Model-based imputation enables improved resolution for identifying differential chromatin contacts in single-cell Hi-C data

Neda Shokraneh NSHOKRAN@SFU.CA
Megan Andrews MAA160@SFU.CA
Maxwell Libbrecht MAXWL@SFU.CA
Computing Science Department, Simon Fraser University, 8888 University Dr, BC, Canada

Abstract

Recent advances in single-cell Hi-C (scHi-C) assays allow studying the chromatin conformation at the resolution of a single cell or a cluster of cells. A key question is to identify changes in the contact strength between two cell types, known as differential chromatin contacts (DCCs). While existing statistical methods can identify changes in contact strength in bulk Hi-C data, these methods cannot be effectively applied to scHi-C data due to its severe sparsity. Thus it is necessary to develop methods for identifying differential chromatin contacts in scHi-C data.

Recently-developed scHi-C imputation approaches can mitigate the issue of sparsity. We propose an approach for identifying differential chromatin contacts using these imputation approaches. We build upon the existing SnapHiC-D method by replacing its imputation step with recent learning-based imputation approaches. We show that, via analysis of real scHi-C datasets with different coverages and at different resolutions, imputation approaches that consider the spatial correlation between bin pairs, Higashi, and random walk with restart, outperform other approaches. Furthermore, we show that careful considerations are needed when imputation is done in preprocessing steps as it may invalidate downstream statistical approaches. Finally, our results indicate that model-based imputations greatly improve performance when analyzing chromatin contacts at moderate resolution (100kb); however, current imputation approaches are inefficient in terms of both accuracy and computational complexity when being applied to high-resolution scHi-C resolution (10kb).

Keywords: scHi-C, differential chromatin contacts, sparsity, imputation

1. Introduction

The 3D chromatin structure is one of the regulators of cellular processes such as gene expression, DNA replication, and splicing (12), and its dynamic plays a critical role in disease (9) and development (5). Advances in chromatin conformation capture with high-throughput sequencing (Hi-C) revealed chromatin fibers' multi-scale and dynamic nature in high resolution. However, bulk Hi-C assays measure the average contacts among many cells and cannot capture cell-type-specific conformation in complex tissues. Recent advances in single-cell Hi-C (scHi-C) assays (16; 17; 18; 7; 21; 10; 13) enable profiling chromatin conformations in individual cells, and it was shown that cells can be clustered according to their Hi-C contact maps (27; 7; 26; 25).

An essential statistical question for scHi-C analysis is to identify differential chromatin contacts (DCCs) between two clusters of cells to characterize them. However, the sparsity of observed contact counts given limited sequencing depths obscures the biological variation, and their distributions do not fit assumptions of existing DCC callers proposed for bulk Hi-C data (15; 20; 4; 6).

Current approaches for the identification of DCCs from scHi-C data are based on pseudo-bulking (26) or imputation (11; 23). Pseudo-bulking- or aggregation-based methods utilize the large sample size of single-cell assays and split scHi-C contact maps corresponding to one cluster into two sets and make two pseudo-bulk Hi-C contact maps by summation. Aggregation of a large number of

cells diminishes the technical variability and reduces the noise and sparsity. Then, existing DCC callers for bulk Hi-C data can be used to identify DCCs. On the other hand, the imputation-based approach keeps the large sample size property for increased statistical power (more reliable estimation of within and between clusters' variation) and addresses noise and sparsity challenges with imputation.

Each of these approaches has its own limitations. Pseudo-bulking-based methods require a large sample size depending on the level of sparsity and the resolution of analysis to fit the distributional assumption of counts in bulk methods. Imputation methods induce false positive signals besides true ones and careful considerations should be taken on the distribution of imputed counts and if post-normalization is required to remove original and after-imputation biases.

SnapHiC-D (11) proposed an imputation-based DCC caller (utilizing Random walk with restart (RWR)) and showed that it is more sensitive and accurate compared to pseudo-bulking approaches. On the other hand, recent learning-based scHi-C computational tools, Higashi (25) and scVI-3D (26), show superior performance compared to scHiCluster (27) which utilizes RWR in other scHi-C downstream analysis tasks such as cell embedding and imputation quality. Therefore, such learning-based imputation methods have the potential to enhance the performance of DCC callers as well.

In this paper, we comprehensively evaluate the effect of imputation methods on calling significant DCCs from scHi-C datasets. We show that model-based imputation approaches, Higashi and RWR, which assume the existence of a correlation between spatial neighbors, improve the power of single-cell DCC caller and identify significant DCCs that are more precise and consistent with reliable significant DCCs identified by deeply sequenced bulk Hi-C data. We show that this is particularly important when identifying DCC at a moderate resolution (100 Kb in our datasets) or from a low-coverage dataset, as imputation is required to achieve reasonable statistical power in those settings. Finally, we discuss the inefficiency of current imputation approaches for high-resolution scHi-C data (10 Kb in our dataset).

2. Methods

2.1. Problem formulation

A scHi-C experiment generates read pairs associated with single cells from the profiled sample that are processed and binned into squared Hi-C matrices at a specific resolution. The cells from a single scHi-C experiment can be clustered (or grouped) into a discrete number of clusters. Examples of these clusters are discrete cell lines, cell types from tissues, cell cycles, etc. The DCC identification problem is to identify differences between such clusters that distinguish them from each other.
Our pipeline takes either matrices from two specific clusters or a whole annotated scHi-C dataset and two cluster names to be compared from the dataset as input depending on the imputation method to be applied. It outputs a p-value for each contact bin pair, evaluating whether the pair has different contact strength between the two groups.

2.2. Method

Our framework is based on the first (to our knowledge) DCC caller for scHi-C data, SnapHiC-D (11). This method has 5 steps:

(1) Impute single-cell contact matrices using random walk with restart (RWR).

(2) Normalize imputed counts distance-wise, by transforming imputed counts within the same genomic distance into z-scores.

