRNA-mRNA e mRNA-mRNA baseada em dados obtidos posteriormente, o que nos facilitou uma análise direcionada.

Informações mais detalhadas sobre a construção da rede são fornecidas na própria seção de análise de redes na Subseção 4.2.3.

3.4 Sumarização da Metodologia: Fluxograma de Dados

Nesta Seção, ilustraremos, na forma de fluxogramas, as etapas de coleta e transformação de dados descritas nas Seções 3.2 e 3.3. Abaixo, na Figura 3.2, temos a primeira fase da metodologia, composta de prospecção dos dados de RNA-seq, metilação e de miRseq. A legenda da Figura 3.2 trás mais informações sobre o processo:

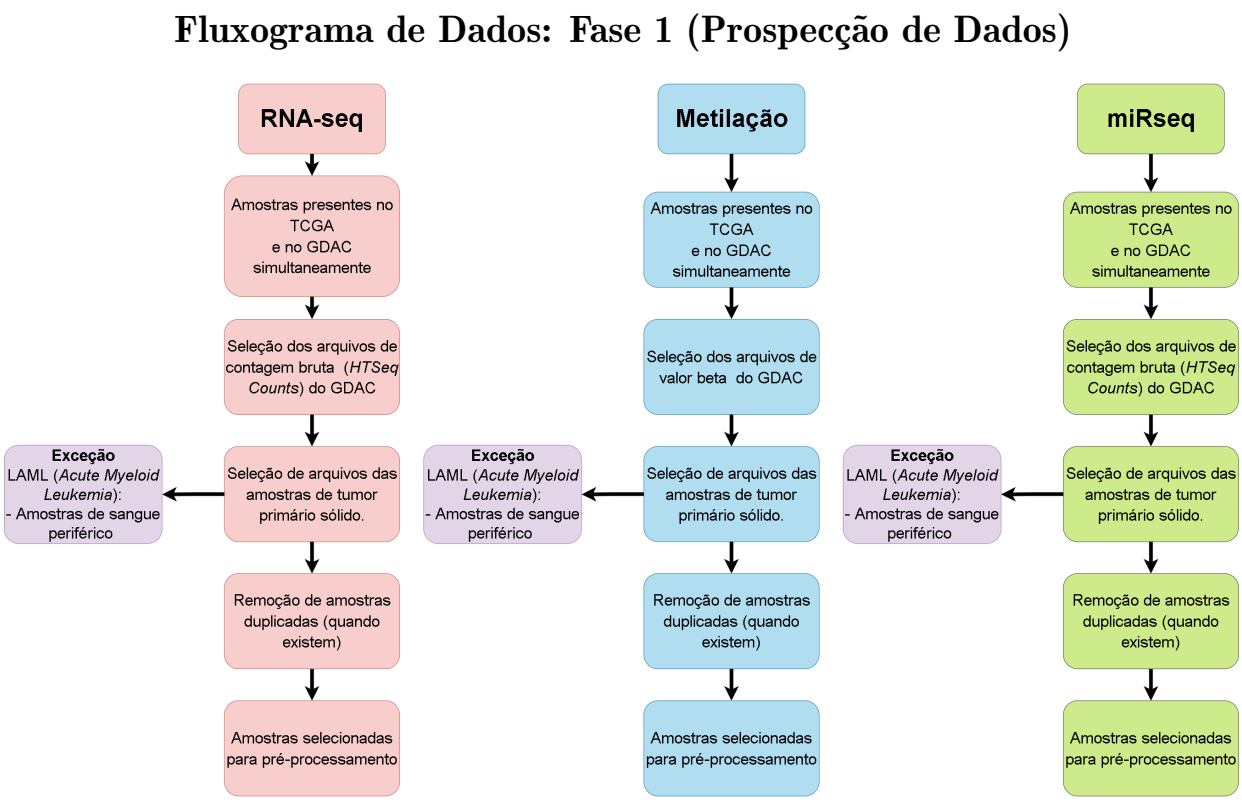


Figura 3.2: Fluxograma de Dados: Fase 1 (Prospecção de Dados). Os fluxogramas de cada um dos 3 tipos de dados são representados com cores diferentes: RNA-seq (rosa), metilação (azul) e miRseq (verde). Cada um dos fluxogramas estão organizados numa hierarquia descendente, sendo assim, a ordem flui de cima para baixo. Na primeira fase, para cada um dos 3 tipos de dados, são prospectadas as amostras existentes no TCGA (*The Cancer Genome Atlas*) e no GDAC (Genomic Data Commons). Em seguida, as amostras que existiam em ambos bancos de dados, foram prospectadas do GDAC. Este banco de dados, realiza diversas filtragens que são descritas na Seção 3.2. É importante relembrar aqui, que o GDAC tem um protocolo para sumarizar as replicatas de expressão ou metilação em uma única amostra. Sendo assim, depois que as amostras do GDAC são obtidas, realizamos a remoção de possíveis amostras duplicadas. Assim, é finalizada a obtenção dos dados brutos iniciais.

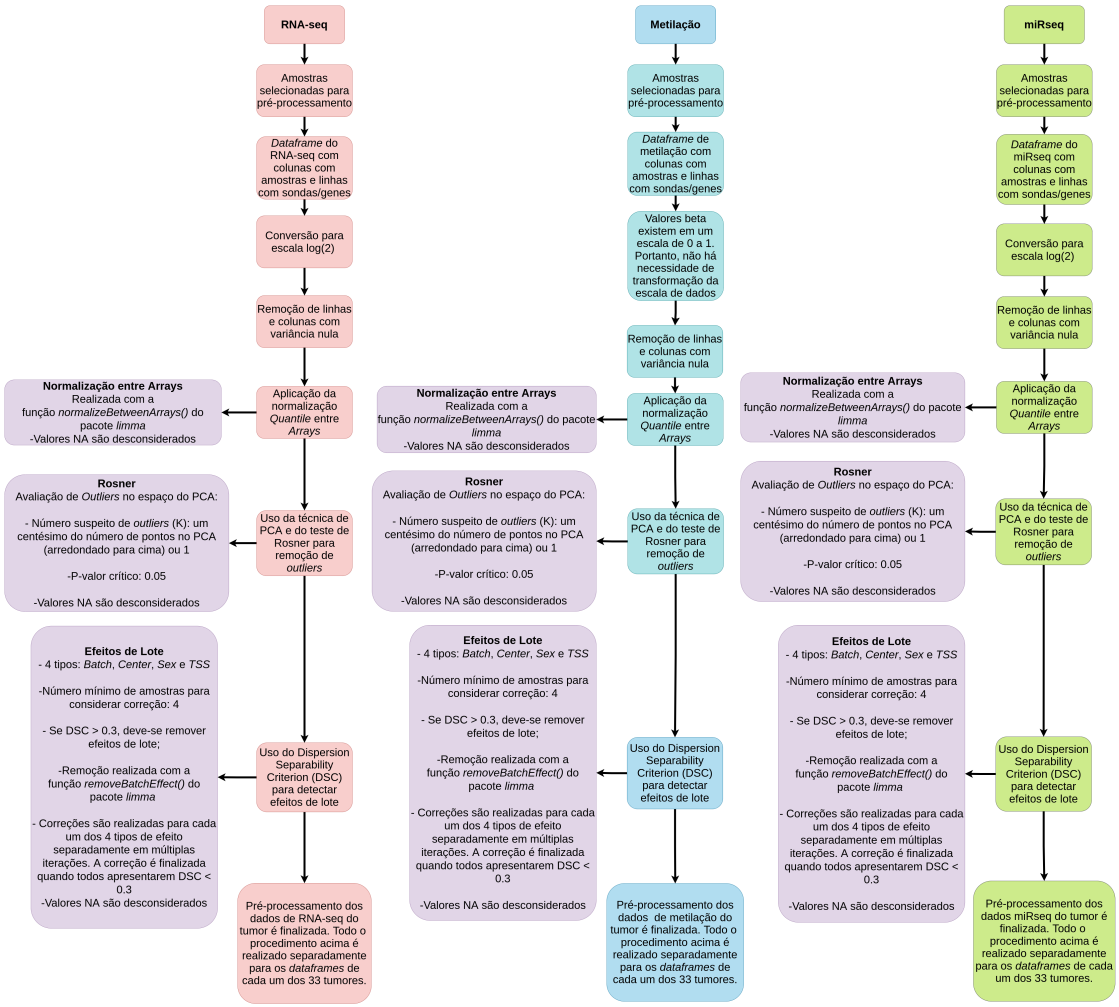


Figura 3.3: Fluxograma de Dados: Fase 2 (Pré-processamento). Os fluxogramas de cada um dos 3 tipos de dados são representados com cores diferentes: RNA-seq (rosa), metilação (azul) e miRseq (verde). Tais fluxogramas estão organizados em uma hierarquia descendente, sendo assim, a ordem flui de cima para baixo. Note que alguns passos dos fluxogramas possuem uma seta fluindo para um retângulo roxo a sua esquerda, tais retângulos contêm informações detalhadas a respeito de seu passo respectivo. Cada um dos fluxogramas são iniciados a partir dos dados prospectados na Figura 3.2, os quais são organizados em *dataframes* tumor-a-tumor, compondo 33 *dataframes* de dados brutos para cada um dos 3 tipos de dados, totalizando 99 *dataframes* para pré-processamento. Esses 99 *dataframes* passarão pela longa fase 2 de pré-processamento descrita pelo seu respectivo fluxograma. Primeiramente, os dados brutos de cada *dataframe* são transformados para uma escala log(2), no caso dos dados de RNA-seq e miRseq. Essa transformação não é necessária para os dados de metilação, pois os valores beta existem em uma escala de 0 a 1. Posteriormente, é realizada uma normalização entre *Arrays* do tipo *Quantile* para os valores de cada sonda/gene. Essa normalização é aplicada separadamente em cada um dos 99 *dataframes*, com auxílio da função *normalizeBetweenArrays()* do pacote *limma*. Os valores NA são desconsiderados no cálculo da normalização, tais valores serão tratados mais a frente. Em seguida, os dados dos 99 *dataframes* são transformados para o espaço do PCA (*Principal Component Analysis*), considerando as componentes principais PC1 e PC2. Com auxílio do teste de Rosner é determinado se existem *outliers* dentro desse espaço, cujos pontos/valores correspondem às amostras de um determinado tumor. O teste de Rosner precisa de dois parâmetros, o número estimado de *outliers* "K" e um p-valor crítico (α). Empiricamente foi determinado que "K" seria igual ao valor de um centésimo do número de amostras no espaço do PCA, arredondado para cima, ou 1. Quanto ao valor "α", foi determinado, também empiricamente, que esse seria igual a 0.05. Após a remoção das amostras consideradas *outliers*, cada uma das 99 *dataframes* passa pela identificação e correção de efeitos de lote. São 4 tipos de efeitos de lote considerados: "*Batch*", "*Center*", "*Sex*" e "*TSS*". Para cada uma das 99 *dataframes*, as amostras são separadas de acordo com sua classificação no que tange seu lote ("*batch*"), o centro de processamento ("*Center*"), sexo do indivíduo ("*Sex*") e local de origem ("*TSS*"). De maneira iterativa é avaliada se a classificação das amostras, de acordo com os 4 tipos mencionados, é responsável pela indução de algum efeito de lote nos dados. Essa avaliação é feita com o "*Dispersion Separability Criterion*" (DSC), definido na Subseção 3.2.1. Quando o DSC de uma das 99 *dataframes*, cujas amostras estão classificadas de acordo com um dos 4 tipos de efeitos, for maior que 0.3, a função "*removeBatchEffect()*" do pacote "*limma*" é utilizada para remoção dos efeitos de lote. Posteriormente o DSC é medido novamente para garantir que este esteja menor que 0.3, pois, isto foi determinado como condição para que o efeito de lote seja considerado desprezível. O processo de medição do DSC e correção de efeitos de lote só acorre quando classificação em questão apresenta mais de 4 amostras. Por exemplo, considerando o efeito "*batch*". Em um determinado tumor existem 500 amostras divididas entre 10 tipos de lotes. Se um destes tipos for constituído por 2 amostras, tal tipo, e portanto, as amostras nele contidas, são excluídas da análise de efeitos de lote. O processo de medição do DSC e de correção dos efeitos de lote foi realizado de maneira iterativa e separadamente para cada um dos 4 tipos de efeitos de lote em cada uma das 99 *dataframes* até garantir que o DSC de cada uma das 4 classificações estivesse abaixo de 0.3, garantindo que os 4 tipos de efeitos de lote são desprezíveis em cada uma das 99 *dataframes*. Ao final temos 98 *dataframes* devidamente pré-processadas: 33 tumores (RNA-seq), 33 tumores (metilação) e 32 tumores (miRseq). Um tumor, o glioblastoma (GBM), foi excluído por não ter amostras suficientes ao fim do pré-processamento.

A Figura 3.3, contém 3 fluxogramas, ilustrando a longa fase de pré-processamento dos dados brutos. Cada uma dos 3 tipos de dados têm seu próprio fluxograma, representado por uma cor: RNA-seq em rosa, metilação em azul e miRseq em verde. Note que alguns passos tem uma seta fluindo para sua esquerda em direção a um retângulo roxo, este contém informações específicas acerca do passo em questão. Em resumo, as etapas da fase 2 são: transformação de escala, remoção de linhas e colunas com variância nula, normalização entre *Arrays*, detecção e remoção de *outliers* e por fim, a detecção e correção de efeitos de lote. Cada uma dessas etapas são mencionadas na Subseção 3.2.1 e estão detalhadas na Figura 3.3 e na sua respectiva legenda. Lembrando que todas essas etapas são aplicadas separadamente aos 99 *dataframes*. No fim, foram obtidas 98 *dataframes*, pois o tumor glioblastoma não apresentou amostras para os dados de miRseq. A terceira fase é composta pela concatenação dos 33 (ou 32) *dataframes* pré-processados em um único objeto, aqui chamado de DG (*Dataframe* Geral). Este contém os todas as amostras de todos os tumores de um tipo de dado, portanto, 3 DGs foram preparados. A Figura 3.4 abaixo, apresenta os fluxogramas das etapas desta fase.

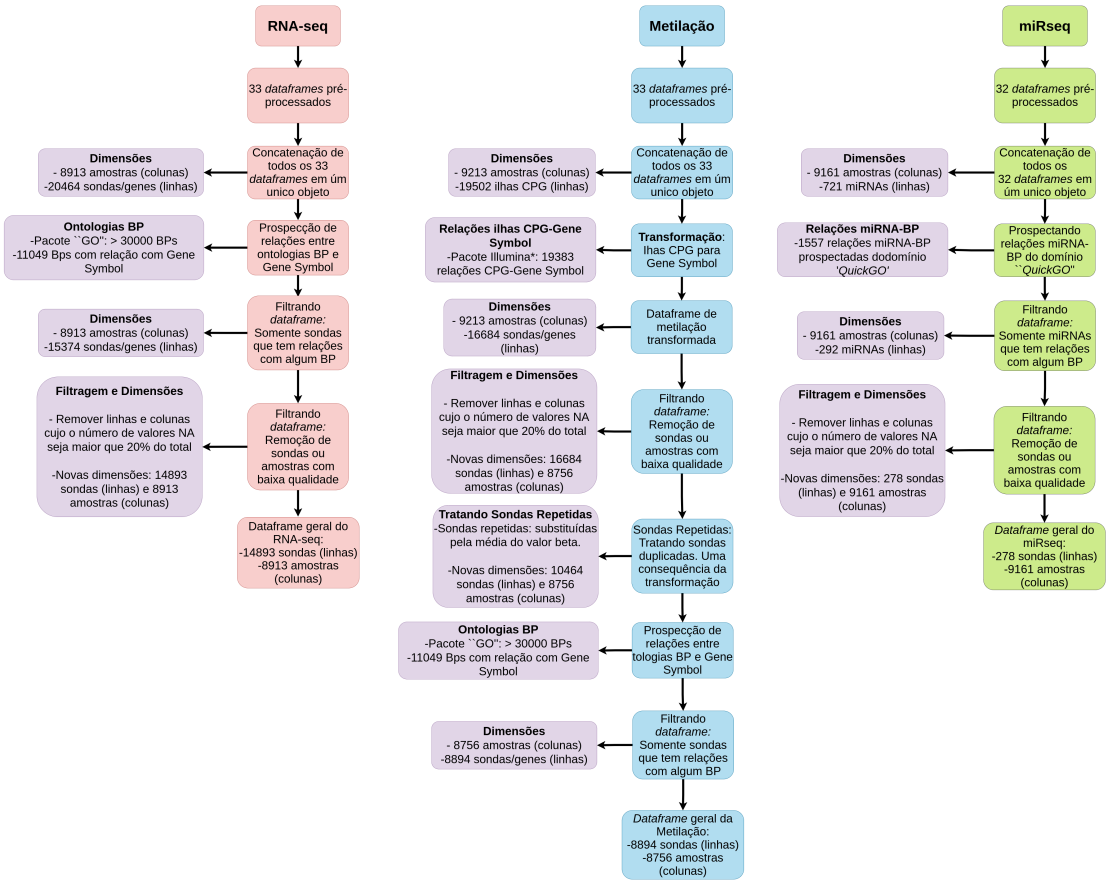


Figura 3.4: Fluxograma de Dados: Fase 3 (concatenação de Dados). Os fluxogramas de cada um dos 3 tipos de dados são representados com cores diferentes: RNA-seq (rosa), metilação (azul) e miRseq (verde). Tais fluxogramas estão organizados em uma hierarquia descendente, sendo assim, a ordem flui de cima para baixo. Note que alguns passos dos fluxogramas possuem uma seta fluindo para um retângulo roxo a sua esquerda, tais retângulos contêm informações detalhadas a respeito de seu passo respectivo. Cada um dos fluxogramas são iniciados a partir dos 33 *dataframes* pré-processados segundo a Figura 3.3. O objetivo desta fase foi a concatenação dos 33 (ou 32) *dataframes* em um único objeto. A primeira etapa é justamente a concatenação, que cria 3 grandes objetos contendo todos os dados de expressão/metilação pré-processados de cada um dos 3 tipos de dados, aqui chamados de "*dataframes* gerais" (DG). O DG do RNA-seq, inicialmente, tinha 8913 amostras e 20464 sondas. O DG de metilação tinha 9213 amostras e 19512 sondas. O DG de miRseq tinha 9161 amostras e 721 sondas. Em seguida é necessário filtrar os DGs, para que estes contenham apena sondas (mRNAs, CPG ou miRNA) que tenham relações com processos biológicos conhecidas. Para os dados de RNA-seq, bastou prospectar as relações Gene Symbol-BP do pacote "GO". Um total de 11049 relações Gene Symbol-BP foram prospectadas e a partir delas, foram removidos as sondas que não possuiam nenhuma relação Gene Symbol-BP conhecida do DG de RNA-seq, criando um objeto com 8913 amostras e 15374 sondas. O DG de miRseq passou por um processo quase idêntico, mas foi necessário realizar primeiro, a prospecção de 1557 relações miRNA-BP do domínio "*QuickGO*". Em seguida forma filtradas do DG de miRseq, as sondas sem nenhuma relação miRNA-BP conhecida, criando um objeto de 9161 amostras e 292 sondas. A metilação teve um processo pouco diferente, pois não foram encontradas relações do tipo CPG-BP. Portanto, foi necessário prospectar 19383 relações CPG-Gene Symbol do pacote Illumina450kkanno e realizar um transformação na DG de metilação onde as sondas (CPGs) foram substituídas pelo seu respectivo Gene Symbol, criando um objeto de 9213 amostras e 16684 sondas. Posteriormente foram removidas amostras e colunas de baixa qualidade, Por "baixa qualidade", se entende que o vetor em questão possuia 20% ou mais de seus elementos como nulos (NA). Os valores NA que ainda existiam nos 3 DGs, mesmo após a filtragens anterior, foram substituídos pela mediana da sonda a que pertenciam. Após essa etapa os DGs tinham as seguintes dimensões: 8913 amostras e 14893 sondas (RNA-seq), 8756 amostras e 16684 sondas (metilação) e 9161 amostras e 278 sondas (miRseq). Os DGs de RNA-seq e de miRseq foram finalizados na etapa anterior, mas o DG de metilação apresentou um problema, sondas repetidas, uma consequência da etapa de transformação. Para corrigir o problema, os vetores das sondas repetidas foram substituídos por um único vetor de médias de metilação. Essa etapa resultou em um DG de metilação com 8756 amostras e 10464 sondas. Posteriormente, foram removidas as sondas sem relação Gene Symbol-BP conhecidas do DG de metilação, criando um objeto de 8756 amostras e 8894 sondas. Assim, os 3 DGs estão finalizados e prontos para a aplicação do mapa modular.

Os DGs (*Dataframe* Geral) dos 3 tipos de dados (RNA-seq, metilação e miRseq) são preparados segundo as etapas ilustradas na Figura 3.4. Após os processos de concatenação e de filtragens foram obtidas 3 DGs: o de RNA-seq com 14893 sondas e 8913 amostras,o de metilação com 8894 sondas e 8756 amostras e, por fim, o de miRseq com 278 sondas e 9161 amostras. O número baixo de sondas para o DG de miRseq se deve à falta de informações de miRNAs que estão disponíveis. A partir dos DGs é possível continuar com a metodologia, o próximo objetivo é a determinação dos BPs cujas alterações são estatisticamente significantes. Isto é realizado segundo as etapas ilustradas na Figura 3.5:

Fluxograma de Dados: Fase 4 (BPs com alterações Significantes)

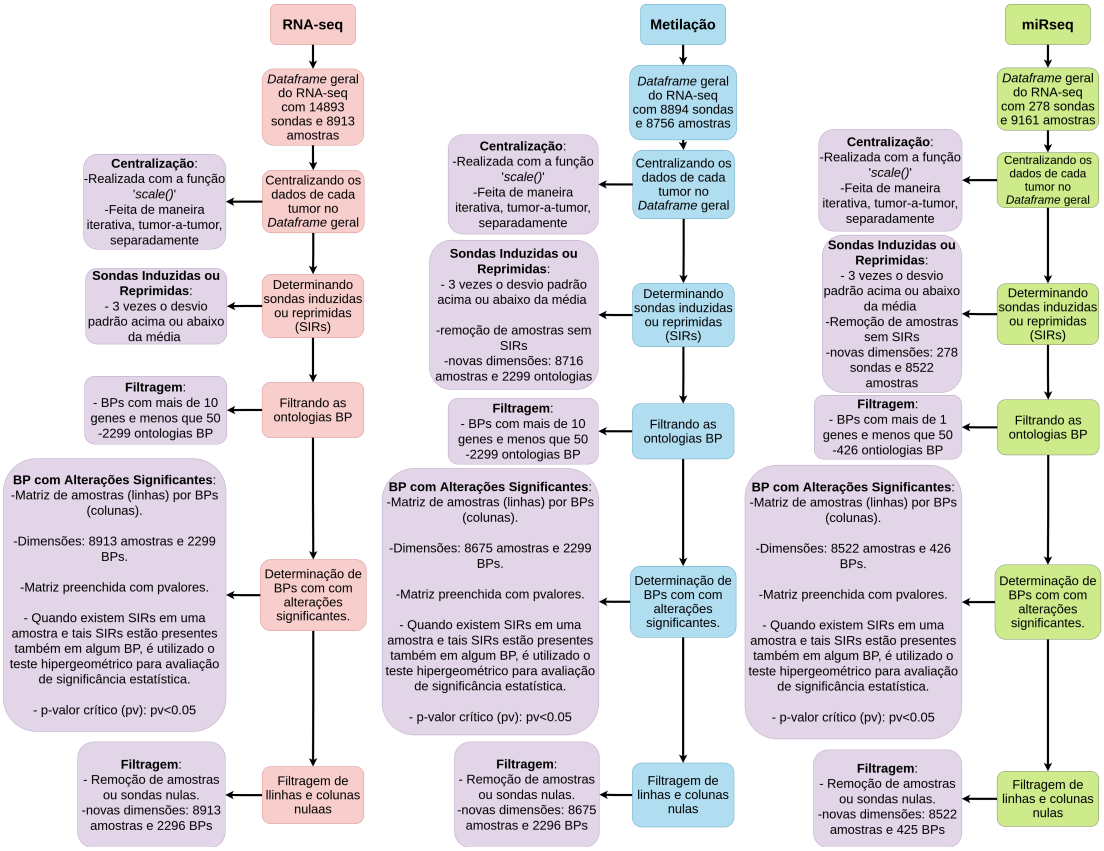


Figura 3.5: Fluxograma de Dados: Fase 4 (BPs com alterações Significantes). Os fluxogramas de cada um dos 3 tipos de dados são representados com cores diferentes: RNA-seq (rosa), metilação (azul) e miRseq (verde). Tais fluxogramas estão organizados em uma hierarquia descendente, sendo assim, a ordem flui de cima para baixo. Note que alguns passos dos fluxogramas possuem uma seta fluindo para um retângulo roxo a sua esquerda, tais retângulos contêm informações detalhadas a respeito de seu passo respectivo. Cada um dos fluxogramas são iniciados a partir dos 3 DGs construídos segundo a Figura 3.4. A primeira etapa dessa fase é a centralização dos valores de expressão/metilação dos DGs. Isto é feito com auxílio da função "*scale()*" que é aplicada as linhas da matriz. Note que esse procedimento é realizado tumor-a-tumor, ou seja, primeiro são selecionadas as amostras pertencentes ao tumor em questão para somente depois realizar a centralização. A segunda etapa é a determinação de sondas induzidas ou reprimidas (SIRs) em cada um dos DGs centralizados. Foram consideradas SIRs as sondas com *z-score* maior que 3 ou menor que -3, lembrando que o *z-score* são agora os valores do DG centralizado e mede quantos desvios padrão o valor antigo estava acima ou abaixo da média. Essa etapa alterou o tamanho dos DGs de metilação e miRseq, pois houveram amostras que não apresentaram nenhuma SIR e tiveram que ser removidas do respectivo DG. A próxima etapa foi a filtragem dos BPs que possuíam menos de 10 genes ou mais de 50 genes. Isso é feito para reduzir o universo de BPs, mantendo um equilíbrio entre generalização e especificidade, isto é discutido na Subseção 3.2.4. Observe, entretanto, que o miRseq filtrou apenas os BPs que possuíam mais de 50 genes, isto foi necessário pois existe pouca informação disponível sobre os miRNAs e remover os BPs com menos de 10 genes iria, como consequência, resultar na remoção de mais de 80% das amostras por falta de SIRs. A seguir, foi realizada a determinação dos BPs cuja alteração é significante, por meio da aplicação de um teste hipergeométrico. Assim foi criada uma nova DG, constituída de amostras (linhas) e de BPs (colunas), que é preenchida com os p-valores calculados pelo teste hipergeométrico. Por fim, realizamos a filtragem dessa nova DG de p-valores, onde são removidas linhas colunas nulas. Os DGs de p-valores tem as seguintes dimensões: 8913 amostras e 2296 BPs (RNA-seq), 8756 amostras e 2296 BPs (metilação) e de 8522 amostras e 426 BPs (miRseq).

Após as etapas da Figura 3.5, foram obtidas 3 DGs de p-valores, que por sua vez foram convertidas em matrizes de porcentagem de expressão segundo as etapas da Figura 3.6:

Fluxograma de Dados: Fase 5 (Matriz de Porcentagem de Expressão)

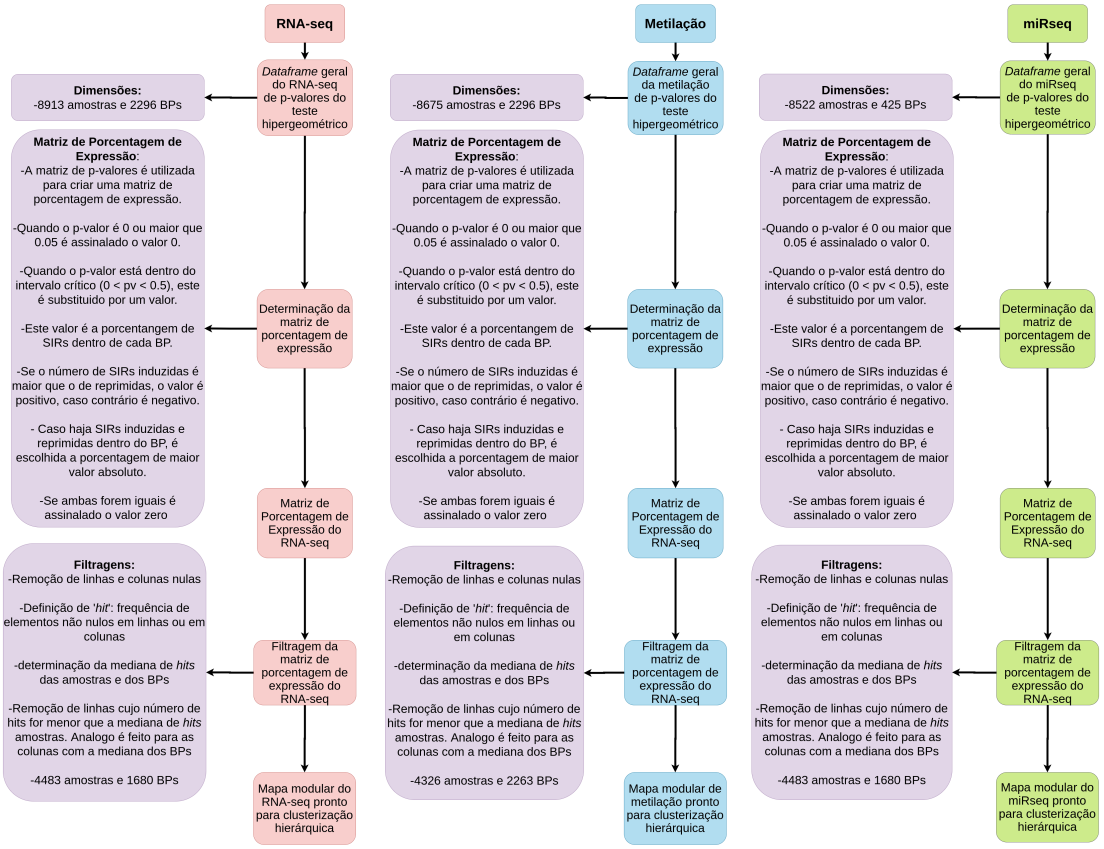


Figura 3.6: Fluxograma de Dados: Fase 5 (Matriz de Porcentagem de Expressão). Os fluxogramas de cada um dos 3 tipos de dados são representados com cores diferentes: RNA-seq (rosa), metilação (azul) e miRseq (verde). Tais fluxogramas estão organizados em uma hierarquia descendente, sendo assim, a ordem flui de cima para baixo. Note que alguns passos dos fluxogramas possuem uma seta fluindo para um retângulo roxo a sua esquerda, tais retângulos contêm informações detalhadas a respeito de seu passo respectivo. Cada um dos fluxogramas são iniciados a partir dos 3 DGs de p-valores construídos segundo a Figura 3.5. A primeira etapa dessa fase é a conversão da matriz de p-valores do teste hipergeométrico para uma matriz de porcentagem de expressão. Isso é feito substituição dos p-valores por um valor de porcentagem de expressão positivo, negativo ou nulo. Caso o p-valor seja 0 ou maior que 0.05, o valor de porcentagem de expressão assinalado é nulo. Caso contrário é calculado um valor de porcentagem de SIRs existentes dentro do BP. Tal valor é positivo se o número de SIR induzidas for maior que o de reprimidas e vice-versa. Caso haja SIRs induzidas e reprimidas dentro de um mesmo BP, foi escolhida a porcentagem de expressão que possuía o maior valor em módulo, se ambas forem idênticas o valor de porcentagem de expressão era anulado. Assim, 3 DGs de porcentagem de expressão são construídas, a próxima etapa é a filtragem das mesmas. O processo de filtragem visa remover linhas e colunas de baixa qualidade ou de pouco valor informacional, reduzindo o tamanho do mapa modular e possibilitando uma clusterização mais rápida, um tópico que sera discutido mais a frente. O processo de filtragem incluí a remoção de linhas e colunas inteiramente nulas. Posteriormente foi determinado os valores de "*hits*" nas linhas e nas colunas, e então, determinado a mediana de "*hits*". Entende-se por "*hit*" a frequência de elementos não nulos no vetor em questão (linha ou coluna). Linhas e colunas que apresentarem um valor de "hits" menores que a mediana foram removidas por serem consideradas de baixo valor informacional. Por fim, temos 3 DGs de porcentagem de expressão devidamente filtrados e com as seguintes dimensões: 4483 amostras e 1680 BPs (RNA-seq), 4326 amostras e 2263 BPs e , por fim, 4600 amostras e 213 BPs.

