Investigation of the performance of driver mutation identification methods using biological networks and enriched biological networks

Alfredo Guilherme da Silva Souza[1], **Adenilso Simao**[1]

[1]Institute of Mathematics and Computer Sciences
University of Sao Paulo (ICMC-USP)
São Carlos – SP – Brazil

alfredo@usp.br, adenilso@icmc.usp.br

Abstract. *Several computational methods allow identifying genes related to cancer (driver mutation) through patient mutation data and biological networks. Usually, networks are not built focusing on biological activities associated with cancer because they are designed for general use. In this study, we investigate the performance of methods for identifying driver mutations using biological networks and enriched biological networks, applying a gene prioritization method to classify genes associated with cancer understudy in the biological network. The results indicated that employing the enrichment method helped identify different driver genes in all cases.*

1. Introduction

Cancer is a disease in a constant evolutionary process. Its birth and manifestation are related to mutations in a set of genes. Several mutations can occur in genes, but only a few are relevant to cancer development. These are classified as driver mutations, while non-significant mutations are classified as passenger mutations. These mutations can be hereditary or acquired during a person's lifetime. They can also vary in different sources, making cancer more difficult to understand; many variables are involved [Nussbaum et al. 2015].

Networks are generally used to represent complex biological systems. With them, it is possible to model a biological system according to the interactions of each element, be it a gene, protein, or other biomolecules [Ozturk et al. 2018]. In the literature, it is possible to find different biological networks, among them: KEEG [Kanehisa et al. 2009], Reactome [Joshi-Tope et al. 2005] [Croft et al. 2013], BioGRID [Chatr-Aryamontri et al. 2017], HPRD [Keshava Prasad et al. 2008], STRING [Szklarczyk et al. 2014], HINT [Das and Yu 2012], and others. Biological networks are extensively used in network-based driver gene identification methods. These networks are built and updated over time to represent interactions associated with various biological mechanisms, which allows us to consider that it has a general rather than a specific purpose. Thus, the network is not defined focusing only on the interactions associated with a type of disease but rather interactions representing various biological activities.

The interest in identifying and discovering new driver mutations has contributed to the development of computational methods that identify genes drivers, cancer genome databases creation and maintenance, and biological networks. These methods also have

the primary function of distinguishing between driver and passenger mutations. These methods use mutation data from patients (MAF) and biological networks as input data. Some also allow the use of gene expression data. These input data are used internally in the methods. Each method adopts a computational approach, which through the use of genes from frequently mutated patients extracted from the MAF associated with the network, can identify possible driver genes. The output of these methods is usually ordered and scored a list of possible driver genes [Hristov and Singh 2017].

One of the computational techniques used to identify possible associations between genes and diseases is gene prioritization. The main objective of the gene prioritization method is to classify a set of candidate genes using different algorithms, which can use data integration approaches from different databases, network-based approaches, and machine learning-based approaches. These methods generally allow the use of seeds gene lists, or disease selection or desired phenotype terms, to be used during the training process. They also allow the use of candidate gene lists or complete genomes, which are the genes to be prioritized. After sorting genes, the methods can generate lists with sorted and scored genes [Zolotareva and Kleine 2019]. Some methods allow the use of seeds gene lists for training, such as: Endeavour [Tranchevent et al. 2016], GeneMANIA [Mostafavi et al. 2008], GPS [Meshkin et al. 2019], MaxLink [Guala et al. 2014], pBRIT [Kumar et al. 2018], ToppGene [Chen et al. 2009], and ToppNet [Chen et al. 2009]). In contrast, in other methods, instead of seed gene lists, the user can define disease or phenotype terms as training data, such as: DisGeNET [Piñero et al. 2016], GLAD4U [Jourquin et al. 2012], OpenTargets [Koscielny et al. 2017], Polysearch [Liu et al. 2015], Phenolyzer [Yang et al. 2015], and PhenoRank [Cornish et al. 2018]. Various methods allow the use of candidate genes lists, such as: OpenTargets, Phenolyzer, Endeavour, GPS, pBRIT, ToppGene, and ToppNet. But, some methods only allow the use of a complete genome instead of a list of candidate genes, such as: DisGeNET, GLAD4U, Polysearch, PhenoRank, GeneMANIA, MaxLink [Raj and Sreeja 2018].

This study investigates whether there is an improvement in precision in the results of driver gene identification methods when using networks enriched with genes associated with each type of cancer compared to networks without enrichment. To enrich the genes in the network, we used prioritization methods to score each set of genes in the network according to each type of cancer, using genes from MAF.

For this study, the driver mutation identification method selected was the nCOP [Hristov and Singh 2017]. Because it allows us to assign weights to each gene in the network; it is easy to execute; and generates a list of potential driver genes, making it possible to compare them with genes in the literature. Many driver gene identification methods currently use complete biological networks (general-purpose). The nCop method by default uses the HPRD [Keshava Prasad et al. 2008] network, and this was the network adopted in this study.

