

UCSF

UC San Francisco Electronic Theses and Dissertations

Title

Machine Learning Insights into the 3D Genome: Diversity and Gene Regulation in Human Populations

Permalink

<https://escholarship.org/uc/item/7zk8d4rk>

Author

Gilbertson, Erin Nicole

Publication Date

2024

Peer reviewed|Thesis/dissertation

Machine Learning Insights into the 3D Genome:
Diversity and Gene Regulation in Human Populations

by
Erin Nicole Gilbertson

DISSERTATION
Submitted in partial satisfaction of the requirements for degree of
DOCTOR OF PHILOSOPHY

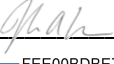
in

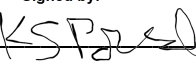
Biological and Medical Informatics

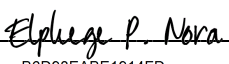
in the

GRADUATE DIVISION
of the
UNIVERSITY OF CALIFORNIA, SAN FRANCISCO

Approved:

Signed by:  _____ John A. Capra
FEE00BDBE74841E... Chair

Signed by:  _____ Katherine S. Pollard
DECUSIGNEDBYE6...

 _____ Elphege P. Nora
B6D98EABE1914FD...

Committee Members

Copyright 2024

by

Erin Nicole Gilbertson

For Grandma Vicki

ACKNOWLEDGEMENTS

Writing this acknowledgments section has been, in some ways, both the most challenging and meaningful part of my dissertation. Each time I've attempted to put my gratitude into words over the past two months, I've been overwhelmed by memories of the incredible people who have made this journey possible. Now, in this final week, I'm filled with deep appreciation for those without whom I would not have reached this point.

First and foremost, I'd like to thank my incredible support system at UCSF in the Capra Lab. Dr. Tony Capra, you have been an exceptional mentor and advisor these past three years. After finding myself lab-less following my second year at UCSF, you welcomed me into your lab and guided me through the challenge of completing my qualifying exam on time with the rest of my cohort. Together, we developed a project I'm truly proud of, and I've learned so much from the experience. You foster an incredible lab environment filled with some of the best people I know.

A heartfelt thank you to Evonne McArthur, Sarah Fong, Keila Velázquez-Arcelay, David Rinker, and all the Vanderbilt folks who welcomed me so warmly as the first UCSF student in your lab. Your mentorship and inspiration have been invaluable over the years. Colin Brand, I couldn't have asked for a better labmate to join at the same time. You are an outstanding postdoc, an incredible friend, an amazing mentor, and simply one of the best people I know.

To Manu, Oge, and Chimno—I wish I'd had more time with you, but I know the lab is fortunate to have you all! And to my fellow grad students—Grace, Seb, Ava, Hannah, Yaen, Allison, Gyasu, and Ashvin—you are all brilliant, kind, and thoughtful. I can't wait to see the amazing things you'll continue to accomplish!

I'm incredibly grateful to my committee members, Katie Pollard and Elphege Nora, whose support and feedback made our meetings something I actually looked forward to. Your insights and encouragement have been essential in shaping my research and helping me grow as a scientist.

A big thank you to Ryan Hernandez and Tony Capra, the directors of the Biological and Medical Informatics program, for your leadership and guidance throughout my PhD. Your dedication to the program and your students has been truly inspiring, and I'm so thankful for the opportunities you've provided.

Rebecca Dawson, our amazing program manager, deserves a special shoutout for keeping everything running smoothly—which is honestly nothing short of a miracle. Your organizational skills and constant support have been a lifesaver.

And to my cohort—Laura Shub, Scott Nanda, Zach Cutts, Tianna Grant, Aiden Winters, and Hasuni Alkhairo—thank you for the laughs, camaraderie, and shared experiences that made this journey so much more enjoyable. I'm so lucky to have gone through this with such an awesome group of people.

This journey started long before my time at UCSF. I would definitely not be here today without Dr. Suzanne McGaugh and Dr. Yaniv Brandvain at the University of Minnesota. Suzanne saw something in me as a clueless freshman and she, Yaniv and their labs welcomed me immediately. I learned all of my first lessons about being a scientist and computational genomics while working with them. An additional thanks to Dr. Chad Myers who served alongside Suzanne and Yaniv on my undergraduate honors thesis committee and also visited UCSF as my first invited guest speaker this past winter.

Of course, none of this would have been possible without the unwavering support of my family. To my mom and dad, Renee and Brian Gilbertson, thank you for always believing in me and supporting me throughout my life. From helping me move to San Francisco to keeping me sane during the pandemic, your love and encouragement have been my foundation.

To my younger brother Jeremy, you are the absolute best. Your constant smile, humor, and ability to make any situation brighter have been a true gift. I'm especially grateful for your amazing naturalist skills and hiking guidance when I needed to escape the city. And thank you for choosing your incredible wife, Bri, who has become such a cherished part of our family. To my sisters-in-law Bri, Julia, and Olivia thank you for being rays of sunshine in my life. I'm so lucky to have three amazing sisters-in-law who I genuinely love spending time with.

To my entire extended family—the Gilbertsons, Bakers, Tollefsons, Richards, Jols, Sterners, Menzels, and Blooms—your support has meant so much to me. Knowing I have such a strong, loving network behind me has been incredibly important.

There are too many friends I love to list you all here, but I'll do my best. Maggie, Katie, Lindsey, Laura, and Danna you stood with my on my wedding day and every day these last five years, I love you all so much (along with Max, Grant, Jack, Mikhail, Greg, Neil and Brandon)! Madelyn, Neil, Mikhail and Grant—having friends outside of UCSF who understand the PhD student life has been an incredible gift, and source of companionship. Nat Rose, I'm not sure I would have started or stuck with computer science if you were not right alongside me in the first few classes. You inspire me to change the plans to fit my passions to do what makes me happy.

A special shoutout to my parents' golden retrievers who brought endless energy and happiness into my life. Tucker, who sadly passed away the day of my qualifying exam, Libby, and Belle were always there with a tail wag or cuddle when I needed motivation or comfort to keep going. Especially during the pandemic, some of the best support I received was FaceTime calls home where I got to see your wagging tails!

And last but certainly not least, my husband Mark. Your unwavering love and support throughout this wild ride of grad school have been everything. Marrying a PhD student is no simple task, but you stuck with me through all of it and kept a smile on both of our faces. Your support means the world to me, and I'm so grateful to have you by my side.

Finally, to everyone else who has been a part of this journey, whether mentioned here or not—please know that your support, encouragement, and kindness have made all the difference. This list could never fully capture all the people who have helped me along the way, but I am deeply grateful to each and every one of you.

CONTRIBUTIONS

This dissertation was supervised by Dr. John A. Capra. Additional guidance was provided by Dr. Katherine S. Pollard and Dr. Elphege P. Nora.

Chapter 2 contains material from a manuscript currently in press at Molecular Biology and Evolution and available in open-access format:

Gilbertson, E.N., Brand, C.M., McArthur, E., Rinker, D.C., Kuang, S., Pollard, K.S., and Capra, J.A. 2023, December 23. Machine learning reveals the diversity of human 3D chromatin contact patterns. bioRxiv. doi:10.1101/2023.12.22.573104.

What I want to know: is how does the song go...
...What I want to know: is where does the time go
- The Grateful Dead, Uncle John's Band

**Machine Learning Insights into the 3D Genome:
Diversity and Gene Regulation in Human Populations**

Erin Nicole Gilbertson

ABSTRACT

The 3D organization of the human genome plays a crucial role in gene regulation, influencing interactions between genes and regulatory elements. Despite significant progress in genomics, the diversity of 3D chromatin contact patterns across human populations remains underexplored. This dissertation describes the use of machine learning to predict 3D chromatin contact maps from genome sequences, revealing new insights into genome architecture among diverse populations. In Chapter 1, I provide a literature review and overview of human population and regulatory genetics in relationship to the 3D genome with a focus on machine learning techniques. In Chapter 2, I present the results of my study using a machine learning model to predict 3D genome for thousands of individuals, uncovering substantial 3D genomic diversity, particularly within African populations. I also identified regions where 3D divergence occurs despite relatively low sequence variation, especially in areas under low functional constraint. In Chapter 3, I provide a perspective on my work and future directions. My findings underscore the importance of considering 3D genome organization in understanding gene regulation and its implications for health and disease.

TABLE OF CONTENTS

Chapter 1: Using machine learning methods to explore the evolution of gene

regulation.....	1
Introduction	1
Human Population Genetics	3
<i>Genetic variation in human populations</i>	<i>3</i>
<i>Methods for studying population genetics</i>	<i>7</i>
<i>Bias and inequity in population genetic studies.....</i>	<i>10</i>
Gene Regulation and the 3D Genome	11
<i>Introduction to gene regulatory elements</i>	<i>11</i>
<i>Disease implications of regulatory disruption</i>	<i>14</i>
<i>3D chromatin conformation in gene regulation.....</i>	<i>15</i>
<i>Conclusion.....</i>	<i>23</i>
Evolution of Gene Regulatory Functions.....	24
<i>Evolutionary conservation of regulatory elements.....</i>	<i>24</i>
<i>Mechanisms of regulatory evolution</i>	<i>29</i>
<i>Regulatory variation in diverse human populations.....</i>	<i>30</i>
<i>Disease implications of regulatory disruption</i>	<i>31</i>
<i>Impact of evolution on 3D genome structure.....</i>	<i>32</i>
DNA Sequence-Based Machine Learning.....	34
<i>Historical context and motivation.....</i>	<i>34</i>

<i>Types of machine learning models</i>	36
<i>Applications in genomic regulation</i>	37
<i>Recent advances</i>	41
<i>Conclusion</i>	44
Conclusion	44
<i>Innovation</i>	44
<i>Significance</i>	45
Chapter 2: Machine learning reveals the diversity of human 3d chromatin	
contact patterns	46
Abstract	46
Introduction	47
Results	49
<i>Accurate prediction of 3D contact maps for diverse individuals</i>	49
<i>3D divergence differs from sequence divergence</i>	51
<i>African populations have the highest predicted 3D genome diversity</i>	51
<i>Most variation in 3D chromatin contact patterns is shared across populations</i>	52
<i>3D divergence is highest in regions with the lowest functional constraint</i>	53
<i>3D chromatin contact constrains sequence evolution</i>	55
<i>392 windows have significantly greater 3D divergence than expected</i>	56
<i>In silico mutagenesis reveals that multiple SNVs contribute to common 3D</i> <i>genome variation</i>	57
<i>31% of the genome has rare 3D genome variation</i>	59

<i>Rare 3D genome variation is usually the result of a single large-effect variant</i>	59
Discussion	61
<i>3D chromatin contact divergence vs. sequence divergence</i>	61
<i>Influence of 3D chromatin contact on sequence evolution and functional constraint</i>	62
<i>In silico mutagenesis identifies SNVs likely to drive 3D divergence</i>	63
<i>Machine learning addresses challenges with experimental Hi-C</i>	64
<i>Limitations</i>	64
<i>Conclusions</i>	66
Methods.....	67
<i>Modern human and ancestral genomes</i>	67
<i>3D chromatin contact prediction with Akita</i>	68
<i>Scaling 3D genome predictions</i>	68
<i>3D and sequence genome comparisons</i>	69
<i>Empirical distribution of expected 3D divergence</i>	69
<i>Comparison of function and conserved elements with 3D divergence</i>	70
<i>Shared divergent windows across populations</i>	72
<i>Hierarchical clustering of 3D chromatin contact maps</i>	72
<i>In silico mutagenesis</i>	73
<i>Analysis of experimental Hi-C data</i>	74
<i>Significance reporting</i>	75
<i>Data availability</i>	75

<i>Code availability</i>	76
<i>Acknowledgements</i>	76
<i>Author contributions</i>	77
<i>Competing interests</i>	77
Figures	78
Supplemental Figures	90
Chapter 3: Final thoughts and future directions	104
Summary of key findings.....	104
Significance of the work	105
Challenges and limitations	106
Future directions	107
Concluding remarks	109
References	110

LIST OF FIGURES

Figure 2.1: Strategy for investigating 3D chromatin contact patterns in diverse humans.....	78
Figure 2.2: Genome-wide 3D divergence follows known population structure	80
Figure 2.3: 3D Divergence is variable across the genome and highest in less functional regions.	82
Figure 2.4: 3D divergence is lower than expected in 89% of genomics windows, but 392 have significantly greater 3D divergence than expected.	83
Figure 2.5: Experimental Hi-C data confirms predicted contacts in highly divergent windows.....	84
Figure 2.6: Most common divergent windows cannot be explained by a single nucleotide variant.	86
Figure 2.7: Genomic windows with rare variation in 3D contact patterns are common.	88
Supplemental Figure 2.1: Generating “flattened” genome sequences.	90
Supplemental Figure 2.2: Correlation between read count and prediction accuracy.....	91
Supplemental Figure 2.3: 3D divergence estimates at different window sizes.....	92
Supplemental Figure 2.4: MSE recapitulates divergence patterns calculated using 3D divergence.	93
Supplemental Figure 2.5: Genome-wide standard deviation of 3D divergence between all 1KG individuals.	95
Supplemental Figure 2.6: Chromosome distributions of 3D divergence.....	96
Supplemental Figure 2.7: Upset plot of population representation in top 10% 3D divergent windows.	97

Supplemental Figure 2.8: SPIN state and repeat element content across 3D	
divergence deciles.....	98
Supplemental Figure 2.9: Generating empirical distribution of expected 3D	
divergence.....	99
Supplemental Figure 2.10: Full upset plot for more divergent than expected	
windows.....	100
Supplemental Figure 2.11: Distributions of 3D divergence from <i>in silico</i>	
mutagenesis on common variants from the 392 divergent windows and 392	
non-divergent windows.....	101
Supplemental Figure 2.12: Genomic annotations for 3D modifying variants in	
divergent windows are consistent across cutoffs.....	102
Supplemental Figure 2.13: Genomic annotations for 3D modifying variants in	
windows with rare 3D divergence are consistent across cutoffs.	103

LIST OF ABBREVIATIONS

1KG	1000 Genomes Project
3C	Chromatin Conformation Capture
3D	3-Dimensional
4C	Circular Chromatin Conformation Capture
4DN	4-Dimensional Nucleome Project
5C	Chromosome Conformation Capture Carbon Copy
ATAC	Assay for Transposase-Accessible Chromatin
BNN	Bayesian Neural Network
ChIP	Chromatin Immunoprecipitation
CNE	Conserved Noncoding Element
CNN	Convolutional Neural Network
CNV	Copy Number Variant
CTCF	CCCTC-binding Factor
eQTL	Expression Quantitative Trait Locus
GLAD	Genetics of Latin American Diversity
GWAS	Genome-wide Association Study
H3Africa	Human Heredity and Health in Africa
HAR	Human Accelerated Region
hCONDEL	Human Conserved Deletion
HFF	Human Foreskin Fibroblast
hg38	GRCh38 Human Reference Genome
HGDP	Human Genome Diversity Project

HMM	Hidden Markov Model
indel	Insertion or Deletion
LSTM	Long Short-term Memory
MEI	Mobile Element Insertion
ML	Machine Learning
MPRA	Massively Parallel Reporter Assay
MSE	Mean-Squared Error
PCA	Principal Component Analysis
QTL	Quantitative Trait Locus
SGDP	Simons Genome Diversity Project
SHAP	Shapley Additive Explanations
SNP	Single Nucleotide Polymorphism
SNV	Single Nucleotide Variant
SV	Structural Variant
SVM	Support Vector Machine
TAD	Topologically Associated Domain
TE	Transposable Element
TF	Transcription Factor
TFBS	Transcription Factor Binding Site
WGS	Whole Genome Sequencing

CHAPTER 1: USING MACHINE LEARNING METHODS TO EXPLORE THE EVOLUTION OF GENE REGULATION

Introduction

Deciphering the three-dimensional (3D) organization of the genome is a crucial component of understanding the mechanisms of gene regulation. Chromatin, the complex of DNA and proteins within the nucleus, is not randomly arranged but is organized in a highly structured manner that influences gene expression and cellular function. The spatial arrangement of chromatin, including loops, domains, and compartments, modulates the proximity of regulatory elements such as enhancers and promoters, facilitating or inhibiting their ability to control gene expression. This 3D architecture is dynamic and changes in response to various cellular signals, playing a vital role in development, differentiation, and disease.

Genetic variants that affect genome folding can alter chromatin interactions and cause disease. However, there is a lack of understanding of how 3D genome contacts vary within a species and when genetic variants will perturb 3D folding. Even though DNA sequence changes that influence genome folding can have massive regulatory impacts, current methods for interpreting non-coding variants do not consider this mechanism. Therefore, a comprehensive understanding of 3D chromatin contact patterns is essential for advancing our knowledge of gene regulation and its implications in health and disease.

Machine learning has emerged as a powerful tool in genomics, enabling the analysis and interpretation of large and complex datasets. In the context of 3D genome organization, machine learning algorithms can be used to predict chromatin interactions,

identify regulatory elements, and infer the functional consequences of genetic variants. These models leverage vast amounts of genomic data to learn patterns and make predictions that would be difficult, if not impossible, to achieve through traditional methods.

The integration of machine learning with genomic data has already led to significant advances in the field. For example, models such as Akita, DeepC, and Orca have been developed to predict 3D chromatin interactions from DNA sequence data, providing insights into the underlying principles of genome organization. These advancements have not only improved our understanding of chromatin dynamics but also hold the potential to uncover new regulatory mechanisms and therapeutic targets.

To address the challenge of experimentally assaying chromatin interactions across diverse populations and tissues, I propose a machine learning approach to quantify variation in 3D chromatin contact patterns from genome sequence alone. This approach aims to understand the diversity of chromatin contact patterns in modern humans, hypothesizing that there are quantifiable differences in the 3D genome organization that constrain sequence evolution and contribute to regulatory variation.

The primary objective of this dissertation is to explore the diversity of human 3D chromatin contact patterns using a machine learning approach. The following sections lay out the key concepts and references necessary to contextualize the work.

Human Population Genetics

Genetic variation in human populations

Genetic variation refers to differences in DNA sequences among individuals within a population. Understanding this variation is crucial for studying human evolution, population genetics, and disease mechanisms. This section provides an overview of the different types of genetic variation, the roles of mutation, recombination, and gene flow, and the mechanisms generating genetic diversity.

Genetic variation can be classified into several types, each with distinct characteristics and implications for genome function and evolution. Single Nucleotide Polymorphisms (SNPs) are the most common type of genetic variation, involving a single base change in the DNA sequence. SNPs occur approximately once every 1,000 base pairs in the human genome and can have significant functional impacts, particularly when they occur in coding or regulatory regions (1000 Genomes Project Consortium et al. 2015; Karczewski et al. 2020). Insertions and Deletions (indels), which involve the addition or removal of small DNA segments, are less frequent than SNPs but can have substantial effects on gene function, especially if they cause frameshift mutations or alter regulatory elements (Karczewski et al. 2020).

Structural Variants (SVs) encompass a wide range of large-scale genomic alterations, including deletions, duplications, inversions, and translocations. These variants can disrupt gene function and regulatory networks and are often challenging to detect due to their size and complexity. Recent advancements in sequencing technologies and computational methods have improved the detection and

characterization of SVs, highlighting their significant role in human genetic diversity (Collins et al. 2020). Copy Number Variants (CNVs), a subtype of SVs where segments of the genome are present in variable copy numbers among individuals, can influence gene expression levels and have been linked to various diseases and phenotypic traits (Telenti et al. 2016). Mobile Element Insertions (MEIs) include the insertion of transposable elements (TEs) into new genomic locations. TEs contribute to genome evolution and can impact gene function by disrupting coding sequences or regulatory regions.

Genetic variation arises through mutation, recombination, and gene flow, each playing a crucial role in shaping the genetic landscape of human populations. Mutation is the primary source of new genetic variation. Mutations can occur spontaneously due to errors in DNA replication or be induced by environmental factors. The mutation rate varies across different regions of the genome and among populations, influencing genetic diversity and evolutionary trajectories (Besenbacher et al. 2019). Beneficial mutations may be subject to positive selection, while deleterious mutations can be eliminated through purifying selection. Neutral mutations, which do not affect fitness, can drift through the population, contributing to genetic diversity (Karczewski et al. 2020). Recombination shuffles genetic material during meiosis, creating new allele combinations and contributing to genetic diversity. Recombination rates are not uniform across the genome; they tend to occur more frequently at specific hotspots. These hotspots can vary between populations, affecting linkage disequilibrium patterns and genetic diversity (Pratto et al. 2014). This process can bring together beneficial alleles from different loci, facilitating adaptation and evolution. Recombination also breaks down linkage

disequilibrium, enhancing the potential for natural selection to act on individual alleles (Pratto et al. 2014). Gene flow introduces new alleles into a population through migration and interbreeding between populations. Gene flow plays a vital role in maintaining genetic diversity by introducing alleles from other populations. Gene flow can counteract the effects of genetic drift, which tends to reduce genetic diversity in small populations. Historical examples of gene flow, such as the admixture events between modern humans and Neanderthals, have left significant genetic footprints in contemporary human genomes (Nielsen et al. 2017). Understanding these mechanisms provides insights into the evolutionary history and genetic diversity of human populations.

Genetic variation is distributed unevenly across human populations and geographic regions, reflecting complex histories of migration, selection, and drift. Studies like the 1000 Genomes Project (1KG) have illustrated how genetic diversity diminishes with geographical distance from East Africa, believed to be the origin of modern humans. This pattern, described by the "serial founder effect," occurs as successive groups migrate farther from their origin, each carrying only a subset of the genetic diversity of their source population (Ramachandran et al. 2005; 1000 Genomes Project Consortium et al. 2015).

Several factors influence the distribution of genetic diversity. Smaller populations often exhibit less genetic diversity due to stronger effects of genetic drift, where random changes can have a more pronounced impact on the genetic structure (Charlesworth 2009). Migration plays a critical role in introducing new genetic variations into populations. For example, the expansion of humans out of Africa and subsequent migrations have significantly shaped the genetic landscapes of modern populations (Cavalli-Sforza 1997).

Natural selection also affects genetic diversity, as certain genes may undergo selective pressures that enhance or reduce variation, such as those involved in disease resistance or skin pigmentation, reflecting adaptations to environmental challenges (Barreiro et al. 2008).

To effectively study and quantify genetic variation, researchers employ various metrics and methods that reveal how genetic differences are structured within and between populations. Heterozygosity measures the likelihood that two randomly chosen alleles at a locus are different, indicating the genetic diversity within a population. The fixation index, or F_{ST} , helps quantify genetic differentiation between populations, with higher values suggesting greater genetic variance among groups (Weir and Cockerham 1984). Principal Component Analysis (PCA) is another critical tool, reducing the complexity of genetic data to visualize and analyze patterns of similarity and diversity across populations (Patterson et al. 2006).

Advancements in genotyping and sequencing technologies have revolutionized our ability to capture detailed genetic variation at both the individual and population levels. These technologies are instrumental in identifying both common and rare genetic variants, providing a comprehensive picture of genetic diversity (1000 Genomes Project Consortium et al. 2015). Additionally, population genetic simulations offer a powerful approach to modeling the effects of evolutionary processes on genetic variation, helping researchers understand how mutation, recombination, and natural selection have shaped the genetic diversity observed in contemporary human populations (Hudson 2002; Rodrigues et al. 2024).

Methods for studying population genetics

Genotyping arrays. Genotyping arrays, or SNP chips, are a cost-effective method for assessing genetic variation across populations. They target specific alleles, allowing for the analysis of thousands of SNPs simultaneously. This technology has facilitated large-scale genetic studies and genome-wide association studies (GWAS), identifying genetic markers linked to various traits and diseases. However, genotyping arrays are limited to detecting a subset of known variants, potentially missing novel or rare variants that may be important in certain populations (McCarthy et al. 2008).

Whole-genome sequencing. Whole-genome sequencing (WGS) offers a comprehensive method for analyzing genetic variation, providing the complete DNA sequence of an organism's genome at a single time. WGS has revolutionized genetics by allowing the discovery of new genetic variants, including SNPs, indels, and structural variations that are not covered by genotyping arrays. This method has become increasingly accessible and cost-effective, although it still represents a significant investment compared to other techniques (Goodwin et al. 2016).

Population specific studies. Studying genetic variation in specific populations can reveal adaptations and evolutionary histories that are not evident in broader studies. Such research has highlighted the importance of including diverse populations in genetic research, as findings from one group may not be applicable to all. Population-specific studies have also identified unique genetic variants that contribute to disease risk or drug

response, which are critical for personalized medicine approaches (Bustamante et al. 2011).

Research initiatives focused on African populations are critical given Africa's status as the most genetically diverse continent. The Human Heredity and Health in Africa (H3Africa) initiative has been pivotal in enhancing the representation of African populations in genetic studies. This consortium has led to significant discoveries by sequencing genomes from diverse African populations, highlighting unique genetic traits and variants that are underrepresented in global databases (The H3Africa Consortium et al. 2014). Projects like the African Genome Variation Project continue to expand our understanding of genetic diversity and disease susceptibility unique to African populations, underscoring the need for a vast increase in genome sequencing efforts across the continent (Gurdasani et al. 2015).

The genetic complexity of South Asian populations has been explored through projects like the GenomeAsia 100K Project. This initiative has cataloged extensive genetic variation across the region, which is influenced by a long history of migrations and complex social structures such as the caste system. The data derived from these studies are crucial for understanding the genetic basis of various diseases prevalent in South Asian populations and for developing targeted treatments (Wall et al. 2019).

Admixed populations in the Americas, particularly those in Latin America, offer unique insights into the genetic outcomes of historical mixing between Indigenous peoples, Europeans, and Africans. Projects like the Mexican Biobank have been pivotal in cataloging the genetic diversity of the Mexican population, illuminating complex genetic structures and their implications for health and disease. Recent findings from this biobank

have shed light on fine-scale genetic ancestries and demographic histories across Mexico, enhancing our understanding of trait-relevant genetic variation (Sohail et al. 2023). In addition, the "Genetics of Latin American Diversity (GLAD) Project" has made significant contributions by examining genetic diversity across various Latin American countries. This study has identified key genetic variants that influence health and disease in these populations, providing insights crucial for the development of medical treatments tailored to the genetic profiles of Latin American individuals. The GLAD project highlights the importance of studying recently admixed groups to better understand the complex interplay of genetic factors contributing to health disparities and treatment responses in these populations (Borda et al. 2023).

Global projects. The endeavor to map human genetic diversity on a global scale has been propelled by a series of foundational projects, each building on the findings of the previous. The Human Genome Diversity Project (HGDP) set the early stage by focusing on indigenous populations to elucidate their genetic relationships and health patterns, providing vital insights into genetic risk factors and human migrations (Cavalli-Sforza 2005). Following this, the International HapMap Project aimed to develop a haplotype map of the human genome, clarifying common patterns of DNA sequence variations. This project was pivotal in enabling genome-wide association studies, linking genetic variations to health, disease, and responses to drugs and environmental factors (Altshuler et al. 2005).

Building on this groundwork, the 1KG Project expanded the scope by mapping detailed genetic variation among 2,500 individuals from 26 populations around the world.

This project provided a comprehensive resource that has been critical for understanding human diversity and its implications for medicine (1000 Genomes Project Consortium et al. 2015). More recently, the Simons Genome Diversity Project (SGDP) advanced these efforts by sequencing 300 genomes from 142 diverse populations, uncovering detailed patterns of human migration and ancient demographic histories, thus providing deeper insights into our evolutionary history (Mallick et al. 2016).

Bias and inequity in population genetic studies

Human population genetics has historically faced significant biases, particularly the oversampling of European populations. This bias has implications for genetic research and medical applications, limiting the generalizability of findings and potentially exacerbating health disparities. Studies have shown that genomic data from predominantly European cohorts have been overly represented in genetic research databases, which skews our understanding of human genetic diversity and disease susceptibility (Popejoy and Fullerton 2016).

The implications of such biases are profound. They impact the accuracy of polygenic risk scores and other genetic tools that are increasingly used in personalized medicine. As these tools are based on data derived largely from European populations, their predictive power and relevance can be significantly diminished for individuals of other ancestries. This results in a healthcare gap, where non-European populations receive less benefit from advances in genomics (Martin et al. 2019).

In response to these challenges, there have been concerted efforts in the last few years to rethink how we cluster and label human cohorts in genetic studies. Recent initiatives aim to increase the diversity of genomic studies by including more underrepresented populations. Projects like the All of Us Research Program in the United States and the H3Africa initiative are examples of efforts to collect and analyze genetic data from a broader array of human populations (The H3Africa Consortium et al. 2014; Bick et al. 2024).

These recent efforts are not only about increasing sample diversity but also involve re-evaluating the methodologies used to categorize and interpret genetic data. There is a growing recognition of the need to move beyond simplistic racial or geographical labels and instead consider a more nuanced understanding of genetic ancestry and its implications for health and disease (Diaz-Papkovich et al. 2019, 2023; Lewis et al. 2022).

Gene Regulation and the 3D Genome

Introduction to gene regulatory elements

Gene regulatory elements are essential components of the genome that control the spatial and temporal expression of genes. These elements play a crucial role in development, differentiation, and the response to environmental stimuli. They orchestrate the complex regulation of gene expression through various mechanisms. Promoters, enhancers, silencers, and insulators are key regulatory elements that interact with transcription factors and other regulatory proteins to ensure precise gene expression regulation (Deplancke et al. 2016; Reilly and Noonan 2016; Long et al. 2016; Lambert et

al. 2018; Agrawal et al. 2018; Xu et al. 2022). This dissertation is primarily focused on diversity of the 3D conformation of the genome and impacts this may have on gene regulation.

Transcription factors. Transcription factors (TFs) are proteins that bind to specific DNA sequences within regulatory elements to control gene expression. TFs can act as activators, enhancing the recruitment of the transcriptional machinery, or as repressors, inhibiting transcription. The activity of TFs is modulated by interactions with co-factors and other proteins, forming complexes that influence transcriptional activity. The binding of TFs to regulatory elements is a dynamic process that responds to cellular signals and environmental changes (Deplancke et al. 2016; Reilly and Noonan 2016; Lambert et al. 2018). TFs bind to the regulatory sequences described below.

Promoters. Promoters are DNA sequences where RNA polymerase binds to initiate the transcription of a gene. They are typically located upstream of the transcription start site and contain specific motifs recognized by transcription factors and RNA polymerase (Reilly and Noonan 2016; Chatterjee and Ahituv 2017).

Enhancers. Enhancers are regulatory sequences that can be located far from the gene they regulate. They enhance the transcriptional activity of promoters by looping through the 3D genome architecture to come into proximity with their target promoters. Enhancers contain binding sites for transcription factors that facilitate the recruitment of the

transcriptional machinery (Reilly and Noonan 2016; Chatterjee and Ahituv 2017; Dean et al. 2021).

Silencers. Silencers are DNA elements that can repress the transcription of a gene. They function by binding transcriptional repressors, which inhibit the recruitment of the transcriptional machinery or promote the formation of repressive chromatin structures (Segert et al. 2021; Pang et al. 2023).

Insulators. Insulators are DNA elements that function in two main ways: as enhancer blockers and as chromatin barriers. Enhancer blockers prevent enhancers from activating promoters when positioned between them, ensuring that gene regulation remains specific and insulated from nearby regulatory elements. Chromatin barriers, also known as chromatin boundaries, prevent the spread of heterochromatin, thereby maintaining distinct chromatin domains. These two functions of insulators are mediated through largely distinct molecular pathways, reflecting the complexity of their roles in genome organization and gene regulation (Bell et al. 2001; Raab and Kamakaka 2010; Ghirlando et al. 2012; Phillips-Cremins and Corces 2013).

Chromatin modifications. Chromatin modifications, such as histone modifications and DNA methylation, play a crucial role in regulating gene expression by altering chromatin structure and accessibility. Histone modifications, including methylation and acetylation, can either promote a more open chromatin structure, facilitating transcription, or lead to a more compact structure, repressing transcription. DNA methylation typically acts as a

repressive mark that inhibits transcription factor binding and promotes the formation of repressive chromatin (Reilly and Noonan 2016; Chatterjee and Ahituv 2017; Millán-Zambrano et al. 2022).

Chromatin remodeling. Chromatin remodeling complexes are proteins that reposition or restructure nucleosomes, making DNA more or less accessible to transcription factors and other regulatory proteins. These complexes use energy from ATP hydrolysis to slide, eject, or restructure nucleosomes, thereby modulating the accessibility of DNA to the transcriptional machinery (Ghirlando et al. 2012; Nodelman and Bowman 2021; Kamat et al. 2023).

Disease implications of regulatory disruption

Mutations in regulatory elements, such as enhancers and promoters, can lead to misregulation of gene expression, contributing to various diseases. For example, mutations in enhancers have been linked to cancer, cardiovascular diseases, and neurodevelopmental disorders (Sanyal et al. 2012; Chatterjee and Ahituv 2017; Maurya 2021). The disruption of these regulatory elements can cause aberrant gene expression, resulting in disease phenotypes. Recent studies have identified non-coding mutations associated with increased disease susceptibility and severity, underscoring the critical role of regulatory elements in maintaining normal cellular function (Quang et al. 2015; Zhang and Lupski 2015; Chatterjee and Ahituv 2017). GWAS have identified numerous non-coding variants associated with complex diseases. These variants often lie in

regulatory regions, highlighting the importance of regulatory elements in disease etiology (Maurano et al. 2012; Finucane et al. 2015).

3D chromatin conformation in gene regulation

Techniques for studying 3D genome structure. Chromosome Conformation Capture (3C) was the first method developed to study the 3D structure of chromatin. It captures a single chromatin interaction by crosslinking interacting DNA segments, digesting the DNA with a restriction enzyme, ligating the interacting fragments, and then using PCR to detect the ligated products. While 3C provided the first glimpse into chromatin interactions, it is limited to analyzing a single interaction at a time, making it low-throughput and not suitable for genome-wide studies (Dekker et al. 2002).

Circular Chromosome Conformation Capture (4C) expanded upon 3C by allowing the detection of all chromatin interactions involving a specific genomic locus. In 4C, after the initial 3C procedure, the ligated DNA is further digested and circularized, enabling the use of inverse PCR to amplify interactions. This technique provides a comprehensive view of the chromatin environment around the focal locus, making it more informative than 3C for studying chromatin organization around specific loci (Simonis et al. 2006; Zhao et al. 2006).

Chromosome Conformation Capture Carbon Copy (5C) further extended the capabilities of 3C by enabling the detection of multiple chromatin interactions simultaneously. 5C uses ligation-mediated amplification with a set of designed primers to capture and amplify interactions between multiple loci. This method allows for higher-

throughput analysis of chromatin interactions, providing detailed interaction maps within selected genomic regions (Dostie et al. 2006).

Hi-C represents a significant advance in the study of chromatin interactions by enabling genome-wide analysis of chromatin conformation. Hi-C captures interactions across the entire genome by crosslinking chromatin, digesting the DNA, ligating the interacting fragments, and sequencing the resulting chimeric molecules. The resulting data are represented as a contact matrix, showing interaction frequencies between all pairs of loci in the genome. Hi-C has revealed the presence of TADs, chromatin loops, and nuclear compartments, providing a comprehensive view of genome organization (Lieberman-Aiden et al. 2009).

Several advanced versions of Hi-C have been developed to increase resolution or focus on specific aspects of genome organization. Micro-C improves the resolution of Hi-C by using micrococcal nuclease instead of restriction enzymes to digest chromatin, allowing for finer mapping of chromatin interactions at the nucleosome level (Hsieh et al. 2015). HiChIP, on the other hand, combines Hi-C with ChIP-seq to enrich for interactions involving specific proteins, such as transcription factors or histone modifications, providing targeted insights into protein-mediated chromatin interactions (Mumbach et al. 2016). These technological advancements have expanded our understanding of chromatin architecture and its role in gene regulation, allowing researchers to study genome organization at increasingly detailed levels.

Introduction to genome architecture. The organization of chromatin within the nucleus is fundamental to its function and regulation (Dekker and Mirny 2016; Ibrahim and Mundlos

2020). Chromatin is not randomly arranged but is structured into distinct, higher-order features that facilitate its various biological roles (Felsenfeld et al. 1996; Dixon et al. 2016). The 3D structure of the genome is crucial for the regulation of gene expression (Ibrahim and Mundlos 2020). One of the key roles of 3D genome organization is to facilitate interactions between regulatory elements, such as enhancers, silencers, and insulators, and their target genes. These interactions are essential for the precise control of gene expression in response to developmental cues and environmental signals (Dean et al. 2021).

The 3D folding of chromatin brings enhancers into proximity with their target promoters, allowing transcription factors and coactivators bound to the enhancer to interact with the transcriptional machinery at the promoter (Dekker and Mirny 2016; Robson et al. 2019). Similarly, the 3D genome structure can segregate elements neighboring insulators into distinct domains, ensuring that enhancers do not activate unintended genes and that silencers effectively repress their targets. This spatial organization is critical for maintaining proper gene expression patterns and preventing aberrant transcription (Ghirlando et al. 2012; Ghirlando and Felsenfeld 2016). Three key levels in this organization are chromatin loops, TADs, and nuclear compartments.

Chromatin loops. Chromatin loops are formed and stabilized by protein complexes, such as cohesin and CCCTC-binding factor (CTCF), which create physical connections between distant genomic regions. The formation of chromatin loops is crucial for regulating gene expression by enabling or preventing the interaction between regulatory elements and gene promoters (Tolhuis et al. 2002; Nora et al. 2017; Grubert et al. 2020;

Misteli 2020; Jerkovic´ and Cavalli 2021). Cohesin plays a central role in loop formation through a process known as loop extrusion, where the cohesin complex moves along the chromatin fiber, extruding a loop until it is halted by CTCF, which binds to specific DNA motifs and acts as a barrier (Sanborn et al. 2015; Fudenberg et al. 2016). This mechanism is critical in the formation of topologically associating domains (TADs), which are essentially collections of chromatin loops that create self-interacting genomic regions. However, other mechanisms also contribute to chromatin looping, such as homotypic interactions between transcription factors and adaptor proteins. For example, the LIM domain-binding protein 1 (LDB1) can mediate chromatin looping through its interaction with multiple transcription factors, bringing distant regulatory elements into proximity without involving cohesin or CTCF (Deng et al. 2012; Krivega et al. 2014).

