MINTsC learns multi-way chromatin interactions from single cell high throughput chromatin conformation data

Kwangmoon Park[1], **Tianchuan Gao**[3], **Jingwen Yan**[3], & **Sündüz Keleş**[1,2,*]

[1]Department of Statistics, University of Wisconsin - Madison, Madison, WI 53706, USA

[2]Department of Biostatistics and Medical Informatics, University of Wisconsin - Madison, Madison, WI 53706, USA

[3] Department of Biomedical Engineering and Informatics, Indiana University, Indianapolis, IN 46202, USA.

[*] Corresponding author

Abstract. A number of foundational analysis methods have emerged for single cell high throughput chromatin conformation (scHi-C) datasets capturing 3D organizations of genomes with pairwise measurements at the single cell or nuclei resolution; however, these datasets are currently under-utilized. The canonical analyses of scHi-C data encompass, beyond standard cell type identification, inference of chromosomal structures and pairwise interactions. However, multi-way chromatin interactions among genomic elements are entirely overlooked. We introduce **MINTsC** to learn **M**ulti-way **INT**eractions from **s**ingle cell Hi-**C**. MINTsC builds on a dirichlet-multinomial spline model and yields multi-way interaction scores by aggregating pairwise interactions across cells of a context and summarizing them using order statistics of pairwise test statistics. MINTsC yields well-calibrated p-values for controlling the false discovery rate. Evaluation of MINTsC with scHi-C datasets from cell lines and complex tissues using multiple external genomic and epigenomic datasets support multi-way interactions inferred by MINTsC. Application of MINTsC to scHi-C data from human prefrontal cortex revealed multi-way chromatin interactions with biological implications of gene regulation by multiple enhancers. Most notably, MINTsC-inferred multi-way interactions demonstrate its potential for probing molecular QTL and association studies for epistatic SNP effects by significantly reducing the multiple-testing burden. MINTsC is publicly available at `https://github.com/keleslab/mintsc`.

1 Introduction

Advances in high-throughput single-nucleus mapping of 3D genome organization via pairwise contacts between proximal loci now enable genome-wide characterization of loops, TADs, A/B compartments, and subcompartments at single-nucleus resolution. Yet these datasets have not been systematically leveraged to resolve multi-way chromatin interactions (simultaneous contacts among multiple loci within the same nucleus). Identifying such interactions is crucial for assessing the epistatic potential of risk variants from genome-wide association studies (GWAS) and molecular quantitative trait loci (QTL) studies. Disease-linked genes are often regulated by multiple distal enhancers spread over large genomic distances, e.g., nested enhancer cooperativity of the oncogene MYC was dissected by multiplexed CRISPRi screening [1]. Genome-wide association studies have shown that more than 90% of disease-associated variants are noncoding variants[2,3], and can be dispersed over large distances. Furthermore, while single enhancer

variants may exert modest effects individually, combinations can interact to substantially influence traits and disease.

A suite of experimental innovations now measure multi-way chromatin contacts: GAM [4], ChIA-drop [5], SPRITE [6,7], Tri-C, multi-contact 4C [8], COLA [9], and Pore-C [10]. These platforms established feasibility mainly in cell lines and mESCs, but Hi-C and scHi-C remain the most widely generated and used for complex tissues. Notably, the NIH BRAIN Initiative has produced extensive high-quality scHi-C from human and mouse brains. Yet analyses of scHi-C have largely focused on pairwise contacts, often adapting bulk Hi-C methods (with snapHi-C [11,12] as a notable exception). This underutilizes scHi-C for elucidating how multiple enhancers cooperate over long genomic distances to regulate target genes and confer disease risk. Long-read sequencing-based scNanoHi-C [13] demonstrated that multi-way contacts can be captured in single cells, but cellular heterogeneity and experimental noise make them inconsistently observable across cells. scMicro-C [14] likewise revealed dynamic variability even in relatively homogeneous cell lines. These observations underscore the need for a principled statistical framework that models the underlying multi-way chromatin interaction structure and reconstructs it from noisy, heterogeneous single-cell data.

In this study, we introduce MINTsC, the first method to learn "**M**ulti-way Chromatin **INT**eractions from scHi-**C**". We cast multi-way chromatin interaction discovery as clique detection in a multilayer network: each layer is a cell, nodes are genomic loci, and edges denote locus-locus contacts (**Fig.** 1a). MINTsC models pairwise contacts with an empirical Bayes spline null that accounts for the well-known genomic distance bias of scHi-C assays, and then derives a clique-level test statistic with an analytic null distribution. Multiple testing is controlled via the Benjamini-Hochberg (BH)[15] procedure to call multi-way chromatin interactions at a target false discovery rate (FDR). Because the clique statistic aggregates pairwise signals, MINTsC also applies targeted pre- and post-processing filters to minimize spurious multi-way chromatin interaction calls.

Evaluations of MINTsC with external multi-way chromatin interaction data, imaging-based 3D genome data, and data from methylation assays support the multi-way interactions identified by MINTsC. Applied to scMicro-C data, MINTsC calls remain consistent with inferred 3D genomic distances despite the dynamic nature of multi-way chromatin contacts representing the contributions of individual enhancers to the regulation of genes. Compared with a baseline derived from state-of-the-art scHi-C loop calling, MINTsC identifies more multi-way chromatin interactions with markedly better FDR control in human prefrontal cortex and synthetic data. Strikingly, interrogation of MINTsC-inferred multi-way interactions from human prefrontal cortex in an eQTL study of human cortex reveals epistatic SNP effects for genes linked to amyloid-β pathology and synaptic plasticity.

2 Results

2.1 Overview of MINTsC

Bulk Hi-C's aggregate nature cannot resolve multi-way chromatin interactions. In contrast, scHi-C enables a graph formulation in which multi-way contacts correspond to cliques in a multilayer network: each layer is a cell, nodes are genomic loci (bins; typically 10 kb-1 Mb depending on coverage), and edges encode locus-locus contacts [16] (**Fig.** 1a). Despite ligation limits and sparsity, scHi-C yields abundant cliques (**Fig.** S1), confirming its potential to infer lower-order multi-way chromatin interactions (empirically order 3-6), even if very high-order events remain challenging.

We developed MINTsC to learn multi-way chromatin interactions from scHi-C by testing cliques for significance. First, MINTsC fits an empirical-Bayes spline null model for pairwise contacts that accounts for genomic-distance bias, yielding bias-adjusted pairwise test statistics (**Fig.** 1b; Methods). Next, within a homogeneous context (e.g., a cell type or treatment), MINTsC aggregates pairwise evidence across cells to form a clique-level statistic ("clique p-score") using order statistics of the pairwise p-values, and derives an analytic null distribution for this statistic (**Fig.** 1c; Methods). Under the homogeneity assumption (Supplementary Note), the clique score reflects the strength of the corresponding clique in the population-level contact matrix. Because the statistic is built from distance-adjusted pairwise tests, it is stable to the distribution of intra-clique genomic distances (**Fig.** S2). This yields well-calibrated *clique p-values* based on parametric null distribution of the clique p-scores for FDR control (**Fig.** 1d, **Method**) and facilitates downstream analysis, e.g., interaction eQTLs, by drastically reducing the multiple testing burden. For example, the expression variability of a gene with promoter within locus v_3 can be studied for contributions of the interactions between SNPs (e.g., SNP_1 in v_1 and SNP_2 in v_2) that reside in loci v_1 and v_2 rather than exhaustively over all combinations of gene's *cis* SNPs (**Fig.** 1e).

Because multi-way calls aggregate pairwise contacts, MINTsC includes pre- and post-filtering to guard against spurious multi-way chromatin interactions. A pre-filter defines candidate cliques as those fully observed in at least a minimum number of cells, protecting against spurious cliques formed by merging pairwise interactions from different groups of cells (**Fig.** 1a; Methods). An optional post-filter applies a statistical co-occurrence test on clique edges to further eliminate artifacts (Supplementary Note).

2.2 Validation of MINTsC multi-way chromatin interactions for same-cell co-localization

Since MINTsC relies on aggregation of pairwise contacts, we first evaluated fidelity of MINTsC-identified multi-way interactions at the single-cell resolution. Specifically, by leveraging orthogonal 3D spatial distance measurements and targeted simulations, we assessed whether inferred multi-way interactions reflect same-cell co-localization. In all of the applications, we grouped the candidate cliques evaluated by MINTsC into two based on MINTsC clique p-values resulting from the parametric Beta distribution of the clique p-scores

(**Methods**). Clique p-values were then subjected to FDR thresholding[15] and labeled as *multi-way chromatin interactions*.

(i) 3D spatial distance–based validation with train-test split on GM12878 scMicro-C. We applied MINTsC to scMicro-C contact matrices[14] of 10kb resolution from GM12878 cells, fitting on one group of cells (48 training cells) and evaluating on the remaining cells (48 test cells). Pre-filtering step of MINTsC generated 43,469 three-way candidate cliques and MINTsC applied to training set identified 904 multi-way chromatin interactions (at FDR of 0.1). Using the 3D genome coordinates constructed for the test scMicro-C cells with the Dip-C algorithm[17], we computed, for each candidate clique in every test cell, the maximum pairwise 3D distance (Euclidean) among the clique loci. Across the individual test cells, cliques identified as multi-way chromatin interaction by MINTsC showed significantly smaller maximum pairwise 3D distances than other cliques, supporting same-cell colocalization of the interacting loci (Fig. 2a). Notably, for nearly all cells, within clique maximum pairwise 3D distances of MINTsC-multi-way chromatin interactions were below 3.5 particle radii ($\sim$240 nm),which is the distance cutoff used in scMicro-C to define pairwise chromatin interactions. Next, to obtain an estimate of empirical FDR, we applied MINTsC to the whole set of cells and quantified the proportion of cells each clique had a maximum pairwise 3D distance below 3.5 particle radii. This revealed that MINTsC-multi-way chromatin interactions co-localized in a significantly larger proportion of cells compared to cliques that were not significant by MINTsC (FDR of 0.1, **Fig.** 2b). We show that the association based on scMicro-C still hold after matching the cliques with similar total pairwise interaction strength, further demonstrating that MINTsC scores are indeed signatures of multi-way chromatin interactions (**Fig.** S3a,b). When cliques with maximum pairwise 3D distance less than 3.5 particle radii in at least 10% of the cells were considered as true multi-way interactions, MINTsC yielded an empirical FDR of 8% (FDR of 10% when the requirement on the proportion of cells was raised to 50%).

(ii) Cross-technology validation in mouse cortex and hippocampus. We next utilized MINTsC to infer multi-way chromatin interactions from mouse cortex and hippocampus scHi-C (Dip-C) data[18]. The contact maps from Dip-C were binned at 25Kb resolution to match the resolution of the publicly available mouse brain DNA seqFISH+ data[19] for evaluation. The DNA seqFISH+ dataset harbored spatial coordinates of targeted 1.5Mb genomic region of each chromosome in individual cells and enabled calculation of spatial distances between loci pairs.

We considered the set of candidate cliques with loci in the profiled genomic regions of the DNA SeqFISH+ data and asked whether, within individual cells, MINTsC-identified multi-way chromatin interactions had smaller maximum pairwise 3D genome distance compared to rest of the cliques. In the top 20% (17/81) of cells with the most profiled cliques (cells with $\geq 6,000$ candidate cliques), MINTsC-identified multi-way interactions had significantly smaller within-clique distances than the remainder (Fig. 2c, p-value < 0.01). With a relaxed threshold (cells with $\geq 2,000$ candidate cliques), 71% of the 62 cells showed the same pattern (p-value < 0.001). Furthermore, pooling data across all cells further supported this observation: MINTsC-identified interactions in the scHi-C (Dip-C) data corresponded closely to

spatial proximity patterns in the DNA seqFISH+ dataset (**Fig.** S4b). We show that the association based on DNA seq FISH+ still hold after matching the cliques with similar total pairwise interaction strength (**Fig.** S3c,d). Compared to the within-assay validation in (i) with scMicro-C, the comparatively reduced separation observed here likely reflects additional cross-modality and between-experiment variability inherent to the DNA seqFISH+ comparison, rather than a limitation of either platform. Lastly, to evaluate the false discovery rates of the method applied to Dip-C data, we leveraged a proximity matrix construction method [19] for converting DNA seqFISH+ imaging data to pairwise interaction data. Specifically, for each individual cell in the DNA seqFISH+ dataset, all the candidate cliques with maximum within clique DNA-seqFISH+ distance less than 150 nm [19] were labelled as true multi-way interactions. Using this gold standard, MINTsC false positive rate did not exceed 3% at the individual cell level.

