

RESSALVA

Atendendo solicitação do(a)
autor(a), o texto completo desta
tese será disponibilizado
somente a partir de 26/08/2024.

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 BIOCÊNCIAS 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

1.1	Dogma central da biologia molecular	p. 1
1.2	Ciências Ômicas	p. 2
1.3	Diagrama do conceito de <i>Big Data</i>	p. 4
1.4	Arquitetura de dados do GDC	p. 7
1.5	Métodos de Acesso aos Dados do GDC	p. 8
1.6	Página inicial do GDC	p. 9
1.7	Dados armazenados no TCGA	p. 11
1.8	Distribuição dos dados armazenados no TCGA	p. 13
1.9	Experimento de RNA-seq	p. 17
1.10	Experimento de microarranjo com dois canais (duas cores)	p. 19
1.11	Fluxograma de submissão de dados de RNA-seq no TCGA	p. 20
1.12	Metilação da Citosina	p. 22
1.13	Experimento de Metilação (Illumina)	p. 23
1.14	Aprendizado de Máquina e suas Subdivisões	p. 25
1.15	Exemplo de Regressão Linear	p. 26
1.16	Regressão Linear <i>vs</i> Regressão Logística	p. 28
1.17	Regressão Logística	p. 28
1.18	Distância Euclidiana	p. 30
1.19	Clusterização K- <i>means</i>	p. 32
1.20	Clusterização Hierárquica	p. 34

1.21	Cálculo das componentes principais	p. 35
1.22	Gráfico das 2 primeiras PCAs	p. 38
3.1	<i>Workflow</i> dos mapas modulares	p. 44
3.2	Fluxograma de Dados: Fase 1 (Prospecção de Dados)	p. 63
3.3	Fluxograma de Dados: Fase 2 (Pré-processamento)	p. 64
3.4	Fluxograma de Dados: Fase 3 (Concatenação de Dados)	p. 66
3.5	Fluxograma de Dados: Fase 4 (BPs com alterações Significantes)	p. 68
3.6	Fluxograma de Dados: Fase 5 (Matriz de Porcentagem de Expressão) . . .	p. 69
3.7	Fluxograma de Dados: Fase 6 (Clusterização Hierárquica)	p. 70
4.1	Legenda de cores do mapa modular de RNA-seq para 32 tumores	p. 74
4.2	<i>Heatmap</i> do mapa modular do RNA-seq para 32 tumores com clusterização modular interna	p. 75
4.3	Versão reduzida do mapa modular de RNA-seq	p. 77
4.4	Versão reduzida do mapa modular de metilação	p. 86
4.5	Versão reduzida do mapa modular de miRseq	p. 92
4.6	Rede PPI dos genes significantes (Expressão Positiva)	p. 103
4.7	Rede PPI dos genes significantes (Expressão Negativa)	p. 104
4.8	Rede PPI dos genes significantes (Metilação Positiva)	p. 115
4.9	Rede PPI dos genes significantes (Metilação Negativa)	p. 116
4.10	Rede miRNA-mRNA dos miRNAs significantes (Expressão Positiva)	p. 130
4.11	Rede miRNA-mRNA dos miRNAs significantes (Expressão Negativa) . . .	p. 131
4.12	Rede miRNA-mRNA dos miRNAs significantes (Metilação Positiva)	p. 134
4.13	Rede miRNA-mRNA dos miRNAs significantes (Expressão Negativa) . . .	p. 135

Lista de Tabelas

1.1	Programas do GDC	p. 10
1.2	Os 33 tumores do TCGA	p. 12
1.3	Tipos de dados contemplados pelo TCGA.	p. 14
1.4	Critérios de Similaridade em Clusterização	p. 33
1.5	Conjunto de dados <i>wdbc.data</i>	p. 37
1.6	Sumário de Componentes Principais	p. 37
4.1	RNA-seq: Módulos Seleccionados	p. 101
4.2	RNA-seq: <i>Overlaps</i> estatisticamente significantes	p. 101
4.3	RNA-seq: Top 10 Induzidos	p. 105
4.4	RNA-seq: Top 10 Reprimidos	p. 105
4.5	RNA-seq: Top 10 valores de <i>Betweenness</i>	p. 108
4.6	RNA-seq: Top 10 <i>Diff</i>	p. 108
4.7	Metilação: Módulos Seleccionados	p. 111
4.8	Metilação: <i>Overlaps</i> estatisticamente significantes	p. 112
4.9	Metilação: Top 10 Induzidos	p. 117
4.10	Metilação: Top 10 Reprimidos	p. 117
4.11	RNA-seq: Top 10 valores de <i>Betweenness</i>	p. 121
4.12	Metilação: Top 10 <i>Diff</i>	p. 122
4.13	miRseq: Módulos Seleccionados	p. 126
4.14	miRseq: <i>Overlaps</i> estatisticamente significantes	p. 126