TADs. TADs are segments of the genome that interact more frequently with each other than with adjacent regions, creating distinct 3D neighborhoods. TADs are not distinct from chromatin loops but rather represent an average of multiple loops created by cohesin and CTCF (Dixon et al. 2012). TADs help organize the 3D structure of the genome, ensuring that regulatory elements interact with their appropriate target genes while preventing inappropriate interactions (Ibrahim and Mundlos 2020; Misteli 2020; Acemel and Lupiáñez 2023). Within TADs, interactions between enhancers and promoters are more frequent, while interactions between regions in different TADs are relatively rare. This compartmentalization ensures that regulatory elements predominantly influence genes within the same TAD, contributing to the specificity and robustness of gene regulation. Disruptions in TAD boundaries can lead to changes in gene expression and are

associated with various diseases, including cancer, congenital disorders as well as common disease (Nora et al. 2012; Dixon et al. 2012; Sauerwald et al. 2020; McArthur and Capra 2021). The boundaries of TADs are often marked by binding sites for CTCF and are conserved across cell types and species, highlighting their functional importance in genome regulation (Dixon et al. 2012; Ibrahim and Mundlos 2020; Misteli 2020 p. 333; McArthur and Capra 2021; Acemel and Lupiáñez 2023).

Compartmentalization. Compartmentalization contributes to the spatial organization of the genome, influencing gene expression patterns (Xiong and Ma 2019; Kim et al. 2020; Kamat et al. 2023).. The genome is segregated into active (A compartment) and inactive (B compartment) regions, reflecting differences in gene density, chromatin accessibility, and transcriptional activity. A compartments, which are gene-rich and transcriptionally active, tend to be located toward the interior of the nucleus. In contrast, B compartments are gene-poor, transcriptionally inactive, and are more frequently associated with the nuclear lamina, a structure at the nuclear periphery that anchors chromatin to the nuclear envelope. This association with the nuclear lamina helps to create a repressive environment, often seen in lamina-associated domains, which contribute to the radial positioning of heterochromatin and late-replicating DNA in the periphery of the nucleus (Briand and Collas 2020; Keough et al. 2020; Kamat et al. 2023). Furthermore, the compartmentalization of the genome is influenced by subnuclear structures and phase-separated domains, which can be identified by methods such as Spatial Position Inference of the Nuclear genome (SPIN) (Wang et al. 2021). SPIN states help to delineate nuclear compartments based on their spatial positioning relative to nuclear structures like

speckles and the lamina. For example, certain SPIN states correspond to regions near nuclear speckles, which are more active, while others correspond to regions near the lamina, which are more repressive (Wang et al. 2021).

Cell-type-specificity. These regulatory mechanisms are often cell-type specific, reflecting the unique transcriptional needs and environmental responses of different cell types (Kim-Hellmuth et al. 2020; Qiu et al. 2021). Understanding how chromatin modifications and 3D contacts contribute to cell-type-specific gene regulation is essential for elucidating the complexities of cellular function and identity (Zheng and Xie 2019; Song et al. 2020; Liu et al. 2023). The interplay between these organizational features is crucial for maintaining genome integrity and regulating gene expression. Disruptions in chromatin organization can lead to mis-regulation of genes and contribute to various diseases, including cancer, developmental disorders, and other genetic conditions.

Integrating 3D genome data with phenotypic data. Linking 3D genome data with phenotypic outcomes involves integrating chromatin interaction maps with gene expression profiles and phenotypic data. This can be achieved through various approaches, such as eQTL mapping, GWAS, and multi-omics approaches.

eQTL mapping identifies genetic variants that influence gene expression levels. By combining eQTL data with 3D genome maps, researchers can determine whether chromatin interactions affect the regulation of genes associated with specific traits. This integration helps to identify the regulatory architecture underlying complex traits and diseases, as it links genetic variants to their target genes through chromatin interactions.

Studies have shown that many eQTLs are located at or near chromatin interaction sites, emphasizing the importance of 3D genome organization in gene regulation (Javierre et al. 2016; The GTEx Consortium 2020).

GWAS identifies genetic variants associated with phenotypic traits across large populations. Integrating 3D genome data into GWAS enhances the understanding of how non-coding variants affect gene regulation through chromatin interactions. For example, GWAS3D uses multiple genome-wide datasets to connect genetic variants with underlying regulatory mechanisms by analyzing high-dimensional chromatin interactions. This integration provides insights into the regulatory roles of SNPs and helps elucidate the mechanisms by which genetic variants contribute to phenotypic diversity and disease susceptibility (Li et al. 2013, 2022a).

Another approach involves fine mapping with epigenetic information and 3D structure. Fine mapping using 3D chromatin interaction data is a powerful approach to pinpoint causal variants within regions of linkage disequilibrium. This method overlays GWAS data with chromatin interaction data to identify how genetic variants affect regulatory elements such as promoters and enhancers. This approach helps to identify tissue-specific regulatory mechanisms, and the causal variants associated with diseases. For example, using 3D chromatin maps, researchers have identified interactions between disease-related genes and enhancers, aiding in the fine mapping of complex traits and diseases (Hu et al. 2021b; Li et al. 2022b).

Integrating 3D genome data with other omics data, such as transcriptomics, epigenomics, and proteomics, provides a more comprehensive view of gene regulation. These multi-omics approaches can reveal how chromatin interactions influence gene

expression, epigenetic modifications, and protein production, thereby linking 3D genome organization to phenotypic outcomes (Dekker et al. 2013; Dekker and Mirny 2016; Ibrahim and Mundlos 2020). Below is a selection of specific examples where 3D changes have been linked to phenotype and disease.

Limb malformations. Disruptions in TAD boundaries have been linked to limb malformations (Boyling et al. 2022; Weischenfeldt and Ibrahim 2023). For instance, structural variants that alter TAD boundaries can lead to ectopic enhancer-promoter interactions, resulting in misexpression of genes critical for limb development. This misregulation can cause congenital limb malformations (Lupiáñez et al. 2015).

Cancer. Chromatin loops involving oncogenes have been associated with various cancers (Boyling et al. 2022; Weischenfeldt and Ibrahim 2023). For example, the formation of new chromatin loops can bring enhancers in proximity to oncogenes, driving their overexpression and contributing to tumorigenesis. Understanding these interactions provides insights into the mechanisms of cancer development and potential therapeutic targets (Hnisz et al. 2016; Jablonski et al. 2021).

Neurodevelopmental disorders. Changes in 3D genome organization have also been implicated in neurodevelopmental disorders (Sánchez-Gaya et al. 2020; Hu et al. 2021a; Boyling et al. 2022; Weischenfeldt and Ibrahim 2023). For example, alterations in chromatin architecture at loci containing neurodevelopmental genes can affect their expression, leading to disorders such as autism and intellectual disability (Werling et al.

2016). These findings emphasize the importance of 3D genome studies in understanding the genetic basis of complex neurological conditions.

Conclusion

The study of 3D genome folding has unveiled critical insights into the structural organization of chromatin and its profound implications for gene regulation and phenotypic expression. The intricate architecture, comprising chromatin loops, TADs, and nuclear compartments, orchestrates the spatial proximity of regulatory elements and their target genes, thereby fine-tuning gene expression in response to various biological cues. Experimental techniques such as Hi-C and its variants, along with integrative computational approaches combining 3D genome data with eQTL mapping and GWAS, have significantly enhanced our understanding of the functional landscape of the genome.

Studies highlighting the role of 3D genome changes in limb malformations, cancer, and neurodevelopmental disorders further emphasize the crucial link between genome architecture and health. The ongoing integration of 3D genome data with multi-omics approaches promises to provide deeper insights into the molecular mechanisms underlying complex traits and diseases, paving the way for innovative therapeutic strategies and personalized medicine.

The continued exploration of 3D genome folding not only enriches our comprehension of genetic regulation but also holds the potential to unlock new frontiers in biomedical research, fostering a holistic understanding of the genome's dynamic and multifaceted nature.

Evolution of Gene Regulatory Functions

Evolutionary conservation of regulatory elements

Comparative genomics is a powerful approach used to identify conserved regulatory elements across different species. By comparing genomes, researchers can pinpoint sequences that have been preserved throughout evolution, suggesting they play crucial roles in essential functional processes, like gene regulation. This method helps to understand the evolutionary pressures acting on regulatory elements and provides insights into their functions. It is important to distinguish between sequence conservation and functional conservation. While some regulatory elements show high sequence conservation, others may exhibit conservation at the functional level despite sequence divergence. For instance, enhancers that regulate the expression of key developmental genes can maintain their regulatory functions even if their sequences evolve (Woolfe et al. 2004; Pennacchio et al. 2006).

Several techniques are employed in comparative genomics to identify conserved regulatory elements. Multiple sequence alignment involves aligning DNA sequences from different species to identify regions of similarity. Conserved sequences are often indicative of important regulatory elements. For example, the use of tools like MULTIZ and Clustal Omega allows researchers to perform high-quality multiple sequence alignments, revealing conserved regions that may function as enhancers or promoters (Blanchette et al. 2004; Pollard et al. 2006a; Sievers et al. 2011). In the UCSC Genome Browser, multiple sequence alignment methods are commonly used to compare genomes across different species. The browser utilizes tools like MULTIZ, TBA (Threaded Blockset Aligner), and

Cactus for genome alignments (Blanchette et al. 2004; Armstrong et al. 2020; Nassar et al. 2022). MULTIZ aligns multiple genomes by iteratively combining pairwise alignments and refines these alignments into blocks of highly conserved sequences (Blanchette et al. 2004). These blocks are then visualized to highlight conserved regulatory elements across species, aiding in the identification of functionally important genomic regions.

Multiple sequence alignment is an early step in identifying regions of similarity across species, but the recognition of conserved regulatory elements often requires further analysis using more specialized tools. As the field has developed, additional tools have emerged to build on the results of multiple sequence alignments. Methods such as PhyloP and PhastCons, part of the PHAST (Phylogenetic Analysis with Space/Time models) suite, have become instrumental in detecting evolutionary conservation across genomes, particularly by assessing conservation across individual nucleotide positions and entire genomic regions. These tools help to identify functional elements by distinguishing them from neutrally evolving sequences (Siepel et al. 2005; Pollard et al. 2010).

Conserved regulatory elements. Recent advancements have improved our ability to identify conserved non-coding elements (CNEs). It is also important to recognize that many conserved elements identified in genome-wide studies are exons, which are critical not only for protein coding but can also have regulatory functions (Margulies et al. 2007). Early work demonstrated the significance of CNEs in regulatory functions through comparative genomics, showing their importance in gene regulation and evolutionary conservation (Bejerano et al. 2004; Siepel et al. 2005). The Zoonomia Consortium has

expanded on these efforts, using large-scale comparative genomics to identify and analyze CNEs across multiple species, providing further insights into how these conserved regions contribute to regulatory landscapes (Genereux et al. 2020).

Phylogenetic footprinting identifies regulatory elements by comparing these non-coding regions across multiple species. Regions that are highly conserved are likely to have regulatory functions (Blanchette and Tompa 2002; Ganley and Kobayashi 2007). Studies using phylogenetic footprinting have successfully identified conserved enhancers in the genomes of mammals, birds, and fish (Tagle et al. 1988; Loots et al. 2000). More recent advancements, such as the integrative framework for phylogenetic footprinting, have enhanced the accuracy and applicability of this method by optimizing orthologous data selection and reducing false positives (Glenwinkel et al. 2014; Liu et al. 2016).

Promoters with conserved sequences and regulatory functions are also common. The promoter regions of housekeeping genes, which are required for basic cellular functions, tend to be highly conserved. For example, the promoter of the ribosomal RNA (rRNA) gene is conserved across different species, reflecting its essential role in ribosome biogenesis (Paule and White, 2000).

Many enhancers are conserved across species, indicating their essential regulatory roles. For example, the Sonic hedgehog (Shh) limb enhancer is highly conserved among vertebrates and plays a critical role in limb development (Lettice et al. 2002). Similarly, the even-skipped (eve) stripe 2 enhancer is conserved between *Drosophila* species and is crucial for proper segmentation during embryogenesis (Ludwig et al. 1998). Enhancers from the sponge *Amphimedon queenslandica* were shown to drive consistent, cell type-specific gene expression in zebrafish and mouse embryos,

despite the lack of sequence similarity. This study highlights the ancient and conserved nature of enhancer regulatory codes across the animal kingdom (Wong et al. 2020). An analysis of conserved non-coding elements in the genomes of 29 mammals identified thousands of conserved enhancers. While many of these enhancers are involved in developmental processes and exhibit high sequence conservation, there is significant turnover of enhancers between species. This turnover highlights the dynamic evolutionary changes in regulatory landscapes, suggesting that some enhancers are crucial for regulating key developmental genes, while others evolve rapidly to meet species-specific regulatory needs (Villar et al. 2015).

Divergence of regulatory elements. Regulatory element divergence plays a role in adaptation and survival. Changes in regulation can contribute to fitness and environmental adaptability. For instance, regulatory elements diverge between humans and rhesus macaques due to changes in both *cis* and *trans*, contributing to species-specific regulatory programs and adaptation (Hansen et al. 2023). Other studies have used CRISPR-based screens to map the functional impact of non-coding regulatory elements, highlighting their role in immune cell function and disease susceptibility (Catalinas et al. 2023). Other studies have shown that changes in regulatory regions can result in observable traits such as limb development, pigmentation, and morphological variations (Carroll 2008). For example, the divergence of regulatory elements has been linked to changes in the expression of genes involved in the development of specific morphological traits in various species (Wong et al. 2020).

Understanding how changes in regulatory elements drive species-specific traits and adaptations is crucial for elucidating the mechanisms behind evolutionary processes. These changes can provide insights into how different species develop unique characteristics and adapt to their environments. For example, human accelerated regions (HARs) are sequences that are conserved in vertebrates but show rapid evolution in humans. These regions, such as HAR2 are potentially involved in human-specific development in the limbs and brain (Pollard et al. 2006c, 2006b; Bird et al. 2007; Prabhakar et al. 2008; Capra et al. 2013). Recent studies have confirmed that HARs often function as enhancers that regulate neurodevelopmental genes, playing a significant role in human brain evolution (Capra et al. 2013; Whalen and Pollard 2022; Whalen et al. 2023). These regions have been found to influence the expression of genes involved in neurogenesis and synaptic transmission, thereby contributing to the development of human-specific cognitive traits (Girskis et al. 2021; Whalen et al. 2023; Keough et al. 2023; Pal et al. 2024).

Human-specific deletions in conserved non-coding regions (hCONDELs) have been shown to affect regulatory elements that are conserved in other species, contributing to human-specific traits (McLean et al. 2011). Recent work has further explored these hCONDELs by examining conserved regions across diverse vertebrate genomes. They showed that hCONDELs are enriched in regions associated with neuronal and cognitive functions. Functional assessments revealed that many hCONDELs perturb transcription factor-binding sites in active enhancers, with some inducing gains of regulatory activity, particularly in genes related to neurodevelopment (Xue et al. 2023).

Mechanisms of regulatory evolution

Sequence motifs. Evolutionary changes in transcription factor binding sites can significantly alter regulatory activity. The gain or loss of binding sites contributes to gene expression divergence, which can contribute to phenotypic variations among species (Dowell 2010; Diehl and Boyle 2018). Transcription factor binding sites in mammalian genomes exhibit both evolutionary conservation and divergence, reflecting their critical roles in maintaining and modifying gene regulatory networks (Schmidt et al. 2010; Villar et al. 2015).

Chromatin accessibility. Chromatin dynamics contribute to regulatory evolution, as open chromatin regions correlate with active regulatory elements. However, closed chromatin regions that change across species also play a significant role, potentially indicating regions of repressive regulation that have evolved differently in various lineages (Gao et al. 2018; Peng et al. 2019). Variations in chromatin accessibility across different tissues and developmental stages provide insights into the role of chromatin dynamics in regulatory element function and evolution (Buenrostro et al. 2013; Corces et al. 2017; Gao et al. 2018). For instance, studies have shown that changes in chromatin accessibility in brain tissue between humans and chimpanzees are linked to differences in enhancer activity, which may contribute to species-specific cognitive traits (Prescott et al. 2015). Additionally, research comparing the chromatin landscapes of various mammalian tissues has demonstrated that shifts in accessibility are often associated with species-specific adaptations and evolutionary innovations (Degner et al. 2012; Shibata et al. 2012).

Epigenetic modifications. Epigenetic changes to DNA methylation and histone modifications are essential mechanisms that impact regulatory element activity. These modifications are associated with the underlying DNA sequences and influence gene expression across generations, playing a critical role in evolutionary processes. DNA methylation typically represses gene activity by adding methyl groups to cytosine residues, leading to a closed chromatin conformation. Histone modifications, such as acetylation and methylation, can either activate or repress gene expression depending on the specific modification and its context. The dynamic landscape of DNA methylation in the human genome reveals extensive variation in methylation patterns across different cell types and developmental stages. Comprehensive maps of human epigenomes highlight the evolutionary significance of epigenetic modifications in regulating gene expression and contributing to phenotypic diversity (Ziller et al. 2013; Roadmap Epigenomics Consortium et al. 2015). Recent studies have expanded our understanding of these mechanisms, demonstrating their roles in the evolution of regulatory elements in response to environmental factors and developmental processes (Reilly and Noonan 2016).

Regulatory variation in diverse human populations

In addition to understanding how regulatory changes drive differences between species, it is equally important to study how such variations occur within a species. GWAS have identified numerous regulatory variants associated with complex traits and diseases, such as height, obesity, and coronary artery disease (Chatterjee and Ahituv 2017). These

variants often lie in non-coding regions and influence phenotypes by affecting gene expression. For instance, variants in the FTO gene are associated with obesity, while those in the 9p21 region are linked to coronary artery disease, despite the absence of any known genes in that region (Maurano et al. 2012; Zhang and Lupski 2015; Finucane et al. 2015).

Different populations exhibit unique regulatory variants that contribute to phenotypic diversity. A notable example is the TRPM8 gene variant, which helps North Europeans adapt to cold temperatures but also increases susceptibility to certain conditions. This variant illustrates how population-specific regulatory elements can drive adaptation to local environments while influencing health (Chun and Fay 2011; Key et al. 2018).

Recent studies on diverse human populations have uncovered regulatory variants that influence chromatin accessibility and gene expression. These findings highlight the importance of genetic diversity in understanding phenotypic variation. For example, fine-mapping studies of cis-regulatory variants in different populations have revealed how these variants contribute to phenotypic traits and disease susceptibility (Degner et al. 2012; Tehrani et al. 2019).

Disease implications of regulatory disruption

Mutations in regulatory elements, such as enhancers and promoters, can lead to misregulation of gene expression, contributing to various diseases. For example, mutations in enhancers have been linked to cancer, cardiovascular diseases, and

neurodevelopmental disorders (Sanyal et al. 2012; Chatterjee and Ahituv 2017; Maurya 2021). The disruption of these regulatory elements can cause aberrant gene expression, resulting in disease phenotypes. Recent studies have identified non-coding mutations associated with increased disease susceptibility and severity, underscoring the critical role of regulatory elements in maintaining normal cellular function (Quang et al. 2015; Zhang and Lupski 2015; Chatterjee and Ahituv 2017). GWAS have identified thousands non-coding variants associated with complex diseases. These variants often lie in regulatory regions, highlighting the importance of regulatory elements in disease etiology (Maurano et al. 2012; Finucane et al. 2015).

Impact of evolution on 3D genome structure

The 3D structure of the genome is shaped by evolutionary forces that act on both sequence and structural features, leading to conservation and divergence in genome folding patterns across species. Understanding these evolutionary changes provides insights into the functional constraints and adaptive flexibility of chromatin organization.

Several studies have highlighted the role of 3D genome organization in evolutionary processes. For instance, in the cotton tribe (Gossypieae), genomic rearrangements have been linked to changes in 3D chromatin topologies, influencing gene regulation and species-specific traits (Li et al. 2022a). Similarly, research on vertebrates has shown that differences in chromosome length and organization can affect 3D genome structure, impacting the frequency of long-range chromatin interactions and gene regulation (Li et al. 2022a).

Comparative genomics studies have revealed that certain aspects of 3D genome organization, such as TADs and specific chromatin loops, are highly conserved across species. For example, many TAD boundaries are preserved between humans and mice, suggesting that these structural features play essential roles in regulating gene expression and maintaining genome integrity. The conservation of these elements implies strong evolutionary constraints, likely due to their critical function in coordinating regulatory interactions and gene expression (Dixon et al. 2012). TAD boundaries that are stable across multiple cell types have been shown to be evolutionarily constrained and enriched for heritability. These stable TAD boundaries are associated with higher levels of CTCF binding and are enriched for housekeeping genes, suggesting their crucial role in maintaining genome function. This stability and heritability underscore the functional importance of these boundaries and their role in genome organization and regulation (McArthur and Capra 2021).

Despite the conservation of certain elements, there are also significant differences in genome folding patterns between species. These differences can arise from species-specific regulatory needs and adaptations. For instance, variations in enhancer-promoter interactions and the arrangement of regulatory elements reflect the unique evolutionary pressures and functional requirements of different organisms. These species-specific features highlight the adaptive potential of genome architecture and its role in facilitating diverse regulatory landscapes (Li et al. 2022a). A study on HARs, which are conserved genomic loci that evolved rapidly in the human lineage, reveals significant insights into genome folding and regulatory evolution. Using deep learning combined with chromatin capture experiments, researchers discovered that HARs are enriched in TADs with

human-specific genomic variants that alter 3D genome organization. This rewiring of regulatory interactions between HARs and neurodevelopmental genes suggests that enhancer hijacking may explain the rapid evolution of HARs, contributing to human-specific traits (Keough et al. 2023).

DNA Sequence-Based Machine Learning

Historical context and motivation

Early attempts to use machine learning in genomics faced significant challenges due to limited data and computational power. Initial models were often simplistic and struggled with the complexity of genomic data. Tasks such as predicting gene function or expression levels and protein structure were limited by the availability of using smaller models and datasets. Initial applications included the use of Hidden Markov Models (HMMs) for gene prediction and identifying protein secondary structures, as well as naive Bayes classifiers and two-layer Bayesian neural networks (BNNs) that utilized sequence-based features (Krogh et al. 1994; Baldi and Brunak 2001; Ding et al. 2012). Another early application involved the identification of transcription start sites and splice sites, which demonstrated the potential of machine learning to handle genomic data (Burge and Karlin 1997; Degroeve et al. 2002). The use of support vector machines (SVMs) and neural networks for these tasks further highlighted the growing utility of machine learning in genomics (Furey et al. 2000; Zhang 2002; Sonnenburg et al. 2006). The development of ensemble methods and decision trees for classifying gene expression data also played a significant role in early machine learning applications (Breiman 2001; Huang et al. 2010; Libbrecht

and Noble 2015; Angermueller et al. 2016), further laying the groundwork for more advanced techniques.

The motivations for using machine learning in genomics were driven by several factors. The explosive growth in genomic data, particularly from projects like the Human Genome Project and various GWAS, created a need for advanced computational tools to analyze this type of data effectively (Collins and Fink 1995; Sollis et al. 2023). Machine learning provided a means to improve predictions of gene expression, identify regulatory elements, and understand the effects of genetic variants on phenotypes. Advances in computational power, algorithm development, and the availability of large annotated genomic datasets further supported the integration of machine learning into genomics, enabling researchers to tackle previously intractable problems (Libbrecht and Noble 2015; Angermueller et al. 2016). The high-dimensionality and complexity of genomic data, which traditional statistical methods struggled to manage, necessitated the use of more sophisticated machine learning approaches (Libbrecht and Noble 2015; Angermueller et al. 2016). Furthermore, integrating various types of genomic data, such as sequence data, expression data, and epigenetic data, became essential to provide a comprehensive understanding of gene regulation and function (Ritchie et al. 2015; Libbrecht and Noble 2015). There was also a growing recognition that non-coding regions of the genome play critical roles in gene regulation, requiring advanced tools to decipher their functions (ENCODE Project Consortium 2012; Roadmap Epigenomics Consortium et al. 2015; Davis et al. 2018; ENCODE Project Consortium et al. 2020).

Types of machine learning models

Supervised learning involves training models on labeled datasets where each input is paired with a known output. This approach is commonly used for tasks such as classification and regression, where the goal is to predict specific outcomes based on input features. In genomics, supervised learning is applied to tasks like predicting transcription factor binding sites and identifying regulatory elements (Zeng et al. 2016; Schrider and Kern 2018; Smith et al. 2023). Common algorithms used in supervised learning for genomics include decision trees, K-nearest neighbors, SVMs, logistic regression, and neural networks (Libbrecht and Noble 2015; Angermueller et al. 2016).

Convolutional neural networks (CNNs) are a type of neural network architecture that is particularly effective for supervised learning tasks, such as identifying patterns in DNA sequences that predict transcription factor binding sites and DNA-protein interactions (Zeng et al. 2016; Kelley et al. 2016, 2018). CNNs have also been used to predict 3D genome folding from DNA sequences, highlighting the importance of nucleotide-level features and enabling rapid *in silico* predictions (Fudenberg et al. 2020; Schwessinger et al. 2020).

In contrast, unsupervised learning involves training models on unlabeled datasets to uncover hidden patterns or structures without predefined labels. This method is used for tasks such as clustering, dimensionality reduction, and association. In genomics, unsupervised learning is used to cluster gene expression profiles and discover new regulatory motifs (Oyelade et al. 2016). Common unsupervised learning algorithms include K-means clustering, hierarchical clustering, PCA, and neural networks as autoencoders (Libbrecht and Noble 2015).

Applications in genomic regulation

Predicting gene expression. Understanding gene expression patterns is crucial for deciphering biological processes and disease mechanisms. Gene expression levels can reveal insights into cellular functions, tissue differentiation, and the molecular basis of diseases. Predicting gene expression can aid in the development of novel therapeutics by identifying target genes and pathways involved in disease. Moreover, accurate prediction models can facilitate personalized medicine by tailoring treatments based on individual gene expression profiles (Libbrecht and Noble 2015; Sasse et al. 2023).

Recent advances in predicting gene expression include the development of CNN-based models such as ExPecto, BigRNA, Xpresso, Basenji and Basenji2, which have significantly improved the accuracy of these predictions (Kelley et al. 2018; Zhou et al. 2018; Agarwal and Shendure 2020; Kelley 2020; Celaj et al. 2023). Enformer, another deep learning model, leverages both proximal and distal regulatory elements for gene expression prediction (Avsec et al. 2021a). Additionally, multi-task learning frameworks like MTM are capable of predicting individualized tissue-specific gene expression profiles, which enhances personalized medicine approaches (He et al. 2023).

Predicting regulatory mechanisms. Machine learning models have become invaluable tools in genomics, particularly for predicting regulatory mechanisms. These models, trained on epigenomic data, are adept at identifying enhancers and transcription factor binding sites, including genetic variants that impact chromatin accessibility and 3D conformation (Zeng et al. 2016; Kelley et al. 2016, 2018; Fudenberg et al. 2020; Schwessinger et al. 2020; Zhou 2021; Smith et al. 2023). Analyzing large datasets with

machine learning leads to the discovery of complex patterns and interactions that traditional methods might miss. This capability is crucial for understanding gene regulation at a detailed level, enabling researchers to predict how changes in DNA sequences can affect gene expression and ultimately influence phenotypes. Moreover, these predictions assist in identifying potential therapeutic targets and understanding disease mechanisms, making machine learning a powerful ally in precision medicine and functional genomics. This section includes descriptions of a non-exhaustive set of exciting applications of machine learning models to regulatory genetic data.

High-throughput methods like Massively Parallel Reporter Assays (MPRA) are frequently used for the functional validation of predicted regulatory elements, providing critical insights into enhancer activity. These assays allow for the parallel measurement of thousands of sequences, helping to identify which DNA elements can drive gene expression (Smith et al. 2023; Gosai et al. 2023). The Malinois CNN model exemplifies this application of machine learning. It predicts enhancer activity by modeling MPRA results, thereby enhancing the prediction of regulatory elements from sequence data (Gosai et al. 2023). Additionally, Malinois can design synthetic cis-regulatory elements, leveraging unique sequence syntax to promote activity in target cell types while reducing off-target effects. MPRA-DracoNN is another model that uses deep neural networks to predict the activity of regulatory elements from MPRA data, further demonstrating the utility of machine learning in understanding enhancer landscapes (Movva et al. 2019; Whalen et al. 2023; Deng et al. 2024). This integration of machine learning and high-throughput validation techniques highlights the power of combining computational and experimental approaches to decode the complexities of the genome.

Accurate prediction of transcription factor binding sites (TFBSs) is crucial for understanding gene regulation and its implications for cellular function and disease. Machine learning models, particularly those based on deep learning, have significantly advanced the prediction of TFBSs by leveraging large datasets and complex algorithms. Recent models have employed various innovative techniques to enhance the accuracy and interpretability of TFBS predictions. For example, DeepSTF integrates CNNs, an improved transformer encoder structure, and bidirectional long short-term memory (LSTM) networks. By combining DNA sequence and multiple types of DNA shape information, DeepSTF outperforms several state-of-the-art predictors in identifying TFBSs (Ding et al. 2023). Similarly, DNABERT-Cap, a transformer-based capsule network, uses bidirectional encoders to predict TFBSs. This model has shown high performance across multiple cell lines and provides robust cross-cell predictions, demonstrating its versatility and accuracy (Ji et al. 2021; Ghosh et al. 2024). Another example is DeepGenBind, a deep learning model specifically designed for predicting TF binding sites. It utilizes a dense neural network architecture to capture the complex interactions between TFs and DNA sequences, improving the understanding of transcriptional regulation (Wang et al. 2022).

Machine learning models have become pivotal in predicting chromatin accessibility and histone modifications, which are essential for understanding gene regulation. These models leverage large datasets from assay for transposase-accessible chromatin (ATAC)-seq and chromatin immune-precipitation (ChIP)-seq experiments to analyze the state of chromatin and predict how genetic variants can influence regulatory regions.

One notable model is the Basset framework, which uses CNNs to learn the regulatory code of the accessible genome. Basset has shown great accuracy in predicting chromatin accessibility across different tissues by analyzing sequence data (Kelley et al. 2016). Similarly, ExplainNN combines the expressiveness of CNNs with the interpretability of linear models, making it possible to predict TF binding, chromatin accessibility, and de novo motifs while providing insights into model predictions (Novakovsky et al. 2023).

Predicting 3D genome structure. In recent years, several machine learning methods for predicting 3D genome chromatin contact maps from sequence have been published. These include Akita and DeepC with similar CNN architectures to predict 1 Mb windows of contacts, and Orca with a hierarchical encoder/decoder structure to predict multi-range contacts from 1 to 256 Mb (Fudenberg et al. 2020; Schwessinger et al. 2020; Zhou 2021). Currently we have access large amounts of publicly available sequence data from hundreds of thousands of individuals, but only a handful of Hi-C maps. Thus, predicting Hi-C maps from sequence is a promising avenue to computationally close the gap in our understanding of 3D genome diversity. I use Akita, because it requires only sequence information as input, it has a flexible architecture, and I am supported by the original developers. Other methods for making cell-type specific predictions exist but they require additional genomic information beyond sequence. This approach enabled me to quantify the diversity of 3D genome organization and its contribution to both rare and common phenotypes, without the need for costly, time-consuming chromatin interaction assays

Recent advances

Accuracy and interpretability. Recent advancements in machine learning algorithms have significantly enhanced the accuracy and interpretability of DNA sequence-based models. The introduction of deep learning models, such as Enformer and BPNet, has revolutionized the field by providing more precise predictions of gene expression and transcription factor binding. Enformer uses a multi-scale architecture and showed advancements in capturing both local and distal regulatory elements, achieving state-of-the-art performance in predicting gene expression (Avsec et al. 2021a). Similarly, BPNet, which employs deep CNNs, has demonstrated superior accuracy in identifying transcription factor binding sites and elucidating regulatory grammar (Avsec et al. 2021b).

Improvements in the interpretability of machine learning models have been crucial for using machine learning in genomics. One notable development is the application of SHAP (SHapley Additive exPlanations) values, which provide a unified framework to interpret model predictions by assigning importance scores to input features (Lundberg and Lee 2017). This method allows researchers to pinpoint specific nucleotide sequences that influence gene expression predictions, thereby linking genetic variants to potential functional outcomes (Levy et al. 2020; Tasaki et al. 2020; Yap et al. 2021). Another significant improvement is the use of attention mechanisms in transformer-based models like Enformer. These mechanisms enable models to attend to relevant regions of the input sequence, offering insights into the regulatory elements that contribute to gene expression changes. By visualizing which parts of the sequence the model focuses on, researchers can better understand the underlying biological processes (Avsec et al. 2021a).

Generalizability. Recent advancements in the generalizability of machine learning models have significantly impacted the field of genomics. One key strategy has been the inclusion of diverse and large-scale datasets. A growing body of research has shown that models trained on data from a wide range of populations perform better across different genetic backgrounds. For instance, initiatives like the HGDP, 1KG and SGDP have provided valuable data that has been instrumental in training models capable of capturing the genetic diversity present across populations (Cavalli-Sforza 2005; 1000 Genomes Project Consortium et al. 2015; Mallick et al. 2016). These datasets have been crucial for developing models that can accurately predict genetic risk across different ancestries, which is essential for equitable healthcare (Wojcik et al. 2019).

Another approach to enhancing generalizability involves developing ancestry-aware models. These models are specifically designed to account for population stratification, which can bias predictions if not properly addressed. This approach addresses the long-standing issue of the underrepresentation of non-European populations in genetic studies and has shown promising results in improving prediction accuracy for diseases like Alzheimer's (Martin et al. 2017, 2019; Gyawali et al. 2023). DisPred, for example, separates ancestry from phenotype-specific information, improving prediction accuracy without requiring self-reported ancestry data (Gyawali et al. 2023). Applied to Alzheimer's disease genetics, DisPred outperformed existing models, demonstrating better prediction accuracy for minority populations. The framework's ability to handle individual-level heterogeneity highlights its potential for more equitable genetic risk assessments across diverse ancestries.

Recent advancements in transfer learning allow models trained on extensive datasets to be adapted for smaller, diverse datasets, improving their predictive performance. For instance, Enformer uses transfer learning to integrate both proximal and distal regulatory elements. By leveraging pre-trained representations on large-scale genomic datasets, Enformer captures complex long-range interactions and improves gene expression predictions across various cell types and conditions (Avsec et al. 2021a). This approach allows the model to apply learned features from one dataset to new, less annotated datasets effectively. Additionally, transfer learning has advanced genetic risk prediction across diverse populations. Recent studies have shown that transfer learning frameworks can leverage large European-ancestry dominated GWAS to pretrain and be fine-tuned to increase performance of genetic risk prediction in non-European cohorts (Tian et al. 2022; Zhao et al. 2022). These developments underscore the effectiveness of transfer learning in making genomic models more robust and adaptable across varied datasets and conditions.

Despite these advances, significant challenges remain in ensuring the generalizability of machine learning models in genomics. One of the key issues is the need for more comprehensive and inclusive datasets that capture the full spectrum of human genetic diversity. Additionally, there is a growing recognition of the importance of incorporating environmental and lifestyle factors into models to better understand gene-environment interactions. Future research is likely to focus on integrating multi-omics data and improving the interpretability of models to ensure they are not only accurate but also applicable across different populations and contexts.

Conclusion

The history of machine learning models in genomics has been marked by significant advancements, addressing early challenges of limited data and computational power. Initial models like Hidden Markov Models and naive Bayes classifiers laid the groundwork, demonstrating the potential of machine learning in genomics. Today, advanced models such as Enformer and BPNet offer high accuracy and interpretability, leveraging deep learning architectures and interpretability tools like SHAP values. Improvements in generalizability through diverse datasets and population-specific models ensure robust applications across varied genetic backgrounds. These developments collectively enhance our understanding of gene regulation and pave the way for breakthroughs in personalized medicine and functional genomics.

Conclusion

Innovation

I use machine learning to predict 3D chromatin conformation from genome sequence to understand its diversity in modern humans and across cell-types. I have generated 3D genome maps across much of the diversity of modern humans by leveraging a sequence-based prediction framework. My results establish the baseline distribution of 3D genome structure and variation. This allows for interrogation of potentially disease implicated divergences and inform interpretation of changes to the 3D genome during development. Beyond making predictions on observed variation, I also to selectively mutate individual base pairs to observe their impact on the predicted contact maps.

Significance

The study of 3D genome folding has unveiled critical insights into the structural organization of chromatin and its implications for gene regulation and phenotypic expression. The intricate architecture, comprising chromatin loops, TADs, and nuclear compartments, orchestrates the spatial proximity of regulatory elements and their target genes, thereby fine-tuning gene expression in response to various biological cues. Advanced techniques such as Hi-C and its variants, along with integrative approaches combining 3D genome data with eQTL mapping and GWAS, have significantly enhanced our understanding of the functional landscape of the genome.

Moreover, the evolutionary conservation and divergence of 3D genome structures across species underscore the adaptive potential and functional constraints of chromatin organization. Studies highlighting the role of 3D genome changes in limb malformations, cancer, and neurodevelopmental disorders further emphasize the crucial link between genome architecture and health. The ongoing integration of 3D genome data with multi-omics approaches promises to provide deeper insights into the molecular mechanisms underlying complex traits and diseases, paving the way for innovative therapeutic strategies and personalized medicine.

The continued exploration of 3D genome folding not only enriches our comprehension of genetic regulation but also holds the potential to unlock new frontiers in biomedical research, fostering a holistic understanding of the genome's dynamic and multifaceted nature.

CHAPTER 2: MACHINE LEARNING REVEALS THE DIVERSITY OF HUMAN 3D CHROMATIN CONTACT PATTERNS

Abstract

Understanding variation in chromatin contact patterns across diverse humans is critical for interpreting non-coding variants and their effects on gene expression and phenotypes. However, experimental determination of chromatin contact patterns across large samples is prohibitively expensive. To overcome this challenge, we develop and validate a machine learning method to quantify the variation in 3D chromatin contacts at 2 kilobase resolution from genome sequence alone. We apply this approach to thousands of human genomes from the 1000 Genomes Project and the inferred hominin ancestral genome. While patterns of 3D contact divergence genome-wide are qualitatively similar to patterns of sequence divergence, we find substantial differences in 3D divergence and sequence divergence in local 1 megabase genomic windows. In particular, we identify 392 windows with significantly greater 3D divergence than expected from sequence. Moreover, for 31% of genomic windows, a single individual has a rare divergent 3D contact map pattern. Using *in silico* mutagenesis we find that most single nucleotide sequence changes do not result in changes to 3D chromatin contacts. However, in windows with substantial 3D divergence just one or a few variants can lead to divergent 3D chromatin contacts without the individuals carrying those variants having high sequence divergence. In summary, inferring 3D chromatin contact maps across human populations reveals variable contact patterns. We anticipate that these genetically diverse maps of 3D chromatin contact will

provide a reference for future work on the function and evolution of 3D chromatin contact variation across human populations.

