GENETIC DISSECTION OF COMPLEX DISEASES AND TRAITS WITH OMICS DATA

Le Huang

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 Curriculum in Bioinformatics and Computational Biology.

Chapel Hill
2024

Approved by:

Yun Li

Laura Raffield

Karen L. Mohlke

Di Wu

Ming Hu

©2024
Le Huang
ALL RIGHTS RESERVED

ABSTRACT

Le Huang: Genetic dissection of complex diseases and traits with omics data
(Under the direction of Yun Li and Laura Raffield)

The human genome consists of approximately three billion DNA base pairs (Rhie et al., 2023), holding the secrets of our evolutionary past, our current biological processes, and the potential determinants of our future health. However, interpreting this genetic data, such as understanding its relationship with diseases, environmental interactions, and inheritance patterns, remains challenging. Consequently, various advanced approaches, such as next-generation high-throughput sequencing (NGS) and liquid chromatography–mass spectrometry (LC-MS), have been developed to allow for the profiling of various small molecules, including RNAs, proteins, and metabolites.

My work integrates multiple topics within the realms of genetics and multi-omics, which are fundamental to our understanding of health and disease. I explore the intricate roles of genetic variants in inheritance differ between monogenic traits, influenced by single-gene mutations, and complex traits, which are shaped by multiple genes and environmental factors.

In my research, I introduce TOP-LD, a new LD resource we provided by using high-coverage whole-genome sequencing (WGS) data from the NHLBI Trans-Omics for Precision Medicine (TOPMed) Program. My TOP-LD significantly outperforms existing LD resources, offering a more comprehensive understanding of the correlation structure among genetic variation, particularly for rare genetic variants and structural variants in diverse populations.

Moving to 3D genomics, I analyzed data from high-throughput chromosome conformation capture technologies like Hi-C, HiChIP, and PLAC-seq, which is important to understand interactions between regulatory elements. Specifically, I evaluated the performance of various deep learning models developed for Hi-C data when applied to HiChIP and PLAC-seq (HP) data, provid-

ing practical guidelines regarding data enhancement for HP data and underscoring the potential of computational methods in data enhancement and thus saving experimental costs.

Lastly, I explore metabolomics through metabolome-wide association study (MWAS) using the UK Biobank (UKBB) data. I developed a method to predict metabolites from genome-wide genetic data, a novel approach given that metabolites do not have corresponding local genetic variants like genes or proteins do. Finally, I tested associations between genetically imputed metabolites and clinical outcomes.

Overall, my dissertation presents the intricate web of genetic variation, 3D genomic organization, and metabolomics, contributing to our understanding of the molecular mechanisms essential for human health.

To my beloved parents, your love and selfless support are my eternal backbone. To my dear boyfriend, Haidong, thank you for every moment spent together.

ACKNOWLEDGEMENTS

I would want to thank everyone who has supported and encouraged me during my academic journey. First and foremost, I want to thank Yun and Laura, my two supervisors. Throughout my research career, they have been both supervisors and role models. Their demanding academic attitude, philanthropic mindset, and incisive thoughts have always been the driving force behind my continued growth and development.

Second, I would like to thank the committee chair Karen, committee members Ming and Di from the bottom of my heart. I really learned a lot from you and especially every useful suggestions to help me PhD career successful. Karen gave me a lot of suggestions and comments for my oral exam, and I really knew that I have a lot to learn for human genetics, which are really helpful for laying great foundation for my internship research. Ming also advised two of my projects. I really enjoy talking with him about the research project, and he also shared that he has the great habbit for marathon, which intrigue me to exercise a lot.

Additionally, I would like to thank my summer internship mentors Judong, Wujuan, and Song at Merck. It was an unforgettable great internship journey. I did a very interesting project in your team under your great supervision. Getting along with you taught me the value of teamwork, and my professional abilities and interpersonal communication have much increased.

I would also like to thank my parents. Your unwavering love has shaped who I am today. Your trust and support in me provide the most firm foundation for me to continue forward. You always support and strengthen me when I face obstacles and challenges.

Similarly, I cannot forget to acknowledge boyfriend Haidong and my friends Keke, Luchao. I am fortunate to meet you all during the final years. Your understanding, patience, and company are the most important aspects of my life experience. You are the ones who allow me to find moments of peace and enjoyment. You are my most important support network.

In addition, I would like to thank all of the labmates in the Li Lab and Raffield Lab. Everyone provided significant assistance, and I truly appreciate every bit of help I received. In this academic family, we learn and grow together. Each individual's distinct viewpoint and information sharing were critical to my abilities to complete my studies and research work. The days and nights we spent together, as well as the scientific research dreams we pursued, will live on in my memory forever.

I want to express my gratitude to everyone in the online exercise chat group. I joined this online community to connect with others about fitness. The group members are incredibly humorous and always make me laugh. We share many common interests and support each other in our journeys toward becoming healthier and better individuals.

Lastly, I am grateful to Genshin Impact, the captivating game that served as wonderful emotional therapy throughout my PhD journey. I particularly enjoyed exploring the Sumeru region, where the characters, much like us, grapple with academia, research funding, and publications. This game not only taught me valuable social skills and humor but also how to address real-life problems effectively. As a result, I became more outgoing and happier during my PhD.

I thank you all again. This degree belongs not only to me, but to all of you. We built this together, and I will always value this experience and everything you have given me.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

AFR	African
ALT	Alternative Alleles
ATAC-seq	Assay for Transposase-Accessible Chromatin with sequencing
BioMe	BioMe Biobank
bp	base pair
CADD	Combined Annotation Dependent Depletion
cCREs	*cis*-regulatory regions
ChIP-seq	Chromatin Immunoprecipitation sequencing
CV	cross validation
CTCF	Transcriptional repressor CTCF or CCCTC-binding factor.
DEL	Deletion
DUP	Duplicated Insertion
E-P	Enhancer-Promoter
EAS	East Asian
EN	Excitatory Neurons
ENCODE	Encyclopedia of DNA Elements
EPACTS	Efficient and Parallelizable Association Container Toolbox
eQTL	expression Quantitative Trait Loci
EUR	European
FDR	False Discovery Rate
FINEMAP	efficient variable selection using summary data
from genome-wide association studies	
FIREs	Frequently Interacting Regions
GAN	generative adversarial networks
GM12878	A cell line from a female donor with Northern and Western European lineage.
GWAS	Genome-wide Association Study

H^2	Heritability score
Hi-C	High-throughput genomic and epigenomic technique to capture chromatin conformation
H3K4me3	A modification of chromatin that entails appending three methyl groups onto the fourth lysine residue found on the histone H3 protein.
HiChIP	Variant of Hi-C
HP	HiChIP and PLAC-seq
ICD-10	the International Classification of Diseases, Tenth Revision
Indel	Duplicated Insertion or Deletion
IN	Interneurons
INV	Inversion Indel
IPC	Intermediate Progenitor Cells
JHS	Jackson Heart Study
Kb	Kilobase
KR	Knight-Ruiz Matrix Balancing
LC-MS	Liquid Chromatography–Mass Spectrometr
LD	Linkage Disequilibrium
MAF	Minor Allele Frequency
Med13l	Mediator Complex Subunit 13L
MESA	Multi-Ethnic Study of Atherosclerosis
mESC	Mouse Embryonic Stem Cells
mGWAS	metabolite Genome-Wide Association Studies
mQTL	metabolite Quantitative Trait Loci
Mtnr1a	Melatonin Receptor 1A
MWAS	Metabolome-wide Association Study
NGS	Next-Generation high-throughput Sequencing
p	p-value

PC	Principal Components
PEA	Proximity Extension Assay
PLAC-seq	Variant of Hi-C
PRS	Polygenic Risk Score
QC	Quality Control
r	Correlation Coefficient
R^2	R-squared
REF	Reference Alleles
RG	Radial Glia
RNA-seq	RNA sequencing
SAS	South Asian
SEN	Sensitivity
Smc1a	Structural Maintenance Of Chromosomes 1A, a protein coding gene
SNP	Single Nucleotide Polymorphism
SNR	Signal to Noise Ratio
SNV	Single Nucleotide Variant
SV	Structural Variant
TAD	Topologically Associated Domains
TOPMed	the NHLBI Trans-Omics for Precision Medicine program
TWAS	Transcriptome-wide Association Study
UKBB	UK Biobank
VAE	Variational Autoencoder
w/o	without
WGS	Whole-Genome Sequencing

CHAPTER 1: INTRODUCTION

1.1 Omics

The human genome consists of approximately three billion DNA base pairs (Rhie et al., 2023). It contains information related to physiological processes and diseases. However, how to reveal the functions of genetic data, including relationships with diseases, environmental interactions, and inheritance patterns, is challenging. However, how to reveal the functions of genetic data, such as understanding genetic data with specific shapes or diseases, environmental interactions, and inheritance patterns, is challenging. Therefore, 'omics' is proposed, which is related to a large amount of advanced analytical methods has been developed, such as next-generation high-throughput sequencing (NGS), liquid chromatography–mass spectrometry (LC-MS), and Proximity Extension Assay (PEA). These techniques enable the quantification of different kinds of small molecules, including RNAs, proteins, and metabolites.

In English terminology, the suffix '-ome' denotes a complete system, while '-omics' pertains to the study of such systems. For instance, 'genomics' is the systematic study of an organism's entire genome and the interrelationships among them. 3D genomics reveals the three-dimensional organization of DNA in the nucleus through sequencing. Metabolomics focuses on quantifying small metabolites in biological systems and subsequently studying relationships of these molecules with diseases or genes. Various omics data bring important evidence and support to the prediction and treatment of diseases, as well as understanding of intracellular regulation and metabolic mechanisms.

1.2 Genetics and Genomics

1.2.1 What are Genetics and Genomics?

Genetics and genomics are important tools in understanding disease. Genetics focuses on the study of genetic variants and how traits are inherited. DNA can be transcribed into transcript, which is then translated into protein to regulate cell activities and body functions. Genetics

generally studies the role of both single gene and complex trait inheritance patterns. Monogenic traits, as the name suggests, are traits that are only affected by a single genetic variation, such as phenylketonuria (Alghamdi et al., 2023), cystic fibrosis (Sun et al., 2022b; Sharma and Cutting, 2020), and Huntington's disease (Gusella et al., 2021). While these conditions are known to follow Mendelian principles of inheritance, accumulating evidence such as identification of modifier genes and incomplete penetrance reveal that even these seemingly straightforward conditions contain complex mechanisms (Scriver and Waters, 1999; Enns et al., 1999; Wright et al., 2022; Shin et al., 1997). Despite their rarity on an individual level, Mendelian disease are collectively relative common, contributing significantly to the disease burden (Antonarakis and Beckmann, 2006). However, most common diseases, such as many cancer subtypes, cardiovascular events, diabetes, asthma, dementia, etc., are complex traits and do not arise from a single gene mutation. Even for monogenic conditions, allelic heterogeneity exists in that the same disease can be caused by mutations at different locations within the same gene. In addition, environmental factors or additional genes may modify disease penetrance and symptom severity. On the other hand, complex traits are influenced by multiple genes and environmental factors. Complex traits do not follow readily predictable patterns of inheritance within families. For many complex traits, hundreds of genetic loci have been identified, but mechanisms by which these loci influence disease risk are often very unclear.

Genomics, on the other hand, studies the entirety of an individual's genome, including interactions of genes with each other and with the person's environment. Genomics can help to develop efficient, cost-effective and robust means of preventing, diagnosing, and treating major diseases.

Genomics has great potential benefits for improving human health and advancing personalized medicine. In oncology, genomics provides candidate targets for drug discovery and drug redirection, as well as identifying specific cancer subtypes for differential treatment. In the field of population health management, genomics can help identify important genetic loci associated with disease prevalence in specific communities and provide information for public health strategies.

Genomics can also contribute to genetic counseling, providing important information about a child's risks for genetic diseases and providing information for reproductive decision-making plans.

1.2.2 Genome-Wide Association Studies (GWAS) and Post-GWAS Analysis

Building on the foundational concepts and benefits of genetics and genomics outlined in Sections 1.2.1, we now transition to a more focused examination of Genome-Wide Association Studies (GWAS) and the subsequent post-GWAS analyses. This exploration is crucial in understanding how genetic variations contribute to complex traits and diseases, and how this knowledge can be further advanced to improve our understanding of the biology and/or clinical practice.

GWAS are designed to identify associations between genotypes and phenotypes, for example, by examining differences in allele frequencies of genetic variants among individuals with similar ancestry but differing phenotypically. Over the past 15 years, GWAS have enabled the identification of polymorphisms associated with disease risk and various complex traits, thereby explaining a portion of familial risk (Abdellaoui et al., 2023). However, the underlying molecular and biological processes of these significant associations between traits and variants are unclear. In this context, post-GWAS analyses have become crucial for deciphering the functional consequences of notable single-nucleotide polymorphisms (SNPs) and other genetic variants (e.g., short insertion or deletions) identified through GWAS (Falola et al., 2023).

It is challenging to identifying the casual variants from the large amount of non-casual variants included in GWAS signals due to linkage disequilibrium (LD) (Chapter 2). Understanding LD's impact on allele associations is critical for interpreting GWAS results and lays the groundwork for more targeted genetic investigations. Fine-mapping is an in-silico process designed to prioritize the set (or sets) of statistically distinct variants most likely to be causal within each genetic locus identified by GWAS. This method uses LD and variant-level summary statistics to infer credible sets, which are further analyzed to identify the putative casual variant(s) associated a specific trait (Grapes et al., 2004).

Post-GWAS analyses aim to understand the biological functions of GWAS signals. In Post-GWAS, there are various approaches such 3D genomics (Uffelmann et al., 2021) and metabolomics.

Using different types of omics data as functional annotations, we can gain understanding of the potential mechanisms by which regulatory elements influence the progress of disease in a specific tissue or organ. There are some practical applications for Post-GWAS analyses. For example, Post-GWAS facilitates identification of new drug targets. A recent study discusses how GWAS findings can be leveraged to repurpose existing drugs. This study utilized various approaches such as linking GWAS signals to genes and pathways, conducting transcriptome-wide association studies, gene-set association, causal inference by Mendelian Randomization, and polygenic scoring prediction (Reay and Cairns, 2021). GWAS provides the initial candidate genetic variants for this type of study. Post-GWAS methods are used to understand those signals and to infer the putative casual variants using multiple methods and functional annotations.

1.3 Linkage Disequilibrium

Linkage Disequilibrium (abbreviated as LD) is an important concept of population genetics and genomics. In a given population, LD is the non-random association of alleles at different loci (Slatkin, 2008). Essentially, LD measures how the frequency of the combination of alleles on a single chromosome differs from what would be expected if their segregation were independent. When two loci are in LD, the inheritance of one locus's allele can predict the inheritance of the allele at the other locus more than by chance alone. Therefore, LD throughout the genome reflects population history and pattern of geographic subdivision, whereas LD in each genomic region reflects the history of natural selection, gene conversion, mutation and other forces that cause gene-frequency evolution (Slatkin, 2008). LD can help to gain insights in a variety of different applications, from population genetic research to disease association studies (Huang et al., 2022).

The current most prevalent tools, HaploReg (Ward and Kellis, 2012) and LDlink (Machiela and Chanock, 2015), utilized 1000 Genome Project (1KGP) Data (Consortium et al., 2015a) to estimate LDs. Although 1KGP Phase 3 sequences genomes from over 2K individuals to provide a large resource for understanding human genetic variation, it still lacks power to capture rare variants and to represent diverse populations. In addition, those two LD query resources did not provide structural variants. Structural variants are critical in understanding genomic diversity and disease

susceptibility. Moreover, variants on chromosome X constitute an important 5% of the genome largely unexplored in GWAS and post-GWAS studies.

However, HaploReg and LDlink include neither structural variants nor genetic variants on X chromosome. Even once credible sets of potentially causal variants are identified at a GWAS locus, more work is needed to link variants to their likely mechanisms, using for example 3D genomics information or high throughput omics platforms.

1.4 3D Genomics

1.4.1 High-throughput chromosome conformation capture (Hi-C) and its variants

Three-dimensional (3D) genomics is a research hotspot in the post-genomics era and the post-GWAS era. In the intricate landscape of the mammalian genomic architecture, the progressive revelation of its 3D structure within the nuclear confines plays a pivotal role in identifying and understanding modulatory functions of cis-regulatory elements, even across extensive mega base (Mb) intervals (Li et al., 2018). The emerging area of Chromosome Conformation Capture (3C) techniques, including 3C, 4C, 5C, and Hi-C, presents a suite of tools dissecting the spatial configurations of chromatin. Among this 3C family, Hi-C technology emerges as a distinguished contender due to its lesser bias and greater reproducibility compared to earlier 3C methods (Lieberman-Aiden et al., 2009) by adopting an unbiased genome-wide approach examining all-to-all chromatin interactions, in contrast to one-to-one interactions in 3C and one-to-many interactions in 4C. In 2016, novel technologies known as HiChIP and PLAC-seq (Mumbach et al., 2016; Fang et al., 2016a) were introduced for the purpose of assessing protein-mediated chromatin interactions. HiChIP and PLAC-seq are variations of Hi-C technologies, which adds a chromatin immunoprecipitation (ChIP) step to capture interactions associated with a protein of interest. Therefore, in contrast to genome-wide unbiased Hi-C sequencing method, those variations of Hi-C improve the signal-to-noise ratios (SNR) of interactions between specific DNA-binding proteins (or histone modifications) and genomic loci. By focusing on the subset of chromatin interactions (of primary interest to investigators), these two technologies can produce substantially higher-depth data for interactions of interest with reduced costs.

1.4.2 *In silico* methods for enhancing sequencing depth

High-quality Hi-C data can provide evidence of interactions between regulatory element regions and target genes (e.g., enhancer promotor interactions). However, it is costly to obtain high sequencing depth bulk Hi-C data.

For example, achieving the kilobase (Kb) resolution demands around 5 billion paired-end reads per sample (Rao et al., 2014a). Even methods like HiChIP and PLAC-seq that focus on chromatin interactions mediated by a pre-selected protein or histone modification mark require 500 million to 1 billion raw reads per sample to map enhancer-promoter interactions at Kb resolution (Mumbach et al., 2016; Fang et al., 2016a).

In an effort to reduce the cost, several computational methods have emerged, offering *in silico* solutions. Inspired by super-resolution, a classic method in computer vision, HiCPlus (Zhang et al., 2018), HiCNN (Liu and Wang, 2019a), HiCNN2 (Liu and Wang, 2019b), DeepHiC (Hong et al., 2020), and VEHiCLE (Highsmith and Cheng, 2021) were developed. These are all deep learning neural networks, but their architectures are a little different. For example, HiCPlus leverages three layers of convolution neural networks to project low-depth Hi-C data into high-depth Hi-C data. HiCNN employs a 54-layer CNN with ResNet while HiCNN2 extends this conception via ensembling three deep learning models. This field is still active with more recent developed methods for Hi-C data enhancement (Hu and Ma, 2021a; Hicks and Oluwadare, 2022).

1.4.3 Metabolomics

Metabolites are important small molecules for cellular and physiological processes in plasma, serum, urine, and cerebrospinal fluid (CSF) (Donatti et al., 2020; Jakkula et al., 2008). Aberrant metabolite levels might indicate the presence of a potential illness (Thysell et al., 2010). Thus, researchers may investigate how metabolite biomarkers associate with some traits of interest (e.g., diseases) with matching tissues and organs, which might offer new insights into human health. While blood plasma is the most common tissue for metabolomic studies, other tissues may for some diseases be even more relevant. For example, some studies have explored the relationship between CSF metabolite biomarkers and neurological/psychiatric phenotypes, such as Alzheimer's disease,

schizophrenia, attention deficit hyperactivity disorder, post-traumatic stress disorder, cognitive performance, and others (Noga et al., 2012; Panyard et al., 2021). In another example, some researchers investigated whether specific urine metabolites (urinary citrate, taurine, and hippurate) can be used to diagnose lupus nephritis (LN), a kidney disease, where the biomarkers just from urine sample might be a replacement of kidney biopsy monitoring LN status (Romick-Rosendale et al., 2011). Plasma is present throughout the body, which is different from urine and CSF, which are localized to specific systems or tissues. Therefore, metabolites associate with various diseases regarding to different tissues, organs, or biological process such as cardiovascular diseases, diabetes, kidney diseases, and metabolic diseases. For example, metabolic profiles can predict cardiovascular events independently of standard predictors. Specifically, medium-chain acylcarnitine, short-chain dicarboxylic acylcarnitine and long-chain dicarboxylic acylcarnitine can independently predict death/myocardial infarction events (Ruiz-Canela et al., 2017; Shah et al., 2012). The non-invasive nature of metabolomics, along with its strong connection to phenotypes of interest, renders it a highly suitable tool for applications in pharmaceuticals, preventive healthcare, agriculture, and various other industries.

Although some biobanks have started to generate metabolomics data in relatively large sample sizes (Group et al., 2023), the usual small sample size in the vast majority of study cohorts and tissues is still an issue, leading to lack of power to identify trait-metabolomic associations. Directly measuring metabolites in larger samples sizes is an obvious solution. For example, a urine metabolomics study concluded that increasing sample size will provide more precise results to determine whether the candidate urine metabolites contribute to LN disease (Romick-Rosendale et al., 2011) since sample sizes of individual urine metabolite studies are small. However, such additional measurements may be costly, and appropriate tissues and samples may be difficult to collect. In the meantime, we can increase effective analysis sample size and study power in alternative ways. For some applications, such as identifying potential genetic mechanisms mediated through metabolite(s), it would be highly informative to study only the genetic component of metabolite variance without having to directly measuring the metabolites in study samples. Since

we have genotype information for much larger sample sizes (e.g., from GWAS studies) than direct measurement of metabolites, we could substantially improve sample size by leveraging genotype data to study genetically determined metabolite levels. To do this, however, there is a critical need to develop a framework to predict metabolites utilizing more widely available and easily measurable germline genetic variant data and perform association with traits of interest. We provide a preliminary method to address this issue in the analyses presented in my thesis.

CHAPTER 2: TOP-LD: A TOOL TO EXPLORE LINKAGE DISEQUILIBRIUM WITH TOPMED WHOLE-GENOME SEQUENCE DATA

This chapter previously appeared as a paper in the Journal of American Journal of Human Genetics. The original citation is as follows: Huang, Le*, Jonathan D. Rosen*, Quan Sun*, Jiawen Chen, Marsha M. Wheeler, Ying Zhou, Yuan-I. Min et al. "TOP-LD: A tool to explore linkage disequilibrium with TOPMed whole-genome sequence data." The American Journal of Human Genetics 109, no. 6 (2022): 1175-1181. * indicates co-first authors who have contributed equally to this work.

2.1 Introduction

Linkage disequilibrium (LD), i.e., the non-random association of alleles at different variant sites in a given population, is an important genetic phenomenon. Patterns of LD between genetic markers can be leveraged to gain insights in a variety of different applications, from population genetic research to disease association studies (Slatkin, 2008; Bush and Moore, 2012). With the growth of whole-genome sequencing (WGS) and high-throughput array and genotype imputation technologies, resources for calculating LD across populations have expanded to encompass multiple populations at variant sites with increasingly rare frequencies (Choudhury et al., 2020; Consortium et al., 2015b, 2010b; Taliun et al., 2021). Due to the centrality of LD in a host of applications, multiple tools exist for querying LD between genetic markers in different populations. The current most widely used LD lookup tools, HaploReg (Ward and Kellis, 2016) and LDlink (Machiela and Chanock, 2015) base their LD estimates on the 1000 Genomes data. Specifically, HaploReg uses phase 1 and LDlink uses phase 3 1000 Genomes data. Although the 1000 Genomes data contains LD information on $> 99\%$ of genetic markers with minor allele frequency (MAF) $> 1\%$ in a variety of populations (Consortium et al., 2015b), there remains a dearth of publicly available information on LD between markers with MAF $< 1\%$. We have created a new LD lookup tool

(called "TOP-LD"), in the spirit of HaploReg and LDlink, that is based on deep ($30\times$) WGS data from the NHLBI Trans-Omics for Precision Medicine (TOPMed) Program. Because the TOPMed data contain much larger sample sizes with greater depth of sequencing than the 1000 Genomes project, TOP-LD provides a significant upgrade in LD information availability, specifically by including single-nucleotide variants and small indels (referred to hereafter simply as "SNVs") with MAF $< 1\%$ as well as structural variants (SVs). Here, we describe the data and methods that went into creating TOP-LD along with specific examples of how TOP-LD can provide essential information that is missed by HaploReg and LDlink.

2.2 Results

2.2.1 Enhanced Variant Coverage

The TOP-LD tool leverages TOPMed WGS data, whose much larger sample size and high depth sequencing lead to LD information for a much larger number of variants compared to the 1000 Genomes Project. As shown in Figure 1A and Table S1, TOP-LD offers 2.6- to 9.1-fold increase in variant coverage compared to the other state-of-the-art resources such as HaploReg 4.0 (Ward and Kellis, 2016) or LDlink (Machiela and Chanock, 2015). For example, for the European population, TOP-LD includes 146.5 million autosomal SNVs, while HaploReg 4.0 or LDlink contains 16.1 million variants. Not surprisingly, the vast majority of the variants in TOP-LD that are not in 1000 Genomes, contributing to the up to 9.1× increase, are low frequency or rare. For example, out of the 146.5 million autosomal SNVs cataloged in the TOP-LD European population, 137.8 million have MAF < 0.01 (Figure 2.1A, Table SA1). Most of the variants have LD proxies. For example, 115.1 out of the 146.5 (78.6%) million autosomal variants have at least one LD tag with R2 ≤ 0.8 and if we further relax the R2 threshold to 0.5 and 0.2, the number increases to 135.3 (92.4%) and 143.5 (98.0%), respectively (Figure 2.1B).

For chromosome X, we have included 6.5 million, 2.4 million, 1.3 million, and 760,000 variants for the European, African, East Asian, and South Asian populations, respectively (Table S2.1). Similar to the autosomal variants, the majority of these variants have at least one LD proxy

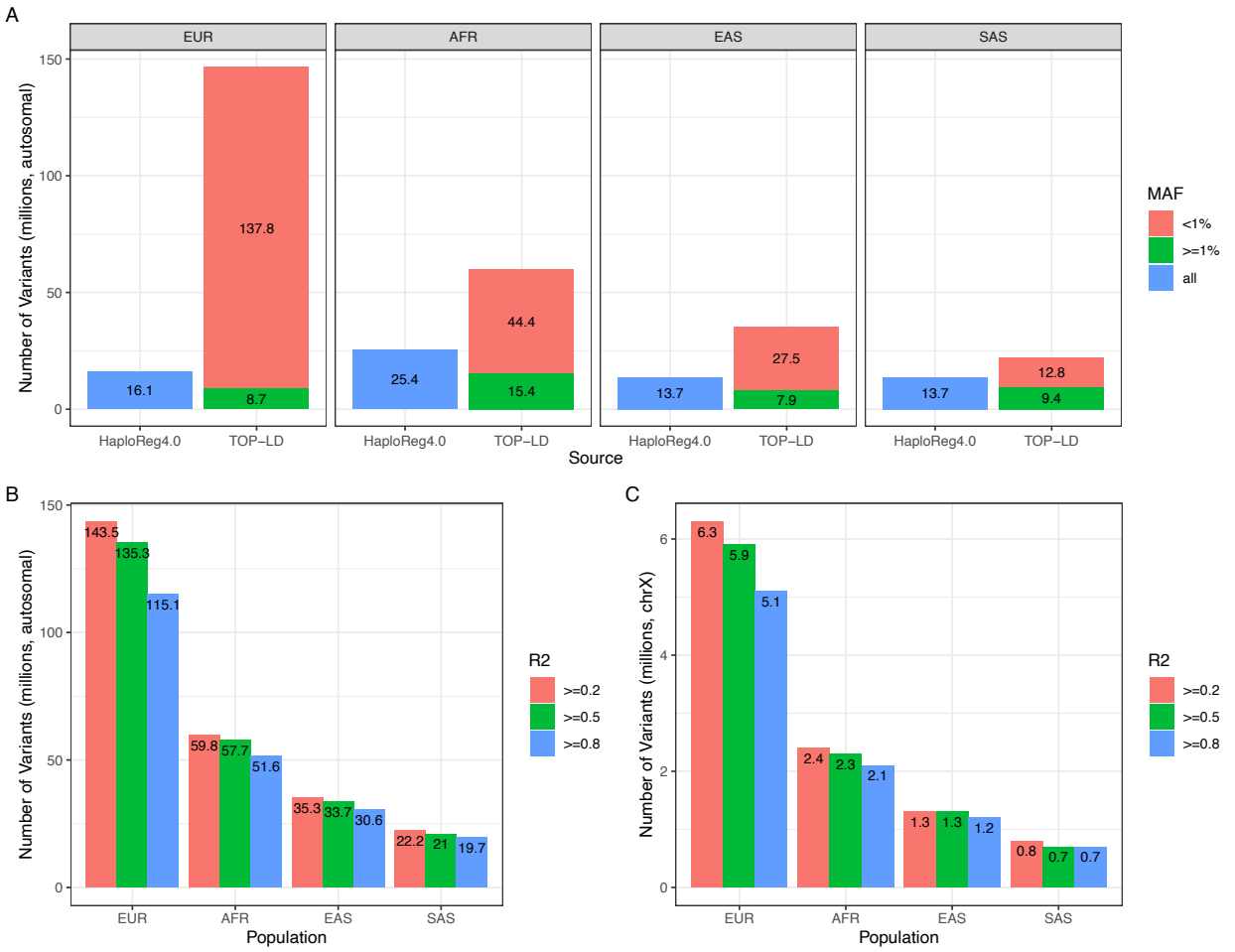

Figure 2.1: Number of variants included in TOP-LD.(A)Comparison of autosomal variants with HaploReg 4.0 by population. Blue bars on the left show total number of autosomal variants in HaploReg4.0. Green and red indicate common (MAF $\geq$ 1%) and uncommon (MAF $<$ 1%) autosomal variants in TOP-LD. Note that HaploReg4.0 provides LD for ASN (Asian) with no separate information for EAS and SAS. Therefore, we used the same 13.7 million ASN variants for comparison in both EAS and SAS. (B) Number of autosomal variants in TOP-LD breaking down by LD R2 threshold. The majority of the variants have at least one LD proxy with R2 $\geq$ 0.8. (C) Number of chrX variants in TOP-LD breaking down by LD R2 threshold. (Note: LD information downloaded from HaploReg4.0 does not contain chromosome X. Therefore, we compared TOP-LD with HaploReg4.0 only for autosomal variants).

with R2 $\geq$ 0.8: 5.1 million, European; 2.1 million, African; 1.2 million East Asian; 690,000, South Asian (Figure A1C, Table AS2).

2.2.2 Consistency Analysis

To evaluate the consistency between TOP-LD estimates and those from Haploreg v4.1, we collected the set of overlapping variants based on rsID with MAF $\geq$ 0.05 for Europeans and Africans. This set of variants was further filtered such that the MAF values were within 10% of

each other because large MAF differences would induce large LD differences. Figures S1 and S2 show high level of agreement between TOP-LD and Haploreg v4.1 LD estimates (e.g., Pearson correlation = 0.972 and 0.962 for European and African chromosome 1, respectively). Similarly, comparison of the chromosome X TOP-LD estimates for females and males again shows high level of consistency (Pearson correlation = 0.992 and 0.975 for European and African population, respectively) (Figures AS3 and AS4).

2.2.3 Application of TOP-LD in Fine-Mapping

To demonstrate the utility of TOP-LD, we performed fine-mapping at the GGT1 locus on chromosome 22, which is known to be associated with gamma glutamyltransferase. We performed sequential conditional analysis with EPACTS (Kang et al., 2010) by using individual-level data among 8,768 UK Biobank participants of African ancestry following the same strategy in our previous work (Raffield et al., 2020) adjusting for the same covariates as in Sun et al (Sun et al., 2022a). The sequential conditional analyses with individual-level data identified seven distinct signals at the GGT1 locus associated with gamma glutamyltransferase (Table 2.1). Because we used individual-level data for this conditional analysis, we considered these seven distinct signals to be the "working truth."

Table 2.1: Summary statistics of distinct working truth at *GGT1* locus associated with gamma glutamyltransferase.

Signal	Variant	Position (hg38)	Effect allele	unconditional p-value	p-value conditional on previous signals*	Effect allele frequency
1	rs4049904	24609759	G	2.82e-61	NA	10.27%
2	rs73404962	24598530	G	4.46e-29	2.00e-36	5.63%
3	rs743369	24588099	A	9.94e-36	7.51e-27	11.94%
4	rs6004193	24598329	C	4.23e-41	3.25e-19	18.27%
5	rs57719575	24609020	C	3.97e-38	1.98e-24	14.86%
6	rs3876101	24607291	A	2.66e-15	1.17e-13	35.45%
7	rs116161010	24585912	T	5.69e-17	7.70e-09	7.13%

*The p-values are reported from the sequential conditional analysis. For example, we report the p-value for rs73404962 conditional on rs4049904, the p-value of rs743369 conditional on both rs4049904 and rs73404962, and so forth.

We then carried out fine-mapping analysis with the FINEMAP method (Benner et al., 2016) by using only GWAS summary statistics from Sun et al (Sun et al., 2022a). We applied FINEMAP with an LD reference either from TOP-LD or from the 1000 Genomes Project and assessed the performance by comparing the results with "working truth" established from the sequential conditional analysis of the individual-level data.

FINEMAP produced 95% credible sets containing five variants when using either the 1000 Genomes (1000G) Project LD panel or the TOP-LD panel (see Table 2.2). However, the 1000G-based credible set contained only one of the seven signals from the "working truth" set. In contrast, the TOP-LD-based credible set contained three of the seven signals from the "working truth" set. In addition, because the lead variant from each conditional analysis (corresponding to each distinct signal) is selected somewhat arbitrarily, we also considered their LD proxies. When we considered any LD proxy (using a lenient R2 threshold of 0.2) of a variant in the working truth set, the 1000G-based results still only identified a single signal from the working truth, whereas the TOP-LD-based results identified four of the seven signals (Table 2.2).

Table 2.2: FINEMAP credible-set variants.

	Variant 1	**Variant 2**	**Variant 3**	**Variant 4**	**Variant 5**
credible set variant	rs4049904	rs147866692	rs570263050	rs115231893	22:24649848:G:A (hg38)
1000G reference	1 (w/ rs4049904 itself)	0.464 (w/ rs4049904)	0.606 (w/ rs4049904)	0.275 (w/ rs4049904)	0.434 (w/ rs4049904)
credible set variant	rs4049904	rs743369	rs57719575	rs2073397	rs5751902
TOP-LD reference	1 (w/ rs4049904 itself)	1 (w/ rs743369 itself)	1 (w/ rs57719575 itself)	0.83 (w/ rs6004193)	0.51 (w/ rs6004193)

*The two five-variant credible sets provided by FINEMAP with either 1000G or TOP-LD as reference. For each credible-set variant, we list the corresponding variant (and the LD Rsq) from the working truth that has the highest LD

2.2.4 Application of TOP-LD in Causal Structural Variants Prioritization

We also used TOP-LD to aid in the identification and prioritization of potentially causal structural variants at GWAS loci. For example, our recent association analysis (Mikhaylova

et al., 2021) with TOPMed data identified an African-specific (MAF = 0.129) variant rs28450540 associated with lower monocyte count (p-value = 3.65×10^{-17}). Query for LD tags via TOP-LD revealed a ∼600 bp deletion near *S1PR3* in perfect LD (R^2 = 1) with rs28450540 in the African population. We performed genome editing in monocytic and primary human HSPCs followed by xenotransplantation, which provides evidence that the deletion disrupts an *S1PR3* monocyte enhancer leading to decreased *S1PR3* expression. These preliminary data from functional experiments suggest that the 600 bp deletion is most likely casual but would have been missed in standard association analysis with only SNVs (Wheeler et al., 2022). TOP-LD offers a simple and efficient approach to rescue such putative causal structural variants.

2.3 Method

2.3.1 Data Preprocessing

We used TOPMed WGS data from the following four cohorts: BioMe Biobank (BioMe), the Multi-Ethnic Study of Atherosclerosis (MESA), the Jackson Heart Study (JHS), and the Women's Health Initiative (WHI). We aimed to provide LD estimates for genetically homogeneous groups of individuals from one of the following four ancestral populations: European (EUR), African (AFR), East Asian (EAS), and South Asian (SAS). To select appropriate samples, we first inferred local and global ancestry for all participants in these four cohorts by using RFMix (Maples et al., 2013), with reference populations including five ancestral groups, namely African, Native American, East Asian, European, and South Asian. After local ancestry inference, we then retained only TOPMed samples with >90% estimated ancestry from a single population, as estimated via RFMix (Maples et al., 2013). We further removed related individuals by using a stringent kinship coefficient threshold of $2^{-5.5}$ obtained via PC-Relate (Conomos et al., 2016). This threshold of $2^{-5.5}$ removes pairs within as far as fifth degree relationship. The final dataset included 1,335 unrelated individuals of African, 844 of East Asian, 13,160 of European, and 239 of South Asian ancestry for pairwise LD inference. Regarding variants, we started with all TOPMed freeze 8 polymorphic variants that passed quality control and retained multi-allelic variants or multiple entries at the same position, resulting in a total of 23.0–153.0 million SNVs in each of the ancestral groups (Figure 2.1, Table 2.1).

2.3.2 LD Inference

We inferred LD separately within each of the four ancestral groups, for all pairs of variants within 1 Mb of each other and retained LD pairs meeting a minimum R^2 threshold of 0.2. The reported R^2 between two variants is the squared Pearson correlation coefficient between their phased haplotypes, where phasing was performed with Eagle 2.4 for all polymorphic variants, similar to phasing of the freeze 5 data (Taliun et al., 2021). No minimum minor allele count thresholding was used, that is, even singletons in our sample were included in LD calculations. We also report the direction of each association as either positive (+) or negative (-) on the basis of the sign of the Pearson correlation coefficient between the corresponding pair of reference (REF) alleles. In addition to R2, we also report D-prime statistics for each pair of variants meeting the R2 of 0.2.

We filtered chromosome X to exclude the pseudo-autosomal regions: PAR1 (bp 10,001 – 2,781,479, GRCh38) and PAR2 (bp 155,701,383 – 156,030,895, GRCh38). Variants that were not coded as homozygous in the males were excluded from the LD calculations. We inferred LD for the remaining variants by using a total of 2F + M haplotypes, where F and M are the numbers of females and males, respectively. The TOPMed structural variant (SV) call-set freeze 1 was merged with a reduced TOPMed SNV call-set where SNVs with MAF $< 0.1\%$ were filtered out before merging, and then the merged SV-SNV dataset was phased with Eagle2. SVs with $>10\%$ missingness were removed prior to phasing. For each ancestry group, we included 16.5–79K SVs (deletions, duplications, and inversion) with the majority being lower frequency (e.g., 7–69K with MAF $< 1\%$) (Table 2.1). LD values were subsequently estimated as the squared Pearson correlation coefficient between the corresponding pair of phased alleles.

