Γεφυρώνοντας μεγάλα ομικά, γενετικά και ιατρικά δεδομένα για την ευρεία εφαρμογή της Ιατρικής Ακριβείας στην Ελλάδα

DELIVERABLE D11.1

«Technical report with data extracted from text mining»

Φορέας	Hellenic Pasteur Institute
Τύπος Παραδοτέου	Report
Ημερομηνία Υποβολής Παραδοτέου	15th February 2026
Ενότητα Εργασίας	Work Package 11 «Information Retrieval and Knowledge Extraction for Mutation–Cancer Associations from Biomedical Literature»

Ελλάδα 2.0
ΕΘΝΙΚΟ ΣΧΕΔΙΟ ΑΝΑΚΑΜΨΗΣ ΚΑΙ ΑΝΘΕΚΤΙΚΟΤΗΤΑΣ
Με τη χρηματοδότηση της Ευρωπαϊκής Ένωσης
NextGenerationEU

Contents

Ελλάδα 2.0 ΕΘΝΙΚΟ ΣΧΕΔΙΟ ΑΝΑΚΑΜΨΗΣ ΚΑΙ ΑΝΘΕΚΤΙΚΟΤΗΤΑΣ

Με τη χρηματοδότηση της Ευρωπαϊκής Ένωσης NextGenerationEU

Related Publications

1. Darling (v2.0): Mining disease-related databases for the detection of biomedical entity associations. Baltoumas FA, Karatzas E, Venetsianou NK, Aplakidou E, Giatras K, Chasapi MN, Chasapi IN, Iliopoulos I, Iconomidou VA, Trougakos IP, Psomopoulos F, Giannakakis A, Georgakopoulos-Soares I, Kontou P, Bagos PG, Pavlopoulos GA. Computational and Structural Biotechnology Journal. 2025 Jun 14;27:2626–2637. doi:10.1016/j.csbj.2025.06.025. PMID:40599243.

Ελλάδα 2.0
ΕΘΝΙΚΟ ΣΧΕΔΙΟ ΑΝΑΚΑΜΨΗΣ ΚΑΙ ΑΝΘΕΚΤΙΚΟΤΗΤΑΣ
Με τη χρηματοδότηση της Ευρωπαϊκής Ένωσης
NextGenerationEU

1 Introduction

The rapid expansion of biomedical literature has created unprecedented opportunities for advancing knowledge in cancer genomics and precision medicine. Thousands of new publications are added to repositories such as PubMed every year, reporting associations between genetic variants, genes, and cancer-related phenotypes. While this growing body of evidence is invaluable, its scale and heterogeneity make systematic manual curation increasingly impractical. As a result, a substantial proportion of potentially actionable knowledge remains fragmented across unstructured textual sources [1].

In the context of precision oncology, the identification and integration of gene-variant-cancer associations is of central importance [2]. Such associations underpin disease risk stratification, prognosis, therapeutic decision-making, and the interpretation of genomic testing results. However, extracting these relationships from free-text scientific articles requires not only the identification of relevant biomedical entities, but also the correct interpretation of complex semantic relationships and study outcomes.

Recent advances in Natural Language Processing (NLP), and in particular Large Language Models (LLMs), provide new opportunities for large-scale knowledge extraction from biomedical literature [3]. LLMs demonstrate strong capabilities in understanding scientific language, reasoning across sentences, and generating structured representations from unstructured text. When combined with targeted corpus construction, entity annotation, and controlled prompting strategies, LLMs offer a promising approach for transforming dispersed literature evidence into structured, computable knowledge.

The objective of this deliverable is to present a technical report on the datasets and knowledge produced through an end-to-end pipeline that integrates literature retrieval, entity recognition, LLM-based semantic extraction, and systematic evaluation.

This report describes the full workflow, beginning with large-scale literature collection from PubMed, followed by corpus preparation and automated entity annotation. It then details the design and implementation of an LLM-based extraction pipeline for identifying explicit gene-variant-cancer associations, along with a quantitative evaluation against manually curated reference data. Finally, it presents the resulting structured knowledge base and analyses its content, coverage, and internal relationships.

By combining scalable data collection, controlled language model inference, and rigorous evaluation, this work demonstrates the feasibility of automated, high-confidence knowledge extraction from biomedical literature and lays the foundation for continuous enrichment of genomic resources relevant to cancer research and precision medicine.

2 Data Collection and Corpus Construction

2.1 Collection of Biomedical Literature from PubMed

As a first step, a large-scale collection of biomedical literature was assembled from PubMed [4] in order to support downstream knowledge extraction using Large Language Models (LLMs). The focus of the collection was on scientific articles reporting associations between genetic variants, genes, and cancer-related phenotypes.

To construct the relevant literature corpus, a targeted PubMed query strategy was employed. Publications were restricted to English-language articles and required to include terminology indicative of genetic variation (e.g. variant, mutant, polymorphism, variation, SNP). In addition, publications were filtered by publication date to support temporal analyses, while non-original research outputs, such as review articles and retracted publications, were explicitly excluded. This query design ensures that the collected corpus primarily consists of original research articles reporting primary findings related to genetic variation, providing a focused input for downstream entity annotation and LLM-based knowledge extraction.

The data collection process followed a keyword- and identifier-based filtering strategy. Publications were initially retrieved from PubMed based on the availability of an abstract and standard bibliographic metadata. From this broader set, articles were further filtered according to two main criteria:

1. **Presence of cancer-related terminology**, including general cancer terms and oncology-specific keywords appearing in the title or abstract using stanza tool [5].

2. **Presence of genetic variant identifiers**, focused on dbSNP [6] reference SNP identifiers (rsIDs), using regular expressions.

This dual filtering strategy ensured that the collected corpus was enriched for

publications likely to describe biologically meaningful relationships between genetic variation and cancer-related outcomes, while excluding a large volume of irrelevant or purely descriptive biomedical literature.

All data collection and preprocessing workflows were implemented using Python-based scripts. The collected data were stored in a MongoDB [7] NoSQL database, selected for its flexibility in handling heterogeneous document structures and its suitability for large-scale text corpora. This design choice facilitates efficient querying, filtering, and iterative enrichment of records during subsequent analysis stages.

2.2 Temporal Trends in PubMed Literature

To characterise the scale and growth of the collected literature, exploratory analyses were performed across publication decades. For each decade, publication counts were computed after applying the filtering criteria described above, allowing the derivation of targeted subtotals reflecting specific aspects of the biomedical literature.

