Statistical Methods for Facilitating 3D Genome Analysis

Ph.D. Defense

Rui Ma

Department of Statistics
University of California, Riverside

May 30, 2024

Outline

Background

Background

- Researchers have explored the organization of the 3D genome.
- The majority of genetic information resides within the nucleus.
- Hi-C technique:
  - is a high-throughput method which explores the 3D genome architecture by extracting the interactions for all loci,
  - uncovers the spatial organization of higher-order chromatin structures,
  - enables high-throughput characterization of chromatin interactions.

Background

- Typical workflow of Hi-C sequencing and Hi-C contact map example:

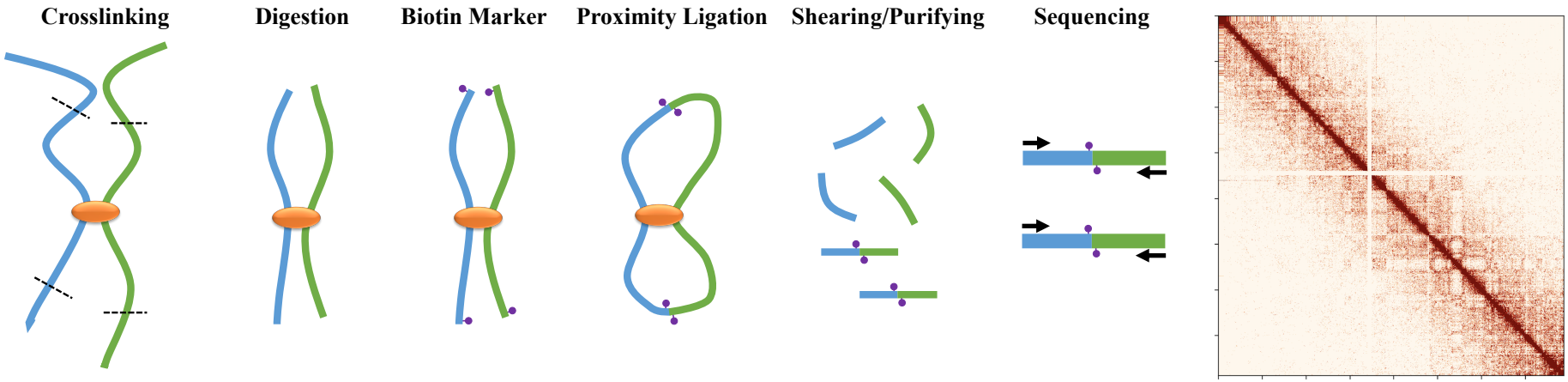


Figure: Illustration of Hi-C Sequencing and Example of Hi-C Contact Map.

Background

- Investigation in 3D genome organization features of Hi-C data:

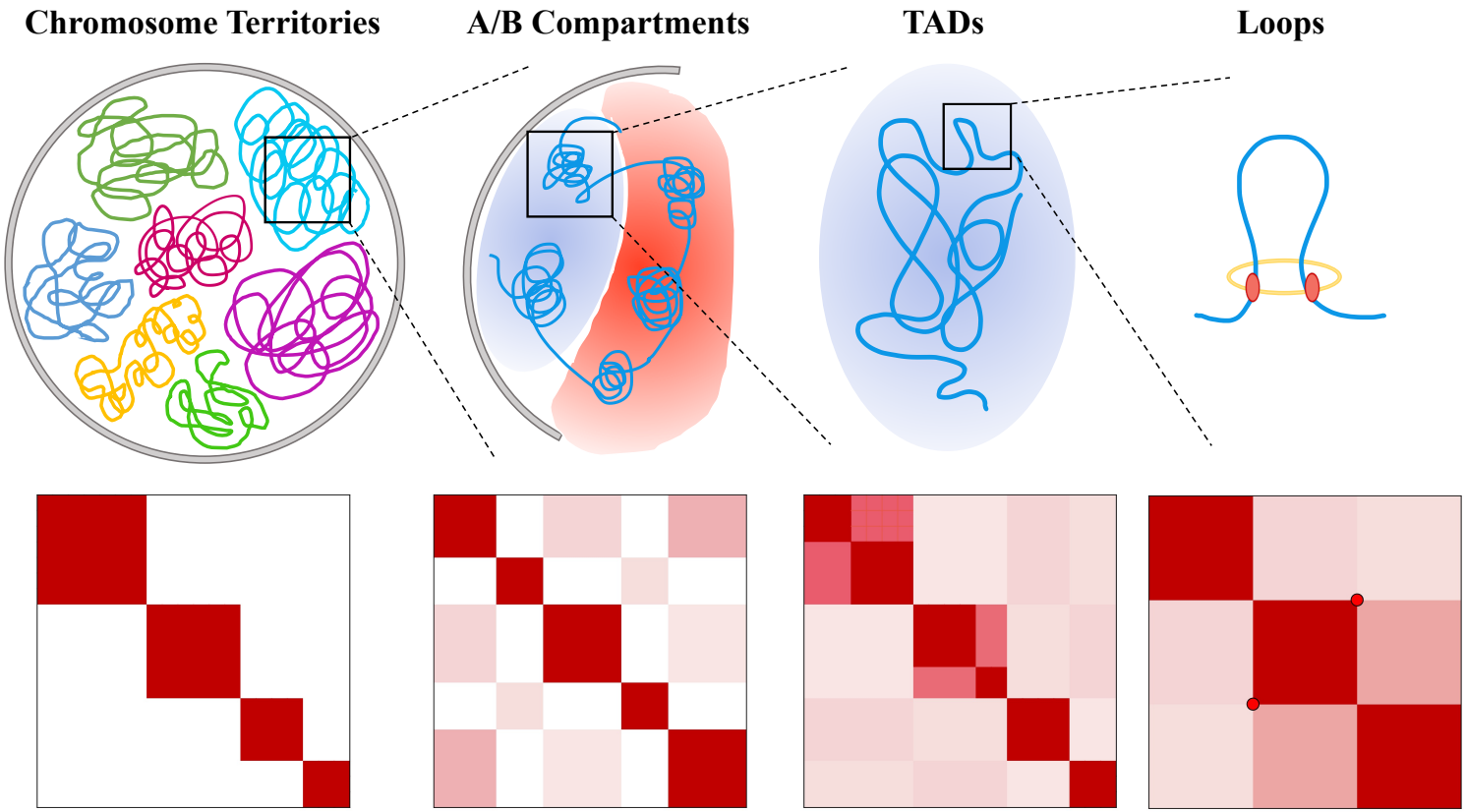


Figure: 3D Genome Structure and Contact Map.

Background

- However, bulk Hi-C techniques cannot investigate 3D genome characterization at single-cell resolution.
- Single-cell techniques:
  - enable the study of multi-scale spatial genome organization at single-cell resolution,
  - provide opportunities to reveal the dynamics and heterogeneity of chromatin conformation.
- Three categories:
  - Proximity ligation
  - Ligation-free
  - Imaging
- Co-assays profiling:
  - scHi-C and DNA methylation
  - scHi-C and scRNA-seq

Background

- Computational approaches are crucial in exploring the hierarchical organization of chromatin in single-cell Hi-C data.
- Two-part analysis process: Data pre-processing & Downstream analysis
- Downstream analysis: A/B compartments assignment, TAD-like boundary detection, and loop detection

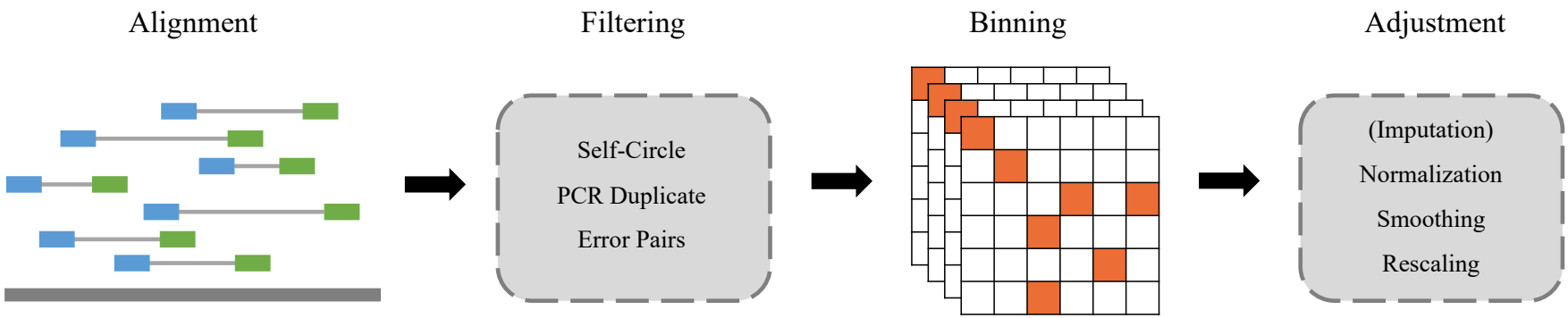


Figure: Illustration of scHi-C Pre-processing Workflow.

scHiCPRSiM: single-cell Hi-C Practical and Rational SiMulator

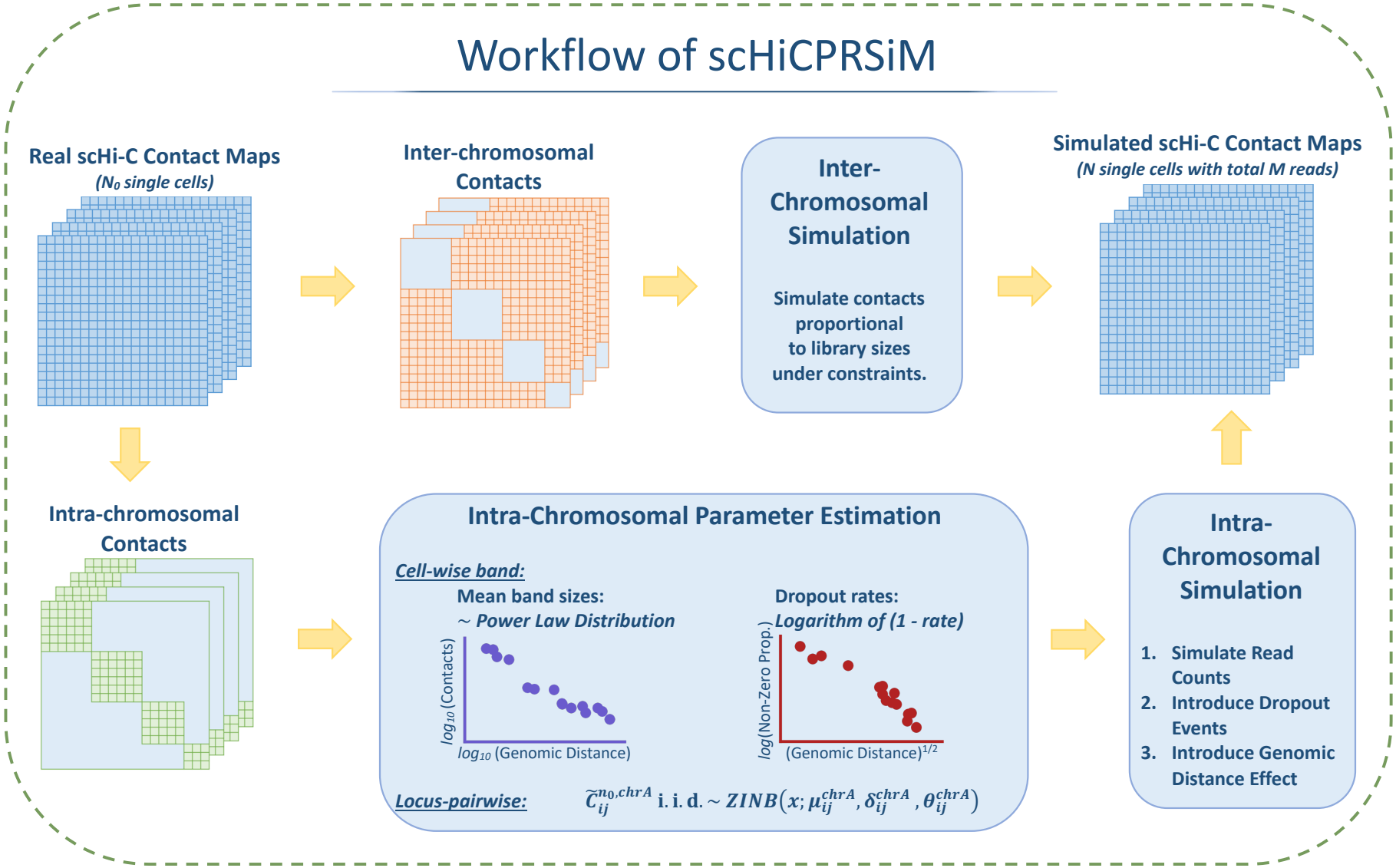
Motivation

- Single-cell Hi-C technologies:
  - provide opportunities to explore 3D genome structure at single-cell level,
  - facilitate integrative studies in gene regulation,
  - motivate computationalists to develop analysis tools.
- High cost in scHi-C sequencing limits the number of scHi-C datasets.
- Simulation, emulation of real scHi-C Data,
  - can help biologist to design cost-effective experiments,
  - is necessary during benchmarking and evaluation.