We chose to use prioritization methods that simultaneously allow seed gene lists for training and candidate gene lists. Only the Endeavour, GPS, pBRIT, ToppGene, and ToppNet methods allow the use of both types of lists. We used this method category to prioritize a list of genes extracted from the network (candidate genes) from a list of genes extracted from MAF (seed genes). This choice made it possible to identify in the network which genes have a functional association with frequently mutated genes for each type of

cancer. After analyzing the methods, we selected Endeavour, because it is one of the most cited methods and compared with other methods in the literature.

We collected MAF for six cancer types and removed outliers from each file to carry out the investigation proposed in this work. We also extracted genes from MAF and the network used in the nCOP method. The extracted genes were used in Endeavor, where genes from the MAF were used as seed genes, and genes extracted from the network were used as candidate genes. We use the prioritization results to enrich the network. We performed simulations with nCOP using MAF for each type of cancer and the enriched and non-enriched networks. This study indicates that the enrichment of networks with prioritized genes helps the driver identification methods discover a more significant gene number recognized in the literature.

This paper is organized in the following order: Section 2 presents a computational approach to investigate the performance of driver mutation identification methods using biological networks and enriched biological networks. Next, Section 3 presents a comparative evaluation according to the literature, presenting the differences in performance using biological networks and enriched biological networks. Finally, Section 4 describes a summary of the results obtained in this study and discusses the contribution of the enrichment process in driver mutation identification methods.

2. Methods

This section describes the methodology adopted to carry out this study, from selecting data collection to the process adopted for the evaluation. Figure 1 graphically represents the steps of the approach adopted in this study, and these are also described in detail in the following subsections: 2.1 **Data collection (1):** We present the data that was collected to be used in the experiment; 2.2 **Data pre-processing (2)**: We describe what processes are performed on the data for removing mutation data outliers, converting the mutation file to nCOP, and extracting the list of genes with mutation frequency $\geq$ 3%, 5%, 10% for each type of cancer, to be used in Endeavour; 2.3 **Gene list prioritization (3)**: We detail the parameters, criteria considered, and data generated as input for the Endeavour to perform the prioritization in the candidate gene list, namely: list of genes for training stage, database, and selection of candidate genes; 2.4 **Simulation data preparation (4)**: We adopted a process for the conversion of the list of genes prioritized by Endeavour, in the nCOP weights file, for each experiment, as well as a description of how the network used in nCOP is structurally represented in the file; 2.5 **Simulation (5)**: We detail the mutation data and weight files generated in the experiments for each cancer type and mutation frequency used in each simulation; and 2.6 **Evaluation results (6)**: We adopt two different approaches to evaluate the results: an approach to calculate the precision of each experiment, for each type of cancer, according to a benchmark, and an approach to identify which genes are found only in enrichment, which is considered driver genes according to the literature.

2.1. Data collection

This subsection presents the mutation data samples selected for each cancer type used during the experiments and from which studies each sample was derived. A brief description of the network selected to be used in this study is also presented.