Introduction

Genetic and transcriptomic variation within and between human populations is extensive, and much of the phenotypic diversity across humans is the result of non-protein-coding variation (Storey et al. 2007; Ho et al. 2008; Novembre et al. 2008; Alemu et al. 2014; 1000 Genomes Project Consortium et al. 2015; Mallick et al. 2016; ENCODE Project Consortium et al. 2020; The GTEx Consortium 2020). However, given the complex and incompletely understood control of gene regulation, linking non-coding variants to effects on gene expression and phenotypes remains challenging (Schipper and Posthuma 2022). Nonetheless, given the importance of variation in gene expression, quantifying the effects of non-coding genetic variants is key to advancing our understanding of gene regulation and disease.

The 3D spatial organization of chromosomes within the nucleus influences gene expression regulation through enhancer modulated transcription (Tolhuis et al. 2002; Tang et al. 2015). Quantifying 3D chromatin contact patterns has provided insights into chromatin structure and interactions within the nucleus (Dekker et al. 2017, 2023). For example, disruption of the structural organization and contacts of distal regulatory elements within the genome has been linked to complex diseases and genomic rearrangements, such as those observed in certain cancers (Roix et al. 2003; Zhang et

al. 2012; Maurano et al. 2012). Despite its importance, our knowledge of the breadth of 3D genome variation across humans is limited.

Previous studies have shown that 3D chromatin contact varies both within and among populations (McArthur et al. 2022; Li et al. 2023). However, experimental determination of chromatin interactions at large scale is expensive, especially at high enough spatial resolution to reveal differences in contacts between specific regulatory elements. This has limited the extent to which chromatin contact variation has been studied across individuals; often a single map is used to represent all individuals. Recent advances in machine learning methods have enabled the prediction of 3D genome chromatin contact maps from DNA sequences (Fudenberg et al. 2020; Schwessinger et al. 2020; Zhou 2021). These methods predict 3D chromatin contact based solely on sequence information, offering a promising approach to computationally study 3D genome variation.

In this study, we used Akita (Fudenberg et al. 2020), a convolutional neural network that requires only DNA sequence information as input, to predict 3D contact maps for 2,457 human individuals. We compared these contact maps between individuals and to the predicted map of an inferred ancestral hominin genome sequence. This revealed regions with significant divergence in 3D contact maps within and between populations. We found that 3D contact divergence genome-wide follows similar patterns as sequence divergence and that pressure to maintain 3D contact patterns has broadly constrained sequence evolution. However, 3D contact divergence is very different from sequence divergence at local scales (1 Mb and below) and is highest in regions under low functional and evolutionary constraint. We also used *in silico* mutagenesis to identify single

nucleotide variants with large effects on contact map variation. We find that rare 3D contact map divergence is often the result of a single large-effect variant, while common 3D divergence usually is influenced by multiple variants. Our results establish the baseline distribution of 3D chromatin contact and variation in diverse populations. They also provide context in which to interpret new human 3D chromatin contact data and the effects of variants identified in disease cohorts on 3D chromatin contact.

Results

Accurate prediction of 3D contact maps for diverse individuals

To quantify variation in the 3D genome of modern humans, we predicted chromatin contact maps for 2,457 unrelated individuals from the 1KG data (1000 Genomes Project Consortium et al. 2015) using Akita (**Figure 2.1**) (Fudenberg et al. 2020). As Akita was trained on the hg38 human reference genome and unphased Hi-C data, we generated “flattened” pseudo-haploid genome sequences for each individual by inserting all their single nucleotide variants (SNVs) into the hg38 human reference sequence (**Supplemental Figure 2.1**). Akita takes an approximately 1 Mb window of DNA sequence as input and outputs local 3D contact patterns for the input region at 2,048 bp resolution. We divided the genome into 1 Mb sliding windows, overlapping by half, and retained windows with 100% sequence coverage from the hg38 reference genome (N=4,873). We then used Akita to predict local chromatin contacts genome-wide for individuals from five continental populations and 26 sub-populations distributed across the globe defined by 1KG (1000 Genomes Project Consortium et al. 2015).

To confirm that Akita performs well on individuals from different continents, we compared Hi-C data from the 4D Nucleome Project (4DN) to predicted maps for 11 Africans, 2 Americans, 1 East Asian, and 1 European (Dekker et al. 2017). We focused on held-out windows from the Akita test set and scaled predictions to 10 kb resolution to be roughly comparable to the lower resolution of the experimental contact maps. The European individual (NA12878) was the basis for the GM12878 lymphoblastoid cell line which was used in the training of Akita. Its Hi-C library was also sequenced to the highest coverage (**Supplemental Figure 2.2**). Thus, it serves as an upper bound on the expected performance. Our predictions for the 11 African individuals were only slightly less accurate (mean Spearman's $\rho = 0.43$) than what was observed for Europeans ($\rho = 0.48$) (**Figure 2.1**), and these approach the similarity observed between replicates in Hi-C (Fudenberg et al. 2020). While the accuracy for the East Asian ($\rho = 0.37$) and American ($\rho = 0.36$) individuals is somewhat lower, we believe this is due to low resolution and sequencing depth of the available experimental maps for these individuals. Filtered read count (retrieved from 4DN Data Portal) correlated with Akita prediction accuracy (**Supplemental Figure 2.2**; $R^2 = 0.25$). Visual checks verify that the predicted and experimental contact maps share key patterns (**Figure 2.1**). These results confirm that Akita has learned to predict 3D contact maps in a way that is not specific to any single human or group and thus can be applied across individuals.

3D divergence differs from sequence divergence

To explore how changes in 3D chromatin contacts relate to DNA sequence changes, we quantified levels of 3D divergence in all windows across the genome between modern humans from 1KG and the inferred hominin ancestor. We then compared sequence divergence from the ancestral sequence with 3D divergence from the ancestral map for each window. Correlation between sequence and 3D divergence across individuals for a window is generally low (mean Spearman's $\rho = 0.113$), and varies greatly across windows ($SD = 0.20$) (**Figure 2.2, Supplemental Figure 2.3**). Genome-wide average sequence and 3D divergence were only moderately correlated across individuals (**Supplemental Figures 2.3, 2.4**; $R^2 = 0.31$ for 3D divergence, $R^2 = 0.34$ for MSE). This aligns with our expectation that most sequence variants do not impact 3D chromatin contacts, and it shows that in most windows sequence divergence is not a proxy for 3D divergence. These relationships between sequence and 3D divergence are maintained across window sizes from 1 Mb to 65 kb (**Supplemental Figure 2.3**).

African populations have the highest predicted 3D genome diversity

African individuals have higher sequence diversity than non-African individuals (1000 Genomes Project Consortium et al. 2015). We tested whether African individuals also have higher predicted 3D divergence from the ancestral state than in other individuals. We calculated 3D divergence from ancestral sequence for each window and took the mean across all genomic windows for each individual.

Mean genome-wide 3D divergence varies significantly among populations (**Figure 2.2, Supplemental Figure 2.4**; Kruskal-Wallis: $P = 2.34 \times 10^{-145}$ for 3D divergence, Kruskal-Wallis: $P = 3.96 \times 10^{-160}$ for MSE). African individuals have significantly greater mean 3D divergence (0.0045) than individuals from all other populations (post-hoc Conover: $P < 1.35 \times 10^{-57}$), and non-African populations have on average 5% lower 3D divergence. While this trend is consistent with patterns of sequence divergence, the size of the difference is smaller: non-African individuals have approximately 20% fewer SNVs on average (1000 Genomes Project Consortium et al. 2015).

Most variation in 3D chromatin contact patterns is shared across populations

To explore the similarity of 3D contact maps within and between humans from different 1KG populations, we hierarchically clustered individuals based on their pairwise 3D divergence from one another. Averaging 3D divergence over all 4,873 genomic windows, individuals largely clustered by population of origin (**Figure 2.2**). In contrast, clustering each window of the genome separately revealed patterns that did not follow global population relationships expected from sequence divergence. To summarize the patterns across windows, we computed the posterior probability of the tree derived from flattened sequence comparisons based on all of the window-specific 3D divergence trees using ASTRAL (Zhang et al. 2018; Rabiee et al. 2019). Branches leading to each modern 1KG population are not strongly supported, reflecting the sharing of contact patterns between individuals from different populations (**Figure 2.2**). In contrast, the branches leading to inferred human-archaic hominin and human-chimpanzee ancestors each have posterior

probabilities of 1.00. These results collectively indicate that 3D genome variation among modern humans is typically not stratified by population in any given genomic locus, but population structure emerges over longer evolutionary time periods and genomic distances.

3D divergence is highest in regions with the lowest functional constraint

To quantify local patterns of 3D divergence along the genome for each genomic window, we computed the 3D divergence of each 1KG individual from the ancestral map. The mean 3D divergence in each window is highly variable across the genome, with many distinct peaks and valleys in both the mean (**Figure 2.3, Supplemental Figure 2.4**) and standard deviation (**Supplemental Figure 2.5**). The distributions of 3D divergence for each chromosome are largely overlapping with slight, but statistically significant differences (**Supplemental Figure 2.6**; Kruskal-Wallis: $P = 3.00 \times 10^{-10}$). The median 3D divergence by chromosome ranges from 0.0014 to 0.0028. Such differences are likely due to variation in gene content, abundance of CTCF binding sites, and other genomic features across chromosomes, all of which can influence 3D genome organization. The majority of the top 10% most divergent windows are shared by all five continental populations (**Supplemental Figure 2.7**). Taken together, these results demonstrate that some windows harbor substantial 3D divergence among individuals, while others exhibit only slight variations on a widely shared contact pattern.

Given the variation in 3D divergence from ancestral across the genome, we tested whether the level of 3D divergence associates with functional annotations or evolutionary

sequence conservation between species. We stratified the genomic windows into deciles based on increasing 3D divergence and quantified the gene count, CTCF site count (ENCODE Project Consortium et al. 2020), and PhastCons 100-way conserved elements (Siepel et al. 2005) distributions for each decile.

Increasing 3D divergence consistently correlates with decreases in sequence identity, gene content, CTCF binding sites and PhastCons conserved bases (**Figure 2.3**). Conversely, recombination rate decreases with increasing 3D divergence (**Figure 2.3**). We also considered SPIN state (Wang et al. 2021; Kamat et al. 2023) predictions and repeat element annotations (Smit 1999; Genereux et al. 2020), but did not observe an overall trend in SPIN state or repeat element prevalence (**Supplemental Figure 2.8**). However, “Lamina” and “Lamina-like” SPIN states are more prevalent in higher 3D divergence windows, while active states are less prevalent (**Supplemental Figure 2.8**). We also compared the 3D divergence of windows containing SNPs that tag 45 common inversions and did not find a relationship between inversion tagging SNPs and 3D divergence (**Supplemental Figure 2.8**) (Giner-Delgado et al. 2019). These results indicate that regions with many functional elements or high sequence conservation overall have less 3D divergence, while those with less functional activity and conservation are more tolerant of variation in 3D contacts. This suggests that 3D chromatin contacts may contribute to evolutionary pressures on sequence divergence.

3D chromatin contact constrains sequence evolution

Next, we explored whether the amount of 3D divergence between humans and the human-archaic hominin ancestor is more or less than expected given the observed sequence divergence. To estimate the expected 3D divergence distribution for each window, we generated 500 sequences with the number of sequence variants from the ancestral matched to the distribution across 1KG individuals and applied Akita to predict the resulting 3D genome divergence (**Supplemental Figure 2.9**) (McArthur et al. 2022). We preserved the tri-nucleotide context of all variants in each window for each sequence to account for variation in the mutation rate across sequence contexts. For each window, we compared the observed 3D divergence with the expected 3D divergence from the 500 sequences with the matched level of nucleotide divergence. If the pressure to maintain 3D chromatin contact patterns does not influence sequence divergence, the observed 3D divergence would be similar to the expected 3D divergence. If the observed 3D divergence deviates from the expected based on sequence divergence, higher observed 3D divergence would suggest positive selection on variants causing 3D differences, while lower observed 3D divergence would suggest negative selection on variants causing 3D differences.

The observed 3D divergence is significantly less than expected based on sequence divergence (**Figure 2.4**). 88.7% of windows have less 3D divergence than expected based on their sequence divergence (binomial test $P < 2.23 \times 10^{-308}$). Genome-wide, the mean expected 3D divergence is 70% higher than the observed 3D divergence (t -test $P = 1.68 \times 10^{-74}$). This suggests that pressure to maintain 3D genome organization constrained sequence divergence in recent human evolution. This aligns with previous

studies that demonstrated depletion of variation at 3D genome-defining elements (e.g., TAD boundaries, CTCF sites) (Fudenberg and Pollard 2019; McArthur and Capra 2021), and it specifically supports maintenance of 3D chromatin contacts as a driver of sequence constraint.

392 windows have significantly greater 3D divergence than expected

Even though most windows have lower 3D divergence than expected, 392 out of the total 4,873 windows have observed 3D divergence significantly greater (t -test $P \leq 0.05$) than the 3D divergence expected based on sequence divergence (**Figure 2.4**). These windows usually have many individuals with high 3D divergence, and we refer to them as “3D divergent windows”. For example, a 3D divergent window on chromosome 1 (chr1:88,604,672-89,653,248) exhibits a multi-modal 3D divergence distribution: a portion of the individuals fall within the expected 3D divergence levels and a portion are much more divergent (**Figure 2.5**). When stratified by populations, the vast majority of 3D divergent windows are divergent in all five continental populations, followed by African-specific divergent windows and divergent windows specific to non-African populations (**Figure 2.5, Supplemental Figure 2.10**). In the example window, our predictions show a group of individuals with a notable loss in contact compared to the other group of individuals with contact maps similar to the ancestral map (**Figure 2.5**). Using lower resolution experimental data from the 4DN we confirmed the presence of both predicted patterns in experimental Hi-C maps (**Figure 2.5**). These results demonstrate that some genomic windows have substantial 3D genome variation within human populations.

In silico mutagenesis reveals that multiple SNVs contribute to common 3D genome variation

To identify the variants underlying the differences observed in each 3D divergent window, we performed *in silico* mutagenesis. *In silico* mutagenesis is a computational technique that uses the ability of Akita to rapidly make predictions on any input DNA sequence to identify and interpret potential causal variants. First, we extracted 616,222 very common (non-ancestral allele frequency > 10%) 1KG SNVs from the 392 divergent windows. We focused on common variants because large numbers of individuals have divergent 3D contact patterns in these windows. We inserted these variants one-by-one into the human-archaic hominin ancestral genome and used Akita to generate chromatin contact predictions for the mutated sequences in each window. Next, we calculated 3D divergence between the ancestral and mutated contact maps (**Figure 2.6**) and quantified the effect of each SNV as the 3D divergence it produces from the ancestral map divided by the maximum 3D divergence between a modern human from ancestral for the window.

A single SNV is not sufficient to explain the 3D divergence observed in most of these windows. For example, the maximum 3D divergence explained by a SNV for each window is less than 10% of the overall 3D divergence (**Figure 2.6**, orange) in more than 40% of windows. We also find that summing the individual effects of all SNVs in a window does not recover substantially more of the observed 3D divergence from ancestral (**Figure 2.6**, grey). This suggests that the 3D divergence is not simply the result of additive combinations of the effects of common SNVs. To illustrate one of the strongest 3D-modifying variants, a SNV on chromosome 7 decreases the strength of an insulating region, causing overall structure in the window to be much less defined (**Figure 2.6**). This

SNV explains 38% of the 3D divergence between an African individual and the ancestor. We also generated *in silico* predictions of 3D divergence for all SNVs > 10% non-ancestral allele frequency in 392 randomly selected windows that are less divergent than expected. Comparing to the 3D divergences for SNVs in the 392 divergent windows, we see an overall similar distribution; however, a larger fraction of SNVs in divergent windows cause high 3D divergence (0.06%) compared to the SNVs from non-divergent windows (0.03%) (**Supplemental Figure 2.11**).

We designated the 176 variants that explain greater than 20% of the maximum observed 3D divergence in a window as “3D-modifying variants”. We quantified the number of 3D modifying variants overlapping CTCF peaks, genes, and conserved bases as called by phyloP (**Figure 2.6**) (Pollard et al. 2010). 82% of 3D modifying variants are found in CTCF binding sites and 60% are in conserved loci. Conversely, only 36% are found within genes. These results are qualitatively maintained when using different cutoffs on the 3D divergence explained (10, 20, 50 and 80%) (**Supplemental Figure 2.12**). Our results suggest that a single common 3D-modifying variant is rarely responsible alone for high 3D divergence, as the maximum impact common SNV for each window contributes modestly to the predicted 3D divergence. Furthermore, these variants predominantly occur at CTCF binding sites and conserved loci, rather than within genes. This underscores their potential significance of SNVs in combination in sculpting the 3D genomic architecture, especially considering the role of 3D chromatin contact in constraining sequence evolution.

31% of the genome has rare 3D genome variation

In the previous section, we investigated windows in which there was common variation in 3D chromatin contact patterns between individuals. We also observed a high occurrence of rare 3D genome variation—where one or a small number of individuals differ from a common contact pattern. To discern underlying patterns in windows with rare 3D divergence, we implemented a classification scheme based on clustering contact maps. Strikingly, the most prevalent pattern was a single individual harboring rare variation that distinguished them from the remainder of the cohort (**Figure 2.7**). This distinctive pattern was observed in approximately 31% of the windows (N = 1,494). Furthermore, the majority of windows exhibiting rare variation were primarily driven by individuals of African ancestry, characterized by substantial 3D divergence from all other individuals in the study cohort (**Figure 2.7**). The prevalence of individuals exhibiting rare variation in a substantial proportion of windows underscores the potential of individual-specific genomic alterations to shape 3D genome architecture. Additionally, the prominent contribution of individuals of African descent to windows with rare variation highlights the importance of considering diverse genetic backgrounds when studying 3D genomic diversity.

Rare 3D genome variation is usually the result of a single large-effect variant

To identify the variants contributing to the most prominent differences in 3D architecture in windows with rare 3D variation, we used *in silico* mutagenesis to test rare SNVs in the windows with rare 3D variation. We selected 12,175 variants that are private to the highly divergent individual (in the context of all 1KG individuals used in this study) to be inserted

these one-by-one into the hg38 human reference genome and calculated 3D divergence between the reference and mutated contact maps (**Figure 2.6**). We then quantified the effect of a SNV by calculating the percentage of the 3D divergence between the highly divergent individual from the reference maps that is generated by inserting the SNV alone into the reference sequence.

In contrast to cases of common 3D divergence, the maximum effect SNV for each window often generates a large fraction of the observed 3D divergence from reference (**Figure 2.7**). In cases in which the 3D divergence induced by a single SNV is greater than 100%, this suggests other variants present in the window temper the effect of the maximum effect SNV and thus reduce the 3D divergence compared to the sequence with the SNV alone. We identified 1,482 variants that explain at least 20% of the 3D divergence between the rare individual and the reference genome. 71% of these 3D modifying variants are found in CTCF binding sites and 69% are in conserved loci. Conversely, only 38% are found within genes (**Figure 2.7**). These results are qualitatively maintained at other cutoffs on 3D divergence explained (10, 20, 50 and 80%) (**Supplemental Figure 2.13**). To illustrate this pattern, we highlight an example SNV that decreases the strength of an insulating region, causing overall structure in the window to be much less defined (**Figure 2.7**). This SNV explains 78% of the 3D divergence between an African individual and the ancestral genome. In contrast to our results in 3D divergent windows, these results suggest that rare 3D variation is often caused by a single, strongly 3D modifying variant.

Discussion

Our study explores the interplay between genetic sequence variation and 3D chromatin contact patterns using machine learning to predict 3D chromatin contacts (Fudenberg et al. 2020) for thousands of modern humans from around the globe (1000 Genomes Project Consortium et al. 2015). Quantifying 3D chromatin contact on this scale is necessary to capture its variation across humans, but given the logistical and technical challenges of generating high-resolution Hi-C data at population-scale, this is not currently possible without computational methods. The perspective provided by our dataset enabled us to make several novel observations not seen in previous small-scale studies of human 3D chromatin contact diversity (Li et al. 2023).

3D chromatin contact divergence vs. sequence divergence

Our results show that 3D chromatin contact divergence follows similar genome-wide trends as sequence divergence. For example, African populations exhibited consistently higher average 3D divergence in comparison to other populations that experienced the out-of-Africa genetic bottleneck, which corresponds to Africans' greater sequence diversity (1000 Genomes Project Consortium et al. 2015). However, the correlation between 3D chromatin contact similarity and sequence divergence ($R^2 = 0.31$) is only moderate, suggesting the existence of differing influences and regulatory mechanisms shaping the interplay between sequence divergence and 3D genome organization across diverse individuals. Indeed, quantification of local window-specific 3D divergence showed that 3D contact map variation in most genomic regions is shared across populations, and

no windows have contact map patterns that stratify by population. Moreover, it revealed that rare 3D contact variation is common—31% of windows have an individual with a rare divergent contact pattern.

Influence of 3D chromatin contact on sequence evolution and functional constraint

We also found that the observed 3D divergence between modern humans and the human-archaic hominin ancestor is significantly less than expected based on observed sequence divergence. This suggests that constraint imposed by the pressure to maintain 3D chromatin contacts shaped sequence divergence during recent human evolution. The findings are consistent with prior studies indicating a depletion of variation at key 3D genome determining elements (Krefting et al. 2018; Fudenberg and Pollard 2019; Whalen and Pollard 2019; Sauerwald et al. 2020; McArthur and Capra 2021; McArthur et al. 2022) and suggest that pressure to preserve 3D chromatin contact contributes to sequence constraint in human evolution. By comparing the observed and expected 3D divergence derived from sequence divergence, we underscore the potential role of 3D genome organization in influencing recent human sequence evolution.

Examining local patterns of 3D divergence along the genome revealed substantial variability, indicating varied tolerance for 3D divergence. Regions exhibiting elevated 3D divergence consistently had reduced gene content, fewer CTCF binding sites, and fewer conserved bases than other genomic windows. These results are consistent with previous work that investigated two cell lines and found variation along chromosomes that correlates with compartment, GC content, transcription rate and repeat element

prevalence (Gunsalus et al. 2023a). This pattern underscores the importance of maintaining 3D chromatin contacts, especially in regions with many functional elements.

In silico mutagenesis identifies SNVs likely to drive 3D divergence

Another strength of the sequence-based machine learning approach is that it enables rapid screening of the effects of individual genetic variants in different genetic backgrounds (McArthur et al. 2022; Gunsalus et al. 2023b; Brand et al. 2023). We used this *in silico* mutagenesis to unravel the influence of SNVs on 3D genome variation. We discovered that the 3D divergence in windows with common 3D variation was rarely the result of the independent additive effects of common SNVs. Instead, our analyses suggest that combinations of SNVs likely interact to produce much of the common variation in the 3D genome. In contrast, for windows with only rare 3D variation, a single, high-impact variant was often sufficient to produce the observed 3D divergence. This suggests that individual variants with strong impacts on 3D contact are rarely tolerated at high frequencies. However, the 3D-modifying variants observed in both types of windows predominantly influenced crucial functional sites such as CTCF binding sites and evolutionarily conserved loci. The sharp contrast in variant contributions to common and rare 3D variation underscores the complex mechanisms governing 3D chromatin contact and its variation.

Machine learning addresses challenges with experimental Hi-C

Traditional Hi-C experiments often compromise resolution for coverage, resulting in representations that lack finer details pivotal for understanding 3D genome architecture at the scale of differences observed between healthy individuals. This drawback limits our ability to capture potentially functional chromatin interactions within larger structures and impedes comprehensive genomic comparisons. To overcome these limitations, our study harnesses Akita, an accurate machine learning prediction model. Akita demonstrates robust performance in generating local 3D contact patterns from DNA sequences at a higher resolution (2 kb), enabling a finer-scale analysis of chromatin interactions that compensates for limitations in available experimental data (Fudenberg et al. 2020). Our findings showcase Akita's efficacy in predicting 3D chromatin architecture not only in its original training data of European-derived cell lines, but also in diverse individuals, particularly among Africans. This ability to perform consistently across diverse individuals is critical, as it allows us to investigate chromatin organization in groups where Hi-C data is limited. Our research thereby offers a more comprehensive view of the 3D genome landscape, crucial for understanding chromatin organization and its functional implications.

Limitations

While our study increases understanding of chromatin contact variation, it is important to acknowledge several limitations due to current Hi-C data quality, resolution, and quantity. First, while we validate example predictions with experimental Hi-C data when possible,

the scope of our predictions means that we are not able to validate most given available data. Second, the Hi-C data used for training and validation of the prediction models are unphased and represent a combination of both alleles. As a result, the prediction model was trained on haploid sequences due to this lack of phased Hi-C data. In previous work, this approach has been demonstrated to accurately capture contact patterns (e.g., Fudenberg et al. 2020). While inferring allele-specific contact maps with phased SNV data is feasible in some cases, high-resolution haplotype-resolved contact maps are not readily available due to limits imposed by both the read depth needed to obtain high-quality Hi-C data and the additional complexity required in analytical methods. Therefore, there is no clear approach for combining phased 3D map predictions in a way that would represent the haploid Hi-C data. Thus, we have focused our predictions and sequence comparisons on pseudo-haploid versions of the phased variant calls from the 1KG Phase 3 data. Moving forward, we envision close integration between computational predictions and new experimental data for discovery and validation of the mapping between DNA sequence variation and 3D chromatin contact.

Additionally, there are some limitations in the variant sets we consider. First, the 1KG dataset, while extensive, does not encompass the entirety of human genetic diversity. Specifically, the African individuals included in 1KG do not capture more deeply divergent African lineages; expanding to additional datasets would further increase the genetic diversity covered (Mallick et al. 2016; Fan et al. 2023). Hence, future studies should aim to incorporate a wider array of individuals to provide a more comprehensive understanding of the interplay between 3D chromatin contact and genetic sequence divergence. Our study is also focused on SNVs, excluding larger structural variants, which

have been shown to contribute to 3D chromatin contact differences (Norton and Phillips-Cremins 2017; Spielmann et al. 2018; Sánchez-Gaya et al. 2020). However, we show that the 3D modifying variants we discover do not tag large inversions (**Supplemental Figure 2.8**). We also note that the *in silico* mutagenesis analysis considers single SNVs at a time. Given the lack of additive effects observed in the commonly 3D divergent windows, future work is needed to evaluate the combinatorial impact of variants and test for interaction effects. Finally, our analysis did not explore the potential impact of differences between cell-types, which could influence the observed 3D chromatin contact patterns.

Further validation of the relationships between sequence variation, 3D chromatin contact, and functional implications presented in this study will require additional data. We are optimistic that ongoing efforts to expand Hi-C data resolution, cell-type coverage, and availability will enable comprehensive understanding of the mechanisms and variation of chromatin organization and its functional outcomes.

Conclusions

Our study uses machine learning to map the relationship between genetic sequence variation and 3D chromatin contact across diverse human populations. Our findings pave the way for future research exploring the mechanisms governing chromatin organization and its functional implications in disease and evolution.

Methods

Modern human and ancestral genomes

All analysis was conducted using the GRCh38 (hg38) genome assembly and coordinates. Genetic variants in modern humans came from 1KG, Phase 3 from (1000 Genomes Project Consortium et al. 2015). The human-archaic hominin ancestral genome was extracted using ancestral allele calls for each position in the tree sequence from an ancestral recombination graph of modern and ancient humans with archaic hominins (Wohns et al. 2022). Tree sequences are an efficient data format for representing the ancestral relationships between sets of DNA sequences and were analyzed using tskit (Kelleher et al. 2018). We identified ancestral allele calls for every available position and used these to generate a VCF file. We retrieved ancestral allele calls for the human-chimpanzee common ancestor from the Great Apes Genome Project (GAGP) (Prado-Martinez et al. 2013). We then constructed full-length genomes for each modern individual and inferred ancestor based upon the genotyping information in their respective VCF file. We constructed the sequence for each 1KG individual's genome using GATK's FastaAlternateReferenceMaker tool (Van der Auwera and O'Connor 2020). If an individual had an alternate allele (homozygous or heterozygous), we inserted it into the reference genome to create a pseudo-haploid, or “flattened” genome for each individual (**Supplemental Figure 2.1**). To maintain the window and overlap size required by Akita, we included all SNVs, but not SVs, in these genomes.

3D chromatin contact prediction with Akita

Akita is a CNN model designed and trained to predict 3D genome contacts at 2 kb resolution in approximately 1 Mb windows from one-hot encoded DNA sequence inputs. After the input genome sequences were prepared, we made predictions using Akita for ~1 Mb ($2^{20} = 1,048,576$ bp) sliding windows overlapping by half (e.g., 524,288–1,572,864, 1,048,576–2,097,152, 1,572,864–2,621,440). Although Akita was trained simultaneously on Hi-C and Micro-C across five cell types in a multi-task framework to achieve greater accuracy, we focused on predictions in the highest resolution maps, human foreskin fibroblast (HFF) as in McArthur et al. 2022. Akita considers the full window, but predictions are generated for only the middle 917,504 bp. Each cell in the contact map predicted from an individual's DNA sequence represents physical 3D contacts between pairs of regions at 2,048 bp resolution. The predicted value in each cell quantifies the observed contact frequency over the expected contact frequency given the distance of the two genomic regions ($\log_2(\text{obs}/\text{exp})$); comparing to expected contacts enables accounting for the distance-dependent nature of chromatin contacts. For all analyses, we only considered windows with 100% coverage in the hg38 reference genome for a total of 4,873 autosomal windows. Fudenberg et al., 2020 provides further details on the CNN architecture and training data used.

Scaling 3D genome predictions

To explore the sensitivity of our results to different window sizes, we compared predicted maps at the native prediction window size (1,048,576 bp) and four additional window

sizes (524,288, 262,144, 131,072, and 65,536 bp). These window sizes correspond to decreasing powers of two (2^{20} , 2^{19} , 2^{18} , 2^{17} , 2^{16}). We created sub-maps from the original 1 Mb Akita predictions by dividing the predicted contact maps into smaller contact maps of each size. To maintain consistency with the previous analyses, the smaller windows overlap by half within the 1 Mb of the original predictions.

3D and sequence genome comparisons

After predictions were made on all 1 Mb windows for all individuals, we compared the resulting predictions using mean-squared error and Spearman and Pearson correlations. All measures are scaled to indicate divergence: high values represent difference and low values represent similarity. We subtracted the Spearman's rank correlation coefficient from one ($1-\rho$) to quantify 3D divergence (Gunsalus et al. 2023b). Some analyses compared 3D divergence with sequence divergence. To calculate the sequence divergence between two individuals, we counted the number of bases at which the two individuals differ in the 1 Mb window and divided by the total number of bases compared. This was done only for windows with 100% coverage in hg38, as with the 3D chromatin contact predictions.

Empirical distribution of expected 3D divergence

We generated genomes with shuffled nucleotide differences to compute the expected 3D divergence in a window given the observed sequence divergence (**Supplemental Figure 2.9**). This approach was adapted from (McArthur et al. 2022). We matched these shuffled

differences to the same number and tri-nucleotide context of the observed sequence differences between 1KG individual genomes and the inferred ancestral genome. The observed sequences differences were extracted from the individual from the set closest to the mean of the 3D divergence distribution for each population (HG03105 [African], HG01119 [American], NA06985 [European], HG00759 [East Asian], HG03007 [South Asian]). For each 1 Mb window of the genome (N = 4,873) we generated 500 shuffled sequences—100 for each population. We applied Akita to each of these shuffled sequences and calculated an empirical distribution of expected 3D divergence by comparing the contact maps of the shuffled sequences with the ancestral sequence. Finally, we compared the expected 3D divergence from this distribution to the observed ancestral-modern 3D divergence. The procedure for generating the null distribution from shuffled sequences has two caveats. First, tri-nucleotide context is commonly used to capture variation in mutation rate, but considering higher-order contexts would result in even better modeling of the mutational process (Aggarwala and Voight 2016). Second, it does not account for factors such as linkage disequilibrium and recombination rate that may vary between populations. This is appropriate for our application because our goal is to broadly survey the range of possible 3D contact variation and we construct null distributions for each population.

Comparison of function and conserved elements with 3D divergence

Evolutionary conservation estimates and genomic functional annotations were obtained from publicly available data sources. Gene annotations are from GENCODE version 24

(Frankish et al. 2019). CTCF binding sites were determined through ChIP-seq analyses from ENCODE (ENCODE Project Consortium 2012; Davis et al. 2018). We downloaded all CTCF ChIP-seq data with the following criteria: experiment, released, ChIP-seq, human (hg38), all tissues, adult, BED NarrowPeak file format. We excluded any experiments with biosample treatments. Across all files, CTCF peaks were concatenated, sorted, and merged into a single file; overlapping peaks were combined into a single larger peak. We quantified the number of CTCF ChIP-seq peaks per genomic window (peaks per window) and the number of CTCF peak base pairs overlapping each window (base pairs per window). Evolutionary constraint was quantified by PhastCons. The PhastCons elements (Siepel et al. 2005) were intersected with 1 Mb genomic windows. The overlap was quantified as the number of PhastCons base pairs per boundary regardless of score (base pairs per window). Conserved base pairs were identified by PhyloP (Pollard et al. 2010). PhyloP scores and PhastCons elements for multiple alignments of 100 vertebrate species were downloaded from the conservation track of the comparative genomics data group on the UCSC Genome Table Browser using the “phyloP100way” and “phastConsElements100way” tables respectively (<https://genome.ucsc.edu/cgi-bin/hgTables>). Recombination rate average annotations for 1KG samples were also downloaded from the UCSC Genome Table Browser (<https://genome.ucsc.edu/cgi-bin/hgTables>). We retrieved positions of inversion tagging SNPs from the supplemental files of (Giner-Delgado et al. 2019).

We used the pybedtools wrapper for BEDtools (Quinlan and Hall 2010; Dale et al. 2011) to perform intersections of genomic regions for the above annotations (genes, CTCF peaks, PhastCons, recombination rate, inversions) with the 1 Mb windows used

for Akita predictions. These windows were stratified by mean 3D divergence from ancestral for all 1KG individuals and by the difference in the mean of the observed distribution of 3D divergence from the expected as described above.

Shared divergent windows across populations

The top 10% of windows for each 1KG population were chosen based on the mean 3D divergence from the ancestral for all individuals in the respective populations. Overlap was calculated using a python implementation of UpSet plots, a tool to visualize set overlaps (Lex et al. 2014; Nothman 2023).

Hierarchical clustering of 3D chromatin contact maps

All pairs of 1KG individuals were compared for each of the 4873 genomic windows. Pairwise 3D divergence score matrices for each 1 Mb window were used to cluster these individuals, plus the human-archaic hominin and human-chimpanzee ancestral genomes, using hierarchical clustering with complete linkage as implemented in scipy (Virtanen et al. 2020). The clustering generated dendrograms (“trees”) that describe the relationships between individuals. The Python API for ETE ToolKit was used to identify any trees that are monophyletic for a given population, meaning that any population clustered entirely and exclusively together. To establish support for known population patterns based on the 3D divergence trees, we generated a baseline tree representing the sequence similarity of two inferred ancestors and 1KG individuals from all 1KG populations but the Americas (AMR) population, the African American Southwest (ASW) sub-population, and the

African Caribbean Barbados (ACB) sub-population which exhibit substantially more admixture than other 1KG populations (Li et al. 2008; Gravel et al. 2011; 1000 Genomes Project Consortium et al. 2015; Duda and Jan Zrzavý 2016; Bergström et al. 2020). This sequence tree was generated using the same flattened haploid sequences that were used to infer 3D chromatin contacts, and it produced similar patterns as expected from diploid comparisons. We used ASTRAL (Zhang et al. 2018; Rabiee et al. 2019) to calculate support for the branches of the baseline tree based on the trees for each window. This analysis treated the tree for each window as a ‘gene tree’ and the baseline tree as the ‘species’ tree.

In silico mutagenesis

We estimated the effects of individual variants on 3D divergence using *in silico* mutagenesis (**Figure 2.6**). For commonly divergent windows, we identified common non-ancestral alleles (AF>10%) among the 1KG individuals, consisting of 616,222 unique variants in 392 genomic windows. For each variant-window pair, we inserted the variant into the ancestral sequence for that window and calculated the 3D divergence between the ancestral map and the ancestral with variant map.

We quantified the effects of variants via “explained 3D divergence”, dividing the 3D divergence for the variant by the maximum ancestral to 1KG 3D divergence for the window. Values near zero indicate that the variant explains minimal 3D divergence among the observed comparisons, while values near one indicate the variant explains most of the 3D divergence among observed comparisons. Values greater than one indicate that

variant creates more 3D divergence than observed among any ancestral to 1KG comparison, suggesting that other variants may “buffer” against the variant's effect. “3D-modifying variants” were defined as variants that produced greater than 20% of the maximum observed 3D divergence in the window. However, we demonstrate that our results are robust to different thresholds (**Supplemental Figure 2.12**).

We also applied our *in silico* mutagenesis approach to rare variants private to the highly divergent individual in each of the 1,494 windows with rare 3D variation. Private variants were defined at positions where only the divergent individual carried a copy of the alternate allele in the 1KG individuals used for the clustering analysis. We considered 12,175 variants across the 1,494 windows. In this case, explained 3D divergence was calculated with respect to the hg38 reference genome as this analysis focuses on within human variation.

Analysis of experimental Hi-C data

We downloaded preprocessed cooler files from the 4DN Data Portal (<https://data.4dnucleome.org>) and quantified contacts at 10 kb resolution. Visualization was done using custom code adapted from Fudenberg et al. 2020, Gunsalus et al. 2023b and Brand et al. 2023.

Significance reporting

The machine used to run analyses had a minimum value for representing floating numbers of $2.2250738585072014 \times 10^{-308}$. Therefore, we abbreviate values less than this as 2.23×10^{-308} .

