UNIVERSIDADE DE SÃO PAULO
Instituto de Ciências Matemáticas e de Computação

Computational approaches for the discovery of significant genes in cancer

Jorge Francisco Cutigi

Tese de Doutorado do Programa de Pós-Graduação em Ciências de Computação e Matemática Computacional (PPG-CCMC)

ICMC USP
SÃO CARLOS

SERVIÇO DE PÓS-GRADUAÇÃO DO ICMC-USP

Data de Depósito:

Assinatura: ______________________

Jorge Francisco Cutigi

Computational approaches for the discovery of significant genes in cancer

Thesis submitted to the Instituto de Ciências Matemáticas e de Computação – ICMC-USP – in accordance with the requirements of the Computer and Mathematical Sciences Graduate Program, for the degree of Doctor in Science. *FINAL VERSION*

Concentration Area: Computer Science and Computational Mathematics

Advisor: Prof. Dr. Adenilso da Silva Simão
Co-advisor: Profa. Dra. Adriane Feijó Evangelista

USP – São Carlos
August 2021

Ficha catalográfica elaborada pela Biblioteca Prof. Achille Bassi
e Seção Técnica de Informática, ICMC/USP,
com os dados inseridos pelo(a) autor(a)

C988c

Cutigi, Jorge Francisco
Computational approaches for the discovery of significant genes in cancer / Jorge Francisco Cutigi; orientador Adenilso Simão; coorientadora Adriane Feijó Evangelista. -- São Carlos, 2021.
134 p.

Tese (Doutorado - Programa de Pós-Graduação em Ciências de Computação e Matemática Computacional) -- Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2021.

1. Cancer Bioinformatics. 2. Cancer Genomics. 3. Significant genes in cancer. 4. Significant mutations in cancer. 5. Computational approach. I. Simão, Adenilso, orient. II. Evangelista, Adriane Feijó , coorient. III. Título.

Bibliotecários responsáveis pela estrutura de catalogação da publicação de acordo com a AACR2:
Gláucia Maria Saia Cristianini - CRB - 8/4938
Juliana de Souza Moraes - CRB - 8/6176

Jorge Francisco Cutigi

Abordagens computacionais para a descoberta de genes significativos para o câncer

Tese apresentada ao Instituto de Ciências Matemáticas e de Computação – ICMC-USP, como parte dos requisitos para obtenção do título de Doutor em Ciências – Ciências de Computação e Matemática Computacional. *VERSÃO REVISADA*

Área de Concentração: Ciências de Computação e Matemática Computacional

Orientador: Prof. Dr. Adenilso da Silva Simão
Coorientadora: Profa. Dra. Adriane Feijó Evangelista

USP – São Carlos
Agosto de 2021

Aos meus pais, João e Fátima

To my parents, João and Fátima

ACKNOWLEDGEMENTS

A todas as pessoas que passaram pela minha vida e que tive a oportunidade de conhecer e conviver, mesmo que brevemente. Carrego um pedaço de cada um comigo. A Deus, por estar sempre à frente.

Aos meus pais, João Cutigi e Dona Fátima. Todo o esforço que eu tive em toda minha trajetória acadêmica e profissional é muito pequeno se comparado com esforço deles para que eu tivesse oportunidades. Aos meus queridos e inesquecíveis Vô Nicola, Vó Ana, Vô Frederico e Vó Bárbara que sempre estiveram presentes e foram muito importantes em minha vida. Aos meus tios, tias, primos e primas, em especial meus primos-irmãos Pati e Dan.

À minha esposa Helena, pelo carinho, companheirismo, paciência durante a condução deste trabalho e por todos os momentos compartilhados. À Rita, Sergio, Rafaela, Guilherme e ao meu sobrinho e afilhado Otto que logo chega.

Ao meu professor, orientador e amigo Adenilso por todos os ensinamentos (não apenas do mundo da computação), pelo impacto significativo que teve na minha vida e por todo o tempo em que trabalhamos juntos (muito tempo, aliás). À minha coorientadora Adriane pela inúmeras conversas sobre o mundo da genômica do câncer e pela sua paciência no meu aprendizado. A todos os professores de todos os níveis de ensino que fizeram parte da minha trajetória.

Aos amigos da minha querida cidade de Ibaté pelo companheirismo, amizade e pelos inesquecíveis encontros. Aos meus sempre presentes amigos da Info04. Ao pessoal do LabES, em especial à velha guarda, os quais mantemos contato até hoje. Aos meus colegas servidores do IFSP São Carlos, em especial aos professores da área de computação por sempre me incentivarem a cursar o doutorado. Aos meus alunos por todos os momentos que dividimos em sala de aula e nos corredores do IFSP São Carlos.

Ao pessoal do grupo de pesquisa ACDBio, por todas as reuniões, conversas, discussões, as quais foram muito importantes nesse trabalho. À Angela pelas criteriosas e inúmeras revisões de inglês neste trabalho e nos artigos submetidos.

À Universidade de São Paulo (USP) por me receber durante a maior parte da minha trajetória acadêmica. Aos funcionários do ICMC/USP, que tanto contribuem para o bom ambiente de trabalho. Ao Hospital de Câncer de Barretos por possibilitar a parceria tão importante para o desenvolvimento deste projeto. Ao CeMEAI e à FAPESP (processo 2013/07375-0) pelo apoio nas atividades de colaboração com o hospital e na disponibilização do cluster para experimentos.

Ao Instituto Federal de Educação e Tecnologia de São Paulo (IFSP) pela oportunidade e incentivo à qualificação dos seus servidores.

Um sincero obrigado a todos!!!

"Se eu vi mais longe, foi por estar sobre ombros de gigantes"
"If I have seen further it is by standing on the shoulders of giants"
(Isaac Newton, 1675)

RESUMO

CUTIGI, J. F. **Abordagens computacionais para a descoberta de genes significativos para o câncer**. 2021. 134 p. Tese (Doutorado em Ciências – Ciências de Computação e Matemática Computacional) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2021.

O câncer é uma doença complexa provocada por alterações genéticas que se acumulam por toda a vida do indivíduo. A essas alterações dá-se o nome de mutação genética, as quais podem ser divididas em dois grupos: 1) *Passenger mutations*: mutações que não alteram o comportamento da célula; 2) *Driver mutations*: mutações significativas para o câncer, ou seja, que provocam a carcinogênese na célula. Células de câncer possuem um elevado número de mutações, das quais a maioria delas são *passenger mutations* e um pequeno número delas são *driver mutations*. A identificação de genes significativamente mutados, isto é, genes com mutações significativas, é essencial para a compreensão dos mecanismos de iniciação e progressão do câncer. Essa tarefa é um desafio chave na genômica do câncer, uma vez que estudos mostram que genes significativos podem sofrer mutação em uma frequência muito baixa. Com o sequenciamento de nova geração, uma extensa quantidade de conjuntos de dados genômicos foram gerados, criando o desafio de analisar e interpretar esses dados. Para identificar genes relacionados ao câncer com taxa de mutação baixa, redes de interação gênica combinadas com dados de mutação têm sindo exploradas. Neste contexto, esta pesquisa apresenta abordagens computacionais para a descoberta de genes significativos para o câncer. O genes são priorizados por um método baseado em redes que combina frequência de mutação ponderada e influência de vizinhos na rede, e possíveis falsos positivos são detectados por método baseado em aprendizado de máquina, o qual utiliza-se de dados de mutação e redes de interação gênica para induzir modelos preditivos. Um estudo experimental conduzido com seis tipos de câncer revelou o potencial das abordagens na descoberta de genes já conhecidos e de possíveis novos genes significativos para o câncer.

Palavras-chave: Bioinformática do Câncer; Genômica do Câncer; Genes Significativos para o Câncer; Mutações Significativas para o Câncer; Abordagem Computacional; Redes de Interação Gênica; Dados de Mutação em Câncer. .

ABSTRACT

CUTIGI, J. F. **Computational approaches for the discovery of significant genes in cancer**. 2021. 134 p. Tese (Doutorado em Ciências – Ciências de Computação e Matemática Computacional) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2021.

Cancer is a complex disease caused by the accumulation of genetic alterations during the individual's life. These alterations are named genetic mutations, which may be divided into two groups: 1) Passenger mutations: mutations that do not change the behavior of the cell; 2) Driver mutations: significant mutations for cancer, that cause carcinogenesis. Cancer cells have a large number of mutations, in which the large majority of them are passenger, and few mutations are drivers. The identification of significant mutated genes, i.e., genes with driver mutations, is essential for the understanding of the mechanisms of cancer initiation and progression. Such a task is a key challenge in cancer genomics, since several studies have shown many significant genes are mutated at a very low frequency. With the next generation DNA sequencing, large and complex genomic datasets have been generated, creating the challenge of analyzing and interpreting this data. Towards uncovering infrequently mutated genes, gene interaction networks combined with mutation data have been explored. This research presents computational approaches for the discovery of reliable significant cancer genes. Such a genes are prioritized by a network-based method which combines weighted mutation frequency and network neighbors influence, and possible false-positives are detected by machine learning-based method which uses mutation data and gene interaction networks to induce predictive models. An experimental study conducted with six types of cancer revealed the potential of the approaches on the discovering of known and possible novel reliable significant cancer genes.

Keywords: Cancer Bioinformatics; Cancer Genomics; Significant Genes in Cancer; Significant Mutations in Cancer; Computational Approach; Gene interaction networks; Cancer Mutation Data. .

LIST OF FIGURES

LIST OF ALGORITHMS

LIST OF TABLES

LIST OF ABBREVIATIONS AND ACRONYMS

BRCA	Breast invasive carcinoma
CGC	Cancer Gene Census
CNAs	Copy Number Alterations
CNVs	Copy Number Variations
COADREAD	Colorectal adenocarcinoma
COSMIC	Catalogue Of Somatic Mutations In Cancer
DCG	Discounted cumulative gain
DNA	Deoxyribonucleic Acid
GBM	Glioblastoma multiforme
HINT	High-quality INTeractomes
HPRD	Human Protein Reference Database
HRN	Human Reference Network
ICGC	International Cancer Genome Consortium
ILP	Integer Linear Programming
InDels	Insert and Deletions
IntOGen	Integrative OncoGenomics
iRefIndex	interaction Reference Index
KEGG	Kyoto Encyclopedia of Genes and Genomes
LUAD	Lung adenocarcinoma
MCMC	Markov Chain Monte Carlo
mRNA	messenger RNA
NCG	Network of Cancer Genes
NGS	Next-Generation Sequencing
PPI	Protein-Protein Interaction
PRAD	Lung adenocarcinoma
Reactome	Reactome Function Interaction Knowledge base
RNA	Ribonucleic Acid
SNVs	Single Nucleotide Variants
STAD	Stomach adenocarcinoma
SVs	Structural Variant
TCGA	The Cancer Genome Atlas
UTR	Untranslated Region

CONTENTS

CHAPTER 1

INTRODUCTION

1.1 Contextualization

Cancer, one of the main causes of death worldwide and responsible for approximately 9.6 million deaths in 2018 (BRAY *et al.*, 2018), is a somatic evolutionary process that causes alterations in the normal behavior of cells. As a result, the cells start a fast and uncontrolled division process, leading to the formation of tumors in many parts of the human body (e.g., lung cancer, leukemia, breast cancer, and melanoma).

Cancer is known to be caused by the accumulation of genetic alterations during an individual's life, which range from small changes in nucleotides to more considerable variations in genetic material. Such changes, called genetic mutations, are caused by variable factors, which may be internal to the organism (e.g., cell division failure), or external (e.g., excessive exposure to the sun). They result in a disordered growth of cells, which invade tissues and organs, thus causing cancer (STRATTON, 2009).

Genetic mutations in cancer have been long studied through DNA/RNA sequencing, and a high number of recurrent mutations has been identified (VOGELSTEIN *et al.*, 2013). New genome-sequencing technologies, called Next-Generation Sequencing (NGS), enable fast and cost-effective genomic sequencing, as well as the generation of a large volume of biological data in short time. Such data help the study and analyses of genetic alterations in many diseases, including cancer, and the development of personalized medicine (SOON; HARIHARAN; SNYDER, 2013; MERIC-BERNSTAM; MILLS, 2012).

However, the abundance of genomic data hampers the processing of NGS data for useful clinical information. In this sense, clinical bioinformatics develops and uses computational methods and techniques for the interpretation of data, thus obtaining information and providing subsidies for health professionals and researchers.

One of the categories of computational methods includes those that aim to identify significant mutations (or driver mutations) and their associated genes (or driver genes) for cancer development. A cancer cell may have two types of mutations, namely 1) passenger mutations, which do not change the cell behavior, and 2) driver mutations, which cause harmful behavior and are responsible for the development of cancer, i.e., they provide cells with a selective advantage in comparison to the other cells, increasing their survival and reproduction. The identification of driver mutations and their associated genes is one of the most significant challenges in the area of Cancer Genomics (HOU; MA, 2013; RAPHAEL *et al.*, 2014). In the present research, the following terms are considered as synonymous: 1): “significant mutations” and “driver mutations”; and 2)“significantly mutated genes” and “driver genes”.

Several studies of computational methods and their algorithms for the identification of significantly mutated genes in cancer have been conducted in recent years (HOU; MA, 2013; RAPHAEL *et al.*, 2014; CHENG; ZHAO; ZHAO, 2015; DIMITRAKOPOULOS; BEERENWINKEL, 2017; CUTIGI; EVANGELISTA; SIMAO, 2020a), and various methods have been proposed (MILLER *et al.*, 2011; VANDIN; UPFAL; RAPHAEL, 2011; CIRIELLO *et al.*, 2012; VANDIN; UPFAL; RAPHAEL, 2012; DEES *et al.*, 2012; BASHASHATI *et al.*, 2012; HODIS *et al.*, 2012; LAWRENCE *et al.*, 2013; LEISERSON *et al.*, 2013; HOU; MA, 2014; LEISERSON *et al.*, 2015; LEISERSON *et al.*, 2015; KIM *et al.*, 2015; LEISERSON; REYNA; RAPHAEL, 2016; CHO *et al.*, 2016; HOU *et al.*, 2016; HRISTOV; SINGH, 2017; HORN *et al.*, 2018; REYNA; LEISERSON; RAPHAEL, 2018; WU *et al.*, 2019; ZHU *et al.*, 2019; CUTIGI; EVANGELISTA; SIMAO, 2020b; YANG *et al.*, 2021). Each method displays different characteristics, from computational and biological perspectives. New associated cancer genes, since a single gene to a group of related genes, have been discovered through analyses of NGS data and application of specific algorithms for finding relevant information.

Computational methods have adopted various strategies to uncover significantly mutated genes in cancer, e.g., gene interaction networks, used for studies of mutated genes’ interactions and their influence on networks. Network analysis is essential, since genes affected by driver mutations tend to participate in common biological activities (OZTURK *et al.*, 2018). Furthermore, somatic mutations in cancer can alter the mutant gene and the entire pathways where such a gene is (VOGELSTEIN *et al.*, 2013).

1.2 Motivation

The knowledge about genes that cause cancer initiation and progression is a critical issue for Cancer Genomics and can significantly impact Cancer Medicine. Both cancer diagnosis and treatment could be substantially improved if doctors knew the mutated genes responsible for cells’ carcinogenic behavior, towards personalizing the treatment of a given patient (MERIC-BERNSTAM; MILLS, 2012), which characterizes a personalized cancer medicine

(CHIN; ANDERSEN; FUTREAL, 2011). In this sense, tumors would be identified and clearly characterized, thus enabling the most appropriate treatment (VINCENT, 2017). Furthermore, the discovery of novel cancer genes can lead new biomedical directions on the cancer treatment.

The identification of significant genes for cancer can be supported by computational methods based on several types of data currently available. Among such methods, mutation data analysis has taken a prominent position after the advent of next-sequencing generation technologies (NGS) and due to projects such as TCGA (The Cancer Genome Atlas) (WEINSTEIN *et al.*, 2013), which collects plenty of mutation data available. Gene interaction information is also explored and plays an important role in several computational methods (OZTURK *et al.*, 2018), providing essential information about complex interactions among genes and their related proteins. Furthermore, data modeled in networks represent prior knowledge based on decades of research (CREIXELL *et al.*, 2015; DENG *et al.*, 2017), thus being a reliable source of data for work on biological problems. In comparison to a sequencing-only approach, the integration of available data, such as sequencing mutation data and gene interaction networks, can enable the finding of novel candidate cancer (NUSSINOV *et al.*, 2019).

Although computational methods have been extensively used for the identification of significant genes for cancer, they can misclassify some genes as significant, thus requiring expert curation to filter their findings (BAILEY *et al.*, 2018). Such a misclassification is due to some genes (referred to as false-positive-drivers, or false-drivers) exhibiting characteristics of being significant for cancer, despite not being actually involved in its initiation and progression. The avoidance of the misclassification of false drivers as drivers is still a challenge, and the development of tools for a further screening of the findings and detection of possible misclassified genes is, therefore, crucial.

1.3 Problem, objective, hypothesis and research questions

The discovery of significant genes is a challenge in Cancer Genomics. Although several computational methods have been developed towards addressing it, they have failed in predicting all clinically diagnosed cancer genes mainly because the complexity of the conceptual biological cancer basis (NUSSINOV; TSAI; JANG, 2019) and the misclassification of some genes. In this context, the general objective of this thesis is to discover reliable significant cancer genes with the use of two computational approaches. The objective is based on the hypotheses that significantly mutated genes in cancer can be discovered through the combination of weighted mutation frequency and network neighbors influence, and possible false-positives can be detected by mutation data and gene interaction networks. Weighted mutation, extracted from mutation data, is based on the functional impact of each different type mutation on a cell's behavior. The influence of neighbors, extracted from gene interaction networks, is obtained from asymmetric spreading strength measurements between all node pairs, which take into account direct and indirect neighbors on the network. Such a proposal is based on a known and classic local hypothesis (BARABÁSI; GULBAHCE; LOSCALZO, 2011) that genes involved in cancer tend to interact with each other.

The following research questions were investigated:

RQ1: *Can significant genes for cancer be discovered through the combination of weighted mutation frequency and network neighbors influence?*

RQ2: *Can false-positive cancer genes be detected in a set of significant candidates for cancer with the use of mutation and gene network data?*

Towards answering to such questions, the following specific objectives were defined:

1. Select, study and analyze computational and biological perspectives from a set of computational methods that identify significant genes for cancer.
2. Identify specific biological and computational issues towards the proposal of new computational approaches.
3. Implement new computational approaches for both discovery of significant genes for cancer and avoidance of possible false-positives.
4. Define *in-silico* evaluation pipeline to assess the results of the proposed computational approaches.
5. Select data to be applied in the proposed computational methods.
6. Conduct experimental evaluation of the proposed approaches by defined evaluation pipelines and selected data.

The present project promoted the union of the Computer Science expertise from the University of Sao Paulo and Cancer Genomics expertise from Barretos Cancer Hospital towards the proposal of a new perspective and way of thinking and dealing with a relevant ongoing and challenging problem in Cancer Bioinformatics and Genomics.

1.4 Contributions

The central contribution of this thesis is the discovery of significant genes for cancer. Such genes are identified and a detection of possible false positives can be performed, in order to obtain more reliable results. The contribution was achieved through the proposal of two computational approaches. First, a computational method, called DiSCaGe (**Di**scovering **S**ignificant **Ca**ncer

Genes), prioritizes significant genes for cancer directly related to the impact of different mutation types and gene interactions on networks. It was applied and evaluated in six types of cancer with the use of their mutation data sets and two gene interaction networks. Cancer mutation data were subjected to a preprocessing routine, and networks underwent a link prediction process. Lists of prioritized genes were evaluated through precision and discounted cumulative gain by six recent cancer driver genes benchmarks, and an automated literature review of discovered genes. The results showed DiSCaGe's potential for discovering known and possible novel cancer genes, including very low frequency mutated genes. The other computational approach, called DFDriver (**D**etecting **F**alse **Driver**), detects possible false positives in a set of significant genes candidates. The classification is performed by using a supervised machine learning approach that employs the combination of mutation data of 33 cancer data sets and two gene interaction networks to induce models to efficiently classify cancer genes. The evaluation was performed using classical machine learning assessment, which showed the predictive potential of the models and the benefits on the combination of mutation and gene network data.

1.5 Thesis Organization

This thesis is organized as follows. In Chapter 2 the main concepts related to Cancer Genomics is presented. The objective of this chapter is to provide the biological cancer background to understand the next chapters and the development of the project. Next, in Chapter 3 the characteristics of driver mutations and the related challenges in cancer research are described. It is also presented existing computational methods for identifying driver mutations, showing their biological backgrounds and computational approaches. A summary of the methods is illustrated, showing the relations among the methods. Chapter 4 presents DiSCaGe, a computational method for the discovery significant genes for cancer. An cancer mutation data preprocessing, exploratory analysis and a network enrichment and characterization study are also performed. An extensive experimental study is presented, showing the potential of the proposed computational approach. Chapter 5 describes DFDriver, a machine learning-based approach to detect possible false-positive drivers, from data collection to the induction of predictive models. It presents a thorough evaluation of the models using classification metrics. Finally, Chapter 6 presents the conclusion, contributions, limitations and future work extracted from this research.

CHAPTER

2

CANCER GENOMICS

2.1 Initial Considerations

Cancer Genomics is the study of cancer cells genome. Basically, it focuses on how recent technological advances in Genomics have deepened the understanding of the genetic basis of appearance and evolution of cancer from a genomic perspective (DELLAIRE; BERMAN; ARCECI, 2014; NCI, 2018). The comprehension of the molecular biology of cancer is fundamental for its diagnosis and treatment, thus contributing to advances in precision oncology.

The chapter discusses the main topics related to Cancer Genomics, providing a background for the understanding and development of this thesis. Some parts of this chapter (e.g., mutation and heterogeneity concepts) were based on a paper published at the *Journal of Bioinformatics and Computational Biology (JBCB)* (CUTIGI; EVANGELISTA; SIMAO, 2020a) as follows:

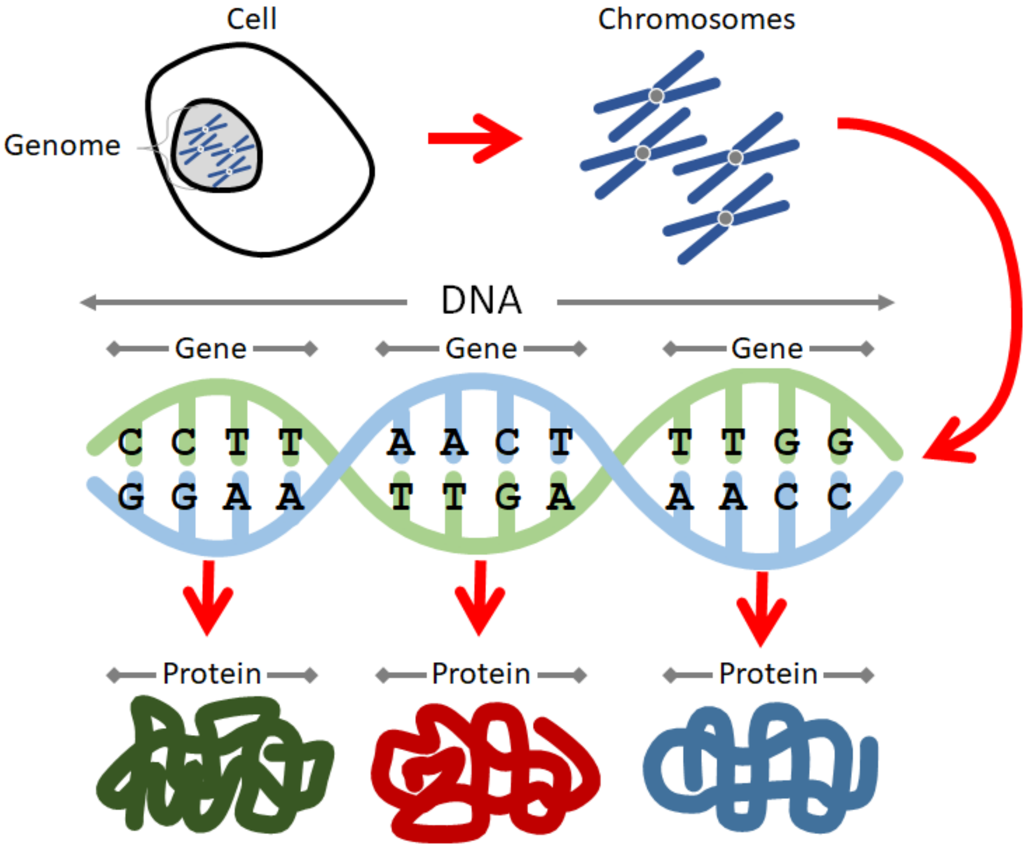
- CUTIGI, J. F.; EVANGELISTA, A. F.; SIMAO, A. Approaches for the identification of driver mutations in cancer: A tutorial from a computational perspective. **Journal of Bioinformatics and Computational Biology**, v. 18, n. 03, p. 2050016, 2020. PMID: 32698724. Available: <https://doi.org/10.1142/S021972002050016X>.

The chapter is organized as follows: Section 2.2 provides an overview on some basic concepts of molecular biology related to this research, and Section 2.3 addresses some biologically important characteristics and concepts of cancer frequently used in Cancer Bioinformatics and employed in the computational aspects presented here.

2.2 An Overview on Molecular Biology

A cell, the smallest unit of life, is composed of genetic material and many parts, called organelles, wrapped by a membrane. It contains the genetic material, comprised of chromosomes with a DNA molecule, which expresses the characteristics of each individual. Genome is the name of the full information from DNA for all proteins to be produced by the individual over his lifetime (ALBERTS *et al.*, 2008). Genes are portions of the DNA with biological information that code for proteins made of a large number of amino acids. Figure 1 illustrates the organization levels of a cell.

Figure 1 – Genome Schema: Genome is composed of chromosomes formed by DNA molecules. A specific portion of the DNA is a gene, which produces a protein.



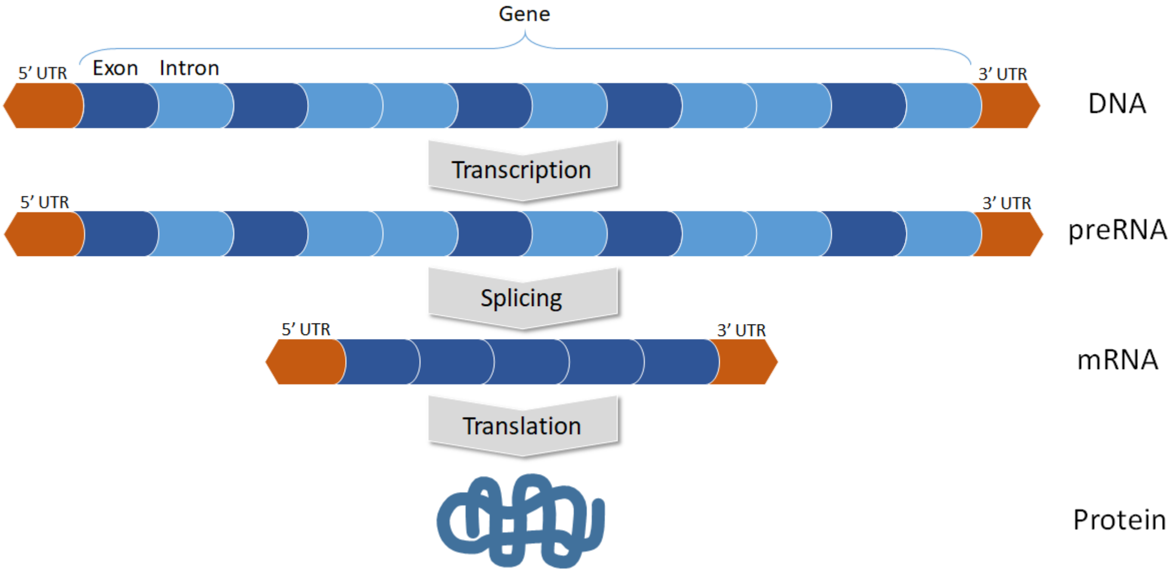
Source: Adapted from ThinnerGene (2018).

DNA is a molecule with two strands in a double-helix form (ALBERTS *et al.*, 2008). Each strand stores a code formed by four basic types of nucleotides, namely adenine (A), thymine (T), cytosine (C) and guanine (G). A single sequence of a strand of nucleotides usually represents a DNA, because the second strand can be derived from the first, where each nucleotide pairs up with each other, following a simple rule: A pairs up with T and C pairs up with G. The pair of nucleotides is called base pair.

A continuous sequence of nucleotides in the DNA can represent a gene. A gene is a portion of DNA that works as instructions for the production of amino acids of proteins responsible for the morphological and physiological characteristics of the individual. It is estimated that each human cell has 20,000 to 25,000 genes (NHGRI, 2021). However, not all parts a gene have instructions (or a code) to amino acids. A gene is composed of coding and non-coding regions, called exons and introns, respectively.

Genes do not produce proteins directly from DNA, but through mRNA (messenger RNA), which is an intermediary molecule in the process (ALBERTS *et al.*, 2008). When a protein is to be produced, a DNA region from a specific gene is copied for an RNA molecule, starting in a promoter region. This phase, called *transcription*, produces the RNA molecule, called preRNA, which has introns, exons, and untranslated regions (UTRs). Subsequently, in a phase called *splicing*, introns are removed from preRNA, thus resulting in mRNA, which produces the protein through molecules called ribosomes, in the *translation* phase. In this same phase, a sequence of three nucleotides (called codon or trinucleotide) is read in the mRNA, which generates a single amino acid. The set of sequenced amino acids generates a protein. All this process is known as Central Dogma, which is a paradigm of molecular biology (LEWIN, 2007). Figure 2 shows an overview of the Central Dogma processes.

Figure 2 – Central dogma of molecular biology.



Source: Elaborated by the author.

2.3 Cancer

Cancer is a disease acquired through an evolutionary process in a population of cells (NOWELL, 1976; STRATTON, 2009). According to Stratton (2009), its development is based on two processes: 1) a continuous acquisition of heritable genetic variation in cells, due to mutations, and 2) a natural selection, which results in diversity in a cell population.

Term “cancer” describes a large number of complex diseases, and comprehends hundreds of types and subtypes (e.g., skin cancer, or melanoma, breast cancer, and lung cancer), which can develop in any organ in the body. Some organs are more susceptible to it and can be affected by different types of tumor, more or less aggressive (INCA, 2019; INCA, 2020).

The following sections describe some cancer characteristics and the main concepts related to the present research.

2.3.1 Stages of Cancer

The evolutionary process in a population of cells can confer them a selective advantage, so that they reproduce indefinitely, which may lead to the development of cancer. During the process, cancer usually continues its evolution, and some stages are established. Among the several classifications of such stages, we have chosen the one defined by ASCO (2018) and described as follows:

The evolutionary process in a population of cells can confer them a selective advantage, which can make these cells to reproduce indefinitely, thus leading to the development of cancer. During the process, cancer usually continues its evolution, and some stages are established. Among the several classifications of such stages, in this works we have chosen the one defined by ASCO (2018) and described as follows:

Stage 0: the cancer is usually small, and contained in the organ of origin. It has not spread to other tissues, and is highly curable.

Stage I: this stage is similar to Stage 0, however, the cancer is larger. It has not spread to other surrounding tissues and can be surgically removed. It is often called early-stage cancer.

Stages II and III: the cancer is usually larger, and has started to spread to the surrounding tissues and lymph nodes. Classification into Stages II or III depends on the type of cancer.

Stage IV: the cancer has spread from its organ of origin to other organs or parts of the body. It is an advanced stage, and cancer is called advanced or metastatic.

The determination of the cancer stage is crucial for the prescription of the best treatment, since the stage is related to severity and diagnosis time.

2.3.2 Hallmarks of Cancer

Research has evidenced cancer is a complex process with multiple phases. According to Hanahan and Weinberg (2000), six essential alterations in the cells lead to a cancer behavior. In 2011, the authors updated their studies, proposing four new items, or characteristics, called *Hallmarks of Cancer*, shared by most, or perhaps, all types of cancer (HANAHAN; WEINBERG, 2011). Below is their description:

1. **Self-sufficiency in growth signals:** cancer cells acquire an autonomy to divide uncontrollably, due to alterations in essential genes (HANAHAN; WEINBERG, 2000).

2. **Insensitivity to anti-growth signals:** cancer cells have deactivated tumor suppressor genes, which cause the cells to not control division and growth (HANAHAN; WEINBERG, 2000).

3. **Evasion of programmed cell death (apoptosis):** cancer cells deactivate genes and pathways that can cause their death. As a result, sick cells do not die, and continue to reproduce indefinitely (HANAHAN; WEINBERG, 2000).

4. **Limitless replicative potential:** cancer cells activate specific genes and pathways that make them immortal, even after having grown for many generations (HANAHAN; WEINBERG, 2000).

5. **Sustained angiogenesis:** cancer cells can start an angiogenesis process (new blood vessels formation) and, therefore, their source of blood and blood vessels are continuously supplied with oxygen and other nutrients (HANAHAN; WEINBERG, 2000).

6. **Tissue invasion and metastasis:** cancer cells can leave their original tissue and invade others, thus spreading to other organs in the body (HANAHAN; WEINBERG, 2000).

7. **Deregulating cellular energetics:** cancer cells require more energy than normal cells. Thereby, they can change themselves for obtaining more energy to survive and divide (HANAHAN; WEINBERG, 2011).

8. **Avoidance of immune destruction:** cancer cells can deceive cells of the immune system, which does not recognize or kill cancer cells (HANAHAN; WEINBERG, 2011).

9. **Genome instability and mutation:** cancer cells have combinations of damaged genome, which deregulate them (HANAHAN; WEINBERG, 2011).

10. **Tumor-promoting inflammation:** inflammations in the human body can drive the emergence of cancer cells (HANAHAN; WEINBERG, 2011).

In general, such hallmarks indicate cancer cells are imortal. They can grow indefinitely, and move to other tissues in the human body, due to their efficient energy-extraction mechanism. However, a cell rarely shows all hallmarks. The older the individual, the more exposure to external environments (e.g., the sun, smoking environments), and the more cell division, the higher the probability of the cells acquiring the hallmarks.

2.3.3 Oncogenes and Tumor Suppressor Genes

Cancer develops from failures in the cellular process that controls the division and reproduction of cells. Such failures can result in gene mutations and development (or not) of cancer. A mutated gene related to cancer can be classified into *proto-oncogene* and *tumor suppressor gene*.

Proto-oncogenes are a group of genes that, when mutated, can potentially incite cancer (WEINBERG, 2013), causing cells to reproduce uncontrollably. The mutated version of a proto-oncogene is called oncogene, which is an important molecular target for anti-cancer drugs design (CHIAL, 2008). On the other hand, tumor suppressor genes can control the cell division, i.e., they are antigrowth genes (WEINBERG, 2013), of which some can repair the DNA of a cell. In this case, if a mutation occurs during the cell division, proteins created by tumor suppressor genes can fix the DNA.

RAS gene, which produces proteins that control the transcription of genes related to the cell growth, is an example of proto-oncogene. When *RAS* is mutated, the protein produced is altered, and the cell can no longer interrupt the regulatory process that controls its growth (CANCERQUEST, 2020). *TP53* is a well-known tumor suppressor gene that controls the cell division and repairs problems in the DNA occurred during the division (CANCERQUEST, 2020).

Oncogenes usually have some few regions, called “hotspots”, which are more mutated than others (LEWIN, 2007). For example, region V600 in *BRAF* gene is a hotspot, i.e., it is altered in a large number of samples with mutations in *BRAF*. On the other hand, tumor suppressor genes usually do not have few hotspots. Information on hotspots is important for studies on the significance of a mutation.

In short, the difference between oncogenes and tumor suppressor genes is that the former can cause cancer as a result of the activation of proto-oncogenes, whereas tumor suppressor genes can not avoid cancer when they are inactivated in the cell (ACS, 2014). Therefore, the identification of proto-oncogenes, tumor suppressor genes, and their associated mutations is mandatory in Cancer Genomics.

2.3.4 Mutations

Genes contained in cells can undergo alterations, called mutations, in comparison with original cells. The genome sequence has revealed a huge number of gene mutations can occur across cancer cells of any type of cancer (STRATTON, 2011). Such mutations are classified according to their origin into *germline mutations*, which are passed from parents to their children, and *acquired mutations*, called somatic mutations, which occur throughout an individual’s life by several causes, such as excessive exposure to sunlight, use of cigarettes, failure of cell division, among other factors.

2.3.4.1 Somatic Mutations

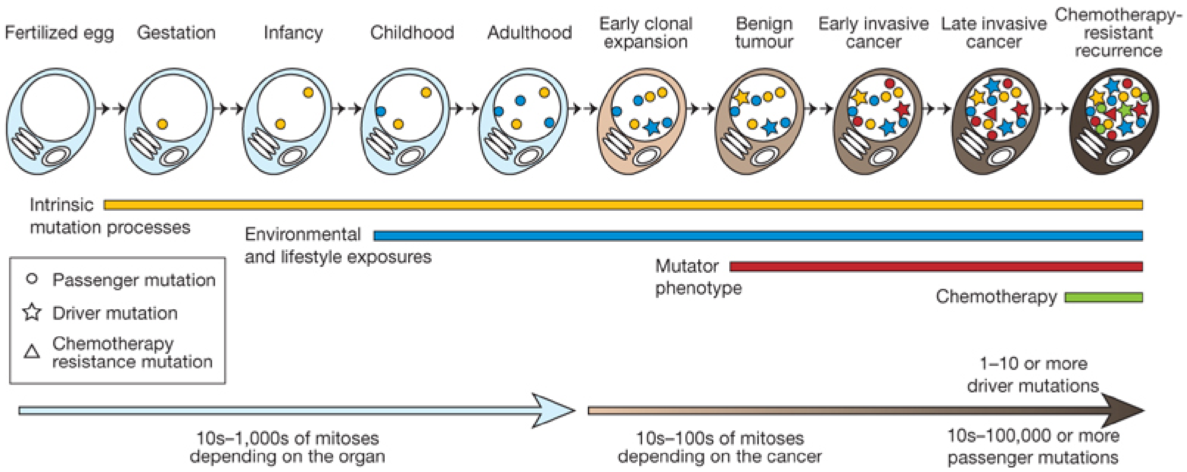
Cancer is caused by somatic mutations that occur all the time. Most mutations are known to be benign, i.e., they do not contribute to carcinogenesis (formation of a cancer) (VOGELSTEIN *et al.*, 2013; PLEASANCE *et al.*, 2009). However, others can cause cells to grow faster or evade to other healthy tissues, thus changing the gene expression. According to Stratton (2009), somatic mutations in a cancer cell genome can be classified into driver or passenger mutations, according to their consequences for cancer development, as follows:

Driver Mutations: they confer cells the advantage of growing uncontrollably, thus promoting the cancer development. They allow cancer cells to divide more than normal cells, and spread to other tissues (STRATTON, 2009; STRATTON, 2011). In general, they are responsible for the cancer initiation and progression, and, therefore, considered significant for cancer.

Passenger Mutations: they do not alter the behavior of cells, i.e., they do not confer a growth advantage to them (STRATTON, 2009; STRATTON, 2011). In general, their impact is believed to be neutral, therefore, they are not significant for cancer.

Figure 3 displays the process of cell divisions from a fertilized egg to a cancer cell (STRATTON, 2009). It shows the timing of the somatic mutations occurred in the cell and the processes that contribute to cancer development.

Figure 3 – Processes that lead a normal cell to cancer through somatic mutations.


Fertilized egg
Gestation
Infancy
Childhood
Adulthood
Early clonal expansion
Benign tumour
Early invasive cancer
Late invasive cancer
Chemotherapy-resistant recurrence
Intrinsic mutation processes
Environmental and lifestyle exposures
Mutator phenotype
Chemotherapy
○ Passenger mutation
☆ Driver mutation
△ Chemotherapy resistance mutation
1–10 or more driver mutations
10s–1,000s of mitoses depending on the organ
10s–100s of mitoses depending on the cancer
10s–100,000 or more passenger mutations

Source: Stratton (2009).

In summary, a single cancer cell usually undergoes a large number of mutations, which comprehend few drivers and many passenger mutations, of which the former provide cells with a selective advantage. An important goal of cancer genome analysis is the identification of driver mutations, therefore, a key challenge is to distinguish between driver and passenger mutations (STRATTON, 2009).

2.3.4.2 Types of Mutations

Mutations in cancer occur on different scales, i.e., from a simple variation of a single nucleotide to a huge alteration in a significant part of the chromosome, or even in the whole chromosome (aneuploidy). Some of the main types of mutations are described below.

Single nucleotide variant (SNV): the smallest unit of a mutation that occurs in a non-coding or coding region when a single nucleotide is substituted by another.

When a mutation occurs in a coding region, the SNV can be classified into synonymous or nonsynonymous. The former refers to an SNV that does not affect the amino acid matched and, consequently, does not change the protein produced. On the other hand, a nonsynonymous SNV alters both amino acid and protein produced.

A nonsynonymous mutation can be categorized as either missense, or nonsense. The former results in the substitution of an amino acid for another, whereas a nonsense mutation creates a signal for the cell to stop building the protein, thus resulting in a shortened protein.

Insert and Deletion (InDel): a single or small sequence of nucleotides can be inserted into or deleted from part of the DNA sequence. This type of mutation can be considered an SNV mutation if the alteration is a single base pair.

Copy Number Alteration (CNA): a middle mutation level, in which the DNA can either gain, or lose a large segment of the genome. In the first case, the mutation is called amplification; otherwise, i.e., when DNA loses a segment, the mutation is called deletion.

