MOVING BEYOND GENOME-WIDE ASSOCIATION STUDIES

Jonathan D. Rosen

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
2021

Approved by:

Yun Li

Ming Hu

Michael I. Love

Hyejung Won

Di Wu

ABSTRACT

Jonathan D. Rosen: Moving beyond Genome-wide Association Studies
(Under the direction of Yun Li)

In the last two decades, thousands of genome-wide association studies (GWAS) have been published, describing hundreds of thousands of variant-trait associations across a diverse set of phenotypes. The ubiquity of these studies, however, does not mitigate their significant limitations, including the inability, in many cases, to illustrate the molecular mechanisms underlying these associations. To bridge this gap between association and biological function, a plethora of methodologies have been introduced that move beyond interrogation of the genome at the variant level.

Transcriptome-wide association studies (TWAS) examine the association between imputed gene expression and traits of interest, and in doing so reduce the multiple testing burden that plagues GWAS while offering biological rationales for such associations. Many such methods have been introduced in the last five years, however most do not account for the uncertainty in genotype that arises from imputation. We present a new Bayesian TWAS method, inspired by the BayesR framework, that explicitly models well- and poorly-imputed variants under differing assumptions, allowing for more flexibility in the training step where models to predict gene expression values are built. This method is compared to existing methods using simulated data, demonstrating improved accuracy and power in certain scenarios as well as conservation of Type I error. Predictive performance versus elastic net, which is utilized by PrediXcan, a popular state-of-the-art TWAS method, is measured using real RNA sequencing (RNA-seq) data generated by the Depression Genes and Network (DGN) consortium.

Chromosome conformation capture (3C) techniques have allowed for analysis of the spatial organization of chromatin within the cell nucleus, and the identification of regions that are in close 3-dimensional (3D) proximity provides insight into regulatory pathways that would be hidden from strictly 1-dimensional (1D) analyses such as GWAS or 1D epigenetic footprints similar to those generated by the ENCODE or Roadmap Epigenomics consortia. HiChIP and PLAC-seq (collectively referred to as HP) are emerging 3C technologies for studying genome-wide long-range chromatin interactions mediated by proteins of interest, enabling more sensitive and cost-efficient interrogation of protein-centric chromatin conformation compared to previous Hi-C methods. We present a stratified and weighted correlation metric, derived from normalized contact counts, for quantification of reproducibility in HP data. Our method is applied to multiple real datasets and is shown to outperform existing methods developed for data generated from Hi-C, a widely used genome-wide 3C technology. Furthermore, in a complex PLAC-seq dataset consisting of 11 samples from four types of human brain cells, our method demonstrates expected clustering of data that could not be reproduced using existing methods developed for Hi-C data.

Continuing work in the arena of HP data analysis, we present HPTAD, a method for the identification of topologically associating domains (TADs) using HP data. TADs are contiguous regions of the genome characterized by a higher frequency of within-region interactions relative to between-region interactions; they are implicated in gene regulation and their disruption is associated with a variety of diseases, including cancer. We compare HPTAD to several publicly available tools used to identify TADs from Hi-C input data and demonstrate improved performance relative to "ground truth" TAD regions and boundaries in both mouse and human cell lines. Furthermore, we demonstrate excellent consistency between results obtained from biological replicates and also observe CTCF enrichment at TAD boundaries identified using HPTAD.

ACKNOWLEDGEMENTS

I would like to gratefully acknowledge my advisor, Yun Li, for their guidance and support along with all of the members of the Li group. I would also like to thank Ming Hu for their scientific insights along with the other members of my doctoral committee, Mike Love, Hyejung Won, and Di Wu. This work was supported in part by NIH Grants 3-T32-ES007018 3-T90-DE021986.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

1000G	The 1000 Genomes Project Consortium
1D	one-dimensional
3C	chromosome conformation capture
4C	chromosome conformation capture on a chip
5C	chromosome conformation capture carbon copy
BLUP	best linear unbiased predictor
BSLMM	Bayesian sparse linear mixed model
BVSR	Bayesian variable selection regression
ChIA-PET	chromatin interaction analysis using paired end tag sequencing
ChIP	chromatin immunoprecipitation
CTCF	CCCTC-binding factor
DGN	Depression Genes and Networks
DPR	Dirichlet process regression
EM	expectation maximization
ENCODE	Encyclopedia of DNA Elements
eQTL	expression quantitative trait loci
FISH	fluorescence in-situ hybridization
fQTL	factored polygenic quantitative trait loci
GReX	genetically regulated gene expression
GTEx	Genotype-Tissue Expression Project
GWAS	genome-wide association study
HP	HiChIP and PLAC-seq
IDR	irreproducibile disovery rate
IGVF	Impact of Genomic Variation on Function Corsortium
IS	insulation score
LD	linkage disequilibrium

LDL-C	low-density lipoprotein cholesterol
MCMC	Markov chain Monte Carlo
MESA	Multi-Ethnic Study of Atherosclerosis
mESC	mouse embryonic stem cell
MoC	measure of concordance
MR	Mendelian randomization
ROS/MAP	Religious Orders Study and Rush Memory Aging Project
RNA-seq	RNA sequencing
SNP	single nucleotide polymorphism
TAD	topologically associating domains
TOPMed	Trans-Omics for Precision Medicine
TWAS	transcriptome-wide association study
VC	vanilla coverage
WTCCC	Wellcome Trust Case Control Consortium

CHAPTER 1: LITERATURE REVIEW

1.1 Introduction

For the last 15 years, genome-wide association studies (GWAS) have been a popular tool for exploring the relationship between genotype and phenotype. These studies typically interrogate the genome at the variant level by performing individual tests of association between single nucleotide polymorphisms (SNPs) and phenotype. The first GWAS, which focused on age-related macular degeneration (Klein *et al.*, 2005), was published in 2005 and since then there have been almost 4,800 publications detailing over 200,000 associations, according to the National Human Genome Research Institute and European Bioinformatics Institute (NHGRI-EBI) catalog (Buniello *et al.*, 2019).

Despite the success of GWAS in revealing associations between genotype and a diverse variety of traits including psychiatric (Li *et al.*, 2017), cardiovascular (Zhao *et al.*, 2017), and endocrinologic disorders (Nikpay *et al.*, 2015), among others, these studies are not without notable shortcomings. Many traits are highly polygenic, implying a large number of variants each contribute very small individual effects on the expression of a trait (Hindorff *et al.*, 2011). Consequently, the true underlying genetic architecture of such traits is difficult to understand, and very large cohorts are necessary to achieve adequate power for detection of subtle signals. GWAS also suffers from a multiple testing burden given the large number of SNPs interrogated. Linkage disequilibrium (LD) obscures which variant or variants represent true associations (Edwards *et al.*, 2013). Even if the truly associated variant(s) could be disentangled, a majority of discovered variant-trait associations have ambiguous causal links; approximately 88% of GWAS identified variants reside in non-coding regions of the genome (Buniello *et al.*, 2019).

It is intuitively attractive to assign intergenic variants to the closest gene in 1D space, but it has been demonstrated that intergenic variants can regulate the expression of genes that are quite far apart (Maurano *et al.*, 2012). As one example, a variant associated with low-density lipoprotein cholesterol (LDL-C) via regulation of the *SORT1* gene is not located within or adjacent to *SORT1*, but rather physically resides between two other genes, *CELSR2* and *PSRC1* (Musunuru *et al.*, 2010). However, it has been observed that GWAS associated variants in non-coding regions are enriched in transcriptional regulatory regions defined by chromatin accessibility, transcription factor binding, and histone marks such as H3K27ac, H3Kme1, and H3K4me3 (Schaub *et al.*, 2012; Maurano *et al.*, 2012), suggesting that they affect phenotype via alteration of gene expression. Additionally, there is a significant overlap between these non-coding region variants and expression quantitative trait loci (eQTL), further supporting their role in gene expression (Nicolae *et al.*, 2010).

To bridge the gap between variant-trait association and function, several methodologies have been developed (Gallagher and Chen-Plotkin, 2018; Cano-Gamez and Trynka, 2020), and we will focus on two of these. Transcriptome-wide association studies (TWAS) utilize transcriptomic reference panels to build predictive models of gene expression based on genotype (Zhu and Zhou, 2020). Using these models it is possible to predict gene expression in large cohorts for which transcriptomic data is not available, and this predicted expression can be tested for association with a trait of interest. Since these models aggregate SNP effects reducing the analysis unit to the gene level, the multiple testing burden is reduced by orders of magnitude. Since 2002, chromosome conformation capture (3C) techniques have allowed for analysis of the spatial organization of the cell nucleus (de Wit and de Laat, 2012). By identifying regions that are in close 3D proximity, we can gain insight into regulatory pathways that would not be discernible from a strictly 1D perspective.

1.2 Transcriptome-wide Association Studies

The number of GWAS identified variant-trait associations is continually increasing, however elucidating the biological processes by which these variants affect phenotype remains a challenge. In the last five years, TWAS have emerged as a new set of tools to help further our understanding of these processes. As opposed to GWAS, which interrogates the genome at the variant level, TWAS tests for associations at the gene level. By treating the gene as the functional unit of interest, the multiple testing burden is reduced by orders of magnitude, and any discovered associations imply a mechanistic biological interpretation, even if these interpretations do require further validation.

A number of diverse TWAS methodologies have been published in the last few years, but all share the same basic framework. The standard TWAS analysis can be broken into three distinct steps:

1. Train a model to predict genetically regulated gene expression (GReX)

2. Predict gene expression into a large (GWAS) cohort

3. Test for association between predicted GReX and phenotype

What differentiates most TWAS methods are the basic model assumptions used in the training step. We will focus most of our attention on examining these models, however some methods do differ, to varying degrees, in the other two steps.

Model training necessitates the availability of gene expression and genotype data, and several consortia have made such data publicly available in the last several years. The Genotype-Tissue Expression (GTEx, Carithers *et al.* 2015) project provides whole genome sequencing (WGS) and RNA-seq data for over 50 tissue types in almost 1,000 subjects, although the number of samples varies widely per tissue. The DGN (Battle *et al.*, 2014) consortium provides whole blood RNA-seq data for 922 genotyped European subjects. The GEUVADIS database (Lappalainen *et al.*, 2013) provides messenger and micro RNA sequencing data from lyphoblastoid cell lines in 462 individuals from The 1000 Genomes Project (1000G, The 1000 Genomes Project Consortium Institution/Organization *et al.* 2015).

The aim of the training step is to construct a prediction model that assigns a weight to each SNP that reflects its effect on GReX. Prediction and testing are typically straightforward; the weights determined in the training step are applied to a large genotype panel providing predicted GREx for this cohort. Testing for association can be accomplished using standard models, such as simple ordinary least squares regression if phenotype is continuous or logistic regression if binary.

Training models differ in distributional assumptions on the weights assigned to SNPs, what regions of the genome to consider in the analysis of each gene, and the extent of polygenicity of GReX. Most TWAS methodologies consider only *cis*-SNPs when fitting models, utilizing a fixed window (1 Mb, e.g.) around the start and stop positions of each gene, despite evidence that a significant amount of variance in gene expression can be explained by *trans*-SNPs (Brynedal *et al.*, 2017; Liu *et al.*, 2019). Certain TWAS models explicitly assume high sparsity in which most variants have no effect on GReX, while others assign full polygenicity, assigning some non-zero effect to all variants in the model.

1.2.1 TWAS Training Models

The first TWAS method was published in 2015 (PrediXcan, Gamazon *et al.* 2015), formally introducing the framework outlined in the previous section. PrediXcan utilizes the elastic net (Zou and Hastie, 2005) illustrated in Equation 1.1, which consists of a linear combination of LASSO ($\alpha = 1$, Tibshirani 1996) and ridge regressions ($\alpha = 0$, Hoerl and Kennard 1970). However, the general TWAS framework the authors present is compatible with other machine learning approaches (Gamazon *et al.*, 2015), such as Random Forest and OmnicKriging (Wheeler *et al.*, 2014).

$$\hat{\boldsymbol{\beta}} = \arg\min_{\beta} \|y - \boldsymbol{X\beta}\|^2 + \lambda \left[\alpha\|\beta\|_1 + (1-\alpha)/2\|\beta\|_2^2\right] \tag{1.1}$$

In addition to minimizing the mean squared error, the two additional penalty terms serve dual purposes. First, by placing a higher penalty on weights with large absolute values, single SNPs with large effects are discouraged. Second, sparsity is induced by favoring many zero-valued weights following a similar rationale. This, then, is ideologically consistent with the assumption that GReX consists of moderate effects from a relatively small number of SNPs.

Using R^2 between predicted and true expression as criteria, a pure LASSO ($\alpha = 1$) model performed similarly to elastic net ($\alpha = 0.5$), however the latter was more robust to slight changes in input SNPs (Gamazon *et al.*, 2015). While imputed genotypes outperformed less dense typed data, the use of denser imputation panels like 1000G did not provide significant benefit over less dense panels (HapMap, International HapMap Consortium *et al.* 2007). Interestingly, the inclusion of *trans*-eQTLs improved performance by such a marginal amount they were not included in the ultimate pipeline.

Assessing performance via cross-validation R^2 with DGN whole blood data, PrediXcan outperformed polygenic risk score (Dudbridge, 2013), the single most significant SNP, and the gene-based testing methods VEGAS (Liu *et al.*, 2010) and SKAT (Wu *et al.*, 2011), with cross-validation R^2 approaching the narrow-sense heritability calculated using GCTA (Yang *et al.*, 2011) for many genes. Previously reported GWAS results from the Wellcome Trust Case Control Consortium (WTCCC) study (Burton *et al.*, 2007) were able to be validated, and a novel gene with an established biological precedent was identified that is associated with type 1 diabetes and rheumatoid arthritis.

One criticism of TWAS is that many methods fail to account for error in the prediction step, instead treating predicted expression as a single point mass in the association step. To address this concern Bhutani *et al.* proposed BAY-TS, which bootstraps elastic net models to estimate the mean and variance of each component of β. These estimates are incorporated into the prior distributions of effect sizes used in a Bayesian association step. Using area under the precision recall curve (AUPRC), BAY-TS was compared to ordinary least square regression using β from a single elastic net model, the mean of 50 bootstrapped models, a multiple imputation framework

(Little and Rubin, 1987), and regression calibration (Fuller, 1987), demonstrating improved performance over the other methods. Efforts to repeat the PrediXcan analysis of the WTCCC dataset resulted in replicating less than half of the reported genes, although the authors do report certain inconsistencies attributable to expression normalization techniques.

Shortly following the publication of PrediXcan, Gusev *et al.* published another TWAS methodology, which actually was the first printed usage of the term TWAS. The software associated with this paper is referred to as TWAS/FUSION and we will adopt that nomenclature when discussing this method. TWAS/FUSION incorporates three models in associating genotype with GREx: Bayesian Sparse Linear Mixed Model (BSLMM, Zhou *et al.* 2013), best linear unbiased predictor (BLUP, Robinson 1991), and the most signficant eQTL. It was reported that BSLMM outperforms the other two methods in a majority of examaples and so we will focus our attention on this method.

$$\boldsymbol{y} = 1_n\mu + \boldsymbol{X}\tilde{\boldsymbol{\beta}} + \boldsymbol{u} + \boldsymbol{\epsilon} \tag{1.2}$$

In Equation 1.2 $\boldsymbol{y}$ is a length n vector of expression values, μ is a scalar representing mean expression, $\boldsymbol{X}$ is an $n \times p$ genotype matrix which is typically standardized such that all p columns have mean 0 and unit variance, $\boldsymbol{u}$ is length n vector of random effects with known covariance matrix, and $\boldsymbol{\epsilon}$ is a length n vector of error terms. The term $\tilde{\boldsymbol{\beta}}$ is a length n vector of weights such that

$$\tilde{\beta}_i \sim \pi N(0, \sigma_a^2) + (1 - \pi)\delta_0 \tag{1.3}$$

where δ_0 is a point mass at zero. This method represents a hybrid between linear mixed models and sparse regression models; the point mass in Equation 1.3 induces sparsity, and indeed the number of β values set to zero is controlled through the parameter π, while the random effect vector $\boldsymbol{u}$ captures the combined small effects from all SNPs.

This methodology was evaluated by comparison to standard GWAS as well as restricting association testing to the single best eQTL in each gene. Using simulated causal scenarios with real genotype data from METSIM (Nuotio *et al.*, 2014), YFS (Raitakari *et al.*, 2008), and NTR (Wright *et al.*, 2014), TWAS/FUSION demonstrated the most power in polygenic scenarios and performed similarly to the best eQTL method in scenarios with a single causal SNP. Similarly to PrediXcan, the BSLMM was able to recover most of the theoretical *cis*-heritabilty as determined by GCTA, and while recovery of some *trans*-heritability was observed, many models failed to converge given the subtle effects and small sample sizes. Cross-cohort performance was also demonstrated; training in one dataset and imputing into another yielded correlations between predicted and true expression comparable to those for within-cohort performance after adjusting for heritability in the test set. This has important practical considerations since real-world applications will entail different training and testing cohorts.

The year after the publication of TWAS/FUSION, a factored polygenic QTL (fQTL) method was proposed (Park *et al.*, 2017) that also utilizes a Bayesian framework, similarly using a "spike and slab" prior on β as in Equation 1.3. In contrast to BSLMM, no random effects are assumed, so fQTL specifically models sparse effects. Here, the response considered is an $n \times m$ matrix of expression values $\boldsymbol{Y}$ for n subjects over m tissues. The per variant effects are partitioned into tissue-invariant and tissue-dependent effects, and by pooling information across multiple tissues fQTL not only improves the power of causal SNP identification, but also helps identify specific tissues in the causal pathway of expression in a particular phenotype. Computational efficiency over the Markov chain Monte Carlo (MCMC) algorithm implemented in BSLMM is achieved via use of stochastic variational parameterization (Paisley *et al.*, 2012; Ranganath *et al.*, 2014).

In simulations with multiple causal tissues, fQTL unsurprisingly demonstrated increased power relative to single tissue fQTL, LASSO, and elastic net, but suffered a loss of power in scenarios with a single causal tissue, assumedly due to diluting the signal over several null tissues. This loss of power was mitigated with an increase in the number of causal SNPs, eventually matching that of elastic net. Real-world testing was done employing GTeX for model training, which provides multi-tissue expression data, and GWAS data from both Alzheimer's disease and schizophrenia studies. fQTL identified 107 significant genes, only 10 of which had been previously reported in the NHGRI GWAS catalog. However, applying fQTL in a series of single tissue analyses identified more than twice as many significant genes, most of which were significant only in a single tissue, highlighting the utility of using this method in conjunction with, not in replacement of, single tissue methods.

Other multiple-tissue TWAS methods followed fQTL. MultiXcan (Barbeira *et al.*, 2019) utilizes an ideologically distinct framework to integrate information from more than one tissue. The central premise of this method is the utilization of single tissue expression models to test phenotype association using a joint model:

$$\boldsymbol{y} = \boldsymbol{\mu} + \boldsymbol{t_1} g_1 + \boldsymbol{t_2} g_2 + \cdots + \boldsymbol{t_p} g_p + \boldsymbol{\epsilon} \tag{1.4}$$

where $\boldsymbol{y}$ is a vector of phenotypes and $\boldsymbol{t}_i$ denotes a vector of imputed gene expression for tissue p. In practice, MultiXcan obtains the weights for the expression prediction models from those supplied by PrediXcan and the joint significance of the regression is assessed via F-test. In contrast to fQTL, this method is not specifically testing tissue-dependent SNP effects but rather aggregating effects over multiple tissues. Recognizing that predicted expression for certain tissues can be highly correlated, pricipal component regularization is employed to deal with potential collinearity issues.

In simulation studies, MultiXcan outperformed PrediXcan in all scenarios except in cases of a single causal tissue, similar to what was demonstrated with fQTL. Both MultiXcan and PrediXcan were applied to 222 traits from UK Biobank (Bycroft *et al.*, 2018), which contains deep genotyping for nearly half a million participants, and the former was able to detect more significant gene-trait associations, including many that had not been previously reported. Admittedly, more associations does not imply more true positives, but in conjunction with simulations results showing type I error control these do warrant further investigation.

Yet another multi-tissue TWAS method is the Unified Test for MOlecular SignaTures (UTMOST, Hu *et al.* 2019), which, like fQTL, fits a multivariate model on expression for multiple tissues simultaneously. As opposed to the BSLMM however, the estimates of $\boldsymbol{\beta}$ are determined by solving

$$\hat{\boldsymbol{\beta}} = \arg\min_{\beta} \sum_{i=1}^{p} \frac{1}{2N_i} \|y_i - \boldsymbol{X}_i \boldsymbol{\beta}_{\cdot i}\|_2^2 + \lambda_1 \sum_{i=1}^{p} \frac{1}{N_i} \|\beta_{\cdot i}\|_1 + \lambda_2 \sum_{j=1}^{m} \|\beta_{j \cdot}\|_2 \tag{1.5}$$

Using the $\hat{\boldsymbol{\beta}}$ values from this model it is possible to test for tissue-specific gene-trait associations, and these Z-values can be combined using a modified generalized Berk-Jones test (Sun and Lin, 2017) to provide an omnibus statistic for gene-trait association.

Similar to the aforementioned multi-tissue methods, UTMOST displayed increased power relative to single tissue methods in simulation scenarios with more than one causal tissue. Within- and cross-cohort imputation accuracy, with models trained using GTEx data, exceeded that of PrediXcan and BSLMM in all 44 tissues analyzed, with the largest gains observed in tissues with the smallest sample size; additionally, there was no observed inflation of type I error. Real-world data analysis using UTMOST revealed more significant associations compared to the other two methods, with the same caveats as previously discussed.

In an extension to PrediXcan, Zhou *et al.* proposed a method for joint-tissue transcriptome imputation (JTI, Zhou *et al.* 2020). The motivation driving JTI is borrowing information from other tissues while retaining the variable selection properties of the elastic net. This is achieved by retaining the L_1 and L_2 penalties in Equation 1.1 and adding a weight to the mean squared error term, where now there are observations for each sample-tissue pair. Weights are derived according to the similarity of gene expression profiles and epigenomic factors (such as DNase-I hypersensitivity sites) between tissues, with larger weights corresponding to greater similarity. In this fashion, cross-tissue information is borrowed in a highly strategic fashion, and by setting all cross-tissue weights to unity, PrediXcan becomes a special case of JTI.

Expected performance gains over single-tissue methods were observed using JTI, with the biggest gains coming from tissues with small sample size and high expression correlation with other tissues, and in a direct comparison to another muti-tissue method, JTI outperformed UT-MOST with regard to prediction accuracy. We note that JTI includes a Mendelian randomization (Smith and Ebrahim, 2004; Pierce and Burgess, 2013; Burgess and Thompson, 2013) component for causal inference on gene-trait relationships that is beyond the scope of this discussion.

The framework for a non-parametric Bayesian method, latent Dirichlet process regression (DPR), was proposed by Zeng and Zhou and later formally incorporated into a TWAS package named TIGAR (Nagpal *et al.*, 2019). The motivation behind DPR is the extension of methodologies such as BVSR to include a linear combination of infinite normal distributions:

$$\begin{aligned} \beta_i &\sim \sum_{k=0}^{\infty} \pi_k N(0, \sigma_k^2), \quad \sigma_k^2 \sim IG(a_k, b_k), \\ \pi_k &= \nu_k \prod_{l=1}^{k-1} (1 - \nu_l), \quad \nu_k \sim Beta(1, \xi) \end{aligned} \tag{1.6}$$

where IG is the inverse gamma distribution and parameter ξ controls the number of non-zero β values.

In simulations with genotype data from the Religious Orders Study and Rush Memory Aging Project (ROS/MAP, A. Bennett *et al.* 2012a,b), TIGAR was compared to PrediXcan with respect to both imputation accuracy and TWAS power. The DPR framework outperformed elastic net in most scenarios except low proportion of causal SNPs (< 0.01) and high expression heritability (> 0.2) in both cross-validation and cross-cohort experiments.

Similar results were observed comparing PrediXcan to a collaborate mixed model (CoMM) for dissecting genetic contributions to genetic traits (Yang *et al.*, 2019). Previously discussed methods have been two-stage strategies, namely imputation of gene expression followed by a distinct association step. CoMM performs both steps simultaneously via an expectation maximization algorithm (EM, Dempster *et al.* 1977), which is accelerated using parameter expansion (Liu *et al.*, 1998), and ultimately tests for association via likelihood ratio test. This framework, like BAY-TS, addresses one criticism of many TWAS methods, namely that little or no consideration is given to the error in GReX prediction.

Since CoMM does not impute gene expression in a distinct step, performance comparisons are restricted to TWAS power, and CoMM outperforms both the elastic net and ridge models of PrediXcan in simulated (NFBC1966, Sabatti *et al.* 2009 and real-world (GERA Hoffmann *et al.* 2011) GWAS datasets. Similar to DPR, the biggest power gains are observed in low heritability, highly polygenic scenarios.

1.2.2 Extensions of TWAS Methods

Recently, the CoMM methodology has been extended to multi-tissue analysis (TisCoMM, Shi *et al.* 2020, again accounting for prediction uncertainty by jointly solving both stages via EM algorithm. Of note is the ability to test the joint null (no gene-trait association across all tissues) and the individual nulls individually. In joint testing, TisCoMM outperformed both UTMOST and MetaXcan, notably at low heritability and high sparsity, with power for all tests roughly equivalent at higher heritability. In single tissue testing, TisCoMM was slightly less powerful than CoMM, PrediXcan, and TWAS/FUSION, but with a lower false positivity rate than the other single tissue methods.

While often not explicitly described as such, the aforementioned TWAS methods have been framed as a form of two-sample Mendelian randomization (MR) analysis (Zhu and Zhou, 2020) which aims to determine the causal relationship between an exposure variable, gene expression in the case of TWAS, and an outcome variable (phenotype). The two-sample component arises from the fact that expression imputation is done on a different sample than association testing. Here the *cis*-SNPs serve as instrument variables for the exposure variable, yet one of the underlying assumptions of MR is that instruments affect the outcome solely through the exposure. The well documented horizontal pleiotropy (Verbanck *et al.*, 2018) of variants invalidates this assumption, necessitating sophisticated MR methods to apply in the TWAS setting. Recently, a spate of such methods have been published (PMR-Egger Yuan *et al.* 2020, TWMR Porcu *et al.* 2019, MR-

Robin Gleason *et al.* 2020, PTWAS Zhang *et al.* 2020, SMR Zhu *et al.* 2016, and eQTLWAS Taylor *et al.* 2019), but we forgo discussion of these because of their primary focus on issues outside of gene expression prediction.

Thus far we have restricted our discussion to methods that utilize *cis*-eQTLs for model construction despite considerable evidence that much, if not most, of the variability in GReX can be explained by distal genetic traits (Brynedal *et al.*, 2017; Liu *et al.*, 2019). The justification for focusing on local variants usually proceeds along two arguments: incorporation of *trans*-eQTLs can be computationally intractable in some cases and *trans*-effects can be too noisy to model adequately. However, methods are beginning to emerge in the literature that successfully incorporate both *cis*- and *trans*-eQTLs into the TWAS framework.

MOSTWAS (Multi-Omic Strategies for TWAS, Bhattacharya *et al.* 2020) extends the typical TWAS paradigm by incorporating distal information via one of two methods: 1) identification of distal regulatory elements, training prediction models that incorporate these using their local SNPs, and including these models in the TWAS prediction model and 2) inclusion of distal eQTLs that demonstrate large indirect mediation effect on the gene of interest through mediating local regulatory elements. Predictive models are fit using previously described elastic net or linear mixed model methods.

In simulation studies, MOSTWAS demonstrated improved prediction accuracy and power with respect to detecting gene-trait associations over local-only methods, with larger gains observed with increased distal expression heritability. Under the null case of no distal expression heritability in the testing panel, MOSTWAS performed similarly to local-only methods, with a modest loss of power in low heritability, high sparsity scenarios. Using ROS/MAP data, MOSTWAS demonstrated improved mean predictive and cross-validation R^2 compared to local-only models.

A Bayesian genome-wide TWAS (BGW-TWAS, Luningham *et al.* 2020) was recently proposed that, unlike MOSTWAS, does not discriminate which *trans*-SNPs are included in the model. This method relies on a BVSR framework, fitting separate models for *cis*- and *trans*-

effects, but standard computational solutions for such models are impractical due to both the memory demands of using the entire genome and slow MCMC convergence. To address both of these concerns, BGW-TWAS implements a scalable EM-MCMC algorithm (Yang *et al.*, 2017a) using only summary GWAS statistics based on single variant tests as opposed to individual genotype information. In simulation studies BGW-TWAS demonstrated improved predictive performance and TWAS power compared to *cis*-only BVSR, PrediXcan, and TIGAR over a range of heritability and sparsity scenarios with two exceptions: in highly polygenic simulations with low expression heritability all methods performed equally poorly, and when the majority of signal was due to *cis*-SNPs under high expression heritability, PrediXcan and TIGAR outperformed BGW-TWAS. Using a ROS/MAP trained model, BGW-TWAS identified an Alzheimer's gene *ZC3H12B* whose signal was driven entirely by *trans*-eQTLs.

Since individual level genotype data is not typically available for large GWAS, many of the aforementioned methods can directly accommodate GWAS summary statistics, or have extensions allowing for their use (S-PrediXcan Barbeira *et al.* 2019, CoMM-S^2 Yang *et al.* 2020). TWAS/FUSION was the first method to demonstrate the viability of using expression weights, GWAS Z-scores, and a suitable LD panel for *cis*-SNPs to construct a gene-trait association Z-score. This process was validated by comparison with an identical analysis using individual genotype data, yielding nearly identical results, with an overall slight underestimation observed using summary data.

1.3 Chromosome Conformation Capture

GWAS has undeniably been successful in identifying variant-trait relationships, but genomic function is known to depend on more than the 1D structure of DNA. Human DNA contains over 3 billion nucleotides, which when unwound has a linear length greater than 2 meters, yet it resides within the nucleus, a cellular structure many orders of magnitude smaller. The organizational architecture required to achieve this necessitates complex folding (Bickmore, 2013) which, in conjunction with its 1D structure, confers functionality to DNA (Ong and Corces, 2014). Early investigations into the 3D structure of DNA relied on simple light microscopy, and subsequent technological advances enabled more sophisticated microscopy methods such as 3D fluorescence in-situ hybridization, which allowed for visualization of individual contacts within the nucleus (3D-FISH, Cremer *et al.* 2012). Microscopy based methods, however, are limited to studying a particular region of the genome rather than the genome as a whole (Bonev and Cavalli, 2016).

A breakthrough study at the beginning of the 21st century (Dekker *et al.*, 2002) demonstrated a technique to explore the 3D structure of DNA organization that did not rely on any form of microscopy. This so-called chromosome conformation capture (3C) technique is based on a simple premise: by immobilizing DNA at points where different sections are in very close spatial proximity to each other, it should be possible to quantify the frequency of interactions at different genetic loci. In practice, this is performed by using formaldehyde to freeze interactions between chromatin bound regions (cross-linking) followed by digestion of DNA using enzymes that cut DNA strands at specific base pair (bp) sequences, typically 4 or 6 bp in length. Subsequent to digestion, the loose ends are ligated in dilute solution to promote ligation between segments that are cross-linked over ligation between non-linked segments. Finally, the ligated segments are reverse cross-linked and the signal amplified using polymerase chain reaction (PCR) methods. These paired ends can then be sequenced to reveal genetic loci that interact in 3D space.

Since the initial Dekker *et al.* procedure, several additional 3C methods have been developed, which differ mainly in the scope of the interactions they detect. While the original procedure quantifies interactions between two specific genetic loci, chromosome conformation capture on a chip (4C, Simonis *et al.* 2006) quantifies interactions between one specific locus and all other loci in the genome. Chromosome conformation capture carbon copy (5C, Dostie *et al.* 2006) quantifies all pairwise interactions between loci confined to a specific region of the genome.

Perhaps the most ubiquitous 3C methodology in the last decade has been Hi-C (Lieberman-Aiden *et al.*, 2009), which incorporates high-throughput sequencing into the 3C workflow, allowing for the genome-wide quantification of all pairwise interactions between genetic loci. Since its introduction, Hi-C has been widely used to explore the 3D structure of both the human and non-human genomes (Han *et al.*, 2018). Data visualization for these studies often consists of a contact map, a symmetric $n \times n$ matrix whose i, j element represents the number of interactions between loci i and j. The length of these loci in bp is referred to as the resolution of the experiment and is a function of the restriction enzyme used and, indirectly, to the read depth. Too fine of a resolution relative to number of reads would result in an overly sparse contact map.

In order to provide scientific rigor in Hi-C experimentation it is critical to have computational tools available to quantify data quality and measure reproducibility between samples. Not only does Hi-C suffer from technical noise that is inherent in all biological experiments, the use of many different tools to perform alignment and many choices among restriction enzymes, resolutions, and normalization techniques add additional variability. Consequently, it is particularly important to have methods that ensure data quality from Hi-C experiments.

1.3.1 Measuring Hi-C Reproducibility

A natural, intuitive, and computationally simple method for quantifying the similarity between two Hi-C contact matrices is Pearson correlation between vectors of contact frequencies, and indeed this was used in practice in the early days of Hi-C. However, one of the characteristic features of Hi-C data is that contact frequency, on average, decreases with increasing genomic distance (Lieberman-Aiden *et al.*, 2009), and Pearson correlation does not account for this. Yang *et al.* have demonstrated the failure of Pearson correlation in distinguishing between biological replicates and non-replicates, specifically showing that the latter can be more correlated than the former, contrary to biological plausibility.

Additional biases further invalidate a naïve correlation analysis, which implicitly assumes that all measurements in a contact map are independent. Contact frequency has been demonstrated to be dependent on factors such as restriction fragment length, GC content, and mappability (Juric *et al.*, 2019). Chromosomes exhibit compartmentalization on several scales, including alternating regions open and closed chromatin (A/B compartments, Lieberman-Aiden *et al.* 2009) and topologically associated domains (TADs, Dixon *et al.* 2012; A. Bennett *et al.* 2012a; Nora *et al.* 2012), regions that exhibit frequent interactions within the domain but much lower levels of interaction outside the domain. Additionally, the choice of resolution is usually done in an ad hoc fashion, yet is correlated with the noise due to signal dropout.

There exists the need then for computational methods that quantify how "close", in some sense, two contact matrices are to each other while accounting for some, if not all, of these biases. At minimum, a reasonable measure of reproducibility should determine that replicates are closer than non-replicates, and ideally this closeness would have some meaningful biological interpretability. We will discuss several such methods that have been developed in the last few years.

GenomeDISCO (Ursu *et al.*, 2018) reports a concordance measure of DIfferences between Smoothed COntact, which comprises three steps: conversion of a Hi-C contact matrix to a transition matrix, smoothing via a random walk process, and comparison of two such transformed matrices to yield a reproducibility metric.

Contact matrices are converted to transition matrices by first normalizing the number of contracts by a procedure known as vanilla coverage (VC, Lieberman-Aiden *et al.* 2009) square root followed by scaling such that the rows all sum to one. In this manner the i, j element is interpretable as the probability of transitioning from locus i to j. A random walk of t steps is then simply calculated by raising the matrix to the t power, the result of which is smoothing the original contact matrix such that between element variability is reduced.

For two such matrices, $A1^t$ and $A2^t$, the distance between them is then measured as the sum of the absolute difference between each entry divided by the average number of non-zero entries prior to smoothing:

$$d_t(A1, A2) = \frac{\sum_i \sum_j \left|(A1)_{ij}^t - (A2)_{ij}^t\right|}{\frac{1}{2}\left(\left|\{A1_i | \sum_j A1_{ij} > 0\}\right| + \left|\{A2_i | \sum_j A2_{ij} > 0\}\right|\right)} \tag{1.7}$$

This metric is converted to the domain of $[-1, 1]$, with larger values representing "closer" matrices. The smoothing is tuned heuristically with high quality Hi-C datasets, using half of the dataset for training and the remaining half for testing in order to find a value for t such that sufficient denoising is performed but not to the extent that domain specific information is lost due to over-smoothing. In practice this is accomplished by maximizing the area under the precision-recall curve (auPRC).

Rather than focusing on smoothing to reduce the noise in Hi-C data, HiC-spector (Yan *et al.*, 2017a) utilizes eigendecomposition to reduce the dimensionality of contact matrices, focusing on eigenvectors containing the most information. As in GenomeDISCO, contact frequency is used as a proxy for spatial distance such that larger values of the i, j element of a contact matrix imply closer proximity of loci i and j in 3D space (Wang *et al.*, 2016). Normalized Laplacian matrix ℓ is defined as $I - D^{-1/2}AD^{-1/2}$ where A is the contact matrix, I is the $n \times n$ identity matrix and D is a diagonal matrix such that $D_{ii} = \sum_j A_{ij}$. In this formulation D_{ii} represents the coverage of locus i, the total contact frequency across the entire chromosome.

Laplacian ℓ is guaranteed to be positive semi-definite with at least one eigenvalue equal to 0, so the eigenvalues λ satisfy $\{0 \leqslant \lambda_0 \leqslant \lambda_1 \leqslant \cdots \leqslant \lambda_{n-1}\}$. For two contact matrices $A1$ and $A2$, if we define the aforementioned eigenvalues as $\{\lambda_0^{A1}, \lambda_1^{A1}, \ldots, \lambda_{n-1}^{A1}\}$ and $\{\lambda_0^{A2}, \lambda_1^{A2}, \ldots, \lambda_{n-1}^{A2}\}$ with corresponding eigenvectors $\{\nu_0^{A1}, \nu_1^{A1}, \ldots, \nu_{n-1}^{A1}\}$ and $\{\nu_0^{A2}, \nu_1^{A2}, \ldots, \nu_{n-1}^{A2}\}$, then the distance between them is defined as the sum of Euclidean norms of the differences between r eigenvectors:

$$S_d(A1, A2) = \sum_{i=0}^{r-1} ||\nu_i^{A1} - \nu_i^{A2}|| \tag{1.8}$$

Using a pair of random vectors as a reference, this metric can be normalized to lie in the range $[0, 1]$.

The Euclidean distance between two high order eigenvalues was found to be nearly identical to that of two randomly selected unit vectors from a multivariate normal distribution, suggesting they are merely capturing noise. Therefore, the inclusion of excess terms would serve only to downward bias the similarity metric between two matrices, increasing their similarity. The authors report the use of 20 as an empirically derived acceptable value for r.

QuASAR-Rep (Sauria and Taylor, 2017) reduces the problem of reproducibility to calculating the Pearson correlation between two vectorized matrices; however, rather than comparing raw or normalized contact frequencies, scaled correlations are the unit of comparison. Since the majority of contacts occur within small linear distances and long-range interactions are very noisy, a local correlation matrix is computed such that only a narrow, pre-defined distance between loci is considered.

For contact matrix A, correlation matrix C is defined such that $C_{ij} = \text{corr}(A_{ij,local}, A_{ji,local})$ where $A_{ij,local} = \{A_{ik} | j - 100 \leqslant k \leqslant i + 100, k \neq i, k \neq j, I(k) = 1\}$ and $I(k)$ is the indicator for valid rows or columns. A valid row is defined as one that contains more than three non-zero terms with a standard deviation not equal to zero. Transformation matrix T is defined as the element-wise product of C and the scaled raw counts of A. The final reproducibility metric is the correlation between two transformation matrices, $\text{corr}(\{T_{A_1ij} | I_{A_1}(i,j) = 1, I_{A_2}(i,j) = 1\}, \{T_{A_2ij} | I_{A_1}(i,j) = 1, I_{A_2}(i,j) = 1\})$ where the indicator functions take on the value of 1 for valid values of the corresponding transformation matrices.

Unlike the previously discussed methods, QuASAR-Rep does not attempt denoising by smoothing or dimensionality reduction, but rather by summarizing a function of contact frequency within genomic neighborhoods. Local correlation structures are scaled proportionally to contact frequency, and the correlation between these is put forward as an indicator of matrix similarity.

HiCRep (Yang *et al.*, 2017b) is perhaps the most mathematically straightforward of the Hi-C reproducibility methods introduced in the last few years, however complexity should not be conflated with performance. HiCRep has been cited far more often than the aforementioned methods and, not surprisingly, shares some elements with them as well. Like GenomeDISCO, contact frequency is smoothed to reduce noise. This is accomplished by simply taking the arithmetic mean of counts within a defined 2-dimensional square region.

