MUNIn (Multiple cell-type UNifying long-range chromatin Interaction detector): a statistical framework for identifying long-range chromatin interactions from multiple cell types

Weifang Liu,[1] Armen Abnousi,[2] Qian Zhang,[3] Yun Li,[1,4,5*] Ming Hu,[2*] Yuchen Yang[4,6,7*]

[1] Department of Biostatistics, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA

[2] Department of Quantitative Health Sciences, Lerner Research Institute, Cleveland Clinic Foundation, Cleveland, OH 44195, USA

[3] Department of Statistics, Purdue University, West Lafayette, IN 47907, USA

[4] Department of Genetics, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA

[5] Department of Computer Science, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA

[6] Department of Pathology and Laboratory medicine, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA

[7] McAllister Heart Institute, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA

* Corresponding author: Yun Li

Email: yunli@med.unc.edu

Tel: 1(919)843-2832

Ming Hu

Email: hum@ccf.org

Tel: 1(216)445-5057

Yuchen Yang

Email: yyuchen@email.unc.edu

Tel: 1(984)215-9019

Abstract

Chromatin spatial organization (interactome) plays a critical role in genome function. Deep understanding of chromatin interactome can shed insights into transcriptional regulation mechanisms and human disease pathology. One essential task in the analysis of chromatin interactomic data is to identify long-range chromatin interactions. Existing approaches, such as HiCCUPS, FitHiC/FitHiC2 and FastHiC, are all designed for analyzing individual cell types. None of them accounts for unbalanced sequencing depths and heterogeneity among multiple cell types in a unified statistical framework. To fill in the gap, we have developed a novel statistical framework MUNIn (**M**ultiple cell-type **UN**ifying long-range chromatin **In**teraction detector) for identifying long-range chromatin interactions from multiple cell types. MUNIn adopts a hierarchical hidden Markov random field (H-HMRF) model, in which the status (peak or background) of each interacting chromatin loci pair depends not only on the status of loci pairs in its neighborhood region, but also on the status of the same loci pair in other cell types. To benchmark the performance of MUNIn, we performed comprehensive simulation studies and real data analysis, and showed that MUNIn can achieve much lower false positive rates for detecting cell-type-specific interactions (33.1 - 36.2%), and much enhanced statistical power for detecting shared peaks (up to 74.3%), compared to uni-cell-type analysis. Our data demonstrated that MUNIn is a useful tool for the integrative analysis of interactomic data from multiple cell types.

Introduction

Chromatin spatial organization plays a critical role in genome function associated with many important biological processes including transcription, DNA replication, and development.[1,2] Recently, the ENCODE and the NIH Roadmap Epigenomics project have identified millions of *cis*-regulatory elements (CREs; e.g., enhancers, silencers and insulators) in mammalian genomes. Notably, the majority of genes are not regulated by CREs in one-dimensional (1D) close vicinity. Instead, by forming three-dimensional (3D) long-range chromatin interactions, CREs are able to regulate the expression of genes hundreds of kilobases (kb) away. Deep understanding of chromatin interactome can shed light on gene regulation mechanisms, and reveal functionally causal genes underlying human complex diseases and traits. Comprehensive characterization of chromatin interactome has become an active research area since the development of Hi-C technology in 2009.[3] Later on, Hi-C and other chromatin conformation capture (3C)-derived technologies (e.g., capture Hi-C, ChIA-PET, PLAC-Seq and HiChIP) have been widely used and great strides have been made to link chromatin interactome to mechanisms of transcriptional regulation and complex human diseases, including autoimmune diseases, neuropsychiatric disorders and cancers.[4–7]

Recent studies have shown that interactomes are highly dynamic across tissues, cell types, cell lines, experimental conditions, environmental triggers, and/or biological samples.[8,9] Better characterization of such interactomic dynamics will substantially advance our understanding of transcription regulation across these conditions. To achieve this goal, one could use methods developed for

single cell type (for brevity, we use cell types to denote multiple datasets across tissues, cell types, cell lines, experimental conditions etc). However, such uni-cell-type analysis would fail to borrow information across cell types, thus losing information for shared features, as well as resulting in false positives for cell-type-specific features. Presumably, as shown in eQTL analysis, shared (among at least two cell types) features typically contribute to a considerable proportion and increase with the number of cell types measured.[10] For delineating shared and cell-type-specific features, Bayesian modeling has been shown repeatedly to boast the advantage of adaptively borrowing information such that little power loss incurs for cell-type-specific features while power to detect shared features increases substantially, as demonstrated in many genomic applications including gene expression, GWAS, ChIP-seq, population genetics, and microbiome.[11–15]

In this paper, we focus on the identification of statistically significant long-range chromatin interactions ("peaks" for short) from Hi-C data generated from multiple cell types. The primary goal is the detection of both shared (i.e., shared by more than one cell types) and cell-type-specific peaks. Existing Hi-C peak calling methods, such as HiCCUPS,[16] FitHiC/FitHiC2[17,18] and FastHiC,[19] are all designed for calling peaks from single cell type. None of them is able to account for unbalanced sequencing depths and heterogeneity among multiple cell types in a unified statistical framework. To fill in the methodological gap, we propose MUNIn (**M**ultiple cell-type **un**ifying long-range chromatin **in**teraction detector) for multiple cell type Hi-C peak calling analysis. MUNIn adopts a hierarchical hidden Markov random field (H-HMRF) model, an extension of our previous HMRF peak caller.[20] Specifically, in MUNIn, the status of each interacting chromatin loci pair (peak or background) depends not only on the status of loci pairs in its neighborhood region, but also on the status of the same loci pair in other closely related cell types (Figure 1). In addition, we assume that the signal-noise-ratio (peak vs. background) among different cell types follows a common prior distribution. Compared to uni-cell-type analysis, the H-HMRF approach adopted by MUNIn has the following three key advantages: (1) MUNIn can achieve lower false positive rates for the detection of cell-type-specific peaks. (2) MUNIn can achieve high power for the detection of shared peaks. (3) MUNIn can borrow information across all cell types proportional to the corresponding sequencing depths. We have conducted comprehensive simulation studies and real data analysis to showcase the advantages of MUNIn over other Hi-C peak calling approaches.