2.3.3 TOP-LD website

TOPMed LD information was then loaded into the TOP-LD website, which is powered by a combination of MySQL, PHP, JavaScript, and Apache2 under the Cloud SQL and Compute Engine of Google Cloud Platform. The web interface provides access to all precomputed LD estimates. Users have the option to either paste or upload a file containing variant(s) of interest. Users can specify the population (East Asian, European, African, or South Asian) in which LD was

estimated. In TOP-LD, markers are identified by rsID, or chr:position, or chr:position:REF:ALT for SNVs, or TOPMed variant names for SVs (in the format of INDEL_chr:startPosition-endPosition which INDEL can be DEL, DUP, or INV, for example, DEL_10:85001–97300). TOP-LD returns all variants within a pre-specified LD threshold (ranging from R2 values of 0.2 to 1.0) with the query variant. TOP-LD supports fast batch queries (Figure 2.2); querying a single variant takes ∼0.5 seconds, while a batch query of 500 variants takes ∼2.3 seconds. TOP-LD currently allows a maximum of 500 variants in one query.

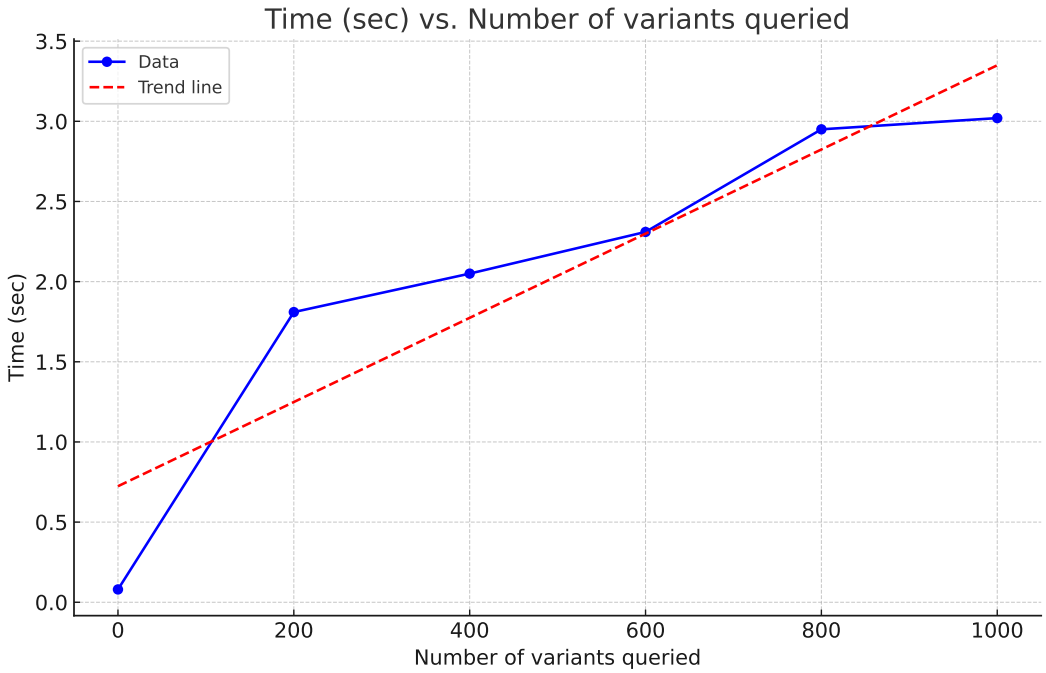

Figure 2.2: The x axis represents the number of variants queried, and the y axis represents the elapsed time.

After submitting the query, the website auto-directs to a result page that contains two parts: LD information on the top panel and variant information on the bottom panel. The latter provides basic information for the queried variants, including position, marker name, alleles (REF and alternative allele, ALT), and minor allele frequency (MAF). Markers not in the database will have "none" for all fields except marker names. The LD panel displays related LD metrics, one pair of variants on each line, including both R^2, D', and the sign of LD (measured between REF alleles of the two variants), along with marker name, marker position, alleles, and frequency for both variants

in the pair (Figure 2.3). In addition, we provide the following pieces of information for SNVs from WGSA annotation : CADD score (phred-scaled), fathmm_XF_coding_or_noncoding classification, FANTOM5 enhancer annotations, gene name, and relative location to gene as well as a link to GWAS catalog query results (MacArthur et al., 2017). For SVs, we provide a variety of annotations including gene(s) overlapping the SV, the SV's location relative to gene, the gene's probability of loss-of-function intolerance (pLI) score, overlapping candidate cis-regulatory regions (cCREs) from ENCODE SCREEN (Geoffroy et al., 2018; Moore et al., 2020). The query results can be sorted, searched, copied, exported, and printed for further analyses.

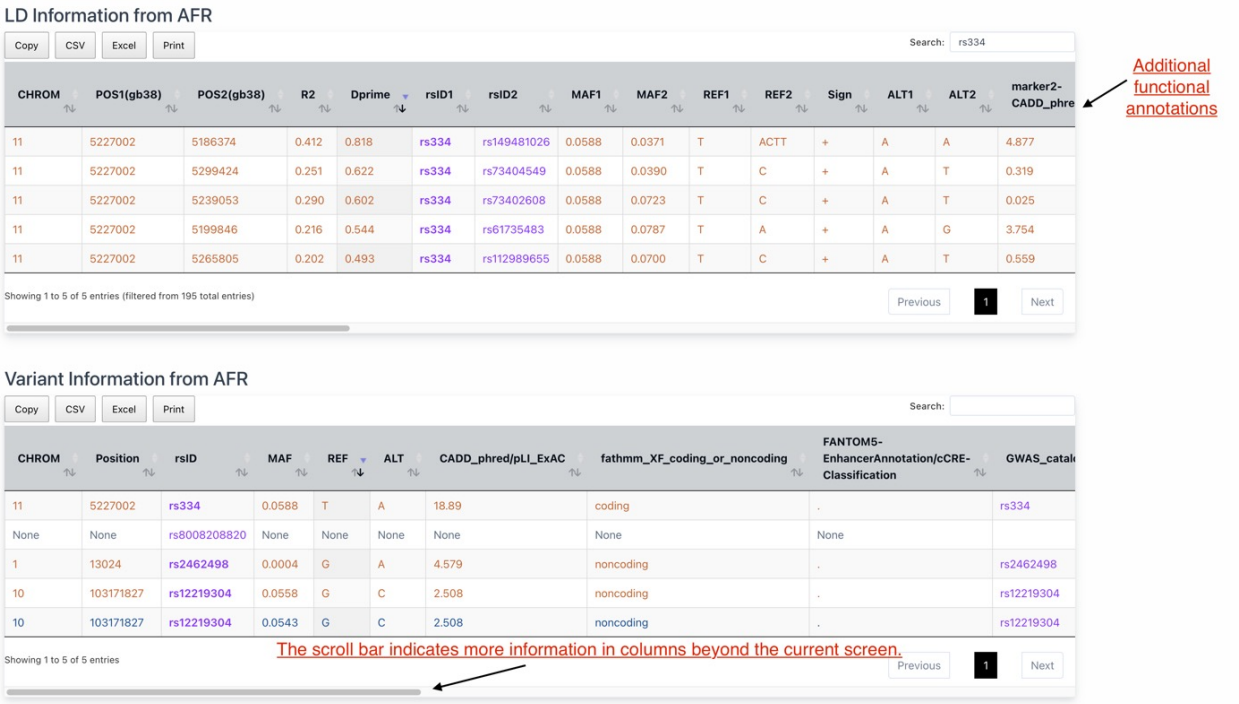

Figure 2.3: An example query result. The result contains two parts. The top part "LD information from AFR" shows the LD information where each line provides information between a query variant (rsID1) and one of its corresponding LD proxies (rsID2). The bottom part "variant information from AFR" provides variant information, which shows basic information for each query variant. From the bottom part, we know that the user's query includes four variants: rs334, rs8008208820, rs2462498, and rs12219304. Variants not included in LD calculation will have "none" records. For instance, rs8008208820 in this example query is not involved in LD inference and therefore will not have any LD proxies in the top part simply because of no data. Records from SV inference are in blue and those from SNV data are in orange. Some variants may appear twice because they are included in both SNV LD calculation and SV calculation. For example, in this example, rs12219304 appeared twice with MAF 0.0558 from the SNV source (second last record in orange) and MAF 0.0543 from the SV source (last record in blue).

2.4 Conclusion

LD information, reflecting recombination, natural selection, and demographic history, has always been of intense interest in population genetics and complex trait association studies. LD information is also indispensable for a wide range of other applications, including GWAS follow-up and many summary-statistics-based inferences including fine-mapping, imputation of association summary statistics, construction of polygenic risk scores (PRSs), and interpretation and prioritization of GWAS results for further functional and clinical studies. TOP-LD significantly boosts the coverage of lower frequency variants by harnessing the power of high-coverage ($\sim 30\times$) WGS data of over 15,000 individuals primarily of a single continental ancestry. We demonstrate the utility of TOP-LD in fine-mapping at the *GGT1* locus and variant prioritization at the *S1PR3* locus. The LD information provided by TOP-LD will facilitate a range of essential inferences for common and rare variation across a diverse range of populations.

Web resources:

EPACTS, https://genome.sph.umich.edu/wiki/EPACTS

FINEMAP, http://www.christianbenner.com/

HaploReg, https://pubs.broadinstitute.org/mammals/haploreg/

LDlink, https://ldlink.nci.nih.gov/

TOP-LD, http://topld.genetics.unc.edu/

TOPMed, https://topmed.nhlbi.nih.gov/

2.5 Supplementary Materials

Supplementary Materials provide Suplemantary Methods, Supplementary tables and figures.

2.5.1 Supplementary Methods

2.5.1.1 TOPMed samples

NHLBI's TOPMed program is comprised of many parent studies, including four ancestrally diverse studies that contributed to our analyses including BioMe Biobank (BioMe) (Gottesman et al., 2013), Jackson Heart Study (JHS) (Taylor et al., 2005; Wilson et al., 2005), Multi-Ethnic Study of Atherosclerosis (MESA) (Bild et al., 2002), and Women's Health Initiative (Group, 1998). Additional information about the design of each study and the sampling of individuals within each cohort for WGS is available in the 2.5.2 Cohort Descriptions section below. All studies were approved by the appropriate institutional review boards (IRBs), and informed consent was obtained from all participants.

2.5.1.2 TOPMed whole genome sequencing and quality control

WGS was performed at an average depth of 38$\times$ by six sequencing centers (Broad Genomics, Northwest Genome Institute, Illumina, New York Genome Center, Baylor, and McDonnell Genome Institute) using Illumina X10 technology and DNA from blood. Here we report analyses from the 'Freeze 8' dataset where reads were aligned to human-genome build GRCh38 using a common pipeline across all sequencing centers. To perform variant quality control (QC) within the 'Freeze 8' dataset, a support vector machine (SVM) classifier was trained on known variant sites (positive labels) and Mendelian inconsistent variants (negative labels). Further variant filtering was done for variants with excess heterozygosity and Mendelian discordance. Sample QC measures included: concordance between annotated and inferred genetic sex, concordance between prior array genotype data and TOPMed WGS data, and pedigree checks. Details regarding the genotype 'freezes', laboratory methods, data processing, and quality control are described on the TOPMed website and in a common document accompanying each study's dbGaP accession.

2.5.1.3 TOPMed structural variant calling and quality control

TOPMed structural variation (SV) callset release 1 was generated by Parliament2-muCNV pipeline across 138,134 multi-ethnic TOPMed WGS samples. The sample list overlaps largely with 'Freeze 8' callset except for the samples removed due to SV specific quality control issues. Parliament2 (Zarate et al., 2020) is a multi-tool SV discovery pipeline that employs SV callers that have strengths in different SV types and sizes to maximize the detection sensitivity and accuracy. SVs detected by individual tools are then merged first across the callers and then across the samples using SURVIVOR (Jeffares et al., 2017) to generate a 'discovery' SV callset. The 'discovery' set is then genotyped and filtered by muCNV, a multi-sample SV genotyping software that performs joint genotyping based on multi-sample statistics across >100,000 samples (Jun et al., 2021). Joint genotyping removes false discoveries by evaluating cluster separations using multi-sample distribution of read pair, split read, soft clips, and GC-corrected sequencing depth distributions. Parliament2, SURVIVOR, and muCNV are available for public access on GitHub:

- https://github.com/slzarate/parliament2
- https://github.com/fritzsedlazeck/SURVIVOR
- https://github.com/gjun/muCNV

2.5.2 Cohort Descriptions

BioMeThe Charles Bronfman Institute for Personalized Medicine at Mount Sinai Medical Center (MSMC), BioMe Biobank, founded in September 2007, is an ongoing, broadly consented electronic health record-linked clinical care biobank that enrolls participants non-selectively from the Mount Sinai Medical Center patient population. The MSMC serves diverse local communities of upper Manhattan, including Central Harlem (86% African American), East Harlem (88% Hispanic/Latino), and Upper East Side (88% Caucasian/White) with broad health disparities.

JHSThe Jackson Heart Study (JHS, https://www.jacksonheartstudy.org/jhsinfo/) is a large, community-based, observational study whose participants were recruited from urban

and rural areas of the three counties (Hinds, Madison and Rankin) that make up the Jackson, MS metropolitan statistical area (MSA). Participants were enrolled from each of 4 recruitment pools: random, 17%; volunteer, 30%; currently enrolled in the Atherosclerosis Risk in Communities (ARIC) Study, 31% and secondary family members, 22%. Recruitment was limited to non-institutionalized adult African Americans 35-84 years old, except in a nested family cohort where those 21 to 34 years of age were also eligible. The final cohort of 5,306 participants included 6.59% of all African American Jackson MSA residents aged 35-84 during the baseline exam (N-76,426, US Census 2000). Among these, approximately 3,700 gave consent that allows genetic research and deposition of data into dbGaP. Major components of three clinic examinations (Exam 1 – 2000-2004; Exam 2 – 2005-2008; Exam 3 – 2009-2013) include medical history, physical examination, blood/urine analytes and interview questions on areas such as: physical activity; stress, coping and spirituality; racism and discrimination; socioeconomic position; and access to health care. A fourth exam is ongoing. Extensive clinical phenotyping includes anthropometrics, electrocardiography, carotid ultrasound, ankle-brachial blood pressure index, echocardiography, CT chest and abdomen for coronary and aortic calcification, liver fat, and subcutaneous and visceral fat measurement, and cardiac MRI. At 12-month intervals after the baseline clinic visit (Exam 1), participants have been contacted by telephone to: update information; confirm vital statistics; document interim medical events, hospitalizations, and functional status; and obtain additional sociocultural information. Questions about medical events, symptoms of cardiovascular disease and functional status are repeated annually. Ongoing cohort surveillance includes abstraction of medical records and death certificates for relevant International Classification of Diseases (ICD) codes and adjudication of nonfatal events and deaths. CMS data are currently being incorporated into the dataset.

MESAThe MESA study is a study of the characteristics of subclinical cardiovascular disease (disease detected non-invasively before it has produced clinical signs and symptoms) and the risk factors that predict progression to clinically overt cardiovascular disease or progression of subclinical disease. MESA researchers study a diverse, population-based sample of 6,814 asymptomatic men and women aged 45-84. Thirty-eight percent of the recruited participants are white, 28 percent

African American, 22 percent Hispanic, and 12 percent Asian, predominantly of Chinese descent. Participants were recruited from six field centers across the United States: Wake Forest University, Columbia University, Johns Hopkins University, University of Minnesota, Northwestern University and the University of California - Los Angeles.

WHIThe Women's Health Initiative (WHI) is a long-term, prospective, multi-center cohort study that investigates post-menopausal women's health 8. WHI was funded by the National Institutes of Health and the National Heart, Lung, and Blood Institute to study strategies to prevent heart disease, breast cancer, colon cancer, and osteoporotic fractures in women 50-79 years of age. WHI involves 161,808 women recruited between 1993 and 1998 at 40 centers across the US. The study consists of two parts: the WHI Clinical Trial which was a randomized clinical trial of hormone therapy, dietary modification, and calcium/Vitamin D supplementation, and the WHI Observational Study, which focused on many of the inequities in women's health research and provided practical information about the incidence, risk factors, and interventions related to heart disease, cancer, and osteoporotic fractures.

Table 2.3: Acknowledgement for the funding

TOPMed Accession #	**TOPMed Project**	**Parent Study Name**	**TOPMed Phase**	**Omics Center**	**Omics Support**
phs001644	BioMe	BioMe	3	Baylor	HHSN268201600033I
phs001644	BioMe	BioMe	3	MGI	HHSN268201600037I
phs000964	JHS	JHS	1	NWGC	HHSN268201100037C
phs001416	AA_CAC	MESA AA_CAC	2	Broad Genomics	HHSN268201500014C
phs001416	MESA	MESA	2	Broad Genomics	3U54HG003067-13S1
phs001237	WHI	WHI	2	Broad Genomics	HHSN268201500014C

BioMe: The Mount Sinai BioMe Biobank has been supported by The Andrea and Charles Bronfman Philanthropies and in part by Federal funds from the NHLBI and NHGRI (U01HG00638001; U01HG007417; X01HL134588). We thank all participants in the Mount Sinai Biobank. We also thank all our recruiters who have assisted and continue to assist in data collection and manage-

ment and are grateful for the computational resources and staff expertise provided by Scientific Computing at the Icahn School of Medicine at Mount Sinai.

JHS: The Jackson Heart Study (JHS) is supported and conducted in collaboration with Jackson State University (HHSN268201800013I), Tougaloo College (HHSN268201800014I), the Mississippi State Department of Health (HHSN268201800015I) and the University of Mississippi Medical Center (HHSN268201800010I, HHSN268201800011I and HHSN268201800012I) contracts from the National Heart, Lung, and Blood Institute (NHLBI) and the National Institute on Minority Health and Health Disparities (NIMHD). The authors also wish to thank the staffs and participants of the JHS. The views expressed in this manuscript are those of the authors and do not necessarily represent the views of the National Heart, Lung, and Blood Institute; the National Institutes of Health; or the U.S. Department of Health and Human Services.

MESA: MESA and the MESA SHARe project are conducted and supported by the National Heart, Lung, and Blood Institute (NHLBI) in collaboration with MESA investigators. Support for MESA is provided by contracts 75N92020D00001, HHSN268201500003I, N01-HC-95159, 75N92020D00005, N01-HC-95160, 75N92020D00002, N01-HC-95161, 75N92020D00003, N01-HC-95162, 75N92020D00006, N01-HC-95163, 75N92020D00004, N01-HC-95164, 75N92020D00007, N01-HC-95165, N01-HC-95166, N01-HC-95167, N01-HC-95168, N01-HC-95169, UL1-TR-000040, UL1-TR-001079, UL1-TR-001420. MESA Family is conducted and supported by the National Heart, Lung, and Blood Institute (NHLBI) in collaboration with MESA investigators. Support is provided by grants and contracts R01HL071051, R01HL071205, R01HL071250, R01HL071251, R01HL071258, R01HL071259, by the National Center for Research Resources, Grant UL1RR033176. The provision of genotyping data was supported in part by the National Center for Advancing Translational Sciences, CTSI grant UL1TR001881, and the National Institute of Diabetes and Digestive and Kidney Disease Diabetes Research Center (DRC) grant DK063491 to the Southern California Diabetes Endocrinology Research Center.

WHI: The WHI program is funded by the National Heart, Lung, and Blood Institute, National Institutes of Health, U.S. Department of Health and Human Services through contracts

HHSN268201600018C, HHSN268201600001C, HHSN268201600002C, HHSN268201600003C, and HHSN268201600004C.

The views expressed in this manuscript are those of the authors and do not necessarily represent the views of the National Heart, Lung, and Blood Institute; the National Institutes of Health; or the U.S. Department of Health and Human Services.

2.5.3 Supplementary Figures and Tables

Table S2.1: Summary of SNVs and small indels by population by MAF.

Population	**#TOP-LD variants[a](MAF >0%) in millions (chrX[b])**	**#TOP-LD variants (MAF <1%) in millions (chrX[b])**	**#autosomal variants in HaploReg4.0 in millions**
EUR	153.0 (6.5)	144.0 (6.2)	16.1
AFR	62.2 (2.4)	46.2 (1.8)	25.4
SAS	23.0 (0.8)	13.3 (0.5)	13.7[d]
EAS	36.7 (1.3)	28.6 (1.1)	-

[a] number of unique variants, genome-wide (including autosomes and chromosome X)
[b] number of unique variants on chromosome X
[c] based on HaploReg LD information downloaded from https://pubs.broadinstitute.org/mammals/haploreg/haploreg_data/, which does not contain chromosome X.
[d] HaploReg4.0 provides LD for ASN (Asian), with no separate information for SAS and EAS.

Table S2.2: Summary of SNVs and small indels by population by varying LD R^2 thresholds.

Population	**#variants[a]($R^2 \geq 0.2$), in millions (chrX[b])**	**#variants[a]($R^2 \geq 0.5$), in millions (chrX[b])**	**#variants[a]($R^2 \geq 0.8$) in millions (chrX[b])**
EUR	149.8 (6.3)	141.2 (5.9)	120.2 (5.1)
AFR	62.2 (2.4)	60.0 (2.3)	53.7 (2.1)
SAS	23.0 (0.8)	21.7 (0.7)	20.4 (0.7)
EAS	36.6 (1.3)	35.0 (1.3)	31.8 (1.2)

[a] number of unique variants, genome-wide (including autosomes and chromosome X) from LD pairs with R^2 greater or equal to a certain threshold
[b] number of unique variants on chromosome X

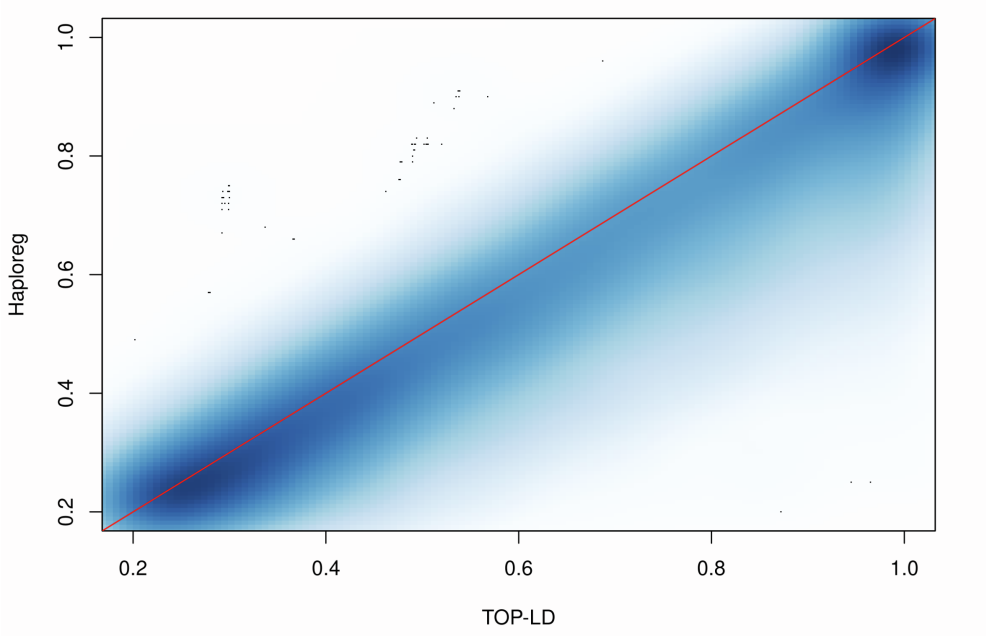

Figure S2.1: Smooth scatter plot of LD R^2 values from TOP-LD (x-axis) and Haploreg (y- axis) for pairs of variants with MAF>5% on chromosome 1 in European populations.

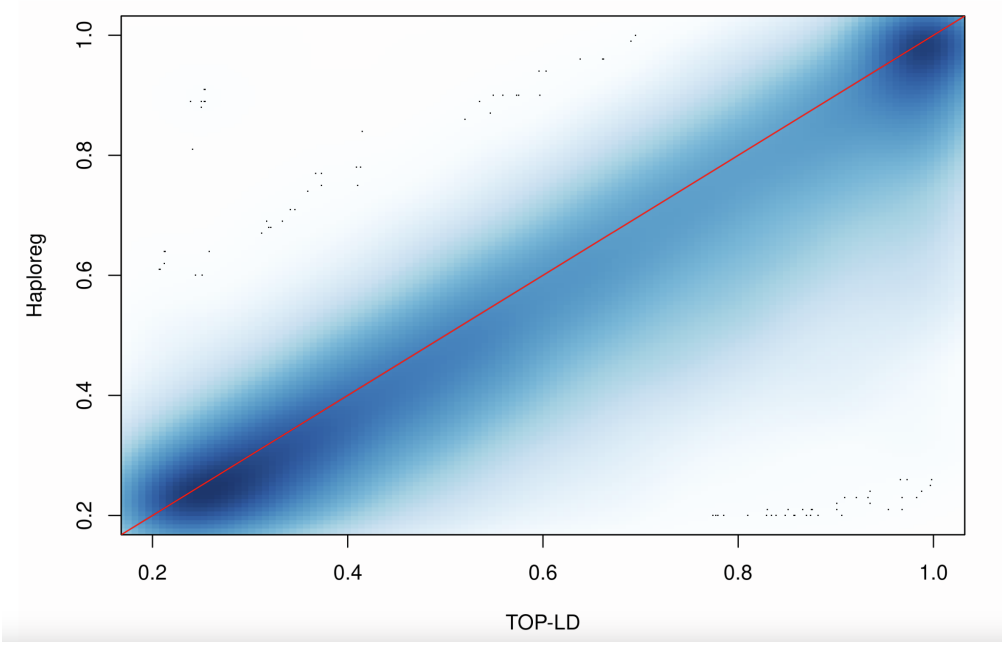

Figure S2.2: Smooth scatter plot of LD R^2 values from TOP-LD (x-axis) and Haploreg (y-axis) for pairs of variants with MAF>5% on chromosome 1 in African populations.

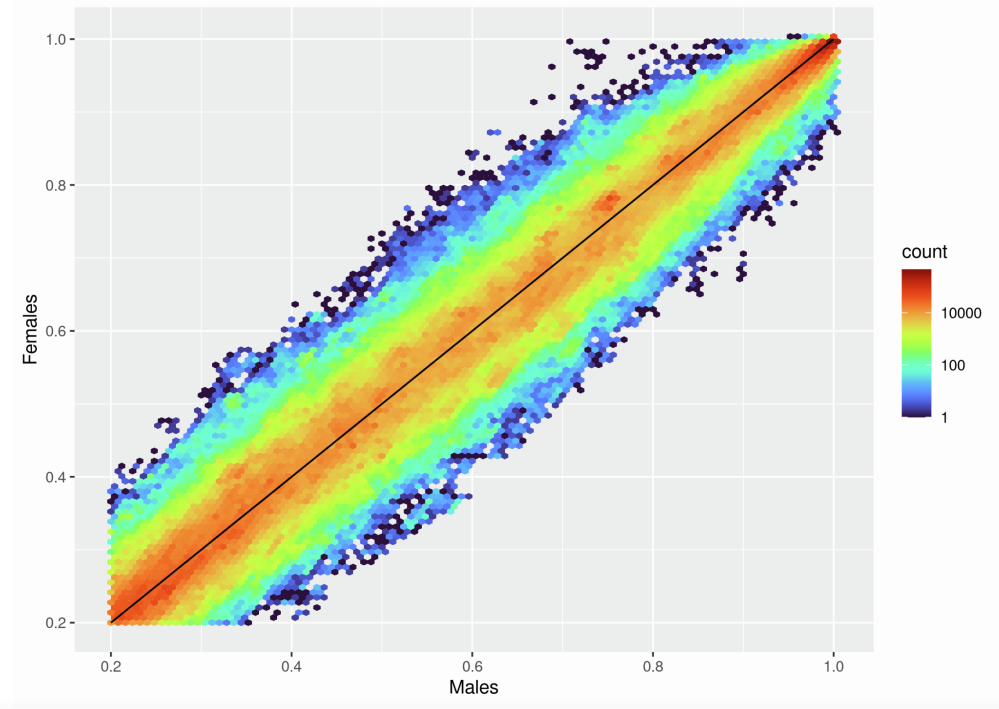

Figure S2.3: Hexbin plot of LD R^2 values between males (x-axis) and females (y-axis) between pairs of variants with MAF>5%, not in PAR1 or PAR2, on chromosome X in European populations

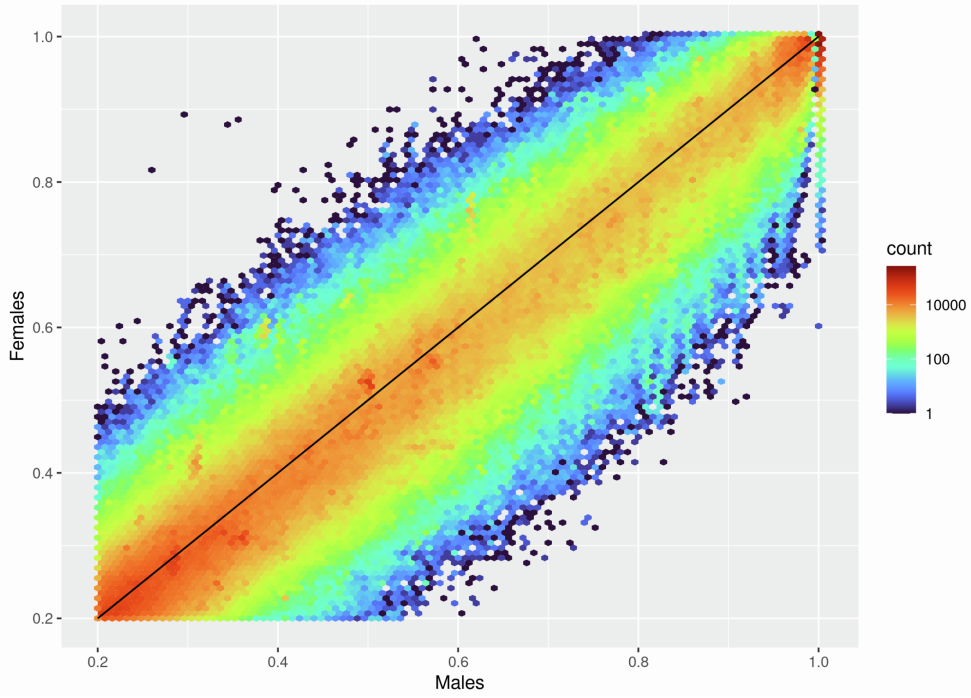

Figure S2.4: Hexbin plot of LD R^2 values between males (x-axis) and females (y-axis) between pairs of variants with MAF>5%, not in PAR1 or PAR2, on chromosome X in African populations.

CHAPTER 3: DEEPCOMPARE: A SYSTEMATIC EVALUATION OF HI-C DATA ENHANCEMENT METHODS FOR ENHANCING PLAC-SEQ AND HICHIP DATA

This chapter previously appeared as a paper in the Journal of Briefings in Bioinformatics. The original citation is as follows: Huang, Le*, Yuchen Yang*, Gang Li, Minzhi Jiang, Jia Wen, Armen Abnousi, Jonathan D. Rosen, Ming Hu, and Yun Li. "A systematic evaluation of Hi-C data enhancement methods for enhancing PLAC-seq and HiChIP data." Briefings in Bioinformatics 23, no. 3 (2022): bbac145. * indicates co-first authors who have contributed equally to this work.

3.1 Introduction

Mammalian genome folds into complex 3D structure in the nucleus, facilitating cis-regulatory elements to regulate genes up to megabase away (Li et al., 2018). The unbiased genome-wide Hi-C technology has been widely adopted for studying chromatin spatial organization (Lieberman-Aiden et al., 2009). However, Hi-C usually requires billions of reads to achieve kilobase (Kb) resolution, which is cost prohibitive (Rao et al., 2014a; Bonev et al., 2017a). Most existing Hi-C data are of $\sim$500 million or fewer raw reads, preventing subsequent Kb resolution analysis. To enhance Hi-C data, several computational methods, including HiCPlus (Zhang et al., 2018), HiCNN (Liu and Wang, 2019a), HiCNN2 (Liu and Wang, 2019b), DeepHiC (Hong et al., 2020), and VEHiCLE (Highsmith and Cheng, 2021) have been recently proposed. All five methods are based on deep neural network with different architectures. Specifically, HiCPlus uses three layers of convolution neural networks (CNN) (LeCun et al., 2015) to construct the mapping from low depth Hi-C data to high depth Hi-C data; HiCNN adopts a 54-layer CNN with skip connections (He et al., 2016); HiCNN2 extends HiCNN and ensembles three deep learning models (Liu and Wang, 2019b); DeepHiC utilizes generative adversarial networks (GAN) framework (Goodfellow et al., 2014); and VEHiCLE pre-trains a variational autoencoder (VAE) (Kingma and Welling, 2013) model

and fine-tunes a GAN model. This is an active research area with multiple more recent methods developed for enhancing Hi-C data (Hu and Ma, 2021b).

In 2016, HiChIP and PLAC-seq technologies (Mumbach et al., 2016; Fang et al., 2016a) were proposed to measure protein-mediated chromatin interactions. While offering higher signal-to-noise ratio and better cost-efficiency over genome-wide unbiased Hi-C data, HP data are still sparse at Kb resolution with the current sequencing depth of typically several hundred million raw reads per sample. Computationally enhancing the depth of HP data can facilitate downstream analysis, such as identification of long-range enhancer-promoter interactions, and prioritization of putative casual genes of genetic variants associated with human complex diseases and traits. No method has been developed for HP data enhancement and the aforementioned methods developed for Hi-C data (i.e., HiCPlus, HiCNN, HiCNN2, DeepHiC, and VEHiCLE) have not been evaluated for their performance on HP data yet.

To benchmark the performance of these methods when applied to HP data, we conducted systematic evaluation using seven publicly available HP datasets, namely Smc1a HiChIP data from the human lymphoblastoid cell line GM12878 (Mumbach et al., 2016), and H3K4me3 PLAC-seq data from five cell types including the mouse embryonic stem cells (mESC) (Juric et al., 2019) and four human fetal brain cell types (Song et al., 2020), and mESC CTCF PLAC-seq data (Juric et al., 2019). We focused on three aspects in our evaluation: A) the relative performance among the assessed methods; B) whether training with HP data leads to improved performance than training with Hi-C data; and C) transferability of the trained models across datasets, here specifically referring to the two cell lines GM12878 and mESC.

3.2 Results

3.2.1 Overview of the evaluation framework

In this study, we mainly applied three existing methods (HiCNN2, HiCPlus, and DeepHiC) designed for Hi-C data to enhance the sequencing depth of HP data (Figure 3.1). We also explored HiCNN and VEHiCLE but chose not to include them for most assessments because HiCNN has highly similar performance as HiCNN2 (Supplementary Figures S3.1-S3.2); and VEHiCLE's

specific features tailored for Hi-C data render it sub-optimal for HP data (Supplementary Figure S3.3). First, we generated low depth HP ("Baseline") datasets from mESC H3K4me3 PLAC-seq data and GM12878 Smc1a HiChIP data with different down-sampling ratios (Methods 1. Data preprocessing). We similarly generated low depth mESC CTCF PLAC-seq data and H3K4me3 PLAC-seq data for four human brain cell types with down-sampling ratio 0.125. Next, we split each dataset into training and testing datasets with the training dataset consisting of chromosomes 1, 2, 3, 5, 7 and 9 (chromosome 2 was used as the validation data, as part of the training procedure, to select the best model); and the testing dataset containing all the other chromosomes (chromosomes 4, 6, 8, 10-19 for mESC or chromosomes 4, 6, 8, 10-22 for human cell types). Then on the training datasets, we applied each method to train models using the low depth (i.e., baseline) input data and the high depth (i.e., full data without down-sampling) target data, and subsequently applied the trained models to the low depth testing data to obtain an enhanced high depth data (Figure 3.1). Finally, we calculated similarity between the enhanced HP data and the high depth HP data (i.e., full data without down-sampling for the testing chromosomes, which serves as the working truth). Specifically, we assessed similarity using four metrics: Pearson's correlation coefficient, Spearman's rank correlation coefficient, and Brownian distance covariance (Székely and Rizzo, 2022) (or distance correlation). For presentation brevity, we have decided to only show the Pearson correlation coefficient results in the main text as the other statistics reach qualitatively same conclusions. In addition, we performed 3D peak calling before and after enhancement to assess the impact of enhancement on the detection of chromatin interactions.

Next, we compared the performance of each method using the model trained on HP data to that trained on Hi-C data (detailed in later section Hi-C or HP data for training). Lastly, we evaluated the transferability of each method by enhancing HP data across different cell types detailed in later section Model transferability).

Since HP data measure protein-mediated chromatin interactions, we evaluated enhancement results only for bin pairs where at least one bin contains the protein of interest. Specifically, we defined bin pairs where both bins contain the protein of interest as the "AND" set, and bin pairs

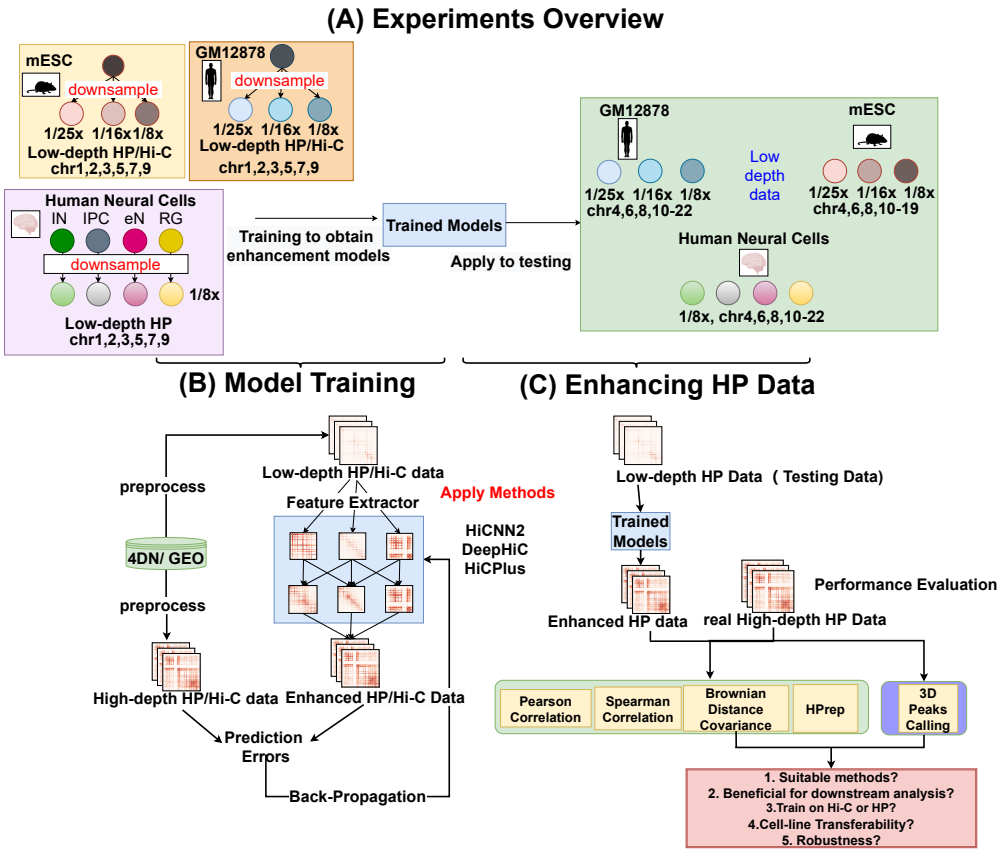

Figure 3.1: **Overview of experimental design**. (A) Experiments Overview: we applied each deep learning method to Hi-C or HP dataset from GM12878 cell line, mESC, and four different human neural cell types to train enhancement models. We then applied those models to enhance the testing datasets. The yellow, orange, and purple blocks on the left side represent training datasets (trained on chromosome 1,3,5,7,9 and validated on chromosome 2) from mESCs, GM12878, and human brain cells, respectively. The tree structure in each block contains high depth datasets (which are the target datasets, shown as parental nodes in darker colors) and low depth datasets (which are the input datasets, shown as offspring nodes in lighter colors). Note that the low depth datasets were created by down-sampling from high depth datasets (see details in Methods). On the right side, the green block represents the testing datasets. (B) Model Training: This panel shows the overall training procedure. The deep learning models learn the features (blue block) which can enhance the sequencing depth of input dataset. The loss (prediction error) measures the difference between an estimated value and its true value, and the gradient of loss can optimize the parameters of the neural networks (see "The principle of Deep Learning" section under Methods). (C) Enhancing HP data. This panel shows that we first applied pre-trained models on testing datasets and then evaluated the performance of each model by comparing the enhanced datasets (prediction) with their corresponding high depth datasets (ground truth) with four metrics (Pearson correlation coefficient, Spearman's rank correlation coefficient, and Brownian distance covariance (Székely and Rizzo, 2022)). We additionally evaluated the impact of enhancement on 3D peak calling.

where only one bin contains the protein of interest as the "XOR" set, following our previous work (Juric et al., 2019). We removed bin pairs where neither of two bins contains the protein of interest

from our downstream analysis, and only applied abovementioned similarity metrics (Pearson's correlation, Spearman's correlation, Brownian distance and HPrep) to bin pairs in the "AND" and "XOR" sets. Noticeably, in HP data, bin pairs in the "AND" set usually show higher contact frequency than bin pairs in the "XOR" set due to double ChIP enrichment. We thus evaluated similarity for the "AND" and "XOR" sets separately.

