Towards an automated pipeline to model a complex-network-driven analysis of microRNAs in cancer: a TCGA-BRCA case study

Mylena Roberta[1], **Murilo Vieira Geraldo**[2], **André Santanchè**[1]

[1]Laboratório de Sistemas de Informação (LIS) — Instituto de Computação
Universidade Estadual de Campinas (UNICAMP) — Campinas, SP — Brasil

[2]Departamento de Biologia Estrutural e Funcional — Instituto de Biologia
Universidade Estadual de Campinas (UNICAMP) — Campinas, SP — Brasil

m222687@dac.unicamp.br, {murilovg,santanch}@unicamp.br

Abstract. *We present the Transparent Reproducible Pipeline (TRP), a core component of our framework for systematizing the comparative analysis of cancer microRNA networks. The TRP is an open, stepwise pipeline for modeling these networks. It provides transparency in materializing intermediary artifacts as tables with schemas, affording explicit semantics and annotation-based provenance. It also offers reproducibility through its open-source code and comprehensive documentation, all accessible without restrictions. To apply and validate the TRP, we conducted a controlled study on breast cancer based on the Breast Invasive Carcinoma (BRCA) project of The Cancer Genome Atlas (TCGA), achieving promising results.*

1. Introduction

The advent of network science dates back to the year 2000. This young field of research is based on the characterization of complex networks and the development of analytical techniques that exploit their topology [da F. Costa et al. 2007]. Complex networks offer a natural framework to systematically represent and analyze the relationships between components of biological systems. Although network representations have been widely adopted in molecular biology for many decades, their convergence with network science is much more recent, giving rise to subfields such as network biology and network medicine [Vulliard and Menche 2021].

Network medicine is an emerging field that fuses systems biology and network science to study human diseases. By applying analytical techniques related to complex networks, it enables a systemic investigation of the perturbations in biological networks that result in disease. Its approaches have contributed to advances in research and treatment development for a variety of complex diseases, including asthma, type 2 diabetes, and several different types of cancer [Barabási et al. 2011].

Cancer is recognized as a systems biology disease, requiring an integrated view to enable continuous advances in its study — a perspective provided by network medicine [Hornberg et al. 2006, Laubenbacher et al. 2009]. Initiatives such as The Cancer Genome Atlas (TCGA) have been essential in supporting the development of complex-network-driven approaches in cancer research. It has generated multi-omic molecular portraits of multiple cancer types. Among its major contributions is the Pan-Cancer project, which

aims to compare diverse cancer types to comprehensively characterize their molecular similarities and differences [Network et al. 2013]. Pan-cancer, or inter-cancer, analyses are a significant component of network medicine's approach to the disease.

MicroRNAs (miRNAs) play a fundamental role in cancer biology and are well established as important disease biomarkers and therapeutic targets. These small non-coding RNAs regulate gene expression by interacting with messenger RNAs (mRNAs) to inhibit their translation [Hayes et al. 2014]. The interactions between miRNAs and mRNAs are complex: each miRNA can target multiple mRNAs, while many miRNAs can target the same mRNA. Complex-network-based analysis has emerged as a key strategy to model and gain deeper insight into these regulatory relationships in cancer [Dragomir et al. 2018]. The role of miRNAs — as well as other non-coding RNAs — continues to increase in significance within network medicine [Gysi and Barabási 2023].

However, as network medicine remains a developing field, there are still open challenges in using complex-network-driven approaches in cancer research, especially in pan-cancer contexts associated with miRNA representations. In our research, we argue that a central challenge lies in exploring network topological analysis to systematically compare different cancer types. Although multiple topological metrics and structures — such as centralities, hubs, modules, and motifs — are widely used in the study of cancer networks, there is a lack of systematic approaches that utilize them in the comparisons inherent to pan-cancer analyses.

In this paper, we present the Transparent Reproducible Pipeline (TRP), which is a core component of our framework for systematizing the comparative analysis of cancer miRNA networks. The TRP provides transparency through an open, stepwise pipeline for modeling these networks, in which each step produces structured, reusable, and reconfigurable artifacts with explicit semantics and traceable actions through annotation-based provenance. The TRP also affords reproducibility through its open-source code in Jupyter Notebooks and comprehensive documentation, all accessible without restrictions at https://github.com/LIS-Unicamp/pan-cancer-analysis/tree/v0.1.0.

To apply and validate the TRP, we conducted a controlled study on breast cancer (BC). We selected BC due to its social relevance and its characterization into distinct molecular subtypes. In recent years, female BC has been identified as the most commonly diagnosed cancer and a leading cause of cancer death worldwide [Sung et al. 2021]. We believe that the biological complexity of BC molecular subtypes allows this case study to closely approximate a pan-cancer investigation, minimizing the variation related to the tissue specificity of the disease. As a basis for our study, we used data from the original TCGA breast cancer project, the Breast Invasive Carcinoma (BRCA) [Network 2012]. The resulting networks show a promising direction towards systematizing complex-network-driven cancer comparative analysis.

