

UC San Diego

UC San Diego Electronic Theses and Dissertations

Title

Interpreting human genetic variations through transcriptional regulation and 3D genome organization

Permalink

<https://escholarship.org/uc/item/8wh4q1sw>

Author

Qiu, Yunjiang

Publication Date

2019

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA SAN DIEGO

Interpreting human genetic variations through transcriptional regulation and 3D genome organization

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

in

Bioinformatics and Systems Biology

by

Yunjiang Qiu

Committee in charge:

Professor Bing Ren, Chair
Professor Sheng Zhong, Co-Chair
Professor Kelly Frazer
Professor Christopher Glass
Professor Graham McVicker
Professor Cornelis Murre

2019

Copyright

Yunjiang Qiu, 2019

All rights reserved.

The Dissertation of Yunjiang Qiu is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

Co-Chair

Chair

University of California San Diego

2019

TABLE OF CONTENTS

SIGNATURE PAGE.....	iii
TABLE OF CONTENTS	iv
LIST OF FIGURES.....	vi
ACKNOWLEDGMENTS.....	viii
VITA.....	x
ABSTRACT OF THE DISSERTATION.....	xiii
INTRODUCTION	1
References	5
CHAPTER 1: Systematic analysis of differential transcription factor binding to	9
non-coding variants in the human genome	9
1.1 Abstract.....	9
1.2 Introduction.....	10
1.3 Results.....	11
1.4 Discussion	21
1.5 Methods.....	23
1.6 Figures.....	39
1.7 Supplemental Figures.....	46
1.8 Acknowledgments	56
1.9 References	57
CHAPTER 2: Common DNA sequence variation influences 3-dimensional conformation	73
of the human genome	73
2.1 Abstract.....	73
2.2 Introduction.....	74
2.3 Results.....	77
2.4 Discussion	91
2.5 Methods.....	94
2.6 Figures.....	115
2.7 Supplemental Figures.....	124
2.8 Acknowledgements	143

2.9 References	144
Chapter 3: Dynamic 3D chromatin organization during differentiation of human embryonic stem cells to pancreatic progenitor cells	152
3.1 Abstract.....	152
3.2 Introduction	154
3.3 Results.....	156
3.4 Discussion	164
3.5 Methods.....	166
3.6 Figures.....	170
3.7 Supplemental Figures.....	176
3.8 Acknowledgments	179
3.9 References	180

LIST OF FIGURES

Figure 1.1. Determination of differential DNA binding of human TFs to common sequence variants by SNP-SELEX.	39
Figure 1.2. Comparison of differentially bound SNPs identified by SNP-SELEX and PWM models.	40
Figure 1.3. pbSNPs uncover potential mode of action for likely T2D causal variants. ...	42
Figure 1.4. KEIS better predicts differential TF binding to non-coding variants in vitro and in vivo than PWM.	43
Figure 1.5. KEIS models identify candidate master TFs involved in complex traits and diseases.	45
Figure S1.1. Quality controls and reproducibility of SNP-SELEX data.	46
Figure S1.2. SNP-SELEX results are correlated with TF binding in vivo and enhancer activity from high through reporter assays.	48
Figure S1.3. SNP-SELEX results are correlated with enhancer activity from high through reporter assays.	50
Figure S1.4. Exploring the mode of action of likely T2D causal SNPs with the help of SNP-SELEX and Hi-C.	52
Figure S1.5. KEIS more accurately predicts non-coding variants affecting TF binding in vivo than PWM.	53
Figure S1.6. Analysis of differentially expressed genes upon knockdown of HLF and MAFG in HepG2 cells.	54
Figure S1.7. KEIS models help identify master TFs involved in complex traits and diseases.	55
Figure 2.1. Biological variability in multiple aspects of 3D chromatin.	115
Figure 2.2. Variable regions of 3D chromatin conformation.	116
Figure 2.3. Coordinated variation of the 3D genome, epigenome, and transcriptome.	118
Figure 2.4. A genetic contribution to variations in 3D chromatin conformation.	120
Figure 2.5. Contribution of 3D chromatin QTLs to other molecular and organismal phenotypes.	122
Figure S2.1. Hi-C derived molecular phenotypes measured across 20 LCLs.	124
Figure S2.2. FIRE measures density of local interactions.	126
Figure S2.3. Aggregate looping interactions in each sample.	128
Figure S2.4. 3D chromatin variation among 20 LCLs and H1-derived lineages.	130
Figure S2.5. Characterization of variable regions of 3D chromatin conformation.	132

Figure S2.6. Additional characterization of variable regions of 3D chromatin conformation.....	134
Figure S2.7. Coordinated variation between 3D chromatin conformation and multiple molecular phenotypes.	135
Figure S2.8. Correlations between DI, INS and multiple molecular phenotypes.....	137
Figure S2.9. Correlations between FIRE, interaction frequency and multiple molecular phenotypes.....	138
Figure S2.10. 3D chromatin QTLs.....	139
Figure S2.11. Influence of 3D chromatin QTLs on epigenomic and disease phenotypes.	141
Figure 3.1. Characterization of three-dimension chromatin organization during pancreatic differentiation.	170
Figure 3.2. Dynamic chromatin loops are associated with stage-specific transcription regulation.....	172
Figure 3.3. Dynamic chromatin loops are associated with lineage-determining TFs. ..	173
Figure 3.4. Identification and characterization of hubs in 3D interaction networks.	174
Figure S3.1. Characterization of three-dimension chromatin organization during pancreatic differentiation.	176
Figure S3.2. Quality control and characterization of hubs.....	177

ACKNOWLEDGMENTS

I would like to thank my advisor Bing Ren for his guidance and encouragement over the years. Bing has brought me to the field of genomics and provided countless support during my Ph.D. study.

I would like to thank my dissertation committee for their support and advice, including Sheng Zhong, Kelly Frazer, Christopher Glass, Graham McVicker, and Cornelis Murre. In particular, Graham McVicker provided informed guidance regarding human genetic.

I would like to thank my colleagues, David Gorkin, Ming Hu, Jian Yan, André Mauricio Ribeiro dos Santos, Ryan Geusz, Francesca Mulas, and Maike Sander for their close collaboration and in-depth insights across several projects. In particular, Dave has been an excellent mentor for me and helped me to develop myself as a scientist in various aspects, including how to approach a problem and how to make better presentations. I would also like to thank Kyle Gaulton and Josh Chiou for sharing their expertise in human genetics.

I would like to thank all past and present members of Bing Ren's lab for their support and helpful discussions especially Feng Yue, Anthony Schmitt, Inkyung Jung, Yarui Diao, Jason Li, Yanxiao Zhang, Miao Yu, and Yang Li. In particular, Feng has been my bioinformatics mentor when I just started in the lab as an undergraduate intern.

I would like to thank my family for their support and endless love. My parents, D. Qiu and Z. Wang, have always been supportive of every decision I made and encouraged me to pursue higher education. My wife, Rong Huang, has been the most important part of my life. She has always been there for me and supported me as we have gong through

ups and downs. My life during the six years wouldn't be such an incredible journey without her.

I would like to thank my close friends inside and outside the lab, including Mengchi Wang, Jia Lv, Jenhan Tao, Kai Zhang, Bingfei Yu, Yiren Hu, Zhanglong Ji, Xiang Fan, Miao Yu, Hui Huang, Rongxin Fang, Rong Hu, Qi Ma, and Jia Shen. Their friendship made life outside of the lab fun and memorable.

Chapter 1, in full, is a manuscript submitted as “Systematic analysis of transcription factor binding to non-coding variants in the human genome”. Jian Yan, Yunjiang Qiu, André M Ribeiro dos Santos, Yimeng Yin, Yang E. Li, Nick Vinckier, Naoki Nariai, Anugraha Raman, Zhe Liu, Joshua Chiou, Kelly A. Frazer, Kyle J. Gaulton, Maike Sander, Jussi Taipale, and Bing Ren. The dissertation author was the primary investigator and author of this paper.

Chapter 2, in full, is a manuscript submitted as "Common DNA sequence variation influences 3-dimensional conformation of the human genome". David U. Gorkin, Yunjiang Qiu, Ming Hu, Kipper Fletez-Brant, Tristin Liu, Anthony D. Schmitt, Amina Noor, Joshua Chiou, Kyle J Gaulton, Jonathan Sebat, Yun Li, Kasper D. Hansen, and Bing Ren. The dissertation author was the primary investigator and author of this paper.

Chapter 3, in full, is a manuscript in preparation as “Dynamic 3D chromatin organization during differentiation of human embryonic stem cells to pancreatic progenitor cells”. Yunjiang Qiu, Francesca Mulas, Ryan J Geusz, Jian Yan, Nick Vinckier, Allen Wang, Maike Sander, and Bing Ren. The dissertation author is the primary investigator and author of this paper.

VITA

EDUCATION

- 2013 Bachelor of Science, Peking University
Biological Science
- 2019 Doctor of Philosophy, University of California San Diego
Bioinformatics and Systems Biology

PUBLICATIONS

1. **Qiu, Y.***, Mulas, F.* , Geusz, R.* , Yan, J.* , Vinckier, N., Wang, A., Sander, M. & Ren, B. Dynamic 3D chromatin interaction during pancreatic differentiation of human embryonic stem cells. (In preparation)
2. Yan, J.* , **Qiu, Y.*** , Ribeiro-dos-Santos, A. M.* , Yin, Y., Li, Y. E., Vinckier, N., Nariyai, N., Raman, A., Liu, Z., Chiou, J., Frazer, K. A., Gaulton, K. J., Sander, M., Taipale, J. & Ren, B. Systematic analysis of differential transcription factor binding to non-coding variants in the human genome. (In review)
3. Greenwald, W. W. * , Chiou, J. * , Yan, J. * , **Qiu, Y.*** , Dai, N., Wang, A., Nariyai, N., Aylward, A., Han, J. Y., Kadakia, N., Regue, L., Okino, M.-L., Drees, F., Kramer, D., Vinckier, N., Minichiello, L., Gorkin, D., Avruch, J., Frazer, K. A., Sander, M., Ren, B. & Gaulton, K. J. Pancreatic islet chromatin accessibility and conformation reveals distal enhancer networks of type 2 diabetes risk. *Nature Communications* 10, 2078–12 (2019).
4. Crowley, C., Yang, Y., **Qiu, Y.**, Hu, B., Won, H., Ren, B., Hu, M. & Li, Y. FIREcaller: an R package for detecting frequently interacting regions from Hi-C data. *bioRxiv* 27, 619288 (2019).
5. Chaisson, M. J. P., Sanders, A. D., Zhao, X., Malhotra, A., Porubsky, D., Rausch, T., Gardner, E. J., Rodriguez, O. L., Guo, L., Collins, R. L., Fan, X., Wen, J., Handsaker, R. E., Fairley, S., Kronenberg, Z. N., Kong, X., Hormozdiari, F., Lee, D., Wenger, A. M., Hastie, A. R., Antaki, D., Anantharaman, T., Audano, P. A., Brand, H., Cantsilieris, S., Cao, H., Cerveira, E., Chen, C., Chen, X., Chin, C.-S., Chong, Z., Chuang, N. T., Lambert, C. C., Church, D. M., Clarke, L., Farrell, A., Flores, J., Galeev, T., Gorkin, D. U., Gujral, M., Guryev, V., Heaton, W. H., Korlach, J., Kumar, S., Kwon, J. Y., Lam, E. T., Lee, J. E., Lee, J., Lee, W.-P., Lee, S. P., Li, S., Marks, P., Viaud-Martinez, K., Meiers, S., Munson, K. M., Navarro, F. C. P., Nelson, B. J., Nodzak, C., Noor, A., Kyriazopoulou-Panagiotopoulou, S., Pang, A. W. C., **Qiu, Y.**, Rosanio, G., Ryan, M., Stütz, A., Spierings, D. C. J., Ward, A., Welch, A. E., Xiao, M., Xu, W., Zhang, C., Zhu, Q., Zheng-Bradley, X., Lowy, E., Yakneen, S., McCarroll, S., Jun, G., Ding, L., Koh, C. L., Ren, B., Flicek, P., Chen, K., Gerstein, M. B., Kwok, P.-Y., Lansdorp, P. M., Marth, G. T., Sebat, J., Shi, X., Bashir, A., Ye, K., Devine, S. E., Talkowski, M. E., Mills, R. E., Marschall, T., Korb, J. O., Eichler, E. E. & Lee, C.

- Multi-platform discovery of haplotype-resolved structural variation in human genomes. *Nature Communications* 10, 1784–16 (2019).
6. Juric, I., Yu, M., Abnousi, A., Raviram, R., Fang, R., Zhao, Y., Zhang, Y., **Qiu, Y.**, Yang, Y., Li, Y., Ren, B. & Hu, M. MAPS: Model-based analysis of long-range chromatin interactions from PLAC-seq and HiChIP experiments. *PLoS Comput Biol* 15, e1006982–24 (2019).
 7. Zhang, Y., Li, T., Preissl, S., Grinstein, J., Farah, E., Destici, E., Lee, A. Y., Chee, S., **Qiu, Y.**, Ma, K., Ye, Z., Zhu, Q., Huang, H., Hu, R., Fang R., Yu, L., Belmonte, J. C., Wu, J., Evans, S., Chi, N. & Ren, B. 3D Chromatin Architecture Remodeling during Human Cardiomyocyte Differentiation Reveals A Role Of HERV-H In Demarcating Chromatin Domains. *bioRxiv* 485961 (2019).
 8. Gorkin, D. U.* , **Qiu, Y.***, Hu, M.* , Fletez-Brant, K., Liu, T., Schmitt, A. D., Noor, A., Chiou, J., Gaulton, K. J., Sebat, J., Li, Y., Hansen, K. D. & Ren, B. Common DNA sequence variation influences 3-dimensional conformation of the human genome. *bioRxiv* 592741 (2019).
 9. Patel, L., Kang, R., Rosenberg, S. C., **Qiu, Y.**, Raviram, R., Chee, S., Hu, R., Ren, B., Cole, F. & Corbett, K. D. Dynamic reorganization of the genome shapes the recombination landscape in meiotic prophase. *Nature structural & molecular biology* 26, 164 (2019).
 10. Yan, J., Chen, S.-A. A., Local, A., Liu, T., **Qiu, Y.**, Dorighi, K. M., Preissl, S., Rivera, C. M., Wang, C., Ye, Z., Ge, K., Hu, M., Wysocka, J. & Ren, B. Histone H3 lysine 4 monomethylation modulates long-range chromatin interactions at enhancers. *Cell Res.* 28, 387–387 (2018).
 11. Plouffe, S. W., Lin, K. C., Moore, J. L., Tan, F. E., Ma, S., Ye, Z., **Qiu, Y.**, Ren, B. & Guan, K.-L. The Hippo pathway effector proteins YAP and TAZ have both distinct and overlapping functions in the cell. *Journal of Biological Chemistry* 293, 11230–11240 (2018).
 12. Meng, Z., **Qiu, Y.**, Lin, K. C., Kumar, A., Placone, J. K., Fang, C., Wang, K.-C., Lu, S., Pan, M. & Hong, A. W. RAP2 mediates mechanoresponses of the Hippo pathway. *Nature* 560, 655 (2018).
 13. Diao, Y., Fang, R., Li, B., Meng, Z., Yu, J., **Qiu, Y.**, Lin, K. C., Huang, H., Liu, T., Marina, R. J., Jung, I., Shen, Y., Guan, K.-L. & Ren, B. A tiling-deletion-based genetic screen for cis-regulatory element identification in mammalian cells. *Nat Meth* 14, 629–635 (2017).
 14. Fletez-Brant, K., **Qiu, Y.**, Gorkin, D. U., Hu, M. & Hansen, K. D. Removing unwanted variation between samples in Hi-C experiments. *bioRxiv* 214361 (2017).
 15. Schmitt, A. D., Hu, M., Jung, I., Xu, Z., **Qiu, Y.**, Tan, C. L., Li, Y., Lin, S., Lin, Y., Barr,

C. L. & Ren, B. A Compendium of Chromatin Contact Maps Reveals Spatially Active Regions in the Human Genome. *Cell Reports* 17, 2042–2059 (2016).

16. Wang, A., Yue, F., Li, Y., Xie, R., Harper, T., Patel, N. A., Muth, K., Palmer, J., **Qiu, Y.**, Wang, J., Lam, D. K., Raum, J. C., Stoffers, D. A., Ren, B. & Sander, M. Epigenetic priming of enhancers predicts developmental competence of hESC-derived endodermal lineage intermediates. *16*, 386–399 (2015).
17. Leung, D., Jung, I., Rajagopal, N., Schmitt, A., Selvaraj, S., Lee, A. Y., Yen, C.-A., Lin, S., Lin, Y., **Qiu, Y.**, Xie, W., Yue, F., Hariharan, M., Ray, P., Kuan, S., Edsall, L., Yang, H., Chi, N. C., Zhang, M. Q., Ecker, J. R. & Ren, B. Integrative analysis of haplotype-resolved epigenomes across human tissues. *Nature* 518, 350–354 (2015).
18. Hu, L., Di, C., Kai, M., Yang, Y.-C. T., Li, Y., **Qiu, Y.**, Hu, X., Yip, K. Y., Zhang, M. Q. & Lu, Z. J. A common set of distinct features that characterize noncoding RNAs across multiple species. *Nucleic Acids Res.* 43, 104–114 (2015).

* **co-first authors**

ABSTRACT OF THE DISSERTATION

Interpreting human genetic variations through transcriptional regulation and 3D genome organization

by

Yunjiang Qiu

Doctor of Philosophy in Bioinformatics and Systems Biology

University of California San Diego, 2019

Professor Bing Ren, Chair
Professor Sheng Zhong, Co-Chair

It has been more than a decade since the human genome was sequenced, but a complete understanding of the functional elements in the human genome is still lacking, especially for the non-coding part of the genome. The lack of complete understanding of the genome makes interpreting the function of genetic variants a daunting challenge.

Here I exploited multiple ways to decipher the function of genetic variants by leveraging knowledge about transcriptional regulation and three-dimension genome organization.

First, we developed SNP-SELEX, a high throughput method to assess the effect of SNPs on transcription factor (TF) binding. I demonstrated the superior performance of SNP-SELEX over previous delta PWM models, and applied results of SNP-SELEX to identify putative causal variants for type 2 diabetes. Furthermore, I employed deltaSVM algorithm to develop models that could predict the effect of SNPs on TF binding for any non-coding variants. Those models not only outperform delta PWM models *in vitro* and *in vivo* but also could help identify novel master regulator for complex traits and diseases.

Next, I co-led a study to investigate the effect of genetic variants on three-dimensional (3D) chromatin conformation. I identified thousands of regions across the genome where 3D chromatin conformation varies between individuals and found those variations often accompany changes in other genome functions. Moreover, I found DNA sequence variations could influence 3D chromatin conformation and mapped hundreds of Quantitative Trait Loci (QTLs) associated with 3D chromatin features, some of which confer disease risk.

Finally, I analyzed Hi-C data from human embryonic stem cells differentiated to beta cell progenitors to characterize changes in chromatin organizations during differentiation. I identified chromatin loops that are dynamic during different stages and found those loops are also associated with transcriptional regulation. Further, I revealed that chromatin loops form interaction hubs that are related to the establishment of stage-specific transcriptional programs.

INTRODUCTION

A long-standing question in the field of biology is how to interpret phenotypic variations due to genotypic variations. The initial sequencing of the human genome¹ decades ago provided an unprecedented opportunity to approach the question. In particular, with the advance of next-generation sequencing technologies, millions of genetic variants have been identified. These variants, including single nucleotide polymorphism (SNP), short indels, and structural variations (SV), have widely broadened our understanding of human genetic variations².

Genome-wide association studies (GWAS) have proven to be a powerful tool to link genetic variants to various phenotypes, especially complex traits and human diseases. To date, about 20,000 associations between genetic variants and traits have been identified by GWAS³. A better understanding of the mechanisms underlying trait-associated variants would likely provide valuable insights into the biological function of those genetic variants and may eventually lead to a more comprehensive picture of the relationships between genotype and phenotype⁴.

However, most of the genetic variants associated with traits or disease lie in the non-coding part of the genome^{5,6}, making it a daunting challenge to decipher their biological function and molecular mechanisms. It is believed that many of those non-coding variants lead to phenotypic changes through changes of gene expression, which are highly regulated processes involving multiple layers. For instance, Hnisz et al found that trait-associated variants are highly enriched in super-enhancers, which play vital roles in shaping cell-type-specific gene expression programs⁵. In another example, Huang et al showed that a genetic variant associated with prostate cancer risks affects

the expression level of RFX6 by changing in transcription factor HOXB13 binding and leads to differences in prostate cancer risk⁷.

Sequence-specific transcription factors (TFs) bind to *cis*-regulatory elements such as enhancers and promoters⁸ to modulate transcription of target genes. Many TFs are master regulators for cell identity. Over-expression of TFs could readily induce trans-differentiation across distinct cell types⁹. Because of the critical roles of TFs, it is crucial to understand how non-coding variants affect TF binding and gene expression. Despite a few case studies focused on specific variants, only less than 100 differential TF-DNA interactions have been linked to phenotypic variations¹⁰, leaving most of the trait-associated variants uncharacterized in terms of their effect on TF binding.

The most popular approach to quantify allelic TF binding is delta PWM scores based on position weight matrix (PWM) models. While it is convenient to compute delta PWM scores for any variants for TFs with known motifs, the prediction made by delta PWM scores suffer from pretty high false positives¹¹, limiting its broad applications. To overcome this challenge, as described in the first chapter of my thesis, my collaborators and I have systematically characterized the effect of 95,886 SNPs on the binding of 270 TF using a high-throughput assay termed as SNP-SELEX. In addition to experimentally assayed SNPs, I developed computational models that significantly outperform delta PWM models to predict the effect of TF binding to DNA variants. I demonstrated that the information about allelic TF binding could be leveraged to identify causal SNPs and pave the way to a complete picture of genetic variations' effect on various phenotypes.

The human genome adopts complicated higher-order structures in three-dimensional (3D) nuclear space. The higher-order chromatin structure has been studied

for decades. Early observations of 3D genome structure was primarily made using microscopy. For example, different chromosomes reside in distinct spatial territories¹². With the advent of C-techniques, including 3C¹³, 4C¹⁴, 5C¹⁵, and Hi-C¹⁶, that can map chromatin interactions in an unprecedented resolution and throughput, more features of chromatin conformation have been revealed. Within each chromosome territory, genome segments form the co-associated active or inactive compartments¹⁶. At the megabase scale, there are self-interacting blocks of DNA known as topological associated domains (TADs) that are conserved between different cell types¹⁷. At a finer scale, *cis*-regulatory elements that are distal from each other could form specific contacts¹⁸. At all levels, the 3D genome organization is related to numerous cellular and molecular events in cells such as transcription regulation, DNA replication, X chromosome inactivation, and DNA repair¹⁸⁻²¹.

As gene regulation is believed to be important in interpreting the relationship between genotype and phenotype, many molecular phenotypes have been mapped across individuals such as gene expressions²², histone modifications^{23,24}, and other epigenetic marks²⁵⁻²⁷, and extensive variations of those molecular phenotypes have been found. Since chromatin conformation is highly related to gene regulation, it is very likely that the difference in chromatin conformation also contributes to the difference in phenotype. There is also evidence showing that disruption of normal chromatin conformation leads to a phenotypical effect²⁸. However, how the 3D genome organization differs between normal individuals and its impact on other genome functions remain unexplored. With decreased cost of next-generation sequencing and more efficient Hi-C protocol, the study of variation of the chromatin conformation in multiple individuals

becomes feasible. The second chapter of my thesis focused on the study of the effect of genetic variations on 3D genome organization, which provides initial discoveries of genetic influence on 3D chromatin conformation and will facilitate future efforts to unravel the molecular basis of genetic disease risk.

To better understand the function of genetic variants, it is crucial to gain more insight about 3D genome organization, since the three-dimensional structure has been shown to be important in many genome functions. However, only a few studies focused on dynamics in the three-dimensional structure. One study focused on human ES cells and four ES-derived lineages and found that interactions within TADs change in conjunction with transcription activity while TADs are maintained²⁹. Another study used 5C to look at 3D interaction patterns during differentiation of pluripotent mouse embryonic stem cells (mES) cells along the neuroectoderm lineage and observed reorganization at the sub-megabase scale at seven loci³⁰. However, previous studies either lacked the resolution to study dynamics in interaction at the sub-megabase scale or lacked coverage to yield a complete result, and lots of questions are still unanswered. For instance, how enhancer-promoter interactions reorganize during differentiation and which transcriptional factors help establish stage-specific chromatin organization remain unveiled. The third chapter of my thesis described the study of the dynamic chromatin conformation in fine-resolution over a time course of human embryonic stem cell (hESC) differentiation. This not only expanded our knowledge regarding dynamics in 3D genome organization during differentiation but also helped us understand the relationship between chromatin organization and transcription regulation.

References

1. Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C. & al, E. Initial sequencing and analysis of the human genome. *Nature* (2001).
2. 1000 Genomes Project Consortium, Abecasis, G. R., Auton, A., Brooks, L. D., DePristo, M. A., Durbin, R. M., Handsaker, R. E., Kang, H. M., Marth, G. T. & McVean, G. A. An integrated map of genetic variation from 1,092 human genomes. *Nature* 491, 56–65 (2012).
3. Mills, M. C. & Rahal, C. A scientometric review of genome-wide association studies. *Communications Biology* 2, 1–11 (2018).
4. Visscher, P. M., Wray, N. R., Zhang, Q., Sklar, P., McCarthy, M. I., Brown, M. A. & Yang, J. 10 Years of GWAS Discovery: Biology, Function, and Translation. *Am. J. Hum. Genet.* 101, 5–22 (2017).
5. Hnisz, D., Abraham, B. J., Lee, T. I., Lau, A., Saint-André, V., Sigova, A. A., Hoke, H. A. & Young, R. A. Super-Enhancers in the Control of Cell Identity and Disease. *Cell* 155, 1–28 (2013).
6. Maurano, M. T., Humbert, R., Rynes, E., Thurman, R. E., Haugen, E., Wang, H., Reynolds, A. P., Sandstrom, R., Qu, H., Brody, J., Shafer, A., Neri, F., Lee, K., Kutyavin, T., Stehling-Sun, S., Johnson, A. K., Canfield, T. K., Giste, E., Diegel, M., Bates, D., Hansen, R. S., Neph, S., Sabo, P. J., Heimfeld, S., Raubitschek, A., Ziegler, S., Cotsapas, C., Sotoodehnia, N., Glass, I., Sunyaev, S. R., Kaul, R. & Stamatoyannopoulos, J. A. Systematic Localization of Common Disease-Associated Variation in Regulatory DNA. *Science* 337, 1190–1195 (2012).
7. Huang, Q., Whittington, T., Gao, P., Lindberg, J. F., Yang, Y., Sun, J., Väisänen, M.-R., Szulkin, R., Annala, M., Yan, J., Egevad, L. A., Zhang, K., Lin, R., Jolma, A., Nykter, M., Manninen, A., Wiklund, F., Vaarala, M. H., Visakorpi, T., Xu, J., Taipale, J. & Wei, G.-H. A prostate cancer susceptibility allele at 6q22 increases RFX6 expression by modulating HOXB13 chromatin binding. *Nat. Genet.* 46, 126–135 (2014).
8. Lambert, S. A., Jolma, A., Campitelli, L. F., Das, P. K., Yin, Y., Albu, M., Chen, X., Taipale, J., Hughes, T. R. & Weirauch, M. T. The Human Transcription Factors. *Cell* 172, 650–665 (2018).
9. Lee, T. I. & Young, R. A. Transcriptional regulation and its misregulation in disease. *Cell* 152, 1237–1251 (2013).
10. Deplancke, B., Alpern, D. & Gardeux, V. The Genetics of Transcription Factor DNA Binding Variation. *Cell* 166, 538–554 (2016).
11. Svetlichnyy, D., Imrichova, H., Fiers, M., Kalender Atak, Z. & Aerts, S. Identification of High-Impact cis-Regulatory Mutations Using Transcription Factor Specific Random

Forest Models. *PLoS Comput Biol* 11, e1004590 (2015).

12. Bolzer, A., Kreth, G., Solovei, I., Koehler, D., Saracoglu, K., Fauth, C., Müller, S., Eils, R., Cremer, C., Speicher, M. R. & Cremer, T. Three-dimensional maps of all chromosomes in human male fibroblast nuclei and prometaphase rosettes. *PLoS Biol.* 3, e157 (2005).
13. Dekker, J., Rippe, K., Dekker, M. & Kleckner, N. Capturing chromosome conformation. *Science* 295, 1306–1311 (2002).
14. van de Werken, H. J. G., Landan, G., Holwerda, S. J. B., Hoichman, M., Klous, P., Chachik, R., Splinter, E., Valdes-Quezada, C., Oz, Y., Bouwman, B. A. M., Verstegen, M. J. A. M., de Wit, E., Tanay, A. & De Laat, W. Robust 4C-seq data analysis to screen for regulatory DNA interactions. *Nat Meth* 9, 969–972 (2012).
15. 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. & Dekker, J. Chromosome Conformation Capture Carbon Copy (5C): a massively parallel solution for mapping interactions between genomic elements. *Genome Res.* 16, 1299–1309 (2006).
16. Lieberman-Aiden, E., van Berkum, N. L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., Amit, I., Lajoie, B. R., Sabo, P. J., Dorschner, M. O., Sandstrom, R., Bernstein, B., Bender, M. A., Groudine, M., Gnirke, A., Stamatoyannopoulos, J., Mirny, L. A., Lander, E. S. & Dekker, J. Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome. *Science* 326, 289–293 (2009).
17. Dixon, J. R., Selvaraj, S., Yue, F., Kim, A., Li, Y., Shen, Y., Hu, M., Liu, J. S. & Ren, B. Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* 485, 376–380 (2012).
18. Gorkin, D. U., Leung, D. & Ren, B. The 3D Genome in Transcriptional Regulation and Pluripotency. *Cell Stem Cell* 14, 762–775 (2014).
19. Pope, B. D., Ryba, T., Dileep, V., Yue, F., Wu, W., Denas, O., Vera, D. L., Wang, Y., Hansen, R. S., Canfield, T. K., Thurman, R. E., Cheng, Y., Gülsoy, G., Dennis, J. H., Snyder, M. P., Stamatoyannopoulos, J. A., Taylor, J., Hardison, R. C., Kahveci, T., Ren, B. & Gilbert, D. M. Topologically associating domains are stable units of replication-timing regulation. *Nature* 515, 402–405 (2015).
20. Engreitz, J. M., Pandya-Jones, A., McDonel, P., Shishkin, A., Sirokman, K., Surka, C., Kadri, S., Xing, J., Goren, A., Lander, E. S., Plath, K. & Guttman, M. The Xist lncRNA exploits three-dimensional genome architecture to spread across the X chromosome. *Science* 341, 1237973–1237973 (2013).
21. Misteli, T. & Soutoglou, E. The emerging role of nuclear architecture in DNA repair and genome maintenance. *Nat. Rev. Mol. Cell Biol.* 10, 243–254 (2009).

