frontiers in Big Data

TECHNOLOGY AND CODE
published: 16 September 2021
doi: 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*

OPEN ACCESS

Edited by:
*Huixiao Hong,
United States Food and Drug Administration, United States*

Reviewed by:
*Ehsan Ullah,
Qatar Computing Research Institute, Qatar
Shailesh Tripathi,
Tampere University of Technology, Finland*

***Correspondence:**
*Jake Y. Chen
jakechen@uab.edu*

†ORCID:
*Christopher D. Willey
orcid.org/0000-0001-9953-0279*

Specialty section:
This article was submitted to Medicine and Public Health, a section of the journal Frontiers in Big Data

Received: *15 June 2021*
Accepted: *06 September 2021*
Published: *16 September 2021*

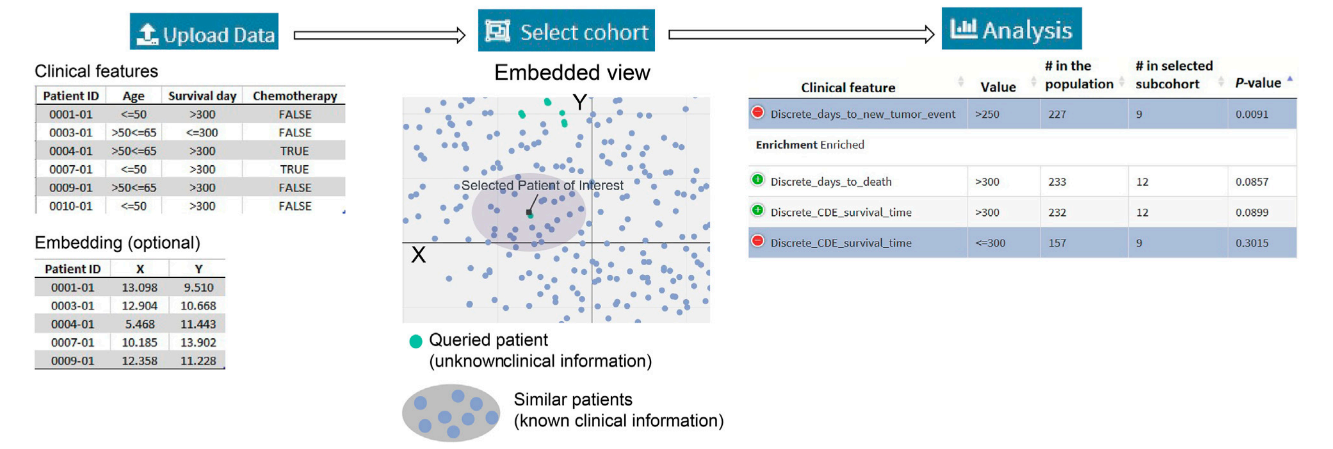
Citation:
Nguyen TM, Bharti S, Yue Z, Willey CD and Chen JY (2021) Statistical Enrichment Analysis of Samples: A General-Purpose Tool to Annotate Metadata Neighborhoods of Biological Samples. Front. Big Data 4:725276. doi: 10.3389/fdata.2021.725276

Unsupervised learning techniques, such as clustering and embedding, have been increasingly popular to cluster biomedical samples from high-dimensional biomedical data. Extracting clinical data or sample meta-data shared in common among biomedical samples of a given biological condition remains a major challenge. Here, we describe a powerful analytical method called Statistical Enrichment Analysis of Samples (SEAS) for interpreting clustered or embedded sample data from omics studies. The method derives its power by focusing on sample sets, i.e., groups of biological samples that were constructed for various purposes, e.g., manual curation of samples sharing specific characteristics or automated clusters generated by embedding sample omic profiles from multi-dimensional omics space. The samples in the sample set share common clinical measurements, which we refer to as "clinotypes," such as age group, gender, treatment status, or survival days. We demonstrate how SEAS yields insights into biological data sets using glioblastoma (GBM) samples. Notably, when analyzing the combined The Cancer Genome Atlas (TCGA)—patient-derived xenograft (PDX) data, SEAS allows approximating the different clinical outcomes of radiotherapy-treated PDX samples, which has not been solved by other tools. The result shows that SEAS may support the clinical decision. The SEAS tool is publicly available as a freely available software package at https://aimed-lab.shinyapps.io/SEAS/.

Keywords: sample enrichment analysis, clinotype, SEAS, glioblastoma multiforme, patient-derived xenograft, patient-derived xenograft

INTRODUCTION

Systematic software platforms to organize large metadata and clinical data [also called "clinotype" (Nguyen et al., 2021)] is essential in biomedical research (Burgun and Bodenreider, 2008; Ohmann and Kuchinke, 2009). These software platforms, such as (Ta et al., 2018; Kim et al., 2019; Hume et al., 2020), have two key objectives. First, it allows the biomedical researcher to perform manual cohort selection quickly. Here, the researcher inputs the filtering query and gets the data from all patients meeting the filtering criteria. Second, it allows quick data exploration, including data visualization and simple aggregated analysis. Here, the researcher may view the basic characteristic of the selected


Upload Data
Select cohort
Analysis
Embedded view
Y
X
Selected Patient of Interest
Queried patient
(unknownclinical information)
Similar patients
(known clinical information)

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

Embedding (optional)

Patient ID	X	Y
0001-01	13.098	9.510
0003-01	12.904	10.668
0004-01	5.468	11.443
0007-01	10.185	13.902
0009-01	12.358	11.228