The structure of the paper is as follows. In Section 2, we discuss related work. In Section 3, we describe the stages and data artifacts that comprise the TRP. In Section 4, we analyze and validate the application of TRP. In Section 5, we present the main aspects of this work and point out directions for future research.

2. Related Work

MicroRNA (miRNA) networks have been widely used to investigate the regulatory roles of these molecules, particularly in relation to their interactions with messenger RNAs (mRNAs). Below, we discuss studies that address the modeling of cancer miRNA networks and that influenced the development of our work.

[Dragomir et al. 2018] reviewed the main types of networks — monopartite, bipartite, and association — used to analyze biological data related to miRNA function to illustrate that network-based approaches can help improve the selection of miRNAs for therapeutic targeting in cancer. A summary of their discussions on each of them follows.

Monopartite network A graph containing only miRNA nodes and constructed using the expression levels of the molecules. Commonly, an edge connects two nodes with correlated expression above a threshold value. Although this type of network has advantages, the biological meaning of its edges is unclear.

Bipartite network Typically, a graph representing miRNAs regulating mRNAs. This network can be constructed based on anti-correlation analysis of expression data or on published and validated interactions. While each approach has its particular limitations, both produce edges with clear biological meaning.

Association network A graph that is a hybrid between the two previous approaches. It consists of transforming a miRNA-mRNA bipartite network into a miRNA monopartite network using an association index, such as Jaccard, Geometric, or Cosine. An edge connects two nodes that are associated above a threshold value, indicating the specific number of targets shared between a pair of miRNAs.

[Na and Kim 2013] investigated miRNA networks to understand the cooperativity between these molecules. The authors used four co-expression datasets, two of which were related to human cancer (prostate adenocarcinomas and lung cancer). Combining static and functional information, they modeled four network types. (i) A miRNA-mRNA bipartite network of interactions predicted by TargetScan. (ii) Condition-specific miRNA-mRNA bipartite networks, whose edges were the intersection between the predicted interactions and those inferred from expression based on Pearson's correlation coefficient. (iii) Condition-specific miRNA association networks, whose edges were defined by the Jaccard index calculated from the previous networks. (iv) A combined miRNA association network, whose nodes and edges resulted from the union of the preceding networks. The latter was the main target of the analyses proposed by the authors.

Regarding miRNA-mRNA bipartite networks, [Jacobsen et al. 2013] modeled cancer-specific ones and then a pan-cancer one to investigate the interactions shared across cancer types; [Fu et al. 2012] and [Xu et al. 2021] built one to uncover the interactions in colorectal cancer and osteosarcoma, respectively; and [Andrés-León et al. 2017] constructed one to identify the conserved interactions in essential cancer pathways.

Inspired by the works described, we decided to build two types of networks in our pipeline: the miRNA-mRNA bipartite network and the miRNA association network. None of these works adopted an open and reproducible pipeline, such as our Transparent Reproducible Pipeline. To the best of our knowledge, the proposition of a systematic approach to apply complex-network-driven analysis and comparison in the cancer transcriptomic context remains an open challenge, including the development of transparent and reproducible pipelines, as proposed in this work.

3. Methodology

Figure 1 outlines the stages of our pipeline. Following a data flow architecture, each stage represents a process connected to adjacent ones via data artifact transfer. A core aspect of the presented pipeline is our Transparent Reproducible Pipeline (TRP) model, based on artifacts documented with ontology-based semantics and provenance metadata: (i) each artifact has a schema annotated with explicit semantics, and (ii) every field of an artifact B derived from an artifact A receives the following provenance annotations in B: the original field from A (*descendant of*) and the transformation applied, if transformed. In this way, it is possible to interpret the role of each field and trace its evolution along the stages.

