UNIVERSITY OF CALIFORNIA SAN DIEGO

Building Multiomics Analysis Tools For a Holistic Understanding of Biological Systems

A dissertation submitted in partial satisfaction of the
requirements for the degree Doctor of Philosophy

in

Bioinformatics & Systems Biology

by

Joaquin Reyna

Committee in charge:

Professor Ferhat Ay, Chair
Professor Lukas Werner Chavez Kuss, Co-Chair
Professor Kyle J. Gaulton
Professor Anjana Rao
Professor Sheng Zhong

2024

The Dissertation of Joaquin Reyna is approved, and it is acceptable in quality and form for publication on microfilm and electronically.

University of California San Diego

2024

DEDICATION

This dedication is written in Spanish, Italian and English to make sure my family and friends can easily understand their importance to me and this dissertation.

En Espanol:

Dedico esta tesis a mi esposa Sara. La vida nos ofrece muchas oportunidades si estamos listos. Nuestro cuento comenzó con una carta y terminó en una vida. Muchas gracias por todo el amor que me das y por los momentos más bonitos de mi vida y nuestra vida juntos. Hemos vivido tantas cosas bellas que nunca olvidaré. Qué bonito fue conocer Big Sur, el Gran Cañón, Inyo, Utah, París, Cinque Terre, Trentino y muchos otros lugares contigo. Me gusta tanto tu risa y la manera en que te inspira la naturaleza. Ves el mundo y sus colores como nadie más, ves lo que está más allá, y me da orgullo verte convertida en una artista. Sé que tendremos muchas más aventuras, y juntos podremos superar cualquier obstáculo. Grazie, mia vita, come prima, più di prima ti amerò.

También me gustaría dar mil gracias a mis papás. Gracias papá, como siempre dices, la vida está llena de capítulos, y este ha terminado. Ahora es tiempo de abrir y vivir otros nuevos capítulos. Con tus enseñanzas pude abrir mis ojos al mundo. Gracias por llevarme a trabajar contigo. Tal vez me quejaba porque me perdía un episodio de Pokémon, pero valió la pena. Te vi, te escuché, y siempre tendré tus enseñanzas conmigo. Mamá, muchas gracias. Tal vez parece que solo ayer era un chiquilín, y ahora soy un adulto. Gracias por todas las veces que me tomaste de la mano y me enseñaste por dónde ir. A veces me perdía, pero siempre estabas esperándome con un gran abrazo. Los quiero mucho, y gracias por ser mis papás.

Quiero también dar gracias a toda mi familia. Gracias, hermanos y hermanas: Celeste, Porfirio, Raiza, Celes y Laiza y mi cuñado Juan. En especial, quiero dar gracias a mi abuela Amparo, mis padrinos Pablo y Enrique, y mi madrina Marty. También doy estas gracias a mis tíos Celestino, Claudia, Ana, primos y absolutamente todos mis parientes. Gracias a mi familia italiana: Mario, Bruna, Paolo, Emanuele, Paolo y Francesca, Giacoma, y María, Patrizio y Christina, Barbara e Iván, Corrado y Morena, Anna, y muchos más.

¡Por último, gracias a todos mis amigos y compañeros! Todos sus apoyos, pequeños o grandes, me ayudan a realizar mi sueños.

In Italiano:

Dedico questa tesi a mia moglie Sara. La vita ci offre molte opportunità se siamo pronti. La nostra storia è iniziata con una lettera e si è trasformata in una vita. Grazie mille per tutto l'amore che mi dai e per i momenti più belli della mia vita e della nostra vita insieme. Abbiamo vissuto tante cose meravigliose che non dimenticherò mai. Che bello è stato scoprire Big Sur, il Grand Canyon, Inyo, Utah, Parigi, le Cinque Terre, il Trentino e tanti altri posti con te. Amo il tuo sorriso e il modo in cui la natura ti ispira. Vedi il mondo e i suoi colori come nessun altro, vedi oltre, e sono orgoglioso di vederti diventata un'artista. So che avremo molte altre avventure e insieme potremo superare qualsiasi ostacolo. Grazie mia vita, come prima, più di prima, ti amerò.

Vorrei anche ringraziare mille volte i miei genitori. Grazie papà, come dici sempre, la vita è piena di capitoli, e questo si è concluso. Adesso è ora di aprire e vivere nuovi capitoli. Con i tuoi insegnamenti ho potuto aprire gli occhi sul mondo. Grazie per avermi portato a lavorare con te. Forse mi lamentavo perché mi perdevo un episodio dei Pokémon, ma ne è valsa la pena. Ti ho visto, ti ho ascoltato e porterò sempre con me le tue lezioni. Mamma, grazie mille. Forse sembra che solo ieri fossi un bambino e ora sono un adulto. Grazie per tutte le volte che mi hai preso per mano e mi hai mostrato la strada. A volte mi perdevo, ma eri sempre lì ad aspettarmi con un abbraccio. Vi voglio tanto bene, e grazie per essere i miei genitori.

Voglio anche ringraziare tutta la mia famiglia. Grazie ai miei fratelli e sorelle: Celeste, Porfirio, Raiza, Celes, Laiza e mio cognato Juan. Grazie a mia nonna Amparo e i miei padrini Pablo ed Enrique e la mia madrina Marty. Ringrazio anche mio zio Celestino e le mie zie Claudia e Ana, e tutti gli altri famigliari. Grazie alla mia famiglia italiana: Mario, Bruna, Emanuele, Paolo e Francesca, Giacoma e Maria, Patrizio e Cristina, Barbara e Ivan, Corrado e Morena, Anna, e tanti altri.

Infine, grazie a tutti i miei amici e colleghi! Ogni vostri supporto, piccolo o grande, mi aiutanno a realizzare i miei sogni.

In English:

I dedicate this thesis to my wife Sara. Life offers us many opportunities if we are ready. Our story began with a letter and turned into a life together. Thank you so much for all the love you give me and for the most beautiful moments of my and our lives. We have experienced so many beautiful things that I will never forget. How amazing it was to discover Big Sur, the Grand Canyon, Inyo, Utah, Paris, Cinque Terre, Trentino, and many other places with you. I love your smile and the way nature inspires you. You see the world and its colors like no one else, you see beyond the expected, and it makes me proud to see you become an artist. I know we will have many more adventures, and together we will be able to overcome any obstacle. Grazie, mia vita, come prima, più di prima ti amerò.

I would also like to thank my parents a million times over. Thank you Dad, as you always say, life is full of chapters, and this one has come to an end. Now is the time to open and live out completely new chapters. Through your teachings, I was able to open my eyes to the world. Thank you for taking me to work with you. Maybe I complained because I missed a Pokémon episode, but it was worth it. I spent time with you, I listened to you, and I will always carry your lessons with me. Mom, thank you so much. Maybe it feels like only yesterday that I was your little kid, and now I'm an adult. Thank you for all the times you took me by the hand and showed me where to go. Sometimes I got lost, but you were always waiting for me with a big hug. I love you both so much, and thank you for being my parents.

I also want to thank my entire family. Thank you, siblings: Celeste, Porfirio, Raiza, Celes, and Laiza and my brother-in-law Juan. I want to give a special thanks to my grandma Amparo and my godparents Pablo, Enrique and Marty. I also extend this thanks to my uncles Celestino, Claudia, Ana and absolutely all of my relatives. Sincere thanks to my Italian family: Mario, Bruna, Paolo, Emanuele, Paolo and Francesca, Giacoma, and María, Patrizio and Christina, Barbara and Iván, Corrado and Morena, Anna, and many more.

Lastly, thank you to all my friends and colleagues! Every bit of support, big or small, helps me realize my dreams.

EPIGRAPH

It is like a voyage of discovery into unknown lands, seeking not for new territory but for new knowledge. It should appeal to those with a good sense of adventure.

Federick Sanger

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF SUPPLEMENTAL TABLES

LIST OF ABBREVIATIONS

4DN 4D Nucleome

ChIP-seq Chromatin Immunoprecipitation Followed by Sequencing

ChIA-PET Chromatin Interaction Analysis with Paired-End Tag

CTCF CCCTC-Binding Factor

dbGaP Database of Genotypes and Phenotypes

EBI European Bioinformatics Institute

eQTL Expression Quantitative Trait Loci

ENCODE ENCyclopedia Of DNA Elements

FDR False Discovery Rate

FC Loops FitHiChIP with ChIP-seq Peaks

FH Loops FitHiChIP with HiChIP Peaks

GEO Gene Expression Omnibus

GWAS Genome-wide Association Study

HCRegLoops High Confidence Regulatory Loops

Hi-C High-Throughput Chromosome Conformation Capture

HiCCUPS Hi-C Computational Unbiased Peak Search

HiChIP High-Throughput Chromosome Conformation Capture with ChIP

KR Knight and Ruiz (normalization)

MAPQ Mapping Quality (score)

QC Quality Control

SCC Stratum-Adjusted Correlation Coefficient

VC Vanilla Coverage (normalization)

AD Atopic Dermatitis

FIMO (MEME Suite) Find Individual Motif Occurrences

LD Linkage Disequilibrium

NK Natural Killer Cells

RA Rheumatoid Arthritis

SEA (MEME Suite) Simple Enrichment Analysis

SGL SNP-to-Gene Linking

SNP Single Nucleotide Polymorphism

SRA Sequence Read Archive

PS Psoriasis

T1D Type 1 Diabetes

TF Transcription Factor

TSS Transcription Start Site

ZNF Zinc Finger Protein

jDR joint Dimensionality Reduction

MCIA Multiple Co-Inertia Analysis

PCA Principal Components Analysis

NIPALS Non-linear Iterative Partial Least Squares

NMF Non-negative Matrix Factorization

SVD Singular Value Decomposition

ACKNOWLEDGEMENTS

I am grateful to my mentor, Dr. Ferhat Ay, and the entire lab for their unwavering support and guidance. It is through the shoulders of giants that we are able to climb the next height. I also extend my sincere thanks to Dr. Anna Konstorum and the nipalsMCIA team for their collaborative spirit. Lastly, I wish to expression my appreciation to the Bjoern Peters lab, whose creation of the CMI-PB multiomics data provided me with an invaluable opportunity to learn about methodologies such as joint dimensionality reduction and machine learning.

Chapters 1 and 2, in full, has been submitted for publication of the material as it may appear in Genome Biology 2024, Joaquin Reyna, Kyra Fetter, Romeo Ignacio, Cemil Can Ali Marandi, Nikhil Rao, Zichen Jiang, Daniela Salgado Figueroa, Sourya Bhattacharyya, Ferhat Ay. The dissertation author was one of the primary investigators and authors of this paper.

Chapter 3, in full, has been submitted for publication of the material as it may appear in Bioinformatics 2024, Max Mattessich, Joaquin Reyna, Edel Aron, Ferhat Ay, Misha Kilmer, Steven H. Kleinstein, Anna Konstorum. The dissertation author was one of the primary investigators and authors of this paper.

VITA

2011–2015	Bachelor of Science, Bioengineering: Bioinformatics University of California San Diego
2015–2018	Postbaccalaureate, Laboratory of Dr. Kelly Frazer University of California San Diego
2018–2024	Doctor of Philosophy, Bioinformatics & Systems Biology University of California San Diego

PUBLICATIONS

First Author
"Loop Catalog: a comprehensive HiChIP database of human and mouse samples." *In Review: Genome Biology*. 2024

Other Authorship
"nipalsMCIA: Flexible Multi-Block Dimensionality Reduction in R via Non-linear Iterative Partial Least Squares" *In Review: Bioinformatics*. 2024
"A multi-omics systems vaccinology resource to develop and test computational models of immunity" *Cell Reports*. 2024
"Rewiring of the 3D genome during acquisition of carboplatin resistance in a triple-negative breast cancer patient-derived xenograft" *Scientific Reports*. 2023

ABSTRACT OF THE DISSERTATION

Building Multiomics Analysis Tools For a Holistic Understanding of Biological Systems

by

Joaquin Reyna

Doctor of Philosophy in Bioinformatics & Systems Biology

University of California San Diego, 2024

Professor Ferhat Ay, Chair
Professor Lukas Werner Chavez Kuss, Co-Chair

The massive generation of genetic, epigenetic, transcriptomic, and other sources of data, allows us to pursue biological questions at scale while simultaneously adding a systems-level context to hypotheses in biology. Questions about gene expression have driven us to understand various chromatin components, most recently that has lead to the study of chromatin conformation via high-throughput methods such as HiC or HiChIP. To obtain a full understanding of chromatin conformation, integration with genetics variants (e.g. SNPs from GWAS and eQTL studies) and epigenetics signals (e.g. histone acetylation, open chromatin regions, transcription factor binding, etc) is essential. Similarly, complex diseases such as cancer can advance via

a system of distinct factors that interact to form a deliberate and potent pathogenic regulatory network. Thus, it is imperative we build the resources and tools necessary to integrate multiomics signals together.

Here, I present three chapters derived from two major works that demonstrate the importance of data integration for a holistic understanding of biology. First, I present a database of HiChIP data for over 1000 samples (chapter 1) with important applications for the analysis of motifs, GWAS and eQTL studies, and network analysis (chapter 2). Second, I showcase and described the nipalsMCIA R package which reduces datasets for a systems level analysis of multiomics data (chapter 3).

Introduction

0.1 Integrating Diverse Biological Signals for a Holistic Understanding of Disease

Disease susceptibility and progression is driven by a variety of factors including both internal (e.g. genetic variation) and external (e.g. environment, bacteria). In the case of diseases associated with genetics, the root cause may be a single monogenic mutation as in the case of Duchenne muscular dystrophy [1] or multifactorial in the case of Type 1 Diabetes [2]. To better understand multifactorial disease we must generate and integrate diverse biological signals. Studies have steadily worked toward utilizing multiple sources of data, eventually leading researchers to coin the term "multiomics". Despite the enthusiasm, it remains a challenge to systematically and meaningfully integrate multiomic datasets. Depending on the data sources it may be possible to perform enrichment and/or correlation analyses for overlapping signals [3–7]. When applicable more sophisticated techniques can also be utilized such as joint dimensionality reduction [8–19]. Indeed, this has been the trajectory of the genetics field, driven by sequencing technologies that allow the interrogation of genetic variants, gene expression, chromatin marks, open chromatin region, chromatin interactions, and beyond. In the broader biology field, these high-throughput methodologies include technologies such as those measuring serological levels (antibody titers), cell frequency estimations, protein abundance and cytokine levels [20, 21]. The growing maturity of these techniques has also encouraged large-scale efforts and databases such as the Roadmap [22], ENCODE [23], 4DN [24], CMI-PB [25] and similar efforts [16, 17, 20, 26]. As alluded to previously, the utilization of multiomics data in genetic and epigenetic studies has enabled several advancements that will be discussed in the following sections.

0.2 The Expanding Role of Multiomics Datasets in Uncovering Genetic Contributions to Disease

Genome-wide association studies (GWAS) have revealed pinpoint regions of the genome with disease-associated variants that predispose us to disease [27, 28]. However, approximately 90% of these variants are located in non-coding regions making it difficult to understand their true function [29]. This has prompted researchers to develop techniques to study epigenetic signals in search of regulatory elements such as enhancers. ChIP-seq has contributed substantially to this endeavour by extracting regions of the genome that are bound by transcription factors or histones with specialized acetylation or methylation-based modifications [30]. Initial studies associated different histone marks with expression inducing marks such as H3K27ac, H3K9ac, H3K4me3 [31]. Ultimately, H3K27ac has been the preferred target for many researchers given its strong association with enhancer elements [31]. Naturally, several studies applying H3K27ac ChIP-seq have demonstrated the specificity of enhancer elements within different cell types thus meaningful integration with GWAS data could reveal whether candidate SNPs are enriched at these elements and provide evidence as to the cell type or tissue of action [32]. Indeed, methodologies [3–7] and studies testing the enrichment between ChIP-seq and GWAS signals have helped prioritize tissues/cell types [33, 34] and find radical difference between cancer and complex disease-associated variants [35]. Overall, the integration of genetic associations with enhancer activity has proven quite important however the goal of interpreting GWAS variants is far from concluded and we need to keep developing resources and frameworks that leverage other types of epigenetic data [32, 36].

Similarly, the integration of transcriptomics with genomics has been essential to detecting genetic variants with a strong influence on gene expression levels. Expression quantitative trait loci (eQTL) studies have undoubtedly been a powerful framework for identifying SNPs that are highly correlated with expression of their nearby genes [37–39]. Like ChIP-seq, transciptomics has revealed eQTLs specific for different tissues, contexts, and timing [40]. Altogether – GWAS, eQTLs and 1D epigenetic signals have made important advances thereby motivating the creation of large consortium such as the GWAS Catalog [28] and eQTL Catalogue [38] that systematically reprocess and make available data on functional variants. Other databases such as the Open Targets have further built upon omics datasets to provide genetically supported drug targets in the form of several QTL signals, enhancer-promoter correlation analyses and targeted chromatin conformation [41]. The design and implementation of powerful integrating methodologies such as those employed by Open Targets plays an essential role in functional variant characterization and investigation.

0.3 Linking Genetics with Chromatin Architecture to Connect Genes and Regulatory Regions

Chromatin conformation enables the genome to interact across long distances, defying the constraint of the linear genome [42, 43]. This is accomplished through the formation of compartments, topologically associated domains and loops [44–48]. To study this phenomena at a genome-wide scale, the HiC assay was developed which captures interactions between all regions simultaneously, creating comprehensive maps of chromatin conformation [49, 50]. Briefly, HiC utilizes DNA crosslinking, digestion, biotin marking, re-ligation, and biotin pulldown to then build a sequencing library that, when fully processed, generates a matrix of interactions between loci at various resolutions. Studies have associated this organization with biological processes as diverse as chromatin accessibility [50], conservation of loops between human and mice [49], 1D epigenetic signals [46], transcriptional changes [51, 52], cell development and differentiation [53], and disease [54, 55]. To expand upon this methodology, HiChIP has fused HiC together with ChIP-seq to extract protein-centric interaction pairs [56]. This method has gone on contribute to similar findings such as changes in loop formation conditional on the presence of specific transcription factors [57] and applied to variant prioritization studies [39, 58, 59]. The latter has been extensively applied to GWAS signals derived from skin [60], kidney [61], ovarian [62–64], heart [35, 65], and immune-associated [66–70] diseases. The linking nature of chromatin conformation will continue to promote its popularity within omics studies and the effective development of resources and tools will be essential to continuing this trend.

0.4 The Evolution of Multiomics within the Broader Biology Field

Biological events do not occur in isolation, rather, they are the result of complex interactions between environments, cellular programs, and molecular events. The field of systems biology attempts to capture a holistic understanding of biology through integration of diverse yet complementary multiomics datasets. Within this broader field, multiomics stands as a significant topic that includes systems biology subfields such as vaccinology [25, 26, 71–74], pharmacology [75] as well as cellular [76, 77] and mechanistic biology [78]. Realizing this growth, several systems biology researchers have devoted considerable effort to build resources capable of analyzing and hosting multiomics datasets such as HIPC [21], CMI-PB [25], the Human Microbiome Project [79, 80], among others [15, 81]. Alongside the buildup of these datasets, the research community has developed and refined several integrative techniques, including methods based on joint dimensionality reduction, kernels, networks and deep learning [81]. Notably, joint dimensionality reduction (jDR) based methodologies can be further categorized into their underlying methods such as tensor extensions of PCA, canonical correlation analysis, non-negative matrix factorization, tri-factorization, Gaussian latent variable models, principle component analysis, co-inertia analysis and factor analysis [11]. With such a diverse array of analytical tools it is imperative we correctly utilize and interpret the results from multiomics tools like jDR methods. Moreover, the eagerness to conduct studies using multiomics datasets has also motivated their use for computational modeling, such as predicting vaccine responses

[25]. These trends underscore the significance of multiomics data to drive the next generation of biological discoveries and insights into disease.

0.5 Contributions

As previously discussed, the biology field is rapidly evolving towards a landscape abundant with multiomics datasets, where each new assay introduces new analytical and resource challenges. This work highlights my contributions to large-scale multiomics analysis, particularly within the chromatin conformation and broader multiomics space. Specifically, I have developed a comprehensive database of HiChIP data, which I utilized for several applications including the dissection of GWAS and eQTL signals, as well as analyses of DNA motifs and the clustering of significant chromatin interactions. Additionally, I will showcase my contributions to the nipalsMCIA package, where my primary focus was enhancing the interpretability and visualization of jDR results.

Note: Supplementary tables have been uploaded as additional files. They are briefly mentioned at the end of chapters 1 and 2 with a title and short description.

Chapter 1

Developing a comprehensive HiChIP database of human and mouse samples

1.1 Abstract

HiChIP enables cost-effective and high-resolution profiling of regulatory and structural loops. To leverage the increasing number of publicly available HiChIP datasets from diverse cell lines and primary cells, we developed the Loop Catalog (https://loopcatalog.lji.org), a web-based database featuring HiChIP loop calls for 1319 samples across 133 studies and 44 high-resolution Hi-C loop calls. Our comprehensive catalog, spanning over 4M unique 5kb loops constitutes an important resource for studies in gene regulation and genome organization.

1.2 Background

Chromatin folding can impact cell-type-specific function and disease risk via altered 3D interactions between genetic loci [55, 82]. An important feature of chromatin folding is the existence of chromatin loops that connect regions separated by large genomic distances, often up to a megabase but with notable cases spanning even larger distances [42, 83–85]. These loops can be broadly categorized into: i) structural loops demarcating domains of interactions such as topological domains (TADs) and marked by the binding of CTCF and cohesin at their anchors, and ii) regulatory loops which join distal gene regulatory elements such as enhancers and promoters to modulate gene expression [44, 47, 54, 57, 86–92]. Our understanding of the exact role these loops/interactions play in cell-type specific gene regulation and ultimately disease susceptibility is far from complete.

To capture such interactions, the Hi-C procedure was developed as a high-throughput genome-wide assay that carries proximity ligation either in dilution [50] or in-situ within intact nuclei [49]. As a subset of interactions, regulatory loops are often associated with regions enriched for histone modifications such as H3K27ac and H3K4me3, transcription factors or for accessible chromatin [56, 93–97]. The development of the HiChIP assay represents an extension of the in-situ Hi-C methodology that can facilitate capture of these protein-specific regulatory loops by using immunoprecipitation to enrich for interactions involving anchors associated with the binding of transcription factors (TFs) or histone modifications of interest. Given the ability to enrich signal for a subset of the genome (e.g., active regulatory elements), HiChIP enables high-resolution profiling of chromatin interactions of interest with lower sequencing depth compared to Hi-C [56, 94, 98]. Another advantage of HiChIP, at the time of its initial development, was that it allowed working with lower numbers of cells as input (1-5M), hence enabling the characterization of 3D organization in primary cells. Since the introduction of HiChIP in 2016, the number of publicly available human and mouse HiChIP studies published has consistently increased each year (Figure 1.1A). This indicates the growing popularity of the HiChIP assay, particularly as a method of investigating distal interactions between genetic variants, often in non-coding regions of the genome, and their potential target genes in a cell-type and context-specific manner [60–63, 66, 83, 99].

Here, we developed the Loop Catalog, a database containing the largest set of uniformly processed HiChIP data to date, curated from over 130 publications and a total of 1319 samples (from 2016 to January 2024) leading to over 5 million unique looping interactions at 5 kb resolution (10M total across all resolutions with 7.7M for human, 3.1M for mouse) with an accompanying web server that enables visualization, querying and bulk download of looping data (Figure 1.1B, Supp Figure 1.1). Recent publications and databases, namely HiChIPdb

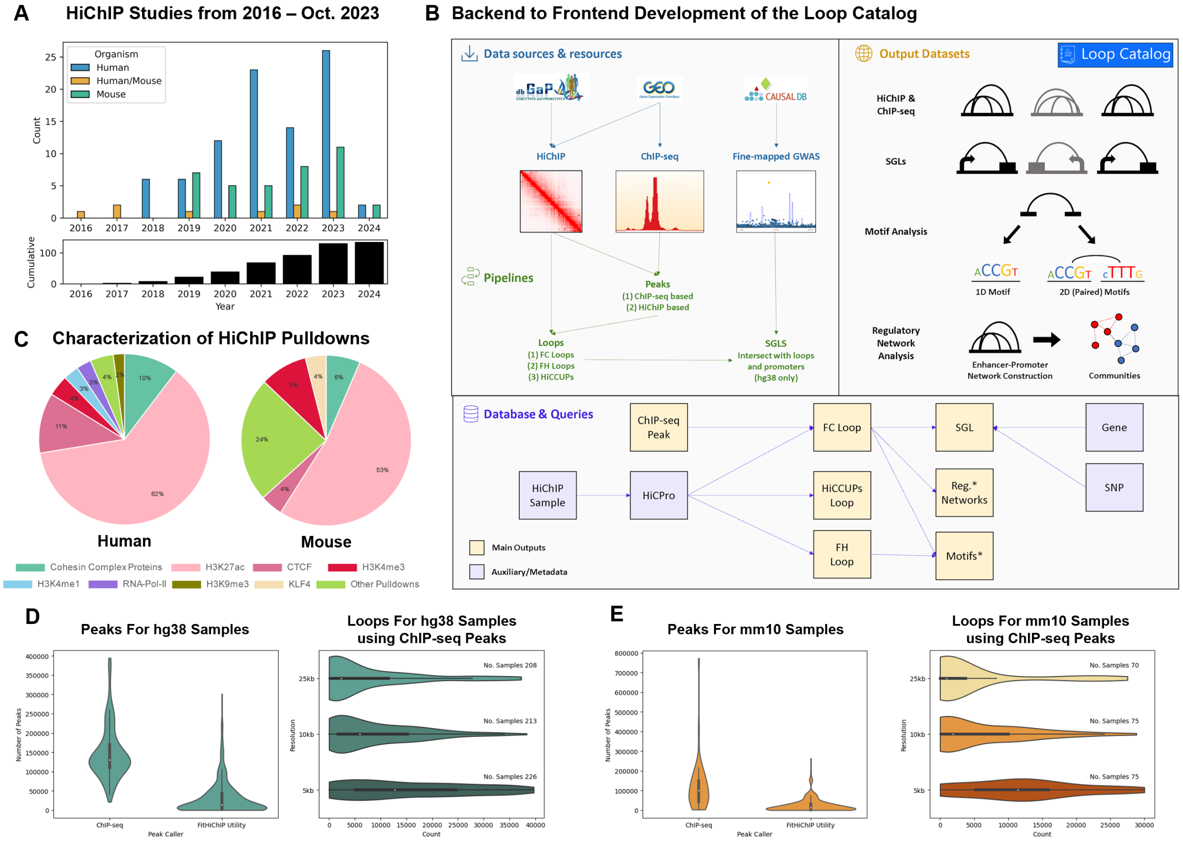


Figure 1.1. Data overview and high-level summary of the Loop Catalog. A) Breakdown of the number of HiChIP samples generated from 2016 to 2024. The top panel shows the number of studies broken down by human (blue), mouse (teal), or both (orange). Bottom panel shows the cumulative trend. **B)** Breakdown of samples by target protein or histone modification and by organism. **C)** Schema for the development of the Loop Catalog starting from raw sequencing files to processing (top left), database storage (bottom) and web accessible analyses (top right). **D-E)** Number of peak calls (left) and FitHiChIP loop calls at three different resolutions (right) for HiChIP samples with ChIP-seq data for **D)** human and **E)** mouse samples.

