A streamlined method for signature score calculation

Yiming Yang, Bo Li

Sep 2020 (First Draft)
Mar 2025 (Last Update)

Signature score is a useful tool to study the activities of gene modules at the single-cell level. The conventional method calculates approximate signature scores by random sampling. Users need to carefully find a trade-off between accurate approximation (more sampling) and computational efficiency (less sampling). We instead propose a closed-form solution to compute exact signature scores, which achives both high accuracy and high efficiency by eliminating the requirement of the sampling step. In the following sections, we will describe the conventional method, give our closed-form method and compare the performance of the two methods using real data.

1 Conventional method

In this section, we describe the conventional method by following [1], which used a modified method from [2].

Assume that we have N cells and M genes. We denote the expression (e.g. $\log(TP100K+1)$) of gene i at cell j as e_{ij}. Then the average expression μ of each gene across N cells can be defined as:

$$\mu_i = \frac{1}{N}\sum_{j=1}^{N} e_{ij}.$$

We bin the M genes into n bins (e.g. $n = 50$) based on their average expressions (i.e. μs). We additionally assume that we have a gene signature S. S consists of K genes, with k_b genes in expression bin b:

$$S = \bigcup_{b=1}^{n} S_b, \qquad |S_b| = k_b, \qquad |S| = \sum_{b=1}^{n} k_b = K.$$

The signature score $\mathcal{S}$ is defined as the difference between the raw score $\mathcal{S}_{raw}$ and the control score $\mathcal{S}_{control}$, which we will define separately.

The **raw score** of cell j, $\mathcal{S}^j_{raw}$, is defined as follows:

$$\mathcal{S}^j_{raw} = \frac{1}{K}\sum_{i\in S} c_{ij}, \qquad c_{ij} = e_{ij} - \mu_i,$$

where c_{ij} is the centered expression. Using centered expression in the raw score helps prevent highly expressed genes from dominating the score.

The **control score** is useful to control technical noise that depends on gene abundance. To calculate this score, we first need to define *S-compatible* random signature. This is a set of K genes sampled without replacement from all M genes, such that there are exactly k_b genes in the set for each bin b. The score of random signature $\mathcal{S}_r$ on cell j is

$$\mathcal{S}^j_r = \frac{1}{K}\sum_{i\in S_r} c_{ij}.$$

We define the **control score** of cell j, $\mathcal{S}^j_{control}$, as the expectation of the random signature on cell j:

$$\mathcal{S}^j_{control} = \mathbb{E}[\mathcal{S}^j_r].$$

In [1], the expectation is approximated by randomly sampling L ($L = 1000$) *S-compatible* signatures:

$$\mathcal{S}^j_{control} = \mathbb{E}[\mathcal{S}^j_r] \approx \frac{1}{L}\sum_{l=1}^{L}\mathcal{S}^j_{rl}.$$

Because thee sampling process is time consuming, the *S-compatible* random signatures are not sampled independently for each cell j. Instead, L random signatures are first sampled and then applied for all N cells.

Once we have the raw and control scores, we can calculate the signature score of cell j, $\mathcal{S}^j$:

$$\mathcal{S}^j = \mathcal{S}^j_{raw} - \mathcal{S}^j_{control}.$$

2 Our streamlined, closed-form solution

After a careful inspection, we find that there is a closed-form solution for calculating the expectation. Let us first rewrite the random signature score $\mathcal{S}^j_r$ so that we can see the random variables clearly:

$$\mathcal{S}^j_r = \frac{1}{K}\sum_{i\in S_r} c_{ij} = \frac{1}{K}\sum_{b=1}^{n}\sum_{p=1}^{k_b} c_{s_{bp},j},$$

where s_{bp} is a random variable denoting the p-th sampled gene in bin b. Let us also define s_b as a random variable denoting one sampled gene in bin b.