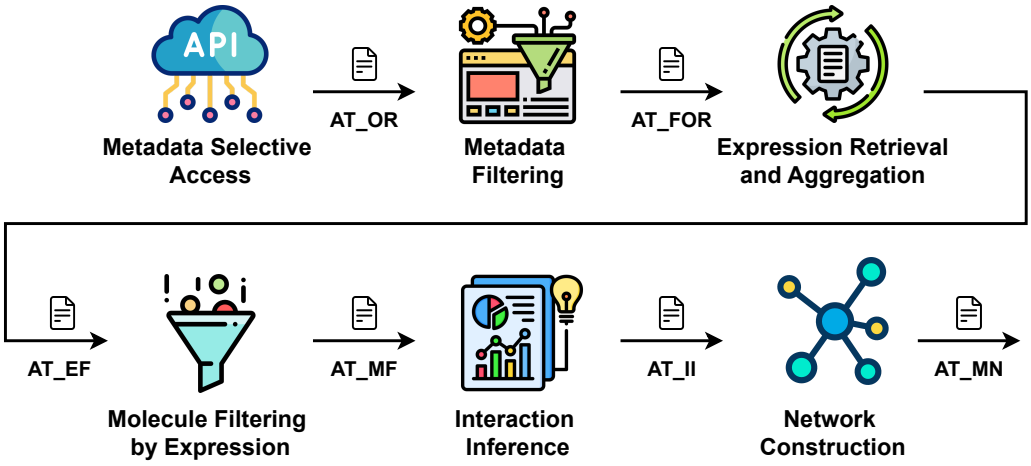

Figure 1. Overview of the stages of our Transparent Reproducible Pipeline. This diagram has been designed using resources from Flaticon.com.

An open GitHub repository presents the schemas of the artifacts and the respective artifacts: https://github.com/LIS-Unicamp/pan-cancer-analysis/tree/v0.1.0/data. As shown in Figure 1, each stage of the TRP produces a family of artifacts identified by the AT_ prefix — e.g., the Metadata Selective Access stage produces the family AT_OR, described in the repository as TCGA Origin Artifacts. Every artifact of the family, documented in the repository, receives a slash and suffix in the ID — e.g., the Project Metadata artifact is AT_OR/PM. The family and artifact IDs are consistently referred to throughout the paper. Field names inside the artifacts' tables will appear in italics.

Explicit semantics rely on references to ontologies based on the role of each field. For example, the Systems Biology Ontology[1] annotates biological agents and processes; the Statistical Methods Ontology[2] documents statistical methods and metrics; and EDAM: the bioscientific data analysis ontology[3] annotates transformations and provenance. Each pipeline stage corresponds to a Jupyter Notebook of the same name in our open GitHub repository. Key parameters for these notebooks are centralized in a Python configuration file and documented in the AT_SETUP artifact family. By modifying these parameters, the pipeline can be tailored to specific requirements.

In the following subsections, we detail each of the six stages presented and the artifact families associated with them.

[1] https://github.com/EBI-BioModels/SBO
[2] https://stato-ontology.org/
[3] https://edamontology.org

3.1. Metadata Selective Access

This stage is responsible for selectively retrieving metadata from external sources and storing it in the artifact family AT_OR. In this study, the pipeline selectively retrieves metadata from the Breast Invasive Carcinoma (BRCA) project of The Cancer Genome Atlas (TCGA). All metadata and data generated by TCGA projects are available on the Genomic Data Commons (GDC) Data Portal[4]. We used its programmatic interface, the GDC Application Programming Interface (API), to implement this stage.

Communicating with the GDC API involves making calls to several endpoints, each representing specific functionality. To discover which data was related to TCGA-BRCA, we needed to explore the metadata contained in the API. The GDC API metadata is organized hierarchically: each project has many associated cases, and each case has multiple related files. Therefore, a top-down selection of the responses from the Projects, Cases, and Files endpoints was necessary to retrieve the TCGA-BRCA metadata.

Based on this selection, we produced the AT_OR family of artifacts: AT_OR/PM, AT_OR/CM, and AT_OR/FM. Each of these artifacts stores the TCGA-BRCA metadata that are relevant to the next stages in our pipeline — AT_OR/PM stores project-related metadata, such as disease type and case and file count; AT_OR/CM stores clinical and biospecimen metadata for the project's cases; and AT_OR/FM stores project-file metadata, such as experimental strategy and sample type.

3.2. Metadata Filtering

When investigating the cases and files related to the TCGA-BRCA project, we found that a significant portion of them fell outside the scope we defined for this case study. This stage addresses the filtering of the metadata of interest from the AT_OR family, creating the AT_FOR artifact family. This filtering does not result in the exclusion of AT_OR/CM case or AT_OR/FM file records; it results in the addition of binary columns to the artifacts, where the records of interest are marked with the value 1 and the others with 0. Combined with the documentation of the parameters used in the filtering, these columns maintain traceability of the fields and record our decisions.

Regarding the files, we searched for those generated by microRNA and RNA sequencing (miRNA- and RNA-Seq) (*experimental_strategy*) and classified under the transcriptome profiling category (*data_category*). Considering the open-access files in the GDC API, there are three types (*data_type*) that correspond to these parameters: miRNA expression quantification (MEQ), isoform expression quantification (IEQ), and gene expression quantification (GEQ). We selected the IEQ and GEQ files as the data to be used in our pipeline. The choice of IEQ over MEQ is due to the fact that only the first data type has reads associated with a specific strand —5p or 3p — of the miRNAs. Furthermore, the files should have been generated from primary tumor or normal tissue samples (*sample_type*). Filtering for AT_OR/FM resulted in the AT_FOR/FM artifact, which consists of the previous artifact added from binary columns.