3.2.2 Performance comparison of different methods

We benchmarked the performance of the three methods (HiCPlus, HiCNN2, and DeepHiC) in terms of enhancing low depth HP data at 10Kb resolution. Note that we decided not include HiCNN for most evaluations because HiCNN and HiCNN2 perform highly similarly (Supplementary Figures S3.1-S3.2). For HiCNN2, which is an ensemble method with three models (HiCNN2-1, HiCNN2-2, and HiCNN2-3), we present only HiCNN2-1 for the rest of the manuscript because the three HiCNN2 models perform almost indistinguishably (Supplementary Figures S3.1-S3.2). Evaluations of the three methods (namely HiCPlus, HiCNN2-1, and DeepHiC) suggest that they perform reasonably, all significantly outperforming the low depth HP data (Figure 3.2, Supplementary Figures S3.4-S3.5). For example, when using HiCNN2-1 to enhance the GM12878 HiChIP data by 25x (i.e., from 0.04 depth to full), Pearson correlation coefficients for the "AND" set are 0.70-0.81, which are 0.09-0.24 higher than the low depth data, when the genomic distance is 20-250Kb (Figure 3.2A). Similarly, when using HiCPlus to enhance the mESC PLAC-seq data by 25x, Pearson correlation coefficients for the "XOR" set are 0.44-63, 0.16-0.21 higher than the low depth data, when the genomic distance is 50Kb-500Kb (Supplementary Figure S3.5). VEHiCLE shows inferior performance when enhancing GM12878 HiChIP data (Supplementary Figure S3.3) possibly due to certain features tailored for Hi-C data that are no longer suitable for HP data. For instance, VEHiCLE performs KR-normalization for Hi-C data. However, because HP technologies enrich chromatin contacts at the region mediated by the protein of interest, the equal visibility assumption made in KR-normalization is invalid for HP data. In addition, VEHiCLE requires all diagonal bin pairs to have non-zero counts. Therefore, we decided not to pursue further with VEHiCLE enhancement.

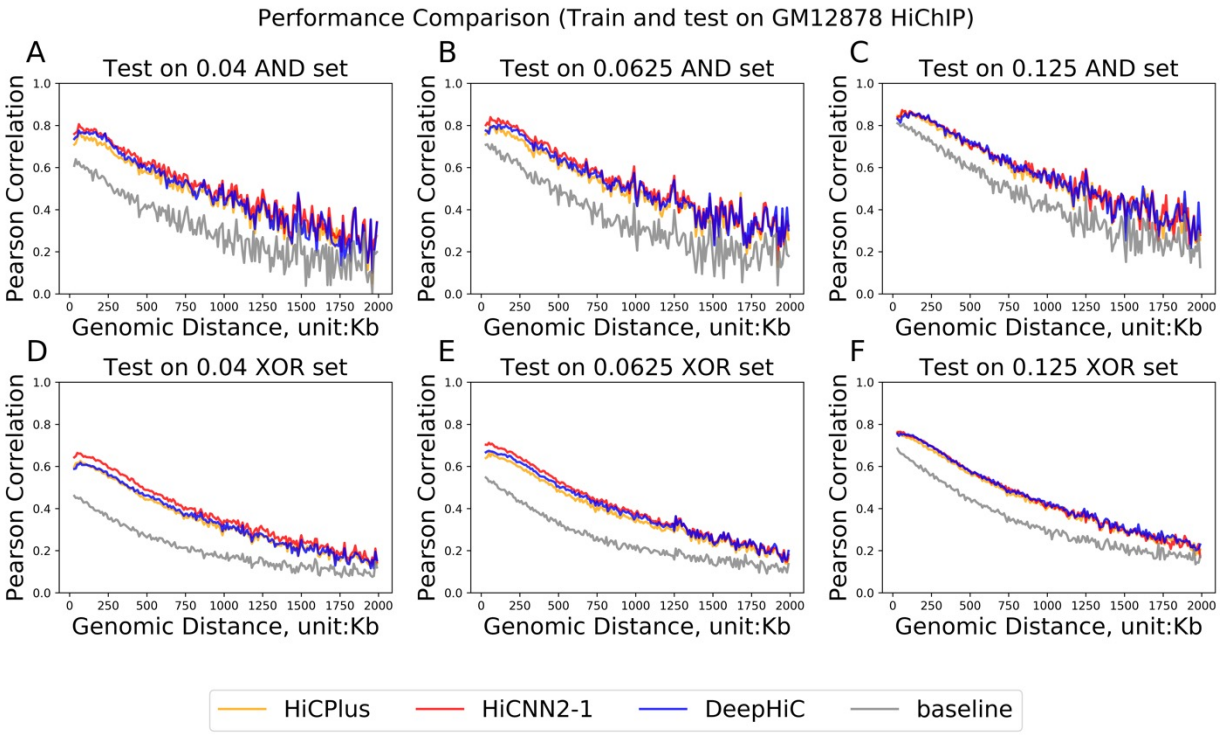

Figure 3.2: Methods comparison when enhancing GM12878 HiChIP data. Three enhancement methods are compared: HiCPlus, HiCNN2-1, and DeepHiC. Left panel (A and D) shows performance in 0.04 down-sampled data, middle panel (B and E) in 0.0625 down-sampled data, and right panel (C and F) in 0.125 down-sampled data. Performance is quantified with Pearson correlation coefficient (Y-axis). X-axis is genomic distance in Kb unit. Top row (A-C) shows performance among bin pairs in the AND set and bottom row (D-F) shows performance among bin pairs in the XOR set. The gray line represents the baseline (i.e., low depth data without any enhancement).

In addition, the three methods perform similarly for most of the seven HP datasets evaluated, with Pearson correlation differences largely within a difference of 0.1 (Figure 3.3-3, Supplementary Figures S3.4-S3.5). For example, when down-sampling ratio is 1/25 and the distance is 20Kb-1.25Mb for the GM12878 HiChIP data, HiCNN2-1 improves Pearson correlation by 0.024-0.048 and 0.012-0.059 on the "XOR" set, compared to HiCPlus and DeepHiC, respectively (Figure 3.2D). For another example, when down-sampling ratio is 1/16 and the distance is 250Kb-1.5Mb for the mESC PLAC-seq data, HiCNN2-1 and HiCPlus show highly similar performance and improve Pearson correlation by 0.01-0.09 on the "AND" set, compared to DeepHiC (Supplementary Figure S3.5B). When enhancing some cell types, for instance radial glia (RG) and intermediate progenitor cells (IPC), DeepHiC is substantially worse than HiCNN2-1 and HiCPlus with a difference of 0.2

in Pearson correlation (Figure 3.3C-D, H-I). The inferior performance of DeepHiC may be due to mode collapse issues (Salimans et al., 2016a; Srivastava et al., 2017) for GAN models (more details in Discussion section).

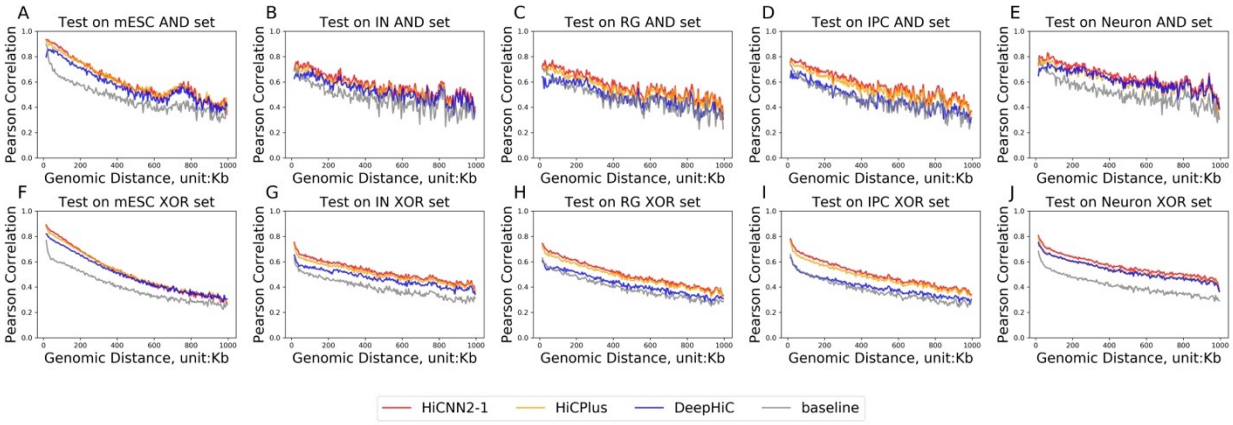

Figure 3.3: Three enhancement methods are compared: HiCPlus, HiCNN2-1, and DeepHiC. Down-sampling ratio is 0.125 for all five cell types evaluated: mESC (1st column), interneurons (IN, 2nd column), radial glia (RG, 3rd column), intermediate progenitor cells (IPC, 4th column), and excitatory neurons (eN, 5th and rightmost column). Performance is quantified with Pearson correlation coefficient (Y-axis). X-axis is genomic distance in Kb unit. Top row (A-E) shows performance among bin pairs in the AND set and bottom row (F-J) shows performance among bin pairs in the XOR set. The gray line represents the baseline (i.e., low depth data without any enhancement).

3.2.3 3D peak calling

3D peak calling, or the detection of statistically significant long-range chromatin interactions, is one of the important downstream analyses for various types of chromatin conformation data, including HP data. To evaluate the impact of HP data enhancement, we further applied our MAPS pipeline (Juric et al., 2019) to identify significant chromatin interactions before and after enhancement and compared them with chromatin interactions detected from the full data. We treated 3D peak calling results derived from the full data (without any down-sampling) as the truth. Specifically, we defined true peaks as bin pairs with MAPS FDR $< 1\%$, contacts ≥ 12, and signal to noise ratio (SNR, i.e., the ratio of observed count over expected count) ≥ 2; and we defined true background bin pairs as those with MAPS FDR $> 10\%$ and contacts ≥ 12. We found that even high-depth input HP data (e.g., 0.5 down-sampled GM12878 HiChIP data in Figure 3.4, or 0.5 down-sampled mESC PLAC-seq data in Supplementary Figure S3.6, where the raw

sequencing depth is ∼322 million and ∼568 million, respectively) benefit from enhancement in that enhanced datasets can improve power of 3D peak calling. For example, for 0.5 down-sampled GM12878 HiChIP data (Figure 3.4), baseline (i.e., down-sampled data before enhancement) has a low sensitivity of 0.32, while the enhanced data (using HiCPlus or HiCNN2) improve sensitivity to 0.71-0.72, while maintaining the desired FDR 1%. DeepHiC increases sensitivity even more drastically but fails to maintain the desired 1% FDR. Similarly, we observed clear improvement with enhanced datasets for 0.5 down-sampled mESC PLAC-seq data (Supplementary Figure S3.6). Observing that the FDRs from baseline is essentially 0, we relaxed the MAPS-FDR threshold to 0.2 for the baseline, which led to an actual FDR of 0.02, comparable to that after enhancement. With the relaxed FDR threshold, the power of the baseline increased substantially, from 0.32 to 0.6 but still clearly lower than >0.71 after enhancement (Figure 3.4). Similarly, we observed sensitivity increasing for baseline but still lower than enhanced data for 0.5 down-sampled mESC PLAC-seq data (Supplementary Figure S3.6).

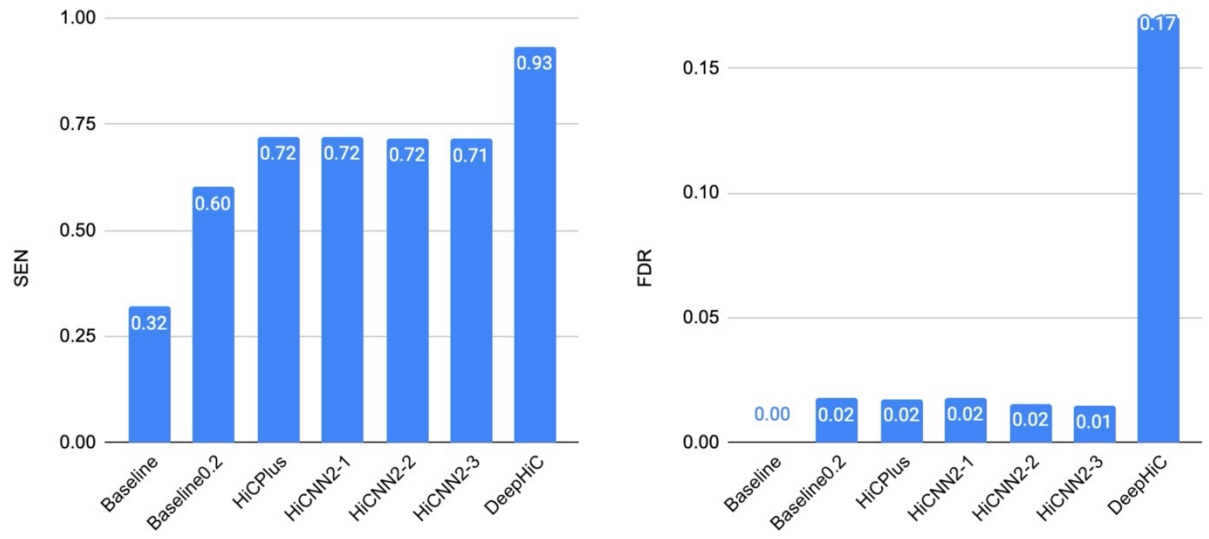

Figure 3.4: **3D peak calling in 0.5 down-sampled GM12878 HiChIP data**. Left panel (A) shows sensitivity (SEN). Right panel (B) shows FDR. The truth (3D peaks or not) is established by peak calling via MAPS from full data without any down-sampling. Specifically, true peaks are bin pairs with MAPS FDR $< 1\%$, contacts ≥ 12, and signal to noise ratio (SNR, i.e., the ratio of observed count over expected count) ≥ 2. True background bin pairs are those with MAPS FDR $> 20\%$ and contacts ≥ 12. Baseline0.2 bars show the 3D peak calling performance when relaxing MAPS-FDR threshold from 1% to 20%.

We additionally examined signal-to-noise ratio (SNR, again defined as the ratio of observed count over expected count) before and after enhancement, both compared to SNRs from the full data without enhancement. We found that SNR estimates from baseline data without enhancement are significantly lower than those from full data (Supplementary Figure S3.7B). Treating the estimates from full data as the working truth, these results indicate that baseline data tend to under-estimate the magnitude of 3D peaks. Data enhancement mitigates the under-estimation issue, with enhanced data producing SNR estimates more closely approaching the working truth (Supplementary Figure S3.7C). Although we observed significant difference in SNR estimates both at 3D peaks (Supplementary Figure S3.7E) and at background bin pairs (Supplementary Figure S3.7F), we noticed that the absolute difference is more pronounced among 3D peaks. Specifically, mean and median SNR at 3D peaks are 4.43 and 4.07 at baseline, 4.90 and 4.37 after HiCNN2-1 enhancement, and 4.99 and 4.38 when using the full data (Supplementary Figure S3.7E). In contrast, the mean and median SNR at background bin pairs are 0.97 and 0.94, 0.98 and 0.97, 1.01 and 0.99 respectively, with only ≤ 0.05 absolute difference. The statistical significance at background bin pairs is driven primarily by the huge number of background bin pairs (Supplementary Figure S3.7F).

Encouraged by the power improvement in 3D peak calling genome-wide, we proceeded to examine two specific loci in mESCs where previous studies (Schoenfelder et al., 2015; Zhou et al., 2013b) have established enhancer-promoter (E-P) interactions. These two loci are *Med13l* and *Mtnr1a* loci (Figure 3.5). From Figure 3.5, we observe that baseline without enhancement fails to identify many 3D peaks, including the most important E-P interactions. After HP data enhancement, we were able to rescue some of the E-P interactions. For example, for the two bin pairs corresponding to E-P interactions at the Med13l locus (illustrated with black arrows), full data identified both; baseline identified only one; while every enhanced data was able to rescue the missed one (Figure 3.5A-E). Similarly, for the bin pair corresponding to E-P interaction at the Mtnr1a locus (illustrated with black arrows), full data identified it; baseline failed to detect it, while again every enhanced data managed to rescue the signal (Figure 3.5G-K). Interesting, simply

multiplying the baseline matrix with a constant of 8 can rescue the E-P interactions at the Med13l locus (Figure 3.5F), suggesting that simple amplification of the contact frequency matrix may help 3D peak calling when the input data is of low-depth. However, this simple strategy still fails to detect the E-P interaction at the Mtnr1a locus (Figure 3.5L), showcasing the advantage of data enhancement.

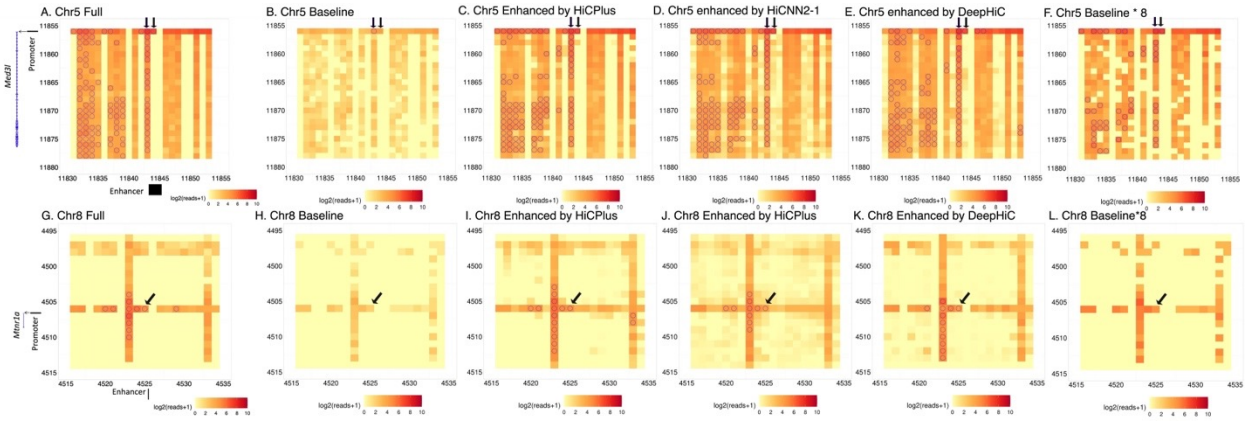

Figure 3.5: **3D peak calling at *Med13l* and *Mtnr1a* loci**. 3D peaking calling results from MAPS are shown. Top panel (A-F) is for the *Med13l* locus and bottom panel (G-L) is for the Mtnr1a locus. From left to right, we show MAPS peak calling results from the full data (without any down-sampling), baseline (0.125 down-sampled mESC data without enhancement), HiCPlus enhanced data, HiCNN2-1 enhanced data, DeepHiC enhanced data, and baseline*8 (by simply multiplying the baseline matrix with a constant 8). 3D Peaks, bin pairs with MAPS FDR $< 1\%$, are indicated by blue circles. For the full data (leftmost column), we further require contacts ≥ 12, while for baseline and enhanced data, we relax the criterion to contacts ≥ 2. The gene track is shown on the very left margin and the enhancer regions are shown at the bottom of the left panel as black rectangles. Gene and enhancer information are visualized with the help of WashU Epigenome Browser (Zhou et al., 2013b) Bin pairs corresponding to the annotated enhancer-promoter regions are marked by black arrows.

Finally, we assessed whether the identified chromatin interactions relate to gene expression. As shown in Supplementary Figure S3.8, we found that genes with promoters involving 3D peaks show significantly higher expression levels than genes whose promoters do not involve in any 3D peaks. The fact that genes with 3D peaks identified from baseline data without any enhancement is expected as they tend to be the lower hanging fruits with stronger magnitude of chromatin interactions that can be detected by low depth data.

3.2.4 Hi-C or HP data for training?

We then evaluated the robustness and relative performance of HP depth enhancement for each method when training model with different assays, specifically Hi-C or HP. For enhancing the GM12878 HiChIP data, HiCPlus, HiCNN2, and DeepHiC all showed comparable or improved enhancement by using models trained on HP data than trained on Hi-C data (Figure 3.6; Supplementary Figures S3.9-S3.11). For example, when enhancing the GM12878 HiChIP data by 8× (i.e., enhancing 1/8 down-sampled data to full data), Pearson correlation coefficients using HiCPlus model trained on HP data improved by up to 0.11 for the "AND" set and 0.04 for the "XOR" set, compared to models trained on Hi-C data (distance: 250Kb-2Mb, Figure 3.6A, Figure 3.6D). Similarly, HiCNN2 and DeepHiC models trained on HP data demonstrated overall improved or comparable performance than those trained on Hi-C data, with more obvious improvement than HiCPlus. For example, within 250Kb-2Mb distance, HiCNN2 (Figure 3.6B and 3.6E) and DeepHiC (Figure 3.6C and 3.6F) improve Pearson correlation coefficient by 0.15 and 0.18 for the "AND" set, and 0.08 and 0.11 for the "XOR" set, compared to the aforementioned 0.11 and 0.04 for HiCPlus (Figure 3.6A and 3.6D).

However, when enhancing the mESC PLAC-seq data, we observed mixed results using models trained on HP data vs. those trained on Hi-C data, (Supplementary Figures S3.12-S3.14). Specifically, HiCPlus showed similar performance using HP data (light yellow) or Hi-C data (dark yellow) for training (Supplementary Figures S3.12-S3.14 left panels); HiCNN2-1's HP trained models (light red) outperformed its Hi-C trained models (dark red) (Supplementary Figures S3.12-S3.14 middle panels) while DeepHiC's HP trained models (light blue) were inferior to its Hi-C trained models (dark blue) (Supplementary Figures S3.12-S3.14 right panels).

One possible explanation for DeepHiC's better performance of mESC Hi-C-trained models is the much higher sequencing depth of mESC Hi-C data relative to mESC PLAC-seq data. Specifically, mESC Hi-C data is 4.59x that of mESC PLAC-seq data, in terms of informative reads (Table 3.1). Such drastic depth difference could render the models trained on Hi-C data more advantageous than those trained on HP data for all three methods (Supplementary Figures S3.12-S3.14. Particularly

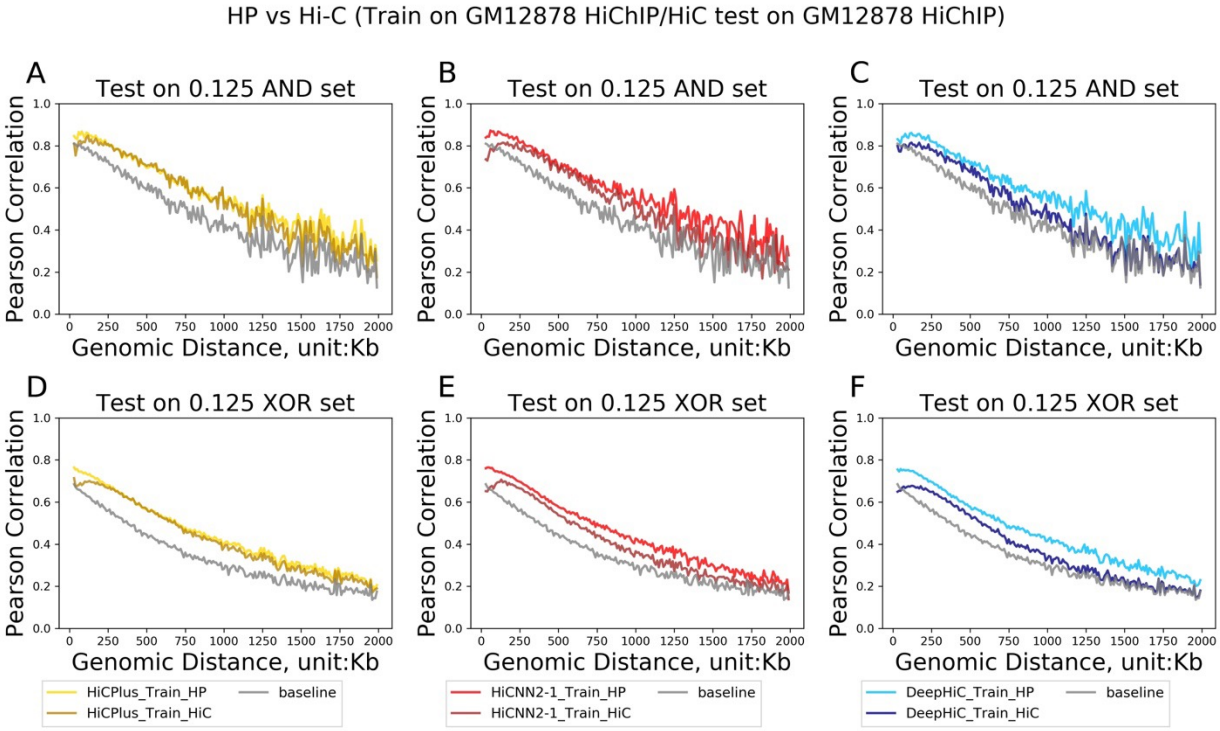

Figure 3.6: **HiChIP trained vs Hi-C trained models when enhancing GM12878 HiChIP data by 8×.** Performance is assessed by Pearson correlation coefficient. Each subfigure represents the performance of one of three read depth enhancement methods (HiCNN2-1, HiCPlus, and DeepHiC) for a certain set (AND set or XOR set). In each subfigure, we show how Pearson correlation coefficient (Y-axis) changes with genomic distance (X-axis), where the distance ranges from 20Kb-2Mb with an increment of 10Kb. The gray line represents the baseline (i.e., low depth data without any enhancement).

in DeepHiC, we observed obvious advantage of Hi-C trained models over HP trained models. To reduce the impact of the different sequencing depths, we down-sampled mESC Hi-C data so that its informative reads are comparable to those in mESC PLAC-seq data (Methods section 3 Generating HiC_Downsampled data). After down-sampling, we obtained 59.2 million informative reads in down-sampled Hi-C data (HiC_downsampled), matching that (also 59.2 million) in PLAC-seq data (Table 3.1). We then re-trained DeepHiC models using the down-sampled Hi-C data. With comparable number of informative reads, models trained on the down-sampled Hi-C data showed worse or comparable performance than those trained on HP data (Supplementary Figure S3.15). Although worse performance is expected, the magnitude of performance impairment is drastic. For example, when enhancing by 8× (Middle panel) within distance 500Kb-1Mb, the Pearson correlation is 0.22-0.37 with DeepHiC models trained on the down-sampled Hi-C data, compared

to 0.50-0.67 with models trained on HP data, and 0.55-0.76 with models trained on full Hi-C data. These results suggest that DeepHiC method is more sensitive to the sequencing depth of Hi-C data than HiCNN2 and HiCPlus.

Table 3.1: Read counts for HP and Hi-C dataset

Down-sampling ratio	**None**		**0.04**	**0.0625**	**0.125**	**0.25**	**0.5**
	#raw reads in full data	**# informative reads[1] in full data**	**# informative reads[1] in down-sampled data**				
GM12878 HiChIP#	643,644,994	28,762,260	1,149,807	1,797,116	3,593,523	7,191,222	14,187,319
mESC H3K4me3 PLAC-seq#	1,135,198,787	59,229,165	2,370,623	3,700,378	7,404,377	14,807,009	29,613,854
mESC CTCF PLAC-seq#	345,816,091	17,372,561	694,895	1,085,777	2,171,563	4,343,135	8,686,277
IN H3K4me3 PLAC-seq*	2,747,206,906	13,387,780	535,499	836,728	1,673,465	3,346,938	6,693,884
IPC H3K4me3 PLAC-seq*	1,837,960,692	15,171,775	606,862	948,226	1,896,464	3,792,937	7,585,882
eN H3K4me3 PLAC-seq*	1,740,000,000	20,547,587	821,895	1,284,213	2,568,439	5,136,889	10,273,789
RG H3K4me3 PLAC-seq*	1,487,624,144	14,180,735	567,217	886,285	1,772,582	3,545,177	7,090,363
GM12878 Hi-C	6,524,520,477	256,378,089	10,257,207	16,023,348	32,057,797	64,087,648	128,177,761
mESC Hi-C	7,260,480,082	272,146,960	10,885,391	17,002,222	34,005,844	68,045,585	136,067,650
GM12878 ratio (Hi-C/HP)	NA	8.914	8.921	8.916	8.921	8.912	9.035
mESC ratio (Hi-C/HP)	NA	4.595	4.592	4.595	4.593	4.595	4.595
mESC Hi-C_downsampled#	NA	59,228,684	2,369,794	3,700,223	7,405,280	13,369,886	26,757,578

3.2.5 Model transferability

Although many HP datasets have been generated recently (Mumbach et al., 2016; Juric et al., 2019; Song et al., 2020), deeply sequenced HP datasets are only available to limited cell types, making it infeasible to train models separately for each cell type. One potential solution is to use models pre-trained on available datasets from other cell type(s).

For enhancing the GM12878 HiChIP data, HiCPlus performed similarly with either GM12878-trained or mESC-trained models (the left panel of Figure 3.7; the left panels of Supplementary Figures S3.16-S3.18). Comparatively, HiCNN2-1 and DeepHiC showed slightly higher or higher accuracy using GM12878-trained models than mESC-trained models (the middle and right panels of Figure 3.7), with the difference more obvious for DeepHiC (the right panel of Figure 3.7).

For enhancing the mESC PLAC-seq data, the performance of HiCPlus, HiCNN2 and DeepHiC using models trained on the GM12878 HiChIP data was comparable to or even slightly better than using models trained on the mESC PLAC-seq data (Supplementary Figures S3.19-S3.21). Specifically, for HiCPlus and HiCNN2-1, the two sets of models were nearly indistinguishable in terms of enhancing the mESC PLAC-seq data (left and middle panels of Supplementary Figures S3.19-S3.21). Interestingly, DeepHiC achieved even slightly better performance when using models

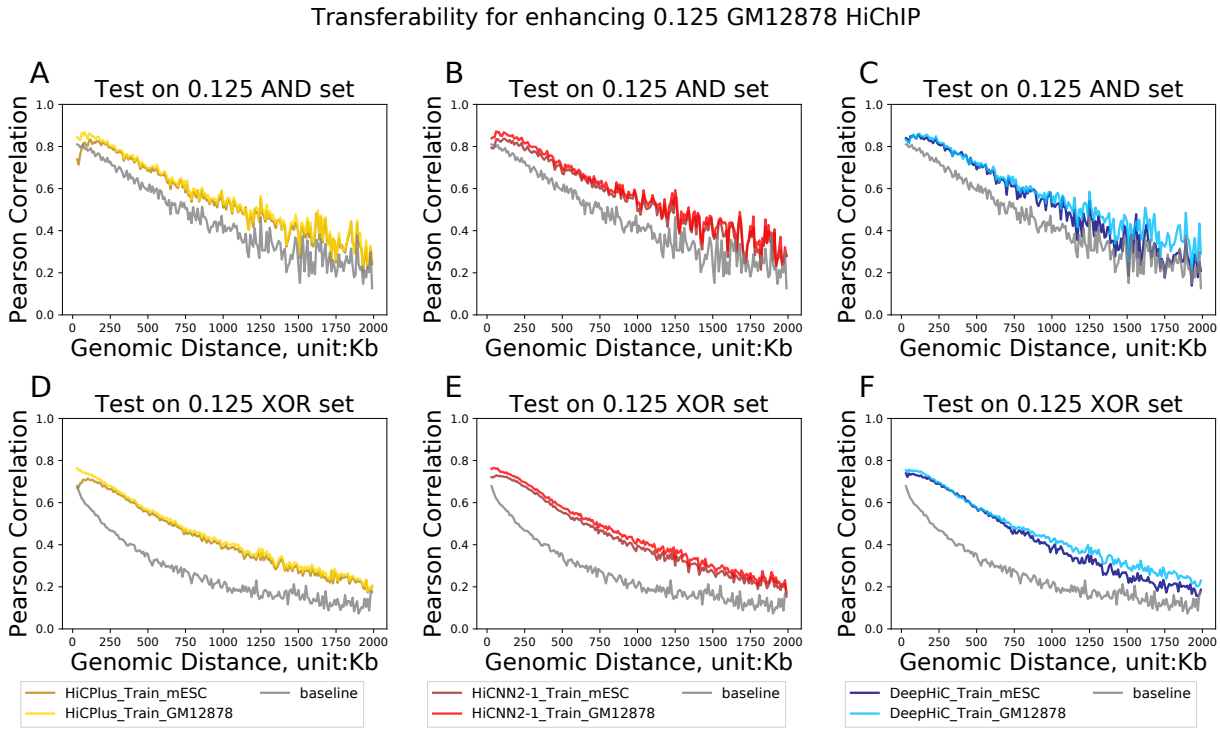

Figure 3.7: Model transferability when enhancing GM12878 HiChIP data by 8x. Each subfigure compares two enhanced GM12878 HiChIP data: one using models trained with GM12878 HiChIP data (Train_GM12878) and the other using models trained with mESC PLAC-seq data (Train_mESC). The evaluation metric is Pearson correlation coefficient. Different colors in the subfigures represent different methods (yellow: HiCPlus, red: HiCNN2-1, blue: DeepHiC) while darker color represents models trained with mESC PLAC-seq and lighter color represents models trained with GM12878 HiChIP data. In each subfigure, we show how the evaluation metric (Y-axis) changes with genomic distance (X-axis), where the distance ranges from 20Kb-2Mb with an increment of 10Kb. The gray line represents the baseline (i.e., low depth data without any enhancement).

trained on the GM12878 HiChIP data (right panel of Supplementary Figures S3.19-S3.21). One plausible reason is that the models for all three methods are originally developed and fine-tuned for the GM12878 Hi-C data.

Encouraged by the promising transferability results between mESC H3K4me3 PLAC-seq data and GM12878 Smc1a HiChIP data, we proceeded with transferability assessment across more cell types. Since HiCNN2-1 and HiCPlus achieved similarly best transferability performance, we presented only HiCNN2-1 results for brevity. Specifically, we enhanced 0.125 down-sampled HP data for each of the six cell types (namely GM12878 and mESC, and four human brain cell types (Song et al., 2020) including radial glia [RG], intermediate progenitor cells [IPC], excitatory neurons

[eN], and interneurons [IN]) using models trained from the corresponding cell type as well as using models trained from each of the other five cell types. For training and testing, we used the same chromosome splitting as illustrated in Figure 3.1. Results shown in Figure 3.8 and Supplementary Figures S3.22-S3.23 further support that the models learned are transferable across cell types. Specifically, Pearson correlation coefficients are almost indistinguishable when enhancing with models trained from the matching cell type or from other cell types (Figure 3.8 and Supplementary Figure S3.22), and all the models lead to similar performance in 3D peak calling (Supplementary Figure S3.23). For example, to detect chromatin interactions in IN, 0.125 down-sampled PLAC-seq data (before any enhancement) had essentially no power at all (sensitivity to detect IN of IN-specific 3D peaks is 0.00, left most bars labeled "Baseline" in Supplementary Figure S3.23C and S3.23G); in contrast, enhanced RG data using models trained with IN data resulted in a sensitivity of 0.49 (or 0.48) for IN (or IN-specific) 3D peaks (magenta bars in Supplementary Figure S3.23C and S3.23G); similarly and importantly, enhanced IN data using models trained with data in any of the other five cell types resulted in comparable sensitivity of 0.47-0.59 (or 0.43-0.63) for IN (or IN-specific) 3D peaks (blue bars in Supplementary Figure S3.23C and S3.23G).

