Code, Chat and Collab
13:00, Feb 17, 2023

A brief tutorial on exploring clinical associations in cancer samples using SEAS

Samuel Bharti
sbharti@uab.edu

SEAS: Statistical Enrichment Analysis of Samples

Analysis in Systems Biology using genomic and clinical features of samples

Genomic Insights such as: Enriched Pathways, Genesets, PAGS, etc.

Co-membership and regulatory gene AND/OR protein networks.

Metabolic models with Gene perturbation.

Genomic Analysis

Samples

Genes

GeneId	GSM1830157	GSM1830158	GSM1830159	GSM1830160	GSM1830161
7896740	4.43131576	4.275443909	4.648786441	4.338786228	4.490090291
7896742	10.66813534	9.922750217	10.80000918	10.53882706	10.64450314
7896744	6.534268579	5.654910258	6.444003383	4.877549558	6.32258658
7896746	9.052738077	9.980675586	9.9079715	10.92179457	9.406636666
7896754	8.067484144	8.093611973	8.206836993	7.961044976	8.039682658
7896817	6.579191435	9.512344595	6.55335871	8.861964325	6.611406201
7896929	6.814974641	7.039245327	7.008061788	7.153730165	6.48483708
7897006	7.149977624	7.315062558	7.554309153	7.568372433	7.216541476
7897172	9.392863432	9.111028395	8.86701238	9.169225582	9.269245563
7897236	9.961729512	9.952991249	9.528521413	9.785365865	10.02552042
7897288	7.252916672	7.758321144	7.682695403	7.684813881	6.93196358
7897370	9.126591428	9.327058196	9.435640016	9.25880796	9.26567564
7897404	9.346825155	9.419133453	9.613147348	9.232721072	9.543719501
7897426	5.880945152	5.537980553	5.605303871	5.586494195	5.816928829
7897482	9.714582765	9.303807831	9.495530109	9.418758607	9.473900066
7897527	9.119710227	8.737328147	8.942785829	8.826439493	8.961604638
7897561	8.103481955	8.060516612	8.476394035	8.228756776	8.028488764
7897620	11.58961496	11.51082781	11.67322239	11.34210418	11.5495066

Gene Expression Profiling of patient samples

Gene Expression Matrix

Correlation, Enrichment, etc. Analysis

Clinical Analysis

Clinical Features i.e., Clinotypes

sample_id	age	disease	fev1_fvc	fev1_predicted	sex	smoking_status	statin_user
GSM1830157	57	COPD	43.13	48.4	F	Former smoker	N
GSM1830158	72	COPD	48.21	54	M	Former smoker	Y
GSM1830159	70	COPD	59.93	61.8	F	Former smoker	N
GSM1830160	57	COPD	40.2	38.9	F	Former smoker	Y
GSM1830161	62	Control	76.93	109.2	M	Former smoker	Y
GSM1830162	67	COPD	43.07	75.1	F	Former smoker	N
GSM1830163	60	COPD	28.97	31.9	M	Former smoker	Y
GSM1830164	66	COPD	43.52	40.6	F	Former smoker	N
GSM1830165	74	COPD	66.02	62.8	F	Former smoker	N
GSM1830166	61	COPD	42.04	31.1	F	Former smoker	N
GSM1830167	70	COPD	36.72	32.2	M	Former smoker	N
GSM1830168	68	COPD	45.52	60.1	F	Former smoker	N
GSM1830169	71	COPD	57.02	66	F	Former smoker	N
GSM1830170	49	Control	81.57	93.3	F	Former smoker	Y
GSM1830171	70	COPD	34.85	53.9	F	Former smoker	N

UMAP, PCA, tSNE

Genotypically AND/OR Phenotypically resolved patient embedding.