Ao concluir todas as etapas descritas na Figura 3.6, foram obtidos os 3 DGs de porcentagem de expressão, que na verdade são os mapas modulares, porém estes ainda precisam passar pelo processo de clusterização hierárquica para definir os módulos. Esse procedimento é feito de acordo com as etapas ilustradas na Figura 3.7:

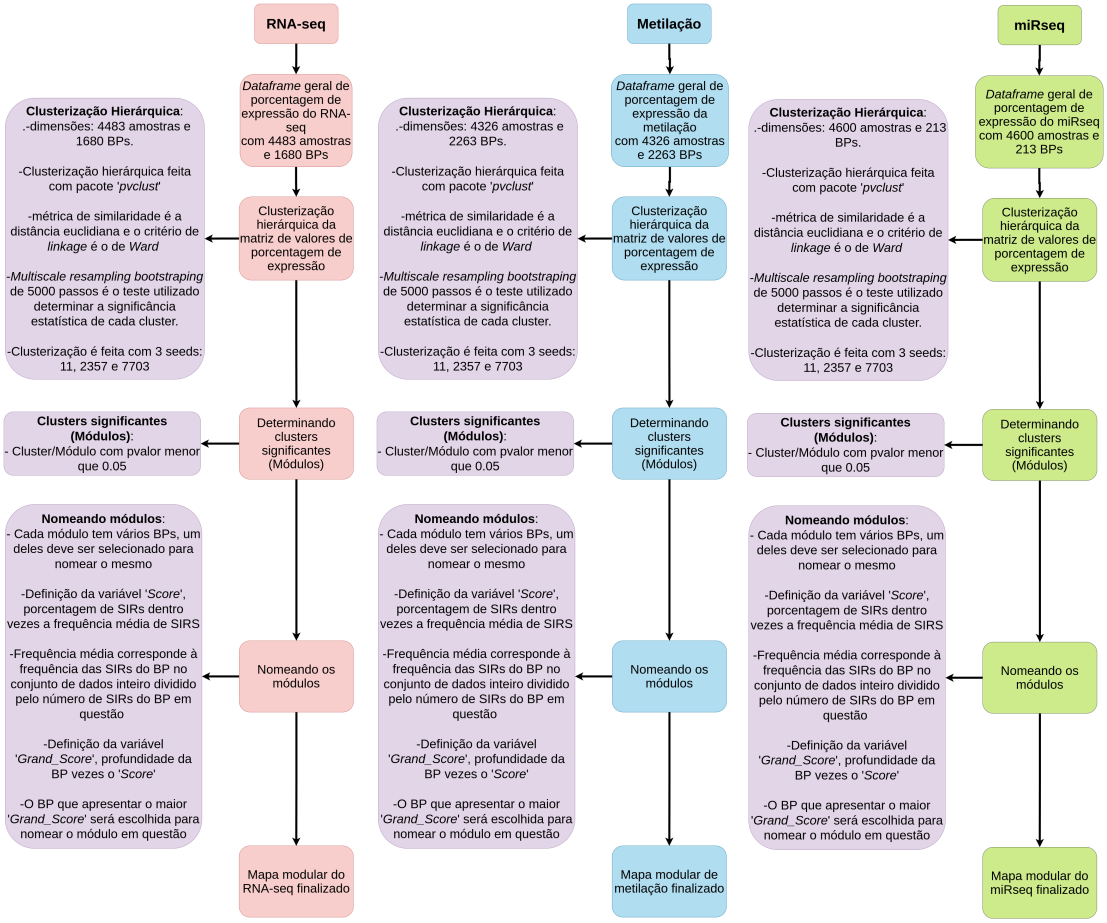


Figura 3.7: Fluxograma de Dados: Fase 6 (Clusterização Hierárquica). Os fluxogramas de cada um dos 3 tipos de dados são representados com cores diferentes: RNA-seq (rosa), metilação (azul) e miRseq (verde). Tais fluxogramas estão organizados em uma hierarquia descendente, sendo assim, a ordem flui de cima para baixo. Note que alguns passos dos fluxogramas possuem uma seta fluindo para um retângulo roxo a sua esquerda, tais retângulos contêm informações detalhadas a respeito de seu passo respectivo. Cada um dos fluxogramas são iniciados a partir dos 3 DGs de porcentagem de expressão segundo a Figura 3.6. A primeira etapa dessa fase é a clusterização hierárquica, feita com o pacote "*pvclust*", utilizando distância euclidiana como métrica de similaridade e critério de *Ward* como critério de similaridade (*Linkage*). Para determinar a significância estatística dos clusters (módulos) definidos foi utilizado o *multiscale resampling bootstraping* com 5000 passos para determinação do "*approximately unbiased p-value*" (au). Os clusters que apresentarem um au menor que 0.05 serão cosiderados como significantes. O procedimento de clusterização hierárquica é realizado 3 vezes para cada mapa modular, fixando *seeds*: 11, 2357 e 7703. Após a determinação dos módulos é necessário nomeá-los. Cada módulo possui diversos BPs dentro de sí e um deles foi selecionado para nomear cada módulo. Para realizar tal procedimento foi criado uma medida chamada "*Grand_Score*". Primeiramente foi determinado o número de SIRs dentro de cada BP dentro de cada módulo. Depois foi determinado a frequência média de SIRs no mapa modular. Isto é feito determinando os SIRs existentes dentro de cada BP do módulo e calculando a frequência destes SIRs no mapa modular, ou seja, o número de amostras em que a SIR em questão é considerada como uma SIR também. Esta frequência é então dividida pelo número de SIrs existentes dentro do BP. Ao multiplicar o número de SIRs do BP pela sua frequência média, calculamos o "*Score*". A seguir, o "*Score*" é multiplicado pelo valor de profundidade do BP, criando assim o "*Grand_Score*". Dessa maneira o "*Grand_Score*" é uma medida que considera: o número SIRs do BP, o quão prevalente essa sir esta no mapa modular e a especificidade do BP. O BP que apresentar o maior valor de "*Grand_Score*" dentro do módulo é selecionado para nomeá-lo. Ao fim, temos os 3 mapas modulares clusterizados e finalizados. Com os módulos devidamente selecionados e nomeados é possível prosseguir para as análises feitas no capítulo 4.

4 Resultados e Discussão

A primeira parte dos resultados constitui uma breve avaliação dos parâmetros utilizados para realizar o processo de clusterização dos dados, onde é explicado o porque de se utilizar tais parâmetros. Na segunda parte é realizada a descrição e análise dos 3 mapas modulares construídos (RNA-seq, metilação e miRNA), além da visualização dos mapas construídos será discutido quais módulos encontrados são interessantes em um contexto de câncer e também quais os módulos mais consistentemente induzidos ou reprimidos nos 32 tumores contemplados. Na terceira e última parte, é realizada uma análise mais minuciosa do tumor de mama (BRCA), onde foi discutido o perfil de expressão de SIRs do tumor.

4.1 Meta Análise da Clusterização Hierárquica e Visualização do Mapa Modular

Antes de realizar qualquer interpretação e validação biológica, é necessário avaliar se as metodologias estatísticas produzem resultados congruentes. Na Subseção 3.2.8, é mencionado a possibilidade de que o método aglomerativo e a métrica de distância utilizados por *Segal* podiam não apresentar a melhor convergência de resultados. Nesta Subseção, foi feita uma meta análise dos resultados do processo de clusterização hierárquica de um objeto formado com os dados dos 32 tumores, o mesmo que será analisado nos passos a frente, considerando a distância euclidiana e critério de *Ward*.

O mapa modular foi desenvolvido para ser uma metodologia baseada em visualização e sumarização de grandes quantidades de dados de expressão. A saída do método é um *heatmap* mostrando como os módulos com alto grau de significância estão expressos nas condições experimentais enriquecidas com anotações de alguma natureza. Por fim, deve ficar claro, então, que a principal função do mapa modular é análise por meio da sumarização e

visualização de dados segundo os seguintes aspectos:

- Quais são os processos biológicos (módulos) mais significativos encontrados no universo de genes/sondas fornecido?
- Qual o comportamento da expressão desses módulos em função do tipo de tumor?

Em materiais e métodos (Subseção 3.2.8) é mencionado o método do *bootstrap resampling* possui uma aleatoriedade intrínseca, mas possui também uma garantia de convergência para um número de passos de *bootstrap* suficientemente grande (SHIMODAIRA, 2013). Dessa maneira é possível conseguir inferir a partir de uma amostra o comportamento geral da população total. Cada conjunto de dados tem uma taxa de convergência diferente a depender de seu tamanho, interdependência das variáveis e complexidade geral. Cabe ao pesquisador determinar qual o número de passos a utilizar observando a convergência dos resultados (SHIMODAIRA, 2013).

A convergência dos resultados, independente da *seed* utilizada, é algo de absoluta importância para a metodologia do mapa modular, pois é a partir do *bootstrap resampling* que o *Approximatelly Unbiased p-value* (AU) é estimado e consequentemente os módulos que são estatisticamente significantes no conjunto de dados em questão (SHIMODAIRA, 2013). Se utilizar de resultados (módulos) não convergentes nessa metodologia seria o equivalente de simplesmente selecionar um número x de módulos aleatoriamente dentro de um universo com n módulos para realizar a análise visual, algo que não nos interessa, pois deseja-se fazê-lo para um conjunto de módulos que de fato sejam significantes segundo os dados fornecidos.

Um desafio que surgiu para execução da clusterização hierárquica foi o custo computacional do procedimento. Por custo computacional se entende o tempo necessário para realizar algum procedimento. O processo de clusterização hierárquica foi feito no cluster de computação do núcleo de computação ciêntifica da UNESP, chamado de "GridUnesp". Estavam disponíveis para uso um total de 55 *threads* ou núcleos de processamento, todos utilizados para realização da clusterização hierárquica, que estava devidamente paralelizada no *script* em R. No caso do mapa modular do RNA-seq, cujo DG de porcentagem de expressão tinha 4483 amostras e 1680 BPs (7531440 elementos), foi necessário em média 50 horas ou pouco mais de 2 dias para realizar a clusterização hierárquica de 5000 passos. Para os

dados de metilação, cujo DG de porcentagem de metilação tinha 4326 amostras e 2263 BPs (9789738 elementos), foi necessário em média 120 horas ou 5 dias para realizar a clusterização hierárquica de 5000 passos. Por fim, para os dados de miRseq, cujo DG de porcentagem de expressão tinha 4600 amostras e 213 BPs (979800 elementos), foi necessário em média 40 minutos para realizar a clusterização hierárquica de 5000 passos. Os dados miRseq demoram significantemente menos para serem clusterizados devido ao baixo número de BPs, não apresentando tanto problema no quesito custo computacional. O mesmo não pode ser dito sobre os dados de RNA-seq e de metilação, ambos levando dias para a conclusão do procedimento. Isto apresenta um problema, pois o GridUnesp é de uso compartilhado entre diversos alunos da UNESP, o que significa que pode existir uma fila de procedimentos (*jobs*) de outros usuários esperando para serem executados e cuja a duração de execução é desconhecida para terceiros.

O alto custo computacional da utilização de um número muito elevado de passos de *bootstrap*, nos levou a seguinte abordagem para seleção de módulos. Primeiramente, foi realizada a clusterização hierárquica da matriz de porcentagem expressão (DG de porcentagem de expressão) dos 32 tumores para 3 *seeds* (11, 2357 e 7703), com 5000 passos de *bootstrap*, distância euclidiana e critério de *Ward*. Para os dados de RNA-seq foram necessários 6 dias para a execução das 3 clusterizações hierárquicas de 5000 passos, 15 dias para os dados de metilação e apenas 1 dia para os dados de miRseq. Isso totaliza um mínimo de 22 dias para execução completa da clusterização hierárquica, contudo, foram necessários pouco mais 60 dias (2 meses) para a execução devido a fila de *jobs* de outros usuários. Por fim, os módulos estatisticamente significantes foram prospectados segundo o *bootstrap resampling* de *Shimodaira* e colaboradores (*Aproximately Unbiased (au)*, $\boldsymbol{au} \geq 0.95$) e, então foram filtrados os resultados convergentes entre as 3 *seeds*. Dessa maneira, foi possível utilizar apenas a convergência dos resultados das *seeds* testadas, ou seja, módulos estatisticamente significantes independente da aleatoriedade intrínseca do método.

Foram selecionados 69 módulos para o mapa modular de RNA-seq. Devido ao elevado número de módulos foram escolhidos aqueles cuja ontologia representada tivesse um nível de profundidade maior que 10, essa questão do nível de profundidade de uma ontologia é abordado na Subseção 3.2.4. Dessa maneira, um total de 35 módulos foram selecionados, uma legenda de cores foi criada para identificá-los, essa legenda pode ser vista na Figura 4.1. Em seguida, foi realizada a visualização e interpretação dos dados com o *heatmap* do

mapa modular, a saída bruta da metodologia pode ser observada na Figura 4.2, junto com uma descrição mais técnica sobre a organização da imagem.

Legenda de cores do mapa modular de RNA-seq para 32 tumores

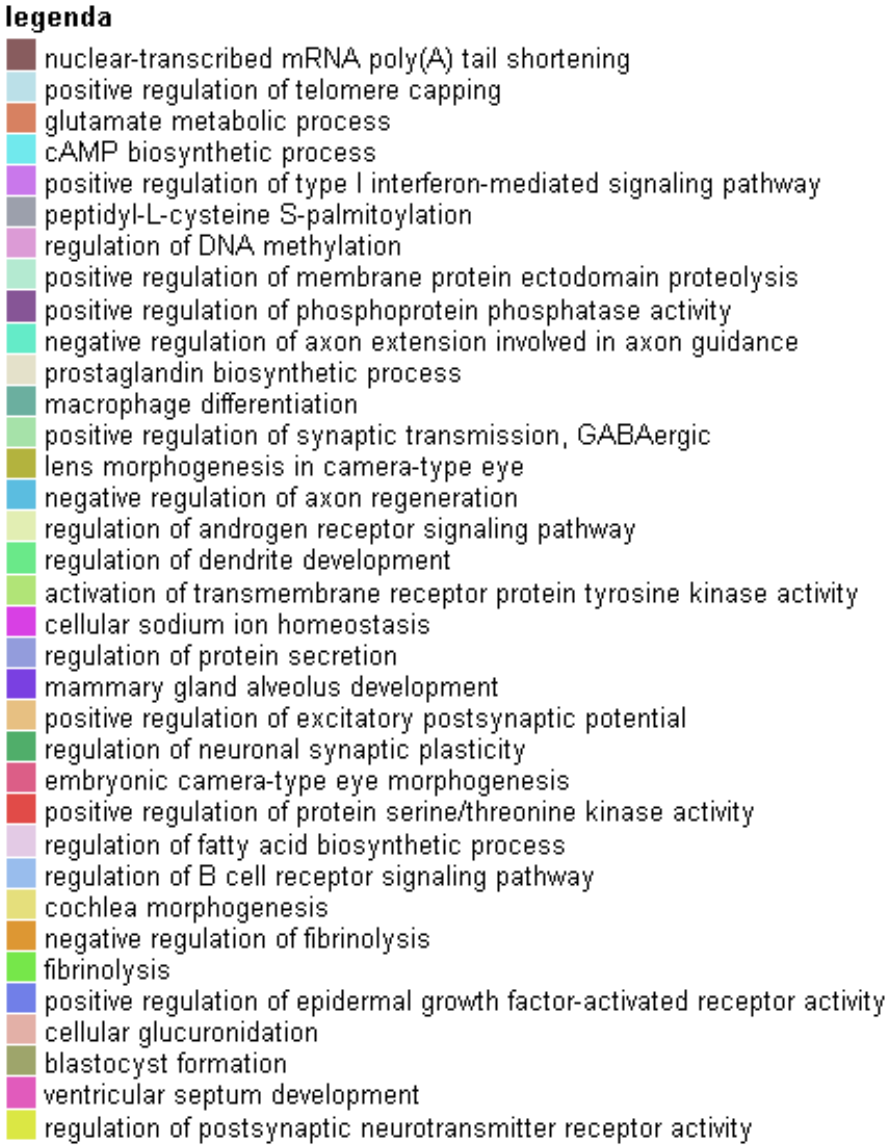


Figura 4.1: Legenda de cores do mapa modular de RNA-seq para 32 tumores. Pode-se observar na figura os 35 módulos selecionados e sua respectiva cor assinalada.

Heatmap do mapa modular de RNA-seq para 32 tumores com clusterização modular interna

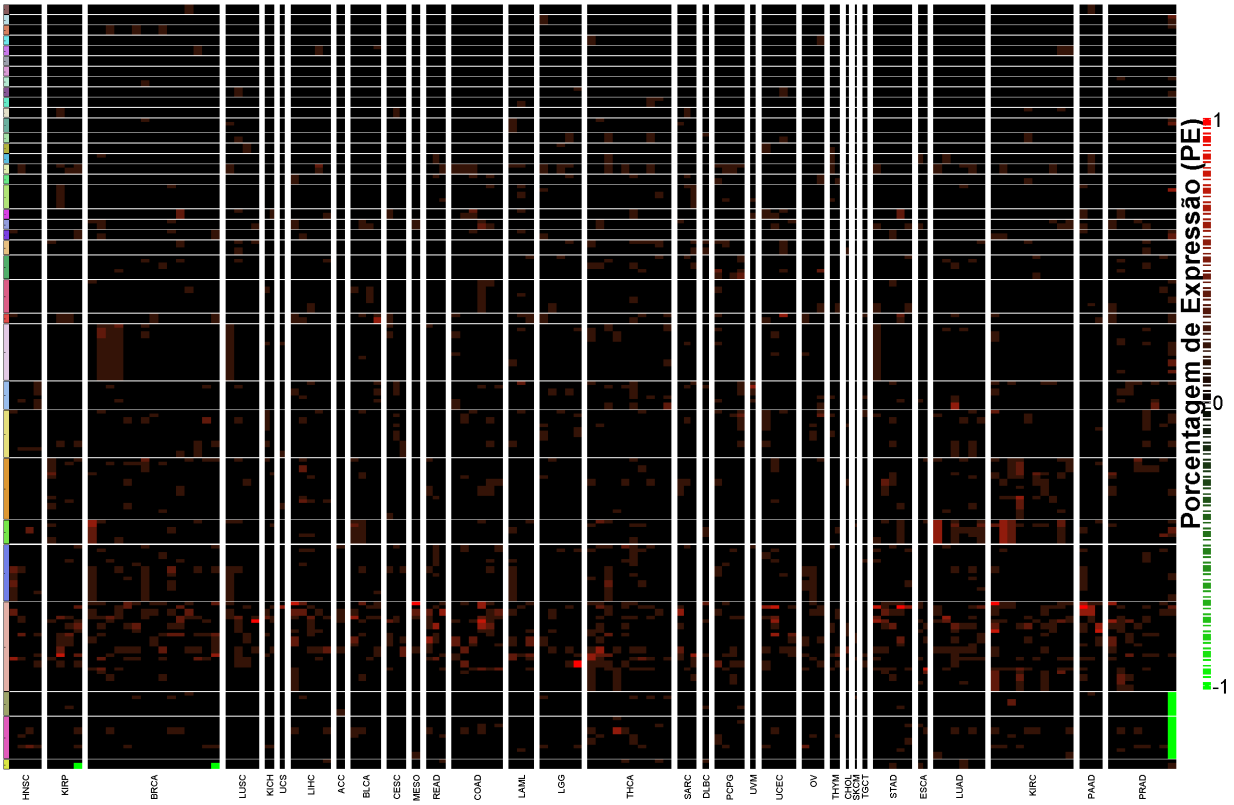


Figura 4.2: *Heatmap* do mapa modular para 32 tumores com clusterização modular interna. As linhas representam os processos biológicos enquanto as colunas representam os *arrays* (ou experimento). Nesse *heatmap*, os 35 módulos selecionados, que são conjuntos de processos biológicos, são representados pela barra colorida, na lateral esquerda do *heatmap* e portanto constituem o eixo *y* da figura. Existem 32 colunas na figura, cada uma nomeada de acordo com um tumor na parte inferior, este constitui o eixo *x* da figura. A respectiva associação de cores a seus módulos é feita na Figura 4.1. As colunas desse *heatmap* foram organizadas de tal forma que é possível dividi-las de acordo com o tumor a que os experimentos pertencem, de fato, o nome de cada uma das colunas definidas pelas divisões verticais do *heatmap* é referente a cada um dos 32 tumores utilizados. As divisões horizontais são referentes aos módulos, de maneira que é possível dividir o *heatmap* original (dos 32 tumores para todos os 35 módulos) em 1120 partes menores, que são também *heatmaps* específicos para tumor/módulo. Após a divisão, clusterização hierárquica é aplicada separadamente a cada um dos 1120 *heatmaps menores* (clusterização modular interna) para fins visuais. Cada *heatmap* é preenchido por valores de Porcentagem de Expressão (PE), os quais são definidos na Subseção 3.2.7. Estes valores estão localizados em um intervalo de [-1,1], como mostra a legenda de cores à direita da figura. Sendo que 1 indica indução, representada pela cor vermelha, enquanto -1 indica repressão, representada pela cor verde. Os valores foram amplificados em 2000 vezes devido ao baixo módulo da absoluta maioria das células não nulas, facilitando a visualização na escala de valores utilizada [-1,1].

O *heatmap* na Figura 4.2 foi gerado a partir do objeto de dados de RNA-seq saída da metodologia do mapa modular adaptado. A seleção de módulos estatisticamente significantes (35) foi previamente realizada, selecionando um total de 156 processos biológicos agrupados em 35 módulos, representados pela legenda de cores na lateral esquerda do *heatmap*.

Os valores de expressão gênica relativa (ER) são obtidos a partir da determinação sondas induzidas e reprimidas (SIRs), processo descrito na Subseção 3.2.6, e da construção da matriz

de *gene sets* x *Arrays*, também chamada de DG (*Dataframe* Geral) de Porcentagem de Expressão, segundo os critérios estabelecidos na Subseção 3.2.7. Assim cada valor anotado nas células dessa matriz é na verdade a porcentagem de sondas do processo biológico que são considerados como SIRs, essa porcentagem é positiva para indução e negativa para repressão. Esse valor que leva em consideração a expressão de SIRs e sua representação no processo biológico em questão é aqui chamado de Porcentagem de Expressão (PE) e corresponde a escala de valores entre -1 a 1 que é ilustrada na matriz da figura 4.2. A escala de Porcentagem de Expressão (PE) pode ser observada na legenda a direita. Elementos da matriz com PE positiva indicam indução (em vermelho) e valores negativos (em verde) indicam repressão da transcrição gênica.

Após essa segmentação, foi realizada uma segunda clusterização hierárquica interna em cada um dos 1120 *heatmaps*, essa clusterização também utiliza a distância euclidiana e critério de *Ward*. Note que esta segunda clusterização tem fins puramente visuais e não tem relação direta alguma com a clusterização hierárquica que foi realizada para determinação dos módulos estatisticamente significantes descrito na Subseção 3.2.8. Em função disso foi observado o comportamento de Figura 4.2 em que os elementos de cada um dos 1120 *heatmaps* se agrupam de acordo com a magnitude de sua expressão, pois o critério de *Ward* tenta minimizar as variâncias internas dos *clusters* gerados no dendrograma.

O *heatmap* bruto do mapa modular do RNA-seq na Figura 4.2 deixa algo bem claro: essa forma bruta do mapa modular tem pouco valor como ferramenta de visualização. Isso é consequência natural da grande quantidade de amostras representadas no *heatmap*. Especificamente é consequência do fato do valor de ER ser uma minuscula parcela (SIRs) é um grande universo de sondas. Assim pouquíssimos elementos podem ser visualizados considerando a escala de possíveis valores [-1,1]. A absoluta maioria das células do *heatmap* são nulas, como é esperado. Algumas poucas células que possuem valores diferentes de 0 mas ainda assim esses são muito baixos (próximos de 0). Como consequência, multiplicamos a todos os valores na figura por 2000, valor determinado empiricamente, para facilitar a visualização.

Para contornar o problema de quantidade de dados na figura, foi construída a versão reduzida do mapa modular. Esta versão resumida do mapa modular, para a Figura 4.2, pode ser observada na figura 4.3:

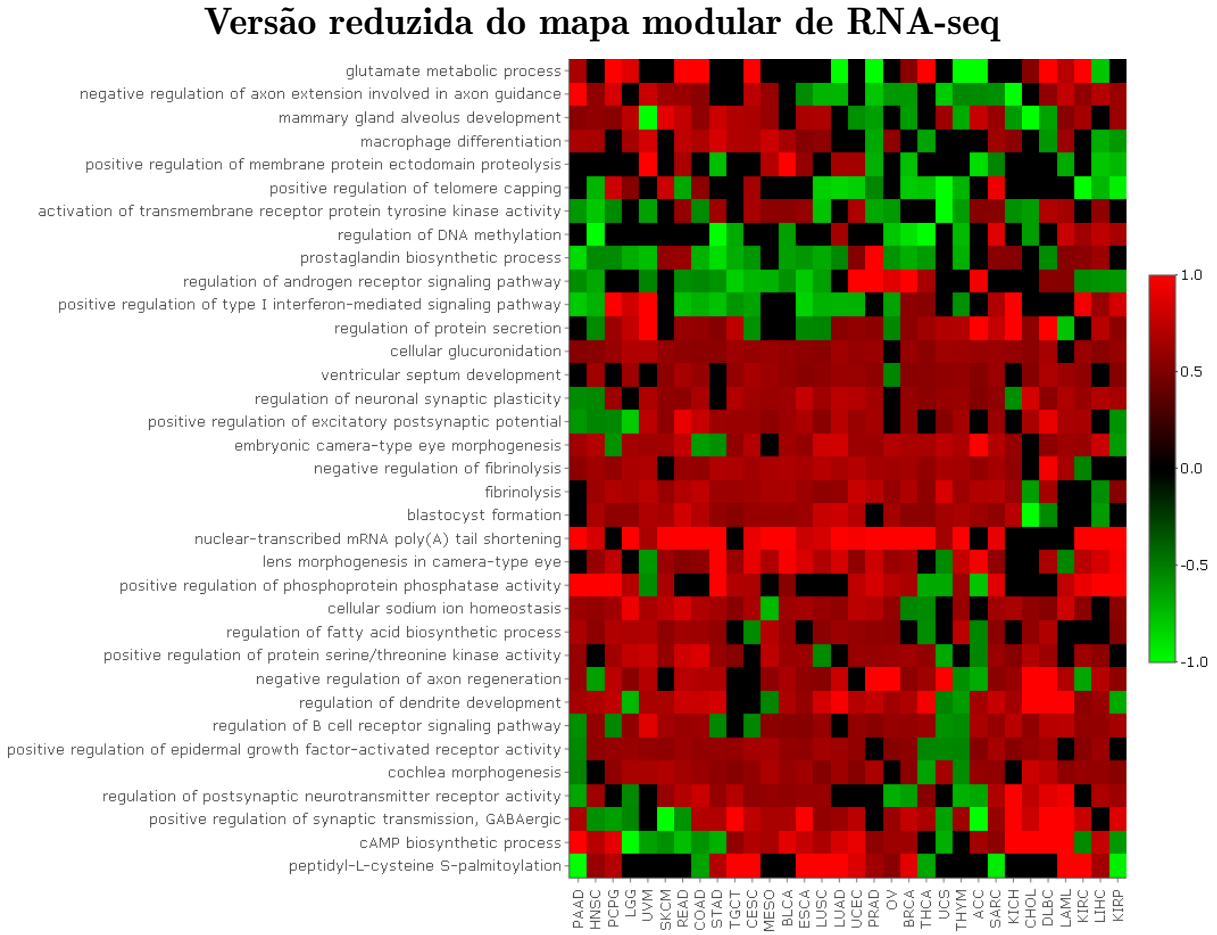


Figura 4.3: Versão reduzida do mapa modular. Essa versão do mapa modular foi desenvolvida com intenção de facilitar ainda mais a visualização do comportamento da expressão dos módulos selecionados. Ele é construído a partir dos 1120 *heatmaps* menores que compõe o *heatmap* original, sendo que no eixo y tem-se os 35 módulos selecionados e no eixo x os tumores. O valor *PEG*, chamado de Porcentagem de Expressão Geral (PEG), de cada célula da figura é determinado por meio das porcentagens de elementos induzidos e reprimidos, a eles é assinalado o maior valor dos dois. Caso os valores não tenham uma diferença percentual maior que 0.1 (10%), então é assinalado o valor zero para o mini *heatmap*, indicando não haver diferença significativa entre as porcentagens de valores induzidos e reprimidos. Portanto os valores *PEG* estão localizados em uma escala de [-1,1], representada na legenda a direita do *heatmap*. Os valores *PEG* entre $-1 \leq PEG < 0$ (verde) indicam repressão, $PEG = 0$ (preto) indica ausência de expressão dominante e valores entre $0 < PEG \leq 1$ (vermelho) indicam indução.

A Figura 4.3 deixa a análise visual de múltiplos tumores mais exequível, nos permitindo enxergar o grande objetivo da metodologia proposta na tese: visualizar módulos consistentemente induzidos ou reprimidos segundo os dados de expressão. Na próxima subseção foram discutidos os módulos consistentemente induzidos ou reprimidos segundo os dados de RNA-seq de 32 tumores.

4.1.1 RNA-seq: Módulos Consistentemente Induzidos ou Reprimidos

Existe uma maior quantidade de elementos induzidos do que reprimidos de acordo com o mapa modular reduzido do RNA-seq, note a predominância da cor vermelha na Figura 4.3, que indica PEG positiva. Alguns dos módulos estão quase que totalmente induzidos em todos os 32 tumores ou pelo menos na maior parte deles. Foram selecionados 9 módulos com um elevado número de tumores induzidos (vermelho):

1. *fibrinolysis.*
2. *negative regulation of fibrinolysis.*
3. *cellular glucuronidation.*
4. *ventricular septum development.*
5. *nuclear-transcribed mRNA poly(A) tail shortening.*
6. *positive regulation of phosphoprotein phosphatase activity.*
7. *cellular sodium ion homeostasis.*
8. *regulation of fatty acid biosynthetic process.*
9. *positive regulation of protein serine/threonine kinase activity.*
10. *regulation of B cell receptor signaling pathway.*
11. *positive regulation of epidermal growth factor-activated receptor activity.*

Os módulos "*fibrinolysis*" e "*negative regulation of fibrinolysis*" mostram como o processo de fibrinólise está negativamente regulado em quase todos os 32 tumores. Fatores hemostáticos e componentes do sistema fibrinolítico como fibrinogênio e a fibrina têm uma complexa interação com o desenvolvimento do tumor e da metástase. A mais de 100 anos já se sabe que células tumorais se envolvem em um trombo de fibrina, promovendo uma matriz pela qual a migração e o processo metastático podem ter início através de moléculas adesivas e integrinas (MAHMOOD; RABBANI, 2021). Outros agentes do sistema fibrinolítico têm efeitos diferentes em células tumorais, a enzima plasmina pode catalisar a degradação dos

coágulos de fibrina e da matriz por eles formada, no entanto, também pode ativar fatores de crescimento latentes.

