Analyzing different cancer mutation data sets from breast invasive carcinoma (BRCA), lung adenocarcinoma (LUAD), and prostate adenocarcinoma (PRAD)

Rodrigo Henrique Ramos[1], **Jorge Francisco Cutigi**[1,2], **Cynthia de Oliveira Lage Ferreira**[2], **Adriane Feijó Evangelista**[3], **Adenilso Simão**[2]

[1]Instituto Federal de Educação, Ciência e Tecnologia de São Paulo – IFSP São Carlos
São Carlos – SP – Brasil

[2]Instituto de Ciências Matemáticas e de Computação – Universidade de São Paulo
São Carlos – SP – Brasil

[3]Centro de Pesquisa em Oncologia Molecular – Hospital de Câncer de Barretos
Barretos – SP – Brasil

ramos@ifsp.edu.br

Abstract. *With the advancements of next-generation sequencing (NGS) technologies, a massive volume of genetic data has been generated. It makes possible the study of complex disease by computational approaches. In the context of cancer, there is a huge variety of mutation data in public databases. However, it is not feasible to use all available data in every analysis; thus, a data subset must be selected. This work is aiming to investigate and understand the mutational characteristics presented in different cancer mutation data sets of the same type of cancer. To achieve this goal, exploration and visualization of cancer mutation data were performed. Several analyses are presented for three common types of cancer: 1) Breast Invasive Carcinoma (BRCA); 2) Lung Adenocarcinoma (LUAD); and Prostate Adenocarcinoma (PRAD). For each cancer type, three distinct data sets were analyzed in order to understand if there are significant differences or similarities among them. The analyses show that BRCA and LUAD have evidence of similarity among their data sets, while PRAD is likely heterogeneous.*

1. Introduction

Cancer is a complex disease characterized by genetic mutations that happen in a cell and lead to uncontrolled growth and division. The study of such mutations may contribute to understand the disease's early phase and evolution, which enables personalized therapies. With the advent of next-generation sequencing (NGS) technologies, a large number of DNA sequencing has been generated [Demkow and Ploski 2015]. Several databases were created as a result of NGS, such as "The Cancer Genome Atlas" (TCGA) and "International Cancer Genome Consortium" (ICGC). Based on these databases, cancer mutation data sets have been widely used by researchers to study mutations in cancer, genomic instability, and tumor evolution. Such studies are performed by computational methods that load and analyze NGS data.

With the massive volume of data, the use of all available mutation data for each type of cancer in analyses is not practicable. In this context, it is desirable to select a

reduced number of cancer studies of the same type, that could significantly represent the whole set of available mutation data of its type. Furthermore, using smaller data sets may reduce the computational time of the analysis.

The goal of this work is to perform data exploration and visualization among different cancer mutation data sets from the same type of cancer. It is expected with this study to contribute with analyses that enable a better understanding of mutational characteristics and specificities of different types of cancer. One of the main motivations for carrying out this study is to investigate whether a single data set is enough for making consistent conclusions about the mutational characteristic of the analyzed cancer type. In this sense, we selected three of the most common types of cancer: 1) Breast Invasive Carcinoma (BRCA); 2) Lung Adenocarcinoma (LUAD); and Prostate Adenocarcinoma (PRAD). For each type of cancer, we chose three data sets. For each data set, we performed some analyses aiming to explore and compare the data sets. A set of visual analyses were used to understand the similarity among the data sets, including a visualization of the breast data sets following an approach of topological data analysis (TDA) [Chazal and Michel 2017]. The analyses show that BRCA and LUAD present similarity among their three data sets, while the data sets of PRAD were understood as heterogeneous.

The remaining of this paper is organized as follows. The process we used to work with the data, the data sets, and the preprocessing routine are presented in Section 2. The data analyses exploration and visualization are presented in Section 3, accompanied by discussions about the exploratory studies. Next, in Section 4 is presented the final considerations of this work, conclusions, limitations and possible future works.

2. Cancer mutation data

To develop this work, we defined a simple process to perform the analysis. In Figure 1 is presented the process conducted in this work. We choose different types of cancer, then three data sets for each type were selected. Second, we apply a preprocessing routine to all data sets. Finally, several analyses were performed in the preprocessed data sets.