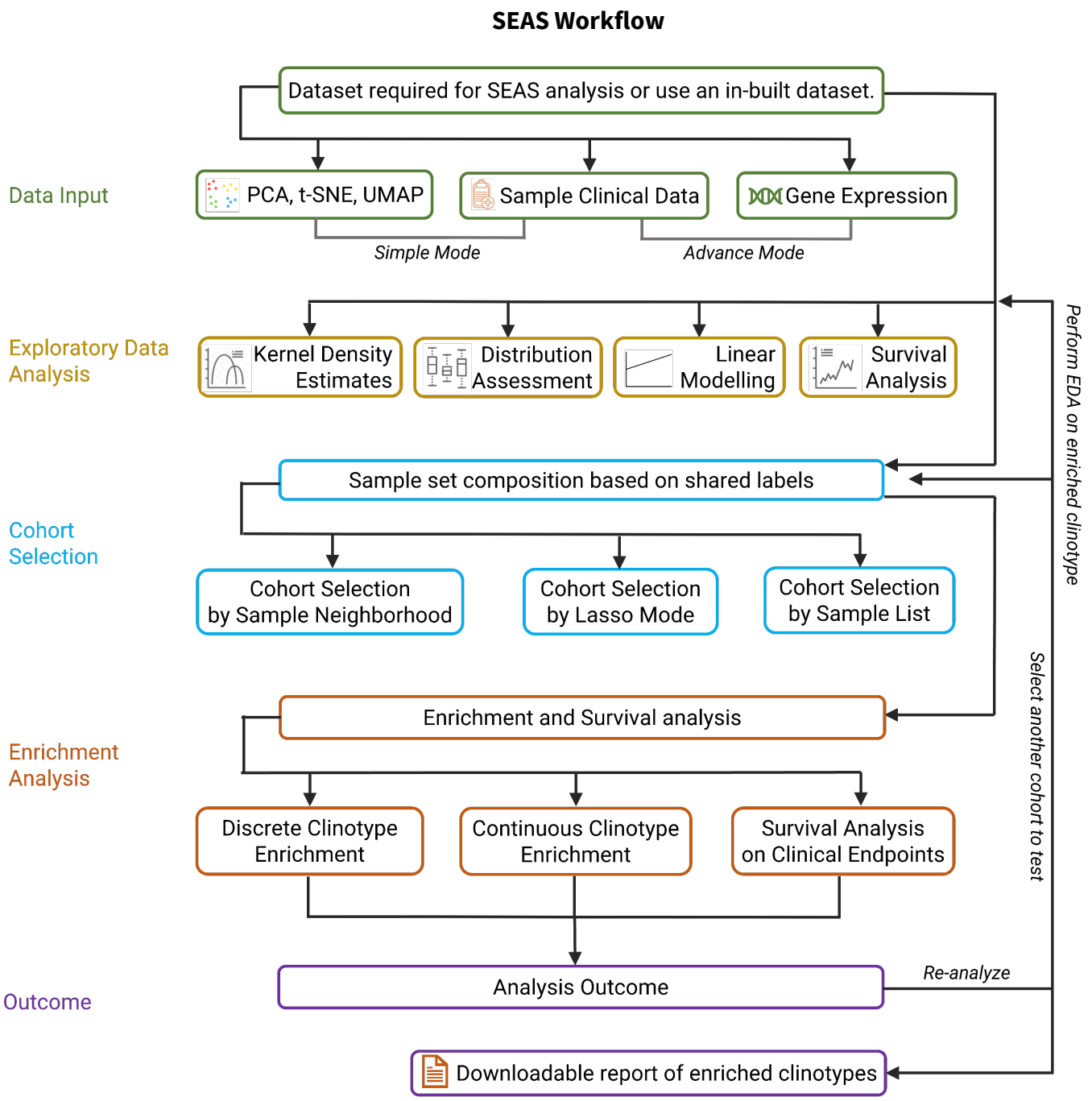
sample_id	X1	X2
GSM1830157	2.998425653	-1.391358321
GSM1830158	-1.401068876	1.81903131
GSM1830159	-0.396056858	-2.896121068
GSM1830160	1.55616542	-1.620665901
GSM1830161	0.160953406	2.76096781
GSM1830162	0.030356896	-2.519666312
GSM1830163	-2.740846083	-0.013063805
GSM1830164	0.939161057	-1.217158655
GSM1830165	-1.645087191	-0.856133376
GSM1830166	1.132666571	-0.6928069
GSM1830167	-1.612448305	-0.466482063
GSM1830168	1.175554088	-0.42713845
GSM1830169	2.117619142	0.180591641
GSM1830170	2.284750197	1.673612929
GSM1830171	2.215471982	1.133465544