Materials and methods

Overview of Statistical Modeling of MUNIn

Let x_{ijk} and e_{ijk} represent the observed and expected chromatin contact frequency spanning between bin i and bin j in cell type k $(1 \leq i < j \leq N, 1 \leq k \leq K)$, respectively, where N is the total number of bins, and K is the total number of cell types. e_{ijk} is pre-calculated by FitHiC.[17] We assume that x_{ijk} follows a negative binomial (NB) distribution with mean μ_{ijk} and over-dispersion ϕ_k:

$$\log(\mu_{ijk}) = \log(e_{ijk}) + I(z_{ijk} = 1)\theta_k. \quad (1)$$

Here $z_{ijk} \in \{-1,1\}$ is the peak indicator for bin pair (i,j), where $z_{ijk} = 1$ indicates (i,j) is a peak in cell type k and $z_{ijk} = -1$ otherwise. θ_k is the signal-to-noise ratio in cell type k. In other word, if (i,j) is an peak in cell type k, x_{ijk} follows the NB distribution $NB(e_{ijk} * \exp\{\theta_k\}, \phi_k)$. If (i,j) is a background (i.e., non-peak) in cell type k, x_{ijk} follows the NB distribution $NB(e_{ijk}, \phi_k)$.

Then, we use a full Bayesian approach for statistical inference, and assign priors for all parameters $(z_{ijk}, \theta_k, \phi_k)$. Specifically, we adopt a hierarchical Ising prior to simultaneously model spatial dependency among z_{ijk}'s within the same cell type (i.e., for z_{ijk}, borrowing information from $z_{i'j'k}$: $\{|i'-i| + |j'-j| = 1\}$), and the dependency across cell type for the same pair (i.e., borrowing information from $z_{ijk'}$ with $k' \in \{1, \dots, k-1, k+1, \dots, K\}$). First of all, to model spatial dependency of peak indicator within cell type k, we assume that:

$$p\left(\{z_{ijk}\}_{1\le i<j\le N} \middle| \psi_k, \gamma_k\right) = C(\gamma_k, \psi_k) * \exp\left\{\gamma_k \sum_{1\le i<j\le N} I(z_{ijk} = 1) + \psi_k \sum_{|i'-i|+|j'-j|=1} z_{ijk} * z_{i'j'k}\right\}, \quad (2)$$

where $\psi_k > 0$ is the inverse temperature parameter modeling the level the spatial dependency in cell type k, and γ_k models the peak proportion in cell type k. $C(\gamma_k, \psi_k)$ is the normalization constant. In addition, we model the heterogeneity of peak status for a given bin pair (i,j) among multiple cell types, where the vector $\boldsymbol{z_{ij\cdot}} \triangleq (z_{ij1}, z_{ij2}, \dots, z_{ijK})$ can take 2^K possible configurations. We model them using a multinomial distribution $Mult(1, \alpha) \triangleq Mult(1, \alpha_{\{-1,-1,\dots,-1\}}, \alpha_{\{1,-1,\dots,-1\}}, \dots, \alpha_{\{1,1,\dots,1\}})$. Here $\alpha_{\{-1,-1,\dots,-1\}}$ is the probability that the (i,j) pair is background in all K cell types. $\alpha_{\{1,-1,\dots,-1\}}$ is the probability that the (i,j) pair is a peak in the first cell type, but background in all the other $K-1$ cell types. $\alpha_{\{1,1,\dots,1\}}$ is the probability that the (i,j) pair is a peak in all K cell types. Let $n_{\boldsymbol{z_{ij\cdot}}}$ represent the frequency of a specific configuration $\alpha_{\boldsymbol{z_{ij\cdot}}}$. The joint distribution is as follows:

$$p\left(\{\boldsymbol{z_{ij\cdot}}\}_{1\le i<j\le N} | \alpha\right) = \prod_{\boldsymbol{z_{ij\cdot}} \in \{-1,1\}^K} \alpha_{\boldsymbol{z_{ij\cdot}}}^{n_{\boldsymbol{z_{ij\cdot}}}}. \quad (3)$$

In this prior distribution, the peak probability of the (i,j) pair in cell type k (z_{ijk}) depends on the status of same (i,j) pair in the other $K-1$ cell types:

$$p\left(z_{ijk} \middle| z_{ij,-k}, \alpha\right) = \frac{\alpha_{\{z_{ij1},\dots,z_{ijk},\dots,z_{ijK}\}}}{\alpha_{\{z_{ij1},\dots,z_{ijk},\dots,z_{ijK}\}} + \alpha_{\{z_{ij1},\dots,-z_{ijk},\dots,z_{ijK}\}}}. \quad (4)$$

From the Bayes formulation, we have the joint posterior distribution as follow:

$$P\left(z_{ijk}, \theta_k, \phi_k, \psi_k, \gamma_k \middle| x_{ijk}, e_{ijk}\right) \propto P\left(x_{ijk} \middle| e_{ijk}, z_{ijk}, \theta_k, \phi_k\right) * P\left(z_{ijk} \middle| \psi_k, \gamma_k, \alpha\right) * P(\theta_k | \mu_\theta, \sigma_\theta^2) * P\left(\phi_k \middle| \mu_\phi, \sigma_\phi^2\right). \quad (5)$$

One key computational challenge is that in the proposed hierarchical Ising prior, the normalization constant involves ψ_k, γ_k and α. Evaluating such normalization constant requires evaluating all $2^{K*N(N-1)/2}$ possible configurations of the peak indicator $\{z_{ijk}\}$, which is computationally prohibitive. To address this challenge, we adopt a pseudo-likelihood approach, using the product of marginal likelihood to approximate the full joint likelihood. We have shown that such approximation leads to gains in both statistical and computational efficiency.[19]

$$p\left(z_{ijk}\middle|\psi_k,\gamma_k,\alpha\right) \propto \prod_{k=1}^{K}\prod_{1\le i<j\le N} p\left(z_{ijk}\middle|z_{-i,-j,k},\psi_k,\gamma_k\right) * p\left(z_{ijk}\middle|z_{ij,-k},\alpha\right).(6)$$

We use the Gibbs sampling algorithm to iteratively update each parameter. Details of statistical inference can be found in **Supplementary Material Section 1**.