Previous Work

- Single-cell Hi-C Simulator:
  - Downsampling Method from scHiCluster (Zhou et al., 2019)
    - uses the pseudo-bulk Hi-C contact matrix
    - has strict constrains of library sizes and sparsity
  - scHi-CSim (Fan et al., 2023)
    - generates single-cell fragment-interactions
    - learns properties by merging adjacent single cells
- Issues: low variation, low sparsity level
- Single-cell RNA-seq Simulator: scDesign (Li and Li, 2019)
  - overcomes the dropout events
  - considers gene-wise parameters (Gamma-Normal mixture model)
- Proposed Method: a ZINB-based statistical simulator, scHiCPRSiM, for scHi-C data inspired by scDesign

Proposed Method



Proposed Method

- Given real scHi-C contact matrices from one cell-type/state with N_0 cells
- Goal: to generate a new series of scHi-C count matrices with N cells and a total of M reads
- Assume each generated single cell n mimics the genomic features of one cell g_n in original dataset, that is

$$g_n \sim Unif(1, ..., N_0).$$

Proposed Method

- Chromosomal parameters:
  - intra/inter-chromosomal (cis/trans) read proportions: $R = (\hat{p}_{chr1}, ..., \hat{p}_{chrX}, \hat{p}_{trans})$
  - simulates total cis/trans-read counts by

$$(S_{chr1}, ..., S_{chrX}, S_{trans}) \sim Multinomial(M, R)$$

- Cell-wise parameters:
  - For each real cell n_0, calculates
    - library size: $\hat{S}^{n_0}$
    - proportions of cis/trans contact: $p^{n_0,chrxx}$, $p^{n_0,trans}$
  - For nth simulated cell, estimates
    - library size of the cell, S_n, using non-parametric kernel density
    - library size of cis/trans as

$$s^{n,chrxx} = S_n * p^{g_n,chrxx} \text{ and } s^{n,trans} = S_n * p^{g_n,trans}$$

Proposed Method

For inter-chromosomal (trans) contacts in each simulated cell,

- simulated contacts are proportional to real data read counts constrained by the total trans contacts and the trans library sizes for each simulated cells, S_{trans} and $s^{n,trans}$.

Overview of Intra-Chromosomal (cis) Simulation Procedure:

- Step 1: Parameter estimation
  - Cell-wise band parameters: to capture the genomic distance effects
  - Locus-pairwise parameters: to handle the dropout events
- Step 2: Primary matrices simulation
- Step 3: Final matrices simulation

Proposed Method: Intra-chromosomal

Single-cell Intra-chromosomal Contacts Simulation:

- For a specific chromosome xx with chromosome length L_{chrxx},
- Goal: to generate a series of count matrices with N cells and a total of S_{chrxx} reads.
- $v \in \{0, 1, 2, ..., L_{chrxx} - 1\}$: bands (genomic distances)
- (i, j): locus-pairs with genomic distance $j - i$
- $C_{i,j}^{n_0,chrxx}$: locus-pair contacts in the contact matrix of n_0th cell

Proposed Method: Intra-chromosomal

Step 1: Estimate the parameters

- Cell-wise band parameters:
  - Mean band size for band v: $\hat{m}_v^{n_0,chrxx}$, $v \in \{0, 1, 2, ..., L_{chrxx} - 1\}$
  - Fit using power law

$$\hat{m}_v^{n_0,chrxx} \sim \alpha^{n_0,chrxx} v^{\beta^{n_0,chrxx}} \tag{1}$$

  - Band dropout rate for band v: $\hat{q}_v^{n_0,chrxx}$, $v \in \{0, 1, 2, ..., L_{chrxx} - 1\}$
  - Fit using

$$log(1 - \hat{q}_v^{n_0,chrxx}) = \gamma^{n_0,chrxx} \sqrt{v} + \eta^{n_0,chrxx}. \tag{2}$$

- Cell-wise parameter bundle:
  $\hat{B}^{n_0,chrxx} = (\hat{\alpha}^{n_0,chrxx}, \hat{\beta}^{n_0,chrxx}, \hat{\gamma}^{n_0,chrxx}, \hat{\eta}^{n_0,chrxx})$

Proposed Method: Intra-chromosomal

Step 1: Estimate the parameters

- Locus-pairwise parameters:
  - For bin pair (i, j) in band v, the read count is normalized by

$$\tilde{C}_{i,j}^{n_0,chrxx} = \left\lceil \frac{C_{i,j}^{n_0,chrxx} mean(\hat{m}_v^{1,chrxx}, ..., \hat{m}_v^{N_0,chrxx})}{\hat{m}_v^{n_0,chrxx}} \right\rceil, \quad (3)$$

  - Assume $\tilde{C}_{i,j}^{1,chrxx}, \ldots, \tilde{C}_{i,j}^{N_0,chrxx} \overset{\text{iid}}{\sim}$ ZINB distribution:

$$P_{ij}^{chrxx}(x) = \theta_{ij}^{chrxx}\mathbb{I}(x=0) + (1-\theta_{ij}^{chrxx})NB(x; \mu_{ij}^{chrxx}, \delta_{ij}^{chrxx}), \quad (4)$$

- Set of locus-pairwise parameters: $(\hat{\theta}_{ij}^{chrxx},\ \hat{\mu}_{ij}^{chrxx},\ \hat{\delta}_{ij}^{chrxx})$

Proposed Method: Intra-chromosomal

Step 2: Simulate primary single-cell Hi-C matrices

- Recall: the nth simulated single cell mimics g_n cell in the original dataset.
- 1. Simulates read counts
  - For each bin pair (i, j), we generate N read counts using

$$\dot{C}_{i,j}^{n,chrxx} \sim NB(\hat{\mu}_{ij}^{chrxx}, \hat{\delta}_{ij}^{chrxx})$$

  - Obtain ideal matrices: $\dot{C}^{1,chrxx}, ..., \dot{C}^{N,chrxx}$.

Proposed Method: Intra-chromosomal

Step 2: Simulate primary single-cell Hi-C matrices

- 2. Introduce dropout events
  - The number of dropout events:

$$k_{ij}^{chrxx} \sim Binomial(N, \hat{\theta}_{ij}^{chrxx})$$

  - Sample k_{ij}^{chrxx} cells from N cells, with probability $\frac{q_{|j-i|}^{n,chrxx}}{\sum_{n=1}^{N} q_{|j-i|}^{n,chrxx}}$
  - scHi-C matrices with dropout events, $\ddot{C}^{1,chrxx}, ..., \ddot{C}^{N,chrxx}$:

$$\ddot{C}_{ij}^{n,chrxx} = \dot{C}_{ij}^{n,chrxx} \mathcal{I}(I_{ij}^{n,chrxx} = 1)$$

Proposed Method: Intra-chromosomal

Step 2: Simulate primary single-cell Hi-C matrices

- 3. Introduce genomic distance effect
  - Adjust the scHi-C matrices by introducing genomic distance effect by

$$\breve{C}_{ij}^{n,chrxx} = \left[\frac{\ddot{C}_{ij}^{n,chrxx} m_{|j-i|}^{n,chrxx}}{mean(\hat{m}_v^{1,chrxx}, ..., \hat{m}_v^{N_0,chrxx})} \right], \tag{5}$$

where $m_v^{n,chrxx} = \alpha^{n,chrxx} v^{\beta^{n,chrxx}}$

Proposed Method: Intra-chromosomal

Step 3: Simulate Final Single-cell Hi-C Matrices

- Expected proportion of each entry in the count matrices by

$$P_{ij}^{n,chrxx} = \frac{s^{n,chrxx}\breve{C}_{ij}^{n,chrxx}}{\sum_{n=1}^{N}\sum_{i=1}^{L_{chrxx}}\sum_{j=i}^{L_{chrxx}} s^{n,chrxx}\breve{C}_{ij}^{n,chrxx}}. \tag{6}$$

- Final set of cis contact matrices are obtained using the multinomial distribution with:
  - probabilities: $P_{ij}^{n,chrxx}$
  - total sequencing depth: S_{chrxx}

Results

- To compare with existing methods, we utilized two scHi-C datasets:
  - mouse-cycle dataset: 1171 dynamic cells with 4 cell stages (Nagano et al., 2017)
  - human brain development dataset: large-scale set of 2438 cells with 14 different cell types (Lee et al., 2019)
- Benchmarking aspects:
  - key statistics: library size, sparsity, etc.
  - cell-type clustering pattern
  - real and synthetic data overlapping

Results

- scHiCPRSiM resembles real scHi-C data and surpasses other methods.

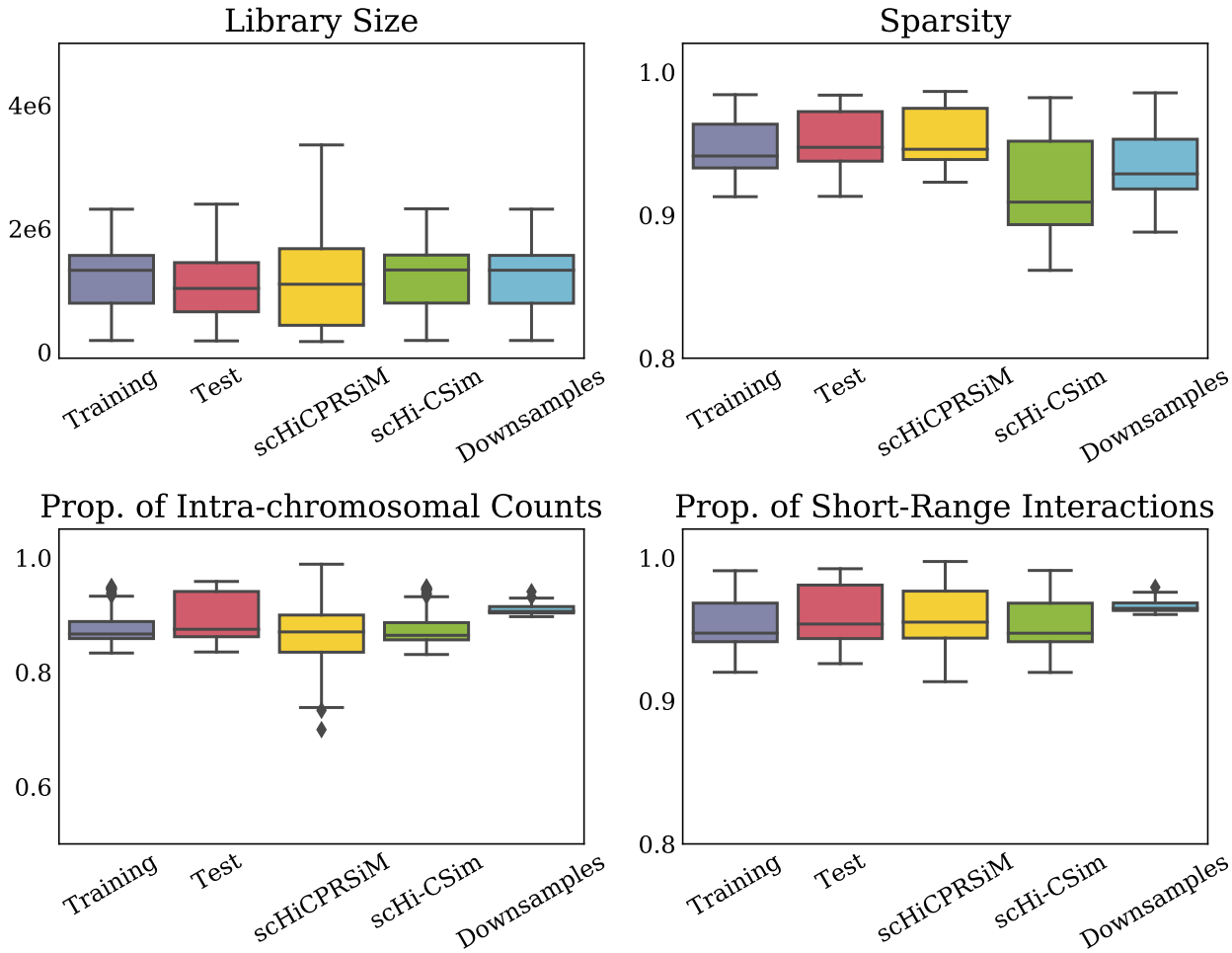


Figure: Boxplots of Summary Statistics for Human Brain Data.

