2023 IEEE International Conference on Bioinformatics and Biomedicine (BIBM) | 979-8-3503-3748-8/23/$31.00 ©2023 IEEE | DOI: 10.1109/BIBM58861.2023.10385540

PLNorm: A Power-Law Distribution-Based Normalization Method for Single-Cell Hi-C Data Analysis

1st Bin Zhao
Department of Computer Science, Department of Statistics
North Dakota State University
Fargo, United States
bin.zhao@ndsu.edu

2nd Lu Liu
Department of Computer Science
North Dakota State University
Fargo, United States
lu.liu.2@ndsu.edu

***Abstract*—Similar to bulk Hi-C data, the frequency distribution of single-cell Hi-C data also adheres to a power law distribution in relation to the genomic distance between chromatin interaction endpoints. In light of this, we introduce an innovative normalization approach for single-cell Hi-C data that capitalizes on this power-law distribution. An extensive comparative study, employing three publicly accessible single-cell Hi-C datasets, underscores the robustness of PLNorm, critically evaluated against established normalization techniques including BandNorm, scVI-3D, and scHiCNorm. A diverse range of metrics, including changes in cell similarities, cell embeddings, and scalability, were utilized in the assessment. The results highlight the distinct advantages of PLNorm: it not only mitigates biases but also adeptly preserves cell-type information, enhances the precision of clustering outcomes, and demonstrates impressive scalability, making it a prime choice for large-scale data analysis. PLNorm is available at https://github.com/bignetworks2019/PLNorm/.**

***Index Terms*—PLNorm, power law, scHi-C, normalization, scalability**

I. Introduction

The evolution of chromosome conformation capture (3C)-based technologies [1]–[5], including Hi-C and related methods [6], has provided crucial insights into 3D genome organization inside the nucleus. Technologies like Hi-C have revealed higher-order chromatin structures including A/B compartments [6], topologically associating domains (TADs) [7], and chromatin loops [8]. These 3D features across different scales are interconnected with vital genome functions like gene transcription and DNA replication [9], [10]. The emergence of single-cell Hi-C (scHi-C) assays [11]–[16] has allowed 3D genome structures to be probed at single-cell resolution. Pioneering scHi-C studies have revealed cell-to-cell variability in chromatin structures [11], [17], suggesting functional implications. However, substantial computational challenges remain due to the high-dimensional yet sparse nature of scHi-C data.

The sparsity and noise characteristic of single-cell data can obscure important biological signals within and between cells. Technical biases related to fragment length, GC content, and mappability also permeate scHi-C experiments [18], [19]. As the number of cells and resolution increases, the complexity intensifies. The surge in data points strains computational requirements for both memory and time.

Normalization methods for bulk Hi-C data can be broadly classified into two categories: explicit factor correction and matrix balancing methods. Explicit factor correction methods, such as HiCNorm [20] and Hi-Corrector [21], necessitate prior knowledge of Hi-C systematic biases and use this information as input for normalization procedures.

On the contrary, matrix balancing algorithms represent the implicit normalization approaches. Methods such as Iterative Correction (ICE) [22], Knight-Ruiz Matrix Balancing (KR) [23], and Vanilla-Coverage (VC) [6] operate under the premise that the Hi-C matrix should exhibit symmetry and maintain equal row and column sums after the normalization process. This fundamental assumption about the characteristics of a normalized Hi-C matrix drives these implicit normalization strategies, reducing the requirement for specific bias information.

Directly applying bulk normalization methods on scHi-C data results in sub-optimal results due to noteworthy distinctions stemming from the inherent features of single-cell data. These differences are predominantly due to the need to account for the sparsity of single-cell contact maps and the variability between cells in scHi-C methods.

In the wake of advancing computational methodologies for scHi-C, several important studies have proposed normalization techniques for single-cell Hi-C data. The scHiCNorm [19] implements normalization of scHi-C data using zero-inflated and hurdle models, while BandNorm [24] offers a fast-scaling normalization approach that exploits stratified off-diagonals of the contact matrix and its variants. Additionally, scVI-3D [24] introduced a generative model structure that systematically incorporates structural properties.

Despite the innovative strategies and pioneering efforts demonstrated by methods such as scHiCNorm, BandNorm, and scVI-3D, these methods fall short of fully capturing the fundamental properties of single-cell and bulk Hi-C data. Previous studies on bulk Hi-C data have demonstrated that the likelihood of observing a contact between two chromosomal

elements decays linearly with genomic distance, adhering to a power law regime within a certain distance interval [6], [25]. This property has inspired the development of computational methods [26], [27] to identify unique chromatin linkages and statistically significant chromatin interactions. More recent investigations into a range of scHi-C assays have confirmed that scHi-C data similarly conform to the power law distribution [17], [28].

