Machine-learning-based inference of clinical metadata from gene expression data

Abstract 7441 Valentin Bernu[1], Guillaume Appé[1], Abdelkader Behdenna[1], Antoine Gaston[1], Julien Haziza[1], Léa Meunier[1], Akpéli Nordor[1]. [1]Epigene Labs, Paris, France

Contact
Akpéli Nordor, PharmD, PhD
akpeli@epigenelabs.com

The authors have no conflict of interest to declare.

AACR American Association for Cancer Research

Can a single generalist pipeline infer multiple clinical variables from transcriptomic data?

CHECK US OUT!

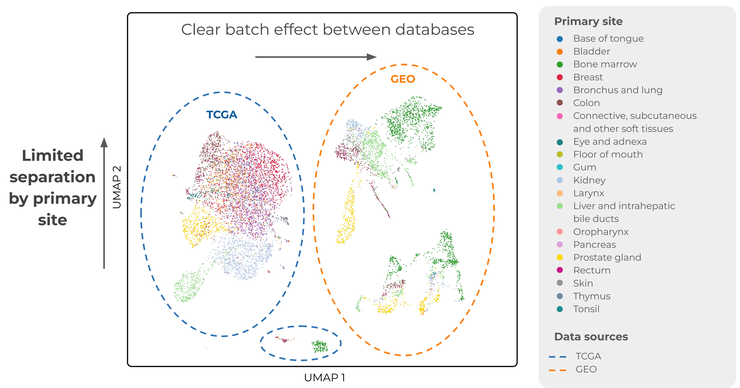
Inferring clinical information from transcriptomic data

Challenge

Metadata label transfer across independent transcriptomic datasets remains challenging, primarily due to batch effects which are particularly significant between datasets originating from different reference databases (Figure A).

- To address this, we present a **pipeline that infers clinical variables** (sex, stage, cancer type, primary site, and histological type) **from transcriptomic profiles**.

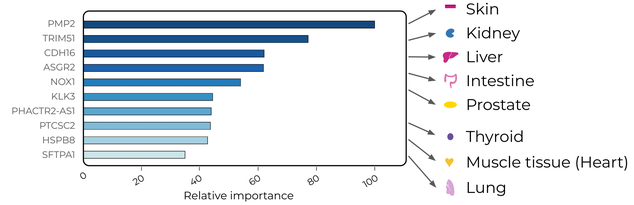
A UMAP projection of transcriptomic data from GEO and TCGA with 28286 genes



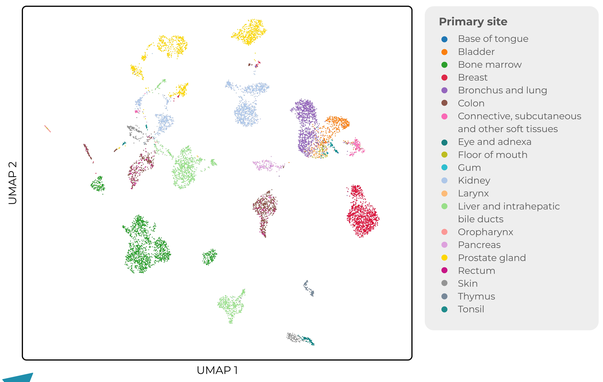
Validation of the models through biological evidence

Primary site model (B & C)

B Top 10 most important features

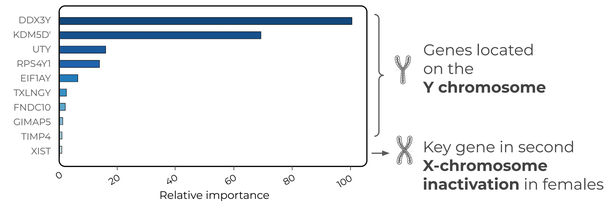


C UMAP projection of transcriptomic data with 100 genes with the highest feature importance for the primary site model

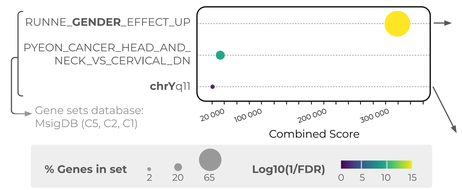


Sex model (D & E)

D Top 10 most important features



E Gene enrichment analysis (EnrichR) of 100 genes with the highest importance for sex model vs 10k background genes



- **High-importance genes identified by the model improve primary site separation** (Figures A & C).
- **Identified genes and pathways in each model** align with key biological mechanisms, supporting their **biological relevance and validity** (Figures B, D, E).

