PedScAtlas

Pediatric **S**ingle **C**ell Cancer **Atlas**: An Integrative Web-Based Resource for Single Cell Transcriptome Data from Pediatric Cancers

Version 2.0.0

https://doi.org/10.1182/blood-2021-154282

Hope L. Mumme, Swati S. Bhasin, Beena E. Thomas, Bhakti Dwivedi, Deborah DeRyckere, Sharon M. Castellino, Daniel S. Wechsler, Sunil S. Raikar, Christopher C. Porter, Douglas K. Graham, Manoj Bhasin

Contact: hmumme@emory.edu, manoj.bhasin@emory.edu

1 Datasets

1.1 Pediatric Leukemias and Young Adult Healthy Controls (Dataset L)

The dataset viewed in the *PediatricCancer* module contains the single-cell rna-sequencing profiles of 82 samples (260k cells): 76 number of pediatric leukemia bone marrow samples taken at diagnosis from 5 different pediatric acute leukemia subtypes (240k cells) and 6 bone marrow samples from healthy young adults (19k cells). See **Table 1** below for the different leukemia subtype group sample and cell amounts and the sources of each sample.

1.2 Adult Healthy Bone Marrow (Dataset H)

The dataset viewed in the *ImmuneCell* module contains the single-cell rna-sequencing profiles of 390k cells from the Human Cell Atlas Census of Immune Cells Project [1]. These cells are all from healthy bone marrow samples.

1.3 Bulk Pediatric Leukemia mRNA-seq from TARGET [X] (Dataset T)

The dataset viewed in the *BulkExpression* module contains the bulk mRNA expression profiles of X samples from the Therapeutically Applicable Research to Generate Effective Treatments (TARGET) program [2] from the National Cancer Institute. Dataset T contains bone marrow diagnosis samples from the ALL-Phase II (n=275), ALL-Phase III (n=52), and AML (n=274) projects. See **Table 1** below for the different leukemia subtype group sample and cell amounts and the sources of each sample. The results published here are in whole or part based upon data generated by the Therapeutically Applicable Research to Generate Effective Treatments (https://ocg.cancer.gov/programs/target) initiative, phs000218. The data used for this analysis are available at https://portal.gdc.cancer.gov/projects.

PedScAtlas Dataset	RNA Seq. Type	Source	Disease Type	N Samples	N Cells
Dataset L	SC	ScPCA [3]	AML	20	86,133
			YA HC	2	13,000
			T/M MPAL	2	12,155
		Bhasin Lab [4]	AML	20	65,732
			B/M MPAL	3	10,961
			T/M MPAL	2	5,443
			T-ALL	10	28,218
			B-ALL	4	8,349
		GSE154109 [5]	AML	8	12,267
			B-ALL	7	10,892
			YA HC	4	5,944
Dataset H		HCA [1]	A HC	-	391,505
Dataset T	Bulk	TARGET [2]	AML	274	-
			B-ALL	11	-
			T-ALL	264	-
			B/M MPAL	23	-
			T/M MPAL	29	-

Table 1. Sample distributions among each PedScAtlas dataset. RNA Sequencing types include Single Cell (SC) and Bulk. Disease subtypes include the pediatric leukemias: acute myeloid leukemia (AML), T-cell acute lymphoid leukemia (T-ALL), B-cell acute lymphoid leukemia (B-ALL), T/Myeloid mixed phenotype acute leukemia (T/M MPAL), and B/Myeloid mixed phenotype acute leukemia (B/M MPAL). The healthy control groups: young adult healthy control (YA HC), and adult healthy control (A HC). Dataset sources include ScPCA [3] (Single-cell Pediatric Cancer Atlas by Alex's Lemonade), Bhasin Lab [4], GSE154109 [5], Human Cell Atlas [1] (HCA), and Therapeutically Applicable Research to Generate Effective Treatments [2] (TARGET).

2 Modules

2.1 PediatricCancer

The *PediatricCancer* module allows users to interact with dataset L, the pediatric leukemia and young adult healthy control dataset. The general steps users will take when accessing this module is (1) choose which disease types they want to access, (2) choose which format they want to view the data in, and (3) choose which disease aspect they want to group the data by.

During step 1, the user chooses which disease types (ex: AML, B-ALL, HC …) under the "Choose Disease Type to Show" they want to access and view; the disease types boxes checked will show on the right-hand plot. Next, during step 2 the user will choose to view the data in a Uniform Manifold Approximation and Projection (UMAP) plot, violin plot, or feature plot under the "Viewing Format" section. The UMAP option shows cell locations along with data aspect information as colors, the violin plot shows single gene expression grouped by a data aspect, and the feature plot shows cell location along with gene expression. And finally, during step 3 the user can choose which disease aspects to show as colors on the UMAP or as groups on the violin plot using the "Data Aspect to Show" section. The drop-down reveals the following options: cell clusters, cell type, disease type, sample id, and future relapse or remission status if the information is available.

To view a gene's expression on the violin or feature plot, the user must input a gene in the "Enter Gene Name to show Expression" text box. If this gene is found in the dataset, the expression plot will be generated on the right side of the screen. Also, there is an option to filter the cells shown on the feature plot via the sliding bar on the bottom left of the screen. By increasing the filter above zero, only cells with expression in the top n^{th} percentile will show on the feature plot. For example, if the user wants to only view cells with expression in the top 90th percentile of cells, they would increase the slide to 90.

2.2 ImmuneCell

The *ImmuneCell* module allows users to interact with dataset H, the adult healthy bone marrow dataset. The user inputs a gene of their choice in the "Enter Gene to show Expression" section, and if this gene is present in the dataset the expression will be plotted on a feature plot on the right side of the screen.