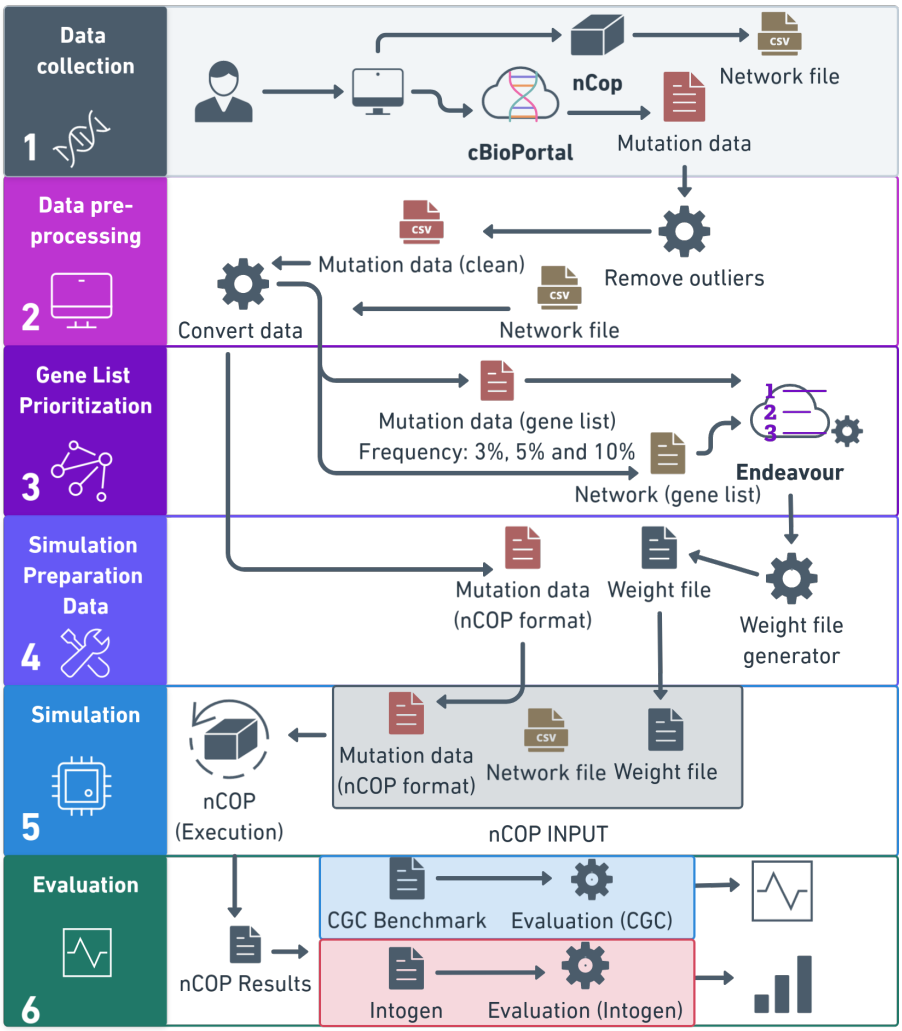

Figure 1. A presentation of the approach.

2.1.1. Mutation data and Biological Network

In this study, we choose types of cancer frequently used in several studies and considered quite popular. After an analysis of the literature, we selected mutation data for six types of cancer to be used as a case study: Bladder Cancer (BLCA) with 413 samples [Ciriello et al. 2015], Breast Invasive Carcinoma (BRCA) with 818 samples [Robertson et al. 2017], Glioblastoma (GBM) with 206 samples [Network et al. 2008], Pancreatic Adenocarcinoma (PAAD) with 184 samples, Prostate Adenocarcinoma (PRAD) with 334 samples [Abeshouse et al. 2015], and Stomach Adenocarcinoma (STAD) with 295 samples [Bass et al. 2014]. This data was captured in Mutation Annotation Format (MAF)[1] through cBioPortal[2] [Gao et al. 2013]. A MAF file is a tab-delimited text file that defines each type of information for each sample in each column. The `Hugo_Symbol` (gene name) and `Tumor Sample Barcode` (sample) fields are relevant and used in this study.

[1]https://docs.gdc.cancer.gov/Data/File_Formats/MAF_Format
[2]https://www.cbioportal.org/datasets

HPRD is a protein interaction network extracted from the human protein reference database, containing experimentally generated information with human proteome data. All data entered into HPRD goes through a curation process, which increases its reliability and avoids redundant protein insertion [Keshava Prasad et al. 2008]. HPRD is a network used by default in the nCOP method. The HPRD used in this study is a preprocessed version for the nCop method. The study [Hristov and Singh 2017] indicates that nodes with high connectivity have been removed, being nodes with degrees > 900 and 10 nodes with standard deviation outside the mean. Nine longer genes were also excluded, namely TTN, MUC16, SYNE1, NEB, MUC19, CCDC168, FSIP2, OBSCN, GPR98, as they had many mutations in extensive coverage of patients. The HPRD used has 9,379 nodes and 36,638 edges.

2.2. Data pre-processing

We extract all mutation data in MAF format, then pre-process to remove outliers and convert the data to standard nCOP and Endeavour input format. A method of removing hypermutated patients from the MAF, implemented by [Cutigi et al. 2020b], was used, according to the method proposed by [Tamborero et al. 2013]. We identified all samples that had a number of somatic mutations greater than $(Q3 + 4.5 \times IQR)$, where $Q3$ is the third quartile and the IQR is the interquartile range of the distribution of mutations in the MAF samples. To generate the mutation data for nCOP, we identified in the MAF which patients each gene appears mutated. We generate a list of all patients for each gene, then create a mutation file extracted from each MAF, according to the mutation data standard accepted by nCOP.

Endeavour allows the user to submit gene lists to the gene prioritization process: gene lists for training and candidate gene lists. In this study, we used MAF genes as a training set. However, we observed that the method has a computational limit, preventing sending an extensive list of genes, making it necessary to select a smaller list of genes to be used in the method. To define a smaller list of genes, we followed a long tail distribution, where few genes appear mutated in many patients, and many genes appear mutated in few patients. Based on the study of [Armenia et al. 2018], where it is indicated that many genes appear mutated in less than 3% of patients, we took as motivation not to select genes with a mutation rate lower than 3%.