Structural variant (SV): the highest mutation level, which occurs when a significant part of the genome has been altered. For example, interchromosomal translocation, an SV type, occurs when a large segment of a chromosome leaves its chromosome of origin and goes to another one. Other types include intrachromosomal translocation and inversions.

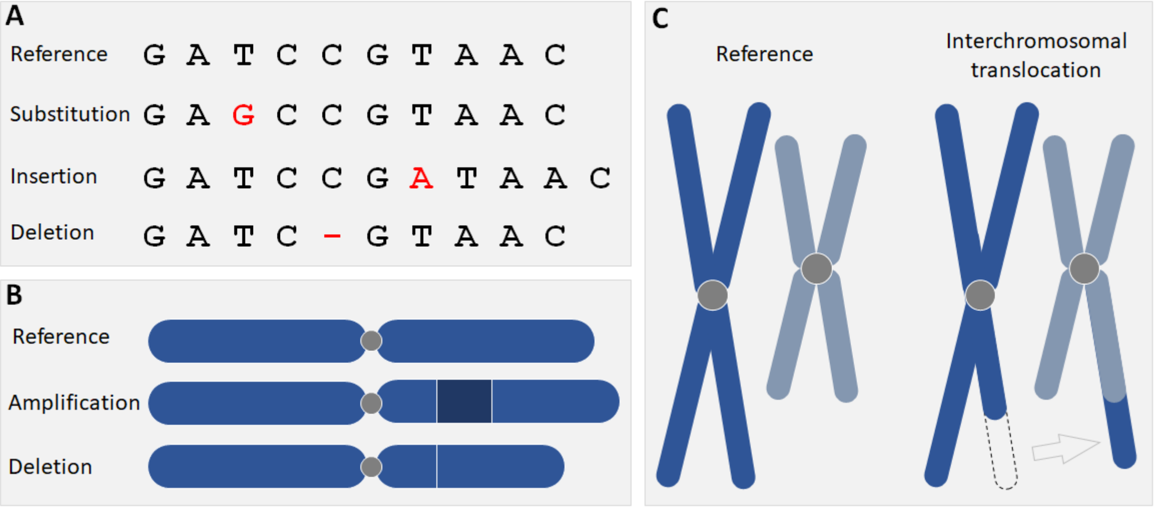
Most mutations can cause cells to produce proteins different than those expected. In such cases, the generated proteins can be nonfunctional or perform an abnormal function. Figure 4 shows some examples of types of mutations.

2.3.4.3 Mutation Matrix

Mutation Matrix is the usual way of representing genes (mutated or not) and patients. It can be defined as a matrix M with m rows denoting patients (or samples), and n columns representing genes. Each cell M_{ij} is given value 1 if gene g_j of patient m_i has been mutated, and 0, otherwise. For a set of m patients and n genes, a mutation matrix is defined as:

$$M_{ij} = \begin{cases} 1 & \text{if gene } j \text{ has been mutated in patient } i \\ 0 & \text{otherwise} \end{cases} \tag{2.1}$$

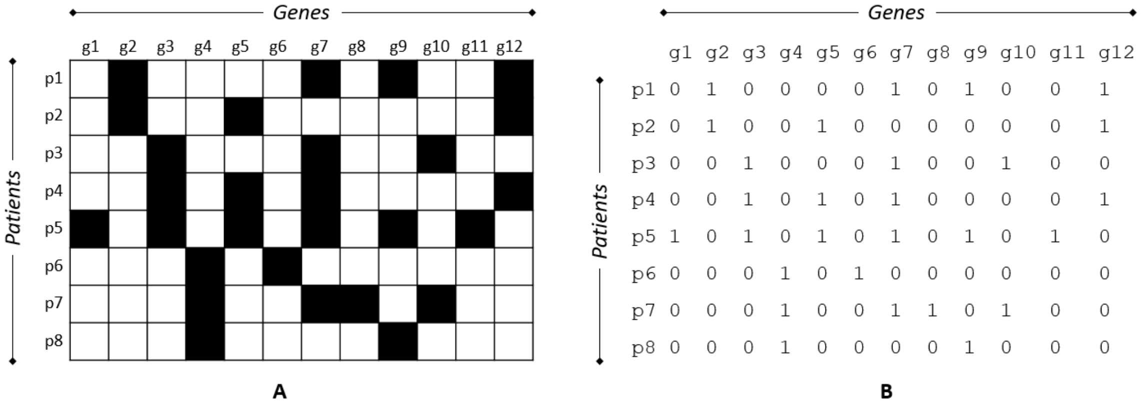
Figure 4 – Non-exhaustive view of types of mutations: (A) SNVs, exemplifying substitution, insertion, and deletion. (B) CNAs, exemplifying amplification and deletion. (C) SVs, exemplifying a case of an interchromosomal translocation.


A
Reference G A T C C G T A A C
Substitution G A G C C G T A A C
Insertion G A T C C G A T A A C
Deletion G A T C – G T A A C
B
Reference
Amplification
Deletion
C
Reference
Interchromosomal translocation

Source: Elaborated by the author.

A mutation matrix is built from a list of somatic mutations from a cohort of patients. Figure 5 illustrates a typical mutation matrix in two perspectives, i.e., graphical and raw data representations.

Figure 5 – Same data in two different representations: (A) Mutation matrix in a graphical perspective, where black cells indicate the gene mutated in the patient. (B) Mutation matrix in a raw data perspective.



	g1	g2	g3	g4	g5	g6	g7	g8	g9	g10	g11	g12
p1	0	1	0	0	0	0	1	0	1	0	0	1
p2	0	1	0	0	1	0	0	0	0	0	0	1
p3	0	0	1	0	0	0	1	0	0	1	0	0
p4	0	0	1	0	1	0	1	0	0	0	0	1
p5	1	0	1	0	1	0	1	0	1	0	1	0
p6	0	0	0	1	0	1	0	0	0	0	0	0
p7	0	0	0	1	0	0	1	1	0	1	0	0
p8	0	0	0	1	0	0	0	0	1	0	0	0

Source: Elaborated by the author.

2.3.5 Next-Generation Sequencing and Databases

Term Next-Generation Sequencing (NGS) denotes a set of technologies for the reading and sequencing of the human genome. They enable the sequencing of an entire human genome in a single day. NGS can be used to sequence the whole genome or some specific area of interest, including all coding genes or a set of individual genes (BEHJATI; TARPEY, 2013). A set of platforms (e.g, Illumina[1], Ion Torrent[2] and PacBio[3]) implements NGS, and different sequence platform vendors employ different strategies to sequence the human genome (BUERMANS; DUNNEN, 2014).

The advent of NGS technologies has changed the studies of Cancer Genomics (STRATTON, 2013) and generated a massive volume of genomic data, including cancer genomic data. Public repositories of cancer mutation data have been created and continuously updated, thus providing the scientific community with fast and easy access to a large variety of cancer data. The primary goal of such repositories is to provide real data towards enabling the development of research on cancer and its characteristics and behaviors. The following repositories have excelled:

Catalogue Of Somatic Mutations In Cancer (COSMIC): it contains large amounts of cancer mutation data and supports the exploration of the impact of mutations on several types of cancer (COSMIC, 2021).

The Cancer Genome Atlas (TCGA): created from the efforts of research entities, this project supplies information on genetic mutations for several types of cancer. It has already generated and stored genomic mutations of 33 cancer types in 11,000 patients, holding approximately 2.5 petabytes of data (WEINSTEIN *et al.*, 2013; TCGA, 2021).

International Cancer Genome Consortium (ICGC): it contains a database with genomic data of more than 20,000 tumor genomes (ICGC, 2019).

Pan-Cancer Analysis of Whole Genomes (PCAWG): it has been created from an international collaboration for the study of patterns of mutation in more than 2,600 cancer whole genomes (PCAWG, 2021).

The creation of large repositories of genomic data and their constant updates provide the scientific community with opportunities for research, however, the huge volume of data hampers data interpretation. Therefore, the current challenge refers to a way of interpreting such large datasets.

2.3.6 Heterogeneity

Although the hallmarks of cancer (Section 2.3.2) are important and similar for most cancer types, they are related to the phenotype. From a genotype point of view, cancer cells in individuals rarely share the same set of mutations (PE'ER; HACOHEN, 2011).

[1] https://www.illumina.com/

[2] https://www.thermofisher.com/br/en/home/brands/ion-torrent.html

[3] https://www.pacb.com/

Cancer is known to be characterized by a high heterogeneity of genetic changes, including cancer of the same type. Although some genes are known to contribute to the development of cancer (oncogenes and mutated tumor suppressor genes), patients rarely undergo the same genetic changes. Moreover, carcinogenesis after a mutation depends on the other changes already existing in the cell (ASHWORTH; LORD; REIS-FILHO, 2011).

Heterogeneity can be classified into the following two levels (BURRELL *et al.*, 2013):

Inter Tumor Heterogeneity: different somatic mutations occur in a same type of tumor, in different patients, i.e., the cancer cells of two patients with the same type of cancer can undergo a different set of mutations.

Intra Tumor Heterogeneity: different somatic mutations occur in the cells of a tumor in a same patient, i.e., the tumor of a single patient can contain a set of both different mutated cells and normal cells.

Heterogeneity leads to low frequency somatic mutations in a cohort of patients. It leads an intrinsic difficulty to identify common mutations through their mutation frequency. Such a heterogeneity also hinders the understanding and treatment of cancer. For example, one out of two patients with the same type of cancer may respond positively to a specific drug, whereas the other may not. Genetic differences in cancer cells can lead to two different diseases, even in the same tissue and with similar characteristics.

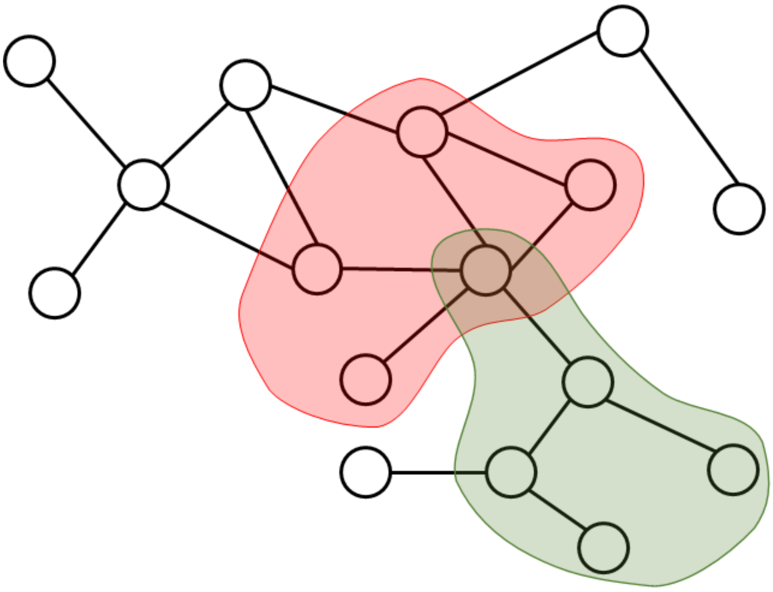
2.3.7 Networks and Pathways

A gene does not work alone, but establishes complex interactions with other genes and their produced proteins. Networks provide a natural representation of complex biological systems(KIM; CHO; PRZYTYCKA, 2016) for showing such interactions. Gene interaction networks are largely used in Cancer Genomics. In such networks, genes are nodes, and edges connect genes that are physically interacting or functionally related (KIM; CHO; PRZYTYCKA, 2016).

Likewise, pathways represent interactions around genes from a group. However, they are small networks of well-studied processes, in which interactions are usually related to some biological function. According to Creixell *et al.* (2015), pathways represent consensus systems, and are based on decades of research. They can be visualized in small diagrams, while networks comprise interactions around genes derived from large-scale screens or integrative analyses of multiple datasets. Figure 6 shows a general representation of network and pathways.

Many databases are sources of information about networks and pathways, e.g., Human Protein Reference Database (HPRD) (PERI *et al.*, 2003; PRASAD *et al.*, 2009), High-quality INTeractomes (HINT) (DAS; YU, 2012), HI-II-14 (ROLLAND *et al.*, 2014), Interaction Reference

Figure 6 – Simple representation of a network and pathways. All nodes (genes) and edges (gene interaction) represent networks. The red and green areas illustrate two different pathways in the network.



Source: Elaborated by the author.

Index (iRefIndex) (RAZICK; MAGKLARAS; DONALDSON, 2008), MutiNet (KHURANA *et al.*, 2013), Reactome Functional Interactions (ReactomeFI) (WU; HAW, 2017; FABREGAT *et al.*, 2018; JASSAL *et al.*, 2020), Kyoto Encyclopedia of Genes and Genomes (KEGG) (KANEHISA; GOTO, 2000; KANEHISA *et al.*, 2012), and Human Reference Interactome (HuRI) (LUCK *et al.*, 2020)

According to Ozturk *et al.* (2018), network and pathway analyses are essential, since genes affected by driver mutations tend to participate in common biological activities, described by network diagrams. Such diagrams can better identify patterns of driver events. Furthermore, somatic mutations in cancer can alter not only the mutant gene, but the entire pathway where such the gene is (VOGELSTEIN *et al.*, 2013). Creixell *et al.* (2015) stated that both pathway and network analyses can reduce the genomic data dimensionality. Information about gene interaction enables the selection of a smaller group of genes for the analysis and interpretation of a set of altered processes.

2.3.8 Mutual Exclusivity

Some patterns have been discovered in the massive volume of cancer genomic data. One of the most important is mutual exclusivity in cancer driver genes, widely observed in cancer genomes(DENG *et al.*, 2017). It shows the existence of relatively few drivers mutations in a cancer cell and such mutations contained in many pathways. Moreover, mutations in the same pathway are usually mutually exclusive (THOMAS *et al.*, 2007; YEANG; MCCORMICK; LEVINE, 2008).

The mutual exclusivity pattern is related to a group of two or more genes rarely mutated in a same patient, i.e., simultaneous mutations of genes in a same patient are less frequent than expected by chance (KIM; MADAN; PRZYTYCKA, 2017). On the other hand, such a group of genes can be mutated in different patients. According to Deng *et al.* (2017), more than one-quarter of known cancer genes are related to the mutual exclusivity pattern. The study conducted by Deng *et al.* (2017) refers to an example of *BRAF* and *NRAS* oncogenes (genes in *MAPK* pathway) showing genetic alterations in 40% and 25% of melanoma patients, respectively, while few patients undergo both genetic alterations (DAVIES *et al.*, 2002).

The following two hypotheses have been raised from the study of the mutual exclusivity pattern:

1. The mutual exclusivity pattern is usually found in a pathway (CISOWSKI; BERGO, 2017). Mutation in only one gene from a pathway is sufficient to perturb the pathway and its function, thus leading the cell to a cancer behavior. A cancer pathway is also expected to be mutated in a large number of patients (VANDIN; UPFAL; RAPHAEL, 2012).

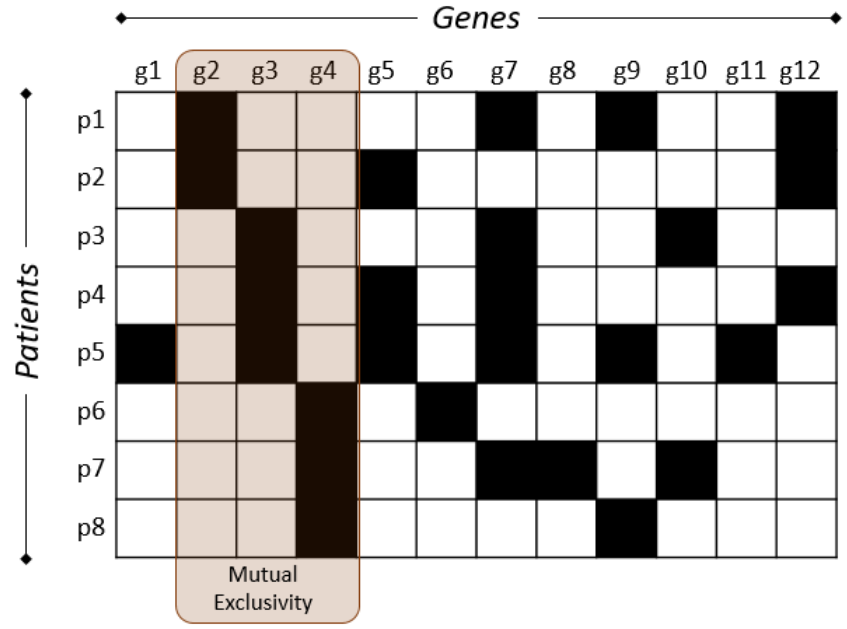
2. The co-occurrence of mutations in mutually exclusive genes is directly related to the cancer cell survival (DENG *et al.*, 2017). If two or more mutually exclusive genes are mutated, the damage in the cell is significant, and leads the cell to death. This situation does not characterize a cancer behavior, since the cell does not divide and proliferate.

Such hypotheses comprehend some important and specific cancer characteristics. For example, it is interesting to find a group of genes that shows a mutual exclusivity pattern. In this way, this group can be part of a known pathway or even a still unknown one. For example, genes *g2*, *g3* and *g4* of the mutation matrix of Figure 7 have mutual exclusivity characteristics, i.e., they are not mutated in more than one patient. When one of the genes is mutated more than once, the cell probably dies.

Many computational methods work with the mutual exclusivity pattern towards identifying driver mutations. According to Deng *et al.* (2017), such methods are divided into two groups: 1) *De Novo* methods: use only genomic data to identify genes with a mutual exclusivity pattern applying some strategies, such as pairwise test, combinatorial score and statistical score; and 2) Knowledge-based methods: use genomic data and prior knowledge information, such as known pathways, networks, and functional phenotype (e.g. gene expression data).

Both *de novo* and knowledge-based methods identify mutual exclusivity patterns in a group of genes. The main difference is knowledge-based is more useful for the study of the characteristics of known pathways or networks, or for the testing of the exclusivity of a group of genes. *De novo* methods can explore not well-known genes and find novel information.

Figure 7 – Mutual exclusivity among genes *g2*, *g3* and *g4*



Source: Elaborated by the author.

2.4 Final Considerations

This chapter has presented some biological concepts of cancer genomics towards the understanding the contribution of this thesis. Some biological concepts have been described and some computational approaches are shown, bringing a computational bias to the presented concept.

Although the full understanding of cancer aspects remains a challenge, significant advances have been made, of which some are related to the genomic area, where NGS technologies have generated a vast volume of data, and computational approaches have been developed for their analyses and interpretation. One of such approaches refers to methods that identify significant mutated genes (driver genes) in cancer. Since driver genes are the basis for this research, the next chapter addresses their main concepts and associated computational approaches.

CHAPTER 3

SIGNIFICANT MUTATIONS IN CANCER

3.1 Initial Considerations

A cancer cell has a large number of somatic mutations, but most of them are not significant for cancer, while a small fraction is responsible for the cancer initiation and progression. In this context, a key point in cancer genomics is to distinguish significant mutations from unimportant ones, and their associated genes, for cancer. This is a complex task due to the complexity of biological data and concepts. Computational approaches have been developed and applied to identify driver genes using NGS data.

The chapter discusses existing computational approaches to identify significant mutations in cancer, and it is based based on a paper published at the *Journal of Bioinformatics and Computational Biology (JBCB)* (CUTIGI; EVANGELISTA; SIMAO, 2020a) as follows:

- CUTIGI, J. F.; EVANGELISTA, A. F.; SIMAO, A. Approaches for the identification of driver mutations in cancer: A tutorial from a computational perspective. **Journal of Bioinformatics and Computational Biology**, v. 18, n. 03, p. 2050016, 2020. PMID: 32698724. Available: <https://doi.org/10.1142/S021972002050016X>.

In that paper, some classical computational methods for the identification of significant mutations in cancer are described from a computational perspective.

The chapter is organized as follows: Section 3.2 introduces characteristics about driver mutations and this significance for cancer. Section 3.3 presents an overview of some specific and related computational methods showing their main characteristics and relations among them.

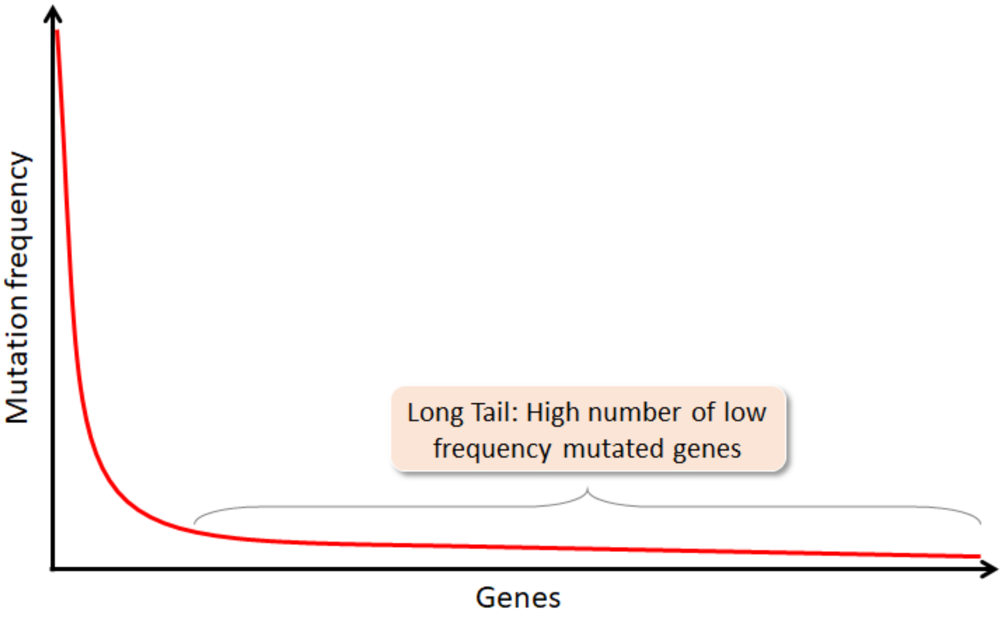
3.2 Identification of Significant Mutations in Cancer

The identification of mutations that cause cancer in the human body is a key challenge in the area of Cancer Genomics. It happens because the cell has mutations in its DNA that were acquired over the lifetime, which are called somatic mutations. These mutations are, for the most of times, random and do not contribute to the carcinogenic behavior of the cell, the so called passenger mutations. While most somatic mutations are passenger mutations, there is a smaller group of mutations that are significant for cancer, the so called driver mutations (STRATTON, 2009; STRATTON, 2011; HANAHAN; WEINBERG, 2011).

A key question in Cancer Genomics is to distinguish driver from passenger mutations (GREENMAN *et al.*, 2007; HABER; SETTLEMAN, 2007; TRAN *et al.*, 2012). In other words, it is a crucial point to determine the significance of the genetic alterations, identifying which set of alterations confer the selective advantage for the cancer cells. It is necessary for understanding the molecular mechanisms of carcinogenesis and improve treatments for patients (HOU; MA, 2013).

Studies on Cancer Genomics have shown that a small number of genes are mutated with high frequency in a given set of patients and a high number of genes are low-frequency mutated (GARRAWAY; LANDER, 2013; BAILEY *et al.*, 2018). This phenomenon is known as “long tail”, and it is illustrated in Figure 8. Some mutated genes in the tail (with a low frequency of mutation) can be genes that are significant for cancer, which brings a statistical difficulty because it is not enough to mention that genes with the highest frequency of mutation are a driver mutation.

Figure 8 – Long tail phenomenon: Few genes are highly mutated; High number of genes are few mutated.



Source: Elaborated by the author.

One of the causes of "long tail" phenomena is the inter-tumor heterogeneity of cancer, which is the fact that two genomes of the same type of cancer do not necessarily have the same set of mutations. All this context shows that many significant genes have not yet been discovered since many of these genes appear at low-frequency (GARRAWAY; LANDER, 2013).

3.3 Related Computational Methods

For the identification of significant mutations in cancer (or driver mutations), it is necessary to understand the molecular mechanisms involved. The identification of these mutations is of extreme importance for the understanding of these mechanisms, besides generating subsidies for the personalized treatment for each type of cancer or individual. The creation of methods for identifying these mutations is a relatively new field of research and has had many recent contributions. Hou and Ma (2013) group the methods into four categories:

Pathway-Based Approach: these methods works with gene networks and pathways, focusing on the interaction among the genes. Thus, they use graph algorithms to identify driver mutations, based on the impact that the mutated gene has on the pathways or network.

Mutation Frequency-Based Approach: these methods use statistical analysis to compare the number of mutations in a cancer cell in contrast to the number in normal cells. Mutations observed more than expected by chance are identified as driver mutations.

Sequence-Based Approach: these methods evaluate the functional impact in the protein generated after a mutation in a gene.

Machine Learning-Based approach: these methods use machine learning in order to create models from existing knowledge in driver mutations.

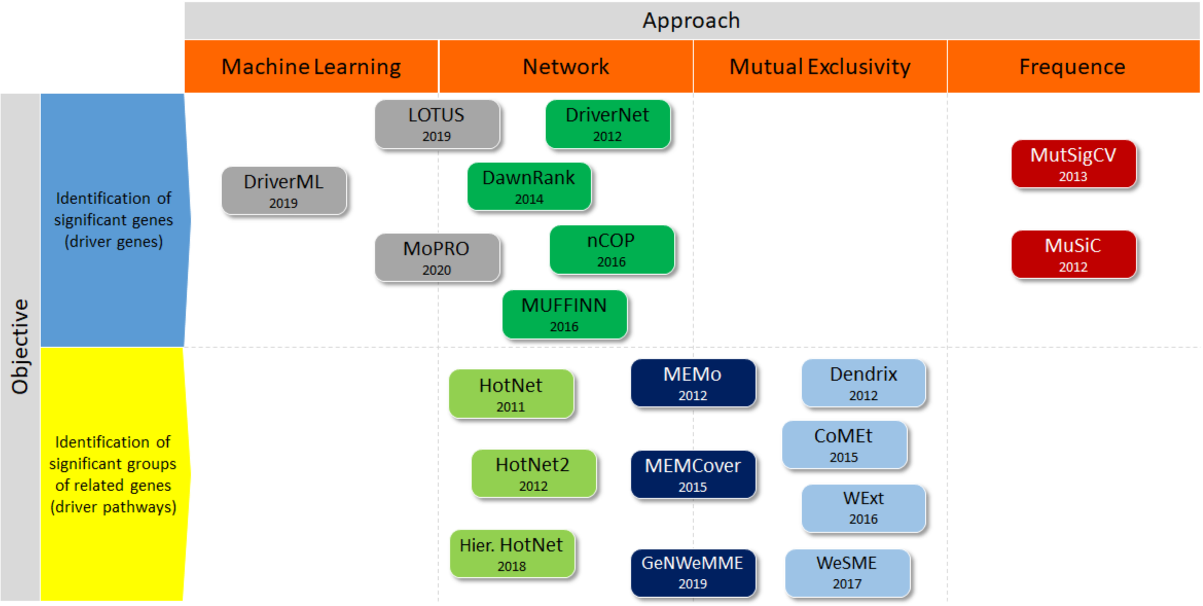
In this research we group the computational methods considering two perspectives:

From their goal: the computational methods have two basic goals: 1) Prioritizing driver genes, i.e., methods produces a ranking of genes in order of significance for cancer; and 2) Suggesting of driver pathways, i.e., methods produces a list of set of related genes significantly mutated, which can be called of driver pathways.

From their approaches: the computational methods have distinct approaches (form computational or biological point of views) to reach their goals. Among the existing several approaches, it can be cited: frequency, network, mutual exclusivity and machine learning.

Figure 9 presents a representation and classification of some computational methods in the perspectives described above. Next, we briefly describe the methods.

Figure 9 – An overview of related methods according to their main goal and based approach.



Source: Elaborated by the author.

MutSigCV (LAWRENCE *et al.*, 2013) is a classical method that incorporates the heterogeneity of mutations in the analysis and identification of driver mutations. The method determines whether the number of mutations observed in a gene is significantly higher than expected Background Mutation Rate (BMR), which is the probability of observing a passenger mutation by chance at a specific location in the genome (RAPHAEL *et al.*, 2014). MutSigCV considers the level of gene expression and the DNA replication time for taking account the heterogeneity of mutations. MuSiC (DEES *et al.*, 2012) seeks to distinguish passenger from driver mutation through a extensive pipeline, which uses several data and tools, e.g., BMR estimation, gene length information, and clinical correlation test.

MUFFINN (CHO *et al.*, 2016) takes into consideration the influence from neighbors to identify significant genes based on this influence, which according to the authors, if a gene has low mutation frequency, but its neighbors have a higher one, such a gene is a highly probable candidate to be a driver mutation. Another network-based method, nCOP (HRISTOV; SINGH, 2017) considers individual mutational profiles in a gene interaction network context towards identifying connected subnetworks that comprise pathways that are significant altered across many samples. Such a method employs an integer linear programming for solving the problem and a greedy heuristic algorithm, in order to get a better performance. DriverNet (BASHASHATI *et al.*, 2012) and DawnRank (HOU; MA, 2014) combines data about mutation, gene expression levels, and gene networks to rank the significance of the mutated genes.

“HotNet family” is composed of three similar approaches that use a diffusion process and scores of mutation to find significantly mutated subnetworks. HotNet (VANDIN; UPFAL; RAPHAEL, 2011) applies subnetwork discovery in an undirected network, while HotNet2

(LEISERSON *et al.*, 2015) uses a directed network. Hierarchical HotNet (REYNA; LEISERSON; RAPHAEL, 2018) groups the genes in hierarchical levels and finds hierarchical significance subnetworks. Recently, a new and improved approach, called NetMix (REYNA *et al.*, 2021), were proposed for the identification of altered subnetworks, which could be applied in HotNet family methods.

MEMo (CIRIELLO *et al.*, 2012) and MEMCover (KIM *et al.*, 2015) are classified as both network-based and mutual exclusivity pattern-based, since they find a mutual exclusivity pattern in a set of genes in an interaction network. The main difference is MEMCover can be applied to many types of cancer in a single analysis. GeNWeMME (CUTIGI; EVANGELISTA; SIMAO, 2020b) is a flexible method that uses an extensive biological base (mutations, type of mutations, gene interaction networks and mutual exclusivity pattern) for prioritizing groups of significant and related genes in cancer, which can be considered according to the objective of the analysis.

Dendrix (VANDIN; UPFAL; RAPHAEL, 2012) applies a weight function and an MCMC approach to find a single set of genes mutually exclusive of high-coverage. Its extension, called MultiDendrix (LEISERSON *et al.*, 2013) uses an ILP approach and the same weight function of Dendrix to find multiple sets of genes. CoMEt (LEISERSON *et al.*, 2015) proposes changing the weight function for a probabilistic score to identify sets of mutually exclusive genes and avoid the coverage bias of some highly mutated gene. WExT (LEISERSON; REYNA; RAPHAEL, 2016) implements an weighted test for mutual exclusivity for taking into account mutation frequency with the the probability in which a mutation can occur in each sample. WeSME (KIM; MADAN; PRZYTYCKA, 2017) employs a fast heuristic for estimating statistical significance of mutual exclusivity sets of genes, computing the significance for a subset of genes, thus not requiring whole genome permutations. Such method also is capable to estimate the significance of co-occurrence of mutated genes.

Regarding machine learning-based approaches have been taking advantage of massive volume of digital biological data and previous knowledge towards training models to find novel biological insights. LOTUS (COLLIER; STOVEN; VERT, 2019) uses one-class support vector machine (OC-SVM) to define two score functions to classify driver genes into oncogenes or tumor-suppressor genes. The model is trained with data from mutation (mutation frequency and functional impact) and protein-protein interactions, using a multitask learning strategy to share information across cancer types. DriverML (HAN *et al.*, 2019) identifies cancer driver genes by combining a weighted score test and machine learning approach. The score test takes account mutation data and somatic variant types, aiming at quantifying the mutation functional impact. Known driver genes are used to define weights of mutations, based on a machine learning approach. Score value of each gene was obtained using the weighted score statistic with the learned weight parameters. MoProEmbeddings (GUMPINGER *et al.*, 2020) employs four supervised machine learning algorithms for the classification of drivers genes. Models are trained under the data set that combines gene mutation score distribution and interaction network, in a node's local neighborhood with network propagation. The learning process takes account the data distributions of known cancer genes to improve its prediction task.

Table 1 displays information about the methods described here. For every method it is presented: 1) Name; 2) Main objective, related to the identification of significant mutations in cancer; 3) Main computational approach used; and 4) Biological knowledge or data used by the method.

Table 1 – Summary of Methods

Method	**Main objective**	**Main computational approach**	**Biological knowledge-based**
MutSigCV	Identification of significant genes	Statistical test	Gene-specific BMR estimation
MuSiC	Identification of significant genes	Ensemble of computational tools	Combination of biological concepts
MUFFINN	Identification of significant related genes	Influence of network neighbors	Interaction network data
nCOP	Identification of significant related genes	Integer linear programming and greedy algorithm	Interaction network data
DriverNet	Generation of a ranking with likely significant mutations	Greedy algorithm	Interaction network data and gene expression data
DawnRank	Generation of a ranking with likely significant mutations	Random walk	Interaction network data and gene expression data
HotNet	Identification of significant sets of related genes	Network diffusion process by diffusion kernel	Interaction network data
Hotnet2	Identification of significant sets of related genes	Network diffusion process by random walk	Interaction network data
Hierarchical HotNet	Identification of significant sets of related genes	Network diffusion process by random walk	Interaction network data
MEMo	Identification of significant sets of related genes with mutual exclusivity pattern	Node graph similarity approaches and MCMC	Interaction network data and mutual exclusivity pattern
MEMCover	Identification of significant sets of related genes with mutual exclusivity pattern in the same or across different type of cancer	Greedy algorithm	Interaction network data and mutual exclusivity pattern
GeNWeMME	Identification of significant sets of related genes with mutual exclusivity pattern	Finding connected components	Interaction network data and mutual exclusivity pattern
Dendrix	Identification of significant sets of genes with mutual exclusivity pattern	Greedy algorithm and MCMC	Mutual exclusivity pattern
CoMEt	Identification of significant sets of genes with mutual exclusivity pattern	MCMC	Mutual exclusivity pattern
WExT	Identification of significant sets of genes with mutual exclusivity pattern	Weighted test	Mutual exclusivity pattern
WeSME	Identification of significant sets of genes with mutual exclusivity pattern	Permutation-based test	Mutual exclusivity pattern
LOTUS	Classification of significant genes	Multi-task learning	Interaction network
DriverML	Classification of significant genes	Statistical score and supervised machine learning	Functional impact
MoPro	Classification of significant genes	Network propagation	Interaction network

3.4 Final Considerations

Many computational methods for identifying significant mutated genes in cancer have been developed over the years. Many computational approaches and biological knowledge are used in the algorithms of these methods. Most of them are summarized in some reviews performed by researchers in the area (HOU; MA, 2013; RAPHAEL *et al.*, 2014; ZHANG *et al.*, 2014; CHENG; ZHAO; ZHAO, 2015; DENG *et al.*, 2017; DIMITRAKOPOULOS; BEERENWINKEL, 2017; OZTURK *et al.*, 2018; CUTIGI; EVANGELISTA; SIMAO, 2020a).

In this chapter, which is based based on a paper published at the *Journal of Bioinformatics and Computational Biology (JBCB)* (CUTIGI; EVANGELISTA; SIMAO, 2020a) the importance of studying significant mutations in cancer is discussed. Some important and classic computational methods to identify such mutations are presented. The methods were briefly described, followed by a summary of related methods. The next two chapters describes the development of this thesis, with a proposal and evaluation of two computational methods: 1) DisCaGe, used for the discovery of driver genes, i.e., significantly mutated genes for cancer; and 2) DFDriver, used for the detection of possible false drivers.

CHAPTER

4

COMPUTATIONAL APPROACH FOR THE DISCOVERY OF SIGNIFICANT GENES FOR CANCER

4.1 Initial Considerations

The identification of significantly mutated genes in cancer is essential for understanding the mechanisms of tumor initiation and progression. Such a task is a key challenge, since large-scale genomic studies have reported an endless number of genes mutated at a shallow frequency. Towards uncovering infrequently mutated genes, gene interaction networks combined with mutation data have been explored. This chapter addresses the investigation of the following defined research question: ***RQ1: Can significant genes for cancer be discovered through the combination of weighted mutation frequency and network neighbors influence?***.

Figure 10 shows a summary of the general process applied to the investigation and described in this chapter. In Step 1, cancer mutation data and a set of gene networks are selected from reliable and widely used sources. In Step 2, an extensive exploratory analysis of mutation data is performed, leading to data preprocessing routines. A network characterization study and a link prediction approach are applied to selected gene networks towards the understanding of their characteristics and obtaining of an enriched version of the networks. Finally, Step 3 refers to the main contribution of this chapter, i.e., the proposal of a computational method, named DiSCaGe ((**Di**scovering **S**ignificant **Ca**ncer **Ge**nes)), that discover driver mutations and generates a ranking of prioritized driver genes. An experimental study is also conducted to evaluate the potential of DiSCaGe on the discovering of significant genes in six types of cancer.

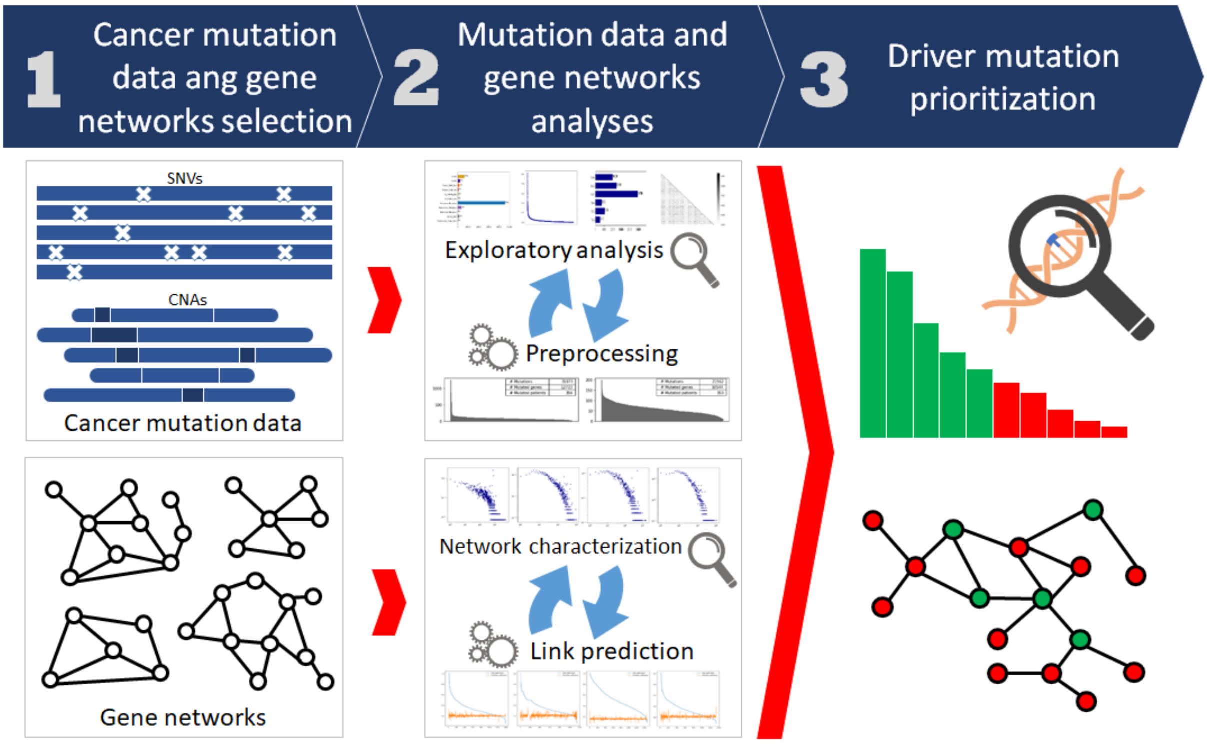

1
Cancer mutation data ang gene networks selection
2
Mutation data and gene networks analyses
3
Driver mutation prioritization
SNVs
CNAs
Cancer mutation data
Gene networks
Exploratory analysis
Preprocessing
Network characterization
Link prediction

Figure 10 – General process of discovering significant genes for cancer.

The chapter is based on two papers: The first was published at the *Brazilian Symposium on Bioinformatics (BSB 2019)* (CUTIGI; EVANGELISTA; SIMAO, 2020b), which presents the preliminary concepts of weighted mutations, and gene network enrichment. The second was submitted to Nature Scientific Reports (CUTIGI *et al.*, 2021), and describes the whole approach presented in this chapter:

- CUTIGI, J. F.; EVANGELISTA, A. F.; SIMAO, A. GeNWeMME: A network-based computational method for prioritizing groups of significant related genes in cancer. In: SPRINGER. **Advances in Bioinformatics and Computational Biology**. [S.l.], 2020. p. 29–40. ISBN 978-3-030-46417-2.

- CUTIGI, J. F.; EVANGELISTA, A. F.; REIS, R. M.; SIMAO, D. A. A computational approach for the discovery of significant cancer genes by weighted mutation and asymmetric spreading strength in networks. **Submitted to Nature Scientific Reports**, 2021.

Additionally, the exploratory study of cancer data is based on analyses published at the *Simposio Brasileiro de Computação Aplicada a Saude (SBCAS 2020)* (RAMOS *et al.*, 2020).