Therefore, we introduce PLNorm, a method that utilizes the power law property of scHi-C data. We extensively evaluate PLNorm against leading scHi-C normalization methods using three public datasets. Various metrics assess changes in cell similarity, clustering performance, and scalability. The results showcase PLNorm's advantages in removing biases while retaining biological information and clustering accuracy. Critically, PLNorm also demonstrates excellent scalability.

II. Materials and methods

A. Data Collection and Preprocessing

This study utilized three publicly available scHi-C datasets: Ramani et al. [12] (Gene Expression Omnibus (GEO): GSE84920), 4DN sci-Hi-C [29] (4DN Data Portal: 4DNES4D5MWEZ, 4DNESUE2NSGS, 4DNESIKGI39T, 4DNES1BK1RMQ, and 4DNESTVIP977), and Lee et al. [14] (Gene Expression Omnibus (GEO): GSE130711).

The first two datasets underwent additional processing steps as described in [30]. For the processing steps of the Lee et al. dataset, please refer to the corresponding paper. All datasets were processed to 1 megabases (Mb) resolution from the raw data to calculate the metrics used in this study.

B. Normalization Methods

This paper evaluated several scHi-C normalization methods, denoted as BandNorm [24], scVI-3D [24], scHiCNorm [19], and PLNorm, which is the normalization method we propose in this study.

BandNorm is a normalization method specifically designed for scHi-C data. Its main purpose is to eliminate genomic distance bias within a cell and sequencing depth bias between cells. Additionally, BandNorm incorporates a common band-dependent contact decay profile to restore the contact matrices' overall structure across cells. By applying BandNorm, it is possible to enhance cell-type clustering, accurately identify interacting loci, and improve the recovery of cell-type relationships in single-cell Hi-C data analysis.

scVI-3D is a deep generative modeling framework designed specifically for scHi-C data analysis. It utilizes parametric count models, such as Poisson and Negative Binomial distributions, which have been proven effective in bulk chromatin conformation capture data analysis. By leveraging variational autoencoders, scVI-3D learns nonlinear mappings and effectively addresses various biases specific to scHi-C data, including genomic distance bias, sequencing depth effects, zero inflation, sparsity impact, and batch effects. One of the key capabilities of scVI-3D is its ability to impute sparse scHi-C contacts and accurately recover cell-type relationships.

scHiCNorm is a specialized software package designed to address systematic biases in scHi-C data. It employs zero-inflated and hurdle models to effectively mitigate biases arising from factors such as cutting sites, GC content, and mappability. By removing these biases, scHiCNorm enhances the ability to accurately assess cell-to-cell variances in chromosomal structures. This allows for a more comprehensive and reliable characterization of chromosomal interactions and structural variations in scHi-C data.

PLNorm is a normalization method specifically designed for scHi-C data. It addresses systematic biases present in scHi-C data using the power-law distribution.

Power law distributions are frequently found across a range of natural, social, and biological systems. For instance, the distribution of connections within a protein-protein interaction network manifests a power-law degree distribution [31]. This can be represented by the following equation:

$$P(k) \propto k^{-\alpha}$$

In this formula, $P(k)$ stands for the probability of a node establishing k connections, while α denotes the scaling exponent. This distribution implies that a small number of nodes (referred to as hubs) within the network are highly interconnected, whereas the majority of nodes have only a few connections.

Prior to the formulation of our PLNorm method, an extensive analysis of the power law characteristics present in single-cell Hi-C (scHi-C) data was undertaken. This examination utilized both the Kolmogorov-Smirnov (K-S) test and the likelihood ratio test to scrutinize the scaling of interaction frequency with genomic distance. Four scHi-C datasets served as the focus of this study: those documented by Ramani et al., 4DN sci-Hi-C, Lee et al., and the additional dataset presented by Li et al. [13].

The hypotheses for the K-S test and the likelihood ratio test were formulated as follows:

k-S test:

H_0 : The interaction frequency between genomic regions scales as a power law with respect to genomic distance

H_a : The interaction frequency between genomic regions does not scale as a power law with respect to genomic distance

Likelihood ratio test:

H_0 : The reference model assumes a power-law relationship between interaction frequency and genomic distance adequately describes the scHi-C data

H_a : The alternative model provides a better fit to the scHi-C data

The p-values from the K-S tests for the datasets from Ramani et al., 4DN sci-Hi-C, Li et al., and Lee et al. all exceeded the 95% confidence threshold, signaling a power law pattern in these datasets. Conversely, the likelihood ratio tests yielded p-values significantly below the confidence level, implying a truncated power law—a variant of the conventional power law—as a more apt model for these datasets. Notably,

the Lee et al. dataset appeared to be better characterized by a stretched exponential distribution.

