A streamlined method for signature score calculation

Bo Li

1 Background

Signature score is a useful tool to study the activities of gene modules at the single-cell level. In this section, we describe the current common practice 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:

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

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:

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

where c_{ij} is the centered expression. Using centered expression in the raw score helps to 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 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 the sampling process is time consuming, the *S-compatible* random signatures are not sampled independently for each cell j. Instead, L random signatures are first samples 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 A streamlined method

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 pth sampled gene in bin b.

Then the control score (expectation) becomes

$$\begin{aligned}\mathcal{S}^j_{control} = \mathbb{E}[\mathcal{S}^j_r] &= \mathbb{E}[\frac{1}{K}\sum_{b=1}^{n}\sum_{p=1}^{k_b} c_{s_{bp},j}] \\ &= \frac{1}{K}\sum_{b=1}^{n}\sum_{p=1}^{k_b} \mathbb{E}[c_{s_{bp},j}] \\ &= \frac{1}{K}\sum_{b=1}^{n} k_b\mathbb{E}[c_{s_{b1},j}].\end{aligned}$$

Note that in the above equations, we use the fact that $\mathbb{E}[c_{s_{b1},j}] = \mathbb{E}[c_{s_{bp},j}]$, which can be proved as follows. 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 pth 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 $\mathbb{E}[c_{s_{b1},j}] = \mathbb{E}[c_{s_{bp},j}]$.

$\mathbb{E}[c_{s_{b1},j}]$ can be easily calculated as

$$\mathbb{E}[c_{s_{b1},j}] = \frac{1}{\left\lceil\frac{M}{n}\right\rceil}\sum_{i\in \text{bin } b} c_{ij},$$

and we can precompute $\mathbb{E}[c_{s_{b1},j}]$ for all bines and all cells.

In conclusion, given a closed-form formula for computing the control score and precomputed $\mathbb{E}[c_{s_{b1},j}]$ terms, we can calculate any signature score instantly.

References

[1] L. Jerby-Arnon, P. Shah, 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. Cañadas, 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, and A. Regev. A cancer cell program promotes t cell exclusion and resistance to checkpoint blockade. *Cell*, 175(4):984–997, 2018.

[2] I. Tirosh, B. Izar, S. M. Prakadan, M. H. n. Wadsworth, D. Treacy, J. J. Trombetta, A. Rotem, C. Rodman, 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. Flaherty, D. T. Frederick, J. Jané-Valbuena, C. H. Yoon, O. Rozenblatt-Rosen, A. K. Shalek, A. Regev, and L. A. Garraway. Dissecting the multicellular ecosystem of metastatic melanoma by single-cell rna-seq. *Science*, 352(6282):189–196, 2016.