The chapter is organized as follows: Section 4.2 describes the cancer mutation data used in the research, which were subjected to a preprocessing routine and an extensive exploratory analysis. Section 4.3 introduces two gene networks extracted from the literature. An enrichment process was applied to the networks, and a characterization study was performed in the enriched networks. Section 4.4 describes the computational method proposed for the prioritization of significant mutated genes. Finally, Section 4.5 reports on an experimental study that evaluated the method applied to the selected cancer data and gene networks.

4.2 Cancer data

The following six cancer mutation data sets were selected for the experiments that aimed at finding significant mutations. Such types are among the most common types of cancer, according to World Health Organization (WHO, 2021), and GBM commonly appears in method's evaluation in research papers. The abbreviations were defined by TCGA, which assigns codes to each study on cancer[1].

Breast invasive carcinoma (BRCA): the most common subtype of cancer in women and one of the main causes of mortality (BING *et al.*, 2016).

Colorectal adenocarcinoma (COADREAD): represents almost 10% of the global cancer incidence. It is mainly related to old age and lifestyle factors, as dietary composition, obesity, and lack of physical activity (WILD; STEWART; WILD, 2014).

Glioblastoma multiforme (GBM): the most common, aggressive and lethal subtype of brain cancer, which shows a high growth rate and usually occurs in adults (PARSONS *et al.*, 2008).

Lung adenocarcinoma (LUAD): the leading cause of cancer death worldwide. It usually develops in smokers, however, it can also occur in non-smokers (NETWORK *et al.*, 2014).

Prostate adenocarcinoma (PRAD): one of the most common types of cancer in men. It occurs in the reproductive system, mostly in the elderly (HUANG; HE; MO, 2018).

Stomach adenocarcinoma (STAD): the fifth most common type of cancer in the world and third related to deaths (ZHOU *et al.*, 2020). It is commonly associated with a bad dietary composition.

Data sets of the following two types of mutations were selected for each type of cancer: 1) Single Nucleotide Variants (SNVs); and 2) Insertions and Deletions (InDels). The sets belong to a Pan-Cancer study of TCGA (TCGA, Cell 2018)[2] and were extracted from cBioPortal[3], which is an interactive platform for the exploration of cancer data (CERAMI *et al.*, 2012; GAO *et al.*, 2013).

[1] <https://gdc.cancer.gov/resources-tcga-users/tcga-code-tables/tcga-study-abbreviations>

[2] <https://www.ncbi.nlm.nih.gov/pubmed/29625048,29596782,29622463,29617662,29625055,29625050,29617662>

[3] <https://www.cbioportal.org/datasets>

SNVs and InDels were accessed through a mutation file in a format called MAF (Mutation Annotation Format), which is a tab-separated text file with structured mutation data. Each of its rows is a somatic mutation with more than one hundred columns of information[4], e.g., gene name (`Hugo_Symbol`), type of mutation (`Variant_Classification`), sample/patient) id (`Tumor_Sample_Barcode`), among other. Table 2 shows an example of a MAF file containing seven mutations in six different genes in three distinct patients.

Table 2 – Example of a MAF file with six columns.

Hugo_ Symbol	Chromosome	Variant_ Classification	Reference_ Allele	Tumor_Seq_ Allele2	Tumor_Sample_ Barcode
TP53	17	Missense_Mutation	G	A	TCGA-02-0003-01
NF1	17	Splice_Site	G	A	TCGA-02-0003-01
RB1	13	Nonsense_Mutation	C	T	TCGA-06-0140-01
PIK3C2A	11	Frame_Shift_Ins	-	C	TCGA-06-5416-01
TP53	17	Missense_Mutation	C	G	TCGA-06-5416-01
PTEN	10	Nonstop_Mutation	C	A	TCGA-06-0140-01
EGRF	7	In_Frame_Ins	-	CTAC	TCGA-02-0003-01

4.2.1 Preprocessing

The selected data sets were subjected to a systematic preprocessing routine. This is a crucial activity in cancer data analyses, since such data contain a large amount of information that can be suppressed (intron region, for example) when exome mutation data are analyzed. Furthermore, outlier samples can also be removed from the original data sets. The preprocessing routine involves the following two steps:

1. **Maintenance of specific somatic variants:** only specific somatic variants were kept in MAF file: `3'UTR`, `5'UTR`, `Frame_Shift_Del`, `Frame_Shift_Ins`, `In_Frame_Del`, `In_Frame_Ins`, `Missense_Mutation`, `Nonsense_Mutation`, `Nonstop_Mutation`, `Splice_Site`, and `Translation_Start_Site`. These variants were selected because they are non-silent mutations and from coding regions, i.e., they are likely to be mutations that lead to a functional impact. These selection was also validated on consultation with experts.

2. **Removal of hypermutated samples:** patients are considered hypermutated when they have a much greater number of mutations than most patients in the set. Hypermutated samples should be removed, since they are usually noisy or outliers, which can bias the analyses. Among the several strategies that identify such samples, we used the one proposed by Tamborero *et al.* (2013), according to which a sample is hypermutated when it contains more than $(Q3 + 4.5 \times IQR)$ somatic mutations, where $Q3$ is the third quartile, and IQR is the interquartile range of the distribution of mutations across all data samples. If a set of hypermutated samples is identified, it is removed from the MAF file.

[4] <https://docs.gdc.cancer.gov/Data/File_Formats/MAF_Format/>

Figure 11 displays the distribution of the mutations before and after the preprocessing task for each cancer data set. It can be noticed that the preprocessed data set the number of mutations is better distributed. For example, regarding PRAD cancer data, in the original data set, the most mutated sample has approximately 10000 mutations, whereas in the preprocessed data set, the most mutated sample had around 100 mutations. The number of mutations, mutated genes, and samples, before and after the preprocessing routine, is also shown in Table 3.

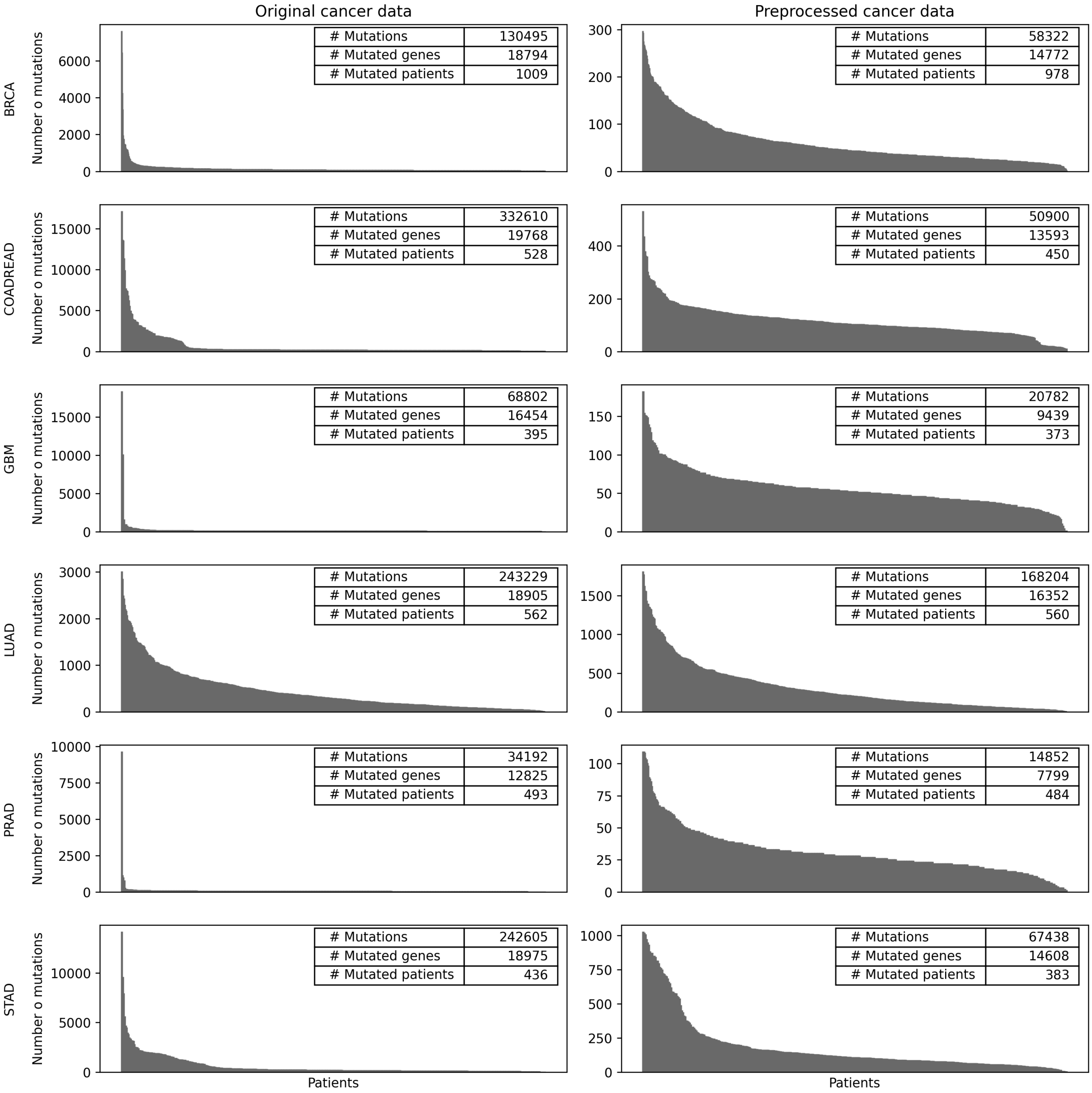


Figure 11 – Preprocessing routine: 1) Removing specific somatic mutation variants; and 2) Removal of hypermutated samples.

Table 3 – Data set characteristics

	Original data sets			Preprocessed data sets		
Code	**Mutations**	**Mutated genes**	**Samples**	**Mutations**	**Mutated genes**	**Samples**
BRCA	130495	18794	1009	58322	14772	978
COADREAD	332610	19768	528	50900	13593	450
GBM	68802	16454	395	20782	9439	373
LUAD	243229	18905	562	168204	16352	560
PRAD	34192	12825	493	14852	7799	484
STAD	242605	18975	436	67438	14608	383

No specific genes were removed from the data set. For example, FLAGS (frequently mutated genes) (SHYR *et al.*, 2014) were kept in the analyses, since the aim was the evaluation of the proposed approach with all genes of the preprocessed data set.

4.2.2 Exploratory analysis of somatic mutation cancer data

The understanding of the characteristics of cancer mutation data is crucial for a proper analysis. In this sense, an exploratory analysis with the preprocessed data was performed towards an overall view of the cancer data.

Long tail phenomenon

The long tail phenomenon shows a small number of genes is mutated in many patients, while most genes are mutated at low-frequency. Figure 12 illustrates the mutation frequency in patients for each gene. As expected, all data sets display long-tail characteristics. For example, regarding BRCA, only five genes are mutated in more than 10% of the patients.

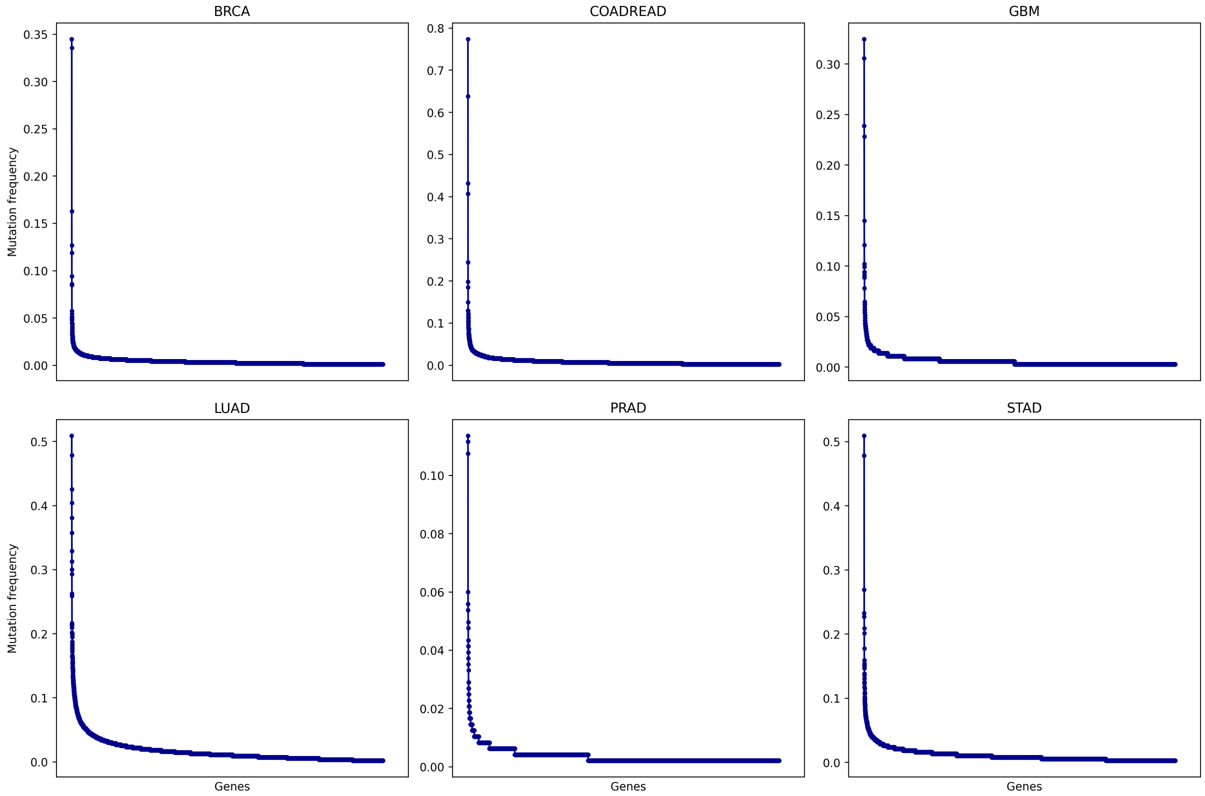


Figure 12 – Evidences of the long tail phenomenon.

Inter-tumor heterogeneity

The long tail phenomenon is directly related to the inter-tumor heterogeneity of cancer, in which two patients with the same type of cancer do not necessarily have the same set of mutations. Jaccard coefficient (ETUDE, 1901) was applied to each pair of samples (s_i, s_j) to measure the similarity between the set of mutations. Two samples s_i and s_j are considered similar if they share a large number of common mutated genes. Considering the set of mutated genes of s_i and s_j as $M(s_i)$ and $M(s_j)$, respectively, the Jaccard coefficient is calculated as follows: $J(s_i, s_j) = \frac{|M(s_i) \cap M(s_j)|}{|M(s_i) \cup M(s_j)|}$.

A set of heatmaps was created for each cancer data set to show heterogeneity. Figure 13 displays the heatmaps, in which a color scale (from white to black) indicates how similar the samples are. It can be noticed that the similarity is small between all pairs of samples in all cancer data sets, thus demonstrating their high heterogeneity.

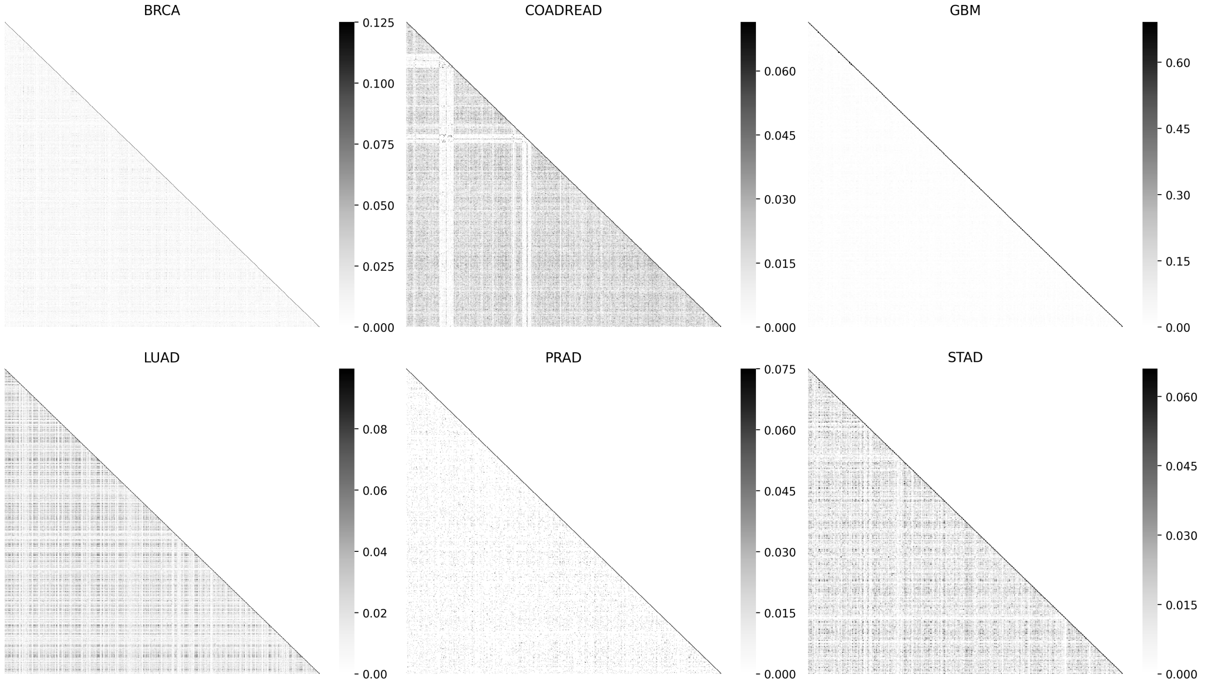


Figure 13 – Heatmaps showing the high inter-tumor heterogeneity of each type of cancer.

Distribution of somatic mutation variants

Each SNV and InDel can be found in several classes, called somatic mutation variants. Figure 14 shows the analysis in which the distribution of each variant is presented. It is possible to notice a similar distribution for all types of cancer, in which most mutations are `missense`.

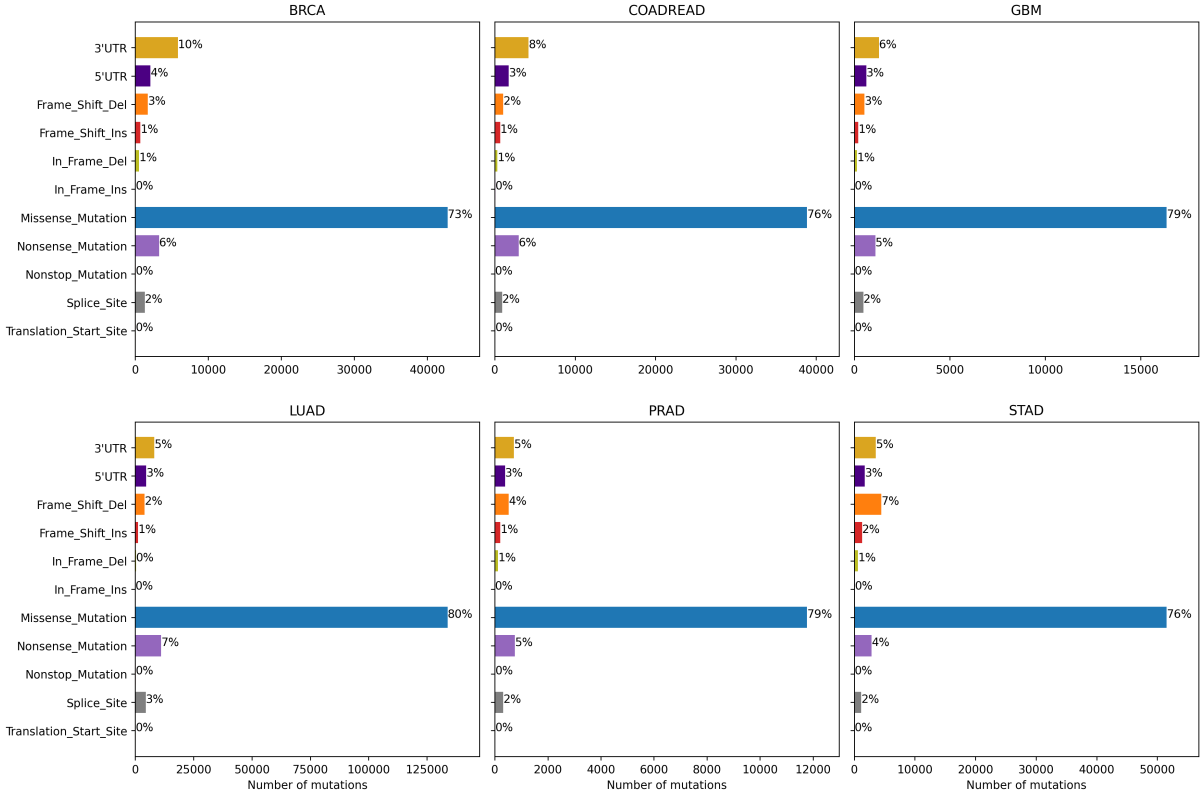


Figure 14 – Somatic mutation variant distribution.

Distribution of SNV classes

Six SNV classes represent a single change in a nucleotide. Such classes are C>A, C>G, C>T, T>A, T>C, T>G. Figure 15 shows the number of SNV classes for each type of cancer. This analysis is related to mutational signatures, which are combinations of mutations generated by different mutational processes (COSMIC, 2021). For example, class C>A is associated with tobacco smoking and the most common in LUAD.

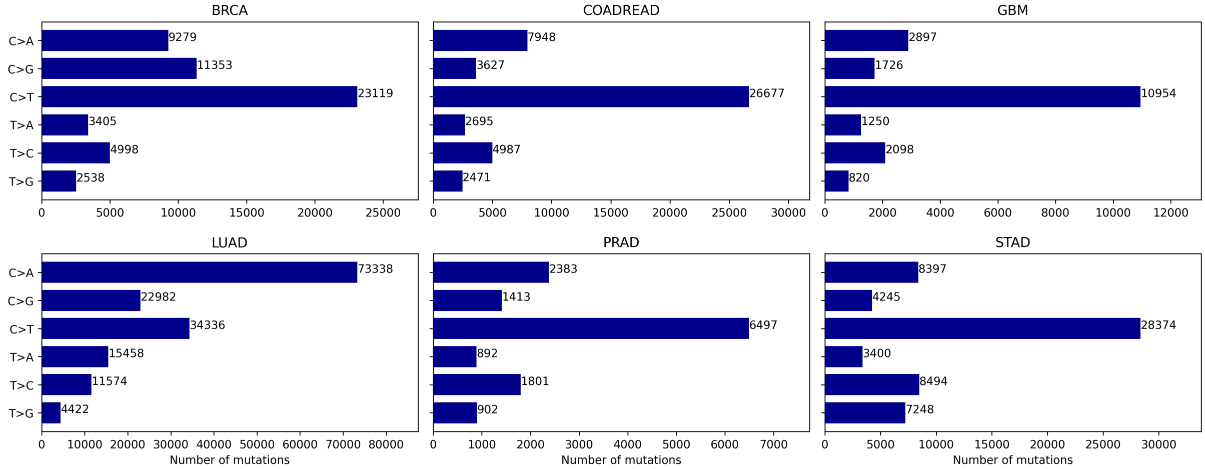


Figure 15 – SNV classes for each type of cancer.

4.3 Gene networks

Understanding the complex interactions among biological entities is fundamental for the acquisition of knowledge in Biology. Such knowledge can be modeled as a complex network for the analyses of biological systems (LIU *et al.*, 2020). In the present thesis, two gene networks that use protein-protein (PPIs) as the main source of interactions were selected for the development of the computational approach and the experiments. Below is their brief description.

Reactome Functional Interactions (Reactome): an extensive gene network built from curated pathways from many sources, and whose data are obtained mainly from reliable PPI and known pathways. Machine learning algorithms are used to train functional interactions. The curation process is performed by domain experts, following a systematic process of reviews, similarly to the editing of a scientific review. The combination of computational approaches and expert‘s review is essential for the generation of a reliable set of high-probability functional interactions (WU; HAW, 2017; FABREGAT *et al.*, 2018; JASSAL *et al.*, 2020). The Reactome group consists of an international multidisciplinary team[5], who has updated versions of Reactome since 2016. The 2019 version, released in February, 2020, was used in this work.

Human Protein Reference Database (HPRD): classical and curated human protein interactions, built from PPI, post-translational modifications, enzyme-substrate relationships, and disease associations. Such interactions were manually extracted through critical readings of studies published by Biology experts and bioinformatics analyses of the protein sequence (PERI *et al.*, 2003; PRASAD *et al.*, 2009). More than 70 laboratories have already participated in the construction of the network[6]. The last available network (Release 9) was used in this research.

Each selected gene network was treated as undirected and unweighted network $G = (V, E)$, where set of vertices $V = \{g_1, g_2, ..., g_n\}$ are genes and $(g_i, g_j) \in E$ if gene g_i interacts with gene g_j. The selected networks and subjected to an enrichment process towards inferences about possible new interactions among the genes.

4.3.1 Gene network enrichment

Biological networks are known to be incomplete (MERING *et al.*, 2002; ALOY; RUSSELL, 2004). To address this issue, a link prediction approach was used in this research for inferring interactions among genes in the network. According to the local hypothesis (BARABÁSI; GULBAHCE; LOSCALZO, 2011), two functionally related gene are likely to share common neighbors (CIRIELLO *et al.*, 2012). Szymkiewicz–Simpson coefficient ($ss(g_i, g_j)$)

[5] <https://reactome.org/about/team>
[6] <https://hprd.org/>

(SZYMKIEWICZ, 1934), also known as overlap coefficient, was used to determine how similar two genes g_i and g_j can be:

$$ss(g_i, g_j) = \frac{|N(g_i) \cap N(g_j)|}{\min(|N(g_i)|, |N(g_j)|)}$$

where $N(g_i)$ is g_i union the set of neighbors of g_i, and $N(g_j)$ is g_j union the set of neighbors of gene g_j; i.e., $N(g_i) = Neighbors(g_i) \cup \{g_i\}$ and $N(g_j) = Neighbors(g_j) \cup \{g_j\}$. The union operator considers the direct link between g_i and g_j. The overlap coefficient is extracted for each pair of nodes of a gene network *GN*, thus resulting in a new weighted gene network *wGN*, where the weight on the links is the overlap coefficient.

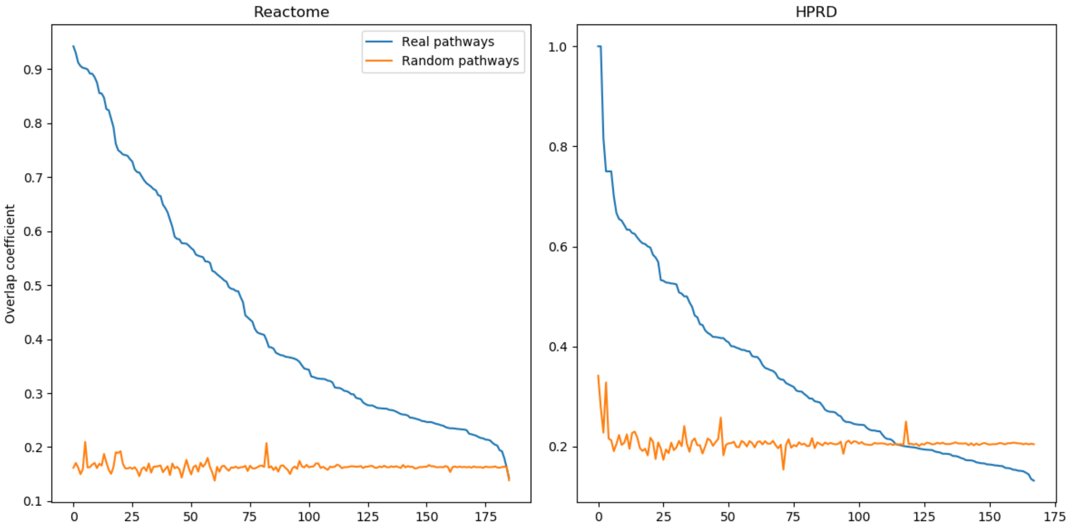
A threshold γ was defined for keeping the most significant links in *wGN*, in which only edges of a coefficient higher than γ are maintained in the network. A similar approach used by Ciriello *et al.* (2012) was applied for the choice of an appropriate γ threshold. For this, 186 known pathways derived from KEGG (KANEHISA; GOTO, 2000; KANEHISA *et al.*, 2012) and extracted from MSigBD (SUBRAMANIAN *et al.*, 2005; LIBERZON *et al.*, 2015) (database v7.2, updated September 2020[7]) were used. For each pathway p, all links among the genes of p are selected and the weight is verified in *wGN*. The average overlap coefficient among all links of p is calculated. In parallel, ten random pathways are extracted in *wGN*, with the same size of p. The link weight average is obtained for each random pathway. As a result, the average overlap coefficient in the network is extracted for each real pathway, considering the known pathway and the set of random pathways. Figure 16 shows the chart of overlap coefficient for both real and random pathways, according to the two networks used. As expected, the known pathways showed a higher overlap coefficient.

The results of this analysis suggest only links whose overlap coefficient is higher than random choices should be kept in *wGN* for maintaining interactions likely to participate in biological processes. Therefore, the median of values of all random pathways was considered. The thresholds obtained were $\gamma = 0.16$ and $\gamma = 0.20$, for Reactome and HPRD, respectively. A non-weighted enriched gene network *eGN* was extracted for each network *wGN*, in which the weight of all links in *eGN* is equal or higher than γ in the respective *wGN*. In order to avoid the possible removal of edges of the original network, existing edges were replaced in the enriched network. Algorithm 1 shows the process for the generation of an enriched gene network.

The enriched gene networks were called by their original names, with prefix *e* (e.g., the enriched version of HPRD was called eHPRD). Table 4 shows some measures of both original and enriched networks. Such measures are the number of nodes, number of edges, mean degree (average connectivity of each node), density (ratio of number of edges and number of possible edges), and number of components (subnetworks in which all nodes are connected by paths).

[7] <https://software.broadinstitute.org/cancer/software/gsea/wiki/index.php/MSigDB_Latest_Release_Notes>

Figure 16 – Choice of an appropriate threshold for the networks after link prediction with the use of overlap coefficient.



Source: Elaborated by the author.

Table 4 – Comparison of the original networks (Reactome and HPRD) with their enriched versions (eReactome and eHPRD).

Network	Nodes	Edges	Mean degree	Density	Components
Reactome	14058	268323	38.17	0.002716	84
eReactome	14058	3067180	436.36	0.031042	84
HPRD	9465	37039	7.83	0.000827	110
eHPRD	9465	415295	87.75	0.009272	110

Algorithm 1: Obtaining of enriched gene network

```
Data: A gene network $GN$; A set of known pathways $KP$.
Result: An enriched gene network $eGN$.
$wGN \leftarrow$ an empty weighted gene network;
$eGN \leftarrow$ an empty gene network;
forall $pairs\ (g_i, g_j) of genes \in GN$ do
    $N(g_i) \leftarrow$ set of neighbors of $g_i$;
    $N(g_i) \leftarrow N(g_i) \cup \{g_i\}$;
    $N(g_j) \leftarrow$ set of neighbors of $g_j$;
    $N(g_j) \leftarrow N(g_j) \cup \{g_j\}$;
    $ss(g_i, g_j) \leftarrow \frac{|N(g_i) \cap N(g_j)|}{\min(|N(g_i)|, |N(g_j)|)}$;
    $wGN \leftarrow$ link $(g_i, g_j)$ with weight $ss(g_i, g_j)$;
end
$avg\_set \leftarrow \emptyset$;
forall $pathway\ p \in KP$ do
    $ng \leftarrow$ number of genes in $p$;
    $ni \leftarrow$ number of interactions in $p$;
    $rp\_set \leftarrow$ 10 random pathways with $ng$ genes and $ni$ interactions;
    $avg\_sum \leftarrow 0$;
    forall $rp \in rp\_set$ do
        $avg\_rp \leftarrow$ average of link weights of $rp$ in $wGN$;
        $avg\_sum \leftarrow avg\_sum + avg\_rp$;
    end
    $avg\_set \leftarrow avg\_set \cup (avg\_sum/10)$;
end
$\gamma \leftarrow$ median of values in $avg\_set$;
forall $pairs\ (g_i, g_j)\ of\ genes\ in\ wGN$ do
    $w \leftarrow$ weight of the link $(g_i, g_j) \in wGN$;
    if $w > \gamma$ then
        $eGN \leftarrow$ unweighted link $(g_i, g_j)$;
    end
end
forall $pairs\ (g_i, g_j) of genes \in GN$ do
    $eGN \leftarrow$ unweighted link $(g_i, g_j)$;
end
return $eGN$;
```

Source: Elaborated by the author.

The gene network enrichment process increased the number of edges almost ten times in each network. Such a process increased the density of the networks and placed more interactions among the genes. The overall results of our approach were improved by the gene network enrichment process.

4.3.2 Characterization of gene networks

The characterization of networks is an important point for comprehending the elements and their interactions. Such networks can display distinct characteristics, depending on the way they are built. Some analyses were performed towards supporting their study.

Network measures

All enriched gene networks used in this research were characterized, and some measures and analyses performed aimed at knowing their topological characteristics. Such measures were extracted from the main connected component of each network, since some of them are applied only to connected networks. Table 5 shows a set of measures for each gene network: number of nodes (N), number of edges (E), mean degree (MD), density (DS), average shortest path (ASP), which is the average of all shortest paths between all pairs of nodes, diameter (DM), which is the shortest path of longest length, assortativity (A), which is the tendency of a node to be linked to another of a similar degree, normalized Shannon entropy (NSE), which is the heterogeneity, considering the number of connections, and average clustering (AC), which is a tendency of the nodes to form clusters.

Table 5 – Set of measures for the main component of each gene network.

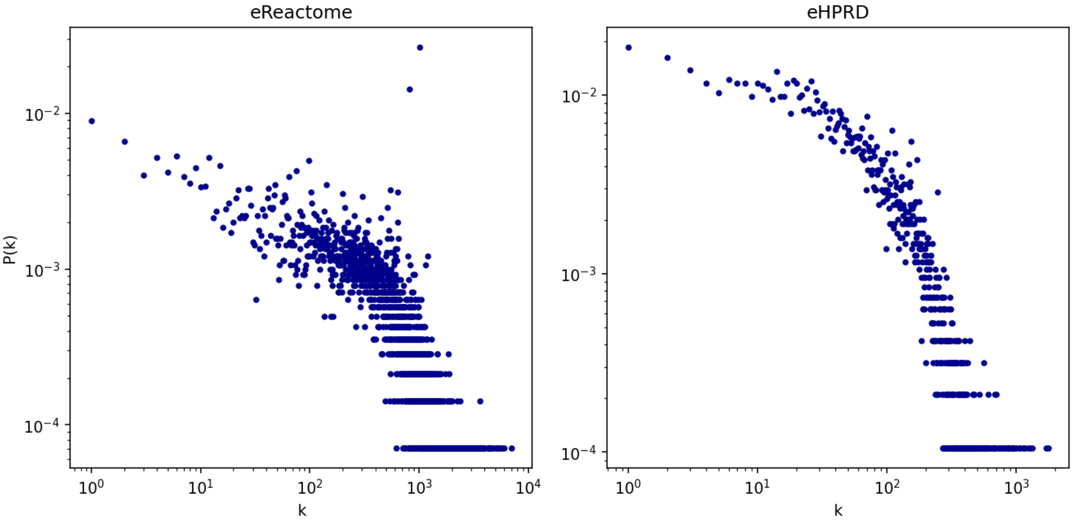
Network	N	E	MD	DS	ASP	DM	A	NSE	AC
eReactome	13864	3067033	442.45	0.031916	2.23	7	-0.07	0.71	0.52
eHPRD	9219	415125	90.06	0.009770	2.68	8	-0.01	0.59	0.43

The main component shows similar measures in comparison to the full network (see Table 4). The average shortest path (ASP) and density (DS) are small in relation to the number of nodes, which shows information in the network can be easily transferred. Assortativity (A) is not strong, since the values of the two networks are near zero, although biological networks are usually disassortative ($A < 0$). The heterogeneity of the networks is evidenced by the normalized Shannon entropy (NSE), because the NSE values are high for all networks. Average clustering (AC) shows a high value, related to the low density of the networks, i.e., such networks tend to form clusters.

Degree distribution

Degree distribution is the relation between degree (k), and the probability of a node has such a degree ($P(k)$). Figure 17 shows the degree distribution of the enriched networks on a logarithmic scale. A power-law distribution is observed in the tail of the graph, where $P(k) \sim k^{-\lambda}$, and λ is the degree exponent. Probability $P(k)$ decreases, as degree k increases, i.e., the higher the degree, the less likely a node with that degree. Such a distribution displays a scale-free network characteristic (BARABÁSI; ALBERT, 1999), which is common in many biological networks (LIU *et al.*, 2020).

Figure 17 – Degree distribution of each enriched gene network.



Source: Elaborated by the author.

The degree distribution shows the enriched gene networks are heterogeneous, i.e., most of the genes have few interactions, whereas a small number of nodes are heavily connected.

4.4 Method

This section introduces DiSCaGe (**Di**scovering **S**ignificant **Ca**ncer **Ge**nes), a network-based computational method for discovering significant genes for cancer. Such significance is directly related to the impact of different mutation types and gene interactions on networks. DiSCaGe is based on the hypothesis that genes involved in cancer tend to interact with each other (BARABÁSI; GULBAHCE; LOSCALZO, 2011), and the mutations they undergo can influence their neighborhood. This influence is extracted from asymmetric spreading strength measures of all node pairs, which take into account direct and indirect neighbors on the network. Such spreading strength is used to quantify how much its neighborhood's mutated genes can perturb a gene. The following questions and answers summarize the method and its characteristics and goals:

1. **What is the biological problem that the method seeks to solve?**

   The prioritization of genes and mutations that are significant for cancer initiation and progression.

2. **Why is this method necessary?**

   With the knowledge of significant mutations for cancer, it is possible to understand the mechanisms of the disease and personalize the cancer treatment

3. **What are the input data and hyperparameters to run the method?**

   Data: Mutation data (MAF file with SNVs and InDels); and Interaction gene networks.

   Hyperparameters: weights for each type of mutation.

4. **How is the problem computationally formulated?**

   A union and enrichment are performed in the networks, and aggregated with mutation data in each node of the network. From it, a directed and weighted network is created, with information on the propagation of the mutation strength among the genes of the network, based on direct and indirect neighbors.

4.4.1 Running example

On this chapter, a running example will be performed. For this, it is necessary to provide input data and hyperparameters that are mandatory to run the proposed method. For the running example, MAF file of Table 6 will be considered, and four gene networks illustrated in Figure 18. The required hyperparameters are presented in Table 7, which shows some defined weights for each variant classification. Such weights will be used on the running example and on the experiments, which are defined based on a consultation with expert. However, the final user is able to change such weights through a input file for hyperparameters definition.

Table 6 – MAF file for the running example.

Hugo_ Symbol	Chromosome	Variant_ Classification	Reference_ Allele	Tumor_Seq_ Allele2	Tumor_Sample_ Barcode
g_1	7	Missense_Mutation	A	T	p_3
g_1	7	Nonsense_Mutation	G	C	p_3
g_1	7	Frame_Shift_Ins	-	T	p_3
g_1	7	Frame_Shift_Del	C	-	p_3
g_1	7	Translation_Start_Site	A	T	p_3
g_1	7	Missense_Mutation	A	T	p_4
g_2	13	Frame_Shift_Ins	-	C	p_2
g_2	13	In_Frame_Ins	-	TTGTGCTTG	p_2
g_2	13	In_Frame_Del	ATTGG	-	p_2
g_2	13	Nonsense_Mutation	C	T	p_4
g_2	13	3'UTR	G	A	p_4
g_2	13	Missense_Mutation	C	T	p_5
g_3	18	Nonsense_Mutation	C	T	p_2
g_3	18	Frame_Shift_Ins	-	A	p_3
g_3	18	Translation_Start_Site	C	T	p_3
g_3	18	Missense_Mutation	C	T	p_5
g_4	1	Frame_Shift_Del	GC	-	p_3
g_4	1	Translation_Start_Site	C	T	p_3
g_4	1	Missense_Mutation	C	T	p_4
g_4	1	Missense_Mutation	G	A	p_5
g_5	11	Missense_Mutation	T	G	p_2
g_6	12	Nonsense_Mutation	C	G	p_1
g_6	12	Nonstop_Mutation	C	G	p_1
g_6	12	Missense_Mutation	C	G	p_1
g_6	12	Translation_Start_Site	G	A	p_1
g_6	12	In_Frame_Ins	-	GAA	p_3
g_7	17	Nonsense_Mutation	C	T	p_5
g_7	17	Frame_Shift_Ins	-	T	p_3
g_7	17	In_Frame_Ins	-	TTGTGCTTG	p_3
g_7	17	In_Frame_Del	CTGGCT	-	p_3
g_7	17	Frame_Shift_Ins	-	G	p_6
g_7	17	Translation_Start_Site	C	T	p_6
g_8	15	Splice_Site	C	T	p_5
g_8	15	Nonsense_Mutation	C	T	p_5
g_8	15	Frame_Shift_Ins	-	A	p_5
g_{10}	18	Frame_Shift_Del	A	-	p_1
g_{10}	18	Translation_Start_Site	C	T	p_1
g_{10}	18	In_Frame_Ins	-	TAT	p_3
g_{10}	18	Nonsense_Mutation	C	T	p_6
g_{10}	18	3'UTR	A	G	p_6
g_{11}	10	Frame_Shift_Del	A	-	p_2
g_{11}	10	In_Frame_Ins	-	TGTA	p_2
g_{11}	10	In_Frame_Del	CTA	-	p_2

Table 7 – Mutation weights.