4.15	Top 10 miRNAs com maior <i>Diff</i>	p. 127
4.16	Top 10 <i>Diff</i> dos nodos da rede miRNA-mRNA para RNA-seq	p. 132
4.17	Top 10 <i>Diff</i> dos nodos da rede miRNA-mRNA para metilação	p. 136

Sumário

Agradecimentos	
Resumo	
Abstract	
1 Introdução	p. 1
1.1 Ciências Ômicas e <i>Big Data</i>	p. 1
1.2 Câncer e Bases de Dados (GDC e TCGA)	p. 5
1.2.1 <i>Genomic Data Commons</i> (GDC)	p. 6
1.3 RNA-seq, miRseq e Metilação: <i>Design</i> Experimental	p. 16
1.3.1 RNA-Seq	p. 16
1.3.2 miRseq	p. 20
1.3.3 Metilação	p. 21
1.4 Aprendizado de Máquina (AM)	p. 24
1.5 Bioinformática: Aplicações de AM e Tendências	p. 38
1.6 Câncer: Integrando Dados Biológicos com Anotações Funcionais e Clínicas	p. 40
1.7 Justificativa	p. 41
2 Objetivos	p. 42
2.1 Objetivos Gerais	p. 42
2.2 Objetivos Específicos	p. 42
3 Material e Métodos	p. 43

3.1	O Mapa Modular	p. 43
3.2	Coleta e Pré-processamento de dados	p. 45
3.2.1	RNA-seq: Coleta e Pré-processamento	p. 45
3.2.2	Metilação: Coleta e Pré-processamento	p. 48
3.2.3	miRseq: Coleta e Pré-processamento	p. 49
3.2.4	Determinação de <i>Gene Sets: Gene Ontology</i> (GO)	p. 50
3.2.5	<i>CpG Sets</i> e <i>miRNA Sets</i>	p. 51
3.2.6	Pré-processamento e Elementos Significativamente Alterados	p. 52
3.2.7	Matriz de Porcentagem de Expressão: Determinação dos <i>arrays</i> em que a Alteração dos Elementos é Significante	p. 53
3.2.8	A Heurística do Novo Mapa Modular: Identificação de <i>Clusters</i> de <i>Sets</i>	p. 54
3.2.9	Definição dos Módulos	p. 56
3.2.10	Finalização do Mapa Modular	p. 57
3.3	Análise Final dos Mapas Modulares	p. 57
3.3.1	Avaliação da Clusterização Hierárquica	p. 57
3.3.2	Visualização dos Mapas Modulares	p. 58
3.3.3	Estudo de Caso com Auxílio de Redes	p. 58
3.3.4	Construção de Redes de Integração para Análise de Expressão	p. 59
3.3.4.1	STRING: Um Banco de Dados de Interações Proteína- Proteína (PPI)	p. 59
3.3.4.2	Starbase ENCORI: Um Banco de Dados de Interações miRNA-Alvo	p. 61
3.4	Sumarização da Metodologia: Fluxograma de Dados	p. 63
4	Resultados e Discussão	p. 71

4.1	Meta Análise da Clusterização Hierárquica e Visualização do Mapa Modular	p. 71
4.1.1	RNA-seq: Módulos Consistentemente Induzidos ou Reprimidos . . .	p. 78
4.1.2	Metilação: Módulos Consistentemente Induzidos ou Reprimidos . .	p. 85
4.1.3	miRseq: Módulos Consistentemente Induzidos ou Reprimidos	p. 92
4.1.4	Sumarização da Visualização	p. 97
4.2	Tumor de mama (BRCA): Um Estudo de Caso	p. 98
4.2.1	BRCA: Análise do RNA-seq	p. 101
4.2.2	BRCA: Análise da Metilação	p. 111
4.2.3	BRCA: Análise do miRseq	p. 126
5	Conclusão	p. 138
	Referências Bibliográficas	p. 140

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 significativamente 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.

Dogma central da biologia molecular



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.