22. Lappalainen, T., Sammeth, M., Friedländer, M. R., 't Hoen, P. A. C., Monlong, J., Rivas, M. A., González-Porta, M., Kurbatova, N., Griebel, T., Ferreira, P. G., Barann, M., Wieland, T., Greger, L., van Iterson, M., Almlöf, J., Ribeca, P., Pulyakhina, I., Esser, D., Giger, T., Tikhonov, A., Sultan, M., Bertier, G., MacArthur, D. G., Lek, M., Lizano, E., Buermans, H. P. J., Padioleau, I., Schwarzmayr, T., Karlberg, O., Ongen, H., Kilpinen, H., Beltran, S., Gut, M., Kahlem, K., Amstislavskiy, V., Stegle, O., Pirinen, M., Montgomery, S. B., Donnelly, P., McCarthy, M. I., Flicek, P., Strom, T. M., Geuvadis Consortium, Lehrach, H., Schreiber, S., Sudbrak, R., Carracedo, A., Antonarakis, S. E., Häslér, R., Syvänen, A.-C., van Ommen, G.-J., Brazma, A., Meitinger, T., Rosenstiel, P., Guigo, R., Gut, I. G., Estivill, X. & Dermitzakis, E. T. Transcriptome and genome sequencing uncovers functional variation in humans. *Nature* 501, 506–511 (2013).
23. Kasowski, M., Kyriazopoulou-Panagiotopoulou, S., Grubert, F., Zaugg, J. B., Kundaje, A., Liu, Y., Boyle, A. P., Zhang, Q. C., Zakharia, F., Spacek, D. V., Li, J., Xie, D., Olarerin-George, A., Steinmetz, L. M., Hogenesch, J. B., Kellis, M., Batzoglou, S. & Snyder, M. Extensive variation in chromatin states across humans. *Science* 342, 750–752 (2013).
24. McVicker, G., van de Geijn, B., Degner, J. F., Cain, C. E., Banovich, N. E., Raj, A., Lewellen, N., Myrthil, M., Gilad, Y. & Pritchard, J. K. Identification of genetic variants that affect histone modifications in human cells. *Science* 342, 747–749 (2013).
25. 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. & Pritchard, J. K. DNase I sensitivity QTLs are a major determinant of human expression variation. *Nature* 482, 390–394 (2012).
26. Ding, Z., Ni, Y., Timmer, S. W., Lee, B.-K., Battenhouse, A., Louzada, S., Yang, F., Dunham, I., Crawford, G. E., Lieb, J. D., Durbin, R., Iyer, V. R. & Birney, E. Quantitative genetics of CTCF binding reveal local sequence effects and different modes of X-chromosome association. *PLoS Genet.* 10, e1004798 (2014).
27. Tehranchi, A. K., Myrthil, M., Martin, T., Hie, B. L., Golan, D. & Fraser, H. B. Pooled ChIP-Seq Links Variation in Transcription Factor Binding to Complex Disease Risk. *Cell* 165, 730–741 (2016).
28. 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. & Mundlos, S. Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions. *Cell* 161, 1012–1025 (2015).
29. Dixon, J. R., Jung, I., Selvaraj, S., Shen, Y., Antosiewicz-Bourget, J. E., Lee, A. Y., Ye, Z., Kim, A., Rajagopal, N., Xie, W., Diao, Y., Liang, J., Zhao, H., Lobanenko, V. V., Ecker, J. R., Thomson, J. A. & Ren, B. Chromatin architecture reorganization during stem cell differentiation. *Nature* 518, 331–336 (2015).

30. Phillips-Cremins, J. E., Sauria, M. E. G., Sanyal, A., Gerasimova, T. I., Lajoie, B. R., Bell, J. S. K., Ong, C.-T., Hookway, T. A., Guo, C., Sun, Y., Bland, M. J., Wagstaff, W., Dalton, S., McDevitt, T. C., Sen, R., Dekker, J., Taylor, J. & Corces, V. G. Architectural protein subclasses shape 3D organization of genomes during lineage commitment. *Cell* 153, 1281–1295 (2013).

CHAPTER 1: Systematic analysis of differential transcription factor binding to non-coding variants in the human genome

1.1 Abstract

A large number of sequence variants have been linked to complex human traits and diseases^{1,2}, but deciphering their biological function remains a daunting challenge especially for the non-protein-coding variants. To fill this gap, we have systematically assessed the differential binding of transcription factors (TF) to different alleles of noncoding variants in the human genome. Using an ultra-high throughput multiplex protein-DNA binding assay, we examined the binding of 270 human TFs to 95,886 common sequence variants within the 110 type 2 diabetes (T2D) risk loci. We then employed a machine-learning approach to derive computational models to predict differential DNA binding of 124 TFs to other common non-coding variants in the human genome. We showed that the newly derived models outperformed current position-weight matrices (PWM) in describing TF binding to non-coding variants and facilitated discovery of potential causal variants and dysregulated molecular pathways in human diseases.

1.2 Introduction

Sequence-specific TFs shape cell-type specific gene expression programs by binding to *cis*-regulatory sequences and modulating transcription of target genes. Mutations in the *cis*-regulatory sequences are believed to underlie the genetic basis of most complex human traits and disease³. Currently, we have very limited understanding of how disease associated non-coding variants affect binding of TFs and expression of target genes.

Tremendous efforts have been devoted to mapping of disease risk genetic variants, resulting in discovery of more than 10,000 single nucleotide polymorphisms (SNPs) associated with various disease and traits^{1,2}. One of the most extensively studied genetic diseases is Type 2 diabetes mellitus (T2D), which affects over 350 million people worldwide (by WHO: <http://www.who.int/news-room/fact-sheets/detail/diabetes>). Genome-Wide Association Study (GWAS) has identified 243 susceptible loci for T2D⁴. However, very few causal variants have been reported, with only a handful mechanistically characterized⁴⁻⁹. This is in large part because existing approaches to predicting non-coding variant effects on TF binding suffer from a high false positive rate and yet incomplete knowledge of the binding specificity for many human TFs. To better define the mode of action of non-coding genetic variants associated with human diseases and physiological traits, it is necessary to systematically assess TF binding to noncoding variants and develop quantitative and predictive models of TF binding.

1.3 Results

In this study, we adopted an ultra-high-throughput, multiplex TF-DNA binding assay, HT-SELEX (for High Throughput Systematic Evolution of Ligands by EXponential enrichment) to examine the binding of human TFs to common sequence variants in the human genome, with an initial focus on those associated with T2D^{10,11}. Compared to previous approaches, which employed randomized DNA sequences as the first cycle input¹¹, our strategy, referred hereafter as SNP-SELEX, used a library consisting of double-stranded custom-design DNA oligonucleotides. Each oligo included a 40-bp-sequence matching the reference human genomic DNA sequence with the center position corresponding to a single nucleotide polymorphism (SNP), permuted with all four nucleotides (**Figure 1.1a; Figure S1.1a**)^{10,11}. At the time when this project began, 110 distinct tagging SNPs were reported linked to T2D susceptibility^{5,12}. We designed 6,724 DNA oligos to represent these tagging variants and the SNPs in linkage disequilibrium (LD) with them. Additionally, we designed 89,162 oligos representing common SNPs located in annotated candidate *cis*-regulatory sequences within 500 kb of these 110 T2D tagging SNPs^{5,12}. A total of 768 SNP-SELEX experiments were conducted with 751 recombinant TF proteins, each with six cycles of consecutive binding, washing, elution and sequencing. Out of these experiments, 360 experiments passed QC, corresponding to 270 distinct TFs. Biological replicates were performed for 43 of these TFs. Additionally, DNA binding for 47 full length (FL) transcription factors was compared to their DNA-binding domains (DBDs). The results are accessible through a searchable web resource (GVAT database <http://renlab.sdsc.edu/GVATdb/>).

To determine the differential binding of TFs to the reference and alternative alleles of each SNP, we first identified the oligo sequences that exhibited significant binding to the TFs. We computed the relative enrichment of oligo sequences in the DNA oligo pool as an odds ratio after each cycle of binding, washing, elution and sequencing. We then defined the Oligo Binding Score (OBS) as the accumulative area under the curve (AUC) of the enrichment values across the six cycles. Using Monte-Carlo Randomization ($p < 0.05$; $n = 250,000$), we determined the significance of OBS for each oligo, finding 89,171 oligos that displayed binding to at least one TF (**Figure 1.1b, c**). We next defined the Preferential Binding Score (PBS) to describe the differential TF binding to the reference or alternative allele of each SNP, by subtracting OBS of its alternative allele from that of the reference allele. A total of 11,079 SNPs exhibited significant differential binding to at least one TF ($p < 0.01$; $n = 25,000$). We termed them pbSNPs hereafter. Among the 270 TFs that passed QC, 251 showed preferential binding to one or more pbSNPs. Overall, each TF bound differentially to a median number of 53 pbSNPs (**Figure 1.1d**), and each pbSNP was differentially bound by just one TF on average (**Figure 1.1e**).

Several lines of evidence support the high quality of the SNP-SELEX results. First, both OBS and PBS were highly reproducible in biological or technical replicative experiments for the same TF (**Figure S1.1b-d**), where data was available. Second, the PBS and OBS of the full-length TFs matched very well with those of the corresponding DNA-binding domains (DBD), to a similar degree as those between the biological replicates (**Figure S1.1b, d**), as noted previously¹¹. Third, the correlation between different TFs within the same structural family was significantly lower than biological replicates but higher than random pairs of TFs (Mann-Whitney U test, $p < 2 \times 10^{-16}$), also as

noted previously^{11,13}. The majority of TFs from the same families except for zinc finger family TFs, tended to have similar pbSNPs, consistent with previous reports^{11,14}(**Figure 1.1f**). Overall, our results suggest that the SNP-SELEX assay is a cost-effective and highly reproducible platform for systematic study of TF binding to non-coding variants *in vitro*.

The Position Weight Matrix (PWM) has been a standard tool to predict effects of SNPs on TF binding, but the accuracy of this approach has yet to be systematically validated. We compared the preferential binding scores (PBS) of TFs to differential PWM scores for 191 TFs for which both datasets were available. In general, the PBS and differential PWM scores (delta PWM score) correlated very well (**Figure 1.2a**). When differential PWM scores greater than two were used to assign preferential DNA binding to a SNP, PWM and SNP-SELEX agreed in more than 80% of cases (634,527 TF-SNP pairs) (**Figure 1.2b**). However, in a substantial fraction of cases (19.75%), predictions using the PWM models did not match those from SNPSELEX assays (153,411 TF-SNP pairs). These discordant cases mainly corresponded to weak TF-DNA binding events, as evaluated by both PWM scores and OBS (**Figure 1.2c, d**). Interestingly, the degrees of correlation between SNP-SELEX and PWM prediction varied dramatically among different TF structural families. For example, PBSs of the TFAP family were highly correlated with the differential PWM scores whereas MADS and E2F families showed poor concordance, despite similar information content of the PWM models (**Figure 1.2e**). One reason for the discrepancy could be interdependency between nucleotides within the DNA binding sites, which was not considered by PWM models¹¹. Another reason might be that some TFs could form heterodimers when binding to DNA, a scenario not

accounted for in the current design of SNP-SELEX experiments. Third, this could also be due to flanking sequence features favored by some TF families, such as A-stacking, yielding poorer predictions by the PWM models¹³. In any case, the above comparison suggests that the current PWM are imperfect models for assessing the impact of genetic variants on TF binding *in vitro*.

We also found that SNP-SELEX could more accurately predict the impact of a SNP on TF binding *in vivo* than PWM models. First, SNP-SELEX results better predicted allelic biases of DNA binding by sequence-specific transcription factors *in vivo*. We examined 14 ChIP-seq datasets, generated in-house or obtained from public databases, of 12 transcription factors in a human hepatocyte-derived cell line HepG2 or a lymphoblastoid cell line GM12878. Among the 85 SNPs displaying allelic biases in binding to one or more TFs in HepG2 cells and assayed in SNP-SELEX, the allelic imbalance ratios were significantly correlated with PBS from SNP-SELEX experiments (t-test $p=3.45\times 10^{-5}$, $r=0.42$; **Figure 1.2f**) while their correlation with differential PWM scores was much weaker (t-test $p=0.001$, $r=0.21$; **Figure S1.2a**). The same trend was observed in ChIP-seq datasets from GM12878 cells (**Figure S1.2b, c**). Second, pbSNPs better predicted allelic chromatin state at *cis*-regulatory elements than PWMs. Indeed, we quantified the enrichment of pbSNPs derived from SNP-SELEX experiments or predicted by differential PWM scores in genomic regions showing allelic biases in chromatin accessibility¹⁵ and active chromatin mark histone H3 lysine 27 acetylation¹⁶(H3K27ac hereafter) relative to non-pbSNPs. While significant enrichment was found for pbSNPs, no significant enrichment was detected for PWM predictions, and the difference was much weaker for PWM predictions (**Figure 1.2g, h; Figure S2d, e**). We also predicted the regulatory effect

of genetic variants assayed by SNP-SELEX using two well-established computational algorithms deltaSVM¹⁷ and DeepSEA¹⁸. Both algorithms predicted higher impact of pbSNPs on regulatory activity than the non-pbSNPs controls (deltaSVM $p=0.008$; DeepSEA $p=6.66e-07$; see also **Figure S1.2f, g**). By contrast, when using allelic SNPs predicted from PWM models, the regulatory activity difference between allelic SNPs and non-allelic SNPs was no longer significant (deltaSVM $p=0.745$; DeepSEA $p=0.014$). Third, pbSNPs better predicted a SNP's effect on enhancer activity than PWMs. Using a high throughput reporter assay STARR-seq¹⁹, we examined the enhancer activity of 2,246 pbSNPs and 1,697 non-pbSNPs containing genomic fragments in HepG2 cells and HEK293T cells (**Figure S1.3a**). We found genomic DNA corresponding to 424 and 527 pbSNPs showed significant enhancer activity in HepG2 and HEK293 cells, respectively (empirical FDR<0.05). Additionally, 200 SNP containing fragments displayed allelic biases on enhancer activity in HepG2 cells and 206 in HEK293T cells (FDR<0.05) and termed these SNPs paSNPs (**Figure S1.3b**). We found that the pbSNPs were more likely to be associated with allelic enhancer activity than non-pbSNPs (Fisher exact test $p=0.02$, OR=1.62; **Figure 1.2i**). By contrast, SNPs predicted by PWMs to be differentially bound by a TF were not enriched for SNPs with differential enhancer activity (**Figure S1.3c**). These results, taken together, strongly suggest that SNP-SELEX results more accurately predict TF binding and regulatory activity *in vivo* than PWMs. Therefore, SNP-SELEX is a valuable tool to study the interactions of transcription factors with non-coding variants in the human genome.

To further demonstrate the utility of SNP-SELEX, we explored further the T2D risk variants in terms of differential TF binding and consequences on target gene expression.

First, supporting the hypothesis that non-coding SNPs may contribute directly to T2D susceptibility by affecting TF binding, we found that the aforementioned pbSNPs were indeed enriched in a set of likely T2D causal variants defined in a recent genetic fine-mapping study²⁰ (Fisher exact test $p=0.014$, $OR=1.08$). As a matter of fact, 1,538 out of 70,975 likely causal variants are pbSNPs of the TFs that we tested. Next, we defined candidate target genes for these likely functional pbSNPs. Using high resolution *in situ* Hi-C²¹, we identified 9,108 and 9,789 long-range chromatin interactions in HepG2 cells and human pancreatic islet tissues, respectively (**Figure S1.4a**), and then used this information to assign target genes to SNPs when a SNP and a gene promoter were located within the two anchors of a chromatin loop (**Figure 1.3a**). We also assigned a SNP to a gene when it was within 2 kb upstream of the gene's transcription starting site (TSS). With this approach, we assigned candidate target genes to 205 pbSNPs in HepG2 cells, and 250 pbSNPs in pancreatic islet tissues. For example, SNP rs7578326, located in a super-enhancer, was predicted to affect the binding of a liver-specific TF CEBPB (**Figure 1.3b; Figure S1.4b**). The super-enhancer that harbored this SNP was linked to Insulin Receptor Substrate 1 (IRS1) gene located ~500kb downstream through long-range chromatin interaction in HepG2 cells. To confirm the regulatory role of the underlying SNP-harboring enhancer in HepG2 cells, we silenced the region in HepG2 and HEK293T cells using CRISPR interference (CRISPRi) with a guide RNA (sgRNA) targeting the sequence adjacent to the SNP rs7578326. Significant reduction of IRS1 was observed in HepG2 cells, which expressed a high level of TF CEBPB protein, whereas no reduction was detected in the control HEK293T cells, where the expression of CEBPB was much lower (**Figure 1.3c**). This result was consistent with an independent study

showing that this SNP was an eQTL of IRS1 gene in liver and adipose tissue (**Figure 1.3d; Figure S1.4c**). The same SNP has also been reported to be genetically associated with fasting insulin level and insulin sensitivity²⁴, suggesting its role in T2D pathogenesis likely through regulation of insulin sensitivity in metabolic organs^{22,23}. In another case, a candidate SNP rs231361 displayed allelic binding for several RFX TFs including RFX1 and RFX2, consistent with a previous report that disruption of RFX motif increases the susceptibility of T2D²⁴. The region enclosing rs231361 interacted with genes SLC22A18 and CDKN1C located ~500Kb away, evidenced by in situ Hi-C data (**Figure S1.4d**). An earlier study has suggested that CDKN1C can mediate T2D susceptibility through its regulatory function in beta cell early development²⁵. Together, our findings extended previous knowledge on T2D risk variants and demonstrated allelic TF binding as a valuable and unique resource to prioritize casual variants and understand their underlying mechanisms.

The number of SNPs functionally tested in the SNP-SELEX assay is still finite and far less than the non-coding SNPs in the human genome²⁶. To be able to predict differential DNA binding by a TF to any genetic variant in the human genome, we used a machine learning approach trained with our SNP-SELEX raw data. Specifically, we employed the deltaSVM algorithm¹⁷ to train our computational models using the enriched oligo sequences from each SNP-SELEX experiment (**Figure 1.4a**). We named these computational models as k-mer based estimation of impact of SNP on TF binding (KEIS) (**Figure 1.4a**). We successfully obtained KEIS models for 167 TFs with excellent performance in five-fold cross-validation. The median area under the receiver operating characteristic curve (AUROC) of these TFs is 0.980, while the median area under the

Precision-Recall (PR) curve (AUPRC) is 0.830 (**Figure 1.4b**). By contrast, the performance of PWM in predicting differentiation DNA binding was much lower, with AUROC and AUPRC 0.898 and 0.470 respectively (**Figure 1.4b**). Among the 167 TFs with KEIS models, 124 TFs had an AUPR score higher than 0.75, and were used for subsequent genome-wide prediction (**Figure 1.4c**). KEIS models of these 124 TFs predicted 1,827,007 out of 10,679,051 common variants in the human genome as differentially recognized by one or more of the TFs. To avoid being confused with SELEX derived allelic SNPs (pbSNPs), we termed these SNPs predicted by KEIS models k-SNPs.

Like SNP-SELEX, KEIS models also outperformed PWMs in predicting differential TF binding to SNPs *in vivo* (**Figure 1.2f; Figure S1.2b**). Analyzing the allelic DNA TF binding in HepG2 cells from CHIP-seq datasets, KEIS models recovered twice as many SNPs with allelic DNA binding than PWM models (**Figure 1.4d; Figure S1.5a**). Similarly, KEIS models could explain a more significant percentage of allelic DNA binding for ATF2, PKNOX1 and NR2F1 in GM12878 cells than PWM models (**Figure S1.5b, c**). Additionally, KEIS models recovered SNPs with allelic regulatory effects at an equivalent odds ratio to the original SNP-SELEX data, further supporting the reliability of the computational models of TF binding specificity (**Figure 1.4e, f; Figure 1.2g, h**).

Master transcription factors are key nodes in the transcriptional network of each cell lineage. We reasoned that SNPs affecting the DNA binding of master TFs were more likely to cause disease than other categories of SNPs. In line with predicting the TFs most likely affected by the genetic variants associated with a particular trait or disease, we used stratified LD score regression (S-LDSC)²⁷ to determine the enrichment of SNPs showing

differential binding to a TF in SNPs associated with traits and diseases in order to associate the TF with the corresponding phenotypes 31,37-45^{23,28-36}. As a first step, we focused on T2D-relevant traits as well as several brain-related traits and diseases. As expected, TFs previously known to be associated with these traits showed strong enrichment (**Figure 1.5a**).

In our analysis, we found TFAP2B, a known regulator of insulin resistance and central adiposity³⁷, was enriched in the set of non-coding variants that had been associated with fasting glucose traits. CCAAT-Enhancer Binding Protein- β (CEBPB), a pivotal factor in Alzheimer's disease (AD) pathogenesis³⁸ that was up-regulated in the AD cortex³⁹, was also enriched in the set of non-coding SNPs associated with this trait. Similarly, ELK1, recently found to be selectively increased in depressed patient and mouse models of depression⁴⁰, was significantly affected by SNPs associated with heritability of major depressive disorders. These examples suggest that we could use the k-SNPs to predict master regulators involved in a particular trait or disease phenotype.

Furthermore, we identified novel candidate TFs associated with additional human traits and diseases. For instance, MAFG is predicted to act in regulating fasting insulin, which was a well-known sign of insulin sensitivity²⁴. To validate this prediction, we identified the genes differentially expressed following shRNA knockdown of MAFG in HepG2 cells and found that genes in the PPAR signaling pathway were most affected (**Figure 1.5b; Figure S1.6a, b**). It was known that the activation of PPAR signaling pathway regulated the insulin signaling cascade and insulin sensitivity⁴¹. This result therefore suggests that MAFG could regulate expression of genes in PPAR signaling pathway and modulate insulin sensitivity and the fasting insulin level.

Another master TF predicted by our analysis to be associated with circulating triglycerides level (**Figure 1.5a**) is the hepatic leukemia factor (HLF), previously known to be involved in childhood B-lineage acute lymphoid leukemia⁴². Consistent with this prediction, knockdown of HLF in HepG2 cells by RNA interference resulted in changes of mRNA expression in genes significantly involved in metabolic pathways and PPAR signaling pathway (**Figure 1.5c; Figure S1.6c, d**), both of which contributed to the regulation of blood triglycerides. In particular, HLF directly regulated the expression of APOC3, a gene encoding apolipoprotein C III (APOC-III) and known to be important for triglyceride-rich lipoprotein (TRL) metabolism^{43,44}. APOC-III is among the most affected genes after HLF knockdown (**Figure 1.5d**). ChIP-seq experiment further showed that HLF bound to a candidate enhancer enclosing SNP rs7118999, located approximately 70 kb upstream of, but was spatially close to, the APOC3 gene promoter (**Figure 1.5e**). Importantly, allelic binding of HLF to the heterozygous SNP rs7118999 was accompanied with the allelic expression of APOC-III in HepG2 cells, where higher binding of HLF corresponded to higher expression of APOC-III in *cis* (**Figure 1.5e**). Therefore, HLF could regulate APOC-III expression and in turn mediates the size of triglyceride-rich lipoprotein, which is a major risk factor for coronary artery disease (CAD)^{44,45}. Since APOC-III has already been considered as a target to reduce the risk of CAD in a variety of clinical studies⁴⁶, our analysis supports that HLF could be a novel therapeutic target for CAD.

1.4 Discussion

Overall, the above results demonstrated the power of high throughput SNP-SELEX approach in the study of human disease and traits. While GWAS can locate disease-associated genetic variants, determining the molecular mechanisms of these variants remains difficult⁴⁷⁻⁴⁹. Currently, fewer than 50 risk SNPs have been mechanistically characterized for TF binding (reviewed in Deplancke et al⁵⁰), while the mode of action for more than 99% of disease associated non-coding genetic variants is unknown. Here, we have established a general strategy and public resource to address this challenge. Using SNP-SELEX method, we have systematically investigated the differential binding of 270 distinct human TFs, from 25 different structural families, to 95,886 common SNPs. The current study has not only improved our knowledge of TF binding specificity for nearly 200 human TFs, but also defined DNA binding specificity for additional 79 TFs, for which PWM models are not yet available. More importantly, using machine-learning techniques, we built highly quantitative and predictive models for 124 TFs and expanded the differential TF binding analysis to the rest of non-coding variants in the human genome. We demonstrated the superior performance of our TF binding models over the standard PWM models in predicting the influence of SNPs on TF binding both *in vitro* and *in vivo*.

Additionally, we used these models to assess the impact of risk variants on TF binding, and predict master TFs potentially involved in a variety of complex traits and diseases. This allows us to analyze the enrichment of allelic SNPs of various TFs in a variety of trait-associated genetic variants and identify potential master TFs involved in these phenotypic traits. It is important to note that if we perform enrichment analysis for the presence of trait-associated SNPs in TF binding sites alone without including the

allelic binding information, we won't be able to recover the trait-associating master TFs (**Figure S1.7**), demonstrating crucial roles of allelic TF binding information.

In summary, our results addressed a critical gap in our knowledge about the impact of variants on TF binding. By combining data from population genetic studies and quantitative models of TF binding to non-coding SNPs, we will be in an excellent position to unravel the mechanisms of human disease and identify new therapeutic targets.

1.5 Methods

SNP Selection

In total, 110 leading SNPs were selected from previous T2D GWAS^{5,6}. Common SNPs (minor allele frequency > 1%) within 500 kb of the 110 leading SNPs were extracted from 1000 Genome Project from all available populations, resulting in 379,895 unique SNPs. From these SNPs, 6,724 SNPs were selected in Linkage Disequilibrium with leading SNPs in East Asian and Caucasian populations ($r^2 \geq 0.8$) from 1000 Genome Project Pilot 1⁵¹, and 89,162 SNPs were selected based on their distance (≤ 500 kb) to the accessible chromatin regions in ENCODE DHS sites⁵² or FANTOME 5⁵³ permissive enhancers for all cell and tissue types. Altogether, 95,886 SNPs were included in the current study.

SNP-SELEX Experiments

Oligo design was adapted to illumina TruSeq dual-index system (**Figure S1.1a**) and synthesized as a 92,000 pool by CustomArray (Seattle, WA). The oligos were amplified using 20 cycles of PCR and sequenced with illumina HiSeq 2500 to verify the identities. The cDNAs of TF proteins were cloned to pET20a plasmids⁵⁴ and expressed using Rosetta (DE3) pLysS strains.

The HT-SELEX experiments were performed essentially the same as previously described¹¹. Briefly, the E. coli expressed 6xHis-tagged TF proteins were immobilized to Ni sepharose beads (GE, 17-5318-01) in Promega binding buffer (10mM Tris pH7.5, 50mM NaCl, 1mM MgCl₂, 4% glycerol, 0.5mM EDTA, 5 μ g/ml poly-dIdC). Oligos from input or previous HT-SELEX cycles were added into the protein beads mixture and incubated at ambient temperature for 30 min. After binding, the beads were consecutively

washed for 12 times with the Promega binding buffer. After final wash, TE (10mM Tris pH 8.0, 1mM EDTA) was used to re-suspend the beads and for PCR amplification. The PCR products from each HT-SELEX cycle were purified (Qiagen, 28004) and sequenced with illumina HiSeq 2500.

Cell Culture and Transfection

The HEK293T cells (ATCC, CRL-3216) and HepG2 (ATCC, HB-8065) cells were cultured under normal condition with 5% CO₂ at 37 °C. Fugene HD (Promega, E2311) was used for plasmid transfection. Specifically, 2 µg of STARR-seq plasmids were mixed with 5 µl of transfection reagents for transfection into 300,000 cells cultured in a single well of 6-well plate. For siRNA transfection, HiPerfect transfection was used following the manufacture guidance. For each experiment, 50 nM of siRNA was used with 5 ul of HiPerfect reagent to make the transfection complex for 1-3×10⁴ cells. Cells were continued to be cultured for 72 hours. The siRNAs targeting human HLF (cat. #GS3131) and MAFG (cat. #GS4097) were commercially available from Qiagen. Silencer negative control siRNA was commercially manufactured and order from Thermo Fisher (cat. #AM4635).

STARR-seq Experiments

To directly evaluate pbSNP impact on enhancer activity, STARR-seq¹⁹ was conducted with human embryonic kidney (HEK293T) and human hepatocarcinoma (HepG2) cell lines. In total, 11,961 genomic sequences harboring 2,246 pbSNPs and 1,697 non-pbSNPs were designed. In addition, we included 37 true positive controls which are known enhancers and 2,998 negative controls that are random yeast open read frames (ORFs) sequences.

Oligo design was adapted from the previously published STARR-seq work¹⁹ and synthesized from Agilent (Santa Clara, CA). Briefly, each oligo contains 190 bp of genomic sequence enclosing the SNP and 20 bp constant flanking sequences (upstream: 5'- ACACGACGCTCTTCCGATCT; downstream: AGATCGGAAGAGCACACGTC-3') on both sides which were used for amplification and cloning. The generic PCR primers including illumina Truseq adapter sequences and different indexes were used to amplify the oligo pool and cloned into the human STARR-seq plasmid (a gift from the Stark lab, Austria). PCR amplification from the plasmids was performed and sequenced for 2x100 paired-end cycles with illumina HiSeq 4000 sequencer as input control.

The plasmid pool was transfected into HEK293T or HepG2 cell lines using Fugene HD and continued culturing for 48 hours before harvest. Total RNA was extracted with RNeasy kit (Qiagen, 74104) and mRNA was enriched with poly(dT)25 Dynabeads (Invitrogen, 61002). First strand cDNA was synthesized using a specific primer (5'- CAAACTCATCAATGTATCTTATCATG) with high High-Capacity cDNA Reverse Transcription kit (ThermoFisher Scientific, 4368814). The nested PCR was used to amplify the SNP specific fragments from cDNA, first using two reporter-specific PCR primers (5'-GGGCCAGCTGTTGGGGTGTCCAC & 5'-CTTATCATGTCTGCTCGAAGC) and then generic primers used in HT-SELEX. DNA was purified with AMPure beads and sequenced for 2x100 paired-end cycles with illumina HiSeq 2500 sequencer. In total, three biological replicates were performed with two technical replicates each for both HepG2 and HEK293T cells.

in situ Hi-C Experiments

The *in situ* Hi-C was performed according to a previously described protocol²¹ with slight modifications. Briefly, the human islets were washed with cold PBS and cut into small pieces. For HepG2 cells, the cells were trypsinized and washed with PBS. The chromatin was cross-linked with 1% formaldehyde (Sigma) at ambient temperature for 10 min and quenched with 125mM glycine for 5 min. PBS washed tissue was homogenized with loose fitting douncer for 30 strokes before centrifugation to isolate the nuclei.