(iii) Simulation-based evaluation of MINTsC for same-cell co-localization. Although MINTsC infers multi-way chromatin contacts by aggregating pairwise interactions, held-out and per-cell 3D distance evaluations support the co-localization of the identified interactions within cells and their robustness to cellular heterogeneity. We next evaluated this observation with a simulation experiment. We simulated scHi-C contact counts under a multinomial model parameterized from the Lee *et al.* dataset (chromosome 19; 5,582 10 kb bins; L2/3 neurons) (see **Methods** for details on data generation). We spiked in two types of of three-way cliques into the scHi-C contact matrices: (i) triads whose loci co-occurred within individual cells (true cliques) and (ii) triads formed by two pairs co-occurring in one cell group and the remaining pair in another (spurious cliques) (**Fig.** S5a; Methods). Applying MINTsC at an FDR of 0.05 while varying pairwise interaction strengths, the empirical FDR remained well controlled at 5% across settings (**Fig.** S5b). This experiment, specifically designed to lead to spurious aggregation of pairwise contacts as multi-way interactions, supports that the MINTsC framework is robust against aggregation of pairwise interactions originating from different groups of cells.

2.3 SPRITE and scNanoHi-C validation of MINTsC multi-way chromatin interactions with both low and high resolution scHi-C data from human cell lines and mouse embryonic stem cells

After evaluating MINTsC for same cell co-localization, we turned our attention to validation of applications to low and high resolution scHi-C data. We first evaluated MINTsC's ability to infer multi-way chromatin interactions supported by multiple external datasets using GM12878 [20] and mESCs (Serum/Lif) [21] at resolutions of 500Kb and 1Mb, respectively (**Tab.** S1, **Tab.** S2 for details). As in the Section 2.2 scMicro-C and Dip-C applications, we grouped the cliques from scHi-C into two, based on MINTsC clique p-values. Cliques with adjusted[15] clique p-values < 0.01 were labelled as *multi-way interactions*. For evaluation, we utilized the enrichment scores of the SPRITE multi-way interactions introduced in Quinodoz *et al.* [6]. These scores aim to denoise raw SPRITE counts by comparing observed counts of a multi-way interaction against counts of randomly constructed multi-way interactions of the same size while matching the genomic distances within the random multi-way interactions and the observed one. Higher

enrichment scores are interpreted as strong evidence for multi-way interactions[6]. For both the GM12878 and mESC datasets, for which SPRITE data was available, cliques labelled as multi-way interactions by MINTsC exhibited significantly larger enrichment scores compared to cliques that were not significant (**Fig.** 2d,S6a; Wilcoxon rank-sum test p-values < 0.01). Further investigation with clique size stratification revealed that significant cliques from MINTsC exhibit a larger enrichment score mode than that of the insignificant cliques for both GM12878 and mESC (**Fig.** S6b,c). In addition to the alignment of MINTsC with SPRITE, MINTsC inferred multi-way interactions in mESCs are also similarly supported by the GAM triplet score[4]-based evaluation (**Fig.** S7). Furthermore, comparisons of MINTsC *clique z-scores* (**Methods**) against clique-level summarizations of pairwise SPRITE and external bulk Hi-C data revealed strong positive associations (**Fig.** S8, Supplementary Note).

Next, we turned our attention to higher resolution (10Kb) mESC (2i/Lif) scHi-C data and leveraged the recent long-read based single cell high throughput chromatin conformation (scNanoHi-C) data of mESC (2i/Lif)[13] for evaluating MINTsC multi-way chromatin interactions. scNanoHi-C assay generated long sequencing reads that contained multiple DNA monomers, referred to as "concatemers". Consequently, concatemers that include DNA monomers originating from three or more different genomic locations signify interactions that are multi-way. We investigated the genomic distance and depth normalized concatemer counts of multi-way interactions (**Fig.** 2e) and observed that multi-way interactions inferred from MINTsC have significantly higher scNanoHi-C counts across clique sizes compared to cliques that are not deemed as multi-way interactions (one-sided Wilcoxon rank-sum test p-value $< 2 \times 10^{-16}$). We also show that these associations based on SPRITE and scNanoHi-C still hold after matching cliques with similar total pairwise interaction strength (**Fig.** S9). Further stratification of the cliques based on their within clique maximum genomic distance revealed that cliques with maximum within clique genomic distance less than 200Kb are well supported by the scNanoHi-C data (**Fig.** S10), suggesting an empirical genomic distance constraint for post-processing the significant cliques from MINTsC.

Collectively, these results suggest that MINTsC can identify multi-way interactions in both high and low resolution scHi-C data from cell lines. In what follows, we showcase the applicability of MINTsC with more recent human and mouse brain tissue scHi-C datasets.

2.4 Evaluation of MINTsC with high-resolution scHi-C data from complex human and mouse brain tissues

We applied MINTsC to infer multi-way interactions in human prefrontal cortex neuronal cell types (excitatory: L2/3, L4, L5, L6 and inhibitory: Nndf, Pvalb, Sst, Vip) using scHi-C data from sn-m3C-seq assay which simultaneously profiles chromatin conformation capture and methylation at the single cell level [22]. Data for each cell type was binned at 10Kb resolution to generate contact matrices for each cell. While gold standard multi-way chromatin interactions for this system are not available via other technologies, we designed an approach to assess whether the resulting MINTsC clique z-score, i.e., the rth smallest locus pair level z-

value within each clique (Methods), associated with multi-way interactions by leveraging the interplay between methylation and chromatin conformation capture at the single cell level. Specifically, we utilized the well-established observation that DNA methylation level correlations (either positive or negative) are indicative of regulatory relationships among the loci [23-25]. In fact, recent work established that DNA methylation profiles and chromatin accessibility as measured by ATAC-seq at the single nuclei level validated each other at the cell-subtype level [26]. Hence, genomic loci within a multi-way interaction can be expected to exhibit a dependency structure in their DNA methylation profiles. To assess such dependence, we leveraged partial correlation coefficient and quantified the level of conditional dependence between pairs of genomic loci given the other loci in the multi-way interactions. For example, for a three way interaction, if any of the pairs exhibit low partial correlation in absolute value given the third locus in the candidate interaction, this would indicate that pairwise dependencies within the candidate multi-way interaction are not supported by the co-methylation patterns. With this line of reasoning, we then compared MINTsC clique z-scores with the average DNA methylation-based partial correlations between all the loci of individual cliques, at clique sizes of 3 and 4 and across all the eight cell types (**Fig.** 2f,S11). This analysis revealed that cliques with larger MINTsC clique z-scores tend to have larger multi-way dependency levels across all clique sizes for each cell type. We further assessed the association between the MINTsC clique z-scores and partial-correlation based multi-way dependencies (slopes of the fitted regression lines, **Fig.** 2g). In this quantification, majority of the comparisons yielded a significant and positive association between the MINTsC clique z-scores and partial-correlation based evidence for multi-way interaction.

Having observed that the MINTsC multi-way chromatin interactions exhibit single-cell level co-localization based on DNA seqFISH+ 3D distances in Section 2.2, we performed a comparative analysis between the MINTsC clique z-scores and the DNA seqFISH+ distances summarized at the clique level. We focused on the cell types shared by both datasets (Dip-C[18] and DNA seqFISH+[19]) and compared the clique z-scores with the $\left(\binom{N-1}{2}+1\right)$th closest spatial distance between loci pairs within the N-way cliques. Specifically, we chose the order number as $\binom{N-1}{2}+1$ to ensure that all the N loci are participating to form a connected graph. This comparison indicated a consistent inverse relationship (high MINTsC z-score with low spatial distance) across the cell types (asctrocyte, microglia, and oligodendrocyte progenitor cells) and clique sizes (3,4) (**Fig.** S12a,S13). A permutation-based evaluation confirmed that loci within MINTsC-inferred multi-way interactions were significantly closer spatially than expected by chance (**Fig.** S12b). Assessment of this inverse association with slopes of the fitted regression lines (**Fig.** S12c) further yielded a significant negative association between the MINTsC clique z-scores and the spatial distances of loci within multi-way interactions for majority of the cell types.

2.5 Application of MINTsC to haplotype-resolved Dip-C data

It is worth remarking that the mouse brain scHi-C data from the Dip-C assay[18] is haplotype-resolved. This allowed us to investigate multi-way interactions in a haplotype-aware manner. Specifically, we deployed MINTsC on single cell contact counts of individual haplotypes. This yielded that MINTsC clique p-values are well calibrated (**Fig.** S14), making it readily applicable to haplotype-resolved scHi-C data. A comparative analysis of the MINTsC results for the two haplotypes revealed that the p-values for most cliques are concordant between the two haplotypes across cell types. Further comparison of the haplotype-aware analysis with the aggregate analysis that sums over contact matrices of the two haplotypes suggested that while aggregate analysis results in some power loss in detecting haplotype-specific multi-way chromatin interactions as expected, it does not lead to inflation of false positives (Supplementary Note, **Tab.** S3).

2.6 MINTsC identifies significantly more multi-way interactions with higher reliability compared to baseline approach

To the best of our knowledge, MINTsC is the first approach for inferring multi-way chromatin interactions from scHi-C contact matrices. We compared MINTsC to two baseline approaches that we devised from standard analysis of scHi-C data. Specifically, we considered multi-way interactions in the form of cliques from **1)** the significant loop calls of SnapHiC [11] (i.e., significant pairwise interactions) and from **2)** strong pairwise interactions with pseudo-bulk O/E (Observed/Expected) ≥ 2 based on cell type specific scHi-C data. We deployed these baseline approaches on the human prefrontal cortex scHi-C data. The first baseline approach did not yield cliques with order larger than 3; hence, we retained the comparisons to cliques of size 3. We thresholded the candidate cliques from MINTsC at FDR 0.05 to obtain significant 3-way interactions.

Next, we employed external PsychENCODE data [27] to evaluate the resulting multi-way interactions of MINTsC and the baseline methods. Specifically, we utilized the compiled neuronal cell type gene-enhancer interactions which encompassed one or more interactions per gene. After filtering the cliques of each method so that at least one locus in the clique harbored at least one gene, we quantified how many of the method cliques were of the form $g - e_1 - e_2$ structure (v-structure), where g refers to the gene with enhancers e_1 and e_2 in the gene-enhancer link list from the PsychENCODE source. Specifically, for each gene g within a locus of a clique q, we assigned the clique q a score of 1, if the other loci in the clique individually overlapped the PsychENCODE enhancers of the gene g. Under this scoring scheme, a clique could have score s if it contains s number of $g - e_1 - e_2$ v-structures. Next, we quantified proportion of method cliques with scores $s \geq 1$ (**Fig.** 2h). This calculation was carried out separately for each cell type and yielded that MINTsC outperforms the baseline approaches in terms of capturing the $g - e_1 - e_2$ structures across cell types (**Fig.** S15).