Results

- scHiCPRSiM resembles real scHi-C data and surpasses other methods.
- Transforms high-dimensional data into a low-dimensional space
- How close cells are within each cluster in 2D plots

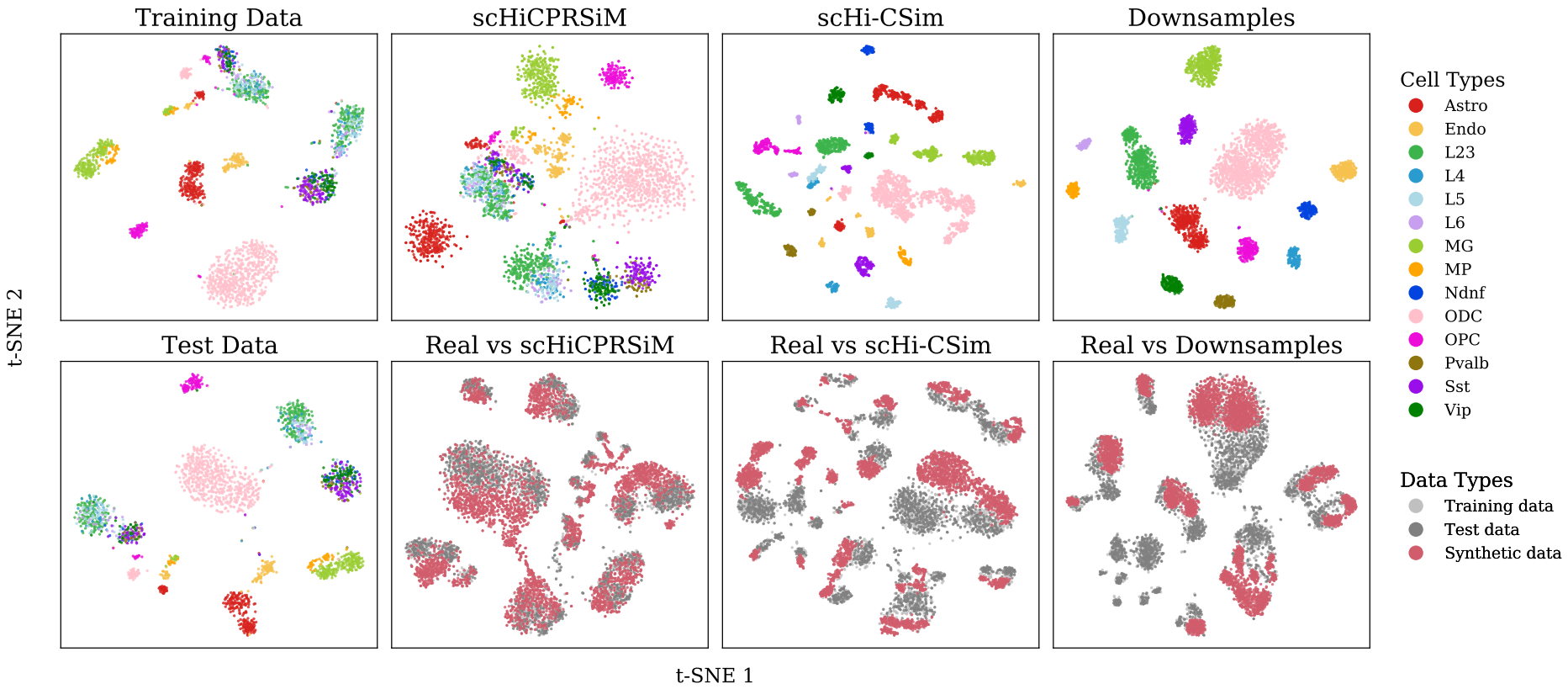


Figure: Embeddings Visualizations of Real and Synthetic Data for Human Brain.

Results

- scHiCPRSiM preserves the compartmental structure observed in real scHi-C data.

	Training vs Test	Test vs Simulation
G1	75.85%	72.74%
Early-S	80.02%	79.05%
Mid-S	76.82%	74.59%
Late-S/G2	71.43%	71.60%

Table: Consistent Rates of A/B Compartments in Mouse-cycle Dataset.

- scHiCPRSiM preserves the natural hierarchical structure from real scHi-C data.

	Training vs Test	Test vs Simulation
G1	70.98%	67.78%
Early-S	72.60%	69.31%
Mid-S	74.02%	71.30%
Late-S/G2	75.30%	72.38%

Table: Consistent Rates of TAD-like Boundaries in Mouse-cycle Dataset.

Results

- scHiCPRSiM can assist the development of computational methods.
- Clustering approaches: BandNorm, CellScale, FastHigashi, scHiCluster, scVI-3D
- ARI and AMI $\longrightarrow$ cell-type separation: expect to be closer to 1.0
- miLISI $\longrightarrow$ batch-mixing: expect to be closer to 2.0

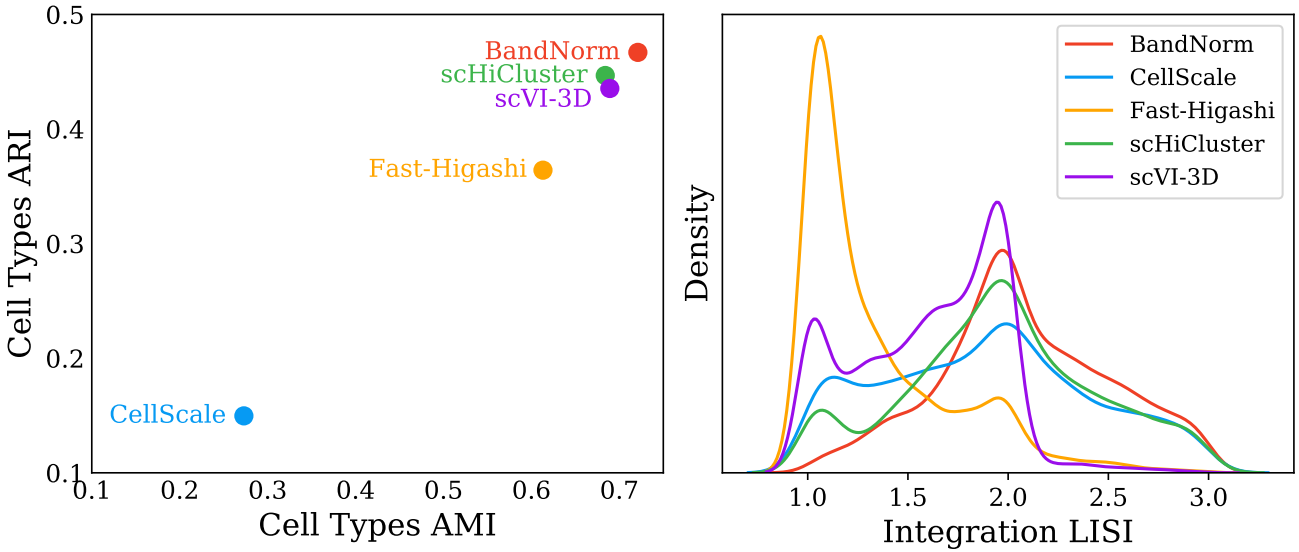


Figure: Clustering Evaluations of Batch Effect Correction Methods.

Results

- scHiCPRSiM guides experimental design in cell clustering.
- Original sequencing depths are highlighted using blue dashed lines.
- Convergence point shows the optimal sequencing depth.

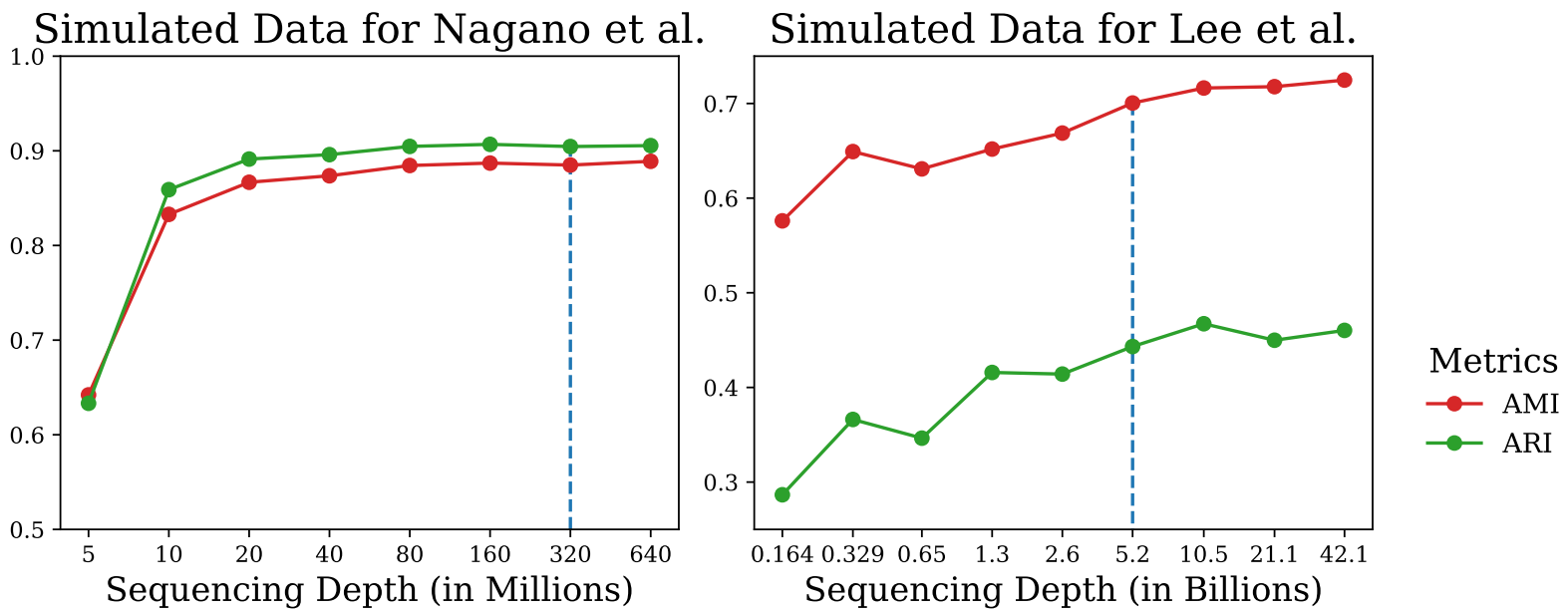


Figure: Clustering Accuracy of scHiCPRSiM Synthetic Data.

Discussion

- Advantages of scHiCPRSiM:
  - Locus-pairwise ZINB models: dropout events and high sparsity
  - Power law relationship: genomic distance variability
  - Cell-wise parameters and models learned locus-pairwisely: cell-to-cell heterogeneity and variation

Discussion

- The capacities of scHiCPRSiM:
  - resembles real scHi-C data and surpasses existing methods
  - maintains the key chromatin structure features
  - guides experimental design in cell clustering at cost-effective level
  - assists in the development of computational methods

SHIM: Single-cell Hi-C Integration using Mutual nearest neighbors

Motivation

- Single-cell Hi-C datasets have expanded significantly in scale.
- The large-scale data enables researchers to:
  - reveal the spatial genome organization, and
  - provide robust insights into developmental processes and differentiation,
  - motivates biological studies with other omics data to unveil the spatial organization's relationship with regulatory elements and cellular functions.
- Batch effects:
  - artificial effects and noises
  - non-biological reasons: different times, different experiments, different laboratories
- Solution: batch effect correction methods (integration)
  - to remove batch effects
  - to enhance statistical power for downstream analyses

Previous Work

- Batch correction has been studies a lot in scRNA-seq
- Similarity-score-based approaches
  - find anchor pairs between datasets to project cells onto latent space
  - Canonical Correlation Analysis (CCA)
  - Mutual Nearest Neighbor (MNN)
- Deep generative models
  - learn the non-linear embeddings
  - Variational Autoencoders (VAE)

Previous Work

- scHi-C integration methods:
  - BandNorm (Zheng et al., 2022) and scHiCluster (Zhou et al., 2019)
    - perform normalization or imputation
    - coupled with Harmony, scRNA-seq method
  - scVI-3D (Zheng et al., 2022)
    - takes batch information in the neural network-modeled distribution
    - estimates zero inflation variables in ZINB model
  - Higashi (Zhang et al., 2022)
    - connects one cell and two bins using hyperedges
    - treats batch label as an additional feature during imputation
- Proposed method: SHIM, a user-friendly and robust batch effect correction tool designed explicitly for chromatin contact maps