Then the control score (expectation) becomes

$$\begin{aligned}\mathcal{S}^j_{control} = \mathbb{E}[\mathcal{S}^j_r] &= \mathbb{E}[\frac{1}{K}\sum_{b=1}^{n}\sum_{p=1}^{k_b} c_{s_{bp},j}] \\ &= \frac{1}{K}\sum_{b=1}^{n}\sum_{p=1}^{k_b}\mathbb{E}[c_{s_{bp},j}] \\ &= \frac{1}{K}\sum_{b=1}^{n} k_b\mathbb{E}[c_{s_{bp},j}].\end{aligned}$$

Note that in the above equations, we use the fact that $\mathbb{E}[c_{s_b,j}] = \mathbb{E}[c_{s_{b1},j}] = \mathbb{E}[c_{s_{bp},j}]$. For each random signature that $s_{bp} = v$, we can map it to a signature with $s_{b1} = v$ by swapping the 1st and the p-th genes. Thus we have a one-to-one mapping between random signatures with $s_{b1} = v$ and random signatures with $s_{bp} = v$. Thus we have $\mathbb{E}[c_{s_{b1},j}] = \mathbb{E}[c_{s_{bp}}, j]$.

$\mathbb{E}[c_{s_b}, j]$ can be easily calculated as

$$\mathbb{E}[c_{s_b,j}] = \frac{1}{[\frac{M}{n}]}\sum_{i\in\text{bin } b} c_{ij},$$

and we can precompute $\mathbb{E}[c_{s_b}, j]$ for all bins and all cells.

In conclusion, given a closed-form formula for computing the control score and precomputed $\mathbb{E}[c_{s_b}, j]$ terms, we can calculate any signature score instantly.

3 A statistical view of the signature score calculation

Let us consider this question: if the observed expression c_{ij} for gene i is specific to cell j (i.e. ultra high or ultra low)? To answer this question, we first need to have an expression distribution for non-specific expressions (null distribution). To construct a null distribution, we partition all genes into n bins as described above.

For each bin b, the emperical distribution consisting of $\{c_{ij}|i \in \text{bin } b\}$ captures non-specificity and technical artifacts in cell j. The sample mean $\hat{\mu}_{bj}$ and sample standard deviation $\hat{\sigma}_{bj}$ of this distribution are

$$\hat{\mu}_{bj} = \mathbb{E}[c_{s_b,j}] = \frac{1}{[\frac{M}{n}]} \sum_{i \in \text{bin } b} c_{ij},$$
$$\hat{\sigma}_{bj} = \sqrt{\frac{\sum_{i \in \text{bin } b} c_{ij}^2 - [\frac{M}{n}]\hat{\mu}_{bj}^2}{[\frac{M}{n}] - 1}}.$$

We use this distribution as our null distribution for genes $i \in$ bin b and we have

$$c_{ij} \sim \text{Dist}(\hat{\mu}_{bj}, \hat{\sigma}_{bj}^2).$$

Signature score as a weighted sum of standard scores We can rewrite the signature score $\mathcal{S}^j$ based on standard scores as follows:

$$\mathcal{S}^j = \frac{1}{K} \sum_{b=1}^{n} \sum_{i \in S_b} \hat{\sigma}_{bj} z_{ij}.$$

The above equation means the signature score we described previously is a standard deviation weighted sum of z scores. Since standard deviations in the null distributions might not represent any interesting biology, we can consider an unweighted version:

$$\mathcal{S}_{new}^j = \frac{1}{K} \sum_{i \in S} z_{ij}.$$

More importantly, by Lindeberg Central Limit Theorem (Linderberg's condition; needs to be checked), $\mathcal{S}_{new}^j \sim \mathcal{N}(0,1)$ asymptotically when $K \to \infty$.

Thus, we can calculate P-value for each cell independently based on $\mathcal{S}_{new}^j$ and control False Discovery Rate for a whole dataset.

4 Experiment results

We tested our closed-form solution (implemented in Pegasus) and conventional methods (implemented in SCANPY and Seurat, respectively) using the full bone marrow dataset (274,182 cells) from the Immune Cell Atlas project.

The benchmark platform has the following specifications:

- macOS 14.7.4 x86_64, with 12 vCPUs and 16 GB memory
- Python 3.11.8
- R 4.4.3
- Enforce single-thread computation by setting the corresponding environment variables.

