HGG Advances

ARTICLE

MUNIn: A statistical framework for identifying long-range chromatin interactions from multiple samples

Weifang Liu,[1,10] Yuchen Yang,[6,8,9,10] Armen Abnousi,[2] Qian Zhang,[3] Naoki Kubo,[4] Joshua S. Martin Beem,[5] Yun Li,[1,6,7,*] and Ming Hu[2,*]

Summary

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 or samples. None of them accounts for unbalanced sequencing depths and heterogeneity among multiple cell types or samples in a unified statistical framework. To fill in the gap, we have developed a novel statistical framework MUNIn (multiple-sample unifying long-range chromatin-interaction detector) for identifying long-range chromatin interactions from multiple samples. 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 samples. 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 sample-specific interactions (33.1%–36.2%), and much enhanced statistical power for detecting shared peaks (up to 74.3%), compared to uni-sample analysis. Our data demonstrated that MUNIn is a useful tool for the integrative analysis of interactomic data from multiple samples.

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 projects 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 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] Since then, 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 samples (for brevity, we use samples to denote multiple datasets across tissues, cell types, cell lines, experimental conditions, etc.). However, such uni-sample analysis would fail to borrow information across samples, thus losing information for shared features as well as resulting in false positives for sample-specific features. Presumably, as shown in expression quantitative trait loci (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 sample-specific features, Bayesian modeling has been shown repeatedly to boast the advantage of adaptively borrowing information,

[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 Cellular and Case Molecular Medicine, University of California San Diego School of Medicine, La Jolla, CA, USA; [5]Duke Human Vaccine Institute, Duke University School of Medicine, Durham, NC 27710, USA; [6]Department of Genetics, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA; [7]Department of Computer Science, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA; [8]Department of Pathology and Laboratory Medicine, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA; [9]McAllister Heart Institute, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA
[10]These authors contributed equally
*Correspondence: yunli@med.unc.edu (Y.L.), hum@ccf.org (M.H.)
https://doi.org/10.1016/j.xhgg.2021.100036.

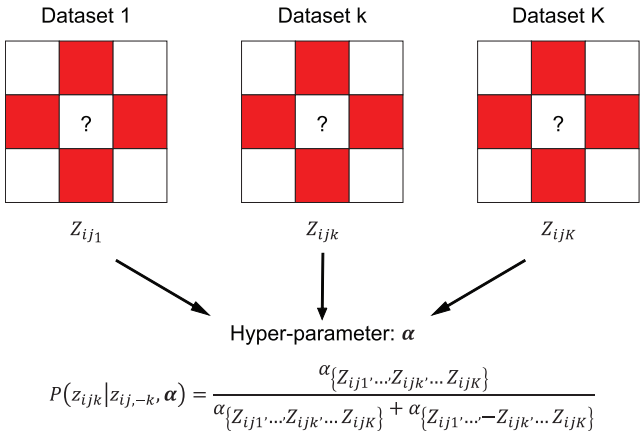

Figure 1. Statistical schematics of MUNIn
In MUNIn, the chromatin interaction status (illustrated with question marks) of each loci pair (i, j) in a sample depends on not only the status of loci pairs in its neighborhood region (red blocks) but also the status of the same loci pair in other samples. Specifically, we model sample dependency by α, where the status of the (i, j)th pair in sample k, z_{ijk} depends on the status of the same (i, j)th pair in the other $K - 1$ samples, given by the formula shown in the figure. Dependency on neighboring loci pairs is captured by the hierarchical Ising prior. See Material and methods and Supplemental section 1 for details.

such that little power loss incurs for sample-specific features, while power to detect shared features increases substantially, as demonstrated in many genomic applications, including gene expression, genome-wide association studies (GWASs), chromatin immunoprecipitation sequencing (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 samples. The primary goal is the detection of both shared (i.e., shared by more than one sample) and sample-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 sample. None of them is able to account for unbalanced sequencing depths and heterogeneity among multiple samples in a unified statistical framework. To fill in the methodological gap, we propose MUNIn (multiple-sample unifying long-range chromatin interaction detector) for multiple-sample 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 samples (Figure 1). Compared to uni-sample 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 sample-specific peaks, (2) MUNIn can achieve high power for the detection of shared peaks, and (3) MUNIn can borrow information across all samples 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.