O ativador do plasminogênio uroquinase (uPA) e seu receptor (uPAR) estão envolvidos em diversas vias importantes para manutenção do tumor incluindo: formação de complexos com vitronectina e integrinas para adesão celular a ECM, formação de complexos com o receptor do fator de crescimento epitelial nas membranas celulares para sinalização das vias *RAF-MEK-ERK* e então forma complexos com o receptor da proteína acoplada G para regulação de vias associadas a proliferação celular. O inibidor do plasminogênio I (PAI-1) inibe vias apoptóticas, aumentando as chances de sobrevivência das células tumorais, também está associado à senescência celular e aumento da secreção de citocinas carcinogênicas encontradas no secretoma da senescência (MAHMOOD; RABBANI, 2021).

Em geral, os estudos mostram que presença excessiva de uPA/uPAR e PAI-1 estão positivamente correlacionadas com agressividade tumoral e prognósticos ruins. No mapa modular reduzido, a fibrinólise está amplamente inibida, sugerindo uma elevada expressão dos agentes mencionados acima (uPA/uPAR/PAI-1) (KWAAN; LINDHOLM, 2019). Essas são observações interessantes, considerando como o módulo *positive regulation of epidermal growth factor-activated receptor activity* também se encontra induzido na maioria dos tumores, talvez uma consequência da regulação negativa do processo de fibrinólise. É necessário uma avaliação dos SIRs contidos nos módulos e de sua expressão, isso foi realizado a frente para o tumor BRCA.

O módulo *cellular glucuronidation* também é de interesse num contexto de câncer. A glucuronidação é uma importante reação da via de metabolismo de fase 2 responsável pela limpeza de vários compostos exógenos e endógenos. Essa reação é catalisada por enzimas chamadas UDP-glucuronosiltransferases (UGP), que realiza a adição do ácido glucurônico a moléculas lipofílicas para transporte. No câncer, a principal função creditada ao processo de glucuronidação é justamente a resistência às drogas de tratamento, embora estudos mostrem que possa haver outros meios pelo qual essa via possa afetar o câncer em um cenário independente das drogas, considerando metabólitos endógenos como lipídios e esteroides (ALLAIN et al., 2020). *Allain* e colaboradores (2020) sugerem, então, que a glucoronidação tem uma dupla função em patologias, primeiro a expressão aberrante desta via pode dar inicio e sustentar a progressão de doenças e segundo tem-se a sua conhecida função de promover resistência às drogas, concedendo à patologia um fenótipo de resistência/reincidência

(ALLAIN et al., 2020). *Zimmer* e colaboradores (2021), demonstram isso em um estudo sobre o câncer de próstata, que revela como níveis elevados da enzima UGP promoveu a inibição dos receptores de androgênio e a transcrição por eles mediada, aumentou a síntese de proteoglicanos, elevou significativamente a taxa de crescimento celular e induziu um fenótipo de resistência às drogas nas células (ZIMMER et al., 2021).

Curiosamente o módulo *regulation of androgen receptor signaling pathway* é um dos módulos mais reprimidos no mapa modular. Para compreender melhor o que essa inibição sugere é necessário avaliar a expressão das SIRs em cada tumor. O tumor PRAD que foi estudado por *Zimmer* e colaboradores (2021), é um dos poucos que apresenta indução desse módulo, assim, é possível presumir que essa indução se refere a agentes inibidores dessa via. O tumor BRCA também está altamente induzido para este módulo, no entanto, *Dates* e colaboradores mostram como a expressão de UGP esta reprimida neste tipo de tumor e que o acumulo de lipídios bioativos na célula tumoral favorece seu desenvolvimento. Essa situação com a enzima UGP do tumor de mama é a oposta do que é relatada por *Zimmer* e colaboradores no tumor de próstata, entretanto, ambos tumores apresentam o que, possivelmente, é uma regulação negativa da transcrição mediada por receptores de androgênio (ZIMMER et al., 2021).

O módulo "*regulation of B cell receptor signaling pathway*" é algo notável num contexto de câncer pois representa uma maneira pela qual o sistema imunológico possa ter alguma influência sobre a doença. Pode-se observar que o módulo está majoritariamente induzido (24) e possui poucos elementos pretos (2) na Figura 4.3, indicando que o processo biológico representado está, de fato, em evidência em quase todos os 32 tumores. Células B são linfócitos constituintes do sistema imunológico que possuem a capacidade de se diferenciar em células secretoras de anticorpos mediante reconhecimento de antígenos por receptores de células B (BCR), que se encontram na membrana dessas células, uma característica única dessas células em comparação com os outros dois tipos de linfócitos, células T e células NK (*natural killers*). *Burger* e colaboradores (2018), comentam como as células T são amplamente estudadas no contexto do câncer em preferência às células B, no entanto está ultima tem ganhado atenção da comunidade científica nos últimos anos. (BURGER; WIESTNER, 2018). Células B, como qualquer outra célula, também podem sofrer de processos tumorais, principalmente através da expressão aberrante da via de sinalização dos BCRs, que modulam vias de desenvolvimento e de viabilidade dessas células. Isso afeta profundamente

a imunidade adaptativa do organismo além das neoplasias derivadas das células B como linfomas e leucemias. Entretanto, o estudo de *Burger* e colaboradores (2018), chama atenção para como as vias de sinalização mediadas pelas BCRs podem afetar a manutenção do microambiente tumoral de outros tipos de tumores sólidos, uma possibilidade com grande potencial terapêutico e pouco explorada segundo o autor (BURGER; WIESTNER, 2018).

Fridman e colaboradores (2021), também discorrem sobre a possibilidade de uso terapêutico das vias mediadas pelos BCRs, e como células T podem estar envolvidas tanto em efeitos promotores de carcinogênese quanto em efeitos repressores. *Fridman* e colaboradores (2021), também reiteram a natureza variante que BCRs podem ter não apenas entre diferentes tecidos mas também entre diferentes patologias em um mesmo tecido, indicando um desafio a ser superado para a implementação terapêutica de inibidores de BCRs devido a variedade de situações (FRIDMAN et al., 2021). Os alvos terapêuticos mais comuns nesse contexto são inibidores das cinases tirosina cinase *Bruton* (BTK) e fosfoinositídeo 3-cinase (PI3k), como demonstrado em *Burger* e colaboradores (2018), e isso nos leva a outra observação, os módulos *positive regulation of phosphoprotein phosphatase activity* e *positive regulation of protein serine/threonine kinase activity* (BURGER; WIESTNER, 2018).

Cinases são um conjunto de proteínas (enzimas) cuja função é catalizar a transferência de um grupo de fosfato para um substrato específico. Geralmente, esse substrato se trata de outra proteína, no final do processo, se obtêm uma fosfoproteína. Analogamente, as enzimas fosfatases catalizam a transferência do grupo fosfato de uma fosfoproteína para outra molécula. Os processos de fosforilação e desfosforilação são componentes essenciais das cascatas de sinalização celulares, pois o estado de fosforilação ou não de uma molécula (proteínas, lipídios e carboidratos) afeta a sua atividade, reatividade e capacidade de interação com outras moléculas (CAPRA et al., 2006). De acordo com *Gelens* e colaboradores (2018), é de conhecimento da comunidade científica, que a fosforilação e desfosforilação estão diretamente envolvidos na regulação de diversos processos fisiológicos como: metabolismo, sinalização celular, atividade de proteínas, transporte celular e secreção (GELENS; SAURIN, 2018).

O módulo *positive regulation of protein serine/threonine kinase activity* no mapa modular reduzido do RNA-seq chama atenção para um grupo específico de cinases que tem como alvo o grupo hidroxila (OH) em cadeias de serina e treonina, chamadas de STK. Estudos mostram que das mais de 500 cinases humanas quase 125 são STKs e que estas estão associadas a regulação de processos de: proliferação celular, apoptose diferenciação celular

e desenvolvimento embrionário (CAPRA et al., 2006).

Todos esses são processos chaves para início e manutenção do câncer, e no mapa o módulo *positive regulation of protein serine/threonine kinase activity* se encontra altamente induzido (23/3) na maioria dos 32 tumores com algums elementos pretos (9). Isso demonstra como a regulação mediada pelas STKs está em evidência para o processo tumoral como um todo. As vias específicas podem ser diferentes a depender do tumor/tecido e para avaliar isso seria necessário observar a expressão dos SIRs em cada módulo. Nossos resultados mostram que o módulo *positive regulation of phosphoprotein phosphatase activity* também está majoritariamente induzido, remetendo-nos a relação entre fosforilação e desfosforilação e que ambos os processos podem alterar a atividade e reatividade das moléculas alvo (CAPRA et al., 2006).

O módulo *cellular sodium ion homeostasis* está induzido na maior parte dos tumores, algo curioso, pois alterações bruscas da concentração de íons de sódio seria algo esperado de células nervosas ou musculares. No entanto, *Roger* e colaboradores (2015), discutem como canais de sódio voltagem-dependente (NAv) estão anormalmente ativos em células cancerígenas de diversos tecidos em comparação com seu cognato normal. *Roger* também demonstra que essa atividade anormal dos NAvs está associada a reaquisição de fenótipos embrionários pelas células tumorais e essa observação foi especialmente significante em células tumorais epiteliais. De acordo com *Roger* e colaboradores (2015), a atividade anormal dos NAvs pode ter importantes implicações na regulação das propriedades proliferativas, migrativas e invasivas dessas células.

Os resultados mostram como o módulo *ventricular septum development* está altamente induzido na maioria dos tumores. Isso está de acordo com a observação de *Roger* e colaboradores (2015), pois o desenvolvimento do septo ventricular é tipicamente uma fase de desenvolvimento embrionária, na qual é um processo altamente regulado (ROGER et al., 2015). Não se pode dizer o mesmo sobre o fenótipo readquirido por células tumorais, na qual diversas vias associadas a crescimento e viabilidade celular apresentam expressão aberrante. O módulo *ventricular septum development* pode inclusive apresentar vias associadas ao conhecido fenômeno de transição epitelial-mesenquimal (EMT) que é muito observado em células tumorais (ROCHE, 2018).

Pode-se observar o módulo *nuclear-transcribed mRNA poly(A) tail shortening* altamente induzido na maioria dos 32 tumores. Após a transcrição, sequências de mRNA sofrem o

processo de regulação pós-transcricional chamado de poliadenilação, no qual uma sequência de adenosinas monofosfatadas (cauda poli-A) é adicionada ao final 3' da sequência de mRNA. Esse processo tem diversas consequências incluindo estabilidade da molécula, transporte nuclear e o processo de tradução. Essa cauda é desgastada conforme o tempo e quando está suficientemente curta a sequência de mRNA é degradada por enzimas. Em algumas células, as sequencias com caudas encurtadas podem ser armazenadas para para ativação posterior mediante o processo de repoliadenilação (ZHANG et al., 2020).

Transcritos de mRNA também podem sofrer um processo de poliadenilação alternativa, gerando diferentes transcritos poliadenilados (isoformas) que podem traduzir diferentes proteínas e possuir estabilidade e tempo de vida diferentes, tudo a partir de uma mesma sequência de mRNA. Além disso, essas isoformas apresentarão diferentes sítios de ligação para outras moléculas com função reguladora como os micro RNAs (miRNAs), que tendem a reprimir a tradução dos transcritos aos quais se ligam, embora existam exemplos de miRNAs estabilizadores.

A seleção do ponto de clivagem da sequência de mRNA e do ponto de ligação da cauda poli-A são mediadas por diversos fatores e incluem: proteínas de clivagem, concentração de fatores de processamento de mRNA, proteinas de ligação ao RNA (*RNA binding proteins*), fatores de *splicing* e metilação do DNA. Dito isso, é simples considerar que um ponto com tantas vias de regulação de diversos processos biológicos possa ser essencial para compreender os mecanismos moleculares pro trás de uma patologia.

Gruber e colaboradores (2019), discutem como a poliadenilação alternativa não é apenas um mecanismo muito utilizado para regulação em tecidos sadios, mas também amplamente presente em diversas patologias (GRUBER; ZAVOLAN, 2019). *Zlotorsky* e colaboradores (2018), também chamam atenção para o encurtamento da cauda poli-A, que geralmente é um indicador para a degradação de sequências de mRNA, mas esta é uma característica mantida em sequências que são amplamente traduzidas (ZLOTORYNSKI, 2018). Essas observações apontam para um cenário onde uma intensa expressão de sequências isoformas estão sendo traduzidas, algo plausível no contexto de células tumorais que constantemente passam por processos de grande estresse molecular como a divisão e a migração celular.

Existem poucos módulos amplamente reprimidos, aqueles um maior número de tumores reprimidos (verde):

1. *activation of transmembrane receptor protein tyrosine kinase activity.*
2. *regulation of DNA Methylation.*
3. *regulation of androgen receptor signaling pathway.*
4. *positive regulation of type I interferon-mediated signaling pathway.*

O módulo *regulation of androgen receptor signaling pathway* foi discutido anteriormente após ao módulo induzido *cellular glucuronidation*, onde foi sugerido como a inibição dos receptores de androgênio pode ser importante para manutenção do processo tumoral.

O módulo *regulation of DNA Methylation* é interessante num contexto de câncer, visto que a metilação é um mecanismo de regulação epigenética que costuma interromper a transcrição do DNA. As observações apontam para um cenário de intensa tradução de isoformas, portanto, era esperada uma menor regulação por parte da metilação. O módulo *regulation of DNA Methylation* está de fato reprimido em um grande número de tumores, corroborando para a observação que foi feita acima. No entanto, ainda existem muitos módulos pretos, cuja expressão geral não pode ser determinada, assim como há alguns módulos induzidos. Assim, avaliar cada uma dos SIRs em cada tumor é essencial para compreender melhor as vias que podem estar sendo induzidas ou reprimidas em cada tecido.

O módulo "*activation of transmembrane receptor protein tyrosine kinase activity*" mostra uma regulação negativa dos receptores transmembranares da classe tirosina cinase, que são responsáveis por catalizar o início de diversas cascatas de sinalização celular. Essa classe de receptores está associada a regulação de processos chave para o desenvolvimento do câncer como a divisão e crescimento celular. Ainda existem um número considerável de tumores que apresentam regulação positiva desse módulo segundo o mapa modular, isso apenas reforça que a regulação de via tão importante para o início de diversas cascatas de sinalização tem uma regulação com alta especificidade tecidual/celular. Assim sendo, seria necessário avaliar as SIRs em cada tumor para tecer hipóteses sobre as vias que estão sendo alteradas no contexto do câncer.

O módulo "*positive regulation of type I interferon-mediated signaling pathway*" esta associado aos interferons de tipo I, uma classe de citocinas associadas a regulação do sistema imune. Segundo *Snell* e colaboradores (2017), essas citocinas e seus receptores são essenciais para o recrutamento de células imunes para vigilância do sistema imune (SNELL;

MCGAHA; BROOKS, 2017).

Esta classe de citocinas é considerada como um potencial alvo terapêutico desde os anos 80, mas os mecanismos específicos das vias de sinalização associadas a essa classe de citocinas não são bem conhecidos. Outro problema é que de acordo com a literatura, o papel dos interferons tipo I pode ser bem contexto especifico e frequentemente controverso.

Em um estudo, *Zhang* e colaboradores (2020), mostram como a regulação negativa de interferons do tipo I tem efeitos positivos no microambiente tumoral, evitando o recrutamento de células imunes e facilitando o desenvolvimento do tumor (ZHANG et al., 2020). Segundo *Bhattacharya* e colaboradores (2013), múltiplos tipos de câncer em estágio avançado apresentam uma considerável redução da expressão de interferons de tipo I em relação ao tecido normal, e isso é consequência da exposição gradual a múltiplas fontes de estresse. Isto deixa o tecido propício para a evasão dos mecanismos anti-câncer dos interferons tipo I, diretos (indução de apoptose) ou indiretos (recrutamento de células do sistema imune) (BHATTACHARYA et al., 2013). Essa baixa expressão dos interferons tipo I também deixa o tecido mais exposto a infecções virais, e esse fato pode ser utilizado para criação de novas estratégias de tratamento. *Zhang* e colaboradores (2010), mostram como a regulação negativa dos interferons tipo I pode ser aproveitada para um tratamento único de tumores que envolve a destruição de células tumorais mediante a aplicação controlada do vírus da estomatite vesicular (ZHANG et al., 2010).

Apesar das informações acima sugerirem que uma regulação positiva dos interferons tipo I pode ser invariavelmente positiva no que tange a cura do câncer, esse nem sempre pode ser o caso. *Budhwani* e colaboradores (2018), demonstram isso em seu estudo, mostrando que a ativação elevada ou crônica dos interferons tipo I induz uma poderosa resistência a diversos tipos de tratamentos para diversas patologias em um tecido (BUDHWANI; MAZZIERI; DOLCETTI, 2018).

4.1.2 Metilação: Módulos Consistentemente Induzidos ou Reprimidos

A Figura 4.5 abaixo ilustra a versão reduzida do mapa modular para 32 tumores com os dados de metilação:

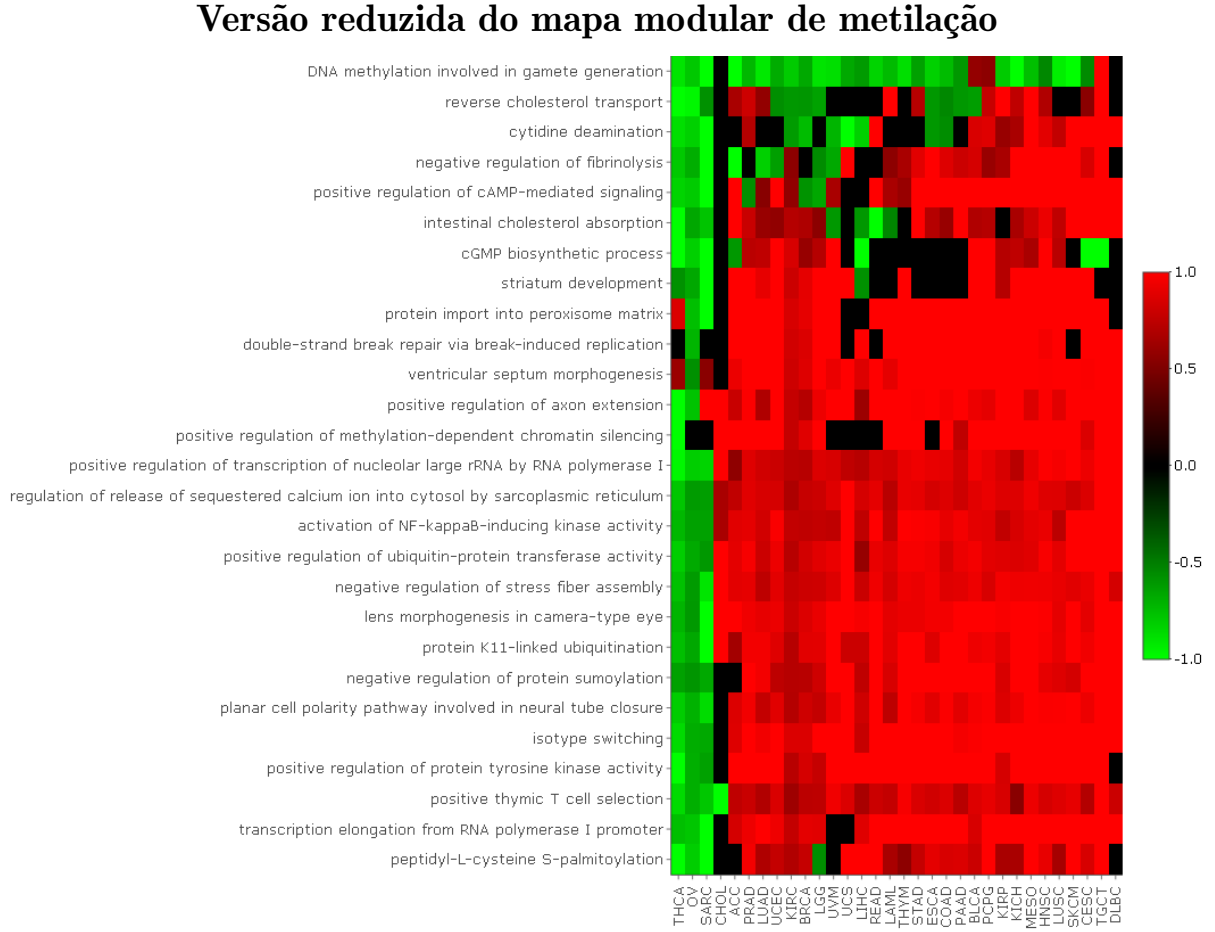


Figura 4.4: Versão reduzida do mapa modular de metilação. Essa versão do mapa modular foi desenvolvida com intenção de facilitar ainda mais a visualização do comportamento da metilação dos módulos selecionados. Ele é construído a partir dos 864 *heatmaps* menores que compõe o *heatmap* original, sendo que no eixo y tem-se os 27 módulos selecionados e no eixo x os tumores. O valor PEG, chamado de Porcentagem de Expressão Geral (PEG), de cada célula da figura é determinado por meio das porcentagens de elementos induzidos e reprimidos, a eles é assinalado o maior valor dos dois. Caso os valores não tenham uma diferença percentual maior que 0.1 (10%), então é assinalado o valor zero para o mini *heatmap*, indicando não haver diferença significativa entre as porcentagens de valores induzidos e reprimidos. Portanto os valores PEG estão localizados em uma escala de [-1,1], representada na legenda a direita do *heatmap*. Os valores PEG entre $-1 \leq PEG < 0$ (verde) indicam repressão, $PEG = 0$ (preto) indica ausência de expressão dominante e valores entre $0 < PEG \leq 1$ (vermelho) indicam indução.

A análise de metilação oferece um certo desafio em relação ao RNA-seq e miRseq, esses últimos nos permitem ter ideia da quantidade de transcritos presentes nas amostras, na metilação por si só isso não é possível. A metilação é um mecanismo de regulação epigenética que atua principalmente no DNA e nas histonas. Em teoria, a metilação elevada (hipermetilação) está associada ao silenciamento dos genes e, portanto, à diminuição de sua expressão. Analogamente a metilação reduzida está associada ao aumento da expressão gênica. No entanto, existem diversos mecanismos de regulação de expressão de genes que não envolvem a metilação. Sendo assim, apenas observar o estado de metilação dos genes não é a melhor maneira de avaliar a expressão dos genes, embora seja de considerável relevância

acessar o estado de metilação dos genes em diversas situações (PHILLIPS et al., 2008). Com isso é possível determinar perfis de metilação para condições diferentes, realizar predições de prognósticos e descobrir vias de regulação da expressão de genes chaves para determinada patologia. Sendo assim esta seção foi limitada a discutir o estado de metilação geral dos módulos e o significado de módulos de interesse, sem entrar nas especificidades de qual via pode estar sendo alterada, pois isso iria requerer análise intensa dos 32 tumores, o que foge ao propósito desta análise inicial. Uma análise minuciosa foi realizada posteriormente para tumor de mama *BRCA*, que pode ser encontrada na seção 4.2.2.

O mapa modular reduzido da metilação está intensamente ativado (em vermelho), mostrando que no contexto de câncer os módulos selecionados apresentam uma hipermetilação generalizada. Existem poucos módulos significantemente hipometilados:

1. *DNA methylation involved in gamete generation*
2. *reverse cholesterol transport*

O restante dos módulos estão majoritariamente hipermetilados ou com grande presença de elementos pretos, onde não foi possível determinar qual das metilações, hipometilação ou hipermetilação, é a dominante. Em geral, todos os tumores a partir do CHOL (da esquerda para direita) estão globalmente hipermetilados. Analogamente, 3 tumores estão quase que universalmente hipometilados, são eles: THCA (Thyroid carcinoma), OV (Ovarian Cancer) e SARC (Sarcoma).

Os dados sugerem que o módulo "*DNA methylation involved in gamete generation*" tende a estar hipometilado num contexto de câncer. É de conhecimento que existe uma reprogramação do perfil epigenético durante a formação de uma célula gamética (OKAE et al., 2014). Fundamentando-se nessas observações pode-se sugerir um possível estudo futuro onde é avaliado se as células tumorais estão se utilizando dessa via para realizar uma reprogramação epigenética que possa favorecer sua divisão, sobrevivência e metástase.

Seria necessário avaliar as SIRs de cada módulo e sua coexpressão (expressão em relação a metilação) para se discutir de maneira mais profunda o que pode estar acontecendo. De qualquer forma, a observação de que genes associados a reprogramação epigenética na geração de gametas estão amplamente hipometilados em células tumorais é algo notável,

uma vez que a hipometilação está, a princípio, associada a uma expressão elevada do gene alvo.

O módulo "*reverse cholesterol transport*" também está hipometilado no contexto de câncer. Tanto o mapa modular de RNA-seq quanto o de miRseq apresentaram módulos associados com a síntese e transporte de lipídios, e novamente foi observado um módulo com essa temática (lipídios/colesterol). Esse módulo está bem dividido entre hipermetilação, hipometilação e elementos pretos, dificultando uma generalização de sua expressão, isso pode indicar que os elementos contidos no módulo podem ser muito específicos para cada tumor.

O colesterol tem importância significativa em vários processos biológicos e recentemente cientistas têm estudado sua importância no processo tumoral. *Cruz* e colaboradores (2013), mencionam que em geral, espera-se que as células tumorais necessitem de maiores quantidades de colesterol, devido a seu metabolismo acelerado em relação as células normais. Segundo *Cruz* e colaboradores (2013), o colesterol é essencial na síntese das balsas lipídicas utilizadas na membrana celular e também é necessário para síntese de esteroides, e isso mostra como qualquer alteração na síntese, captura e efluxo de colesterol podem alterar muito o comportamento da célula.

As balsas lipídicas, em especial, são essenciais para regulação de diversas cascatas de sinalização mediadas por receptores, segundo *Cruz* e colaboradores (2013). Muitos dos quais podem regular processos de desenvolvimento celular e apoptose como o receptor "Fas" e o *TNF-related apoptosis-inducing ligand* (TRAIL) que, frequentemente, se encontram subexpressos em células tumorais por falta de balsas lipídicas, consequência da alteração na síntese e metabolismo do colesterol promovida pelo câncer. Segundo *Cruz* e colaboradores (2013), outra via afetada pela concentração de balsas lipídicas e a da cinase serina-treonina Akt, associada a sobrevivência e crescimento celular, sua ativação é dependente das balsas lipídicas (CRUZ et al., 2013).

No mapa modular reduzido da metilação, foi observado que o módulo trata especificamente do efluxo de colesterol (transporte reverso), que aparenta estar hipometilado em boa parte dos tecidos. Isso pode gerar a falta de colesterol nas membranas e portanto das balsas lipídicas. A depender do tecido e, portanto, dos receptores encontrados na membrana, isso pode ter efeitos tanto carcinogênicos como anti-carcinogênicos. Goossesns e colaboradores (2019), demonstram como o efluxo de colesterol é responsável pela depleção de balsas lipídicas em macrófagos associados ao tumor em ovários de ratas e como isso provoca uma

reprogramação dos macrófagos via interleucina 4 (IL-4), fazendo com que estas células deixem de exercer atividades anti tumorais e passem a apresentar um fenótipo pró-tumoral (GOOSSENS et al., 2019).

O módulo "*negative regulation of fibrinolysis*" também é importante, pois, assim como o colesterol, este é mencionado em outros mapas modulares, corroborando para a observação de que este é um processo chave nos 32 tumores, mesmo que por razões diferentes.

Como mencionado na Subseção 4.1.1, a fibrinólise é um processo de grande importância na carcinogênese, associado à adesão celular a hemóstase e a metástase tumoral. No mapa modular da metilação, esse módulo está mais hipermetilado do que hipometilado, é difícil dizer o que pode estar acontecendo sem analisar as SIRs em cada tumor, uma vez que a expressão elevada ou inibida das SIRs deste módulos têm efeitos diferentes a depender do receptor da membrana celular, ou seja, do tecido ao qual a célula pertence. *Mahmood* e colaboradores (2021), mencionam que um dos principais mecanismos que a metilação pode estar envolvida nesse contexto de fibrinólise é através da modulação da expressão do ativador de plasminogênio uroquinase (uPA). *Mahmood* e colaboradores (2021), demonstram que há uma relação inversa entre metilação do gene uPA e sua expressão. Tumores mais agressivos costumam apresentar hipometilação desse gene e portanto apresentam uma expressão elevada do mesmo (MAHMOOD; RABBANI, 2021). Assim, *Mahmood* e colaboradores (2021), concluem que acessar o estado de metilação do gene uPA pode constituir uma possível ferramenta de prognóstico do tumor, bem como apresentar um potencial alvo de tratamento.

Mahmood e colaboradores (2021), também mencionam que a expressão da proteína (MBD2) esta elevada na maioria dos tumores. Isso faz sentido, pois esta é uma proteína associada a regulação do perfil epigenético do DNA, sua principal função é identificar as marcas de metilação na fita de DNA. Em seu estudo, *Mahmood*, mostra que realizar o *knockdow* dessa proteína impedia a hipometilação do gene uPA e mantinha sua expressão baixa. Naturalmente não é possível afirmar que a expressão dessa proteína tem relevância nos mapas modulares sem primeiro checar as SIRs. No entanto, existe uma observação a ser feita, muitos dos módulos estão associados a regulação da metilação ou de mecanismos diretamente por ela regulado (como a transcrição). Alguns exemplos:

- *DNA methylation involved in gamete generation.*
- *positive regulation of methylation dependent chromatin silencing.*

- *positive regulation of transcription of nucleolar large rRNA by RNA polymerase I.*
- *transcription elongation from RNA polymerase I promoter.*

Com exceção do primeiro módulo, que foi discutido anteriormente, todos os outros módulos atuam na fase da transcrição, diretamente impactada pelo estado de metilação do DNA. Percebe-se que estes módulos, ao contrário do primeiro, estão majoritariamente hipermetilados, o que em teoria indica uma expressão menos elevada dos elementos (genes/SIRs) contidos nesses módulos.

A hipermetilação do módulo *positive regulation of methylation dependent chromatin silencing* é algo para se destacar, pois, o silenciamento de genes pela cromatina é um dos principais mecanismos de inibição de transcrição utilizados pela metilação, isso indica que, no contexto de câncer, muitos genes antes silenciados passam a ser expressos, causando um aumento da atividade transcricional.

A partir desses resultados, pode-se assim deixar uma outra sugestão de um futuro estudo, onde é avaliado se a hipermetilação observada nos módulos "*positive regulation of transcription of nucleolar large rRNA by RNA polymerase I*" e "*transcription elongation from RNA polymerase I promoter*" pode estar associada a genes reguladores do processo de transcrição e assim diminuindo a expressão de genes reguladores transcricionais e aumentando a atividade transcricional. De fato, de acordo com mapa modular do RNA-seq na Subseção 4.1.1, pode-se observar módulos associados ao aumento da atividade transcricional positivamente regulados, bem como módulos associados ao aumenta de transcrição de isoformas.

