From data disparity to data harmony: A comprehensive pan-cancer omic data collection

Abstract 6209

Lea Meunier[1], Guillaume Appe[1], Abdelkader Behdenna[1], Valentin Bernu[1], Helia Brull Corretger[1], Prashant Dhillon[1], Eleonore Fox[1], Julien Haziza[1], Charles Lescure[1],
Camille Marijon[1], Clemence Petit[1], Solene Weill[1], Akpeli Nordor[1] - [1]Epigene Labs, Paris, France

AACR
American Association for Cancer Research

INTRODUCTION

- The exponential growth of omics datasets offers a significant opportunity for scientific advancement in cancer research.
- However, though the **lack of uniform standards**, in both clinical and omic data, hinder the effective utilization of these datasets, thus impeding our understanding of cancer biology and the development of innovative therapies.
- We have created a **novel collection of pan-cancer datasets** with **extensive clinical data harmonization** and **consistent omic data normalization.**
- This approach **enhances data quality**, and is also **cost-effective**, offering significant advantages in the realm of cancer research.

Here, we focused on patient-derived gene expression microarray datasets from the Gene Expression Omnibus[1] (GEO) database.

DATA COLLECTION PRESENTATION

470 Datasets	Average of 82 samples per dataset	30 Cancer types	45 Clinical data elements	32,825 Samples	20,347 Genes

Our data collection aims to encompass numerous cancer types alongside their corresponding non-tumoral tissue counterparts. **Healthy tissue** was favored over tumor adjacent tissue, to **minimize the risk of introducing biases** related to cancer patient background into downstream analyses.