Clinical feature	Value	# in the population	# in selected subcohort	*P*-value
⊖ Discrete_days_to_new_tumor_event	>250	227	9	0.0091
Enrichment Enriched				
⊕ Discrete_days_to_death	>300	233	12	0.0857
⊕ Discrete_CDE_survival_time	>300	232	12	0.0899
⊖ Discrete_CDE_survival_time	<=300	157	9	0.3015

FIGURE 1 | Overview of data processing and analysis.

subcohort, find potential clinical bias, and adjust the filtering criteria to obtain a better subcohort. Integrating Biology and the Bedside (Murphy et al., 2010) is a typical example of a clinical metadata software system. Some systems and techniques may offer more in-depth and specific analysis. For example, Weng et al. (2017) implemented a machine-learning based system to estimate the patients' cardiovascular risk from the routine checkup records. Fang et al. (Fang et al., 2014) implemented a visual analytic system to view patient's geographical demographic and disease comorbidities.

On the other hand, the state-of-the-art clinical data software still has three limitations. First, the simple aggregated analysis has not been well-developed for categorical clinical attributes. Therefore, the researcher may not easily find whether a specific categorical attribute is explicit for the selected cohort compared to the whole population. Second, methods to quantify and visualize patients' similarities have not been implemented. Therefore, the existing clinical software is likely ineffective in clinical support scenarios such as "finding the clinical outcome data about previous patients that are the most similar to the under-treatment patients". Third, the existing software does not support patient clustering. Therefore, they may not automatically recommend subcohort to the researcher. This feature could provide new insights to biomedical research; for example, a tool that quickly shows two clusters in a treatment-selected cohort may enable a new hypothesis about the treatment outcome.

This work introduces Statistical Enrichment Analysis of Samples (https://aimed-lab.shinyapps.io/SEAS/), a software tool with both online and standalone versions to tackle the above limitations. SEAS graphical user interface is user-friendly, where the user interacts by uploading datafile, primarily uses mouse operations, and requires a very limited amount of typing. Furthermore, SEAS implements methods to analyze numerical and categorical data, compute patient similarity, and automatically cluster the patients. For the demo, we use SEAS to analyzing the glioblastoma multiforme (GBM) patients' clinical metadata in The Cancer Genome Atlas Program (TCGA) (Verhaak et al., 2010) and estimate the clinical outcome of patient-derived xenograft (PDX) models data.

SEAS FUNCTIONS

Figure 1A summarizes a SEAS session. The required input is the clinical metadata that is organized in one table. The user may choose to let SEAS automatically compute and represent the patients' similarity in a 2D embedding space or optionally upload another patients' scatterplot. Here, each plot represents a patient, and the distance among the plots should represent patients' similarities. Then, the user may manually enter a subcohort, automatically let SEAS select a subcohort, or semi-automatically choose a subcohort. After selecting a subcohort, SEAS performs clinical feature enrichment analysis (CFEA) and reports all enriched features in the selected subcohort.

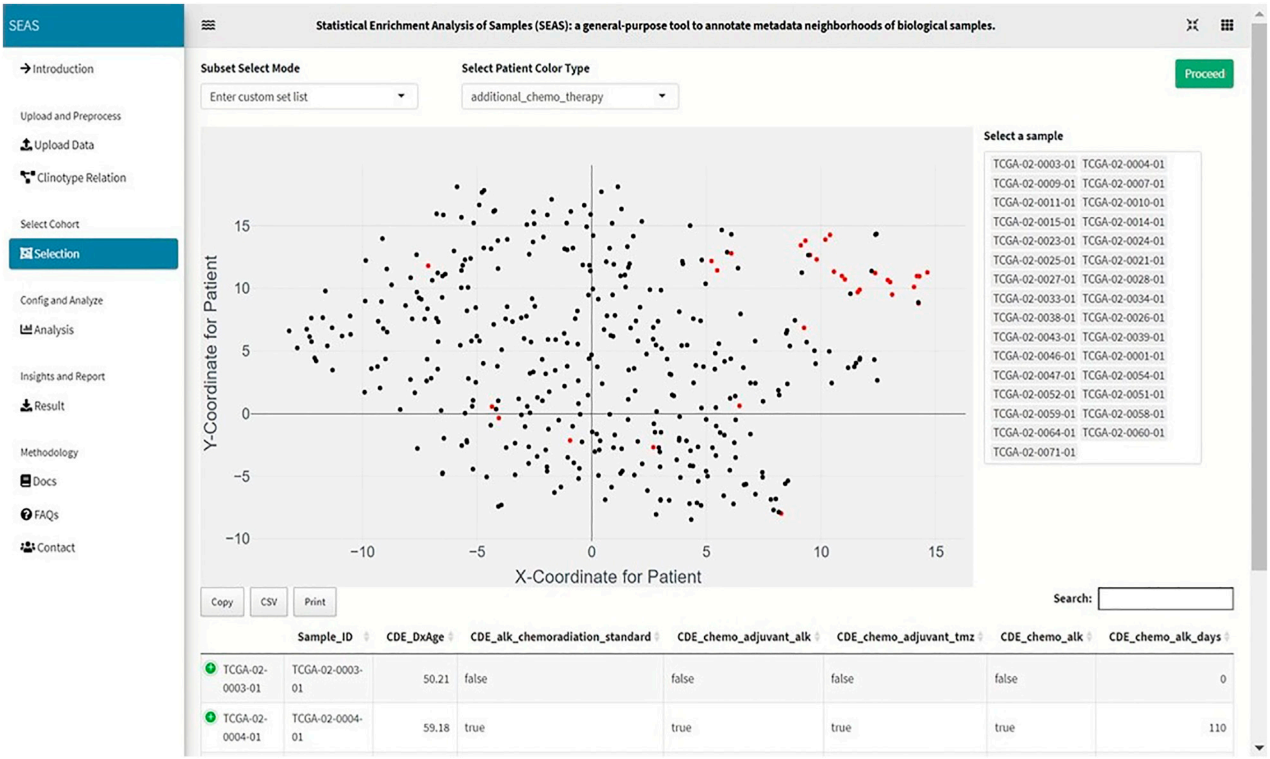
Automatically Compute Patients' Similarity and Embedding