Data availability

The publicly available data used for analysis are available in the following repositories:

1KG VCFs (1000 Genomes Project Consortium et al. 2015):
(ftp://1000genomes.ebi.ac.uk/vol1/ftp/data_collections/1000_genomes_project/release/20190312_biallelic_SNV_and_INDEL/)

Human-chimpanzee ancestral alleles (Prado-Martinez et al. 2013):
(https://eichlerlab.gs.washington.edu/greatape/data/Ancestral_Alleles/)

CTCF-bound open chromatin candidate cis-regulatory elements (cCREs) in all cell types:
(<https://screen.encodeproject.org/> > Downloads > Download Human CTCF-bound cCREs)

phastCons elements and PhyloP scores were retrieved from the UCSC Genome Browser:
(<https://hgdownload.soe.ucsc.edu/goldenPath/hg38/database/phastConsElements100way.txt.gz>,

<https://hgdownload.soe.ucsc.edu/goldenPath/hg38/database/phyloP100way.txt.gz>).

Recombination rate annotations for 1KG samples were also downloaded from the UCSC Genome Table Browser: (<https://genome.ucsc.edu/cgi-bin/hgTables>, Recomb1000GAvg table).

Inversion tagging SNPs: Supplemental table 12 from (Giner-Delgado et al. 2019).

Experimental Hi-C available at the 4D Nucleome data portal:

(https://data.4dnucleome.org/browse/?dataset_label=Hi-C+on+lympoblastoid+cell+lines+from+1000G+individuals&experiments_in_set.experiment_type.experiment_category=Sequencing&experimentset_type=replicate&type=ExperimentSetReplicate)

Code availability

All code used to conduct analyses and generate figures is publicly available on GitHub (<https://github.com/egilbertson-ucsf/3DGenome-diversity>). Akita is available from the basenji repository on GitHub

(<https://github.com/calico/basenji/tree/master/manuscripts/akita>).

Acknowledgements

This work was supported by the National Institutes of Health (NIH) General Medical Sciences Institute award R35GM127087 to JAC, NIH National Heart, Lung, and Blood Institute award U01HL157989 to KSP, and NIH National Human Genome Research Institute award F30HG011200 to EM. The work as also supported by funds from the Gladstone Institutes and the Bakar Computational Health Sciences Institute.

This work was conducted in part using the resources of the Wynton High Performance Computer at the University of California San Francisco.

We thank Jian Ma's lab for sharing SPIN State annotations for HFF cells.

We also thank members of the Capra and Pollard Labs who gave helpful feedback throughout this project.

Author contributions

Conceptualization: ENG, CMB, EM, DCR, KSP and JAC

Formal Analysis: ENG

Visualization: ENG, CMB, and JAC

Resources and Software: ENG, CMB, EM, DCR, SK

Writing – Original Draft: ENG and JAC

Writing – Review & Editing: ENG, CMB, EM, DCR, SK, KSP and JAC.

Competing interests

The authors declare no competing interests

Figures

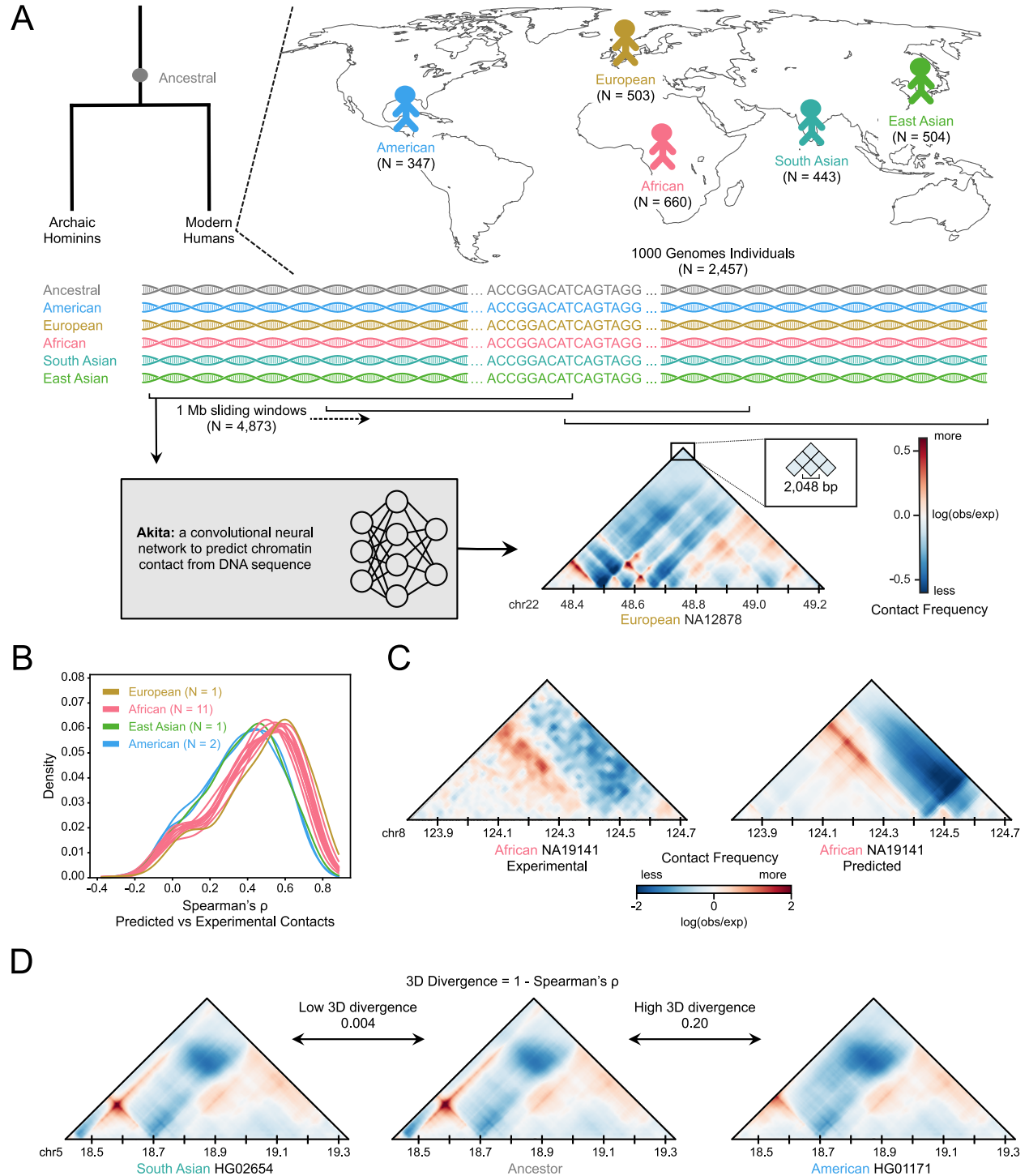


Figure 2.1: Strategy for investigating 3D chromatin contact patterns in diverse humans.

(A) Schematic of the generation of genome-wide 3D contact maps for 2,457 unrelated individuals from five population groups defined by the 1000 Genomes Project. Akita is a deep neural network that takes approximately 1 Mb of DNA sequence as input and generates a 3D contact map of the window. The map consists of contacts for all pairs of 2,048 bp regions within the window. We applied Akita to the DNA sequence of each individual in sliding windows overlapping by half across the genome. We discarded windows without full sequence coverage in the hg38 reference sequence, resulting in 4,873 windows. We also applied Akita to an inferred human--archaic hominin ancestral sequence (Wohns et al. 2022). **(B)** Density of Spearman correlations between experimental and predicted contact maps at 10 kb resolution for windows in the Akita held-out test set of 413 windows across 15 individuals from 4 populations. This includes a European individual (GM12878) used as part of the Akita training data as a benchmark. The strong performance on African individuals suggests that Akita is accurate across populations. The lower performance on the East Asian and admixed American individuals is likely due to lower resolution of their experimental maps (**Supplemental Figure 2.2**). **(C)** Example experimental and predicted maps for a representative window on chromosome 8 (chr8:123,740,160-124,788,736) from an African individual. **(D)** Example predictions and comparisons of 3D chromatin contact maps between pairs of individuals on chromosome 5 (chr5:18,350,080-19,398,656). To quantify “3D divergence”, we calculated the Spearman correlation coefficient over the corresponding cells for a given pair of maps subtracted from 1. Low 3D divergence scores indicate high similarity between contact maps and high 3D divergence scores indicate low similarity between maps. We also compute the mean squared error (MSE) between contact maps (**Supplemental Figure 2.4**).

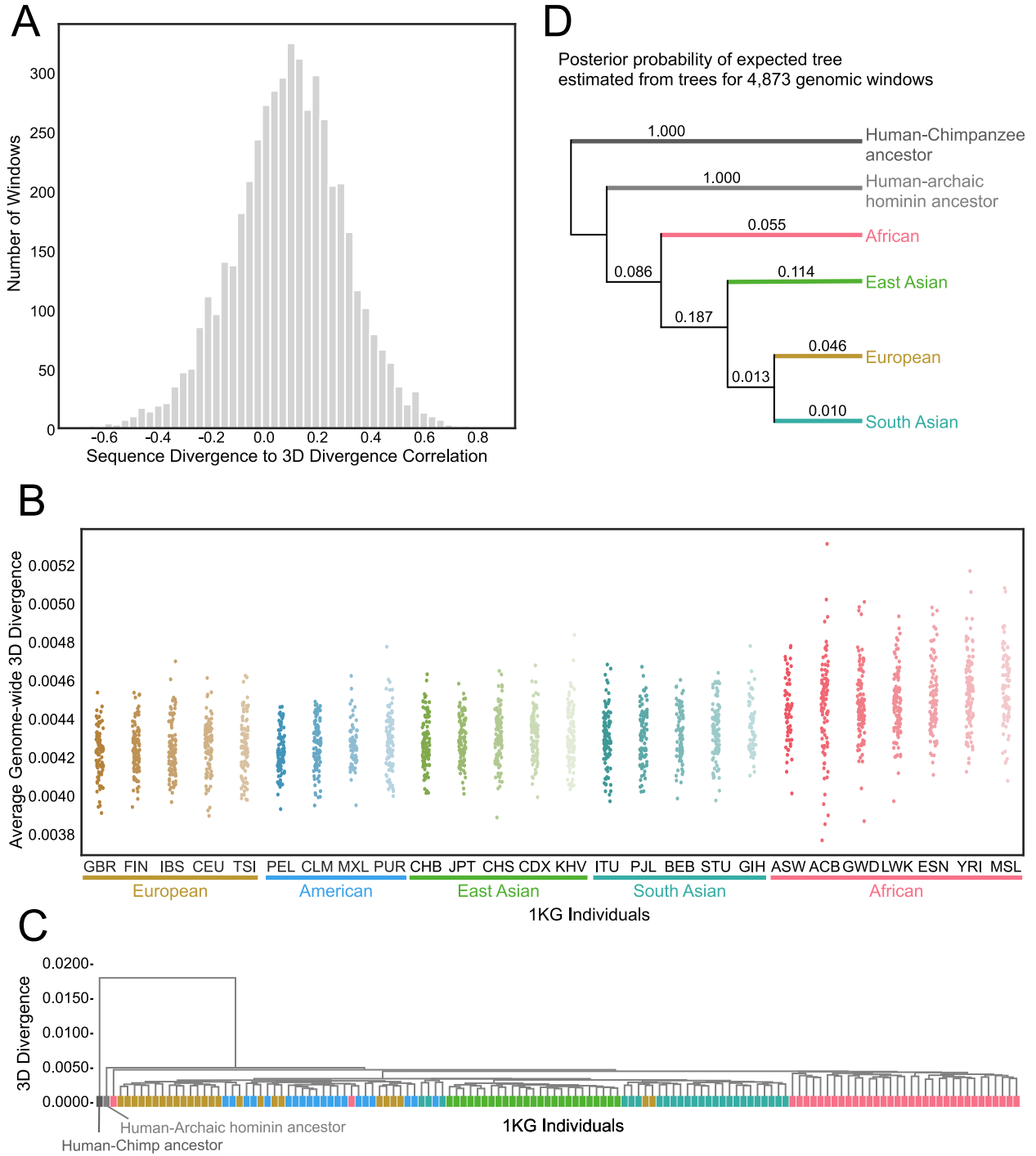


Figure 2.2: Genome-wide 3D divergence follows known population structure

(A) Distribution of Spearman correlation of pairwise SNV sequence divergence and 3D divergence between 1KG individuals and the human archaic-hominin ancestor for all 4,873 windows. **(B)** Genome-wide average 3D divergence for each individual, stratified by continental and sub-continental 1KG populations. Color indicates super-population and hue indicates sub-population. **(C)** Genome-wide hierarchical clustering of 1KG individuals with the inferred human-archaic hominin ancestor, and the human-chimpanzee ancestor using pairwise average genome-wide 3D divergence. Color indicates super-population. We plot 130 individuals who represent the overall patterns for visual simplicity. **(D)** Branch support (posterior probability) for the population tree inferred from 1KG sequences estimated using ASTRAL (Zhang et al. 2018) from the topologies of trees constructed for each window based on 3D divergence.

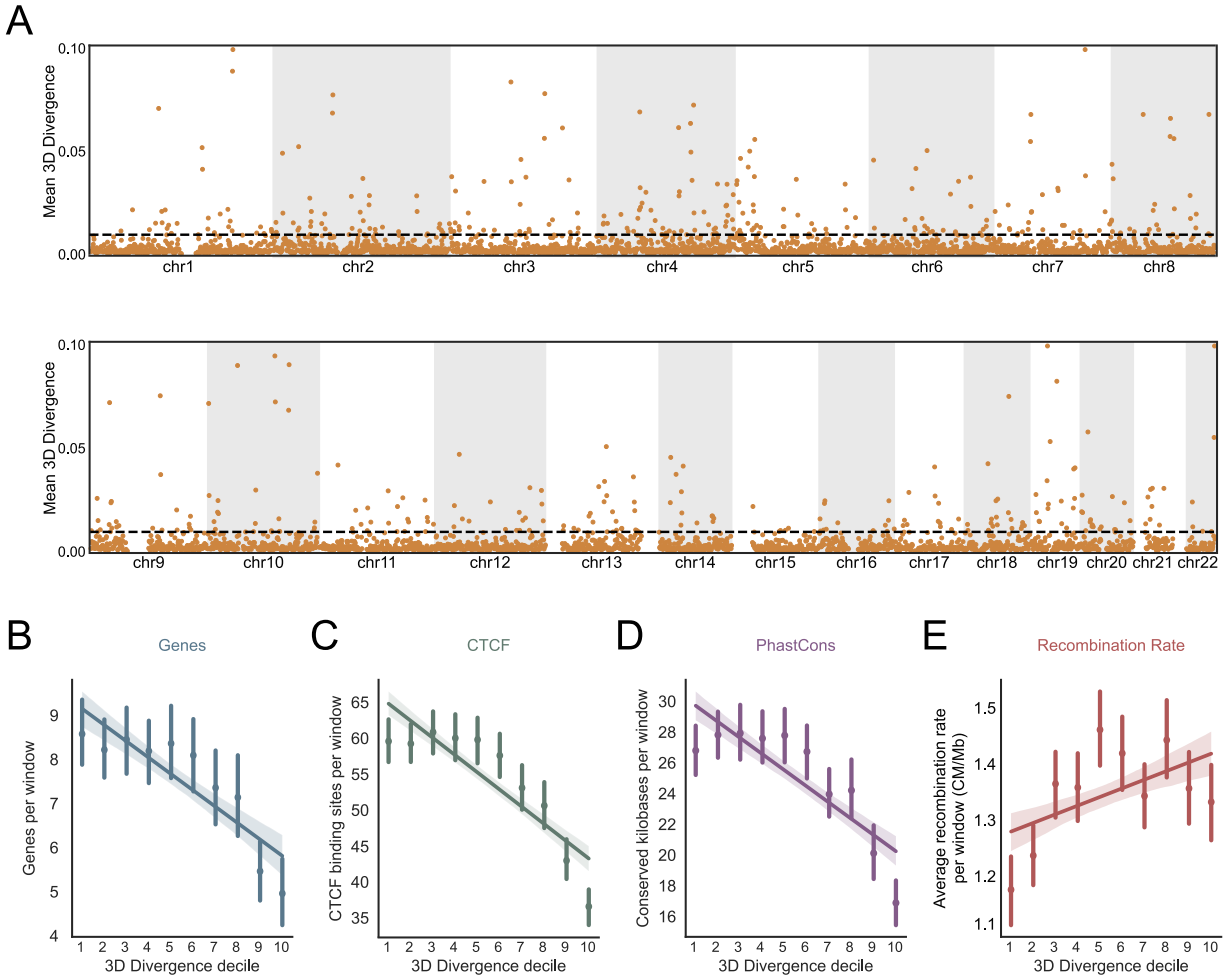


Figure 2.3: 3D Divergence is variable across the genome and highest in less functional regions.

(A) Mean 3D divergence from the human-archaic hominin ancestor across 2,457 individuals from 1KG for each of the 4,873 genomic windows. Each point represents the mean 3D divergence of all individuals from the ancestral genome for a single genomic window. The dotted line indicates the top 10% of 3D divergence. 3D divergences greater than 0.10 are plotted at 0.10 to aid visualization. **(B)** Average number of genes per window in deciles based on mean 3D divergence from the hominin ancestor (bin 1 has the lowest 3D divergence and 10 highest). Bars indicate bootstrapped 95% confidence intervals. Gene annotations are from GENCODE version 24 in each 3D divergence decile. **(C)** Average number of CTCF binding sites per window. CTCF peaks come from merging CTCF ChIP-seq peaks across all cell types from the ENCODE Consortium. Visualized as in **B**. **(D)** Average PhastCons 100-way conserved bases (in kb) per window in each 3D divergence decile. Visualized as in **B**. **(E)** Average recombination rate (centimorgans/Mb) per window in each 3D divergence decile. Visualized as in **B**.

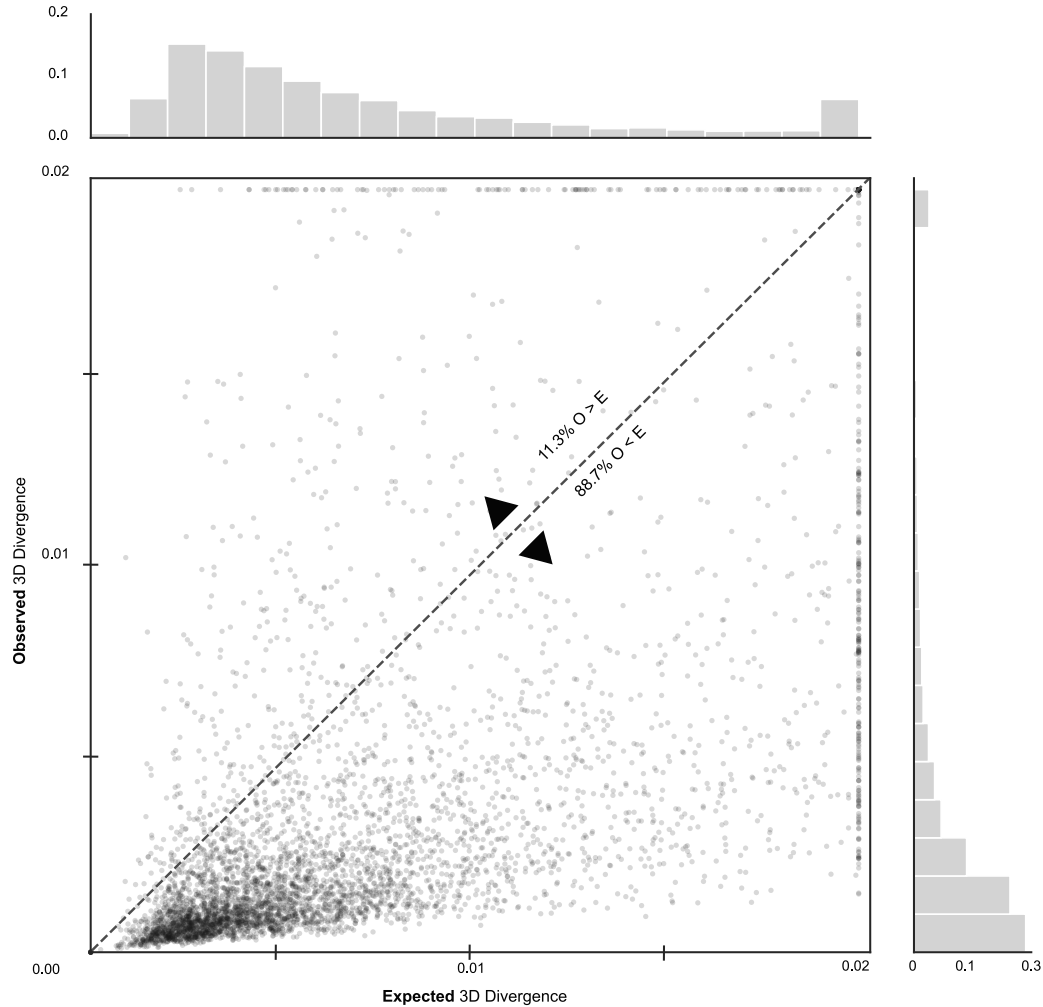


Figure 2.4: 3D divergence is lower than expected in 89% of genomics windows, but 392 have significantly greater 3D divergence than expected.

Mean observed 3D divergence between 1KG individuals and the human-archaic hominin ancestor compared to 3D divergence expected from the amount of sequence variation. The expected 3D divergence distribution for a window is based predicted 3D genome contact maps for 500 simulated sequences for each window (Methods). Points above the line represent windows more divergent than expected, which suggests more observed variants that alter 3D divergence than expected. Points below represent windows less divergent than expected, which suggests constraint on sequence variation to maintain 3D chromatin contact patterns. Observed 3D divergence is significantly less than the mean expected 3D divergence based on sequence (O < E) for 88.7% of windows (N = 4322; binomial-test $P < 2.23 \times 10^{-308}$). The mean expected 3D divergence is on average 70-times higher than the observed 3D divergence (t-test $P = 1.68 \times 10^{-74}$). Nonetheless, we identified 392 windows with observed 3D divergence distributions significantly greater than the 3D divergence expected based on sequence divergence (O > E; t-test $P \leq 0.05$). 3D divergence scores greater than 0.02 are plotted at 0.02 for visualization.

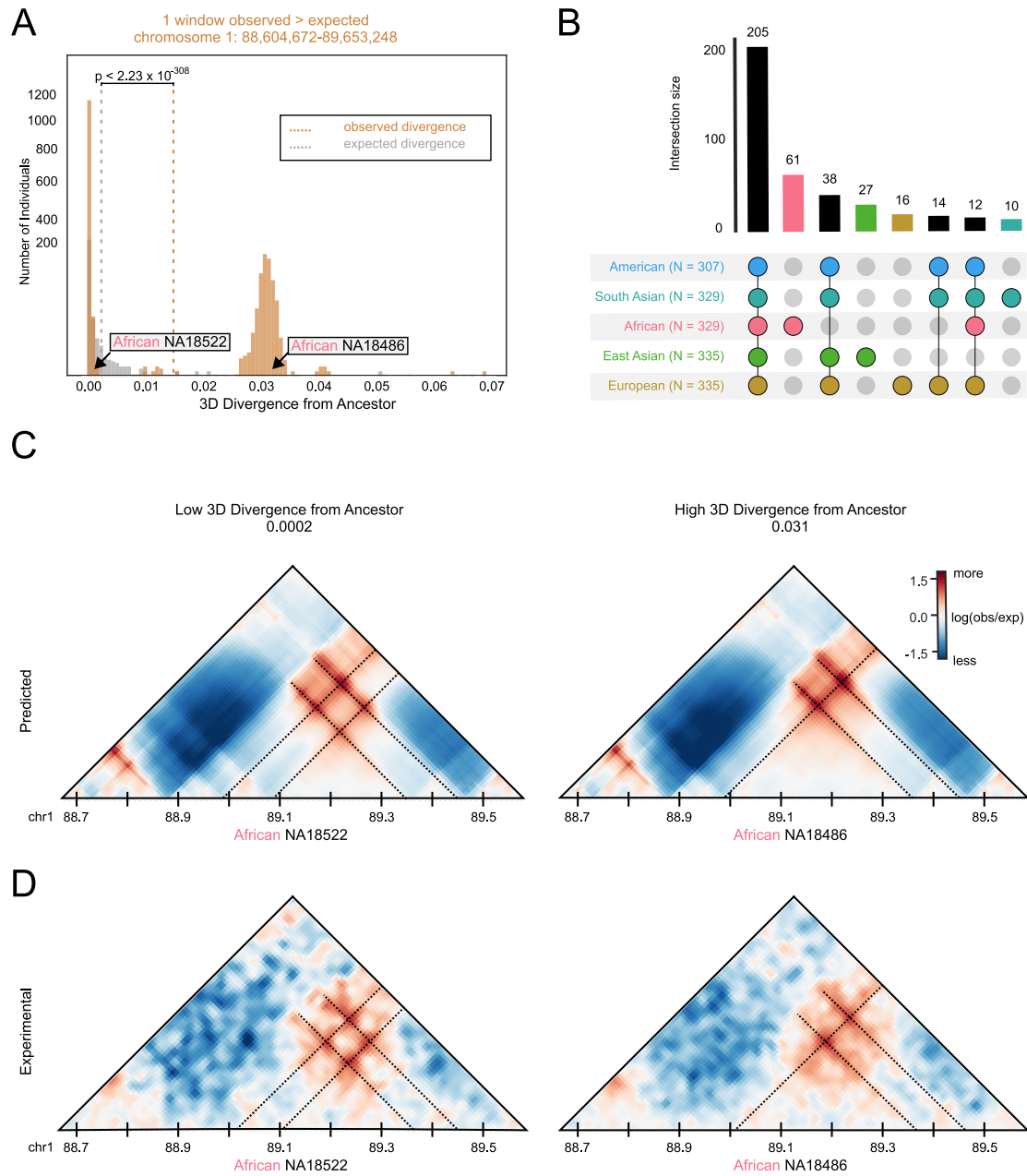


Figure 2.5: Experimental Hi-C data confirms predicted contacts in highly divergent windows.

(A) Distributions of the 3D divergence between humans and the archaic hominin ancestor (orange) and the expected 3D divergence (gray) 3D divergence for an example highly divergent window. Dotted lines represent the mean of the respective distributions. **(B)** Sharing of highly divergent windows among 1KG super-populations. Bars indicate the number of highly divergent windows present in each combination of populations indicated by the dot matrix. Population combinations with fewer than 10 windows are not plotted; see **Supplemental Figure 2.10** for the full plot. **(C)** Example predicted maps for two African Yoruba individuals at the example window, one with low 3D divergence from the ancestor (NA18522; 3D divergence = 0.0002) and one with high 3D divergence (NA18486; 3D divergence = 0.031). The predicted maps are scaled to 10 kb resolution to be comparable to the resolution of the experimental Hi-C maps. The dotted lines highlight strong contacts. **(D)** Experimentally determined Hi-C contact maps for this example window for the two Yoruba individuals. These experimental maps confirm the predicted high 3D divergence and contact pattern differences.

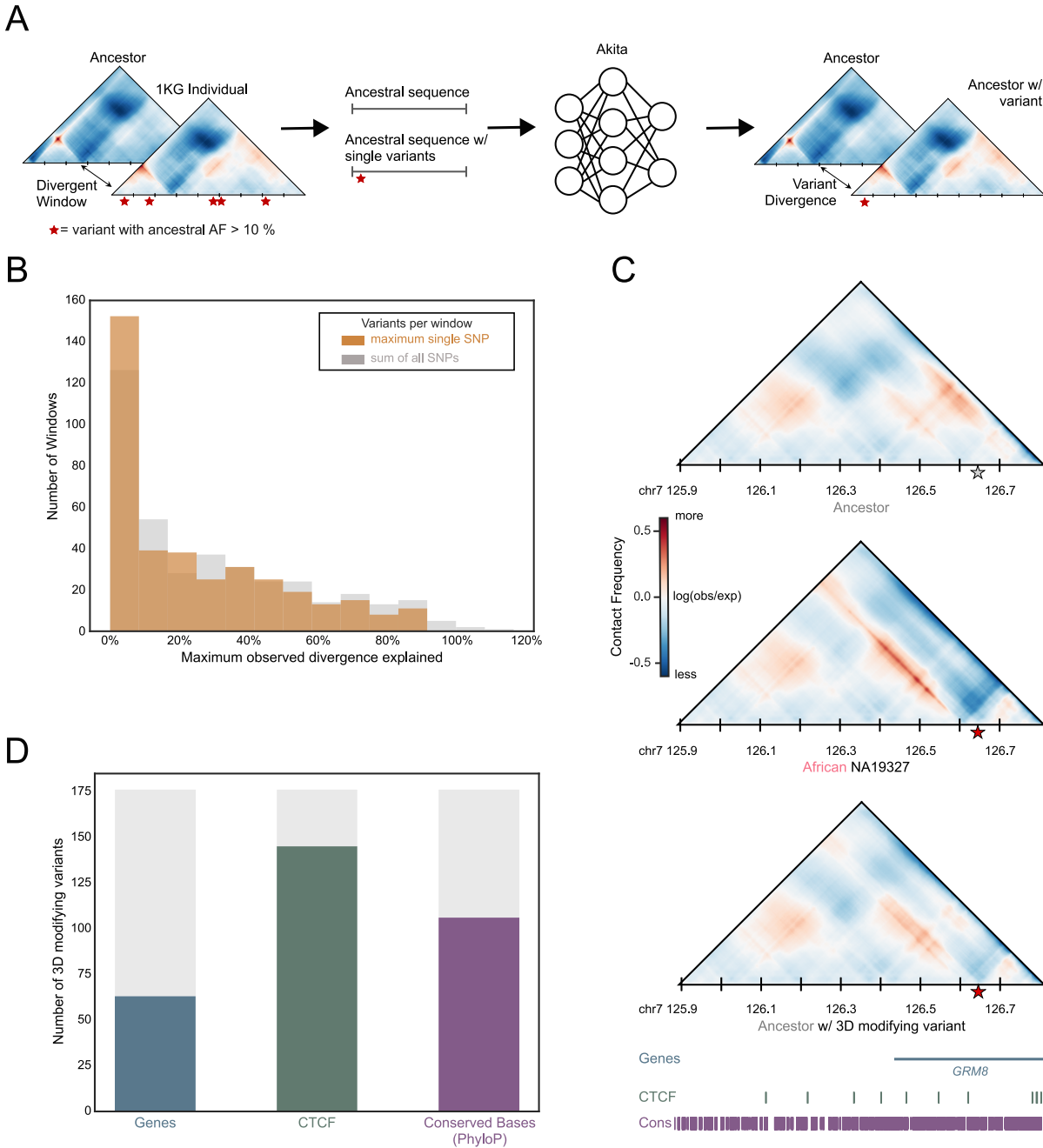


Figure 2.6: Most common divergent windows cannot be explained by a single nucleotide variant.

(A) We used *in silico* mutagenesis to estimate the contribution of individual SNVs to 3D genome differences in highly divergent windows. First, we extracted common—non-ancestral allele frequency (AF) > 10%—1KG SNVs (red stars) from the 392 windows with significantly greater 3D divergence across individuals than expected. We inserted the variants one-by-one into the human-archaic hominin ancestral genome and used Akita to generate chromatin contact predictions for the mutated sequences. Next, we calculated 3D divergence between the ancestral and mutated contact maps. **(B)** Distribution of single SNV effects for the maximally disruptive SNV per window (gray) and for the linear sum of all SNV effects (orange). SNV effects are calculated as the percent of maximum 3D divergence in a window between any 1KG individual and the ancestor that is observed in the mutated map. Variants that produce greater than 20% of the maximum observed 3D divergence in the window were designated 3D-modifying variants (N = 176). **(C)** Example SNV that recapitulates some, but not all, of the observed 3D divergence from ancestral in a 3D divergent window. The tracks below the contact map show locations of genes (blue), CTCF binding sites (green) and phastCons elements (purple). **(D)** Number of the 176 3D modifying variants that are in CTCF binding peaks, genes, and conserved bases (phyloP).

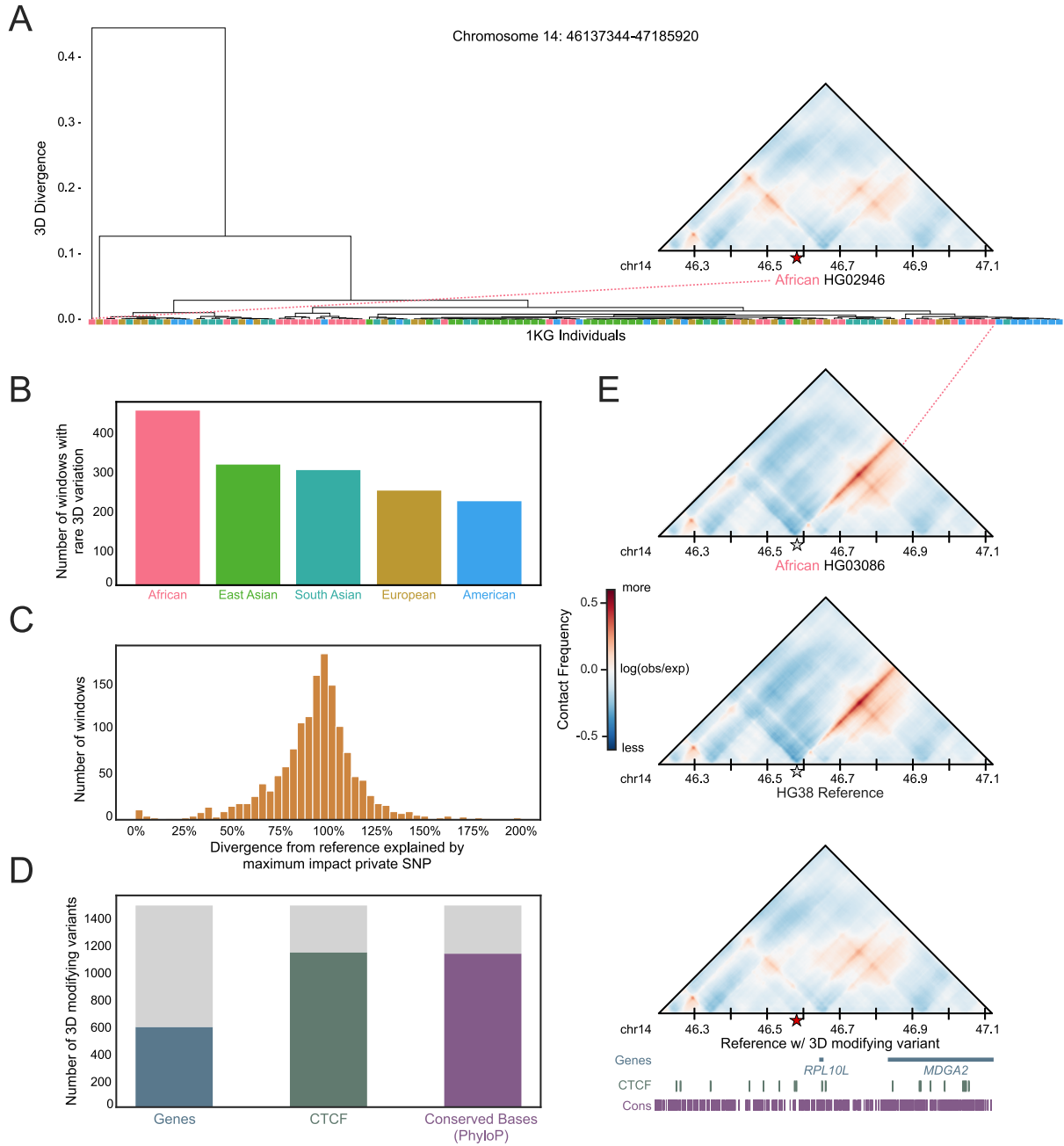
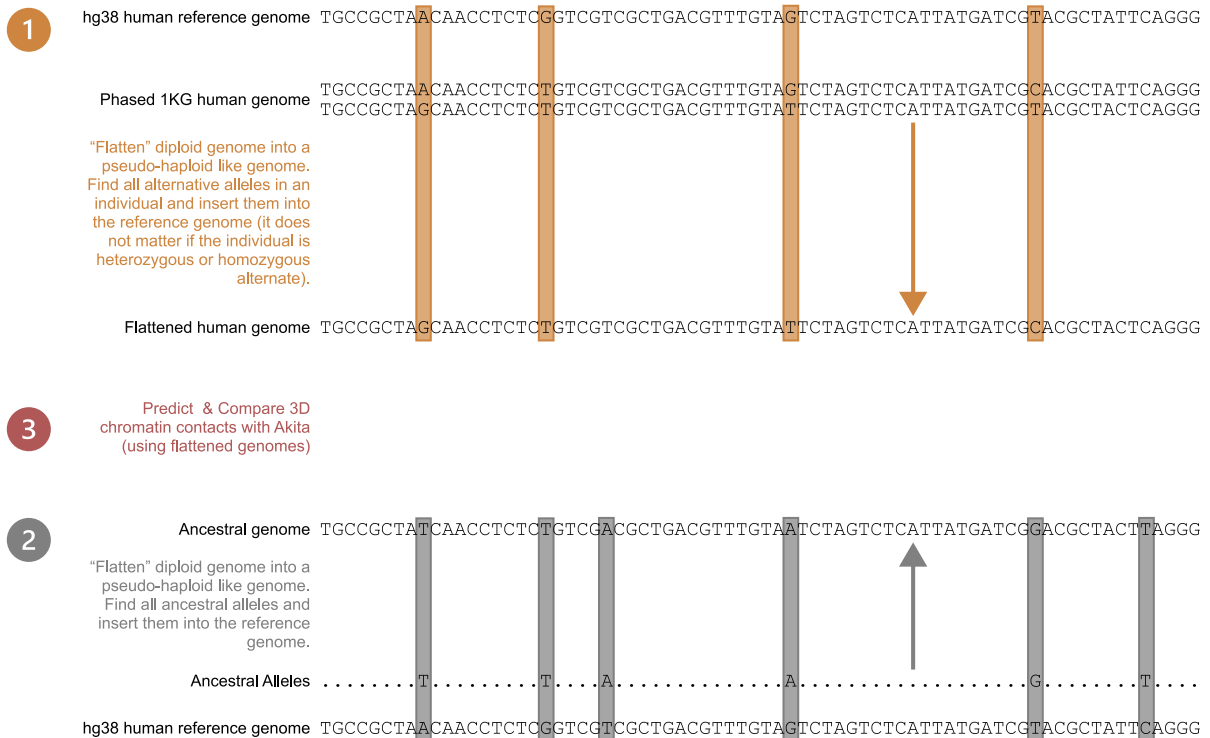


Figure 2.7: Genomic windows with rare variation in 3D contact patterns are common.