Variant classification - *vc*	Weight - $w(vc)$
Nonsense_Mutation	1.0
Missense_Mutation	0.4
Splice_Site	0.4
Frame_Shift_Del	1.0
Frame_Shift_Ins	1.0
In_Frame_Del	0.4
In_Frame_Ins	0.4
3'UTR	0.2
5'UTR	0.4
Nonstop_Mutation	0.4
Translation_Start_Site	0.2

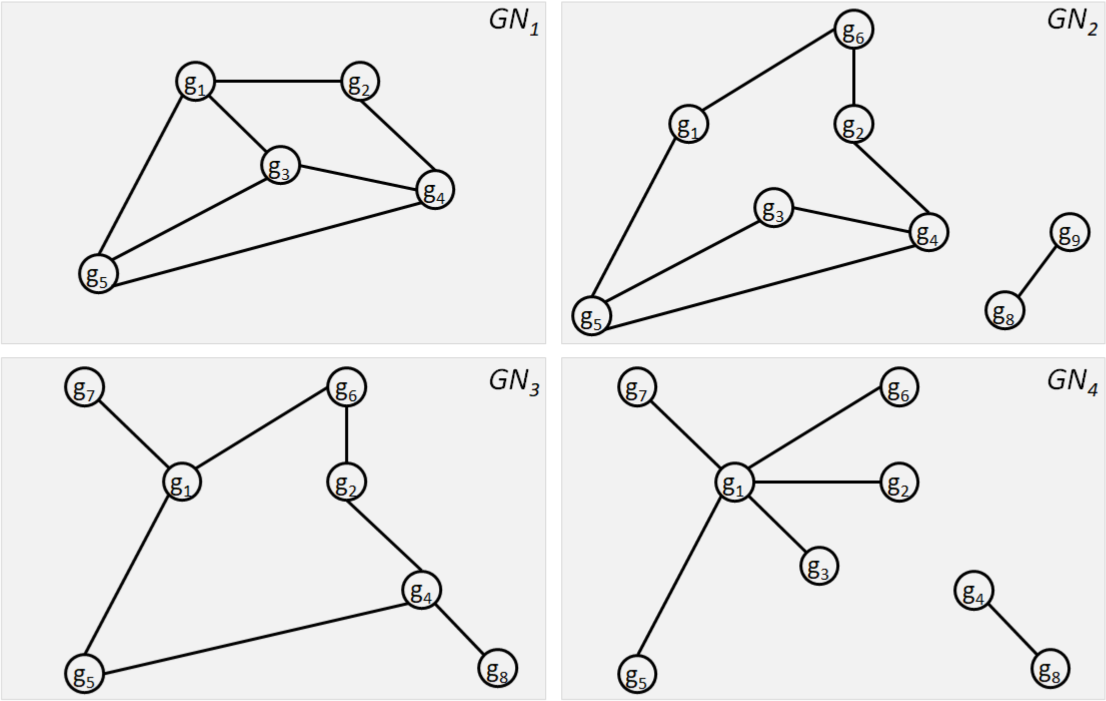
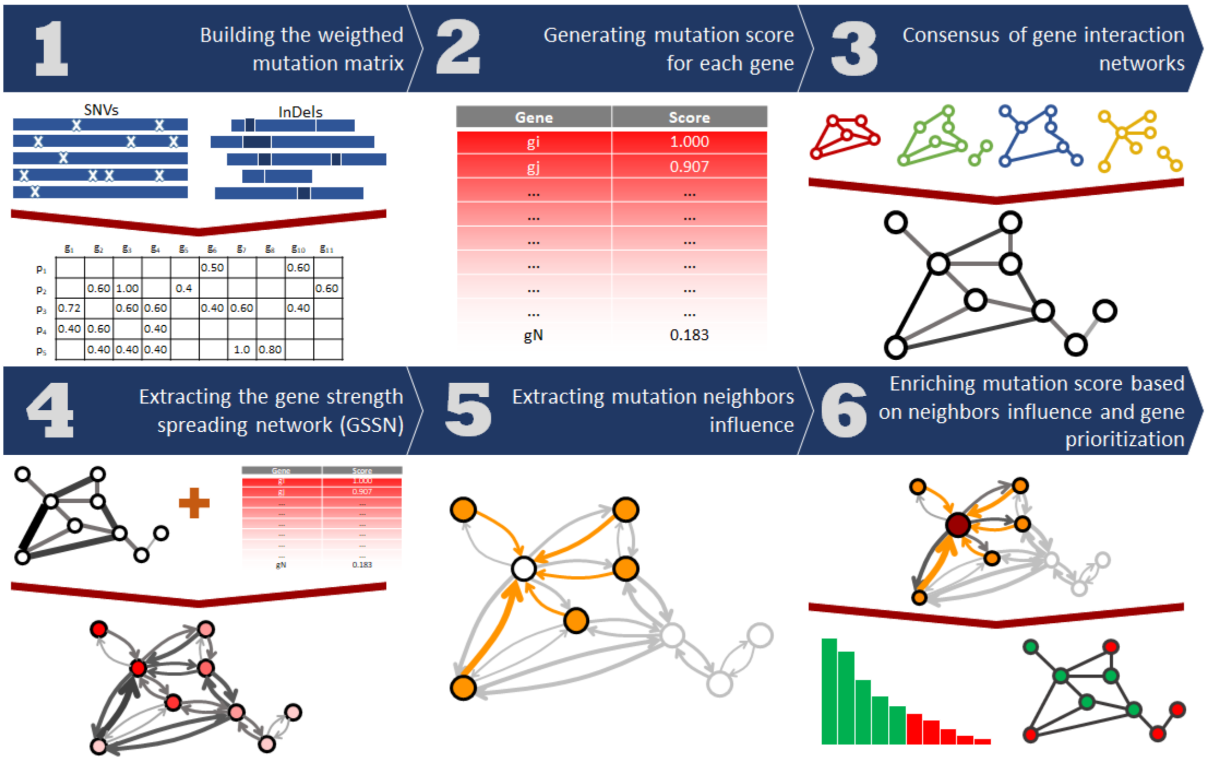


Figure 18 – Networks for the running example.

4.4.2 Method description

DiSCaGe uses cancer mutation data (SNVs and InDels in an MAF file format) and a set of $N \geq 1$ undirected and unweighted gene interaction networks (in edge lists) as input, whereas the output is a ranking of prioritized mutated cancer genes. DiSCaGe is composed of 6 steps, as illustrated in Figure 19. In Step 1, a weighted mutation matrix (WMM) is built and assigned a real value for each patient-gene pair, according to the weight defined for the variant classification of the mutation and number of mutated patients. Step 2 uses WMM to obtain a mutation score for each gene, called weighted mutation frequency. Next, in Step 3, a union operation is performed on the gene interactions networks, resulting in an undirect and weighted consensus gene interaction network. Based on such a network, in Step 4, a gene spreading strength network (GSSN) is obtained, according to the spreading strength from a gene to its direct and indirect neighbors. Step 5 extracts a mutation influence exerted on all genes by their neighbors, based on GSSN and gene mutation scores. Finally, in Step 6, each gene mutation score is enriched with the neighbors' influence, and a sorted list of prioritized genes is obtained.


1
Building the weigthed mutation matrix
2
Generating mutation score for each gene
3
Consensus of gene interaction networks
SNVs
InDels
4
Extracting the gene strength spreading network (GSSN)
5
Extracting mutation neighbors influence
6
Enriching mutation score based on neighbors influence and gene prioritization

	g1	g2	g3	g4	g5	g6	g7	g8	g10	g11
p1						0.50			0.60	
p2		0.60	1.00		0.4					0.60
p3	0.72		0.60	0.60		0.40	0.60		0.40	
p4	0.40	0.60		0.40						
p5		0.40	0.40	0.40			1.0	0.80		

Gene	Score
gi	1.000
gj	0.907
...	...
...	...
...	...
...	...
...	...
...	...
gN	0.183

Figure 19 – Approach overview.

Step 1: Building the weighted mutation matrix

In this first step, the preprocessed MAF file is used as a source for the construction of the Weighted Mutation Matrix (WMM), in which rows are patients and columns are genes. In WMM matrix *wmm*, entry $wmm_{p_i g_j}$, for each pair of patient p_i and gene g_j, a score is obtained according to its type of mutation *vc* (`Variant_Classification` from MAF input file) and in a weight $w(vc)$ assigned for each vc, as seen in Table 7. Such weights are defined in the input of the method.

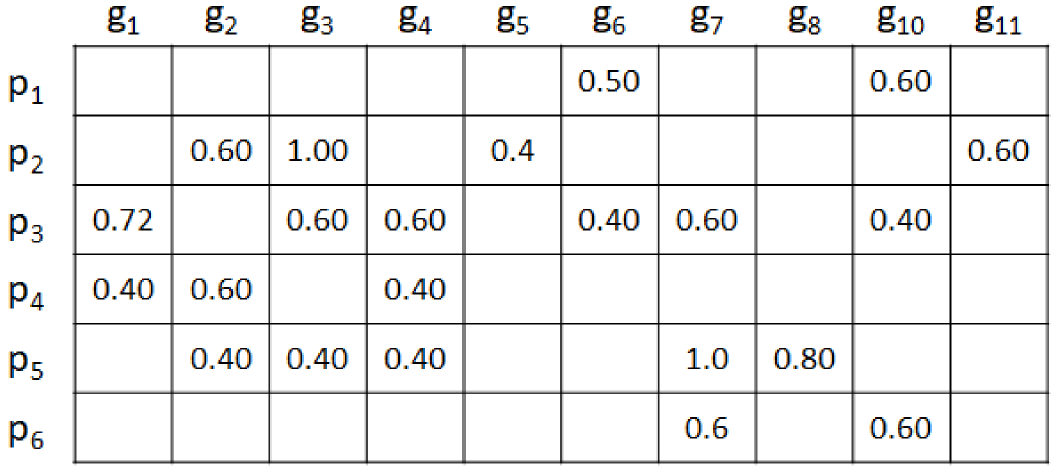
Considering a patient p_i, and all mutations in a gene g_j, entry $wmm_{p_i g_j}$ is defined as

$$wmm_{p_i g_j} = \frac{1}{|VC_{p_i g_j}|} \sum_{vc \in VC_{p_i g_j}} w(vc)$$

where $VC_{p_i g_j}$ is the list of mutations that patient p_i undergoes in gene g_j, and $w(vc)$ is the weight defined for the type of specific mutation. For example, Considering MAF file from Table 6, patient p_3 has the following list of mutations types in gene g_1: {`Missense_Mutation`, `Nonsense_Mutation`, `Frame_Shif_Ins`, `Frame_Shif_Del`, `Translation_Start_Site`}. As a result, the score $w(p_3 g_1) = \frac{0.4+1.0+1.0+1.0+0.2}{5} = 0.72$.

Such a process is performed for all patient-gene pairs. Therefore, all pairs of a mutated gene g_j in a patient p_i have score $wmm_{p_i g_j}$, which represents the importance of that mutation in that patient. The weighted average of mutations are used to consider the mutations and the possible functional impact of them, and because sometime it is necessary a set of mutations to initiate the cell carcinogenesis. Furthermore, the use of average avoids possible errors and noise in the sequencing data.

Figure 20 – An Weighted Mutation Matrix (WMM).



	g_1	g_2	g_3	g_4	g_5	g_6	g_7	g_8	g_{10}	g_{11}
p_1						0.50			0.60	
p_2		0.60	1.00		0.4					0.60
p_3	0.72		0.60	0.60		0.40	0.60		0.40	
p_4	0.40	0.60		0.40						
p_5		0.40	0.40	0.40			1.0	0.80		
p_6							0.6		0.60	

Source: Elaborated by the author.

Algorithm 2: Building *WMM*

```
Data: Mutation data $\mathscr{M}$ in MAF format; Dictionary $wc$ of variant classification
      weight
Result: Weighted Mutation Matrix (WMM)
$P \leftarrow$ set of patients of $\mathscr{M}$;
$G \leftarrow$ set of genes of $\mathscr{M}$;
$wmm \leftarrow P \times G = \{(p_i, g_j) : p_i \in P$ and $g_j \in G\}$;
forall pairs $wmm_{p_i g_j} \in wmm$ do
    $VC_{p_i g_j} \leftarrow$ list of variant classifications of pair $(p_i, g_j) \in \mathscr{M}$;
    $wmm_{p_i g_j} \leftarrow 0$;
    forall $vc \in VC_{p_i g_j}$ do
        $wmm_{p_i g_j} \leftarrow wmm_{p_i g_j} + wc(vc)$;
    end
    $wmm_{p_i g_j} \leftarrow \frac{wmm_{p_i g_j}}{|VC_{p_i g_j}|}$;
end
return $wmm$;
```

Step 2: Generating mutation score for each gene

A single score for each gene is extracted from *wmm*, and called weighted frequency $wf(g_i)$, which is the sum of the gene scores for all patients, divided by the number of patients, defined as

$$wf(g_i) = \frac{1}{|P|} \sum_{p_j \in P} wmm_{p_j g_i}$$

where P is the set of patients. For example, considering gene g_1, the weighted frequency $wf(g_1) = (0+0+0.72+0.4+0+0)/6 = 0.187$. Such a frequency is extracted for all genes, generating set wf with weighted frequencies for all genes. The final single score is normalized by the largest value in wf, thus yielding a normalized weighted frequency $nwf(g_i)$, defined as

$$nwf(g_i) = \frac{wf(g_i)}{\max_{g \in G}(wf(g))}$$

For example, considering gene g_1, $nwf(g_1) = 0.187/0.367 = 0.510$. Table 8 presents the weighted frequencies wf and normalized weighted frequencies nwf for all genes on the running example.

Algorithm 3 presents the process to obtain the mutation score from WMM generated in Step 1.

Table 8 – Weighted frequencies.

	wf	nwf
g_1	0.187	0.510
g_2	0.267	0.728
g_3	0.333	0.907
g_4	0.233	0.635
g_5	0.067	0.183
g_6	0.150	0.409
g_7	0.367	1.000
g_8	0.133	0.362
g_{10}	0.267	0.728
g_{11}	0.100	0.272

Algorithm 3: Extracting mutation score for each gene

```
Data: Weighted Mutation Matrix $wmm$
Result: A set $nwf$ of mutation score for each gene
$P \leftarrow$ set of patients of $wmm$;
$G \leftarrow$ set of genes of $wmm$;
$wf = \{\emptyset\}$;
forall $g_j \in G$ do
    $wf_{g_j} \leftarrow 0$;
    forall $p_i \in P$ do
        $wf_{g_j} \leftarrow wf_{g_j} + wmm_{p_i g_j}$;
    end
    $wf \leftarrow wf \cup \{wf_{g_j}\}$;
end
$nwf = \{\emptyset\}$;
forall $wf_{g_j} \in wf$ do
    $nwf_{g_j} \leftarrow \frac{wf_{g_j}}{max(wf)}$;
    $nwf \leftarrow nwf \cup \{nwf_{g_j}\}$;
end
return $nwf$;
```

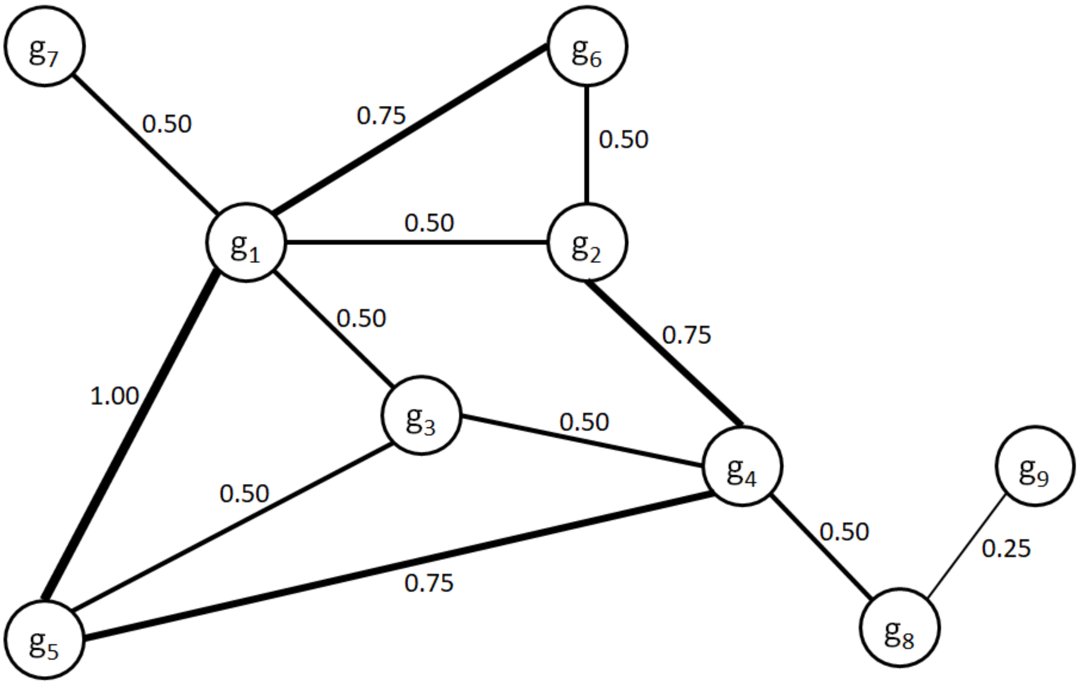
Source: Elaborated by the author.

Step 3: Consensus of gene interaction networks

An important component of DiSCaGe is gene network, which significantly impacts on the method result. Towards reducing the bias of choice of a single network, DiSCaGe accepts multiple networks as input, i.e., the method can be executed with one or more networks.

Each gene network GN_i of input set $GN_1, ..., GN_N$ was treated as undirected and unweighted networks. The union operation on these networks generates an undirected and weighted network UGN. Weights on UGN interactions are the average of times an interaction occurs in each network. For example, considering gene networks of Figure 18, the resulting UGN is illustrated in Figure 21. It can be notice that interaction (g_1, g_5) are contained in all individual networks, then such interaction has weight $w((g_1, g_5)) = 1$, while $w((g_1, g_2)) = 0.5$ because interaction (g_1, g_2) is presented in two networks (GN_1 and GN_4). Algorithm 4 shows the union process of networks to generate a consensus network UGN.

Figure 21 – A consensus network UGN, extracted from the union of gene networks of Figure 18



Source: Elaborated by the author.

Algorithm 4: Building consensus network

```
Data: A set $GN_{SET}$ of gene networks
Result: A weighted network $UGN(V, E, w)$
$UGN(V, E, w) \leftarrow$ empty graph;
$V \leftarrow$ set of all genes of $GN_{SET}$;
forall $GN(V_{GN}, E_{GN}) \in GN_{SET}$ do
    forall interaction $(g_i, g_j) \in E_{GN}$ do
        if $(g_i, g_j) \notin E$ then
            $E \leftarrow E \cup \{(g_i, g_j)\}$;
            $w((g_i, g_j)) \leftarrow \frac{1}{|GN_{SET}|}$;
        else
            $w((g_i, g_j)) \leftarrow w((g_i, g_j)) + \frac{1}{|GN_{SET}|}$;
        end
    end
end
return $UGN$;
```

Source: Elaborated by the author.

Step 4: Extraction of the Gene Strength Spreading Network (GSSN)

According to the local hypothesis (BARABÁSI; GULBAHCE; LOSCALZO, 2011), genes (and their associated proteins) involved in a certain disease tend to interact with each other, and some mutations can influence other genes in the same pathway (DING *et al.*, 2015). In this context, if a gene is mutated, such a mutation can impact its neighbors and propagate to the network.

An adapted spreading strength measure proposed by (LIU *et al.*, 2017) was defined for quantifying the spreading strength of a mutated gene through the neighborhood in UGN. Such a measure takes into account both direct and indirect neighbors, and quantifies the spread of a mutation from a node g_i to a node g_j, defined as

$$ss(g_i, g_j) = (1 + r_i \times r_j^{out}) \times p_{ij}$$

where r_i is the sum of the edge weights of g_i; r_j^{out} is the sum of the edge weights of g_j that are not edges of g_i; and p_{ij} is the weight of edge (g_i, g_j). The spreading strength is an asymmetric measure, i.e, $ss(g_i, g_j) \neq ss(g_j, g_i)$. Considering term $(1 + r_i \times r_j^{out})$, value 1 represents a single spreading from g_i to g_j and $r_i \times r_j^{out}$ denotes the impact of g_i through g_j, taking into account their indirect neighbors which are direct neighbors of g_j. At the end, such a value is tuned by the weight of the edge (g_i, g_j). The final spreading strength measure is normalized by the largest value of ss, thus obtaining a normalized spreading strength $nss(g_i, g_j)$, defined as

$$nss(g_i, g_j) = \frac{ss(g_i, g_j)}{\max_{(g,g') \in G \times G}(ss(g, g'))}$$

For example, considering UGN of Figure 21 and the strength spreading from g_3 to g_4:

$r_{g_3} = p_{g_3g_1} + p_{g_3g_4} + p_{g_3g_5} = 0.5 + 0.5 + 0.5 = 1.50$

$r_{g_4}^{out} = p_{g_4g_2} + p_{g_4g_8} = 0.75 + 0.5 = 1.25$

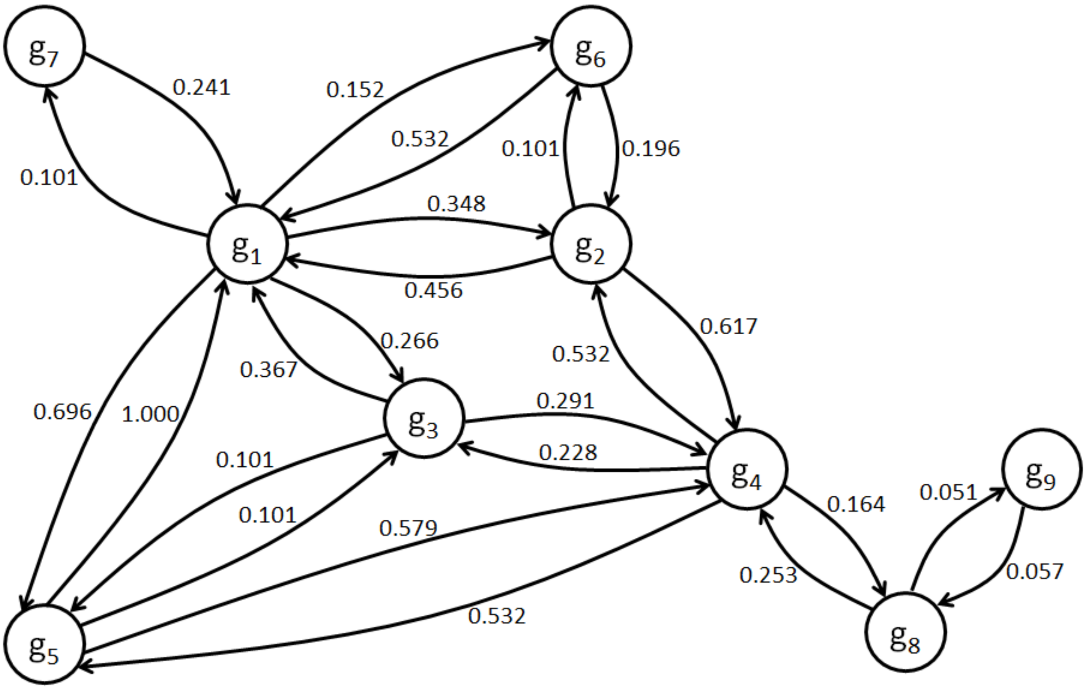
$p_{g_3g_4} = 0.50$

$ss(g_3, g_4) = (1 + r_{g_3} \times r_{g_4}^{out}) \times p_{g_3g_4} = (1 + 1.50 \times 1.25) \times 0.50 = 1.438.$

Performing the normalization by the maximum value of the set ss of spreading strength measure, $nss(g1) = 1.438/4.938 = 0.291$.

After the extraction of the normalized spreading strength measure for all neighbor genes, a directed and weighted network, called Gene Spreading Strength Network (GSSN), is obtained. In GSSN, directed interaction weights represent the degree of spreading at which a mutation in a gene g_i can pass through a gene g_j. Finally, the mutation score of each gene is assigned to the GSSN network. Figure 22 presents the Gene Strength Spreading Network $GSSN$ generated from UGN of Figure 21. Algorithm 5 shows the process to build the $GSSN$.

Figure 22 – A Gene Strength Spreading Network *GSSN*, extracted from the network of Figure 21



Source: Elaborated by the author.

```
Algorithm 5: Building Gene Strength Spreading Network GSSN
    Data: An weighted network $UGN(V', E', w')$
    Result: A directed and weighted network $GSSN(V, E, w)$
    $GSSN(V, E, w) \leftarrow$ empty directed and weighted graph;
    $V \leftarrow V'$;
    forall interaction $(g_i, g_j) \in E_{GN}$ do
        $N_{g_i} \leftarrow$ set of neighbors of $g_i$ in $UGN$;
        $N_{g_j} \leftarrow$ set of neighbors of $g_j$ in $UGN$;
        $r_{g_i} \leftarrow \sum_{g}^{N_{g_i}} w'(g_i, g)$;
        $r_{g_j} \leftarrow \sum_{g}^{N_{g_j}} w'(g_j, g)$;
        $r_{g_i}^{out} \leftarrow \sum_{g}^{N_{g_i} \setminus N_{g_j}} w'(g_i, g)$;
        $r_{g_j}^{out} \leftarrow \sum_{g}^{N_{g_j} \setminus N_{g_i}} w'(g_j, g)$;
        $p_{g_i g_j} \leftarrow w'(g_i, g_j)$;
        $ss(g_i, g_j) \leftarrow (1 + r_{g_i} \times r_{g_j}^{out}) \times p_{g_i g_j}$;
        $ss(g_j, g_i) \leftarrow (1 + r_{g_j} \times r_{g_i}^{out}) \times p_{g_i g_j}$;
        $E \leftarrow E \cup (g_i, g_j)$ of weight $w(g_i g_j) = ss(g_i, g_j)$;
        $E \leftarrow E \cup (g_j, g_i)$ of weight $w(g_j g_i) = ss(g_j, g_i)$;
    end
    return $GSSN$;
```

Source: Elaborated by the author.

Step 5: Extraction of mutation neighbors influence

The spreading strength among genes represents how much a single gene can be affected by mutations of its neighborhood, and how much it can affect its neighbors. In this step, the received influence of a mutated gene is extracted by a function $r(g_i)$, which represents how much influence g_i receives from its neighbors, defined as

$$r(g_i) = \sum_{g_k \in N(g_i)} nwf(g_k) \times nss(g_k, g_i)$$

where $N(g_i)$ are direct neighbors of g_i on GSSN.

For example, considering gene g_2 and and its neighbors $N(g_2) = \{g_1, g_4, g_6\}$:

$r(g_2) = \sum_{g_k}^{N(g_2)} nwf(g_k) \times w(g_k, g_2)$

$r(g_2) = nwf(g_1) \times w(g_1, g_2) + nwf(g_4) \times w(g_4, g_2) + nwf(g_6) \times w(g_6, g_2)$

$r(g_2) = 0.510 \times 0.348 + 0.635 \times 0.532 + 0.409 \times 0.196$

$r(g_2) = 0.177 + 0.338 + 0.080$

$r(g_2) = 0.595$

With this, gene g_2 receive from its neighbors a mutation influence score of 0.595. After the calculation of $r(g_i)$ for all genes of *GSSN*, a maximum value normalization is applied on $r(g_i)$, as follows:

$$nr(g_i) = \frac{r(g_i)}{\max_{g \in G}(r(g))}$$

For example, considering gene g_2, $nr(g_2) = 0.595/1.306 = 0.455$. Table 9 presents neighbors mutation influence r and normalized influence nr for all genes on the running example. Algorithm 6 shows the process to extracted the neighbors influence.

Table 9 – Neighbors mutation influence.

	r	nr
g_1	1.306	1.000
g_2	0.595	0.455
g_3	0.299	0.229
g_4	0.911	0.697
g_5	0.784	0.600
g_6	0.151	0.116
g_7	0.052	0.040
g_8	0.105	0.080
g_9	0.018	0.014
g_{10}	0.000	0.000
g_{11}	0.000	0.000

Algorithm 6: Extraction of mutation neighbors influence

```
Data: Mutation score $nwf$ for each gene; A directed and weighted network
      $GSSN(V,E,w)$
Result: A mutation influence $nr$ received for each gene
$r \leftarrow \{\emptyset\}$;
forall $g_i \in V$ do
    $N(g_i) \leftarrow$ neighbors of $g_i$;
    $r(g_i) \leftarrow 0$;
    forall $g_k \in N(g_i)$ do
        $r(g_i) \leftarrow nwf(g_k) \times w(g_k, g_i)$;
    end
    $r \leftarrow r \cup \{r(g_i)\}$;
end
$nr \leftarrow \{\emptyset\}$;
forall $r_{g_i} \in r$ do
    $nr_{g_i} \leftarrow \frac{r_{g_i}}{max(r)}$;
    $nr \leftarrow nr \cup \{nr_{g_i}\}$;
end
return $nr$;
```

Source: Elaborated by the author.

Step 6: Gene mutation score enrichment based on GSSN and gene prioritization

In this step, the final mutation score of each mutated gene g is obtained, taking into account the individual mutation score of gene $nwf(g)$ and the influence $nr(g)$ score from its neighbors. The final mutation score $ms(g_i)$ of a gene g_i is the sum of its mutation score and its neighbors mutations influence, given by

$$ms(g_i) = nwf(g_i) + nr(g_i)$$

For example, gene g_2 has own mutation score $nwf(g_2)$ of 0.728, and receives from its neighbors $r(g_2)$ of 0.455, thus resulting in a final mutation score $ms(g_2) = nwf(g_2) + nr(g_2) = 0.728 + 0.455 = 1.183$.

After $ms(g)$ has been obtained for all mutated genes, the final ranking of prioritized genes is extracted through their sorting by ms. Mutated genes with the highest ms values are likely to be significantly mutated and related with significantly mutated neighbors. Table 10 shows the final mutation score $ms(g_i)$ for every mutated gene g_i of the running example, sorted by $ms(g_i)$.

Table 10 – Final mutation score.

gene	$nwf(g_i)$	$r(g_i)$	$ms(g_i)$
g_1	0.510	1.000	1.510
g_4	0.635	0.697	1.332
g_2	0.728	0.455	1.183
g_3	0.907	0.229	1.136
g_7	1.000	0.040	1.040
g_5	0.183	0.600	0.783
g_{10}	0.728	0.000	0.728
g_6	0.409	0.116	0.525
g_8	0.362	0.080	0.442
g_{11}	0.272	0.000	0.272

4.5 Experimental study

The evaluation of computational methods that identify significant mutations in cancer remains a challenging task (CUTIGI; EVANGELISTA; SIMAO, 2020a). The lack of gold standard databases for driver and passenger genes hampers the obtaining of an optimal measure of the output. *In-vivo* or *in-vitro* biological laboratory experiments could be performed to analyze the suggested cancer genes found by the computational methods. However, they require considerable time and are costly.

In this context, prior to laboratory experiments, prioritized cancer genes should be considered highly reliable, and *in-silico* experiments can be performed. For this, in this research, the results of the method will be evaluates as follows:

4.5.1 Evaluation metrics

4.5.1.1 Precision

Precision is the fraction of prioritized genes that are known related to cancer. The precision of the ranking can be computed, and is obtained by

$$Precision = \frac{|PG \cap D|}{|PG|}$$

where PG is the set of prioritized genes and D is the set of known driver genes.

To extract precision of the results, a set of known driver genes D is necessary. Despite the lack of a gold standard for driver and passenger genes, some gene databases are widely used and continuously updated. Towards a well-defined set of known cancer genes, the following four reliable and recent available benchmarks were considered:

1. A D_{NCG} set of 711 known cancer drivers extracted from Network of Cancer Genes (NCG)[8] (REPANA *et al.*, 2019).

2. A D_{CGC} set of 723 driver genes extracted from Cancer Gene Census (CGC)[9] (FUTREAL *et al.*, 2004; SONDKA *et al.*, 2018).

3. A $D_{IntOGen}$ set of 568 driver genes extracted from Integrative OncoGenomics (IntOGen)[10] (MARTÍNEZ-JIMÉNEZ *et al.*, 2020).

4. A D_{Bailey} set of 299 driver genes extracted from the recent and extensive study conducted by (BAILEY *et al.*, 2018)[11].

Considering the described four set of known drivers, a union of all lists was performed, thus resulting in a single list D of 951 known cancer drivers, i.e., $D = D_{NCG} \cup D_{CGC} \cup D_{IntOGen} \cup D_{Bailey}$. For example, considering the following list of five ($p = 5$) prioritized genes: $PG_5 = \{TP53, TTN, EGFR, SEPT9, CDKN2A\}$, four genes $PG_5 \cap D = \{TP53, EGFR, SEPT9, CDKN2A\}$ are known to be driver genes. Thus, considering $P = 5$, the $Precision_5 = 0.8$, i.e, 80% of the prioritized genes are known to be related to cancer.

It is important to discuss that is a hard task to determine an ideal value for the precision. For example, if all the prioritized genes are known drivers, the result was not able to bring any possible novel information, thus, the gene prioritization method is nos useful. On the other hand, if none of the prioritized genes are known driver, the result seems to be random. So, the results should be analyzed not in a binary way, but in perspective with other analysis and specialist validation.

4.5.1.2 Discounted cumulative gain (DCG)

Discounted cumulative gain (DCG) (JäRVELIN; KEKäLäINEN, 2002) is a measure of the set prioritized genes that considers their position on the ranking and relevance of each gene. It can be used for the analysis of how good a set of genes is. Two same size ranking lists of genes can be compared, even if the genes are the same, but are not in the same position.

To allow the extraction of DCG, a relevance must be assigned for each gene. For this, it is used information about the driver gene benchmarks D_{NCG}, D_{CGC}, $D_{IntOGen}$, and D_{Bailey} presented in previous section. In addition, other gene benchmarks were considered:

1. An FD_{NCG} set of 250 genes listed as possible false positive drivers[12] by the Network of Cancer Genes (NCG).

[8] Version 6.0 – http://ncg.kcl.ac.uk/download.php

[9] Version 92, 27-AUG-20 – https://cancer.sanger.ac.uk/census

[10] Release 2020-02-01 – https://www.intogen.org/search

[11] Baylei et. al 2018 – https://pubmed.ncbi.nlm.nih.gov/29625053/

[12] Version 6.0 – http://ncg.kcl.ac.uk/false_positives.php

2. Six $SD_{IntOGen}^{cancer\_type}$ sets of specific drivers for each type of cancer, based on the known specific driver from IntOGen, i.e., each set contains genes related to a specific type of cancer, whose specific benchmark is essential, since many genes are important in specific types of cancer, and probably irrelevant in others (LEVER *et al.*, 2019). The sets $SD_{IntOGen}^{BRCA}$, $SD_{IntOGen}^{COADREAD}$, $SD_{IntOGen}^{GBM}$, $SD_{IntOGen}^{LUAD}$, $SD_{IntOGen}^{PRAD}$, and $SD_{IntOGen}^{STAD}$ have 99, 72, 35, 42, 82 and 61 specific drivers, respectively.

With all driver genes benchmarks, it is possible to have the relevance of all genes, considering all types of cancer. In this context, to get the relevance $rel_{g_i}^{ct_j}$, of a gene g_i in the cancer type ct_j a score for the presence of g_i in each gene benchmark is assigned. The relevance $rel_{g_i}^{ct_j}$ is incremented by 1 for each time that g_i are contained in D_{NCG}, D_{CGC}, $D_{IntOGen}$, or D_{Bailey}. If g_i is contained in the specific driver benchmark SD_{ct_j}, $rel_{g_i}^{ct_j}$ is incremented by 4, in order to specific cancer drivers have the same weight if it appears in all four general benchmarks. If g_i is contained in FD_{NCG}, $rel_{g_i}^{ct_j}$ is decreased by 1. Finally, if g_i is not in any gene set, its value is zero.

For example, Table 11 presents the relevance of five genes for STAD. The relevance of gene TP53 for STAD $rel_{TP53}^{STAD} = 8$. Such relevance are used to extracted the DCG for a ranking of prioritized genes.

Table 11 – Relevance of five genes for STAD

Gene	D_{NCG}	D_{CGC}	$D_{IntOGen}$	D_{Bailey}	FD_{NCG}	$SD_{IntOGen}^{STAD}$	rel_g^{STAD}
TP53	1	1	1	1	0	4	8
TTN	0	0	0	0	-1	0	-1
EGFR	1	1	1	1	0	0	4
SEPT9	1	1	1	0	0	0	3
CDKN2A	1	1	1	1	0	4	8

Discounted cumulative gain (DCG_p) of a ranking of genes up to position p is a measure that takes account the relevance of the genes and their position in the ranking, being reduced logarithmically proportional to this position. DCG_p is defined as follows:

$$DCG_p = \sum_{i=1}^{PG_p} \frac{rel_{g_i}^{ct_j}}{\log_2(i+1)}$$

where PG_p is the ranking list of p prioritized genes. For example, considering $PG_5 = \{TP53, TTN, EGFR, SEPT9, CDKN2A\}$, and the gene relevance rel_g^{STAD} of Table 11.

$DCG_5 = \frac{rel_{TP53}^{STAD}}{\log_2(1+1)} + \frac{rel_{TTN}^{STAD}}{\log_2(2+1)} + \frac{rel_{EGFR}^{STAD}}{\log_2(3+1)} + \frac{rel_{SEPT9}^{STAD}}{\log_2(4+1)} + \frac{rel_{CDKN2A}^{STAD}}{\log_2(5+1)}$

$DCG_5 = \frac{8}{\log_2 2} + \frac{-1}{\log_2 3} + \frac{4}{\log_2 4} + \frac{3}{\log_2 5} + \frac{8}{\log_2 6}$

$DCG_5 = 13.756$

4.5.2 Results

A set of experiments was designed to show the potential of the computational approach proposed in this research. Such experiments were evaluated using the criteria defined on the previous section and cancer data and gene interaction network presented in Section 4.2 and 4.3 respectively.

For the DiSCaGe hyperparameter, i.e., weights of each type of mutation, the values of Table 7 were considered. However, a study of the influence of hyperparameters is presented in Section 4.5.2.10.

4.5.2.1 Precision and DCG

Figure 23 shows a plot with precision for each type of cancer, from the first prioritized genes up to position 200. For example, considering BRCA and a list of 50 prioritized genes, nearly 60% of genes are contained in the driver benchmarks, i.e., they are likely to be a known driver.

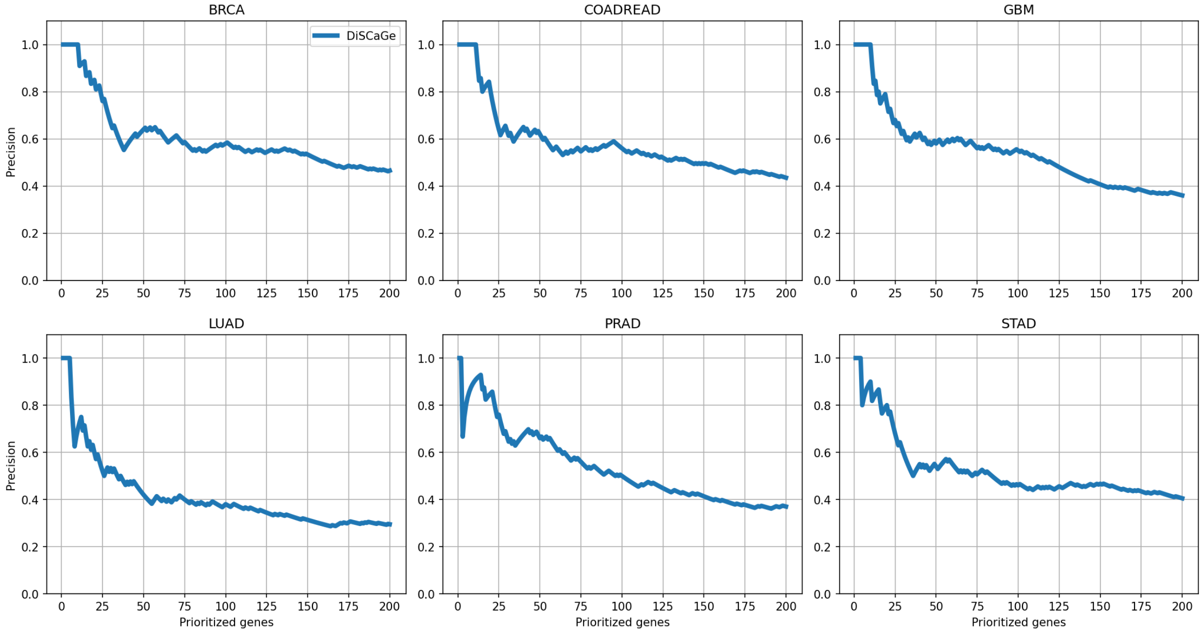


Figure 23 – Precision

Figure 24 shows a plot with DCG at a specific position. For example, considering BRCA and a list of 25 prioritized genes, DCG is 40. It can be noticed that is not natural to interpret DCG when presented alone. However, such measure can be used together with precision to infer the quality of a rank. Furthermore, it can be used to compare different rankings. For example, ranking with the same genes but in different positions can be compared using DCG.

