The evolving landscape of transcriptomics data

Abstract 1087 Valentin Bernu[1], Hèlia Brull Corretger[1], Charles Lescure[1], Abdelkader Behdenna[1], Julien Haziza[1], Léa Meunier[1], Clémence Petit[1], Akpéli Nordor[1]. [1]Epigene Labs, Paris, France

AACR American Association for Cancer Research

Cancer scientists can find available public data for their project thanks to our AI-powered tool

Why is public transcriptomic data untapped?

- **Lack of structure** - Plain text use limits standardization.
- **Inconsistent ontologies** – Synonyms, abbreviations, etc.
- **Heterogeneous databases** – Different structures make comprehensive searches challenging.

How valuable is secondary analysis of public data?

- It maximizes the **research impact** and **cost efficiency** of studies.
- It enhances **statistical power** and **generalizability.**
- It encourages **open science** & **collaboration.**

How can public data be used in research projects?

- **Define your criteria** – Select cancer type, mutations, treatments, etc.
- **Explore available data** – Request a live demo of our tool.
- **Integrate & analyze** – Leverage our capabilities for seamless use.

Contact

Akpéli Nordor, PharmD, PhD (akpeli@epigenelabs.com).

The authors have no conflict of interest to declare.

CHECK US OUT!

Ask us for a live demo!

AI techniques unlock access to public data

Harmonization

- **Raw data is structured and cleaned using NLP (Natural Language Processing) techniques** to ensure consistency and usability.
- Key metadata such as dataset name, publication date, technology, and location are retrieved directly from public databases without additional inference.

Classification models

- **Some attributes** (treatment, mutations, cancer type) **are not explicitly provided and require inference.**
- To address this, we developed both **rule-based and machine learning (ML) models in collaboration with cancer scientists,** some of them relying on proprietary ontologies.
- Our models (Table 1) classify datasets using ground truth labels curated by domain experts.

Classifier	Methods	Output	N test (datasets)	Metric	Performance
Cancer type	Custom embeddings and ML classifier	Various indications (21)	295	F1-score (weighted)	**0.91**
RNA-Seq resolution	Custom embeddings and ML classifier	Bulk / Single-cell	55	AUC	**0.97**
Specimen type	Custom embeddings and ML classifier	Patient / Preclinical model	378	AUC	**0.96**
Drug information	Pattern matching	Yes / No*	21	Accuracy	0.81
Drug type (multilabel)	Pattern matching	900 possible drug types*	14	Precision	0.84
Mutation information	Pattern matching	Yes / No*	21	Accuracy	0.76
Mutated genes (multilabel)	Pattern matching	X possible mutated genes*	10	Precision	0.64

Work in progress

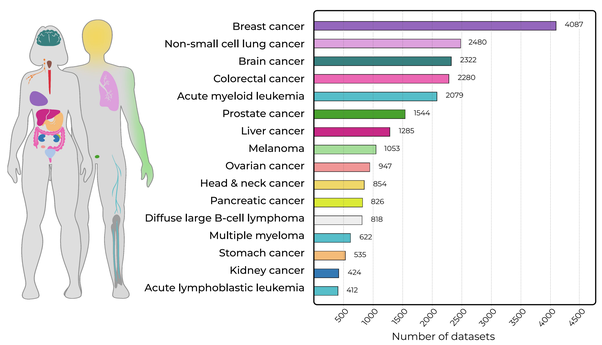
Table 1: Performance of the machine learning models, (*) indicates the use of proprietary ontologies.

Abundant and diverse public data is growing steadily

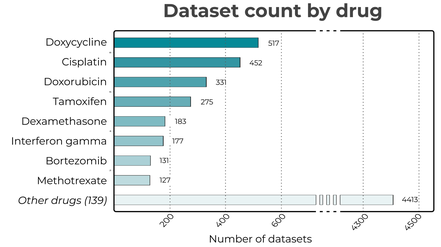
32,708 Datasets	1,264,622 Samples	33 Cancer types

Panels show transcriptomic datasets from GEO and ArrayExpress, filtered by model-predicted criteria and/or time, each illustrating a different distribution aspect.

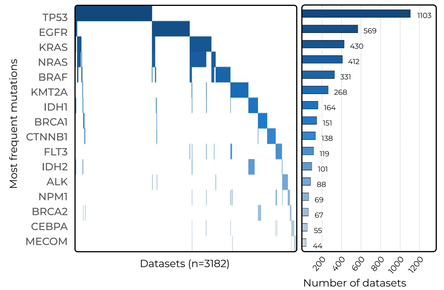
Dataset count for the most frequent cancer types