Pode-se identificar outros módulos dentre os hipermetilados, como por exemplo:

- *double-strand break repair via break induced replication.*
- *ventricular septum morphogenesis .*
- *striatum development.*
- *negative regulation of stress fiber assembly.*
- *positive regulation of protein tyrosine kinase activity.*

O módulo *double-strand break repair via break induced replication* está associado a mecanismos de reparação do DNA, e se encontra hipermetilado na maior parte dos tumores. Além do reparo da molécula de DNA, essa via contém genes capazes de dar início a um processo de apoptose mediante falha da reparação da molécula original. Células tumorais, em geral, apresentam uma elevada resistência a ativação de vias apoptóticas. As SIRs presentes nesse módulo podem fornecer informações sobre quais vias podem estar sendo reprimidas em cada tecido (HELLEDAY et al., 2007).

Os módulos *ventricular septum morphogenesis* e *striatum development* estão associados a mecanismos de desenvolvimento celular e tecidual, algo que está altamente alterado em células tumorais, módulos com a mesma temática podem ser encontrados nos mapas modulares de RNA-seq e miRseq. Tais módulos estão, em geral, positivamente regulados e isso é uma observação esperada de células tumorais, que possuem elevada taxa de divisão em relação às células sadias. No mapa modular de metilação pode-se observar uma hipermetilação desses dois módulos. Aqui, pode-se sugerir um outro estudo futuro para avaliar se isso trata da hipermetilação de genes reguladores de etapas da divisão celular ou de processos associados a essa via.

O módulo *negative regulation of stress fiber assembly* está associado ao módulo *negative regulation of fibrinolysis*, que foi discutido ainda nesta seção. Ambos esses módulos estão associados a adesão celular, hemostase, migração e metástase. Uma vez que ambos estão hipermetilados, pode-se presumir, a princípio, que esses módulos contêm genes/SIRs associados à regulação negativa de tais processos, como foi mencionado várias vezes nos resultados. No entanto, avaliar essas SIRs para 32 tumores é algo demasiadamente trabalhoso que foge ao propósito dessa análise geral, posteriormente esse procedimento foi feito para o tumor BRCA na Subseção 4.2.2.

Por fim, o módulo *positive regulation of protein tyrosine kinase activity*, mostra como SIRs associadas com a atividade de tirosina cinases estão altamente hipermetiladas nos tumores. Os receptores de tirosina cinase são conhecidos por estarem associados a vias de regulação do processo de divisão celular, essencial para progressão do tumor. A temática da atividade das proteínas tirosina cinase também aparece mapa modular de RNA-seq, no qual aparentam possuir expressão reprimida na maior parte dos tumores. Isso faz sentido considerando a hipermetilação do módulo referente a regulação positiva da atividade de proteínas tirosina cinase, mas é difícil dizer quais vias especificamente apresentam essa

metilação aberrante sem olhar as SIRs em cada tumor.

Finalizando a discussão inicial, segundo o mapa modular de metilação, as células tumorais apresentam hipermetilação aberrante de diversos processos chaves para progressão do tumor e hipometilação de genes associados com a reprogramação epigenética.

4.1.3 miRseq: Módulos Consistentemente Induzidos ou Reprimidos

A Figura 4.5 abaixo ilustra a versão reduzida do mapa modular para 32 tumores com os dados miRseq:

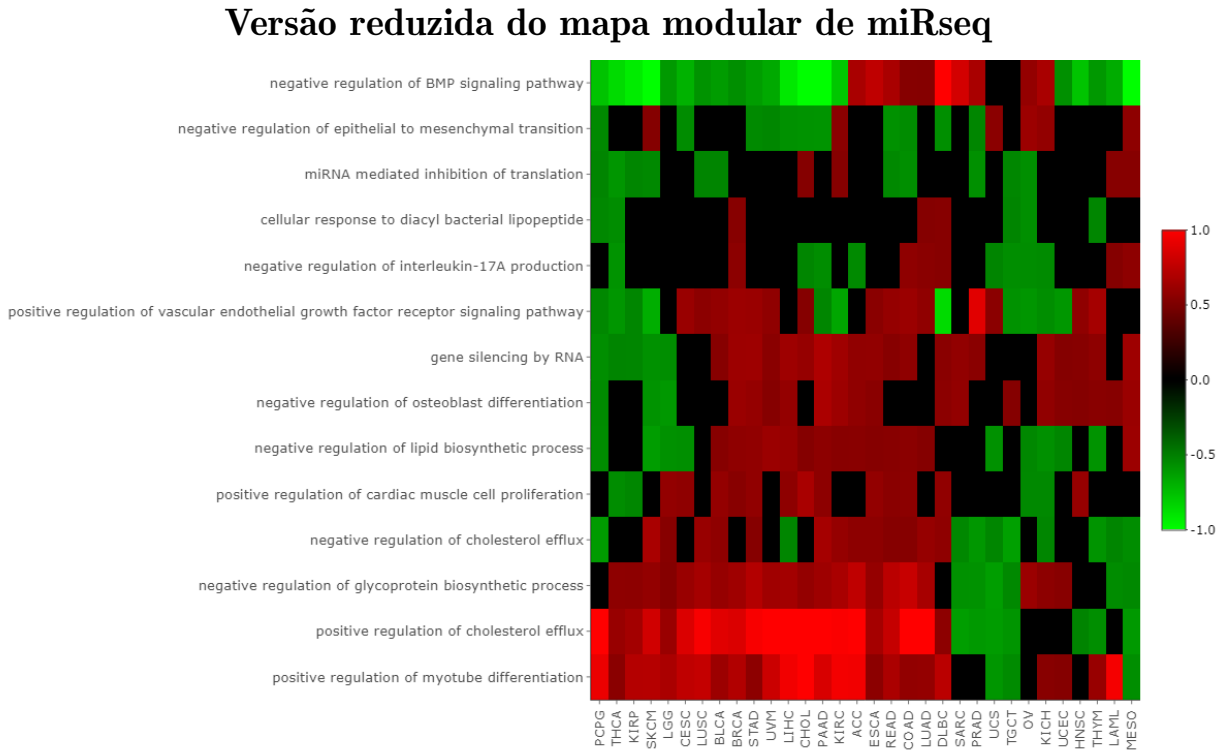


Figura 4.5: Versão reduzida do mapa modular de miRseq. Essa versão do mapa modular foi desenvolvida com intenção de facilitar a visualização da expressão dos módulos selecionados. Ele é construído a partir dos 448 *heatmaps* menores que compõe o *heatmap* original, sendo que no eixo y tem-se os 14 módulos selecionados e no eixo x os tumores. O valor PEG, aqui chamado de Porcentagem de Expressão Geral (PEG), de cada célula da figura é determinado por meio das porcentagens de elementos induzidos e reprimidos, a eles é assinalado o maior valor dos dois. Caso os valores não tenham uma diferença percentual maior que 0.1 (10%), então é assinalado o valor zero para o mini *heatmap*, indicando não haver diferença significativa entre as porcentagens de valores induzidos e reprimidos. Portanto os valores PEG estão localizados em uma escala de [-1,1], representada na legenda a direita do *heatmap*. Os valores PEG eentre $-1 \leq PEG < 0$ (verde) indicam repressão, $PEG = 0$ (preto) indica ausência de expressão dominante e valores entre $0 < PEG \leq 1$ (vermelho) indicam indução.

Existem poucos módulos encontrados para os dados de miRseq, apenas 15, em contrapartida aos mais de 20 no RNA-seq e na metilação. Isso é esperado pois a quantidade de dados disponível para os miRNAs é bem menor em comparação aos dados para mRNAs

e ilhas CPG. A primeira característica notável da Figura 4.5 é a existência de um elevado número de elementos nulos (pretos), indicando que não é possível determinar uma PEG para os SIRs contidos dentro do par módulo-Tumor, tornando necessária uma avaliação interna do elemento em questão para retirar conclusões significativas. Outra observação oportuna é que mesmo os módulos mais consistentemente induzidos ou reprimidos possuem em sí um considerável número de tumores cuja PEG destoa da maioria, como no módulo *negative regulation of BMP signaling pathway*. Essas observações contribuem para uma conclusão esperada: a expressão dos miRNAs é algo altamente tecido específica e mesmo dentro de um tecido a expressão dos miRNAs pode variar de tal maneira que é impossível determinar um tipo de "perfil geral de expressão".

Apesar disso, o mapa modular reduzido do miRseq nos permite realizar algumas observações no que diz respeito aos módulos consistentemente inibidos ou reprimidos. Abaixo foram listados alguns módulos de interesse que estão consistentemente induzidos no mapa modular:

- *negative regulation of cholesterol efflux*.
- *negative regulation of glycoprotein biosynthetic process*.
- *positive regulation of cholesterol efflux*.
- *positive regulation of myotube differentiation*.

A primeira observação de interesse é a existência dos módulos *negative regulation of cholesterol efflux* e *positive regulation of cholesterol efflux*. São módulos opostos, e fariam sentido se todos os 32 tumores apresentassem PEGs opostas nesses dois, como no caso do hepatocarcinoma (LIHC). Entretanto, esse não é o caso da maioria dos tumores, novamente, uma análise minuciosa dos SIRs em cada par módulo-Tumor se faz necessária para maior compreensão.

De qualquer forma, a expressão aberrante de genes associados a homeostase do colesterol é documentada em diversas células tumorais, embora a exata função do mesmo ainda permaneça desconhecida. *Kuzu* e colaboradores (2016) demonstram, com dados do próprio TCGA, que elevados níveis de síntese de colesterol estão associados a prognósticos ruins e a baixa sobrevivência em diversos tumores do TCGA, e que a expressão de mRNAs associados

a homeostase do colesterol variam muito entre tecidos diferentes. Algo que os resultados do mapa modular reduzido do miRseq demonstra também (KUZU; NOORY; ROBERTSON, 2016). Embora a exata função do colesterol ainda seja desconhecida, e também possa variar de tecido a tecido, tem-se evidências de que a homeostase do colesterol tem papel fundamental na manutenção do processo tumoral.

O mapa modular demonstra o efluxo de colesterol como processo chave, células normais não conseguem degradar essa molécula e seu acúmulo pode levar uma célula a morte por estresse metabólico. *Liu* e colaboradores (2021) publicaram, na *Nature Communications* em 2021, uma pesquisa demonstrando como células do câncer de mama (ER negativas) de ratos que sobreviviam ao acúmulo de colesterol podiam adquirir resistência a via da ferroptose. Assim, essas células resistentes ao estresse metabólico se reproduzem rapidamente devido ao seu fenótipo tumoral e rapidamente dão início aos processos metastáticos. *Liu* e colaboradores (2021), também mencionam que esse é um fenômeno possível de ocorrer em outros tecidos e pode constituir um importante alvo terapêutico (LIU et al., 2021).

O colesterol também é um importante componente das balsas lipídicas, constituintes da membrana celular juntamente com glicoproteínas e os fosfolipídeos. *Kuzu* e colaboradores (2016), mencionam brevemente que as balsas lipídicas são um importante meio de transdução de sinais intracelulares, e como tal existe uma infinidade de mecanismos pelos quais o processo tumoral pode se beneficiar (KUZU; NOORY; ROBERTSON, 2016). *Yan* e colaboradores (2020), discutem vários deles em seu artigo e demonstra como esses podem conferir resistência a drogas para as células tumorais. Um deles é justamente pela regulação das balsas lipídicas, as quais modulam sinais do EGFR (*Epithelial Growth Factor Receptor*) e da caveolina-1, o que pode levar a célula tumoral a entrar em um processo de transição epitelial-mesenquimal ou lhe conferir um fenótipo de células não diferenciadas (YAN et al., 2020).

De fato, o efluxo de colesterol como mecanismo de aquisição de resistência a drogas é algo explorado há algumas décadas. *Liscovitch* e colaboradores, ainda no ano de 2000, discutiam justamente como o acumulo e efluxo de colesterol, juntamente com a caveolina-1 podem conferir um fenótipo de resistência a drogas (LISCOVITCH; LAVIE, 2000).

Yan e colaboradores (2020), também mencionam como a elevação dos níveis de colesterol altera composição das balsas lipídicas também muda, fatalmente, a composição da membrana celular. Assim, não apenas a transdução de sinais intracelulares mediadas pelas

balsas lipídicas são afetadas por essa mudança, mas a permeabilidade geral da membrana celular também. Isso pode ter efeitos imprevisíveis na captação celular de drogas e de outros agentes podendo constituir uma outra fonte de resistência a drogas (YAN et al., 2020).

O módulo *negative regulation of glycoprotein biosynthetic process*, se encontra majoritariamente induzido (em vermelho) em diversos tumores na Figura 4.5. As glicoproteínas são um dos principais constituintes da membrana celular, sendo responsáveis tanto por transdução de sinais celulares como por funções estruturais (como por exemplo: transporte de moléculas). Uma vez que este módulo está consistentemente induzido nos tumores, pode-se supor uma regulação positiva do processo de biossíntese de glicoproteínas, embora alguns tumores apresentem a repressão do mesmo módulo (como o LAML), novamente relembrando a natureza tecido-específica da expressão dos miRNAs. É difícil tecer hipóteses sem checar primeiro quais miRNAs estão envolvidos, pois existe uma infinidade de glicoproteínas e de mecanismos mediadas pelas mesmas. Sendo assim, esse módulo será analisado exclusivamente no tumor principal (BRCA) mais adiante.

Abaixo estão os módulos consistentemente inibidos no mapa modular do miRseq, esses incluem:

- *negative regulation of BMP signaling pathway.*
- *negative regulation of epithelial to mesenchymal transition.*
- *miRNA mediated inhibition of translation.*

O módulo *negative regulation of BMP signaling pathway* está consistentemente inibido na maioria dos tumores (20), induzido em alguns outros (10) e em apenas dois sua expressão não esta clara. As *BMPs* (*Bone Morphogenetic Proteins*) são um grupo de mais de 20 proteínas que pertencem a família do fator de crescimento $\boldsymbol{\beta}$ ($\boldsymbol{TGF-\beta}$). As *BMPs* são citocinas de sinalização extracelular multifuncionais e muitos estudos demonstram a importância no desenvolvimento de tecidos na fase embrionária e pós-natal. Recentemente, essas proteínas têm recebido atenção devido ao seu potencial envolvimento com surgimento e manutenção de processos cancerígenos, especialmente na última década, onde os transcritos não codificantes abrem uma camada de regulação nova pela qual diversos mecanismos moleculares podem operar (BACH; PARK; LEE, 2018).

Estudos mostram que as vias de sinalização das *BMPs* estão associadas a agressividade de tumores primários e o desenvolvimento de processos metastáticos através de mecanismos como angiogênese, transição epitelial-mesenquimal (EMT) e aquisição do fenótipo de células não diferenciadas. Muitos dos sinais celulares nascidos da interação das *BMPs* com seus receptores e antagonistas ainda são desconhecidos, principalmente, em uma especificidade em nível de tecido. *Bach* e colaboradores (2018), discutem essa questão, demonstrando o paradoxal papel que as *BMPs* parecem ter em diversos tipos de câncer, ora agindo como indutores ou como supressores do processo tumoral (BACH; PARK; LEE, 2018).

Nossos dados, de fato, corroboram para uma expressão com alta especificidade, essa especificidade é esperada de transcritos como os miRNAs, os quais têm parte na regulação da expressão das *BMPs*. Analisar a expressão dos SIRs dentro de cada um dos módulos seria o próximo passo para melhor entender quais vias de sinalização estão sendo alteradas no contexto de cada tipo de tecido. Isso foi realizado tal análise para o tumor BRCA mais adiante.

O módulo *negative regulation of epithelial to mesenchymal transition* também está consistentemente inibido, embora apresente um número bem maior de elementos pretos, onde a expressão das SIRs é muito variável. A transição epitelial-mesenquimal (TEM) é um fenômeno que faz células epiteliais adquirirem fenótipo de células mesenquimais, e ocorre de maneira natural e regulada durante o desenvolvimento tecidual na fase embrionária e durante processos de regeneração na fase adulta. *Ribatti* e colaboradores (2020), fizeram uma revisão histórica da TEM em câncer, demonstrando como esse processo ocorre de maneira desregulada em células cancerígenas e como está desregulação esta positivamente correlacionada ao início da metástase (RIBATTI; TAMMA; ANNESE, 2020).

Pode-se observar que tanto o módulo *negative regulation of BMP signaling pathway* quanto o módulo *negative regulation of epithelial to mesenchymal transition* estão ambos consistentemente inibidos, segundo os dados de miRseq. Como foi mencionado as *BMPs* estão associadas a regulação de diversos processos importantes para desenvolvimento do câncer, um deles é justamente a TEM. Existem vários estudos que exploram os mecanismos pelos quais as BMPs podem modular a TEM. Choi e colaboradores (2019), discutem como BMP-4 pode realizar esse processo, utilizando modelos xenoenxertos de câncer mamário através da bem conhecida via de sinalização *notch* (CHOI et al., 2019). *Huang* e colaboradores (2017), demonstram como a BMP-2 pode modular esse processo em xenoenxertos

mamários através da via de sinalização AKT juntamente com os genes Rb e CD44 (HUANG et al., 2017). Uma análise minuciosa das SIRs no tumor BRCA é necessária para se tirar conclusões significativas, tal análise será feita mais adiante.

Por fim, pode-se observar o módulo *miRNA mediated inhibition of translation* como consistentemente inibido, mas com elevado número de elementos pretos, uma situação similar a do módulo *negative regulation of epithelial to mesenchymal transition*. Esse é um módulo muito geral, e uma análise de especificidade tumoral é necessária para identificar as vias envolvidas. O que pode ser afirmado é que miRNAs geralmente atuam justamente no silenciamento da expressão de genes, portanto, desregulação dos transcritos contidos nesse módulo têm papel chave na modulação de qualquer processo dependente da fase de tradução de proteínas.

4.1.4 Sumarização da Visualização

Para finalizar a análise inicial (visual) dos mapas modulares, foi realizada uma listagem dos processos biológicos que foram considerados chave segundo os 3 mapas modulares, dando especial atenção àqueles cuja temática apareceu em mais de um mapa:

- síntese e transporte de lipídios;
- regulação da atividade transcricional (silenciamento gênico e isoformas);
- diferenciação e desenvolvimento tecidual;
- Transição epitelial-mesenquimal;
- regulação da fibrinólise;
- homeostase de íons;
- sinalização mediada por proteínas tirosina cinase;
- regulação de receptores das células do sistema imune;
- regulação da metilação.

Note que os itens acima se tratam de vias biológicas generalizadas, e não de módulos específicos definidos para cada mapa modular, pois existe um número muito pequeno de

módulos em comum entre os mapas. No entanto, existe uma convergência no que tange a natureza das vias alteradas, e tal convergência é observada nos itens citados acima. Isso essencialmente cumpre o primeiro objetivo da metodologia: o de encontrar vias significativamente alteradas simultaneamente em diversos tumores.

4.2 Tumor de mama (BRCA): Um Estudo de Caso

Esta é a etapa final da análise, onde foram reutilizados os dados gerados com a construção dos mapas modulares. Na etapa de visualização, foi discutido de maneira superficial, o comportamento dos módulos em cada um dos 3 mapas e foi possível derivar *insights* importantes sobre vias que aparentam ser processos chaves para células tumorais e que, frequentemente, estavam positiva ou negativamente regulados. Por meio dos resultados, foram sugeridos temas para possíveis estudos futuros, usando a literatura científica como base, mas as SIRs de cada módulo em cada tumor nunca foram visualizadas. Fazer uma análise minuciosa das SIRs é de grande importância para avaliação de vias alteradas e determinação de hipóteses, mas realizar esse procedimento para todos os 32 tumores é algo excessivamente extenso que foge ao objetivo principal da tese: identificar módulos consistentemente alterados segundo os dados de 32 tipos de tumores. Esse objetivo foi concluído na análise visual, mas para demonstrar que esses valiosos *insights* ainda podem ser utilizados para extrair mais informações, foi analisado também o comportamento das SIRs para o tumor mamário (BRCA), pois é o tumor com maior quantidade de amostras e com ampla literatura científica. Pretende-se analisar a expressão das SIRs dos 3 mapas modulares e construir redes de expressão gênica para determinar as vias alteradas como foi feito na análise visual, com a diferença de que a expressão das SIRs agora será levada em consideração.

Usaremos, para cada um dos 3 mapas modulares, as seguintes etapas de análise:

I seleção dos módulos de interesse;

II determinação da significância estatística das sobreposições *overlaps*;

III breve descrição das SIRs selecionadas em II;

IV análise de redes com as SIRS detectadas em II.

O item I é auto explanatório, é realizada uma seleção manual de módulos presentes nos mapas modulares apresentados. Essa seleção foi baseada nas vias definidas como importantes para a maior parte dos tumores, tais vias podem ser observadas na Subseção 4.1.4 e, naturalmente, isso resultará em um conjunto de módulos selecionados diferentes para cada um dos mapas. Em II, foi determinada a existência de elementos com presença significantemente estatística no conjunto de módulos selecionados no item II. Isso foi feito computando p-valores para todos os elementos com sobreposição (*overlaps*) entre módulos diferentes utilizando um teste de reamostragem com reposição (*bootstrap*) com 10000 passos, os elementos que apresentarem p-valor inferior a 0.01 serão considerados significantes para a análise de redes no item IV.

No item III, foram determinadas as SIRs mais positiva e negativamente reguladas para os módulos selecionados dentro do conjunto de dados do BRCA. Para fazer isso, foram selecionadas todas as sondas cuja expressão é maior que 3 vezes o desvio padrão da amostra, e então, se a expressão geral se encontra majoritariamente positiva ou negativamente regulada seguindo os seguintes passos:

I determinar a frequência com que cada sonda está positiva ou negativamente regulada no conjunto de dados do BRCA;

II determinar a diferença entre frequência positiva e negativa;

III filtrar SIRs cujo maior número de frequência for menor que 10% do valor de frequência máximo encontrado;

IV filtrar SIRs cuja diferença entre frequência positiva e negativa seja menor que 40% do valor máximo entre ambas;

Ao final, foram construídas duas tabelas, uma para sondas positivamente reguladas e outra para sondas negativamente reguladas. Assim, foram selecionadas as SIRs mais pertinentes para análise de redes de acordo com seu perfil de expressão no conjunto de dados do BRCA. Após a seleção das sondas de interesse de acordo com os itens II e III, foi utilizado o banco de dados *string* para montar redes de interação proteína-proteína (PPI). Tanto o *string* quanto o protocolo para geração de redes são descritos na Subseção 3.3.4.1.

4.2.1 BRCA: Análise do RNA-seq

Essa seção é iniciada com a lista dos módulos selecionados para análise de *overlaps* na Tabela 4.1 abaixo:

RNA-seq: Módulos Selecionados

Tabela 4.1: Módulos selecionados para análise de *overlaps* na primeira coluna (coluna módulo). Na coluna "Ontologias" tem-se o número de ontologias contidas dentro do módulo e na "Sondas" o número de sondas/genes.

	Módulo	Ontologias	Sondas
1	*activation of transmembrane receptor protein tyrosine kinase activity*	5	86
2	*cellular sodium ion homeostasis*	2	21
3	*negative regulation of fibrinolysis*	13	218
4	*nuclear-transcribed mRNA poly(A) tail shortening*	2	46
5	*positive regulation of epidermal growth factor-activated receptor activity*	12	299
6	*positive regulation of phosphoprotein phosphatase activity*	2	39
7	*positive regulation of protein serine/threonine kinase activity*	2	87
8	*positive regulation of type I interferon-mediated signaling pathway*	2	25
9	*regulation of B cell receptor signaling pathway*	6	157
10	*regulation of DNA methylation*	2	34
11	*regulation of fatty acid biosynthetic process*	12	265
12	*ventricular septum development*	9	255

A Tabela 4.2 abaixo mostra as sondas/genes que foram selecionados de acordo com o teste de *bootstrap*.

RNA-seq: *Overlaps* estatisticamente significantes

Tabela 4.2: *Overlaps* nos dados de RNA-seq. A tabela mostra os genes estatisticamente significantes de acordo com um teste de reamostragem cm reposição de 10000 passos. Na primeira coluna, estão os genes (*Symbol*), na segunda, o número de *overlaps* e na terceira tem-se o p-valor computado.

	Genes	Freq	P-valor
1	RHOA	4	0.0000
2	HSP90AB1	3	0.0001
3	AGT	4	0.0002
4	DAB2IP	5	0.0013
5	GRHL2	4	0.0023

O gene RHOA (*Ras Homolog Family Member A*) codifica uma GTPase pequena. Estudos apontam que essa molécula tem importância em diversas cascatas de sinalização celular. De acordo com a literatura, essa proteína está associada a processos envolvendo atividade

do citoesqueleto, regulação da forma celular, adesão celular e motilidade (MCBEATH et al., 2004). No câncer, a super expressão dessa proteína esta associada à proliferação celular e metástase (LI et al., 2012).

A HSP90AB1 (*Heat shock protein HSP 90-beta*) é uma chaperona molecular, cuja função é se ligar a diversas proteínas para estabilizá-las ou degradá-las de uma maneira ATP-dependente (HAASE; FITZE, 2016). Essa proteína também tem importância na regulação da maquinaria de transcrição celular: modulando níveis de fatores de transcrição, modulando modificadores epigenéticos como DNA metiltransferases e deacetilases de histonas, e também pela ativação de genes através da remoção de histonas das regiões promotoras (KHURANA; BHATTACHARYYA, 2015). Devido a todas essas funções, a HSP90AB1 está presente em uma elevada gama de vias de sinalização, além disso, estudos mostram que, no câncer, essa proteína tem importância através da estabilização de proteínas aberrantes, tendo impacto direto na sobrevivência e desenvolvimento de células tumorais (CHENG et al., 2012).

O gene AGT (*Angiotensinogen*) codifica um precursor do angiotensinogênio, o qual pode ser clivado pela enzima ACE (*Angiotensinogen Converting Enzyme*) para originar a enzima fisiologicamente ativa, angiotensina II (KUMAR; SINGH; BAKER, 2007). O AGT está associado à manutenção da pressão sanguínea e homeostase iônica (KUMAR; SINGH; BAKER, 2007).

O DAB2IP *Disabled homolog 2-interacting protein* é GTPase de ativação de diversas proteínas da família Ras (*Rat Sarcoma*). Por essa razão, o DAB2IP está associado a modulação de diversas vias chave no contexto de câncer: resposta imune, inflamação, crescimento celular, apoptose, sobrevivência celular, angiogênese e migração celular (BELLAZZO; MININ; COLLAVIN, 2017).

Com esses 5 genes determinados (Tabela 4.2), foi construída uma rede PPI via *String* segundo o protocolo descrito na Subseção 3.3.4.1. Abaixo, existem duas figuras contendo essa mesma rede PPI, cada uma tem uma escala de cores com valores de “Fator de Indução/Repressão”. Os valores utilizados para criar a escala de cores são determinados ao ajustar o conjunto de medianas das porções positiva/negativa dos valores de “Expressão Relativa” dos genes presentes na rede entre os valores 1 e 0. Estes novos valores são denominados: “Fator de Indução” para a porção positiva de “Expressão Relativa” e “Fator de Repressão” para a porção negativa. Um gradiente entre 1 e 0 é criado utilizando 4 cores, o que facilita a identificação do módulo do “Fator de Indução/Repressão”. O gradiente é

formado por: 0 (preto), 0.25 (amarelo), 0.75 (laranja) e 1 (vermelho). Note que o tamanho dos nodos da rede é ajustado pelo parâmetro *betweenness*, nodos maiores possuem um valor de *betweenness* maior.

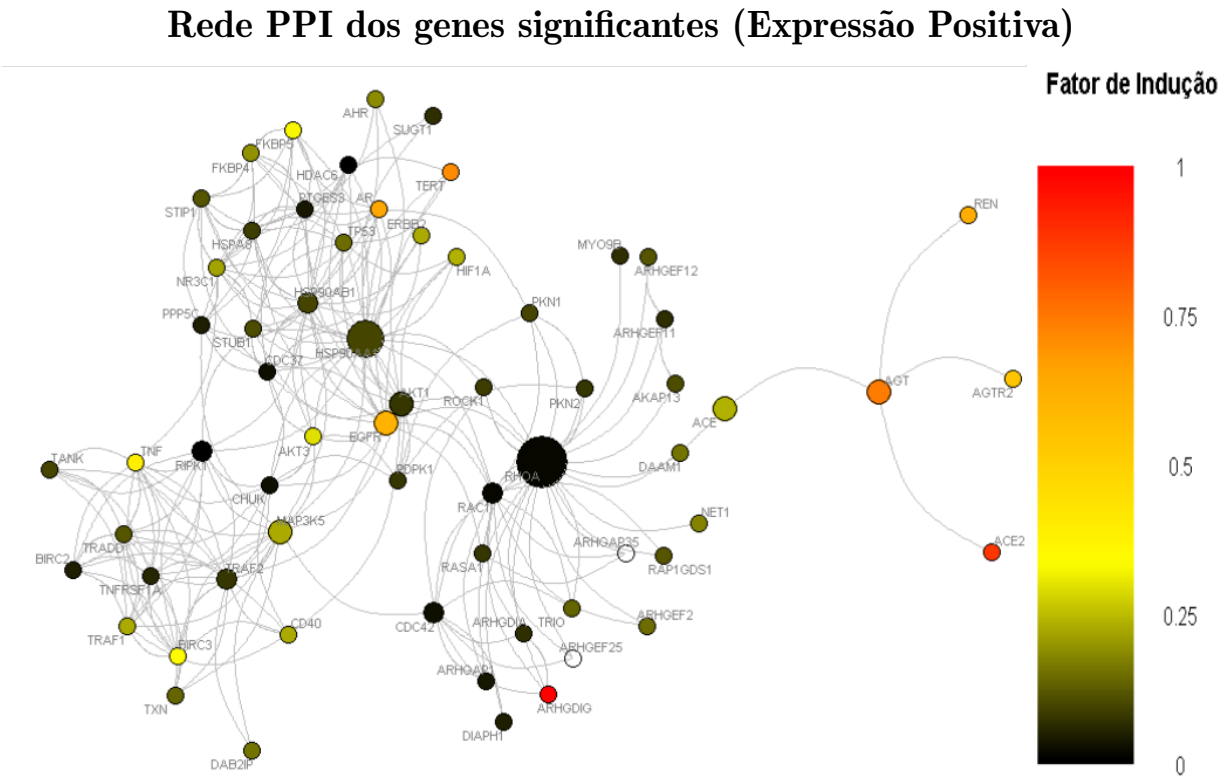


Figura 4.6: Rede de Expressão Gênica Induzida do RNA-seq. A rede é obtida ao utilizar os 5 genes descritos na Tabela 4.2 no banco de dados *string*, seguindo o protocolo descrito na Subseção 3.3.4.1. A escala de cores, aqui denominada "Fator de Indução", é construída a partir do conjunto de medianas das expressões relativas (ER) positivas dos genes presentes na rede, considerando apenas o conjunto de dados do BRCA, a definição de ER é fornecida na Subseção 3.2.6. Assim, a escala reflete a magnitude da expressão positiva do conjunto, onde 1 (vermelho) representa módulo elevado e 0 (preto) módulo baixo. O gradiente de cores entre esses valores tem a função de facilitar a comparação visual com sua análoga para expressão negativa a seguir. Note que o tamanho dos nodos é ajustado pelo parâmetro *betweenness* do nodo.