We benchmark three implementations (Pegasus, SCANPY and Seurat), with versions specified in the following table:

	Seurat	SCANPY	Pegasus
Version	5.2.1	1.11.0	1.10.2
Release date	2025-01-24	2025-02-14	2025-03-27

Table 1: Software for benchmark

We first benchmarked the three implementations for calculating B cell signature scores on the bone marrow data. The B cell signature S = { CD19, MS4A1, CD79A, CD79B, BANK1, BLK, RALGPS2, ARHGAP24, AFF3, BCL11A }.

For all the three implementations, we set the number of bins $n = 50$. For SCANPY, we calculated B cell signature scores by ranging the number of sampled random signatures from $L = 50$ to $L = 1000$, with step size 50. For Seurat, we varied L from $L = 50$ to $L = 500$ with step size 50, as Seurat crashed for any $L \geq 550$. For Pegasus, we didn't standardize the scores to be consistent with the results of the other two software.

We then plotted the Spearman's rank correlation between Pegasus-calculated (non-standardized) scores and SCANPY or Seurat scores in Figure 1. We can observe from the plot that 1) both SCANPY and Seurat scores approached Pegasus scores when L increases; 2) Seurat scores have better correlations with Pegasus scores compared to SCANPY when L is fixed.

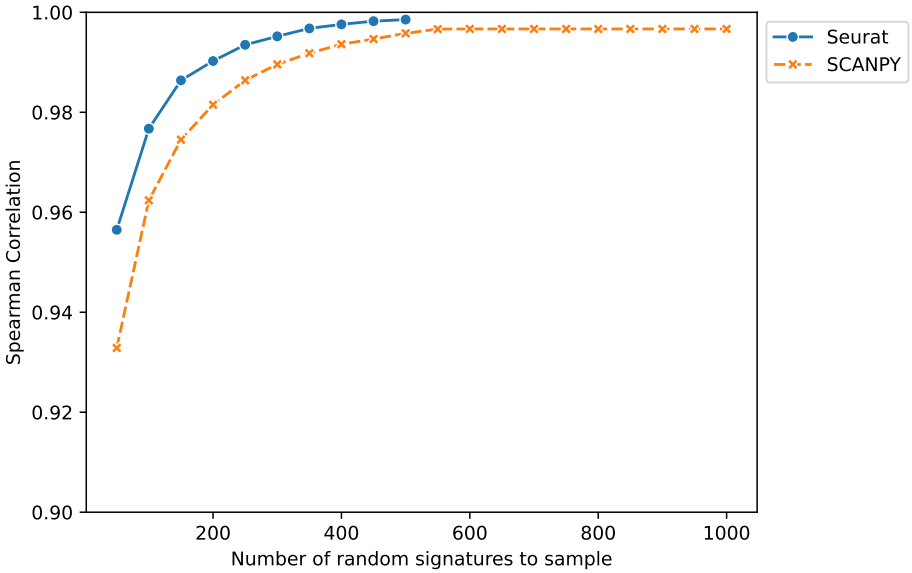


Figure 1. Spearman correlation of Pegasus (non-standardized) scores and SCANPY/Seurat scores under different signature sampling sizes L.

After that, we also benchmarked the three implementations with respect to computational efficiency. In this case, we ask each tool to calculate 5 signature scores: B cell, Plasma cell, $CD4^+$ T cell, $CD8^+$ T cell, and Natural Killer cell. All the tools still use 50 bins. For SCANPY and Seurat, we set $L = 100$. For Pegasus, we calculate the non-standardized scores to be consistent with the steps in SCANPY and Seurat. In addition, we ran each tool 10 times to estimate error bars. The execution time results are shown in Figure 2:

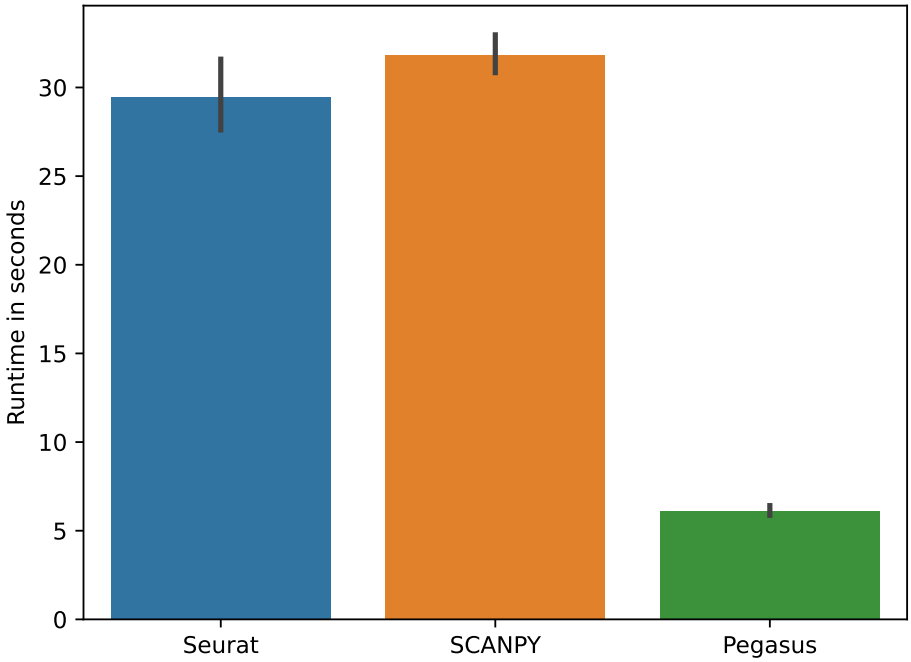


Figure 2. Bar plots showing the runtime in seconds for each tool to calculate 5 signature scores (B, Plasma, CD4$^+$ T, CD8$^+$ T, NK cells). Error bars were calculated from 10 independent runs.

For the rest part of this section, we use Pegasus standardized scores for benchmarking. In Figure 3, it's easy to see that Pegasus standardized scores' correlation with SCANPY/Seurat scores under different L sizes still roughly holds: the lower correlation coefficients are due to the Pegasus standardization process regarding mean expression of each gene respectively.

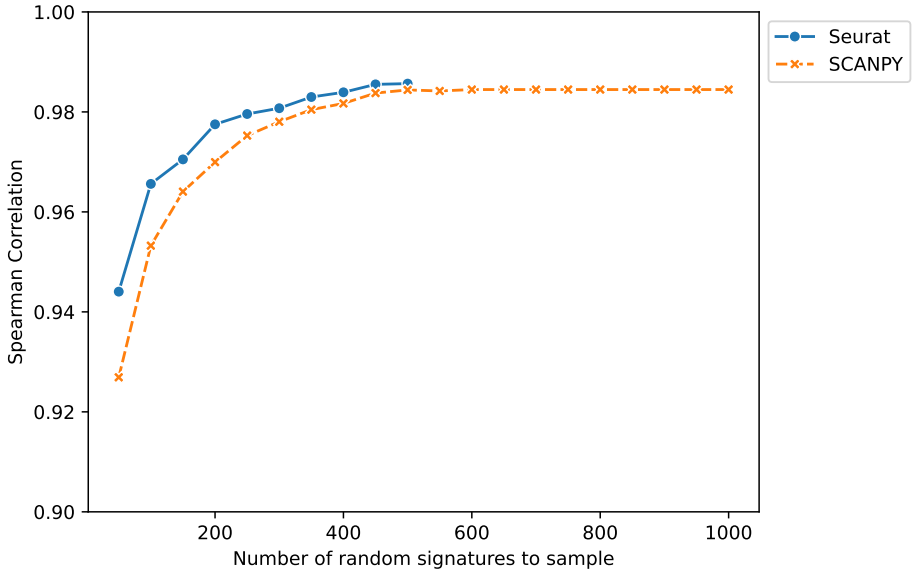
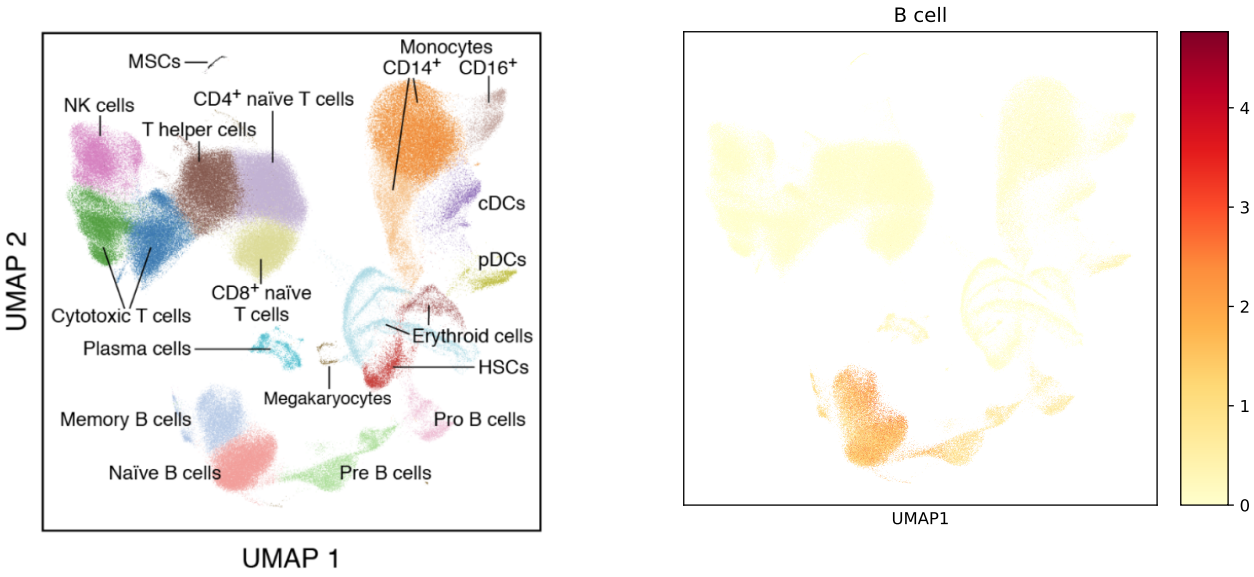