Nuclei were isolated and directly applied for digestion using 4 cutter restriction enzyme MboI (NEB) at 37 °C o/n. The single strand overhang was filled with biotinylated-14-ATP (Life Tech.) using Klenow DNA polymerase (NEB). Different from tradition Hi-C, with *in situ* protocol the ligation was performed when the nuclear membrane was still intact. DNA was ligated for 4h at 16 °C using T4 ligase (NEB). Protein was degraded by proteinase K (NEB) treatment at 55 °C for 30 min. The crosslinking was reversed with 500 mM of NaCl and heated at 68 °C o/n. DNA was purified and sonicated to 300-700 bp small fragments. Biotinylated DNA was selected with Dynabeads My One T1 Streptavidin beads (Life Tech.). Sequencing library was prepared on beads and intensive wash was performed between different reactions. Libraries were checked with Agilent TapeStation and quantified using Qubit (Life Tech.). Libraries were sequenced with illumina HiSeq 4000 100 cycles of paired-end reads.

ChIP-seq Experiments

The ChIP-seq experiment was carried out using an established protocol⁵⁵. Briefly, the cells were crossed linked with 1% formaldehyde at ambient temperature for 10 min. The reaction was quenched by 125mM glycine for 5 min at room temperature. Cells were washed with PBS and treated with hypotonic buffer (20mM HEPES pH7.9, 10mM KCl,

1mM EDTA, 10% Glycerol and 1mM DTT with additional protease inhibitor (Roche)) to isolate nuclei. The nuclei were suspended with RIPA buffer (10 mM Tris-HCl pH 8.0, 140 mM NaCl, 1 mM EDTA, 1% Triton X-100, 0.1% SDS, 0.1% sodium deoxycholate with protease inhibitor) and sonicated using Covaris S220 Focused-ultrasonicator. Fragmented chromatin was pre-cleared with protein G conjugated sepharose beads (GE).

Antibodies against HLF (Santa Cruz, sc-134359), MAFG (Santa Cruz, sc-166548 X), Histone H3K4me1 (Abcam, ab8895), H3K4me3 (Abcam, ab8580) and H3K27ac (Abcam, ab4729) were used to pull down the respective proteins and their associated chromatin. Washes with different concentration of NaCl were performed. The enriched protein-DNA complexes were reverse crosslinked at 65 °C over night with proteinase K (NEB). DNA was purified with Qiagen MinElute kit.

Sequencing library was prepared using an in-house kit, including end-repair, “A” addition and adapter ligation. The library was sequenced with illumina HiSeq 4000 for 50bp single reads or 100bp pair-end reads.

Whole Genome Sequencing

The genomic DNA was extracted using Qiagen kit (cat. no. 69506). The DNA was then fragmented with Covaris S220 ultrasonicator to 300-500 bp long. Sequencing library was then prepared using the same in-house kit as ChIP-seq, including end-repair, “A” addition and adapter ligation. The library was sequenced with illumina HiSeq 4000 sequenced for 100 bp paired-end reads to achieve an average coverage of 30-40 times of the human genome.

RNA-seq Experiments

The total RNA was isolated using Qiagen RNeasy mini kit. The sequencing library was prepared using Truseq illumine RNA Library Prep Kit v2 (cat. #RS-122-2001). The library was sequenced using illumina HiSeq 4000 for 100bp paired-end reads.

CRISPRi

CRISPR/dCas9 fused with KRAB domain (addgene cat. no. 71236) was introduced to genomic locus enclosing the SNP rs7578326 using sgRNA (targeting sequence TCCGTTGGTGACACAGTTGG) in HepG2 cells. CRISPR/dCas9 with the same sgRNA was used as negative control. Similarly, both plasmids were transfected in 293T cells as control. RNA was extracted using Qiagen RNeasy kit and reverse transcribed using High-Capacity cDNA Reverse Transcription Kit (Thermo). Quantitative PCR was performed to measure the expression of IRS1 gene using pre-designed primers (Qiagen QT00074144) and beta actin for internal control (Qiagen QT00095431). Triplicates were carried out for each experiment to compute the statistical significance using t-test.

SNP-SELEX Data Analysis

Sequencing data of each HT-SELEX cycle was aligned to the oligo library using BWA⁵⁶. Several filters were applied to aligned reads after alignments: 1) Reads of low quality, containing ambiguous bases, unaligned to reference and aligned outside of the oligo boundaries were filtered out and experiments with less than 10,000 reads were excluded from further analysis; 2) To control for PCR-duplication bias, the frequency of all PCR bias control (PDC) sequences (256 combinations) of each cycle were compared to the input library (cycle 0) using a linear regression model. PDC whose difference

between expected and observed frequency exceeded 30% of the observed values were considered biased and all reads containing the biased PDC were removed.

De novo motif discovery was then conducted using the cycle six reads with Homer toolset⁵⁷. Motifs were then compared to JASPAR 2016 non-redundant vertebrates' motifs⁵⁸ and HT-SELEX models to examine quality of the experiments⁵⁴. Only SNP-SELEX experiments whose motif models match either its TF or TF of same structural family¹¹ were kept for further analysis. The frequencies of reads supporting each SNP oligo and its alleles were obtained from the remaining dataset.

Aiming to quantify the TF binding to genomic oligo, oligo binding score (OBS) was defined as area under the curve (AUC) of logarithmic odds-ratio along HT-SELEX cycles. We first estimated odds ratio of observing oligo o , OR_c at cycle c compared to the input library, where $o.Odds_c$ is the odds of observing oligo at cycle c regarding all other oligos. OBS was then computed as AUC of log base 10 of $o.OR_c$ over six HT-SELEX cycles.

Likewise, preferential binding score (PBS) was introduced to quantify each SNP allele preferential binding as difference of AUC between reference and alternative alleles in terms of logarithmic odds-ratio along HT-SELEX cycles. PBS was obtained by estimating the odds ratio of observing allele a at cycle c compared to cycle 0, where $Odd_{a,c}$ is the odds of observing allele a at cycle c given other alleles. Given the difference of reference and mutant alleles (r and m , respectively) logarithmic odds-ratio at cycle c , PBS was obtained by computing AUC of ΔLOR_c over six HT-SELEX cycles.

The statistical significance of both PBS and OBS in each experiment was measured by Monte-Carlo randomization, where the oligo and allele read counts were shuffled within each cycle and the scores were recomputed for 250,000 times. Oligos

were considered significantly bound to the TF for OBS p-value < 0.05. Oligos were considered significantly preferentially bound for SNPs for PBS p-value < 0.01 and OBS p-value < 0.05.

PWM Binding Preference Determination

Using JASPAR 2016 non-redundant vertebrate motif database⁵⁸, the score for reference and alternative genomic oligo sequences was measured for 208 distinct TFs using Biopython⁶². The PWM score of each sequence was obtained by computing the maximum motif score of a sliding window over sequence in both forward and reverse strand. Only oligos with at least one allele passed threshold false positive rate (FPR) 0.01 were considered for further analysis. Specifically, the FPR threshold was computed based on the human genome GC content (0.4) for each TF separately. The difference of PWM score between reference allele r and alternative allele a respectively (Δ PWM) was used to estimate allele preferential binding.

The PWM model allele preference was determined as follow: (i) preferred reference allele, when only PWM_r is above 0 or Δ PWM is above 2; (ii) preferred mutant allele, when only PWM_a is above 0 or Δ PWM is below -2; (iii) no allele preference otherwise.

Genotyping of HepG2 cells

Reads from whole genome sequencing (WGS) were aligned using BWA MEM⁵⁹ in pair-end model with default parameters. PCR duplicates were removed using Picard tools (<http://broadinstitute.github.io/picard>). Variants were then called according to the GATK best practice pipeline using GATK 3.6-0⁶⁰. Briefly, reads were realigned locally, and base pair qualities were recalibrated. Variants were then called using HaplotypeCaller with

default parameters. Variants were then recalibrated based on known gold standard variants. Only variants that passed filters were used in the downstream analysis.

ChIP-seq Data Analysis

Reads were aligned using BWA MEM⁵⁹ with either single-end or pair-end model to the hg19 reference genome. Reads with low mapping quality ($\text{mapq} < 10$) were filtered out, and PCR duplicates were removed using Picard tool (<http://broadinstitute.github.io/picard/>). MACS2⁶¹ were then applied to call peaks and generate signal tracks to view in the genome browser.

Determination of Allele Imbalance of TF binding from ChIP-seq data

In addition to ChIP-seq performed in this study, ChIP-seq for additional TFs were also collected from ENCODE project. For allelic analysis, reads were aligned using WASP mapping pipeline to control potential allelic mapping bias⁶². Specifically, heterozygous SNPs called using WGS data were used for HepG2 cells, and heterozygous SNPs from 1000 genome project were used for GM12878 cells. Allelic read counts for each phased heterozygous SNP within the 300bp window in TF ChIP-seq data and corresponding control data were obtained using custom python scripts. To remove sampling biases, SNPs that are covered by less than 20 reads in either the treatment or the control were filtered out. Odds ratios were then computed for each SNPs comparing allelic counts between the treatment and control to measure allelic imbalance. SNPs were tested for allelic imbalance using binomial test using background ratio derived from control data. SNPs with Benjamin-Hochberg adjusted p-value < 0.05 were considered as allelic imbalanced.

STARR-seq Data Analysis

STARR-seq reads were aligned to the oligo libraries using BWA⁵⁶ with default parameters. Read counts for each oligos were then counted using custom scripts. Counts for technical replicates were merged. Oligos were filtered to keep only oligos covered by more than 25 reads in the input library and more than five reads in at least three libraries.

We first identified oligo that were enriched compared to the input library. Enriched oligos were determined by a negative binomial regression from R package edgeR⁶³. Common biological dispersion was estimated using only yeast oligos where no real variation is expected. The resulting p-values were adjusted by Benjamin-Hochberg procedure, and the significance cutoff for enriched oligos was set to limit the rate of enriched yeast oligos to 5%.

We then focused on SNPs for which at least one allele were significantly enriched, and calculated the difference of log fold-change activity between two alleles using paired t-test from R package limma⁶⁴, shirking the variance with an empirical Bayesian method. The p-values were adjusted by Benjamin-Hochberg procedure and SNPs were considered significant with adjusted p-value < 0.01.

RNA-seq Data Analysis

Reads were aligned to the hg19 reference genome using STAR 2.4.2a⁶⁵ with default parameters in pair-end model. Only uniquely aligned reads were kept for further analysis. Cufflinks 2.2.1⁶⁶ was used to compute FPKM for each gene.

Determination of Allele Imbalance in Gene Expression

For allelic analysis, reads were aligned to the hg19 reference genome using STAR and WASP⁶² pipeline to control allelic mapping bias. The same set of SNPs and haplotypes were used for RNA-seq as ChIP-seq as described above in HepG2 cells.

Allelic counts for each gene were generated using htseq-count 0.6.0⁶⁷. Genes with at least 10 allelic reads were tested for allelic imbalance using the Binomial test using background ratio derived from whole genome sequencing data. Genes with Benjamin-Hochberg adjusted p-value < 0.1 were considered allelic imbalanced.

Enrichment of pbSNP in Allele Imbalanced Chromatin Features

To evaluate if pbSNPs were enriched for allele imbalanced chromatin features, fraction of pbSNP between imbalanced and balanced SNPs were compared using Fisher-test for allele imbalanced sites from CHIP-seq data (as described above), DHS¹⁵, H3K27ac¹⁶ and paSNPs.

Compare SNP-SELEX with DeltaSVM and DeepSEA

DeltaSVM¹⁷ and DeepSEA¹⁸ are two other computational tools used to predict the regulatory activity and chromatin effects of variants. Briefly, machine learning models were trained using predefined positive and negative sequences, where positive sequences are sequences overlapping with CHIP-seq or DHS peaks and negative sequences are random sampled genomic sequences without peaks. Then the models were applied to score genomic sequences surrounding SNPs of interest for both alleles. For DeepSEA, log₂ fold change between two alleles were used as score for SNPs. For DeltaSVM, weights for all 10mers surrounding the SNP were summed as the score for SNPs. SNPs with higher scores were more likely to affect regulatory activity.

We applied both DeltaSVM and DeepSEA models to measure chromatin effects of all 95,886 SNPs evaluated by SNP-SELEX experiments. DeltaSVM evaluated SNP scores using the weights obtained from GM12878, K562 and HepG2 DHS data²⁹. DeepSEA evaluated SNP log fold-change of DNase I sensitivity on AoAF, CD20+, Caco-

2, Fibrobl, HepG2, Hepatocytes, K562, Myometr, PANC-1, PanIsletD, and PanIslets cell lines.

Hi-C Data Analysis

Hi-C data was processed as previously described⁶⁸. Briefly, each end of read pairs were aligned separately using BWA MEM to the hg19 reference genome with default parameters. Chimeric read ends were further processed to keep only the five-prime alignment. Read ends with low mapping quality ($\text{mapq} < 10$) were removed, and remaining read ends were paired using custom scripts. PCR duplicates were removed using Picard tool (<http://broadinstitute.github.io/picard>). Aligned reads were further transformed to the juicer format and processed into hic format using juicebox tool⁶⁹. Chromatin loops were called using Hiccups with default parameters.

Haplotype Phasing of HepG2 cells using Hi-C and WGS

Aligned Hi-C bam files were processed through GATK realignment pipeline the same as WGS data describe above. Two filters were applied to SNPs to keep only high-quality SNPs: 1) Only bi-allelic SNPs were kept; 2) Only heterozygous SNPs with high genotype quality ($\text{GQ} > 20$) were kept. WGS and Hi-C data were then parsed to extract informative fragments with extractHAIRs⁷⁰ using filtered SNPs. The fragments from Hi-C and WGS data were combined, and HAPCUT2⁷⁰ was used to derive haplotypes. Results from HAPCUT2 were then paired with SNPs in 1000 Genome Project Phase 3 data, and Beagle 4.1⁷¹ was used to impute haplotypes for SNPs that were not phased by HAPCUT2. We obtained chromosome-span haplotypes for all auto chromosomes except for chr22. Phasing quality was further examined by computing fraction of homologous trans (h-trans) reads in RNA-seq data from HepG2 cells. Specifically, h-trans reads were

read pairs that contain SNPs from both haplotypes. Chromosome-span haplotypes with high accuracy were obtained.

Predicting target genes of non-coding SNPs

To assign potential target genes for SNPs, two approaches were taken: 1) SNPs within 2Kb upstream region of a TSS were assigned to the TSS; 2) SNPs overlapping one anchor of chromatin loops (with in 25Kb window) were assigned to the TSS overlapping the other anchor (with in 25Kb window). Similar approaches were used to connect TF binding sites to target genes.

Training and Validation of KEIS Models

TFs with at least 20 pbSNPs were selected to build KEIS models. For each TF, two separate models with different k-mer size were built for each experiment in each cycle using SELEX reads as positive sequences and other sequences in cycle 0 as negative sequences. Both positive and negative sequences were randomly down-sampled to 20,000 sequences. The models were trained using lsgkm with two k-mer size using parameters “-l 10 -k 6 -d 3” and “-l 8 -k 5 -d 3” respectively. The models were then used to score SNPs using deltasvm.pl script. For each SNPs, deltaSVM scores were computed using 40bp sequences with SNP at the center.

For each model, we measured the performance of the model based on SNPs tested directly by SNP-SELEX. Specifically, pbSNPs (p-value < 0.01) were treated as true positive sets and non-pbSNPs (p-value > 0.5) were used as true negative sets. We computed AUROC and AUPR for each model using R package PPROC. For each TF, we select the best model based on AUPR for genome-wide prediction.

Five-fold cross validation was also performed to measure the performance of KEIS models. Specifically, SNPs for benchmarking were divided equally into five folds. Sequences containing SNPs in the one-fold used to test were removed when training the models, and the remaining four-fold of SNPs were used for testing.

Genome-wide Prediction using KEIS Models

For each SNP to test, a pair of 40bp genomic sequences from the hg19 reference genome with the SNP to test in the center were generated. We first scored both sequences using gkm models and determined if at least one of oligos can be bound by the TF. Threshold was determined based on bound oligos identified using SNP-SELEX experiments. Specifically, we computed gkm scores for all bound oligos and used the medium of the scores for bound oligos for each TF as the threshold to determine binding. Only bound oligos were further predicted for allelic TF binding.

The pair of sequences were then scored using deltasvm script. Similarly, threshold was determined based on pbSNPs measured by SNP-SELEX. Specifically, we computed KEIS scores for all pbSNPs and used the medium of pbSNPs' scores for each TF as the threshold to determine allelic TF binding.

Performance Comparison of PWM and KEIS

For PWM models, delta PWM scores were calculated as described in the previous section for all TF-SNP pairs. PWM models were applied to the same set of SNPs as KEIS models. Only 125 TFs for which models for both PWM models and KEIS models are available were included in the comparison.

Comparison with TF ChIP-seq Data

We made predictions for heterozygous SNPs covered by at least 20 allelic reads in ChIP-seq experiments in HepG2 and GM12878 cells respectively. For each TF ChIP-seq experiment, we computed the percentage of allelic imbalanced SNP in k-SNPs and non-k-SNPs. Allelic imbalanced SNPs were determined as described in the previous section.

For PWM models, we determined threshold for oligo binding and k-SNPs using the exact procedure as KEIS models. For ATF2 data, we used ATF3 motif because there is no motif for ATF2 in JASPAR database.

Partition Heritability of Complex Traits using k-SNPs

To partition heritability of diseases and complex traits in context of k-SNPs, we applied previously established method LDSC²⁷. Briefly, LDSC models the casual effect of each SNP for a given trait as a linear additive contribution by a list of annotations and then estimates per-SNP heritability for each annotation as regression coefficient considering not only the SNP to test but also all SNPs in LD. Then p-value was computed to test if regression coefficient for annotation i is positive, which means annotation i explains additional heritability in addition to other annotations.

We made predictions for 124 TFs with good KEIS models for all common SNPs in 1000 genome project phase 3 for European population as mentioned above. The list of SNPs was downloaded from ldsc website (<https://data.broadinstitute.org/alkesgroup/LDSCORE/>). K-SNP predictions for each TF within DHS regions in any Roadmap tissues were then used as annotation to estimate annotation-specific LD scores for each TF. We then run LDSC using k-SNPs for each TF along with 53 baseline models including genic regions, enhancer regions and conserved

regions. To rule out the effect of TF binding rather than allelic TF binding, we also included predictions for SNPs bound by the TF in the regression model. In summary, we run LDSC using 55 annotations including k-SNP prediction, binding SNP prediction, and 53 baseline models, and p values for regression coefficient of k-SNP prediction for each TF were used to measure if k-SNP explains additional heritability. The p-values for the term of binding SNP prediction were used in **Figure S1.6**.

Differentially Gene Expression Analysis

Read counts for each gene were obtained using htseq-count using GENCODE human annotation release 24 as reference. DESeq2⁷² was used to identify differentially expressed genes using default parameters. Genes with Benjamin-Hochberg adjusted p-value < 0.2 were considered as differentially expressed. KEGG pathway enrichment analysis was performed with DAVID.

1.6 Figures

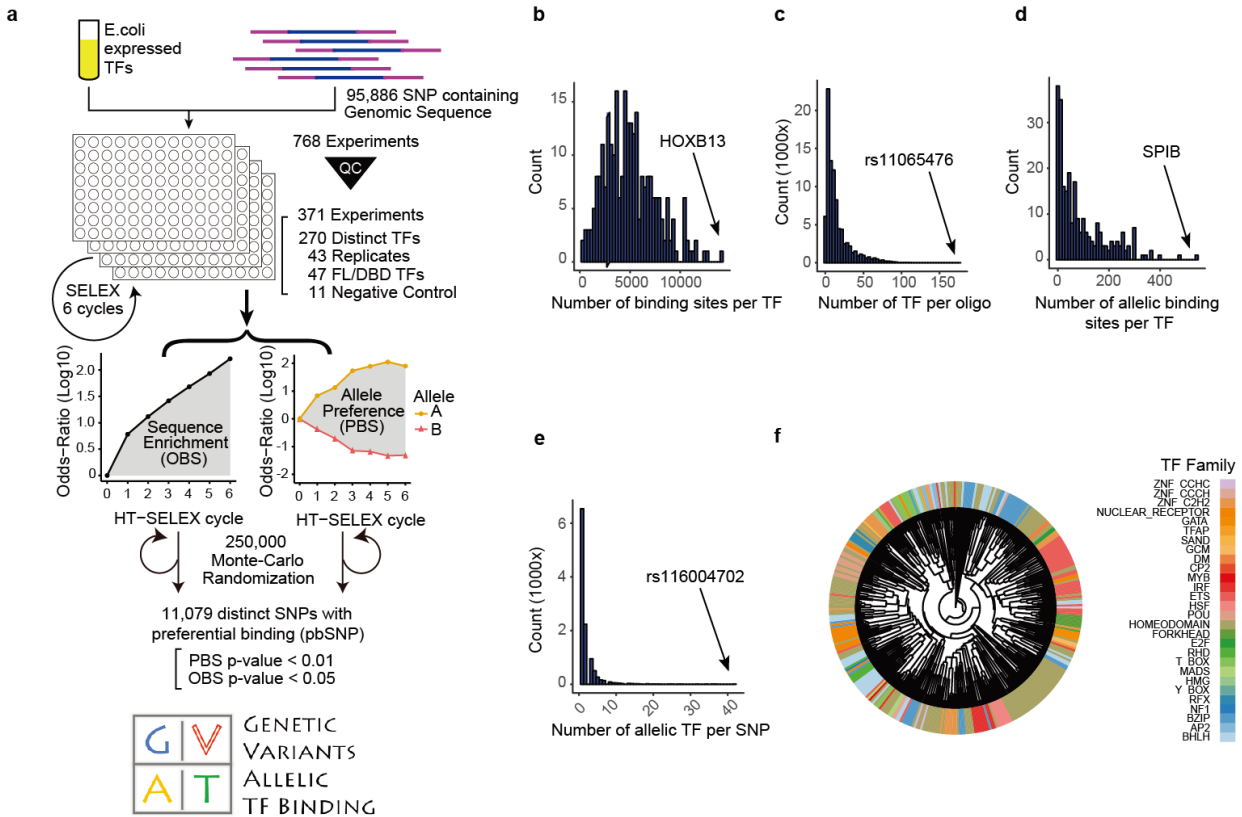
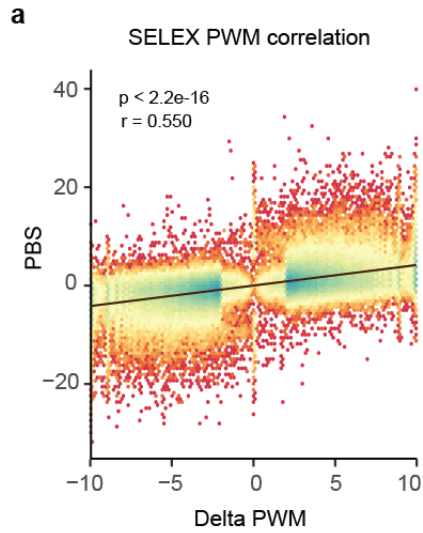


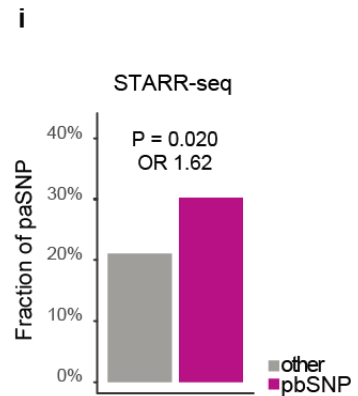
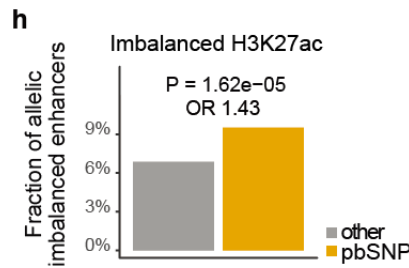
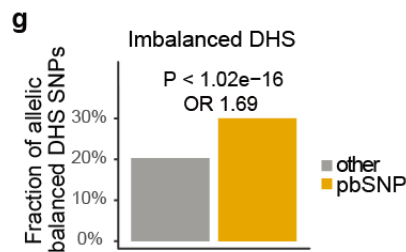
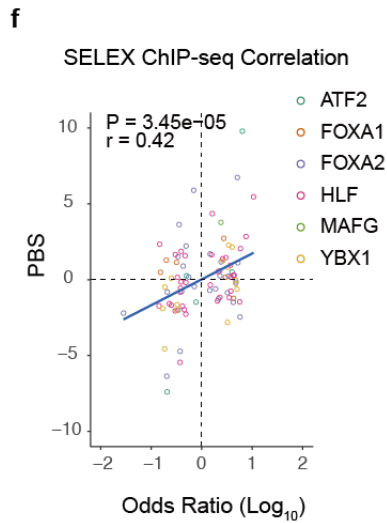
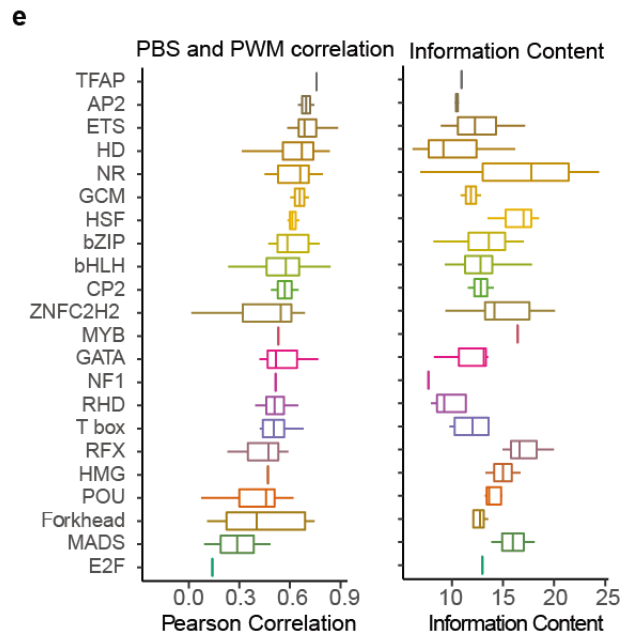
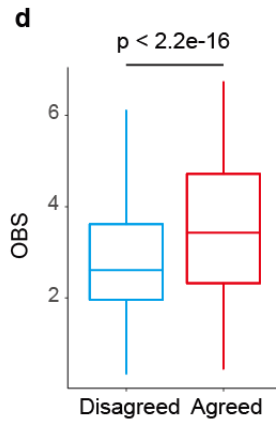
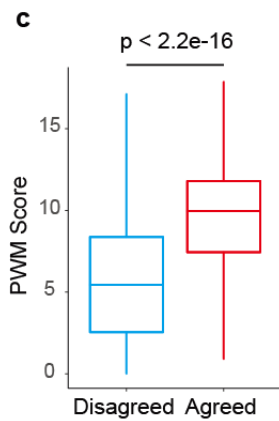
Figure 1.1. Determination of differential DNA binding of human TFs to common sequence variants by SNP-SELEX. (a) An overview of the SNP-SELEX procedure. (b-e) Histograms show the number of oligos bound by each TF (b), the number of binding TFs for each oligo (c), the number of pbSNPs bound by each TF (d), and the number of TFs showing allelic binding for each pbSNP (e). (f) A clustering diagram of TFs tested in this study was generated based on the pairwise correlation of their DNA binding specificity from the SNP-SELEX data.