Regarding the cases, we searched for those characterized as "ductal and lobular neoplasms" (*disease_type*) and with a defined molecular subtype (*pam50_mrna*). We recovered the molecular subtype classification of the TCGA-BRCA cases through Supplementary Table 1 of the article of the project [Network 2012] — artifacts AT_FOR/PC and

[4]https://portal.gdc.cancer.gov

AT_FOR/FPC. To filter them, we aggregated the files of each case, looking for those that presented a co-transcriptomic profiling of primary tumor or paired normal tissue. This refers to a pair of IEQ and GEQ files for at least one of these tissue types. Filtering for AT_OR/CM resulted in the AT_FOR/CM artifact, which consists of the previous artifact added from binary columns and a column characterizing the molecular subtype of the case: basal-like, HER2-enriched, luminal A, or luminal B.

3.3. Expression Retrieval and Aggregation

This stage comprises retrieving expression reads of the files selected in the previous stage and aggregating them according to the miRNA or gene, creating the AT_EF family. Files are retrieved via the GDC API Data endpoint. We dealt with two different file types in this pipeline, one related to miRNA-Seq and the other to RNA-Seq, and each of them requires distinct processing. In this step, the processing consists of specific aggregations for each group — basal-like, HER2-enriched, luminal A, luminal B, and normal.

An IEQ file consists of raw and normalized read counts for miRNA isoforms (AT_EF/MR). Isoforms are identified by their coordinates (*isoform_coords*), miRNA ID (*miRNA_ID*), region type, and, in some cases, MIMAT ID (both in *miRNA_region*). Since only the MIMAT ID indicates the molecule and its strand, we chose to use it as the primary identifier in the aggregation. We filtered the isoforms with associated MIMAT IDs and then summed the raw and normalized read counts by each MIMAT ID (AT_EF/MPR). We aggregated the results in two data artifacts, one for raw read counts (AT_EF/AMR) and the other for the normalized ones (AT_EF/AMN). In both, rows correspond to miRNAs and columns to files — read counts are in the intersections of this matrix format.

A GEQ file consists of raw and normalized read counts for genes (AT_EF/RR). Genes are identified by their Ensembl ID (*gene_id*), name (*gene_name*), and type (*gene_type*). We filtered the protein-coding genes and extracted data from two count methods: unstranded raw (*unstranded*) and TPM unstranded normalized (*tpm_unstranded*) (AT_EF/RPR). Again, we aggregated the results in two data artifacts, one for raw read counts (AT_EF/ARR) and the other for the normalized ones (AT_EF/ARN). In both, rows correspond to genes and columns to files, with read counts at the intersections.

3.4. Molecule Filtering by Expression

Observing the aggregated read counts produced in the previous stage, we noted the presence of molecules, both miRNAs and mRNAs, with low or almost no expression. We concluded that these molecules would not contribute to network construction and future analyses. This stage consists of filtering the relevant miRNAs and mRNAs according to their expression in the aggregated read count files of all groups, creating the AT_MF family. We opted for a widely used approach in the community: the Filter Genes by Expression Level (filterByExpr)[5] function from the edgeR package [Chen et al. 2025].

The processing is separate for miRNAs and mRNAs. For each molecule type, we apply filterByExpr with its default parameters to the concatenation of the aggregated raw read counts of all five groups. This resulted in the artifacts AT_MF/EM and AT_MF/ER, which represent, respectively, the expressed miRNAs and mRNAs. We used these arti-

[5]https://rdrr.io/bioc/edgeR/man/filterByExpr.html

facts to flag the expressed molecules in the aggregated read count files, giving rise to the artifacts AT_MF/AMR, AT_MF/AMN, AT_MF/ARR, and AT_MF/ARN.

3.5. Interaction Inference

When designing this case study, we established that the interaction inference would result from a combination of static and functional information. Static information is associated with the use of sequence complementarity-based target prediction data, while functional information is related to the anti-correlation analysis of the co-transcriptomic profiling data. This stage is responsible for inferring miRNA-mRNA interactions according to the described approach, producing the AT_II artifact family.