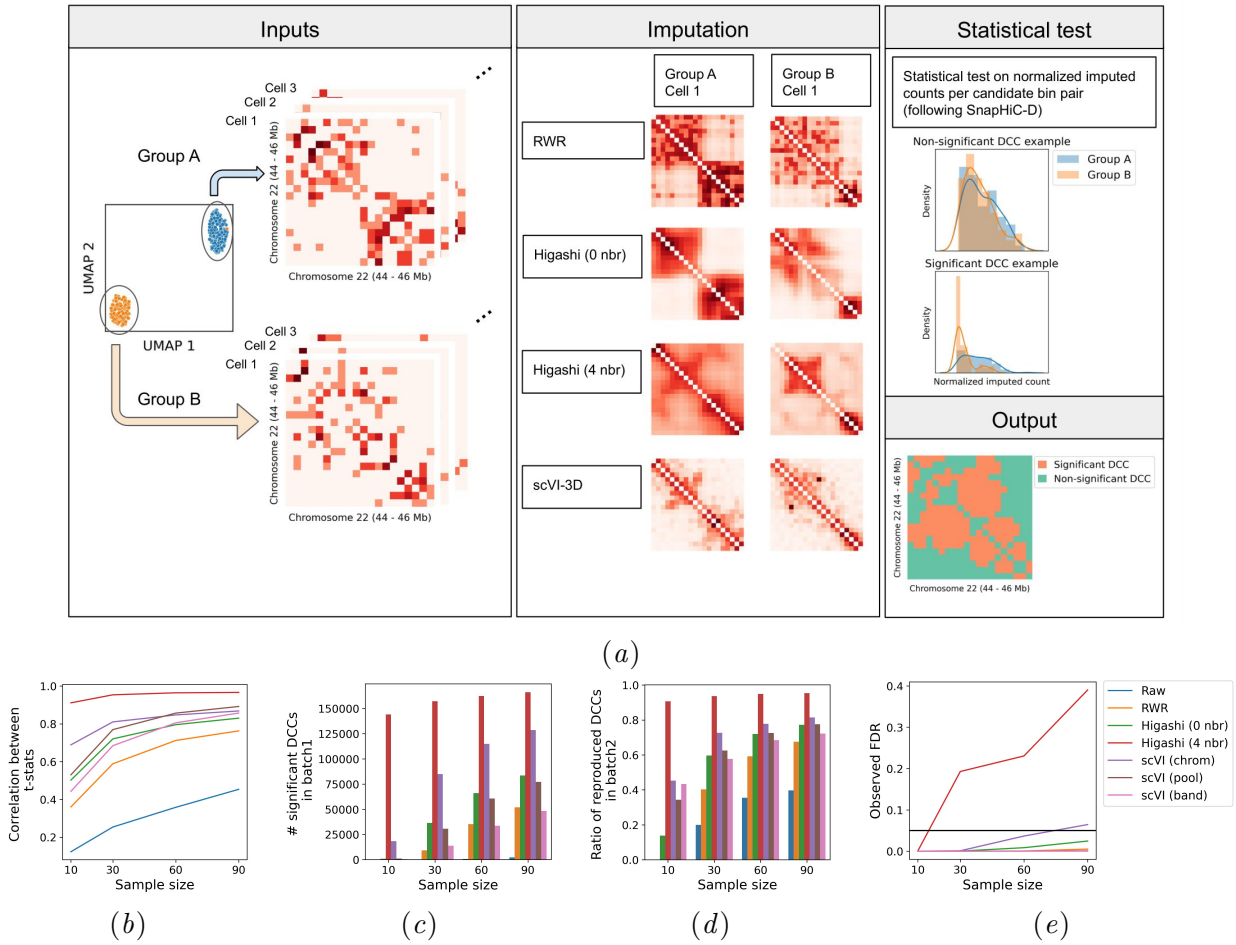


Figure 1: Overview of the framework for single-cell DCC caller. Two groups of cells are imputed by one of the RWR, Higashi, and scVI-3D methods. The two-sided, two-sample t-test is applied on normalized imputed counts following SnapHiC-D (11). The output is a t-statistic and p-value per bin pair indicating whether the pair has different contact strength between the two groups. (b,c,d,e) DCC analysis between Astro and MG is done twice for two batches, 190315_21yr and 190315_29yr from *Lee2019* dataset. (b) The correlation between t-statistics from two replicates. (c) The number of called DCCs from batch 1. (d) The ratio of called DCCs from batch 1 that are reproduced in batch 2. (e) The observed false discovery rate (FDR) when comparing two groups of Astro cells, one from batch 190315_21yr, and another one from batch 190315_29yr. All experiments are done on chromosomes {11..22} at 100 Kb resolution for different numbers of cells in the groups (sample size). The imputed counts are normalized with distance-based normalization.

(3) Select candidate bin pairs with filtering based on 3 criteria. First, remove bin pairs with anchors within annotated filter regions; second, include only bin pairs with at least one anchor covering a TSS region; and finally, only consider bin pairs for which at least 10% of cells from one group of cells have normalized imputed count greater than 1.96 standard deviations.

(4) Apply a two-sided, two-sample t-test on each candidate bin pair to calculate the t-statistic and corresponding p-value of being a DCC.

(5) Perform multiple hypothesis adjustment of p-values given the large number of candidates to control the false discovery rate.

Our contribution is to benchmark recent imputation methods, Higashi (25) and scVI-3D (26), on calling DCCs. To do this, we replaced random walk with restart (RWR) imputation in step (2) above with each of these methods respectively.

- Higashi (25) uses a hypergraph representation learning framework, Hyper-SAGNN (24), to model scHi-C data, learn single-cell embeddings, and impute Hi-C contacts in single-cell resolution. In Higashi, the scHi-C data is transformed to a hypergraph including two node types, cell and bin, and the Hi-C contact between genomic bins j and k in the cell i, is represented as a hyperedge between nodes corresponding to genomic bins j and k and the cell i. Then, a hypergraph neural network, Hyper-SAGNN, is trained to reconstruct the hypergraph. Using the trained hypergraph neural network, the learned embeddings for cell nodes are used as cell embeddings and predicted hyperedges are used as imputed Hi-C contacts at the single cell resolution which we use as imputed contact counts.

  Furthermore, Higashi can enhance imputation by sharing information between neighbor cells in the latent space. They calculate the pairwise distance of cell embeddings and find the k-nearest neighbors of each cell. Then, they construct the new imputed contact map as the weighted sum of contact maps of itself and its neighbors, where the weights are proportional to the distance between cell embeddings. We use both Higashi without and with neighbors in our benchmarking.

- scVI-3D (26) utilizes a deep generative model for the single-cell RNA-seq data, scVI (14), to embed, normalize and impute scHi-C dataset. As scVI takes 1-dimensional vector observation per cell, scVI-3D splits contact matrices into multiple bands (or sets of bands, pools), where each band corresponds to bin pairs within the same genomic distance. Then, one scVI model is trained per band (or pool) to normalize and impute the contact counts within that band (or pool), and learned embeddings from all trained models are concatenated to construct cell embeddings.

  Since there is a lack of research on the most effective strategy for flattening 2D contact matrices, we employ three distinct flattening approaches—chromosome-based, pool-based, and band-based—in our experiments. Chromosome-based flattening involves transforming the entire chromatin contact map into a 1-dimensional vector, and a separate scVI model is trained for each chromosome.

We modified the filtering step slightly. Different genomic bins might have different visibilities such that the total number of interactions corresponding to them is different. If the visibility bias exists, highly visible genomic bins consistently receive a high z-score, and candidates will be limited to bin pairs with anchors from such genomic bins. Therefore, we skipped this filtering step. Second, we only applied TSS filtering for the analysis of the mouse ESC-NPC dataset at 10 Kb resolution. A summary of filtering steps for each of the resolutions is provided in Table 3.

We also tried quantile normalization on the whole data together with distance-wise normalization because of three reasons. First, distance-wise normalization removes global differences between samples such as distance effect, which might not be proper for DCC analysis at coarser resolution. Second, we compare our called DCCs to DCCs from bulk data called by diffHiC (15). The diffHiC method normalizes data with total sequencing effect, therefore normalization based on whole data makes called DCCs more comparable to each other. Third, we can assess the sensitivity of the

results to the normalization approach. Thus, we also employ a total normalization where we apply quantile normalization on *cells* $\times$ *all candidate bin pairs* matrix to transform the statistical distribution across cells to be the same without removing the distance effect.

2.3. Evaluation

2.3.1. Accuracy and recall of called DCCs given ground truth from bulk data

For the experiments where bulk Hi-C data exists for the cell groups, we evaluate significant DCCs according to ground truth DCCs identified by a bulk DCC caller, diffHiC (15), applied to deeply sequenced bulk data. We characterize significant vs. non-significant DCCs ground truth by thresholding on diffHiC log fold-change (LFC) and evaluate precision by visualizing the ROC curve.

2.3.2. Reproducibility of called DCCs across different batches

For the experiments where ground truth does not exist, we identify DCCs between two cell types from the two most abundant batches and calculate the reproducibility of called DCCs across different batches.

2.4. Dataset

In our experiments, we used 3 recent scHi-C datasets with different coverage and throughput (10; 7; 11).