We further conducted baseline comparisons using synthetic scHi-C datasets, mimicking the characteristics of actual scHi-C data (**Fig.** S16a). Specifically, we leveraged the Lee *et*

al. [22] scHi-C dataset (chromosome 19 with 5,582 10Kb-sized bins, and L2/3 neurons) and multinomial distribution for the scHi-C counts in generating synthetic data for the comparisons. To ensure a more heterogeneous background contact count distribution than accounted for by MINTsC (i.e., not strictly following the MINTsC background), we randomly implanted true (300) and false (2,100) cliques in the estimated background contact probability by varying the signal sizes of the pairwise contacts within true cliques (**Methods** for details on the synthetic data generation). We then evaluated the False/True Positive Rates (FPR/TPR) across the simulation replications. While the TPRs of the considered approaches were comparable, MINTsC exhibited significantly lower FPRs (< 0.05) (**Fig.** 2i) across signal sizes of the true cliques, due to its well-calibrated pairwise and multi-way interaction p-values (**Fig.** S16b,c,d and **Fig.** S17). These results underscore MINTsC's capability to identify multi-way chromatin interactions with more reliable FDR control while maintaining comparable power to natural baseline methods.

2.7 Biological implications of MINTsC-inferred multi-way interactions involving genes and multiple enhancers

After establishing that multi-way interactions inferred by MINTsC are supported by external data, we turned our attention to specific multi-way interactions that involve genes and enhancers. For these investigations, we leveraged (i) Activity-By-Contact (ABC) scores [28] that infer gene-enhancer interactions based on genomewide profiling of chromatin accessibility and activating histone mark H3K27ac; and (ii) gene-enhancer interactions inferred from multi-modal profiling of single cells for chromatin accessibility and gene expression [29].

We considered the mESC[30] cliques evaluated by MINTsC and leveraged the ABC scores from Narita *et al.* [31] to asses whether multi-way interactions with at least one gene are supported by the ABC scores of the involved enhancers. A total of 6,030 cliques harbored genes regulated by multiple enhancers within the same clique, i.e., cliques containing a full $g-e_1-e_2-\cdots-e_N$ structure, where g refers to a gene with ABC scores evaluated for enhancers $e_1, e_2, \cdots, e_N$. Specifically, we required the cliques to have one locus overlapping with a gene and have the other clique loci each overlap with at least one enhancer for which the gene has a reported ABC score. For each clique satisfying this requirement and each $g-e_1-e_2-\cdots-e_N$ structure in the clique (6,584 cases in total), we took the average of the ABC scores assigned for all $g-e_j$ links, $j \in \{1, \cdots, N\}$, to summarize ABC scores at the clique level. Then, we assessed if the multi-way interactions inferred from MINTsC (FDR cutoff of 0.05) are supported by the multi-way gene regulation level, which is represented by the clique level ABC score. This comparison revealed a significant difference in summarized ABC scores of significant and insignificant cliques (**Fig.** S18, one-sided Wilcoxon rank-sum test $p < 2e-16$).

Next, for the Lee *et al.*[22] human prefrontal cortex scHi-C data analyzed at 10Kb resolution, we conducted a similar analysis for the cliques with a full $g - e_1 - e_2 - \cdots - e_N$ structure, with the enhancers e_js from the set of prefrontal cortex enhancers identified by an integrative analysis of snATAC-seq and scRNA-seq data (further details in **Fig.** 2 of Li *et al.*

[29]). Specifically, Li *et al.* [29] provides a set of $g - e_j$ links referred to as positively correlated gene-cCRE connections. We compared the MINTsC $-\log_{10}$(p-scores) of the cliques containing a $g - e_1 - e_2 - \cdots - e_N$ (with positively correlated gene-cCRE connections) against those of randomly selected cliques and observed that cliques supported by the gene-cCRE connections tend to have markedly stronger MINTsC multi-way interaction levels (**Fig.** S19). A consistent result can be observed across cell types based on a more systematic comparison with multiple replicates of the randomly selected cliques (**Fig.** 3a).

To probe biological implications of multi-way interactions, we asked whether genes with MINTsC-inferred multi-way interactions exhibited different gene expression patterns than those without multi-way interactions by leveraging the Lee *et al.*[22] scHi-C data. Utilizing SMART-seq data [32] of the same system, we compared expression levels of the genes with and without multi-way interactions for individual cell types. **Fig.** 3b compares the expression distribution of the genes within the two groups and shows that the genes with multi-way interactions on average exhibit significantly higher expression levels throughout most of the cell types (Wilcoxon rank-sum test p-values ≤ 0.003). This agrees well with the findings from the long-read based scHi-C which revealed a similar result for GM12878 cells [33].

2.8 MINTsC-derived multi-way interactions in human prefrontal cortex enable discovery of potential epistatic SNP effects

To establish the broader biological discovery potential of identifying multi-way interactions, we investigated whether MINTsC could facilitate discovery of epistatic SNP effects in eQTL analysis. Specifically, we focused on pre-frontal cortex multi-way interactions (FDR$<$0.05) involving genes and their enhancers supported by cCRE scores (**Tab.** S4). We leveraged the ROS/MAP study (**Methods**) of Alzheimer's disease, and assessed whether variation in cerebral cortex expression of genes involved in multi-way interactions inferred by MINTsC were better explained by an eQTL model with SNP-SNP interactions compared to a linear additive model. We specifically investigated the p-values of the ANOVA F-test statistics for the null hypothesis $H_0 : \beta_{12} = 0$ comparing the full model with a two-way SNP interaction term, i.e., Expression $= \text{SNP}_1\beta_1 + \text{SNP}_2\beta_2 + (\text{SNP}_1 : \text{SNP}_2)\beta_{12} + \text{error}$, against the reduced model without the SNP interaction term, i.e., Expression $= \text{SNP}_1\beta_1 + \text{SNP}_2\beta_2 + \text{error}$, of each $(\text{Gene}, \text{SNP}_1, \text{SNP}_2)$ tuple within a MINTsC multi-way interaction supported by non-zero cCRE scores. Evaluation of such 39 genes with multi-way interactions revealed that SNPs within multi-way interactions exhibited significantly stronger SNP-SNP interaction effects than expected by chance. Specifically, the SNP-SNP interaction F-test $-\log_{10}$(p-values) of the 321 $(\text{Gene}, \text{SNP}_1, \text{SNP}_2)$ tuples were notably greater overall than those obtained from permuting the gene expression levels (**Fig.** 3c). A striking example is observed for the gene DKK3, which is linked to Amyloid-β pathology and synapse restoration [34] (**Fig.** S20, **Fig.** 3d,e). MINTsC identified a multi-way interaction involving $\pm$ 200bp of the DKK3 TSS and SNPs rs7480026 and rs16910272 in the 3rd and 5th introns of DKK3, respectively. Statistical testing in a linear model of DKK3 expression revealed a significant interaction effect for the DKK3 expression

(FDR$\leq$0.05). Specifically, this analysis highlighted that while each SNP individually has a weak effect, their interaction significantly contributes to DKK3 expression variation. A similar epistatic effect was revealed for other genes (**Tab.** S5), including CPLX2, linked to cognitive resilience and synaptic plasticity [35] (**Fig.** S21). These results underscore MINTsC's potential to discover epistatic effects in molecular QTL studies by generating candidate interactions to test for and, thereby, significantly reducing the multiple testing burden.

3 Discussion

We presented MINTsC as an empirical Bayes framework to infer multi-way interactions from pairwise interaction signals quantified in scHi-C data. Despite the limitations due to data sparsity and proximity ligation, scHi-C data harbors cliques representing potential multi-way interactions (albeit of lower sizes ≤ 6) and presents an exciting opportunity to learn higher-order chromatin interactions. MINTsC builds on a dirichlet-multinomial model that takes into account inherent genomic distance bias of scHi-C data by a natural spline model. We developed a computationally efficient estimation strategy for MINTsC and showed that null generative model of MINTsC results in well calibrated p-values, enabling appropriate false discovery rate control (**Fig.** S22, S23, S24, S25, S26). Evaluations of MINTsC using a wide range of real scHi-C datasets yielded multiple-levels of support for MINTsC multi-way interactions from independent external genomic data resources. In addition, multi-way interactions involving genes and their enhancers were strongly supported by their corresponding ABC scores of gene regulation. This finding highlights the potential of MINTsC to infer multi-way gene regulation. Despite the lack of computational methods aimed at identifying multi-way interactions from scHi-C data, we considered a baseline approach by utilizing state-of-the art scHi-C loop calling tool SnapHi-C[11] and generated multi-way interactions as cliques from significant loop calls. Applications of the baseline approach and MINTsC on scHi-C from human prefrontal cortex cells and synthetic data mimicking actual data yielded that MINTsC is able to identify more multi-way interactions with markedly higher reliability in terms of FDR control. As biological discovery potential of MINTsC, we showcased that MINTsC-derived multi-way interactions in human prefrontal cortex enable discovery of potential epistatic SNP effects. Striking examples included strong SNP-SNP interactions on the expression levels of the genes related to Amyloid-β pathology (DKK3) and synaptic plasticity (CPLX2) based on the ROS/MAP study of Alzheimer's disease. We expect that with the continued development of scHi-C technologies, MINTsC analysis of scHi-C data will complement data from other multi-way interaction assays.

While MINTsC applications presented here relied on known cell type labels, this is not an inherent limitation because there is a plethora of computational tools for inferring cell type labels from scHi-C data [11,36-42]. Furthermore, once we quantify the uncertainty of label assignments (e.g., by soft-threshold in clustering of low-dimensional embeddings of single cell data), these uncertainity estimates can be naturally incorporated into the generalized linear model fit that MINTsC employs. More generally, from a methodological point of view, the two-stage

(band level and locus pair level) modeling by MINTsC offers a computationally useful framework for any type of high-dimensional and sparse single cell genomics/epigenomics data where exponential family of distributions is a valid generative distribution choice. Specifically, if a group of genomic loci (or genes) can be bundled into a larger class (analogous to a band in MINTsC) for a GLM fitting, then the original locus level estimation (analogous to locus pair level estimation in MINTsC) can be carried out based on estimate allocation procedure introduced in **Fig.** 1a.

Competing Interests

The authors declare no competing financial interests.

Correspondence

Sündüz Keleş (keles@wisc.edu).

Acknowledgments

This work was supported by National Institutes of Health (NIH) grants (HG003747, HG012881) and a Chan Zuckerberg Initiative (CZI) single-cell data insights grant (cycle # 1). We thank Dr. Longzhi Tan from Stanford University for insightful discussions regarding the method validation, Dr. Jiansen Lu from Peking University for sharing detailed information and discussions about the scNanoHi-C data, Dr. Takeo Narita from Copenhagen University for sharing the ABC scores of the mESC system, and Siqi Shen from the University of Wisconsin-Madison for sharing processed Nagano *et al.* scHi-C data. We also thank Dr. Ye Zheng from the University of Texas MD Anderson Cancer Center for a useful discussion regarding the baseline method comparison. The results published here are in whole or in part based on data obtained from the AD Knowledge Portal (https://adknowledgeportal.org). Study data were provided by the Rush Alzheimer's Disease Center, Rush University Medical Center, Chicago. Data collection was supported through funding by NIA grants P30AG10161 (ROS), R01AG15819 (ROSMAP; genomics and RNAseq), R01AG17917 (MAP), R01AG30146, R01AG36042 (5hC methylation, ATAC-seq), RC2AG036547 (H3K9Ac), R01AG36836 (RNAseq), R01AG48015 (monocyte RNAseq) RF1AG57473 (single nucleus RNAseq), U01AG32984 (genomic and whole exome sequencing), U01AG46152 (ROSMAP AMP-AD, targeted proteomics), U01AG46161(TMT proteomics), U01AG61356 (whole genome sequencing, targeted proteomics, ROSMAP AMP-AD), the Illinois Department of Public Health (ROSMAP), and the Translational Genomics Research Institute (genomic). Additional phenotypic data can be requested at www.radc.rush.edu.