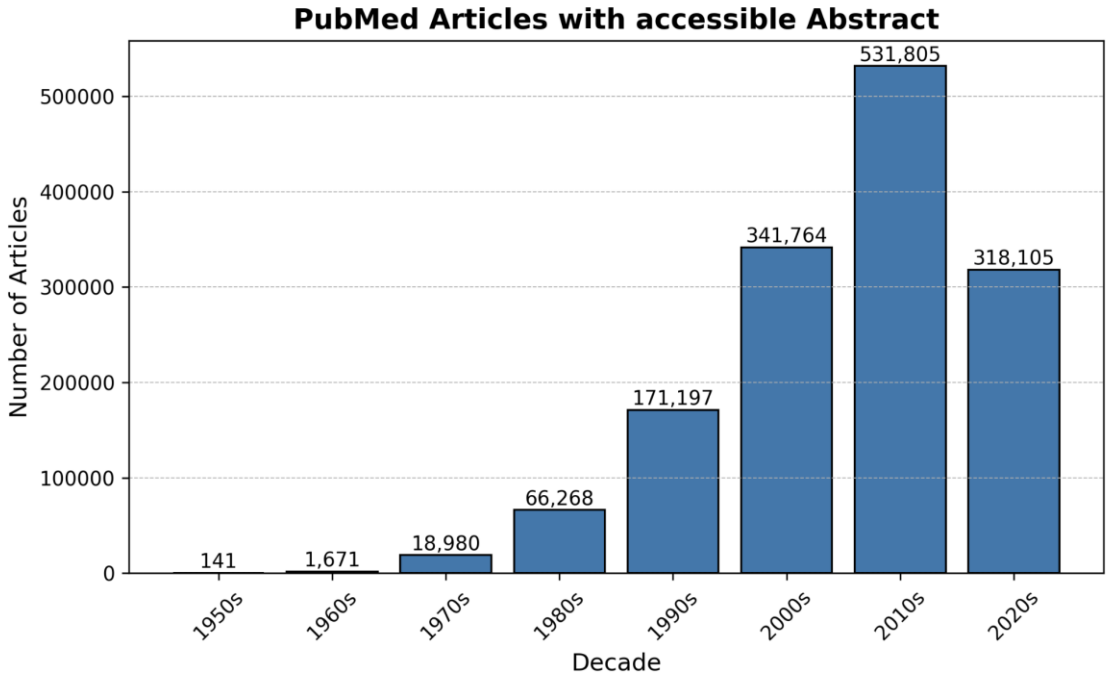


Figure 1 presents the total number of PubMed articles per decade with an accessible abstract, provided by the main query. This figure provides a reference baseline and illustrates the overall growth of biomedical publishing over time.

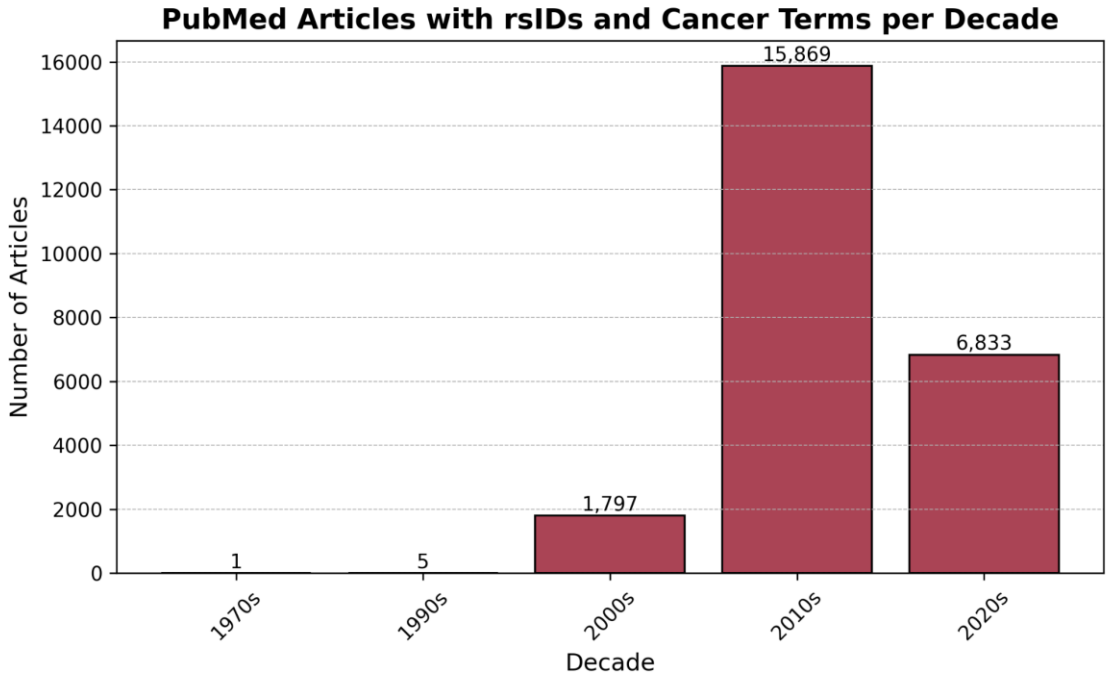


Figure 2 focuses on publications that mention genetic variants and cancer type terms, identified through the presence of rsID patterns, and keywords through name entity recognition pipeline.

These figures collectively motivate the use of automated, LLM-based approaches, as manual curation of such a rapidly expanding body of literature is no longer feasible.

2.3 Construction of an LLM-Ready Corpus

Following data collection and filtering, the resulting literature corpus was prepared for automated parsing and information extraction by Large Language Models. By restricting the corpus to articles that explicitly reference cancer-related concepts and genetic variants, the downstream extraction process is both more efficient and more likely to yield biologically and clinically meaningful associations.

Each publication record was structured to include:

- PubMed Identifier (PMID),
- publication metadata (title, journal, year),
- the abstract text,
- and automatically derived indicators reflecting the presence of cancer terminology and variant identifiers.

This structured representation enables efficient ingestion by LLM-based pipelines and supports scalable extraction of complex relationships, such as associations between genes, genetic variants, cancer types, and reported clinical or biological outcomes.

2.4 Named Entity Recognition and Normalisation Using Tool EXTRACT

In order to further characterise the collected literature and to identify publications containing relevant biomedical entities, an automated Named Entity Recognition (NER) and Named Entity Normalisation (NEN) step was applied to the corpus using the EXTRACT text-mining tool [8].

EXTRACT is a biomedical text-mining system designed to identify and normalise mentions of biological entities in scientific text. Within the scope of this Work Package, EXTRACT was employed to automatically detect and annotate gene- and cancer-related entities within PubMed abstracts. For variants a regular expression approach was employed to identify rsIDs.

The objective of this step was to identify publications that contain explicit mentions of both genetic variants and cancer-related entities. This process enabled a more targeted selection of articles for downstream analysis and ensured that the subsequent Large Language Model (LLM) processing stage focuses on publications with high informational relevance.

2.5 Distribution of Cancer Mentions Across the Literature Corpus

Following entity annotation with EXTRACT, aggregated statistics were generated to analyse the distribution of cancer-related mentions across the corpus. Cancer entities were grouped into broader cancer categories in order to provide a high-level overview of disease coverage within the collected publications and a simplified method for querying the mentions later on.