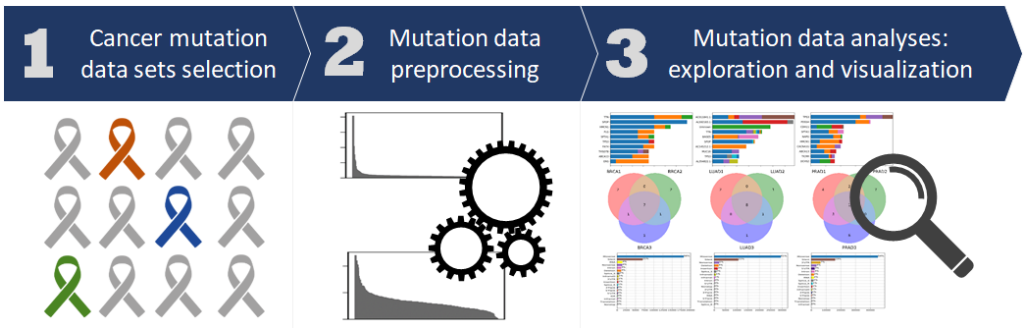

Figure 1. Process followed in this work.

2.1. Data sets

Cancer is the result of several mutations that occurs during a person's life, such as errors in DNA replication and environmental exposures. Such mutations in cancer occur in different scales: 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.

Public repositories of cancer mutation data have been created and continuously updated. Thus, the scientific community has fast and easy access to a large variety of cancer data. These repositories have as their primary goal to provide real data to support researches involving cancer and its characteristics and behaviors. Among the existing repositories, it can be cited "The Cancer Genome Atlas" (TCGA), which is a project created by the efforts of research entities to provide information on genetic mutations for various types of cancer. So far, this project has already generated and stored genomic mutations of 33 cancer types in 11,000 patients, holding approximately 2.5 petabytes of data [TCGA 2020]. Other platforms also have cancer data sets and provide essential tools for analyzing and visualizing mutation data. For example, the cBioPortal is an interactive platform for the exploration of cancer data [Cerami et al. 2012, Gao et al. 2013].

For this study, we worked with cancer mutation data of three types of cancer: 1) Breast Invasive Carcinoma (BRCA); 2) Lung Adenocarcinoma (LUAD); and Prostate Adenocarcinoma (PRAD). These types are among the most common cancers, according to the World Health Organization [WHO 2018]. We selected data sets with two types of mutations: 1) Single Nucleotide Variants (SNVs); and 2) Insertions and Deletions (InDels). For each type of cancer, we selected three distinct data sets, which were extracted from CBioPortal. The selected data sets and their references are presented in Table 1.

Table 1. Selected data sets and their references

	Data set A	Data set B	Data set C
BRCA	TCGA, Nature 2012 [Koboldt et al. 2012] id: BRCA1	TCGA, Cell 2015 [Ciriello et al. 2015] id: BRCA2	TCGA, Cell 2018 [TCGA 2018] id: BRCA3
LUAD	Imielinksi et al. Cell 2012 [Imielinski et al. 2012] id: LUAD1	TCGA, Nature 2014 [Collisson et al. 2014] id: LUAD2	TCGA, Cell 2018 [TCGA 2018] id: LUAD3
PRAD	Barbieri et al. Nat Genet 2012 [Barbieri et al. 2012] id: PRAD1	TCGA, Cell 2015 [Abeshouse et al. 2015] id: PRAD2	Kumar et al. Nat Med 2016 [Kumar et al. 2016] id: PRAD3

For each data set, SNVs and InDels mutation data were extracted, in which the data are contained in one mutation file in a format called MAF file (Mutation Annotation Format)[1]. We assign an ID for each file in order to make the identification easier. In Table 1 is also presented the ID of each data set.

2.2. Preprocessing

The selected data sets were submitted for a simple preprocessing routine. We remove hypermutated samples from the datasets because they are usually outliers and may biasing the analysis.

Several strategies can be used to remove the hypermutated samples. For this work, we used a strategy proposed by [Tamborero et al. 2013], in which the authors considered a hypermutated sample 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 samples of the data. In Figure 2 is presented the distribution of the mutations before and after such a preprocessing task. The number of samples, before and after removal, is also presented on the Y-axis of each chart.