In the literature, it is possible to find several approaches where genes with a high mutation rate, strongly associated with a type of cancer, are preferentially considered. Also, for this study, we chose to select genes with mutation frequency $\geq 3\%, 5\%, 10\%$, to identify whether, from genes that appear mutated very frequently in patients, it is possible through of them to identify possible new genes associated with each type of cancer. We generate lists of genes for each type of cancer and each mutation frequency to do this. We also extract all genes from the network and generate candidate genes from these genes.

2.3. Gene list prioritization

There are different methodologies used by gene prioritization methods, such as methods based on networks, methods based on data aggregation, and others. Endeavour is based on data aggregation. It uses sources of evidence to calculate the score for each gene, which indicates the probability that the gene is possibly responsible for the phenotype.

[Zolotareva and Kleine 2019]. The Endeavour gene prioritization method allows the execution of prioritization through four stages: 1) Species selection, 2) Selection of training genes list (seed genes), 3) Selection databases, and 4) Selection of candidate genes [Tranchevent et al. 2016]. In this study we work with the species Homo sapiens.

2.3.1. Selection of list of training genes (seed genes):

Endeavour uses the genes from the training gene list to train its model according to the biological processes associated with each seed gene [Tranchevent et al. 2016]. In this study, we prioritize the genes extracted from the network according to the potential genes associated with the biological activities of the genes in the training set. We extracted seed genes from the mutation data for each type of cancer to be used as a training genes list.

2.3.2. Databases selection:

In Endeavour, the databases are used to help identify possible other genes that share the same biological activities as the seed genes, those specified as a training set. In the method, for each species type, it is possible to select one or a set of different databases, ranging from pathway databases to pharmaceuticals databases. In this study, only the Reactome [Croft et al. 2013] database was used, which is a Pathways database. Reactome was chosen because it is widely used, accepted by the scientific community, updated regularly, and all its data goes through a rigorous curation process.

2.3.3. Selection of candidate genes:

In the context of gene prioritization, a candidate gene list is a list of genes that should be prioritized, according to the generated training model. Each gene can be prioritized according to biological activities, phenotypes, or other characteristics of interest. In Endeavour, biological activities are identified through selected databases. For each biological activity, it is possible to locate which seed genes are present and which other genes from the candidate list are also present and have great functional potential associated with each seed gene [Tranchevent et al. 2016]. This study defined a list of candidate genes from the extraction of genes of the biological network also used in nCOP. A reason for choosing this approach to define the list of candidates is to prioritize the genes in the network of the seed genes extracted from the mutation data for each type of cancer. For each type of cancer, the genes in the network are prioritized for the cancer being studied.

2.4. Simulation data preparation

We generate a weights file by joining the weights assigned to the nCOP file and the scores generated by Endeavour in the prioritization. From this process, we created a new weight file containing all the genes of the network used in nCOP, with weights updated according to the prioritization. We generate a weights file for each type of cancer and each type of experiment. The network data used in nCOP has not changed. The network is represented as a list of undirected edges. The weight file used in nCOP and the network have the same genes.

2.5. Simulation

For each type of cancer, we performed four simulations, one simulation used only the mutation data (complete) and the network, no gene prioritization. In comparison, the other three simulations used mutation data (complete), the network, and the prioritized weight file generated with a list of genes with mutation frequency $\geq 3\%, 5\%, 10\%$ in the source MAF file. In all, 24 simulations were performed with nCOP.

2.6. Evaluation

Identifying driver mutations generates a sorted and scored list, from the most relevant to the least relevant driver gene found for the cancer type studied. Still, some genes found by the method may be a driver gene, but not for cancer studied. Validations were done using studies from the literature that indicate whether the gene is a gene driver or not.

We divided the results evaluations into two stages: **1)** We calculated the precision of the results using a benchmark to verify the experiments' performance to find significant genes for each type of cancer. In this case, we use the benchmark because it is constantly updated and well consolidated, CGC [3]. **2)** We evaluated whether genes found only with enrichment are driver genes for the type of cancer in the experiment and how often they appeared in samples obtained from intOGen [4] for cancer studied. The results of the evaluation process are described in the next section.

3. Results

To evaluate the results, we adopted a precision calculation method by the literature reference to identify and compare the precision between the types of experiments performed for each type of cancer. We performed the precision calculation according to the presence or absence of genes in the Cancer Gene Sensus (CGC). The following calculation defines a precision: $Precision = nGiB/(nGiB + nGoB)$, where $nGiB$ represents the number of genes found in the benchmark and $nGoB$ represents the number of genes not found. This evaluation model was proposed by [Cutigi et al. 2020a]. At this stage of the evaluation, we analyzed the three types of experiments individually compared to the no-enrichment experiment for each type of cancer. In this step, we had the objective of calculating the precision of each experiment to identify in which situations the experiments carried out with enrichment have greater precision than the experiments without.