In this step, the categorical clinical attributes are digitized as in (Zaki et al., 2014). For example, if the categorical attribute X has three discrete values: low, normal, and high, it can be decomposed into three binary attributes: is_X_low, is_X_normal, is_X_high. If a patient has a "high" categorical value for X, then the patient's digital representation is (0, 0, 1). On the other hand, the numerical attributes are normalized using the z-score approach.

After digitizing the clinical attributes, SEAS applies the embedding method (**Figures 2–7**) to represent the patients in a 2D space. By default, SEAS uses the umap (McInnes et al., 2018) algorithm. Alternatively, the user may also select tSNE (Hinton and Roweis, 2002) for embedding. SEAS computes patients' similarities using the 2D embedded coordinate.

Automatically Select a Subcohort

In SEAS, the user can manually define a subcohort by typing the list of patient IDs (**Figure 2**). Besides, the user may use SEAS to



Statistical Enrichment Analysis of Samples (SEAS): a general-purpose tool to annotate metadata neighborhoods of biological samples.

Subset Select Mode: Enter custom set list

Select Patient Color Type: additional_chemo_therapy

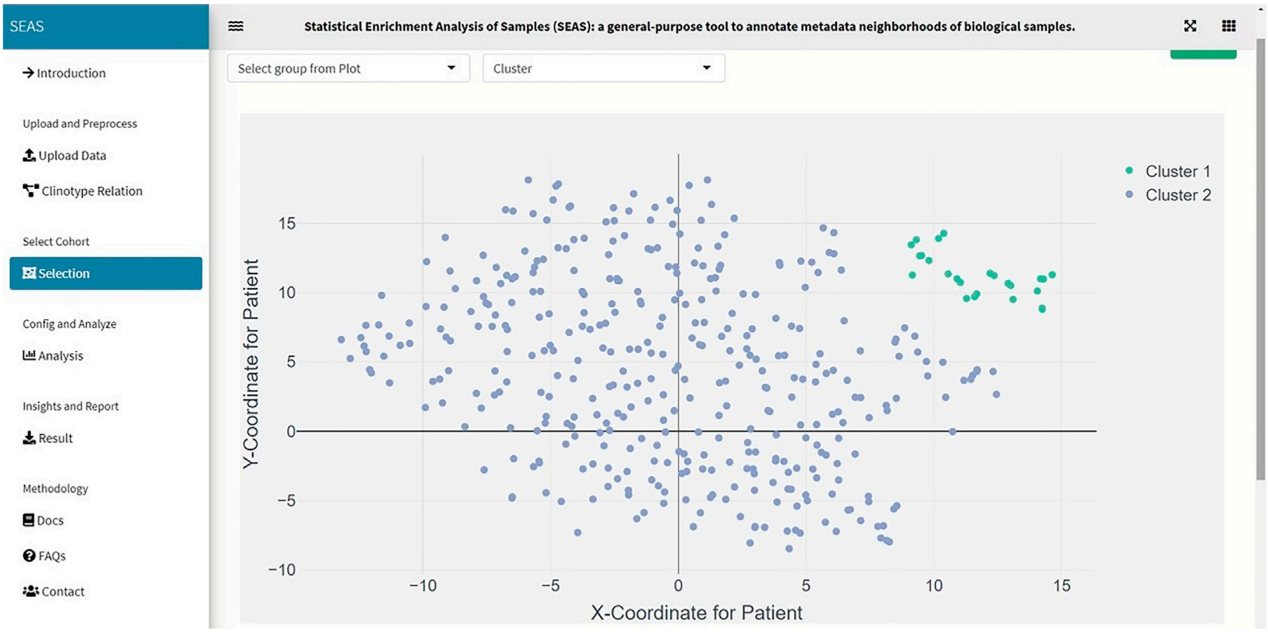
Select a sample

TCGA-02-0003-01	TCGA-02-0004-01
TCGA-02-0009-01	TCGA-02-0007-01
TCGA-02-0011-01	TCGA-02-0010-01
TCGA-02-0015-01	TCGA-02-0014-01
TCGA-02-0023-01	TCGA-02-0024-01
TCGA-02-0025-01	TCGA-02-0021-01
TCGA-02-0027-01	TCGA-02-0028-01
TCGA-02-0033-01	TCGA-02-0034-01
TCGA-02-0038-01	TCGA-02-0026-01
TCGA-02-0043-01	TCGA-02-0039-01
TCGA-02-0046-01	TCGA-02-0001-01
TCGA-02-0047-01	TCGA-02-0054-01
TCGA-02-0052-01	TCGA-02-0051-01
TCGA-02-0059-01	TCGA-02-0058-01
TCGA-02-0064-01	TCGA-02-0060-01
TCGA-02-0071-01	