Supplementary visual analyses of the interaction frequencies against genomic distances corroborated these statistical outcomes. The data from Ramani et al., 4DN sci-Hi-C, and Li et al. conformed to power law scaling, whereas the Lee et al. dataset exhibited certain discrepancies. Nevertheless, the K-S test results suggest that, while not optimal when compared to the stretched exponential model, a power law distribution still provides a commendable fit for the Lee et al. data.

This in-depth investigation substantiates the presence of power law dynamics within scHi-C datasets. Such evidence substantiates the exploitation of power law properties in the development of novel computational methodologies for the analysis of scHi-C data, thereby underpinning the foundation of our PLNorm approach.

Specifically, PLNorm performs normalization on the raw contact matrix X_i of cell i, resulting in the transformed matrix $\hat{X}_i$. This transformation is achieved by estimating the power-law scaling parameter α by applying Maximum Likelihood Estimation (MLE) on its probability density function (PDF) using the following expression:

$$\hat{X}_i = X_i \times \hat{\alpha} = \frac{X_i \times n}{\sum_{j=1}^{n} \log_2(x_j + 1)} \tag{1}$$

where X is a n-by-n matrix representing the contact frequency, n denotes the chromosome length divided by the predefined resolution, and x_j represents the column sum of column j within the matrix X_i.

C. Evaluation Criteria

We employed various metrics to assess the performance of the normalization methods used in this study. These evaluation tools include **HiCRep** [32] and **cell embeddings**.

HiCRep is a method specifically designed for calculating the correlation between two contact matrices. It applies a smoothing technique to the contact maps, which enhances contiguity and promotes the identification of domain structures. Within each genomic distance stratum, HiCRep computes the Pearson correlation coefficient and aggregates the stratum-specific correlation coefficients to obtain the Stratum-adjusted Correlation Coefficient (SCC). In our study, we calculated SCCs between pairwise single cells and derived the overall mean SCC as well as cell type-specific mean SCCs using cell type information. To assess the similarity change between the normalization methods and the raw data, we computed the absolute difference between the SCC values obtained from the different normalization methods and the raw data. A smaller difference indicates a normalization method can reduce biases and at the same time does not change the similarity of raw scHi-C data dramatically. To the contrary, a larger difference means that a normalization method changes the raw scHi-C data significantly.

In the context of this research, **cell embeddings** refer to the derived latent vectors of single cells, generated via the "InterProduct" similarity method encapsulated within **scHiCTools** [33]. To critically evaluate the performance of different normalization methods using these embeddings, our study implemented two distinct evaluation approaches:

1) *t-SNE Clustering Evaluation:* We procured cell embeddings from the scHi-C data normalized by each respective method. These embeddings were then subject to dimensionality reduction using t-SNE [34], resulting in a two-dimensional representation conducive to visual interpretation. Subsequent clustering via the K-means algorithm was conducted on these reduced embeddings, and the concordance with known cell labels was statistically quantified using the Adjusted Rand Index (ARI). The ARI metric was utilized to gauge the accuracy of the clustering results in alignment with the ground truth labeling.
2) *Silhouette Score Assessment:* The second approach entailed computing the average Silhouette score directly from the t-SNE reduced embeddings, eschewing any clustering. The Silhouette score serves as a measure of how distinct the clusters are, with higher values indicating more pronounced separation.

III. Results and Discussion

A. HiCRep Results

In order to measure the ability to reveal the cell-to-cell variances, we calculated the SCC values for each method of the three datasets in this study. Figure 1 illustrates the absolute difference in the SCC values between the normalization methods and the raw data for the Ramani et al. dataset. It is evident from the plot that PLNorm, BandNorm, and scHiCNorm exhibit small differences compared to the raw data across various cell types and the pseudo bulk data. On the other hand, scVI-3D demonstrates the largest deviation from the raw data.

Figure 2 presents a similar pattern to Figure 1, demonstrating the absolute difference in the SCC values between the normalization methods and the raw data for the Lee et al. dataset. In this case, PLNorm and BandNorm exhibit the lowest absolute differences compared to the raw data, indicating their effectiveness in preserving the correlation structure. On the other hand, both scHiCNorm and scVI-3D display some degree of deviation from the raw data.

The 4DN sc-Hi-C dataset, possessing the highest number of cells, displays a trend in Figure 3 consistent with previous observations. This trend reaffirms that PLNorm and BandNorm provide results closest to the raw data, while scVI-3D diverges the most.

B. Cell Embeddings Results

As described in the methodology section, we obtained latent cell embeddings for each method across the three datasets to assess their capacity to enhance cell-to-cell variances and distinguish between different cell types.