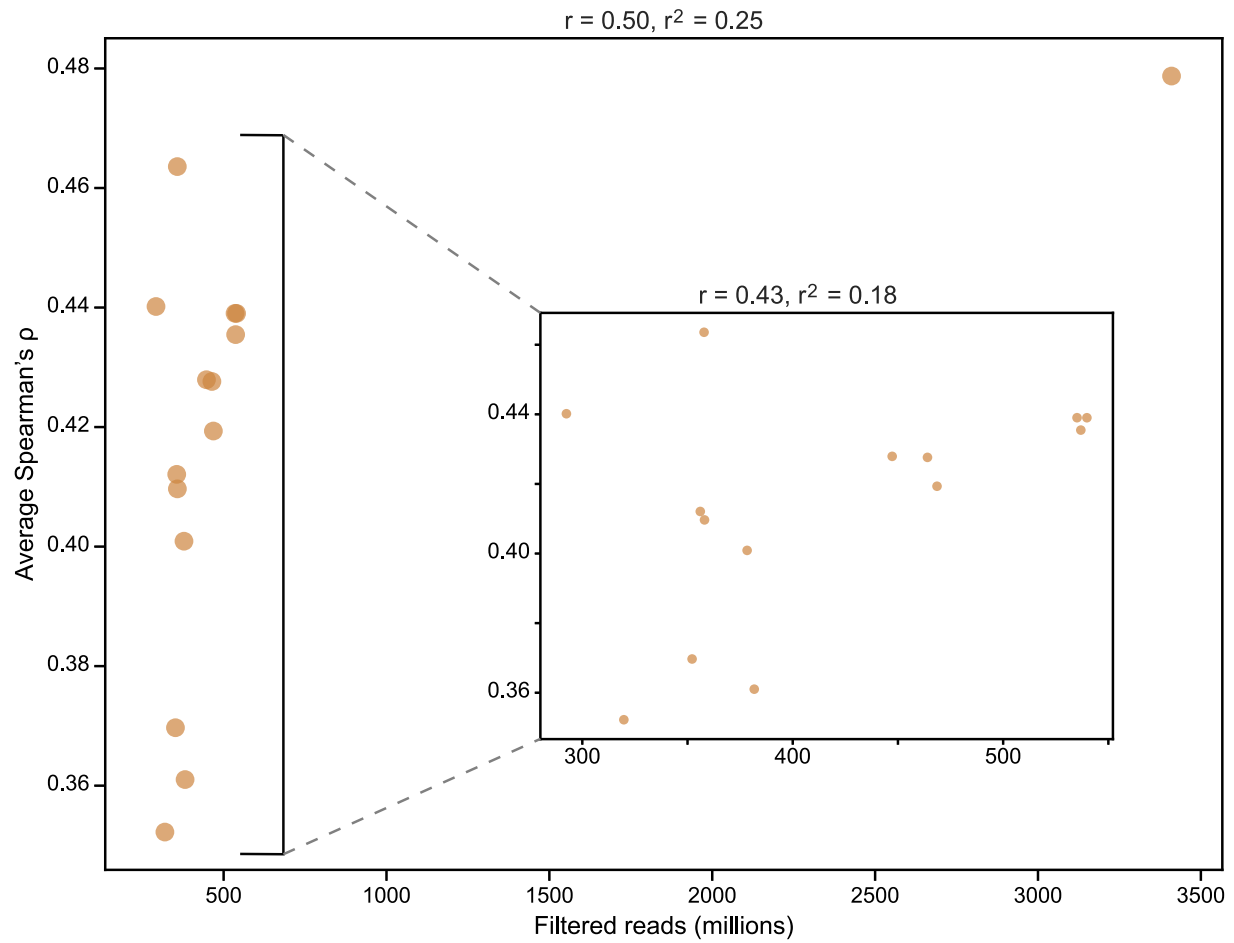
(A) Hierarchical clustering of individuals based on pairwise 3D divergence for an example window in which one individual is highly divergent from all others. The contact map for the divergent individual is shown in the top right. **(B)** Number of windows with rare 3D divergence stratified by continental origin of the rare individual. In total, 31% of windows in the genome have a rare divergent 3D contact pattern. **(C)** Distribution of single SNV effects estimated from *in silico* mutagenesis for the maximally disruptive SNV per window. SNV effects are calculated as the percent of maximum 3D divergence observed between a 1KG individual and the hg38 reference for a given window. **(D)** Number of the 3D modifying variants that are within CTCF binding peaks, genes, and conserved bases (phyloP). **(E)** Example of a single SNV that recovers the 3D divergence observed in the individual with rare 3D variation from **A** when placed into the reference sequence. The tracks below the contact map show the locations of genes (blue), CTCF binding sites (green) and phastCons elements (purple). The star represents the position of the tested SNV; it is red when the alternate allele is present and unfilled with the reference allele is present.

Supplemental Figures

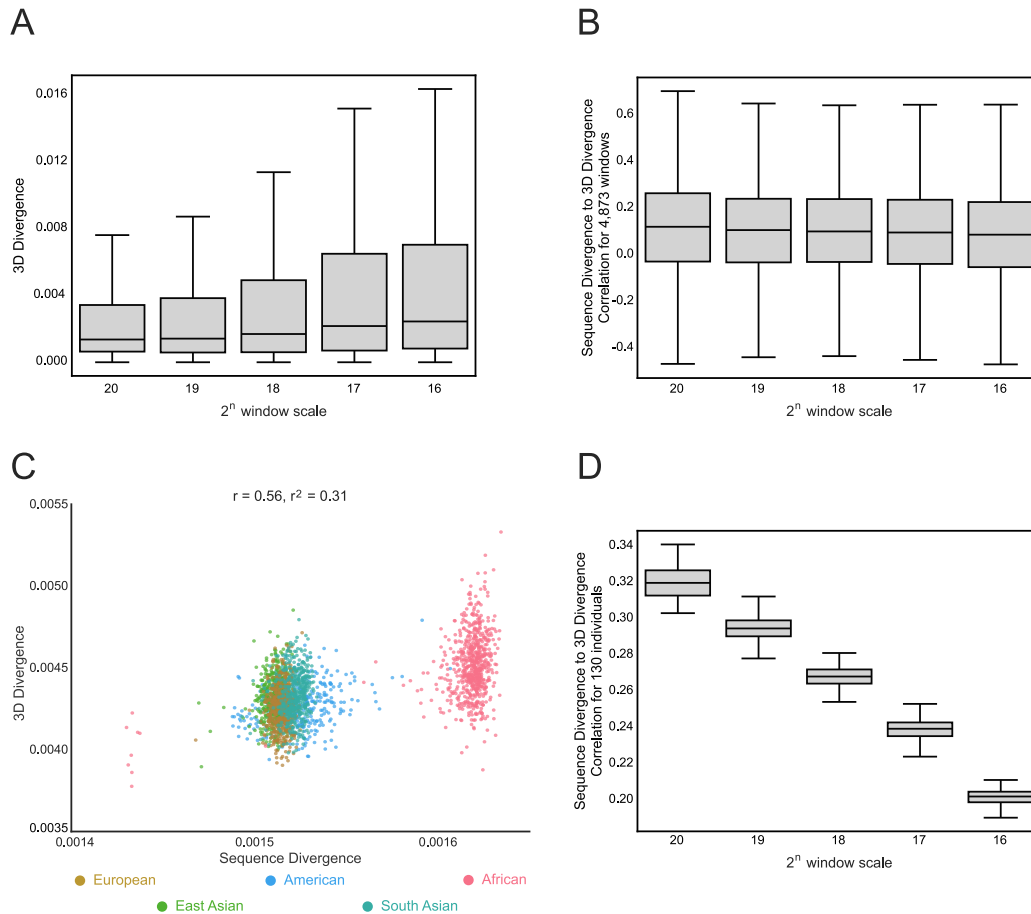


Supplemental Figure 2.1: Generating “flattened” genome sequences.

We constructed full-length genome sequences for each 1KG individual based on their genotyping information and for two ancestral sequences: human-archaic hominin and human-chimpanzee. Here, we illustrate a schematic of the procedure used. **(1)** If an individual had an alternate allele (homozygous or heterozygous), we inserted it into the reference genome to create a pseudo-haploid, or “flattened” genome for each individual (highlighted in orange boxes). **(2)** We also do this for ancestral alleles from both the human-archaic hominin and human-chimpanzee ancestors (highlighted in grey boxes) to facilitate appropriate comparisons. **(3)** We run Akita on each processed genome sequence separately and then compare the resulting contact maps.

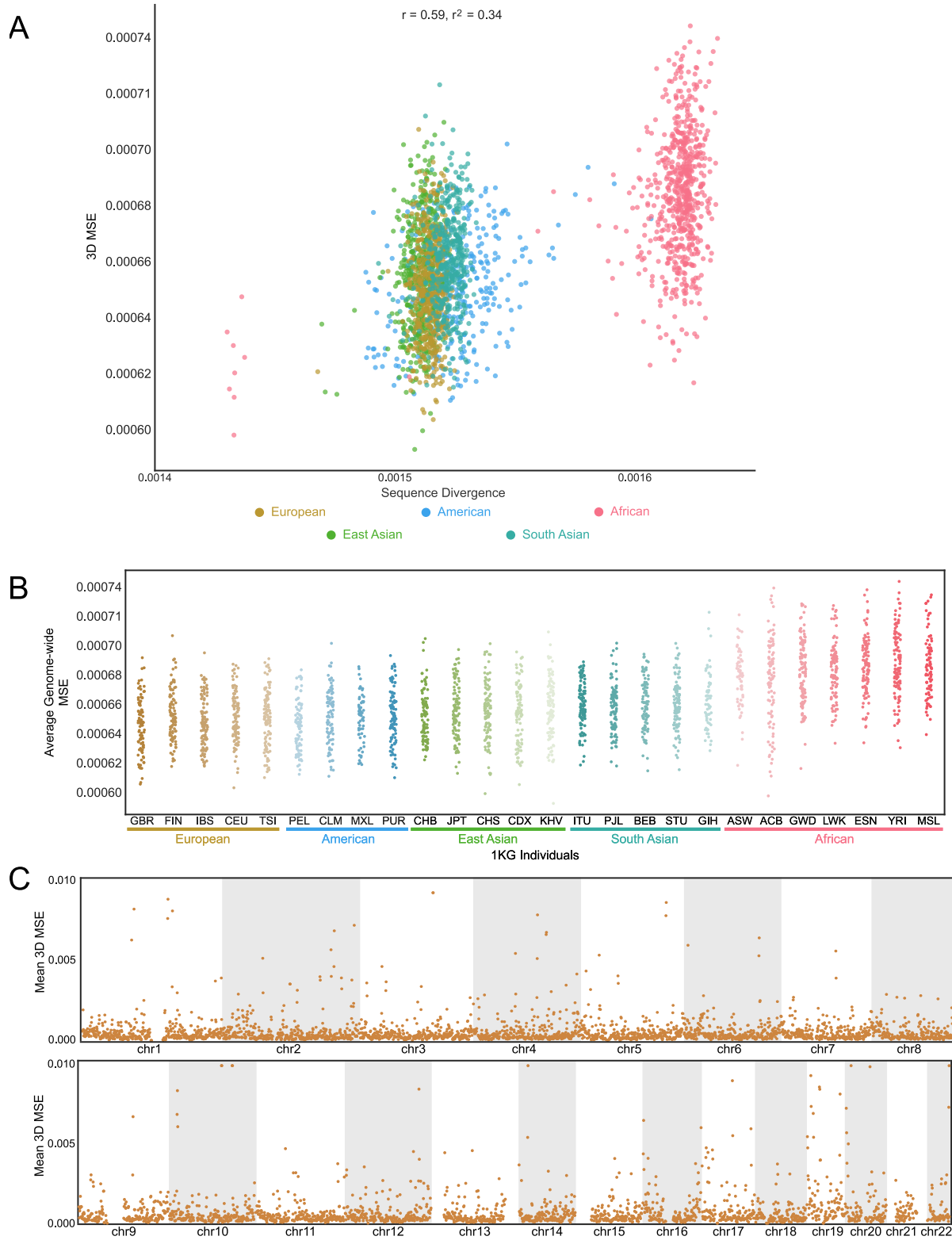


Supplemental Figure 2.2: Correlation between read count and prediction accuracy. Filtered read count in millions moderately correlated with genome-wide average Spearman's ρ (predicted vs experimental) for 15 individuals for which we have experimental Hi-C data and Akita predictions at 10 kb resolution.



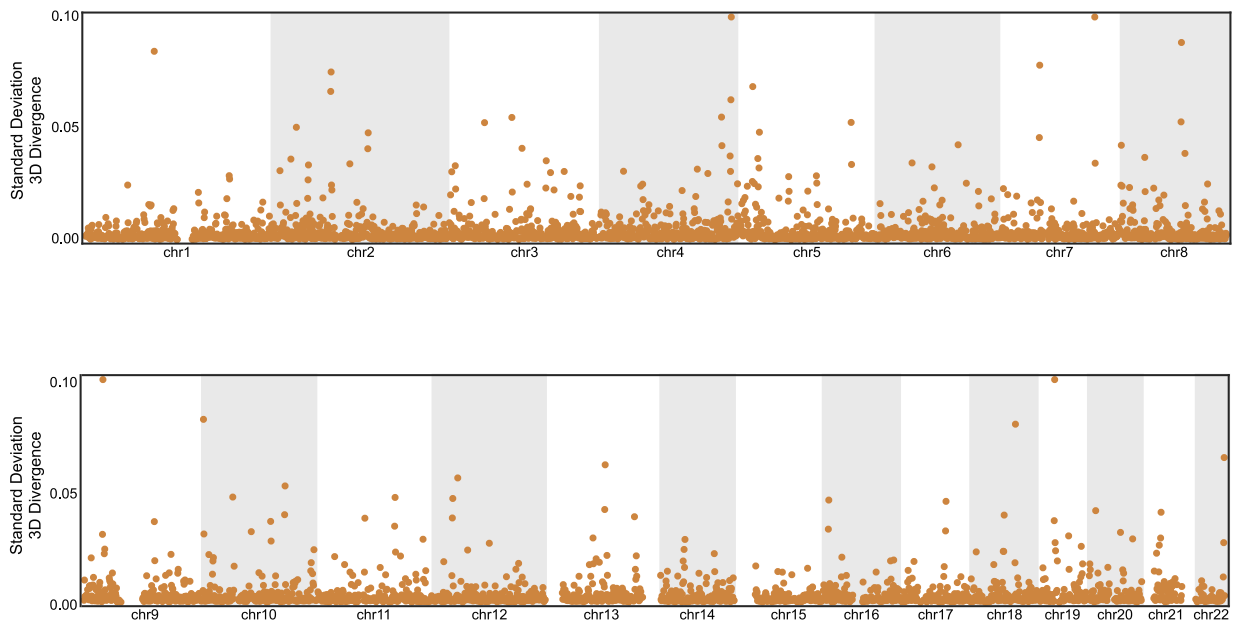
Supplemental Figure 2.3: 3D divergence estimates at different window sizes.

(A) To explore the sensitivity of our results to different window sizes, we compare predicted maps at five scales (2^{20} – 2^{16} bp). 3D divergence between 1KG individuals and the human-archaic hominin ancestor for all 1KG individuals increases on average with decreasing window size. **(B)** Distribution of Spearman correlation of pairwise SNV sequence divergence and 3D divergence between 1KG individuals and the human archaic-hominin ancestor for each window size. The low correlation between sequence and 3D divergence is not sensitive to window size. **(C)** The relationship between genome-wide average sequence and 3D divergence from the human–archaic hominin common ancestor for each 1KG individual at ~1 Mb window size. This correlation is primarily driven by the increased 3D divergence in African samples; when these are removed the correlation decreases substantially ($R^2 = 0.02$). Variation in correlation strength by population suggests that the relationship between sequence divergence and 3D genome organization is complex and may be influenced by population-specific factors. **(D)** The relationship between genome-wide average sequence and 3D divergence from the human–archaic hominin common ancestor for each of 130 1KG individuals at the five different window sizes. As sequence divergence is relatively constant with changes in window size, the genome-wide correlation between sequence and 3D divergence decreases with decreasing window size due to the increase in 3D divergence.



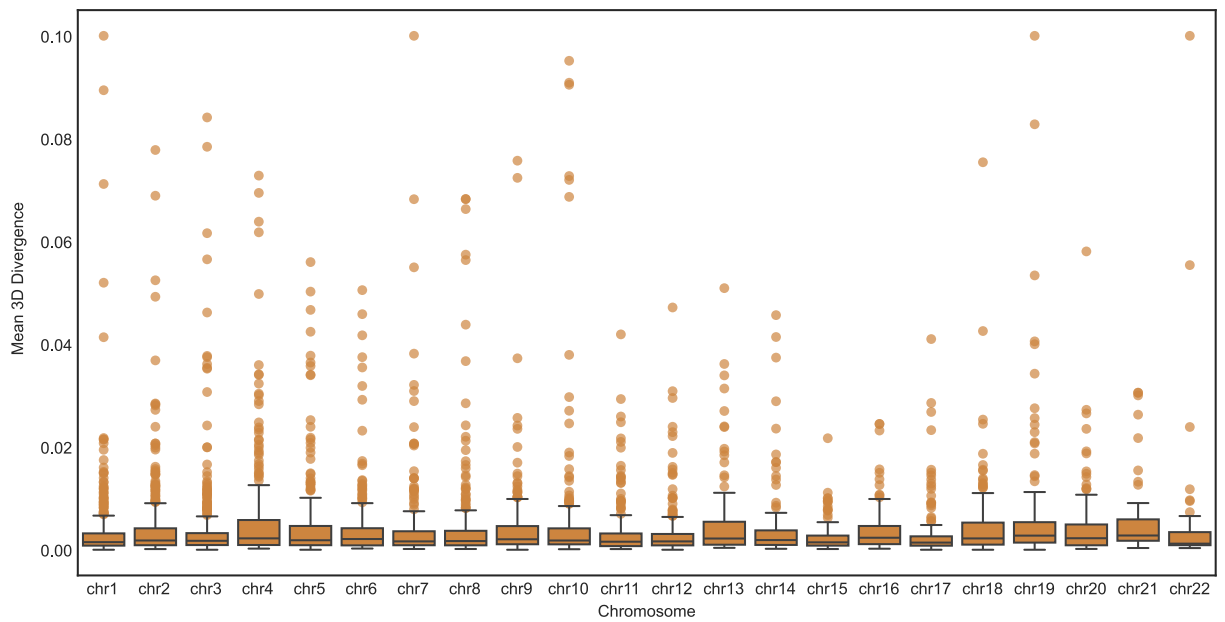
Supplemental Figure 2.4: MSE recapitulates divergence patterns calculated using 3D divergence.

(A) The relationship between genome-wide average sequence difference and 3D MSE from the human--archaic hominin common ancestor for each 1KG individual. **(B)** Genome-wide average 3D MSE for each 1KG individual, stratified by continental and sub-continental populations. Color indicates super-population and hue indicates sub-population. **(C)** Mean of genome-wide 3D MSE from the human-archaic hominin ancestor across 4,873 genomic windows of 2,457 individuals from 1KG. Each point represents the mean MSE of all individuals from the ancestral genome for a single genomic window. All points greater than 0.010 are clipped to 0.010 for visualization



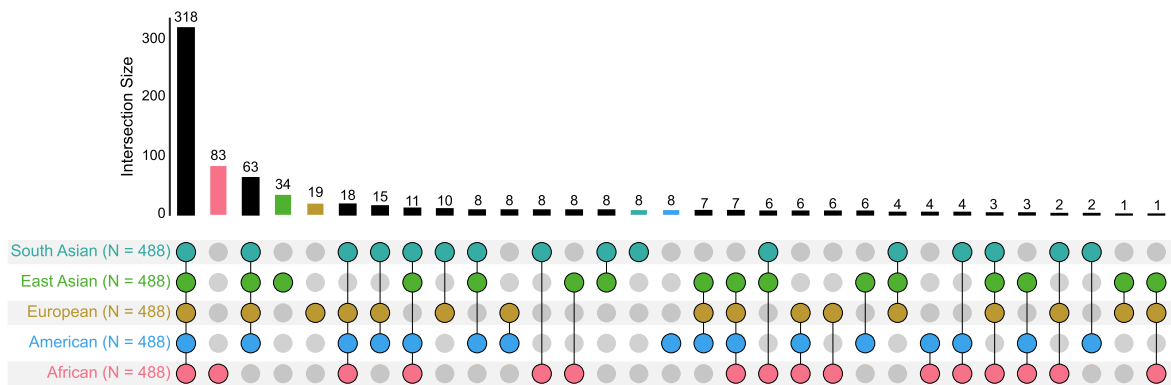
Supplemental Figure 2.5: Genome-wide standard deviation of 3D divergence between all 1KG individuals.

Standard deviation of genome-wide 3D divergence from the human-archaic hominin ancestor across 4,873 genomic windows of 2,457 individuals from 1KG. Each point represents the standard deviation of 3D divergence of all individuals from the ancestral genome for a single genomic window. All points greater than 0.10 are clipped to 0.10 for visualization.



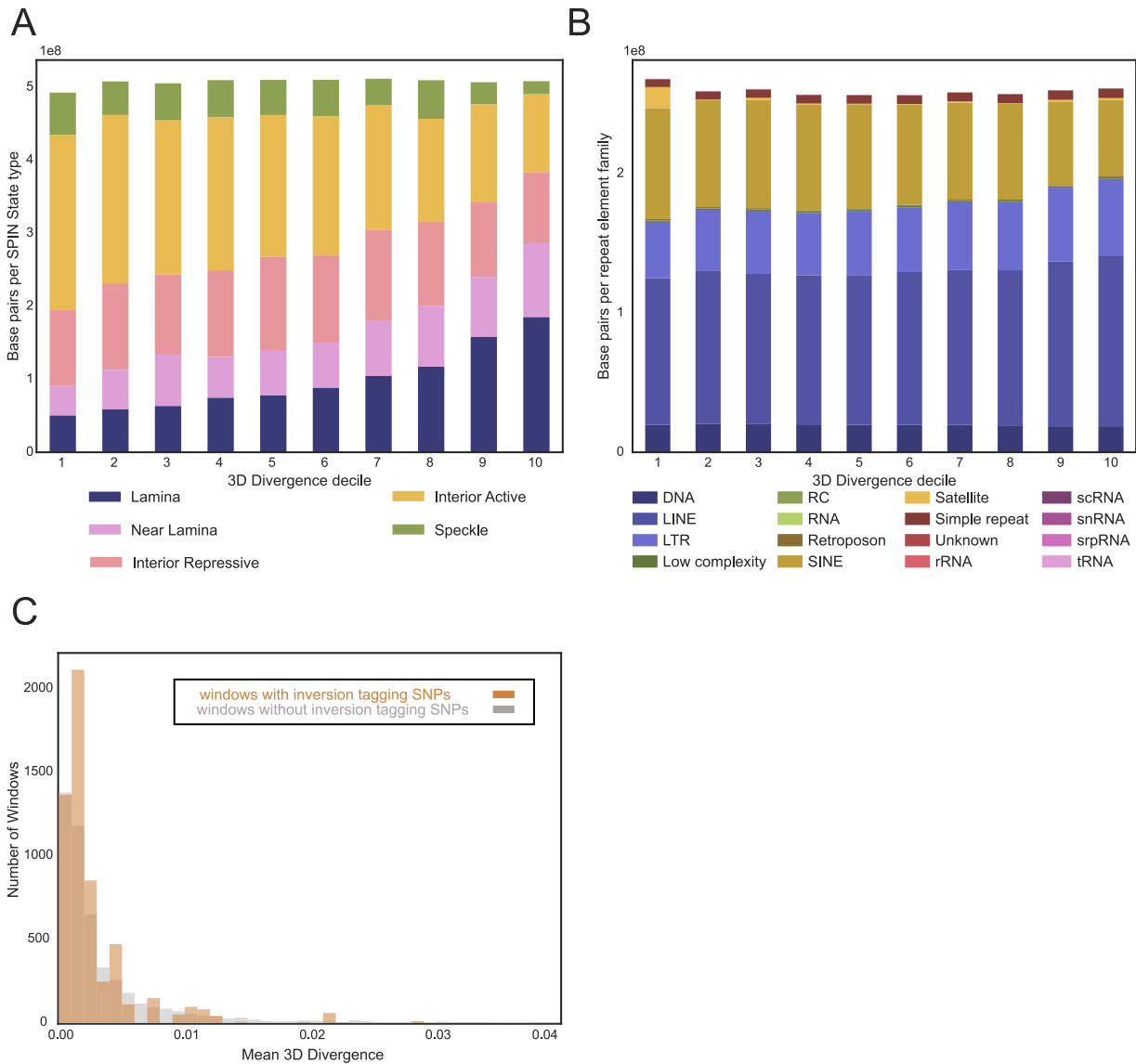
Supplemental Figure 2.6: Chromosome distributions of 3D divergence.

Mean 3D divergence from the human-archaic hominin ancestor across the 22 autosomal chromosomes of 2,457 individuals from 1KG. Each point represents the mean of 3D divergence of all individuals from the ancestral genome for a single genomic window on the specified chromosome. All points greater than 0.10 are clipped to 0.10 for visualization.



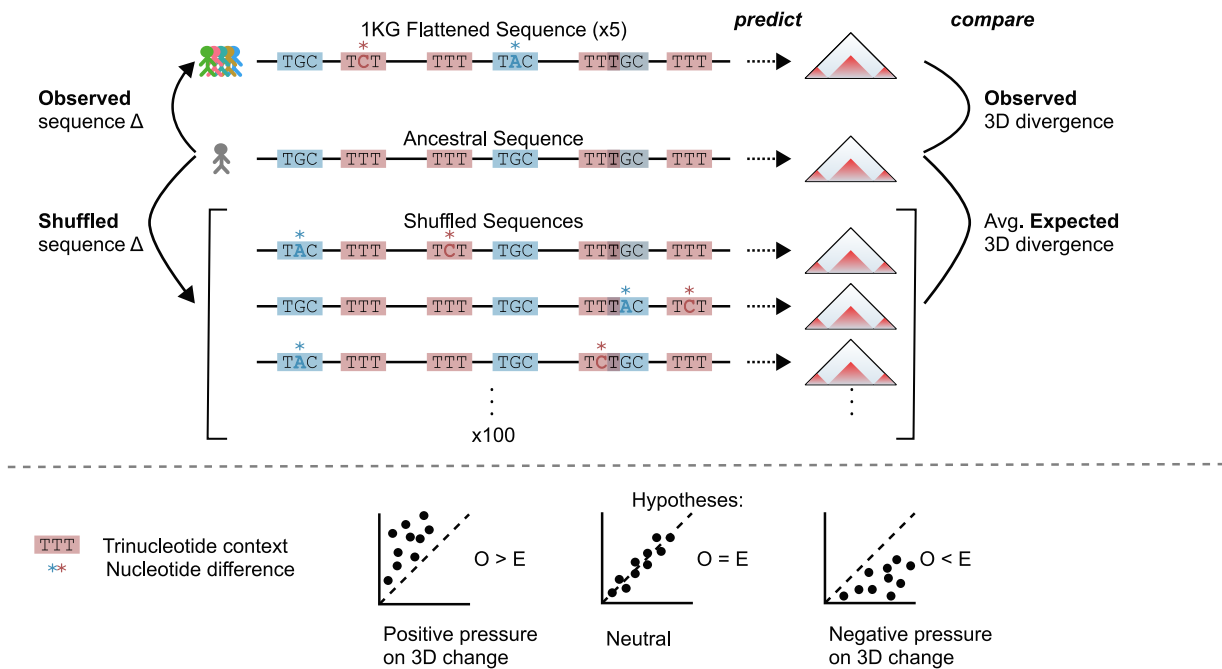
Supplemental Figure 2.7: Upset plot of population representation in top 10% 3D divergent windows.

Unique and shared top 10% divergent windows among 1KG super-populations. Bars indicate the number of windows and the dot matrix indicates the populations represented by each set.



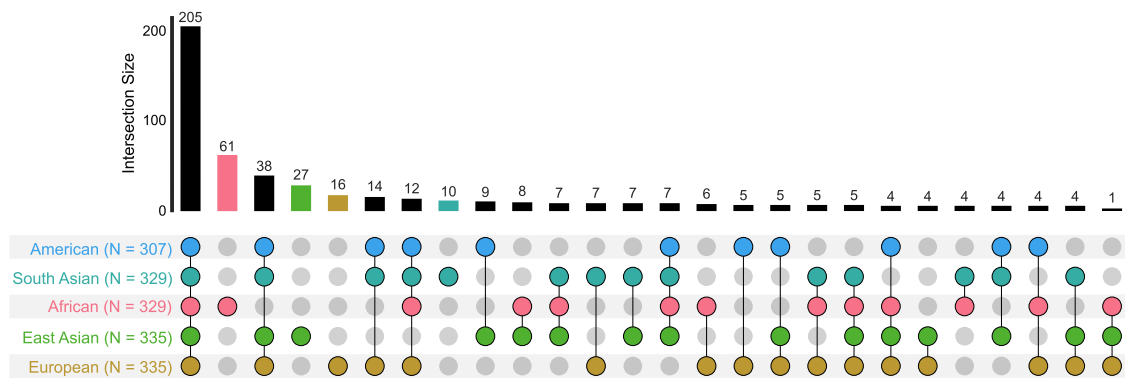
Supplemental Figure 2.8: SPIN state and repeat element content across 3D divergence deciles.

(A) SPIN states as in Wang et al. 2021, from the HFF cell line. The number of base pairs assigned to each SPIN state across all windows in a given decile of 3D divergence. **(B)** The number of RMSK repeat element base pairs of each repeat family is shown according to presence each decile of 3D divergence. **(C)** The distribution of 3D divergence from the ancestor for windows with (orange) and without (grey) inversion tagging SNPs.



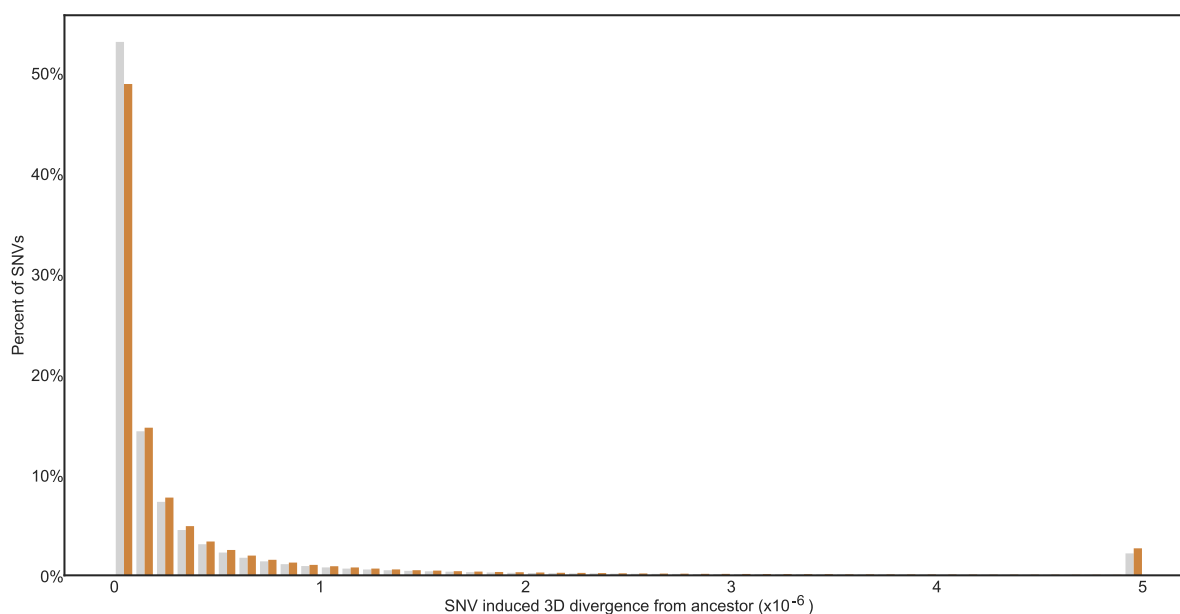
Supplemental Figure 2.9: Generating empirical distribution of expected 3D divergence.

To evaluate whether 3D genome organization constrained sequence divergence, we estimated the null distribution of expected 3D divergence based on sequence differences between the 1KG (HG03105 [African], HG01119 [American], NA06985 [European], HG00759 [East Asian], HG03007 [South Asian]) and human-archaic hominin ancestral genomes. We shuffled observed nucleotide differences (stars) while preserving trinucleotide context (colored rectangles) and predicted 3D genome organization for 100 shuffled sequences for each window. This is done with variants from each of 5 individuals for a total of 500 shuffled sequences per window. If there is no sequence constraint to maintain 3D organization, observed 3D divergence would equal the expected 3D divergence ($O = E$). Alternatively, observing more 3D divergence than expected would suggest positive selection on sequence changes that cause 3D divergence ($O > E$). Finally, observing less 3D divergence than expected would suggest negative pressure on sequence changes that cause 3D divergence ($O < E$).



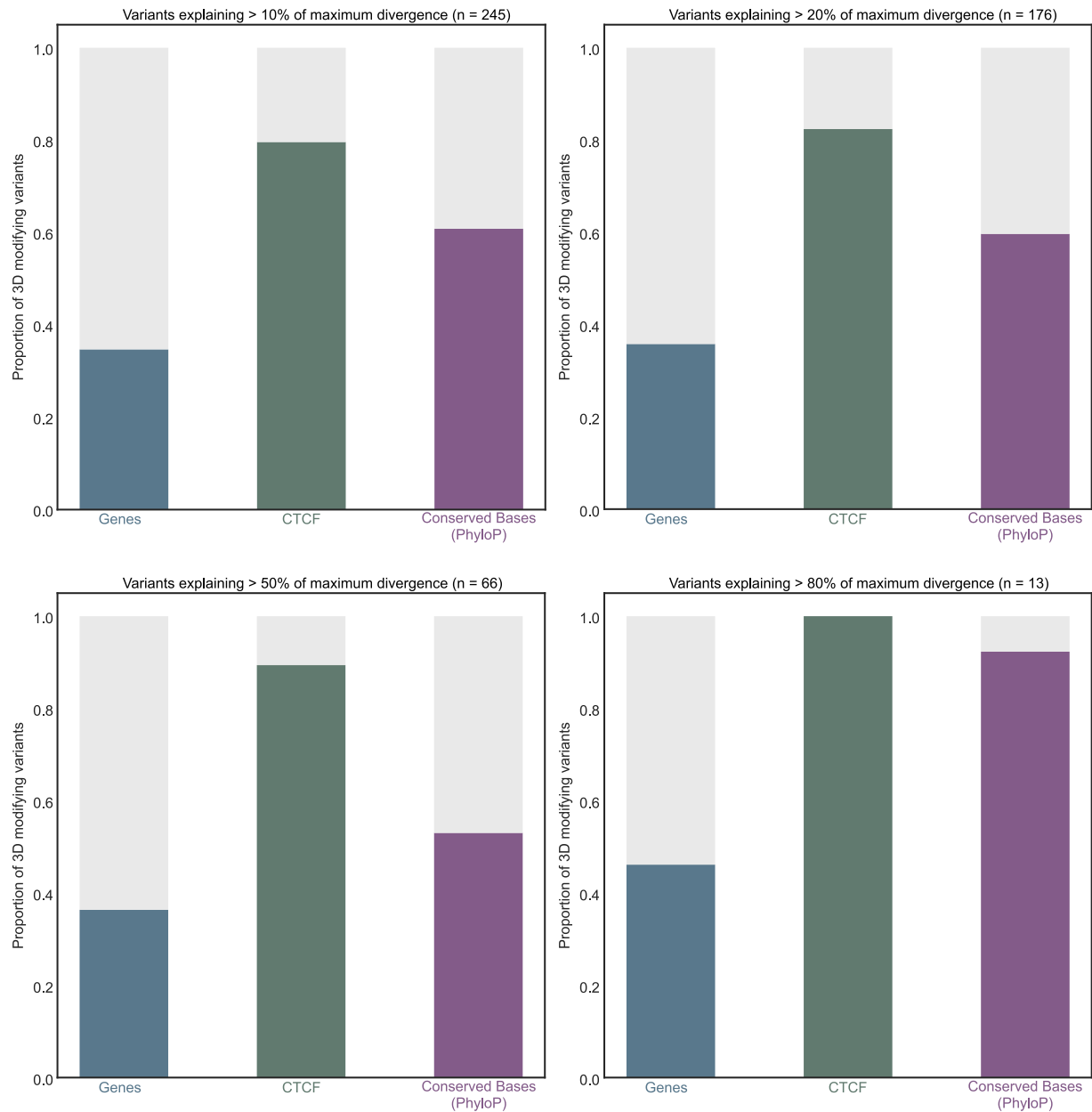
Supplemental Figure 2.10: Full upset plot for more divergent than expected windows.

Unique and shared divergent windows among 1KG super-populations. Bars indicate the number of windows and the dot matrix indicates the populations represented by each set.