The static information we used in this work comes from the miRWalk database [Sticht et al. 2018]. After downloading the miRNA-target files of all expressed miRNAs through a web scraping approach (AT_II/MT), we filtered the interactions of interest (AT_II/MPT). The parameters for this filtering were the following: binding probability greater than 0.9 (*bindingp*), prediction by TargetScan (*TargetScan*), and occurrence in the 3UTR position of the mRNA (*position*). It is worth mentioning that we used the MIMAT IDs for querying in miRWalk and, based on the miRNA-target files, we created an artifact to map each MIMAT ID to the respective miRNA name: AT_II/MNP.

The aforementioned anti-correlation analysis is specific to each of the five study groups. Based on the miRWalk interactions of interest, we calculated the Spearman correlation coefficient for each potential miRNA-mRNA pair. We chose to use Spearman rather than Pearson because we consider the former to be more robust. We also applied the Benjamini-Hochberg false discovery rate control method to calculate the q-values — the adjusted version of Spearman's p-values. The artifact AT_II/IS represents the inferred interactions, associating each with a correlation value and a q-value.

3.6. Network Construction

We chose to construct two types of miRNA network: the miRNA-mRNA interaction network (MRIN) and the miRNA association network (MRAN). Figure 2 illustrates the architecture of these networks. This stage is responsible for the creation of the AT_MN family, whose artifacts are used to construct the networks related to each group. Each of the five study groups has its own MRIN and MRAN. We used the software Cytoscape to visualize the networks generated from these artifacts [Shannon et al. 2003].

The MRIN is a bipartite graph that represents miRNAs regulating mRNAs. The edge between each miRNA-mRNA pair displays an inferred interaction, with a correlation value below -0.3 (*correlation*) and a q-value below 0.05 (*q_value*). Based on the literature, we consider these thresholds appropriate for the identification of statistically significant interactions demonstrating at least moderate correlation. The artifacts AT_MN/IS, AT_MN/INE, and AT_MN/INN are related to the construction of this network type.

The MRAN is an association graph of miRNAs. The edge between each pair of miRNAs displays an indirect interaction defined and weighted by the Jaccard index. We filtered the associations with a Jaccard index above 0.1 (*association)* to build the MRANs. We believe that this threshold is sufficient to select relevant associations. The artifacts AT_MN/AS, AT_MN/FAS, AT_MN/ANE, and AT_MN/ANN are related to the construction of this network type.

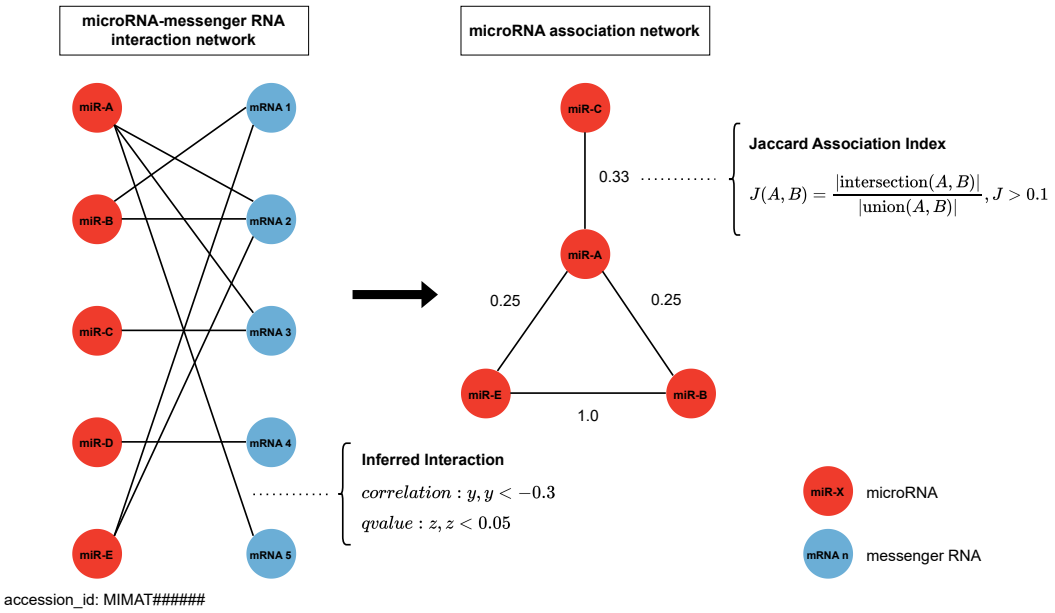

Figure 2. Architectures of the two types of networks built in this work: the miRNA-mRNA interaction network and the miRNA association network.

The MRIN supports analytical questions such as which equivalent mRNAs are regulated by two miRNAs, the degree centrality of miRNAs, or miRNA–mRNA interaction motifs. The MRAN, on the other hand, while losing information on individual mRNAs, expands the analysis to a broader range of metrics, including a richer set of centrality measures and module detection.

