DECIPHERING ASSOCIATION SIGNALS FROM GENOME-WIDE ASSOCIATION STUDIES

Weifang Liu

A dissertation submitted to the faculty of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Biostatistics in the Gillings School of Global Public Health.

Chapel Hill
2023

Approved by:

Ming Hu

Yun Li

Michael I. Love

Hudson P. Santos, Jr.

Di Wu

ABSTRACT

Weifang Liu: Deciphering association signals from genome-wide association studies
(Under the direction of Dr. Yun Li)

Over 90% of disease-associated variants detected from Genome-wide Association Studies (GWAS) are in non-coding regions of the genome, making the interpretation of GWAS signals a daunting challenge. Recent advances in genome-wide experimental assays like high-throughput chromosome conformation capture (Hi-C) have greatly improved our understanding of chromatin folding principles and revealed the structural basis of gene regulation and genome function. One essential task in the analysis of chromatin interactome data is the identification of long-range chromatin interactions. However, existing computational tools are all designed for analyzing individual cell types or samples, ignoring unbalanced sequencing depths and heterogeneity among multiple samples. In my first dissertation project, I present MUNIn, a novel statistical framework for identifying long-range chromatin interactions from multiple samples. MUNIn achieves lower false positive rates for sample-specific interactions and enhanced statistical power for shared interactions. Following MUNIn, I will briefly illustrate how I used chromatin conformation data with functional genomics data to identify potential functional genes underlying GWAS signals for cognitive impairment among children born extremely preterm.

In my second dissertation project, I propose SnapHiC-G, a new computational approach based on a global background model to identify cell-type-specific long-range enhancer-promoter interactions from single-cell HiC (scHi-C) data. SnapHiC-G outperforms existing methods designed for both single-cell and bulk Hi-C data with higher sensitivity in identifying long-range enhancer-promoter interactions. SnapHiC-G is a powerful tool for characterizing cell-type-specific enhancer-promoter interactions in single cells from complex tissue samples and facilitating the interpretation of non-coding GWAS variants.

In my third project, I explore the effect of using a more comprehensive LD reference panel for LD score calculation and LD score regression (LDSC) estimates to control for false positives in GWAS results. Results showed that more polygenic signals could be captured by including more variants in the regression, and low-frequency variants exhibited less inflation compared with common variants. Assessing the impact of using a more comprehensive LD reference panel in LD score computation as well as LDSC estimates has important practical implications, and it will guide the choices of the most appropriate LD scores to use and sets of variants to be included in the analysis.

ACKNOWLEDGEMENTS

I feel incredibly fortunate to have met some truly amazing individuals during my life and throughout my Ph.D. journey. I would like to express my heartfelt gratitude to the following individuals:

Firstly, I would like to extend my deepest appreciation to my advisor, Dr. Yun Li, for her support, guidance, patience, and encouragement throughout my time at UNC. I am particularly grateful for the opportunity she provided me to work on many exciting projects. Her willingness to go above and beyond in offering guidance and assistance has been extremely valuable in helping me to develop myself both as a researcher and a professional.

My sincere thanks to the members of my thesis committee. I would like to thank the close mentorship and guidance provided by Dr. Ming Hu, the constant support of Dr. Hudson Santos, and the invaluable feedback and suggestions by Drs. Di Wu and Mike Love. The completion of my Ph.D. would not have been possible without them.

I am deeply grateful to my parents and my boyfriend for their love, encouragement, and support throughout my life. Their support has been my source of strength and happiness.

I would also like to thank my past mentors and advisors. Their support and guidance have been invaluable, and I honestly could not have made it this far without them. Their influence has meant so much to me and has played a critical role in shaping my academic and personal development.

Finally, I would like to express my sincere appreciation to my friends, lab members, and peers. I feel blessed to be surrounded by such exceptional individuals who have made my academic journey a fulfilling and enjoyable experience.

TABLE OF CONTENTS

CHAPTER 1: LITERATURE REVIEW

1.1 High-throughput chromosome conformation capture (Hi-C)

1.1.1 The three-dimensional (3D) genome organization

The human genome consists of approximately 3 billion nucleotides, which can form a $\sim$2-meter-long polymer if stretched in 1D space. However, the average diameter of the nucleus in human cells is only $\sim$6 μm. The five orders of magnitude compaction from 1D space to 3D space results in highly complex chromatin spatial organization. A deep understanding of the principles of chromatin folding holds great promise to reveal the structural basis of gene regulation and genome function [1, 2, 3, 4, 5, 6].

Recent advances in imaging- and sequencing-based technologies have revealed chromatin folding patterns at multiple scales. At the chromosome resolution, different chromosomes occupy distinct locations in the nucleus, termed chromosome territories (CTs) [7]. Transcriptionally active regions are near the nuclear center, while transcriptionally inactive regions are near the nuclear periphery. Zooming in, each chromosome consists of mega-base (Mb) resolution A/B compartments [8] and sub-compartments [9] that are cell-type-specific and correspond to open and closed chromatin [8, 10]. The compartments can be further divided into topologically associating domains (TADs) [11, 12], which are self-interacting domains that are typically several hundred kilobases to $\sim$1 Mb in size, dictating most chromatin interactions to be within the same TAD. Finally, at the kilobase resolution, two types of chromatin loops have been discovered, mostly inside of TADs. One type is structural loops mediated by the convergent CTCF motif pairs and largely conserved among different cell types [9]. The other type is functional loops, which are formed by enhancer-promoter interactions and exhibit high cell-type specificity [13, 14, 15]. Extensive studies have demonstrated that these 3D genome features are closely related to transcriptional regulation mechanisms, the determination of cell identity, and organism-level health and disease outcomes [16, 17, 18].

1.1.2 High-throughput chromosome conformation capture (Hi-C)

Since the advent of Hi-C [8] and Hi-C-derived technologies [19, 20, 21], knowledge of genome-wide chromatin spatial organization has been significantly advanced. Harnessing the power of next-generation sequencing technologies, Hi-C has been widely applied to cultured cell lines, purified cell types, and complex tissues [9, 2, 22], and has revealed the 3D genome features described above. Hi-C is a sequencing-based approach that relies on proximity ligation to quantify pairwise chromatin contact frequency among the cell population under study. The Hi-C protocol involves first crosslinking the chromatin with formaldehyde such that spatially close DNA segments are fixed. Cells are then lysed to allow for the release of nucleic materials. Next, the chromatin is digested by a restriction enzyme that recognizes short DNA sequences (e.g., 4bp or 6bp) and cuts the DNA into fragmented pieces. This step generates sticky ends of DNA fragments which are then filled in with biotins. Those sticky ends are then ligated under dilute conditions where ligation products between cross-linked DNA fragments are enriched. A Hi-C library can be constructed by purifying and shearing DNA and then selecting biotinylated junctions with streptavidin beads. Finally, a catalog of genome-wide pairwise interacting fragments can be obtained by massively parallel DNA sequencing.

Hi-C provides an unbiased view of the entire genome with high throughput and sequence coverage. However, for high-resolution inference, Hi-C approaches require ultra-deep sequencing depth, which can be cost-prohibitive. For example, we usually need several billion raw reads to detect chromatin interactions at Kb resolution. Moreover, Hi-C requires a large amount of input materials with a typical Hi-C bulk sample containing $10^5 \sim 10^6$ cells. Another key weakness of Hi-C is that it does not directly measure the spatial distance between genomic loci of interest but rather gauges the frequency of the loci coming in spatial proximity, which is an indirect measure of the 3D distance among sequencing reads. Therefore, computational tools are needed to infer spatial relationships among genomic regions.

1.1.3 Hi-C Data processing workflow

A typical data processing workflow for Hi-C data involves (i) read alignment, (ii) quality filtering, (iii) binning, and (iv) normalization. After these preprocessing steps, Hi-C data are represented as a contact matrix where each entry represents the number of contacts observed between a pair of genomic regions.

Read alignment Hi-C experiments generate chimeric read pairs lying far apart along the linear genome that need to be aligned on the reference genome. Many Hi-C read alignment algorithms have been proposed, including iterative mapping, split read alignment (BWA MEM), and read clipping [23, 24, 25, 26, 27, 28].

Quality filtering Quality control is performed at both the read level and read-pair level to filter out low-quality alignment. At the read level, standard filters similar to other sequencing-based assays can be applied to the number of mismatches, mapping quality, and uniqueness of mapped reads [29, 30]. Each read is then assigned to the nearest restriction site with a distance that should agree with the molecular size after DNA shearing. At the read-pair level, strand and distance filters can be applied to filter out *de novo* ligations products from the Hi-C protocol and duplicated pairs from polymerase chain reaction (PCR) in the library preparation step.

Binning After filtering, contacts are aggregated into fixed-size genomic bins to construct a contact matrix, where each entry represents the number of contacts observed between a pair of genomic bins. The size of each bin, which is defined by the user, is referred to as the resolution. A smaller bin size is helpful for discovering more refined chromatin features with the cost of more severe sparsity and increased noises, while a larger bin size gives a coarser representation of the 3D genome that might be sufficient for clustering and visualization purposes and detecting features at a large scale.

Normalization Hi-C data are subject to systematic biases including but not limited to sequencing-dependent features, including GC content, mappability, and Hi-C-specific features such as effective fragment length. Several existing normalization methods for Hi-C data [31, 32, 24, 33, 34, 35] can be classified into explicit-factor correction, matrix balancing, and joint correction approaches.

1.1.4 Identifying significant chromatin loops and interactions with Hi-C data

As discussed previously, Hi-C is a sequencing-based method that can capture only relative spatial relationships among sequencing reads instead of exact spatial positions of genomic loci. Therefore, special computational tools are needed to reconstruct 3D conformation from a 2D contact matrix. Recent advances in computational approaches have enabled the identification of chromatin interactions and structures using genome-wide contacts generated by Hi-C. Many computational tools have been developed for investigating chromatin structures from Hi-C contact matrices [31, 33, 6]. This section briefly summarizes existing methods for the detection of chromatin loops and interactions from Hi-C data.

Chromatin loops and interactions are contacts between regions that are far from each other in the 1D genomic distance but close in 3D space [36]. Many methods exist to detect chromatin loops or statistically significant/enriched chromatin interactions. They can be broadly grouped into two classes, namely global background methods and local background methods. Global background methods fit a global statistical model based on 1D genomic distance and assign p-values to each bin pair in the contact matrix by comparing the observed count to the expected under the global background. These methods include HiC-DC [37], HiC-DC+ [38], FitHiC [39], FitHiC2 [40], HMRF [41], FastHiC [41], HiC-ACT [15]. Local background methods identify peaks in the contact map that are local maxima with respect to their neighboring bin pairs. These methods include HiCCUPs [9], cLoops [42], Significant Interaction Peak caller [43], and Mustache [44]. Not surprisingly, global background methods tend to detect a much larger number of chromatin interactions than local background methods, e.g., 10^5-10^6 versus $\sim10^4$ significant chromatin loops reported by the two categories of methods. Since there is no consensus or gold standard method, investigators should choose the appropriate methods tailored to their research questions [36].

1.1.5 From Hi-C to GWAS

GWAS has identified thousands of genetic variants associated with complex human diseases and traits [45]. However, most of these GWAS variants ($>90\%$) reside in non-coding regions producing no proteins, making the interpretation of these variants a daunting challenge [46, 47]. Prior studies observed significant enrichment of non-coding GWAS variants within cis-regulatory elements (CREs, e.g., promoters, enhancers, silencers, and insulators), which play critical roles in disease etiology by regulating the expression of target genes in a cell-type-specific manner [48, 49]. Instead of directly changing the protein-coding DNA sequences, these non-coding variants may disrupt the functional roles of CREs, resulting in the dysregulation of relevant genes.

The comprehensive annotation of CREs is a substantial step forward in understanding non-coding GWAS variants. However, many genes are not regulated merely by CREs in close one-dimensional (1D) vicinity but also by those that form DNA loops with the promoter of their target gene(s) from hundreds of kilobase (Kb) or further away [50, 51]. Advanced genomics technologies like chromosome conformation capture (3C), together with powerful computational methods, have enabled the comprehensive characterization of regulatory DNA interactions and substantially improved our understanding of the three-dimensional (3D) genome architecture. Characterizing 3D chromatin structure has the potential to prioritize disease causal genes, particularly those spatially close but distal in the 1D genomic distance from their CREs, and reveal mechanistic insights underlying non-coding GWAS variants. One of the earliest and most renowned examples was reported by Smemo *et al.* [52], where the authors elegantly elucidated molecular mechanisms underlying the noncoding obesity-associated GWAS variants at the *FTO* locus with chromatin interactions identified from a Hi-C alike technology 4C-seq [53]. Specifically, long-range chromatin interactions link *FTO* intronic variants to their target gene

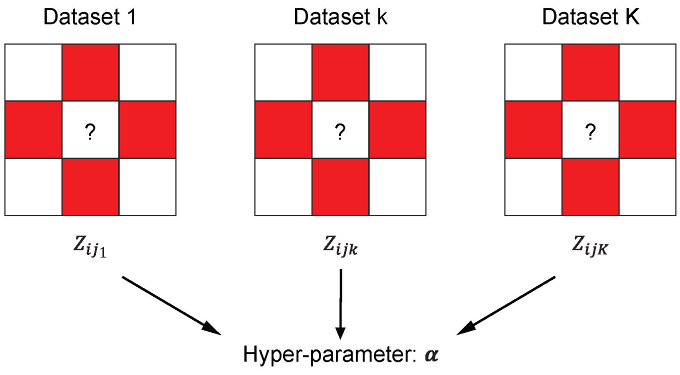
CHAPTER 2: MUNIN (MULTIPLE SAMPLE UNIFYING LONG-RANGE CHROMATIN INTERACTION DETECTOR)

2.1 Introduction

Chromatin spatial organization plays a critical role in genome function associated with many important biological processes, including transcription, DNA replication, and development [54, 55]. 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 (kb) away. A 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 [8]. 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 [56, 57, 58].

Recent studies have shown that interactomes are highly dynamic across tissues, cell types, cell lines, experimental conditions, environmental triggers, and/or biological samples [2]. 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 a single sample (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 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 [59]. For delineating shared and sample-specific features, Bayesian modeling has been shown repeatedly to boast the advantage of adaptively borrowing information such that little power loss incurs for sample-specific features while the power to detect shared features increases substantially, as demonstrated in many genomic applications including gene expression, GWAS, ChIP-seq, population genetics, and microbiome [60, 61, 62].

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 [9], FitHiC/FitHiC2 [39] and FastHiC [41], are all designed for calling peaks from a 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 (**M**ultiple-sample **un**ifying long-range chromatin **in**teraction detector) for multiple samples Hi-C peak calling analysis. MUNIn adopts a hierarchical hidden Markov random field (H-HMRF) model, an extension of our previous HMRF peak caller [41]. 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. (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 advantage of MUNIn over other Hi-C peak calling approaches.



$$P(z_{ijk}|z_{ij,-k},\boldsymbol{\alpha}) = \frac{\alpha_{\{Z_{ij1},\ldots,Z_{ijk},\ldots Z_{ijK}\}}}{\alpha_{\{Z_{ij1},\ldots,Z_{ijk},\ldots Z_{ijK}\}} + \alpha_{\{Z_{ij1},\ldots,-Z_{ijk},\ldots Z_{ijK}\}}}$$

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 z_{ijk}, the status of the (*i*, *j*)th pair in sample *k*, 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 Methods and **Appendix A** for details.

2.2 Materials and methods

2.2.1 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 *k* $(1 \leqslant i < j \leqslant N, 1 \leqslant k \leqslant K)$, respectively, where N is the total number of bins, and K is the total number of samples. e_{ijk} is pre-calculated by FitHiC. Briefly, FitHiC uses a non-parametric approach to estimate the empirical null distribution of contact frequency (**Appendix A**). We assume that x_{ijk} follows a negative binomial (NB) distribution with mean μ_{ijk} and over-dispersion ϕ_k:

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

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

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

where $\psi_k > 0$ is the inverse temperature parameter modeling the level the spatial dependency in sample k, γ_k models the peak proportion in sample k, and $C\left(\gamma_k, \psi_k\right)$ 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 $\mathbf{z_{ij\cdot}} \triangleq (z_{ij1}, z_{ij2}, \ldots, z_{ijK})$ 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)th pair is background in all K samples, $\alpha_{\{1,-1,\ldots,-1\}}$ is the probability that the (i, j)th pair is a peak in the first sample, but a background in all the other $K-1$ samples, and similarly $\alpha_{\{1,1,\ldots,1\}}$ is the probability that the (i, j)th pair is a peak in all K samples. Let $n_{\mathbf{z_{ij\cdot}}}$ represent the frequency of a specific configuration $\alpha_{\mathbf{z_{ij\cdot}}}$. The joint distribution is given by

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

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

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

From the Bayes rule, we have the joint posterior distribution as

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

We use uniform prior distributions for $\theta_k, \phi_k, \psi_k, \gamma_k$, which are initialized from estimates from the uni-sample analysis in our implementation (**Appendix A**). 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 indicators $\{z_{ijk}\}$. 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 in our previous work [41].

Let $\{z_{-i,-j,k}\}$ denote the set $\{z_{i'j'k}|i^{'} \neq i, j^{'} \neq j\}$ and $\{z_{ij,-k}\}$ denote the set $\{z_{ijk'}|k' \neq k\}$, the posterior probability can be approximated by

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

We use the Gibbs sampling algorithm to iteratively update each parameter. Details of statistical inference can be found in **Appendix A**.

2.2.2 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 $\times$ 100 contact matrix. To control the level of sample dependency, we first simulated the peak status for one "hidden" sample using the Ising prior, where 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 the peak status. The level of sample dependency is modeled by $p_0 = P(z_{ijk} = 0|z_{ijk'} = 0)$ and $p_1 = P(z_{ijk} = 1|z_{ijk'} = 1)$. 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 $p_0 = p_1 = 0.8or0.9$ indicate the peak indicators of three samples are of median and high correlation. To simulate Hi-C data with an equal sequencing depth, we specified the 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), where $e_{ijk} = \frac{40}{j-i}(1 < i < j < 100)$, 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 observed counts 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, respectively.

Simulations under each scenario were performed 100 times with different random seeds. We then applied both MUNIn and the uni-sample analysis using the single-sample HMRF model (**Appendix A**) on simulated data of each scenario and compared to the ground truth. Performance was evaluated by ROC curves using the *pROC* package [63], the overall percentage of error in peak status z_{ijk}, power, and type I error for four types of peak status (i.e., shared, sample1-specific, sample2-specific, and sample3-specific peaks), respectively.

2.2.3 Performance evaluation

To evaluate the performance of MUNIn in real data, we first compared MUNIn to the uni-sample analysis of two biological replicates of Hi-C data from human embryonic stem cells at 10kb resolution [64] (**Table A1**), 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 [64]. To measure the consistency between these two replicates, we computed the Adjusted Rand Index (ARI) [65] for the peak status within each TAD.

Additionally, we also analyzed Hi-C data from two different cell lines, GM12878 and IMR90 at 10kb resolution [9] (**Table A1**), 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 overlapping TAD regions as the TADs shared between two two samples and only retained shared TADs spanning at least 200kb for the downstream analysis. Sample dependency was inferred for each TAD based on results from the 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) [66] in GM12878 and IMR90 cell lines as the working truth (**Table A1**). Since significant interactions identified from PC-HiC data are enriched of promoters, we filtered our significant peaks to retain only bin pairs where at least one of the two bins overlaps with a promoter region. Detailed evaluation framework can be found in **Appendix A**. We did additional performance evaluation by running MUNIn by a sliding window approach instead of shared TADs, and also performed peak calling on samples under different conditions from mouse embryonic stem cells for both wild-type (without CTCF depletion) and after CTCF deletion resolution [67] (**Table A1**; **Appendix A**).

2.3 Results

2.3.1 Simulation results

To evaluate the performance of MUNIn, we conduct 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 2-3; Figures A1-A4**). In the first scenario, when all three samples are independent ($p_0 = p_1 = 0.5$), MUNIn achieves comparable results to the uni-sample 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-sample analysis range from 17.2 - 17.3$\%$ (**Figure 2a**). With increased sample dependency, MUNIn achieves a lower $\%$error than uni-sample analysis. With a higher sample dependency, 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 has substantial power gain in shared peaks across samples than the uni-sample analysis (85.9$\%$ vs. 54.1$\%$; **Figure 2c**), at the cost of a mild increase in error rate (20.6$\%$ vs. 9.1$\%$; **Figure 2d**). In addition, MUNIn reduces the type I error in calling sample-specific peaks by 33.1 - 34.3$\%$ on top of uni-sample results (45.5 - 46.3$\%$ vs. 69.3 - 69.5$\%$; **Figure A1a**), at the cost of power loss (36.4 - 37.1$\%$ vs. 57.3 - 58.5$\%$; **Figure A1b**). The ROC curves show that MUNIn better detects shared peaks than uni-sample analysis (**Figure 2b**), and these two methods performed comparably in sample-specific peaks (**Figure A2**).

Furthermore, when three samples have different sequencing depths, we observe consistent patterns that MUNIn outperforms uni-sample analysis, especially for sample 3 with a shallower sequencing depth (**Figure 3; Figures A3-A4**). Similar to scenario 1, the ROC curves show that MUNIn exhibits better performance in shared peaks (**Figure 3b**). Consistently, MUNIn substantially improves the power in calling shared peaks than the uni-sample analysis (84.0$\%$ vs 48.2$\%$

by MUNIn and uni-sample analysis, respectively) with a mild increase of type I error (22.7% vs 11.4%) (**Figure 3c; Figure 3d)**. More importantly, MUNIn achieves 36.2% reduction of %error for sample 3 with shallower sequencing depth on top of the uni-sample analysis results with high sample dependence (15.7% vs 24.6%; **Figure 3a**). MUNIn also attains a lower type I error in calling sample3-specific peaks (51.1% vs 74.4%) with a loss in power (26.7% vs 48.1%) (**Figure A3a; Figure A3b**). These results indicate that MUNIn can accurately identify peaks in the shallower sequenced sample by borrowing information from deeper sequenced samples. We further evaluated the robustness and scalability of MUNIn using simulated data where we evaluated results with non-zero γ_k's and an increased sample size (**Appendix A**; **Figure A5 and A6**).

2.3.2 Real data analysis

To assess the performance of MUNIn in real data, we compare the consistency of peak status between two replicates of human embryonic stem cells between MUNIn and the uni-sample analysis. Comparatively, ARI values of MUNIn are significantly higher than those of the uni-sample analysis (Wilcoxon test, *p*-value $<$ 2.2e-16; **Figure 4**; **Figure A7**). Specifically, the median value of ARI in MUNIn is 0.993, which shows a 48.9% improvement over the uni-sample analysis (**Figure A7**), suggesting an improved consistency between two replicates by MUNIn.

We further compare the accuracy of peak calling in GM12878 and IMR90 cell lines between MUNIn and the uni-sample analysis. In total, 439,412 and 432,394 shared peaks are detected by MUNIn and uni-sample analysis, respectively, where 376,658 of them are shared by both methods (85.7 and 87.1% of the shared peaks identified MUNIn and uni-sample analysis, respectively) (**Figure A8a**). In addition, 217,400 and 82,614 GM12878 and IMR90-specific peaks are identified by MUNIn, while 315,849 and 141,708 GM12878 and IMR90-specific peaks are detected by uni-sample analysis. Among them, 77.5 and 75.7% of GM12878- and IMR90-specific peaks called by MUNIn are also identified by the uni-sample analysis (**Figure A8b and c**). The ROC curves show that MUNIn obtains more accurate results for both GM12878 and IMR90-specific peaks (**Figures 5a and d)**, while its performance in shared peaks is comparable to the

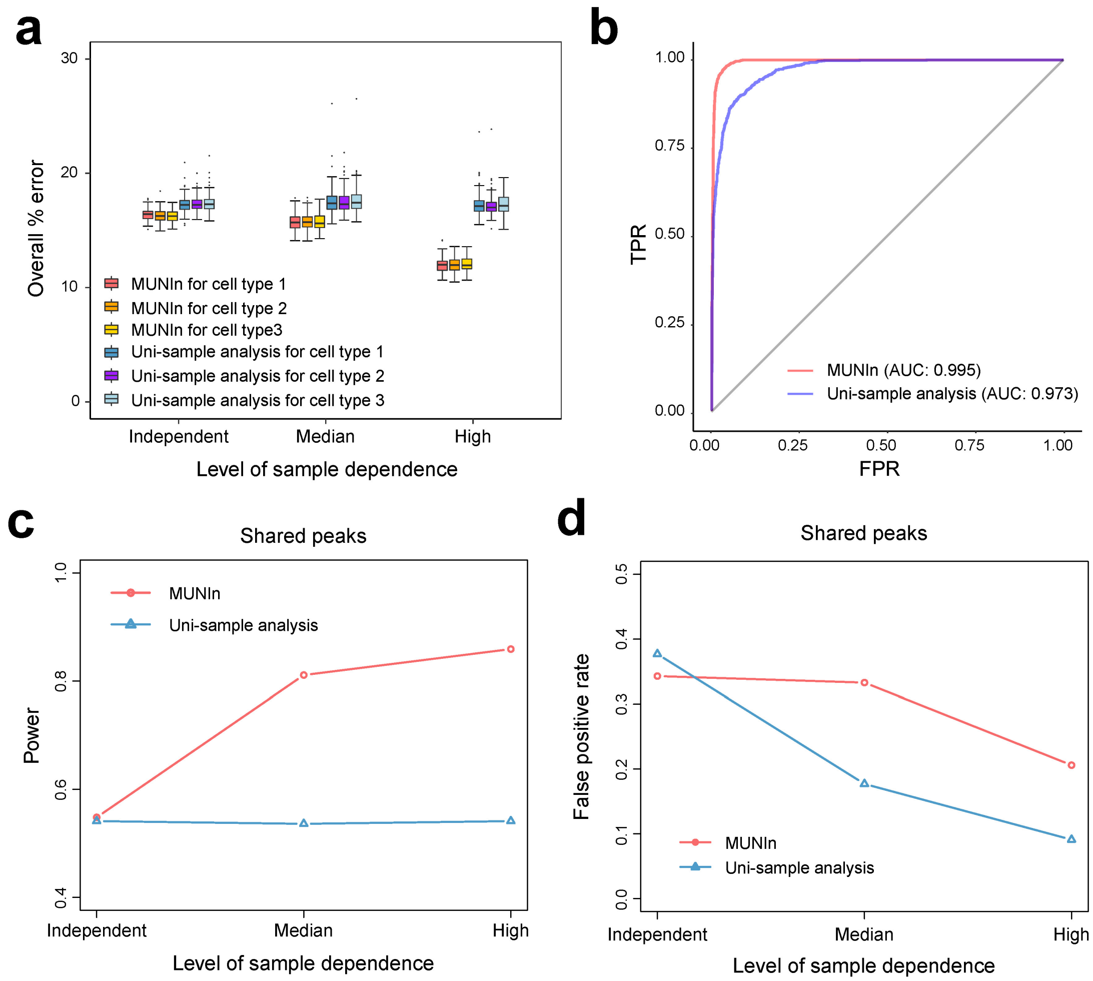


Figure 2. Performance comparison between MUNIn and uni-sample analysis in the simulation data where all three samples have equal sequencing depth. (**a**) The overall error rate (denoted as "%error") in peak identification in each sample using MUNIn and uni-sample analysis. (**b**) ROC curves for shared 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.

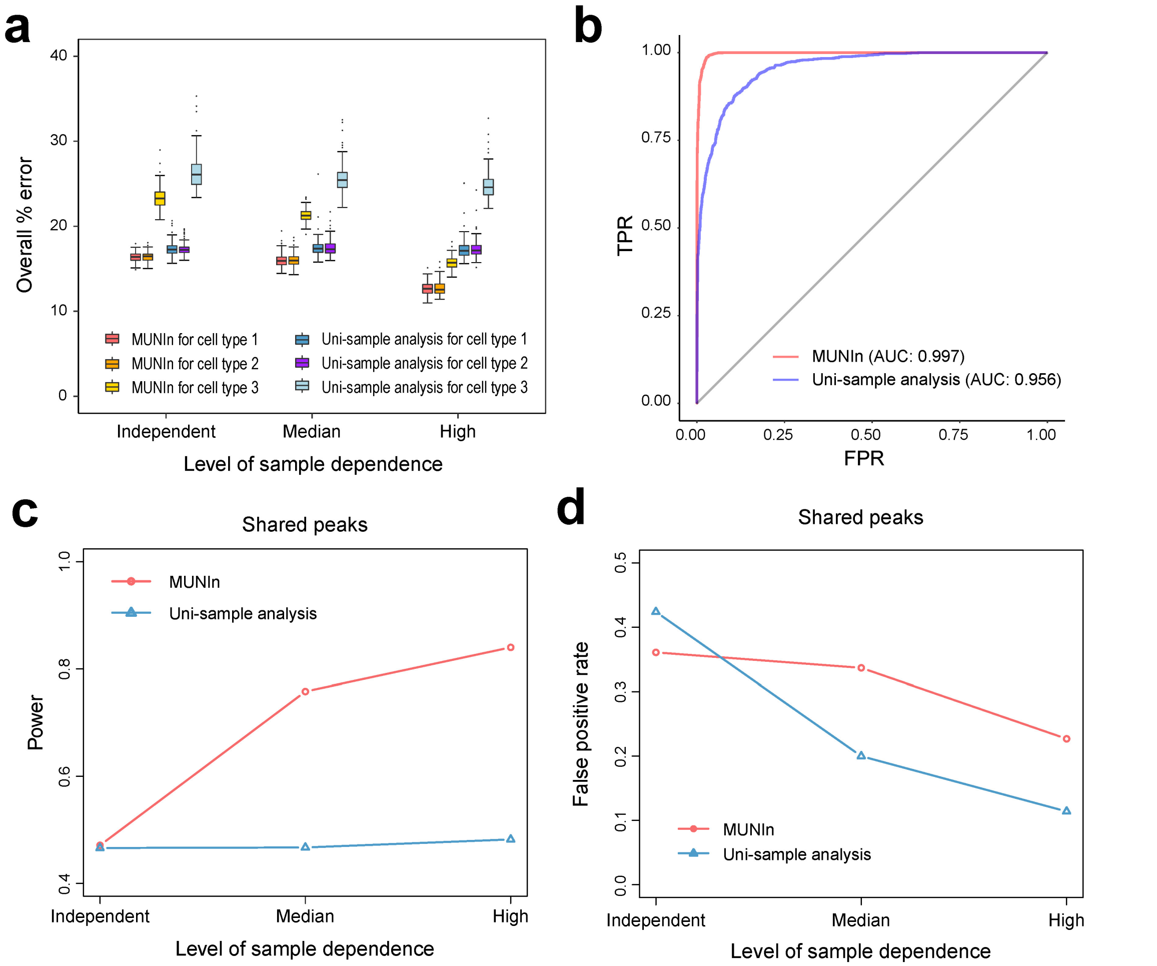


Figure 3. Performance comparison between MUNIn and uni-sample analysis in the simulation data where the sequencing depth in sample 3 is half of that in samples 1 and 2. (**a**) The overall error rate (denoted as "%error") in peak identification in each sample using MUNIn and uni-sample analysis. (**b**) 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 rates for the shared peaks identified by MUNIn and uni-sample analysis.

uni-sample analysis (**Figure A9**). The area under the curve (AUC) for GM12878 and IMR90-specific peaks of MUNIn increases by 3.0% and 4.5%, respectively (**Figure 5a and d**). One example of a GM12878-specific peak exclusively identified by MUNIn is shown in **Figure 5b (Figure A10).** One bin of this pair is overlapped with the promoter of the *ZNF827* gene (transcription start site (TSS) +/- 500 bp), while the other bin is overlapped with a known typical enhancer in GM12878 cells (**Figure A11**) [13]. In addition, *ZNF827* shows a higher gene expression in GM12878 cells than in IMR90 cells (**Figure 5c**), which further suggests the potential role of this GM12878-specific peak in a cell-type-specific transcriptional regulation gene. 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 the *F3* gene (**Figure 5e; Figure A12**). *F3* encodes the tissue factor coagulation factor III and 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 (**Appendix A**; **Figure A13-A17**).

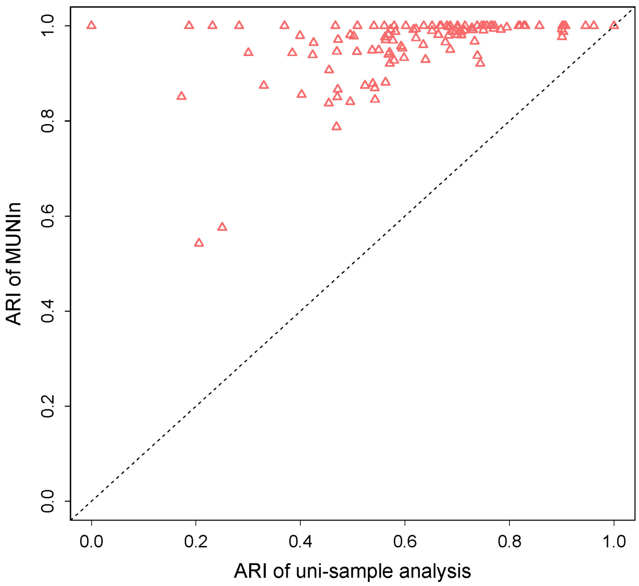


Figure 4. Adjusted Rand Index (ARI). The consistency of peak calling by MUNIn and uni-sample analysis between the two replicates of human embryonic stem cells. Each triangle represents a TAD. The x- and y-axis: ARI of uni-sample analysis and MUNIn, respectively.

2.4 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 extends previously developed methods HMRF peak caller and FastHiC to jointly model multiple samples and explicitly account for the dependency across samples. It simultaneously accounts for both spatial dependency within each sample and dependency across samples. By borrowing information in both aspects, MUNIn can enhance the power of detecting shared peaks, and reduce the type I error of detecting sample-specific peaks.