Proposed Method

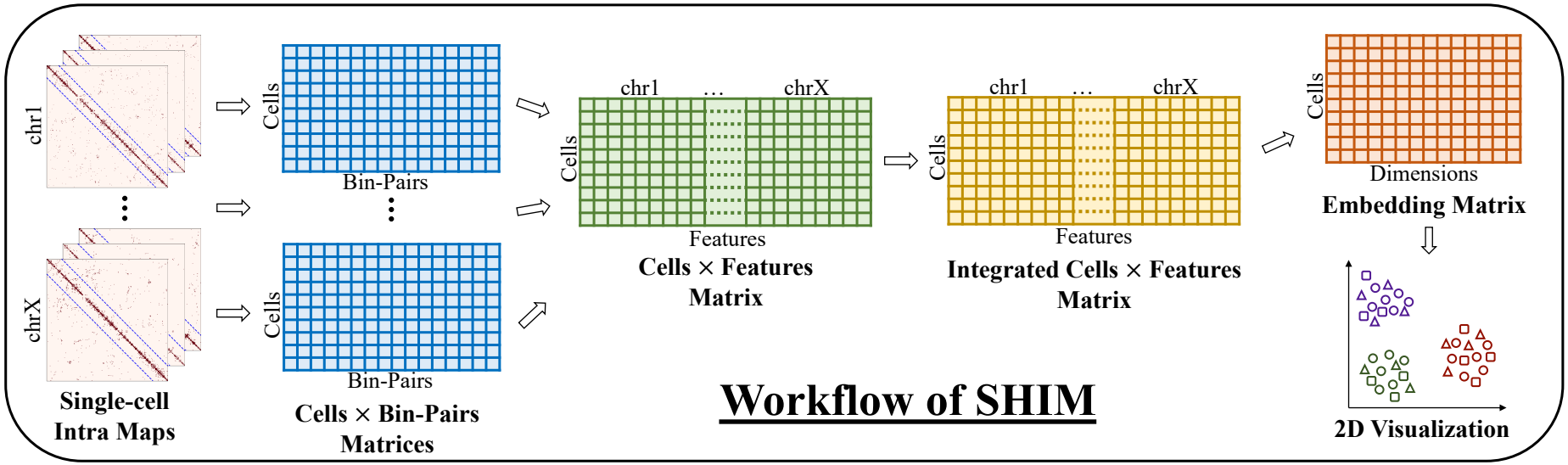


Figure: SHIM Workflow.

Proposed Method

- Mutual Nearest Neighbor (MNN)
  - Assumption: orthogonality of batch effects in data in high dimensional space
  - MNN pairs: observations that are nearest neighbor of each other from two datasets
  - i.e. an observation X from dataset 1 is the neighbor to observation X' in dataset 2, and the reverse is also true.

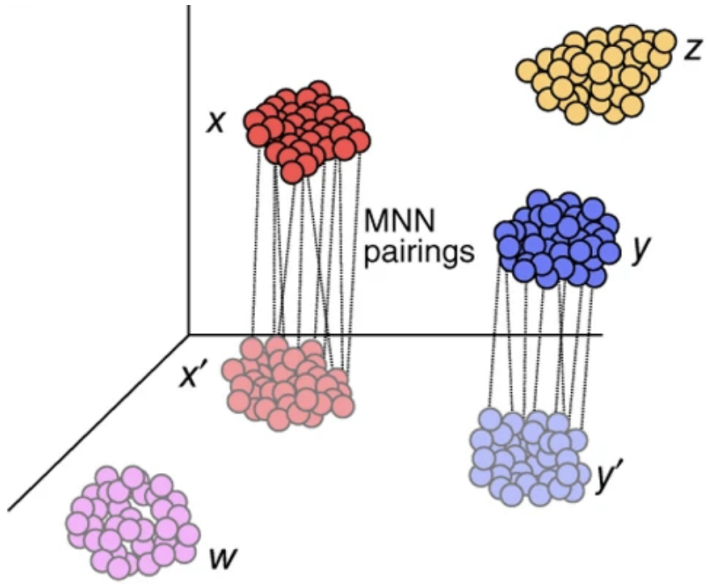


Figure: Illustration of Mutual Nearest Neighbors. (Haghverdi et al., 2018)

Proposed Method

- Using the similarity-based feature matrix, $\tilde{S}$
  - define reference set, D_a, such that

$$a = \text{argmax}_l |D_l|, \tag{7}$$

  - $|.|$ denotes the cardinality of a set.
- For each cell i in D_a,
  - find k nearest neighbors for each datasets D_l, where $l \in \{1, ..., L\}/a$
  - all KNN pairs for D_l: M_l^a
- For each cell j in D_l, $l \in \{1, ..., L\}/a$
  - find k nearest neighbors in D_a, forming M_l

Proposed Method

- Set of MNN pairs for D_a and D_l, for each $l \in \{1, ..., L\}/a$

$$\text{MNN}_l = \{(i, j) \in M_l^a \cap M_l\} \tag{8}$$

- For each $l \in \{1, ..., L\}/a$,
  - obtains a set of pair-wise difference

$$\bar{M}_l = \{\tilde{S}_i - \tilde{S}_j \mid (i, j) \in M_l^a \cap M_l\} \tag{9}$$

  - estimates correction vector m_l as the weighted linear combination of $\bar{M}_l$

$$w_l = e^{-\frac{1}{2\sigma^2}(\|\bar{M}_l\|_2^2)}, \tag{10}$$

  - $\sigma = 15$ and $\|.\|$ represents the pair-wise norm.

Results

Comparison with existing methods in three aspects:

- batch mixing
- cell-type separation
- cell clustering

On datasets with three different batch effects:

- across experiments: human developmental brain dataset (Lee et al., 2019)
- across laboratories: two human developmental brain datasets (Lee et al., 2019 and Heffel et al., 2022)
- across regions: Dip-C mouse developmental brain dataset (Tan et al., 2021)

Results

- SHIM surpasses the state-of-the-art methods with significant improvements in batch mixing.

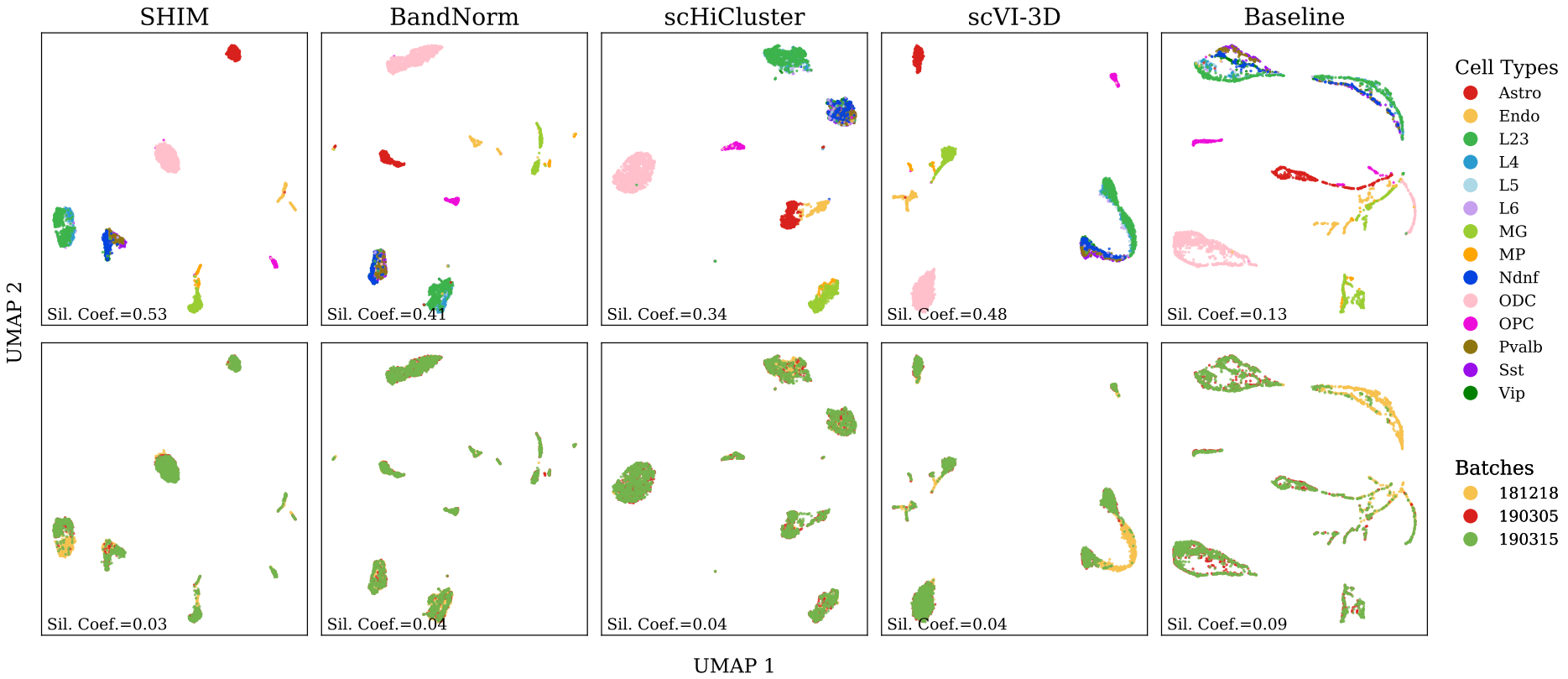


Figure: Embeddings for Non-integrated and Integrated Lee *et al.* Dataset.

Results

- SHIM outperforms the state-of-the-art methods with significant improvements in cell-type isolation.