We observed that the number of genes found without enrichment is smaller than those found with enrichment in some experiments. This fact limits the validation accuracy to the limit of the number of genes available. The results of the evaluation are shown in Figure 2, where you can see the lines that indicate the precision for each type of experiment. In some cases, enrichment experiments obtained greater accuracy than without enrichment. In other cases, lower performance. However, the enrichment process helped discover more genes in all cases.

To validate that each enrichment-only gene is a significant gene for the type of cancer, we collected data from intOGen samples for each experiment, which allowed us to identify which genes appear in the samples and how often. For this validation step, we

[3] https://cancer.sanger.ac.uk/census

[4] https://www.intOGen.org/

selected data sets from six types of Cancer in intOGen, namely: BLCA (411 samples), BRCA (973 samples), GBM (391 samples), PAAD (176 samples), PRAD (492 samples), and STAD (436 samples). All datasets selected in intOGen come from the TCGA.

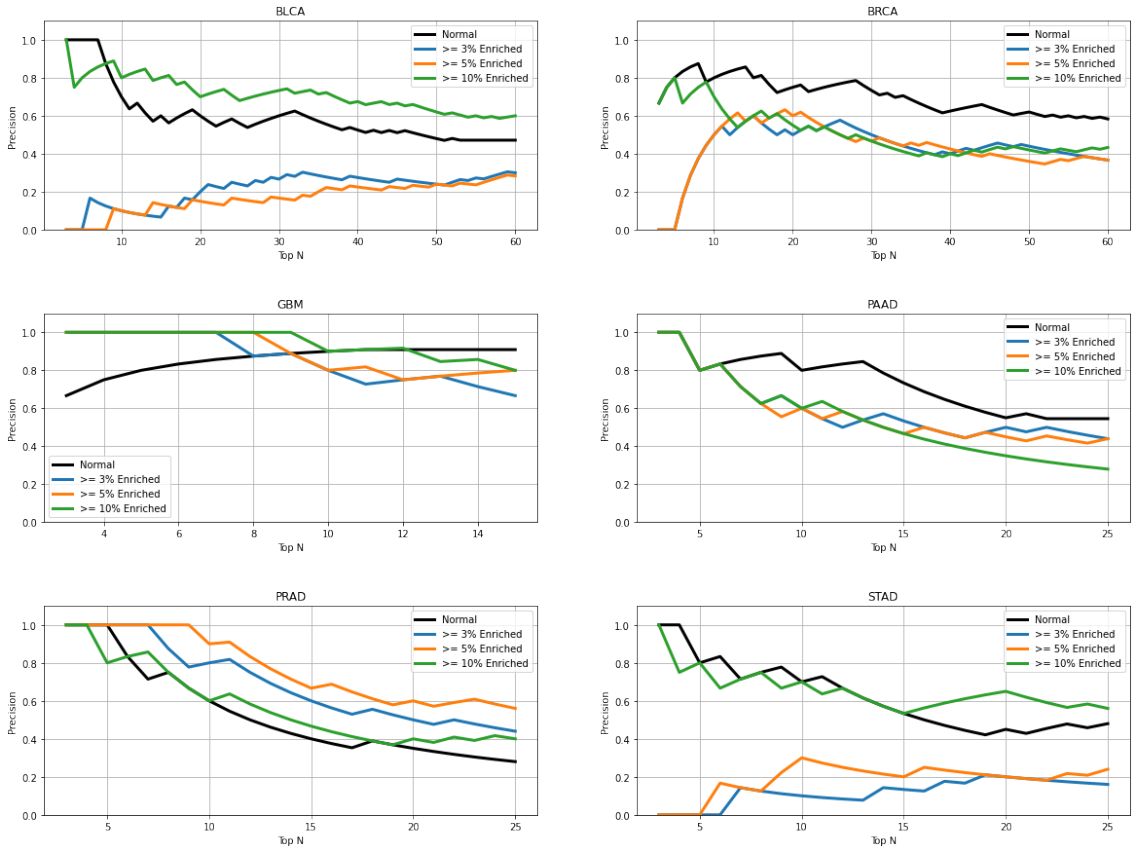

Figure 2. Comparison of precision between results. No enrichment (normal), and gene enrichment with mutation frequency $\geq$ 3%, 5%, 10%.

