Genome Analysis

SCAN-ATAC Sim: a scalable and efficient method to simulate single-cell ATAC-seq from bulk-tissue experiments

Zhanlin Chen[1,2†], Jing Zhang[1,* †] , Jason Liu[1], Zixuan Zhang[2], Jiangqi Zhu[2], Min Xu[4], and Mark Gerstein [1,2,*]

[1]Department of Molecular Biophysics and Biochemistry

[2]Department of Computer Science, Yale University, New Haven, CT 06520, USA

[3]International School, Beijing University of Posts and Telecommunications

[4]Computational Biology Department, Carnegie Mellon University, Pittsburgh, PA 15213, USA

*To whom correspondence should be addressed.

†Authors make equal contributions

Abstract

Summary: Single-cell ATAC-seq (scATAC-seq) is a powerful approach that characterizes cell-type-specific regulatory landscapes. However, it is difficult to benchmark the performance of various scATAC-seq analysis tools (such as clustering and deconvolution) without knowing beforehand a set of gold-standard cell types. To simulate scATAC-seq experiments with known cell type labels, we introduce an efficient and scalable scATAC-seq simulation method (SCAN-ATAC-Sim) that down-samples bulk-tissue ATAC-seq (e.g. from representative cell lines) in an organized fashion. Our simulation protocol recreates the homogeneous signal-to-noise ratio in a single scATAC-seq experiment by integrating the different amounts of background noise in separate bulk-tissue experiments and independently samples twice without replacement to account for the diploid genome. Our implementation in C++ allows millions of cells to be simulated under an hour on a laptop computer, as it uses an efficient weighted reservoir sampling algorithm and is highly parallelizable with OpenMP.

Availability: SCAN-ATAC Sim is available at scan-atac-sim.gersteinlab.org.

Contact: pi@gersteinlab.org

1 Introduction

High-resolution single-cell Assay for Transposase-Accessible Chromatin using sequencing (scATAC-seq) techniques reveal transcriptional landscapes in a cell-type-specific manner. More specifically, scATAC-seq experiments classify individual cells according to their unique regulatory landscape. Many scATAC-seq data analysis approaches -- for instance, for calling and defining the active regions in various cell types, clustering, and deconvolution -- have been published (Bravo Gonzalez-Blas, et al., 2019; Fang, et al., 2019; Schep, et al., 2017; Xiong, et al., 2019; Zamanighomi, et al., 2018). However, it has been difficult to evaluate their efficacy since we do not have *a priori* knowledge of gold-standard cell types. For instance, how do we know if a clustering method really works correctly if we do not have *a priori* knowledge of the cell types and their fractional composition beforehand? Thus, simulation from known cell types can be a valuable component to evaluation and benchmarking of different tools.

