



Mar 26, 2025

🌐 Developmental State Quantification from Bulk Transcriptomes in B-ALL

📖 [Nature Cancer](#)

DOI

<https://dx.doi.org/10.17504/protocols.io.dm6gpde9jgzp/v1>

Andy Zeng¹

¹Princess Margaret Cancer Centre, University Health Network, Toronto, Canada



Andy Zeng

University of Toronto

Create & collaborate more with a free account

Edit and publish protocols, collaborate in communities, share insights through comments, and track progress with run records.

Create free account

OPEN  ACCESS



DOI: <https://dx.doi.org/10.17504/protocols.io.dm6gpde9jgzp/v1>

External link: <https://doi.org/10.1038/s43018-025-00987-2>

Protocol Citation: Andy Zeng 2025. Developmental State Quantification from Bulk Transcriptomes in B-ALL. **protocols.io** <https://dx.doi.org/10.17504/protocols.io.dm6gpde9jgzp/v1>

Manuscript citation:

Iacobucci, I., Zeng, A.G.X., Gao, Q. *et al.* Multipotent lineage potential in B cell acute lymphoblastic leukemia is associated with distinct cellular origins and clinical features. *Nat Cancer* **6**, 1242–1262 (2025). <https://doi.org/10.1038/s43018-025-00987-2>

License: This is an open access protocol distributed under the terms of the **Creative Commons Attribution License**, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited

Protocol status: Other

This protocol contains supplementary methods for the corresponding manuscript in Nature Cancer which could not be incorporated into the manuscript due to length restrictions. Please see the published manuscript, methods, and main / extended figures for more information and context.

Created: March 18, 2025

Last Modified: March 26, 2025

Protocol Integer ID: 124613

Keywords: developmental states in bulk transcriptome profile, developmental state quantification from bulk transcriptome, bulk transcriptome profile, bulk transcriptome, transcriptome, distinct cellular origin, cell acute lymphoblastic, multipotent lineage potential, developmental state quantification, nature cancer, developmental state

Abstract

This protocol contains methods pertaining to the quantification of B-ALL developmental states in bulk transcriptome profiles from primary B-ALL patient samples.

This pertains to the following manuscript:

MULTIPOTENT LINEAGE POTENTIAL IN B-CELL ACUTE LYMPHOBLASTIC LEUKEMIA IS ASSOCIATED WITH DISTINCT CELLULAR ORIGINS AND CLINICAL FEATURES

by Iacobucci, Zeng, Gao, Garcia-Prat, *et al.*

This manuscript is in press at Nature Cancer.

An earlier version of this manuscript can also be found on bioRxiv: <https://pubmed.ncbi.nlm.nih.gov/38106088/>

Note 1 - Developmental State Quantification in B-ALL bulk RNA-seq

1 **Background.**

We start with a single cell transcriptomic data set in which leukemia cells from 89 patient samples have been assigned to specific cell states (termed B-ALL Developmental States within our study). This will be used to develop the gene expression scores which will be used to quantify Developmental State abundance within bulk transcriptomes from B-ALL patients.

The resulting scores are validated using independent bulk RNA-seq profiles from 85 / 89 of the patient samples for which scRNA-seq was performed.

2 For each developmental state, a metric is required to quantify the relative abundance of this state within each patient sample. Within our study, developmental states are derived from cell composition analysis of more granular cell states, with correlated cell states grouped into a broader developmental state using non-negative matrix factorization (NMF). Thus, the relative abundance of each developmental state is represented by the score for each corresponding composition-based NMF component within each patient. To estimate the abundance of each developmental state in bulk RNA-seq data, we identified biologically relevant genes which were correlated to the abundance of each developmental state by integrating three approaches of feature selection.

2.1 **First**, we created patient-level pseudo-bulk profiles from the 89 patient samples and identified genes that were significantly associated with abundance of each developmental state. Briefly, pseudo-bulk samples were vst normalized using DESeq2 (v1.32.0) and a Pearson test was performed between normalized gene expression and abundance of each developmental state across 89 patient samples. Filtering at FDR < 0.05 resulted in thousands of significantly associated genes for each developmental state, thus further filtering was required. First, adaptive K1 thresholding, adapted from the AUCell R package, was performed within significantly positively correlated genes and within significantly negatively correlated genes.