2.3 Biomarkers

The *Biomarkers* module allows users to interact with dataset L via pre-built in biomarker sets for each leukemia subtype. The user can either view a feature plot or violin plot with the gene set enrichment for the chosen biomarker set. Once again, they choose the viewing format in the “Viewing Format” section. The user chooses a biomarker set for a leukemia subtype in the “Biomarker Set” section. They can group the violin plot by a data aspect of their choice in the “Data Aspect to Show” section.

2.4 BulkExpression

The *BulkExpression* module allows users to interact with dataset T, the bulk RNA expression dataset from TARGET [2]. The module contains two ways to view gene expression in dataset T, by showing a box plot with a single gene’s normalized expression or with the enrichment of a gene set for each disease group (AML, B-ALL, B/M MPAL, T-ALL, and T/M MPAL). The user would choose whether they want to input a single gene or a gene set by choosing either “Single Gene” or “Gene Set” in the “Select Gene Expression Viewing” section.

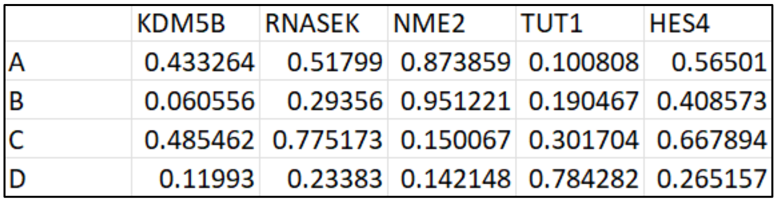
If the user chooses the single gene option, a text box will show under the “Enter Gene Name to Show Bulk Expression” section. If they choose the gene set option, a text box will show under the “Enter Gene Names, One per Line” section; for this option, the user must input their gene list, only one gene per line and not separated by commas or spaces.

They can also choose whether to include statistical comparisons on the box plot using the drop down in the “Choose which statistical comparisons to perform” section. Possible statistical comparisons include a one-way Analysis of Variance (ANOVA) and Wilcox T-tests comparing all subtypes individually to a single subtype. If comparisons are chosen, the p-values will show on the generated box plot.

2.5 CancerTypePrediction

The *CancerTypePrediction* module predicts the cancer subtype of a user’s bulk RNA-seq expression data. Currently, the module uses a Random Forest Classifier that predicts a leukemia subtype class based on the expression of the biomarker sets shown in the *Biomarkers* module. The classifier was trained using dataset T to classify samples as AML, B-ALL, T-ALL, B/M MPAL, or T/M MPAL.

The user uploads the bulk-rna seq expression data as a csv file with the genes as columns and the sample(s) as rows in the “Input Sample(s) section”. See **Fig. 1** below for a screenshot of a properly formatted csv file for this step. The list of the necessary genes is available for download by clicking the “Download Leukemia Biomarker Gene Names” button. The user presses the “Predict Leukemia Subtype of Sample(s)” button to generate the predictions for the samples in their data. The predicts will be shown on a table on the right side of the screen.



	KDM5B	RNASEK	NME2	TUT1	HES4
A	0.433264	0.51799	0.873859	0.100808	0.56501
B	0.060556	0.29356	0.951221	0.190467	0.408573
C	0.485462	0.775173	0.150067	0.301704	0.667894
D	0.11993	0.23383	0.142148	0.784282	0.265157

Figure 1. The first five columns of an example user input for the *CancerTypePrediction* module. Columns are the 37 leukemia biomarker gene names and rows are the 4 samples (A-D).

4 Error Messages

4.1 Gene Expression

If a gene is not found in the expression data, the following error messages will show: in the *PediatricCancer* and *ImmuneCell* modules, “The gene entered was not found in the expression data”, and in the *BulkExpression* module, “The gene entered was not found in the expression data” for the single gene option and “The following genes were not found in the expression data, please remove from input:” for the gene set option.

If the update plot button is pushed in any of the modules without inputting a gene or gene list in the text entry plot (when applicable viewing option is chosen), then the “Please enter gene to view expression” error will show.

For the *BulkExpression* module, the gene set entry option requires a precise format to plot the enrichment of the user’s gene set. If there are extra lines, spaces, or commas the “Please check formatting. There should be one gene per line, no extra spaces or commas.” error will show. If the user inputs only one gene in the gene set entry box, the “If you would like to view the expression of one gene, please use the ‘Single Gene’ selection” error will show.

4.2 User Data

If the user does not input a file in the csv format in the *CancerTypePrediction* module, then the error message “Please upload a csv file” will show. If the user does not enter at least 10 genes found in the biomarker set needed for prediction, then the error message "Please include at least 10 of the required biomarkers in your uploaded dataset" will show.

References

[1] Human Cell Atlas Data Portal. Census of Immune Cells. https://data.humancellatlas.org/explore/projects/cc95ff89-2e68-4a08-a234-480eca21ce79 (2022).

[2] National Cancer Institute. TARGET: Therapeutically Applicable Research To Generate Effective Treatments. https://ocg.cancer.gov/programs/target (2022).

[3] Childhood Cancer Data Lab, Alex's Lemonade Stand. Single-cell Pediatric Cancer Atlas (ScPCA) Data Portal. https://scpca.alexslemonade.org/ (2022).

[4] Bhasin Systems Biomedicine Lab, Emory University School of Medicine and Children's Healthcare of Atlanta. http://www.bhasinlab.org/ (2022).

[5] Bailur JK, McCachren SS, Pendleton K, Vasquez JC et al. Risk-associated alterations in marrow T cells in pediatric leukemia. JCI Insight 2020 Aug 20;5(16). GSE154109. PMID: 32692727