Samples by Major Biopsy Site

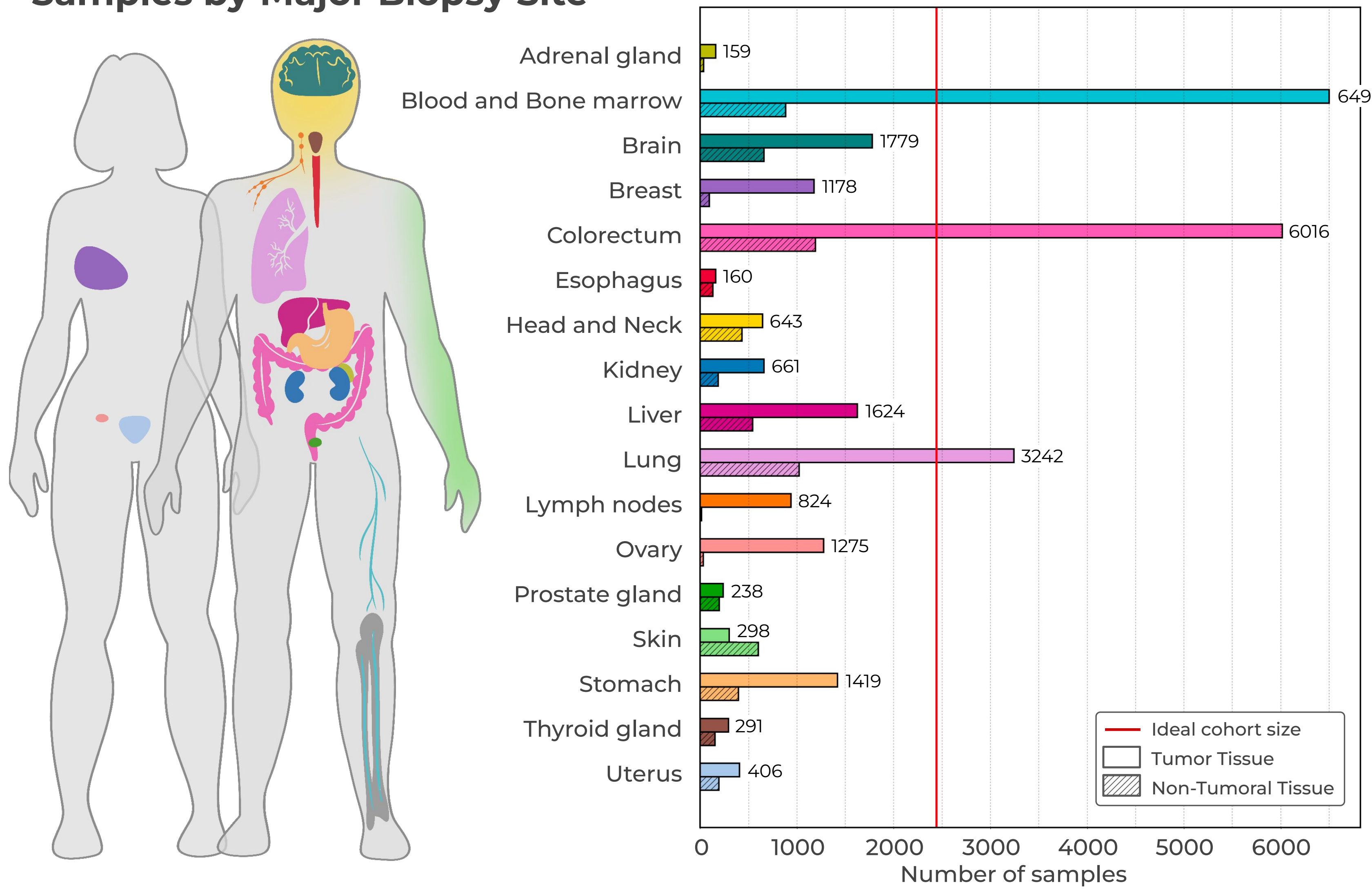


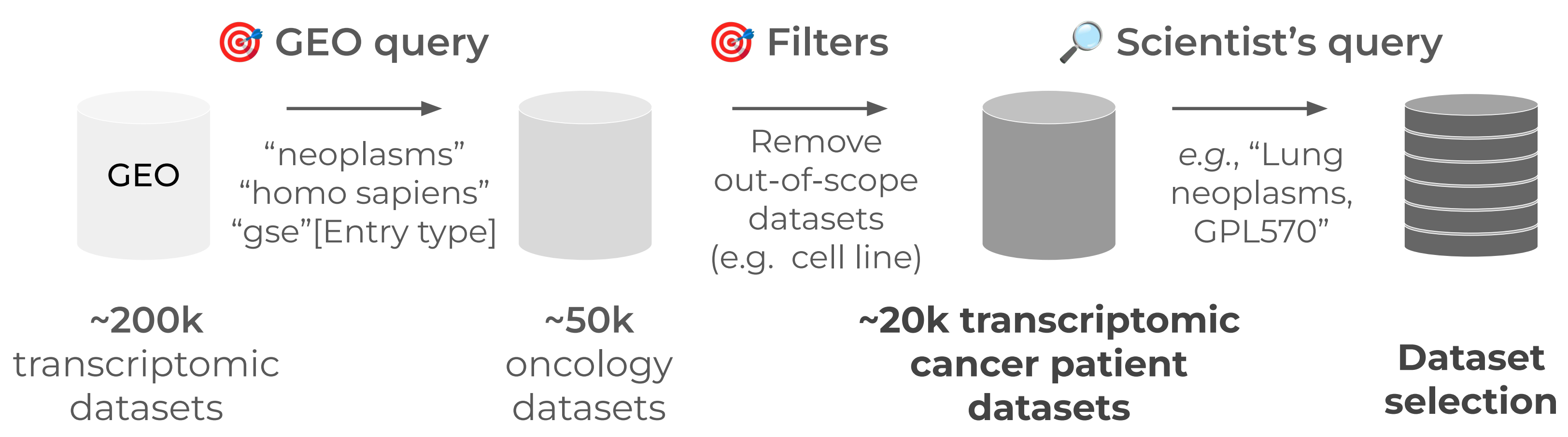
Figure 1: Distribution of samples by major biopsy site (n>150) and sample type

On average, GEO individual datasets typically hold around 60 samples. However, by adjusting the expected outcome of a Kolmogorov-Smirnov test to a target p-value of 5%, **we estimated the ideal cohort size** to study sub-population composition (theoretically set at 5) to be 2,441 samples (Fig. 1).

Surpassing the size of popular databases, 3 biopsy sites in our data meet the high cohort size limit. With ongoing data integration, **we anticipate surpassing this limit for various biopsy sites**, enhancing the robustness of our analyses.