[1] https://docs.gdc.cancer.gov/Data/File_Formats/MAF_Format/

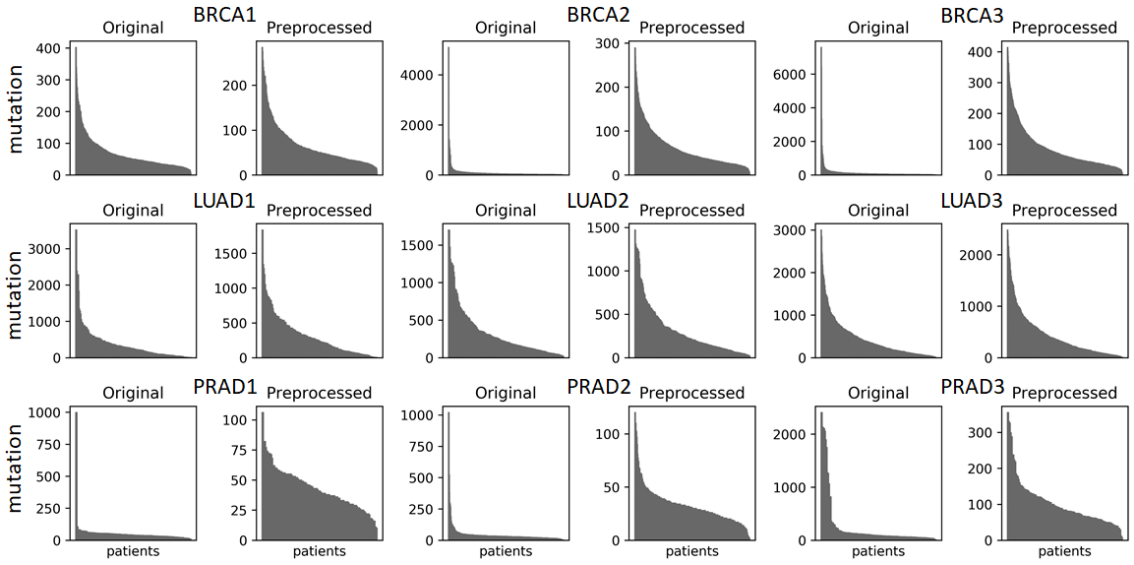

Figure 2. Preprocessing routine: Removing hypermutated samples.

For example, considering the data set PRAD1, before the data preprocessing, the most mutated patient had nearly 1000 mutations. In contrast, after removing hypermutated samples, the most mutated patient is represented by approximately 100 mutations. It can be noticed that after the preprocessing routine, the number of mutations is better distributed among the patients of each data set. In Table 2 is presented some metrics about the data sets, before and after the preprocessing routine.

Table 2. Data set characteristics

	Original data sets			Preprocessed data sets		
id	#mutations	#mutated genes	#samples	#mutations	#mutated genes	#samples
BRCA1	33990	13415	507	32627	13169	503
BRCA2	69968	16178	817	48085	14616	795
BRCA3	130495	18794	1009	83258	17588	978
LUAD1	65767	14770	183	55316	14012	179
LUAD2	72566	15132	230	69176	14977	228
LUAD3	243229	18905	562	237387	18868	560
PRAD1	5764	4298	112	4767	3682	111
PRAD2	14045	8176	333	10546	6679	322
PRAD3	32764	10412	141	13069	5484	129

It is important to mention that no genes were removed from the data set. Genes FLAGS [Shyr et al. 2014] were kept in the analyses, once we intended to show original characteristics of the data, thus minimizing significant modifications.

3. Mutation data analyses: exploration and visualization

A set of analyses were performed using the preprocessed data. Aiming to explore and compare each data set, a series of visual analyses were used to understand the similarity among the data sets, including a visualization of the breast data sets following an approach of topological data analysis (TDA).

3.1. Distribution of the classification of mutation

Each mutation may belong to several classes. In this analysis, we studied how is the distribution of each mutation class in the data sets, in which is presented in Figure 3. It can be noticed that *missense* is the most common mutation class in all data sets, presenting around 60% in total, thus being more than double the second position.

The distribution of the mutation class is similar, considering BRCA and LUAD. Such similarity can be seen in the data sets of the same type of cancer and across all six data sets. On the other hand, the data set PRAD3 does not present an obvious similarity; for example, the second most frequent mutation class is intron, while in the other data sets, the second most common is silent.

Figure 3. Distribution of mutation class in each data set.

3.2. Distribution of SNV classes

There are six SNVs classes, which represent a single change in a nucleotide. Such classes are C>A, C>G, C>T, T>A, T>C, T>G. In Figure 4 is presented the number of SNVs

classes in each data set. It can be noticed that data sets from the same type of cancer have a high similarity in the distribution of SNVs classes. Such observation has a strong relation with mutational signatures, which are combinations of mutations generated by different mutational processes [COSMIC 2019]. For example, classes C>T and C>A are associated with ultraviolet light exposure and tobacco smoking, respectively.