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 A13**; **Appendix A**). Our results suggested that including those regions did not have a significant impact of the performance of MUNIn (**Figure A13**). Additionally, we assessed MUNIn's performance on the Hi-C datasets from mouse embryonic

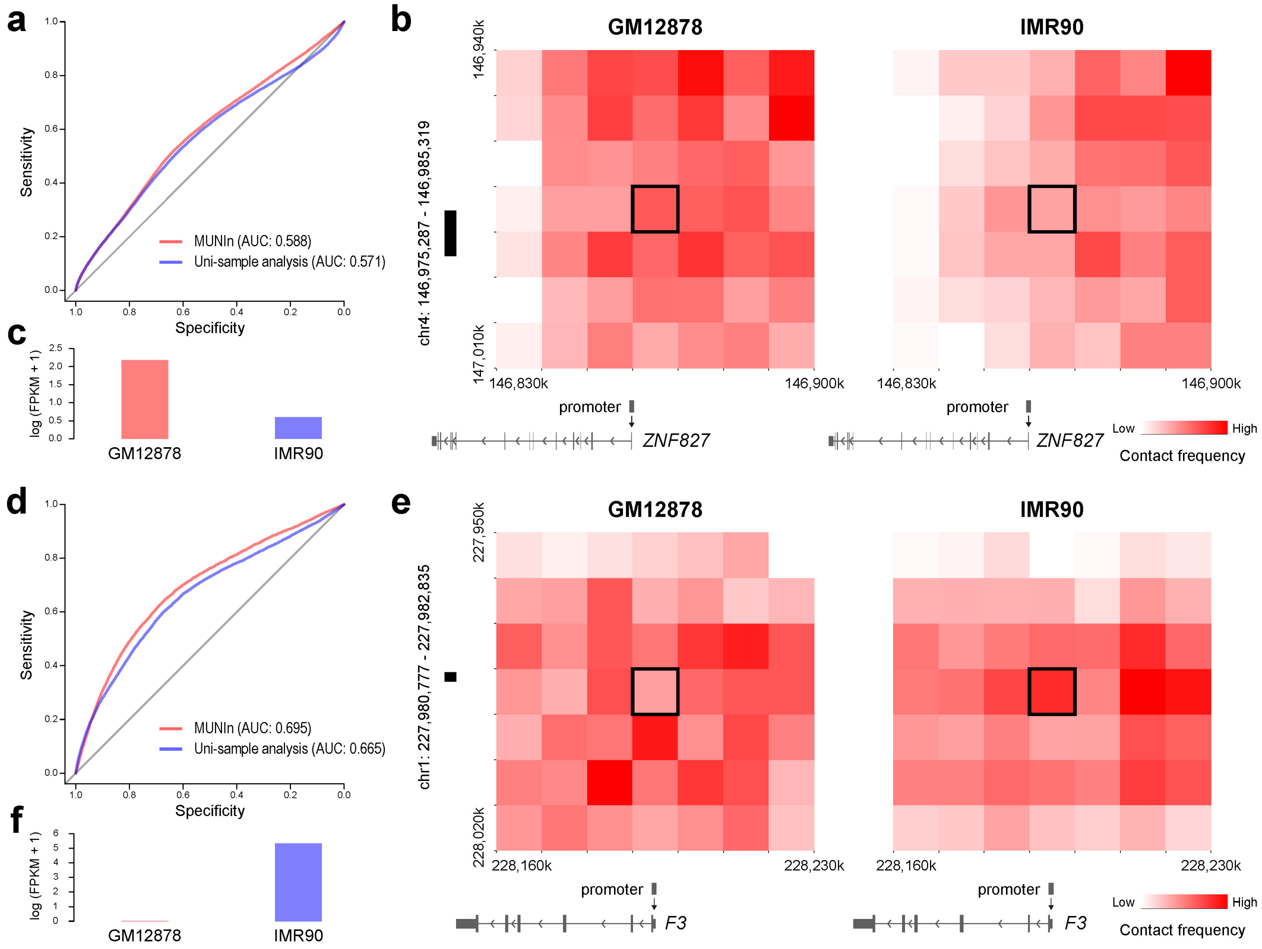


Figure 5. Performance comparison between MUNIn and uni-sample analysis in the Hi-C data of GM12878 and IMR90 cell lines. (**a**) ROC for GM12878-specific peaks identified by MUNIn and uni-sample analysis. (**b**) Heatmap showing one example of the GM12878-specific peaks in GM12878 (left) and IMR90 (right) Hi-C data. One bin of this pair (highlighted in black) is overlapped with the promoter of *ZNF827* gene (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. The Gene model is obtained from the WashU epigenome browser (PMID:31165883) (**c**) Gene expression profiles of *ZNF827* in GM12878 and IMR90 cells. (**d**) ROC for IMR90-specific 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 the *F3* gene, while the other is overlapped with a known typical enhancer (chr1:227,980,777-227,982,835) in IMR90 cells. Gene mode is obtained from the WashU epigenome browser. (**f**) Gene expression profiles of *F3* gene in GM12878 and IMR90 cell lines.

stem cells for both wild-type (without CTCF depletion) and after CTCF deletion at 10kb resolution (**Table A1**). The results showed that MUNIn better captured the wild-type-specific pattern in mESC Hi-C data than uni-sample analysis and FitHiC (**Figure A14 and A15**; **Appendix A**), demonstrating the power of MUNIn to reveal peaks more powerfully and accurately by borrowing information from another sample.

Taking the advantage of jointly modeling multiple samples, MUNIN can easily accommodate many more samples simultaneously. MUNIn shows a high computational efficiency that MUNIn takes $\sim$36 minutes to perform peak calling in a 2MB TAD of 10kb resolution (**Figure A16 and A17**; **Appendix A**). Moreover, MUNIn is able to handle multiple samples with differential levels of dependency, for example, when samples forming 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.

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 improved power in detecting shared peaks, and 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.

CHAPTER 3: A GWAS TO STUDY COGNITIVE IMPAIRMENT AMONG PRETERM CHILDREN

3.1 Introduction

Extreme prematurity (birth<28 weeks of gestation) remains one of the leading causes of neonatal morbidity and mortality in the US [68]. Although survival rates for infants born extremely preterm have improved dramatically in recent decades, children born extremely preterm remain at higher risk for cognitive impairment, with lower average general intelligence and executive function deficit [69, 70, 71, 72, 73] and 9-fold higher risk of severe cognitive impairment compared to children born full-term [74, 75, 76, 77, 78, 79, 80].

Despite substantial research efforts to understand neurodevelopment outcomes, we know remarkably little about genetic factors and molecular mechanisms influencing cognitive function in preterm children. Some genetic studies have evaluated genetic risk factors for neurodevelopmental outcomes for preterm children or children with low birth weight [81, 82, 83, 84]. However, previous studies do not explain the pathways through which genetic variants or genes might influence the risk of poor cognitive outcomes, and few genome-wide association studies (GWAS) examined the genomic regions associated with cognitive function among children born extremely preterm. Therefore, identifying genetic factors that are associated with children's cognitive function and understanding related mechanisms are necessary to develop earlier screening assessments and effective precision interventions and understand why some preterm children of the same gestational age do worse than others. To advance along these directions, we utilized samples from the Extremely Low Gestational Age Newborns (ELGAN) cohort [85], the largest US-based study of children born extremely preterm, to identify genetic factors associated with cognitive impairment at age 10 years. Integrative analysis with brain expression quantitative trait loci (eQTL) and chromatin interactome data was performed to identify potential causal variants and functional genes underlying the GWAS associations.

3.2 Methods

3.2.1 Study participants

ELGAN is a multicenter cohort study originally designed to identify exposures increasing the risk of structural and functional neurologic disorders in children born extremely preterm [85]. We included 528 children in the ELGAN2 cohort born before the 28th week of gestation who had genotype data available for analysis. **Table B1** summarizes demographic information for study participants (**Appendix B**).

3.2.2 Cognitive function at age 10 years

Cognitive function at age 10 years was assessed with Latent Profile Analysis (LPA) [86], which empirically identifies subgroups of children who share similar profiles on a set of measures. The LPA included 9 cognitive measures including verbal and nonverbal IQ and several measures of executive function (EF). LPA classifies subjects who share a similar pattern of scores on the measured variables while maximizing the difference in scoring patterns across distinct profiles [87]. It assigns subjects to a finite number of profiles by identifying the most likely model that describes the heterogeneity of data, which is known as finite mixture models. For our analysis, we used a binary classification that grouped participants into two previously validated distinct profile groups (LPAx) [86]: no or low cognitive impairment and moderate-to-severe cognitive impairment.

3.2.3 Genome-wide association analysis

Quality control and genotype imputation are described in **Appendix B**. For the association analysis, we used EPACTS 3.3.0 [88] for single variant association testing. To account for the relatedness among samples, we used the EMMAX (Efficient Mixed Model Association eXpedited) test [89], which is an efficient implementation of mixed model association accounting for sample structure including population structure and hidden relatedness. Biallelic SNPs with MAF>2% (did not account for relatedness) and Rsq>0.8 were included in the analysis. In total, 8,535,130 variants were included in the association analysis. For the 528 samples that had genotype and covariates data available, we inferred the kinship matrix using EPACTS and the top 10 principal components (PCs) from the genotype data using PLINK. We performed the association test on the outcome LPAx, a binary outcome that classifies children into no or low cognitive impairment and moderate-severe cognitive impairment groups. The covariates for the single variant association analysis included gestational age, maternal education, maternal race, sex of the infant, and top 10 PCs.

3.3 Results

3.3.1 Association analysis results

We conducted a GWAS on LPAx of 528 samples from the ELGAN2 cohort. We identified two genome-wide significant loci from the 8,535,130 variants tested: *STX18* and *TEAD4*, which are located on chromosome 4 and chromosome 12, respectively (**Figure 1**). The index SNPs are rs79453226 (MAF=0.036) and rs11829294 (MAF=0.145) at the *STX18* and *TEAD4* loci, respectively. **Table 1** shows genome-wide significant variants.

3.3.2 Functional annotations

To further investigate the two loci identified for potential mechanisms, we examined several functional annotation metrics, including the CADD phred score [91] and the fathmm MKL score [92]. We also looked at the Genehancer feature [93] and the genes predicted by Genehancer. **Table B2** shows functional annotations for variants that passed the genome-wide significant p-value threshold (p-value<5e-8) (**Appendix B**). We observed that variants rs9424366, rs79946490,

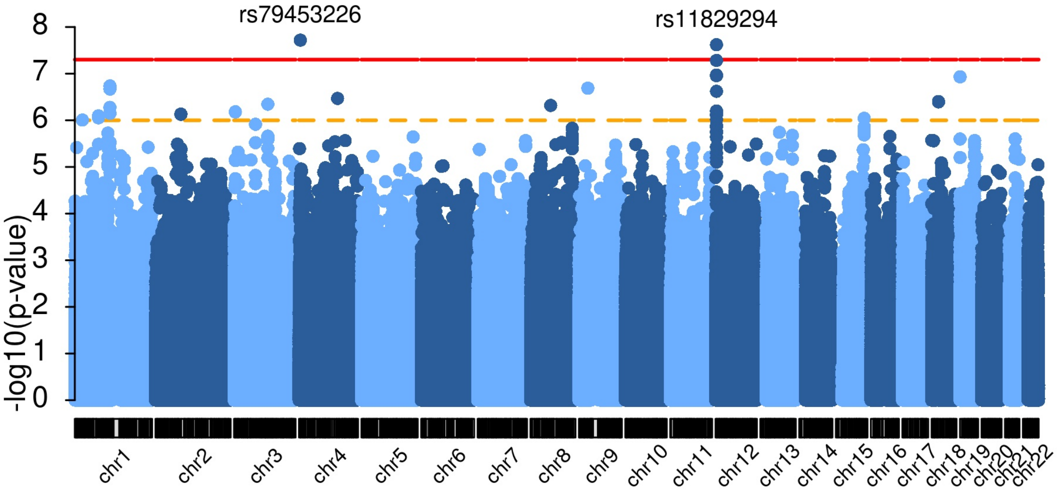


Figure 1. Manhattan plot The Manhattan plot visualizes the association of SNPs along the genome with the LPAx trait. The X-axis represents the genomic location and the y-axis represents -log10(p-value). Each dot represents a SNP tested. SNPs above the red horizontal line, which marks the 5×10^{-8} are considered genome-wide significant. This plot was generated using the R package *karyoploteR* [90]. NCBI build 38.

rs58545250, and rs17031018 were among the top 10% most deleterious in the human genome, and variant rs16913588was predicted to be deleterious (with a fathmm MKL score of 0.97). Several variants were assigned by Genehancer as falling into enhancer regions with target genes *TSPAN9*, *ITPR1*, and *CLIC4*. These results provide evidence that some of the variants might have deleterious effects that are relevant to neurocognitive development in preterm children and suggest additional genes that might be functionally related.

3.3.3 Chromatin interactions

We examined chromatin conformation data for additional functional implications based on physical contacts from Hi-C and alike technologies. **Figure 2** shows virtual 4C plots generated by HUGIn2 [13] for the top two loci in adult cortex and fetal cortex Hi-C data [58]. We examined ±500kb regions around each locus and observed significant chromatin interactions between the putative regulatory regions (harboring GWAS variant(s)) and promoters of likely causal or effector genes. Specifically, the variant rs79453226 at the *STX18* locus was linked to the promoter regions of several genes, including *STX18* and *NSG1* (**Figure 2a**), and the variant rs12322215(p-

Table 1 Significant association results for LPAx. Ordered by significance. $*$: NCBI build 38

rsID	Chr$*$	Position$*$	REF	ALT	P-value	MAF	Locus	Effect size (s.e.)
rs79453226	chr4	4483114	G	C	1.91e-08	0.036	*STX18* (intron)	0.421 (0.074)
rs11829294	chr12	3014153	C	T	2.40e-08	0.145	*TEAD4* (intron)	-0.231 (0.041)
rs10774094	chr12	3014630	C	A	5.21e-08	0.160	*TEAD4* (intron)	-0.214 (0.039)

value=1.08e-07) in high LD with the lead SNP rs11829294 (LD r^2=0.88) at the *TEAD4* locus was linked to a number of genes including *TSPAN9* and *PRMT8* (**Figure 2b**).

3.3.4 Overlapping with brain eQTL

Next, we investigated whether we could find any brain eQTL signals among the top variants. We examined all variants with LD with variants that passed the suggestive p-value threshold (p-value $<$1e-6) using LD calculated from TOPMed European ancestry samples. **Table B3** shows variants overlapped with commonMind eQTL [94] with FDR$<5\%$ (**Appendix B**). Multiple brain eQTLs for *PRMT8* on chromosome 12 in LD with the index SNP rs11829294 were identified.

3.4 Discussion

Leveraging an LPA-derived phenotype and genetics data, we identified two genome-wide significant loci in our genome-wide association analysis for LPAx (a data-derived cognitive impairment outcome): *TEAD4* (rs11829294, p-value=2.40e-8) and *STX18* (rs79453226, p-value=1.91e-8). We utilized chromatin conformation data from multiple human cell lines and primary tissues to see whether there are significant chromatin interactions between the two genome-wide significant loci and their neighboring regions. In the adult cortex and fetal cortex, we found that variant rs12322215 (p-value=1.08e-07) in high LD with rs11829294 (r^2=0.883) is linked to promoter regions of a few genes including *TSPAN9* and *PRMT8* (**Figure 2**). Furthermore, the association at the *TEAD4* locus rs11829294 and a few other variants that showed suggestive significance at the same locus were assigned by Genehancer as falling into the enhancer region of *TSPAN9* (**Table B2**). We also observed *TSPAN9* is highly expressed in both the adult cortex and fetal cortex but not in the hippocampus, and we did not observe similar chromatin interactions in the hippocampus. These pieces of evidence suggest the potential regulatory role of rs11829294 and its LD buddies on the *TSPAN9* gene that could impact cognitive development among children born extremely preterm.

We also performed an integrative analysis with brain eQTL to identify potential functional genes underlying the genome-wide significant association. A few brain eQTL for *PRMT8* were found to be in high LD with rs11829294 (**Table B3**). Along with the evidence that variant rs12322215 is linked to the promoter region of *PRMT8*, we conclude that *PRMT8* is another biologically plausible gene regulated by eQTL at the *TEAD4* locus that could have potential effects on cognitive impairment among preterm children. For rs79453226, we found that it is linked to promoter regions of *STX18* and *NSG1* (**Figure 2**). We did not find as much evidence for the *STX18* locus supporting the significant association as for the *TEAD4* locus.

One limitation of our analysis is that our results may not be generalizable to children who are not extremely premature. Another issue is the small sample size, although we were able to impute most variants well, it limits the statistical power of the association analysis. The few genome-wide significant single variant associations we found, and the non-statistically significant heritability estimate also suggest the need for better-powered analyses (Appendix B). It is also possible that variants included in our analyses are in low or moderate LD with true causal variants which are rare and cannot be well-imputed in the ELGAN2 cohort. While ELGAN2 is the largest cohort with genotype and long-term cognitive assessment for extremely preterm children currently available in the US, in the future we hope to study a larger population with longitudinal data of cognitive function, to investigate whether there are genetic variants that interact with perinatal and neonatal immune factors to increase the risk for development of trajectories of impaired cognitive function.

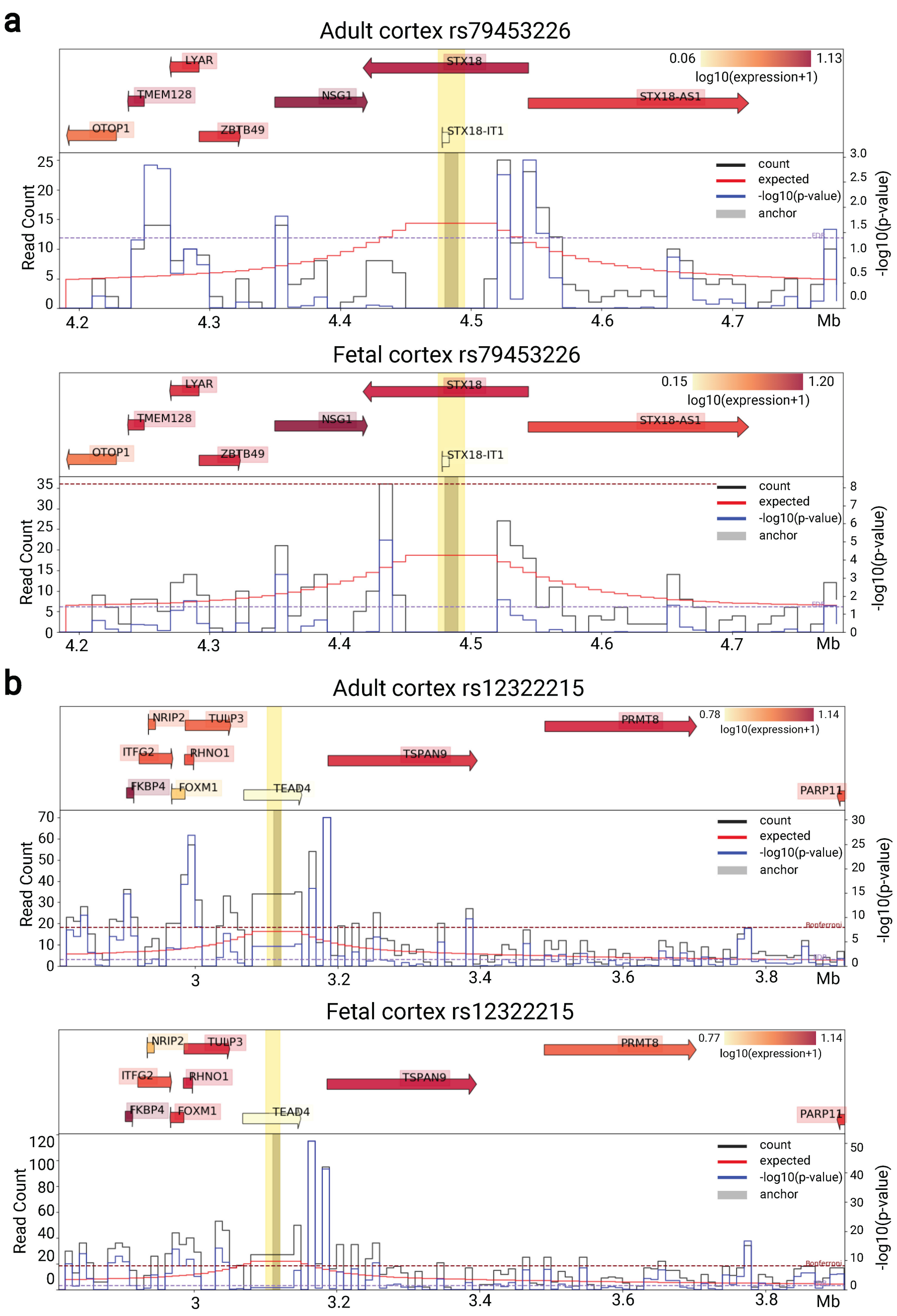


Figure 2. Virtual 4C plots centered at (a) rs79453226 (b) rs12322215 in the adult cortex and fetal cortex. The bin containing the anchor position is indicated as a thick grey vertical bar. On the top is gene expression data with gene locations. Each gene is indicated by an arrow pointing in the direction of transcription. The start site is indicated by the tail of the arrow. On the bottom is the chromatin interaction Hi-C data that is plotted as a virtual 4C plot with the given anchor position. The black line shows the observed counts, the red line shows the expected counts, and the blue line shows the -log10(p-value). The range of the -log10(p-value) is plotted on the y-axis on the right while the range of the count data is shown on the left. The x-axis is the genomic location in Mb. NCBI build 37.

CHAPTER 4: SNAPHiC-G: IDENTIFYING ENHANCER-PROMOTER INTERACTIONS FROM SCHI-C DATA

4.1 Introduction

As discussed in Section 1.2, single-cell Hi-C and its derived co-assays, such as sc-methyl-Hi-C and sn-m3c-seq, provide powerful tools to measure spatial proximity between cis-regulatory elements and their target genes in individual cells. The recent work SnapHiC [95] and SnapHiC2 [96] perform data imputation by RWR first and then combine global and local background models to identify chromatin loops from single cells of the same cell type. While being the first and only existing method to detect chromatin loops from scHi-C data, most SnapHiC-identified chromatin loops are CTCF-anchored structural loops due to the local background model, while the sensitivity to identify enhancer-promoter interactions is relatively low. Currently, no method exists for explicitly identifying enhancer-promoter interactions from scHi-C data.

To fill this gap, we propose SnapHiC-G, a new computational approach based solely on a global background model to identify long-range enhancer-promoter interactions from scHi-C data. We applied SnapHiC-G to re-analyze scHi-C datasets generated from mouse embryonic stem cells (mESCs) and human brain cortical cells. We showed that SnapHiC-G outperformed SnapHiC and existing methods designed for bulk Hi-C data, achieving higher sensitivity with comparable precision in identifying long-range enhancer-promoter interactions.

4.2 Results

4.2.1 Overview of the SnapHiC-G algorithm

The SnapHiC-G algorithm consists of four components: (i) imputing chromatin contact probabilities in every single cell, (ii) distance-stratified normalization of imputed contact probabilities,

(iii) filtering candidate bin pairs, and (iv) identifying statistically significant long-range enhancer-promoter interactions. Given single-cell contact matrices, SnapHiC-G first applies the RWR algorithm to generate imputed contact probability matrices for every single cell using a sliding window approach (Methods: SnapHiC-G algorithm). Next, the imputed contact probabilities are converted into distance stratified Z-scores to account for the dependence between contact probability and the 1D genomic distance between two bins. To identify enhancer-promoter interactions, SnapHiC-G filters bin pairs based on transcript start sites (TSS) and the available epigenetic annotations (e.g., H3K27ac ChIP-seq peaks or ATAC-seq peaks) to obtain a set of candidate bin pairs that span between gene promoters and cis-regulatory regions (Methods: SnapHiC-G algorithm). SnapHiC-G then defines enhancer-promoter interactions based on the global background by applying the one-sample t-test for each tested bin pair across all single cells belonging to the same cell type. Specifically, for each bin pair, the one-sided hypothesis test is conducted where the null hypothesis states that the bin pairs average value of normalized contact probability across all single cells equals zero. The alternative hypothesis states that the average normalized contact probability is greater than zero. By default, bin pairs with FDR<0.1 and t-statistics>3 are identified as significant enhancer-promoter interactions (Methods: SnapHiC-G algorithm). In our analysis, all scHi-C data are binned into 10Kb resolution unless stated otherwise.

4.2.2 Benchmarking with mouse embryonic stem cells (mESCs)

We applied SnapHiC, SnapHiC-G, FitHiC2, FastHiC, HiC-ACT, and HiC-DC+ single-cell Hi-C data generated from 742 mouse embryonic stem cells (mESCs) [27], where the latter four are methods designed for bulk Hi-C data. We aggregated single-cell Hi-C data for all cells as a pseudo-bulk Hi-C sample as input for bulk Hi-C methods (Methods: Identification of loops/interactions using other Hi-C methods). To benchmark against other methods, a reference list of significant interactions was constructed using HiCCUPS-identified loops from deeply sequenced bulk Hi-C data [22] and MAPS-identified significant interactions from H3K4me3 PLAC-seq [97], cohesion HiChIP [98], and H3K27ac HiChIP data [99]. We took the union of all reference interactions and kept only bin pairs with a genomic distance between 20Kb and 1Mb for evaluation. The same filtering step implemented in SnapHiC-G was applied to outputs from other methods to ensure a fair comparison.

When applying to the complete set of 742 mESCs, SnapHiC-G identified notably more significant enhancer-promoter interactions than other methods and reached a genome-wide power of 80%, recovering most of the 38,588 interactions in the reference list (Figure 1A; Table 1). Due to extreme data sparsity, bulk Hi-C methods missed most reference interactions without imputing single-cell contact probabilities. FitHiC2 performed the best among other methods, calling 3,476 interactions with a genome-wide power of 16%. FastHiC identified more interactions than FitHiC2 with a lower genome-wide power of 12%. As expected, SnapHiC identified few enhancer-promoter interactions due to the local background model. Although already tailored for sparse scHi-C data, SnapHiC performed similarly to HiC-ACT, a global background method. HiC-DC+ identified the fewest interactions among all methods with the lowest genome-wide power.

Since the number of input cells is critical in scHi-C data analysis, we assessed whether SnapHiC-G could retain its performance with fewer cells. Among all 742 mESCs, 100 mESCs were randomly selected, and the same performance evaluation of the six methods mentioned above was repeated. As shown in Figure 1B and Table 1, all methods had reduced power with 100 mESCs, while SnapHiC-G was least affected by the number of cells and showed more significant power gain over other methods compared with results from 742 cells. With only 100 mESCs, SnapHiC-G retained a genome-wide power of 61%, while all the other methods had genome-wide power below 10%. FastHiC still performed the best among others, with 3,576 significant interactions identified, comparable to the complete data, but the genome-wide power reduced almost by half to 7%. On the other hand, SnapHiC reached a 3% genome-wide power with 265 loops called; however, still more sensitive than other bulk Hi-C methods.

Due to the large number of significant interactions detected by SnapHiC-G, we evaluated the precision of identified interactions across the six methods with the same reference list. To control

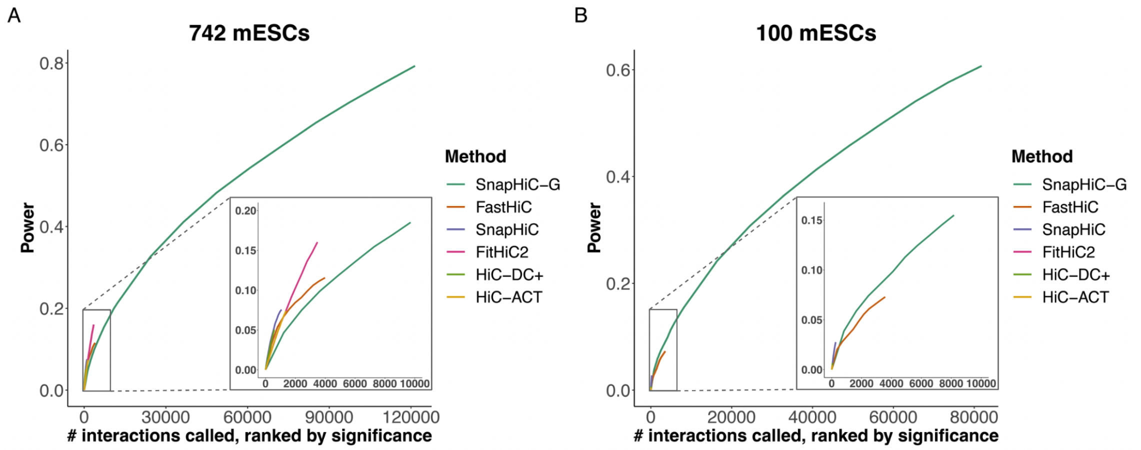


Figure 1. Power curves with (a) 742 and (b) 100 mES cells. Interactions were ranked by significance on the x-axis and power was evaluated with the corresponding number of top interactions. The lower right-hand corner sub-figures are zoomed-in views of the top 10,000 interactions for each method.

for the number of bin pairs compared, we ranked identified interactions in each method by their significance (i.e., p-value for HiC-ACT or FDR for FitHiC2, HiC-DC+, SnapHiC) or posterior probability (FastHiC) and calculated precision for the top 1,000, 2,000, 5,000, and 10,000 interactions. SnapHiC-G showed a comparable or better performance in terms of precision among the most significant interactions, even with a much larger number of interactions called (Figure 2). For example, with 742 mESCs, SnapHiC-G attained a precision of 0.93 for the top 1,000 interactions, which was comparable with HiC-DC+ (0.96) and FastHiC (0.96) and substantially higher than HiC-ACT (0.73), FitHiC2 (0.73), and SnapHiC (0.72). With 100 mESCs, SnapHiC-G had a precision of 0.60-0.76 for the top 1,000 to 10,000 interactions. We performed the same evaluation with three cell types from human brain cortical cells (oligodendrocytes, microglia, and L2/3 neurons) and had similar observations (Supplementary notes; Figure C3; Figure C4). Taken together, we have shown that SnapHiC-G had much higher sensitivity than other methods while maintaining precision among top enhancer-promoter interactions even with a small number of cells.

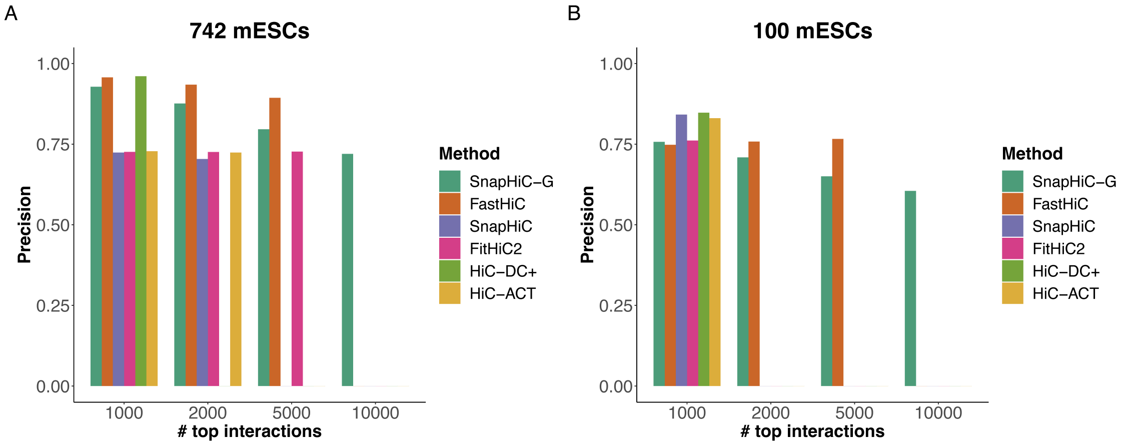


Figure 2. Precision bar plots with (a) 742 and (b) 100 mES cells. Results shown were for the top 1,000, 2,000, 5,000, and 10,000 interactions ranked by significance. Some bars are missing because the number exceeds the number of interactions called by that method.

Table 1 Genome-wide power and the number of interactions called for mESCs and three brain cell types.

	742 mESCs		100 mESCs		L2/3 neurons		Microglia		Oligodendrocytes	
Method	# of interactions	Sensitivity	# of interactions	Sensitivity	# of interactions	Sensitivity	# of interactions	Sensitivity	# of interactions	Sensitivity
HiCDC+	582	0.050	59	0.007	4359	0.283	4321	0.245	9881	0.466
FastHiC	3988	0.116	3576	0.073	6585	0.207	7852	0.245	15583	0.492
HiC-ACT	1365	0.074	53	0.005	1471	0.159	1881	0.172	9166	0.575
FitHiC2	3476	0.160	67	0.005	4019	0.298	4834	0.310	18708	0.723
SnapHiC	1070	0.076	265	0.028	2918	0.203	1766	0.147	3605	0.242
SnapHiC-G	12147	0.213	8183	0.155	23680	0.322	16590	0.346	65389	0.587

4.2.3 Enrichment of brain eQTL-TSS pairs in human cortical cells

To evaluate the functional characteristics of SnapHiC-G-identified enhancer-promoter interactions, we re-analyzed single-nucleus methyl-3C-seq (sn-m3c-seq) data from 2,869 human prefrontal cortical cells [100]. We collected eQTL data from various sources, including eQTL data from brain cell types released from Netherlands Brain Bank (NBB), the MS UK Tissue Bank (UKTB), and the Edinburgh Brain Bank (EBB) [101], the CommonMind Consortium brain eQTL data [94], and GTEx consortium v7 liver eQTL data [102] as a control sample. For sn-m3c-seq data, cell types were classified based on DNA methylome as described in the original study. We applied SnapHiC-G to four major cell types (astrocytes [n=338], microglia [n=323], oligodendrocytes [n=1038], and L2/3 neurons [n=261]) separately, where single cells from the same cell type were pooled (Figure 3A; Table 1; Table C1). We created control bin pairs to compute the enrichment of SnapHiC-G-identified enhancer-promoter interactions as eQTL-TSS of eGene pairs for each cell type. Specifically, we generated a pseudo bin pair for each significant enhancer-promoter interaction by retaining the bin with the promoter as the center but flipping the other bin to the opposite side of the center (Methods: Enrichment of eQTL-TSS pairs). We found that the odds for a bin pair to be an eQTL-TSS pair were significantly higher for significant interactions than pseudo bin pairs by overlapping with true eQTL-TSS bin pairs (Figure 3A). The wide confidence intervals (CI) of the odds ratios (OR) calculated from the Bryois et al. data [101] were due to a smaller sample size to detect eQTLs compared with the CommonMind brain eQTL data (192 versus 467 brain samples, respectively). The two sets of brain eQTL data were highly consistent in terms of the order of enrichment for the four cell types, where oligodendrocytes showed the strongest enrichment, followed by astrocytes and microglia, and L2/3 neurons the weakest. As expected, SnapHiC-G-identified interactions were more enriched in brain eQTLs than in liver eQTLs. For example, the odds ratio for eQTLs from Bryois et al., CommonMind brain, and CommonMind liver in astrocytes were 2.35 (95% CI: 1.66 - 3.36), 1.95 (95% CI: 1.92 - 1.98), and 1.42 (95% CI: 1.29 - 1.55), respectively. These results validated the functional importance of SnapHiC-G-identified enhancer-promoter interactions in relevant cell types and tissues.

4.2.4 Gene expression patterns of cell-type-specific enhancer-promoter interactions

To further access the biological relevance of SnapHiC-G-identified enhancer-promoter interactions, we evaluated gene expression patterns in cell-type-specific interactions from the four brain cell types. Cell-type-specific interactions were defined as exclusively present in only one cell type, not including those shared by two or more cell types (Figure 3B). To avoid the impact of the number of cells in each cell type, we down-sampled astrocytes, microglia, and oligodendrocytes to 261 cells each to match the number of L2/3 neuron cells and identified cell-type-specific interactions with SnapHiC-G (Methods: Down-sampling of scHi-C data). Next, we selected genes with promoters overlapping with cell-type-specific interactions and extracted gene expression levels from RNA sequencing data from corresponding cell types [103]. As shown in Figure 4A, gene expression levels were higher for the genes from cell-type-specific enhancer-promoter interactions in the corresponding cell type. Moreover, after removing genes that overlapped with enhancer-promoter interactions detected by SnapHiC, a substantial number (above 65%) of the selected genes remained. We also observed increased expression of these genes in a cell-type-specific manner (Figure 4B). These results add another line of evidence that predicted enhancer-promoter interactions could provide valuable information in a cell-type-specific manner on top of the eQTL enrichment analysis.