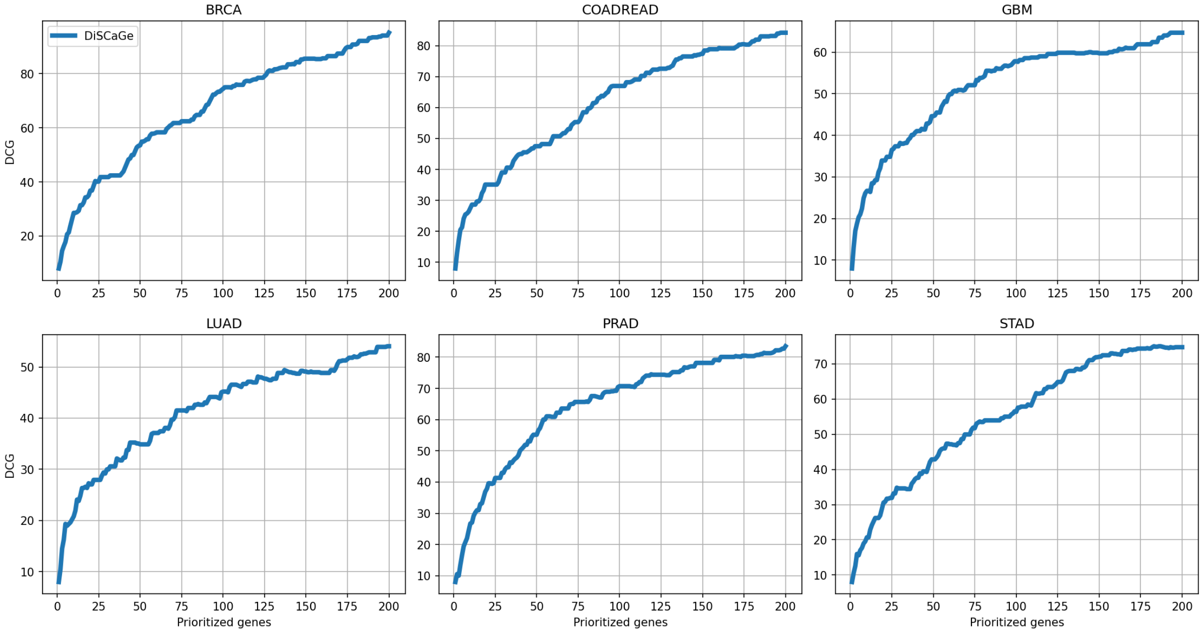


Figure 24 – DCG

4.5.2.2 Comparison with mutation frequency

Not all most frequent mutated genes are considered drivers. Frequency-based methods use complex approaches to estimate background mutation rate to extracted frequency based on this estimation, which is a challenging task. In this context, the prioritized genes should outperform a ranking of genes sorted by their simple mutation frequency. Figure 25 and 26 presents precision and DCG, respectively, of the results of DiSCaGe in comparison with simple frequency. It can be noticed that DiSCaGe outperforms the results obtained with the simple mutation frequency for all types of cancer.

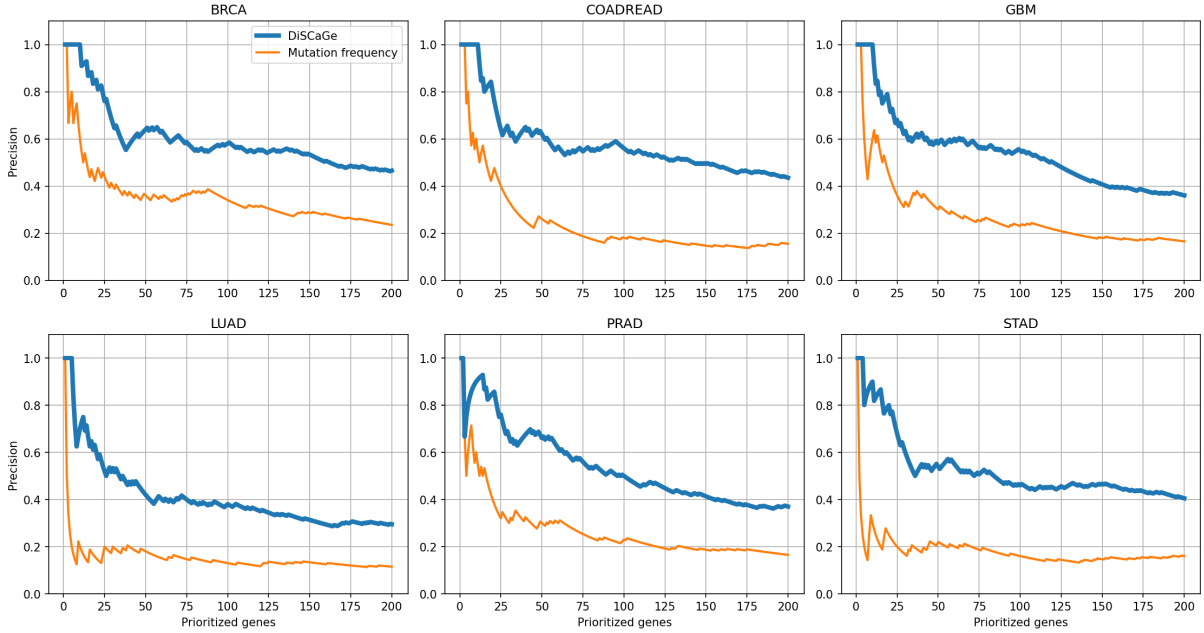


Figure 25 – Precision: DiSCaGe x Mutation frequency

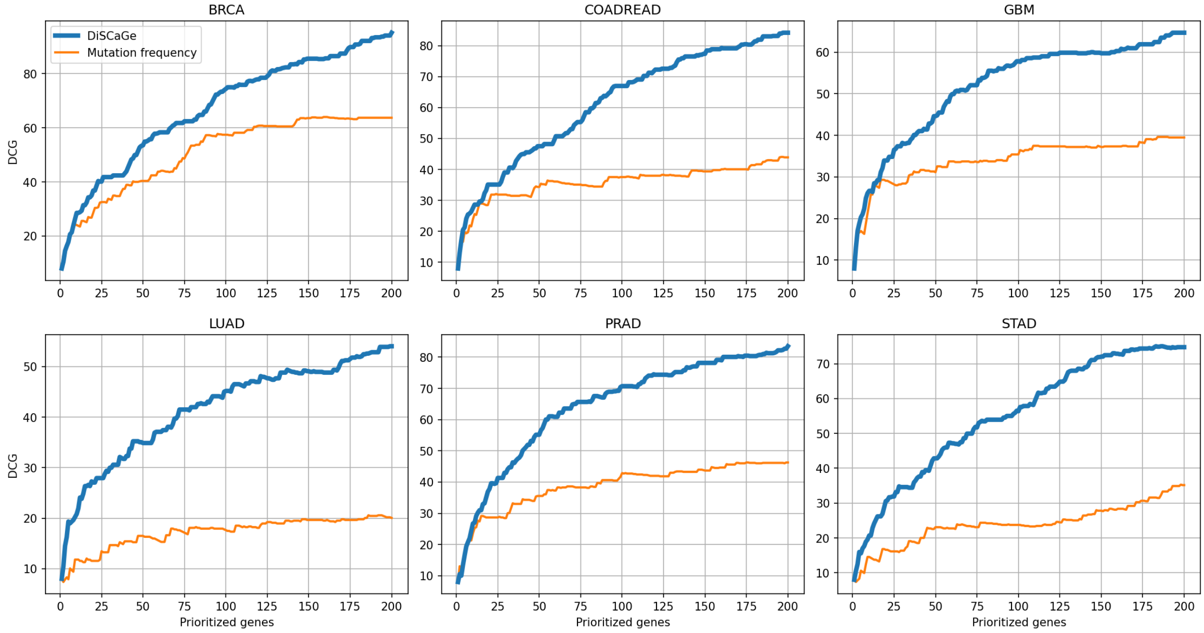


Figure 26 – DCG: DiSCaGe x Mutation frequency

4.5.2.3 Comparison with weighted frequency

In Step 2 of the DiSCaGe, a mutation score is generated for each gene. Such score is a normalized weighted frequency $nwf(g)$, based on the type of mutations that gene g has on the mutation data. In this experiment, a comparison between the weighted frequency nwf, and the final mutation score ms is performed. Figure 27 and 28 presents precision and DCG, respectively. It can be noticed that the final mutation score obtained by DiSCaGe outperforms the weighted score based only on the gene mutation. It suggests that the neighbors' influence has a significant impact on the performance of the method.

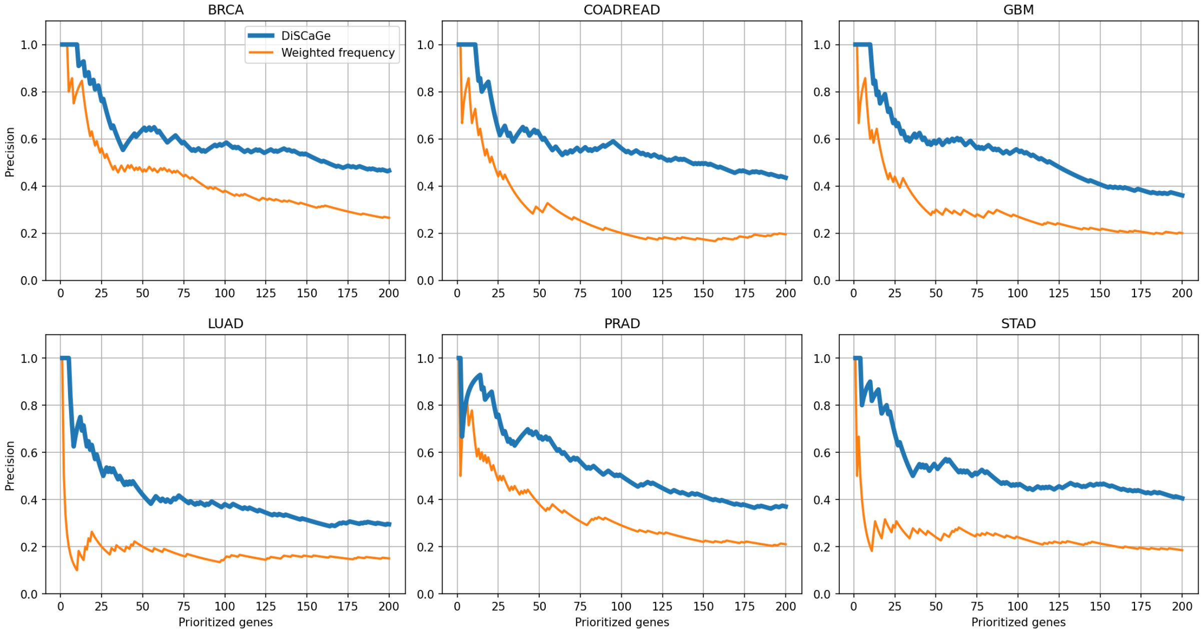


Figure 27 – Precision: DiSCaGe x Weighted frequency

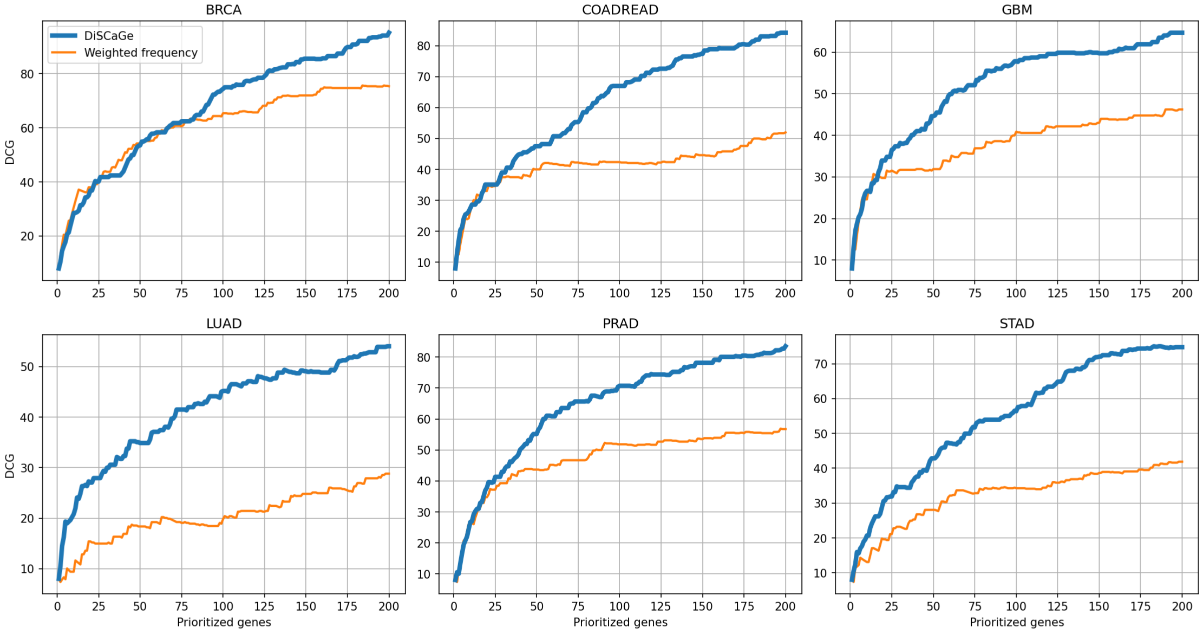


Figure 28 – DCG: DiSCaGe x Weighted frequency

4.5.2.4 Comparison with mutation influence from neighbors

Step 4 and Step 5 of the approach extract information to measure the influence of the direct and indirect neighbors on the mutated genes. In this experiment, the neighbors influence $r(g)$ that a gene g receive from its neighbors is compared with the final mutation score $ms(g)$.

Figure 29 and 30 presents precision and DCG, respectively. It can be noticed that the precision is similar, considering both scores. However, comparing DCG for the score, it can be noticed that the DiSCaGe outperforms the neighbors' influence in all types of cancer, except in LUAD. It suggests that, although the precision is similar, DiSCaGe is likely to return a ranking of genes with high quality. However, it is important to notice that the neighbors' influence has a significant impact on the final result of the method.

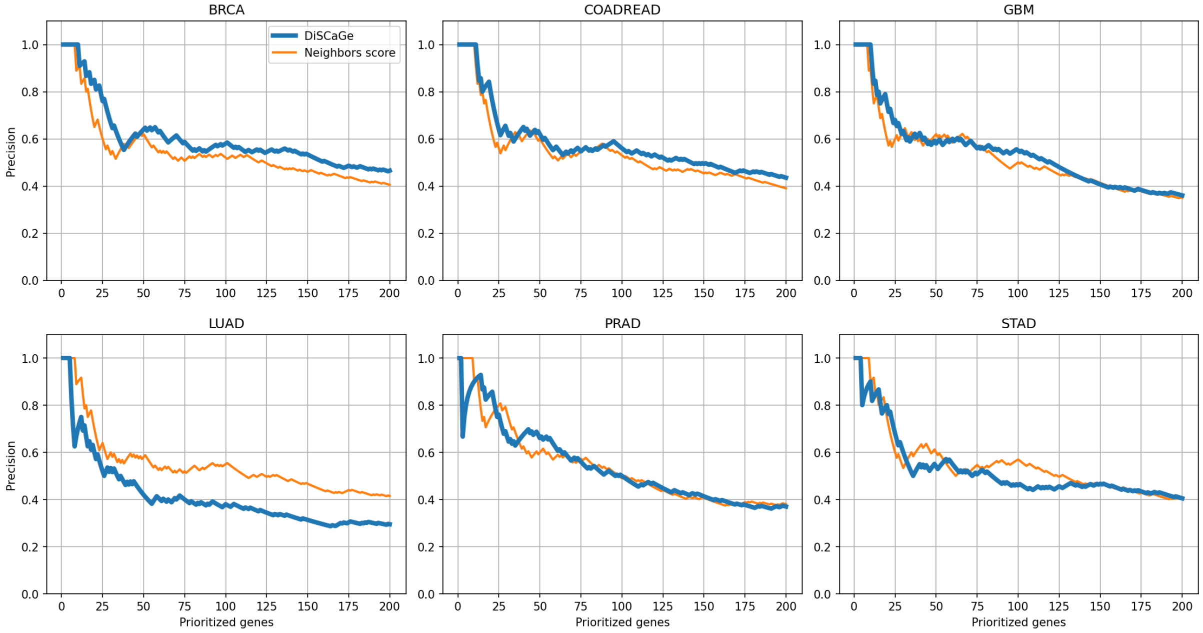


Figure 29 – Precision: DiSCaGe x Neighbors influence

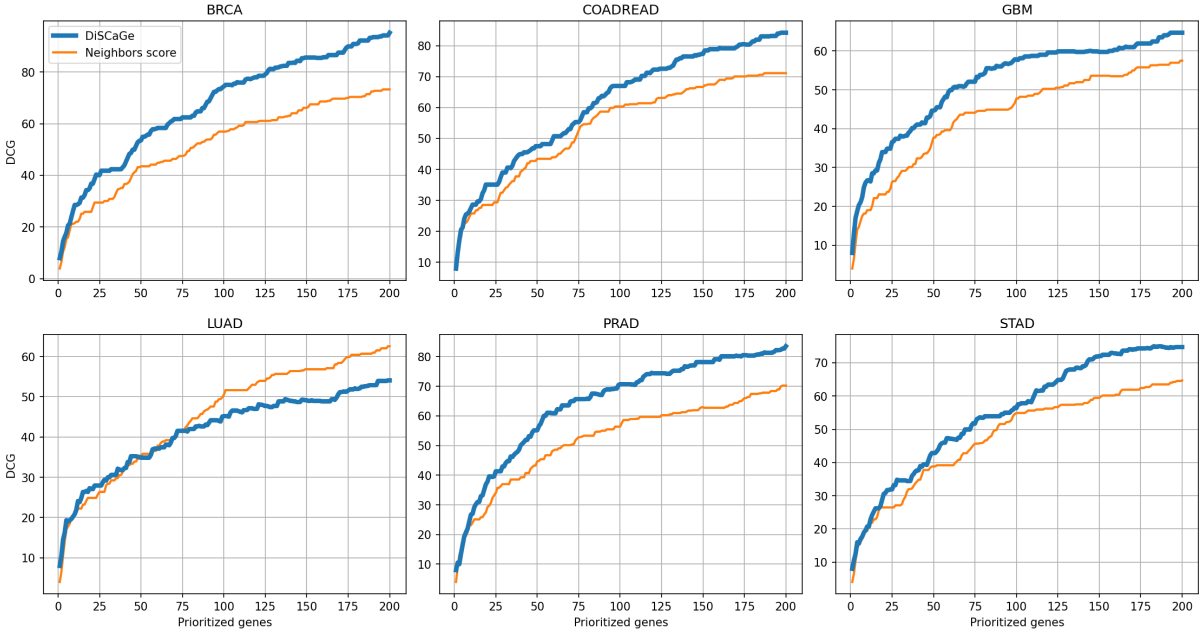


Figure 30 – DCG: DiSCaGe x Neighbors influence

4.5.2.5 Prioritization of low-frequency mutated genes

DiSCaGe can find low-frequency cancer mutated genes. Figure 31 shows the long-tail chart for each cancer data set studied. The top 30 prioritized genes are highlighted in the charts, where red dots are genes known to be related to cancer, and blue ones are possible cancer genes prioritized by DiSCaGe. The gene names can be observed in the matrix of the Figure 34, in which the blue genes are not contained in any benchmark, i.e., the matrix column are composed by only white cells. Several prioritized genes are on the tail of the graph, thus showing the potential of DiSCaGe for prioritizing known and low-frequency cancer genes and possible novel ones.

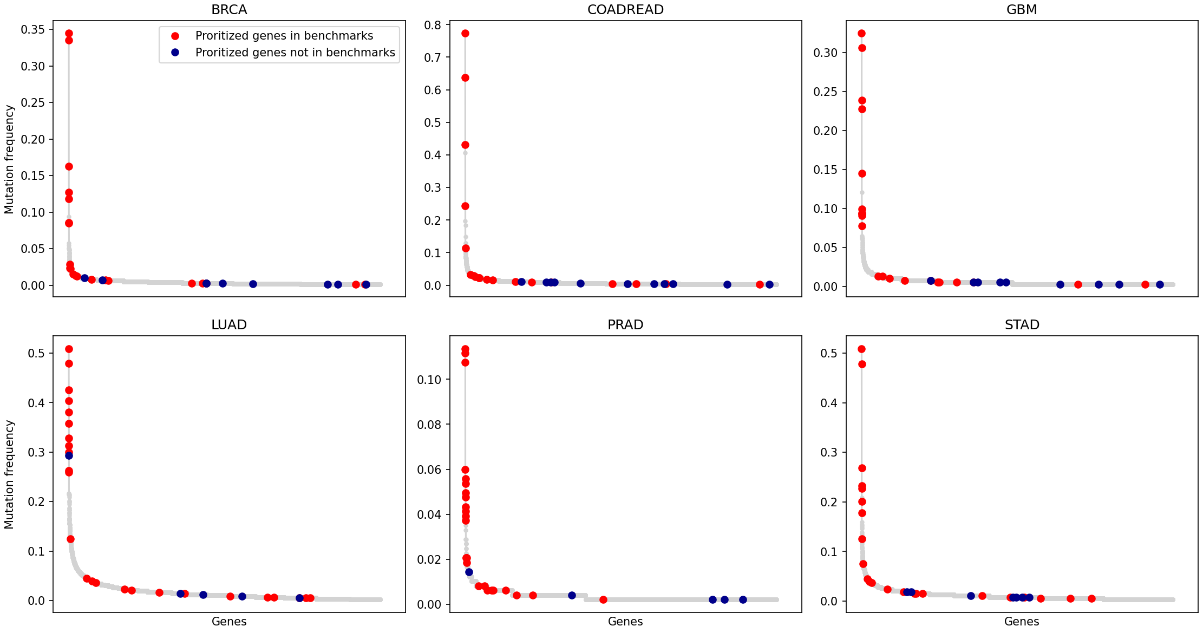


Figure 31 – Low-frequency mutated genes

4.5.2.6 Evidencing the potential of combination of weighted mutation and asymmetric spreading strength

This experiment shows the potential of the combination of the use of weighted mutation and asymmetric spreading strength. For this, A comparison between DiSCaGe and an alternative version was performed. Such an alternative version uses simple mutation frequency for mutation score and value one for the weights of edges on GSSN. As seen in Figure 29 and 30, DiSCaGe outperforms the alternative version on most of cancer types. For COADREAD the results are similar and for LUAD the alternative version had better results.

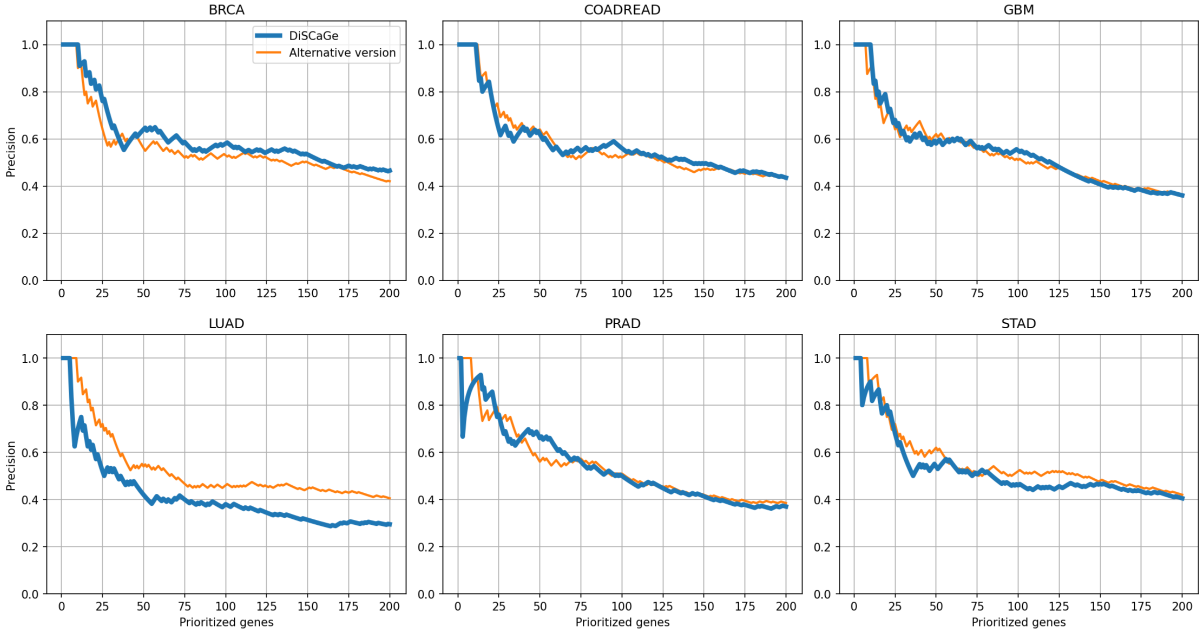


Figure 32 – Precision: DiSCaGe x Alternative version

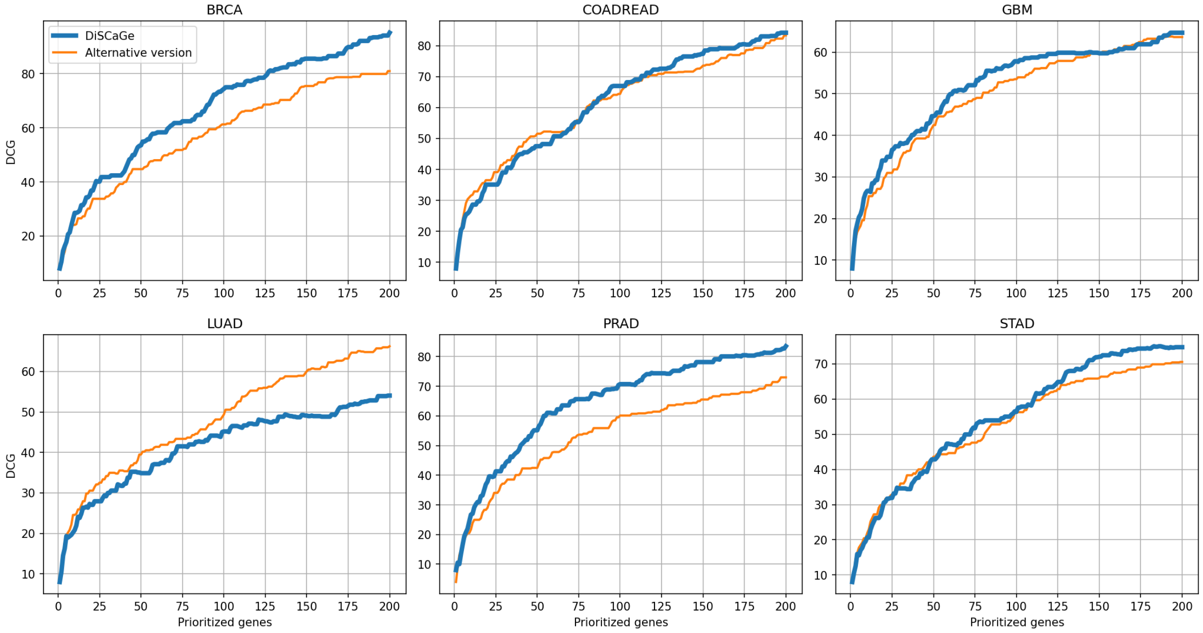


Figure 33 – DCG: DiSCaGe x Alternative version

4.5.2.7 Top 50 genes on benchmarks

In this experiment, the presence of prioritized genes on the benchmarks is illustrated. The 50 prioritized genes for each type of cancer were selected to be presented in a colored matrix, as seen in Figure 34. Each driver gene benchmark is presented in a row with a specific color, and genes discovered by DiSCaGe are presented in the columns, for example, NCG is the first row and has color green. If a gene is presented in the benchmark, the matrix cell is colored, otherwise the color is white. It can be noticed how the genes are arranged on the benchmarks and show that results are consistent. The top 50 genes were chosen due to provide a better visualization.

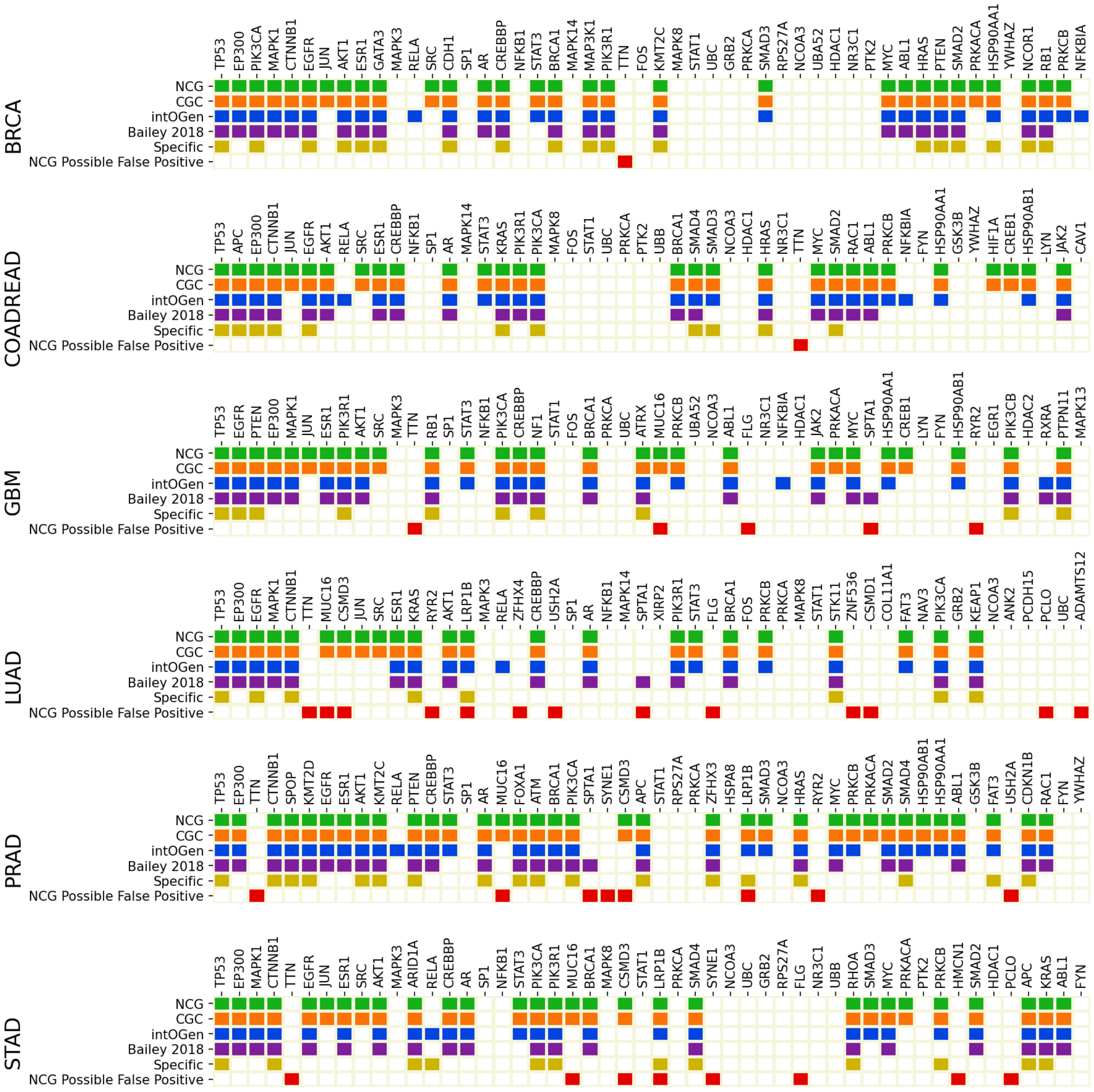


Figure 34 – How top 50 genes appear in benchmarks

4.5.2.8 Automated literature-based analysis

As shown in Figures 31 and 34, some genes are not in any the driver gene benchmarks. Their prioritization suggests they can potentially be novel cancer genes. Towards a secondary study on those genes, an automated literature review was performed using CancerMine (LEVER *et al.*, 2019). Figure 35 displays, for each type of cancer and for the top 50 genes, the ones that are not in the driver benchmarks, as shown in Figure 34, and their respective number of citations found by CancerMine[13].

[13] Query performed on January 2021

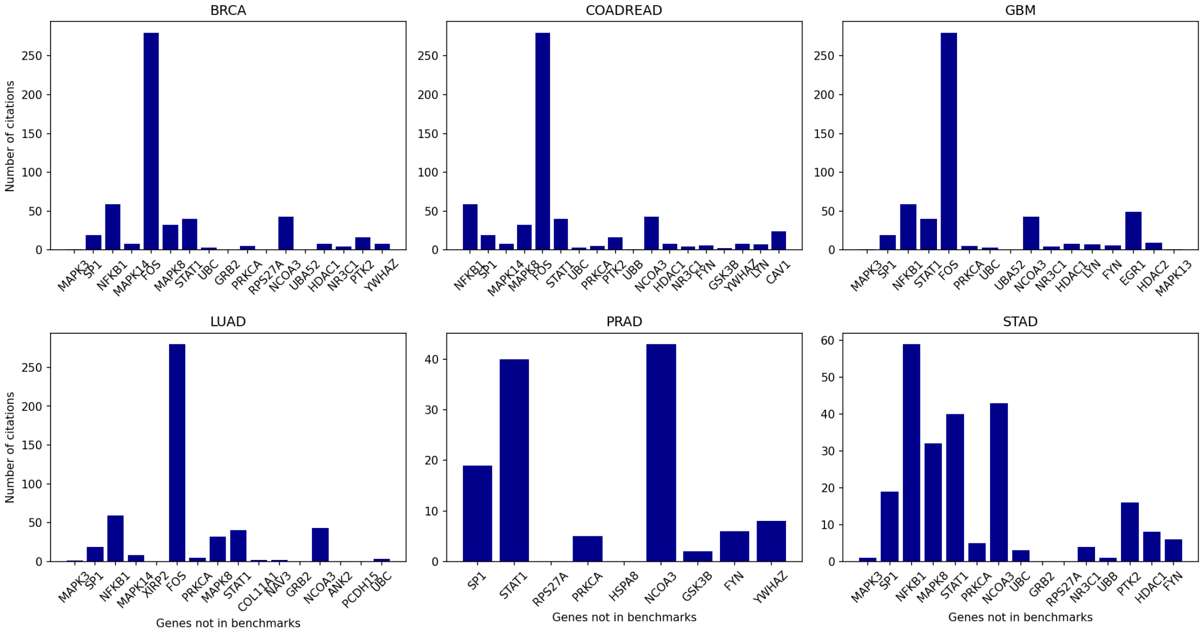


Figure 35 – Number of citations reported by CancerMine

Most genes were cited as cancer genes at least once in the research papers, which suggests even prioritized genes that are not in driver benchmarks can be related to cancer. The remaining of non-classified genes should be further evaluated and suggested as possible novel genes for their respective cancer types. Although it is a secondary study, this experiment allows comparing the relationship of genes known to be drivers with genes not yet known.

4.5.2.9 Comparison with related methods

In this section DiSCaGe is compared with related methods. The comparative study was performed in two perspectives: 1) Quantitative comparison: DiSCaGe is compared with related methods using precision and DCG. Furthermore, the potential of finding low-frequency genes and the ability to suggest possible novel cancer genes was also compared; and 2) Qualitative comparison: the main differences and novelties of DiSCaGe were pointed out in comparison to related methods.

Quantitative comparison

In this experiments DiSCaGe method is compared with three related methods: MutSigCV (LAWRENCE *et al.*, 2013), MUFFINN (CHO *et al.*, 2016), and nCOP (HRISTOV; SINGH, 2017). Such selected methods are chosen based on the classification of methods presented in Section 3.3, which returns a ranking list of prioritized genes and use mutation data. To avoid possible running influence, the methods were executed using standard parameters and configurations, which are described as follows:

MutSigCV: it was run using only the MAF file[14]. A difference on the preprocessing routine was assigned to MAF files for MutSigCV, in which `Silent` mutation were kept on MAF, because it is a mandatory data for the method, that uses it to extract the background mutation rate.

MUFFINN: it was run using the number of mutated patients for each gene as the gene mutation score (`mutation occurrence data`)[15]. MUFFINN has four variations, that is a combination of the approach (`DNmax` or `DNsum`) with the gene network (`STRING` or `HumanNet`). For this experiment, `DNmax + HumanNet` was selected, because it presented the best results.

nCOP: it was run using its preprocessed HPRD gene network[16] with no weight for each node in the network. The optimal value for alpha was obtained for the method itself.

Precision

Figure 36 displays precision for all methods. The results show variations according to the type of cancer and the top *N* value. For example, considering precision, for BRCA, nCOP is better for *N* nearly from 20 to 50. In general, DiSCaGe outperformed all methods for most values of *N* for BRCA, COADREAD, GBM, PRAD and STAD, which is evidenced in Figure 37. Such a figure shows boxplots of precision presented in the curves of Figure 36. It can be noticed the median for the DiSCaGe precision is better for five methods, except for LUAD, in which MUFFINN is slightly better.

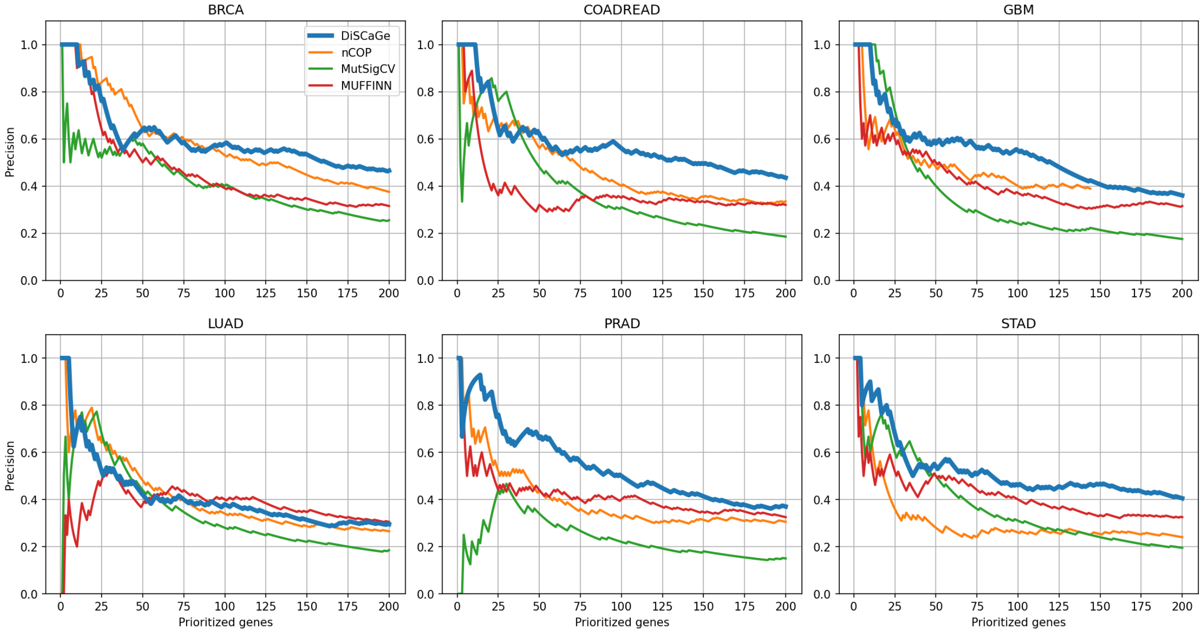


Figure 36 – Precision: related methods

[14] https://software.broadinstitute.org/cancer/cga/mutsig_run
[15] https://www.inetbio.org/muffinn/download.php
[16] https://github.com/Singh-Lab/nCOP

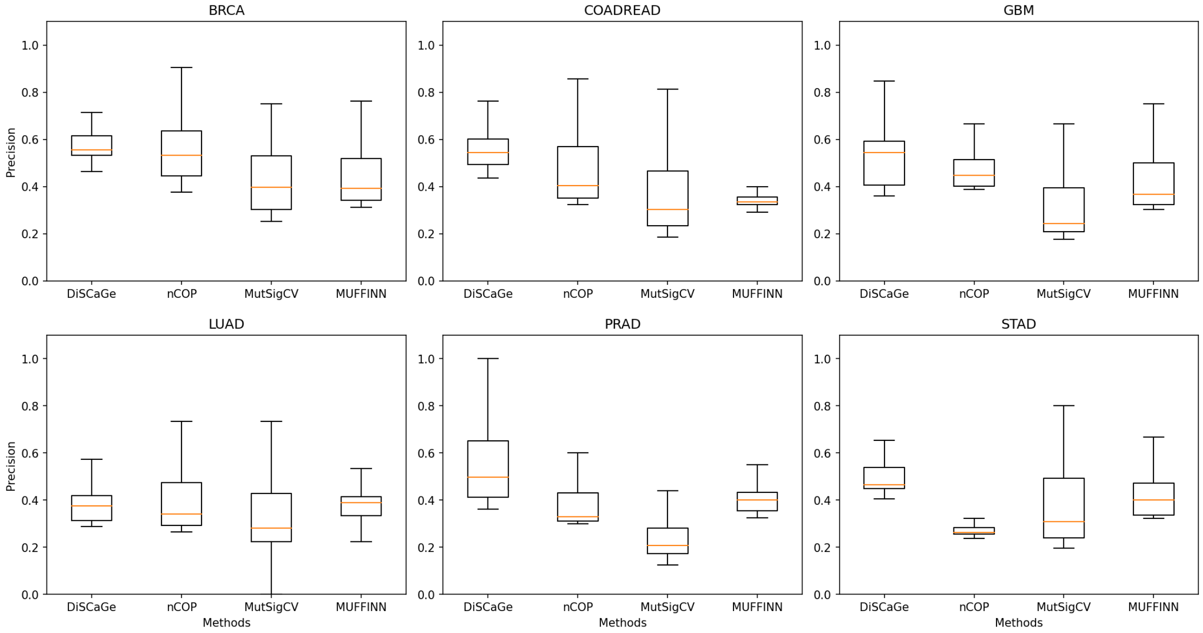


Figure 37 – Precision boxplots: related methods

DCG

Figure 38 displays DCG for all methods, which showed a significant variation in the results. Although DiSCaGe outperformed all methods for PRAD, this performance was neither dominant, nor explicit for the other types of cancer. nCOP is clearly better for BRCA and LUAD, and MutSigCV yielded promising results for STAD and GBM up to *N* nearly 50. DiSCaGe outperformed all methods for COADREAD, GBM, and STAD for *N* larger than nearly 50. The boxplots of DCG, presented in Figure 39 show that the DCG median of DiSCaGe is better for COADREAD, GBM, PRAD and STAD. For BRCA, nCOP is the only method better than DiSCaGe, and for LUAD, nCOP and MutSigCV outperfom it.