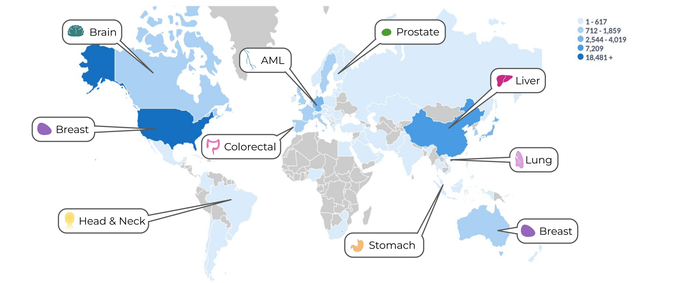
Dataset count by drug



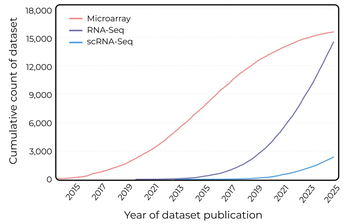
Dataset count by mutated gene



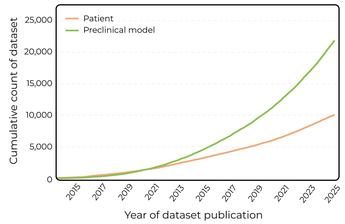
Number of datasets with the most frequent cancer type by country



Cumulative dataset count by technology



Cumulative dataset count by specimen type



Our user-friendly tool allows to find the data matching a research question

Queries examples

Seamless searches

Automated reporting

Search of a specific dataset with defined characteristics

- Breast neoplasms
- scRNA-Seq
- Olaparib drug
- BRCA1 mutation

Technology: scRNA-seq | Omic: Transcriptomic | Cancer Type: Breast cancer | Drug information: True | Drug: Olaparib | Mutated genes: BRCA1 | Batch | Source | Platform | Patient sample | Status

Mapped from public databases

1	1	16
Nb of databases	Datasets	Samples

Identify datasets matching the query

NIH GDC | GEO Gene Expression Omnibus | ArrayExpress

Dataset GSEXXXX match the query.

View all available datasets for a cancer type

- Acute myeloid leukemia
- Get the technology distribution
- Get the drug distribution

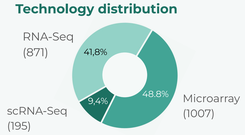
Technology | Omic: Transcriptomic | Cancer Type: Acute myeloid leukemia | Platform | Mutated genes | Batch | Drug information | Source | Drug | Patient sample | Status

Mapped from public databases

4	2,079	74,352
Nb of databases	Datasets	Samples

Overview of datasets features

Technology distribution



Drugs identified

- Imatinib
- Doxycycline
- Arsenic trioxide
- Bortezomib
- Interferon gamma
- Dexamethasone
- Cyclophosphamide
- Rituximab
- Dasatinib
- Sorafenib
- Tamoxifen

See Other (58)

The evolving landscape of transcriptomics data

Abstract 1087 Valentin Bernu[1], Hèlia Brull Corretger[1], Charles Lescure[1], Abdelkader Behdenna[1], Julien Haziza[1], Léa Meunier[1], Clémence Petit[1], Akpéli Nordor[1]. [1]Epigene Labs, Paris, France

AACR American Association for Cancer Research

Cancer scientists can find available public data for their project thanks to our AI-powered tool

Why is public transcriptomic data untapped?

- **Lack of structure** - Plain text use limits standardization.
- **Inconsistent ontologies** – Synonyms, abbreviations, etc.
- **Heterogeneous databases** – Different structures make comprehensive searches challenging.

How valuable is secondary analysis of public data?

- It maximizes the **research impact** and **cost efficiency** of studies.
- It enhances **statistical power** and **generalizability.**
- It encourages **open science** & **collaboration.**

How can public data be used in research projects?

- **Define your criteria** – Select cancer type, mutations, treatments, etc.
- **Explore available data** – Request a live demo of our tool.
- **Integrate & analyze** – Leverage our capabilities for seamless use.

Contact

Akpéli Nordor, PharmD, PhD (akpeli@epigenelabs.com).

The authors have no conflict of interest to declare.

CHECK US OUT!

Ask us for a live demo!

AI techniques unlock access to public data

Harmonization

- **Raw data is structured and cleaned using NLP (Natural Language Processing) techniques** to ensure consistency and usability.
- Key metadata such as dataset name, publication date, technology, and location are retrieved directly from public databases without additional inference.

Classification models

- **Some attributes** (treatment, mutations, cancer type) **are not explicitly provided and require inference.**
- To address this, we developed both **rule-based and machine learning (ML) models in collaboration with cancer scientists,** some of them relying on proprietary ontologies.
- Our models (Table 1) classify datasets using ground truth labels curated by domain experts.