Search:

	Sample_ID	CDE_DxAge	CDE_alk_chemoradiation_standard	CDE_chemo_adjuvant_alk	CDE_chemo_adjuvant_tmz	CDE_chemo_alk	CDE_chemo_alk_days
TCGA-02-0003-01	TCGA-02-0003-01	50.21	false	false	false	false	0
TCGA-02-0004-01	TCGA-02-0004-01	59.18	true	true	true	true	110

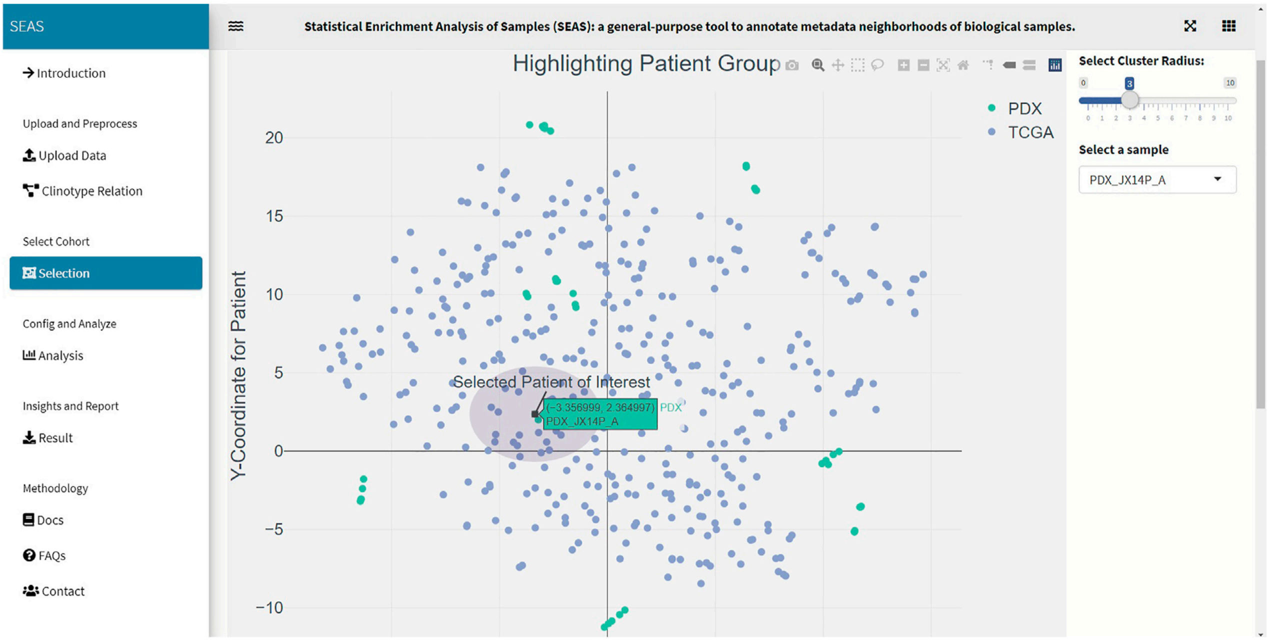
FIGURE 2 | Screenshot showing that SEAS visualizes the TCGA-GBM patients using embedding, and the user manually selects the subcohort.



Statistical Enrichment Analysis of Samples (SEAS): a general-purpose tool to annotate metadata neighborhoods of biological samples.

Select group from Plot: Cluster

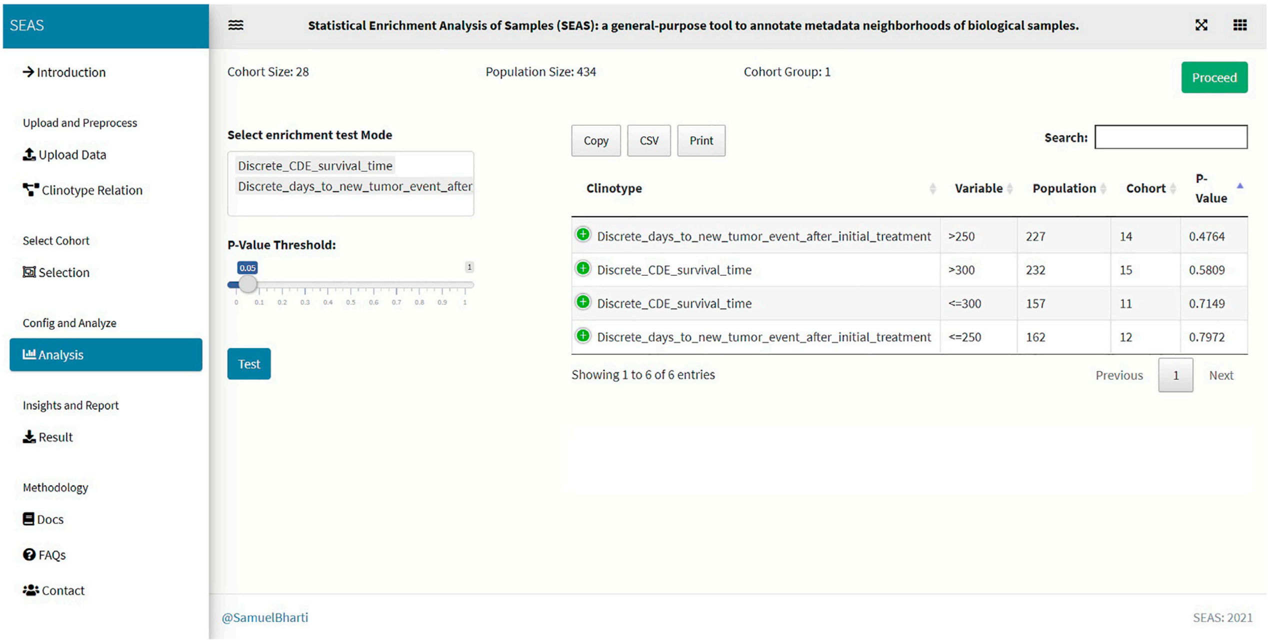
FIGURE 3 | SEAS identifies a subcohort by clustering the TCGA-GBM patients (green dots on the top-right of the embedding scatterplot).