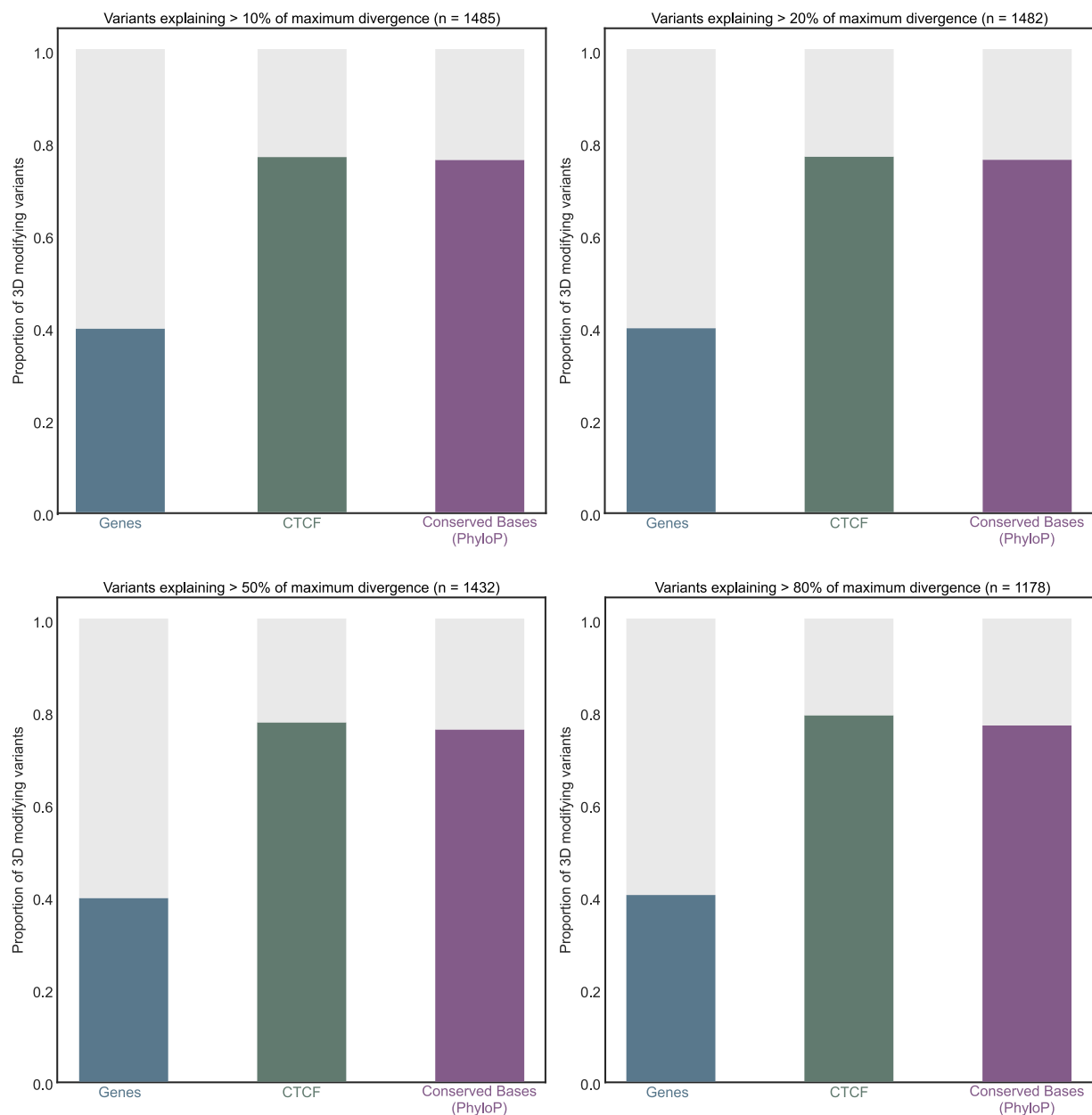
Supplemental Figure 2.11: Distributions of 3D divergence from *in silico* mutagenesis on common variants from the 392 divergent windows and 392 non-divergent windows.

The distribution of 3D divergence from the ancestor induced by a SNV in the 392 divergent windows (orange) and 392 randomly selected non-divergent windows (grey). All points greater than 0.000005 are clipped to 0.000005 for visualization. The overall distribution of 3D divergence for SNVs from less divergent windows is qualitatively similar to that from divergent windows, as expected, since we anticipate that most SNVs do not influence 3D divergence. However, we observe that a larger fraction of SNVs in divergent windows cause high 3D divergence (0.06%) compared to the SNVs from non-divergent windows (0.03%).



Supplemental Figure 2.12: Genomic annotations for 3D modifying variants in divergent windows are consistent across cutoffs.

Number of the 3D modifying variants that are in CTCF binding peaks, genes, and conserved bases (phyloP) at 4 different 3D divergence explained cutoffs: 10% (n = 245), 20% (n = 176), 50% (n = 66) and 80% (n = 13).



Supplemental Figure 2.13: Genomic annotations for 3D modifying variants in windows with rare 3D divergence are consistent across cutoffs.

Number of the 3D modifying variants that are in CTCF binding peaks, genes, and conserved bases (phyloP) at 4 different 3D divergence explained cutoffs: 10% (n = 1,485), 20% (n = 1,482), 50% (n = 1,432) and 80% (n = 1,178)

CHAPTER 3: FINAL THOUGHTS AND FUTURE DIRECTIONS

Summary of key findings

This dissertation has explored the diversity of 3D chromatin contact patterns across human populations using machine learning methods to predict these interactions from genome sequence data, which is easier and less costly to generate than directly measuring 3D genome folding via chromatin capture assay. A critical finding of this research is the identification of significant variation in 3D genome organization among different individuals, with African populations exhibiting the highest levels of 3D genome diversity. This observation aligns with previous studies on sequence diversity but introduces a new dimension by examining the spatial organization of the genome.

The study also reveals that 3D chromatin contact divergence does not strongly correlate with sequence divergence. This dissociation is particularly evident in specific genomic regions where 3D divergence is significantly greater than expected based on sequence variation alone. These findings suggest that 3D genome structure is subject to distinct evolutionary pressures that may not be captured by sequence data alone, highlighting the importance of considering 3D chromatin architecture in studies of human genetics and gene expression variation.

Moreover, the research underscores the functional implications of 3D chromatin variation, particularly in regions with low functional constraint. These regions, where 3D divergence was found to be highest, may indicate that such areas are more permissive to 3D changes without disrupting essential functions. This flexibility in chromatin organization could allow for regulatory innovation, though this is more likely due to a lack

of purifying selection rather than direct evidence of adaptive processes. This insight contributes to our understanding of how gene regulation can vary and evolve within different human populations, emphasizing the potential for 3D genome variation to influence regulatory landscapes.

Significance of the work

The findings of this dissertation contribute to the field of genomics, particularly in the study of 3D genome folding by leveraging machine learning to predict 3D chromatin structures from sequence data. This work addresses a significant gap in our understanding of 3D genome diversity. The ability to generate 3D genome maps for over 2,400 diverse humans from all major continental populations provides a foundational resource for future studies investigating the relationship between genome structure, gene regulation, and phenotypic diversity.

While the direct links between 3D genome organization and gene expression across populations remain to be fully elucidated, the implications of this work extend beyond 1D genomic studies. By providing a deeper understanding of 3D chromatin structure across diverse human populations, this research highlights the substantial shared 3D genome variation that exists among populations. These findings underscore the idea that human genetic diversity exists on a continuous spectrum rather than within discrete population groups. Recognizing the full scope of this diversity is crucial for advancing our understanding of genome functionality and avoiding the creation of artificial or arbitrary group labels.

The integration of 3D genome data into broader genomic studies promises to reveal specific regulatory mechanisms, such as how enhancer-promoter interactions vary across different populations, potentially explaining population-specific gene expression patterns. This approach could also illuminate the evolutionary pressures that have shaped chromatin architecture over time, offering insights into how these 3D structures have influenced genetic adaptation and survival. Furthermore, by examining the spatial organization of the genome, we may uncover patterns that contribute to phenotypic diversity, revealing how 3D genome organization impacts traits and disease susceptibility across diverse human populations. Collectively, these findings could significantly enhance our understanding of the complex relationship between genome structure and function.

Challenges and limitations

Despite the significant contributions of this research, several challenges and limitations should be acknowledged. One of the primary challenges was the reliance on current machine learning models, which, while powerful, have inherent limitations in accurately predicting 3D chromatin interactions across different cell types and contexts. These models are trained on existing data, which is limited by resolution and cell-type availability.

Data availability and resolution also posed limitations. Although the use of machine learning allowed for the prediction of 3D chromatin contact maps across many individuals, the resolution of these maps is still limited compared to what can be achieved with high-resolution experimental techniques like Hi-C. The availability of more comprehensive and

higher-resolution experimental data would enable more accurate predictions and a better understanding of 3D genome diversity.

Moreover, while the predicted 3D folding patterns have been shown to be generally accurate, the validation of these predictions at the individual level remains a crucial step. Validation through experimental methods, such as high-resolution chromatin conformation capture techniques, is necessary to confirm the accuracy of these predictions and to ensure that the inferred structures correspond to the true chromatin organization in various cell types and conditions. However, even if the 3D predictions are validated, the current work does not directly link these folding patterns to gene expression. Thus, further research is needed to establish a connection between the variation in 3D structures and gene expression variation.

Future directions

There are several promising directions for future research that build on the findings of this dissertation. First, expanding the diversity of the populations studied is essential. Including more underrepresented groups, particularly from regions with high levels of genetic diversity, such as Africa, Oceania, and South Asia, will provide a more complete picture of global 3D genome variation.

Second, integrating 3D genome data with other -omics data, such as gene expression patterns, epigenetic modifications, and protein levels, is another crucial step towards uncovering the intricate regulatory networks that drive phenotypic diversity and disease. Multi-omics approaches can offer a more holistic understanding of how 3D

genome organization influences gene regulation and phenotype. Exciting future work will include directly linking these 3D genome changes to changes in gene regulation. The work presented here will provide an important foundation that can be combined with new gene expression data in 1KG individuals from the MAGE dataset (Taylor et al. 2024) to prioritize regions of the genome for further experimentation.

Third, advancing machine learning models to improve the accuracy of 3D chromatin contact predictions is also a critical area for future research. Developing models that can account for cell-type-specific interactions and integrating additional layers of genomic information, such as histone modifications and transcription factor binding sites, will enhance the predictive power and applicability of these tools. In addition, these models should be paired with experimental validation to confirm the predicted chromatin structures and their functional relevance.

Finally, future studies could focus on characterizing the effects of 3D chromatin changes on various aspects of genome function, such as 1D chromatin states, gene expression, and transcription factor binding. Understanding how changes in 3D genome architecture influence these factors will provide deeper insights into the regulatory mechanisms that underlie phenotypic diversity and could inform the development of new therapeutic strategies. Applying these approaches to disease-specific studies holds significant potential. By focusing on diseases where regulatory disruptions are known to play a role, such as cancer and neurodevelopmental disorders, researchers can uncover new insights into disease mechanisms. Predicting how genetic variants influence 3D chromatin structure will be invaluable in identifying novel biomarkers and therapeutic targets.

Concluding remarks

In conclusion, this dissertation has provided insights into the diversity of 3D chromatin contact patterns across human populations and their implications for gene regulation and disease. The use of machine learning to predict 3D genome structures represents a significant advancement in the field, offering new avenues for exploring the relationship between genome organization and phenotype.

The findings of this study underscore the importance of considering 3D genome organization in genetic research and personalized medicine. By expanding our understanding of how 3D chromatin structures vary across populations, we can better appreciate the complexity of gene regulation and its role in human health and disease.

As we move forward, the integration of 3D genome data with other genomic and phenotypic data, the development of more sophisticated predictive models, and the application of these tools to disease research will continue to drive the field toward a more comprehensive understanding of the human genome. The potential for this research to contribute to personalized medicine and targeted therapies is immense. Key to realizing this potential will be the validation of predicted 3D structures, integration with clinical data, and the development and validation of models that connect 3D genome organization with regulatory functions. Ultimately, fostering collaborations across disciplines will be essential in translating these insights into clinical practice, promising a future where the full complexity of the genome is harnessed to improve human health.

REFERENCES

- 1000 Genomes Project Consortium, Auton, A., Brooks, L.D., Durbin, R.M., Garrison, E.P., Kang, H.M., Korbel, J.O., Marchini, J.L., McCarthy, S., McVean, G.A., and Abecasis, G.R. 2015. A global reference for human genetic variation. *Nature* **526**(7571): 68–74. doi:10.1038/nature15393.
- Acemel, R.D., and Lupiáñez, D.G. 2023. Evolution of 3D chromatin organization at different scales. *Curr. Opin. Genet. Dev.* **78**: 102019. doi:10.1016/j.gde.2022.102019.
- Agarwal, V., and Shendure, J. 2020. Predicting mRNA Abundance Directly from Genomic Sequence Using Deep Convolutional Neural Networks. *Cell Rep.* **31**(7): 107663. doi:10.1016/j.celrep.2020.107663.
- Aggarwala, V., and Voight, B.F. 2016. An expanded sequence context model broadly explains variability in polymorphism levels across the human genome. *Nat. Genet.* **48**(4): 349–355. Nature Publishing Group. doi:10.1038/ng.3511.
- Agrawal, P., Heimbruch, K.E., and Rao, S. 2018. Genome-wide maps of transcription regulatory elements and transcription enhancers in development and disease. *Compr. Physiol.* **9**(1): 439–455. doi:10.1002/cphy.c180028.
- Alemu, E.Y., Carl, J.W., Jr, Corrada Bravo, H., and Hannenhalli, S. 2014. Determinants of expression variability. *Nucleic Acids Res* **42**(6): 3503–3514. doi:10.1093/nar/gkt1364.
- Altshuler, D., Donnelly, P., and The International HapMap Consortium. 2005. A haplotype map of the human genome. *Nature* **437**(7063): 1299–1320. Nature Publishing Group. doi:10.1038/nature04226.

- Angermueller, C., Pärnamaa, T., Parts, L., and Stegle, O. 2016. Deep learning for computational biology. *Mol. Syst. Biol.* **12**(7): 878. John Wiley & Sons, Ltd. doi:10.15252/msb.20156651.
- Armstrong, J., Hickey, G., Diekhans, M., Fiddes, I.T., Novak, A.M., Deran, A., Fang, Q., Xie, D., Feng, S., Stiller, J., Genereux, D., Johnson, J., Marinescu, V.D., Alföldi, J., Harris, R.S., Lindblad-Toh, K., Haussler, D., Karlsson, E., Jarvis, E.D., Zhang, G., and Paten, B. 2020. Progressive Cactus is a multiple-genome aligner for the thousand-genome era. *Nature* **587**(7833): 246–251. Nature Publishing Group. doi:10.1038/s41586-020-2871-y.
- Avsec, Ž., Agarwal, V., Visentin, D., Ledsam, J.R., Grabska-Barwinska, A., Taylor, K.R., Assael, Y., Jumper, J., Kohli, P., and Kelley, D.R. 2021a. Effective gene expression prediction from sequence by integrating long-range interactions. *Nat. Methods* **18**(10): 1196–1203. Nature Publishing Group. doi:10.1038/s41592-021-01252-x.
- Avsec, Ž., Weilert, M., Shrikumar, A., Krueger, S., Alexandari, A., Dalal, K., Fropf, R., McAnany, C., Gagneur, J., Kundaje, A., and Zeitlinger, J. 2021b. Base-resolution models of transcription-factor binding reveal soft motif syntax. *Nat. Genet.* **53**(3): 354–366. Nature Publishing Group. doi:10.1038/s41588-021-00782-6.
- Baldi, P., and Brunak, S. 2001. *Bioinformatics The Machine Learning Approach*. In 2nd edition. The MIT Press.
- Barreiro, L.B., Laval, G., Quach, H., Patin, E., and Quintana-Murci, L. 2008. Natural selection has driven population differentiation in modern humans. *Nat. Genet.* **40**(3): 340–345. Nature Publishing Group. doi:10.1038/ng.78.

- Bejerano, G., Pheasant, M., Makunin, I., Stephen, S., Kent, W.J., Mattick, J.S., and Haussler, D. 2004. Ultraconserved Elements in the Human Genome. *Science* **304**(5675): 1321–1325. American Association for the Advancement of Science. doi:10.1126/science.1098119.
- Bell, A.C., West, A.G., and Felsenfeld, G. 2001. Insulators and Boundaries: Versatile Regulatory Elements in the Eukaryotic Genome. *Science* **291**(5503): 447–450. American Association for the Advancement of Science. doi:10.1126/science.291.5503.447.
- Bergström, A., McCarthy, S.A., Hui, R., Almarri, M.A., Ayub, Q., Danecek, P., Chen, Y., Felkel, S., Hallast, P., Kamm, J., Blanché, H., Deleuze, J.-F., Cann, H., Mallick, S., Reich, D., Sandhu, M.S., Skoglund, P., Scally, A., Xue, Y., Durbin, R., and Tyler-Smith, C. 2020. Insights into human genetic variation and population history from 929 diverse genomes. *Science* **367**(6484). doi:10.1126/science.aay5012.
- Besenbacher, S., Hvilsom, C., Marques-Bonet, T., Mailund, T., and Schierup, M.H. 2019. Direct estimation of mutations in great apes reconciles phylogenetic dating. *Nat. Ecol. Evol.* **3**(2): 286–292. Nature Publishing Group. doi:10.1038/s41559-018-0778-x.
- Bick, A.G., Metcalf, G.A., Mayo, K.R., Lichtenstein, L., Rura, S., Carroll, R.J., Musick, A., Linder, J.E., Jordan, I.K., Nagar, S.D., Sharma, S., Meller, R., Basford, M., Boerwinkle, E., Cicek, M.S., Doheny, K.F., Eichler, E.E., Gabriel, S., Gibbs, R.A., Glazer, D., Harris, P.A., Jarvik, G.P., Philippakis, A., Rehm, H.L., Roden, D.M., Thibodeau, S.N., Topper, S., Blegen, A.L., Wirkus, S.J., Wagner, V.A., Meyer, J.G., Cicek, M.S., Muzny, D.M., Venner, E., Mawhinney, M.Z., Griffith, S.M.L.,

Hsu, E., Ling, H., Adams, M.K., Walker, K., Hu, J., Doddapaneni, H., Kovar, C.L., Murugan, M., Dugan, S., Khan, Z., Boerwinkle, E., Lennon, N.J., Austin-Tse, C., Banks, E., Gatzen, M., Gupta, N., Henricks, E., Larsson, K., McDonough, S., Harrison, S.M., Kachulis, C., Lebo, M.S., Neben, C.L., Steeves, M., Zhou, A.Y., Smith, J.D., Frazar, C.D., Davis, C.P., Patterson, K.E., Wheeler, M.M., McGee, S., Lockwood, C.M., Shirts, B.H., Pritchard, C.C., Murray, M.L., Vasta, V., Leistritz, D., Richardson, M.A., Buchan, J.G., Radhakrishnan, A., Krumm, N., Ehmen, B.W., Schwartz, S., Aster, M.M.T., Cibulskis, K., Haessly, A., Asch, R., Cremer, A., Degatano, K., Shergill, A., Gauthier, L.D., Lee, S.K., Hatcher, A., Grant, G.B., Brandt, G.R., Covarrubias, M., Banks, E., Able, A., Green, A.E., Carroll, R.J., Zhang, J., Condon, H.R., Wang, Y., Dillon, M.K., Albach, C.H., Baalawi, W., Choi, S.H., Wang, X., Rosenthal, E.A., Ramirez, A.H., Lim, S., Nambiar, S., Ozenberger, B., Wise, A.L., Lunt, C., Ginsburg, G.S., Denny, J.C., The All of Us Research Program Genomics Investigators, Manuscript Writing Group, All of Us Research Program Genomics Principal Investigators, Biobank, M., Genome Center: Baylor-Hopkins Clinical Genome Center, Genome Center: Broad, C., and Mass General Brigham Laboratory for Molecular Medicine, Genome Center: University of Washington, Data and Research Center, All of Us Research Demonstration Project Teams, and NIH All of Us Research Program Staff. 2024. Genomic data in the All of Us Research Program. *Nature* **627**(8003): 340–346. Nature Publishing Group. doi:10.1038/s41586-023-06957-x.

Bird, C.P., Stranger, B.E., Liu, M., Thomas, D.J., Ingle, C.E., Beazley, C., Miller, W., Hurles, M.E., and Dermitzakis, E.T. 2007. Fast-evolving noncoding sequences in the human genome. *Genome Biol.* **8**(6): R118. doi:10.1186/gb-2007-8-6-r118.

Blanchette, M., Kent, W.J., Riemer, C., Elnitski, L., Smit, A.F.A., Roskin, K.M., Baertsch, R., Rosenbloom, K., Clawson, H., Green, E.D., Haussler, D., and Miller, W. 2004. Aligning Multiple Genomic Sequences With the Threaded Blockset Aligner. *Genome Res.* **14**(4): 708–715. doi:10.1101/gr.1933104.

Blanchette, M., and Tompa, M. 2002. Discovery of Regulatory Elements by a Computational Method for Phylogenetic Footprinting. *Genome Res.* **12**(5): 739–748. doi:10.1101/gr.6902.

Borda, V., Loesch, D.P., Guo, B., Laboulaye, R., Veliz-Otani, D., French-Kwawu, J.N., Leal, T.P., Gogarten, S.M., Ikpe, S., Gouveia, M.H., Mendes, M., Abecasis, G.R., Alvim, I., Arboleda-Bustos, C.E., Arboleda, G., Arboleda, H., Barreto, M.L., Barwick, L., Bezzer, M.A., Blangero, J., Borges, V., Caceres, O., Cai, J., Chana-Cuevas, P., Chen, Z., Custer, B., Dean, M., Dinardo, C., Domingos, I., Duggirala, R., Dieguez, E., Fernandez, W., Ferraz, H.B., Gilliland, F.D., Guio, H., Horta, B., Curran, J.E., Johnsen, J.M., Kaplan, R.C., Kelly, S., Kenny, E.E., Konkle, B.A., Kooperberg, C., Lescano, A., Lima-Costa, M.F., Loos, R.J.F., Manichaikul, A., Meyers, D.A., Naslavsky, M.S., Nickerson, D.A., North, K.E., Padilla, C., Preuss, M., Raggio, V., Reiner, A.P., Rich, S.S., Rieder, C.R., Rienstra, M., Rotter, J.I., Rundek, T., Sacco, R.L., Sanchez, C., Sankaran, V.G., Santos-Lobato, B.L., Schumacher-Schuh, A.F., Scliar, M.O., Silverman, E.K., Sofer, T., Lasky-Su, J., Tumas, V., Weiss, S.T., Disease (LARGE-PD), L.A.R.C. on the G. of P.,

- Consortium, N.S.G.N. (SiGN), Group, Topm.P.G.W., Mata, I.F., Hernandez, R.D., Tarazona-Santos, E., and O'Connor, T.D. 2023, March 29. Genetics of Latin American Diversity (GLAD) Project: insights into population genetics and association studies in recently admixed groups in the Americas. bioRxiv. doi:10.1101/2023.01.07.522490.
- Boyling, A., Perez-Siles, G., and Kennerson, M.L. 2022. Structural Variation at a Disease Mutation Hotspot: Strategies to Investigate Gene Regulation and the 3D Genome. *Front Genet* **13**: 842860. doi:10.3389/fgene.2022.842860.
- Brand, C.M., Kuang, S., Gilbertson, E.N., McArthur, E., Pollard, K.S., Webster, T.H., and Capra, J.A. 2023, October 26. Sequence-based machine learning reveals 3D genome differences between bonobos and chimpanzees. bioRxiv. doi:10.1101/2023.10.26.564272.
- Breiman, L. 2001. Random Forests. *Mach. Learn.* **45**(1): 5–32. doi:10.1023/A:1010933404324.
- Briand, N., and Collas, P. 2020. Lamina-associated domains: peripheral matters and internal affairs. *Genome Biol.* **21**(1): 85. doi:10.1186/s13059-020-02003-5.
- Buenrostro, J.D., Giresi, P.G., Zaba, L.C., Chang, H.Y., and Greenleaf, W.J. 2013. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat. Methods* **10**(12): 1213–1218. doi:10.1038/nmeth.2688.
- Burge, C., and Karlin, S. 1997. Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.* **268**(1): 78–94. doi:10.1006/jmbi.1997.0951.

- Bustamante, C.D., De La Vega, F.M., and Burchard, E.G. 2011. Genomics for the world. *Nature* **475**(7355): 163–165. Nature Publishing Group. doi:10.1038/475163a.
- Capra, J.A., Erwin, G.D., McKinsey, G., Rubenstein, J.L.R., and Pollard, K.S. 2013. Many human accelerated regions are developmental enhancers. *Philos. Trans. R. Soc. B Biol. Sci.* **368**(1632): 20130025. Royal Society. doi:10.1098/rstb.2013.0025.
- Carroll, S.B. 2008. Evo-devo and an expanding evolutionary synthesis: a genetic theory of morphological evolution. *Cell* **134**(1): 25–36. doi:10.1016/j.cell.2008.06.030.
- Catalinas, C.A., Ibarra-Soria, X., Flouri, C., Gordillo, J.E., Cousminer, D., Hutchinson, A., Krejci, A., Cortes, A., Acevedo, A., Malla, S., Fishwick, C., Drewes, G., and Rapiteanu, R. 2023, May 14. Mapping the functional impact of non-coding regulatory elements in primary T cells through single-cell CRISPR screens. bioRxiv. doi:10.1101/2023.05.14.540711.
- Cavalli-Sforza, L.L. 1997. Genes, peoples, and languages. *Proc. Natl. Acad. Sci.* **94**(15): 7719–7724. *Proceedings of the National Academy of Sciences*. doi:10.1073/pnas.94.15.7719.
- Cavalli-Sforza, L.L. 2005. The Human Genome Diversity Project: past, present and future. *Nat. Rev. Genet.* **6**(4): 333–340. Nature Publishing Group. doi:10.1038/nrg1596.
- Celaj, A., Gao, A.J., Lau, T.T.Y., Holgersen, E.M., Lo, A., Lodaya, V., Cole, C.B., Denroche, R.E., Spickett, C., Wagih, O., Pinheiro, P.O., Vora, P., Mohammadi-Shemirani, P., Chan, S., Nussbaum, Z., Zhang, X., Zhu, H., Ramamurthy, E., Kanuparthi, B., Iacocca, M., Ly, D., Kron, K., Verby, M., Cheung-Ong, K., Shalev,

Z., Vaz, B., Bhargava, S., Yusuf, F., Samuel, S., Alibai, S., Baghestani, Z., He, X., Krastel, K., Oladapo, O., Mohan, A., Shanavas, A., Bugno, M., Bogojeski, J., Schmitges, F., Kim, C., Grant, S., Jayaraman, R., Masud, T., Deshwar, A., Gandhi, S., and Frey, B.J. 2023, September 26. An RNA foundation model enables discovery of disease mechanisms and candidate therapeutics. *bioRxiv*. doi:10.1101/2023.09.20.558508.

Charlesworth, B. 2009. Effective population size and patterns of molecular evolution and variation. *Nat. Rev. Genet.* **10**(3): 195–205. Nature Publishing Group. doi:10.1038/nrg2526.

Chatterjee, S., and Ahituv, N. 2017. Gene Regulatory Elements, Major Drivers of Human Disease. *Annu. Rev. Genomics Hum. Genet.* **18**(Volume 18, 2017): 45–63. Annual Reviews. doi:10.1146/annurev-genom-091416-035537.

Chun, S., and Fay, J.C. 2011. Evidence for Hitchhiking of Deleterious Mutations within the Human Genome. *PLOS Genet.* **7**(8): e1002240. Public Library of Science. doi:10.1371/journal.pgen.1002240.

Collins, F.S., and Fink, L. 1995. The Human Genome Project. *Alcohol Health Res. World* **19**(3): 190–195.

Collins, R.L., Brand, H., Karczewski, K.J., Zhao, X., Alföldi, J., Francioli, L.C., Khera, A.V., Lowther, C., Gauthier, L.D., Wang, H., Watts, N.A., Solomonson, M., O'Donnell-Luria, A., Baumann, A., Munshi, R., Walker, M., Whelan, C.W., Huang, Y., Brookings, T., Sharpe, T., Stone, M.R., Valkanas, E., Fu, J., Tiao, G., Laricchia, K.M., Ruano-Rubio, V., Stevens, C., Gupta, N., Cusick, C., Margolin, L., Taylor, K.D., Lin, H.J., Rich, S.S., Post, W.S., Chen, Y.-D.I., Rotter, J.I.,

- Nusbaum, C., Philippakis, A., Lander, E., Gabriel, S., Neale, B.M., Kathiresan, S., Daly, M.J., Banks, E., MacArthur, D.G., and Talkowski, M.E. 2020. A structural variation reference for medical and population genetics. *Nature* **581**(7809): 444–451. Nature Publishing Group. doi:10.1038/s41586-020-2287-8.
- Corces, M.R., Trevino, A.E., Hamilton, E.G., Greenside, P.G., Sinnott-Armstrong, N.A., Vesuna, S., Satpathy, A.T., Rubin, A.J., Montine, K.S., Wu, B., Kathiria, A., Cho, S.W., Mumbach, M.R., Carter, A.C., Kasowski, M., Orloff, L.A., Risca, V.I., Kundaje, A., Khavari, P.A., Montine, T.J., Greenleaf, W.J., and Chang, H.Y. 2017. An improved ATAC-seq protocol reduces background and enables interrogation of frozen tissues. *Nat. Methods* **14**(10): 959–962. doi:10.1038/nmeth.4396.
- Dale, R.K., Pedersen, B.S., and Quinlan, A.R. 2011. Pybedtools: a flexible Python library for manipulating genomic datasets and annotations. *Bioinformatics* **27**(24): 3423–3424. doi:10.1093/bioinformatics/btr539.
- Davis, C.A., Hitz, B.C., Sloan, C.A., Chan, E.T., Davidson, J.M., Gabdank, I., Hilton, J.A., Jain, K., Baymuradov, U.K., Narayanan, A.K., Onate, K.C., Graham, K., Miyasato, S.R., Dreszer, T.R., Strattan, J.S., Jolanki, O., Tanaka, F.Y., and Cherry, J.M. 2018. The Encyclopedia of DNA elements (ENCODE): data portal update. *Nucleic Acids Res.* **46**(D1): D794–D801. doi:10.1093/nar/gkx1081.
- Dean, A., Larson, D.R., and Sartorelli, V. 2021. Enhancers, gene regulation, and genome organization. *Genes Dev.* **35**(7–8): 427–432. doi:10.1101/gad.348372.121.
- Degner, J.F., Pai, A.A., Pique-Regi, R., Veyrieras, J.-B., Gaffney, D.J., Pickrell, J.K., De Leon, S., Michelini, K., Lewellen, N., Crawford, G.E., Stephens, M., Gilad, Y., and

- Pritchard, J.K. 2012. DNase I sensitivity QTLs are a major determinant of human expression variation. *Nature* **482**(7385): 390–394. doi:10.1038/nature10808.
- Degroeve, S., De Baets, B., Van de Peer, Y., and Rouzé, P. 2002. Feature subset selection for splice site prediction. *Bioinformatics* **18**(suppl_2): S75–S83. doi:10.1093/bioinformatics/18.suppl_2.S75.
- Dekker, J., Alber, F., Aufmkolk, S., Beliveau, B.J., Bruneau, B.G., Belmont, A.S., Bintu, L., Boettiger, A., Calandrelli, R., Disteché, C.M., Gilbert, D.M., Gregor, T., Hansen, A.S., Huang, B., Huangfu, D., Kalhor, R., Leslie, C.S., Li, W., Li, Y., Ma, J., Noble, W.S., Park, P.J., Phillips-Cremins, J.E., Pollard, K.S., Rafelski, S.M., Ren, B., Ruan, Y., Shav-Tal, Y., Shen, Y., Shendure, J., Shu, X., Strambio-De-Castillia, C., Vertii, A., Zhang, H., and Zhong, S. 2023. Spatial and temporal organization of the genome: Current state and future aims of the 4D nucleome project. *Mol. Cell* **83**(15): 2624–2640. doi:10.1016/j.molcel.2023.06.018.
- 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., and 4D Nucleome Network. 2017. The 4D nucleome project. *Nature* **549**(7671): 219–226. doi:10.1038/nature23884.
- Dekker, J., Marti-Renom, M.A., and Mirny, L.A. 2013. Exploring the three-dimensional organization of genomes: interpreting chromatin interaction data. *Nat. Rev. Genet.* **14**(6): 390–403. doi:10.1038/nrg3454.
- Dekker, J., and Mirny, L. 2016. The 3D Genome as Moderator of Chromosomal Communication. *Cell* **164**(6): 1110–1121. doi:10.1016/j.cell.2016.02.007.

- Dekker, J., Rippe, K., Dekker, M., and Kleckner, N. 2002. Capturing Chromosome Conformation. *Science* **295**(5558): 1306–1311. American Association for the Advancement of Science. doi:10.1126/science.1067799.
- Deng, C., Whalen, S., Steyert, M., Ziffra, R., Przytycki, P.F., Inoue, F., Pereira, D.A., Capauto, D., Norton, S., Vaccarino, F.M., PsychENCODE Consortium, Pollen, A.A., Nowakowski, T.J., Ahituv, N., and Pollard, K.S. 2024. Massively parallel characterization of regulatory elements in the developing human cortex. *Science* **384**(6698): eadh0559. American Association for the Advancement of Science. doi:10.1126/science.adh0559.
- Deng, W., Lee, J., Wang, H., Miller, J., Reik, A., Gregory, P.D., Dean, A., and Blobel, G.A. 2012. Controlling long-range genomic interactions at a native locus by targeted tethering of a looping factor. *Cell* **149**(6): 1233–1244. doi:10.1016/j.cell.2012.03.051.
- Deplancke, B., Alpern, D., and Gardeux, V. 2016. The Genetics of Transcription Factor DNA Binding Variation. *Cell* **166**(3): 538–554. Elsevier. doi:10.1016/j.cell.2016.07.012.
- Diaz-Papkovich, A., Anderson-Trocmé, L., Ben-Eghan, C., and Gravel, S. 2019. UMAP reveals cryptic population structure and phenotype heterogeneity in large genomic cohorts. *PLOS Genet.* **15**(11): e1008432. Public Library of Science. doi:10.1371/journal.pgen.1008432.
- Diaz-Papkovich, A., Zabad, S., Ben-Eghan, C., Anderson-Trocmé, L., Femerling, G., Nathan, V., Patel, J., and Gravel, S. 2023, July 7. Topological stratification of

- continuous genetic variation in large biobanks. *bioRxiv*.
doi:10.1101/2023.07.06.548007.
- Diehl, A.G., and Boyle, A.P. 2018. Conserved and species-specific transcription factor co-binding patterns drive divergent gene regulation in human and mouse. *Nucleic Acids Res.* **46**(4): 1878–1894. doi:10.1093/nar/gky018.
- Ding, J., Bashashati, A., Roth, A., Oloumi, A., Tse, K., Zeng, T., Haffari, G., Hirst, M., Marra, M.A., Condon, A., Aparicio, S., and Shah, S.P. 2012. Feature-based classifiers for somatic mutation detection in tumour-normal paired sequencing data. *Bioinforma. Oxf. Engl.* **28**(2): 167–175. doi:10.1093/bioinformatics/btr629.
- Ding, P., Wang, Y., Zhang, X., Gao, X., Liu, G., and Yu, B. 2023. DeepSTF: predicting transcription factor binding sites by interpretable deep neural networks combining sequence and shape. *Brief. Bioinform.* **24**(4): bbad231. doi:10.1093/bib/bbad231.
- Dixon, J.R., Gorkin, D.U., and Ren, B. 2016. Chromatin Domains: The Unit of Chromosome Organization. *Mol. Cell* **62**(5): 668–680. Elsevier.
doi:10.1016/j.molcel.2016.05.018.
- Dixon, J.R., Selvaraj, S., Yue, F., Kim, A., Li, Y., Shen, Y., Hu, M., Liu, J.S., and Ren, B. 2012. Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* **485**(7398): 376–380. Nature Publishing Group.
doi:10.1038/nature11082.
- Dostie, J., Richmond, T.A., Arnaout, R.A., Selzer, R.R., Lee, W.L., Honan, T.A., Rubio, E.D., Krumm, A., Lamb, J., Nusbaum, C., Green, R.D., and Dekker, J. 2006. Chromosome Conformation Capture Carbon Copy (5C): A massively parallel

solution for mapping interactions between genomic elements. *Genome Res.* **16**(10): 1299–1309. doi:10.1101/gr.5571506.

Dowell, R.D. 2010. Transcription factor binding variation in the evolution of gene regulation. *Trends Genet.* **26**(11): 468–475. Elsevier.
doi:10.1016/j.tig.2010.08.005.

Duda, P. and Jan Zrzavý. 2016. Human population history revealed by a supertree approach. *Sci. Rep.* **6**(1): 29890. Nature Publishing Group.
doi:10.1038/srep29890.

ENCODE Project Consortium. 2012. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**(7414): 57–74. doi:10.1038/nature11247.

ENCODE Project Consortium, Moore, J.E., Purcaro, M.J., Pratt, H.E., Epstein, C.B., Shores, 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., Lécuycer, 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., and Weng, Z. 2020. Expanded encyclopaedias of DNA elements in the human and mouse genomes. *Nature* **583**(7818): 699–710. doi:10.1038/s41586-020-2493-4.

Fan, S., Spence, J.P., Feng, Y., Hansen, M.E.B., Terhorst, J., Beltrame, M.H., Ranciaro, A., Hirbo, J., Beggs, W., Thomas, N., Nyambo, T., Mpoloka, S.W., Mokone, G.G., Njamnshi, A.K., Fokunang, C., Meskel, D.W., Belay, G., Song, Y.S., and Tishkoff, S.A. 2023. Whole-genome sequencing reveals a complex African population demographic history and signatures of local adaptation. *Cell* **186**(5): 923-939.e14. doi:10.1016/j.cell.2023.01.042.

Felsenfeld, G., Boyes, J., Chung, J., Clark, D., and Studitsky, V. 1996. Chromatin structure and gene expression. *Proc. Natl. Acad. Sci.* **93**(18): 9384–9388. *Proceedings of the National Academy of Sciences.* doi:10.1073/pnas.93.18.9384.

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., ReproGen Consortium, Schizophrenia Working Group of the Psychiatric Genomics Consortium, RACI Consortium, Purcell, S., Stahl, E., Lindstrom, S., Perry, J.R.B., Okada, Y., Raychaudhuri, S., Daly, M.J., Patterson, N., Neale, B.M., and Price, A.L. 2015. Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat. Genet.* **47**(11): 1228–1235. doi:10.1038/ng.3404.