We obtained favorable results in all three types of experiments at this validation stage, experiment with enrichment using genes with mutation frequency $\geq$ 3%, 5%, 10%. In all three types of experiments, we found the same driver genes in each cancer analyzed, such as: **BLCA**: PIK3CA (19.95%), ERBB2 (11.92%), ERBB3 (9.98%), TSC1 (6.33%), RHOA (4.62%), HRAS (4.14%), KRAS (3.89%), PTEN (3.16%), PIK3CB (2.19%), RAF1(0.97%); **BRCA**: RB1 (2.26%), CTCF (2.16%), CREBBP (1.44%), FBXW7 (1.34%), SMAD4 (0.51%); **GBM**: PTPN11 (2.81%), NRAS (1.02%); **PAAD**: GNAS (4.55%), TGFBR2 (3.98%), PIK3CA (2.84%), U2AF1 (1.14%); **PRAD**: LRP1B (5.08%), KDM6A (1.22%); and **STAD**: KRAS (5.73%), PTEN (3.21%), ERBB2 (2.98%), SDC4 (0.92%). These results are graphically represented in the Figure 3.

4. Conclusion and Discussion

This study investigated the performance of driver mutation identification methods using biological networks and enriched networks. We selected two computational methods for the experiment stage: 1) nCOP as a driver mutation identification method and 2) Endeavour as a gene prioritization method from training gene set, pathways database, and candidate gene set. As data set were MAFs for 6 cancer types (BLCA, BRCA, GBM, PAAD, PRAD, and STAD), and a biological functional interaction network (HPRD).

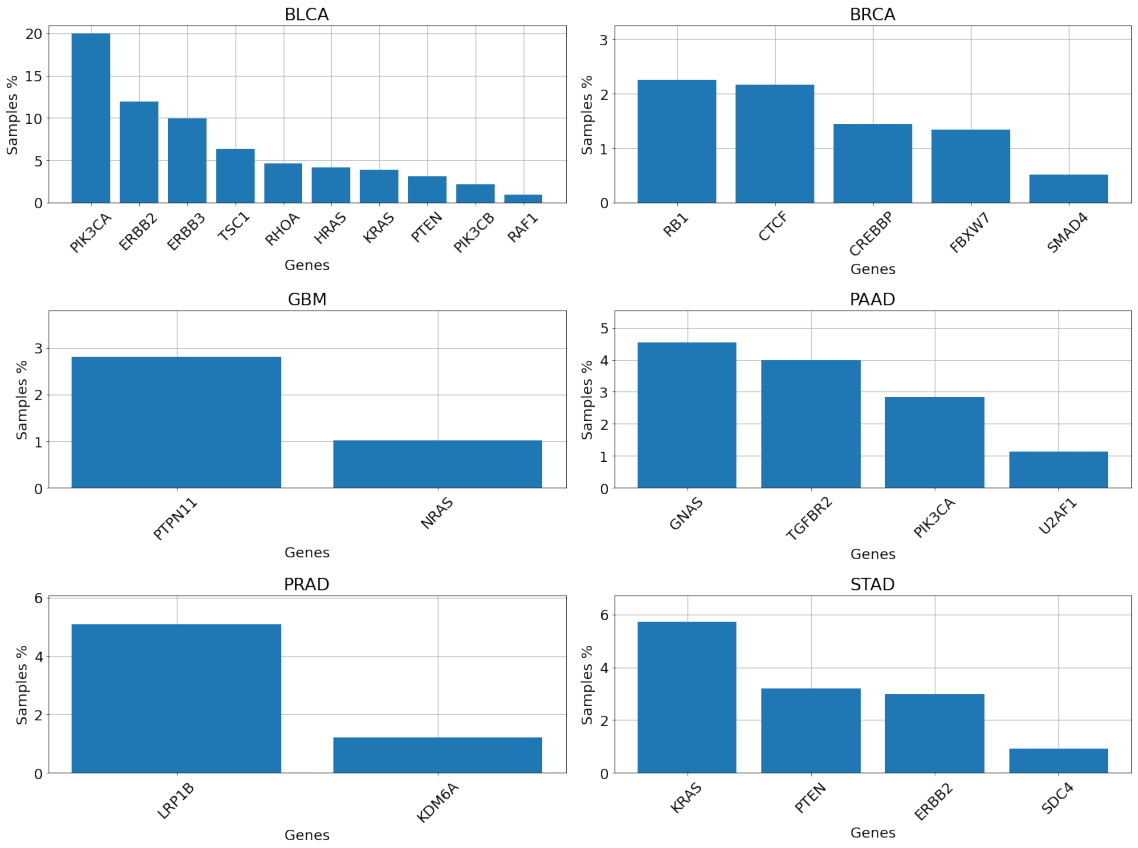

Figure 3. Genes present in the intOGen.

We used nCop as a driver gene identification method in this study and Endeavour as a gene prioritization method. We selected nCop because it allows us to assign weights to each gene in the network, but the same experiment can be performed using other methods to assign weights to the genes in the network.

We note Endeavour has computational limitations regarding the size of the data input, which would make it impossible to use all the genes from the MAF's, as some have more than 15,000 genes. Therefore, we extracted from each MAF only genes with mutation frequency greater than or equal to 3%, 5%, and 10%. We adopted this process to remove from the samples only genes with low mutation frequency and analyze whether even genes with a higher mutation rate allow the discovery of more driver genes in conjunction with nCOP.