A snapshot of Clinical Features associated to 389 TCGA GBM patients

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA	AB
1	sampleID	Dataset	Cluster	Discrete_CDE_DxAge	Discrete_CDE_chemo_	Discrete_C	Discrete_C	Discrete_c	Discrete_c	Discrete_c	Discrete_c	Discrete_kar	CDE_DxAge	CDE_alk_c	CDE_chemo_adjuvar	CDE_chem	CDE_chem	CDE_chemo_alk	CDE_chem	CDE_chem	CDE_chemo_t	CDE_chem	CDE_missi	CDE_previ	CDE_radia	CDE_radia	CDE_radia	CDE_ra
2	TCGA-02-C	TCGA	Cluster 1	<=50	<=100	<=50	>300	>300	<=250	>250	<=200	>80<=100	44.3	FALSE	FALSE	FALSE	FALSE	0	FALSE	FALSE	0	FALSE	FALSE	FALSE	TRUE	TRUE	TRUE	TRUE
3	TCGA-02-C	TCGA	Cluster 1	>50<=65	<=100	<=50	<=300	<=300	<=250	>250	<=200		50.21	FALSE	FALSE	FALSE	FALSE	0	FALSE	FALSE	0	FALSE	FALSE	FALSE	TRUE	TRUE	TRUE	TRUE
4	TCGA-02-C	TCGA	Cluster 2	>50<=65	>100	>50	>300	>300	>250	>250	>200	>80<=100	59.18	TRUE	TRUE	TRUE	TRUE	110	TRUE	TRUE	110	TRUE	FALSE	FALSE	TRUE	TRUE	TRUE	TRUE
5	TCGA-02-C	TCGA	Cluster 1	<=50	>100	>50	>300	>300	>250	>250	>200	>80<=100	40.53	TRUE	TRUE	TRUE	TRUE	306	TRUE	TRUE	306	TRUE	FALSE	TRUE	TRUE	TRUE	TRUE	TRUE
6	TCGA-02-C	TCGA	Cluster 1	>50<=65	<=100	<=50	>300	>300	>250	>250	>200	>80<=100	61.48	FALSE	FALSE	FALSE	FALSE	0	FALSE	FALSE	0	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	TRUE
7	TCGA-02-C	TCGA	Cluster 1	<=50	<=100	>50	>300	>300	>250	>250	>200	>80<=100	20.4	FALSE	TRUE	TRUE	TRUE	61	TRUE	TRUE	61	TRUE	FALSE	TRUE	TRUE	TRUE	TRUE	TRUE
8	TCGA-02-C	TCGA	Cluster 1	<=50	>100	>50	>300	>300	<=250	<=250	>200	>80<=100	18.96	TRUE	TRUE	TRUE	TRUE	125	TRUE	TRUE	125	TRUE	FALSE	FALSE	TRUE	TRUE	TRUE	TRUE
9	TCGA-02-C	TCGA	Cluster 2	<=50	<=100	<=50	>300	>300	>250	>250	>200		25.65	FALSE	FALSE	FALSE	TRUE	0	FALSE	TRUE	0	FALSE	FALSE	FALSE	TRUE	TRUE	TRUE	TRUE
10	TCGA-02-C	TCGA	Cluster 2	>50<=65	<=100	<=50	>300	>300	>250	>250	>200	>80<=100	50.39	FALSE	TRUE	FALSE	TRUE	0	FALSE	FALSE	0	FALSE	FALSE	FALSE	TRUE	TRUE	TRUE	TRUE
11	TCGA-02-C	TCGA	Cluster 1	<=50	>100	>50	>300	>300	>250	>250	>200	>80<=100	43.9	FALSE	TRUE	TRUE	TRUE	119	TRUE	TRUE	119	TRUE	FALSE	FALSE	TRUE	TRUE	TRUE	TRUE
12	TCGA-02-C	TCGA	Cluster 2	<=50	<=100	<=50	>300	>300	>250	>250	>200	>80<=100	38.34	TRUE	TRUE	TRUE	TRUE	44	TRUE	TRUE	44	TRUE	FALSE	FALSE	TRUE	TRUE	TRUE	TRUE
13	TCGA-02-C	TCGA	Cluster 1	<=50	>100	>50	>300	>300	>250	>250	>200		35.91	FALSE	TRUE	TRUE	TRUE	551	TRUE	TRUE	551	TRUE	FALSE	TRUE	TRUE	TRUE	TRUE	TRUE
14	TCGA-02-C	TCGA	Cluster 1	<=50	>100	>50	>300	>300	>250	>250	>200		47.64	FALSE	TRUE	TRUE	TRUE	539	TRUE	TRUE	539	TRUE	FALSE	FALSE	TRUE	TRUE	TRUE	TRUE
15	TCGA-02-C	TCGA	Cluster 2	<=50	>100	>50	>300	>300	>250	>250	>200	>80<=100	27.44	FALSE	TRUE	TRUE	TRUE	327	TRUE	TRUE	327	TRUE	FALSE	FALSE	TRUE	TRUE	TRUE	TRUE