Table 1.1. Comparison of HiChIP processing methods and website features of Loop Catalog, ChromLoops, HiChIPdb, and Cohesin-DB. Categories include specification and quantification of data types, implementation of loop calling (software, configurations, replicates), ease of data download and visualization, including the ability to select and visualize multiple samples simultaneously, and embedded biological analysis modules.* annotation of genes, SNPs, E/Ps, silencers, circRNA, TWAS, chromosome open access data, alternative splicing, transcription factors ** annotation of genes, SNPs *** annotation of genes

General Feature	**Specific Feature**	**Loop Catalog**	**ChromLoops (Zhou et al., 2022)**	**HiChIPdb (Zeng et al., 2022)**	**Cohesin-DB (Wang et al., 2022)**
Data Types	Organism	Homo sapiens, Mus musculus	Homo sapiens, Mus musculus, 11 others	Homo sapiens	Homo sapiens
	Reference Genome	hg38, mm10	hg38, mm10, 11 others	hg19	hg38
HiChIP Processing	Total HiChIP Samples	1319 total (1031 distinct)	772	200	42
	Loop Calling	FitHiChIP, HiCCUPS	ChIA-PET Tool (V3)	FitHiChIP, hichipper	HiCCUPS
	Loop Resolutions	5kb, 10kb, 25kb	variable	1kb, 5kb, 10kb, 50kb, variable	5kb, 10kb, 25kb
	Peak Type Used for Loop Calling	HiChIP-Inferred, ChIP-seq when available	HiChIP-Inferred	HiChIP-Inferred	N/A
	Replicate Handling	Technical/Biological Reps Merged, Multiple Donors Merged	Biological Replicates Merged	Technical/Biological Reps Merged	Technical Reps Merged
	Pipeline Code Released	☑	☑	□	☑
Data Download	Loop Calls	☑	☑	☑	☑
	Peak Calls Used for Loop Calling	☑	□	□	□
	Browser Track Files	☑	□	□	□
Data Visualization	Embedded Visualization	Epigenome Browser (WashU)	Epigenome Browser (WashU)	IGV	Epigenome Browser (WashU)
	Multi-Sample Selection	☑	□	☑	☑
	Multi-Sample Visualization	☑	□	□	☑
Data Analysis	Community Structure Analysis, Loop Motif Pair Analysis, SGL Analysis	Embedded Tools	Functional Anchor Annotation*, Cancer High-Frequency Loops, Species-Specific High Frequency Loops	Functional Anchor Annotation**	Functional Anchor Annotation***, Loop discovery given genome regions, Prediction of regulatory sites and target genes
Additional Data	High Resolution Hi-C (n=44)	Hi-C and Other Conformation Capture Data	High Resolution Hi-C (n=44)	PLAC-seq (n=89) ChIAPET (n=789)	Hi-C (n=385) ChIAPET (n=119)

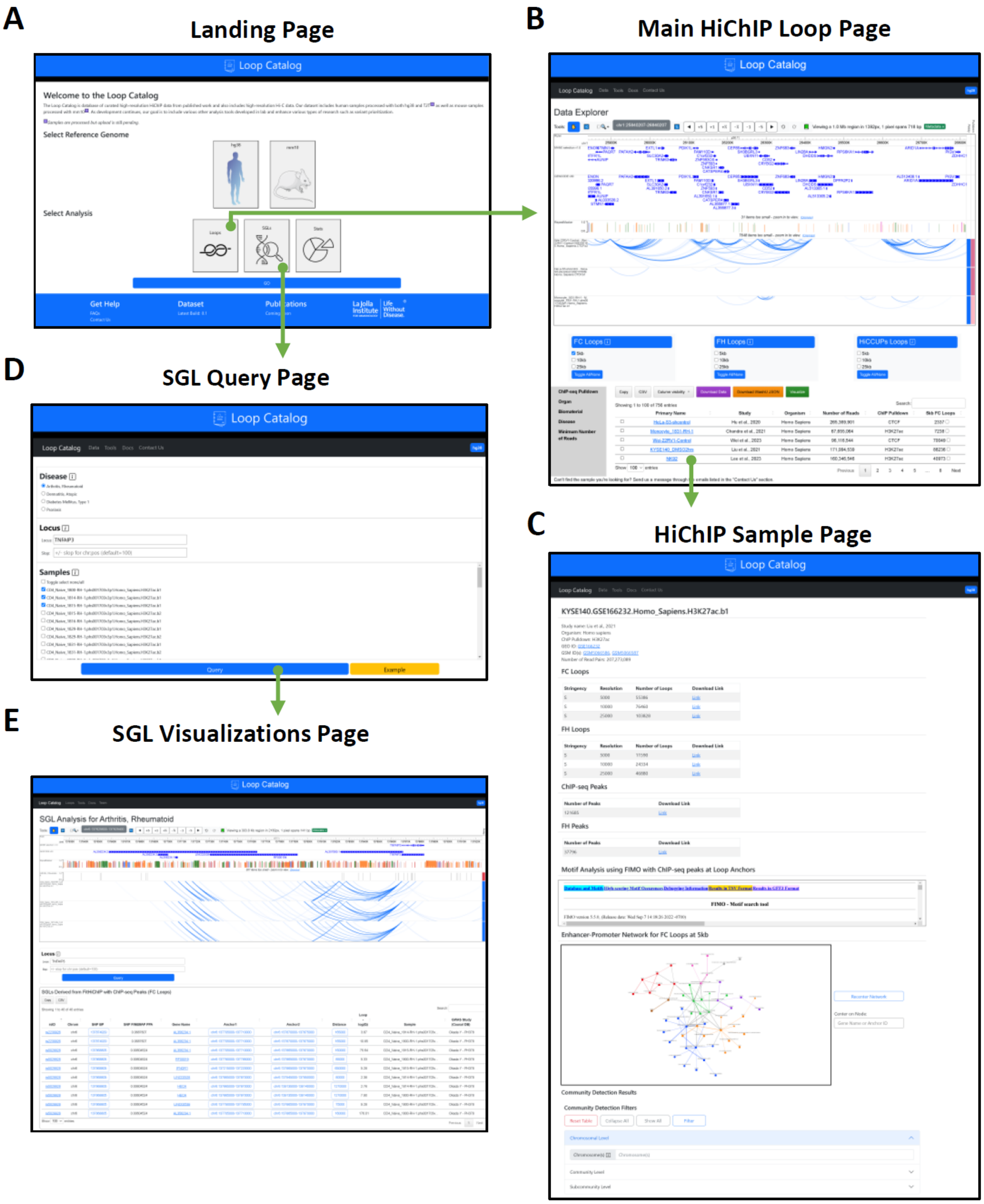
[100], ChromLoops [101], and Cohesin-DB [102], were first attempts to catalog HiChIP loops by curating them from the broader literature. We provide a detailed comparison of Loop Catalog with these previous databases on the basis of the number of samples processed, the choice of tools for data processing including loop calling, integrated visualization and additional data types compiled (Table 1.1). In addition to having the largest number of HiChIP samples, unique features of Loop Catalog include utilization of matched ChIP-seq data when available, enabling a broader set of data download and visualization capabilities and additional modalities including SGL mapping, traditional and pairwise motif enrichment analysis and regulatory network analysis (Table 1.1; additional modalities will be covered in Chapter 2). The large set of loops cataloged in this work will enable not only 3D-informed prioritization of genetic variants and enhancers involved in gene regulation but also will stimulate development of machine learning, deep learning, and network construction approaches that require large scale data.

1.3 Results

1.3.1 Curating a set of publicly available HiChIP samples for human and mouse cells

In total, we collected 750 human and 281 mouse samples from 133 studies that cover a diverse set of cell types and cell lines. For primary human samples there is a concentration of immune cell types that include monocytes, natural killer (NK) cells, T cell and B cell subsets, among others (175 designated samples; 140 and 35 for human and mouse, respectively). As expected, cancer cell lines are well represented (e.g., HCC15, NCI-H1105, MCF7) together with other cell lines from normal tissue including heart-derived samples (e.g., aortic valve interstitial, coronary artery smooth muscle, and aortic smooth muscle cells) as listed in Supp Table 1.1. Regarding the target protein in the HiChIP experiment, active regulatory element-associated histone mark H3K27ac was the most highly represented, accounting for 62% of human samples and 53% of mouse samples. Other frequently represented ChIP targets include CTCF for human datasets and H3K4me3 for mouse datasets. Cohesin subunits such as SMC1A was also a frequent protein of choice in both human and mouse samples (Figure 1.1C). It is evident that, while the majority of studies target functional interactions via H3K27ac or H3K4me3 pulldown, structural interactions via CTCF and cohesin pull-down are also well-represented among human and mouse HiChIP studies. Overall, this dataset provides a comprehensive coverage of HiChIP samples that investigate structural as well as regulatory loops profiled across hundreds of samples.

Figure 1.2. Layout of the Loop Catalog portal. **A)** Screenshot of the main entry page to the Loop Catalog. **B)** Main data page which includes an embedding of the WashU Epigenome Browser followed by a table of HiChIP samples with various metadata fields. **C)** Screenshot of a HiChIP sample page with statistics on the loop calls and the enhancer-promoter network visualization with an associated table on ranking of identified network communities. Subcommunities are color coded with nodes belonging to either enhancers (circles), promoters (squares) or other (triangles). **D)** Screenshot of the SGL entry page where disease, locus/gene and HiChIP samples are selected. **E)** Screenshot of the SGL analysis page which includes an embedded browser and a table with SGL metadata including buttons for navigation.


A
Landing Page
Loop Catalog
Welcome to the Loop Catalog
Select Reference Genome
hg38
mm10
Select Analysis
Loops
SGL
Stats
GO
Get Help
Dataset
Publications
La Jolla Institute
Life Without Disease
B
Main HiChIP Loop Page
Loop Catalog
Data Explorer
FC Loops
FH Loops
HiCCUPs Loops
5kb
10kb
25kb
C
HiChIP Sample Page
Loop Catalog
KYSE140.GSE166232.Homo_Sapiens.H3K27ac.b1
FC Loops
FH Loops
ChIP-seq Peaks
FH Peaks
Motif Analysis using FIMO with ChIP-seq peaks at Loop Anchors
FIMO - Motif search tool
Enhancer-Promoter Network for FC Loops at 5kb
Community Detection Results
Community Detection Filters
D
SGL Query Page
Loop Catalog
Disease
Arthritis, Rheumatoid
Dermatitis, Atopic
Diabetes Mellitus, Type 1
Psoriasis
Locus
TNFAIP3
Samples
Query
Example
E
SGL Visualizations Page
Loop Catalog
SGL Analysis for Arthritis, Rheumatoid
Locus
Query

1.3.2 Uniform processing of HiChIP samples and quality controls

The raw HiChIP data aligned to human (hg38) or mouse (mm10) genomes were processed to generate high-confidence loops utilizing multiple loop calling approaches: i) FitHiChIP with ChIP-seq peaks (FC), ii) FitHiChIP with HiChIP-inferred peaks (FH), and iii) HiCCUPS (Supp Figure 1.1). In addition, we performed loop calling using Mustache [103] for 44 high-resolution Hi-C samples available from the 4DN data portal (Supp Table 1.9). To visualize these loops, we developed a webpage with a highly interactive table that allows selection of multiple samples, loop calling settings, resolutions and corresponding peaks (Figure 1.2). These selections can then be plugged into an embedding of the WashU Epigenome Browser, downloaded as track files or WashU hub files for additional use cases (Figure 1.2B). The Loop Catalog was initialized with loops called for 1031 unmerged HiChIP samples and 282 merged samples that were created after combining biological replicates from the same study that pass a correlation threshold for pairwise similarity of individual replicates (Supp Figure 1.3). An additional 6 immune cell-based mega-merged samples were created by merging across all donors and all biological replicates using data from two previous publications [39, 104]. Overall, the Loop Catalog provides access to loops for 1319 HiChIP samples with FH loop calls for all samples and, of these, 386 samples had matching ChIP-seq data that enable loop calling using FC. In addition, we called HiCCUPS loops genome-wide for 426 samples, for which we had at least a certain number of loop calls from chr1 (200 for human, 100 for mouse) (Supp Figure 1.4). In particular, when comparing these figures to other databases, the Loop Catalog stands out with the highest number of HiChIP samples, totaling 1031 distinct and 1319 overall with merged samples. ChromLoops provides loop calls for 816 samples (772 HiChIP), however this total is spread across 13 different species, whereas HiChIPdb and CohesinDB offer significantly fewer HiChIP samples, with only 200 and 42, respectively (Table 1.1). It is noteworthy that various loop-calling methods were employed across all databases. With this consideration, HiChIPdb is most similar to the Loop Catalog, using mainly FitHiChIP for loop calling (Table 1.1). However, unlike the Loop Catalog, HiChIPdb derives peaks solely from HiChIP data, which tends to have a low recall rate with respect to ChIP-seq based peaks (Supp Figure 1.2). Conversely, Chromloops opted to use ChIA-PET Tool (V3), a tool specifically designed to address concerns for ChIA-PET experiments and may fail to correct for HiChIP biases. With regards to Cohesin-DB, this database used HiCCUPS, a software that is originally designed for high-depth Hi-C data [105] (Table 1.1). Despite using HiCCUPS in addition to FitHiChIP, we found that, for most of our curated HiChIP samples, the number of HiCCUPS loops were quite low (e.g., less than 200 for human chr1) likely limiting the utility of those loop calls for downstream analysis. Another important feature unique to Loop Catalog is the extent of quality control that is performed on both peak (ChIP-seq or HiChIP-derived) and loop level to provide QC flags for each set of loop calls to inform users of the inferred quality of the data they are about to visualize or download for further analysis (Supp Figure 1.5, Supp Table 1.7).

1.3.3 Visualization and exploration of loop calls through a web interface

The Loop Catalog is underpinned by a comprehensive database that incorporates processed HiChIP data from GEO and dbGaP, high-resolution Hi-C data from the 4DN data portal, along with fine-mapped GWAS data for a number of immune-associated diseases and eQTLs for major immune cell types (Figure 1.1). When first accessing the platform, users are presented with an entry page featuring a selection for reference genome and analysis type including loops, SNP-Gene linking (SGLs) and statistics (Figure 1.2A). Subsequently, once an analysis has been selected, a secondary page will render the corresponding analysis with a navigation menu to switch between other analyses and website-related information. The data page offers immediate and extensive visualization of HiChIP and Hi-C samples with their associated loop calls, spanning various methods and resolutions, as illustrated in Figure 1.2B. For specific sample information, each sample is linked to a dedicated page displaying metadata, loop data, regulatory network analyses and a motif scanning analysis (Figure 1.2C). Furthermore, the SGL page grants access to the integration of immune-based HiChIP samples and fine-mapped GWAS SNPs and eQTLs. Lastly, the statistics page provides a higher-level overview of all HiChIP data stored for a given reference genome.

1.4 Conclusion

The HiChIP assay has empowered the field of chromatin structure by providing a relatively inexpensive, high-resolution and targeted approach to mapping chromatin interactions. It is no surprise that the number of HiChIP studies published annually keeps growing each year and will continue to do so until superior methods are developed. The Loop Catalog is a public hub that centralizes 1319 HiChIP samples from these studies and, through the use of user-facing tools, lowers the bar for exploring 3D genome organization datasets, which would require substantial bioinformatics skills otherwise. As previewed by our applications, we foresee the Loop Catalog becoming a valuable resource for a broad range of chromatin studies including but not limited to variant-gene prioritization, machine learning, and deep learning approaches for loop prediction or utilizing loops to predict other functional measurements including gene expression as well as benchmarking analyses of such methods.

At its current state, the Loop Catalog allows the community to access and bulk download uniformly processed chromatin looping data from HiChIP experiments and intermediate files, as desired. Looking ahead, we plan to continue expanding the Loop Catalog as newly published datasets become available. Our overarching and ongoing goal is to build a chromatin-centric database that seamlessly expands in terms of data as well as computational utilities offered that will eventually become the go-to platform for access and analysis of all published HiChIP data.

1.5 Methods

1.5.1 Curating HiChIP and ChIP-seq Samples

To identify a comprehensive list of publicly-released HiChIP datasets, we developed a pipeline that scans NCBI's Gene Expression Omnibus (GEO) database [106] for studies performing HiChIP experiments. To extract information on these studies the BioPython.Entrez [107] and metapub.convert (https://pypi.org/project/metapub/) packages were used. Raw sequencing data associated to these studies was then identified from the SRA database using the pysradb Python package (https://github.com/saketkc/pysradb) and the results were manually examined to extract HiChIP samples. ChIP-seq samples corresponding to these studies were also extracted if there was a record of them within the same GEO ID as the HiChIP sample.

1.5.2 Populating Sample Metadata

To automatically add metadata, such as organ and cell type, to the curated HiChIP data, the BioPython.Entrez package was used to perform a search query within the GEO database. GSM IDs from the previous query were used with an esummary query to the biosample database [107]. From these results organism, biomaterial, celltype, GSM ID, SRA ID, disease, organ, treatment, tissue, and strain (only for mouse) were extracted. To improve classes for organ and biomaterial, a dictionary of classes as keys and synonyms as value (e.g. heart is a key and its synonyms are cardiovascular, atrium, aorta, etc) as used to search across GEO and Biosample reports. "N/A " was used to indicate fields that are not applicable for a given sample and for the remaining fields, "Undetermined" was assigned when the information could not be retrieved.

1.5.3 Downloading Raw Sequencing Datasets

HiChIP FASTQ files were systematically downloaded using one of three methods and prioritized in the following order: SRA-Toolkit's prefetch and fasterq-dump (2.11.2) (https://hpc.nih.gov/apps/sratoolkit.html), grabseqs (0.7.0) [108], or EBI URLs generated from SRA Explorer (1.0) [109]. In addition, FASTQ files for phs001703v3p1 and phs001703v4p1 were downloaded that pertain to two previously published HiChIP studies from the Database of Genotypes and Phenotypes (dbGaP) [39, 104, 110] (Supp Table 1.1). Additionally, ChIP-seq FASTQ files were downloaded from GEO using grabseqs (0.7.0).

1.5.4 Processing Reads from HiChIP Data

Human samples were aligned to hg38 while mouse samples were aligned to mm10 using the HiC-Pro (3.1.0) pipeline [111] (Supp Table 1.2). Reference genomes were downloaded for hg38 (https://hgdownload.soe.ucsc.edu/goldenPath/hg38/bigZips/), and for mm10 (https://hgdownload.soe.ucsc.edu/goldenPath/mm10/bigZips/) and indexed with bowtie2 (2.4.5) using default parameters. All technical replicates were grouped into their respective biological replicate before read alignment. Restriction enzymes and ligation sequences were determined during the HiChIP literature search (Supp Table 1.3) and restriction fragments were generated with the HiC-Pro digestion.py tool using default parameters. HiC-Pro was then run in parallel mode by splitting FASTQ files into chunks of 50,000,000 reads using HiC-Pro's split_reads.py utility with a minimum MAPQ threshold of 30. HiC-Pro generated contact maps at the 1kb, 2kb, 5kb, 10 kb, 20kb, 40kb, 100kb, 500kb, and 1Mb resolutions. For all other HiC-Pro configuration parameters, the defaults were used.

1.5.5 Processing Reads and Calling Peaks for ChIP-seq Data Using ChIPLine

Our previously developed ChIPLine pipeline (https://github.com/ay-lab/ChIPLine) was used to align ChIP-seq reads with Bowtie2 [112] and call peaks using MACS2 (2.2.7.1) [113] (Supp Table 1.4). If available, the corresponding input ChIP-seq files were used by MACS2 in ChIPLine for peak calling using input vs treatment mode. MACS2 peak calls were made without an input file if no such file was available for the sample. For ChIP-seq datasets with multiple biological replicates, the replicate with the largest number of peak calls was selected as the peak set to be used in downstream analyses for loop calling.

1.5.6 Calling Peaks from Reads of HiChIP Data Using FitHiChIP

In the absence of matching ChIP-seq data, we turned to inferring peaks directly from HiChIP data. We called HiChIP peaks the PeakInferHiChIP.sh utility function of FitHiChIP 10.0

version [114], which takes processed interaction pairs generated by HiC-Pro as input (i.e., one valid read pair corresponds to two individual reads) and utilizes MACS2 for peak calling using the PeakInferHiChIP.sh script by specifying the correct HiC-Pro output directory, reference genome, and read length as reported by SRA Run Selector entry for a given sample (Supp Table 1.4). If read lengths differed across technical replicates for a single HiChIP biological replicate, the mode read length was used. For cases in which there was no single mode read length, the longest read length was selected. For merged biological replicates, the mode read length of all individual biological replicates for a given sample was chosen, and similarly, the longest read length was selected in the case of multiple modes. We additionally used HiChIP-Peaks (0.1.2) [60] with its default parameters to infer peaks from HiChIP data; however, these peak calls spanned untypically large-sized genomic regions (median 2.15 kb and up to 122 kb) and we decided to not utilize these peak calls.

1.5.7 Performing Recall Analysis of 1D Peaks Inferred from HiChIP Data

Peaks called from ChIP-seq datasets were considered the ground truth set and peaks inferred directly from HiChIP datasets were assessed for their overlap with ChIP-seq peaks (Supp Figure 1.2). To measure the validity of HiChIP-inferred peaks, we compared them to the peaks from ChIP-seq data and computed the percentage of ChIP-seq peaks recovered using HiChIP-inferred peaks. To obtain the intersection, each pair of corresponding HiChIP and ChIP-seq peak sets were intersected using bedtools allowing for 1kb slack on both sides.

1.5.8 Integration of Biological Replicates of HiChIP Data

In order to generate more deeply sequenced contact maps from the initial set of samples, we grouped together samples originating from the same study, pulldown and biological replicate set for both human and mouse datasets (Supp Table 1.1). Before merging the biological replicates, the reproducibility was assessed using hicreppy (0.1.0) which generates a stratum-adjusted correlation coefficient (SCC) as a measure of similarity for a pair of HiChIP contact maps (https://github.com/cmdoret/hicreppy) (Supp Table 1.5). Briefly, contact maps in hic format were converted to cool format for hicreppy input at 1kb, 5kb, 10kb, 25kb, 50kb, 100kb, 250kb, 500kb, and 1mb resolutions using hic2cool (0.8.3) (https://github.com/4dn-dcic/hic2cool). hicreppy was run on all pairwise combinations of biological replicates from a given HiChIP experiment as follows: first, hicreppy htrain was used to determine the optimal smoothing parameter value, or h, for a pair of input HiChIP contact matrices at 5kb resolution. htrain was run on a subset of chromosomes (chr1, chr10, chr17, and chr19) and with a maximum possible h-value of 25. Default settings were used otherwise. Next, hicreppy scc was ran to generate a SCC for the matrix pair using the optimal h-value reported by htrain and at 5kb resolution considering chr1, chr10, chr17, and chr19 only. A group of biological replicates were merged if all pairwise combinations of replicates in that group resulted in a SCC greater than 0.8 (Supp Figure 1.3). Merging of biological replicates was performed by concatenating HiC-Pro validpairs files from which we performed all downstream peak and loop calling steps of our pipeline.