Simulation Framework

To evaluate the performance of MUNIn, we first performed simulation studies with three cell types, considering two scenarios: 1) all three cell types have the same sequencing depth, and 2) the sequencing depth in cell type 3 is half of that in cell type 1 and cell type 2. Each simulated cell types consisted of a 100 × 100 contact matrix. To ensure the three cell types are symmetric, we first simulated the peak status for one "hidden" cell type using the Ising prior, where the two parameters ψ_k and γ_k were set to 0.2 and 0, respectively, assuming that peak and background pairs are equally likely. 10,000 Gibbs sampling steps were carried out to update peak status. Let: $p_0 = P(z_{ijk} = 0|z_{ijk'} = 0)$ and $p_1 = P(z_{ijk} = 1|z_{ijk'} = 1)$ denote the level of dependence across cell types. The peak status of the three testing cell types were simulated from the hidden cell type following three different cell-type-dependence levels $p_0 = p_1 = 0.5, 0.8\ and\ 0.9$, where $p_0 = p_1 = 0.5$ indicates the peak status of three cell types are independent, while $p_0 = p_1 = 0.8\ and\ 0.9$ indicate the peak status of three cell types are of median and high correlation. To simulate Hi-C data with equal sequencing depth, we specified expected contact frequency for the bin pair (i, j) to be inversely proportional to the genomic distance between two interacting anchor bins, following the same formula in each cell type k (note the formula does not depend on k):

$$e_{ijk} = \frac{40}{j-i}\ \ (1 < i < j < 100),$$

To simulate Hi-C data with different sequencing depths, we defined the expected count for bin pair (i, j) in cell type 3 as:

$$e_{ij3} = \frac{20}{j-i}\ \ (1 < i < j < 100).$$

Next, we simulated the observed count from a negative binomial distribution:

$$NB\left(e_{ijk}\exp\left\{\frac{\theta_k(Z_{ijk}+1)}{2}\right\},\phi_k\right).$$

Here, the signal-to-noise ratio parameter θ_k and the over-dispersion parameter ϕ_k were set to be 1.5 and 10.0, respectively. The two Ising parameters were inferred according to the maximum likelihood estimation (MLE) (detailed in **Supplementary Section 1**).

Simulations under each scenario were performed 100 times with different random seeds. We then applied both MUNIn and uni-cell-type analysis using single cell type HMRF model (detailed in **Supplementary Section 2**) on simulated data of each scenario. The peak status was identified from the simulated data using both MUNIn and uni-cell-type method, and compared to the ground truth. ROC curve was computed using *pROC* package.[21] Furthermore, the performance of MUNIn was also evaluated according to the overall percentage of error in peak status z_{ijk}, and the power and type I error for four types of peak status (i.e., shared, cell-type1-specific, cell-type2-specific and cell-type3-specific peaks), respectively.

Performance evaluation

To evaluate the performance of MUNIn in real data, we first compared MUNIn to uni-cell-type method to two biological replicates of Hi-C data from human embryonic stem cells [22], where the peak status is expected to be highly similar. For each biological replicate, both MUNIn and uni-cell-type analysis were implemented for peak calling within each topologically associating domain (TAD) of chromosome 1. To measure the consistency between these two replicates, we computed Adjusted Rand Index (ARI)[23] for the peak status within each TAD.

Additionally, we also analyzed Hi-C data from two different cell lines, GM12878 and IMR90 at 10 kb resolution[16], again using both MUNIn and uni-cell-type analysis. Analyses were performed with each TAD in all chromosomes. Since some TAD boundaries are different between GM12878 and IMR90, we first defined the overlapped TAD regions as the shared TADs between two cell types, and only retained the shared TADs spanning at least 200Kb for the downstream analysis. Cell type dependency was inferred for each TAD based on the results of uni-cell-type analysis. Since there is no ground truth for peaks, we selected significant chromatin interactions (p-value < 0.01 and raw interaction frequency > 5) identified by promoter-capture Hi-C (PC-HiC)[9] in GM12878 and IMR90 cells as the working truth. Since significant interactions identified from PC-HiC data are enriched of promoters, we filtered our significant peaks to only remain bin pairs where at least one of two bins overlaps with a promoter.

Results

Simulation results

To evaluate the performance of MUNIn, we first conducted simulation studies with three cell types, considering two scenarios: (1) all three cell types have equal sequencing depth, and (2) the sequencing depth in cell type 3 is half of that in cell type 1 and 2. In both scenarios, MUNIn outperforms uni-cell-type analysis (**Figures 2-3; Figures S1-S2**). In the first scenario, when all three cell types are independent ($p_0 = p_1 = 0.5$), MUNIn achieves comparable results to uni-cell-type analysis, where the medians of the overall error rate (denoted as "%error") in peak identification of MUNIn range from 16.3 to 16.4% and those of uni-cell-type analysis are 17.2 - 17.3% (**Figure 2a**). With increased cell type dependency, MUNIn achieves lower %error than uni-cell-type analysis. When the cell type dependency becomes high, MUNIn reduces %error by approximately 30.3% on top of uni-cell-type results (11.9 - 12.0% for MUNIn and 17.0 - 17.2% for uni-cell-type analysis) (**Figure 2a**). We then assessed the power and type I error for detecting shared and cell-type-specific peaks by MUNIn and uni-cell-type analysis. When three cell types are highly correlated, MUNIn achieves substantial power gain in shared peaks across cell types than uni-cell-type analysis (85.9% vs. 54.1%; **Figure 2c**), at the cost of a slight increase in error rate (20.6% vs. 9.1%; **Figure S1a**). In addition, MUNIn reduces the type I error in calling cell-type-specific peaks by 33.1 - 34.3% on the top of uni-cell-type results (45.5 - 46.3% vs. 69.3 - 69.5%; **Figure 2d**), at the cost of power loss (36.4 -

37.1% vs. 57.3 - 58.5%; **Figure S1b**). The ROC curves showed that MUNIn better detects shared peaks than uni-cell-type analysis (**Figure 2b**), and these two methods performed comparably in cell-type-specific peaks (**Figure S2**).