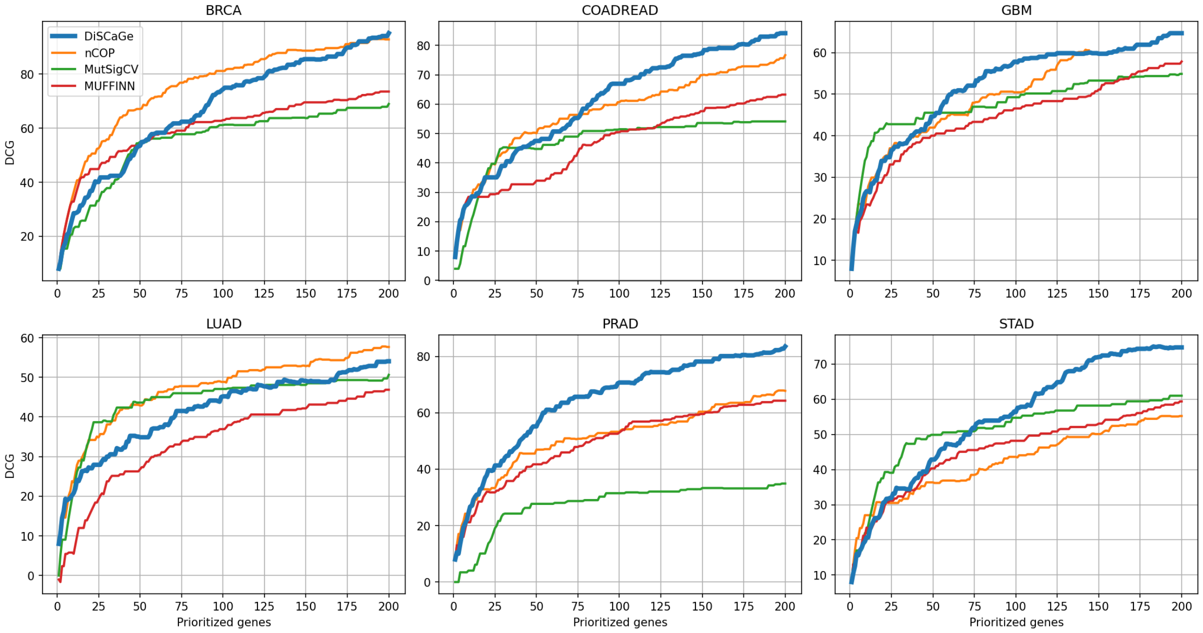


Figure 38 – DCG: related methods

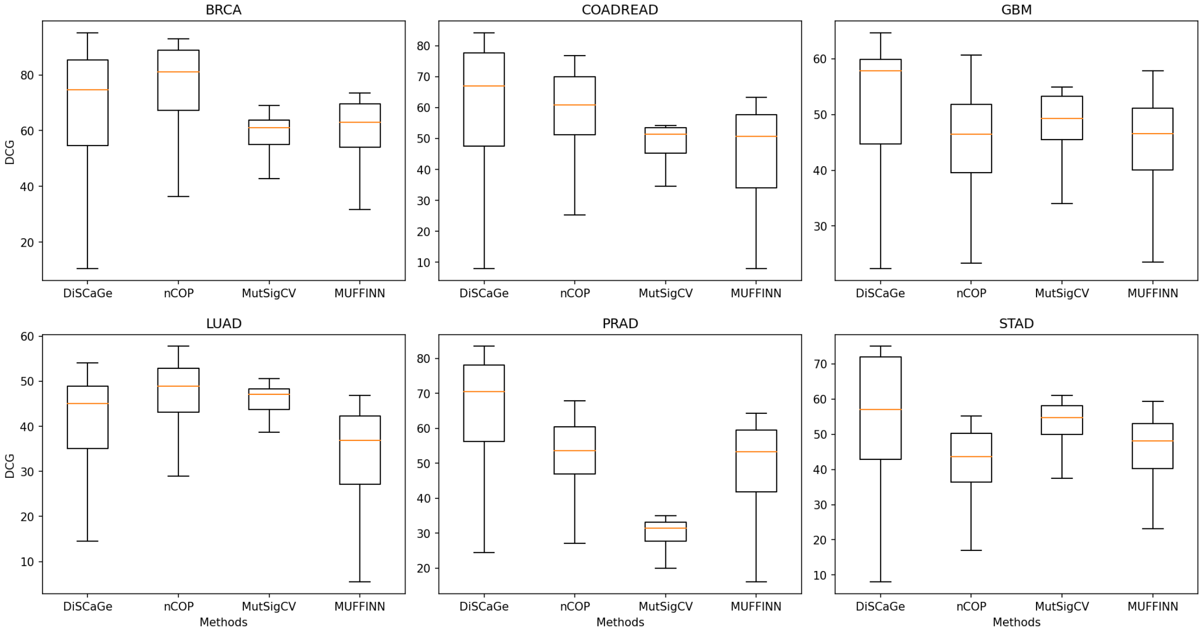


Figure 39 – DCG boxplots: related methods

Low-frequency mutated genes

Section 4.5.2.5 shows that DiSCaGe is able to suggest possible cancer genes with very low mutation frequency. In order to compare this ability with previous methods, Figure 40 shows boxplots of mutation frequency of top 200 genes prioritized for DiSCaGe and related methods. The mutation frequency median is lower for COADREAD and similar to MUFFINN for BRCA and STAD. It is important to mention that this analysis is quite relative, because the fact of finding more genes with low frequency is good or bad is dependable on the objective of analysis.

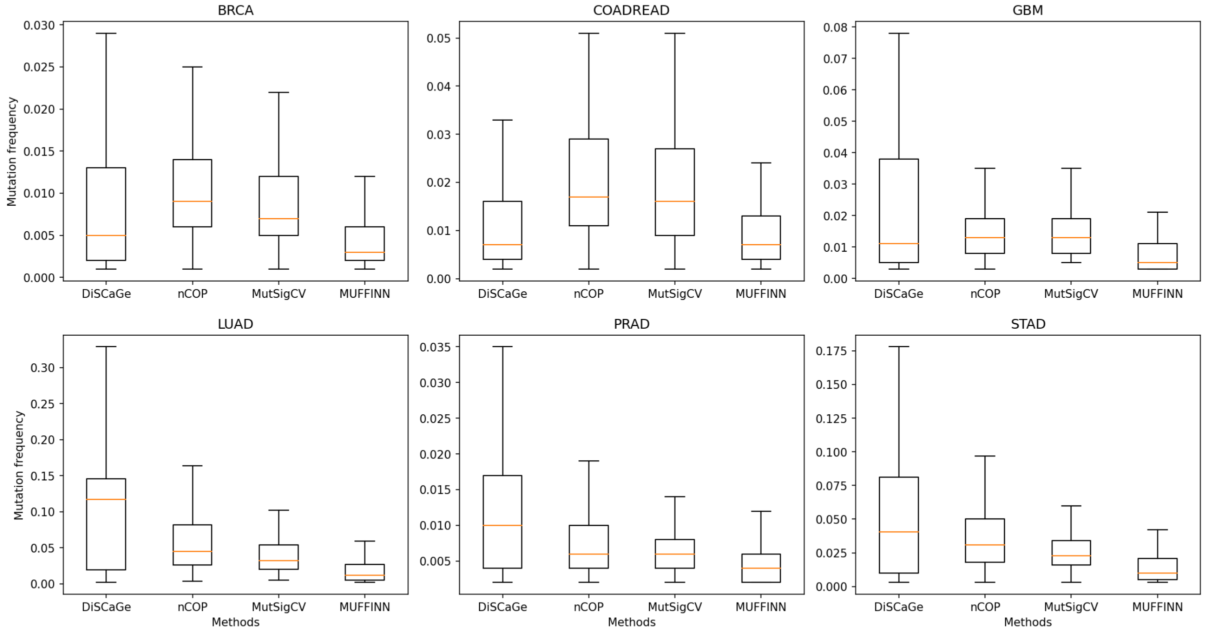


Figure 40 – Boxplots of mutation frequencies of prioritized genes

Potential on discovering possible novel cancer genes

Figure 41 shows stacked bar plots of the frequency of top 200 genes prioritized by DiSCaGe and related methods appears in driver benchmarks (green bar), false-positive benchmark (red bar), in CancerMine cited at least one time (blue bar), and not appearance in any of them (gray bar). It can be noticed that MUFFINN prioritized most known cancer-related genes, including evidences from CancerMine, but DiSCaGe outperforms MUFFINN on finding genes on driver benchmarks. DiSCaGe finds a significant number of false-positives for LUAD, PRAD and STAD, but DiSCaGe was the better of finding genes in driver benchmarks in such cancer types.

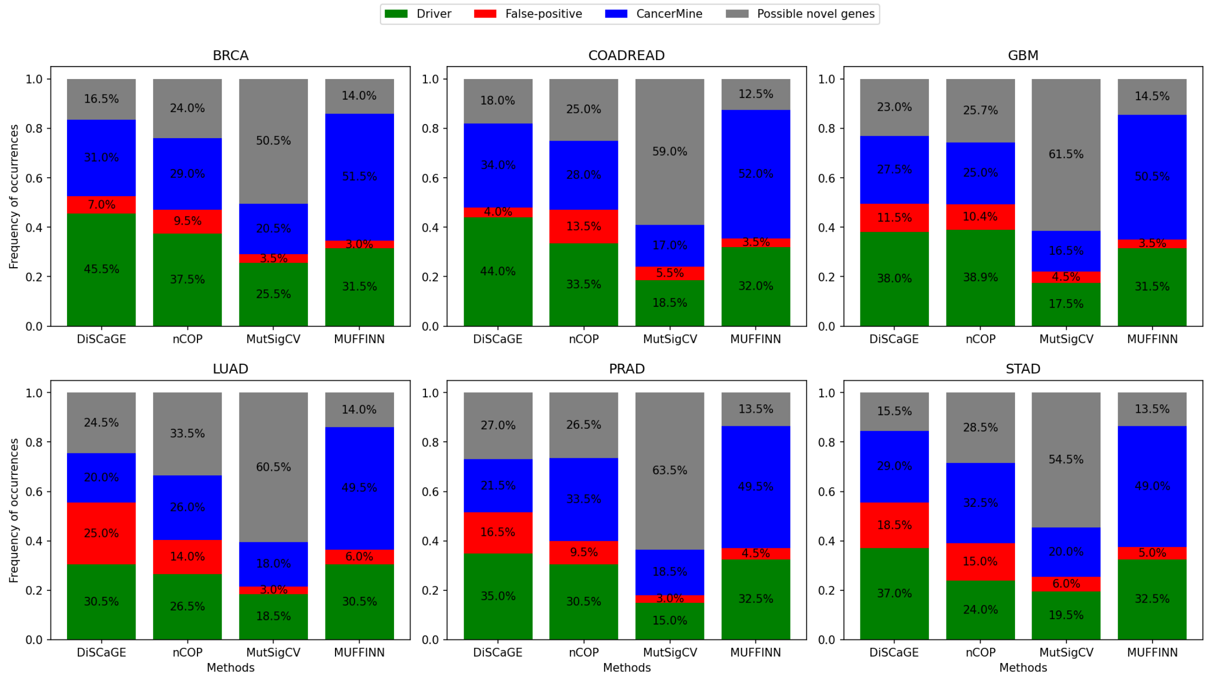


Figure 41 – Frequency of occurrences of prioritized genes on benchmarks and CancerMine

Qualitative comparison

DiSCaGe presents several differences in comparison with related methods. In this chapter a qualitative comparison and discussion is performed with some related methods presented in Section 3.3.

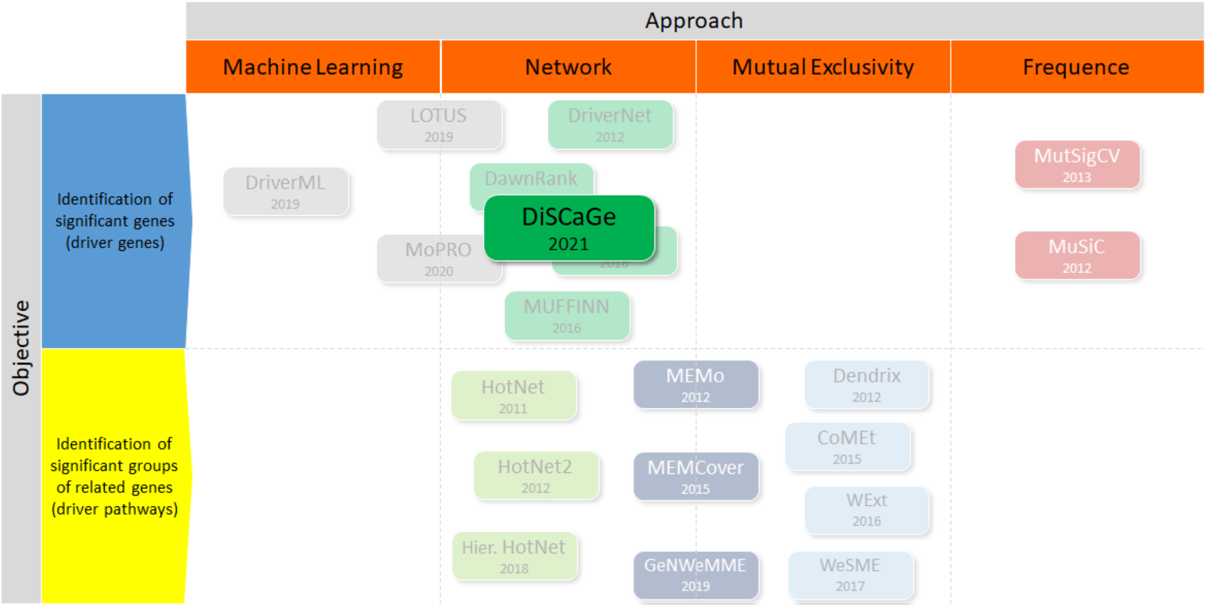
Most of methods use the simple frequency or a binary mutation matrix to compute the mutation score, such as MUFFINN, nCOP, and MEMo. DiSCaGe employs a simple way to get weighted frequency, based on the definition of weights on the impact of each type o mutation. Such impact are user-centric, i.e., the final user can define the weights based on the objective of analysis. DriverML automatizes the definition of the impact of mutation types through a machine-learning approach, based on previous information.

Related to the use gene interaction networks, nCOP seeks to identify connected subnetworks that are significant altered across the patients, using theses finding to ranking the genes, while DiSCaGe dos not considers the subnetworks, only the gene neighborhood. In this way, MUFFINN is closely related to DiSCaGe, but the neighborhood influence is obtained through the maximum of the direct neighbor mutation score or the sum of the direct neighbor, divided by its degree. Also, DiSCaGe differs on the using of the union network to infer the spread strength from a mutated gene, thus considering it on the neighbor influence.

Gene interaction networks are the only previous knowledge information that DiSCaGe employs for the cancer genes prioritization. None previous information about the gene cancer significance is used. The machine-learning methods uses this known cancer genes in order to discover possible novel ones. MutSiGCV and MuSiC are frequency-based methods that estimates the background mutation, while other methods, such as Dendrix and WExT uses only mutation data do infer group of mutual exclusive genes. DiSCaGe do not employ such features. DriverNet and DawnRank use gene expression data on their algorithms.

According to the map of Figure 9, DiSCaGe can be classified as a method of identification of significant genes for cancer that use network-based approach to discover such genes. The map with DiSCaGe is displayed in Figure 42.

Figure 42 – Classification of DiSCaGe on the methods map.



Source: Elaborated by the author.

4.5.2.10 Influence of hyperparameter

The hyperparameter of DiSCaGe is the definition of weights of each type of mutation. Although this hyperparameter is clear and can be defined logically by the expert user, in this experiment an analysis of the influence of the hyperparameter values was performed. For this, the mutation types were divided in three groups, that were defined based on the rationale that mutation of same group have similar functional impact:

Group 1: `Nonsense_Mutation`, `Frame_Shift_Ins`, and `Frame_Shift_Del`.

Group 2: `Missense_Mutation`, `Splice_Site`, `In_Frame_Ins`, `In_Frame_Del`, `5'UTR`, and `Nonstop_Mutation`.

Group 3: `3'UTR` and `Translation_Start_Site`.

In the analysis, a fixed value was defined for two groups, while the other group was variate. The fixed values follow the presented on Table 7, i.e., 1.0, 0.4 and 0.2 for the Groups 1, 2 and 3, respectively. The variation of values was on the range from 0.2 up 1.0.

Figures 43 and 44 show the precision and DCG, respectively, for the variation of the weights of mutation of Group 1. The boxplots represents all values of precision and DCG for the top 200 prioritized genes for each type of cancer. In a general way, it can be noticed that the DiSCaGe performance increases when the weights values are higher. For COADREAD and PRAD this evidence is not so clear after value 0.6.

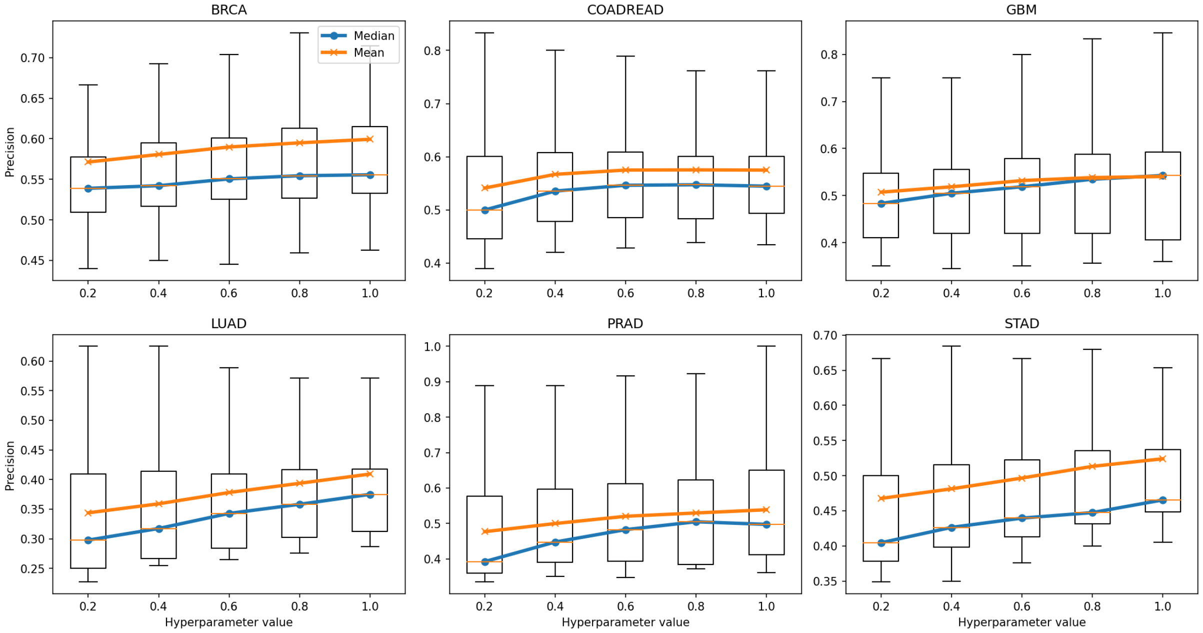


Figure 43 – Precision considering the variation on the weights of mutations of Group 1

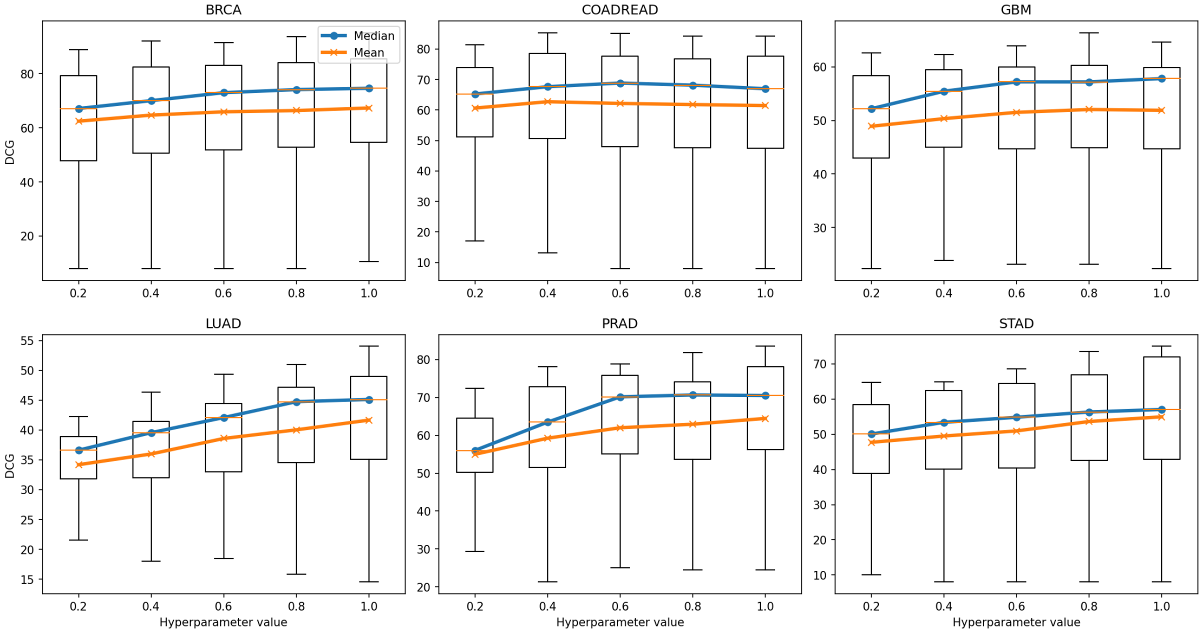


Figure 44 – DCG considering the variation on the weights of mutations of Group 1

Related to the weights for the mutation of Group 2, Figures 45 and 46 show the precision and DCG, respectively. It can be noticed that there is no a clear tendency for all cancer types.

For example, for LUAD and STAD, the value 0.2 have a better performance. These lack of a clear tendency can be related to the number of missense mutations, which are the majority of the mutations. LUAD is a hypermutated cancer type, and has large number of missense that can be with no impact, then a lower weight for missense can improve results for LUAD.

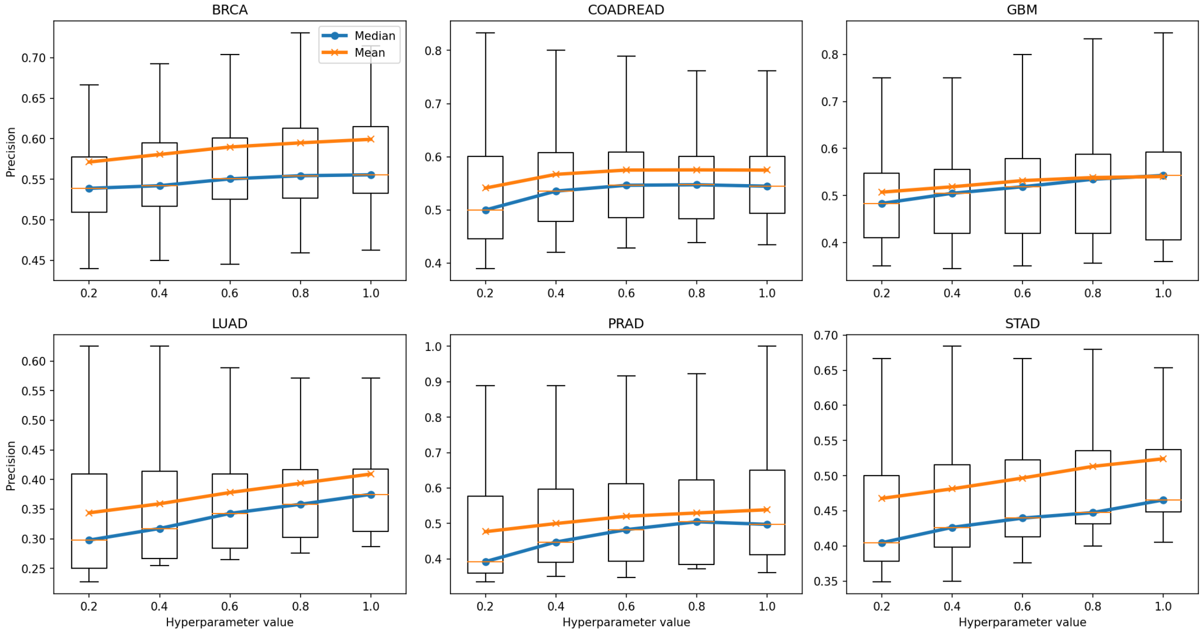


Figure 45 – Precision considering the variation on the weights of mutations of Group 2

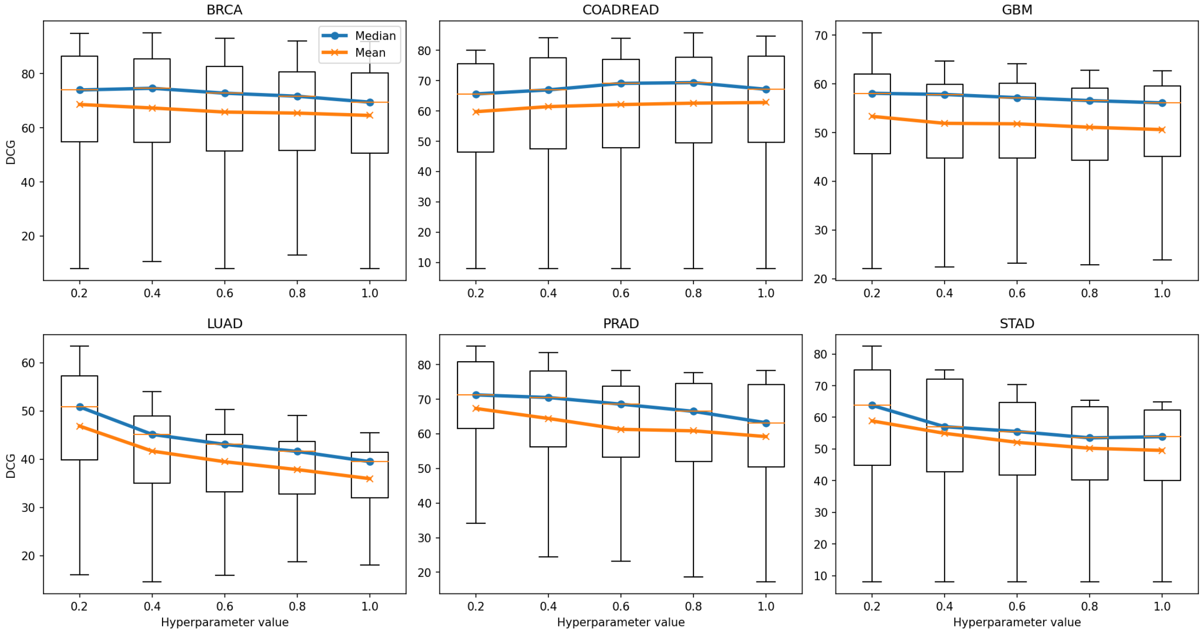


Figure 46 – DCG considering the variation on the weights of mutations of Group 2

For the mutation types o Group 3, the performance of DiSCaGe have no significant difference, as shown in Figures 47 and 48. The cause of this behavior can be due to the number of mutations for the types of Group 3 in small, then the impact is less significant.

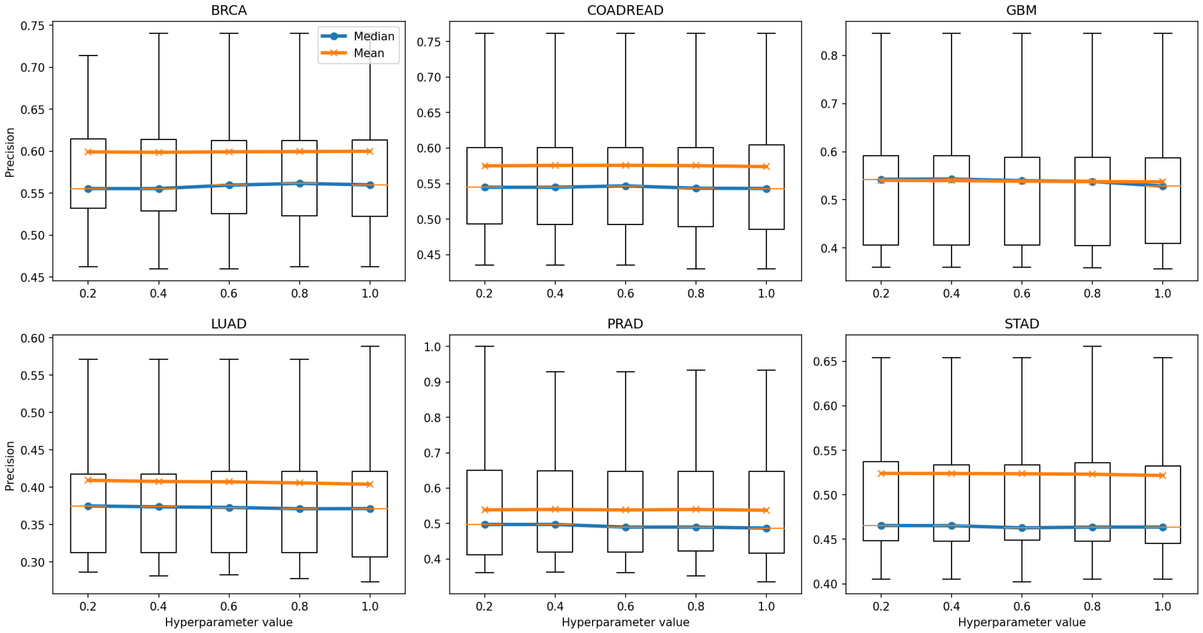


Figure 47 – Precision considering the variation on the weights of mutations of Group 3

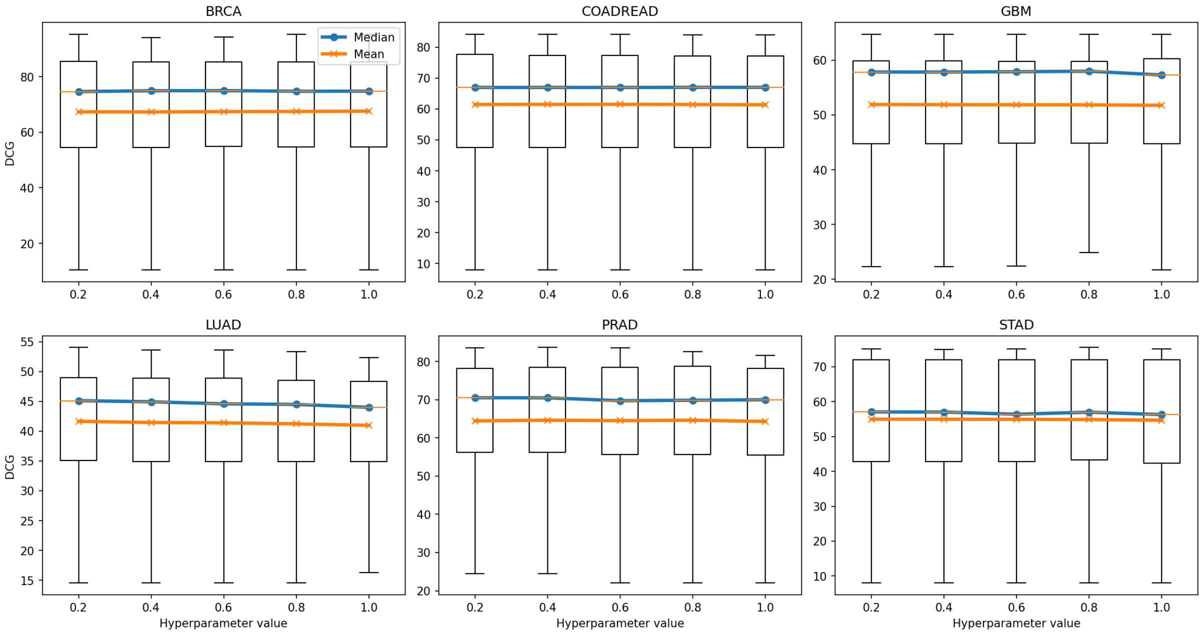


Figure 48 – DCG considering the variation on the weights of mutations of Group 3

4.5.2.11 Impact of the union of networks

An important aspect of DiSCaGe is the possibility of using many gene interaction networks. They are combined using union of edges across networks. Figures 49 and 50 shows the impact when combining networks. The first and second boxplot are the results for DiSCaGe using eHPRD and eReactome, respectively. After that, a union of them was performed, which is displayed in the third boxplot.

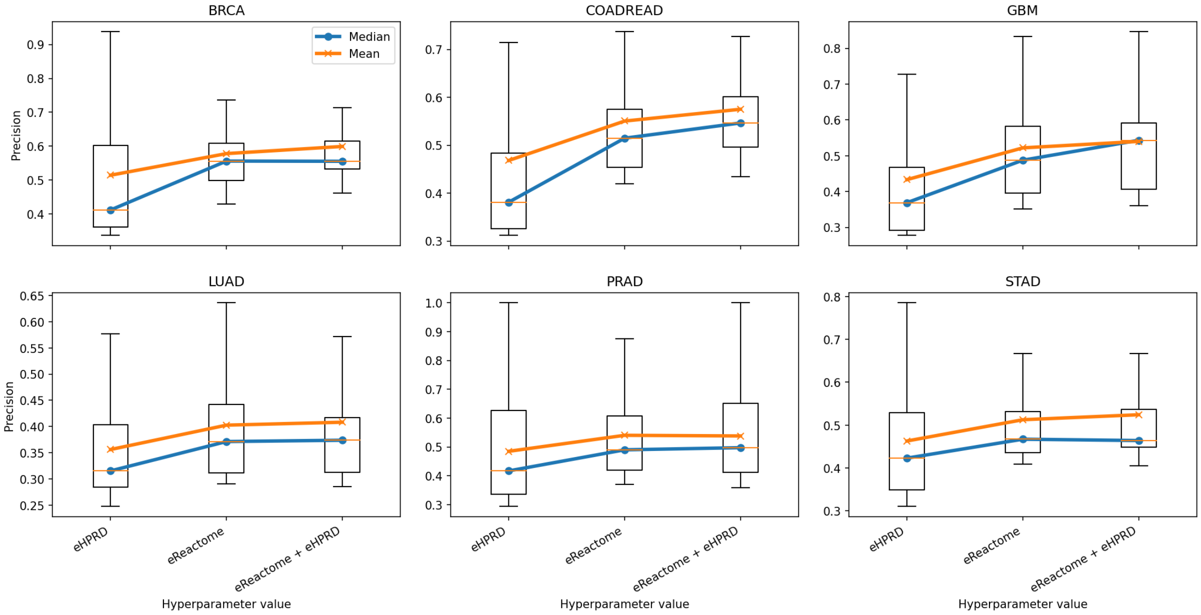


Figure 49 – Precision considering the union of gene interaction networks

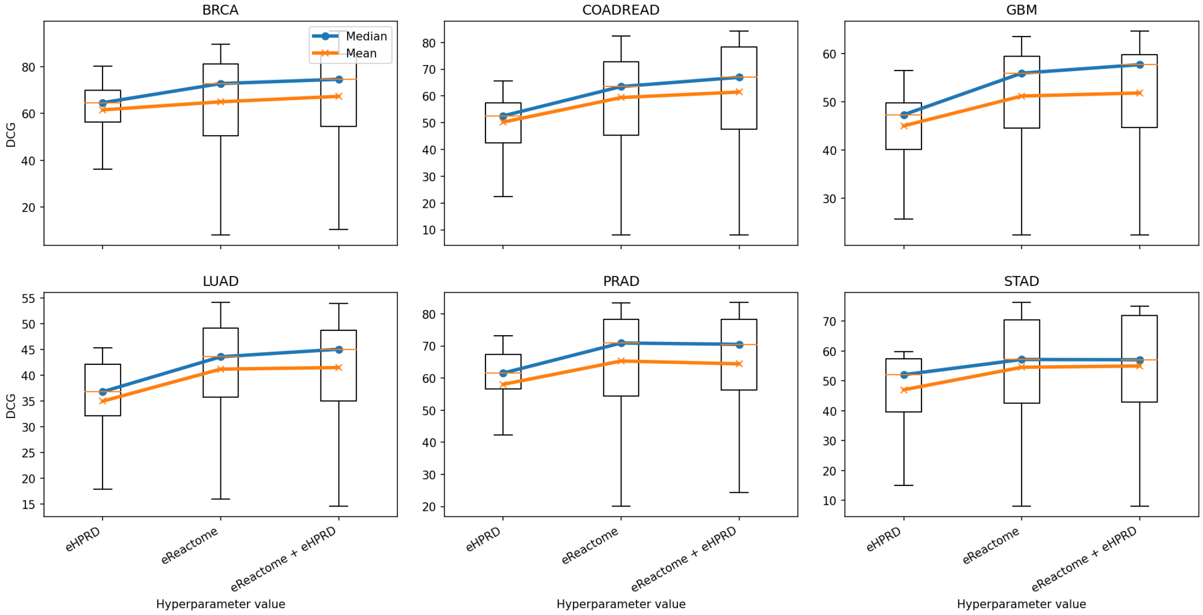


Figure 50 – DCG considering the union of gene interaction networks

It can be noticed that the performance of DiSCaGe have a slightly impacted with the union of the networks. Although that procedure is a interesting way to avoid a bias on the chose of a single network, the results shows the results using eReactome is similar with the union. It can occur because Reactome is a extensively studied and developed network, with a high confidence on the existing gene interactions. Maybe, with the discovery of new interactions in other networks, the union operation could impact the results directly.

4.6 Final Considerations

This chapter described DiSCaGe, a computational method for the discovery of significant genes for cancer, which is based on ideas of two papers (CUTIGI; EVANGELISTA; SIMAO, 2020b; CUTIGI *et al.*, 2021).

First, the cancer mutation data is collected, preprocessed and analyzed, followed by the selection of gene interaction networks, their enrichment and characterization study. Next, DisCaGe is described as a computational method for the discovery of significant genes for cancer which takes into account weighted mutations in genes and the way they can affect network neighborhood through an asymmetric spreading strength measure. The method presentation is accompanied by a running example, and algorithms description. An experimental evaluation was conducted and presented, with a set of known cancer genes benchmarks and an automated literature review of genes prioritized by the proposed method. The method was able to 1) prioritize genes known to be related to cancer, 2) prioritize genes related to cancer with low mutation frequency, 3) suggest genes that are not in benchmarks, but are cited in research papers as cancer-related ones, and 4) suggest possible novel cancer genes.

In the next chapter a machine learning approach, called DFDriver, is proposed to identify possible false-positives significant genes for cancer. Such an approach could be used on the output of DiSCaGE, to produce a more reliable cancer gene list.

CHAPTER 5

COMPUTATIONAL APPROACH FOR THE DETECTION OF FALSE-POSITIVE SIGNIFICANT GENES FOR CANCER

5.1 Initial Considerations

An increasing interest in Cancer Genomics research emerged from the advent and widespread use of next-generation sequencing technologies, which have generated a large amount of digital biological data. However, not all of this information in fact contributes to cancer studies. For instance, false-positive-driver genes may contain characteristics of cancer genes but are not actually relevant to the cancer initiation and progression. Including this type of genes in cancer studies may lead to identifying unrealistic trends in the data and mislead biomedical decisions. This chapter reports on an investigation of the following research question: ***RQ2: Can false-positive genes be detected in a set of significant candidates for cancer with the use of mutation and gene network data?***.

Towards answering the question, a machine learning-based approach, named DFDriver (**D**etecting **F**alse **Driver**), is proposed and described. It aims to induce predictive models to classify supposedly driver genes as real drivers or false-positive drivers based on both mutation data and gene network interactions. Figure 51 shows a summary of the approach established for the research. In Step 1, cancer mutation data, gene interaction networks, and gene labels are selected from reliable and widely used sources. In Step 2, data are preprocessed, and features are extracted towards composing a labeled data set created from the combination of somatic mutation data of 33 types of cancer and centrality measures of a union of four enriched gene interaction networks. Finally, in Step 3, a hyperparameters tuning is performed so that optimized models can be induced and evaluated through stratified k-fold cross-validation, according to a set of evaluation metrics. Experimental results show the combination of mutation data and gene interaction data can improve the models' prediction potential.