1.5.9 Identifying Significant Chromatin Loops from HiChIP Data

Loop calling was performed for both unmerged and merged HiChIP biological replicates by (a) HiCCUPS (JuicerTools 1.22.01) [49], (b) FitHiChIP with HiChIP-inferred peaks (FH loops), and (c) FitHiChIP with ChIP-seq peaks (FC loops), when available [105] (Supp Table 1.6, Supp Figure 1.4). Briefly, HiC-Pro validpairs files for each sample were converted to .hic format using HiC-Pro's hicpro2juicebox utility with default parameters. HiCCUPS loop calling was initially performed for chr1 only with the following parameters: –cpu, –ignore-sparsity, -c chr1, -r 5000,10000,25000, and -k VC. Samples which passed the thresholds of at least 200 loops from chromosome 1 for human samples and at least 100 loops from chromosome 1 for mouse samples, both at 10kb resolution, were processed further with HiCCUPS for genome-wide loop calling using the same parameters. FitHiChIP loop calling was run with HiC-Pro validpairs as input at the 5kb, 10kb, and 25kb resolutions for the Stringent (S) background model with coverage bias regression, merge filtering, peak-to-all interactions and an FDR threshold of 0.01. Default values were used for all other FitHiChIP parameters.

1.5.10 Assigning QC Flags to Peak and Loops Calls

To assign a set of comprehensive QC flags to all HiChIP samples (unmerged and merged biological replicates), we derived criteria from quantifying numbers of peak and loop calls (Supp Table 1.7, Supp Figure 1.5). For both FH and ChIP-seq peaks, we assigned a “Poor” flag to samples with <=1000 peaks, “Warning” to samples with >1000 and <5000 peaks, and “Good” to samples with >=5000 peaks. To establish intermediate flags for loops, each one of the six FitHiChIP loop call configurations (Stringent 5kb, Loose 5kb, Stringent 10kb, Loose 10kb, Stringent 25kb, and Loose 25kb) was assigned a flag as follows: “Poor” if the number of Stringent loops was greater than the number of Loose loops at a given resolution or the sample possessed <=100 loops, “Warning” for >100 and <1000 loops, or “Good” for >= 1000 loops. Lastly, we set a final loop flag on each loop call configuration which considers both the peak and intermediate loop flags. This final loop flag similarly takes on the values of “Poor”, “Warning”, or “Good”. “Good” was only assigned in cases where both the peak and intermediate flags were “Good”. All other possible cases are further described in Supp Table 1.7.

1.5.11 Establishing a Set of High-Confidence Regulatory HiChIP Loops

We established a high-confidence set of 53 unique human H3K27ac HiChIP samples from the pool of 120 human H3K27ac HiChIP datasets with either merged or unmerged biological replicates and over 10,000 stringent 5kb FC loops, henceforth called the HCRegLoops-All sample set. We additionally established two subsets of the HCRegLoops-All sample set as follows: HCRegLoops-Immune contains 27 samples from immune-associated cell types and HCRegLoops-Non-Immune contains 26 samples from non-immune cell types (Supp Table 1.8).

1.5.12 Identifying Significant Chromatin Loops from High Resolution Hi-C Data

Loop calling was performed for 44 high resolution Hi-C samples gathered from the 4DN data Portal's (https://data.4dnucleome.org) "High-Resolution Hi-C Datasets" Collection using Mustache (Roayaei Ardakany et al., 2020). Processed .hic files were downloaded and Mustache was run with default parameters across chr1 through chr22 using raw, KR, and VC normalized contact matrices (-norm) to determine loops at 1kb, and 5kb resolutions (-r) (Supp Table 1.9).

1.5.13 Designing the Internal Database and Filesystem

The database is composed of two main parts, the first contains high sample-level and low-level loop information while the second contains SGL specific tables with auxiliary tables used to add additional annotations. For the first part, we atomized the data into the following tables: hic_sample, celltype, hicpro, chipseq_merged, fithichip_cp, fithichip_cp_loop, hiccup, fithichip_hp and reference which allowed us to capture important metadata and the uniqueness of loops using different loop callers. The second part includes the gwas_study, gwas_snp, snp, gene, fcp_fm_sgl, eqtl_study, eqtl, and fcp_eqtl_sgl tables which allowed us to capture important relationships for SGLs and to facilitate their query. The full database schema can be found in Supp Figure 1.6.

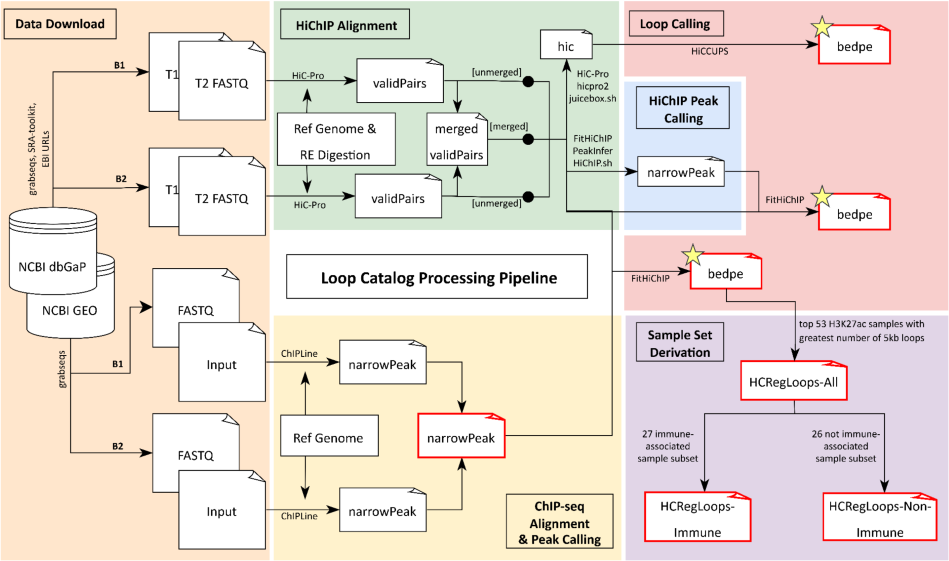
1.5.14 Designing and Implementing the Web Interface

Loop Catalog was built using Django v3.2 (https://www.djangoproject.com/) as the backend framework with all data stored using the Postgresql v9 (https://www.postgresql.org/) database management system. To style the frontend interface, we used Bootstrap 5.2 (https://getbootstrap.com/docs/5.2/). For implementing advanced tables and charts DataTables v1.12.1 (https://datatables.net), Charts.js v4.0 (https://www.chartjs.org/), and D3 v4.13.0 (https://d3js.org/) were used. To visualize genetic and epigenetic data the WashU Epigenome Browser v53.6.3 (https://epigenomegateway.wustl.edu/) was used which maintains an easily accessible web embedding. Lastly, CytoscapeJS v3.26.0 (https://js.cytoscape.org/) was used to visualize enhancer-promoter networks.

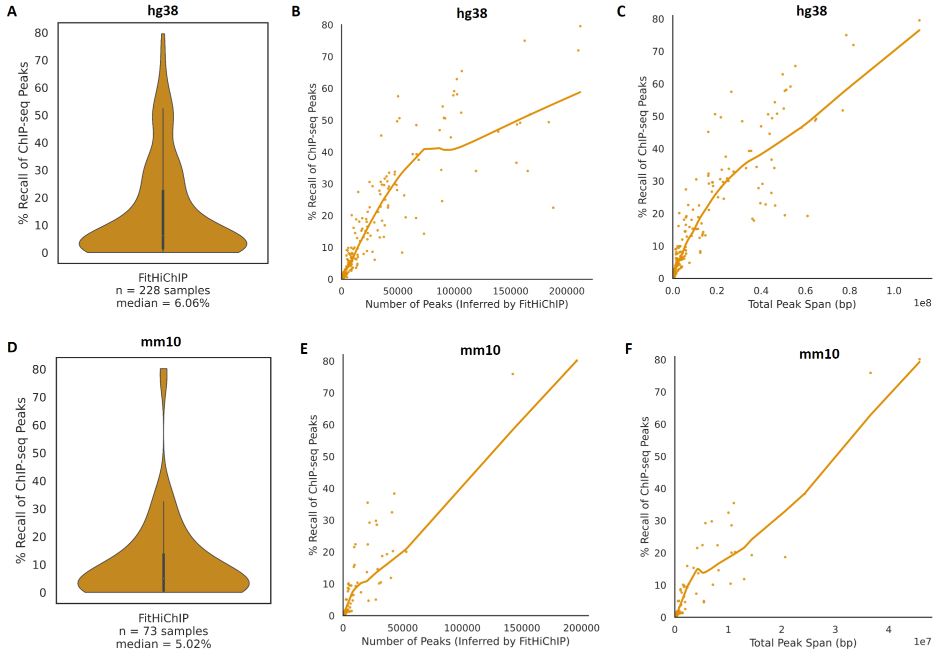
1.6 Availability of data and materials

The Loop Catalog is freely available at https://loopcatalog.lji.org to all users without any log-in or registration requirements. The main processing pipeline has been released on Github as Loop-Catalog-Pipelines (https://github.com/ay-lab-team/Loop-Catalog-Pipelines). Similarly, we developed GEO-Resources (https://github.com/ay-lab-team/GEO-Resources) to locate HiChIP datasets in NCBI GEO. Versions of all software included in our pipeline are recorded (Supp Table 1.10). Raw sequencing reads for HiChIP and ChIP-seq were downloaded from NCBI GEO (https://www.ncbi.nlm.nih.gov/geo/) and NCBI dbGaP (https://www.ncbi.nlm.nih.gov/gap). Hi-C contact matrices were retrieved from the 4DN Data Portal (https://data.4dnucleome.org).

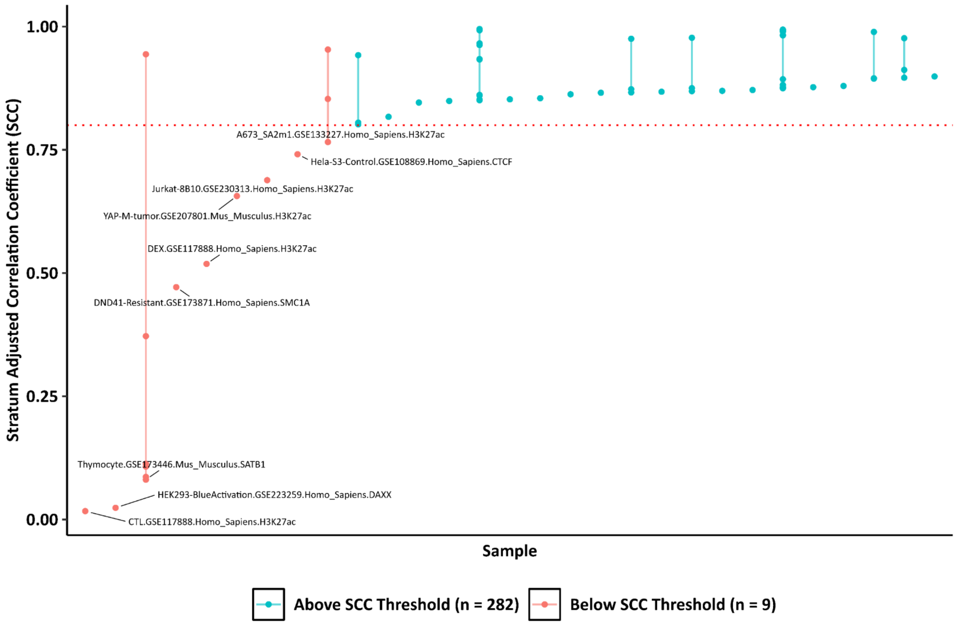
1.7 Supplementary Figures



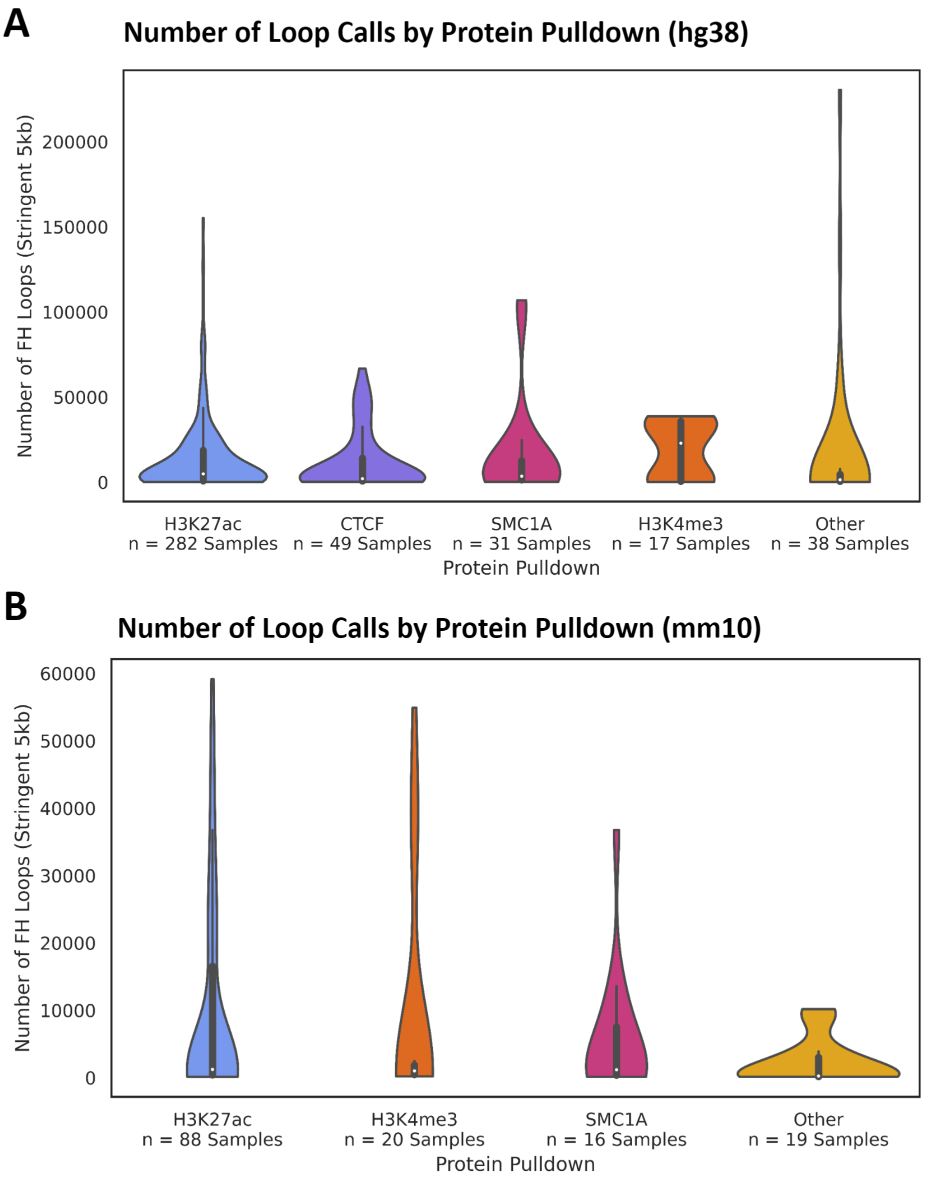
Supp Figure 1.1. Schematic of the HiChIP and ChIP-seq data processing pipeline we developed for the Loop Catalog. Raw sequencing reads are downloaded from NCBI GEO and dbGaP and are aligned to the reference genome (hg38 or mm10). Loops are called for HiChIP (both unmerged and merged biological replicates, as indicated by the shaded circles) using HiCCUPS and FitHiChIP at the 5kb, 10kb, and 25kb resolutions. Peaks derived from both ChIP-seq and HiChIP are used for FitHiChIP loop calling. High-confidence sample sets of top H3K27ac HiChIP samples are curated from the final set of loop calls. Red borders indicate usage in downstream database application analyses. Yellow stars indicate that the data type is available for download from the Loop Catalog web platform.



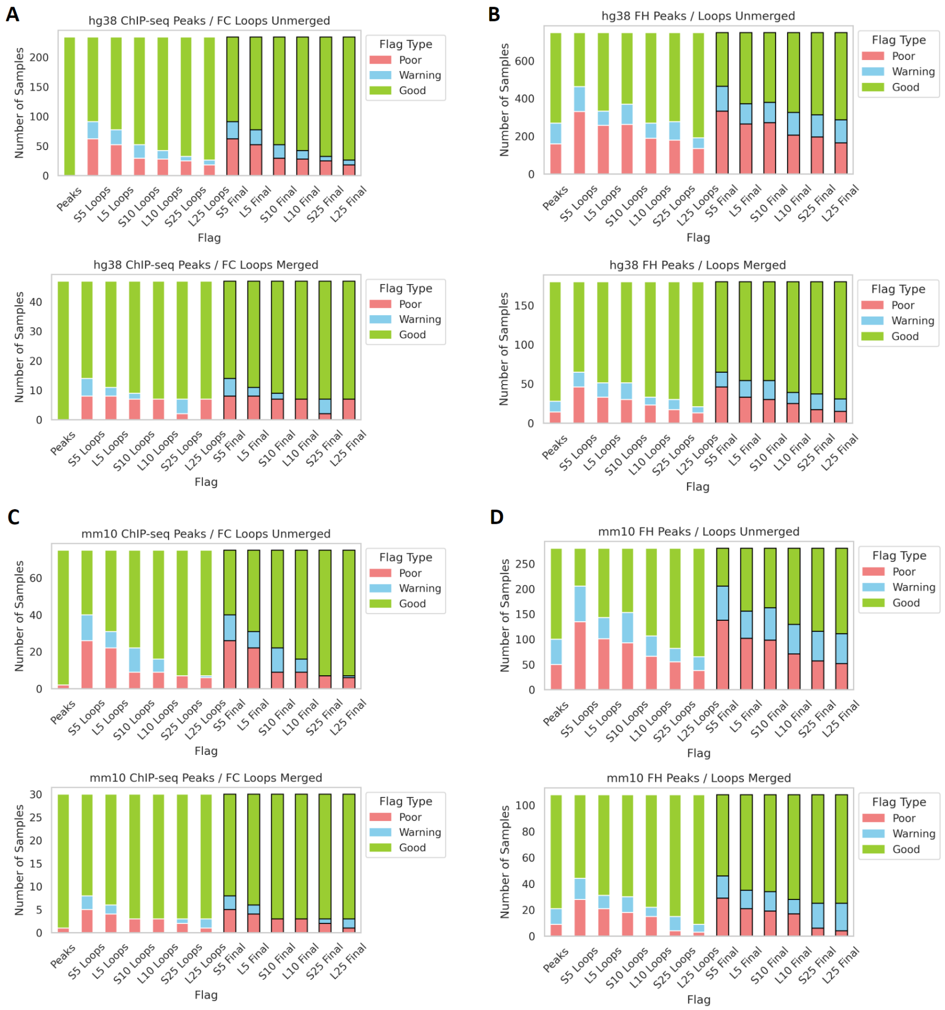
Supp Figure 1.2. Recall analysis of ChIP-seq peaks using peaks called from HiChIP reads. **A**) Distribution of recall rates of ChIP-seq peaks by peaks inferred from HiChIP by FitHiChIP in hg38. **B**) Scatter plot of recall rate versus the number of peaks in hg38. Samples with zero percent recall or with zero peak calls derived from HiChIP are not included. **C**) Scatter plot of the recall rate versus the total peak span in hg38. Samples with zero peak calls derived from HiChIP are not included. **D-F**) Represent the same analysis as A-C but using mouse HiChIP samples.



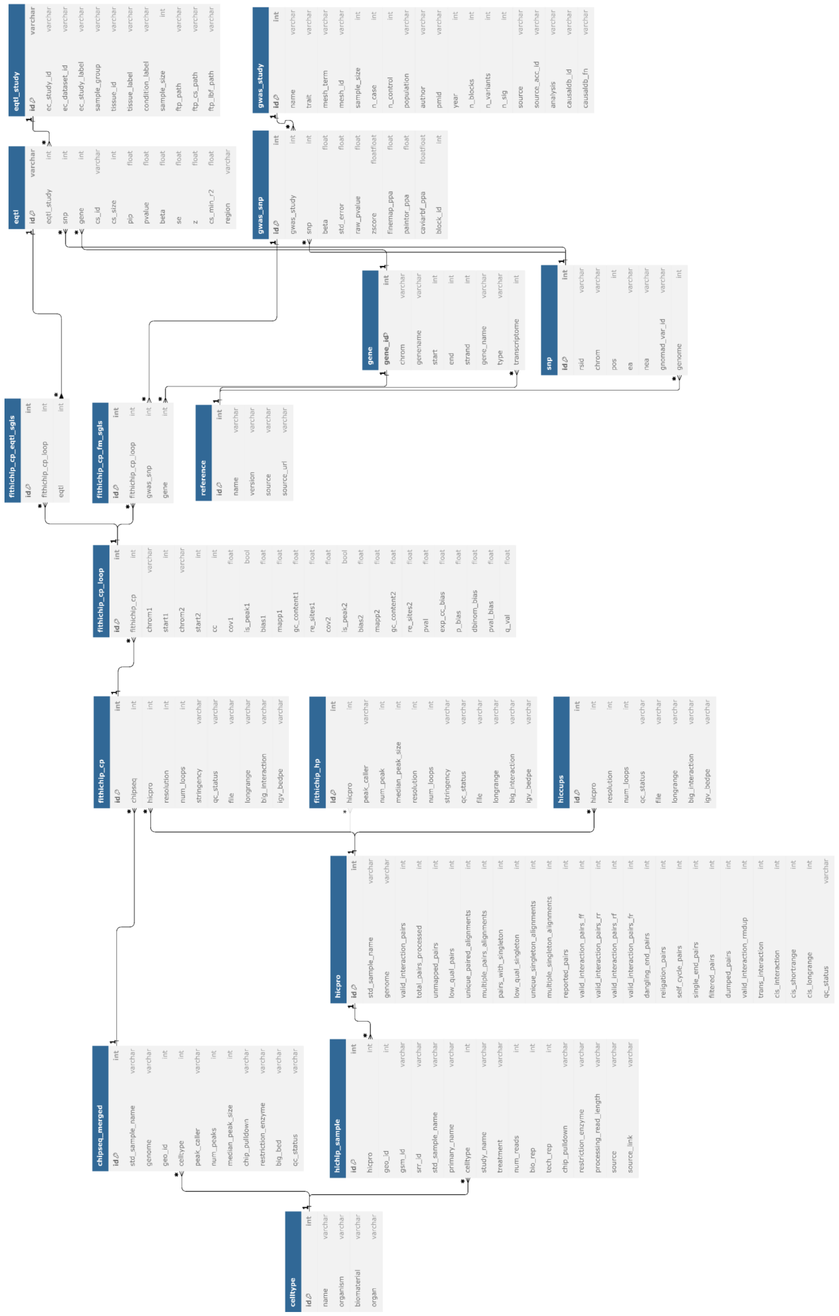
Supp Figure 1.3. Reproducibility analysis for HiChIP biological replicates. Stratum adjusted correlation coefficient scores (SCCs) for all pairwise combinations of HiChIP biological replicates for 291 HiChIP samples with at least two biological replicates were generated using hicreppy. SCCs for pairwise combinations of HiChIP biological replicates are displayed for the 29 HiChIP samples with a minimum SCC below 0.90. SCCs for pairwise combinations of replicates for samples with more than 2 replicates are connected with vertical lines. Samples passed the SCC threshold if the SCC of all pairwise combinations of replicates was greater than 0.80 (n = 282 samples, a subset of 20 passing samples shown in blue). Samples which possessed at least one replicate combination with a SCC less than 0.80 did not pass the threshold (n = 9 samples, all shown in red).



Supp Figure 1.4. Number of FH loop calls by protein pulldown. Distributions of number of FH loop calls by protein pulldown are displayed for **A**) human and **B**) mouse samples with >100 loops. Proteins represented by >10 samples are individually displayed while all others are grouped into the "Other" category. These include proteins such as RNA-Pol-II, GATA1, STAG1, STAG2, RAD21, etc.



Supp Figure 1.5. Assignment of QC Flags to HiChIP Samples. 13 flags were assigned to all samples. For peak calls, a flag of "Poor" was assigned if the sample possessed <=1000 peaks, "Warning" to >1000 and <5000 peaks, and "Good" to >=5000 peaks. Each configuration of FitHiChIP loop calls were each similarly assigned "Poor" (number of Stringent loops > number Loose loops at that resolution, or <=100 loops), "Warning" (>100 and <1000 loops), or "Good" (>= 1000 loops). A final flag was assigned based on the logic in Supplemental Table 7. The distribution of each flag type is displayed for **A**) human FC loops, unmerged and merged, **B**) human FH loops, unmerged and merged, **C**) mouse FC loops, unmerged and merged, and **D**) mouse FH loops, unmerged and merged. Distributions of final flags are highlighted with black borders.



Supp Figure 1.6. Schema for the Loop Catalog database tables. Lines connect tables based on relationships with children tables represented by a forked edge and asterisk.