We further evaluated transferability in terms of capturing cell-type-specific features, examining gene expression and open chromatin status in the corresponding cell types. Specifically, we compared distribution of gene expressions for three groups of genes: (1) genes with 3D peak(s) looping to their promoters identified at "baseline" (low-depth data without enhancement); (2) genes without any 3D peaks at "baseline" but with 3D peak(s) after enhancement, separately for enhanced data using models trained with each of the six cell types; and (3) genes without any 3D peaks even with the full data ("background"). Not surprisingly, as shown in Supplementary Figure S3.24 A-D, "baseline" identified only few lower-hanging fruits, and thus expression levels are the highest; genes with 3D peaks identified only after enhancement (whether using models trained with the matching cell type [yellow boxplots] or different cell types [non-baseline and non-background cyan boxplots]), reassuringly, had only slightly lower expression levels, drastically higher than those "background" genes. Similar patterns are observed when restricting only to cell-type-specifically

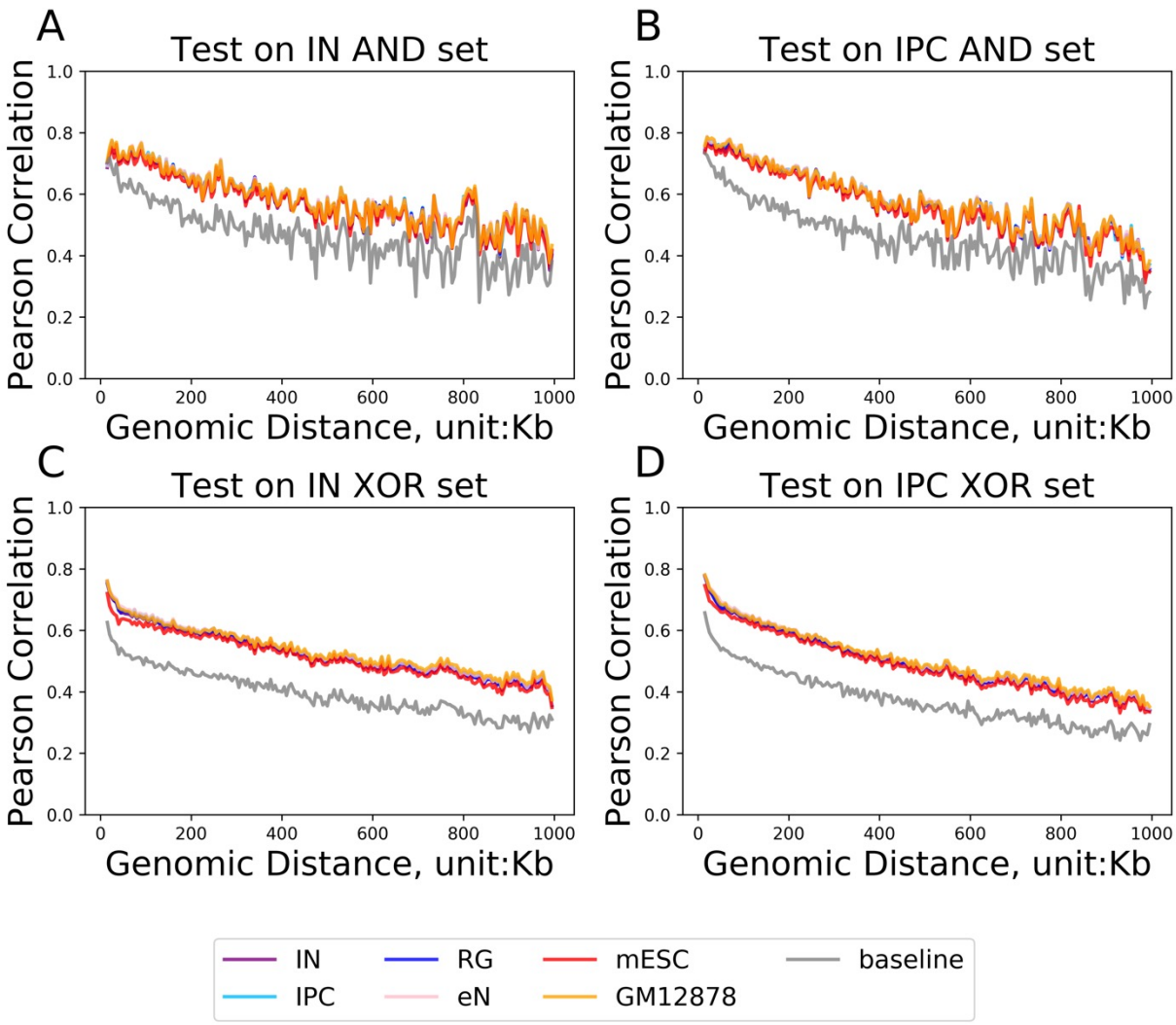

Figure 3.8: Model transferability when enhancing two neural cell types. All results are from HiCNN2-1 models. We test (i.e., perform enhancement) on two cell types: interneurons (IN, left sub-figures A and C) and intermediate progenitor cells (IPC, right sub-figures B and D), with down-sampling ratio 0.125. The enhancement models are trained using HP data from each of the following six cell types: IN, IPC, radial glia (RG), excitatory neurons (eN), mESC or GM12878. The gray line represents the baseline (i.e., low depth data without any enhancement).

expressed genes (Supplementary Figure S3.24 E-H). Following similar logic, we assessed 3D peaks in terms of their overlap with cell-type-specific ATAC-seq peaks, observing similar pattern (Supplementary Figure S3.24 I-L). In particular, models trained with matching cell type (yellow bars) resulted in similar proportion of overlap with cell-type-specific ATAC-seq peaks as those

trained with different cell types (non-baseline and non-background cyan bars), suggesting that models trained from un-matching cell types can similarly retain cell-type-specific features.

Finally, we explored model transferability across different proteins of interest by cross-applying models learned from H3K4me3 and CTCF PLAC-seq data in mESC. We used the same mESC CTCF PLAC-seq data as in Juric et al (Juric et al., 2019). We observed almost indistinguishable performance when using models trained with the same protein of interest or the other protein (Supplementary Figure S3.25), both visibly better than without enhancement (i.e., baseline). These results suggest that the enhancement models learned are likely transferable also across different proteins of interest, with the caveat that our assessments only involved three different proteins: Smc1a above for GM12878, CTCF here for mESC, and H3K4me3. In the future, more high-depth HP data with different proteins of interest will allow us to perform more comprehensive assessment across various proteins.

3.2.6 Model robustness

Throughout the manuscript so far, we have used chromosomes chr1, 3, 5, 7 and 9 as training; chromosome 2 as the validation (part of the training procedure to select the best model); and the remaining chromosomes as testing. In addition, when creating low-depth input data, we performed down-sampling only once. We evaluated model robustness by swapping training and training, by using leaving-one-chromosome-out, and by performing down-sampling five times. Results presented in Supplementary Figures S3.26-S3.27 show that the models trained are robust, resulting in highly similar Pearson correlation coefficient decay profiles.

3.3 Methods

3.3.1 Data Preprocessing

All our assessed deep learning methods require training data, testing data and validation data as input. In our study, for the mESC PLAC-seq data, we assigned chromosomes 1, 3, 5, 7 and 9 as the training data, chromosome 2 as the validation data, and chromosomes 4, 6, 8, 10-19 as the testing data. For the GM12878 HiChIP data, we assigned chromosomes 1, 3, 5, 7, and 9 as the training data, chromosome 2 as the validation data, and chromosomes 4, 6, 8, 10-22 as the testing

data. Here, validation data was used to select the best model (details in Methods 2. The principle of Deep Learning). Both mESC PLAC-seq data and GM12878 HiChIP data consist of two parts: high depth data and low depth data, referring to the original/full HP data without down-sampling and down-sampled data, respectively. We used the low depth data as the input for each deep learning method, and the high depth data to calculate the loss function (details in Methods 2. The Principle of Deep Learning).

Specifically, we applied the following steps to generate low depth data:

- Converting read pairs into bin pairs. We followed our previous study (Juric et al., 2019) to preprocess the HP data to retain only long-range read pairs (intra-chromosomal contacts >1Kb). We then randomly selected a subset of the read pairs with down-sampling ratios 1/25, 1/16 or 1/8. Down-sampling was implemented using command (Li et al., 2009):

```
samtools view -s ratio
```

  Next, we binned the original read pairs and down-sampled read pairs into 10Kb bin pairs, resulting in high depth data and low depth data, respectively.

- (2) Converting high depth and low depth bin pairs into contact matrices. For each chromosome, based on whether the bins containing the protein of interest (H3K4me3 ChIP-seq peaks for mESC PLAC-seq data (Juric et al., 2019), and Smc1a ChIP-seq peaks for GM12878 HiChIP data (Mumbach et al., 2016)), we further grouped bin pairs into three categories: the "AND" set (bin pairs where both bins contain the protein of interest), the "XOR" set (bin pairs where only one bin contains the protein of interest), and the "NOT" set (bin pairs where neither bins contains the protein of interest). Since HP technologies measure protein-mediated long-range chromatin interactions, we only focused on the "AND" and "XOR" sets for downstream analysis. In addition, we filtered out bin pairs with either end overlapping with the ENCODE blacklist regions (Amemiya et al., 2019) or with low mappability (mappability score < 0.9)

(Hu et al., 2012). After filtering, we created 10Kb bin resolution contact matrix for each chromosome. In this work, we only used 10Kb bin pairs with 1D genomic distance less than 2Mb in our analysis.

- Converting contact matrices into training data, testing data and validation data. According to the required format of each deep learning method, we split the contact matrix for each chromosome into multiple sub-matrices. Different deep learning methods adopt different splitting strategies as their default configuration. DeepHiC splits the high depth and low depth contact matrices into non-overlapping 40×40 sub-matrices with stride size 40×40. In contrast, HiCNN, HiCNN2, and HiCPlus partition the low depth data with overlapping 40x40 sub-matrices with stride size 34×34 (the overlapping region between two consecutive sub-matrices is 6×40). Next, HiCNN, HiCNN2, and HiCPlus partition the high depth data into non-overlapping 28x28 sub-matrices with stride size 28×28. The overlapping sub-matrices split by HiCNN, HiCNN2, and HiCPlus imply that all inferred regions (i.e., the 34×34 core regions) have flanking information. We applied each method with its default matrix splitting strategy. With those sub-matrices, we constructed three types of tensors: for training data, testing data, and validation data, respectively. Here, training data is the tensor concatenating data from five chromosomes (1, 3, 5, 7 and 9), validation data is a tensor of chromosome 2, and testing data contains data from chromosomes 4, 6, 8, 10-19 or chromosomes 4, 6, 8, 10-22, for the mESC PLAC-seq data or the GM12878 HiChIP data, respectively.

3.3.2 The Principle of Deep Learning

All three deep learning methods (HiCNN2, HiCPlus, and DeepHiC) evaluated in this study are supervised learning algorithms, which can be formulated as the following:

$$y = f_\theta(x) \tag{3.1}$$

where x represents the low depth data in the training dataset (see 3.3.1 Data Preprocessing), y represents the enhanced data, and θ represents the parameters of neural network $f(\cdot)$, which

approximates the mapping $f : x \mapsto y$ by learning from the training dataset. Each parameter θ_i ($i = 1, 2, \ldots, n$) (with n being the total number of parameters of the neural network) can be optimized by the gradient descent algorithm (e.g. stochastic gradient descent (SGD) or Adam (Kingma and Ba, 2014)) in equation 3.2,

$$\theta_i = \theta_i - r\frac{\partial}{\partial\theta_i}J(\theta) \tag{3.2}$$

where r is the learning rate, which controls the step size of gradient descending. $J(\theta)$ is the pre-defined loss function and $\frac{\partial}{\partial\theta_i}J(\theta)$ is the gradient of θ_i which is calculated by

Backpropagation algorithm (LeCun et al., 2015). In HiCNN2 and HiCPlus, $J(\theta)$ is the mean squared error as specified in equation 3.3,

$$J(\theta) = \frac{1}{m}\sum_{j=1}^{m}(f_\theta(x_j) - y_j)^2 \tag{3.3}$$

where m is the sample size and it is the product of batch size (hyperparameter), the width of the sub-matrix N, and the height of the sub-matrix N; $y = \{y_1, y_2, \ldots, y_m\}$ is a batch of target data (high depth sub-matrices, see 3.3.1 Data Preprocessing (3)), $y_j \in \mathbb{R}^{N\times N}$ is the high depth matrix; $x_j \in \mathbb{R}^{N\times N}$ represents the low depth matrix; $f_\theta(x_j) \in \mathbb{R}^{N\times N}$ is the enhanced matrix; $f_\theta(\cdot)$ is the neural network which represents the mapping f of x to y ($f : x \mapsto y$).

In HiCNN2 or HiCPlus, the training loss (represented by equation 3.3) is optimized by the gradient descent algorithm (Equation 3.2) iteratively. Each method trains the neural network using multiple epochs, with the default being 500 and 40,000 for HiCNN2 and HiCPlus respectively. One epoch involves passing all batches completely through the neural network. In each epoch, HiCNN2 or HiCPlus uses validation loss to evaluate whether to retain the current trained model or not. Specifically, the algorithm calculates the validation loss between full data (viewed as the target data) and the enhanced data using equation 3.3, and updates to the current model only when the validation loss decreases.

3.3.3 Generating HiC_downsampled data

In addition, we conducted additional experiments for Evaluation of transferability, where models were trained on the mESC Hi-C data with comparable sequencing depth as the mESC PLAC-seq data, which we referred to as HiC_downsampled data. We generated the HiC_downsampled data by down-sampling read counts within 2Mb genomic distance of mESC Hi-C data to 59.2 million, matching the total number of reads (59.2 million) in the "AND" and "XOR" sets of the corresponding mESC PLAC-seq data (Table 3.1).

3.4 Discussion

While several computational methods have been developed for enhancing the depth of Hi-C data, tools tailored for HP data depth enhancement are still lacking. In this study, we evaluated three methods (HiCPlus, HiCNN2 and DeepHiC) developed for Hi-C data, when applying them to enhance HP data. Our results showed that all three methods performed similarly on enhancing HP datasets when training on the HP data from the same cell type, with HiCNN2 and HiCPlus outperforming DeepHiC in most scenarios. We further assessed the robustness of enhancement when models were trained with Hi-C or HP data from the same cell type. We found that enhancement using models trained on ultra-high depth Hi-C data achieved similar or even better performance than using models trained on HP data. However, when the sequencing depth of Hi-C data and HP data used for training were comparable, models trained on HP data exhibited better performance than those trained on Hi-C data. These results suggest that users can train models with high depth Hi-C data for HP data enhancement if similar high depth HP data are not available for training. We note that the terminology "Hi-C data resolution enhancement" prevails in the literature. We have, however, decided to use "data depth enhancement" to avoid ambiguity since resolution is also commonly used to indicate the bin size of the analysis unit.

Transferability across datasets (e.g., cell types, proteins of interest) is important because in practice there are limited cell types sequenced with HiChIP or PLAC-seq techniques. Our analysis across six cell types, three proteins of interest, and two organisms, showed promising transferability results for enhancing HP data, consistent with the existing literatures (Zhang et al., 2018; Liu and

Wang, 2019a,b) for enhancing Hi-C data. For example, models trained using high depth GM12878 data can lead to better enhancement results in mESC than models trained with mESC data. More evaluations are needed in the future to draw stronger conclusions. Such evaluations will become possible when more high depth HP data are generated both for training better models and for evaluations. Note that the actual meaningful information, specifically where the non-zero or zero contacts reside or where the chromatin interactions locate, differs across cell types, organisms, and proteins of interest in HP dataset. The observed promising transferability results suggest that the rules learned to enhance lower depth data to higher depth are shared across cell types (even across organisms). Together, results presented under sections Hi-C or HP data for training and Model transferability suggest that HiCNN2 or HiCPlus models pre-trained from high-depth Hi-C or HP data can be directly applied to enhance HP from various cell types.

For performance evaluation, we used three standard metrics, Pearson correlation coefficients, Spearman correlation coefficients, and Brownian distance covariance (Székely and Rizzo, 2022). Brownian distance covariance is a multivariate dependence coefficient which measures dependency of two random vectors of arbitrary and not necessarily equal dimension, providing more general quantification of independence than linear correlation by Pearson correlation (Székely and Rizzo, 2022). In addition, similarity metrics tailored for Hi-C data, such as HiCRep (Yang et al., 2017) and HiC-spector (Yan et al., 2017) have been widely used. We have recently extended HiCRep to HPrep, tailored for HP data after adjusting for ChIP enrichment biases (Rosen et al., 2021). Applying HPrep to evaluate similarity between enhanced data and full data led to findings consistent with what revealed by Pearson correlation: for example, the three deep learning methods behave better than baseline and they all have similar performance. Overall, all three methods evaluated are able to generate enhanced data exceeding the baseline (i.e., low depth data without enhancement), both in terms of enhancing the contact frequency matrix (as quantified by the correlation metrics) and probably more importantly in terms of improving power to detect chromatin interactions. Among the three, we recommend HiCNN2 and HiCPlus, both consistently exhibiting similar performance, superior to DeepHiC and VEHiCLE, when applied to enhance HP data. Note that DeepHiC and

VEHiCLE both employ the GAN model, which has been known to suffer from mode collapse problem (Salimans et al., 2016b; Srivastava et al., 2017). Due to the nature of HP data, multi-modal distribution is expected because of the systematic difference between AND and XOR bin pairs, which might explain why DeepHiC and VEHiCLE perform sub-optimally in HP data enhancement. Not surprisingly, with increased down-sampling ratio, enhanced data from very shallow depth data showed more pronounced improvement over the baseline. When the sequencing depth is high, there is less room for improvement, particularly when using methods developed for Hi-C data that do not consider ChIP enrichment bias of HP data. For example, when we enhanced HP data on higher depth data (e.g., 1/4 and 1/2 down-sampled data), we found that "enhanced" data from all three methods are comparable or even slightly worse than the baseline when measured by correlation metrics, while theoretically, enhancement methods can still improve 1/4 and 1/2 data. In addition, we observed HP data enhanced using these methods show lower correlation than Hi-C data enhanced by these methods. For example, the Pearson correlation coefficients are in the range of 0.95-0.96 within 500Kb for HiCNN and HiCPlus on GM12878 1/8 ratio on chromosomes 6 and 12 (Figure BS3 in HiCNN paper (Liu and Wang, 2019a)) but enhanced HP data in our evaluations showed Pearson correlation $<$0.81. Furthermore, the improvement in HP data (as reflected by the correlation decay with distance figures) is not as smooth as in Hi-C data, which might be caused by unbalanced read distribution due to protein immunoprecipitation in HP data. Therefore, methods developed for Hi-C data are not optimal for HP data. Development of methods tailored to HP data is warranted.

3.5 Data Availability

We downloaded GM12878 Smc1a HiChIP dataset (Mumbach et al., 2016), H3K4me3 PLAC-seq dataset in mouse embryonic stem cells (mESCs) (Fang et al., 2016b) (GSE119663), H3K4me3 PLAC-seq datasets for four human brain cell types (Song et al., 2020), mESC CTCF PLAC-seq data (Juric et al., 2019), GM12878 Hi-C dataset (Rao et al., 2014b), and mESC Hi-C dataset (Bonev et al., 2017b). We also obtained ChIP-seq peaks for different cell lines (GM12878 Smc1a ChIP-seq peaks: https://www.encodeproject.org/files/ENCFF686FLD/, mESC H3K4me3 ChIP-seq peaks: https://www.ncbi.nlm.nih.gov/geo/query/acc.

cgi?acc=GSM3380558) as well as RNA-seq and ATAC-seq data for the human brain cell types from (Song et al., 2020).

3.6 Supplementary Materials

This chapter provides supplementary figures for this work.

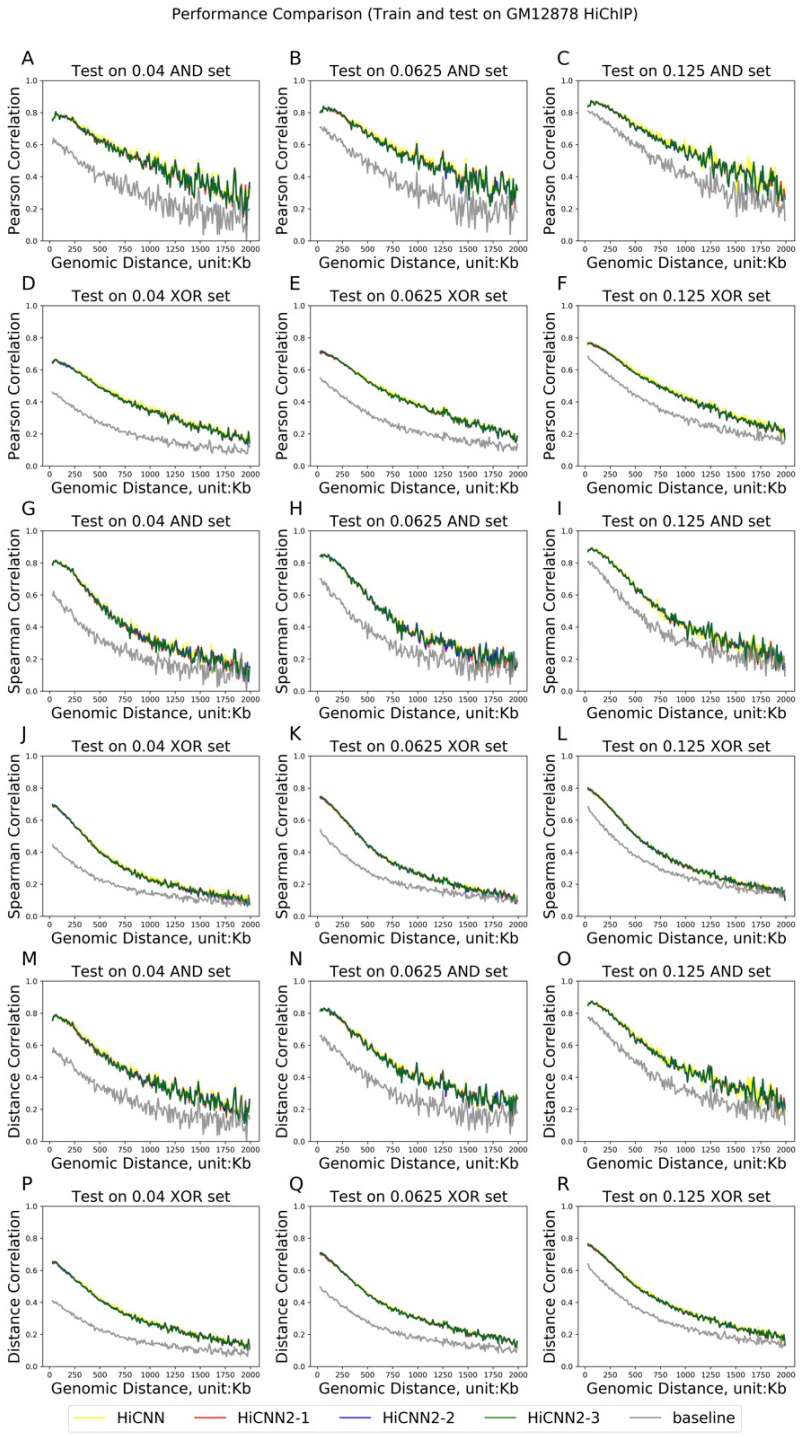

Figure S3.1: **Performance of HiCNN2 and HiCNN when enhancing GM12878 HiChIP data**. Note that HiCNN2 has three models: HiCNN2-1, HiCNN2-2, HiCNN2-3. Left panel shows performance in 0.04 down-sampled data, middle panel in 0.0625 down-sampled data, and right panel in 0.125 down-sampled data. Top two rows quantify performance with Pearson correlation coefficient, middle two rows with Spearman correlation, and bottom two rows with distance correlation (Y-axis). X-axis is genomic distance in Kb unit. The gray line represents the baseline (i.e., low depth data without any enhancement).

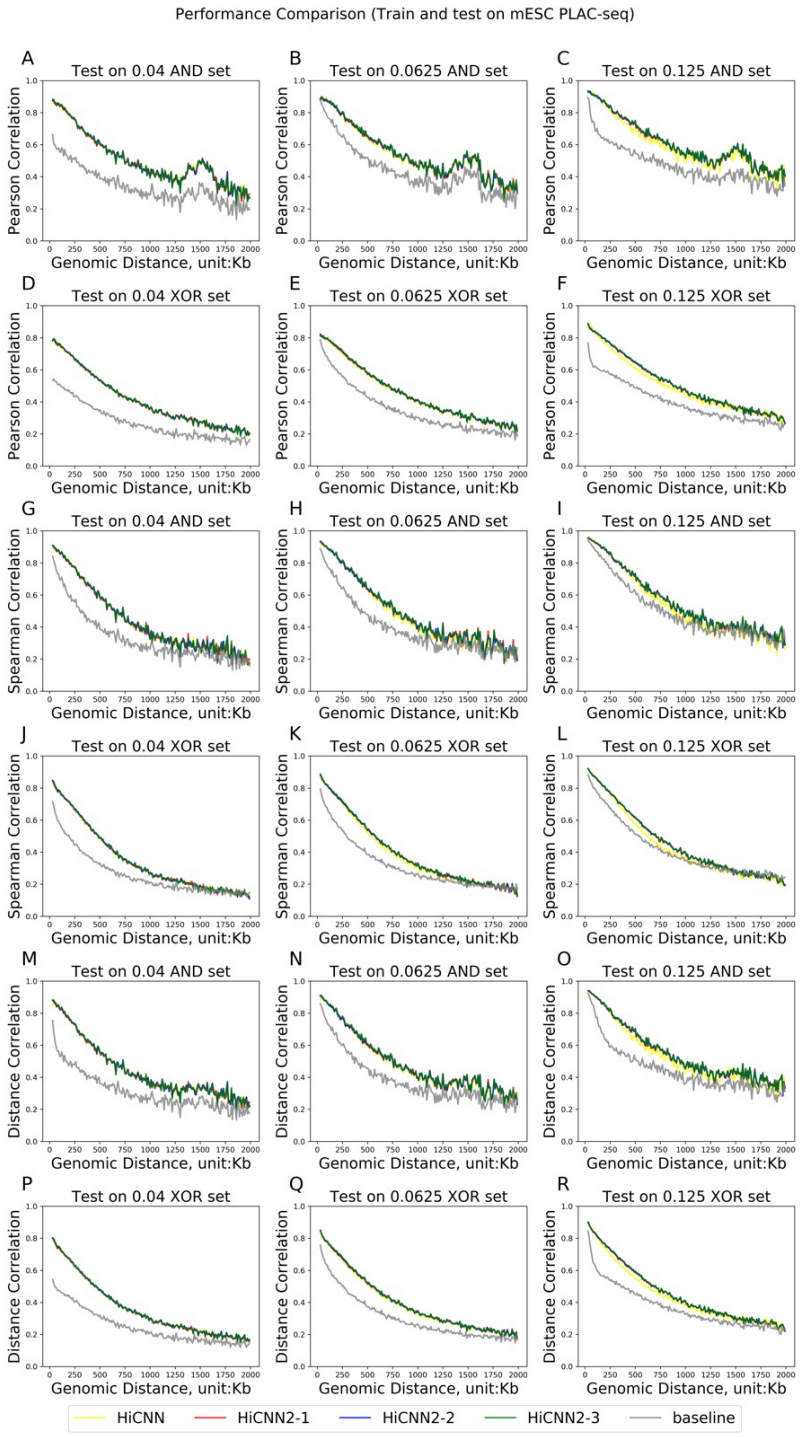

Figure S3.2: **Performance of HiCNN2 and HiCNN when enhancing mESC PLAC-seq data**. Note that HiCNN2 has three models: HiCNN2-1, HiCNN2-2, HiCNN2-3. Left panel shows performance in 0.04 down-sampled data, middle panel in 0.0625 down-sampled data, and right panel in 0.125 down-sampled data. Top two rows quantify performance with Pearson correlation coefficient, middle two rows with Spearman correlation, and bottom two rows with distance correlation (Y-axis). X-axis is genomic distance in Kb unit. The gray line represents the baseline (i.e., low depth data without any enhancement).

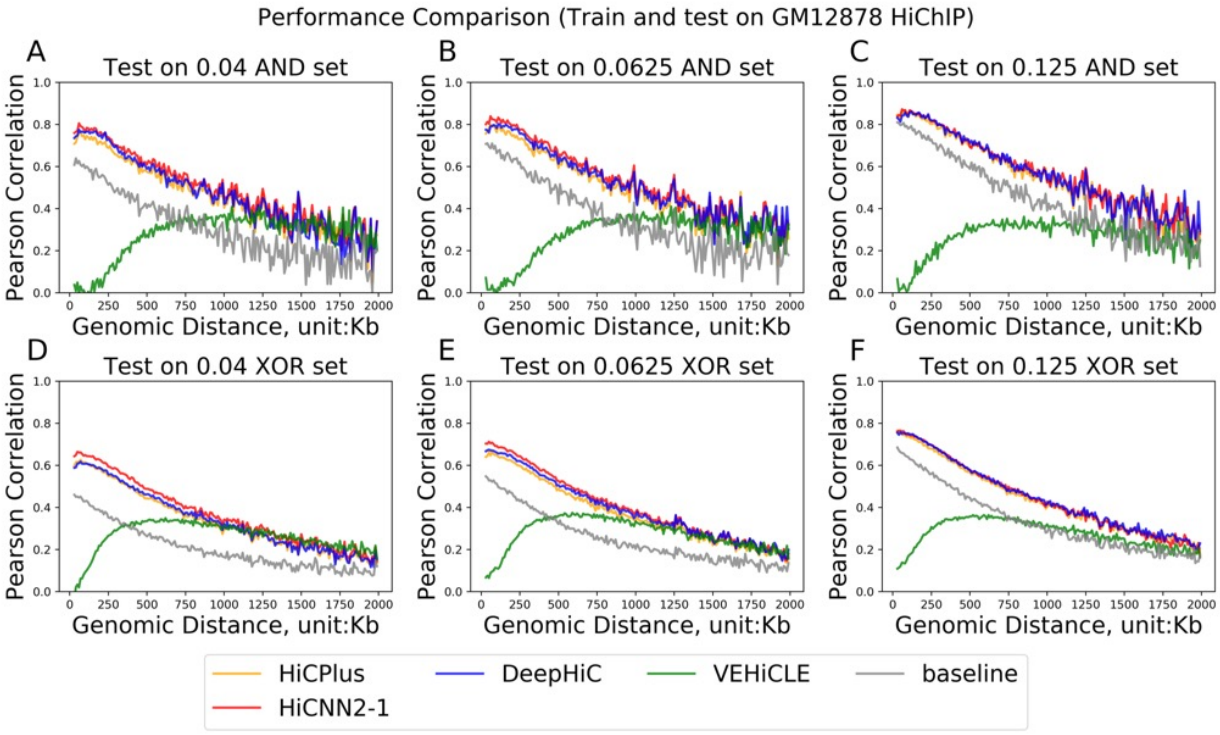

Figure S3.3: **Methods comparison when enhancing GM12878 HiChIP data (with VEHiCLE)**. Four enhancement methods are compared: HiCPlus, HiCNN2-1, and DeepHiC. Left panel (A and D) shows performance in 0.04 down-sampled data, middle panel (B and E) in 0.0625 down-sampled data, and right panel (C and F) in 0.125 down-sampled data. Performance is quantified with Pearson correlation coefficient (Y-axis). X-axis is genomic distance in Kb unit. Top row (A-C) shows performance among bin pairs in the AND set and bottom row (D-F) shows performance among bin pairs in the XOR set. The gray line represents the baseline (i.e., low depth data without any enhancement).

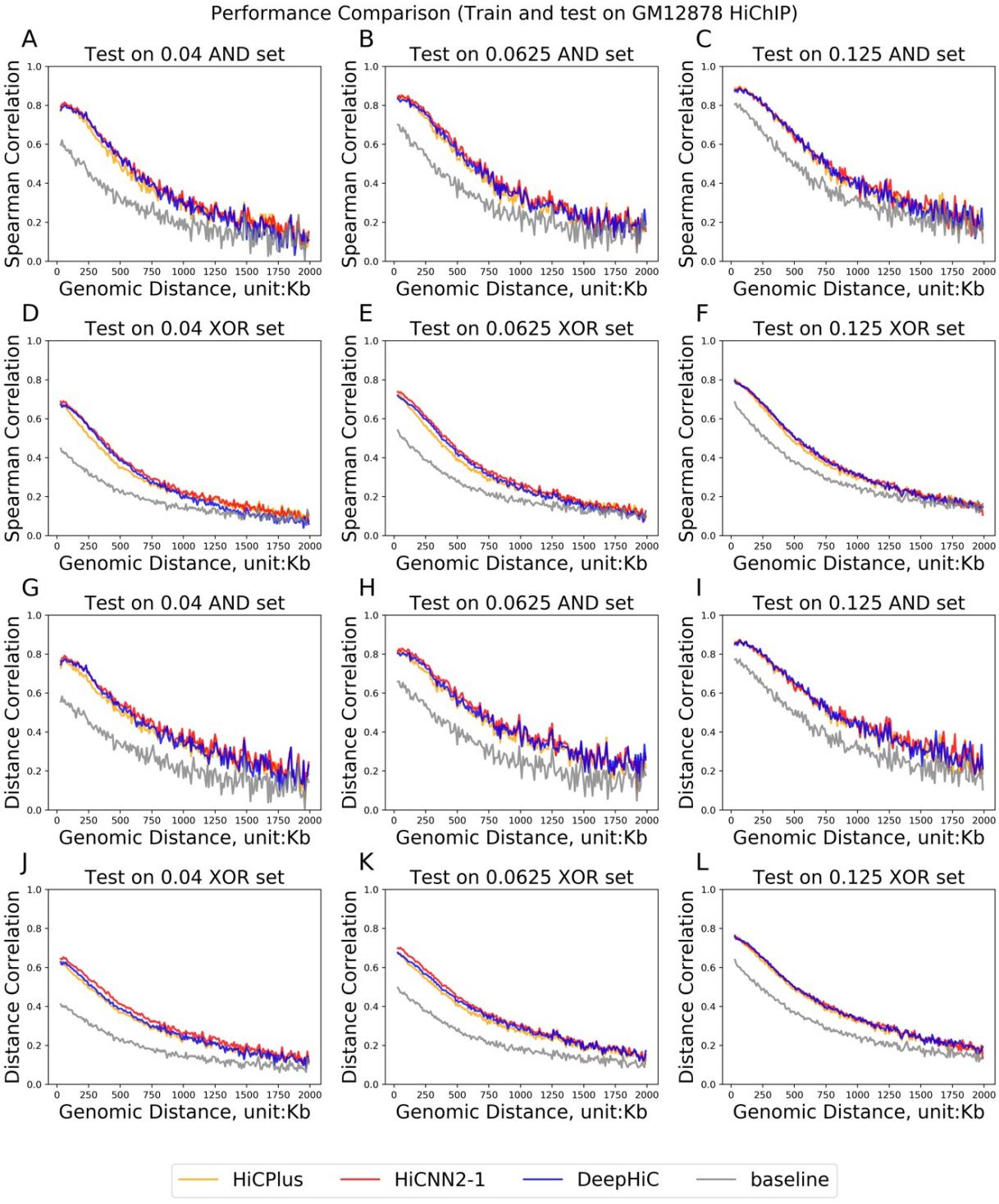

Figure S3.4: **Performance comparison when enhancing GM12878 HiChIP data, quantified with Spearman and distance correlations.** Left panel shows performance in 0.04 down-sampled data, middle panel in 0.0625 down-sampled data, and right panel in 0.125 down-sampled data. Top two rows quantify performance with Spearman correlation, and bottom two rows with distance correlation (Y-axis). X-axis is genomic distance in Kb unit. The gray line represents the baseline (i.e., low depth data without any enhancement).

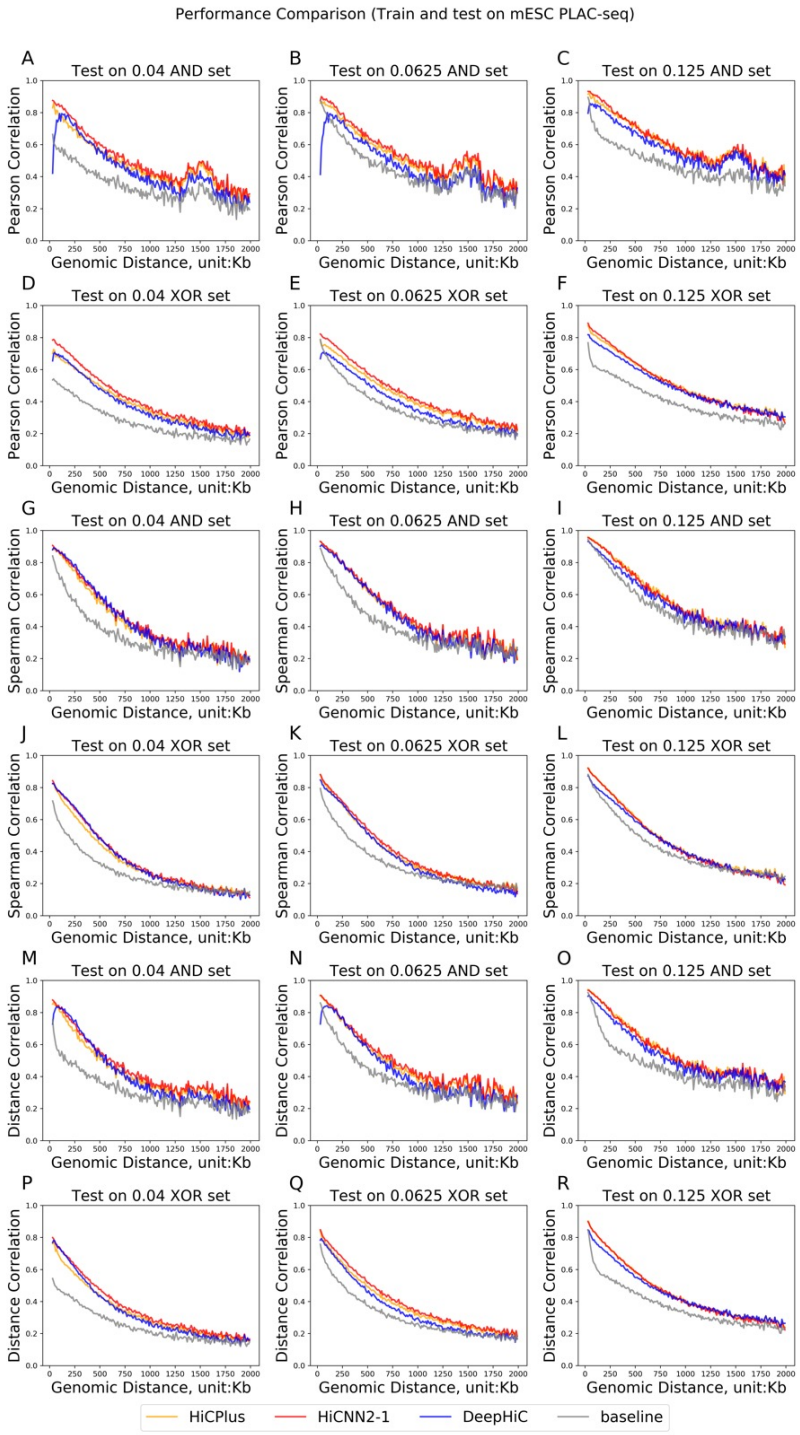

Figure S3.5: **Performance comparison when enhancing mESC PLAC-seq data**. Left panel shows performance in 0.04 down-sampled data, middle panel in 0.0625 down-sampled data, and right panel in 0.125 down-sampled data. Top two rows quantify performance with Pearson correlation coefficient, middle two rows with Spearman correlation, and bottom two rows with distance correlation (Y-axis). X-axis is genomic distance in Kb unit. The gray line represents the baseline (i.e., low depth data without any enhancement).

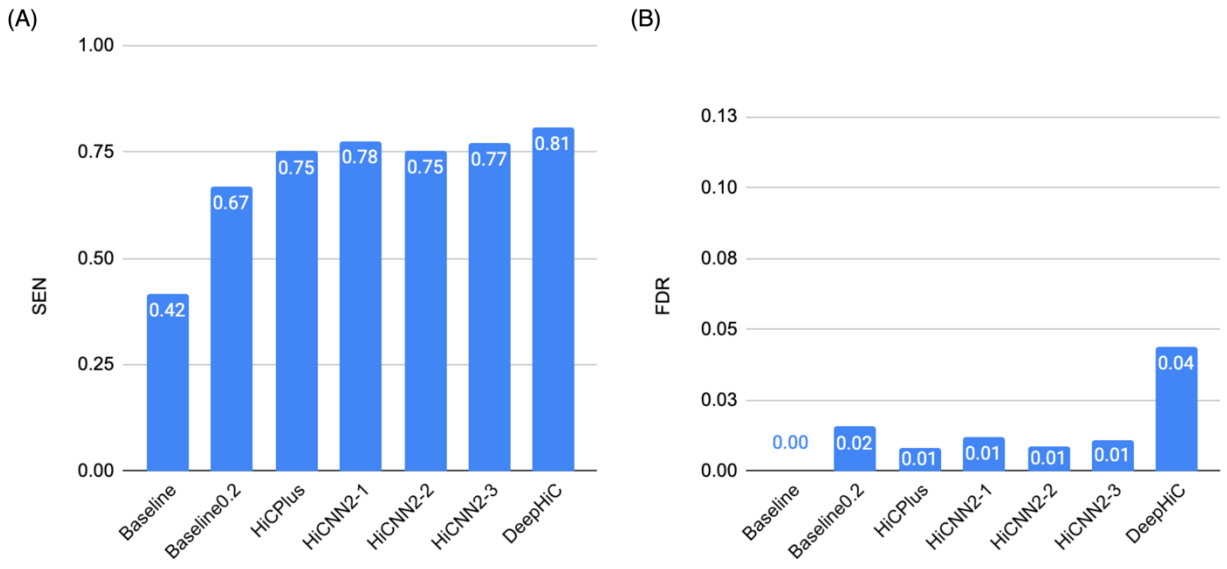

Figure S3.6: **3D peak calling in 0.5 down-sampled mESC PLAC-seq data**. Left panel (A) shows sensitivity (SEN). Right panel (B) shows FDR. The truth (3D peaks or not) is established by peak calling via MAPS from full data without any down-sampling. Specifically, true peaks are bin pairs with MAPS FDR $< 1\%$ and contacts ≥ 12. True background bin pairs are those with MAPS FDR $> 20\%$ and contacts ≥ 12. Baseline0.2 bars show the 3D peak calling performance when relaxing MAPS-FDR threshold from 1% to 20%.