Figure 1.2. Comparison of differentially bound SNPs identified by SNP-SELEX and PWM models. (a) A scatterplot shows the preferential binding score (PBS) from SNP-SELEX experiments on the y-axis and differential PWM scores (Δ PWM) on the x-axis. (b) Comparison of the SNPs with differential TF binding determined by SNP-SELEX and PWM. (c-d) Comparison of the PWM scores (b) and the OBS scores (c) between SNPs with consistent (Agreed) and inconsistent (Disagreed) prediction. (e) Boxplot showing Pearson correlation coefficients of PBS and Δ PWM (left) and information content (right) for each TF family. (f) Scatterplot showing the correlation of allelic biases of DNA binding detected from ChIP-seq data in HepG2 cells and those predicted by PBS and SNP-SELEX, respectively. (g) Bar plot showing the comparison of the fraction of allelic imbalanced DHS sites in pbSNPs and non-pbSNPs. (h) Bar plot comparing the fraction of enhancers showing allelic imbalance in H3K27ac with regard to pbSNPs and non-pbSNPs. (i) Bar plot comparing the fractions of paSNPs determined using STARR-seq in pbSNPs and non-pbSNPs.



b Comparison Between SELEX and PWM

PWM vs SELEX	No. of TF-SNPs Pairs
Agreed	634,527
PWM + / SELEX -	153,411
PWM - / SELEX +	2,568
Contradictory	185



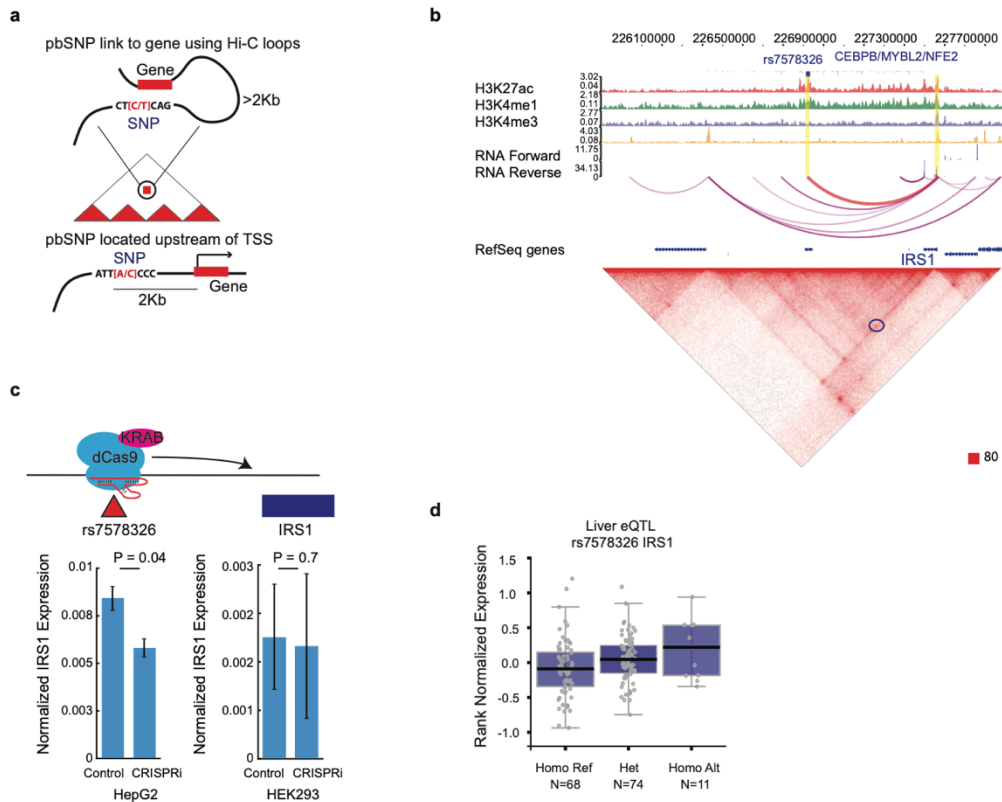
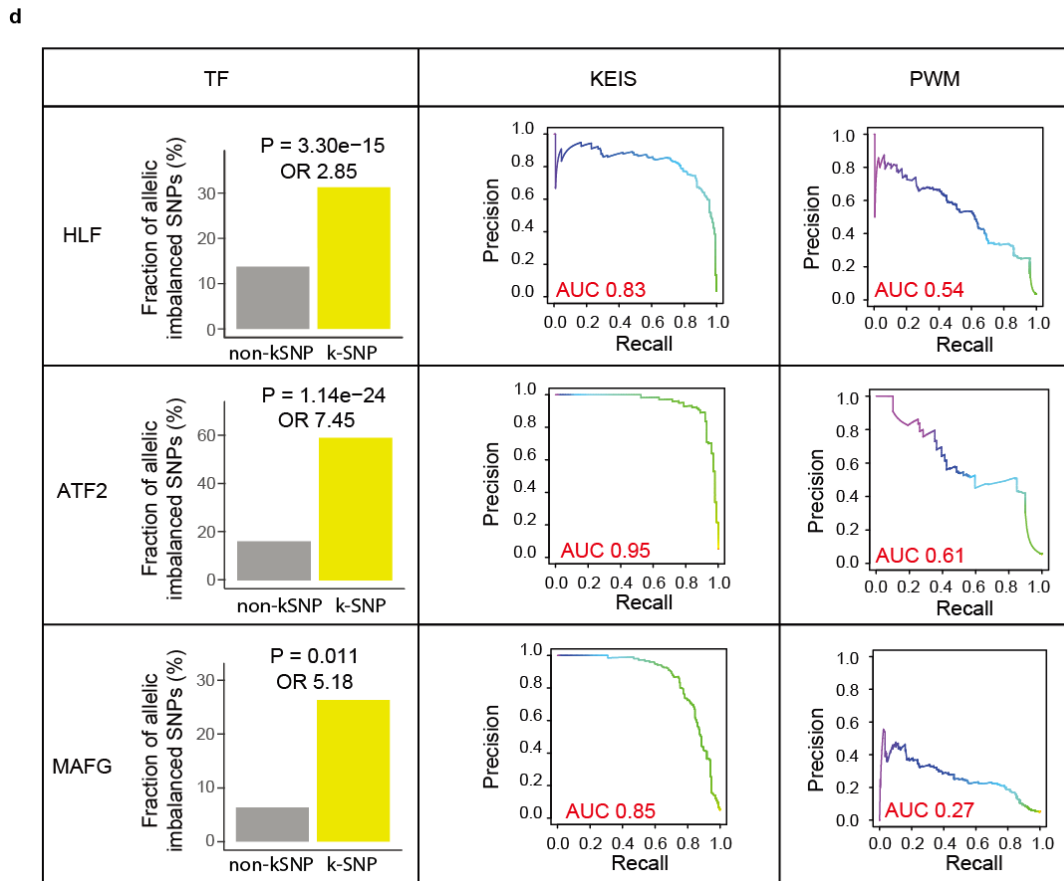
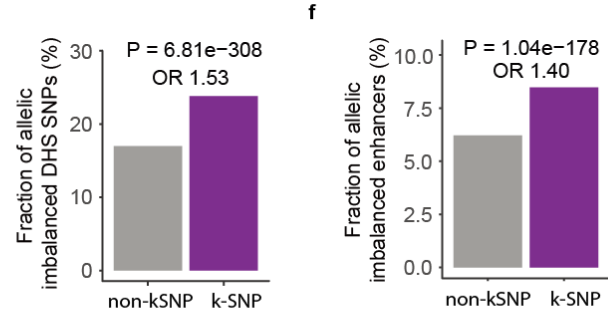
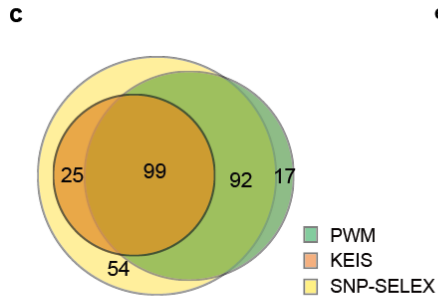
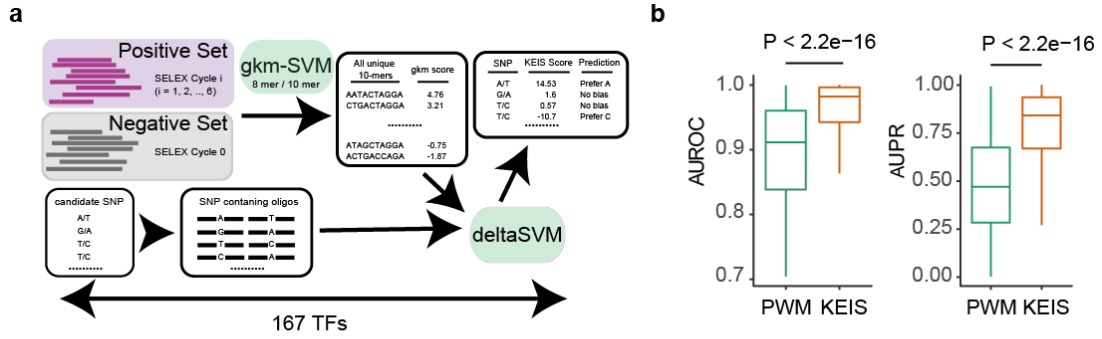


Figure 1.3. pbSNPs uncover potential mode of action for likely T2D causal variants. (a) Semantic plot showing two approaches to link SNPs to target genes: (top) chromatin loops where DNA forms a loop and brings the SNP in proximity to TSS, which can be identified as a red dot in the Hi-C contact map (dot circled in black); (bottom) A SNP is located within 2Kb upstream of TSS. (b) A T2D GWAS leading SNP rs7578326, a pbSNP differentially bound by TFs CEBPB, CEBPE, MYBL2 and NFE2, is predicted to target the IRS1 gene based on Hi-C analysis (circled in blue in bottom panel) in HepG2 cells. (c) CRISPRi using dCas9 fused with repressive KRAB domain and guide RNA targeting the locus of SNP rs7578326 (upper) leads to reduced expression of IRS1 gene in HepG2 but not in HEK293T cells. (d) SNP rs7578326 is an eQTL in the liver. Normalized expression value in liver from 153 individuals from GTEx project for IRS1 gene is grouped based on individual's genotype of SNP rs7578326.

Figure 1.4. KEIS better predicts differential TF binding to non-coding variants in vitro and in vivo than PWM. (a) A schematic graph for development of KEIS models for the TFs. (b) Boxplot showing the comparison of performance for PWM and KEIS for 167 TFs with KEIS models based on AUROC and AUPR in five-fold cross-validation. AUROC, area under the receiver operating characteristic curve. AUPR, area under the precision-recall curve. (c) Venn diagram showing the number of TFs with DNA binding specificities defined by PWM, KEIS, and SNP-SELEX, respectively. (d) KEIS outperforms PWM models in predicting differential DNA binding in vivo. (e) Bar plot showing the comparison of the fraction of allelic imbalanced DHS sites for k-SNPs and non-kSNPs. (f) Bar plot showing the comparison of the fraction of allelic imbalanced enhancers measured by H3K27ac around k-SNPs and non-kSNPs.



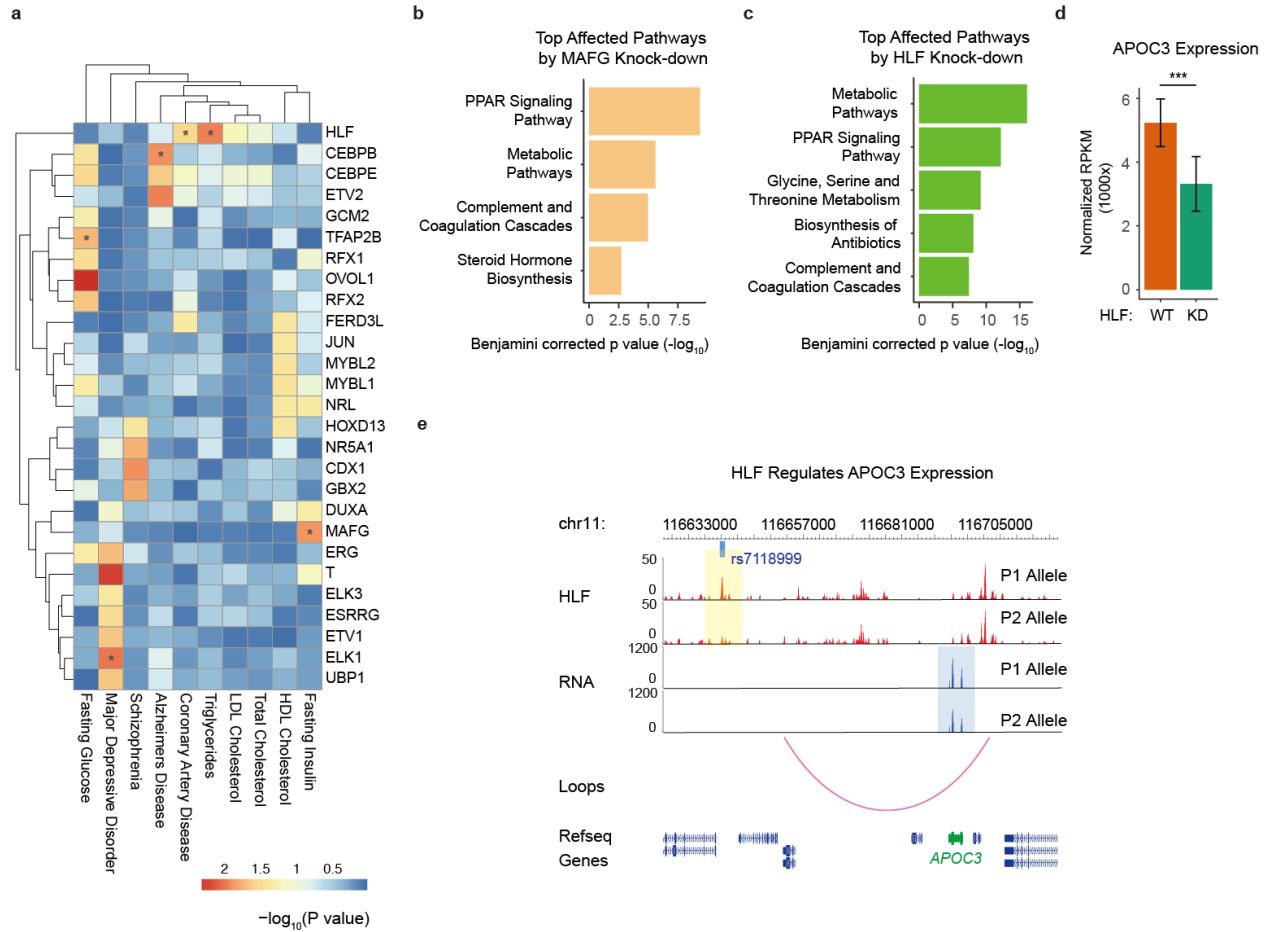
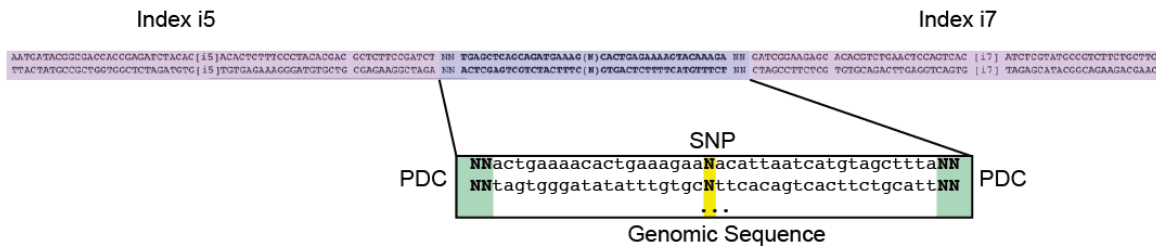


Figure 1.5. KEIS models identify candidate master TFs involved in complex traits and diseases. a) Heatmap showing the significance of enrichment of SNPs with differential binding to TFs among traits- or disease-associated SNP. Only TFs showing significant enrichment ratio in at least one trait are shown in the figure for clarity. (b, c) Bar plot showing enriched KEGG pathways affected by MAFG (b) and HLF (c) knockdown in HepG2 cells. (d) Bar plot showing normalized gene expression for APOC3 in HLF KO and WT HepG2 cells. P-value is 6.67×10^{-5} as computed by DESeq2. (e) Genome browser shot showing differential HLF binding to rs7118999 is linked to allelic gene expression of APOC3, which is predicted to be targeted by the SNP based on chromatin looping in HepG2.

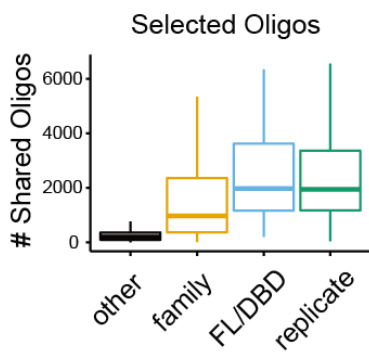
1.7 Supplemental Figures

Figure S1.1. Quality controls and reproducibility of SNP-SELEX data. (a) Two random nucleotides were added to each end of the oligos as unique molecule identifiers (UMIs) to remove over-represented PCR duplicates. Illumina TruSeq dual-index system was adapted for oligo design. (b-c) Comparison of oligo selection (b) and allele preference (c) between different biological replicates (replicates), of full length (FL) and DNA Binding Domain (DBD), members of the same structural family (family), and random pairs (others). (d) An example illustrating differential DNA binding at six SNPs, in four SNP-SELEX experiments, including (i) two full-length ELK1 replicates, on the first two lines; (ii) one DNA binding domain (DBD) ELK1, on the third line; and one full-length ELK4 TF which belongs to the same structure family, on the last line.

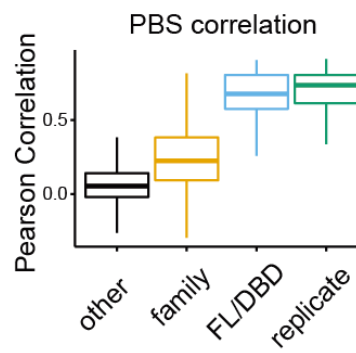
a



b



c



d

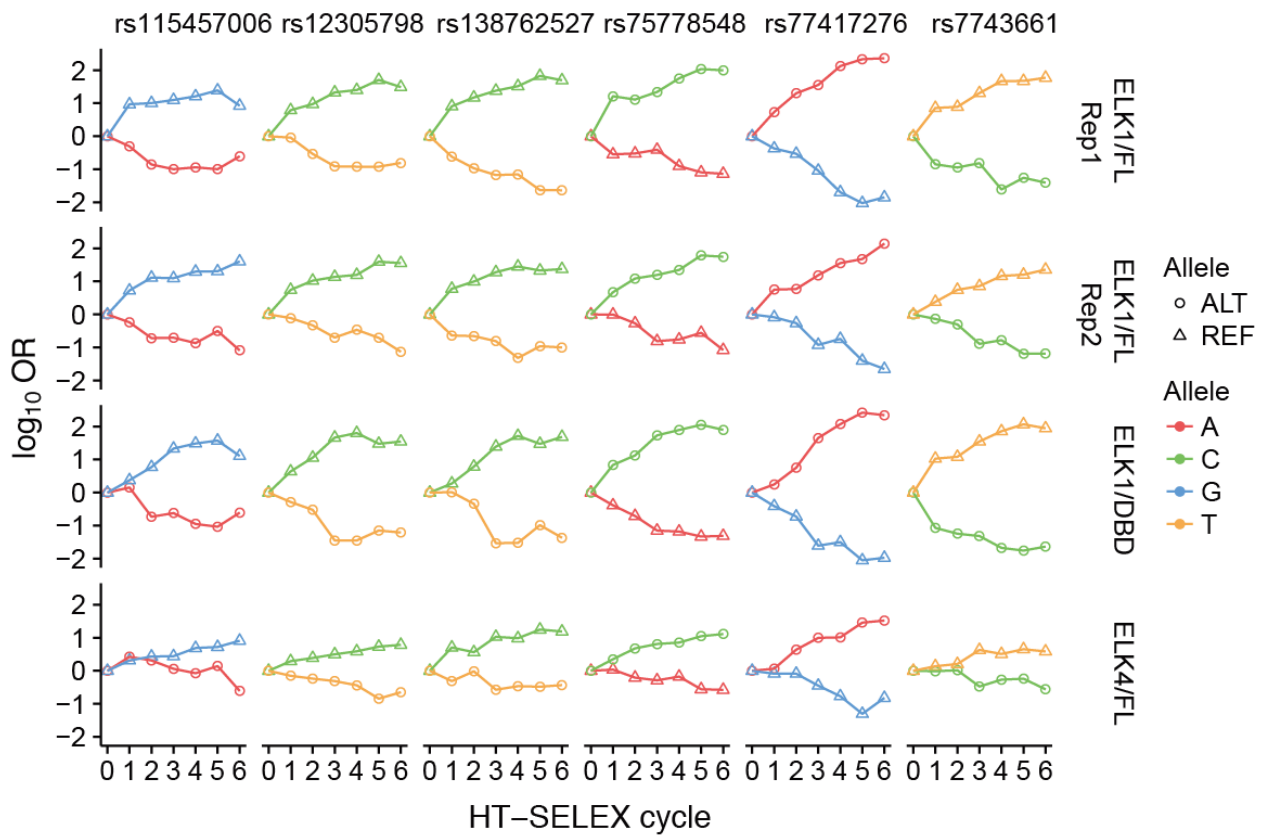


Figure S1.2. SNP-SELEX results are correlated with TF binding in vivo and enhancer activity from high through reporter assays. (a, c) Scatterplot showing the correlation of delta PWM score and allelic binding ratio and ChIP-seq in HepG2 (a) and GM12878 (c) cells respectively. (b) Scatterplot showing correlation of PBS and allelic binding ratio derived from SNP-SELEX and ChIP-seq in GM12878 cells respectively. (d) Bar plot showing the comparison of the fraction of allelic imbalanced DHS sites in pbSNPs and non-pbSNPs predicted by PWM models instead of SNP-SELEX. (e) Bar plot comparing the fraction of enhancers showing allelic biased in H3K27ac with regard to pbSNPs and non-pbSNPs predicted by PWM models instead of SNP-SELEX. (f-g) Predictions of differential TF binding using deltaSVM (d) and DeepSea (e) are well correlated with SNP-SELEX results.

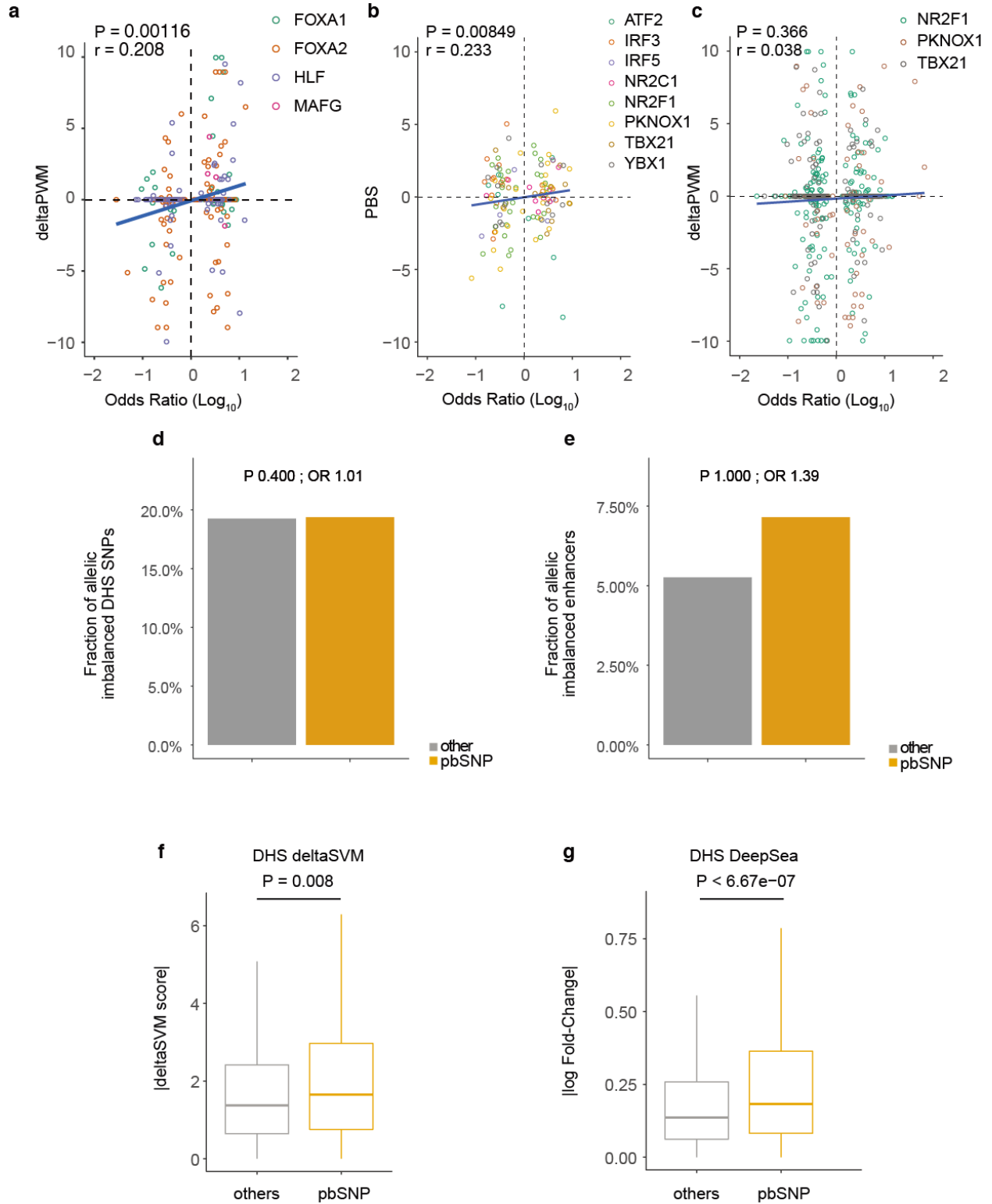
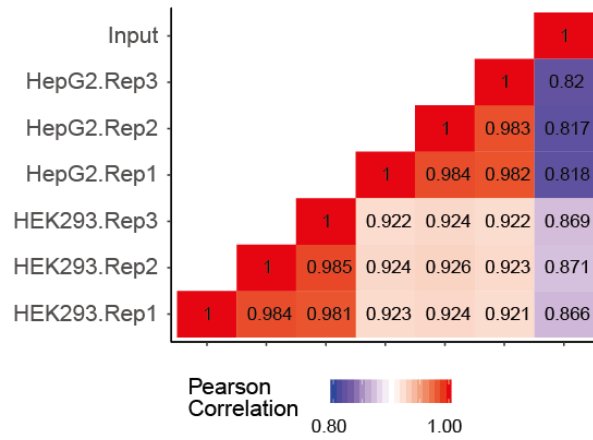
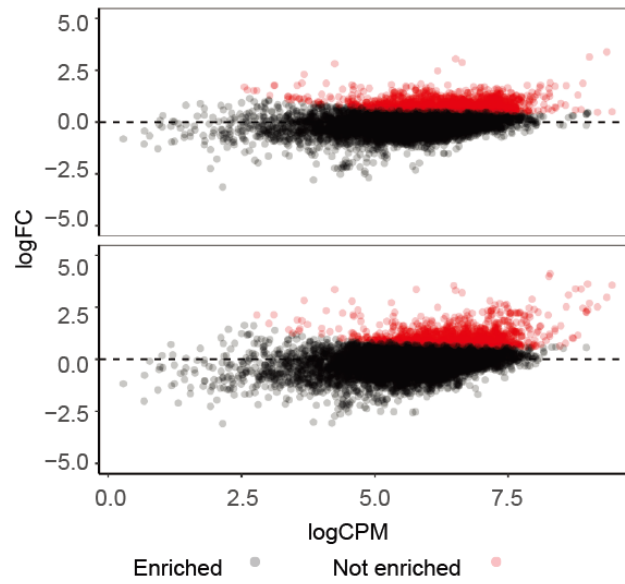


Figure S1.3. SNP-SELEX results are correlated with enhancer activity from high through reporter assays. (a) Heatmap of Pearson's correlation coefficient calculated among STARR-seq read counts in the input library, three HepG2 replicates, and three HEK293T replicates. (b) An illustration of the oligo logarithmic fold-change (y-axis) over the library logarithmic counts per million (CPM) (x-axis) for HEK293T, on the top panel, and HepG2, on the bottom panel. (c) Bar plot comparing the fractions of paSNPs determined using STARR-seq in pbSNPs and non-pbSNPs predicted by PWM models instead of SNP-SELEX. SNPs with absolute PWM changes larger than three were considered as pbSNPs.

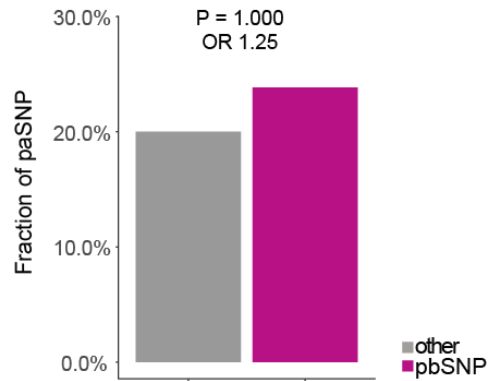
a



b



c



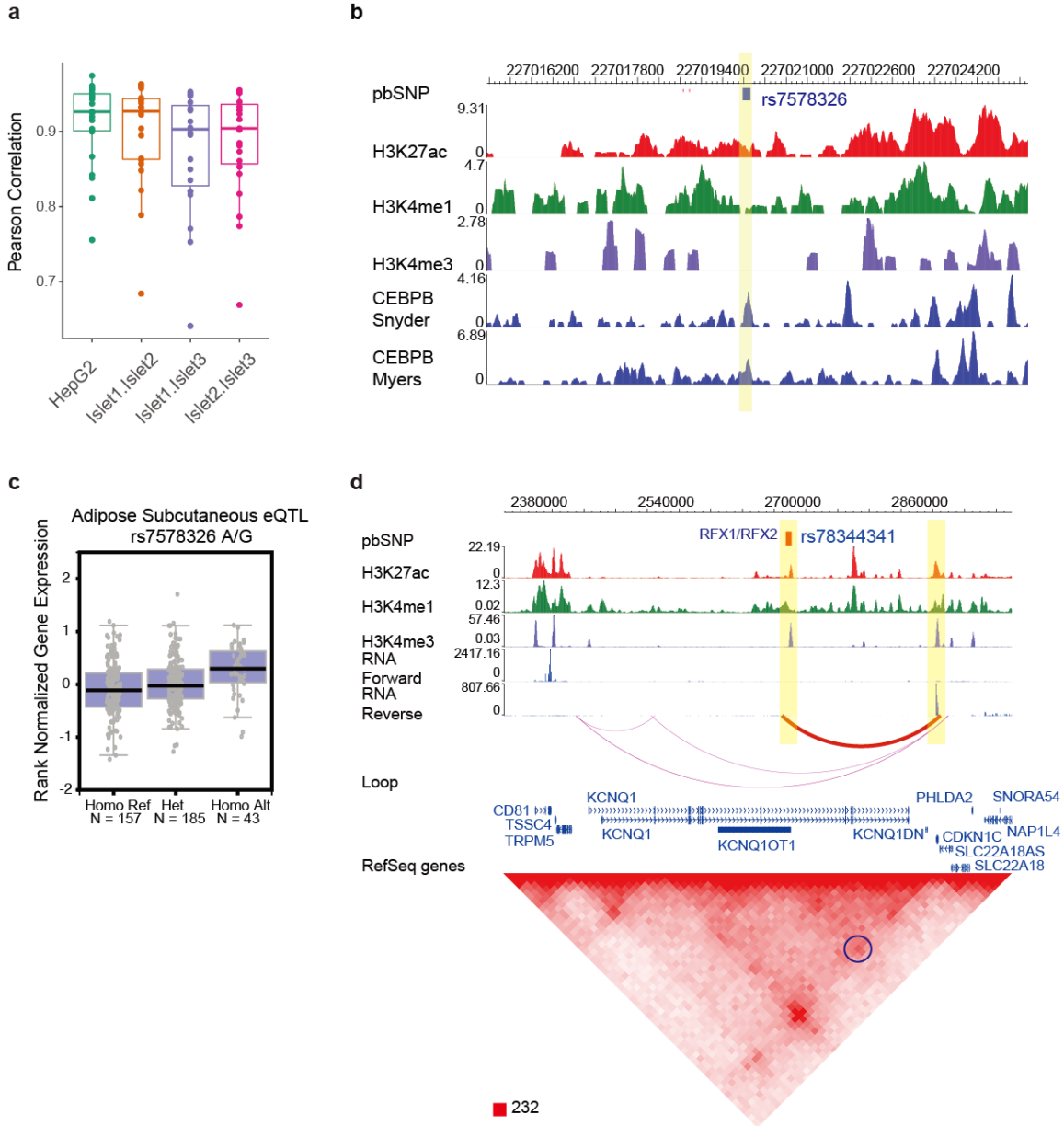


Figure S1.4. Exploring the mode of action of likely T2D causal SNPs with the help of SNP-SELEX and Hi-C. (a) Contact matrix for the pair of three human islets and two replicates for HepG2 cells are highly reproducible (see Methods for details). (b) Browser view of CEBPB binding at rs7578326, which is located in a CEBPB binding site. Two independent ChIP-seq data from ENCODE were shown. (c) SNP rs7578326 is an eQTL in adipose tissues. Normalized expression value in adipose tissues from 385 individuals from GTEx project for IRS1 gene is grouped based on individual's genotype of SNP rs7578326. Linear regression p-value and effect size are noted on the top. (d) A T2D GWAS leading SNP rs231361 differentially bound by RFX1, and RFX2 is predicted to regulate SLC22A18 and CDKN1C gene based on chromatin loops (circled in blue in bottom panel) in islets.