1.8 Supplementary Tables

Supp Table 1.1: Summary of HiChIP Sequence Files

Supp Table 1.2: HiC-Pro Quality Control Statistics

Supp Table 1.3: Restriction Enzyme Sequences Used in the HiC-Pro Pipeline

Supp Table 1.4: Peak Call Statistics for HiChIP and ChIP-seq

Supp Table 1.5: Assessment of HiChIP Reproducibility by Stratum-Adjusted Correlation Coefficients

Supp Table 1.6: HiChIP Loop Call Statistics for FitHiChIP and HiCCUPS Loop Calling

Supp Table 1.7: Quality Control Flag Definitions for FitHiChIP Loop Calling

Supp Table 1.8. List of High Confidence H3K27ac HiChIP Sample Sets

Supp Table 1.9: Hi-C Loop Call Statistics for Mustache Loop Calling

Supp Table 1.10: Software and Package Versions

Chapters 1 and 2, in full, has been submitted for publication of the material as it may appear in Genome Biology 2024, Joaquin Reyna, Kyra Fetter, Romeo Ignacio, Cemil Can Ali Marandi, Nikhil Rao, Zichen Jiang, Daniela Salgado Figueroa, Sourya Bhattacharyya, Ferhat Ay. The dissertation author was one of the primary investigators and authors of this paper.

Chapter 2

Utilizing the Loop Catalog to Identify Important Regulatory Elements

2.1 Abstract

Chromatin conformation is the infrastructure of the genome, from loop to compartments, the organization of elements within the genome is essential to cellular specialization and function. Thus, detection of regulatory elements within HiChIP data and integrating with functional variants can deliver meaningful insights. In this chapter we look at the following: i) linking of target genes for GWAS signals, ii) prioritizing of eQTL signals with an overlapping loop, iii) 1D motif analysis, iv) paired motif analysis, v) community detection, and vi) the incorporation of these analyses to the Loop Catalog website for wider use. From these efforts we can see that the Loop Catalog provides much more that just HiChIP loop calls, these can be widely analyzed and integrated with various other datasets to derive a new understanding of genetics and disease.

2.2 Background

In the context of human disease, HiChIP and similar assays provide a 3D view for the annotation of disease associations of non-coding genetic variants identified from GWAS [58, 59, 99, 115, 116]. Combined with efforts from multiple large consortia for cataloging putative regulatory elements spanning distinct cell types (e.g., ENCODE, BLUEPRINT and Roadmap

Epigenomics), mapping such 3D maps of chromatin organization has become a critical piece of the epigenetics puzzle, which led to formation of the 4D Nucleome Consortium [24, 117]. In parallel, other large-scale efforts identified SNPs associated with gene expression (i.e., expression quantitative trait loci or eQTLs) for different tissues and primary cell types [22, 23, 37, 104, 118, 119].

Motivated by these large scale analyses that demonstrate the importance of chromatin conformation, we expanded upon the Loop Catalog to develop three new analyses and full-fledged features that leverage our high confidence loop calls. In the first application, we intersected fine-mapped GWAS SNPs from CAUSALdb for four autoimmune diseases [27] with HiChIP loops derived from various immune cell types to identify potential target genes for these disease associated SNPs in each cell type (SNP to gene linking with loops or SGLs). Across all diseases we located 3048 unique SNPs, 1486 genes, 3411 loops and 13672 SGLs that span a median genomic distance of ~140kb. One example was a fine-mapped Type 1 Diabetes (T1D) GWAS SNP which was linked to multiple genes through HiChIP loops in multiple lymphoid cells but not in monocytes. As the second application, we investigated TF motifs in loop anchors using motif enrichment analysis [120] and pairs of TF motifs at loop anchors through pairwise motif enrichment using a bootstrapping approach. Our analysis of regulatory loop anchors conserved across a large majority of samples identified hundreds of enriched motifs including known and novel zinc finger transcription factors some of which were also enriched for significantly overrepresented combinations of motif pairs across samples (e.g., ZNF460, ZNF135). For the last application, we constructed enhancer-promoter networks using H3K27ac HiChIP loops as the underlying connections between these regulatory elements. Analysis of this network via community detection and community ranking algorithms resulted in the identification of tightly connected network communities, which may act as modular units of gene regulation. These features showcase the richness of chromatin conformation data from the Loop Catalog. From integration with GWAS, eQTLs, ChIP-seq and DNA motifs, the Loop Catalog enables the investigation of regulatory connectivity across the genome, that may influence molecular mechanisms and disease.

2.3 Results

2.3.1 Expanding the use of HiChIP data for annotating GWAS SNPs and eQTLs using SNP-to-Gene Loops

Leveraging the Loop Catalog we integrated loops for immune-related HiChIP samples, we identified 79 samples with nonzero FC loops at 5kb (63 unmerged, 10 bio-rep merged, and 6 all donors merged). We then used these loop calls together with fine-mapped GWAS SNPs from CAUSALdb to find target genes for each SNP that we term a SNP-gene pair with a loop (SGL) (Figure 2.1A). Briefly, across T1D, RA, PS and AD there are 7729, 1121, 590, and 674 unique fine-mapped GWAS SNPs, respectively. After overlapping these SNPs with our loop anchors and genes connected through such loops, we found 74 samples with at least one SGL and a total of 182,306 SGL instances across 18 studies covering the four diseases. When removing duplicate SGLs we found 28,162 distinct SGLs which included 4241 SNPs, 2354 genes and 7269 loops (Figure 2.1B, Supp Table 2.1). To store these results into the database and to allow querying, we constructed SNP, GWAS, gene and loop-level tables with genomic coordinates. To then browse these results, an SGL entry page allows users to first select their target disease, locus and samples from which the loops will be derived (Figure 1.2D). Subsequently, the Loop Catalog returns an SGL analysis page that includes an embedded WashU Epigenome Browser element loaded with a track for fine-mapped SNPs and loops tracks for each sample. Below the browser users will find an interactive table that lists all mapped SGLs for their selection. This table also allows navigating between loops and within loops including the left or right anchor and SNP positions (Figure 1.2E).

Similar to GWAS variants, we expanded our annotation of genomic elements using eQTL SNP-gene pairs. We started by downloading uniformly processed eQTL studies from the eQTL Catalogue and focused on cell types from Schmiedel et al 2018 dataset that include eQTLs for

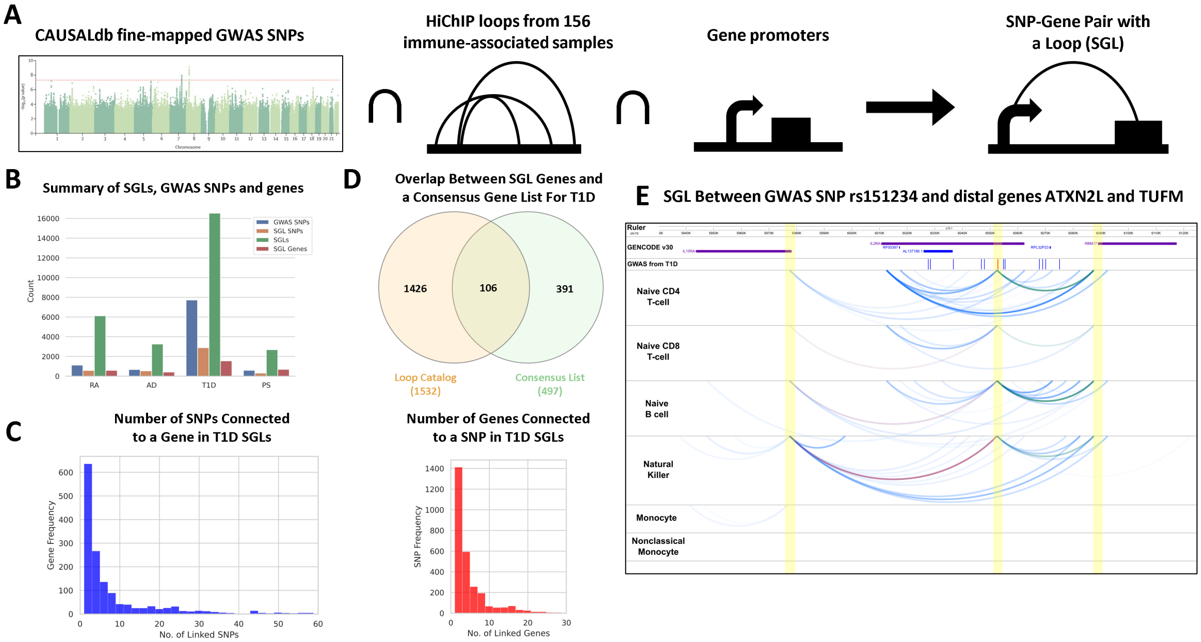


Figure 2.1. SGL analysis overview and results. A) Schema of the SGL analysis using fine-mapped SNPs from CAUSALdb, 156 Loop Catalog immune-related samples, and TSS coordinates. **B)** Summary of results across all 4 diseases including the total number of GWAS hits (blue), SNPs found in a SGL (orange), genes found in SGL (green) and total SGLs (red). **C)** Distribution of SNP counts with respect to each SGL gene (left) and the distribution of gene counts with respect to each SGL SNP (right) for T1D. **D)** Evaluating the number of SGL genes which belong to a consensus list of T1D genes (green) and unique (orange). **E)** Example of an SGL between rs61839660 (red) and the genes *IL15RA* (red arc) and **RBM17** (blue arc). Tracks contain H3K27ac HiChIP tracks for naïve CD4 T-cell, naïve CD8 T-cell, naïve B cell, Natural Killer, monocytes, nonclassical monocytes derived from the Chanra et al 2021 and Schmiedel et al 2021 samples that were merged across all donors.

naive CD4 T-cells (n=64386), naive CD8 T-cells (n=67793), naive B cells (n=60629), NK cells (n=48221), monocytes (n=66024) and nonclassical monocytes (58069). For these cell types, we also have the HiChIP data derived from a subset of the same donors [39, 104, 121]. For these cell types, we then located 11604 unique eQTL-SGLs that cover 11053 SNPs and 1128 genes (Supp Figure 2.2, Supp Table 2.2). These results are made available through a similar web interface as SGLs derived from GWAS (Supp Figure 2.3). It is possible to extend our eQTL-based SGLs analysis to the remainder of the eQTL Catalogue, however, matching cell types from eQTL studies to those from HiChIP studies is not a trivial task for most of the cases.

2.3.2 Utilizing T1D SGLs for SNP and Gene Prioritization

To better understand the utility of using SGLs for linking GWAS SNPs and genes, we focused our attention on analyzing SGLs in T1D. When compared to all other diseases, T1D has the highest number of unique SGLs (n=16,534) mainly due to the high number of fine-mapped GWAS SNPs as our starting point (Figure 2.1B). In post-GWAS analyses, it is important to distinguish putative causal SNPs among those that are in linkage disequilibrium (high LD). In cases where the phenotypic effect of the SNP is mainly through regulation of a distal gene, SGLs can corroborate important information to accurately annotate SNP function and to prioritize GWAS genes and SNPs while utilizing information on their 3D proximity. Investigating this for T1D, we observed that at least half of the CAUSALdb SNPs participate in an SGL, these SNPs are often in contact with multiple genes (Figure 2.1B). We further explored the multiplicity of SNP and gene links within T1D and found that 26% of SGL genes are linked to a single SNP, and the median number of SNP links per gene is 3. On the other hand, 27% of SGL SNPs are linked to a single gene and the median number of gene links per gene is 3 (Figure 2.1C).

To understand if SGL genes overlap genes with known T1D associations, we built a consensus gene list using MalaCards [122], eDGAR [123], OpenTargets [41], and GWAS Catalog [28] and a T1D review paper [124]. The union of the gene lists across these five resources had 497 genes in total, of which 106 overlapped with our 1532 SGL genes identified for T1D. The 1426 genes uniquely found by our SGL approach, although likely to involve false positives, are potential targets for future investigations (Figure 2.1D). One of these genes was *IL15RA*, a cytokine receptor that binds the pro-inflammatory cytokine IL-15 with high affinity and through cis and trans presentation of IL-15 impacts cellular functions of CD8 T cells as well as Natural Killer (NK) cells [125]. Despite *IL15RA* not being within the consensus T1D gene list, the IL15RA/IL-15 axis has been associated with T1D but whether this axis has a pathogenic or protective role has not been clear [126–128]. Through our SGL analysis of H3K27ac HiChIP loop calls from major immune cells, we identified looping between the *IL15RA* promoter and rs61839660, a SNP 75kb away that is highly associated with T1D and has been further prioritized by fine-mapping in three out of four T1D-GWAS studies with a posterior probability greater than 0.70. The corresponding loops are found for T cells, Naive B cells and NK cells but not for monocytes suggesting an important role for this SGL within the adaptive immune system. The more likely scenario is that rs61839660's T1D association is mediated through *IL2RA* given that this SNP falls within a constituent intron. However, specific loops connecting rs61839660 to *IL15RA* (P-value < 10-9) as well as to *RBM17* (P-value < 10-11) suggest the possibility of a pleiotropic effect for this SNP (Figure 2.1E, Supp Table 2.1). In addition, *RBM17*, an RNA-binding protein, has been previously shown to affect other autoimmune diseases such as rheumatoid arthritis [129]. As exemplified here, SGL analysis with the Loop Catalog may provide further evidence and/or mechanisms of action for a genetic variant and its target gene. In addition, it may help to find targets for GWAS variants whose target gene remains elusive.

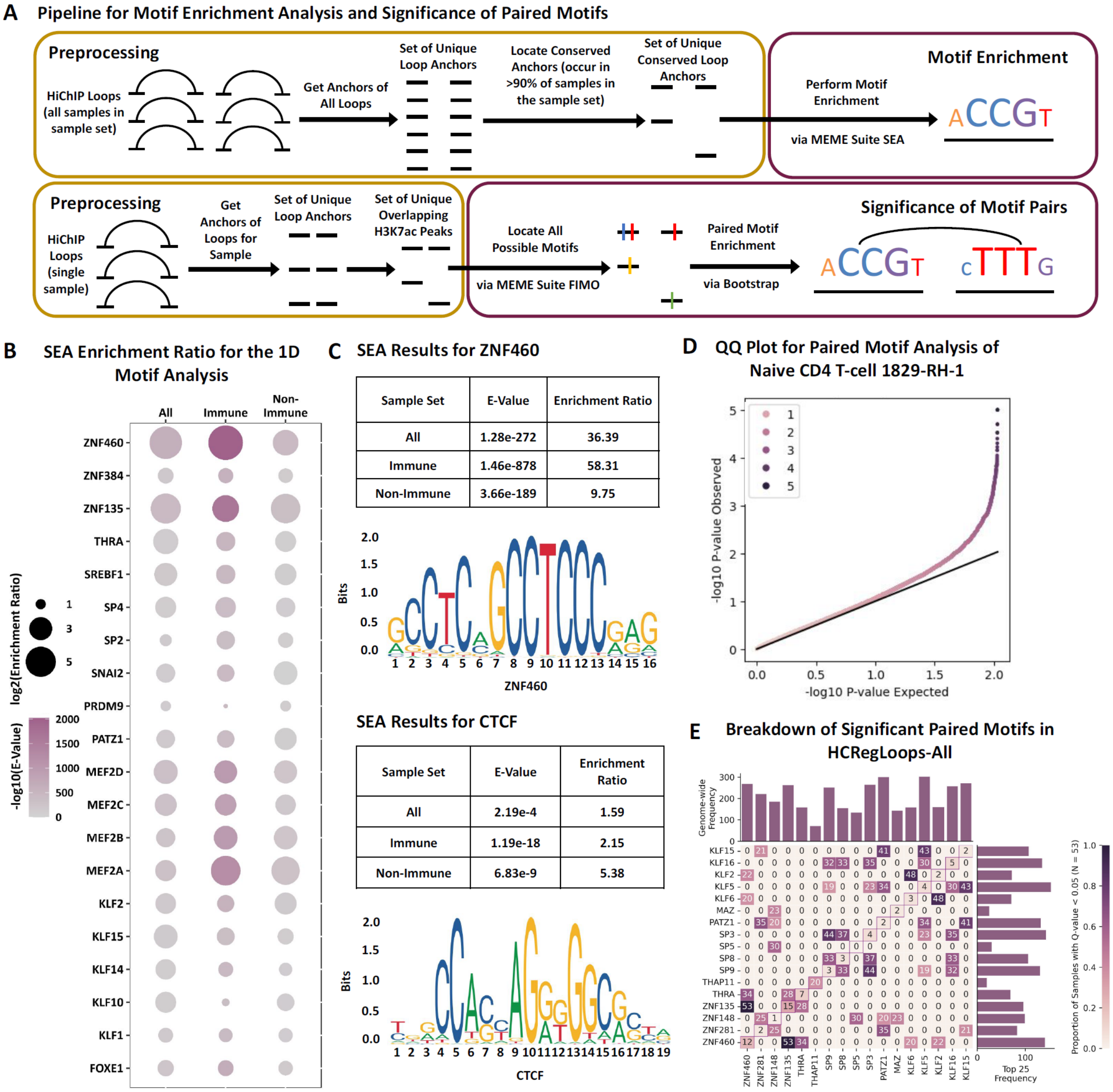
2.3.3 Identifying Significant Sequence Motifs at Regulatory Loop Anchors

In order to examine binding patterns of TFs in regulatory loops (H3K27ac HiChIP), we performed 1D motif enrichment analysis on highly conserved regulatory loop anchors from three high-confidence (HC) sample sets (Figure 2.2A). The HCRegLoops-All sample set contains the 53 H3K27ac HiChIP samples with over 10000 stringent 5kb FC loops. These samples encompass diverse cell types including immune cells, heart cells, and various cancer cell lines. The HCRegLoops-Immune sample set contains the subset of 27 samples from the HCRegLoops-All set which are from immune-associated cell types and the remaining 26 samples are contained in HCRegLoops-Non-Immune. In terms of conserved anchors that were derived for motif enrichment analysis, we annotated an anchor as conserved if was involved in at least one loop call in 90 or more of the samples from the given sample set and identified 879, 2766, and 721 anchors fitting this criterion for HCRegLoops-All, HCRegLoops-Immune, and HCRegLoops-Non-Immune, respectively. Using the Simple Enrichment Analysis tool SEA from the MEME

Suite, we identified 205, 313, and 177 significantly enriched motifs (p-value < 1e-6) for the three sample sets, respectively (Supp Table 2.3). The top 3 most significantly enriched motifs were *ZNF460*, *ZNF135*, and *MEF2D* or *MEF2A* across all three sets (Figure 2.2B). *ZNF460* demonstrated strikingly high enrichment relative to other motifs, especially in the HCRegLoops-Immune set, with over 58-fold enrichment compared to control regions in this sample set, in comparison to 2.15-fold enrichment for CTCF (Figure 2.2C). Although similar to CTCF in regard to possessing 11 zinc finger domains [130], *ZNF460* also harbors a KRAB domain (like one third of all ZNFs [131] that is associated with transcriptional repression. It is therefore surprising to have *ZNF460* motif enrichment in anchors of loops detected through enrichment of H3K27ac, an active histone mark. It is possible that *ZNF460* may be playing an important role in looping and one that is more pronounced for immune cells. However, recent studies highlight the importance of caution and the requirement for functional validation by showing that, for *ZNF143*, another ZNF with a presumed role in looping, such an association was a result of antibody cross-reactivity [132, 133]. Our motif enrichment analysis also identified two other ZNFs, MAZ and PATZ1, which are recently shown to have novel chromatin insulating activities, similar to CTCF [134, 135]. Regardless of the specific transcription factors, our provided analysis of TF enrichment at loop anchors for a diverse set of samples is useful for developing hypotheses about generic and cell-type-specific regulators or correlates of looping.

2.3.4 Identifying Significant Motif-Pairs Across Loop Anchors

Transcription factors often work together, through co-binding, dimerization, or multimerization, to promote gene expression. With this motivation, we expanded our motif analysis at conserved anchors to a search for significantly overrepresented motif pairs connected by loops. Since the brute-force approach of scanning all possible motifs across the genome for enrichment would have been infeasible, we performed bootstrap analysis to calculate an empirical p-value. We used the frequency of a given motif pair across the entire set of loops as our statistic and tested whether the frequency of this motif pair appeared greater than expected by chance for



SEA Results for ZNF460

Sample Set	E-Value	Enrichment Ratio
All	1.28e-272	36.39
Immune	1.46e-878	58.31
Non-Immune	3.66e-189	9.75

SEA Results for CTCF

Sample Set	E-Value	Enrichment Ratio
All	2.19e-4	1.59
Immune	1.19e-18	2.15
Non-Immune	6.83e-9	5.38

Breakdown of Significant Paired Motifs in HCRegLoops-All

	ZNF460	ZNF281	ZNF148	ZNF135	THRA	THAP11	SP9	SP8	SP5	SP3	PATZ1	MAZ	KLF6	KLF5	KLF2	KLF16	KLF15
KLF15	0	21	0	0	0	0	0	0	0	0	41	0	0	43	0	0	2
KLF16	0	0	0	0	0	0	32	33	0	35	0	0	0	30	0	5	0
KLF2	22	0	0	0	0	0	0	0	0	0	0	0	48	0	2	0	0
KLF5	0	0	0	0	0	0	19	0	0	23	34	0	0	4	0	30	43
KLF6	20	0	0	0	0	0	0	0	0	0	0	0	3	0	48	0	0
MAZ	0	0	23	0	0	0	0	0	0	0	0	2	0	0	0	0	0
PATZ1	0	35	20	0	0	0	0	0	0	0	2	0	0	34	0	0	41
SP3	0	0	0	0	0	0	44	37	0	4	0	0	0	23	0	35	0
SP5	0	0	30	0	0	0	0	0	0	0	0	0	0	0	0	0	0
SP8	0	0	0	0	0	0	33	3	0	37	0	0	0	0	0	33	0
SP9	0	0	0	0	0	0	3	33	0	44	0	0	0	19	0	32	0
THAP11	0	0	0	0	0	20	0	0	0	0	0	0	0	0	0	0	0
THRA	34	0	0	28	7	0	0	0	0	0	0	0	0	0	0	0	0
ZNF135	53	0	0	15	28	0	0	0	0	0	0	0	0	0	0	0	0
ZNF148	0	25	1	0	0	0	0	0	30	0	20	23	0	0	0	0	0
ZNF281	0	2	25	0	0	0	0	0	0	0	35	0	0	0	0	0	21
ZNF460	12	0	0	53	34	0	0	0	0	0	0	0	20	0	22	0	0

Figure 2.2. Motif and paired-motif analysis of loop anchors. A) Schematic of the 1D and 2D (paired) motif enrichment analysis. **B)** Bubble plot for the union of the top 15 motifs from each sample set (20 total motifs). The E-value is represented on a range from 0 (gray) to 2000 (magenta) and the log2(enrichment ratio) is represented by a circle radius from 1 to 7. **C-D)** Motif plot of ZNF460 and CTCF with the significance and enrichment ratio values as reported for each sample set. **D)** Q-Q plot testing p-values for Naive CD4 T-cell 1829-RH-1 sample. **E)** Heatmap of significant motif pairs (center) where rows and columns represent motifs on opposite anchors and each cell represents the proportion of samples where the given motif pair is significant. The distributions of a given motif across the whole genome and within the top 25 motif pairs are represented on the top and right, respectively.

a given loop set. Bootstrap was used to build a simulated null distribution by shuffling loop anchors and thus motif pairs (see methods "Performing Enrichment Analysis for Motif Pairs"). A p-value was then calculated as the fraction of simulations greater than our observed frequency (Figure 2.2A (bottom); Supp Figure 2.1). We performed this analysis on the HCRegLoops-ALL sample sets (Supp Table 2.4) and, to ensure our bootstrapping approach does not lead to inflated p-values, we investigated the Q-Q plots of our samples. We observed that most p-values lie along the diagonal for non-significant p-values suggesting no inflation (Figure 2.2D). Summarizing this analysis across all of our samples, we found signals for ZNF460 paired with motifs such as ZNF135 and THRA across a majority of samples (Figure 2.2E). Self pairs of ZNF135 and ZNF460 were also overrepresented across multiple samples. Overall, this is an interesting finding which suggests an association for zinc finger proteins other than CTCF with chromatin loops and their anchor regions.

2.3.5 Investigating topological properties of enhancer-promoter interactions networks derived from HiChIP loops

In addition, for a subset of human H3K27ac HiChIP samples, we provide a network analysis with nodes being genomic regions (including regulatory elements such as enhancers and promoters) and edges representing connections by significant loops we detected. The network analysis was run for 240 human samples with 5kb FC loops of which 151 had at least one community (strongly connected set of nodes also referred to as network modules) detected. For these, we performed a two-level community detection for each chromosome where we first defined communities and then defined sub-communities for those with a large number of nodes. For both communities and sub-communities, we used CRank, an unsupervised technique for prioritization of network communities depending on their connectivity and topology parameters such as conductance, modularity and randomness [136]. Sub-communities are then visualized using CytoscapeJS to provide a dynamic interface and users can switch between sub-communities using an interactive table that allows filtering on various CRank variables (Figure 1.2C). The provided functionalities for interrogating and ranking network communities at multiple levels will allow downstream analysis such as hyperconnected cliques [137] and multi-enhancer hubs [138] that were previously studied in the context of oncogenes and immune cell development, respectively.