Furthermore, when three cell types are with different sequencing depths, we observe consistent patterns that MUNIn outperforms uni-cell-type analysis, especially for cell type 3 with shallower sequencing depth (**Figure 3; Figures S3-S4**). Similar to scenario 1, the ROC curves show that MUNIn exhibits better calling in shared peaks (**Figure 3b**). Consistently, MUNIn substantially improves the power in calling shared peaks than uni-cell-type analysis (84.0% vs 48.2% by MUNIn and uni-cell-type analysis, respectively) with a slightly increase of type I error (22.7% vs 11.4%) (**Figure 3c; Figure S3a)**. More importantly, MUNIn achieves 36.2% reduction of %error for cell type 3 with shallower sequencing depth on the top of uni-cell-type analysis results (15.7% vs 24.6%; **Figure 3a**). MUNIn also attains lower type I error in calling cell-type3-specific peaks (51.1% vs 74.4%) with a loss in power (26.7% vs 48.1%) (**Figure 3d; Figure S3b**). These results indicate that MUNIn can accurately identify peaks in the shallowly sequenced cell type by adaptively borrowing information from deeply sequenced cell types.

Real data analysis

To assess the performance of MUNIn in real data, we first applied both MUNIn and uni-cell-type analysis to two biological replicates of Hi-C data from human embryonic stem cells,[22] and then measured the consistency between two replicates by ARI for the peak status within each TAD. Comparatively, the ARI values of MUNIn are significantly higher than those of uni-cell-type analysis (Wilcoxon test, *p*-value < 2.2e-16; **Figure 4**; **Figure S5**). Specifically, the median value of ARI in MUNIn is 0.993, which shows 48.9% improvement over that of uni-cell-type analysis (**Figure S5**). Our results suggest improved consistency between two replicates by MUNIn, compared to uni-cell-type analysis.

We further compared the accuracy of peak calling in GM12878 and IMR90 cell lines between MUNIn and uni-cell-type analysis. In total, 439,412 and 432,394 shared peaks were detected by MUNIn and uni-cell-type analysis, respectively, 376,658 of which were shared by both methods (85.7 and 87.1% of the shared peaks identified MUNIn and uni-cell-type analysis, respectively) (**Figure S6a**). 217,400 and 82,614 GM12878 and IMR90-specific peaks were identified by MUNIn, while 315,849 and 141,708 GM12878 and IMR90-specific peaks were detected by uni-cell-type analysis. Among them, 77.5 and 75.7% of GM12878- and IMR90-specfic peaks called by MUNIn were also identified by uni-cell-type-analysis (**Figure S6b and c**). The ROC curves show that MUNIn obtains more accurate results for both GM12878 and IMR90-specific peaks (**Figure 5a and d)**, while its performance in shared peaks is comparable to uni-cell-type analysis (**Figure S7**). The area under the curve (AUC) for GM12878 and IMR90-specific peaks of MUNIn increases by 3.7 and 4.6%, respectively, on top of uni-cell-type analysis (**Figure 5a and d**). One example of GM12878-specific peak exclusively identified by MUNIn is shown in **Figure 5b.** One bin of this pair is overlapped with the promoter of *ZNF827* gene (transcription start site (TSS) +/- 500 bp), while the other is overlapped with a known typical enhancer in GM12878

cells (**Figure S8**).[24] In addition, *ZNF827* is more highly expressed in EBV-transformed lymphocytes than in fibroblasts (**Figure 5c; Figure S8**) (the data is obtained from the GTEx Portal), which further suggests the potential role of this GM12878-specific peak in cell-type-specific transcriptional regulation gene. Similarly, the MUNIn-exclusively identified peak between bin 95,005,000 and 95,175,000 is specific to IMR90, which is involved in the regulation of *F3* gene (**Figure 5e**). *F3* gene encodes the tissue factor coagulation factor III and it is usually expressed in the fibroblasts surrounding blood vessels. Consistently, we observed a higher expression level of *F3* gene in cultured fibroblasts than in EBV-transformed lymphocytes (**Figure 5f; Figure S9**)

Discussion

In this study, we present MUNIn, a statistical framework to identify long-range chromatin interactions for Hi-C data from multiple tissues, cell lines or cell types. MUNIn is built on our previously developed methods, HMRF peak caller and FastHiC.[19,20] On top of HMRF, MUNIn jointly models multiple cell types and explicitly accounts for the dependency across cell types. It simultaneously accounts for both spatial dependency within each cell type, and dependency across cell types. By adaptively borrowing information in both aspects, MUNIn can enhance the power of detecting shared peaks, and reduce type I error of detecting cell-type-specific peaks.

MUNIn exhibits substantial advantages in calling peaks shared across cell types compared to uni-cell-type analysis (**Figure 2b**), which are more pronounced with the increased level of across cell-type dependency. In addition, with imbalanced sequencing depth among different cell types, uni-cell-type analysis may mis-classify shared peaks as cell-type-specific due to differential power across cell types. Comparatively, MUNIn can more accurately identify shared peaks (**Figure 3b**). Noticeably, MUNIn resulted in reduced false positives when calling cell-type-specific peaks for the cell type with shallower depth (**Figure 3d**). This is because MUNIn can borrow information from cell types with higher sequencing depth based on the level of dependency across cell types, which is also learned from the data. In our real data evaluations, MUNIn also outperformed uni-cell-type analysis. Specifically, for Hi-C data from human embryonic stem cells, MUNIn exhibited significantly higher consistency between the two biological replicates than the uni-cell-type analysis (**Figure 4; Figure S5**). For Hi-C data from GM12878 and IMR90 cell lines, MUNIn more accurately identified cell-line specific peaks, in terms of both sensitivity and specificity (**Figure 5a and d**). In addition, GM12878- and IMR90-specific peaks exclusively identified by MUNIn showed in **Figure 5** may play a potential role in regulating gene *ZNF827* and *F3* respectively, which are differentially expressed between these two cell lines in the expected direction (**Figure 5c and f**).