2.2 **Second**, we identified differentially expressed genes across normal B cell development. Cells belonging to each normal cell type within a healthy donor from the B cell development reference map were pooled into pseudo-bulk profiles. Only profiles comprised of ≥ 5 cells were retained. One-vs-all differential expression was performed between each cell state and all others using DESeq2 (v1.32.0), correcting for Study ID and Tissue Source (Fetal Liver, Fetal Bone Marrow, Cord Blood, Bone Marrow) as covariates. Among the correlated genes, only those that were differentially expressed across normal B cell development at FDR < 0.01 were retained.

2.3 **Third**, we identified differentially expressed genes across B-ALL developmental states. Cells belonging to each developmental state within a B-ALL patient were pooled into pseudo-bulk profiles. Only profiles comprised of ≥ 100 cells were retained. One-vs-all

differential expression was performed between each developmental state and all others using DESeq2 (v1.32.0), correcting for Patient ID as a covariate. Among the correlated genes, only those that were differentially expressed across B-ALL developmental states at $FDR < 0.01$ were retained.

3 For each developmental state, only genes meeting all of the following criteria were used:

Criteria 1: Positively or Negatively correlated with abundance of this B-ALL developmental state at $FDR < 0.05$ and exceeding an adaptive positive or negative threshold

Criteria 2: Significantly differentially expressed across normal B cell development

Criteria 3: Significantly differentially expressed across B-ALL developmental states

Using genes that met these criteria, LASSO regression was performed on developmental state abundance using vst normalized gene expression data from pseudo-bulk profiles of 89 B-ALL patient samples, using R package glmnet (v4.1.2). Briefly, to optimize model parameters 5-fold cross-validation was performed within each dataset with 10 repeats. Data was shuffled in between repeats, resulting in performance estimates for 50 random train/test splits. Pearson correlations across each train/test split was used as a measure of accuracy on unseen data.

4 Final models for each developmental state were trained using LASSO, with leave-one-out cross-validation to determine the lambda value corresponding to the lowest root mean square prediction error (RMSE). To further reduce the number of features in the model, the largest lambda within one standard error of the lowest RMSE was used for the training the final model. This resulted in a different gene expression model for inferring abundance of each B-ALL developmental state, with all six models captured within 145 genes.

To validate these gene expression scores, we obtained matched bulk RNA-seq profiles for 85 B-ALL samples profiled by scRNA-seq. Inferred Developmental State abundance calculated from bulk RNA-seq profiles were compared against actual Developmental State abundance derived from the original scRNA-seq composition analysis.

Note 2 - Deconvolution Benchmarking in B-ALL

5 Background.

To accurately quantify B-ALL-specific developmental states within bulk transcriptomes, we sought to benchmark multiple gene expression deconvolution methods and evaluate them against our new regression-based approach for bulk RNA-seq quantification. For deconvolution analysis, we evaluated the following approaches: BayesPrism (Chu *et al.* Nature Cancer 2022), CIBERSORTx (Newman *et al.* Nature Biotechnology 2019),

Dampened Weighted Least Squares (DWLS) (Tsoucas *et al.* Nature Communications 2019), and Support Vector Regression (SVR). As a reference for deconvolution, we utilize the 10 broad Developmental states in B-ALL to minimize co-linearity between similar cell types.

6 Feature selection

Feature selection is a critical step for maximizing the accuracy of gene expression deconvolution. To approach this, we utilized multiple approaches in parallel. For feature selection, we utilized pseudobulk profiles from pooling scRNA-seq count data across cells within each B-ALL developmental state or normal immune states within each patient sample, hereafter Dev State Pseudobulks. These spanned 10 broad malignant and normal states (HSC/MPP, Myeloid Progenitor, Pre-pDC, Early Lymphoid, Pro-B, Pre-B, Mature B, T/NK, Monocyte, and Erythroid) across 89 patient samples. We utilize a pseudobulk reference for deconvolution to reduce noise from scRNA-seq data and manage inter-patient heterogeneity. From these Dev State Pseudobulks, we applied **two feature selection approaches** in parallel:

6.1 Feature Selection Approach #1 - Batch-aware highly variable genes.

To derive batch-aware highly variable genes, we first excluded Dev State Pseudobulks comprised of less than 100 cells. We also filtered out genes that were present in less than 10 Dev State Pseudobulk profiles. After this, we excluded patients with less than two distinct Dev State Pseudobulk profiles. After performing logCPM normalization we identified highly variable genes across Dev State Pseudobulks within scanpy. using `sc.pp.highly_variable_genes` and specifying `n_bins=100`, `min_mean=0.1`, `max_mean=12`, `min_disp=0.5`, and controlling for inter-patient variation with the parameter `batch_key='Patient'`. This returned a set of 3,583 highly variable genes across Dev State Pseudobulk samples.