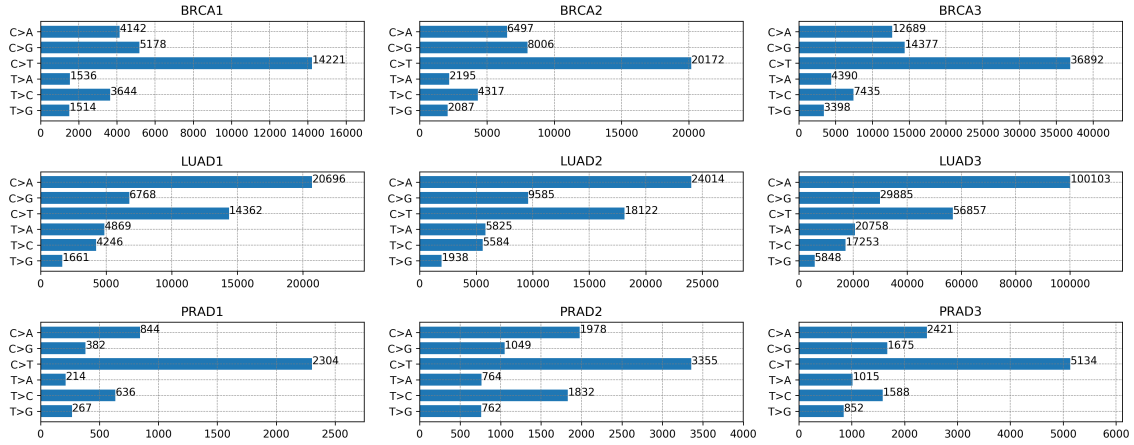

Figure 4. Number of mutations considering only SNVs.

3.3. Exploration of mutated genes

To understand the mutational characteristics of each data set, we performed some analyses on mutated genes.

First, a coefficient of similarity is calculated to understand how similar are two sets of mutated genes, considering different data sets. For this study, the Jaccard coefficient was chosen to determine how similar are two sets of genes. The Jaccard coefficient is defined as follows: $j(s_i, s_j) = \frac{|s_i \cap s_j|}{|s_i \cup s_j|}$, where s_i and s_j are two sets of genes from different data sets. In Table 3 is presented the Jaccard coefficient for all pair of data sets, in which yellow cell represents coefficients in the same type of cancer.

It can be noticed that LUAD3 and BRCA3 present a high similarity coefficient. It occurs because both data sets have almost all known genes in their mutated genes set. Considering the PRAD data sets, there are few mutated genes, but the similarity among them is lower in comparison on BRCA and LUAD. It can show the heterogeneity of PRAD.

Table 3. Jaccard coefficient for each pair of data sets.

	BRCA1	BRCA2	BRCA3	LUAD1	LUAD2	LUAD3	PRAD1	PRAD2	PRAD3
BRCA1		0.69	0.66	0.53	0.60	0.61	0.21	0.34	0.26
BRCA2			0.77	0.60	0.68	0.72	0.21	0.36	0.27
BRCA3				0.63	0.73	0.84	0.20	0.34	0.26
LUAD1					0.65	0.65	0.21	0.34	0.26
LUAD2						0.78	0.22	0.36	0.27
LUAD3							0.19	0.34	0.25
PRAD1								0.21	0.18
PRAD2									0.22
PRAD3									

Second, the top 10 mutated genes were analyzed in each data set. In the charts of Figure 5 we presented the top 10 analyses, which is also presented the distribution of mutation type. Genes within top 10 of all data sets are marked as bold. It can be noticed that BRCA data sets have seven common genes in the top 10, while in LUAD, such number is eight. On the other hand, in PRAD, only one gene is common in data sets. It can be seen as a heterogeneity among the PRAD data sets, as shown previously in Table 3.