4.2.5 Assigning GWAS variants to putative target genes

With over 90% of GWAS variants associated with human complex diseases and traits residing in non-coding regions yet enriched in cis-regulatory elements (e.g., promoters, enhancers, silencers, and insulators), enhancer-promoter interactions have the potential to prioritize disease-relevant genes for non-coding variants, particularly those in close spatial proximity that are far away in the 1D genomic distance with the promoter of their target genes. To assign putative target genes to non-coding GWAS variants based on predicted enhancer-promoter interactions in brain cell types, we collected the latest GWAS summary statistics for eight neurodevelopmental and neurodegenerative disorders: Alzheimer's disease (AD) [105], attention deficit hyperactivity

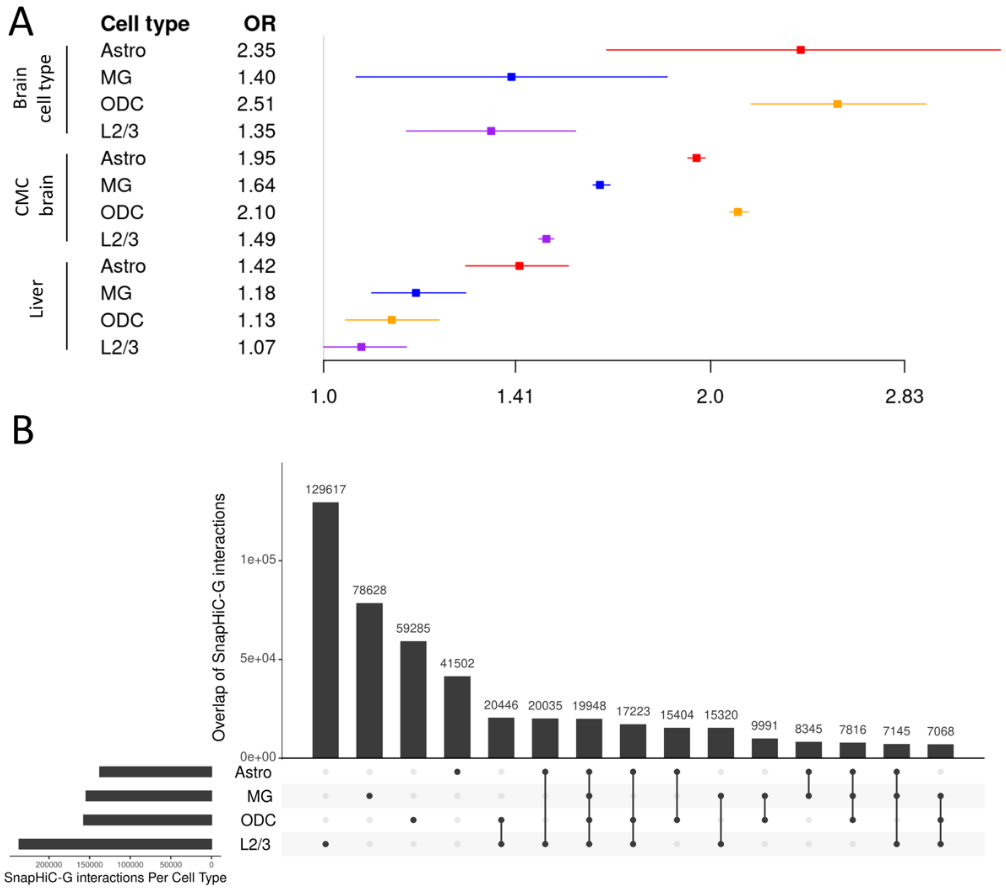


Figure 3. (A) Enrichment of SnapHiC-G-identified interactions for each brain cell type in CommonMind and liver eQTL-TSS pairs. The squares denote point estimates for odds ratios (ORs) and the error bars denote 95% confidence intervals for ORs. OR was defined as the odds of SnapHiC-G-identified interactions overlapping with eQTL-TSS bin pairs to the odds of pseudo bin pairs overlapping with eQTL-TSS bin pairs. **B) UpSet plot for SnapHiC-G interactions.** Number of exact overlapped interactions between four cell types: astrocytes (Astro), microglia (MG), oligodendrocytes (ODC), and L2/3 neurons (L2/3). The interactions were identified from 261 cells from each cell type. The UpSet figure is generated using the R package UpSetR [104].

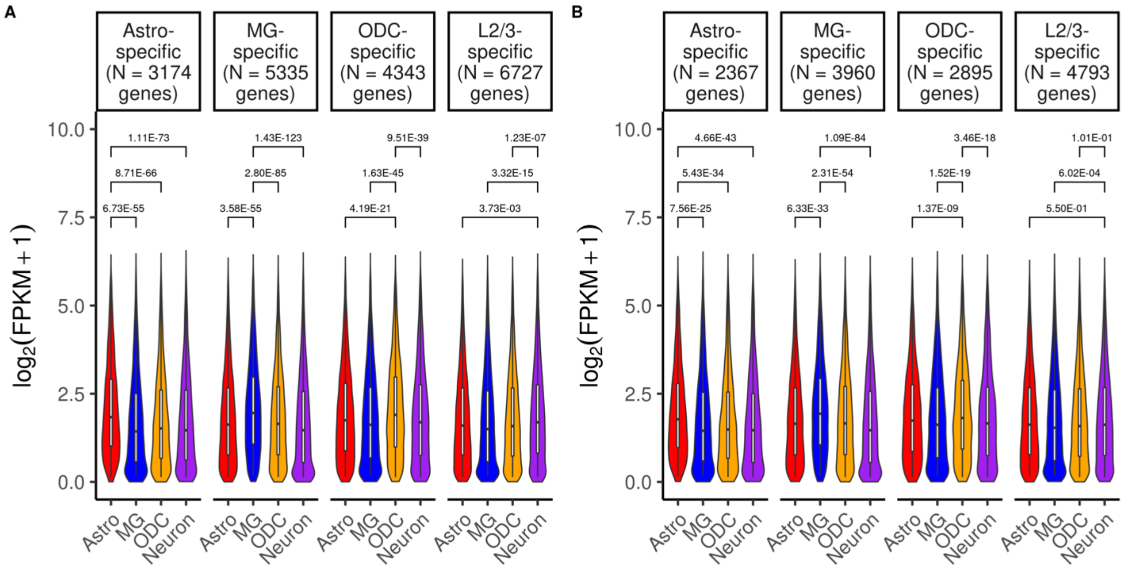


Figure 4. Cell-type-specific gene expression showing violin plots of RNA-seq expression levels (log2(FPKM+1) value) for selected genes. (A) Genes overlapping with cell-type-specific SnapHiC-G enhancer-promoter interactions. (B) Genes overlapping with cell-type-specific SnapHiC-G enhancer-promoter interactions, and those overlapping with enhancer-promoter interactions detected by SnapHiC were subsequently removed. P-values were calculated from paired Wilcoxon signed-rank tests. Gene expression outliers for each cell type were removed for visualization.

disorder (ADHD) [106], autism spectrum disorders (ASD) [107], bipolar disorder (BP) [108], schizophrenia (SCZ) [109], Parkinson's disease (PD) [110], major depressive disorder (MDD) [111], and neuroticism (NEU) [112], and two complex traits: educational attainment (EDU) [113] and intelligence quotient (IQ) [114]. We again focused on cell-type-specific enhancer-promoter interactions to predict target genes in a cell-type-specific manner using down-sampled sn-m3c-seq data. Specifically, SnapHiC-G identified 137,418, 154,261, 157,181, and 236,802 enhancer-promoter interactions in astrocytes, microglia, oligodendrocytes, and L2/3 neurons, respectively (Table 2). After excluding interactions shared among cell types, 14,440, 39,220, 23,853, and 65,819 cell-type-specific interactions were left correspondingly. To facilitate the interpretation of the GWAS variants, we focused on non-coding GWAS variants that reside in an active enhancer region in astrocytes, microglia, oligodendrocytes, or L2/3 neurons [115]. When matching GWAS variants and SnapHiC-G results, for each cell-type-specific enhancer-promoter interaction, we required that one bin contains GWAS variant(s) and the other bin overlaps with a genes TSS, and we annotated this gene as the putative target gene. Furthermore, we required that the corresponding gene is highly expressed (FPKM >1) in this cell type and lowly expressed (FPKM $\leqslant 1$) in the other three cell types. We found 35, 82, 7, and 98 matched enhancer-promoter interactions (222 in total) for astrocytes, microglia, oligodendrocytes, and L2/3 neurons, respectively, and resolved over 600 SNP-disease associations (Table 2). Moreover, the average number of target genes for each variant was close to 1, much smaller than the number of nearby genes (+/- 1 Mb region), ranging from 25 to 83. For example, in astrocytes, the average number of target genes and nearby genes per variant were 1.1 and 38.3, respectively. These results showed that we could pinpoint target genes of non-coding variants in a cell-type-specific manner by integrating SnapHiC-G-identified enhancer-promoter interactions with GWAS results.

4.2.6 Examples of cell-type-specific enhancer-promoter interactions

From the 222 matched cell-type-specific enhancer-promoter interactions mentioned in the previous section, we were able to map a GWAS variant residing in an active enhancer with the

Table 2 Summary of SnapHiC-G interactions for each brain cell type. SnapHiC-G interactions: number of interactions identified from SnapHiC-G; cell type-specific interactions: number of interactions identified only in this cell type and not in any other three cell types; matched interactions: number of cell type-specific interactions with one bin containing the GWAS SNP residing in an active enhancer while the other bin overlapping with a gene's TSS and the corresponding gene is highly expressed (FPKM > 1) in this cell type and lowly expressed (FPKM <= 1) in other three cell types; unique GWAS SNPs: number of unique SNPs contained in the matched interactions; SNP-disease associations: number of GWAS disease-SNP associations in the matched interactions; the average number of target genes per GWAS SNP: average number of targeted genes for each GWAS SNP based on the matched interactions; the average number of (+/- 1 Mb) genes per GWAS SNP: average number of nearby genes (within 1 Mb) for each GWAS SNP.

	SnapHiC-G Interactions	Cell-Type-Specific Interactions	Matched interactions	Unique GWAS SNPs	SNP-Disease Associations	Avg.# of Target Genes per SNP	Avg.# of +/- 1Mb Genes per SNP
Astrocyte	137,418	14,440	35	72	72	1.07	38.32
Microglia	154,261	39,220	82	279	288	1.03	82.86
Oligodendrocytes	157,181	23,853	7	22	39	1.00	46.91
L2/3 Neurons	236,802	65,819	98	181	202	1.02	24.61

promoter of a cell-type-specific gene (Table 2). Notably, most of these interactions were not identified by SnapHiC with the local background approach, as shown in Figure 4B. We demonstrate how SnapHiC-G-identified enhancer-promoter interactions can elucidate functional genes of GWAS loci in relevant cell types with a few examples.

The first example locates at a locus on chromosome 8, showing cell-type-specific interactions in L2/3 neurons and astrocytes (Figure 5). Specifically, a SCZ-associated GWAS SNP rs2565064 (chr8: 27,327,841) interacts with the promoter of *PNOC* in neurons, while another SCZ-associated SNP rs28541694 (chr8: 27,462,008) interacts with the promoter of *ZNF395* in astrocytes. In addition, ten AD-associated SNPs located in the chr8:27.4Mb-27.5Mb region are also connected to the promoter region of *ZNF395* in astrocytes, while none of the AD-associated SNPs interact with the *PNOC* promoter. These two genes also showed consistent cell-type-specific gene expression patterns, where *PNOC* was highly expressed in neurons (FPKM = 4.75 in neurons vs. $\leqslant$ 1 in the other three cell types) and *ZNF395* was highly expressed in astrocytes (FPKM = 2.26 in astrocytes vs. $\leqslant$ 1 in the other three cell types). Moreover, rs2565064 resides in a neuron-specific enhancer, and SNPs interacting with *ZNF395* reside in an astrocyte-specific enhancer. These results indicated that *PNOC* was the putative target gene for SCZ-associated SNP rs2565064 in neurons and *ZNF395* was the putative target gene for both SCZ and AD-associated SNPs in this 27.4Mb-27.5Mb region on chromosome 8. *PNOC* is primarily transcribed in the brain and spinal cord in the central nervous system [116] and encodes the precursor for bioactive neuropeptides that influence a broad range of physiological roles, including memory, learning, and neuronal development, fear, anxiety, and sleep [117]. *PNOC* was also associated with PTSD, whose transcriptome significantly correlated with SCZ [118]. On the other hand, *ZNF395*, a gene involved in inflammation and cancer progression57, is upregulated in the SCZ network from the integrative network analysis [119].

Next, we focused on a microglia-specific interaction between the promoter of *ARPC1B*, which is highly expressed in microglia (FPKM = 7.12 in microglia vs. $\leqslant 1$ in the other three cell types), and an AD-GWAS locus located in an active enhancer on chromosome 7 at 99.7 Mb (Figure C1). Our results showed that *ARPC1B* was the predicted target gene for this AD-associated GWAS variant rs1880949, consistent with prior findings that *ARPC1B* was active in microglia in AD patients but not in healthy controls [120]. At the same locus, SnapHiC-G detected another microglia-specific enhancer-promoter interaction between an active enhancer containing an EDU-associated GWAS variant rs10241492 (chr7: 99,994,813) and the *STAG3* genes promoter region. Moreover, rs10241492 was identified as an eQTL for *STAG3* from CommonMind in brain tissues [113]. In addition, *STAG3* was predicted to be the target gene for rs10241492 from a previous GWAS study of EDU50. These results together showed that the *STAG3* gene was potentially a target gene for rs10241492, specifically in microglia, consistent with findings in the original paper by Lee et al. [113].

As a final example, Figure C2 illustrates enhancer-promoter interactions identified specifically in microglia and neurons on chromosome 10. One microglia-specific enhancer-promoter interaction links a SCZ-associated GWAS locus in an active enhancer with *SFXN*2s promoter. At the same time, *SFXN2* is highly expressed only in microglia (FPKM = 4.41 in microglia vs.

$\leqslant 1$ in the other three cell types). While a previous study suggested that *SFXN2* was a potential SCZ risk gene because of linkage or pleiotropic effects [121], our results further indicated that *SFXN2* was the putative target gene for this particular SCZ-associated GWAS locus. Additionally, multiple neuron-specific enhancer-promoter interactions connect SCZ-associated GWAS loci to *INA* (FPKM = 46.21 in neurons vs. $\leqslant 1$ in the other three cell types), suggesting this gene is a potential novel target gene for this locus with the evidence that corresponding GWAS variants (rs11191557, rs11191558, rs11191559, rs10883832, rs12413046) in this locus were also CommonMind eQTLs for *INA*.

Together, these examples showcase how SnapHiC-G-identified enhancer-promoter interactions can aid the interpretation of non-coding GWAS variants and reveal underlying mechanistic insights. Integrating with gene expression data and epigenetic annotations, SnapHiC-G was able to decipher the critical roles of non-coding variants in disease etiology in relevant cell types.

4.2.7 Elucidating relevant cell types by heritability enrichment analysis

Next, we evaluated whether genetic heritability for complex diseases and traits was enriched for SNPs within the anchors of SnapHiC-G-identified interactions in specific cell types. Using the same set of GWAS summary statistics data together with two other complex traits: body mass index (BMI) [122] and white blood cell count (WBC) [123], we performed stratified linkage disequilibrium score regression (S-LDSC) analysis [124]. In brief, S-LDSC estimates the proportion of SNP heritability from predefined SNP-level functional annotations using GWAS summary statistics while accounting for linkage disequilibrium (LD) to identify functional categories enriched in SNP heritability and hence of functional relevance to the trait. In our case, the functional categories correspond to enhancer-promoter interactions from SnapHiC-G results in brain cell types, and our goal is to identify disease-relevant cell types for GWAS traits.

We first obtained SnapHiC-G-identified interactions from astrocytes, microglia, oligodendrocytes, and L2/3 neurons, all with 261 cells, to construct functional categories. As previously described, SnapHiC-G requires at least one bin overlapping with a TSS to ensure that the algo-

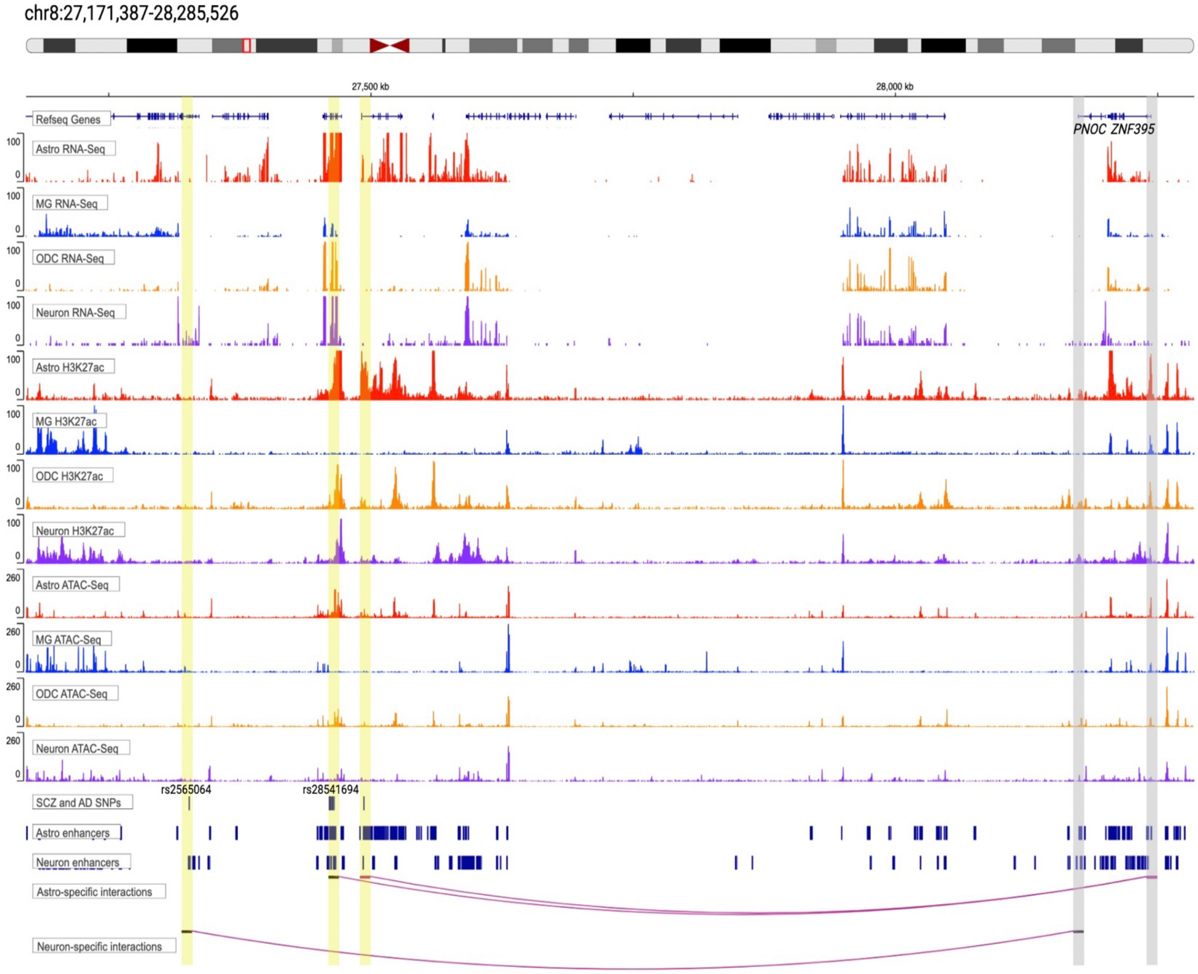


Figure 5. An illustrative example at the *PNOC-ZNF395* locus on chromosome 8 with neuron- and astrocyte-specific interactions. The top panel shows the gene track. The middle panels show RNA-seq, H3K27ac, and ATAC-seq tracks for the four brain cell types. The bottom panels show SCZ and AD GWAS SNPs, enhancer regions in astrocytes and neurons, and cell-type-specific interactions identified by SnapHiC-G but not SnapHiC. Astrocyte-specific interactions link the promoter region of ZNF395 (highlighted in grey) with a SCZ-associated SNP rs28541694 and ten AD-associated SNPs located in the chr8:27.4Mb-27.5Mb locus (highlighted in yellow); a neuron-specific interaction links the promoter region of *PNOC* (highlighted in grey) with another SCZ-association SNP rs2565064 (highlighted in yellow). Both genes highlighted showed cell-type-specific gene expression in corresponding cell types. The anchors of these interactions also showed stronger H3K27ac ChIP-Seq and ATAC-seq signals in matched cell types.

rithm captures promoter-anchored interactions. Therefore, both bins may overlap with a TSS. However, to distinguish bins that overlap with a TSS and those that do not overlap with a TSS, we focused on the case where only one bin overlaps with a TSS. We focused on significant interactions in the four brain cell types to partition the genome. We defined SnapHiC-G anchor regions as the bins that overlap with TSS and SnapHiC-G target regions as the bins that do not overlap with TSS for each cell type. GWAS SNPs were annotated based on whether they fall into SnapHiC-G target regions for each cell type. Since the anchor regions all overlap with TSS, we evaluated whether SNPs located in SnapHiC-G target regions were enriched in SNP heritability in specific cell types.

Figure 6 shows SNP heritability enrichment results for GWAS variants from the twelve traits analyzed, using SnapHiC-G target bins in four brain cell types as the functional categories. For example, AD SNP heritability was most strongly and significantly enriched in microglia target regions, which is consistent with the fact that the majority of AD GWAS risk loci are found close to genes highly expressed in microglia, and that microglia play a vital role in the pathogenesis of AD64. Notably, we observed two neuropsychiatric traits with high genetic correlation, namely bipolar disorder and schizophrenia, had enrichment scores and significance in the same order, where neuronal target regions were most strongly enriched, followed by astrocytes and oligodendrocytes. In contrast, enrichment in microglia was much lower. The strong enrichment in oligodendrocytes for Parkinsons disease, although not the most significant, was consistent with the literature as well [125]. Results for white blood cell count showed that microglia were most strongly enriched, in agreement with the crucial roles that white blood cells have in the immune system.

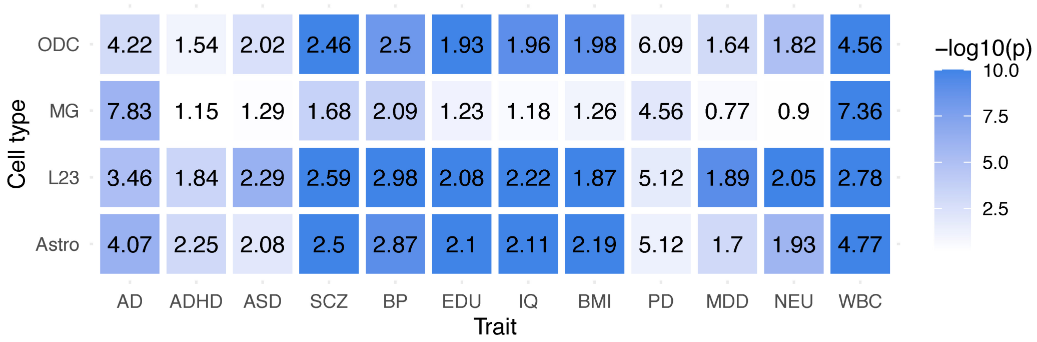


Figure 6. Heritability enrichment analysis results using LDSC. Cells were down-sampled to match the number of cells in L2/3 neurons. Numbers in the figure represent the enrichment score and colors represent the significance level in the -log10(p-value) scale. A higher enrichment score represents stronger enrichment in the corresponding cell type. Abbreviations: Alzheimers disease (AD), attention-deficit/hyperactivity disorder (ADHD), autism spectrum disorder (ASD), schizophrenia (SCZ), bipolar disorder (BP), educational attainment (EDU), intelligent quotient (IQ), body mass index (BMI), Parkinsons disease (PD), major depressive disorder (MDD), neuroticism (NEU), white blood cell count (WBC).

4.3 Conclusion and discussion

In this work, we developed SnapHiC-G, a computational pipeline to detect enhancer-promoter interactions from scHi-C data based on the global background model. To our knowledge, a method has yet to be proposed to address this task in scHi-C data. Our previous work, SnapHiC, the first computational pipeline to detect chromatin loops in scHi-C data, utilizes a combination of global and local background models to identify loop summits, therefore having limited power to detect enhancer-promoter interactions. Other existing global background methods, such as FitHiC2, HiC-DC+, FastHiC, and HiC-ACT, were designed for bulk Hi-C data and lacked the power to analyze sparse scHi-C contact data. To overcome the data sparsity challenge in scHi-C data, SnapHiC-G first applies the RWR algorithm to the observed raw contacts and then constructs a normalized contact probability matrix against the linear genomic distance. For each candidate bin pair, SnapHiC-G applies a one-sample t-test to test whether its average normalized contact frequency across single cells is significantly greater than zero (global background). Combined with epigenetic annotations such as enhancer and promoter marks, SnapHiC-G enables the profiling of cell-type-specific enhancer-promoter interactions by analyzing scHi-C data from multiple cell types.

Data from mESCs and human brain cortical cells showed that SnapHiC-G identified more significant interactions than existing global background-based methods designed for bulk Hi-C data with notably higher sensitivity. This was mainly due to the RWR imputation step, which significantly reduced the sparsity of the scHi-C data, especially when the number of cells was low. Aggregating a small number of single cells can result in very sparse pseudo-bulk Hi-C data, which may lead to poor performance of methods designed for bulk Hi-C. We evaluated the precision of top interactions identified by each method against a reference list of interactions and showed that SnapHiC-G did not suffer from a higher false discovery rate compared with other methods. The extra interactions called by SnapHiC-G may need further investigation, but we were not able to perform a systematic evaluation due to the lack of a gold-standard dataset.

Using GWAS summary statistics, we have also demonstrated the utility of enhancer-promoter interactions identified from SnapHiC-G in human brain cell types in prioritizing functionally relevant genes and cell types for complex human traits and diseases. SnapHiC-G can predict putative target genes for GWAS variants and help elucidate functionally relevant cell types of complex diseases.

For the human brain cell types, SnapHiC-G did not improve substantially over other methods compared with mESCs, especially when we look at the most significant interactions for oligodendrocytes (Figure C3; Figure C4). Several reasons can explain this. First, there was no gold-standard data for evaluating cell-type-specific enhancer-promoter interactions. The reference interaction lists inferred from H3K4me3 PLAC-seq data were suggestive rather than optimal, meaning that the reference might miss many cell-type-specific enhancer-promoter interactions. However, it was the best available data we could find. Second, SnapHiC-G-identified interactions had much lower FDRs than other bulk Hi-C methods. For identified oligodendrocyte interactions, the median of -$\log10$(FDR) was 17.5 for SnapHiC-G, while the medians of other bulk Hi-C methods ranged from 1.7 to 7.5. The highly significant results of SnapHiC-G were due to both the RWR imputation and its global background nature; therefore, it can be hard to distinguish among identified interactions by ranking them by significance. Third, the number of cells for oligodendrocytes was relatively large ($>$1,000). After aggregating the single cells to construct bulk Hi-C data, it had a comparable sequencing depth to traditional bulk Hi-C data with 278 million intra-chromosomal contacts $>$20Kb. Consistent with SnapHiC, we observed a more significant power gain from SnapHiC-G when the number of cells was relatively small, which was consistent for other cell types.

SnapHiC-G performs the statistical test across all cells from the same cell type, which means that signals from the input cells are aggregated, and the identified enhancer-promoter interactions are still at the population level, similar to bulk Hi-C analysis. However, chromatin folding can be highly variable and dynamic even among cells of similar identities, which is an exciting future direction. As more scHi-C data become available, multimodal data integration is another promising area of research to study the complexity and heterogeneity of chromatin interactions in single cells and can provide new insights into the regulatory mechanisms that underlie gene expression and cell differentiation. scHi-C data also has great potential for predicting structural variations in cancer genomes, which is beyond the scope of this work.

In our analysis, cell-type-specific epigenetic data considerably narrowed down candidate bin pairs. While such data aid the detection of cell-type-specific enhancer-promoter interactions, when they are not available, users can input only the TSS files to define the promoter regions and apply SnapHiC-G to identify promoter-interacting regions. With the rapid development of new technologies and more data available, scHi-C can trigger the study of fundamental questions about chromatin spatial organizations in individual cells during development, cancer cells, and different organs. Being able to identify cell-type-specific enhancer-promoter interactions from scHi-C data, SnapHiC-G results can be combined with the widely available GWAS results and epigenetic data, and it has a great potential to facilitate the discovery of regulatory chromatin interactions that are important for gene regulation in biologically relevant cell types.

4.4 Methods

4.4.1 SnapHiC-G algorithm

Step A. Imputation of contact probability using RWR with a sliding window approximation

We followed SnapHiC for imputing intra-chromosomal contact probability in every cell using the RWR algorithm following scHiCluster [126]. Each autosome was divided into consecutive 10 Kb bins, and each bin pair was converted to a binary representation with values 1 representing nonzero contact and 0 representing no contact observed. An unweighted and undirected graph modeled each chromosome by defining bins as the nodes and adjacent bin pairs or bin pairs with nonzero contact as edges. The RWR method was then used with a restart probability of 0.05 to estimate the likelihood of traveling between two nodes allowing for the imputation of contact probability between all intra-chromosomal bin pairs. The random-walk step captures information from global network structures, while the restart step captures information from local network structures. Following SnapHiC2, we adopted a sliding window approach when imputing missing contacts to reduce computational costs. Specifically, instead of performing RWR over the entire chromosome, we divided the original contact matrix into partially overlapping matrices of size 2 Mb by 2 Mb along the diagonal line with overlapping areas of 1 Mb by 1 Mb. We then performed RWR for all 10 Kb bin pairs within each 2 Mb by 2 Mb submatrix along the diagonal to approximate contact probability. Only imputed contact probability in the middle rectangle areas was kept to avoid artifacts near corners.

Step B. Normalization of contact probability using one-dimensional (1D) genomic distance

All bin pairs were stratified based on the genomic distance between two bins to account for the dependency between imputed contact probability and 1D genomic distance. Let k=1,2,...,K be the index of K input cells. For a bin pair (i, j) in cell k with a genomic distance of d, let $A_d^{(k)}$ represent the strata including bin pairs in cell k with 1D genomic distance d. The normalized contact probability (Z-score) $z_{ij}^{(k)}$ was calculated as $z_{ij}^{(k)} = (x_{ij}^{(k)} - \mu_d^{(k)})/\sigma_d^{(k)}$, where $x_{ij}^{(k)}$ is the contact probability between bin i and bin j in cell k, $\mu_d^{(k)}$ and $\sigma_d^{(k)}$ are mean and standard deviation of the contact probability of bin pairs in cell k within the strata $A_d^{(k)}$.

Step C. Filtering the significant chromatin interactions

First, we define the AND bin pair as both sides overlapping with TSS and genes promoter regions, determined from the users input file or +/-500bp of TSS. Next, we define the XOR bin pair as only one side overlapping with TSS and genes promoter regions while the other side overlaps with enhancer regions. Significant chromatin interactions categorized as AND or XOR are the candidate SnapHiC-G enhancer-promoter interactions. Furthermore, candidate enhancer-promoter interactions with low mappability score ($\leqslant$0.8) or overlapping with the ENCODE blacklist regions (mm10: `http://mitra.stanford.edu/kundaje/akundaje/release/blacklists/mm10-mouse/mm10.blacklist.bed.gz`; hg19: `https://www.encodeproject.org/files/ENCFF001TDO/`) were excluded. HiCNorm was used to calculate each 10 kb bin's sequence mappability [33].

Step D. Detecting interaction candidates

For each bin pair (i, j), we applied a one-sample *t*-test across all K cells to evaluate whether the contact probability is significantly higher than zero. We further converted one sample *t*-test p-values into false discovery rates (FDRs), again stratified by 1D genomic distance. We defined a bin pair as a significant chromatin interaction if its (1) average of *z*-scores across all input cells > 0, (2) the proportion of cells with a *z*-score >1.96 (outlier cells) $>10\%$, (3) FDR$<10\%$, and (4) *t*-test statistic >3.

4.4.2 The computational cost of SnapHiC-G

Adopting a similar parallel computing strategy, SnapHiC-G is highly efficient regarding memory and computational time. We evaluated the computational cost of SnapHiC-G-specific steps (Step C and Step D) with both mouse and human scHi-C data under different settings and summarized the results in Table C2.

4.4.3 Processing scHi-C data

We used the same procedure as in SnapHiC to process the single-cell Hi-C data of mESCs and the human prefrontal cortex sn-m3C-seq data. For mESCs, we aligned the raw read pairs in fastq format to the mm10 genome, removed duplications, and then chose the top 742 cells ($>$150,000 contacts in each cell) for downstream analysis. For the human cortex, we used reference genome hg19 to process the data and then removed duplications. We chose the top 2,869 cells ($>$150,000 contacts in each cell) for downstream analysis and conducted cell type annotation as described in the original work.

4.4.4 Analyzing sn-m3C-seq data from the human cortex

We only analyzed data generated from the same cell type. For sn-m3C-seq data generated from complex tissue samples of heterogeneous cell types, we used cell type annotations reported by the original study.

4.4.5 Down-sampling of scHi-C data

We randomly permuted 742 quality-controlled cells for the mESC scHi-C data to evaluate with down-sampled data. We then performed down-sampling by selecting the first 100 and 300 cells from the pool of 742 cells. As for the astrocytes (338 cells), microglia (323 cells), and oligodendrocytes (1,038 cells) from the human prefrontal cortex sn-m3C-seq data, we permuted the cells and selected the first 261 cells from each cell type to match the number of L2/3 neurons.

4.4.6 Existing methods for bulk Hi-C data

Many methods, including FitHiC2, FastHiC, HiC-ACT, and HiC-DC+ have been developed for identifying long-range chromatin interactions from bulk Hi-C data. Specifically, FitHiC2 is a spline regression-based approach to identify intra-chromosomal chromatin interactions. FastHiC is a Bayesian hidden Markov random field method, which models the spatial dependency structure in high-resolution Hi-C data, for detecting biologically meaningful chromatin interactions (i.e. peak calling). HiC-ACT is an aggregated Cauchy test-based approach that combines p values from other methods (e.g., FitHiC2) without knowing the underlying spatial dependency structure. HiC-DC+ is a negative binomial regression-based method to predict chromatin interactions based on genomic distance, GC content, and mappability features. These bulk Hi-C data-based methods cannot handle the sparsity of the scHi-C data.

4.4.7 Identification of loops/interactions using other Hi-C methods

We identified chromatin loops/interactions with HiCCUPS, FastHiC, FitHiC2, HiC-ACT, and HiC-DC+ from aggregated contact matrix and SnapHiC from the single-cell contact matrix. To apply the methods developed for bulk Hi-C data, we generated pseudo bulk Hi-C data by aggregating 10Kb resolution scHi-C contact matrices across single cells (i.e., sum up the contacts). Then we applied FitHiC2, FastHiC, HiC-ACT, and HiC-DC+ to the pseudo bulk Hi-C data with lenient significance thresholds to consider the sparsity of scHi-C data. We used the following criteria for bulk Hi-C methods to detect significant interactions: FDR $<$10% from FitHiC2;

posterior probability $>$0.9 from FastHiC; local neighborhood smoothed p-values $< 10^{-6}$ from HiC-ACT; FDR $<$ 10% from HiC-DC+. For SnapHiC, we used FDR $<$10% and t-statistics $>$3 to select significant interactions. We further required interactions within the 20Kb-1Mb 1D genomic distance, with high mappability ($>$0.8) and no overlapping with ENCODE blacklist regions for standard quality control. To fairly compare with SnapHiC-G-identified interactions, we required interactions to be ”AND” or ”XOR” as an additional filtering criterion. Note that the number of interactions for SnapHiC was reduced compared with the original SnapHiC paper because of this additional filtering criterion.