16	TCGA-02-C	TCGA	Cluster 1	<=50	>100	>50	>300	>300	>250	>250	>200		33.86	TRUE	TRUE	TRUE	TRUE	230	TRUE	TRUE	230	TRUE	FALSE	FALSE	TRUE	TRUE	TRUE	TRUE
17	TCGA-02-C	TCGA	Cluster 1	<=50	>100	>50	>300	>300	>250	>250	>200		39.16	FALSE	TRUE	TRUE	TRUE	366	TRUE	TRUE	366	TRUE	FALSE	TRUE	TRUE	TRUE	TRUE	TRUE
18	TCGA-02-C	TCGA	Cluster 1	>50<=65	<=100	<=50	<=300	<=300	<=250	<=250	>200		54.95	FALSE	FALSE	FALSE	FALSE	0	FALSE	FALSE	0	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	TRUE
19	TCGA-02-C	TCGA	Cluster 2	>50<=65	<=100	>50	>300	>300	>250	>250	>200	>80<=100	60.69	FALSE	TRUE	TRUE	TRUE	71	TRUE	TRUE	71	TRUE	FALSE	FALSE	TRUE	TRUE	TRUE	TRUE
20	TCGA-02-C	TCGA	Cluster 1	<=50	<=100	<=50	>300	>300	<=250	<=250	>200	>80<=100	48.59	FALSE	FALSE	FALSE	FALSE	0	FALSE	FALSE	0	FALSE	FALSE	FALSE	TRUE	TRUE	TRUE	TRUE
21	TCGA-02-C	TCGA	Cluster 2	>50<=65	<=100	<=50	>300	>300	<=250	<=250	>200	>60<=80	54.93	FALSE	FALSE	FALSE	FALSE	0	FALSE	FALSE	0	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	TRUE
22	TCGA-02-C	TCGA	Cluster 1	>50<=65	<=100	<=50	>300	>300	>250	>250	>200	>80<=100	54.43	FALSE	TRUE	FALSE	TRUE	63	TRUE	FALSE	0	FALSE	FALSE	TRUE	TRUE	TRUE	TRUE	TRUE
23	TCGA-02-C	TCGA	Cluster 1	>50<=65	<=100	<=50	<=300	<=300	<=250	<=250	>200	>60<=80	61.37	FALSE	FALSE	FALSE	FALSE	0	FALSE	FALSE	0	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	TRUE
24	TCGA-02-C	TCGA	Cluster 1	>65	<=100	<=50	>300	>300	<=250	<=250	>200	>80<=100	78.74	FALSE	FALSE	FALSE	FALSE	0	FALSE	FALSE	0	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	TRUE
25	TCGA-02-C	TCGA	Cluster 2	>65	<=100	<=50	<=300	<=300	<=250	<=250	>200		80.22	FALSE	FALSE	FALSE	FALSE	0	FALSE	FALSE	0	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
26	TCGA-02-C	TCGA	Cluster 2	<=50	<=100	>50	>300	>300	<=250	<=250	>200		43.76	TRUE	TRUE	TRUE	TRUE	69	TRUE	TRUE	69	TRUE	FALSE	FALSE	TRUE	FALSE	FALSE	TRUE
27	TCGA-02-C	TCGA	Cluster 1	<=50	>100	>50	>300	>300	<=250	<=250	>200	>80<=100	49.45	TRUE	TRUE	TRUE	TRUE	171	TRUE	TRUE	171	TRUE	FALSE	FALSE	TRUE	TRUE	TRUE	TRUE
28	TCGA-02-C	TCGA	Cluster 1	<=50	<=100	<=50	<=300	<=300	<=250	<=250	>200	>80<=100	44.42	TRUE	TRUE	TRUE	TRUE	46	TRUE	TRUE	46	TRUE	FALSE	FALSE	TRUE	TRUE	TRUE	TRUE
29	TCGA-02-C	TCGA	Cluster 1	>65	>100	>50	>300	>300	>250	>250	>200	>80<=100	66.09	FALSE	TRUE	FALSE	TRUE	331	TRUE	TRUE	331	TRUE	FALSE	TRUE	TRUE	TRUE	TRUE	TRUE
30	TCGA-02-C	TCGA	Cluster 1	<=50	>100	>50	<=300	<=300	<=250	<=250	>200	>80<=100	28.79	TRUE	TRUE	TRUE	TRUE	123	TRUE	TRUE	123	TRUE	FALSE	TRUE	TRUE	TRUE	TRUE	TRUE
31	TCGA-02-C	TCGA	Cluster 2	>65	<=100	<=50	<=300	<=300	<=250	>250	<=200	>80<=100	68.71	TRUE	TRUE	TRUE	TRUE	43	TRUE	TRUE	43	TRUE	FALSE	FALSE	TRUE	FALSE	FALSE	TRUE