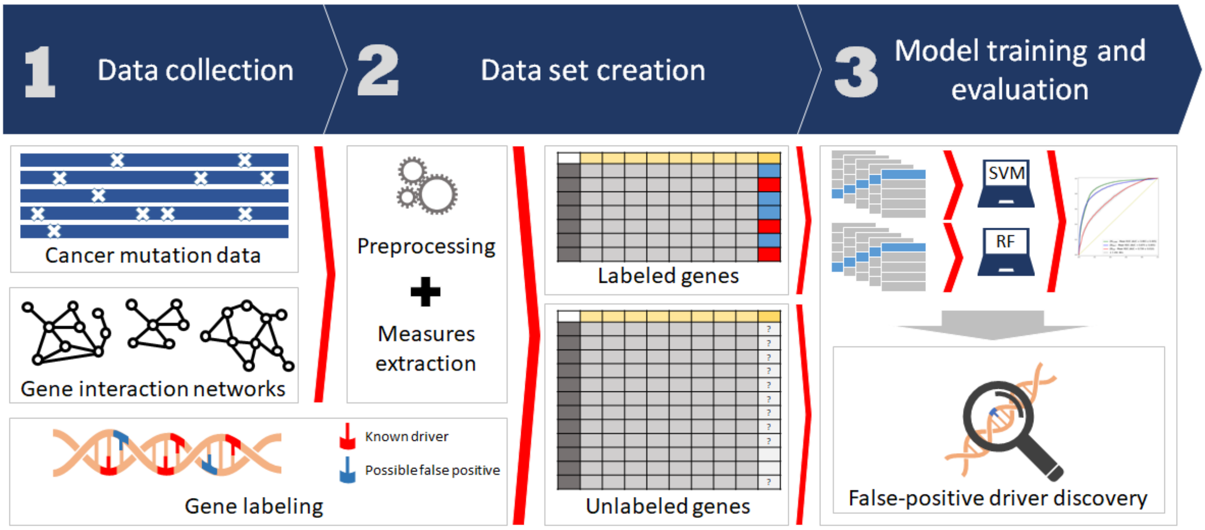


Figure 51 – An overview of the approach.

The chapter is based in a paper published at the *Brazilian Symposium on Bioinformatics (BSB 2020)* (CUTIGI *et al.*, 2020), as follows:

- CUTIGI, J. F.; EVANGELISTA, R. F.; RAMOS, R. H.; FERREIRA, C. d. O. L.; EVANGELISTA, A. F.; CARVALHO, A. C. de; SIMAO, A. Combining mutation and gene network data in a machine learning approach for false-positive cancer driver gene discovery. In: SPRINGER. **Brazilian Symposium on Bioinformatics**. [S.l.], 2020. p. 81–92.

The chapter is organized as follows: Section 5.2 describes the collection of mutation data, gene interaction network, and gene labels; Section 5.3 reports the data set preparation for the models training; Section 5.4 introduces the machine learning-based approach, with the selection of algorithms, training process, and hyperparameter selection; Section 5.5 is devoted to an evaluation of the models by classical classification metrics and discussion and summarization of the results.

5.2 Data collection

An essential step in a machine learning process is properly data collection, which involves data acquisition, data preprocessing, and data labeling. Such activity is crucial for the obtaining of a reliable data set for inducing useful models. The next sections address the collection of three sources of information, necessary for the training data set preparation.

5.2.1 Cancer mutation data

Data sets of 33 types of cancer were selected according to a TCGA Pan-Cancer study (BAILEY *et al.*, 2018) and downloaded from cBioPortal[1] (CERAMI *et al.*, 2012) by a web API[2]. The collection contains mutation data comprehended of single nucleotide variants (SNVs), and insertions and deletions (InDels). The mutation data for each type of cancer are structured in an MAF format, as described in Section 4.2.

Each MAF file was subjected to a preprocessing routine similar to the process described in Section 4.2.1. The difference is only nine specific somatic variants were kept in MAF file, namely: `Frame_Shift_Del`, `Frame_Shift_Ins`, `In_Frame_Del`, `In_Frame_Ins`, `Missense_-Mutation`, `Nonsense_Mutation`, `Nonstop_Mutation`, `Splice_Site` and `Translation_-Start_Site`. Other variants (`3'UTR`, and `5'UTR`) were not available on MAF file obtained through the API. Hypermutated samples were also removed, according to the same process described in Section 4.2.1, i.e., the strategy defined by Tamborero *et al.* (2013).

All preprocessed MAFs were merged into a single MAF file, thus generating a MAF file with mutation data of 33 types of cancer. Table 12 shows some metrics of the consolidated mutation data, before and after the preprocessing routine, with the number of patients, genes, and mutations.

Table 12 – Mutation data before and after preprocessing routine.

	Non-preprocessed mutation data	Preprocessed mutation data
Patients	10429	9741
Genes	20072	19183
Mutations	2192073	1228102

5.2.2 Gene interaction network data

The enriched gene interaction networks described in Section 4.3 were selected to be used as a source of information of gene interactions. Similarly to Step 3 of DiSCaGe (see Section 16), a union operation was applied to the enriched networks, resulting in a single network. However, the resulting network is treated here as an unweighted network. Only the main component of the network was considered towards the extractions of all defined centrality measures. Table 13 shows some metrics of the full network and its main component, which is similar to the full network.

[1] <https://www.cbioportal.org/datasets>

[2] <https://www.cbioportal.org/api/swagger-ui.html>

Table 13 – Comparison of the full network and its main component.

Network	Nodes	Edges	Mean degree	Density	Components
Full network	15432	3388292	439.13	0.028457	61
Main component	15294	3388191	443.07	0.028972	1

5.2.3 Gene labels

The proposed machine learning approach aims to classify driver candidates as real or false-drivers, which requires genes labeled in such classes.

Set FD_{NCG} with 250 genes (described in Section 4.5.1.2) listed as possible false-positive drivers was used as a reference to label genes in the data set as false-drivers (FD) for the induction of predictive models by the supervised machine learning algorithms. For driver class D, the four sets, i.e., D_{NCG}, D_{CGC}, $D_{IntOGen}$, and D_{Bailey} (described in Section 4.5.1.1) of known driver genes were joined, resulting in a set of 951 genes listed as drivers. However, 65 genes in the set were also present in FD, therefore, they were removed from D. The remaining 886 genes were then used as a reference to label the drivers.

5.3 Data set preparation

The data set was properly structured for the training of supervised machine learning algorithms. The samples in the unlabeled data set are the genes, while the features are the measures extracted from the mutation and gene network data, as summarized below:

Mutation data set DS_{MUT}: nine features were extracted from the MAF file for each gene for the creation of a mutation data set DS_{MUT}. Such features comprehends the number of mutations of each specific somatic variant. Therefore, DS_{MUT} is composed of 19183 samples and nine features.

Gene network data set DS_{GN}: nine features were extracted for each gene (node) in network for the creation of a data set DS_{GN}. The features are centrality measures, presented as follows and described according to the characteristics of a central node (OLDHAM *et al.*, 2019):

Degree: a node connected to many other nodes, i.e, a node with a high number of edges.

Betweenness: a node that is part of many shortest paths linking all pairs of nodes in the network.

Closeness: a node with lower average shortest path length to other nodes in the network.

Eigenvector: a node connected to many other nodes and to other high-degree nodes.

Coreness: a node connected with many other nodes in a peripheral region in the network.

Clustering coefficient: a node with a high fraction of edges among the neighbors.

Average of neighbors' degree: a node connected with a high-degree node.

Leverage: a node with a higher degree than its neighbors.

Bridging: a node that is a key link between high-degree nodes.

Such measures consider distinct aspects of the network structure and topology to characterize the importance of a node, thus highlighting its central role (OLDHAM *et al.*, 2019). DS_{GN} is composed of 15294 samples and nine features.

Combined data set DS_{COMB}: features from DS_{MUT} and DS_{GN} were merged towards creating a combined data set. Some of the genes were not contained in the data sets, therefore, only their intersection was taken. The merging resulted in a data set DS_{COMB}, composed of 13988 samples and 18 features.

Finally, the genes in data set DS_{COMB} in the FD and D lists were extracted and properly labeled. Considering only the labeled samples, the resulting DS_{COMB_L} is composed of 1033 samples, 18 numeric features, and one class label (819 drivers and 214 false-drivers). The same process was applied to DS_{MUT} and DS_{GN}, thus resulting in DS_{MUT_L} and DS_{GN_L}.

5.4 Machine learning approach

In DFDriver approach, supervised machine learning algorithms were trained with data set DS_{COMB_L} towards inducing predictive models to classify genes as drivers or false-drivers. They were also applied to DS_{MUT_L} and DS_{GN_L} for comparing models induced with combined data (mutation + gene network) to those induced with a single source of information (mutation or gene network). Scikit-learn (PEDREGOSA *et al.*, 2011), a Python module for machine learning, was used in all processes described in this section. The following questions and answers summarize DFDriver and its characteristics and goals:

1. **What is the biological problem that the method seeks to solve?**

   The detection of possible false positives drivers among a set of genes candidates to be driver.

2. **Why is this method necessary?**

   Improve the the reliability of driver genes candidates, because including false positive drivers in cancer studies may lead to identifying unrealistic trends and mislead biomedical decisions.

3. **What are the input data and hyperparameters to run the method?**

Data: Mutation data (MAF file with SNVs and InDels); and centrality measures of genes in interaction networks.

Hyperparameters: machine learning algorithms hyperparameters.

4. **How is the problem computationally formulated?**

   An extraction of useful information from both mutation data and gene network interactions is performed and used as features for the models. Random Forest and Support Vector Machine models were induced using the selected data.

5.4.1 Predictive models

The following two machine learning algorithms were selected to induce the predictive models: 1) Support Vector Machine (SVM), a statistical learning algorithm that seeks the identification of a hyperplane that can separate the classes of a problem, and 2) Random Forest, an ensemble learning algorithm that generates several random decision trees, taking the combination of their outputs as the classification.

SVM and Random Forest were selected for this study because they enable significant flexibility in the induced models' architecture through the adjustment of their hyperparameters. This selection was performed considering that the structure required to represent the problem and perform the classification was not known.

The models were induced by a stratified 5-fold cross-validation scheme with re-sampling applied to every training portion of folds towards avoiding overfitting and addressing class-imbalance. The re-sampling was performed through a combination of over and under-sampling For over and under-sampling, SMOTE (Synthetic Minority Over-sampling Technique) (CHAWLA *et al.*, 2002) and ENN (Edited Nearest Neighbors Undersampling) (WILSON, 1972) were used, respectively, which can be called SMOTEENN (BATISTA; PRATI; MONARD, 2004). SMOTE generates synthetic examples from the minority class by placing them among samples that are close in the feature space. In the other hand, ENN removes examples from the majority class by removing them if a certain number of neighbors of minority class is higher.

Both re-sampling and folds split procedures were repeated, taking different random states in each new training process. Also, a z-score standardization was applied to all features of data sets.

5.4.2 Hyperparameter selection

Different hyperparameter sets were assessed through the training of multiple models in a grid-search process, using procedures described before towards addressing overfitting and class-imbalance problems. The process was repeated ten times for accounting for a possible influence of randomness on both models training and re-sampling. The grid-search process induced models for all combinations of hyperparameters into a defined range, thus evaluating each model according to a defined metric. An optimal hyperparameter set was obtained in the training. Accuracy was used as an evaluation metric for comparing the induced models with different configurations.

The following hyperparameters were considered for SVM, and Table 14 shows the range of each hyperparameter conducted in the grid-search process.

- *C*: a regularization parameter that controls the error rate. Higher values imply higher tolerance for the misclassified samples.
- *gamma*: a parameter for nonlinear hyperplanes that controls the curvature on boundaries of classes separation. Higher values imply more curvature on boundaries. Value *auto* is equal to $1/n\_features$, where `n_features` are all features of the data set, and value *scale* is equal to $q/n\_features * X.var()$, where $X.var()$ is the variance of the training data set.
- *kernel*: a type of hyperplane used in classes separation. Value *linear* is a liner hyperplane, while *rbf* and *sigmoid* are nonlinear.

Table 14 – Grid-search process for SVM.

Hyperparameter	Values
C	1, 2, 4, 6, 8, 10
gamma	auto, scale, 1, 2, 3, 4, 6, 8, 10
kernel	linear, rbf, sigmoid

The following hyperparameters were considered for Random Forest, and Table 15 shows the range of each hyperparameter conducted in the grid-search process.

- *n_estimators*: number of decision trees in the forest.
- *max_depth*: max number of levels of each decision tree. Value *None* considers there is no limit for expanding the tree (for example, an expansion can occur until all leaves are pure according to the criterion).
- *max_features*: max number of features considered by each tree. Value *None* considers all features (`n_features`) of the data set, while *sqrt* and *log2* consider a random set of features of size equal to the square root of `n_features` and logarithm equal to base 2 of `n_features`, respectively.
- *criterion*: measure of the quality of each split on the trees. Values *gini* and *entropy* can be used to decide the split in the decision tree considering impurity and information gain, respectively.

Table 15 – Grid-search process for Random Forest.

Hyperparameter	Values
n_estimators	20, 50, 100, 150, 200
max_depth	None, 3, 4, 5, 6, 8, 10, 12, 14, 16, 18, 20
max_features	None, autor, sqrt, log2
criterion:	gini, entropy

The optimal hyperparameters obtained from the grid-searches performed for the three labeled data sets, with different features, are provided in Tables 16 and 17 for SVM and Random Forest, respectively. The data sets are referred to as follows: DS_{COMB_L}: labeled data set with features from mutation data and gene network data; DS_{MUT_L}: labeled data set with features only from mutation data; and DS_{GN_L}: labeled data set with features only from gene network data.

Table 16 – Optimal hyperparameters for data sets containing different features in SVM. (Note: gamma hyperparameter is not applicable for kernel `linear`)

	DS_{COMB_L}	DS_{MUT_L}	DS_{GN_L}
Kernel	linear	linear	rbf
C	4	1	1
gamma	NA	NA	10

Table 17 – Optimal hyperparameters for data sets containing different features in Random Forest.

	DS_{COMB_L}	DS_{MUT_L}	DS_{GN_L}
n_estimators	200	200	200
max_depth	16	16	16
max_features	auto	auto	auto
criterion	entropy	entropy	entropy

5.5 Experimental study

An experimental study was conduct to evaluate the potential of DFDriver on the detection of driver candidates into false or real drivers. For this, the induced models were evaluated through classical machine learning classification metrics. Also, genes prioritized by DisCaGe were submitted to the classification using DFDriver.

5.5.1 Evaluation criteria

A set of metrics was selected for the assessment of the trained models' performances. They are based on the classification of a driver gene as true positive (TP), true negative (TN), false positive (FP), or false negative (FN). Such metrics, described in what follows, are important because they can help the identification of possible systematic trends in the miss-classifications.

1. Accuracy: fraction of genes correctly classified. Calculation: $accuracy = \frac{TP+TN}{TP+TN+FP+FN}$

2. Precision: fraction of genes correctly classified as drivers among the total of genes classified as drivers. Calculation: $precision = \frac{TP}{TP+FP}$
3. Recall: fraction of genes correctly classified as drivers among the total of drivers. Calculation: $recall = \frac{TP}{TP+FN}$
4. F1 score: harmonic average of the precision and recall measurements. Calculation: $F1 = 2 \times \frac{precision \times recall}{precision + recall}$

Receiver operating characteristic (ROC) curves were generated for the models trained, according to both mutation and gene network features, and compared to the ROC curves obtained from models trained using only a single source of features (i.e., either mutation, or gene network data). The models with a single source of features were also trained by the same methodology and with hyperparameters selected through new grid-searches. The areas under the ROC curves (AUC) were also calculated for comparisons.

5.5.2 Results

The models induced using optimal hyperparameters evaluated using the metrics described before, and the mean of each metric was calculated over 30 repetitions of the whole training process. Tables 18 and 19 show the results for SVM and Random Forest, respectively. Averages and standard deviations were calculated for each selected evaluation metric.

Both models trained with combined features provided satisfactory results according to the selected metrics. Such models outperform the other models induced with a single source of features for the most of metrics. This trend was also observed in the analysis of ROC curves, depicted in Figures 52 and 53.

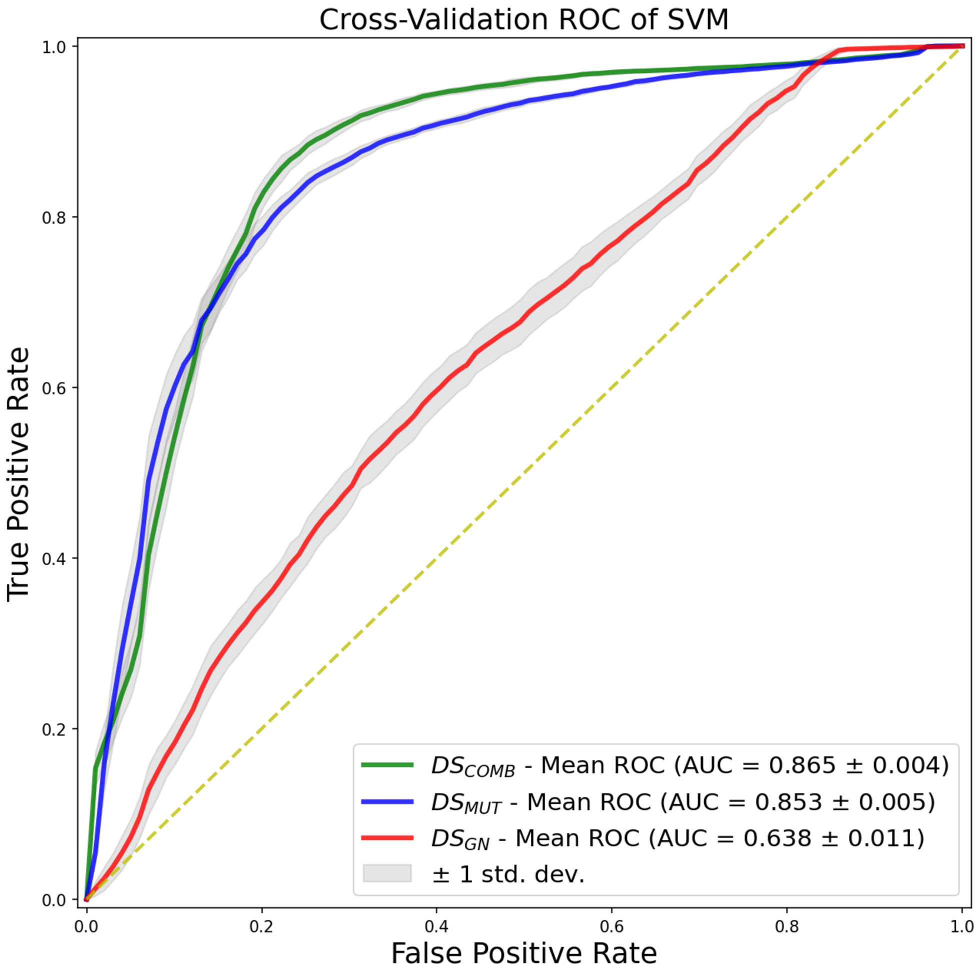
Table 18 – Comparison among SVM models induced by different features: DS_{COMB_L}, DS_{MUT_L}, and DS_{GN_L}.

	DS_{COMB_L}	DS_{MUT_L}	DS_{GN_L}
Accuracy	0.857 ± 0.006	0.839 ± 0.004	0.796 ± 0.005
Precision	0.928 ± 0.003	0.911 ± 0.003	0.819 ± 0.002
Recall	0.888 ± 0.006	0.884 ± 0.004	0.954 ± 0.005
F1	0.908 ± 0.004	0.897 ± 0.003	0.881 ± 0.003

Table 19 – Comparison among Random Forest models induced by different features: DS_{COMB_L}, DS_{MUT_L}, and DS_{GN_L}.

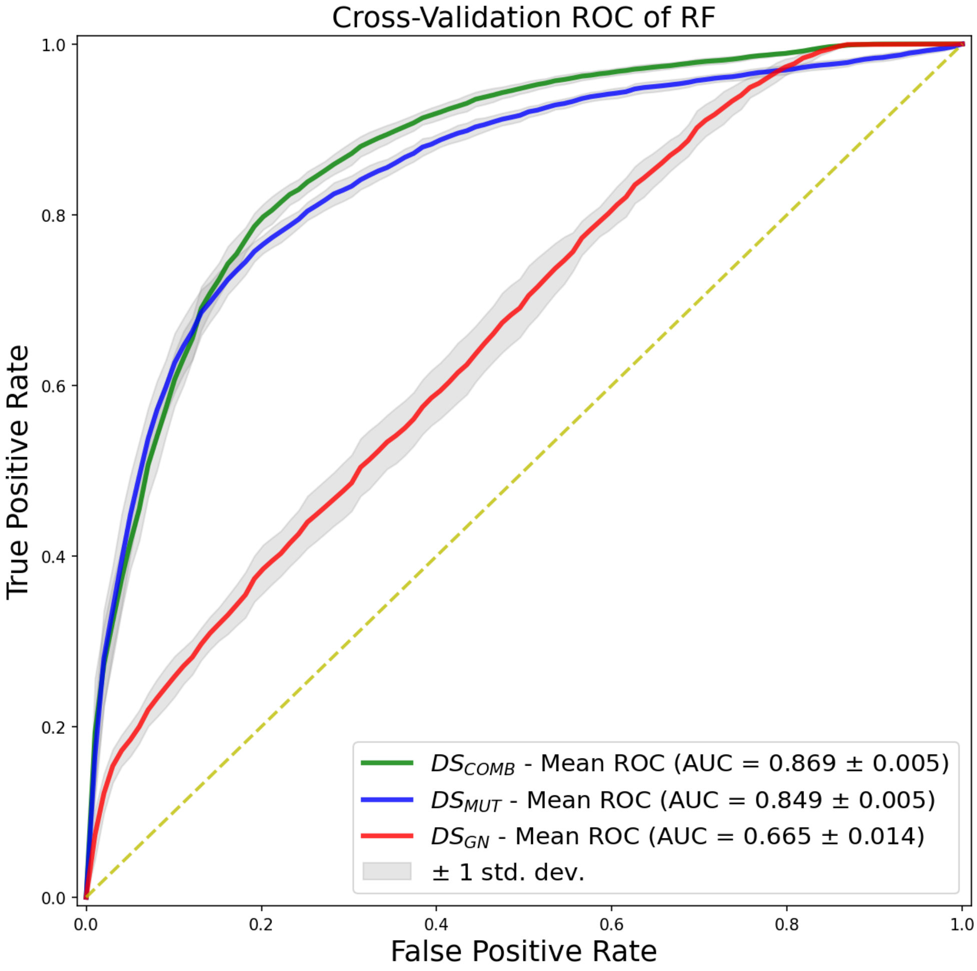
	DS_{COMB_L}	DS_{MUT_L}	DS_{GN_L}
Accuracy	0.850 ± 0.006	0.832 ± 0.006	0.780 ± 0.007
Precision	0.903 ± 0.005	0.885 ± 0.005	0.831 ± 0.004
Recall	0.907 ± 0.006	0.905 ± 0.005	0.905 ± 0.008
F1	0.905 ± 0.004	0.895 ± 0.004	0.867 ± 0.005

Figure 52 – ROC curve comparison for SVM models



Source: Elaborated by the author.

Figure 53 – ROC curve comparison for Random Forest models



Source: Elaborated by the author.

The possible novel cancer genes discovered by DiSCaGe shown in Figure 35 were subjected to the machine learning approach for the suggestion of possible false-positives on the DiSCaGe results. Both SVM and Random Forest models detect the same set of false-positive for LUAD, which are the genes: XIRP2, COL11A1, NAV3, ANK2, and PCDH15. For the other cancer cancer types no false-positives were detected.

The natural application of this discovery approach is to avoid the misclassification of false-positive-drivers as drivers and possibly eliminate unnecessary further analysis. Detecting false-drivers is also crucial to prevent their inclusion in data analyses or on the development of models, which could lead to the identification of unrealistic patterns. However, it is important to note that this concept has been implemented considering the currently available data, which is scarce and still under continuous investigation. Therefore, it is expected that the proposed approach can be eventually revisited and improved as new information becomes available.

5.6 Final Considerations

This chapter described DFDriver, a machine learning-based approach for the classification driver gene candidates in real or false-drivers. The chapter is based in a paper published at the *Brazilian Symposium on Bioinformatics (BSB 2020)* (CUTIGI *et al.*, 2020).

Nine measures from mutation data and nine from gene interactions were extracted, and Support Vector Machines and Random Forest models were induced by a combined source of features. Data were properly preprocessed, and stratified k-fold cross-validation was applied to the models' training. Moreover, a grid-search process was employed for hyperparameters optimization.

In general, DFDriver achieved satisfactory classification performance due to the combination of mutation and gene interaction features in both RF and SVM models.

The next chapter is devoted to the conclusions, contributions, and limitations of this thesis, as well as directions for possible future studies, and other secondary results.

CHAPTER 6

CONCLUSIONS

This PhD thesis has addressed a classical and ongoing problem in Cancer Bioinformatics and Genomics: the discovery of significant genes for cancer through computational approaches. The thesis describes two computational approaches to deal with this challenging problem. The first approach, called DiSCaGe (**Di**scovering **S**ignificant **Ca**ncer **Ge**nes), discovers significant cancer genes taking into account weighted mutations in genes and the way they can affect a network neighborhood through an asymmetric spreading strength measure. The second approach, called DFDriver (**D**etecting **F**alse **Driver**), identifies possible false-positive driver gene candidates through machine learning models induced by a combination of mutation and gene network data. Experimental evaluations were conducted for both approaches, with a set of known cancer genes benchmarks and an automated literature review of discovered genes.

The combination of weighted mutation frequency and network neighbors influence shows the potential of discovering significant genes for cancer, thus it was possible to investigate and answer the research question ***RQ1***. The results of experimental study shows DiSCaGe is able to prioritize known cancer-related genes, including genes with low mutation frequency, and cited in research papers as cancer-related genes. Furthermore, DiSCaGe also suggests possible novel cancer genes.

The potential of the combination of features from mutation and gene interaction network was confirmed on the training of machine learning models to detect possible false-positive significant genes for cancer, thus answering the research question ***RQ2***. The results of the experimental study with DFDriver shows that models trained with combined features outperform the other models induced with a single source of features for the most of metrics and by the observation of ROC curves.

The investigation of both ***RQ1*** and ***RQ2*** leads to reach the general objective of this thesis, which was to discover reliable significant cancer genes with the use of two computational approaches. Furthermore the hypothesis was confirmed, which was significantly mutated genes in cancer can be discovered through the combination of weighted mutation frequency and network neighbors influence, and possible false-positives can be detected by mutation data and gene interaction networks.

6.1 Contributions

The general contribution of this thesis was an investigation on the problem of discovering significant genes for cancer, which resulted in two computational approaches for the discovery of significant genes for cancer, avoiding possible false-positive results. With the proposed approaches is possible to:

1. **Prioritize genes known to be related to cancer:** Results shown the proposed approach can discover genes that are in reliable benchmarks of driver genes. It outperforms related methods for most types of cancer selected in the experimental study for most top N ranges of genes.

2. **Prioritize genes related to cancer with low mutation frequency:** The long tail analysis shown that the proposed approach can discover genes in the tail of the graph, i.e., genes with low mutation frequency, which is a challenging task.

3. **Suggest genes that are not in benchmarks but are cited in research papers as cancer-related ones:** An automated literature-based analysis on discovered genes shown that genes out of benchmarks are cited in research papers as cancer genes. It suggests such genes can be related to cancer; even they were not in benchmarks.

4. **Suggest possible novel cancer genes:** Some discovered genes by the proposed computational approach were not in benchmarks and are not cited. Such genes are suggestions for a further investigation through *in-vitro* and *in-vivo* experiments.

5. **Suggest possible false-positive cancer gene candidates:** Results shown the potential of the machine learning-based approach in classifying driver gene candidates in real or false drivers.

The contributions of the proposed approaches can be summarized in three perspectives:

1) Computational perspective: complex networks and their algorithms were used in gene interaction networks in combination with mutation data. Especially, an adapted asymmetric spreading strength was employed to quantify how a mutation can influence a gene neighborhood.

2) Biological perspective: The local hypothesis defined by Barabási, Gulbahce and Loscalzo (2011) was used as the basis of the method, together with weighted mutations, based on the functional impact of distinct types of mutation in the genes. Additionally, a link prediction approach was performed on the networks to deal with problem of incomplete gene interactions.

3) User's perspective: the approaches were built towards being easily adopted by end-users, demanding mutation data and gene networks as input in standard formats. Additionally, the user must define mutation weights, with no definition of unclear hyperparameters, thus facilitating the use.

Such a combination of computational, biological and user's perspectives enables the definition and development of efficient computational methods for the discovery of novel cancer genes.

Another important contribution is the evaluation of the results of computational methods that discover cancer genes. A systematic pipeline was defined that uses four recent known cancer genes benchmarks. Precision and DGC were used together to evaluate the methods. Such a pipeline can be used in future works to evaluate and compare existing and related methods.

As secondary result, a research group composed of researchers from University of Sao Paulo (Sao Carlos campus), Federal Institute of Sao Paulo (Sao Carlos and Barretos campus), and Barretos Cancer Hospital, have been established during the development of this research. A partnership with Barretos Cancer Hospital has been started, and has led to some initial collaborations.

The following research papers have been submitted and published in conferences and journals:

1. A short paper published at the *Simposio Brasileiro de Computacao Aplicada a Saude (SBCAS 2019)* (CUTIGI; EVANGELISTA; SIMAO, 2019), which is a preliminary proposal of a flexible computational method for ranking significant set of related genes in cancer, by considering data about mutations, type of mutations, gene interaction networks and mutual exclusivity pattern.

    - CUTIGI, J. F.; EVANGELISTA, A. F.; SIMAO, A. A proposal of a graph-based computational method for ranking significant set of related genes in cancer. In: **Anais Principais do XIX Simpósio Brasileiro de Computação Aplicada à Saúde**. Porto Alegre, RS, Brasil: SBC, 2019. p. 300–305. Available: <https://sol.sbc.org.br/index.php/sbcas/article/view/6266>.

2. A full paper published at the *Brazilian Symposium on Bioinformatics (BSB 2019)* (CUTIGI; EVANGELISTA; SIMAO, 2020b), which is the improvement and implementation of the idea published before.

   - CUTIGI, J. F.; EVANGELISTA, A. F.; SIMAO, A. GeNWeMME: A network-based computational method for prioritizing groups of significant related genes in cancer. In: SPRINGER. **Advances in Bioinformatics and Computational Biology**. [S.l.], 2020. p. 29–40. ISBN 978-3-030-46417-2.

3. A full paper published at the *Journal of Bioinformatics and Computational Biology (JBCB)* (CUTIGI; EVANGELISTA; SIMAO, 2020a), which addresses significant mutations in cancer and classical computational methods. It details some methods, presenting their approaches and algorithms, and briefly describes some other related works, this providing a summary of such methods. It also discusses their computational complexity and the way they can be evaluated and compared.

   - CUTIGI, J. F.; EVANGELISTA, A. F.; SIMAO, A. Approaches for the identification of driver mutations in cancer: A tutorial from a computational perspective. **Journal of Bioinformatics and Computational Biology**, v. 18, n. 03, p. 2050016, 2020. PMID: 32698724. Available: <https://doi.org/10.1142/S021972002050016X>.

4. A full paper published, as secondary author, at the *Simposio Brasileiro de Computacao Aplicada a Saude (SBCAS 2020)* (RAMOS *et al.*, 2020), which is a exploratory work that investigates the mutational characteristics presented in different cancer mutation data sets of the same type of cancer.

   - RAMOS, R. H.; CUTIGI, J. F.; FERREIRA, C. de O. L.; EVANGELISTA, A. F.; SIMAO, A. Analyzing different cancer mutation data sets from breast invasive carcinoma (brca), lung adenocarcinoma (luad), and prostate adenocarcinoma (prad). In: **Anais Principais do XX Simpósio Brasileiro de Computação Aplicada à Saúde**. Porto Alegre, RS, Brasil: SBC, 2020. p. 37–48. Available: <https://sol.sbc.org.br/index.php/sbcas/article/view/11500>.

5. A full paper published at the *Brazilian Symposium on Bioinformatics (BSB 2020)* (CUTIGI *et al.*, 2020), which presents a machine learning-based approach to induce predictive models able to classify driver gene candidates as real drivers or false-drivers.

   - CUTIGI, J. F.; EVANGELISTA, R. F.; RAMOS, R. H.; FERREIRA, C. d. O. L.; EVANGELISTA, A. F.; CARVALHO, A. C. de; SIMAO, A. Combining mutation and gene network data in a machine learning approach for false-positive cancer driver gene discovery. In: SPRINGER. **Brazilian Symposium on Bioinformatics**. [S.l.], 2020. p. 81–92.

6. A full paper submitted at the *Nature Scientific Reports* (CUTIGI *et al.*, 2021), which describes a computational method for the discovery of significant genes for cancer, which takes into account weighted mutations in genes and the way they can affect network neighborhood through an asymmetric spreading strength measure.

- CUTIGI, J. F.; EVANGELISTA, A. F.; REIS, R. M.; SIMAO, D. A. A computational approach for the discovery of significant cancer genes by weighted mutation and asymmetric spreading strength in networks. **Submitted to Nature Scientific Reports**, 2021.

All materials to allow the reproducibility of experiments and results (eg.: source codes, instructions of use, scripts of experiments and the complete list of libraries and versions) are available on the following link: <https://github.com/jcutigi/Thesis_ComputationalApproaches>.

6.2 Limitations and future directions

The limitations of the research refer to a lack of a systematic biological evaluation of the findings. Despite being a hard task, further *in-vitro* and *in-vivo* investigations can be performed to confirm the results of both approaches. Related to the computational approach, DiSCaGe is highly dependent on the networks, then the quality and the assertiveness of gene interactions are determinants for the result. Although the gene network enrichment and spreading strength minimize some outliers in the local neighborhood (e.g., star topology), the high connected genes can still receive a strong influence from their neighbors, leading to possible false positives. In this way, a smooth approach could be addressed to deal with high-connected genes. Additionally, a different preprocessing in the networks could be performed to deal with high-degree genes.

As future work, both approaches can be subjected to a pan-cancer study for their evaluation in a large number of cancer types, thus providing subsidies for their characterization and understanding of cases in which they can be properly adopted. A natural extension of the approaches is to allow the method to suggest possible driver pathways with the use of the final network and gene mutation score for finding significantly related genes. The development of an online tool will facilitate the use of DiSCaGe and DFDriver by end-users. The study of impact of different networks and their combination could be better addressed and evaluate in both methods. Furthermore, such an impact could be performed in network-based methods, since networks have significant impact on the methods results.

In sum, both approaches are not definitive solutions for the problem on identifying reliable cancer genes. The proposed approaches and existing related method can be complementary, i.e., some methods with different approaches should be selected and used together in order to get robust results.

BIBLIOGRAPHY

ALBERTS, B.; JOHNSON, A.; LEWIS, J.; RAFF, M.; ROBERTS, K.; WALTER, P. **Molecular biology of the cell, 5th edition**. [S.l.]: Garland Science, 2008. ISBN 978-0815341055. Citations on pages 34 and 35.

ALOY, P.; RUSSELL, R. B. Taking the mystery out of biological networks. **EMBO reports**, John Wiley & Sons, Ltd, v. 5, n. 4, p. 349–350, 2004. Citation on page 64.

AMERICAN CANCER SOCIETY. **Oncogenes and tumor suppressor genes**. 2014. [Online; accessed August-2021]. Available: <https://www.cancer.org/cancer/cancer-causes/genetics/genes-and-cancer/oncogenes-tumor-suppressor-genes.html>. Citation on page 38.

AMERICAN SOCIETY OF CLINICAL ONCOLOGY. **Stages of Cancer**. 2018. [Online; accessed August-2021]. Available: <https://www.cancer.net/navigating-cancer-care/diagnosing-cancer/stages-cancer>. Citation on page 36.

ASHWORTH, A.; LORD, C.; REIS-FILHO, J. Genetic interactions in cancer progression and treatment. **Cell**, v. 145, n. 1, p. 30 – 38, 2011. ISSN 0092-8674. Available: <http://www.sciencedirect.com/science/article/pii/S0092867411002972>. Citation on page 43.

BAILEY, M. H.; TOKHEIM, C.; PORTA-PARDO, E.; SENGUPTA, S.; BERTRAND, D.; WEERASINGHE, A.; COLAPRICO, A.; WENDL, M. C.; KIM, J.; REARDON, B.; NG, P. K.-S.; JEONG, K. J.; CAO, S.; WANG, Z.; GAO, J.; GAO, Q.; WANG, F.; LIU, E. M.; MULARONI, L.; RUBIO-PEREZ, C.; NAGARAJAN, N.; CORTéS-CIRIANO, I.; ZHOU, D. C.; LIANG, W.-W.; HESS, J. M.; YELLAPANTULA, V. D.; TAMBORERO, D.; GONZALEZ-PEREZ, A.; SUPHAVILAI, C.; KO, J. Y.; KHURANA, E.; PARK, P. J.; ALLEN, E. M. V.; LIANG, H. Comprehensive characterization of cancer driver genes and mutations. **Cell**, v. 173, n. 2, p. 371 – 385.e18, 2018. ISSN 0092-8674. Available: <http://www.sciencedirect.com/science/article/pii/S009286741830237X>. Citations on pages 29, 48, 83, and 107.

BARABÁSI, A.-L.; ALBERT, R. Emergence of scaling in random networks. **science**, American Association for the Advancement of Science, v. 286, n. 5439, p. 509–512, 1999. Citation on page 69.

BARABÁSI, A.-L.; GULBAHCE, N.; LOSCALZO, J. Network medicine: a network-based approach to human disease. **Nature reviews genetics**, Nature Publishing Group, v. 12, n. 1, p. 56–68, 2011. Citations on pages 30, 64, 69, 78, and 118.

BASHASHATI, A.; HAFFARI, G.; DING, J.; HA, G.; LUI, K.; ROSNER, J.; HUNTSMAN, D. G.; CALDAS, C.; APARICIO, S. A.; SHAH, S. P. Drivernet: uncovering the impact of somatic driver mutations on transcriptional networks in cancer. **Genome Biology**, v. 13, n. 12, p. R124, 2012. ISSN 1474-760X. Available: <http://dx.doi.org/10.1186/gb-2012-13-12-r124>. Citations on pages 28 and 50.

BATISTA, G. E.; PRATI, R. C.; MONARD, M. C. A study of the behavior of several methods for balancing machine learning training data. **ACM SIGKDD explorations newsletter**, ACM New York, NY, USA, v. 6, n. 1, p. 20–29, 2004. Citation on page 110.

BEHJATI, S.; TARPEY, P. S. What is next generation sequencing? **Archives of disease in childhood - Education & practice edition**, BMJ Publishing Group Ltd and Royal College of Paediatrics and Child Health, v. 98, n. 6, p. 236–238, Dec. 2013. ISSN 1743-0593. Available: <http://dx.doi.org/10.1136/archdischild-2013-304340>. Citation on page 42.

BING, Z.; TIAN, J.; ZHANG, J.; LI, X.; WANG, X.; YANG, K. An integrative model of mirna and mrna expression signature for patients of breast invasive carcinoma with radiotherapy prognosis. **Cancer Biotherapy and Radiopharmaceuticals**, Mary Ann Liebert, Inc. 140 Huguenot Street, 3rd Floor New Rochelle, NY 10801 USA, v. 31, n. 7, p. 253–260, 2016. Citation on page 57.

BRAY, F.; FERLAY, J.; SOERJOMATARAM, I.; SIEGEL, R. L.; TORRE, L. A.; JEMAL, A. Global cancer statistics 2018: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. **CA: A Cancer Journal for Clinicians**, v. 0, n. 0, 2018. Available: <https://onlinelibrary.wiley.com/doi/abs/10.3322/caac.21492>. Citation on page 27.

BUERMANS, H.; DUNNEN, J. den. Next generation sequencing technology: Advances and applications. **Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease**, v. 1842, n. 10, p. 1932 – 1941, 2014. ISSN 0925-4439. From genome to function. Available: <http://www.sciencedirect.com/science/article/pii/S092544391400180X>. Citation on page 42.

BURRELL, R. A.; MCGRANAHAN, N.; BARTEK, J.; SWANTON, C. The causes and consequences of genetic heterogeneity in cancer evolution. **Nature**, v. 501, p. 338–45, 09 2013. Citation on page 43.

CATALOGUE OF SOMATIC MUTATIONS IN CANCER. **Catalogue Of Somatic Mutations In Cancer**. 2021. [Online; accessed August-2021]. Available: <http://cancer.sanger.ac.uk/>. Citation on page 42.

CERAMI, E.; GAO, J.; DOGRUSOZ, U.; GROSS, B. E.; SUMER, S. O.; AKSOY, B. A.; JACOBSEN, A.; BYRNE, C. J.; HEUER, M. L.; LARSSON, E.; ANTIPIN, Y.; REVA, B.; GOLDBERG, A. P.; SANDER, C.; SCHULTZ, N. The cbio cancer genomics portal: An open platform for exploring multidimensional cancer genomics data. **Cancer Discovery**, American Association for Cancer Research, v. 2, n. 5, p. 401–404, 2012. ISSN 2159-8274. Citations on pages 57 and 107.

CHAWLA, N. V.; BOWYER, K. W.; HALL, L. O.; KEGELMEYER, W. P. Smote: synthetic minority over-sampling technique. **Journal of artificial intelligence research**, v. 16, p. 321–357, 2002. Citation on page 110.