4. Pipeline Application

In this section, we describe the analysis and validation of the Transparent Reproducible Pipeline (TRP). As stated earlier, TRP models cancer microRNA networks for a systematic, network-topology-driven confrontation, analysis, and discovery. In the case study of this paper, we applied TRP in the context of breast cancer (BC) molecular subtypes. The basis for this intra-cancer work is the Breast Invasive Carcinoma (BRCA) project of The Cancer Genome Atlas (TCGA), in which BC is characterized into basal-like, HER2-enriched, luminal A, and luminal B subtypes. In addition to these four tumor groups, we used data from the TCGA-BRCA normal tissue analysis to define a control group — a baseline for future comparative analyses.

The TCGA-BRCA has 1,098 cases and 71,079 files. After filtering the metadata according to the mentioned parameters, we obtained 488 cases ($\approx 44.4\%$) and 1,084 files ($\approx 1.5\%$). The proportions of cases and files of interest justify the need for the metadata filtering step. Table 1 shows the quantification of cases and files of interest for each of the analyzed groups. The number of cases per group is representative of the prevalence of BC molecular subtypes [Malhotra et al. 2010]. Since we are dealing only with cases with co-transcriptomic profiling, half of the files in each group were produced by microRNA sequencing and the other half by RNA sequencing. It is relevant to mention that, of all cases of interest, there are 56 cases — luminal A (29), luminal B (14), basal-like (8), and HER2-enriched (5) — that present co-transcriptome profiling of paired normal tissue.

After processing the files, we noticed that a total of 2,128 microRNAs (miRNAs) and 19,314 messenger RNAs (mRNAs) presented at least one read associated with one of the study groups. By filtering these molecules by expression, we calculated that 436 miRNAs ($\approx 20.5\%$) and 16,906 mRNAs ($\approx 87.5\%$) were expressed. Notably, the filtering was more restrictive for miRNAs than for mRNAs. We hypothesize that this is related to the lower expression patterns of miRNAs compared to mRNAs and the high tissue and condition specificity of miRNA expression. Based on the expressed miRNAs, we extracted their predicted interaction files from miRWalk. By filtering these files, we obtained 41,169 interactions as a "static baseline" for the anti-correlation analyses of the functional information of each of the five groups. The results of interaction inference served directly as the basis for the construction of the networks.

To validate that TRP is modeling biologically meaningful networks, we chose to analyze the networks we constructed for the basal-like tumor group. We selected this group because, according to the quantification of nodes and edges, it has middle-sized networks. Figures 3 and 4 illustrate, respectively, the miRNA-mRNA interaction network (MRIN) and the miRNA association network (MRAN) generated by our pipeline for this group. This MRIN has 291 nodes (58 miRNAs and 233 mRNAs) and 296 edges, and this MRAN has 19 nodes and 13 edges. We further provide preliminary observations of the resulting networks to indicate the validity of our pipeline and its potential to support a network-centered analysis. The TRP is a pillar towards systematizing a network-topology-driven analysis in our framework.

In the MRIN, we found several genes important for BC, with oncogenic or tumor suppressor activity already described in the literature. For example, the genes TGFB2, CCND2, and FGF1 are involved in proliferative stimulation (control or promotion), and BCLB11 is involved in escaping programmed cell death. There are also genes significant for tumor aggressiveness or progression, such as TSN1 (cell-cell adhesion), MMP2, ADAM19, and MSN (matrix remodeling for invasion and metastasis), HDAC9 (epigenetic remodeling), and transcription factors linked to tumor stem cell characteristics, such as FOXO1. Well-known suppressor genes such as KLF2 and KLF3 and inflammatory modulators such as TLR4 and CXCL14 also appear in the network. The ZEB1 + miR-200c and miR-200b pairs are widely known to influence tumor aggressiveness in numerous types of cancer, including BC. The JAK1 and ETS1 members are transcription factors classically known to be activated by the MAPK pathway in several types of cancer, including BC, which has participation of the MAPK pathway, mainly in HER+. Finally, the estrogen receptor ESR1 is very important clinically, being a marker of tumor subtype and inducing therapeutic conduct.

In the MRAN, we observed that many of the edges occur between (i) miRNA isoforms or (ii) clusters. (i) miR-29a and miR-29b are isoforms with the same seed region, so their targets are very similar. The same applies to miR-200b and miR-200c. (ii) miR-221 and miR-222 are co-expressed clusters and likely arose from the same ancestor, so they have similar targets. The same applies to miR-17-5p and miR-20a. Furthermore, the association between miR-106b and miR-17-5p has been previously explored in relation to the PI3K pathway [Lee et al. 2019]. Likewise, the association between miR-93 and miR-106b-5p has also been explored previously in relation to the TGFB pathway — both target TGFB2 in our bipartite network [Li et al. 2017].