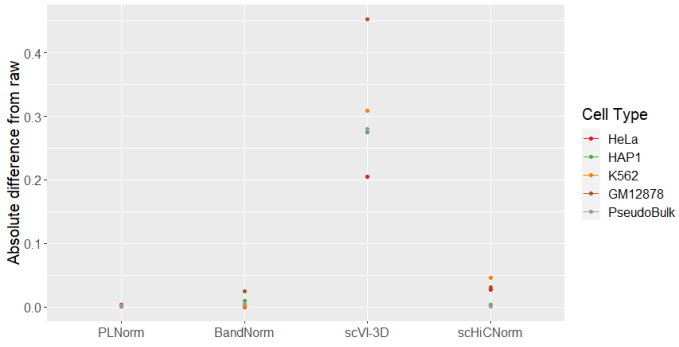

Fig. 1. Absolute difference of the SCC value between normalization methods from the raw data for Ramani et al.

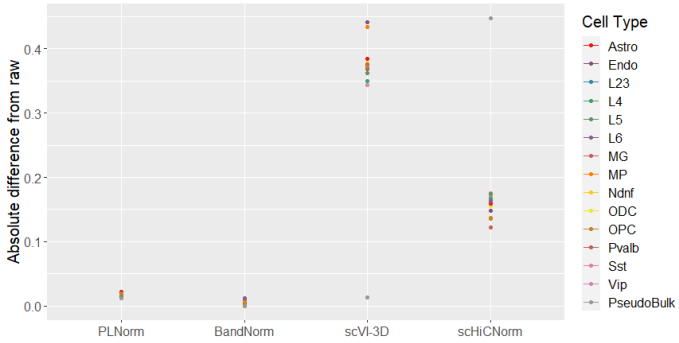

Fig. 2. Absolute difference of the SCC value between normalization methods from the raw data for Lee et al.

Table I demonstrates the superior performance of our method, PLNorm, compared to other methods in terms of the ARI and silhouette score for the Ramani et al. dataset, with an ARI of 0.627 and a silhouette score of 0.373. BandNorm achieves comparable results to the raw data, with an ARI of 0.450 and a silhouette score of 0.337. On the other hand, methods such as scVI-3D and scHiCNorm exhibit significantly poorer performance in clustering cell types compared to the raw data.

Table II presents the results for the Lee et al. dataset, which differ slightly from the results in Table I. In this dataset, both PLNorm and scHiCNorm outperform the raw data in clustering the cell types. However, PLNorm remains the top-performing method, achieving an ARI of 0.361 and a silhouette score of 0.125. This indicates the effectiveness of our method, PLNorm, in preserving cell-type information and enhancing the clustering performance compared to other methods.

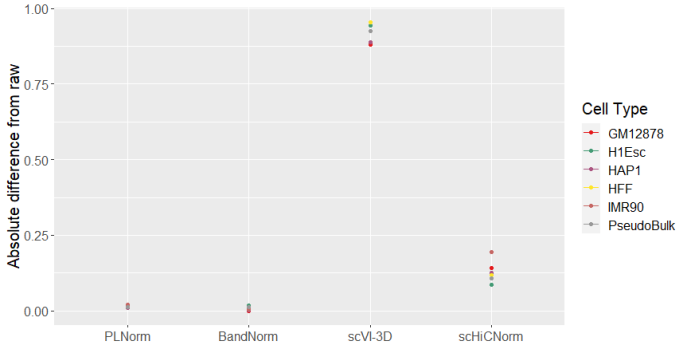

Fig. 3. Absolute difference of the SCC value between normalization methods from the raw data for 4DN sc-Hi-C

Table III displays the results for the 4DN sc-Hi-C dataset, where all methods except scHiCNorm outperform the raw data in terms of ARI and the silhouette score. The ARI and silhouette scores for PLNorm, BandNorm, and scVI-3D are relatively similar. However, it is worth noting that all methods achieve low ARI values and silhouette scores, indicating the difficulty of clustering cell types in this dataset due to its larger size.

TABLE I
CELL EMBEDDINGS RESULTS FOR RAMANI ET AL.

method	ARI_K-means	silhouette
raw	0.453	0.343
PLNorm	**0.627**	**0.373**
BandNorm	0.450	0.337
scVI-3D	0.162	-0.107
scHiCNorm	0.114	0.069

TABLE II
CELL EMBEDDINGS RESULTS FOR LEE ET AL.

method	ARI_K-means	silhouette
raw	0.006	-0.072
PLNorm	**0.361**	**0.125**
BandNorm	0.005	-0.072
scVI-3D	-0.001	-0.081
scHiCNorm	0.352	0.120

TABLE III
CELL EMBEDDINGS RESULTS FOR 4DN SC-HI-C.

method	ARI_K-means	silhouette
raw	0.015	-0.097
PLNorm	0.032	-0.088
BandNorm	0.031	-0.089
scVI-3D	0.033	-0.080
scHiCNorm	0.015	-0.158

In Figure 4, we present the cell embeddings for both raw data and the various scHi-C normalization methods evaluated in this study across three scHi-C datasets. Notably, our method, PLNorm, distinctly differentiates various cell lines for the first two datasets. In contrast, some methods, such as scHiCNorm, exhibit a suboptimal performance on this same task. This visualization of the last dataset illustrates the difficulty of separating a larger scHi-C dataset.

