# Model-based imputation enables improved resolution for identifying differential chromatin contacts in single-cell Hi-C data (supplementary information)

# 1 Supplementary tables

Table. S 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. S 2: scHi-C datasets statistics.

| Dataset | # of cells | average total contact counts | average off-diagonal contact counts | group 1 | group 2 |
|---|---|---|---|---|---|
| Lee2019 | 4238 | 1.08 M | 190 K | 126 Astro cells from batch 190315_21yr | 112 MG cells from batch 190315_21yr |
| | | | | 90 Astro cells from batch 190315_29yr | 102 MG cells from batch 190315_29yr |
| | | | | 126 Astro cells from batch 190315_21yr | 90 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 |

Table. S 3: Filtering criteria for different resolutions.

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

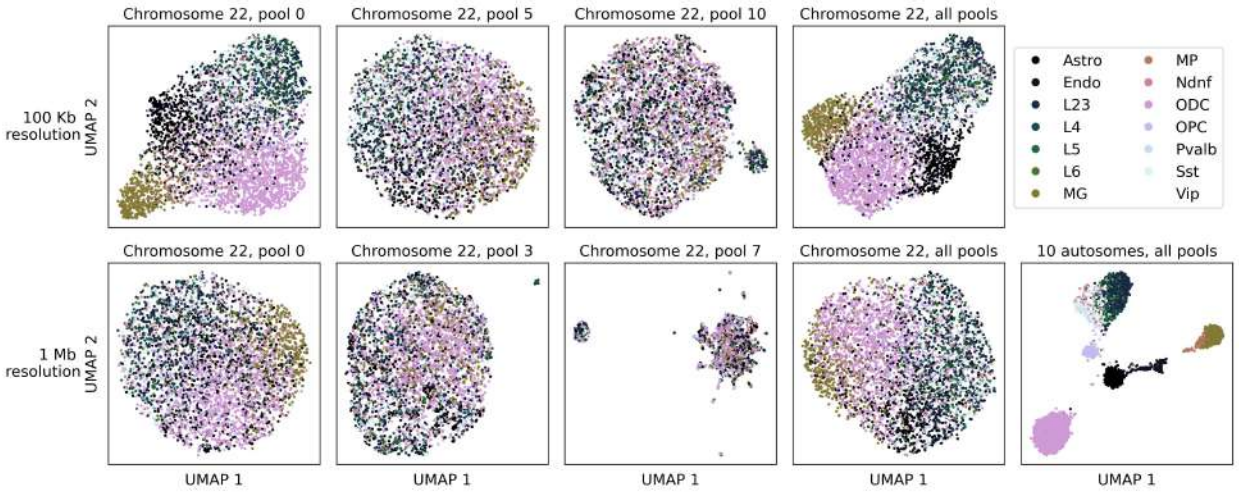
# 2 Supplementary notes

## 2.1 Data Preprocessing

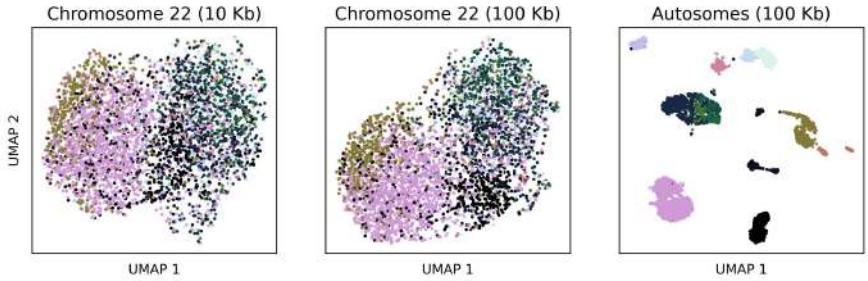
All Hi-C and scHi-C datasets (Table. S 1), except bulk Hi-C data for mESC, NPC, and HFF were processed and mapped to hg19 or mm10 and stored in tab-separated, pairs, or cool format. The bulk Hi-C data for HFF was mapped to hg38, and we used HiCLift [1] to lift it to hg19 to compare it with other datasets. 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.