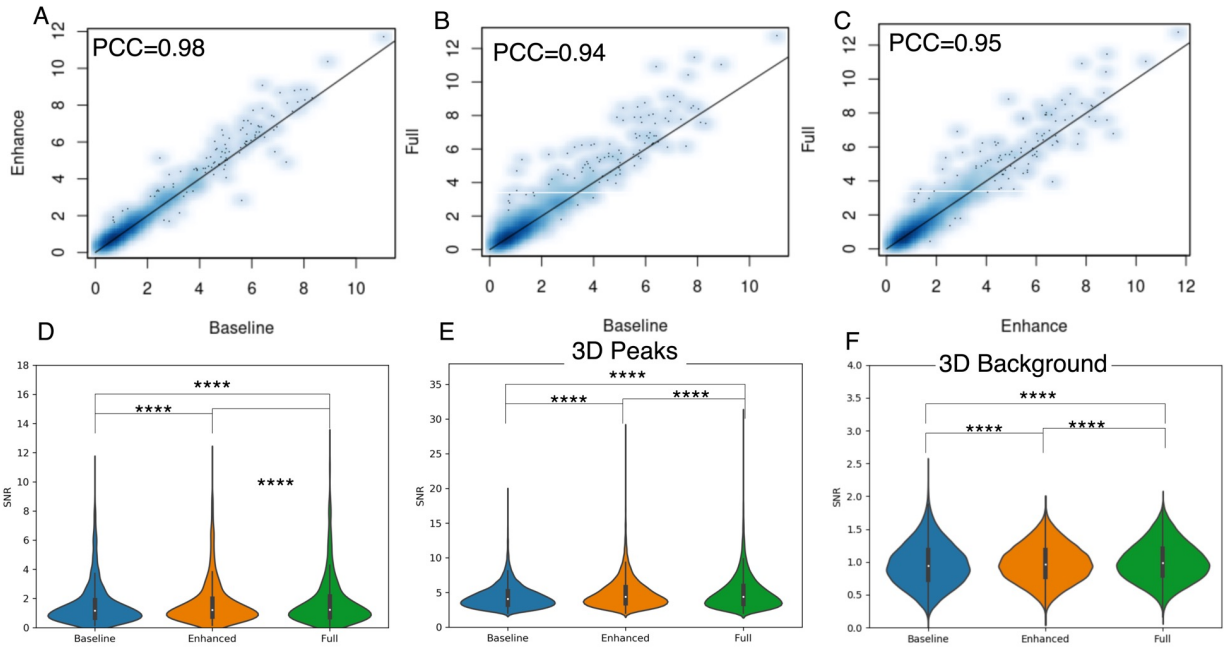

Figure S3.7: **Signal-to-noise-ratio (SNR) comparison**. We compare SNR when enhancing 0.5 down-sampled GM12878 HiChIP dataset with HiCNN2-1. Specifically, we compare SNR of bin pairs estimated at baseline (low depth data without any enhancement), after enhancement, or in full data (without any down-sampling). True peaks are established by peak calling via MAPS from full data. SNR is defined as observed count over expected count, output from MAPS. Left panel shows all bin pairs regardless or 3D peak status established by full data. Middle panel shows 3D peak bin pairs identified by MAPS from full data. Right panel shows 3D non-peaks or background bin pairs as classified by MAPS inference. A-C are scatter plots. D-F are side-to-side violin plots. We assess statistical significance by Wilcoxon test. ****, ***, and ** indicate Wilcoxon p-value $<$ 1e-4, 1e-3, and 1e-2 respectively.

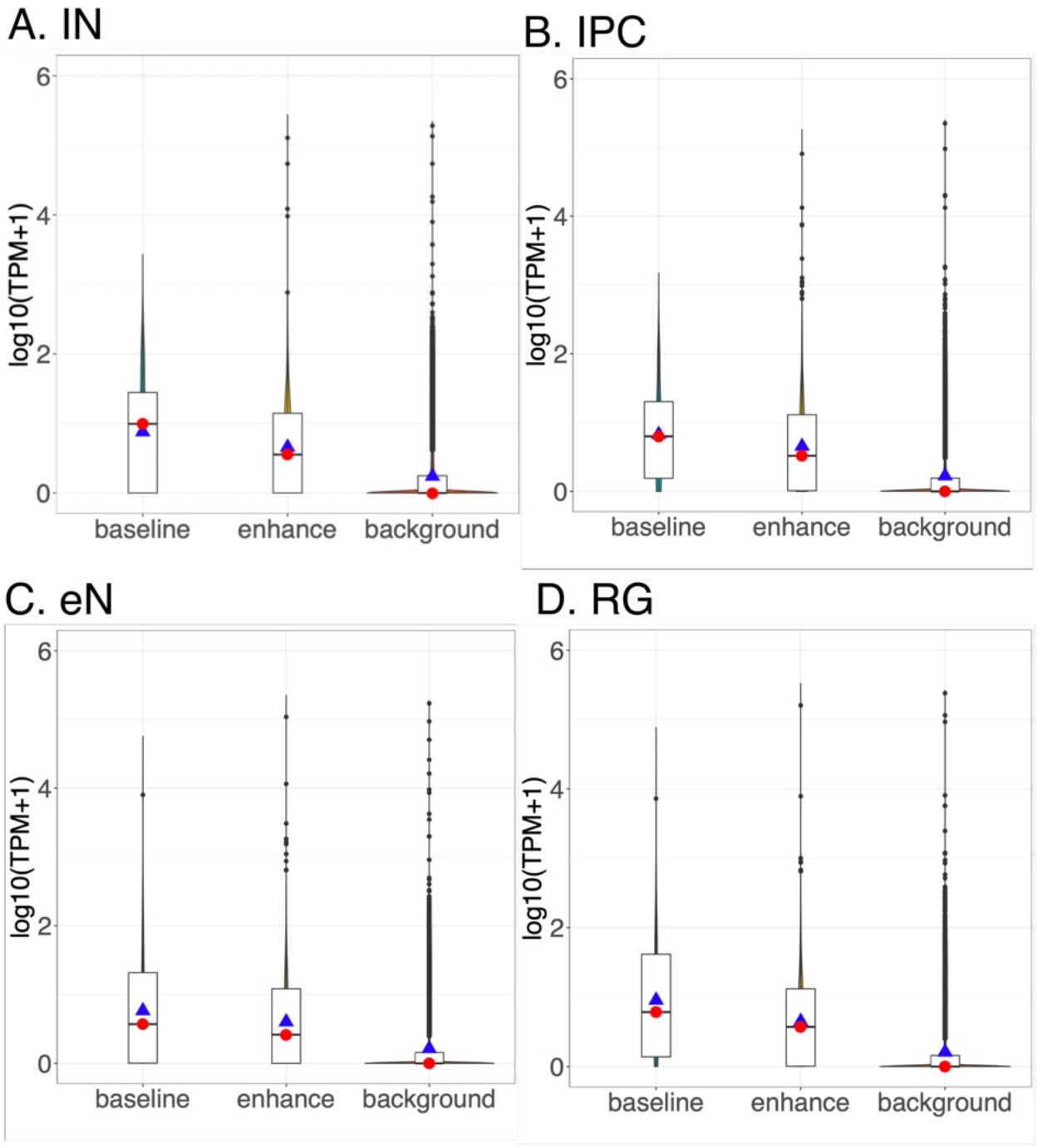

Figure S3.8: **Expression of genes with identified chromatin interaction(s) in neural cells**. Data enhancement was performed on each of the four neural cell types (interneurons [IN], intermediate progenitor cells [IPC], excitatory neurons [eN], and radial glia [RG]), from 0.125 down-sampled data to full data. All results are from HiCNN2-1 models. For each cell type, we compare gene expression (measured by log10(TPM+1)) for three sets of genes: (1) "baseline": genes whose promoter region (+/-500bp of TSS) involves in some significant chromatin interaction(s) at baseline (i.e., from low depth data before enhancement; (2) "enhance": genes whose promoter region does not involve in any significant chromatin interaction at baseline, but in some significant chromatin interactions after enhancement; and (3) "background": genes whose promoter region does not involve in any significant chromatin interactions even with the full data (without any down-sampling).

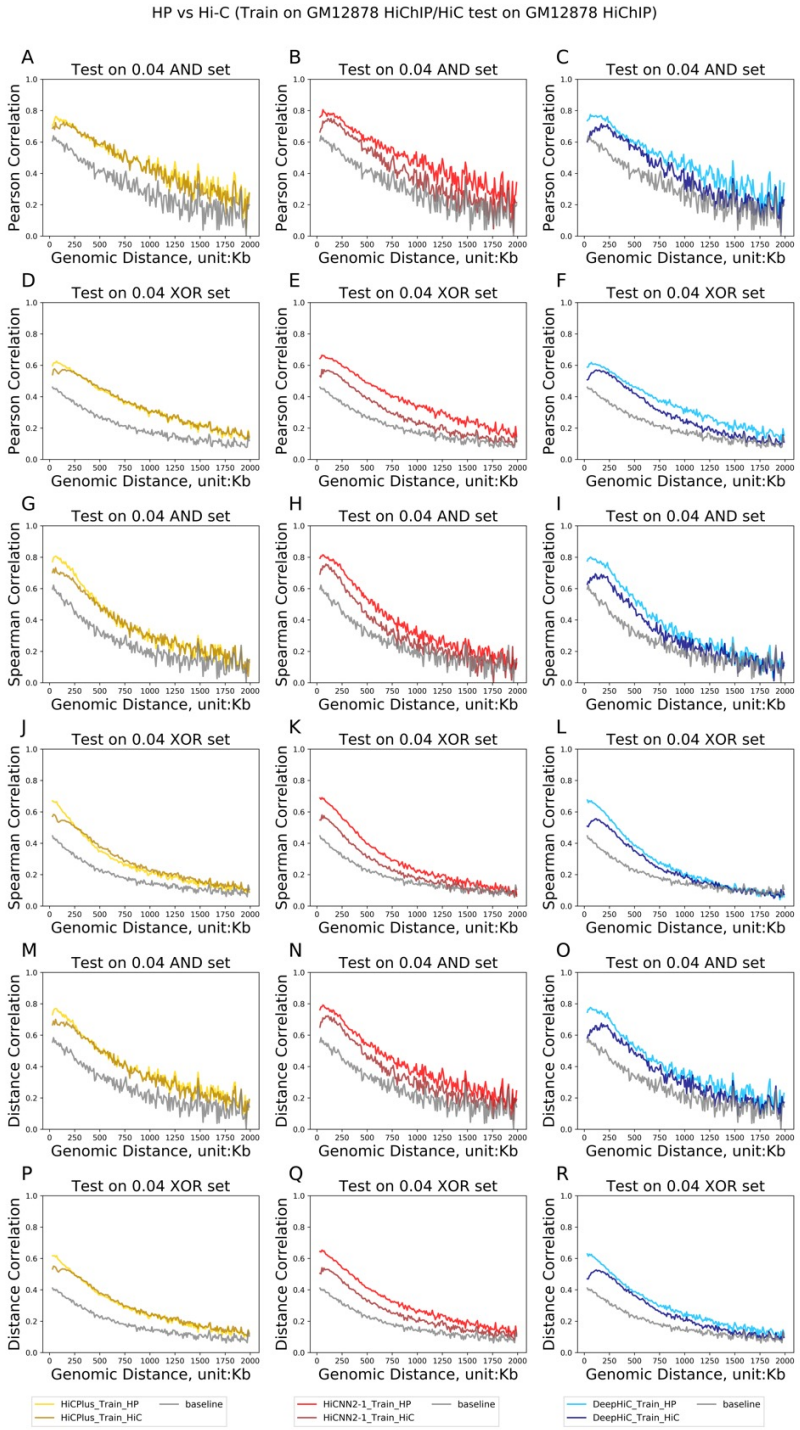

Figure S3.9: **HiChIP trained vs Hi-C trained models when enhancing GM12878 HiChIP data by 25×.** The evaluation metrics are (1) Pearson correlation coefficient (A-F), (2) Spearman correlation coefficient (G-L), and (3) Brownian distance correlation (M-R). Each subfigure represents the performance of one of three read depth enhancement methods (HiCNN2-1, HiCPlus, and DeepHiC) for a certain set (AND set or XOR set). In each subfigure, we show how the evaluation metric (Y-axis) changes with genomic distance (X-axis), where the distance ranges from 20Kb-2Mb with an increment of 10Kb. The gray line represents the baseline (i.e., low depth data without any enhancement).

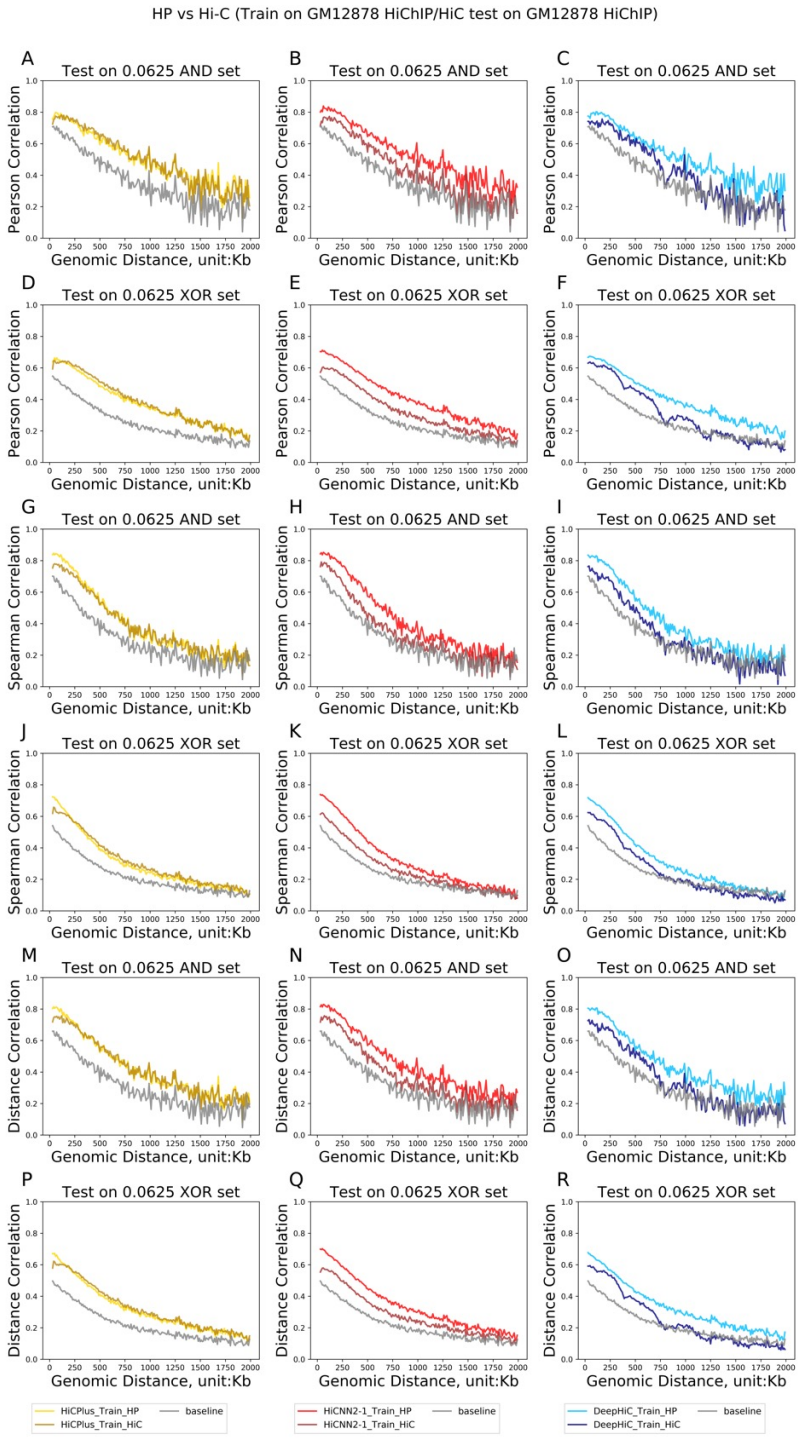

Figure S3.10: **HiChIP trained vs Hi-C trained models when enhancing GM12878 HiChIP data by 16×.** The evaluation metrics are (1) Pearson correlation coefficient (A-F), (2) Spearman correlation coefficient (G-L), and (3) Brownian distance correlation (M-R). Each subfigure represents the performance of one of three read depth enhancement methods (HiCNN2-1, HiCPlus, and DeepHiC) for a certain set (AND set or XOR set). In each subfigure, we show how the evaluation metric (Y-axis) changes with genomic distance (X-axis), where the distance ranges from 20Kb-2Mb with an increment of 10Kb. The gray line represents the baseline (i.e., low depth data without any enhancement).

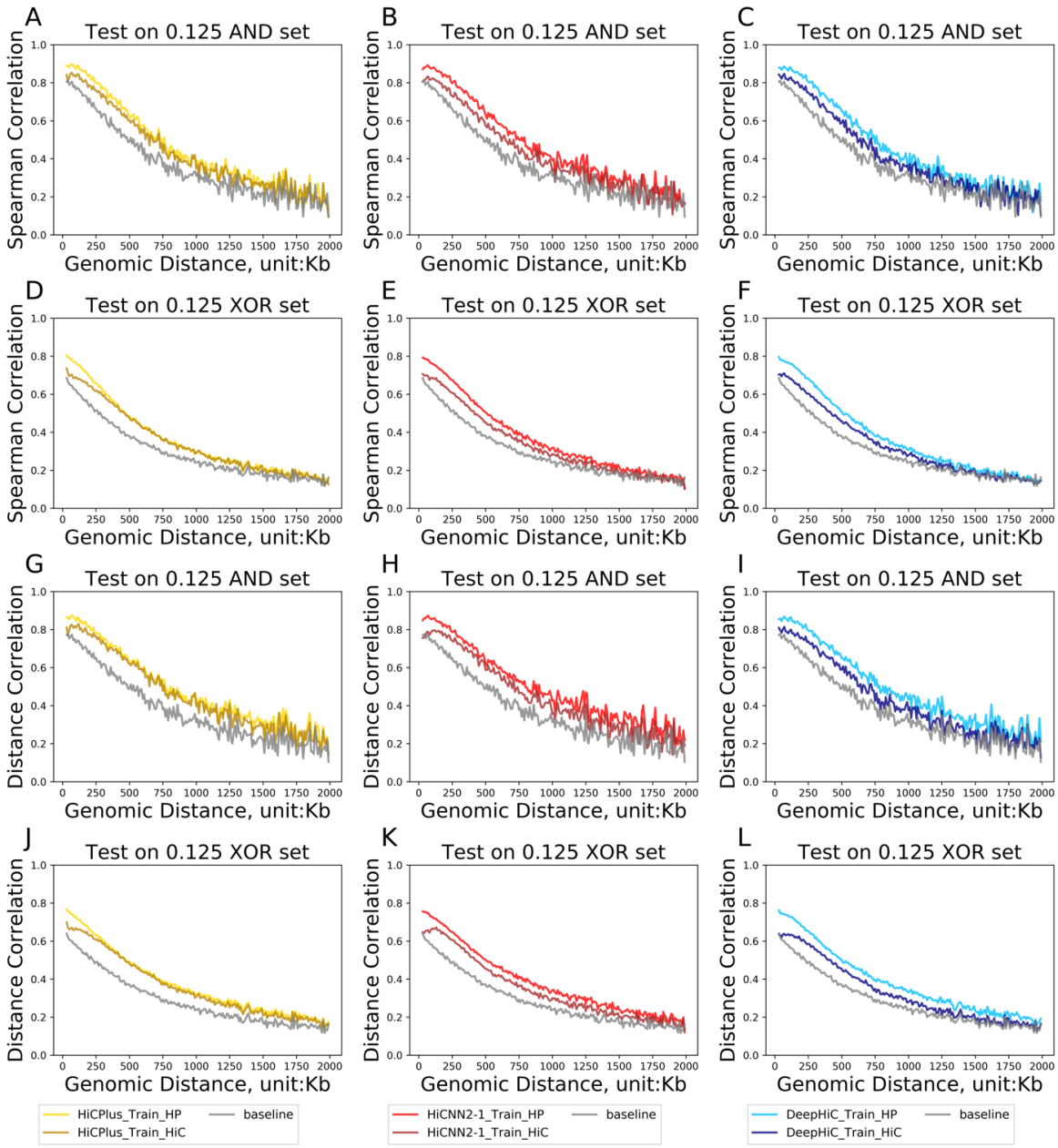

Figure S3.11: **HiChIP trained vs Hi-C trained models when enhancing GM12878 HiChIP data by 8×.** The evaluation metrics are (1) Spearman correlation coefficient (A-F), and (2) Brownian distance correlation (G-L). Each subfigure represents the performance of one of three read depth enhancement methods (HiCNN2-1, HiCPlus, and DeepHiC) for a certain set (AND set or XOR set). In each subfigure, we show how the evaluation metric (Y-axis) changes with genomic distance (X-axis), where the distance ranges from 20Kb-2Mb with an increment of 10Kb. The gray line represents the baseline (i.e., low depth data without any enhancement).

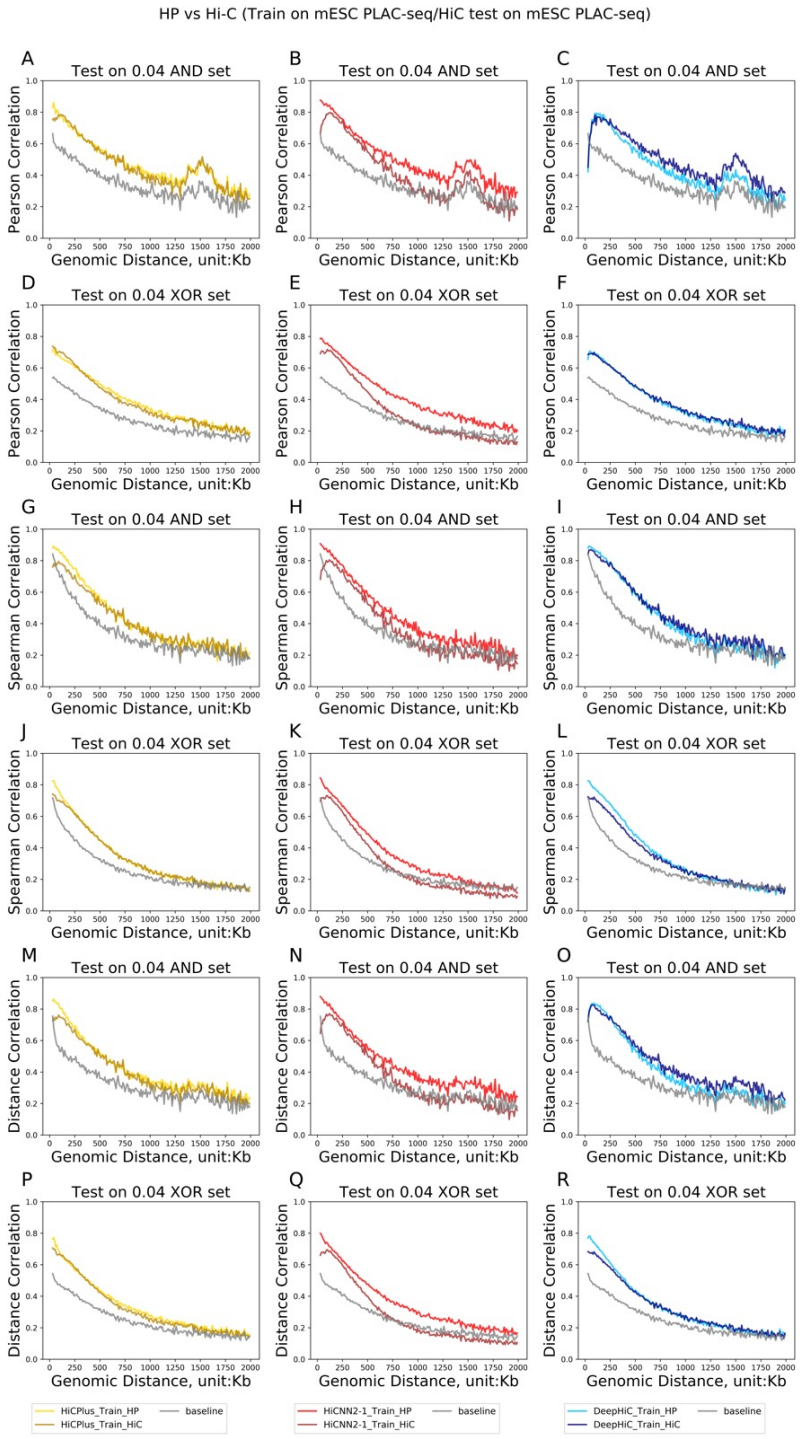

Figure S3.12: **PLAC-seq trained vs Hi-C trained models when enhancing mESC PLAC-seq data by 25×.** The evaluation metrics are (1) Pearson correlation coefficient (A-F), (2) Spearman correlation coefficient (G-L), and (3) Brownian distance correlation (M-R). Each subfigure represents the performance of one of three read depth enhancement methods (HiCNN2-1, HiCPlus, and DeepHiC) for a certain set (AND set or XOR set). In each subfigure, we show how the evaluation metric (Y-axis) changes with genomic distance (X-axis), where the distance ranges from 20Kb-2Mb with an increment of 10Kb. The gray line represents the baseline (i.e., low depth data without any enhancement).

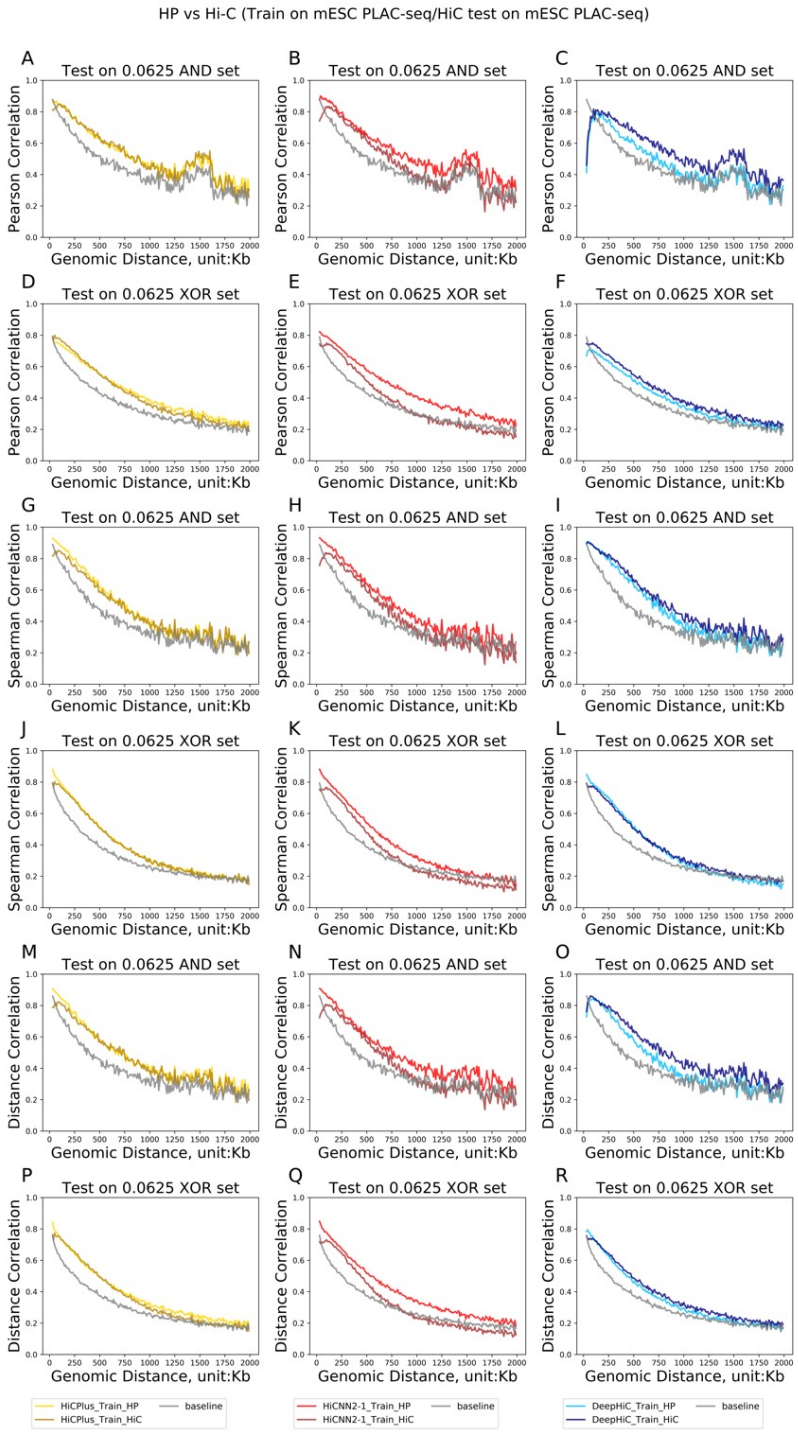

Figure S3.13: **PLAC-seq trained vs Hi-C trained models when enhancing mESC PLAC-seq data by 16×.** The evaluation metrics are (1) Pearson correlation coefficient (A-F), (2) Spearman correlation coefficient (G-L), and (3) Brownian distance correlation (M-R). Each subfigure represents the performance of one of three read depth enhancement methods (HiCNN2-1, HiCPlus, and DeepHiC) for a certain set (AND set or XOR set). In each subfigure, we show how the evaluation metric (Y-axis) changes with genomic distance (X-axis), where the distance ranges from 20Kb-2Mb with an increment of 10Kb. The gray line represents the baseline (i.e., low depth data without any enhancement).

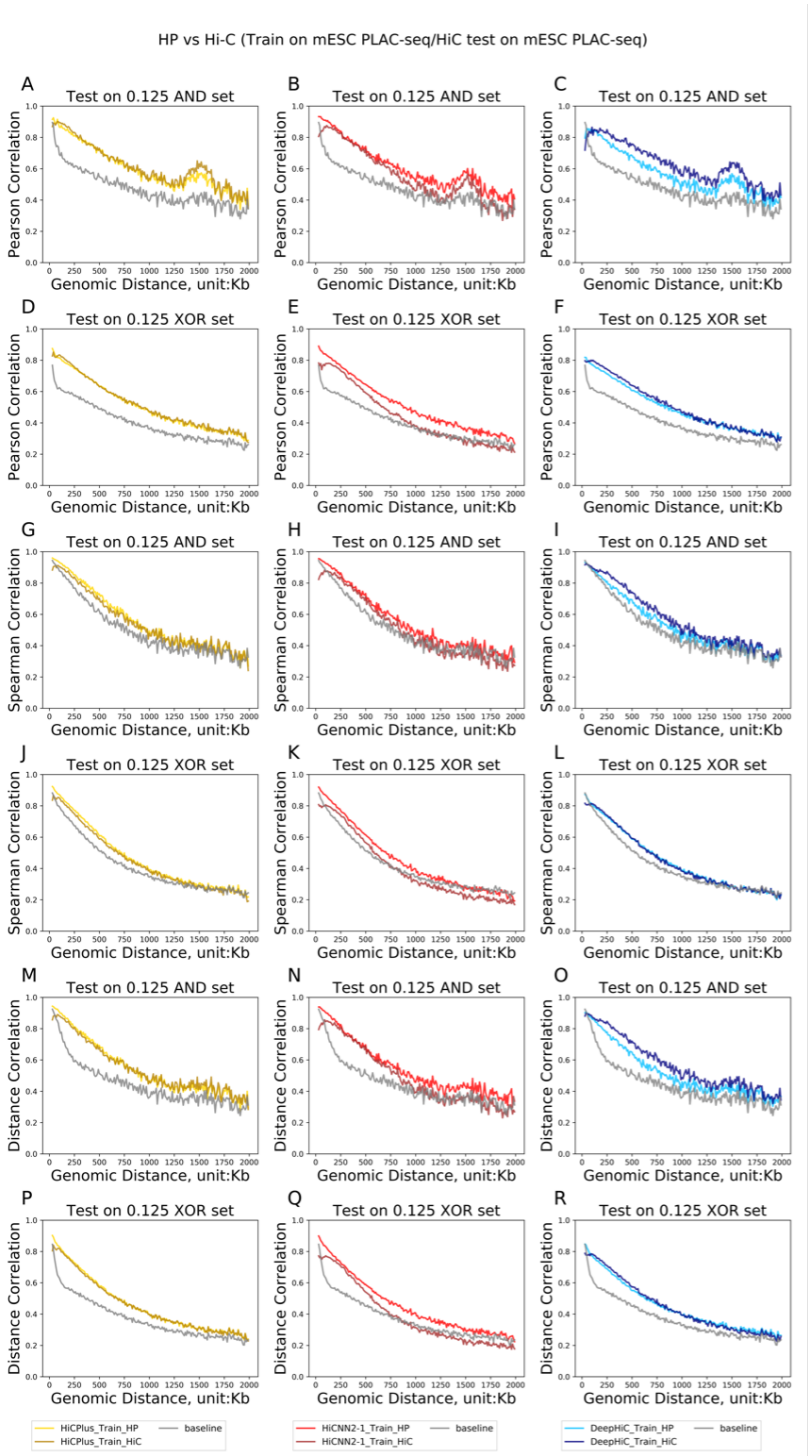

Figure S3.14: **PLAC-seq trained vs Hi-C trained models when enhancing mESC PLAC-seq data by 8×.** The evaluation metrics are (1) Pearson correlation coefficient (A-F), (2) Spearman correlation coefficient (G-L), and (3) Brownian distance correlation (M-R). Each subfigure represents the performance of one of three read depth enhancement methods (HiCNN2-1, HiCPlus, and DeepHiC) for a certain set (AND set or XOR set). In each subfigure, we show how the evaluation metric (Y-axis) changes with genomic distance (X-axis), where the distance ranges from 20Kb-2Mb with an increment of 10Kb. The gray line represents the baseline (i.e., low depth data without any enhancement).

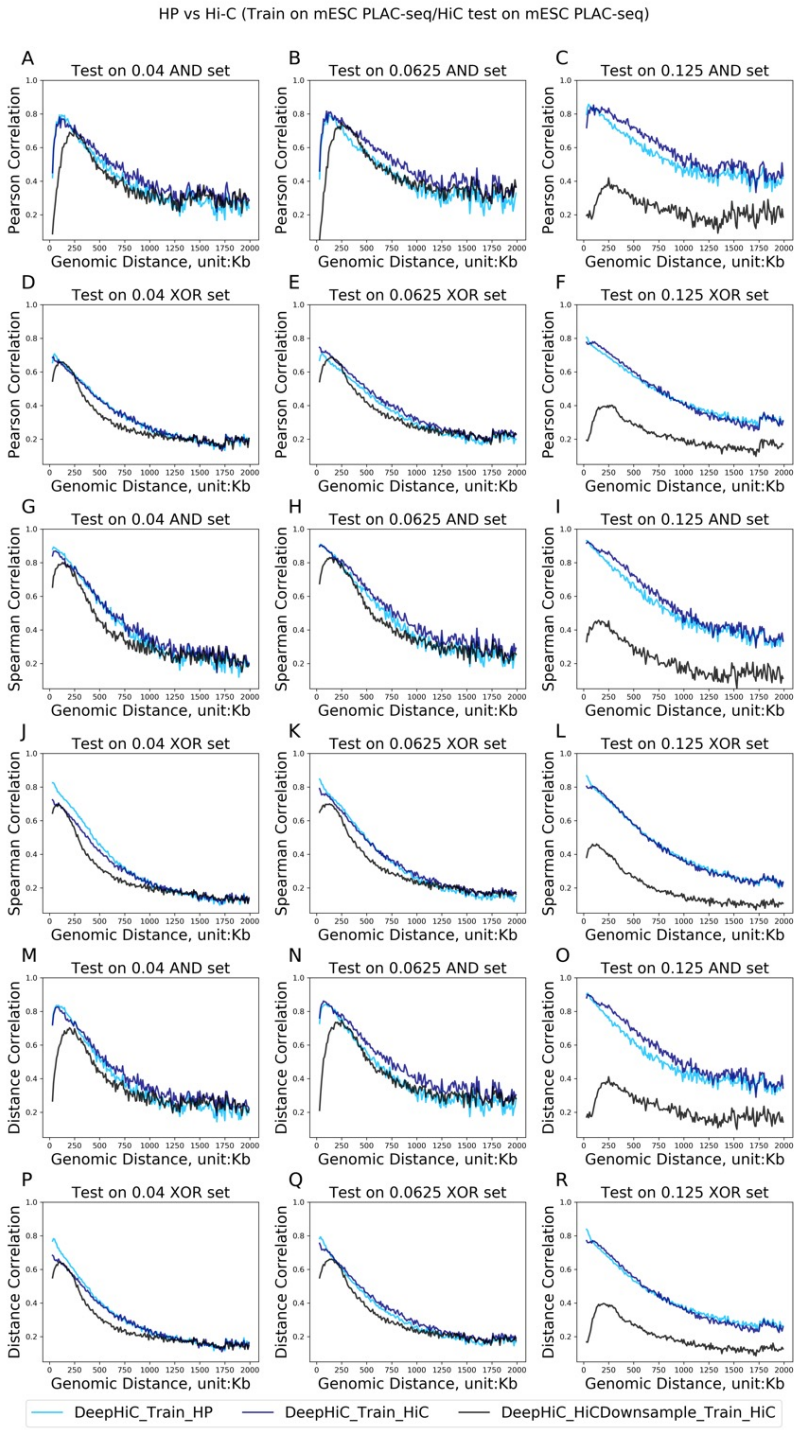

Figure S3.15: **Impact of training data sequencing depth on DeepHiC Performance**. We compare models trained on mESC PLAC-seq (light blue), mESC HiC (deep blue) and mESC HiCDownsample (black) datasets. The evaluation metrics are (1) Pearson correlation coefficient (A-F), (2) Spearman's rank correlation coefficient (G-L), and (3) Brownian distance correlation (M-R). Each subfigure represents the performance of the DeepHiC model at a certain down-sampling ratio (0.04, 0.0625, or 0.125) and for a certain set (AND set or XOR set). In each subfigure, we show how the evaluation metric (Y-axis) changes with genomic distance (X-axis), where the distance ranges from 20Kb-2Mb with an increment of 10Kb.