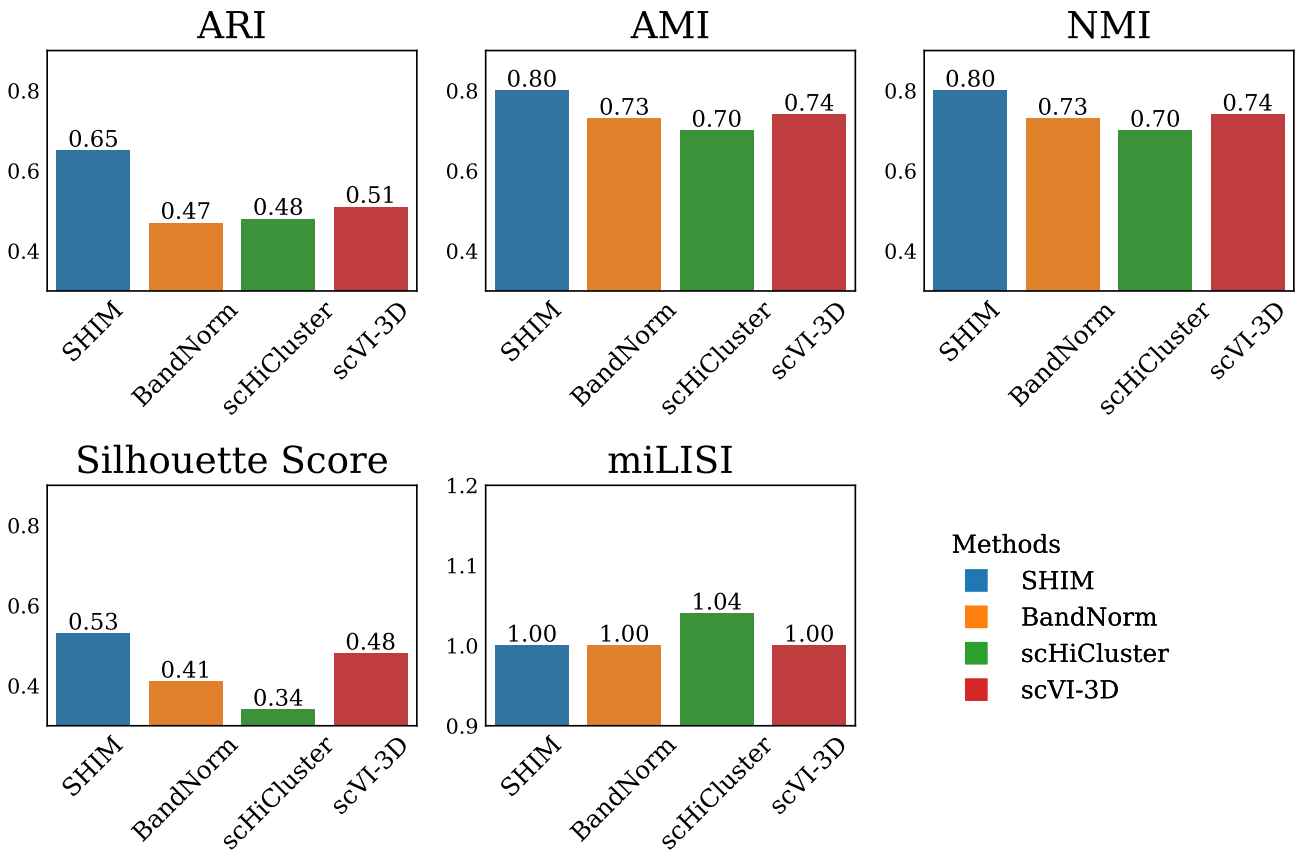
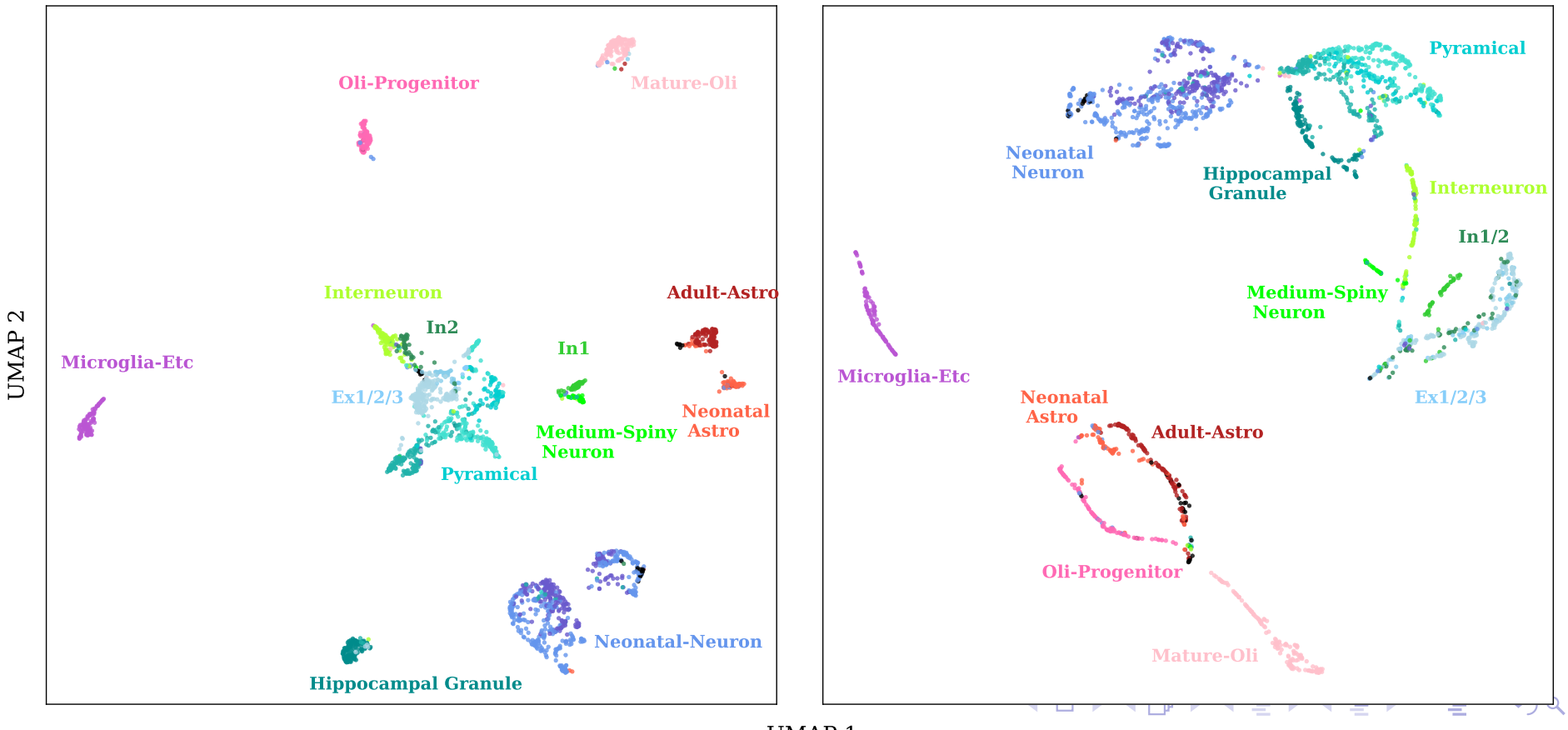


Figure: Clustering for Well-integrated Lee *et al.* Dataset.

Results

- Validation on cell maturity projection.
- HiRES: mature cells, 8- to 9-weeks
- Dip-C: both mature and immature cells throughout first postnatal year

Integrated *Non-integrated*



Results

- Structural similarity between true pseudo-bulk and inferred pseudo-bulk
- Three aspects: HiCRep, A/B Compartments, and TAD Boundaries

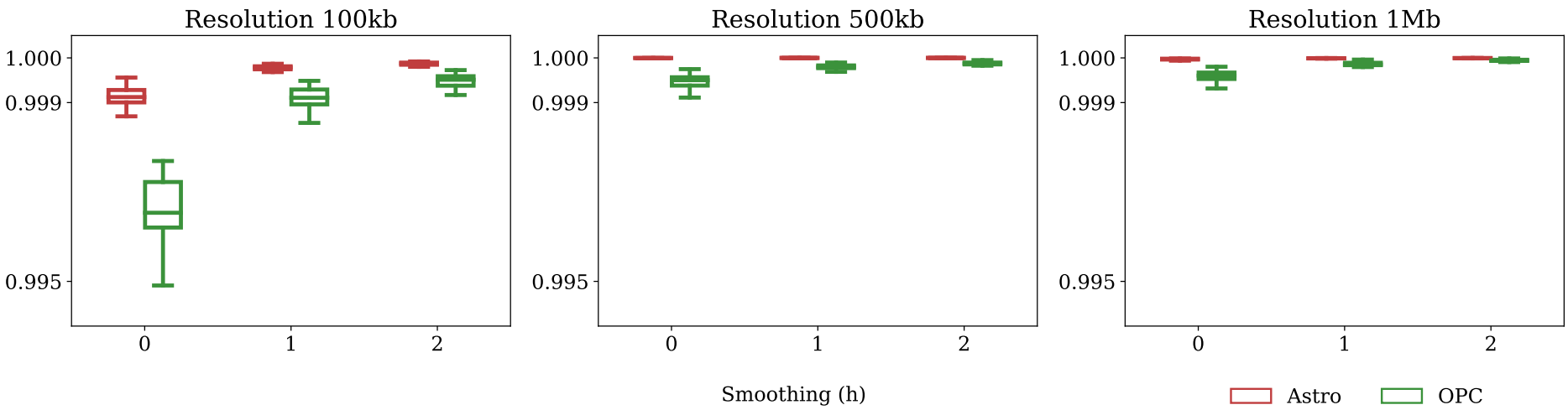


Figure: Pseudo-bulk Reproducibility.

Discussion

- Preservation of natural cell-type separation: MNN
- Artificial noises and biases: diagonal contact exclusion
- High sparsity and skewed distribution: scaled contact maps
- SHIM surpasses the state-of-the-art methods with significant improvements in cell-type isolation and batch mixing.
- We have also validated that SHIM can assist the cell annotation in scHi-C data.
- Integration by SHIM does not affect the cell-type-specific hierarchical structures.

Conclusion

Conclusion

- We provide a robust and rational scHi-C simulator, scHiCPRSiM, to address the problem of limited scHi-C datasets and budgets.
- scHiCPRSiM can generate synthetic scHi-C contact maps that maintains the original structural features.
- Compared to existing methods, scHiCPRSiM is not restricted by the real data, considers the natural sparsity level in real data, and provides more flexible data.
- scHiCPRSiM can be used as guide for experimental design and benchmarking for method developments.

Conclusion

Possible extensions:

- Independent and identical distribution assumption of scHi-C counts
  - Consider bin pair-wise dependency
- Dynamic and Development
  - Single cells are profiled throughout the developmental period.
  - Similar to scRNA-seq, pseudotime in the lineage is taken into consideration.
- Sequencing Reads Simulation
  - scHiCPRSiM is a count-based simulator for scHi-C.
  - Facilitate the method development for upstream analyses
  - Barrier: customized demultiplexing methods

Conclusion

- We introduce SHIM, a powerful statistical scHi-C integration that corrects batch effects under multiple situations.
- SHIM can facilitate 3D genome studies by removing batch effects, which significantly impact results, prior to downstream analysis.
- Compared to existing methods, SHIM is user-friendly, time-efficient, and resource-efficient.

Conclusion

Possible extensions:

- Detection of rare cell types
  - Uneven number of cell subtypes
  - Few cell subtypes with relatively low numbers of cells
- Cross-species scHi-C integration
  - Variations in chromosome lengths across different species
  - Oversimplify: ignoring such cross-species effects
- Integrative alignment
  - Instead of one-step correction between a pair of sets
  - Issue: empty sets of MNNs for some single cells

Acknowledgment

scHiCPRSiM was collaborated with the former lab member, Huiling Liu, under the supervision of Dr. Wenxiu Ma.

I would like to thank all the members in Dr. Ma's lab for their helpful feedback and suggestions.

References I

Fan, S., Dang, D., Ye, Y., Zhang, S.-W., Gao, L., & Zhang, S. (2023).Schi-csim: A flexible simulator that generates high-fidelity single-cell hi-c data for benchmarking. *Journal of Molecular Cell Biology*, *15*(1), mjad003.

Haghverdi, L., Lun, A. T., Morgan, M. D., & Marioni, J. C. (2018).Batch effects in single-cell rna-sequencing data are corrected by matching mutual nearest neighbors. *Nature biotechnology*, *36*(5), 421–427.

Heffel, M. G., Zhou, J., Zhang, Y., Lee, D.-S., Hou, K., Alonso, O. P., Abuhanna, K., Schmitt, A. D., Li, T., Haeussler, M., et al. (2022).Epigenomic and chromosomal architectural reconfiguration in developing human frontal cortex and hippocampus. *bioRxiv*, 2022–10.

References II

Lee, D.-S., Luo, C., Zhou, J., Chandran, S., Rivkin, A., Bartlett, A., Nery, J. R., Fitzpatrick, C., O'Connor, C., Dixon, J. R., et al. (2019).Simultaneous profiling of 3d genome structure and dna methylation in single human cells. *Nature methods*, *16*(10), 999–1006.

Li, W. V., & Li, J. J. (2019).A statistical simulator scdesign for rational scrna-seq experimental design. *Bioinformatics*, *35*(14), i41–i50.

Nagano, T., Lubling, Y., Várnai, C., Dudley, C., Leung, W., Baran, Y., Mendelson Cohen, N., Wingett, S., Fraser, P., & Tanay, A. (2017).Cell-cycle dynamics of chromosomal organization at single-cell resolution. *Nature*, *547*(7661), 61–67.

References III

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

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

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

Zhou, J., Ma, J., Chen, Y., Cheng, C., Bao, B., Peng, J., Sejnowski, T. J., Dixon, J. R., & Ecker, J. R. (2019).Robust single-cell hi-c clustering by convolution-and random-walk–based imputation. *Proceedings of the National Academy of Sciences*, *116*(28), 14011–14018.

Supplementary: scHiCPRSiM

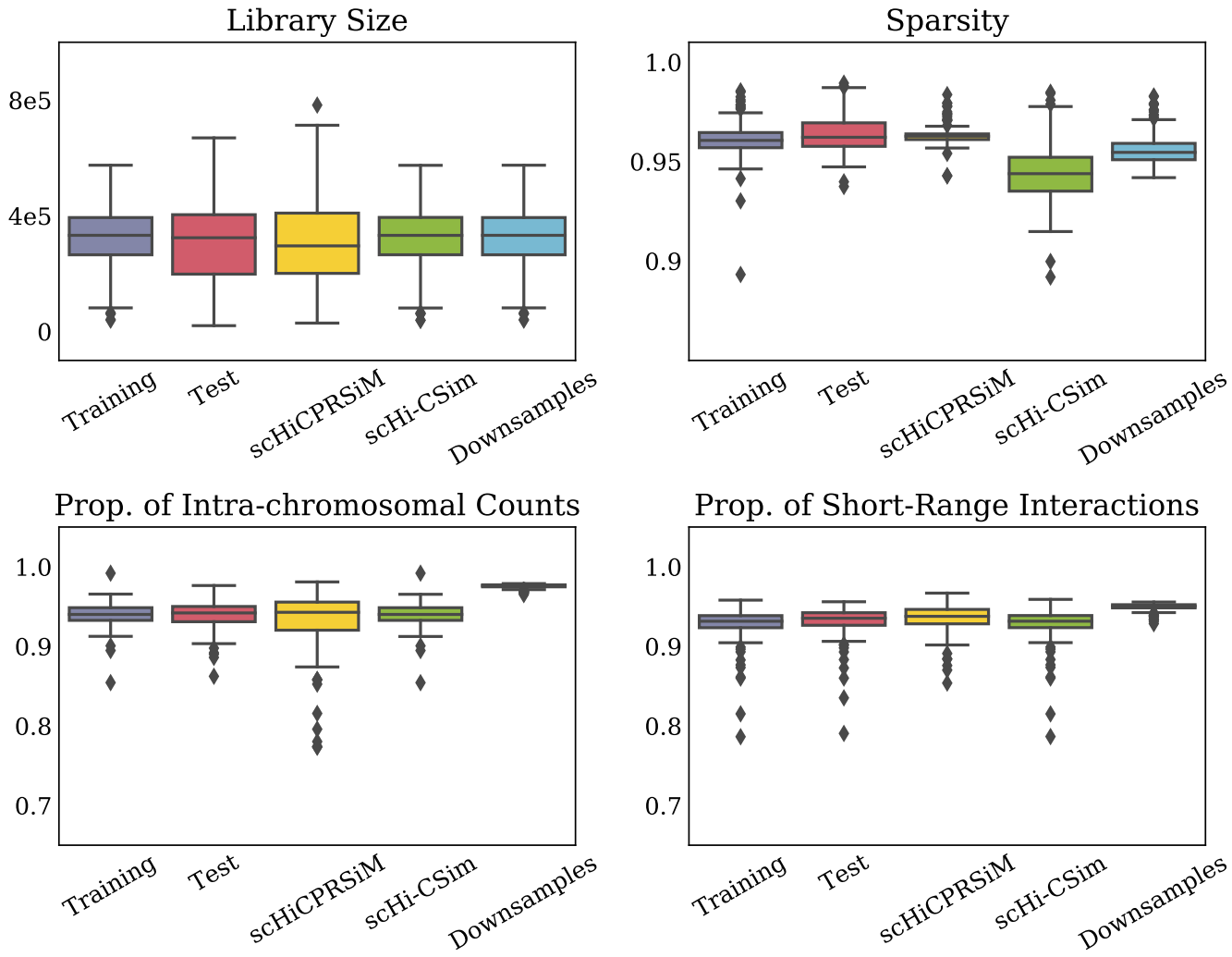


Figure: Boxplots of Summary Statistics for Mouse Cell-cycle Data.