6.2 Feature Selection Approach #2 - Differentially expressed genes across normal and malignant B-cell development.

In line with the approach outlined in Supplemental Note 2 for developmental state quantification in bulk B-ALL RNA-seq profiles, this approach utilizes genes that meet all three of the following criteria:

Criteria 1: Differentially expressed across B-ALL developmental states.

Positively enriched within a given B-ALL developmental state at $FDR < 0.05$, based on DESeq2 differential expression analysis among Dev State Pseudobulk samples.

Criteria 2: Differentially expressed within normal B-cell development.

Positively enriched within a given cell type along our B cell development atlas, at $FDR < 0.01$, based on DESeq2 differential expression analysis among pseudobulk profiles from each normal cell type within each donor from our B cell development atlas.

Criteria 3: Significantly correlated with developmental state abundance among 89 sample-level pseudobulk samples.

Genes for which bulk expression is significantly positively or negatively correlated with the abundance of a B-ALL developmental state at $FDR < 0.01$. Among significantly positively and negatively correlated genes for any given B-ALL developmental state, adaptive thresholding through the AUCCell R package is applied to identify subsets of genes with the strongest correlations.

The intersection of all three criteria returned a list of 2,671 biologically relevant genes that are significantly differentially expressed across normal and malignant B-cell development as well as significantly correlated with the abundance of B-ALL developmental states among sample-level pseudobulk profiles.

7 Signature matrix generation:

For deconvolution with CIBERSORTx, DWLS, and SVR, an additional pre-processing step was applied to derive signature matrices from each of the two feature sets. Specifically, an additional round of feature selection using multi-objective optimization was performed with the AutoGeneS tool (Ailee & Theis, Cell Systems 2021), starting from each of these two feature sets. Dev State Pseudobulks were CPM-normalized and mean expression values for each population ('centroids') were used by AutoGeneS to refine the feature set and maximize separation between populations.

Critically, starting from each feature set, AutoGeneS was run in hierarchical mode constituting three independent rounds of feature selection. The first round (500 generations) sought to separate all normal and malignant populations. The second round (100 generations) focused on discerning between populations along B cell development. The third round (100 generations) focused on discerning between Pro-B and Pre-B populations. Feature sets from each of the three rounds were combined to constitute the final signature matrix.

After applying AutoGeneS to the original set of 3,583 highly variable genes, a refined subset consisting of 1,135 genes was used for signature matrix generation. After applying AutoGeneS to the original set of 2,671 biologically relevant genes, a refined subset consisting of 726 genes was used for signature matrix generation. For each set of features, the final signature matrix was represented by CPM-normalized centroids for each normal and malignant population.

8 Bulk RNA-seq samples for benchmarking analysis

Following signature matrix generation, we sought to benchmark the accuracy of gene expression deconvolution. To do so, we identified 85 B-ALL patient samples from our scRNA-seq cohort for which bulk RNA-seq profiles were also available. We next applied deconvolution to these 85 bulk RNA-seq profiles to evaluate the association between scRNA-seq abundance and deconvolution-inferred abundance for each malignant population.

9 Deconvolution with CIBERSORTx, DWLS, and SVR

Using these two signature matrices (matrix 1: highly variable genes + AutoGeneS; matrix 2: biologically relevant genes + AutogeneS), we applied deconvolution to CPM-normalized bulk transcriptomes from the 85 matched patient samples. For CIBERSORTx deconvolution, we applied B-mode batch correction given that our signature matrix was derived from pseudobulk profiles rather than scRNA-seq data. For DWLS and SVR, we used deconvolution functions provided by the original DWLS publication (Tsoucas *et al.* Nature Communications 2019).

