UAB MEDICINE
INFORMATICS INSTITUTE

ai.med

Statistical Enrichment Analysis of Sample Clinical Attributes Using SEAS

Jake Y. Chen, PhD, FACMI, FAIMBE
Professor & Chief Bioinformatics Officer
The University of Alabama at Birmingham, USA
jakechen@uab.edu | bio.informatics.uab.edu

July 29th 2022

TECHNOLOGY AND CODE article

Front. Big Data, 16 September 2021 | https://doi.org/10.3389/fdata.2021.725276

Statistical Enrichment Analysis of Samples: A General-Purpose Tool to Annotate Metadata Neighborhoods of Biological Samples

Thanh M. Nguyen[1], **Samuel Bharti**[2], **Zongliang Yue**[1], **Christopher D. Willey**[3†] and **Jake Y. Chen**[1*]

[1]Informatics Institute, School of Medicine, The University of Alabama at Birmingham, Birmingham, AL, United States

[2]Centre for Computational Biology and Bioinformatics, Amity Institute of Biotechnology, Amity University, Noida, India

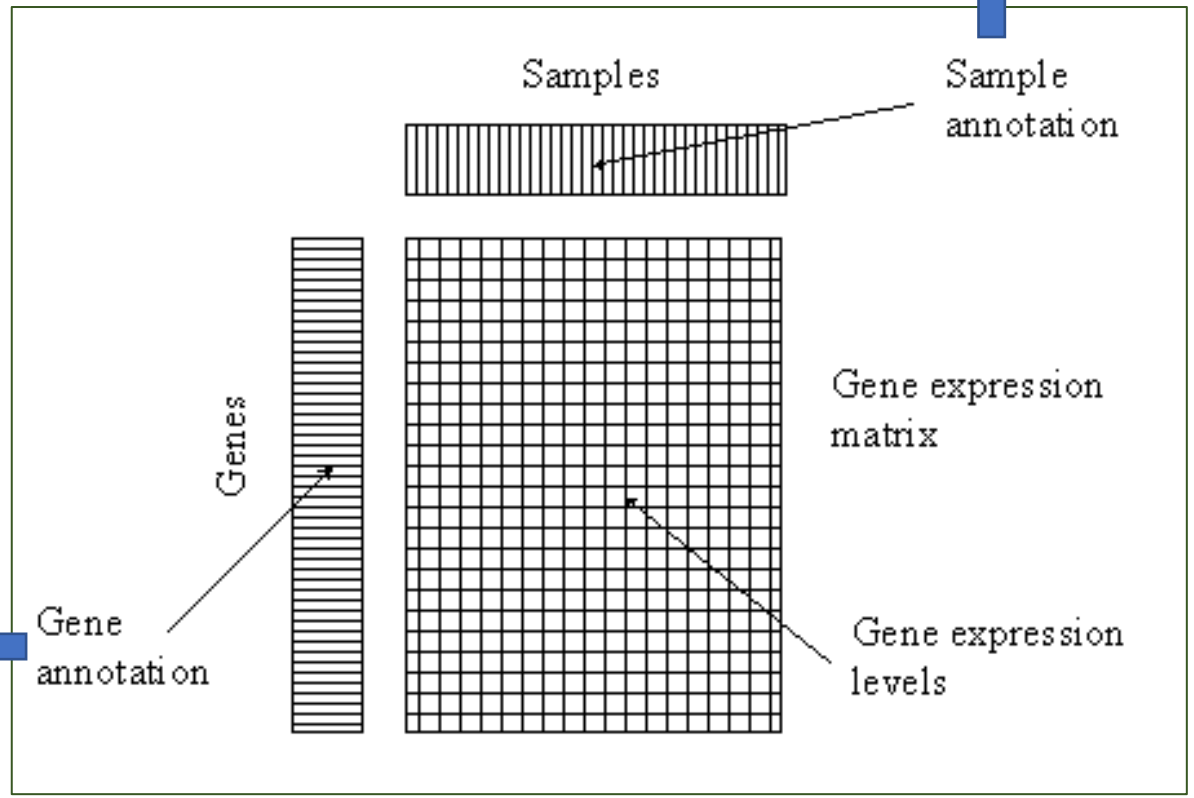
[3]Department of Radiation Oncology, School of Medicine, The University of Alabama at Birmingham, Birmingham, AL, United States