4.4.8 SnapHiC-G cell-type-specific interactions

The cell-type-specific enhancer-promoter interactions were a subset of the SnapHiC-G enhancer-promoter interactions detected from down-sampled 261 cells in each of the four cell types: astrocytes, oligodendrocytes, microglia, and L2/3 excitatory neurons. Specifically, an enhancer-promoter interaction detected from a cell type was defined as a cell-type-specific SnapHiC-G enhancer-promoter interaction if none of the other three cell types' identified enhancer-promoter interactions overlapped with it.

4.4.9 Enrichment of eQTL-TSS pairs

We evaluated whether SnapHiC-G-identified interactions are enriched in eQTL-TSS bin pairs in brain cell types, CommonMind brain eQTL-TSS bin pairs, and GTEx consortium liver eQTL-TSS bin pairs for all the cell types we considered. Specifically, for the eQTL-TSS bin pairs in brain cell types, we used eQTLs with Bonferroni corrected p-value $<$ 0.05 within each cell type. Because the eQTL data did not have results in L2/3 neurons, we used eQTL results of excitatory neurons to compare with SnapHiC-G-identified bin pairs in L2/3 neurons. For each SnapHiC-G-identified interaction, we constructed a matched pseudo bin pair as a control: if only one bin contains the promoter, we kept the bin with the promoter as the center and flipped the other bin to be on the opposite side of the center but with the same distance from the center; if both bins contain promoters, we randomly selected one bin with probability 0.5 and kept it as the center, and then flipped the other bin to the other side of the center. Next, we removed the duplicates between controls and SnapHiC-G-identified bin pairs. We constructed a two-by-two table for the union of SnapHiC-G-identified interactions and pseudo bin pairs, categorizing each bin pair by whether it is a SnapHiC-G-identified interaction or an eQTL-TSS bin pair. The p-value for independence between these two features was calculated using a two-sided Fishers exact test.

4.4.10 Gene expression analysis at cell-type-specific interactions

The fragments per kilobase of transcript per million mapped reads (FPKM) values of each protein-coding gene in human astrocytes, neurons, microglia, and oligodendrocytes were acquired from Zhang et al. [103]. We used average FPKM values across biological replicates of the same cell type to quantify cell-type-specific gene expression levels.

4.4.11 Selection of GWAS SNPs

First, we gathered significant ($P<5\times10^{-8}$) GWAS SNPs from ten brain-related traits, including Alzheimer's disease, attention deficit hyperactivity disorder, autism spectrum disorder, bipolar disorder, educational attainment, intelligence quotient, neuroticism, Parkinson's disease, schizophrenia, major depressive disorder. Next, we took an overlap between these GWAS SNPs and active enhancers of astrocytes, neurons, microglia, or oligodendrocytes and we had 9,764 SNP-trait associations (8,516 unique GWAS SNPs).

CHAPTER 5: LD SCORE REGRESSION WITH MORE COMPREHENSIVE CATALOGS OF LD SCORES

As discussed in Chapter 1, LDSC is a widely used statistical method to estimate the degree of inflation in GWAS test statistics that can distinguish between inflation from true polygenic signals and bias. Public LD scores are available for $\sim$1.3 million HapMap3 (HM3) variants for European and East Asian samples from the 1000 Genomes Project (1000G) [127, 128]; however, these variants cove only a subset of common variants (there are $\sim$5-7 million variants with MAF>0.05) and few low-frequency or rare variants (LFRV). In this chapter, we aim to investigate how LDSC estimates change when including additional variants and using a more comprehensive LD reference panel built from whole genome sequencing data from the NHLBI Trans-Omics for Precision Medicine (TOPMed) Program [129].

5.1 Calculation of TOPMed LD scores

First, we computed LD scores using LD calculated from the TOPMed freeze 8 reference panel using unrelated European samples (RFMix score $>$ 0.9, kinship score $< 2^{-5.5}$, n = 13,160 from MESA, BioMe, and WHI cohorts [130]). LD values were computed as the squared Pearson correlation (r^2) between pairs of variants on a phased haplotype basis. The LD score for variant i was computed as $LDSC_i = \sum r_{ij}^2$ for $r_{ij}^2 > 1e^{-7}$ for all variants j within 1Mb of variant i.

5.2 GWAS summary statistics

We collected GWAS summary statistics for 36 phenotypes, including BMI, height, brain-related traits and diseases, and 20 blood cell traits [131, 132, 110, 133, 134, 109, 135, 113, 122, 106, 107, 136, 137, 123]. All GWASes were conducted in European-ancestry individuals except for blood cell traits being a trans-ethnic study including a small portion of individuals from African American and East Asian ancestries.

5.3 Preliminary results

We ran LDSC with different LD reference panels and sets of variants included in the regression and performed multiple comparisons to evaluate their impact on the estimation of LD scores and LDSC intercept. Specifically, we were interested in addressing the following questions: (i) How does the choice of LD reference panel affect LD score estimation itself? (ii) What is the effect of using LD scores calculated from different LD reference panels in LDSC intercept estimation? (iii) Hoes does the choice of variants included in the regression affect LDSC intercept estimation? (iv) Is there a difference in LDSC intercept estimation when we include common vs. low-frequency and rare variants in the regression?

To address the first question, we compared LD scores calculated from 1000G and TOPMed for HapMap3 variants. The 1000G LD scores are publicly available for HapMap3 variants where 378 European-ancestry individuals were included, and the TOPMed LD scores were computed with 13,160 European-ancestry individuals as previously described. From Figure 1, we observed that 1000G and TOPMed LD scores for HapMap3 variants are similar in general, especially for variants with relatively low or moderate LD scores. For variants with high LD scores (e.g., $>$400), the LD score estimation tends to be higher with the 1000G reference panel.

To investigate the effect of the LD reference panel in LDSC intercept estimation, we ran LDSC with 1000G and TOPMed LD scores while including only HapMap3 variants ($\sim$1.3 million variants) in the regression and calculated regression weights accordingly following the original LDSC method. Each variant was weighted by the reciprocal of its LD Score, counting LD only with other SNPs included in the regression. Figure 2 shows LDSC intercept estimates with LD scores calculated from 1000G and TOPMed were highly consistent for the 36 GWAS summary statistics, which was expected because of the similar LD scores calculated from two LD reference panels.

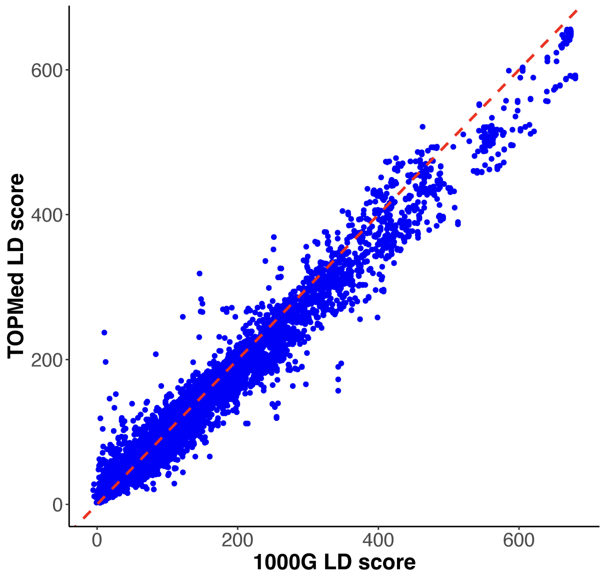


Figure 1. TOPMed vs 1000G LD scores (chr20). Comparision between TOPMed and 1000G LD scores for HapMap3 variants. Each dot represents a variant. The red dashed line represents the diagonal.

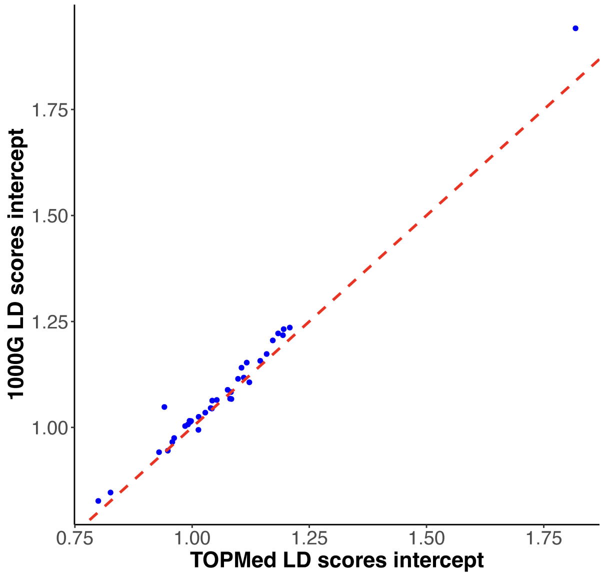


Figure 2. LDSC intercept estimates with 1000G and TOPMed LD scores(chr20). Only HapMap3 variants were included in the regression. Each dot represents a GWAS. The red dashed line represents the diagonal.

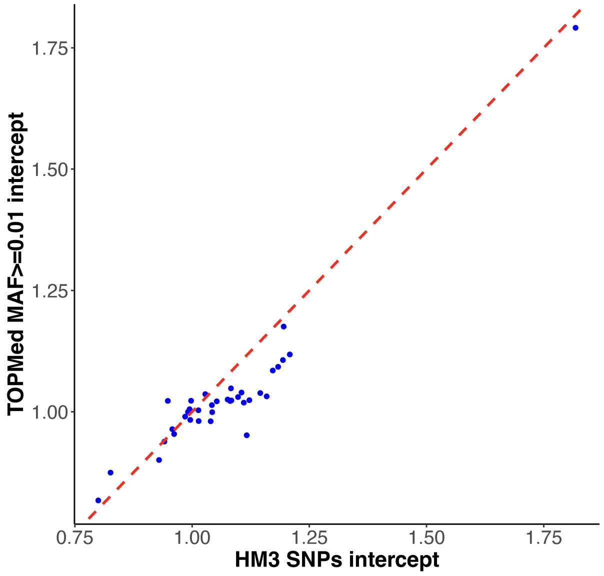


Figure 3. LDSC intercept estimates with TOPMed LD scores for MAF⩾0.01 and HapMap3 variants (chr20). LD scores were calculated using the TOPMed reference panel. Different sets of variants were included in the regression for comparison. Each dot represents a GWAS. The red dashed line represents the diagonal.

Next, we compared LDSC intercept estimates when including HapMap3 variants vs. including TOPMed variants with MAF>0.01 in the regression, using TOPMed LD scores. Results showed that using TOPMed variants with MAF⩾0.01 gave smaller LDSC intercept estimates for most traits than including only HapMap3 variants in the regression (Figure 3), suggesting a non-negligible impact of including additional variants in LD scores on LDSC estimates.

Since HapMap3 covers only a subset of common variants in Europeans and few LFRV, we further investigated the impact of including LFRV on LDSC intercept estimates. TOPMed SNPs were partitioned by MAF with a 0.01 cutoff in Europeans, and then LDSC was run with variants for the MAF<0.01 and MAF⩾0.01 sets separately. We focused on 20 blood cell traits that had a large number of LFRV available in the GWAS summary statistics. From Figure 4, we can see that using TOPMed LD scores with MAF<0.01 gave smaller estimates for most blood cell traits than including only common variants.

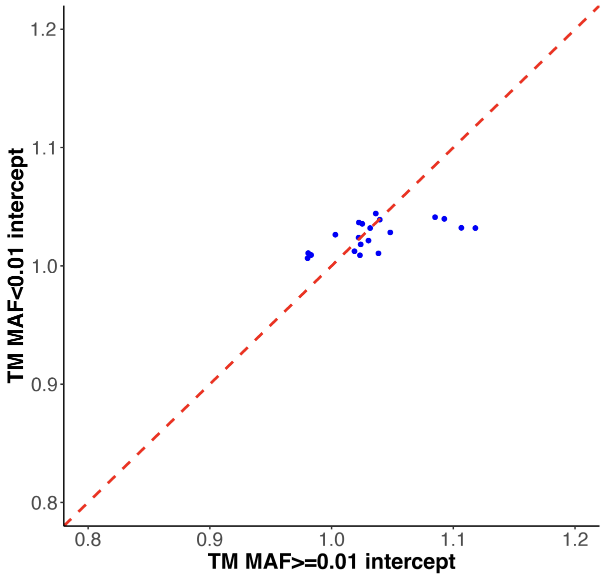


Figure 4. LDSC estimates with TOPMed LD scores with MAF$<$0.01 and MAF$\geqslant$0.01 variants (chr20). LD scores were calculated using the TOPMed reference panel. Common and LFRV variants were included in the regression for comparison. Each dot represents a blood cell trait. The red dashed line represents the diagonal.

5.4 Discussion

In GWAS, it is essential to adjust for the inflation in test statistics resulting from confounding biases, such as population stratification and cryptic relatedness to false positives. On the other hand, when including more common variants in the regression, LDSC intercepts decrease and better account for polygenicity, avoiding missing potential GWAS signals by the overcorrection of test statistics. In other words, true GWAS signals that might have been overcorrected and missed after the LDSC intercept correction when using only HapMap3 variants can be rescued by incorporating additional common variants in the regression.

Results suggest that LFRV have less inflation compared with common variants, which can be explained by a couple of reasons. First, for rare variants, GWAS test statistics are skewed towards the null because of the low discovery power. Second, as the number of LFRV is greater than common variants, by including more variants, the model is able to capture more polygenic signals. If low-frequency variants are indeed different from common variants, care needs to be taken when analyzing common and rare variants jointly. One option is to separate the two sets of variants in the analysis to avoid overcorrection of LFRV variants and undercorrection of common variants.

The authors of LDSC claimed that LD scores calculated from HapMap3 variants tended to give LDSC intercepts that were closer to 1 in studies believed to have little confounding from population stratification than LD scores calculated over all 1000G variants. One reason could be that 1000G LD scores give too much weight to LFRV, and HapMap3 LD scores give more weight to common variants. However, further analysis is needed to investigate which variant sets should be included in LD score calculation and LD score regression. Other future directions include conducting the analyses with the East Asian ancestry population, where public LD scores are also available. Besides LDSC intercepts, it would be interesting to look at heritability and genetic correlation estimates as well.

Finally, and perhaps most importantly, as the public LD scores are released for HapMap3 variants only, the LD scores we computed using TOPMed freeze 8 as the reference panel for both common variants and LFRV will be a valuable resource for the research community to investigate a broad range of important scientific questions.

APPENDIX A: SUPPLEMENTARY MATERIALS FOR CHAPTER 2

Details of inferred expected contacts from FitHiC

FitHiC models the expected contact count conditioned on observed contact counts and the genomic distance between interacting regions [39]. Specifically, all locus pairs with a non-zero contact count are sorted with respect to increasing genomic distance between the two ends of the pair and this sorted list is broken into *b* bins. Then, the total number of contact counts is divided into the *b* bins using an equal occupancy binning strategy. Thus, all bins have an approximately equal number of contacts. In each bin, the average genomic distance (x-axis) and contact probability (y-axis) are computed among all pairs (including possible pairs with zero contact counts) in this bin. Then, FitHiC fits a cubic smoothing spline (third-degree polynomial) to these x and y values (one per bin) to learn a continuous function that relates these two entities. The inferred expected contact probabilities are further corrected through the bias values, which are computed per locus/bin by the KR normalization method [138].

Details of statistical inference

Based on the Bayes rule, we have the joint posterior distribution as follows:

$$\begin{gathered}P\left(\{z_{ijk}\},\{\theta_k\},\{\phi_k\},\{\psi_k\},\{\gamma_k\}|\{x_{ijk}\},\{e_{ijk}\}\right)\propto P\left(\{x_{ijk}\}|\{e_{ijk}\},\{z_{ijk}\},\{\theta_k\},\{\phi_k\}\right)*\\ Prior\left(\{z_{ijk}\}|\{\psi_k\},\{\gamma_k\},\alpha\right)*Prior\left(\{\theta_k\}\right)*Prior\left(\{\phi_k\}\right)*Prior\left(\{\psi_k\}\right)*Prior\left(\{\gamma_k\}\right).\end{gathered}$$

Note that we used uniform prior distributions for $\theta_k, \phi_k, \psi_k, \gamma_k$, which were initialized from estimates from uni-sample analysis in our implementation (see section below). Let $NB(x|\mu, \phi)$ represent the probability mass function of negative binomial distribution with mean μ and over-dispersion ϕ, then

$$P\left(\{x_{ijk}\}|\{e_{ijk}\},\{z_{ijk}\},\{\theta_k\},\{\phi_k\}\right)=\prod_{k=1}^{K}\prod_{1\leqslant i<j\leqslant N}NB\left(x_{ijk}|e_{ijk}\exp\{I\left(z_{ijk}=1\right)\theta_k\},\phi_k\right).$$

In addition, we use pseudo-likelihood approximation to calculate the hierarchical Ising prior where the joint density is approximated by the product of conditional densities:

$$P\left(\{z_{ijk}\}|\{\psi_k\},\{\gamma_k\},\alpha\right) \approx \prod_{k=1}^{K}\prod_{1\leqslant i<j\leqslant N} P\left(z_{ijk}|\{z_{-(ijk)}\},\psi_k,\gamma_k,\alpha\right).$$

Here

$$\begin{aligned}
P\left(z_{ijk}|\{z_{-(ijk)}\},\psi_k,\gamma_k,\alpha\right) &\propto P\left(z_{ijk}|\{z_{-i,-j,k}\},\psi_k,\gamma_k\right) * P\left(z_{ijk}|\{z_{ij,-k}\},\alpha\right) \\
&\propto \exp\left\{\gamma_k I(z_{ijk}=1) + \psi_k * z_{ijk} * \sum_{\left|i'-i\right|+\left|j'-j\right|=1} z_{i'j'k}\right\} \\
&\quad * \frac{\alpha_{\{Z_{ij1},\ldots,Z_{ijk},\ldots,Z_{ijK}\}}}{\alpha_{\{Z_{ij1},\ldots,Z_{ijk},\ldots,Z_{ijK}\}} + \alpha_{\{Z_{ij1},\ldots,-Z_{ijk},\ldots,Z_{ijK}\}}} \\
&\propto \exp\left\{\gamma_k I(z_{ijk}=1) + \psi_k * z_{ijk} * \sum_{|i'-i|+|j'-j|=1} z_{i'j'k}\right\} \\
&\quad * \alpha_{\left\{z_{ij1},\ldots z_{ijk},\ldots,z_{ijK}\right\}}.
\end{aligned}$$

Where $\{z_{-(ijk)}\}$ denotes the set $\{z_{i'j'k'}|i^{'} \neq i, j^{'} \neq j, k' \neq k\}$, $\{z_{-i,-j,k}\}$ denotes the set $\{z_{i'j'k}|i^{'} \neq i, j^{'} \neq j\}$, and $\{z_{ij,-k}\}$ denotes the set $\{z_{ijk'}|k' \neq k\}$.

Define $A\left(z_{ijk}\right) = \exp\left\{\gamma_k I(z_{ijk}=1) + \psi_k * z_{ijk} * \sum_{|i'-i|+|j'-j|=1} z_{i'j'k}\right\} * \alpha_{\left\{z_{ij1},\ldots z_{ijk},\ldots,z_{ijK}\right\}}$, we have $P\left(z_{ijk}|\{z_{-(ijk)}\},\psi_k,\gamma_k,\alpha\right) = \frac{A\left(z_{ijk}\right)}{A\left(z_{ijk}\right)+A\left(-z_{ijk}\right)} = \frac{1}{1+A\left(-z_{ijk}\right)/A\left(z_{ijk}\right)}$. Taken together,

$$P\left(\{z_{ijk}\}|\{z_{-(ijk)}\},\{\psi_k\},\{\gamma_k\},\alpha\right) = \prod_{k=1}^{K}\prod_{1\leqslant i<j\leqslant N} \frac{1}{1+A\left(-z_{ijk}\right)/A\left(z_{ijk}\right)}.$$

The log of the joint posterior distribution is approximated as follows:

$$\begin{aligned}
&\log P\left(\{z_{ijk}\},\{\theta_k\},\{\phi_k\},\{\psi_k\},\{\gamma_k\}|\{x_{ijk}\},\{e_{ijk}\}\right. \approx Constant \\
&+\sum_{k=1}^{K}\sum_{1\leqslant i<j\leqslant N}\log NB\left(x_{ijk}|e_{ijk}\exp\left\{\frac{z_{ijk}+1}{2}\theta_k\right\},\phi_k\right) \\
&+\log P\left(\{z_{ijk}\}|\{z_{-(ijk)}\},\{\psi_k\},\{\gamma_k\},\alpha\right) \\
&= Constant+\sum_{k=1}^{K}\sum_{1\leqslant i<j\leqslant N}\{\log\Gamma\left(x_{ijk}+\phi_k\right)-\log\Gamma\left(\phi_k\right) \\
&+x_{ijk}\left(\log e_{ijk}+\theta_k\frac{z_{ijk}+1}{2}\right) \\
&+\phi_k\log\phi_k-\left(x_{ijk}+\phi_k\right)\log\left(e_{ijk}\exp\left\{\theta_k\frac{z_{ijk}+1}{2}\right\}+\phi_k\right) \\
&-\sum_{k=1}^{K}\sum_{1\leqslant i<j\leqslant N}\log\left\{1+\frac{A\left(-z_{ijk}\right)}{A\left(z_{ijk}\right)}\right\}
\end{aligned}$$

In the equation above,

$$\begin{aligned}
\frac{A\left(-z_{ijk}\right)}{A\left(z_{ijk}\right)} &= \frac{\exp\left\{-\gamma_k z_{ijk}-\psi_k*z_{ijk}*\sum_{\left|i'-i\right|+\left|j'-j\right|=1}z_{i'j'k}\right\}*\alpha_{\left\{z_{ij1},\ldots-z_{ijk},\ldots,z_{ijK}\right\}}}{\exp\left\{\gamma_k z_{ijk}+\psi_k*z_{ijk}*\sum_{\left|i'-i\right|+\left|j'-j\right|=1}z_{i'j'k}\right\}*\alpha_{\left\{z_{ij1},\ldots z_{ijk},\ldots,z_{ijK}\right\}}} \\
&= \exp\left\{-2\gamma_k z_{ijk}-2\psi_k*z_{ijk}*\sum_{|i'-i|+|j'-j|=1}z_{i'j'k}\right\}*\frac{\alpha_{\left\{z_{ij1},\ldots-z_{ijk},\ldots,z_{ijK}\right\}}}{\alpha_{\left\{z_{ij1},\ldots z_{ijk},\ldots,z_{ijK}\right\}}}.
\end{aligned}$$

In the Gibbs sampler, the conditional distribution of z_{ijk} follows a Bernoulli distribution. We have

$$\begin{aligned}
\log P\left(z_{ijk}=1|\theta_k,\phi_k,\psi_k,\gamma_k,x_{ijk},e_{ijk}\right) &= x_{ijk}\theta_k-\left(x_{ijk}+\phi_k\right)\log\left(e_{ijk}\exp\left\{\theta_k\right\}+\phi_k\right) \\
-\log\left\{1+\exp\left(-2\gamma_k-2\psi_k\sum_{\left|i'-i\right|+\left|j'-j\right|=1}z_{i'j'k}\right)\frac{\alpha_{\left\{z_{ij1},\ldots-1,\ldots,z_{ijK}\right\}}}{\alpha_{\left\{z_{ij1},\ldots1,\ldots,z_{ijK}\right\}}}\right\}&; \\
\log P\left(z_{ijk}=-1|\theta_k,\phi_k,\psi_k,\gamma_k,x_{ijk},e_{ijk}\right) &= -\left(x_{ijk}+\phi_k\right)\log\left(e_{ijk}+\phi_k\right) \\
-\log\left\{1+\exp\left(2\gamma_k+2\psi_k\sum_{\left|i'-i\right|+\left|j'-j\right|=1}z_{i'j'k}\right)\frac{\alpha_{\left\{z_{ij1},\ldots1,\ldots,z_{ijK}\right\}}}{\alpha_{\left\{z_{ij1},\ldots-1,\ldots,z_{ijK}\right\}}}\right\}&.
\end{aligned}$$

Consider a special case where $K = 2$, we hope to calculate the probability mass function for $P(z_{ij1}, z_{ij2} | \{\theta_k\}, \{\phi_k\}, \{\psi_k\}, \{\gamma_k\}, \{x_{ijk}\}, \{e_{ijk}\})$. For a fixed (i, j) pair, we have

$$\begin{aligned}
&\log P\left(z_{ij1}, z_{ij2} \middle| \{z_{-(ij1,ij2)}\}, \{\theta_k\}, \{\phi_k\}, \{\psi_k\}, \{\gamma_k\}, \{x_{ijk}\}, \{e_{ijk}\}\right) \\
&\qquad = Constant + \sum_{k=1}^{2} \{\log\Gamma\left(x_{ijk} + \phi_k\right) - \log\Gamma\left(\phi_k\right) \\
&\qquad + x_{ijk}\left(\log e_{ijk} + \theta_k \frac{z_{ijk}+1}{2}\right) + \phi_k \log\phi_k \\
&\qquad - \left(x_{ijk} + \phi_k\right)\log\left(e_{ijk}\exp\left\{\theta_k\frac{z_{ijk}+1}{2}\right\} + \phi_k\right)\Bigg\} - \sum_{k=1}^{2}\log\left\{1 + \frac{A\left(-z_{ijk}\right)}{A\left(z_{ijk}\right)}\right\}
\end{aligned}$$

and denote it as $B(z_{ij1}, z_{ij2})$.

Therefore,

$$P(z_{ij1} = 1, z_{ij2} = 1 | \{\theta_k\}, \{\phi_k\}, \{\psi_k\}, \{\gamma_k\}, \{x_{ijk}\}, \{e_{ijk}\}) \propto \exp\{B(1,1)\}$$

Considering all 4 possibilities, we have

$$P\left(z_{ij1} = 1, z_{ij2} = 1 | \{\theta_k\}, \{\phi_k\}, \{\psi_k\}, \{\gamma_k\}, \{x_{ijk}\}, \{e_{ijk}\}\right) = \frac{\exp\{B(1,1)\}}{\exp\{B(1,1)\} + exp\{B(1,-1)\} + \exp\{B(-1,1)\} + \exp\{B(-1,-1)\}}$$

Similarly, we can calculate the conditional probabilities for the other three configurations for $z_{ij1}$and z_{ij2}. We use the Gibbs sampler to update all the other parameters (θ_k, ϕ_k, ψ_k, γ_k). The hyper-parameter α can be estimated from empirical data.

Implementation details of MUNIn and uni-sample analysis

Uni-sample analysis was implemented following Xu *et al.* [41], where the initial peak status was randomly assigned. For MUNIn analysis, both peak status and parameters of each cell type, i.e., (θ_k, ϕ_k, ψ_k, γ_k), were initialized according to results from uni-sample analysis. Specifically, we searched within the range of +20% and -20% of estimates from uni-sample analysis for θ_k, ϕ_k, ψ_k, γ_k, which was equivalent to uniform priors. The across-sample dependency parameter α was estimated based on uni-sample inference. Then, the peak status and parameters were updated following the procedures described above, and 10,000 Gibbs sampling steps were performed.

Real data evaluation framework

To avoid false positive calls at a close distance, we further filtered the MUNIn-called peaks by excluding the bin pairs less than 50 kb apart. ROC curve was applied to illustrate the performance of MUNIn and uni-sample analysis, to help avoid the bias of unbalanced peak calling number between the two methods. For the ROC curve, we excluded bin pairs with corresponding posterior probabilities less than the 10^{th} percentile, since including bin pairs with very low posterior probabilities is not meaningful. We partitioned the TADs within each chromosome into three categories (shared, GM-specific, and IMR-specific peaks) according to their prior probabilities of four types of peak status (shared peaks, GM-specific peaks, IMR-specific peaks, and shared backgrounds). If the prior probability is highest in shared peaks or cell-type-specific peaks, then we assigned this TAD to the corresponding category. If the prior probability is highest in shared backgrounds, then we assigned this TAD to the category where its prior probability is the second highest. Each ROC curve was then plotted only including TADs in that category.

Additional performance evaluation

In this study, we further evaluated several other performance aspects of MUNIn. In our simulation, with loss of generality, we only considered the case that $\gamma_k = 0$, where the proportion of peak and background is the same. To assess the robustness of MUNIn to the simulation parameters, we simulated data with different γ_k values, -0.02, -0.05, -0.2, and -0.4. In all four scenarios, we observed that MUNIn achieved a lower error rate in peak calling in all three simulated samples than uni-sample analysis when there is moderate or high dependency among samples (**Figure A5**). Our results demonstrated that MUNIn is robust to the simulation parameters.

In addition, we evaluated the scalability of MUNIn for a moderate sample size. We performed a simulation study with five samples. For all five samples, the overall error rate for the peaks identified by MUNIn is substantially lower than that of uni-sample analysis when there is moderate or high dependency among samples (**Figure A6**).

To assess the robustness of MUNIn to the TAD boundaries, we re-ran the real data analysis for Hi-C data from GM12878 and IMR90 cell lines [9] at 10kb resolution using a sliding window approach, instead of focusing on shared TADs. Specifically, we divided the genome into 1Mb windows (core region) with 200kb flanking regions on each side of the window and applied both MUNIn and uni-sample analysis to call interactions within each window. Similar to the results focusing on shared TADs, we observed that MUNIn obtained more accurate results for both GM12878 and IMR90-specific peaks (**Figure 5a and d**), while its performance in shared peaks was comparable to uni-sample analysis (**Figure A13**).

We then examined the overlapping between the cell-type-specific interactions and promoters and enhancers [139]. We identified 535,908 and 674,194 GM12878- and IMR90-specific interactions genome-wide. 109,406 and 41,178 (20.4 and 6.1%) of them overlapped with promoter regions, which are significantly higher than those of all bin pairs (p-value < 2.2e-16 for both samples). Similarly, 71,978 and 48,485 GM12878- and IMR90-specific interactions (13.4 and 7.2%) overlapped with enhancer regions, which are also significantly higher than the genome background (p-value < 2.2e-16 for both samples). These results indicate that the overlap between the significant interactions and promotor/enhancer region is not a coincidence.

To further compare MUNIn with uni-sample analysis, we applied MUNIn and uni-sample analysis on Hi-C datasets of mouse embryonic stem cells for both wild-type and after CTCF deletion at 10kb resolution, and compared the peaks identified by uni-sample analysis and MUNIn with those identified by HiCCUPS [9]. First, we overlapped bin pairs detected from uni-sample analysis and MUNIn with the union of HiCCUPS loops from the wild-type and the CTCF-depleted sample. Then, we looked at wild-type-specific peaks identified by uni-sample analysis or by MUNIn among those overlapping bin pairs. A bin pair was called a wild-type-specific peak in the uni-sample analysis if it was called a peak in the wild-type sample and called a background in the CTCF-depleted sample. Similarly, a bin pair is called a wild-type-specific peak in Munin if the configuration of being a peak in the wild-type sample while being a non-peak in the CTCF-depleted sample has the highest posterior probability. We ran both methods on chromosome 1 and found that among the overlapping bin pairs, uni-sample analysis called 17 wild-type-specific peaks, while MUNIn called 10 wild-type-specific peaks. We used HiCCUPS wild-type-specific loops as the ground truth and defined HiCCUPS wild-type-specific loops by first taking loops called by HiCCUPS in the wild-type mESC sample, and then excluding those that were also called loops in the CTCF-depleted mESC sample by HiCCUPS. The uni-sample analysis had one false positive out of the 17 wild-type specific peaks, while all the ten wild-type-specific peaks called by MUNIn were also identified by HiCCUPS. **Figure A14** shows aggregate peaks plots for wild-type-specific peaks identified by uni-sample analysis and MUNIn. We can see that MUNIn better captured the wild-type-specific pattern in mESC Hi-C data.

We used the same mESC HiC data [67] to further compare MUNIn with FitHiC [39]. Specifically, we performed FitHiC peak calling on the wild-type data (i.e., without CTCF depletion). We performed MUNIn peak calling by jointly analyzing Hi-C data before and after CTCF depletion (by inputting them as two samples into MUNIn) and focused only on peaks called in the wild-type sample (regardless of the status in the sample after CTCF deletion). We therefore were able to compare MUNIn and FitHiC peak calls in the wild-type sample. Specifically, we took the same number of top peaks from each method and compared the percent overlapping with HiCCUPS loops (treated as the truth). For MUNIn, we first identified peaks by their inferred peak status, and then ranked them by the posterior probability of being a peak in the wild-type sample from largest to smallest; for FitHiC, we ranked the peaks by their FitHiC p-values from smallest to largest. We compared the top 1,000 to 5,000 peaks called by each method and found that MUNIn had a higher number of overlaps with HiCCUPS loops than FitHiC (**Figure A15**). We chose this example and focus on the wild-type sample because HiCCUPS on the previously generated GM12878 wild-type data is believed to rather accurately reflect wild-type peaks. That our MUNIn results showing better performance than FitHiC demonstrates the power of MUNIn to more powerfully reveal peaks by borrowing information from another sample. In this case, it is particularly interesting because our results suggest that we attain better power detecting wild-type peaks in the wild-type sample even by borrowing information from the CTCF-depleted sample.

Finally, we estimated the computational time of MUNIn. Uni-sample analysis was first implemented on four shared TADs of different sizes from GM12878 and IMR90 cell lines, which contain 50, 100, 150, and 200 bins, respectively, and then MUNIn was performed based on uni-sample analysis results. The running time of uni-sample analysis and MUNIn was summed as the total computational time of MUNIn. For each TAD, the procedure was executed 10 times. The results showed that MUNIn takes ∼2 and 31 minutes to perform peak calling in a TAD consisting of 50 and 200 bins, respectively (**Figure A16**). We also assess the computational time of MUNIn at different resolutions. MUNIn was implemented on four 2 MB TADs of 10, 20, and 40 kb resolution ten times. The results showed that MUNIn takes only ∼5 minutes for a TAD of 40 kb resolution, and ∼36 minutes for a TAD of 10 kb resolution (**Figure A17**).

Table A1 Major characteristics of the benchmarking datasets.

Dataset	**Cell type**	**Data type**	**GEO accession number/-Download URLs**	**Reference**
Simulations	-	Source codes	https://github.com/yycunc/MUNIn	
Dixon *et al.* (2015)	Human embryonic stem cells	Dilute Hi-C	GSE52457	[64]
Rao *et al.* (2014)	GM12878 IMR90	In situ Hi-C	GSE63525	[9]
Jung *et al.* (2019)	GM12878 IMR90	Promoter-capture Hi-C	GSE86189	[66]
Kubo *et al.* (2021)	Mouse embryonic stem cells	In situ Hi-C	GSE94452	[67]

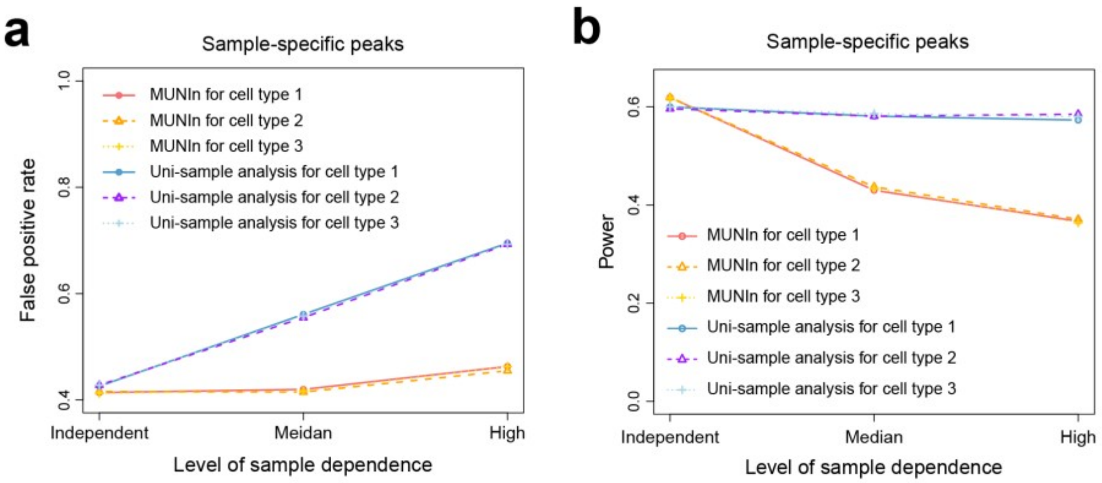


Figure A1. Performance comparison between MUNIn and uni-sample analysis in the simulation data where all three samples have equal sequencing depth. (**a**) False positive rate for the sample-specific peaks identified in each sample using MUNIn and uni-sample analysis. (**b**) Power for the sample-specific peaks identified by MUNIn and uni-sample analysis.