Frankish, A., Diekhans, M., Ferreira, A.-M., Johnson, R., Jungreis, I., Loveland, J., Mudge, J.M., Sisu, C., Wright, J., Armstrong, J., Barnes, I., Berry, A., Bignell, A., Carbonell Sala, S., Chrast, J., Cunningham, F., Di Domenico, T., Donaldson, S., Fiddes, I.T., García Girón, C., Gonzalez, J.M., Grego, T., Hardy, M., Hourlier, T., Hunt, T., Izuogu, O.G., Lagarde, J., Martin, F.J., Martínez, L., Mohanan, S., Muir, P., Navarro, F.C.P., Parker, A., Pei, B., Pozo, F., Ruffier, M., Schmitt, B.M., Stapleton, E., Suner, M.-M., Sycheva, I., Uszczyńska-Ratajczak, B., Xu, J.,

- Yates, A., Zerbino, D., Zhang, Y., Aken, B., Choudhary, J.S., Gerstein, M., Guigó, R., Hubbard, T.J.P., Kellis, M., Paten, B., Reymond, A., Tress, M.L., and Flicek, P. 2019. GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Res.* **47**(D1): D766–D773. doi:10.1093/nar/gky955.
- Fudenberg, G., Imakaev, M., Lu, C., Goloborodko, A., Abdennur, N., and Mirny, L.A. 2016. Formation of Chromosomal Domains by Loop Extrusion. *Cell Rep.* **15**(9): 2038–2049. doi:10.1016/j.celrep.2016.04.085.
- Fudenberg, G., Kelley, D.R., and Pollard, K.S. 2020. Predicting 3D genome folding from DNA sequence with Akita. *Nat Methods* **17**(11): 1111–1117. doi:10.1038/s41592-020-0958-x.
- Fudenberg, G., and Pollard, K.S. 2019. Chromatin features constrain structural variation across evolutionary timescales. *Proc Natl Acad Sci U S A* **116**(6): 2175–2180. doi:10.1073/pnas.1808631116.
- Furey, T.S., Cristianini, N., Duffy, N., Bednarski, D.W., Schummer, M., and Haussler, D. 2000. Support vector machine classification and validation of cancer tissue samples using microarray expression data. *Bioinformatics* **16**(10): 906–914. doi:10.1093/bioinformatics/16.10.906.
- Ganley, A.R.D., and Kobayashi, T. 2007. Phylogenetic footprinting to find functional DNA elements. *Methods Mol. Biol. Clifton NJ* **395**: 367–380. doi:10.1007/978-1-59745-514-5_23.
- Gao, L., Wu, K., Liu, Z., Yao, X., Yuan, S., Tao, W., Yi, L., Yu, G., Hou, Z., Fan, D., Tian, Y., Liu, J., Chen, Z.-J., and Liu, J. 2018. Chromatin Accessibility Landscape in

Human Early Embryos and Its Association with Evolution. *Cell* **173**(1): 248-259.e15. Elsevier. doi:10.1016/j.cell.2018.02.028.

Genereux, D.P., Serres, A., Armstrong, J., Johnson, J., Marinescu, V.D., Murén, E., Juan, D., Bejerano, G., Casewell, N.R., Chemnick, L.G., Damas, J., Di Palma, F., Diekhans, M., Fiddes, I.T., Garber, M., Gladyshev, V.N., Goodman, L., Haerty, W., Houck, M.L., Hubley, R., Kivioja, T., Koepfli, K.-P., Kuderna, L.F.K., Lander, E.S., Meadows, J.R.S., Murphy, W.J., Nash, W., Noh, H.J., Nweeia, M., Pfenning, A.R., Pollard, K.S., Ray, D.A., Shapiro, B., Smit, A.F.A., Springer, M.S., Steiner, C.C., Swofford, R., Taipale, J., Teeling, E.C., Turner-Maier, J., Alfoldi, J., Birren, B., Ryder, O.A., Lewin, H.A., Paten, B., Marques-Bonet, T., Lindblad-Toh, K., Karlsson, E.K., and Zoonomia Consortium. 2020. A comparative genomics multitool for scientific discovery and conservation. *Nature* **587**(7833): 240–245. Nature Publishing Group. doi:10.1038/s41586-020-2876-6.

Ghirlando, R., and Felsenfeld, G. 2016. CTCF: making the right connections. *Genes Dev.* **30**(8): 881–891. doi:10.1101/gad.277863.116.

Ghirlando, R., Giles, K., Gowher, H., Xiao, T., Xu, Z., Yao, H., and Felsenfeld, G. 2012. Chromatin domains, insulators, and the regulation of gene expression. *Biochim. Biophys. Acta BBA - Gene Regul. Mech.* **1819**(7): 644–651. doi:10.1016/j.bbagr.2012.01.016.

Ghosh, N., Santoni, D., Saha, I., and Felici, G. 2024. Predicting Transcription Factor Binding Sites with Deep Learning. *Int. J. Mol. Sci.* **25**(9): 4990. Multidisciplinary Digital Publishing Institute. doi:10.3390/ijms25094990.

- Giner-Delgado, C., Villatoro, S., Lerga-Jaso, J., Gayà-Vidal, M., Oliva, M., Castellano, D., Pantano, L., Bitarello, B.D., Izquierdo, D., Noguera, I., Olalde, I., Delprat, A., Blancher, A., Lalueza-Fox, C., Esko, T., O'Reilly, P.F., Andrés, A.M., Ferretti, L., Puig, M., and Cáceres, M. 2019. Evolutionary and functional impact of common polymorphic inversions in the human genome. *Nat. Commun.* **10**(1): 4222. Nature Publishing Group. doi:10.1038/s41467-019-12173-x.
- Girskis, K.M., Stergachis, A.B., DeGennaro, E.M., Doan, R.N., Qian, X., Johnson, M.B., Wang, P.P., Sejourne, G.M., Nagy, M.A., Pollina, E.A., Sousa, A.M.M., Shin, T., Kenny, C.J., Scotellaro, J.L., Debo, B.M., Gonzalez, D.M., Rento, L.M., Yeh, R.C., Song, J.H.T., Beaudin, M., Fan, J., Kharchenko, P.V., Sestan, N., Greenberg, M.E., and Walsh, C.A. 2021. Rewiring of human neurodevelopmental gene regulatory programs by human accelerated regions. *Neuron* **109**(20): 3239-3251.e7. doi:10.1016/j.neuron.2021.08.005.
- Glenwinkel, L., Wu, D., Minevich, G., and Hobert, O. 2014. TargetOrtho: A Phylogenetic Footprinting Tool to Identify Transcription Factor Targets. *Genetics* **197**(1): 61–76. doi:10.1534/genetics.113.160721.
- Goodwin, S., McPherson, J.D., and McCombie, W.R. 2016. Coming of age: ten years of next-generation sequencing technologies. *Nat. Rev. Genet.* **17**(6): 333–351. Nature Publishing Group. doi:10.1038/nrg.2016.49.
- Gosai, S., Castro, R., Fuentes, N., Butts, J., Kales, S., Noche, R., Mouri, K., Sabeti, P., Reilly, S., and Tewhey, R. 2023. Machine-guided design of synthetic cell type-specific cis-regulatory elements. *bioRxiv*: 2023.08.08.552077. doi:10.1101/2023.08.08.552077.

- Gravel, S., Henn, B.M., Gutenkunst, R.N., Indap, A.R., Marth, G.T., Clark, A.G., Yu, F., Gibbs, R.A., 1000 Genomes Project, and Bustamante, C.D. 2011. Demographic history and rare allele sharing among human populations. *Proc Natl Acad Sci U S A* **108**(29): 11983–11988. doi:10.1073/pnas.1019276108.
- Grubert, F., Srivas, R., Spacek, D.V., Kasowski, M., Ruiz-Velasco, M., Sinnott-Armstrong, N., Greenside, P., Narasimha, A., Liu, Q., Geller, B., Sanghi, A., Kulik, M., Sa, S., Rabinovitch, M., Kundaje, A., Dalton, S., Zaugg, J.B., and Snyder, M. 2020. Landscape of cohesin-mediated chromatin loops in the human genome. *Nature* **583**(7818): 737–743. Nature Publishing Group. doi:10.1038/s41586-020-2151-x.
- Gunsalus, L.M., Keiser, M.J., and Pollard, K.S. 2023a. In silico discovery of repetitive elements as key sequence determinants of 3D genome folding. *Cell Genomics* **3**(10). Elsevier. doi:10.1016/j.xgen.2023.100410.
- Gunsalus, L.M., McArthur, E., Gjoni, K., Kuang, S., Pittman, M., Capra, J.A., and Pollard, K.S. 2023b. Comparing chromatin contact maps at scale: methods and insights. *bioRxiv*: 2023.04.04.535480. doi:10.1101/2023.04.04.535480.
- Gurdasani, D., Carstensen, T., Tekola-Ayele, F., Pagani, L., Tachmazidou, I., Hatzikotoulas, K., Karthikeyan, S., Iles, L., Pollard, M.O., Choudhury, A., Ritchie, G.R.S., Xue, Y., Asimit, J., Nsubuga, R.N., Young, E.H., Pomilla, C., Kivinen, K., Rockett, K., Kamali, A., Doumatey, A.P., Asiki, G., Seeley, J., Sisay-Joof, F., Jallow, M., Tollman, S., Mekonnen, E., Ekong, R., Oljira, T., Bradman, N., Bojang, K., Ramsay, M., Adeyemo, A., Bekele, E., Motala, A., Norris, S.A., Pirie, F., Kaleebu, P., Kwiatkowski, D., Tyler-Smith, C., Rotimi, C., Zeggini, E., and

- Sandhu, M.S. 2015. The African Genome Variation Project shapes medical genetics in Africa. *Nature* **517**(7534): 327–332. Nature Publishing Group. doi:10.1038/nature13997.
- Gyawali, P.K., Le Guen, Y., Liu, X., Belloy, M.E., Tang, H., Zou, J., and He, Z. 2023. Improving genetic risk prediction across diverse population by disentangling ancestry representations. *Commun. Biol.* **6**: 964. doi:10.1038/s42003-023-05352-6.
- Hansen, T., Fong, S., Capra, J.A., and Hodges, E. 2023, February 15. Human gene regulatory evolution is driven by the divergence of regulatory element function in both cis and trans. *bioRxiv*. doi:10.1101/2023.02.14.528376.
- He, G., Chen, M., Bian, Y., and Yang, E. 2023. MTM: a multi-task learning framework to predict individualized tissue gene expression profiles. *Bioinformatics* **39**(6): btad363. doi:10.1093/bioinformatics/btad363.
- Hnisz, D., Weintraub, A.S., Day, D.S., Valton, A.-L., Bak, R.O., Li, C.H., Goldmann, J., Lajoie, B.R., Fan, Z.P., Sigova, A.A., Reddy, J., Borges-Rivera, D., Lee, T.I., Jaenisch, R., Porteus, M.H., Dekker, J., and Young, R.A. 2016. Activation of proto-oncogenes by disruption of chromosome neighborhoods. *Science* **351**(6280): 1454–1458. doi:10.1126/science.aad9024.
- Ho, J.W.K., Stefani, M., dos Remedios, C.G., and Charleston, M.A. 2008. Differential variability analysis of gene expression and its application to human diseases. *Bioinformatics* **24**(13): i390-8. doi:10.1093/bioinformatics/btn142.

- Hsieh, T.-H.S., Weiner, A., Lajoie, B., Dekker, J., Friedman, N., and Rando, O.J. 2015. Mapping Nucleosome Resolution Chromosome Folding in Yeast by Micro-C. *Cell* **162**(1): 108–119. doi:10.1016/j.cell.2015.05.048.
- Hu, B., Won, H., Mah, W., Park, R.B., Kassim, B., Spiess, K., Kozlenkov, A., Crowley, C.A., Pochareddy, S., PsychENCODE Consortium, Li, Y., Dracheva, S., Sestan, N., Akbarian, S., and Geschwind, D.H. 2021a. Neuronal and glial 3D chromatin architecture informs the cellular etiology of brain disorders. *Nat Commun* **12**(1): 3968. Springer Science and Business Media LLC. doi:10.1038/s41467-021-24243-0.
- Hu, M., Cebola, I., Carrat, G., Jiang, S., Nawaz, S., Khamis, A., Canouil, M., Froguel, P., Schulte, A., Solimena, M., Ibberson, M., Marchetti, P., Cardenas-Diaz, F.L., Gadue, P.J., Hastoy, B., Almeida-Souza, L., McMahon, H., and Rutter, G.A. 2021b. Chromatin 3D interaction analysis of the *STARD10* locus unveils *FCHSD2* as a regulator of insulin secretion. *Cell Rep.* **34**(5): 108703. doi:10.1016/j.celrep.2021.108703.
- Huang, J., Fang, H., and Fan, X. 2010. Decision forest for classification of gene expression data. *Comput. Biol. Med.* **40**(8): 698–704. doi:10.1016/j.combiomed.2010.06.004.
- Hudson, R.R. 2002. Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinforma. Oxf. Engl.* **18**(2): 337–338. doi:10.1093/bioinformatics/18.2.337.

- Ibrahim, D.M., and Mundlos, S. 2020. The role of 3D chromatin domains in gene regulation: a multi-faceted view on genome organization. *Curr. Opin. Genet. Dev.* **61**: 1–8. doi:10.1016/j.gde.2020.02.015.
- Jablonski, K.P., Carron, L., Mozziconacci, J., Forné, T., Hütt, M.-T., and Lesne, A. 2021, July 29. Contribution of 3D genome topological domains to genetic risk of cancers. doi:10.1101/2021.07.26.453813.
- Javierre, B.M., Burren, O.S., Wilder, S.P., Kreuzhuber, R., Hill, S.M., Sewitz, S., Cairns, J., Wingett, S.W., Várnai, C., Thiecke, M.J., Burden, F., Farrow, S., Cutler, A.J., Rehnström, K., Downes, K., Grassi, L., Kostadima, M., Freire-Pritchett, P., Wang, F., Martens, J.H., Kim, B., Sharifi, N., Janssen-Megens, E.M., Yaspo, M.-L., Linser, M., Kovacsovics, A., Clarke, L., Richardson, D., Datta, A., Flicek, P., Stunnenberg, H.G., Todd, J.A., Zerbino, D.R., Stegle, O., Ouwehand, W.H., Frontini, M., Wallace, C., Spivakov, M., and Fraser, P. 2016. Lineage-Specific Genome Architecture Links Enhancers and Non-coding Disease Variants to Target Gene Promoters. *Cell* **167**(5): 1369-1384.e19. doi:10.1016/j.cell.2016.09.037.
- Jerkovic, I., and Cavalli, G. 2021. Understanding 3D genome organization by multidisciplinary methods. *Nat. Rev. Mol. Cell Biol.* **22**(8): 511–528. Nature Publishing Group. doi:10.1038/s41580-021-00362-w.
- Ji, Y., Zhou, Z., Liu, H., and Davuluri, R.V. 2021. DNABERT: pre-trained Bidirectional Encoder Representations from Transformers model for DNA-language in genome. *Bioinformatics* **37**(15): 2112–2120. doi:10.1093/bioinformatics/btab083.

Kamat, K., Lao, Z., Qi, Y., Wang, Y., Ma, J., and Zhang, B. 2023. Compartmentalization with nuclear landmarks yields random, yet precise, genome organization.

Biophys. J. **122**(7): 1376–1389. doi:10.1016/j.bpj.2023.03.003.

Karczewski, K.J., Francioli, L.C., Tiao, G., Cummings, B.B., Alföldi, J., Wang, Q., Collins, R.L., Laricchia, K.M., Ganna, A., Birnbaum, D.P., Gauthier, L.D., Brand, H., Solomonson, M., Watts, N.A., Rhodes, D., Singer-Berk, M., England, E.M., Seaby, E.G., Kosmicki, J.A., Walters, R.K., Tashman, K., Farjoun, Y., Banks, E., Poterba, T., Wang, A., Seed, C., Whiffin, N., Chong, J.X., Samocha, K.E., Pierce-Hoffman, E., Zappala, Z., O'Donnell-Luria, A.H., Minikel, E.V., Weisburd, B., Lek, M., Ware, J.S., Vittal, C., Armean, I.M., Bergelson, L., Cibulskis, K., Connolly, K.M., Covarrubias, M., Donnelly, S., Ferriera, S., Gabriel, S., Gentry, J., Gupta, N., Jeandet, T., Kaplan, D., Llanwarne, C., Munshi, R., Novod, S., Petrillo, N., Roazen, D., Ruano-Rubio, V., Saltzman, A., Schleicher, M., Soto, J., Tibbetts, K., Tolonen, C., Wade, G., Talkowski, M.E., Neale, B.M., Daly, M.J., and MacArthur, D.G. 2020. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**(7809): 434–443. Nature Publishing Group. doi:10.1038/s41586-020-2308-7.

Kelleher, J., Thornton, K.R., Ashander, J., and Ralph, P.L. 2018. Efficient pedigree recording for fast population genetics simulation. *PLOS Comput. Biol.* **14**(11): e1006581. Public Library of Science. doi:10.1371/journal.pcbi.1006581.

Kelley, D.R. 2020. Cross-species regulatory sequence activity prediction. *PLOS Comput. Biol.* **16**(7): e1008050. Public Library of Science. doi:10.1371/journal.pcbi.1008050.

- Kelley, D.R., Reshef, Y.A., Bileschi, M., Belanger, D., McLean, C.Y., and Snoek, J. 2018. Sequential regulatory activity prediction across chromosomes with convolutional neural networks. *Genome Res* **28**(5): 739–750. doi:10.1101/gr.227819.117.
- Kelley, D.R., Snoek, J., and Rinn, J.L. 2016. Basset: learning the regulatory code of the accessible genome with deep convolutional neural networks. *Genome Res* **26**(7): 990–999. doi:10.1101/gr.200535.115.
- Keough, K.C., Shah, P.P., Wickramasinghe, N.M., Dundes, C.E., Chen, A., Salomon, R.E.A., Whalen, S., Loh, K.M., Dubois, N., Pollard, K.S., and Jain, R. 2020, July 24. An atlas of lamina-associated chromatin across thirteen human cell types reveals cell-type-specific and multiple subtypes of peripheral heterochromatin. doi:10.1101/2020.07.23.218768.
- Keough, K.C., Whalen, S., Inoue, F., Przytycki, P.F., Fair, T., Deng, C., Steyert, M., Ryu, H., Lindblad-Toh, K., Karlsson, E., Zoonomia Consortium, Nowakowski, T., Ahituv, N., Pollen, A., and Pollard, K.S. 2023. Three-dimensional genome rewiring in loci with human accelerated regions. *Science* **380**(6643): eabm1696. American Association for the Advancement of Science. doi:10.1126/science.abm1696.
- Key, F.M., Abdul-Aziz, M.A., Mundry, R., Peter, B.M., Sekar, A., D'Amato, M., Dennis, M.Y., Schmidt, J.M., and Andrés, A.M. 2018. Human local adaptation of the TRPM8 cold receptor along a latitudinal cline. *PLoS Genet.* **14**(5): e1007298. doi:10.1371/journal.pgen.1007298.
- Kim, H.-J., Yardımcı, G.G., Bonora, G., Ramani, V., Liu, J., Qiu, R., Lee, C., Hesson, J., Ware, C.B., Shendure, J., Duan, Z., and Noble, W.S. 2020. Capturing cell type-

specific chromatin compartment patterns by applying topic modeling to single-cell Hi-C data. *PLoS Comput Biol* **16**(9): e1008173.

doi:10.1371/journal.pcbi.1008173.

Kim-Hellmuth, S., Aguet, F., Oliva, M., Muñoz-Aguirre, M., Kasela, S., Wucher, V., Castel, S.E., Hamel, A.R., Viñuela, A., Roberts, A.L., Mangul, S., Wen, X., Wang, G., Barbeira, A.N., Garrido-Martín, D., Nadel, B.B., Zou, Y., Bonazzola, R., Quan, J., Brown, A., Martinez-Perez, A., Soria, J.M., GTEx Consortium, Getz, G., Dermitzakis, E.T., Small, K.S., Stephens, M., Xi, H.S., Im, H.K., Guigó, R., Segrè, A.V., Stranger, B.E., Ardlie, K.G., and Lappalainen, T. 2020. Cell type-specific genetic regulation of gene expression across human tissues. *Science* **369**(6509): eaaz8528. American Association for the Advancement of Science.

doi:10.1126/science.aaz8528.

Krefting, J., Andrade-Navarro, M.A., and Ibn-Salem, J. 2018. Evolutionary stability of topologically associating domains is associated with conserved gene regulation.

BMC Biol. **16**(1): 87. doi:10.1186/s12915-018-0556-x.

Krivega, I., Dale, R.K., and Dean, A. 2014. Role of LDB1 in the transition from chromatin looping to transcription activation. *Genes Dev.* **28**(12): 1278–1290.

doi:10.1101/gad.239749.114.

Krogh, A., Brown, M., Mian, I.S., Sjölander, K., and Haussler, D. 1994. Hidden Markov Models in Computational Biology: Applications to Protein Modeling. *J. Mol. Biol.*

235(5): 1501–1531. doi:10.1006/jmbi.1994.1104.

- Lambert, S.A., Jolma, A., Campitelli, L.F., Das, P.K., Yin, Y., Albu, M., Chen, X., Taipale, J., Hughes, T.R., and Weirauch, M.T. 2018. The Human Transcription Factors. *Cell* **172**(4): 650–665. Elsevier. doi:10.1016/j.cell.2018.01.029.
- Lettice, L.A., Horikoshi, T., Heaney, S.J.H., van Baren, M.J., van der Linde, H.C., Breedveld, G.J., Joosse, M., Akarsu, N., Oostra, B.A., Endo, N., Shibata, M., Suzuki, M., Takahashi, E., Shinka, T., Nakahori, Y., Ayusawa, D., Nakabayashi, K., Scherer, S.W., Heutink, P., Hill, R.E., and Noji, S. 2002. Disruption of a long-range cis-acting regulator for Shh causes preaxial polydactyly. *Proc. Natl. Acad. Sci. U. S. A.* **99**(11): 7548–7553. doi:10.1073/pnas.112212199.
- Levy, J.J., Titus, A.J., Petersen, C.L., Chen, Y., Salas, L.A., and Christensen, B.C. 2020. MethylNet: an automated and modular deep learning approach for DNA methylation analysis. *BMC Bioinformatics* **21**(1): 108. doi:10.1186/s12859-020-3443-8.
- Lewis, A.C.F., Molina, S.J., Appelbaum, P.S., Dauda, B., Di Rienzo, A., Fuentes, A., Fullerton, S.M., Garrison, N.A., Ghosh, N., Hammonds, E.M., Jones, D.S., Kenny, E.E., Kraft, P., Lee, S.S.-J., Mauro, M., Novembre, J., Panofsky, A., Sohail, M., Neale, B.M., and Allen, D.S. 2022. Getting genetic ancestry right for science and society. *Science* **376**(6590): 250–252. American Association for the Advancement of Science. doi:10.1126/science.abm7530.
- Lex, A., Gehlenborg, N., Strobel, H., Vuillemot, R., and Pfister, H. 2014. UpSet: Visualization of Intersecting Sets. *IEEE Trans. Vis. Comput. Graph.* **20**(12): 1983–1992. doi:10.1109/TVCG.2014.2346248.

- Li, C., Bonder, M.J., Syed, S., Consortium (HGSC), H.G.S.V., Group, H.F.A.W., Zody, M.C., Chaisson, M.J.P., Talkowski, M.E., Marschall, T., Korb, J.O., Eichler, E.E., Lee, C., and Shi, X. 2023, May 15. A comprehensive catalog of 3D genome organization in diverse human genomes facilitates understanding of the impact of structural variation on chromatin structure. *bioRxiv*. doi:10.1101/2023.05.15.540856.
- Li, D., He, M., Tang, Q., Tian, S., Zhang, J., Li, Y., Wang, D., Jin, L., Ning, C., Zhu, W., Hu, S., Long, K., Ma, J., Liu, J., Zhang, Z., and Li, M. 2022a. Comparative 3D genome architecture in vertebrates. *BMC Biol.* **20**(1): 99. doi:10.1186/s12915-022-01301-7.
- Li, J., Xiang, Y., Zhang, L., Qi, X., Zheng, Z., Zhou, P., Tang, Z., Jin, Y., Zhao, Q., Fu, Y., Zhao, Y., Li, X., Fu, L., and Zhao, S. 2022b. Enhancer-promoter interaction maps provide insights into skeletal muscle-related traits in pig genome. *BMC Biol.* **20**(1): 136. doi:10.1186/s12915-022-01322-2.
- Li, J.Z., Absher, D.M., Tang, H., Southwick, A.M., Casto, A.M., Ramachandran, S., Cann, H.M., Barsh, G.S., Feldman, M., Cavalli-Sforza, L.L., and Myers, R.M. 2008. Worldwide Human Relationships Inferred from Genome-Wide Patterns of Variation. *Science* **319**(5866): 1100–1104. American Association for the Advancement of Science. doi:10.1126/science.1153717.
- Li, M.J., Wang, L.Y., Xia, Z., Sham, P.C., and Wang, J. 2013. GWAS3D: detecting human regulatory variants by integrative analysis of genome-wide associations, chromosome interactions and histone modifications. *Nucleic Acids Res.* **41**(W1): W150–W158. doi:10.1093/nar/gkt456.

- Libbrecht, M.W., and Noble, W.S. 2015. Machine learning applications in genetics and genomics. *Nat. Rev. Genet.* **16**(6): 321–332. Nature Publishing Group. doi:10.1038/nrg3920.
- Lieberman-Aiden, E., van Berkum, N.L., Williams, L., Imakaev, M., Ragooczy, T., Telling, A., Amit, I., Lajoie, B.R., Sabo, P.J., Dorschner, M.O., Sandstrom, R., Bernstein, B., Bender, M.A., Groudine, M., Gnirke, A., Stamatoyannopoulos, J., Mirny, L.A., Lander, E.S., and Dekker, J. 2009. Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome. *Science* **326**(5950): 289–293. American Association for the Advancement of Science. doi:10.1126/science.1181369.
- Liu, B., Zhang, H., Zhou, C., Li, G., Fennell, A., Wang, G., Kang, Y., Liu, Q., and Ma, Q. 2016. An integrative and applicable phylogenetic footprinting framework for cis-regulatory motifs identification in prokaryotic genomes. *BMC Genomics* **17**(1): 578. doi:10.1186/s12864-016-2982-x.
- Liu, S., Zheng, P., Wang, C.Y., Jia, B.B., Zemke, N.R., Ren, B., and Zhuang, X. 2023. Cell-type-specific 3D-genome organization and transcription regulation in the brain. *bioRxiv*: 2023.12.04.570024. doi:10.1101/2023.12.04.570024.
- Long, H.K., Prescott, S.L., and Wysocka, J. 2016. Ever-Changing Landscapes: Transcriptional Enhancers in Development and Evolution. *Cell* **167**(5): 1170–1187. Elsevier. doi:10.1016/j.cell.2016.09.018.
- Loots, G.G., Locksley, R.M., Blankespoor, C.M., Wang, Z.E., Miller, W., Rubin, E.M., and Frazer, K.A. 2000. Identification of a coordinate regulator of interleukins 4,

13, and 5 by cross-species sequence comparisons. *Science* **288**(5463): 136–140. doi:10.1126/science.288.5463.136.

Ludwig, M.Z., Patel, N.H., and Kreitman, M. 1998. Functional analysis of eve stripe 2 enhancer evolution in *Drosophila*: rules governing conservation and change. *Dev. Camb. Engl.* **125**(5): 949–958. doi:10.1242/dev.125.5.949.

Lundberg, S.M., and Lee, S.-I. 2017. A Unified Approach to Interpreting Model Predictions. *In* *Advances in Neural Information Processing Systems*. Curran Associates, Inc. Available from https://papers.nips.cc/paper_files/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html [accessed 29 July 2024].

Lupiáñez, D.G., Kraft, K., Heinrich, V., Krawitz, P., Brancati, F., Klopocki, E., Horn, D., Kayserili, H., Opitz, J.M., Laxova, R., Santos-Simarro, F., Gilbert-Dussardier, B., Wittler, L., Borschiwer, M., Haas, S.A., Osterwalder, M., Franke, M., Timmermann, B., Hecht, J., Spielmann, M., Visel, A., and Mundlos, S. 2015. Disruptions of Topological Chromatin Domains Cause Pathogenic Rewiring of Gene-Enhancer Interactions. *Cell* **161**(5): 1012–1025. doi:10.1016/j.cell.2015.04.004.

Mallick, S., Li, H., Lipson, M., Mathieson, I., Gymrek, M., Racimo, F., Zhao, M., Chennagiri, N., Nordenfelt, S., Tandon, A., Skoglund, P., Lazaridis, I., Sankararaman, S., Fu, Q., Rohland, N., Renaud, G., Erlich, Y., Willems, T., Gallo, C., Spence, J.P., Song, Y.S., Poletti, G., Balloux, F., van Driem, G., de Knijff, P., Romero, I.G., Jha, A.R., Behar, D.M., Bravi, C.M., Capelli, C., Hervig, T., Moreno-Estrada, A., Posukh, O.L., Balanovska, E., Balanovsky, O., Karachanak-Yankova,

S., Sahakyan, H., Toncheva, D., Yepiskoposyan, L., Tyler-Smith, C., Xue, Y.,
Abdullah, M.S., Ruiz-Linares, A., Beall, C.M., Di Rienzo, A., Jeong, C.,
Starikovskaya, E.B., Metspalu, E., Parik, J., Villems, R., Henn, B.M., Hodoglugil,
U., Mahley, R., Sajantila, A., Stamatoyannopoulos, G., Wee, J.T.S., Khusainova,
R., Khusnutdinova, E., Litvinov, S., Ayodo, G., Comas, D., Hammer, M.F., Kivisild,
T., Klitz, W., Winkler, C.A., Labuda, D., Bamshad, M., Jorde, L.B., Tishkoff, S.A.,
Watkins, W.S., Metspalu, M., Dryomov, S., Sukernik, R., Singh, L., Thangaraj, K.,
Pääbo, S., Kelso, J., Patterson, N., and Reich, D. 2016. The Simons Genome
Diversity Project: 300 genomes from 142 diverse populations. *Nature* **538**(7624):
201–206. doi:10.1038/nature18964.

Margulies, E.H., Cooper, G.M., Asimenos, G., Thomas, D.J., Dewey, C.N., Siepel, A.,
Birney, E., Keefe, D., Schwartz, A.S., Hou, M., Taylor, J., Nikolaev, S., Montoya-
Burgos, J.I., Löytynoja, A., Whelan, S., Pardi, F., Massingham, T., Brown, J.B.,
Bickel, P., Holmes, I., Mullikin, J.C., Ureta-Vidal, A., Paten, B., Stone, E.A.,
Rosenbloom, K.R., Kent, W.J., Bouffard, G.G., Guan, X., Hansen, N.F., Idol, J.R.,
Maduro, V.V.B., Maskeri, B., McDowell, J.C., Park, M., Thomas, P.J., Young,
A.C., Blakesley, R.W., Muzny, D.M., Sodergren, E., Wheeler, D.A., Worley, K.C.,
Jiang, H., Weinstock, G.M., Gibbs, R.A., Graves, T., Fulton, R., Mardis, E.R.,
Wilson, R.K., Clamp, M., Cuff, J., Gnerre, S., Jaffe, D.B., Chang, J.L., Lindblad-
Toh, K., Lander, E.S., Hinrichs, A., Trumbower, H., Clawson, H., Zweig, A., Kuhn,
R.M., Barber, G., Harte, R., Karolchik, D., Field, M.A., Moore, R.A., Matthewson,
C.A., Schein, J.E., Marra, M.A., Antonarakis, S.E., Batzoglou, S., Goldman, N.,
Hardison, R., Haussler, D., Miller, W., Pachter, L., Green, E.D., and Sidow, A.

2007. Analyses of deep mammalian sequence alignments and constraint predictions for 1% of the human genome. *Genome Res.* **17**(6): 760–774. doi:10.1101/gr.6034307.
- Martin, A.R., Gignoux, C.R., Walters, R.K., Wojcik, G.L., Neale, B.M., Gravel, S., Daly, M.J., Bustamante, C.D., and Kenny, E.E. 2017. Human Demographic History Impacts Genetic Risk Prediction across Diverse Populations. *Am. J. Hum. Genet.* **100**(4): 635–649. Elsevier. doi:10.1016/j.ajhg.2017.03.004.
- Martin, A.R., Kanai, M., Kamatani, Y., Okada, Y., Neale, B.M., and Daly, M.J. 2019. Clinical use of current polygenic risk scores may exacerbate health disparities. *Nat Genet* **51**(4): 584–591. doi:10.1038/s41588-019-0379-x.
- Maurano, M.T., Humbert, R., Rynes, E., Thurman, R.E., Haugen, E., Wang, H., Reynolds, A.P., Sandstrom, R., Qu, H., Brody, J., Shafer, A., Neri, F., Lee, K., Kutuyavin, T., Stehling-Sun, S., Johnson, A.K., Canfield, T.K., Giste, E., Diegel, M., Bates, D., Hansen, R.S., Neph, S., Sabo, P.J., Heimfeld, S., Raubitschek, A., Ziegler, S., Cotsapas, C., Sotoodehnia, N., Glass, I., Sunyaev, S.R., Kaul, R., and Stamatoyannopoulos, J.A. 2012. Systematic localization of common disease-associated variation in regulatory DNA. *Science* **337**(6099): 1190–1195. doi:10.1126/science.1222794.
- Maurya, S.S. 2021. Role of Enhancers in Development and Diseases. *Epigenomes* **5**(4): 21. Multidisciplinary Digital Publishing Institute. doi:10.3390/epigenomes5040021.
- McArthur, E., and Capra, J.A. 2021. Topologically associating domain boundaries that are stable across diverse cell types are evolutionarily constrained and enriched

- for heritability. *Am. J. Hum. Genet.* **108**(2): 269–283.
doi:10.1016/j.ajhg.2021.01.001.
- McArthur, E., Rinker, D.C., Gilbertson, E.N., Fudenberg, G., Pittman, M., Keough, K., Pollard, K.S., and Capra, J.A. 2022, February 8. Reconstructing the 3D genome organization of Neanderthals reveals that chromatin folding shaped phenotypic and sequence divergence. doi:10.1101/2022.02.07.479462.
- McCarthy, M.I., Abecasis, G.R., Cardon, L.R., Goldstein, D.B., Little, J., Ioannidis, J.P.A., and Hirschhorn, J.N. 2008. Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nat. Rev. Genet.* **9**(5): 356–369. Nature Publishing Group. doi:10.1038/nrg2344.
- McLean, C.Y., Reno, P.L., Pollen, A.A., Bassan, A.I., Capellini, T.D., Guenther, C., Indjeian, V.B., Lim, X., Menke, D.B., Schaar, B.T., Wenger, A.M., Bejerano, G., and Kingsley, D.M. 2011. Human-specific loss of regulatory DNA and the evolution of human-specific traits. *Nature* **471**(7337): 216–219. Nature Publishing Group. doi:10.1038/nature09774.
- Millán-Zambrano, G., Burton, A., Bannister, A.J., and Schneider, R. 2022. Histone post-translational modifications — cause and consequence of genome function. *Nat. Rev. Genet.* **23**(9): 563–580. Nature Publishing Group. doi:10.1038/s41576-022-00468-7.
- Misteli, T. 2020. *The Self-Organizing Genome: Principles of Genome Architecture and Function.* *Cell* **183**(1): 28–45. Elsevier. doi:10.1016/j.cell.2020.09.014.
- Movva, R., Greenside, P., Marinov, G.K., Nair, S., Shrikumar, A., and Kundaje, A. 2019. Deciphering regulatory DNA sequences and noncoding genetic variants using

- neural network models of massively parallel reporter assays. *PLOS ONE* **14**(6): e0218073. Public Library of Science. doi:10.1371/journal.pone.0218073.
- Mumbach, M.R., Rubin, A.J., Flynn, R.A., Dai, C., Khavari, P.A., Greenleaf, W.J., and Chang, H.Y. 2016. HiChIP: efficient and sensitive analysis of protein-directed genome architecture. *Nat. Methods* **13**(11): 919–922. Nature Publishing Group. doi:10.1038/nmeth.3999.
- Nassar, L.R., Barber, G.P., Benet-Pagès, A., Casper, J., Clawson, H., Diekhans, M., Fischer, C., Gonzalez, J.N., Hinrichs, A.S., Lee, B.T., Lee, C.M., Muthuraman, P., Nguy, B., Pereira, T., Nejad, P., Perez, G., Raney, B.J., Schmelter, D., Speir, M.L., Wick, B.D., Zweig, A.S., Haussler, D., Kuhn, R.M., Haeussler, M., and Kent, W.J. 2022. The UCSC Genome Browser database: 2023 update. *Nucleic Acids Res.* **51**(D1): D1188–D1195. doi:10.1093/nar/gkac1072.
- Nielsen, R., Akey, J.M., Jakobsson, M., Pritchard, J.K., Tishkoff, S., and Willerslev, E. 2017. Tracing the peopling of the world through genomics. *Nature* **541**(7637): 302–310. Nature Publishing Group. doi:10.1038/nature21347.
- Nodelman, I.M., and Bowman, G.D. 2021. Biophysics of Chromatin Remodeling. *Annu. Rev. Biophys.* **50**(Volume 50, 2021): 73–93. Annual Reviews. doi:10.1146/annurev-biophys-082520-080201.
- Nora, E.P., Goloborodko, A., Valton, A.-L., Gibcus, J.H., Uebersohn, A., Abdennur, N., Dekker, J., Mirny, L.A., and Bruneau, B.G. 2017. Targeted Degradation of CTCF Decouples Local Insulation of Chromosome Domains from Genomic Compartmentalization. *Cell* **169**(5): 930-944.e22. doi:10.1016/j.cell.2017.05.004.

