scientific reports

Check for updates

OPEN

A computational framework for extracting biological insights from SRA cancer data

Paul Anderson Souza Guimarães[1,2], Maria Gabriela Reis Carvalho[1,2,3]✉ & Jeronimo Conceição Ruiz[1,2,3]✉

The integration of sequenced samples and clinical data from independent yet related studies from public domain databases, such as The Sequence Read Archive (SRA), has the potential to increase sample sizes and enhance the statistical power needed for more precise bioinformatic analysis. Data mining and sample grouping are the starting points in this process and still present several challenges, including the presence of structured and unstructured data, missing deposited data, and varying experimental conditions and techniques applied across the studies. Designed to address the main challenges of data mining and sample grouping for biomarkers research, the proposed methodology employs a computational approach integrating relational database construction, text and data mining, natural language processing, network analysis, search by Pubmed publications, and combining MeSH, TTD and WordNet database to identify groups of samples with the same characteristics. As a result, it identifies and illustrates relationships among sample collections, aiming to discover potential cancer biomarkers. In colorectal cancer (CRC) and acute lymphoblastic leukemia (ALL) case studies, this methodology effectively navigates SRA metadata, retrieving, extracting, and integrating data. It highlights significant connections between samples and patient clinical data, revealing important biological insights. The study grouped 2,737 (CRC) and 3,655 (ALL) samples into potential comparison groups, demonstrating the method's power in identifying relationships and aiding biomarker discovery.

Keywords SRA mining tool, Neoplasms, Cancer, Biomarker, Data mining, Natural language processing

In the context of today's scientific advancements, the scientific community generates an immense volume of biological data on a daily basis. These data arise from studies exploring a wide range of topicsand areas of knowledge, encompassing information on genes, proteins, metabolites, and various interactions, including biological, chemical, cellular, and parasitic, among others[1,2]. A significant portion of this data is stored in databases and made available for consultation through various protocols, including websites, datasets, web services, and more. However, even with public, unrestricted access and the current tools available such as websites, web services, APIs, graphical interfaces, command-line interfaces, scripts and pipelines, identifying useful biological information can be challenging and requires specific computational approaches[3–6].

The challenges associated with this endeavor can be broadly categorized into several key areas. First, the heterogeneity and volume of data present a significant obstacle. Public biological databases house information derived from diverse technologies and experimental methodologies, including genomic, transcriptomic, and metagenomic sequencing. This diversity results in heterogeneous data formats—such as RAW, FASTQ, BAM, and VCF.

Second, metadata quality and curation remain critical issues. While raw data are generally well-structured, associated metadata such as experimental conditions, organism information, methodologies, and clinical outcomes are often poorly standardized, inconsistently populated, or incomplete. These deficiencies complicate data retrieval and integration across different studies, as they frequently fail to adhere to international standards and guidelines such as MIAME (Minimum Information About a Microarray Experiment)[7] or MINSEQE (Minimum Information About a Next-generation Sequencing Experiment)[8].

[1]Grupo Informática de Biossistemas, Bioengenharia e Genômica, Instituto René Rachou, Fiocruz Minas, Av. Augusto de Lima, 1715, Barro Preto, Belo Horizonte, MG, Brazil. [2]Biologia Computacional e Sistemas (BCS), Instituto Oswaldo Cruz (IOC), Fiocruz, Rio de Janeiro, Brazil. [3]Maria Gabriela Reis Carvalho and Jeronimo Conceição Ruiz contributed equally to this work. ✉email: maria.gabriela@fiocruz.br; jeronimo.ruiz@fiocruz.br

Third, many tools, including graphical interfaces, APIs, and predefined pipelines, lack the flexibility needed to address specific research needs. They are often designed for generic analyses and may not support more complex use cases[9–13], such as biomarker discovery in particular cancer subtypes. Additionally, many of these tools require advanced programming skills or proficiency in command-line environments[14–19], which can pose accessibility barriers for non-specialists.

Scalability and performance represent another key challenge. Computational methods must be capable of managing ever-growing datasets, which involve tasks such as processing billions of sequencing reads, genome alignments, and large-scale statistical analyses. These operations demand significant computational resources, such as high-performance computing clusters or cloud-based services, which may not always be accessible, particularly in resource-limited settings.

Another significant hurdle lies in the interoperability and integration of data. Different databases and tools often employ distinct naming conventions, taxonomies, and data formats, creating barriers to cross-platform analyses and meta-analyses[20–22]. Although open standards are key to addressing these issues, their adoption remains inconsistent across resources[8,23].

Extracting biologically meaningful information from raw genomic data is a multifaceted challenge requiring advanced analytical methodologies to identify patterns, correlate genetic variations with phenotypes, and validate findings experimentally. Despite the progress in bioinformatics, the absence of standardized pipelines for these tasks often compromises reproducibility and impedes the broader validation of results. Furthermore, contextualizing extracted data within functional biological frameworks is equally demanding. This step necessitates the integration of genomic insights with existing knowledge bases, such as protein-protein interaction networks, metabolic pathways, and functional ontologies[20–22]. Achieving this requires interdisciplinary expertise and access to diverse, complementary databases, underscoring the complexity of translating genomic data into actionable biological insights. Addressing these challenges is pivotal for advancing the identification and validation of disease-specific biomarkers and improving translational research outcomes.Among other public domain databases such as Gene Expression Omnibus (GEO)[24] and The Cancer Genome Atlas (TCGA)[25], the Sequence Read Archive (SRA)[26] is recognized as a crucial data source for biological research. This repository was established with the aim of maintaining and providing access to raw next-generation sequencing data from various technologies, thereby promoting reproducibility, enabling meta-analyses, and even facilitating new conclusions from already sequenced samples[27]. The database was developed following the guidelines of the International Nucleotide Sequence Database Collaboration (INSDC). Instances of the SRA are maintained by the National Center for Biotechnology Information (NCBI), the European Bioinformatics Institute (EBI), and the DNA Data Bank of Japan (DDBJ), persevering in the mission to provide data in a permanent, free, and unrestricted manner to the community[27–29].

Although the database instances are geographically independent, they adhere to strict data synchronization and maintenance protocols, ensuring that data published in any one region is quickly made available in the others. While the primary goal is the maintenance of raw sequencing data, the SRA database also aggregates sequencing-related data and metadata, including protocols, samples, techniques, analysis results, patient clinical data, and more, pertaining to sequencing experiments from platforms such as the Roche 454 GS System®, Illumina Genome Analyzer®, Applied Biosystems SOLiD System®, Helicos Heliscope®, Complete Genomics®, and Pacific Biosciences SMRT®[27,29].

Among the various tools currently available to facilitate the access, analysis, and annotation of data and metadata from next-generation sequencing (NGS) repositories, those specifically focused on Sequence Read Archive (SRA) data operate on principles aimed at improving accessibility, enhancing metadata quality, and integrating advanced computational approaches to address the challenges inherent in large-scale biological data management. Tools such as MetaRNA-Seq [30] and those employing deep learning-based named entity recognition prioritize improving metadata coverage and quality through advanced computational techniques, including machine learning and interactive browsing interfaces. These methodologies empower users to derive more comprehensive and biologically meaningful annotations, significantly enhancing both the interpretability and practical utility of the datasets. These tools are also designed to simplify data access and retrieval. For example, **grabseqs** [31] provides a user-friendly solution for downloading reads and metadata from multiple repositories, focusing on accessibility for researchers with limited programming skills. In contrast, **SRAdb**[19] integrates seamlessly with the R programming environment, offering programmatic access to SRA metadata and appealing to users proficient in statistical computing and data manipulation. While both tools prioritize usability, they are tailored to distinct user groups depending on their technical expertise. The robustness of these platforms depends on their specific focus and technological foundation. For example, tools like **SRA Down Under**[32], which integrates caching with analytical functionalities, provide a framework for managing large datasets and conducting domain-specific analyses, particularly in infectious disease research. This seamless integration of data management and analytical capabilities enhances workflow efficiency and supports targeted investigations. On the other hand, tools leveraging natural language processing (NLP), such as those combining NLP with metabarcoding, exhibit robustness in automating the extraction of pathogen-environment associations from unstructured data sources[33,34]. However, their effectiveness is highly contingent on the quality and scope of the training data used to develop these computational models. These tools relying on NLP often encounter challenges when dealing with domain-specific or poorly curated metadata, which can limit their ability to generalize across diverse datasets. This underscores the critical need for curated and comprehensive training datasets to ensure consistent performance across varying research contexts and applications.

In general, the limitations of these tools stem from their scope and specialization. For example, tools designed for specific applications, such as **MetaRNA-Seq**[30], may lack adaptability to other types of sequencing data or analytical objectives. Similarly, platforms like **SRAdb**[19], which depend on user expertise in R programming, can be less accessible to researchers with limited technical skills, restricting their usability for a broader audience.

The application of metadata mining approaches through the aforementioned tools is particularly impactful in cancer biomarker research. High-throughput RNA sequencing data, accessible through repositories such as SRA, has revolutionized the identification and validation of potential biomarkers.

The term "biomarkers" encompasses a wide range of measurable indicators, including molecular, histologic, radiographic, and physiologic characteristics. These biomarkers serve various purposes, such as diagnosing diseases, predicting disease progression, monitoring treatment responses and guiding therapeutic decisions.

It is also important to underline that, despite the benefit of various RNA-seq datasets[35,36] deposited in a database, RNA-seq experiments typically involve sequencing a small number of samples, primarily due to their cost, which, despite having decreased over the last decade, still remains high (USD 200 to USD 1,000 per sample). This context leads to the issue associated with the low statistical power of studies analyzing the differential expression of thousands of genes in a number of samples that may vary only from dozens to hundreds. To overcome this issue, the combination of data and/or results from independent yet related studies has been widely used to increase the sample size and consequently the statistical power required for a more accurate inference of differential gene expression[37,38]. Consequently, various approaches have been proposed, initially for the integration of microarray studies[39,40] and more recently adapted for RNA-Seq studies[41,42].

Some important issues in this process, however, include the variability in data quality that arises from inconsistent quality control measures, leading to concerns regarding data reliability. Moreover, the lack of standardization is evident, with studies employing diverse protocols, platforms, and annotations, complicating the integration of data from multiple sources. Additionally, data annotation and metadata often lack completeness, omitting crucial clinical information and sample details. Furthermore, inconsistent annotations impede comparative analysis and integration efforts[43,44].

To address these challenges and contribute to cancer biomarker research this paper presents a new methodology designated to facilitate the identification, selection, categorization, and grouping of collections of massively sequenced samples deposited in the SRA, aiming to integrate data associated with these samples with clinical patient data. This methodology was also designed to address and overcome the limitations identified in existing platforms used for accessing, analyzing, and annotating next-generation sequencing (NGS) data and metadata, particularly from the Sequence Read Archive (SRA). Through the integration of SRA-associated metadata and patient clinical data using a data-mining, text-mining, NLP and network integrated approach, the proposed methodology integrates solutions to these shortcomings by enhancing metadata quality, ensuring adaptability to various data types and analytical goals, and offering an intuitive user interface that balances accessibility with advanced functionality. By combining robust computational frameworks with innovative strategies for metadata standardization and cross-platform integration, this tool aims to streamline workflows, support more complex biological inquiries, and provide a versatile solution for researchers working with large-scale NGS datasets.

Results

The methodology proposed in this study was implemented as a Python package to automate the processes of querying SRA data, constructing a local database, indexing data, performing natural language processing and grouping of samples with shared characteristics. This process is divided into two steps which can be executed either via the command line or through Python code.

Firstly, local databases are constructed containing all SRA metadata and Pubmed publications related to the sequences. After database construction, users can navigate through local data to prospect interest groups of sequenced samples in a graphical tool that could be implemented at a web server.