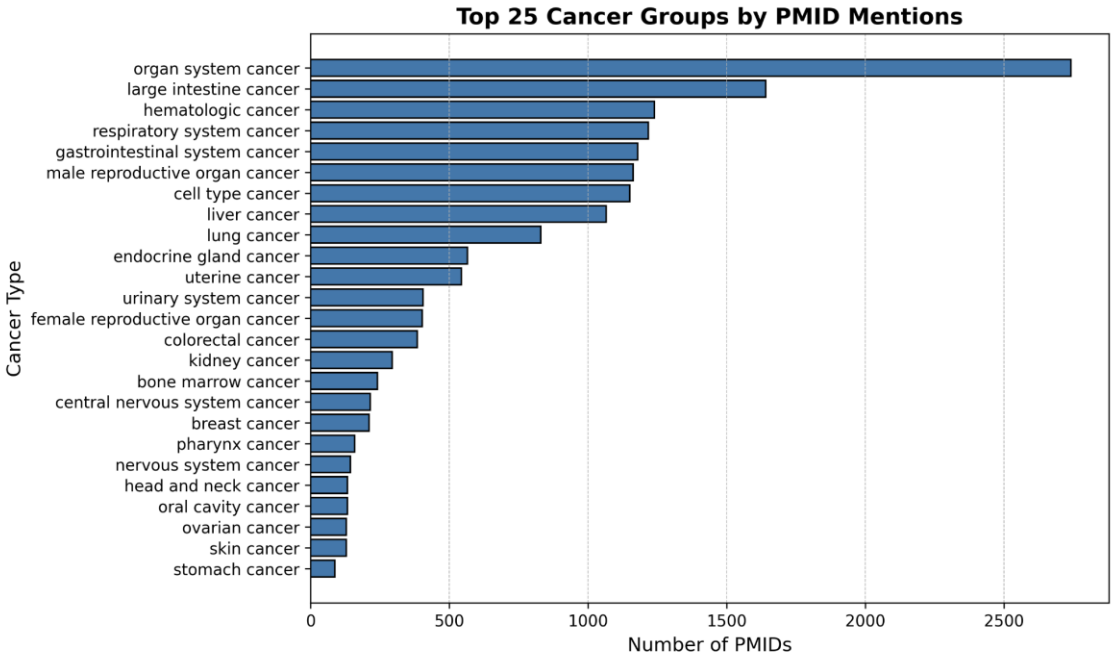


Figure 3 presents the top 25 cancer groups ranked by the number of associated PubMed identifiers (PMIDs). The results demonstrate broad coverage across multiple organ systems and cancer types, with particularly high representation for organ system cancers, gastrointestinal and hematologic malignancies, respiratory system cancers, and reproductive organ cancers.

2.6 Identification of Publications for LLM-Based Parsing

The annotations produced by EXTRACT were used as a filtering and prioritisation mechanism to identify publications suitable for deep semantic parsing by the LLM-based extraction pipeline.

Publications that contained:

- at least one recognised genetic variant mention, and
- at least one cancer-related entity,

were flagged as candidates for further processing. This step significantly reduces noise by excluding articles unlikely to contain explicit variant-cancer associations, while retaining a diverse and information-rich subset of the literature.

Finalized dataset of publications ready to be parsed: 24505

At the end of this process, a final set of publications was defined as the input corpus for LLM-based extraction, from which structured records of the form:

PMID - Gene - Variant - Cancer

will be derived.

2.7 Creation of an Evaluation Dataset

In parallel with dataset construction, an evaluation dataset was created to support qualitative and quantitative assessment of the LLM-based extraction pipeline.

A representative subset of 200 scientific articles was selected from the full corpus. Selection criteria ensured coverage across:

- different cancer types,
- a range of genetic variants and genes,
- multiple publication periods,
- and diverse study designs.

This evaluation set serves as a reference benchmark for assessing the accuracy and completeness of extracted entities and relations, including:

- Gene-Variant mappings
- Variant-Cancer associations

The evaluation dataset is stored alongside the main corpus in the MongoDB database, allowing seamless comparison between automatically extracted results and curated reference information during model development and validation.

2.8 Implementation Overview

All stages of data collection, preprocessing, and corpus management were implemented using Python, leveraging widely adopted libraries for biomedical data access and text processing. Data persistence and management were handled through a MongoDB NoSQL database, enabling flexible schema evolution as additional annotations and extracted knowledge are incorporated in later stages of the Work Package. The NER/NEN annotation workflow using EXTRACT, as well as the aggregation and analysis of annotation results, were implemented using Python-based scripts.

3 LLM-Based Extraction of Gene-Variant-Cancer Associations

Following corpus construction and entity annotation, an LLM-based extraction step was applied to identify explicit gene-variant-cancer associations reported in PubMed titles and abstracts. The goal of this stage is to transform unstructured text into structured records of the form:

PMID - Gene - Variant - Cancer

This structured representation enables downstream analysis, aggregation across cancer types, and systematic evaluation against curated reference annotations.

3.1 Controlled and Scalable LLM Inference Configuration

For scalable processing of a large literature corpus, we used the DeepSeek-R1-Distill-Llama-8B model via the HuggingFace Transformers framework [9] (see Section 4 for evaluation).

To ensure reliable and scalable extraction across a large corpus of biomedical publications, the LLM inference process was executed under a controlled configuration. Memory-optimised model loading was employed to enable stable execution on available GPU resources, while supporting sustained processing of thousands of abstracts.

Inference was performed in small, fixed-size batches, allowing predictable resource utilisation and preventing memory saturation during execution. This batch-based strategy also facilitates incremental processing and intermediate result storage, supporting fault tolerance and reproducibility.

To minimise variability in model outputs and ensure consistent interpretation of extracted associations, generation parameters were configured to favour deterministic behaviour. In particular, low-temperature, non-sampling generation was used, ensuring that identical inputs yield consistent outputs across runs. This is essential for systematic evaluation, comparison across model configurations, and alignment with curated reference annotations [10].

3.2 Entity-Primed One-Shot Prompting Strategy

To improve extraction precision and ensure consistent structured outputs, an entity-primed one-shot prompting strategy was adopted. For each publication, the LLM was provided with pre-identified candidate entities derived from earlier annotation steps, including gene mentions, genetic variant identifiers, and cancer-related disease terms.

Rather than requiring the model to independently identify entities from free text, the prompt explicitly constrained the extraction process to these supplied entity lists. This approach reduces ambiguity in entity naming, limits spurious associations, and ensures alignment with standardised identifiers used throughout the pipeline.

To further guide model behaviour and enforce adherence to the desired output schema, a one-shot prompting approach was employed. Each prompt included a single illustrative example demonstrating how gene-variant-cancer associations should be extracted from a publication and encoded in a structured JSON format. The example explicitly defined how to interpret evidence for association and how to label cases where associations were previously reported, newly observed, or not supported by the text.

Together, entity priming and one-shot prompting enable controlled semantic extraction while preserving flexibility in language understanding. This strategy ensures that the LLM focuses on relation extraction rather than entity discovery, resulting in consistent, reproducible, and evaluation-ready structured outputs.

3.3 Illustrative LLM Prompt Used for Gene-Variant-Cancer Extraction

System role:
You are an expert biomedical researcher.

Instructions:
Use only the entity names provided below.
If a field is not supported by the text, output "0".

Provided entity mentions (pre-annotated):

- **Gene mentions:** *[list of gene names]*
- **Variant mentions:** *[list of rsIDs]*
- **Cancer mentions:** *[list of cancer terms]*

Task:
Extract each unique, direct, and novel gene-variant-cancer association from the title and abstract provided below.

For each extracted association, return a JSON object with the following fields:

- **Gene_name**: use only names from the *Gene mentions* list, or "0"
- **Variant_name**: use only names from the *Variant mentions* list, or "0"
- **Cancer_name**: use only names from the *Cancer mentions* list, or "0"
- **Cancer_Risk**:
    - "1" if the abstract reports an association (increase or decrease of risk),
    - "2" if the association was previously reported (i.e. not novel), **or** if the study investigated the association but found no significant relationship,

- "0" if no evidence of association is stated in the abstract.

Return the result as a single JSON array.
If no associations are found, return only an empty array: [].

One-Shot Example Provided to the Model

Example input:

Title:
Association of BRCA1 genotypes with susceptibility to breast cancer.

Abstract:
The rs3847 variant in the BRCA2 gene has been previously associated with breast cancer (BC).
In this study, we investigated the association of two BRCA1 variants, rs1293 and rs3822, with BC.
Our results demonstrate that the rs3822 (A>T) variant in BRCA1 is significantly associated with breast cancer in the Chinese population.