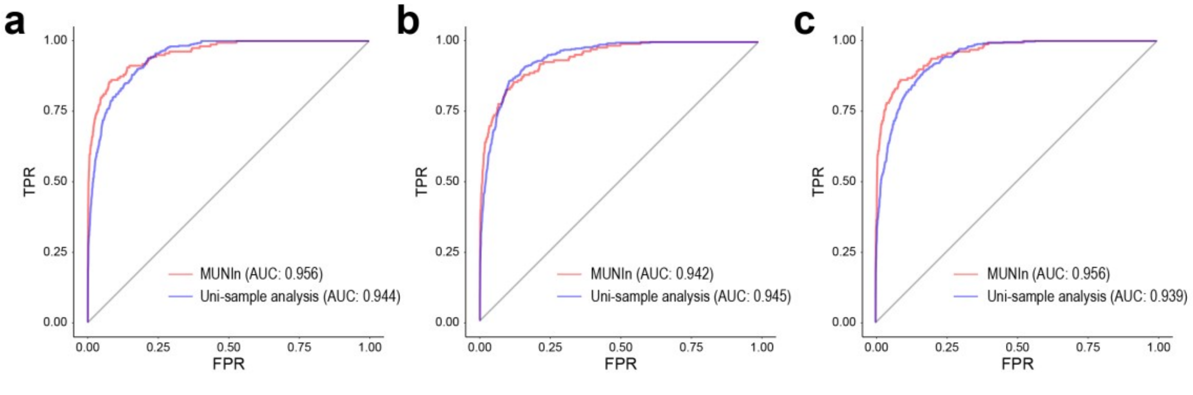


Figure A2. ROC curves for MUNIn and uni-sample analysis in the simulation data where all three samples have equal sequencing depth. ROC curves for **(a)** sample 1-, **(b)** 2-, and **(c)** 3-specific peaks identified by MUNIn and uni-sample analysis.

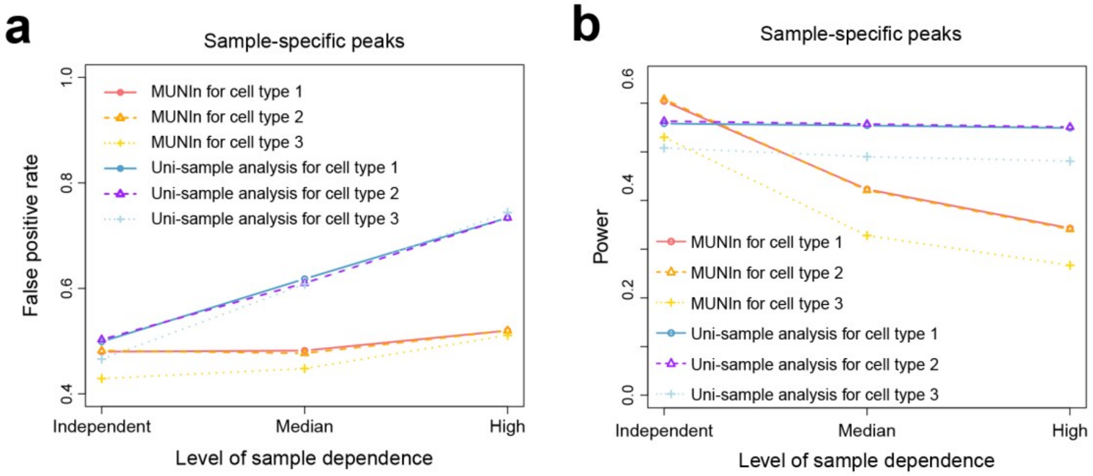


Figure A3. Performance comparison between MUNIn and uni-sample analysis in the simulation data where the three samples have different sequencing depths. (**a**) False positive rates for the sample-specific peaks identified in each sample using MUNIn and uni-sample analysis. (**b**) Power for the sample-specific peaks identified by MUNIn and uni-sample analysis.

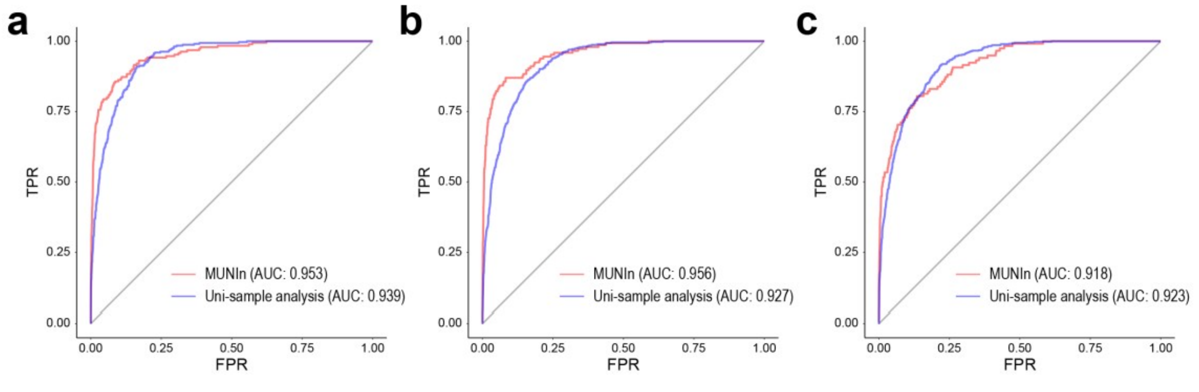


Figure A4. ROC curves for MUNIn and uni-sample analysis in the simulation data where all three samples have different sequencing depths. ROC curves for **(a)** sample 1-, **(b)** 2-, and **(c)** 3-specific peaks identified by MUNIn and uni-sample analysis in the simulated data.

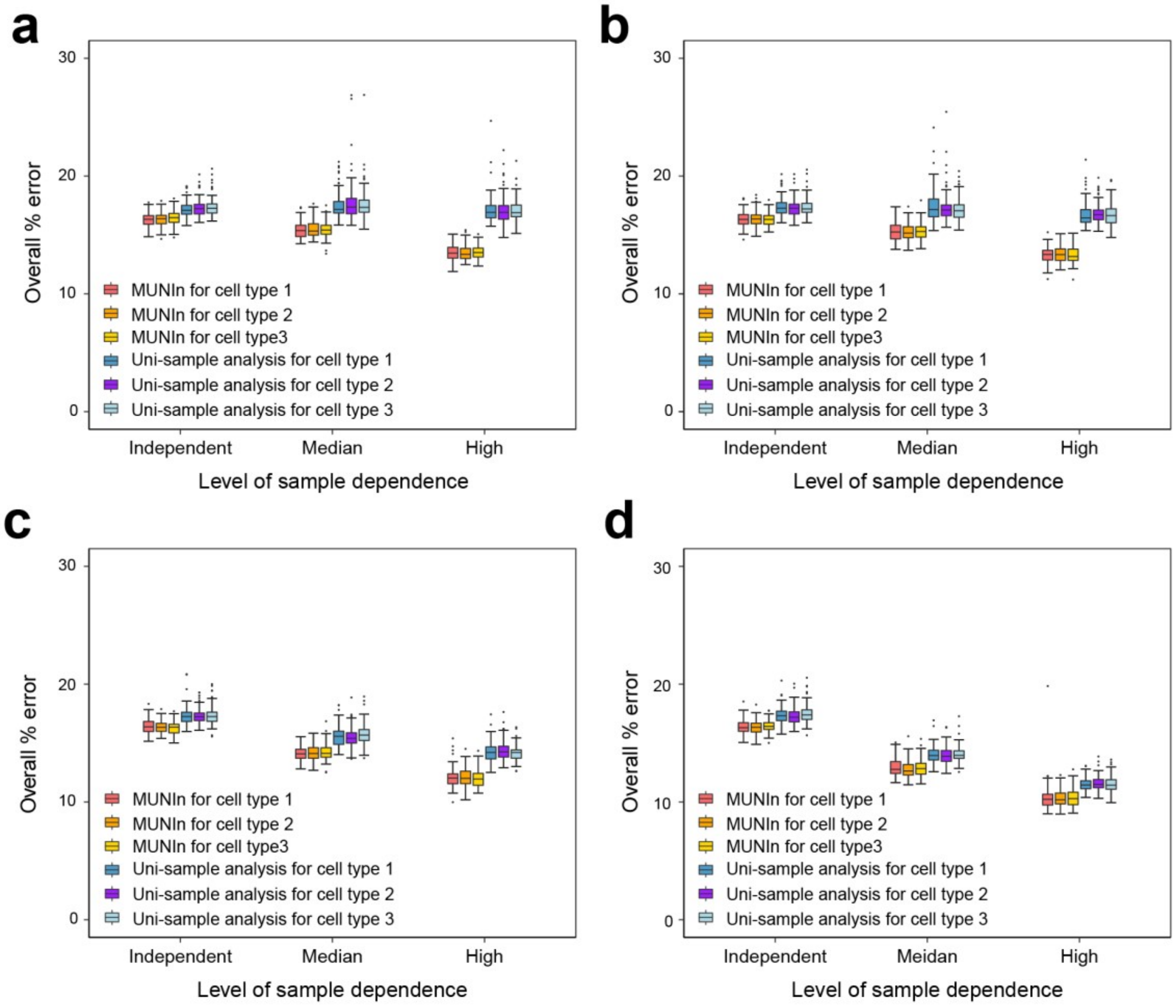


Figure A5. The overall error rate (denoted as “%error”) in peak identification in each sample of the simulations using different γ_k values. (a) $\gamma_k = 0.02$, **(b)** γ_k =-0.05, **(c)** γ_k = -0.2, and **(d)** γ_k = -0.4.

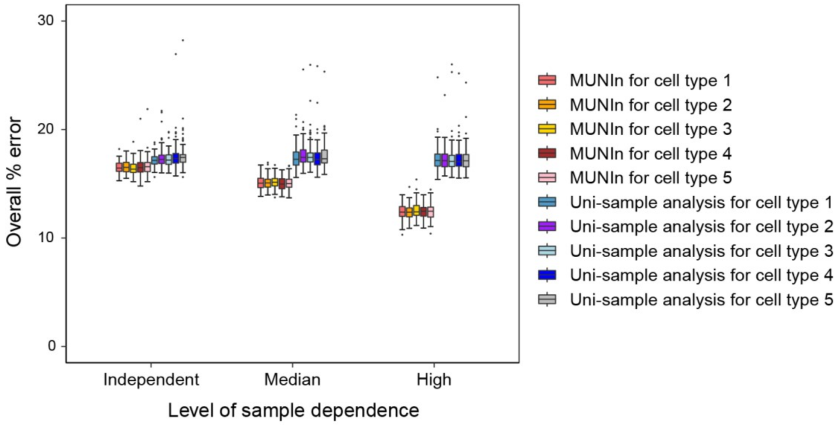


Figure A6. The overall error rate (denoted as "%error") in peak identification. The overall error rate for each of the five simulated samples was computed by performing MUNIn and uni-sample analysis, respectively.

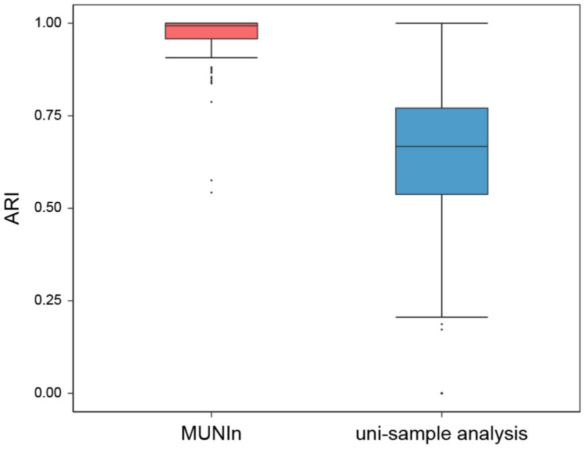


Figure A7. Adjusted Rand Index (ARI) for MUNIn and uni-sample analysis. The ARI shows the level of consistency between the interactions detected in the two biological replicates of human embryonic stem cells.

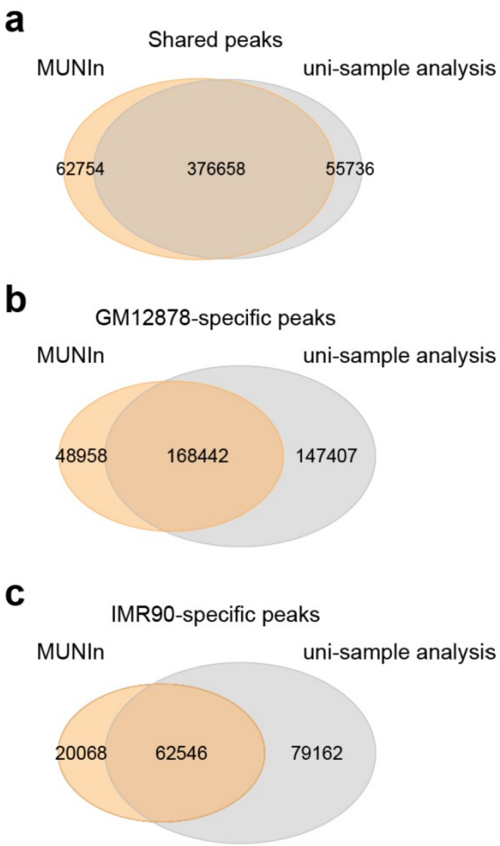


Figure A8. Venn diagram of overlap between MUNIn and uni-sample analysis. The diagram shows the overlap of the **(a)** shared, **(b)** GM12878-specific, and **(c)** IMR90-specific peaks identified by MUNIN and uni-sample analysis.

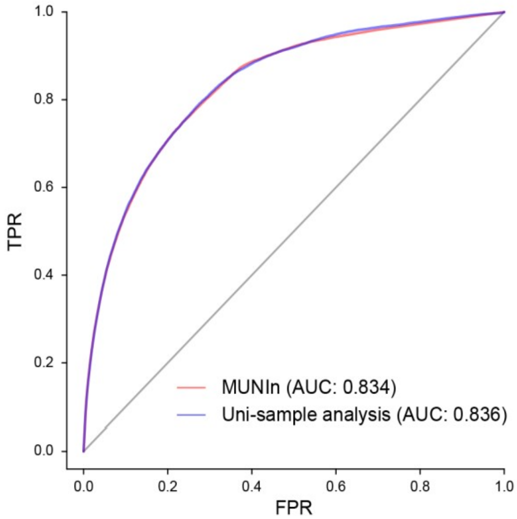


Figure A9. ROC curves for shared peaks. ROC curve for the shared peaks detected in both GM12878 and IMR90 using MUNIn and uni-sample analysis.

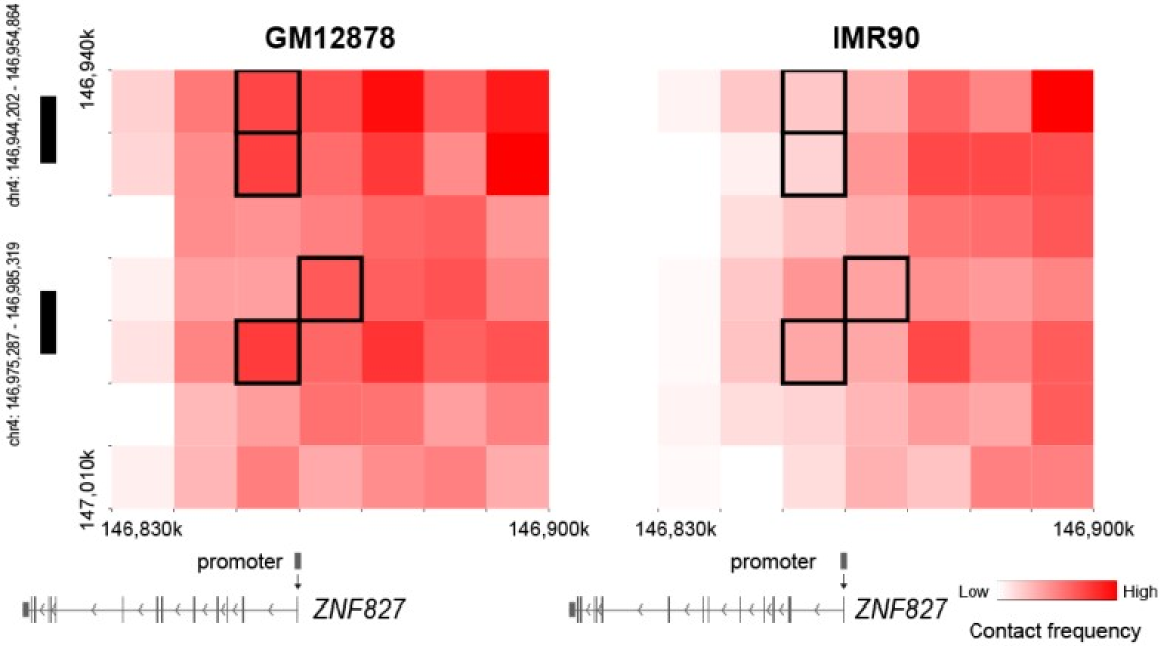


Figure A10. Heatmap showing the GM12878-specific peaks in GM12878 (left) and IMR90 (right) Hi-C data. One bin of these pairs (highlighted in black) is overlapped with the promoter of *ZNF827* gene (transcription start site (TSS) +/- 500bp), while the others are overlapped with known typical enhancers (chr4:146,975,287-146,985,319 and chr4:146,944,202-146,954,864) in GM12878 cells. Gene model is obtained from WashU epigenome browser [140].

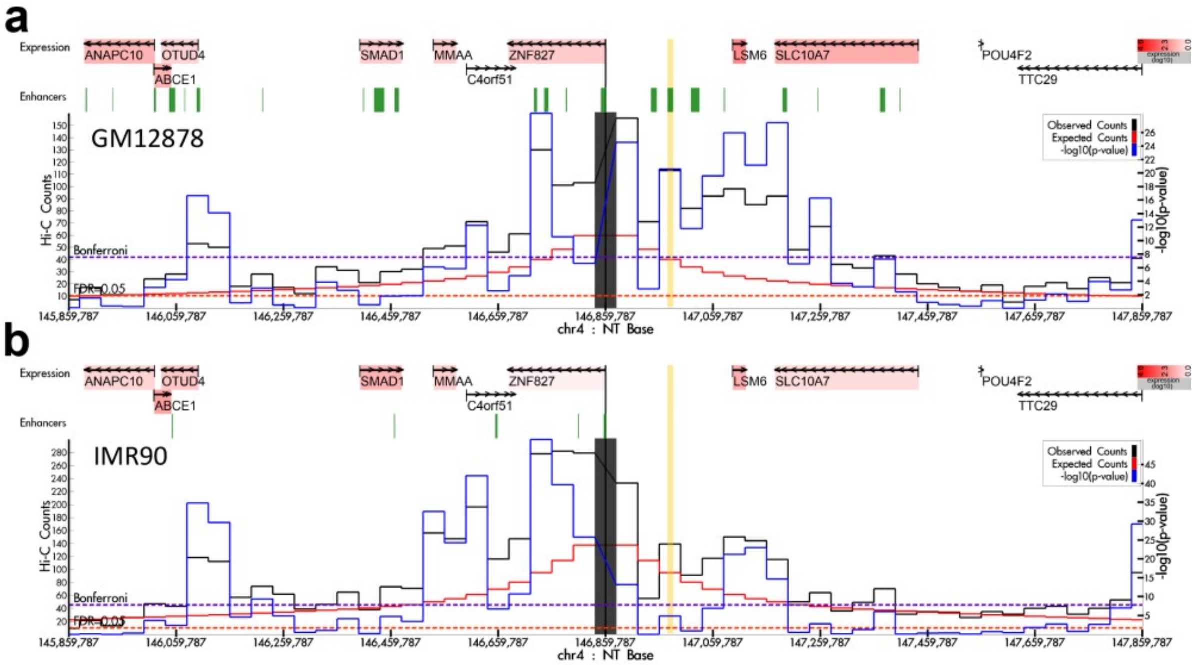


Figure A11. Virtual 4C plot showing one example of the GM12878-specific peaks in (a) GM12878 and (b) IMR90 Hi-C data using HUGIn [13]**.** On the top of each panel, the genes are colored according to their expression level (the deeper red, the higher the expression level) with arrows indicating the direction of transcription and a vertical bar indicating the transcription start site (TSS). On the bottom of each panel, the chromatin interaction from Hi-C data is shown by a virtual 4C plot. The anchor bin overlapped with the promoter region of gene *ZNF827* is indicated as a thick grey vertical bar at the center. The bin overlapped with the GM12878-specific enhancer region is highlighted in yellow. The black line shows the observed counts; the red line shows the expected counts, and the blue line shows the -log10(p-value). The range of observed and expected counts is shown on the left Y-axis, and the range of the -log10(p-value) is plotted on the right Y-axis. The X-axis is the genomic location on chromosome 4.

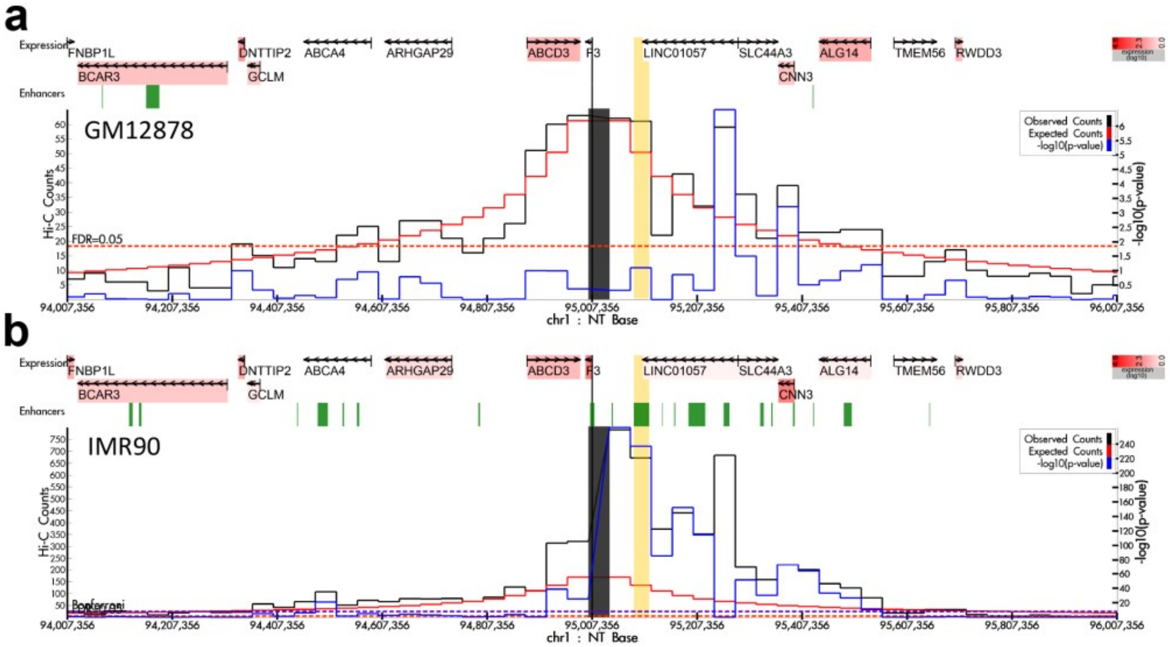


Figure A12. Virtual 4C plot showing one example of the IMR90-specific peaks in (a) GM12878 and (b) IMR90 Hi-C data using HUGIn [13]. On the top of each panel, the genes are colored according to their expression level (the deeper red, the higher the expression level) with arrows indicating the direction of transcription and a vertical bar indicating the transcription start site (TSS). On the bottom of each panel, the chromatin interaction from Hi-C data is shown by a virtual 4C plot. The anchor bin overlapped with the promoter region of gene *F3* is indicated as a thick grey vertical bar at the center. The bin overlapped with the IMR90-specific enhancer region is highlighted in yellow. The black line shows the observed counts; the red line shows the expected counts, and the blue line shows the -log10(p-value). The range of observed and expected counts is shown on the left Y-axis, and the range of the -log10(p-value) is plotted on the right Y-axis. The X-axis is the genomic location on chromosome 1.

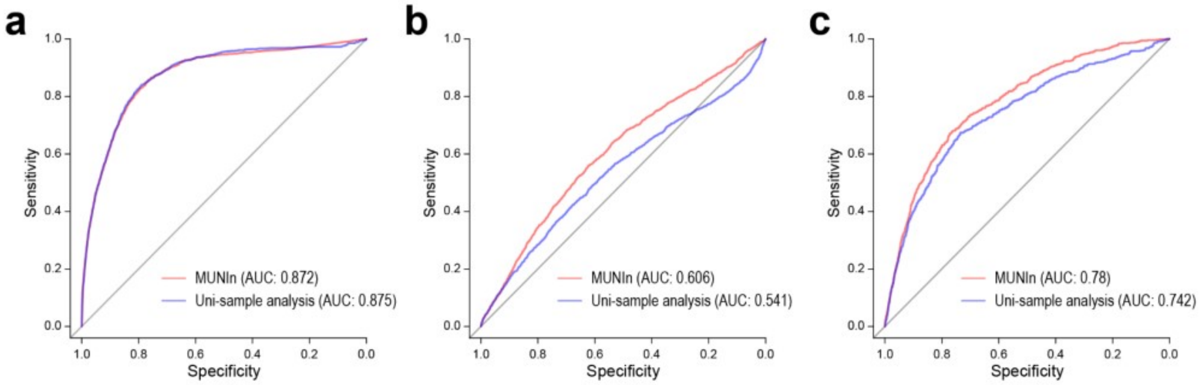


Figure A13. ROC curves for MUNIn and uni-sample analysis in the simulation data where all three samples have equal sequencing depth. ROC curves for (**a**) shared, (**b**) GM12878-specific, and (**c**) IMR90-specific peaks identified by MUNIn and uni-sample analysis using the sliding window approach.

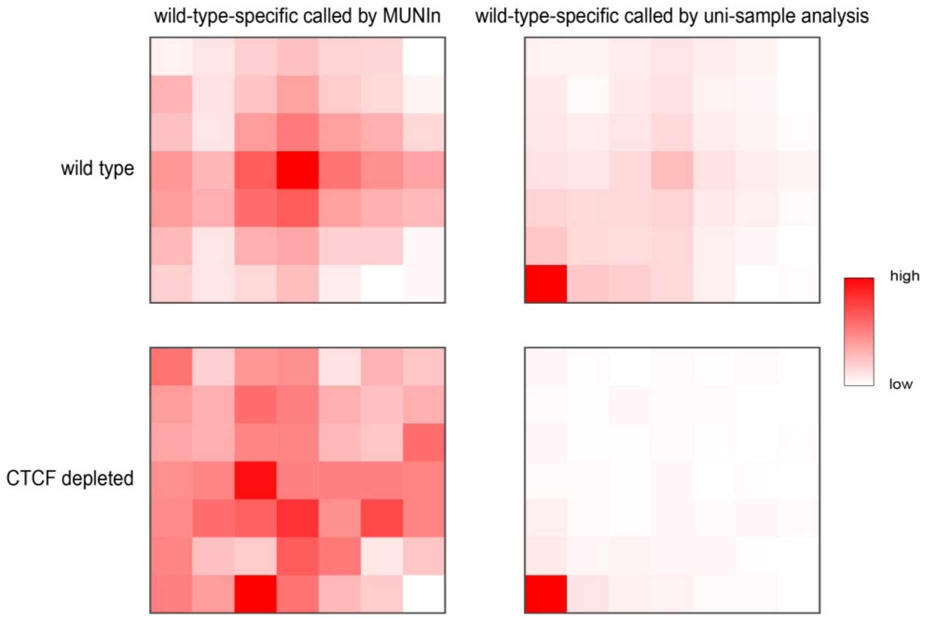


Figure A14. Aggregate peaks plots on mESC wild-type-specific peak loci identified by MUNIn and uni-sample analysis. Observed contact counts are aggregated over the peak bin pairs and their 7 x 7 neighbors. Each bin has a length of 10kb. The first row shows aggregate counts for the wild-type sample and the second row shows aggregate counts for the CTCF-depleted sample. For wild-type-specific peaks, we expect to see a peak pattern in the wild-type sample but not in the CTCF-depleted sample**.**

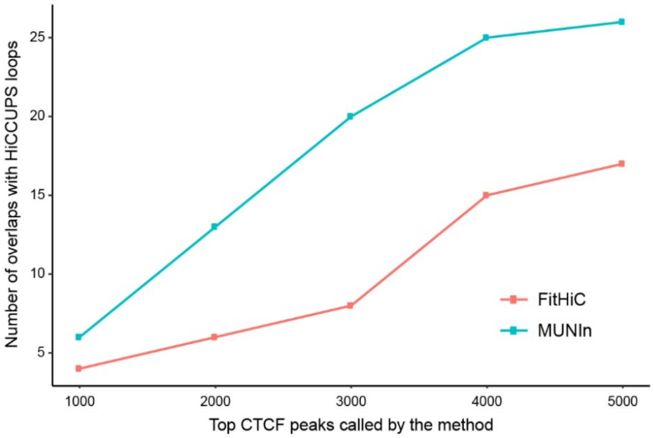
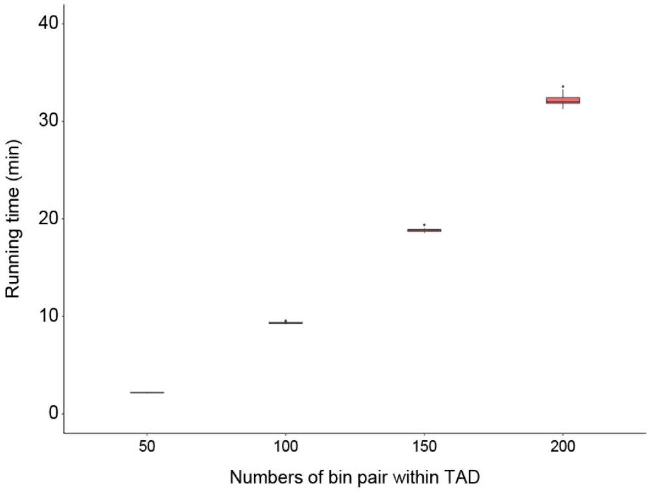


Figure A15. The number of peaks overlapped with HiCCUPS loops. The number of peaks overlapped with HiCCUPS loops at the top 1000, 2000, 3000, 4000, and 5000 CTCF peaks called by MUNIn and Fithic.



FigureA16. Running time of MUNIn. Four shared TADs of GM12878 and IMR90 cell lines, which contain 50, 100, 150 and 200 10kb bins, respectively. For each TAD, uni-sample analysis and MUNIn were executed 10 times. Running time (Y-axis) is in minutes (min). The computing time is from running on a Linux-based computing cluster.

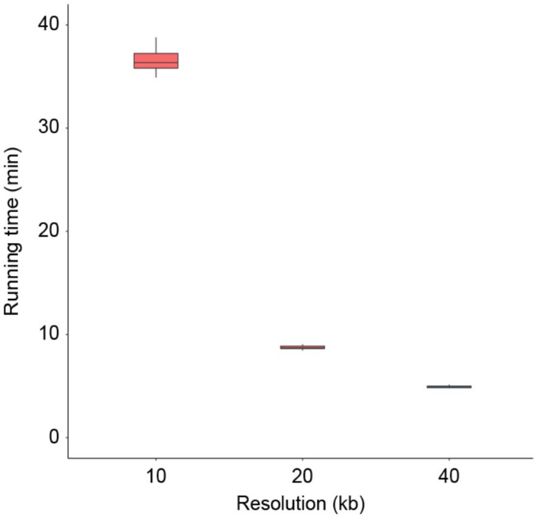


Figure A17. Running time of MUNIn. For four shared TADs of GM12878 and IMR90 cell lines at different resolutions, 10, 20, and 40kb. Running time (Y-axis) is in minutes (min). The computing time is from running on a Linux-based computing cluster.

APPENDIX B: SUPPLEMENTARY MATERIALS FOR CHAPTER 3

Study participants

A total of 1,506 infants born before the 28th week of gestation and 1,249 mothers were enrolled during the years 2002-2004. Study participants were enrolled at 14 hospitals in the United States to achieve a large enough sample size and generalizability. The enrollment and consent procedures were approved by the individual institutional review boards. At the age of 10 years, 889 of the surviving children returned for follow-up (ELGAN2, 92% of the 966 who were recruited for this phase of the ELGAN Study) and were assessed for cognition capacity, learning abilities, and impairments in executive function [79].

Genotype data and quality control

Genomic DNA was isolated from umbilical cords and genotyping was performed using Illumina 1 Million Quad (Illumina Inc, San Diego, California). This work was done as part of the candidate gene analysis of severe intraventricular hemorrhage (IVH) in preterm born infants [141], where infants with birth weights 500-1250g and severe grades IVH and neonates with normal cranial ultrasounds were enrolled prospectively at 24 universities. A subset of ELGAN participants were provided as additional samples along with samples from a few other studies in the IVH study.

We performed variant level and sample level quality control (QC) on genotype data. For variant level QC, we excluded variants with call rate $< 90\%$ or minor allele frequency (MAF) $< 1\%$. For sample level QC, we excluded samples with missing rate $> 10\%$. These resulted in 700,845 SNPs and 528 samples using plink v.1.90 [142] [143].

Genotype imputation

Starting with the quality controlled (QCed) genotype data, we used the Michigan imputation server [144] for phasing and imputation using TOPMed freeze 5 [129] as the reference panel. Specifically, Eagle [145] was used for phasing and Minimac4 [146] was used for imputation. We performed strand matching by dropping ambiguous (i.e., A/T or C/G) SNPs and by flipping nonambiguous SNPs that were initially in strand when compared to alleles in the + strand observed in the TOPMed freeze 5 reference panel. Genotype data was lifted over to genome build hg38. In total, we obtained $\sim$34 million well-imputed variants.

Functional annotations

CADD phred score measures the deleteriousness of variants and is computed as

$$-10 * \log 10(\text{rank/total}).$$

A CADD phred score of $\geqslant$10 indicates that the variant is predicted to be among the 10$\%$ most deleterious variants in the human genome, a score of $\geqslant$20 indicates among the 1$\%$ most deleterious. The fathmm MKL score predicts the functional consequences of variants where values above 0.5 are generally considered deleterious, and values below 0.5 neutral or benign.