One local database, described in the sections below, is constructed using colorectal cancer data through this Python package. The implementation of the methodology in the Python package is described in the "Methods" section. The execution times for each stage are provided in the Supplementary file 1.

Centralized database

Using the methodology developed in this work, a centralized database containing all metadata related to high-throughput data sequencing of human colorectal cancer samples was constructed. All metadata related to public accessible studies were transferred from the SRA database. In total, approximately 45,000 experiment packages were downloaded and inserted into the local database, including 44,884 experiments, 2,000 experimental designs, 200 organizations, 421 articles, 54,698 runs, 39,833 samples, 1,043 studies, and 1,165 submissions.

For all 132 tables inserted into the database, a numeric internal key was assigned to preserve relationships between database entities. All tables were protected with cascade instructions for updates and restrict instructions for deletions. Table names were shortened to comply with the maximum character limits of the DBMS, and a name translation table was added to the data schema.

This database is automatically constructed using Python scripts described in the following sections and is used in case study presented in the results section.

Indexing submissions from the SRA database

Developed within Python, the indexing methodology includes executable scripts and a set of specialized modules for acquiring, processing, organizing, and indexing data related to high-throughput data sequencing deposited in the SRA database. The methodology facilitates seamless execution of all stages, integrating: (a) search capabilities through the E-utilities package; (b) query functionalities in PubMed; (c) constructing centralized databases; and (d) following a relational model, and extracting groups of samples sharing similar characteristics.

The methodology incorporates a variety of objects designed for feature tracking, text and data mining (including PubMed), sample indexing, and clustering through a network approach. Additionally, it can be directly utilized via Python module implementation or command-line scripts.

The database package comprises three command-line scripts. The 'sra.py' script enables direct searches within the SRA database, transferring, parsing and storing metadata in a local database. The 'prepare_database.py' script maps structures and data within the local database, constructing all necessary tables for data handling, indexing, and sample clustering. The 'network.py' script constructs networks based on mapped data from the local database.

Both scripts perform tasks by utilizing modules implemented in the data package. The 'database', 'mesh', and 'sra' modules enable manipulation of the local database, access to MeSH data, and interaction with SRA, respectively. The 'xml', 'mining', 'synonym', and 'util' modules provide necessary objects, methods, and functions for handling SRA search results, data and text mining, and grouping of synonymous data. 'Network' and 'pyvisnetwork' modules contain elements for constructing and displaying sample networks, generating tables, graphs, and other files useful for searching similar samples.

Querying local database

Table and field evaluation is facilitated by the web package developed. By executing the 'manage.py' script (the web service manager of the Django framework), a web application, mapping the entire database described in the "Centralized database" and in the "Methods" sections, accessible via a web browser is launched (Fig. 1). Once loaded, database structures can be viewed in the Filter menu as a directory tree. The tree visually represents the hierarchical definition of structures within the database.

All data is organized under the base element 'experiment_package', represented as a folder. Tables constructed from data extracted from SRA are denoted by a yellow star for base tables and an orange star for sub-tables (tables referencing other tables). Attributes are represented by a leaf icon, while tables resulting from data grouping and mining are depicted by a constellation icon. Within this interface, searches can be conducted using entire tables or by selecting attributes from each table. By default, all searches target the table containing the list of samples. This behavior, along with table or attribute selection, can be adjusted via a dropdown menu accessible through right-clicking. Once the search criteria are defined, clicking 'Get network' initiates the process .

The search results consist of a set of files containing an index summarizing the located sample groups, data used for network generation in CSV and JSON formats, and a file containing the search parameters in JSON format.

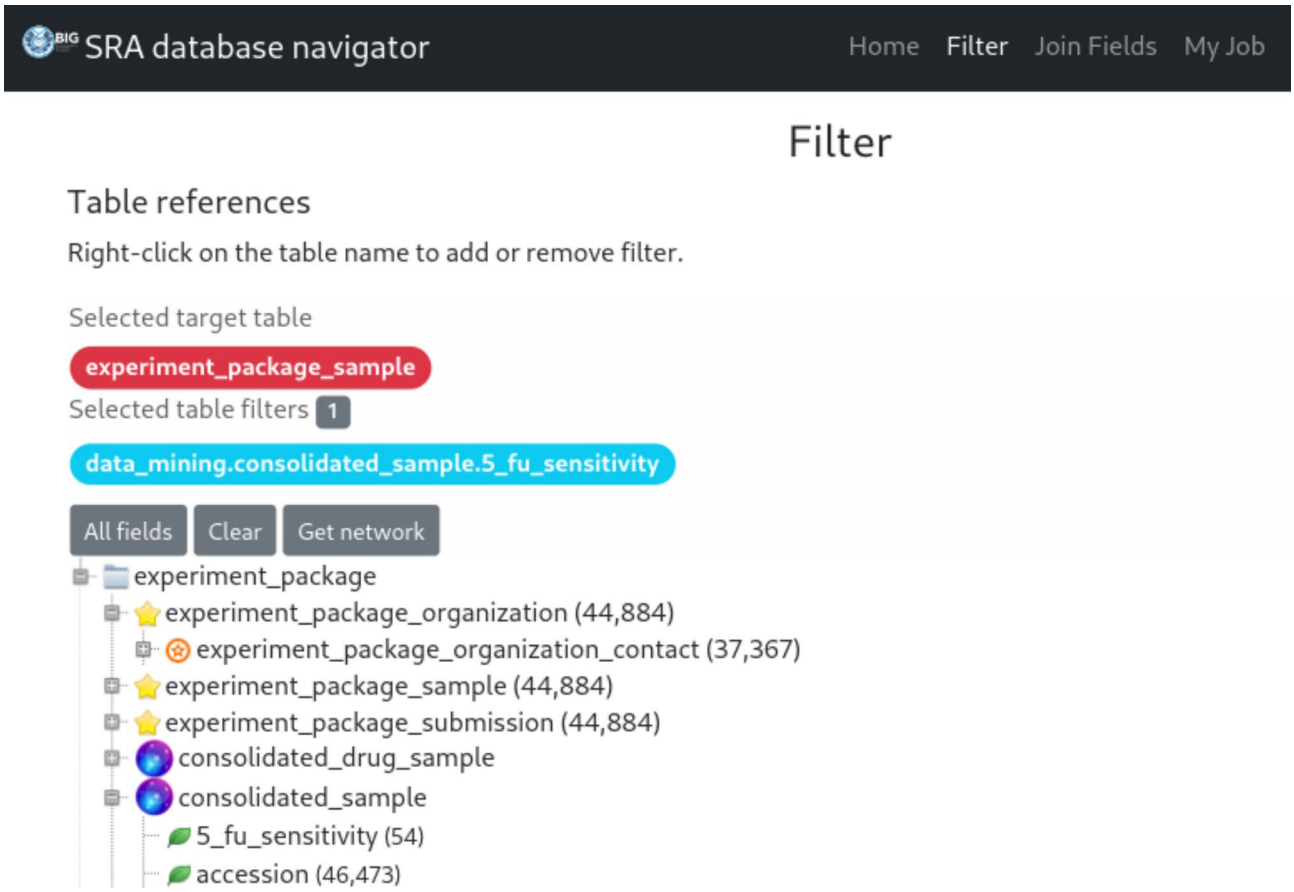

Fig. 1. Search interface for sample groupings. Searches are conducted against the target table and can be based on one or more attributes or entire tables. *Yellow star*: base table. *Orange star*: sub-table. *Constellation*: set of attributes obtained through normalized data and text and data mining. *Leaf*: attribute or column. Screenshot captured using Debian GNU/Linux 12 (bookworm) and cropped with GIMP 2.10.34.

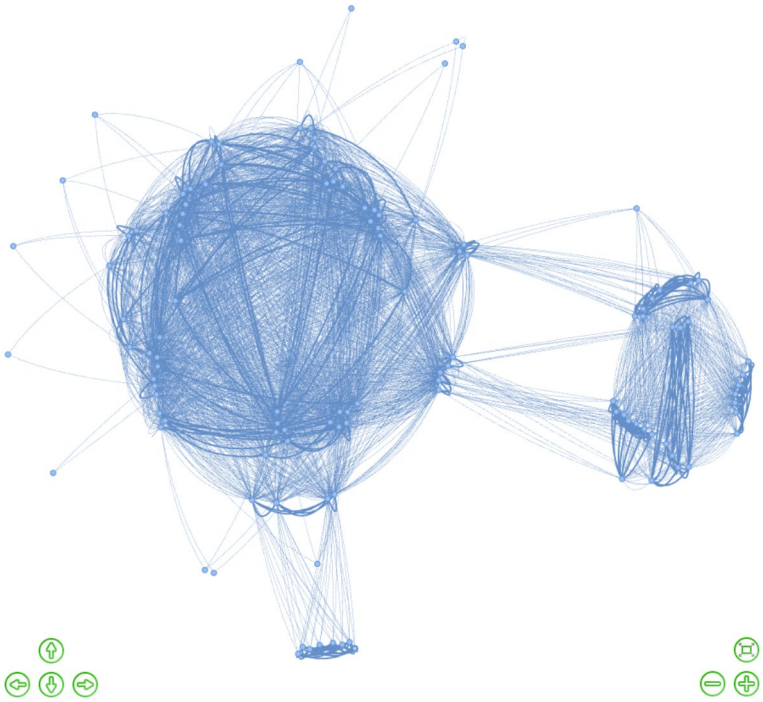

Fig. 2. Network inspect area. Samples are represented by blue circles, and the edges, shown as solid blue lines, indicate shared characteristics between samples. The edge thickness represents the connection weight, highlighting the sharing of multiple characteristics. Image generated using the Python package developed in this study, implementing a customized version of the PyVis library version 0.3.2.

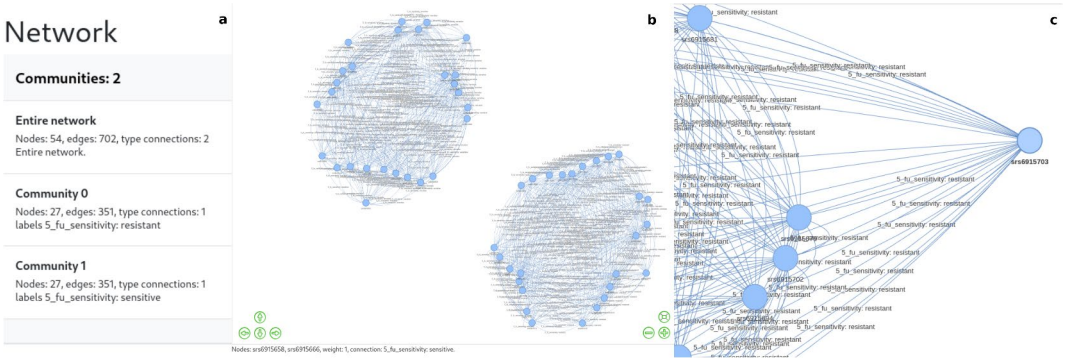

Fig. 3. Network sample results. Samples are represented by blue circles, and the edges, shown as solid blue lines, indicate shared characteristics between samples. The edge thickness represents the connection weight, highlighting the sharing of multiple characteristics. Users can interact with the figure using mouse movement and buttons. Clicking on the description opens the corresponding network for visualization. (A) Summary of data returned in the search, showing the number of edges, communities, and nodes. (B) Groups of samples sharing resistance or sensitivity to a drug in the context of colorectal cancer search. (C) Displaying a portion of the resilient group after zooming in and manually moving an individual to inspect details. Image generated using the Python package developed in this study, implementing a customized version of the PyVis library version 0.3.2 and the Inkscape software version 1.2.