32	TCGA-02-C	TCGA	Cluster 1	>65	>100	>50	<=300	<=300	<=250	<=250	>200	>80<=100	66.12	TRUE	TRUE	TRUE	TRUE	168	TRUE	TRUE	168	TRUE	FALSE	FALSE	TRUE	TRUE	TRUE	TRUE
33	TCGA-02-C	TCGA	Cluster 2	>50<=65	>100	>50	>300	>300	>250	>250	>200		50.05	TRUE	TRUE	TRUE	TRUE	336	TRUE	TRUE	336	TRUE	FALSE	FALSE	TRUE	TRUE	TRUE	TRUE
34	TCGA-02-C	TCGA	Cluster 2	>50<=65	>100	>50	>300	>300	<=250	>250	<=200	>80<=100	57.93	TRUE	TRUE	TRUE	TRUE	166	TRUE	TRUE	166	TRUE	FALSE	FALSE	TRUE	TRUE	TRUE	TRUE
35	TCGA-02-C	TCGA	Cluster 2	>50<=65	<=100	<=50	<=300	<=300	<=250	<=250	>200	>80<=100	53.18	FALSE	FALSE	FALSE	FALSE	0	FALSE	FALSE	0	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
36	TCGA-02-C	TCGA	Cluster 2	>65	>100	>50	>300	>300	<=250	<=250	>200	>80<=100	68.19	TRUE	TRUE	TRUE	TRUE	233	TRUE	TRUE	233	TRUE	FALSE	FALSE	TRUE	TRUE	TRUE	TRUE
37	TCGA-02-C	TCGA	Cluster 2	>50<=65	<=100	<=50	>300	>300	>250	>250	>200	>80<=100	63.53	TRUE	TRUE	TRUE	TRUE	21	TRUE	TRUE	21	TRUE	FALSE	FALSE	TRUE	TRUE	TRUE	TRUE
38	TCGA-02-C	TCGA	Cluster 2	>50<=65	<=100	<=50	>300	>300	>250	>250	>200	>80<=100	57.97	TRUE	TRUE	TRUE	TRUE	42	TRUE	TRUE	42	TRUE	FALSE	FALSE	TRUE	FALSE	FALSE	TRUE
39	TCGA-02-C	TCGA	Cluster 2	<=50	<=100	>50	>300	>300	>250	>250	>200	>80<=100	28.22	FALSE	TRUE	TRUE	TRUE	92	TRUE	TRUE	92	TRUE	FALSE	FALSE	TRUE	TRUE	TRUE	TRUE
40	TCGA-02-C	TCGA	Cluster 2	>50<=65	>100	>50	>300	>300	>250	>250	>200	>80<=100	59.21	FALSE	TRUE	TRUE	TRUE	337	TRUE	TRUE	337	TRUE	FALSE	TRUE	TRUE	TRUE	TRUE	TRUE
41	TCGA-02-C	TCGA	Cluster 2	<=50	<=100	<=50	>300	>300	>250	>250	>200	>80<=100	36.31	FALSE	FALSE	FALSE	FALSE	0	FALSE	FALSE	0	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
42	TCGA-02-C	TCGA	Cluster 2	>50<=65	<=100	<=50	>300	>300	>250	>250	>200	>80<=100	63.76	TRUE	TRUE	TRUE	TRUE	46	TRUE	TRUE	46	TRUE	FALSE	FALSE	TRUE	TRUE	TRUE	TRUE
43	TCGA-02-C	TCGA	Cluster 2	<=50	<=100	>50	<=300	<=300	<=250	<=250	>200		45.89	TRUE	TRUE	TRUE	TRUE	73	TRUE	TRUE	73	TRUE	FALSE	FALSE	TRUE	TRUE	TRUE	TRUE
44	TCGA-02-C	TCGA	Cluster 2	>50<=65	>100	>50	>300	>300	>250	>250	>200		52.66	TRUE	TRUE	TRUE	TRUE	216	TRUE	TRUE	216	TRUE	FALSE	TRUE	TRUE	TRUE	TRUE	TRUE
45	TCGA-02-C	TCGA	Cluster 2	<=50	>100	>50	<=300	<=300	>250	>250	>200	>80<=100	46.76	FALSE	TRUE	TRUE	TRUE	137	TRUE	TRUE	137	TRUE	FALSE	FALSE	TRUE	TRUE	TRUE	TRUE
46	TCGA-02-C	TCGA	Cluster 2	<=50	>100	<=50	>300	>300	>250	>250	>200		42.87	FALSE	TRUE	FALSE	TRUE	246	TRUE	FALSE	0	FALSE	FALSE	TRUE	TRUE	TRUE	TRUE	TRUE
47	TCGA-02-C	TCGA	Cluster 2	<=50	>100	>50	>300	>300	>250	>250	>200	>80<=100	29.29	TRUE	TRUE	TRUE	TRUE	251	TRUE	TRUE	251	TRUE	FALSE	FALSE	TRUE	TRUE	TRUE	TRUE