Ciências Ômicas

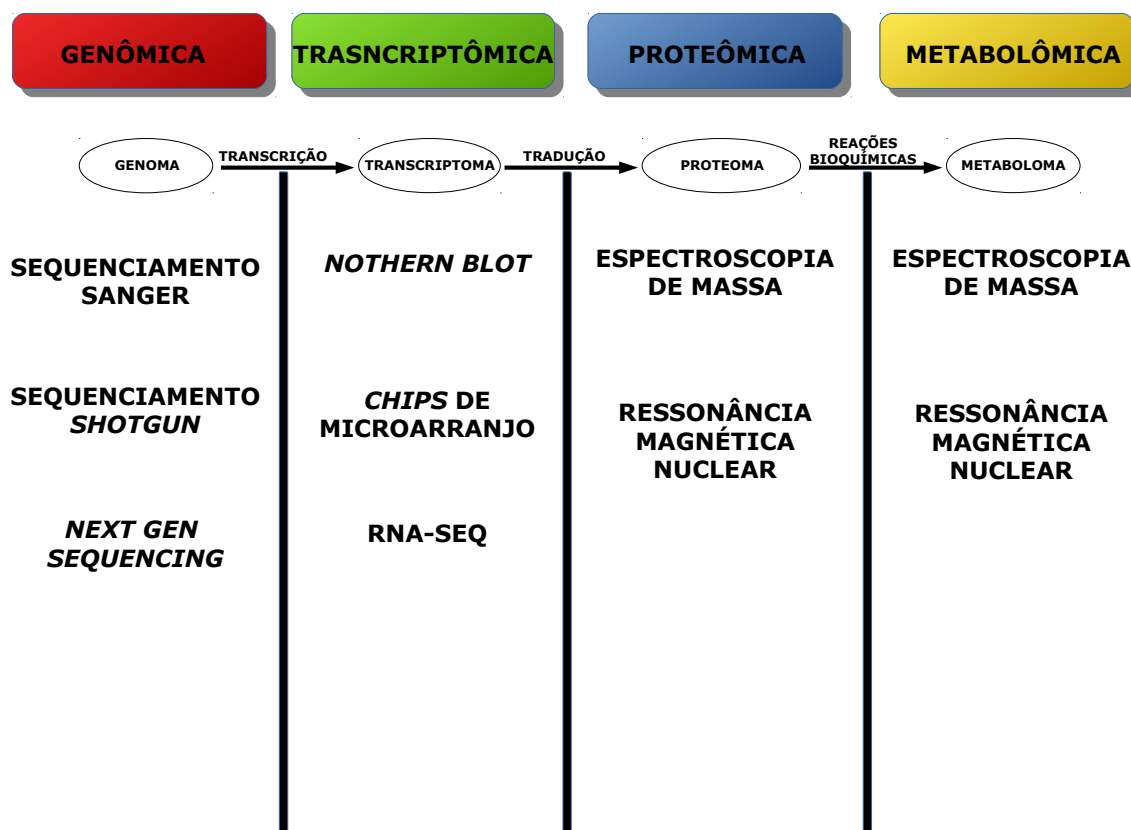


Figura 1.2: Na figura, está ilustrado o fluxo de informação biológica do genoma para o metaboloma, identificando o nome do processo responsável por catalisar a transição de um nível de informação para o outro. Esses níveis de informação, genoma, transcriptoma, proteoma e metaboloma estão separados por 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 si 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 auxílio de *software* e metodologias tradicionais. O *Big Data* é uma área que lida com 4 principais conceitos (os 4 **Vs**) (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 **Vs** 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).