Considering that the predominant feature of Hi-C data is the decay of interaction freqency as linear genomic distance increases (Lajoie *et al.*, 2015), rather than treating all data in a smoothed contact matrix as equally informative, the HiCRep procedure explicitly takes this relationship into account. This is accomplished by extracting equidistant loci (stratum) from smoothed contract matrices and comparing these via Pearson correlation. Strata specific values are then combined via weighted sum, with weights derived as a function of sample variances. If A_{1k} and A_{2k} are the k^{th} strata from smoothed contract matrices A_1 and A_2 respectively, then ρ_k is corr(A_{1k}, A_{2k}). The weight for statum k is given by

$$w_k = \frac{N_k\sqrt{\text{var}(A_{1k})\text{var}(A_{2k})}}{\sum_{k=1}^{K} N_k\sqrt{\text{var}(A_{1k})\text{var}(A_{2k})}} \tag{1.9}$$

where K is the total number of strata. By definition the weights fall in the range $[0, 1]$ and sum to one, and from Equation 1.9 we see that more weight is placed on strata containing more observations and which are more highly variable. The final reproducibility metric is simply $\sum_{k=1}^{K} w_k\rho_k$.

1.3.2 Comparison of Hi-C Reproducibility Methods

In their original publications, some of the aforementioned methods were directly compared to each other; recently, however, a comprehensive comparison of the methods described in the previous section, including Pearson correlation, was published to assess their relative performance (Yardimci *et al.*, 2019). Testing data consisted of two replicate Hi-C experiments on cells from 13 immortalized human cancer cell lines, with read depth ranging from 10 million to 400 million paired reads per experiment. Simulated noise was generated arising from two sources: genomic distance noise (Lieberman-Aiden *et al.*, 2009) and random ligation noise (Lajoie *et al.*, 2015). Two ratios of these noise types were injected into real data to examine the effect on performance.

The reproducibility metrics monotonically decreased with increasing noise for all methods compared. Qualitatively, there is evidence that HiCRep and QuASAR-Rep might be more robust to noise evidenced by the concavity of their plotted curves of metric vs. noise, in contrast to convex patterns observed with the other methods. That is, steep declines in reproducibility metrics (relative to other methods) were only observed at the highest noise levels for HiCRep and QuASAR-Rep. Despite consistency in the monotonicity of decay trends, the methods did differ in which noise ratio resulted in higher reproducibility metrics, suggesting that some methods are more sensitive to specific types of noise.

Performance using real-world data focused primarily on differentiating biological replicates, non-replicates, and pseudo-replicates (further details in Chapter 2), with the expectation that a reproducibility metric should rank pseudo-, biological, and non-replicates respectively in order of decreasing metric. While all five methods were able to correctly rank the replicate types on average, intuitively a measure of method performance would be the greatest separation between replicate types. In many test scenarios HiCRep and HiC-Spector outperformed the other measures in this regard.

Performance at different coverages was assessed by downsampling, and all methods demonstrated a consistent pattern of smaller metrics and worse separation between replicate types at lower coverages. However, performance did plateau for all methods at approximately 25 million read pairs. HiCRep and HiC-Spector outperformed the other methods at most depths, while QuASAR-Rep was unable to differentiate cell pair types below 20 million read pairs.

Using two deeply sequenced cell lines (400 million read pairs) all methods were compared at different binning resolutions (10, 40, and 500 Kb) and found to be robust to changes in resolution. Interestingly, QuASAR-Rep demonstrated greater separation between the three types of replicate pairs at lower resolution, potentially implying that the method is capturing more large domain structure than the other methods. However, when binning resolutions of 5, 10, 20, and 40 Kb were applied to pairs of biological replicates, the relative performances of the methods showed variable trends. Consequently, using reproducibility measures to determine appropriate experiment resolutions is not recommended.

The main conclusion from the various comparisons is not that one method is wholly superior to the others, but rather that selection of reproducibility measure should be guided by the nature of the analysis.

1.3.3 Chromatin Immunoprecipitation Methods

While Hi-C has allowed for tremendous gains in the understanding of the 3D architecture of the human genome at high resolution, one drawback is the requirement for very deep sequencing to achieve such resolution since Hi-C detects proximity ligations in a genome-wide fashion. Such deep sequencing not only increases cost, but often makes investigation of specific regions at high resolution impractical. Gene transcription regulation has been shown to be controlled by distal interaction between enhancers and promoters (West and Fraser, 2005) but Hi-C is not ideally suited for a strict focus restricted to such interactions. To enhance specificity of Hi-C, a novel strategy for chromatin interaction analysis using paired end tag sequencing (ChIA-PET, Fullwood *et al.* 2009) has been proposed that combines Hi-C with chromatin immunoprecipitation (ChIP) to enrich long-range contacts associated with a protein or histone modification of interest.

ChIA-PET indeed allows for improved resolution of interactions implicated in transcriptional regulation, however these experiments still require hundreds of millions of cells with a small fraction of informative reads (Tang *et al.*, 2015). Two related methods, HiChIP (Mumbach *et al.*, 2016) and PLAC-seq (Fang *et al.*, 2016), have recently been introduced that aim to alleviate this issue by utilizing the principles of *in situ* Hi-C (Rao *et al.*, 2014); specifically, proximity ligation is performed within intact nuclei prior to lysis reducing the number of false positive interactions. Both methods boast improved signal-to-noise ratios over *in situ* Hi-C. HiChiP provides 10 times the number of informative reads with a 100-fold reduction in input material relative to ChIA-PET (Mumbach *et al.*, 2016), and PLAC-seq similarly provides 10 times the number of reads with 20-fold fewer cells relative to a similar ChIA-PET study (Fang *et al.*, 2016).

Several tools originally designed for Hi-C have been applied to HiChIP and PLAC-seq data (Phanstiel *et al.*, 2015; Lareau and Aryee, 2018; Juric *et al.*, 2019) for the detection of long-range peaks, but no methods for assessing reproducibility have been designed specifically for this type of data. In addition to biases inherent in all Hi-C data, such as effective fragment length and GC content (Yaffe and Tanay, 2011), the chromatin immunoprecipitation step introduces additional biases. Furthermore, since the HiChIP/PLAC-seq methods allow for detection of interactions between two protein-bound regions as well as one bound and one unbound region, some biases may not be consistent across all interactions.

GenomeDISCO was successfully applied to one HiChIP dataset (Ursu *et al.*, 2018), however that methodology was designed for Hi-C data analysis and does not account for the aforementioned ChIP bias, nor does it restrict attention solely to interactions that contain at least one protein-bound region. Recently, an extension to the irreproducible discovery rate (IDR, Li *et al.* 2011) method has been proposed that allows for the principles of the original method to be applied to 2-dimensional data (IDR2D, Krismer *et al.* 2020). By use of a two-component gaussian copula mixture model, the IDR method assigns a probability of being irreproducible to each peak in a ChIP-seq experiment. IDR2D applies this to significant ChIA-PET and HiChIP interactions as called by Mango (Phanstiel *et al.*, 2015) and hichipper (Lareau and Aryee, 2018) respectively, however neither method considers interactions between protein-bound and non-bound regions. Moreover, IDR2D considers interactions individually, and while these results could theoretically be combined to produce a single genome-wide reproducibility metric, no formal method for this exists. It is specifically recommended that IDR2D should be used in conjunction with, not in lieu of, existing Hi-C specific methods.

1.4 Topologically Associating Domains

The development of chromatin conformation capture technologies has led to the discovery of larger scale organizational features of DNA beyond chromatin loops. When the Hi-C method was originally described (Lieberman-Aiden *et al.*, 2009) the authors discussed an apparent segmentation of the genome into arbitrarily labeled A and B compartments. These compartments represent areas of open and closed chromatin (A and B respectively), and have been shown to be cell-type specific. Furthermore, compartmentalization of the genome can change during the course of an organism's development (Dixon *et al.*, 2015). In 2012, another type of compartmentalization was described (Dixon *et al.*, 2012); topologically associating domains (TADs) are defined as chromatin regions where within-region interactions are more frequent than between-region interactions. TADs are believed to regulate gene expression by limiting the interaction between distal regulatory elements such as enhancers and promoters (Dixon *et al.*, 2016), while genes within the same TAD have demonstrated correlated expression (Nora *et al.*, 2012).

Since their initial description, there has been an increase in the examination of the functional relevance of TADs. These regions have been shown to be highly conserved between cell types (Schmitt *et al.*, 2016; McArthur and Capra, 2021) and species (Dixon *et al.*, 2012). Additionally, they have been implicated in number of diseases. Disruption of TAD boundaries has been linked to adult-onset demyelinating leukodystrophy (Giorgio *et al.*, 2015), developmental limb defects (Lupiáñez *et al.*, 2015; Ren and Dixon, 2015) and cancer (Valton and Dekker, 2016; Li *et al.*, 2019; Akdemir *et al.*, 2020; Pinoli *et al.*, 2020). It should not be surprising then that a number of methods to detect TADs and TAD boundaries have been reported in the last decade. In the following section we will examine some of these methods and compare their underlying assumptions.

1.4.1 TAD Calling Methods

The first formalized TAD calling method was based on the directionality index described by Dixon *et al.*. As descrbed in the previous section, a Hi-C experiment yields the number of read pairs between two genetic loci which are binned at a specific resolution. Consequently, using these (possibly normalized) counts as input, it is straightforward to compute the total number of read counts between any locus i all loci within a specified window upstream or downstream of i. If we let A_i and B_i represent these upstream and downstream counts respectively, the directionality index is defined as:

$$DI = \frac{B_i - A_i}{|B_i - A_i|}\left(\frac{(A_i - E_i)^2}{E_i} + \frac{(B_i - E_i)^2}{E_i}\right) \tag{1.10}$$

where E_i represents the expected counts at locus i, defined as the arithmetic mean of A_i and B_i.

The fundamental assumption is that the difference between upsteam and downstream contacts should be maximized at TAD boundaries, specifically upstream counts are expected to exceed downstream counts at the beginning of a TAD region while the opposite is expected at the end of a TAD. Actual boundaries are called using the vector of indices as input into a hidden markov model consisting of a Gaussian mixture model with one to twenty components. The optimal number of mixture components is chosen using AIC to assess the best fit model and domains are called based on changes in hidden state. TADs that do not meet specific criteria (such as minimum length) are designed as unorganized chromatin and represent inter-TAD regions.

Another widely adopted early TAD calling method is referred to as insulation score (Crane *et al.*, 2015). Scanning the diagonal of a contact matrix, for each locus an insulation score it computed representing the sum of interactions spanning that locus within a specified neighborhood. Intuitively, a TAD boundary should be reprented by a local minima insulation score, and the method employs an algorithmic process by which such minima are selected and evaluated. It should be noted that this process involves parameters whose value can have a significant impact on the number of TADs identified.

TopDom (Shin *et al.*, 2016) is another TAD calling method that bears some ideological similarities to both the directionality index and insulation score. Similar to the directionality index, upstream and downstream contacts from a specific locus are counted, but these are averaged ("bin score") instead of evaluated separately. Similar to the insulation score, potential TAD boundaries are identified as inflection points in the series of "bin scores", however these are determined by a piecewise linear function. TAD boundaries are selected from these potential ones by testing the upstream and downstream contact frequency difference using a Wilcoxan rank sum test.

The aforementioned methods all identify discrete TADs, however high resolution Hi-C studies have pointed to the existence of hierarchically nested TADs (Weinreb and Raphael, 2016). OnTAD (An *et al.*, 2019) was developed specifically to identify such nested TADs (subTADs) and TAD-containing TAD hierarchies (metaTADs). Similar to insulation score and TopDom, the average contact frequencies within a diamond-shaped window are determined by sliding along the diagonal. The process is repeated varying the window size, and local minima are selected for each window size. Hierarchical TADs are called from the union of potential boundaries using a dynamic programming algorithm.

More recent TAD calling methods include Grinch (Lee and Roy, 2021), which is reported to be particularly well suited to sparse Hi-C data. Unlike the previously discussed methods, Grinch relies on contact matrix factorization rather than a linear score based method for TAD detection. Non-negative factorization of the contact matrix is followed by a local smoothing procedure to account for the previously discussed distance dependence of Hi-C contact frequency. One of the smoothed factor matrices is treated as a set of latent features, and TADs represent clustered regions found by applying k-medoids clustering.

Other methods use statistical inference to call TADs, making assumptions about the distribution of integer counts or normalized contact frequency. HiCseg (Lévy-Leduc *et al.*, 2014) assumes counts follow a negative binomial (integer) or Gaussian distribution (normalized), and by assuming that interactions within a TAD arise from a common distribution, the 2-dimension segmentation problem is reduced to a 1-dimensional problem. Boundaries are determined using a dynamic programming algorithm that iteratively tests sets of boundaries to determine the set that maximizes the likelihood of the observed data.

There are over two dozen published TAD calling methods to date, and it is outside of the scope of this review to provide detail for all of them. In addition to linear score based methods such as insulation score, directionality index, TopDom, and OnTAD, numerous others have been reported. These include armatus (Filippova *et al.*, 2014), arrowhead (Durand *et al.*, 2016), CaTCH (Zhan *et al.*, 2017), EAST (Ardakany and Lonardi, 2017), GMAP (Yu *et al.*, 2017), HiCDB (Chen *et al.*, 2018), HiCExplorer (Ramírez *et al.*, 2018), HiTAD (Wang *et al.*, 2017), matryoshka (Malik and Patro, 2019), and TADBD (Lyu *et al.*, 2020).

Several clustering based methods have also been published, such as CHDF (Wang *et al.*, 2015), ClusterTAD (Oluwadare and Cheng, 2017), ICFinder (Haddad *et al.*, 2017), and TADpole (Soler-Vila *et al.*, 2020). Other statistical based methods include HiCKey (Xing *et al.*, 2021), PSYCHIC (Ron *et al.*, 2017), TADbit (Serra *et al.*, 2017), and TADtree (Weinreb and Raphael, 2016). Viewing Hi-C contacts as a network where bins are nodes and contact pairs are connected by an edge, several network based TAD detection methods have also been proposed: 3DNetMod (Norton *et al.*, 2018), deDoc (Li *et al.*, 2018a), MrTADFinder (Yan *et al.*, 2017b), and spectral (Chen *et al.*, 2016). Recently, a machine learning approach has been proposed which utilizes epigenetic data (such as CTCF ChIP-seq) to train predictive models (Stilianoudakis and Dozmorov, 2020).

None of the aforementioned methods are designed for use with HiChIP or PLAC-seq as input, which leaves the question of whether biologically meaningful TAD information can be obtained from such data currently unanswered. It is reasonable to asssume that the biases introduced by the chromatin immunoprecipitation step would make Hi-C specific methodologies inappropriate to use, at least without modification, similar to what discussed in the previous section.

CHAPTER 2: BAYES QN: A TWAS TRAINING MODEL

2.1 Introduction

GWAS has been undeniably successful in revealing a tremendous number of SNP-trait associations, however the multiple testing burden of genome-wide analyses, difficulty in detecting small effects even with very large sample sizes, and frequent lack of mechanistic insight provided by such results have all contributed to the search for improved methods to test genotype-phenotype relationships. TWAS has been proposed as one method to address these GWAS shortcomings (Gamazon *et al.*, 2015; Gusev *et al.*, 2016), and a number of methods have emerged in the last five years to implement the TWAS strategy, as discussed in Chapter 1.

Despite the diverse modeling strategies employed, few have explicitly incorporated sources of uncertainty, outside of random noise, into TWAS models. Consequently, most TWAS methods restrict training and testing to high-quality SNPs, typically those meeting stringent criteria, such as high imputation R^2 (> 0.80, e.g.). Variants that fail to meet these thresholds often do so as a result of low allele frequency, and are frequently discarded from analyses despite evidence that less common variants may carry significant information (Dickson *et al.*, 2010; Gibson, 2012). The BAY-TS method (Bhutani *et al.*, 2017) estimates prediction error from the training models by bootstrapping the posterior distributions of SNP effect coefficients, but still only utilizes high-quality genotype data.

In this chapter we introduce BayesQN, a TWAS model framework motivated by BayesR (Erbe *et al.*, 2012) that is designed to incorporate low-quality variants. Rather than applying identical distributional assumptions to each SNP in the model, BayesQN allows for the dichotomization of variants based on imputation quality, each with attributes unique to its category. Proof of concept is demonstrated using simulated data and conservation of Type I error is verified. Predictive performance training with data from DGN is compared to the methodology used by PrediXcan (Gamazon *et al.*, 2015), highlighting the benefit of using poor-quality SNPs and revealing genes for which this strategy is preferred over elastic net. Lastly, we compare cross-cohort predictive performance training with data from DGN and imputing into an unrelated dataset.

2.2 Methods

2.2.1 Training Model

The BayesQN model is built upon an assumption common to many TWAS methods; SNPs individually contribute to gene expression in a linear fashion:

$$\boldsymbol{Y} = \mu + \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{2.1}$$

where $\boldsymbol{X}$ is an $n \times p$ matrix whose elements are dosages for n subjects and p variants whose columns have been centered and scaled to have unit variance. Dosage values capture the post-imputation genotype probability with respect to a reference allele; if p_0, p_1 and p_2 represent the probabilities of homozygous for the alternate allele, heterozygous, and homozygous for the reference allele, the dosage given by $p_1 + 2p_2$, and consequently is restricted to the domain $[0, 2]$. For each gene, the set of variants that comprise $\boldsymbol{X}$ are chosen from a 1 Mb window around the gene, a *cis*-SNP range common to many TWAS methods. The response $\boldsymbol{Y}$ is a length n vector of gene expression values that have been adjusted for covariate effects and transformed to be approximately normally distributed, $\boldsymbol{\beta}$ is a length p vector of per-variant effects on gene expression, μ is a scalar mean effect and $\boldsymbol{\epsilon}$ is a length n vector of random noise with assumed distribution $\boldsymbol{\epsilon} \sim N(0, \sigma_g^2 I_n)$.

One of the primary motivations for BayesQN was the ability to incorporate the increased imputation uncertainty of low-quality variants by not forcing a uniform set of distributional assumptions across all $\boldsymbol{X}$. In practice, this is accomplished by partitioning the variants *a priori* into two categories and assuming

$$\beta_j|\boldsymbol{\pi_q}, \sigma_q^2, q_j = q \sim \pi_{q1}N(0,0) + \pi_{q2}N(0, 0.001\sigma_q^2) + \pi_{q3}N(0, 0.01\sigma_q^2) + \pi_{q4}N(0, 0.1\sigma_q^2) \quad (2.2)$$

where $q \in \mathbb{N}$ in theory, but will be restricted to 1 and 2 in the examples that follow. The vector $\boldsymbol{\pi_q}$, whose elements sum to unity, contains the proportions each of the components of Equation 2.2 contributes to the distribution of variant j.

2.2.2 Genotype Simulation

Genotypes were simulated such that samples would possess realistic linkage disequilibrium (LD) structures and minor allele frequency profiles. First, 1 Mb sections of the genome were generated using a coalescent simulation model (Schaffner *et al.*, 2005), each section containing ten-thousand chromosomes. These samples were subsequently thinned to HapMap variant density and matching HapMap allele frequency spectrum, and pairwise LD values were computed on a random subset of haplotypes. From this set of SNPs, tags were chosen to mimic the coverage of the Illumna 300K panel. Finally, two-thousand haplotypes were randomly chosen to provide a reference panel for imputation of the remaining eight-thousand haplotypes, reduced to the set of tag SNPs, with Minimac3. This procedure not only simulated realistic genotypes, but also imputation results, including “Rsq” ($\hat{r}^2$) for each variant, which is defined as

$$\hat{r}^2 = \frac{\frac{1}{2n} \times \sum_{i=1}^{2n}(D_i - \hat{p})^2}{\hat{p}(1-\hat{p})} \quad (2.3)$$

where n is the number of samples, D_i is the dosage of sample i, and $\hat{p}$ is the allele frequency. This $\hat{r}^2$ value is the metric used to dichotomize variants in the BayesQN experiments with two categories (herein referred to as BayesQ2).

2.2.3 Phenotype Simulations

Two sources of variation need to be specified in a standard TWAS framework, the percent variance of gene expression explained by genotype (PVE_g) and the percent variance of phenotype explained by gene expression (PVE_e). Gene expression values are simulated by first randomly selecting the desired number of causal SNPs from their respective categories. The value of each β corresponding to a causal SNP is then simulated from a standard Normal distribution for scenarios in which all effect sizes of genotype on expression come from a common distribution. In cases of differential distributions for causal SNP β values, the standard distributions are modified accordingly. For example, if one set of causal SNPs is assumed to have twice the variance of another, the βs for the former are simulated from independent $N(0, \sqrt{2})$ distributions while the latter are simulated from standard Normal. Non-causal SNPs correspond to $\beta = 0$. We then define simulated gene expression $\boldsymbol{Y_{sim}}$ as

$$\boldsymbol{Y_{sim}} = \boldsymbol{X\beta} \times \sqrt{PVE_g/v} + N(0, I \times \sqrt{1 - PVE_g}) \tag{2.4}$$

where $v = Var(\boldsymbol{X\beta})$.

Phenotype values are simulated in a related fashion. We define $\boldsymbol{P_{sim}}$ as

$$\boldsymbol{P_{sim}} = \boldsymbol{Y_{sim}} \times \alpha + N\left(0, I \times \sqrt{u\frac{1 - PVE_e}{PVE_e}}\right) \tag{2.5}$$

where $u = Var(\boldsymbol{Y_{sim}} \times \alpha)$ and α is a scalar indicating the strength of association between gene expression and phenotype.

2.3 Results

2.3.1 Variation in Number of Causal SNPs

In order to establish the viability of the BayesQN methodology, we compared predictive performance between several established methods using simulated data, with the number of

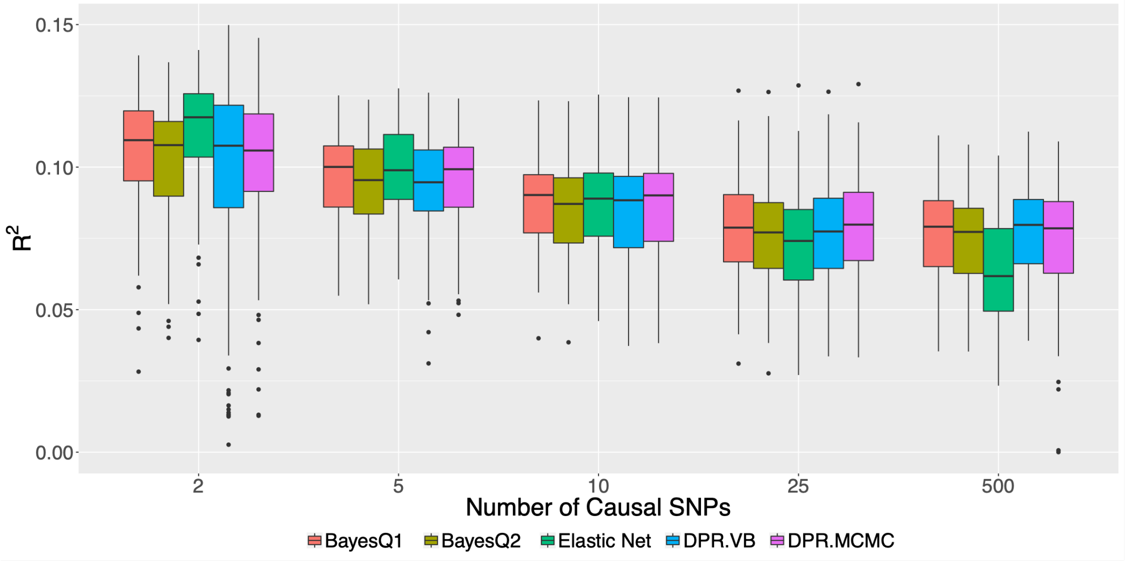

Figure 2.1: Comparison of five TWAS gene expression prediction methods using simulated data. The number of causal SNPs is displayed on the x-axis, and the predictive R^2 is on the y-axis. Center lines of the boxplots indicate median over all simulations, and the whiskers represent the IQR, with outliers displayed as points. The PVE_g is set to 12.5%.

causal SNPs ranging from 2 to 500. Predictive R^2 was determined by splitting each sample into 4:1 training/testing sets and measuring the correlation between predicted expression and the truth in the testing set. Elastic net was chosen due to its ubiquity and the contrast between penalized regression to the “spike and slab” Bayesian framework of our method, as was similarly done in the TIGAR publication (Nagpal *et al.*, 2019). BayesR is a special case of BayesQN with $q = 1$, and is referred to as BayesQ1 for the remainder of this text. The Dirichlet process regression (DPR) is conceptually similar to the BayesQN method with an infinite number of mixtures, and two implementations are compared, one using the computationally intense Monte Carlo Markov Chain (DPR.MCMC) and another using a variational Bayes (DPR.VB) approximation to reduce computation time.

Considering all causal SNPs were assigned to a single category in these experiments, we expected performance to lag behind that of BayesQ1 due to signal dilution arising from the second category containing no true signal. Indeed, this is the case; the relative mean predictive R^2 of

BayesQ2 consistently lagged BayesQ1 by 3.1, 2.8, 3.0, 2.5, and 3.0% in scenarios with 2, 5, 10, 25, and 500 causal SNPs respectively.

The theoretical R^2 limit in these experiments is PVE_g, which was set at 12.5%. Under the scenario with two causal SNPs, the results of all five methods performed similarly well; BayesQ1 and BayesQ2 captured 84.2 and 81.7% of the theoretical maximum, while elastic net outperformed both capturing 89.9%. The two DPR methods exhibited the poorest performance of the five methods, recovering 77.2 and 80.6% of the theoretical maximum R^2 for the variational Bayes and MCMC methods respectively. While it is expected that DPR.MCMC would outperform DPR.VB, it underperformed both elastic net and BayesQN in the sparsest setting.

There is a trend of monotonically decreasing R^2 with an increasing number of causal SNPs across all five methods, except comparing the two most polygenic scenarios for the Bayesian methods (Figure 2.1). In this simulation framework, the effect sizes are drawn from the same distribution, so the increased polygenicity places smaller effects over a larger number of variants. Elastic net has been reported to underperform in highly polygenic scenarios, and we observe a decline in performance relative to other methods as the number of causal SNPs increases. As mentioned above, in the most sparse scenario elastic net outperforms all other methods, yet in the most polygenic scenario it underperforms the next best method by 16.1% (mean R^2 0.0624 vs. 0.0745 for DPR.MCMC). The other methods, however, perform similarly in the most polygenic scenario, with mean R^2 values of 0.0774, 0.0750, 0.0783, and 0.0745 for BayesQ1, BayesQ2, DPR.VB, and DPR.MCMC respectively. It is interesting to note that this is the only scenario under which the DPR.VB DPR method outperforms DPR.MCMC, hinting that slow convergence under high polygenicity could be hindering performance, however BayesQN similarly involves Monte Carlo simulation and this effect is not observed.

2.3.2 Variation in Quality of Causal SNPs

Having demonstrated that BayesQN performs comparable to established methods when causal variants do not span multiple designated categories, we sought to explore performance when causal SNPs are both well and poorly imputed ($\hat{r}^2$ cutoff of 0.8). The average proportion of poorly-imputed variants is only 8.4% across all samples implying that $\pi_1 \neq \pi_2$ and stochastic variation in a small sample would mean $\sigma_1 \neq \sigma_2$ with high probability. Given these considerations we chose 10 causal SNPs for our simulations under three scenarios: all well-imputed, all poorly-imputed, and 5 well-imputed / 5 poorly-imputed variants, again setting $PVE_g = 12.5\%$.

Table 2.1: A summary of R^2, MSE, and power comparing TWAS training methods varying the quality of causal SNPs.

Method	10 well-imputed SNPs			Half poorly-imputed SNPs			10 poorly-imputed SNPs		
	R^2	MSE	Power	R^2	MSE	Power	R^2	MSE	Power
BayesQ1	0.0887	0.056	0.823	0.0794	0.118	0.717	0.0436	0.343	0.293
BayesQ2	0.0860	0.049	0.800	0.0794	0.067	0.722	0.0533	0.099	0.407
Elastic Net	0.0877	0.225	0.810	0.0806	0.317	0.735	0.0525	1.323	0.399
DPR.VB	0.0852	0.118	0.787	0.0749	0.258	0.653	0.0224	3.920	0.120
DPR.MCMC	0.0869	0.062	0.801	0.0744	0.116	0.657	0.0383	0.250	0.256

Considering only well-imputed variants as causal, predictive perfomance is similar across all methods (Table 2.1), with mean R^2 values ranging only 0.0035. Surprisingly, the predictive performance using half well-imputed, half poorly-imputed variants is also quite uniform, ranging from a minimum of 0.0744 (DPR.MCMC) to 0.0806 (elastic net). This range increases when all SNPs are poorly-imputed: the DPR methods perform the worst (R^2 = 0.0224 and 0.0383 for VB and MCMC respectively), while BayesQ2 performs the best ($R^2 = 0.0533$). Interestingly, elastic net barely lags BayesQ2 ($R^2 = 0.0525$), but BayesQ1 unsurprisingly performs worse than ($R^2 = 0.0436$) both.

To assess the association testing performance, simulated phenotype data was generated with $PVE_e = 10\%$ and an effect size $\alpha = 1$, and predicted gene expression was regressed on phenotype using ordinary least squares regression. An increase in the variance of estimated association coefficients with increasing numbers of poorly-imputed causal variants was evident across all methods (Figure 2.2). This was most pronounced for elastic net and DPR.VB, which is partially

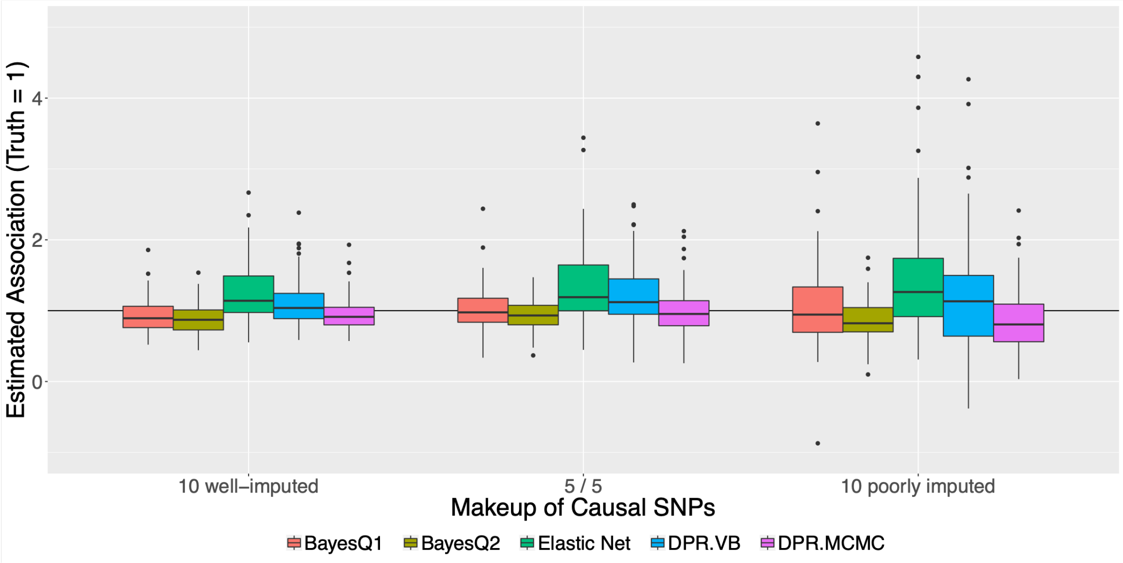

Figure 2.2: Comparison of five TWAS gene expression prediction methods using simulated data. The makeup of causal SNPs is displayed on the x-axis, and the predictive R^2 is on the y-axis. Center lines of the boxplots indicate median over all simulations, and the whiskers represent the IQR, with outliers displayed as points. The PVE_g is set to 12.5% and PVE_e is set to 10%. Several extreme outliers for DPR.VB are omitted to maintain scale.

reflected in the mean squared error (MSE), which increased 5.9 and 33-fold respectively comparing entirely poorly-imputed to entirely well-imputed causal SNPs. The massive increase observed in the MSE for DPR.VB is mainly driven by a small number of extreme outliers, which are not plotted. Slight negative bias is consistently observed with both BayesQN models as well as DPR.MCMC, with stronger positive bias observed with the other two methods.

Given the aim of TWAS, power in detecting signficant gene expression-trait associations is another important consideration in evaluating any TWAS method. When all causal variants are well-imputed, all methods perform relatively similar in terms of power (0.787 - 0.823) with BayesQ1 outperforming the others, however power decreases as much as 85% when comparing to entirely poorly-imputed causal variants. This decrease in power reflects the difficulty of modeling gene expression from poorly-imputed variants; some of this effect presumably being due to the confounding of MAF with imputation quality. When all causal variants are poorly-imputed, BayesQ2 is the only method to exceed 40% power, although elastic net falls just below this value.

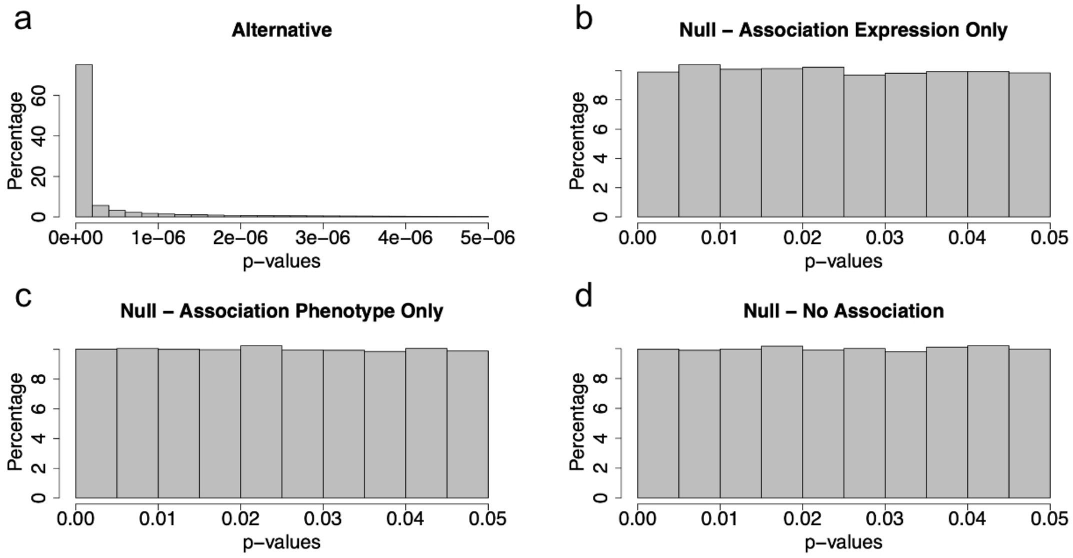

Figure 2.3: Calibration of BayesQ2 method under the alternative (a); the null hypothesis of association between genotype and gene expression but no trait-expression association (b); the null hypothesis of no association between genotype and gene expression but trait-expression association (c); the combined null (d). Histograms are truncated at 5×10^{-6} (a) and 0.05 (b-d)

The two DPR methods underperform the others under this scenario, with DPR.VB exhibiting only 12% power.

2.3.3 Calibration of BayesQN

Given the two-stage nature of TWAS, there are multiple null scenarios to consider: no association between genotype and gene expression, no association between predicted gene expression and trait, and the redundant null when both nulls are true. We found BayesQN to be well calibrated under all three null scenarios by examining the p-value distribution under each; additionally we contrasted these distributions against p-values under the single alternative (Figure 2.3).

2.3.4 Comparison to Elastic Net Using DGN

We next sought to examine the performance of our method using real-world data, opting for the publicly available DGN data. This dataset includes whole blood gene expression for over 900 subjects, and was particularly attractive considering these subjects were genotyped by panel, not whole genome sequencing as was done in other expression datasets. Consequently, we imputed the cohort's genotype, without resorting to artificial downsampling, to TOPMed freeze 5b using the Michigan Imputation Server, not only affording accurate imputation quality data but reproducing a representative pipeline likely to be followed by other groups.

Considering the similarity between BayesQN and elastic net in terms of both predictive performance and power using simulated data, we restricted our comparison to elastic net in this analysis as was done in the original TIGAR publication (Nagpal *et al.*, 2019). To ensure technical consistency, we fit models using elastic net rather than using the pre-trained weights offered by PrediXcan, resulting in 11,687 genes for which complete gene expression values were available and fitted models contained more than one SNP with a non-zero effect. Genomic regions for each predictive model consisted of $\pm$ 1 Mb from each gene's start and stop position, respectively, trimming the genotype to randomly remove one variant of each perfectly collinear pair.

Using the aforementioned set of genes, we directly compared BayesQ1 to BayesQ2 (Figure 2.4) in order to assess if any performance gain is observed using real-world data. Based on the previously discussed results using simulated data, the possibility existed that partitioning data would lead to overall predictive performance degradation if the majority GReX was driven solely by well-imputed variants. This does not appear to be the case when comparing model R^2; the mean values are 0.275 and 0.274, and median values are 0.221 and 0.220 for BayesQ2 and BayesQ1 respectively. A one-sided paired t-test of the alternative that BayesQ2 yields higher predictive R^2 values is marginally significant ($p = 0.039$). It should be noted R^2 values are not only similar in the aggregate, the majority of genes yield models whose predictive performances are close on an individual level. Of all genes tested, 85.8% have R^2 values within 0.025 of each other, and of those outside this range, BayesQ2 outperforms BayesQ1 in 190 additional genes.