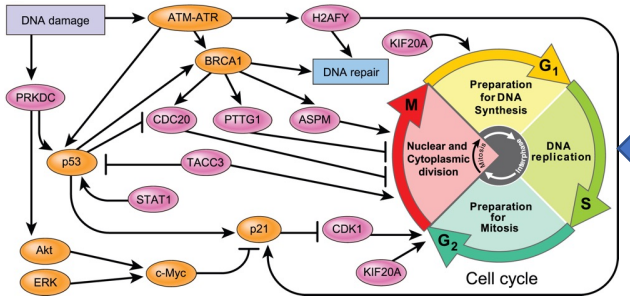
Unsupervised learning techniques, such as clustering and embedding, have been increasingly popular to cluster biomedical samples from high-dimensional biomedical data. Extracting

Background: *Exploratory Analysis of Samples*

Clinical features

Patient ID	Age	Survival day	Chemotherapy
0001-01	<=50	>300	FALSE
0003-01	>50<=65	<=300	FALSE
0004-01	>50<=65	>300	TRUE
0007-01	<=50	>300	TRUE
0009-01	>50<=65	>300	FALSE
0010-01	<=50	>300	FALSE



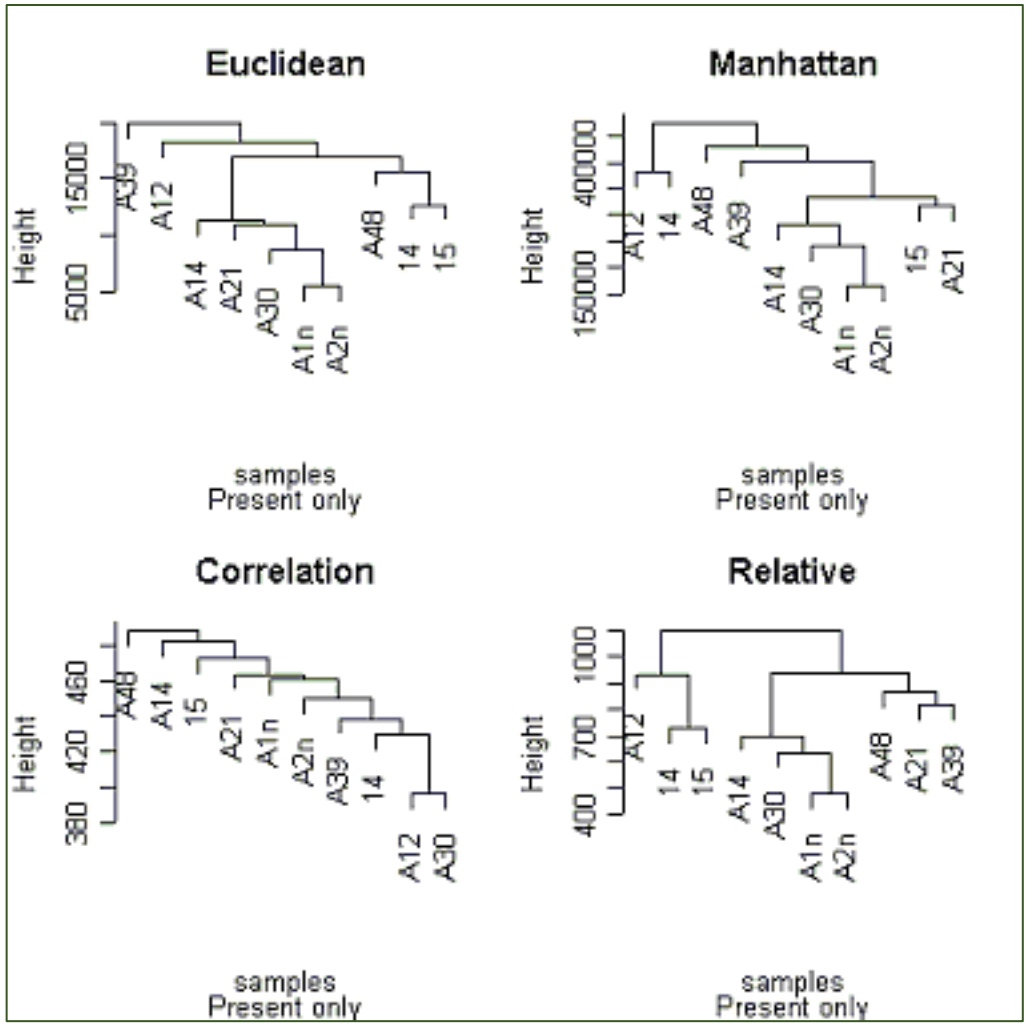


Gene Expression Profiling of multiple patient samples