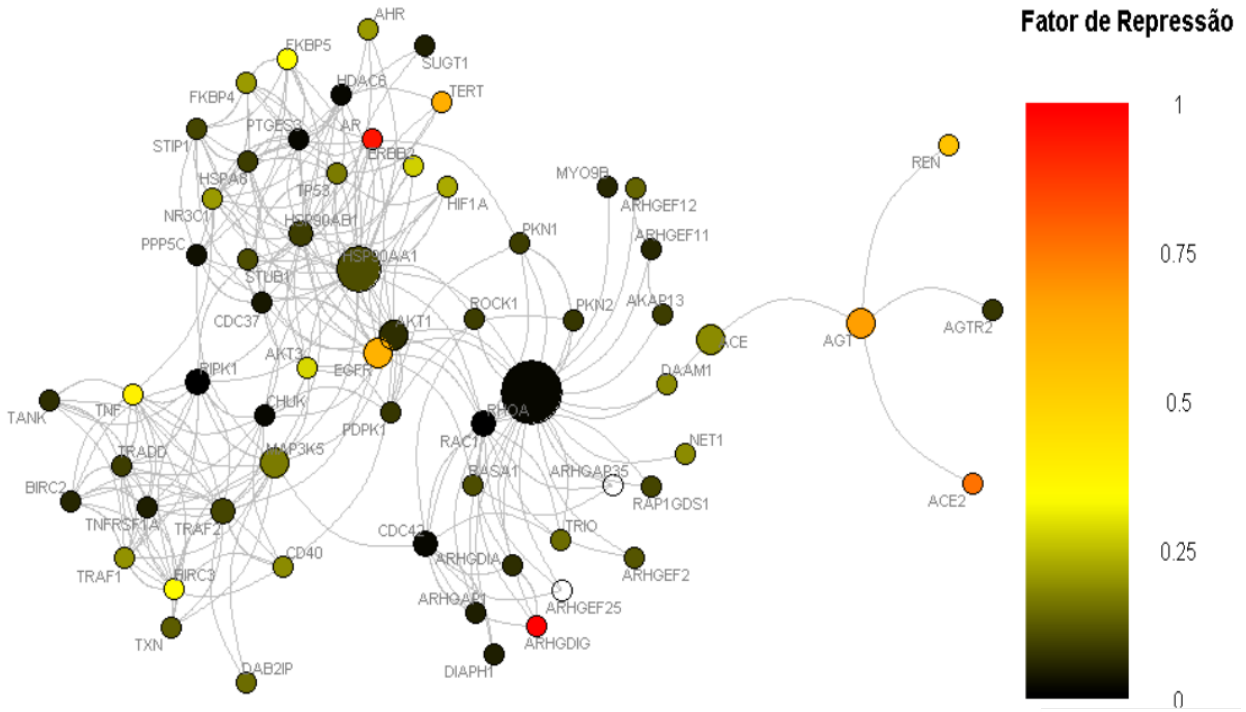


Figura 4.7: Rede de Expressão Gênica Reprimida do RNA-seq. A rede é obtida ao utilizar os 5 genes descritos na Tabela 4.2 no banco de dados *string*, seguindo o protocolo descrito na Subseção 3.3.4.1. A escala de cores, aqui denominada "Fator de Repressão", é construída a partir do conjunto de medianas das expressões relativas (ER) negativas dos genes presentes na rede, considerando apenas o conjunto de dados do BRCA, a definição de ER é fornecida na Subseção 3.2.6. Assim, a escala reflete a magnitude da expressão positiva do conjunto, onde 1 (vermelho) representa módulo elevado e 0 (preto) módulo baixo. O gradiente de cores entre esses valores tem a função de facilitar a comparação visual com sua análoga para expressão positiva anterior. Note que o tamanho dos nodos é ajustado pelo parâmetro *betweenness* do nodo.

Antes de analisar as redes das Figuras 4.6 e 4.7 é oportuno relembrar a definição da "Fator de Indução/Repressão". Por "Fator de Indução/Repressão", se entende como o módulo da mediana da porção positiva/negativa dos valores de "Expressão Relativa" da sonda/gene entre as amostras de tumor mamário. A definição de "Expressão Relativa" é fornecida nas Subseções 3.2.6, e corresponde a porcentagem de SIRs existentes dentro das amostras, nesse caso, SIRs correspondentes aos dados de RNA-seq do tumor mamário. Sendo que uma "Expressão Relativa" positiva indica indução e uma negativa indica repressão. Assim sendo, "Fator de Indução/Repressão" corresponde a um vetor contendo em sí os valores das medianas das porções positivas ou negativas das sondas/genes, esse vetor é ajustado para uma escala de 0 a 1, criando assim, os valores de "Fator de Indução/Repressão". Na Figura 4.6 os valores de "Fator de Indução" correspondem as medianas da porção positiva (maior que 0) dos valores de "Expressão Relativa" para cada gene/sonda, tais valores são descritos mais afrente na tabela 4.3. Na Figura 4.7 os valores de "Fator de Repressão" correspondem as

medianas da porção positiva (menor que 0) dos valores de "Expressão Relativa" para cada gene/sonda, tais valores são descritos mais afrente na tabela 4.4.

A primeira observação notável que pode ser feita em relação as Figuras 4.6 e 4.7 é a existência de um conjunto de genes que apresentam um módulo de expressão elevado de expressão em ambas figuras: ACE2, AGT, AR, ARHGDIG, EGFR, REN e TERT . Isso denota que, no contexto do câncer, esses genes consistentemente apresentam um elevado módulo de expressão. Naturalmente, essa mesma observação sugere que tais genes tem uma expressão contexto específica, pois, podem estar tanto positiva quanto negativamente regulado. Para observar melhor esse comportamento, observe as tabelas abaixo:

RNA-seq: Top 10 Induzidos/Reprimidos

Tabela 4.3: Top 10 sondas induzidas para os dados de RNA-seq. Na tabela temos as sondas da rede na Figura 4.6. Estas sondas estão organizadas de maneira decrescente de acordo com os valores da coluna "Fator de Indução", a qual contém o valor, normalizado em uma escala de 0 a 1, do módulo da mediana da porção positiva de valores de "Expressão Relativa" da sonda segundo os dados de RNA-seq do tumor mamário (BRCA). Lembrando que a definição de "Expressão Relativa" é apresentada na Subseção 3.2.6. Assim sendo, na tabela, estão os 10 genes/sondas com maior indução, segundo a rede da Figura 4.6.

Genes	Fator de Indução
ARHGDIG	1.00
ACE2	0.88
AGT	0.75
TERT	0.72
AR	0.66
REN	0.63
EGFR	0.62
AGTR2	0.54
TNF	0.40
BIRC3	0.34

Tabela 4.4: Top 10 sondas reprimidas para os dados de RNA-seq. Na tabela temos as sondas da rede na Figura 4.7. Estas sondas estão organizadas de maneira decrescente de acordo com os valores da coluna "Fator de Repressão", a qual contém o valor, normalizado em uma escala de 0 a 1, do módulo da mediana da porção negativa de valores de "Expressão Relativa" da sonda segundo os dados de RNA-seq do tumor mamário (BRCA). Lembrando que a definição de "Expressão Relativa" é apresentada na Subseção 3.2.6. Assim sendo, na tabela, estão os 10 genes/sondas com maior repressão, segundo a rede da Figura 4.7.

Genes	Fator de Repressão
ARHGDIG	1.00
AR	0.95
ACE2	0.76
AGT	0.68
TERT	0.62
EGFR	0.62
REN	0.55
TNF	0.38
BIRC3	0.35
FKBP5	0.34

Pode-se observar nas Tabelas 4.3 e 4.4, a razão do comportamento visual da expressão gênica nas redes das Figuras 4.6 e 4.7. Os genes que foram mencionados, ACE2, AGT, AR, ARHGDIG, EGFR, REN e TERT, estão nas duas tabelas e, portanto, têm um módulo de

expressão consistentemente elevado no conjunto de dados do BRCA.

Os genes ACE2 (*Angiotensinogen Converting Enzyme*) e AGT (*Angiotensinogen*) já foram discutidos anteriormente, e estão associados a manutenção da pressão sanguínea, homeostase iônica e vascularização de tecidos. É importante observar como esses genes possuem tanto o "Fator de Indução" quanto o "Fator de Repressão" elevados, pois a vascularização é de fundamental importância para manutenção e desenvolvimento do tumor (KUMAR; SINGH; BAKER, 2007). Lembrando que os dados permitem-nos concluir apenas que esses dois genes têm uma importância estatística em relação ao seu módulo de "Expressão Relativa" (positiva/negativa), uma vez que não havia controles pareados não é possível afirmar que estão diferencialmente expressos.

O gene ARHGDIG *Rho GDP-dissociation inhibitor 3* tem os maiores "Fatores de Indução/Repressão", segundo as Tabelas 4.3 e 4.4. Este gene codifica um inibidor de dissociação GDP (GDI), que tem a função de modular a ativação de GTPases através da inibição da troca de GDP (guanosina difosfato) por GTP (guanosina trifosfato) (ZALCMAN et al., 1996). Foi mencionado que GTPases funcionam como mediadores moleculares para diversos processos: transdução de sinais de receptores membranares, biossíntese de proteínas, diferenciação, proliferação, divisão, movimento e translocação de proteínas. É simples pressupor que o ARHGDIG tenha uma modulação elevada (positiva e negativa) considerando o elevado número de processos em que podem estar envolvidos, alguns muito importantes para o processo carcinogênico.

O gene AR (*Androgen Receptor*) codifica um fator de transcrição que é ativado pelo hormônio esteroide androgênio. Este fator de transcrição está envolvido na transcrição de genes que modulam diferenciação e proliferação celular, ambos processos importantes no câncer (RAHIM; O'REGAN, 2017). Por essa razão, estudos mostram que a expressão desse gene pode servir como um importante biomarcador para prognóstico do BRCA e potencial alvo para tratamento, dado que uma das principais subclassificações dos tumores mamários é justamente a prevalência dos receptores de esteroides.

O gene TERT (*Telomerase Reverse Transcriptase*) codifica uma ribonucleoproteína responsável pela manutenção de telômeros, regiões terminais dos cromossomos, durante a replicação (HAIMAN et al., 2011). Normalmente, esse gene está reprimido em células somáticas e induzido em células proliferativas ou cancerígenas (HAIMAN et al., 2011).

O gene REN (*Renin*) codifica uma enzima responsável pela primeira etapa de formação da angiotensina II, sua função é catalizar a clivagem do angiotensinogênio para formar a angiotensina I (VINSON; BARKER; PUDDEFOOT, 2012). Como foi mencionado, a angiotensina tem um papel importante na regulação da pressão sanguínea, homeostase iônica e vascularização de tecidos. Assim, é interessante observar que tanto o angiotensinogênio quanto suas enzimas de ativação ACE2 e REN estão igualmente positiva ou negativamente moduladas segundo os dados.

O gene EGFR (*Epidermal Growth Factor Receptor*) codifica uma glicoproteína transmembranar da família tirosina cinase, age como um receptor para diversos ligantes e da inicio a diversas cascatas de sinalização, convertendo estímulos externos em respostas celulares (NAKAI; HUNG; YAMAGUCHI, 2016). Super expressão ou mutação desse gene é reconhecido como causa principal de diversos eventos no processos tumoral, principalmente os que envolvem proliferação celular, transição epitélio-mesênquima e aquisição de resistência a drogas (NAKAI; HUNG; YAMAGUCHI, 2016). O gene TNF (*Tumor Necrosis Factor*) codifica uma citocina pró-inflamatória e multifuncional (CRUCERIU et al., 2020). Geralmente, essa proteína é secretada por macrófagos e está associada à modulação de diversas vias como: proliferação, diferenciação, apoptose, metabolismo de lipídeos e coagulação. Estudos atuais sugerem que o subgrupo de tumores mamários que apresentam elevada expressão desse gene tem grande correlação com proliferação celular, malignância, prognostico pobre e metástase (CRUCERIU et al., 2020).

Continuando a análise com duas tabelas. Temos na Tabela 4.5, os 10 genes com maior valor de *betweenness* e, a seguir, a Tabela 4.6 com os 10 genes com maior diferença entre “Fator de Indução” e “Fator de Repressão”:

RNA-seq: *Betweenness*

Tabela 4.5: Top 10 com maiores valores de *Betweenness* para os dados de RNA-seq. Na tabela temos as sondas das redes na Figura 4.6. Estas sondas estão organizadas de maneira decrescente de acordo com os valores da coluna "*Betweenness*".

Genes	*Betweenness*
RHOA	841.37
HSP90AA1	580.12
MAP3K5	236.67
ACE	236.00
AKT1	208.27
AGT	183.00
EGFR	177.62
RAC1	163.17
CDC42	102.83
RIPK1	96.79

RNA-seq: *Diff*

Tabela 4.6: Top 10 sondas com maiores valores de *Diff* para os dados de RNA-seq. Na tabela temos as sondas das redes nas Figuras 4.6 e 4.7. Estas sondas estão organizadas de maneira decrescente de acordo com os valores da coluna "*Diff*". Esta contém o valor, normalizado em uma escala de 0 a 1, do módulo da diferença entre o "Fator de Indução" e o "Fator de Repressão" da sonda segundo os dados de RNA-seq do tumor mamário (BRCA). A definição destes fatores pode ser encontrada nesta mesma Subseção, nas redes das Figuras 4.6 e 4.7. Assim sendo, na tabela, estão os 10 genes/sondas com maior valor de *Diff*, segundo as redes das Figuras 4.6 e 4.7.

Genes	Fator de Indução	Fator de Repressão	*Diff*
AGTR2	0.54	0.08	1.00
AR	0.66	0.95	0.70
ACE2	0.88	0.76	0.22
TERT	0.72	0.62	0.17
REN	0.63	0.55	0.13
ERBB2	0.22	0.27	0.13
MAP3K5	0.22	0.16	0.10
AGT	0.75	0.68	0.10
RASA1	0.07	0.11	0.09
ACE	0.23	0.18	0.08

Na Tabela 4.5 pode-se observar que dois genes possuem valor de *betweenness* especialmente altos em relação aos outros: RHOA e HSP90AA1. O gene RHOA (*Ras homolog family member A*), é um dos genes significantes do mapa modular de RNA-seq, como descreve a Tabela 4.2. É esperado que um dos genes que ligam um grande número de grupos diferentes apresente um alto valor de *betweenness* calculado e, também, que seja considerado estatisticamente significante, estando envolvido em diversas vias. O mesmo pode ser dito para o gene HSP90AA1 (*Heat Shock Protein 90 Alpha Family Class A Member 1*), ele não é um dos genes estatisticamente significantes, mas é um parálogo de um deles, o HSP90AB1. Estes dois genes têm funções similares, sendo ambas classificadas como chaperonas, responsáveis pela estabilização de diversas proteínas.

Na tabela 4.6 observa-se, ajustada para uma escala de 0 à 1, a diferença entre os módulos positivo e negativo dos genes presentes na rede. Assim como na Tabela 4.5, dois genes estão com valores especialmente elevados em relação aos outros: AGTR2 e AR. Ambos estão associados ao sistema RAS (*Renin Angiotensin System*), sendo que o AGTR2 tem uma expressão mais positiva ("Fator de Indução" > "Fator de Repressão") no conjunto de dados e o AR uma expressão negativa ("Fator de Indução" < "Fator de Repressão"). Lembrando que por "Fator de Indução", se entende como o módulo da mediana da porção positiva (maior que 0) dos valores de expressão relativa da sonda/gene entre as amostras de tumor mamário. A definição de "Expressão Relativa" é fornecida nas Subseção 3.2.6. Sendo que uma "Expressão Relativa" positiva indica indução e uma negativa indica repressão. Assim sendo, a coluna "Fator de Indução" corresponde a um vetor contendo em sí os valores das medianas das porções positivas das sondas/genes, esse vetor é ajustado para uma escala de 0 a 1, criando assim, os valores de "Expressão Positiva" observados na Tabela 4.3. Um processo análogo, com a porção de valores de "Expressão Relativa" negativos, cria os valores de "Fator de Repressão" da Tabela 4.4.

Assim, tendo estas definições em mente, é possível sugerir que o módulo da expressão gênica geral desses genes, representados pelas variáveis "Fator de Indução" e "Fator de Repressão", pode constituir um biomarcador para caracterização de subgrupos do BRCA, em especial, o gene AR, pois tanto seu módulo positivo quanto negativo tem uma magnitude de tamanho considerável. É oportuno mencionar que a expressão do androgênio (AR), atualmente, é utilizada como biomarcador de prognóstico no câncer mamário triplo negativo (ZHANG et al., 2015).

Em suma, análise de redes e de SIRs dos dados de RNA-seq apontam para um cenário onde genes associados a vascularização, manutenção da pressão sanguínea e homeostase iônica e genes ligados ao sistema RAS (*Renin Angiotensin System*) apresentam grande relevância estatística (ACE2, AGT, AR, DAB2IP, REN). Isso é um resultado que deve ser destacado, pois a vascularização é fundamental para manutenção do ambiente microtumoral e portanto para o desenvolvimento tumoral. Também encontramos genes que mediam diversas vias biológicas como ARHGDIG, RHOA e HSP90AA1, esses são genes com elevado valor de *betweenness*, indicando serem importantes para funcionamento da comunicação entre vias.

4.2.2 BRCA: Análise da Metilação

Começamos essa seção listando os módulos selecionados para análise de *overlaps* na tabela 4.7 abaixo:

Metilação: Módulos Selecionados

Tabela 4.7: Metilação: Módulos selecionados. Na tabela temos os módulos selecionados para análise de *overlaps* na primeira coluna. Na segunda coluna temos o número de ontologias contidas dentro do módulo e na terceira o número de sondas.

Módulo	Ontologias	Sondas
artery morphogenesis	5	112
cell proliferation in forebrain	5	73
DNA methylation involved in gamete generation	2	53
double-strand break repair via break-induced replication	2	21
establishment or maintenance of transmembrane electrochemical gradient	3	21
intestinal cholesterol absorption	4	46
negative regulation of endothelial cell migration	2	49
negative regulation of fibrinolysis	2	21
negative regulation of stress fiber assembly	6	173
positive regulation of JAK-STAT cascade	6	121
positive regulation of methylation-dependent chromatin silencing	2	18
positive regulation of protein tyrosine kinase activity	2	47
positive regulation of transcription of nucleolar large rRNA by RNA polymerase I	10	125
positive thymic T cell selection	11	140
regulation of keratinocyte differentiation	5	84
regulation of microtubule polymerization or depolymerization	2	20
regulation of release of sequestered calcium ion into cytosol by sarcoplasmic reticulum	10	202
reverse cholesterol transport	2	20
striatum development	3	29
transcription elongation from RNA polymerase I promoter	3	34
ventricular septum morphogenesis	2	68

A Tabela 4.8 mostra todas as sondas/genes que foram selecionados de acordo com o teste de *bootstrap*.

Metilação: *Overlaps* estatisticamente significantes

Tabela 4.8: Metilação: *Overlaps* estatisticamente significantes. A tabela mostra os genes estatisticamente significantes de acordo com um teste de reamostragem cm reposição de 10000 passos. Na primeira coluna estão os genes (*Symbol*) na segunda o número de *overlaps* e na terceira tem-se o p-valor computado.

Genes	Freq	P-valor
APOE	6	0.0000
FGF8	4	0.0000
VEGFA	6	0.0000
NR2F2	4	0.0002
AGT	3	0.0003
DRD2	4	0.0007
TGFBR1	4	0.0017
SLIT2	3	0.0020
CCL5	3	0.0024
INPP5K	3	0.0048
TBX1	6	0.0058
HIF1A	3	0.0083
MAPK14	3	0.0100

O gene APOE (*Apolipoprotein E*) é uma proteína importante para o transporte e metabolismo de lipídios. Um estudo mostra que esse gene pode promover proliferação celular e metástase no câncer mamário triplo negativo, mas pode impedir a progressão tumoral no caso do câncer ER+, mostrando como a atuação desse gene é contexto específica (HASSEN et al., 2020).

O gene FGF8 (*Fibroblast Growth Factor 8*) está associado com desenvolvimento embrionário, crescimento celular, morfogênese, reparo tecidual e crescimento tumoral (MARSH et al., 1999). Vendo como o FGF8 está associado a tantas vias associadas com a viabilidade tecidual, e portanto, também está associado à viabilidade dos tumores.

O gene VEGFA *Vascular Endothelial Growth Factor A* é fator de crescimento de células endoteliais vasculares, envolvido no processo de angiogênese, fisiológico e patológico. Além de sua importância para vascularização de tecidos, este gene também está associado à inibição da apoptose e migração celular, e a literatura mostra evidências de que este gene está, frequentemente, correlacionado com elevado estágio tumoral (ADAMS et al., 2000).

O gene NR2F2 (*Nuclear Receptor Subfamily 2 Group F Member 2*) é um receptor de hormônios esteroides que funciona como fator de transcrição para uma variedade de genes. Um estudo por Zhang e colaboradores (2014), mostra que esse gene é essencial par mediar a transição epitélio-mesênquima (EMT), tão importante para progressão tumoral, além de inibir diferenciação de células tumorais para mantê-las com um fenótipo pluripotente, aumentando sua viabilidade e possibilidade de metástase (ZHANG et al., 2014).

O gene AGT é estatisticamente significante na análise do mapa modular da metilação. Esse gene codifica o angiotensinogênio, uma forma anterior da enzima fisiologicamente ativa angiotensina II. Sendo assim, é um gene essencial do sistema Ras (Renin Angiotensin System), cujas vias são essenciais para modulação da vascularização de tecidos, tanto em um processo fisiológico quanto patológico (LADD et al., 2007).

O gene DRD2 (*Dopamine Receptor D2*) é um receptor de dopamina pertencente a classe das proteínas G-acopladas, a literatura sugere que este gene tem o potencial de suprimir o desenvolvimento tumoral facilitando a piroptose através da via NF-κB (TAN et al., 2021). O gene TGFBR1 (*Transforming Growth Factor Beta Receptor 1*) é um receptor de fatores de crescimento pertencente a classe serina/treonina, importante para transdução de sinais de receptores TGF de superfície (membrana celular) (MOORE-SMITH; PASCHE, 2011). Segundo a literatura, esse gene tende a estar hipermetilado em diversos tipos de câncer e essa inibição de sua atividade promove a proliferação de células tumorais por meio da redução dos mecanismos moleculares de regulação de fatores de crescimento (MOORE-SMITH; PASCHE, 2011).

O gene SLIT2 (*Slit Guidance Ligand 2*) é uma glicoproteína que funciona como um guia molecular na matriz extracelular para desenvolvimento e migração de células, em especial, no desenvolvimento de axônios. Dallol e colaboradores (2002), mostram que o silenciamento desse gene via hipermetilação de sua região promotora está associado a proliferação de células tumorais e o oposto tende a acontecer no caso de sua reativação (DALLOL et al., 2002). O gene CCL5 (*C-C Motif Chemokine Ligand 5*) é uma quimiocina envolvida no regulação de processos inflamatórios e imunossupressores através do fenômeno da quimiotaxia de células. Soria e colaboradores (2008), mostram que este gene está superexpresso em vários tumores e, no caso do tumor mamário, sua elevada expressão tende a estar correlacionada com malignância e proliferação celular. Estes estudos também sugerem que esse gene é um possível candidato para ser um alvo terapêutico no tratamento da patologia (SORIA;

BEN-BARUCH, 2008).

O gene INPP5K (*Inositol Polyphosphate-5-Phosphatase K*) codifica uma fosfatase que é ativa no citosol, especialmente, em áreas com poucas fibras de *stress*, compostas de actina, que ligam a célula à matriz extracelular. Uma das funções deste gene é a regulação negativa da formação de fibras de *stress* e do citoesqueleto. Por essa razão, estudos sugerem que esse gene pode se tratar de um gene anti-câncer, embora exista pouca literatura abordando essa temática (IJUIN, 2019).

O gene TBX1 (*T-Box Transcription Factor 1*), pertence a uma família de genes filogeneticamente conservadas que possuem o mesmo sítio de ligação com DNA, o T-Box. Um estudo por Huang e colaboradores (2020), mostram que esse genes têm importância na regulação de processos de desenvolvimento. Este mesmo estudo também mostram que o *knockdown* desse gene inviabiliza a proliferação de células tumorais no câncer mamário via regulação de genes associados à modulação do ciclo celular (HUANG et al., 2020).

O gene HIF1A (*Hypoxia Inducible Factor 1 Subunit Alpha*) é um fator de transcrição que pode modular uma grande variedade de genes em resposta a hipóxia celular, incluindo: metabolismo energético, angiogênese e apoptose. Por essa razão, esse gene é considerado um oncogene, envolvido na viabilização do processo cancerígeno. A literatura mostra como que a expressão desse gene está positivamente correlacionado com prognóstico negativo no câncer mamário triplo negativo (WANG; ZHANG; HAN, 2019).

O gene MAPK14 (*Mitogen-Activated Protein Kinase 14*) é uma proteína do tipo cinase pertencente à superfamília de proteínas MAP. Essa proteína age como um ponto de integração para múltiplos sinais bioquímicos e, portanto, está envolvida na regulação de múltiplos processos incluindo: proliferação, diferenciação, regulação de transcrição e desenvolvimento. Existe pouca informação sobre a expressão do MAPK14 ou de seu estado de metilação no contexto do tumor mamário, dificultando a análise deste gene em tal contexto. Entretanto, foi encontrado um estudo por Dashti e colaboradores (2020), que afirma haver grande diferença na expressão de miRNAs e lncRNAs que regulam a expressão do MAPK14, uma observação interessante considerando o elevado número de vias que este gene pode regular (DASHTI et al., 2020).

Com esses 13 genes determinados na Tabela 4.8, foi construída uma rede PPI via *String* segundo o protocolo descrito na Subseção 3.3.4.1. Abaixo, existem duas figuras contendo

essa mesma rede PPI, cada uma tem uma escala de cores com valores de “Fator de Indução/Repressão”. Os valores utilizados para criar a escala de cores são determinados ao ajustar o conjunto de medianas das porções positiva/negativa dos valores de “Metilação Relativa” dos genes presentes na rede entre os valores 1 e 0. Estes novos valores são denominados: “Fator de Indução” para a porção positiva de “Metilação Relativa” e “Fator de Repressão” para a porção negativa. Um gradiente entre 1 e 0 é criado utilizando 4 cores, o que facilita a identificação do módulo do “Fator de Indução/Repressão”. O gradiente é formado por: 0 (preto), 0.25 (amarelo), 0.75 (laranja) e 1 (vermelho). Note que o tamanho dos nodos da rede é ajustado pelo parâmetro *betweenness*, nodos maiores possuem um valor de *betweenness* maior. Cabe lembrar aqui que os valores beta de metilação das ilhas CPG foram convertidos para seu respectivo gene segundo o protocolo descrito na subseção 3.2.5.

Rede PPI dos genes significantes (Metilação Positiva)

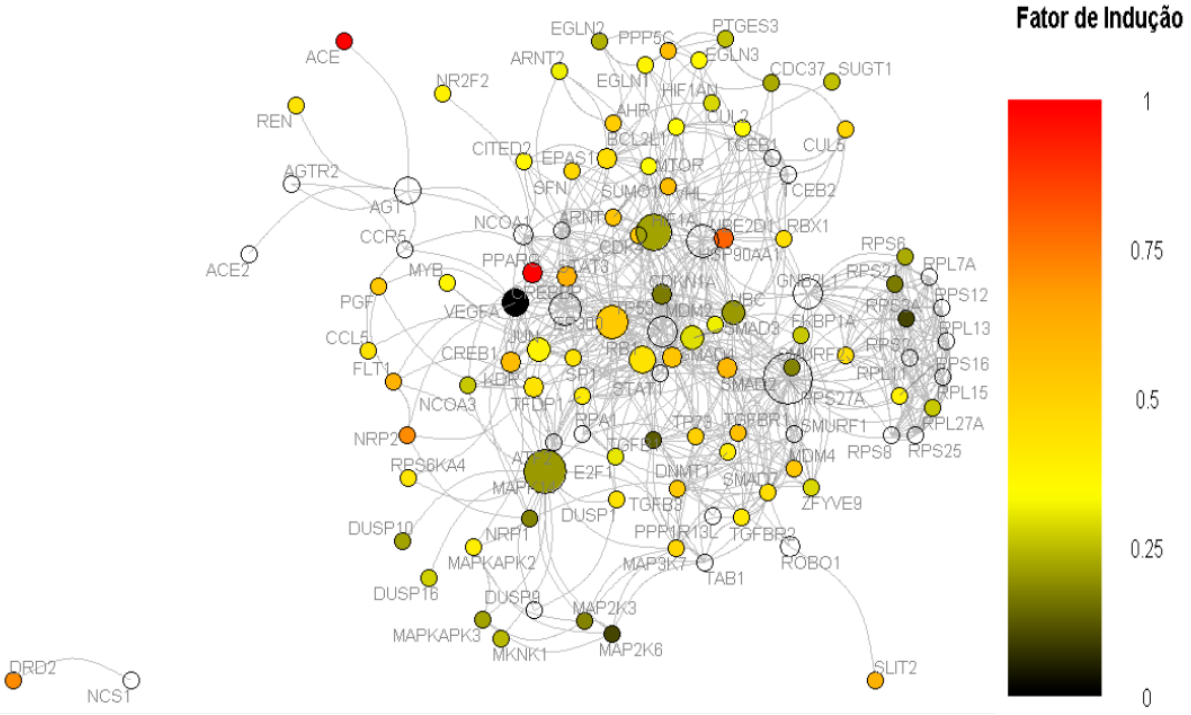


Figura 4.8: Rede PPI dos genes significantes (Metilação Positiva). A rede é obtida ao utilizar os 13 genes descritos na tabela 4.8 no banco de dados *String*, seguindo o protocolo descrito na subseção 3.3.4.1. A escala de cores, aqui denominada “Fator de Indução”, é construída a partir do conjunto de medianas das metilações relativas (também defido com ER) positivas dos genes presentes na rede, considerando apenas o conjunto de dados do BRCA, a definição de ER é fornecida na Subseção 3.2.6. Assim a escala reflete o módulo da metilação positiva (hipermetilação) do conjunto, onde 1 (vermelho) representa módulo elevado e 0 (preto) módulo baixo. O gradiente de cores entre esses valores tem a função de facilitar a comparação visual com sua análoga para expressão negativa abaixo. Note também a presença de nodos transparentes, tais nodos representam genes não encontrados no conjunto de dados segundo as relações GPG-gene prospectadas na subseção 3.2.5.