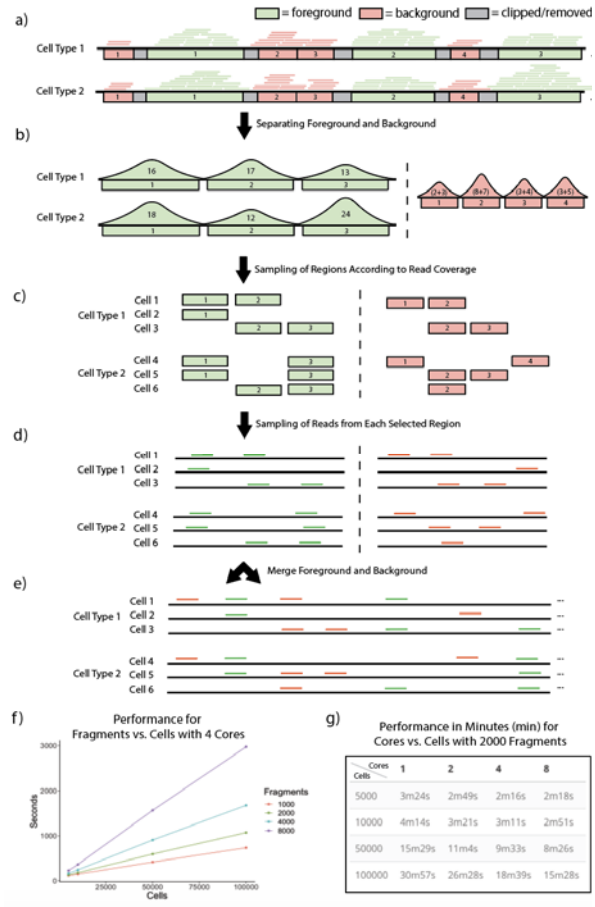
Currently, there are two different approaches to construct scATAC-seq experiments from bulk-tissue ATAC-seq, and both approaches are limited in accurately representing scATAC-seq. The first approach is to randomly and directly sample reads from bulk-tissue ATAC-seq to represent cells from a particular cell type (e.g. sampling from a cell line or FACS separated cells), then pool cells from various tissue/cell lines at a pre-defined proportion to mimic the mixture of cell types from bulk-tissue experiments (Fang, et al., 2019). However, ATAC-seq reads usually come from both peak foreground regions and random genomic background regions, and the signal-to-noise ratio usually varies dramatically across experiments. In contrast, cells of various types in a scATAC-seq experiment undergoes similar procedure and treatments, resulting in a similar signal-to-noise ratio. A naïve down-sampling strategy could introduce cell-type-specific background reads, which may confound down-stream analyses. Moreover, the human genome is diploid. Direct down-sampling could extract more than two reads within an overlapping genomic locus and thus falsely represent the nature of scATAC-seq experiments and further confound the data analysis techniques.

The second class of methods simulate individual cells by selecting a maximum of two peaks from bulk-tissue ATAC-seq and strictly controlling the signal-to-noise ratio (Chen, et al., 2019; Xiong, et al., 2019). However, these methods do not sample reads and only sample from the merged peaks, which is only a subset of the whole genome. In reality, the peak and background regions are indistinguishable before efficient clustering. Furthermore, simulating representative peak regions for one cell requires repeating the binomial number generator for all merged peak

regions (up to hundreds of thousands), making it computationally expensive. Hence, it is difficult to realistically scale up to millions of cells for performance benchmark on large-scale single-cell data. There is a need to accurately simulate scATAC-seq experiments in an efficient and scalable way for scATAC-seq analysis.

The lack of a standard, representational synthetic scATAC-seq dataset motivated our SCAN-ATAC Sim, which takes previous factors into consideration when reconstructing scATAC-seq data. SCAN-ATAC Sim generates a representative and evaluative dataset from bulk-tissue ATAC-seq for benchmarking different scATAC-seq analysis techniques with an efficient reservoir sampling algorithm and parallelization. In the following sections, we describe SCAN-ATAC Sim's implementation and analyze its computational complexity.

Figure 1. a) The bulk-tissue ATAC-seq reads are partitioned into foreground and background based on overlap with merged peaks. The number in each region indicate the read coverage. b) The cell-type-specific foreground reads are separated, and a unified background is created by combining background reads across all cell lines. c) The regions are sampled without replacement with read coverage as weights for foreground and background. d) Reads are sampled from the selected regions using a uniform distribution. e) Sampled foreground and background reads are combined to simulate individual cells. f) Performance for region number vs cell number with 4 cores. g) Performance for cores vs cell number with regions.



2 Methods

The SCAN-ATAC Sim consists of two main steps: data preprocessing and single-cell simulation. The process starts from aligned bam files of bulk-tissue ATAC-seq for a desired cell type group (Fig. 1a). The data preprocessing step defines a cell-type-specific foreground from the merged peaks and a unified background (Fig. 1b). For each cell, the single-cell simulation step samples the foreground and background regions twice without replacement with the probability proportional to the read coverage (Fig. 1c) and randomly selects one read for each sampled region (Fig. 1d). Then, reads from both foreground and background are combined to form reads covering one cell (Fig. 1e). This single-cell simulation step is then repeated for a large number of cells as specified by the user-provided parameter.