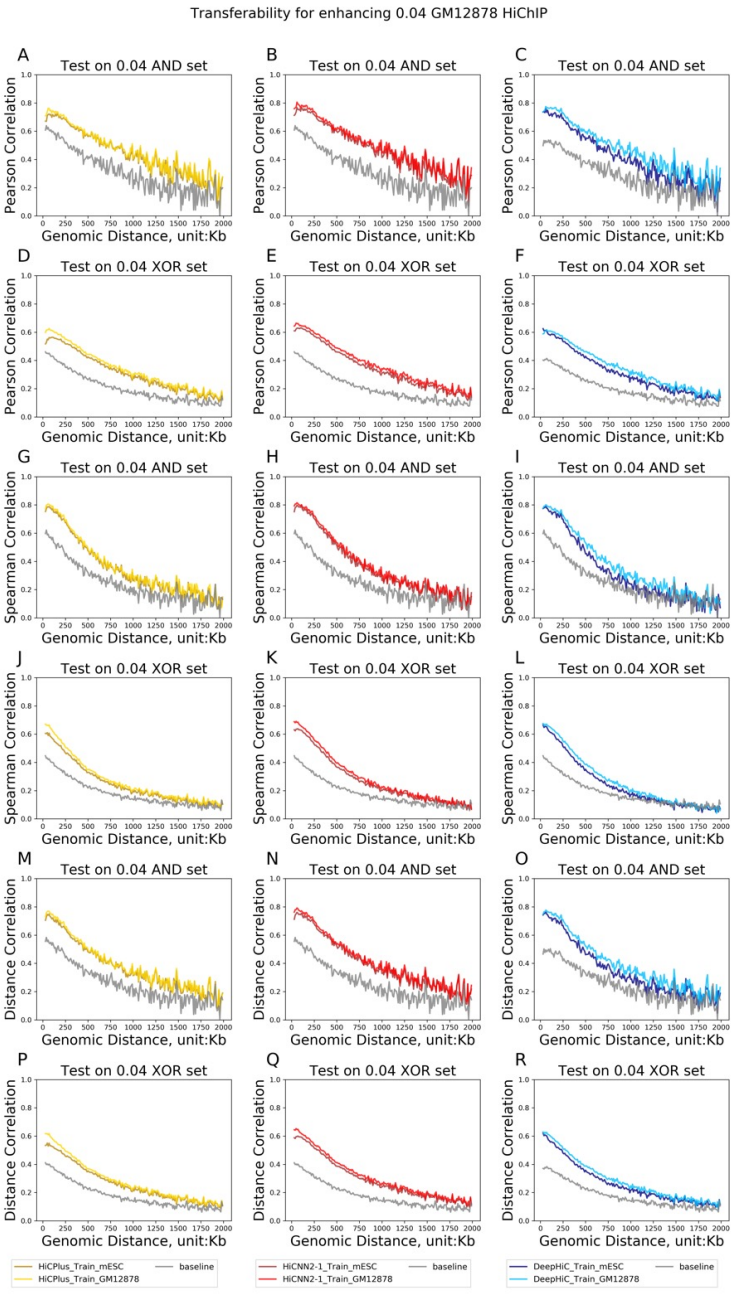

Figure S3.16: **Model transferability when enhancing GM12878 HiChIP data by 25**×. Each subfigure compares two enhanced GM12878 HiChIP data: one using models trained with GM12878 HiChIP data and the other using models trained with mESC PLAC-seq data. The evaluation metrics are (1) Pearson correlation coefficient (A-F), (2) Spearman's rank correlation coefficient (G-L), and (3) Brownian distance coefficient (M-R). Different colors in the subfigures represent different methods (yellow: HiCPlus, red: HiCNN2-1, blue: DeepHiC) while darker color represents models trained with mESC PLAC-seq and lighter color represents models trained with GM12878 HiChIP data. In each subfigure, we show how the evaluation metric (Y-axis) changes with genomic distance (X-axis), where the distance ranges from 20Kb-2Mb with an increment of 10Kb. The gray line represents the baseline (i.e., low depth data without any enhancement).

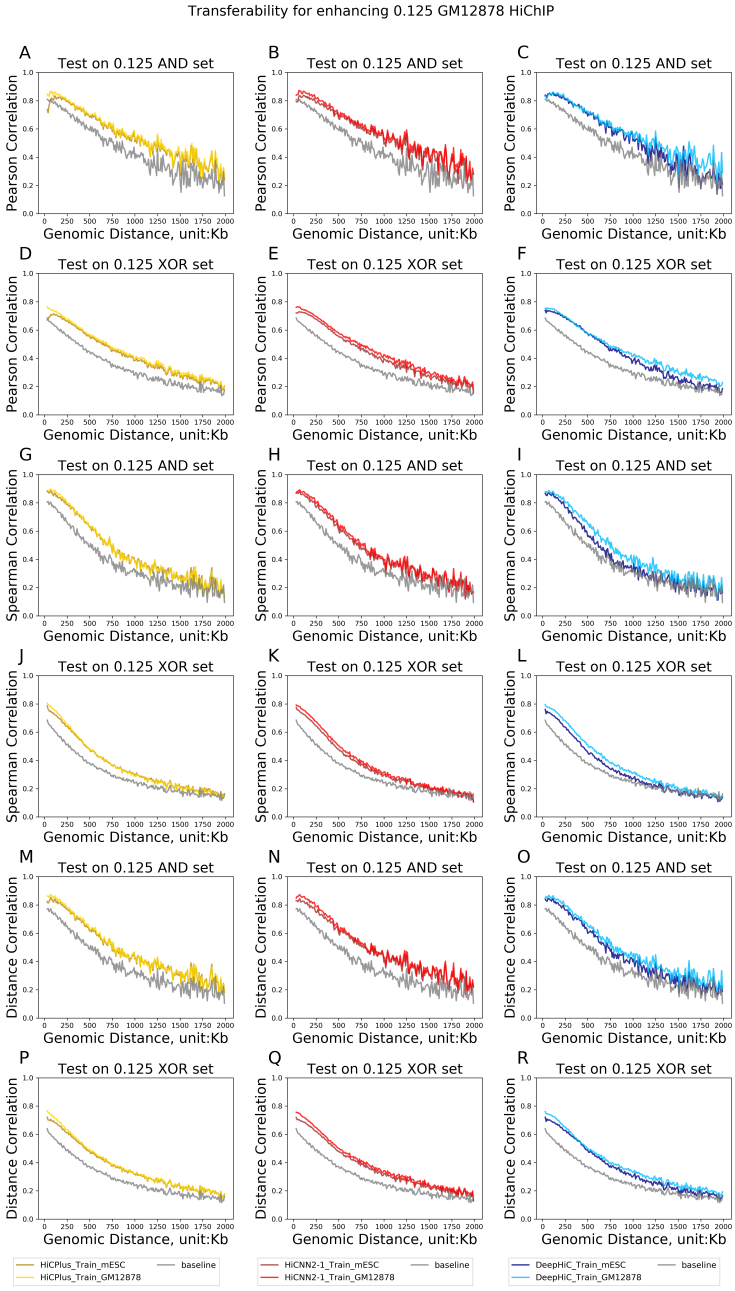

Figure S3.17: **Model transferability when enhancing GM12878 HiChIP data by 16×**. Each subfigure compares two enhanced GM12878 HiChIP data: one using models trained with GM12878 HiChIP data and the other using models trained with mESC PLAC-seq data. The evaluation metrics are (1) Pearson correlation coefficient (A-F), (2) Spearman's rank correlation coefficient (G-L), and (3) Brownian distance coefficient (M-R). Different colors in the subfigures represent different methods (yellow: HiCPlus, red: HiCNN2-1, blue: DeepHiC) while darker color represents models trained with mESC PLAC-seq and lighter color represents models trained with GM12878 HiChIP data. In each subfigure, we show how the evaluation metric (Y-axis) changes with genomic distance (X-axis), where the distance ranges from 20Kb-2Mb with an increment of 10Kb. The gray line represents the baseline (i.e., low depth data without any enhancement).

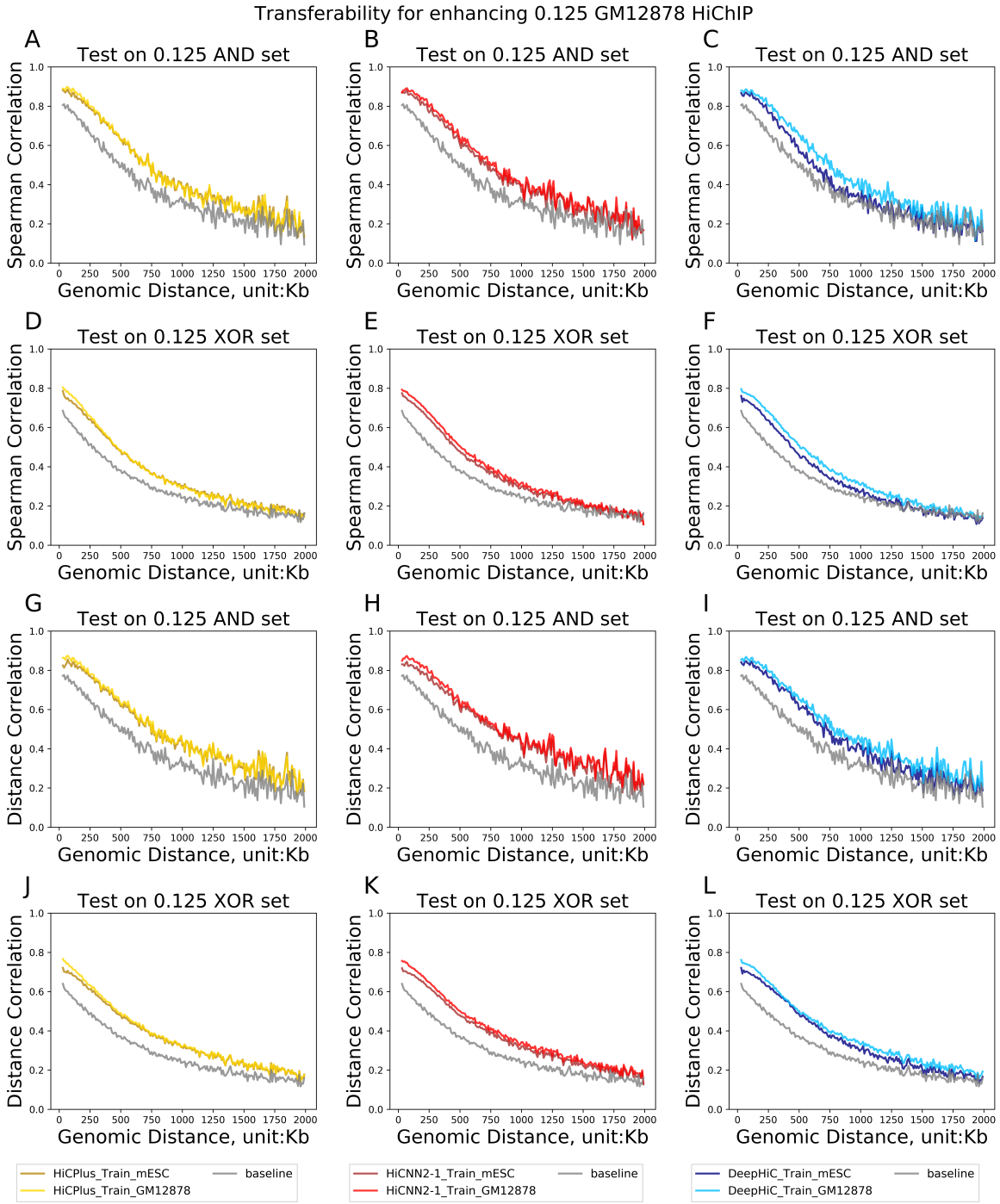

Figure S3.18: **Model transferability when enhancing GM12878 HiChIP data by 8×**. Each subfigure compares two enhanced GM12878 HiChIP data: one using models trained with GM12878 HiChIP data and the other using models trained with mESC PLAC-seq data. The evaluation metrics are (1) Spearman's rank correlation coefficient (A-F), and (2) Brownian distance coefficient (G-L). Different colors in the subfigures represent different methods (yellow: HiCPlus, red: HiCNN2-1, blue: DeepHiC) while darker color represents models trained with mESC PLAC-seq and lighter color represents models trained with GM12878 HiChIP data. In each subfigure, we show how the evaluation metric (Y-axis) changes with genomic distance (X-axis), where the distance ranges from 20Kb-2Mb with an increment of 10Kb. The gray line represents the baseline (i.e., low depth data without any enhancement).

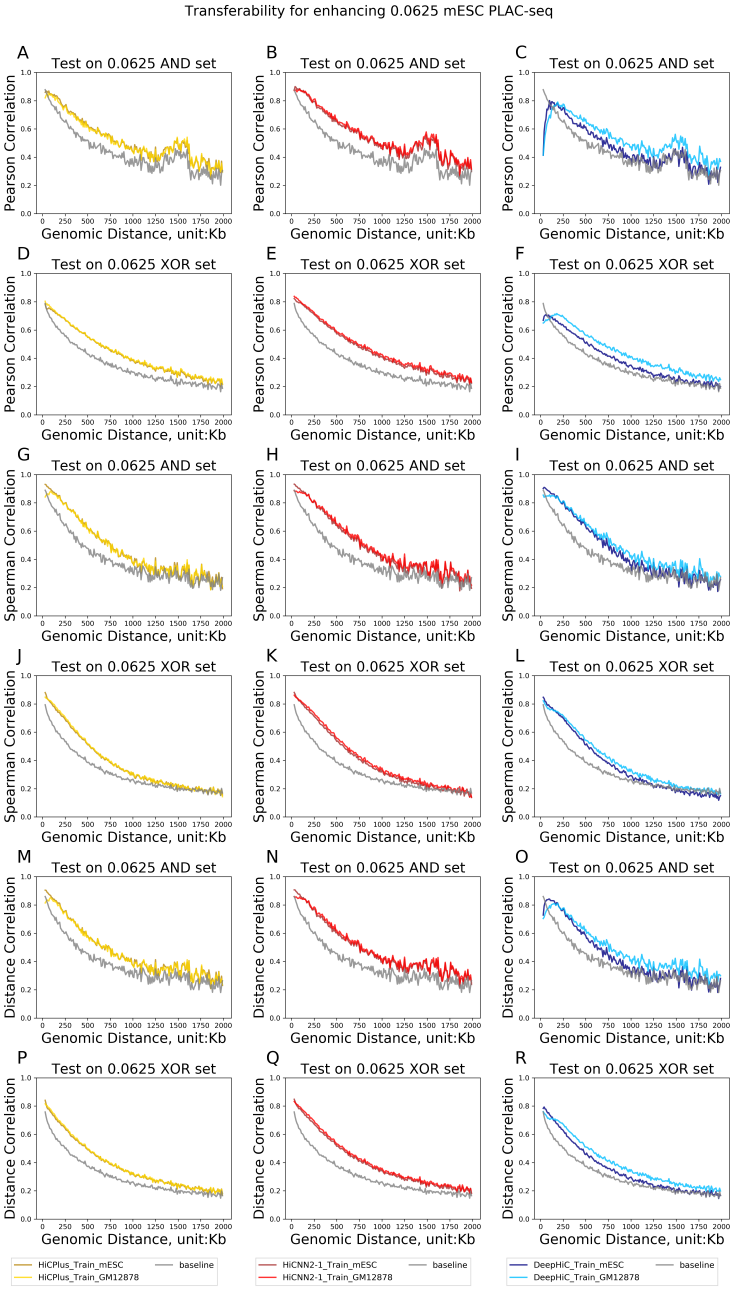

Figure S3.19: **Model transferability when enhancing mESC PLAC-seq data by 25×**. Each subfigure compares two enhanced mESC PLAC-seq data: one using models trained with GM12878 HiChIP data and the other using models trained with mESC PLAC-seq data. The evaluation metrics are (1) Pearson Correlation coefficient (A-F), (2) Spearman's rank correlation coefficient (G-L), and (3) Brownian distance coefficient (M-R). Different colors in the subfigures represent different methods (yellow: HiCPlus, red: HiCNN2, blue: DeepHiC) while darker color represents models trained with mESC PLAC-seq and lighter color represents models trained with GM12878 HiChIP data. In each subfigure, we show how the evaluation metric (Y-axis) changes with genomic distance (X-axis), where the distance ranges from 20Kb-2Mb with an increment of 10Kb. The gray line represents the baseline (i.e., low depth data without any enhancement).

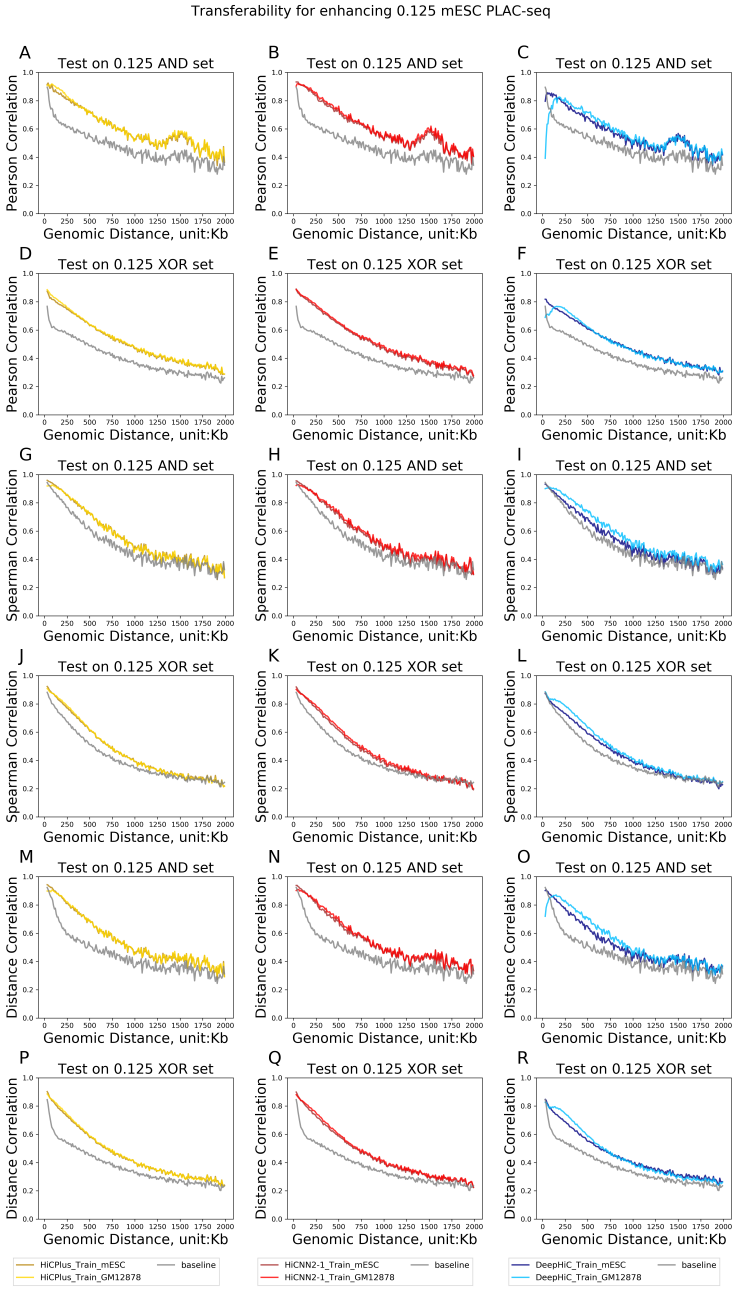

Figure S3.20: **Model transferability when enhancing mESC PLAC-seq data by 16×**. Each subfigure compares two enhanced mESC PLAC-seq data: one using models trained with GM12878 HiChIP data and the other using models trained with mESC PLAC-seq data. The evaluation metrics are (1) Pearson Correlation coefficient (A-F), (2) Spearman's rank correlation coefficient (G-L), and (3) Brownian distance coefficient (M-R). Different colors in the subfigures represent different methods (yellow: HiCPlus, red: HiCNN2, blue: DeepHiC) while darker color represents models trained with mESC PLAC-seq and lighter color represents models trained with GM12878 HiChIP data. In each subfigure, we show how the evaluation metric (Y-axis) changes with genomic distance (X-axis), where the distance ranges from 20Kb-2Mb with an increment of 10Kb. The gray line represents the baseline (i.e., low depth data without any enhancement).

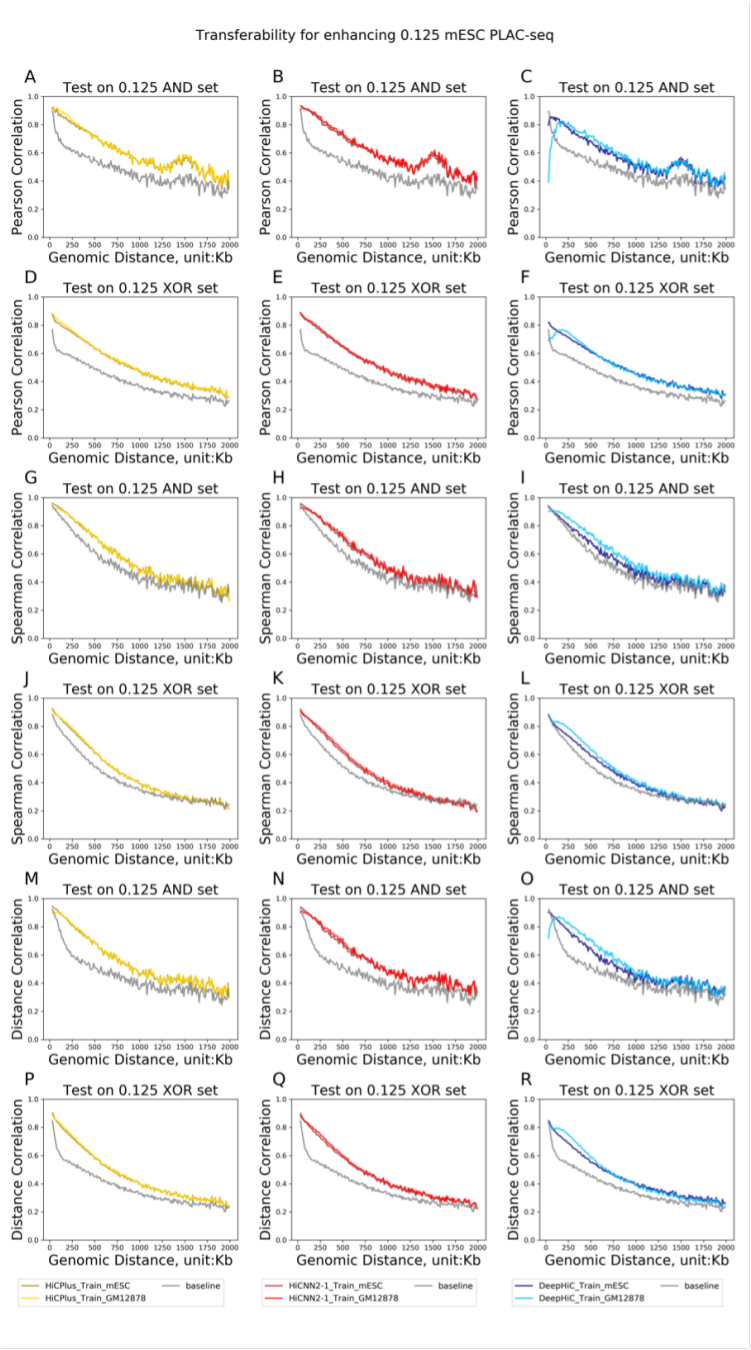

Figure S3.21: **Model transferability when enhancing mESC PLAC-seq data by 8×**. Each subfigure compares two enhanced mESC PLAC-seq data: one using models trained with GM12878 HiChIP data and the other using models trained with mESC PLAC-seq data. The evaluation metrics are (1) Pearson Correlation coefficient (A-F), (2) Spearman's rank correlation coefficient (G-L), and (3) Brownian distance coefficient (M-R). Different colors in the subfigures represent different methods (yellow: HiCPlus, red: HiCNN2, blue: DeepHiC) while darker color represents models trained with mESC PLAC-seq and lighter color represents models trained with GM12878 HiChIP data. In each subfigure, we show how the evaluation metric (Y-axis) changes with genomic distance (X-axis), where the distance ranges from 20Kb-2Mb with an increment of 10Kb. The gray line represents the baseline (i.e., low depth data without any enhancement).

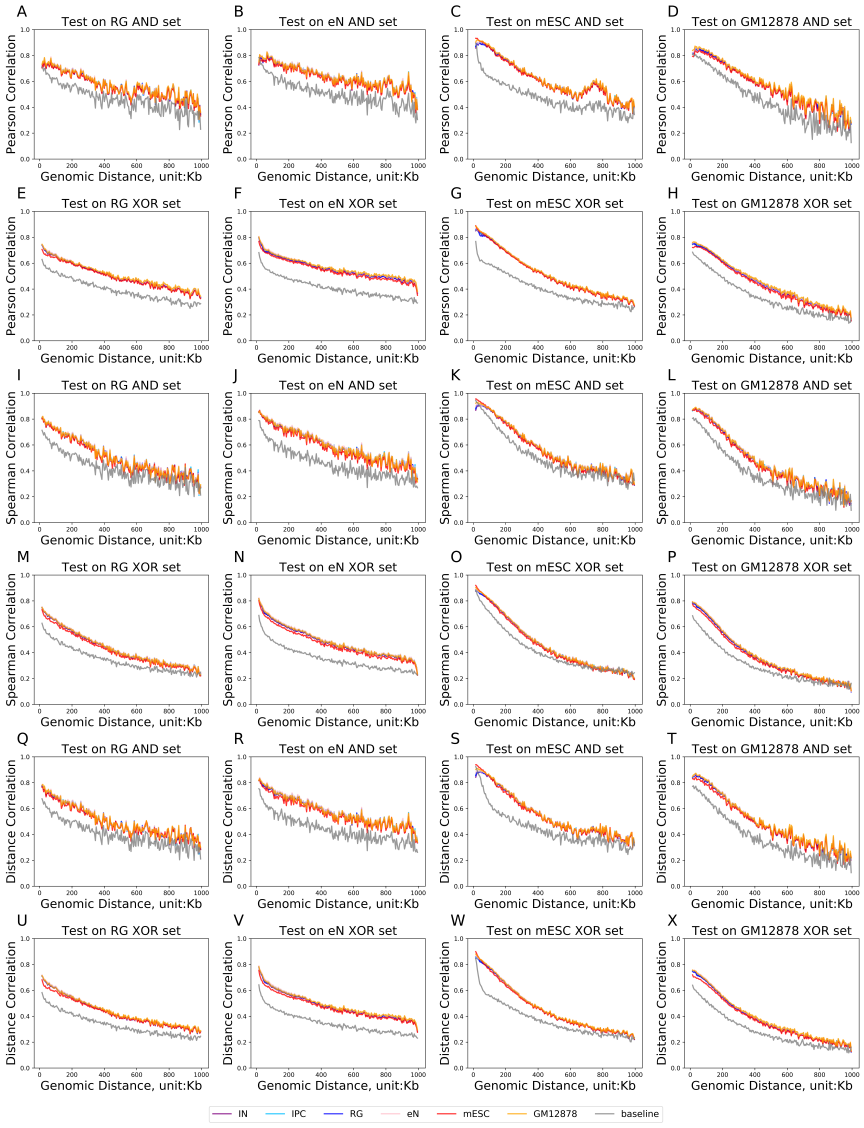

Figure S3.22: **Model transferability among six cell types, measured by correlation of contact matrices**. All results are from HiCNN2-1 models. We test (i.e., perform enhancement) on four cell types: radial glia (RG, 1st column), excitatory neurons (eN, 2nd column), mESC (3rd column) and GM12878 (4th column), all with down-sampling ratio 0.125. The enhancement models are trained using HP data from each of the following six cell types: interneurons (IN), intermediate progenitor cells (IPC), RG, eN, mESC or GM12878. The gray line represents the baseline (i.e., low depth data without any enhancement).

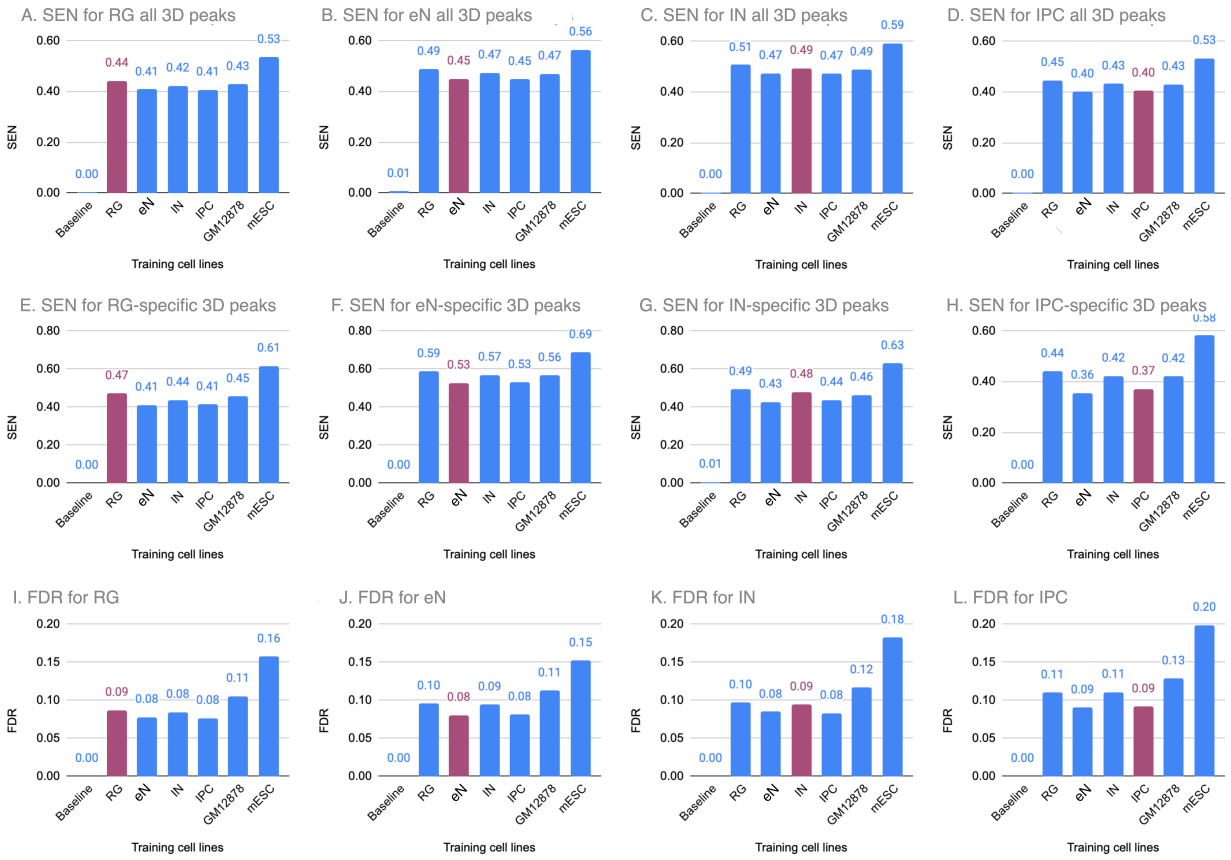

Figure S3.23: **Model transferability among six cell types, measured by 3D peak calling performance**. All results are from HiCNN2-1 models. We test (i.e., perform enhancement) on four neural cell types: radial glia (RG, 1st column), excitatory neurons (eN, 2nd column), interneuron (IN, 3rd column) and intermediate progenitor cells (IPC, 4th column), all with down-sampling ratio 0.125. The enhancement models are trained using HP data from each of the following six cell types: RG, eN, IN, IPC, GM12878 or mESC. Top panel (A-D) shows sensitivity (SEN) for calling all the 3D peaks (i.e., chromatin interactions) identified in a cell type from full data before down-sampling. Middle panel (E-H) shows SEN for calling 3D peaks identified exclusively in the testing cell type, again from full data before down-sampling. Bottom panel (I-L) shows FDR in the four testing cell types where the non-peaks or background bin pairs are also established from full data before down-sampling. Magenta color indicates scenarios when testing and training cell type matches. Baseline represents performance with low depth data without any enhancement. mESC-trained models tend to result in both higher sensitivity and higher FDR possibly due to a combination of two reasons. First, mESC's full data has the highest depth ($\sim$1.1 billion raw reads). Second, we adopt a minimum contact criterion when calling 3D peaks with the MAPS pipeline.

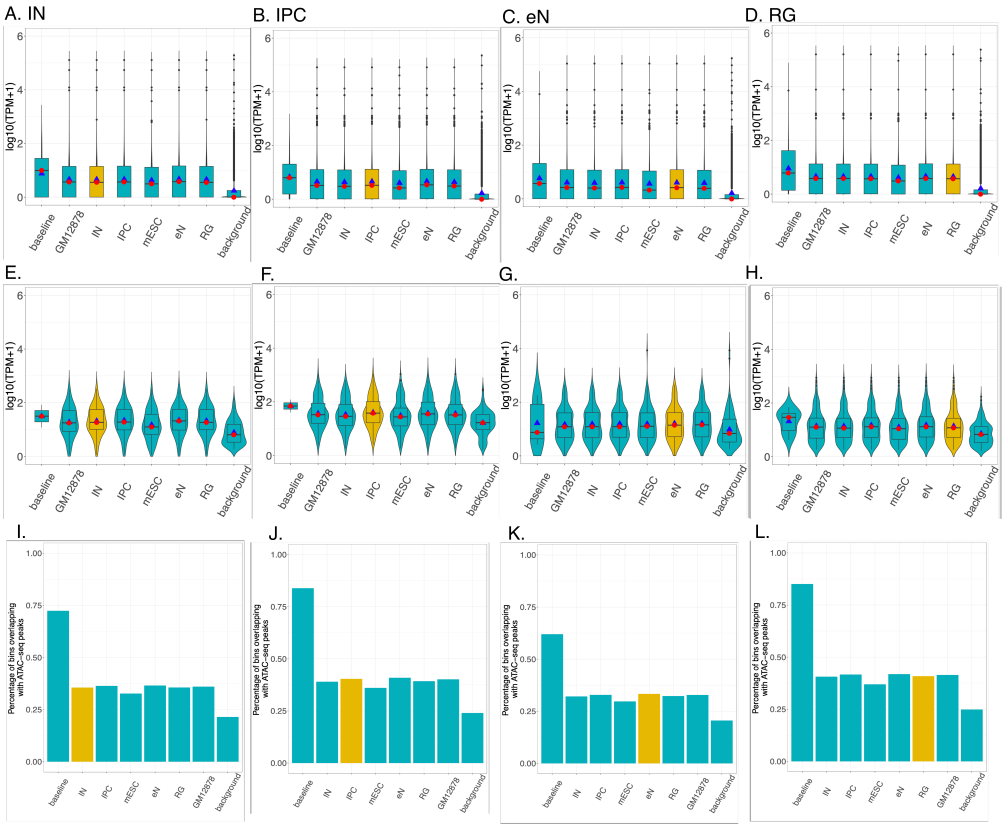

Figure S3.24: **Model transferability assessed by cell-type-specific gene expression and open chromatin status**. All results are from HiCNN2-1 models. We test (i.e., perform enhancement) on four neural cell types: interneuron (IN, 1st column), intermediate progenitor cells (IPC, 2nd column), excitatory neurons (eN, 3rd column) and radial glia (RG, 4th column), all with down-sampling ratio 0.125. The enhancement models are trained using HP data from each of the following six cell types: IN, IPC, mESC, eN, RG, or GM12878. Top panel shows gene expression profiles for eight sets of genes: (1) "baseline": genes whose promoter region (+/-500bp of TSS) involves in some significant chromatin interaction(s) at baseline (i.e., from low depth data before enhancement; (2-7) the six "enhance" sets, each enhanced with a model trained from one of the six cell types: genes whose promoter region does not involve in any significant chromatin interaction at baseline, but in some significant chromatin interactions after enhancement; and (8) "background": genes whose promoter region does not involve in any significant chromatin interactions even with the full data (without any down-sampling). Middle panel is identical to the top panel, except that now we restrict only the cell type specific genes identified in Song et al 2020. Note that baseline contains a very small number of cell type specific genes (2, 2, 5, and 8 for IN, IPC, eN and RG respectively). Background also reduces to a much smaller set of genes (ranging 28,424-30,583 with mean/median 29,562/29,620 in the top panel to ranging 105-237 with mean/median 145/119 in the middle panel). Bottom panel shows the proportion of bins involved in significant chromatin interactions (except for "background") that overlap cell ATAC-seq peaks in the cell type, where significant interactions are defined either from data without enhancement ("baseline"), or from enhanced data with models trained using one of the six cell types. Background bins are bins which are not involved in any significant chromatin interactions even using the full data without any enhancement. Yellow bars indicate that the enhanced data are generated from models trained using the same cell type as the testing cell type.

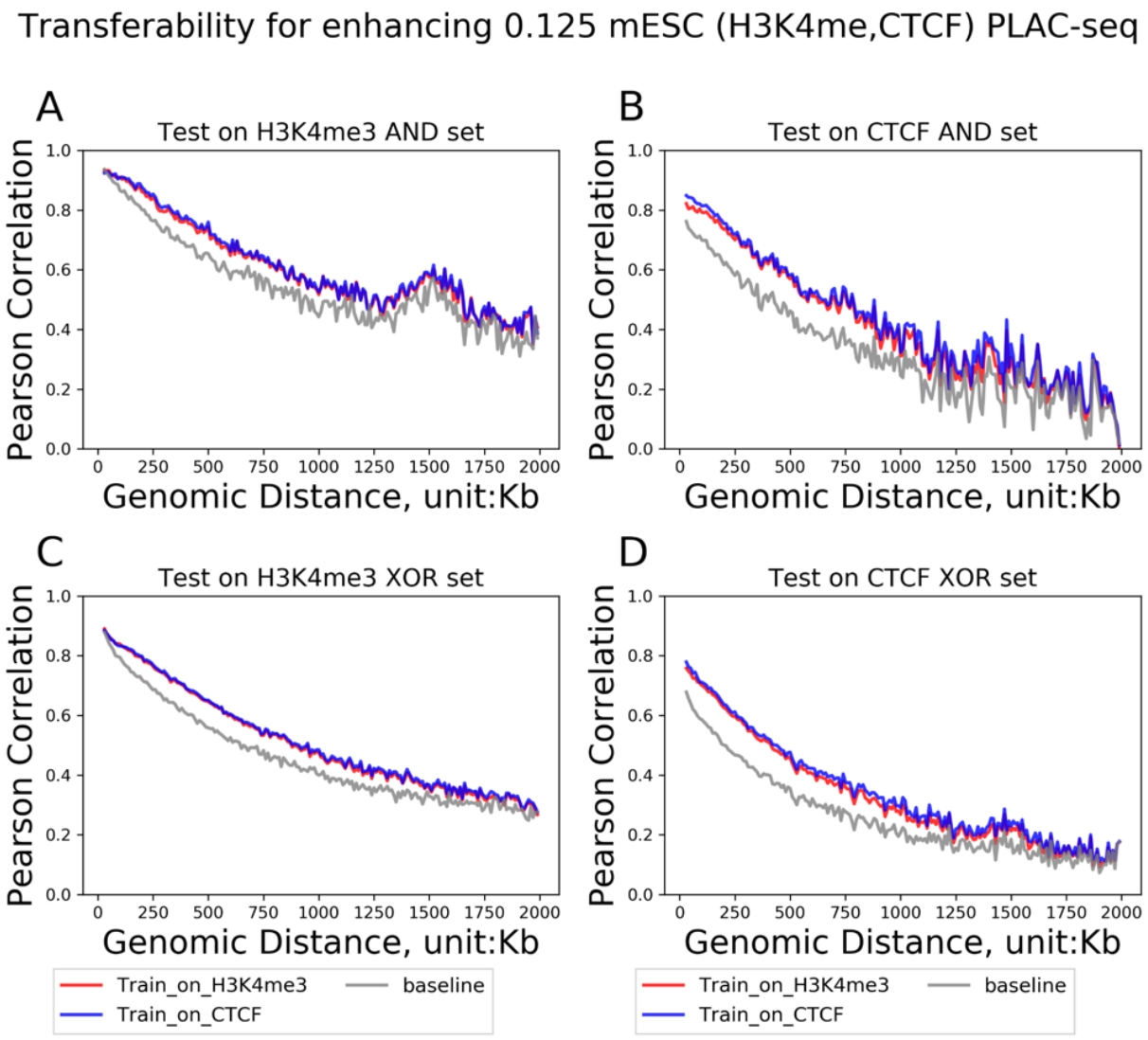

Figure S3.25: **Transferability across different proteins of interest**. All results are based on HiCNN2-1 models. Transferability is assessed between two PLAC-seq datasets in mESC: one H3K4me3 and the other CTCF. Down-sampling ratio is 0.125 for either dataset.

