

Analysing single-cell bisulfite sequencing data with scbs

Lukas P. M. Kremer, Leonie Küchenhoff, ..., Simon Anders
University of Heidelberg, Germany

March 2022

Comparison

To demonstrate an improvement of results, we compare previously published analyses described above with our analysis method. First, we use a dataset by [Argelaguet et al., 2019](#). In their work, they generated a single cell methylome dataset of mouse embryo cells across different stages of development (E4.5 – E7.5). They calculated the correlation between RNA expression intensity and corresponding DNA methylation at the promoter region to elucidate the implications of the epigenome on RNA expression during development. As promoters, they used DNA regions of $\pm 2\text{kb}$ around transcriptions start sites (TSS).

–The following paragraph potentially needs to be changed when new plots are made–

In our analysis, we compare correlations between RNA expression intensity and DNA methylation of these promoter regions to correlations between RNA expression intensity and DNA methylation of the closest variable methylation regions that we identified with scbs. In figure 1, we show that correlation of RNA expression intensity to gene methylation around the TSS $\pm 2\text{kb}$ is less strong compared to the correlation at the closest variable region. This effect can be observed with methylation fractions (figure 1A) or shrunken residuals (figure 1B) as data inputs.

As an example to illustrate this effect more clearly, we picked a gene with significantly stronger correlation at the closest variable region than at the TSS. This gene is marked on figure 1A and B with a yellow dot. We divided all cells into four groups of cells with no, low, medium and high expression of this exemplary gene Ube216 and plotted a smoothed methylation level per group. As shown in figure 1C, the promoter region is sparsely methylated in all four RNA expression groups and is therefore a poor region to make conclusions about the relation of DNA methylation and RNA expression intensity. However, a region close to the TSS, identified as the closest highly variable methylation region with scbs, shows a much clearer negative correlation between RNA expression intensity and DNA methylation.

To further demonstrate the benefit of detecting variable methylation regions, we used a second dataset, published by [Luo et al., 2017](#). In their work, [Luo et al., 2017](#) generated methylomes from single cell neuronal nuclei in mice to identify genomic regulatory elements across neuron types. To compare how well cell types are separated by different analysis techniques, we processed the same dataset into 2-dimensional UMAPs several times. Subsequently, we compared the results by assigning a score

to each output based on cell type separation in the final 15 principal components. We see an improvement of cell separation when comparing the analysis of variable methylation regions to 100 kb regions (2A and C). The difference of using shrunken residuals instead of methylation fractions to assess the methylation per region is not as significant. However, when emulating a sparse dataset with less cells by randomly reducing the dataset to 500 cells, it becomes clear that using shrunken residuals over methylation fractions also improves the outcome of our analysis (2B and D).

Discussion and Conclusion

...

Methods

Data processing

To analyze CpG methylation data by [Luo et al., 2017](#), the mouse genome was first separated into regions. Regions were either obtained by tiling the genome into non-overlapping 100 kb bins or by acquiring the most variable methylation regions with the scbs script as described above. For each region, the mean methylation fraction or methylation residual was calculated. Each region that contained less than four measured CpGs or that was measured in less than 10% of cells was removed from the dataset. To select n regions for further analysis, top n regions were picked based on the cell number the region was measured in. For dimensionality reduction, missing values were imputed with an iterative PCA as described below. To visualize data, a UMAP was performed on the final 15 principal components.

Scoring

To assess data processing performance of several DNA methylation analysis techniques, we processed single cell DNA methylation data with several approaches and scored each result. Scores evaluate the ability of each method to separate cell types in a 15-dimensional space of the first 15 principal components based on their CpG methylation. Each score varies between 0 and 1, where higher scores reflect better data processing performance.

Nearest Neighbor Count Score

The nearest neighbor count score $nnc(k)$ is based on the $\Gamma - \text{Score}$ published by [Kireeva et al., 2014](#). Each cell c_j^i

belongs to a cell type j . For each cell, k nearest neighbors nn_j^k were computed based on their Euclidean Distance. Based on the cell type j of the k nearest neighbors of each cell, the nnc score $nnc(k)$ was computed with fomula (1), in which $d(c_j^i, k_j^i) = 1$ if at least 90% of neighbor classes match the class of cell c_j^i and $d(c_j^i, k_j^i) = 0$ if less than 90% of classes match, while n are the total number of cells.

$$nnc(k) = \frac{1}{n}d(c_j^i, k_j^i) \quad (1)$$

References

- Argelaguet R, Clark SJ, Mohammed H, Stapel LC, Krueger C, Kapourani Ca, Imaz-rosshandler I, Lohoff T, Xiang Y, Hanna CW, et al.. 2019. Multi-omics profiling of mouse gastrulation at single-cell resolution. *Nature* **576**: 487–491.
- Kireeva NV, Ovchinnikova SI, Tetko IV, Asiri AM, Balakin KV, and Tsivadze AY. 2014. Nonlinear dimensionality reduction for visualizing toxicity data: Distance-based versus topology-based approaches. *ChemMedChem* **9**: 1047–1059.
- Luo C, Keown CL, Kurihara L, Zhou J, He Y, Li J, Castanon R, Lucero J, Nery JR, Sandoval JP, et al.. 2017. Single-cell methylomes identify neuronal subtypes and regulatory elements in mammalian cortex **604**: 600–604.