C. Computational Performance

We conducted a comparative evaluation of computational performances under different conditions. The RAM usage and elapsed time for each method in single-core mode, except for scVI-3D, the only method supporting multi-core parallel

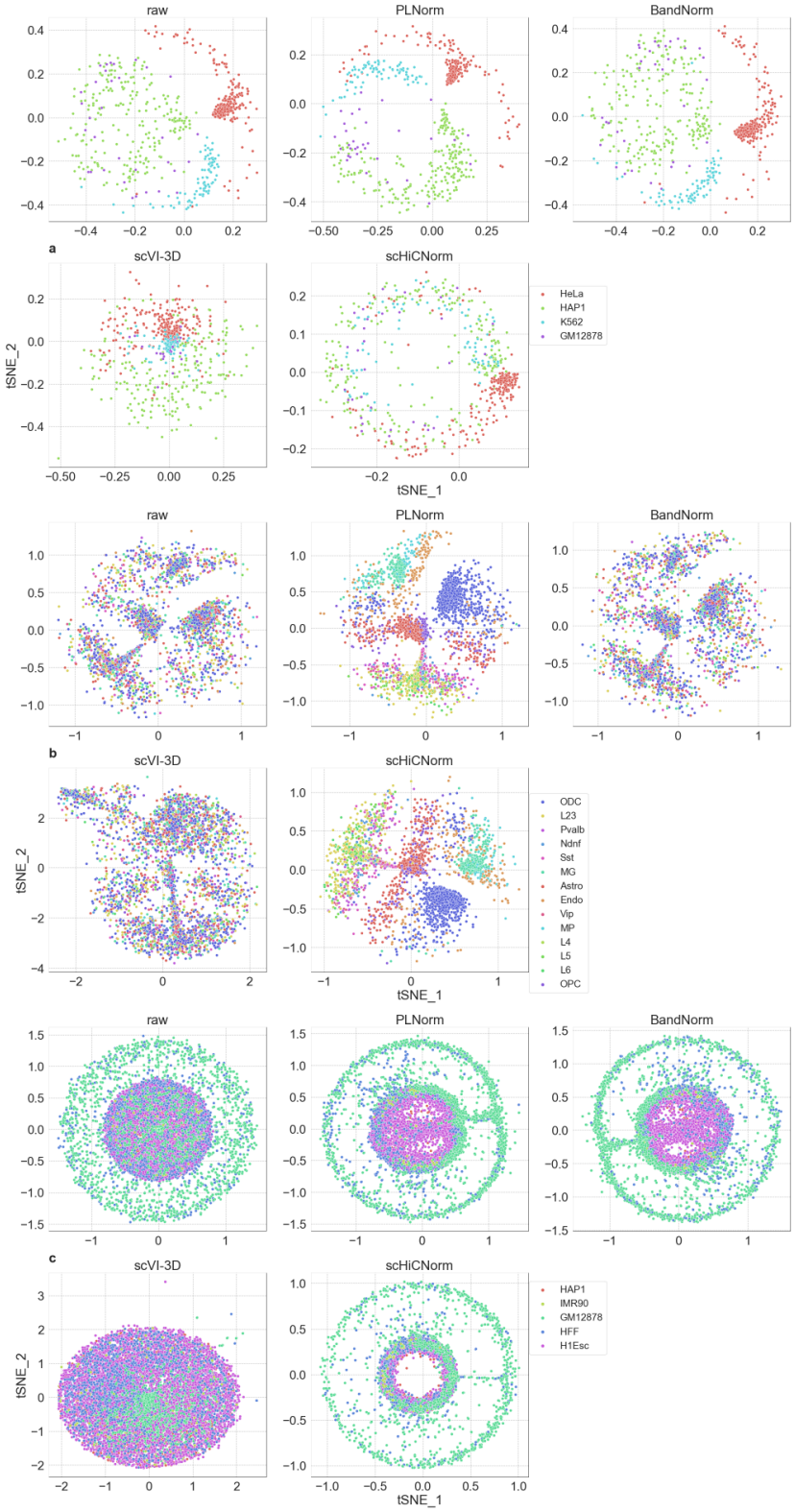

Fig. 4. Comparison of scHi-C normalization methods for their performances in separating cells from major cell lines. Application of the scHi-C data normalization methods on Ramani2017 (a) with 4 cell types and 620 cells, Lee2019 (b) with 14 cell types and 4,238 cells, and 4DN sc-Hi-C (c) data sets with 5 cell types and 16,707 cells. The results are displayed using scatter plots of the first two t-SNE coordinates. The colors of the plotting symbols depict the cell types

computing, for which we tested the performance in the case of 10-core modes, are illustrated in Figure 5 and Table IV, respectively. As anticipated, PLNorm exhibits the highest scalability (Figure 6), primarily due to its reliance on basic operations like averaging and multiplication for data scaling. While scVI-3D matches PLNorm's speed, it demands significantly more memory.