Entity lists:

- Gene mentions: BRCA2, BRCA1
- Variant mentions: rs3847, rs3822, rs1293
- Cancer mentions: breast cancer

Expected output:

```
[
 {
  "Gene_name": "BRCA2",
  "Variant_name": "rs3847",
  "Cancer_name": "breast cancer",
  "Cancer_Risk": "2"
 },
 {
  "Gene_name": "BRCA1",
```

```
  "Variant_name": "rs3822",
  "Cancer_name": "breast cancer",
  "Cancer_Risk": "1"
 },
 {
  "Gene_name": "BRCA1",
  "Variant_name": "rs1293",
  "Cancer_name": "breast cancer",
  "Cancer_Risk": "0"
 }
]
```


Actual input provided to the model:

Title: *[Publication title]*
Abstract: *[Publication abstract]*

3.4 Implementation Overview

The LLM-based extraction workflow was implemented using Python and the HuggingFace Transformers framework. The pipeline integrates pre-annotated entity information with controlled LLM prompting scalable and reproducible extraction of gene-variant-cancer associations from biomedical literature.

Publications were processed incrementally in fixed-size batches, allowing robust execution across large corpora and facilitating intermediate result storage. LLM outputs were generated in a structured JSON format and stored for subsequent post-processing, validation, and evaluation against curated reference datasets.

4 Evaluation

To assess the quality and robustness of the LLM-based extraction pipeline, a systematic evaluation was conducted using a manually curated reference dataset and multiple language model configurations. The objective of this evaluation was twofold:

(i) to quantify extraction performance using standard information retrieval metrics, and

(ii) to identify the most suitable model for large-scale deployment in subsequent pipeline stages.

4.1 Curated Evaluation Dataset

Evaluation was performed using the benchmark set of 200 manually curated PubMed publications as described in Section 2, selected to represent a diverse range of cancer types, genes, and genetic variants. For each publication, expert-curated annotations were available in the form of structured tuples:

PMID - Gene - Variant - Cancer

This dataset serves as ground truth and enables objective comparison between automatically extracted associations and reference annotations.

4.2 Model Configurations Evaluated

Three different large language models were evaluated under the same extraction framework, prompt structure, and inference configuration, ensuring a controlled comparison:

- DeepSeek-R1-Distill-Llama-8B
- DeepSeek-R1-Distill-Qwen-7B
- Hermes-3-Llama-3.1-8B (NousResearch)

All models were evaluated using the same entity-primed one-shot prompting strategy and deterministic inference settings described in Section 3.

4.3 Evaluation Methodology and Metrics

For each model, extracted gene-variant-cancer associations were compared against the curated reference dataset at the tuple level. Performance was quantified using standard information extraction metrics, including:

- True Positives (TP): correctly extracted associations
- False Positives (FP): extracted associations not present in the reference set
- False Negatives (FN): reference associations not recovered by the model
- Precision: proportion of extracted associations that are correct
- Recall: proportion of reference associations successfully recovered
- F1-score: harmonic mean of precision and recall

Overall accuracy was also computed to provide a high-level summary of model performance.

4.4 Comparative Results

Table 1 summarises the comparative performance of the three evaluated models.

Table 1.1a - Overall Extraction Performance

Model	Accuracy (%)	Precision	Recall	F1-score
Hermes-3-Llama-3.1-8B	86.09	0.86	0.81	**0.83**
DeepSeek-R1-Distill-Qwen-7B	89.20	0.89	0.70	0.78
DeepSeek-R1-Distill-Llama-8B	**93.63**	**0.94**	0.68	0.79

Table 1.1b - Error Analysis and Record Coverage

Model	Extracted Records	Ground Truth Records	TP	FP	FN
Hermes-3-Llama-3.1-8B	345	368	297	48	71

DeepSeek-R1-Distill-Qwen-7B	287	368	256	31	112
DeepSeek-R1-Distill-Llama-8B	267	368	250	**17**	118

4.5 Model Selection Rationale

The evaluation results reveal distinct performance trade-offs across the evaluated models, particularly with respect to precision and false positive rates. Given the downstream use of the extracted associations, priority was placed on minimising false positives, even at the expense of reduced recall. In this context, it is preferable to recover fewer associations with high confidence rather than a larger number of potentially incorrect records.

The DeepSeek-R1-Distill-Llama-8B model demonstrated the most conservative extraction behaviour, achieving the lowest number of false positives (FP = 17) and the highest precision (0.94) among all evaluated models. This indicates a strong ability to extract associations only when supported by clear evidence in the text, thereby reducing the risk of introducing spurious gene-variant-cancer relationships into the final corpus.

Based on these considerations, the DeepSeek-R1-Distill-Llama-8B model was selected as the primary model for large-scale corpus processing in subsequent stages of the pipeline, as it best satisfies the requirement for high-confidence, low-noise extraction.

5 Resulting Knowledge Base of Gene-Variant-Cancer Associations

The application of the selected LLM-based extraction pipeline resulted in a large, structured knowledge base of gene-variant-cancer associations derived from PubMed abstracts. This section summarises the key characteristics of the resulting database and provides an overview of its content, coverage, and internal structure.

5.1 Dataset Size and Coverage

Following deduplication of identical associations, the final dataset comprises 13,680 unique gene-variant-cancer associations, supported by a broad set of PubMed publications. The extracted knowledge base spans multiple decades of biomedical research and covers a wide spectrum of cancer types, genes, and genetic variants.

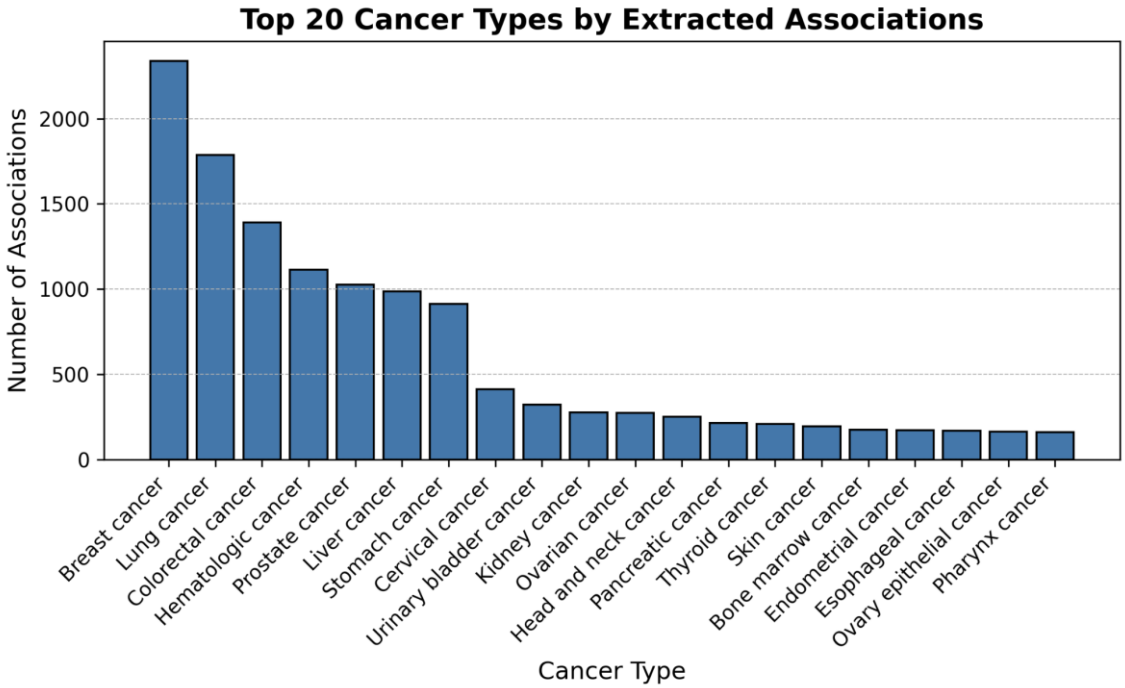
Overall summary statistics for the dataset are:

- Total extracted associations - 13680
- Unique PMIDs - 7319
- Unique variants - 7191
- Unique cancer types - 73
- Unique genes - 2406
- Year range - 2002 -2025

5.2 Descriptive Analysis of the Extracted Knowledge Base

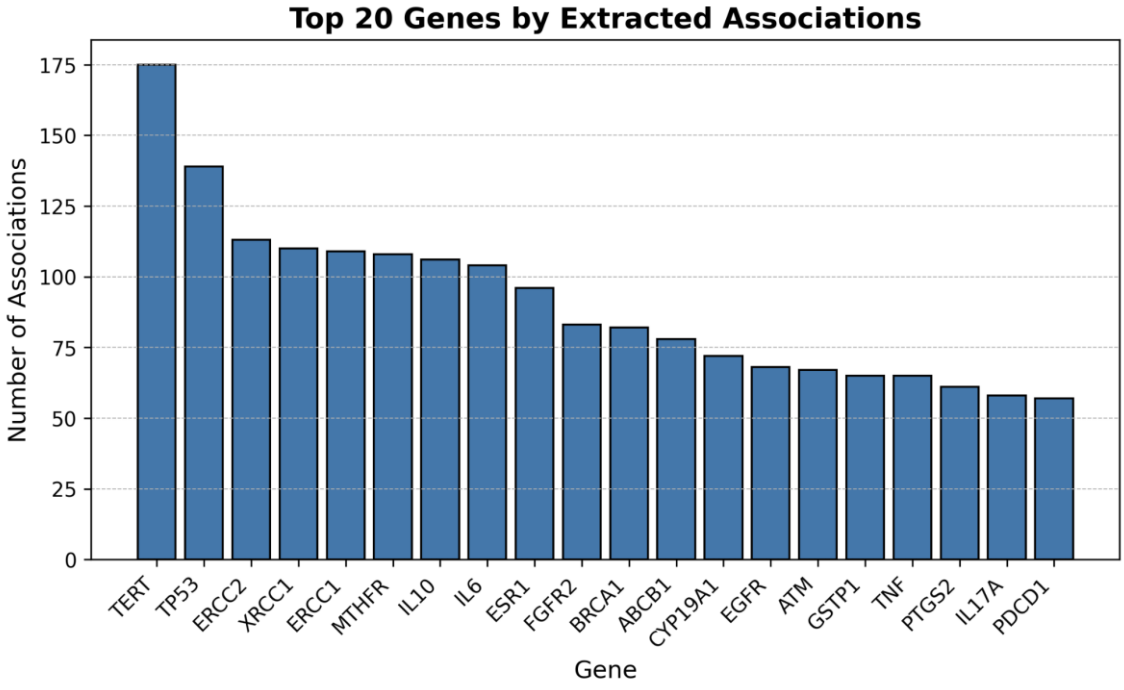
An overview of the most frequently represented cancer types in the extracted database is shown in **Figure 4**. The distribution is dominated by common and well-studied cancers, including breast cancer, lung cancer, colorectal cancer, and hematologic malignancies. Importantly, the presence of a long tail of additional cancer types indicates that the extraction pipeline captures associations beyond a narrow subset of diseases, supporting broad applicability across oncology domains.

Figure 4



A complementary gene-centric perspective is provided in **Figure 5**, which illustrates the distribution of associations across genes. Several biologically well-established cancer-related genes, such as *TP53*, *TERT*, *BRCA1*, *ERCC1*, and *XRCC1*, appear among the most frequently represented entries. The enrichment of these genes provides qualitative validation of the extraction process, indicating that the LLM-based approach prioritises meaningful genetic signals reported in the literature.

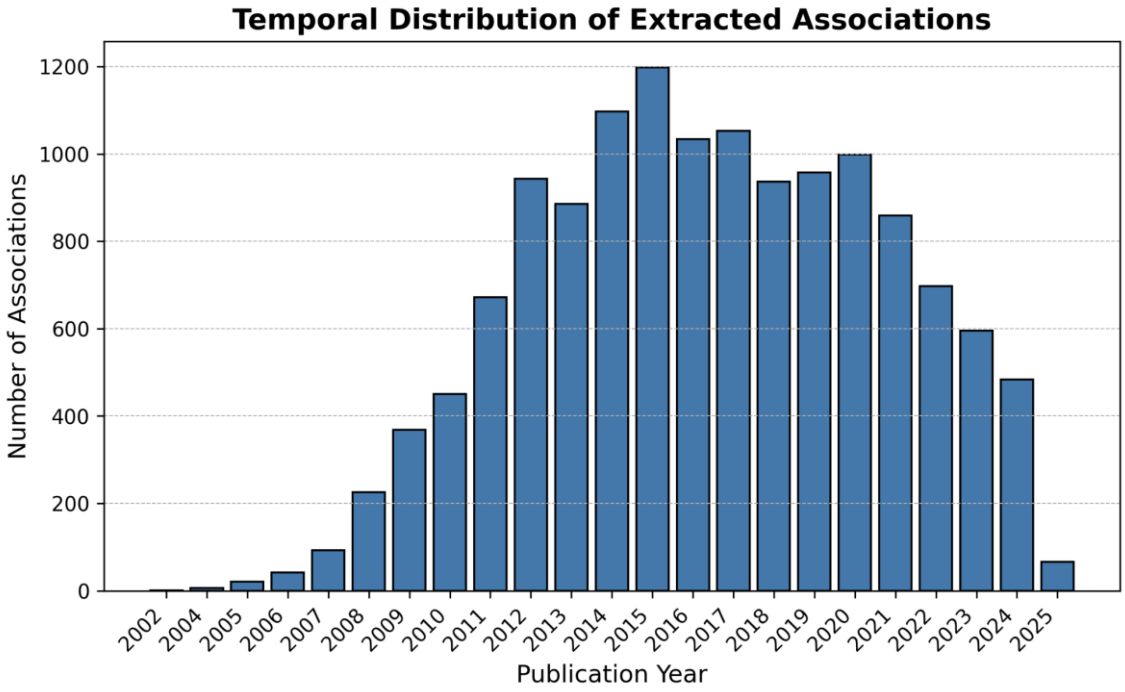
Figure 5



Finally, the temporal evolution of extracted gene-variant-cancer associations is shown

in **Figure 6**. The number of associations increases steadily from the early 2000s, peaks during the mid-to-late 2010s, and remains substantial in more recent years. This trend mirrors the expansion of genomic research and the widespread adoption of high-throughput genetic technologies in oncology. The sustained extraction of associations from recent publications demonstrates that the pipeline remains applicable to contemporary literature and is well suited for continuous updates as new studies emerge.