Statistical Enrichment Analysis of Sample Clinical Attributes Using SEAS

- Embedding techniques has gained popularity in visualizing the high-dimensional gene expression profiles of patient samples yet the systematic extraction of sample set composition based on shared labels instead of shared embedding neighbourhood remains a major challenge.
- SEAS can be used to perform exploratory analysis of embedded sample data by focusing on the “clinotypes” of selected sample sets.
- Clinotypes: Clinotypes are referred as the clinical/phenotypical features of a sample. For SEAS analysis clinotypes are classified in two i.e., discrete and continuous clinotypes.
- Discrete Clinotypes are the clinotypes which take specific value in quantitative or qualitative data. For examples, age groups, cancer subtypes, treatment method, etc.
- Continuous Clinotypes are the clinotypes which take continuous quantitative values. For example, age, survival days, treatment days, dose levels, etc.
- CFEA: Clinical Feature Enrichment Analysis is a method defined in SEAS to identify clinotypes which are over-represent in a selected cohort from population.
- We used Hypergeometric Test, KS-test, and Kaplan-Meier Method to perform discrete clinotype enrichment, continuous clinotype enrichment and survival analysis, respectively.



https://github.com/informaticsclub/ccc_presentations

Functional Enrichment Analysis of GBM patients uncovers clinical/phenotypic difference in additional chemotherapy lacking cohort.

GBM: glioblastoma multiforme

We acquired and preprocessed TCGA-GBM dataset, which consists of 389 patients, according to the pipeline in Jia et al. (2018). The dataset had both the genetic and the clinical sections. We also used 45 GBM tumor-samples hosted in patient-derived xenograft (PDX) models (Willey et al., 2020). We performed SEAS analysis to test enrichment in the patient samples where no additional chemotherapy was given

5 of 50 Enriched Discrete Clinotypes

Show 5 rows | Copy | Excel | Search:

Clinotype	Variable	P-Value	Adjusted P-value	Enriched	N	n	K	k
Discrete_CDE_DxAge	>65	1.826e-06	3.579e-05	Yes	389	192	128	85
Discrete_CDE_chemo_alk_days	<=100	2.571e-06	4.725e-05	Yes	389	192	245	143
Discrete_CDE_chemo_tmz_days	<=50	1.073e-03	9.279e-03	Yes	389	192	245	136
Discrete_CDE_survival_time	<=300	4.200e-15	2.470e-13	Yes	389	192	157	115
Discrete_days_to_death	<=300	1.667e-15	1.226e-13	Yes	389	192	156	115