Supplementary: scHiCPRSiM

- The silhouette coefficient measures how similar an object is to its own cluster compared to other clusters.
  - +1 means clusters are distinguishable.
  - 0 means clusters are indifferent.
  - -1 means clusters are assigned incorrectly.

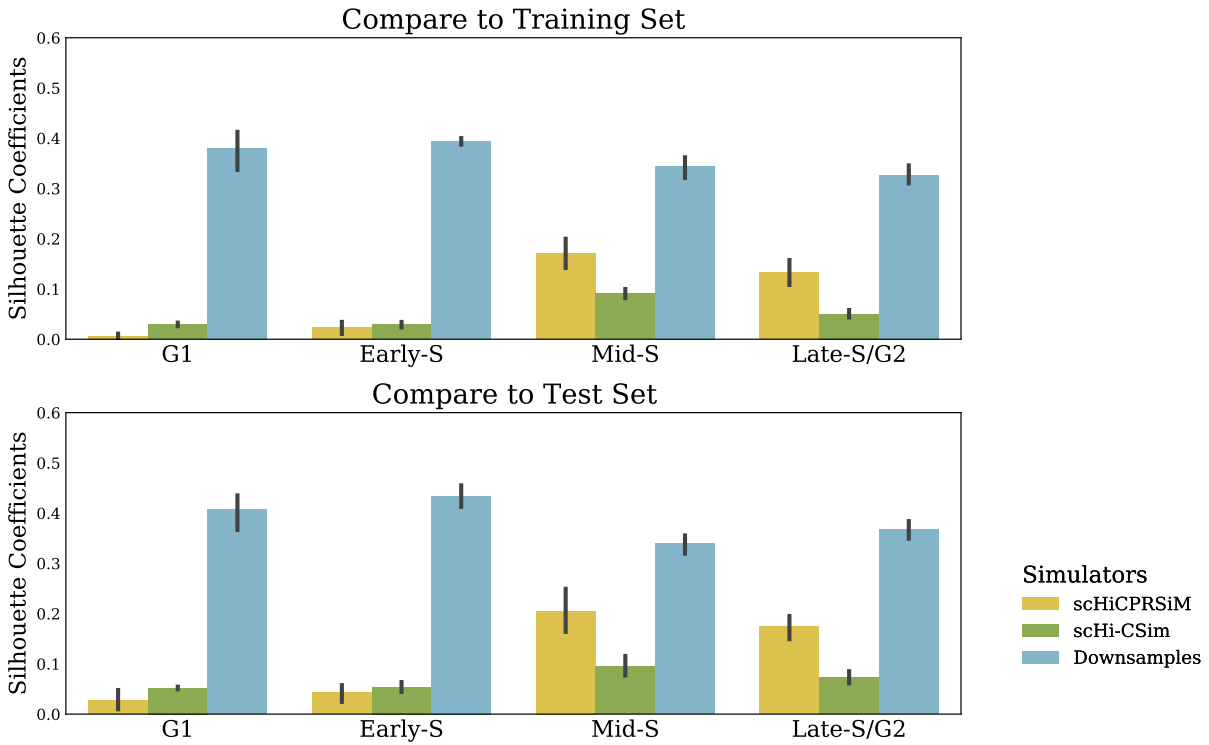


Figure: Silhouette Coefficients for Real v.s. Synthetic Mouse Cell-cycle Data.

Supplementary: scHiCPRSiM

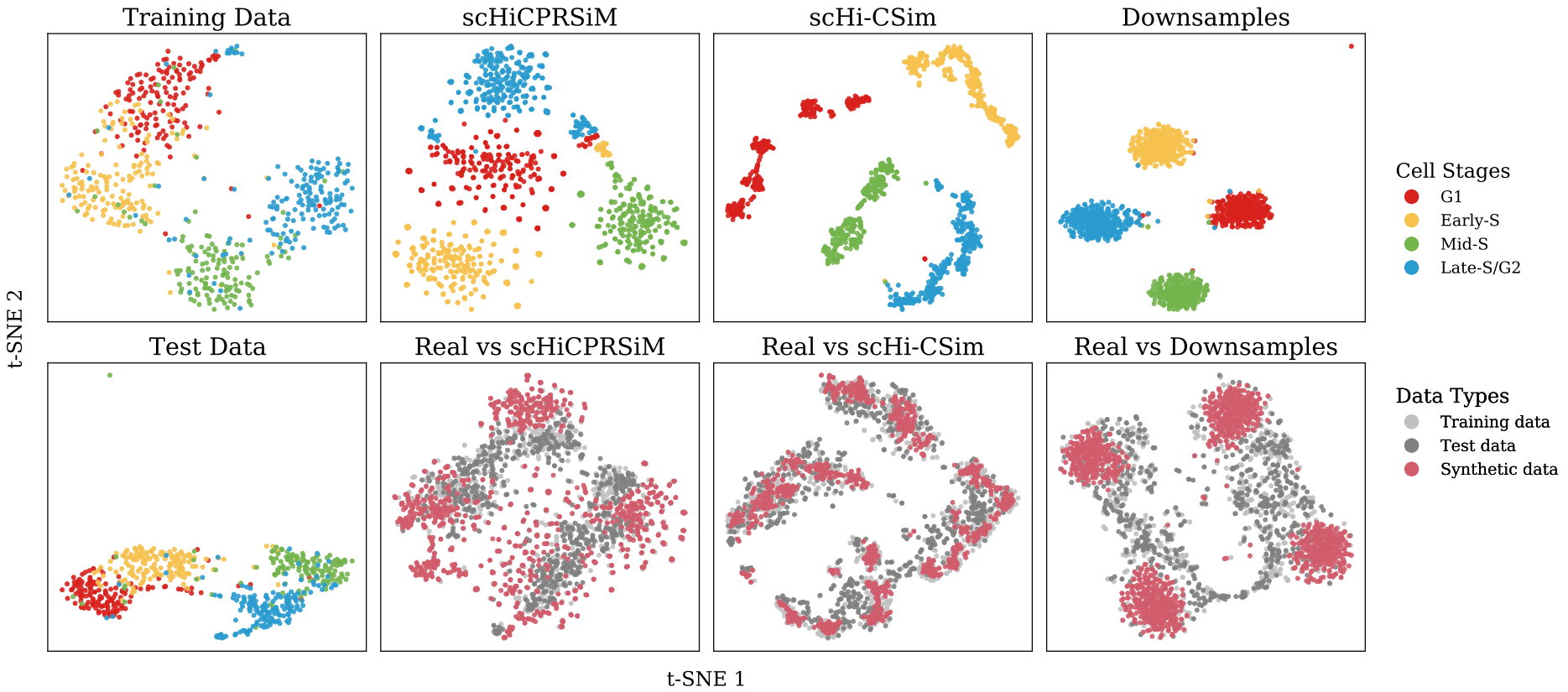


Figure: 2D Visualizations on Embeddings of Real and Synthetic Data for Mouse Cell-cycle.

Supplementary: scHiCPRSiM

Figure: Silhouette Coefficients for Real v.s. Synthetic Human Brain Data.

Results

- scHi-C maps generated by scHiCPRSiM at different resolutions are consistent.

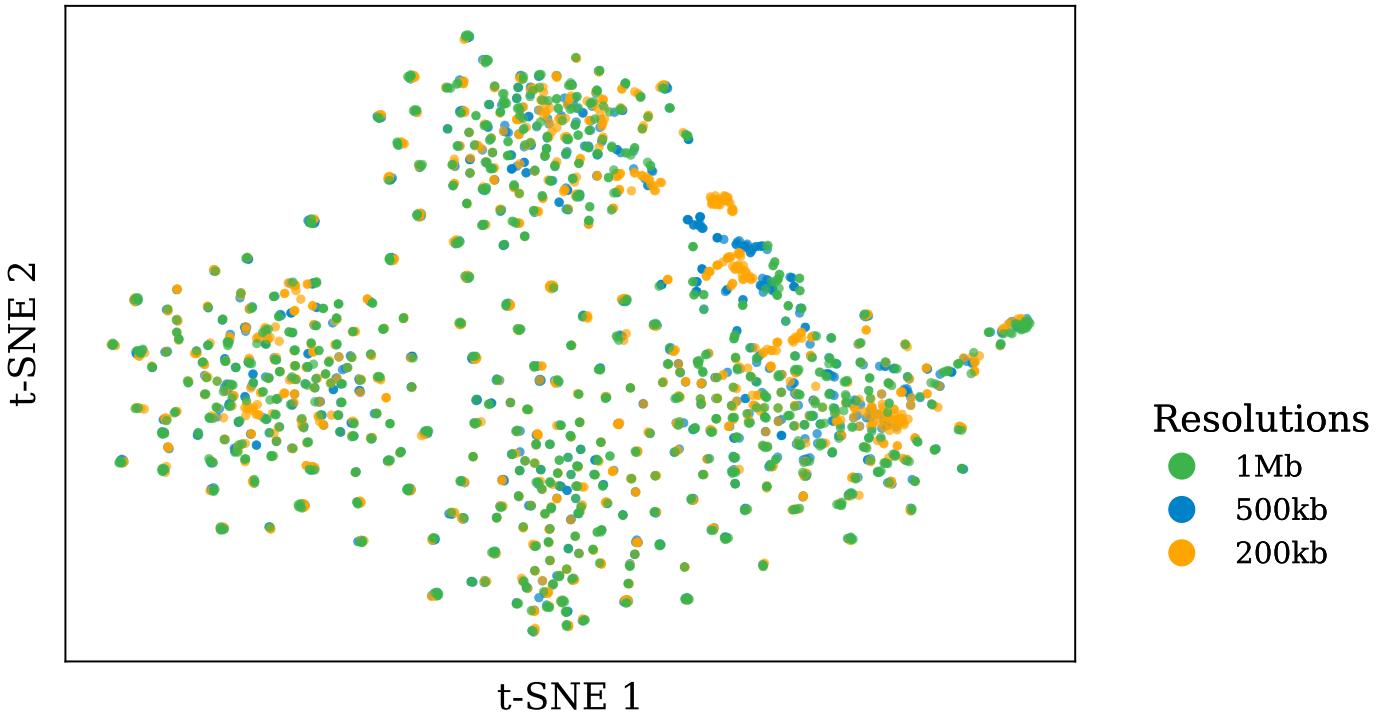


Figure: Consistency at Different Resolutions.