Comparing BayesQ2 to elastic net, again on the basis of model R^2 without cross-validation (that is, the model is both trained and tested using all subjects), yielded less similar results. Excluding low imputation quality variants from the elastic net model fitting resulted in predictive

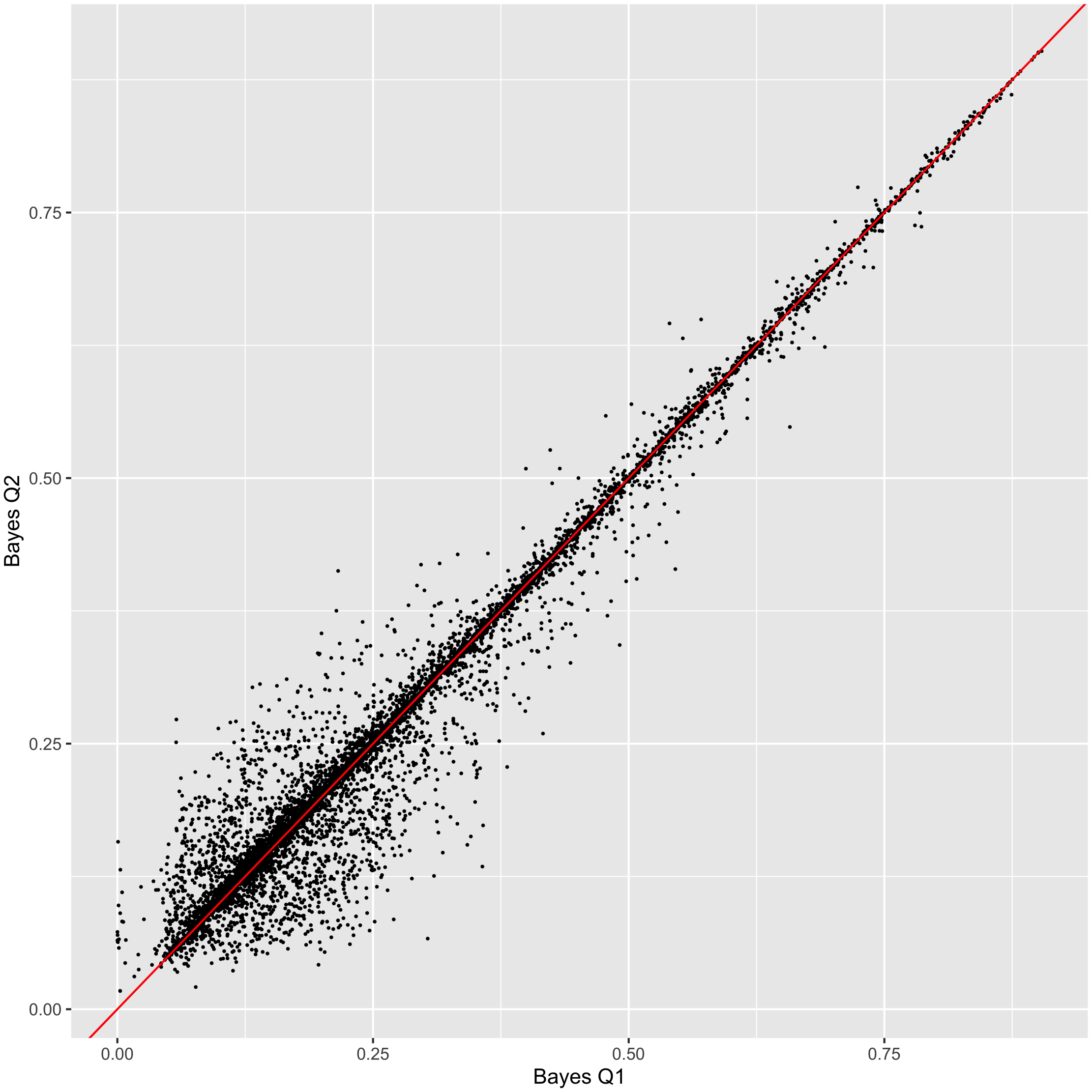

Figure 2.4: Scatterplot comparing the model R^2 values of BayesQ1 and BayesQ2 using 11,687 select DGN genes. The model R^2 values were obtained testing and training on the entire cohort. BayesQ1 R^2 values are on the x-axis and BayesQ2 R^2 values are on the y-axis. The red line represents equality.

performance that lagged behind BayesQ2 (Figure 2.5a), illustrating that these variants do indeed carry useful information for the prediction of gene expression. Mean R^2 was 0.255 for BayesQ2 vs. 0.238 for elastic net using only high quality variants, and median R^2 was 0.198 and 0.176 respectively for the two methods. However, the opposite relationships were observed when comparing the methods using the same set of variants (Figure 2.5b): mean and median R^2 values

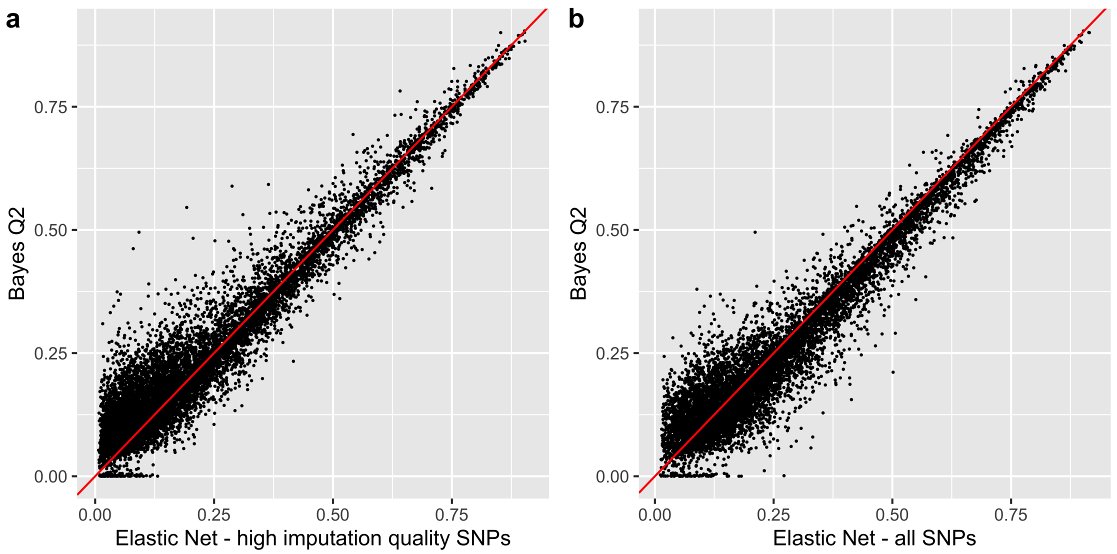

Figure 2.5: Scatterplot comparing the model R^2 values of elastic net and BayesQ2 using 11,687 select DGN genes. The model R^2 values were obtained testing and training on the entire cohort. Elastic net R^2 values using only high imputation quality SNPs ($\hat{r}^2 \geqslant 0.8$) (a) and using the same set of variants as BayesQ2 (b) are on the x-axis and BayesQ2 R^2 values are on the y-axis. The red line represents equality.

for BayesQ2 were both lower than those compared to elastic net (mean 0.255 vs. 0.266; median 0.198 vs. 0.212). Two-sided paired t-tests against the alternative that mean differences are not zero resulted in $p < 2.2 \times 10^{-6}$ for both cases. It is interesting to note, however, that the increased imputation accuracy of elastic net over BayesQ2 is not consistent over the range of R^2 values. Indeed, while elastic net outperforms BayesQ2 at high R^2, it underperforms at low R^2. This can be quantified by comparison of mean values conditioned on elastic net R^2 above or below 0.2; the mean model R^2 values are 0.401 and 0.375 for elastic net and BayesQ2 respectively under the former condition, and 0.116 and 0.128 under the latter. The low prevalence of very highly heritable GReX allows for the potential utility of the BayesQ2 method, while the possibility of model over-fitting in the case of high model R^2 motivated us to repeat these comparisons using cross-validation.

As expected, using 5-fold cross-validation reduced the number of genes with high model R^2 values. The number of genes with model $R^2 > 0.2$ either via elastic net (using all variants)

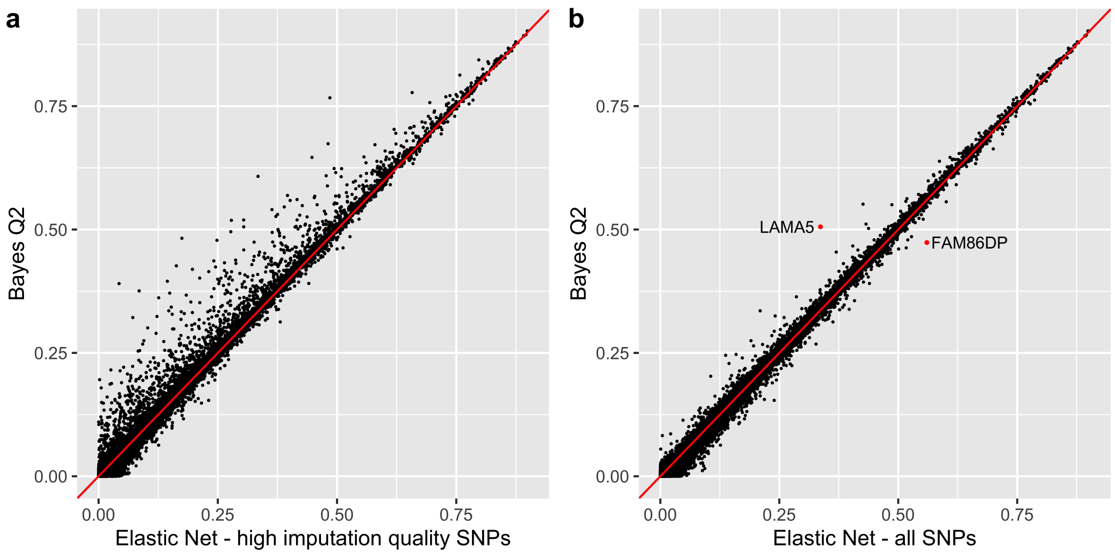

Figure 2.6: Scatterplot comparing the 5-fold cross-validation R^2 values of elastic net and BayesQ2 using 11,687 select DGN genes. Elastic net R^2 values using only high imputation quality SNPs ($\hat{r}^2 \geqslant 0.8$) (a) and using the same set of variants as BayesQ2 (b) are on the x-axis and BayesQ2 values are on the y-axis. The red line represents equality.

or BayesQ2 decreased from 5,442 before using cross-validation to 3,739 after, and a similar decrease was observed comparing to elastic net using only high quality variants (3,724 to 5,102). Furthermore, results between the two methods are overall more consistent, both when incorporating all or only high quality variants into elastic net models (Figure 2.6). While BayesQ2 does outperform elastic net using on high quality variants, the extent to which this occurs is visually overstated by Figure 2.6a. The difference in mean model R^2 is 0.004, more than half of which is accounted for by 168 genes where the BayesQ2 model R^2 exceeds that of elastic net by more than 0.1. There were no genes where the elastic net model R^2 exceeded that of BayesQ2 by more than 0.1.

In an attempt to gain insight into potential drivers for differential predictive performance between the two models, we examined the models for two genes: *LAMA5*, which represents the greatest R^2 differential in favor of BayesQ2, and *FAM86DP*, the analogous gene in favor of elastic net. For *FAM86DP*, there were three SNPs consistently in the top five with respect to absolute effect size over all BayesQ2 folds: 75395451, 75397556, and 75434806 (using bp position to identify SNPs). The first two are in very high LD ($R^2 = 0.975$) but the third SNP is essentially independent of the other two (highest $R^2 = 0.002$). Similarly, the top two variants with respect to absolute effect size over all elastic net folds were 75599696 and 75618257, which are independent signals ($R^2 = 0.001$). We hypothesized that given the largely similar overall predictive performance between the two methods, it is plausible that the lead variants in each pair of models would be correlated. This was not the observed however with the two models for *FAM86DP*; the four cross-model correlations between 75395451 and 75434806 from BayesQ2 and the two lead variants from elastic net were all < 0.001.

A similar analysis on *LAMA5* provided contrasting results. The top two BayesQ2 model SNPs were identical across all folds (61898661 and 61898954), and the SNP with the next largest effect size was 62339136, or in very high LD with it. Again, the top two variants were in very high LD with each other ($R^2 = 0.862$), but in low LD with 62339136 (highest $R^2 < 0.001$). However, the elastic model was not as consistent across folds; while SNP 61886187 had the most or second most absolute effect across all five folds, the next four highest effect sizes were very similar, and there was little consistency between folds. Furthermore, these variants were largely independent from each other and the BayesQ2 SNPs. The top variant for the BayesQ2 and elastic net model were highly correlated though ($R^2 = 0.743$).

While this model comparison is admittedly not rigorous, both in terms of scope (examining only two genes) and depth (comparing only top variants), it nonetheless shows some contrast between two models with opposite relative performances. Specifically, BayesQ2 outperforms elastic net in the case of similar models and lags in the case of dissimilar models. Paradoxically, the lead variants in both models for *FAM86DP* all fall below the imputation quality threshold of 0.8 yet elastic net outperforms BayesQ2, however the lead variants for *LAMA5* have high imputation quality and the opposite is true.

2.3.5 Cross-Cohort Evaluation

Finally, in order to most closely approximate a real-world TWAS application of training in one cohort and imputing into an independent cohort, we utilized the Multi-Ethnic Study of Atherosclerosis (MESA, Bild *et al.* 2002) dataset. This multi-site, longitudinal study was designed to investigate the prevalence and progression of cardiovascular disease and RNAseq data for a subset of the entire cohort over several of the five exams that comprise the entire study has been made available.

In the following analysis we compare BayesQ2 to elastic net by training each model with the previously described DGN population and then imputing into 356 European MESA subjects using TOPMed freeze 5b WGS genotype data, and comparing imputed GReX with the RNAseq measured values, adjusted for age, sex, study site, and the first ten principle components, followed by inverse normalization. One-thousand genes were randomly chosen from a set of those with model $R^2 > 0.1$ under both models, also adding *FAM86DP* and *LAMA5* which were not included in the original random sample. Of these 1,002 genes, we analyzed only 891 for which expression values for $> 25\%$ of subjects were available.

Given the similar predictive performance between BayesQ2 and elastic net in cross-validation experiments, it is not surprising that the cross-cohort R^2 values remain similarly close (Figure 2.7). The relative predictive performances of both models on *FAM86DP* and *LAMA5* remain unchanged; specifically, elastic net outperformed BayesQ2 for *FAM86DP* (R^2 = 0.239 vs. 0.185 respectively) while BayesQ2 outperformed elastic net for *LAMA5* (R^2 = 0.046 vs. 0.030 respectively). The gene for which we observed the largest disparity between imputation accuracy was CEACAM3 (R^2 = 0.533 for elastic net, 0.438 for BayesQ2). Interestingly, the model R^2 values in the training models were 0.164 and 0.248 for elastic net and BayesQ2 respectively. Not only were the relative performances with regard to cross-cohort imputation reversed, the imputed values greatly exceeded the respective model R^2 values.

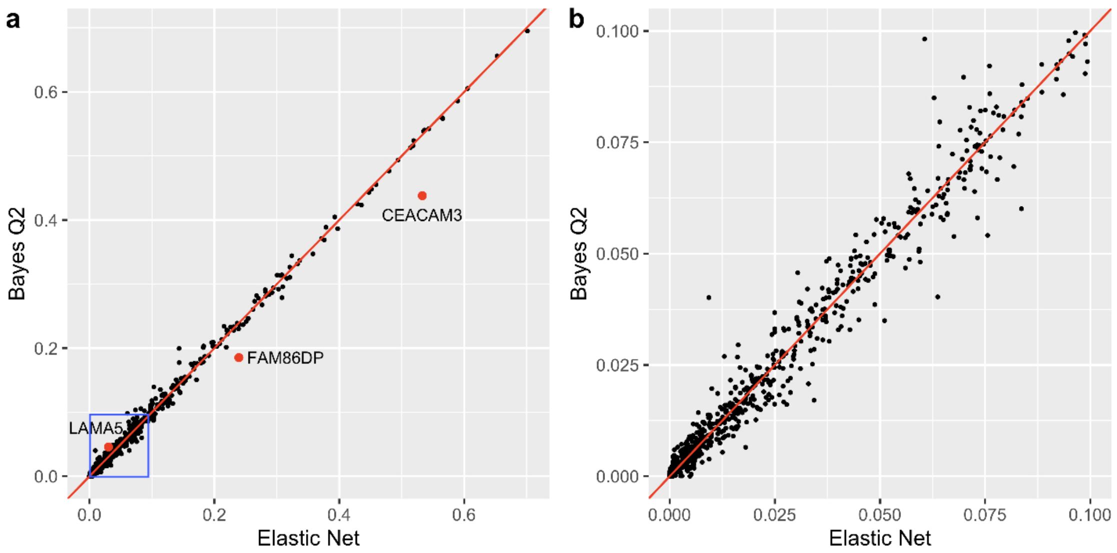

Figure 2.7: Scatterplot comparing the R^2 values of elastic net and BayesQ2 training in 891 select DGN genes and imputing into European subjects from the MESA cohort. (a) Each point represents one gene; three are highlighted for discussion. The blue box represents the area of (a) plotted in (b).

2.4 Discussion

We have demonstrated BayesQN, a TWAS gene expression prediction model that attempts to leverage the information in SNPs with both high and low imputation quality by allowing for separate distributional assumptions for the two groups. This method has several significant shortcomings however, perhaps most notably that it does not appear to outperform existing methods such as PrediXcan in real-world settings. The requirement for MCMC sampling necessitates a computational burden that grows linearly with the number of variants in the model, and processing times using BayesQN signficantly exceed that for for elastic net. The current implementation is in C++ via the Rcpp R package, while the original BayesR software improves computational time with a Fortran implementation but requires typed genotypes, not post-imputation dosages. While there is mounting evidence of significant *trans*-SNP contributions to GReX (Brynedal *et al.*, 2017; Liu *et al.*, 2019) we restrict our attention to *cis*-SNPs given the computational intractability of our current MCMC sampling genome-wide.

The comparison between the *FAM86DP* and *LAMA5* models hints at another potential shortcoming. We would expect that relative model performance would be greater in scenarios where effect alleles were categorized separately from the majority of the alleles, as in *FAM86DP*, and simulation results support this expectation. However, we observe the opposite effect, but noting that *FAM86DP* models were driven by a very small number of high effect SNPs using both BayesQ2, while the *LAMA5* models were far more polygenic. As demonstrated in simulations, elastic net performance declines with increasing polygenicity, and it is possible that in this one example these competing factors are dominated by polygenicity. A more fair comparison would require two genes with similar effect variant sparsity but differing in imputation quality, a goal for future work.

As imputation panels increase in size and population diversity, and whole genome sequencing for large cohorts becomes more financially feasible, the density of well-imputed SNPs in experiment cohorts will continue to increase, diminishing the need for categorization based on imputation quality. While we focus on this as the discriminating factor, there is no reason this cannot be extended to any factor suitable to categorization. Opportunities for future work include exploring other such factors, such as stratification based on population; our cross-cohort analysis was limited to training and testing in European populations. The current methodology requires that categorization be done *a priori*; extending the scope of BayesQN to adaptively categorize variants in a data-driven fashion is another opportunity for future efforts.

Given the failure to demonstrate any significant real-world predictive performance gains over elastic net, we did not continue with GReX-trait association testing. These analyses provide even more opportunities for future work.

CHAPTER 3: HPREP: QUANTIFYING REPRODUCIBILITY IN HICHIP/PLAC-SEQ DATASETS

3.1 Introduction

Chromatin spatial organization plays a critical role in genome structure and transcriptional regulation (Li *et al.*, 2018b; Schmitt *et al.*, 2016; Schoenfelder and Fraser, 2019). During the last decade, great strides have been made in the mapping of long-range chromatin interactions, thanks to the rapid development of chromatin conformation capture (3C) based technologies. Among them, Hi-C enables genome-wide measure of chromatin spatial organization and has been widely used in practice. To ensure scientific rigor, various methods have been developed to assess the reproducibility of Hi-C data (Yang *et al.*, 2017a; Ursu *et al.*, 2018; Yan *et al.*, 2017a; Sauria and Taylor, 2017; Yardimci *et al.*, 2019), as discussed in detail in Chapter 1. To recapitulate, HiCRep (Yang *et al.*, 2017a) first performs 2D smoothing to reduce the stochastic noise resulting from the sparsity of Hi-C data, and then quantifies reproducibility by calculating a weighted average of correlation coefficients between contact frequencies across specific one-dimensional (1D) genomic distance bands. Similar to HiCRep, GenomeDISCO (Ursu *et al.*, 2018) relies on data smoothing, which is performed over a range of steps of the random walk to determine an optimal separation between biological replicates and non-replicates as measured by area under the precision-recall curve. The reproducibility measure is a function of distances between two contact matrices smoothed using this optimized number of steps. HiC-Spector (Yan *et al.*, 2017a) adopts a different approach, transforming symmetric Hi-C contact matrices to their corresponding Laplacian matrices and then calculating similarity as the average of the distances between normalized eigenvectors. QuASAR-Rep (Sauria and Taylor, 2017) determines a local correlation matrix by comparing observed interaction counts to background signal-distance values within a 100-bin range. This local correlation matrix is transformed by element-wise multiplication with a matrix of scaled interaction counts and the reproducibility between two samples is defined as the Pearson correlation coefficient between the corresponding transformed matrices. These methodologies generally share the conceptual framework of data smoothing (with the exception of HiC-Spector) followed by a correlation calculation.

Recently, HiChIP (Mumbach *et al.*, 2016) and PLAC-seq (Fang *et al.*, 2016) technologies (hereafter collectively referred to as HP for brevity) have been developed to study protein-mediated long-range chromatin interactions at much reduced cost and greatly enhanced resolution relative to Hi-C. While the chromatin immunoprecipitation (ChIP) step involved in HP technologies allows for the cost and resolution benefits, it also introduces additional layers of systematic biases which make analysis methods developed for Hi-C data potentially unsuitable for HP data. To date, no method is available for quantifying reproducibility of HP data.

To fill in this gap, we propose a novel method, HPRep, to measure the similarity or reproducibility between two HP datasets. HPRep is motivated by HiCRep (Yang *et al.*, 2017a), the previously described method developed for quantifying reproducibility of Hi-C data. Similar to HiCRep, HPRep leverages the dependence of chromatin contact frequency on 1D genomic distance. In particular, HPRep models different ChIP enrichment levels, which contributes to the systematic biases specific to HP data.

3.2 Methods

Currently available methods to quantify reproducibility in Hi-C datasets, such as HiCRep, HiC-Spector, GenomeDISCO, and QuASAR-Rep (systematically evaluated in Yardimci *et al.* (2019)), all involve derivation of a similarity metric between two contact frequency matrices. The input Hi-C data consists of $n \times n$ symmetric matrices of non-negative integers, where each row/column represents one genomic locus (i.e., bin). The ij element of such a matrix represents the number of paired-end reads spanning between bin pair i and j.

These existing methods are conceptually inappropriate for HP data due to the ChIP enrichment bias introduced in the HP data. In addition, while Hi-C data consist of interactions between all bin pairs, HP data is restricted to bin pairs where at least one bin overlaps a binding region of the protein of interest. Such overlapping bins are referred to as the anchor bins. We further define bin pairs consisting of two anchor bins as the AND pairs, while those consisting of only one anchor bin are defined as the XOR pairs. In contrast, the NOT pairs, for which neither bin is an anchor bin, are not meaningful due to the nature of HP technologies and therefore are not used in HP data analysis (Juric *et al.*, 2019).

The basic data structure we consider is an $N \times m$ matrix, where N represents the number of anchor bins and m is 2 x 1 Mb/resolution. The ij element represents the normalized contact count between anchor i and the bin j units away, $j \in \{-m/2, \ldots, -1, 1, \ldots, m/2\}$. The number of anchor bins considered (N) is the union set of anchor bins derived from all samples in the study under analysis. Normalization is performed via a multi-step procedure: 1) Integer counts are adjusted for the biases introduced by effective fragment length, GC content, mappability, and ChIP efficiency by fitting a positive Poisson regression model (see Section B.1), following the approach detailed in the MAPS method (Juric *et al.*, 2019). Separate models are fit to the AND and XOR sets since the AND pairs are expected to have significantly higher contact frequencies due to double ChIP enrichment. 2) Using the fitted models, the data are normalized by taking the $\log_2$ value of (1 + observed / expected counts).

Similar to HiCRep, the distance metric used by our method is a weighted Pearson correlation that is stratified by distance (see Section B.2), which is identical for columns of the matrix which are symmetric from the center (Figure 3.1). Due to the sparsity of HP data, especially at longer distances, the normalized count values are smoothed (see Section B.3). The smoothing procedure used is a 1D arithmetic mean of values within a window of d bins away along the same row. Each of the $m/2$ correlations is weighted based on the variation of the smoothed values at that distance such that the weights sum to one. Therefore, the resultant metric is restricted to lie in $[-1, 1]$ and has a similar interpretation as a standard Pearson correlation coefficient.

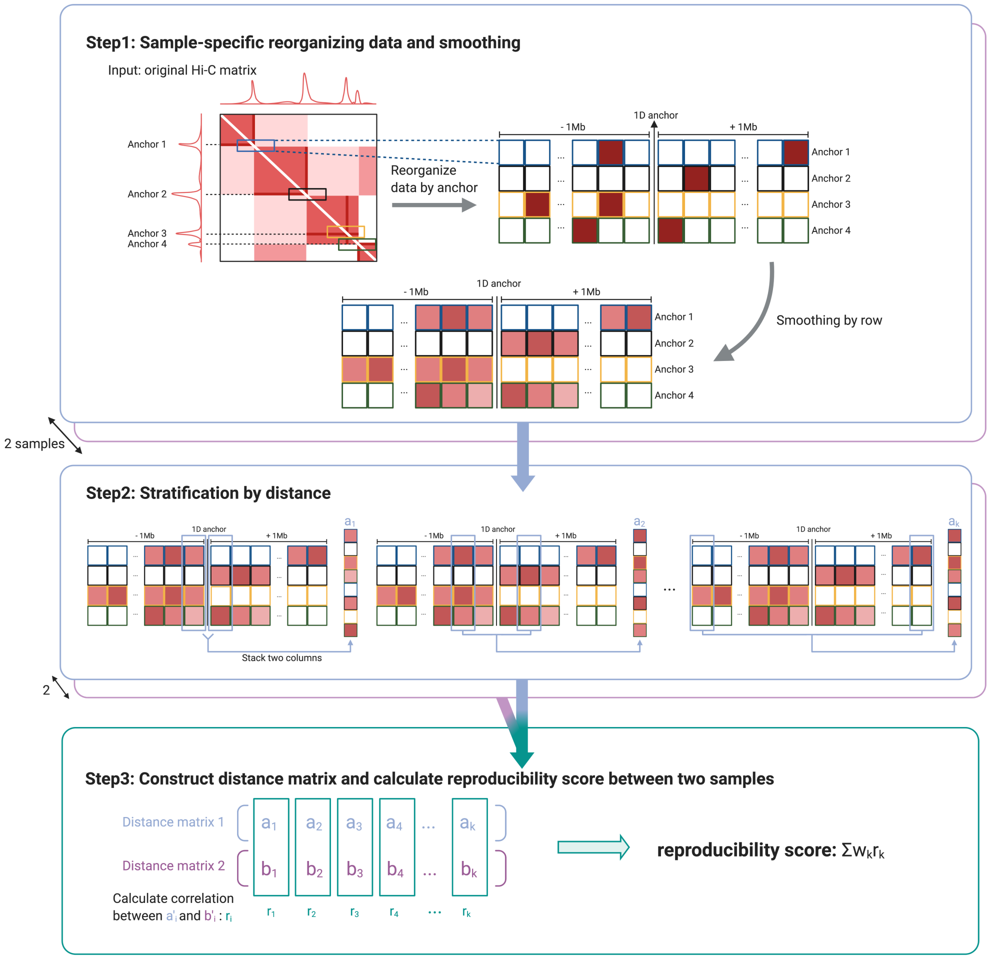
Step1: Sample-specific reorganizing data and smoothing
Input: original Hi-C matrix
Anchor 1
Anchor 2
Anchor 3
Anchor 4
Reorganize data by anchor
1D anchor
- 1Mb
+ 1Mb
Smoothing by row
2 samples
Step2: Stratification by distance
a1
a2
ak
Stack two columns
2
Step3: Construct distance matrix and calculate reproducibility score between two samples
Distance matrix 1
Distance matrix 2
a1 a2 a3 a4 ... ak
b1 b2 b3 b4 ... bk
Calculate correlation between a'i and b'i : ri
r1 r2 r3 r4 ... rk
reproducibility score: Σwkrk

Figure 3.1: Step 1 involves first identifying anchors (i.e., 1D ChIP peak sites) and then extracting all interactions between these anchors and bins within a specified genomic distance from the anchors. This is followed by a one-dimensional smoothing procedure. Stratification by distance is performed in step 2 such that the elements of vector a_k represent interactions that are equidistant from their respective anchors, k bins apart. In the final step, the Pearson correlation coefficients are calculated between a'_k from one sample and b'_k from another for all k, and these are combined in a weighted average to yield the final reproducibility metric.

Let a_k and b_k be two vectors of length $2N$ from samples a and b, whose elements are smoothed normalized contact counts, where N represents the number of anchor bins in the union set of anchor bins from all samples in a study, and k indexes bins that are $\pm k$ units away. Let a'_k and b'_k be the resulting vectors of length $N_k \leqslant 2N$ after removing any elements that are 0 in identical positions in both vectors. The weight for stratum k (w_k) is defined as

$$w_k = \frac{N_k \sqrt{\frac{\sum_{i=1}^{N_k} a_i'^2}{N_k} - \left(\frac{\sum_{i=1}^{N_k} a_i'^2}{N_k}\right)^2} \sqrt{\frac{\sum_{i=1}^{N_k} b_i'^2}{N_k} - \left(\frac{\sum_{i=1}^{N_k} b_i'^2}{N_k}\right)^2}}{\sum_{k=1}^{K} N_k \left(\sqrt{\frac{\sum_{i=1}^{N_k} a_i'^2}{N_k} - \left(\frac{\sum_{i=1}^{N_k} a_i'^2}{N_k}\right)^2} \sqrt{\frac{\sum_{i=1}^{N_k} b_i'^2}{N_k} - \left(\frac{\sum_{i=1}^{N_k} b_i'^2}{N_k}\right)^2}\right)} \tag{3.1}$$

where K is the total number of strata, which is analogous to the weighting used in HiCRep (Yang *et al.*, 2017a). The numerator of w_k is the product of strata size and the standard deviations of each stratum, while the denominator is the sum of these values over all strata. Consequently, the weights are restricted to $[0, 1]$ and sum to 1, where larger and more variable strata carry more weight than smaller and less variable strata.

3.3 Results

3.3.1 Mouse H3K4me3 PLAC-seq data

In order to examine the performance of HPRep, we first analyzed published H3K4me3 PLAC-seq datasets from mouse embryonic stem cells (mESCs) (Juric *et al.*, 2019) and mouse brain tissues (Yamada *et al.*, 2019), both consisting of two samples analyzed at 10 Kb resolution. Samples from the same cell type or tissue were labeled as biological replicates while those cross cell type or tissue were labeled non-replicates, yielding two pairs of biological replicates and four pairs of non-replicates. Pseudo replicates were generated by pooling the two samples of the same cell type or tissue and partitioning the pooled contact frequency in each bin pair randomly via Binomial (p = 0.5) sampling.

We would expect that pseudo replicates are most similar, followed by biological replicates, and that non-replicates are least similar. Indeed, this expected pattern is observed using HPRep (Figure 3.2), with results also exhibiting highly consistent patterns across chromosomes (Figure B.9). The higher metric for replicate mESC samples relative to mouse brain samples is due to the higher sampling depth of the former.

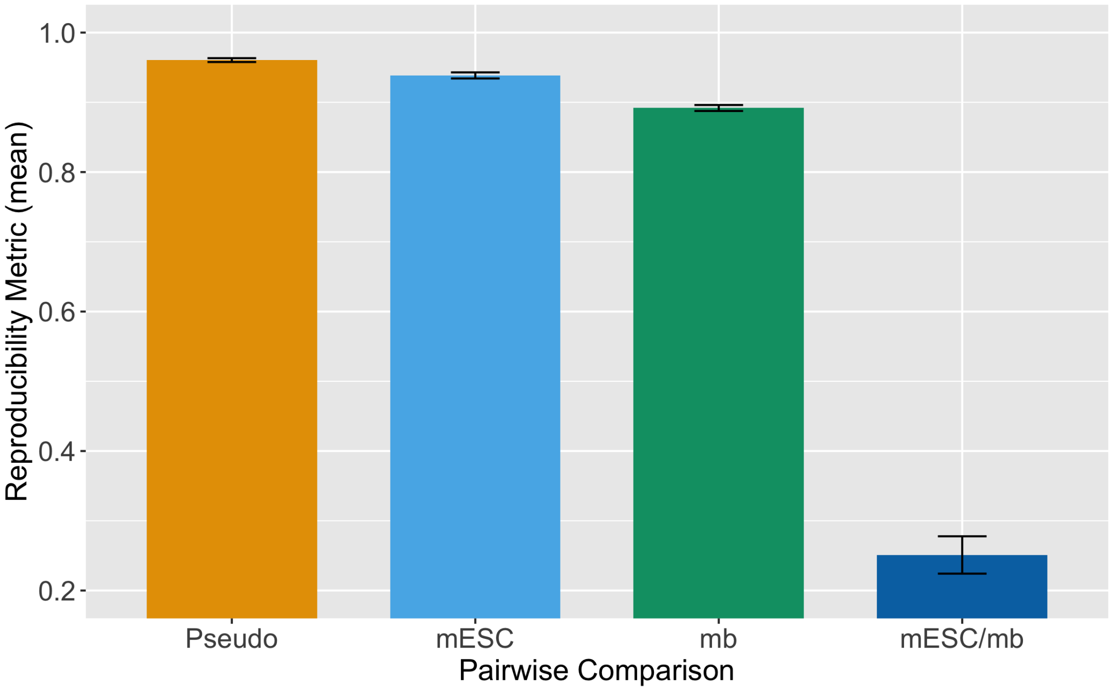

Figure 3.2: Metrics obtained applying HPRep to PLAC-seq data from mESC and mouse brain (mb) tissues. Pseudo replicates are generated by pooling two mESC samples followed by random sampling. Cross sample results represent the mean of four pairings. Results are presented as the mean value over 19 autosomal chromosomes with error bar representing $\pm$ 1 standard deviation.

We next compared HPRep with alternative methods, specifically two Hi-C reproducibility methods: HiCRep and HiC-Spector as well as a naïve Pearson correlation (see Section B.4). Since the Hi-C specific methods are designed using $n \times n$ symmetric contact matrices as the standard input, for these comparisons, in addition to restricting to bin pairs in the AND and XOR sets, we generated a pseudo Hi-C dataset from a HP dataset by also using bin pairs in the NOT set. The naïve Pearson correlation consisted simply of converting the entire upper triangular Hi-C contact matrices for each sample to single vectors and calculating the Pearson correlation coefficient between them. The methods were performed separately on all 19 autosomal chromosomes and the resulting metrics were reported as the arithmetic mean. The HiCRep and HiC-Spector methods were applied with the default parameters. The results are displayed in Figure 3.3.

All methods except for naïve Pearson correlation yielded results consistent with what we expected, namely higher similarity for the biological replicates and lower similarity for the non-replicates. The similarity or reproducibility metrics for the biological replicates were similar among these three methods, which is expected for HPRep and HiCRep, since both methods are based on stratified Pearson correlation, but is noteworthy for HiC-Spector since it is based on a rather different method, and furthermore is restricted to a different domain ($[0, 1]$ as opposed to $[-1, 1]$). The difference among these methods, with the exclusion of HiC-Spector when including the NOT set, manifests largely in values for non-replicates, with HPRep yielding much smaller values relative to the others, although in each case the four non-replicate pair results were very consistent. Interestingly, the naïve Pearson correlation nearly fails to correctly rank replicates with the mouse brain sample, yielding a reproducibility score almost identical to those of the non-replicates, whereas the result from mESC replicates is consistent with the other three methods. This failure is obviated in HiCRep and HPRep, the other Pearson based methods. For example, for biological replicates, HPRep yields a mean reproducibility metric of 0.92 compared to a mean value of 0.25 for non-replicates. For the experiments using bin pairs in the AND, XOR and NOT sets, the mean reproducibility metrics comparing replicates and non-replicates are 0.80 vs. 0.51, 0.99 vs. 0.73, and 0.88 vs. 0.76 for HiC-Spector, HiCRep, and Pearson correlation coefficients, respectively.

3.3.2 Human HiChIP data

In addition, we applied HPRep to measure the reproducibility of H3K27ac HiChIP data from GM12878 cells (two biological replicates) and K562 cells (three biological replicates) at 10 Kb resolution (Mumbach *et al.*, 2016) resulting in 4 pairs of biological replicates (1 pair from GM12878, 3 pairs from K562) and 6 pairs of non-replicates (Figure 3.4). We anticipated *a priori* that differences between replicates and non-replicates would be more pronounced in this human dataset than the previous mouse H3K4me3 PLAC-seq dataset due to the greater dissimilarity in H3K27ac anchor bins between GM12878 cells and K562 cells. Specifically, the GM12878 and K562 cell lines contain 31,980 and 26,963 H3K27ac anchor bins genome-wide (autosomal) respectively, with only 14,304 shared (Jaccard index 0.32). By contrast, mESC and mouse brain

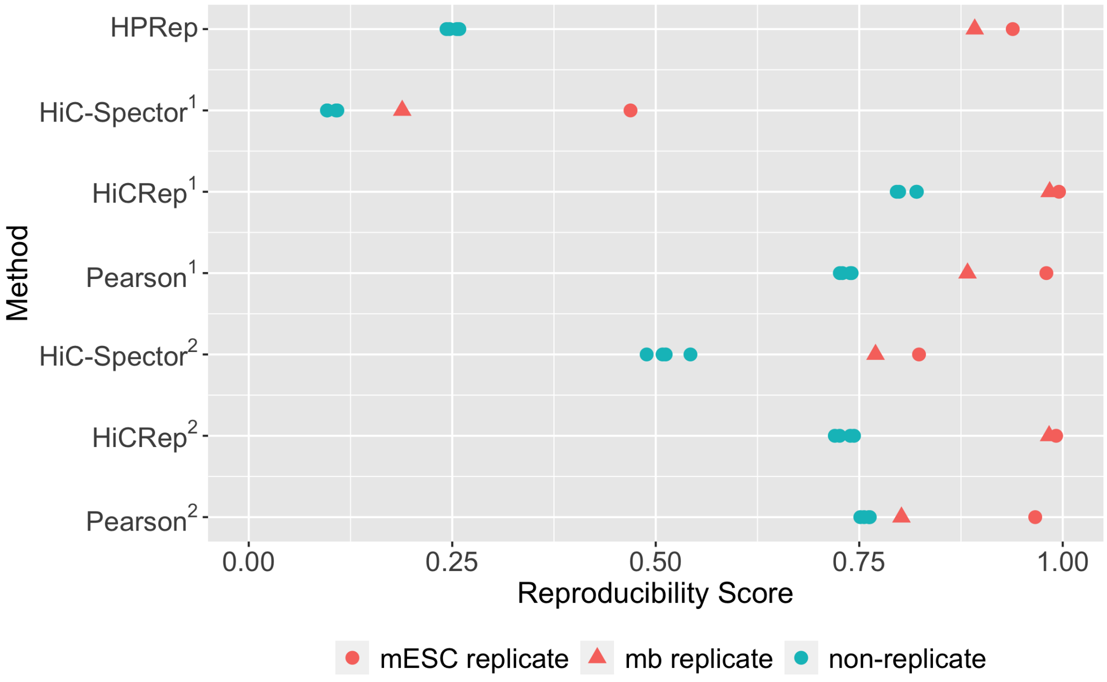

Figure 3.3: HPRep compared to Hi-C specific methods HiC-Spector and HiCRep as well as Pearson correlation. ^{1}: All methods using bin pairs in the AND and XOR sets. ^{2}: Methods other than HPRep using all bin pairs in the AND, XOR and NOT sets. PLAC-seq dataset consisted of two mESC and two mouse brain replicates.

have 28,903 and 21,778 H3K4me3 anchor bins, with 17,722 overlapping, (Jaccard index 0.54) which is not surprising given active promoters are largely shared across tissues and cell lines. For this human dataset, again, the methods were performed individually on all 22 autosomal chromosomes and the resulting metrics were averaged across chromosomes.

The results from the human HiChIP data are consistent with those from mouse PLAC-seq data: the biological replicates yield high similarity (close to 1) while the non-replicates yield uniformly lower similarity. While all autosomal chromosomes were used in these analyses and results were largely consistent across them using HPRep, HiCRep, and Pearson correlation coefficients, results were quite inconsistent using HiC-Spector (Figure B.9). Specifically, HiC-Spector used 20 eigenvectors in the computation of a reproducibility metric, yet for several chromosomes convergence failed so fewer eigenvectors were used which yielded erratic results (Table B.1).

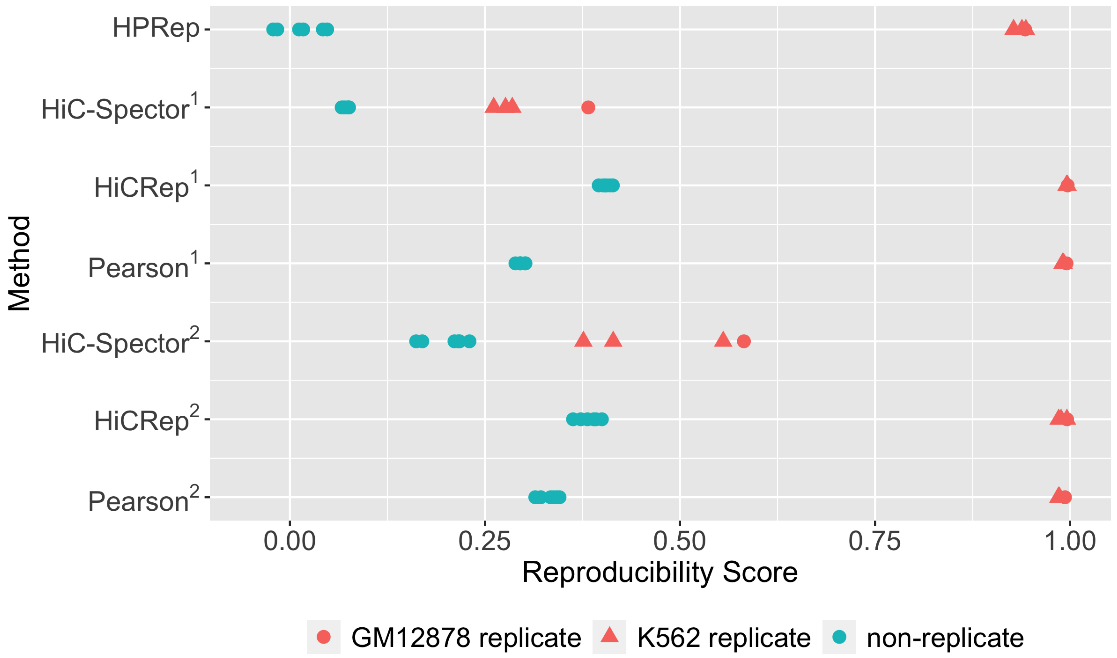

Figure 3.4: HPRep compared to Hi-C specific methods HiC-Spector and HiCRep as well as Pearson correlation. ^{1}: All methods using bin pairs in the AND and XOR sets. ^{2}: Methods other than HPRep using all bin pairs in the AND, XOR and NOT sets. HiChIP dataset consisted of two GM12878 replicates and three K562 replicates.