- Nora, E.P., Lajoie, B.R., Schulz, E.G., Giorgetti, L., Okamoto, I., Servant, N., Piolot, T., van Berkum, N.L., Meisig, J., Sedat, J., Gribnau, J., Barillot, E., Blüthgen, N., Dekker, J., and Heard, E. 2012. Spatial partitioning of the regulatory landscape of the X-inactivation centre. *Nature* **485**(7398): 381–385. doi:10.1038/nature11049.
- Norton, H.K., and Phillips-Cremins, J.E. 2017. Crossed wires: 3D genome misfolding in human disease. *J Cell Biol* **216**(11): 3441–3452. doi:10.1083/jcb.201611001.
- Nothman, J. 2023, October 30. UpSetPlot documentation. Python. Available from <https://github.com/jnothman/UpSetPlot> [accessed 31 October 2023].
- Novakovsky, G., Fornes, O., Saraswat, M., Mostafavi, S., and Wasserman, W.W. 2023. ExplainNN: interpretable and transparent neural networks for genomics. *Genome Biol.* **24**(1): 154. doi:10.1186/s13059-023-02985-y.
- Novembre, J., Johnson, T., Bryc, K., Kutalik, Z., Boyko, A.R., Auton, A., Indap, A., King, K.S., Bergmann, S., Nelson, M.R., Stephens, M., and Bustamante, C.D. 2008. Genes mirror geography within Europe. *Nature* **456**(7218): 98–101. doi:10.1038/nature07331.
- Oyelade, J., Isewon, I., Oladipupo, F., Aromolaran, O., Uwoghiren, E., Ameh, F., Achas, M., and Adebisi, E. 2016. Clustering Algorithms: Their Application to Gene Expression Data. *Bioinforma. Biol. Insights* **10**: 237–253. doi:10.4137/BBI.S38316.
- Pal, A., Noble, M.A., Morales, M., Pal, R., Baumgartner, M., Yang, J.W., Yim, K.M., Uebbing, S., and Noonan, J.P. 2024, June 26. Resolving the three-dimensional interactome of Human Accelerated Regions during human and chimpanzee neurodevelopment. *bioRxiv*. doi:10.1101/2024.06.25.600691.

- Pang, B., van Weerd, J.H., Hamoen, F.L., and Snyder, M.P. 2023. Identification of non-coding silencer elements and their regulation of gene expression. *Nat. Rev. Mol. Cell Biol.* **24**(6): 383–395. Nature Publishing Group. doi:10.1038/s41580-022-00549-9.
- Patterson, N., Price, A.L., and Reich, D. 2006. Population structure and eigenanalysis. *PLoS Genet.* **2**(12): e190. doi:10.1371/journal.pgen.0020190.
- Peng, P.-C., Khoueiry, P., Girardot, C., Reddington, J.P., Garfield, D.A., Furlong, E.E.M., and Sinha, S. 2019. The Role of Chromatin Accessibility in cis-Regulatory Evolution. *Genome Biol. Evol.* **11**(7): 1813–1828. doi:10.1093/gbe/evz103.
- Pennacchio, L.A., Ahituv, N., Moses, A.M., Prabhakar, S., Nobrega, M.A., Shoukry, M., Minovitsky, S., Dubchak, I., Holt, A., Lewis, K.D., Plajzer-Frick, I., Akiyama, J., De Val, S., Afzal, V., Black, B.L., Couronne, O., Eisen, M.B., Visel, A., and Rubin, E.M. 2006. In vivo enhancer analysis of human conserved non-coding sequences. *Nature* **444**(7118): 499–502. doi:10.1038/nature05295.
- Phillips-Cremins, J.E., and Corces, V.G. 2013. Chromatin Insulators: Linking Genome Organization to Cellular Function. *Mol. Cell* **50**(4): 461–474. doi:10.1016/j.molcel.2013.04.018.
- Pollard, D.A., Moses, A.M., Iyer, V.N., and Eisen, M.B. 2006a. Detecting the limits of regulatory element conservation and divergence estimation using pairwise and multiple alignments. *BMC Bioinformatics* **7**(1): 376. doi:10.1186/1471-2105-7-376.

- Pollard, K.S., Hubisz, M.J., Rosenbloom, K.R., and Siepel, A. 2010. Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Res.* **20**(1): 110–121. doi:10.1101/gr.097857.109.
- Pollard, K.S., Salama, S.R., King, B., Kern, A.D., Dreszer, T., Katzman, S., Siepel, A., Pedersen, J.S., Bejerano, G., Baertsch, R., Rosenbloom, K.R., Kent, J., and Haussler, D. 2006b. Forces Shaping the Fastest Evolving Regions in the Human Genome. *PLoS Genet.* **2**(10): e168. doi:10.1371/journal.pgen.0020168.
- Pollard, K.S., Salama, S.R., Lambert, N., Lambot, M.-A., Coppens, S., Pedersen, J.S., Katzman, S., King, B., Onodera, C., Siepel, A., Kern, A.D., Dehay, C., Igel, H., Ares, M., Vanderhaeghen, P., and Haussler, D. 2006c. An RNA gene expressed during cortical development evolved rapidly in humans. *Nature* **443**(7108): 167–172. doi:10.1038/nature05113.
- Popejoy, A.B., and Fullerton, S.M. 2016. Genomics is failing on diversity. *Nature* **538**(7624): 161–164. Nature Publishing Group. doi:10.1038/538161a.
- Prabhakar, S., Visel, A., Akiyama, J.A., Shoukry, M., Lewis, K.D., Holt, A., Plajzer-Frick, I., Morrison, H., FitzPatrick, D.R., Afzal, V., Pennacchio, L.A., Rubin, E.M., and Noonan, J.P. 2008. Human-Specific Gain of Function in a Developmental Enhancer. *Science* **321**(5894): 1346–1350. American Association for the Advancement of Science. doi:10.1126/science.1159974.
- Prado-Martinez, J., Sudmant, P.H., Kidd, J.M., Li, H., Kelley, J.L., Lorente-Galdos, B., Veeramah, K.R., Woerner, A.E., O'Connor, T.D., Santpere, G., Cagan, A., Theunert, C., Casals, F., Laayouni, H., Munch, K., Hobolth, A., Halager, A.E., Malig, M., Hernandez-Rodriguez, J., Hernando-Herraez, I., Prüfer, K., Pybus, M.,

- Johnstone, L., Lachmann, M., Alkan, C., Twigg, D., Petit, N., Baker, C., Hormozdiari, F., Fernandez-Callejo, M., Dabad, M., Wilson, M.L., Stevison, L., Camprubí, C., Carvalho, T., Ruiz-Herrera, A., Vives, L., Mele, M., Abello, T., Kondova, I., Bontrop, R.E., Pusey, A., Lankester, F., Kiyang, J.A., Bergl, R.A., Lonsdorf, E., Myers, S., Ventura, M., Gagneux, P., Comas, D., Siegismund, H., Blanc, J., Agueda-Calpena, L., Gut, M., Fulton, L., Tishkoff, S.A., Mullikin, J.C., Wilson, R.K., Gut, I.G., Gonder, M.K., Ryder, O.A., Hahn, B.H., Navarro, A., Akey, J.M., Bertranpetit, J., Reich, D., Mailund, T., Schierup, M.H., Hvilsom, C., Andrés, A.M., Wall, J.D., Bustamante, C.D., Hammer, M.F., Eichler, E.E., and Marques-Bonet, T. 2013. Great ape genetic diversity and population history. *Nature* **499**(7459): 471–475. Nature Publishing Group. doi:10.1038/nature12228.
- Pratto, F., Brick, K., Khil, P., Smagulova, F., Petukhova, G.V., and Camerini-Otero, R.D. 2014. Recombination initiation maps of individual human genomes. *Science* **346**(6211): 1256442. American Association for the Advancement of Science. doi:10.1126/science.1256442.
- Prescott, S.L., Srinivasan, R., Marchetto, M.C., Grishina, I., Narvaiza, I., Selleri, L., Gage, F.H., Swigut, T., and Wysocka, J. 2015. Enhancer Divergence and cis-Regulatory Evolution in the Human and Chimp Neural Crest. *Cell* **163**(1): 68–83. Elsevier. doi:10.1016/j.cell.2015.08.036.
- Qiu, Y., Wang, J., Lei, J., and Roeder, K. 2021. Identification of cell-type-specific marker genes from co-expression patterns in tissue samples. *Bioinformatics* **37**(19): 3228–3234. doi:10.1093/bioinformatics/btab257.

- Quang, D., Chen, Y., and Xie, X. 2015. DANN: a deep learning approach for annotating the pathogenicity of genetic variants. *Bioinforma. Oxf. Engl.* **31**(5): 761–763. doi:10.1093/bioinformatics/btu703.
- Quinlan, A.R., and Hall, I.M. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**(6): 841–842. doi:10.1093/bioinformatics/btq033.
- Raab, J.R., and Kamakaka, R.T. 2010. Insulators and promoters: closer than we think. *Nat. Rev. Genet.* **11**(6): 439–446. Nature Publishing Group. doi:10.1038/nrg2765.
- Rabiee, M., Sayyari, E., and Mirarab, S. 2019. Multi-allele species reconstruction using ASTRAL. *Mol. Phylogenet. Evol.* **130**: 286–296. doi:10.1016/j.ympev.2018.10.033.
- Ramachandran, S., Deshpande, O., Roseman, C.C., Rosenberg, N.A., Feldman, M.W., and Cavalli-Sforza, L.L. 2005. Support from the relationship of genetic and geographic distance in human populations for a serial founder effect originating in Africa. *Proc. Natl. Acad. Sci.* **102**(44): 15942–15947. Proceedings of the National Academy of Sciences. doi:10.1073/pnas.0507611102.
- Reilly, S.K., and Noonan, J.P. 2016. Evolution of Gene Regulation in Humans. *Annu. Rev. Genomics Hum. Genet.* **17**(Volume 17, 2016): 45–67. Annual Reviews. doi:10.1146/annurev-genom-090314-045935.
- Ritchie, M.D., Holzinger, E.R., Li, R., Pendergrass, S.A., and Kim, D. 2015. Methods of integrating data to uncover genotype–phenotype interactions. *Nat. Rev. Genet.* **16**(2): 85–97. Nature Publishing Group. doi:10.1038/nrg3868.

Roadmap Epigenomics Consortium, Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., Heravi-Moussavi, A., Kheradpour, P., Zhang, Z., Wang, J., Ziller, M.J., Amin, V., Whitaker, J.W., Schultz, M.D., Ward, L.D., Sarkar, A., Quon, G., Sandstrom, R.S., Eaton, M.L., Wu, Y.-C., Pfening, A.R., Wang, X., Claussnitzer, M., Liu, Y., Coarfa, C., Harris, R.A., Shores, N., Epstein, C.B., Gjoneska, E., Leung, D., Xie, W., Hawkins, R.D., Lister, R., Hong, C., Gascard, P., Mungall, A.J., Moore, R., Chuah, E., Tam, A., Canfield, T.K., Hansen, R.S., Kaul, R., Sabo, P.J., Bansal, M.S., Carles, A., Dixon, J.R., Farh, K.-H., Feizi, S., Karlic, R., Kim, A.-R., Kulkarni, A., Li, D., Lowdon, R., Elliott, G., Mercer, T.R., Neph, S.J., Onuchic, V., Polak, P., Rajagopal, N., Ray, P., Sallari, R.C., Siebenthall, K.T., Sinnott-Armstrong, N.A., Stevens, M., Thurman, R.E., Wu, J., Zhang, B., Zhou, X., Beaudet, A.E., Boyer, L.A., De Jager, P.L., Farnham, P.J., Fisher, S.J., Haussler, D., Jones, S.J.M., Li, W., Marra, M.A., McManus, M.T., Sunyaev, S., Thomson, J.A., Tlsty, T.D., Tsai, L.-H., Wang, W., Waterland, R.A., Zhang, M.Q., Chadwick, L.H., Bernstein, B.E., Costello, J.F., Ecker, J.R., Hirst, M., Meissner, A., Milosavljevic, A., Ren, B., Stamatoyannopoulos, J.A., Wang, T., and Kellis, M. 2015. Integrative analysis of 111 reference human epigenomes. *Nature* **518**(7539): 317–330. doi:10.1038/nature14248.

Robson, M.I., Ringel, A.R., and Mundlos, S. 2019. Regulatory Landscaping: How Enhancer-Promoter Communication Is Sculpted in 3D. *Mol. Cell* **74**(6): 1110–1122. Elsevier. doi:10.1016/j.molcel.2019.05.032.

- Rodrigues, M.F., Kern, A.D., and Ralph, P.L. 2024. Shared evolutionary processes shape landscapes of genomic variation in the great apes. *Genetics* **226**(4): iyae006. doi:10.1093/genetics/iyae006.
- Roix, J.J., McQueen, P.G., Munson, P.J., Parada, L.A., and Misteli, T. 2003. Spatial proximity of translocation-prone gene loci in human lymphomas. *Nat Genet* **34**(3): 287–291. doi:10.1038/ng1177.
- Sanborn, A.L., Rao, S.S.P., Huang, S.-C., Durand, N.C., Huntley, M.H., Jewett, A.I., Bochkov, I.D., Chinnappan, D., Cutkosky, A., Li, J., Geeting, K.P., Gnirke, A., Melnikov, A., McKenna, D., Stamenova, E.K., Lander, E.S., and Aiden, E.L. 2015. Chromatin extrusion explains key features of loop and domain formation in wild-type and engineered genomes. *Proc. Natl. Acad. Sci.* **112**(47): E6456–E6465. Proceedings of the National Academy of Sciences. doi:10.1073/pnas.1518552112.
- Sánchez-Gaya, V., Mariner-Faulí, M., and Rada-Iglesias, A. 2020. Rare or overlooked? Structural disruption of regulatory domains in human neurocristopathies. *Front Genet* **11**: 688. Frontiers Media SA. doi:10.3389/fgene.2020.00688.
- Sanyal, A., Lajoie, B.R., Jain, G., and Dekker, J. 2012. The long-range interaction landscape of gene promoters. *Nature* **489**(7414): 109–113. Nature Publishing Group. doi:10.1038/nature11279.
- Sasse, A., Ng, B., Spiro, A.E., Tasaki, S., Bennett, D.A., Gaiteri, C., De Jager, P.L., Chikina, M., and Mostafavi, S. 2023. Benchmarking of deep neural networks for predicting personal gene expression from DNA sequence highlights

- shortcomings. *Nat. Genet.* **55**(12): 2060–2064. Nature Publishing Group.
doi:10.1038/s41588-023-01524-6.
- Sauerwald, N., Singhal, A., and Kingsford, C. 2020. Analysis of the structural variability of topologically associated domains as revealed by Hi-C. *NAR Genomics Bioinforma.* **2**(1): lqz008. doi:10.1093/nargab/lqz008.
- Schipper, M., and Posthuma, D. 2022. Demystifying non-coding GWAS variants: an overview of computational tools and methods. *Hum. Mol. Genet.* **31**(R1): R73–R83. doi:10.1093/hmg/ddac198.
- Schmidt, D., Wilson, M.D., Ballester, B., Schwalie, P.C., Brown, G.D., Marshall, A., Kutter, C., Watt, S., Martinez-Jimenez, C.P., Mackay, S., Talianidis, I., Flicek, P., and Odom, D.T. 2010. Five-vertebrate ChIP-seq reveals the evolutionary dynamics of transcription factor binding. *Science* **328**(5981): 1036–1040. doi:10.1126/science.1186176.
- Schrider, D.R., and Kern, A.D. 2018. Supervised Machine Learning for Population Genetics: A New Paradigm. *Trends Genet.* **34**(4): 301–312. Elsevier. doi:10.1016/j.tig.2017.12.005.
- Schwessinger, R., Gosden, M., Downes, D., Brown, R.C., Oudelaar, A.M., Telenius, J., Teh, Y.W., Lunter, G., and Hughes, J.R. 2020. DeepC: predicting 3D genome folding using megabase-scale transfer learning. *Nat Methods* **17**(11): 1118–1124. doi:10.1038/s41592-020-0960-3.
- Segert, J.A., Gisselbrecht, S.S., and Bulyk, M.L. 2021. Transcriptional Silencers: Driving Gene Expression with the Brakes On. *Trends Genet.* **37**(6): 514–527. Elsevier. doi:10.1016/j.tig.2021.02.002.

- Shibata, Y., Sheffield, N.C., Fedrigo, O., Babbitt, C.C., Wortham, M., Tewari, A.K., London, D., Song, L., Lee, B.-K., Iyer, V.R., Parker, S.C.J., Margulies, E.H., Wray, G.A., Furey, T.S., and Crawford, G.E. 2012. Extensive Evolutionary Changes in Regulatory Element Activity during Human Origins Are Associated with Altered Gene Expression and Positive Selection. *PLOS Genet.* **8**(6): e1002789. Public Library of Science. doi:10.1371/journal.pgen.1002789.
- Siepel, A., Bejerano, G., Pedersen, J.S., Hinrichs, A.S., Hou, M., Rosenbloom, K., Clawson, H., Spieth, J., Hillier, L.W., Richards, S., Weinstock, G.M., Wilson, R.K., Gibbs, R.A., Kent, W.J., Miller, W., and Haussler, D. 2005. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.* **15**(8): 1034–1050. doi:10.1101/gr.3715005.
- Sievers, F., Wilm, A., Dineen, D., Gibson, T.J., Karplus, K., Li, W., Lopez, R., McWilliam, H., Remmert, M., Söding, J., Thompson, J.D., and Higgins, D.G. 2011. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol. Syst. Biol.* **7**: 539. doi:10.1038/msb.2011.75.
- Simonis, M., Klous, P., Splinter, E., Moshkin, Y., Willemsen, R., de Wit, E., van Steensel, B., and de Laat, W. 2006. Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture–on-chip (4C). *Nat. Genet.* **38**(11): 1348–1354. Nature Publishing Group. doi:10.1038/ng1896.
- Smit, A.F. 1999. Interspersed repeats and other mementos of transposable elements in mammalian genomes. *Curr. Opin. Genet. Dev.* **9**(6): 657–663. doi:10.1016/S0959-437X(99)00031-3.

- Smith, G.D., Ching, W.H., Cornejo-Páramo, P., and Wong, E.S. 2023. Decoding enhancer complexity with machine learning and high-throughput discovery. *Genome Biol.* **24**(1): 116. doi:10.1186/s13059-023-02955-4.
- Sohail, M., Palma-Martínez, M.J., Chong, A.Y., Quinto-Cortés, C.D., Barberena-Jonas, C., Medina-Muñoz, S.G., Ragsdale, A., Delgado-Sánchez, G., Cruz-Hervert, L.P., Ferreyra-Reyes, L., Ferreira-Guerrero, E., Mongua-Rodríguez, N., Canizales-Quintero, S., Jimenez-Kaufmann, A., Moreno-Macías, H., Aguilar-Salinas, C.A., Auckland, K., Cortés, A., Acuña-Alonzo, V., Gignoux, C.R., Wojcik, G.L., Ioannidis, A.G., Fernández-Valverde, S.L., Hill, A.V.S., Tusié-Luna, M.T., Mentzer, A.J., Novembre, J., García-García, L., and Moreno-Estrada, A. 2023. Mexican Biobank advances population and medical genomics of diverse ancestries. *Nature* **622**(7984): 775–783. Nature Publishing Group. doi:10.1038/s41586-023-06560-0.
- 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., and Harris, L.W. 2023. The NHGRI-EBI GWAS Catalog: knowledgebase and deposition resource. *Nucleic Acids Res.* **51**(D1): D977–D985. doi:10.1093/nar/gkac1010.
- Song, M., Peabworth, M.-P., Yang, X., Abnousi, A., Fan, C., Wen, J., Rosen, J.D., Choudhary, M.N.K., Cui, X., Jones, I.R., Bergenholtz, S., Eze, U.C., Juric, I., Li, B., Maliskova, L., Lee, J., Liu, W., Pollen, A.A., Li, Y., Wang, T., Hu, M.,

- Kriegstein, A.R., and Shen, Y. 2020. Cell-type-specific 3D epigenomes in the developing human cortex. *Nature* **587**(7835): 644–649. doi:10.1038/s41586-020-2825-4.
- Sonnenburg, S., Zien, A., and Rättsch, G. 2006. ARTS: accurate recognition of transcription starts in human. *Bioinformatics* **22**(14): e472–e480. doi:10.1093/bioinformatics/btl250.
- Spielmann, M., Lupiáñez, D.G., and Mundlos, S. 2018. Structural variation in the 3D genome. *Nat Rev Genet* **19**(7): 453–467. doi:10.1038/s41576-018-0007-0.
- Storey, J.D., Madeoy, J., Strout, J.L., Wurfel, M., Ronald, J., and Akey, J.M. 2007. Gene-expression variation within and among human populations. *Am J Hum Genet* **80**(3): 502–509. doi:10.1086/512017.
- Tagle, D.A., Koop, B.F., Goodman, M., Slightom, J.L., Hess, D.L., and Jones, R.T. 1988. Embryonic epsilon and gamma globin genes of a prosimian primate (*Galago crassicaudatus*). Nucleotide and amino acid sequences, developmental regulation and phylogenetic footprints. *J. Mol. Biol.* **203**(2): 439–455. doi:10.1016/0022-2836(88)90011-3.
- Tang, Z., Luo, O.J., Li, X., Zheng, M., Zhu, J.J., Szalaj, P., Trzaskoma, P., Magalska, A., Włodarczyk, J., Ruszczycki, B., Michalski, P., Piecuch, E., Wang, P., Wang, D., Tian, S.Z., Penrad-Mobayed, M., Sachs, L.M., Ruan, X., Wei, C.-L., Liu, E.T., Wilczynski, G.M., Plewczynski, D., Li, G., and Ruan, Y. 2015. CTCF-Mediated Human 3D Genome Architecture Reveals Chromatin Topology for Transcription. *Cell* **163**(7): 1611–1627. doi:10.1016/j.cell.2015.11.024.

- Tasaki, S., Gaiteri, C., Mostafavi, S., and Wang, Y. 2020. Deep learning decodes the principles of differential gene expression. *Nat. Mach. Intell.* **2**(7): 376–386. Nature Publishing Group. doi:10.1038/s42256-020-0201-6.
- Taylor, D.J., Chhetri, S.B., Tassia, M.G., Biddanda, A., Yan, S.M., Wojcik, G.L., Battle, A., and McCoy, R.C. 2024. Sources of gene expression variation in a globally diverse human cohort. *Nature* **632**(8023): 122–130. Nature Publishing Group. doi:10.1038/s41586-024-07708-2.
- Tehranchi, A., Hie, B., Dacre, M., Kaplow, I., Pettie, K., Combs, P., and Fraser, H.B. 2019. Fine-mapping cis-regulatory variants in diverse human populations. *eLife* **8**: e39595. eLife Sciences Publications, Ltd. doi:10.7554/eLife.39595.
- Telenti, A., Pierce, L.C.T., Biggs, W.H., di Iulio, J., Wong, E.H.M., Fabani, M.M., Kirkness, E.F., Moustafa, A., Shah, N., Xie, C., Brewerton, S.C., Bulsara, N., Garner, C., Metzker, G., Sandoval, E., Perkins, B.A., Och, F.J., Turpaz, Y., and Venter, J.C. 2016. Deep sequencing of 10,000 human genomes. *Proc. Natl. Acad. Sci.* **113**(42): 11901–11906. Proceedings of the National Academy of Sciences. doi:10.1073/pnas.1613365113.
- The GTEx Consortium. 2020. The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science* **369**(6509): 1318–1330. American Association for the Advancement of Science. doi:10.1126/science.aaz1776.
- The H3Africa Consortium, Matovu, E., Bucheton, B., Chisi, J., Enyaru, J., Hertz-Fowler, C., Koffi, M., Macleod, A., Mumba, D., Sidibe, I., Simo, G., Simuunza, M., Mayosi, B., Ramesar, R., Mulder, N., Ogendo, S., Mocumbi, A.O., Hugo-Hamman, C., Ogah, O., El Sayed, A., Mondo, C., Musuku, J., Engel, M., De Vries, J., Lesosky,

- M., Shaboodien, G., Cordell, H., Paré, G., Keavney, B., Motala, A., Sobngwi, E., Mbanya, J.C., Hennig, B., Balde, N., Nyirenda, M., Oli, J., Adebamowo, C., Levitt, N., Mayige, M., Kapiga, S., Kaleebu, P., Sandhu, M., Smeeth, L., McCarthy, M., and Rotimi, C. 2014. Enabling the genomic revolution in Africa. *Science* **344**(6190): 1346–1348. American Association for the Advancement of Science. doi:10.1126/science.1251546.
- Tian, P., Chan, T.H., Wang, Y.-F., Yang, W., Yin, G., and Zhang, Y.D. 2022. Multiethnic polygenic risk prediction in diverse populations through transfer learning. *Front. Genet.* **13**. Frontiers. doi:10.3389/fgene.2022.906965.
- Tolhuis, B., Palstra, R.J., Splinter, E., Grosveld, F., and de Laat, W. 2002. Looping and interaction between hypersensitive sites in the active beta-globin locus. *Mol Cell* **10**(6): 1453–1465. doi:10.1016/s1097-2765(02)00781-5.
- Van der Auwera, G.A., and O'Connor, B.D. 2020. *Genomics in the Cloud*. O'Reilly Media, Inc. Available from <https://www.oreilly.com/library/view/genomics-in-the/9781491975183/> [accessed 22 May 2023].
- Villar, D., Berthelot, C., Aldridge, S., Rayner, T.F., Lukk, M., Pignatelli, M., Park, T.J., Deaville, R., Erichsen, J.T., Jasinska, A.J., Turner, J.M.A., Bertelsen, M.F., Murchison, E.P., Flicek, P., and Odom, D.T. 2015. Enhancer Evolution across 20 Mammalian Species. *Cell* **160**(3): 554–566. doi:10.1016/j.cell.2015.01.006.
- Virtanen, P., Gommers, R., Oliphant, T.E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S.J., Brett, M., Wilson, J., Millman, K.J., Mayorov, N., Nelson, A.R.J., Jones, E., Kern, R., Larson, E., Carey, C.J., Polat, İ., Feng, Y., Moore, E.W., VanderPlas, J., Laxalde,

- D., Perktold, J., Cimrman, R., Henriksen, I., Quintero, E.A., Harris, C.R., Archibald, A.M., Ribeiro, A.H., Pedregosa, F., van Mulbregt, P., and SciPy 1.0 Contributors. 2020. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat. Methods* **17**(3): 261–272. doi:10.1038/s41592-019-0686-2.
- Wall, J.D., Stawiski, E.W., Ratan, A., Kim, H.L., Kim, C., Gupta, R., Suryamohan, K., Gusareva, E.S., Purbojati, R.W., Bhangale, T., Stepanov, V., Kharkov, V., Schröder, M.S., Ramprasad, V., Tom, J., Durinck, S., Bei, Q., Li, J., Guillory, J., Phalke, S., Basu, A., Stinson, J., Nair, S., Malaichamy, S., Biswas, N.K., Chambers, J.C., Cheng, K.C., George, J.T., Khor, S.S., Kim, J.-I., Cho, B., Menon, R., Sattibabu, T., Bassi, A., Deshmukh, M., Verma, A., Gopalan, V., Shin, J.-Y., Pratapneni, M., Santhosh, S., Tokunaga, K., Md-Zain, B.M., Chan, K.G., Parani, M., Natarajan, P., Hauser, M., Allingham, R.R., Santiago-Turla, C., Ghosh, A., Gadde, S.G.K., Fuchsberger, C., Forer, L., Schoenherr, S., Sudoyo, H., Lansing, J.S., Friedlaender, J., Koki, G., Cox, M.P., Hammer, M., Karafet, T., Ang, K.C., Mehdi, S.Q., Radha, V., Mohan, V., Majumder, P.P., Seshagiri, S., Seo, J.-S., Schuster, S.C., Peterson, A.S., and GenomeAsia100K Consortium. 2019. The GenomeAsia 100K Project enables genetic discoveries across Asia. *Nature* **576**(7785): 106–111. Nature Publishing Group. doi:10.1038/s41586-019-1793-z.
- Wang, W., Jiao, X., Sun, B., Liang, S., Wang, X., and Zhou, Y. 2022. DeepGenBind: a novel deep learning model for predicting transcription factor binding sites. *In* 2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). pp. 3629–3635. doi:10.1109/BIBM55620.2022.9994984.

- Wang, Y., Zhang, Y., Zhang, R., van Schaik, T., Zhang, L., Sasaki, T., Peric-Hupkes, D., Chen, Y., Gilbert, D.M., van Steensel, B., Belmont, A.S., and Ma, J. 2021. SPIN reveals genome-wide landscape of nuclear compartmentalization. *Genome Biol.* **22**(1): 36. doi:10.1186/s13059-020-02253-3.
- Weir, B.S., and Cockerham, C.C. 1984. Estimating F-Statistics for the Analysis of Population Structure. *Evolution* **38**(6): 1358–1370. Oxford University Press. doi:10.2307/2408641.
- Weischenfeldt, J., and Ibrahim, D.M. 2023. When 3D genome changes cause disease: the impact of structural variations in congenital disease and cancer. *Curr. Opin. Genet. Dev.* **80**: 102048. doi:10.1016/j.gde.2023.102048.
- Werling, D.M., Parikshak, N.N., and Geschwind, D.H. 2016. Gene expression in human brain implicates sexually dimorphic pathways in autism spectrum disorders. *Nat. Commun.* **7**: 10717. doi:10.1038/ncomms10717.
- Whalen, S., Inoue, F., Ryu, H., Fair, T., Markenscoff-Papadimitriou, E., Keough, K., Kircher, M., Martin, B., Alvarado, B., Elor, O., Laboy Cintron, D., Williams, A., Hassan Samee, Md.A., Thomas, S., Krencik, R., Ullian, E.M., Kriegstein, A., Rubenstein, J.L., Shendure, J., Pollen, A.A., Ahituv, N., and Pollard, K.S. 2023. Machine learning dissection of human accelerated regions in primate neurodevelopment. *Neuron* **111**(6): 857-873.e8. doi:10.1016/j.neuron.2022.12.026.
- Whalen, S., and Pollard, K.S. 2019. Most chromatin interactions are not in linkage disequilibrium. *Genome Res.* **29**(3): 334–343. doi:10.1101/gr.238022.118.

- Whalen, S., and Pollard, K.S. 2022. Enhancer Function and Evolutionary Roles of Human Accelerated Regions. *Annu. Rev. Genet.* **56**(Volume 56, 2022): 423–439. Annual Reviews. doi:10.1146/annurev-genet-071819-103933.
- Wohns, A.W., Wong, Y., Jeffery, B., Akbari, A., Mallick, S., Pinhasi, R., Patterson, N., Reich, D., Kelleher, J., and McVean, G. 2022. A unified genealogy of modern and ancient genomes. *Science* **375**(6583): eabi8264. American Association for the Advancement of Science. doi:10.1126/science.abi8264.
- Wojcik, G.L., Graff, M., Nishimura, K.K., Tao, R., Haessler, J., Gignoux, C.R., Highland, H.M., Patel, Y.M., Sorokin, E.P., Avery, C.L., Belbin, G.M., Bien, S.A., Cheng, I., Cullina, S., Hodonsky, C.J., Hu, Y., Huckins, L.M., Jeff, J., Justice, A.E., Kocarnik, J.M., Lim, U., Lin, B.M., Lu, Y., Nelson, S.C., Park, S.-S.L., Poisner, H., Preuss, M.H., Richard, M.A., Schurmann, C., Setiawan, V.W., Sockell, A., Vahi, K., Verbanck, M., Vishnu, A., Walker, R.W., Young, K.L., Zubair, N., Acuña-Alonso, V., Ambite, J.L., Barnes, K.C., Boerwinkle, E., Bottinger, E.P., Bustamante, C.D., Caberto, C., Canizales-Quinteros, S., Conomos, M.P., Deelman, E., Do, R., Doheny, K., Fernández-Rhodes, L., Fornage, M., Hailu, B., Heiss, G., Henn, B.M., Hindorff, L.A., Jackson, R.D., Laurie, C.A., Laurie, C.C., Li, Y., Lin, D.-Y., Moreno-Estrada, A., Nadkarni, G., Norman, P.J., Pooler, L.C., Reiner, A.P., Romm, J., Sabatti, C., Sandoval, K., Sheng, X., Stahl, E.A., Stram, D.O., Thornton, T.A., Wassel, C.L., Wilkens, L.R., Winkler, C.A., Yoneyama, S., Buyske, S., Haiman, C.A., Kooperberg, C., Le Marchand, L., Loos, R.J.F., Matisse, T.C., North, K.E., Peters, U., Kenny, E.E., and Carlson, C.S. 2019.

- Genetic analyses of diverse populations improves discovery for complex traits. *Nature* **570**(7762): 514–518. doi:10.1038/s41586-019-1310-4.
- Wong, E.S., Zheng, D., Tan, S.Z., Bower, N.I., Garside, V., Vanwalleghem, G., Gaiti, F., Scott, E., Hogan, B.M., Kikuchi, K., McGlenn, E., Francois, M., and Degnan, B.M. 2020. Deep conservation of the enhancer regulatory code in animals. *Science* **370**(6517): eaax8137. American Association for the Advancement of Science. doi:10.1126/science.aax8137.
- Woolfe, A., Goodson, M., Goode, D.K., Snell, P., McEwen, G.K., Vavouri, T., Smith, S.F., North, P., Callaway, H., Kelly, K., Walter, K., Abnizova, I., Gilks, W., Edwards, Y.J.K., Cooke, J.E., and Elgar, G. 2004. Highly Conserved Non-Coding Sequences Are Associated with Vertebrate Development. *PLOS Biol.* **3**(1): e7. Public Library of Science. doi:10.1371/journal.pbio.0030007.
- Xiong, K., and Ma, J. 2019. Revealing Hi-C subcompartments by imputing inter-chromosomal chromatin interactions. *Nat Commun* **10**(1): 5069. doi:10.1038/s41467-019-12954-4.
- Xu, J., Pratt, H.E., Moore, J.E., Gerstein, M.B., and Weng, Z. 2022. Building integrative functional maps of gene regulation. *Hum. Mol. Genet.* **31**(R1): R114–R122. doi:10.1093/hmg/ddac195.
- Xue, J.R., Mackay-Smith, A., Mouri, K., Garcia, M.F., Dong, M.X., Akers, J.F., Noble, M., Li, X., ZOONOMIA CONSORTIUM, Lindblad-Toh, K., Karlsson, E.K., Noonan, J.P., Capellini, T.D., Brennand, K.J., Tewhey, R., Sabeti, P.C., and Reilly, S.K. 2023. The functional and evolutionary impacts of human-specific deletions in

- conserved elements. *Science* **380**(6643): eabn2253. American Association for the Advancement of Science. doi:10.1126/science.abn2253.
- Yap, M., Johnston, R.L., Foley, H., MacDonald, S., Kondrashova, O., Tran, K.A., Nones, K., Koufariotis, L.T., Bean, C., Pearson, J.V., Trzaskowski, M., and Waddell, N. 2021. Verifying explainability of a deep learning tissue classifier trained on RNA-seq data. *Sci. Rep.* **11**(1): 2641. Nature Publishing Group. doi:10.1038/s41598-021-81773-9.
- Zeng, H., Edwards, M.D., Liu, G., and Gifford, D.K. 2016. Convolutional neural network architectures for predicting DNA–protein binding. *Bioinformatics* **32**(12): i121–i127. doi:10.1093/bioinformatics/btw255.
- Zhang, C., Rabiee, M., Sayyari, E., and Mirarab, S. 2018. ASTRAL-III: polynomial time species tree reconstruction from partially resolved gene trees. *BMC Bioinformatics* **19**(6): 153. doi:10.1186/s12859-018-2129-y.
- Zhang, F., and Lupski, J.R. 2015. Non-coding genetic variants in human disease. *Hum. Mol. Genet.* **24**(R1): R102-110. doi:10.1093/hmg/ddv259.
- Zhang, M.Q. 2002. Computational prediction of eukaryotic protein-coding genes. *Nat. Rev. Genet.* **3**(9): 698–709. Nature Publishing Group. doi:10.1038/nrg890.
- Zhang, Y., McCord, R.P., Ho, Y.-J., Lajoie, B.R., Hildebrand, D.G., Simon, A.C., Becker, M.S., Alt, F.W., and Dekker, J. 2012. Spatial organization of the mouse genome and its role in recurrent chromosomal translocations. *Cell* **148**(5): 908–921. doi:10.1016/j.cell.2012.02.002.

- Zhao, Z., Fritsche, L.G., Smith, J.A., Mukherjee, B., and Lee, S. 2022. The construction of cross-population polygenic risk scores using transfer learning. *Am. J. Hum. Genet.* **109**(11): 1998–2008. Elsevier. doi:10.1016/j.ajhg.2022.09.010.
- Zhao, Z., Tavoosidana, G., Sjölander, M., Göndör, A., Mariano, P., Wang, S., Kanduri, C., Lezcano, M., Singh Sandhu, K., Singh, U., Pant, V., Tiwari, V., Kurukuti, S., and Ohlsson, R. 2006. Circular chromosome conformation capture (4C) uncovers extensive networks of epigenetically regulated intra- and interchromosomal interactions. *Nat. Genet.* **38**(11): 1341–1347. Nature Publishing Group. doi:10.1038/ng1891.
- Zheng, H., and Xie, W. 2019. The role of 3D genome organization in development and cell differentiation. *Nat. Rev. Mol. Cell Biol.* **20**(9): 535–550. Nature Publishing Group. doi:10.1038/s41580-019-0132-4.
- Zhou, J. 2021, May 20. Sequence-based modeling of genome 3D architecture from kilobase to chromosome-scale. doi:10.1101/2021.05.19.444847.
- Zhou, J., Theesfeld, C.L., Yao, K., Chen, K.M., Wong, A.K., and Troyanskaya, O.G. 2018. Deep learning sequence-based ab initio prediction of variant effects on expression and disease risk. *Nat Genet* **50**(8): 1171–1179. doi:10.1038/s41588-018-0160-6.
- Ziller, M.J., Gu, H., Müller, F., Donaghey, J., Tsai, L.T.-Y., Kohlbacher, O., De Jager, P.L., Rosen, E.D., Bennett, D.A., Bernstein, B.E., Gnirke, A., and Meissner, A. 2013. Charting a dynamic DNA methylation landscape of the human genome. *Nature* **500**(7463): 477–481. doi:10.1038/nature12433.

Publishing Agreement

It is the policy of the University to encourage open access and broad distribution of all theses, dissertations, and manuscripts. The Graduate Division will facilitate the distribution of UCSF theses, dissertations, and manuscripts to the UCSF Library for open access and distribution. UCSF will make such theses, dissertations, and manuscripts accessible to the public and will take reasonable steps to preserve these works in perpetuity.

I hereby grant the non-exclusive, perpetual right to The Regents of the University of California to reproduce, publicly display, distribute, preserve, and publish copies of my thesis, dissertation, or manuscript in any form or media, now existing or later derived, including access online for teaching, research, and public service purposes.

Signed by:

Erin Nicole Gilbertson

2650AB23F68A467...

Author Signature

8/27/2024

Date