Material 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 sample $(1 \le i < j \le N,\ 1 \le k \le K)$, respectively, where N is the total number of bins and K is the total number of samples. e_{ijk} is pre-calculated by FitHiC.[17] Briefly, FitHiC used a non-parametric approach to estimate the empirical null distribution of contact frequency (detailed in Supplemental section 1). 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 sample k, and $z_{ijk} = -1$ otherwise. θ_k is the signal-to-noise ratio in sample k. In other words, if (i, j) is a peak in sample 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 sample 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 modeling spatial dependency among z_{ijk}s within the same sample (i.e., for z_{ijk}, borrowing information from $z_{i'j'k} : \{|i' - i| + |j' - j| = 1\}$) and the dependency across samples for the same pair (i.e., borrowing information from $z_{ijk'}$ with $k' \in \{1, \ldots, k-1, k+1, \ldots, K\}$). First of all, to model spatial dependency of peak indicator within sample 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 of the spatial dependency in sample k, γ_k models the peak proportion in sample k, and $C(\gamma_k, \psi_k)$ is the normalization constant. In addition, we model the heterogeneity of peak status for a given bin pair (i, j) among multiple samples, where the vector $\boldsymbol{z_{ij\cdot}} \triangleq (z_{ij1}, z_{ij2}, \ldots, z_{ijK})$ can take 2^K possible configurations. We model them using a multinomial distribution $Mult(1, \alpha) \triangleq Mult(1, \alpha_{\{-1,-1,\ldots,-1\}}, \alpha_{\{1,-1,\ldots,-1\}}, \ldots, \alpha_{\{1,1,\ldots,1\}})$. Here $\alpha_{\{-1,-1,\ldots,-1\}}$ is the probability that the (i, j) pair is background in all K samples, $\alpha_{\{1,-1,\ldots,-1\}}$ is the probability that the (i, j) pair is a peak in the first sample but background in all the other $K - 1$ samples, and similarly $\alpha_{\{1,1,\ldots,1\}}$ is the probability that the (i, j) pair is a peak in all K samples. Let $n_{\boldsymbol{z_{ij\cdot}}}$ represent the frequency of a specific configuration $\alpha_{\boldsymbol{z_{ij\cdot}}}$. The joint distribution is as follows:

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

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

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

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

$$P(z_{ijk},\theta_k,\phi_k,\psi_k,\gamma_k|x_{ijk},e_{ijk}) \propto P(x_{ijk}|e_{ijk},z_{ijk},\theta_k,\phi_k)*P(z_{ijk}|\psi_k,\gamma_k,\alpha) *Prior(\theta_k)*Prior(\phi_k)*Prior(\psi_k)*Prior(\gamma_k). \quad (5)$$

We used uniform prior distributions for θ_k, ϕ_k,ψ_k,γ_k, which were initialized from estimates from uni-sample analysis in our implementation (Supplemental sections 2 and 3). One key computational challenge is that in the proposed hierarchical Ising prior, the normalization constant involving ψ_k, γ_k, and α is computationally prohibitive, since evaluating such a normalization constant requires evaluating all $2^{K*N(N-1)/2}$ possible configurations of the peak indicator $\{z_{ijk}\}$. 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] Let $\{z_{-i,-j,k}\}$ denote the set $\{z_{i'j'k}\big|i'\neq i,\ j'\neq j\}$ and $\{z_{ij,-k}\}$ denote the set $\{z_{ijk'}\big|\ k'\neq k\}$; the posterior probability can be approximated by:

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

We use the Gibbs sampling algorithm to iteratively update each parameter. Details of statistical inference can be found in Supplemental section 2.

Simulation framework

To benchmark the performance of MUNIn, we first performed simulation studies with three samples, where each sample represents a cell type, considering two scenarios: (1) all three samples had the same sequencing depth, and (2) the sequencing depth in sample 3 was half of that in sample 1 and sample 2. Each simulated sample consisted of a 100 × 100 contact matrix. To ensure the three samples were symmetric, we first simulated the peak status for one "hidden" sample using the Ising prior, where the parameter ψ_k was set to 0.2 and γ_k was set to {0, −0.02, −0.05, −0.2, −0.4}, respectively. 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 samples. The peak status of the three testing samples was simulated from the hidden sample following three different sample-dependence levels, $p_0=p_1=0.5$, 0.8, or 0.9, where $p_0=p_1=0.5$ indicates the peak status of three samples are independent, while $p_0=p_1=0.8$ or 0.9 indicates the peak status of three samples is 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 sample 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 sample 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.