Again, HPRep results in the lowest metrics for the non-replicates which are all close to zero, highlighting the influence on anchor bin identity in this method.

3.3.3 Human PLAC-Seq data

We next applied HPRep to a more complex H3K4me3 PLAC-seq dataset at 5 Kb resolution, consisting of 11 samples from four brain cell types in human fetal brain obtained via fluorescence-activated cell sorting (Song *et al.*, 2020): 3 samples from neurons (N), 3 samples from interneurons (IN), 2 samples from radial glial (RG), and 3 samples from intermediate progenitor cells (IPC). These samples have varying sequencing depths (detailed in Supplementary Table 2 of Song *et al.* 2020), with number of *cis* reads ranging from 47.5 million for RG2 (second replicate for RG cell type) to 390 million for RG1. The anchor bins are the union of 1D H3K4me3 peaks from all 4 cell types. In Figure 3.5a, reproducibility obtained by HiCRep shows

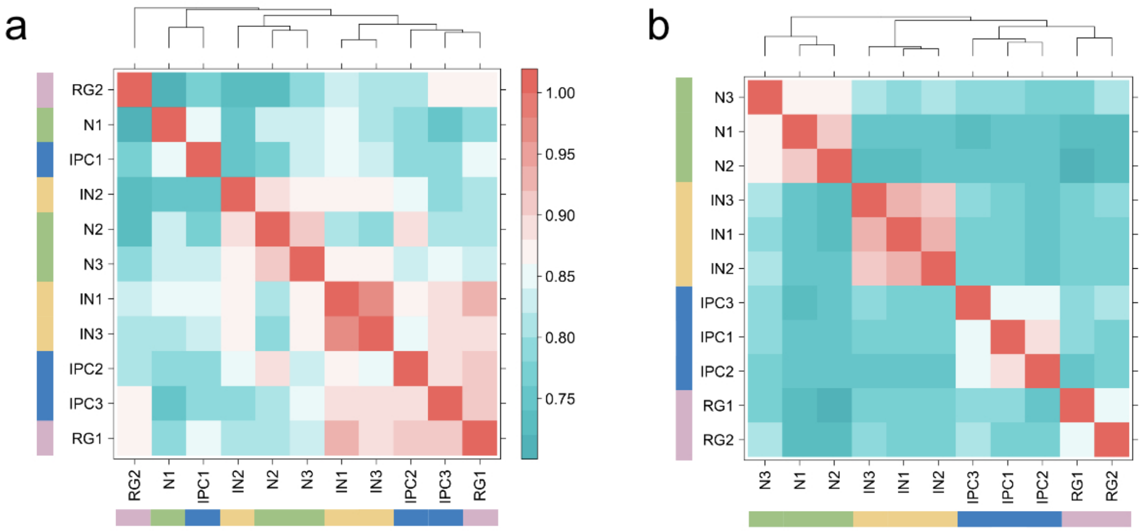

Figure 3.5: HPRep compared to HiCRep. HiChIP dataset consisted of three neuron (N), 3 interneuron (IN), 2 radial glial (RG), and 3 intermediate progenitor cell (IPC) samples. The color scale indicates low (blue) to high (red) reproducibility metric.

no differentiation between inter- and intra-cell types. In contrast, HPRep shows a clear pattern of higher similarity for replicates from the same cell type compared to those from different cell types (Figure 3.5b).

Focusing on bin pairs in the AND and XOR sets highlights the effect of normalizing ChIP enrichment level bias. Figure 3.6 is analogous to 3.5a excluding bin pairs in the NOT set. The cell type clustering is more in line with the known truth, however, still has misspecifications according to the dendrogram: neuron and interneuron cells are correctly grouped, but radial glial cells are not.

Recent studies have shown that HiCRep is sensitive to sequencing depth (Yardimci *et al.*, 2019). To evaluate the robustness of HPrep with respect to different sequencing depths, we performed down-sampling to the original PLAC-seq data from 4 human brain cell types. This was performed by sampling from a multinomial distribution with n equal to the new total count and count probabilities set to match the distribution in the corresponding original data (see Section B.5).

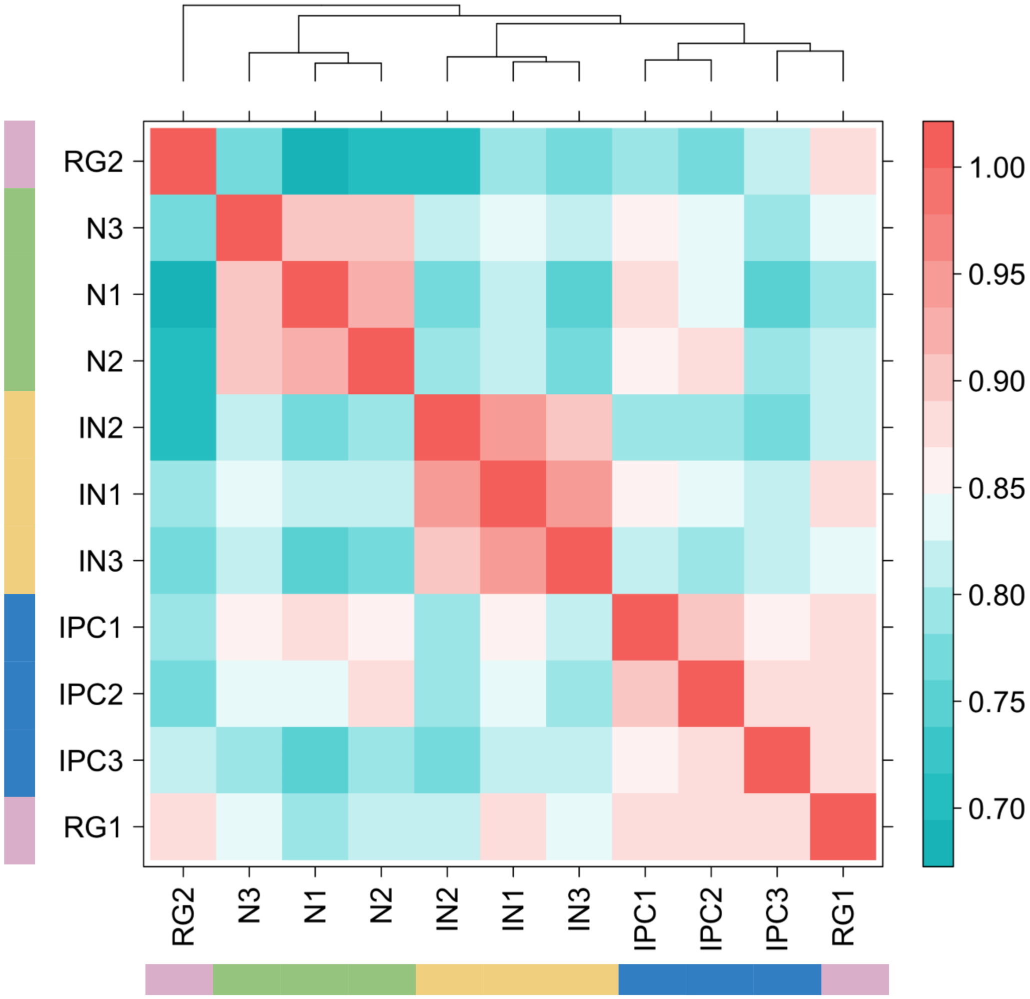

Figure 3.6: HiCRep. HiChIP dataset consisted of three neuron, 3 interneuron, 2 radial glial, and 3 intermediate progenitor cell samples excluding interactions where neither bin overlapped with an anchor. Red color signifies results indicating stronger correlation.

The first down-sampling was performed such that all samples matched the depth of the sample which had the lowest sequencing depth (RG2). Note the identical color scales for Figures 3.5b and Figure 3.7, but the decrease in metric values for many pairwise comparisons for samples of the same cell type such as the interneuron cells. In order to quantify this reduced discernibility

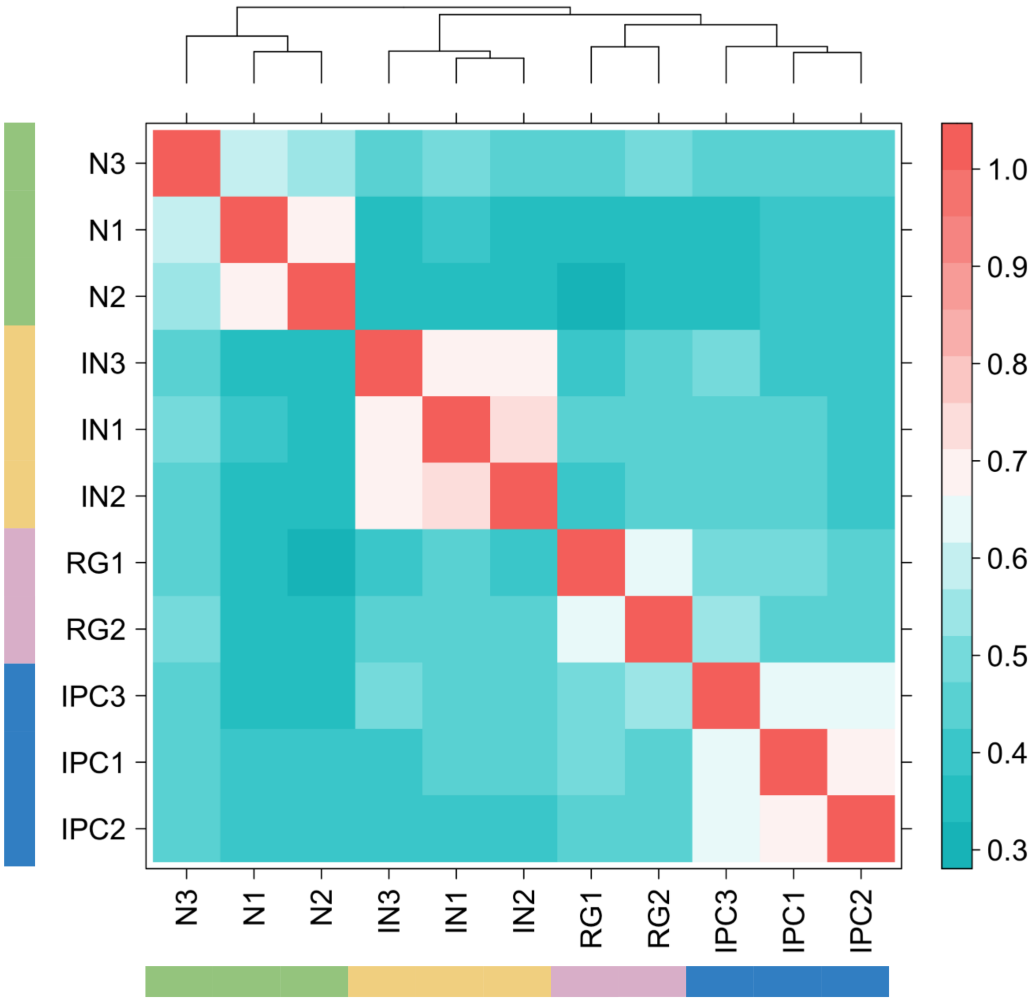

Figure 3.7: HPRep results obtained after down-sampling all eleven samples to read depth of that of the lowest sample.

between samples, we utilized the silhouette procedure (Rousseeuw, 1987), treating reproducibility score as a distance metric and reporting the average of the 11 silhouette values, one for each sample (see Section B.6). We obtain 0.717 and 0.685 for the original experiment and down-sampled results respectively, where smaller numbers indicate worse clustering performance.

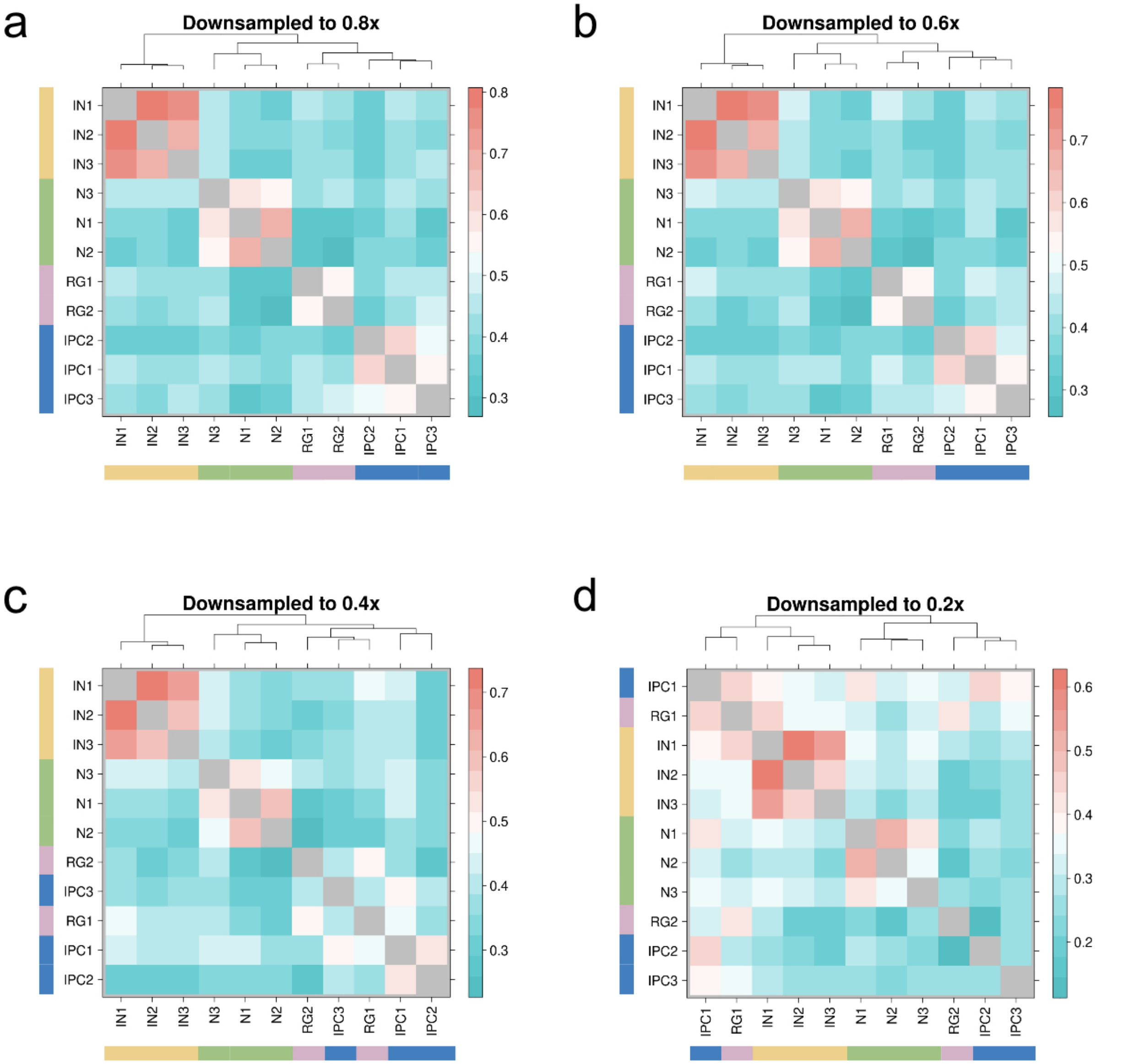

Figure 3.8: HPRep results obtained after down-sampling each sample to by specified factor. a) 80% of original depth of each sample, b) 60% of original depth, c) 40% of original depth, d) 20% of original depth. Note that the diagonal is now gray to remove it from the scaling in order to better highlight differences.

Subsequent down-sampling was performed uniformly across all samples such that total counts were reduced to 80%, 60%, 40%, and 20% of their original values following the same sampling protocol as described above. As expected, in Figure 3.8 we observe decreased discernibility among samples from different cell types, most strikingly with IPC and RG where the within sample HPRep reproducibility metric dropped to as low as 0.26 and 0.43, respectively. Applying the modified silhouette procedure described above to these four down-sampled datasets, we obtained a silhouette score of 0.700, 0.678, 0.634, and 0.518 for down-sampling to 80%, 60%, 40%, and 20% respectively.

We next sought to investigate the extent to which our HPRep metric was driven by the 1D ChIP (anchor) signals relative to the 3D bin contact signals. To this end, we compared the irreproducible discovery rate (IDR, Li *et al.* 2011) (see Section B.8) to the HPRep results utilizing the highest read depth brain sample (RG1). This was accomplished by pairwise comparisons between the original ChIP-Seq data (IDR) or AND/XOR data (HPRep) and corresponding samples that had been downsampled to 80%, 60%, 40%, and 20% to the original depth. As expected, both IDR and HPRep metrics decreased with more aggressive downsampling, however, the effect on IDR, as measured by fraction of peaks passing a false discovery rate threshold of 5%, was far more pronounced. HPRep metrics were 0.97, 0.96, 0.93, and 0.88 compared to IDR of 0.80, 0.68, 0.24, and 0.06 at 80%, 60%, 40%, and 20% of the original depth, respectively. This effect difference suggests that 1D information does not dominate our results; if the HPRep results were merely a reflection of anchor similarity, we would expect a more consistent trend between the two experiments.

3.4 Discussion

Quantification of data reproducibility is critical to ensure scientific rigor, yet methods tailored for HiChIP and PLAC-seq data are still lacking. Here, we propose HPRep, the first model-based approach to account for ChIP enrichment biases in measuring HP data reproducibility. Given the lack of HP specific tools, we compare HPRep to existing methods designed for Hi-C data, specifically HiCRep and HiC-Spector. Additionally, since our method, similar to HiCRep, relies on a weighted average of Pearson correlation coefficients, we also compare HPRep to the naïve Pearson correlation coefficient.

Our HPRep method, improving on existing Hi-C specific methods, is tailored to HP data for the measurement of reproducibility in two fundamental ways. First, HPRep is designed around the specific structure of HP data, namely that while Hi-C data consists of contact frequencies for all bin pairs, HP data focuses on bin pairs where at least one bin overlaps with a ChIP-seq peak for a protein of interest. This is different from the standard $n \times n$ symmetric Hi-C contract matrix. We focus the data matrix on anchor bins, regions that overlap with ChIP-seq peaks, and pairs between bins within a specified window of these anchors as illustrated in Figure 3.1.

Second, HPRep fits a positive Poisson regression model to normalize HP-specific biases from ChIP enrichment level, and used the residuals as the normalized contact frequencies. It also analyzes bin pairs in the AND and XOR sets separately, effectively accounting for ChIP enrichment bias for the two different types of bin pairs.

Our results from mouse H3K4me3 PLAC-seq data demonstrated very low variability in metrics between chromosomes (Figure B.9), which is consistent with HiCRep (Figure B.9). In addition, we also compared HPRep with other existing methods using H3K27ac HiChIP data from GM12878 and K562 cells, as well as H3K4me3 PLAC-seq data from 4 human brain cell types. Our results demonstrated the superior performance of HPRep, in terms of accurate clustering of samples from the human brain cell types which was not achievable using HiCRep, although better clustering accuracy was observed when excluding bin pairs in the NOT set.

Future work involves exploring the potential of using this method to determine minimum per-sample sequencing depth or maximum allowable (if any) differential depth across samples for accurate quantification or HP data reproducibility. We show that sample differentiation and expected clustering can survive down-sampling, but rigorous experimentation needs to be conducted in order to demonstrate practical use, as more high-depth HP data become available from more tissues, cell lines or cell types. Additionally, we plan to examine the use of this general framework with capture Hi-C datasets, including those targeting at a relatively small number of loci centered at regions identified from genome-wide association studies, and those genome-wide promoter capture Hi-C experiments. These extensions are highly warranted, but are beyond the scope of our current HPRep work, due to the complexities of these data and their specific features.

In terms of computational efficiency, for the human PLAC-seq data consisting of 11 samples, tuning the smoothing parameter and determining all 55 pairwise reproducibility metrics for all 22 autosomal chromosomes took 1 hour and 5 minutes using a single core on a 2.50 GHz Intel processor with 4GB of RAM. One can choose to apply HPRep to one chromosome to obtain nearly identical results. On the same data, HPRep takes 35 minutes to perform tuning and analysis on solely chromosome 1 using the same single core.

CHAPTER 4: HPTAD: TAD DETECTION IN HICHIP/PLAC-SEQ DATASETS

4.1 Introduction

The examination of the spatial organizational structure of DNA has been greatly enhanced by the development of chromatin conformation capture technologies (Dekker *et al.*, 2002). Specifically, Hi-C (Lieberman-Aiden *et al.*, 2009) has allowed for the analysis of genome-wide interactions between genetic loci at very high resolution. This has led to the identification of various structural elements such as chromatin loops (Rao *et al.*, 2014), chromosomal compartments (Lieberman-Aiden *et al.*, 2009), and topologically associating domains (TADS, Dixon *et al.* (2012)).

TADs are characterized as contiguous sections of the genome where within-region interactions are more frequent than between-region interactions. Since their first formal description in 2012 (Dixon *et al.*, 2012; Nora *et al.*, 2012), the functional relevance of TADs continues to be investigated; they have been implicated in multiple cellular contexts (Sikorska and Sexton, 2020), and modification or destruction of TAD boundaries have been associated with cellular dysfunction (Lupiáñez *et al.*, 2015; Ren and Dixon, 2015) and cancer (Valton and Dekker, 2016; Li *et al.*, 2019; Akdemir *et al.*, 2020; Pinoli *et al.*, 2020).

In the last decade many methods have been proposed for the detection of TADs. The approaches taken by these methods vary widely; many use metrics derived directly from contact frequency to detect TAD regions and boundaries (Filippova *et al.*, 2014; Durand *et al.*, 2016; Zhan *et al.*, 2017; Ardakany and Lonardi, 2017; Dixon *et al.*, 2012; Yu *et al.*, 2017; Ramírez *et al.*, 2018; Wang *et al.*, 2017; Crane *et al.*, 2015; Malik and Patro, 2019; Shin *et al.*, 2016; An *et al.*, 2019) while others use statistical (Lévy-Leduc *et al.*, 2014; Ron *et al.*, 2017; Xing *et al.*, 2021; Serra *et al.*, 2017; Weinreb and Raphael, 2016), cluster (Wang *et al.*, 2015; Oluwadare and

Cheng, 2017; Haddad *et al.*, 2017; Soler-Vila *et al.*, 2020), network-based (Norton *et al.*, 2018; Yan *et al.*, 2017b; Lyu *et al.*, 2020), and machine learning (Stilianoudakis and Dozmorov, 2020) methods. Results obtained using these methods have been shown to vary greatly in terms of number, size, and enrichment of measures of biological relevance (Zufferey *et al.*, 2018) even when controlling for factors such as technical variation, read depth, etc.

While high-depth, unbiased Hi-C datasets will remain the preferred input data for TAD detection, cost limitations can make acquisition of such input data infeasible. Recent advances in chromatin conformation capture technologies have provided alternatives that can achieve kilobase resolution while requiring less biological input material and at reduced cost relative to Hi-C. Two such methods, HiChIP (Mumbach *et al.*, 2016) and PLAC-seq (Fang *et al.*, 2016) (herein referred to HP for brevity) achieve this by combining chromatin immunosuppression (ChIP) with in-situ Hi-C to specifically target interactions bound by specific proteins or histone modifications. HP data is primarily used for the detection of enhancer-promoter interactions at high resolution (5 or 10 Kb), however considering the similarity to Hi-C data we explored the use of HP data to identify low resolution (40 Kb) TADs.

We present HPTAD, a TAD caller designed for use with HP, rather than Hi-C, data as input. We compare its performance against several publically available TAD callers using mESC and GM12878 HP datasets relative to ground truth TAD boundaries called from Hi-C experiments in the respective cell types. Additionally, we demonstrate good consistency between biological replicates and CTCF enrichment at TAD boundaries called with HPTAD, a biological feature frequently associated with TAD boundaries (Dixon *et al.*, 2012).

4.2 Methods

4.2.1 HPTAD Method

During the pre-processing step, intra-chromosomal reads are split into two groups: short-range reads ($\leqslant$ 1 Kb) and long-range reads ($>$ 1Kb). The short-range reads are used as a measure of ChIP efficiency in the regression framework described in Equation 4.1. Long-range reads are used to determine long-range interactions, which are extracted and classified as either AND, XOR, or NOT sets based on whether 2, 1, or 0 (respectively) read ends overlap with a ChIP-seq identified peak for the protein of interest. Additional details can be found in the MAPS paper (Juric *et al.*, 2019).

The subsequent HPTAD method follows a four step procedure:

1. We model the non-zero intra-chromosomal contacts as a zero-truncated Poisson model with mean μ_{ij}. We consider the following covariates: effective fragment length (FL), GC content (GC), mappability (MS), ChIP enrichment level (IP), and 1-D distance (D). The values used in the regression are log($x_i \times x_j$), where x_i and x_j are the corresponding covariates for bins i and j respectively. Unless otherwise stated, the bin size is 40 Kb. We fit regression models for the AND, XOR, and NOT sets separately.

   $$\log(\mu_{ij}) = \beta_0 + \beta_1 \cdot FL_{ij} + \beta_2 \cdot GC_{ij} + \beta_3 \cdot MS_{ij} + \beta_4 \cdot IP_{ij} + \beta_5 \cdot D_{ij} \tag{4.1}$$

   Fitted values are normalized as the ratio of observed to expected (fitted) count.

2. Let X_{ij} be the normalized count between bins i and j. Next, for each bin b we calculate the mean of X_{ij} for all (i, j) such that

   $$\{(i, j) : b - w < i \leqslant b, b < j \leqslant b + w\}$$

   for specified window size w measured in bin units. Let X_b represent this value, then we record the score as

   $$\text{Score} = \log_2\left(\frac{X_b}{\bar{X}}\right) \tag{4.2}$$

   where

   $$\bar{X} = \frac{\sum_{b=1}^{B} X_b}{B} \tag{4.3}$$

   and B is equal to the total number of bins.

3. The resulting vector of scores is smoothed using the Nadaraya-Watson kernel regression estimate as implemented in the R function “ksmooth”, using a box kernel with bandwith 3. From this smoothed vector of scores we select candidate TAD boundaries as the set of points meeting the following criteria: 1) Using an approximation to the second derivative the score is identified as a minimum with positive second derivative and 2) the value is below zero (Figure 4.1b).

4. From this set of candidate boundaries we choose final TAD boundaries by first assuming $X_{ij} \sim N(\mu_{ij}, \sigma^2)$. We define b_t as the locus (bin) corresponding to candidate TAD boundary t, and define domain (D_T) as the TAD region bounded by candidate boundaries $t-1$ and t. Therefore,

$$D_T = \{(i,j) : b_{t-1} \leqslant i < j < b_t\} \tag{4.4}$$

We make the assumption that intra-TAD interactions share a common mean, that is

$$\mu_{ij} = \mu_T \text{ if } (i,j) \in D_T \tag{4.5}$$

Further, we define an exterior region between two tads (E_T) as

$$E_T = \{(i,j) : b_{t-1} \leqslant i < j < b_{t+1}\} - D_T - D_{T+1} \tag{4.6}$$

as illustrated in Figure 4.1c. Now we assume

$$\mu_{ij} = \mu_{E_T} \text{ if } (i,j) \in E_T \tag{4.7}$$

and therefore we would expect $\mu_T = \mu_{T+1} = \mu_{E_T}$ if t is a boundary. We formally test this null hypothesis against the alternative using a likelihood ratio test.

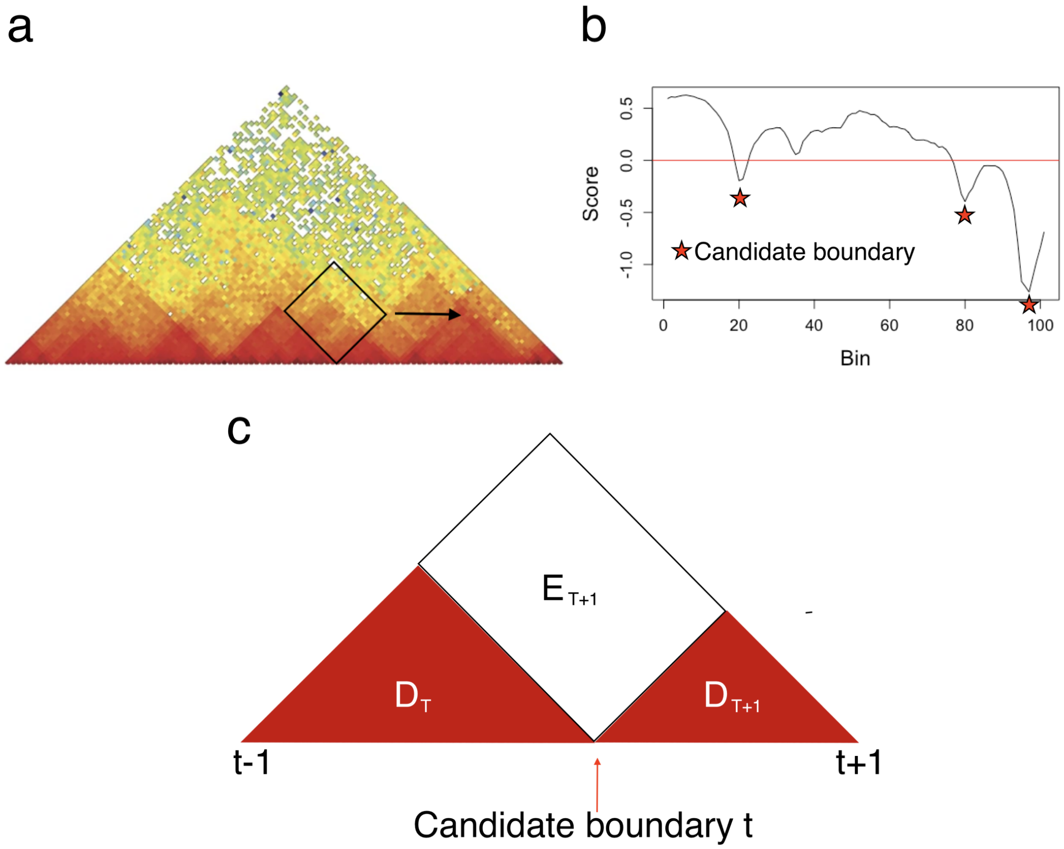

Figure 4.1: Cartoon illustration of HPTAD pipeline. We calculate the mean normalized contact frequency in a square region, sliding the window along the diagonal (a). We determine candidate boundaries from scores derived from these values (b) and then formally test these candidates. Diagram (c) is provided to illustrate the regions defined by Equations 4.4 and 4.6. The red areas are separate TAD regions assuming candidates $t-1$, t, and $t+1$ are actual boundaries, however all three illustrated regions are a single TAD if candidate t isn't a boundary.

4.2.2 Jaccard Index

We utilize a modified Jaccard index for comparing sets of TAD boundaries. The intersecting set contains boundaries that are within $\pm$ 1 bin unit. Recognizing that such an offset could introduce double counting of a boundary within 1 bin unit of two others, we disallow any boundary being counted more than once.

Let A and B be two sets of TAD boundaries, and let r represent the resolution of the analyses in base pairs. The intersecting set I is comprised of elements $b \in B$ such that there exists at least one element $a \in A$ where $a - r \leqslant b \leqslant a + r$. We sequentially test elements b; if one $a \in A$ satisfies $a - r \leqslant b \leqslant a + r$ it is removed from A to prevent double counting and b is added to intersecting set I. If more than one $a \in A$ satisfies $a - r \leqslant b \leqslant a + r$ the lowest value is removed from A.

The modified Jaccard index J is then defined as:

$$J = \frac{|I|}{|A| + |B| - |I|} \tag{4.8}$$

where $| \cdot |$ represents the cardinality of the corresponding set.

4.2.3 Measure of Concordance

The measure of concordance (MoC) is a metric typically used to compare clustering assignments (Pfitzner *et al.*, 2009), but has been used to compare TAD regions (Zufferey *et al.*, 2018), considering them clusters of contiguous bins. Let A and B be two sets of TAD regions with cardinality N_A and N_B respectively, with each region being defined as a range of contiguous bin intervals. Let I_{ij} be the set of bins in both TAD regions A_i and B_j, and let $| \cdot |$ indicate the size of the corresponding TAD region in bins. We then define the MoC between A and B as:

$$MoC(A, B) = \frac{1}{\sqrt{N_A N_B} - 1} \left(\sum_{i=1}^{N_A} \sum_{j=1}^{N_B} \frac{|I_{ij}|^2}{|A_i||B_j|} - 1 \right) \tag{4.9}$$

This metric is a formal measure, and is restricted to the domain [0, 1], which makes it appealing as an easily interpretable metric. Values closer to 1 indicate better agreement between TAD regions, with 1 achieved if and only if $A = B$.

4.3 Results

4.3.1 Performance assessment using “ground truth”

A persistent issue that arises when analyzing or comparing TAD regions and boundaries is the lack of an objective ground truth. The definition of a TAD as a section of the genome in which within-region interactions are more frequent than between-region interactions does not specify a quantitative threshold by which to judge this differential. Consequently, significant variability exists in what constitutes a TAD depending on what criteria are used to define them, as evidenced by the widely varying results obtained from numerous published TAD callers (Zufferey *et al.*, 2018). Despite this, some boundaries have been reported to be reproducible not only across samples, but are also preserved between different cell types in the same organism (Dixon *et al.*, 2015). We used TAD boundaries identified from deeply sequenced Hi-C data as the working truth to compare the relative performances of TAD callers.

Another issue when comparing TAD regions and boundaries is that different methodologies define boundaries as points (bp) or regions (bin). Moreover, many methodologies define TADs such that they span the entire chromosome while others specifically identify undefined regions not classified as part of any TAD. To ensure consistent comparisons, we define TAD boundaries as bins and exclude undefined regions in methods where they are specified rather than folding them into a neighboring TAD or splitting between two neighbors. Consequently, we utilize a modified Jaccard index for the purpose of comparing TAD boundaries. Specific details can be found in Methods 4.2.2, but the key modification to the standard Jaccard index is that we consider boundaries that overlap within $\pm$ one bin to be matched. The interpretability of this modified Jaccard index is unchanged from the standard index.

In addition to the Jaccard index we use the measure of concordance (MoC, Methods 4.2.3) to compare the TAD regions, originally described in Pfitzner *et al.* (2009). This is akin to viewing TAD regions as a clustering of bins with the constraint that clusters contain only contiguous bins. One advantage of using the MoC to compare TAD regions is its ease in interpretability: the response is restricted to [0, 1], with higher values indicating closer alignment of TAD regions. A value of 1 is achievable if and only if the set of regions overlap perfectly.

We compared the performance of HPTAD to four publically available TAD callers (OnTAD, Grinch, TopDom, and the insulation score (IS). We applied each method to mouse embryonic stem cell (mESC) H3K4me3 PLAC-seq data, using TADs reported in Dixon *et al.* (2012) as the

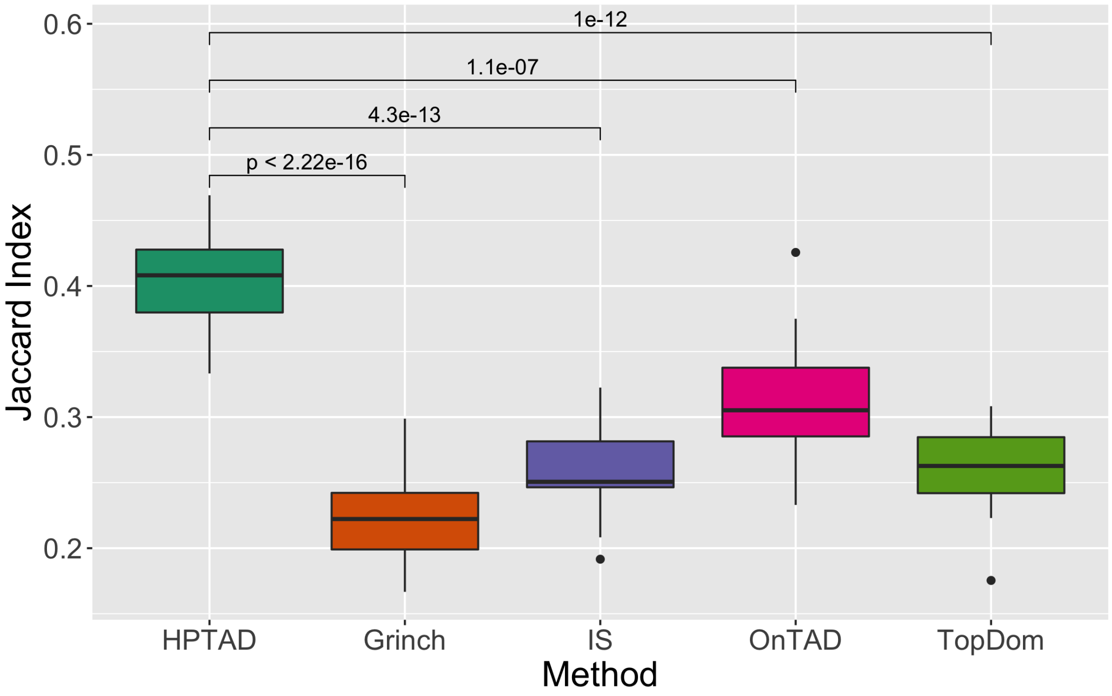

Figure 4.2: mESC H3K4me3 PLAC-seq experiment - Boxplots displaying modified Jaccard index results for HPTAD vs. the indicated methods using raw contact counts as input. The top, middle, and bottom lines of the boxes represent the third, second, and first quartiles respectively and whiskers extend to 1.5 times the interquartile range. Outliers are plotted as points. Displayed results are for all 20 mouse chromosomes. The numbers above pairs of boxes represent the p-value of a paired t-test comparing two methods.

ground truth. For the four TAD callers designed for use with Hi-C data we initially used raw counts for the AND, XOR, and NOT sets (Methods 4.2.1) as input. The results are displayed in Figure 4.2.

Considering the biases introduced by the chromatin immunoprecipitation step, it is not surprising that HPTAD outperforms the other methods since the HPTAD input is normalized to account for these biases. The largest p-value obtained from paired t-tests comparing HPTAD to the other methods was 1×10^{-7}. We repeated the experiment, normalizing the counts for effective fragment length, GC content, mappability, and ChIP enrichment level (Figure 4.3). As expected, the mean Jaccard indices for the Hi-C specific methods increased, with the exception of TopDom, which was unchanged. For Grinch, the mean Jaccard index improved from 0.225 to 0.291, for

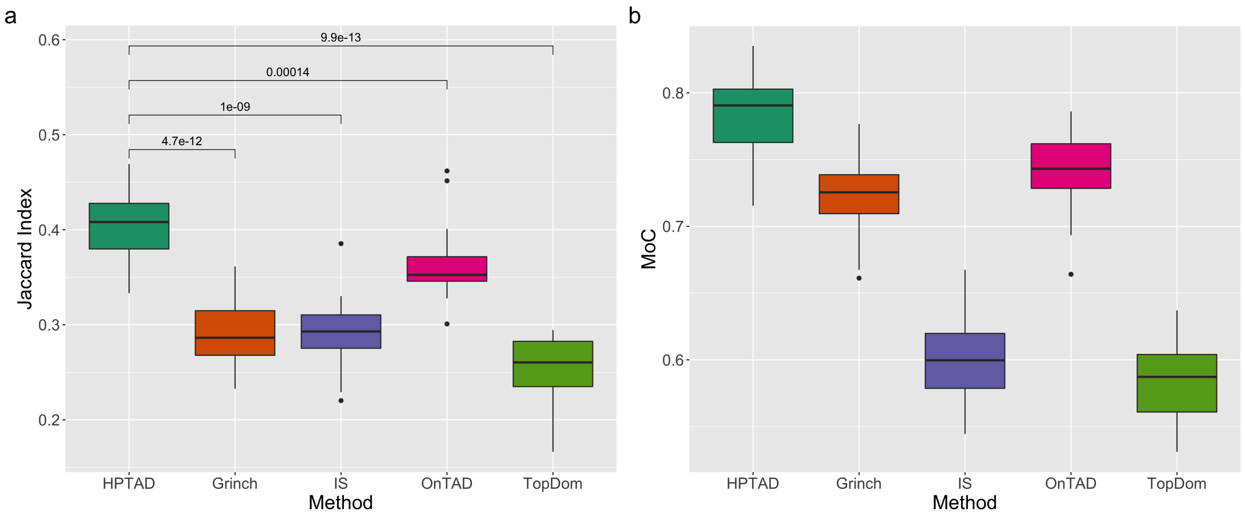

Figure 4.3: mESC H3K4me3 PLAC-seq experiment - Boxplots displaying modified Jaccard index results (a) and Measure of Concordance (b) for HPTAD vs. the indicated methods using normalized counts as input. The top, middle, and bottom lines of the boxes represent the third, second, and first quartiles respectively and whiskers extend to 1.5 times the interquartile range. Outliers are plotted as points. Displayed results are for all 20 mouse chromosomes. The numbers above pairs of boxes represent the p-value of a paired t-test comparing two methods.