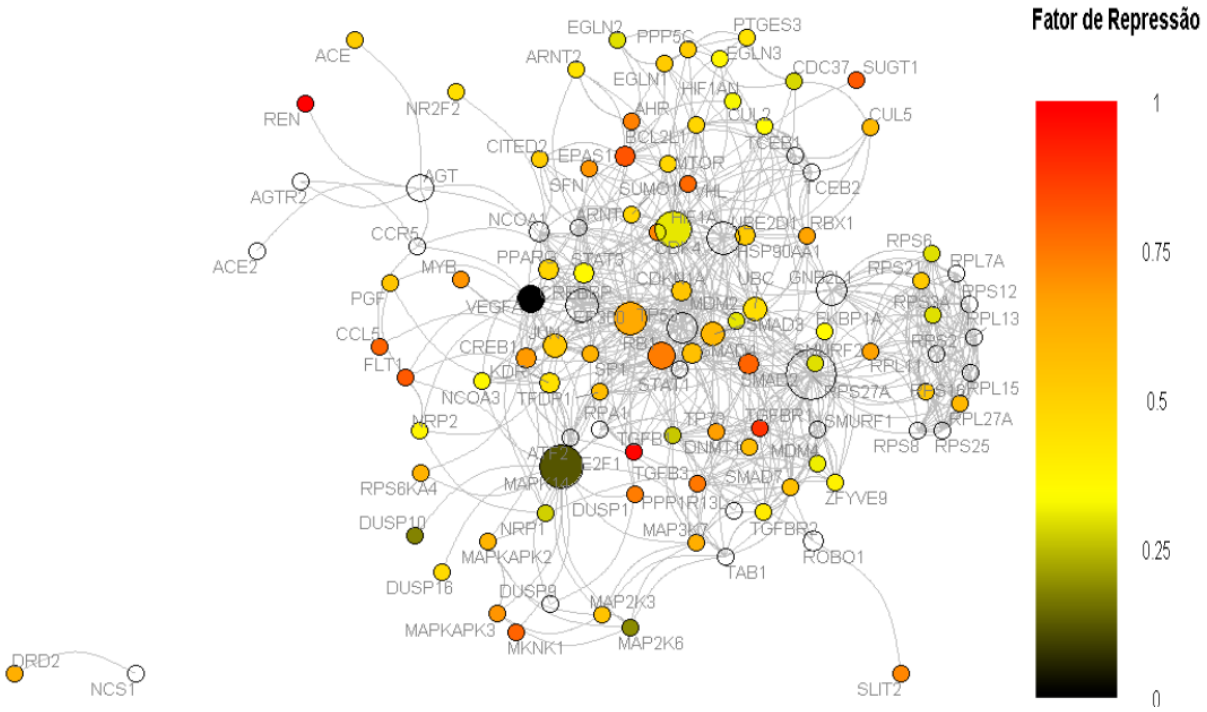


Figura 4.9: Rede PPI dos genes significantes (Metilação Negativa). A rede é obtida ao utilizar os 13 genes descritos na tabela 4.8 no banco de dados *String*, seguindo o protocolo descrito ena subseção 3.3.4.1. A escala de cores, aqui denominada de "Fator de Repressão", é construída a partir do conjunto de medianas das metilações relativas (também definido como ER) negativas dos genes presentes na rede, considerando apenas o conjunto de dados do BRCA, a definição de ER é fornecida na Subseção 3.2.6. Assim a escala reflete o módulo da metilação negativa (hipometilação) do conjunto, onde 1 (vermelho) representa módulo elevado e 0 (preto) módulo baixo. O gradiente de cores entre esses valores tem a função de facilitar a comparação visual com sua análoga para expressão positiva acima. Note também a presença de nodos transparentes, tais nodos representam genes não encontrados no conjunto de dados segundo as relações GPG-gene prospectadas na subseção 3.2.5.

As Figuras 4.8 e 4.9 mostram uma mesma rede, que recebe uma escala de cores diferente em cada figura de acordo com a metilação que é representada, positiva na figura 4.8 e negativa na figura 4.9. Como foram utilizados mais genes do que na tabela 4.2 esta rede ficou maior do que a representada nas figuras 4.6 e 4.7. A primeira observação que pode ser feita em relação as figuras 4.8 e 4.9 é que a figura 4.9 aparenta ter uma coloração mais avermelhada, indicando uma metilação negativa mais elevada do que positiva para a maior parte dos nodos representados na rede. Naturalmente existem exceções a essa generalização, o gene ACE por exemplo tem um "Fator de Indução" maior que o "Fator de Repressão". Para melhor analisar as informações presentes nas figuras 4.8 e 4.9 foram disponibilizadas algumas tabelas abaixo.

Pode-se observar na tabela 4.10 que os genes hipometilados possuem, em geral, um módulo maior do que os hipermetilados, como foi observado de maneira visual anteriormente.

Metilação: Top 10 Induzidos/Reprimidos

Tabela 4.9: Top 10 sondas induzidas para os dados de Metilação. Na tabela temos as sondas da rede na Figura 4.8. Estas sondas estão organizadas de maneira decrescente de acordo com os valores da coluna "Fator de Indução", a qual contém o valor, normalizado em uma escala de 0 a 1, do módulo da mediana da porção positiva de valores de "Metilação Relativa" da sonda segundo os dados de metilação do tumor mamário (BRCA). Lembrando que a definição de "Metilação Relativa" é apresentada na Subseção 3.2.6. Assim sendo, na tabela, estão os 10 genes/sondas com maior indução, segundo a rede da Figura 4.8.

Genes	Fator de Indução
ACE	1.00
PPARG	1.00
UBE2D1	0.79
NRP2	0.71
DRD2	0.71
SLIT2	0.61
STAT3	0.61
FLT1	0.61
SMAD2	0.60
VHL	0.57

Tabela 4.10: Top 10 sondas reprimidas para os dados de metilação. Na tabela temos as sondas da rede na Figura 4.9. Estas sondas estão organizadas de maneira decrescente de acordo com os valores da coluna "Fator de Repressão", a qual contém o valor, normalizado em uma escala de 0 a 1, do módulo da mediana da porção negativa de valores de "Metilação Relativa" da sonda segundo os dados de metilação do tumor mamário (BRCA). Lembrando que a definição de "Metilação Relativa" é apresentada na Subseção 3.2.6. Assim sendo, na tabela, estão os 10 genes/sondas com maior repressão, segundo a rede da Figura 4.9.

Genes	Fator de Repressão
REN	1.00
E2F1	0.98
TGFBR1	0.89
EPAS1	0.82
FLT1	0.82
SUGT1	0.82
CCL5	0.80
MKNK1	0.79
SMAD2	0.79
VHL	0.79

Também pode-se notar a presença de genes que já foram posteriormente discutidos nessa seção. Nos hipermetilados, na Tabela 4.9, foram encontrados os genes ACE, DRD2 e SLIT2 e nos hipometilados foram encontrados o REN, TGFBR1 e o CCL5. O comportamento de metilação que foi observada nas redes corrobora com o que ja foi dito sobre estes genes. DRD2 e SLIT2 tendem a estar silenciados em câncer, como os estudos mencionaram anteriormente e se encontram majoritariamente hipermetilados no conjunto de dados do BRCA.

Segundo o que foi discutido anteriormente, TGFBR1 e CCL5 tendem a estar mais expressos em tumores e se encontram majoritariamente hipometilados segundo os dados, novamente corroborando com observações feitas com diversos estudos de bancada.

ACE é uma outra observação a ser notada, este gene codifica a enzima de conversão da

angiotensina I e faz parte do sistema RAS (*Renin angiotensin system*), que foi discutido tanto no mapa modular de metilação quanto de RNA-seq. Segundo os dados, o ACE se encontra majoritariamente hipermetilado no dados de BRCA, o que sugere um silenciamento da expressão desse gene. De maneira oposta, o gene ACE2 se encontra positivamente regulado nos dados de RNA-seq segundo as figuras 4.6 e 4.7 assim como o gene AGT (angiotensinogenio), sugerindo que o sistema RAS tenha uma importante função na manutenção do processo tumoral por meio da vascularização e manutenção de homeostase. O gene REN (*Renin*) é o gene com maior módulo de metilação negativa, sugerindo uma elevada expressão do mesmo, algo que novamente é confirmado pelas figuras 4.6 e 4.7 nos dados de RNA-seq.

Existem outros genes para serem discutidos na tabela 4.9. O PPARG (*Peroxisome Proliferator Activated Receptor Gamma*), por exemplo, é um gene envolvido no armazenamento de lipídios, e vias associadas com transporte e metabolismo de lipídios foram identificados como essenciais para maior parte dos tumores na seção 4.1. De fato, a literatura confirma que esse gene é, tipicamente, pouco expresso no tumor mamário e isso tem correlação positiva com prognósticos ruins (AUGIMERI et al., 2020). Nossos dados demonstram que esse gene está altamente hipermetilado, corroborando para as observações feitas nos estudos mencionados (AUGIMERI et al., 2020).

O gene UBE2D1 (*Ubiquitin Conjugating Enzyme E2 D1*) codifica uma ubiquitina, uma proteína que tem como função marcar proteínas aberrantes ou de baixa vida útil para degradação molecular (BUI et al., 2021). Portanto, observar sua elevada hipermetilação no conjunto de dados do BRCA é algo importante, pois seu silenciamento pode ser uma das causas de progressão da patologia.

O gene NRP2 (*Neuropilin-2*) codifica um receptor para o fator de crescimento vascular endotelial (VEGF) e pode estar associado com a regulação de múltiplos processos incluindo: desenvolvimento vascular, desenvolvimento de axônios e tumorigênese (YASUOKA et al., 2009). Yasuoka e colaboradores (2009), sugerem que isso possa ser um prognóstico ruim, no entanto, os dados de metilação sugerem que esse gene está ativamente sendo silenciado, algo que é corroborado pelos dados de RNA-seq, pois o NRP2 não foi selecionado como uma SIR segundo o método utilizado. É possível que essa discrepância em relação a expressão do NRP2 seja fruto da limitação intrínseca do método, uma vez que o mesmo não trabalha com controles (YASUOKA et al., 2009).

O gene STAT3 (*Signal Transducer And Activator Of Transcription 3*) codifica uma

proteína da família STAT, que atuam como fatores de transcrição para citocinas e fatores de crescimento (MA; QIN; LI, 2020). Esse gene está tipicamente ativado no câncer mamário, mas os dados de metilação sugerem uma elevada hipermetilação do mesmo e os dados de RNA-seq sequer chega a selecionar esse gene como uma SIR de relevância (MA; QIN; LI, 2020). Essa é uma situação parecida com a do gene NRP2, talvez causada pela mesma limitação.

O gene FLT1 (*Fms Related Receptor Tyrosine Kinase 1*) codifica uma tirosina cinase membro da família VEGF (*Vascular Endothelial Growth Factor*) (QIAN et al., 2015). Sua expressão está associada a regulação de várias vias, incluindo: angiogênese, viabilidade celular, migração celular, quimiotaxia, macrófagos e metástase (QIAN et al., 2015). A literatura sugere que a expressão desse gene está associada a metástase e, portanto, a melhores prognósticos(QIAN et al., 2015). Entretanto, a ausência de expressão de FLT1 não parece ser primordial para a metástase, embora isso possa diminuir a eficiência do processo.

Nossos dados mostram a hipermetilação desse gene, o que sugere um silenciamento de sua expressão. E, de fato, a Tabela 4.9 confirma isso, mostrando que o FLT1 é um dos genes com maior nível de hipometilação no conjunto. Essa situação demonstra que, segundo os dados, o FLT1 seja um gene de considerável importância no que tange sua metilação uma vez que tanto seu "Fator de Indução" quanto seu "Fator de Repressão" mostram um módulo elevado. Esta observação poderia constituir um marcador para subgrupos diferentes de câncer mamário, contudo seria necessário realizar outro tipo de análise para testar essa hipótese.

O gene SMAD2 (*SMAD Family Member 2*) codifica uma proteína da família SMAD, essas proteínas funcionam como transdutoras de sinais e moduladoras transcricionais para diversas vias (TIAN et al., 2003). A SMAD2, em especifico, interage com o TGF-beta e, portanto, pode regular vias associadas como apoptose, proliferação e diferenciação (TIAN et al., 2003). Um estudo confirma que a expressão do SMAD2 tem um efeito ambíguo no câncer mamário, simultaneamente aumentando a proliferação de células e diminuindo sua capacidade de metástase (TIAN et al., 2003). Isso é uma observação a ser destacada pois é possível observar que o SMAD2 também se apresenta entre hipometilados segundo a tabela 4.10, tal qual o gene FLT1. Assim, o SMAD2 também pode ser um candidato a marcação de subgrupo tal como o FLT1. E se for levado em consideração o estudo mencionado anteriormente, pode-se sugerir que na hipermetilação observa-se um aumento

da proliferação celular em detrimento da metástase e vice-versa no caso da hipometilação.

Por fim, o gene VHL (*Von Hippel-Lindau*) codifica uma proteína com atividade de ubiquitinase, seu principal alvo são proteínas da família HIF (*Hipoxia Inducible Factor*) (ZIA et al., 2007). Esse gene tem o seu perfil de metilação parecido com o do gene FLT1, estando presente nas Tabelas 4.9 e 4.10, onde pode-se ver que o VHL está claramente mais hipometilado. Devido a relação do VHL com a via da hipóxia é importante relembrar que o gene HIF1A é dos genes estatisticamente significantes na Tabela 4.8.

Agora são avaliados os genes que apareceram exclusivamente como hipometilados na Tabela 4.10 e que ainda não foram mencionados. O gene E2F1 (*E2F Transcription Factor 1*) codifica um fator de transcrição pertencente à família E2, suas funções envolvem a modulação de algumas vias incluindo: ciclo celular, proliferação, apoptose por via TP53 e diferenciação de adipócitos. Um estudo por Hollern e colaboradores (2019), mostra que a expressão elevada desse gene é um biomarcador para metástase do câncer mamário, algo que ocorre devido sua função moduladora da progressão do ciclo celular (HOLLERN et al., 2019). De maneira complementar, um estudo por Al-Sharif e colaboradores (2013), mostrou que o silenciamento da gene promove a apoptose celular e diminuição da metástase (AL-SHARIF; REMMAL; ABOUSSEKHRA, 2013). Nossos dados mostram uma elevada hipometilação do E2F1 oque sugere uma elevada expressão do mesmo, isso corrobora com as informações dos estudos citados, pois a elevada expressão desse gene promove o desenvolvimento do tumor.

O gene EPAS1 (*Endothelial PAS Domain Protein 1*) codifica um fator de transcrição envolvido na indução de genes ativados por níveis de oxigênio e o mesmo é ativado perante queda na concentração de oxigênio disponível (hipóxia). Um estudo por Cui e colaboradores (2016), mostrou que este gene, em geral, está pouco metilado no câncer mamário o que contribui para sua expressão e isso é corroborado pelos resultados na tabela 4.9 (CUI et al., 2016).

O gene SUGT1 (*SGT1 Homolog*) codifica uma cochaperona da proteína HSP90, que possuem a função de estabilizar proteínas ou degradar proteínas marcadas. Um estudo por Ogi e colaboradores (2015), mostrou que esta proteína tende a estar superexpressa em diversos tipos de tumores, incluindo mamário e os dados corroboram para essa observação, considerando o elevado nível de hipometilação do SUGT1. O mesmo estudo demonstra que o *knockdown* do SUGT1 suprimiu a formação de tumores em ratos e aumentou o tempo de

sobrevivência dos mesmos (OGI et al., 2015). Levando em consideração essas observações pode-se sugerir a baixa metilação do SUGT1 como um biomarcador para presença de um processo tumoral.

O gene MKNK1 (*MAPK Interacting Serine/Threonine Kinase 1*) codifica um proteína serina/treonina cinase que esta envolvida na resposta a estresse ambiental e à atuação de citocinas. Um estudo por Astanehe e colaboradores (2012), demonstra que a expressão desse gene está associada a diminuição da sensibilidade a drogas por parte das células tumorais (ASTANEHE et al., 2012). O MKNK1 está hipometilado, portanto, pode-se presumir uma elevada expressão do mesmo e, por consequência, também que isso possa induzir certa resiliência às células tumorais nas amostras. Isso é apenas uma sugestão, mas os dados corroboram com as observações feitas nos estudos, de que esse gene tende estar expresso em células tumorais e, portanto, é possível presumir um baixo nível de metilação para o mesmo.

Continuando a análise com duas tabelas. Temos na Tabela 4.11, os 10 genes com maior valor de *betweenness* e, a seguir, a Tabela 4.12 com os 10 genes com maior diferença entre "Fator de Indução" e "Fator de Repressão":

Metilação: *Betweenness*

Tabela 4.11: Top 10 com maiores valores de *Betweenness* para os dados de metilação. Na tabela temos as sondas da rede nas Figuras 4.8 e 4.9. Estas sondas estão organizadas de maneira decrescente de acordo com os valores da coluna "*Betweenness*".

Genes	*Betweenness*
MAPK14	817.46
HIF1A	574.79
EP300	532.27
VEGFA	335.51
RB1	289.35
JUN	251.06
SMAD3	218.56
UBC	214.22
STAT3	155.01
SMAD4	137.52

Metilação: *Diff*

Tabela 4.12: Top 10 sondas com maiores valores de *Diff* para os dados de metilação. Na tabela temos as sondas das redes nas Figuras 4.8 e 4.9. Estas sondas estão organizadas de maneira decrescente de acordo com os valores da coluna "*Diff*". Esta contém o valor, normalizado em uma escala de 0 a 1, do módulo da diferença entre o "Fator de Indução" e o "Fator de Repressão" da sonda segundo os dados de RNA-seq do tumor mamário (BRCA). A definição destes fatores pode ser encontrada nesta mesma Subseção, nas redes das Figuras 4.8 e 4.9. Assim sendo, na tabela, estão os 10 genes/sondas com maior valor de *Diff*, segundo as redes das Figuras 4.8 e 4.9.

Genes	Fator de Indução	Fator de Repressão	*Diff*
PPARG	1.00	0.48	1.00
ACE	1.00	0.52	0.97
NRP2	0.71	0.37	0.67
UBE2D1	0.79	0.52	0.65
E2F1	0.30	0.98	0.54
STAT3	0.61	0.36	0.52
SUGT1	0.25	0.82	0.46
MKNK1	0.24	0.79	0.45
MDM4	0.53	0.31	0.45
DRD2	0.71	0.62	0.43

Na tabela 4.12 acima, na parte direita, observa-se a seleção dos top 10 genes com maior diferencial entre as medianas da hipermetilação e da hipometilação, convertida para uma escala de 1 a 0. Quase todos os genes aqui foram mencionados anteriormente, com exceção do gene MDM4. O gene PPARG foi mencionado como potencial regulador de vias associadas à regulação do armazenamento e transporte de lipídios, um processo identificado como importante na Seção 4.1. Esse gene está tipicamente subexpresso em tumores mamários, portanto, observar essa discrepância entre hipermetilação e hipometilação é algo que corrobora para estas observações, visto que o módulo de sua hipermetilação é maior que de sua hipometilação, possuindo o maior diferencial (*Diff*) entre todos os genes selecionados.

Os genes ACE, UBE2D1, STAT3 e DRD2 também apresentam esse perfil de metilação, com o "Fator de Repressão", em geral, menor do que o "Fator de Indução", como pode ser observado na Tabela 4.12. Os genes NRP2 e STAT3 (Ma e colaboradores) são casos especiais, pois, a literatura sugere que os mesmos se encontram altamente expressos no câncer mamário, mas tanto os dados de metilação quanto de RNA-seq sugerem o contrário.

Observe a hipermetilação dos mesmos na tabela 4.12 (YASUOKA et al., 2009; MA; QIN; LI, 2020).

Novamente, é lembrado que isso pode ser uma limitação do método, que trabalha apenas com sumarização estatística dos casos, sem considerar os controles. Teria de se realizar um estudo de expressão diferencial formal (RNA-seq) para avaliar se a expressão desses dois genes estão ativadas como dizem os estudos mencionados. O mesmo pode ser dito em relação à metilação, mas é necessário lembrar que nem sempre a hipermetilação garante o silenciamento de um gene, uma vez que existem outras maneiras de modular a expressão de um gene. Sendo assim a hipermetilação observada na Tabela 4.12 para os genes NRP2 e STAT3 pode sim refletir a realidade biológica e não ser apenas um artefato das limitações impostas pelo método de análise de dados.

Os genes E2F1, SUGT1 e MKNK1 corroboram com oque foi discutido sobre hipometilação na tabela 4.10, pois a Tabela 4.12, mostra claramente a hipometilação desses ("Fator de Indução" < "Fator de Repressão"), o que a princípio pode implicar numa atividade de expressão maior para tais genes e essa regulação positiva foi discutida anteriormente.

Por fim, tem-se o gene MDM4 (*MDM4 Regulator Of P53*) que codifica uma proteína nuclear com um domínio de ligação p53 que pode inibir a atuação de proteínas com efeitos anti-tumoral (HAUPT et al., 2017). Haupt e colaboradores (2017), reportam a super expressão desse gene, mas os dados de metilação sugerem que o mesmo tende a estar mais hipermetilado do que hipometilado, algo que a principio contraria essa observação. Nossos dados de RNA-seq na subseção 4.2.1 não mostram a expressão do MDM4 como positiva ou negativamente regulada, sendo assim essa discrepância pode ser fruto da limitação do método (HAUPT et al., 2017).

Na tabela 4.11, estão os top 10 genes com maior valor de *Betweenness*. Alguns desses genes foram mencionados anteriormente: MAPK14, HIF1A, VEGFA e STAT3. O gene MAPK14 em especial apresenta o maior valor de *Betweenness* do conjunto, e isso condiz com sua função como ponto de integração de diversos sinais bioquímicos. Os outros genes mencionados, de alguma forma, têm papel fundamental em vias de sinalização que regulam aspectos fundamentais para manutenção do processo tumoral e incluem: proliferação, diferenciação, angiogênese e apoptose.

Os outros genes ainda não mencionados serão brevemente discutidos a seguir. O gene

EP300 (*E1A Binding Protein P300*) codifica uma acetil-transferase de histonas que viabiliza o processo de transcrição de inúmeras vias de sinalização perante remodelamento da cromatina (RING; KAUR; LANG, 2020). Essa função de remodelamento da cromatina garante um papel especial para esse gene durante processos de intensa atividade transcricional como durante a proliferação e diferenciação. Sendo assim, é esperado encontrar um valor elevado de *Betweenness* para esse gene, mesmo que sua expressão ou metilação não tenham sido identificado como significante.

O gene RB1 (*RB Transcriptional Corepressor 1*) codifica uma proteína com atividade anti-tumoral, sendo esta, uma reguladora negativa da progressão do ciclo celular, O RB1 está diretamente envolvido com a manutenção da heterocromatina através da estabilização de histonas, sendo assim, vital para momentos de grande atividade transcricional (JIANG et al., 2011). Essa proteína também media a repressão transcricional epigenética através do recrutamento de metil-transferases e também pode mediar a repressão transcricional pelo recrutamento do complexo de desacetilação de histonas (HDAC) (JIANG et al., 2011). Sendo o RB1 necessário para regulação de tantos processos envolvendo a estabilização da heterocromatina e para repressão transcricional (epigenética ou não), é esperado que este gene seja encontrado com um valor elevado de *Betweenness* mesmo quando seu "Fator de Indução" ou "fator de Repressão" não sejam particularmente elevados segundo a metodologia.

O gene JUN (*Jun Proto-Oncogene, AP-1 Transcription Factor Subunit*) é um importante fator de transcrição que modula a regulação de diversas vias associadas a progressão do ciclo celular e à apoptose (VLEUGEL et al., 2006). O gene JUN foi um dos primeiros fatores de transcrição oncogênicos a ser descoberto e sua ativação tipicamente se dá através de resposta a certos estímulos de superfície que incluem: fatores de crescimento, citocinas pró-inflamatórias, estresse oxidativo e físico (VLEUGEL et al., 2006). Considerando o intenso estresse metabólico que as células tumorais estão normalmente submetidas e a identificação da elevada expressão de fatores de crescimento na subseção 4.2.1 é esperado que esse gene possua uma elevada métrica de *Betweenness*.

Os genes SMAD3 e SMAD4 são proteínas que fazem parte da mesma família que o SMAD2, mencionada anteriormente como transdutor de sinais de superfície e como modulador de transcrição de genes. Considerando o que foi sugerido sobre os efeitos que o SMAD2 pode ter nas células, especificamente, sobre proliferação e metástase, pode-se supor que uma importância similar pode ser esperada das proteínas SMAD3 e SMAD4, o que explicaria

seus elevados valores de *Betweenness*.

Por fim, tem-se o gene UBC (*Ubiquitin C*) que codifica uma ubiquitina. Como mencionado anteriormente, essa classe de proteínas tem a função de estabilizar outras proteínas ou marca-las para degradação (BIANCHI et al., 2018). Assim sendo, é esperado que ubiquitinas sejam parte de diversos processos de regulação. A UBC, em especifico, é ativada em situações de intenso estresse, algo comum em células tumorais, sendo assim o elevado valor de *Betweenness* observado é algo previsto.

4.2.3 BRCA: Análise do miRseq

Essa seção é iniciada com a lista dos módulos selecionados para análise de *overlaps* na Tabela 4.13 abaixo:

miRseq: Módulos Selecionados

Tabela 4.13: miRseq: Módulos Selecionados. Na tabela estão os módulos selecionados para análise de *overlaps* na primeira coluna. Na segunda coluna estão o número de ontologias contidas dentro do módulo e na terceira o número de sondas/genes.

Módulo	Ontologias	Sondas
cellular response to diacyl bacterial lipopeptide	103	158
gene silencing by RNA	2	6
miRNA mediated inhibition of translation	2	40
negative regulation of BMP signaling pathway	2	1
negative regulation of cholesterol efflux	2	7
negative regulation of epithelial to mesenchymal transition	2	6
negative regulation of lipid biosynthetic process	2	3
negative regulation of osteoblast differentiation	2	5
positive regulation of cardiac muscle cell proliferation	2	17
positive regulation of cholesterol efflux	2	1
positive regulation of myotube differentiation	2	4
positive regulation of vascular endothelial growth factor receptor signaling pathway	2	5

A Tabela 4.14 abaixo, mostrando todas as sondas/genes que foram selecionadas de acordo com o teste de *bootstrap*.

miRseq: *Overlaps* estatisticamente significantes

Tabela 4.14: miRseq: *Overlaps* estatisticamente significantes. A tabela mostra os genes estatisticamente significantes de acordo com um teste de reamostragem cm reposição de 10000 passos. Na primeira, coluna encontram-se os genes (*Symbol*), na segunda, o número de *overlaps* e na terceira, encontra-se o p-valor computado.

miRNAs	Freq	P-valor
hsa-miR-133b	6	0.0000
hsa-miR-138-5p	7	0.0000
hsa-miR-144-3p	10	0.0000
hsa-miR-590-5p	4	0.0000
hsa-miR-665	4	0.0000
hsa-miR-675-5p	5	0.0001
hsa-miR-146b-5p	4	0.0004
hsa-miR-33a-5p	3	0.0004
hsa-miR-33b-5p	4	0.0004
hsa-miR-10a-5p	3	0.0005

Foi encontrada pouca informação no que tange a relação miRNA-miRNA, de fato, estudos sobre miRNAs, em geral, têm interesse em avaliar principalmente as relações que

esses transcritos tem com mRNAs. Por essa razão, não foram encontradas bases de dados com relações miRNA-miRNA, assim sendo, as redes desta seção serão compostas pela relação miRNA-mRNA. Entretanto, primeiro é apresentada análise do perfil de expressão dos transcritos da Tabela 4.14, que pode ser observado na tabela abaixo:

Tabela 4.15: Top 10 miRNAs com maior *Diff*. A variável "*Diff*" é determinada como o módulo da diferença entre os valores de "Fator de Indução" e de "Fator de Repressão", que posteriormente são convertidos para uma escala de 1 a 0. A tabela foi ordenada para demonstrar os genes com os 10 maiores valores de *Diff*.

miRNAs	Fator de Indução	Fator de Repressão	Diff
hsa-miR-133b	0.33	0.00	1.00
hsa-miR-138-5p	0.43	0.12	0.96
hsa-miR-665	0.39	0.10	0.91
hsa-miR-204-5p	1.00	1.00	0.21
hsa-miR-10a-5p	0.44	0.46	0.19
hsa-miR-33b-5p	0.74	0.92	0.18
hsa-miR-144-3p	0.45	0.61	0.10
hsa-miR-590-5p	0.11	0.19	0.09
hsa-miR-675-5p	0.44	0.58	0.07
hsa-miR-33a-5p	0.40	0.49	0.07
hsa-miR-132-3p	0.00	0.12	0.01
hsa-miR-146b-5p	0.10	0.22	0.00

Pode-se observar que poucos miRNAs possuem um valor de *Diff* realmente elevado (os 3 primeiros). Sendo assim, a discussão e posterior análise de redes miRNA-mRNA foram limitadas aos 6 primeiros miRNAs, que possuem um *Diff* maior que 0.1. Como foi mencionado nessa Subseção, existe pouca informação disponível sobre os miRNAs e certamente alguns desses transcritos possuem mais informações disponíveis do que outros.

O miRNA hsa-miR-133b é o transcrito com maior valor de *Diff* nos dados de câncer mamário. A Tabela 4.15 mostra que esse transcrito está positivamente regulado no conjunto de dados, ou seja, o "Fator de Indução" é maior que o "Fator de Repressão". De acordo com *Wang* e colaboradores (2018), o hsa-miR-133b costuma estar subexpresso no câncer mamário e que esta subexpressão tem correlação com um prognóstico negativo (WANG et al., 2018). *Wang* e colaboradores (2018), mencionam que o hsa-miR-133b é um supressor de vias associadas com a metástase para vários tumores em experimentos *in vitro* (WANG et al., 2018). Sendo assim, pode-se pressupor que a regulação positiva observada nos dados é um prognóstico positivo, pois segundo *Wang* e colaboradores, a expressão de hsa-miR-133b é indiretamente correlacionada com o estagio de metástase do tumor (WANG et al., 2018).

O miRNA hsa-miR-138-5p possui mais informações disponíveis que indicam um papel

ambíguo para a expressão deste transcrito no que tange o processo tumoral. Segundo um estudo, esse transcrito se encontra subexpresso no câncer mamário e isso se trata de um prognóstico ruim, pois a expressão do hsa-miR-138-5p está associada a redução da proliferação de células tumorais e da transição epitélio-mesênquima (EMT) (RASOOLNEZHAD et al., 2021). No entanto, outro estudo mostra que a atuação desse transcrito no ambiente microtumoral mediante transporte por vesículas promove fenótipos pró-cancerígenos como indução de resistência a drogas e polarização de macrófagos associados ao câncer (CAMs)(XUN et al., 2021).

A Tabela 4.15 mostra a regulação positiva desse miRNA no conjunto de dados de câncer mamário ("Fator de Indução" > "Fator de Repressão"). É possível tanto sugerir que se pode se tratar de um prognóstico positivo ou negativo a depender da papel que esse transcrito desempenham, algo que não é possível determinar somente com os dados utilizados.

O miRNA hsa-miR-665 tem informações convergentes, um estudo mostra que sua superexpressão em tumores mamários constitui um prognóstico ruim, sendo tal superexpressão positivamente correlacionada com proliferação, migração, invasão e transição epitélio-mesênquima (ZHAO et al., 2019). A Tabela 4.15 mostra regulação positiva do hsa-miR-665 ("Fator de Indução" > "Fator de Repressão"), sugerindo um prognóstico ruim. Segundo *Zhao* e colaboradores (2019), isso pode acontecer mediante inibição do gene NR4A3 (*nuclear receptor subfamily 4 group A member 3*) e posterior ativação da via MAPK/ERK (ZHAO et al., 2019).

O miRNA hsa-miR-204-5p é um caso particularmente importante, pois segundo a Tabela 4.15 ele simultaneamente possui o maior "Fator de Indução" e "Fator de Repressão" no conjunto de dados, o que pode implicar que este pode ser um bom marcador de subgrupos de amostras. De acordo com dois estudos, esse miRNA está consistentemente subexpresso em tumores mamários e isso constitui um prognóstico negativo, pois a superexpressão desse miRNA esta associada supressão do crescimento tumoral através de diversas vias incluindo: proliferação, metástase, ciclo celular e remodelação do microambiente tumoral (HONG et al., 2019; SHEN et al., 2017).

O miRNA hsa-miR-10a-5p possui pouca diferença entre seus módulos de expressão positiva e negativa e de fato, o valor da respectiva variável *Diff* é o segundo mais baixo dentre os 6 miRNAs selecionados. Essa baixa variação entre os módulos positivo e negativo, torna difícil realizar especulações sobre a função que este miRNA pode ter. De acordo com a

literatura, o hsa-miR-10a-5p tende a estar reprimido em tumores mamários particularmente agressivos (WORST et al., 2019). A expressão do hsa-miR-10a-5p está associada a inibição da proliferação celular, migração celular e promoção da apoptose mitocondrial (WORST et al., 2019).