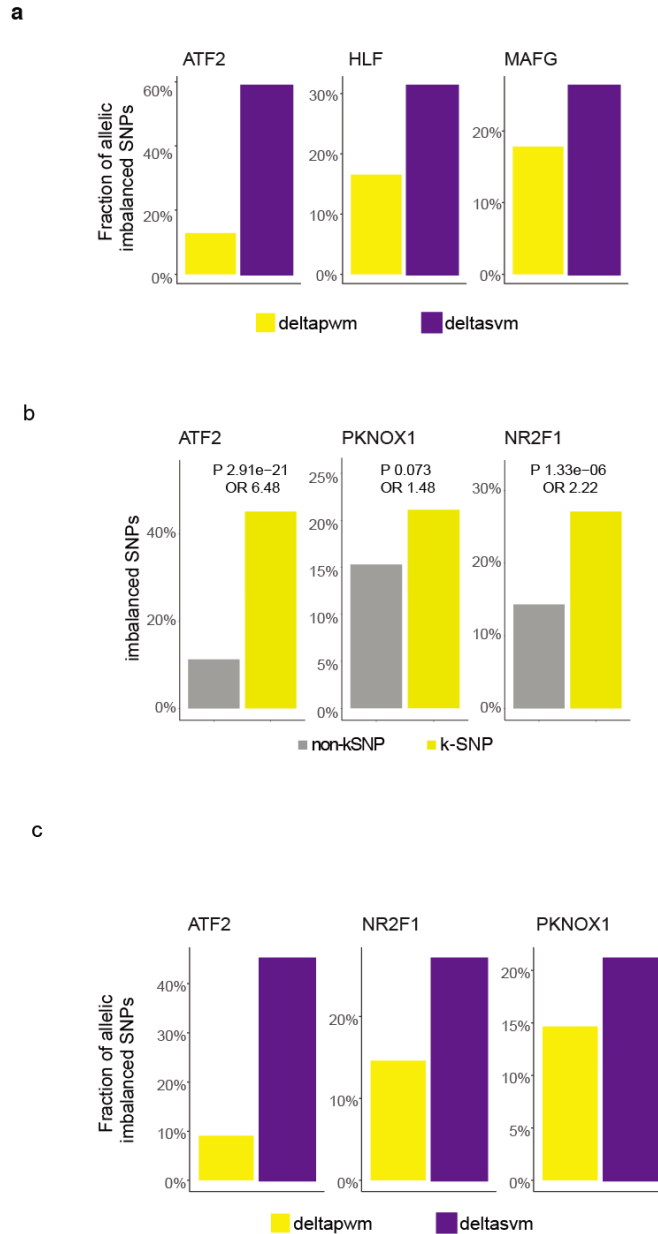


Figure S1.5. KEIS more accurately predicts non-coding variants affecting TF binding in vivo than PWM. (a) Comparison of the fraction of allelic imbalanced SNPs measured by TF ChIP-seq in pbSNPs predicted by KEIS model and PWM models in HepG2 cells. (b) Bar plot showing the comparison of the allelic imbalance of SNPs as measured by TF ChIP-seq for k-SNPs and non-k-SNPs for KEIS models in GM12878 cells, where k-SNPs are SNPs are predicted to affect TF binding by KEIS models.(c) Comparison of the fraction of allelic imbalance of SNPs as measured by TF ChIP-seq in pbSNPs and predicted by KEIS model and PWM models in GM12878 cells.

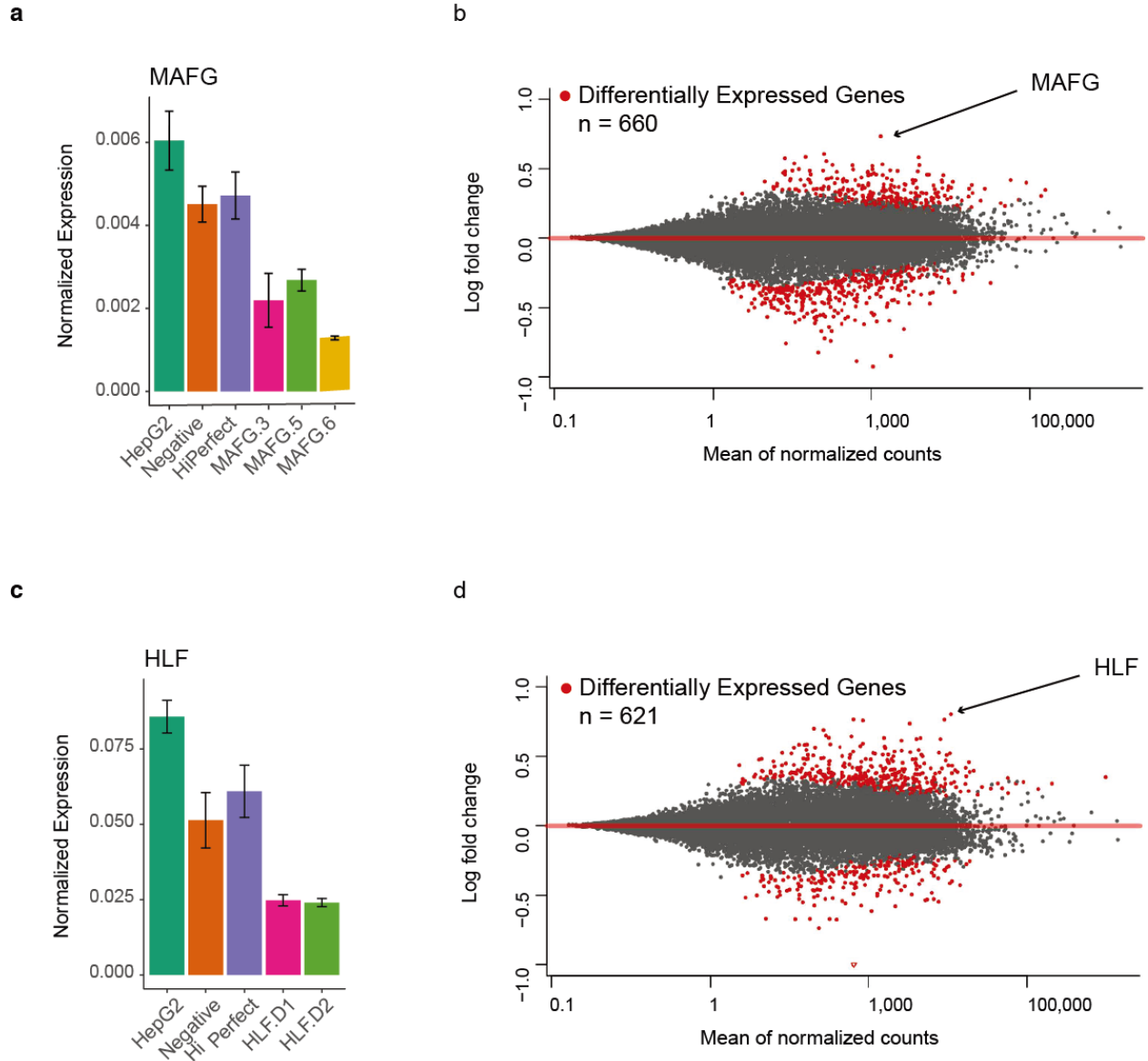


Figure S1.6. Analysis of differentially expressed genes upon knockdown of HLF and MAFG in HepG2 cells. (a, c) qPCR results of MAFG (a) and HLF (c) in WT (HepG2), Control (Negative and HiPerfect), and cells treated with different siRNAs. (b, d) MA-plot showing differentially expressed genes comparing MAFG knockdown (b) and HLF knockdown (d) versus controls. Significant differentially expressed genes (FDR<0.2) were marked in red.

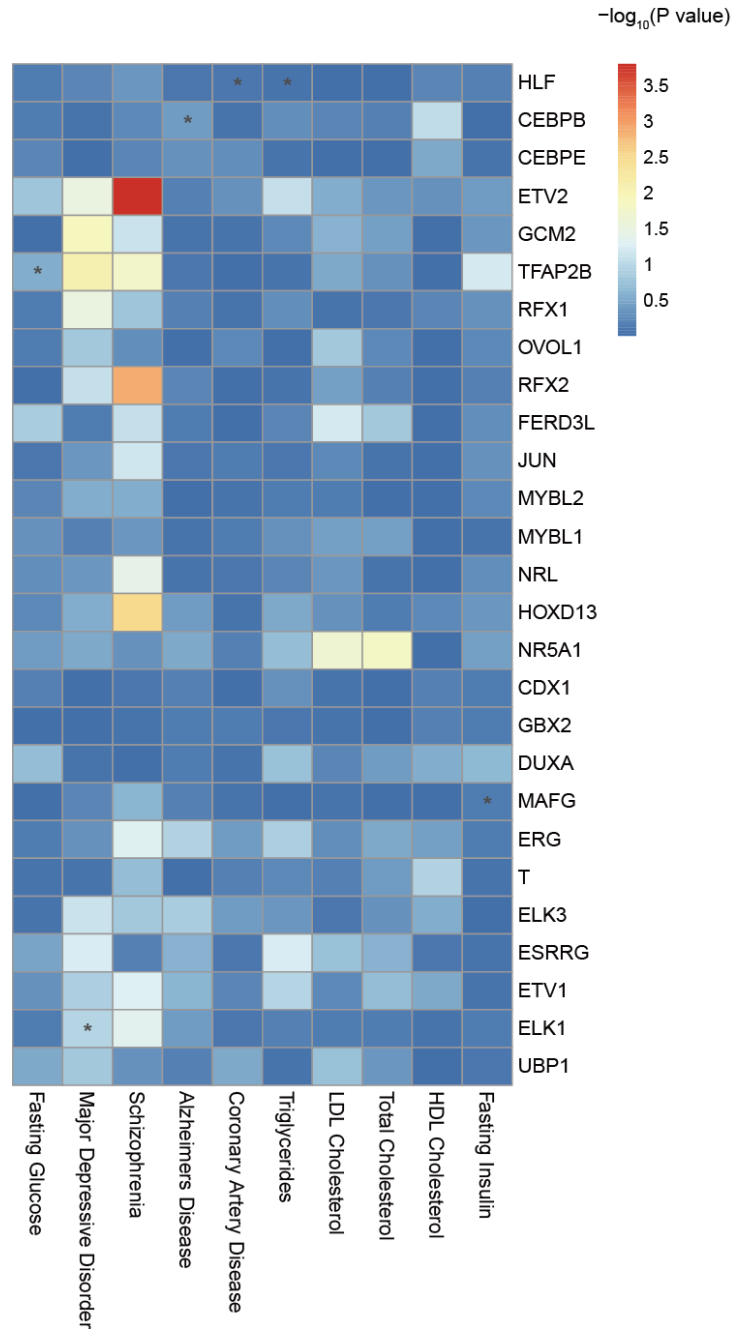


Figure S1.7. KEIS models help identify master TFs involved in complex traits and diseases. Heatmap showing the significance of enrichment of TFs showing differential DNA binding to traits- or disease-associated SNP. The color key is shown, and the value represents the -log₁₀ p-value. TF-trait pairs mentioned in the text were marked with *. Note that the master regulator we observed and validated (Figure 1.5a) could not be identified here if we only use the presence of SNPs at the binding sites without taking into account the impact of SNP on binding affinity.

1.8 Acknowledgments

Chapter 1, in full, is a manuscript submitted as “Systematic analysis of transcription factor binding to non-coding variants in the human genome”. Jian Yan, Yunjiang Qiu, André M Ribeiro dos Santos, Yimeng Yin, Yang E. Li, Nick Vinckier, Naoki Nariai, Anugraha Raman, Zhe Liu, Joshua Chiou, Kelly A. Frazer, Kyle J. Gaulton, Maïke Sander, Jussi Taipale, and Bing Ren. The dissertation author was the primary investigator and author of this paper.

1.9 References

1. MacArthur, J., Bowler, E., Cerezo, M., Gil, L., Hall, P., Hastings, E., Junkins, H., McMahon, A., Milano, A., Morales, J., Pendlington, Z. M., Welter, D., Burdett, T., Hindorff, L., Flicek, P., Cunningham, F. & Parkinson, H. The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). *Nucleic Acids Research* 45, D896–D901 (2017).
2. Visscher, P. M., Wray, N. R., Zhang, Q., Sklar, P., McCarthy, M. I., Brown, M. A. & Yang, J. 10 Years of GWAS Discovery: Biology, Function, and Translation. *Am. J. Hum. Genet.* 101, 5–22 (2017).
3. Altshuler, D., Daly, M. J. & Lander, E. S. Genetic mapping in human disease. *Science* 322, 881–888 (2008).
4. Mahajan, A., Taliun, D., Thurner, M., Robertson, N. R., Torres, J. M., Rayner, N. W., Payne, A. J., Steinthorsdottir, V., Scott, R. A., Grarup, N., Cook, J. P., Schmidt, E. M., Wuttke, M., Sarnowski, C., Mägi, R., Nano, J., Gieger, C., Trompet, S., Lecoeur, C., Preuss, M. H., Prins, B. P., Guo, X., Bielak, L. F., Below, J. E., Bowden, D. W., Chambers, J. C., Kim, Y. J., Ng, M. C. Y., Petty, L. E., Sim, X., Zhang, W., Bennett, A. J., Bork-Jensen, J., Brummett, C. M., Canouil, M., Ec Kardt, K.-U., Fischer, K., Kardia, S. L. R., Kronenberg, F., Läll, K., Liu, C.-T., Locke, A. E., Luan, J., Ntalla, I., Nylander, V., Schönherr, S., Schurmann, C., Yengo, L., Bottinger, E. P., Brandslund, I., Christensen, C., Dedoussis, G., Florez, J. C., Ford, I., Franco, O. H., Frayling, T. M., Giedraitis, V., Hackinger, S., Hattersley, A. T., Herder, C., Ikram, M. A., Ingelsson, M., Jørgensen, M. E., Jørgensen, T., Kriebel, J., Kuusisto, J., Ligthart, S., Lindgren, C. M., Linneberg, A., Lyssenko, V., Mamakou, V., Meitinger, T., Mohlke, K. L., Morris, A. D., Nadkarni, G., Pankow, J. S., Peters, A., Sattar, N., Stančáková, A., Strauch, K., Taylor, K. D., Thorand, B., Thorleifsson, G., Thorsteinsdottir, U., Tuomilehto, J., Witte, D. R., Dupuis, J., Peyser, P. A., Zeggini, E., Loos, R. J. F., Froguel, P., Ingelsson, E., Lind, L., Groop, L., Laakso, M., Collins, F. S., Jukema, J. W., Palmer, C. N. A., Grallert, H., Metspalu, A., Dehghan, A., Köttgen, A., Abecasis, G. R., Meigs, J. B., Rotter, J. I., Marchini, J., Pedersen, O., Hansen, T., Langenberg, C., Wareham, N. J., Stefansson, K., Gloyn, A. L., Morris, A. P., Boehnke, M. & McCarthy, M. I. Fine-mapping type 2 diabetes loci to single-variant resolution using high-density imputation and islet-specific epigenome maps. *Nat. Genet.* 50, 1505–1513 (2018).
5. DIAbetes Genetics Replication And Meta-analysis (DIAGRAM) Consortium, Asian Genetic Epidemiology Network Type 2 Diabetes (AGEN-T2D) Consortium, South Asian Type 2 Diabetes (SAT2D) Consortium, Mexican American Type 2 Diabetes (MAT2D) Consortium, Type 2 Diabetes Genetic Exploration by Nex-generation sequencing in muylti-Ethnic Samples (T2D-GENES) Consortium, Mahajan, A., Go, M. J., Zhang, W., Below, J. E., Gaulton, K. J., Ferreira, T., Horikoshi, M., Johnson, A. D., Ng, M. C. Y., Prokopenko, I., Saleheen, D., Wang, X., Zeggini, E., Abecasis, G. R., Adair, L. S., Almgren, P., Atalay, M., Aung, T., Baldassarre, D., Balkau, B., Bao, Y., Barnett, A. H., Barroso, I., Basit, A., Been, L. F., Beilby, J., Bell, G. I.,

Benediktsson, R., Bergman, R. N., Boehm, B. O., Boerwinkle, E., Bonnycastle, L. L., Burt, N., Cai, Q., Campbell, H., Carey, J., Cauchi, S., Caulfield, M., Chan, J. C. N., Chang, L.-C., Chang, T.-J., Chang, Y.-C., Charpentier, G., Chen, C.-H., Chen, H., Chen, Y.-T., Chia, K. S., Chidambaram, M., Chines, P. S., Cho, N. H., Cho, Y. M., Chuang, L.-M., Collins, F. S., Cornelis, M. C., Couper, D. J., Crenshaw, A. T., van Dam, R. M., Danesh, J., Das, D., de Faire, U., Dedoussis, G., Deloukas, P., Dimas, A. S., Dina, C., Doney, A. S., Donnelly, P. J., Dorkhan, M., van Duijn, C., Dupuis, J., Edkins, S., Elliott, P., Emilsson, V., Erbel, R., Eriksson, J. G., Escobedo, J., Esko, T., Eury, E., Florez, J. C., Fontanillas, P., Forouhi, N. G., Forsen, T., Fox, C., Fraser, R. M., Frayling, T. M., Froguel, P., Frossard, P., Gao, Y., Gertow, K., Gieger, C., Gigante, B., Grallert, H., Grant, G. B., Grrop, L. C., Groves, C. J., Grundberg, E., Guiducci, C., Hamsten, A., Han, B.-G., Hara, K., Hassanali, N., Hattersley, A. T., Hayward, C., Hedman, A. K., Herder, C., Hofman, A., Holmen, O. L., Hovingh, K., Hreidarsson, A. B., Hu, C., Hu, F. B., Hui, J., Humphries, S. E., Hunt, S. E., Hunter, D. J., Hveem, K., Hydrie, Z. I., Ikegami, H., Illig, T., Ingelsson, E., Islam, M., Isomaa, B., Jackson, A. U., Jafar, T., James, A., Jia, W., Jöckel, K.-H., Jonsson, A., Jowett, J. B. M., Kadowaki, T., Kang, H. M., Kanoni, S., Kao, W.-H. L., Kathiresan, S., Kato, N., Katulanda, P., Keinänen-Kiukaanniemi, K. M., Kelly, A. M., Khan, H., Khaw, K.-T., Khor, C.-C., Kim, H.-L., Kim, S., Kim, Y. J., Kinnunen, L., Klopp, N., Kong, A., Korpi-Hyövälti, E., Kowlessur, S., Kraft, P., Kravic, J., Kristensen, M. M., Krithika, S., Kumar, A., Kumate, J., Kuusisto, J., Kwak, S.-H., Laakso, M., Lagou, V., Lakka, T. A., Langenberg, C., Langford, C., Lawrence, R., Leander, K., Lee, J.-M., Lee, N. R., Li, M., Li, X., Li, Y., Liang, J., Liju, S., Lim, W. Y., Lind, L., Lindgren, C. M., Lindholm, E., Liu, C.-T., Liu, J. J., Lobbens, S., Long, J., Loos, R. J. F., Lu, W., Luan, J., Lyssenko, V., Ma, R. C. W., Maeda, S., Mägi, R., Männistö, S., Matthews, D. R., Meigs, J. B., Melander, O., Metspalu, A., Meyer, J., Mirza, G., Mihailov, E., Moebus, S., Mohan, V., Mohlke, K. L., Morris, A. D., Mühleisen, T. W., Müller-Nurasyid, M., Musk, B., Nakamura, J., Nakashima, E., Navarro, P., Ng, P.-K., Nica, A. C., Nilsson, P. M., Njølstad, I., Nöthen, M. M., Ohnaka, K., Ong, T. H., Owen, K. R., Palmer, C. N. A., Pankow, J. S., Park, K. S., Parkin, M., Pechlivanis, S., Pedersen, N. L., Peltonen, L., Perry, J. R. B., Peters, A., Pinidiyapathirage, J. M., Platou, C. G., Potter, S., Price, J. F., Qi, L., Radha, V., Rallidis, L., Rasheed, A., Rathman, W., Rauramaa, R., Raychaudhuri, S., Rayner, N. W., Rees, S. D., Rehnberg, E., Ripatti, S., Robertson, N., Roden, M., Rossin, E. J., Rudan, I., Rybin, D., Saaristo, T. E., Salomaa, V., Saltevo, J., Samuel, M., Sanghera, D. K., Saramies, J., Scott, J., Scott, L. J., Scott, R. A., Segrè, A. V., Sehmi, J., Sennblad, B., Shah, N., Shah, S., Shera, A. S., Shu, X.-O., Shuldiner, A. R., Sigurdsson, G., Sijbrands, E., Silveira, A., Sim, X., Sivapalaratnam, S., Small, K. S., So, W. Y., Stančáková, A., Stefansson, K., Steinbach, G., Steinthorsdóttir, V., Stirrups, K., Strawbridge, R. J., Stringham, H. M., Sun, Q., Suo, C., Syvänen, A.-C., Takayanagi, R., Takeuchi, F., Tay, W. T., Teslovich, T. M., Thorand, B., Thorleifsson, G., Thorsteinsdóttir, U., Tikkanen, E., Trakalo, J., Tremoli, E., Trip, M. D., Tsai, F. J., Tuomi, T., Tuomilehto, J., Uitterlinden, A. G., Valladares-Salgado, A., Vedantam, S., Veglia, F., Voight, B. F., Wang, C., Wareham, N. J., Wennauer, R., Wickremasinghe, A. R., Wilsgaard, T., Wilson, J. F., Wiltshire, S., Winckler, W., Wong, T. Y., Wood, A. R., Wu, J.-Y., Wu, Y., Yamamoto, K., Yamauchi, T., Yang, M., Yengo, L., Yokota, M., Young, R., Zabaneh, D., Zhang,

- F., Zhang, R., Zheng, W., Zimmet, P. Z., Altshuler, D., Bowden, D. W., Cho, Y. S., Cox, N. J., Cruz, M., Hanis, C. L., Kooner, J., Lee, J.-Y., Seielstad, M., Teo, Y. Y., Boehnke, M., Parra, E. J., Chambers, J. C., Tai, E. S., McCarthy, M. I. & Morris, A. P. Genome-wide trans-ancestry meta-analysis provides insight into the genetic architecture of type 2 diabetes susceptibility. *Nat. Genet.* 46, 234–244 (2014).
6. Kato, M., Goto, A., Tanaka, T., Sasaki, S., Igata, A. & Noda, M. Effects of walking on medical cost: A quantitative evaluation by simulation focusing on diabetes. *J Diabetes Investig* 4, 667–672 (2013).
 7. Morris, A. P., Voight, B. F., Teslovich, T. M., Ferreira, T., Segrè, A. V., Steinthorsdottir, V., Strawbridge, R. J., Khan, H., Grallert, H., Mahajan, A., Prokopenko, I., Kang, H. M., Dina, C., Esko, T., Fraser, R. M., Kanoni, S., Kumar, A., Lagou, V., Langenberg, C., Luan, J., Lindgren, C. M., Müller-Nurasyid, M., Pechlivanis, S., Rayner, N. W., Scott, L. J., Wiltshire, S., Yengo, L., Kinnunen, L., Rossin, E. J., Raychaudhuri, S., Johnson, A. D., Dimas, A. S., Loos, R. J. F., Vedantam, S., Chen, H., Florez, J. C., Fox, C., Liu, C.-T., Rybin, D., Couper, D. J., Kao, W.-H. L., Li, M., Cornelis, M. C., Kraft, P., Sun, Q., van Dam, R. M., Stringham, H. M., Chines, P. S., Fischer, K., Fontanillas, P., Holmen, O. L., Hunt, S. E., Jackson, A. U., Kong, A., Lawrence, R., Meyer, J., Perry, J. R. B., Platou, C. G. P., Potter, S., Rehnberg, E., Robertson, N., Sivapalaratnam, S., Stančáková, A., Stirrups, K., Thorleifsson, G., Tikkanen, E., Wood, A. R., Almgren, P., Atalay, M., Benediktsson, R., Bonnycastle, L. L., Burt, N., Carey, J., Charpentier, G., Crenshaw, A. T., Doney, A. S. F., Dorkhan, M., Edkins, S., Emilsson, V., Eury, E., Forsen, T., Gertow, K., Gigante, B., Grant, G. B., Groves, C. J., Guiducci, C., Herder, C., Hreidarsson, A. B., Hui, J., James, A., Jonsson, A., Rathmann, W., Klopp, N., Kravic, J., Krjutškov, K., Langford, C., Leander, K., Lindholm, E., Lobbens, S., Männistö, S., Mirza, G., Mühleisen, T. W., Musk, B., Parkin, M., Rallidis, L., Saramies, J., Sennblad, B., Shah, S., Sigurðsson, G., Silveira, A., Steinbach, G., Thorand, B., Trakalo, J., Veglia, F., Wennauer, R., Winckler, W., Zabaneh, D., Campbell, H., van Duijn, C., Uitterlinden, A. G., Hofman, A., Sijbrands, E., Abecasis, G. R., Owen, K. R., Zeggini, E., Trip, M. D., Forouhi, N. G., Syvänen, A.-C., Eriksson, J. G., Peltonen, L., Nöthen, M. M., Balkau, B., Palmer, C. N. A., Lyssenko, V., Tuomi, T., Isomaa, B., Hunter, D. J., Qi, L., Wellcome Trust Case Control Consortium, Meta-Analyses of Glucose and Insulin-related traits Consortium (MAGIC) Investigators, Genetic Investigation of ANthropometric Traits (GIANT) Consortium, Asian Genetic Epidemiology Network Type 2 Diabetes (AGEN-T2D) Consortium, South Asian Type 2 Diabetes (SAT2D) Consortium, Shuldiner, A. R., Roden, M., Barroso, I., Wilsgaard, T., Beilby, J., Hovingh, K., Price, J. F., Wilson, J. F., Rauramaa, R., Lakka, T. A., Lind, L., Dedoussis, G., Njølstad, I., Pedersen, N. L., Khaw, K.-T., Wareham, N. J., Keinanen-Kiukaanniemi, S. M., Saaristo, T. E., Korpi-Hyövälti, E., Saltevo, J., Laakso, M., Kuusisto, J., Metspalu, A., Collins, F. S., Mohlke, K. L., Bergman, R. N., Tuomilehto, J., Boehm, B. O., Gieger, C., Hveem, K., Cauchi, S., Froguel, P., Baldassarre, D., Tremoli, E., Humphries, S. E., Saleheen, D., Danesh, J., Ingelsson, E., Ripatti, S., Salomaa, V., Erbel, R., Jöckel, K.-H., Moebus, S., Peters, A., Illig, T., de Faire, U., Hamsten, A., Morris, A. D., Donnelly, P. J., Frayling, T. M., Hattersley, A. T.,

- Boerwinkle, E., Melander, O., Kathiresan, S., Nilsson, P. M., Deloukas, P., Thorsteinsdottir, U., Groop, L. C., Stefansson, K., Hu, F., Pankow, J. S., Dupuis, J., Meigs, J. B., Altshuler, D., Boehnke, M., McCarthy, M. I. DIABetes Genetics Replication And Meta-analysis (DIAGRAM) Consortium. Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes. *Nat. Genet.* 44, 981–990 (2012).
8. Wang, X., Strizich, G., Hu, Y., Wang, T., Kaplan, R. C. & Qi, Q. Genetic markers of type 2 diabetes: Progress in genome-wide association studies and clinical application for risk prediction. *J Diabetes* 8, 24–35 (2016).
 9. Herder, C. & Roden, M. Genetics of type 2 diabetes: pathophysiologic and clinical relevance. *Eur. J. Clin. Invest.* 41, 679–692 (2011).
 10. Jolma, A., Kivioja, T., Toivonen, J., Cheng, L., Wei, G., Enge, M., Taipale, M., Vaquerizas, J. M., Yan, J., Sillanpää, M. J., Bonke, M., Palin, K., Talukder, S., Hughes, T. R., Luscombe, N. M., Ukkonen, E. & Taipale, J. Multiplexed massively parallel SELEX for characterization of human transcription factor binding specificities. *Genome Res.* 20, 861–873 (2010).
 11. Jolma, A., Yan, J., Whittington, T., Toivonen, J., Nitta, K. R., Rastas, P., Morgunova, E., Enge, M., Taipale, M., Wei, G., Palin, K., Vaquerizas, J. M., Vincentelli, R., Luscombe, N. M., Hughes, T. R., Lemaire, P., Ukkonen, E., Kivioja, T. & Taipale, J. DNA-Binding Specificities of Human Transcription Factors. *Cell* 152, 327–339 (2013).
 12. Kato, N. Insights into the genetic basis of type 2 diabetes. *J Diabetes Investig* 4, 233–244 (2013).
 13. Dror, I., Golan, T., Levy, C., Rohs, R. & Mandel-Gutfreund, Y. A widespread role of the motif environment in transcription factor binding across diverse protein families. *Genome Res.* 25, 1268–1280 (2015).
 14. Najafabadi, H. S., Mnaimneh, S., Schmitges, F. W., Garton, M., Lam, K. N., Yang, A., Albu, M., Weirauch, M. T., Radovani, E., Kim, P. M., Greenblatt, J., Frey, B. J. & Hughes, T. R. C2H2 zinc finger proteins greatly expand the human regulatory lexicon. *Nature Publishing Group* 33, 555–562 (2015).
 15. 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. & Stamatoyannopoulos, J. A. Systematic Localization of Common Disease-Associated Variation in Regulatory DNA. *Science* 337, 1190–1195 (2012).
 16. Leung, D., Jung, I., Rajagopal, N., Schmitt, A., Selvaraj, S., Lee, A. Y., Yen, C.-A., Lin, S., Lin, Y., Qiu, Y., Xie, W., Yue, F., Hariharan, M., Ray, P., Kuan, S., Edsall, L.,

- Yang, H., Chi, N. C., Zhang, M. Q., Ecker, J. R. & Ren, B. Integrative analysis of haplotype-resolved epigenomes across human tissues. *Nature* 518, 350–354 (2015).
17. Lee, D., Gorkin, D. U., Baker, M., Strober, B. J., Asoni, A. L., McCallion, A. S. & Beer, M. A. A method to predict the impact of regulatory variants from DNA sequence. *Nat. Genet.* 47, 955–961 (2015).
 18. Zhou, J. & Troyanskaya, O. G. Predicting effects of noncoding variants with deep learning–based sequence model. *Nat Meth* 12, 931–934 (2015).
 19. Arnold, C. D., Gerlach, D., Stelzer, C., Boryń, Ł. M., Rath, M. & Stark, A. Genome-wide quantitative enhancer activity maps identified by STARR-seq. *Science* 339, 1074–1077 (2013).
 20. Greenwald, W. W., Chiou, J., Yan, J., Qiu, Y., Dai, N., Wang, A., Nariyai, N., Aylward, A., Han, J. Y., Kadakia, N., Regue, L., Okino, M.-L., Drees, F., Kramer, D., Vinckier, N., Minichiello, L., Gorkin, D., Avruch, J., Frazer, K. A., Sander, M., Ren, B. & Gaulton, K. J. Pancreatic islet chromatin accessibility and conformation reveals distal enhancer networks of type 2 diabetes risk. *Nature Communications* 10, 2078–12 (2019).
 21. Rao, S. S. P., Huntley, M. H., Durand, N. C., Stamenova, E. K., Bochkov, I. D., Robinson, J. T., Sanborn, A. L., Machol, I., Omer, A. D., Lander, E. S. & Aiden, E. L. A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping. *Cell* 159, 1665–1680 (2014).
 22. Scott, R. A., Lagou, V., Welch, R. P., Wheeler, E., Montasser, M. E., Luan, J., Mägi, R., Strawbridge, R. J., Rehnberg, E., Gustafsson, S., Kanoni, S., Rasmussen-Torvik, L. J., Yengo, L., Lecoeur, C., Shungin, D., Sanna, S., Sidore, C., Johnson, P. C. D., Jukema, J. W., Johnson, T., Mahajan, A., Verweij, N., Thorleifsson, G., Hottenga, J.-J., Shah, S., Smith, A. V., Sennblad, B., Gieger, C., Salo, P., Perola, M., Timpson, N. J., Evans, D. M., Pourcain, B. S., Wu, Y., Andrews, J. S., Hui, J., Bielak, L. F., Zhao, W., Horikoshi, M., Navarro, P., Isaacs, A., O'Connell, J. R., Stirrups, K., Vitart, V., Hayward, C., Esko, T., Mihailov, E., Fraser, R. M., Fall, T., Voight, B. F., Raychaudhuri, S., Chen, H., Lindgren, C. M., Morris, A. P., Rayner, N. W., Robertson, N., Rybin, D., Liu, C.-T., Beckmann, J. S., Willems, S. M., Chines, P. S., Jackson, A. U., Kang, H. M., Stringham, H. M., Song, K., Tanaka, T., Peden, J. F., Goel, A., Hicks, A. A., An, P., Müller-Nurasyid, M., Franco-Cereceda, A., Folkersen, L., Marullo, L., Jansen, H., Oldehinkel, A. J., Bruinenberg, M., Pankow, J. S., North, K. E., Forouhi, N. G., Loos, R. J. F., Edkins, S., Varga, T. V., Hallmans, G., Oksa, H., Antonella, M., Nagaraja, R., Trompet, S., Ford, I., Bakker, S. J. L., Kong, A., Kumari, M., Gigante, B., Herder, C., Munroe, P. B., Caulfield, M., Antti, J., Mangino, M., Small, K., Miljkovic, I., Liu, Y., Atalay, M., Kiess, W., James, A. L., Rivadeneira, F., Uitterlinden, A. G., Palmer, C. N. A., Doney, A. S. F., Willemsen, G., Smit, J. H., Campbell, S., Polasek, O., Bonnycastle, L. L., Hercberg, S., Dimitriou, M., Bolton, J. L., Fowkes, G. R., Kovacs, P., Lindstrom, J., Zemunik, T., Bandinelli, S., Wild, S. H., Basart, H. V., Rathmann, W., Grallert, H., DIAbetes Genetics Replication And Meta-analysis

(DIAGRAM) Consortium, Maerz, W., Kleber, M. E., Boehm, B. O., Peters, A., Pramstaller, P. P., Province, M. A., Borecki, I. B., Hastie, N. D., Rudan, I., Campbell, H., Watkins, H., Farrall, M., Stumvoll, M., Ferrucci, L., Waterworth, D. M., Bergman, R. N., Collins, F. S., Tuomilehto, J., Watanabe, R. M., de Geus, E. J. C., Penninx, B. W., Hofman, A., Oostra, B. A., Psaty, B. M., Vollenweider, P., Wilson, J. F., Wright, A. F., Hovingh, G. K., Metspalu, A., Uusitupa, M., Magnusson, P. K. E., Kyvik, K. O., Kaprio, J., Price, J. F., Dedoussis, G. V., Deloukas, P., Meneton, P., Lind, L., Boehnke, M., Shuldiner, A. R., van Duijn, C. M., Morris, A. D., Toenjes, A., Peyser, P. A., Beilby, J. P., Körner, A., Kuusisto, J., Laakso, M., Bornstein, S. R., Schwarz, P. E. H., Lakka, T. A., Rauramaa, R., Adair, L. S., Smith, G. D., Spector, T. D., Illig, T., de Faire, U., Hamsten, A., Gudnason, V., Kivimaki, M., Hingorani, A., Keinanen-Kiukaanniemi, S. M., Saaristo, T. E., Boomsma, D. I., Stefansson, K., van der Harst, P., Dupuis, J., Pedersen, N. L., Sattar, N., Harris, T. B., Cucca, F., Ripatti, S., Salomaa, V., Mohlke, K. L., Balkau, B., Froguel, P., Pouta, A., Jarvelin, M.-R., Wareham, N. J., Bouatia-Naji, N., McCarthy, M. I., Franks, P. W., Meigs, J. B., Teslovich, T. M., Florez, J. C., Langenberg, C., Ingelsson, E., Prokopenko, I. & Barroso, I. Large-scale association analyses identify new loci influencing glycemic traits and provide insight into the underlying biological pathways. *Nat. Genet.* 44, 991–1005 (2012).