Supplementary: scHiCPRSiM

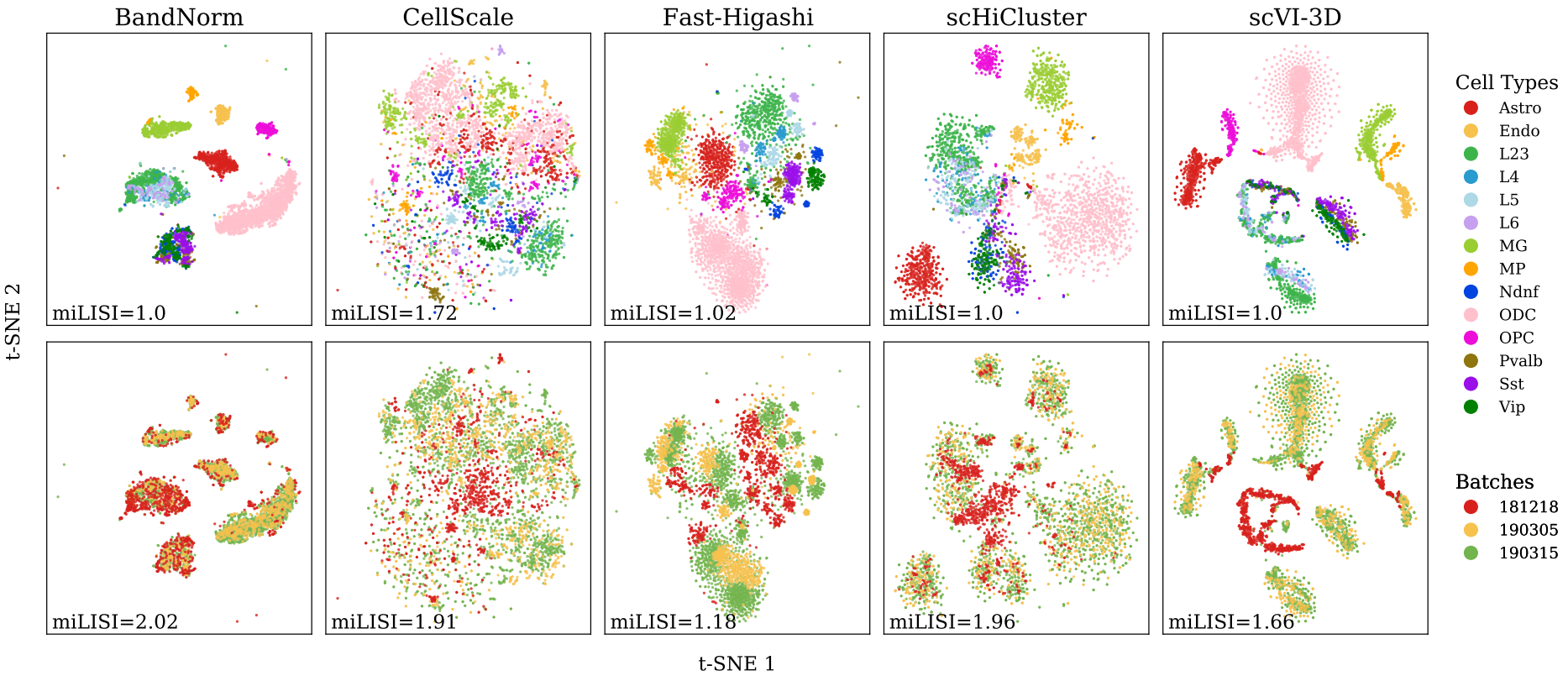


Figure: Embedding Results of Batch Effect Correction Methods.

Supplementary: SHIM

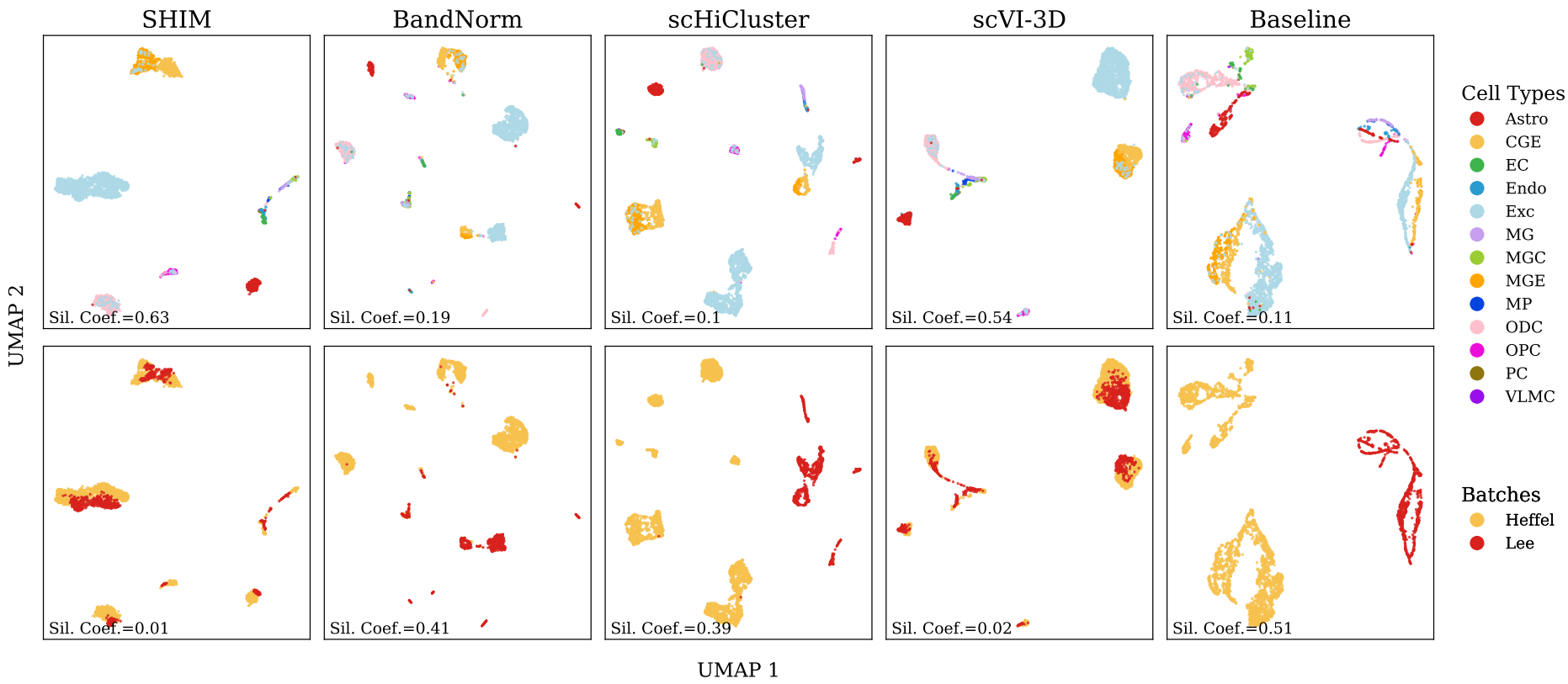


Figure: Embeddings for Non-integrated and Integrated Human Brain Datasets.

Supplementary: SHIM

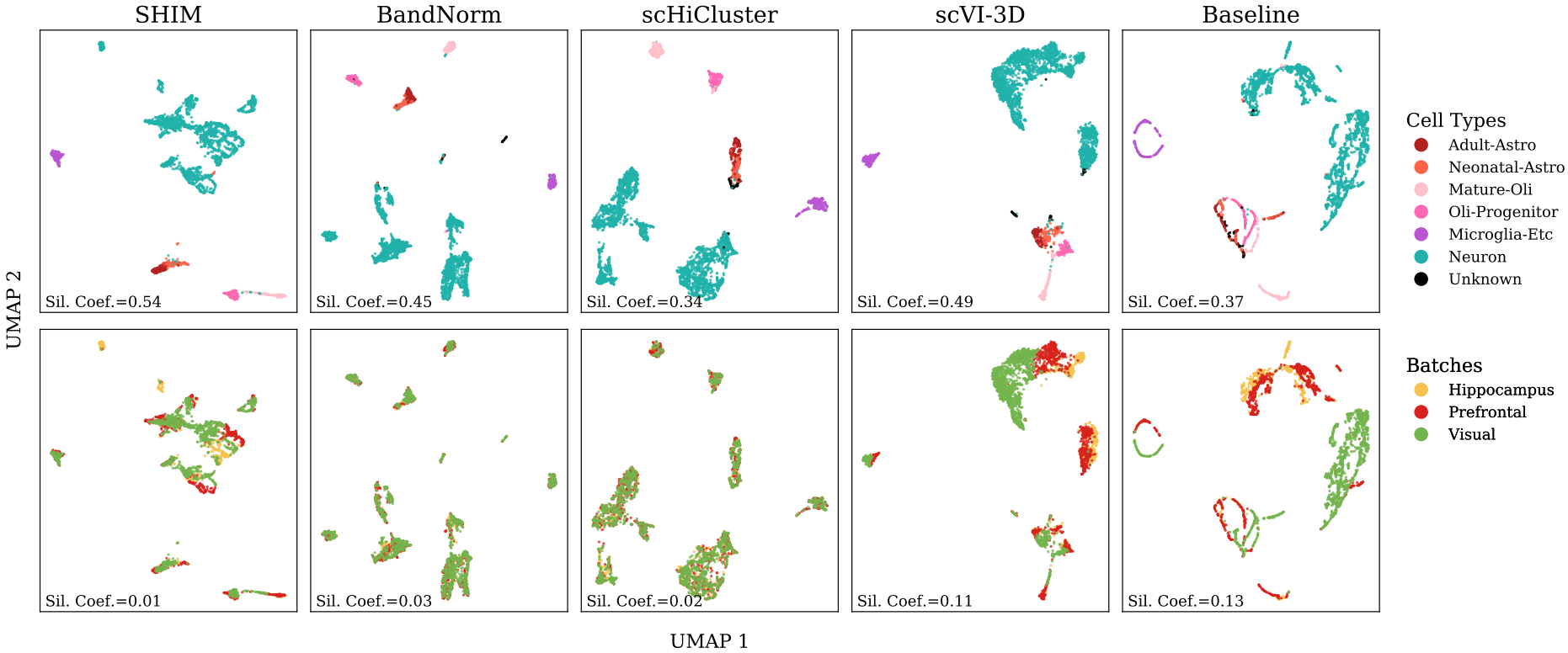


Figure: Embedding for Non-integrated and Integrated Dip-C Dataset.

Supplementary: SHIM

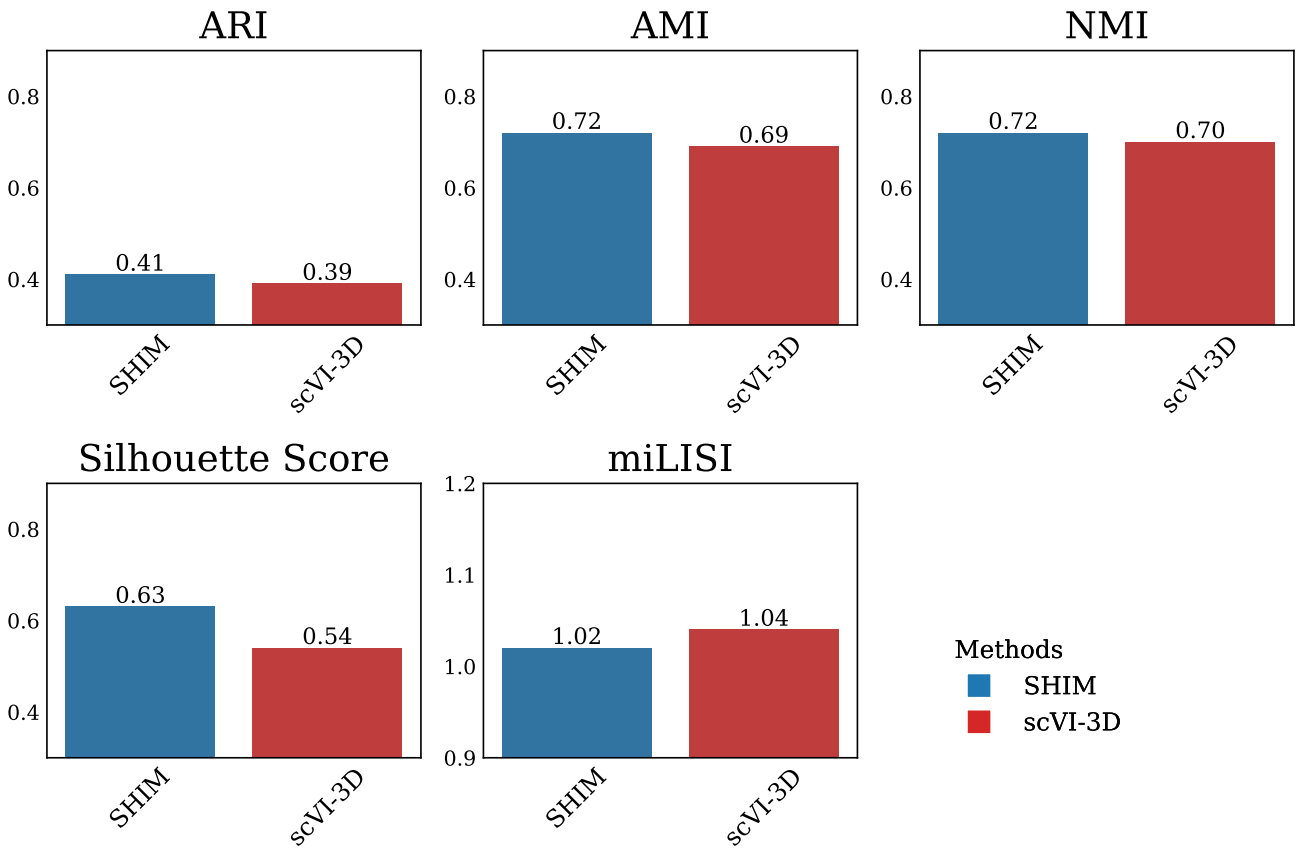


Figure: Clustering for Well-integrated Human Brain Datasets.