IS it improved from 0.257 to 0.294, and for OnTAD it improved from 0.312 to 0.364. The mean Jaccard index for HPTAD was 0.404. All comparisons by paired t-test were again significant at the 0.05 significance level, with the largest p-value now 1×10^{-4}.

With respect to the MoC, HPTAD also outperformed the other four methods. The mean MoC over all 20 mouse chromosomes was 0.784 for HPTAD compared to 0.723, 0.601, 0.741, and 0.584 for Grinch, IS, OnTAD, and TopDom respectively.

We next repeated the previous experiment applying each method to human lymphoblastoid cell line GM12878 H3K27ac HiChIP data, using TADs reported in Schmitt *et al.* (2016) as the ground truth. We observed similar relative performances of the methods with respect to both Jaccard index and MoC, with the notable exception of Grinch's improved performance relative to OnTAD (Figure 4.4). In the mESC experiment Grinch outperformed IS and TopDom and underperformed OnTAD with respect to both Jaccard and MoC. However, in the GM12878 experiment Grinch modestly outperformed OnTAD with respect to both metrics (0.201 to 0.190 Jaccard; 0.670 to 0.650 MoC). Again, HPTAD outperformed the other methods with respect to both met-

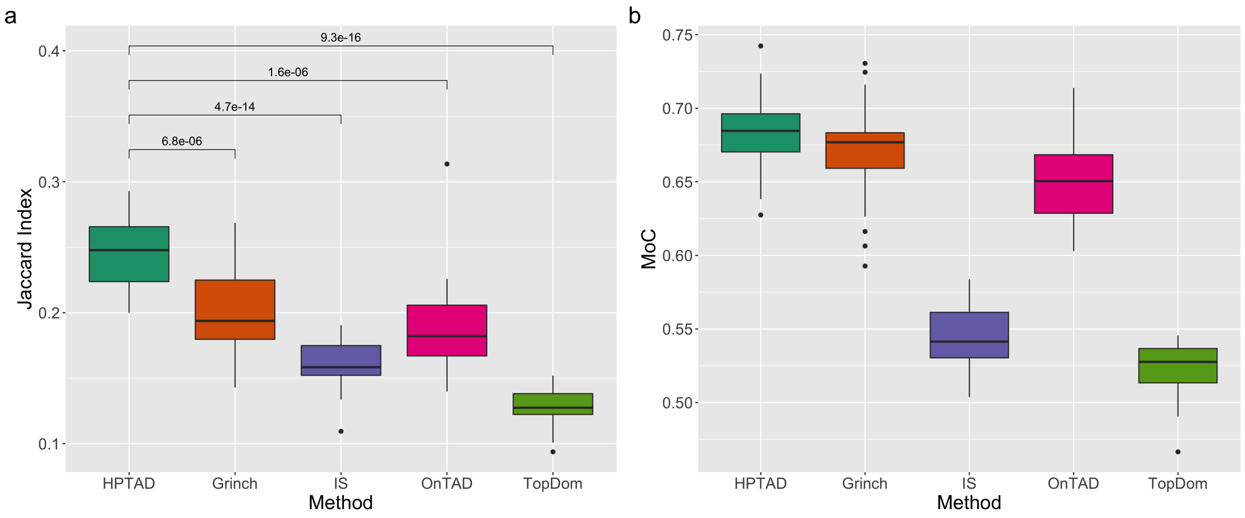

Figure 4.4: GM12878 H3K27ac HiChIP experiment - Boxplots displaying modified Jaccard index results (a) and Measure of Concordance (b) for HPTAD vs. the indicated methods using raw contact counts as input. The top, middle, and bottom lines of the boxes represent the third, second, and first quartiles respectively and whiskers extend to 1.5 times the interquartile range. Outliers are plotted as points. Displayed results are for all 20 mouse chromosomes.

rics. We observed that for all methods both the Jaccard indices and measures of concordance are lower for the GM12878 results relative to the mESC experiment. This is likely attributable to the smaller number of read counts for the GM12878 experiment.

The consistency of relative performance between the two experiments is particularly encouraging. The "ground truth" TADs for both experiments were called using different methods (mESC: directinality index; GM12878: IS). Inconsistent relative performances would have weakened justification for our choice of "ground truth" TADs by suggesting that our observed results were a function of the method used to call the reference TADs. It is interesting that the IS method only outperforms TopDom in the GM12878 experiment given that the "truth" was called using that method. This result highlights the difference between using Hi-C and HP data as input for a TAD caller designed for use with the former technology.

4.3.2 Model consistency

Considering evidence that TAD boundaries are conserved even across different cell lines in the same organism, we hypothesized that a reasonable TAD calling methodology should be able to produce consistent results across biological replicates of the same cell type. We compared results from the five TAD calling methodologies using two biological replicates each of previously described mESC and GM12878 HP samples.

We observed strong agreement between biological replicates for all methods except for Grinch. The mean Jaccard indices for the other four methods ranged from 0.756 (OnTAD) to 0.859 (HPTAD) and the MoC values ranged from 0.909 (IS) to 0.923 (HPTAD) for the mESC replicates. Similarly, for the GM12878 replicates the mean Jaccard indices for the same methods ranged from 0.684 (OnTAD) to 0.817 (IS) and the MoC values ranged from 0.870 (TopDom) to 0.895 (HPTAD). The poor consistency observed with Grinch suggests that some aspect of the matrix factorization used by that method is particularly sensitive to HP data.

4.3.3 CTCF enrichment

Thus far we have compared TAD boundaries and regions to a chosen "ground truth" with the understanding that no true standard actually exists. The transcription factor protein CTCF is reported to be enriched at TAD boundaries ((Dixon *et al.*, 2012)). Since the methods we are comparing vary in the number and identity of TADs they call, we presumed examining the overlap of boundaries with CTCF signals would support the biological signficance of called TADs.

Repeating an analysis from Dixon *et al.* we first examined the number of CTCF peaks as a function of distance from TAD boundaries within a window of ± 500 Kb. As observed in the aforementioned reference, we see a peak at the TAD boundary and a rapid decrease in average peak density with increasing distance from a boundary (Figure 4.6a).

We chose to compare CTCF enrichment between methods using two metrics: the mean number of CTCF peaks per TAD boundary and the fold change enrichment based on number of boundaries that overlap CTCF peaks. The motivation for looking at the second metric was to

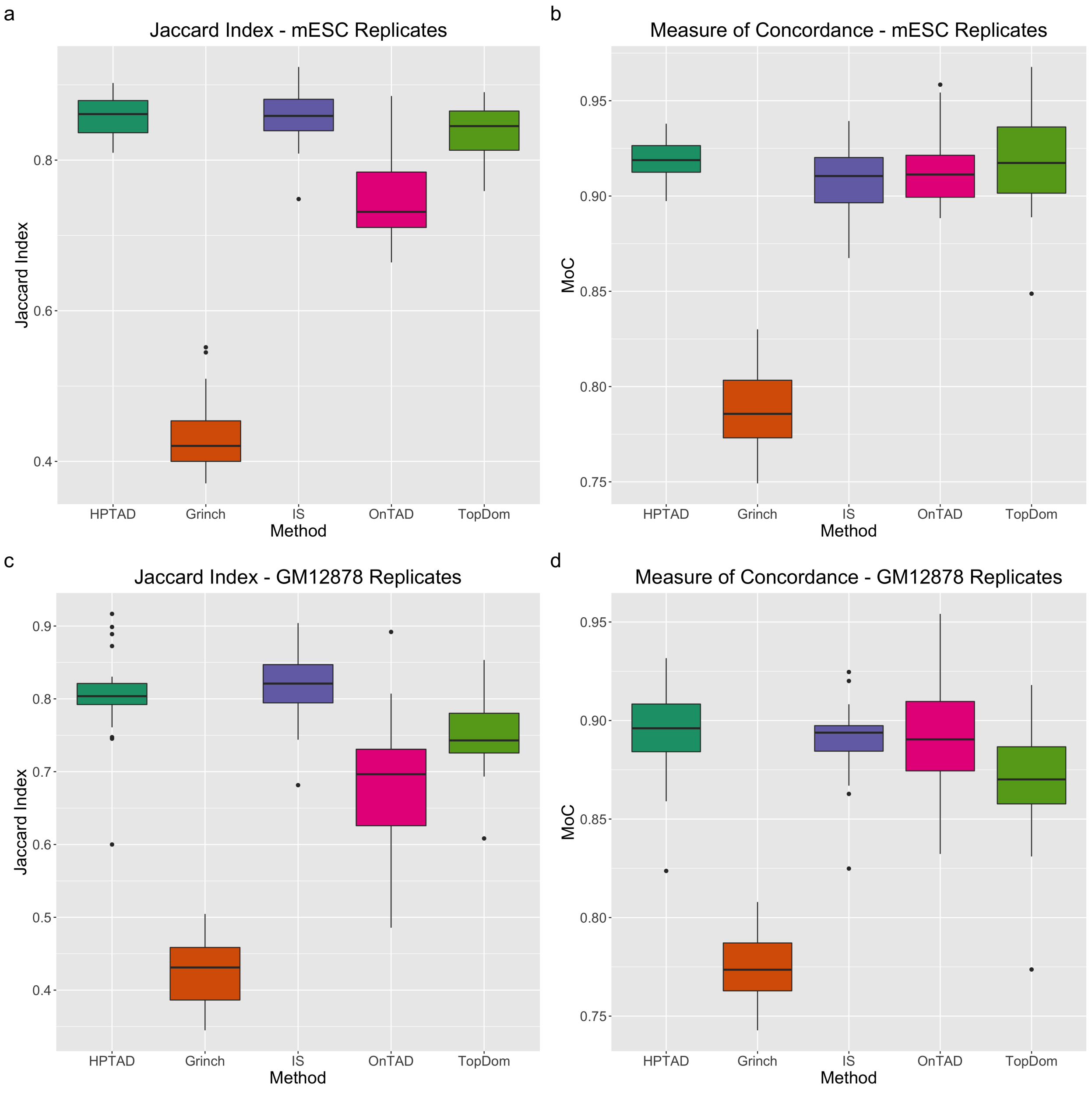

Figure 4.5: Consistency between biological replicates - mESC Jaccard index (a) and measure of concordance (b) and GM12787 Jaccard index (c) and measure of concordance (d). Boxplots displaying results for HPTAD vs. the indicated methods using normalized contact counts as input. The top, middle, and bottom lines of the boxes represent the third, second, and first quartiles respectively and whiskers extend to 1.5 times the interquartile range. Outliers are plotted as points. Displayed results are for all 20 mouse or 23 human chromosomes.

mitigate against the possibility that a method was capturing a few dense CTCF peak regions but many non-overlapping regions. This does not appear to the case however, considering the relative

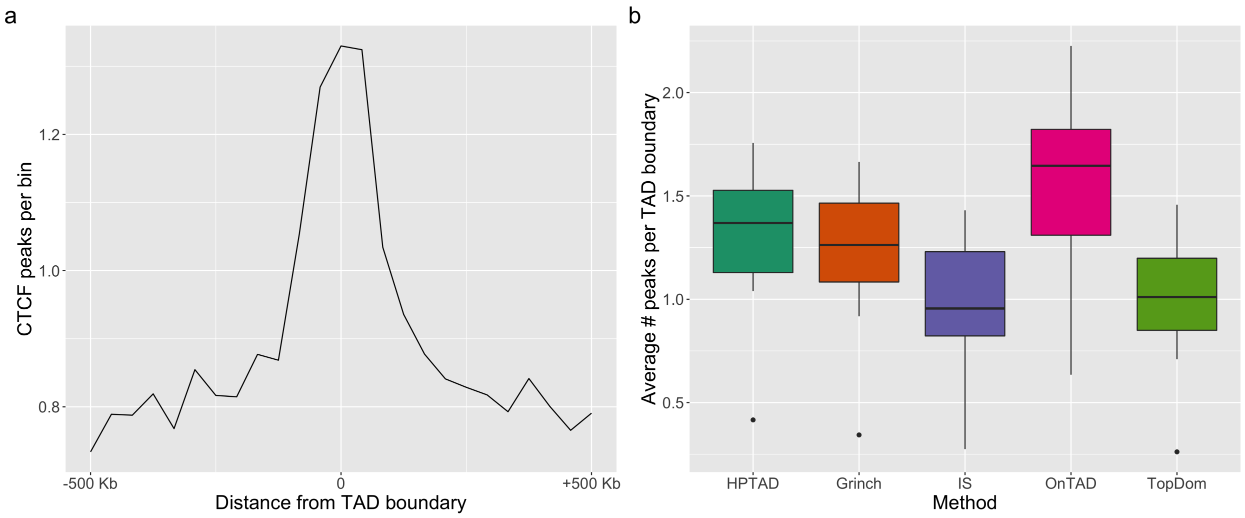

Figure 4.6: CTCF enrichment for mESC experiment - Number of CTCF peaks as a function of distance from HPTAD boundaries (a) and average number of peaks per TAD boundary (b). Boxplots displaying results for HPTAD vs. the indicated methods using normalized contact counts as input. The top, middle, and bottom lines of the boxes represent the third, second, and first quartiles respectively and whiskers extend to 1.5 times the interquartile range. Outliers are plotted as points. Displayed results are for all 20 mouse chromosomes.

performances of the methods are identical under both metrics (Figures 4.6b and C.2). OnTAD exhibits the highest average peak density per TAD boundary and fold enrichment (1.59 and 2.16 respectively), followed by HPTAD (1.32, 1.90), Grinch (1.23, 1.76), TopDom (1.00, 1.56), and IS (1.00, 1.49). Similar relative results were observed repeating the experiment with GM12878 data (Figure C.1).

4.3.4 Number of TADs called

To better understand the differences in results for the various methods we visualized the TAD regions in the mESC experiments by plotting them against the “ground truth” values from Dixon *et al.*, overlaid on a heatmap of the normalized counts (Figure 4.7). Expectedly, TopDOM and IS, the methods that call the greatest number of TADs, show the highest level of partitioning TADs into sub-TAD regions. Based on the lower CTCF enrichment observed in these methods compared to HPTAD, OnTAD, and Grinch, there is not compelling evidence that these sub-TADs

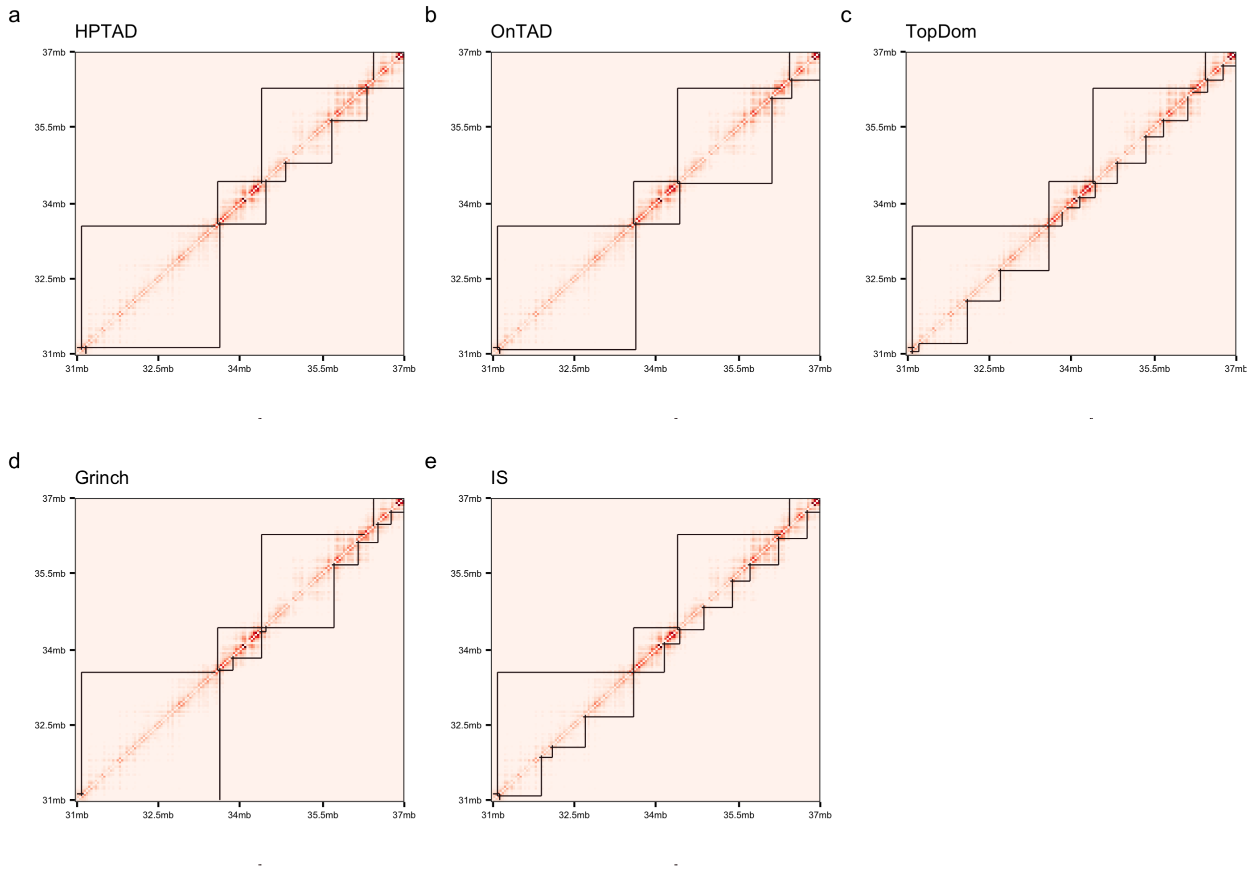

Figure 4.7: Visualization of TAD region 31 - 37 Mbp on chromosome 1 for mESC experiments. TADs from Dixon *et al.* (2012) on top, indicated method on bottom.

are biologically relevant. This implies that those methods could be overly sensitive using HP input data.

The increased numbers of TADs called by TopDom and IS (Figure C.4) are undoubtedly contributing to their poor performance measured by Jaccard index and MoC. For example, if two sets have relative sizes 1:2, their Jaccard index has a maximum possible value of only 0.5. The measure of consistency metric decreases as a function of the inverse square root of set size.

Both the TopDom and IS implementations utilize method-specific tuning parameters. While we conducted our primary analyses using default options (see Section C.4), we modified parameters to intentionally reduce the number of TADs called to closer match the "ground truth" numbers (Figure C.5). We then compared the Jaccard indices and MoCs obtained using the results with fewer called TADs.

TopDom has one user-defined parameter, “window.size”, which defines the number of bins to extend for locus evaluation. Larger window sizes lead to fewer called TADs so we extended this to the maximum recommended value of 20. This reduced the mean number of TADs called per chromosome from 319 to 181, which is still greater than the mean number of “ground truth” TADs (158). For the IS method we modified two parameters independently, window size and minimum score, a threshold used to determine whether the change in insulation score between loci is sufficient to indicate a TAD boundary (default is 0). Both parameters are inversely related to the number of called TADs, that is, increasing their values results in fewer TADs. By increasing the window size we reduced the mean number of TADs called from 246 to 164, and by increasing the minimum score we similarly reduced the number of called TADs to 166. As expected, we observed improvements in both the Jaccard index and MoC in all three cases, however the performance still lagged behind HPTAD (Figure 4.8). For TopDom, the mean Jaccard index increased only modestly from 0.255 to 0.272, while for IS the mean Jaccard index increased more substantially from 0.294 to 0.364 (window size adjustment) and 0.360 (minimum score adjustment). All of these are below 0.404, the mean Jaccard index for HPTAD. The mean MoC for TopDom increased from 0.584 to 0.781, and for IS increased from 0.601 to 0.719 (window size adjustment) and 0.697 (minimum score adjustment), but again, these are all below 0.786, the mean MoC for HPTAD.

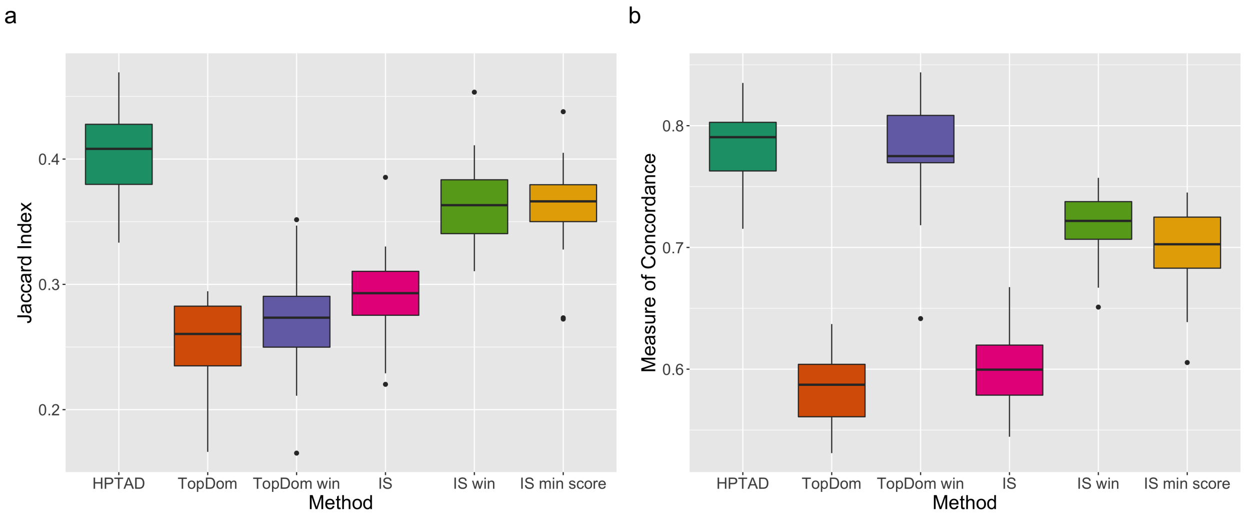

Figure 4.8: Intentional reduction in number of TADs called - Jaccard index (a) and measure of concordance (b) for HPTAD, TopDom, and IS (repeated) plotted with results obtained with adjusted window sizes for TopDom (TopDom win) and IS (IS win) and adjusted minimum score for IS (IS min score). The top, middle, and bottom lines of the boxes represent the third, second, and first quartiles respectively and whiskers extend to 1.5 times the interquartile range. Outliers are plotted as points. Displayed results are for all 20 mouse chromosomes.

4.4 Discussion

While we anticipate that dense read count Hi-C data will remain the gold standard for TAD identification in the foreseeable future, practical considerations such as experiment expense can preclude this as an option for some researchers. Lower-cost methods such as HiChIP and PLAC-seq (HP) provide similar data to Hi-C experiments, however the chromatin immunoprecipitation step introduces an additional bias that must be accounted for. Our aim was to explore the feasibility of using HP data, rather than Hi-C data, to identify TADs.

We present HPTAD, a novel method for TAD identification in HP data. After standard pre-processing we normalize the data as the ratio of expected to observed counts, where the expected counts are derived from a zero-truncated Poisson model. We select candidate TAD boundaries using mean normalized counts within a neighborhood of each locus, and from these candidates we select boundaries based on statistical significance from an assumed model.

We compare the performance of HPTAD to four other publicly available TAD calling methods that were designed for use with Hi-C data and demonstrate improved performance with respect to the Jaccard index applied to TAD boundaries and the measure of concordance applied to TAD regions. In addition, we demonstrate excellent consistency between results from biological replicates, matching or exceeding those of the other methods compared. We even manipulated the number of TADs called by two methods assuming oracular knowledge of the true number of TADs. Even with such matching HPTAD still outperforms these methods.

An obvious shortcoming of our comparisons is that we use published TADs as “ground truth” with the understanding that these are not actually truth. Nonetheless, we use reference TADs called from mouse embryonic stem cells and human lympoblastoic cells using different methods, yet demonstrate consistent relative performance of the TAD callers. This reduces the possibility that HPTAD is merely doing a better job mimicking one particular method. Additional support for the biological relevance of TADs called by our method was provided by CTCF enrichment.

Still, more convincing evidence for the relevance of TADs called from HP data would greatly strengthen an argument for our method’s utility to the scientific community. To that end we plan to apply HPTAD to a complex H3K4me3 PLAC-seq dataset consisting of samples from four brain cell types in human fetal brain obtained via fluorescence-activated cell sorting (Song *et al.*, 2020).

In terms of computational efficiency, we processed the mESC data in 1 hours and 12 minutes using a single core on a 2.50 GHz Intel processor with 9GB of RAM. This represents the time to run all 20 chromosomes; it is possible to run each chromosome in parallel to accelerate processing.

CHAPTER 5: CONCLUSION

While undeniably successful in terms of identifying a tremendous number of variant-trait associations, GWAS results have left open many questions concerning the biological mechanisms underlying these associations. In this dissertation we have proposed several novel methods to analyze genomic data that move beyond one-dimensional, variant-level interrogation of the genome. In a broad sense, these methods represent zooming out from examination of the genome at the allelic level to varying degrees.

First, we pull back from the GWAS variant-level analysis of the genome in favor of the gene-based vantage of transcriptome-wide association studies (TWAS), which examine the association between imputed gene expression and traits of interest. In Chapter 2 we present BayesQN, a Bayesian TWAS method that explicitly models well- and poorly-imputed variants under different distributional assumptions. BayesQN is compared to existing TWAS training methods in both simulated and actual RNA-seq data. Performance gains over several TWAS training methods were observed in specific simulation scenarios, however these gains attenuated in real-world data. Nonetheless, we identified scenarios in which our Bayesian methodology outperformed the commonly used elastic net framework. Future research endeavors include moving from dichotomization of variants based on imputation quality in favor of population-based metrics. For example, local ancestry can be assigned on the variant level using software such as RFMix (Maples *et al.*, 2013), potentially allowing for improved predictive performance in admixed populations.

Next, we pull back even further and examine long-range interactions between genomic loci, which help link GWAS variants to their most likely effector genes in relevant tissues or cell types. Specifically, we concern ourselves with Hi-ChIP and PLAC-seq (HP) data, two evolutions of Hi-C technology that incorporate chromatin immunoprecipitation to select for loci bound by a protein of interest, which enables high-resolution analyses at reduced cost and using less input material compared to Hi-C. In Chapter 3 we present HPRep, a method that uses zero-truncated Poisson regression to normalize contact frequencies adjusting for HP-specific biases and calculates a stratified and weighted correlation metric to quantify the reproducibility in HP datasets. While several methods for quantifying reproducibility in Hi-C datasets exist, none of these account for biases introduced in the chromatin immunoprecipitation step of HP methods. We compare HPRep to some of these methods developed for Hi-C data and demonstrate improved performance in both mouse and human data. Furthermore, we demonstrate the ability to differentiate cell types using HPRep applied to a complex set of 11 samples of human brain cells representing four cell types. Future research will involve applying our method to similar chromatin conformation capture technologies, such as promoter capture Hi-C.

Finally, we shift vantage point once again and examine topologically association domains (TADs), large contiguous regions of the genome characterized by a higher frequency of within-region interactions relative to between-region interactions. Continuing our analyses of HP data introduced in Chapter 3, in Chapter 4 we explore the possibility of using such data to identify TADs, thereby expanding the utility of HP data. We introduce HPTAD, which uses a regression-based framework to normalize contact frequencies and select a set of candidate TAD boundaries from which final boundaries are chosen using a statistical test. We compare the performance of HPTAD to several publicly available tools for identifying TADs from Hi-C data and demonstrate improved performance compared to "ground truth" data in both mouse and human cell lines. Additionally, we demonstrate good consistency between results obtained from biological replicates and CTCF enrichment in TAD boundaries called with HPTAD. Future research plans are to apply this method to the complex PLAC-seq human brain cell dataset described in Chapter 3 with the aim of identifying TADs whose biological relevance can be supported with other omics data (e.g. gene expression) in matched tissue(s) and/or cell type(s).

Over the last twenty years GWAS has continued to be utilized for the identification of variants associated with a diverse array of phenotypes, however the need for new tools to move beyond mere correlation in favor of a deeper understanding of causation still exists. A recent initiative was just funded by the NHGRI to address this need: the Impact of Genomic Variation on Function (IGVF) Consortium's stated goal is to "develop a framework for systematically understanding the effects of genomic variation on genome function and how these effects shape phenotypes." While far less expansive in scope than anticipated IGVF-driven advances, we present three projects that represent additions to the continually expanding toolkit of methods aimed at the same goal.

APPENDIX A: ADDITIONAL RESULTS FOR CHAPTER 2

A.1 Terminology Definitions

Let us define some terminology to be used throughout this section:

- $\boldsymbol{Y}$ is a length p vector of responses

- $\boldsymbol{X}$ is an $n \times p$ matrix of genotypes for n subjects. The p SNPs can be represented by dosages or $\{0, 1, 2\}$, and the matrix is centered and scaled to unit variance. $\boldsymbol{X}_{i\cdot}$ and $\boldsymbol{X}_{\cdot j}$ represent the i^{th} row and j^{th} column of $\boldsymbol{X}$ respectively. $\boldsymbol{X}_{(-j)}$ represents matrix $\boldsymbol{X}$ with the j^{th} column removed.

- For a BayesQN model, we assign SNPs to category $q \in \{1, 2, \ldots, N\}$; we focus on the case where $N = 2$. The categories are mutually exclusive; therefore if m_q is the number of SNPs in category q then $\sum_{q=1}^{N} m_q = p$.

- $\boldsymbol{\beta}$ is a length p vector of SNP effect sizes. $\boldsymbol{\beta}_{(-j)}$ is a length $p-1$ vector representing $\boldsymbol{\beta}$ with the j^{th} element (β_j) removed. $\boldsymbol{\beta}_q$ is a length m_q vector of category q SNP effect sizes and β_{qj} is the j^{th} element of $\boldsymbol{\beta}_q$. The value q_j simply denotes the category of SNP j.

- $\boldsymbol{J}_n$ is a length n vector of ones.

- $\boldsymbol{\pi}_q$ is a length 4 vector whose elements are probabilities that β_{qj} follows distribution $k \in \{1, 2, 3, 4\}$, and sum to 1. π_{q1} is the first element of the vector, π_{q2} the second, etc.

- σ^2_{qk} is a scalar representing 0, $0.001\sigma^2_q$, $0.01\sigma^2_q$, or $0.1\sigma^2_q$ for $k = \{1, 2, 3, 4\}$ respectively.

Following the framework established in the BayesR paradigm, the response variable $\boldsymbol{Y}$ is modeled as

$$\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{\beta}, \mu, \sigma^2_e \sim N(\mu J_n + \boldsymbol{X}\boldsymbol{\beta}, \boldsymbol{I}_n\sigma^2_e)$$

A.2 Prior Distributions

We set the following priors:

- Prior on μ is uninformative

$$p(\mu)\propto 1$$

- Prior on β_j

$$\beta_j|\boldsymbol{\pi_q}, \sigma_q^2, q_j \sim \pi_{q_1}N(0,0) + \pi_{q_2}N(0, 0.001\sigma_q^2) + \pi_{q_3}N(0, 0.01\sigma_q^2) + \pi_{q_4}N(0, 0.1\sigma_q^2)$$

- Prior on σ_e^2

$$p(\sigma_e^2) \sim \text{Scaled inverse } \chi^2(\nu_o, S_o^2)$$

- Prior on σ_q^2

$$p(\sigma_q^2) \sim \text{Scaled inverse } \chi^2(\nu_o, S_o^2)$$

- Prior on $\boldsymbol{\pi_q}$

$$p(\boldsymbol{\pi_q}) \sim \text{Dirichlet}(1,1,1,1)$$

A.3 Derivation of Posterior Distributions

So for the conditional posterior distribution of μ

$$p(\mu|\cdot)\propto p(\boldsymbol{Y}|\cdot)p(\mu)\propto(2\pi)^{-n/2}\left(\frac{1}{\sigma_e^2}\right)^n\exp\{-1/2\sigma_e^2(\boldsymbol{Y}-\mu\boldsymbol{J}_n-\boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{Y}-\mu\boldsymbol{J}_n-\boldsymbol{X}\boldsymbol{\beta})\}$$

Therefore

$$\mu|\cdot\sim N\left(1/n\sum_{i=1}^{n}(Y_i-\boldsymbol{X_{i\cdot}}\boldsymbol{\beta}),\sigma_e^2/n\right)$$

where $\boldsymbol{X_{i\cdot}}$ is the length p row vector for the i^{th} subject.

Next we need to calculate the conditional probability that SNP$_j$ comes from distribution $k\in\{1,\ldots,4\}$ given the data

$$\text{Likelihood}\propto p(\boldsymbol{Y}|\cdot,\text{SNP}_j\text{ from dist }k)p(\text{SNP}_j\text{ from dist }k)$$

where the first term is

$$\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{\beta}_{(-\boldsymbol{j})},\mu,\sigma_e^2,\sigma_q^2,\boldsymbol{\pi}_q,q_j\sim N(\mu\boldsymbol{J}_n+\boldsymbol{X}_{(-j)}\boldsymbol{\beta}_{(-\boldsymbol{j})},\boldsymbol{I}_n\sigma_e^2+\boldsymbol{X}_{\cdot j}\boldsymbol{X}_{\cdot j}^T\sigma_{qk}^2)$$

Therefore the log liklihood is

$$\begin{aligned}&\log(\boldsymbol{\pi}_{qk})+\log(C)-1/2\log(\det(\boldsymbol{I}_n\sigma_e^2+\boldsymbol{X}_{\cdot j}\boldsymbol{X}_{\cdot j}^T\sigma_{qk}^2))\\&\qquad-1/2(\boldsymbol{Y}-\mu\boldsymbol{J}_n-\boldsymbol{X}_{(-j)}\boldsymbol{\beta}_{(-j)})^T(\boldsymbol{I}_n\sigma_e^2+\boldsymbol{X}_{\cdot j}\boldsymbol{X}_{\cdot j}^T\sigma_{qk}^2)^{-1}(\boldsymbol{Y}-\mu\boldsymbol{J}_n-\boldsymbol{X}_{(-j)}\boldsymbol{\beta}_{(-j)})\end{aligned}$$

Next we make use of the Woodbury identity:

$$(A+XBX^T)^{-1}=A^{-1}-A^{-1}X(B^{-1}+X^TA^{-1}X)^{-1}X^TA^{-1}$$

where $A = \mathbf{I_n \sigma_e^2}$ and $B = \sigma_{qk}^2$. Therefore the inverted matrix becomes

$$\mathbf{I_n}\left(\frac{\mathbf{1}}{\mathbf{\sigma_e^2}}\right) - \mathbf{I_n}\left(\frac{\mathbf{1}}{\mathbf{\sigma_e^2}}\right)\mathbf{X_{\cdot j}}\left(\frac{\mathbf{1}}{\mathbf{\sigma_{qk}^2}} + \mathbf{X_{\cdot j}^T X_{\cdot j}}\frac{\mathbf{1}}{\mathbf{\sigma_e^2}}\right)^{-\mathbf{1}}\mathbf{X_{\cdot j}^T}\left(\frac{\mathbf{1}}{\mathbf{\sigma_e^2}}\right)$$

Now let $\tilde{\boldsymbol{Y}} = \boldsymbol{Y} - \mu\boldsymbol{J}_n - \boldsymbol{X}_{(-j)}\boldsymbol{\beta}_{(-j)}$ and let $\boldsymbol{W} = \boldsymbol{I}_n\sigma_e^2 + \boldsymbol{X}_{\cdot j}\boldsymbol{X}_{\cdot j}^T\sigma_{qk}^2$. Substituting into above we get:

$$\log(\pi_{qk}) + \log(C) - 1/2\log(\det(\boldsymbol{W})) - \frac{1}{2\sigma_e^2}\tilde{\boldsymbol{Y}}^T\tilde{\boldsymbol{Y}} + \frac{1}{2\sigma_e^2}\tilde{\boldsymbol{Y}}^T\boldsymbol{X}_{\cdot j}\left(\frac{\sigma_e^2 + \sigma_{qk}^2 X_{\cdot j}^T\boldsymbol{X}_{\cdot j}}{\sigma_e^2\sigma_{qk}^2}\right)^{-1}\frac{1}{2\sigma_e^2}\tilde{\boldsymbol{Y}}^T\boldsymbol{X}_{\cdot j}$$

$$= \log(\pi_{qk}) + \log(C) - 1/2\log(\det(\boldsymbol{W})) - \frac{1}{2\sigma_e^2}\tilde{\boldsymbol{Y}}^T\tilde{\boldsymbol{Y}} + \frac{1}{2\sigma_e^2}(\tilde{\boldsymbol{Y}}^T X_{\cdot j})^2\frac{\sigma_{qk}^2}{\sigma_e^2 + \sigma_{qk}^2\boldsymbol{X}_{\cdot j}^T\boldsymbol{X}_{\cdot j}}$$

Making use of the identity $\det(A + uv^T) = (1 + v^T A^{-1}u)\det(A)$ we let $A = \mathbf{I_n\sigma_e^2}$ and let $u = v = \boldsymbol{X}_{\cdot j}\sigma_{qk}$ and obtain

$$\det(\boldsymbol{W}) = (1 + \sigma_{qk}X_{\cdot j}^T\frac{1}{\sigma_e^2}\boldsymbol{I}_n\boldsymbol{X}_{\cdot j}\sigma_{qk}) = \left(1 + \frac{\sigma_{qk}^2}{\sigma_e^2}\boldsymbol{X}_{\cdot j}^T\boldsymbol{X}_{\cdot j}\right)\sigma_e^{2n}$$

Therefore

$$\begin{aligned}\text{Loglik} \propto \log(\pi_{qk}) + \log(C) - 1/2\log\left(1 + \frac{\sigma_{qk}^2}{\sigma_e^2}\boldsymbol{X}_{\cdot j}^T\boldsymbol{X}_{\cdot j}\right)\\ - n/2\log(\sigma_e^2) - \frac{1}{2\sigma_e^2}\left[\tilde{\boldsymbol{Y}}^T\tilde{\boldsymbol{Y}} - \tilde{\boldsymbol{Y}}^T\boldsymbol{X}_{\cdot j}\left(\frac{\tilde{\boldsymbol{Y}}^T\boldsymbol{X}_{\cdot j}}{\sigma_e^2/\sigma_{qk}^2 + \boldsymbol{X}_{\cdot j}^T\boldsymbol{X}_{\cdot j}}\right)\right]\end{aligned}$$

For the posterior distribution of $\beta_j|\cdot$, j from dist k we have:

$$\beta_j|\cdot, \text{j from dist k} \propto p(\boldsymbol{Y}|\cdot)p(\beta_j|\text{j from dist k})p(\text{j from dist k})$$

$$
\begin{aligned}
&\propto \exp\left(-\frac{1}{2\sigma_e^2}(\boldsymbol{Y}-\mu\boldsymbol{J}_n-\boldsymbol{X}_{(-j)}\boldsymbol{\beta}_{(-j)}-\boldsymbol{X}_{\cdot j}\boldsymbol{\beta}_j)^T(\boldsymbol{Y}-\mu\boldsymbol{J}_n-\boldsymbol{X}_{(-j)}\boldsymbol{\beta}_{(-j)}-\boldsymbol{X}_{\cdot j}\boldsymbol{\beta}_j)\right)\\
&\quad \exp\left(\frac{1}{2\sigma_{qk}^2}\beta_j^2\right)\\
&= \exp\left(-\frac{1}{2}\left[\frac{1}{\sigma_e^2}(\tilde{\boldsymbol{Y}}-\boldsymbol{X}_{\cdot j}\beta_j)^T(\tilde{\boldsymbol{Y}}-\boldsymbol{X}_{\cdot j}\beta_j)+\frac{1}{\sigma_{qk}^2}\beta_j^2\right]\right)\\
&\propto \exp\left(-\frac{1}{2}\left[\frac{1}{\sigma_e^2}(\beta_j\boldsymbol{X}_{\cdot j}^T\boldsymbol{X}_{\cdot j}\beta_j-2\beta_j\boldsymbol{X}_{\cdot j}^T\tilde{\boldsymbol{Y}})+\frac{1}{\sigma_{qk}^2}\beta_j^2\right]\right)\\
&\propto \exp\left(-\frac{1}{2}\left(\frac{1}{\sigma_{qk}^2}+\frac{\boldsymbol{X}_{\cdot j}^T\boldsymbol{X}_{\cdot j}}{\sigma_e^2}\right)\left(\beta_j-\left(\frac{1}{\sigma_{qk}^2}+\frac{\boldsymbol{X}_{\cdot j}^T\boldsymbol{X}_{\cdot j}}{\sigma_e^2}\right)^{-1}\boldsymbol{X}_{\cdot j}^T\tilde{\boldsymbol{Y}}\frac{1}{\sigma_e^2}\right)^2\right)
\end{aligned}
$$