The results indicate that even for the experiment with enrichment that obtained a lower precision than without enrichment, it was possible to discover a more significant number of driver genes recognized by the literature. In all cases, the results were positive concerning the number of driver genes discovered when we used enrichment in the experiment.

For future work, we will identify if the genes discovered with enrichment and that by the literature are not yet recognized as drivers, if they are cited in recent studies as a driver, oncogene, or tumor suppressor. This future work may indicate that possible genes found with enrichment are not recognized as drivers, but studies are emerging on them

that indicate their association with cancer.

Acknowledgments

The authors acknowledge Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) for their financial support for the conclusion of this study.

References

Abeshouse, A., Ahn, J., Akbani, R., Ally, A., Amin, S., Andry, C. D., Annala, M., Aprikian, A., Armenia, J., Arora, A., et al. (2015). The molecular taxonomy of primary prostate cancer. *Cell*, 163(4):1011–1025.

Armenia, J., Wankowicz, S. A., Liu, D., Gao, J., Kundra, R., Reznik, E., Chatila, W. K., Chakravarty, D., Han, G. C., Coleman, I., et al. (2018). The long tail of oncogenic drivers in prostate cancer. *Nature genetics*, 50(5):645–651.

Bass, A. J., Thorsson, V., Shmulevich, I., Reynolds, S. M., Miller, M., Bernard, B., Hinoue, T., Laird, P. W., Curtis, C., Shen, H., et al. (2014). Comprehensive molecular characterization of gastric adenocarcinoma. *Nature*, 513(7517):202.

Chatr-Aryamontri, A., Oughtred, R., Boucher, L., Rust, J., Chang, C., Kolas, N. K., O'Donnell, L., Oster, S., Theesfeld, C., Sellam, A., et al. (2017). The biogrid interaction database: 2017 update. *Nucleic acids research*, 45(D1):D369–D379.

Chen, J., Bardes, E. E., Aronow, B. J., and Jegga, A. G. (2009). Toppgene suite for gene list enrichment analysis and candidate gene prioritization. *Nucleic acids research*, 37(suppl_2):W305–W311.

Ciriello, G., Gatza, M. L., Beck, A. H., Wilkerson, M. D., Rhie, S. K., Pastore, A., Zhang, H., McLellan, M., Yau, C., Kandoth, C., et al. (2015). Comprehensive molecular portraits of invasive lobular breast cancer. *Cell*, 163(2):506–519.

Cornish, A. J., David, A., and Sternberg, M. J. (2018). Phenorank: reducing study bias in gene prioritization through simulation. *Bioinformatics*, 34(12):2087–2095.

Croft, D., Mundo, A. F., Haw, R., Milacic, M., Weiser, J., Wu, G., Caudy, M., Garapati, P., Gillespie, M., Kamdar, M. R., et al. (2013). The reactome pathway knowledgebase. *Nucleic acids research*, 42(D1):D472–D477.

Cutigi, J. F., Evangelista, A. F., and Simao, A. (2020a). Genwemme: a network-based computational method for prioritizing groups of significant related genes in cancer. In *Advances in Bioinformatics and Computational Biology: 12th Brazilian Symposium on Bioinformatics, BSB 2019, Fortaleza, Brazil, October 7–10, 2019, Revised Selected Papers*, volume 11347, page 29. Springer Nature.

Cutigi, J. F., Evangelista, R. F., Ramos, R. H., Ferreira, C. d. O. L., Evangelista, A. F., de Carvalho, A. C., and Simao, A. (2020b). Combining mutation and gene network data in a machine learning approach for false-positive cancer driver gene discovery. In *Brazilian Symposium on Bioinformatics*, pages 81–92. Springer.

Das, J. and Yu, H. (2012). Hint: High-quality protein interactomes and their applications in understanding human disease. *BMC systems biology*, 6(1):92.

Gao, J., Aksoy, B. A., Dogrusoz, U., Dresdner, G., Gross, B., Sumer, S. O., Sun, Y., Jacobsen, A., Sinha, R., Larsson, E., et al. (2013). Integrative analysis of complex cancer genomics and clinical profiles using the cbioportal. *Science signaling*, 6(269):pl1–pl1.

Guala, D., Sjölund, E., and Sonnhammer, E. L. (2014). Maxlink: network-based prioritization of genes tightly linked to a disease seed set. *Bioinformatics*, 30(18):2689–2690.

Hristov, B. H. and Singh, M. (2017). Network-based coverage of mutational profiles reveals cancer genes. *Cell systems*, 5(3):221–229.

