A machine learning-powered dashboard for the exploration of high-throughput transcriptomic datasets

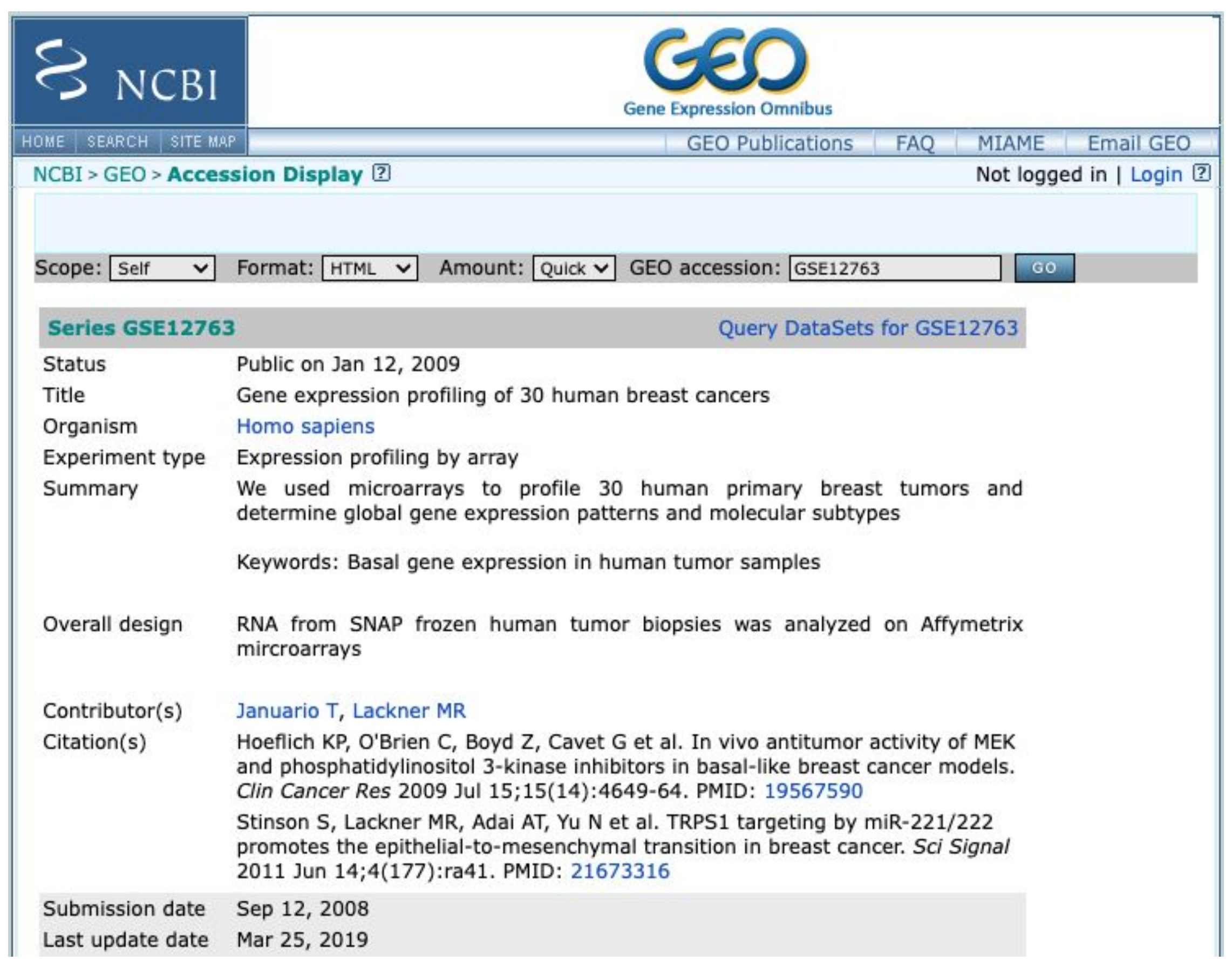
V. Bernu[1], C. Lescure[1], H. Brull Corretger[1], P. Dhillon[1], E. Fox[1], C. Marijon[1], A. Nordor[1], C. Petit[1], A. Behdenna[1]
[1]Epigene Labs, Paris, France