Therefore the posterior distribution of $\beta_j|\cdot$, j from dist k is

$$
N\left(\frac{\boldsymbol{X}_{\cdot j}^T\tilde{\boldsymbol{Y}}\sigma_{qk}^2}{\sigma_e^2+\sigma_{qk}^2\boldsymbol{X}_{\cdot j}^T\boldsymbol{X}_{\cdot j}},\frac{\sigma_e^2\sigma_{qk}^2}{\sigma_e^2+\sigma_{qk}^2\boldsymbol{X}_{\cdot j}^T\boldsymbol{X}_{\cdot j}}\right)
$$

The posterior distribution of σ_e^2 is:

$$
\begin{aligned}
p(\sigma_e^2|\cdot)&\propto p(\boldsymbol{Y}|\cdot)p(\sigma_e^2)\\
&= (2\pi)^{-n/2}\left(\frac{1}{\sigma_e^2}\right)^n\exp\left(-\frac{1}{2\sigma_e^2}(\boldsymbol{Y}-\mu\boldsymbol{J}_n-\boldsymbol{X\beta})^T(\boldsymbol{Y}-\mu\boldsymbol{J}_n-\boldsymbol{X\beta})\right)\\
&\quad \frac{(S_o^2\nu_o/2)^{\nu_o/2}}{\Gamma(\nu_o/2)}\exp\left(-\frac{1}{2\sigma_e^2}\nu_oS_o^2\right)\left(\sigma_e^2\right)^{-1-\nu_o/2}\\
&= \left(\sigma_e^2\right)^{-1-(\nu_o+n)/2}\exp\left(-\frac{1}{2\sigma_e^2}[\nu_oS_o^2+(\boldsymbol{Y}-\mu\boldsymbol{J}_n-\boldsymbol{X\beta})^T(\boldsymbol{Y}-\mu\boldsymbol{J}_n-\boldsymbol{X\beta})]\right)
\end{aligned}
$$

So

$$
\sigma_e^2|\cdot \sim \text{Scaled inverse }\chi^2\left(\nu_o+n,\frac{\nu_oS_o^2+(\boldsymbol{Y}-\mu\boldsymbol{J}_n-\boldsymbol{X\beta})^T(\boldsymbol{Y}-\mu\boldsymbol{J}_n-\boldsymbol{X\beta})}{\nu_o+n}\right)
$$

The posterior distribution of σ_q^2 is a function of the β_j for which $q_j = q$. Specifically,

$$p(\sigma_q^2|\cdot) \propto p(\sigma_q^2) \prod_{j=1}^{p} p(\beta_j|\sigma_q^2)^{I(q_j=q)}$$

For clarity, let us introduce a new variable Z_j which takes on the value of $k \in \{1, 2, 3, 4\}$ if SNP j follows distribution k. For $k = 1$ the corresponding βs are zero, hence they do not factor into the derivation.

$$\begin{aligned}
p(\sigma_q^2|\cdot) &\propto p(\sigma_q^2) \prod_{k=2}^{4} \left\{ \prod_{j \in \{j:Z_j=k\}} \left(\frac{1}{\sigma_{qk}\sqrt{2\pi}} \exp\left\{ -\frac{\beta_j^2}{2\sigma_{qk}^2} \right\} \right)^{I(q_j=q)} \right\} \\
&\propto \exp\left(-\frac{\nu_o S_o^2}{2\sigma_q^2} \right) \left(\sigma_q^2\right)^{-1-\nu_o/2} \prod_{k=2}^{4} {\sigma_q}^{-m_{qk}} \exp\left\{ -\frac{\sum_{j \in \{j:Z_j=k\}} \beta_j^2 I(q_j = q)}{2\sigma_{qk}^2} \right\} \\
&\propto \sigma_q^{-2-\nu_o-m_{q2}-m_{q3}-m_{q4}} \exp\left\{ \sum_{k=2}^{4} -\frac{\sum_{j \in \{j:Z_j=k\}} \beta_j^2 I(q_j = q)}{2c_k\sigma_q^2} - \frac{\nu_o S_o^2}{2\sigma_q^2} \right\} \\
&\propto \sigma_q^{-2-\nu_o-m_{q2}-m_{q3}-m_{q4}} \exp\left\{ -\frac{1}{2\sigma_q^2} \left(\nu_o S_o^2 + \sum_{j=1}^{p} \frac{\beta_j I(q_j = q) I(Z_j = 2)}{c_2} \right.\right. \\
&\qquad \left.\left. + \sum_{j=1}^{p} \frac{\beta_j I(q_j = q) I(Z_j = 3)}{c_3} + \sum_{j=1}^{p} \frac{\beta_j I(q_j = q) I(Z_j = 4)}{c_4} \right) \right\} \\
&\propto \sigma_q^{-2-\nu_o-m_{q2}-m_{q3}-m_{q4}} \exp\left\{ -\frac{1}{2\sigma_q^2} \left(\nu_o S_o^2 + \sum_{k=2}^{4} \sum_{j=1}^{p} \frac{\beta_j I(q_j = q) I(Z_j = k)}{c_k} \right) \right\}
\end{aligned}$$

Therefore, the posterior distribution is

$$\sigma_g^2|\cdot \sim \text{Scaled inverse } \chi^2 \left(\nu_o + m_{q2} + m_{q3} + m_{q4}, \frac{\nu_o S_o^2 + \sum_{k=2}^{4} \sum_{j=1}^{p} \frac{\beta_j I(q_j=q) I(Z_j=k)}{c_k}}{\nu_o + m_{q2} + m_{q3} + m_{q4}} \right)$$

Finally, it is a known result that the Dirichlet distribution is conjugate in this setting, so the posterior distribution of π_q is given by

$$p(\pi_q|\cdot) \propto \prod_{k=1}^{4} \pi_k^{1-1} {\pi_k}^{m_{qk}}$$

So the posterior distribution is

$$p(\pi_q|\cdot) \sim \text{Dirichlet}\left(m_{q1}+1, m_{q2}+1, m_{q3}+1, m_{q4}+1\right)$$

APPENDIX B: ADDITIONAL RESULTS FOR CHAPTER 3

B.1 Details for Step 1 and 2 of HPRep

During the pre-processing step, intra-chromosomal reads are split into two groups: short-range reads ($\leqslant$ 1 Kb) and long-range reads ($>$ 1Kb). The short-range reads are used as a measure of ChIP efficiency in the regression framework described later in the pipeline. Long-range reads are used to determine long-range interactions, which are extracted and classified as either AND, XOR, or NOT sets based on whether 2, 1, or 0 (respectively) read ends overlap with a ChIP-seq identified peak for the protein of interest. Additional details can be found in the MAPS paper: (Juric *et al.*, 2019).

The regression and normalization step (step 2) follows a multi-step procedure:

1. We model the non-zero intra-chromosomal contacts as a zero-truncated Poisson model with mean μ_{ij}. The covariates for effective fragment length (FL), GC content (GC), mappability (MS), and ChIP enrichment level (IP) are provided by the feather pre-processing step (as implemented in the MAPS pipeline), and represent log($x_i \times x_j$), where x_i and x_j are the corresponding covariates for bins i and j respectively. We fit regression models for the AND and XOR sets separately.

$$\log(\mu_{ij}) = \beta_0 + \beta_1 \cdot FL_{ij} + \beta_2 \cdot GC_{ij} + \beta_3 \cdot MS_{ij} + \beta_4 \cdot IP_{ij} \tag{B.1}$$

2. Fitted values are determined for each bin pair based on the resulting model for AND and XOR sets in each chromosome, resulting in $2 \times n$ files where n is the number of autosomal chromosomes.

3. Normalized values are defined as $\log_2$(1 + observed / fitted) and all bin pairs are combined into one file. Additionally, the ChIP-seq peaks are binned to analysis resolution and sup-

plied as a file containing a list of these anchor bins. Peaks that span a bin boundary are assigned to all bins they span.

B.2 Details for Step 3 of HPRep

The final step involves data smoothing and sample comparison to calculate a final reproducibility metric between each pair of samples as a weighted Pearson correlation. The combined AND and XOR normalized data is stored in a matrix which is used as an input for the comparison algorithm. The basic data structure we consider is an $N \times m$ matrix, where N represents the number of anchor bins in the union set of anchors from all samples and m is 2 x binning distance/resolution, where binning distance is recommended to be set at 1 Mb but can be user specified. Interactions further than 1 Mb are typically sparse and highly variable. The ij element of the matrix represents the normalized contact frequency between the anchor i and the bin j bin widths away, $j \in \{-m/2, \ldots, -1, 1, \ldots, m/2\}$. In Figure B.1, $N = 4, m = 400$ at 5 Kb resolution, 200 at 10 Kb resolution.

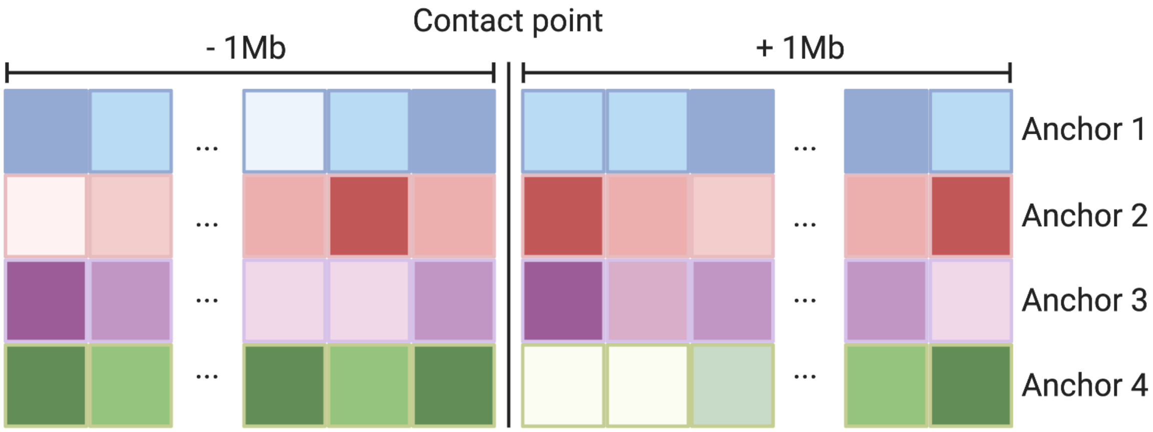

Figure B.1: Each cell represents the smoothed normalized contact frequency between the anchor bin corresponding to the row and the bin x bin units away, where x is the number of cells to the right or left of the midpoint. Cells to the left and right of the midpoint represent bins upstream and downstream of the anchor bin respectively.

The normalized values undergo a 1-D smoothing procedure as follows: for a specified window size d, the ij element (x_{ij}) is transformed such that the smoothed value is

$$x_{ij}^{smoothed} = \frac{\sum_{k=j-d}^{j+d} x_{ik}}{2d+1} \tag{B.2}$$

Let a_k and b_k be two vectors of length $2N$ from samples a and b respectively, whose elements consist of the values from the smoothed data matrix from columns $\pm k$ units symmetrically from the center. All these values represent normalized and smoothed contacts that are $\pm k$ bins from their respective anchors. Let a'_k and b'_k be the resulting vectors of length $N_k \leqslant 2N$ after removing any elements satisfying $a_i = b_i = 0$, where a_{ki} is the i^{th} element of vector a_k. We define r_k as

$$r_k = \frac{N_k \sum_{i=1}^{N_k} a'_i b'_i - \sum_{i=1}^{N_k} a'_i \sum_{i=1}^{N_k} b'_i}{\sqrt{N_k \sum_{i=1}^{N_k} a'^2_i - (\sum_{i=1}^{N_k} a'_i)^2}\sqrt{N_k \sum_{i=1}^{N_k} b'^2_i - (\sum_{i=1}^{N_k} b'_i)^2}} \tag{B.3}$$

namely the empirical correlation between a'_k and b'_k. We then define the weights for each of the k strata as

$$w_k = \frac{N_k \sqrt{\frac{\sum_{i=1}^{N_k} a_i'^2}{N_k} - \left(\frac{\sum_{i=1}^{N_k} a_i'^2}{N_k}\right)^2} \sqrt{\frac{\sum_{i=1}^{N_k} b_i'^2}{N_k} - \left(\frac{\sum_{i=1}^{N_k} b_i'^2}{N_k}\right)^2}}{\sum_{k=1}^{K} N_k \left(\sqrt{\frac{\sum_{i=1}^{N_k} a_i'^2}{N_k} - \left(\frac{\sum_{i=1}^{N_k} a_i'^2}{N_k}\right)^2} \sqrt{\frac{\sum_{i=1}^{N_k} b_i'^2}{N_k} - \left(\frac{\sum_{i=1}^{N_k} b_i'^2}{N_k}\right)^2}\right)} \tag{B.4}$$

The reproducibility score between two matrices is then the weighted average of the stratified correlations r_k

$$\text{reproducibility score} = \sum_{k=1}^{K} r_k w_k \tag{B.5}$$

B.3 Smoothing Parameter Optimization

The smoothing parameter d is tuned using the method similar to the HiCRep protocol with modification to the sampling scheme and search termination criterion. The following algorithm is used:

Two samples to be analyzed are selected, preferably ones that are dissimilar such as non-biological replicates. Twenty-five percent of the non-zero contacts from one are randomly sampled and used to populate a contact matrix as previously diagrammed, with the remaining entries set to zero. The analogous positions in the other sample are used to populate a corresponding matrix. The reproducibility score is calculated for these matrices and the sampling procedure is repeated a total of ten times with no smoothing performed. The average of these ten values is recorded.

The smoothing parameter is then iterated, repeating the above procedure until the average metric using smoothing parameter $d + 1$ compared to d exhibits less than a one percent increase. The value of d is recorded and used as the smoothing parameter for all analyses with the particular dataset.

B.4 Procedures for Comparative Methods

- **HiCRep**
  - All results obtained using HiCRep were conducted using R (3.6.0) and using version 1.12.0 of the HiCRep package obtained from `https://github.com/MonkeyLB/hicrep`. Default parameters were used for all experiments. Note that the documentation recommends a smoothing parameter of 20 for 10 Kb resolution but does not specify a recommended parameter for 5 Kb resolution. We used 20 for 5 Kb as well since marginal difference was reported when tuning beyond 20.
- **HiC-Spector**
  - The Python version of HiC-Spector was used rather than the Julia version since the former readily accepts Hi-C data in genomic coordinates rather than .hic format. The program used was “run_reproducibility_v2.py found at `https://github.com/gersteinlab/HiC-spector`. Experiments that included solely AND and XOR sets of contacts were prepared by extracting bin pairs and observed (integer) contacts

from the corresponding AND / XOR files. Note, the bin positions had to be converted to indices starting at 1, so the global minimum bin position was determined, and all bins positions scaled by (genomic position minimum position) / resolution. Experiments also including NOT sets were generated similarly.

- **Pearson correlation**
  - The upper triangular component of a standard symmetric $n \times n$ contact matrix was flattened to a vector for each sample. The Pearson correlation between two samples was computed as the correlation between these vectors.

B.5 Down-sampling Procedure

The generalized down-sampling procedure was performed on the AND and XOR contact files for each chromosome separately. Let n be the total number of counts for all bin pairs in the specific file and let d be the down-sampling coefficient. That is, to down-sample to $0.8x$ depth, $d = 0.8$. The vector v of counts for all bin pairs is down-sampled to depth d utilizing the R function rbinom where the size parameter is set to floor($n \times d$) and the probability vector is the element-wise division of v by n. These down-sampled AND and XOR files then intersect the pipeline as usual with the removal of bins that now have counts of 0.

B.6 Determination of Silhouette Values

Silhouette values were calculated via the method in Rousseeuw (1987). Let $d(i, j)$ be the similarity between sample i and j, which in this analysis is the scaled reproducibility metric between the two samples. The silhouette method requires that the similarity (or distance) quantities be comparable on a ratio scale, that is, if the distance between two points is doubled that implies the points are twice as far apart. Pearson correlation does not have such a property, so for each experiment the values were standardized to [0, 1] by subtracting the lowest value and dividing by (max - min) value.

Let sample i be a member of cluster A. Furthermore, let $a(i)$ be the average similarity of i to all other samples in the same cluster. Let $d(i, C)$ be the average similarity of sample i to all other samples in cluster C and let $b(i)$ be the maximum value of $d(i, C)$ over all clusters C distinct from cluster A. Then the silhouette value is defined as

$$s(i) = \frac{a(i) - b(i)}{\max\{a(1), b(i)\}} \tag{B.6}$$

We report the average $s(i)$ over all 11 samples. The closer this value is to 1 the better the clustering performance.

B.7 Data Details

For the human brain PLAC-seq data, fastp (`https://github.com/OpenGene/fastp`) was used to trim the fastq files to 100 bp. No additional modifications to the described pipeline were performed on any of the datasets used in this paper. Default software options described in `https://github.com/yunliUNC/HPRep` were used for alignment and merging for all samples analyzed. Resolutions used for each dataset were: 10 Kb for mouse embryonic stem cell and mouse brain tissue H3K4me3 PLAC-seq; GM12878 and K562 H3K27ac HiChIP; and 5 Kb for human brain H3K4me3 PLAC-seq.

B.8 Irreproducible Discovery Rate

ChIP-Seq data processing followed the procedure outlined in Juric *et al.* (2019). Specifically, MACS2 (v 2.1.2) was used to provide the narrowPeak input files using flags: –nolambda, –nomodel, –extsize 147, –call-summits, -B, –SPMR, and -q 1e-2. These files were processed using IDR (v 2.0.4.2) with default parameters. Results reported represent the fraction of peaks that exceed a false discovery rate of 5%. Downsampling was performed on the MACS2 input files by randomly selecting an appropriately sized subset of reads.

B.9 Additional Results by Chromosome

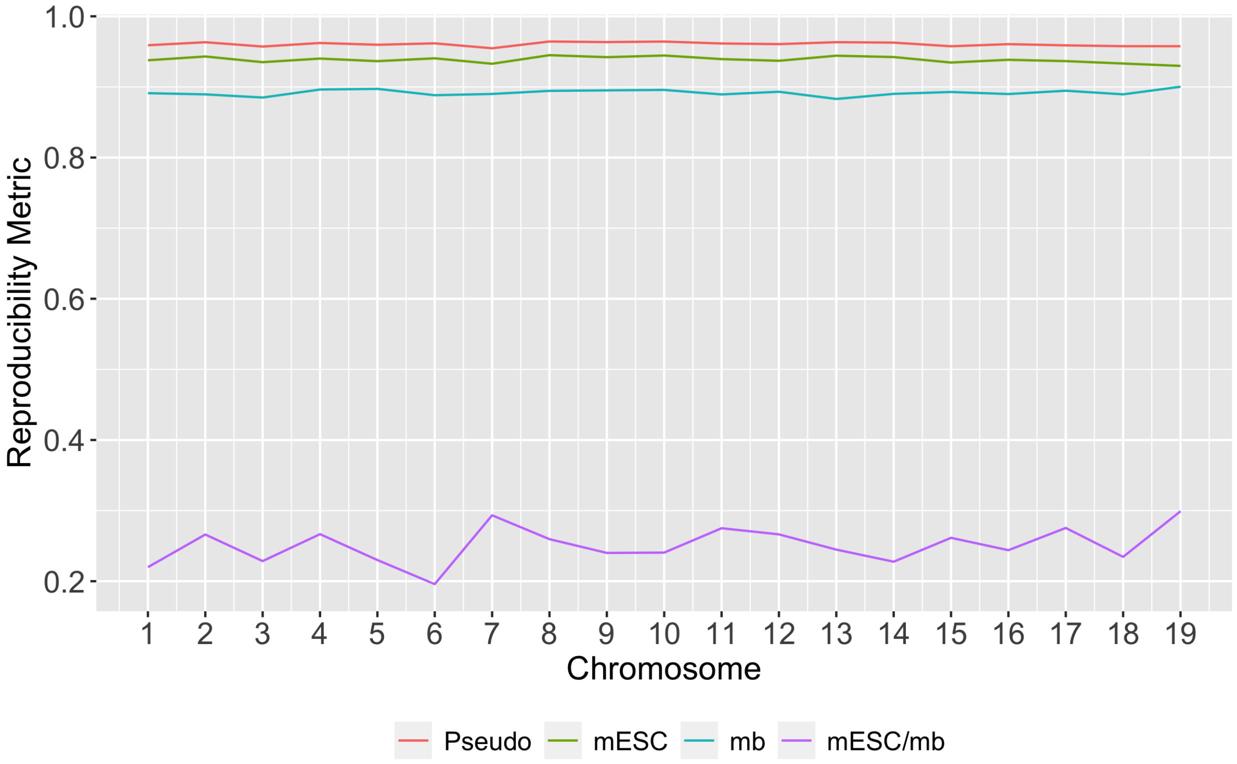

Figure B.2: Metrics obtained applying HPRep to two mouse embryonic stem cell (mESC) and mouse brain (mb) tissues H3K4me3 PLAC-seq samples. Pseudo replicates were generated from pooling mESC samples followed by random sampling via a Binomial (p=0.5) distribution. Cross sample results represent the mean of four cross-tissue pairings.

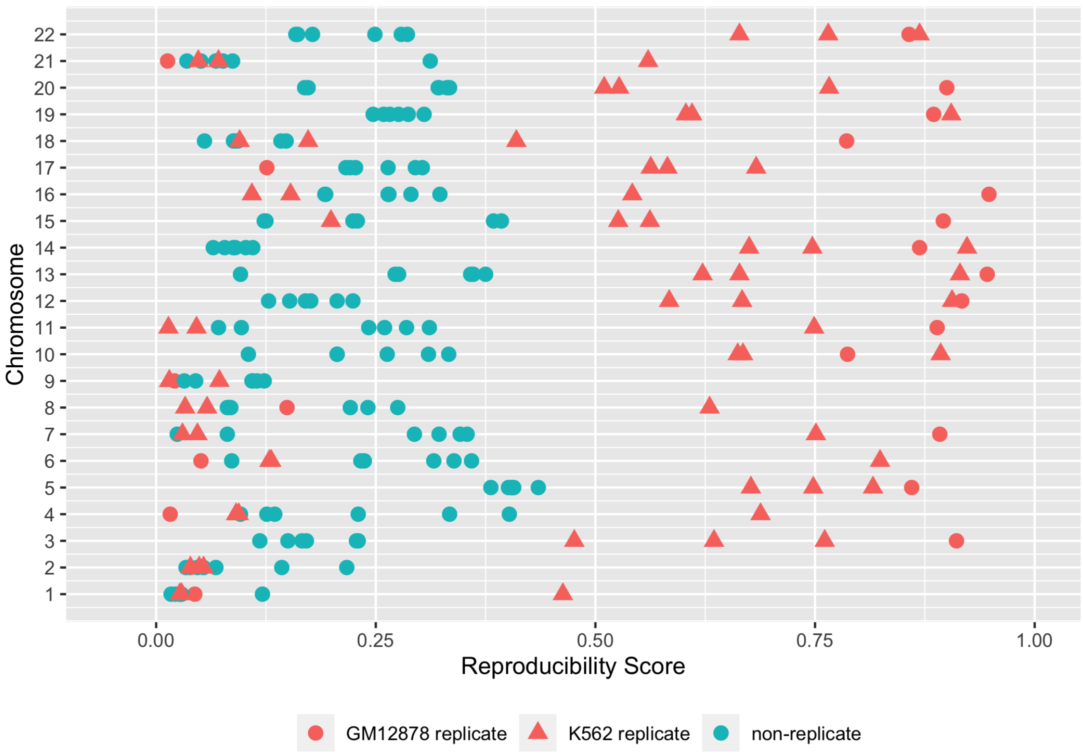

Figure B.3: HiC-Spector results by chromosome. The plotted results clearly demonstrate the chromosome to chromosome variability we do not see with HiCRep or HPRep with this data. For example, the chromosome 22 results are as expected whereas the chromosome 21 results fail to distinguish between 5 of the 6 non-replicates and 3 of the 4 replicates.

Table B.1: The number of eigenvectors used by HiC-Spector by chromosome in the GM12878 and K562 H3K27ac HiChIP experiment. Subscripts are used to denote the six non-repicate and K563 biological replicate sample pairs respectively.

Chr	Sample pair									
	GM	NR_1	NR_2	NR_3	NR_4	NR_5	NR_6	$K562_1$	$K562_2$	$K562_3$
1	18	19	18	19	17	17	17	16	19	18
2	19	19	19	19	19	20	18	19	19	18
3	20	20	20	20	20	20	20	20	20	20
4	18	19	19	18	18	18	18	19	19	19
5	20	20	20	20	20	20	20	20	20	20
6	19	19	19	19	19	18	19	19	19	19
7	20	20	18	19	20	18	19	19	19	19
8	20	18	20	20	18	20	20	20	20	20
9	18	19	19	19	19	19	19	19	20	19
10	19	19	20	20	19	19	19	19	19	20
11	20	20	18	20	20	18	20	18	20	19
12	20	20	20	20	20	20	20	20	20	20
13	20	20	20	20	19	20	20	19	19	20
14	19	19	19	19	20	20	19	20	19	19
15	20	20	19	20	20	19	20	18	20	18
16	20	19	19	20	19	19	20	19	20	19
17	19	19	19	19	19	19	19	20	20	20
18	19	20	19	19	18	20	20	18	18	18
19	20	20	20	20	20	20	20	20	20	20
20	20	20	20	20	20	20	20	20	20	20
21	16	16	16	16	18	18	18	19	19	18
22	19	20	20	20	20	20	20	19	19	20

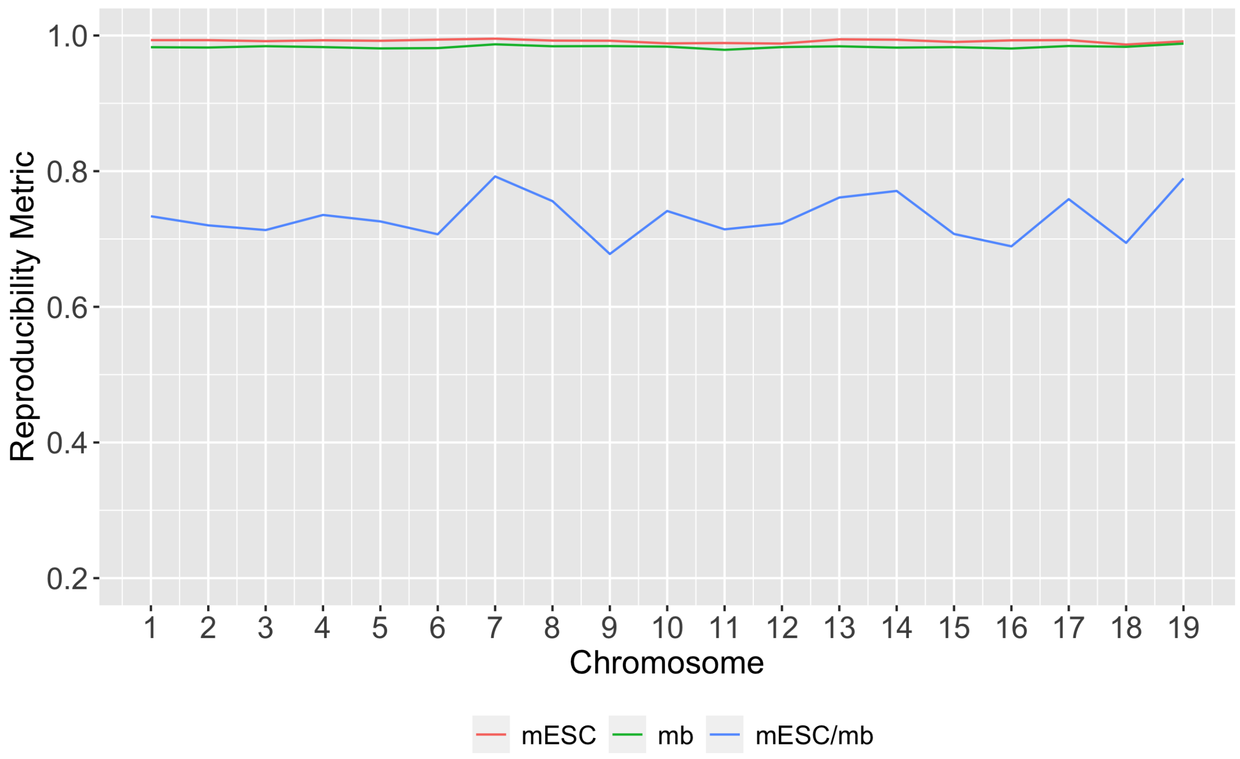

Figure B.4: Metrics obtained applying HiCRep to two mouse embryonic stem cell (mESC) and mouse brain (mb) tissues H3K4me3 PLAC-seq samples. Cross sample results represent the mean of four cross-tissue pairings.

APPENDIX C: ADDITIONAL RESULTS FOR CHAPTER 4

C.1 HP and ChIP-seq data sources

Table C.1: Source HP and CTCF ChIP-seq data for experiments described in HPTAD manuscript.

Data description	Reference (PMID)	GEO accession number or other sources
mESC H3K4me3 PLAC-seq	30986246	GSE119663
GM12878 H3K27ac HiChIP	28945252	GSE101498
mESC CTCF ChIP-seq	ENCFF508CKL	ENCODE ENCSR000CCB
GM12878 CTCF ChIP-seq	ENCFF217EAX	ENCODE ENCSR000AKB

C.2 TAD sources

mESC TADs: (Dixon *et al.*, 2012) `https://static-content.springer.com/esm/art%3A10.1038%2Fnature11082/MediaObjects/41586_2012_BFnature11082_MOESM330_ESM.xls`

GM12878 TADs: (Schmitt *et al.*, 2016) `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5478386/bin/NIHMS828671-supplement-3.xlsx`

The mESC TADs were reported referenced to the mm9 genome but HP data were referenced to the mm10 genome. Consequently, the TAD boundaries were lifted over to mm10 using the USCS liftover tool: `https://genome.ucsc.edu/cgi-bin/hgLiftOver`.

C.3 Likelihood Ratio Test

We begin by assuming that normalized interaction counts between loci i and j (X_{ij}) have a Gaussian distribution with mean μ_{ij} and known common variance σ^2. We further assume that intra-TAD interactions share a common mean as described in Equation 4.5 and that interactions between loci in the exterior regions share a common mean as described in Equation 4.7.

The likelihood ratio statistic λ is defined as

$$\lambda = -2 \ln \left[\frac{\sup_{\theta \in \Theta_0} \mathcal{L}(\theta)}{\sup_{\theta \in \Theta} \mathcal{L}(\theta)} \right]$$

Under the null hypothesis that candidate t is not a TAD boundary, we expect $\mu_T = \mu_{T+1} = \mu_{E_T}$. Therefore, the supremum of the likelihood under the null is obtained by setting θ_0 equal to the mean of X_{ij} for all $\{(i,j) : b_{t-1} \leqslant i < j < b_{t+1}\}$

Under the alternative hypothesis, the supremum of the likelihood is achieved if $\mu_T = \bar{X}_{ij}$ for all $(i,j) \in D_T$, if $\mu_{T+1} = \bar{X}_{ij}$ for all $(i,j) \in D_{T+1}$ and if $\mu_{E_T} = \bar{X}_{ij}$ for all $(i,j) \in E_T$.

Define $D_0 = D_T + D_{T+1} + E_T$. Therefore, we can write the likelihood ratio as

$$-2[l(\theta_0) - l(\hat{\theta})]$$

where $\hat{\theta} = (\mu_T, \mu_{T+1}, \mu_{E_T})$. This is equivalent to

$$-2 \left[\sum_{(i,j) \in D_T} (X_{ij} - \mu_T)^2 + \sum_{(i,j) \in D_{T+1}} (X_{ij} - \mu_{T+1})^2 + \sum_{(i,j) \in E_T} (X_{ij} - \mu_{E_T})^2 - \sum_{(i,j) \in D_0} (X_{ij} - \theta_0)^2 \right]$$

Asymptotically, $\lambda \sim \chi^2_2$ and we reject the null at the $\alpha = 0.05$ significance level.

C.4 Procedures for Comparative Methods

- **TopDom (Shin *et al.*, 2016)**
  - Similar to the directionality index, upstream and downstream contacts from a specific locus are counted, but these are averged ("bin score") instead of evaluated separately. Similar to the insulation score, potential TAD boundaries are identified as inflection points in the series of bin scores, however these points are determined using a piecewise linear function. TAD boundaries are selected from these potential ones by testing

  the upstream and downstream contact frequency difference using a Wilcoxan rank sum test.

  - All results obtained using the "TopDom" R package version 0.10.0 downloaded from the CRAN repository. The source code and documentation for the package is also available at https://github.com/HenrikBengtsson/TopDom. Analyses were conducted using R version 3.6.0. Default parameters were used for experiments unless otherwise noted. The window size was set to 500 kb to closely match the average window size used by HPTAD after tuning. This was extended as described in Results.

- **Insulation Score (Crane *et al.*, 2015)**

  - Scanning the diagonal of a contact matrix, for each locus an insulation score it computed representing the sum of interactions spanning that locus within a specified neighborhood. Intuitively, a TAD boundary should be reprented by a local minima insulation score, and the method employs an algorithmic process by which such minima are selected and evaluated. It should be noted that this process involves parameters whose value can have a significant impact on the number of TADs identified.
  - All results obtained using the command line tools from FAN-C version 0.9.1 obtained from https://vaquerizaslab.github.io/fanc/index.html. Specifically, the "insulation" and "boundaries" functions were used sequentially to call TADs. Default parameters were used for experiments unless otherwise noted. The window size for the "insulation" function was adjusted to 20 and the minimum score of the "boundaries" function was set to 0.7 for the two experiments described in Results.

- **Grinch (Lee and Roy, 2021)**

  - Non-negative factorization of the contact matrix is followed by a local smoothing procedure to account for the distance dependence of Hi-C contact frequency. One of the smoothed factor matrices is treated as a set of latent features, and TADs represent clustered regions found by applying k-medoids clustering.

  - All results obtained using GRiNCH version 1.0.0 obtained from `https://roy-lab.github.io/grinch/`. Default parameters were used for all experiments.

- **OnTAD (An *et al.*, 2019)**
  - Similar to insulation score and TopDom, the average contact frequencies within a diamond-shaped window are determined by sliding along the diagonal. The process is repeated varying the window size, and local minima are selected for each size. Hierarchical TADs are called from the union of potential boundaries using a dynamic programming algorithm.
  - All results obtained using OnTAD version 1.4 obtained from `https://github.com/anlin00007/OnTAD`. Default parameters were used for all experiments.

C.5 Window Size Tuning

The HPTAD process has a single adjustable parameter which determines the evaluation window size used to determine average normalized contact frequency. We tune this parameter using a data-driven approach that obviates the need for user specification. This is beneficial considering a lack of criteria for evaluating appropriate values in an agnostic fashion, that is without "dialing in" specific TAD characteristics.

Let N_{ij} represent the raw number of counts between loci i and j and further define set $S_t = \{(i,j) : i < j; j - i = t\}$ for $t \in \{1, 2, \ldots, 20\}$. We define

$$N_t = \frac{\sum_{(i,j)\in S_t} N_{ij}}{|S_t|}$$

where $|S_t|$ represents the cardinality of set S_t. One of the characteristics of Hi-C data is its strong distance dependence; average contact frequency decreases (quickly) as a function of genomic distance. Consequenty, the values of N_t are expected to decrease monotonically with increasing t.

We set the window size to the smallest value of t such that

$$1 - \frac{N_{t+1}}{N_t} \leqslant 0.1$$

C.6 CTCF Enrichment

The CTCF ChIP-seq peaks were obtained from the sources referenced in Supplementary Table S4.1. The GM12878 peaks were reported referenced to build hg38 but HP data were referenced to build hg19 genome. Consequently, the CTCF peaks were lifted over to hg19 using the USCS liftover tool: `https://genome.ucsc.edu/cgi-bin/hgLiftOver`.

Peak density is reported as the number of peaks whose start and end points fall entirely within a specific 40 Kb locus. The fold change is reported as the ratio of (# of bins containing at least one peak / # of bins) to (# of peaks / # TADs).

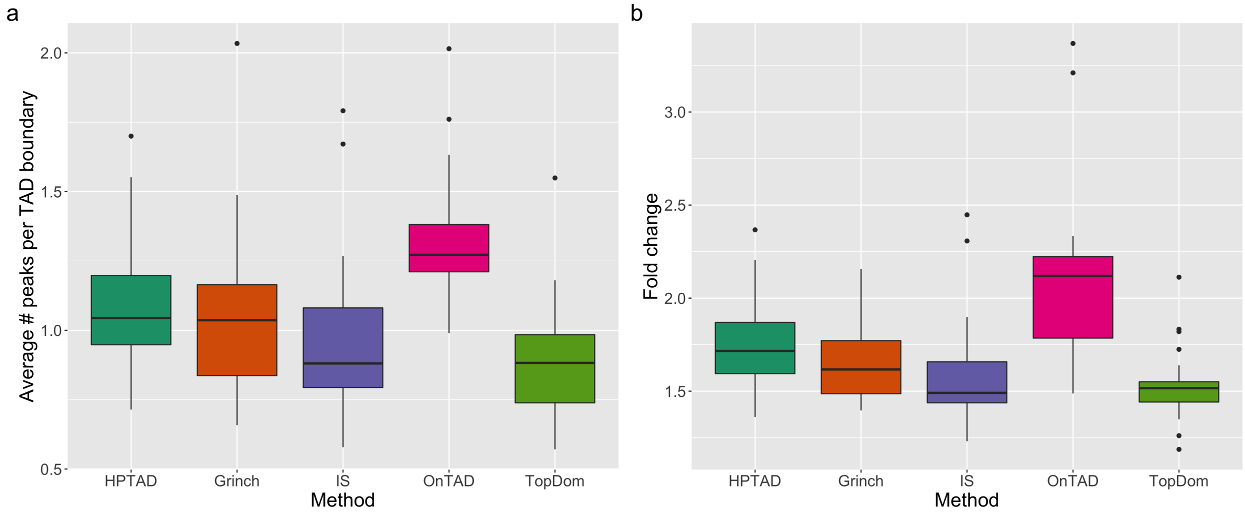

Figure C.1: CTCF enrichment for GM12878 experiment - Average number of peaks per TAD boundary (a) and fold change (b). Boxplots displaying results for HPTAD vs. the indicated methods using normalized contact counts as input. The top, middle, and bottom lines of the boxes represent the third, second, and first quartiles respectively and whiskers extend to 1.5 times the interquartile range. Outliers are plotted as points. Displayed results are for all 23 human chromosomes.