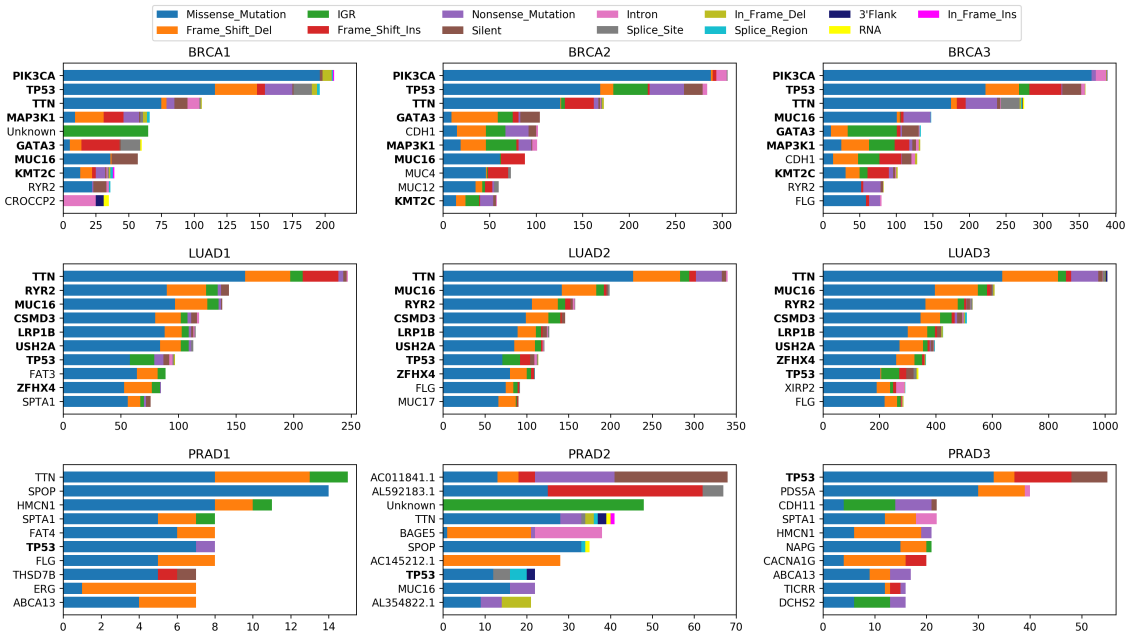

Figure 5. Top 10 analysis.

3.4. Identification of driver mutations

Somatic mutations in cancer can be classified into two types: 1) Driver Mutations: significant mutations for cancer, i.e., they confer cells the advantage of growing uncontrollably, thus promoting the cancer development; and 2) Passenger Mutations: they do not alter the natural behavior of cells, i.e., they are not significant for cancer. The identification of driver mutations in cancer is a challenge in cancer genomics since a single cancer cell usually undergoes a large number of mutations, which comprehend few drivers and many passenger mutations.

In this context, many computational methods have been developed for the prioritization of driver mutations in cancer. Such methods usually output a gene ranking, where the top genes in the ranking are most likely to be a driver mutation. In this experiment, three methods for the identification of driver mutations were selected, as follows:

1. MutSigCV [Lawrence et al. 2013]: executed using only the MAF file from the data sets.
2. MUFFINN [Cho et al. 2016]: it was considered results from the DNmax approach using the String gene network. As gene score was considered the number of mutations in each gene.
3. nCOP [Hristov and Singh 2017]: it was considered the results using the HPRD gene network.

Such methods were executed considering all mutation data sets. From the result gene ranking of each method, we selected the top 50 genes. Then, we compare the result among data sets of the same type of cancer considering the overlap of genes, thus creating a Venn diagram for each method. In Figure 6 is presented the result of the experiment.

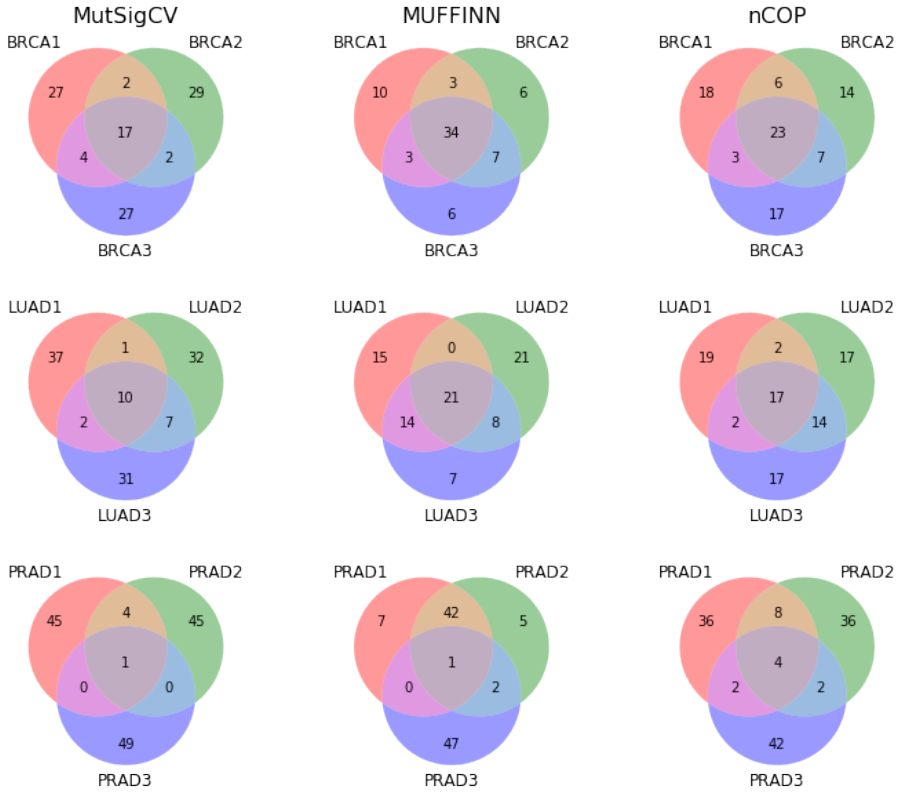

Figure 6. Venn diagram for prioritized genes in three computational methods.