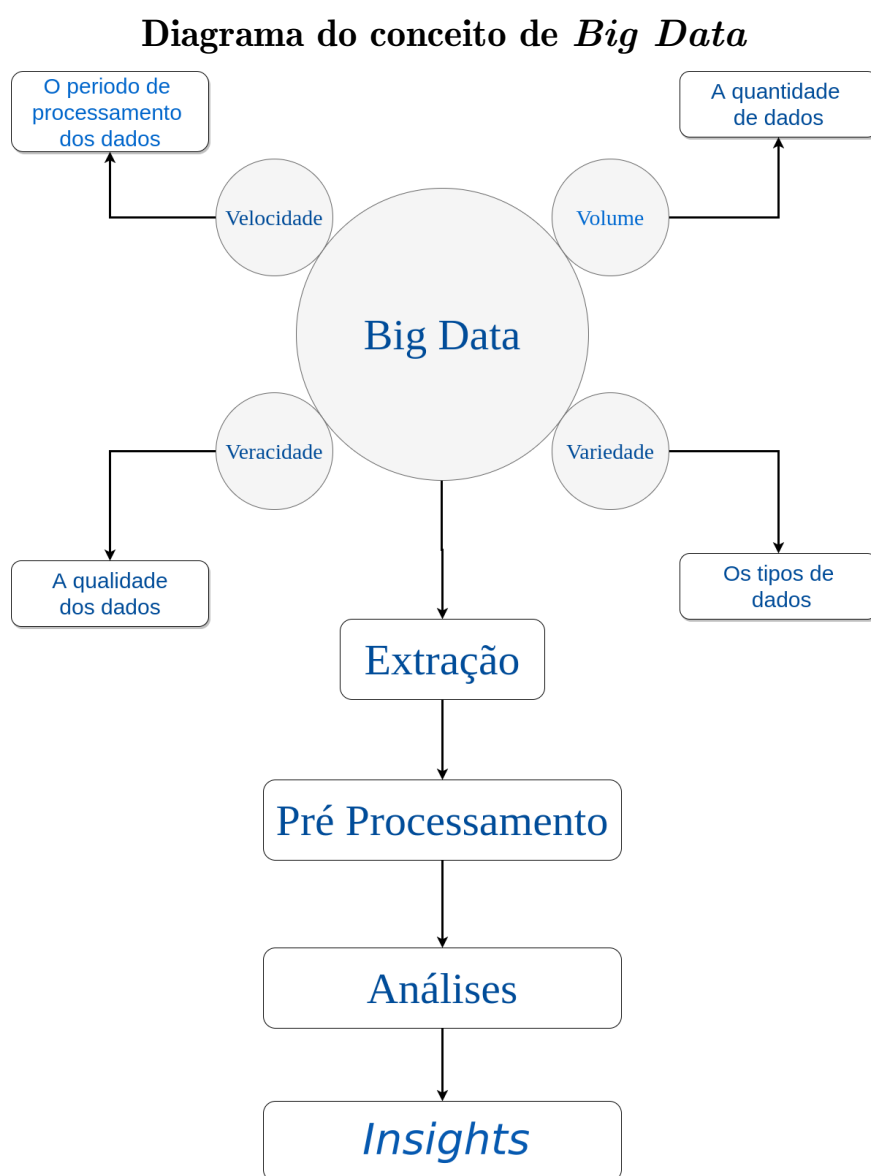


Figura 1.3: O diagrama ilustra o conceito de *Big Data* segundo suas 4 principais características (volume, variedade, velocidade e veracidade) e os procedimentos necessários para derivação de *insights*, a principal motivação desta área.

Os desafios associados aos 4 Vs, 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 si 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 si 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 na 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 manejo clínico de pacientes (prognóstico, predição de sobrevivência, recorrência e susceptibilidade) e estudo de mecanismos moleculares da doença como é apresentado por Hanahan e colaboradores (HANAHAAN; 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 clusterização¹ 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; GOLETIS; 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

¹O 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.