Simulations under each scenario were performed 100 times with different random seeds. We then applied both MUNIn and uni-sample analysis using a single-sample HMRF model (detailed in Supplemental section 3) on simulated data of each scenario. The peak status was identified from the simulated data using both MUNIn and uni-sample methods and compared to the ground truth. Receiver operating characteristic (ROC) curve was computed using the *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, sample1-specific, sample2-specific, and sample3-specific peaks), respectively.

Performance evaluation

To evaluate the performance of MUNIn in real data, we first compared MUNIn to uni-sample analysis to two biological replicates of Hi-C data from human embryonic stem cells at 10 kb resolution[22] (Table S1), where the peak status is expected to be highly similar. For each biological replicate, both methods were implemented for peak calling within each topologically associating domain (TAD) of chromosome 1, where TADs were directly obtained from the original paper defined by the insulation score.[22] 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] (Table S1), again using both MUNIn and uni-sample 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 samples and only retained the shared TADs spanning at least 200 kb for the downstream analysis. Sample dependency was inferred for each TAD based on the results of uni-sample 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 (Table S1). Since significant interactions identified from PC-HiC data are enriched of promoters, we filtered our significant peaks to only remaining bin pairs where at least one of two bins overlaps with a promoter. The detailed evaluation framework is in Supplemental section 4. We did additional performance evaluation by running MUNIn by a sliding window approach instead of shared TADs, and we also performed peak calling on samples under different conditions from mouse embryonic stem cells for both wild-type (without

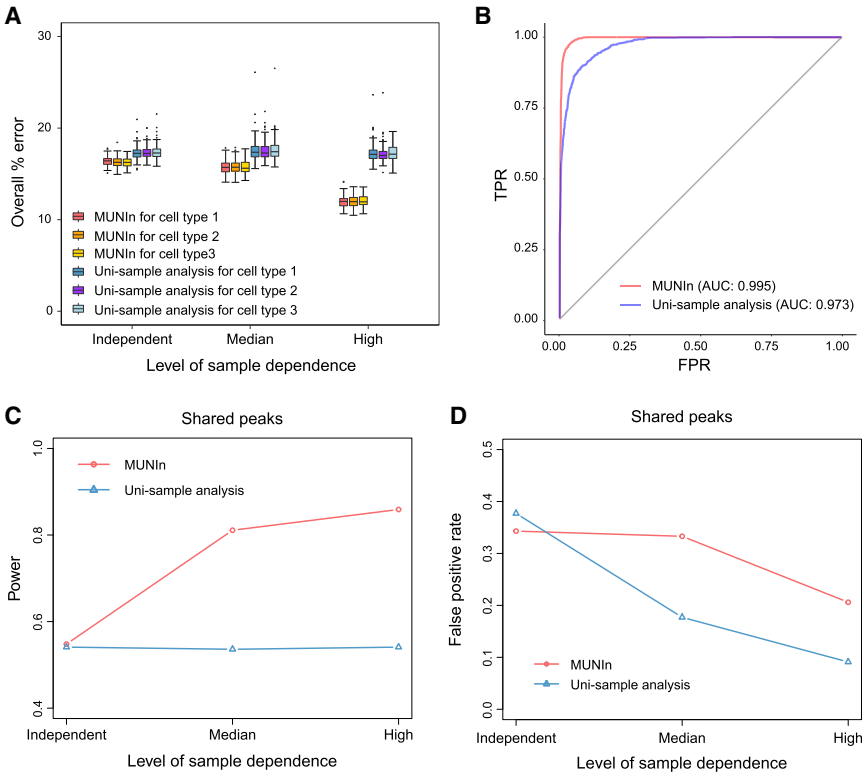

Figure 2. Performance comparison between MUNIn and uni-sample analysis in the simulation data where all three samples have equal sequencing depth

(A) The overall error rate (denoted as "% error") in peak identification in each sample using MUNIn and uni-sample analysis. On each box, the line in the middle is the median across simulations, the lower edge of the box is the 25th percentile, the upper edge of the box is the 75th percentile, the whiskers extend to the smallest and largest values that are not considered outliers, and the outliers are plotted as dots.