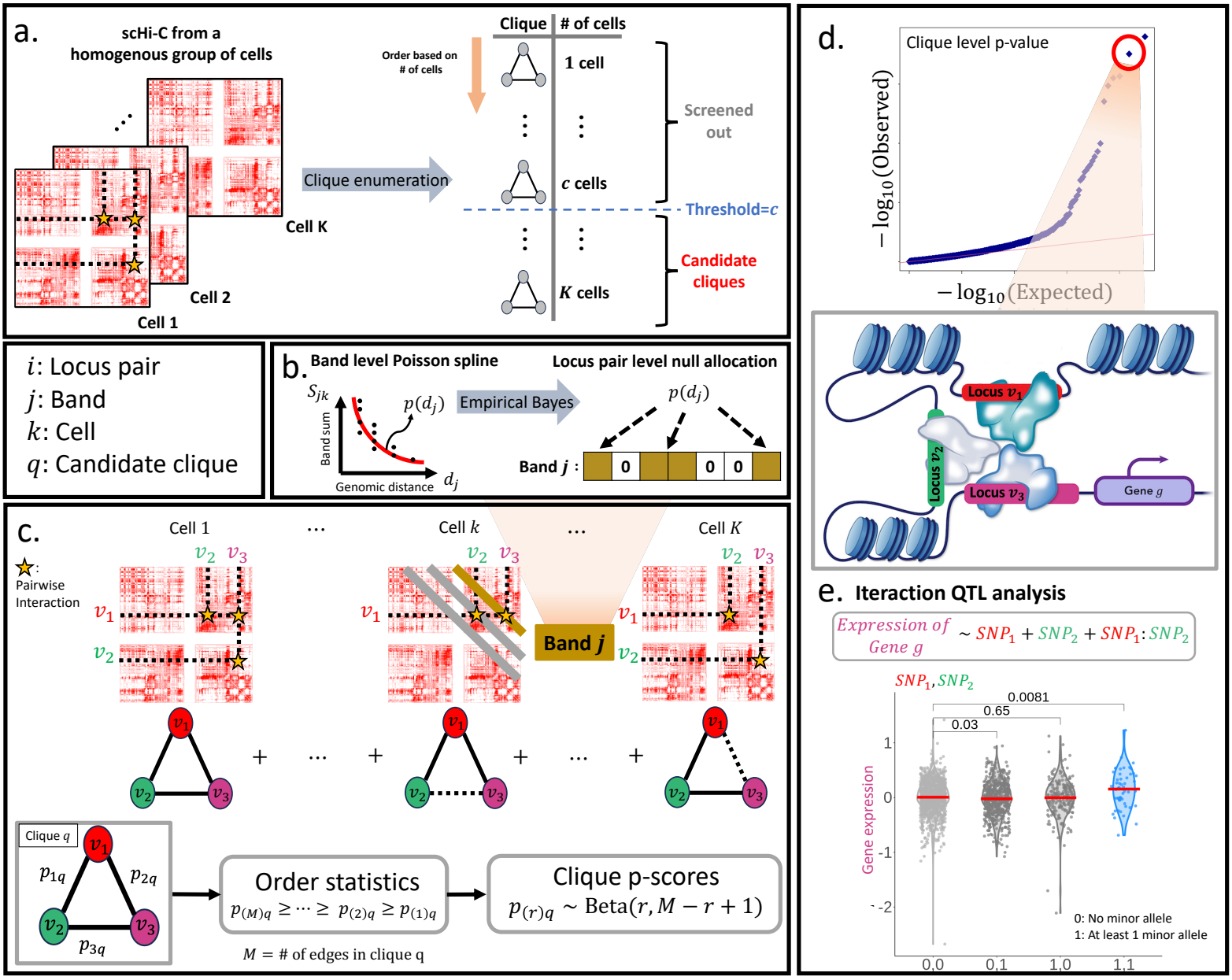
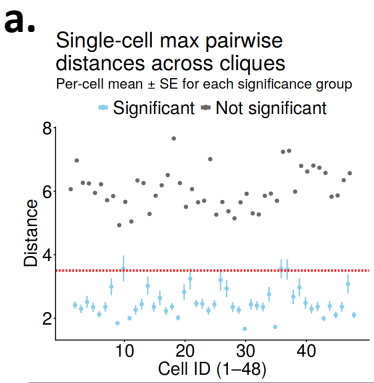
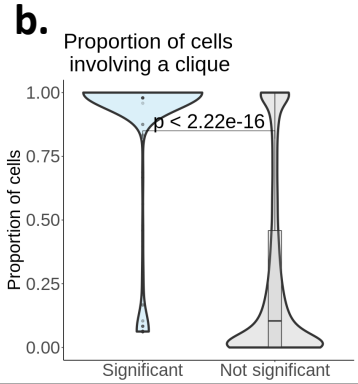

a.
scHi-C from a homogenous group of cells
Cell 1
Cell 2
Cell K
Clique enumeration
Order based on # of cells
Clique
# of cells
1 cell
c cells
K cells
Screened out
Threshold=c
Candidate cliques
i: Locus pair
j: Band
k: Cell
q: Candidate clique
b.
Band level Poisson spline
Locus pair level null allocation
Empirical Bayes
Band sum
Genomic distance
Band j :
c.
Cell 1
Cell k
Cell K
Pairwise Interaction
Band j
Clique q
Order statistics
Clique p-scores
M = # of edges in clique q
d.
Clique level p-value
−log10(Observed)
−log10(Expected)
Locus v1
Locus v2
Locus v3
Gene g
e. Iteraction QTL analysis
Expression of Gene g ~ SNP1 + SNP2 + SNP1:SNP2
SNP1, SNP2
0.0081
0.65
0.03
Gene expression
0: No minor allele
1: At least 1 minor allele
0,0
0,1
1,0
1,1

Figure 1 A schematic overview of MINTsC for learning multi-way interactions from scHi-C data. a. MINTsC first employs a clique search algorithm on the scHi-C contact matrices of a homogeneous group of cells, e.g., a cell type, and enumerate how many times each clique appears across cells. Then, a cell number threshold is applied to the set of cliques searched to define *candidate cliques*. **b.** MINTsC fits a locus-pair level multinomial model to scHi-C contact counts by a multinomial aggregation and poisson approximation strategy at the band level. Contact counts y_{ijk} of locus pairs, $i \in [N_j]$, across bands, $j \in [J]$, of a cell $k \in [K]$ follow a multinomial distribution. Band-level aggregated counts for each band j within cell k , $S_{jk} := \sum_{i=1}^{N_j} y_{ijk}$, follows a poisson distribution with a rate parameter that depends on the band-level count probability which is modeled as natural cubic spline function $p_j = p(d_j)$ of genomic distance d_j between the locus pairs in the band. Estimated band level probabilities, p_j, are distributed across the locus-pairs within a band with under a null Dirichlet prior to the multinomial counts (**Methods**). **c.** MINTsC multi-way interaction scores (i.e., clique-level test statistics) are based on the order statistics of pairwise interaction test statistics, e.g., p-values. Square matrices denote scHi-C contact matrices of K cells for a fixed chromosome. The starred entries depict the pairwise contacts among three genomic loci (i.e., nodes) v_1, v_2, v_3, displayed as an undirected graph in the bottom. The gray or gold bars along the off diagonals denote bands. MINTsC aggregates the pairwise contact counts within a given candidate clique across the cells of the cell type, and generates clique-level statistic, which are clique p-score $p_{(r)q}$ (**Eqn.** 7 in **Methods**) and clique z-score $Z_{(M-r+1)q}$ (**Eqn.** 8 in **Methods**). **d.** Q-Q plot of the $-\log_{10}$(clique p-values), with x- and y-axes denoting $-\log_{10}$ transformed uniform random variables and the observed clique p-values (**Eqn.** 9 in **Methods**), respectively. **e.** Illustration of a multi-way interaction among genomic loci harboring two enhancers and a gene promoter. MINTsC multi-way interactions can be probed for epistatic SNP effects in molecular QTL studies or GWAS.