Adapted from: Mark Reimers (2008) An Opinionated Guide to Microarray Data Analysis

Background:

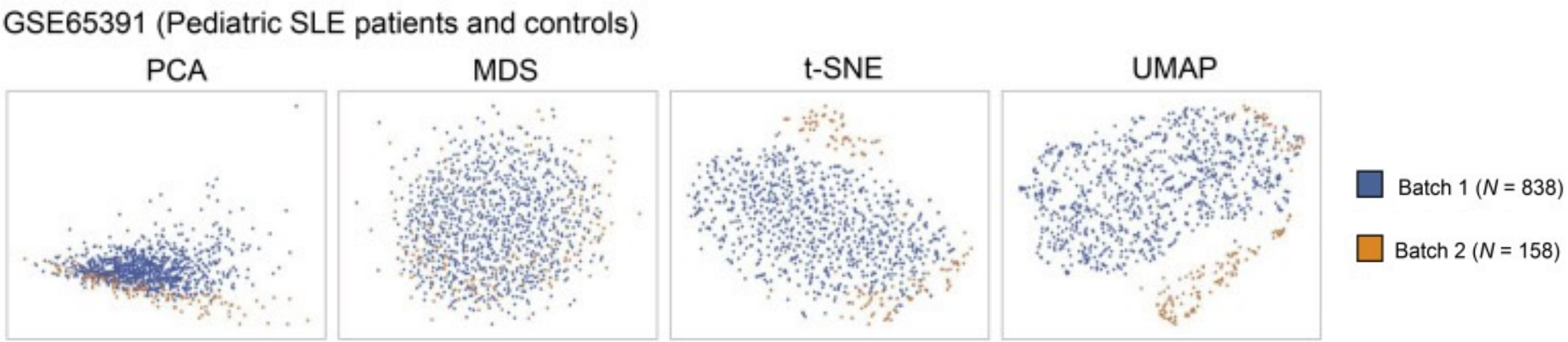
High-dimensional data Clustering can be unstable and algorithm-dependent



Adapted from: Mark Reimers (2008) An Opinionated Guide to Microarray Data Analysis

Dimensionality reduction techniques: Embeddings

To detect sample-to-sample heterogeneity in "bulk analysis" of patient samples



Linear embeddings do not separate samples well

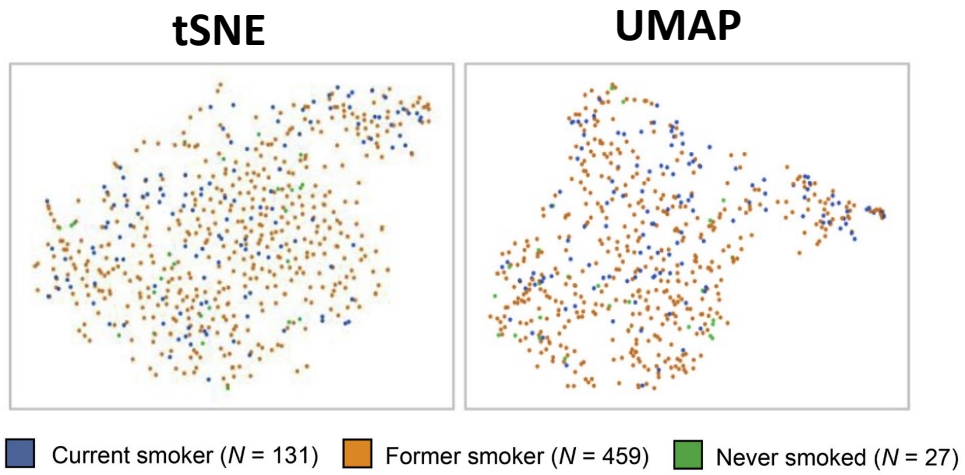
Non-linear embeddings can separate samples better

