Supplementary Materials for

scAB detects multiresolution cell states with clinical significance by integrating single-cell genomics and bulk sequencing data

Qinran Zhang¹ ², Suoqin Jin¹ ²* and Xiufen Zou¹ ²*

¹ School of Mathematics and Statistics, Wuhan University, Wuhan, 430072, China
² Hubei Key Laboratory of Computational Science, Wuhan University, Wuhan, 430072 China

* Correspondence: S.J. (sqjin@whu.edu.cn) and X.Z. (xfzou@whu.edu.cn)

This file includes the following subsections:

- Supplementary Text
  - Optimization algorithm for scAB
  - Proof of convergence of the proposed optimization problem
  - Implementation of the scAB model in the bulk RNA-seq data
  - Details of method comparisons
  - Exploration of different similarity metrics

- Supplementary Figures 1-6

- Supplementary Tables 1-3
Supplementary Text

Optimization algorithm for scAB

The optimization problem of scAB model is solved by a multiplicative update rule, which updates variables $W$ and $H$ iteratively according to the following equations:

$$w_{ik} \leftarrow w_{ik} \frac{(XH^T)_{ik}}{(WHH^T + \alpha_1SSW)_{ik}} \quad \text{(S1)}$$

$$h_{kj} \leftarrow h_{kj} \frac{(W^TX + \alpha_2HD^{-\frac{1}{2}}A^TD^{-\frac{1}{2}})_{kj}}{(W^TWH + \alpha_2H)_{kj}} \quad \text{(S2)}$$

where $w_{ik}$ represents the entry in the $i$th row and $k$th column of $W$, and $h_{kj}$ represents the entry in the $k$th row and $j$th column of $H$. The complete description of the scAB algorithm is presented as follows:

1. Calculate $S$, $D$, and $L$
2. Initialize $W$, $H$ using a 0-1 uniform distribution
3. Fix $W$, $H$ and then update $W$ by Eq. (S1)
4. Fix $W$, $H$ and then update $H$ by Eq. (S2)
5. Repeat steps 3-4 until satisfying the stop criterion

There are two stop criterions for scAB algorithm. On the one hand, the algorithm will be terminated if the objective function tends to be stable, that is $|F(t) - F(t+1)| \leq 10^{-4}$, where $F(t)$ is the objective function in Eq.(1) of the $t$-th iteration. On the other hand, the algorithm will be terminated when the number of iterations is large than 2000.

Proof of convergence of the proposed optimization problem

**Definition 1** $G(h, h')$ is an auxiliary function for $F(h)$ if the following conditions are true: $G(h, h') \geq F(h), G(h, h) = F(h)$.

**Lemma 1** If $G(h, h')$ is an auxiliary function of $F(h)$, $F(h)$ is a non-increasing function under the update formula

$$h^{t+1} = \arg\min_h G(h, h') \quad \text{(S3)}$$
Proof:

\[ F(h'^{+1}) < G(h'^{+1}, h') \leq G(h', h') = F(h') \]

Furthermore, we rewrite the objective function as follows:

\[
F(W, H) = \|X - WH\|_F^2 + \alpha_1 \|SW\|_2^2 + \alpha_2 \text{tr}(HLH^T)
\]

\[
= \sum_{i=1}^n \sum_{j=1}^m \left(x_{ij} - \sum_{k=1}^K w_{ik} h_{ij} \right)^2 + \alpha_1 \sum_{i=1}^n \left(s_{ij} \sum_{k=1}^K w_{ik} \right)^2 + \alpha_2 \sum_{k=1}^K \sum_{j=1}^m \sum_{i=1}^n h_{ij} L_{ij} h_{ik} \tag{S4}
\]