O hsa-miR-33b-5p possui o menor valor de *Diff*, e sua expressão está pouco mais negativa como pode-se ver na Tabela 4.15. Um estudo de *Lin* e colaboradores (2015), demonstra que, de fato, esse gene tende a estar reprimido em tumores mamários, essa supressão tem correlação positiva com fenótipos mais agressivos da doença (LIN et al., 2015). O mesmo estudo menciona que uma expressão elevada do hsa-miR-33b-5p está associada à inibição do fenótipo de células pluripotentes associado à células tumorais e por extensão à capacidade de sobrevivência dessas células (LIN et al., 2015).

Prosseguindo com a análise da expressão de miRNAs, utilizando redes de interação miRNA-mRNA. Inicialmente, a ideia era criar uma rede de interação miRNA-miRNA, mas esse tipo de informação é escassa. Por essa razão, foi seguido o protocolo descrito na Subseção 3.3.4.2 para definir redes de interação miRNA-mRNA para os miRNAs de interesse: hsa-miR-133b, hsa-miR-138-5p, hsa-miR-665, hsa-miR-204-5p, hsa-miR-10a-5p e hsa-miR-33b-5p.

Após a prospecção das relações miRNA-mRNA, foram selecionados apenas aquelas cujos mRNAs aparecem nas redes 4.6 e 4.7 (redes de RNA-seq) e também nas redes 4.8 e 4.9 (redes de metilação). Depois disso, as relações miRNA-mRNA filtradas foram concatenadas às relações mRNA-mRNA de sua respectiva rede (RNA-seq ou metilação) cujos mRNAs estejam presentes nas relações miRNA-mRNA. Assim foi obtida uma rede com interações miRNA-mRNA e mRNA-mRNA, segundo dados que foram obtidos posteriormente. Abaixo encontram-se as Figuras 4.10 e 4.11 com a rede referente aos dados miRNA-mRNA construída a partir da rede de RNA-seq.

Rede miRNA-mRNA (RNA-seq) dos miRNAs significantes (Expressão Positiva)

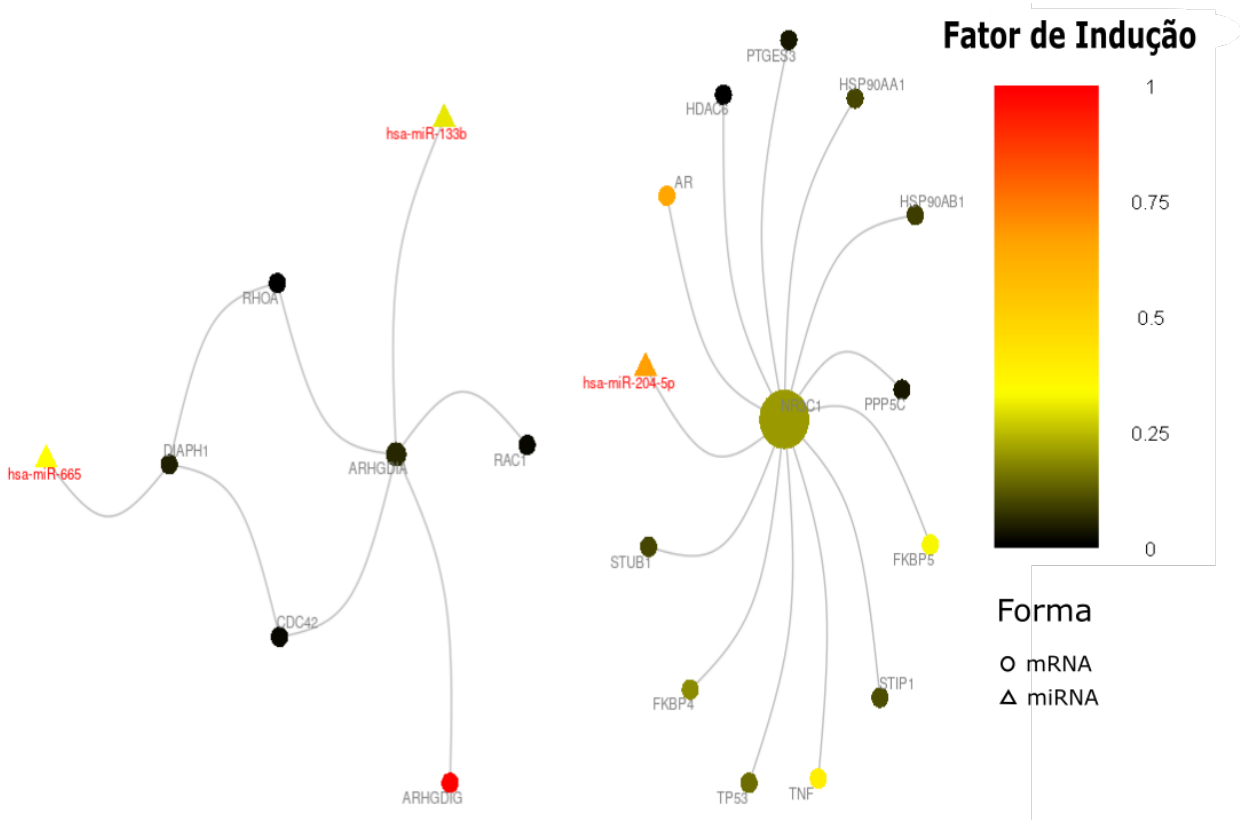


Figura 4.10: [Rede miRNA-mRNA dos miRNAs significantes (Expressão Positiva). A rede é obtida ao utilizar os 6 primeiros miRNAs descritos na Tabela 4.14 no banco de dados *Starbase ENCORI*, seguindo o protocolo descrito ena subseção 3.3.4.2. A escala de cores, aqui chamada de "Fator de Indução", é construída a partir do conjunto de medianas das expressões relativas (ER) positivas dos nodos presentes na rede, considerando apenas o conjunto de dados do BRCA, a definição de ER é fornecida na Subseção 3.2.6. Assim a escala reflete o módulo da expressão positiva do conjunto, onde 1 (vermelho) representa módulo elevado e 0 (preto) módulo baixo. O gradiente de cores entre esses valores tem a função de facilitar a comparação visual com sua análoga para expressão negativa abaixo. Note que o tamanho dos nodos é ajustado pelo parâmetro *betweenness* do nodo.

Rede miRNA-mRNA (RNA-seq) dos miRNAs significantes (Expressão Negativa)

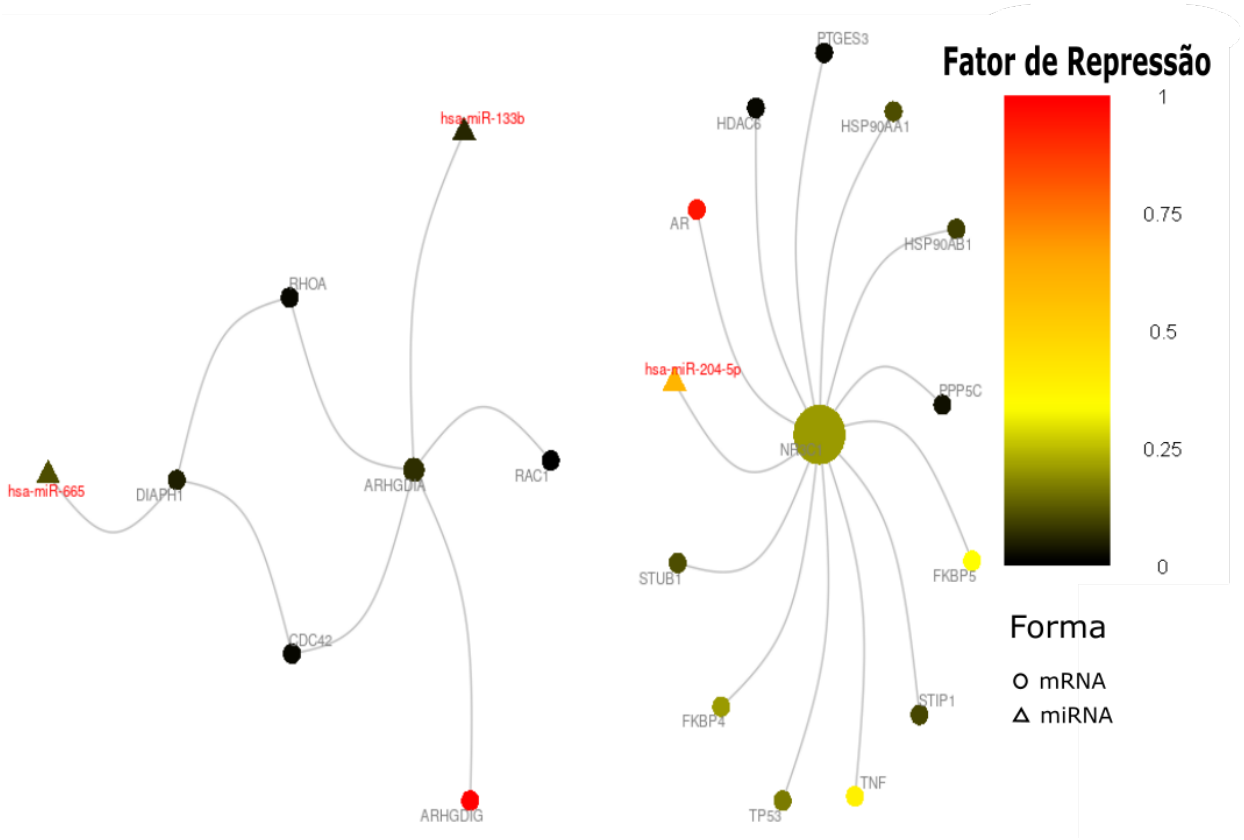


Figura 4.11: Rede miRNA-mRNA dos miRNAs significantes (Expressão Negativa). A rede é obtida ao utilizar os 6 primeiros miRNAs descritos na tabela 4.14 no banco de dados *Starbase ENCORI*, seguindo o protocolo descrito na subseção 3.3.4.2. A escala de cores, aqui chamada de "Fator de Repressão", é construída a partir do conjunto de medianas das expressões relativas (ER) negativas dos nodos presentes na rede, considerando apenas o conjunto de dados do BRCA, a definição de ER é fornecida na Subseção 3.2.6. Assim a escala reflete o módulo da expressão negativa do conjunto, onde 1 (vermelho) representa módulo elevado e 0 (preto) módulo baixo. O gradiente de cores entre esses valores tem a função de facilitar a comparação visual com sua análoga para expressão positiva acima. Note que o tamanho dos nodos é ajustado pelo parâmetro *betweenness* do nodo.

Nas Figuras 4.10 e 4.11 acima, pode-se ver a rede miRNA-mRNA construída a partir dos dados de RNA-seq, coloridas de acordo com o "Fator de Indução" e "Fator de Repressão", respectivamente. A primeira observação a ser feita é o tamanho dessa rede, que contém poucos elementos em relação as redes das Seções 4.1.1 e 4.1.2. Isso ocorre devido as rigorosas etapas de filtragens descritas na Subseção 3.3.4.2 junto ao fato das informações sobre miRNAs serem em geral mais escassas em relação aos mRNAs.

Outra observação a ser feita é que apenas 3 dos 6 miRNAs de interesse (hsa-miR-133b, hsa-miR-665 e hsa-miR-204-5p) estão presentes na rede. Isso é consequência da etapa de filtragem onde foram selecionados somente as relações miRNA-mRNA cujos mRNAs estejam

presentes na rede dos dados de RNA-seq na Subseção 4.1.1.

Por fim, também pode-se observar que temos duas redes claramente separadas na Figura 4.10 e na Figura 4.11, sendo que a rede à direita aparenta se tratar de um único nodo ligado a todos os outros enquanto a rede à esquerda aparenta ter um formato mais parecido com uma rede convencional como as das Subseções 4.1.1 e 4.1.2. Em geral, as redes mostram que os elementos que possuem a maior magnitude de expressão positiva também são aqueles que apresentam a maior magnitude de expressão negativa. A Tabela 4.16 mostra esse dado.

Tabela 4.16: Top 10 *Diff* dos nodos da rede miRNA-mRNA para RNA-seq. A variável "*Diff*" é determinada como o módulo da diferença entre os valores de "Fator de Indução" e de "Fator de Repressão", que posteriormente são convertidos para uma escala de 1 a 0. A tabela foi ordenada para demonstrar os genes com os 10 maiores valores de *Diff*.

Nodos	Fator de Indução	Fator de Repressão	*Diff*
AR	0.66	0.95	1.00
hsa-miR-133b	0.30	0.05	0.79
hsa-miR-665	0.34	0.11	0.72
hsa-miR-204-5p	0.68	0.61	0.17
ARHGDIG	1.00	1.00	0.09
FKBP4	0.18	0.20	0.07
FKBP5	0.32	0.34	0.06
TP53	0.14	0.16	0.06
PTGES3	0.03	0.02	0.05
RAC1	0.01	0.00	0.05

Pode-se observar, na tabela acima, que poucos nodos têm um *Diff* maior que 0.1, portanto a maior parte deles têm o "Fator de Indução" e o "Fator de Repressão" bem próximos. O gene AR (*androgen receptor*) foi previamente descrito como negativamente regulado na Subseção 4.2.1, sendo um fator de transcrição ativado por hormônios esteroides. A Tabela 4.15 mostra que esse é o gene com maior regulação negativa entre os nodos da rede, sugerindo a presença de amostras negativadas em relação à expressão do AR (AR-). Nas redes ilustradas nas Figuras 4.10 e 4.11 o AR não apresenta ligação direta com nenhum miRNA, entretanto o mesmo está ligado ao miRNA has-miR-204-5p por meio do gene NR3C1.

O NR3C1 é um gene que age de duas maneiras: como um fator de transcrição mediado por elementos de respostas ao glucocorticoide e como modulador de fatores de transcrição associados à processos inflamatórios, proliferação celular e diferenciação de tecidos. Este gene não aparece na Tabela 4.16 acima, mostrando não haver uma grande diferença entre seus módulos positivo e negativo. Essa situação mostra um relação importante entre esses

3 nodos (AR, NR3C1 e hsa-miR-204-5p), pois as redes nas Figuras 4.10 e 4.11 mostram como os seus módulos de expressão estão sempre significantemente elevados em relação ao restante.

A rede da esquerda ilustrada nas figuras 4.10 e 4.11 mostra uma situação parecida com essa, envolvendo os nodos hsa-miR-665, hsa-miR-133b e ARHGDIG. Os miRNAs hsa-miR-665 e hsa-miR-133b possuem um *Diff* consideravelmente maior que o já mencionado hsa-miR-204-5p, sugerindo que estes se encontram positivamente regulados nas amostras.

O gene ARHGDIG é um nodo bem aparente, pois possui os maiores módulos de "Fator de Indução" e de "Fator de Repressão" entre todos os nodos, entretanto seu valor de *Diff* é muito baixo, sugerindo não estar majoritariamente positiva ou negativamente regulado segundo as amostras. Isso apenas sugere a existência de diferentes subgrupos de amostras nos dados, um com expressão positiva do ARHGDIG e outro com expressão negativa, entretanto, não foram encontrados estudos que suportem essa sugestão. Portanto, existe aqui uma outra oportunidade para um futuro estudo, onde o objetivo seria avaliar o que estes subgrupos de expressão do ARHGDIG representam clinicamente (estágio prognóstico,etc...).

É importante salientar que o módulo de expressão dos miRNAs hsa-miR-665 e hsa-miR-133b também acompanham a expressão do ARHGDIG. Em geral, os miRNAs hsa-miR-665 e hsa-miR-133b têm um "Fator de indução" maior que o "Fator de Repressão", como mostra a variável *Diff*. Estes mesmos miRNAs apresentam um "Fator de Repressão" de baixo módulo quando o ARHGDIG também apresenta. Essa relação entre os nodos ARHGDIG, hsa-mir665 e hsa-miR-133b também não aparenta possuir muitas informações para consulta na literatura. Assim sendo, esta também se torna uma observação que merece estudos adicionais. Isso finaliza a análise de redes dos miRNAs a partir os dados de RNA-seq, a análise prossegue com as redes de miRNAs segundo os dados de metilação.

Foi seguido o protocolo descrito na Subseção 3.3.4.2 para definir redes de interação miRNA-mRNA para os miRNAs de interesse: hsa-miR-133b, hsa-miR-138-5p, hsa-miR-665, hsa-miR-204-5p, hsa-miR-10a-5p e hsa-miR-33b-5p. Após a prospecção das relações miRNA-mRNA, são selecionadas apenas aquelas cujos mRNAs aparecem nas redes dos dados de metilação, ilustradas nas Figuras 4.8 e 4.9.

Depois disso as relações miRNA-mRNA filtradas são concatenadas às relações mRNA-mRNA de sua respectiva rede (metilação, no caso) cujos mRNAs estejam presentes nas rela-

ções miRNA-mRNA. Assim, foi obtida uma rede com interações miRNA-mRNA e mRNA-mRNA segundo dados que foram obtidos posteriormente. Abaixo, temos as Figuras 4.12 e 4.13 com a rede referente aos dados miRNA-mRNA construída com auxílio da rede de metilação.

Rede miRNA-mRNA (Metilação) dos miRNAs significantes (Expressão Positiva)

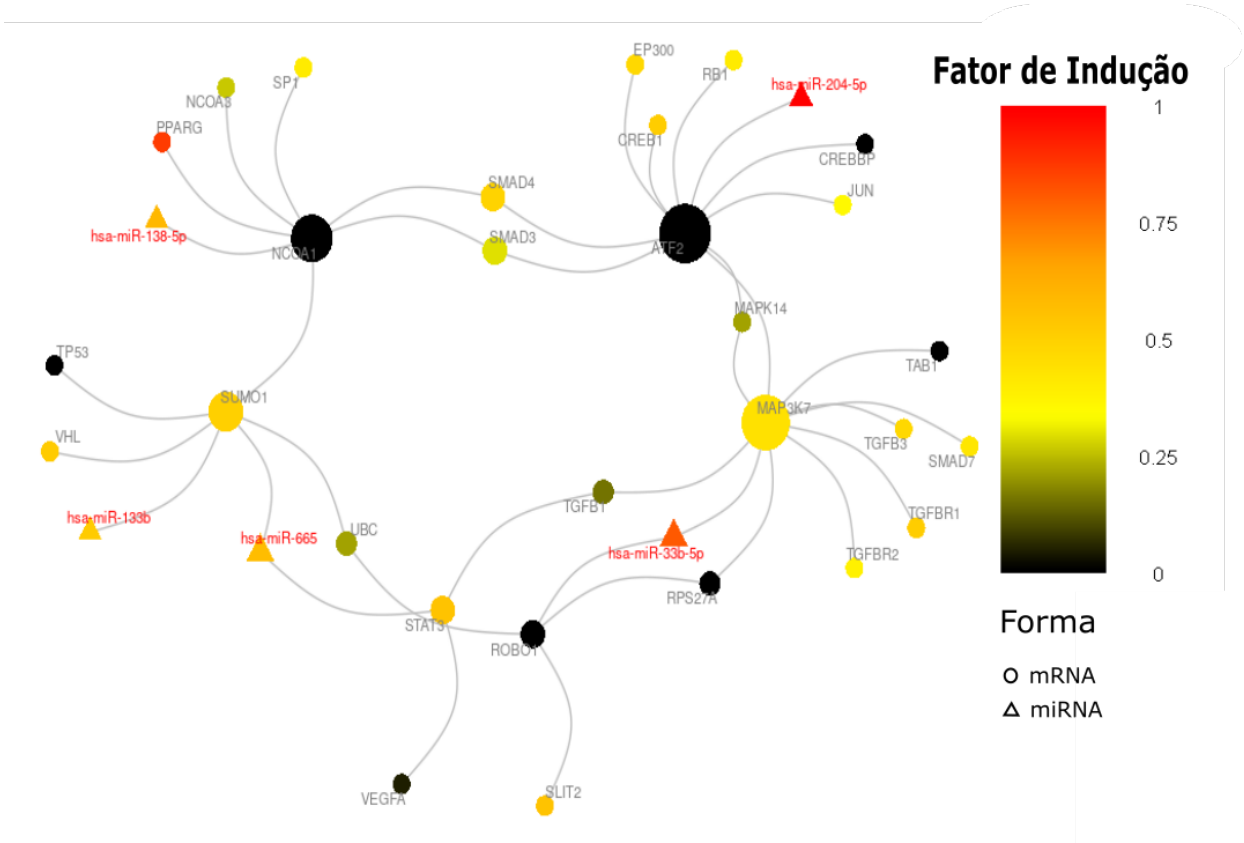


Figura 4.12: Rede miRNA-mRNA dos miRNAs significantes (Metilação Positiva). A rede é obtida ao utilizar os 6 primeiros miRNAs descritos na Tabela 4.14 no banco de dados *Starbase ENCORI*, seguindo o protocolo descrito na Subseção 3.3.4.2. A escala de cores, aqui chamada de "Fator de Indução", é construída a partir do conjunto de medianas das metilações relativas (também definida como ER) positivas dos nodos presentes na rede, considerando apenas o conjunto de dados do BRCA, a definição de ER é fornecida na Subseção 3.2.6. Assim a escala reflete o módulo da metilação positiva (hipermetilação) do conjunto, onde 1 (vermelho) representa módulo elevado e 0 (preto) módulo baixo. O gradiente de cores entre esses valores tem a função de facilitar a comparação visual com sua análoga para expressão negativa abaixo. Note que o tamanho dos nodos é ajustado pelo parâmetro *betweenness* do nodo.

Rede miRNA-mRNA (Metilação) dos miRNAs significantes (Expressão Negativa)

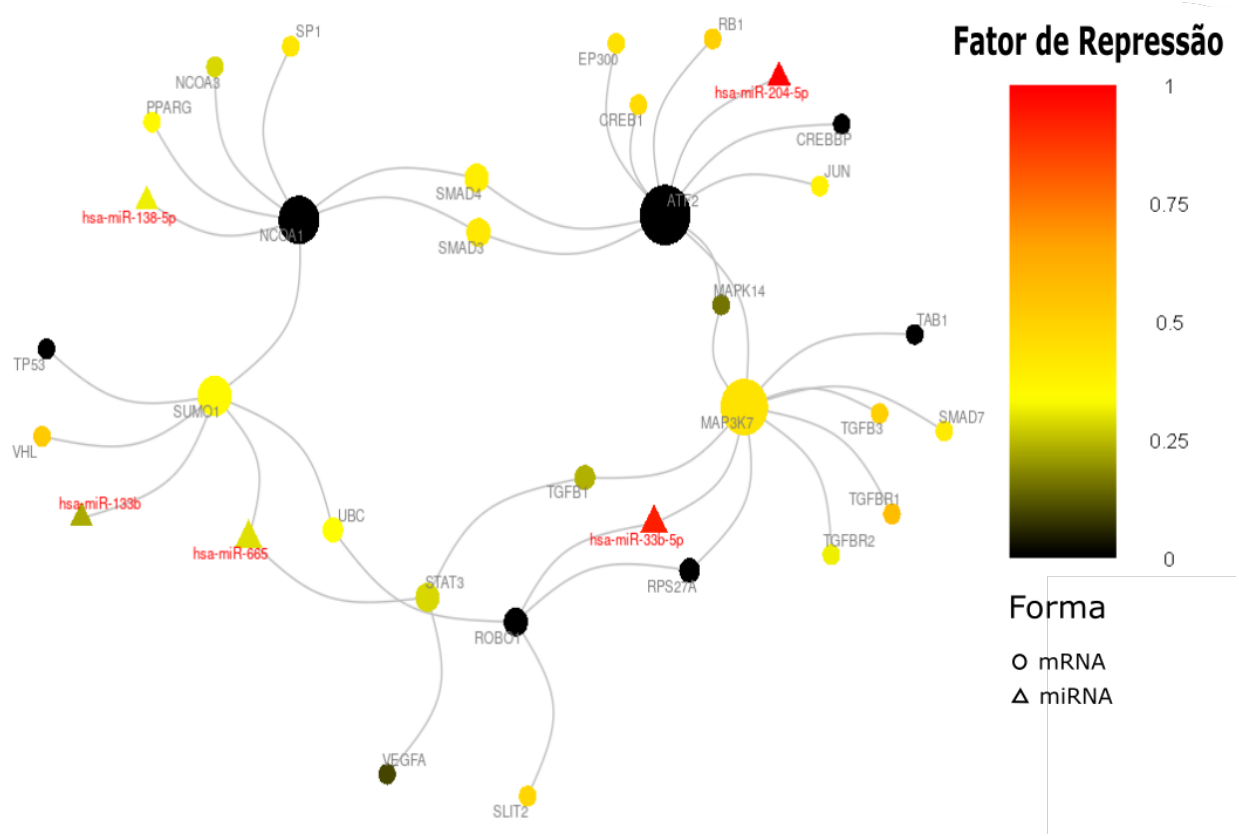


Figura 4.13: Rede miRNA-mRNA dos miRNAs significantes (Metilação Negativa). A rede é obtida ao utilizar os 6 primeiros miRNAs descritos na Tabela 4.14 no banco de dados *Starbase ENCORI*, seguindo o protocolo descrito na Subseção 3.3.4.2. A escala de cores, aqui chamada de "Fator de Repressão", é construída a partir do conjunto de medianas das metilações relativas (também definida como ER) negativas dos nodos presentes na rede, considerando apenas o conjunto de dados do BRCA, a definição de ER é fornecida na Subseção 3.2.6. Assim a escala reflete o módulo da metilação negativa (hipometilação) do conjunto, onde 1 (vermelho) representa módulo elevado e 0 (preto) módulo baixo. O gradiente de cores entre esses valores tem a função de facilitar a comparação visual com sua análoga para expressão positiva acima. Note que o tamanho dos nodos é ajustado pelo parâmetro *betweenness* do nodo.

As Figuras 4.12 e 4.13 mostram uma única rede ao contrário da rede que foi derivada dos dados de RNA-seq, ilustradas nas Figuras 4.10 e 4.11. A rede miRseq-mRNA (metilação) é composta por 4 principais *hubs* ligados de maneira escassa por alguns nodos. O núcleo desses 4 *hubs* é composto pelos genes NCOA1, ATF2, SUMO1 e MAP3K7. Para prosseguir com a análise de expressão relativa, observe a Tabela 4.17 abaixo:

Assim como a tabela 4.16, a tabela 4.17 acima apresenta 3 miRNAs, com valor de *Diff* relativamente elevado. Quase todos os nodos selecionados têm um "Fator de Indução" maior que o "Fator de Repressão", indicando um deslocamento da metilação para hipermetilação geral, embora os nodos UBC e SMAD3 apresentem evidências de hipometilação geral.

Tabela 4.17: *Diff* dos nodos da rede miRNA-mRNA para metilação. A variável "*Diff*" é determinada como o módulo da diferença entre o "Fator de Indução" e o "Fator de Repressão", que posteriormente são convertidos para uma escala de 1 a 0. A tabela foi ordenada para demonstrar os genes com os 10 maiores valores de *Diff*.

Nodos	Expressão Positiva	Expressão Negativa	*Diff*
PPARG	0.87	0.36	1.00
hsa-miR-133b	0.53	0.22	0.59
hsa-miR-138-5p	0.60	0.31	0.57
hsa-miR-665	0.57	0.30	0.54
STAT3	0.55	0.29	0.52
SUMO1	0.50	0.36	0.31
SMAD4	0.49	0.40	0.22
UBC	0.21	0.34	0.19
SMAD3	0.29	0.42	0.18
SLIT2	0.55	0.49	0.18

Os miRNAs hsa-miR-204-5p e hsa-miR-33b-5p aparecem na rede ilustrada nas Figuras 4.12 e 4.13, mas não na Tabela 4.17. Isso ocorre pois a variação entre o "Fator de Indução" e o "Fator de Repressão" é muito pequena, e a tabela mencionada contém os nodos com maior variação entre os módulos, mensurada pela variável *Diff*.

Pode-se observar que o lado esquerdo da rede tem vários nodos com forte metilação positiva e uma metilação negativa mais fraca ("Fator de Indução" > "Fator de Repressão"). Nesta observação, estão justamente os 3 miRNAs da Tabela 4.17: hsa-miR-138-5p, hsa-miR-133b e hsa-miR-665. O hsa-miR-133b, por exemplo, foi mencionado ter sua expressão associada ao estágio tumoral de maneira inversamente proporcional, também foi mencionado que tumores mamários tipicamente apresentam uma subexpressão desse miRNA.

Nossa rede de metilação ilustrada nas Figuras 4.12 e 4.13 mostram que a expressão positiva do hsa-miR-133b está associada a hipermetilação de genes como PPARG, SUMO1 e STAT3. O gene SUMO1 é especial, neste caso, pois além da sua própria hipermetilação, pode-se observar que este liga dois dos miRNAs de interesse: hsa-miR-133b e hsa-miR-665. O hsa-miR-665 também foi mencionado anteriormente como um miRNA cuja superexpressão constitui um prognóstico ruim para o câncer mamário, e de fato os dados sugerem que o miRNA tem uma elevada expressão positiva. Tal expressão está, segundo a rede e aos dados de expressão, correlacionada a hipermetilação dos genes SUMO1 e STAT3.

O gene SUMO1 (*small ubiquitin-like modifier*) ainda não foi discutido no trabalho. Este

gene codifica uma proteína que tem função similar às proteínas ubiquitinas, ligando-se a outras proteínas com a finalidade de marcá-las para mudanças pós-transcricionais. Entretanto a proteína SUMO1, não está associada à degradação molecular, como as ubiquitinas, a SUMO1 está associada a outros processos celulares que incluem regulação transcricional, transporte nuclear, apoptose e estabilidade proteica. A hipermetilação geral do SUMO1 nos leva a presumir uma expressão reduzida do mesmo, embora os dados de RNA-seq na Subseção 4.1.1 não tenham sequer identificado a expressão do SUMO1 como significante. Informações sobre a metilação do SUMO1 são escassas, assim como potenciais relações que o mesmo pode ter com os miRNAs hsa-miR-133b e hsa-miR-665. Portanto, essa´é uma relação de interesse para estudos futuros, considerando tanto o papel regulatório dos miRNAs quanto as diversos processos biológicos mediados pelo SUMO1.

O gene STAT3, foi mencionado na Subseção 4.1.2, pertence a uma família de proteínas que atuam como fatores de transcrição para citocinas e fatores de crescimento. Um estudo por MA e colaboradores (2020), menciona que este gene costuma estar superexpresso em câncer mamário, no entanto, os dados na Subseção 4.1.2 sugerem que este se encontra positivamente metilado (hipermetilado), o que nos leva a presumir uma expressão reduzida/silenciada (MA; QIN; LI, 2020). É difícil comentar mais sobre a expressão desse gene, pois, a Subseção 4.1.1 não o identifica como um gene de interesse. Entretanto, a hipermetilação do STAT3 é confirmada novamente nas redes ilustradas nas Figuras 4.12 e 4.13 assim como uma relação direta entre a expressão positiva do hsa-miR-665 e a hipermetilação do STAT3.

5 Conclusão

O mapa modular de Eran Segal foi adaptado com sucesso para uso em 32 tipos de tumores e utilizando 3 tipos de dados: RNA-seq, metilação e miRseq. Por meio de análise estatística e visual a metodologia foi capaz de determinar vias biológicas de importância geral nos 32 tumores. Tais achados envolvem vias como: síntese e transporte de lipídios, regulação da metilação, fibrinólise, regulação da atividade transcricional, transição epitélio-mesênquima, homeostase iônica, regulação de receptores de células do sistema imune e regulação de vias de sinalização mediadas por proteínas tirosina cinase.