2.4 Conclusion

The Loop Catalog portal offers a powerful way to access HiChIP data while providing additional analyses that are highly valuable to the scientific community. One such analysis is the SGL analysis, initially developed for GWAS studies in four immune-related diseases. This tool enables researchers to query variants or genes of interest, including GWAS SNPs or prioritized subsets like fine-mapped SNPs. SGL analysis has also been extended to eQTLs, allowing for the exploration of how gene expression influences chromatin conformation. For instance, the SNP rs61839660, associated with T1D, overlaps the intron of *IL2RA*. While many researchers might consider this association sufficient, chromatin conformation data from the Loop Catalog reveals that rs61839660 also interacts with the promoters of *IL15RA* and *RBM17* in various adaptive immune cells. This example showcases the invaluable insights that can be gained when integrating the Loop Catalog with genetic and other epigenetic datasets. Moreover, HiChIP data is rich in information, enabling analyses such as motif and paired motif analysis, which have uncovered the high prevalence of zinc finger family motifs. Additionally, enhancer-promoter networks facilitated by HiChIP data offer a deeper understanding of interaction clusters. Overall, we have leveraged the foundational HiChIP data from Chapter 1 to develop integrative analyses that significantly broaden our understanding of chromatin conformation. These additional features will benefit the scientific community, and as shown in this chapter, the Loop Catalog facilitates researchers' efforts to achieve a system-level understanding of their biological questions.

2.5 Methods

2.5.1 Identifying SGLs in Immune-based Diseases

SGLs were identified utilizing fine-mapped GWAS SNPs from CAUSALdb for Type 1 Diabetes (T1D), Rheumatoid Arthritis (RA), Psoriasis (PS), and Atopic Dermatitis (AD) which included 7, 7, 3, and 1 individual studies, respectively (Supp Table 2.1). The fine-mapped data was downloaded and lifted over from hg19 to hg38 using the MyVariantInfo Python package [139]. We downloaded the GENCODE v30 transcriptome reference, filtered transcripts for type equal to "gene" and located coordinates of the transcription start site (TSS) [140]. For genes on the plus strand, the TSS would be assigned as a 1bp region at the start site and, for the minus strand, the 1bp region at the end site. Lastly, we extracted all HiChIP samples whose organ was classified as "Immune-associated". To integrate these datasets, loop anchors were intersected with fine-mapped GWAS SNPs and TSSs independently using bedtools pairtobed. Subsequently, loops were extracted as an SGL if at least one anchor contained a GWAS SNP and the opposing anchor contained a TSS.

2.5.2 Identifying SGLs with Immune-associated eQTL Studies

SGLs were identified utilizing eQTLs from the eQTL Catalog which included CD4 T-cells, CD8 T-cells, B cells, Natural Killer cells and monocytes (Supp Table 2.2). Similarly to GWAS-SGLs, we used GENCODE v30 to locate the TSS and extracted a subset of the GWAS-SGL HiChIP samples whose cell type matched the eQTL studies. To integrate these datasets, loops were intersected with pairs of SNP-gene pairs using bedtools pairtopair. Subsequently, loops were extracted as an SGL if at least one anchor contained an eQTL SNP and the opposing anchor contained a promoter.

2.5.3 Building a Consensus Gene List for T1D

From the MalaCards database (https://www.malacards.org/) a query was made using the term “type_1_diabetes_mellitus” and the list of associated genes was downloaded. For eDGAR (https://edgar.biocomp.unibo.it/gene_disease_db/), “diabetes mellitus, 1” was queried under the Main Tables tab and all gene symbols were extracted. OpenTargets hosts a disease based search with gene association scores and a query was made to the MONDO ID for T1D (MONDO_0005147). The corresponding genes were downloaded and filtered for an association score > 0.5. For the GWAS Catalog, a query was made using the previous MONDO ID and all associated genes were extracted. Table 1 of Klak et al [124] summarizes genes that have been associated with T1D and gene symbols were extracted from this resource.

2.5.4 Performing Motif Enrichment Analysis across Conserved Anchors

Motif enrichment analysis on the HCRegLoops-All, HCRegLoops-Immune, and HCRegLoops-Non-Immune sample sets was performed. Briefly, for each sample set, we identified highly conserved loop anchors by compiling loops across all samples in the set, extracting anchors, and filtering the anchors for those involved in at least one loop call in at least 90% of samples from the given sample set. We downloaded 727 known human motifs from the 2022 JASPAR CORE database [141]. Motif enrichment analysis was directly applied to the conserved anchor sites using MEME Suite SEA (Simple Enrichment Analysis) version 5.5.0 [120] using a match p-value threshold of 1e-6 and default values for all other parameters (Supp Table 2.3).

2.5.5 Identifying Pairs of Motifs Overlapping Loop Anchors

Briefly, for each sample in HCRegLoops-All, unique loop anchors were extracted and these were intersected with sample-specific corresponding ChIP-seq peaks using bedtools intersect with no slack. We selected the peak with the highest signal value to represent that anchor for cases in which multiple H3K27ac peaks overlapped one loop anchor. The resulting peak sets were deduplicated since a single peak may overlap multiple loop anchors. To determine the genomic coordinates of motifs from the 2022 JASPAR CORE database (n = 727 motifs) in these representative peak sets, MEME Suite FIMO (Find Individual Motif Occurrences) version 5.5.0 [142] was applied to the HCRegLoops-All sample set on a sample-by-sample basis using a match p-value threshold of 1e-6 and default values for all other parameters. For each sample, we intersected motif coordinates with loop coordinates using bedtools pairtobed and annotated loop anchors with the motifs falling within the anchor.

2.5.6 Performing Enrichment Analysis for Motif Pairs

Statistical analyses were performed via a bootstrapping method for HCRegLoops-All HiChIP samples (Supp Table 2.4, Supp Figure 2.1). Bootstrapping was performed using a block bootstrap with anchors as the unit analysis and anchors were randomly shuffled within their corresponding chromosomes. To get the null distribution, we shuffled the dataset through the location of unique instances of anchors. From there, we assigned a uniform probability of drawing a given anchor within each chromosome with replacement and used a total of 100,000 simulations. P-values were then determined for each sample by counting up the total number of simulated pairs that were greater than or equal to observed pairs in the original sample. Due to the large number of pairs ranging from tens to hundreds of thousands for each sample, we only focused on motif pairs with both motifs within the top 50 most frequently enriched motifs for that sample. A multiple testing correction using Benjamini-Hochberg was then utilized with these filtered pairs to obtain adjusted p-values.

2.5.7 Constructing Chromatin Interaction Networks using Loops

To construct a network from chromatin loops, each anchor was considered a node and each significant loop an edge. In addition, anchors were labeled as promoters by intersecting with TSS coordinates (slack of +/- 2.5kb) and allowing the promoter label to take priority over any other possible label. For H3K27ac HiChIP data, non-promoter nodes/anchors are labeled as enhancers when they overlap with ChIP-seq peaks (no slack). All other nodes that are not a promoter or an enhancer were designated as “other”. After obtaining the annotated anchors, we created a weighted undirected graph using the loops as edges and loop strength as edge weights calculated as the -log10(q-value) of FitHiChIP loop significance. To trim outliers, we set values larger than 20 to 20 and further scaled these values to between 0 and 1 for ease of visualization.

2.5.8 Detecting and Prioritizing Communities

Community detection was applied to the networks created using FC loops at 5kb. Two levels of community detection were applied, the first detected communities within the overall network created for each chromosome (high-level) followed by a second round that detected sub-communities of the communities reported in the first round (low-level). High-level communities were located by running the Louvain algorithm using default parameters as implemented by the CDlib Python package. Starting with each high-level community, subcommunities were called using the same Louvain parameters. CRank [136] was then applied at both levels to obtain a score that aggregates several properties related to the connectivity of a community into a single score for ranking.

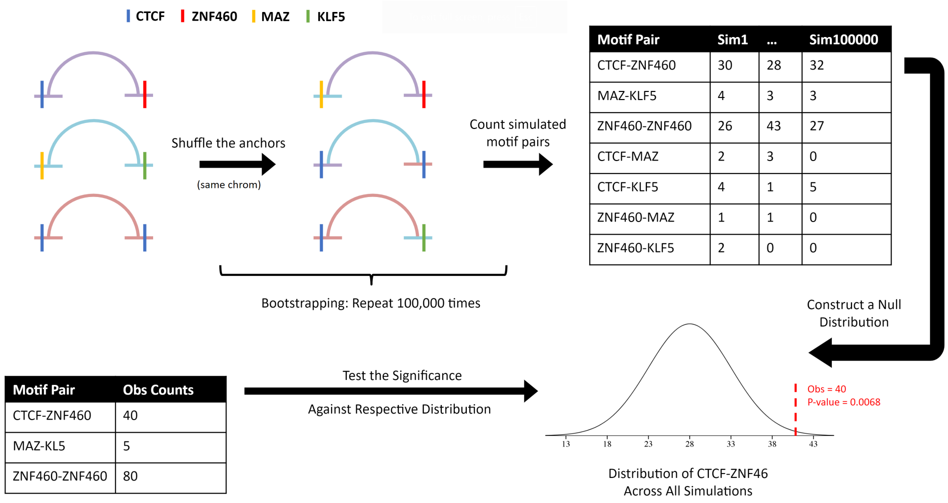
2.6 Availability of data and materials

Relevant code for this chapter has been released on Github at motif_pair_enrichment (https://github.com/ay-lab-team/motif_pair_enrichment) to perform enrichment analyses for motif pairs, Community-Detection-Using-Chromatin-Loops (https://github.com/ay-lab-team/Community-Detection-Using-Chromatin-Loops) to locate communities formed by chromatin loops, and T1D-Loop-Catalog (https://github.com/ay-lab-team/T1D-Loop-Catalog) to detect immune-associated SGLs. Versions of all software included in our pipeline are recorded (Supp Table 1.10). Fine-mapped GWAS SNPs were retrieved from CAUSALdb (http://www.mulinlab.org/causaldb).

2.7 Authors’ contributions

Joaquin Reyna: Code development, Web Development, Formal analysis, Writing–original draft. Kyra Fetter: Data collection, Formal analysis, Writing—original draft. Romeo Ignacio: Formal analysis, Writing—original draft. Cemil Can Ali Marandi: Formal analysis, Writing—original draft. Nikhil Rao: Data collection, Web development. Zichen Jiang: Data collection, Web development. Daniela Salgado Figueroa: Data collection. Sourya Bhattacharya: Conceptualization, Writing—review & editing. Ferhat Ay: Conceptualization, Writing—review & editing.

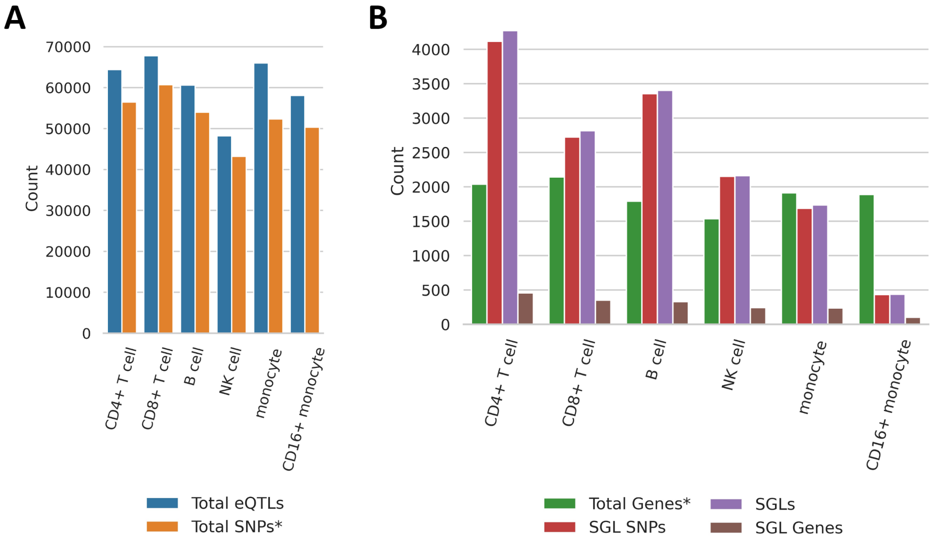
2.8 Supplementary Figures


CTCF
ZNF460
MAZ
KLF5
Shuffle the anchors
(same chrom)
Count simulated motif pairs
Bootstrapping: Repeat 100,000 times
Construct a Null Distribution
Test the Significance
Against Respective Distribution
Obs = 40
P-value = 0.0068
13
18
23
28
33
38
43
Distribution of CTCF-ZNF46
Across All Simulations

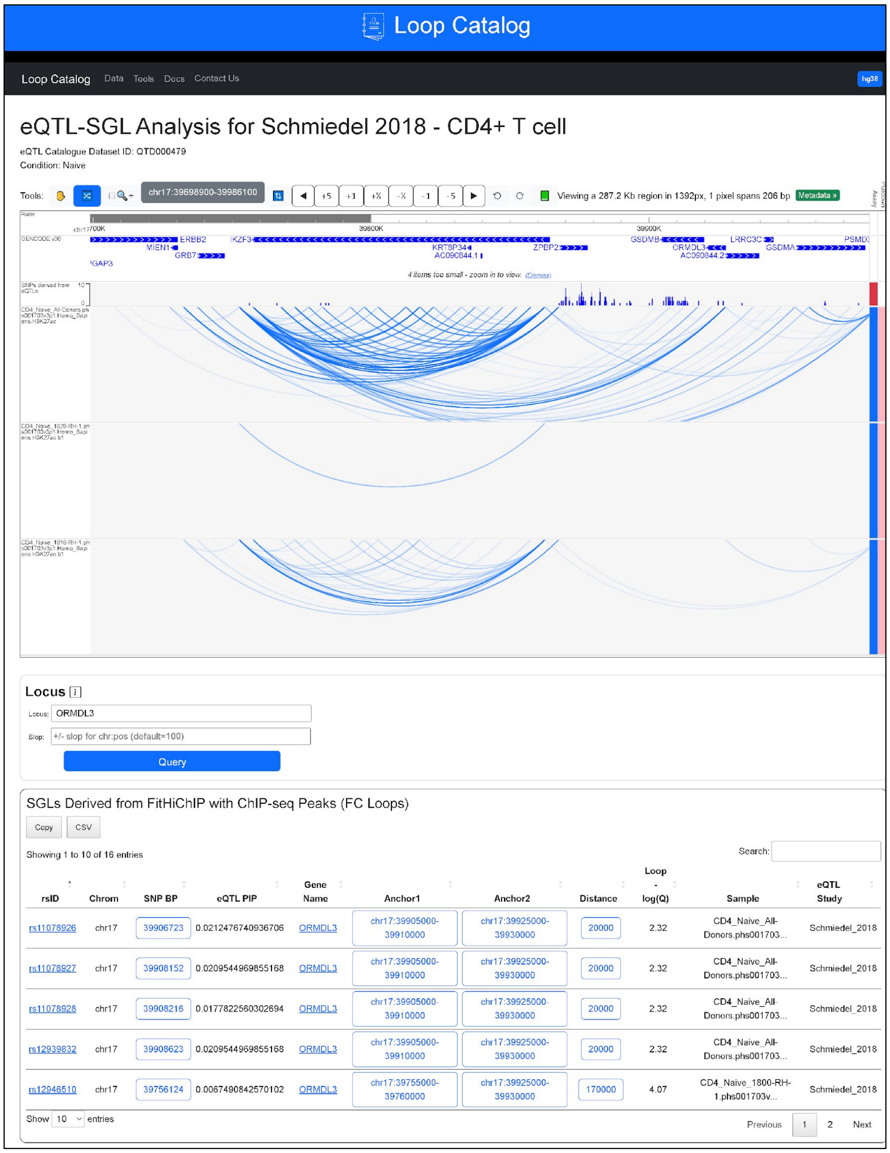
Motif Pair	Sim1	...	Sim100000
CTCF-ZNF460	30	28	32
MAZ-KLF5	4	3	3
ZNF460-ZNF460	26	43	27
CTCF-MAZ	2	3	0
CTCF-KLF5	4	1	5
ZNF460-MAZ	1	1	0
ZNF460-KLF5	2	0	0

Motif Pair	Obs Counts
CTCF-ZNF460	40
MAZ-KL5	5
ZNF460-ZNF460	80

Supp Figure 2.1. Schema of Bootstrap Analysis for Motif Pairs. Loops are depicted with an arc and horizontal lines are used to denote their anchors. Each loop has a distinct color to emphasize the true loop composition and overlapping motifs are depicted with horizontal lines and a separate color for each (top right). Observed motif pairs are tabulated (bottom left). To simulate a new set of loops, anchors are shuffled thereby bringing together new combinations of motifs (top middle). Motif pairs from these simulations are tabulated across 100,000 simulations (top right) to then build a null distribution (bottom right). The observed counts for a given motif are then evaluated against their respective null distribution (bottom).



Supp Figure 2.2. Summary of SGLs derived from HiChIP and eQTL intersections. **A**) Breakdown by cell type for the total number of eQTLs (blue) and SNPs (orange) within the original eQTL dataset. **B**) Breakdown by cell type for the total number of genes from the eQTL study (green) followed by SGL summaries for unique SGLs (red), SNPs (purple) and genes (brown).


Loop Catalog
Loop Catalog
Data
Tools
Docs
Contact Us
hg38
Tools:
chr17:39698900-39986100
◀
+5
+1
+½
-½
-1
-5
▶
Viewing a 287.2 Kb region in 1392px, 1 pixel spans 206 bp
Metadata »
Ruler
chr17700K
39800K
39900K
GENCODE v30
ERBB2
IKZF3
MIEN1
GRB7
PGAP3
KRT8P34
AC090844.1
ZPBP2
GSDMB
ORMDL3
AC090844.2
LRRC3C
GSDMA
PSMD3
4 items too small - zoom in to view. (Dismiss)
SNPs derived from eQTLs
10
0
CD4_Naive_All-Donors.phs001703v3p1.Homo_Sapiens.H3K27ac
CD4_Naive_1829-RH-1.phs001703v3p1.Homo_Sapiens.H3K27ac.b1
CD4_Naive_1816-RH-1.phs001703v3p1.Homo_Sapiens.H3K27ac.b1

eQTL-SGL Analysis for Schmiedel 2018 - CD4+ T cell

eQTL Catalogue Dataset ID: QTD000479

Condition: Naive

Locus

Locus: ORMDL3

Slop: +/- slop for chr:pos (default=100)

Query

SGLs Derived from FitHiChIP with ChIP-seq Peaks (FC Loops)

Copy CSV

Showing 1 to 10 of 16 entries

Search:

rsID	Chrom	SNP BP	eQTL PIP	Gene Name	Anchor1	Anchor2	Distance	Loop - log(Q)	Sample	eQTL Study
rs11078926	chr17	39906723	0.0212476740936706	ORMDL3	chr17:39905000-39910000	chr17:39925000-39930000	20000	2.32	CD4_Naive_All-Donors.phs001703...	Schmiedel_2018
rs11078927	chr17	39908152	0.0209544969855168	ORMDL3	chr17:39905000-39910000	chr17:39925000-39930000	20000	2.32	CD4_Naive_All-Donors.phs001703...	Schmiedel_2018
rs11078928	chr17	39908216	0.0177822560302694	ORMDL3	chr17:39905000-39910000	chr17:39925000-39930000	20000	2.32	CD4_Naive_All-Donors.phs001703...	Schmiedel_2018
rs12939832	chr17	39908623	0.0209544969855168	ORMDL3	chr17:39905000-39910000	chr17:39925000-39930000	20000	2.32	CD4_Naive_All-Donors.phs001703...	Schmiedel_2018
rs12946510	chr17	39756124	0.0067490842570102	ORMDL3	chr17:39755000-39760000	chr17:39925000-39930000	170000	4.07	CD4_Naive_1800-RH-1.phs001703v...	Schmiedel_2018

Show 10 entries

Previous 1 2 Next

Supp Figure 2.3. Example analysis for CD4 T cells using SGLs derived from their corresponding eQTLs. Depicted is the IKZF3/ORMDL3 locus with the top gray track containing SNPs derived from eQTLs followed by H3K27ac HiChIP loops for six CD4 T cell samples.

2.9 Supplementary Tables

Supp Table 2.1: GWAS SGL Statistics for RA, AD, T1D, and PS

Supp Table 2.2: eQTL SGL Statistics for 6 Immune Cell Types

Supp Table 2.3: 1D Conserved Anchor Motif Enrichment Analysis Statistics

Supp Table 2.4: Paired Motif Enrichment Analysis Statistics

Chapters 1 and 2, in full, has been submitted for publication of the material as it may appear in Genome Biology 2024, Joaquin Reyna, Kyra Fetter, Romeo Ignacio, Cemil Can Ali Marandi, Nikhil Rao, Zichen Jiang, Daniela Salgado Figueroa, Sourya Bhattacharyya, Ferhat Ay. The dissertation author was one of the primary investigators and authors of this paper.

Chapter 3

nipalsMCIA: Flexible Multi-Block Dimensionality Reduction in R via Non-linear Iterative Partial Least Squares

3.1 Abstract

With the increased reliance on multiomics data for bulk and single cell analyses, the availability of robust approaches to perform unsupervised analysis for clustering, visualization, and feature selection is imperative. Joint dimensionality reduction methods can be applied to multiomics datasets to derive a global sample embedding analogous to single-omic techniques such as Principal Components Analysis (PCA). We introduce `nipalsMCIA`, an MCIA implementation that solves the objective function using an extension to Non-linear Iterative Partial Least Squares (NIPALS), and shows significant speed-up over earlier implementations that rely on eigendecompositions for single cell multi-omics data. It also removes the dependence on an eigendecomposition for calculating the variance explained, and allows users to perform out-of-sample embedding for new data. `nipalsMCIA` provides users with a variety of pre-processing and parameter options, as well as ease of functionality for down-stream analysis of single-omic and global-embedding factors.

3.2 Background

Prior to the birth of high-throughput methods, the biology field has long relied on analyzing biological phenomena with limited data and assays per study. DNA sequencing, when first established in the 1970's, was a slow process that only sequenced about 10kb per day [143]. Fast forward to today and sequencing is available to study RNA [144], DNA binding proteins [145, 146], chromatin conformation [50], DNA accessibility [147, 148] and much more [149]. The same growth has been seen for other biological signals such as proteomics and cell frequencies via methods such mass spectrometry [150] and mass cytometry [151]. This surge in data generation has also results in the ability to ask more complex questions, demanding well-conceived integration strategies [149, 152, 153]. To pave the way, several methods known as joint dimensionality reduction (jDR) methods, have constructed mathematical formulation for deriving signals shared between blocks or derived from within a single block of data [11]. Canonical correlation analysis was an early pioneer of these methods that analyzes two blocks of data together, utilizing an optimization function that attempts to optimize for the covariance between blocks of data [154]. Naturally, there has been an extension of CCA and other methods to much larger blocks of data as our capacity to assay different phenomena has increased [11, 155].

The biology community needs a comprehensive and systems-wide understanding of mechanisms for various biological phenomena and sophisticated jDR methods can play a crucial role. The complexity of immune responses, cancer progression, human development and many other biological events has prompted the use of multiple assays with some advancements being attributed to the use of multiomics data [156–167]. MCIA is a member of the jDR family that extends unsupervised dimension reduction techniques such as Principal Components Analysis (PCA) and Non-negative Matrix Factorization (NMF) to datasets with multiple data *blocks* (alternatively called *views*) ([11, 168]). Such methods, also known as multi-block or multi-view analysis algorithms, are becoming increasingly important in the field of bioinformatics, where data is often collected simultaneously using multiple *omics* technologies such as transcriptomics, proteomics, epigenomics, metabolomics, etc [14].