Considering any element \( h_{ab} \) in F, we use \( F_{ab} \) to denote the part of F, which is only relevant to \( h_{ab} \). The partial derivative and second-order partial derivative of \( F_{ab} (h_{ab}) \) with respect to element \( h_{ab} \) are

\[
F'_{ab} (h_{ab}) = \left( \frac{\partial F}{\partial H} \right)_{ab} = \left( -2W^T X + 2W^T WH + 2\alpha_2 HL \right)_{ab} \tag{S5}
\]

\[
F''_{ab} (h_{ab}) = 2(W^T W)_{bb} + 2\alpha_2 (L)_{aa} \tag{S6}
\]

**Lemma 2** \( G(h, h') \) is an auxiliary function for \( F_{ab} (h) \) when \( G(h, h') \) is as follow:

\[
G(h, h') = F_{ab} (h_{ab}') + (h - h') F'_{ab} (h_{ab}') + (h - h')^2 \left( \frac{(W^T WH)}{h_{ab}'} + \alpha_2 (H)_{ab} \right) \tag{S7}
\]

Proof:

\( G(h, h) = F_{ab} (h) \) is obvious, we only need to prove that \( G(h, h') \geq F_{ab} (h) \). We compare the Taylor expansion of the \( F_{ab} (h) \),

\[
F_{ab} (h) = F_{ab} (h_{ab}') + (h - h') F'_{ab} (h_{ab}') + (h - h')^2 \left[ (W^T W)_{bb} + \alpha_2 L_{aa} \right] \tag{S8}
\]

with \( G(h, h') \) to find that \( G(h, h') \geq F_{ab} (h) \) is equivalent to

\[
\left( \frac{(W^T WH)}{h_{ab}'} + \alpha_2 (H)_{ab} \right) \geq (W^T W)_{bb} + \alpha_2 L_{aa} \tag{S9}
\]

We have

\[
(W^T WH)_{ab} = \sum_{i=1}^k h_{ij}' (W^T W)_{ib} \geq h_{ab}' (W^T W)_{bb} \tag{S10}
\]

and \( \alpha_2 (H)_{ab} = \alpha_2 \sum_{j=1}^n h_{kj}' \geq \alpha_2 h_{ab}' \geq \alpha_2 (I - W)_{aa} h_{ab}' = \alpha_2 L_{aa} h_{ab}' \tag{S11} \)
Thus, \( G(h, h'_\text{ab}) \geq F_{ab}(h) \).

**Theorem 1** The objective function Eq. (1) in the main text is nonincreasing under the updating rules in Eq. (S1) and Eq. (S2) in the main text.

**Proof:**

Replacing \( G(h, h'_\text{ab}) \) in (S3) by (S7) results in the update rule

\[
\begin{align*}
    h'^{t+1}_{ab} &= h'_t - h'_{t-1} \\
    &= h'_a - \frac{F'_{ab}(h'_\text{ab})}{2(W^TWH)_{ab} + 2\alpha_z (H)_{ab}} \\
    &= h'_a - \left( \frac{W^T X + \alpha_z HD^{-\frac{1}{2}} A^{-\frac{1}{2}} D^{-\frac{1}{2}}}{W^T WH + \alpha_z H} \right)_{ab} \\
    \text{(S12)}
\end{align*}
\]

By reversing the roles of W and H in Lemma 1 and 2, the objection function F can similarly be shown to be nonincreasing under the update rules for W.

**Implementation of the scAB model in the bulk RNA-seq data**

The scAB model in (Eq. (1)) can be naturally degenerated to the scenario when only bulk expression matrix and survival information are available. Briefly, given a normalized bulk expression matrix \( X \), we replace the symmetric normalized Laplacian matrix \( L \) in the optimization problem (Eq. (1)) by an identity matrix \( I_{m\times m} \) and solve the following optimization problem

\[
\begin{align*}
    \min_{W,H} & \|X - WH\|_F^2 + \alpha_z \|SW\|_F^2 + \alpha_z \|H\|_F^2, \\
    \text{s.t.} & \ W \geq 0, H \geq 0 \\
    \text{(S13)}
\end{align*}
\]

The expression matrix \( X \) is decomposed into a sample loading matrix \( W \) and a gene loading matrix \( H \), and the gene loading matrix is used to determine phenotype-associated gene signatures. The loading values represent the contributions of each gene in each program, and genes with high loading values are defined as phenotype-associated genes. The definition of high loading value is consistent with the previous definition in single-cell analysis.