Table 1. Quantification of cases and files of interest by group.

Group	# of Cases	# of Files
Luminal A	224	446
Luminal B	120	240
Basal-like	88	174
HER2-enriched	56	112
Normal	-	112

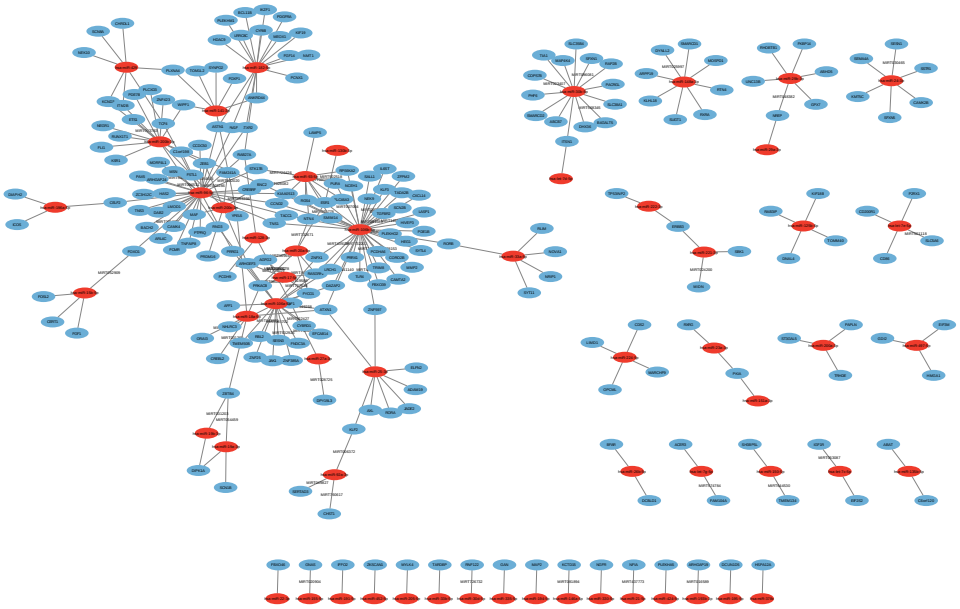

Figure 3. Basal-like miRNA-mRNA interaction network produced by our pipeline. Red nodes represent miRNAs and blue nodes represent mRNAs.

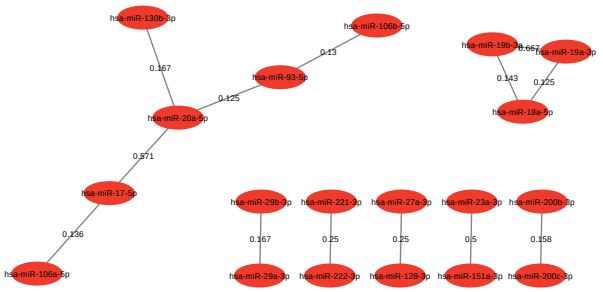

Figure 4. Basal-like miRNA association network produced by our pipeline.

5. Final Remarks

This paper presented the Transparent Reproducible Pipeline for modeling cancer microRNA networks. By analyzing and validating its application in a controlled study on breast cancer, we concluded that our pipeline was capable of producing networks with significant biological meaning. These networks represent appropriate targets for our framework for systematizing the comparative analysis of cancer microRNA networks.

Our TRP, by providing traceable information and transparent, reconfigurable decisions throughout the process, allows not only for open interpretation and inspection but also for reconfiguration and reuse. Each intermediate artifact can be reused in new pipelines and selections, and decisions throughout the pipeline can be adapted according to the context and problem — for example, another cancer type.

In future work, we will address current limitations, including batch effects and the methods used to infer interactions. We also plan to analyze the robustness of the pipeline to parameter changes and optimize stages that involve downloading data. In addition, we aim to expand the pipeline with new stages for differential expression analysis, functional enrichment, and network-topology-based comparative analysis.

Acknowledgements

This study was financed, in part, by the São Paulo Research Foundation (FAPESP), Brasil. Process Number #2024/14197-5; by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Finance Code 001. Process Number #88887.950665/2024-00; by the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), Brasil. Process Number #400062/2023-2; and by INCT ICo-NIoT, funded by CNPq (Process Number #405940/2022-0) and CAPES (Process Number #88887.954253/2024-00). This work was also supported by Amazon Web Services, Inc.

References

Andrés-León, E., Cases, I., Alonso, S., and Rojas, A. M. (2017). Novel mirna-mrna interactions conserved in essential cancer pathways. *Scientific Reports*, 7:46101.