Classifier	Methods	Output	N test (datasets)	Metric	Performance
Cancer type	Custom embeddings and ML classifier	Various indications (21)	295	F1-score (weighted)	**0.91**
RNA-Seq resolution	Custom embeddings and ML classifier	Bulk / Single-cell	55	AUC	**0.97**
Specimen type	Custom embeddings and ML classifier	Patient / Preclinical model	378	AUC	**0.96**
Drug information	Pattern matching	Yes / No*	21	Accuracy	0.81
Drug type (multilabel)	Pattern matching	900 possible drug types*	14	Precision	0.84
Mutation information	Pattern matching	Yes / No*	21	Accuracy	0.76
Mutated genes (multilabel)	Pattern matching	X possible mutated genes*	10	Precision	0.64

Work in progress

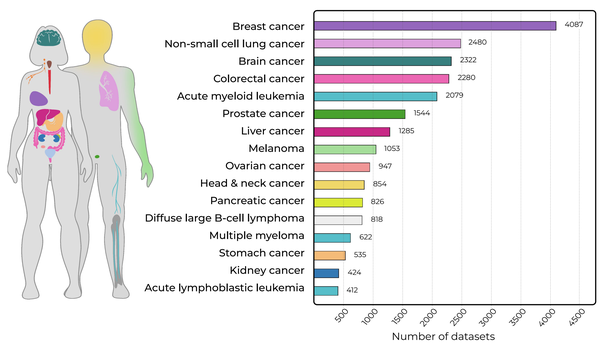
Table 1: Performance of the machine learning models, (*) indicates the use of proprietary ontologies.

Abundant and diverse public data is growing steadily

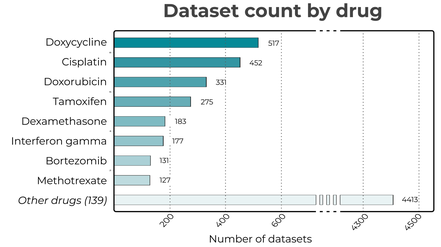
32,708 Datasets	1,264,622 Samples	33 Cancer types

Panels show transcriptomic datasets from GEO and ArrayExpress, filtered by model-predicted criteria and/or time, each illustrating a different distribution aspect.

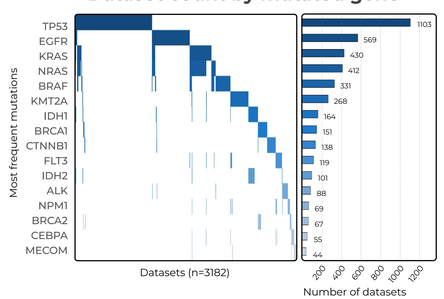
Dataset count for the most frequent cancer types



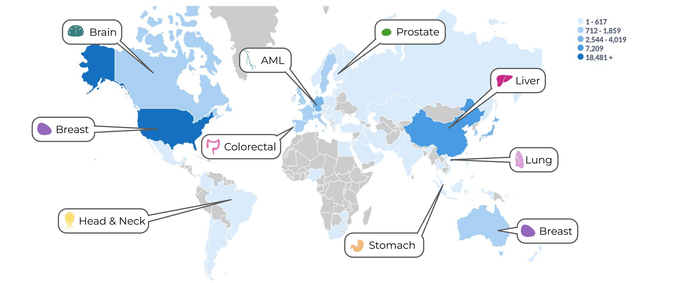
Dataset count by drug



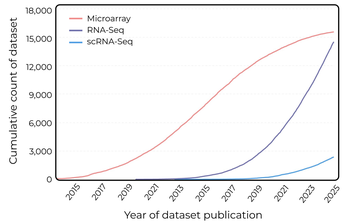
Dataset count by mutated gene



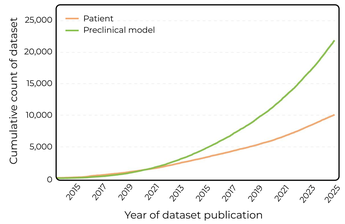
Number of datasets with the most frequent cancer type by country



Cumulative dataset count by technology



Cumulative dataset count by specimen type



Our user-friendly tool allows to find the data matching a research question

Queries examples

Seamless searches

Automated reporting

Search of a specific dataset with defined characteristics

- Breast neoplasms
- scRNA-Seq
- Olaparib drug
- BRCA1 mutation

Technology: scRNA-seq | Omic: Transcriptomic | Cancer Type: Breast cancer | Drug information: True | Drug: Olaparib | Mutated genes: BRCA1 | Batch | Source | Platform | Patient sample | Status

Mapped from public databases

1	1	16
Nb of databases	Datasets	Samples

Identify datasets matching the query

NIH GDC | GEO Gene Expression Omnibus | ArrayExpress

Dataset GSEXXXX match the query.

View all available datasets for a cancer type

- Acute myeloid leukemia
- Get the technology distribution
- Get the drug distribution

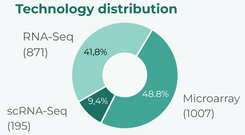
Technology | Omic: Transcriptomic | Cancer Type: Acute myeloid leukemia | Platform | Mutated genes | Batch | Drug information | Source | Drug | Patient sample | Status

Mapped from public databases

4	2,079	74,352
Nb of databases	Datasets	Samples

Overview of datasets features

Technology distribution



Drugs identified

- Imatinib
- Doxycycline
- Arsenic trioxide
- Bortezomib
- Interferon gamma
- Dexamethasone
- Cyclophosphamide
- Rituximab
- Dasatinib
- Sorafenib
- Tamoxifen

See Other (58)