**Details of method comparisons**

**Details of data analysis by Scissor.** Scissor v2.0.0 was used for the analysis of scRNA-seq data based on the vignette provided in https://sunduanchen.github.io/Scissor/vignettes/Scissor_Tutorial.html. For all the bulk and single-cell datasets, the same datasets used in scAB were used as inputs. Default parameters were used in all the analyses. After identifying phenotype-associated cells, the downstream analysis pipelines were consistent with scAB.
Details of data analysis by scPrognosis. scPrognosis was used for the analysis of scRNA-seq data based on the code provided in https://github.com/XiaomeiLi1/scPrognosis/blob/master/R/main.R. Default parameters were used in all the analyses. The EMT Pseudotime for individual cells was estimated by their expression values of VIM. In the liver cancer dataset, the genes identified by scPrognosis were used to predict prognostic scores by the cox model. In the melanoma dataset, scPrognosis-identified genes were used to predict scores by Gene Set Variation Analysis (GSVA).

Details of data analysis by DEGAS. DEGAS v0.1.0 was used for the analysis of scRNA-seq data based on the vignette provided in https://github.com/tsteelejohnson91/DEGAS/blob/master/MM_example/MM_example.md. The bulk and single-cell data normalized by the "normalizeScale" function from DEGAS package were used as inputs. Default parameters were used in all the analyses. Cox proportional hazard output from DEGAS model was used as the prediction score of patients.

Exploration of different similarity metrics
We explored other similarity calculation strategies, including Spearman correlation, mutual information (MI), Euclidean similarity 1 (ES1) and Euclidean similarity 2 (ES2). Here ES1 was defined as follows

\[ ES1(x_{bulk}, x_{sc}) = \frac{1}{1 + dist(x_{bulk}, x_{sc})} \]

where "dist" represented the Euclidean distance between the gene expression profile of bulk data and that of single-cell data. ES2 was defined in a similar way, but the Euclidean distance was calculated in the PCA-space.

We run scAB by taking these similarity matrices as inputs and then assessed the performance of different similarity metrics by using either concordance index (C-index) or the p-values of Wilcoxon tests. On the liver cancer dataset with survival information, we observed higher C-indexes when using Pearson correlation and MI compared to Spearman correlation, ES1 and ES2 (Supplementary Figure 5). On the melanoma dataset with immunotherapy information, after running scAB on the training dataset PRJEB23709, we used the candidate biomarker sets identified by different similarity metrics to calculate the gene set variation analysis (GSVA) score for each patient in the testing sets of bulk RNA-seq (MGSP, GSE91061, GSE181815). We then evaluated the ability of different methods in distinguishing responders and non-responders using the p-values of Wilcoxon tests on the GSVA scores. We observed
relatively better and stable performance of Pearson correlation compared to other similarity metrics in the three bulk datasets, as reflected by the smaller p-value when using Pearson correlation (Supplementary Figure 5). Spearman correlation and ES2 showed better performance than ES1 and MI. Taken together, these results indicate that Pearson correlation is a promising metric when integrating single-cell RNA-seq data with the bulk RNA-seq data.