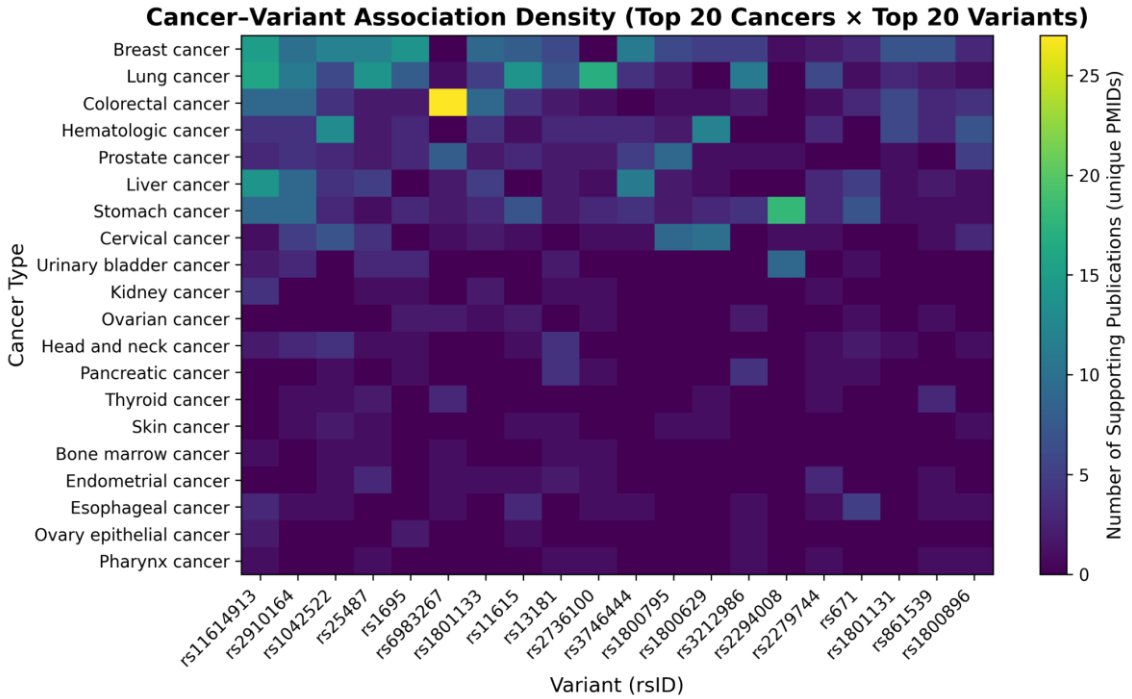
Figure 6



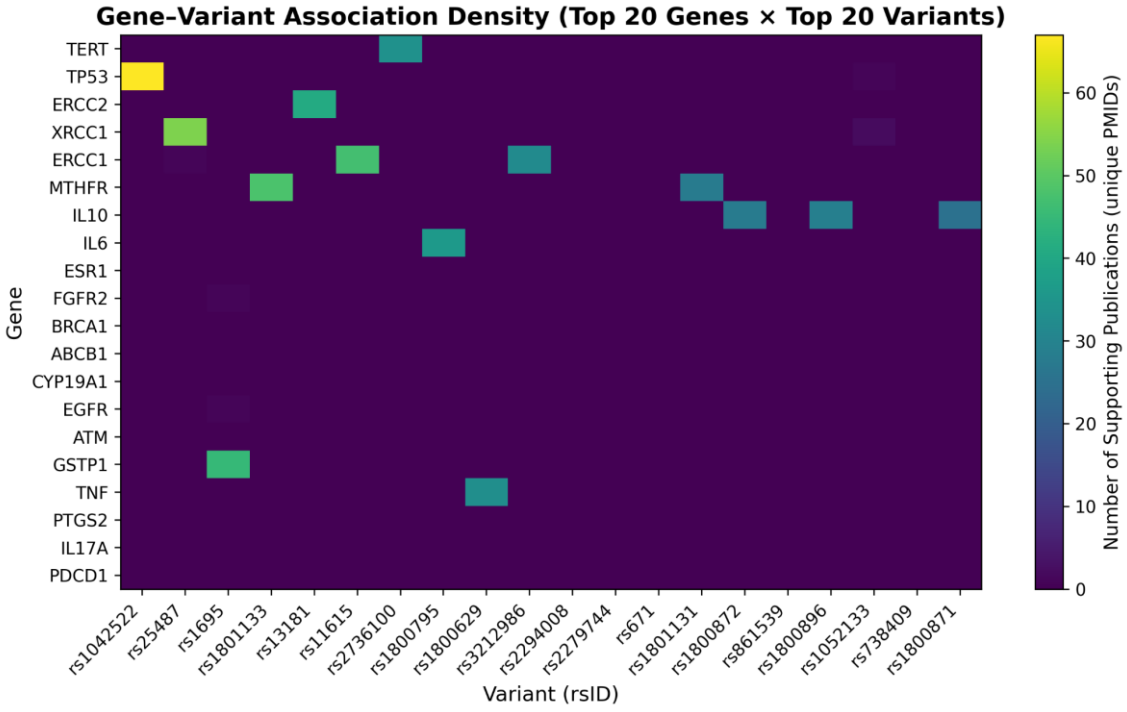
5.3 Relationship Density Across Entity Types

To further characterise the internal structure of the resulting knowledge base, we examined pairwise association densities between key entity types.

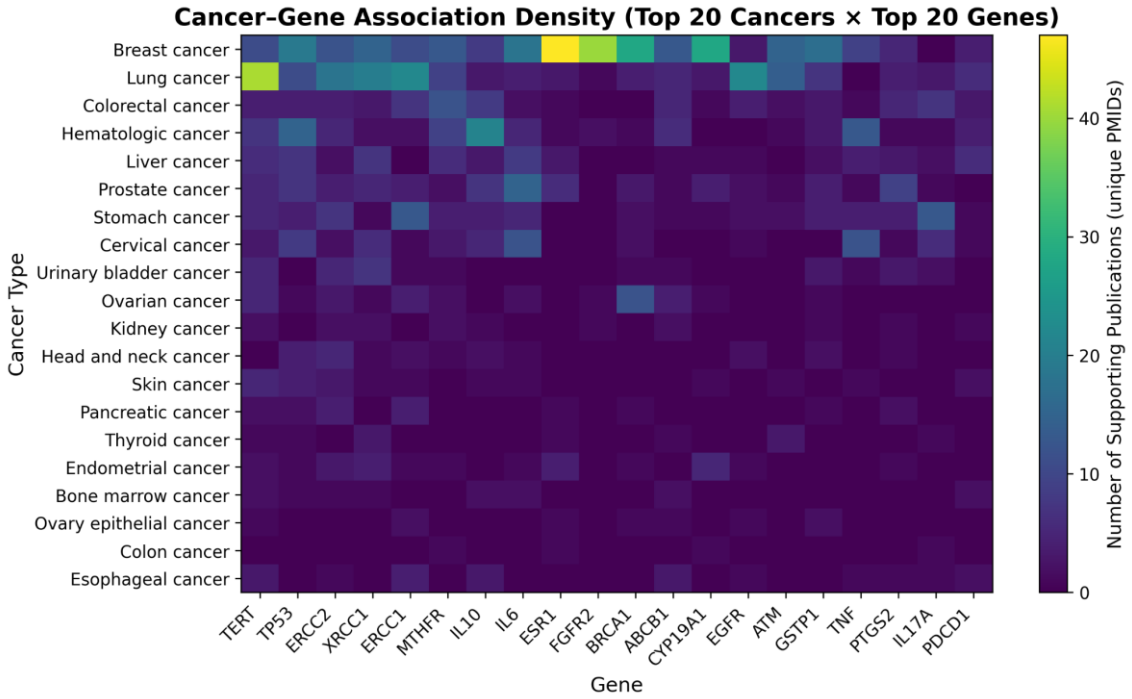
- **Figure 7** presents a heatmap of cancer-variant associations, highlighting variants that recur across multiple cancer types and revealing cancer-specific variant patterns.



- **Figure 8** shows gene-variant association density, illustrating how particular variants cluster around well-studied genes while many gene-variant pairs remain sparsely represented.