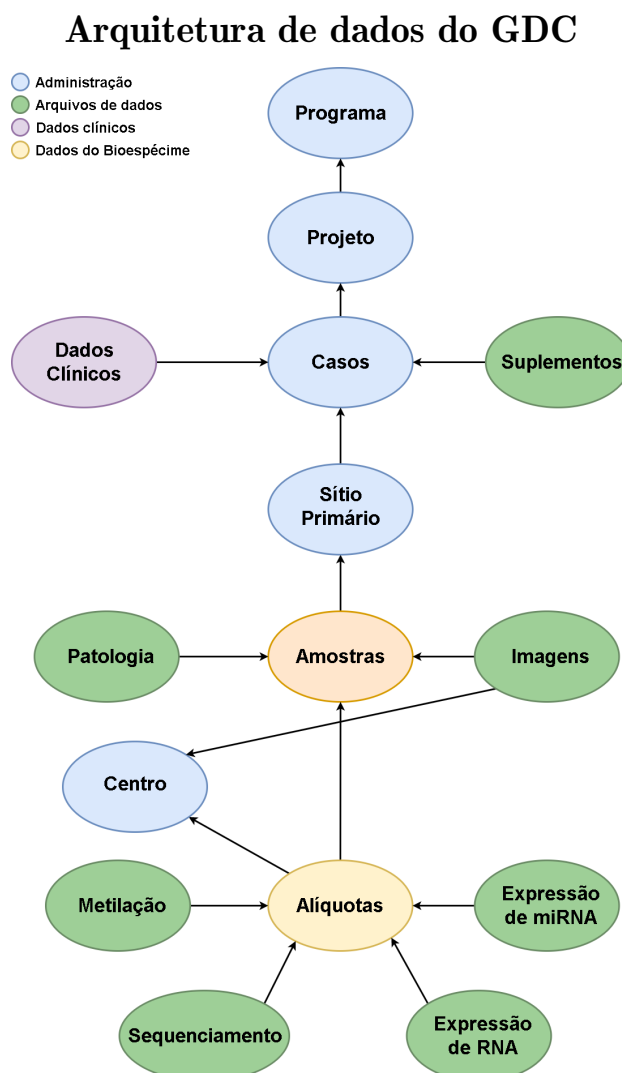


Figura 1.4: A figura ilustra a arquitetura de dados utilizada pelo GDC (*Genomic Data Commons*). As setas indicam a direção de um elemento de menor hierarquia da estrutura para um de maior hierarquia. Note que as cores tem significados: os elementos azuis são os de maior hierarquia da estrutura e representam 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 si, 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 si, 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:

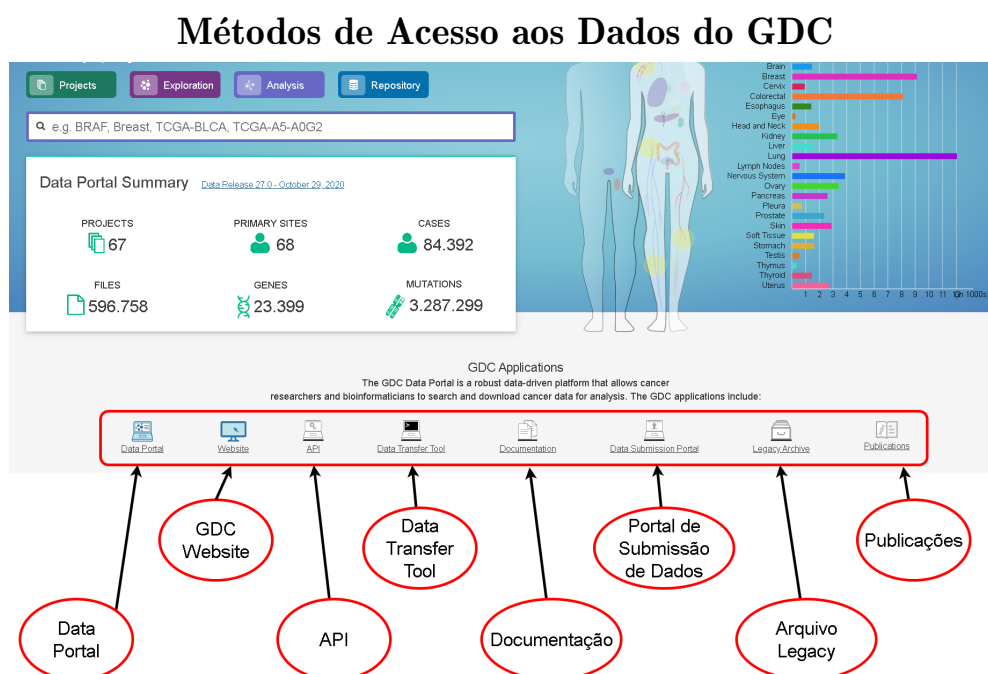


Figura 1.5: Na figura acima, temos a página inicial ao se acessar o endereço eletrônico <https://portal.gdc.cancer.gov/>, na qual pode se observar uma sumarização dos dados contidos no domínio. A parte inferior da página contém um menu com ícones, circulado em vermelho. Esses ícones representam: maneiras 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 (“Documentação”) 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