We further evaluated whether one similarity metric was suitable for integrating single-cell RNA-seq data with the bulk RNA-seq data by generating a pseudo bulk RNA-seq dataset. To do it, we randomly shuffled the gene expression values of each sample in the real bulk dataset. We reason that this pseudo bulk data is unrelated to the corresponding scRNA-seq data, and cannot be used for prioritizing clinically relevant cell subsets and predictive signatures. We thus calculated the similarity between the pseudo bulk data and the single-cell data using the four metrics, including Pearson correlation, Spearman correlation, mutual information (MI) and Euclidean similarity 2 (ES2). We randomly generated 100 pseudo bulk datasets, and computed the mean values of similarity values given by each metric. For each metric, we computed the median similarity value and then compared it against the one computed using real bulk RNA-seq data (Supplementary Table 3). For Pearson correlation metric, we observed a clear positive correlation (liver cancer dataset: 0.3; melanoma dataset: 0.28) between the real bulk data and the scRNA-seq data, but almost no correlation (liver cancer dataset: -0.00057; melanoma dataset: -0.00042) between the pseudo bulk data and the scRNA-seq data. Similar results were observed for Spearman correlation. However, for the Euclidean-based metrics, we observed comparable similarity values (liver cancer dataset: 0.005 vs. 0.005, melanoma dataset: 0.006 vs. 0.003) when using real bulk data versus pseudo bulk data, suggesting a risk of Euclidean-based metrics in determining whether bulk RNA-seq data was correlated with scRNA-seq data. These results indicated that correlation-based metrics rather than Euclidean-based metrics can help to determine whether a bulk dataset was suitable to use for integrating scRNA-seq data and thus predicting clinically relevant cell subsets.
Supplementary Figures

Supplementary Figure 1. The bar plot showing the percentage of scAB+ cells over each cell type in the liver cancer single-cell dataset.

Supplementary Figure 2 UMAP visualization of fatty acid metabolic pattern based on the computed gene scores of cells in liver microenvironment.

Supplementary Figure 3 The violin plot showing the CD274 (i.e. PD-L1) expression in the scAB_IR macrophages vs. other macrophage cells.
Supplementary Figure 4 Overall survival curves of patients with low or high risks according to 23 KEGG pathways in training and test sets.
Supplementary Figure 5. Performance regarding different similarity strategies in the testing set. (A) Performance regarding different similarity strategies in the liver cancer dataset. (B) Performance regarding different similarity strategies in the melanoma dataset.

Supplementary Figure 6. Performance of scAB in the bulk testing set (ICGC-LIRI) with respect to different sizes of the bulk training samples (TCGA-LIHC).
### Supplementary Tables

**Supplementary Table 1.** The calculation of the relative score from survival information.

- $T(A)$ and $T(B)$ represents last observation time of sample A and B, respectively.
- $\tau_A$ and $\tau_B$ represents status of sample A and B, respectively.
- $r(t)$ is the proportion of survival patients estimated using the KM method at time $t$.

| Cases | Observation time | Status | Rank of samples |
|-------|------------------|--------|-----------------|
| Case1 | $T_A < T_B$      | $\tau_A = 1$, $\tau_B = 0$ | Score of A + 1, Score of B + 0 |
| Case2 | $T_A < T_B$      | $\tau_A = 1$, $\tau_B = 1$ | Score of A + 1, Score of B + 0 |
| Case3 | $T_A < T_B$      | $\tau_A = 0$, $\tau_B = 1$ | Score of A $+ \frac{r(T_A) - r(T_B)}{r(T_A)}$, Score of B $+ \frac{r(T_B)}{r(T_A)}$ |
| Case4 | $T_A < T_B$      | $\tau_A = 0$, $\tau_B = 0$ | Score of A $+ 0.5 \times \left(1 + \frac{r(T_A) - r(T_B)}{r(T_A)}\right)$, Score of B $+ 0.5 \times \left(1 - \frac{r(T_A) - r(T_B)}{r(T_A)}\right)$ |
| Case5 | $T_A = T_B$      | $\tau_A = \tau_B$ | Score of A + 0.5, Score of B + 0.5 |
| Case6 | $T_A = T_B$      | $\tau_A = 1$, $\tau_B = 0$ | Score of A + 1, Score of B + 0 |
**Supplementary Table 2** The correlation coefficients and p-values between the sensitivity IC50 value and GSVA score in 54 drugs.