SNP-heritability estimation with GCTA

GCTA [147] [148] was used to estimate SNP-heritability for LPAx. We used well-imputed SNPs (Rsq>0.8) with MAF $>$ 1$\%$. Since there were many closely related individuals in our sample, we utilized a method that can estimate pedigree-based and SNP-based heritability simultaneously in one model [148]. The main advantage of this method is that it allows us to estimate SNP-based heritability without having to remove related individuals. A genetic relationship matrix (GRM) was first derived using a total of 9,817,454 variants for 22 autosomes using all 528 samples. Another GRM was then made setting the first GRM off-diagonal elements that were below 0.05 to 0. Heritability was estimated by REML (restricted maximum likelihood) analysis with these two GRMs adjusting for covariates (sex, race, maternal education, gestational age) and the first 10 principal components. Using all 528 genotyped samples, the estimated SNP-heritability of LPAx is 0.38 (s.e.$=$1.38) with a prespecified prevalence of 25$\%$. The point estimates of SNP heritability were moderate but were not significantly different from zero, likely due to our relatively small sample size with closely related individuals.

Table B1 Participant characteristics of the ELGAN2 subset and ELGAN2 cohort.

Variable name	**ELGAN2 subset (N=528)** n (% or SD)	**ELGAN2 (N=889)** n (% or SD)
Infant sex	274 (51.9%)	455 (51.2%)
Male	254 (48.1%)	434 (48.8%)
Female		
Cognitive impairment	390 (73.9%)	660 (74.2%)
No/Low	138 (26.1%)	214 (24.1%)
Moderate/Severe	0	15 (1.7%)
Not reported		
Gestational age	26.1 (1.27)	26.1 (1.28)
Maternal education	205 (38.8%)	355 (39.9%)
<=12 years	119 (22.5%)	202 (22.7%)
13-15 years	204 (38.6%)	306 (34.4%)
16+ years	0	26 (2.9%)
Not reported		
Maternal Smoking	128 (24.2%)	215 (24.2%)
Yes	400 (75.8%)	655 (73.7%)
No	0	19 (2.1%)
Not reported		
Race	342 (64.8%)	554 (62.3%)
White	133 (25.2%)	227 (25.5%)
Black	53 (10.0%)	98 (11.0%)
Other	0	10 (1.1%)
Not reported		
Public insurance	167 (31.6%)	307 (34.5%)
Yes	361 (68.4%)	568 (63.9%)
No	0	14 (1.6%)
Not reported		
Multiple births	189 (35.8%)	313 (35.2%)
Yes	339 (64.2%)	576 (64.8%)
No		

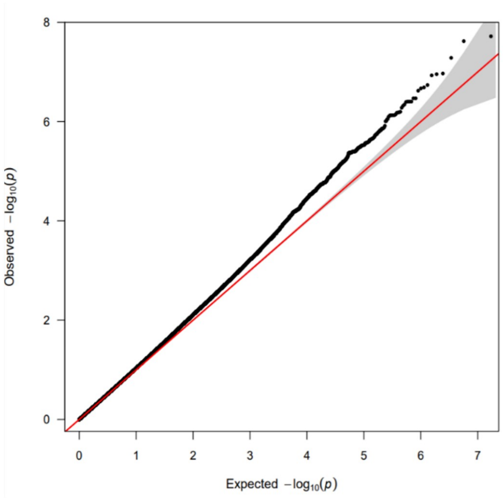


Figure B1. QQ plot. A quantile-quantile (Q-Q) plot is used to characterize the extent to which the observed distribution of the test statistics follows the expected null distribution. This plot was generated using the R package *qqman* [149].

Table B2 Epigenetic functional annotations for selected genome-wide significant and suggestive variants.

rsID	P-value	CADD phred	FATHMM-MKL	Genehancer feature	Genehancer connected gene	Locus
rs11829294	2.40e-08	3.728	0.21	enhancer	*TSPAN9*	*TEAD4* (intron)
rs10774094	5.21e-08	0.805	0.10	enhancer	*TSPAN9*	*TEAD4* (intron)
rs16913588	2.05e-07	7.525	0.97	-	-	intergenic
rs58545250	2.14e-07	9.661	0.49	-	-	*RP11-389O22.4* (downstream)
rs61114884	2.39e-07	3.602	0.15	enhancer	*TSPAN9*	*TEAD4* (intron)
rs17031018	5.27e-07	9.16	0.30	-	-	*LRIG2* (intron)
rs79946490	6.64e-07	10.19	0.23	enhancer	*ITPR1*	*SUMF1* (intron)
rs11062457	7.44e-07	0.362	0.13	enhancer	*TSPAN9*	*TEAD4* (intron)
rs2286647	7.46e-07	0.16	0.07	enhancer	*TSPAN9*	*TEAD4* (intron)
rs143923810	7.73e-07	1.518	0.04	enhancer	*TSPAN9*	*TEAD4* (intron)
rs9424366	9.86e-07	13.82	0.13	enhancer	*CLIC4*	*NIPAL3* (downstream)

Table B3 Variants overlapped with commonMind eQTL*: NCBI build 38.

rsID	Gene	Chr*	Position*	FDR	Index SNP	LD r^2 with the index SNP
rs143923810	*PRMT8*	chr12	2988024	0.010	rs11829294	0.724
rs7302783	*PRMT8*	chr12	2989245	0.010	rs11829294	0.724
rs7302789	*PRMT8*	chr12	2989254	0.010	rs11829294	0.720
rs10082968	*PRMT8*	chr12	2990125	0.025	rs11829294	0.720
rs12322215	*PRMT8*	chr12	3001421	0.048	rs11829294	0.883
rs10128796	*PRMT8*	chr12	3003552	0.045	rs11829294	0.883

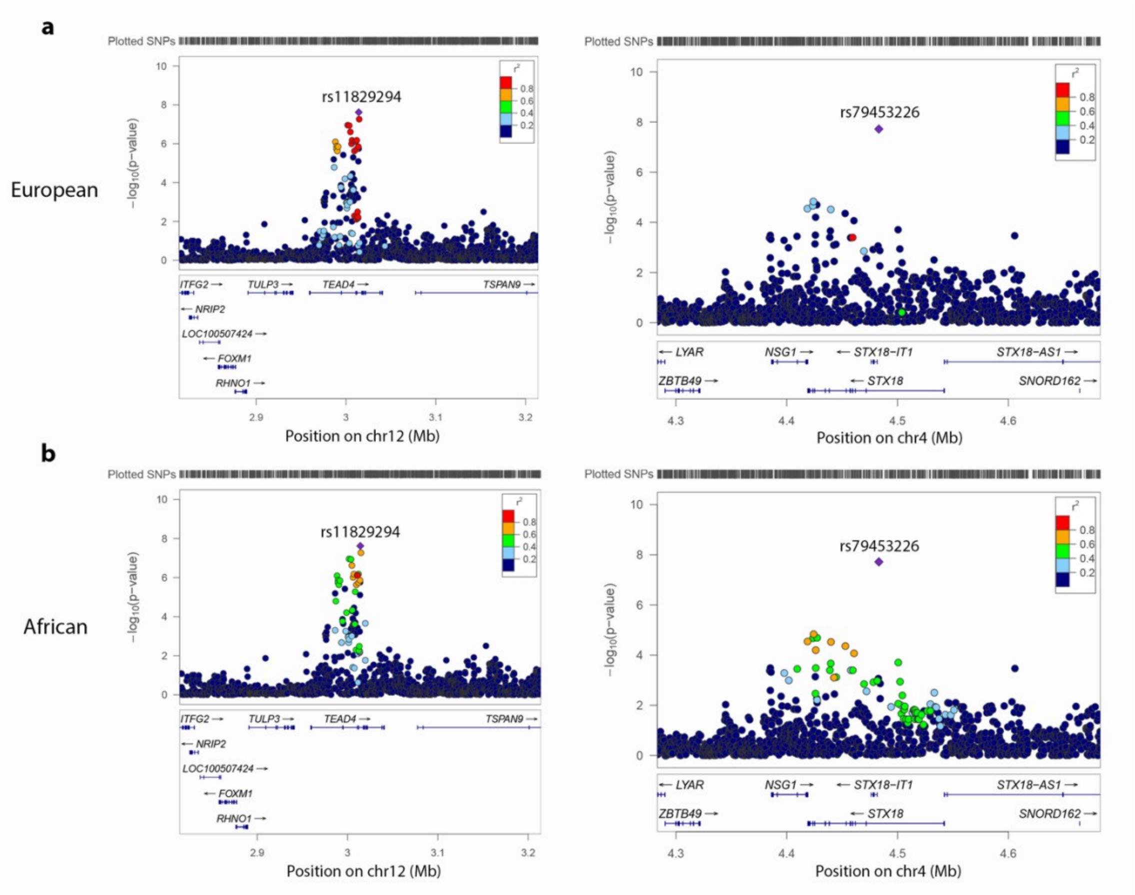


Figure B2. Locus zoom plots for the two genome-wide significant loci. Colors represent linkage disequilibrium r^2 values calculated from TOPMed individuals with the lead SNP in each plot. **(a)** Locus zoom plots with linkage disequilibrium r^2 values calculated from TOPMed European ancestry individuals. **(b)** Locus zoom plots with linkage disequilibrium r^2 values calculated from TOPMed African ancestry individuals. NCBI build 38.

APPENDIX C: SUPPLEMENTARY MATERIALS FOR CHAPTER 4

Identification of SnapHiC-G chromatin enhancer-promoter interactions

SnapHiC-G was applied to 742 mESCs (and down-sampled to 100 and 300 mESCs) scHi-C data and four cell types of the human prefrontal cortex sn-m3C-seq data to identify 10Kb bin chromatin enhancer-promoter interactions on autosomal chromosomes with 20Kb to 1Mb distance range. Bin pairs within 20 Kb were excluded from analyses.

Definition of overlapped enhancer-promoter interactions

We followed the same definition of overlapped loops as in SnapHiC. We define d_{im} as the 1D genomic distance between the center of bin i and the center of bin m, and we define the distance between (i, j) and (m, n) as the maximum of d_{im} and d_{jn}. For an enhancer-promoter interaction (i, j), if there exists an enhancer-promoter interaction (m, n) in set S such that the distance between (i, j) and (m, n) is within 20Kb, we define that the enhancer-promoter interaction (i, j) overlaps with the set S.

Benchmarking with human brain cortical cells

To evaluate the performance of SnapHiC-G in human brain cortical cells, we tested SnapHiC, FitHiC2, FastHiC, HiC-ACT, and HiC-DC+ along with SnapHiC-G on three cell types (oligodendrocytes, microglia, and L2/3 neurons) from 2,869 human brain cortical cells, each with more than 150,000 contacts, from the Lee et al. study [100], where bulk H3K4me3 PLAC-Seq data was available for reference (oligodendrocytes, microglia, and neurons) [115]. We applied each method for the three brain cell types separately and pooled single cells from the same cell type as the pseudo-bulk Hi-C data for bulk Hi-C methods. As shown in Figures C3-C4, SnapHiC-G detected the largest number of significant interactions and achieved higher sensitivity than alternative methods for the analyses of all three brain cell types, where the power gain was more substantial for L2/3 neurons and microglia. Similar to mESCs, there was a large difference between the number of significant interactions identified from different methods based on comparable significance thresholds (Methods: Identification of loops/interactions using other Hi-C methods).

For example, with 261 L2/3 neurons, the number of significant interactions ranged from 1,471 for HiC-ACT to 6,585 for FastHiC, while SnapHiC-G identified more than 20,000 significant interactions (Table 1).

Difference between SnapHiC and SnapHiC-G

For each bin pair, SnapHiC-G applies the one-sample t-test to test whether its average normalized contact frequency across single cells is significantly greater than zero (global background), while SnapHiC requires the average value to be positive and larger than the mean value of its surrounding bin pairs (local background). As a result, SnapHiC-G called more interactions than SnapHiC, with a significantly higher sensitivity for detecting enhancer-promoter interactions with the same FDR cutoff.

The reduced sensitivity of SnapHiC can be explained by three reasons. First, gene promoters can form wide-spread interaction clusters with multiple narrow typical enhancers or broad super-enhancers, and enhancer-promoter interactions may not be located exactly at the summit of such clusters. In addition, CTCF-anchored loops can bring enhancers to the proximity of distal promoters to facilitate enhancer-promoter interaction, as suggested by the loop extrusion model [150, 151]. Finally, enhancer-promoter interactions can be formed by the phase separation mechanism, which is independent of CTCF binding. In all these three scenarios, applying a local background model and selecting interaction summits will miss many of the enhancer-promoter interactions. Another key distinction between SnapHiC-G and SnapHiC is using cell-type-specific epigenetic data to narrow down candidate bin pairs. While such data greatly aid the detection of cell-type-specific enhancer-promoter interactions, when they are not available, users can input only the TSS files to define the promoter regions and apply SnapHiC-G to identify promoter-interacting regions.

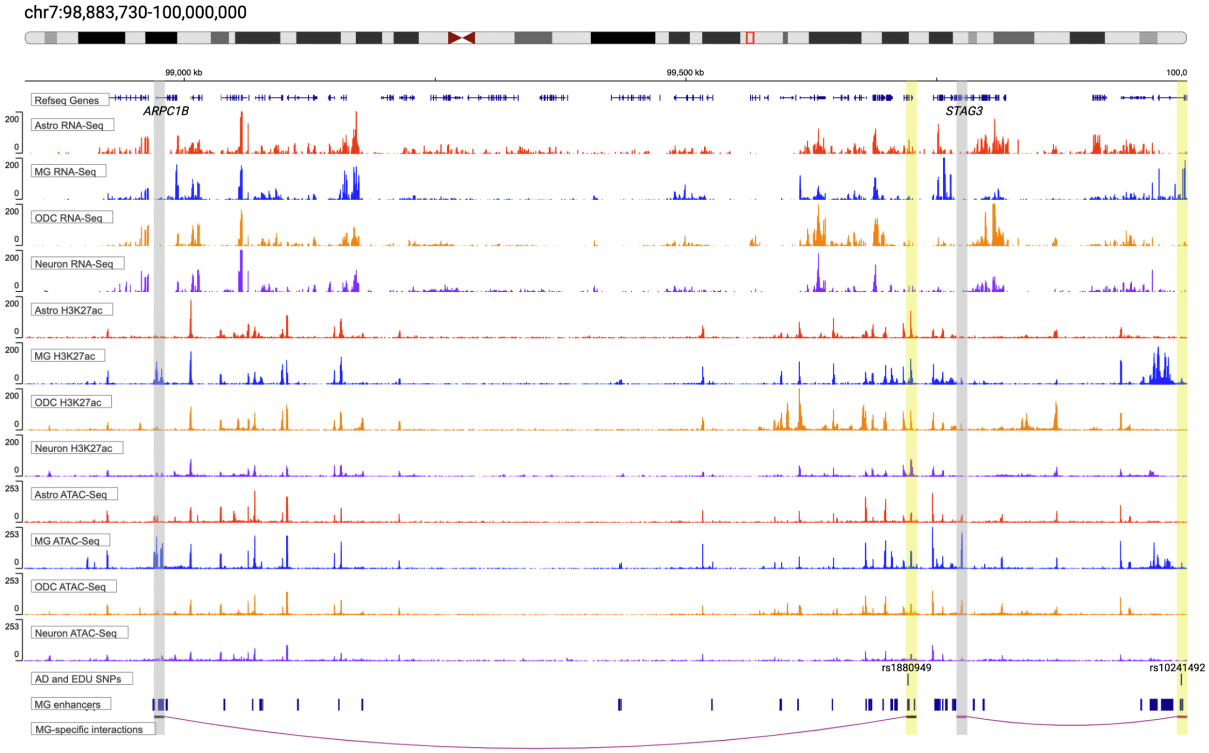


Figure C1. An illustrative example at the *APRC1B*-*STAG3* locus on chromosome 7 with a microglia-specific interaction. The top panel shows the gene track. The middle panels show RNA-seq, H3K27ac, and ATAC-seq tracks for the four brain cell types. The bottom panels show AD and EDU GWAS SNPs, microglia enhancer regions, and two microglia-specific interactions identified by SnapHiC-G but not SnapHiC. One microglia-specific interaction links the promoter region of *ARPC1B* (highlighted in grey) with an AD-associated SNP rs1880949 (highlighted in yellow); the other one links the promoter region of *STAG3* (highlighted in grey) with an EDU-association SNP rs10241492 (highlighted in yellow). Both genes highlighted showed microglia-specific gene expression. The anchors of these interactions also showed stronger H3K27ac ChIP-seq and ATAC-seq signals in microglia.

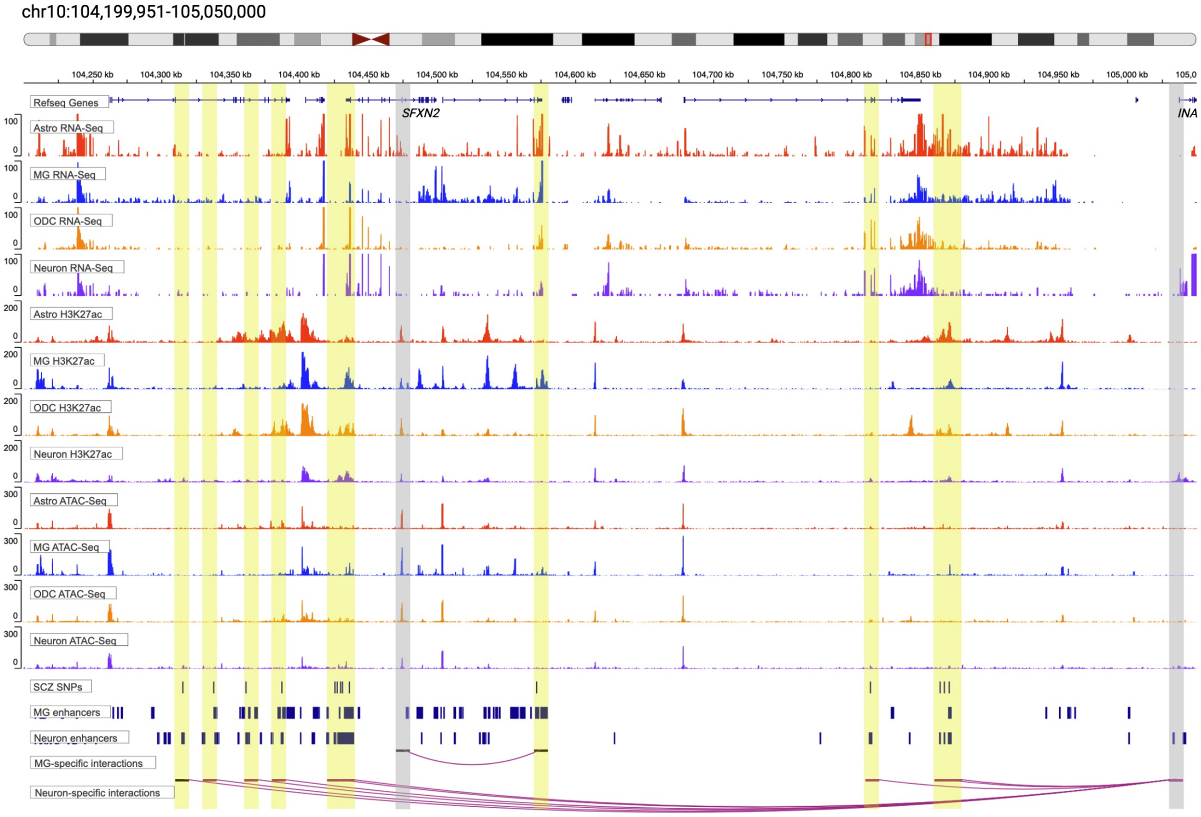


Figure C2. An illustrative example at the *SFXN2-INA* locus on chromosome 10 with neuron- and microglia-specific interactions. The top panel shows the gene track. The middle panels show RNA-Seq, H3K27ac, and ATAC-seq tracks for the four brain cell types. The bottom panels show SCZ GWAS SNPs, enhancer regions in microglia and neurons, and cell-type-specific interactions identified by SnapHiC-G but not SnapHiC. A microglia-specific interaction links the promoter region of *SFXN2* (highlighted in grey) with a SCZ-associated SNP (highlighted in yellow); multiple neuron-specific interactions link the promoter region of *INA* (highlighted in grey) to SCZ-association loci (highlighted in yellow). Both genes highlighted showed cell-type-specific gene expression in corresponding cell types. The anchors of these interactions also showed stronger H3K27ac ChIP-seq and ATAC-seq signals in matched cell types.

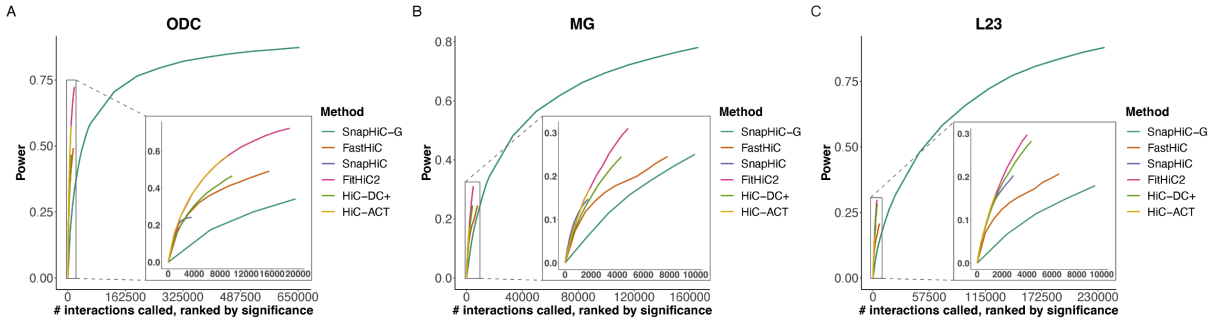


Figure C3. Power curves with (a) oligodendrocytes, (b) microglia, and (c) L2/3 neurons. Interactions were ranked by significance on the x-axis, and power was evaluated with the corresponding number of top interactions. Smaller figures in the lower right-hand corner are zoomed-in views of the top 20,000 interactions from oligodendrocytes and the top 10,000 interactions from microglia and L2/3 neurons.

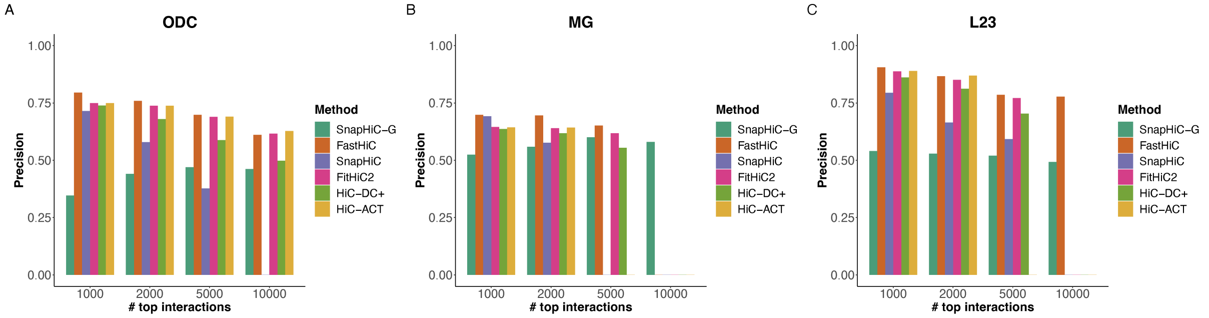
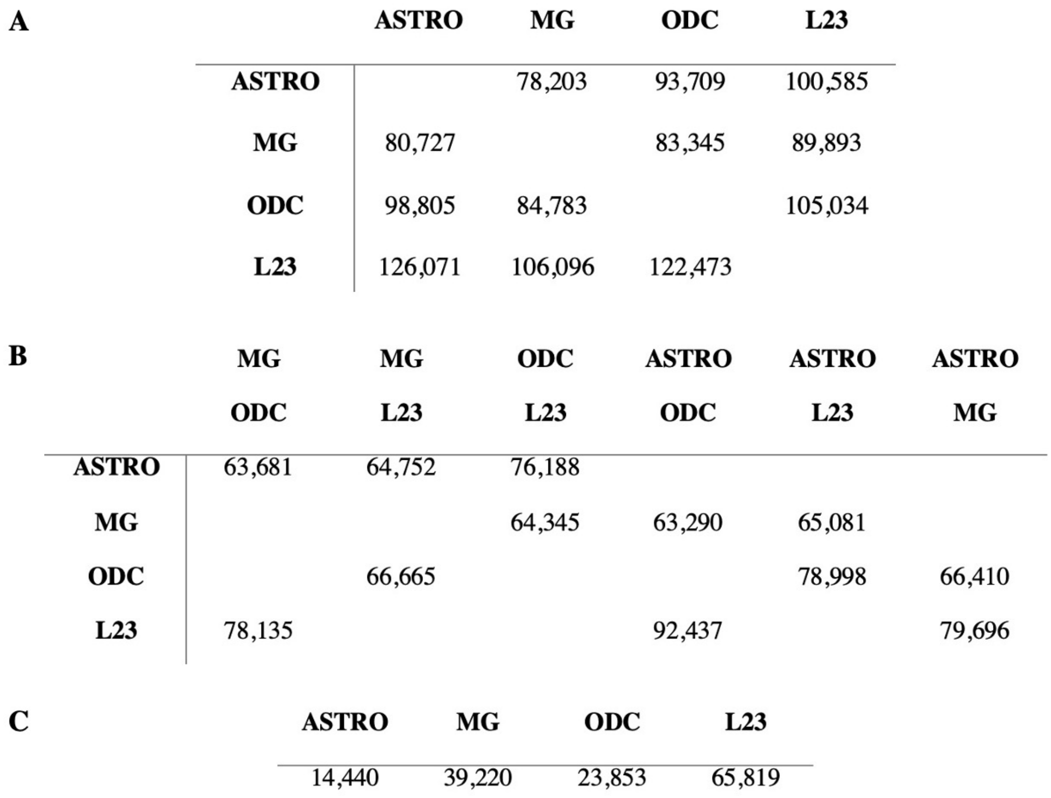


Figure C4. Precision barplots for three brain cell types. Precision bar plots with (a) oligodendrocytes, (b) microglia, and (c) L2/3 neurons. Results shown were for the top 1,000, 2,000, 5,000, and 10,000 interactions ranked by significance. Some bars are missing because the number exceeds the number of interactions called by that method.



A

	ASTRO	MG	ODC	L23
ASTRO		78,203	93,709	100,585
MG	80,727		83,345	89,893
ODC	98,805	84,783		105,034
L23	126,071	106,096	122,473	

B

	MG	MG	ODC	ASTRO	ASTRO	ASTRO
	ODC	L23	L23	ODC	L23	MG
ASTRO	63,681	64,752	76,188			
MG			64,345	63,290	65,081	
ODC		66,665			78,998	66,410
L23	78,135			92,437		79,696

C

ASTRO	MG	ODC	L23
14,440	39,220	23,853	65,819

Table C1. Overlapped interactions in human brain cell types. (A) Four-by-four table of two-way overlapped interactions. Element (i, j) (e.g. (1, 3)) is the number of the interactions identified in the i-th (e.g., astrocytes) cell type that overlap with the interactions identified in the j-th (e.g., oligodendrocytes) cell type. (B) Four-by-seven table of three-way overlapped interactions. Element (i, j) (e.g., (1, 3)) is the number of the interactions identified in the i-th (e.g., astrocytes) cell type that overlaps with the interactions identified in both cell types j (e.g., oligodendrocytes and L2/3 neurons). (C) The number of cell-type specific interactions. Column i is the number of interactions that are identified in the i-th cell type that does not overlap with interactions in any of the other cell types.

Genome	# cells	Cell type	MaxRSS/node (Gb)	Runtime per CPU (hour)	# CPUs	# Tasks/node	# Nodes	Memory-per-task (Gb)
mm10	500	mESC	12.72	0.59	100	25	4	0.51
mm10	742	mESC	9.00	0.78	150	10	15	0.90
hg19	338	Astro	11.19	0.57	60	20	3	0.56
hg19	323	MG	7.61	0.68	60	10	6	0.76
hg19	1038	ODC	10.84	1.27	80	10	8	1.08
hg19	261	ODC	7.88	1.77	60	10	6	0.79

Table C2. Running time and memory for SnapHiC-G step C and step D. All scHi-C data are at 10 Kb resolution.

REFERENCES

[1] Bonev B and Cavalli G. Organization and function of the 3D genome. *Nature reviews. Genetics*, 17:661–678, Oct 2016.

[2] Schmitt AD, Hu M, and Ren B. Genome-wide mapping and analysis of chromosome architecture. *Nature reviews. Molecular cell biology*, 17:743–755, Dec 2016.

[3] Rowley MJ and Corces VG. Organizational principles of 3D genome architecture. *Nature reviews. Genetics*, 19:789–800, Dec 2018.

[4] Zheng H and Xie W. The role of 3D genome organization in development and cell differentiation. *Nature reviews. Molecular cell biology*, 20:535–550, Sep 2019.

[5] Marchal C, Sima J, and Gilbert DM. Control of DNA replication timing in the 3D genome. *Nature reviews. Molecular cell biology*, 20:721–737, Dec 2019.

[6] Jerkovic I and Cavalli G. Understanding 3D genome organization by multidisciplinary methods. *Nature reviews. Molecular cell biology*, 22:511–528, Aug 2021.

[7] Cremer T and Cremer M. Chromosome territories. *Cold Spring Harbor perspectives in biology*, 2:a003889, Mar 2010.

[8] Lieberman-Aiden E, van Berkum NL, Williams L, Imakaev M, Ragoczy T, Telling A, Amit I, Lajoie BR, Sabo PJ, Dorschner MO, Sandstrom R, Bernstein B, Bender MA, Groudine M, Gnirke A, Stamatoyannopoulos J, Mirny LA, Lander ES, and Dekker J. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science (New York, N.Y.)*, 326:289–93, Oct 2009.

[9] Rao SS, Huntley MH, Durand NC, Stamenova EK, Bochkov ID, Robinson JT, Sanborn AL, Machol I, Omer AD, Lander ES, and Aiden EL. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell*, 159:1665–80, Dec 2014.

[10] Fortin JP and Hansen KD. Reconstructing A/B compartments as revealed by hi-c using long-range correlations in epigenetic data. *Genome biology*, 16:180, Aug 2015.

[11] Dixon JR, Selvaraj S, Yue F, Kim A, Li Y, Shen Y, Hu M, Liu JS, and Ren B. Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature*, 485:376–80, Apr 2012.

[12] Nora EP, Lajoie BR, Schulz EG, Giorgetti L, Okamoto I, Servant N, Piolot T, van Berkum NL, Meisig J, Sedat J, Gribnau J, Barillot E, Blthgen N, Dekker J, and Heard E. Spatial partitioning of the regulatory landscape of the x-inactivation centre. *Nature*, 485:381–5, Apr 2012.

[13] Martin JS, Xu Z, Reiner AP, Mohlke KL, Sullivan P, Ren B, Hu M, and Li Y. Hugin: Hi-c unifying genomic interrogator. *Bioinformatics (Oxford, England)*, 33:3793–3795, Dec 2017.

[14] Schoenfelder S and Fraser P. Long-range enhancer-promoter contacts in gene expression control. *Nature reviews. Genetics*, 20:437–455, Aug 2019.

[15] Lagler TM, Abnousi A, Hu M, Yang Y, and Li Y. Hic-act: improved detection of chromatin interactions from hi-c data via aggregated cauchy test. *American journal of human genetics*, 108:257–268, Feb 2021.

[16] Krijger PH and de Laat W. Regulation of disease-associated gene expression in the 3D genome. *Nature reviews. Molecular cell biology*, 17:771–782, Dec 2016.

[17] Schulz JM, Al-Khazraji BK, and Shoemaker JK. Sodium nitroglycerin induces middle cerebral artery vasodilatation in young, healthy adults. *Experimental physiology*, 103:1047–1055, Aug 2018.

[18] Stadhouders R, Filion GJ, and Graf T. Transcription factors and 3D genome conformation in cell-fate decisions. *Nature*, 569:345–354, May 2019.

[19] Fullwood MJ, Liu MH, Pan YF, Liu J, Xu H, Mohamed YB, Orlov YL, Velkov S, Ho A, Mei PH, Chew EG, Huang PY, Welboren WJ, Han Y, Ooi HS, Ariyaratne PN, Vega VB, Luo Y, Tan PY, Choy PY, Wansa KD, Zhao B, Lim KS, Leow SC, Yow JS, Joseph R, Li H, Desai KV, Thomsen JS, Lee YK, Karuturi RK, Herve T, Bourque G, Stunnenberg HG, Ruan X, Cacheux-Rataboul V, Sung WK, Liu ET, Wei CL, Cheung E, and Ruan Y. An oestrogen-receptor-alpha-bound human chromatin interactome. *Nature*, 462:58–64, Nov 2009.

[20] Mifsud B, Tavares-Cadete F, Young AN, Sugar R, Schoenfelder S, Ferreira L, Wingett SW, Andrews S, Grey W, Ewels PA, Herman B, Happe S, Higgs A, LeProust E, Follows GA, Fraser P, Luscombe NM, and Osborne CS. Mapping long-range promoter contacts in human cells with high-resolution capture hi-c. *Nature genetics*, 47:598–606, Jun 2015.

[21] Fang R, Yu M, Li G, Chee S, Liu T, Schmitt AD, and Ren B. Mapping of long-range chromatin interactions by proximity ligation-assisted chip-seq., Dec 2016.

[22] Bonev B, Mendelson Cohen N, Szabo Q, Fritsch L, Papadopoulos GL, Lubling Y, Xu X, Lv X, Hugnot JP, Tanay A, and Cavalli G. Multiscale 3D genome rewiring during mouse neural development. *Cell*, 171:557–572.e24, Oct 2017.

[23] Li H and Durbin R. Fast and accurate long-read alignment with burrows-wheeler transform. *Bioinformatics (Oxford, England)*, 26:589–95, Mar 2010.

[24] Imakaev M, Fudenberg G, McCord RP, Naumova N, Goloborodko A, Lajoie BR, Dekker J, and Mirny LA. Iterative correction of hi-c data reveals hallmarks of chromosome organization. *Nature methods*, 9:999–1003, Oct 2012.

[25] Ramani V, Deng X, Qiu R, Gunderson KL, Steemers FJ, Disteche CM, Noble WS, Duan Z, and Shendure J. Massively multiplex single-cell hi-c. *Nature methods*, 14:263–266, Mar 2017.

[26] Stevens TJ, Lando D, Basu S, Atkinson LP, Cao Y, Lee SF, Leeb M, Wohlfahrt KJ, Boucher W, O'Shaughnessy-Kirwan A, Cramard J, Faure AJ, Ralser M, Blanco E, Morey L, Sans M, Palayret MGS, Lehner B, Di Croce L, Wutz A, Hendrich B, Klenerman D, and

Laue ED. 3D structures of individual mammalian genomes studied by single-cell hi-c. *Nature*, 544:59–64, Apr 2017.

[27] Nagano T, Lubling Y, Vrnai C, Dudley C, Leung W, Baran Y, Mendelson Cohen N, Wingett S, Fraser P, and Tanay A. Cell-cycle dynamics of chromosomal organization at single-cell resolution. *Nature*, 547:61–67, Jul 2017.

[28] Galitsyna AA and Gelfand MS. Single-cell hi-c data analysis: safety in numbers. *Briefings in bioinformatics*, 22, Nov 2021.

[29] Ay F and Noble WS. Analysis methods for studying the 3D architecture of the genome. *Genome biology*, 16:183, Sep 2015.