(B) ROC curves for shared peaks identified by MUNIn and uni-sample analysis.

(C) Power for the shared peaks identified using MUNIn and uni-sample analysis.

(D) False-positive rate for the shared peaks identified by MUNIn and uni-sample analysis.

CTCF depletion) and after CTCF deletion resolution[24] (Table S1; Supplemental section 5).

Results

Simulation results

To evaluate the performance of MUNIn, we first conducted simulation studies with three samples, considering two scenarios: (1) all three samples have equal sequencing depth, and (2) the sequencing depth in sample 3 is half of that in sample 1 and 2. In both scenarios, MUNIn outperforms uni-sample analysis (Figures 2 and 3; Figures S1–S4). In the first scenario, when all three samples are independent ($p_0 = p_1 = 0.5$), MUNIn achieves comparable results to uni-sample analysis, where the medians of the overall error rate (denoted as "% error") in peak identification of MUNIn range from 16.3%–16.4% and those of uni-sample analysis are 17.2%–17.3% (Figure 2A). With increased sample dependency, MUNIn achieves lower % error than uni-sample analysis. When the sample dependency becomes high, MUNIn reduces % error by approximately 30.3% on top of uni-sample results (11.9%–12.0% for MUNIn and 17.0%–17.2% for uni-sample analysis) (Figure 2A). We then assessed the power and type I error for detecting shared and sample-specific peaks by MUNIn and uni-sample analysis. When three samples are highly correlated, MUNIn achieves substantial power gain in shared peaks across samples compared with uni-sample analysis (85.9% versus 54.1%; Figure 2C), at the cost of a slight increase in error rate (20.6% versus 9.1%; Figure 2D). In addition, MUNIn reduces the type I error in calling sample-specific peaks by 33.1%–34.3% on the top of uni-sample results (45.5%–46.3% versus 69.3%–69.5%; Figure S1A), at the cost of power loss (36.4%–37.1% versus 57.3%–58.5%; Figure S1B). The ROC curves showed that MUNIn better detects shared peaks than uni-sample analysis (Figure 2B), and these two methods performed comparably in sample-specific peaks (Figure S2).

Furthermore, when three samples are with different sequencing depths, we observe consistent patterns that MUNIn outperforms uni-sample analysis, especially for sample 3 with shallower sequencing depth (Figure 3; Figures S3 and 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-sample analysis (84.0% versus 48.2% by MUNIn and uni-sample analysis, respectively) with a slight increase of type I error (22.7% versus 11.4%) (Figures 3C and 3D). More importantly, MUNIn achieves 36.2% reduction of % error for sample 3 with shallower sequencing depth on the top of uni-sample analysis results with high sample dependence (15.7% versus 24.6%; Figure 3A). MUNIn also attains lower type I error in calling sample-3-specific peaks (51.1% versus 74.4%) with a loss in power (26.7% versus 48.1%) (Figures S3A and S3B). These results indicate that MUNIn can accurately identify peaks in the shallowly sequenced sample by adaptively borrowing information from deeply sequenced samples. We further evaluated the robustness and scalability of MUNIn using simulation data where we evaluated results with non-zero γ_ks and increased sample size (Supplemental section 5; Figures S5 and S6).