Lee2019 (10) simultaneously profiled DNA methylation and 3D conformation of human brain prefrontal cortex cells, and annotated cells into 14 cell types according to methylation profiles. We compare two non-neuronal cell types, microglia (MG) and astrocyte (Astro), following SnapHiC-D paper (11). We analyzed this dataset at 100 Kb resolution following the original paper.

Kim2020 (7) includes multiple single-cell combinatorially-indexed Hi-C libraries from 5 human cell lines. We compared three cell lines, GM12878 , HFF, and H1Esc, in a pairwise manner (3 pairs of cell lines). We analyzed this dataset at 500 Kb resolution because of its limited sequencing depth. For the evaluation of called DCCs in this dataset, we used two replicates of bulk data generated for each of the GM12878, HFF, and H1Esc cell lines (19; 8; 1).

Lee2023 (11) generated high coverage scHi-C data from 94 mouse embryonic stem cells (ESCs) and 188 mouse neuron progenitor cells (NPCs) to assess the performance of SnapHiC-D at finer resolution (10 Kb) given ground truth DCCs called from bulk data (3). We analyzed this dataset across three resolutions, 10 Kb, 100 Kb, and 1 Mb to study the effect of imputation methods across different resolutions and sparsity levels. For the evaluation of called DCCs in this dataset, we used four replicates of bulk Hi-C data for each of the ESC and NPC.

Details about coverage, throughput, and compared groups of cells of each dataset and data sources are provided in Table 2 and Table 1, respectively.

3. Results

3.1. All imputation methods except Higashi with neighbors control false discovery rate under null hypothesis

Because of the excess of zero contact counts in scHi-C data, the identification of significant DCCs from raw data results in a low recall and a large number of false negatives. Therefore, DCC callers from scHi-C data can benefit from imputation methods to fill technical zeros by sharing the information between neighbor cells and bin pairs.

One downside of imputation methods is inducing false signals into imputed counts and consequently a large number of false positives in downstream differential analysis (2). We assessed the performance of imputation methods in terms of controlling the false discovery rate (FDR) by comparing two groups of cells from the same cell type (Astro) but different batches from *Lee2019*. Ideally, we expect to see no significant DCCs; all called DCCs are considered to be false positives.

The observed FDR under the null hypothesis for different imputations shows that all of them except Higashi with neighbors control FDR under target FDR, 0.05 (Figure 1(*e*)). However, Higashi with neighbors (Higashi (4 nbr)) borrows information from neighbor cells in Higashi's embedding space to enhance the imputation process. This sharing violates the independence assumption of the t-test so its p-values are not reliable.

In our experiments, we observed that Higashi with neighbors results in a larger range of t-statistics compared to t-statistics from other imputation approaches. This is the result of smoothing with neighbors such that imputed counts are highly similar to each other and the denominator of the t-statistic, the standard error of imputed counts, is low. We calculated the coefficient of variation (CV) of imputed contact counts per bin pair, once given all Astro cells, and once for Astro cells from one batch. Figure 4(*b*) shows that the CV of Higashi with neighbor's imputed counts is much smaller than other imputations meaning that they are less variable around the mean. This result suggests that a different statistical test is required for calling significant DCCs from Higashi with neighbor's imputed counts (Discussion).

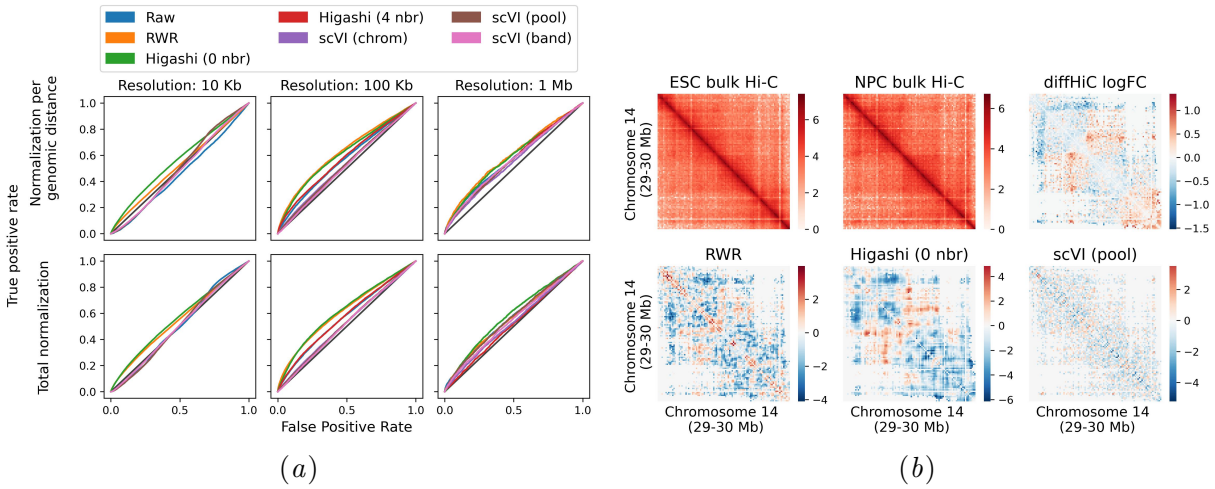


Figure 2: Evaluation of called DCCs from *Liu2023* at different resolutions. (a) Comparison of ROC curves for called DCCs by two different normalization and imputation approaches at three different resolutions. The results for 100 Kb and 1 Mb resolutions are from chromosomes {12..15}. The results for 10 Kb resolution is from chromosome 14. (b) The heatmap of bulk Hi-C contact maps for mESC and mNPC, and diffHiC log fold-change (logFC) and single-cell t-statistics from the comparison of these two cell lines (10 Kb resolution).

3.2. Model-based imputation improves the recall, precision, and reproducibility of called DCCs for moderately sparse contact maps

The main goal of the imputation process is to increase the number of called significant DCCs while controlling the precision and reproducibility. First, we assessed the number and reproducibility of

called DCCs between two cell types (Astro, MG) from *Lee2019*. We use reproducibility to assess the performance of called DCCs because there is no bulk ground truth for this dataset. This dataset has 5 batches, and we calculate the reproducibility of significant DCCs by comparing Astro and MG from the two most abundant batches. We employ two ways to calculate the reproducibility; first, the correlation between t-statistics driven from two different batches, second, the ratio of significant DCCs from the first batch that are called significant in the second batch.

In the *Lee2019* dataset, all imputations result in a significant increase in the number of called DCCs (Figure 1(*c*)). We show that both t-statistics (Figure 1(*b*)) and significantly called DCCs (Figure 1(*d*)) are better reproduced after imputation, particularly learning-based imputation approaches, Higashi and scVI-3D. These results indicate that findings are more reliable after imputation. However, for learning-based imputation approaches (Higashi 4nbr and scVI), the reproducibility might reflect how well the model learned the latent space that separates cell types and mixes different batches, because, as noted above, a t-test is not a valid statistical test.

For example, Higashi with neighbors results in a correlation close to 1 between t-statistics from two batches. However, t-statistics are not reliable as most candidates are called significant given their t-statistics and corresponding adjusted p-values. In addition to the smaller CV of Higashi with neighbor's imputed counts (discussed in Section 3.1), we observed that the total CV, CV of a set of cells including both Astro and MG, is much higher than within-cell-type CV (Figure 4(*a*)). After visualization of a few random bin pairs, we found that the distribution of Higashi with neighbor's imputed counts is bimodal (Figure 4(*c*)), which means almost all bin pairs are differentially expressed after Higashi imputation that incorporates neighbors' information. Since such artifact results in high reproducibility metrics, we should be careful about the biological interpretation of t-statistics after learning- and smoothing-based imputation approaches.