In Cell Reports, Vol. 36, No. 4, 27 July 2021, 109442

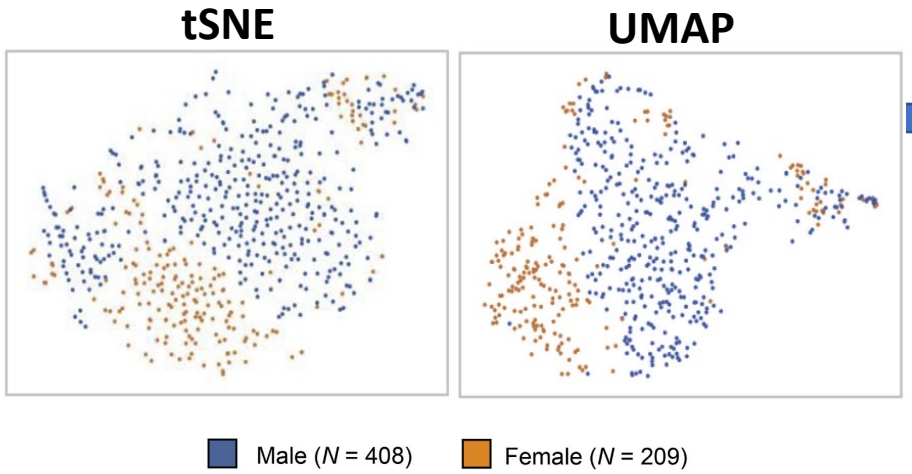
Exploration Case Study: by **smoker status** vs. by **gender**

Yielding different functional genomic profile insights

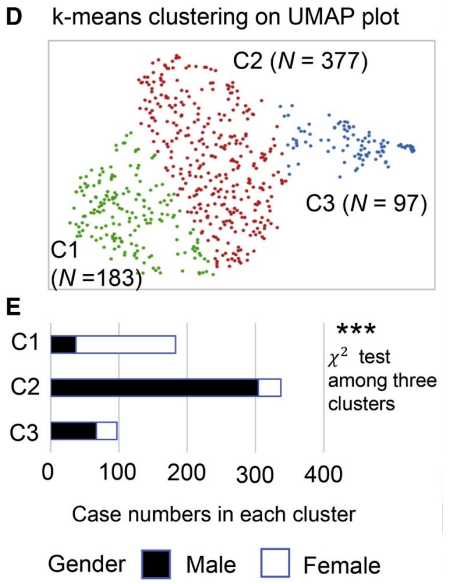
by **smoker status**

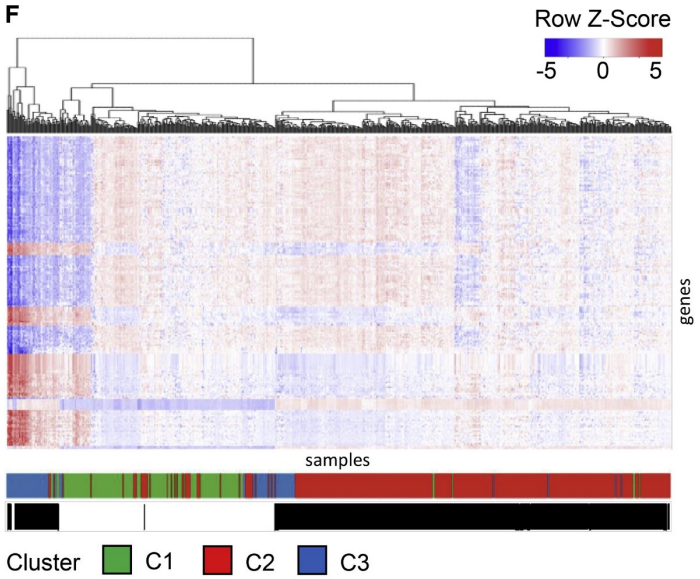


by **gender**



GSE71220 (COPD Patients with Different Smoking History)