Introduction

- NCBI's[1] GEO[2] database is a major repository for **high-throughput transcriptomic datasets[3] referencing ~7,000,000 samples[4]**, including a significant number of tumor biopsy samples[4].
- Today, this invaluable database is underused because of technical roadblocks[I].
- We demonstrate that recent AI developments unlock this vast potential to **infer new biological understanding and shape future clinical study designs.**

Key results

- Our high-performance and state-of-the-art **AI models** identify the most relevant GEO datasets in oncology, representing the **top 10%**, which we refer to as **Epigene Labs data lake**.
- These models are integrated into a filtering and prioritization pipeline (Fig. 2), designed and developed for high scalability, modularity, and updatability.
- **A GEO[2] dashboard** (Fig. 3) enables a user-friendly interface with the Epigene Labs data lake by aggregating the results of the above-mentioned models.



NCBI > GEO > **Accession Display**

Not logged in | Login

Scope: Self Format: HTML Amount: Quick GEO accession: GSE12763 GO

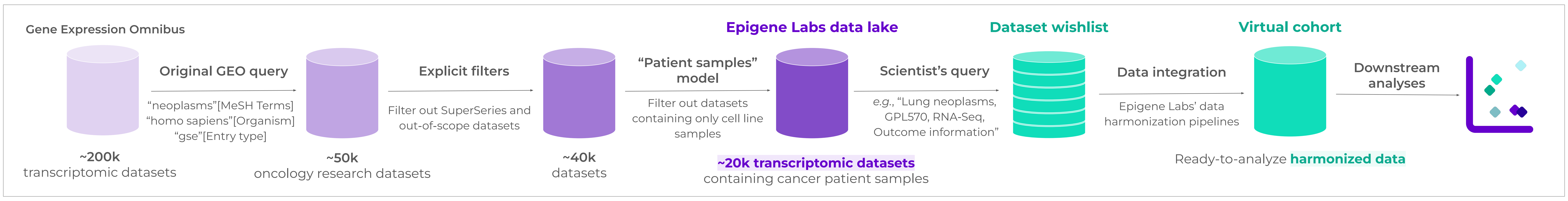
Series GSE12763 Query DataSets for GSE12763

Status	Public on Jan 12, 2009
Title	Gene expression profiling of 30 human breast cancers
Organism	Homo sapiens
Experiment type	Expression profiling by array
Summary	We used microarrays to profile 30 human primary breast tumors and determine global gene expression patterns and molecular subtypes Keywords: Basal gene expression in human tumor samples
Overall design	RNA from SNAP frozen human tumor biopsies was analyzed on Affymetrix mircroarrays
Contributor(s)	Januario T, Lackner MR
Citation(s)	Hoeflich KP, O'Brien C, Boyd Z, Cavet G et al. In vivo antitumor activity of MEK and phosphatidylinositol 3-kinase inhibitors in basal-like breast cancer models. *Clin Cancer Res* 2009 Jul 15;15(14):4649-64. PMID: 19567590 Stinson S, Lackner MR, Adai AT, Yu N et al. TRPS1 targeting by miR-221/222 promotes the epithelial-to-mesenchymal transition in breast cancer. *Sci Signal* 2011 Jun 14;4(177):ra41. PMID: 21673316
Submission date	Sep 12, 2008
Last update date	Mar 25, 2019

Figure 1: Screenshot of the description of a GEO[2] dataset

Lexicon

1. **NCBI**: National Center for Biotechnology Information[IV].
2. **GEO**: The Gene Expression Omnibus is a public genomics data repository.
3. **Dataset**: Refers to a GEO[2] Series, (GSExxx), it is "an original submitter-supplied record that summarizes a study"[IV].
4. **Sample**: "A Sample record describes the conditions under which an individual Sample was handled. Each Sample record is assigned a unique and stable GEO accession number (GSMxxx)"[IV].
5. **Technology**: Method for the production of the gene expression profiles (*e.g.*, Microarray or RNA-seq).
6. **Platform:** "A Platform record is composed of a summary description of the array or sequencer and, for array-based Platforms, a data table defining the array template. Each Platform record is assigned a unique and stable GEO accession number (GPLxxx)"[IV].
7. **NER and NEN**: Named Entity Recognition and Named Entity Normalization are natural language processing methods.
8. **LLM**: Large Language Model (*e.g.*, GPT-4[II] or Mistral[III] models).
9. **AUC**: Area Under the receiver operating characteristic Curve. It is a score between 0 and 1.