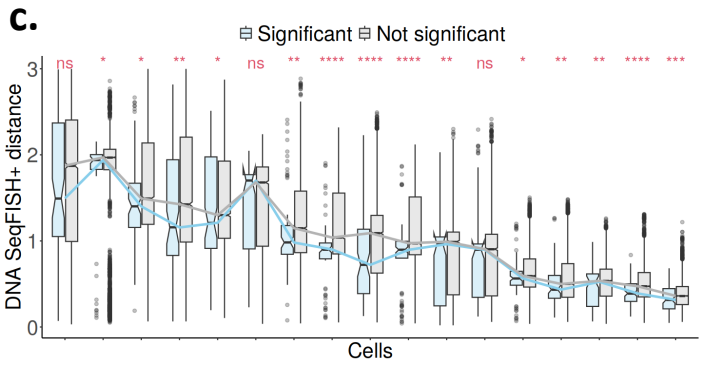
Evaluation

scMicro-C 3D spatial distance

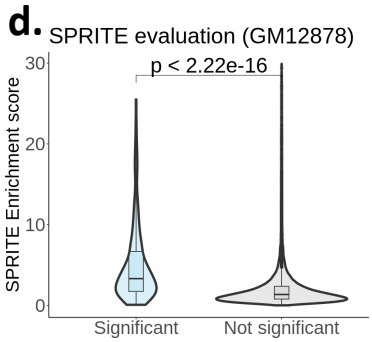


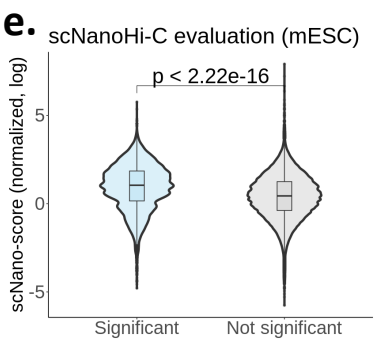


DNA seqFISH+ spatial distance

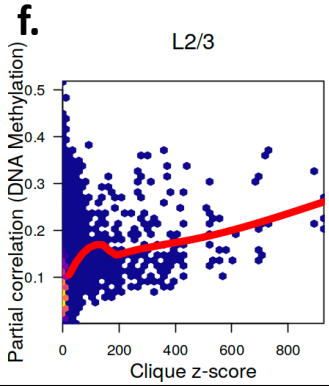


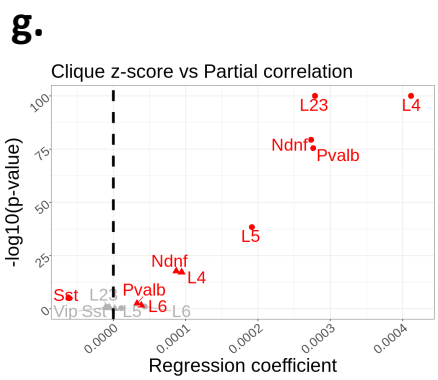
Multi-way chromatin contacts





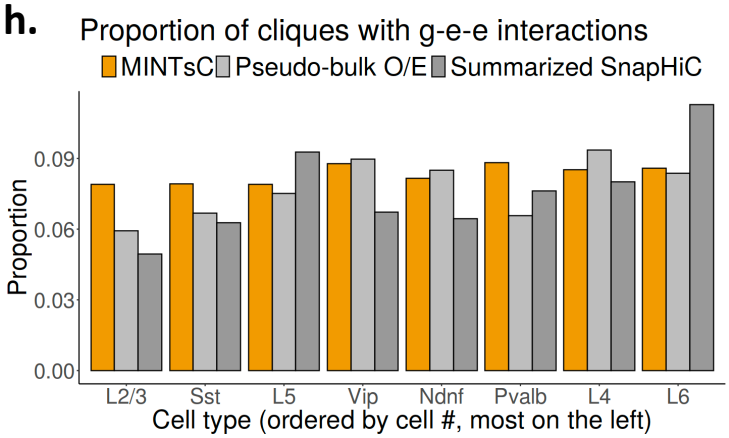
DNA methylation from snm3C-seq





Baseline comparison

PsychENCODE gene-enhancer links



Synthetic data

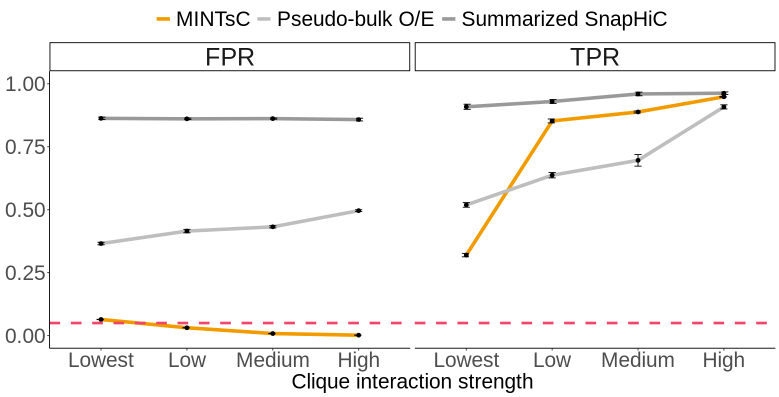


Figure 2 MINTsC identifies multi-way interactions supported by a multitude of external data and with biological implications in gene regulation. a. Per-cell mean and two-step standard deviation of the maximum pairwise 3D spatial distance within each clique, stratified by significance group. The red dashed line marks the 3.5 particle radii cutoff suggested by Wu *et al.*[14] **b.** Proportion of cells where a clique's maximum 3D distance is below 3.5 particle radii ($\sim$240 nm) [14], comparing significant and non-significant MINTsC cliques in GM12878 scMicro-C contact matrices[14]. **c.** Comparison of within-clique DNA seqFISH+ distances between significant and insignificant MINTsC candidate cliques for the top 20% of cells with the highest number of candidate cliques (as indicated by the red vertical line in **Fig.** S4a. The x-axis indexes individual cells. Blue and gray lines denote the trend of DNA seqFISH+ distances across cells for significant and insignificant MINTsC cliques, respectively. **d.** Comparison of SPRITE enrichment scores [6] between the significant MINTsC cliques (FDR$<$0.01) and the others for GM12878 [43]. **e.** Similar comparison for MINTsC mESC multi-way interactions (FDR$<$0.05) using normalized scNanoHi-C [13] concatemer counts. **e.** Comparison of clique z-scores (x-axis) with average absolute partial correlations of locus pairs within each clique based on single cell DNA methylation (y-axis) for L2/3 cells (**Fig.** S11 for other cell types). **f.** $-\log_{10}$(p-values) (y-axis) of the estimated regression coefficients (x-axis) from regressing the MINTsC clique z-scores on the partial correlation summary of DNA methylation. The cases with BH FDR$<$0.05 are labelled as significant. **h.** Proportion of gene harboring cliques with the full $g - e_1 - e_2$ based on PsychENCODE gene-enhancer interactions. Cell types are ordered based on the number of cells from left to right. **i.** Comparison of MINTsC with baseline approaches using synthetic data (**Methods** for details). FPR and TPR for each method are quantified under different signal scenarios for multi-way interactions (x-axis) (Red line : $y = 0.05$).

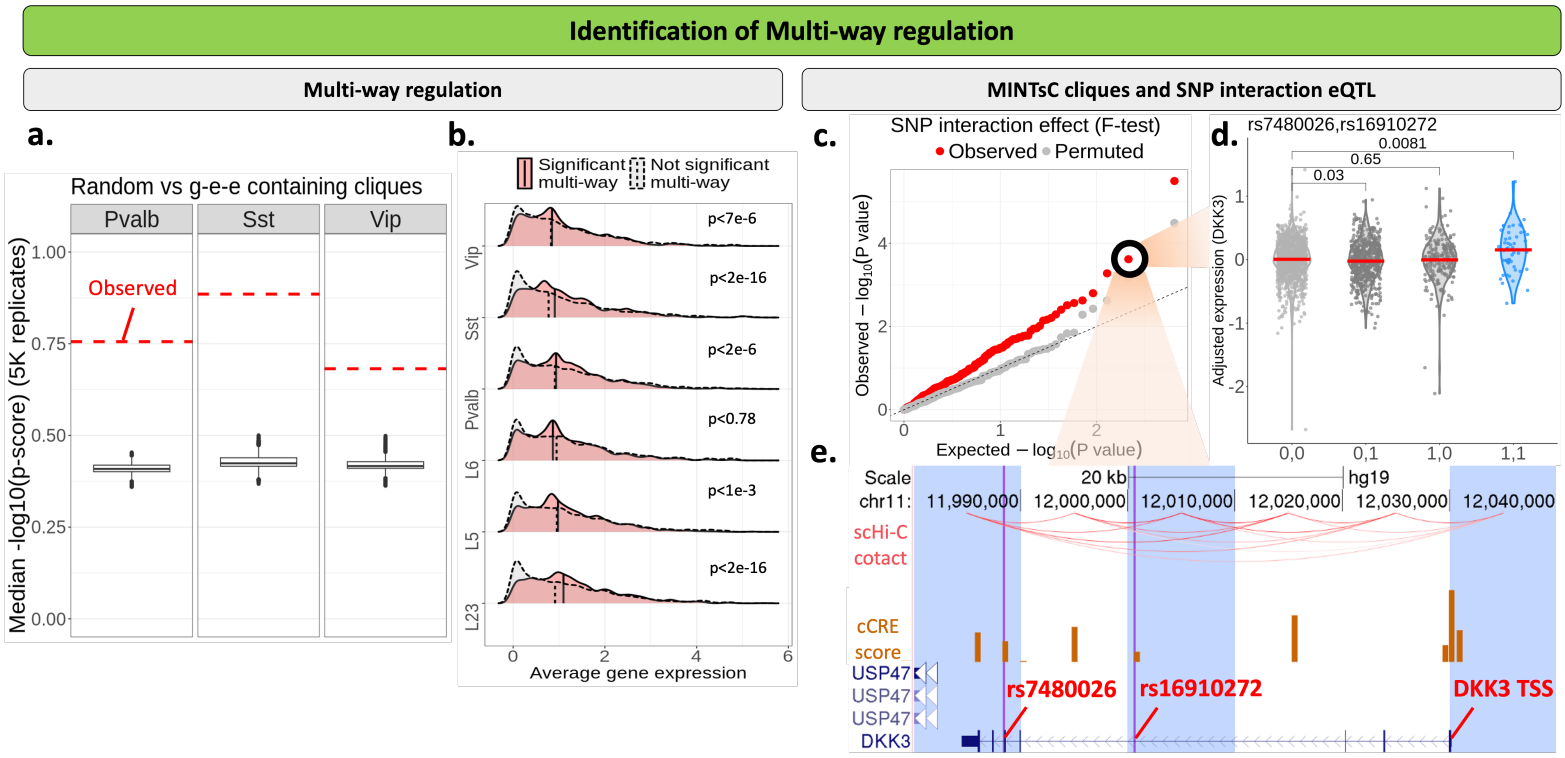


Figure 3 MINTsC identifies multi-way interactions supported by biological implications in gene regulation. **a.** Summarization of the median MINTsC $-\log_{10}$(p-scores) of the randomly sampled cliques across 5K replicates. The red line on each panel (cell type) denotes the median $-\log_{10}$(p-scores) of the $g - e_1 - e_2$ (from the gene-cCRE list[29]) harboring cliques. **b.** Comparison of average (normalized) gene expression levels among cliques with (FDR$<$0.1) and without significant multi-way interactions across the cell types in human brain prefrontal cortex[22]. p-values above each density plot is based on one-sided Wilcoxon rank-sum test. **c.** QQ-plot displaying F-test $-\log_{10}$(p-values) of SNP interaction effects on gene expression from AD ROS/MAP study (red) and that of under random permutation (grey). Each red dot is from a two-way SNP interaction eQTL model, where the gene TSS and SNP locations are identified as a multi-way interaction by MINTsC (Vip cells). **d.** Covariate-adjusted expression levels of DKK3 across samples for combinations of alleles of

SNPs rs16910272 and rs7480026. For each SNP, ‘0’ indicates no minor allele, and ‘1’ indicates existence of at least one minor allele. BH adjusted p-values are reported for each comparison. **e.** UCSC Genome Browser display of the DKK3 multi-way interaction. Orange bars depict ATAC-seq signal summarized by cCRE scores[29] and the red arcs denote the pseudo-bulk scHi-C contacts [22] for Vip cells. Purple vertical lines within each blue shaded 10Kb bin denote the SNP locations with significant interaction effect on DKK3 expression.

1. Lin, X., Liu, Y., Liu, S., Zhu, X., Wu, L., Zhu, Y., Zhao, D., Xu, X., Chemparathy, A., Wang, H., Cao, Y., Nakamura, M., Noordermeer, J.N., La Russa, M., Wong, W.H., Zhao, K., Qi, L.S.: Nested epistasis enhancer networks for robust genome regulation. Science **377**(6610) (September 2022) 1077–1085

2. Deplancke, B., Alpern, D., Gardeux, V.: The genetics of transcription factor dna binding variation. Cell **166**(3) (2016) 538–554

3. Buniello, A., MacArthur, J.A.L., Cerezo, M., Harris, L.W., Hayhurst, J., Malangone, C., McMahon, A., Morales, J., Mountjoy, E., Sollis, E., et al.: The nhgri-ebi gwas catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. Nucleic acids research **47**(D1) (2019) D1005–D1012

4. Beagrie, R.A., Scialdone, A., Schueler, M., Kraemer, D.C.A., Chotalia, M., Xie, S.Q., Barbieri, M., De Santiago, I.and Lavitas, L.M., Branco, M.R., Fraser, J., Dostie, J., Game, L., Dillon, N., Edwards, P.A.W., Nicodemi, M., Pombo, A.: Complex multi-enhancer contacts captured by genome architecture mapping. Nature **543**(7646) (March 2017) 519–524

5. Zheng, M., Tian, S.Z., Capurso, D., Kim, M., Maurya, R., Lee, B., Piecuch, E., Gong, L., Zhu, J.J., Li, Z., Wong, C.H., Ngan, C.Y., Wang, P., Ruan, X., Wei, C.L., Ruan, Y.: Multiplex chromatin interactions with single-molecule precision. Nature **566**(7745) (February 2019) 558–562

6. Quinodoz, S.A., Ollikainen, N., Tabak, B., Palla, A., Schmidt, J.M., Detmar, E., Lai, M.M., Shishkin, A.A., Bhat, P., Takei, Y., Trinh, V., Aznauryan, E., Russell, P., Cheng, C., Jovanovic, M., Chow, A., Cai, L., McDonel, P.and Garber, M., Guttman, M.: Higher-Order Inter-chromosomal Hubs Shape 3D Genome Organization in the Nucleus. Cell **174**(3) (July 2018) 744–757.e24

7. Arrastia, M.V., Jachowicz, J.W., Ollikainen, N., Curtis, M.S., Lai, C., Quinodoz, S.A., Selck, D.A., Ismagilov, R.F., Guttman, M.: Single-cell measurement of higher-order 3D genome organization with scSPRITE. Nature Biotechnology **40**(1) (January 2022) 64–73

8. Vermeulen, C., Allahyar, A., Bouwman, B.A.M., Krijger, P.H.L., Verstegen, M.J.A.M., Geeven, G., Valdes-Quezada, C., Renkens, I., Straver, R., Kloosterman, W.P., De Ridder, J., De Laat, W.: Multi-contact 4C: long-molecule sequencing of complex proximity ligation products to uncover local cooperative and competitive chromatin topologies. Nature Protocols **15**(2) (February 2020) 364–397

9. Darrow, E.M., Huntley, M.H., Dudchenko, O., Stamenova, E.K., Durand, N.C., Sun, Z., Huang, S., Sanborn, A.L., Machol, I., Shamim, M., Seberg, A., Lander, E.S., Chadwick, B.P., Aiden, E.L.: Deletion of *DXZ4* on the human inactive X chromosome alters higher-order genome architecture. Proceedings of the National Academy of Sciences **113**(31) (August 2016)