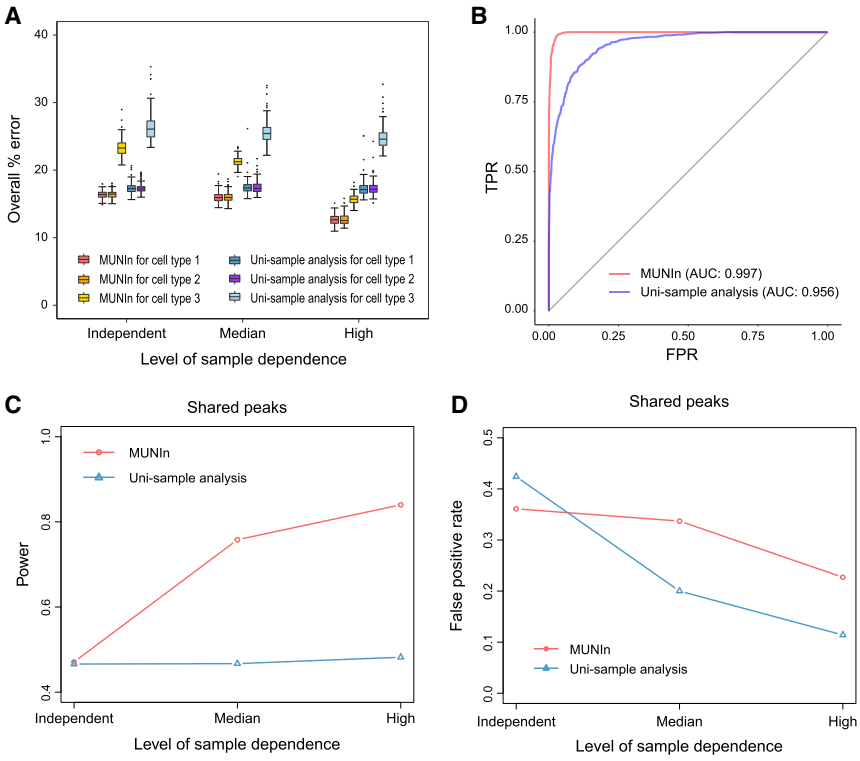

Figure 3. Performance comparison between MUNIn and uni-sample analysis in the simulation data where the sequencing depth in sample 3 is half of that in sample 1 and 2

(A) The overall error rate (denoted as "% error") in peak identification in each sample using MUNIn and uni-sample analysis. On each box, the line in the middle is the median across simulations, the lower edge of the box is the 25th percentile, the upper edge of the box is the 75th percentile, the whiskers extend to the smallest and largest values that are not considered outliers, and the outliers are plotted as dots.

(B) ROC curves for shared peaks identified by MUNIn and uni-sample analysis.

(C) Power for the shared peaks identified using MUNIn and uni-sample analysis.

(D) False-positive rate for the shared peaks identified by MUNIn and uni-sample analysis.

Real data analysis

To assess the performance of MUNIn in real data, we compared the consistency of peak status between two replicates of human embryonic stem cells between MUNIn and uni-sample analysis. Comparatively, the ARI values of MUNIn are significantly higher than those of uni-sample analysis (Wilcoxon test, p value < 2.2e−16; Figure 4; Figure S7). Specifically, the median value of ARI in MUNIn is 0.993, which shows 48.9% improvement over that of uni-sample analysis (Figure S7). Our results suggest improved consistency between two replicates by MUNIn, compared to uni-sample analysis.

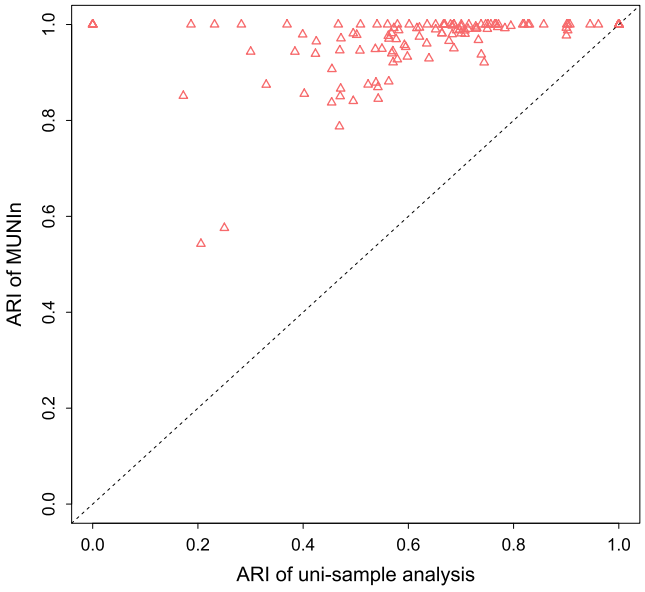

Figure 4. Adjusted Rand index (ARI) showing the consistency of peak calling by MUNIn and uni-sample analysis between the two replicates of human embryonic stem cells

Each triangle represents a TAD. The x and y axes show ARI of uni-sample analysis and MUNIn, respectively.