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

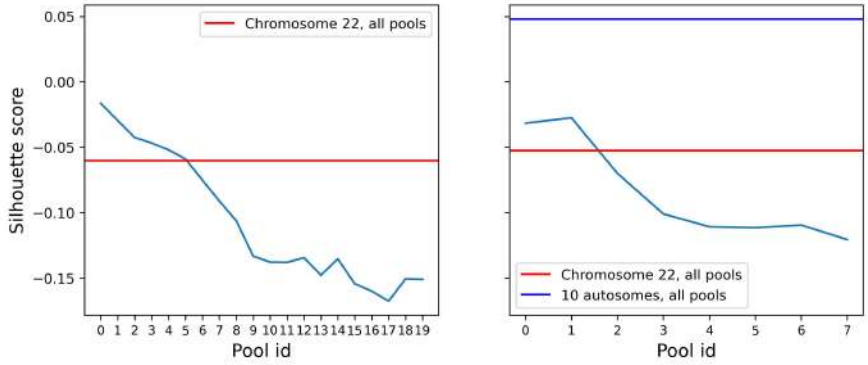
# 3 Supplementary figures




(a)




(b)




(c)

Fig. S1: (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.

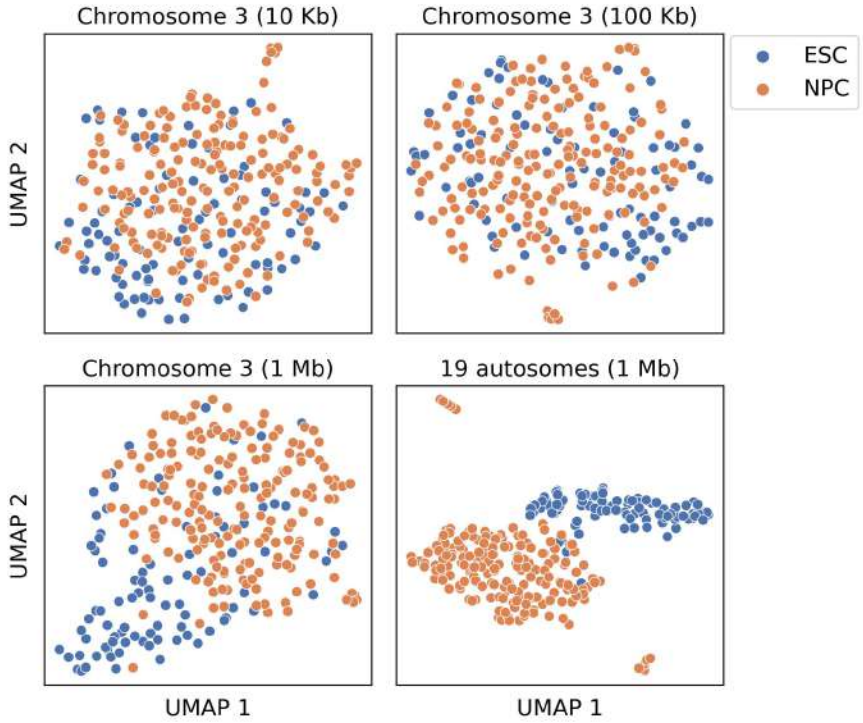
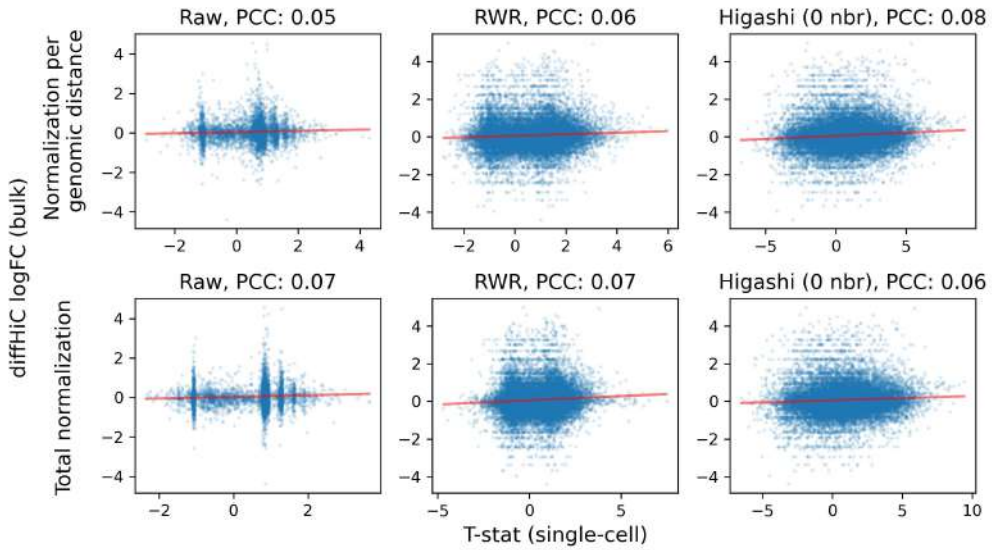
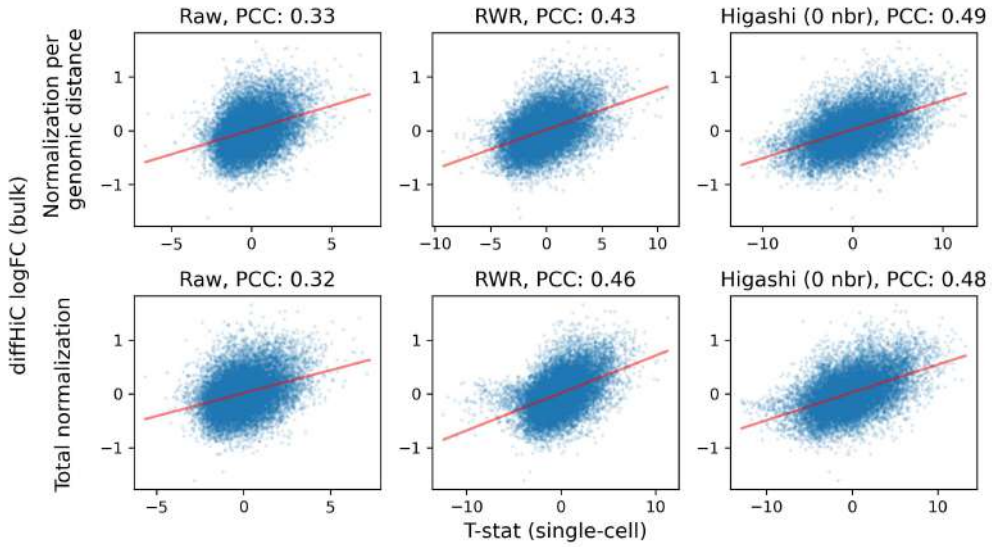