MATERIAL AND METHODS

Dataset prioritization



AI-powered clinical data harmonization

Heterogeneous clinical file

Proprietary ontology

Format and attributes harmonization → AI-based data element values prediction → Quality control by human experts

Transcriptomic data processing

Microarray expression CEL file

Microarray platform description

Gene name databases (GENCODE, NCBI, Ensembl, HPA, UniProt, HGNC, UCSC)

Format and attributes harmonization → Expression data normalization (RMA) → Gene names harmonization (gene names mapping on gencode v36)

Data Aggregation

To aggregate data and build larger cohort, we use pyComBat[2] to **rectify for batch effects** on expression data. By including crucial covariates, such as phenotype, in the parameters, we **ensure the preservation of the biological signal**.

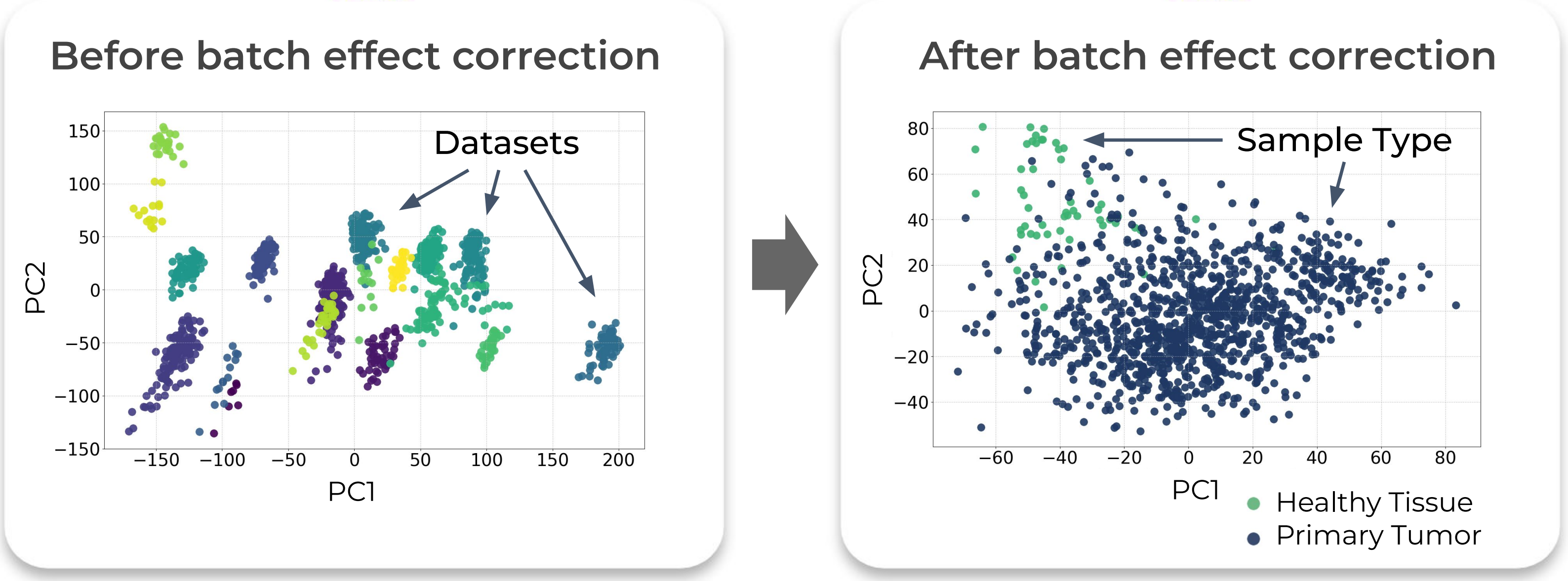


Figure 2: Principal component analysis on gene expression of breast invasive carcinoma cohort (n=1,146)

RESULTS

Data collection comparison with The Cancer Genome Atlas[3] (TCGA)

Cohorts were constructed based on cancer types, and then aligned with the TCGA projects. On average, these cohorts comprise **4.2x more samples** ([min 0.3; max 45.5], median 3.4).

Detailed comparisons were conducted on 8 cancer types, involving on average 19,129 shared genes. Notably, we observed a **100% overlap in gender-associated differentially expressed genes** between TCGA and our cohort.