10. Zhong, J.Y., Niu, L., Lin, Z.B., Bai, X., Chen, Y., Luo, F., Hou, C., Xiao, C.L.: High-throughput Pore-C reveals the single-allele topology and cell type-specificity of 3D genome folding. Nature Communications **14**(1) (March 2023) 1250

11. Yu, M., Abnousi, A., Zhang, Y., Li, G., Lee, L., Chen, Z., Fang, R., Wen, J., Sun, Q., Li, Y., Ren, B., Hu, M.: SnapHiC: a computational pipeline to map chromatin contacts from single cell Hi-C data. . https://www.biorxiv.org/content/10.1101/2020.12.13.422543v1 (12 2020)

12. Li, X., Lee, L., Abnousi, A.and Yu, M., Liu, W., Huang, L.and Li, Y., Hu, M.: SnapHiC2: A computationally efficient loop caller for single cell Hi-C data. Computational and Structural Biotechnology Journal **20** (2022) 2778–2783

13. Li, W., Lu, J., Lu, P., Gao, Y., Bai, Y., Chen, K., Su, X., Li, M., Liu, J., Chen, Y., Wen, L., Tang, F.: scNanoHi-C: a single-cell long-read concatemer sequencing method to reveal high-order chromatin structures within individual cells. Nature Methods **20**(10) (October 2023) 1493–1505

14. Wu, H., Zhang, J., Tan, L., Xie, X.S.: Single-cell micro-c profiles 3d genome structures at high resolution and characterizes multi-enhancer hubs. Nature Genetics (2025) 1–10

15. Benjamini, Y., Hochberg, Y.: Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal statistical society: series B (Methodological) **57**(1) (1995) 289–300

16. Zhang, R., Ma, J.: MATCHA: Probing Multi-way Chromatin Interaction with Hypergraph Representation Learning. Cell Systems **10**(5) (May 2020) 397–407.e5

17. Tan, L., Xing, D., Chang, C.H., Li, H., Xie, X.S.: Three-dimensional genome structures of single diploid human cells. Science **361**(6405) (2018) 924–928

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

19. Takei, Y., Zheng, S., Yun, J., Shah, S., Pierson, N., White, J., Schindler, S., Tischbirek, C.H., Yuan, G.C., Cai, L.: Single-cell nuclear architecture across cell types in the mouse brain. Science **374**(6567) (2021) 586–594

20. Ramani, V., Deng, X., Qiu, R., Gunderson, K.L., Steemers, F.J., Disteche, C.M., Noble, W.S., Duan, Z., Shendure, J.: Massively multiplex single-cell Hi-C. Nature methods **14**(3) (2017) 263–266

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

22. 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.: Simultaneous profiling of 3D genome structure and DNA methylation in single human cells. Nature methods **16**(10) (2019) 999–1006

23. Akulenko, R., Helms, V.: Dna co-methylation analysis suggests novel functional associations between gene pairs in breast cancer samples. Human molecular genetics **22**(15) (2013) 3016–3022

24. Wang, F., Xu, H., Zhao, H., Gelernter, J., Zhang, H.: Dna co-methylation modules in postmortem prefrontal cortex tissues of european australians with alcohol use disorders. Scientific reports **6**(1) (2016) 19430

25. Kachroo, P., Morrow, J.D., Kho, A.T., Vyhlidal, C.A., Silverman, E.K., Weiss, S.T., Tantisira, K.G., DeMeo, D.L.: Co-methylation analysis in lung tissue identifies pathways for fetal origins of copd. European Respiratory Journal **56**(4) (2020)

26. Liu, H., Zhou, J., Tian, W., Luo, C., Bartlett, A., Aldridge, A., Lucero, J., Osteen, J.K., Nery, J.R., Chen, H., Rivkin, A., Castanon, R.G., Clock, B., Li, Y.E., Hou, X., Poirion, O.B., Preissl, S., Pinto-Duarte, A., O'Connor, C., Boggeman, L., Fitzpatrick, C., Nunn, M., Mukamel, E.A., Zhang, Z., Callaway, E.M., Ren, B., Dixon, J.R., Behrens, M.M., Ecker, J.R.: DNA methylation atlas of the mouse brain at single-cell resolution. Nature **598**(7879) (October 2021) 120–128

27. Wang, D., Liu, S., Warrell, J., Won, H., Shi, X., Navarro, F.C., Clarke, D., Gu, M., Emani, P., Yang, Y.T., et al.: Comprehensive functional genomic resource and integrative model for the human brain. Science **362**(6420) (2018) eaat8464

28. Fulco, C.P., Nasser, J., Jones, T.R., Munson, G., Bergman, D.T., Subramanian, V., Grossman, S.R., Anyoha, R., Doughty, B.R., Patwardhan, T.A., et al.: Activity-by-contact model of enhancer–promoter regulation from thousands of crispr perturbations. Nature genetics **51**(12) (2019) 1664–1669

29. Li, Y.E., Preissl, S., Miller, M., Johnson, N.D., Wang, Z., Jiao, H., Zhu, C., Wang, Z., Xie, Y., Poirion, O., et al.: A comparative atlas of single-cell chromatin accessibility in the human brain. Science **382**(6667) (2023) eadf7044

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

31. Narita, T., Higashijima, Y., Kilic, S., Liebner, T., Walter, J., Choudhary, C.: Acetylation of histone h2b marks active enhancers and predicts cbp/p300 target genes. Nature Genetics **55**(4) (2023) 679–692

32. Bakken, T.E., Jorstad, N.L., Hu, Q., Lake, B.B., Tian, W., Kalmbach, B.E., Crow, M., Hodge, R.D., Krienen, F.M., Sorensen, S.A., et al.: Comparative cellular analysis of motor cortex in human, marmoset and mouse. Nature **598**(7879) (2021) 111–119

33. Li, W., Lu, J., Lu, P., Gao, Y., Bai, Y., Chen, K., Su, X., Li, M., Liu, J., Chen, Y., et al.: scnanohi-c: a single-cell long-read concatemer sequencing method to reveal high-order chromatin structures within individual cells. Nature Methods (2023) 1–13

34. Zhang, L., Sun, C., Jin, Y., Gao, K., Shi, X., Qiu, W., Ma, C., Zhang, L.: Dickkopf 3 (dkk3) improves amyloid-β pathology, cognitive dysfunction, and cerebral glucose metabolism in a transgenic mouse model of alzheimer's disease. Journal of Alzheimer's Disease **60**(2) (2017) 733–746

35. Hass, J., Walton, E., Kirsten, H., Turner, J., Wolthusen, R., Roessner, V., Sponheim, S.R., Holt, D., Gollub, R., Calhoun, V.D., Ehrlich, S.: Complexin2 modulates working memory-related neural activity in patients with schizophrenia. European Archives of Psychiatry and Clinical Neuroscience **265**(2) (March 2015) 137–145

36. Zhou, J., Ma, J., Chen, Y., Cheng, C., Bao, B., Peng, J., Sejnowski, T.J., Dixon, J.R., Ecker, J.R.: Robust single-cell Hi-C clustering by convolution-and random-walk–based imputation

37. Kim, H.J., Yardımcı, G.G., Bonora, G., Ramani, V., Liu, J., Qiu, R., Lee, C., Hesson, J., Ware, C.B., Shendure, J., et al.: Capturing cell type-specific chromatin compartment patterns by applying topic modeling to single-cell Hi-C data. PLoS computational biology **16**(9) (2020) e1008173

38. Zhang, R., Zhou, T., Ma, J.: Multiscale and integrative single-cell Hi-C analysis with Higashi. Nature biotechnology **40**(2) (2022) 254–261

39. Zheng, Y., Shen, S., Keleş, S.: Normalization and De-noising of Single-cell Hi-C Data with BandNorm and scVI-3D. To appear in *Genome Biology*. https://www.biorxiv.org/content/10.1101/2021.03.10.434870v1 (2022)

40. Park, K., Keleş, S.: Joint tensor modeling of single cell 3d genome and epigenetic data with muscle. Journal of the American Statistical Association (just-accepted) (2024) 1–20

41. Li, X., Feng, F., Pu, H., Leung, W.Y., Liu, J.: scHiCTools: A computational toolbox for analyzing single-cell Hi-C data. PLoS computational biology **17**(5) (2021) e1008978

42. Li, X., Zeng, G., Li, A., Zhang, Z.: DeTOKI identifies and characterizes the dynamics of chromatin TAD-like domains in a single cell. Genome biology **22**(1) (2021) 1–26

43. Ramani, V.and Deng, X., Qiu, R., Gunderson, K., Steemers, F., Disteche, C., Noble, W., Duan, Z., Shendure, J.: Massively multiplex single-cell Hi-C. Nature Methods **14** (01 2017)

44. Ay, F., Bailey, T.L., Noble, W.S.: Statistical confidence estimation for hi-c data reveals regulatory chromatin contacts. Genome research **24**(6) (2014) 999–1011

45. Stansfield, J.C., Cresswell, K.G., Dozmorov, M.G.: multihiccompare: joint normalization and comparative analysis of complex hi-c experiments. Bioinformatics **35**(17) (2019) 2916–2923

46. Baker, S.G.: The multinomial-poisson transformation. Journal of the Royal Statistical Society: Series D (The Statistician) **43**(4) (1994) 495–504

47. Song, C., Tseng, G.C.: Hypothesis setting and order statistic for robust genomic meta-analysis. The annals of applied statistics **8**(2) (2014) 777

48. Rao, S.S., Huntley, M.H., Durand, N.C., Stamenova, E.K., Bochkov, I.D., Robinson, J.T., Sanborn, A.L., Machol, I., Omer, A.D., Lander, E.S., et al.: A 3d map of the human genome at kilobase resolution reveals principles of chromatin looping. Cell **159**(7) (2014) 1665–1680

49. McDaniell, R., Lee, B.K., Song, L., Liu, Z., Boyle, A.P., Erdos, M.R., Scott, L.J., Morken, M.A., Kucera, K.S., Battenhouse, A., et al.: Heritable individual-specific and allele-specific chromatin signatures in humans. Science **328**(5975) (2010) 235–239

50. Hao, Y., Hao, S., Andersen-Nissen, E., Mauck III, W.M., Zheng, S., Butler, A., Lee, M.J., Wilk, A.J., Darby, C., Zager, M., et al.: Integrated analysis of multimodal single-cell data. Cell (2021)

51. Bennett, D.A., Buchman, A.S., Boyle, P.A., Barnes, L.L., Wilson, R.S., Schneider, J.A.: Religious orders study and rush memory and aging project. Journal of Alzheimer's disease **64**(s1) (2018) S161–S189

52. Flores, N.M., Podpolny, M., McLeod, F., Workman, I., Crawford, K., Ivanov, D., Leonenko, G., Escott-Price, V., Salinas, P.C.: Downregulation of dickkopf-3, a wnt antagonist elevated in alzheimer's disease, restores synapse integrity and memory in a disease mouse model. Elife **12** (2024) RP89453

53. Mostafa, G.A., Shehab, A.A., Al-Ayadhi, L.Y.: The link between some alleles on human leukocyte antigen system and autism in children. Journal of neuroimmunology **255**(1-2) (2013) 70–74

54. Durand, N.C., Robinson, J.T., Shamim, M.S., Machol, I., Mesirov, J.P., Lander, E.S., Aiden, E.L.: Juicebox provides a visualization system for hi-c contact maps with unlimited zoom. Cell systems **3**(1) (2016) 99–101

55. Kent, W.J., Sugnet, C.W., Furey, T.S., Roskin, K.M., Pringle, T.H., Zahler, A.M., Haussler, D.: The human genome browser at UCSC. Genome research **12**(6) (2002) 996–1006

Methods

MINTsC model representation

MINTsC, presented in **Fig.** 1, is an empirical Bayes model for inferring multi-way interactions from scHi-C contact matrices. It leverages pairwise interactions quantified in contact matrices to construct a test statistic for each candidate multi-way interaction, i.e., clique of bins/genomic-intervals, and evaluates these against a generative null distribution for contact matrices. Here, the candidate multi-way interactions are the cliques whose nodes (loci) are simultaneously observed in at least certain number of cells in scHi-C contact matrices. This step is designed to rule out false cliques that can be detected by merely aggregating pairwise interactions observed across different groups of cells. Specifically, a clique finding algorithm was deployed on each individual cell's contact matrix. Then, the union of set of these cliques were filtered to retain the cliques that were observed in at least c cells across all the cell types. Different c values (c_0 and c_1) were employed for 3-way cliques and higher order cliques (with size larger than 4), due to the overwhelming number of size 3 cliques. Consequently, for each cell type, the cliques that were not fully observed (with all of its pairwise interactions) at least once in any of the cells within that cell type were filtered. We defer further details on the choice of these parameters to the Supplementary Note.