It can be noticed that PRAD data sets presented a high difference considering all the computational methods, thus reinforcing that data from PRAD data sets are heterogeneous. Considering BRCA and LUAD, the previous analyses showed there are similarities in their data sets, but on the identification of driver mutation, the results do not have high similarity. It occurs because the computational methods present different approaches and take into consideration external information for prioritizing genes, such as gene networks.

3.5. Network visualization

The data sets studied were converted into a point cloud in a high-dimensional space in which each point represents a gene. Clinical information from patients was considered for this purpose. Considering the nature of the problem, we normalized the data in a scale from 0 to 100, then we performed this analysis only for breast cancer data. In this sense, we considered each mutated gene as a point in R^6 with coordinates: the number of patients with mutation in this gene, the average age of diagnosis, the percentage of survival, sex, chromosome and the average position of the mutation in the genome of all samples that present such mutated gene. The chromosome and the average position of the mutation in the genome were included as coordinates to prevent two or more genes being represented by the same point. Those coordinates, therefore, work as a unique geometric attribute for each gene.

Visualizing data on a large scale presents several difficulties and challenges. In particular, it is very difficult to deal with noise or with the vast size of the data set. Most dimension reduction maps, such as the Principal Component Analysis (PCA), end up generating a loss of information. In this context, we chose the Mapper algorithm [Singh et al. 2007] to obtain a visualization that describes the topological characteristics of the data.

The Mapper algorithm has been developed to identify topological features and provides a simple and convenient way to view a summary of data sets based on the distance between data points after applying a filter function. The main idea of the algorithm is, given a data set $X \in R^n$, and a function $f : X \to R^d$, to summarize X through a network constructed from $f^{-1}(U_i)$ in which U_i define a coverage of $f(X)$. Originally, it was applied to extract descriptors for the recognition of 3D objects [Singh et al. 2007]. In Biology, the Mapper algorithm was applied to study the structure of the space of expression of breast tumors [Nicolau et al. 2011] and to explore the relationships between genetic pathways and their association with brain function [Patania et al. 2019].

To perform the visualizations of this study, we used the Kepler-Mapper package [Veen and Saul 2017] developed in Python. The Mapper algorithm involves the choice of several parameters in the construction of the network that best describes the topology of the data. For this study, we considered as filter function the projection of data in the two principal components. We divided each principal component by 20, getting 400 windows and considered the percentage of overlap of these windows to be 50%. The K-Means algorithm was chosen to separate 5 clusters of the inverse image from the filter function of each window.

The networks obtained with the data sets related to breast cancer, considering the parameters described above, are presented in Figure 7. They have a similar degree of distribution, as we can see in Figure 8. Also, we can see in Table 4 that the networks obtained have very similar topological measures. For example, the average coefficient clustering (cc), a measure relative to how complete the neighborhood of a node is, refers to the count of triangles in this neighborhood. A triangle means that a set of three nodes are connected. In the three networks, this measure is around 0.5. Other measures, such as the number of connected components (CoC), diameter (d), shortest path length (sp), the average number of neighbors (ng), number of nodes (N) and number of nodes of the two largest connected components ($Nc1$, $Nc2$) are quite similar.

Finally, we would like to highlight the intersection between the three networks, considering the largest connected component in relation to the total number of genes in each study. The BRCA1, BRCA2 and BRCA3 networks have, respectively, 12140, 13737 and 16981 genes grouped in all nodes (clusters) in their related components. There are 11032 genes common to the three networks. This means that around 90% of the BRCA1 network genes are also represented in the largest connected component of the BRCA2 and BRCA3 networks. This suggests that the networks obtained by the Mapper algorithm applied to the set of genes common to the three studies would have even more similar topological measures.