The networks generated by the package can be viewed directly in web browsers such as Firefox or Google Chrome (Fig. 2), allowing users to interact with the display environment, nodes, and edges. Users can move elements or the display window, adjust zoom, view details, and temporarily reposition nodes and edges (Fig. 3).

Some searches may result in groups containing a high number of nodes and edges, which could overload web browsers during the display process. Additionally, inspecting large networks with an excessive amount of elements can be laborious. Therefore, a series of scripts has been included to assist in this process. The scripts 'check_community.py' and 'check_community_by_csv.py', 'plot_networkx.py' and 'sort_indexes.py' extract data

from search results to create summary files containing descriptions of connections, nodes, weights, attributes, values, diagram, and non-interactive visualizations.

Usage and functionalities through a colorectal cancer search: a case study

According to GLOBOCAN data, in 2022, colorectal cancer (CRC) was the third most common cancer in terms of incidence (1,142,286 cases in the colon and 729,833 in the rectum) and the second most common cancer in terms of mortality (538,167 deaths in the colon and 343,817 in the rectum worldwide[45,46]. The disease is highly complex, and the prognosis, pathology, and treatment can vary based on factors such as sex, age, diet, organ localization affected, molecular mechanisms, and more. Most patients develop the disease from a polyp, which results from aberrant crypts that can evolve into cancer over a period of 10 to 15 years[47–49]. Cancer can be stratified on a scale from 1 to 4, with the histological cell grade ranging from 1 (well-differentiated) to 4 (undifferentiated), or by stage, considering the tumor, lymph nodes, and metastasis. Considering the problem of cancer, Pierre Denoix developed the Tumor-Node-Metastasis (TNM) staging system for cancer stratification in the 1940s. Thereafter the American Joint Committee on Cancer and the Union for International Cancer Control maintain the TNM Staging System as a recommended international standard cancer classification system assisting physicians and other professionals[50–52].

Cancer staging ranges from stages 0 to 4 (AJCC eighth edition) according to disease progression. The TNM system describes the primary tumor (T), the affected lymph nodes (N) and the presence of metastasis (M). T values ranging from TX to T4, NX to N3, M0 to M1 do describe tumor, lymph nodes and metastasis impact respectively in an international standard form (Table 1)[50,52,53].

Despite many reported and ongoing studies, several major scientific questions about CRC remain only partially elucidated and could benefit from the computational methodology proposed.

Aiming to verify the functionality of the developed methodology, we have focused on searches that could identify attributes to integrate more samples and help to answer the following questions: (a) Genetic and molecular mechanisms: What are the specific genetic mutations and molecular pathways that initiate and drive the progression of colorectal cancer? How do these genetic and molecular changes vary among different populations and colorectal subtypes? (b) Metastasis: What are the key factors that drive colorectal cancer metastasis, particularly to the liver and lungs?

Through the developed methodology, we performed a SRA search for sequencing data related to colorectal cancer using the search terms '(colorectal cancer) AND "Homo sapiens"[orgn: __txid9606]'. The search resulted in the construction of a local database containing 44,884 packages of experiments. The database structure comprised two relational schemas, consisting of 132 related tables storing experiments metadata.

After fully processing the metadata, we extracted a list of all registered drug names from TTD and mapped them to the local database. Then, we performed a search using the consolidated drug samples (a database table that stores sample accession numbers and maps them to the TTD approved drug names). This facilitates the association of specific drug treatments with individual samples, aiding in comprehensive data analysis for research and clinical purposes. The table ensures accurate identification and retrieval of drug information corresponding to each sample's unique identifier, supporting effective study and evaluation of treatment outcomes. Integrated with this table is a series of attributes that would characterize the pathology, sample, and patients. A total of 120 different attributes (drug name of consolidated drug samples and all 119 attributes of consolidated sample mining terms) were used (see Supplementary file 2). As a result, we obtained a network containing 30,899 nodes, 764,222,174 edges, and 11,159 different connection types, representing 363 communities. Fourteen different weights ranging from 0.0714 to 1 were identified. The largest and smallest numbers of nodes were 22,830 and 2, respectively.

Interested in the networks that exhibited the highest number of different attributes and consequently greater weight, we selected the largest community and reconstructed it by removing all edges with a weight less than 0.2, the minimum weight in this CRC network, which represents edges with more than one attribute (applying this exclusion criterion removes edges that are connected by a single attribute). As a result, we obtained a new network consisting of 400 communities, 8,991 nodes, 528,552 edges, and 8,262 different types of connections. The largest and smallest numbers of nodes were 6,195 and 2, respectively. After applying this filter, the network displayed 16 high-throughput sequencing technologies in use: 3,883 RNA-seq samples (144 studies), 713 Whole Exome Sequencing (WXS) samples (19 studies) and 4.722 samples from other technologies (95 studies). Table 2 shows the number of samples and studies found in the filtered network.

This new network was utilized for all subsequent analyses and ranked in descending order by the number of connections, attributes, and non-redundant values. The ten highest-ranked communities are shown in Fig. 4 illustrating some of the communities generated by the computational methodology developed in this study. See the "Code availability" and "Data Availability" sections for complete results and source code.

To shed light on how genetic and molecular changes vary among different populations and colorectal cancer subtypes, we explored the information contained in the above mentioned network. This analysis provided an association list between samples and patient clinical information, with the potential to offer researchers a comprehensive understanding of the variations and their implications for diagnosis and treatment in further studies.

Using samples extracted from these networks, either individually or collectively, researchers could investigate genetic, molecular, or mutational profiles in groups of individuals sharing similar characteristics. To address the first question, groups of samples could be constructed by considering cancer stage and grade (AJCC (American Joint Committee on Cancer stratification system) and CRC grade), affected organs/tissues, patient age, and gender identity. In many cases, information about the affected tissue and tumor malignancy could be inferred from the T, N, and M values (TNM system), which describe the primary tumor site and size, regional lymph node involvement, and the presence or absence of distant metastatic spread, respectively. Thus, groups of

Stage	TNM system		
	T	N	M
0	Tis	N0	M0
I	T1	N0	M0
	T2	N0	M0
IIA	T3	N0	M0
IIB	T4a	N0	M0
IIC	T4b	N0	M0
IIIA	T1	N1	M0
	T1	N1c	M0
	T2	N1	M0
	T2	N1c	M0
	T1	N2a	M0
IIIB	T3	N1	M0
	T3	N1c	M0
	T4a	N1	M0
	T4a	N1c	M0
	T2	N2a	M0
	T3	N2a	M0
	T1	N2b	M0
	T2	N2b	M0
IIIC	T4a	N2a	M0
	T3	N2b	M0
	T4a	N2b	M0
	T4b	N1	M0
	T4b	N2	M0
IVA	Any T	Any N	M1a
IVB	Any T	Any N	M1b

Table 1. Colorectal cancer staging. The AJCC Colorectal Cancer staging system ranges from 0 to V. This system evaluates the tumor (T), lymph nodes (N), and presence of metastasis (M) for cancer stratification. TNM is the recommended international system for stratifying colorectal cancer patients. T values (primary tumor): X indicates the tumor cannot be evaluated; 0 indicates no evidence of tumor; IS refers to carcinoma in situ; 1–4 denote the extent and size of the tumor. N values (nearby lymph nodes): X indicates the lymph nodes cannot be evaluated; 0 indicates no lymph nodes are affected; 1–3 indicate the extent and number of affected lymph nodes. M values (distant metastasis): 0 indicates no distant metastasis; 1 indicates the presence of distant metastasis. AJCC: American Joint Committee on Cancer. Extracted and adapted from[51–53].

patients in stages I, II, and III could be further evaluated to find genetic or mutational characteristics that might explain their respective phenotypes, considering affected organs, progression, and tumor size. Other possible comparisons could consider the profiles of samples from the right side versus the left side of the intestine to understand differences in diagnosis and prognosis.

Addressing the case study question related to molecular mechanisms and metastasis

To investigate the case study question concerning molecular mechanisms and metastasis, we organized various attributes to examine how elements could be interconnected by using the sample attributes (the number of samples sharing attribute values). We begin by identifying characteristics that could be associated with the highest number of samples, which could then be used to describe social and clinical data. The "AJCC" attribute (câncer staging) grouped 2,737 RNA-seq samples from 85 studies (Supplementary file 3). These nodes could help answer questions related to samples of different cancer stages, from normal to metastatic scenarios. Polyps samples (74 nodes, 20 positive polyps and 54 negative polyps, one study, (supplementary file 4)), should be included as control and comparison points for earlier cancer stages.

Additionally, 827 RNA-seq nodes from 5 studies were connected by attributes such as "age", "AJCC", "crispr_therapy", "drug", "ethnicity", "gender_identity" and "health_state" attributes (Supplementary file 5). These nodes comprised male and female patients aged 26 to 85 years who were treated with bevacizumab, cetuximab and fluorouracil. These samples could be grouped by 7 ethnicities: "Asian", "White", "Black or African American", "More than one race", "American Indian or Alaska Native", "Asian; Taiwanese" and "Mongoloid; Taiwanese". The samples were classified into different metastatic colorectal cancer stages according to the "AJCC" attribute, ranging from I to IV, and TNM cancer staging from T2 to T4 (There were no samples at T0, T1 or Tis TNM stage), N0 to N2 and TNM 0 to 4. Furthermore, 82 negative control samples were obtained from another 3 studies in the same network (Supplementary file 6). These nodes included individuals aged 26 to 83 years from both sex,

Strategy	Samples	Studies
amplicon	772	16
atac-seq	103	3
bisulfite-seq	145	4
chip-seq	113	5
mbd-seq	21	1
mirna-seq	295	8
mre-seq	96	1
ncrna-seq	100	1
other	1969	28
rip-seq	24	2
rna-seq	3883	144
ssrna-seq	51	2
targeted-capture	426	13
wga	78	4
wgs	529	7
wxs	713	19

Table 2. Technologies found in the filtered network. This table lists the number of samples, studies and high-throughput sequencing technologies found in a network filtered by weighted connections greater than or equal to 0.2.

classified as "control", "health" and "adult normal" samples. From the same network, we extracted 834 metastatic samples from 19 studies (Supplementary file 7), considering stages IV, TNM = 4 and metastasis annotation on "AJCC", "diagnosis", "organ" and "M" attributes. Together or separately, these nodes could contribute to research investigating the particularities of colorectal cancer across different stages from normal to metastatic.

In the same form that we could obtain samples from the entire network or big communities, samples could also be extracted from smaller communities formed by fewer characteristics. For example, we prospect small and homogeneous communities to obtain metastatic samples from communities 42 (12 nodes, 66 edges), 49 (113 nodes, 4156 edges), 124 (3 nodes, 3 edges), and 341 (3 nodes, 3 edges). Those communities are formed after filtering the network to keep only edges representing two or more characteristics according to section "Usage and functionalities through a colorectal cancer search: a case study" (weight >= 0.2). The nodes of community 42 (Supplementary File 8) included twelve female samples representing, four skin metastasis samples and eight bronchus metastasis samples. In community 49 (Supplementary File 9), we identified 113 metastatic samples that could be grouped by TNM stages T3 and T4 (there are no samples at stage T2 or lower), and N0 to N2. These nodes could also be grouped by cancer location, including the ascending colon (9 samples), cecum (74 samples), splenic flexure (6 samples), rectum (7 samples), and sigmoid colon (17 samples). Communities 124 (Supplementary File 10) and 341 (Supplementary File 11) consisted of six metastasis samples classified as stage IV and T3N2M1, respectively.