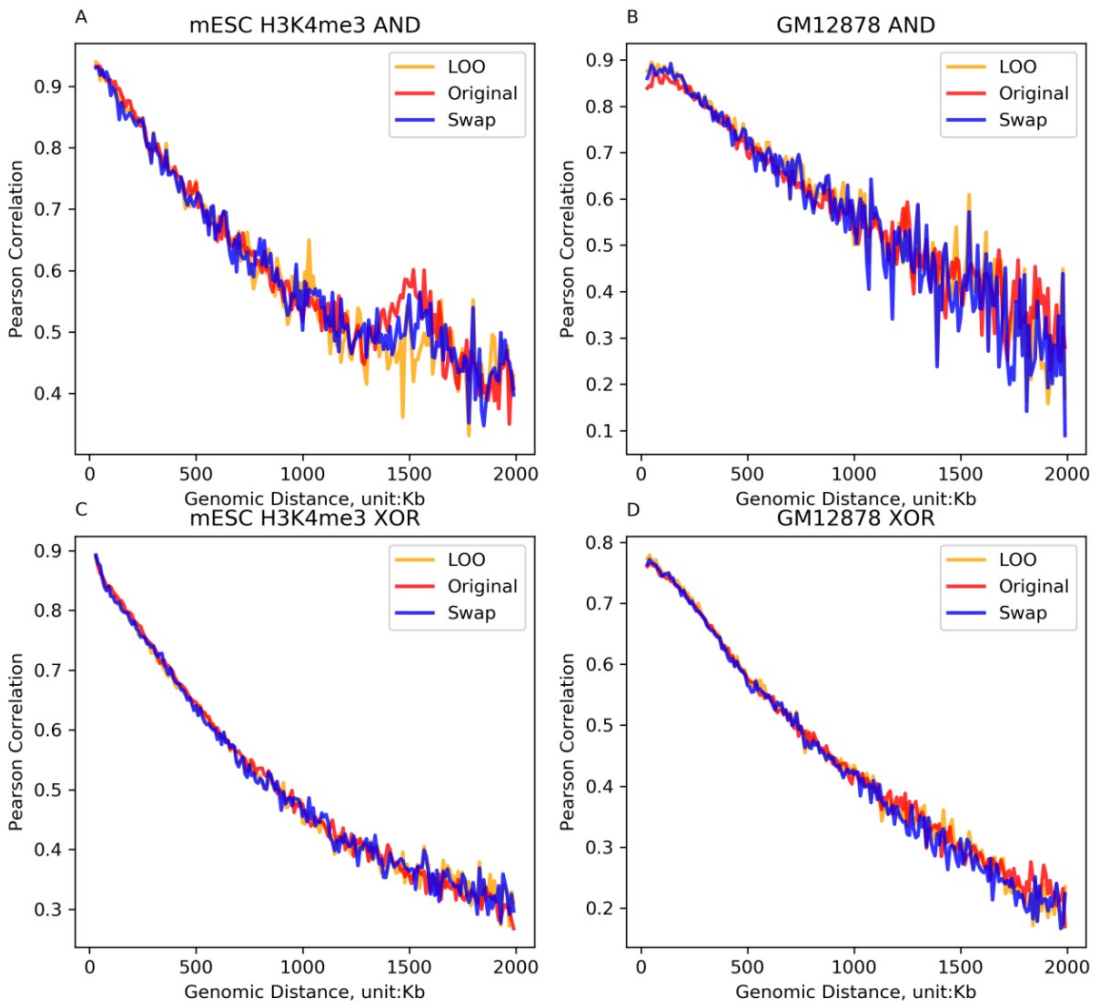

Figure S3.26: **Impact of choice of chromosomes used for training and testing on enhancement performance**. All results are based on HiCNN2-1 models. For all three choices (i.e., "Original", "Swap", "LOO" [leave one out]), chromosome 2 is reserved as the validation dataset which is used for selecting the best model as part of the training process. In "Original", chromosomes 1,3,5,7,9 are used as training and the rest as testing; in "Swap", chromosomes 1,3,5,7,9 are used as testing and the rest as training; in "LOO", we consider only five chromosomes 1 and 3-6 (in addition to the validation chromosome 2), training a model using four chromosome at a time (by leaving one chromosome out), and testing on the left-out chromosome (in "LOO", we repeat the LOO five times to obtain enhanced data for all five chromosomes: 1 and 3-6). Panel A: Performance for AND pairs when enhancing 1/8 depth mESC H3K4me3 PLAC-seq data; Panel B: Performance for AND pairs when enhancing 1/8 depth GM12878 Smc1a HiChIP data; Panel C: Performance for XOR pairs when enhancing 1/8 depth mESC H3K4me3 PLAC-seq data; Panel D: Performance for XOR pairs when enhancing 1/8 depth GM12878 Smc1a HiChIP data. The larger variation in LOO is likely due to smaller number of bin pairs evaluated.

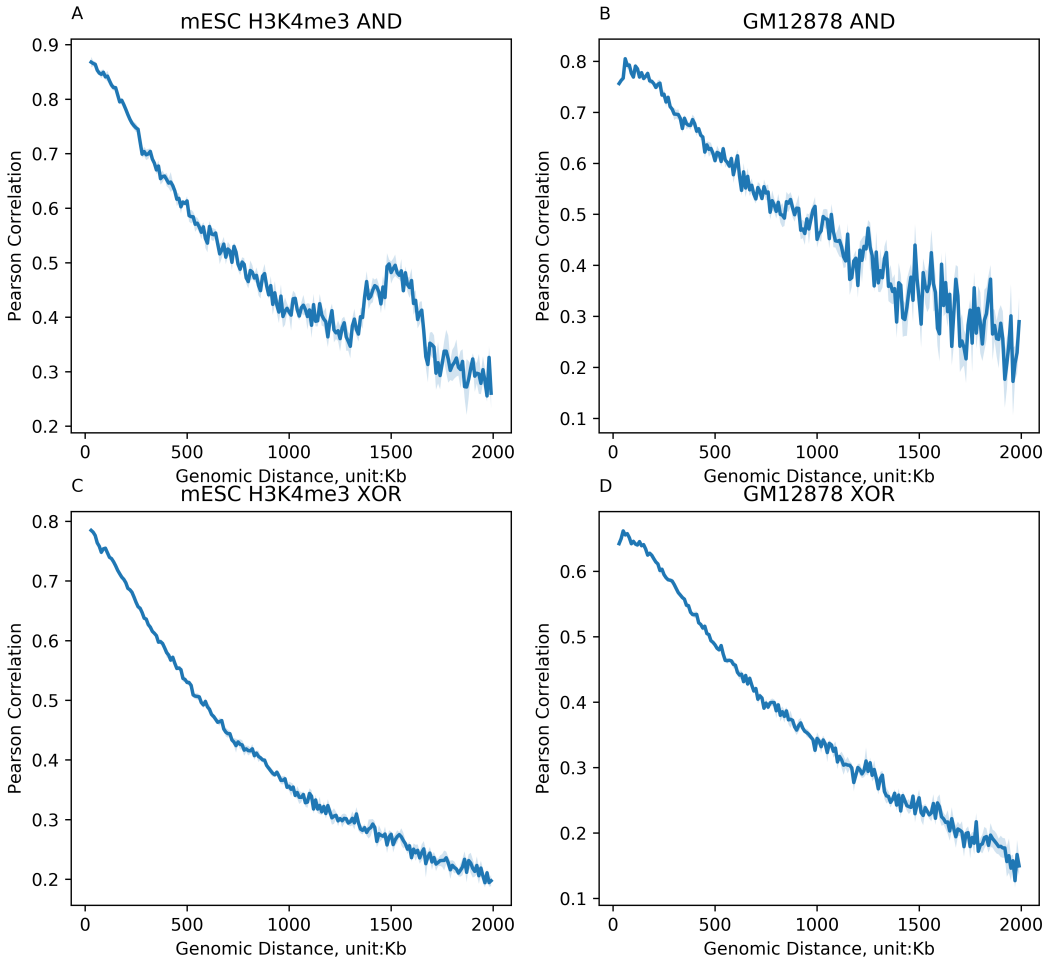

Figure S3.27: **Performance of same 0.04 down-sampling ratio, across five times, on GM12878 HiChIP data**. To evaluate the variability due to random down-sampling, we repeated the down-sampling five times, enhanced data, and assessed the performance using Pearson correlation. All results are based on HiCNN2-1 models. The dark blue curve and light blue bars represent the mean and standard deviation respectively.

CHAPTER 4: METABOLITE PREDICTION MODELS IN UK BIOBANK: A METABOLOME-WIDE ASSOCIATION STUDY (MWAS)

4.1 Introduction

Metabolites are tiny molecules that are essential to physiological and biological processes. They may be detected in a wide range of sample types, including tissues, cells, and biofluids (plasma, urine, and cerebrospinal fluid or CSF) (Donatti et al., 2020). Changes in the amounts of metabolites might indicate presence of a potential illness by impacting cellular processes. A wide range of diseases affecting various tissues, organs, and biological pathways, including metabolic syndromes, diabetes, cardiovascular diseases, and renal disorders, have been linked to plasma metabolites (Kim et al., 2019; Guasch-Ferré et al., 2017; Rios et al., 2023; Wang et al., 2022; Carioca et al., 2021), with plasma circulating throughout the body and containing metabolites produced by multiple tissues. Limited sample size in customized metabolite data present a barrier as small sample size implies impaired statistical power and less accurate or even unstable estimate of effect sizes. Although one could increase sample size by directly measuring metabolites in more samples, such an approach entails extra and often prohibitive costs. In the meantime, we have large numbers of samples with genotype data, presenting a missed opportunity. To take this opportunity, developing a strategy for metabolite prediction from genotype data is crucial. After we obtain predicted metabolite levels from genotype data, we will be able to perform trait-metabolite association in all these samples for whom we have genotype data.

The approach described above, predicting metabolites from genotypes, and then performing association between genotype-predicted metabolites and phenotypic traits, is highly similar to the widely adopted TWAS approach where gene expression (rather than metabolite) is the molecular phenotype in between genotype and traits. Specifically, TWAS techniques were developed to utilize reference expression quantitative trait loci (eQTL) datasets, such as GTEx (Consortium,

2020) and DGN (Battle et al., 2014), for selecting genetic variants that collectively predict gene expression levels. These techniques involve applying weights, derived from eQTL datasets to cohorts with genotype data but lacking gene expression data. This allows for the inference of gene expression levels among individuals without directly measured gene expression data, and subsequent association between genotype-inferred gene expression and phenotypes. Like gene expression, serum metabolite levels and proteins are also heritable (Kiddle et al., 2015; Rhee et al., 2016; Li et al., 2022), with an average genetic contribution to phenotypic variance around 50%, varying across metabolite classes (Hagenbeek et al., 2020). Moreover, metabolomics GWAS studies have discovered metabolite quantitative trait loci (mQTL) linked to serum and plasma metabolites (Shin et al., 2014; Long et al., 2017; Yin et al., 2022; Richardson et al., 2022). This indicates the potential for predicting metabolites using genetic data, akin to transcript prediction in TWAS. However, genetic variants that can contribute to the prediction models differ for these two contexts. In gene expression prediction models, most published studies focus primarily on local (or "cis") variants, which substantially reduces the dimensionality when building the prediction models. However, for metabolites, it is non-trivial, if not impossible, to define reasonable local SNPs. Rather, one has to search for variants from anywhere in the genome.

In this study, we aim to refine and apply the TWAS methodology, devising an innovative metabolome-wide association study (MWAS) approach, in particular to accommodate the mQTLs have to be searched genome-wide. Utilizing targeted high-throughput metabolomics data in the UK Biobank (UKBB), which includes 161 identified metabolites following quality control(Sudlow et al., 2015), we build metabolite prediction models from 96,494 participants of European ancestry (EUR) who have directly measured metabolites and genotype data. This is achieved by using an elastic net modeling framework, and we benchmark the model's performance against the recently calculated genomic risk scores for metabolites from OMICSPRED (`https://www.omicspred.org/`). We also assess the performance of the prediction models (again, built using EUR participants) on 4,780 non-EUR ancestry UKBB participants with measured metabolite data, including 554 East Asian (EAS), 2,198 South Asian (SAS), 2,028 African (AFR) participants, among others.

Furthermore, we conduct an MWAS analysis on 333,501 EUR UKBB individuals who have genotype but lack metabolite data and compare MWAS results with associations obtained from participants with actual measured metabolites. My MWAS involves performing associations between predicted metabolites and 29 hematological traits, as well as 798 diseases categorized under the International Classification of Diseases, Tenth Revision (ICD-10) codes.

4.2 Results

4.2.1 Overview of MWAS

As shown in Figure 4.1, the construction of the MWAS reference panel initiates by segmenting the UKBB samples, who have measured metabolite data, into subsets designated for mQTL identification, model training, and testing. This methodology is necessitated by the computational infeasibility of using genome-wide common SNPs in most TWAS-like models (including elastic net). As afore-mentioned, distinctly, MWAS requires a genome-wide approach as metabolites, unlike gene expression, are not linked to specific cis-encoding regions. Initially, metabolite Genome-Wide Association Studies (mGWAS) are conducted on roughly 45% of the sample to identify a set of genome-wide nominally significant ($p < 1e\text{-}6$) variants for each metabolite. Subsequently, in another 45% of the sample, elastic net models are trained using these mGWAS-identified variants, coupled with a comparative analysis of each metabolite's heritability against the model's R^2 performance. The final stage involves evaluating the effectiveness of the elastic net-derived weights on the remaining 10% of the sample by calculating the correlation between observed and predicted metabolite measures. Additionally, the performance of our derived weights is benchmarked against those obtained using metabolomics data and genomic risk scores from OMICSPRED (https://www.omicspred.org/), a recent atlas for predicting metabolomics and other multi-omics data.

4.2.2 Metabolite GWAS

In our study, mGWAS were conducted on each of the 161 metabolites present in the UKBB dataset, utilizing a GWAS unrelated sample from the cohort (N = 43,447; relatedness < 0.2, as detailed in the 4.3 Methods section). Across the genome, we observed a considerable variation

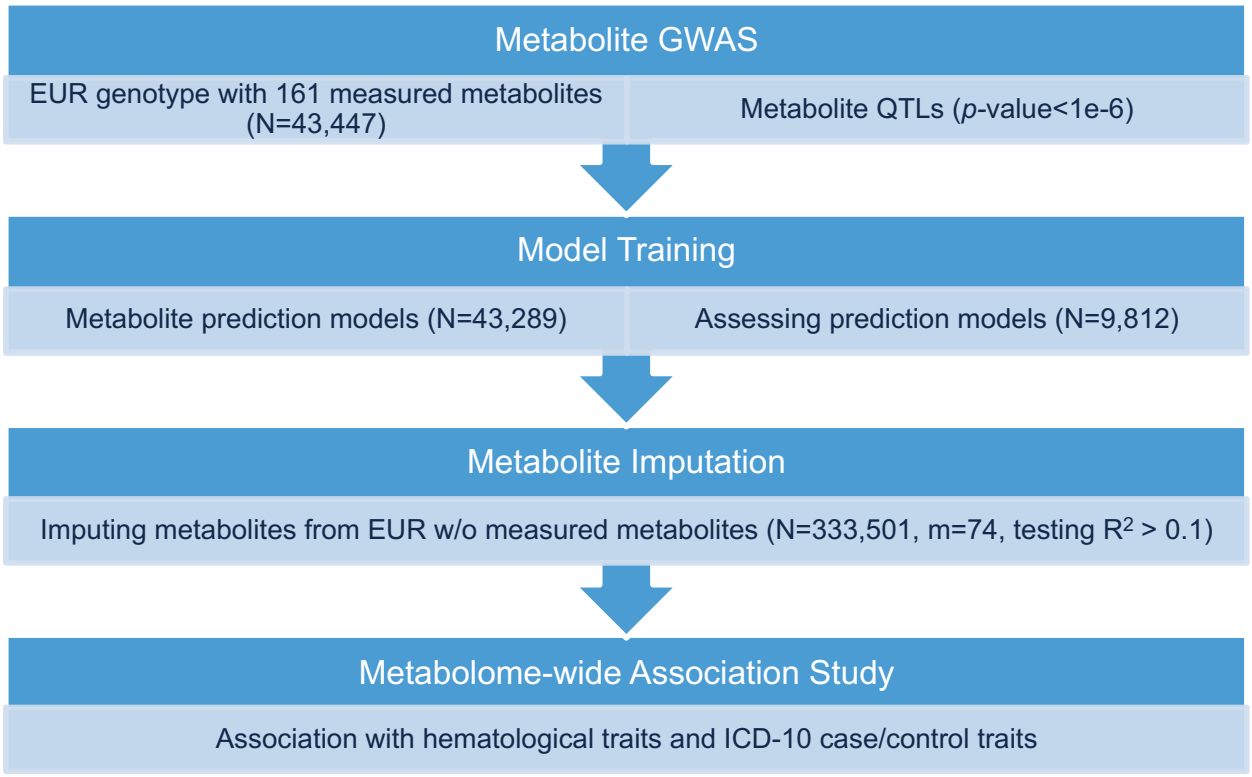

Figure 4.1: Metabolome-wide Association Study Framework

in the number of significant (p $<$ 1e-6) variants associated with these metabolites. The minimum number of significant variants for any metabolite was 120, while the maximum reached 10,006. Notably, the median number of significant variants was 4,624, as depicted in Supplementary Figure S4.1. This range underscores the diverse genetic architecture underlying metabolite variation.

4.2.3 Model Training to Predict Metabolites Using Genetic Variants

Utilizing the significant variants identified from the mGWAS, we started elastic net model training to build predictive models, one for each metabolite. This process of model selection was conducted using the training subset of the cohort (N = 43,289, relatedness $<$ 0.2, as detailed in the 4.3 Methods section). All 161 metabolites included in the model training yielded corresponding elastic net models. The mean and median genetic model R^2 across the 161 metabolites was 0.090 and 0.100, respectively, among UKBB European ancestry participants (Supplementary Figure S4.2).

4.2.4 Metabolite Prediction in Testing Samples

EUR sample With the weights obtained from elastic net models trained using N = 43,289 UKBB EUR participants, we predicted metabolite measures in the testing EUR sample (N=9,812; see 4.3 Methods section).

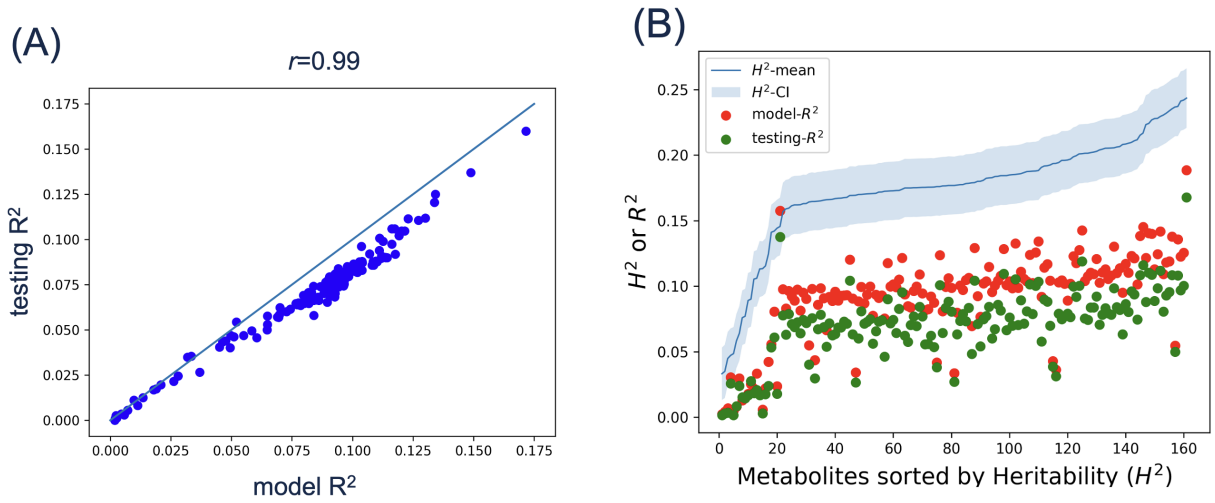

Figure 4.2: The performance of metabolite prediction model. (A) The comparison between the model genetic R^2 and testing genetic R^2. (B) The comparison of three metrics: the upper bound (heritability score, H^2, as estimated by GCTA (Yang et al., 2011)) , model R^2 and testing R^2. In each panel, individual dot represents the model R^2 or testing R^2 for a specific metabolite, representing the distribution and correlation of these datasets.

In general, the genetically regulated component of metabolites was reasonably predicted, with the mean and median genetic prediction R^2 for metabolites 0.071 and 0.074 in the testing dataset, respectively (Figure 4.2). Testing genetic R^2 values, assessing the relationship between observed metabolite measures and genetically predicted metabolite measures are presented in Figure 4.2. Figure 4.2 demonstrates a high correlation between elastic net model R^2 and testing R^2 (r = 0.99), indicating that model overfitting is unlikely an issue in our data. As expected, the right panel of Figure 4.2 reveals that model performance is higher for more heritable metabolites (H^2 estimated by GCTA(Yang et al., 2011)), and there is a strong association between heritability and model performance.

Non-EUR Ancestry Participants With the model weights obtained from EUR participants, we also deployed the models trained in EUR samples to predict metabolites in non-EUR ancestry

participants (N = 4,673), including individuals of predominantly African (AFR, N=1,908), South Asian (SAS, N=2,212), and East Asian (EAS, N=553) ancestry. The mean and median testing genetic R^2 are shown in Table 4.1. Figure 4.3 displays the relationship between testing genetic R^2 values in the EUR sample compared to the non-EUR samples. Generally, prediction in non-EUR samples performed worse than in the EUR sample, which is consistent with our expectations and with the prior TWAS literature (Rowland et al., 2022).

Table 4.1: Mean and Median of prediction values for non-European ancestry group.

	AFR	**SAS**	**EAS**
Mean	0.035	0.041	0.030
Median	0.029	0.038	0.028

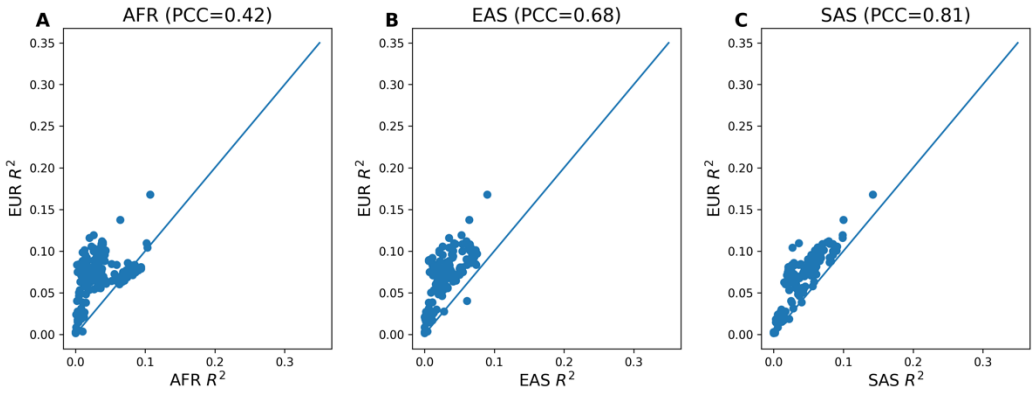

Figure 4.3: Evaluate the relative transferability (in terms of correlation between true and imputed metabolite levels) in held out data from EUR versus non-EUR UKBB participants. (A) represents the correlation between SAS genetic R^2 and EUR genetic R^2 (n=2,212 SAS participants). (B) represents the correlation between AFR genetic R^2 and EUR genetic R^2 (n=1,908 AFR participants). (C) represents the correlation between EAS genetic R^2 and EUR genetic R^2 (n=553 EAS participants). PCC represents the Pearson Correlation Coefficient between non-EUR genetic R^2 and EUR genetic R^2.

4.2.5 Comparison with OMICSPRED

The comparison between our proposed method and the genome-wide risk score model derived from INTERVAL, referred to as OMICSPRED, indicates a strong correlation between the two methods. This is evident from the Pearson Correlation Coefficient (PCC) of 0.85, as displayed in the accompanying figure (Figure 4.4). Both the R^2 from OMICSPRED and our method demonstrate a high degree of similarity, suggesting consistent predictive performance. Notably, our

method exhibits superior predictive capabilities compared to OMICSPRED. This enhancement in performance is expected, given that OMICSPRED was initially trained on the smaller INTERVAL cohort (N=13,668), while our method was pretrained on the UK Biobank dataset, which may contribute to its improved predictive accuracy.

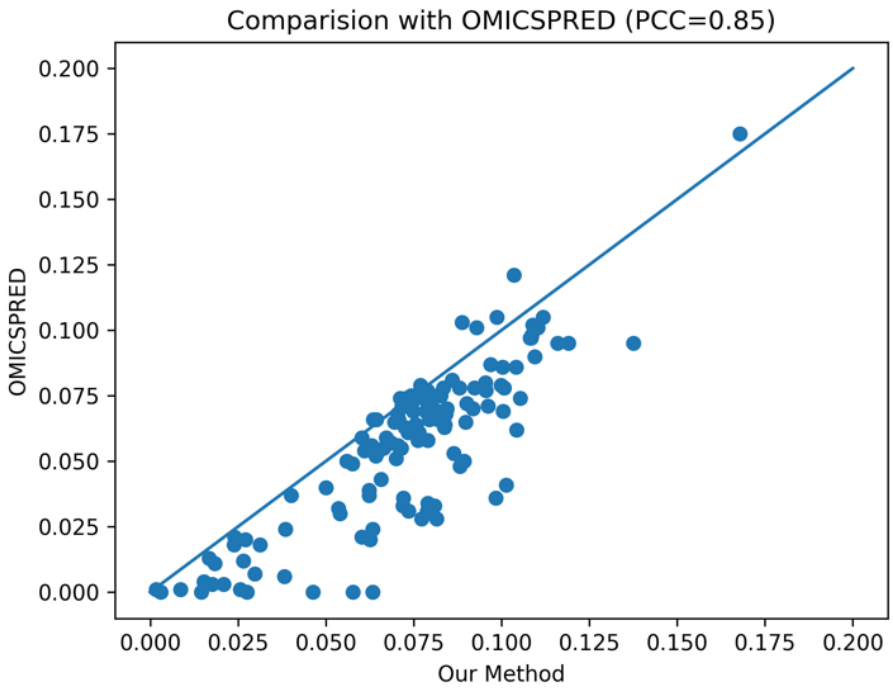

Figure 4.4: OMICSPRED vs Elastic Net Models from UK Biobank. Scatter plot illustrating the correlation between my proposed elastic net model outcomes and OMICSPRED predictions. Data points represent individual metabolites, with a PCC of 0.85, highlighting a strong linear relationship between the two models' predictions.

4.2.6 Phenotype Regressed on Predicted and Measured Metabolites

To assess metabolite associations with various phenotypes, we regressed phenotypic traits (including hematological traits as well as ICD-10 codes for various diseases) on both measured and predicted metabolites. We then compared the two sets of association results.

Blood cell phenotypes We first predicted metabolites for UKBB European participants (N=333,501) without measured metabolites using my pre-trained elastic net models. We regressed 29 blood cell traits (see 4.3 Methods), on these predicted metabolites and compared these associations to those using measured metabolites from 96,494 UKBB European participants. We selected 91 metabolites

(56.5%) with testing $R^2 > 0.1$ in EUR (see 4.3 Methods). Among measured metabolites, we found a median of 83 statistically significant associations (Bonferroni-corrected p-value threshold=1.89e-05) with each blood cell trait, ranging from 48 to 90 significant associations. Among predicted metabolites, we found a median of 58 statistically significant associations (Bonferroni-corrected p-value threshold = 1.89e-05) per blood cell trait, ranging from 3 to 72 significant associations. Additionally, among the shared significant associations (associations significant not only in measured metabolites but also in imputed metabolites), 82.01% of them have the same direction of effect.

ICD-10 code disease phenotype We further performed MWAS on ICD-10 code defined diseases status. We regressed 798 ICD-10 code disease status, on these predicted metabolites and compared these associations to those using measured metabolites from 96,494 UKBB European participants.

For measured metabolites, we found an average of 45 statistically significant associations (Bonferroni-corrected p-value threshold=5.22e-07) with ICD-19 code defined disease status, ranging from 1 to 128. Among predicted metabolites, we found an average of 40 statistically significant associations (Bonferroni-corrected p-value threshold = 5.22e-07) with each ICD-10 disease, ranging from 1 to 119. Additionally, among the shared significant associations (associations significant not only in measured metabolites but also in imputed metabolites), only 60.5% of these associations have the same direction of effect.

4.3 Methods

4.3.1 UK Biobank Data Preprocessing

UKBB recruited approximately 500,000 individuals aged 40-69 years in 2006-2010 to form a biobank study. This cohort's imputed genotype data includes 90 million genetic variants. Genome-wide genotyping and genotype imputation data are available for all participants. Furthermore, 117,981 participants have 249 metabolic measures quantified using the Nightingale Health nuclear magnetic resonance (NMR) platform (for the first release of metabolomics data).

The initial step in our MWAS framework involves quality control for both the metabolite and genotype data. Specifically, for metabolite data, we filtered out samples flagged with "Low Protein" suggesting a possible significant problem with sample dilution (N=867) according to the tutorial

of UKBB nightingale metabolic biomarkers phase 1 release information (https://biobank.ndph.ox.ac.uk/showcase/showcase/docs/nmrm_companion_doc.pdf), 81 metabolites which are ratio measurements, 7 withdrawn metabolites based on UKBB quality control (QC) documentation, and metabolites with $\geq$10% of values below the lower limit of detection. We also removed samples with missing metabolite measurements (N=7,065) and those who have withdrawn from UKBB (N=5). Regarding genotype data, we retained SNPs with minor allele frequency (MAF) $\geq$ 0.1% and INFO $\geq$ 0.5 (computed by QCTOOL V2, INFO scores were precomputed by UKBB). After QC, our analysis involved 96,494 unrelated European (EUR), 2,028 African (AFR) ancestry individuals, 2,198 South Asian, 554 Eastern Asian ($<$ 0.2 relatedness using GCTA, based on k-means clustering for similarity to 1000 Genomes continental superpopulations, ancestry clusters estimated as described in our previous TWAS research (Consortium et al., 2010a)), 161 QC+ metabolites, and 19,709,367 QC+ SNPs. Individuals who did not fall into any k-means based ancestry cluster with 1000 Genomes populations were excluded.

4.3.2 Identifying mQTLs

To select the suitable predictors for a metabolite prediction model, we hypothesized that nominally significant metabolite QTLs (mQTLs) identified in the first sub-group can serve as predictors when training elastic net prediction models in the second sub-group. To identify mQTLs, we first performed rank-based inverse normal transformation (INT) to normalize each metabolite. We then used REGENIE (Mbatchou et al., 2021), a fast GWAS tool, to perform association analysis between each metabolite and the QC+ SNPs, adjusting for continues variables age and PC1-PC10, and binary variables sex and Center1-Center22, as well as genotyping array. Based on REGENIE results, we considered variants with p-value $<$ 1e-6 as mQTLs. In our preliminary data, we observed on average 4,624 mQTLs per metabolite using this nominal threshold (Supplementary Figure S4.1).

4.3.3 Developing Metabolite Prediction Models

Based on the mQTLs identified in the previous step, we extracted dosage from UKBB genotype imputation data in the second group of 43,289 individuals and combined them to train an elastic net model. We fine-tuned the model with 10-fold cross validation (CV). Our model contains

genetic variants defined by mQTLs and same covariates as in metabolite GWAS (e.g., age, sex, recruitment centers, top 10 principal components (PCs), and genotyping array). When building the elastic net models, there are several technical details to consider:

1) Although elastic net models can achieve variable selection, we still need to perform LD pruning (removing SNPs with $r > 0.95$) before running elastic net. This is because variants in strong LD ($r > 0.95$) might both end up in the final trained models (based on our preliminary experiments), which will lead to high instability in the estimated regression coefficients for these highly correlated predictors.

2) When tuning the parameters, we employed a grid search to look of optimal alpha and lambda. We then used the best pair of parameters (alpha and lambda) to further fine-tune the model.

4.3.4 Assessing Model Performance

Lastly, we assessed the performance of the models constructed by applying them to an independent held-out group of 9,812 EUR, 553 EAS, 2,212 SAS, and 1,908 AFR individuals. For each metabolite prediction model, we determined the testing genetic R^2, which quantifies the squared correlation between the imputed and measured metabolite values in this testing dataset. The testing dataset was selected to be independent of the samples used in the initial two groups for building the prediction models, with a relatedness threshold of less than 0.2 as estimated by GCTA (Yang et al., 2011). Genetic R^2 is a measure of the genetic contribution to metabolite variance after adjusting for covariates. Each genetic R^2 is calculated by predicting metabolite levels using our model that includes both genetic variants and covariates and then we compute the residuals by regressing out covariates. A similar process is applied to the measured metabolite levels to obtain a second set of residuals. The genetic R^2 is the R^2 between these two sets of residuals, reflecting how well the genetic variants alone can explain the variation in metabolite levels. To investigate the presence of overfitting, we compare the model genetic R^2 from the training set with the testing genetic R^2 using Pearson's Correlation Coefficient.

4.3.5 Phenotype of Interest

Blood Cell Traits We proceeded with MWAS analysis on 29 hematological traits for metabolites with testing $R^2 > 0.1$. We referred to these metabolites as predicted metabolites. Hematological traits or blood cell traits are critical characteristics of overall health and can offer insights into various physiological and pathological states. Since our metabolites data were measured from plasma, it is naturally to examine whether the imputed metabolites are associated with blood cell traits to test our MWAS framework.

In the context of our study, we performed an analysis on twenty-nine different blood cell traits: 14 indices related to red blood cells, 4 indices associated with platelets, and 11 indices for white blood cells (Supplementary Table S4.1). These traits were measured using blood samples obtained during visits to the UKBB assessment centers by approximately 470,000 individuals. We then adjusted these blood cell trait measures for age, sex, and assessment centers.

ICD-10 code defined disease We also tested associations between imputed metabolites and a battery of ICD-10 code defined disease status. Specifically, the occurrence of a disease was identified by the first appearance of a 3-character ICD-10 or ICD-9 code, extracted from UKB's hospital inpatient records or death registry data (refer to Table 4.1 for more details). Subsequently, we converted ICD-9 codes to ICD-10 codes using the general equivalence mappings provided by the Center for Disease Control (https://ftp.cdc.gov/pub/Health_Statistics/NCHS/Publications/ICD10CM/2018/). Diseases with < 200 cases were excluded from the analysis. This resulted in a total of 798 binary disease outcomes for the population involved in the study. Supplementary Table S4.2 indicates the related ICD-10 code fields from UK Biobank dataset.

4.3.6 MWAS

We performed MWAS with predicted metabolites and traits of interest listed in Supplementary Table S4.1-S4.2. Specifically, our analysis procedure contained the following two steps:

(1) We performed metabolites prediction for UKBB individuals with genotype data but without measured metabolite data. There is a total of 333,501 European ancestry, unrelated (relatedness < 0.2, estimated by GCTA (Yang et al., 2011)) individuals. We name it as target data.

(2) We fitted linear regression model (for continuous traits, e.g., hematological traits) or logistic regression model (for binary traits, e.g., ICD-10 code defined disease status) to investigate the association between each imputed metabolite and each trait of interest, adjusted for age, sex, PC1-PC10, genotyping array, and recruitment centers. Bonferroni correction will be used to adjust for multiple testing (alpha=0.05 / (number of tested metabolites $\times$ number of traits)). To evaluate our implicated imputed metabolites, we compared our MWAS association results with results derived from measured metabolites based on $\sim$100K UKBB individuals.

4.3.7 Discussion

Here we adapted the conceptually similar TWAS approach to perform MWAS that uses genetic variants to infer heritable metabolite levels and subsequently associate them with a large set of phenotypic traits of interest. Although building metabolite prediction models by leveraging genome-wide genotypes is a natural thought, having millions of predictors is computationally burdensome and not feasible with elastic net or similar sparse model commonly used for TWAS like approaches. Moreover, the majority of variants are not associated with metabolites levels and do not need to enter the prediction models. We therefore perform a mGWAS first to pre-screen mQTLs before training elastic net models.

We also notice that models pretrained in EUR samples performed less well when applied to non-EUR ancestry participants, consistent with the poor transferability in PRS literature (Martin et al., 2019). We also found that testing R^2s from Omicspred are smaller than testing R^2s from our elastic net model. One potential explanation is that Omicspred used their own INTERVAL cohort to pretrain the models, while we used UKBB participants to train elastic net models.

In addition, we found some discrepancies in association results when comparing MWAS to results derived from measured metabolites. We hypothesize this is likely due to differences in effects between parts of metabolite variation that are and are not genetically regulated. Integration of

environmental and clinical covariate data may provide additional insights in future. Future studies of larger sample size and in diverse samples may mitigate the inconsistencies, leading to consistent or complementary information that provide biological insights into the role of metabolites in shaping health and disease related outcomes.

Web ResourcesQCTOOL V2: https://www.well.ox.ac.uk/~gav/qctool_v2/

4.4 Supplementary Materials

This chapter provides supplementary figures and tables for this work

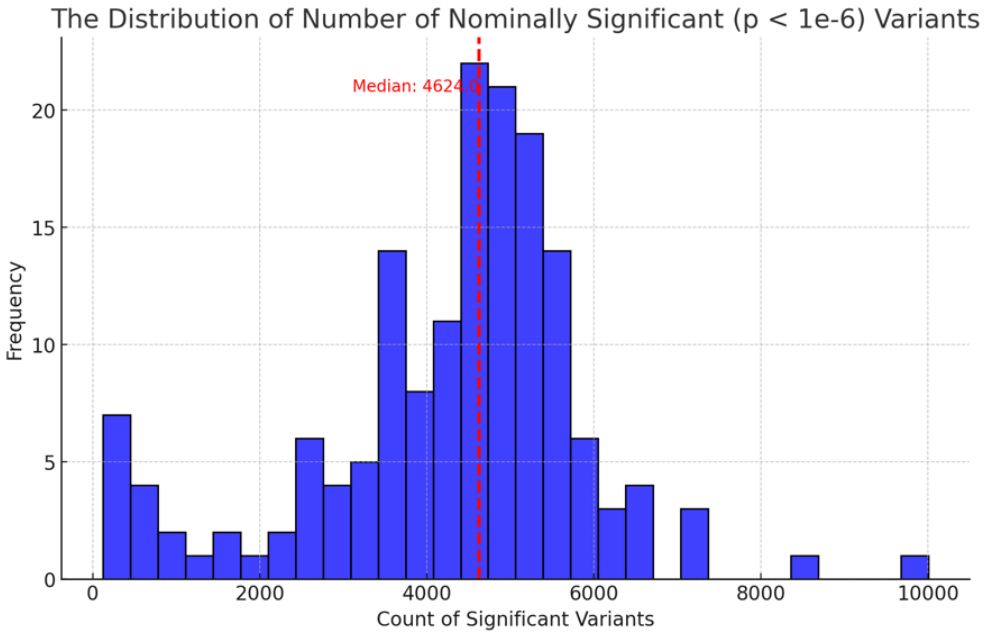

Figure S4.1: Distribution of the number of nominally significant (p $<$ 1e-6) GWAS variants (SNPs) per metabolite used for model training.