Applications

- **Clinical metadata quality control.**
- **Labeling samples with missing clinical metadata.**

Limitations

The lower performance on rare classes is not visible in overall accuracy and accuracy can be inflated for variables with fewer classes.

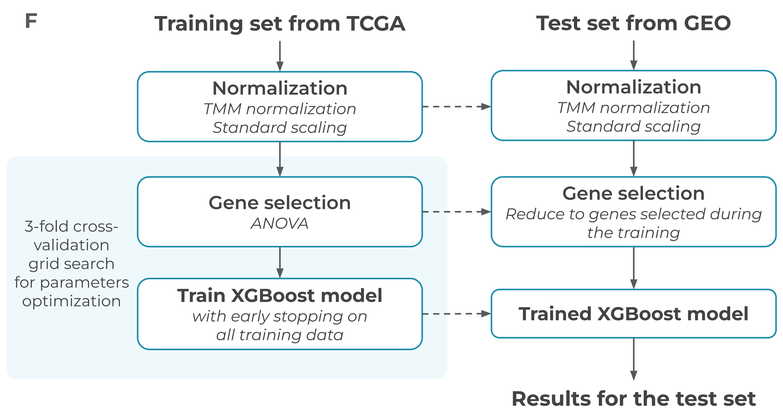
Future Work

- **Explore batch effect correction methods** without covariates or using predicted primary site as a proxy covariate to improve performance.
- **Develop cancer-specific models**, particularly to improve Stage inference where general models seem to underperform.
- Employ stricter criteria for gene selection to **enhance model applicability**.

Building novel predictive models for clinical data inference

Description of the workflow used to generate the five new models

F



Cross-database pan-cancer inference approach

- **This study** focus on predicting **five clinical variables** in a generalized pan-cancer approach to infer clinical data.

Metadata	Genes (test)	Cross-validation			Test		
		Samples	Classes	Performance	Samples	Classes	Performance
Sex	28286	9887	2	100% accuracy	3187	2	**99%** accuracy
Cancer type	10292	9925	27	97% accuracy	4233	11	**81%** accuracy
	28286	-	-	-	5868	4	**91%** accuracy
Primary Site	28286	9887	33	92% accuracy	3938	20	**89%** accuracy
Histological Type	28286	9913	20	93% accuracy	3458	7	**94%** accuracy
Stage	28286	9887	2	71% accuracy	440	2	49% accuracy (work in progress)

Models trained and validated on TCGA

Models tested on GEO datasets
(e.g. 100 datasets for histological type)

Good performance across all variables
-except for Stage

Consistent results

Comparison with existing cross-dataset models trained on TCGA

- Most **previous studies** focus on **predicting the tissue of origin, often in metastatic cancers with unknown origin**, to improve diagnosis and outcomes.

Papers	Metadata	Cross-validation			Test		
		Samples	Classes	Performance	Samples	Classes	Performance
Van et al., 2024	**Tissue type**	7192	14	Undisclosed	876 (ICGC & 4 GEO datasets)	6	0.80 weighted F1-score
Chen et al., 2021	**Cancer type**	7715	21	96.38% R2 score	42 (1 GEO dataset)	5	83.3% R2 score
He et al., 2023	**Cancer type**	9911	32	97.50% accuracy	1988 (ICGC)	10	82.67% accuracy
Zhao et al., 2020	**Cancer type**	18217 (+ ICGC)	32	98.54% accuracy	23 & 69 (JAX & Melbourne)	16 & 18	86.96% & 72.46% accuracy
	Molecular subtype	3367	11	4 classifiers 60% - 83.5% (acc.)	1784 & 215 (1 EGA & 1 GEO dataset)	4 & 4	84.19% ovarian subtype 79.88% breast subtype
Elmahy et al., 2021	**Stage**	TCGA-KIRC	4	82% accuracy	-	-	-

To our knowledge, **no existing approach performs pan-cancer inference across multiple clinical variables**.

Effective within databases

Limited representativeness of test datasets compared to the TCGA distribution
→ Difficulty to properly evaluate cross-database performance