In Cell Reports, Vol. 36, No. 4, 27 July 2021, 109442

Motivation for developing SEAS

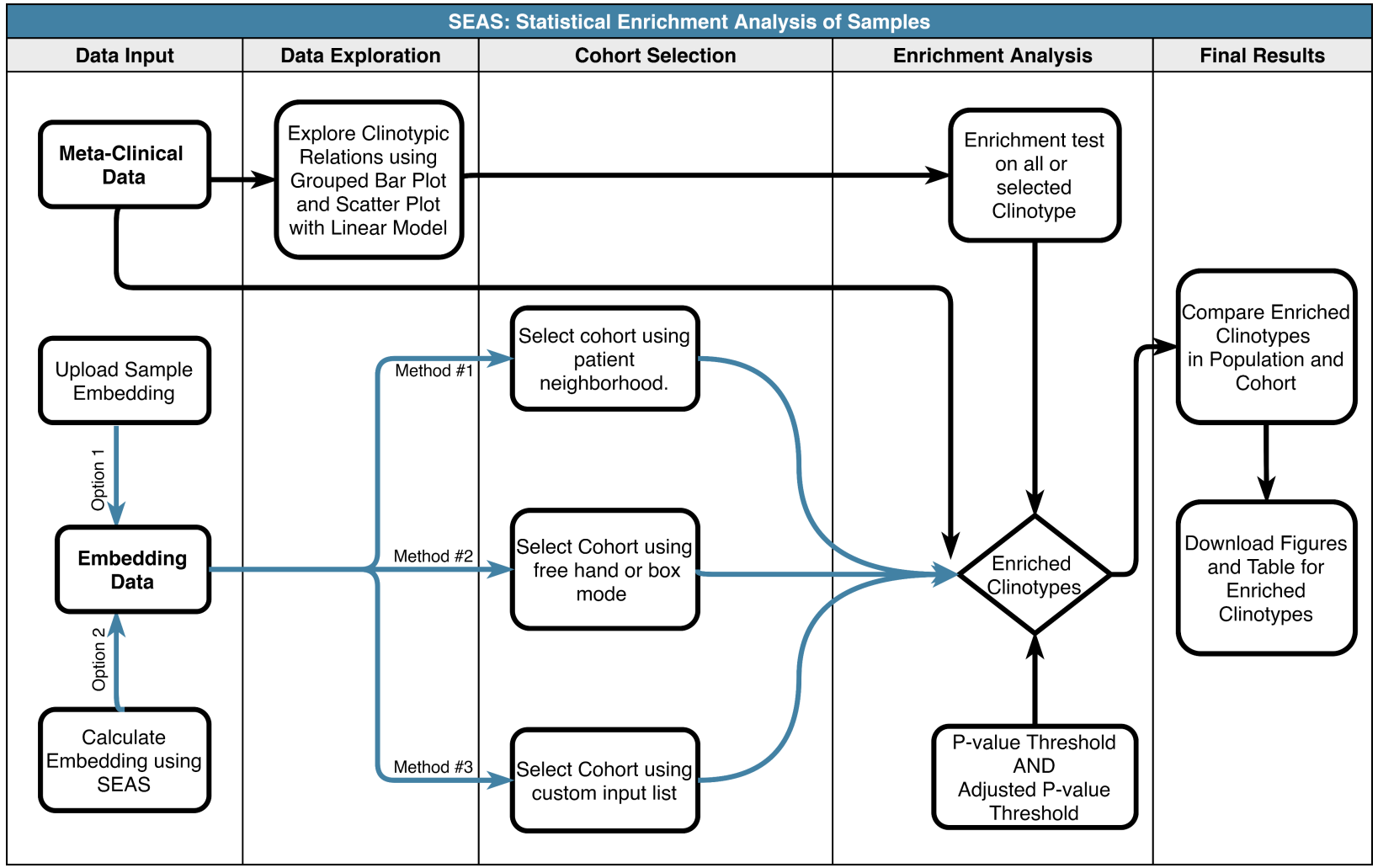
Statistical Enrichment Analysis of Samples (SEAS) is developed to perform exploratory analysis of embedded sample data from omics studies.