Statistical Enrichment Analysis of Samples (SEAS): a general-purpose tool to annotate metadata neighborhoods of biological samples.

Select Cluster Radius: 3

Select a sample: PDX_JX14P_A

FIGURE 4 | SEAS identifies a subcohort by a circle region around PDX JX14P_A datapoint.


Statistical Enrichment Analysis of Samples (SEAS): a general-purpose tool to annotate metadata neighborhoods of biological samples.

Cohort Size: 28 Population Size: 434 Cohort Group: 1

Select enrichment test Mode: Discrete_CDE_survival_time, Discrete_days_to_new_tumor_event_after

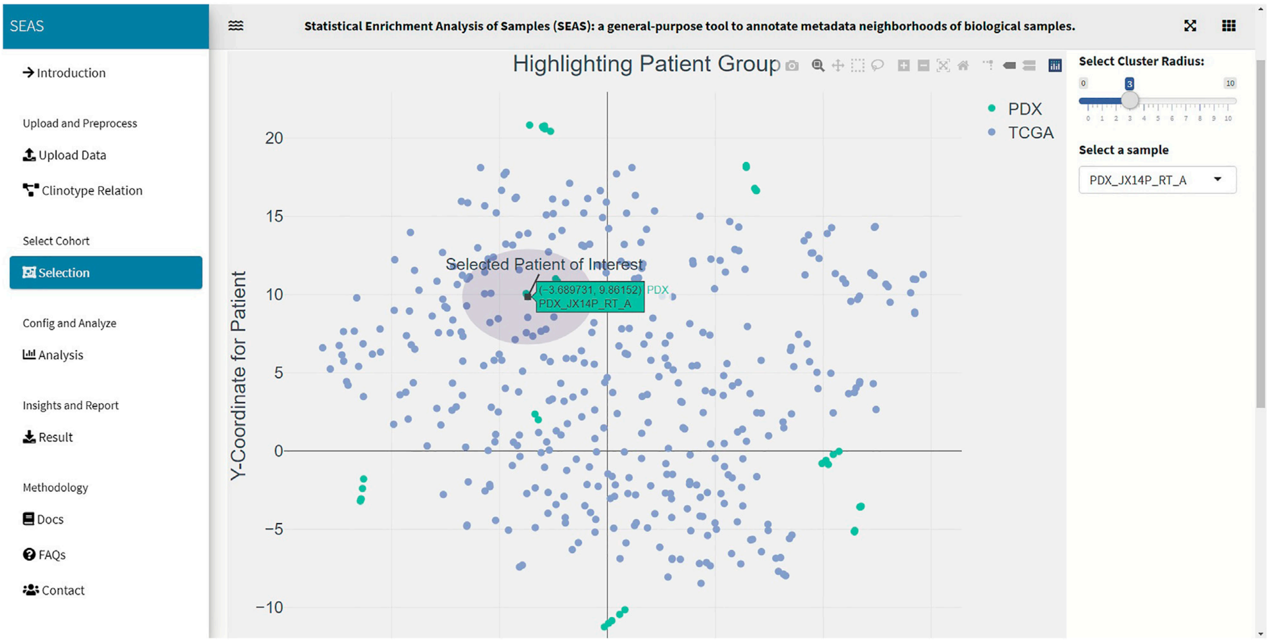
P-Value Threshold: 0.05

Clinotype	Variable	Population	Cohort	P-Value
Discrete_days_to_new_tumor_event_after_initial_treatment	>250	227	14	0.4764
Discrete_CDE_survival_time	>300	232	15	0.5809
Discrete_CDE_survival_time	<=300	157	11	0.7149
Discrete_days_to_new_tumor_event_after_initial_treatment	<=250	162	12	0.7972

Showing 1 to 6 of 6 entries

FIGURE 5 | SEAS identifies enriched clinical features for the subcohort in **Figure 4**.