We further compared the accuracy of peak calling in GM12878 and IMR90 cell lines between MUNIn and uni-sample analysis. In total, 439,412 and 432,394 shared peaks were detected by MUNIn and uni-sample analysis, respectively, 376,658 of which were shared by both methods (85.7% and 87.1% of the shared peaks identified by MUNIn and uni-sample analysis, respectively) (Figure S8A). 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-sample analysis. Among them, 77.5% and 75.7% of GM12878- and IMR90-specfic peaks called by MUNIn were also identified by uni-sample analysis (Figures S8B and S8C). The ROC curves show that MUNIn obtains more accurate results for both GM12878- and IMR90-specific peaks (Figures 5A and 5D), while its performance in shared peaks is comparable to uni-sample analysis (Figure S9). The area under the curve (AUC) for GM12878- and IMR90-specific peaks of MUNIn increases by 3.0% and 4.5%, respectively, on top of uni-sample analysis (Figures 5A and 5D). One example of a GM12878-specific peak exclusively identified by MUNIn is shown in Figure 5B (Figure S10). One bin of this pair is overlapped with the promoter of *ZNF827* (transcription start site [TSS] ± 500 bp), while the other is overlapped with a known typical enhancer in GM12878 cells (Figure S11).[26] In addition, *ZNF827* showed higher gene expression in GM12878 cells than in IMR90 cells (Figure 5C; GTEx Portal), which further suggests the potential role of this GM12878-specific