23. Manning, A. K., Hivert, M.-F., Scott, R. A., Grimsby, J. L., Bouatia-Naji, N., Chen, H., Rybin, D., Liu, C.-T., Bielak, L. F., Prokopenko, I., Amin, N., Barnes, D., Cadby, G., Hottenga, J.-J., Ingelsson, E., Jackson, A. U., Johnson, T., Kanoni, S., Ladenvall, C., Lagou, V., Lahti, J., Lecoeur, C., Liu, Y., Martinez-Larrad, M. T., Montasser, M. E., Navarro, P., Perry, J. R. B., Rasmussen-Torvik, L. J., Salo, P., Sattar, N., Shungin, D., Strawbridge, R. J., Tanaka, T., van Duijn, C. M., An, P., de Andrade, M., Andrews, J. S., Aspelund, T., Atalay, M., Aulchenko, Y., Balkau, B., Bandinelli, S., Beckmann, J. S., Beilby, J. P., Bellis, C., Bergman, R. N., Blangero, J., Boban, M., Boehnke, M., Boerwinkle, E., Bonnycastle, L. L., Boomsma, D. I., Borecki, I. B., Böttcher, Y., Bouchard, C., Brunner, E., Budimir, D., Campbell, H., Carlson, O., Chines, P. S., Clarke, R., Collins, F. S., Corbatón-Anchuelo, A., Couper, D., de Faire, U., Dedoussis, G. V., Deloukas, P., Dimitriou, M., Egan, J. M., Eiriksdottir, G., Erdos, M. R., Eriksson, J. G., Eury, E., Ferrucci, L., Ford, I., Forouhi, N. G., Fox, C. S., Franzosi, M. G., Franks, P. W., Frayling, T. M., Froguel, P., Galan, P., de Geus, E., Gigante, B., Glazer, N. L., Goel, A., Groop, L., Gudnason, V., Hallmans, G., Hamsten, A., Hansson, O., Harris, T. B., Hayward, C., Heath, S., Hercberg, S., Hicks, A. A., Hingorani, A., Hofman, A., Hui, J., Hung, J., Jarvelin, M.-R., Jhun, M. A., Johnson, P. C. D., Jukema, J. W., Jula, A., Kao, W. H., Kaprio, J., Kardina, S. L. R., Keinanen-Kiukaanniemi, S., Kivimaki, M., Kolcic, I., Kovacs, P., Kumari, M., Kuusisto, J., Kyvik, K. O., Laakso, M., Lakka, T., Lannfelt, L., Lathrop, G. M., Launer, L. J., Leander, K., Li, G., Lind, L., Lindstrom, J., Lobbens, S., Loos, R. J. F., Luan, J., Lyssenko, V., Mägi, R., Magnusson, P. K. E., Marmot, M., Meneton, P., Mohlke, K. L., Mooser, V., Morken, M. A., Miljkovic, I., Narisu, N., O'Connell, J., Ong, K. K., Ben A Oostra, Palmer, L. J., Palotie, A., Pankow, J. S., Peden, J. F., Pedersen, N. L., Pehlic, M., Peltonen, L., Penninx, B., Pericic, M., Perola, M., Perusse, L., Peyser, P. A., Polasek, O., Pramstaller, P. P., Province, M. A., Rääkkönen, K., Rauramaa, R., Rehnberg, E.,

- Rice, K., Rotter, J. I., Rudan, I., Ruukonen, A., Saaristo, T., Sabater-Lleal, M., Salomaa, V., Savage, D. B., Saxena, R., Schwarz, P., Seedorf, U., Sennblad, B., Serrano-Rios, M., Shuldiner, A. R., Sijbrands, E. J. G., Siscovick, D. S., Smit, J. H., Small, K. S., Smith, N. L., Smith, A. V., Stančáková, A., Stirrups, K., Stumvoll, M., Sun, Y. V., Swift, A. J., Tonjes, A., Tuomilehto, J., Trompet, S., Uitterlinden, A. G., Uusitupa, M., Vikström, M., Vitart, V., Vohl, M.-C., Voight, B. F., Vollenweider, P., Waeber, G., Waterworth, D. M., Watkins, H., Wheeler, E., Widen, E., Wild, S. H., Willems, S. M., Willemsen, G., Wilson, J. F., Witteman, J. C. M., Wright, A. F., Yaghootkar, H., Zelenika, D., Zemunik, T., Zgaga, L., Wareham, N. J., McCarthy, M. I., Barroso, I., Watanabe, R. M., Florez, J. C., Dupuis, J., Meigs, J. B. & Langenberg, C. A genome-wide approach accounting for body mass index identifies genetic variants influencing fasting glycemic traits and insulin resistance. *Nat. Genet.* 44, 659–669 (2012).
24. Varshney, A., Scott, L. J., Welch, R. P., Erdos, M. R., Chines, P. S., Narisu, N., Albanus, R. D., Orchard, P., Wolford, B. N., Kursawe, R., Vadlamudi, S., Cannon, M. E., Didion, J. P., Hensley, J., Kirilusha, A., NISC Comparative Sequencing Program, Bonnycastle, L. L., Taylor, D. L., Watanabe, R., Mohlke, K. L., Boehnke, M., Collins, F. S., Parker, S. C. J. & Stitzel, M. L. Genetic regulatory signatures underlying islet gene expression and type 2 diabetes. *Proc. Natl. Acad. Sci. U.S.A.* 114, 2301–2306 (2017).
25. Travers, M. E., Mackay, D. J. G., Dekker Nitert, M., Morris, A. P., Lindgren, C. M., Berry, A., Johnson, P. R., Hanley, N., Groop, L. C., McCarthy, M. I. & Gloyn, A. L. Insights into the molecular mechanism for type 2 diabetes susceptibility at the KCNQ1 locus from temporal changes in imprinting status in human islets. *Diabetes* 62, 987–992 (2013).
26. Chakravarti, A., La Vega, De, F. M., Donnelly, P., Egholm, M., Knoppers, B. M., Nickerson, D. A., Peltonen, L., Schafer, A. J., Deiros, D., Metzker, M., Muzny, D., Reid, J., Wang, J., Li, J., Jian, M., Liang, H., Tian, G., Wang, B., Wang, W., Zhang, X., Zheng, H., Lander, E. S., Ambrogio, L., Jaffe, D. B., Sougnez, C. L., Bentley, D. R., Gormley, N., Kingsbury, Z., Koko-Gonzales, P., Stone, J., McKernan, K. J., Costa, G. L., Ichikawa, J. K., Lee, C. C., Lehrach, H., Borodina, T. A., Dahl, A., Davydov, A. N., Marquardt, P., Mertes, F., Nietfeld, W., Rosenstiel, P., Schreiber, S., Soldatov, A. V., Timmermann, B., Tolzmann, M., Affourtit, J., Ashworth, D., Attiya, S., Bachorski, M., Buglione, E., Burke, A., Caprio, A., Celone, C., Clark, S., Conners, D., Gu, L., Guccione, L., Kao, K., Kebbel, A., Knowlton, J., Labrecque, M., McDade, L., Mealmaker, C., Minderman, M., Nawrocki, A., Niazi, F., Pareja, K., Ramenani, R., Riches, D., Song, W., Turcotte, C., Wang, S., Wilson, R. K., Fulton, L., Burton, J., Churcher, C., Coffey, A., Cox, A., Quail, M., Skelly, T., Swerdlow, H. P., Turner, D., De Witte, A., Giles, S., Wheeler, D., Challis, D., Sabo, A., Yu, F., Yu, J., Fang, X., Guo, X., Li, R., Tai, S., Wu, H., Zheng, H., Zheng, X., Zhou, Y., Li, G., Wang, J., Yang, H., Marth, G. T., Garrison, E. P., Huang, W., Indap, A., Lee, W.-P., Stromberg, M. P., Mills, R. E., Daly, M. J., Ball, A. D., Bloom, T., Browning, B. L., Grossman, S. R., Handsaker, R. E., Hanna, M., Hartl, C., Kernytsky, A. M., Korn, J. M., Maguire, J. R., McCarroll, S. A., McKenna, A., Philippakis, A. A., Poplin, R. E., Price, A., Rivas,

M. A., Sabeti, P. C., Schaffner, S. F., Shlyakhter, I. A., Cooper, D. N., Ball, E. V., Mort, M., Phillips, A. D., Stenson, P. D., Makarov, V., Boyko, A., Degenhardt, J., Gravel, S., Gutenkunst, R. N., Kaganovich, M., Keinan, A., Lacroute, P., Ma, X., Reynolds, A., Clarke, L., Flicek, P., Cunningham, F., Herrero, J., Keenen, S., Kulesha, E., Leinonen, R., McLaren, W. M., Radhakrishnan, R., Smith, R. E., Zalunin, V., Zheng-Bradley, X., Korbel, J. O., Stütz, A. M., Bauer, M., Keira Cheetham, R., Cox, T., James, T., Hyland, F. C. L., Manning, J. M., McLaughlin, S. F., Peckham, H. E., Sakarya, O., Tsung, E. F., Konkel, M. K., Walker, J. A., Sudbrak, R., Albrecht, M. W., Amstislavskiy, V. S., Herwig, R., Parkhomchuk, D. V., Sherry, S. T., Agarwala, R., Khouri, H. M., Morgulis, A. O., Paschall, J. E., Phan, L. D., Rotmistrovsky, K. E., Sanders, R. D., Shumway, M. F., Auton, A., Iqbal, Z., Lunter, G., Marchini, J. L., Moutsianas, L., Myers, S., Tumian, A., Desany, B., Knight, J., Winer, R., Craig, D. W., Beckstrom-Sternberg, S. M., Christoforides, A., Kurdoglu, A. A., Pearson, J. V., Sinari, S. A., Tembe, W. D., Haussler, D., Hinrichs, A. S., Kern, A., Kuhn, R. M., Przeworski, M., Hernandez, R. D., Howie, B., Kelley, J. L., Cord Melton, S., Li, Y., Anderson, P., Chen, W., Cookson, W. O., Ding, J., Min Kang, H., Lathrop, M., Liang, L., Moffatt, M. F., Scheet, P., Sidore, C., Zhan, X., Zöllner, S., Awadalla, P., Casals, F., Idaghdour, Y., Keebler, J., Stone, E. A., Zilvermit, M., Jorde, L., Hajirasouliha, I., Cenk Sahinalp, S., Sudmant, P. H., Mardis, E. R., Chen, K., Chinwalla, A., Koboldt, D. C., Weinstock, G., Wendl, M. C., Zhang, Q., Albers, C. A., Ayub, Q., Barrett, J. C., Carter, D. M., Chen, Y., Conrad, D. F., Danecek, P., Dermitzakis, E. T., Hu, M., Huang, N., Hurler, M. E., Jin, H., Jostins, L., Quang Le, S., Lindsay, S., Long, Q., Montgomery, S. B., Parts, L., Walter, K., Snyder, M., Abyzov, A., Balasubramanian, S., Bjornson, R., Du, J., Grubert, F., Habegger, L., Haraksingh, R., Jee, J., Lam, H. Y. K., Jasmine Mu, X., Zhang, Z., Li, Y., Luo, R., Kural, D., Quinlan, A. R., Ward, A. N., Lee, C., Shi, X., Banks, E., DePristo, M. A., Li, H., Nemesh, J. C., Sebat, J., Ye, K., Yoon, S. C., Humphray, S., Eberle, M., Kahn, S., Murray, L., Ye, K., Fu, Y., Sun, Y. A., Batzer, M. A., Xiao, C., Snyder, M., Xing, J., Eichler, E. E., Aksay, G., Alkan, C., Hormozdiari, F., Kidd, J. M., Ding, L., McLellan, M. D., Wallis, J. W., Zhang, Y., Khurana, E., Leng, J., Urban, A. E., Bainbridge, M., Coafra, C., Dinh, H., Kovar, C., Lee, S., Nazareth, L., Fung Leong, W., Stewart, C., Wu, J., Cibulskis, K., Fennell, T. J., Gabriel, S. B., Garimella, K. V., Shefler, E., Wilkinson, J., Clark, A. G., McVean, G. A., Katzman, S. J., Blackwell, T., Dooling, D., Fulton, R., Balasubramanian, S., Coffey, A., Keane, T. M., MacArthur, D. G., Palotie, A., Scott, C., Stalker, J., Tyler Smith, C., Gerstein, M. B., Bustamante, C. D., Gharani, N., Kaye, J. S., Kent, A., Li, T., McGuire, A. L., Ossorio, P. N., Rotimi, C. N., Su, Y., Toji, L. H., Brooks, L. D., Felsenfeld, A. L., McEwen, J. E., Abdallah, A., Juenger, C. R., Clemm, N. C., Collins, F. S., Duncanson, A., Green, E. D., Guyer, M. S., Peterson, J. L., Abecasis, G. R., Altshuler, D. L., Durbin, R. M. & Gibbs, R. A. A map of human genome variation from population-scale sequencing. *Nature* 467, 1061–1073 (2010).

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

28. Lambert, J. C., Ibrahim-Verbaas, C. A., Harold, D., Naj, A. C., Sims, R., Bellenguez, C., DeStafano, A. L., Bis, J. C., Beecham, G. W., Grenier-Boley, B., Russo, G., Thorton-Wells, T. A., Jones, N., Smith, A. V., Chouraki, V., Thomas, C., Ikram, M. A., Zelenika, D., Vardarajan, B. N., Kamatani, Y., Lin, C. F., Gerrish, A., Schmidt, H., Kunkle, B., Dunstan, M. L., Ruiz, A., Bihoreau, M. T., Choi, S. H., Reitz, C., Pasquier, F., Cruchaga, C., Craig, D., Amin, N., Berr, C., Lopez, O. L., De Jager, P. L., Deramecourt, V., Johnston, J. A., Evans, D., Lovestone, S., Letenneur, L., Morón, F. J., Rubinsztein, D. C., Eiriksdottir, G., Sleegers, K., Goate, A. M., Fiévet, N., Huentelman, M. W., Gill, M., Brown, K., Kamboh, M. I., Keller, L., Barberger-Gateau, P., McGuinness, B., Larson, E. B., Green, R., Myers, A. J., Dufouil, C., Todd, S., Wallon, D., Love, S., Rogaeva, E., Gallacher, J., St George-Hyslop, P., Clarimon, J., Lleo, A., Bayer, A., Tsuang, D. W., Yu, L., Tsolaki, M., Bossù, P., Spalletta, G., Proitsi, P., Collinge, J., Sorbi, S., Sanchez-Garcia, F., Fox, N. C., Hardy, J., Deniz Naranjo, M. C., Bosco, P., Clarke, R., Brayne, C., Galimberti, D., Mancuso, M., Matthews, F., European Alzheimer's Disease Initiative (EADI), Genetic and Environmental Risk in Alzheimer's Disease, Alzheimer's Disease Genetic Consortium, Cohorts for Heart and Aging Research in Genomic Epidemiology, Moebus, S., Mecocci, P., Del Zompo, M., Maier, W., Hampel, H., Pilotto, A., Bullido, M., Panza, F., Caffarra, P., Nacmias, B., Gilbert, J. R., Mayhaus, M., Lannefelt, L., Hakonarson, H., Pichler, S., Carrasquillo, M. M., Ingelsson, M., Beekly, D., Alvarez, V., Zou, F., Valladares, O., Younkin, S. G., Coto, E., Hamilton-Nelson, K. L., Gu, W., Razquin, C., Pastor, P., Mateo, I., Owen, M. J., Faber, K. M., Jonsson, P. V., Combarros, O., O'Donovan, M. C., Cantwell, L. B., Soininen, H., Blacker, D., Mead, S., Mosley, T. H., Bennett, D. A., Harris, T. B., Fratiglioni, L., Holmes, C., de Bruijn, R. F., Passmore, P., Montine, T. J., Bettens, K., Rotter, J. I., Brice, A., Morgan, K., Foroud, T. M., Kukull, W. A., Hannequin, D., Powell, J. F., Nalls, M. A., Ritchie, K., Lunetta, K. L., Kauwe, J. S., Boerwinkle, E., Riemenschneider, M., Boada, M., Hiltunen, M., Martin, E. R., Schmidt, R., Rujescu, D., Wang, L. S., Dartigues, J. F., Mayeux, R., Tzourio, C., Hofman, A., Nöthen, M. M., Graff, C., Psaty, B. M., Jones, L., Haines, J. L., Holmans, P. A., Lathrop, M., Pericak-Vance, M. A., Launer, L. J., Farrer, L. A., van Duijn, C. M., Van Broeckhoven, C., Moskvina, V., Seshadri, S., Williams, J., Schellenberg, G. D. & Amouyel, P. Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease. *Nat. Genet.* 45, 1452–1458 (2013).
29. Willer, C. J., Schmidt, E. M., Sengupta, S., Peloso, G. M., Gustafsson, S., Kanoni, S., Ganna, A., Chen, J., Buchkovich, M. L., Mora, S., Beckmann, J. S., Bragg-Gresham, J. L., Chang, H.-Y., Demirkan, A., Hertog, Den, H. M., Do, R., Donnelly, L. A., Ehret, G. B., Esko, T., Feitosa, M. F., Ferreira, T., Fischer, K., Fontanillas, P., Fraser, R. M., Freitag, D. F., Gurdasani, D., Heikkilä, K., Hyppönen, E., Isaacs, A., Jackson, A. U., Johansson, Å., Johnson, T., Kaakinen, M., Kettunen, J., Kleber, M. E., Li, X., Luan, J., Lytikäinen, L.-P., Magnusson, P. K. E., Mangino, M., Mihailov, E., Montasser, M. E., Müller-Nurasyid, M., Nolte, I. M., O'Connell, J. R., Palmer, C. D., Perola, M., Petersen, A.-K., Sanna, S., Saxena, R., Service, S. K., Shah, S., Shungin, D., Sidore, C., Song, C., Strawbridge, R. J., Surakka, I., Tanaka, T., Teslovich, T. M., Thorleifsson, G., Van den Herik, E. G., Voight, B. F., Volcik, K. A.,

Waite, L. L., Wong, A., Wu, Y., Zhang, W., Absher, D., Asiki, G., Barroso, I., Been, L. F., Bolton, J. L., Bonnycastle, L. L., Brambilla, P., Burnett, M. S., Cesana, G., Dimitriou, M., Doney, A. S. F., Döring, A., Elliott, P., Epstein, S. E., Ingi Eyjolfsson, G., Gigante, B., Goodarzi, M. O., Grallert, H., Gravito, M. L., Groves, C. J., Hallmans, G., Hartikainen, A.-L., Hayward, C., Hernandez, D., Hicks, A. A., Holm, H., Hung, Y.-J., Illig, T., Jones, M. R., Kaleebu, P., Kastelein, J. J. P., Khaw, K.-T., Kim, E., Klopp, N., Komulainen, P., Kumari, M., Langenberg, C., Lehtimäki, T., Lin, S.-Y., Lindstrom, J., Loos, R. J. F., Mach, F., McArdle, W. L., Meisinger, C., Mitchell, B. D., Müller, G., Nagaraja, R., Narisu, N., Nieminen, T. V. M., Nsubuga, R. N., Olafsson, I., Ong, K. K., Palotie, A., Papamarkou, T., Pomilla, C., Pouta, A., Rader, D. J., Reilly, M. P., Ridker, P. M., Rivadeneira, F., Rudan, I., Ruukonen, A., Samani, N., Scharnagl, H., Seeley, J., Silander, K., Stančáková, A., Stirrups, K., Swift, A. J., Tiret, L., Uitterlinden, A. G., van Pelt, L. J., Vedantam, S., Wainwright, N., Wijmenga, C., Wild, S. H., Willemssen, G., Wilsgaard, T., Wilson, J. F., Young, E. H., Zhao, J. H., Adair, L. S., Arveiler, D., Assimes, T. L., Bandinelli, S., Bennett, F., Bochud, M., Boehm, B. O., Boomsma, D. I., Borecki, I. B., Bornstein, S. R., Bovet, P., Burnier, M., Campbell, H., Chakravarti, A., Chambers, J. C., Chen, Y.-D. I., Collins, F. S., Cooper, R. S., Danesh, J., Dedoussis, G., de Faire, U., Feranil, A. B., Ferrières, J., Ferrucci, L., Freimer, N. B., Gieger, C., Groop, L. C., Gudnason, V., Gyllensten, U., Hamsten, A., Harris, T. B., Hingorani, A., Hirschhorn, J. N., Hofman, A., Hovingh, G. K., Hsiung, C. A., Humphries, S. E., Hunt, S. C., Hveem, K., Iribarren, C., Jarvelin, M.-R., Jula, A., Kähönen, M., Kaprio, J., Kesäniemi, A., Kivimäki, M., Kooner, J. S., Koudstaal, P. J., Krauss, R. M., Kuh, D., Kuusisto, J., Kyvik, K. O., Laakso, M., Lakka, T. A., Lind, L., Lindgren, C. M., Martin, N. G., März, W., McCarthy, M. I., McKenzie, C. A., Meneton, P., Metspalu, A., Moilanen, L., Morris, A. D., Munroe, P. B., Njølstad, I., Pedersen, N. L., Power, C., Pramstaller, P. P., Price, J. F., Psaty, B. M., Quertermous, T., Rauramaa, R., Saleheen, D., Salomaa, V., Sanghera, D. K., Saramies, J., Schwarz, P. E. H., Sheu, W. H.-H., Shuldiner, A. R., Siegbahn, A., Spector, T. D., Stefansson, K., Strachan, D. P., Tayo, B. O., Tremoli, E., Tuomilehto, J., Uusitupa, M., van Duijn, C. M., Vollenweider, P., Wallentin, L., Wareham, N. J., Whitfield, J. B., Wolfenbittel, B. H. R., Ordovas, J. M., Boerwinkle, E., Palmer, C. N. A., Thorsteinsdottir, U., Chasman, D. I., Rotter, J. I., Franks, P. W., Ripatti, S., Cupples, L. A., Sandhu, M. S., Rich, S. S., Boehnke, M., Deloukas, P., Kathiresan, S., Mohlke, K. L., Ingelsson, E., Abecasis, G. R. Global Lipids Genetics Consortium. Discovery and refinement of loci associated with lipid levels. *Nat. Genet.* 45, 1274–1283 (2013).