Gene Expression Omnibus

~200k
transcriptomic datasets

Original GEO query
"neoplasms"[MeSH Terms]
"homo sapiens"[Organism]
"gse"[Entry type]

~50k
oncology research datasets

Explicit filters
Filter out SuperSeries and out-of-scope datasets

~40k
datasets

"Patient samples" model
Filter out datasets containing only cell line samples

Epigene Labs data lake
~20k transcriptomic datasets
containing cancer patient samples

Scientist's query
e.g., "Lung neoplasms, GPL570, RNA-Seq, Outcome information"

Dataset wishlist

Data integration
Epigene Labs' data harmonization pipelines

Virtual cohort

Downstream analyses

Ready-to-analyze **harmonized data**

Figure 2: End-to-end pipeline, from heterogeneous datasets[3] to harmonized virtual cohorts

Methods

Classification models

Our models (Table 1) classify datasets[3] based on ground truth data that were labeled by cancer scientists. They are divided into two categories:

- **Dataset description**-based models:
  - Utilize only dataset-level descriptions (Fig. 1).
  - Offer high scalability to encompass the entirety of Epigene Labs data lake.
- **Sample description**-based models:
  - Require detailed sample-level descriptions.
  - Demand increased computational power but deliver enhanced precision in results.

Large Language Models (LLMs[8])

Adaptation for treatment-related models:

- Traditional training approaches were infeasible due to limited labeled data.
- Prompt engineering strategies harness the capabilities of **pre-trained LLMs** from OpenAI[II] and Mistral[III].

Performance metrics

The models are evaluated by the metrics below, depending on the format of their respective output. All three are scores between 0 and 1.

- Classic binary classification: **AUC[9]** (the higher the better).
- Multiclass and LLM-based binary classification: **F1-score** (the higher the better).
- Multilabel classification: **Hamming loss** (the lower the better).

Classifier	Methods	Input	Output	N train	N test	Performance
Cancer type	NER[7], NEN[7] and ML classifier	Dataset description	Various indications (21 possibilities)	1254	295	F1-score 0.91
RNA-Seq resolution	Custom embeddings and ML classifier	Dataset description and custom features. Only for RNA-Seq	Bulk / Single-cell	216	55	AUC[9] 0.97
Treatment information	LLM[8]	Dataset description and PubMed abstract	True / False	NA	30	F1-score 0.87
Treatment timing	LLM	Dataset description and PubMed abstract	Pre-treatment / Post-treatment / Both / Not available	NA	12	F1-score 0.75
Treatment type	LLM	Dataset description and PubMed abstract	List of treatment classes	NA	8	Hamming 0.28
Patient samples (dataset version)	Custom embeddings and ML classifier	Dataset description and custom features	True / False	2657	378	AUC 0.96
Patient samples (sample version)	Custom embeddings and ML classifier	Dataset & sample descriptions	True / False	206	95	AUC 0.97
Outcome information	Custom embeddings and ML classifier	Dataset & sample descriptions	Available / Not available	199	108	AUC 0.98
Donor type	Custom embeddings and ML classifier	Dataset & sample descriptions	Healthy donor and cancer patient / Other	237	96	AUC 0.85

Table 1: Description, methods, and performance evaluation of the models

GEO dashboard

- **Intuitive data mapping:** Our Metabase[V] dashboard visualizes the vast landscape of the Epigene Labs data lake, integrating diverse GEO explicit attributes and model-predicted attributes.
- **Dynamic filtering system:** With advanced filtering capabilities, the dashboard also permits in-depth interactive exploration of the data.
- **Seamless research integration:** The dashboard, combined with the AI models, empowers researchers to create tailored virtual cohorts for computational oncology (Fig. 2), sourced from the GEO database.