Showing 1 to 5 of 50 entries | Previous 1 2 3 4 5 ... 10 Next

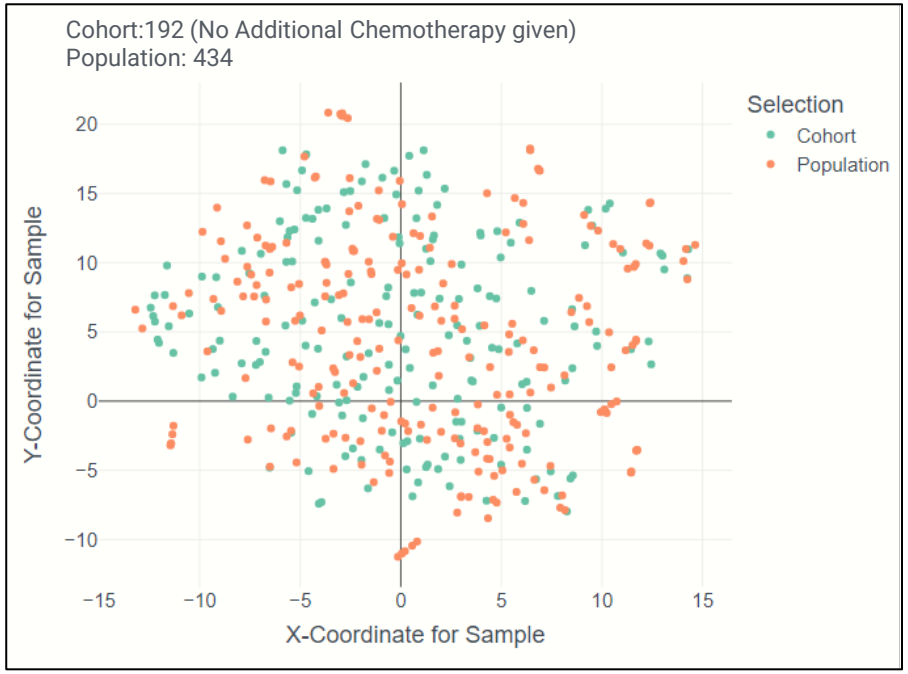
5 Enriched Continuous Clinotypes

Show 5 rows | Copy | Excel | Search:

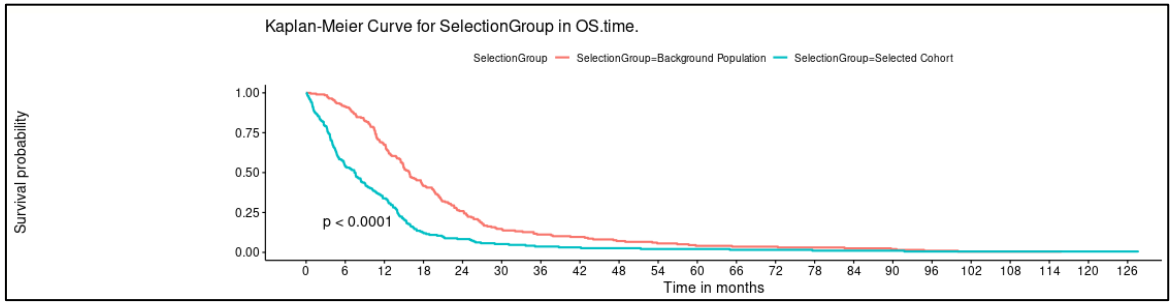
Clinotype	P-value	Adjusted P-value	Enriched
days_to_death	3.588e-04	1.507e-03	Yes
days_to_last_followup	3.508e-05	1.842e-04	Yes
days_to_tumor_progression	3.508e-05	1.842e-04	Yes
intermediate_dimension	3.508e-05	1.842e-04	Yes
karnofsky_performance_score	3.508e-05	1.842e-04	Yes

Showing 1 to 5 of 5 entries | Previous 1 Next

UMAP Embedding of 434 GBM cancer samples



Survival Difference Between Cohort and Population.



- We found 50 discrete and 5 continuous enriched clinotypes in the selected cohort. Performing survival analysis we found a significant p-value of 0.0001 suggesting there's a significant effect of additional chemotherapy on survival of GBM cancer patients.
- **GBM cancer patients who didn't receive additional chemotherapy died earlier.**