2.1 Input and Output

The input is aligned bulk-tissue ATAC-seq bam files of a cell group consisting of either cell lines or other representations targeted for scATAC-seq format, and the outputs are sampled reads for each cell based on user-provided parameters such as the number of reads per cell, total cell number, and signal-to-noise ratio. We also provide the regions from which the reads are selected as an intermediate output (Fig. 1c).

2.2 Data Preprocessing

One of the main goals for data preprocessing is to remove the variance in signal-to-noise ratios across cell line ATAC-seq and create a unified background with the same signal-to-noise ratio for all sampled cells represented in scATAC-seq. The bulk-tissue ATAC-seq peaks for all cell lines are merged to define the foreground regions. The background regions are defined as the complement of the 1kb-extended foreground divided in bins of fixed sizes (i.e. 1kb). Paired and deduplicated reads from each cell type intersecting a foreground region are defined as foreground reads, whereas the same paired and deduplicated reads from all cell types intersecting with a background region are defined as background reads. Only the midpoint of background reads is used in counting read coverage to prevent coverage of across multiple regions. The foreground reads are cell-line-specific, and the unified background reads are created by combining background reads from all cell lines (Fig. 1b), which more realistically resembles scATAC-seq. This step yields two sets of information required for the simulation step: the read counts associated with each foreground and background region and the reads overlapping those regions.

2.3 Single-Cell Simulation

For scATAC-seq in an individual cell, the reads are from both the foreground (peaks) and the background (non-peaks). First, let and represent the number of foreground and background reads in a cell of cell type , such that the total number of reads in an individual cell is . We assumed that and could vary across cells in scATAC-seq, so users could provide the mean and variance for a log-normal distribution of in each cell.

The ratio of foreground versus background sampled reads affects the uniqueness and separability between simulated cells. More foreground reads increases the separability between cells. We have designated a signal-to-noise ratio to conduct a custom allocation of foreground and background reads (1). Signal-to-noise ratio of simulated cells is unified across cells and could be adjusted to test the efficacy of data analysis methods.

(1)

Once an allocation between foreground and background reads is made, the number of representative regions are further divided evenly to mimic reads coming from maternal (,) and paternal (,) alleles. Hence, for any given region in the genome, we can guarantee that there are at most two reads sampled.

SCAN-ATAC Sim down-samples reads to simulate each cell in scATAC-seq. First, and regions are separately and randomly selected without replacement for both foreground and background, using

read coverage from its corresponding bam files as weights (Fig. 1c). Then, $N_{c,P}^{F}$ and $N_{c,P}^{B}$ regions are independently sampled again without replacement under the assumption that the two allele observations are independent events (Zhang, et al., 2015). Lastly, one read is randomly sampled form each selected region using a uniform distribution, resulting in a total of N_c reads (Fig. 1d).

It is worth noting that we used cell-type-specific foreground sampling probability to sample the foreground regions but non-cell-type-specific background probabilities for the background regions to mimic similar random cut probabilities for all cells in one scATAC-seq experiment. Specifically, if $r_{i,c}^{F}$ represents the total number of reads in the ith foreground region in cell type c, then the probability to sample region i can be calculated by:

$$p_{i,c}^{F} = \frac{r_{i,c}^{F}}{\sum_i r_{i,c}^{F}} \quad (2).$$

For the background region, we merge all the reads from different experiments and calculated a uniform background sampling probability. Specifically, $r_{i,c}^{B}$ represents the read count in the ith background region in cell type c, then the uniform background sampling rate for the ith background region is $p_i^{B} = \sum_c r_{i,c}^{B} / \sum_i \sum_c r_{i,c}^{B}$.

Lastly, the foreground and background reads are merged for a total of $N_c^{F} + N_c^{B} = N_c$ reads in one cell (Fig. 1e). The process is repeated for a large number of times to massively simulate each cell.

3 Results

There are two major computational challenges to simulating scATAC-seq data. First, there are up to millions of representative foreground and background regions to be selected with varying probabilities without replacement. Second, one experiment generally contains tens of thousands of cells. To address these challenges, we used two techniques to implement a highly efficient and scalable software for the sampling procedures mentioned in section 2.3.