[30] Servant N, Varoquaux N, Lajoie BR, Viara E, Chen CJ, Vert JP, Heard E, Dekker J, and Barillot E. Hic-pro: an optimized and flexible pipeline for hi-c data processing. *Genome biology*, 16:259, Dec 2015.

[31] Yaffe E and Tanay A. Probabilistic modeling of hi-c contact maps eliminates systematic biases to characterize global chromosomal architecture. *Nature genetics*, 43:1059–65, Oct 2011.

[32] Cournac A, Marie-Nelly H, Marbouty M, Koszul R, and Mozziconacci J. Normalization of a chromosomal contact map. *BMC genomics*, 13:436, Aug 2012.

[33] Hu M, Deng K, Selvaraj S, Qin Z, Ren B, and Liu JS. Hicnorm: removing biases in hi-c data via poisson regression. *Bioinformatics (Oxford, England)*, 28:3131–3, Dec 2012.

[34] Shavit Y and Lio' P. Combining a wavelet change point and the bayes factor for analysing chromosomal interaction data. *Molecular bioSystems*, 10:1576–85, Jun 2014.

[35] Li W, Gong K, Li Q, Alber F, and Zhou XJ. Hi-corrector: a fast, scalable and memory-efficient package for normalizing large-scale hi-c data. *Bioinformatics (Oxford, England)*, 31:960–2, Mar 2015.

[36] Forcato M, Nicoletti C, Pal K, Livi CM, Ferrari F, and Bicciato S. Comparison of computational methods for hi-c data analysis. *Nature methods*, 14:679–685, Jul 2017.

[37] Carty M, Zamparo L, Sahin M, Gonzlez A, Pelossof R, Elemento O, and Leslie CS. An integrated model for detecting significant chromatin interactions from high-resolution hi-c data. *Nature communications*, 8:15454, May 2017.

[38] Sahin M, Wong W, Zhan Y, Van Deynze K, Koche R, and Leslie CS. Hic-dc+ enables systematic 3D interaction calls and differential analysis for hi-c and hichip. *Nature communications*, 12:3366, Jun 2021.

[39] Ay F, Bailey TL, and Noble WS. Statistical confidence estimation for hi-c data reveals regulatory chromatin contacts. *Genome research*, 24:999–1011, Jun 2014.

[40] Kaul A, Bhattacharyya S, and Ay F. Identifying statistically significant chromatin contacts from hi-c data with fithic2. *Nature protocols*, 15:991–1012, Mar 2020.

[41] Xu Z, Zhang G, Jin F, Chen M, Furey TS, Sullivan PF, Qin Z, Hu M, and Li Y. A hidden markov random field-based bayesian method for the detection of long-range chromosomal interactions in hi-c data. *Bioinformatics (Oxford, England)*, 32:650–6, Mar 2016.

[42] Cao Y, Chen Z, Chen X, Ai D, Chen G, McDermott J, Huang Y, Guo X, and Han JJ. Accurate loop calling for 3D genomic data with cloops. *Bioinformatics (Oxford, England)*, 36:666–675, Feb 2020.

[43] Rowley MJ, Poulet A, Nichols MH, Bixler BJ, Sanborn AL, Brouhard EA, Hermetz K, Linsenbaum H, Csankovszki G, Lieberman Aiden E, and Corces VG. Analysis of hi-c data using SIP effectively identifies loops in organisms from ¡i¿c. elegans¡/i¿ to mammals. *Genome research*, 30:447–458, Mar 2020.

[44] Roayaei Ardakany A, Gezer HT, Lonardi S, and Ay F. Mustache: multi-scale detection of chromatin loops from hi-c and micro-c maps using scale-space representation. *Genome biology*, 21:256, Sep 2020.

[45] Buniello A, MacArthur JAL, Cerezo M, Harris LW, Hayhurst J, Malangone C, McMahon A, Morales J, Mountjoy E, Sollis E, Suveges D, Vrousgou O, Whetzel PL, Amode R, Guillen JA, Riat HS, Trevanion SJ, Hall P, Junkins H, Flicek P, Burdett T, Hindorff LA, Cunningham F, and Parkinson H. The NHGRI-EBI GWAS catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic acids research*, 47:D1005–D1012, Jan 2019.

[46] Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS, and Manolio TA. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proceedings of the National Academy of Sciences of the United States of America*, 106:9362–7, Jun 2009.

[47] Sun Q, Crowley CA, Huang L, Wen J, Chen J, Bao EL, Auer PL, Lettre G, Reiner AP, Sankaran VG, Raffield LM, and Li Y. From GWAS variant to function: A study of 148,000 variants for blood cell traits. *HGG advances*, 3:100063, Jan 2022.

[48] Trynka G, Sandor C, Han B, Xu H, Stranger BE, Liu XS, and Raychaudhuri S. Chromatin marks identify critical cell types for fine mapping complex trait variants. *Nature genetics*, 45:124–30, Feb 2013.

[49] Zhang F and Lupski JR. Non-coding genetic variants in human disease. *Human molecular genetics*, 24:R102–10, Oct 2015.

[50] Dekker J, Marti-Renom MA, and Mirny LA. Exploring the three-dimensional organization of genomes: interpreting chromatin interaction data. *Nature reviews. Genetics*, 14:390–403, Jun 2013.

[51] Fulco CP, Munschauer M, Anyoha R, Munson G, Grossman SR, Perez EM, Kane M, Cleary B, Lander ES, and Engreitz JM. Systematic mapping of functional enhancer-promoter connections with CRISPR interference. *Science (New York, N.Y.)*, 354:769–773, Nov 2016.

[52] Smemo S, Tena JJ, Kim KH, Gamazon ER, Sakabe NJ, Gmez-Marn C, Aneas I, Credidio FL, Sobreira DR, Wasserman NF, Lee JH, Puviindran V, Tam D, Shen M, Son JE, Vakili NA, Sung HK, Naranjo S, Acemel RD, Manzanares M, Nagy A, Cox NJ, Hui CC, Gomez-Skarmeta JL, and Nbrega MA. Obesity-associated variants within FTO form long-range functional connections with IRX3. *Nature*, 507:371–5, Mar 2014.

[53] van de Werken HJ, Landan G, Holwerda SJ, Hoichman M, Klous P, Chachik R, Splinter E, Valdes-Quezada C, Oz Y, Bouwman BA, Verstegen MJ, de Wit E, Tanay A, and de Laat W. Robust 4c-seq data analysis to screen for regulatory DNA interactions. *Nature methods*, 9:969–72, Oct 2012.

[54] Yu M and Ren B. The three-dimensional organization of mammalian genomes. *Annual review of cell and developmental biology*, 33:265–289, Oct 2017.

[55] Li Y, Hu M, and Shen Y. Gene regulation in the 3D genome. *Human molecular genetics*, 27:R228–R233, Aug 2018.

[56] Zhou X, Chen Y, Mok KY, Kwok TCY, Mok VCT, Guo Q, Ip FC, Mullapudi N, Giusti-Rodrguez P, Sullivan PF, Hardy J, Fu AKY, Li Y, and Ip NY. Non-coding variability at the APOE locus contributes to the alzheimer's risk. *Nature communications*, 10:3310, Jul 2019.

[57] Song M, Yang X, Ren X, Maliskova L, Li B, Jones IR, Wang C, Jacob F, Wu K, Traglia M, Tam TW, Jamieson K, Lu SY, Ming GL, Li Y, Yao J, Weiss LA, Dixon JR, Judge LM, Conklin BR, Song H, Gan L, and Shen Y. Mapping cis-regulatory chromatin contacts in neural cells links neuropsychiatric disorder risk variants to target genes. *Nature genetics*, 51:1252–1262, Aug 2019.

[58] Paola MD Giusti-Rodriguez and Patrick F Sullivan. Using three-dimensional regulatory chromatin interactions from adult and fetal cortex to interpret genetic results for psychiatric disorders and cognitive traits. 8 2018.

[59] Flutre T, Wen X, Pritchard J, and Stephens M. A statistical framework for joint eqtl analysis in multiple tissues. *PLoS genetics*, 9:e1003486, May 2013.

[60] Beaumont MA and Rannala B. The bayesian revolution in genetics. *Nature reviews. Genetics*, 5:251–61, Apr 2004.

[61] Neal S. Grantham, Yawen Guan, Brian J. Reich, Elizabeth T. Borer, and Kevin Gross. MIMIX: A bayesian mixed-effects model for microbiome data from designed experiments. *Journal of the American Statistical Association*, 115(530):599–609, 7 2019.

[62] Chen X, Jung JG, Shajahan-Haq AN, Clarke R, Shih IeM, Wang Y, Magnani L, Wang TL, and Xuan J. Chip-bit: Bayesian inference of target genes using a novel joint probabilistic model of chip-seq profiles. *Nucleic acids research*, 44:e65, Apr 2016.

[63] Robin X, Turck N, Hainard A, Tiberti N, Lisacek F, Sanchez JC, and Mller M. proc: an open-source package for R and S+ to analyze and compare ROC curves. *BMC bioinformatics*, 12:77, Mar 2011.

[64] Dixon JR, Jung I, Selvaraj S, Shen Y, Antosiewicz-Bourget JE, Lee AY, Ye Z, Kim A, Rajagopal N, Xie W, Diao Y, Liang J, Zhao H, Lobanenkov VV, Ecker JR, Thomson JA, and Ren B. Chromatin architecture reorganization during stem cell differentiation. *Nature*, 518:331–6, Feb 2015.

[65] Lawrence Hubert and Phipps Arabie. Comparing partitions. *Journal of Classification*, 2(1):193–218, 12 1985.

[66] Jung I, Schmitt A, Diao Y, Lee AJ, Liu T, Yang D, Tan C, Eom J, Chan M, Chee S, Chiang Z, Kim C, Masliah E, Barr CL, Li B, Kuan S, Kim D, and Ren B. A compendium of promoter-centered long-range chromatin interactions in the human genome. *Nature genetics*, 51:1442–1449, Oct 2019.

[67] Kubo N, Ishii H, Xiong X, Bianco S, Meitinger F, Hu R, Hocker JD, Conte M, Gorkin D, Yu M, Li B, Dixon JR, Hu M, Nicodemi M, Zhao H, and Ren B. Promoter-proximal CTCF binding promotes distal enhancer-dependent gene activation. *Nature structural molecular biology*, 28:152–161, Feb 2021.

[68] Mathews TJ and Driscoll AK. Trends in infant mortality in the united states, 2005-2014. *NCHS data brief*, pages 1–8, Mar 2017.

[69] Moore T, Hennessy EM, Myles J, Johnson SJ, Draper ES, Costeloe KL, and Marlow N. Neurological and developmental outcome in extremely preterm children born in england in 1995 and 2006: the epicure studies. *BMJ (Clinical research ed.)*, 345:e7961, Dec 2012.

[70] Serenius F, Klln K, Blennow M, Ewald U, Fellman V, Holmstrm G, Lindberg E, Lundqvist P, Marl K, Norman M, Olhager E, Stigson L, Stjernqvist K, Vollmer B, and Strmberg B and. Neurodevelopmental outcome in extremely preterm infants at 2.5 years after active perinatal care in sweden. *JAMA*, 309:1810–20, May 2013.

[71] Pascal A, Govaert P, Oostra A, Naulaers G, Ortibus E, and Van den Broeck C. Neurodevelopmental outcome in very preterm and very-low-birthweight infants born over the past decade: a meta-analytic review. *Developmental medicine and child neurology*, 60:342–355, Apr 2018.

[72] Ding S, Lemyre B, Daboval T, Barrowman N, and Moore GP. A meta-analysis of neurodevelopmental outcomes at 4-10 years in children born at 22-25 weeks gestation. *Acta paediatrica (Oslo, Norway : 1992)*, 108:1237–1244, Jul 2019.

[73] Korologou-Linden R, Anderson EL, Jones HJ, Davey Smith G, Howe LD, and Stergiakouli E. Polygenic risk scores for alzheimer's disease, and academic achievement, cognitive and behavioural measures in children from the general population. *International journal of epidemiology*, 48:1972–1980, Dec 2019.

[74] Johnson S, Fawke J, Hennessy E, Rowell V, Thomas S, Wolke D, and Marlow N. Neurodevelopmental disability through 11 years of age in children born before 26 weeks of gestation. *Pediatrics*, 124:e249–57, Aug 2009.

[75] Russ SA, Larson K, and Halfon N. A national profile of childhood epilepsy and seizure disorder. *Pediatrics*, 129:256–64, Feb 2012.

[76] Van Naarden Braun K, Christensen D, Doernberg N, Schieve L, Rice C, Wiggins L, Schendel D, and Yeargin-Allsopp M. Trends in the prevalence of autism spectrum disorder, cerebral palsy, hearing loss, intellectual disability, and vision impairment, metropolitan atlanta, 1991-2010. *PloS one*, 10:e0124120, 2015.

[77] Chan E, Leong P, Malouf R, and Quigley MA. Long-term cognitive and school outcomes of late-preterm and early-term births: a systematic review. *Child: care, health and development*, 42:297–312, May 2016.

[78] Kuban KC, Joseph RM, O'Shea TM, Allred EN, Heeren T, Douglass L, Stafstrom CE, Jara H, Frazier JA, Hirtz D, and Leviton A and. Girls and boys born before 28 weeks gestation: Risks of cognitive, behavioral, and neurologic outcomes at age 10 years. *The Journal of pediatrics*, 173:69–75.e1, Jun 2016.

[79] Joseph RM, O'Shea TM, Allred EN, Heeren T, Hirtz D, Jara H, Leviton A, and Kuban KC and. Neurocognitive and academic outcomes at age 10 years of extremely preterm newborns. *Pediatrics*, 137, Apr 2016.

[80] Johnson S and Marlow N. Early and long-term outcome of infants born extremely preterm. *Archives of disease in childhood*, 102:97–102, Jan 2017.

[81] Clark EA, Mele L, Wapner RJ, Spong CY, Sorokin Y, Peaceman A, Iams JD, Leveno KJ, Harper M, Caritis SN, Miodovnik M, Mercer BM, Thorp JM, Ramin SM, Carpenter M, and Rouse DJ and. Association of fetal inflammation and coagulation pathway gene polymorphisms with neurodevelopmental delay at age 2 years. *American journal of obstetrics and gynecology*, 203:83.e1–83.e10, Jul 2010.

[82] Dutt A, Shaikh M, Ganguly T, Nosarti C, Walshe M, Arranz M, Rifkin L, McDonald C, Chaddock CA, McGuire P, Murray RM, Bramon E, and Allin MP. COMT gene polymorphism and corpus callosum morphometry in preterm born adults. *NeuroImage*, 54:148–53, Jan 2011.

[83] Costantine MM, Clark EA, Lai Y, Rouse DJ, Spong CY, Mercer BM, Sorokin Y, Thorp JM Jr, Ramin SM, Malone FD, Carpenter M, Miodovnik M, O'Sullivan MJ, Peaceman AM, and Caritis SN. Association of polymorphisms in neuroprotection and oxidative stress genes and neurodevelopmental outcomes after preterm birth. *Obstetrics and gynecology*, 120:542–50, Sep 2012.

[84] Blair LM, Pickler RH, and Anderson C. Integrative review of genetic factors influencing neurodevelopmental outcomes in preterm infants. *Biological research for nursing*, 18:127–37, Mar 2016.

[85] O'Shea TM, Allred EN, Dammann O, Hirtz D, Kuban KC, Paneth N, and Leviton A and. The ELGAN study of the brain and related disorders in extremely low gestational age newborns. *Early human development*, 85:719–25, Nov 2009.

[86] Heeren T, Joseph RM, Allred EN, O'Shea TM, Leviton A, and Kuban KCK. Cognitive functioning at the age of 10 years among children born extremely preterm: a latent profile approach. *Pediatric research*, 82:614–619, Oct 2017.

[87] Kongsted A and Nielsen AM. Latent class analysis in health research. *Journal of physiotherapy*, 63:55–58, Jan 2017.

[88] Kang HM et al. Epacts, 2016.

[89] Kang HM, Sul JH, Service SK, Zaitlen NA, Kong SY, Freimer NB, Sabatti C, and Eskin E. Variance component model to account for sample structure in genome-wide association studies. *Nature genetics*, 42:348–54, Apr 2010.

[90] Gel B and Serra E. karyoploter: an r/bioconductor package to plot customizable genomes displaying arbitrary data. *Bioinformatics (Oxford, England)*, 33:3088–3090, Oct 2017.

[91] Rentzsch P, Witten D, Cooper GM, Shendure J, and Kircher M. CADD: predicting the deleteriousness of variants throughout the human genome. *Nucleic acids research*, 47:D886–D894, Jan 2019.

[92] Shihab HA, Rogers MF, Gough J, Mort M, Cooper DN, Day IN, Gaunt TR, and Campbell C. An integrative approach to predicting the functional effects of non-coding and coding sequence variation. *Bioinformatics (Oxford, England)*, 31:1536–43, May 2015.

[93] Fishilevich S, Nudel R, Rappaport N, Hadar R, Plaschkes I, Iny Stein T, Rosen N, Kohn A, Twik M, Safran M, Lancet D, and Cohen D. Genehancer: genome-wide integration of enhancers and target genes in genecards. *Database : the journal of biological databases and curation*, 2017, Jan 2017.

[94] Fromer M, Roussos P, Sieberts SK, Johnson JS, Kavanagh DH, Perumal TM, Ruderfer DM, Oh EC, Topol A, Shah HR, Klei LL, Kramer R, Pinto D, Gm ZH, Cicek AE, Dang KK, Browne A, Lu C, Xie L, Readhead B, Stahl EA, Xiao J, Parvizi M, Hamamsy T, Fullard JF, Wang YC, Mahajan MC, Derry JM, Dudley JT, Hemby SE, Logsdon BA, Talbot K, Raj T, Bennett DA, De Jager PL, Zhu J, Zhang B, Sullivan PF, Chess A, Purcell SM, Shinobu LA, Mangravite LM, Toyoshiba H, Gur RE, Hahn CG, Lewis DA, Haroutunian V, Peters MA, Lipska BK, Buxbaum JD, Schadt EE, Hirai K, Roeder K, Brennand KJ, Katsanis N, Domenici E, Devlin B, and Sklar P. Gene expression elucidates functional impact of polygenic risk for schizophrenia. *Nature neuroscience*, 19:1442–1453, Nov 2016.

[95] Yu M, Abnousi A, Zhang Y, Li G, Lee L, Chen Z, Fang R, Lagler TM, Yang Y, Wen J, Sun Q, Li Y, Ren B, and Hu M. Snaphic: a computational pipeline to identify chromatin loops from single-cell hi-c data. *Nature methods*, 18:1056–1059, Sep 2021.

[96] Li X, Lee L, Abnousi A, Yu M, Liu W, Huang L, Li Y, and Hu M. Snaphic2: A computationally efficient loop caller for single cell hi-c data. *Computational and structural biotechnology journal*, 20:2778–2783, 2022.

[97] Juric I, Yu M, Abnousi A, Raviram R, Fang R, Zhao Y, Zhang Y, Qiu Y, Yang Y, Li Y, Ren B, and Hu M. MAPS: model-based analysis of long-range chromatin interactions from plac-seq and hichip experiments. *PLoS computational biology*, 15:e1006982, Apr 2019.

[98] Mumbach MR, Rubin AJ, Flynn RA, Dai C, Khavari PA, Greenleaf WJ, and Chang HY. Hichip: efficient and sensitive analysis of protein-directed genome architecture. *Nature methods*, 13:919–922, Nov 2016.

[99] Mumbach MR, Satpathy AT, Boyle EA, Dai C, Gowen BG, Cho SW, Nguyen ML, Rubin AJ, Granja JM, Kazane KR, Wei Y, Nguyen T, Greenside PG, Corces MR, Tycko J, Simeonov DR, Suliman N, Li R, Xu J, Flynn RA, Kundaje A, Khavari PA, Marson A, Corn JE, Quertermous T, Greenleaf WJ, and Chang HY. Enhancer connectome in primary human cells identifies target genes of disease-associated DNA elements. *Nature genetics*, 49:1602–1612, Nov 2017.

[100] Lee DS, Luo C, Zhou J, Chandran S, Rivkin A, Bartlett A, Nery JR, Fitzpatrick C, O'Connor C, Dixon JR, and Ecker JR. Simultaneous profiling of 3D genome structure and DNA methylation in single human cells. *Nature methods*, 16:999–1006, Oct 2019.

[101] Bryois J, Calini D, Macnair W, Foo L, Urich E, Ortmann W, Iglesias VA, Selvaraj S, Nutma E, Marzin M, Amor S, Williams A, Castelo-Branco G, Menon V, De Jager P, and Malhotra D. Cell-type-specific cis-eqtls in eight human brain cell types identify novel risk genes for psychiatric and neurological disorders. *Nature neuroscience*, 25:1104–1112, Aug 2022.

[102] Carithers LJ, Ardlie K, Barcus M, Branton PA, Britton A, Buia SA, Compton CC, DeLuca DS, Peter-Demchok J, Gelfand ET, Guan P, Korzeniewski GE, Lockhart NC, Rabiner CA, Rao AK, Robinson KL, Roche NV, Sawyer SJ, Segr AV, Shive CE, Smith AM, Sobin LH, Undale AH, Valentino KM, Vaught J, Young TR, and Moore HM and. A novel approach to high-quality postmortem tissue procurement: The gtex project. *Biopreservation and biobanking*, 13:311–9, Oct 2015.

[103] Zhang Y, Sloan SA, Clarke LE, Caneda C, Plaza CA, Blumenthal PD, Vogel H, Steinberg GK, Edwards MS, Li G, Duncan JA 3rd, Cheshier SH, Shuer LM, Chang EF, Grant GA, Gephart MG, and Barres BA. Purification and characterization of progenitor and mature human astrocytes reveals transcriptional and functional differences with mouse. *Neuron*, 89:37–53, Jan 2016.

[104] Jake R Conway, Alexander Lex, and Nils Gehlenborg. Upsetr: an R package for the visualization of intersecting sets and their properties. *Bioinformatics*, 33(18):2938–2940, 6 2017.

[105] Schwartzentruber J, Cooper S, Liu JZ, Barrio-Hernandez I, Bello E, Kumasaka N, Young AMH, Franklin RJM, Johnson T, Estrada K, Gaffney DJ, Beltrao P, and Bassett A. Genome-wide meta-analysis, fine-mapping and integrative prioritization implicate new alzheimer's disease risk genes. *Nature genetics*, 53:392–402, Mar 2021.

[106] Demontis D, Walters RK, Martin J, Mattheisen M, Als TD, Agerbo E, Baldursson G, Belliveau R, Bybjerg-Grauholm J, Bkvad-Hansen M, Cerrato F, Chambert K, Churchhouse C, Dumont A, Eriksson N, Gandal M, Goldstein JI, Grasby KL, Grove J, Gudmundsson OO, Hansen CS, Hauberg ME, Hollegaard MV, Howrigan DP, Huang H, Maller JB, Martin AR, Martin NG, Moran J, Pallesen J, Palmer DS, Pedersen CB, Pedersen MG, Poterba T, Poulsen JB, Ripke S, Robinson EB, Satterstrom FK, Stefansson H, Stevens C, Turley P, Walters GB, Won H, Wright MJ, Andreassen OA, Asherson P, Burton CL, Boomsma DI, Cormand B, Dalsgaard S, Franke B, Gelernter J, Geschwind D, Hakonarson H, Haavik J, Kranzler HR, Kuntsi J, Langley K, Lesch KP, Middeldorp C, Reif A, Rohde LA, Roussos P, Schachar R, Sklar P, Sonuga-Barke EJS, Sullivan PF, Thapar A, Tung JY, Waldman ID, Medland SE, Stefansson K, Nordentoft M, Hougaard DM, Werge T, Mors O, Mortensen PB, Daly MJ, Faraone SV, Brglum AD, and Neale BM. Discovery of the first genome-wide significant risk loci for attention deficit/hyperactivity disorder. *Nature genetics*, 51:63–75, Jan 2019.

[107] Grove J, Ripke S, Als TD, Mattheisen M, Walters RK, Won H, Pallesen J, Agerbo E, Andreassen OA, Anney R, Awashti S, Belliveau R, Bettella F, Buxbaum JD, Bybjerg-Grauholm J, Bkvad-Hansen M, Cerrato F, Chambert K, Christensen JH, Churchhouse C, Dellenvall K, Demontis D, De Rubeis S, Devlin B, Djurovic S, Dumont AL, Goldstein JI, Hansen CS, Hauberg ME, Hollegaard MV, Hope S, Howrigan DP, Huang H, Hultman CM, Klei L, Maller J, Martin J, Martin AR, Moran JL, Nyegaard M, Nrland T, Palmer DS, Palotie A, Pedersen CB, Pedersen MG, dPoterba T, Poulsen JB, Pourcain BS, Qvist P, Rehnstrm K, Reichenberg A, Reichert J, Robinson EB, Roeder K, Roussos P, Saemundsen E, Sandin S, Satterstrom FK, Davey Smith G, Stefansson H, Steinberg S, Stevens CR, Sullivan PF, Turley P, Walters GB, Xu X, Stefansson K, Geschwind DH, Nordentoft M, Hougaard DM, Werge T, Mors O, Mortensen PB, Neale BM, Daly MJ, and Brglum AD. Identification of common genetic risk variants for autism spectrum disorder. *Nature genetics*, 51:431–444, Mar 2019.

[108] Stahl EA, Breen G, Forstner AJ, McQuillin A, Ripke S, Trubetskoy V, Mattheisen M, Wang Y, Coleman JRI, Gaspar HA, de Leeuw CA, Steinberg S, Pavlides JMW, Trzaskowski M, Byrne EM, Pers TH, Holmans PA, Richards AL, Abbott L, Agerbo E, Akil H, Albani D, Alliey-Rodriguez N, Als TD, Anjorin A, Antilla V, Awasthi S, Badner JA, Bkvad-Hansen M, Barchas JD, Bass N, Bauer M, Belliveau R, Bergen SE, Pedersen CB, Ben E, Boks MP, Boocock J, Budde M, Bunney W, Burmeister M, Bybjerg-Grauholm J, Byerley W, Casas M, Cerrato F, Cervantes P, Chambert K, Charney AW, Chen D, Churchhouse C, Clarke TK, Coryell W, Craig DW, Cruceanu C, Curtis D, Czerski PM, Dale AM, de Jong S, Degenhardt F, Del-Favero J, DePaulo JR, Djurovic S, Dobbyn AL, Dumont A, Elvsshagen T, Escott-Price V, Fan CC, Fischer SB, Flickinger M, Foroud TM, Forty L, Frank J, Fraser C, Freimer NB, Frisn L, Gade K, Gage D, Garnham J, Giambartolomei C, Pedersen MG, Goldstein J, Gordon SD, Gordon-Smith K, Green EK, Green MJ, Greenwood TA, Grove J, Guan W, Guzman-Parra J, Hamshere ML, Hautzinger M, Heilbronner U, Herms S, Hipolito M, Hoffmann P, Holland D, Huckins L, Jamain S, Johnson JS, Jurus A, Kandaswamy R, Karlsson R, Kennedy JL, Kittel-Schneider S, Knowles JA, Kogevinas M, Koller AC, Kupka R, Lavebratt C, Lawrence J, Lawson WB, Leber M, Lee PH,

Levy SE, Li JZ, Liu C, Lucae S, Maaser A, MacIntyre DJ, Mahon PB, Maier W, Martinsson L, McCarroll S, McGuffin P, McInnis MG, McKay JD, Medeiros H, Medland SE, Meng F, Milani L, Montgomery GW, Morris DW, Mhleisen TW, Mullins N, Nguyen H, Nievergelt CM, Adolfsson AN, Nwulia EA, O'Donovan C, Loohuis LMO, Ori APS, Oruc L, sby U, Perlis RH, Perry A, Pfennig A, Potash JB, Purcell SM, Regeer EJ, Reif A, Reinbold CS, Rice JP, Rivas F, Rivera M, Roussos P, Ruderfer DM, Ryu E, Snchez-Mora C, Schatzberg AF, Scheftner WA, Schork NJ, Shannon Weickert C, Shehktman T, Shilling PD, Sigurdsson E, Slaney C, Smeland OB, Sobell JL, Sholm Hansen C, Spijker AT, St Clair D, Steffens M, Strauss JS, Streit F, Strohmaier J, Szelinger S, Thompson RC, Thorgeirsson TE, Treutlein J, Vedder H, Wang W, Watson SJ, Weickert TW, Witt SH, Xi S, Xu W, Young AH, Zandi P, Zhang P, Zllner S, Adolfsson R, Agartz I, Alda M, Backlund L, Baune BT, Bellivier F, Berrettini WH, Biernacka JM, Blackwood DHR, Boehnke M, Brglum AD, Corvin A, Craddock N, Daly MJ, Dannlowski U, Esko T, Etain B, Frye M, Fullerton JM, Gershon ES, Gill M, Goes F, Grigoroiu-Serbanescu M, Hauser J, Hougaard DM, Hultman CM, Jones I, Jones LA, Kahn RS, Kirov G, Landn M, Leboyer M, Lewis CM, Li QS, Lissowska J, Martin NG, Mayoral F, McElroy SL, McIntosh AM, McMahon FJ, Melle I, Metspalu A, Mitchell PB, Morken G, Mors O, Mortensen PB, Mller-Myhsok B, Myers RM, Neale BM, Nimgaonkar V, Nordentoft M, Nthen MM, O'Donovan MC, Oedegaard KJ, Owen MJ, Paciga SA, Pato C, Pato MT, Posthuma D, Ramos-Quiroga JA, Ribass M, Rietschel M, Rouleau GA, Schalling M, Schofield PR, Schulze TG, Serretti A, Smoller JW, Stefansson H, Stefansson K, Stordal E, Sullivan PF, Turecki G, Vaaler AE, Vieta E, Vincent JB, Werge T, Nurnberger JI, Wray NR, Di Florio A, Edenberg HJ, Cichon S, Ophoff RA, Scott LJ, Andreassen OA, Kelsoe J, and Sklar P and. Genome-wide association study identifies 30 loci associated with bipolar disorder. *Nature genetics*, 51:793–803, May 2019.

[109] Pardias AF, Holmans P, Pocklington AJ, Escott-Price V, Ripke S, Carrera N, Legge SE, Bishop S, Cameron D, Hamshere ML, Han J, Hubbard L, Lynham A, Mantripragada K, Rees E, MacCabe JH, McCarroll SA, Baune BT, Breen G, Byrne EM, Dannlowski U, Eley TC, Hayward C, Martin NG, McIntosh AM, Plomin R, Porteous DJ, Wray NR, Caballero A, Geschwind DH, Huckins LM, Ruderfer DM, Santiago E, Sklar P, Stahl EA, Won H, Agerbo E, Als TD, Andreassen OA, Bkvad-Hansen M, Mortensen PB, Pedersen CB, Brglum AD, Bybjerg-Grauholm J, Djurovic S, Durmishi N, Pedersen MG, Golimbet V, Grove J, Hougaard DM, Mattheisen M, Molden E, Mors O, Nordentoft M, Pejovic-Milovancevic M, Sigurdsson E, Silagadze T, Hansen CS, Stefansson K, Stefansson H, Steinberg S, Tosato S, Werge T, Collier DA, Rujescu D, Kirov G, Owen MJ, O'Donovan MC, and Walters JTR. Common schizophrenia alleles are enriched in mutation-intolerant genes and in regions under strong background selection. *Nature genetics*, 50:381–389, Mar 2018.

[110] Pankratz N, Beecham GW, DeStefano AL, Dawson TM, Doheny KF, Factor SA, Hamza TH, Hung AY, Hyman BT, Ivinson AJ, Krainc D, Latourelle JC, Clark LN, Marder K, Martin ER, Mayeux R, Ross OA, Scherzer CR, Simon DK, Tanner C, Vance JM, Wszolek ZK, Zabetian CP, Myers RH, Payami H, Scott WK, and Foroud T and. Meta-analysis of

parkinson's disease: identification of a novel locus, RIT2. *Annals of neurology*, 71:370–84, Mar 2012.

[111] Howard DM, Adams MJ, Clarke TK, Hafferty JD, Gibson J, Shirali M, Coleman JRI, Hagenaars SP, Ward J, Wigmore EM, Alloza C, Shen X, Barbu MC, Xu EY, Whalley HC, Marioni RE, Porteous DJ, Davies G, Deary IJ, Hemani G, Berger K, Teismann H, Rawal R, Arolt V, Baune BT, Dannlowski U, Domschke K, Tian C, Hinds DA, Trzaskowski M, Byrne EM, Ripke S, Smith DJ, Sullivan PF, Wray NR, Breen G, Lewis CM, and McIntosh AM. Genome-wide meta-analysis of depression identifies 102 independent variants and highlights the importance of the prefrontal brain regions. *Nature neuroscience*, 22:343–352, Mar 2019.

[112] Nagel M, Jansen PR, Stringer S, Watanabe K, de Leeuw CA, Bryois J, Savage JE, Hammerschlag AR, Skene NG, Muoz-Manchado AB, White T, Tiemeier H, Linnarsson S, Hjerling-Leffler J, Polderman TJC, Sullivan PF, van der Sluis S, and Posthuma D. Meta-analysis of genome-wide association studies for neuroticism in 449,484 individuals identifies novel genetic loci and pathways. *Nature genetics*, 50:920–927, Jul 2018.

[113] Lee JJ, Wedow R, Okbay A, Kong E, Maghzian O, Zacher M, Nguyen-Viet TA, Bowers P, Sidorenko J, Karlsson Linnr R, Fontana MA, Kundu T, Lee C, Li H, Li R, Royer R, Timshel PN, Walters RK, Willoughby EA, Yengo L, Alver M, Bao Y, Clark DW, Day FR, Furlotte NA, Joshi PK, Kemper KE, Kleinman A, Langenberg C, Mgi R, Trampush JW, Verma SS, Wu Y, Lam M, Zhao JH, Zheng Z, Boardman JD, Campbell H, Freese J, Harris KM, Hayward C, Herd P, Kumari M, Lencz T, Luan J, Malhotra AK, Metspalu A, Milani L, Ong KK, Perry JRB, Porteous DJ, Ritchie MD, Smart MC, Smith BH, Tung JY, Wareham NJ, Wilson JF, Beauchamp JP, Conley DC, Esko T, Lehrer SF, Magnusson PKE, Oskarsson S, Pers TH, Robinson MR, Thom K, Watson C, Chabris CF, Meyer MN, Laibson DI, Yang J, Johannesson M, Koellinger PD, Turley P, Visscher PM, Benjamin DJ, and Cesarini D. Gene discovery and polygenic prediction from a genome-wide association study of educational attainment in 1.1 million individuals. *Nature genetics*, 50:1112–1121, Jul 2018.