For the evaluation of the pre-defined candidate multi-way interactions, we specifically consider cell k, $k \in [K]$, and its symmetric (*cis*-interaction) contact matrices (maps) with dimensions $l_{chr} \times l_{chr}$ for each chromosome $chr \in [Chr]$, where l_{chr} denotes the number of loci for chromosome chr and the bracket notation $[K]$ denotes a set of sequential numbers $\{1, \cdots, K\}$ (**Fig.** 1). Each entry (edge) of a contact matrix (graph) quantifies the level of contact (edge weight) between two genomic loci (nodes), e.g., $v_1 \in [l_{chr}], v_2 \in [l_{chr}]$. None-zero pairwise contacts among more than two loci, e.g., a size 3 clique of loci v_1, v_2, v_3 (starred entries in **Fig.** 1c), provide evidence for a potential multi-way interaction among the loci v_1, v_2, v_3 (triangles in **Fig.** 1c) within a single cell. Under a noise free setting, all pairwise interactions stemming from a multi-way interaction would be observed and generate a clique. However, owing to sparsity levels of contact matrices (different edge widths in each triangle in **Fig.** 1c) and potential proximity ligation limitations, all pairwise interactions may not be observable within an individual cell. To address this issue for detecting multi-way interactions from contact matrices, we view each contact matrix of a cell as an independent sample of the true unknown contact matrix of a context, e.g., cell type, pseudo-time, or treatment. This inherently requires that the cells, to which the method is applied, should be homogeneous enough (Supplementary Note on the homogeneity assumption). Therefore, a user needs to first start with identifying a group of relatively homogeneous cells and apply MINTsC for each group. Then, we compute a clique level score (z-score and p-score) for each candidate clique $q \in [Q]$, by summarizing the pairwise interaction levels of the clique throughout the cells (box with order statistics in **Fig.** 1b). This score aggregates the evidence for individual cliques across the cells of the same type by leveraging all pairwise interactions.

To assess the significance of the cliques, we evaluate whether the clique scores are higher than expected under a contact matrix generating model which considers counts in each upper triangular entry of the symmetric contact matrix of a cell k as a random draw from a multinomial distribution with the sequencing depth C_k of the cell as the number of trials. In order to take into account the well-known genomic distance bias, i.e., band bias indicating that contact counts between two loci systematically vary with the genomic distance between them, observed in both the bulk and single cell Hi-C data [39,44,45], we model the parameters of the multinomial distribution with a natural cubic spline (**Fig.** 1b). Specifically, we define the set of locus pairs within each off-diagonal of the contact matrix as a "band" (gray and gold bars in **Fig.** 1c), where the locus pairs $i \in [N_j]$ in the same band $j \in [J]$ have the same genomic distance d_j.

Then, the contact counts y_{ijk} of locus pairs $i \in [N_j]$ of bands $j \in [J]$ of cell $k \in [K]$ follow a multinomial random vector $Y_k := (\boldsymbol{y}_{1k}^T, \cdots, \boldsymbol{y}_{Jk}^T)^T$ with total count C_k, where $\boldsymbol{y}_{jk} = (y_{ijk}, \cdots, y_{N_j jk})^T$. Without loss of generality, we dropped the chromosome subscript $chr \in [Chr]$ and cell type subscript $t \in [T]$ for multinomial sample Y_k; however, this model applies for each chromosome and for each cell type, separately. Additionally, while the contact counts from different cells can be assumed to be independent, dependency among contact counts of a clique within a single cell is plausible. This aspect is naturally incorporated by the multinomial model.

In addition to the multinomial likelihood that serves to estimate the band-level parameters $p(d_j)$ (Band level Poisson Spline in **Fig.** 1a), we further impose a Dirichlet prior distribution (Gray arrow in **Fig.** 1a) on the locus-pair level contact probabilities $\tilde{p}_{ij}$. Hyper-parameters α_{ij} of this Dirichlet prior are utilized to distribute the band-level contact probability p_j onto locus pairs within the band while taking into account the overall null hypothesis (Locus pair level null allocation in **Fig.** 1a). This statistical model for the scHi-C contact counts, along with the null distributional assumption introduced in the next section, enables constructing clique level test statistics, including clique z-score and clique p-score, which enable false discovery rate (FDR) control (**Fig.** 1b, c).

Details on statistical framework and estimation of MINTsC

MINTsC relies on a Dirichlet-multinomial distribution that is fitted for each chromosome and cell type separately. In the presentation below, without loss of generality, we drop the chromosome and cell type index. In the MINTsC model, for each locus pair $i \in [N_j]$ within band $j \in [J]$ and a cell $k \in K$, pairwise contact levels y_{ijk} follow a multinomial distribution

$$\begin{bmatrix} [\boldsymbol{y}_{1k}] \\ \vdots \\ [\boldsymbol{y}_{Jk}] \end{bmatrix} \overset{indep}{\sim} \text{Mult}\left(C_k, \begin{bmatrix} [\boldsymbol{p}_1] \\ \vdots \\ [\boldsymbol{p}_J] \end{bmatrix} \right), \tag{1}$$

where $\boldsymbol{y}_{jk} := (y_{1jk}, \cdots, y_{N_j jk})^T$, $\boldsymbol{p}_j := (p_{1j}, \cdots, p_{N_j j})^T$, and the number of trials C_k is the sequencing depth of cell k. The bracket $[\cdot]$ for each $\boldsymbol{y}_{jk}$ (or $\boldsymbol{p}_j$) emphasizes that this component is vector-valued for each band $j \in [J]$. We further parametrize the contact probabilities p_{ij} as $p_{ij} = \tilde{p}_{ij} p_j$, where for $\tilde{\boldsymbol{p}}_j = (\tilde{p}_{1j}, \cdots, \tilde{p}_{N_j j})^T$, $\tilde{\boldsymbol{p}}_j \overset{indep}{\sim} \text{Dir}\left((\alpha_{1j}, \cdots, \alpha_{N_j j})\right)$. This formulation

includes a band-level probability p_j and within band allocation weights $\tilde{p}_{ij}$. Since $\sum_{i=1}^{N_j} \tilde{p}_{ij} = 1$, by the aggregation property of multinomial distribution, we have the following band-level model

$$\begin{bmatrix} S_{1k} \\ \vdots \\ S_{Jk} \end{bmatrix} \overset{indep}{\sim} \text{Mult}\left(C_k, \begin{bmatrix} p_1 \\ \vdots \\ p_J \end{bmatrix} \right),$$

where $S_{jk} := \sum_{i=1}^{N_j} y_{ijk}$ and $p_j = \sum_{i=1}^{N_j} p_{ij}$ for $\forall j \in [J], k \in [K]$. Moreover, by the Poisson transformation of the multinomial distribution [46], we equivalently have

$$S_{jk} \overset{indep}{\sim} \text{Pois}(C_k p_j). \tag{2}$$

In order to incorporate the well-established band bias (or genomic distance bias) of chromatin conformation capture assays, we further impose a spline model on the band-level contact probability as $p_j = p(d_j)$, where d_j is the genomic distance between the loci in a locus pair in band j. A similar distance dependence was adapted for loop identification in bulk Hi-C data [44]. Here, p is a natural-cubic spline function with L number of basis functions and is set to $L = 20, 20, 30, 30, 100$ for Ramani *et al.*, Li *et al.*, Nagano *et al.*, Tan *et al.* and Lee *et al.* data, respectively, in the applications presented.

MINTsC first fits the model in **Eqn.** (2), i.e., obtains $\hat{p}(\cdot)$, by a Poisson generalized linear model on the total band counts S_{jk} (Band level Poisson Spline in **Fig.** 1a). This procedure reduces computational burden significantly (e.g., from a spline GLM fit with the number of experimental units as $\left\{ \binom{l_{chr}}{2} \text{ locus pairs } \forall chr \in [Chr] \right\}$ to only $\{\text{number of bands}\}$, with as many observations as the number of cells for the cell type for each experimental unit. Then, within each band j, MINTsC distributes the estimated overall band contact probability $\hat{p}_j = \hat{p}(d_j)$ across all the locus pairs (Locus pair level null allocation in **Fig.** 1a) under the following null assumption:

$$H_0 : \alpha_{ij} \underset{(\leq)}{=} \begin{cases} \alpha_j, \text{ if } i \in A_j \\ \epsilon, \text{ if } i \notin A_j, \end{cases} \tag{3}$$

where $\epsilon > 0$ is a small positive number, e.g., $\epsilon = 1e-05$. Here, A_j is the set of band j locus pairs that are expected to have nonzero contact counts, and, similarly, A_j^C is the set of locus pairs that are more likely to have zero contacts. This null assumption results in expected contact counts as $\mathbb{E}[y_{ijk}] = C_k p(d_j) \frac{\alpha_j \mathbf{1}(i \in A_j) + \epsilon \mathbf{1}(i \notin A_j)}{N_{A_j} \alpha_j + N_{A_j^C} \epsilon} \approx \frac{C_k p(d_j)}{N_{A_j}} I(i \in A_j)$, where N_{A_j} denotes the cardinality of the set A_j. This intuitively indicates that the locus pairs in A_j will share the non-zero uniform expected contact counts $C_k p(d_j)$, and the ones outside of A_j will have expected count as zero. The set A_j can be obtained from external data including combination of mappability scores and GC content. However, given that we are in a setting where we are considering multiple cell types, these sets can be obtained by evaluating the contact counts of the loci across all the cells excluding the cells from the cell type that is being analyzed (Supplementary Note for details on the construction of A_j set). This exclusion is to ensure that this set, which informs the null assumption, is estimated from independent data. A simpler choice for α_{ij} would be equal distribution of the contact probabilities across all the locus pairs within

a band: $1/N_j$. However, our experiments indicated that this choice of prior null underestimates the mean of the clique statistics and leads to uncalibrated p-values as quantified by significant violations from standard normal in the resulting Q-Q plots. We discuss how A_j can be set by leveraging cells of the other cell types in the next section.