To evaluate the biological validity of identified DCCs, we assess the precision of called DCCs between mouse ESC and NPC from *Lee2023* at the same resolution, 100 Kb, with a comparable sparsity level. The chromatin conformation of these two mouse cell types is richly profiled (3), and we use significantly called DCCs by applying diffHiC on this bulk Hi-C data as a ground truth. Analysis of the ROC curve of single-cell t-statistics given ground truth (Figure 2(*a*), Resolution: 100 Kb) shows that both model-based imputation methods, RWR and Higashi (0 nbr), which use spatial correlation between bin pairs enhance the precision of called DCCs. These findings demonstrate that only RWR and Higashi (0 nbr) improve both the precision and recall of called DCCs from single-cell Hi-C datasets with higher coverage and at domain-scale resolution such as 100 Kb.

3.3. The effect of imputation is dependent on the resolution of the analysis

Depending on the coverage of the dataset and the resolution of the analysis, scHi-C counts have different sparsity levels. For example, considering one dataset with a specific coverage, contact counts at 10 Kb, 100 Kb, and 1 Mb resolutions are highly, moderately, and lowly sparse, respectively. To assess the effect of imputation at different resolutions, we analyzed *Lee2023* dataset at three different resolutions (10 Kb, 100 Kb, and 1 Mb) because it has the highest coverage among our datasets and its corresponding bulk data exist to evaluate called DCCs.

Our results show that current imputation methods are not able to solve the sparsity of finer-resolution scHi-C contact maps like 10 Kb to improve downstream analysis tasks such as DCC identification (Figures 2(*a*) and 2(*b*)). Figure 2(*b*) shows that t-statistics calculated from imputed contact map with scVI-3D is almost noise and does not capture logFC pattern from bulk data. RWR and Higashi (0 nbr)'s t-statistics are slightly similar to bulk logFCs, however, there are many differences such that ROC curve is very close to the identity line (Figure 2(*a*)).

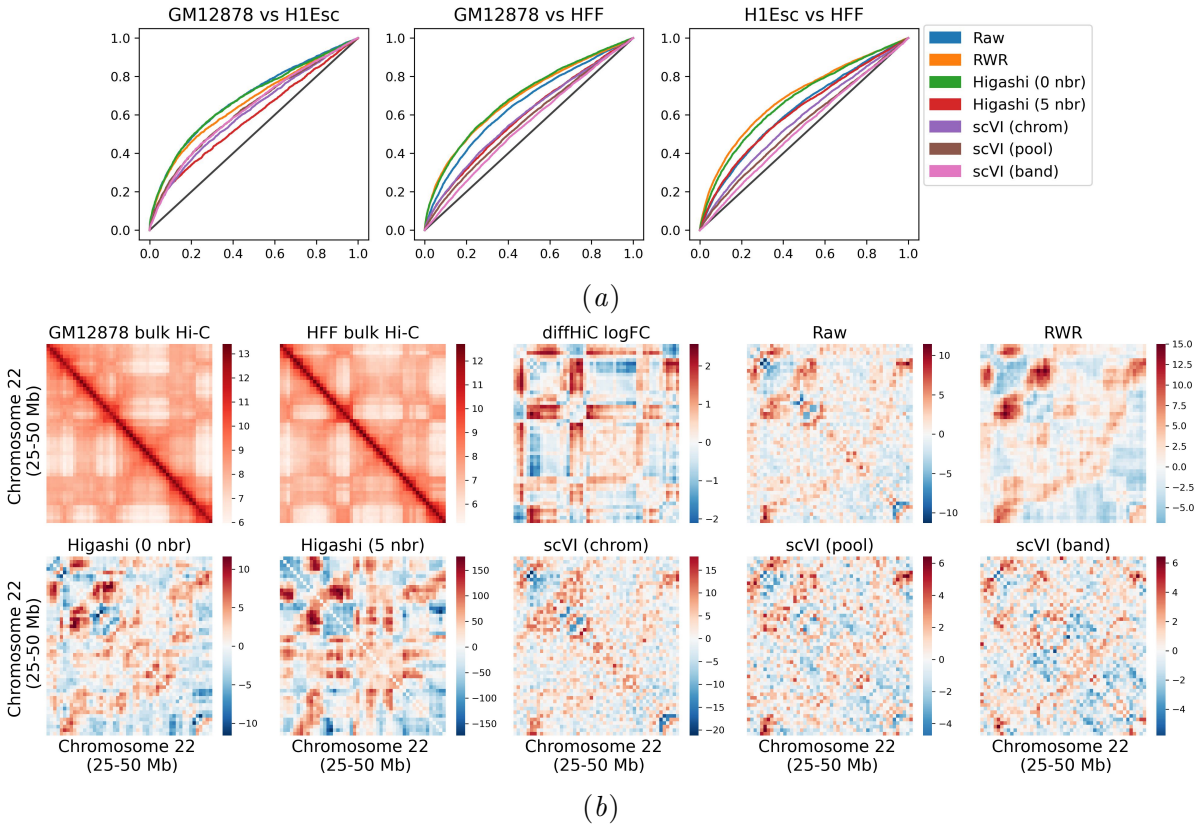


Figure 3: Evaluation of called DCCs from *Kim2020*. (a) Comparison of ROC curves for called DCCs by different imputation approaches from three pairs of cell lines. (b) The heatmap of bulk Hi-C contact maps for GM12878 and HFF cell lines and diffHiC log fold-change (logFC) and single-cell t-statistics from the comparison of these two cell lines. The imputed counts are normalized with distance-based normalization. All experiments are done on chromosomes {11..18} at 500 Kb resolution.

On the other hand, imputation does not significantly improve the performance of DCC caller significantly at 1 Mb resolution (Figure 2(a)). The reason might be the invalidity of spatial correlation assumption at coarser resolution (26) such that model-based imputation approaches, Higashi and RWR, induce more false positives and the overall performance does not improve compared to using raw counts.

Furthermore, we analyzed *Kim2020* dataset with the lowest coverage at 500 Kb resolution, for which corresponding bulk ground truth exists too. We assessed the ROC curve of called DCCs given diffHiC results on their corresponding bulk data for three pairs of cell lines, (GM12878, H1Esc), (GM12878, HFF), and (H1Esc, HFF). Figure 3(a) also shows that model-based imputation approaches, Higashi and RWR, outperform other imputation approaches, particularly for (GM12878, HFF) and (H1Esc, HFF) comparisons where two cell lines are more different from each other.

To better understand the impact of imputations, we conducted a comparison between the heatmap of bulk LFCs and single-cell t-statistics resulting from different imputation approaches for all three comparisons (Figures 3(b), 6(a) and 6(b)). The t-statistic pattern of both RWR and Higashi (0 nbr) appears smoother compared to the t-statistic pattern of scVI-3D and closely

resembles the bulk LFC pattern. These results suggest that model-based imputation approaches, which assume the existence of correlations between spatial neighbors, are essential due to 3D nature of the chromatin.

4. Discussion and Conclusion

Here, we benchmark the effect of recent scHi-C imputation methods (25; 26) on the precision and recall of identified significant DCCs from scHi-C datasets. We demonstrate that imputation is essential and that model-based imputation, Higashi (25) and RWR (27; 11), enhances recall, precision, and reproducibility when applied to datasets (10; 7; 11) with relatively moderate sparsity level.