| Drug              | Cor     | P-value  | adj.P-value |
|-------------------|---------|----------|-------------|
| ICL1100013        | 0.4274  | 0.001576 | 0.02942     |
| QS11              | 0.5638  | 2.01E-05 | 0.005197    |
| XMD15-27          | 0.3986  | 0.004142 | 0.041931    |
| Dactolisib        | 0.3324  | 0.002756 | 0.034291    |
| PLK_6522          | 0.4193  | 0.002434 | 0.033747    |
| Pelitinib         | 0.4806  | 0.000412 | 0.014888    |
| AZD7762           | 0.2378  | 0.005653 | 0.04953     |
| Etoposide         | 0.4413  | 0.001335 | 0.026009    |
| BIBF-1120         | 0.4958  | 0.000292 | 0.014888    |
| Pazopanib         | 0.3973  | 0.004699 | 0.044786    |
| OSU-03012         | 0.4871  | 0.000334 | 0.014888    |
| CP724714          | 0.4050  | 0.003524 | 0.040481    |
| Vorinostat        | 0.3139  | 0.003844 | 0.041931    |
| Sepantronium bromide | 0.3866 | 0.000514 | 0.014888    |
| Mitomycin-C       | 0.4983  | 0.000269 | 0.014888    |
| JNK-9L            | 0.3879  | 0.005889 | 0.049653    |
| Y-39983           | 0.3851  | 0.005749 | 0.04953     |
| GSK650394         | 0.4855  | 0.000352 | 0.014888    |
| BAY-61-3606       | 0.4474  | 0.001121 | 0.024968    |
| BMS-345541        | 0.3312  | 0.002689 | 0.034291    |
| Phenformin        | 0.4322  | 0.002158 | 0.031644    |
| PF-00299804       | 0.4554  | 0.001891 | 0.031057    |
| Vinorelbine       | 0.3307  | 0.002561 | 0.033747    |
| AZD6738           | 0.3208  | 0.003105 | 0.037592    |
| Doxorubicin       | 0.3936  | 4.64E-05 | 0.005197    |
| NPK76-II-72-1     | 0.4346  | 0.001813 | 0.031057    |
| Fulvestrant       | 0.2584  | 0.005723 | 0.04953     |
| Brivanib, BMS-540215 | 0.4024 | 0.004152 | 0.041931    |
| Mirin             | 0.3886  | 0.000479 | 0.014888    |
| Gemcitabine       | 0.2605  | 0.002551 | 0.033747    |
| Dacinostat        | 0.3858  | 0.005655 | 0.04953     |
| Epothilone B      | 0.4795  | 0.000426 | 0.014888    |
| AZD7969           | 0.4667  | 0.000556 | 0.014888    |
| Bleomycin         | 0.4591  | 0.00117  | 0.024968    |
| Thapsigargin      | 0.5117  | 0.000146 | 0.010888    |
| Daporinad         | 0.4914  | 2.79E-05 | 0.005197    |
| Vincristine       | 0.6347  | 0.00125  | 0.010888    |
| CDK9_5576         | 0.5428  | 0.001941 | 0.031057    |
| Oxaliplatin       | 0.4239  | 0.000432 | 0.014888    |
Supplementary Table 3 The median values of different similarity metrics in the real and pseudo bulk data.

| Dataset          | Pearson  | Spearman | MutualInfo | ES2    |
|------------------|----------|----------|------------|--------|
| liver cancer (real) | 0.30679  | 0.22903  | 0.02614    | 0.00466|
| liver cancer (pseudo) | -0.00057 | -0.00051 | 0.00124    | 0.0046 |
| melanoma (real)   | 0.28083  | 0.25327  | 0.01939    | 0.00602|
| melanoma (pseudo) | -0.00042 | -0.00007 | 0.0011     | 0.00343|