To expand group comparisons, an additional 20 communities, from 205 to 225 (Supplementary File 12), comprised 72 samples from patients of both sexes with late-onset colorectal cancer. Additionally, community 143 (Supplementary File 13), containing 20 polyp and 54 non-polyp samples, could be included in the study.

Table 3 presents the association between samples and their attributes, extracted using the developed computational approach to address the case study question related to metastasis. Samples were extracted from the entire filtered network or its communities and organized according to characteristics that could be used to construct comparison groups, such as cancer or pre-cancer stages. Because communities were extracted from the filtered network, the same sample could appear in both examples.

Addressing the case study question on diagnosis and treatment (TTD approved drugs)

Focusing on diagnosis and treatment, we evaluated the new network for node samples connected by TTD approved drugs. We found 1,316 samples from 37 studies (Supplementary File 14) linked by one or more of 24 different substances. These samples could be classified using attributes like cancer or lesion type and stage ("AJCC" and "CRC grade" attributes), and 801 nodes could be further sub-grouped by patient age and sex. Additionally, 145 samples from 5 studies were associated with drug resistance to 6 TTD drugs (Supplementary File 15). Researchers could use these nodes to compare sensitive and resistant groups to elucidate drug resistance mechanisms. Normal and sensitive samples could be used to extract data specific to resistance profiles.

Navigating through the network, we identified communities constructed by a single fluorouracil resistance study. The samples from this study comprised communities 2, 3, 4, 8, 22, and 25 (Supplementary File 16), consisting of 54 sensitive and resistant sample nodes. Additionally, nodes from communities 47, 53, 55, 90, 98, 104, 109, 112, and 129 (Supplementary File 17) could be utilized to increase the number of samples, thereby enhancing the statistical power of potential findings.

Similarly, 58 cetuximab study samples were split into three parts, forming communities 7, 9, and 40 (Supplementary File 18). These samples could also be combined with nodes from community 399 (Supplementary

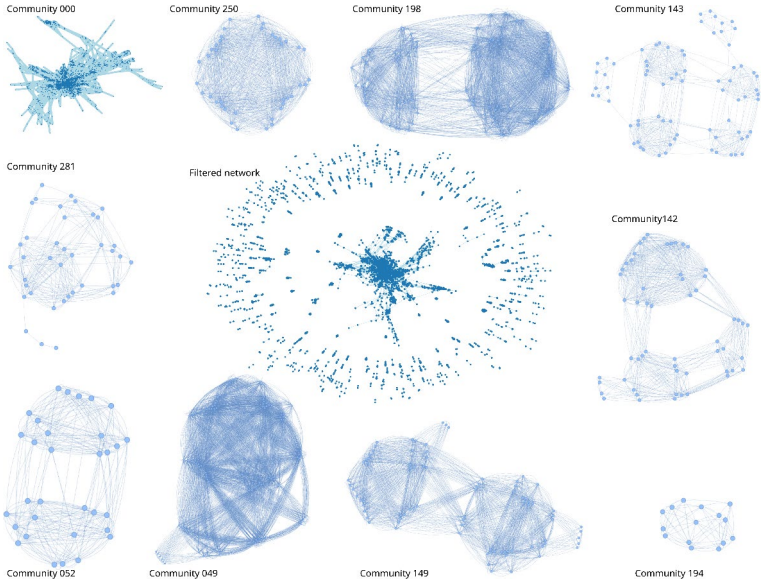

Fig. 4. Top 10 networks in the ranking. The 10 best results from a search in the local database using 120 attributes capable of describing the pathology, samples, and patients. In the center is the initial network after removing all edges with a weight less than 0.2. The surroundings networks display the 10 highest ranked communities, considering the number of connections, attributes, and values. Number of nodes: main network: 8,991; community 000: 6,195; community 250: 40; community 198: 96; community 143: 74; community 281: 39; community 142: 56; community 052: 32; community 049: 113; community 149: 84, and community 194: 15. Networks generated using the PyVis library (version 0.3.2) and NetworkX (version 3.3). Image generated using the Inkscape software version 1.2.

File 19), which is a study involving 10 drugs for treatment choices. Imatinib analysis could be performed by evaluating nodes from communities 13, 14, 16, 24, 33, 44, 46, 47, and 387 (Supplementary File 20).

Table 3 presents the association between samples and their attributes, extracted using the developed computational approach to address the case study question on diagnosis and treatment with Therapeutic Target Database (TTD) approved drugs.

Usage and functionalities through a acute lymphoblastic leukemia search: another case study

Acute Lymphoblastic Leukemia (ALL) is a genetically heterogeneous and complex disease. While it can affect both children and adults, it is more frequently observed in children under the age of five. Over the years, significant advances have been achieved, including the potential for cure in some cases. However, prognosis remains highly variable depending on the type/subtype and age group of the patients, often requiring combination drug therapies to optimize treatment outcomes. With advancements in disease characterization, patient stratification, management, and diagnosis increasingly rely on the genetic and molecular profiles of individuals. Nevertheless, routine genetic screening is not recommended, as many individuals with features considered risk factors will not necessarily develop the disease[54,55].

Leukemias are clonal disorders arising from genetic alterations and transformations in hematopoietic progenitor cells during their natural maturation process. These alterations can occur at various stages, involving mature, immature, or precursor cells. The disease typically originates from a single lineage and can stem from either B or T cells. The majority of cases, approximately 85%, originate from the B-cell lineage, which may include precursor or intermediate cells. The remaining 15% result from disruptions in the maturation process of the T-cell lineage, impacting differentiation in the thymus. T-cell leukemias generally have a worse prognosis compared to those of B-cell origin. However, with advances in patient stratification, the prognosis for T-cell leukemias can be comparable to that of B-cell leukemias, provided they are appropriately stratified and treated[56].

Continuing with the tests, we constructed a second database using sequencing metadata related to acute lymphoblastic leukemia. We executed the computational methodology developed in a manner similar to that described in the case study involving CRC. To run the scripts, we used the search terms '(acute lymphoblastic leukemia) AND "Homo sapiens"[orgn: __txid9606]'. Similarly to what occurred during the processing of CRC metadata, a relational database was constructed, containing 10,741 experiment packages structured into two relational schemas with 129 tables (three fewer than in the other theme).

After constructing and indexing the database (as described in the previous case study and methodology), we utilized 175 attributes (the drug name of consolidated drug samples and all 174 attributes of consolidated sample mining terms) (see the table in supplementary file 21). The processing result was a network containing 8,220 nodes, 33,780,090 edges, and 71,851 different types of connections, forming a large community. A total of 107 distinct weights, ranging from 0.0088 to 1, were identified. In total, the network was composed of 3,655 RNA-seq

samples (196 studies), 503 WXS samples (12 studies), and 4,210 samples (238 studies) from other technologies (see supplementary file 22).

As a result, we analyzed the main network, searching for features that could stratify patients based on disease stages, clinical data, drug usage, and data derived from RNA-seq sequencing. The selected features were "age," "cell line," "cell subtype," "condition," "diagnosis," "disease," "disease stage," "drug," "ethnicity," "gender identity," and "tissues." For each of these attributes, we extracted the corresponding subsets to evaluate which other features co-occurred within the network. The attributes "existing study type," "qualifier name," "registry number," and "title" (It is important to note that the "title" field did not correspond to the original title of the sample or study but rather to terms that were mined from each of these data sources) were excluded from the analysis as they were minimally informative or did not add value to the results.

The first characteristic assessed was the age of the patients. A total of 486 samples (from 33 studies) were included, with ages ranging from 1 to 56 years. Among these, there were samples labeled as "adult," "pediatric," and some recorded in months, likely due to the higher incidence in children compared to adults. The sample group comprised individuals of both sexes, with 176 classified as "female" and 133 as "male." Regarding disease profile ("disease" and "subtype"), the 79 individuals were classified as "acute lymphoblastic leukemia or ALL or common" (8 with "onset"), 101 as "T-cell acute lymphoblastic leukemia," 104 as "B-cell acute lymphoblastic leukemia or B-ALL or B-cell acute lymphoblastic leukemia (AML)," 84 as "T-cell acute lymphoblastic leukemia," 26 as "childhood T acute lymphoblastic leukemia," and 2 as "LL-T." Additionally, 14 samples were classified as "relapse," 9 as "diagnosis," and 5 as "de novo diagnosis." The majority of the samples were obtained from blood and bone marrow (Supplementary file 23).

The characteristic of "cell line" classified 162 samples (22 studies) into 15 distinct categories: 21 samples of "Jurkat," 20 of "RPMI-8402," 20 of "SEM," 18 of "RCH-ACV," 15 of "NALM-6," 13 of "CCRF-CEM," 12 of "MOLT-4," 11 of "NALM6," 9 of "BV-173-SHCDK6," 8 of "PF-382," 6 of "REH," 5 of "SEM cell line," 2 of "THP-1," 1 of "CEM," and 1 of "697" only. This group can be further subdivided by the attributes "cell subtype" and "cell type" into 38 "B cells," 35 "T cells," and 9 with "9:22 translocation generating BCR-ABL1 fusion protein." A total

Question	Sample	Studies	Description	Network	File
Molecular mechanisms and metastasis	2,737	85	Different cancer stages, from normal to metastatic scenarios	Filtered network	Supplementary file 3
	74	1	Polyps and normal samples	Filtered network	Supplementary file 4
	827	5	Samples age, stage, ethnicity, sex, heath and therapy	Filtered network	Supplementary file 5
	82	3	Negative control samples	Filtered network	Supplementary file 6
	834	19	Metastatic samples	Filtered network	Supplementary file 7
	12	2	Metastatic samples, skin and bronchus	Community 42	Supplementary file 8
	113	1	Metastatic samples with cancer location	Community 49	Supplementary file 9
	3	1	Metastatic samples. Female cancer stage IV	Community 124	Supplementary file 10
	3	1	Metastatic samples. Male cancer TNM T3, N2, M1	Community 341	Supplementary file 11
	72	1	Late-onset colorectal cancer samples	Communities 205 to 225	Supplementary file 12
	74	1	Polyps and normal samples	Community 143	Supplementary file 13
Diagnosis and treatment	1,316	37	One or more from 24 different substances	Filtered network	Supplementary file 14
	145	5	Drug resistance to 6 TTD drugs	Filtered network	Supplementary file 15
	54	1	Fluorouracil resistance	Communities 2, 3, 4, 8, 22 and 25	Supplementary file 16
	19	4	Fluorouracil treatment	Communities 47, 53, 55, 90, 98, 104, 109, 112 and 129	Supplementary file 17
	58	1	Cetuximab resistance	Communities 7, 9 and 40	Supplementary file 18
	35	1	10 drugs for treatment choice	Community 399	Supplementary file 19
	109	9	Imatinib analysis	Communities 13, 14, 16, 24, 33, 44, 46, 47 and 387	Supplementary file 20

Table 3. Association between nodes samples and questions of study case. This table lists the number of samples of networks and communities extracted by this computational methodology. The filtered network is the result of a filter applied to the entire network to remove all edges with a weight < 0.2, which represents the sharing of only a single characteristic. Samples of interest were selected based on their characteristics as candidates for the response to molecular phenomena for the characterization of controls, cancer stages, diagnosis, or treatment resistance. The samples were evaluated either from the entire filtered network or from communities, which are groups of individuals connected to each other but disconnected from the rest of the network. Communities were extracted from the filtered network so that the same samples could remain in both groups.

of 110 individuals were classified, with ages ranging from 24 h to 16 years. 95 samples have annotations under the characteristic "disease," the majority of which are classified as "ALL" (48) and "T-cell acute lymphoblastic leukemia" (24). Additionally, the majority of the samples are from "female" (89) compared to "male" (15). This group also includes 28 samples (24 controls) associated with the "NSD2 p.E1099K" mutation. Most of the samples were obtained from blood (23 "blood" and 71 "peripheral blood") and bone marrow (12) (Supplementary file 24).