Taking the advantages of jointly modeling multiple cell types, MUNIN can easily accommodate many more cell types simultaneously. MUNIn shows a high computational efficiency that MUNIn takes ~31 minutes to perform peak calling in a TAD consisting of 200 bins (**Figure S10**). In addition, MUNIn is also able to handle multiple cell types with differential levels of dependency, for example, when cell types forming clusters where cell types within a cluster are more correlated than those across

clusters. The MUNIn framework can be further extended to accommodate time series chromatin conformation data, which will be explored in our future work.

Taken together, our results show the advantages of MUNIn over the uni-cell-type approach when analyzing data from multiple cell types. By adaptively borrowing information both within and across cell types, MUNIn can achieve much improved power in detecting shared peaks, and much reduced type I error in detecting cell-type-specific peaks. MUNIn's ability to reduce false positive cell-type-specific peak calls due to imbalanced sequencing depths across cell types is also appealing. Finally, MUNIn can more effectively identify biologically relevant chromatin interactions with better sensitivity than the uni-cell-type strategy. We anticipate that MUNIn will become a convenient and essential tool in the analysis of multi-cell-type chromatin spatial organization data.

Data and code availability

MUNIn is compiled as a C++ program, and freely available at https://github.com/yycunc/MUNIn and https://yunliweb.its.unc.edu/MUNIn/.

Supplementary Data

Supplementary materials are available online.

Acknowledgements

This research was supported by the National Institute of Health grants [R01 HL129132, U01 DA052713, R01 GM105785, and P50 HD103573].

Conflict of Interest

The authors declare that they have no competing interests.

Web resources

GTEx Portal, https://gtexportal.org/home/

WashU epigenome browser, http://epigenomegateway.wustl.edu/browser/

HUGIN, https://yunliweb.its.unc.edu/hugin/

REFERENCES

1. Yu, M., and Ren, B. (2017). The three-dimensional organization of mammalian genomes. Annu. Rev. Cell Dev. Biol. *33*, 265–289.

2. Li, Y., Hu, M., and Shen, Y. (2018). Gene regulation in the 3D genome. Hum. Mol. Genet. *27*, R228–R233.

3. Lieberman-Aiden, E., Van Berkum, N.L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., Amit, I., Lajoie, B.R., Sabo, P.J., and Dorschner, M.O. (2009). Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science (80-.). *326*, 289–293.

4. Giusti-Rodríguez, P., Lu, L., Yang, Y., Crowley, C.A., Liu, X., Juric, I., Martin, J.S., Abnousi, A., Allred, S.C., and Ancalade, N. (2019). Using three-dimensional regulatory chromatin interactions from adult and fetal cortex to interpret genetic results for psychiatric disorders and cognitive traits. BioRxiv 406330.

5. Zhou, X., Chen, Y., Mok, K.Y., Kwok, T.C.Y., Mok, V.C.T., Guo, Q., Ip, F.C., Chen, Y., Mullapudi, N., and Giusti-Rodríguez, P. (2019). Non-coding variability at the APOE locus contributes to the Alzheimer's risk. Nat. Commun. *10*, 1–16.

6. Song, M., Pebworth, M.-P., Yang, X., Abnousi, A., Fan, C., Wen, J., Rosen, J.D., Choudhary, M.N.K., Cui, X., and Jones, I.R. (2020). Cell-type-specific 3D epigenomes in the developing human cortex. Nature 1–6.

7. Song, M., Yang, X., Ren, X., Maliskova, L., Li, B., Jones, I.R., Wang, C., Jacob, F., Wu, K., and Traglia, M. (2019). Mapping cis-regulatory chromatin contacts in neural cells links neuropsychiatric disorder risk variants to target genes. Nat. Genet. *51*, 1252.

8. Schmitt, A.D., Hu, M., Jung, I., Xu, Z., Qiu, Y., Tan, C.L., Li, Y., Lin, S., Lin, Y., and Barr, C.L. (2016). A compendium of chromatin contact maps reveals spatially active regions in the human genome. Cell Rep. *17*, 2042–2059.

9. Jung, I., Schmitt, A., Diao, Y., Lee, A.J., Liu, T., Yang, D., Tan, C., Eom, J., Chan, M., and Chee, S. (2019). A compendium of promoter-centered long-range chromatin interactions in the human genome. Nat. Genet. 1–8.

10. Flutre, T., Wen, X., Pritchard, J., and Stephens, M. (2013). A statistical framework for joint eQTL analysis in multiple tissues. PLoS Genet *9*, e1003486.

11. Beaumont, M.A., and Rannala, B. (2004). The Bayesian revolution in genetics. Nat. Rev. Genet. *5*, 251–261.

12. Chen, X., Jung, J.-G., Shajahan-Haq, A.N., Clarke, R., Shih, I.-M., Wang, Y., Magnani, L., Wang, T.-L., and Xuan, J. (2016). ChIP-BIT: Bayesian inference of target genes using a novel joint probabilistic model of ChIP-seq profiles. Nucleic Acids Res. *44*, e65–e65.

13. Wang, Q., Chen, R., Cheng, F., Wei, Q., Ji, Y., Yang, H., Zhong, X., Tao, R., Wen, Z., and Sutcliffe, J.S. (2019). A Bayesian framework that integrates multi-omics data and gene networks predicts risk genes from schizophrenia GWAS data. Nat. Neurosci. *22*, 691–699.

14. Wu, J., Gupta, M., Hussein, A.I., and Gerstenfeld, L. (2020). Bayesian modeling of factorial time-course data with applications to a bone aging gene expression study. J. Appl. Stat. 1–25.

15. Grantham, N.S., Guan, Y., Reich, B.J., Borer, E.T., and Gross, K. (2020). Mimix: A bayesian mixed-effects model for microbiome

data from designed experiments. J. Am. Stat. Assoc. *115*, 599–609.

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

17. Ay, F., Bailey, T.L., and Noble, W.S. (2014). Statistical confidence estimation for Hi-C data reveals regulatory chromatin contacts. Genome Res. *24*, 999–1011.