- **Figure 9** depicts cancer-gene association density, capturing both broadly acting cancer genes (associated with multiple cancer types) and genes with more cancer-specific relevance.



5.7 Summary

In summary, the LLM-based extraction pipeline produced a large, high-quality, and literature-supported database of gene-variant-cancer associations. The resulting knowledge base captures well-established oncogenetic relationships while maintaining broad coverage across genes, variants, cancer types, and publication years. These characteristics make it a strong foundation for downstream analysis, integration with curated resources, and future expansion.

6. Darling: disease-centric literature mining

In parallel to the tuple-based knowledge base described above, Work Package 11 also delivers a disease-centric literature-mining component, Darling (v2.0), which supports large-scale retrieval and exploratory analysis of sentence-level co-occurrence networks among biomedical entities. Together, the two components provide complementary access to the literature: Darling enables interactive exploration and hypothesis generation, while the LLM pipeline yields evaluated, machine-readable gene–variant–cancer tuples for downstream computational reuse.

Darling is a web application that leverages literature mining to detect associations among biomedical entities related to disease. Darling identifies sentence-level co-occurrences of biomedical entities—such as genes, proteins, chemicals, functions, tissues, diseases, environments and phenotypes—from disease-centric collections curated across six established resources. In this release we additionally integrate dedicated query channels focusing on COVID-19, GWAS studies, cardiovascular, neurodegenerative and cancer diseases. Compared with its predecessor, users now have

extended search options, including searches by PubMed IDs, disease records, entity names, titles, single-nucleotide polymorphisms, or Entrez syntax. Moreover, relevant literature can be retrieved and mined from recognized terms in free-text input after applying named entity recognition (NER). Term associations are rendered as configurable networks that can be further filtered by term or by co-occurrence frequency and visualized in 2D as weighted graphs or in 3D as multilayer networks. Retrieved terms are organized into searchable tables and clustered, annotated documents. Reported genes can be further analyzed for functional enrichment via external applications invoked from within Darling. The Darling back-end databases, including terms and their associations, are updated annually.

Availability: http://darling.pavlopouloslab.info or https://pavlopoulos-lab.org/shinyapps/app/darling

6.1 Data sources and automated retrieval of publications

The solution leverages open full-text and metadata infrastructures, including the PubMed Central Open Access Subset through its available distribution mechanisms and content catalogues, Europe PMC via REST APIs and bulk downloads for access to open full texts and rich metadata, and the NCBI Entrez E-utilities as a programmatic interface for searching, fetching and linking records across PubMed, PMC and related databases. Retrieval is implemented as an iterative collection pipeline that aggregates open publications, stores them locally and maintains references to primary identifiers (PMID, PMCID and Europe PMC IDs, where available).

Following retrieval, automated curation procedures are applied to harmonize article formats so that content from different sources is converted into a uniform intermediate representation suitable for text mining. In addition, per-publication metadata—including title, authors, journal, year, article type, identifiers and access or licensing information—are systematically extracted and stored in structured files and/or database tables to support efficient querying and downstream analysis.

6.3 Knowledge extraction and analytical flow (text mining and Machine Learning)

Darling supports entity mining and sentence-level co-occurrence analysis, producing association networks among biomedical terms. In this project, the analytical flow leverages these capabilities and specializes them for entities related to DNA variants and polymorphisms (e.g., SNPs), with emphasis on associations with cancer types/subtypes, and to pharmacological response, therapeutic regimens and outcome indicators (including survival).

Extracted knowledge is encoded as associations between entities and categorized by cancer type and association type, supporting queries such as incidence, relapse, metastasis, favourable/unfavourable response and survival.

6.4 Database and user interface

A database and web interface are delivered that integrate: the curated publication corpus (and, where licensed, full texts or parts thereof), the associated metadata and extracted entities, and the association networks with frequency/strength indicators. The UI provides search and filtering by cancer type, DNA variants/SNPs, genes, drugs and related entities, as well as visualization of associations as configurable networks.

Because open collections are updated systematically, the collection and curation pipeline is designed to run periodically and refresh the local corpus, with a suggested quarterly cycle complemented by daily update streams where provided by the data sources.

6.5 Data collection and analysis workflow

An overview of the data collection, handling and annotation process is shown in Figure 1A. Similar to v1.0, the first step involved collecting updated disease records and their literature references from seven recognized repositories (accessed December 2024): OMIM, DisGeNET and HPO (already present in v1.0), plus four additional repositories: (i) MONDO, a semi-automatically constructed disease ontology from the Monarch Initiative; (ii) RNADisease (formerly MNDR), focusing on RNA-associated diseases; (iii) the GWAS Catalog, curated by EMBL-EBI; and (iv) LitCovid, a literature hub hosted by NCBI specializing on SARS-CoV-2/COVID-19. Database files were parsed and the linked publications were extracted, yielding a non-redundant set of 1,464,952 publications (Table 1). Corresponding abstracts were then retrieved from PubMed via the Entrez API [36] and Biopython [11].

Retrieved abstracts were processed with text mining and NER to identify biomedical entities using the EXTRACT tagger, a lexicon-based method that scales to large corpora. EXTRACT maps canonical and synonym terms to database identifiers, ensuring concept normalization. In total, 96,176 unique entity terms were obtained, spanning genes/proteins, chemicals, organisms, GO terms, diseases, tissues, environments and phenotypes.

The corpus was then used for co-occurrence analysis by building a knowledge-based interaction network: each node is a bio-entity, and an edge indicates that the two entities co-occur within the same sentence of an abstract. Edge weights represent the total number of co-mentions across the analyzed corpus.

Finally, abstracts and entities were further filtered to produce annotated literature subsets for the new disease-specific search interfaces in Darling v2.0 (Figure 1B), including cardiovascular diseases, cancer, nervous system diseases, COVID-19 and GWAS. Annotation for cardiovascular/cancer/nervous system subsets used mapping of disease terms to Disease Ontology (DO) identifiers and their hierarchy to classify terms and associated publications (e.g., cancer type/subtype or affected organ). LitCovid metadata were used for COVID-19 categorization. GWAS annotations were retrieved from the GWAS Catalog for publications linked to relevant GWAS IDs, study names, traits, SNPs and loci.

6.6 Search inputs, channels and parameters

Darling provides a comprehensive set of search options supporting disease-related research (Figure 2). Users may run general disease searches or targeted queries for cardiovascular, nervous system, cancer and COVID-19-related literature. Additional channels support PubMed literature searches, keyword queries, GWAS and SNP queries, bio-entity searches (chemicals, genes/proteins, tissues) and free-text mining. Structured Entrez queries with Boolean operators are supported, improving precision over the curated corpus.

6.7 Document clustering

The Entrez REST API was used to retrieve the supported set of PubMed articles for Darling v2.0. A graph representation of article relationships was constructed and HipMCL [12], a graph-based clustering algorithm, was applied to cluster articles into densely connected groups. This precomputed step facilitates more efficient interpretation and exploration. When a user submits a query, retrieved articles are presented in annotated form with highlighted entities and grouped clusters (Figure 3A).

6.8 Enhanced functional enrichment and network visualization

Darling v2.0 was substantially extended by integrating external applications for advanced visualization and functional enrichment. Arena3Dweb [13,14] enables conversion of Darling 2D networks into interactive 3D multilayer networks, with each layer representing an entity type. Additionally, integration with Flame [15,16] provides functional enrichment for gene sets highlighted in Darling, with multiple enrichment engines available (aGOtool [17], g:Profiler [18], WebGestalt [19], enrichR [20]) and multiple visualization options (bar plots, heatmaps, networks, scatter plots).