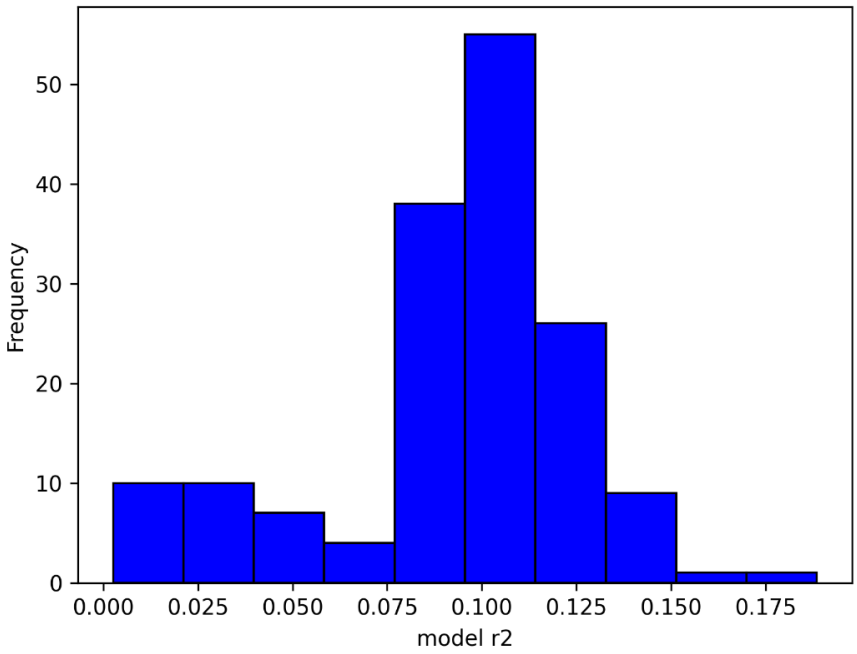

Figure S4.2: The distribution of model R^2. It contains model R^2 from 161 metabolite prediction models. The Mean and median model genetic R^2 are 0.09 and 0.10 respectively.

Table S4.1: List of phenotypes with their corresponding class and abbreviation.

Phenotype	Phenotype Class	Abbreviation
Basophil Count	WBC	BAS#
Basophil Percentage	WBC	BAS%
Eosinophil Count	WBC	EOS#
Eosinophil Percentage	WBC	EOS%
Hematocrit Percentage	RBC	HCT
Hemoglobin Concentration	RBC	HGB
Hls Reticulocyte Count	RBC	HLSR#
Hls Reticulocyte Percentage	RBC	HLSR%
Immature Reticulocyte Fraction	RBC	IRF
Lymphocyte Count	WBC	LYMPH#
Lymphocyte Percentage	WBC	LYMPH%
Mean Cell Hemoglobin	RBC	MCH
Mean Cell Hemoglobin Concentration	RBC	MCHC
Mean Cell Volume	RBC	MCV
Mean Platelet Volume	PLT	MPV
Mean Reticulocyte Volume	RBC	MRV
Mean Sphered Cell Volume	RBC	MSCV
Monocyte Count	WBC	MONO#
Monocyte Percentage	WBC	MONO%
Neutrophil Count	WBC	NEUT#
Neutrophil Percentage	WBC	NEUT%
Platelet Count	PLT	PLT#
Platelet Distribution Width	PLT	PDW
Plateletcrit	PLT	PCT
Red Blood Cell Count	RBC	RBC#
Red Blood Cell Width	RBC	RDW
Reticulocyte Count	RBC	RET#
Reticulocyte Percentage	RBC	RET%
White Blood Cell Count	WBC	WBC#

Table S4.2: The UKBB data resources used for disease outcome definition.

Field ID	Description	Category
40001	Underlying (primary) cause of death: ICD10	Death Register
40002	Contributory (secondary) causes of death: ICD10	Death Register
41202	Diagnoses - main ICD10	Hospital inpatient
41271	Diagnoses - ICD9	Hospital inpatient

CHAPTER 5: CONCLUSION AND FUTURE WORK

This thesis encompasses three distinct studies aimed at unraveling the intricacies of complex diseases utilizing various omics datasets. Specifically, it concentrates on creating a webtool for LD information derived from TOPMed Whole Genome Sequencing (WGS) data (discussed in Chapter 2), examining the effectiveness of deep learning techniques in enhancing sequencing depth for chromatin interactome data (Chapter 3), and formulating a framework for metabolome-wide association studies (Chapter 4).

5.1 TOP-LD

5.1.1 Summary

In Chapter 2, we introduced TOP-LD which is an enhanced database that was developed to query LD based on TOPMed WGS data. Compared to existing LD proxy query tools HaploReg and LDlink, TOP-LD presents more comprehensive genetic variants: 1. larger amounts of genetic variants. 2. genetic variants on chromosome X. 3. structural variants. TOP-LD can offer a larger number of variants due to higher coverage of TOPMed WGS data compared with 1KGP WGS, as well as larger sample size. In addition, including tens of thousands of SVs in TOP-LD can make it helpful for fine-mapping analysis and structural variants prioritization. The fine-mapping at *GGT1* locus associated with gamma glutamyltransferase in African ancestry participants from the UK Biobank by leveraging TOP-LD is a demonstration application.

5.1.2 Future Direction

A primary area of future work involves enlargement sample sizes of non-European ancestry populations. In the construction of TOP-LD, I used 13,160 TOPMed European ancestry individuals but only 2,418 non-European ancestry individuals, which shows non-European individuals are still underrepresented. By incorporating more diverse populations with more comparable sample sizes,

future iterations of TOP-LD could provide valuable insights into the detection of statistically causal variants or the prioritization of structural variants.

In order to further enhance the utility of TOP-LD, the following improvement methods are proposed. First, the sequencing depth of TOPMed WGS data needs to be further improved, which can expand the number of genetic variants included, thereby improving the LD-proxy groupings in TOP-LD data. Secondly, we must increase the data size of non-European population groups in the TOP-LD data set, and then calculate more accurate LD patterns of other populations, thereby helping to understand the specific genetic structure of specific populations. Third, TOP-LD can be combined with other commonly used genome analysis tools (such as LD score (LDSC) regression and fine-mapping tools, etc.), and the results of these additional tools could also be presented in TOP-LD as functional annotations. Finally, we will keep the TOP-LD database up-to-date and will propose a new version annually. We will also maintain the standalone version of our TOP-LD for users to use our database directly with MySQL commands or a RESTful API using a server. In addition, we will resolve user issues and always keep our website usable.

5.2 DeepCompare

5.2.1 summary

In Chapter 3, I have introduced comprehensive evaluation framework DeepCompare to compare available deep learning models for enhancing the sequencing depth of HP data. My findings indicate that these tools indeed can also enhance HP data. I also find that HiCPlus and HiCNN2 generally outperform DeepHiC in enhancing HP data. Interestingly, models trained on ultra-high-depth Hi-C data often show similar or superior performance compared to those trained directly on HP data. This suggests a practical strategy for researchers to utilize high-depth Hi-C data for enhancing HP data when high-depth HP data are not available in the tissue or cell type of interest. The transferability of these models across different cell types and proteins of interest is a notable aspect of my study. For instance, models trained with high-depth GM12878 data provide better enhancement results in mESC than those trained with similar high-depth mESC data. This implies that the principles underlying the enhancement of lower depth data are consistent across different

cell types and even organisms. Evaluation of these methods using standard metrics (e.g., Pearson Correlation Coefficients, Spearman Correlation Coefficients) and HPrep (Rosen et al., 2021), an extension of HiCRep for HP data, showed that deep learning methods significantly improved the detection of chromatin interactions compared to baseline. According to the comparison results, we recommend users to use HiCNN2 and HiCPlus when enhancing the HP data.

5.2.2 Future Directions

We will improve DeepCompare analysis using multiple methods. Firstly, we can compare with more recent methods such as HiCARN (Hicks and Oluwadare, 2022), EnHiC (Hu and Ma, 2021a), etc. Secondly, we can develop a new enhancement algorithm dedicated to HP data. While DeepCompare shows that leveraging the Hi-C enhancement method for HP data is applicable, this approach is not as natural as directly using a super-resolution method from computer vision for HP data. Concretely, HP data are generated from a more biased sequencing method compared with Hi-C data, which is different from Hi-C data and image data. Therefore, leveraging the rectangular kernel, which is often utilized in super-resolution in computer vision, may help improve coverage the invalid regions (see more details in the HP data definition from Chapter 3). Designing a neural network with specific kernels (kernel in cross shape) which only focus on the regions of interest may improve performance. However, there are no current tools that satisfy this feature. Thirdly, if this kind of algorithm is developed, it should not only have higher accuracy for enhancing HP compared with other methods, but also should have improved measurements in multiple downstream analyses, such as the important interactions detection, high-order chromatin spatial structure identification which includes AB compartments, topologically associated domains and frequently interacting regions.

5.3 Metabolome-wide Association Study

5.3.1 Summary

MWAS framework is designed for detecting the significant associations between the genetically imputed metabolites with a specific trait. This framework has several advantages. First, we efficiently utilized UKBB data to perform mGWAS, model pre-training, model testing, and MWAS.

The mQTLs helps to reduce the computational burden for model training, and to select the most relevant predictors in the model using single variant testing . Furthermore, we utilized this model to impute the metabolites for those individuals without measured metabolites, and then we conduct MWAS between those imputed metabolites and blood cell traits or ICD-10 codes defined diseases. We found that our method generally provides consistent association directions when compared with the associations between measured metabolites and traits.

5.3.2 Future Direction

There are multiple potential future directions for MWAS. Firstly, we can consider different approaches for building metabolite prediction models. For example, we can leverage Bayesian sparse linear mixed models (Zhou et al., 2013a). This kind of method is always utilized in polygenic risk score modeling. We can also utilize adapting Non-parametric Dirichlet Process Regression as utilized in TIGAR (Parrish et al., 2022). TIGAR estimates variance parameters based on the input data and provides a flexible alternative to traditional parametric priors.

Another strategy is to incorporate more cohorts besides UKBB. Specifically, we would like to include more non-European ancestry participants in MWAS model training. Additionally, we will expand metabolites from other tissue besides plasma. Those strategies would provide more insights when applying our framework into other tissues and other populations.

We could also consider furthering our framework to model interactions between genes and environ- mental factors. In other words, we will re-frame our MWAS into Metabolome-wide Interaction Studies (MWIS). Genotype-by-Environment (GEI) is increasingly examined in the PRS area. Factors such as smoking and pregnancy status and medication use may be important to examine. GEI might have a larger effect size than Genotype-only prediction model, which would be beneficial for the association analysis between imputed metabolites and traits of interest.

Furthermore, given that most metabolites exhibit high degrees of correlation, I will extend MWAS by incorporating other ‘omics’, namely transcriptomics and proteomics, to reveal more insights into potential pathways underlying normal biological systems and disease progression.

REFERENCES

Abdellaoui, A., Yengo, L., Verweij, K. J., and Visscher, P. M. (2023). 15 years of gwas discovery: Realizing the promise. *The American Journal of Human Genetics*.

Alghamdi, M. A., O'Donnell-Luria, A., Almontashiri, N. A., AlAali, W. Y., Ali, H. H., and Levy, H. L. (2023). Classical phenylketonuria presenting as maternal pku syndrome in the offspring of an intellectually normal woman. *JIMD reports*, 64(5):312–316.

Amemiya, H. M., Kundaje, A., and Boyle, A. P. (2019). The ENCODE blacklist: identification of problematic regions of the genome. *Scientific Reports*, 9(1):9354.

Antonarakis, S. E. and Beckmann, J. S. (2006). Mendelian disorders deserve more attention. *Nature Reviews Genetics*, 7(4):277–282.

Battle, A., Mostafavi, S., Zhu, X., Potash, J. B., Weissman, M. M., McCormick, C., Haudenschild, C. D., Beckman, K. B., Shi, J., Mei, R., et al. (2014). Characterizing the genetic basis of transcriptome diversity through rna-sequencing of 922 individuals. *Genome research*, 24(1):14–24.

Benner, C., Spencer, C. C., Havulinna, A. S., Salomaa, V., Ripatti, S., and Pirinen, M. (2016). Finemap: efficient variable selection using summary data from genome-wide association studies. *Bioinformatics*, 32(10):1493–1501.

Bild, D. E., Bluemke, D. A., Burke, G. L., Detrano, R., Diez Roux, A. V., Folsom, A. R., Greenland, P., Jacob, 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., Cohen, N. M., Szabo, Q., Fritsch, L., Papadopoulos, G. L., Lubling, Y., Xu, X., Lv, X., Hugnot, J.-P., Tanay, A., et al. (2017a). Multiscale 3d genome rewiring during mouse neural development. *Cell*, 171(3):557–572.

Bonev, B., Mendelson Cohen, N., Szabo, Q., Fritsch, L., Papadopoulos, G. L., Lubling, Y., Xu, X., Lv, X., Hugnot, J.-P., Tanay, A., and Cavalli, G. (2017b). Multiscale 3D genome rewiring during mouse neural development. *Cell*, 171(3):557–572.e24.

Bush, W. S. and Moore, J. H. (2012). Chapter 11: Genome-wide association studies. *PLoS computational biology*, 8(12):e1002822.

Carioca, A. A. F., Steluti, J., de Carvalho, A. M., Silva, A. M., da Silva, I. D. C. G., Fisberg, R. M., and Marchioni, D. M. (2021). Plasma metabolomics are associated with metabolic syndrome: A targeted approach. *Nutrition*, 83:111082.

Choudhury, A., Aron, S., Botigué, L. R., Sengupta, D., Botha, G., Bensellak, T., Wells, G., Kumuthini, J., Shriner, D., Fakim, Y. J., et al. (2020). High-depth african genomes inform human migration and health. *Nature*, 586(7831):741–748.

Conomos, M. P., Reiner, A. P., Weir, B. S., and Thornton, T. A. (2016). Model-free estimation of recent genetic relatedness. *The American Journal of Human Genetics*, 98(1):127–148.

Consortium, . G. P. et al. (2015a). A global reference for human genetic variation. *Nature*, 526(7571):68.

Consortium, G. (2020). The gtex consortium atlas of genetic regulatory effects across human tissues. *Science*, 369(6509):1318–1330.

Consortium, G. P., Auton, A., Brooks, L., Durbin, R., Garrison, E., and Kang, H. (2015b). A global reference for human genetic variation. *Nature*, 526(7571):68–74.

Consortium, I. H. ., Altshuler, D. M., Gibbs, R. A., Peltonen, L., Altshuler, D. M., Gibbs, R. A., Peltonen, L., Dermitzakis, E., Schaffner, S. F., Yu, F., Peltonen, L., Dermitzakis, E., Bonnen, P. E., Altshuler, D. M., Gibbs, R. A., de Bakker, P. I. W., Deloukas, P., Gabriel, S. B., Gwilliam, R., Hunt, S., Inouye, M., Jia, X., Palotie, A., Parkin, M., Whittaker, P., Yu, F., Chang, K., Hawes, A., Lewis, L. R., Ren, Y., Wheeler, D., Gibbs, R. A., Muzny, D. M., Barnes, C., Darvishi, K., Hurles, M., Korn, J. M., Kristiansson, K., Lee, C., McCarrol, S. A., Nemesh, J., Dermitzakis, E., Keinan, A., Montgomery, S. B., Pollack, S., Price, A. L., Soranzo, N., Bonnen, P. E., Gibbs, R. A., Gonzaga-Jauregui, C., Keinan, A., Price, A. L., Yu, F., Anttila, V., Brodeur, W., Daly, M. J., Leslie, S., McVean, G., Moutsianas, L., Nguyen, H., Schaffner, S. F., Zhang, Q., Ghori, M. J. R., McGinnis, R., McLaren, W., Pollack, S., Price, A. L., Schaffner, S. F., Takeuchi, F., Grossman, S. R., Shlyakhter, I., Hostetter, E. B., Sabeti, P. C., Adebamowo, C. A., Foster, M. W., Gordon, D. R., Licinio, J., Manca, M. C., Marshall, P. A., Matsuda, I., Ngare, D., Wang, V. O., Reddy, D., Rotimi, C. N., Royal, C. D., Sharp, R. R., Zeng, C., Brooks, L. D., and McEwen, J. E. (2010a). Integrating common and rare genetic variation in diverse human populations. *Nature*, 467(7311):52–58.

Consortium, I. H. . et al. (2010b). Integrating common and rare genetic variation in diverse human populations. *Nature*, 467(7311):52.

Donatti, A., Canto, A. M., Godoi, A. B., da Rosa, D. C., and Lopes-Cendes, I. (2020). Circulating metabolites as potential biomarkers for neurological disorders—metabolites in neurological disorders. *Metabolites*, 10(10):389.

Enns, G. M., Martinez, D. R., Kuzmin, A. I., Koch, R., Wakeem, C. K., Woo, S. L., Eisensmith, R. C., and Packman, S. (1999). Molecular correlations in phenylketonuria: mutation patterns and corresponding biochemical and clinical phenotypes in a heterogeneous california population. *Pediatric research*, 46(5):594–594.

Falola, O., Adam, Y., Ajayi, O., Kumuthini, J., Adewale, S., Mosaku, A., Samtal, C., Adebayo, G., Emmanuel, J., Tchamga, M. S., et al. (2023). Sysbiolpgwas: simplifying post-gwas analysis through the use of computational technologies and integration of diverse omics datasets. *Bioinformatics*, 39(1):btac791.

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

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

Geoffroy, V., Herenger, Y., Kress, A., Stoetzel, C., Piton, A., Dollfus, H., and Muller, J. (2018). Annotsv: an integrated tool for structural variations annotation. *Bioinformatics*, 34(20):3572–3574.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative adversarial nets. *Advances in neural information processing systems*, 27.

Gottesman, O., Kuivaniemi, H., Tromp, G., Faucett, W. A., Li, R., Manolio, T. A., Sanderson, S. C., Kannry, J., Zinberg, R., Basford, M. A., Brilliant, M., Carey, D. J., Chisholm, R. L., Chute, C. G., Connolly, J. J., Crosslin, D., Denny, J. C., Gallego, C. J., Haines, J. L., Hakonarson, H., Harley, J., Jarvik, G. P., Kohane, I., Kullo, I. J., Larson, E. B., McCarty, C., Ritchie, M. D., Roden, D. M., Smith, M. E., Böttinger, E. P., Williams, M. S., and Network, e. (2013). The electronic medical records and genomics (eMERGE) network: past, present, and future. *Genetics in Medicine*, 15(10):761–771.

Grapes, L., Dekkers, J., Rothschild, M., and Fernando, R. (2004). Comparing linkage disequilibrium-based methods for fine mapping quantitative trait loci. *Genetics*, 166(3):1561–1570.

Group, N. H. B. C., Barrett, J. C., Esko, T., Fischer, K., Jostins-Dean, L., Jousilahti, P., Julkunen, H., Jaaskelainen, T., Kerimov, N., Kerminen, S., et al. (2023). Metabolomic and genomic prediction of common diseases in 477,706 participants in three national biobanks. *medRxiv*, pages 2023–06.

Group, T. W. H. I. S. (1998). Design of the women's health initiative clinical trial and observational study. *Controlled clinical trials*, 19(1):61–109.

Guasch-Ferré, M., Hu, F. B., Ruiz-Canela, M., Bullo, M., Toledo, E., Wang, D. D., Corella, D., Gómez-Gracia, E., Fiol, M., Estruch, R., et al. (2017). Plasma metabolites from choline pathway and risk of cardiovascular disease in the predimed (prevention with mediterranean diet) study. *Journal of the American Heart Association*, 6(11):e006524.

Gusella, J. F., Lee, J.-M., and MacDonald, M. E. (2021). Huntington's disease: nearly four decades of human molecular genetics. *Human Molecular Genetics*, 30(R2):R254–R263.

Hagenbeek, F. A., Roetman, P. J., Pool, R., Kluft, C., Harms, A. C., Van Dongen, J., Colins, O. F., Talens, S., van Beijsterveldt, C. E., Vandenbosch, M. M., et al. (2020). Urinary amine and organic acid metabolites evaluated as markers for childhood aggression: the action biomarker study. *Frontiers in psychiatry*, 11:165.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Hicks, P. and Oluwadare, O. (2022). Hicarn: resolution enhancement of hi-c data using cascading residual networks. *Bioinformatics*, 38(9):2414–2421.

Highsmith, M. and Cheng, J. (2021). Vehicle: a variationally encoded hi-c loss enhancement algorithm for improving and generating hi-c data. *Scientific Reports*, 11(1):8880.

Hong, H., Jiang, S., Li, H., Du, G., Sun, Y., Tao, H., Quan, C., Zhao, C., Li, R., Li, W., et al. (2020). Deephic: A generative adversarial network for enhancing hi-c data resolution. *PLoS computational biology*, 16(2):e1007287.

Hu, M., Deng, K., Selvaraj, S., Qin, Z., Ren, B., and Liu, J. S. (2012). HiCNorm: removing biases in hi-c data via poisson regression. *Bioinformatics*, 28(23):3131–3133.

Hu, Y. and Ma, W. (2021a). Enhic: learning fine-resolution hi-c contact maps using a generative adversarial framework. *Bioinformatics*, 37(Supplement_1):i272–i279.

Hu, Y. and Ma, W. (2021b). EnHiC: learning fine-resolution hi-c contact maps using a generative adversarial framework. *Bioinformatics*, 37(Suppl_1):i272–i279.

Huang, L., Rosen, J. D., Sun, Q., Chen, J., Wheeler, M. M., Zhou, Y., Min, Y.-I., Kooperberg, C., Conomos, M. P., Stilp, A. M., et al. (2022). Top-ld: A tool to explore linkage disequilibrium with topmed whole-genome sequence data. *The American Journal of Human Genetics*, 109(6):1175–1181.

Jakkula, E., Rehnström, K., Varilo, T., Pietiläinen, O. P., Paunio, T., Pedersen, N. L., deFaire, U., Järvelin, M.-R., Saharinen, J., Freimer, N., et al. (2008). The genome-wide patterns of variation expose significant substructure in a founder population. *The American Journal of Human Genetics*, 83(6):787–794.

Jeffares, D. C., Jolly, C., Hoti, M., Speed, D., Shaw, L., Rallis, C., Balloux, F., Dessimoz, C., Bähler, J., and Sedlazeck, F. J. (2017). Transient structural variations have strong effects on quantitative traits and reproductive isolation in fission yeast. *Nature Communications*, 8:14061.

Jun, G., Sedlazeck, F., Zhu, Q., English, A., Metcalf, G., Kang, H. M., (HGSVC), H. G. S. V. C., Lee, C., Gibbs, R., and Boerwinkle, E. (2021). muCNV: Genotyping structural variants for population-level sequencing. *Bioinformatics*, 37(14):2055–2057.

Juric, I., Yu, M., Abnousi, A., Raviram, R., Fang, R., Zhao, Y., Zhang, Y., Qiu, Y., Yang, Y., Li, Y., et al. (2019). Maps: Model-based analysis of long-range chromatin interactions from plac-seq and hichip experiments. *PLoS computational biology*, 15(4):e1006982.

Kang, H. M., Sul, J. H., Service, S. K., Zaitlen, N. A., Kong, S.-y., Freimer, N. B., Sabatti, C., and Eskin, E. (2010). Variance component model to account for sample structure in genome-wide association studies. *Nature genetics*, 42(4):348–354.

Kiddle, S., Steves, C., Mehta, M., Simmons, A., Xu, X., Newhouse, S., Sattlecker, M., Ashton, N., Bazenet, C., Killick, R., et al. (2015). Plasma protein biomarkers of alzheimer's disease endophenotypes in asymptomatic older twins: early cognitive decline and regional brain volumes. *Translational psychiatry*, 5(6):e584–e584.

Kim, K., Trott, J. F., Gao, G., Chapman, A., and Weiss, R. H. (2019). Plasma metabolites and lipids associate with kidney function and kidney volume in hypertensive adpkd patients early in the disease course. *BMC nephrology*, 20:1–12.

Kingma, D. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kingma, D. P. and Welling, M. (2013). Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *nature*, 521(7553):436–444.

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature*, 521(7553):436–444.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., and Subgroup, . G. P. D. P. (2009). The sequence alignment/map format and SAMtools. *Bioinformatics*, 25(16):2078–2079.

Li, W., Yang, F., Wang, F., Rong, Y., Wu, B., Zhang, H., and Yao, J. (2022). A versatile deep graph contrastive learning framework for single-cell proteomics embedding. *bioRxiv*, pages 2022–12.

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

Lieberman-Aiden, E., Van Berkum, N. L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., Amit, I., Lajoie, B. R., Sabo, P. J., Dorschner, M. O., et al. (2009). Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *science*, 326(5950):289–293.

Liu, T. and Wang, Z. (2019a). Hicnn: a very deep convolutional neural network to better enhance the resolution of hi-c data. *Bioinformatics*, 35(21):4222–4228.

Liu, T. and Wang, Z. (2019b). Hicnn2: enhancing the resolution of hi-c data using an ensemble of convolutional neural networks. *Genes*, 10(11):862.

Long, T., Hicks, M., Yu, H.-C., Biggs, W. H., Kirkness, E. F., Menni, C., Zierer, J., Small, K. S., Mangino, M., Messier, H., et al. (2017). Whole-genome sequencing identifies common-to-rare variants associated with human blood metabolites. *Nature genetics*, 49(4):568–578.

MacArthur, J., Bowler, E., Cerezo, M., Gil, L., Hall, P., Hastings, E., Junkins, H., McMahon, A., Milano, A., Morales, J., et al. (2017). The new nhgri-ebi catalog of published genome-wide association studies (gwas catalog). *Nucleic acids research*, 45(D1):D896–D901.

Machiela, M. J. and Chanock, S. J. (2015). Ldlink: a web-based application for exploring population-specific haplotype structure and linking correlated alleles of possible functional variants. *Bioinformatics*, 31(21):3555–3557.

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. *The American Journal of Human Genetics*, 93(2):278–288.

Martin, A. R., Kanai, M., Kamatani, Y., Okada, Y., Neale, B. M., and Daly, M. J. (2019). Clinical use of current polygenic risk scores may exacerbate health disparities. *Nature genetics*, 51(4):584–591.

Mbatchou, J., Barnard, L., Backman, J., Marcketta, A., Kosmicki, J. A., Ziyatdinov, A., Benner, C., O'Dushlaine, C., Barber, M., Boutkov, B., et al. (2021). Computationally efficient whole-genome regression for quantitative and binary traits. *Nature genetics*, 53(7):1097–1103.

Mikhaylova, A. V., McHugh, C. P., Polfus, L. M., Raffield, L. M., Boorgula, M. P., Blackwell, T. W., Brody, J. A., Broome, J., Chami, N., Chen, M.-H., et al. (2021). Whole-genome sequencing in diverse subjects identifies genetic correlates of leukocyte traits: The nhlbi topmed program. *The American Journal of Human Genetics*, 108(10):1836–1851.

Moore, J. E., Purcaro, M. J., Pratt, H. E., Epstein, C. B., Shoresh, N., Adrian, J., Kawli, T., Davis, C. A., Dobin, A., et al. (2020). Expanded encyclopaedias of dna elements in the human and mouse genomes. *Nature*, 583(7818):699–710.

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.

Noga, M. J., Dane, A., Shi, S., Attali, A., van Aken, H., Suidgeest, E., Tuinstra, T., Muilwijk, B., Coulier, L., Luider, T., et al. (2012). Metabolomics of cerebrospinal fluid reveals changes in the central nervous system metabolism in a rat model of multiple sclerosis. *Metabolomics*, 8:253–263.

Panyard, D. J., Kim, K. M., Darst, B. F., Deming, Y. K., Zhong, X., Wu, Y., Kang, H., Carlsson, C. M., Johnson, S. C., Asthana, S., et al. (2021). Cerebrospinal fluid metabolomics identifies 19 brain-related phenotype associations. *Communications biology*, 4(1):63.

Parrish, R. L., Gibson, G. C., Epstein, M. P., and Yang, J. (2022). Tigar-v2: Efficient twas tool with nonparametric bayesian eqtl weights of 49 tissue types from gtex v8. *Human Genetics and Genomics Advances*, 3(1).

Raffield, L. M., Iyengar, A. K., Wang, B., Gaynor, S. M., Spracklen, C. N., Zhong, X., Kowalski, M. H., Salimi, S., Polfus, L. M., Benjamin, E. J., et al. (2020). Allelic heterogeneity at the crp locus identified by whole-genome sequencing in multi-ancestry cohorts. *The American Journal of Human Genetics*, 106(1):112–120.

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

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

Reay, W. R. and Cairns, M. J. (2021). Advancing the use of genome-wide association studies for drug repurposing. *Nature Reviews Genetics*, 22(10):658–671.

Rhee, E. P., Yang, Q., Yu, B., Liu, X., Cheng, S., Deik, A., Pierce, K. A., Bullock, K., Ho, J. E., Levy, D., et al. (2016). An exome array study of the plasma metabolome. *Nature communications*, 7(1):12360.

Rhie, A., Nurk, S., Cechova, M., Hoyt, S. J., Taylor, D. J., Altemose, N., Hook, P. W., Koren, S., Rautiainen, M., Alexandrov, I. A., et al. (2023). The complete sequence of a human y chromosome. *Nature*, 621(7978):344–354.

Richardson, T. G., Leyden, G. M., Wang, Q., Bell, J. A., Elsworth, B., Davey Smith, G., and Holmes, M. V. (2022). Characterising metabolomic signatures of lipid-modifying therapies through drug target mendelian randomisation. *PLoS biology*, 20(2):e3001547.

Rios, S., García-Gavilán, J. F., Babio, N., Paz-Graniel, I., Ruiz-Canela, M., Liang, L., Clish, C. B., Toledo, E., Corella, D., Estruch, R., et al. (2023). Plasma metabolite profiles associated with the world cancer research fund/american institute for cancer research lifestyle score and future risk of cardiovascular disease and type 2 diabetes. *Cardiovascular diabetology*, 22(1):252.

Romick-Rosendale, L. E., Brunner, H. I., Bennett, M. R., Mina, R., Nelson, S., Petri, M., Kiani, A., Devarajan, P., and Kennedy, M. A. (2011). Identification of urinary metabolites that distinguish membranous lupus nephritis from proliferative lupus nephritis and focal segmental glomerulosclerosis. *Arthritis research & therapy*, 13:1–10.

Rosen, J. D., Yang, Y., Abnousi, A., Chen, J., Song, M., Jones, I. R., Shen, Y., Hu, M., and Li, Y. (2021). Hprep: quantifying reproducibility in hichip and plac-seq datasets. *Current Issues in Molecular Biology*, 43(2):1156–1170.

Rowland, B., Venkatesh, S., Tardaguila, M., Wen, J., Rosen, J. D., Tapia, A. L., Sun, Q., Graff, M., Vuckovic, D., Lettre, G., et al. (2022). Transcriptome-wide association study in uk biobank europeans identifies associations with blood cell traits. *Human Molecular Genetics*, 31(14):2333–2347.

Ruiz-Canela, M., Hruby, A., Clish, C. B., Liang, L., Martínez-González, M. A., and Hu, F. B. (2017). Comprehensive metabolomic profiling and incident cardiovascular disease: a systematic review. *Journal of the American heart association*, 6(10):e005705.

Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., and Chen, X. (2016a). Improved techniques for training gans. *Advances in neural information processing systems*, 29.

Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., and Chen, X. (2016b). Improved techniques for training gans. *Advances in neural information processing systems*, 29.

Schoenfelder, S., Furlan-Magaril, M., Mifsud, B., Tavares-Cadete, F., Sugar, R., Javierre, B.-M., Nagano, T., Katsman, Y., Sakthidevi, M., Wingett, S. W., et al. (2015). The pluripotent regulatory circuitry connecting promoters to their long-range interacting elements. *Genome research*, 25(4):582–597.

Scriver, C. R. and Waters, P. J. (1999). Monogenic traits are not simple: lessons from phenylketonuria. *Trends in genetics*, 15(7):267–272.

Shah, S. H., Sun, J.-L., Stevens, R. D., Bain, J. R., Muehlbauer, M. J., Pieper, K. S., Haynes, C., Hauser, E. R., Kraus, W. E., Granger, C. B., et al. (2012). Baseline metabolomic profiles predict cardiovascular events in patients at risk for coronary artery disease. *American heart journal*, 163(5):844–850.

Sharma, N. and Cutting, G. (2020). The genetics and genomics of cystic fibrosis. *Journal of Cystic Fibrosis*, 19:S5–S9.

Shin, D., Gilbert, F., Goldstein, M., and Schlegel, P. N. (1997). Congenital absence of the vas deferens: incomplete penetrance of cystic fibrosis gene mutations. *The Journal of urology*, 158(5):1794–1799.

Shin, S.-Y., Fauman, E. B., Petersen, A.-K., Krumsiek, J., Santos, R., Huang, J., Arnold, M., Erte, I., Forgetta, V., Yang, T.-P., et al. (2014). An atlas of genetic influences on human blood metabolites. *Nature genetics*, 46(6):543–550.

Slatkin, M. (2008). Linkage disequilibrium—understanding the evolutionary past and mapping the medical future. *Nature Reviews Genetics*, 9(6):477–485.

Song, M., Pebworth, M.-P., Yang, X., Abnousi, A., Fan, C., Wen, J., Rosen, J. D., Choudhary, M. N. K., Cui, X., Jones, I. R., 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.

Srivastava, A., Valkov, L., Russell, C., Gutmann, M. U., and Sutton, C. (2017). Veegan: Reducing mode collapse in gans using implicit variational learning. *Advances in neural information processing systems*, 30.

Sudlow, C., Gallacher, J., Allen, N., Beral, V., Burton, P., Danesh, J., Downey, P., Elliott, P., Green, J., Landray, M., et al. (2015). Uk biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS medicine*, 12(3):e1001779.

Sun, Q., Graff, M., Rowland, B., Wen, J., Huang, L., Miller-Fleming, T. W., Haessler, J., Preuss, M. H., Chai, J.-F., Lee, M. P., et al. (2022a). Analyses of biomarker traits in diverse uk biobank participants identify associations missed by european-centric analysis strategies. *Journal of human genetics*, 67(2):87–93.

Sun, Q., Liu, W., Rosen, J. D., Huang, L., Pace, R. G., Dang, H., Gallins, P. J., Blue, E. E., Ling, H., Corvol, H., et al. (2022b). Leveraging topmed imputation server and constructing a cohort-specific imputation reference panel to enhance genotype imputation among cystic fibrosis patients. *Human Genetics and Genomics Advances*, 3(2).

Székely, G. J. and Rizzo, M. L. (2022). Brownian distance covariance. In *The energy of data and distance correlation*, pages 249–260. Chapman and Hall/CRC, Boca Raton.

Taliun, D., Harris, D. N., Kessler, M. D., Carlson, J., Szpiech, Z. A., Torres, R., Taliun, S. A. G., Corvelo, A., Gogarten, S. M., Kang, H. M., et al. (2021). Sequencing of 53,831 diverse genomes from the nhlbi topmed program. *Nature*, 590(7845):290–299.

Taylor, H. A., Wilson, J. G., Jones, D. W., Sarpong, D. F., Srinivasan, A., Garrison, R. J., Nelson, C., and Wyatt, S. B. (2005). Toward resolution of cardiovascular health disparities in african americans: design and methods of the jackson heart study. *Ethnicity & Disease*, 15(4 Suppl 6):S6–4.

Thysell, E., Surowiec, I., Hörnberg, E., Crnalic, S., Widmark, A., Johansson, A. I., Stattin, P., Bergh, A., Moritz, T., Antti, H., et al. (2010). Metabolomic characterization of human prostate cancer bone metastases reveals increased levels of cholesterol. *PloS one*, 5(12):e14175.

Uffelmann, E., Huang, Q. Q., Munung, N. S., De Vries, J., Okada, Y., Martin, A. R., Martin, H. C., Lappalainen, T., and Posthuma, D. (2021). Genome-wide association studies. *Nature Reviews Methods Primers*, 1(1):59.

Wang, F., Baden, M. Y., Guasch-Ferré, M., Wittenbecher, C., Li, J., Li, Y., Wan, Y., Bhupathiraju, S. N., Tobias, D. K., Clish, C. B., et al. (2022). Plasma metabolite profiles related to plant-based diets and the risk of type 2 diabetes. *Diabetologia*, 65(7):1119–1132.

Ward, L. D. and Kellis, M. (2012). Haploreg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. *Nucleic acids research*, 40(D1):D930–D934.

Ward, L. D. and Kellis, M. (2016). Haploreg v4: systematic mining of putative causal variants, cell types, regulators and target genes for human complex traits and disease. *Nucleic acids research*, 44(D1):D877–D881.

Wheeler, M. M., Stilp, A. M., Rao, S., Halldórsson, B. V., Beyter, D., Wen, J., Mihkaylova, A. V., McHugh, C. P., Lane, J., Jiang, M.-Z., et al. (2022). Whole genome sequencing identifies structural variants contributing to hematologic traits in the nhlbi topmed program. *Nature communications*, 13(1):7592.

Wilson, J. G., Rotimi, C. N., Ekunwe, L., Royal, C. D. M., Crump, M. E., Wyatt, S. B., Steffes, M. W., Adeyemo, A., Zhou, J., Taylor, H. A., and Jaquish, C. (2005). Study design for genetic analysis in the jackson heart study. *Ethnicity & Disease*, 15(4 Suppl 6):S6–30.

Wright, C. F. et al. (2022). Incomplete penetrance and variable expressivity: from clinical studies to population cohorts. *Frontiers in Genetics*, 13:920390.

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

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

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

Yin, X., Chan, L. S., Bose, D., Jackson, A. U., VandeHaar, P., Locke, A. E., Fuchsberger, C., Stringham, H. M., Welch, R., Yu, K., et al. (2022). Genome-wide association studies of metabolites in finnish men identify disease-relevant loci. *Nature communications*, 13(1):1644.

Zarate, S., Carroll, A., Mahmoud, M., Krasheninina, O., Jun, G., Salerno, W. J., Schatz, M. C., Boerwinkle, E., Gibbs, R. A., and Sedlazeck, F. J. (2020). Parliament2: Accurate structural variant calling at scale. *GigaScience*, 9(12).

Zhang, Y., An, L., Xu, J., Zhang, B., Zheng, W. J., Hu, M., Tang, J., and Yue, F. (2018). Enhancing hi-c data resolution with deep convolutional neural network hicplus. *Nature communications*, 9(1):750.

Zhou, X., Carbonetto, P., and Stephens, M. (2013a). Polygenic modeling with bayesian sparse linear mixed models. *PLoS genetics*, 9(2):e1003264.

Zhou, X., Lowdon, R. F., Li, D., Lawson, H. A., Madden, P. A. F., Costello, J. F., and Wang, T. (2013b). Exploring long-range genome interactions using the WashU epigenome browser. *Nature Methods*, 10(5):375–376.