31 samples (from 2 studies) were grouped by the characteristic "condition", involving REH and Jurkat cells for studies comparing EV (4 controls samples), DUX4 (4 samples), DUX4-IGH (4 samples), DUX4-DEL50 (4 samples), and PSIP1_KD (15 samples, 5 of which are controls). 16 of the 31 samples were treated with "doxycycline." The remaining 15 samples were annotated with the genotype "wt" (Supplementary file 25).

The characteristic "diagnosis" encompasses 564 samples from 8 studies, derived from various sources: 259 samples of "pediatric acute lymphoblastic leukemia" (sourced from "bone marrow," "peripheral blood," and "peripheral blood CD19+"), 231 of "B-lineage ALL" ("bone marrow" and "blood" tissues), 31 from "bone marrow" ("bone marrow" and "bone marrow tumor"), 15 from "peripheral blood," 8 from "blood," and 3 from "pre-B leukemic cell line" (classified under the attributes "source names" and "tissues"). Among these samples, 142 are annotated as "male" and 124 as "female." Furthermore, 46 samples are labeled as "discovery" and 41 as "replication" under the "cohort" attribute. A total of 86 samples were obtained from patients aged 1 to 86 years, with some labeled simply as "pediatric" or "adult" (Supplementary file 26).

The "diseases" category contains 931 samples from 57 studies, with the following classifications: 143 samples of "acute lymphoblastic leukemia" (ALL), 37 of "acute lymphoblastic leukemia" (T-ALL), "T-cell acute lymphoblastic leukemia," or "T ALL", 2 of "T-cell lymphoblastic leukemia", 5 of "acute lymphocytic leukemia; non-T; non-B", 198 of "B-cell acute lymphoblastic leukemia" (B-ALL), 2 of "B-cell acute lymphoblastic leukemia (AML)", 107 of "childhood acute lymphoblastic leukemia" or "childhood B-cell acute lymphoblastic leukemia," 36 of "childhood T acute lymphoblastic leukemia", 17 of "cortical T ALL", 4 of "EBV infection" or "Epstein-Barr virus transformed cells," 15 of "healthy," "healthy donor," or "normal", 13 of "immature T ALL", 3 of "KMT2A-r acute lymphoblastic leukemia", 54 of "KMT2A-rearranged infant ALL", 8 of "leukemia", 5 of "mature T ALL", 15 of "pediatric B-cell acute lymphoblastic leukemia", and 88 of "precursor B-cell acute lymphoblastic leukemia".

These sequences were primarily obtained from samples of "blood" (249) and "bone marrow" (86), followed by "cell line" (58), and other sources (annotated under the "source names" attribute). A total of 204 samples were classified by developmental stage ("dev stage") as "juvenile" or "juvenile stage" (100), "adult" (55), and "child" or "infant" (39). Among the individuals, 85 were neither controls nor treated during collection, while 32 were treated with "DMSO," 42 with "Dexamethasone," 18 with "Azacitidine," 18 with "Decitabine," and 22 with other treatments (Supplementary file 27).

The "disease stage" attribute includes 89 samples from 6 studies, classified as follows: 27 "de novo diagnosis," 23 "relapse," 23 "diagnosis," 6 "remission," 4 "primary," 2 "baseline," and 4 "healthy." Among these, 46 samples are from "male" and 31 from "female" individuals. Disease classifications include 23 samples from "acute lymphoblastic leukemia," 20 from "B-ALL," 12 from "T-cell acute lymphoblastic leukemia," 8 from "pediatric B-cell acute lymphoblastic leukemia," 4 from "childhood B-cell acute lymphoblastic leukemia," and 2 from "EBV infection." Additionally, 20 samples were exposed to "Iron," 12 to "Omaveloxolone," and 1 to "Blinatumomab" (Supplementary file 28).

The "ethnicity" attribute includes 37 samples from 3 studies, comprising 27 "Latin," 8 "Hispanic," and 2 "Caucasian" individuals of both sexes (19 "male" and 18 "female"). These samples were derived from "bone marrow" (27 samples), "blood" (8 samples), and "peripheral blood" (2 samples). Eight samples were subjected to "immunosuppressive" chemotherapy treatment. Seven individuals were between 2 and 19 years of age (Supplementary file 29).

The "gender identity" attribute characteristics grouped 710 samples, including 356 "female," 354 "male," and 1 "not determined." This group primarily includes samples from "tissues," with the majority derived from "bone marrow" (385), "peripheral blood" (101), and "blood" (30). A total of 309 individuals are between 17 months and 56 years of age. Within this group, 218 samples were classified as "9836/3 - pre-B ALL" and 36 as "9837/3 - pre-T ALL." The "disease" attribute further grouped 65 samples as "T-cell acute lymphoblastic leukemia," 94 as "ALL," 26 as "childhood T acute lymphoblastic leukemia," 23 as "B ALL," 5 as "acute lymphocytic leukemia; non-T; non-B," 2 as "B-cell acute lymphoblastic leukemia (AML)," 4 as "healthy," and 2 as "EBV infection" (Supplementary file 30).

To address studies related to treatment response, 1,173 samples from 79 studies were mapped under the "drug" attribute. The most frequent items include "iron" (381), "dexamethasone" (145), "prednisolone" (144), "decitabine" (104), "ethanol" (74), "azacitidine" (66), "doxycycline" (55), "romidepsin" (54), "mercaptopurine" (45), "cytarabine" (45), along with 39 other substances comprising a total of 550 samples. These samples can be reorganized according to "source names" (1,026 samples), "construct therapeutics" (530 samples, including 108 controls), "tissues" (355 samples), "disease" (289 samples, including 15 controls), and "genotype" (238 samples, including 38 controls) (Supplementary file 31).

Landscape of similar algorithms

In this study, we developed a computational package designed to facilitate the localization of sequencing data deposited in the SRA database through the metadata processing. The complete workflow involves directly querying metadata from the SRA, downloading the metadata, searching for publications associated with the identified experiment packages, constructing a local relational database, indexing both structured and unstructured data, and clustering samples by comparing similar detected attributes. The package can be utilized through Python language, command-line scripts and a graphical user interface developed with web technologies compatible with any modern browser.

This approach provides the capability to search for sequencing projects deposited in the SRA, grouped by shared characteristics, thereby enabling the combination of similar samples from different studies. This contributes to enhancing the statistical robustness of analyses related to the topic of interest. The scientific community has developed other tools to assist in the search for data and metadata in the SRA.

SRAdb is a package for the R programming language that provides relational databases using SQLite. In its development, the authors employed PHP to build a parser for processing SRA data. During execution, five relational tables are created: "experiment", "sample", "submission", "study", and "run". Using this package, users can perform searches through SQL queries or conduct "Google-like searches" using the fts3 protocol. The output retains the original structure retrieved from the SRA with minimal modifications, applied only to adapt the data to the relational model[19]. While SRAdb provides a relational database with structured metadata from the SRA, our package, in contrast, creates on-demand relational databases and indexes all data, normalizing it through the use of controlled vocabularies, resolving plurals, synonyms, and expressions formed through word combinations, among other methods. Furthermore, although our primary focus is on sample acquisition, our tool also processes data from all SRA structures. It includes entity recognition of unstructured data.

The pysradb is a Python package designed for querying next-generation sequencing metadata. The authors implemented a command-line interface to search for and download metadata from the SRA. This functionality relies on normalized data from SRAdb, leveraging libraries such as Pandas[57] to display metadata and facilitate sample downloads. Users can perform searches using terms and keywords to locate metadata of interest efficiently[58]. By using data from SRAdb, the results of pysradb are limited to what is available in the SRA without any additional normalization. Although users benefit from easy access to data in a relational format, they still need to address annotation discrepancies and the detection of data, which is often provided in an unstructured format. In contrast, our computational package retrieves, normalizes, and indexes the data, presenting it to the user with hierarchical search options in a graphical interface.

MetaSeek is a structured metadata database accessible via a Web GUI and API. It was developed using Python and React and is hosted on an Amazon EC2 server. The service undergoes weekly updates, during which it evaluates, filters, cleans, and predicts SRA metadata. Users can search the database using controlled vocabulary terms directly on the website, save their findings, and later export their search results for further analysis[59]. Therefore, users can obtain data that is normalized through controlled vocabularies, which facilitates the retrieval of metadata in the SRA. In our package, we also include sample grouping based on the presence of similar characteristics, allowing users to locate metadata, even from studies by different authors, in a consolidated manner.

MetaRNA-Seq is a standardized and semi-automatically curated database that concentrates RNA-Seq data from the SRA. It was built by implementing the Entrez API and utilizing the GEOmetadb framework to centralize SRA data into a relational database using MySQL. The tool was developed using Java EE 7, the VAADIN framework, the Glassfish web server, and a simplified text mining implementation using a regular expression model. The database is updated quarterly, and its focus, unlike the SRA search interface, is on studies rather than experiments. The tool offers various methods for textual or numerical searches. Database performance is ensured through prior indexing[30]. In our computational package, the focus of the analyses is directed towards the samples; however, all data from other entities are assessed in an interconnected manner as complementary data, allowing users to locate samples based on data from any other tables.

Another tool, grabseqs, is a command-line utility for downloading NGS data and metadata from the SRA, MG-RAST, and iMicrobe databases. It was developed in Python and is available via GitHub, Conda, and PyPI. With this tool, users can directly download data and metadata from all three databases. The data is provided in the formats offered by each database without any conversion[31]. This tool is excellent for the rapid retrieval of data, including from other databases; however, the data are obtained and presented in the original formats of each database. In this study, we provide in the computational package the ability to locate data in a normalized form with controlled vocabulary, obtained from three different data dictionaries, structuring everything in a relational model. Moreover, a graphical interface aids in the construction of queries, displaying all available attributes from the centralized database.

MetaSRA is a standardized database inspired by the ENCODE project, which combines biomedical ontologies, metadata from the SRA obtained through SRAdb, non-statistical machine learning, natural language processing, and Text Reasoning Graph to store and index human RNA-seq metadata from the Illumina platform. The database implements a pipeline for data acquisition, preparation, cleaning, and prediction. The tool is available on GitHub, and a web interface allows users to query SRA data[60]. Similarly, we also incorporate natural language processing and data normalization. However, the normalization in our computational package is based on three reference databases: MeSH[61], TTD[62], and WordNet[63]. Furthermore, we provide a search interface that displays to the user query possibilities using terms that have been indexed in the local database, thereby expanding the search to terms that the user may not have initially considered due to the lack of standardization adopted by SRA users.

In this study, we provide a computational package written in Python that can be used programmatically or via command line. Additionally, we have developed an interface to assist with the search process. Users can install the package on their local systems and build their own custom databases to aid in the search for NGS sequencing samples from any platform deposited in the SRA. In our process, in addition to making a significant effort to normalize the data, we utilize the MeSH, TTD, and Wordnet databases as controlled vocabularies. We have incorporated a sample clustering system to identify similar samples, aiming to provide groups of individuals that share similar characteristics to build comparison groups. This is demonstrated in biomarker research, but it has the potential for use in various fields, thereby assisting specialists in their analyses with a larger number of samples.

However, the profound differences in design philosophies, supported use cases, and underlying technologies among these tools preclude a direct performance comparison. Instead, we focus on highlighting their distinguishing features and how they complement the specific needs of diverse research workflows. By elucidating these differences, we underscore the value of our tool in addressing unmet needs, particularly in metadata normalization, integration, and on-demand database creation to support advanced analyses, such as biomarker discovery in cancer research.