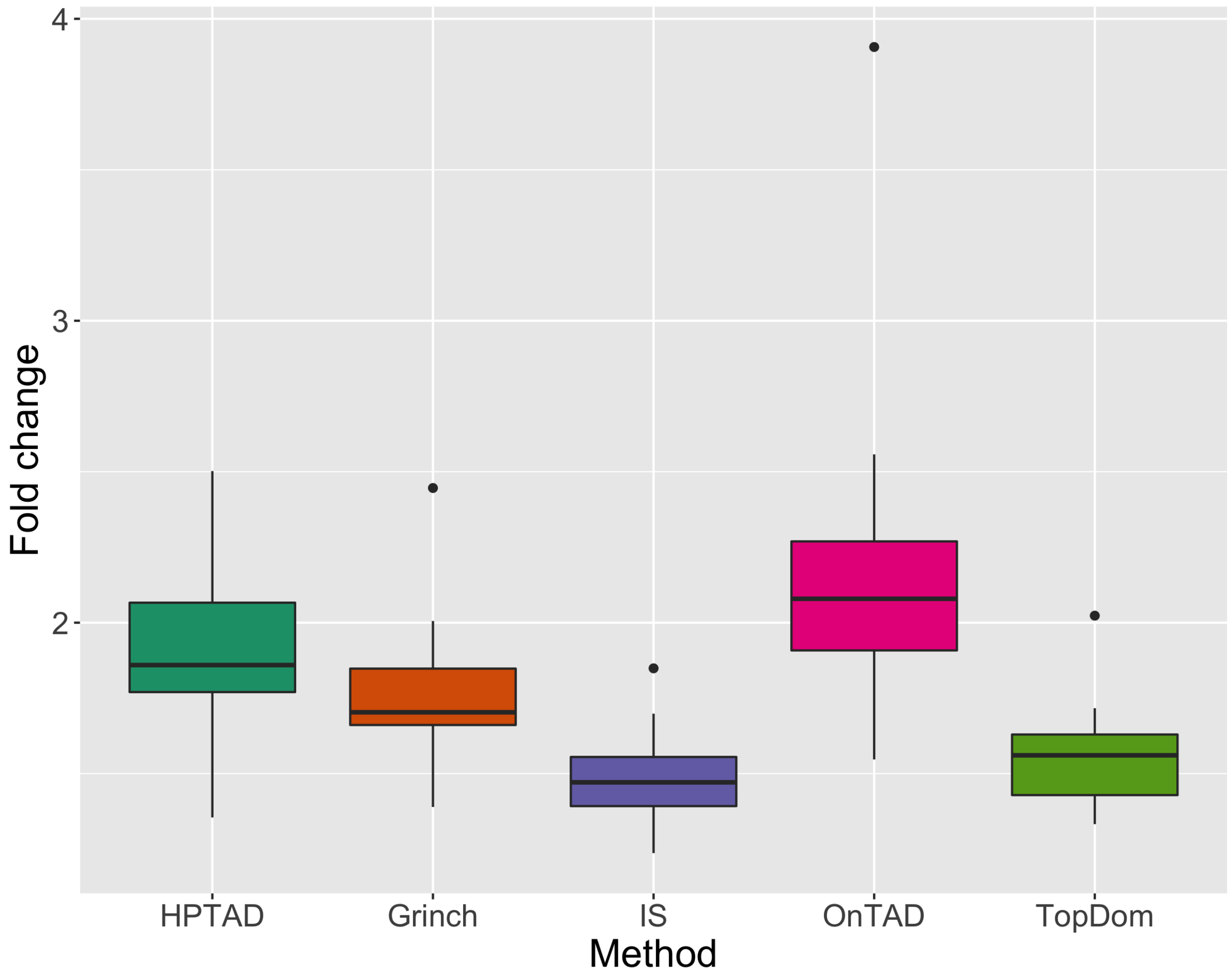

Figure C.2: CTCF fold change for mESC H3K4me3 PLAC-seq experiment - Boxplots displaying results for HPTAD vs. the indicated methods using normalized contact counts as input. The top, middle, and bottom lines of the boxes represent the third, second, and first quartiles respectively and whiskers extend to 1.5 times the interquartile range. Outliers are plotted as points. Displayed results are for all 20 mouse chromosomes.

C.7 Jaccard Index Comparison

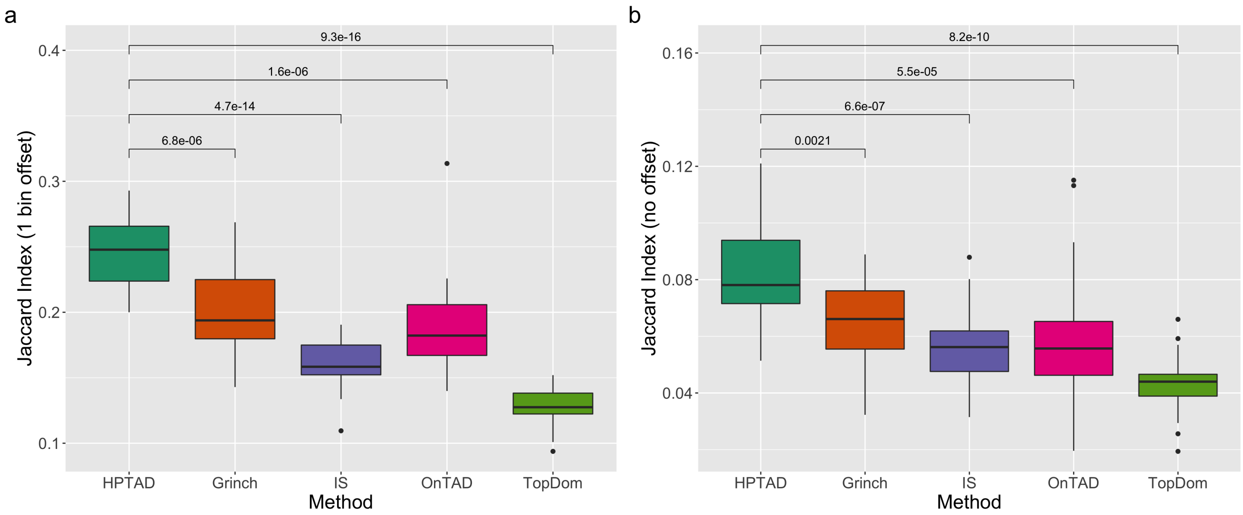

Figure C.3: Comparison of modified Jaccard index (a) and standard Jaccard index (b) applied to GM12878 H3K27ac HiChIP data.

C.8 Number of Reduced TADs

We artificially reduced the number of TADs called by both TopDom and IS by modifying the appropriate input parameters. For TopDom, this consisted of setting the “window.size” parameter to the maximum value in the recommended range, which is 20. Note, the default value is 5.

There were two parameters we modified in the FAN-C implementation of the insulation score method presented by Crane *et al.*. The first paramter, “window-sizes”, was set to 2 Mb; the original experiements were conducted using a window size of 500 Kb consistent with other methods. The second parameter, “min-score”, thresholds the delta vector and was set to 0.7. Note, the default setting is no threshold.

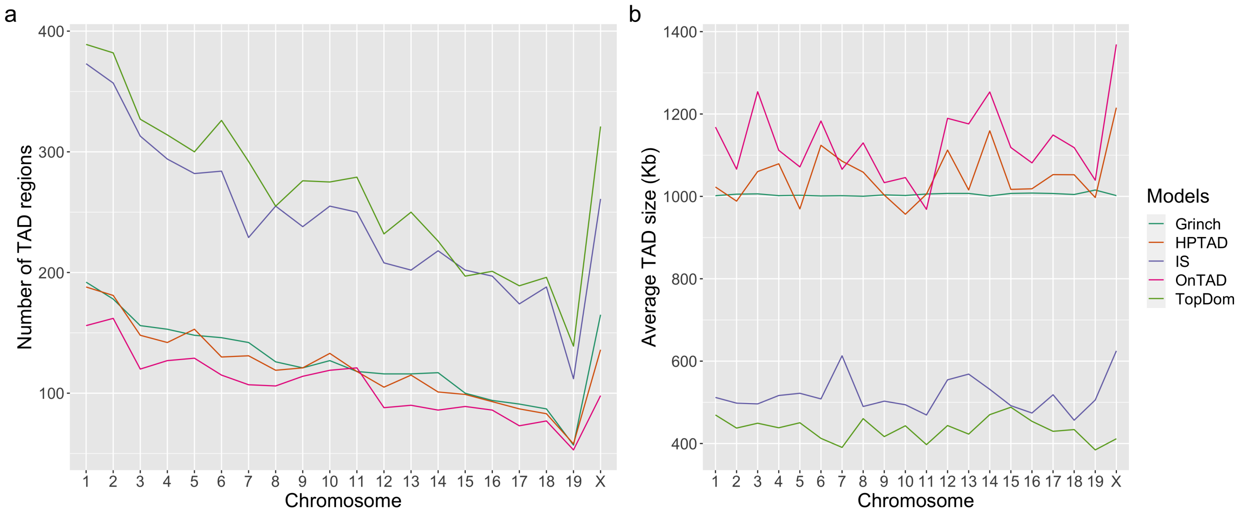

Figure C.4: Scatterplots comparing the number (a) and average size (b) of TADs identified by HPTAD, OnTAD, TopDom, Grinch, and IS. Results are displayed as average per chromosome. For all methods we observed the same trend of decreasing number of TADs with decreasing autosomal chromosome number. Average TAD size remains relatively consistent across chromosomes for all methods. Note that Grinch explicitly models the average TAD size to be 1 Mb.

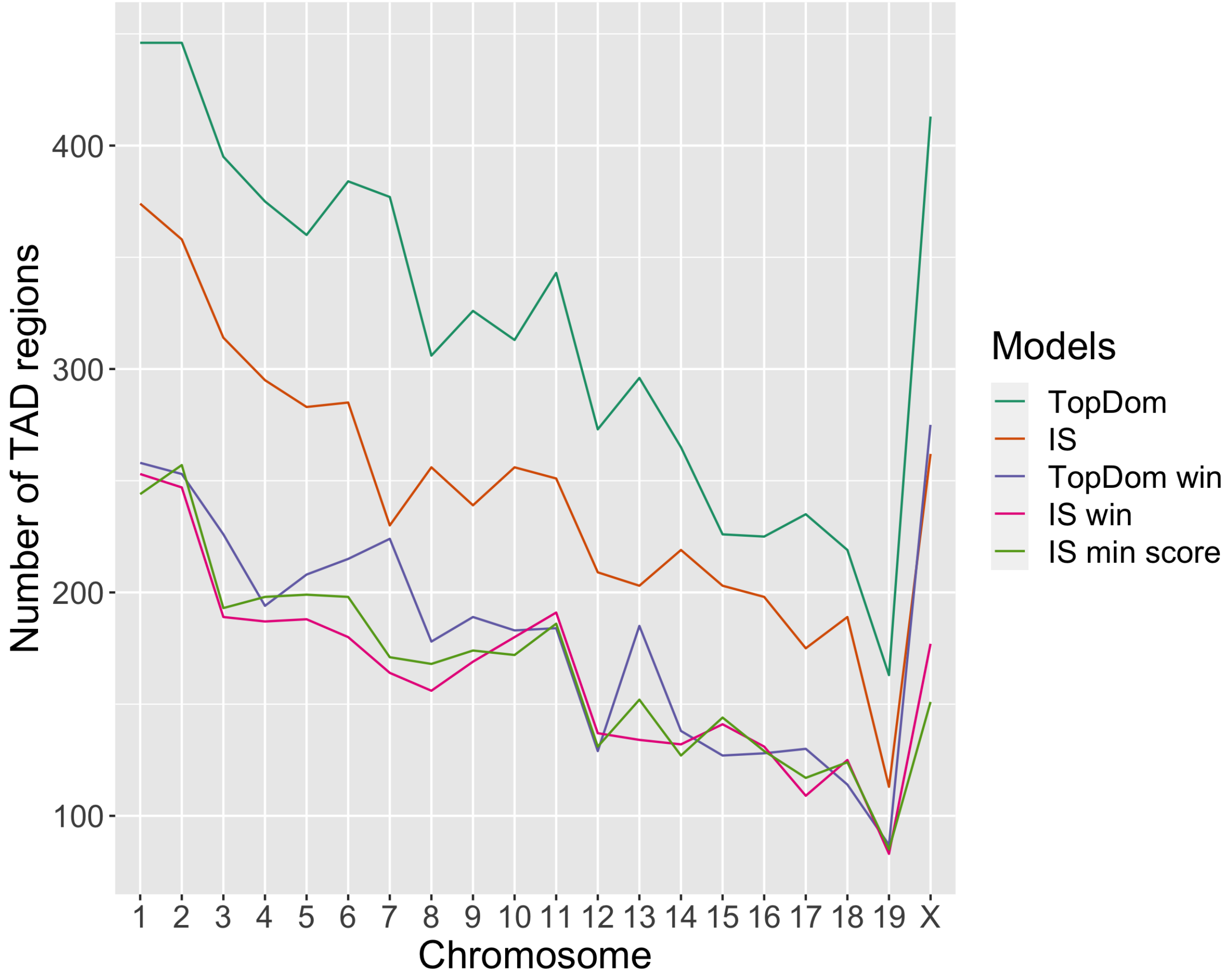

Figure C.5: Scatterplot of number of TADs called before and after reduction achieved by changing default parameters for TopDom and IS methods.

REFERENCES

A. Bennett, D., A. Schneider, J., Arvanitakis, Z., and S. Wilson, R. (2012a). Overview and Findings from the Religious Orders Study. *Current Alzheimer Research*, **9**(6), 628–645.

A. Bennett, D., A. Schneider, J., S. Buchman, A., L. Barnes, L., A. Boyle, P., and S. Wilson, R. (2012b). Overview and Findings from the Rush Memory and Aging Project. *Current Alzheimer Research*, **9**(6), 646–663.

Akdemir, K. C., Le, V. T., Chandran, S., Li, Y., Verhaak, R. G., Beroukhim, R., Campbell, P. J., Chin, L., Dixon, J. R., and Futreal, P. A. (2020). Disruption of chromatin folding domains by somatic genomic rearrangements in human cancer. *Nature Genetics 2020 52:3*, **52**(3), 294–305.

An, L., Yang, T., Yang, J., Nuebler, J., Xiang, G., Hardison, R. C., Li, Q., and Zhang, Y. (2019). OnTAD: hierarchical domain structure reveals the divergence of activity among TADs and boundaries. *Genome Biology 2019 20:1*, **20**(1), 1–16.

Ardakany, A. R. and Lonardi, S. (2017). Efficient and Accurate Detection of Topologically Associating Domains from Contact Maps. In R. Schwartz and K. Reinert, editors, *17th International Workshop on Algorithms in Bioinformatics (WABI 2017)*, volume 88 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 22:1–22:11, Dagstuhl, Germany. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik.

Barbeira, A. N., Pividori, M. D., Zheng, J., Wheeler, H. E., Nicolae, D. L., and Im, H. K. (2019). Integrating predicted transcriptome from multiple tissues improves association detection. *PLoS Genetics*, **15**(1), e1007889.

Battle, A., Mostafavi, S., Zhu, X., Potash, J. B., Weissman, M. M., McCormick, C., Haudenschild, C. D., Beckman, K. B., Shi, J., Mei, R., Urban, A. E., Montgomery, S. B., Levinson, D. F., and Koller, D. (2014). Characterizing the genetic basis of transcriptome diversity through RNA-sequencing of 922 individuals. *Genome Research*, **24**(1), 14–24.

Bhattacharya, A., Li, Y., and Love, M. (2020). MOSTWAS: Multi-Omic Strategies for Transcriptome-Wide Association Studies. *bioRxiv*, page doi:10.1101/2020.04.17.047225.

Bhutani, K., Sarkar, A., Park, Y., Kellis, M., and Schork, N. J. (2017). Modeling prediction error improves power of transcriptome-wide association studies. *bioRxiv*, page doi:10.1101/108316.

Bickmore, W. A. (2013). The spatial organization of the human genome. *Annual Review of Genomics and Human Genetics*, **14**, 67–84.

Bild, D. E., Bluemke, D. A., Burke, G. L., Detrano, R., Diez Roux, A. V., Folsom, A. R., Greenland, P., Jacobs, D. R., Kronmal, R., Liu, K., Nelson, J. C., O'Leary, D., Saad, M. F., Shea, S., Szklo, M., and Tracy, R. P. (2002). Multi-Ethnic Study of Atherosclerosis: Objectives and design. *American Journal of Epidemiology*, **156**(9), 871–881.

Bonev, B. and Cavalli, G. (2016). Organization and function of the 3D genome. *Nature Reviews Genetics*, **17**(11), 661–678.

Brynedal, B., Choi, J. M., Raj, T., Bjornson, R., Stranger, B. E., Neale, B. M., Voight, B. F., and Cotsapas, C. (2017). Large-Scale trans-eQTLs Affect Hundreds of Transcripts and Mediate Patterns of Transcriptional Co-regulation. *American Journal of Human Genetics*, **100**(4), 581–591.

Buniello, A., Macarthur, J. A., Cerezo, M., Harris, L. W., Hayhurst, J., Malangone, C., McMahon, A., Morales, J., Mountjoy, E., Sollis, E., Suveges, D., Vrousgou, O., Whetzel, P. L., Amode, R., Guillen, J. A., Riat, H. S., Trevanion, S. J., Hall, P., Junkins, H., Flicek, P., Burdett, T., Hindorff, L. A., Cunningham, F., and Parkinson, H. (2019). The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Research*, **47**(D1), D1005–D1012.

Burgess, S. and Thompson, S. G. (2013). Use of allele scores as instrumental variables for Mendelian randomization. *International Journal of Epidemiology*, **42**(4), 1134–1144.

Burton, P. R., Clayton, D. G., Cardon, L. R., Craddock, N., Deloukas, P., Duncanson, A., Kwiatkowski, D. P., McCarthy, M. I., Ouwehand, W. H., Samani, N. J., Todd, J. A., Donnelly, P., Barrett, J. C., Burton, P. R., Davison, D., Donnelly, P., Easton, D., Evans, D., Leung, H.-T., Marchini, J. L., Morris, A. P., Spencer, C. C. A., Tobin, M. D., Cardon, L. R., Clayton, D. G., Attwood, A. P., Boorman, J. P., Cant, B., Everson, U., Hussey, J. M., Jolley, J. D., Knight, A. S., Koch, K., Meech, E., Nutland, S., Prowse, C. V., Stevens, H. E., Taylor, N. C., Walters, G. R., Walker, N. M., Watkins, N. A., Winzer, T., Todd, J. A., Ouwehand, W. H., Jones, R. W., McArdle, W. L., Ring, S. M., Strachan, D. P., Pembrey, M., Breen, G., St Clair, D., Caesar, S., Gordon-Smith, K., Jones, L., Fraser, C., Green, E. K., Grozeva, D., Hamshere, M. L., Holmans, P. A., Jones, I. R., Kirov, G., Moskvina, V., Nikolov, I., O'Donovan, M. C., Owen, M. J., Craddock, N., Collier, D. A., Elkin, A., Farmer, A., Williamson, R., McGuffin, P., Young, A. H., Ferrier, I. N., Ball, S. G., Balmforth, A. J., Barrett, J. H., Bishop, D. T., Iles, M. M., Maqbool, A., Yuldasheva, N., Hall, A. S., Braund, P. S., Burton, P. R., Dixon, R. J., Mangino, M., Stevens, S., Tobin, M. D., Thompson, J. R., Samani, N. J., Bredin, F., Tremelling, M., Parkes, M., Drummond, H., Lees, C. W., Nimmo, E. R., Satsangi, J., Fisher, S. A., Forbes, A., Lewis, C. M., Onnie, C. M., Prescott, N. J., Sanderson, J., Mathew, C. G., Barbour, J., Mohiuddin, M. K., Todhunter, C. E., Mansfield, J. C., Ahmad, T., Cummings, F. R., Jewell, D. P., Webster, J., Brown, M. J., Clayton, D. G., Lathrop, G. M., Connell, J., Dominiczak, A., Samani, N. J., Marcano, C. A. B., Burke, B., Dobson, R., Gungadoo, J., Lee, K. L., Munroe, P. B., Newhouse, S. J., Onipinla, A., Wallace, C., Xue, M., Caulfield, M., Farrall, M., Barton, A., and Genomics (BRAGGS), T. B. i. R. A. G., Bruce, I. N., Donovan, H., Eyre, S., Gilbert, P. D., Hider, S. L., Hinks, A. M., John, S. L., Potter, C., Silman, A. J., Symmons, D. P. M., Thomson, W., Worthington, J., Clayton, D. G., Dunger, D. B., Nutland, S., Stevens, H. E., Walker, N. M., Widmer, B., Todd, J. A., Frayling, T. M., Freathy, R. M., Lango, H., Perry, J. R. B., Shields, B. M., Weedon, M. N., Hattersley, A. T., Hitman, G. A., Walker, M., Elliott, K. S., Groves, C. J., Lindgren, C. M., Rayner, N. W., Timpson, N. J., Zeggini, E., McCarthy, M. I., Newport, M., Sirugo, G., Lyons, E., Vannberg, F., Hill, A. V. S., Bradbury, L. A., Farrar, C., Pointon, J. J., Wordsworth, P., Brown, M. A., Franklyn, J. A., Heward, J. M., Simmonds, M. J., Gough, S. C. L., Seal, S., Susceptibility Collaboration (UK), B. C., Stratton, M. R., Rahman, N., Ban, M., Goris, A., Sawcer, S. J., Compston, A., Conway, D., Jallow, M., Newport, M., Sirugo, G., Rockett, K. A., Kwiatkowski, D. P., Bumpstead, S. J., Chaney, A., Downes, K., Ghori, M. J. R., Gwilliam, R., Hunt, S. E., Inouye, M., Keniry, A., King, E., McGinnis, R., Potter, S., Ravindrarajah, R., Whittaker, P., Widden, C., Withers, D., Deloukas, P., Leung, H.-T., Nutland, S., Stevens, H. E., Walker, N. M., Todd, J. A., Easton, D., Clayton, D. G., Burton, P. R., Tobin, M. D., Barrett, J. C., Evans, D., Morris, A. P., Cardon, L. R., Cardin, N. J., Davison, D., Ferreira, T., Pereira-Gale, J., Hallgrimsdóttir, I. B., Howie, B. N., Marchini, J. L., Spencer, C. C. A., Su, Z., Teo, Y. Y., Vukcevic, D., Donnelly, P., Bentley, D., Brown, M. A., Cardon, L. R., Caulfield, M., Clayton, D. G., Compston, A., Craddock, N., Deloukas, P., Donnelly, P., Farrall, M., Gough, S. C. L., Hall, A. S., Hattersley, A. T., Hill, A. V. S., Kwiatkowski, D. P., Mathew, C. G., McCarthy, M. I., Ouwehand, W. H., Parkes, M., Pembrey, M., Rahman, N., Samani, N. J., Stratton, M. R., Todd, J. A., Worthington, J., Consortium, T. W. T. C. C., Committee, M., Committee, D., Analysis, Controls, U. K. B. S., of Cambridge, U., Controls, . B. C., Disorder, B., Disease, C. A., Disease, C., Hypertension, Arthritis, R., Diabetes, T. ., Diabetes, T. ., Tuberculosis, Spondylitis, A., Disease, A. T., Cancer, B., Sclerosis, M., Controls, G., DNA Data QC and Informatics, G., Statistics, and Investigators, P. (2007). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, **447**(7145), 661–678.

Bycroft, C., Freeman, C., Petkova, D., Band, G., Elliott, L. T., Sharp, K., Motyer, A., Vukcevic, D., Delaneau, O., O'Connell, J., Cortes, A., Welsh, S., Young, A., Effingham, M., McVean, G., Leslie, S., Allen, N., Donnelly, P., and Marchini, J. (2018). The UK Biobank resource with deep phenotyping and genomic data. *Nature*, **562**(7726), 203–209.

Cano-Gamez, E. and Trynka, G. (2020). From GWAS to Function: Using Functional Genomics to Identify the Mechanisms Underlying Complex Diseases. *Frontiers in Genetics*, **11**, 424.

Carithers, L. J., Ardlie, K., Barcus, M., Branton, P. A., Britton, A., Buia, S. A., Compton, C. C., Deluca, D. S., Peter-Demchok, J., Gelfand, E. T., Guan, P., Korzeniewski, G. E., Lockhart, N. C., Rabiner, C. A., Rao, A. K., Robinson, K. L., Roche, N. V., Sawyer, S. J., Segrè, A. V., Shive, C. E., Smith, A. M., Sobin, L. H., Undale, A. H., Valentino, K. M., Vaught, J., Young, T. R., Moore, H. M., Barker, L., Basile, M., Battle, A., Boyer, J., Bradbury, D., Bridge, J. P., Brown, A., Burges, R., Choi, C., Colantuoni, D., Cox, N., Dermitzakis, E. T., Derr, L. K., Dinsmore, M. J., Erickson, K., Fleming, J., Flutre, T., Foster, B. A., Gamazon, E. R., Getz, G., Gillard, B. M., Guigó, R., Hambright, K. W., Hariharan, P., Hasz, R., Im, H. K., Jewell, S., Karasik, E., Kellis, M., Kheradpour, P., Koester, S., Koller, D., Konkashbaev, A., Lappalainen, T., Little, R., Liu, J., Lo, E., Lonsdale, J. T., Lu, C., MacArthur, D. G., Magazine, H., Maller, J. B., Marcus, Y., Mash, D. C., McCarthy, M. I., McLean, J., Mestichelli, B., Miklos, M., Monlong, J., Mosavel, M., Moser, M. T., Mostafavi, S., Nicolae, D. L., Pritchard, J., Qi, L., Ramsey, K., Rivas, M. A., Robles, B. E., Rohrer, D. C., Salvatore, M., Sammeth, M., Seleski, J., Shad, S., Siminoff, L. A., Stephens, M., Struewing, J., Sullivan, T., Sullivan, S., Syron, J., Tabor, D., Taherian, M., Tejada, J., Temple, G. F., Thomas, J. A., Thomson, A. W., Tidwell, D., Traino, H. M., Tu, Z., Valley, D. R., Volpi, S., Walters, G. D., Ward, L. D., Wen, X., Winckler, W., Wu, S., Zhu, J., Abdallah, A., Addington, A., Anderson, J. M., Bender, P. K., Cosentino, M., Diaz-Mayoral, N., Engel, T., Garci, F., Green, A., Hammond, T., Jaffe, K., Keen, J., Kennedy, M., Kigonya, P., Lander, B., Nampally, S., Ny, C., Robb, J., Santhanum, V., Sharopova, N., Singh, S., Soria, C., Sturcke, A., Sukari, S., Thomson, E. J., Tomaszewski, M., Trowbridge, C., Udoye, F., Vanscoy, D., Vatanian, N., Wilder, E. L., and Williams, P. (2015). A Novel Approach to High-Quality Postmortem Tissue Procurement: The GTEx Project. *Biopreservation and Biobanking*, **13**(5), 311–317.

Chen, F., Li, G., Zhang, M. Q., and Chen, Y. (2018). HiCDB: a sensitive and robust method for detecting contact domain boundaries. *Nucleic Acids Research*, **46**(21), 11239–11250.

Chen, J., Hero, A. O., III, and Rajapakse, I. (2016). Spectral identification of topological domains. *Bioinformatics*, **32**(14), 2151.

Crane, E., Bian, Q., McCord, R. P., Lajoie, B. R., Wheeler, B. S., Ralston, E. J., Uzawa, S., Dekker, J., and Meyer, B. J. (2015). Condensin-driven remodelling of X chromosome topology during dosage compensation. *Nature*, **523**(7559), 240–244.

Cremer, M., Grasser, F., Lanctôt, C., Müller, S., Neusser, M., Zinner, R., Solovei, I., and Cremer, T. (2012). Multicolor 3D Fluorescence In Situ Hybridization for Imaging Interphase Chromosomes. *Methods in Molecular Biology*, **463**, 205–239.

de Wit, E. and de Laat, W. (2012). A decade of 3C technologies: Insights into nuclear organization. *Genes and Development*, **26**(1), 11–24.

Dekker, J., Rippe, K., Dekker, M., and Kleckner, N. (2002). Capturing chromosome conformation. *Science*, **295**(5558), 1306–1311.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, **39**(1), 1–38.

Dickson, S. P., Wang, K., Krantz, I., Hakonarson, H., and Goldstein, D. B. (2010). Rare Variants Create Synthetic Genome-Wide Associations. *PLoS Biology*, **8**(1).

Dixon, J. R., Selvaraj, S., Yue, F., Kim, A., Li, Y., Shen, Y., Hu, M., Liu, J. S., and Ren, B. (2012). Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature*, **485**(7398), 376–380.

Dixon, J. R., Jung, I., Selvaraj, S., Shen, Y., Antosiewicz-Bourget, J. E., Lee, A. Y., Ye, Z., Kim, A., Rajagopal, N., Xie, W., Diao, Y., Liang, J., Zhao, H., Lobanenkov, V. V., Ecker, J. R., Thomson, J. A., and Ren, B. (2015). Chromatin architecture reorganization during stem cell differentiation. *Nature*, **518**(7539), 331–336.

Dixon, J. R., Gorkin, D. U., and Ren, B. (2016). Chromatin Domains: The Unit of Chromosome Organization. *Molecular Cell*, **62**(5), 668–680.

Dostie, J., Richmond, T. A., Arnaout, R. A., Selzer, R. R., Lee, W. L., Honan, T. A., Rubio, E. D., Krumm, A., Lamb, J., Nusbaum, C., Green, R. D., and Dekker, J. (2006). Chromosome Conformation Capture Carbon Copy (5C): A massively parallel solution for mapping interactions between genomic elements. *Genome Research*, **16**(10), 1299–1309.

Dudbridge, F. (2013). Power and Predictive Accuracy of Polygenic Risk Scores. *PLoS Genetics*, **9**(3), 1003348.

Durand, N. C., Robinson, J. T., Shamim, M. S., Machol, I., Mesirov, J. P., Lander, E. S., and Aiden, E. L. (2016). Juicebox Provides a Visualization System for Hi-C Contact Maps with Unlimited Zoom. *Cell Systems*, **3**(1), 99–101.

Edwards, S. L., Beesley, J., French, J. D., and Dunning, M. (2013). Beyond GWASs: Illuminating the dark road from association to function. *American Journal of Human Genetics*, **93**(5), 779–797.

Erbe, M., Hayes, B. J., Matukumalli, L. K., Goswami, S., Bowman, P. J., Reich, C. M., Mason, B. A., and Goddard, M. E. (2012). Improving accuracy of genomic predictions within and between dairy cattle breeds with imputed high-density single nucleotide polymorphism panels. *Journal of Dairy Science*, **95**(7), 4114–4129.

Fang, R., Yu, M., Li, G., Chee, S., Liu, T., Schmitt, A. D., and Ren, B. (2016). Mapping of long-range chromatin interactions by proximity ligation-assisted ChIP-seq. *Cell Research*, **26**(12), 1345–1348.

Filippova, D., Patro, R., Duggal, G., and Kingsford, C. (2014). Identification of alternative topological domains in chromatin. *Algorithms for Molecular Biology*, **9**(1), 1–11.

Fuller, W. (1987). *Measurement error models*. Wiley, New York.

Fullwood, M. J., Liu, M. H., Pan, Y. F., Liu, J., Xu, H., Mohamed, Y. B., Orlov, Y. L., Velkov, S., Ho, A., Mei, P. H., Chew, E. G., Huang, P. Y. H., Welboren, W. J., Han, Y., Ooi, H. S., Ariyaratne, P. N., Vega, V. B., Luo, Y., Tan, P. Y., Choy, P. Y., Wansa, K. D. A., Zhao, B., Lim, K. S., Leow, S. C., Yow, J. S., Joseph, R., Li, H., Desai, K. V., Thomsen, J. S., Lee, Y. K., Karuturi, R. K. M., Herve, T., Bourque, G., Stunnenberg, H. G., Ruan, X., Cacheux-Rataboul, V., Sung, W. K., Liu, E. T., Wei, C. L., Cheung, E., and Ruan, Y. (2009). An oestrogen-receptor-α-bound human chromatin interactome. *Nature*, **462**(7269), 58–64.

Gallagher, M. D. and Chen-Plotkin, A. S. (2018). The Post-GWAS Era: From Association to Function. *American Journal of Human Genetics*, **102**(5), 717–730.

Gamazon, E. R., Wheeler, H. E., Shah, K. P., Mozaffari, S. V., Aquino-Michaels, K., Carroll, R. J., Eyler, A. E., Denny, J. C., Nicolae, D. L., Cox, N. J., and Im, H. K. (2015). A gene-based association method for mapping traits using reference transcriptome data. *Nature Genetics*, **47**(9), 1091–1098.

Gibson, G. (2012). Rare and common variants: Twenty arguments. *Nature Reviews Genetics*, **13**(2), 135–145.

Giorgio, E., Robyr, D., Spielmann, M., Ferrero, E., Di Gregorio, E., Imperiale, D., Vaula, G., Stamoulis, G., Santoni, F., Atzori, C., Gasparini, L., Ferrera, D., Canale, C., Guipponi, M., Pennacchio, L. A., Antonarakis, S. E., Brussino, A., and Brusco, A. (2015). A large genomic deletion leads to enhancer adoption by the lamin B1 gene: a second path to autosomal dominant adult-onset demyelinating leukodystrophy (ADLD). *Human Molecular Genetics*, **24**(11), 3143–3154.

Gleason, K., Yang, F., and Chen, L. (2020). A robust two-sample Mendelian Randomization method integrating GWAS with multi-tissue eQTL summary statistics. *bioRxiv*, page doi:2020.06.04.135541.

Gusev, A., Ko, A., Shi, H., Bhatia, G., Chung, W., Penninx, B. W., Jansen, R., De Geus, E. J., Boomsma, D. I., Wright, F. A., Sullivan, P. F., Nikkola, E., Alvarez, M., Civelek, M., Lusis, A. J., Lehtimäki, T., Raitoharju, E., Kähönen, M., Seppälä, I., Raitakari, O. T., Kuusisto, J., Laakso, M., Price, A. L., Pajukanta, P., and Pasaniuc, B. (2016). Integrative approaches for large-scale transcriptome-wide association studies. *Nature Genetics*, **48**(3), 245–252.

Haddad, N., Vaillant, C., and Jost, D. (2017). IC-Finder: inferring robustly the hierarchical organization of chromatin folding. *Nucleic Acids Research*, **45**(10), e81–e81.

Han, J., Zhang, Z., and Wang, K. (2018). 3C and 3C-based techniques: The powerful tools for spatial genome organization deciphering. *Molecular Cytogenetics*, **11**(1).

Hindorff, L. A., Gillanders, E. M., and Manolio, T. A. (2011). Genetic architecture of cancer and other complex diseases: lessons learned and future directions. *Carcinogenesis*, **32**(7), 945–954.

Hoerl, A. E. and Kennard, R. W. (1970). Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics*, **12**(1), 55–67.

Hoffmann, T. J., Kvale, M. N., Hesselson, S. E., Zhan, Y., Aquino, C., Cao, Y., Cawley, S., Chung, E., Connell, S., Eshragh, J., Ewing, M., Gollub, J., Henderson, M., Hubbell, E., Iribarren, C., Kaufman, J., Lao, R. Z., Lu, Y., Ludwig, D., Mathauda, G. K., McGuire, W., Mei, G., Miles, S., Purdy, M. M., Quesenberry, C., Ranatunga, D., Rowell, S., Sadler, M., Shapero, M. H., Shen, L., Shenoy, T. R., Smethurst, D., Van den Eeden, S. K., Walter, L., Wan, E., Wearley, R., Webster, T., Wen, C. C., Weng, L., Whitmer, R. A., Williams, A., Wong, S. C., Zau, C., Finn, A., Schaefer, C., Kwok, P. Y., and Risch, N. (2011). Next generation genome-wide association tool: Design and coverage of a high-throughput European-optimized SNP array. *Genomics*, **98**(2), 79–89.

Hu, Y., Li, M., Lu, Q., Weng, H., Wang, J., Zekavat, S. M., Yu, Z., Li, B., Gu, J., Muchnik, S., Shi, Y., Kunkle, B. W., Mukherjee, S., Natarajan, P., Naj, A., Kuzma, A., Zhao, Y., Crane, P. K., Lu, H., and Zhao, H. (2019). A statistical framework for cross-tissue transcriptome-wide association analysis. *Nature Genetics*, **51**(3), 568–576.

International HapMap Consortium, Frazer, K. A., Ballinger, D. G., Cox, D. R., Hinds, D. A., Stuve, L. L., Gibbs, R. A., Belmont, J. W., Boudreau, A., Hardenbol, P., Leal, S. M., Pasternak, S., Wheeler, D. A., Willis, T. D., Yu, F., Yang, H., Zeng, C., Gao, Y., Hu, H., Hu, W., Li, C., Lin, W., Liu, S., Pan, H., Tang, X., Wang, J., Wang, W., Yu, J., Zhang, B., Zhang, Q., Zhao, H., Zhao, H., Zhou, J., Gabriel, S. B., Barry, R., Blumenstiel, B., Camargo, A., Defelice, M., Faggart, M., Goyette, M., Gupta, S., Moore, J., Nguyen, H., Onofrio, R. C., Parkin, M., Roy, J., Stahl, E., Winchester, E., Ziaugra, L., Altshuler, D., Shen, Y., Yao, Z., Huang, W., Chu, X., He, Y., Jin, L., Liu, Y., Shen, Y., Sun, W., Wang, H., Wang, Y., Wang, Y., Xiong, X., Xu, L., Waye, M. M., Tsui, S. K., Xue, H., Wong, J. T. F., Galver, L. M., Fan, J. B., Gunderson, K., Murray, S. S., Oliphant, A. R., Chee, M. S., Montpetit, A., Chagnon, F., Ferretti, V., Leboeuf, M., Olivier, J. F., Phillips, M. S., Roumy, S., Sallée, C., Verner, A., Hudson, T. J., Kwok, P. Y., Cai, D., Koboldt, D. C., Miller, R. D., Pawlikowska, L., Taillon-Miller, P., Xiao, M., Tsui, L. C., Mak, W., You, Q. S., Tam, P. K., Nakamura, Y., Kawaguchi, T., Kitamoto, T., Morizono, T., Nagashima, A., Ohnishi, Y., Sekine, A., Tanaka, T., Tsunoda, T., Deloukas, P., Bird, C. P., Delgado, M., Dermitzakis, E. T., Gwilliam, R., Hunt, S., Morrison, J., Powell, D., Stranger, B. E., Whittaker, P., Bentley, D. R., Daly, M. J., De Bakker, P. I., Barrett, J., Chretien, Y. R., Maller, J., McCarroll, S., Patterson, N., Pe'Er, I., Price, A., Purcell, S., Richter, D. J., Sabeti, P., Saxena, R., Schaffner, S. F., Sham, P. C., Varilly, P., Stein, L. D., Krishnan, L., Smith, A. V., Tello-Ruiz, M. K., Thorisson, G. A., Chakravarti, A., Chen, P. E., Cutler, D. J., Kashuk, C. S., Lin, S., Abecasis, G. R., Guan, W., Li, Y., Munro, H. M., Qin, Z. S., Thomas, D. J., McVean, G., Auton, A., Bottolo, L., Cardin, N., Eyheramendy, S., Freeman, C., Marchini, J., Myers, S., Spencer, C., Stephens, M., Donnelly, P., Cardon, L. R., Clarke, G., Evans, D. M., Morris, A. P., Weir, B. S., Johnson, T. A., Mullikin, J. C., Sherry, S. T., Feolo, M., Skol, A., Zhang, H., Matsuda, I., Fukushima, Y., MacEr, D. R., Suda, E., Rotimi, C. N., Adebamowo, C. A., Ajayi, I., Aniagwu, T., Marshall, P. A., Nkwodimmah, C., Royal, C. D., Leppert, M. F., Dixon, M., Peiffer, A., Qiu, R., Kent, A., Kato, K., Niikawa, N., Adewole, I. F., Knoppers, B. M., Foster, M. W., Clayton, E. W., Watkin, J., Muzny, D., Nazareth, L., Sodergren, E., Weinstock, G. M., Yakub, I., Birren, B. W., Wilson, R. K., Fulton, L. L., Rogers, J., Burton, J., Carter, N. P., Clee, C. M., Griffiths, M., Jones, M. C., McLay, K., Plumb, R. W., Ross, M. T., Sims, S. K., Willey, D. L., Chen, Z., Han, H., Kang, L., Godbout, M., Wallenburg, J. C., L'Archevêque, P., Bellemare, G., Saeki, K., Wang, H., An, D., Fu, H., Li, Q., Wang, Z., Wang, R., Holden, A. L., Brooks, L. D., McEwen, J. E., Guyer, M. S., Wang, V. O., Peterson, J. L., Shi, M., Spiegel, J., Sung, L. M., Zacharia, L. F., Collins, F. S., Kennedy, K., Jamieson, R., and Stewart, J. (2007). A second generation human haplotype map of over 3.1 million SNPs. *Nature*, **449**(7164), 851–861.

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. (2019). MAPS: Model-based analysis of long-range chromatin interactions from PLAC-seq and HiChIP experiments. *PLOS Computational Biology*, **15**(4), e1006982.

Klein, R. J., Zeiss, C., Chew, E. Y., Tsai, J. Y., Sackler, R. S., Haynes, C., Henning, A. K., SanGiovanni, J. P., Mane, S. M., Mayne, S. T., Bracken, M. B., Ferris, F. L., Ott, J., Barnstable, C., and Hoh, J. (2005). Complement factor H polymorphism in age-related macular degeneration. *Science*, **308**(5720), 385–389.

Krismer, K., Guo, Y., and Gifford, D. K. (2020). IDR2D identifies reproducible genomic interactions. *Nucleic acids research*, **48**(6), e31.