18. Kaul, A., Bhattacharyya, S., and Ay, F. (2020). Identifying statistically significant chromatin contacts from Hi-C data with FitHiC2. Nat. Protoc. *15*, 991–1012.

19. Xu, Z., Zhang, G., Wu, C., Li, Y., and Hu, M. (2016). FastHiC: a fast and accurate algorithm to detect long-range chromosomal interactions from Hi-C data. Bioinformatics *32*, 2692–2695.

20. Xu, Z., Zhang, G., Jin, F., Chen, M., Furey, T.S., Sullivan, P.F., Qin, Z., Hu, M., and Li, Y. (2016). A hidden Markov random field-based Bayesian method for the detection of long-range chromosomal interactions in Hi-C data. Bioinformatics *32*, 650–656.

21. Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.-C., and Müller, M. (2011). pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinformatics *12*, 1–8.

22. Dixon, J.R., Jung, I., Selvaraj, S., Shen, Y., Antosiewicz-Bourget, J.E., Lee, A.Y., Ye, Z., Kim, A., Rajagopal, N., and Xie, W. (2015). Chromatin architecture reorganization during stem cell differentiation. Nature *518*, 331–336.

23. Hubert, L., and Arabie, P. (1985). Comparing partitions. J. Classif. *2*, 193–218.

24. Martin, J.S., Xu, Z., Reiner, A.P., Mohlke, K.L., Sullivan, P., Ren, B., Hu, M., and Li, Y. (2017). HUGIn: Hi-C unifying genomic interrogator. Bioinformatics *33*, 3793–3795.

25. Li, D., Hsu, S., Purushotham, D., Sears, R.L., and Wang, T. (2019). WashU epigenome browser update 2019. Nucleic Acids Res. *47*, W158–W165.

Figures

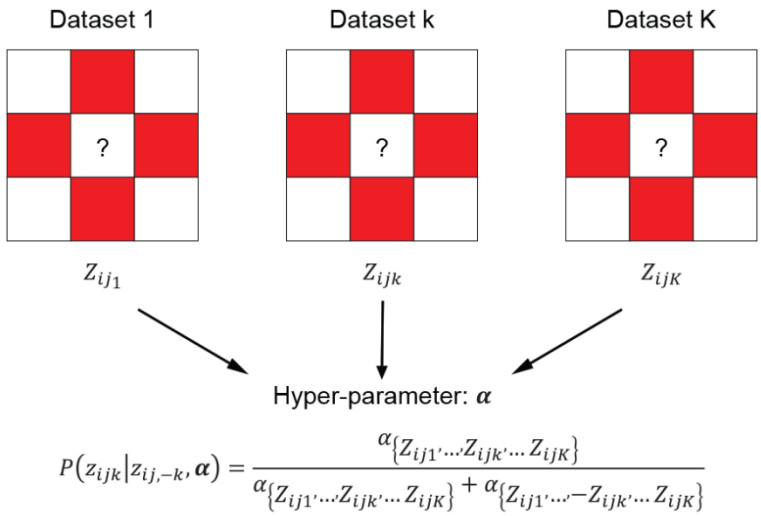

Figure 1. Statistical Schematics of MUNIn.