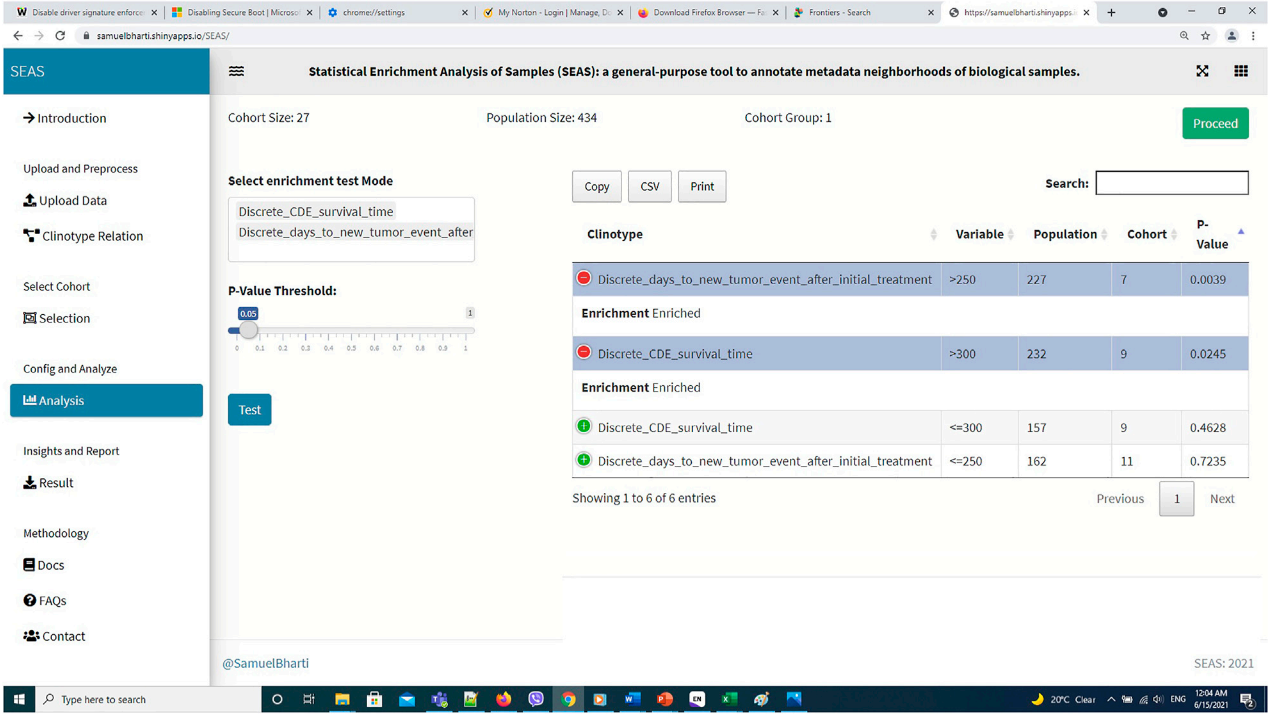
automatically select a subcohort in two ways. In the fully automatic approach, SEAS applies clustering algorithms to divide the patient data into multiple groups. Then, the user selects a group as a subcohort. This approach is preferred because the clustering results can provide the threshold to discretize the numerical attributes into categorical attributes, resulting in the next step. By default, SEAS uses the density-based clustering algorithm (Ester et al., 1996, **Figure 3**). In the semi-automatic approach (**Figures 4**, **6**), the user selects a patient ID, a radius of "similarity area" in the 2D embedding



Select Cluster Radius: 3

Select a sample: PDX_JX14P_RT_A

FIGURE 6 | SEAS identifies a subcohort by a circle region around PDX JX14P_RT_A datapoint.



Cohort Size: 27 Population Size: 434 Cohort Group: 1

Select enrichment test Mode

Discrete_CDE_survival_time
Discrete_days_to_new_tumor_event_after

P-Value Threshold: 0.05

Clinotype	Variable	Population	Cohort	P-Value
Discrete_days_to_new_tumor_event_after_initial_treatment	>250	227	7	0.0039
Enrichment Enriched				
Discrete_CDE_survival_time	>300	232	9	0.0245
Enrichment Enriched				
Discrete_CDE_survival_time	<=300	157	9	0.4628
Discrete_days_to_new_tumor_event_after_initial_treatment	<=250	162	11	0.7235

Showing 1 to 6 of 6 entries

FIGURE 7 | SEAS identifies enriched clinical features for the subcohort in **Figure 6**.

space. All patients in the circle area are centered by the selected patient ID, and the radius becomes the selected subcohort.

Analyze Clinical Feature Enrichment

Besides implementing Wilcoxon-ranksum (Mann and Whitney, 1947) and test between the selected cohort and the whole population for numerical attributes, SEAS defines the CFEA that can be applied for both numerical and categorical attributes. Here, we denoted a patient population **S** and a set of all clinical attributes **C**. Given any cohort **s** in **S**, the main question is which attributes are representative or enriched in **s**. For a categorical attribute, SEAS applies the hypergeometric test, which compares the proportions of patients having the attribute between **s** and **S**. This approach is well-known in gene set enrichment analysis (Falcon and Gentleman, 2008). Here, the null hypothesis is the proportion of patients having attribute C in **s** and **S** is the same. This is analog to the null hypothesis in the Wilcoxon-ranksum (Mann and Whitney, 1947) test, where the median of attribute **C** in **s** and **S** is the same. To apply in numerical data, the numerical attributes are discretized. For example, in our GBM case study, "CDE_survival_time" (survival day), which is a numerical attribute, is discretized into "Discrete_CDE_survival_time <300 days" and "Discrete_CDE_survival_time ≥300 days." As mentioned in the previous section, clustering the patient and using the cluster to determine the numerical thresholds is a good approach. SEAS reports all enriched clinical attributes and their *p*-values and the Bonferroni adjusted *p*-value (for false discovery rate control) (Sedgwick, 2014), as in **Figure 5**.

