scGAD: single-cell gene associating domain scores for exploratory analysis of scHi-C data

Siqi Shen¹, Ye Zheng²,* and Sündüz Keleṣ¹,3,*

¹Department of Biostatistics and Medical Informatics, University of Wisconsin - Madison, Madison, WI 53706, USA, ²Biostatistics, Bioinformatics and Epidemiology Program, Vaccine and Infectious Disease Division, Fred Hutchinson Cancer Research Center, Seattle, WA 98109, USA, ³Department of Statistics, University of Wisconsin - Madison, Madison, WI 53706, USA. * Corresponding authors.

Abstract

Summary: Quantitative tools are needed to leverage the unprecedented resolution of single-cell high-throughput chromatin conformation (scHi-C) data and to integrate it with other single-cell data modalities. We present single-cell gene associating domain (scGAD) scores as a dimension reduction and exploratory analysis tool for scHi-C data. scGAD enables summarization at the gene level while accounting for inherent gene-level genomic biases. Low-dimensional projections with scGAD capture clustering of cells based on their 3D structures. scGAD enables identifying genes with significant chromatin interactions within and between cell types. We further show that scGAD facilitates the integration of scHi-C data with other single-cell data modalities by enabling its projection onto reference low-dimensional embeddings.

Availability: scGAD is part of the BandNorm R package at https://sshen82.github.io/BandNorm/articles/scGAD-tutorial.html.

Contact: keles@stat.wisc.edu, yzheng23@fredhutch.org

Supplementary information: Supplementary data are available at Bioinformatics online.

1 Introduction

Single-cell technologies that profile chromatin conformation at the single-cell level emerged as promising approaches for high-resolution 3D genome characterization (Nagano et al., 2013; Stevens et al., 2017; Ramani et al., 2017; Li et al., 2019; Lee et al., 2019; Tan et al., 2021). Tools for specific scHi-C data inference tasks are gradually emerging (e.g., scHiCluster (Zhou et al., 2019), scHiTopics (Kim et al., 2020), Higashi (Zhang et al., 2021), BandNorm and 3DVI (Zheng et al., 2021) for imputation and normalization of the sparse scHi-C data to advance de facto downstream analysis; SnapHiC (Yu et al., 2020) for chromatin loop detection; scHiTools (Li et al., 2021) for quantifying cell-cell similarities). However, computational tools for extracting salient features of scHi-C data for integration with other single-cell data modalities, such as transcriptomics and epigenomics, are lacking. Here, we generalized the concept of gene-body associating domain (GAD) for bulk Hi-C data (Nagano et al., 2013, 2014; Stevens et al., 2017, 2019; Tan et al., 2019) to investigate whether they can accommodate inherent gene clusters, leading to overestimation of the background (Supplementary Figs. 2-3). We first considered four variations of scGAD (Supplementary Figs. 2-3), scGAD_regression first calibrates the sequencing depths of the cells and utilizes a Generalized Additive Model (GAM) to adjust for well known genomic biases, i.e., gene length, mappability, and GC content, with non-parametric smooth functions $s_1$, $s_2$, $s_3$. Specifically, we have $	ilde{R}_{ij} = \frac{R_{ij}}{\sum_{i=1}^{N} R_{ij}}$.

and $scGAD_{regression}$ is the residual from this model. In contrast, scGAD_global removes potential gene-level biases implicitly by a standardization approach (Fig. 1A):

$$scGAD_{global} = \frac{\tilde{R}_{ij} - mean_i}{\sqrt{\frac{1}{N-1} \sum_{j=1}^{N} (R_{ij} - mean_i)^2}}$$

where $mean_i = \frac{1}{N} \sum_{j=1}^{N} \tilde{R}_{ij}$. Note that gene-wise standardization of the residuals from the GAM model is operationally equivalent to standardization of $\tilde{R}_{ij}$, namely $scGAD_{global}$. Hence, we kept $scGAD_{regression}$ as unstandardized. We evaluated these variations for their performance in cell clustering and revealing relationships between the cell types. These comparisons revealed that $scGAD_{global}$ performs the best by separating the cell types (Fig. 1B), almost as good as the full Hi-C contact matrices (Zheng et al., 2021), and recovers the known cell-type relationships (Supplementary Figs. 4-5). We set $scGAD_{global}$ as the formal definition of the scGAD scores for all the downstream analyses.
3 Result

scGAD: genes with abundant interactions

Recent studies revealed that GAD formation is a chromatin feature of highly expressed genes (Zhang et al., 2020). We evaluated whether this feature is salient in the single-cell data. The correlations between the scGAD scores and expressions of genes were markedly higher, especially for the scRNA-seq marker genes, when both quantities were quantified in the same cell type compared to those in different cell types (p-value < 10^{-26}; Supplementary Fig. 6). Next, we developed a permutation strategy to detect genes with significantly high scGAD scores as a means to infer highly expressed genes in a given cell type (Supplementary Note). When we evaluated genes with significantly high scGAD scores with the corresponding scRNA-seq data, they showed significantly higher expression levels compared to the genes with insignificant scGAD scores (Fig. 1C and Supplementary Fig. 7), illustrating how genes with high scGAD scores are highly expressed.

scGAD: marker genes of cell types

We observed that cell-type-specific marker genes defined from scRNA-seq data displayed elevated scGAD scores (Supplementary Fig. 8). Furthermore, the average scGAD scores of these marker genes revealed cell-type-specific patterns (Fig. 1D and Supplementary Fig. 9). Leveraging the same marker gene detection procedure as in scRNA-seq analysis, we identified marker genes from scHi-C data with scGAD scores. We found that a large proportion of scRNA-seq marker genes overlapped with the scGAD marker genes (Fig. 1E), and the top scGAD significant marker genes yielded cell-type-specific gene expression patterns (Supplementary Fig. 10).

scGAD: projection onto reference low-dimensional embeddings

Finally, we asked whether scGAD scores can be exploited to project cells from scHi-C data onto a given reference low-dimensional embedding (e.g., from scRNA-seq; Fig. 1F). Projection onto the scRNA-seq embedding from the same system (i.e., with the exact same cell types) revealed that the cells originating from the same cell type but quantified by different data modalities were tightly clustered. Next, taking advantage of the Paired-Tag data, which included a larger number of cell types than the scHi-C data, we observed that scGAD facilitated an accurate projection of cells onto this larger space (Supplementary Figs. 11-12). This across-modality projection enables fast cell-type annotation for 3D genomics data and promotes integrative analyses of 3D genome structure, epigenomics, and transcriptomics to decipher gene regulation mechanisms at single-cell resolution.

In summary, scGAD provides a set of analysis tools to address the pressing needs for integrating scHi-C data with other single-cell data modalities.

Funding

This work was supported by NIH grants HG003747 and HG011371 to SK.

References