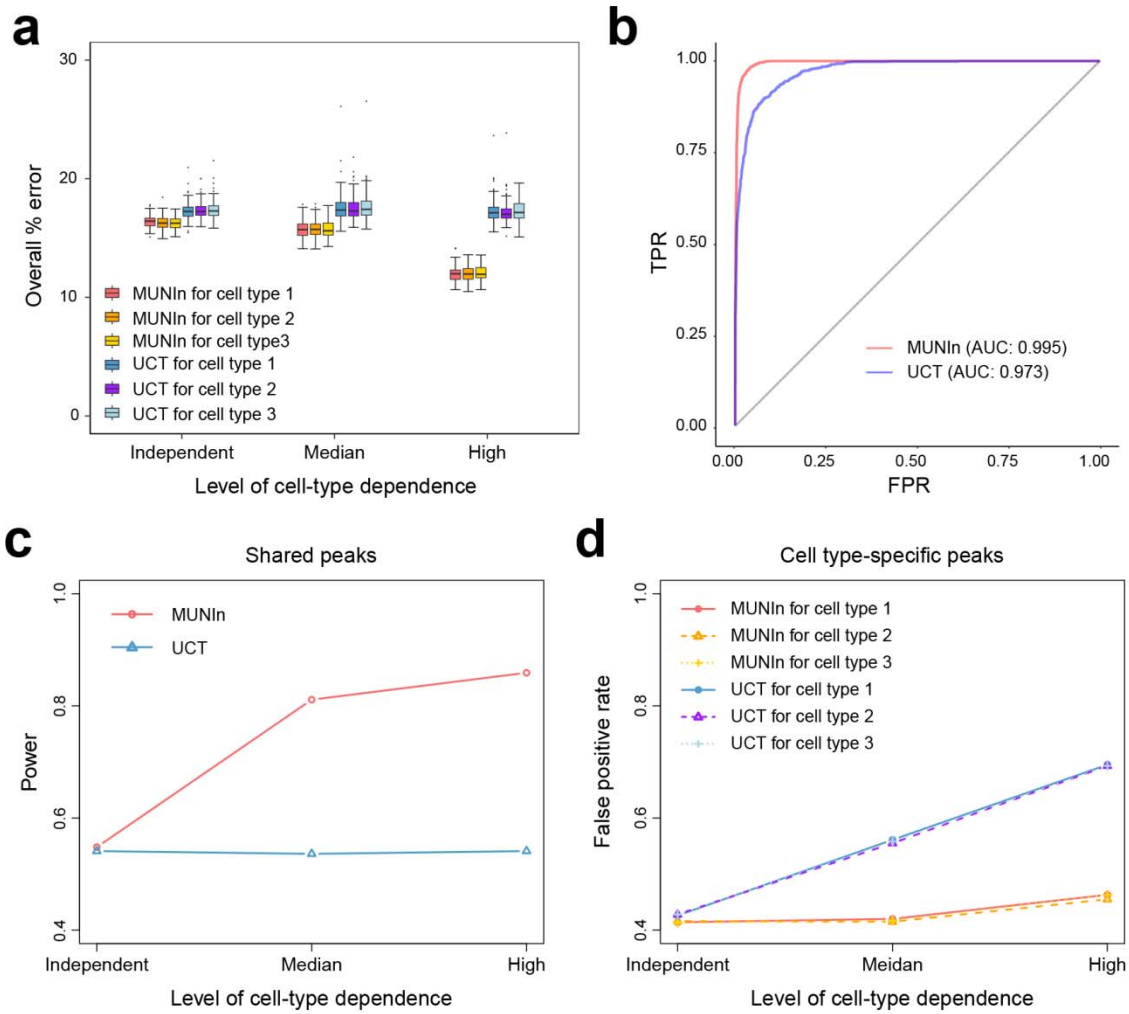

Figure 2. Performance comparison between MUNIn and uni-cell-type analysis (UCT) in the simulation data where all three cell types have equal sequencing depth. (**a**) The overall error rate (denoted as "%error") in peak identification in each cell type using MUNIn and uni-cell-type analysis. (**b**) ROC curves for shared peaks identified by MUNIn and uni-cell-type. (**c**) Power for the shared peaks identified using MUNIn and uni-cell-type (**d**) False positive rate for the cell-type-specific peaks identified in each cell type using MUNIn and uni-cell-type analysis.

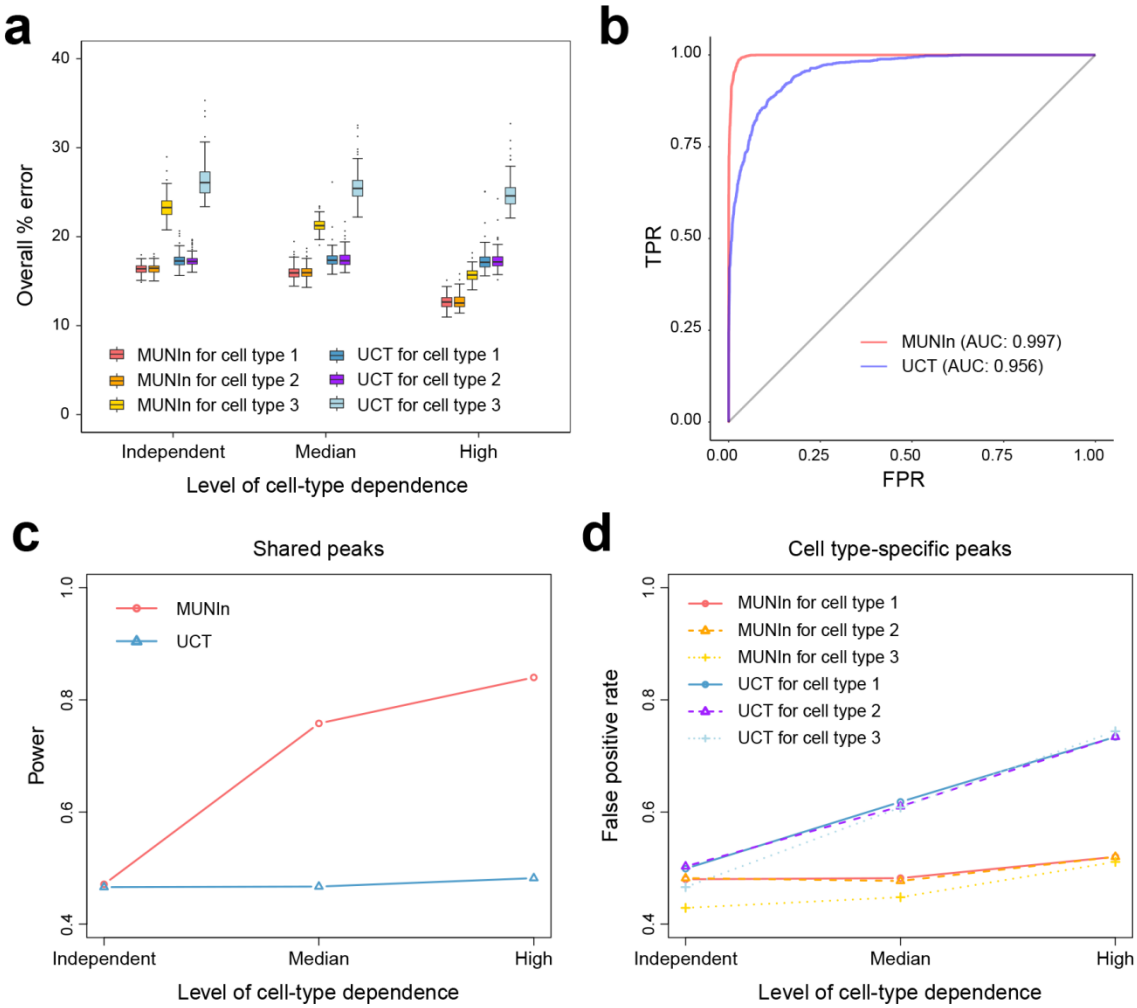

Figure 3. Performance comparison between MUNIn and uni-cell-type analysis (UCT) in the simulation data where the sequencing depth in cell type 3 is half of that in cell type 1 and 2. (**a**) The overall error rate (denoted as "%error") in peak identification in each cell type using MUNIn and uni-cell-type analysis. (**b**) ROC curves for shared peaks identified by MUNIn and uni-cell-type. (**c**) Power for the shared peaks identified using MUNIn and uni-cell-type (**d**) False positive rate for the cell-type-specific peaks identified in each cell type using MUNIn and uni-cell-type analysis.