Implementing the Software

The SEAS web version is built primarily by bs4Dash (https://cran.r-project.org/web/packages/bs4Dash/index.html) and R-shiny (https://shiny.rstudio.com/) packages. Both packages run based on R and can be hosted inside well-known web programming languages: HTML, CSS, and javascript. In addition, the data processing and statistical methods are also implemented in R.

Demo Using TCGA-GBM Dataset

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. Among 108 clinical attributes, 22 categorical and seven numerical ones were used to compute patient similarity and embedding (**Supplementary Data S1**). Also, we used 45 GBM tumor-samples hosted in patient-derived xenograft (PDX) models (Willey et al., 2020). In these samples, the patients were treated by radiation therapy (RT), but did not have clinical information. Besides the automatic embedding using the clinical data, we manually applied tSNE (Hinton and Roweis, 2002) on the combined TCGA-GBM and PDX genetic data as another 2D representation. We checked the quality of the embedding by the close positions of the PDX JX14P_A/JX14P_B sample pair and the PDX JX14P_RT_A/JX14P_RT_B sample pair. These pairs are replicates of the same patient tumor JX14P (before radiation therapy) and JX14P_RT (after radiation therapy—RT), as shown in **Supplementary Figure S1**.

In this case study, to estimate the clinical outcome of an unknown PDX sample, we select a TCGA-GBM subcohort surrounding the PDX sample (**Figures 4**, **6**) and performed SEAS in the selected TCGA subcohort. In **Figures 4**, **5**, SEAS shows no enriched clinical feature for sample PDX JX14P_A. Here, the average survival time among the surrounding TCGA patients was 339 days. In **Figures 6**, **7**, feature "Discrete_CDE_survival_time >300", which means that the patients who survive for more than 300 days, are enriched among the TCGA samples surrounding the PDX JX14P_RT_A sample. Here, the average survival time for these patients was 434 days. This result suggests radiation therapy may improve the clinical condition of the JX14P patient. Thus, SEAS analysis suggests two opposite clinical outcomes for GBM patients even when being treated by the same therapy. The finding could be helpful in further clinical decisions regarding the selected patients.

Other Notes About Similarity Measures and Embedding Options

Similarity Measures

In SEAS, we used the embedded coordinates to compute the Euclidean distance between two patient datapoints

$$d(i,j) = \sqrt{\left(x_i - x_j\right)^2 + \left(y_i - y_j\right)^2} \qquad (1)$$

Here, *i* and *j* denotes two patients, $d(i,j)$ denotes the distance between *i* and *j*, (x_i, y_i) denotes the embedded coordinate for patient *i*, and (x_j, y_j) denotes the embedded coordinate for patient *j*. We did not use any other similarity measure because we assume that the good embedding results already reflect the patient-wise similarity. In case the user's defined similarity could not be reflected by SEAS, the user can manually enter the list of similar patients to perform the enrichment analysis.

Embedding Options

By default, if the user does not supply the embedding input, SEAS may use umap (McInnes et al., 2018) or tSNE (Hinton and Roweis, 2002) to embed the patient from the clinical features. The embedding algorithms, as in (Konopka, 2020), require a pairwise distance or similarity matrix. At this release, SEAS supports the Euclidean distance (default), cosine similarity, and Jaccard index. Besides, the user is encouraged to supply an embedding file for more in-depth analysis. For example, in our GBM case study, the patient pairwise similarity and embedding are computed by the gene expression data instead of the clinical feature. The PDX have gene expression data but do not have clinical attributes; therefore, they could not be embedded correctly with SEAS default option. When the clinical data is insufficient to compute good embedding results, we highly recommend the user to use other tools to compute the embedding prior to using SEAS.

DISCUSSION AND CONCLUSION

To summarize, we developed the user-friendly and online version of SEAS. The tool can provide new and significant insights into clinical data research and may support the clinical decision. In the future, we expect to develop the add-on version of SEAS, which can be integrated into I2B2 clinical data management system.

One limitation in this SEAS first release is that we have not implemented techniques handling missing values in the patients' clinical data. To lower the impact of this limitation, we chose the enrichment methods, such as the hypergeometric test, that do not require a very large data size. In our GBM case study, the population consists of 389 patients, which is a moderate size. However, it is sufficient to perform the statistical test even if the missing data rate for one clinical attribute is 10%. On the other hand, we encourage the user to use the non-clinical data to embed the patients; therefore, the missing clinical data may not impact the quality of SEAS results. In fact, our GBM case study shows an approach to infer unknown clinical attributes in PDX data by SEAS analysis of TCGA-GBM data.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: https://github.com/aimed-uab/SEAS.

AUTHOR CONTRIBUTIONS

TN acquired, processed, and analyzed the combined TCGA-GBM/GBM-PDX data, and wrote the manuscript draft. SB implemented the SEAS software. ZY processed the GBM-PDX data. CW prepared and provided the GBM-PDX data. JC conceptualized the SEAS framework and designed the analytical experiment. All authors read, revised, and approved the manuscript.

FUNDING

The work is partly supported by the National Institute of Health Center for Clinical and Translational Science grant award (No. U54TR002731) to the University of Alabama at Birmingham (UAB) where JC is a co-investigator, a research start-up fund provided by the UAB Informatics Institute to JC, and the National Cancer Institute grant award (No. U01CA223976) in which CW is a principal investigator and JC is a co-investigator.