[114] Savage JE, Jansen PR, Stringer S, Watanabe K, Bryois J, de Leeuw CA, Nagel M, Awasthi S, Barr PB, Coleman JRI, Grasby KL, Hammerschlag AR, Kaminski JA, Karlsson R, Krapohl E, Lam M, Nygaard M, Reynolds CA, Trampush JW, Young H, Zabaneh D, Hgg S, Hansell NK, Karlsson IK, Linnarsson S, Montgomery GW, Muoz-Manchado AB, Quinlan EB, Schumann G, Skene NG, Webb BT, White T, Arking DE, Avramopoulos D, Bilder RM, Bitsios P, Burdick KE, Cannon TD, Chiba-Falek O, Christoforou A, Cirulli ET, Congdon E, Corvin A, Davies G, Deary IJ, DeRosse P, Dickinson D, Djurovic S, Donohoe G, Conley ED, Eriksson JG, Espeseth T, Freimer NA, Giakoumaki S, Giegling I, Gill M, Glahn DC, Hariri AR, Hatzimanolis A, Keller MC, Knowles E, Koltai D, Konte B, Lahti J, Le Hellard S, Lencz T, Liewald DC, London E, Lundervold AJ, Malhotra AK, Melle I, Morris D, Need AC, Ollier W, Palotie A, Payton A, Pendleton N, Poldrack RA, Rikknen K, Reinvang I, Roussos P, Rujescu D, Sabb FW, Scult MA, Smeland OB, Smyrnis N, Starr JM, Steen VM, Stefanis NC, Straub RE, Sundet K, Tiemeier H, Voineskos AN, Weinberger DR, Widen E, Yu J, Abecasis G, Andreassen OA, Breen G, Christiansen L,

Debrabant B, Dick DM, Heinz A, Hjerling-Leffler J, Ikram MA, Kendler KS, Martin NG, Medland SE, Pedersen NL, Plomin R, Polderman TJC, Ripke S, van der Sluis S, Sullivan PF, Vrieze SI, Wright MJ, and Posthuma D. Genome-wide association meta-analysis in 269,867 individuals identifies new genetic and functional links to intelligence. *Nature genetics*, 50:912–919, Jul 2018.

[115] Nott A, Holtman IR, Coufal NG, Schlachetzki JCM, Yu M, Hu R, Han CZ, Pena M, Xiao J, Wu Y, Keulen Z, Pasillas MP, O'Connor C, Nickl CK, Schafer ST, Shen Z, Rissman RA, Brewer JB, Gosselin D, Gonda DD, Levy ML, Rosenfeld MG, McVicker G, Gage FH, Ren B, and Glass CK. Brain cell type-specific enhancer-promoter interactome maps and disease¡b¿-¡/b¿risk association. *Science (New York, N.Y.)*, 366:1134–1139, Nov 2019.

[116] Mollereau C, Simons MJ, Soularue P, Liners F, Vassart G, Meunier JC, and Parmentier M. Structure, tissue distribution, and chromosomal localization of the prepronociceptin gene. *Proceedings of the National Academy of Sciences of the United States of America*, 93:8666–70, Aug 1996.

[117] Darland T, Heinricher MM, and Grandy DK. Orphanin fq/nociceptin: a role in pain and analgesia, but so much more. *Trends in neurosciences*, 21:215–21, May 1998.

[118] Girgenti MJ, Wang J, Ji D, Cruz DA, Stein MB, Gelernter J, Young KA, Huber BR, Williamson DE, Friedman MJ, Krystal JH, Zhao H, and Duman RS. Transcriptomic organization of the human brain in post-traumatic stress disorder. *Nature neuroscience*, 24:24–33, Jan 2021.

[119] Ankur Sahu, Hussain Ahmed Chowdhury, Mithil Gaikwad, Chen Chongtham, Uddip Talukdar, Jadab Kishor Phukan, Dhruba Kumar Bhattacharyya, and Pankaj Barah. Integrative network analysis identifies differential regulation of neuroimmune system in schizophrenia and bipolar disorder. *Brain, Behavior, amp; Immunity - Health*, 2:100023, 2 2020.

[120] Chen WT, Lu A, Craessaerts K, Pavie B, Sala Frigerio C, Corthout N, Qian X, Lalkov J, Khnemund M, Voytyuk I, Wolfs L, Mancuso R, Salta E, Balusu S, Snellinx A, Munck S, Jurek A, Fernandez Navarro J, Saido TC, Huitinga I, Lundeberg J, Fiers M, and De Strooper B. Spatial transcriptomics and in situ sequencing to study alzheimer's disease. *Cell*, 182:976–991.e19, Aug 2020.

[121] Wang JY, Li XY, Li HJ, Liu JW, Yao YG, Li M, Xiao X, and Luo XJ. Integrative analyses followed by functional characterization reveal TMEM180 as a schizophrenia risk gene. *Schizophrenia bulletin*, 47:1364–1374, Aug 2021.

[122] Yengo L, Sidorenko J, Kemper KE, Zheng Z, Wood AR, Weedon MN, Frayling TM, Hirschhorn J, Yang J, and Visscher PM and. Meta-analysis of genome-wide association studies for height and body mass index in 700000 individuals of european ancestry. *Human molecular genetics*, 27:3641–3649, Oct 2018.

[123] Vuckovic D, Bao EL, Akbari P, Lareau CA, Mousas A, Jiang T, Chen MH, Raffield LM, Tardaguila M, Huffman JE, Ritchie SC, Megy K, Ponstingl H, Penkett CJ, Albers PK, Wigdor EM, Sakaue S, Moscati A, Manansala R, Lo KS, Qian H, Akiyama M, Bartz TM, Ben-Shlomo Y, Beswick A, Bork-Jensen J, Bottinger EP, Brody JA, van Rooij FJA, Chitrala KN, Wilson PWF, Choquet H, Danesh J, Di Angelantonio E, Dimou N, Ding J, Elliott P, Esko T, Evans MK, Felix SB, Floyd JS, Broer L, Grarup N, Guo MH, Guo Q, Greinacher A, Haessler J, Hansen T, Howson JMM, Huang W, Jorgenson E, Kacprowski T, Khnen M, Kamatani Y, Kanai M, Karthikeyan S, Koskeridis F, Lange LA, Lehtimki T, Linneberg A, Liu Y, Lyytikinen LP, Manichaikul A, Matsuda K, Mohlke KL, Mononen N, Murakami Y, Nadkarni GN, Nikus K, Pankratz N, Pedersen O, Preuss M, Psaty BM, Raitakari OT, Rich SS, Rodriguez BAT, Rosen JD, Rotter JI, Schubert P, Spracklen CN, Surendran P, Tang H, Tardif JC, Ghanbari M, Vlker U, Vlzke H, Watkins NA, Weiss S, Cai N, Kundu K, Watt SB, Walter K, Zonderman AB, Cho K, Li Y, Loos RJF, Knight JC, Georges M, Stegle O, Evangelou E, Okada Y, Roberts DJ, Inouye M, Johnson AD, Auer PL, Astle WJ, Reiner AP, Butterworth AS, Ouwehand WH, Lettre G, Sankaran VG, and Soranzo N. The polygenic and monogenic basis of blood traits and diseases. *Cell*, 182:1214–1231.e11, Sep 2020.

[124] Finucane HK, Bulik-Sullivan B, Gusev A, Trynka G, Reshef Y, Loh PR, Anttila V, Xu H, Zang C, Farh K, Ripke S, Day FR, Purcell S, Stahl E, Lindstrom S, Perry JR, Okada Y, Raychaudhuri S, Daly MJ, Patterson N, Neale BM, and Price AL. Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nature genetics*, 47:1228–35, Nov 2015.

[125] Bryois J, Skene NG, Hansen TF, Kogelman LJA, Watson HJ, Liu Z, Brueggeman L, Breen G, Bulik CM, Arenas E, Hjerling-Leffler J, and Sullivan PF. Genetic identification of cell types underlying brain complex traits yields insights into the etiology of parkinson's disease. *Nature genetics*, 52:482–493, May 2020.

[126] Zhou J, Ma J, Chen Y, Cheng C, Bao B, Peng J, Sejnowski TJ, Dixon JR, and Ecker JR. Robust single-cell hi-c clustering by convolution- and random-walk-based imputation. *Proceedings of the National Academy of Sciences of the United States of America*, 116:14011–14018, Jul 2019.

[127] Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, Korbel JO, Marchini JL, McCarthy S, McVean GA, and Abecasis GR. A global reference for human genetic variation. 526:68–74, Oct 2015.

[128] The international hapmap project. 426:789–96, Dec 2003.

[129] Taliun D, Harris DN, Kessler MD, Carlson J, Szpiech ZA, Torres R, Taliun SAG, Corvelo A, Gogarten SM, Kang HM, Pitsillides AN, LeFaive J, Lee SB, Tian X, Browning BL, Das S, Emde AK, Clarke WE, Loesch DP, Shetty AC, Blackwell TW, Smith AV, Wong Q, Liu X, Conomos MP, Bobo DM, Aguet F, Albert C, Alonso A, Ardlie KG, Arking DE, Aslibekyan S, Auer PL, Barnard J, Barr RG, Barwick L, Becker LC, Beer RL, Benjamin EJ, Bielak LF, Blangero J, Boehnke M, Bowden DW, Brody JA, Burchard EG, Cade BE, Casella JF, Chalazan B, Chasman DI, Chen YI, Cho MH, Choi SH, Chung MK, Clish CB,

Correa A, Curran JE, Custer B, Darbar D, Daya M, de Andrade M, DeMeo DL, Dutcher SK, Ellinor PT, Emery LS, Eng C, Fatkin D, Fingerlin T, Forer L, Fornage M, Franceschini N, Fuchsberger C, Fullerton SM, Germer S, Gladwin MT, Gottlieb DJ, Guo X, Hall ME, He J, Heard-Costa NL, Heckbert SR, Irvin MR, Johnsen JM, Johnson AD, Kaplan R, Kardia SLR, Kelly T, Kelly S, Kenny EE, Kiel DP, Klemmer R, Konkle BA, Kooperberg C, Kttgen A, Lange LA, Lasky-Su J, Levy D, Lin X, Lin KH, Liu C, Loos RJF, Garman L, Gerszten R, Lubitz SA, Lunetta KL, Mak ACY, Manichaikul A, Manning AK, Mathias RA, McManus DD, McGarvey ST, Meigs JB, Meyers DA, Mikulla JL, Minear MA, Mitchell BD, Mohanty S, Montasser ME, Montgomery C, Morrison AC, Murabito JM, Natale A, Natarajan P, Nelson SC, North KE, O'Connell JR, Palmer ND, Pankratz N, Peloso GM, Peyser PA, Pleiness J, Post WS, Psaty BM, Rao DC, Redline S, Reiner AP, Roden D, Rotter JI, Ruczinski I, Sarnowski C, Schoenherr S, Schwartz DA, Seo JS, Seshadri S, Sheehan VA, Sheu WH, Shoemaker MB, Smith NL, Smith JA, Sotoodehnia N, Stilp AM, Tang W, Taylor KD, Telen M, Thornton TA, Tracy RP, Van Den Berg DJ, Vasan RS, Viaud-Martinez KA, Vrieze S, Weeks DE, Weir BS, Weiss ST, Weng LC, Willer CJ, Zhang Y, Zhao X, Arnett DK, Ashley-Koch AE, Barnes KC, Boerwinkle E, Gabriel S, Gibbs R, Rice KM, Rich SS, Silverman EK, Qasba P, Gan W, Papanicolaou GJ, Nickerson DA, Browning SR, Zody MC, Zllner S, Wilson JG, Cupples LA, Laurie CC, Jaquish CE, Hernandez RD, O'Connor TD, and Abecasis GR. Sequencing of 53,831 diverse genomes from the NHLBI topmed program. *Nature*, 590:290–299, Feb 2021.

[130] Huang L, Rosen JD, Sun Q, Chen J, Wheeler MM, Zhou Y, Min YI, Kooperberg C, Conomos MP, Stilp AM, Rich SS, Rotter JI, Manichaikul A, Loos RJF, Kenny EE, Blackwell TW, Smith AV, Jun G, Sedlazeck FJ, Metcalf G, Boerwinkle E, Raffield LM, Reiner AP, Auer PL, and Li Y. TOP-LD: A tool to explore linkage disequilibrium with topmed whole-genome sequence data. 109:1175–1181, Jun 2022.

[131] Lambert JC, Ibrahim-Verbaas CA, Harold D, Naj AC, Sims R, Bellenguez C, DeStafano AL, Bis JC, Beecham GW, Grenier-Boley B, Russo G, Thorton-Wells TA, Jones N, Smith AV, Chouraki V, Thomas C, Ikram MA, Zelenika D, Vardarajan BN, Kamatani Y, Lin CF, Gerrish A, Schmidt H, Kunkle B, Dunstan ML, Ruiz A, Bihoreau MT, Choi SH, Reitz C, Pasquier F, Cruchaga C, Craig D, Amin N, Berr C, Lopez OL, De Jager PL, Deramecourt V, Johnston JA, Evans D, Lovestone S, Letenneur L, Morn FJ, Rubinsztein DC, Eiriksdottir G, Sleegers K, Goate AM, Fivet N, Huentelman MW, Gill M, Brown K, Kamboh MI, Keller L, Barberger-Gateau P, McGuiness B, Larson EB, Green R, Myers AJ, Dufouil C, Todd S, Wallon D, Love S, Rogaeva E, Gallacher J, St George-Hyslop P, Clarimon J, Lleo A, Bayer A, Tsuang DW, Yu L, Tsolaki M, Boss P, Spalletta G, Proitsi P, Collinge J, Sorbi S, Sanchez-Garcia F, Fox NC, Hardy J, Deniz Naranjo MC, Bosco P, Clarke R, Brayne C, Galimberti D, Mancuso M, Matthews F, Moebus S, Mecocci P, Del Zompo M, Maier W, Hampel H, Pilotto A, Bullido M, Panza F, Caffarra P, Nacmias B, Gilbert JR, Mayhaus M, Lannefelt L, Hakonarson H, Pichler S, Carrasquillo MM, Ingelsson M, Beekly D, Alvarez V, Zou F, Valladares O, Younkin SG, Coto E, Hamilton-Nelson KL, Gu W, Razquin C, Pastor P, Mateo I, Owen MJ, Faber KM, Jonsson PV, Combarros O, O'Donovan MC, Cantwell LB, Soininen H, Blacker D, Mead S, Mosley TH Jr, Bennett DA, Harris TB, Fratiglioni L, Holmes C, de Bruijn RF, Passmore P, Montine TJ, Bettens K, Rotter JI,

Brice A, Morgan K, Foroud TM, Kukull WA, Hannequin D, Powell JF, Nalls MA, Ritchie K, Lunetta KL, Kauwe JS, Boerwinkle E, Riemenschneider M, Boada M, Hiltuenen M, Martin ER, Schmidt R, Rujescu D, Wang LS, Dartigues JF, Mayeux R, Tzourio C, Hofman A, Nthen MM, Graff C, Psaty BM, Jones L, Haines JL, Holmans PA, Lathrop M, Pericak-Vance MA, Launer LJ, Farrer LA, van Duijn CM, Van Broeckhoven C, Moskvina V, Seshadri S, Williams J, Schellenberg GD, and Amouyel P. Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for alzheimer's disease. 45:1452–8, Dec 2013.

[132] Identification of risk loci with shared effects on five major psychiatric disorders: a genome-wide analysis. 381:1371–1379, Apr 2013.

[133] Okbay A, Beauchamp JP, Fontana MA, Lee JJ, Pers TH, Rietveld CA, Turley P, Chen GB, Emilsson V, Meddens SF, Oskarsson S, Pickrell JK, Thom K, Timshel P, de Vlaming R, Abdellaoui A, Ahluwalia TS, Bacelis J, Baumbach C, Bjornsdottir G, Brandsma JH, Pina Concas M, Derringer J, Furlotte NA, Galesloot TE, Girotto G, Gupta R, Hall LM, Harris SE, Hofer E, Horikoshi M, Huffman JE, Kaasik K, Kalafati IP, Karlsson R, Kong A, Lahti J, van der Lee SJ, deLeeuw C, Lind PA, Lindgren KO, Liu T, Mangino M, Marten J, Mihailov E, Miller MB, van der Most PJ, Oldmeadow C, Payton A, Pervjakova N, Peyrot WJ, Qian Y, Raitakari O, Rueedi R, Salvi E, Schmidt B, Schraut KE, Shi J, Smith AV, Poot RA, St Pourcain B, Teumer A, Thorleifsson G, Verweij N, Vuckovic D, Wellmann J, Westra HJ, Yang J, Zhao W, Zhu Z, Alizadeh BZ, Amin N, Bakshi A, Baumeister SE, Biino G, Bnnelykke K, Boyle PA, Campbell H, Cappuccio FP, Davies G, De Neve JE, Deloukas P, Demuth I, Ding J, Eibich P, Eisele L, Eklund N, Evans DM, Faul JD, Feitosa MF, Forstner AJ, Gandin I, Gunnarsson B, Halldrsson BV, Harris TB, Heath AC, Hocking LJ, Holliday EG, Homuth G, Horan MA, Hottenga JJ, de Jager PL, Joshi PK, Jugessur A, Kaakinen MA, Khnen M, Kanoni S, Keltigangas-Jrvinen L, Kiemeney LA, Kolcic I, Koskinen S, Kraja AT, Kroh M, Kutalik Z, Latvala A, Launer LJ, Lebreton MP, Levinson DF, Lichtenstein P, Lichtner P, Liewald DC, Loukola A, Madden PA, Mgi R, Mki-Opas T, Marioni RE, Marques-Vidal P, Meddens GA, McMahon G, Meisinger C, Meitinger T, Milaneschi Y, Milani L, Montgomery GW, Myhre R, Nelson CP, Nyholt DR, Ollier WE, Palotie A, Paternoster L, Pedersen NL, Petrovic KE, Porteous DJ, Rikknen K, Ring SM, Robino A, Rostapshova O, Rudan I, Rustichini A, Salomaa V, Sanders AR, Sarin AP, Schmidt H, Scott RJ, Smith BH, Smith JA, Staessen JA, Steinhagen-Thiessen E, Strauch K, Terracciano A, Tobin MD, Ulivi S, Vaccargiu S, Quaye L, van Rooij FJ, Venturini C, Vinkhuyzen AA, Vlker U, Vlzke H, Vonk JM, Vozzi D, Waage J, Ware EB, Willemsen G, Attia JR, Bennett DA, Berger K, Bertram L, Bisgaard H, Boomsma DI, Borecki IB, Bltmann U, Chabris CF, Cucca F, Cusi D, Deary IJ, Dedoussis GV, van Duijn CM, Eriksson JG, Franke B, Franke L, Gasparini P, Gejman PV, Gieger C, Grabe HJ, Gratten J, Groenen PJ, Gudnason V, van der Harst P, Hayward C, Hinds DA, Hoffmann W, Hyppnen E, Iacono WG, Jacobsson B, Jrvelin MR, Jckel KH, Kaprio J, Kardia SL, Lehtimki T, Lehrer SF, Magnusson PK, Martin NG, McGue M, Metspalu A, Pendleton N, Penninx BW, Perola M, Pirastu N, Pirastu M, Polasek O, Posthuma D, Power C, Province MA, Samani NJ, Schlessinger D, Schmidt R, Srensen TI, Spector TD, Stefansson K, Thorsteinsdottir U, Thurik AR, Timpson NJ, Tiemeier H, Tung JY, Uitterlinden AG, Vitart V, Vollenweider P, Weir

DR, Wilson JF, Wright AF, Conley DC, Krueger RF, Davey Smith G, Hofman A, Laibson DI, Medland SE, Meyer MN, Johannesson M, Visscher PM, Esko T, Koellinger PD, Cesarini D, and Benjamin DJ. Genome-wide association study identifies 74 loci associated with educational attainment. 533:539–42, May 2016.

[134] Astle WJ, Elding H, Jiang T, Allen D, Ruklisa D, Mann AL, Mead D, Bouman H, Riveros-Mckay F, Kostadima MA, Lambourne JJ, Sivapalaratnam S, Downes K, Kundu K, Bomba L, Berentsen K, Bradley JR, Daugherty LC, Delaneau O, Freson K, Garner SF, Grassi L, Guerrero J, Haimel M, Janssen-Megens EM, Kaan A, Kamat M, Kim B, Mandoli A, Marchini J, Martens JHA, Meacham S, Megy K, O'Connell J, Petersen R, Sharifi N, Sheard SM, Staley JR, Tuna S, van der Ent M, Walter K, Wang SY, Wheeler E, Wilder SP, Iotchkova V, Moore C, Sambrook J, Stunnenberg HG, Di Angelantonio E, Kaptoge S, Kuijpers TW, Carrillo de Santa-Pau E, Juan D, Rico D, Valencia A, Chen L, Ge B, Vasquez L, Kwan T, Garrido-Martn D, Watt S, Yang Y, Guigo R, Beck S, Paul DS, Pastinen T, Bujold D, Bourque G, Frontini M, Danesh J, Roberts DJ, Ouwehand WH, Butterworth AS, and Soranzo N. The allelic landscape of human blood cell trait variation and links to common complex disease. 167:1415–1429.e19, Nov 2016.

[135] Howard DM, Adams MJ, Shirali M, Clarke TK, Marioni RE, Davies G, Coleman JRI, Alloza C, Shen X, Barbu MC, Wigmore EM, Gibson J, Hagenaars SP, Lewis CM, Ward J, Smith DJ, Sullivan PF, Haley CS, Breen G, Deary IJ, and McIntosh AM. Genome-wide association study of depression phenotypes in UK biobank identifies variants in excitatory synaptic pathways. 9:1470, Apr 2018.

[136] Kunkle BW, Grenier-Boley B, Sims R, Bis JC, Damotte V, Naj AC, Boland A, Vronskaya M, van der Lee SJ, Amlie-Wolf A, Bellenguez C, Frizatti A, Chouraki V, Martin ER, Sleegers K, Badarinarayan N, Jakobsdottir J, Hamilton-Nelson KL, Moreno-Grau S, Olaso R, Raybould R, Chen Y, Kuzma AB, Hiltunen M, Morgan T, Ahmad S, Vardarajan BN, Epelbaum J, Hoffmann P, Boada M, Beecham GW, Garnier JG, Harold D, Fitzpatrick AL, Valladares O, Moutet ML, Gerrish A, Smith AV, Qu L, Bacq D, Denning N, Jian X, Zhao Y, Del Zompo M, Fox NC, Choi SH, Mateo I, Hughes JT, Adams HH, Malamon J, Sanchez-Garcia F, Patel Y, Brody JA, Dombroski BA, Naranjo MCD, Daniilidou M, Eiriksdottir G, Mukherjee S, Wallon D, Uphill J, Aspelund T, Cantwell LB, Garzia F, Galimberti D, Hofer E, Butkiewicz M, Fin B, Scarpini E, Sarnowski C, Bush WS, Meslage S, Kornhuber J, White CC, Song Y, Barber RC, Engelborghs S, Sordon S, Voijnovic D, Adams PM, Vandenberghe R, Mayhaus M, Cupples LA, Albert MS, De Deyn PP, Gu W, Himali JJ, Beekly D, Squassina A, Hartmann AM, Orellana A, Blacker D, Rodriguez-Rodriguez E, Lovestone S, Garcia ME, Doody RS, Munoz-Fernadez C, Sussams R, Lin H, Fairchild TJ, Benito YA, Holmes C, Karamuji omi H, Frosch MP, Thonberg H, Maier W, Roshchupkin G, Ghetti B, Giedraitis V, Kawalia A, Li S, Huebinger RM, Kilander L, Moebus S, Hernndez I, Kamboh MI, Brundin R, Turton J, Yang Q, Katz MJ, Concari L, Lord J, Beiser AS, Keene CD, Helisalmi S, Kloszewska I, Kukull WA, Koivisto AM, Lynch A, Tarraga L, Larson EB, Haapasalo A, Lawlor B, Mosley TH, Lipton RB, Solfrizzi V, Gill M, Longstreth WT Jr, Montine TJ, Frisardi V, Diez-Fairen M, Rivadeneira F, Petersen RC, Deramecourt V, Alvarez I, Salani F, Ciaramella A, Boerwinkle E, Reiman EM, Fievet N, Rotter JI, Reisch JS, Hanon O, Cupidi C, Andre Uitterlinden AG, Royall DR, Dufouil

C, Maletta RG, de Rojas I, Sano M, Brice A, Cecchetti R, George-Hyslop PS, Ritchie K, Tsolaki M, Tsuang DW, Dubois B, Craig D, Wu CK, Soininen H, Avramidou D, Albin RL, Fratiglioni L, Germanou A, Apostolova LG, Keller L, Koutroumani M, Arnold SE, Panza F, Gkatzima O, Asthana S, Hannequin D, Whitehead P, Atwood CS, Caffarra P, Hampel H, Quintela I, Carracedo , Lannfelt L, Rubinsztein DC, Barnes LL, Pasquier F, Frlich L, Barral S, McGuinness B, Beach TG, Johnston JA, Becker JT, Passmore P, Bigio EH, Schott JM, Bird TD, Warren JD, Boeve BF, Lupton MK, Bowen JD, Proitsi P, Boxer A, Powell JF, Burke JR, Kauwe JSK, Burns JM, Mancuso M, Buxbaum JD, Bonuccelli U, Cairns NJ, McQuillin A, Cao C, Livingston G, Carlson CS, Bass NJ, Carlsson CM, Hardy J, Carney RM, Bras J, Carrasquillo MM, Guerreiro R, Allen M, Chui HC, Fisher E, Masullo C, Crocco EA, DeCarli C, Bisceglio G, Dick M, Ma L, Duara R, Graff-Radford NR, Evans DA, Hodges A, Faber KM, Scherer M, Fallon KB, Riemenschneider M, Fardo DW, Heun R, Farlow MR, Klsch H, Ferris S, Leber M, Foroud TM, Heuser I, Galasko DR, Giegling I, Gearing M, Hll M, Geschwind DH, Gilbert JR, Morris J, Green RC, Mayo K, Growdon JH, Feulner T, Hamilton RL, Harrell LE, Drichel D, Honig LS, Cushion TD, Huentelman MJ, Hollingworth P, Hulette CM, Hyman BT, Marshall R, Jarvik GP, Meggy A, Abner E, Menzies GE, Jin LW, Leonenko G, Real LM, Jun GR, Baldwin CT, Grozeva D, Karydas A, Russo G, Kaye JA, Kim R, Jessen F, Kowall NW, Vellas B, Kramer JH, Vardy E, LaFerla FM, Jckel KH, Lah JJ, Dichgans M, Leverenz JB, Mann D, Levey AI, Pickering-Brown S, Lieberman AP, Klopp N, Lunetta KL, Wichmann HE, Lyketsos CG, Morgan K, Marson DC, Brown K, Martiniuk F, Medway C, Mash DC, Nthen MM, Masliah E, Hooper NM, McCormick WC, Daniele A, McCurry SM, Bayer A, McDavid AN, Gallacher J, McKee AC, van den Bussche H, Mesulam M, Brayne C, Miller BL, Riedel-Heller S, Miller CA, Miller JW, Al-Chalabi A, Morris JC, Shaw CE, Myers AJ, Wiltfang J, O'Bryant S, Olichney JM, Alvarez V, Parisi JE, Singleton AB, Paulson HL, Collinge J, Perry WR, Mead S, Peskind E, Cribbs DH, Rossor M, Pierce A, Ryan NS, Poon WW, Nacmias B, Potter H, Sorbi S, Quinn JF, Sacchinelli E, Raj A, Spalletta G, Raskind M, Caltagirone C, Boss P, Orfei MD, Reisberg B, Clarke R, Reitz C, Smith AD, Ringman JM, Warden D, Roberson ED, Wilcock G, Rogaeva E, Bruni AC, Rosen HJ, Gallo M, Rosenberg RN, Ben-Shlomo Y, Sager MA, Mecocci P, Saykin AJ, Pastor P, Cuccaro ML, Vance JM, Schneider JA, Schneider LS, Slifer S, Seeley WW, Smith AG, Sonnen JA, Spina S, Stern RA, Swerdlow RH, Tang M, Tanzi RE, Trojanowski JQ, Troncoso JC, Van Deerlin VM, Van Eldik LJ, Vinters HV, Vonsattel JP, Weintraub S, Welsh-Bohmer KA, Wilhelmsen KC, Williamson J, Wingo TS, Woltjer RL, Wright CB, Yu CE, Yu L, Saba Y, Pilotto A, Bullido MJ, Peters O, Crane PK, Bennett D, Bosco P, Coto E, Boccardi V, De Jager PL, Lleo A, Warner N, Lopez OL, Ingelsson M, Deloukas P, Cruchaga C, Graff C, Gwilliam R, Fornage M, Goate AM, Sanchez-Juan P, Kehoe PG, Amin N, Ertekin-Taner N, Berr C, Debette S, Love S, Launer LJ, Younkin SG, Dartigues JF, Corcoran C, Ikram MA, Dickson DW, Nicolas G, Campion D, Tschanz J, Schmidt H, Hakonarson H, Clarimon J, Munger R, Schmidt R, Farrer LA, Van Broeckhoven C, C O'Donovan M, DeStefano AL, Jones L, Haines JL, Deleuze JF, Owen MJ, Gudnason V, Mayeux R, Escott-Price V, Psaty BM, Ramirez A, Wang LS, Ruiz A, van Duijn CM, Holmans PA, Seshadri S, Williams J, Amouyel P, Schellenberg GD, Lambert JC, and Pericak-Vance MA and. Genetic meta-analysis of diagnosed alzheimer's disease identifies new risk loci and implicates a, tau, immunity and lipid processing. 51:414–430, Mar 2019.

[137] Chen MH, Raffield LM, Mousas A, Sakaue S, Huffman JE, Moscati A, Trivedi B, Jiang T, Akbari P, Vuckovic D, Bao EL, Zhong X, Manansala R, Laplante V, Chen M, Lo KS, Qian H, Lareau CA, Beaudoin M, Hunt KA, Akiyama M, Bartz TM, Ben-Shlomo Y, Beswick A, Bork-Jensen J, Bottinger EP, Brody JA, van Rooij FJA, Chitrala K, Cho K, Choquet H, Correa A, Danesh J, Di Angelantonio E, Dimou N, Ding J, Elliott P, Esko T, Evans MK, Floyd JS, Broer L, Grarup N, Guo MH, Greinacher A, Haessler J, Hansen T, Howson JMM, Huang QQ, Huang W, Jorgenson E, Kacprowski T, Khnen M, Kamatani Y, Kanai M, Karthikeyan S, Koskeridis F, Lange LA, Lehtimki T, Lerch MM, Linneberg A, Liu Y, Lyytikinen LP, Manichaikul A, Martin HC, Matsuda K, Mohlke KL, Mononen N, Murakami Y, Nadkarni GN, Nauck M, Nikus K, Ouwehand WH, Pankratz N, Pedersen O, Preuss M, Psaty BM, Raitakari OT, Roberts DJ, Rich SS, Rodriguez BAT, Rosen JD, Rotter JI, Schubert P, Spracklen CN, Surendran P, Tang H, Tardif JC, Trembath RC, Ghanbari M, Vlker U, Vlzke H, Watkins NA, Zonderman AB, Wilson PWF, Li Y, Butterworth AS, Gauchat JF, Chiang CWK, Li B, Loos RJF, Astle WJ, Evangelou E, van Heel DA, Sankaran VG, Okada Y, Soranzo N, Johnson AD, Reiner AP, Auer PL, and Lettre G. Trans-ethnic and ancestry-specific blood-cell genetics in 746,667 individuals from 5 global populations. 182:1198–1213.e14, Sep 2020.

[138] P. A. Knight and D. Ruiz. A fast algorithm for matrix balancing. *IMA Journal of Numerical Analysis*, 33(3):1029–1047, 10 2012.

[139] Kundaje A, Meuleman W, Ernst J, Bilenky M, Yen A, Heravi-Moussavi A, Kheradpour P, Zhang Z, Wang J, Ziller MJ, Amin V, Whitaker JW, Schultz MD, Ward LD, Sarkar A, Quon G, Sandstrom RS, Eaton ML, Wu YC, Pfenning AR, Wang X, Claussnitzer M, Liu Y, Coarfa C, Harris RA, Shoresh N, Epstein CB, Gjoneska E, Leung D, Xie W, Hawkins RD, Lister R, Hong C, Gascard P, Mungall AJ, Moore R, Chuah E, Tam A, Canfield TK, Hansen RS, Kaul R, Sabo PJ, Bansal MS, Carles A, Dixon JR, Farh KH, Feizi S, Karlic R, Kim AR, Kulkarni A, Li D, Lowdon R, Elliott G, Mercer TR, Neph SJ, Onuchic V, Polak P, Rajagopal N, Ray P, Sallari RC, Siebenthall KT, Sinnott-Armstrong NA, Stevens M, Thurman RE, Wu J, Zhang B, Zhou X, Beaudet AE, Boyer LA, De Jager PL, Farnham PJ, Fisher SJ, Haussler D, Jones SJ, Li W, Marra MA, McManus MT, Sunyaev S, Thomson JA, Tlsty TD, Tsai LH, Wang W, Waterland RA, Zhang MQ, Chadwick LH, Bernstein BE, Costello JF, Ecker JR, Hirst M, Meissner A, Milosavljevic A, Ren B, Stamatoyannopoulos JA, Wang T, and Kellis M. Integrative analysis of 111 reference human epigenomes. *Nature*, 518:317–30, Feb 2015.

[140] Li D, Hsu S, Purushotham D, Sears RL, and Wang T. Washu epigenome browser update 2019. *Nucleic acids research*, 47:W158–W165, Jul 2019.

[141] Adn U, Lin A, Carlo W, Leviton A, Murray JC, Hallman M, Lifton RP, Zhang H, and Ment LR and. Candidate gene analysis: severe intraventricular hemorrhage in inborn preterm neonates. *The Journal of pediatrics*, 163:1503–6.e1, Nov 2013.

[142] Purcell SM and Chang CC. Plink 1.9, 2015.

[143] Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, and Lee JJ. Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience*, 4:7, 2015.

[144] Das S, Forer L, Schnherr S, Sidore C, Locke AE, Kwong A, Vrieze SI, Chew EY, Levy S, McGue M, Schlessinger D, Stambolian D, Loh PR, Iacono WG, Swaroop A, Scott LJ, Cucca F, Kronenberg F, Boehnke M, Abecasis GR, and Fuchsberger C. Next-generation genotype imputation service and methods. *Nature genetics*, 48:1284–1287, Oct 2016.

[145] Loh PR, Danecek P, Palamara PF, Fuchsberger C, A Reshef Y, K Finucane H, Schoenherr S, Forer L, McCarthy S, Abecasis GR, Durbin R, and L Price A. Reference-based phasing using the haplotype reference consortium panel. *Nature genetics*, 48:1443–1448, Nov 2016.

[146] Fuchsberger C, Abecasis GR, and Hinds DA. minimac2: faster genotype imputation. *Bioinformatics (Oxford, England)*, 31:782–4, Mar 2015.

[147] Jian Yang, Beben Benyamin, Brian P McEvoy, Scott Gordon, Anjali K Henders, Dale R Nyholt, Pamela A Madden, Andrew C Heath, Nicholas G Martin, Grant W Montgomery, Michael E Goddard, and Peter M Visscher. Common snps explain a large proportion of the heritability for human height. *Nature Genetics*, 42(7):565–569, 6 2010.

[148] Noah Zaitlen, Peter Kraft, Nick Patterson, Bogdan Pasaniuc, Gaurav Bhatia, Samuela Pollack, and Alkes L. Price. Using extended genealogy to estimate components of heritability for 23 quantitative and dichotomous traits. *PLoS Genetics*, 9(5):e1003520, 5 2013.

[149] Stephen D. Turner. qqman: an R package for visualizing GWAS results using Q-Q and manhattan plots. *Journal of Open Source Software*, 3(25):731, 5 2018.

[150] Fudenberg G, Imakaev M, Lu C, Goloborodko A, Abdennur N, and Mirny LA. Formation of chromosomal domains by loop extrusion. *Cell reports*, 15:2038–49, May 2016.

[151] Davidson IF and Peters JM. Genome folding through loop extrusion by SMC complexes. *Nature reviews. Molecular cell biology*, 22:445–464, Jul 2021.