Conversely, BandNorm exhibits constraints in scalability. Although its memory requirements remain relatively modest among the methods evaluated—around 30 GB for 4,238 cells as documented in Table V—it demands a significantly extended duration, taking up to 507 minutes for 16,707 cells. For more expansive datasets, we foresee BandNorm's processing time increasing at a considerably sharper rate than scVI-3D, a trend evident when comparing results from Ramani et al. to those from Lee et al. Interestingly, despite the 4DN sc-Hi-C dataset having fewer memory requirements than Lee et al.—even with a larger cell count—we surmise this may be attributed to a substantial reduction in the number of loci pairs in individual cells. scHiCNorm, though displaying commendable memory efficiency, emerges as the most time-intensive method, necessitating a staggering 11,818 minutes for 16,707 cells, as indicated in Table V. Collectively, PLNorm stands out for its exceptional scalability, demonstrating both time and memory efficiency, as illustrated in Figure 5.

TABLE IV
METHOD SCALABILITY IN TERMS OF COMPUTATIONAL TIME (IN MINUTES) VERSUS NUMBER OF CELLS

Computational cost (m)	Ramani et al.	Lee et al.	4DN sc-Hi-C
cell num	620	4,238	16,707
loci pairs (median)	6,143	32,923	3,006
PLNorm	**13**	40	**170**
BandNorm	15	**28**	507
scVI-3D	25	170	472
scHiCNorm	226	3,859	11,818

TABLE V
METHOD SCALABILITY IN TERMS OF RAM USAGE VERSUS NUMBER OF CELLS

RAM (Gb)	Ramani et al.	Lee et al.	4DN sc-Hi-C
cell num	620	4,238	16,707
loci pairs (median)	6,143	32,923	3,006
PLNorm	**0.027**	**0.036**	**0.036**
BandNorm	2.6	29	16.6
scVI-3D	22	42	63
scHiCNorm	0.219	0.232	0.243

D. Discussion and Conclusions

Derived from the Power Law distribution scale parameter estimate, $\hat{\alpha}$, we propose a simple and scalable global normalization method, termed PLNorm. Not only is PLNorm quick, but it also exhibits remarkable memory efficiency.

As PLNorm does not require a reference and operates independently for each cell, it allows normalization application to new out-of-sample data. The ultimate goal of the transformation is to retain and amplify cell-to-cell variances. It is noteworthy that our method, like other global scaling methods, does not eliminate batch effects; these can be addressed with downstream tools (e.g., [35]–[37]).

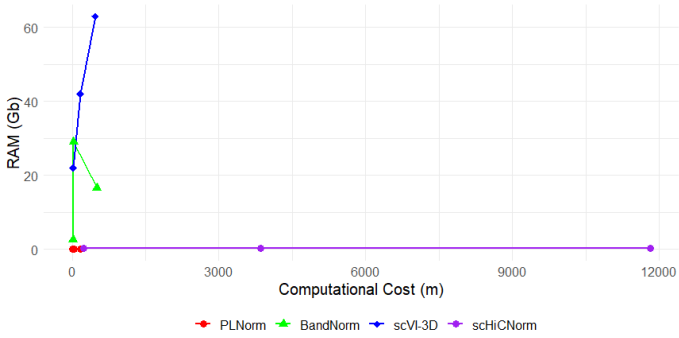

Fig. 5. Comparative evaluation of RAM usage (Gb) and computational time (in minutes) on three data sets with an increasing number of cells

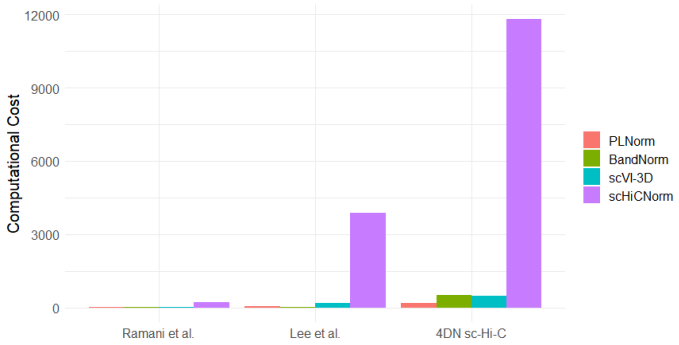

Fig. 6. Method scalability in terms of computational time versus number of cells

We evaluate the performance of several normalization methods—BandNorm, scVi-3D, scHiCnorm, and our proposed PLNorm—in terms of cell embeddings (ARI and silhouette score), changes in cell similarity, and scalability. Among these, PLNorm exhibits commendable performance across all four aspects.