Methods

Computational methodology development

All the methodology discussed here was tested and implemented as a computer package using Python programming language[64] version 3.12.4 and the Integrated Development Environment (IDE) PyCharm Community Edition version 2024.1.3. The research and retrieval of data from the Sequence Read Archive (SRA) and Pubmed databases were implemented through the Entrez Programming Utilities Help (E-utilities) release February 14, 2014 [65] and obtained in XML (eXtensible Markup Language) format. The DBeaver software version 22.1.2 was used to facilitate the visualization, test and maintenance of database structures. We run and test all modules and scripts into a computer with AMD Ryzen™ 5 5600G with Radeon Graphics processor having 12 threads, 48 GB memory ram and 1 SSD Kingston NVMe and a server with AMD Opteron™ Processor 6376 having 64 threads, 252 GB memory ram and 1 RAID HDD 2.74 TB.

The entire process of grouping sequence samples in this computational methodology involves two steps: the construction of a local relational database and querying it to select groups of samples that share the same characteristics. The database construction needs a SRA query search string and an internet connection. After receiving the query string, using E-utilities programs to access NCBI databases, the package searches on the SRA database for all experiments packages that could satisfy the question, then Pubmed publications are included on results. The entire results file is processed and scripts for create and charge databases are constructed and executed on DBMS. This process results in a local database, implementing the relational model, containing all SRA metadata describing interesting sequence experiments.

The local database is indexing using natural language processing (NLP) techniques such as splitting text into sentences and words (tokenization), extracting root forms of words to improve comparison (lemmatization)[66], syntax detection, n-grams (contiguous sequences of words of size n) and the identifying important information and terms (entity recognition)[67]. The MeSH[61] and WordNet[63] databases are used as a dictionary of Medical and English terms, respectively.

The Fig. 5 shows the process of data acquisition, indexing, local database construct and data storage.

Visit the "Code availability" and "Data Availability" sections for access to the complete source code and the results generated in this study.

Data acquisition

The ESearch, EFetch, and ELink programs, part of the E-utilities suite (release February 14, 2014)[65], were used for searching, downloading, and cross-referencing between databases. Using ESearch, searches were conducted in the SRA for sequencing studies related to human colorectal cancers. The metadata, describing all studies publicly accessible in the database, were transferred and stored in a temporary XML file.

Subsequently, ELink was used to search for corresponding publications deposited in PubMed. The located data were included in the temporary XML[68] file along with their respective studies. The related publications are obtained by extracting PubMed IDs (PMIDs) found within the experiment packages contained in the XML file. The data specification files in XSD (XML Schema Definition)[68] were obtained from NCBI documentation[69] and used to help part of the development of the Python scripts and modules.

Centralized database

Using the PostgreSQL Database Management System (DBMS) (version 15.8), a relational database to store all metadata and processing results obtained throughout the study was built. The metadata and data derived from text and data mining (detailed in the subsequent section) were organized into two schemas within the database to facilitate handling and organization. The modeling, construction, and implementation of the database were automated through scripts and modules written in Python 3 (version 3.12.4). The connection with python modules and DBMS are performed with Psycopg2 (Psycopg version 2) library.

The translation of the structure stored in the XML file to the relational model was achieved through recursive calls of methods and functions developed to map entities, keys and relations. The processing of the entire file resulted in Data Definition Language (DDL) scripts used to construct the database structure in the DBMS. Similarly, scripts SQL (Structured Query Language) were generated for loading the data into the database. During this stage, data was extracted from the XML file using XPath (XML Path Language) and stored into the locally relational database. The relational schema representing the constructed automated database, including all table columns, is provided in Supplementary Files 32, 33, and 34.

Data and text mining

Three distinct databases were used to guide the identification of words, terms, and expressions: The WordNet (version 3.0) lexical dictionary of the English language[63,70], the controlled vocabulary dictionary Medical Subject Heading (MeSH) (version 20240101)[61,71], and The Therapeutic Target Database (TTD) (released on January 10th, 2024)[62].

The databases were used to detect English words and expression (WordNet), biomedical or health-related terms (MeSH), and approved drug names (TTD). Data from Wordnet were obtained using the Natural Language Toolkit (NLTK) (version 3.8.1) library[33], while data from MeSH and TTD were directly transferred from their

respective NIH servers using HTTP (Hypertext Transfer Protocol). MeSH data downloaded in ASCII (flat files) format, and TTD data were downloaded in txt (raw format) files.

Methods for natural language processing were implemented to separate sentences and words ("tokenization"), identify grammatical classes, synonyms, hypernyms, and detect entities. The metadata was organized into structured and unstructured data and processed separately.

Unstructured data were first divided into sentences and then into words. Each term in the detected sentences was tagged with its respective grammatical class for entity detection. Entity detection was supported using n-grams (a contiguous sequence of n items in a sentence, such as phonemes, syllables, letters, words, among others[72]) (Fig. 6), with the initial value of n set to the number of words in the longest term in the MeSH dictionary, decreasing by one at each execution step. All n-grams and their respective lemmas (original word in infinitive, singular, and masculine forms) were compared with the dictionary. The process continued until unigrams (n-grams of a single word, $n = 1$) were obtained. Longer matches were prioritized over the short ones. Similarly, structured data were compared using synonyms and lemmas.

After processing, tables were dynamically constructed to assist in visualizing the characteristics contained in the database. During this stage, synonymous terms were grouped under a single term. The results were then inserted into the local database, prioritizing the MeSH term when available.

Data indexing

After processing the structured and unstructured data, auxiliary tables were created containing all data for each SRA theme. At this stage, data were standardized so that synonymous terms were stored consistently, normalized using WordNet, TTD, and MeSH dictionaries. The normalization process follows three steps. Firstly, all terms are compared using lemmatization (extraction root form of words) and n-grams (contiguous sequences of words of size n) techniques evaluating plural forms, synonyms, verb tense, word position, among others, using WordNet as an English dictionary. Secondly, data is translated into MeSH terms. Thirdly, data are compared to approved drugs table extract from TTD data.

Attributes with complementary fields were subdivided into new attributes. Values from the tag and value columns, such as those in the sample_atributes table, allowed for free-form annotations combining key-value pairs. These fields were tagged and restructured into subtables and subsequently grouped by synonymous values. Elements identified through text and data mining, along with metadata and references for each sample, were consolidated into an additional table. A visualization and search interface was implemented using Django (version 4.2.11) framework[73]. This interface included a representation of the relational database and options to resolve synonyms that could not be automatically resolved. After implementing the interface, ambiguities were addressed and recorded in auxiliary tables in the local database.

Using the Python (version 3.12.4) Standard Library[64], the PyVis[74] (version 0.3.2) and the NetworkX[75] (version 3.3) libraries, a network construction was implemented in the search algorithm. In these networks, nodes represent samples, and edges represent shared attributes and values. For network construction, it is necessary to provide as parameters one of the base tables of the database and one or more entire tables (all fields) or partial tables (one or more fields). The initial table is used as a reference, from which the identifiers used to characterize the nodes of the network will be obtained. All other fields from the other tables are compared with the characteristics of the reference table. Data from intermediate tables that may exist between the reference table and the others are included in the comparison as additional attributes. After processing all the selected and intermediate data, attribute values that are the same are used as relationships between nodes, characterizing the edges of the network. In this way, interaction networks between the entities of the main table are built. In this network, two entities (nodes) can have zero, one, or multiple undirected interactions (edges). In this process, consolidated tables (constructed using normalized terms as described in "Data and text mining" section and current section above) can be used to enhance the connectivity power of the network.

To highlight multiple to single relationships, we implemented an edge weighted. When samples shared multiple characteristics, the edge weight, represented by its thickness, was increased proportionally, prioritizing elements connected by more characteristics. The weights of each edge are defined by the sum of the number of attributes characterizing the connection, which determines the thickness of the edge and quantifies the strength of the linkage. Thus, edges with fewer attributes receive lower values, while those with more attributes receive higher weights. Finally, the weights are normalized to a range from 0 to 1. The widths of an edge in the network is given by the equation $t_e = \left(\frac{a_e}{A_{max}}\right) \times k$, where t_e is the edge width, a_e is the number of attributes describing the edge, A_{max} is the maximum number of attributes of any edge in the network, and k is a scale factor used to adjust the edges thickness visually.

Groups of elements that are interconnected but not connected to other groups are separated by the module through the processing of the entire network. After identification, each group is isolated into different files and presented to the user as a community within the network. In this work, we define a community as an interconnected group of nodes that is isolated from others, meaning there is no single node connecting this community to the rest of the network. To facilitate the location of each element, an index file is included in the results package. Finally, the entire network can be filtered by removing edges with low width or weight, using values that represent connections formed by two or more attributes according to the needs of a user study.

Intermediate data generated during searches were included in result files in formats such as CSV (Comma-separated values), HTML5 (HyperText Markup Language), JSON (JavaScript Object Notation), PNG (Portable Network Graphic), SVG (Scalable Vector Graphic), and TXT, allowing for further processing in other tools.