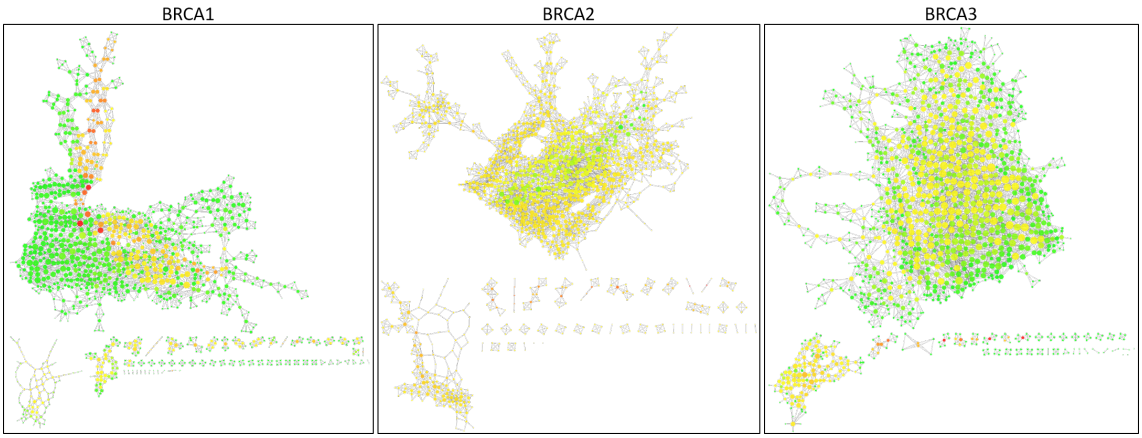

Figure 7. Networks obtained from BRCA studies. The size of the nodes in each network is proportional to the number of its connections (degree). The color scale represents the betweenness centrality of each node, from green (least) to red (greatest).

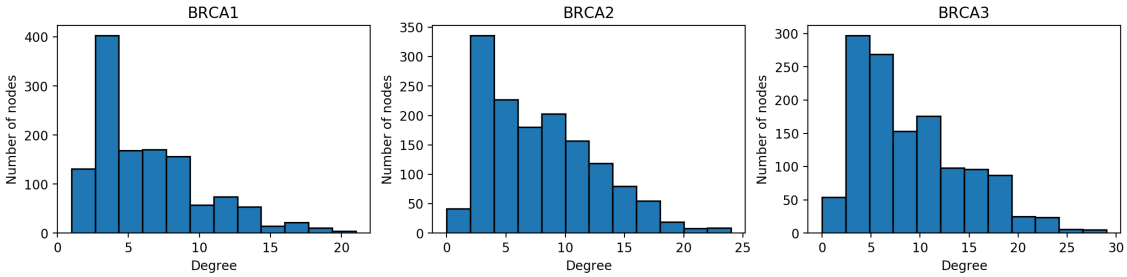

Figure 8. Degree distribution of BRCA data sets.

Table 4. Topological measures of BRCA networks

	cc	CoC	d	sp	ng	N	$Nc1/Nc2$
BRCA1	0.513	59	35	10.041	6.149	1265	875/91
BRCA2	0.494	42	31	9.589	7.599	1435	1105/158
BRCA3	0.512	37	19	7.734	8.933	1290	1019/130

4. Conclusion

This work aimed to make an exploratory and visual analysis of three data sets of three different types of cancer: 1) Breast Invasive Carcinoma (BRCA); 2) Lung Adenocarcinoma (LUAD); and 3) Prostate Adenocarcinoma (PRAD). Our main objective was to understand and explore the mutational characteristics of different studies concerning the same type of cancer. For this purpose, we used visual analyses, with several charts comparing the number of mutations, methods of identification of drivers, and classification of mutations.

Considering the selected data sets, we could observe that there is a similarity between the BRCA and LUAD data sets. However, in the case of PRAD, the similarity is not evident. For example, when we look at the top 10 most mutated genes and the types of mutations involved, only one gene appears in the three PRAD data sets, while there is a

intersection of seven genes in BRCA and eight genes in LUAD. Also, we were able to observe that the application of three computational methods in the identification of drivers showed a more homogeneous and more consensual behavior when we selected the top 50 genes for the BRCA and LUAD cancers. When we consider the distribution of SNV classes, we observed that for the three types of cancer, the same pattern was produced on three data sets.

We also applied a visual topological analysis of BRCA, which can be used as a criterion for choosing the most convenient mutation data set according to the study to be developed. Such analysis revealed topological similarities between the networks obtained from each of the three studies.

This study could not be generalized because several factors interfere in the production of mutation data, such as the date of collection of the study, the quality of the sequencing performed, and also the specificities of each type of cancer. Nonetheless, the analyses we performed in this work have the possibility to be replicated to other cancer mutation data sets. Expanding these analyses to more types of cancer, considering more data sets and exploring more deeply the topological analysis of cancer data are natural continuations of this work. These will contribute further to the understanding of mutation data, helping researchers to make choices according to their study objectives.