Given the credibly non-zero set A_j, we can estimate α_{ij} with an empirical Bayes approach by maximizing the marginal log likelihood $l(\boldsymbol{\alpha}; Y) = \log\left\{\prod_{k\in[K]} f(Y_k;\boldsymbol{\alpha})\right\} = \log\left[\prod_{k\in[K]}\left\{\int f(Y_k|\boldsymbol{p})f(\boldsymbol{p}|\boldsymbol{\alpha})d\boldsymbol{p}\right\}\right]$ as:

$$\max_{\boldsymbol{\alpha}} l(\boldsymbol{\alpha}; Y) \propto \max_{\boldsymbol{\alpha}} \sum_{k\in[K]} \left\{ \sum_{j\in[J]} \left(\log\Gamma(N_{A_j}\alpha_j + N_{A_j^C}\epsilon) + \sum_{i\in A_j} \log\Gamma(y_{ijk}+\alpha_j) - N_{A_j}\log(\alpha_j) \right) \right. \tag{4}$$

$$\left. - \log\Gamma\left(C_k + \sum_{j\in[J]} \{N_{A_j}\alpha_j + N_{A_j^C}\epsilon\}) \right) \right\}, \tag{5}$$

under the null hypothesis **Eqn.** (3) on $\boldsymbol{\alpha}$, where N_{A_j} (or $N_{A_j^C}$) denotes the cardinality of the A_j (or A_j^C) set, $Y = (Y_1, \cdots, Y_K)$, $Y_k = (\boldsymbol{y}_{1k}, \cdots, \boldsymbol{y}_{Jk})$, $\boldsymbol{\alpha} = (\boldsymbol{\alpha}_1, \cdots, \boldsymbol{\alpha}_J)$, $\boldsymbol{\alpha}_j = (\alpha_{1j}, \cdots, \alpha_{N_j j})$, and $\boldsymbol{p} = (\boldsymbol{p}_1, \cdots, \boldsymbol{p}_J)$. Then, using the properties of conditional expectation and covariance, under the null model **Eqn.** (3), we have locus pair level null parameters

$$\begin{aligned}
\mathbb{E}[y_{ijk}] &= \mathbb{E}[\mathbb{E}[y_{ijk}|\tilde{p}_{ij}]] = \mathbb{E}[C_k\tilde{p}_{ij}p_j] = C_k\frac{\alpha_{ij0}}{\alpha_{j0}}p_j, \\
\mathbb{V}[y_{ijk}] &= C_k p_j \frac{\alpha_{ij0}}{\alpha_{j0}}\left[1 - p_j\frac{1+\alpha_{ij0}}{1+\alpha_{j0}} + C_k p_j \frac{1}{1+\alpha_{j0}}\left(1 - \frac{\alpha_{ij0}}{\alpha_{j0}}\right)\right],
\end{aligned}$$

where $\alpha_{ij0} := \alpha_j \mathbf{1}(i \in A_j) + \epsilon\mathbf{1}(i \notin A_j)$, $\alpha_{j0} := N_{A_j}\alpha_j + N_{A_j^C}\epsilon$. For each locus pair i (within band j) within a size N clique q for $m \in [M]$, where $M = \binom{N}{2}$, we define

$$Z_{i,q} = \frac{\sum_{k=1}^K y_{ijk} - \hat{\mathbb{E}}[\sum_{k=1}^K y_{ijk}]}{\sqrt{\hat{\mathbb{V}}[\sum_{k=1}^K y_{ijk}]}}, \tag{6}$$

where both the expectation and the variance terms are estimated under the null assumption on the locus pair. As the random variable $y_{i_m j_m k}$ is independently distributed across cells $k \in [K]$, by the Lindeberg-feller Central Limit Theorem and the $\sqrt{K}-$consistency of the Maximum Likelihood Estimator (MLE), under some regularity conditions required for MLE, we achieve

$$Z_{i,q} \xrightarrow{d} N(0,1),$$

as the number of cells $K \to \infty$ under the null distribution assumption **Eqn.** (3).

Finally, based on the set of locus pair level statistics of a (size N) clique q, we introduce clique level summary statistics. First, considering the vector of locus pair level statistics within a clique q, $\boldsymbol{Z}_q = (Z_{1,q}, Z_{2,q}, \cdots, Z_{M,q})$ with $M := \binom{N}{2}$, we define the rth largest value $Z_{(M-r+1)q}$ as the *clique z-score* of the clique q. Equivalently, for $\boldsymbol{p}_q = (p_{1,q}, p_{2,q}, \cdots, p_{M,q})$, where $p_{i,q} := 1 - \Phi(Z_{i,q})$ denotes the right-tail standard normal cdf value of $Z_{i,q}$, we let the rth smallest value $p_{(r),q}$ as the *clique p-score* of the clique q. Hence, the clique level summary

statistics are defined as follows for clique q:

$$\text{Clique p-score} := p_{(r)q}, \tag{7}$$

$$\text{Clique z-score} := Z_{(M-r+1)q}. \tag{8}$$

The key reasoning behind choosing rth strongest locus pair level statistics of a clique is that we require a clique to have at least r number of significant pairwise edges to represent a multi-way interaction. A small enough $p_{(r)q}$ (or large enough $Z_{(r)q}$) automatically guarantees that r number of locus pairs will have significant interactions to form a multi-way interaction. Under this definition, we have tractable null distribution for the clique p-scores. Theoretically, for independent p_{iq}s following uniform distribution $U(0,1)$, we have $p_{(r),q} \sim \text{Beta}(r, M-r+1)$. We note that the entries of $\boldsymbol{Z}_q$, and therefore those of $\boldsymbol{p}_q$, are not necessarily independent of each other due to the multinomial assumption. However, empirically, the covariance between the entries are sufficiently small (with ratio approximately $1/100$) compared to the variance terms across all the datasets we analyzed. Therefore, we approximate the null distribution of the clique p-score as

$$p_{(r),q} \sim \text{Beta}(r, M-r+1).$$

Analytically tractable p-values of the p-scores facilitate false discovery rate control in the downstream analysis. Clique-level p-values are computed from the parametric distribution $\text{Beta}(r, M-r+1)$, i.e.,

$$\text{Clique p-value} := \mathbb{P}(X \leq p_{(r),q}), \tag{9}$$

where $X \sim \text{Beta}(r, M-r+1)$. Then, the Benjamini-Hochberg procedure is employed with these p-values for FDR control across the candidate cliques.

The most conservative choice for r is M, where the maximum of the locus pair level p-values of a clique is set as the clique p-score. Empirically, we (and others in a different context for combining p-values [47]) observed that when the observed group of p-values, e.g., a clique's locus pair level p-values in this case, exhibit deviations from the assumed theoretical null due to data noise, maximum p-value or a too large choice of r is not robust, i.e., its empirical null distribution deviates from the theoretical null distribution, especially for large M [47]. Moreover, in practice, this leads to detecting too few cliques. Another choice of r, $r = \binom{N-1}{2} + 1$, ensures that all the N loci participate in forming a connected graph, when $p_{(r)q}$ is small enough. However, our empirical results show that this choice is not well calibrated when compared to $\text{Beta}(r, M-r+1)$. Hence, for the datasets analyzed in this paper, we set $r = N-1$ for a size N clique and filter the cliques from further testing if the r strongest locus pairs involve fewer than N loci in order to make sure the significant pairwise edges at least form a connected graph. The latter condition ensures that all the loci within the clique are involved in at least one pairwise interaction. The order statistic summarization to construct a clique score naturally makes the multi-way interactions to involve a sufficient number of strong pairwise interactions, while the non significant cliques would involve a relatively smaller number of significant pairwise interactions.

Details of synthetic data generation

We now introduce details on synthetic data generation for the analyses made in Section 2.2 and 2.6.

For both of the simulations, we leveraged Lee *et al.* [22] scHi-C dataset (chromosome 19 with 5,582 bins of 10Kb, and L2/3 neurons) and the model in **Eqn.** (1), where the multinomial contact probabilities $\hat{\boldsymbol{p}}_j$, $j \in [J]$ were initially estimated under the null assumption **Eqn.** (3). Then, in the null contact matrices learned from Lee *et al.* [22] scHi-C data, we implanted clique structures for the simulation-based analyses.

For the analysis made in Section 2.2, in order to showcase how MINTsC empirically handles spurious cliques that are mere aggregation of pairwise locus pairs appearing in disjoint set of cells, two different sets of clique structures were implanted, which resulted in the population level contact probability matrix as a mixture of two contact matrices. Specifically, both matrices had the same 300 three way cliques (with edges A_i, B_i, C_i, $i = 1, \cdots, 300$) which constituted the true three-way interactions. For the other 300 three-way cliques (referred to as spurious multi-way interactions) with edges (D_i, E_i, F_i, $i = 1, \cdots, 300$), one of the matrices had the D_i and E_i edges, whereas the other had the F_i edge to ensure that two of the three edges were observed in half of the cells, while the third edge was observed in the other half. This mimics the setting where aggregated pairwise data might spuriously suggest a three-way interaction (**Fig.** S5a). Then, cells were divided into two groups, and contact counts for each cell were generated by sampling from one of the two contact matrices, with the first half of the cells sampled from the first matrix and the remaining half from the second. In both cases, contact counts were sampled from a multinomial distribution with probabilities given by the corresponding contact matrix. Across 551 simulated cells in one simulation replicate, each true three-way interaction was observed in an average of 7.4 (median of 7) cells and each spurious three-way interaction was observed in an average of 1.7 (median of 0) cells. To provide practical validity of the simulated true cliques, we analyzed a CRISPRi validated multi-way interaction involving the MYC gene in K562 cells[1]. We quantified how many cells in the K562 scHi-C dataset contained the complete clique corresponding to this multi-way interaction. Out of 135 cells, only 7 cells (5%) harbored the full set of pairwise contacts, which roughly mimics the observation we have for the true cliques. All true multi-way interactions were observed as cliques in at least one cell, while only 11% of the spurious multi-way interactions were observed as cliques in at least one cell. These patterns are consistent with our empirical findings from multiple scHi-C datasets, where the number of cells containing cliques tends to be limited due to sparsity of the data. Specifically, in any given scHi-C dataset, only about 2–5% of the cliques are observed in more than 10 cells.

With the generated data, we assessed MINTsC's ability to distinguish between these two groups across varying levels of interaction signal (i.e., contact probabilities in the population level contact matrices) and ten replicated simulations. With moderate signal strength (**Fig.** S5b), MINTsC achieved an 83% true positive rate (TPR) and an 8% false positive rate

(FPR). With strong signal, it achieved 100% TPR and 6% FPR, consistent with the expected behavior under a 5% FDR threshold (**Fig.** S5c).

For the baseline comparisons made in Section 2.6, to ensure a more heterogeneous background contact count distribution than accounted by MINTsC (i.e., not strictly following the MINTsC background), we considered $\boldsymbol{p}_0 := (\hat{\boldsymbol{p}}_1^T, \cdots, \hat{\boldsymbol{p}}_J^T)^T \in \mathbb{R}^{\sum_{j=1}^J N_j}$, where the multinomial contact probabilities $\hat{\boldsymbol{p}}_j$, $j \in [J]$, of model in **Eqn.** (1) were estimated under the null assumption **Eqn.** (3), and introduced heterogeneity through 2,100 background cliques with pairwise interaction signal size of 0.00001. As the signal contact counts, we considered 300 multi-way interaction cliques with pairwise interaction signal sizes of 0.000025 (Lowest signal), 0.00005 (Low), 0.000075 (Medium), and 0.0001 (High). This construction resulted in a parameter vector $\boldsymbol{p}_a$, each entry of which denoted (population) pairwise interaction value of the implanted background (2,100) and true (300) cliques. Then, we generated the scHi-C contact maps across cells $k \in [K]$ based on

$$\begin{bmatrix} [\boldsymbol{y}_{1k}] \\ \vdots \\ [\boldsymbol{y}_{Jk}] \end{bmatrix} \overset{indep}{\sim} \text{Mult}\left(C_k, \boldsymbol{p}\right), \tag{10}$$

where $\boldsymbol{p} \propto \boldsymbol{p}_0 + \boldsymbol{p}_a$. Specifically, in each simulation replicate, we generated contact maps across 551 cells under each signal setting (Lowest, Low, Medium, High) and kept the number of simulation replications at 10 to ensure computational feasibility of the baseline method (Supplementary Note).

Within each simulation replicate across the four signal settings, we applied MINTsC to infer multi-way interactions (at FDR 0.05) and also fit SnapHi-C, which is designed to exclusively call loops, for the baseline comparison. Significant SnapHi-C loops (at FDR of 0.05) were aggregated into cliques to identify multi-way interactions. Inferred multi-way interactions from both approaches were evaluated for their performance in identifying the true and false multi-way interactions.