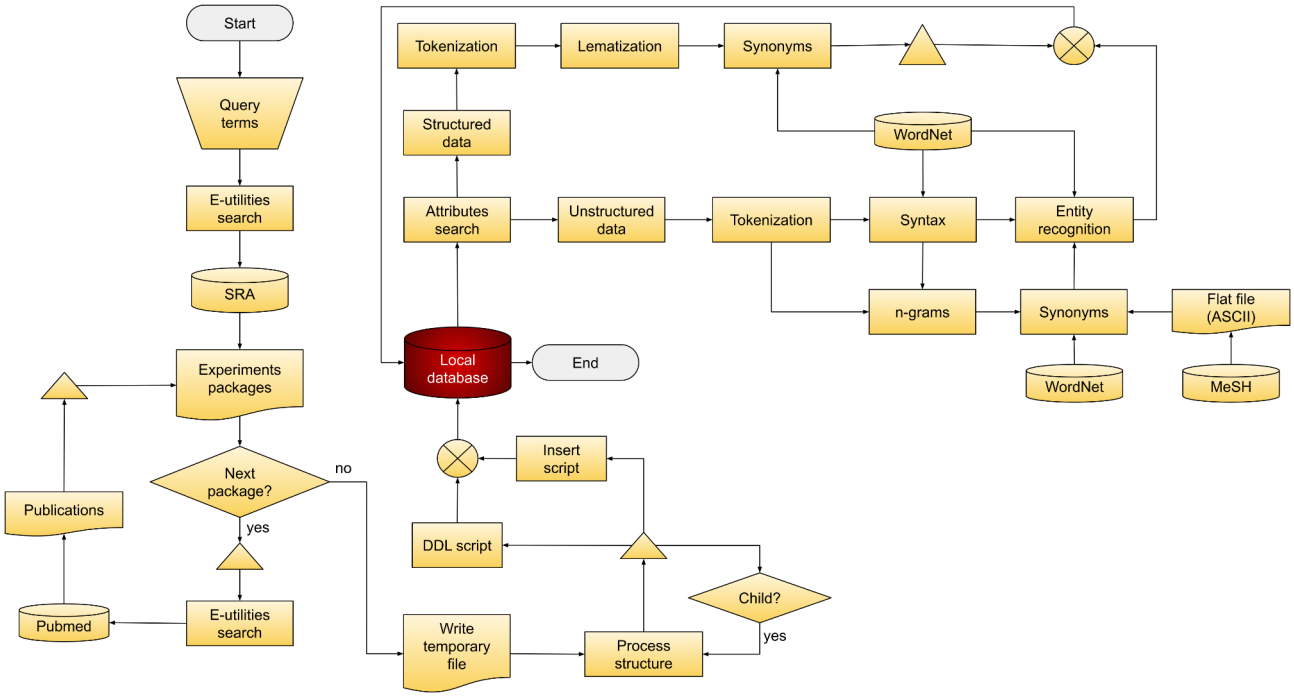

Fig. 5. Python package local database construction flowchart. After receiving user SRA query terms, the computational package performs the search into SRA through E-utilities programs. Experiments packages data retrieved from SRA are entirely processed and queries using E-utilities are executed trying to locate related Pubmed publications. All publications found are included into Experiments packages and stored into a XML temporary file. Then the temp file is processed to generate DDL and insert database scripts. The generated scripts are executed building a local database that implements the relational model. All generated database tables are processed to attribute classification. Structured and unstructured data are processed into different steps. Structured data is split into words, then words are normalized to single root forms (lemmatization). Synonyms terms are grouped and registered into the local database. Unstructured data are split into sentences and then further split into words. Then they are processed for syntax detection. N-grams with different sizes are generated and grouped by synonyms. Then entity recognition processes are started and the results of the entire process are stored into the local database. WordNet and MeSH databases are used like dictionaries of English and medical terms. WordNet data are obtained by NLTK library and MeSH data are accessed by HTTP protocol in flat file (ASCII format). E-utilities: Entrez Programming Utilities Help, SRA: the Sequence Read Archive database, DDL: Data Definition Language, Tokenization: a computational approach to split elements in tokens (words in this work), Lemmatization: a morphological transformation to recovery the root or base form of a word, Entity recognition: a Natural Language technique to detect and classify important information on text, ASCII: American Standard Code for Information Interchange, MeSH: Medical Subject Headings, WordNet: a lexical database of English, NLTK: Natural Language Toolkit. Image generated using Google Drawings software (accessed in November 2024).

Conclusions

The SRA (Sequence Read Archive) is a repository of raw high-throughput sequencing data, maintained across three instances managed by NCBI, EBI, and DDBJ, in accordance with INSDC guidelines. Using SRA data offers several advantages, including enhancing reproducibility, enabling meta-analyses, and facilitating new discoveries from previously sequenced samples. However, the lack of standardization and scarcity of metadata pose significant challenges for automating processes and utilizing tools such as machine learning, statistical models, and meta-analyses. This difficulty arises from the need to locate standardized metadata capable of comprehensively describing samples, patients, studies, and experiments. Considering this context, a new methodology has been developed to facilitate the identification, selection, categorization, and grouping of collections of massively sequenced samples deposited in the SRA. The objective is to integrate data associated with these samples and clinical patient data.

The presented computational methodology combines data and text mining, NLP, complex network, relational databases to index sample metadata from SRA Database. In the process we used MeSH, TTD and WordNet databases to group, normalize synonym data and to identify entities. All the methodology was implemented and tested into Python 3 modules including command line scripts and web interfaces to search automation.

The presented case studies, which focused on colorectal cancer (CRC) and Acute Lymphoblastic Leukemia (ALL), exemplified the effectiveness of this methodology in integrating raw sequencing data and clinical attributes. This integration is crucial to help address significant inquires that remain incompletely understood regarding colorectal cancer, including: (a) Identification of specific genetic mutations and molecular pathways

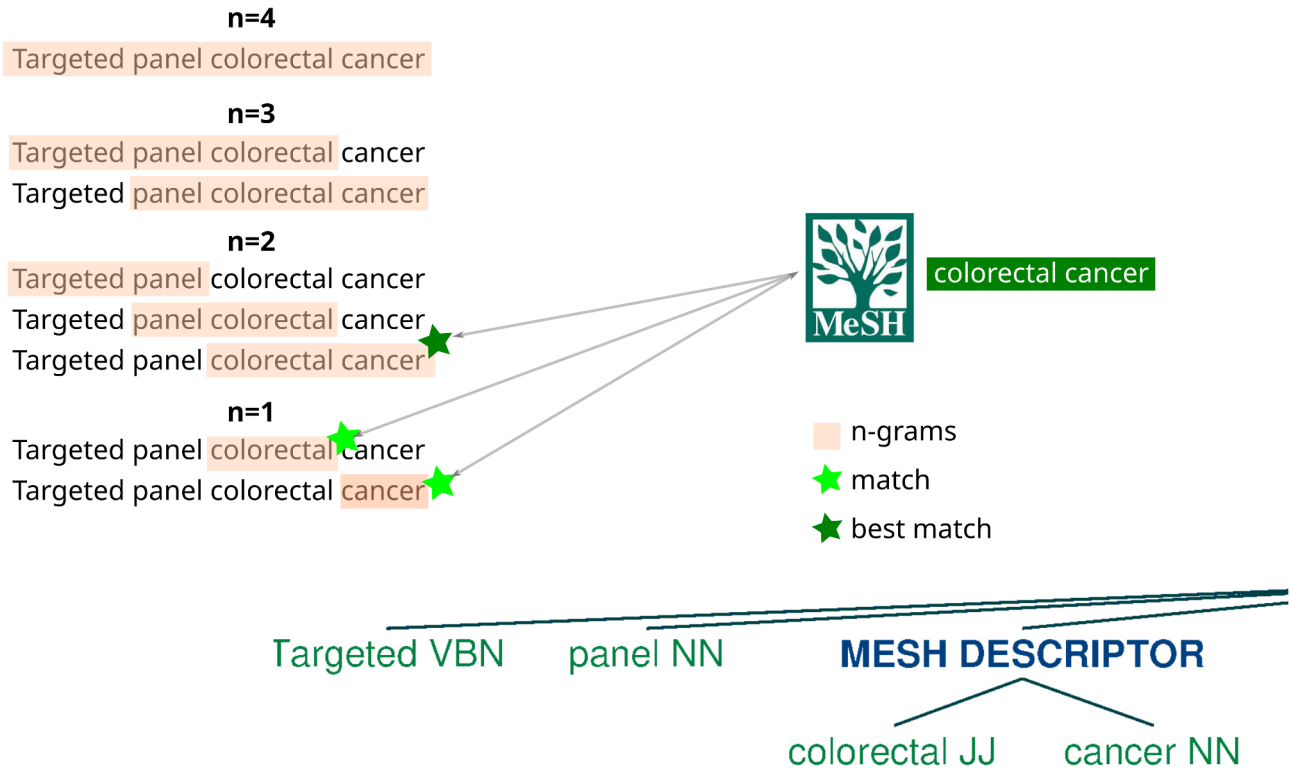

Fig. 6. Term search using n-grams. Unstructured data is split into sentences and then into words. After the grammatical class is tagged using the WordNet dictionary, we begin the entity detection process. We initiate the detection using a sliding window. To do this, we start by processing n-grams (a contiguous sequence of n items in a sentence, such as phonemes, syllables, letters or words) of all sizes, starting with n equal to the maximum number of words in the longest term in the MeSH dictionary. Then, the n value decreases by one unit, and the process is repeated while n is greater than 0. Terms formed by n-grams existing in the MeSH database are considered matches. If two matches share the same words, the longest match is kept as the best match, and the others are discarded. JJ: adjective; NN: singular noun; VBN: past participle verb. Image generated using the Inkscape software version 1.2.

responsible for initiating and driving cancer progression; and (b) Exploration of key factors contributing to colorectal cancer metastasis, particularly to the liver and lungs.

To shed light on groups associating raw data sequence and patients clinical information that can guide future studies on biomarkers, we identified groups within the interaction network comprising individuals sharing clinical characteristics across various stages of cancer, among them the chemotherapy treatment, cancer grade and stage, cancer location (left or right sides of the abdomen), affected organ and tissues, affected organ location, patient sex, tumor size and patient health state. These groups collectively can enhance the statistical power needed for more precise bioinformatic and support analyses such as differential gene expression and mutational profiling, enabling visualization of differences in each profile that could serve as diagnostic or prognostic biomarkers.

For the first question, we categorized groups of individuals into healthy, intermediate stage, and advance stage categories, with disease presence in various organs and abdomen locations. Regarding the second question, we additionally included individuals with characteristics with attributes associated with cancer stage, polyps, affected organs or regions, with affected lymph nodes, therapies utilized and metastasis process presence. In total, hundreds of samples were grouped to form potential comparison groups for evaluating colorectal cancer. In future studies, these samples could serve as a valuable resource for investigating disease progression and identifying potential therapeutic targets.

Considering the analyzed metadata and the applied exclusion criteria (only RNA-seq), we determined that the maximum number of samples analyzed per study in colorectal cancer deposited in the SRA was 579 (study accession number SRP359396), and the minimum was 1 (study accession number SRP183065) with an average of 20.73 samples per SRA-deposited study). In this analysis, we identified clusters comprising 983 samples (with 14 attributes addressing biological question 01) and 1,012 samples (with 20 attributes addressing biological question 02) and 3,655 acute lymphoblastic leukemia samples. These findings clearly illustrate the positive outcomes of the methodology developed for integrating samples from various studies, which could be further explored to address the guiding biological questions outlined in the presented case study.

Future research directions will focus on enhancing algorithmic performance and optimizing user experience. By expanding current functionalities and improving the intuitive nature of the interface, our goal is to better support researchers and experts in effectively analyzing and interpreting data.

Data availability

The developed code and all data generated or analyzed during this study can be accessed in https://ib.minas.fiocruz.br/sradatamining.

Code availability

The source code is available at https://github.com/paul-guimaraes/SRADatabaseNavigator.

Received: 18 July 2024; Accepted: 24 February 2025
Published online: 08 March 2025