30. Dubois, P. C. A., Trynka, G., Franke, L., Hunt, K. A., Romanos, J., Curtotti, A., Zhernakova, A., Heap, G. A. R., Adány, R., Aromaa, A., Bardella, M. T., van den Berg, L. H., Bockett, N. A., la Concha, de, E. G., Dema, B., Fehrmann, R. S. N., Fernández-Arquero, M., Fiatal, S., Grandone, E., Green, P. M., Groen, H. J. M., Gwilliam, R., Houwen, R. H. J., Hunt, S. E., Kaukinen, K., Kelleher, D., Korponay-Szabo, I., Kurppa, K., MacMathuna, P., Mäki, M., Mazzilli, M. C., McCann, O. T., Mearin, M. L., Mein, C. A., Mirza, M. M., Mistry, V., Mora, B., Morley, K. I., Mulder, C. J., Murray, J. A., Núñez, C., Oosterom, E., Ophoff, R. A., Polanco, I., Peltonen, L., Platteel, M., Rybak, A., Salomaa, V., Schweizer, J. J., Sperandeo, M. P., Tack,

- G. J., Turner, G., Veldink, J. H., Verbeek, W. H. M., Weersma, R. K., Wolters, V. M., Urcelay, E., Cukrowska, B., Greco, L., Neuhausen, S. L., McManus, R., Barisani, D., Deloukas, P., Barrett, J. C., Saavalainen, P., Wijmenga, C. & van Heel, D. A. Multiple common variants for celiac disease influencing immune gene expression. *Nat. Genet.* 42, 295–302 (2010).
31. Okada, Y., Wu, D., Trynka, G., Raj, T., Terao, C., Ikari, K., Kochi, Y., Ohmura, K., Suzuki, A., Yoshida, S., Graham, R. R., Manoharan, A., Ortmann, W., Bhangale, T., Denny, J. C., Carroll, R. J., Eyler, A. E., Greenberg, J. D., Kremer, J. M., Pappas, D. A., Jiang, L., Yin, J., Ye, L., Su, D.-F., Yang, J., Xie, G., Keystone, E., Westra, H.-J., Esko, T., Metspalu, A., Zhou, X., Gupta, N., Mirel, D., Stahl, E. A., Diogo, D., Cui, J., Liao, K., Guo, M. H., Myouzen, K., Kawaguchi, T., Coenen, M. J. H., van Riel, P. L. C. M., van de Laar, M. A. F. J., Guchelaar, H.-J., Huizinga, T. W. J., Dieudé, P., Mariette, X., Bridges, S. L., Zhernakova, A., Toes, R. E. M., Tak, P. P., Miceli-Richard, C., Bang, S.-Y., Lee, H.-S., Martin, J., Gonzalez-Gay, M. A., Rodriguez-Rodriguez, L., Rantapaa-Dahlqvist, S., Arlestig, L., Choi, H. K., Kamatani, Y., Galan, P., Lathrop, M., RACI consortium, GARNET consortium, Eyre, S., Bowes, J., Barton, A., de Vries, N., Moreland, L. W., Criswell, L. A., Karlson, E. W., Taniguchi, A., Yamada, R., Kubo, M., Liu, J. S., Bae, S.-C., Worthington, J., Padyukov, L., Klareskog, L., Gregersen, P. K., Raychaudhuri, S., Stranger, B. E., De Jager, P. L., Franke, L., Visscher, P. M., Brown, M. A., Yamanaka, H., Mimori, T., Takahashi, A., Xu, H., Behrens, T. W., Siminovitch, K. A., Momohara, S., Matsuda, F., Yamamoto, K. & Plenge, R. M. Genetics of rheumatoid arthritis contributes to biology and drug discovery. *Nature* 506, 376–381 (2014).
 32. Schizophrenia Working Group of the Psychiatric Genomics Consortium. Biological insights from 108 schizophrenia-associated genetic loci. *Nature* 511, 421–427 (2014).
 33. Bentham, J., Morris, D. L., Graham, D. S. C., Pinder, C. L., Tombleson, P., Behrens, T. W., Martin, J., Fairfax, B. P., Knight, J. C., Chen, L., Replogle, J., Syvänen, A.-C., Rönnblom, L., Graham, R. R., Wither, J. E., Rioux, J. D., Alarcón-Riquelme, M. E. & Vyse, T. J. Genetic association analyses implicate aberrant regulation of innate and adaptive immunity genes in the pathogenesis of systemic lupus erythematosus. *Nat. Genet.* 47, 1457–1464 (2015).
 34. de Lange, K. M., Moutsianas, L., Lee, J. C., Lamb, C. A., Luo, Y., Kennedy, N. A., Jostins, L., Rice, D. L., Gutierrez-Achury, J., Ji, S.-G., Heap, G., Nimmo, E. R., Edwards, C., Henderson, P., Mowat, C., Sanderson, J., Satsangi, J., Simmons, A., Wilson, D. C., Tremelling, M., Hart, A., Mathew, C. G., Newman, W. G., Parkes, M., Lees, C. W., Uhlig, H., Hawkey, C., Prescott, N. J., Ahmad, T., Mansfield, J. C., Anderson, C. A. & Barrett, J. C. Genome-wide association study implicates immune activation of multiple integrin genes in inflammatory bowel disease. *Nat. Genet.* 49, 256–261 (2017).
 35. Nelson, C. P., Goel, A., Butterworth, A. S., Kanoni, S., Webb, T. R., Marouli, E., Zeng, L., Ntalla, I., Lai, F. Y., Hopewell, J. C., Giannakopoulou, O., Jiang, T., Hamby, S. E.,

- Di Angelantonio, E., Assimes, T. L., Bottinger, E. P., Chambers, J. C., Clarke, R., Palmer, C. N. A., Cubbon, R. M., Ellinor, P., Ermel, R., Evangelou, E., Franks, P. W., Grace, C., Gu, D., Hingorani, A. D., Howson, J. M. M., Ingelsson, E., Kastrati, A., Kessler, T., Kyriakou, T., Lehtimäki, T., Lu, X., Lu, Y., März, W., McPherson, R., Metspalu, A., Pujades-Rodriguez, M., Ruusalepp, A., Schadt, E. E., Schmidt, A. F., Sweeting, M. J., Zalloua, P. A., AlGhalayini, K., Keavney, B. D., Kooner, J. S., Loos, R. J. F., Patel, R. S., Rutter, M. K., Tomaszewski, M., Tzoulaki, I., Zeggini, E., Erdmann, J., Dedoussis, G., Björkegren, J. L. M., Schunkert, H., Farrall, M., Danesh, J., Samani, N. J., Watkins, H. & Deloukas, P. Association analyses based on false discovery rate implicate new loci for coronary artery disease. *Nat. Genet.* 49, 1385–1391 (2017).
36. Wray, N. R., Ripke, S., Mattheisen, M., Trzaskowski, M., Byrne, E. M., Abdellaoui, A., Adams, M. J., Agerbo, E., Air, T. M., Andlauer, T. M. F., Bacanu, S.-A., Bækvad-Hansen, M., Beekman, A. F. T., Bigdeli, T. B., Binder, E. B., Blackwood, D. R. H., Bryois, J., Buttenschøn, H. N., Bybjerg-Grauholm, J., Cai, N., Castelao, E., Christensen, J. H., Clarke, T.-K., Coleman, J. I. R., Colodro-Conde, L., Couvy-Duchesne, B., Craddock, N., Crawford, G. E., Crowley, C. A., Dashti, H. S., Davies, G., Deary, I. J., Degenhardt, F., Derks, E. M., Direk, N., Dolan, C. V., Dunn, E. C., Eley, T. C., Eriksson, N., Escott-Price, V., Kiadeh, F. H. F., Finucane, H. K., Forstner, A. J., Frank, J., Gaspar, H. A., Gill, M., Giusti-Rodríguez, P., Goes, F. S., Gordon, S. D., Grove, J., Hall, L. S., Hannon, E., Hansen, C. S., Hansen, T. F., Herms, S., Hickie, I. B., Hoffmann, P., Homuth, G., Horn, C., Hottenga, J.-J., Hougaard, D. M., Hu, M., Hyde, C. L., Ising, M., Jansen, R., Jin, F., Jorgenson, E., Knowles, J. A., Kohane, I. S., Kraft, J., Kretschmar, W. W., Krogh, J., Kutalik, Z., Lane, J. M., Li, Y., Li, Y., Lind, P. A., Liu, X., Lu, L., MacIntyre, D. J., MacKinnon, D. F., Maier, R. M., Maier, W., Marchini, J., Mbarek, H., McGrath, P., McGuffin, P., Medland, S. E., Mehta, D., Middeldorp, C. M., Mihailov, E., Milaneschi, Y., Milani, L., Mill, J., Mondimore, F. M., Montgomery, G. W., Mostafavi, S., Mullins, N., Nauck, M., Ng, B., Nivard, M. G., Nyholt, D. R., O'Reilly, P. F., Oskarsson, H., Owen, M. J., Painter, J. N., Pedersen, C. B., Pedersen, M. G., Peterson, R. E., Pettersson, E., Peyrot, W. J., Pistis, G., Posthuma, D., Purcell, S. M., Quiroz, J. A., Qvist, P., Rice, J. P., Riley, B. P., Rivera, M., Saeed Mirza, S., Saxena, R., Schoevers, R., Schulte, E. C., Shen, L., Shi, J., Shyn, S. I., Sigurdsson, E., Sinnamon, G. B. C., Smit, J. H., Smith, D. J., Stefansson, H., Steinberg, S., Stockmeier, C. A., Streit, F., Strohmaier, J., Tansey, K. E., Teismann, H., Teumer, A., Thompson, W., Thomson, P. A., Thorgeirsson, T. E., Tian, C., Traylor, M., Treutlein, J., Trubetskoy, V., Uitterlinden, A. G., Umbricht, D., Van der Auwera, S., van Hemert, A. M., Viktorin, A., Visscher, P. M., Wang, Y., Webb, B. T., Weinsheimer, S. M., Wellmann, J., Willemsen, G., Witt, S. H., Wu, Y., Xi, H. S., Yang, J., Zhang, F., eQTLGen, 23andMe, Arolt, V., Baune, B. T., Berger, K., Boomsma, D. I., Cichon, S., Dannlowski, U., de Geus, E. C. J., DePaulo, J. R., Domenici, E., Domschke, K., Esko, T., Grabe, H. J., Hamilton, S. P., Hayward, C., Heath, A. C., Hinds, D. A., Kendler, K. S., Kloiber, S., Lewis, G., Li, Q. S., Lucae, S., Madden, P. F. A., Magnusson, P. K., Martin, N. G., McIntosh, A. M., Metspalu, A., Mors, O., Mortensen, P. B., Müller-Myhsok, B., Nordentoft, M., Nöthen, M. M., O'Donovan, M. C., Paciga, S. A., Pedersen, N. L., Penninx, B. W. J. H., Perlis, R. H.,

- Porteous, D. J., Potash, J. B., Preisig, M., Rietschel, M., Schaefer, C., Schulze, T. G., Smoller, J. W., Stefansson, K., Tiemeier, H., Uher, R., Völzke, H., Weissman, M. M., Werge, T., Winslow, A. R., Lewis, C. M., Levinson, D. F., Breen, G., Børglum, A. D., Sullivan, P. F. Major Depressive Disorder Working Group of the Psychiatric Genomics Consortium. Genome-wide association analyses identify 44 risk variants and refine the genetic architecture of major depression. *Nat. Genet.* 50, 668–681 (2018).
37. Nordquist, N., Göktürk, C., Comasco, E., Eensoo, D., Merenäkk, L., Veidebaum, T., Orelund, L. & Harro, J. The transcription factor TFAP2B is associated with insulin resistance and adiposity in healthy adolescents. *Obesity (Silver Spring)* 17, 1762–1767 (2009).
 38. Wang, Z.-H., Gong, K., Liu, X., Zhang, Z., Sun, X., Wei, Z. Z., Yu, S. P., Manfredsson, F. P., Sandoval, I. M., Johnson, P. F., Jia, J., Wang, J.-Z. & Ye, K. C/EBP β regulates delta-secretase expression and mediates pathogenesis in mouse models of Alzheimer's disease. *Nature Communications* 9, 1784–16 (2018).
 39. Strohmeyer, R., Shelton, J., Lougheed, C. & Breitkopf, T. CCAAT-enhancer binding protein- β expression and elevation in Alzheimer's disease and microglial cell cultures. *PLoS ONE* 9, e86617 (2014).
 40. Apazoglou, K., Farley, S., Gorgievski, V., Belzeaux, R., Lopez, J. P., Grenier, J., Ibrahim, E. C., Khoury, El, M.-A., Tse, Y. C., Mongredien, R., Barbé, A., de Macedo, C. E. A., Jaworski, W., Bochereau, A., Orrico, A., Isingrini, E., Guinaudie, C., Mikasova, L., Louis, F., Gautron, S., Groc, L., Massaad, C., Yildirim, F., Vialou, V., Dumas, S., Marti, F., Mechawar, N., Morice, E., Wong, T. P., Caboche, J., Turecki, G., Giros, B. & Tzavara, E. T. Antidepressive effects of targeting ELK-1 signal transduction. *Nat. Med.* 24, 591–597 (2018).
 41. Leonardini, A., Laviola, L., Perrini, S., Natalicchio, A. & Giorgino, F. Cross-Talk between PPAR γ and Insulin Signaling and Modulation of Insulin Sensitivity. *PPAR Res* 2009, 818945–12 (2009).
 42. Hunger, S. P., Ohyashiki, K., Toyama, K. & Cleary, M. L. Hlf, a novel hepatic bZIP protein, shows altered DNA-binding properties following fusion to E2A in t(17;19) acute lymphoblastic leukemia. *Genes & Development* 6, 1608–1620 (1992).
 43. Shachter, N. S. Apolipoproteins C-I and C-III as important modulators of lipoprotein metabolism. *Current Opinion in Lipidology* 12, 297 (2001).
 44. TG and HDL Working Group of the Exome Sequencing Project, National Heart, Lung, and Blood Institute, Crosby, J., Peloso, G. M., Auer, P. L., Crosslin, D. R., Stitzel, N. O., Lange, L. A., Lu, Y., Tang, Z.-Z., Zhang, H., Hindy, G., Masca, N., Stirrups, K., Kanoni, S., Do, R., Jun, G., Hu, Y., Kang, H. M., Xue, C., Goel, A., Farrall, M., Duga, S., Merlini, P. A., Asselta, R., Girelli, D., Olivieri, O., Martinelli, N., Yin, W., Reilly, D., Speliotes, E., Fox, C. S., Hveem, K., Holmen, O. L., Nikpay, M., Farlow, D. N.,

- Assimes, T. L., Franceschini, N., Robinson, J., North, K. E., Martin, L. W., DePristo, M., Gupta, N., Escher, S. A., Jansson, J.-H., Van Zuydam, N., Palmer, C. N. A., Wareham, N., Koch, W., Meitinger, T., Peters, A., Lieb, W., Erbel, R., König, I. R., Kruppa, J., Degenhardt, F., Gottesman, O., Bottinger, E. P., O'Donnell, C. J., Psaty, B. M., Ballantyne, C. M., Abecasis, G. R., Ordovas, J. M., Melander, O., Watkins, H., Orho-Melander, M., Ardissino, D., Loos, R. J. F., McPherson, R., Willer, C. J., Erdmann, J., Hall, A. S., Samani, N. J., Deloukas, P., Schunkert, H., Wilson, J. G., Kooperberg, C., Rich, S. S., Tracy, R. P., Lin, D.-Y., Altshuler, D., Gabriel, S., Nickerson, D. A., Jarvik, G. P., Cupples, L. A., Reiner, A. P., Boerwinkle, E. & Kathiresan, S. Loss-of-function mutations in APOC3, triglycerides, and coronary disease. *N. Engl. J. Med.* 371, 22–31 (2014).
45. Gotto, A. M. Triglyceride as a risk factor for coronary artery disease. *Am. J. Cardiol.* 82, 22Q–25Q (1998).
46. Khetarpal, S. A., Qamar, A., Millar, J. S. & Rader, D. J. Targeting ApoC-III to Reduce Coronary Disease Risk. *Curr Atheroscler Rep* 18, 54 (2016).
47. Ulirsch, J. C., Nandakumar, S. K., Wang, L., Giani, F. C., Zhang, X., Rogov, P., Melnikov, A., McDonel, P., Do, R., Mikkelsen, T. S. & Sankaran, V. G. Systematic Functional Dissection of Common Genetic Variation Affecting Red Blood Cell Traits. *Cell* 165, 1530–1545 (2016).
48. Tewhey, R., Kotliar, D., Park, D. S., Liu, B., Winnicki, S., Reilly, S. K., Andersen, K. G., Mikkelsen, T. S., Lander, E. S., Schaffner, S. F. & Sabeti, P. C. Direct Identification of Hundreds of Expression- Modulating Variants using a Multiplexed Reporter Assay. *Cell* 165, 1519–1529 (2016).
49. Ernst, J., Melnikov, A., Zhang, X., Wang, L., Rogov, P., Mikkelsen, T. S. & Kellis, M. Genome-scale high-resolution mapping of activating and repressive nucleotides in regulatory regions. *Nat Biotechnol* 34, 1180–1190 (2016).
50. Deplancke, B., Alpern, D. & Gardeux, V. The Genetics of Transcription Factor DNA Binding Variation. *Cell* 166, 538–554 (2016).
51. Johnson, A. D., Handsaker, R. E., Pulit, S. L., Nizzari, M. M., O'Donnell, C. J. & de Bakker, P. I. W. SNAP: a web-based tool for identification and annotation of proxy SNPs using HapMap. *Bioinformatics* 24, 2938–2939 (2008).
52. Thurman, R. E., Rynes, E., Humbert, R., Vierstra, J., Maurano, M. T., Haugen, E., Sheffield, N. C., Stergachis, A. B., Wang, H., Vernot, B., Garg, K., John, S., Sandstrom, R., Bates, D., Boatman, L., Canfield, T. K., Diegel, M., Dunn, D., Ebersol, A. K., Frum, T., Giste, E., Johnson, A. K., Johnson, E. M., Kutayavin, T., Lajoie, B., Lee, B.-K., Lee, K., London, D., Lotakis, D., Neph, S., Neri, F., Nguyen, E. D., Qu, H., Reynolds, A. P., Roach, V., Safi, A., Sanchez, M. E., Sanyal, A., Shafer, A., Simon, J. M., Song, L., Vong, S., Weaver, M., Yan, Y., Zhang, Z., Zhang, Z., Lenhard, B., Tewari, M., Dorschner, M. O., Hansen, R. S., Navas, P. A., Stamatoyannopoulos,

- G., Iyer, V. R., Lieb, J. D., Sunyaev, S. R., Akey, J. M., Sabo, P. J., Kaul, R., Furey, T. S., Dekker, J., Crawford, G. E. & Stamatoyannopoulos, J. A. The accessible chromatin landscape of the human genome. *Nature* 489, 75–82 (2012).
53. Andersson, R., Gebhard, C., Miguel-Escalada, I., Hoof, I., Bornholdt, J., Boyd, M., Chen, Y., Zhao, X., Schmidl, C., Suzuki, T., Ntini, E., Arner, E., Valen, E., Li, K., Schwarzfischer, L., Glatz, D., Raithel, J., Lilje, B., Rapin, N., Bagger, F. O., Jørgensen, M., Andersen, P. R., Bertin, N., Rackham, O., Burroughs, A. M., Baillie, J. K., Ishizu, Y., Shimizu, Y., Furuhashi, E., Maeda, S., Negishi, Y., Mungall, C. J., Meehan, T. F., Lassmann, T., Itoh, M., Kawaji, H., Kondo, N., Kawai, J., Lennartsson, A., Daub, C. O., Heutink, P., Hume, D. A., Jensen, T. H., Suzuki, H., Hayashizaki, Y., Müller, F., Forrest, A. R. R., Carninci, P., Rehli, M. & Sandelin, A. An atlas of active enhancers across human cell types and tissues. *Nature* 507, 455–461 (2014).
54. Yin, Y., Morgunova, E., Jolma, A., Kaasinen, E., Sahu, B., Khund-Sayeed, S., Das, P. K., Kivioja, T., Dave, K., Zhong, F., Nitta, K. R., Taipale, M., Popov, A., Ginno, P. A., Domcke, S., Yan, J., Schübeler, D., Vinson, C. & Taipale, J. Impact of cytosine methylation on DNA binding specificities of human transcription factors. *Science* 356, eaaj2239–17 (2017).
55. Yan, J., Enge, M., Whittington, T., Dave, K., Liu, J., Sur, I., Schmierer, B., Jolma, A., Kivioja, T., Taipale, M. & Taipale, J. Transcription Factor Binding in Human Cells Occurs in Dense Clusters Formed around Cohesin Anchor Sites. *Cell* 154, 801–813 (2013).
56. Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* 26, 589–595 (2010).
57. Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y. C., Laslo, P., Cheng, J. X., Murre, C., Singh, H. & Glass, C. K. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Molecular Cell* 38, 576–589 (2010).
58. Mathelier, A., Fornes, O., Arenillas, D. J., Chen, C.-Y., Denay, G., Lee, J., Shi, W., Shyr, C., Tan, G., Worsley-Hunt, R., Zhang, A. W., Parcy, F., Lenhard, B., Sandelin, A. & Wasserman, W. W. JASPAR 2016: a major expansion and update of the open-access database of transcription factor binding profiles. *Nucleic Acids Research* 44, D110–5 (2016).
59. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv q-bio.GN*, (2013).
60. McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M. & DePristo, M. A. The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* 20, 1297–1303 (2010).

61. Zhang, Y., Liu, T., Meyer, C. A., Eeckhoute, J., Johnson, D. S., Bernstein, B. E., Nusbaum, C., Myers, R. M., Brown, M., Li, W. & Liu, X. S. Model-based Analysis of ChIP-Seq (MACS). *Genome Biol* 9, 1–9 (2008).
62. van de Geijn, B., McVicker, G., Gilad, Y. & Pritchard, J. K. WASP: allele-specific software for robust molecular quantitative trait locus discovery. *Nat. Methods* 12, 1061–1063 (2015).
63. Zhou, X., Lindsay, H. & Robinson, M. D. Robustly detecting differential expression in RNA sequencing data using observation weights. *Nucleic Acids Research* 42, e91–e91 (2014).
64. Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W. & Smyth, G. K. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research* 43, e47–e47 (2015).
65. Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M. & Gingeras, T. R. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* 29, 15–21 (2013).
66. Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D. R., Pimentel, H., Salzberg, S. L., Rinn, J. L. & Pachter, L. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat Protoc* 7, 562–578 (2012).
67. Anders, S., Pyl, P. T. & Huber, W. HTSeq—a Python framework to work with high-throughput sequencing data. *Bioinformatics* 31, 166–169 (2015).
68. Dixon, J. R., Jung, I., Selvaraj, S., Shen, Y., Antosiewicz-Bourget, J. E., Lee, A. Y., Ye, Z., Kim, A., Rajagopal, N., Xie, W., Diao, Y., Liang, J., Zhao, H., Lobanenko, V. V., Ecker, J. R., Thomson, J. A. & Ren, B. Chromatin architecture reorganization during stem cell differentiation. *Nature* 518, 331–336 (2015).
69. Durand, N. C., Robinson, J. T., Shamim, M. S., Machol, I., Mesirov, J. P., Lander, E. S. & Aiden, E. L. Juicebox Provides a Visualization System for Hi-C Contact Maps with Unlimited Zoom. *Cell Systems* 3, 99–101 (2016).
70. Edge, P., Bafna, V. & Bansal, V. HapCUT2: robust and accurate haplotype assembly for diverse sequencing technologies. *Genome Res.* gr.213462.116 (2016). doi:10.1101/gr.213462.116
71. Browning, S. R. & Browning, B. L. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *The American Journal of Human Genetics* 81, 1084–1097 (2007).
72. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* 15, 550–21 (2014).

CHAPTER 2: Common DNA sequence variation influences 3-dimensional conformation of the human genome

2.1 Abstract

The 3-dimensional (3D) conformation of chromatin inside the nucleus is integral to a variety of nuclear processes including transcriptional regulation, DNA replication, and DNA damage repair. Aberrations in 3D chromatin conformation have been implicated in developmental abnormalities and cancer. Despite the importance of 3D chromatin conformation to cellular function and human health, little is known about how 3D chromatin conformation varies in the human population, or whether DNA sequence variation between individuals influences 3D chromatin conformation. To address these questions, we performed Hi-C on Lymphoblastoid Cell Lines (LCLs) from 20 individuals. We identified thousands of regions across the genome where 3D chromatin conformation varies between individuals and found that this conformational variation is often accompanied by variation in gene expression, histone modifications, and transcription factor (TF) binding. Moreover, we found that DNA sequence variation influences several features of 3D chromatin conformation including loop strength, contact insulation, contact directionality and density of local cis contacts. We mapped hundreds of Quantitative Trait Loci (QTLs) associated with 3D chromatin features and found evidence that some of these same variants are associated at modest levels with other molecular phenotypes as well as complex disease risk. Our results demonstrate that common DNA sequence variants can influence 3D chromatin conformation, pointing to a more pervasive role for 3D chromatin conformation in human phenotypic variation than previously recognized.

2.2 Introduction

3-dimensional (3D) organization of chromatin is essential for proper regulation of gene expression¹⁻³, and plays an important role in other nuclear processes including DNA replication^{4,5}, X chromosome inactivation⁶⁻⁸, and DNA repair^{9,10}. Many recent insights about 3D chromatin conformation have been enabled by a suite of technologies based on Chromatin Conformation Capture (3C)¹¹. A high-throughput version of 3C called “Hi-C” enables the mapping of 3D chromatin conformation at genome-wide scale¹², and has revealed several key features of 3D chromatin conformation including: 1) compartments (often referred to as “A/B compartments”), which refer to the tendency of loci with similar transcriptional activity to physically segregate in 3D space¹²⁻¹⁴, 2) chromatin domains (often referred to as Topologically Associating Domains, or TADs) demarcated by sharp boundaries across which contacts are relatively infrequent¹⁵⁻¹⁷, 3) chromatin loops, which describe point-to-point interactions that occur more frequently than would be expected based on the linear distance between interacting loci, and often anchored by convergent CTCF motif pairs¹³, and 4) Frequently Interacting Regions (FIREs), which are regions of increased local interaction frequency enriched for tissue-specific genes and enhancers^{18,19}.

Previous studies have used Hi-C to profile 3D chromatin conformation across different cell types^{13,15,20}, different primary tissues¹⁸, different cell states²¹, and in response to different genetic and molecular perturbations²²⁻²⁶, producing a wealth of knowledge about key features of 3D chromatin conformation. However, to our knowledge no study to date has measured variation in 3D chromatin conformation across more than a handful of unrelated individuals. Several observations demonstrate that at least in some

cases DNA sequence variation between individuals can alter 3D chromatin organization with pathological consequences²⁷. Pioneering work by Mundlos and colleagues described several cases in which rearrangements of TAD structure lead to gene dysregulation and consequent developmental malformations^{28,29}. In cancer, somatic mutations and aberrant DNA methylation can disrupt TAD boundaries leading to dysregulation of proto-oncogenes^{30,31}. Moreover, many genetic variants associated with human traits by GWAS occur in distal regulatory elements that loop to putative target gene promoters in 3D, and in some cases, the strength of these looping interactions has been shown to vary between alleles of the associated SNP^{32,33}. Although these studies demonstrate that both large effects as well as more subtle aberrations of 3D chromatin conformation are potential mechanisms of disease, population-level variation in 3D chromatin conformation more broadly has remained unexplored.

In the present study, we set out to characterize inter-individual variation in 3D chromatin conformation by performing Hi-C on Lymphoblastoid Cell Lines (LCLs) derived from individuals whose genetic variation has been cataloged by the HapMap or 1000 Genomes Consortia³⁴. LCLs have been used as a model system to study variation in several other molecular phenotypes including gene expression, histone modifications, transcription factor (TF) binding, and chromatin accessibility³⁵⁻⁴¹. These previous efforts provide a rich context to explore variation in 3D chromatin conformation identified in this model system. Through integrative analyses, we found that inter-individual variation in 3D chromatin conformation occurs on many levels including compartments, TAD boundary strengths, FIREs, and looping interaction strengths. Moreover, we found that variation in 3D chromatin conformation coincides with variation in activity of the underlying genome

sequence as evidenced by transcription, histone modifications, and TF binding. Although our sample size is small, we observe reproducible effects of DNA sequence variation on 3D chromatin conformation and identify hundreds of Quantitative Trait Loci (QTLs) associated with multiple features of 3D chromatin conformation. Our results demonstrate that variation in 3D chromatin conformation is readily detectable from Hi-C data, often overlaps with regions of transcriptomic and epigenomic variability, and is influenced in part by genetic variation that may contribute to disease risk.

2.3 Results

Mapping 3D chromatin conformation across individuals

To generate maps of 3D chromatin conformation suitable for comparison across individuals, we performed “dilution” Hi-C on LCLs derived from 13 Yoruban individuals (including one trio), one Puerto Rican trio, and one Han Chinese trio (19 individuals total). We also include published Hi-C data from one European LCL (GM12878) generated previously by our group using the same protocol⁴², for a total of 20 individuals from four different populations. Many of these same LCLs have been used in previous genomic studies^{37,39,41}, allowing us to leverage multiple transcriptomic and epigenomic datasets in our analysis below. Importantly, 18 of these individuals have had their genetic variation cataloged by the 1000 Genomes Consortium^{34,43}, which allowed us to examine the influence of genetic variation on 3D chromatin conformation. Two replicates of Hi-C were performed on each LCL, with each replicate performed on cells grown independently in culture for at least two passages.

All Hi-C data were processed using a uniform pipeline that incorporates the WASP approach^{39,44} to eliminate allelic mapping biases. For each sample, we mapped a series of well-established Hi-C-derived features including 40Kb resolution contact matrices, Directionality Index (DI)¹⁵, Insulation score (INS)⁷, and compartmentalization¹² (**Figure 2.1a; Figure S2.1a-c**). Compartmentalization is measured by the first Principal Component (PC1) of Hi-C contact matrices, and thus we use the acronym “PC1” below to refer to this measure of compartmentalization. We also identified regions known as Frequently Interacting Regions (FIREs)¹⁸ and their corresponding “FIRE scores”, which measure how frequently a given region interacts with its neighboring regions (15~200kb).

The concept of FIRE is based on the observation that the frequency of contacts at this distance is not evenly distributed across the genome, but rather, tends to peak in regions showing epigenomic signatures of transcriptional and regulatory activity (**Figure S2.2**). As we have shown previously^{18,19}, FIRE regions often overlap putative enhancer elements (**Figure S2.1d-e**). We did not call “chromatin loops” in this study because our data was not of sufficient resolution, but we use a set of loops called previously in the LCL GM1287814 to examine variation in loop strength among the LCLs in our study. Aggregate analysis shows that these published LCL loops are generally reproduced in our data (**Figure S2.3**).

3D chromatin conformation variations between individuals

After uniformly processing all Hi-C data, we compared chromatin conformation across LCLs at the level of contact matrices and multiple derived features (PC1, DI, INS, and FIRE). From a genome-wide perspective, each of these 3D chromatin features shows a signature consistent with reproducible inter-individual variation whereby replicates from the same individual (i.e. same LCL) are more highly correlated than datasets from different individuals (PC1 $p=2.4e-7$, INS $p=1.6e-7$, DI $p=3.3e-7$, FIRE $p=0.0157$ by Wilcoxon rank sum test; **Figure 2.1b-d**, **Figure S2.4a-f**). The Hi-C data also cluster by population (**Figure S2.4f-g**) consistent with an influence from genetic background, but we note that this population-level clustering can be caused by other factors such as batch of sample acquisition⁴⁵.

Despite generally high correlations of Hi-C data across individuals, we frequently observed regions where 3D chromatin conformation varies reproducibly between individuals (example shown in **Figure 2.2a**, **Figure S2.5a**). To more systematically

identify regions of variable 3D chromatin conformation, we used the “limma” package⁴⁶ to identify regions where variation between individuals was more significant than variation between two replicates from the same individual. We applied this approach to DI, INS, FIRE, and PC1. For each metric, we first defined a set of testable 40kb bins across the genome by filtering out bins with low levels of signal across all individuals or near structural variants that can appear as aberrations in Hi-C maps⁴⁷. We then applied a False Discovery Rate (FDR) threshold of 0.1 and merged neighboring variable bins, resulting in the identification of 2,318 variable DI regions, 2,485 variable INS regions, 1,996 variable FIRE regions, and 7,732 variable PC1 regions (**Figure 2.2b, Figure S2.5b**). We note that there is strong overlap between the variable DI, INS, FIRE, and PC1 regions detected across all 20 LCLs and those detected using only the 11 unrelated YRI LCLs, which suggests that potential confounding effects of variation between different populations are not driving the identification of these variable regions (**Figure S2.5c**). Although each metric has a unique set of testable bins, we found significant enrichment for bins that are variable in more than one metric (**Figure 2.2c, Figure S2.5d-e**), indicating that the same regions often vary across multiple features of 3D chromatin conformation.

We next used Fluorescent In Situ Hybridization (FISH) to examine whether variable regions detected by Hi-C are consistent with distance measurements from imaging data (**Figure 2.2d-e**). Focusing on a variable DI region on chromosome 15 (chr15:96720000-96920000; hg19), we performed FISH in LCLs from four individuals with different levels of DI at the variable region being evaluated (YRI-3, YRI-4, YRI-5, YRI-8). We used three BAC probes that hybridize respectively to the variable DI region (“center”, probe covers chr15:96715965-96898793), a region approximately 668Kb upstream

(“upstream”, probe covers chr15:95897555-96047720), or a region approximately 590Kb downstream (“downstream”, probe covers chr15:97488414-97648104). We found that distances between the center probe and these flanking probes vary significantly between individuals with strong upstream contact bias as measured by DI (YRI-4, YRI-8) and individuals without this upstream contact bias (YRI-3, YRI-5)(**Figure 2.2d-e**, center-upstream distance $p=0.017$, center-downstream distance $p=1.7e-5$ by Wilcoxon rank sum test). Moreover, we found that the center probe is closer to the upstream than the downstream probe in the two individuals with strong upstream DI signal at the central variable DI region ($p=3.2e-3$ for YRI-3, $p=1.5e-4$ for YRI-5 by Wilcoxon rank sum test). However, this trend is reversed in individuals without upstream DI signal where the center probe is now closer to the downstream probe ($p=0.021$ for YRI-4, $p=0.1$ for YRI-8 by Wilcoxon rank sum test) (**Figure S2.6a**).