- The method derives its power by focusing on the "clinotypes" of selected sample sets in an embedding neighborhood, i.e., groups of biological samples that may share specific clinical characteristics different from that of the background.

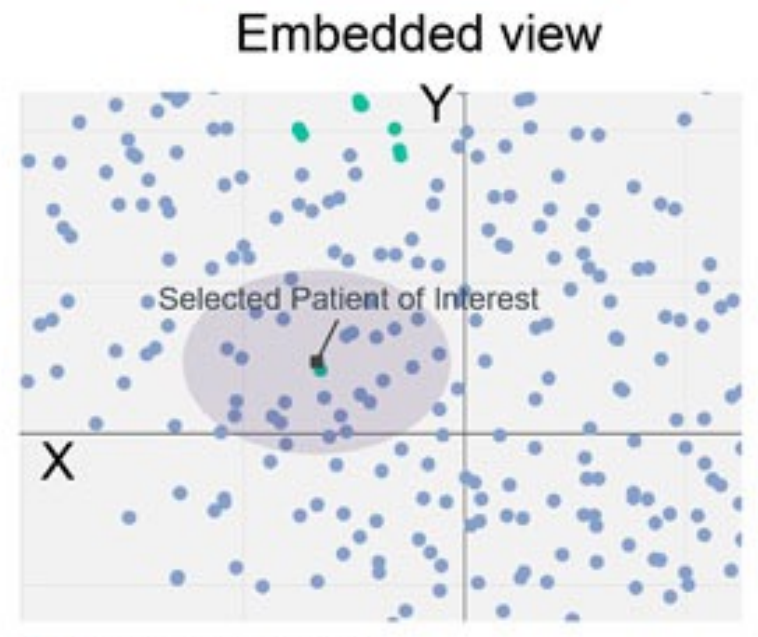
- Examples of "clinotypes" are: age group, gender, treatment status, or survival days.

aimed

SEAS Exploratory Data Analysis Workflow



How does SEAS work?



SEAS currently handles discrete clinotypes and their enrichment in the query sample set. (In future implementation, we will handle continuous clinotypes and their enrichment. The enrichment test is therefore a simple hypergeometric test overall all available clinotypes for each sample.