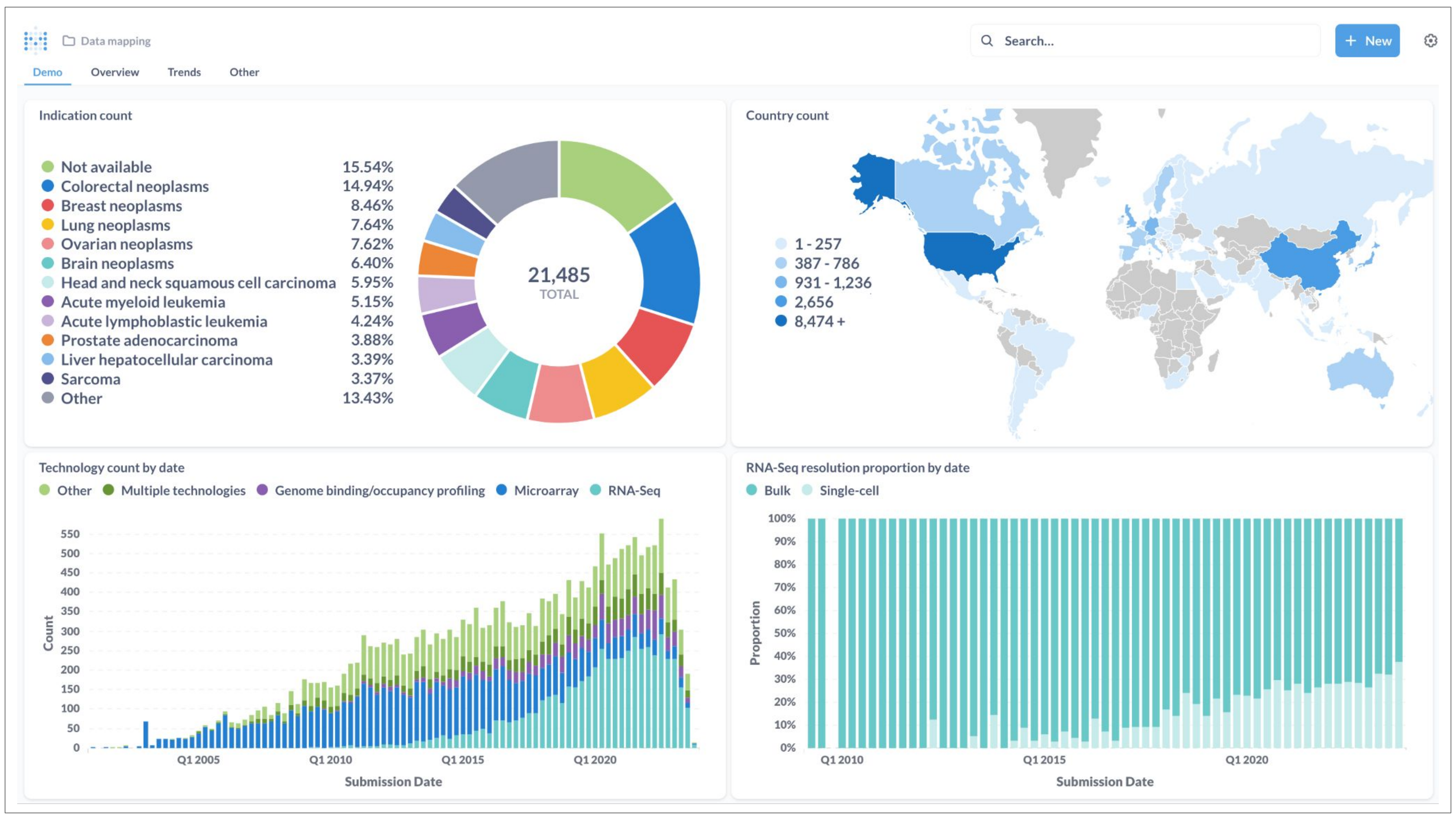


Figure 3: GEO Data Mapping Dashboard visualization example

Conclusion

AI techniques enable the annotation and exploration of the GEO[2] database, facilitating secondary analysis of oncology research omic data.

References

I. Hawkins, N., Maldaver, M., Yannakopoulos, A., Guare, L. & Krishnan, A. Systematic tissue annotations of genomics samples by modeling unstructured metadata. Nature Communications 13, (2022).
II. OpenAI et al. GPT-4 Technical Report. Preprint at http://arxiv.org/abs/2303.08774 (2023).
III. Jiang, A. Q. et al. Mistral 7B. Preprint at http://arxiv.org/abs/2310.06825 (2023).
IV. National Center for Biotechnology Information (NCBI)[Internet]. Bethesda (MD): National Library of Medicine (US), National Center for Biotechnology Information; [1988] – [cited 26 Feb 24]. Available from: https://www.ncbi.nlm.nih.gov/.
V. Metabase | Business Intelligence, Dashboards, and Data Visualization. https://www.metabase.com/

- **Contact: Akpéli Nordor, PharmD, PhD (akpeli@epigenelabs.com).**
- Final publication number: 2P.
- The authors have no conflict of interest to declare.