Ainda foi feito um estudo de caso com os dados do tumor mamário (BRCA), onde foi demonstrado que os dados gerados com a criação do mapa modular ainda podem ser utilizados para analises mais especificas.

Em geral, os dados de RNA-seq mostram a importância de vias associadas aos processos de angiogênese e regulação transcricional para o BRCA. Os dados de metilação mostram também a importância da regulação da vascularização pelo sistema RAS assim como regulação de processos associados a diferenciação tecidual, síntese e transporte de lipídios, fibrinólise e regulação pós-transcrional. Por fim, os dados de miRseq nos sugerem que no BRCA os miRNAs hsa-miR-133b, hsa-miR-665 e hsa-miR-138-5p parecem ter uma expressão positivamente regulada. Tal expressão pode estar correlacionada a expressão de genes (AR e ARHGDIG) e também com a metilação de outros (PPARG, SUMO1 e STAT3).

É importante ressaltar que essa é uma metodologia de análise exploratória alternativa desenvolvida para reaproveitamento de grandes quantidades de dados biológicos, especialmente quando esses dados não apresentem paridade caso/controle. Saliento, que o método desenvolvido não substitui o padrão ouro que considera a paridade caso/controle, ao invés disso deve ser encarado como uma análise complementar para buscar novos caminhos para pesquisa científica. No entanto, até o momento, a metodologia se apresentou eficiente, tendo sido capaz de detectar processos biológicos importantes para vários tipos de tumor, determinar se sua respectiva expressão/metilação está induzida ou reprimida e ainda foi capaz identificar miRNAs de interesse para o tumor mamário. Assim pode-se concluir que a metodologia desenvolvida é capaz de cumprir sua função como ferramenta de análise exploratória.

Referências Bibliográficas

ADAMS, J. et al. Vascular endothelial growth factor (vegf) in breast cancer: comparison of plasma, serum, and tissue vegf and microvessel density and effects of tamoxifen. *Cancer research*, AACR, v. 60, n. 11, p. 2898–2905, 2000.

AGARWAL, V. et al. Predicting effective microrna target sites in mammalian mrnas. *elife*, eLife Sciences Publications Limited, v. 4, p. e05005, 2015.

AL-SHARIF, I.; REMMAL, A.; ABOUSSEKHRA, A. Eugenol triggers apoptosis in breast cancer cells through e2f1/survivin down-regulation. *BMC cancer*, BioMed Central, v. 13, n. 1, p. 1–10, 2013.

ALLAIN, E. P. et al. Emerging roles for udp-glucuronosyltransferases in drug resistance and cancer progression. *British journal of cancer*, Nature Publishing Group, v. 122, n. 9, p. 1277–1287, 2020.

ANG, J. C. et al. Supervised, unsupervised, and semi-supervised feature selection: a review on gene selection. *IEEE/ACM transactions on computational biology and bioinformatics*, IEEE, v. 13, n. 5, p. 971–989, 2015.

ASHBURNER, M. et al. Gene ontology: tool for the unification of biology. *Nature genetics*, Nature Publishing Group, v. 25, n. 1, p. 25–29, 2000.

ASTANEHE, A. et al. Mknk1 is a yb-1 target gene responsible for imparting trastuzumab resistance and can be blocked by rsk inhibition. *Oncogene*, Nature Publishing Group, v. 31, n. 41, p. 4434–4446, 2012.

AUGIMERI, G. et al. The role of pparγ ligands in breast cancer: from basic research to clinical studies. *Cancers*, Multidisciplinary Digital Publishing Institute, v. 12, n. 9, p. 2623, 2020.

AYER, T. et al. Breast cancer risk estimation with artificial neural networks revisited: discrimination and calibration. *Cancer*, Wiley Online Library, v. 116, n. 14, p. 3310–3321, 2010.

BACH, D.-H.; PARK, H. J.; LEE, S. K. The dual role of bone morphogenetic proteins in cancer. *Molecular Therapy-Oncolytics*, Elsevier, v. 8, p. 1–13, 2018.

BAREL, G.; HERWIG, R. Network and pathway analysis of toxicogenomics data. *Frontiers in genetics*, Frontiers Media SA, v. 9, p. 484, 2018.

BELLAZZO, A.; MININ, G. D.; COLLAVIN, L. Block one, unleash a hundred. mechanisms of dab2ip inactivation in cancer. *Cell Death & Differentiation*, Nature Publishing Group, v. 24, n. 1, p. 15–25, 2017.

BEWICK, V.; CHEEK, L.; BALL, J. Statistics review 7: Correlation and regression. *Critical care*, Springer, v. 7, n. 6, p. 451, 2003.

BEWICK, V.; CHEEK, L.; BALL, J. Statistics review 14: Logistic regression. *Critical care*, Springer, v. 9, n. 1, p. 112, 2005.

BHATTACHARYA, S. et al. Anti-tumorigenic effects of type 1 interferon are subdued by integrated stress responses. *Oncogene*, Nature Publishing Group, v. 32, n. 36, p. 4214–4221, 2013.

BIAGI-JUNIOR, C. A. O. de. Meta-análise do projeto toxicogenêmico japonês: diferenças entre modelos in vivo e in vitro.

BIANCHI, M. et al. Induction of ubiquitin c (ubc) gene transcription is mediated by hsf 1: role of proteotoxic and oxidative stress. *FEBS Open Bio*, Wiley Online Library, v. 8, n. 9, p. 1471–1485, 2018.

BIRD, A. The essentials of dna methylation. *Cell*, Cell Press, v. 70, n. 1, p. 5–8, 1992.

BUDHWANI, M.; MAZZIERI, R.; DOLCETTI, R. Plasticity of type i interferon-mediated responses in cancer therapy: from anti-tumor immunity to resistance. *Frontiers in oncology*, Frontiers, p. 322, 2018.

BUI, Q. T. et al. Ubiquitin-conjugating enzymes in cancer. *Cells*, Multidisciplinary Digital Publishing Institute, v. 10, n. 6, p. 1383, 2021.

BURGER, J. A.; WIESTNER, A. Targeting b cell receptor signalling in cancer: preclinical and clinical advances. *Nature reviews Cancer*, Nature Publishing Group, v. 18, n. 3, p. 148–167, 2018.

CAPRA, M. et al. Frequent alterations in the expression of serine/threonine kinases in human cancers. *Cancer research*, AACR, v. 66, n. 16, p. 8147–8154, 2006.

CHENG, Q. et al. Amplification and high-level expression of heat shock protein 90 marks aggressive phenotypes of human epidermal growth factor receptor 2 negative breast cancer. *Breast cancer research*, Springer, v. 14, n. 2, p. 1–15, 2012.

CHENG, W.; HÜLLERMEIER, E. Combining instance-based learning and logistic regression for multilabel classification. *Machine Learning*, Springer, v. 76, n. 2-3, p. 211–225, 2009.

CHOI, S. et al. Bmp-4 enhances epithelial mesenchymal transition and cancer stem cell properties of breast cancer cells via notch signaling. *Scientific reports*, Nature Publishing Group, v. 9, n. 1, p. 1–14, 2019.

CRUCERIU, D. et al. The dual role of tumor necrosis factor-alpha (tnf-α) in breast cancer: molecular insights and therapeutic approaches. *Cellular Oncology*, Springer, v. 43, n. 1, p. 1–18, 2020.

CRUZ, J. A.; WISHART, D. S. Applications of machine learning in cancer prediction and prognosis. *Cancer informatics*, SAGE Publications Sage UK: London, England, v. 2, p. 117693510600200030, 2006.

CRUZ, P. M. et al. The role of cholesterol metabolism and cholesterol transport in carcinogenesis: a review of scientific findings, relevant to future cancer therapeutics. *Frontiers in pharmacology*, Frontiers, v. 4, p. 119, 2013.

CUI, J. et al. Mbd3 mediates epigenetic regulation on epas1 promoter in cancer. *Tumor Biology*, Springer, v. 37, n. 10, p. 13455–13467, 2016.

DALLOL, A. et al. Slit2, a human homologue of the drosophila slit2 gene, has tumor suppressor activity and is frequently inactivated in lung and breast cancers. *Cancer research*, AACR, v. 62, n. 20, p. 5874–5880, 2002.

DALMA-WEISZHAUSZ, D. D. et al. [1] the affymetrix genechip® platform: An overview. *Methods in enzymology*, Elsevier, v. 410, p. 3–28, 2006.

DAS, P. M.; SINGAL, R. Dna methylation and cancer. *Journal of clinical oncology*, American Society of Clinical Oncology, v. 22, n. 22, p. 4632–4642, 2004.

DASHTI, S. et al. In silico identification of mapk14-related lncrnas and assessment of their expression in breast cancer samples. *Scientific reports*, Nature Publishing Group, v. 10, n. 1, p. 1–12, 2020.

ENRIGHT, A. et al. Microrna targets in drosophila. *Genome biology*, Springer, v. 4, n. 11, p. 1–27, 2003.

EXARCHOS, K. P.; GOLETSIS, Y.; FOTIADIS, D. I. Multiparametric decision support system for the prediction of oral cancer reoccurrence. *IEEE Transactions on Information Technology in Biomedicine*, Institute of Electrical and Electronics Engineers, Inc., 3 Park Avenue, 17 th Fl New York NY 10016-5997 United States, v. 16, n. 6, p. 1127–1134, 2012.

FRIDMAN, W. H. et al. B cells and cancer: To b or not to b? *Journal of Experimental Medicine*, The Rockefeller University Press, v. 218, n. 1, 2021.

GELENS, L.; SAURIN, A. T. Exploring the function of dynamic phosphorylation-dephosphorylation cycles. *Developmental cell*, Elsevier, v. 44, n. 6, p. 659–663, 2018.

GILMORE, S.; HOFMANN-WELLENHOF, R.; SOYER, H. P. A support vector machine for decision support in melanoma recognition. *Experimental dermatology*, Wiley Online Library, v. 19, n. 9, p. 830–835, 2010.

GOLDBERG, A. D.; ALLIS, C. D.; BERNSTEIN, E. Epigenetics: a landscape takes shape. *Cell*, Elsevier, v. 128, n. 4, p. 635–638, 2007.

GOLDSTEIN, I.; SPATT, C. S.; YE, M. Big data in finance. *The Review of Financial Studies*, Oxford University Press, v. 34, n. 7, p. 3213–3225, 2021.

GOOSSENS, P. et al. Membrane cholesterol efflux drives tumor-associated macrophage reprogramming and tumor progression. *Cell metabolism*, Elsevier, v. 29, n. 6, p. 1376–1389, 2019.

GRUBER, A. J.; ZAVOLAN, M. Alternative cleavage and polyadenylation in health and disease. *Nature Reviews Genetics*, Nature Publishing Group, v. 20, n. 10, p. 599–614, 2019.

HAASE, M.; FITZE, G. Hsp90ab1: Helping the good and the bad. *Gene*, Elsevier, v. 575, n. 2, p. 171–186, 2016.

HAIMAN, C. A. et al. A common variant at the tert-clptm1l locus is associated with estrogen receptor–negative breast cancer. *Nature genetics*, Nature Publishing Group, v. 43, n. 12, p. 1210–1214, 2011.

HANAHAN, D.; WEINBERG, R. A. Hallmarks of cancer: the next generation. *cell*, Elsevier, v. 144, n. 5, p. 646–674, 2011.

HANSEN, K. Illuminahumanmethylation450kanno. ilmn12. hg19: annotation for illumina's 450k methylation arrays. *R package, version 0.2*, v. 1, 2015.

HASSEN, C. B. et al. Apolipoprotein-mediated regulation of lipid metabolism induces distinctive effects in different types of breast cancer cells. *Breast Cancer Research*, BioMed Central, v. 22, n. 1, p. 1–17, 2020.

HAUPT, S. et al. The role of mdm2 and mdm4 in breast cancer development and prevention. *Journal of molecular cell biology*, Oxford University Press, v. 9, n. 1, p. 53–61, 2017.

HELLEDAY, T. et al. Dna double-strand break repair: from mechanistic understanding to cancer treatment. *DNA repair*, Elsevier, v. 6, n. 7, p. 923–935, 2007.

HILBERT, M. Big data for development: A review of promises and challenges. *Development Policy Review*, Wiley Online Library, v. 34, n. 1, p. 135–174, 2016.

HOEN, P. A. t et al. Fluorescent labelling of crna for microarray applications. *Nucleic acids research*, Oxford University Press, v. 31, n. 5, p. e20–e20, 2003.

HOLLERN, D. P. et al. E2f1 drives breast cancer metastasis by regulating the target gene fgf13 and altering cell migration. *Scientific reports*, Nature Publishing Group, v. 9, n. 1, p. 1–13, 2019.

HONG, B. S. et al. Tumor suppressor mirna-204-5p regulates growth, metastasis, and immune microenvironment remodeling in breast cancer. *Cancer research*, AACR, v. 79, n. 7, p. 1520–1534, 2019.

HUANG, P. et al. Bmp-2 induces emt and breast cancer stemness through rb and cd44. *Cell death discovery*, Nature Publishing Group, v. 3, n. 1, p. 1–12, 2017.

HUANG, S. et al. *TBX1 functions as a potential oncogene in breast cancer*. [S.l.]: AACR, 2020.

HUANG, Y. et al. The behaviour of 5-hydroxymethylcytosine in bisulfite sequencing. *PloS one*, Public Library of Science, v. 5, n. 1, p. e8888, 2010.

IJUIN, T. Phosphoinositide phosphatases in cancer cell dynamic-beyond pi3k and pten. In: ELSEVIER. *Seminars in Cancer Biology*. [S.l.], 2019. v. 59, p. 50–65.

JENSEN, L. J. et al. String 8–a global view on proteins and their functional interactions in 630 organisms. *Nucleic acids research*, Oxford Univ Press, v. 37, n. suppl 1, p. D412–D416, 2009.

JIANG, Z. et al. Rb1 and p53 at the crossroad of emt and triple-negative breast cancer. *Cell cycle*, Taylor & Francis, v. 10, n. 10, p. 1563–1570, 2011.

JOLLIFFE, I. T.; CADIMA, J. Principal component analysis: a review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, The Royal Society Publishing, v. 374, n. 2065, p. 20150202, 2016.

KERTESZ, M. et al. The role of site accessibility in microrna target recognition. *Nature genetics*, Nature Publishing Group, v. 39, n. 10, p. 1278–1284, 2007.

KHURANA, N.; BHATTACHARYYA, S. Hsp90, the concertmaster: tuning transcription. *Frontiers in oncology*, Frontiers, v. 5, p. 100, 2015.

KODINARIYA, T. M.; MAKWANA, P. R. Review on determining number of cluster in k-means clustering. *International Journal*, v. 1, n. 6, p. 90–95, 2013.

KOUROU, K. et al. Machine learning applications in cancer prognosis and prediction. *Computational and structural biotechnology journal*, Elsevier, v. 13, p. 8–17, 2015.

KOZOMARA, A.; GRIFFITHS-JONES, S. mirbase: annotating high confidence micrornas using deep sequencing data. *Nucleic acids research*, Oxford University Press, v. 42, n. D1, p. D68–D73, 2013.

KREK, A. et al. Combinatorial microrna target predictions. *Nature genetics*, Nature Publishing Group, v. 37, n. 5, p. 495–500, 2005.

KUMAR, R.; SINGH, V. P.; BAKER, K. M. The intracellular renin–angiotensin system: a new paradigm. *Trends in Endocrinology & Metabolism*, Elsevier, v. 18, n. 5, p. 208–214, 2007.

KUZU, O. F.; NOORY, M. A.; ROBERTSON, G. P. The role of cholesterol in cancer. *Cancer research*, AACR, v. 76, n. 8, p. 2063–2070, 2016.

KWAAN, H. C.; LINDHOLM, P. F. Fibrin and fibrinolysis in cancer. In: THIEME MEDICAL PUBLISHERS. *Seminars in thrombosis and hemostasis*. [S.l.], 2019. v. 45, n. 04, p. 413–422.

LADD, A. et al. Differential roles of angiotensinogen and angiotensin receptor type 1 polymorphisms in breast cancer risk. *Breast cancer research and treatment*, Springer, v. 101, n. 3, p. 299–304, 2007.

LI, B. et al. Rhoa triggers a specific signaling pathway that generates transforming microvesicles in cancer cells. *Oncogene*, Nature Publishing Group, v. 31, n. 45, p. 4740–4749, 2012.

LI, J.-H. et al. starbase v2. 0: decoding mirna-cerna, mirna-ncrna and protein–rna interaction networks from large-scale clip-seq data. *Nucleic acids research*, Oxford University Press, v. 42, n. D1, p. D92–D97, 2013.

LI, J.-H. et al. starbase v2. 0: decoding mirna-cerna, mirna-ncrna and protein–rna interaction networks from large-scale clip-seq data. *Nucleic acids research*, Oxford University Press, v. 42, n. D1, p. D92–D97, 2014.

LI, Y.; WU, F.-X.; NGOM, A. A review on machine learning principles for multi-view biological data integration. *Briefings in bioinformatics*, Oxford University Press, v. 19, n. 2, p. 325–340, 2016.

LIN, Y. et al. Microrna-33b inhibits breast cancer metastasis by targeting hmga2, sall4 and twist1. *Scientific reports*, Nature Publishing Group, v. 5, n. 1, p. 1–12, 2015.

LISCOVITCH, M.; LAVIE, Y. Multidrug resistance: a role for cholesterol efflux pathways? *Trends in biochemical sciences*, Elsevier, v. 25, n. 11, p. 530–534, 2000.

LIU, W. et al. Dysregulated cholesterol homeostasis results in resistance to ferroptosis increasing tumorigenicity and metastasis in cancer. *Nature communications*, Nature Publishing Group, v. 12, n. 1, p. 1–15, 2021.

LOHER, P.; RIGOUTSOS, I. Interactive exploration of rna22 microrna target predictions. *Bioinformatics*, Oxford University Press, v. 28, n. 24, p. 3322–3323, 2012.

MA, J.-h.; QIN, L.; LI, X. Role of stat3 signaling pathway in breast cancer. *Cell Communication and Signaling*, Springer, v. 18, n. 1, p. 1–13, 2020.

MAHMOOD, N.; RABBANI, S. A. Fibrinolytic system and cancer: diagnostic and therapeutic applications. *International Journal of Molecular Sciences*, Multidisciplinary Digital Publishing Institute, v. 22, n. 9, p. 4358, 2021.

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

MARAGKAKIS, M. et al. Diana-microt web server: elucidating microrna functions through target prediction. *Nucleic acids research*, Oxford University Press, v. 37, n. suppl_2, p. W273–W276, 2009.

MARSH, S. et al. Increased expression of fibroblast growth factor 8 in human breast cancer. *Oncogene*, Nature Publishing Group, v. 18, n. 4, p. 1053–1060, 1999.

MARX, V. *Biology: The big challenges of big data*. [S.l.]: Nature Publishing Group, 2013.

MCBEATH, R. et al. Cell shape, cytoskeletal tension, and rhoa regulate stem cell lineage commitment. *Developmental cell*, Elsevier, v. 6, n. 4, p. 483–495, 2004.

MCKINNEY, S. M. et al. International evaluation of an ai system for breast cancer screening. *Nature*, Nature Publishing Group, v. 577, n. 7788, p. 89–94, 2020.

MERING, C. V. et al. String 7–recent developments in the integration and prediction of protein interactions. *Nucleic acids research*, Oxford Univ Press, v. 35, n. suppl 1, p. D358–D362, 2007.

MERING, C. V. et al. String: known and predicted protein–protein associations, integrated and transferred across organisms. *Nucleic acids research*, Oxford Univ Press, v. 33, n. suppl 1, p. D433–D437, 2005.

MOORE-SMITH, L.; PASCHE, B. Tgfbr1 signaling and breast cancer. *Journal of mammary gland biology and neoplasia*, Springer, v. 16, n. 2, p. 89–95, 2011.

MURTAGH, F.; CONTRERAS, P. Algorithms for hierarchical clustering: an overview. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, Wiley Online Library, v. 2, n. 1, p. 86–97, 2012.

NAKAI, K.; HUNG, M.-C.; YAMAGUCHI, H. A perspective on anti-egfr therapies targeting triple-negative breast cancer. *American journal of cancer research*, e-Century Publishing Corporation, v. 6, n. 8, p. 1609, 2016.

NASRABADI, N. M. Pattern recognition and machine learning. *Journal of electronic imaging*, International Society for Optics and Photonics, v. 16, n. 4, p. 049901, 2007.

NRC et al. *Applications of toxicogenomic technologies to predictive toxicology and risk assessment*. [S.l.]: National Academies Press (US), 2007.

OGI, H. et al. The oncogenic role of the cochaperone sgt1. *Oncogenesis*, Nature Publishing Group, v. 4, n. 5, p. e149–e149, 2015.

OKAE, H. et al. Genome-wide analysis of dna methylation dynamics during early human development. *PLoS genetics*, Public Library of Science San Francisco, USA, v. 10, n. 12, p. e1004868, 2014.

PHILLIPS, T. et al. The role of methylation in gene expression. *Nature Education*, v. 1, n. 1, p. 116, 2008.

QIAN, B.-Z. et al. Flt1 signaling in metastasis-associated macrophages activates an inflammatory signature that promotes breast cancer metastasis. *Journal of Experimental Medicine*, The Rockefeller University Press, v. 212, n. 9, p. 1433–1448, 2015.

RAHIM, B.; O'REGAN, R. Ar signaling in breast cancer. *Cancers*, Multidisciplinary Digital Publishing Institute, v. 9, n. 3, p. 21, 2017.

RASOOLNEZHAD, M. et al. Mirna-138–5p: A strong tumor suppressor targeting pd-l-1 inhibits proliferation and motility of breast cancer cells and induces apoptosis. *European Journal of Pharmacology*, Elsevier, v. 896, p. 173933, 2021.

RIBATTI, D.; TAMMA, R.; ANNESE, T. Epithelial-mesenchymal transition in cancer: a historical overview. *Translational oncology*, Elsevier, v. 13, n. 6, p. 100773, 2020.

RING, A.; KAUR, P.; LANG, J. E. Ep300 knockdown reduces cancer stem cell phenotype, tumor growth and metastasis in triple negative breast cancer. *BMC cancer*, Springer, v. 20, n. 1, p. 1–14, 2020.

ROCHE, J. *The epithelial-to-mesenchymal transition in cancer*. [S.l.]: Multidisciplinary Digital Publishing Institute, 2018. 52 p.

ROGER, S. et al. Voltage-gated sodium channels and cancer: is excitability their primary role? *Frontiers in pharmacology*, Frontiers, v. 6, p. 152, 2015.

SASIREKHA, K.; BABY, P. Agglomerative hierarchical clustering algorithm-a. *International Journal of Scientific and Research Publications*, Citeseer, v. 83, p. 83, 2013.

SCHUSTER, S. C. Next-generation sequencing transforms today's biology. *Nature methods*, Nature Publishing Group, v. 5, n. 1, p. 16–18, 2008.

SEGAL, E. et al. A module map showing conditional activity of expression modules in cancer. *Nature genetics*, Nature Publishing Group, v. 36, n. 10, p. 1090, 2004.

SHEN, S.-Q. et al. mir-204 regulates the biological behavior of breast cancer mcf-7 cells by directly targeting foxa1. *Oncology reports*, Spandidos Publications, v. 38, n. 1, p. 368–376, 2017.

SHIMODAIRA, H. Technical details of the multistep-multiscale bootstrap resampling. *arXiv preprint arXiv:1312.6354*, 2013.

SINWAR, D.; KAUSHIK, R. Study of euclidean and manhattan distance metrics using simple k-means clustering. *Int. J. Res. Appl. Sci. Eng. Technol*, v. 2, n. 5, p. 270–274, 2014.

SNELL, L. M.; MCGAHA, T. L.; BROOKS, D. G. Type i interferon in chronic virus infection and cancer. *Trends in immunology*, Elsevier, v. 38, n. 8, p. 542–557, 2017.

SORIA, G.; BEN-BARUCH, A. The inflammatory chemokines ccl2 and ccl5 in breast cancer. *Cancer letters*, Elsevier, v. 267, n. 2, p. 271–285, 2008.

SZKLARCZYK, D. et al. The string database in 2011: functional interaction networks of proteins, globally integrated and scored. *Nucleic acids research*, Oxford Univ Press, v. 39, n. suppl 1, p. D561–D568, 2011.

SZKLARCZYK, D. et al. String v10: protein–protein interaction networks, integrated over the tree of life. *Nucleic acids research*, Oxford University Press, v. 43, n. D1, p. D447–D452, 2014.

TAM, S.; TSAO, M.-S.; MCPHERSON, J. D. Optimization of mirna-seq data preprocessing. *Briefings in bioinformatics*, Oxford University Press, v. 16, n. 6, p. 950–963, 2015.

TAN, Y. et al. Tumor suppressor drd2 facilitates m1 macrophages and restricts nf-κb signaling to trigger pyroptosis in breast cancer. *Theranostics*, Ivyspring International Publisher, v. 11, n. 11, p. 5214, 2021.

TARCA, A. L. et al. Machine learning and its applications to biology. *PLoS Comput Biol*, Public Library of Science, v. 3, n. 6, p. e116, 2007.

TCGA. *The Cancer Genome Atlas (TCGA)*. jul. 2018. Disponível em: <https://portal-.gdc.cancer.gov>.

TIAN, F. et al. Reduction in smad2/3 signaling enhances tumorigenesis but suppresses metastasis of breast cancer cell lines. *Cancer research*, AACR, v. 63, n. 23, p. 8284–8292, 2003.

VEJNAR, C. E.; ZDOBNOV, E. M. Mirmap: comprehensive prediction of microrna target repression strength. *Nucleic acids research*, Oxford University Press, v. 40, n. 22, p. 11673–11683, 2012.

VINSON, G. P.; BARKER, S.; PUDDEFOOT, J. R. The renin–angiotensin system in the breast and breast cancer. *Endocrine-related cancer*, Society for Endocrinology, v. 19, n. 1, p. R1–R19, 2012.

VLEUGEL, M. M. et al. c-jun activation is associated with proliferation and angiogenesis in invasive breast cancer. *Human pathology*, Elsevier, v. 37, n. 6, p. 668–674, 2006.

WANG, Q.-Y. et al. Mir-133b targets sox9 to control pathogenesis and metastasis of breast cancer. *Cell death & disease*, Nature Publishing Group, v. 9, n. 7, p. 1–14, 2018.

WANG, Y.; ZHANG, G.; HAN, J. Hif1a-as2 predicts poor prognosis and regulates cell migration and invasion in triple-negative breast cancer. *Journal of cellular biochemistry*, Wiley Online Library, v. 120, n. 6, p. 10513–10518, 2019.

WANG, Z.; GERSTEIN, M.; SNYDER, M. Rna-seq: a revolutionary tool for transcriptomics. *Nature reviews genetics*, Nature Publishing Group, v. 10, n. 1, p. 57–63, 2009.

WANG, Z. et al. Improvements to the gastric cancer tumor-node-metastasis staging system based on computer-aided unsupervised clustering. *BMC cancer*, BioMed Central, v. 18, n. 1, p. 706, 2018.

WOO, Y. et al. A comparison of cdna, oligonucleotide, and affymetrix genechip gene expression microarray platforms. *Journal of biomolecular techniques: JBT*, The Association of Biomolecular Resource Facilities, v. 15, n. 4, p. 276, 2004.

WORST, T. S. et al. mir-10a-5p and mir-29b-3p as extracellular vesicle-associated prostate cancer detection markers. *Cancers*, MDPI, v. 12, n. 1, p. 43, 2019.

XING, Q. et al. Integrated analysis of differentially expressed profiles and construction of a competing endogenous long non-coding rna network in renal cell carcinoma. *PeerJ*, PeerJ Inc., v. 6, p. e5124, 2018.

XUN, J. et al. Cancer-derived exosomal mir-138-5p modulates polarization of tumor-associated macrophages through inhibition of kdm6b. *Theranostics*, Ivyspring International Publisher, v. 11, n. 14, p. 6847, 2021.

YAN, A. et al. Cholesterol metabolism in drug-resistant cancer. *International Journal of Oncology*, Spandidos Publications, 2020.

YASUOKA, H. et al. Neuropilin-2 expression in breast cancer: correlation with lymph node metastasis, poor prognosis, and regulation of cxcr4 expression. *BMC cancer*, Springer, v. 9, n. 1, p. 1–7, 2009.

ZALCMAN, G. et al. Rhogdi-3 is a new gdp dissociation inhibitor (gdi): identification of a non-cytosolic gdi protein interacting with the small gtp-binding proteins rhob and rhog. *Journal of Biological Chemistry*, ASBMB, v. 271, n. 48, p. 30366–30374, 1996.

ZHANG, C. et al. High nr2f2 transcript level is associated with increased survival and its expression inhibits tgf-β-dependent epithelial-mesenchymal transition in breast cancer. *Breast cancer research and treatment*, Springer, v. 147, n. 2, p. 265–281, 2014.

ZHANG, F. et al. A network medicine approach to build a comprehensive atlas for the prognosis of human cancer. *Briefings in bioinformatics*, Oxford University Press, v. 17, n. 6, p. 1044–1059, 2016.

ZHANG, H. et al. *Abstract A100: Downregulation of type 1 interferon receptor (IFNAR1) regulates the balance of regulatory T cells (Tregs) and cytotoxic T lymphocytes (CTLs) in tumor microenvironment*. [S.l.]: AACR, 2020.

ZHANG, K.-x. et al. Down-regulation of type i interferon receptor sensitizes bladder cancer cells to vesicular stomatitis virus-induced cell death. *International journal of cancer*, Wiley Online Library, v. 127, n. 4, p. 830–838, 2010.

ZHANG, L. et al. Androgen receptor, egfr, and brca1 as biomarkers in triple-negative breast cancer: a meta-analysis. *BioMed research international*, Hindawi, v. 2015, 2015.

ZHANG, Y. et al. Systemic analysis of the prognosis-associated alternative polyadenylation events in breast cancer. *Frontiers in Genetics*, Frontiers Media SA, v. 11, p. 590770, 2020.

ZHANG, Z. Introduction to machine learning: k-nearest neighbors. *Annals of translational medicine*, AME Publications, v. 4, n. 11, 2016.

ZHAO, X.-G. et al. mir-665 expression predicts poor survival and promotes tumor metastasis by targeting nr4a3 in breast cancer. *Cell death & disease*, Nature Publishing Group, v. 10, n. 7, p. 1–21, 2019.

ZHONG, S. et al. High-throughput illumina strand-specific rna sequencing library preparation. *Cold spring harbor protocols*, Cold Spring Harbor Laboratory Press, v. 2011, n. 8, p. pdb–prot5652, 2011.

ZIA, M. K. et al. The expression of the von hippel-lindau gene product and its impact on invasiveness of human breast cancer cells. *International journal of molecular medicine*, Spandidos Publications, v. 20, n. 4, p. 605–611, 2007.

ZIMMER, B. M. et al. Altered glucuronidation deregulates androgen dependent response profiles and signifies castration resistance in prostate cancer. *Oncotarget*, Impact Journals, LLC, v. 12, n. 19, p. 1886, 2021.

ZLOTORYNSKI, E. The short tail that wags the mrna. *Nature Reviews Molecular Cell Biology*, Nature Publishing Group, v. 19, n. 1, p. 2–3, 2018.