ACKNOWLEDGMENTS

All authors thank the following general technical support that made case studies included for this work possible: Christian Stackhouse for helping CW generate the GBM-PDX model, Jelai Wang for managing the data management and data analysis computing framework, and Christian Stackhouse and Lara Lanov for executing the RNA-seq analysis pipelines.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fdata.2021.725276/full#supplementary-material

Supplementary Figure S1 | Locations of datapoint pairs PDX JX14P_A / JX14P_B sample pair (top) and the PDX JX14P_RT_A / JX14P_RT_B (bottom).

REFERENCES

Burgun, A., and Bodenreider, O. (2008). Accessing and Integrating Data and Knowledge for Biomedical Research. *Yearb. Med. Inform.*, 91–101.

Ester, M., Kriegel, H.-P., Sander, J., and Xu, X. (1996). *A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise*. In KDD conference, Munich, Germany, 226–231. Available at: https://www.aaai.org/Papers/KDD/1996/KDD96-037.pdf.

Falcon, S., and Gentleman, R. (2008). *Hypergeometric Testing Used for Gene Set Enrichment Analysis, Bioconductor Case Studies*. Springer, 207–220. doi:10.1007/978-0-387-77240-0_14

Fang, S., Palakal, M., Xia, Y., Grannis, S. J., and Williams, J. L. (2014). *Health-Terrain: Visualizing Large Scale Health Data*. Indianapolis, IN: INDIANA UNIV INDIANAPOLIS.

Hinton, G., and Roweis, S. T. (2002). *Stochastic Neighbor Embedding*. Citeseer: NIPS, 833–840.

Hume, S., Sarnikar, S., and Noteboom, C. (2020). Enhancing Traceability in Clinical Research Data through a Metadata Framework. *Methods Inf. Med.* 59, 75–85. doi:10.1055/s-0040-1714393

Jia, D., Li, S., Li, D., Xue, H., Yang, D., and Liu, Y. (2018). Mining TCGA Database for Genes of Prognostic Value in Glioblastoma Microenvironment. *Aging* 10, 592–605. doi:10.18632/aging.101415

Kim, H. H., Park, Y. R., Lee, K. H., Song, Y. S., and Kim, J. H. (2019). Clinical MetaData Ontology: a Simple Classification Scheme for Data Elements of Clinical Data Based on Semantics. *BMC Med. Inform. Decis. Mak* 19, 166. doi:10.1186/s12911-019-0877-x

Konopka, T. (2020). *Package 'umap' Version 0.2.7.0*. CRAN. Available at: https://cran.r-project.org/web/packages/umap/index.html.

Mann, H. B., and Whitney, D. R. (1947). On a Test of whether One of Two Random Variables Is Stochastically Larger Than the Other. *Ann. Math. Statist.* 18, 50–60. doi:10.1214/aoms/1177730491

McInnes, L., Healy, J., and Melville, J. (2018), Umap: Uniform Manifold Approximation and Projection for Dimension Reduction. arXiv preprint arXiv:1802.03426.

Murphy, S. N., Weber, G., Mendis, M., Gainer, V., Chueh, H. C., Churchill, S., et al. (2010). Serving the enterprise and beyond with Informatics for Integrating Biology and the Bedside (I2b2). *J. Am. Med. Inform. Assoc.* 17, 124–130. doi:10.1136/jamia.2009.000893

Nguyen, T., Zhang, T., Fox, G., Zeng, S., Cao, N., Pan, C., et al. (2021). Linking Clinotypes to Phenotypes and Genotypes from Laboratory Test Results in Comprehensive Physical Exams. *BMC Med. Inform. Decis. Mak* 21, 51. doi:10.1186/s12911-021-01387-z

Ohmann, C., and Kuchinke, W. (2009). Future Developments of Medical Informatics from the Viewpoint of Networked Clinical Research. Interoperability and Integration. *Methods Inf. Med.* 48, 45–54.

Sedgwick, P. (2014). Multiple Hypothesis Testing and Bonferroni's Correction. *BMJ* 349, g6284. doi:10.1136/bmj.g6284

Ta, C. N., Dumontier, M., Hripcsak, G., Tatonetti, N. P., and Weng, C. (2018). Columbia Open Health Data, Clinical Concept Prevalence and Co-occurrence from Electronic Health Records. *Sci. Data* 5, 180273. doi:10.1038/sdata.2018.273

Verhaak, R. G. W., Hoadley, K. A., Purdom, E., Wang, V., Qi, Y., Wilkerson, M. D., et al. (2010). Integrated Genomic Analysis Identifies Clinically Relevant Subtypes of Glioblastoma Characterized by Abnormalities in PDGFRA, IDH1, EGFR, and NF1. *Cancer Cell* 17, 98–110. doi:10.1016/j.ccr.2009.12.020

Weng, S. F., Reps, J., Kai, J., Garibaldi, J. M., and Qureshi, N. (2017). Can Machine-Learning Improve Cardiovascular Risk Prediction Using Routine Clinical Data? *PLoS One* 12, e0174944. doi:10.1371/journal.pone.0174944

Willey, C. D., Stackhouse, C. T., Rowland, J. R., Langford, C. P., Anderson, J. C., Ianov, L., et al. (2020). Multi-omic Exploration of Inherent and Acquired Radiation Resistance of Glioblastoma Patient-Derived Xenografts. *Int. J. Radiat. Oncology*Biology*Physics* 108, S40. doi:10.1016/j.ijrobp.2020.07.2148

Zaki, M. J., Meira, W., Jr, and Meira, W. (2014). *Data Mining and Analysis: Fundamental Concepts and Algorithms*. Cambridge University Press.

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.