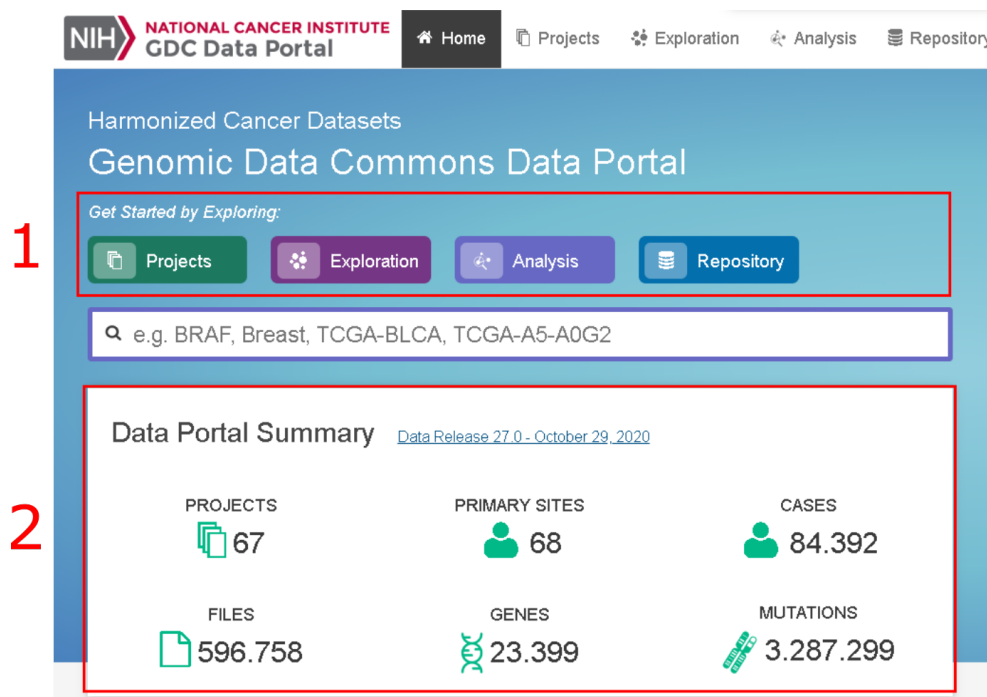


Figura 1.6: Em 1 pode-se observar as 4 principais formas de acesso aos dados armazenados no GDC. No caso, esse se trata do acesso via endereço eletrônico, mas existem outras maneiras de trabalhar com o GDC. Em *Projects*, os dados podem ser acessados como parte de um dos muitos projetos armazenados no 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 **C**ancer). 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
<i>The Cancer Genome Atlas</i>	TCGA	33
<i>Therapeutically Applicable Research to Generate Effective Treatments</i>	TARGET	9
<i>Genomics Evidence Neoplasia Information Exchange</i>	GENIE	8
<i>BETAML1.0</i>	BETAML	2
<i>Cancer Genome Characterization Initiatives</i>	CGCI	2
<i>Count Me In</i>	CMI	2
<i>Clinical Proteomic Tumor Analysis Consortium</i>	CPTAC	2
<i>Clinical Trial Sequencing Project</i>	CTSP	1
<i>FM-AD</i>	FM	1
<i>Human Cancer Models Initiative</i>	HCMi	1
<i>The Multiple Myeloma Research Foundation</i>	MMRF	1
<i>National Cancer Institute Center for Cancer Research</i>	NCICCR	1
<i>Philadelphia-Negative Neutrophilic Leukemias (CNL/aCML/MDS/MPNu)</i>	OHSU	1
<i>Pancreatic Cancer Organoid Profiling</i>	ORGANOID	1
<i>VA Research Precision Oncology Program</i>	VAREPOP	1
<i>Genomic Characterization of Metastatic Castration Resistant Prostate Cancer</i>	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