All routines and codes used to perform this work are available online[2].

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.

Barbieri, C. E., Baca, S. C., Lawrence, M. S., Demichelis, F., Blattner, M., Theurillat, J.-P., White, T. A., Stojanov, P., Van Allen, E., Stransky, N., et al. (2012). Exome sequencing identifies recurrent spop, foxa1 and med12 mutations in prostate cancer. *Nature genetics*, 44(6):685–689.

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., and Schultz, N. (2012). The cbio cancer genomics portal: An open platform for exploring multidimensional cancer genomics data. *Cancer Discovery*, 2(5):401–404.

Chazal, F. and Michel, B. (2017). An introduction to topological data analysis: fundamental and practical aspects for data scientists. *arXiv preprint arXiv:1710.04019*.

Cho, A., Shim, J. E., Kim, E., Supek, F., Lehner, B., and Lee, I. (2016). Muffinn: cancer gene discovery via network analysis of somatic mutation data. *Genome Biology*, 17(1):129.

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.

Collisson, E., Campbell, J., Brooks, A., and others. (2014). Comprehensive molecular profiling of lung adenocarcinoma. *Nature*, 511(7511):543–550.

COSMIC (2019). Mutational signatures. [Online; accessed March-2020].

[2] `https://github.com/RodrigoHenriqueRamos/SBCAS-2020`

Demkow, U. and Ploski, R. (2015). *Clinical applications for next-generation sequencing*. Academic Press.

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. *Sci. Signal.*, 6(269):pl1–pl1.

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

Imielinski, M., Berger, A. H., Hammerman, P. S., Hernandez, B., Pugh, T. J., Hodis, E., Cho, J., Suh, J., Capelletti, M., Sivachenko, A., et al. (2012). Mapping the hallmarks of lung adenocarcinoma with massively parallel sequencing. *Cell*, 150(6):1107–1120.

Koboldt, D., Fulton, R., McLellan, M., et al. (2012). Comprehensive molecular portraits of human breast tumours. *Nature*, 490(7418):61.

Kumar, A., Coleman, I., Morrissey, C., Zhang, X., True, L. D., Gulati, R., Etzioni, R., Bolouri, H., Montgomery, B., White, T., et al. (2016). Substantial interindividual and limited intraindividual genomic diversity among tumors from men with metastatic prostate cancer. *Nature medicine*, 22(4):369.

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., Kiezun, A., Hammerman, P. S., McKenna, A., Drier, Y., Zou, L., Ramos, A. H., Pugh, T. J., Stransky, N., Helman, E., Kim, J., Sougnez, C., Ambrogio, L., Nickerson, E., Shefler, E., Cortés, M. L., Auclair, D., Saksena, G., Voet, D., Noble, M., and DiCara, D. (2013). Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature*, 499.

Nicolau, M., Levine, A. J., and Carlsson, G. (2011). Topology based data analysis identifies a subgroup of breast cancers with a unique mutational profile and excellent survival. *Proceedings of the National Academy of Sciences*, 108(17):7265–7270.

Patania, A., Selvaggi, P., Veronese, M., Dipasquale, O., Expert, P., and Petri, G. (2019). Topological gene expression networks recapitulate brain anatomy and function. *Network Neuroscience*, 3(3):744–762.

Shyr, C., Tarailo-Graovac, M., Gottlieb, M., Lee, J. J., van Karnebeek, C., and Wasserman, W. W. (2014). Flags, frequently mutated genes in public exomes. *BMC medical genomics*, 7(1):64.

Singh, G., Mémoli, F., and Carlsson, G. E. (2007). Topological methods for the analysis of high dimensional data sets and 3d object recognition. In *SPBG*, pages 91–100.

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., and Lopez-Bigas, N. (2013). Comprehensive identification of mutational cancer driver genes across 12 tumor types. *Scientific Reports*, 3:2650–.

TCGA (2020). The cancer genome atlas. [Online; accessed March-2020].

TCGA, C. . (2018). TCGA, cell 2018. [Online; accessed March-2020].

Veen, H. J. V. and Saul, N. (2017). Keplermapper: a python class for visualization of high-dimensional data and 3-D point cloud data. http://doi.org/10.5281/zenodo.1054444. [Online; accessed March-2020].

WHO (2018). Cancer – (world health organization). https://www.who.int/news-room/fact-sheets/detail/cancer. [Online; accessed March-2020].