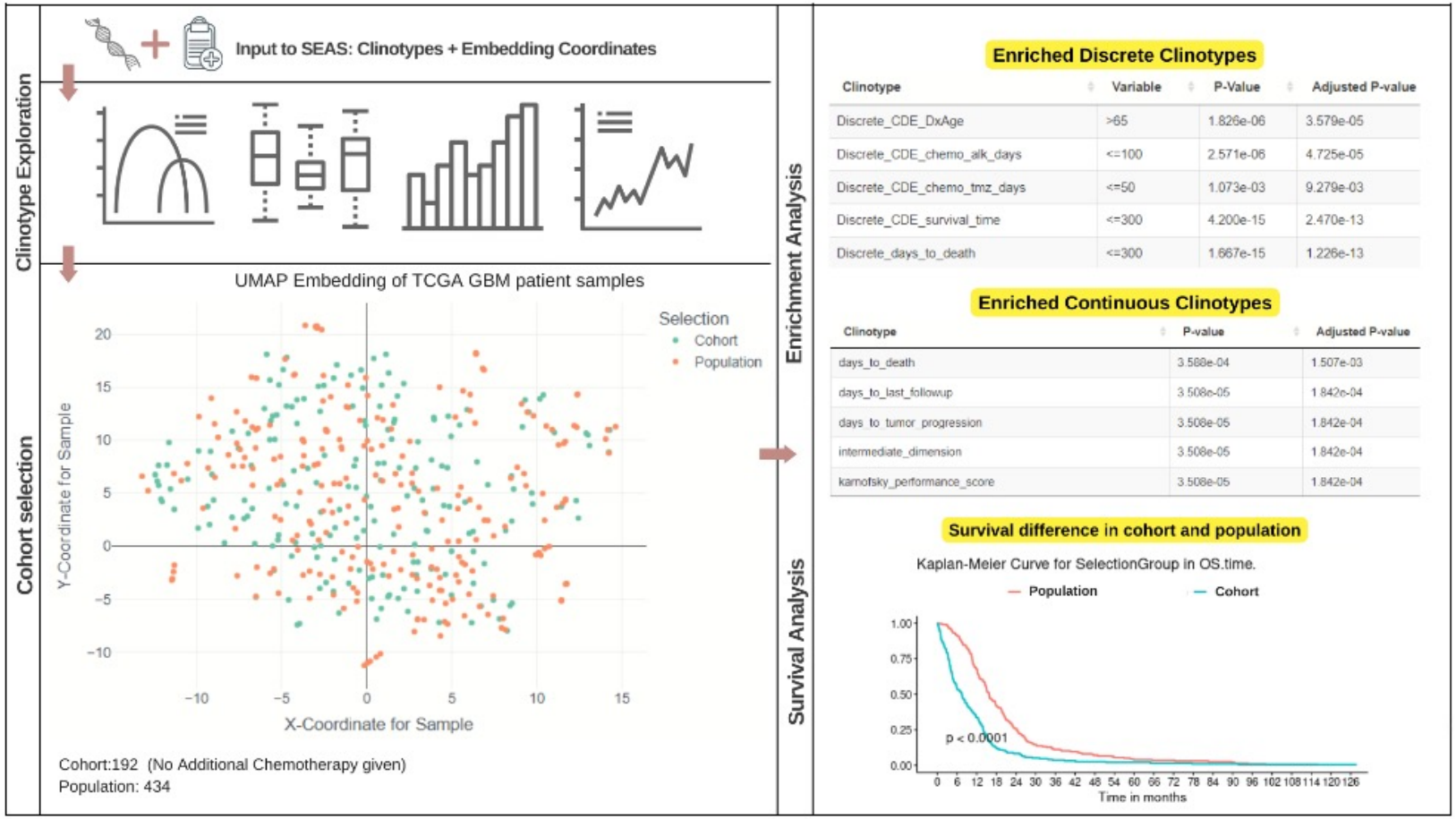
$$p_X(k) = \Pr(X = k) = \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}},$$

N: # of samples in the population
n: # of samples in the selected patient sample embedding neighborhood
K: # of samples with a specific clinotype in the population
k: # of samples in the selected patient sample embedding neighborhood

Clinotype	Variable	P-Value	Adjusted P-value	Enriched	N	n	K	k
Cluster	Cluster 1	2.78780037358447e-37	1.372992e-33	Yes	385	27	25	25
CDE_previously_treated	TRUE	3.01699789251814e-06	2.971743e-03	Yes	388	27	17	8
histological_type	Treated primary GBM	2.9582465599075e-06	2.971743e-03	Yes	389	27	17	8
history_of_neoadjuvant_treatment	Yes	9.47525921338965e-07	1.590266e-03	Yes	389	27	20	9
pretreatment_history	YES	9.68689666939645e-07	1.590266e-03	Yes	388	27	20	9

SEAS also allows sample set composition based on shared labels instead of shared embedding neighborhood

SEAS



Docs | Code | Web Server → https://aimed-lab.shinyapps.io/SEAS/

Stati al Enrichment Analysis of Samples (SEAS) is an online interactive software tool to help users characterize specific sample subsets (cohorts). SEAS enables users to explore all available clinotypes (clin attributes) within a sample subset or cohort, from which they can identify which clinotypes are enriched against all the samples in a data set. SEAS is useful not only for balancing case and control coho n traditional case-control studies, but also for profiling samples and sample subsets in cross-sectional studies.

Nguyen M., Bharti, S., Yue, Z., Willey, C. D., Chen, J. Y. (2021). Statistical Enrichment Analysis of Samples: A General-Purpose Tool to Annotate Metadata Neighborhoods of Biological Samples. Frontiers in big data, 4,

https://aimed-lab.github.io/SEAS/

aimed

Acknowledgment

The AI.*med* Lab

Ehsan Saghapour, PhD
Zongliang Yue, PhD
Thanh Nguyen, PhD
Zhandos Sembay, MS
Samuel Bharti

O'Neal CCC

Chris Willey, MD, PhD
T. Stackhouse, PhD

U54TR002731
U01CA223976
R01AI134023
R01AR073850