Deconvolution with BayesPrism

For deconvolution with BayesPrism, we utilized a different approach for highly variable gene selection than outlined above. Instead, the standard BayesPrism feature selection approach was employed wherein mitochondrial genes, ribosomal genes, MALAT1, chrX/chrY genes, and pseudogenes were removed (Chu *et al.* Nature Cancer 2022). Subsequently, BayesPrism was used to identify markers for each cell type with $pval.max=0.1$ and $lfc.min=0.1$, resulting in a total of 7,215 variable genes used specifically for BayesPrism deconvolution.

Starting from either the set of variable genes or the set of biologically relevant genes, deconvolution with BayesPrism was applied to the CPM-normalized bulk transcriptomes from the 85 matched patient samples using profiles from 423 Dev State Pseudobulks as a reference, and the final theta predictions of population abundance were used as the BayesPrism estimates.

10 Deconvolution using a reference of normal B cell development

Previous works have directly used normal hematopoietic populations along B cell development as a reference for deconvolution rather than populations derived from B-ALL patients (Beder *et al.* Hemasphere 2023; Hu *et al.* Haematologica 2024; Huang *et al.* Cancer Cell 2024; Khabirova *et al.* Nature Medicine 2022). We have shown in the context of myeloid malignancies that this leads to inferior deconvolution performance (Zeng *et al.* Nature Medicine 2022). Nonetheless, we sought to perform deconvolution from a normal hematopoietic reference as an additional benchmark using each of the four approaches (BayesPrism, CIBERSORTx, DWLS, SVR).

To perform deconvolution with a normal hematopoietic reference, we utilized single-cell transcriptomes from our B cell development atlas as a reference. Specifically, we condensed the numerous cell states into the seven developmental states spanning B cell development (HSC/MPP, Myeloid Progenitor, Pre-pDC, Early Lymphoid, Pro-B, Pre-B, Mature B), and generated pseudobulk profiles by pooling cells from each developmental state from each healthy donor, only retaining pseudobulk profiles comprised of 100 or more cells.

For signature matrix generation prior to deconvolution with CIBERSORTx, DWLS, or SVR, AutoGeneS was applied in hierarchical mode to 3,031 batch-aware highly variable genes identified across normal B cell development. The first round of AutoGeneS (500 generations) sought to separate all populations while the second round (100 generations) focused on discerning between Pro-B and Pre-B populations. Feature sets from each round were combined to constitute a final signature matrix of 637 genes representing developmental states spanning normal B cell development. Deconvolution with CIBERSORTx (B-mode), DWLS, and SVR was performed on the 85 bulk RNA-seq profiles as described above.

For BayesPrism deconvolution with the normal hematopoietic reference, pseudobulk profiles from normal B cell developmental states were subject to standard BayesPrism preprocessing and feature selection as described above, resulting in a final matrix of 6550 genes across 233 normal pseudobulk profiles used to run BayesPrism. Final theta predictions of population abundance were used as the BayesPrism estimates.

11 **Benchmarking deconvolution performance in B-ALL**

To benchmark the accuracy of B-ALL Developmental State abundance quantification in bulk RNA-seq samples, we evaluated the correlation between the observed abundance of each B-ALL Developmental State by scRNA-seq with the inferred abundance of each B-ALL Developmental State from our 85 matched bulk RNA-seq profiles. Specifically, we compared our new LASSO-based approach against deconvolution approaches including BayesPrism, CIBERSORTx, DWLS, and SVR, as well as against these same deconvolution approaches utilizing a normal B cell development reference.

Critically, we found that our LASSO-based approach for B-ALL Developmental State quantification significantly outperformed deconvolution methods including CIBERSORTx, DWLS, BayesPrism, and SVR, regardless of the feature selection approach. Across the B-ALL Developmental States, predicted abundance by LASSO yielded a median correlation of 0.87 and median R-squared of 0.75, whereas the next best method, CIBERSORTx, yielded a median correlation of 0.57 and median R-squared of 0.32. Finally, our LASSO approach exhibited minimal variance in performance across B-ALL developmental states and was less susceptible to population dropout than other methods.

Together, this demonstrates that when many scRNA-seq samples are available, the application of LASSO regression to predict population abundance across patient-level pseudobulk profiles from a carefully filtered set of biologically relevant genes can outperform popular deconvolution methods for quantification of population abundance from bulk transcriptomes. This is particularly advantageous in quantifying rare populations, which are susceptible to drop-out with deconvolution approaches. However, this approach faces an important limitation in that it does not directly predict sample composition, and thus caution is warranted in directly comparing abundance scores between populations.