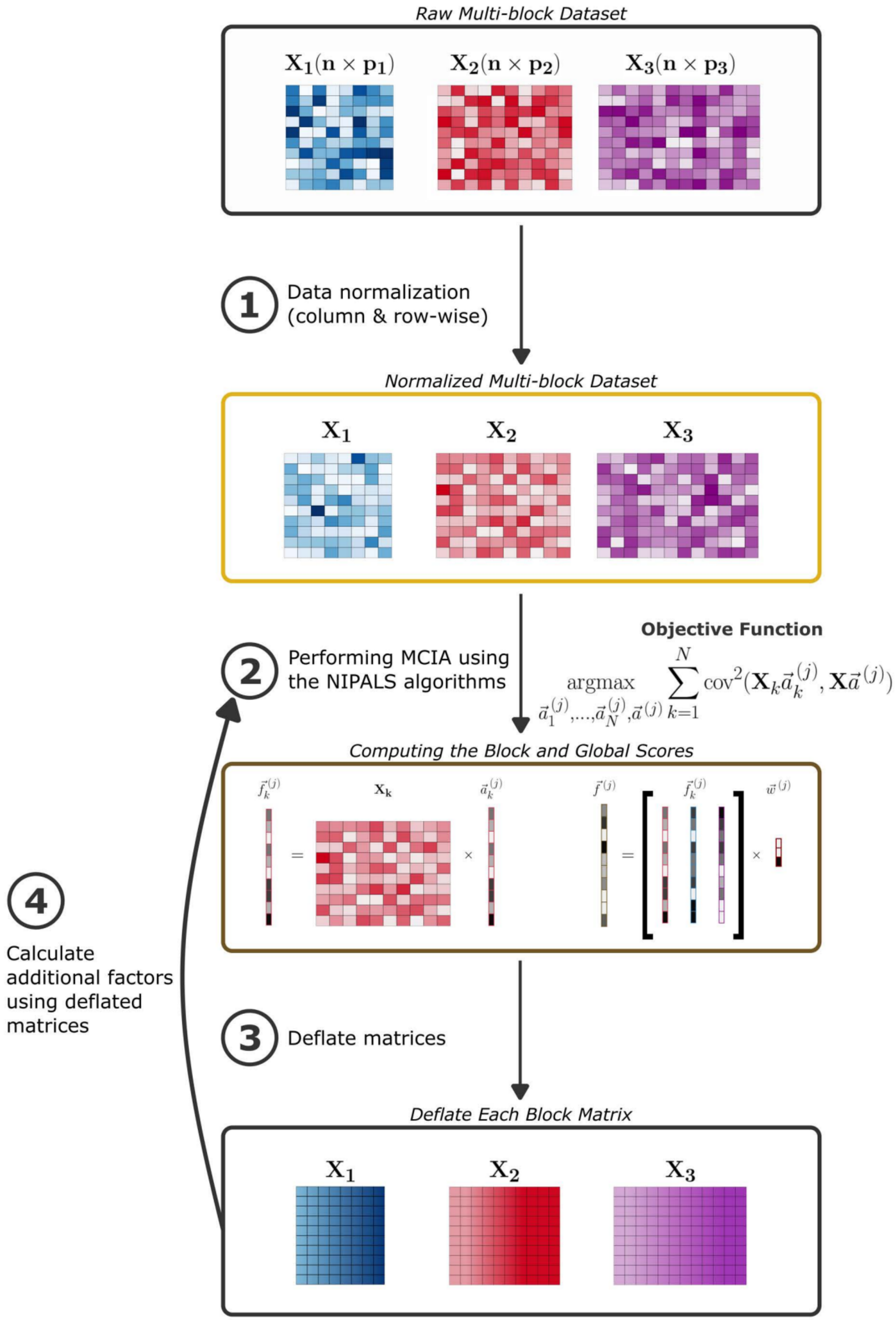
Here, we present a new implementation in R/Bioconductor of MCIA, `nipalsMCIA`, that uses an extension with proof of monotonic convergence of Non-linear Iterative Partial Least Squares (NIPALS) to solve the MCIA optimization problem [169]. This implementation shows significant speed-up over existing Singular Value Decomposition (SVD)-based approaches for MCIA [12, 170] on large datasets. Furthermore, `nipalsMCIA` offers users several options for pre-processing and deflation to customize algorithm performance, methodology to perform out-of-sample global embedding, and analysis and visualization capabilities for efficient results interpretation. We show application of `nipalsMCIA` to both bulk and single cell multi-omics data. The overall workflow that includes the optimization steps and analyses for `nipalsMCIA` is outlined in Figure 3.1.

3.3 Results

3.3.1 Package Design and Goals

The overall goal of nipalsMCIA is to make multiomics analysis accessible and interpretable. Previous attempts at making packages for MCIA existed in the form of Omicade and MOGSA however their implementations made it difficult to use beyond integration and data reduction. To promote the adoption of these methods within the bioinformatics community, `nipalsMCIA` was developed as an extensible software hosted on Bioconductor. This ensured `nipalsMCIA` was rigorously evaluated and revisions were suggested. This process resulted in a highly accessible version of `nipalsMCIA` that utilizes MultiAssayExperiment to load multiomics data, several pre-processing options to allow within- and whole-block normalizations, and a nipalsResults object especially tailored for downstream analyses. The latter feature includes various visualization methods to extract insights from global scores, global loadings, weights as well as their block level counterparts. To succinctly demonstrate `nipalsMCIA`'s use, we included three vignettes that cover: 1) the basics of jDR, 2) prediction of MCIA scores for novel datasets (not shown), and 3) application to single cell analysis. As will be shown, `nipalsMCIA` comes with an example dataset from the National Cancer Institute 60 tumor-cell line screen (NCI60 data) [13, 171]. It includes RNA-Seq, miRNA, and protein data from 21 cell lines that correspond to three cancer subtypes (brain, leukemia, and melanoma). Processed single cell data for is also available and sourced from 10x Genomics. It includes both gene expression and cell surface antibody data obtained from a CITE-seq experiment on approximately 5000 immune cells derived from blood [172].

Figure 3.1. A breakdown of the NIPALS algorithm for performing MCIA. Data blocks are normalized before scores and loadings are computed to satisfy the objective function. Higher-order results are then computed after the data has been deflated with the current scores or loadings.

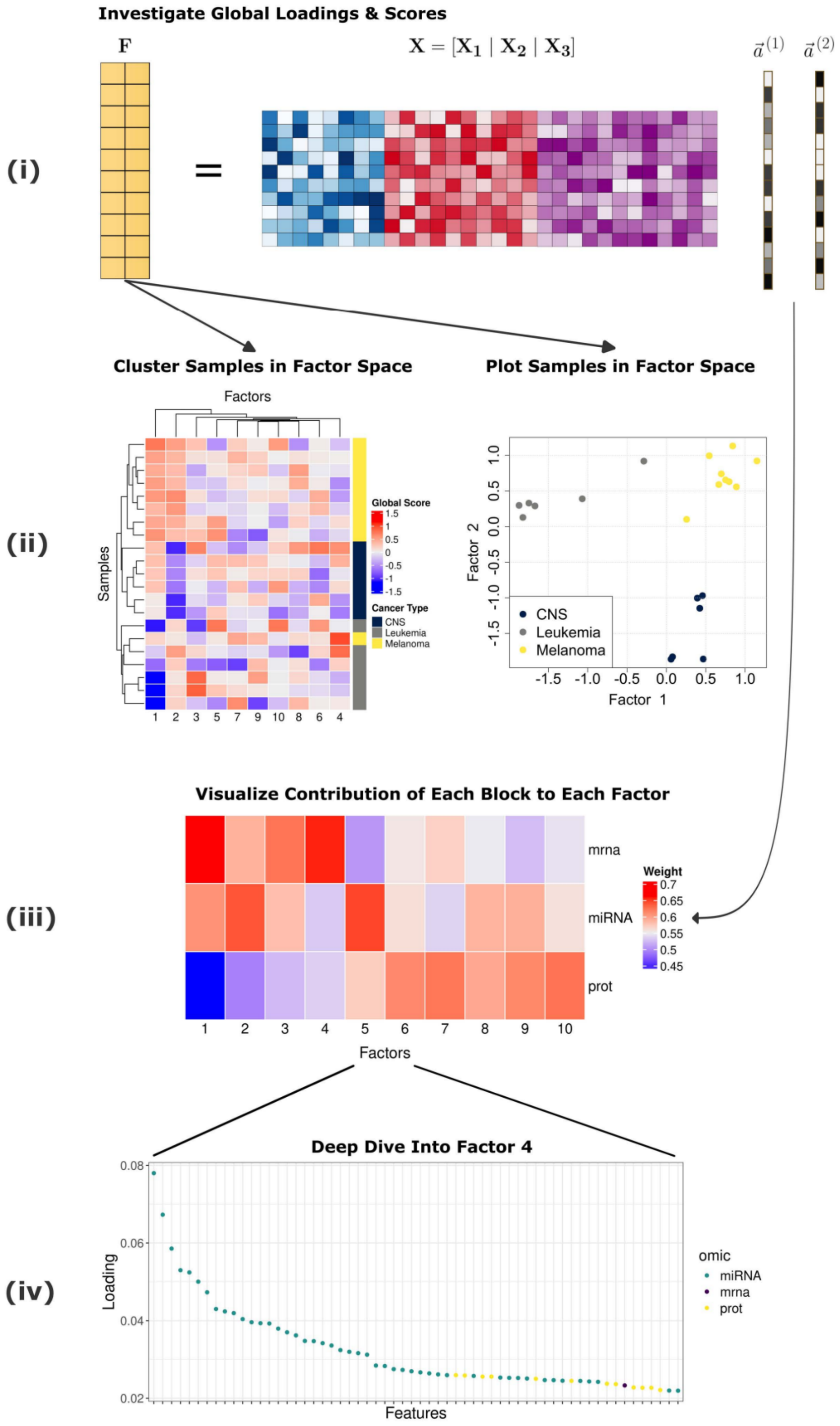

Raw Multi-block Dataset
$\mathbf{X_1(n \times p_1)}$
$\mathbf{X_2(n \times p_2)}$
$\mathbf{X_3(n \times p_3)}$
1 Data normalization (column & row-wise)
Normalized Multi-block Dataset
$\mathbf{X_1}$
$\mathbf{X_2}$
$\mathbf{X_3}$
2 Performing MCIA using the NIPALS algorithms
Objective Function
$\underset{\vec{a}_1^{(j)},\ldots,\vec{a}_N^{(j)},\vec{a}^{(j)}}{\operatorname{argmax}} \sum_{k=1}^{N} \operatorname{cov}^2(\mathbf{X}_k \vec{a}_k^{(j)}, \mathbf{X}\vec{a}^{(j)})$
Computing the Block and Global Scores
$\vec{f}_k^{(j)}$ = $\mathbf{X_k}$ × $\vec{a}_k^{(j)}$
$\vec{f}^{(j)}$ = $\vec{f}_k^{(j)}$ × $\vec{w}^{(j)}$
3 Deflate matrices
Deflate Each Block Matrix
$\mathbf{X_1}$
$\mathbf{X_2}$
$\mathbf{X_3}$
4 Calculate additional factors using deflated matrices

3.3.2 Analysis & Interpretation

The nipals_multiblock function is used to run MCIA in nipalsMCIA. The function outputs an object of the NipalsResult class, which includes the global scores and loadings, block scores and loadings, the global score eigenvalues, and the block score contributions vector for all orders up to the maximum specified via the num_PCs argument. The global scores represent the projection of the multi-block data in the reduced space (Figure 3.2i). MCIA is a completely unsupervised method, however, using the NCI60 dataset it can capture underlying signals that separate sample based on cancer type (Figure 3.2ii, left) (global_scores_heatmap function). In addition, factors can be plotted against each other to further dissect findings within this reduced space (Figure 3.2ii, right) (projection_plot function). This analysis is reminiscent of PC1 versus PC2 plots from PCA and, once again, highlights the ability of nipalsMCIA to derive informative factors for analysis goals such as sample separation. In addition, this last analysis can be extended by overlaying the corresponding block scores. These represent the projection of a sample given a single omic (square, triangle, circle) with lines connected to the global score (middle).

Each multiomics dataset is able to contribute a unique signal that may be captured in a single factor or across a variety of factors. These contributions to the global score can be easily described using the global weights and visualized using the block_weights_heatmap function (Figure 3.2iii). As shown in this figure, factor 1 signals are mainly derived from mRNA levels with some contribution from miRNA but very little contribution from proteins levels. This is in stark contrast to factor 5 where mRNA levels do not contribute but miRNA has a strong contributions with some additional contribution from protein levels. Deeper dives can also be performed for which we took factor 4 as an example. Unlike the other factors, contributions to factor 4 are concentrated most strongly within mRNA levels. The global loadings contains coefficients associated with the contribution of a single feature to the global score. To analyze these coefficients we made available the vis_load_ord function where features are on the X axis and loading values are on the Y axis. Using this plot we can see that out of the top 60 features, 47 are coming from mRNA, 12 from protein and only one from miRNA (Figure 3.2iv).

Figure 3.2. Scheme for interpreting the global loadings and scores performed on the three-block NCI60 data from the main text. **(i)** Global scores are calculated from the global data matrix and global loadings. **(ii)** Global scores represent low-dimensional embeddings of the data used to cluster samples via hierarchical clustering. Colors represent the three different cancer types associated with each sample **(iii)** Block contributions vectors plotted to visualize the weight of each block on each order of global score. **(iv)** The first global loadings vector is plotted to identify the top features for the first global score.



Investigate Global Loadings & Scores
(i)
F
=
$\mathbf{X} = [\mathbf{X_1} \mid \mathbf{X_2} \mid \mathbf{X_3}]$
$\vec{a}^{(1)}$
$\vec{a}^{(2)}$
Cluster Samples in Factor Space
Plot Samples in Factor Space
(ii)
Factors
Samples
1 2 3 5 7 9 10 8 6 4
Global Score
1.5
1
0.5
0
-0.5
-1
-1.5
Cancer Type
CNS
Leukemia
Melanoma
Factor 2
Factor 1
1.0
0.5
0.0
-0.5
-1.0
-1.5
-1.5 -1.0 -0.5 0.0 0.5 1.0
CNS
Leukemia
Melanoma
Visualize Contribution of Each Block to Each Factor
(iii)
mrna
miRNA
prot
1 2 3 4 5 6 7 8 9 10
Factors
Weight
0.7
0.65
0.6
0.55
0.5
0.45
Deep Dive Into Factor 4
(iv)
Loading
0.08
0.06
0.04
0.02
Features
omic
miRNA
mrna
prot

Given the prevalence of RNAseq in many multiomics datasets we also integrated a GSEA analysis via the `gsea_report` function. GSEA analyzes genes utilizing an associated scoring metric for which the global loadings provide an impact score. When interpreted, these scores can reveal important gene expression pathways that are relevant at this system-wide level. Results are provided in table 3.1 where we see that cell cycle and DNA replication processes are enriched, reciprocating the cancer context of the NCI60 dataset.

Table 3.1. GSEA results utilizing mRNA loading scores as input and the Reactome 6.2 gene set.

Pathway	**Adjusted P-value**
REACTOME CELL CYCLE	1.59e-20
REACTOME CELL CYCLE MITOTIC	4.50e-18
REACTOME GENERIC TRANSCRIPTION PATHWAY	5.81e-13
REACTOME DNA REPLICATION	1.92e-09
REACTOME MITOTIC M M G1 PHASES	2.24e-09
REACTOME PROCESSING OF CAPPED INTRON CONTAINING PR...	4.43e-09

3.3.3 Application to scRNA-seq Data

Given the growth of single cell technologies with powerful platforms such as 10X we also demonstrate the utility of `nipalsMCIA` to decompose the resulting high dimensional data into biologically relevant factors. We downloaded CITE-seq data from the 10X website that contains approximately 5000 peripheral blood mononuclear cells from a single donor as well as 33,538 genes and 32 cell surface markers [172] and applied nipalsMCIA using 10 factors. Factors from this analysis are able to separate immune cells from one another (Figure 3.3A). By hierarchically clustering using global factor scores we can further dissect that signals for macrophages are strongly captured in factor 1 whereas B cells signals are captured by factor 3. In addition, Natural Killers appear to be captured by a strong negative signal in factor 2 (Figure 3.3B).

3.3.4 Out-of-sample embedding

The loadings vectors generated by MCIA on a dataset X represent linear combinations of the original features of X. Therefore, after computing MCIA on a training dataset, one can use the associated loadings vectors to predict global embeddings for a test dataset of new observations of the same features. `nipalsMCIA` provides the `predict_gs` function for this task.

This can be valuable for testing the quality of the embedding, as well as embedding new data without rerunning the decomposition. We provide a vignette in the package showing how this can be done using the NCI60 data set, using 70% of the data to train the model, and then deriving global scores for the remaining 30%.

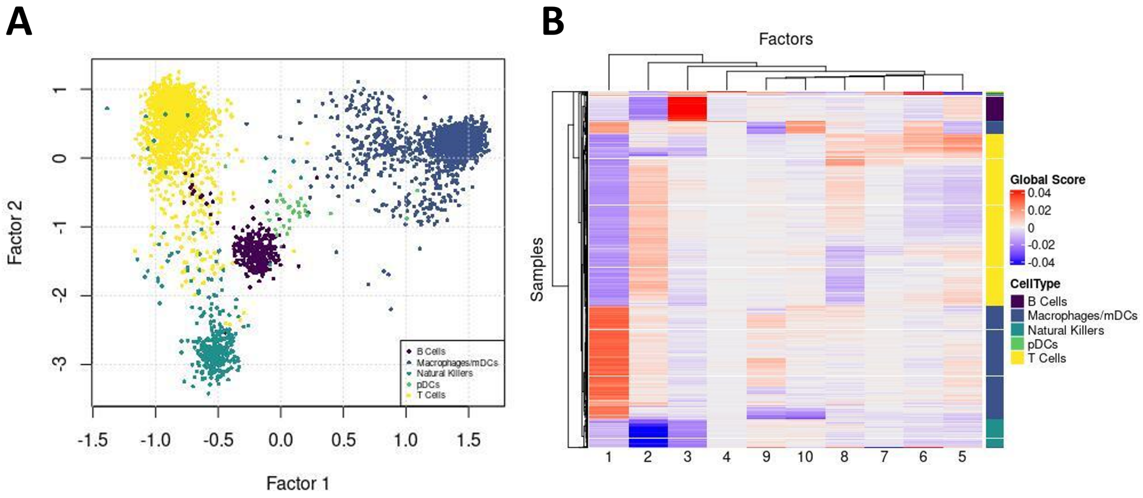


Figure 3.3. Analysis of global loading values derived from a CITE-seq dataset for 5,247 immune cells from the PMBC. A) Global factors scores for factor 1 and 2. **B**) Heatmap of global factor scores across all cells.

3.4 Discussion

The accessibility of next-generation sequencing and other high-throughput biological assays are resulting in an increase of multi-block (or multi-modal) datasets [15–18]. Analysis of these data are facilitated by the application of joint dimensionality reduction methods such as MCIA. `nipalsMCIA` is a comprehensive R package that implements MCIA in a highly efficient manner using the NIPALS algorithm. The package features various pre-processing and analysis options, is much faster for large input datasets compared with existing packages, supports the projection for out-of-sample scores, and offers visualization options for scores and top-magnitude loadings at each order. To showcase its applicability to emerging technologies, we applied `nipalsMCIA` to single cell data as part of our introductory vignettes and released our package on Bioconductor to further enhance its adoption across the wider bioinformatics community.

3.5 Supplementary Materials

The Supplementary Materials include additional information on the NIPALS algorithm implemented in `nipalsMCIA`, in-depth discussion of data pre-processing options, detailed overview of the calculations for variance explained and out-of-sample embedding, as well as a summary of the results (detailed more fully in the vignettes) corresponding to out-of-sample embedding of NCI60 data and application of `nipalsMCIA` to single cell data.

3.6 Acknowledgments

This work was supported in part by National Institutes of Health (NIH) grants R21AI176204 to S.H.K. and M.K., R35GM128938 to F.A., and U01AI150753.

Chapter 3, in full, has been submitted for publication of the material as it may appear in Bioinformatics 2024, Max Mattessich, Joaquin Reyna, Edel Aron, Ferhat Ay, Misha Kilmer, Steven H. Kleinstein, Anna Konstorum. The dissertation author was one of the primary investigators and authors of this paper.

Bibliography

1. Chang, M., Cai, Y., Gao, Z., Chen, X., Liu, B., Zhang, C., Yu, W., Cao, Q., Shen, Y., Yao, X., Chen, X. & Sun, H. Duchenne muscular dystrophy: pathogenesis and promising therapies. *Journal of Neurology* **270,** 3733–3749 (Aug. 2023).

2. Redondo, M. J. & Morgan, N. G. Heterogeneity and endotypes in type 1 diabetes mellitus. *Nature Reviews. Endocrinology* **19,** 542–554 (Sept. 2023).

3. Trynka, G., Westra, H.-J., Slowikowski, K., Hu, X., Xu, H., Stranger, B. E., Klein, R. J., Han, B. & Raychaudhuri, S. Disentangling the Effects of Colocalizing Genomic Annotations to Functionally Prioritize Non-coding Variants within Complex-Trait Loci. *The American Journal of Human Genetics* **97,** 139–152. (2024) (July 2015).

4. Pickrell, J. K. Joint Analysis of Functional Genomic Data and Genome-wide Association Studies of 18 Human Traits. *The American Journal of Human Genetics* **94,** 559–573. (2024) (Apr. 2014).

5. Finucane, H. K., Bulik-Sullivan, B., Gusev, A., Trynka, G., Reshef, Y., Loh, P.-R., Anttila, V., Xu, H., Zang, C., Farh, K., Ripke, S., Day, F. R., Purcell, S., Stahl, E., Lindstrom, S., Perry, J. R. B., Okada, Y., Raychaudhuri, S., Daly, M. J., Patterson, N., Neale, B. M. & Price, A. L. Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nature Genetics* **47,** 1228–1235. (2024) (Nov. 2015).

6. Farh, K. K.-H., Marson, A., Zhu, J., Kleinewietfeld, M., Housley, W. J., Beik, S., Shoresh, N., Whitton, H., Ryan, R. J. H., Shishkin, A. A., Hatan, M., Carrasco-Alfonso, M. J., Mayer, D., Luckey, C. J., Patsopoulos, N. A., De Jager, P. L., Kuchroo, V. K., Epstein, C. B., Daly, M. J., Hafler, D. A. & Bernstein, B. E. Genetic and epigenetic fine mapping of causal autoimmune disease variants. *Nature* **518,** 337–343. (2024) (Feb. 2015).

7. Mathelier, A., Shi, W. & Wasserman, W. W. Identification of altered cis-regulatory elements in human disease. *Trends in genetics: TIG* **31,** 67–76 (Feb. 2015).

8. Kang, M., Ko, E. & Mersha, T. B. A roadmap for multi-omics data integration using deep learning. *Briefings in Bioinformatics* **23,** bbab454 (Jan. 2022).

9. Babu, M. & Snyder, M. Multi-Omics Profiling for Health. *Molecular & cellular proteomics: MCP* **22,** 100561 (June 2023).

10. Chen, C., Wang, J., Pan, D., Wang, X., Xu, Y., Yan, J., Wang, L., Yang, X., Yang, M. & Liu, G.-P. Applications of multi-omics analysis in human diseases. *MedComm* **4,** e315 (Aug. 2023).

11. Cantini, L., Zakeri, P., Hernandez, C., Naldi, A., Thieffry, D., Remy, E. & Baudot, A. Benchmarking joint multi-omics dimensionality reduction approaches for the study of cancer. *Nature Communications* **12,** 124. (2023) (Jan. 2021).

12. Meng, C., Kuster, B., Culhane, A. C. & Gholami, A. M. A multivariate approach to the integration of multi-omics datasets. *BMC Bioinformatics* **15,** 162. (2023) (May 2014).

13. Meng, C., Zeleznik, O. A., Thallinger, G. G., Kuster, B., Gholami, A. M. & Culhane, A. C. Dimension reduction techniques for the integrative analysis of multi-omics data. *Briefings in Bioinformatics* **17,** 628–641. (2023) (July 2016).

14. Karczewski, K. J. & Snyder, M. P. Integrative omics for health and disease. *Nature Reviews. Genetics* **19,** 299–310 (May 2018).

15. Perez-Riverol, Y. Discovering and Linking Public ‘Omics’ Datasets using the Omics Discovery Index. *Nature Biotechnology* (May 2017).

16. Conesa, A. Making multi-omics data accessible to researchers. *Scientific data* (Oct. 2019).

17. Vasaikar, S. V. LinkedOmics: analyzing multi-omics data within and across 32 cancer types. *Nucleic Acids Research* (Jan. 2018).

18. Mohammadi-Shemirani, P. From ’Omics to Multi-omics Technologies: the Discovery of Novel Causal Mediators. *Current atherosclerosis reports* (Jan. 2023).

19. Fouché, A. & Zinovyev, A. Omics data integration in computational biology viewed through the prism of machine learning paradigms. *Frontiers in bioinformatics* (Aug. 2023).

20. HIPC-CHI Signatures Project Team & HIPC-I Consortium. Multicohort analysis reveals baseline transcriptional predictors of influenza vaccination responses. *Science Immunology* **2,** eaal4656. (2024) (Aug. 2017).

21. Diray-Arce, J., Miller, H. E. R., Henrich, E., Gerritsen, B., Mulè, M. P., Fourati, S., Gygi, J., Hagan, T., Tomalin, L., Rychkov, D., Kazmin, D., Chawla, D. G., Meng, H., Dunn, P., Campbell, J., Sarwal, M., Tsang, J. S., Levy, O., Pulendran, B., Sekaly, R., Floratos, A., Gottardo, R., Kleinstein, S. H. & Suárez-Fariñas, M. The Immune Signatures data resource, a compendium of systems vaccinology datasets. *Scientific Data* **9,** 635. (2024) (Oct. 2022).

22. Bernstein, B. E., Stamatoyannopoulos, J. A., Costello, J. F., Ren, B., Milosavljevic, A., Meissner, A., Kellis, M., Marra, M. A., Beaudet, A. L., Ecker, J. R., Farnham, P. J., Hirst, M., Lander, E. S., Mikkelsen, T. S. & Thomson, J. A. The NIH Roadmap Epigenomics Mapping Consortium. *Nature Biotechnology* **28,** 1045–1048 (Oct. 2010).

23. ENCODE Project Consortium, Moore, J. E., Purcaro, M. J., Pratt, H. E., Epstein, C. B., Shoresh, N., Adrian, J., Kawli, T., Davis, C. A., Dobin, A., Kaul, R., Halow, J., Van Nostrand, E. L., Freese, P., Gorkin, D. U., Shen, Y., He, Y., Mackiewicz, M., Pauli-Behn, F., Williams, B. A., Mortazavi, A., Keller, C. A., Zhang, X.-O., Elhajjajy, S. I., Huey, J., Dickel, D. E., Snetkova, V., Wei, X., Wang, X., Rivera-Mulia, J. C., Rozowsky, J., Zhang, J., Chhetri, S. B., Zhang, J., Victorsen, A., White, K. P., Visel, A., Yeo, G. W., Burge, C. B., Lécuyer, E., Gilbert, D. M., Dekker, J., Rinn, J., Mendenhall, E. M., Ecker, J. R., Kellis, M., Klein, R. J., Noble, W. S., Kundaje, A., Guigó, R., Farnham, P. J., Cherry, J. M., Myers, R. M., Ren, B., Graveley, B. R., Gerstein, M. B., Pennacchio, L. A., Snyder, M. P., Bernstein, B. E., Wold, B., Hardison, R. C., Gingeras, T. R., Stamatoyannopoulos, J. A. & Weng, Z. Author Correction: Expanded encyclopaedias of DNA elements in the human and mouse genomes. *Nature* **605,** E3 (May 2022).