Lajoie, B. R., Dekker, J., and Kaplan, N. (2015). The Hitchhiker's guide to Hi-C analysis: practical guidelines. *Methods (San Diego, Calif.)*, **72**, 65–75.

Lappalainen, T., Sammeth, M., Friedländer, M. R., 'T Hoen, P. A., Monlong, J., Rivas, M. A., Gonzàlez-Porta, M., Kurbatova, N., Griebel, T., Ferreira, P. G., Barann, M., Wieland, T., Greger, L., Van Iterson, M., Almlöf, J., Ribeca, P., Pulyakhina, I., Esser, D., Giger, T., Tikhonov, A., Sultan, M., Bertier, G., Macarthur, D. G., Lek, M., Lizano, E., Buermans, H. P., Padioleau, I., Schwarzmayr, T., Karlberg, O., Ongen, H., Kilpinen, H., Beltran, S., Gut, M., Kahlem, K., Amstislavskiy, V., Stegle, O., Pirinen, M., Montgomery, S. B., Donnelly, P., McCarthy, M. I., Flicek, P., Strom, T. M., Lehrach, H., Schreiber, S., Sudbrak, R., Carracedo, Á., Antonarakis, S. E., Häsler, R., Syvänen, A. C., Van Ommen, G. J., Brazma, A., Meitinger, T., Rosenstiel, P., Guigó, R., Gut, I. G., Estivill, X., and Dermitzakis, E. T. (2013). Transcriptome and genome sequencing uncovers functional variation in humans. *Nature*, **501**(7468), 506–511.

Lareau, C. A. and Aryee, M. J. (2018). Hichipper: A preprocessing pipeline for calling DNA loops from HiChIP data. *Nature Methods*, **15**(3), 155–156.

Lee, D.-I. and Roy, S. (2021). GRiNCH: simultaneous smoothing and detection of topological units of genome organization from sparse chromatin contact count matrices with matrix factorization. *Genome Biology 2021 22:1*, **22**(1), 1–31.

Lévy-Leduc, C., Delattre, M., Mary-Huard, T., and Robin, S. (2014). Two-dimensional segmentation for analyzing Hi-C data. *Bioinformatics*, **30**(17), i386.

Li, A., Yin, X., Xu, B., Wang, D., Han, J., Wei, Y., Deng, Y., Xiong, Y., and Zhang, Z. (2018a). Decoding topologically associating domains with ultra-low resolution Hi-C data by graph structural entropy. *Nature Communications 2018 9:1*, **9**(1), 1–12.

Li, L., Barth, N. K. H., Pilarsky, C., and Taher, L. (2019). Cancer Is Associated with Alterations in the Three-Dimensional Organization of the Genome. *Cancers*, **11**(12).

Li, Q., Brown, J. B., Huang, H., and Bickel, P. J. (2011). Measuring reproducibility of high-throughput experiments. *Annals of Applied Statistics*, **5**(3), 1752–1779.

Li, Y., Hu, M., and Shen, Y. (2018b). Gene regulation in the 3D genome. *Human molecular genetics*, **27**(R2), R228–R233.

Li, Z., Chen, J., Yu, H., He, L., Xu, Y., Zhang, D., Yi, Q., Li, C., Li, X., Shen, J., Song, Z., Ji, W., Wang, M., Zhou, J., Chen, B., Liu, Y., Wang, J., Wang, P., Yang, P., Wang, Q., Feng, G., Liu, B., Sun, W., Li, B., He, G., Li, W., Wan, C., Xu, Q., Li, W., Wen, Z., Liu, K., Huang, F., Ji, J., Ripke, S., Yue, W., Sullivan, P. F., O'Donovan, M. C., and Shi, Y. (2017). Genome-wide association analysis identifies 30 new susceptibility loci for schizophrenia. *Nature Genetics*, **49**(11), 1576–1583.

Lieberman-Aiden, E., Van Berkum, N. L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., Amit, I., Lajoie, B. R., Sabo, P. J., Dorschner, M. O., Sandstrom, R., Bernstein, B., Bender, M. A., Groudine, M., Gnirke, A., Stamatoyannopoulos, J., Mirny, L. A., Lander, E. S., and Dekker, J. (2009). Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*, **326**(5950), 289–293.

Little, R. J. A. and Rubin, D. B. (1987). *Statistical analysis with missing data*. Wiley, New York.

Liu, C., Rubin, D. B., and Wu, Y. N. (1998). Parameter expansion to accelerate EM: the PX-EM algorithm. *Biometrika*, **85**(4), 755–770.

Liu, J. Z., McRae, A. F., Nyholt, D. R., Medland, S. E., Wray, N. R., Brown, K. M., Hayward, N. K., Montgomery, G. W., Visscher, P. M., Martin, N. G., Macgregor, S., Mann, G. J., Kefford, R. F., Hopper, J. L., Aitken, J. F., Giles, G. G., and Armstrong, B. K. (2010). A versatile gene-based test for genome-wide association studies. *American Journal of Human Genetics*, **87**(1), 139–145.

Liu, X., Li, Y. I., and Pritchard, J. K. (2019). Trans Effects on Gene Expression Can Drive Omnigenic Inheritance. *Cell*, **177**(4), 1022–1034.e6.

Luningham, J. M., Chen, J., Tang, S., De Jager, P. L., Bennett, D. A., Buchman, A. S., and Yang, J. (2020). Bayesian Genome-wide TWAS Method to Leverage both cis- and trans-eQTL Information through Summary Statistics. *American Journal of Human Genetics*, **107**(4), 714–726.

Lupiáñez, D. G., Kraft, K., Heinrich, V., Krawitz, P., Brancati, F., Klopocki, E., Horn, D., Kayserili, H., Opitz, J. M., Laxova, R., Santos-Simarro, F., Gilbert-Dussardier, B., Wittler, L., Borschiwer, M., Haas, S. A., Osterwalder, M., Franke, M., Timmermann, B., Hecht, J., Spielmann, M., Visel, A., and Mundlos, S. (2015). Disruptions of Topological Chromatin Domains Cause Pathogenic Rewiring of Gene-Enhancer Interactions. *Cell*, **161**(5), 1012–1025.

Lyu, H., Li, L., Wu, Z., Wang, T., Zheng, J., and Wang, H. (2020). TADBD: a sensitive and fast method for detection of typologically associated domain boundaries. *Biotechniques*, **69**(1), 19–26.

Malik, L. and Patro, R. (2019). Rich chromatin structure prediction from Hi-C data. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, **16**(5), 1448–1458.

Maples, B. K., Gravel, S., Kenny, E. E., and Bustamante, C. D. (2013). RFMix: A Discriminative Modeling Approach for Rapid and Robust Local-Ancestry Inference. *American Journal of Human Genetics*, **93**(2), 278.

Maurano, M. T., Humbert, R., Rynes, E., Thurman, R. E., Haugen, E., Wang, H., Reynolds, A. P., Sandstrom, R., Qu, H., Brody, J., Shafer, A., Neri, F., Lee, K., Kutyavin, T., Stehling-Sun, S., Johnson, A. K., Canfield, T. K., Giste, E., Diegel, M., Bates, D., Hansen, R. S., Neph, S., Sabo, P. J., Heimfeld, S., Raubitschek, A., Ziegler, S., Cotsapas, C., Sotoodehnia, N., Glass, I., Sunyaev, S. R., Kaul, R., and Stamatoyannopoulos, J. A. (2012). Systematic localization of common disease-associated variation in regulatory DNA. *Science*, **337**(6099), 1190–1195.

McArthur, E. and Capra, J. A. (2021). Topologically associating domain boundaries that are stable across diverse cell types are evolutionarily constrained and enriched for heritability. *The American Journal of Human Genetics*, **108**(2), 269–283.

Mumbach, M. R., Rubin, A. J., Flynn, R. A., Dai, C., Khavari, P. A., Greenleaf, W. J., and Chang, H. Y. (2016). HiChIP: efficient and sensitive analysis of protein-directed genome architecture. *Nature Methods*, **13**(11), 919–922.

Musunuru, K., Strong, A., Frank-Kamenetsky, M., Lee, N. E., Ahfeldt, T., Sachs, K. V., Li, X., Li, H., Kuperwasser, N., Ruda, V. M., Pirruccello, J. P., Muchmore, B., Prokunina-Olsson, L., Hall, J. L., Schadt, E. E., Morales, C. R., Lund-Katz, S., Phillips, M. C., Wong, J., Cantley, W., Racie, T., Ejebe, K. G., Orho-Melander, M., Melander, O., Koteliansky, V., Fitzgerald, K., Krauss, R. M., Cowan, C. A., Kathiresan, S., and Rader, D. J. (2010). From noncoding variant to phenotype via SORT1 at the 1p13 cholesterol locus. *Nature*, **466**(7307), 714–719.

Nagpal, S., Meng, X., Epstein, M. P., Tsoi, L. C., Patrick, M., Gibson, G., De Jager, P. L., Bennett, D. A., Wingo, A. P., Wingo, T. S., and Yang, J. (2019). TIGAR: An Improved Bayesian Tool for Transcriptomic Data Imputation Enhances Gene Mapping of Complex Traits. *American Journal of Human Genetics*, **105**(2), 258–266.

Nicolae, D. L., Gamazon, E., Zhang, W., Duan, S., Dolan, M. E., and Cox, N. J. (2010). Trait-Associated SNPs Are More Likely to Be eQTLs: Annotation to Enhance Discovery from GWAS. *PLoS Genetics*, **6**(4), e1000888.

Nikpay, M., Goel, A., Won, H. H., Hall, L. M., Willenborg, C., Kanoni, S., Saleheen, D., Kyriakou, T., Nelson, C. P., CHopewell, J., Webb, T. R., Zeng, L., Dehghan, A., Alver, M., MArmasu, S., Auro, K., Bjonnes, A., Chasman, D. I., Chen, S., Ford, I., Franceschini, N., Gieger, C., Grace, C., Gustafsson, S., Huang, J., Hwang, S. J., Kim, Y. K., Kleber, M. E., Lau, K. W., Lu, X., Lu, Y., Lyytikäinen, L. P., Mihailov, E., Morrison, A. C., Pervjakova, N., Qu, L., Rose, L. M., Salfati, E., Saxena, R., Scholz, M., Smith, A. V., Tikkanen, E., Uitterlinden, A., Yang, X., Zhang, W., Zhao, W., De Andrade, M., De Vries, P. S., Van Zuydam, N. R., Anand, S. S., Bertram, L., Beutner, F., Dedoussis, G., Frossard, P., Gauguier, D., Goodall, A. H., Gottesman, O., Haber, M., Han, B. G., Huang, J., Jalilzadeh, S., Kessler, T., König, I. R., Lannfelt, L., Lieb, W., Lind, L., MLindgren, C., Lokki, M. L., Magnusson, P. K., Mallick, N. H., Mehra, N., Meitinger, T., Memon, F. U. R., Morris, A. P., Nieminen, M. S., Pedersen, N. L., Peters, A., Rallidis, L. S., Rasheed, A., Samuel, M., Shah, S. H., Sinisalo, J., EStirrups, K., Trompet, S., Wang, L., Zaman, K. S., Ardissino, D., Boerwinkle, E., Borecki, I. B., Bottinger, E. P., Buring, J. E., Chambers, J. C., Collins, R., Cupples, L., Danesh, J., Demuth, I., Elosua, R., Epstein, S. E., Esko, T., Feitosa, M. F., Franco, O. H., Franzosi, M. G., Granger, C. B., Gu, D., Gudnason, V., SHall, A., Hamsten, A., Harris, T. B., LHazen, S., Hengstenberg, C., Hofman, A., Ingelsson, E., Iribarren, C., Jukema, J. W., Karhunen, P. J., Kim, B. J., Kooner, J. S., Kullo, I. J., Lehtimäki, T., Loos, R. J., Melander, O., Metspalu, A., März, W., Palmer, C. N., Perola, M., Quertermous, T., Rader, D. J., Ridker, P. M., Ripatti, S., Roberts, R., Salomaa, V., Sanghera, D. K., Schwartz, S. M., Seedorf, U., Stewart, A. F., Stott, D. J., Thiery, J., Zalloua, P. A., O'Donnell, C. J., Reilly, M. P., Assimes, T. L., Thompson, J. R., Erdmann, J., Clarke, R., Watkins, H., Kathiresan, S., McPherson, R., Deloukas, P., Schunkert, H., Samani, N. J., and Farrall, M. (2015). A comprehensive 1000 Genomes-based genome-wide association meta-analysis of coronary artery disease. *Nature Genetics*, **47**(10), 1121–1130.

Nora, E. P., Lajoie, B. R., Schulz, E. G., Giorgetti, L., Okamoto, I., Servant, N., Piolot, T., van Berkum, N. L., Meisig, J., Sedat, J., Gribnau, J., Barillot, E., Blüthgen, N., Dekker, J., and Heard, E. (2012). Spatial partitioning of the regulatory landscape of the X-inactivation centre. *Nature*, **485**(7398), 381–385.

Norton, H. K., Emerson, D. J., Huang, H., Kim, J., Titus, K. R., Gu, S., Bassett, D. S., and Phillips-Cremins, J. E. (2018). Detecting hierarchical genome folding with network modularity. *Nature Methods*, **15**(2), 119–122.

Nuotio, J., Oikonen, M., Magnussen, C. G., Jokinen, E., Laitinen, T., Hutri Kähönen, N., Kähönen, M., Lehtimäki, T., Taittonen, L., Tossavainen, P., Jula, A., Loo, B. M., Viikari, J. S., Juonala, M., and Raitakari, O. T. (2014). Cardiovascular risk factors in 2011 and secular trends since 2007: The Cardiovascular Risk in Young Finns Study. *Scandinavian Journal of Public Health*, **42**(7), 563–571.

Oluwadare, O. and Cheng, J. (2017). ClusterTAD: an unsupervised machine learning approach to detecting topologically associated domains of chromosomes from Hi-C data. *BMC Bioinformatics*, **18**(1), 1–14.

Ong, C. T. and Corces, V. G. (2014). CTCF: An architectural protein bridging genome topology and function. *Nature Reviews Genetics*, **15**(4), 234–246.

Paisley, J., Blei, D., and Jordan, M. (2012). Variational Bayesian Inference with Stochastic Search. *Proceedings of the 29th International Conference on Machine Learning, ICML 2012*, **2**, 1367–1374.

Park, Y., Sarkar, A., Bhutani, K., and Kellis, M. (2017). Multi-tissue polygenic models for transcriptome-wide association studies. *bioRxiv*, page doi:10.1101/107623.

Pfitzner, D., Leibbrandt, R., and Powers, D. (2009). Characterization and evaluation of similarity measures for pairs of clusterings. *Knowledge and Information Systems*, **19**, 361–394.

Phanstiel, D. H., Boyle, A. P., Heidari, N., and Snyder, M. P. (2015). Mango: A bias-correcting ChIA-PET analysis pipeline. *Bioinformatics*, **31**(19), 3092–3098.

Pierce, B. L. and Burgess, S. (2013). Efficient design for mendelian randomization studies: Subsample and 2-sample instrumental variable estimators. *American Journal of Epidemiology*, **178**(7), 1177–1184.

Pinoli, P., Stamoulakatou, E., Nguyen, A.-P., Martínez, M. R., and Ceri, S. (2020). Pan-cancer analysis of somatic mutations and epigenetic alterations in insulated neighbourhood boundaries. *PLOS ONE*, **15**(1), e0227180.

Porcu, E., Rüeger, S., Lepik, K., Agbessi, M., Ahsan, H., Alves, I., Andiappan, A., Arindrarto, W., Awadalla, P., Battle, A., Beutner, F., Jan Bonder, M., Boomsma, D., Christiansen, M., Claringbould, A., Deelen, P., Esko, T., Favé, M.-J., Franke, L., Frayling, T., Gharib, S. A., Gibson, G., Heijmans, B. T., Hemani, G., Jansen, R., Kähönen, M., Kalnapenkis, A., Kasela, S., Kettunen, J., Kim, Y., Kirsten, H., Kovacs, P., Krohn, K., Kronberg-Guzman, J., Kukushkina, V., Lee, B., Lehtimäki, T., Loeffler, M., Marigorta, U. M., Mei, H., Milani, L., Montgomery, G. W., Müller-Nurasyid, M., Nauck, M., Nivard, M., Penninx, B., Perola, M., Pervjakova, N., Pierce, B. L., Powell, J., Prokisch, H., Psaty, B. M., Raitakari, O. T., Ripatti, S., Rotzschke, O., Saha, A., Scholz, M., Schramm, K., Seppälä, I., Slagboom, E. P., Stehouwer, C. D. A., Stumvoll, M., Sullivan, P., t Hoen, P. A. C., Teumer, A., Thiery, J., Tong, L., Tönjes, A., van Dongen, J., van Iterson, M., van Meurs, J., Veldink, J. H., Verlouw, J., Visscher, P. M., Völker, U., Võsa, U., Westra, H.-J., Wijmenga, C., Yaghootkar, H., Yang, J., Zeng, B., Zhang, F., Arindrarto, W., Beekman, M., Boomsma, D. I., Bot, J., Deelen, J., Deelen, P., Franke, L., Heijmans, B. T., ’t Hoen, P. A. C., Hofman, B. A., Hottenga, J. J., Isaacs, A., Bonder, M. J., Jhamai, P. M., Jansen, R., Kielbasa, S. M., Lakenberg, N., Luijk, R., Mei, H., Moed, M., Nooren, I., Pool, R., Schalkwijk, C. G., Slagboom, P. E., Stehouwer, C. D. A., Suchiman, H. E. D., Swertz, M. A., Tigchelaar, E. F., Uitterlinden, A. G., van den Berg, L. H., van der Breggen, R., van der Kallen, C. J. H., van Dijk, F., van Dongen, J., van Duijn, C. M., van Galen, M., van Greevenbroek, M. M. J., van Heemst, D., van Iterson, M., van Meurs, J., van Rooij, J., van’t Hof, P., van Zwet, E. W., Vermaat, M., Veldink, J. H., Verbiest, M., Verkerk, M., Wijmenga, C., Zhernakova, D. V., Zhernakova, S., Santoni, F. A., Reymond, A., Kutalik, Z., eQTLGen Consortium, and Consortium, B. (2019). Mendelian randomization integrating GWAS and eQTL data reveals genetic determinants of complex and clinical traits. *Nature Communications*, **10**(1), 3300.

Raitakari, O. T., Juonala, M., Rönnemaa, T., Keltikangas-Järvinen, L., Räsänen, L., Pietikäinen, M., Hutri-Kähönen, N., Taittonen, L., Jokinen, E., Marniemi, J., Jula, A., Telama, R., Kähönen, M., Lehtimäki, T., Åkerblom, H. K., and Viikari, J. S. (2008). Cohort profile: The cardiovascular risk in young Finns study. *International Journal of Epidemiology*, **37**(6), 1220–1226.

Ramírez, F., Bhardwaj, V., Arrigoni, L., Lam, K. C., Grüning, B. A., Villaveces, J., Habermann, B., Akhtar, A., and Manke, T. (2018). High-resolution TADs reveal DNA sequences underlying genome organization in flies. *Nature Communications*, **9**(1), 1–15.

Ranganath, R., Gerrish, S., and Blei, D. (2014). Black Box Variational Inference. In *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics*, pages 814–822. PMLR.

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

Ren, B. and Dixon, J. R. (2015). A CRISPR Connection between Chromatin Topology and Genetic Disorders. *Cell*, **161**(5), 955–957.

Robinson, G. K. (1991). That BLUP is a good thing: The estimation of random effects. *Statistical Science*, **6**(1), 15–32.

Ron, G., Globerson, Y., Moran, D., and Kaplan, T. (2017). Promoter-enhancer interactions identified from Hi-C data using probabilistic models and hierarchical topological domains. *Nature Communications*, **8**(1), 1–12.

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, **20**(C), 53–65.

Sabatti, C., Service, S. K., Hartikainen, A. L., Pouta, A., Ripatti, S., Brodsky, J., Jones, C. G., Zaitlen, N. A., Varilo, T., Kaakinen, M., Sovio, U., Ruokonen, A., Laitinen, J., Jakkula, E., Coin, L., Hoggart, C., Collins, A., Turunen, H., Gabriel, S., Elliot, P., McCarthy, M. I., Daly, M. J., Järvelin, M. R., Freimer, N. B., and Peltonen, L. (2009). Genome-wide association analysis of metabolic traits in a birth cohort from a founder population. *Nature Genetics*, **41**(1), 35–46.

Sauria, M. E. and Taylor, J. (2017). QuASAR: Quality Assessment of Spatial Arrangement Reproducibility in Hi-C Data. *bioRxiv*, page doi:10.1101/204438.

Schaffner, S. F., Foo, C., Gabriel, S., Reich, D., Daly, M. J., and Altshuler, D. (2005). Calibrating a coalescent simulation of human genome sequence variation. *Genome Research*, **15**(11), 1576–1583.

Schaub, M. A., Boyle, A. P., Kundaje, A., Batzoglou, S., and Snyder, M. (2012). Linking disease associations with regulatory information in the human genome. *Genome Research*, **22**(9), 1748–1759.

Schmitt, A. D., Hu, M., and Ren, B. (2016). Genome-wide mapping and analysis of chromosome architecture. *Nature Reviews Molecular Cell Biology*, **17**(12), 743–755.

Schoenfelder, S. and Fraser, P. (2019). Long-range enhancerpromoter contacts in gene expression control. *Nature Reviews Genetics*, **20**(8), 437–455.

Serra, F., Baù, D., Goodstadt, M., Castillo, D., Filion, G. J., and Marti-Renom, M. A. (2017). Automatic analysis and 3D-modelling of Hi-C data using TADbit reveals structural features of the fly chromatin colors. *PLOS Computational Biology*, **13**(7), e1005665.

Shi, X., Chai, X., Yang, Y., Cheng, Q., Jiao, Y., Chen, H., Huang, J., Yang, C., and Liu, J. (2020). A tissue-specific collaborative mixed model for jointly analyzing multiple tissues in transcriptome-wide association studies. *Nucleic Acids Research*, **48**(19), e109.

Shin, H., Shi, Y., Dai, C., Tjong, H., Gong, K., Alber, F., and Zhou, X. J. (2016). TopDom: an efficient and deterministic method for identifying topological domains in genomes. *Nucleic Acids Research*, **44**(7), e70.

Sikorska, N. and Sexton, T. (2020). Defining Functionally Relevant Spatial Chromatin Domains: It is a TAD Complicated. *Journal of Molecular Biology*, **432**(3), 653–664.

Simonis, M., Klous, P., Splinter, E., Moshkin, Y., Willemsen, R., De Wit, E., Van Steensel, B., and De Laat, W. (2006). Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture-on-chip (4C). *Nature Genetics*, **38**(11), 1348–1354.

Smith, G. D. and Ebrahim, S. (2004). Mendelian randomization: Prospects, potentials, and limitations. *International Journal of Epidemiology*, **33**(1), 30–42.

Soler-Vila, P., Cuscó, P., Farabella, I., DiStefano, M., and Marti-Renom, M. (2020). Hierarchical chromatin organization detected by TADpole. *Nucleic Acids Research*, **48**(7), e39–e39.

Song, M., Pebworth, M. P., Yang, X., Abnousi, A., Fan, C., Wen, J., Rosen, J. D., Choudhary, M. N., Cui, X., Jones, I. R., Bergenholtz, S., Eze, U. C., Juric, I., Li, B., Maliskova, L., Lee, J., Liu, W., Pollen, A. A., Li, Y., Wang, T., Hu, M., Kriegstein, A. R., and Shen, Y. (2020). Cell-type-specific 3D epigenomes in the developing human cortex. *Nature*, **587**(7835), 644–649.

Stilianoudakis, S. C. and Dozmorov, M. G. (2020). preciseTAD: A machine learning framework for precise 3D domain boundary prediction at base-level resolution. *bioRxiv*, page doi:10.1101/2020.09.03.282186.

Sun, R. and Lin, X. (2017). Set-Based Tests for Genetic Association Using the Generalized Berk-Jones Statistic. *arXiv*, page arXiv:1710.02469.

Tang, Z., Luo, O. J., Li, X., Zheng, M., Zhu, J. J., Szalaj, P., Trzaskoma, P., Magalska, A., Wlodarczyk, J., Ruszczycki, B., Michalski, P., Piecuch, E., Wang, P., Wang, D., Tian, S. Z., Penrad-Mobayed, M., Sachs, L. M., Ruan, X., Wei, C. L., Liu, E. T., Wilczynski, G. M., Plewczynski, D., Li, G., and Ruan, Y. (2015). CTCF-Mediated Human 3D Genome Architecture Reveals Chromatin Topology for Transcription. *Cell*, **163**(7), 1611–1627.

Taylor, K., Davey Smith, G., Relton, C. L., Gaunt, T. R., and Richardson, T. G. (2019). Prioritizing putative influential genes in cardiovascular disease susceptibility by applying tissue-specific Mendelian randomization. *Genome Medicine*, **11**(1), 6.

The 1000 Genomes Project Consortium Institution/Organization, Auton, A., Abecasis, G. R., Altshuler, D. M., Durbin, R. M., Abecasis, G. R., Bentley, D. R., Chakravarti, A., Clark, A. G., Donnelly, P., Eichler, E. E., Flicek, P., Gabriel, S. B., Gibbs, R. A., Green, E. D., Hurles, M. E., Knoppers, B. M., Korbel, J. O., Lander, E. S., Lee, C., Lehrach, H., Mardis, E. R., Marth, G. T., McVean, G. A., Nickerson, D. A., Schmidt, J. P., Sherry, S. T., Wang, J., Wilson, R. K., Gibbs, R. A., Boerwinkle, E., Doddapaneni, H., Han, Y., Korchina, V., Kovar, C., Lee, S., Muzny, D., Reid, J. G., Zhu, Y., Wang, J., Chang, Y., Feng, Q., Fang, X., Guo, X., Jian, M., Jiang, H., Jin, X., Lan, T., Li, G., Li, J., Li, Y., Liu, S., Liu, X., Lu, Y., Ma, X., Tang, M., Wang, B., Wang, G., Wu, H., Wu, R., Xu, X., Yin, Y., Zhang, D., Zhang, W., Zhao, J., Zhao, M., Zheng, X., Lander, E. S., Altshuler, D. M., Gabriel, S. B., Gupta, N., Gharani, N., Toji, L. H., Gerry, N. P., Resch, A. M., Flicek, P., Barker, J., Clarke, L., Gil, L., Hunt, S. E., Kelman, G., Kulesha, E., Leinonen, R., McLaren, W. M., Radhakrishnan, R., Roa, A., Smirnov, D., Smith, R. E., Streeter, I., Thormann, A., Toneva, I., Vaughan, B., Zheng-Bradley, X., Bentley, D. R., Grocock, R., Humphray, S., James, T., Kingsbury, Z., Lehrach, H., Sudbrak, R., Albrecht, M. W., Amstislavskiy, V. S., Borodina, T. A., Lienhard, M., Mertes, F., Sultan, M., Timmermann, B., Yaspo, M.-L., Mardis, E. R., Wilson, R. K., Fulton, L., Fulton, R., Sherry, S. T., Ananiev, V., Belaia, Z., Beloslyudtsev, D., Bouk, N., Chen, C., Church, D., Cohen, R., Cook, C., Garner, J., Hefferon, T., Kimelman, M., Liu, C., Lopez, J., Meric, P., O'Sullivan, C., Ostapchuk, Y., Phan, L., Ponomarov, S., Schneider, V., Shekhtman, E., Sirotkin, K., Slotta, D., Zhang, H., McVean, G. A., Durbin, R. M., Balasubramaniam, S., Burton, J., Danecek, P., Keane, T. M., Kolb-Kokocinski, A., McCarthy, S., Stalker, J., Quail, M., Schmidt, J. P., Davies, C. J., Gollub, J., Webster, T., Wong, B., Zhan, Y., Auton, A., Campbell, C. L., Kong, Y., Marcketta, A., Gibbs, R. A., Yu, F., Antunes, L., Bainbridge, M., Muzny, D., Sabo, A., Huang, Z., Wang, J., Coin, L. J. M., Fang, L., Guo, X., Jin, X., Li, G., Li, Q., Li, Y., Li, Z., Lin, H., Liu, B., Luo, R., Shao, H., Xie, Y., Ye, C., Yu, C., Zhang, F., Zheng, H., Zhu, H., Alkan, C., Dal, E., Kahveci, F., Marth, G. T., Garrison, E. P., Kural, D., Lee, W.-P., Fung Leong, W., Stromberg, M., Ward, A. N., Wu, J., Zhang, M., Daly, M. J., DePristo, M. A., Handsaker, R. E., Altshuler, D. M., Banks, E., Bhatia, G., del Angel, G., Gabriel, S. B., Genovese, G., Gupta, N., Li, H., Kashin, S., Lander, E. S., McCarroll, S. A., Nemesh, J. C., Poplin, R. E., Yoon, S. C., Lihm, J., Makarov, V., Clark, A. G., Gottipati, S., Keinan, A., Rodriguez-Flores, J. L., Korbel, J. O., Rausch, T., Fritz, M. H., Stütz, A. M., Flicek, P., Beal, K., Clarke, L., Datta, A., Herrero, J., McLaren, W. M., Ritchie, G. R. S., Smith, R. E., Zerbino, D., Zheng-Bradley, X., Sabeti, P. C., Shlyakhter, I., Schaffner, S. F., Vitti, J., Cooper, D. N., Ball, E. V., Stenson, P. D., Bentley, D. R., Barnes, B., Bauer, M., Keira Cheetham, R., Cox, A., Eberle, M., Humphray, S., Kahn, S., Murray, L., Peden, J., Shaw, R., Kenny, E. E., Batzer, M. A., Konkel, M. K., Walker, J. A., MacArthur, D. G., Lek, M., Sudbrak, R., Amstislavskiy, V. S., Herwig, R., Mardis, E. R., Ding, L., Koboldt, D. C., Larson, D., Ye, K., and Gravel, S. (2015). A global reference for human genetic variation. *Nature*, **526**(7571), 68–74.

Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, **58**(1), 267–288.

Ursu, O., Boley, N., Taranova, M., Wang, Y. X., Yardimci, G. G., Noble, W. S., and Kundaje, A. (2018). GenomeDISCO: A concordance score for chromosome conformation capture experiments using random walks on contact map graphs. *Bioinformatics*, **34**(16), 2701–2707.

Valton, A. L. and Dekker, J. (2016). TAD disruption as oncogenic driver. *Current Opinion in Genetics & Development*, **36**, 34–40.

Verbanck, M., Chen, C. Y., Neale, B., and Do, R. (2018). Detection of widespread horizontal pleiotropy in causal relationships inferred from Mendelian randomization between complex traits and diseases. *Nature Genetics*, **50**(5), 693–698.

Wang, S., Su, J. H., Beliveau, B. J., Bintu, B., Moffitt, J. R., Wu, C. T., and Zhuang, X. (2016). Spatial organization of chromatin domains and compartments in single chromosomes. *Science*, **353**(6299), 598–602.

Wang, X.-T., Cui, W., and Peng, C. (2017). HiTAD: detecting the structural and functional hierarchies of topologically associating domains from chromatin interactions. *Nucleic Acids Research*, **45**(19), e163.

Wang, Y., Li, Y., Gao, J., and Zhang, M. Q. (2015). A novel method to identify topological domains using Hi-C data. *Quantitative Biology 2015 3:2*, **3**(2), 81–89.

Weinreb, C. and Raphael, B. J. (2016). Identification of hierarchical chromatin domains. *Bioinformatics*, **32**(11), 1601–1609.

West, A. G. and Fraser, P. (2005). Remote control of gene transcription. *Human Molecular Genetics*, **14**(SPEC. ISS. 1).

Wheeler, H. E., Aquino-Michaels, K., Gamazon, E. R., Trubetskoy, V. V., Dolan, M. E., Huang, R. S., Cox, N. J., and Im, H. K. (2014). Poly-omic prediction of complex traits: OmicKriging. *Genetic Epidemiology*, **38**(5), 402–415.

Wright, F. A., Sullivan, P. F., Brooks, A. I., Zou, F., Sun, W., Xia, K., Madar, V., Jansen, R., Chung, W., Zhou, Y. H., Abdellaoui, A., Batista, S., Butler, C., Chen, G., Chen, T. H., D'Ambrosio, D., Gallins, P., Ha, M. J., Hottenga, J. J., Huang, S., Kattenberg, M., Kochar, J., Middeldorp, C. M., Qu, A., Shabalin, A., Tischfield, J., Todd, L., Tzeng, J. Y., Van Grootheest, G., Vink, J. M., Wang, Q., Wang, W., Wang, W., Willemsen, G., Smit, J. H., De Geus, E. J., Yin, Z., Penninx, B. W., and Boomsma, D. I. (2014). Heritability and genomics of gene expression in peripheral blood. *Nature Genetics*, **46**(5), 430–437.

Wu, M. C., Lee, S., Cai, T., Li, Y., Boehnke, M., and Lin, X. (2011). Rare-variant association testing for sequencing data with the sequence kernel association test. *American Journal of Human Genetics*, **89**(1), 82–93.

Xing, H., Wu, Y., Zhang, M. Q., and Chen, Y. (2021). Deciphering hierarchical organization of topologically associated domains through change-point testing. *BMC Bioinformatics 2021 22:1*, **22**(1), 1–23.

Yaffe, E. and Tanay, A. (2011). Probabilistic modeling of Hi-C contact maps eliminates systematic biases to characterize global chromosomal architecture. *Nature Genetics*, **43**(11), 1059–1065.

Yamada, T., Yang, Y., Valnegri, P., Juric, I., Abnousi, A., Markwalter, K. H., Guthrie, A. N., Godec, A., Oldenborg, A., Hu, M., Holy, T. E., and Bonni, A. (2019). Sensory experience remodels genome architecture in neural circuit to drive motor learning. *Nature*, **569**(7758), 708–713.

Yan, K. K., Yardlmcl, G. G., Yan, C., Noble, W. S., and Gerstein, M. (2017a). HiC-spector: A matrix library for spectral and reproducibility analysis of Hi-C contact maps. *Bioinformatics*, **33**(14), 2199–2201.

Yan, K.-K., Lou, S., and Gerstein, M. (2017b). MrTADFinder: A network modularity based approach to identify topologically associating domains in multiple resolutions. *PLOS Computational Biology*, **13**(7), e1005647.

Yang, C., Wan, X., Lin, X., Chen, M., Zhou, X., and Liu, J. (2019). CoMM: a collaborative mixed model to dissecting genetic contributions to complex traits by leveraging regulatory information. *Bioinformatics*, **35**(10), 1644–1652.

Yang, J., Lee, S. H., Goddard, M. E., and Visscher, P. M. (2011). GCTA: A tool for genome-wide complex trait analysis. *American Journal of Human Genetics*, **88**(1), 76–82.

Yang, J., Fritsche, L. G., Zhou, X., and Abecasis, G. (2017a). A Scalable Bayesian Method for Integrating Functional Information in Genome-wide Association Studies. *American Journal of Human Genetics*, **101**(3), 404–416.

Yang, T., Zhang, F., Yardmci, G. G., Song, F., Hardison, R. C., Noble, W. S., Yue, F., and Li, Q. (2017b). HiCRep: assessing the reproducibility of Hi-C data using a stratum-adjusted correlation coefficient. *Genome Research*, **27**(11), 1939–1949.

Yang, Y., Shi, X., Jiao, Y., Huang, J., Chen, M., Zhou, X., Sun, L., Lin, X., Yang, C., and Liu, J. (2020). CoMM-S2: A collaborative mixed model using summary statistics in transcriptome-wide association studies. *Bioinformatics*, **36**(7), 2009–2016.

Yardimci, G. G., Ozadam, H., Sauria, M. E. G., Ursu, O., Yan, K.-K., Yang, T., Chakraborty, A., Kaul, A., Lajoie, B. R., Song, F., Zhan, Y., Ay, F., Gerstein, M., Kundaje, A., Li, Q., Taylor, J., Yue, F., Dekker, J., and Noble, W. S. (2019). Measuring the reproducibility and quality of Hi-C data. *Genome Biology*, **20**(1), 57.

Yu, W., He, B., and Tan, K. (2017). Identifying topologically associating domains and subdomains by Gaussian Mixture model And Proportion test. *Nature Communications 2017 8:1*, **8**(1), 1–9.

Yuan, Z., Zhu, H., Zeng, P., Yang, S., Sun, S., Yang, C., Liu, J., and Zhou, X. (2020). Testing and controlling for horizontal pleiotropy with probabilistic Mendelian randomization in transcriptome-wide association studies. *Nature Communications*, **11**(1), 3861.

Zeng, P. and Zhou, X. (2017). Non-parametric genetic prediction of complex traits with latent Dirichlet process regression models. *Nature Communications*, **8**(1), 456.

Zhan, Y., Mariani, L., Barozzi, I., Schulz, E. G., Blüthgen, N., Stadler, M., Tiana, G., and Giorgetti, L. (2017). Reciprocal insulation analysis of Hi-C data shows that TADs represent a functionally but not structurally privileged scale in the hierarchical folding of chromosomes. *Genome Research*, **27**(3), 479–490.

Zhang, Y., Quick, C., Yu, K., Barbeira, A., Luca, F., Pique-Regi, R., Kyung Im, H., and Wen, X. (2020). PTWAS: Investigating tissue-relevant causal molecular mechanisms of complex traits using probabilistic TWAS analysis. *Genome Biology*, **21**(1), 232.

Zhao, W., Rasheed, A., Tikkanen, E., Lee, J. J., Butterworth, A. S., Howson, J. M., Assimes, T. L., Chowdhury, R., Orho-Melander, M., Damrauer, S., Small, A., Asma, S., Imamura, M., Yamauch, T., Chambers, J. C., Chen, P., Sapkota, B. R., Shah, N., Jabeen, S., Surendran, P., Lu, Y., Zhang, W., Imran, A., Abbas, S., Majeed, F., Trindade, K., Qamar, N., Mallick, N. H., Yaqoob, Z., Saghir, T., Hasan Rizvi, S. N., Memon, A., Rasheed, S. Z., Memon, F. U. R., Mehmood, K., Ahmed, N., Hussain Qureshi, I., Tanveer-Us-Salam, Iqbal, W., Malik, U., Mehra, N., Kuo, J. Z., Sheu, W. H., Guo, X., Hsiung, C. A., Juang, J. M. J., Taylor, K. D., Hung, Y. J., Lee, W. J., Quertermous, T., Lee, I. T., Hsu, C. C., Bottinger, E. P., Ralhan, S., Teo, Y. Y., Wang, T. D., Alam, D. S., Di Angelantonio, E., Epstein, S., Nielsen, S. F., Nordestgaard, B. G., Tybjaerg-Hansen, A., Young, R., Benn, M., Frikke-Schmidt, R., Kamstrup, P. R., Jukema, J. W., Sattar, N., Smit, R., Chung, R. H., Liang, K. W., Anand, S., Sanghera, D. K., Ripatti, S., Loos, R. J., Kooner, J. S., Tai, E. S., Rotter, J. I., Ida Chen, Y. D., Frossard, P., Maeda, S., Kadowaki, T., Reilly, M., Pare, G., Melander, O., Salomaa, V., Rader, D. J., Danesh, J., Voight, B. F., and Saleheen, D. (2017). Identification of new susceptibility loci for type 2 diabetes and shared etiological pathways with coronary heart disease. *Nature Genetics*, **49**(10), 1450–1457.

Zhou, D., Jiang, Y., Zhong, X., Cox, N. J., Liu, C., and Gamazon, E. R. (2020). A unified framework for joint-tissue transcriptome-wide association and Mendelian randomization analysis. *Nature Genetics*, **52**(11), 1239–1246.

Zhou, X., Carbonetto, P., and Stephens, M. (2013). Polygenic Modeling with Bayesian Sparse Linear Mixed Models. *PLoS Genetics*, **9**(2), 1003264.

Zhu, H. and Zhou, X. (2020). Transcriptome-wide association studies: a view from Mendelian randomization. *Quantitative Biology*, **9**(2), 1–15.

Zhu, Z., Zhang, F., Hu, H., Bakshi, A., Robinson, M. R., Powell, J. E., Montgomery, G. W., Goddard, M. E., Wray, N. R., Visscher, P. M., and Yang, J. (2016). Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. *Nature Genetics*, **48**(5), 481–487.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the Elastic Net. *Journal of the Royal Statistical Society. Series B*, **67**, 301–320.

Zufferey, M., Tavernari, D., Oricchio, E., and Ciriello, G. (2018). Comparison of computational methods for the identification of topologically associating domains. *Genome Biology 2018 19:1*, **19**(1), 1–18.