While imputation methods improve the accuracy by sharing information between either neighbor bin pairs (27; 25; 11) or neighbor cells (25; 26), their performance highly depends on the existence of spatial correlation between neighbor bin pairs and the accuracy of identified neighbor cells. For example, Higashi trains one model given contact counts from all chromosomes, and sharing information from a whole dataset results in good separation of cells from distinct cell types in the cell embedding space and reliable neighbor cells and imputed contact counts. However, scVI-3D trains a model per pool, and the quality of imputed contact counts depends on the chromosome's size and the pool's sparsity. For example, a comparison of the Silhouette scores of scVI-3D embeddings (given ground truth cell annotations) for different pools and their concatenation across one chromosome and more chromosomes shows that a single model on one pool of a small chromosome is poorly trained and we cannot reliably use its learned parameters to generate imputed contact counts (Figures 7(a) and 7(c)).

Another issue with the imputation step is its memory usage and training time which increases severely at finer resolution, particularly for learning-based approaches, Higashi and scVI-3D. One future direction is to develop a faster and more precise imputation method for the scHi-C data that improves its downstream analysis tasks including differential compartment, TAD, and chromatin contact analysis.

Furthermore, an ideal imputation method should preserve the biological variability between cells from the same group. Otherwise, the t-test calls negligible differences significant because of the low standard error between cells. For example, Higashi (n nbr) which smooths single-cell contact maps according to their neighbor cells in the cell embedding space, decreases the biological variability between cells severely (Figures 4(a) and 4(b)); thus applying a t-test on its imputed contact counts results in calling almost all candidate DCCs significant. We should either choose the proper imputation method or a suitable statistical test to avoid such false positives.

Finally, the accuracy and specificity of called significant DCCs can be further evaluated in two ways; first, using more scHi-C datasets with available corresponding bulk datasets, and second, based on their association with other cellular activities such as gene expression and epigenomic signals. This type of evaluation gets more valuable with increased profiling of different modalities from the same tissues (10; 22; 13; 28). The development of an analysis tool for investigating the association between differential signals from different modalities and evaluation functions for quantifying such associations would be another direction for interpreting significant DCCs and evaluating DCC callers from scHi-C data.

Acknowledgments

This work was funded by Compute Canada (kdd-445), NSERC (RGPIN/06150-2018), and Health Research BC (SCH-2021-1734). We thank reviewers for their valuable feedback.

References

[1] Betul Akgol Oksuz, Liyan Yang, Sameer Abraham, Sergey V Venev, Nils Krietenstein, Krishna Mohan Parsi, Hakan Ozadam, Marlies E Oomen, Ankita Nand, Hui Mao, et al. Systematic evaluation of chromosome conformation capture assays. *Nature methods*, 18(9):1046–1055, 2021.

[2] Tallulah S Andrews and Martin Hemberg. False signals induced by single-cell imputation. *F1000Research*, 7, 2018.

[3] Boyan Bonev, Netta Mendelson Cohen, Quentin Szabo, Lauriane Fritsch, Giorgio L Papadopoulos, Yaniv Lubling, Xiaole Xu, Xiaodan Lv, Jean-Philippe Hugnot, Amos Tanay, et al. Multiscale 3d genome rewiring during mouse neural development. *Cell*, 171(3):557–572, 2017.

[4] Kate B Cook, Borislav H Hristov, Karine G Le Roch, Jean Philippe Vert, and William Stafford Noble. Measuring significant changes in chromatin conformation with accost. *Nucleic acids research*, 48(5):2303–2311, 2020.

[5] Jesse R Dixon, Inkyung Jung, Siddarth Selvaraj, Yin Shen, Jessica E Antosiewicz-Bourget, Ah Young Lee, Zhen Ye, Audrey Kim, Nisha Rajagopal, Wei Xie, et al. Chromatin architecture reorganization during stem cell differentiation. *Nature*, 518(7539):331–336, 2015.

[6] Mohamed Nadhir Djekidel, Yang Chen, and Michael Q Zhang. Find: differential chromatin interactions detection using a spatial poisson process. *Genome research*, 28(3):412–422, 2018.

[7] Hyeon-Jin Kim, Galip Gürkan Yardımcı, Giancarlo Bonora, Vijay Ramani, Jie Liu, Ruolan Qiu, Choli Lee, Jennifer Hesson, Carol B Ware, Jay Shendure, et al. Capturing cell type-specific chromatin compartment patterns by applying topic modeling to single-cell hi-c data. *PLoS computational biology*, 16(9):e1008173, 2020.

[8] Nils Krietenstein, Sameer Abraham, Sergey V Venev, Nezar Abdennur, Johan Gibcus, Tsung-Han S Hsieh, Krishna Mohan Parsi, Liyan Yang, René Maehr, Leonid A Mirny, et al. Ultrastructural details of mammalian chromosome architecture. *Molecular cell*, 78(3):554–565, 2020.

[9] Anton Krumm and Zhijun Duan. Understanding the 3d genome: emerging impacts on human disease. In *Seminars in cell & developmental biology*, volume 90, pages 62–77. Elsevier, 2019.

[10] Dong-Sung Lee, Chongyuan Luo, Jingtian Zhou, Sahaana Chandran, Angeline Rivkin, Anna Bartlett, Joseph R Nery, Conor Fitzpatrick, Carolyn O'Connor, Jesse R Dixon, et al. Simultaneous profiling of 3d genome structure and dna methylation in single human cells. *Nature methods*, 16(10):999–1006, 2019.

[11] Lindsay Lee, Miao Yu, Xiaoqi Li, Chenxu Zhu, Yanxiao Zhang, Hongyu Yu, Ziyin Chen, Shreya Mishra, Bing Ren, Yun Li, et al. Snaphic-d: a computational pipeline to identify differential chromatin contacts from single-cell hi-c data. *Briefings in Bioinformatics*, 24(5): bbad315, 2023.

[12] Guoliang Li, Xiaoan Ruan, Raymond K Auerbach, Kuljeet Singh Sandhu, Meizhen Zheng, Ping Wang, Huay Mei Poh, Yufen Goh, Joanne Lim, Jingyao Zhang, et al. Extensive promoter-centered chromatin interactions provide a topological basis for transcription regulation. *Cell*, 148(1):84–98, 2012.

[13] Hanqing Liu, Jingtian Zhou, Wei Tian, Chongyuan Luo, Anna Bartlett, Andrew Aldridge, Jacinta Lucero, Julia K Osteen, Joseph R Nery, Huaming Chen, et al. Dna methylation atlas of the mouse brain at single-cell resolution. *Nature*, 598(7879):120–128, 2021.

[14] Romain Lopez, Jeffrey Regier, Michael B Cole, Michael I Jordan, and Nir Yosef. Deep generative modeling for single-cell transcriptomics. *Nature methods*, 15(12):1053–1058, 2018.

[15] Aaron TL Lun and Gordon K Smyth. diffhic: a bioconductor package to detect differential genomic interactions in hi-c data. *BMC bioinformatics*, 16(1):1–11, 2015.

[16] 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.

[17] Takashi Nagano, Yaniv Lubling, Csilla Várnai, Carmel Dudley, Wing Leung, Yael Baran, Netta Mendelson Cohen, Steven Wingett, Peter Fraser, and Amos Tanay. Cell-cycle dynamics of chromosomal organization at single-cell resolution. *Nature*, 547(7661):61–67, 2017.

[18] Vijay Ramani, Xinxian Deng, Ruolan Qiu, Kevin L Gunderson, Frank J Steemers, Christine M Disteche, William S Noble, Zhijun Duan, and Jay Shendure. Massively multiplex single-cell hi-c. *Nature methods*, 14(3):263–266, 2017.