24. Reiff, S. B., Schroeder, A. J., Kırlı, K., Cosolo, A., Bakker, C., Mercado, L., Lee, S., Veit, A. D., Balashov, A. K., Vitzthum, C., Ronchetti, W., Pitman, K. M., Johnson, J., Ehmsen, S. R., Kerpedjiev, P., Abdennur, N., Imakaev, M., Öztürk, S. U., Çamoğlu, U., Mirny, L. A., Gehlenborg, N., Alver, B. H. & Park, P. J. The 4D Nucleome Data Portal as a resource for searching and visualizing curated nucleomics data. *Nature Communications* **13,** 2365 (May 2022).

25. Shinde, P., Soldevila, F., Reyna, J., Aoki, M., Rasmussen, M., Willemsen, L., Kojima, M., Ha, B., Greenbaum, J. A., Overton, J. A., Guzman-Orozco, H., Nili, S., Orfield, S., Gygi, J. P., Antunes, R. d. S., Sette, A., Grant, B., Olsen, L. R., Konstorum, A., Guan, L., Ay, F., Kleinstein, S. H. & Peters, B. A multi-omics systems vaccinology resource to develop and test computational models of immunity. *Cell Reports Methods* **4.** (2024) (Mar. 2024).

26. Pulendran, B. Systems vaccinology: Probing humanity's diverse immune systems with vaccines. *Proceedings of the National Academy of Sciences* **111,** 12300–12306. (2024) (Aug. 2014).

27. Wang, J., Huang, D., Zhou, Y., Yao, H., Liu, H., Zhai, S., Wu, C., Zheng, Z., Zhao, K., Wang, Z., Yi, X., Zhang, S., Liu, X., Liu, Z., Chen, K., Yu, Y., Sham, P. C. & Li, M. J. CAUSALdb: a database for disease/trait causal variants identified using summary statistics of genome-wide association studies. *Nucleic Acids Research* **48,** D807–D816 (Jan. 2020).

28. Sollis, E., Mosaku, A., Abid, A., Buniello, A., Cerezo, M., Gil, L., Groza, T., Güneş, O., Hall, P., Hayhurst, J., Ibrahim, A., Ji, Y., John, S., Lewis, E., MacArthur, J. A. L., McMahon, A., Osumi-Sutherland, D., Panoutsopoulou, K., Pendlington, Z., Ramachandran, S., Stefancsik, R., Stewart, J., Whetzel, P., Wilson, R., Hindorff, L., Cunningham, F., Lambert, S. A., Inouye, M., Parkinson, H. & Harris, L. W. The NHGRI-EBI GWAS Catalog: knowledgebase and deposition resource. *Nucleic Acids Research* **51,** D977–D985 (Jan. 2023).

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

30. Park, P. J. ChIP–seq: advantages and challenges of a maturing technology. *Nature Reviews Genetics* **10,** 669–680. (2024) (Oct. 2009).

31. Zhao, Y. & Garcia, B. A. Comprehensive Catalog of Currently Documented Histone Modifications. *Cold Spring Harbor Perspectives in Biology* **7,** a025064. (2024) (Sept. 2015).

32. Cano-Gamez, E. & Trynka, G. From GWAS to Function: Using Functional Genomics to Identify the Mechanisms Underlying Complex Diseases. *Frontiers in Genetics* **11.** (2024) (May 2020).

33. Hou, L., Xiong, X., Park, Y., Boix, C., James, B., Sun, N., He, L., Patel, A., Zhang, Z., Molinie, B., Van Wittenberghe, N., Steelman, S., Nusbaum, C., Aguet, F., Ardlie, K. G. & Kellis, M. Multitissue H3K27ac profiling of GTEx samples links epigenomic variation to disease. *Nature Genetics* **55,** 1665–1676. (2024) (Oct. 2023).

34. Dong, S. & Boyle, A. P. Prioritization of regulatory variants with tissue-specific function in the non-coding regions of human genome. *Nucleic Acids Research* **50,** e6. (2024) (Oct. 2021).

35. Ma, M., Ru, Y., Chuang, L.-S., Hsu, N.-Y., Shi, L.-S., Hakenberg, J., Cheng, W.-Y., Uzilov, A., Ding, W., Glicksberg, B. S. & Chen, R. Disease-associated variants in different categories of disease located in distinct regulatory elements. *BMC genomics* **16 Suppl 8,** S3 (2015).

36. Qi, T., Song, L., Guo, Y., Chen, C. & Yang, J. From genetic associations to genes: methods, applications, and challenges. *Trends in Genetics* **0.** (2024) (May 2024).

37. GTEx Consortium. The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science (New York, N.Y.)* **369,** 1318–1330 (Sept. 2020).

38. Kerimov, N., Hayhurst, J. D., Peikova, K., Manning, J. R., Walter, P., Kolberg, L., Samoviča, M., Sakthivel, M. P., Kuzmin, I., Trevanion, S. J., Burdett, T., Jupp, S., Parkinson, H., Papatheodorou, I., Yates, A. D., Zerbino, D. R. & Alasoo, K. A compendium of uniformly processed human gene expression and splicing quantitative trait loci. *Nature Genetics* **53,** 1290–1299. (2024) (Sept. 2021).

39. Chandra, V., Bhattacharyya, S., Schmiedel, B. J., Madrigal, A., Gonzalez-Colin, C., Fotsing, S., Crinklaw, A., Seumois, G., Mohammadi, P., Kronenberg, M., Peters, B., Ay, F. & Vijayanand, P. Promoter-interacting expression quantitative trait loci are enriched for functional genetic variants. *Nature Genetics* **53,** 110–119. (2024) (Jan. 2021).

40. Burgess, D. J. Getting dynamic with eQTLs. *Nature Reviews Genetics* **20,** 500–501. (2024) (Sept. 2019).

41. Ochoa, D., Hercules, A., Carmona, M., Suveges, D., Baker, J., Malangone, C., Lopez, I., Miranda, A., Cruz-Castillo, C., Fumis, L., Bernal-Llinares, M., Tsukanov, K., Cornu, H., Tsirigos, K., Razuvayevskaya, O., Buniello, A., Schwartzentruber, J., Karim, M., Ariano, B., Martinez Osorio, R. E., Ferrer, J., Ge, X., Machlitt-Northen, S., Gonzalez-Uriarte,

A., Saha, S., Tirunagari, S., Mehta, C., Roldán-Romero, J. M., Horswell, S., Young, S., Ghoussaini, M., Hulcoop, D. G., Dunham, I. & McDonagh, E. M. The next-generation Open Targets Platform: reimagined, redesigned, rebuilt. *Nucleic Acids Research* **51,** D1353–D1359 (Jan. 2023).

42. Friman, E. T., Flyamer, I. M., Marenduzzo, D., Boyle, S. & Bickmore, W. A. Ultra-long-range interactions between active regulatory elements. *Genome Research* **33,** 1269–1283 (Aug. 2023).

43. Nurick, I., Shamir, R. & Elkon, R. Genomic meta-analysis of the interplay between 3D chromatin organization and gene expression programs under basal and stress conditions. *Epigenetics & Chromatin* **11,** 49 (Aug. 2018).

44. Ibrahim, D. M. & Mundlos, S. The role of 3D chromatin domains in gene regulation: a multi-facetted view on genome organization. *Current Opinion in Genetics & Development* **61,** 1–8 (Apr. 2020).

45. Bouwman, B. A. & de Laat, W. Getting the genome in shape: the formation of loops, domains and compartments. *Genome Biology* **16,** 154. (2024) (Aug. 2015).

46. Yu, M. & Ren, B. The Three-Dimensional Organization of Mammalian Genomes. *Annual Review of Cell and Developmental Biology* **33,** 265–289. (2024) (Oct. 2017).

47. Chen, L.-F. & Long, H. K. Topology regulatory elements: From shaping genome architecture to gene regulation. *Current Opinion in Structural Biology* **83,** 102723. (2024) (Dec. 2023).

48. Dekker, J., Marti-Renom, M. A. & Mirny, L. A. Exploring the three-dimensional organization of genomes: interpreting chromatin interaction data. *Nature Reviews Genetics* **14,** 390–403. (2024) (June 2013).

49. 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. & Aiden, E. L. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* **159,** 1665–1680 (Dec. 2014).

50. Lieberman-Aiden, E., van Berkum, N. L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., Amit, I., Lajoie, B. R., Sabo, P. J., Dorschner, M. O., Sandstrom, R., Bernstein, B.,

Bender, M. A., Groudine, M., Gnirke, A., Stamatoyannopoulos, J., Mirny, L. A., Lander, E. S. & Dekker, J. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science (New York, N.Y.)* **326,** 289–293 (Oct. 2009).

51. Witten, D. M. & Noble, W. S. On the assessment of statistical significance of three-dimensional colocalization of sets of genomic elements. *Nucleic Acids Research* **40,** 3849–3855. (2024) (May 2012).

52. Fraser, J., Ferrai, C., Chiariello, A. M., Schueler, M., Rito, T., Laudanno, G., Barbieri, M., Moore, B. L., Kraemer, D. C. A., Aitken, S., Xie, S. Q., Morris, K. J., Itoh, M., Kawaji, H., Jaeger, I., Hayashizaki, Y., Carninci, P., Forrest, A. R. R., FANTOM Consortium, Semple, C. A., Dostie, J., Pombo, A. & Nicodemi, M. Hierarchical folding and reorganization of chromosomes are linked to transcriptional changes in cellular differentiation. *Molecular Systems Biology* **11,** 852 (Dec. 2015).

53. Zheng, H. & Xie, W. The role of 3D genome organization in development and cell differentiation. *Nature Reviews Molecular Cell Biology* **20,** 535–550. (2024) (Sept. 2019).

54. Boltsis, I., Grosveld, F., Giraud, G. & Kolovos, P. Chromatin Conformation in Development and Disease. *Frontiers in Cell and Developmental Biology* **9,** 723859 (2021).

55. Chakraborty, A. & Ay, F. The role of 3D genome organization in disease: From compartments to single nucleotides. *Seminars in Cell & Developmental Biology* **90,** 104–113 (June 2019).

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

57. Hu, Y., Salgado Figueroa, D., Zhang, Z., Veselits, M., Bhattacharyya, S., Kashiwagi, M., Clark, M. R., Morgan, B. A., Ay, F. & Georgopoulos, K. Lineage-specific 3D genome organization is assembled at multiple scales by IKAROS. *Cell* **186,** 5269–5289.e22. (2024) (Nov. 2023).

58. Gazal, S., Weissbrod, O., Hormozdiari, F., Dey, K. K., Nasser, J., Jagadeesh, K. A., Weiner, D. J., Shi, H., Fulco, C. P., O'Connor, L. J., Pasaniuc, B., Engreitz, J. M. & Price, A. L. Combining SNP-to-gene linking strategies to identify disease genes and assess disease omnigenicity. *Nature Genetics* **54,** 827–836 (June 2022).

59. Orozco, G., Schoenfelder, S., Walker, N., Eyre, S. & Fraser, P. 3D genome organization links non-coding disease-associated variants to genes. *Frontiers in Cell and Developmental Biology* **10,** 995388 (2022).

60. Shi, C., Ray-Jones, H., Ding, J., Duffus, K., Fu, Y., Gaddi, V. P., Gough, O., Hankinson, J., Martin, P., McGovern, A., Yarwood, A., Gaffney, P., Eyre, S., Rattray, M., Warren, R. B. & Orozco, G. Chromatin Looping Links Target Genes with Genetic Risk Loci for Dermatological Traits. *The Journal of Investigative Dermatology* **141,** 1975–1984 (Aug. 2021).

61. Duan, A., Wang, H., Zhu, Y., Wang, Q., Zhang, J., Hou, Q., Xing, Y., Shi, J., Hou, J., Qin, Z., Chen, Z., Liu, Z. & Yang, J. Chromatin architecture reveals cell type-specific target genes for kidney disease risk variants. *BMC biology* **19,** 38 (Feb. 2021).

62. Wang, W., Song, F., Feng, X., Chu, X., Dai, H., Tian, J., Fang, X., Song, F., Liu, B., Li, L., Li, X., Zhao, Y., Zheng, H. & Chen, K. Functional Interrogation of Enhancer Connectome Prioritizes Candidate Target Genes at Ovarian Cancer Susceptibility Loci. *Frontiers in Genetics* **12,** 646179 (2021).

63. O'Mara, T. A., Spurdle, A. B., Glubb, D. M. & Endometrial Cancer Association Consortium. Analysis of Promoter-Associated Chromatin Interactions Reveals Biologically Relevant Candidate Target Genes at Endometrial Cancer Risk Loci. *Cancers* **11,** 1440 (Sept. 2019).

64. Glubb, D. M., Thompson, D. J., Aben, K. K., Alsulimani, A., Amant, F., Annibali, D., Attia, J., Barricarte, A., Beckmann, M. W., Berchuck, A., Bermisheva, M., Bernardini, M. Q., Bischof, K., Bjorge, L., Bodelon, C., Brand, A. H., Brenton, J. D., Brinton, L. A., Bruinsma, F., Buchanan, D. D., Burghaus, S., Butzow, R., Cai, H., Carney, M. E., Chanock, S. J., Chen, C., Chen, X. Q., Chen, Z., Cook, L. S., Cunningham, J. M., De Vivo, I., deFazio, A., Doherty, J. A., Dörk, T., du Bois, A., Dunning, A. M., Dürst, M., Edwards, T., Edwards, R. P., Ekici, A. B., Ewing, A., Fasching, P. A., Ferguson, S., Flanagan, J. M., Fostira, F., Fountzilas, G., Friedenreich, C. M., Gao, B., Gaudet, M. M., Gawełko, J., Gentry-Maharaj, A., Giles, G. G., Glasspool, R., Goodman, M. T., Gronwald, J., Harris, H. R., Harter, P., Hein, A., Heitz, F., Hildebrandt, M. A., Hillemanns, P., Høgdall, E., Høgdall, C. K., Holliday, E. G., Huntsman, D. G., Huzarski, T., Jakubowska, A., Jensen, A., Jones, M. E., Karlan, B. Y., Karnezis, A., Kelley, J. L., Khusnutdinova, E., Killeen, J. L., Kjaer, S. K., Klapdor, R., Köbel, M., Konopka, B., Konstantopoulou, I., Kopperud, R. K., Koti, M., Kraft, P., Kupryjanczyk, J., Lambrechts, D., Larson, M. C., Le Marchand, L., Lele, S., Lester, J., Li, A. J., Liang, D., Liebrich, C., Lipworth, L., Lissowska, J., Lu, L., Lu, K. H., Macciotta, A., Mattiello, A., May, T., McAlpine, J. N., *et al.* Cross-

Cancer Genome-Wide Association Study of Endometrial Cancer and Epithelial Ovarian Cancer Identifies Genetic Risk Regions Associated with Risk of Both Cancers. *Cancer Epidemiology, Biomarkers & Prevention* **30,** 217–228. (2024) (Jan. 2021).

65. Tan, W. L. W., Anene-Nzelu, C. G., Wong, E., Lee, C. J. M., Tan, H. S., Tang, S. J., Perrin, A., Wu, K. X., Zheng, W., Ashburn, R. J., Pan, B., Lee, M. Y., Autio, M. I., Morley, M. P., Tam, W. L., Cheung, C., Margulies, K. B., Chen, L., Cappola, T. P., Loh, M., Chambers, J., Prabhakar, S., Foo, R. S. & CHARGE-Heart Failure Working Group, CHARGE-EchoGen Consortium. Epigenomes of Human Hearts Reveal New Genetic Variants Relevant for Cardiac Disease and Phenotype. *Circulation Research* **127,** 761–777. (2024) (Aug. 2020).

66. Mumbach, M. R., Satpathy, A. T., Boyle, E. A., Dai, C., Gowen, B. G., Cho, S. W., Nguyen, M. L., Rubin, A. J., Granja, J. M., Kazane, K. R., Wei, Y., Nguyen, T., Greenside, P. G., Corces, M. R., Tycko, J., Simeonov, D. R., Suliman, N., Li, R., Xu, J., Flynn, R. A., Kundaje, A., Khavari, P. A., Marson, A., Corn, J. E., Quertermous, T., Greenleaf, W. J. & Chang, H. Y. Enhancer connectome in primary human cells identifies target genes of disease-associated DNA elements. *Nature Genetics* **49,** 1602–1612 (Nov. 2017).

67. Ray-Jones, H., Duffus, K., McGovern, A., Martin, P., Shi, C., Hankinson, J., Gough, O., Yarwood, A., Morris, A. P., Adamson, A., Taylor, C., Ding, J., Gaddi, V. P., Fu, Y., Gaffney, P., Orozco, G., Warren, R. B. & Eyre, S. Mapping DNA interaction landscapes in psoriasis susceptibility loci highlights KLF4 as a target gene in 9q31. *BMC Biology* **18,** 47. (2024) (May 2020).

68. Piekos, S. N., Gaddam, S., Bhardwaj, P., Radhakrishnan, P., Guha, R. V. & Oro, A. E. Biomedical Data Commons (BMDC) prioritizes B-lymphocyte non-coding genetic variants in Type 1 Diabetes. *PLOS Computational Biology* **17,** e1009382. (2024) (Sept. 2021).

69. Tarbell, E., Jiang, K., Hennon, T. R., Holmes, L., Williams, S., Fu, Y., Gaffney, P. M., Liu, T. & Jarvis, J. N. CD4+ T cells from children with active juvenile idiopathic arthritis show altered chromatin features associated with transcriptional abnormalities. *Scientific Reports* **11,** 4011. (2024) (Feb. 2021).

70. Pelikan, R. C., Kelly, J. A., Fu, Y., Lareau, C. A., Tessneer, K. L., Wiley, G. B., Wiley, M. M., Glenn, S. B., Harley, J. B., Guthridge, J. M., James, J. A., Aryee, M. J., Montgomery, C. & Gaffney, P. M. Enhancer histone-QTLs are enriched on autoimmune risk haplotypes and influence gene expression within chromatin networks. *Nature Communications* **9,** 2905. (2024) (July 2018).

71. Plaça, D. R., Fonseca, D. L. M., Marques, A. H. C., Zaki Pour, S., Usuda, J. N., Baiocchi, G. C., Prado, C. A. d. S., Salgado, R. C., Filgueiras, I. S., Freire, P. P., Rocha, V., Camara, N. O. S., Catar, R., Moll, G., Jurisica, I., Calich, V. L. G., Giil, L. M., Rivino, L., Ochs, H. D., Cabral-Miranda, G., Schimke, L. F. & Cabral-Marques, O. Immunological signatures unveiled by integrative systems vaccinology characterization of dengue vaccination trials and natural infection. *Frontiers in Immunology* **15,** 1282754 (2024).

72. Tsang, J. S., Schwartzberg, P. L., Kotliarov, Y., Biancotto, A., Xie, Z., Germain, R. N., Wang, E., Olnes, M. J., Narayanan, M., Golding, H., Moir, S., Dickler, H. B., Perl, S., Cheung, F., Baylor HIPC Center & CHI Consortium. Global analyses of human immune variation reveal baseline predictors of postvaccination responses. *Cell* **157,** 499–513 (Apr. 2014).

73. Clarke, J. Assessing vaccine responses: you've got to have a system. *Nature Research.* (2024) (Sept. 2020).

74. Cortese, M., Sherman, A. C., Rouphael, N. G. & Pulendran, B. Systems Biological Analysis of Immune Response to Influenza Vaccination. *Cold Spring Harbor Perspectives in Medicine* **11,** a038596. (2024) (June 2021).

75. Chelliah, V., Lazarou, G., Bhatnagar, S., Gibbs, J. P., Nijsen, M., Ray, A., Stoll, B., Thompson, R. A., Gulati, A., Soukharev, S., Yamada, A., Weddell, J., Sayama, H., Oishi, M., Wittemer-Rump, S., Patel, C., Niederalt, C., Burghaus, R., Scheerans, C., Lippert, J., Kabilan, S., Kareva, I., Belousova, N., Rolfe, A., Zutshi, A., Chenel, M., Venezia, F., Fouliard, S., Oberwittler, H., Scholer-Dahirel, A., Lelievre, H., Bottino, D., Collins, S. C., Nguyen, H. Q., Wang, H., Yoneyama, T., Zhu, A. Z. X., van der Graaf, P. H. & Kierzek, A. M. Quantitative Systems Pharmacology Approaches for Immuno-Oncology: Adding Virtual Patients to the Development Paradigm. *Clinical Pharmacology and Therapeutics* **109,** 605–618 (Mar. 2021).

76. Butcher, E. C. Can cell systems biology rescue drug discovery? *Nature Reviews Drug Discovery* **4,** 461–467. (2024) (June 2005).

77. Butcher, E. C., Berg, E. L. & Kunkel, E. J. Systems biology in drug discovery. *Nature Biotechnology* **22,** 1253–1259. (2024) (Oct. 2004).

78. Benson, N. Network-based discovery through mechanistic systems biology. Implications for applications – SMEs and drug discovery: where the action is. *Drug Discovery Today:*

Technologies. SI: Network-based discovery through systems biology **15,** 41–48. (2024) (Aug. 2015).

79. Proctor, L. M., Creasy, H. H., Fettweis, J. M., Lloyd-Price, J., Mahurkar, A., Zhou, W., Buck, G. A., Snyder, M. P., Strauss, J. F., Weinstock, G. M., White, O., Huttenhower, C. & The Integrative HMP (iHMP) Research Network Consortium. The Integrative Human Microbiome Project. *Nature* **569,** 641–648. (2024) (May 2019).

80. Zhou, W., Sailani, M. R., Contrepois, K., Zhou, Y., Ahadi, S., Leopold, S. R., Zhang, M. J., Rao, V., Avina, M., Mishra, T., Johnson, J., Lee-McMullen, B., Chen, S., Metwally, A. A., Tran, T. D. B., Nguyen, H., Zhou, X., Albright, B., Hong, B.-Y., Petersen, L., Bautista, E., Hanson, B., Chen, L., Spakowicz, D., Bahmani, A., Salins, D., Leopold, B., Ashland, M., Dagan-Rosenfeld, O., Rego, S., Limcaoco, P., Colbert, E., Allister, C., Perelman, D., Craig, C., Wei, E., Chaib, H., Hornburg, D., Dunn, J., Liang, L., Rose, S. M. S.-F., Kukurba, K., Piening, B., Rost, H., Tse, D., McLaughlin, T., Sodergren, E., Weinstock, G. M. & Snyder, M. Longitudinal multi-omics of host–microbe dynamics in prediabetes. *Nature* **569,** 663–671. (2024) (May 2019).

81. Athieniti, E. & Spyrou, G. M. A guide to multi-omics data collection and integration for translational medicine. *Computational and Structural Biotechnology Journal* **21,** 134–149. (2024) (Jan. 2023).

82. Wang, S., Luo, Z., Liu, W., Hu, T., Zhao, Z., Rosenfeld, M. G. & Song, X. The 3D genome and its impacts on human health and disease. *Life Medicine* **2,** lnad012. (2024) (Apr. 2023).

83. Chandra, V., Bhattacharyya, S., Schmiedel, B. J., Madrigal, A., Gonzalez-Colin, C., Fotsing, S., Crinklaw, A., Seumois, G., Mohammadi, P., Kronenberg, M., Peters, B., Ay, F. & Vijayanand, P. Promoter-interacting expression quantitative trait loci are enriched for functional genetic variants. *Nature Genetics* **53,** 110–119 (Jan. 2021).

84. Levo, M., Raimundo, J., Bing, X. Y., Sisco, Z., Batut, P. J., Ryabichko, S., Gregor, T. & Levine, M. S. Transcriptional coupling of distant regulatory genes in living embryos. *Nature* **605,** 754–760. (2024) (May 2022).

85. Long, H. K., Osterwalder, M., Welsh, I. C., Hansen, K., Davies, J. O., Liu, Y. E., Koska, M., Adams, A. T., Aho, R., Arora, N., Ikeda, K., Williams, R. M., Sauka-Spengler, T., Porteus, M. H., Mohun, T., Dickel, D. E., Swigut, T., Hughes, J. R., Higgs, D. R., Visel, A.,

Selleri, L. & Wysocka, J. Loss of Extreme Long-Range Enhancers in Human Neural Crest Drives a Craniofacial Disorder. *Cell Stem Cell* **27,** 765–783.e14. (2024) (Nov. 2020).

86. Brandão, H. B., Gabriele, M. & Hansen, A. S. Tracking and interpreting long-range chromatin interactions with super-resolution live-cell imaging. *Current Opinion in Cell Biology* **70,** 18–26. (2024) (June 2021).

87. Ito, S., Das, N. D., Umehara, T. & Koseki, H. Factors and Mechanisms That Influence Chromatin-Mediated Enhancer-Promoter Interactions and Transcriptional Regulation. *Cancers* **14,** 5404 (Nov. 2022).

88. Popay, T. M. & Dixon, J. R. Coming full circle: On the origin and evolution of the looping model for enhancer-promoter communication. *The Journal of Biological Chemistry* **298,** 102117 (Aug. 2022).