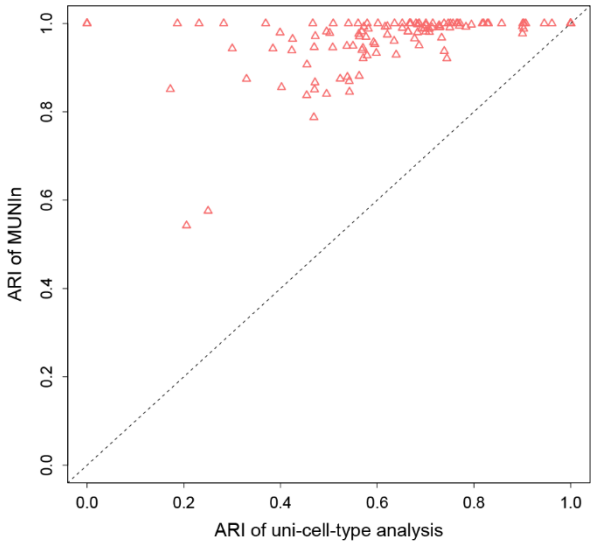

Figure 4. Adjusted Rand Index (ARI) showing the consistency of peak calling by MUNIn and uni-cell-type analysis between the two replicates of human embryonic stem cells.

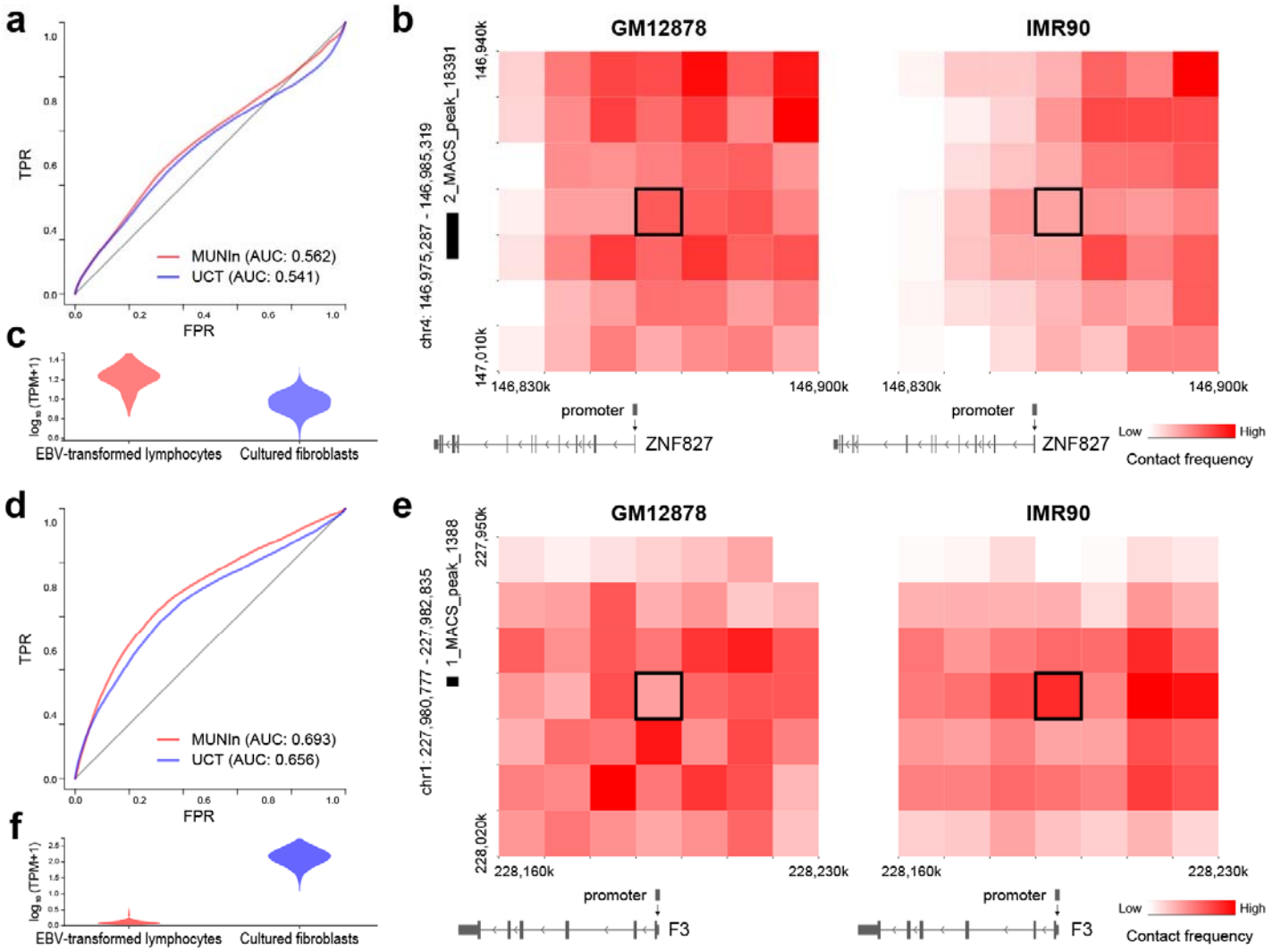

Figure 5. Performance comparison between MUNIn and uni-cell-type analysis (UCT) in the Hi-C data of GM12878 and IMR90 cell lines. (**a**) ROC for GM12878-specfic peaks identified by MUNIn and uni-cell-type. (**b**) Heatmap showing one example of the GM12878-specific peaks in GM12878 (left) and IMR90 (right) Hi-C data. One bin of this pair (highlighted in black) is overlapped with the promoter of *ZNF827* gene (transcription start site (TSS) +/- 500 bp), while the other is overlapped with a known typical enhancer (2_MACS_peak_18391) in GM12878 cells. Gene mode is obtained from WashU epigenome browser.[25] (**c**) Gene expression profiles of *ZNF827* in EBV-transformed lymphocytes and cultured fibroblasts (the data is obtained from the GTEx Portal). (**d**) ROC for IMR90-specfic peaks identified by MUNIn and uni-cell-type. (**e**) Heatmap showing one example of the IMR90-specific peaks in GM12878 (left) and IMR90 (right) Hi-C data. One bin of this pair (highlighted in black) is overlapped with the promoter of *F3* gene, while the other is overlapped with a known typical enhancer (1_MACS_peak_1388) in IMR90 cells. Gene mode is obtained from WashU epigenome browser. (**f**) Gene expression profiles of *F3* in EBV-transformed lymphocytes and cultured fibroblasts (the data is obtained from the GTEx Portal).