[19] Suhas SP Rao, Miriam H Huntley, Neva C Durand, Elena K Stamenova, Ivan D Bochkov, James T Robinson, Adrian L Sanborn, Ido Machol, Arina D Omer, Eric S Lander, et al. A 3d map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell*, 159(7):1665–1680, 2014.

[20] John C Stansfield, Kellen G Cresswell, and Mikhail G Dozmorov. multihiccompare: joint normalization and comparative analysis of complex hi-c experiments. *Bioinformatics*, 35(17): 2916–2923, 2019.

[21] Longzhi Tan, Dong Xing, Chi-Han Chang, Heng Li, and X Sunney Xie. Three-dimensional genome structures of single diploid human cells. *Science*, 361(6405):924–928, 2018.

[22] Longzhi Tan, Wenping Ma, Honggui Wu, Yinghui Zheng, Dong Xing, Ritchie Chen, Xiang Li, Nicholas Daley, Karl Deisseroth, and X Sunney Xie. Changes in genome architecture and transcriptional dynamics progress independently of sensory experience during post-natal brain development. *Cell*, 184(3):741–758, 2021.

[23] Miao Yu, Yun Li, and Ming Hu. Mapping chromatin loops in single cells. *Trends in Genetics*, 2022.

[24] R Zhang, Y Zou, and J Ma. Hyper-sagnn: a self-attention based graph neural network for hypergraphs. In *International Conference on Learning Representations (ICLR)*, 2020.

[25] Ruochi Zhang, Tianming Zhou, and Jian Ma. Multiscale and integrative single-cell hi-c analysis with higashi. *Nature biotechnology*, 40(2):254–261, 2022.

[26] Ye Zheng, Siqi Shen, and Sündüz Keleş. Normalization and de-noising of single-cell hi-c data with bandnorm and scvi-3d. *Genome biology*, 23(1):1–34, 2022.

[27] Jingtian Zhou, Jianzhu Ma, Yusi Chen, Chuankai Cheng, Bokan Bao, Jian Peng, Terrence J Sejnowski, Jesse R Dixon, and Joseph R Ecker. Robust single-cell hi-c clustering by convolution- and random-walk–based imputation. *Proceedings of the National Academy of Sciences*, 116 (28):14011–14018, 2019.

[28] Chenxu Zhu, Yanxiao Zhang, Yang Eric Li, Jacinta Lucero, M Margarita Behrens, and Bing Ren. Joint profiling of histone modifications and transcriptome in single cells from mouse brain. *Nature methods*, 18(3):283–292, 2021.

Appendix A. Data sources, statistics and preprocessing

All downloaded Hi-C and scHi-C datasets (Table 1), except bulk Hi-C data for mESC, NPC, and HFF were processed and mapped to hg19 or mm10 and stored in tab-separated, pairs, or cool format. The bulk Hi-C data for HFF was mapped to hg38, and we used HiCLift [1] to lift it to hg19 to compare it with other datasets. The bulk mESC and NPC data were unbinned genomic tracks, and we used misha package to bin them. First, we followed a vignette [2] to create a misha database for mm10 assembly. Then, we copied the downloaded track data to misha database's track subdirectory, and binned tracks with 'gextract' command.

Table 1: Data sources.

Data type	dataset	link
scHi-C	Lee2019	GSE130711 from GEO
	Kim2020	sci-Hi-C .matrix files
	Lee2023	GSE210585 from GEO
bulk Hi-C	GM12878	GSE63525 from GEO
	HFF	4DNES2R6PUEK from 4DN portal
	H1Esc	4DNESRJ8KV4Q from 4DN portal
	mESC, mNPC	GSE96107 from GEO

Table 2: scHi-C datasets statistics.

Dataset	# cells	average total contact counts	average off-diagonal contact counts	group 1	group 2
Lee2019	4238	1.08 M	190 K	149 Astro cells from batch 190315_21yr	128 MG cells from batch 190315_21yr
				101 Astro cells from batch 190315_29yr	112 MG cells from batch 190315_29yr
				149 Astro cells from batch 190315_21yr	101 Astro cells from batch 190315_29yr
Kim2020	8023	11.4 K	5.7 K	2784 GM12878 cells	908 HFF cells
				2784 GM12878 cells	2436 H1Esc cells
Lee2023	282	1.05 M	191 K	94 mESC cells	188 mNPC cells

1. https://github.com/XiaoTaoWang/HiCLift#installation
2. https://rdrr.io/cran/misha/f/vignettes/Genomes.Rmd

Table 3: Filtering criteria for different resolutions.

Resolution	Excluding filter regions	Including TSS regions	Genomic distance threshold
10 Kb	✓	✓	2 Mb
100 Kb	✓	✗	2 Mb
500 Kb	✓	✗	10 Mb
1 Mb	✓	✗	✗