			Comparison with TCGA matching project	
Cancer type	# datasets	# samples	Proportion of samples compare to TCGA	Expression correlation (Spearman)
Acute lymphoblastic leukemia	1	783	x	x
Acute myeloid leukemia	5	875	5.79	0.74
Ovarian serous carcinoma	14	1006	2.39	0.76
Liver hepatocellular carcinoma	29	725	1.95	0.75
Breast invasive carcinoma	18	1146	1.05	0.78
Uterine corpus endometrioid carcinoma	10	469	0.84	0.72
Skin melanoma	23	388	0.83	0.74
Pancreas adenocarcinoma	9	108	0.61	0.75

Breast invasive carcinoma cohort - Molecular subtype composition

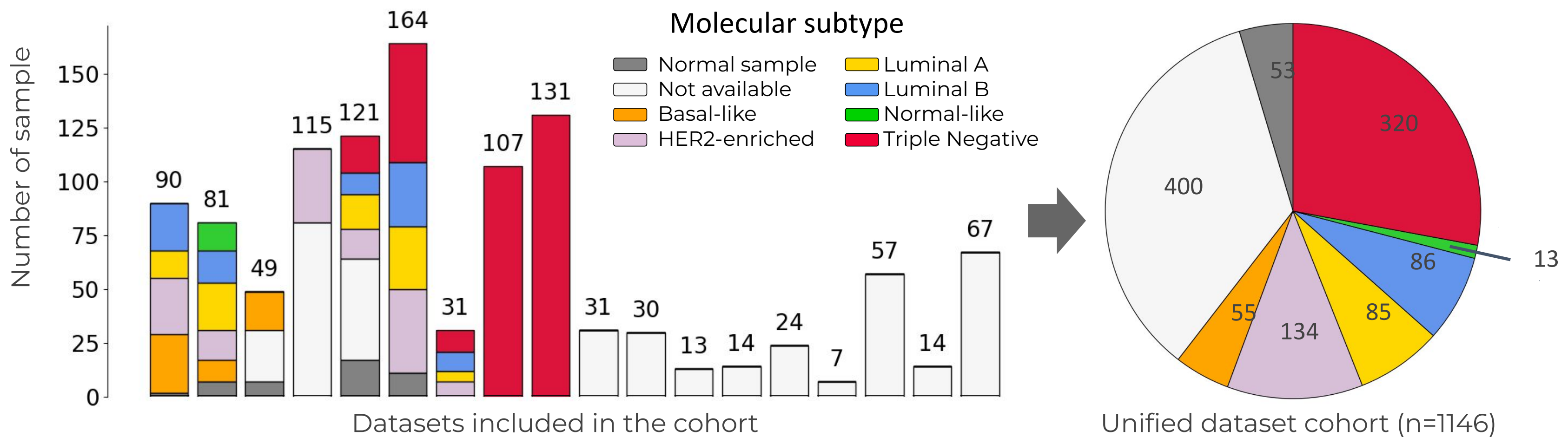


Figure 3: Molecular subtype composition of the breast invasive carcinoma cohort.
Each barplot represents a dataset and its composition. The pie chart represents the composition of the aggregated cohort.

By consolidating diverse datasets, we create a cohort with a more comprehensive molecular subtype composition.

CONCLUSIONS

- **Leveraging diverse cohorts for target discovery**: This study demonstrates the successful utilization of seven unique cohorts within a target discovery project. (see poster #1915)
- **Cross-platform validation:** The observed consistency between RNA-seq and microarray data from these cohorts underscores the reliability and complementary nature of these technologies.
- **Future directions:** Building upon this success, this project will continue to integrate microarray datasets alongside pan-cancer RNA-seq and single-cell data. This initiative paves the way for future expansion, incorporating a wider spectrum of omics datasets.

REFERENCES

[1] Edgar R, Domrachev M, Lash AE. Gene Expression Omnibus: *NCBI gene expression and hybridization array data repository,* Nucleic Acids Res. 2002 Jan 1;30(1):207-10

[2] Behdenna A, Colange M, Haziza J *et al. pyComBat, a Python tool for batch effects correction in high-throughput molecular data using empirical Bayes methods*. BMC Bioinformatics. 2023 Dec 7; 24, 459.

[3] The results shown here are partially based upon data generated by the TCGA Research Network: https://www.cancer.gov/tcga.

CONTACT

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

CHECK
OUT OUR
WEBSITE!

European Innovation Council
Co-funded by the European Union