CHENG, F.; ZHAO, J.; ZHAO, Z. Advances in computational approaches for prioritizing driver mutations and significantly mutated genes in cancer genomes. **Briefings in Bioinformatics**, v. 17, n. 4, p. 642, 2015. Available: <+http://dx.doi.org/10.1093/bib/bbv068>. Citations on pages 28, 53, and 54.

CHIAL, H. Proto-oncogenes to oncogenes to cancer. **Nature Education**, 2008. Citation on page 38.

CHIN, L.; ANDERSEN, J. N.; FUTREAL, P. A. Cancer genomics: from discovery science to personalized medicine. **Nature medicine**, Nature Publishing Group, v. 17, n. 3, p. 297, 2011. Citation on page 29.

CHO, A.; SHIM, J. E.; KIM, E.; SUPEK, F.; LEHNER, B.; LEE, I. Muffinn: cancer gene discovery via network analysis of somatic mutation data. **Genome Biology**, v. 17, n. 1, p. 129, Jun 2016. ISSN 1474-760X. Available: <https://doi.org/10.1186/s13059-016-0989-x>. Citations on pages 28, 50, and 93.

CIRIELLO, G.; CERAMI, E.; SANDER, C.; SCHULTZ, N. Mutual exclusivity analysis identifies oncogenic network modules. **Genome research**, Cold Spring Harbor Lab, v. 22, n. 2, p. 398–406, 2012. Citations on pages 28, 51, 64, and 65.

CISOWSKI, J.; BERGO, M. O. What makes oncogenes mutually exclusive? **Small GTPases**, Taylor & Francis, v. 8, n. 3, p. 187–192, 2017. Citation on page 45.

COLLIER, O.; STOVEN, V.; VERT, J.-P. Lotus: A single- and multitask machine learning algorithm for the prediction of cancer driver genes. **PLOS Computational Biology**, Public Library of Science, v. 15, n. 9, p. 1–27, 09 2019. Citation on page 51.

COSMIC. **Mutational Signatures**. 2021. [Online; accessed August-2021]. Available: <https://cancer.sanger.ac.uk/cosmic/signatures>. Citation on page 63.

CREIXELL, P.; REIMAND, J.; HAIDER, S.; WU, G.; SHIBATA, T.; VAZQUEZ, M.; MUSTONEN, V.; GONZALEZ-PEREZ, A.; PEARSON, J.; SANDER, C.; RAPHAEL, B. J.; MARKS, D. S.; OUELLETTE, B. F. F.; VALENCIA, A.; BADER, G. D.; BOUTROS, P. C.; STUART, J. M.; LINDING, R.; LOPEZ-BIGAS, N.; STEIN, L. D. Pathway and network analysis of cancer genomes. **Nature Methods**, Nature Publishing Group, a division of Macmillan Publishers Limited. All Rights Reserved., v. 12, n. 7, p. 615–621, Jun. 2015. ISSN 1548-7091. Available: <http://dx.doi.org/10.1038/nmeth.3440>. Citations on pages 29, 43, and 44.

CUTIGI, J. F.; EVANGELISTA, A. F.; REIS, R. M.; SIMAO, D. A. A computational approach for the discovery of significant cancer genes by weighted mutation and asymmetric spreading strength in networks. **Submitted to Nature Scientific Reports**, 2021. Citations on pages 56, 104, 120, and 121.

CUTIGI, J. F.; EVANGELISTA, A. F.; SIMAO, A. A proposal of a graph-based computational method for ranking significant set of related genes in cancer. In: **Anais Principais do XIX Simpósio Brasileiro de Computação Aplicada à Saúde**. Porto Alegre, RS, Brasil: SBC, 2019. p. 300–305. Available: <https://sol.sbc.org.br/index.php/sbcas/article/view/6266>. Citation on page 119.

CUTIGI, J. F.; EVANGELISTA, A. F.; SIMAO, A. Approaches for the identification of driver mutations in cancer: A tutorial from a computational perspective. **Journal of Bioinformatics and Computational Biology**, v. 18, n. 03, p. 2050016, 2020. PMID: 32698724. Available: <https://doi.org/10.1142/S021972002050016X>. Citations on pages 28, 33, 47, 53, 54, 82, and 120.

CUTIGI, J. F.; EVANGELISTA, A. F.; SIMAO, A. GeNWeMME: A network-based computational method for prioritizing groups of significant related genes in cancer. In: SPRINGER. **Advances in Bioinformatics and Computational Biology**. [S.l.], 2020. p. 29–40. ISBN 978-3-030-46417-2. Citations on pages 28, 51, 56, 104, 119, and 120.

CUTIGI, J. F.; EVANGELISTA, R. F.; RAMOS, R. H.; FERREIRA, C. d. O. L.; EVANGELISTA, A. F.; CARVALHO, A. C. de; SIMAO, A. Combining mutation and gene network data in a machine learning approach for false-positive cancer driver gene discovery. In: SPRINGER.

Brazilian Symposium on Bioinformatics. [S.l.], 2020. p. 81–92. Citations on pages 106, 116, and 120.

DAS, J.; YU, H. Hint: High-quality protein interactomes and their applications in understanding human disease. **BMC systems biology**, v. 6, p. 92, 2012. Citation on page 43.

DAVIES, H.; BIGNELL, G.; COX, C.; STEPHENS, P.; EDKINS, S.; CLEGG, S.; TEAGUE, J.; WOFFENDIN, H.; GARNETT, M.; BOTTOMLEY, W. *et al.* Mutations of the braf gene in human cancer. **Nature**, Nature Publishing Group, v. 417, n. 6892, p. 949–954, 2002. Citation on page 45.

DEES, N. D.; ZHANG, Q.; KANDOTH, C.; WENDL, M. C.; SCHIERDING, W.; KOBOLDT, D. C.; MOONEY, T. B.; CALLAWAY, M. B.; DOOLING, D.; MARDIS, E. R. *et al.* Music: identifying mutational significance in cancer genomes. **Genome research**, Cold Spring Harbor Lab, v. 22, n. 8, p. 1589–1598, 2012. Citations on pages 28 and 50.

DELLAIRE, G.; BERMAN, J. N.; ARCECI, R. J. **Cancer Genomics, First edition**. Academic Press, 2014. ISBN 978-0-12-396967-5. Available: <http://www.sciencedirect.com/science/article/pii/B978012396967500027X>. Citation on page 33.

DENG, Y.; LUO, S.; DENG, C.; LUO, T.; YIN, W.; ZHANG, H.; ZHANG, Y.; ZHANG, X.; LAN, Y.; PING, Y.; XIAO, Y.; LI, X. Identifying mutual exclusivity across cancer genomes: computational approaches to discover genetic interaction and reveal tumor vulnerability. **Briefings in Bioinformatics**, p. bbx109, 2017. Available: <+http://dx.doi.org/10.1093/bib/bbx109>. Citations on pages 29, 44, 45, 53, and 54.

DIMITRAKOPOULOS, C. M.; BEERENWINKEL, N. Computational approaches for the identification of cancer genes and pathways. **Wiley Interdisciplinary Reviews: Systems Biology and Medicine**, v. 9, n. 1, p. e1364, 2017. Available: <https://onlinelibrary.wiley.com/doi/abs/10.1002/wsbm.1364>. Citations on pages 28, 53, and 54.

DING, J.; MCCONECHY, M. K.; HORLINGS, H. M.; HA, G.; CHAN, F. C.; FUNNELL, T.; MULLALY, S. C.; REIMAND, J.; BASHASHATI, A.; BADER, G. D. *et al.* Systematic analysis of somatic mutations impacting gene expression in 12 tumour types. **Nature communications**, Nature Publishing Group, v. 6, n. 1, p. 1–13, 2015. Citation on page 78.

EMORY UNIVERSITY – WINSHIP CANCER INSTITUTE. **Cancer Biology: Cancer Genes**. 2020. [Online; accessed August-2021]. Available: <https://www.cancerquest.org/cancer-biology/cancer-genes>. Citation on page 38.

ETUDE, P. J. Comparative de la distribution florale dans une portion des alpes et des jura. **Bull. Soc. Vaud. Sci. Nat**, v. 37, p. 547, 1901. Citation on page 61.

FABREGAT, A.; JUPE, S.; MATTHEWS, L.; SIDIROPOULOS, K.; GILLESPIE, M.; GARAPATI, P.; HAW, R.; JASSAL, B.; KORNINGER, F.; MAY, B.; MILACIC, M.; ROCA, C. D.; ROTHFELS, K.; SEVILLA, C.; SHAMOVSKY, V.; SHORSER, S.; VARUSAI, T.; VITERI, G.; WEISER, J.; WU, G.; STEIN, L.; HERMJAKOB, H.; D'EUSTACHIO, P. The reactome pathway knowledgebase. **Nucleic Acids Research**, v. 46, n. D1, p. D649–D655, 2018. Available: <http://dx.doi.org/10.1093/nar/gkx1132>. Citations on pages 44 and 64.

FUTREAL, P. A.; COIN, L.; MARSHALL, M.; DOWN, T.; HUBBARD, T.; WOOSTER, R.; RAHMAN, N.; STRATTON, M. R. A census of human cancer genes. **Nature reviews. Cancer**, v. 4, n. 3, p. 177–83, Mar 2004. Citation on page 83.

GAO, J.; AKSOY, B. A.; DOGRUSOZ, U.; DRESDNER, G.; GROSS, B.; SUMER, S. O.; SUN, Y.; JACOBSEN, A.; SINHA, R.; LARSSON, E. *et al.* Integrative analysis of complex cancer genomics and clinical profiles using the cbioportal. **Sci. Signal.**, American Association for the Advancement of Science, v. 6, n. 269, p. pl1–pl1, 2013. Citation on page 57.

GARRAWAY, L. A.; LANDER, E. S. Lessons from the cancer genome. **Cell**, v. 153, n. 1, p. 17–37, Mar. 2013. Available: <https://www.ncbi.nlm.nih.gov/pubmed/23540688>. Citations on pages 48 and 49.

GREENMAN, C.; STEPHENS, P.; SMITH, R.; DALGLIESH, G.; HUNTER, C.; BIGNELL, G.; DAVIES, H.; TEAGUE, J.; BUTLER, A.; STEVENS, C.; OTHERS; STRATTON, M. Patterns of somatic mutation in human cancer genomes. **Nature**, Nature Publishing Group, v. 446, n. 7132, p. 153–158, 2007. Citation on page 48.

GUMPINGER, A. C.; LAGE, K.; HORN, H.; BORGWARDT, K. Prediction of cancer driver genes through network-based moment propagation of mutation scores. **Bioinformatics**, v. 36, n. Supplement_1, p. i508–i515, 07 2020. ISSN 1367-4803. Citation on page 51.

HABER, D. A.; SETTLEMAN, J. Cancer: drivers and passengers. **Nature**, v. 446, n. 7132, p. 145–6, Mar 2007. Citation on page 48.

HAN, Y.; YANG, J.; QIAN, X.; CHENG, W.-C.; LIU, S.-H.; HUA, X.; ZHOU, L.; YANG, Y.; WU, Q.; LIU, P. *et al.* Driverml: a machine learning algorithm for identifying driver genes in cancer sequencing studies. **Nucleic acids research**, Oxford University Press, v. 47, n. 8, p. e45–e45, 2019. Citation on page 51.

HANAHAN, D.; WEINBERG, R. A. The hallmarks of cancer. **Cell**, v. 100, p. 57–70, 2000. Citations on pages 36 and 37.

HANAHAN, D.; WEINBERG, R. A. Hallmarks of cancer: the next generation. **Cell**, v. 144, n. 5, p. 646–74, 2011. Citations on pages 36, 37, and 48.

HODIS, E.; WATSON, I. R.; KRYUKOV, G. V.; AROLD, S. T.; IMIELINSKI, M.; THEURILLAT, J.-P.; NICKERSON, E.; AUCLAIR, D.; LI, L.; PLACE, C.; DICARA, D.; RAMOS, A. H.; LAWRENCE, M. S.; CIBULSKIS, K.; SIVACHENKO, A.; VOET, D.; SAKSENA, G.; STRANSKY, N.; ONOFRIO, R. C.; WINCKLER, W.; ARDLIE, K.; WAGLE, N.; WARGO, J.; CHONG, K.; MORTON, D. L.; STEMKE-HALE, K.; CHEN, G.; NOBLE, M.; MEYERSON, M.; LADBURY, J. E.; DAVIES, M. A.; GERSHENWALD, J. E.; WAGNER, S. N.; HOON, D. S.; SCHADENDORF, D.; LANDER, E. S.; GABRIEL, S. B.; GETZ, G.; GARRAWAY, L. A.; CHIN, L. A landscape of driver mutations in melanoma. **Cell**, v. 150, n. 2, p. 251–263, 2012. Exported from https://app.dimensions.ai on 2018/07/30. Available: <https://app.dimensions.ai/details/publication/pub.1024635810andhttps://doi.org/10.1016/j.cell.2012.06.024>. Citation on page 28.

HORN, H.; LAWRENCE, M.; CHOUINARD, C. R.; SHRESTHA, Y.; HU, J. X.; WORSTELL, E.; SHEA, E.; ILIC, N.; KIM, E.; KAMBUROV, A.; KASHANI, A.; HAHN, W. C.; CAMPBELL, J. D.; BOEHM, J. S.; GETZ, G.; LAGE, K. Netsig: Network-based discovery from cancer genomes. **Nature Methods**, v. 15, p. 61–66, 01 2018. Citation on page 28.

HOU, J. P.; EMAD, A.; PULEO, G. J.; MA, J.; MILENKOVIC, O. A new correlation clustering method for cancer mutation analysis. **Bioinformatics**, v. 32, n. 24, p. 3717–3728, 2016. Available: <http://dx.doi.org/10.1093/bioinformatics/btw546>. Citation on page 28.

HOU, J. P.; MA, J. Identifying driver mutations in cancer. In: ____. **Bioinformatics for Diagnosis, Prognosis and Treatment of Complex Diseases**. [S.l.]: Springer Netherlands, 2013. p. 33–56. Citations on pages 28, 48, 49, 53, and 54.

HOU, J. P.; MA, J. Dawnrank: discovering personalized driver genes in cancer. **Genome Medicine**, v. 6, n. 7, p. 56, 2014. ISSN 1756-994X. Available: <http://dx.doi.org/10.1186/s13073-014-0056-8>. Citations on pages 28 and 50.

HRISTOV, B. H.; SINGH, M. Network-based coverage of mutational profiles reveals cancer genes. **Cell systems**, Elsevier, v. 5, n. 3, p. 221–229, 2017. Citations on pages 28, 50, and 93.

HUANG, Z.-G.; HE, R.-Q.; MO, Z.-N. Prognostic value and potential function of splicing events in prostate adenocarcinoma. **International journal of oncology**, Spandidos Publications, v. 53, n. 6, p. 2473–2487, 2018. Citation on page 57.

INSTITUTO NACIONAL DE CâNCER. **O que é cancer**. 2019. [Online; accessed August-2021]. Available: <https://www.inca.gov.br/o-que-e-cancer>. Citation on page 36.

INSTITUTO NACIONAL DE CâNCER. **Tipos de Câncer**. 2020. [Online; accessed August-2021]. Available: <https://www.inca.gov.br/tipos-de-cancer>. Citation on page 36.

INTERNATIONAL CANCER GENOME CONSORTIUM. **International Cancer Genome Consortium**. 2019. [Online; accessed August-2021]. Available: <http://icgc.org/>. Citation on page 42.

JäRVELIN, K.; KEKäLäINEN, J. Cumulated gain-based evaluation of ir techniques. **ACM Trans. Inf. Syst.**, Association for Computing Machinery, New York, NY, USA, v. 20, n. 4, p. 422–446, Oct. 2002. ISSN 1046-8188. Available: <https://doi.org/10.1145/582415.582418>. Citation on page 83.

JASSAL, B.; MATTHEWS, L.; VITERI, G.; GONG, C.; LORENTE, P.; FABREGAT, A.; SIDIROPOULOS, K.; COOK, J.; GILLESPIE, M.; HAW, R. *et al.* The reactome pathway knowledgebase. **Nucleic acids research**, Oxford University Press, v. 48, n. D1, p. D498–D503, 2020. Citations on pages 44 and 64.

KANEHISA, M.; GOTO, S. Kegg: kyoto encyclopedia of genes and genomes. **Nucleic acids research**, Oxford University Press, v. 28, n. 1, p. 27–30, 2000. Citations on pages 44 and 65.

KANEHISA, M.; GOTO, S.; SATO, Y.; FURUMICHI, M.; TANABE, M. Kegg for integration and interpretation of large-scale molecular data sets. **Nucleic Acids Research**, v. 40, n. D1, p. D109–D114, 2012. Available: <http://dx.doi.org/10.1093/nar/gkr988>. Citations on pages 44 and 65.

KHURANA, E.; FU, Y.; CHEN, J.; GERSTEIN, M. Interpretation of genomic variants using a unified biological network approach. **PLOS Computational Biology**, Public Library of Science, v. 9, n. 3, p. 1–9, 03 2013. Available: <https://doi.org/10.1371/journal.pcbi.1002886>. Citation on page 44.

KIM, Y.; CHO, D.; PRZYTYCKA, T. M. Understanding genotype-phenotype effects in cancer via network approaches. **PLoS Computational Biology**, v. 12, n. 3, 2016. Available: <https://doi.org/10.1371/journal.pcbi.1004747>. Citation on page 43.

KIM, Y.-A.; CHO, D.-Y.; DAO, P.; PRZYTYCKA, T. M. Memcover: integrated analysis of mutual exclusivity and functional network reveals dysregulated pathways across multiple cancer types. **Bioinformatics**, v. 31, n. 12, p. i284–i292, 2015. Available: <http://dx.doi.org/10.1093/bioinformatics/btv247>. Citations on pages 28 and 51.

KIM, Y.-A.; MADAN, S.; PRZYTYCKA, T. M. Wesme: uncovering mutual exclusivity of cancer drivers and beyond. **Bioinformatics**, v. 33, n. 6, p. 814–821, 2017. Available: <+http://dx.doi.org/10.1093/bioinformatics/btw242>. Citations on pages 45 and 51.

LAWRENCE, M. S.; STOJANOV, P.; POLAK, P.; KRYUKOV, G. V.; CIBULSKIS, K.; SIVACHENKO, A.; CARTER, S. L.; STEWART, C.; MERMEL, C. H.; ROBERTS, S. A. *et al.* Mutational heterogeneity in cancer and the search for new cancer-associated genes. **Nature**, Nature Publishing Group, v. 499, n. 7457, p. 214–218, 2013. Citations on pages 28, 50, and 93.

LEISERSON, M. D.; REYNA, M. A.; RAPHAEL, B. J. A weighted exact test for mutually exclusive mutations in cancer. **Bioinformatics**, v. 32, n. 17, p. i736–i745, 2016. Available: <http://dx.doi.org/10.1093/bioinformatics/btw462>. Citations on pages 28 and 51.

LEISERSON, M. D.; WU, H.-T.; VANDIN, F.; RAPHAEL, B. J. Comet: a statistical approach to identify combinations of mutually exclusive alterations in cancer. **Genome Biology**, v. 16, n. 1, p. 160, 2015. ISSN 1474-760X. Available: <http://dx.doi.org/10.1186/s13059-015-0700-7>. Citations on pages 28 and 51.

LEISERSON, M. D. M.; BLOKH, D.; SHARAN, R.; RAPHAEL, B. J. Simultaneous identification of multiple driver pathways in cancer. **PLOS Computational Biology**, Public Library of Science, v. 9, n. 5, p. 1–15, 05 2013. Available: <https://doi.org/10.1371/journal.pcbi.1003054>. Citations on pages 28 and 51.

LEISERSON, M. D. M.; VANDIN, F.; WU, H.-T.; DOBSON, J. R.; ELDRIDGE, J. V.; THOMAS, J. L.; PAPOUTSAKI, A.; KIM, Y.; NIU, B.; MCLELLAN, M.; LAWRENCE, M. S.; GONZALEZ-PEREZ, A.; TAMBORERO, D.; CHENG, Y.; RYSLIK, G. A.; LOPEZ-BIGAS, N.; GETZ, G.; DING, L.; RAPHAEL, B. J. Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes. **Nature Genetics**, v. 47, n. 2, p. 106–114, 02 2015. Available: <http://dx.doi.org/10.1038/ng.3168>. Citation on page 28.

LEVER, J.; ZHAO, E. Y.; GREWAL, J.; JONES, M. R.; JONES, S. J. M. Cancermine: a literature-mined resource for drivers, oncogenes and tumor suppressors in cancer. **Nature Methods**, v. 16, n. 6, p. 505–507, 2019. ISSN 15487105. Citations on pages 84 and 92.

LEWIN, B. **Genes IX , 9th edition**. [S.l.]: Jones & Bartlett Learning, 2007. ISBN 978-0763740634. Citations on pages 35 and 38.

LIBERZON, A.; BIRGER, C.; THORVALDSDÓTTIR, H.; GHANDI, M.; MESIROV, J. P.; TAMAYO, P. The molecular signatures database hallmark gene set collection. **Cell systems**, Elsevier, v. 1, n. 6, p. 417–425, 2015. Citation on page 65.

LIU, C.; MA, Y.; ZHAO, J.; NUSSINOV, R.; ZHANG, Y.-C.; CHENG, F.; ZHANG, Z.-K. Computational network biology: Data, model, and applications. **Physics Reports**, Elsevier, v. 846, p. 1–66, 2020. ISSN 0370-1573. Citations on pages 64 and 69.

LIU, Y.; TANG, M.; DO, Y.; HUI, P. M. Accurate ranking of influential spreaders in networks based on dynamically asymmetric link weights. **Physical Review E**, APS, v. 96, n. 2, p. 022323, 2017. Citation on page 78.

LUCK, K.; KIM, D.-K.; LAMBOURNE, L.; SPIROHN, K.; BEGG, B. E.; BIAN, W.; BRIGNALL, R.; CAFARELLI, T.; CAMPOS-LABORIE, F. J.; CHARLOTEAUX, B. *et al.* A reference map of the human binary protein interactome. **Nature**, Nature Publishing Group, p. 1–7, 2020. Citation on page 44.

MARTÍNEZ-JIMÉNEZ, F.; MUIÑOS, F.; SENTÍS, I.; DEU-PONS, J.; REYES-SALAZAR, I.; ARNEDO-PAC, C.; MULARONI, L.; PICH, O.; BONET, J.; KRANAS, H. *et al.* A compenium of mutational cancer driver genes. **Nature Reviews Cancer**, Nature Publishing Group, p. 1–18, 2020. Citation on page 83.

MERIC-BERNSTAM, F.; MILLS, G. Overcoming implementation challenges of personalized cancer therapy. **Nature Reviews Clinical Oncology**, Nature Publishing Group, 2012. Citations on pages 27 and 28.

MERING, C. V.; KRAUSE, R.; SNEL, B.; CORNELL, M.; OLIVER, S. G.; FIELDS, S.; BORK, P. Comparative assessment of large-scale data sets of protein–protein interactions. **Nature**, Nature Publishing Group, v. 417, n. 6887, p. 399–403, 2002. Citation on page 64.

MILLER, C. A.; SETTLE, S. H.; SULMAN, E. P.; ALDAPE, K. D.; MILOSAVLJEVIC, A. Discovering functional modules by identifying recurrent and mutually exclusive mutational patterns in tumors. **BMC Med Genomics**, v. 4, 2011. Citation on page 28.

NATIONAL CANCER INSTITUTE. **Cancer Genomics Research**. 2018. [Online; accessed August-2021]. Available: <https://www.cancer.gov/research/areas/genomics>. Citation on page 33.

NETWORK, C. G. A. R. *et al.* Comprehensive molecular profiling of lung adenocarcinoma. **Nature**, Nature Publishing Group, v. 511, n. 7511, p. 543–550, 2014. Citation on page 57.

NOWELL, P. The clonal evolution of tumor cell populations. **Science**, American Association for the Advancement of Science, v. 194, n. 4260, p. 23–28, 1976. ISSN 0036-8075. Available: <http://science.sciencemag.org/content/194/4260/23>. Citation on page 35.

NUSSINOV, R.; JANG, H.; TSAI, C.-J.; CHENG, F. Precision medicine and driver mutations: Computational methods, functional assays and conformational principles for interpreting cancer drivers. **PLoS computational biology**, Public Library of Science San Francisco, CA USA, v. 15, n. 3, p. e1006658, 2019. Citation on page 29.

NUSSINOV, R.; TSAI, C.-J.; JANG, H. Why are some driver mutations rare? **Trends in pharmacological sciences**, Elsevier, v. 40, n. 12, p. 919–929, 2019. Citation on page 29.

OLDHAM, S.; FULCHER, B.; PARKES, L.; ARNATKEVICIUTE, A.; SUO, C.; FORNITO, A. Consistency and differences between centrality measures across distinct classes of networks. **PLOS ONE**, Public Library of Science, v. 14, n. 7, p. 1–23, 07 2019. Citations on pages 108 and 109.

OZTURK, K.; DOW, M.; CARLIN, D. E.; BEJAR, R.; CARTER, H. The emerging potential for network analysis to inform precision cancer medicine. **Journal of molecular biology**, Elsevier, v. 430, n. 18, p. 2875–2899, 2018. Citations on pages 28, 29, 44, 53, and 54.

PAN-CANCER ANALYSIS OF WHOLE GENOMES. **Pan-Cancer Analysis of Whole Genomes**. 2021. [Online; accessed August-2021]. Available: <https://dcc.icgc.org/pcawg>. Citation on page 42.

PARSONS, D. W.; JONES, S.; ZHANG, X.; LIN, J. C.-H.; LEARY, R. J.; ANGENENDT, P.; MANKOO, P.; CARTER, H.; SIU, I.-M.; GALLIA, G. L. *et al.* An integrated genomic analysis of human glioblastoma multiforme. **science**, American Association for the Advancement of Science, v. 321, n. 5897, p. 1807–1812, 2008. Citation on page 57.

PEDREGOSA, F.; VAROQUAUX, G.; GRAMFORT, A.; MICHEL, V.; THIRION, B.; GRISEL, O.; BLONDEL, M.; PRETTENHOFER, P.; WEISS, R.; DUBOURG, V.; VANDERPLAS, J.; PASSOS, A.; COURNAPEAU, D.; BRUCHER, M.; PERROT, M.; DUCHESNAY, E. Scikit-learn: Machine learning in Python. **Journal of Machine Learning Research**, v. 12, p. 2825–2830, 2011. Citation on page 109.

PE'ER, D.; HACOHEN, N. Principles and strategies for developing network models in cancer. **Cell**, Cell Press, v. 144, n. 6, p. 864–873, Mar. 2011. ISSN 1097-4172. Available: <http://dx.doi.org/10.1016/j.cell.2011.03.001>. Citation on page 42.

PERI, S.; NAVARRO, J. D.; AMANCHY, R.; KRISTIANSEN, T. Z.; JONNALAGADDA, C. K.; SURENDRANATH, V.; NIRANJAN, V.; MUTHUSAMY, B.; GANDHI, T. K. B.; GRONBORG, M.; IBARROLA, N.; DESHPANDE, N.; SHANKER, K.; SHIVASHANKAR, H. N.; RASHMI, B. P.; RAMYA, M. A.; ZHAO, Z.; CHANDRIKA, K. N.; PADMA, N.; HARSHA, H. C.; YATISH, A. J.; KAVITHA, M. P.; MENEZES, M.; CHOUDHURY, D. R.; SURESH, S.; GHOSH, N.; SARAVANA, R.; CHANDRAN, S.; KRISHNA, S.; JOY, M.; ANAND, S. K.; MADAVAN, V.; JOSEPH, A.; WONG, G. W.; SCHIEMANN, W. P.; CONSTANTINESCU, S. N.; HUANG, L.; KHOSRAVI-FAR, R.; STEEN, H.; TEWARI, M.; GHAFFARI, S.; BLOBE, G. C.; DANG, C. V.; GARCIA, J. G. N.; PEVSNER, J.; JENSEN, O. N.; ROEPSTORFF, P.; DESHPANDE, K. S.; CHINNAIYAN, A. M.; HAMOSH, A.; CHAKRAVARTI, A.; PANDEY, A. Development of human protein reference database as an initial platform for approaching systems biology in humans. **Genome research**, v. 13, n. 10, p. 2363–71, Oct 2003. Citations on pages 43 and 64.

PLEASANCE, E.; CHEETHAM, R.; STEPHENS, P.; MCBRIDE, D.; HUMPHRAY, S.; GREENMAN, C.; VARELA, I.; LIN, M.; NEZ, G. O.; BIGNELL, G. *et al.* A comprehensive catalogue of somatic mutations from a human cancer genome. **Nature**, Nature Publishing Group, v. 463, n. 7278, p. 191–196, 2009. Citation on page 39.

PRASAD, T. S. K.; GOEL, R.; KANDASAMY, K.; KEERTHIKUMAR, S.; KUMAR, S.; MATHIVANAN, S.; TELIKICHERLA, D.; RAJU, R.; SHAFREEN, B.; VENUGOPAL, A.; BALAKRISHNAN, L.; MARIMUTHU, A.; BANERJEE, S.; SOMANATHAN, D. S.; SEBASTIAN, A.; RANI, S.; RAY, S.; KISHORE, C. J. H.; KANTH, S.; AHMED, M.; KASHYAP, M. K.; MOHMOOD, R.; RAMACHANDRA, Y. L.; KRISHNA, V.; RAHIMAN, B. A.; MOHAN, S.; RANGANATHAN, P.; RAMABADRAN, S.; CHAERKADY, R.; PANDEY, A. Human protein reference database–2009 update. **Nucleic acids research**, v. 37, n. Database issue, p. D767–72, Jan 2009. Citations on pages 43 and 64.

RAMOS, R. H.; CUTIGI, J. F.; FERREIRA, C. de O. L.; EVANGELISTA, A. F.; SIMAO, A. Analyzing different cancer mutation data sets from breast invasive carcinoma (brca), lung adenocarcinoma (luad), and prostate adenocarcinoma (prad). In: **Anais Principais do XX Simpósio Brasileiro de Computação Aplicada à Saúde**. Porto Alegre, RS, Brasil: SBC, 2020. p. 37–48. Available: <https://sol.sbc.org.br/index.php/sbcas/article/view/11500>. Citations on pages 56 and 120.

RAPHAEL, B. J.; DOBSON, J. R.; OESPER, L.; VANDIN, F. Identifying driver mutations in sequenced cancer genomes: computational approaches to enable precision medicine. **Genome**

Medicine, v. 6, n. 1, p. 5, 2014. ISSN 1756-994X. Available: <http://dx.doi.org/10.1186/gm524>. Citations on pages 28, 50, 53, and 54.

RAZICK, S.; MAGKLARAS, G.; DONALDSON, I. M. irefindex: a consolidated protein interaction database with provenance. **BMC bioinformatics**, v. 9, p. 405, 2008. Citation on page 44.

REPANA, D.; NULSEN, J.; DRESSLER, L.; BORTOLOMEAZZI, M.; VENKATA, S. K.; TOURNA, A.; YAKOVLEVA, A.; PALMIERI, T.; CICCARELLI, F. D. The network of cancer genes (ncg): a comprehensive catalogue of known and candidate cancer genes from cancer sequencing screens. **Genome Biology**, v. 20, n. 1, p. 1, Jan 2019. ISSN 1474-760X. Citation on page 83.

REYNA, M. A.; CHITRA, U.; ELYANOW, R.; RAPHAEL, B. J. Netmix: A network-structured mixture model for reduced-bias estimation of altered subnetworks. **Journal of Computational Biology**, Mary Ann Liebert, Inc., publishers 140 Huguenot Street, 3rd Floor New ..., 2021. Citation on page 51.

REYNA, M. A.; LEISERSON, M. D. M.; RAPHAEL, B. J. Hierarchical hotnet: identifying hierarchies of altered subnetworks. **Bioinformatics**, v. 34, n. 17, p. i972–i980, 2018. Available: <http://dx.doi.org/10.1093/bioinformatics/bty613>. Citations on pages 28 and 51.

ROLLAND, T.; FONTANILLO, C.; RIVAS, J. D. L.; VIDAL, M. A proteome-scale map of the human interactome network. **Cell**, Elsevier, v. 159, p. 1212–1226, 2014. Available: <http://digital.csic.es/handle/10261/134690>. Citation on page 43.

SHYR, C.; TARAILO-GRAOVAC, M.; GOTTLIEB, M.; LEE, J. J.; KARNEBEEK, C. van; WASSERMAN, W. W. Flags, frequently mutated genes in public exomes. **BMC medical genomics**, BioMed Central, v. 7, n. 1, p. 64, 2014. Citation on page 60.

SONDKA, Z.; BAMFORD, S.; COLE, C. G.; WARD, S. A.; DUNHAM, I.; FORBES, S. A. The cosmic cancer gene census: describing genetic dysfunction across all human cancers. **Nature Reviews Cancer**, Nature Publishing Group, v. 18, n. 11, p. 696–705, 2018. Citation on page 83.

SOON, W. W.; HARIHARAN, M.; SNYDER, M. P. High-throughput sequencing for biology and medicine. **Mol. Syst. Biol.**, v. 9, p. 640, 2013. Citation on page 27.

STRATTON, M. R. The cancer genome. **Nature**, Nature Publishing Group, v. 458, n. 7239, p. 719–724, 2009. ISSN 0028-0836. Available: <https://www.nature.com/articles/nature07943>. Citations on pages 27, 35, 39, and 48.

STRATTON, M. R. Exploring the genomes of cancer cells: Progress and promise. **Science**, American Association for the Advancement of Science, v. 331, n. 6024, p. 1553–1558, 2011. ISSN 0036-8075. Available: <http://science.sciencemag.org/content/331/6024/1553>. Citations on pages 38, 39, and 48.

STRATTON, M. R. Journeys into the genome of cancer cells. **EMBO Molecular Medicine**, EMBO Press, v. 5, n. 2, p. 169–172, 2013. ISSN 1757-4676. Available: <http://embomolmed.embopress.org/content/5/2/169>. Citation on page 42.

SUBRAMANIAN, A.; TAMAYO, P.; MOOTHA, V. K.; MUKHERJEE, S.; EBERT, B. L.; GILLETTE, M. A.; PAULOVICH, A.; POMEROY, S. L.; GOLUB, T. R.; LANDER, E. S.; MESIROV, J. P. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. **Proceedings of the National Academy of Sciences**, National Academy of Sciences, v. 102, n. 43, p. 15545–15550, 2005. ISSN 0027-8424. Available: <http://www.pnas.org/content/102/43/15545>. Citation on page 65.

SZYMKIEWICZ, D. Une conlribution statistique à la géographie floristique. **Acta Societatis Botanicorum Poloniae**, v. 11, n. 3, p. 249–265, 1934. Citation on page 65.

TAMBORERO, D.; GONZALEZ-PEREZ, A.; PEREZ-LLAMAS, C.; DEU-PONS, J.; KANDOTH, C.; REIMAND, J.; LAWRENCE, M. S.; GETZ, G.; BADER, G. D.; DING, L.; LOPEZ-BIGAS, N. Comprehensive identification of mutational cancer driver genes across 12 tumor types. **Scientific Reports**, The Author(s), v. 3, p. 2650–, Oct. 2013. Citations on pages 58 and 107.

THE CANCER GENOME ATLAS. **The Cancer Genome Atlas**. 2021. [Online; accessed August-2021]. Available: <https://cancergenome.nih.gov/>. Citation on page 42.

THE NATIONAL HUMAN GENOME RESEARCH INSTITUTE. **Chromosome Abnormalities**. 2021. [Online; accessed August-2021]. Available: <https://www.genome.gov/11508982/>. Citation on page 35.

THINNERGENE. **Chromosomes, DNA e Genes**. 2018. [Online; accessed December-2018]. Available: <http://www.thinnergene.com/about-thinnergene/genetics-101/>. Citation on page 34.

THOMAS, R.; BAKER, A.; DEBIASI, R.; WINCKLER, W.; LAFRAMBOISE, T.; LIN, W.; WANG, M.; FENG, W.; ZANDER, T.; MACCONAILL, L. *et al.* High-throughput oncogene mutation profiling in human cancer. **Nature genetics**, Nature Publishing Group, v. 39, n. 3, p. 347–351, 2007. Citation on page 44.

TRAN, B.; DANCEY, J. E.; KAMEL-REID, S.; MCPHERSON, J. D.; BEDARD, P. L.; BROWN, A. M.; ZHANG, T.; SHAW, P.; ONETTO, N.; STEIN, L.; HUDSON, T. J.; NEEL, B. G.; SIU, L. L. Cancer genomics: Technology, discovery, and translation. **Journal of Clinical Oncology**, v. 30, n. 6, p. 647–660, 2012. PMID: 22271477. Available: <https://doi.org/10.1200/JCO.2011.39.2316>. Citation on page 48.

VANDIN, F.; UPFAL, E.; RAPHAEL, B. J. Algorithms for detecting significantly mutated pathways in cancer. **Journal of Computational Biology**, v. 18, n. 3, p. 507–522, 2011. PMID: 21385051. Available: <https://doi.org/10.1089/cmb.2010.0265>. Citations on pages 28 and 50.

VANDIN, F.; UPFAL, E.; RAPHAEL, B. J. De novo discovery of mutated driver pathways in cancer. **Genome research**, Cold Spring Harbor Laboratory Press, v. 22, n. 2, p. 375–385, Feb. 2012. ISSN 1549-5469. Available: <http://dx.doi.org/10.1101/gr.120477.111>. Citations on pages 28, 45, and 51.

VINCENT, J.-L. The coming era of precision medicine for intensive care. **Critical Care**, Springer, v. 21, n. 3, p. 314, 2017. Citation on page 29.

VOGELSTEIN, B.; PAPADOPOULOS, N.; VELCULESCU, V. E.; ZHOU, S.; DIAZ, L. A.; KINZLER, K. W. Cancer genome landscapes. **Science**, American Association for the Advancement of Science, v. 339, n. 6127, p. 1546–1558, 2013. ISSN 0036-8075. Available: <http://science.sciencemag.org/content/339/6127/1546>. Citations on pages 27, 28, 39, and 44.

WEINBERG, R. **The Biology of Cancer, Second Edition**. Taylor & Francis Group, 2013. ISBN 9781317963462. Available: <https://books.google.com.br/books?id=MzMmAgAAQBAJ>. Citation on page 38.

WEINSTEIN, J. N.; COLLISSON, E. A.; MILLS, G. B.; SHAW, K. R. M.; OZENBERGER, B. A.; ELLROTT, K.; SHMULEVICH, I.; SANDER, C.; STUART, J. M.; NETWORK, C. G. A. R. *et al.* The cancer genome atlas pan-cancer analysis project. **Nature genetics**, Nature Publishing Group, v. 45, n. 10, p. 1113, 2013. Citations on pages 29 and 42.

WHO. **Cancer – (World Health Organization)**. 2021. Https://www.who.int/news-room/fact-sheets/detail/cancer. [Online; accessed August-2021]. Citation on page 57.

WILD, C. P.; STEWART, B. W.; WILD, C. **World cancer report 2014**. [S.l.]: World Health Organization Geneva, Switzerland, 2014. Citation on page 57.

WILSON, D. L. Asymptotic properties of nearest neighbor rules using edited data. **IEEE Transactions on Systems, Man, and Cybernetics**, IEEE, n. 3, p. 408–421, 1972. Citation on page 110.

WU, G.; HAW, R. Functional interaction network construction and analysis for disease discovery. In: **Protein Bioinformatics**. [S.l.]: Springer, 2017. p. 235–253. Citations on pages 44 and 64.

WU, J.; CAI, Q.; WANG, J.; LIAO, Y. Identifying mutated driver pathways in cancer by integrating multi-omics data. **Computational Biology and Chemistry**, v. 80, p. 159 – 167, 2019. ISSN 1476-9271. Citation on page 28.

YANG, L.; CHEN, R.; GOODISON, S.; SUN, Y. An efficient and effective method to identify significantly perturbed subnetworks in cancer. **Nature Computational Science**, Nature Publishing Group, v. 1, n. 1, p. 79–88, 2021. Citation on page 28.

YEANG, C.-H.; MCCORMICK, F.; LEVINE, A. Combinatorial patterns of somatic gene mutations in cancer. **The FASEB Journal**, v. 22, n. 8, p. 2605–2622, 2008. PMID: 18434431. Available: <https://doi.org/10.1096/fj.08-108985>. Citation on page 44.

ZHANG, J.; LIU, J.; SUN, J.; CHEN, C.; FOLTZ, G.; LIN, B. Identifying driver mutations from sequencing data of heterogeneous tumors in the era of personalized genome sequencing. **Briefings in Bioinformatics**, v. 15, n. 2, p. 244–255, 2014. Available: <http://dx.doi.org/10.1093/bib/bbt042>. Citations on pages 53 and 54.

ZHOU, L.; HUANG, W.; YU, H.-F.; FENG, Y.-J.; TENG, X. Exploring tcga database for identification of potential prognostic genes in stomach adenocarcinoma. **Cancer Cell International**, BioMed Central, v. 20, n. 1, p. 1–12, 2020. Citation on page 57.

ZHU, C.-Y.; ZHOU, C.; CHEN, Y.-Q.; SHEN, A.-Z.; GUO, Z.-M.; YANG, Z.-Y.; YE, X.-Y.; QU, S.; WEI, J.; LIU, Q. C3: Consensus cancer driver gene caller. **Genomics, Proteomics & Bioinformatics**, v. 17, n. 3, p. 311 – 318, 2019. ISSN 1672-0229. Citation on page 28.

ICMC USP
SÃO CARLOS