To improve the efficiency of sampling one cell, we implemented a fast reservoir sampling algorithm to select the representative regions. Specifically, let n represents the number of regions and m represents the number of regions to be selected, traditional weighted sampling without replacement is $O(n \cdot m)$ (van der Walt, et al., 2011). Weighted reservoir sampling without replacement could perform weighted sampling without replacement in $O(n) + O(m \cdot log(n/m))O(logm)$ (Efraimidis and Spirakis, 2006). Especially in our case where $n \gg m$, weighted reservoir sampling approaches $O(n)$. Moreover, weighted reservoir sampling with exponential jumps would only require $O(m \cdot log(n/m)$ random numbers to be generated instead of $O(n)$, further improving the runtime of the algorithm.

Since the sampling of one cell is an independent event, the simulation steps can be efficiently parallelized in a cell-wise fashion with shared-memory multi-core parallelism using OpenMP (Dagum and Menon, 1998). Multi-core parallelism permits sampling multiple cells simultaneously across multiple cores, which offers a scalable improvement in runtime. Furthermore, shared-memory parallelism shares one copy of regions to reads mapping, which also improves runtime by reducing IO costs.

By utilizing both methods, we optimized the sampling process such that it could sample millions of cells under an hour (Fig. 1f, Fig. 1g).

We demonstrate that SCAN-ATAC Sim achieves a scalable speed up with a cell group of five cell types. The performance corresponds linearly with both cell number and region number N_c (Fig. 1f, Fig. 1g). Furthermore, parameters such as the signal-to-noise ratio ρ does not affect the runtime of the simulation. Overhead from shared-memory communication prevents further speed up within individual cell types, most likely due to IO limitations. Dividing the cell type into separate batches or splitting parallelism across cell types minimizes the overhead and resolves the IO limitations.

4 Discussion

Here, we introduce a new software tool for constructing a labeled scATAC-seq dataset from bulk-tissue ATAC-seq via down-sampling. The tool is pragmatic and useful in evaluating the efficacy of single-cell data analysis techniques by simulating scATAC-seq while adhering to various biological constraints. The acceleration via multi-core parallelism permits simulation of millions of cells under an hour. The tool is highly scalable and offers a space-time tradeoff to match the rate of growth in the number of cells sequenced in single-cell ATAC-seq.

Funding

This work was supported by the National Institutes of Health [grant number U01 MH 116492]

Conflict of Interest: none declared.

References

Bravo Gonzalez-Blas, C., *et al.* cisTopic: cis-regulatory topic modeling on single-cell ATAC-seq data. *Nat Methods* 2019;16(5):397-400.

Chen, H., *et al.* Assessment of computational methods for the analysis of single-cell ATAC-seq data. *Genome Biol* 2019;20(1):241.

Dagum, L. and Menon, R. OpenMP: An industry standard API for shared-memory programming. *Ieee Comput Sci Eng* 1998;5(1):46-55.

Efraimidis, P.S. and Spirakis, P.G. Weighted random sampling with a reservoir. *Information Processing Letters* 2006;97(5):181-185.

Fang, R., *et al.* Fast and Accurate Clustering of Single Cell Epigenomes Reveals Cis-Regulatory Elements in Rare Cell Types. *bioRxiv 615179 [**Preprint**]* 2019.

Schep, A.N., *et al.* chromVAR: inferring transcription-factor-associated accessibility from single-cell epigenomic data. *Nat Methods* 2017;14(10):975-978.

van der Walt, S., Colbert, S.C. and Varoquaux, G. The NumPy Array: A Structure for Efficient Numerical Computation. *Computing in Science & Engineering* 2011;13:8.

Xiong, L., *et al.* SCALE method for single-cell ATAC-seq analysis via latent feature extraction. *Nat Commun* 2019;10(1):4576.

Zamanighomi, M., *et al.* Unsupervised clustering and epigenetic classification of single cells. *Nat Commun* 2018;9(1):2410.

Zhang, C.Z., *et al.* Calibrating genomic and allelic coverage bias in single-cell sequencing. *Nat Commun* 2015;6:6822.