Figure 3. Spearman correlation of Pegasus scores and SCANPY/Seurat scores under different signature sampling sizes L.

Finally, the Pegasus B cell signature scores are shown in UMAP coordinates along with the cell type annotation for the dataset to validate the results (notice that the scores < 0 are clipped to 0 when plotting):



(a) Cell type annotation

(b) Pegasus B cell signature scores

Figure 4. Cell-type annotated bone marrow dataset and B cell signature scores in UMAP coordinates.

References

[1] M. S. Cuoco C. Rodman M. J. Su J. C. Melms R. Leeson A. Kanodia S. Mei J. R. Lin S. Wang B. Rabasha D. Liu G. Zhang C. Margolais O. Ashenberg P. A. Ott E. I. Buchbinder R. Haq F. S. Hodi G. M. Boland R. J. Sullivan D. T. Frederick B. Miao T. Moll K. T. Flaherty M. Herlyn R. W. Jenkins R. Thummalapalli M. S. Kowalczyk I. Can adas B. Schilling A. N. R. Cartwright A. M. Luoma S. Malu P. Hwu C. Bernatchez M. A. Forget D. A. Barbie A. K. Shalek I. Tirosh P. K. Sorger K. Wucherpfennig E. M. Van Allen D. Schadendorf B. E. Johnson A. Rotem O. Rozenblatt-Rosen L. A. Garraway C. H. Yoon B. Izar L. Jerby-Arnon, P. Shah and A. Regev. A cancer cell program promotes t cell exclusion and resistance to checkpoint blockade. *Cell*, 175(4):984–997, 2018.

[2] S. M. Prakadan M. H. n. Wadsworth D. Treacy J. J. Trombetta A. Rotem C. Rod-man C. Lian G. Murphy M. Fallahi-Sichani K. Dutton-Regester J. R. Lin O. Cohen P. Shah D. Lu A. S. Genshaft T. K. Hughes C. G. Ziegler S. W. Kazer A. Gaillard K. E. Kolb A. C. Villani C. M. Johannessen A. Y. Andreev E. M. Van Allen M. Bertagnolli P. K. Sorger R. J. Sullivan K. T. Fla-herty D. T. Frederick J. Jan e-Valbuena C. H. Yoon O. Rozenblatt-Rosen A. K. Shalek A. Regev I. Tirosh, B. Izar and L. A. Garraway. Dissecting the multicellular ecosystem of metastatic melanoma by single-cell rna-seq. *Science*, 352(6282):189–196, 2016.