Fig. S2: Higashi embedding for *Lee2023* dataset across different resolutions, and two training sets, including genomic bins from chromosome 3 or all chromosomes.

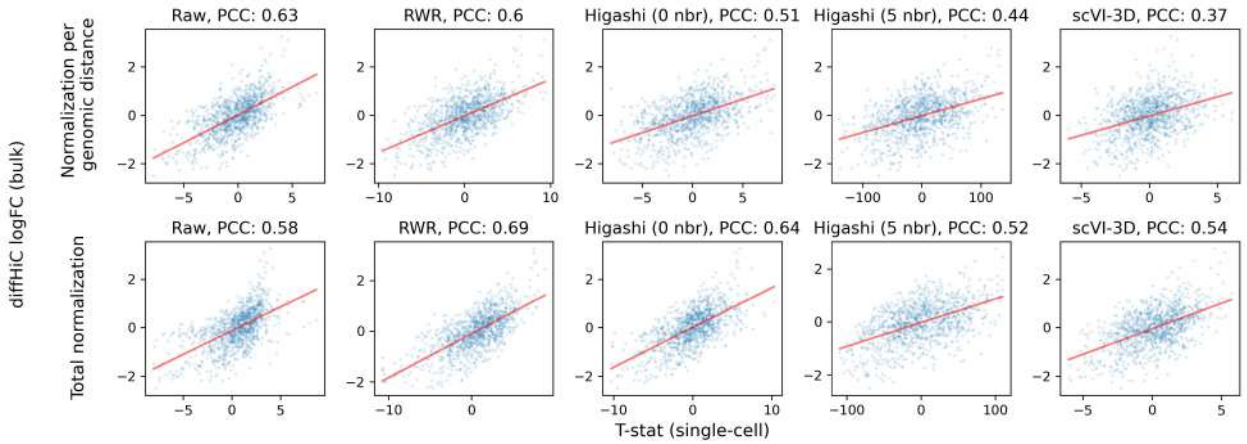



(a)