Appendix B. Supplementary figures

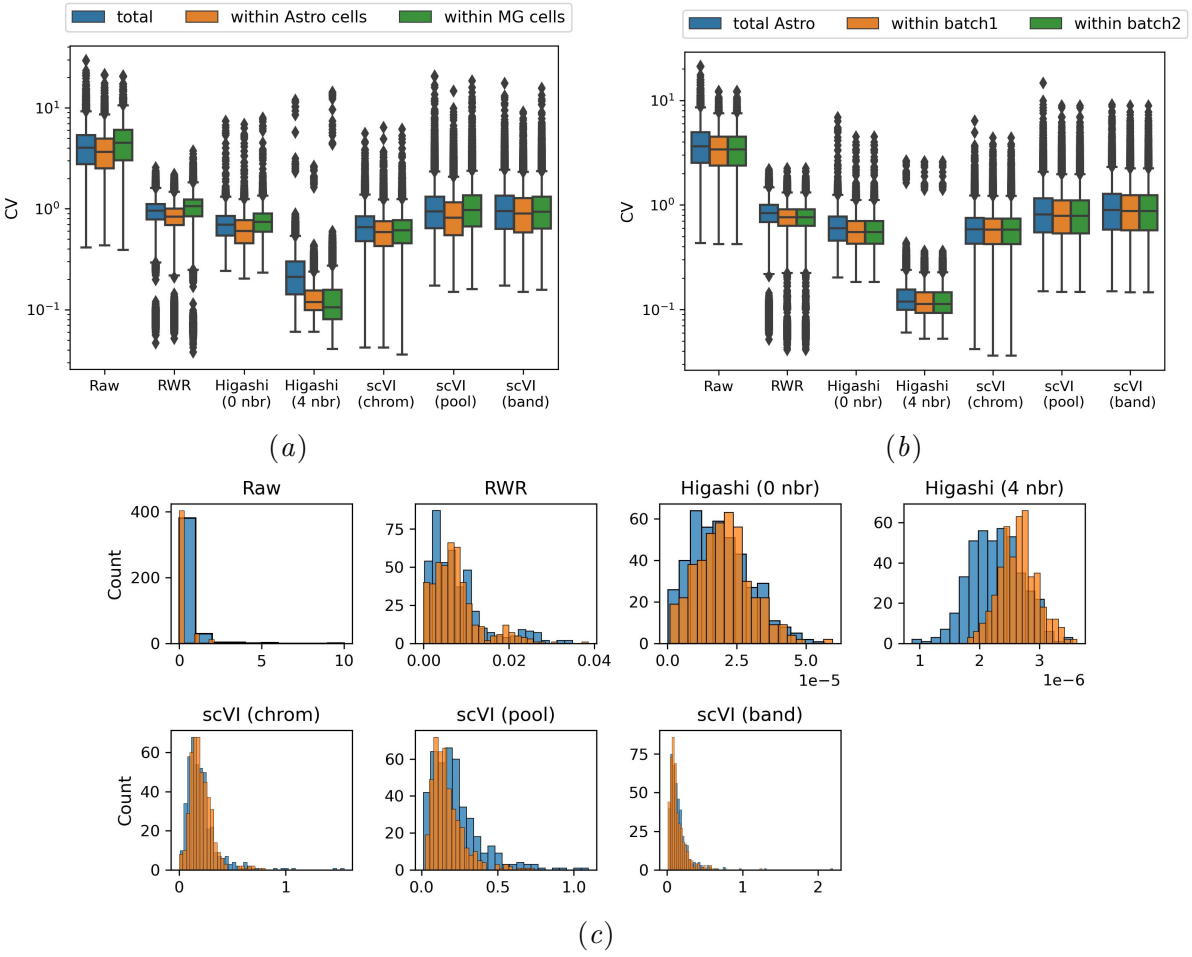


Figure 4: (a) Coefficient of variation (CV) calculated for all the cells from two groups (MG and Astro in this plot) named as 'total', and for each group separately. (b) CV calculated for all Astro cells named as 'total Astro', and for Astro cells in two batches separately. The coefficient of variation is the ratio of standard deviation to the mean of counts. (c) The distribution of imputed counts for a bin pair selected randomly, (chr11:84800000-84900000, chr11:85300000-85400000). Blue and orange colors indicate MG and Astro counts respectively.

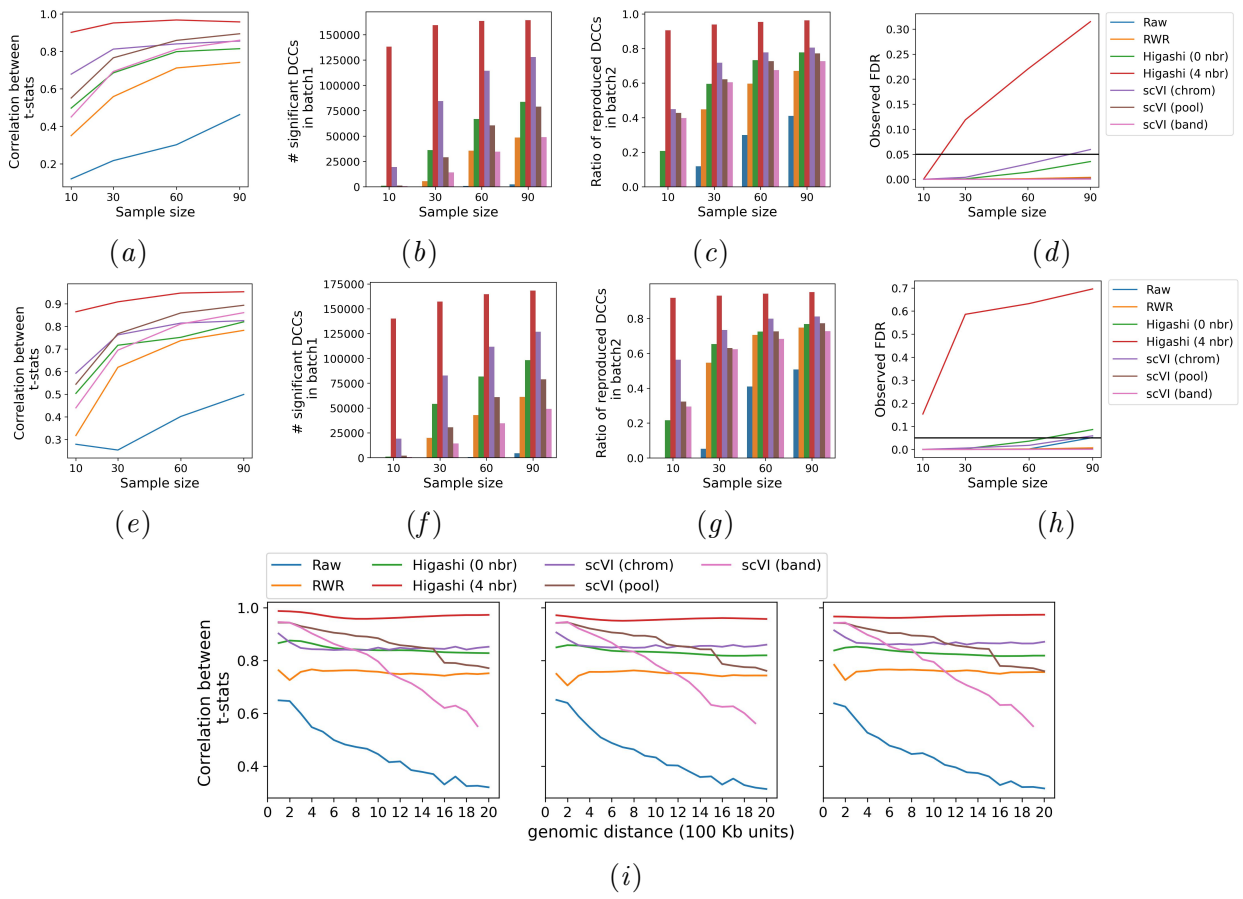
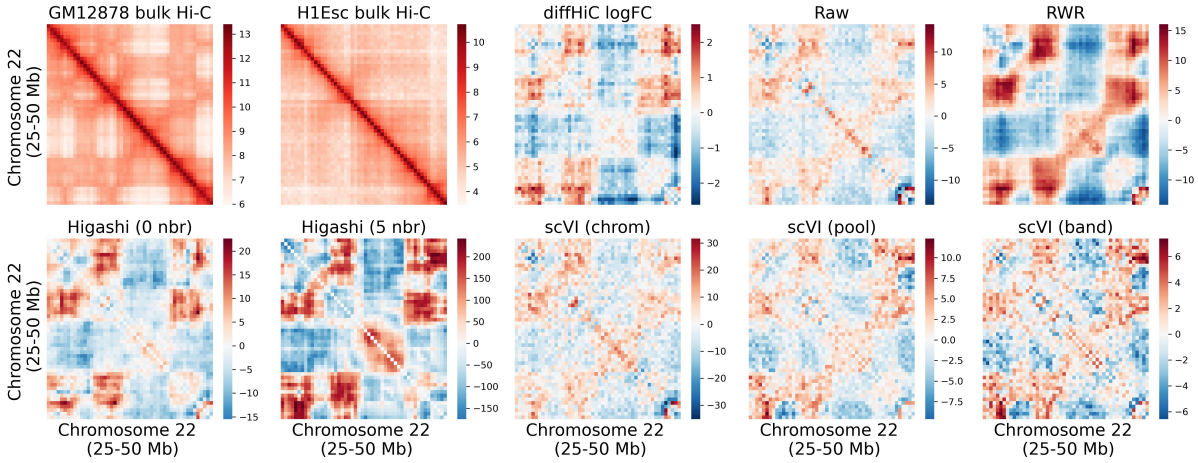
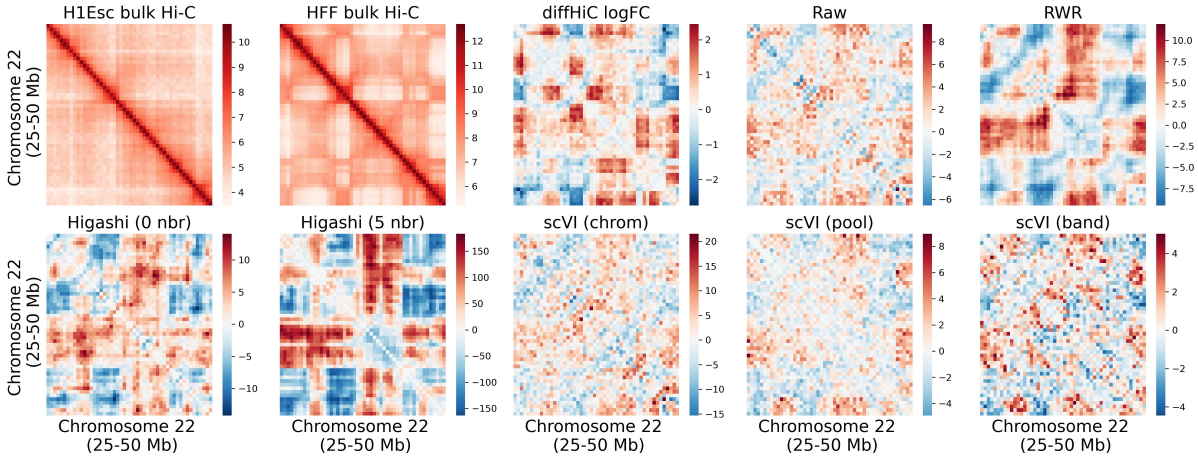