Joshi-Tope, G., Gillespie, M., Vastrik, I., D'Eustachio, P., Schmidt, E., de Bono, B., Jassal, B., Gopinath, G., Wu, G., Matthews, L., et al. (2005). Reactome: a knowledgebase of biological pathways. *Nucleic acids research*, 33(suppl_1):D428–D432.

Jourquin, J., Duncan, D., Shi, Z., and Zhang, B. (2012). Glad4u: deriving and prioritizing gene lists from pubmed literature. *BMC genomics*, 13(8):1–12.

Kanehisa, M., Goto, S., Furumichi, M., Tanabe, M., and Hirakawa, M. (2009). Kegg for representation and analysis of molecular networks involving diseases and drugs. *Nucleic acids research*, 38(suppl_1):D355–D360.

Keshava Prasad, T., Goel, R., Kandasamy, K., Keerthikumar, S., Kumar, S., Mathivanan, S., Telikicherla, D., Raju, R., Shafreen, B., Venugopal, A., et al. (2008). Human protein reference database—2009 update. *Nucleic acids research*, 37(suppl_1):D767–D772.

Koscielny, G., An, P., Carvalho-Silva, D., Cham, J. A., Fumis, L., Gasparyan, R., Hasan, S., Karamanis, N., Maguire, M., Papa, E., et al. (2017). Open targets: a platform for therapeutic target identification and validation. *Nucleic acids research*, 45(D1):D985–D994.

Kumar, A. A., Van Laer, L., Alaerts, M., Ardeshirdavani, A., Moreau, Y., Laukens, K., Loeys, B., and Vandeweyer, G. (2018). pbrit: gene prioritization by correlating functional and phenotypic annotations through integrative data fusion. *Bioinformatics*, 34(13):2254–2262.

Liu, Y., Liang, Y., and Wishart, D. (2015). Polysearch2: a significantly improved text-mining system for discovering associations between human diseases, genes, drugs, metabolites, toxins and more. *Nucleic acids research*, 43(W1):W535–W542.

Meshkin, A., Shakery, A., and Masoudi-Nejad, A. (2019). Gps: Identification of disease genes by rank aggregation of multi-genomic scoring schemes. *Genomics*, 111(4):612–618.

Mostafavi, S., Ray, D., Warde-Farley, D., Grouios, C., and Morris, Q. (2008). Genemania: a real-time multiple association network integration algorithm for predicting gene function. *Genome biology*, 9(1):1–15.

Network, C. G. A. T. R. et al. (2008). Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature*, 455(7216):1061.

Nussbaum, R. L., McInnes, R. R., and Willard, H. F. (2015). *Thompson & Thompson genetics in medicine e-book*. Elsevier Health Sciences.

Ozturk, K., Dow, M., Carlin, D. E., Bejar, R., and Carter, H. (2018). The emerging potential for network analysis to inform precision cancer medicine. *Journal of molecular biology*, 430(18):2875–2899.

Piñero, J., Bravo, À., Queralt-Rosinach, N., Gutiérrez-Sacristán, A., Deu-Pons, J., Centeno, E., García-García, J., Sanz, F., and Furlong, L. I. (2016). Disgenet: a comprehensive platform integrating information on human disease-associated genes and variants. *Nucleic acids research*, page gkw943.

Raj, M. R. and Sreeja, A. (2018). Analysis of computational gene prioritization approaches. *Procedia computer science*, 143:395–410.

Robertson, A. G., Kim, J., Al-Ahmadie, H., Bellmunt, J., Guo, G., Cherniack, A. D., Hinoue, T., Laird, P. W., Hoadley, K. A., Akbani, R., et al. (2017). Comprehensive molecular characterization of muscle-invasive bladder cancer. *Cell*, 171(3):540–556.

Szklarczyk, D., Franceschini, A., Wyder, S., Forslund, K., Heller, D., Huerta-Cepas, J., Simonovic, M., Roth, A., Santos, A., Tsafou, K. P., et al. (2014). String v10: protein–protein interaction networks, integrated over the tree of life. *Nucleic acids research*, 43(D1):D447–D452.

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., et al. (2013). Comprehensive identification of mutational cancer driver genes across 12 tumor types. *Scientific reports*, 3(1):1–10.

Tranchevent, L.-C., Ardeshirdavani, A., ElShal, S., Alcaide, D., Aerts, J., Auboeuf, D., and Moreau, Y. (2016). Candidate gene prioritization with endeavour. *Nucleic acids research*, 44(W1):W117–W121.

Yang, H., Robinson, P. N., and Wang, K. (2015). Phenolyzer: phenotype-based prioritization of candidate genes for human diseases. *Nature methods*, 12(9):841–843.

Zolotareva, O. and Kleine, M. (2019). A survey of gene prioritization tools for mendelian and complex human diseases. *Journal of integrative bioinformatics*, 16(4).