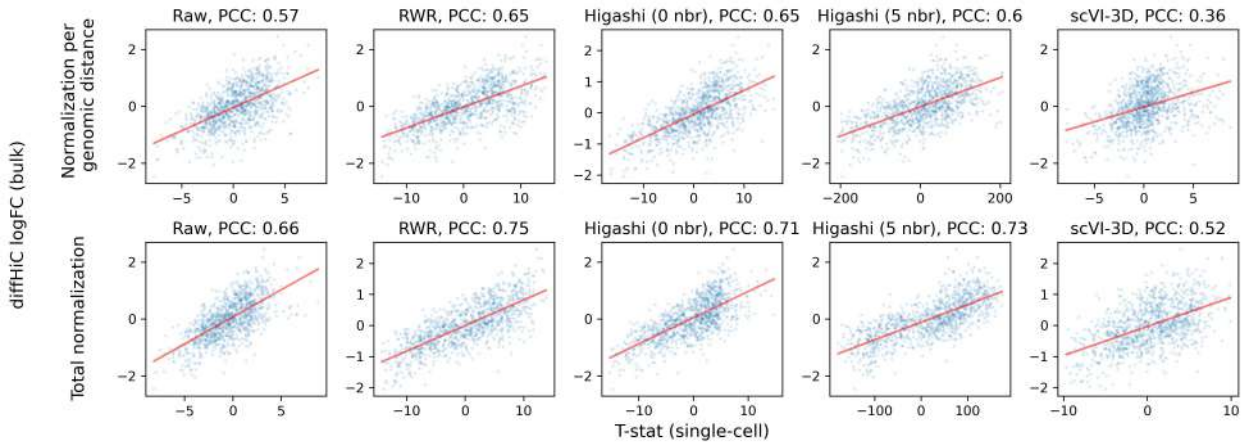



(b)

Fig. S3: The correlation between the log fold-change from bulk DCC caller and t-statistics from single-cell DCC caller after different normalization and imputation approaches for (a) 10 Kb and (b) 100 Kb resolutions (*Lee2023* dataset).