In conclusion, normalization for maintaining cell-to-cell variances appears less critical than normalization for clustering and cell type discovery. Even rudimentary methods such as scHiCNorm perform adequately in this metric, despite their divergence from raw data but a reduction in cell similarities. Therefore, both the ability in cell embeddings and the scalability of methods become crucial factors to consider when selecting a normalization technique. Our proposed PLNorm normalization strikes a favorable balance between clustering accuracy and scalability, displaying superior performance to BandNorm with only a minor increase in the elapsed time in certain cases. This makes PLNorm a promising normalization approach for very large datasets.

ACKNOWLEDGMENT

The work is supported by the National Science Foundation under NSF EPSCoR Track-1 Cooperative Agreement OIA #1946202 and used advanced cyberinfrastructure resources

provided by the University of North Dakota Computational Research Center and the Center for Computationally Assisted Science and Technology (CCAST) at North Dakota State University.

REFERENCES

[1] J. Dekker, K. Rippe, M. Dekker, and N. Kleckner. Capturing chromosome conformation. *Science*, 295(5558):1306–1311, 2002.

[2] M. A. Rubtsov, Y. S. Polikanov, V. A. Bondarenko, Y. H. Wang, and V. M. Studitsky. Chromatin structure can strongly facilitate enhancer action over a distance. *Proceedings of the National Academy of Sciences*, 103(47):17690–17695, 2006.

[3] A. Miele and J. Dekker. Long-range chromosomal interactions and gene regulation. *Molecular biosystems*, 4(11):1046–1057, 2008.

[4] E. De Wit and W. De Laat. A decade of 3c technologies: insights into nuclear organization. *Genes & development*, 26(1):11–24, 2012.

[5] W. De Laat and D. Duboule. Topology of mammalian developmental enhancers and their regulatory landscapes. *Nature*, 502(7472):499–506, 2013.

[6] E. Lieberman-Aiden, N. L. Van Berkum, L. Williams, M. Imakaev, T. Ragoczy, A. Telling, and J. Dekker. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*, 326(5950):289–293, 2009.

[7] J. R. Dixon, S. Selvaraj, F. Yue, A. Kim, Y. Li, Y. Shen, and B. Ren. Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature*, 485(7398):376–380, 2012.

[8] S. S. Rao, M. H. Huntley, N. C. Durand, E. K. Stamenova, I. D. Bochkov, J. T. Robinson, and E. L. Aiden. A 3d map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell*, 159(7):1665–1680, 2014.

[9] Job Dekker, Andrew S Belmont, Mitchell Guttman, Victor O Leshyk, John T Lis, Stavros Lomvardas, Leonid A Mirny, Clodagh C O'shea, Peter J Park, Bing Ren, et al. The 4d nucleome project. *Nature*, 549(7671):219–226, 2017.

[10] Claire Marchal, Jiao Sima, and David M Gilbert. Control of dna replication timing in the 3d genome. *Nature Reviews Molecular Cell Biology*, 20(12):721–737, 2019.

[11] T. J. Stevens, D. Lando, S. Basu, L. P. Atkinson, Y. Cao, S. F. Lee, et al. 3d structures of individual mammalian genomes studied by single-cell hi-c. *Nature*, 544(7648):59–64, 2017.

[12] V. Ramani, X. Deng, R. Qiu, K. L. Gunderson, F. J. Steemers, C. M. Disteche, and J. Shendure. Massively multiplex single-cell hi-c. *Nature Methods*, 14(3):263–266, 2017.

[13] G. Li, Y. Liu, Y. Zhang, N. Kubo, M. Yu, R. Fang, and B. Ren. Joint profiling of dna methylation and chromatin architecture in single cells. *Nature Methods*, 16(10):991–993, 2019.

[14] D. S. Lee, C. Luo, J. Zhou, S. Chandran, A. Rivkin, A. Bartlett, and J. R. Ecker. Simultaneous profiling of 3d genome structure and dna methylation in single human cells. *Nature Methods*, 16(10):999–1006, 2019.

[15] L. Tan, W. Ma, H. Wu, Y. Zheng, D. Xing, R. Chen, and X. S. Xie. Changes in genome architecture and transcriptional dynamics progress independently of sensory experience during post-natal brain development. *Cell*, 184(3):741–758, 2021.

[16] S. V. Ulianov, V. V. Zakharova, A. A. Galitsyna, P. I. Kos, K. E. Polovnikov, I. M. Flyamer, et al. Order and stochasticity in the folding of individual drosophila genomes. *Nature communications*, 12(1):41, 2021.