6.9 Implementation

Darling is built around a MySQL database updated annually. The graphical user interface (GUI) and backend are developed mainly in R/Shiny. Interactive network visualization is implemented with R/visNetwork [60], while network topology analysis uses R/igraph [21]. Plots are produced with R/Plotly and word clouds with R/wordcloud2. The EXTRACT API provides entity pop-ups in annotated abstracts. Arena3Dweb and Flame v2.0 APIs support 3D visualization and functional enrichment, respectively.

6.10 Example use case

Search scenario for rs6983267 and targeted exploration in colorectal cancer

As a demonstration, we outline a use case with SNP rs6983267 (8q24), which has been reported as a risk factor for colorectal cancer in meta-analyses and clinico-biological studies. The user selects the SNP search channel and submits rs6983267. The system normalizes the query into structured information (genomic coordinates and gene mappings) using established identifier-to-genome mapping resources. Relevant publications are retrieved from the supported sources with stored metadata (title, authors, journal, year, identifiers) and, where available, open full text. The user then filters results to cancer and specifically colorectal cancer, ensuring thematic focus. Darling performs NER and sentence-level co-occurrence analysis to generate association networks among entities (genes, proteins, diseases, tissues, chemicals and functional terms). Results include (i) the list of retrieved publications, (ii) searchable tables of co-mentioned entities in the selected corpus, and (iii) configurable 2D/3D networks for iterative exploration. The process can be periodically repeated to incorporate new publications as bibliographic sources are updated.

6.11 Publication

This work has been published as:

Darling (v2.0): Mining disease-related databases for the detection of biomedical entity associations. Baltoumas FA, Karatzas E, Venetsianou NK, Aplakidou E, Giatras K, Chasapi MN, Chasapi IN, Iliopoulos I, Iconomidou VA, Trougakos IP, Psomopoulos F, Giannakakis A, Georgakopoulos-Soares I, Kontou P, Bagos PG, Pavlopoulos GA. *Computational and Structural Biotechnology Journal*. 2025 Jun 14;27:2626–2637. doi:10.1016/j.csbj.2025.06.025. PMID:40599243.

Conclusions

This deliverable demonstrates an end-to-end, scalable workflow for extracting and organizing mutation–cancer evidence from the biomedical literature in support of precision medicine. The LLM-based pipeline enables the transformation of unstructured PubMed abstracts into an evaluated, machine-readable knowledge base of gene–variant–cancer associations, prioritizing high precision to support reliable downstream reuse and integration with other resources. In parallel, Darling (v2.0) provides a complementary disease-centric literature-mining framework that supports large-scale retrieval and interactive exploration of sentence-level co-occurrence networks, enabling hypothesis generation and evidence-oriented navigation across disease domains. Together, these two components provide both structured, high confidence relation tuples and an exploratory environment for contextual interpretation, forming a unified foundation for continuous updates, resource enrichment, and broader adoption of precision oncology workflows within WP11.

REFERENCES

[1] Hunter L, Cohen KB. Biomedical language processing: what's beyond PubMed? Molecular Cell. 2006;21(5):589–594.

doi:10.1016/j.molcel.2006.02.012

[2] Garraway LA, Verweij J, Ballman KV. Precision oncology: an overview. Journal of Clinical Oncology. 2013;31(15):1803–1805.

doi:10.1200/JCO.2013.49.4799

[3] Singhal K, et al. Large language models encode clinical knowledge. Nature. 2023;620:172–180.

doi:10.1038/s41586-023-06291-2

[4] Canese K, Weis S. PubMed: the bibliographic database. Nucleic Acids Research. 2013;41(D1):D20–D26.

doi:10.1093/nar/gks1149

[5] Qi P, et al. Stanza: A Python Natural Language Processing Toolkit for Many Human Languages. ACL System Demonstrations. 2020.

doi:10.18653/v1/2020.acl-demos.14

[6] Sherry ST, et al. dbSNP: the NCBI database of genetic variation. Nucleic Acids Research. 2001;29(1):308–311.

doi:10.1093/nar/29.1.308

[7] MongoDB Inc. MongoDB Documentation. https://www.mongodb.com/docs/

[8] Pafilis E, et al.EXTRACT 2.0: text-mining-assisted interactive annotation of biomedical named entities and ontology terms. (2017) bioRxiv

[9] Wolf T, et al. Transformers: State-of-the-Art Natural Language Processing. EMNLP. 2020.

doi:10.18653/v1/2020.emnlp-demos.6

[10] Reimers N, Gurevych I. Reporting score distributions makes a difference: Performance study of sequence tagging models. EMNLP. 2017.

doi:10.18653/v1/D17-1035

[11] Cock PJA, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, et al. Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics 2009;25:1422–3. https://doi.org/10.1093/bioinformatics/btp163.

[12] Azad A, Pavlopoulos GA, Ouzounis CA, Kyrpides NC, Buluç A. HipMCL: a high-performance parallel implementation of the Markov clustering algorithm for large-scale networks. Nucleic Acids Res 2018;46:e33. https://doi.org/10.1093/nar/gkx1313.

[13] Karatzas E, Baltoumas FA, Panayiotou NA, Schneider R, Pavlopoulos GA. Arena3Dweb: interactive 3D visualization of multilayered networks. Nucleic Acids Res 2021. https://doi.org/10.1093/nar/gkab278.

[14] Kokoli M, Karatzas E, Baltoumas FA, Schneider R, Pafilis E, Paragkamian S, et al. Arena3D web : Interactive 3D visualization of multilayered networks supporting multiple directional information channels, clustering analysis and application integration. Bioinformatics; 2022. https://doi.org/10.1101/2022.10.01.510435.

[15] Thanati F, Karatzas E, Baltoumas FA, Stravopodis DJ, Eliopoulos AG, Pavlopoulos GA. FLAME: A Web Tool for Functional and Literature Enrichment Analysis of Multiple Gene Lists. Biology (Basel) 2021;10:665. https://doi.org/10.3390/biology10070665.

[16] Karatzas E, Baltoumas FA, Aplakidou E, Kontou PI, Stathopoulos P, Stefanis L, et al. Flame (v2.0): advanced integration and interpretation of functional enrichment results from multiple sources. Bioinformatics 2023;39:btad490. https://doi.org/10.1093/bioinformatics/btad490.

[17] Schölz C, Lyon D, Refsgaard JC, Jensen LJ, Choudhary C, Weinert BT. Avoiding abundance bias in the functional annotation of post-translationally modified proteins. Nat Methods 2015;12:1003–4. https://doi.org/10.1038/nmeth.3621.

[18] Raudvere U, Kolberg L, Kuzmin I, Arak T, Adler P, Peterson H, et al. g:Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update). Nucleic Acids Res 2019;47:W191–8. https://doi.org/10.1093/nar/gkz369.

[19] Wang J, Vasaikar S, Shi Z, Greer M, Zhang B. WebGestalt 2017: a more comprehensive, powerful, flexible and interactive gene set enrichment analysis toolkit. Nucleic Acids Research 2017;45:W130–7. https://doi.org/10.1093/nar/gkx356.

[20] Chen EY, Tan CM, Kou Y, Duan Q, Wang Z, Meirelles GV, et al. Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. BMC Bioinformatics 2013;14:128. https://doi.org/10.1186/1471-2105-14-128.

[21] Gabor Csardi, Tamas Nepusz. The igraph software package for complex network research. InterJournal 2006;Complex Systems:1695.