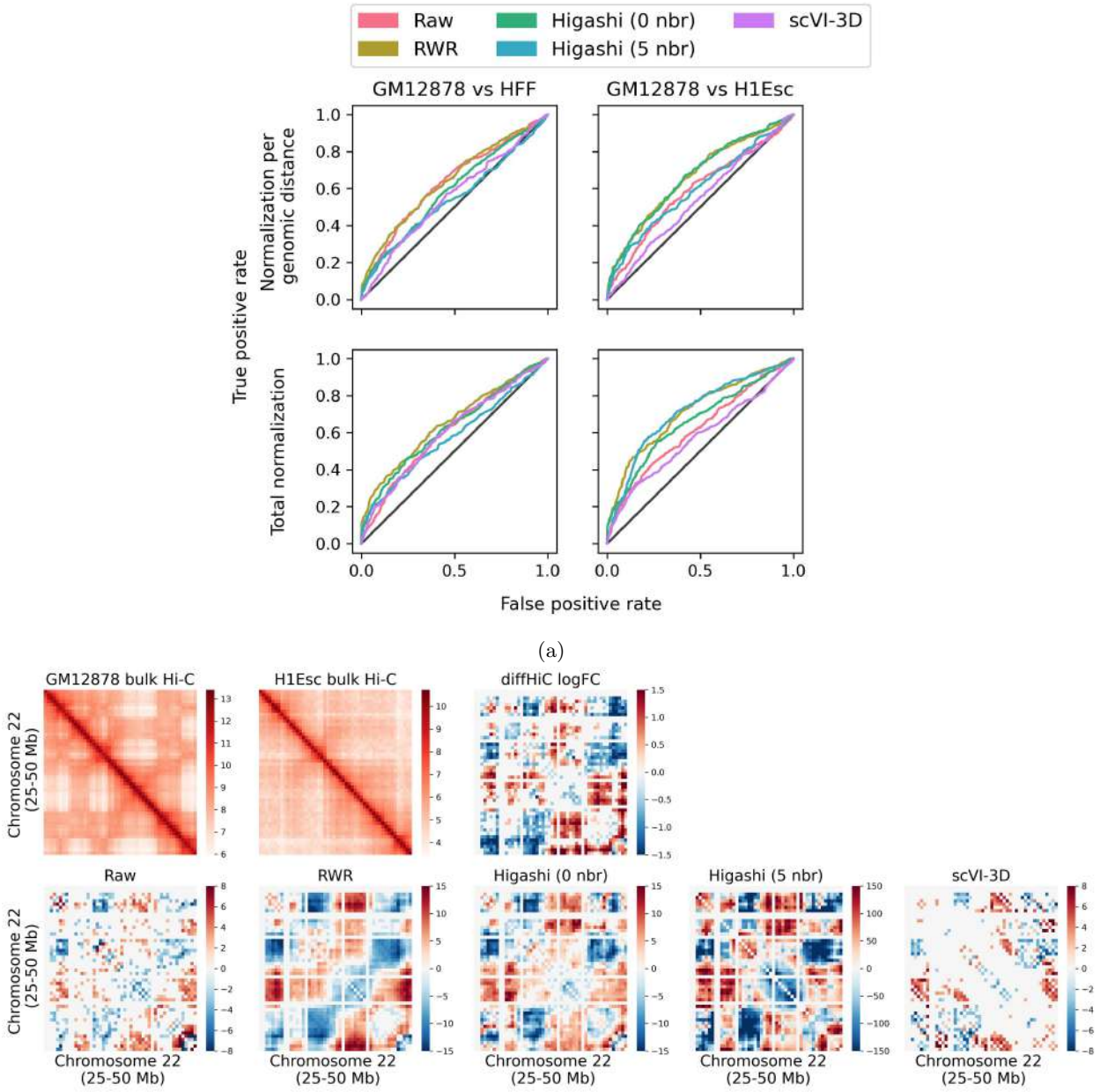



(a)




(b)

Fig. S4: The correlation between the log fold-change from bulk DCC caller and t-statistics from single-cell DCC caller after different normalization and imputation approaches for (a) GM12878 vs HFF and *b* GM12878 vs H1Esc comparisons (*Kim2020* dataset).




(a)

(b)

Fig. S5: (a) Comparison of ROC curves for called DCCs by different imputation and normalization approaches from two comparisons. (b) The heatmap of bulk Hi-C contact maps for GM12878 and H1Esc cell lines and diffHiC log fold-change (logFC) and single-cell t-statistics from the comparison of these two cell lines.

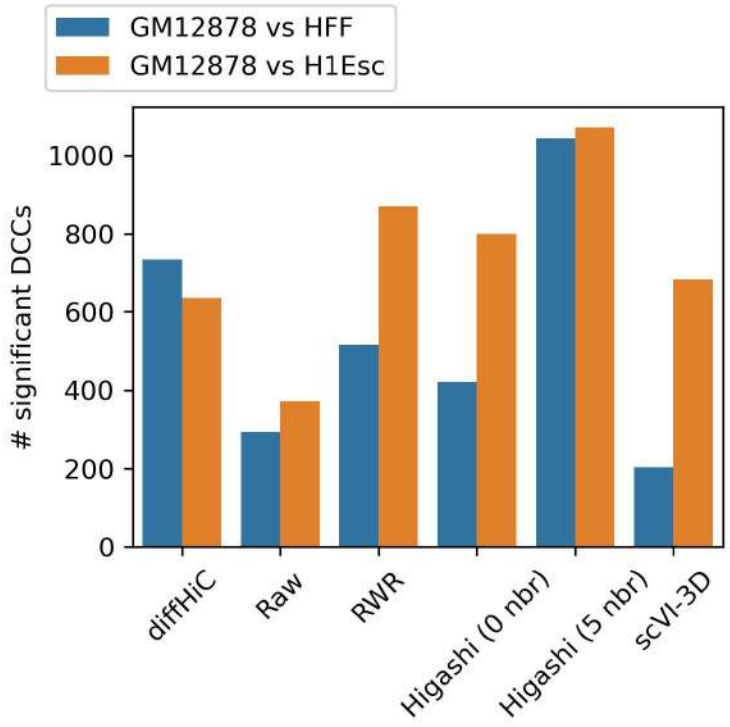



Fig. S6: The number of significant DCCs for two *Kim2020*'s comparisons.