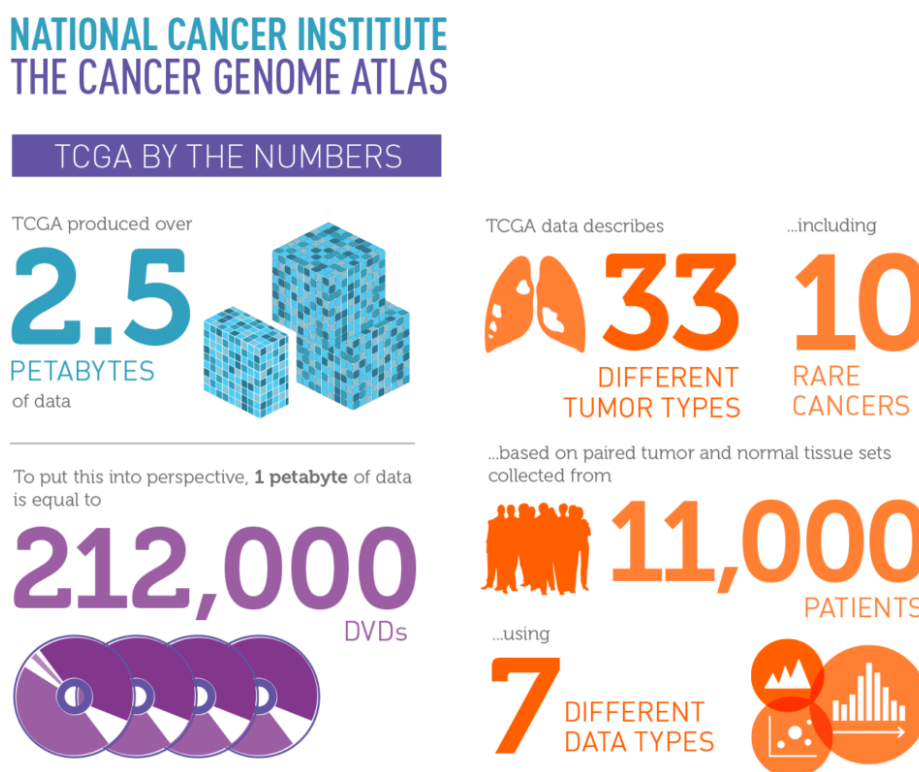


Figura 1.7: Dados armazenados no TCGA: A coluna esquerda, em azul e roxo, contém informações sobre a quantidade bruta de dados que o TCGA armazena no sítio <https://portal.gdc.cancer.gov/>. Na coluna direita, em laranja, pode se observar a classificação desses dados quanto aos tipos de tumor (33), experimentos (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	<i>Acute Myeloid Leukemia</i>
2	ACC	<i>Adrenocortical carcinoma</i>
3	BLCA	<i>Bladder Urothelial Carcinoma</i>
4	LGG	<i>Brain Lower Grade Glioma</i>
5	BRCA	<i>Breast invasive carcinoma</i>
6	CESC	<i>Cervical squamous cell carcinoma and endocervical adenocarcinoma</i>
7	CHOL	<i>Cholangiocarcinoma</i>
8	LCML	<i>Chronic Myelogenous Leukemia</i>
9	COAD	<i>Colon adenocarcinoma</i>
10	CNTL	<i>Controls</i>
11	ESCA	<i>Esophageal carcinoma</i>
12	FPPP	<i>FFPE Pilot Phase II</i>
13	GBM	<i>Glioblastoma multiforme</i>
14	HNSC	<i>Head and Neck squamous cell carcinoma</i>
15	KICH	<i>Kidney Chromophobe</i>
16	KIRC	<i>Kidney renal clear cell carcinoma</i>
17	KIRP	<i>Kidney renal papillary cell carcinoma</i>
18	LIHC	<i>Liver hepatocellular carcinoma</i>
19	LUAD	<i>Lung adenocarcinoma</i>
20	LUSC	<i>Lung squamous cell carcinoma</i>
21	DLBC	<i>Lymphoid Neoplasm Diffuse Large B-cell Lymphoma</i>
22	MESO	<i>Mesothelioma</i>
23	MISC	<i>Miscellaneous</i>
24	OV	<i>Ovarian serous cystadenocarcinoma</i>
25	PAAD	<i>Pancreatic adenocarcinoma</i>
26	PCPG	<i>Pheochromocytoma and Paraganglioma</i>
27	PRAD	<i>Prostate adenocarcinoma</i>
28	READ	<i>Rectum adenocarcinoma</i>
29	SARC	<i>Sarcoma</i>
30	SKCM	<i>Skin Cutaneous Melanoma</i>
31	STAD	<i>Stomach adenocarcinoma</i>
32	TGCT	<i>Testicular Germ Cell Tumors</i>
33	THYM	<i>Thymoma</i>
34	THCA	<i>Thyroid carcinoma</i>
35	UCS	<i>Uterine Carcinosarcoma</i>
36	UCEC	<i>Uterine Corpus Endometrial Carcinoma</i>
37	UVM	<i>Uveal Melanoma</i>

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.

Distribuição dos dados armazenados no TCGA

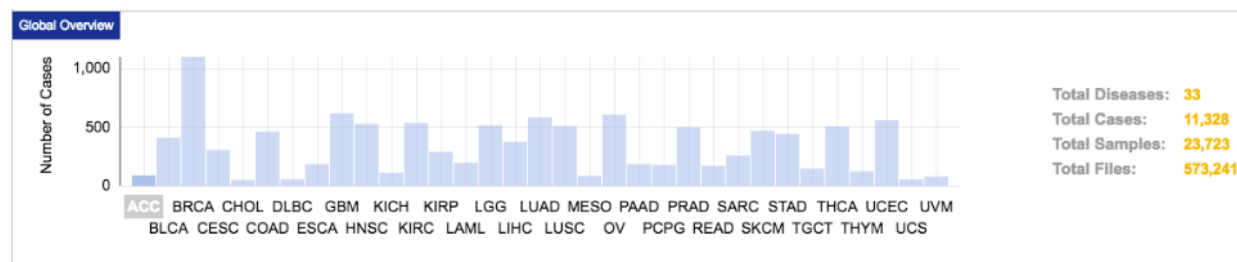


Figura 1.8: A distribuição de dados armazenados no sítio do TCGA. Pode se observar como existe uma grande discrepância na distribuição dos casos entre os tipos de tumores. O BRCA possui mais de 1000 casos disponíveis enquanto o CHOL possui menos de 100. Obtido de <https://www.sevenbridges.com/tcga/>