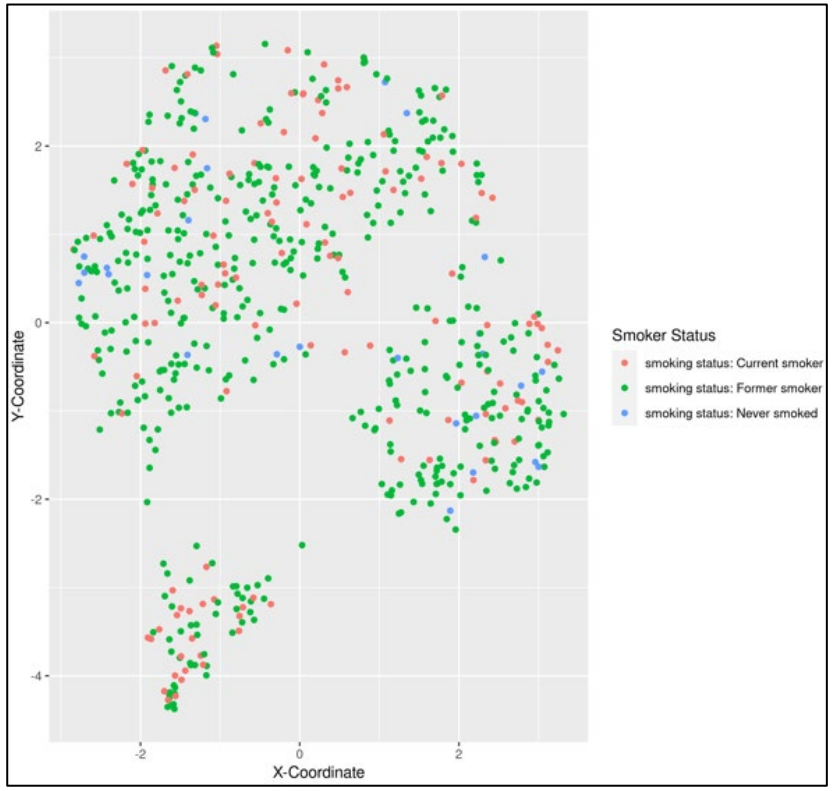
Exploratory Analysis of COPD Patient Profiles using SEAS reveal understandings in clinotype-explained genomic variation.

COPD: Chronic obstructive pulmonary disease

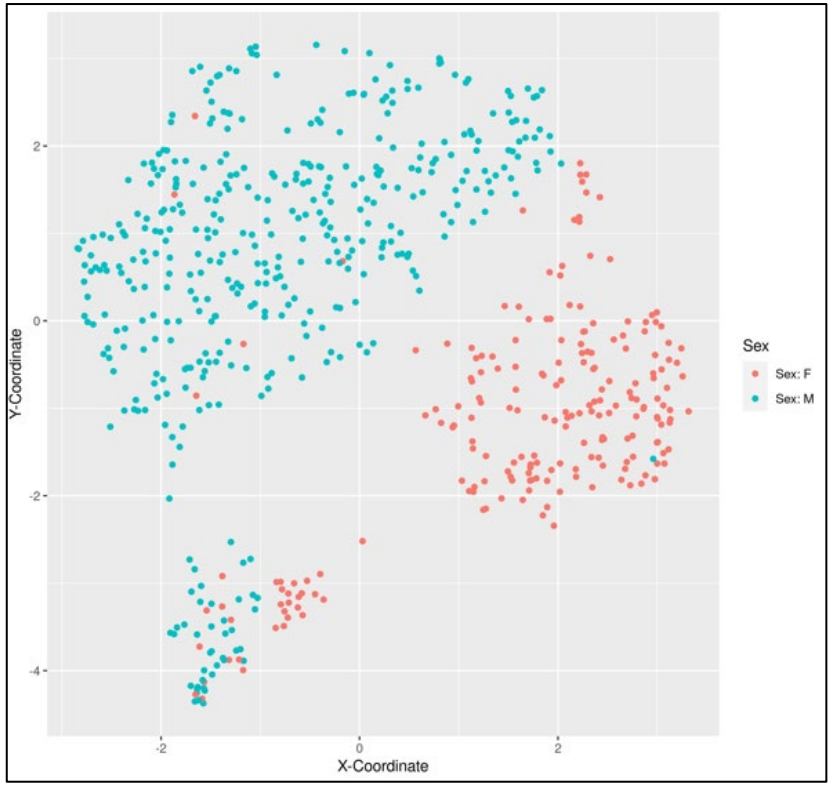
We acquired and preprocessed clinical and genetic data of 617 COPD patient samples publicly accessible at GEO: GSE71220. We perform a basic exploratory analysis of data using SEAS to demonstrate how researchers can identify clinotype showing high association with sample embedding.

UMAP embedding of 617 COPD patient samples

Clinotype: By Smoking Status



Clinotype: By Gender



- We showed a simple exploratory analysis using SEAS on COPD patient samples revealing sample embedding highly associated by gender difference than originally studied smoking status clinotype.
- **COPD patient transcriptomes were highly influenced by gender difference than smoking status.**