89. Portillo-Ledesma, S., Li, Z. & Schlick, T. Genome modeling: From chromatin fibers to genes. *Current Opinion in Structural Biology* **78,** 102506 (Feb. 2023).

90. Qiu, Y., Feng, D., Jiang, W., Zhang, T., Lu, Q. & Zhao, M. 3D genome organization and epigenetic regulation in autoimmune diseases. *Frontiers in Immunology* **14,** 1196123 (2023).

91. Schaeffer, M. & Nollmann, M. Contributions of 3D chromatin structure to cell-type-specific gene regulation. *Current Opinion in Genetics & Development* **79,** 102032 (Apr. 2023).

92. Xu, H., Zhang, S., Yi, X., Plewczynski, D. & Li, M. J. Exploring 3D chromatin contacts in gene regulation: The evolution of approaches for the identification of functional enhancer-promoter interaction. *Computational and Structural Biotechnology Journal* **18,** 558–570 (2020).

93. Davies, J. O. J., Telenius, J. M., McGowan, S. J., Roberts, N. A., Taylor, S., Higgs, D. R. & Hughes, J. R. Multiplexed analysis of chromosome conformation at vastly improved sensitivity. *Nature Methods* **13,** 74–80 (Jan. 2016).

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

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

96. Liu, S., Cao, Y., Cui, K., Tang, Q. & Zhao, K. Hi-TrAC reveals division of labor of transcription factors in organizing chromatin loops. *Nature Communications* **13,** 6679 (Nov. 2022).

97. Wei, X., Xiang, Y., Peters, D. T., Marius, C., Sun, T., Shan, R., Ou, J., Lin, X., Yue, F., Li, W., Southerland, K. W. & Diao, Y. HiCAR is a robust and sensitive method to analyze open-chromatin-associated genome organization. *Molecular Cell* **82,** 1225–1238.e6 (Mar. 2022).

98. Yu, M., Juric, I., Abnousi, A., Hu, M. & Ren, B. Proximity Ligation-Assisted ChIP-Seq (PLAC-Seq). *Methods in Molecular Biology (Clifton, N.J.)* **2351,** 181–199 (2021).

99. Giambartolomei, C., Seo, J.-H., Schwarz, T., Freund, M. K., Johnson, R. D., Spisak, S., Baca, S. C., Gusev, A., Mancuso, N., Pasaniuc, B. & Freedman, M. L. H3K27ac HiChIP in prostate cell lines identifies risk genes for prostate cancer susceptibility. *American Journal of Human Genetics* **108,** 2284–2300 (Dec. 2021).

100. Zeng, W., Liu, Q., Yin, Q., Jiang, R. & Wong, W. H. HiChIPdb: a comprehensive database of HiChIP regulatory interactions. *Nucleic Acids Research* **51,** D159–D166 (Jan. 2023).

101. Zhou, Q., Cheng, S., Zheng, S., Wang, Z., Guan, P., Zhu, Z., Huang, X., Zhou, C. & Li, G. ChromLoops: a comprehensive database for specific protein-mediated chromatin loops in diverse organisms. *Nucleic Acids Research* **51,** D57–D69 (Jan. 2023).

102. Wang, J. & Nakato, R. CohesinDB: a comprehensive database for decoding cohesin-related epigenomes, 3D genomes and transcriptomes in human cells. *Nucleic Acids Research* **51,** D70–D79 (Jan. 2023).

103. Roayaei Ardakany, A., Gezer, H. T., Lonardi, S. & Ay, F. Mustache: multi-scale detection of chromatin loops from Hi-C and Micro-C maps using scale-space representation. *Genome Biology* **21,** 256 (Sept. 2020).

104. Schmiedel, B. J., Rocha, J., Gonzalez-Colin, C., Bhattacharyya, S., Madrigal, A., Ottensmeier, C. H., Ay, F., Chandra, V. & Vijayanand, P. COVID-19 genetic risk variants are associated with expression of multiple genes in diverse immune cell types. *Nature Communications* **12,** 6760 (Nov. 2021).

105. Durand, N. C., Shamim, M. S., Machol, I., Rao, S. S. P., Huntley, M. H., Lander, E. S. & Aiden, E. L. Juicer Provides a One-Click System for Analyzing Loop-Resolution Hi-C Experiments. *Cell Systems* **3,** 95–98 (July 2016).

106. Barrett, T., Wilhite, S. E., Ledoux, P., Evangelista, C., Kim, I. F., Tomashevsky, M., Marshall, K. A., Phillippy, K. H., Sherman, P. M., Holko, M., Yefanov, A., Lee, H., Zhang, N., Robertson, C. L., Serova, N., Davis, S. & Soboleva, A. NCBI GEO: archive for functional genomics data sets–update. *Nucleic Acids Research* **41,** D991–995 (Jan. 2013).

107. Buchmann, J. P. & Holmes, E. C. Entrezpy: a Python library to dynamically interact with the NCBI Entrez databases. *Bioinformatics (Oxford, England)* **35,** 4511–4514 (Nov. 2019).

108. Taylor, L. J., Abbas, A. & Bushman, F. D. grabseqs: simple downloading of reads and metadata from multiple next-generation sequencing data repositories. *Bioinformatics (Oxford, England)* **36,** 3607–3609 (June 2020).

109. Ewels, P. *SRA Explorer* Apr. 2024.

110. Sayers, E. W., Bolton, E. E., Brister, J. R., Canese, K., Chan, J., Comeau, D. C., Connor, R., Funk, K., Kelly, C., Kim, S., Madej, T., Marchler-Bauer, A., Lanczycki, C., Lathrop, S., Lu, Z., Thibaud-Nissen, F., Murphy, T., Phan, L., Skripchenko, Y., Tse, T., Wang, J., Williams, R., Trawick, B. W., Pruitt, K. D. & Sherry, S. T. Database resources of the national center for biotechnology information. *Nucleic Acids Research* **50,** D20–D26 (Jan. 2022).

111. Servant, N., Varoquaux, N., Lajoie, B. R., Viara, E., Chen, C.-J., Vert, J.-P., Heard, E., Dekker, J. & Barillot, E. HiC-Pro: an optimized and flexible pipeline for Hi-C data processing. *Genome Biology* **16,** 259 (Dec. 2015).

112. Langmead, B., Wilks, C., Antonescu, V. & Charles, R. Scaling read aligners to hundreds of threads on general-purpose processors. *Bioinformatics (Oxford, England)* **35,** 421–432 (Feb. 2019).

113. Zhang, Y., Liu, T., Meyer, C. A., Eeckhoute, J., Johnson, D. S., Bernstein, B. E., Nusbaum, C., Myers, R. M., Brown, M., Li, W. & Liu, X. S. Model-based analysis of ChIP-Seq (MACS). *Genome Biology* **9,** R137 (2008).

114. Bhattacharyya, S., Chandra, V., Vijayanand, P. & Ay, F. Identification of significant chromatin contacts from HiChIP data by FitHiChIP. *Nature Communications* **10,** 4221 (Sept. 2019).

115. Fu, Y., Tessneer, K. L., Li, C. & Gaffney, P. M. From association to mechanism in complex disease genetics: the role of the 3D genome. *Arthritis Research & Therapy* **20,** 216 (Sept. 2018).

116. Zhong, W., Liu, W., Chen, J., Sun, Q., Hu, M. & Li, Y. Understanding the function of regulatory DNA interactions in the interpretation of non-coding GWAS variants. *Frontiers in Cell and Developmental Biology* **10,** 957292 (2022).

117. Dekker, J., Belmont, A. S., Guttman, M., Leshyk, V. O., Lis, J. T., Lomvardas, S., Mirny, L. A., O'Shea, C. C., Park, P. J., Ren, B., Politz, J. C. R., Shendure, J. & Zhong, S. The 4D nucleome project. *Nature* **549,** 219–226. (2024) (Sept. 2017).

118. Martens, J. H. A. & Stunnenberg, H. G. BLUEPRINT: mapping human blood cell epigenomes. *Haematologica* **98,** 1487–1489 (Oct. 2013).

119. Ota, M., Nagafuchi, Y., Hatano, H., Ishigaki, K., Terao, C., Takeshima, Y., Yanaoka, H., Kobayashi, S., Okubo, M., Shirai, H., Sugimori, Y., Maeda, J., Nakano, M., Yamada, S., Yoshida, R., Tsuchiya, H., Tsuchida, Y., Akizuki, S., Yoshifuji, H., Ohmura, K., Mimori, T., Yoshida, K., Kurosaka, D., Okada, M., Setoguchi, K., Kaneko, H., Ban, N., Yabuki, N., Matsuki, K., Mutoh, H., Oyama, S., Okazaki, M., Tsunoda, H., Iwasaki, Y., Sumitomo, S., Shoda, H., Kochi, Y., Okada, Y., Yamamoto, K., Okamura, T. & Fujio, K. Dynamic

landscape of immune cell-specific gene regulation in immune-mediated diseases. *Cell* **184,** 3006–3021.e17 (May 2021).

120. Bailey, T. L., Johnson, J., Grant, C. E. & Noble, W. S. The MEME Suite. *Nucleic Acids Research* **43,** W39–49 (July 2015).

121. Schmiedel, B. J., Singh, D., Madrigal, A., Valdovino-Gonzalez, A. G., White, B. M., Zapardiel-Gonzalo, J., Ha, B., Altay, G., Greenbaum, J. A., McVicker, G., Seumois, G., Rao, A., Kronenberg, M., Peters, B. & Vijayanand, P. Impact of Genetic Polymorphisms on Human Immune Cell Gene Expression. *Cell* **175,** 1701–1715.e16 (Nov. 2018).

122. Rappaport, N., Nativ, N., Stelzer, G., Twik, M., Guan-Golan, Y., Stein, T. I., Bahir, I., Belinky, F., Morrey, C. P., Safran, M. & Lancet, D. MalaCards: an integrated compendium for diseases and their annotation. *Database: The Journal of Biological Databases and Curation* **2013,** bat018 (2013).

123. Babbi, G., Martelli, P. L., Profiti, G., Bovo, S., Savojardo, C. & Casadio, R. eDGAR: a database of Disease-Gene Associations with annotated Relationships among genes. *BMC genomics* **18,** 554 (Aug. 2017).

124. Klak, M., Gomółka, M., Kowalska, P., Cichoń, J., Ambrożkiewicz, F., Serwańska-Świętek, M., Berman, A. & Wszoła, M. Type 1 diabetes: genes associated with disease development. *Central-European Journal of Immunology* **45,** 439–453 (2020).

125. Waldmann, T. A. The shared and contrasting roles of IL2 and IL15 in the life and death of normal and neoplastic lymphocytes: implications for cancer therapy. *Cancer Immunology Research* **3,** 219–227 (Mar. 2015).

126. Bobbala, D., Mayhue, M., Menendez, A., Ilangumaran, S. & Ramanathan, S. Trans-presentation of interleukin-15 by interleukin-15 receptor alpha is dispensable for the pathogenesis of autoimmune type 1 diabetes. *Cellular & Molecular Immunology* **14,** 590–596 (July 2017).

127. Chen, J., Feigenbaum, L., Awasthi, P., Butcher, D. O., Anver, M. R., Golubeva, Y. G., Bamford, R., Zhang, X., St Claire, M. B., Thomas, C. J., Discepolo, V., Jabri, B. & Waldmann, T. A. Insulin-dependent diabetes induced by pancreatic beta cell expression of IL-15 and IL-15Rα. *Proceedings of the National Academy of Sciences of the United States of America* **110,** 13534–13539 (Aug. 2013).

128. Lu, J., Liu, J., Li, L., Lan, Y. & Liang, Y. Cytokines in type 1 diabetes: mechanisms of action and immunotherapeutic targets. *Clinical & Translational Immunology* **9,** e1122 (2020).

129. Ibáñez-Costa, A., Perez-Sanchez, C., Patiño-Trives, A. M., Luque-Tevar, M., Font, P., Arias de la Rosa, I., Roman-Rodriguez, C., Abalos-Aguilera, M. C., Conde, C., Gonzalez, A., Pedraza-Arevalo, S., Del Rio-Moreno, M., Blazquez-Encinas, R., Segui, P., Calvo, J., Ortega Castro, R., Escudero-Contreras, A., Barbarroja, N., Aguirre, M. A., Castaño, J. P., Luque, R. M., Collantes-Estevez, E. & Lopez-Pedrera, C. Splicing machinery is impaired in rheumatoid arthritis, associated with disease activity and modulated by anti-TNF therapy. *Annals of the Rheumatic Diseases* **81,** 56–67 (Jan. 2022).

130. Dai, J., Li, Y., Ji, C., Jin, F., Zheng, Z., Wang, X., Sun, X., Xu, X., Gu, S., Xie, Y. & Mao, Y. Characterization of two novel KRAB-domain-containing zinc finger genes, ZNF460 and ZNF461, on human chromosome 19q13.1–>q13.4. *Cytogenetic and Genome Research* **103,** 74–78 (2003).

131. Lupo, A., Cesaro, E., Montano, G., Zurlo, D., Izzo, P. & Costanzo, P. KRAB-Zinc Finger Proteins: A Repressor Family Displaying Multiple Biological Functions. *Current Genomics* **14,** 268–278 (June 2013).

132. Magnitov, M. D., Maresca, M., Saiz, N. A., Teunissen, H., Braccioli, L. & Wit, E. d. *ZNF143 is a transcriptional regulator of nuclear-encoded mitochondrial genes that acts independently of looping and CTCF* Mar. 2024. (2024).

133. Narducci, D. N. & Hansen, A. S. *Putative Looping Factor ZNF143/ZFP143 is an Essential Transcriptional Regulator with No Looping Function* Mar. 2024. (2024).

134. Ortabozkoyun, H., Huang, P.-Y., Cho, H., Narendra, V., LeRoy, G., Gonzalez-Buendia, E., Skok, J. A., Tsirigos, A., Mazzoni, E. O. & Reinberg, D. CRISPR and biochemical screens identify MAZ as a cofactor in CTCF-mediated insulation at Hox clusters. *Nature Genetics* **54,** 202–212 (Feb. 2022).

135. Ortabozkoyun, H., Huang, P.-Y., Cho, H., Tsirigos, A., Mazzoni, E. & Reinberg, D. *Novel Chromatin Insulating Activities Uncovered upon Eliminating Known Insulators in vivo* Apr. 2023. (2024).

136. Zitnik, M., Sosič, R. & Leskovec, J. Prioritizing network communities. *Nature Communications* **9,** 2544 (June 2018).

137. Petrovic, J., Zhou, Y., Fasolino, M., Goldman, N., Schwartz, G. W., Mumbach, M. R., Nguyen, S. C., Rome, K. S., Sela, Y., Zapataro, Z., Blacklow, S. C., Kruhlak, M. J., Shi, J., Aster, J. C., Joyce, E. F., Little, S. C., Vahedi, G., Pear, W. S. & Faryabi, R. B. Oncogenic Notch Promotes Long-Range Regulatory Interactions within Hyperconnected 3D Cliques. *Molecular Cell* **73,** 1174–1190.e12 (Mar. 2019).

138. Chandra, A., Yoon, S., Michieletto, M. F., Goldman, N., Ferrari, E. K., Fasolino, M., Joannas, L., Kee, B. L., Henao-Mejia, J. & Vahedi, G. *A multi-enhancer hub at the Ets1 locus controls T cell differentiation and allergic inflammation through 3D genome topology* Oct. 2022. (2024).

139. Xin, J., Mark, A., Afrasiabi, C., Tsueng, G., Juchler, M., Gopal, N., Stupp, G. S., Putman, T. E., Ainscough, B. J., Griffith, O. L., Torkamani, A., Whetzel, P. L., Mungall, C. J., Mooney, S. D., Su, A. I. & Wu, C. High-performance web services for querying gene and variant annotation. *Genome Biology* **17,** 91 (May 2016).

140. Frankish, A., Diekhans, M., Jungreis, I., Lagarde, J., Loveland, J. E., Mudge, J. M., Sisu, C., Wright, J. C., Armstrong, J., Barnes, I., Berry, A., Bignell, A., Boix, C., Carbonell Sala, S., Cunningham, F., Di Domenico, T., Donaldson, S., Fiddes, I. T., García Girón, C., Gonzalez, J. M., Grego, T., Hardy, M., Hourlier, T., Howe, K. L., Hunt, T., Izuogu, O. G., Johnson, R., Martin, F. J., Martínez, L., Mohanan, S., Muir, P., Navarro, F. C. P., Parker, A., Pei, B., Pozo, F., Riera, F. C., Ruffier, M., Schmitt, B. M., Stapleton, E., Suner, M.-M., Sycheva, I., Uszczynska-Ratajczak, B., Wolf, M. Y., Xu, J., Yang, Y. T., Yates, A., Zerbino, D., Zhang, Y., Choudhary, J. S., Gerstein, M., Guigó, R., Hubbard, T. J. P., Kellis, M., Paten, B., Tress, M. L. & Flicek, P. GENCODE 2021. *Nucleic Acids Research* **49,** D916–D923 (Jan. 2021).

141. Castro-Mondragon, J. A., Riudavets-Puig, R., Rauluseviciute, I., Lemma, R. B., Turchi, L., Blanc-Mathieu, R., Lucas, J., Boddie, P., Khan, A., Manosalva Pérez, N., Fornes, O., Leung, T. Y., Aguirre, A., Hammal, F., Schmelter, D., Baranasic, D., Ballester, B., Sandelin, A., Lenhard, B., Vandepoele, K., Wasserman, W. W., Parcy, F. & Mathelier, A. JASPAR 2022: the 9th release of the open-access database of transcription factor binding profiles. *Nucleic Acids Research* **50,** D165–D173 (Jan. 2022).

142. Grant, C. E., Bailey, T. L. & Noble, W. S. FIMO: scanning for occurrences of a given motif. *Bioinformatics (Oxford, England)* **27,** 1017–1018 (Apr. 2011).

143. Gkazi, A. *An Overview of Next-Generation Sequencing* (2024).

144. Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L. & Wold, B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature Methods* **5,** 621–628. (2024) (July 2008).

145. Johnson, D. S., Mortazavi, A., Myers, R. M. & Wold, B. Genome-wide mapping of in vivo protein-DNA interactions. *Science (New York, N.Y.)* **316,** 1497–1502 (June 2007).

146. Skene, P. J. & Henikoff, S. An efficient targeted nuclease strategy for high-resolution mapping of DNA binding sites. *eLife* **6,** e21856. (2024).

147. Zhou, W., Sherwood, B., Ji, Z., Xue, Y., Du, F., Bai, J., Ying, M. & Ji, H. Genome-wide prediction of DNase I hypersensitivity using gene expression. *Nature Communications* **8,** 1038. (2024) (Oct. 2017).

148. Buenrostro, J., Wu, B., Chang, H. & Greenleaf, W. ATAC-seq: A Method for Assaying Chromatin Accessibility Genome-Wide. *Current protocols in molecular biology / edited by Frederick M. Ausubel ... [et al.]* **109,** 21.29.1–21.29.9. (2024) (Jan. 2015).

149. Graw, S., Chappell, K., Washam, C. L., Gies, A., Bird, J., Robeson, M. S. & Byrum, S. D. Multi-omics data integration considerations and study design for biological systems and disease. *Molecular omics* **17,** 170–185. (2024) (Apr. 2021).

150. Omenn, G. S., Lane, L., Overall, C. M., Corrales, F. J., Schwenk, J. M., Paik, Y.-K., Van Eyk, J. E., Liu, S., Pennington, S., Snyder, M. P., Baker, M. S. & Deutsch, E. W. Progress on Identifying and Characterizing the Human Proteome: 2019 Metrics from the HUPO Human Proteome Project. *Journal of proteome research* **18,** 4098–4107. (2024) (Dec. 2019).

151. Iyer, A. CyTOF® for the Masses. *Frontiers in immunology* (Apr. 2022).

152. Chicco, D., Cumbo, F. & Angione, C. Ten quick tips for avoiding pitfalls in multi-omics data integration analyses. *PLOS Computational Biology* **19,** e1011224. (2024) (July 2023).

153. Vahabi, N. & Michailidis, G. Unsupervised Multi-Omics Data Integration Methods: A Comprehensive Review. *Frontiers in Genetics* **13,** 854752 (2022).

154. Qiu, L. & Chinchilli, V. M. *Probabilistic Canonical Correlation Analysis for Sparse Count Data* May 2020. (2024).

155. Nguyen, N. D. & Wang, D. Multiview learning for understanding functional multiomics. *PLoS Computational Biology* **16,** e1007677. (2024) (Apr. 2020).

156. Kolenc, Ž., Pirih, N., Gretic, P. & Kunej, T. Top Trends in Multiomics Research: Evaluation of 52 Published Studies and New Ways of Thinking Terminology and Visual Displays. *OMICS: A Journal of Integrative Biology* **25,** 681–692. (2024) (Nov. 2021).

157. Wang, S., Yong, H. & He, X.-D. Multi-omics: Opportunities for research on mechanism of type 2 diabetes mellitus. *World Journal of Diabetes* **12,** 1070–1080. (2024) (July 2021).

158. Lussier, Y. A. & Li, H. Breakthroughs in genomics data integration for predicting clinical outcome. *Journal of Biomedical Informatics* **45,** 1199–1201. (2024) (Dec. 2012).

159. Goff, A., Cantillon, D., Muraro Wildner, L. & Waddell, S. J. Multi-Omics Technologies Applied to Tuberculosis Drug Discovery. *Applied Sciences* **10,** 4629. (2024) (Jan. 2020).

160. Lu, M. & Zhan, X. The crucial role of multiomic approach in cancer research and clinically relevant outcomes. *The EPMA Journal* **9,** 77–102. (2024) (Feb. 2018).

161. Zhao, Z., Ding, Y., Tran, L. J., Chai, G. & Lin, L. Innovative breakthroughs facilitated by single-cell multi-omics: manipulating natural killer cell functionality correlates with a novel subcategory of melanoma cells. *Frontiers in Immunology* **14.** (2024) (June 2023).

162. Shalita, R. & Amit, I. The industrial genomic revolution: A new era in neuroimmunology. *Neuron* **110,** 3429–3443 (Nov. 2022).

163. Xu, G., Wu, Y., Xiao, T., Qi, F., Fan, L., Zhang, S., Zhou, J., He, Y., Gao, X., Zeng, H., Li, Y. & Zhang, Z. Multiomics approach reveals the ubiquitination-specific processes hijacked by SARS-CoV-2. *Signal Transduction and Targeted Therapy* **7,** 1–13. (2024) (Sept. 2022).

164. Almeida, L. S., Pereira, C., Aanicai, R., Schröder, S., Bochinski, T., Kaune, A., Urzi, A., Spohr, T. C. L. S., Viceconte, N., Oppermann, S., Alasel, M., Ebadat, S., Iftikhar, S., Jasinge, E., Elsayed, S. M., Tomoum, H., Marzouk, I., Jalan, A. B., Cerkauskaite, A.,

Cerkauskiene, R., Tkemaladze, T., Nadeem, A. M., El Din Mahmoud, I. G., Mossad, F. A., Kamel, M., Selim, L. A., Cheema, H. A., Paknia, O., Cozma, C., Juaristi-Manrique, C., Guatibonza-Moreno, P., Böttcher, T., Vogel, F., Pinto-Basto, J., Bertoli-Avella, A. & Bauer, P. An integrated multiomic approach as an excellent tool for the diagnosis of metabolic diseases: our first 3720 patients. *European Journal of Human Genetics* **30,** 1029–1035. (2024) (Sept. 2022).

165. Agrawal, M., Allin, K. H., Petralia, F., Colombel, J.-F. & Jess, T. Multiomics to elucidate inflammatory bowel disease risk factors and pathways. *Nature Reviews Gastroenterology & Hepatology* **19,** 399–409. (2024) (June 2022).

166. Martino, D., Ben-Othman, R., Harbeson, D. & Bosco, A. Multiomics and Systems Biology Are Needed to Unravel the Complex Origins of Chronic Disease. *Challenges* **10,** 23. (2024) (June 2019).

167. Ota, M. & Fujio, K. Multi-omics approach to precision medicine for immune-mediated diseases. *Inflammation and Regeneration* **41,** 23. (2024) (Aug. 2021).

168. Tenenhaus, A. & Tenenhaus, M. Regularized Generalized Canonical Correlation Analysis. *Psychometrika* **76,** 257–284. (2023) (Apr. 2011).

169. Hanafi, M., Kohler, A. & Qannari, E.-M. Connections between multiple co-inertia analysis and consensus principal component analysis. *Chemometrics and Intelligent Laboratory Systems. Multiway and Multiset Data Analysis* **106,** 37–40. (2023) (Mar. 2011).

170. Meng, C., Basunia, A., Peters, B., Gholami, A. M., Kuster, B. & Culhane, A. C. MOGSA: Integrative Single Sample Gene-set Analysis of Multiple Omics Data. *Molecular & Cellular Proteomics : MCP* **18,** S153–S168. (2023) (Aug. 2019).

171. Shoemaker, R. H. The NCI60 human tumour cell line anticancer drug screen. *Nature Reviews. Cancer* **6,** 813–823 (Oct. 2006).

172. Genomics, 1. 5k Peripheral blood mononuclear cells (PBMCs) from a healthy donor with cell surface proteins (v3 chemistry) (May 2019).