Supplementary: SHIM

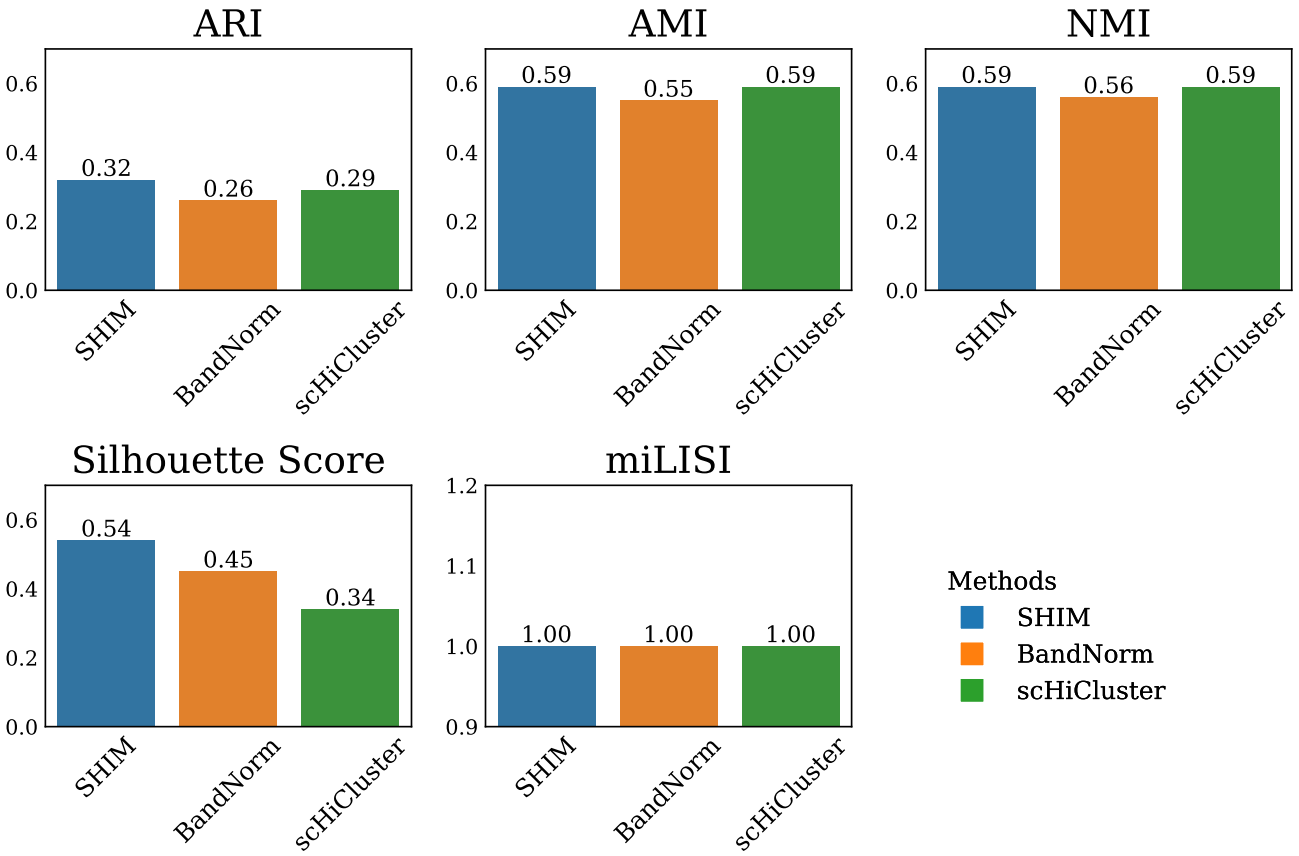


Figure: Clustering for Well-integrated Dip-C Dataset.

Supplementary: SHIM

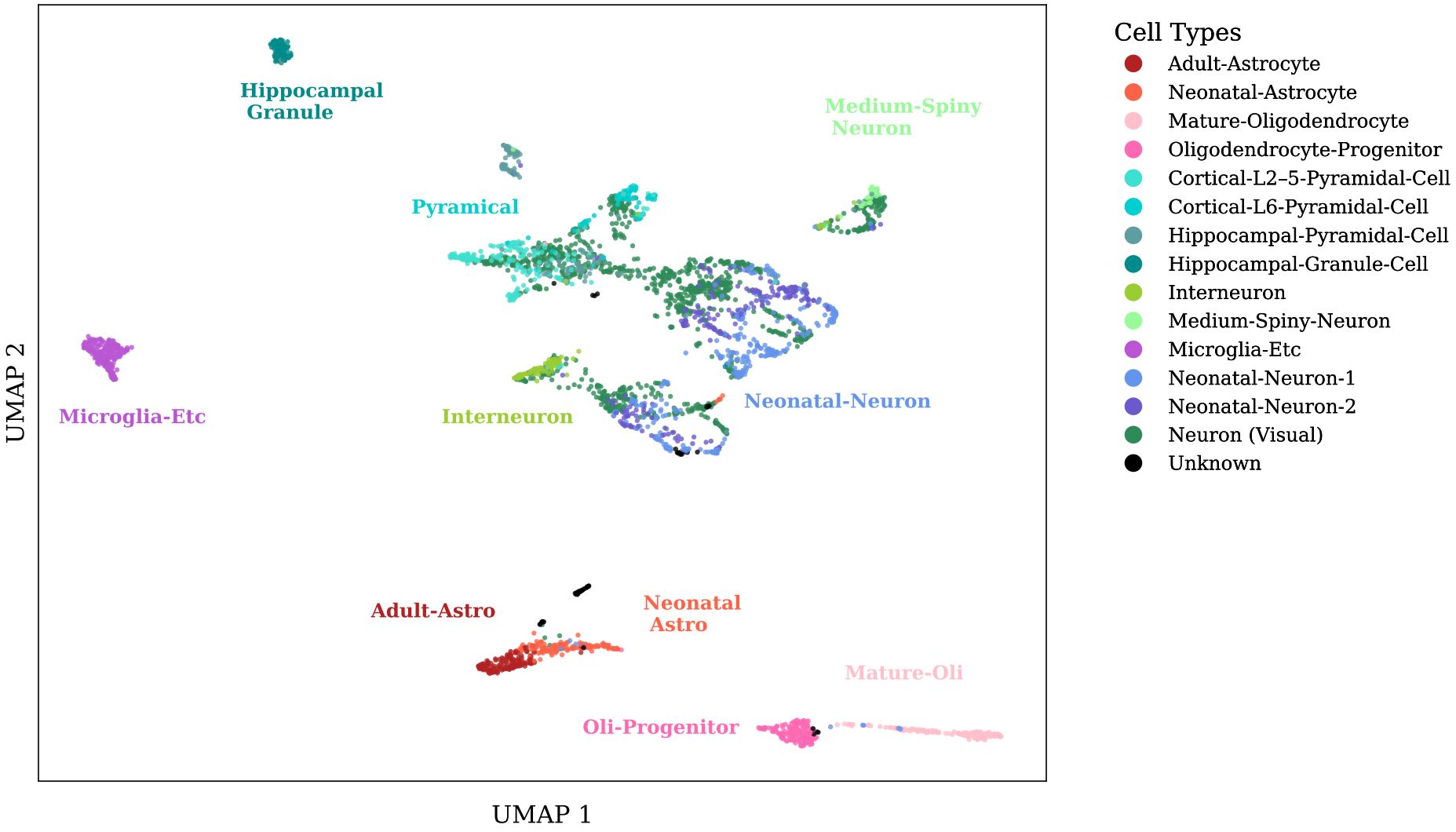


Figure: Embedding for Integrated Dip-C Mouse Brain Dataset using SHIM.

Supplementary: SHIM

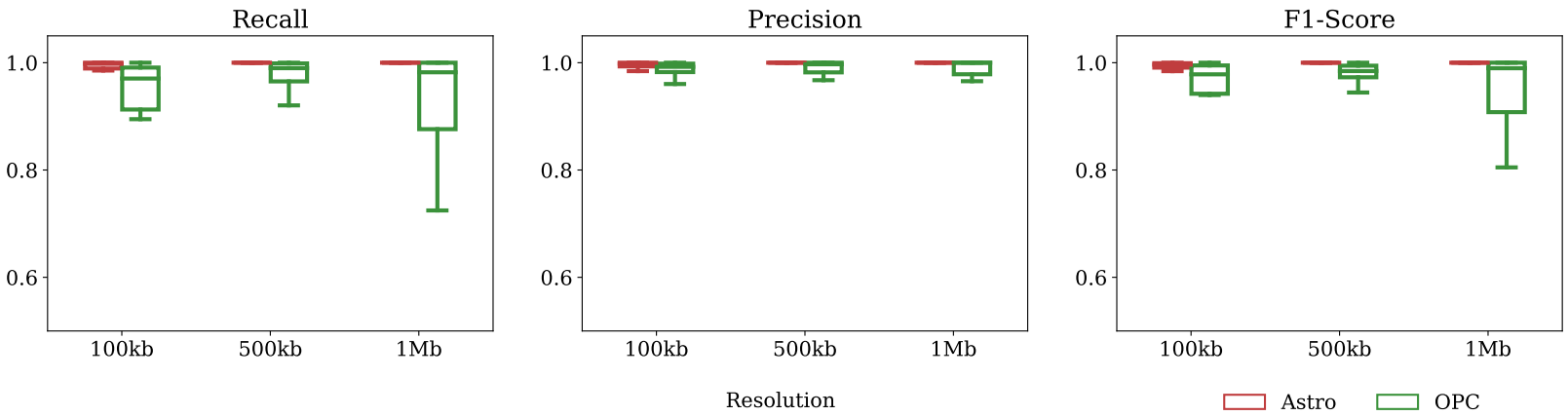


Figure: Pseudo-bulk A/B Compartments Evaluation.

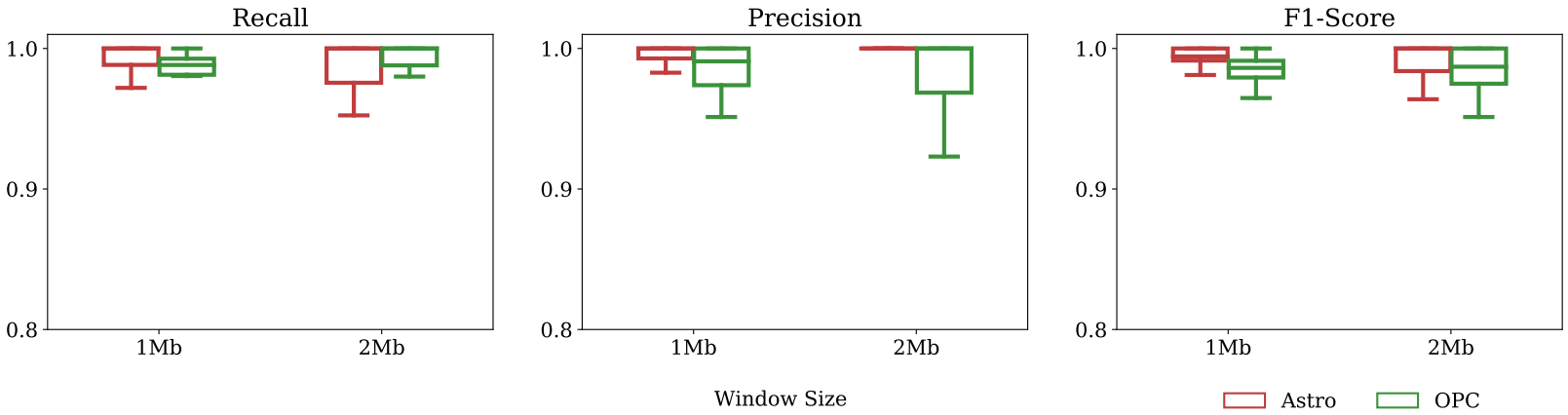


Figure: Pseudo-bulk TAD Detection Evaluation.