Barabási, A.-L., Gulbahce, N., and Loscalzo, J. (2011). Network medicine: a network-based approach to human disease. *Nature reviews. Genetics*, 12:56–68.

Chen, Y., Chen, L., Lun, A. L., Baldoni, P., and Smyth, G. (2025). edger v4: powerful differential analysis of sequencing data with expanded functionality and improved support for small counts and larger datasets. *Nucleic Acids Research*, 53:13–14.

da F. Costa, L., Rodrigues, F. A., Travieso, G., and Boas, P. R. V. (2007). Characterization of complex networks: A survey of measurements. *Advances in Physics*, 56:167–242.

Dragomir, M., Mafra, A. C. P., Dias, S. M. G., Vasilescu, C., and Calin, G. A. (2018). Using microrna networks to understand cancer. *International Journal of Molecular Sciences*, 19:1871.

Fu, J., Tang, W., Du, P., Wang, G., Chen, W., Li, J., Zhu, Y., Gao, J., and Cui, L. (2012). Identifying microrna-mrna regulatory network in colorectal cancer by a combination of expression profile and bioinformatics analysis. *BMC Systems Biology*, 6:68.

Gysi, D. M. and Barabási, A.-L. (2023). Noncoding rnas improve the predictive power of network medicine. *Proceedings of the National Academy of Sciences of the United States of America*, 120:e2301342120.

Hayes, J., Peruzzi, P. P., and Lawler, S. (2014). Micrornas in cancer: biomarkers, functions and therapy. *Trends in molecular medicine*, 20:460–9.

Hornberg, J. J., Bruggeman, F. J., Westerhoff, H. V., and Lankelma, J. (2006). Cancer: A systems biology disease. *Biosystems*, 83:81–90.

Jacobsen, A., Silber, J., Harinath, G., Huse, J. T., Schultz, N., and Sander, C. (2013). Analysis of microrna-target interactions across diverse cancer types. *Nature Structural & Molecular Biology*, 20:1325–1332.

Laubenbacher, R., Hower, V., Jarrah, A., Torti, S. V., Shulaev, V., Mendes, P., Torti, F. M., and Akman, S. (2009). A systems biology view of cancer. *Biochimica et Biophysica Acta (BBA) - Reviews on Cancer*, 1796:129–139.

Lee, J., Kim, H. E., Song, Y.-S., Cho, E. Y., and Lee, A. (2019). mir-106b-5p and mir-17-5p could predict recurrence and progression in breast ductal carcinoma in situ based on the transforming growth factor-beta pathway. *Breast cancer research and treatment*, 176:119–130.

Li, N., Miao, Y., Shan, Y., Liu, B., Li, Y., Zhao, L., and Jia, L. (2017). Mir-106b and mir-93 regulate cell progression by suppression of pten via pi3k/akt pathway in breast cancer. *Cell death & disease*, 8:e2796.

Malhotra, G. K., Zhao, X., Band, H., and Band, V. (2010). Histological, molecular and functional subtypes of breast cancers. *Cancer Biology & Therapy*, 10:955–960.

Na, Y.-J. and Kim, J. H. (2013). Understanding cooperativity of micrornas via microrna association networks. *BMC Genomics*, 14:S17.

Network, C. G. A. (2012). Comprehensive molecular portraits of human breast tumours. *Nature*, 490:61–70.

Network, C. G. A. R., Weinstein, J. N., Collisson, E. A., Mills, G. B., Shaw, K. R. M., Ozenberger, B. A., Ellrott, K., Shmulevich, I., Sander, C., and Stuart, J. M. (2013). The cancer genome atlas pan-cancer analysis project. *Nature genetics*, 45:1113–20.

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., Amin, N., Schwikowski, B., and Ideker, T. (2003). Cytoscape: A software environment for integrated models of biomolecular interaction networks. *Genome Research*, 13:2498–2504.

Sticht, C., Torre, C. D. L., Parveen, A., and Gretz, N. (2018). mirwalk: An online resource for prediction of microrna binding sites. *PLOS ONE*, 13:e0206239.

Sung, H., Ferlay, J., Siegel, R. L., Laversanne, M., Soerjomataram, I., Jemal, A., and Bray, F. (2021). Global cancer statistics 2020: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer Journal for Clinicians*, 71:209–249.

Vulliard, L. and Menche, J. (2021). *Complex Networks in Health and Disease*, volume 1-3, pages 26–33. Elsevier.

Xu, K., Zhang, P., Zhang, J., Quan, H., Wang, J., and Liang, Y. (2021). Identification of potential micro-messenger rnas (mirna–mrna) interaction network of osteosarcoma. *Bioengineered*, 12:3275–3293.