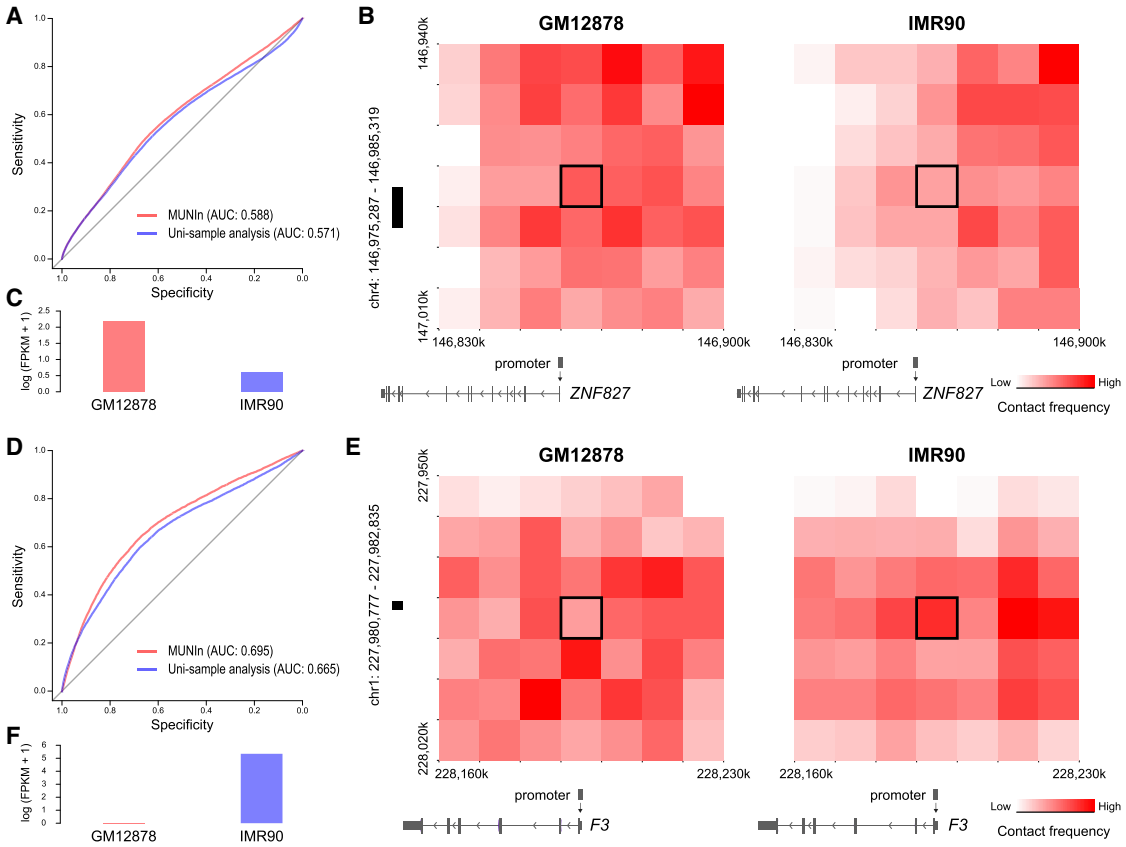

Figure 5. Performance comparison between MUNIn and uni-sample analysis in the Hi-C data of GM12878 and IMR90 cell lines
(A) ROC for GM12878-specfic peaks identified by MUNIn and uni-sample analysis.
(B) Heatmap showing one example of the GM12878-specific peaks in GM12878 (left) and IMR90 (right) Hi-C data. One bin of this pair (highlighted in black) is overlapped with the promoter of *ZNF827* (transcription start site [TSS] ± 500 bp), while the other is overlapped with a known typical enhancer (chr4:146,975,287–146,985,319) in GM12878 cells. Gene model is obtained from WashU epigenome browser.[25]
(C) Gene expression profiles of *ZNF827* in GM12878 and IMR90 cells (GTEx Portal).
(D) ROC for IMR90-specfic peaks identified by MUNIn and uni-sample analysis.
(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*, while the other is overlapped with a known typical enhancer (chr1:227,980,777–227,982,835) in IMR90 cells. Gene model is obtained from WashU epigenome browser.
(F) Gene expression profiles of *F3* in GM12878 and IMR90 cell lines (GTEx Portal).

peak in cell-type-specific transcriptional regulation genes. Similarly, the MUNIn exclusively identified peak between bins chr4:95,000,000–95,010,000 and chr4:95,170,000–95,180,000 is specific to IMR90, which is involved in the regulation of *F3* (Figure 5E; Figure S12). *F3* 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* in IMR90 cells than in GM12878 cells (Figure 5F). Additional real data evaluation also showed the value of borrowing information across samples where we compared MUNIn to uni-sample analysis and FitHiC (Supplemental section 5; Figures S13–S17).

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 samples and explicitly accounts for the dependency across samples. It simultaneously accounts for both spatial dependency within each sample and dependency across samples. By adaptively borrowing information in both aspects, MUNIn can enhance the power of detecting shared peaks and reduce type I error of detecting sample-specific peaks.

MUNIn exhibits substantial advantages in calling peaks shared across samples compared to uni-sample analysis (Figure 2B), which are more pronounced with the increased level of across-sample dependency. In addition, with imbalanced sequencing depth among different samples, uni-sample analysis may mis-classify shared peaks as sample-specific due to differential power across samples. Comparatively, MUNIn can more accurately identify shared peaks (Figure 3B). Noticeably, MUNIn resulted in

reduced false positives when calling sample-specific peaks for the sample with shallower depth (Figure S3A). This is because MUNIn can borrow information from samples with higher sequencing depth based on the level of dependency across samples, which is also learned from the data. In our real data evaluations, MUNIn also outperformed uni-sample analysis. Specifically, for Hi-C data from human embryonic stem cells, MUNIn exhibited significantly higher consistency between the two biological replicates than the uni-sample analysis (Figure 4; Figure S7). 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 (Figures 5A and 5D). In addition, GM12878- and IMR90-specific peaks exclusively identified by MUNIn shown in Figure 5 may play a potential role in regulating *ZNF827* and *F3*, respectively, which are differentially expressed between these two cell lines in the expected direction (Figures 5C and 5F). In our real data analysis, we ran MUNIn in shared TADs across samples instead of the whole chromosomes. We realized that regions outside of TADs or TADs that are not shared across samples may contain sample-specific peaks; therefore, we re-ran the analysis including those regions by a sliding window approach (Figure S13; Supplemental section 5). Our results suggested that including those regions did not have a significant impact on the performance of MUNIn (Figure S13). Additionally, we assessed MUNIn's performance on the Hi-C datasets from mouse embryonic stem cells for both wild-type (without CTCF depletion) and after CTCF deletion at 10 kb resolution[24] (Table S1). The results showed that MUNIn better captured the wild-type-specific pattern in mESC Hi-C data than uni-sample analysis and FitHiC (Figures S14 and S15; Supplemental section 5), demonstrating the power of MUNIn to reveal peaks more powerfully and accurately by borrowing information from another sample.