Figure 5: DCC analysis between Astro and MG is done twice for two batches, 190315_21yr and 190315_29yr from *Lee2019* dataset. (b) The correlation between t-statistics from two replicates (b) after total-based normalization and (e) without normalization. The number of called DCCs from batch 1 (c) after total-based normalization and (f) without normalization. The ratio of called DCCs from batch 1 that are reproduced in batch 2 (c) after total-based normalization and (g) without normalization. The observed false discovery rate (FDR) when comparing two groups of Astro cells, one from batch 190315_21yr, and another one from batch 190315_29yr (d) after total-based normalization and (h) without normalization. (i) The correlation between t-statistics from two replicates calculated per genomic distance after distance-based normalization (right), total-based normalization (middle), and without normalization (left). All experiments are done on chromosomes {11..22} at 100 Kb resolution for different numbers of cells in the groups (sample size). (a-h) demonstrates that reproducibility and FDR analysis results are invariant to the normalization step. FDR is only higher without normalization which is expected as unwanted technical variations are called significant. (i) shows the correlation between t-statistics per genomic distance which decreases with increasing the genomic distance as the sparsity increases. All imputation approaches except scVI (band) and scVI(pool) that impute contact counts per band or pool impute bin pairs within different genomic distances similarly.



(a)



(b)

Figure 6: The heatmap of bulk Hi-C contact maps for cell lines and diffHiC log fold-change (logFC) and single-cell t-statistics from the comparison of (a) (GM12878, H1Esc) and (b) (H1Esc, HFF) cell lines. The imputed counts are normalized with distance-based normalization.

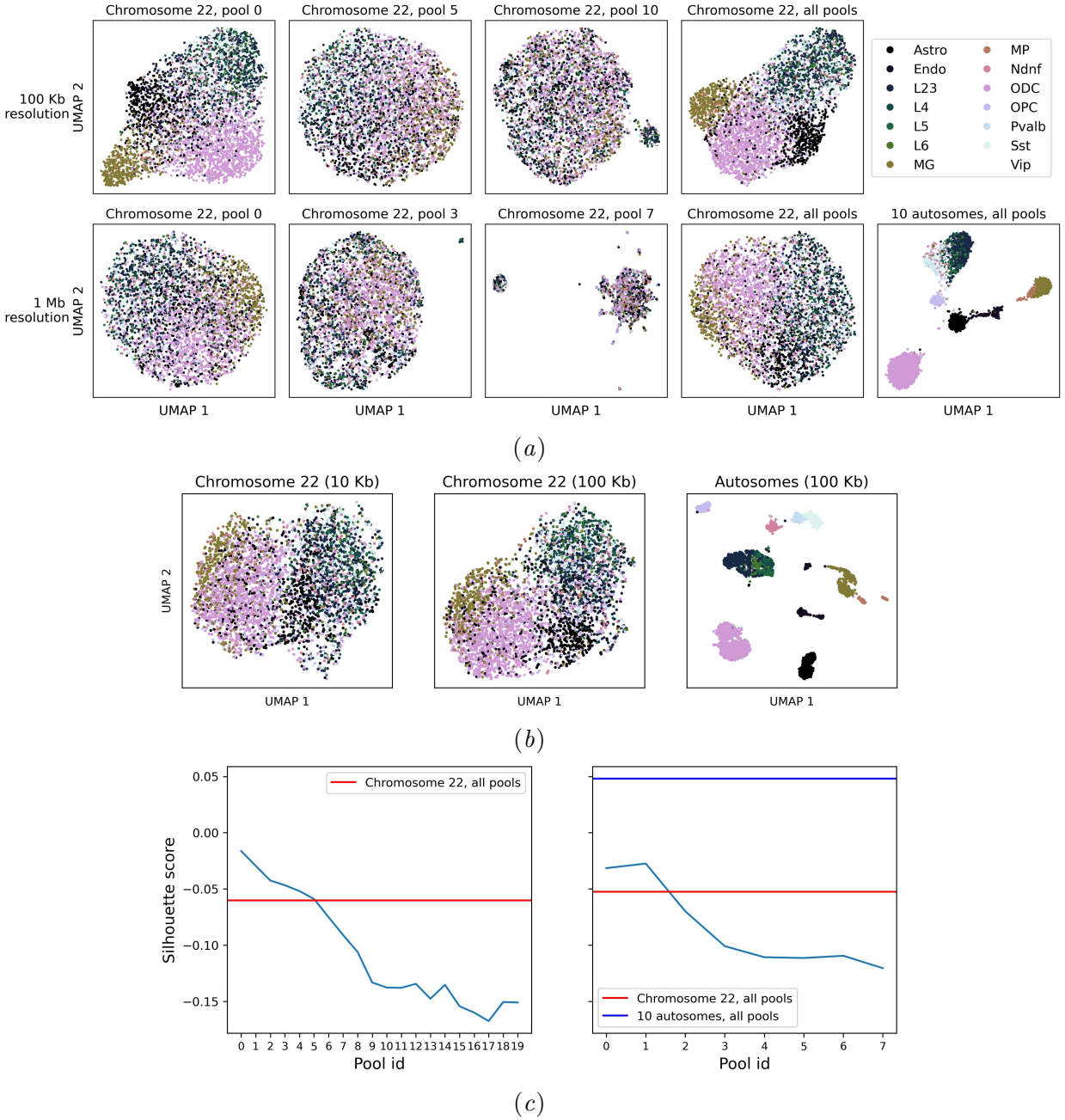


Figure 7: (a) scVI-3D embeddings UMAP for different pools and their concatenation across 1 and 10 chromosomes at two resolutions. (b) Higashi embeddings UMAP at two different resolutions and two training sets, including genomic bins from chromosome 22 or all autosomes. (c) Silhouette Index of scVI-3D embeddings for different pool IDs according to cell annotations. Pool IDs increase by genomic distance. For example, the first pool includes contact counts from the first off-diagonal of a contact matrix, a second pool includes contact counts from a second and third off-diagonal of a contact matrix, etc. The right and left plots are for 100 Kb and 1 Mb resolutions, respectively. All plots are for *Lee2019* dataset.