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)*

Quanto 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	<i>Expression-Genes</i>	<i>transcriptome</i>	Sim
2	CGCC	<i>Expression-Exon</i>	<i>exon</i>	Sim
3	CGCC	<i>Expression-miRNA</i>	<i>mirna</i>	Sim
4	CGCC	<i>CNV (CN Array)</i>	<i>cna</i>	Sim
5	CGCC	<i>DNA Methylation</i>	<i>methylation</i>	Sim
6	CGCC	<i>CNV (SNP Array)</i>	<i>snp</i>	Sim
7	CGCC	<i>SNP Frequencies</i>	<i>snp</i>	Não
8	CGCC	<i>SNP Copy Number Results</i>	<i>cna</i>	Não
9	GSC	<i>Trace-Sample Relationship</i>	<i>tracere1</i>	Sim
10	GSC	<i>Somatic Mutations</i>	<i>mutations</i>	Sim
11	BCR	<i>Complete Clinical Set</i>	<i>clin</i>	Sim
12	BCR	<i>Minimal Clinical Set</i>	<i>clin</i>	Sim
13	GSC	<i>Sequencing Trace</i>	<i>seq</i>	Não
14	BCR	<i>Tissue Slide Images</i>	<i>slide_images</i>	Sim
15	GSC	<i>Short-Read Relationship</i>	<i>sr</i>	Sim
16	GSC	<i>BAM-file Relationship</i>	<i>bam</i>	Sim
17	CGCC	<i>Quantification-miRNA</i>	<i>mirnaseq</i>	Sim
18	CGCC	<i>Quantification-miRNA Isoform</i>	<i>mirnaseq</i>	Sim
19	CGCC	<i>Annotations</i>	<i>annotations</i>	Sim
20	CGCC	<i>miRNA Sequence-Variants</i>	<i>mirnaseq</i>	Sim
21	CGCC	<i>Quantification-Junction</i>	<i>rnaseq</i>	Sim
22	CGCC	<i>Quantification-Gene</i>	<i>rnaseq</i>	Sim
23	CGCC	<i>Quantification-Exon</i>	<i>rnaseq</i>	Sim
24	BCR	<i>Reports</i>	<i>reports</i>	Sim
25	CGCC	<i>RNASeq</i>	<i>rnaseq</i>	Sim
26	CGCC	<i>miRNASeq</i>	<i>mirnaseq</i>	Sim
27	CGCC	<i>Expression-Protein</i>	<i>protein_exp</i>	Sim
28	CGCC	<i>Raw Sequence</i>	<i>raw_sequence</i>	Sim
29	CGCC	<i>Fragment Analysis Results</i>	<i>fragment_analysis</i>	Sim
30	CGCC	<i>GCC Reports</i>	<i>gcc_reports</i>	Sim
31	CGCC	<i>RNA Sequence-Variants</i>	<i>rnaseq</i>	Sim
32	CGCC	<i>RNA Structural-Variants</i>	<i>rnaseq</i>	Sim
33	CGCC	<i>RNASeqV2</i>	<i>rnaseqV2</i>	Sim
34	CGCC	<i>CNV (Low Pass DNaseq)</i>	<i>cna</i>	Sim
35	GSC	<i>Protected Mutations</i>	<i>mutations_protected</i>	Sim
36	CGCC	<i>Methylation - Bisulfite Sequencing</i>	<i>bisulfiteseq</i>	Sim
37	CGCC	<i>TotalRNASeqV2</i>	<i>totalrnaseqV2</i>	Sim
38	GDAC	<i>analyses</i>	<i>analyses</i>	Sim
39	GDAC	<i>stddata</i>	<i>stddata</i>	Sim
40	GDAC	<i>Firehose Reports</i>	<i>reports</i>	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).

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 análises mais específicas.

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-transcricional. 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 ppar γ 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 toxicogênico 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.; GOLETIS, 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.

HANAHAHAN, 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 micrnas 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- κ b 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.