Taking the advantages of jointly modeling multiple samples, MUNIn can easily accommodate many more samples simultaneously. MUNIn shows a high computational efficiency, in that MUNIn takes ~36 minutes to perform peak calling in a 2 MB TAD of 10 kb resolution (Figures S16 and S17; Supplemental section 5). Moreover, MUNIn is also able to handle multiple samples with differential levels of dependency, for example, when samples form clusters where samples 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. Although MUNIn simultaneously models multiple samples, we note that the goal is to detect chromatin interactions of various peak status configurations across samples, rather than differential interactions. Theoretically, while the posterior probabilities of the peak status configurations can inform differential interactions, it is not our objective here and can be a direction for further exploration.

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

Data and code availability

MUNIn is compiled as a C++ program and is freely available at https://github.com/yycunc/MUNIn and https://yunliweb.its.unc.edu/MUNIn/. All datasets used in this study are publicly available. Accession numbers are included in Table S1.

Supplemental information

Supplemental information can be found online at https://doi.org/10.1016/j.xhgg.2021.100036.

Acknowledgments

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

Declaration of interests

The authors declare no competing interests.

Received: November 12, 2020
Accepted: May 11, 2021

Web resources

WashU epigenome browser, http://epigenomegateway.wustl.edu/browser/
GTEx Portal, https://gtexportal.org/home/
HUGIn, https://yunliweb.its.unc.edu/hugin/
yycunc/MUNIn, https://github.com/yycunc/MUNIn
Li Group Home, https://yunliweb.its.unc.edu/MUNIn/

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* (R2), 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., Dorschner, M.O., et al. (2009). Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science *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., Giusti-Rodríguez, P., et al.; Alzheimer's Disease Neuroimaging Initiative (2019). Non-coding variability at the APOE locus contributes to the Alzheimer's risk. Nat. Commun. *10*, 3310.
6. Song, M., Pebworth, M.-P., Yang, X., Abnousi, A., Fan, C., Wen, J., Rosen, J.D., Choudhary, M.N.K., Cui, X., Jones, I.R., et al. (2020). Cell-type-specific 3D epigenomes in the developing human cortex. Nature *587*, 644–649.
7. Song, M., Yang, X., Ren, X., Maliskova, L., Li, B., Jones, I.R., Wang, C., Jacob, F., Wu, K., Traglia, M., et al. (2019). Mapping cis-regulatory chromatin contacts in neural cells links neuropsychiatric disorder risk variants to target genes. Nat. Genet. *51*, 1252–1262.
8. Schmitt, A.D., Hu, M., Jung, I., Xu, Z., Qiu, Y., Tan, C.L., Li, Y., Lin, S., Lin, Y., Barr, C.L., and Ren, B. (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., Chee, S., et al. (2019). A compendium of promoter-centered long-range chromatin interactions in the human genome. Nat. Genet. *51*, 1442–1449.
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, IeM., 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.
13. Wang, Q., Chen, R., Cheng, F., Wei, Q., Ji, Y., Yang, H., Zhong, X., Tao, R., Wen, Z., Sutcliffe, J.S., et al. (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. Published online June 1, 2020. https://doi.org/10.1080/02664763.2020.1772733.
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., Lander, E.S., and Aiden, E.L. (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*, 77.
22. Dixon, J.R., Jung, I., Selvaraj, S., Shen, Y., Antosiewicz-Bourget, J.E., Lee, A.Y., Ye, Z., Kim, A., Rajagopal, N., Xie, W., et al. (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. Kubo, N., Ishii, H., Xiong, X., Bianco, S., Meitinger, F., Hu, R., Hocker, J.D., Conte, M., Gorkin, D., Yu, M., et al. (2021). Promoter-proximal CTCF binding promotes distal enhancer-dependent gene activation. Nat. Struct. Mol. Biol. *28*, 152–161.
25. Li, D., Hsu, S., Purushotham, D., Sears, R.L., and Wang, T. (2019). WashU epigenome browser update 2019. Nucleic Acids Res. *47* (W1), W158–W165.
26. 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.