[17] Takashi Nagano, Yaniv Lubling, Tim J Stevens, Stefan Schoenfelder, Eitan Yaffe, Wendy Dean, Ernest D Laue, Amos Tanay, and Peter Fraser. Single-cell hi-c reveals cell-to-cell variability in chromosome structure. *Nature*, 502(7469):59–64, 2013.

[18] E. Yaffe and A. Tanay. Probabilistic modeling of hi-c contact maps eliminates systematic biases to characterize global chromosomal architecture. *Nature genetics*, 43(11):1059–1065, 2011.

[19] T. Liu and Z. Wang. schicnorm: a software package to eliminate systematic biases in single-cell hi-c data. *Bioinformatics*, 34(6):1046–1047, 2018.

[20] M. Hu, K. Deng, S. Selvaraj, Z. Qin, B. Ren, and J. S. Liu. Hicnorm: removing biases in hi-c data via poisson regression. *Bioinformatics*, 28(23):3131–3133, 2012.

[21] Wentian Li, Ke Gong, Qiye Li, Frank Alber, and Xiaobo J Zhou. Hi-corrector: a fast, scalable and memory-efficient package for normalizing large-scale hi-c data. *Bioinformatics*, 31(6):960–962, 2015.

[22] M. Imakaev, G. Fudenberg, R. P. McCord, N. Naumova, A. Goloborodko, B. R. Lajoie, and L. A. Mirny. Iterative correction of hi-c data reveals hallmarks of chromosome organization. *Nature methods*, 9(10):999–1003, 2012.

[23] P. A. Knight and D. Ruiz. A fast algorithm for matrix balancing. *IMA Journal of Numerical Analysis*, 33(3):1029–1047, 2013.

[24] Y. Zheng, S. Shen, and S. Keleş. Normalization and de-noising of single-cell hi-c data with bandnorm and scvi-3d. *Genome Biology*, 23(1):222, 2022.

[25] B. R. Lajoie, J. Dekker, and N. Kaplan. The hitchhiker's guide to hi-c analysis: practical guidelines. *Methods*, 72:65–75, 2015.

[26] X. Lan, H. Witt, K. Katsumura, Z. Ye, Q. Wang, E. H. Bresnick, ..., and V. X. Jin. Integration of hi-c and chip-seq data reveals distinct types of chromatin linkages. *Nucleic acids research*, 40(16):7690–7704, 2012.

[27] Y. Zhou, X. Cheng, Y. Yang, T. Li, J. Li, T. H. M. Huang, ..., and V. X. Jin. Modeling and analysis of hi-c data by hisif identifies characteristic promoter-distal loops. *Genome Medicine*, 12(1):1–13, 2020.

[28] B. Zhao, P. Shen, and L. Liu. Do single-cell hi-c data follow a power law distribution? *International Conference on Intelligent Biology and Medicine*, 2023.

[29] H. J. Kim, G. G. Yardımcı, G. Bonora, V. Ramani, J. Liu, R. Qiu, and W. S. Noble. Capturing cell type-specific chromatin compartment patterns by applying topic modeling to single-cell hi-c data. *PLoS Computational Biology*, 16(9):e1008173, 2020.

[30] R. Zhang, T. Zhou, and J. Ma. Multiscale and integrative single-cell hi-c analysis with higashi. *Nature Biotechnology*, 40(2):254–261, 2022.

[31] Hawoong Jeong, Stephen P Mason, Albert-László Barabási, and Zoltan N Oltvai. Lethality and centrality in protein networks. *Nature*, 411(6833):41–42, 2001.

[32] Tao Yang, Feipeng Zhang, Galip Gürkan Yardımcı, Fan Song, Ross C Hardison, William Stafford Noble, Feng Yue, and Qunhua Li. Hicrep: assessing the reproducibility of hi-c data using a stratum-adjusted correlation coefficient. *Genome research*, 27(11):1939–1949, 2017.

[33] X. Li, F. Feng, H. Pu, W. Y. Leung, and J. Liu. schictools: a computational toolbox for analyzing single-cell hi-c data. *PLoS computational biology*, 17(5):e1008978, 2021.

[34] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

[35] Andrew Butler, Paul Hoffman, Peter Smibert, Efthymia Papalexi, and Rahul Satija. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nature biotechnology*, 36(5):411–420, 2018.

[36] Laleh Haghverdi, Aaron T.L. Lun, Michael D. Morgan, and John C. Marioni. Batch effects in single-cell rna-sequencing data are corrected by matching mutual nearest neighbors. *Nature biotechnology*, 36(5):421–427, 2018.

[37] Davide Risso, John Ngai, Terence P. Speed, and Sandrine Dudoit. Normalization of rna-seq data using factor analysis of control genes or samples. *Nature biotechnology*, 32(9):896–902, 2014.