References

1. Stephens, Z. D. et al. Big data: astronomical or genomical?? *PLoS Biol.* **13**, e1002195 (2015).
2. Goodwin, S., McPherson, J. D. & McCombie, W. R. Coming of age: ten years of next-generation sequencing technologies. *Nat. Rev. Genet.* **17**, 333–351 (2016).
3. Lapatas, V., Stefanidakis, M., Jimenez, R. C., Via, A. & Schneider, M. V. Data integration in biological research: an overview. *J. Biol. Res-Thessaloniki* **22**, 9 (2015).
4. Wanichthanarak, K., Fahrmann, J. F., Grapov, D. Genomic, proteomic, and metabolomic data integration strategies. *Biomarker Insights* **2015**, 10s4 (2015).
5. Gligorijević, V. & Pržulj, N. Methods for biological data integration: perspectives and challenges. *J. R Soc. Interface* **12**, 20150571 (2015).
6. Xue, B., Khoroshevskyi, O., Gomez, R. A. & Sheffield, N. C. Opportunities and challenges in sharing and reusing genomic interval data. *Front. Genet.* **14**, 1155809 (2023).
7. Rustici, G. et al. Transcriptomics data availability and reusability in the transition from microarray to next-generation sequencing. 12.31.425022 Preprint at (2020). https://doi.org/10.1101/2020.12.31.425022 (2021).
8. MINSEQE. Minimum Information about a high--throughput Nucleotide SeQuencing Experiment—a proposal for standards in functional genomic data reporting. https://cdn.elifesciences.org/articles/48958/elife–48958-repstand1-v2.pdf (2023).
9. Zheng, H. et al. Comprehensive review of web servers and bioinformatics tools for Cancer prognosis analysis. *Front. Oncol.* **10**, 896 (2020).
10. Li, T. et al. TIMER: A web server for comprehensive analysis of Tumor-Infiltrating immune cells. *Cancer Res.* **77**, e108–e110 (2017).
11. Wang, X., Hu, S., Ji, W., Tang, Y. & Zhang, S. Identification of genes associated with clinicopathological features of colorectal cancer. *J. Int. Med. Res.* **48**, 300060520912139 (2020).
12. Győrffy, B., Surowiak, P., Budczies, J. & Lánczky, A. Online survival analysis software to assess the prognostic value of biomarkers using transcriptomic data in Non-Small-Cell lung Cancer. *PLOS ONE* **8**, e82241 (2013).
13. Tang, Z. et al. GEPIA: a web server for cancer and normal gene expression profiling and interactive analyses. *Nucleic Acids Res.* **45**, W98–W102 (2017).
14. Clark, A. J. & Lillard, J. W. A comprehensive review of bioinformatics tools for genomic biomarker discovery driving precision oncology. *Genes* **15**, 1036 (2024).
15. Liu, Z. et al. BEST: a web application for comprehensive biomarker exploration on large-scale data in solid tumors. *J. Big Data* **10**, 165 (2023).
16. Alshawaqfeh, M., Rababah, S., Hayajneh, A., Gharaibeh, A. & Serpedin, E. MetaAnalyst: a user-friendly tool for metagenomic biomarker detection and phenotype classification. *BMC Med. Res. Methodol.* **22**, 336 (2022).
17. Terkelsen, T., Krogh, A. & Papaleo, E. CAncer bioMarker Prediction Pipeline (CAMPP)—a standardized framework for the analysis of quantitative biological data. *PLoS Comput. Biol.* **16**, e1007665 (2020).
18. Netanely, D., Stern, N., Laufer, I. & Shamir, R. PROMO: an interactive tool for analyzing clinically-labeled multi-omic cancer datasets. *BMC Bioinform.* **20**, 732 (2019).
19. Zhu, Y., Stephens, R. M., Meltzer, P. S. & Davis, S. R. SRAdb: query and use public next-generation sequencing data from within R. *BMC Bioinform.* **14**, 19 (2013).
20. Van den Broeck, L. et al. Functional annotation of proteins for signaling network inference in non-model species. *Nat. Commun.* **14**, 4654 (2023).
21. Austin-Tse, C. A. et al. Best practices for the interpretation and reporting of clinical whole genome sequencing. *Npj Genom Med.* **7**, 1–13 (2022).
22. Qi, T., Song, L., Guo, Y., Chen, C. & Yang, J. From genetic associations to genes: methods, applications, and challenges. *Trends Genet.* **40**, 642–667 (2024).
23. Balakrishnan, R., Harris, M. A., Huntley, R., Van Auken, K. & Cherry, J. M. A guide to best practices for Gene Ontology (GO) manual annotation. *Database* **2013**, bat054 (2013).
24. Barrett, T. et al. NCBI GEO: archive for functional genomics data sets—update. *Nucleic Acids Res.* **41**, D991–D995 (2013).
25. The Cancer Genome Atlas Program (TCGA). NCI. https://www.cancer.gov/ccg/research/genome-sequencing/tcga (2022).
26. Sayers, E. W. et al. Database resources of the National center for biotechnology information. *Nucleic Acids Res.* **47**, D23–D28 (2019).
27. The Sequence Read Archive (SRA). Getting Started. https://www.ncbi.nlm.nih.gov/sra/docs/ (2022).
28. Sanitá Lima, M. & Smith, D. R. Don't just dump your data and run. *EMBO Rep.* **18**, 2087–2089 (2017).
29. Leinonen, R., Sugawara, H. & Shumway, M. The sequence read archive. *Nucleic Acids Res.* **39**, D19–D21 (2011).
30. Kumar, P. et al. MetaRNA-Seq: an interactive tool to browse and annotate metadata from RNA-Seq studies. *BioMed Res. Int.* **2015**, 318064 (2015).
31. Taylor, L. J., Abbas, A. & Bushman, F. D. Grabseqs: simple downloading of reads and metadata from multiple next-generation sequencing data repositories. *Bioinformatics* **36**, 3607–3609 (2020).
32. Cuddihy, T. et al. SRA down under: cache and analysis platform for infectious disease. *Stud. Health Technol. Inf.* **266**, 76–82 (2019).
33. NLTK:. Natural Language Toolkit. https://www.nltk.org/ (2022).
34. Bazoge, A., Morin, E., Daille, B. & Gourraud, P. A. Applying natural language processing to textual data from clinical data warehouses: systematic review. *JMIR Med. Inf.* **11**, e42477 (2023).
35. The Cancer Genome Atlas Research Network. The cancer genome atlas pan-cancer analysis project. *Nat. Genet.* **45**, 1113–1120 (2013).
36. Zhao, Z. et al. Comprehensive RNA-seq transcriptomic profiling in the malignant progression of gliomas. *Sci. Data.* **4**, 170024 (2017).

37. Tseng, G. C., Ghosh, D. & Feingold, E. Comprehensive literature review and statistical considerations for microarray meta-analysis. *Nucleic Acids Res.* **40**, 3785–3799 (2012).
38. Wan, Y. W. et al. Meta-Analysis of the Alzheimer's disease human brain transcriptome and functional dissection in mouse models. *Cell. Rep.* **32**, 107908 (2020).
39. Hong, F. & Breitling, R. A comparison of meta-analysis methods for detecting differentially expressed genes in microarray experiments. *Bioinformatics* **24**, 374–382 (2008).
40. Hu, P., Greenwood, C. M. T. & Beyene, J. Statistical methods for meta-analysis of microarray data: a comparative study. *Inf. Syst. Front.* **8**, 9–20 (2006).
41. Rau, A., Marot, G. & Jaffrézic, F. Differential meta-analysis of RNA-seq data from multiple studies. *BMC Bioinform.* **15**, 91 (2014).
42. Toro-Domínguez, D. et al. A survey of gene expression meta-analysis: methods and applications. *Brief. Bioinform.* **22**, 1694–1705 (2021).
43. Marx, V. The big challenges of big data. *Nature* **498**, 255–260 (2013).
44. Katsnelson, A. Big science: the cancer genome challenge. *Nature* **498**(7455), 255–260 (2013).
45. Cancer (IARC), T. I. A. for R on Global Cancer Observatory. https://gco.iarc.fr/ (2023).
46. Global Cancer Observatory. Cancer Today. https://gco.iarc.who.int/today/ (2023).
47. Kuipers, E. J. et al. Colorectal cancer. *Nat. Rev. Dis. Primers*. **1**, 15065 (2015).
48. Dekker, E., Tanis, P. J., Vleugels, J. L. A., Kasi, P. M. & Wallace, M. B. Colorectal cancer. *Lancet* **394**, 1467–1480 (2019).
49. Simon, K. Colorectal cancer development and advances in screening. *Clin. Interv Aging* **11**, 967–976 (2016).
50. Greene, F. L. & Sobin, L. H. The staging of cancer: a retrospective and prospective appraisal. *CA Cancer J. Clin.* **58**(3), 180–190. https://doi.org/10.3322/CA.2008.0001 (2008).
51. The Eighth Edition AJCC Cancer Staging Manual: continuing to build a bridge from a population-based to a more "personalized" approach to cancer staging—Amin—2017—CA: A Cancer Journal for Clinicians—Wiley Online Library. https://acsjournals.onlinelibrary.wiley.com/doi/. https://doi.org/10.3322/caac.21388 (2017).
52. Mahmoud, N. N. Colorectal cancer: preoperative evaluation and staging. *Surg. Oncol. Clin. N. Am.* **31**, 127–141 (2022).
53. Cancer Staging Systems. *ACS* https://www.facs.org/quality-programs/cancer-programs/american-joint-committee-on-cancer/cancer-staging-systems/ (2020).
54. Malard, F. & Mohty, M. Acute lymphoblastic leukaemia. *The Lancet* **395**, 1146–1162 (2020).
55. Aldoss, I., Forman, S. J. & Pullarkat, V. Acute lymphoblastic leukemia in the older adult. *J. Oncol. Pract.* **15**, 67–75 (2019).
56. Chan, K. W. Acute lymphoblastic leukemia. *Curr. Probl. Pediatr. Adolesc. Health Care*. **32**, 40–49 (2002).
57. pandas—Python. Data Analysis Library. https://pandas.pydata.org/ (2023).
58. Choudhary, S. & pysradb A Python package to query next-generation sequencing metadata and data from NCBI sequence read archive. *F1000Res* **8**, 532 (2019).
59. Hoarfrost, A., Brown, N., Brown, C. T. & Arnosti, C. Sequencing data discovery with metaseek. *Bioinformatics* **35**, 4857–4859 (2019).
60. Bernstein, M. N., Doan, A. & Dewey, C. N. MetaSRA: normalized human sample-specific metadata for the sequence read archive. *Bioinformatics* **33**, 2914–2923 (2017).
61. MeSH on the Web. NLM Technical Bulletin. https://wayback.archive-it.org/org–350/20170327163032/https://www.nlm.nih.gov/pubs/techbull/so97/so97_mesh_web.html (1997).
62. Zhou, Y. et al. Therapeutic target database describing target druggability information. *Nucleic Acids Res.* **52**, D1465–D1477 (2024).
63. WordNet. https://wordnet.princeton.edu/ (2010).
64. Python Software Foundation. Python 3.12.4 Documentation. https://docs.python.org/3/ (2022).
65. *Entrez Programming Utilities Help* (National Center for Biotechnology Information US, 2010).
66. Liu, H., Christiansen, T., Baumgartner, W. A. & Verspoor, K. BioLemmatizer: a lemmatization tool for morphological processing of biomedical text. *J. Biomed. Semant.* **3**, 3 (2012).
67. Yang, T., He, Y. & Yang, N. Named entity recognition of medical text based on the deep neural network. *J. Healthcare Eng.* **2022**, 3990563 (2022).
68. W3C XML Schema Definition. Language (XSD) 1.1 Part 1: structures. https://www.w3.org/TR/xmlschema11–1/ (2022).
69. ViewVC SRA docs repository [v3]. Index of /trunk/sra/doc/SRA. https://www.ncbi.nlm.nih.gov/viewvc/v3/trunk/sra/doc/SRA/ (2022).
70. *WordNet: an Electronic Lexical Database*. (Bradford Books, 1998).
71. Download MeSH Data. https://www.nlm.nih.gov/databases/download/mesh.html (2022).
72. Nguyen, V. H., Nguyen, H. T., Duong, H. N. & Snasel, V. n-Gram-based text compression. *Comput. Intell. Neurosci.* **2016**, 9483646 (2016).
73. Django documentation. Django documentation. *Django Project*. https://docs.djangoproject.com/en/5.0/ (2022).
74. WestHealth/pyvis. West Health Institute (2024).
75. Proceedings of the Python in Science Conference (SciPy). Exploring network structure, dynamics, and function using NetworkX. http://conference.scipy.org.s3-website-us-east–1.amazonaws.com/proceedings/scipy2008/paper_2/index.html (2022).

Author contributions

Conception: PAGS and JCR. Methodology: PASG and JCR. Creation of pipeline presented: PASG. Acquisition, analysis and interpretation of data: PASG, MGRC and JCR. Writing: PAGS. Writing—review and editing: MGRC and JCR. Supervision: JCR. Funding acquisition: MGRC and JCR.

Funding

This research was funded by: (a) Conselho Nacional de Desenvolvimento Científico e Tecnológico – CNPq, grant number 313108/2022-6 (JCR) and process number 156904/2021-7 (PASG); (b) Chamada nº 08/2023 - CAPES/PrInt-Fiocruz, grant number 88887.885815/2023-00 (JCR); (c) Call FAPEMIG 07/2021: Redes de Pesquisa Científica e Desenvolvimento Tecnológico com Foco em Demandas Estratégicas, grant number RED-00104-22 (JCR and MGRC);

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-025-91781-8.

Correspondence and requests for materials should be addressed to M.G.R.C. or J.C.R.