We also sought to identify variable entries in the Hi-C contact matrix itself (“matrix cells”). To facilitate this search, we used a method called Bandwise Normalization and Batch Correction (BNBC) that we recently developed to normalize Hi-C data across individuals (Fletez-Brant et al. Pre-print: <https://doi.org/10.1101/214361>). BNBC takes contact distance into account as a co-variate because batch effects in Hi-C data can be distance-dependent. To identify variable matrix cells, we performed a variance decomposition on Hi-C contact matrix cells which exhibited statistically-significant variability between individuals, resulting in a measure of biological variability for each bin in the contact matrix (see example in **Figure 2.2a** and **Figure S2.5a**). To identify matrix cells with significant levels of biological variability, we estimated FDR using the IHW framework⁴⁸ to include the distance between anchor bins as an informative covariate. At

an FDR threshold of 0.1, we identified 115,817 matrix cells showing significant variability between samples. These variable bins are heavily skewed toward shorter contact distances (**Figure 2.2f, Figure S2.6b**), likely due in part to higher read counts and thus increased power at these distances. We observed that the anchor regions of variable matrix cells overlap with variable regions of DI, INS, FIRE, and PC1 more often than would be expected by chance (**Figure 2.2g; Figure S2.6c**). We also observed that variable matrix cells tend to occur in groups (**Figures 2.2a, 2.3a**), suggesting that variation in 3D chromatin conformation often affects more than one adjacent genomic window.

Coordinated variation of the 3D genome, epigenome, and transcriptome

To investigate the relationship between variation in 3D chromatin conformation and gene regulation, we analyzed multiple published datasets including RNA-seq, ChIP-seq, and DNase-seq data generated from some of the same LCLs in our study. Strikingly, for all external datasets examined here, we see an enrichment for regions at which 3D chromatin conformation across individuals is correlated with measures of genome activity in the same 40Kb bin (see example in **Figure 2.3a** and **Figure S2.7a**). To assign a level of statistical significance to these observations, we approximated the null distribution by randomly permuting the sample labels of external datasets, thus disrupting the link between Hi-C and ChIP/RNA/DNase-seq data from the same individual, but not changing the underlying data structure (see schematic in **Figure S2.7b**). We used these permutations to calculate the bootstrap p-values in **Figure 2.3b**. Among variable PC1 regions, we observed a significant enrichment for regions at which PC1 values across individuals are positively correlated with histone modifications indicative of transcriptional

activity including H3K27ac (bootstrap $P < 0.001$), H3K4me1, and H3K4me3 (but notably less so with H3K27me3, which is marker of transcriptional repression) (bootstrap $P < 0.001$ for all histone modifications, **Figure 2.3b**). The correlations between PC1 and marks of transcriptional activity occur in the expected direction -- i.e. higher PC1 values are associated with higher gene expression and more active histone modifications. Similar correlations were apparent in two distinct sets of ChIP-seq data generated by different groups^{39,41}, and observed whether we use variable regions identified across all 20 LCLs or only across the 11 unrelated YRI LCLs (**Figure S2.7c**).

The relationship between variation in 3D chromatin conformation and underlying genome activity extends beyond A/B compartmentalization. At variable FIRE regions, we found an abundance of regions where FIRE score is positively correlated with marks of cis-regulatory activity including H3K27ac and H3K4me1 (Bootstrap $P < 0.001$; **Figure 2.3b, Figure S2.9a**), consistent with the previously reported relationship between FIREs and cis-regulatory activity^{18,19}. DI and INS values at variable regions tend to be correlated histone modification levels as well as CTCF and Cohesin subunit SA1 binding (Bootstrap $P < 0.001$; **Figure 2.3b, Figure S2.8a-b**), which are known to influence these 3D chromatin features^{15,49,50}. For INS, the relationship is directional such that higher CTCF/Cohesin binding corresponds to more contact insulation (i.e. lower INS score). However, at variable DI regions the correlations are not as clearly directional, reflecting current understanding that the direction of DI (i.e. upstream vs downstream contact bias) is arbitrary relative to strength of CTCF/Cohesin binding. We performed similar analysis on variable cells in the contact matrix, and found that the interaction frequency in these matrix cells across individuals tends to be correlated with epigenetic or transcriptional properties

of one or both corresponding “anchor” bins (Bootstrap $P < 0.001$; **Figure 2.3b**, **Figure S2.9b**). Importantly, for all types of variable regions examined here we found correlation with RNA-seq signal, indicating that at least at some regions, variation in 3D chromatin features accompanies variation in gene expression.

We examined further whether 3D chromatin conformation at a given variable region tends to be correlated with only one epigenomic property, or with several properties simultaneously. We found that PC1, FIRE, INS, and DI values across individuals are often correlated with multiple features of active regions (e.g. H3K27ac, H3K4me1, RNA), and anti-correlated with the repressive H3K27me3 histone modification (**Figure 2.3c, d**). For DI, where direction is not as clearly linked to magnitude of gene regulatory activity, we note a larger set of regions with anti-correlation to features of active regions (e.g. H3K27ac, H3K4me1, RNA) and positive correlation with H3K27me3 (**Figure 2.3e, f**). These results demonstrate that variation in 3D chromatin conformation is often accompanied by variation in transcriptional and regulatory activity of the same region. Moreover, the correlations between multiple molecular phenotypes at the same region suggest that shared mechanism(s) underlie variation in these phenotypes across individuals.

Genetic loci influencing 3D chromatin conformation

To examine genetic influence on 3D chromatin conformation we first considered genetic variants overlapping CTCF motifs at chromatin loop anchors¹³, because disruption of these CTCF motifs by genome engineering has been shown to alter chromatin looping²². Focusing on SNPs at variation-intolerant positions in anchor CTCF motifs (“anchor disrupting SNPs”, at sequence weight matrix positions where a single

base has a probability of >0.75 , **Figure 2.4a**), we observed a significant linear relationship between SNP genotype and the strength of corresponding loops ($p=7.6e-5$ by linear regression; **Figure 2.4b,c**). We also examined whether individuals heterozygous for anchor disrupting SNPs showed allelic imbalance in loop strength. To facilitate this analysis, we used the HaploSeq⁴² method to generate chromosome-span haplotype blocks for each LCL. Although few Hi-C read pairs overlap a SNP allowing haplotype assignment (mean 7.89% of usable reads per LCL), we do observe that the haplotype bearing the stronger motif allele tends to show more reads connecting the corresponding loop anchors ($p=5.9e-4$ by one-sided t-test of mean > 0.5 ; **Figure 2.4d**). Our observation that CTCF motif SNPs can modulate 3D chromatin conformation is consistent with similar findings reported from ChIA-PET data⁵¹, and a recent report of haplotype-associated chromatin loop published while this manuscript was in preparation²⁶.

Motivated by these preliminary observations of genetic effects on 3D chromatin conformation, we next searched directly for QTLs associated with Hi-C derived features of 3D chromatin conformation. Power calculations indicated that, despite limited sample size, we were moderately powered to find QTLs with strong effect sizes using a linear mixed effect model (LMM) approach that takes advantage of the Hi-C replicates for each LCL. Thus, we conducted a targeted search for QTLs associated with variation in FIRE, DI, INS, and contact frequency. We did not include PC1 in the QTL search because we reasoned that individual genetic variants would be more likely to have detectable effects on local chromatin conformation rather than large-scale features like compartmentalization. For this same reason, we used modified versions of DI and INS scores for the QTL search calculated with a window size of 200Kb upstream and

downstream of the target bin, rather than the standard 2Mb window size for DI¹⁵ or 480Kb for INS⁸. We also limited our QTL searches to the 11 unrelated YRI individuals in our study (referred to below as the “discovery set”) to mitigate potential confounding differences between populations.

For each 3D genome phenotype under study we identified a list of testable bins that showed appreciable levels of signal in at least one individual in our discovery set. We also identified a set of test SNPs that includes at most one tag SNP among those in perfect LD in each 40Kb bin. Response variables (i.e. 3D chromatin phenotype values) were quantile normalized across the discovery set. For each testable bin, we measured the association of the given 3D chromatin phenotype with all test SNPs in that bin. In cases where multiple SNPs in the same bin were significantly associated with the phenotype, we selected only the most significantly associated SNP per bin for our final QTL list. Ultimately, at an FDR of 0.2, we identified 387 FIRE-QTLs (i.e. testable bins in which FIRE score is associated with at least one SNPs in that bin; comprising 6.6% of tested bins), 545 DI-QTLs (4.2% of tested bins), and 911 INS-QTLs (12.0% of tested bins)(**Figure 2.4e, Figure S2.10a**). For analysis of DI-QTLs, we separated the testable bins into those with upstream bias and those with downstream, because we observed a Simpson's paradox when we analyzed the genotype trend at all DI-QTL regions together (**Figure S2.10b**).

We also searched for QTLs associated directly with interaction frequency in individual contact matrix cells using an LMM approach like that described above for FIRE, DI, and INS. The large number of cells in a Hi-C contact matrix, together with limited sample size, made a true genome-wide QTL search unfeasible. However, power

calculations indicated that if we limited our QTL search to a subset of cells in the matrix, we could have moderate power to detect strong genetic signal. Thus, we limited our QTL search for contact matrix QTLs (“C-QTLs”) to matrix cells that showed significant biological variability in our samples, as described above. We tested for association in our discovery set between the BNBC-normalized interaction frequency in these variable matrix cells and the genotype of test SNPs in either of the two anchor bins. We selected at most one QTL SNP per matrix cell, using association p-value to prioritize, finally yielding 345 C-QTL SNPs associated with 463 matrix cells at an IHW-FDR threshold of 0.2 (**Figure 2.4f**).

To evaluate the reproducibility of each of these QTLs sets (FIRE-QTLs, DI-QTLs, INS-QTLs, and C-QTLs), we examined Hi-C data from 6 individuals who were not included in our discovery set (we refer to these 6 individuals our “validation set”). These individuals represent four different populations (CEU, PUR, CHS, YRI), and they include a child of two individuals in the discovery set (YRI-13/NA19240 is child of YRI-11/NA19238 and YRI-12/NA19239). In each case, we find a significant linear relationship in the validation set between QTL genotype and the corresponding 3D chromatin phenotype ($p=1.8e-14$ for FIRE-QTLs, $p=2.5e-7$ for DI-QTLs at positive DI bins, $p=0.008$ for DI-QTLs at negative DI bins, $p=3e-4$ for INS-QTLs, $p=4.1e-9$ for C-QTLs; **Figure 2.4g**). To provide an additional and more stringent estimate of the significance of these observations, we performed permutations by randomly selecting sets of test SNPs and measuring the linear relationship between genotype and phenotype in the validation set. In all cases, the observed relationship was also significant by this more conservative bootstrap approach ($p<0.001$ for FIRE-QTLs, $p<0.001$ for DI-QTLs at positive DI bins,

$p=0.041$ for DI-QTLs at negative DI bins, $p=0.005$ for INS-QTLs, $p=0.006$ for C-QTLs; **Figure 2.4h**).

There is little direct overlap between our different QTL sets (**Figure S2.10c**), likely due to limited power and the fact that the testable bins were different for each metric. However, we observed genotype-dependent INS score at FIRE-QTLs and C-QTLs, and genotype-dependent FIRE score at INS-QTLs and DI-QTLs (**Figure S2.10d**), which suggested that overlapping signal between different types of 3D chromatin QTLs in present below the level of test-wide significance. To more rigorously assess overlapping signal between our QTL sets we examined shared association below the threshold of multiple test correction, inspired by similar approaches reported elsewhere⁵². Our underlying hypothesis is that genetic association studies of two different phenotypes “X” and “Y” with overlapping (or partially overlapping) genetic architecture may have few direct overlaps between significant hits due to limited power or differing study designs, but the shared signal should become apparent when the full range of association results are considered. To quantify this, we calculated the fraction of QTLs for a given phenotype X that exceed a nominal level of significance ($p < 0.05$) when tested for association with a different phenotype Y. We refer to this value as the “nominal fraction” below and in **Figure 2.4i**. To test whether the nominal fraction of X-QTLs was significantly higher than would be expected by chance, we approximated the null distribution by calculating nominal fractions for 10,000 sets of SNPs selected randomly from among all X test SNPs. In almost all pairwise comparisons between 3D chromatin QTL types examined here, we find that the observed nominal fractions are significantly higher than would be expected in the absence of shared genetic architecture (**Figure 2.4i, j**).

Contribution of 3D chromatin QTLs to molecular phenotypes and disease risk

Given the correlation observed between 3D chromatin variation and epigenome variation, we next investigated whether 3D chromatin QTLs could modulate both the epigenome and 3D genome. Here, we made use of published ChIP-seq data for histone modifications (H3K4me1, H3K4me3, H3K27ac) in a large set of 65 YRI LCLs³⁸, DNase-seq data from 59 YRI LCLs³⁷, and CTCF ChIP-seq data from 15 CEU LCLs⁵³. Notably, most individuals in these datasets were not included in our QTL discovery or validation sets (54/65 for histone modification ChIP-seq, 48/59 for DNase-seq, 15/15 for CTCF ChIP-seq). In many cases, we found a significant linear relationship between 3D chromatin QTL genotypes and these different epigenetic phenotypes (**Figure 2.5a**, **Figure S2.11a**). For example, at FIRE-QTLs, the high-FIRE allele is also associated with higher levels of active histone modifications and chromatin accessibility (**Figure 2.5a**). We note that although these associations are all significant by linear regression, only H3K27ac and H3K4me1 passed more conservative permutation testing in which the null distribution is approximated by selecting random SNPs from the full set of tested SNPs (**Figure 2.5b**). At C-QTLs, the high-contact alleles show higher levels of the enhancer-associated mark H3K4me1 in the two anchor bins that connect the corresponding matrix cell. Moreover, the nominal fraction of C-QTLs (i.e. fraction of c-QTLs with $p < 0.05$) in a published set of H3K4me1-QTLs is significantly higher than expected in the absence of shared genetic association ($p = 6.9 \times 10^{-6}$ by chi square test, bootstrap $p = 0.028$; **Figure S2.11b,d**). At INS-QTLs, the slope of these genotype-phenotype relationships is inverted such that higher levels of histone modifications and chromatin accessibility are associated with the low INS score allele (i.e. more contact insulation), although only the association

with chromatin accessibility is significant by both linear regression and permutation test ($p=1.6e-40$ by linear regression, bootstrap $p=0.023$; **Figure S2.11b,d**). The genotype-phenotype relationships observed at DI-QTLs are not as clear as for other metrics (**Figure 2.5b, Figure S2.11a**), but this is expected because increased histone modifications or chromatin accessibility can influence DI in either direction, potentially confounding this type of aggregate analysis. Anecdotally, we do observe examples of individual DI-QTLs where genotype appears to correlate with epigenomic phenotype (**Figure 2.5c**).

Finally, we sought to examine whether 3D chromatin QTLs might contribute risk for complex diseases. There are 44 direct overlaps between our 3D chromatin QTLs (or SNPs in perfect LD in the same 40Kb bin) and NHGRI-EBI GWAS⁵⁴. However, the significance of these direct overlaps is hard to assess given the differences between the populations and study designs in question. Thus, here again we examined overlaps below the level of genome-wide significance by looking at nominal fractions to assess shared signal between association studies. We compiled full summary statistics for large GWAS (>50,000 individuals) of the related immune-relevant phenotypes Crohn's Disease (CD), Ulcerative Colitis (UC), and Inflammatory Bowel Disease (IBD)⁵⁵, as well as studies of the non-immune phenotypes height⁵⁶ and Body Mass Index (BMI)⁵⁷. We observed striking enrichments for INS-QTLs among variants with nominal associations to UC and IBD risk (1.67- and 1.65-fold, respectively), and these enrichments are significant by both chi square and permutation tests (INS-QTL with UC chi square $p=2.5e-16$ and bootstrap $p=0.024$; INS-QTL with IBD chi square $p=5.5e-17$ and bootstrap $p=0.018$; **Figure 2.5d,e**). We also note a trend in which FIRE-QTLs show nominal association with UC and IBD (1.36- and 1.58-fold enrichment, respectively), although these observations fall just below

the threshold of significance by the more stringent permutation test (FIRE-QTL with UC chi square $p= 7.6e-6$ and bootstrap $p=0.090$; FIRE-QTL with IBD chi square $p= 4.2e-8$ and bootstrap $p=0.056$; **Figure 2.5d,f**).

2.4 Discussion

Our results provide the first systematic characterization of how chromatin conformation varies between unrelated individuals at the population level, and as a consequence of genetic variation. The most important finding of our study is that genetic variation influences multiple features of 3D chromatin conformation and does so to an extent that is detectable even with limited sample size and Hi-C resolution. To the best of our knowledge, this represents the first report of QTLs directly associated with 3D chromatin conformation. However, there are limitations to our QTL search that are important to note here. First, the small sample size means that our power to detect QTLs is limited, and in order to identify QTL sets that could be analyzed in aggregate we tolerated elevated type I error by using an FDR threshold of 0.2 (as done previously for molecular QTL studies with limited power³⁹). Second, the limited resolution of our Hi-C data (40Kb) and extensive LD in our study population prevented us from identifying specific causal variant(s) for validation through genetic perturbation experiments. Nonetheless, we were able to validate the 3D chromatin QTL sets through aggregate analysis of Hi-C data from a small set of individuals who were not included in the QTL search, and with independently generated ChIP-seq and DNase-seq data from a larger set of individuals. Taken together, our results show that genetic variation influences several features of 3D chromatin conformation, which is an important step forward to evaluate the role of 3D chromatin conformation in mediating disease risk.

Another key finding of our study is that regions which vary in 3D chromatin conformation across individuals also tend to vary in measures of transcriptional and regulatory activity. This supports the existence of shared mechanisms that underlie

variation in 3D chromatin conformation, transcription, and epigenomic properties. We suspect that no single mechanism or causal hierarchy applies to all regions of the genome with variation in one or more of these properties. However, in at least some cases, this shared mechanism is likely genetic. This raises the question of whether 3D chromatin QTLs are fundamentally the same as QTLs previously described for other molecular phenotypes (e.g. eQTLs, dsQTLs, histoneQTLs; collectively referred to below as “molQTLs”), or represent a separate set of QTLs not detectable with other methods. This question is difficult to answer in the present study for two main reasons: 1) Our power is limited and thus we cannot say with confidence that a given SNP is not a 3D chromatin QTL. Many molQTL studies also have limited power and are thus prone to type II error. 2) Our QTL searches, like most molQTL studies, are not truly genome-wide because subsets of testable regions and testable SNPs are preselected to focus the search space. These selection criteria can differ widely between studies, making direct QTL-to-QTL comparisons challenging. The observation of genotype dependent epigenetic signal at 3D chromatin QTLs suggest that at least some 3D chromatin QTLs could also be detected as other types of molQTLs if those studies had sufficient statistical power. However, the limited overlap between 3D chromatin QTLs and published molQTLs (even when considering SNPs with only a nominal level of significance) points to a lack of power in current studies, and suggests further that the QTLs with largest effects on 3D chromatin conformation are not necessarily the same as those with large effects on other molecular phenotypes, and vice versa. Therefore, it is likely that QTL studies directed toward different types of molecular phenotypes (including 3D chromatin features) are likely to be complimentary rather than redundant.

Future studies with higher resolution Hi-C data and larger sample sizes will be important to identify functional variants modulating 3D chromatin conformation, and to further dissect the mechanistic relationships between genetics, 3D chromatin conformation, and other molecular phenotypes. We anticipate that these studies will continue to reveal cases in which perturbation of 3D chromatin conformation is a molecular mechanism through which disease-associated genetic variants confer disease risk. The present study provides initial discoveries of genetic influence on 3D chromatin conformation and an analytical framework and that we believe will facilitate future efforts to unravel the molecular basis of genetic disease risk.

2.5 Methods

Hi-C data generation

Hi-C was performed as previously described¹². We note that all Hi-C experiments were performed using a “dilution” HindIII protocol, rather than the newer “in situ” version of the protocol, for consistency because data generation began before the invention of in situ Hi-C. In addition, the resolution of 40kb used here for most analysis was determined primarily by sequencing depth rather than choice of a restriction enzyme. Thus, even if a 4-cutter like Mbol had been used, the prohibitive cost of sequencing would have prevented us taking advantage of the additional possible resolution.

Hi-C data processing

Alignment with WASP Read ends were aligned to the hg19 reference genome using BWA-MEM⁵⁸ v0.7.8 as single-end reads with the following parameters: -L 13,13. We used the WASP pipeline^{39,44} to control for potential allelic mapping biases, which some modifications to account for unique aspects of Hi-C data. BWA-MEM can produce split alignments where different parts of a read are aligned to different parts of the genome. This is critical for Hi-C data, because a read can span a Hi-C ligation junction between two interacting fragments. In the case of a split alignment, BWA-MEM will mark the higher-scoring alignment as the primary alignment. For Hi-C data this is not ideal – we want the five-prime-most alignment (before the ligation junction) to be the primary alignment. To account for this, we further processed the alignments from BWA-MEM to select the five-prime-most alignment in cases where one read was split. Reads without an alignment to the five-prime end of the read were filtered out, as were alignments with low mapping quality (<10). The WASP pipeline was then used to generate alternative

reads by flipping the allele in reads overlapping SNPs, and these reads were then realigned using the same pipeline. As input to WASP, we included all SNPs and indels present in the PUR individuals in our set (HG00731, 732, 733), CHS individuals in 1000 genomes (we included all CHS to account for the fact that no 1000 genomes genotype calls were available for HG00514), YRI individuals in 1000 genomes (we included all YRI individuals to account for the fact that no 1000 genomes genotype calls were available for GM19193), and the H1 cell line²¹ (to facilitate uniform processing and comparisons between LCLs and H1-derived datasets). After alignment of the alternative reads, alignment of the original reads and alternative reads were compared by WASP, and only the original reads for which all alternative reads aligned at the same location with same CIGAR string were kept. Reads overlapping indels were removed. Reads were then repaired, and only pairs in which both reads survived this filtering were kept. PCR duplicates were removed using Picard tool (<http://broadinstitute.github.io/picard/>) with default parameters. To ensure that our adapted WASP pipeline removed allelic mapping biases effectively, we simulated all possible 100bp single end reads spanning SNPs in our LCLs and aligned them back to the genome using our adapted WASP pipeline. We found no SNPs which depart from 50/50 mapping ration between reference and alternative allele in these simulations.

We also took steps to remove any potential artifacts due to HindIII polymorphisms. Hi-C data was obtained by cutting the genome with HindIII, so we reasoned that SNPs or indels that disrupt existing HindIII sites or create novel HindIII sites could lead to differential cutting of two alleles and thus the appearance of differential contact frequency. To mitigate these potential artifacts, we identified all HindIII sites that would be disrupted

or created by genetic variants present in our samples, and removed all reads within 1Kb of these polymorphisms in all individuals.

Contact Matrix Calculations Matrices were generated and normalized as previously described²⁰. Briefly, intra-chromosomal read pairs were divided into 40Kb bin pairs based on five prime positions. The number of read pairs connecting each pair of 40Kb bins were tallied to produce contact matrices for each chromosome. Raw counts in the contact matrices were then normalized using HiCNorm⁵⁹ to correct for known sources of bias in Hi-C contact matrices (GC content, mappability, fragment length). Bins that are unmappable (effective fragment length, GC content or mappability is 0) were assigned NA values. These normalized matrices were further quantile normalized across samples to account for differing read depths and mitigate potential batch effects. One such quantile normalized matrix was generated for each chromosome in each replicate, as well as in each sample (replicates pooled together). We eliminated chromosomes X and Y from all downstream analyses due to the gender differences between our samples.

PC1 Score PC1 scores were computed using methods defined previously¹². Briefly, quantile normalized matrices for each chromosome were transformed to Observed/Expected (O/E) matrices by dividing each entry in the matrix by the expected contact frequency between regions in that matrix at a given genomic distance. For a given matrix, the expected contact frequencies were computed by averaging contact frequencies at the same distance in that each matrix. The O/E matrices were further transformed to Pearson correlation matrices by the “cor” function in R and eigen vectors (principal components) were computed using the “cov” function in R. Generally, the first eigenvector (“PC1”) reflects A/B compartmentalization. However, for some chromosomes

we have seen that the second eigenvector sometimes reflects compartmentalization, while the first eigenvector reflects other features like the two chromosome arms. To systematically account for this effect, we examined the first three eigenvectors for each chromosome in each replicate by correlating them with the gene density (compartmentalization is correlated with gene density, while other properties like chromosome arms generally are not). We required that PC1 show the highest correlation with gene density among the first three eigenvectors in every replicate. If this was not the case for a given chromosome, we eliminated that chromosome from all downstream analyses in all individuals to be conservative. Six chromosomes were eliminated in this way: chr1, chr9, chr14, chr19, chr21 and chr22. For the chromosomes that passed this filter, the sign of the first eigenvector (which is arbitrary) was adjusted such that the correlation between PC1 and gene density is positive, and this positive PC1 values correspond to compartment A. Finally, PC1 tracks were manually inspected to ensure that they are consistent with expected checkerboard patterns of compartmentalization.

Directionality Index Directionality Index was computed as previously described¹⁵. Briefly, upstream and downstream contacts within 2Mb window for each 40Kb bin were counted, and chi-square statistics were calculated under equal assumption. The sign of the chi-square statistics was adjusted such that positive values represent upstream biases. For some bins, there are more than five NA bins within 2Mb window and DI for those bins are not calculated. As noted in the main text, we made a slight variation of these DI scores for the QTL searches in which DI was recalculated using a window size of 200Kb to capture more local features.

Insulation Score Insulation scores were computed as previously described⁷ with

some adjustments. Briefly, contacts linking upstream and downstream 400Kb windows for each 40Kb bin were calculated in the O/E matrices instead of raw matrices. We further divided the contact frequency by the average of upstream and downstream 400Kb windows, to account for differences in contact density across the chromosome. The Insulation Scores were then ranged from 0 to 1, representing absolute insulation and no insulation respectively. Insulation scores for bins, for which more than 50% cells in the 400Kb window as NA values, were not computed. For the QTL search, we also calculated insulation scores using 200Kb window.

TADs Calling TADs were called using the same approach as described previously¹⁵. DI values for each 40Kb bins were used to build a Hidden Markov Model and predict the probability being upstream bias, no bias, and downstream bias. Regions switching from upstream bias to downstream bias were called as boundaries.

FIRE We first calculated FIRE score for each of 20 individuals, as described in our previous study¹⁸. Specifically, we mapped the raw reads to the reference genome hg19 as described above. Next, we removed all intra-chromosomal reads within 15Kb, and created 40Kb raw Hi-C contact matrix for each individual for each autosome. For each 40Kb bin, we calculated the total number of intra-chromosomal reads in the distance range of 15-200Kb. We then filtered bins as follows, starting from 72,036 autosomal 40Kb bins: First, we removed 40Kb bins with zero effective fragment size, zero GC content, or zero mappability score. Next, we filtered out 40Kb bins within 200Kb of the bins removed in the previous step. We further filtered out 40Kb bins overlapping with the chr6 MHC region (chr6:28,477,797-33,448,354; hg19), which has extremely high SNP density that can make it difficult to correct for allelic mapping artifacts. This left 64,222 40Kb bins for

downstream analysis. Next, we applied HiCNormCis to remove systematic biases from local genomic features, including effective fragment size, GC content and mappability. The normalized total number of cis intra-chromosomal reads is defined as FIRE score. We further performed quantile normalization across multiple individuals using R package “preprocessCore”. The final FIRE score is log transformation $\log_2(\text{FIRE score} + 1)$ and converted into a Z-score to create a mean of 0 and standard deviation of one. To identify significant FIRE bins in each individual, we used one-sided P-value < 0.05 . Ultimately, merging across all individuals, we identified 6,980 40Kb bins which are FIRE bin in at least one of 12 YRI individuals. Consistent with our previous findings¹⁸, we observed significant enrichment of GM12878 typical enhancers and super enhancers among these 6,980 40Kb FIRE bins (**Figure S2.1d**). GREAT analysis⁶⁰ further showed immune-related biological pathways and disease ontologies are enriched in these 6,980 40Kb FIRE bins (**Figure S2.1e**).

Comparison of intra-individual vs inter-individual variation

To estimate variability between replicates, we computed Pearson correlation coefficient for all pairs of biological replicates for each score (DI, INS, FIRE and PC1). The pairs then can be divided into two groups based on whether they are from the same individuals as illustrated in **Figure S2.4c**. We then tested if the distribution of Pearson correlation coefficients were different comparing two groups. Similar analysis was performed for contact matrices. For contact matrices, we calculated Pearson correlation coefficient for each distance and each chromosome separately as shown in **Figure 2.1c**.

Variable regions

limma test for variable bin To test regions that are variable across genomes, we

applied limma⁴⁶ with default parameters. First, values for each 40Kb bin in hg19 reference genome were calculated for each metrics tested (DI, FIRE, INS, PC1) as described above. DI, PC1, and INS scores were calculated based on contact matrices quantile normalized across 40 replicates. FIRE scores were calculated based on raw counts using HiCNormCis and then quantile normalized across 40 replicates. Second, we filtered out bins that are not testable. Specifically, FIRE scores were only tested for bins that are FIRE regions (p-value < 0.05) in any of 40 replicates. DI scores were only tested for bins where strong biases are observed ($\text{abs}(\text{DI}) > 10.82757$, which correspond to Chi-squared test p-value 0.001) in any of 40 replicates. INS scores were only tested for bins where strong insulation is observed (z-score transformed INS score < -1) in any of 40 replicates. No filterers were performed for PC1 scores. Third, we filtered out any bins that overlapping large SVs (> 10,000 bp) to avoid effect caused by SVs. Specifically, for FIRE, INS, and DI scores, bins that are within 200Kb, 400Kb, and 2Mb respectively upstream or downstream of large SVs were removed. For PC1 scores, bins overlapping large SVs were removed. Lastly, we applied limma standard model with individual as a fixed factor and eBayes correction. To estimate empirical false positive rate (FDR), we bootstrapped replicates to calculate the number of false positives in random background. Briefly, we random selected 40 or 22 replicates with replacement for LCL20 and YRI11 respectively, and identified variable regions as mentioned above. We performed 1,000 permutations and calculated empirical FDR as the average positive hits in 1,000 permutations divided by number of hits in real data.

Normalizing Hi-C contact matrices using BNBC normalization To directly compare individual Hi-C contact matrix cells across samples, we sought to remove unwanted per-

cell variation owing to date of processing or other unknown ‘batch’ effects. To this end we developed Bandwise Normalization and Batch effect Correction (BNBC), described and evaluated in a separate manuscript (preprint on bioRxiv <https://www.biorxiv.org/content/10.1101/214361v1>). A brief description follows. For each chromosome and for each strata of distance between loci (a matrix “band”, hence the term “bandwise”), we correct for unwanted variation by taking the log counts-per-million-transformed values of all samples and generating a matrix whose entries are the observations for that chromosome’s matrix band across all samples (columns indexes samples and rows indexes contact matrix cells with anchor bins separated by a fixed distance). We then quantile normalize this matrix and regress out the impact of known batches (here, date of processing) using ComBat⁶¹ (specifically we correct both mean and variance). This procedure essentially conditions on genomic distance. We correct the majority of each contact matrix for each chromosome for each sample: we correct all but the 8 most distal matrix bands, for which we set all values to 0. The choice of the last 8 bands is empirical and reflects the small number of observations in each band matrix. The procedure is implemented in the `bnbc` package available through Bioconductor (<http://www.bioconductor.org/packages/bnbc>). Correction of contact matrices was performed on replicate-level data using the following LCLs: GM18486 (YRI-1), GM18505 (YRI-2), GM18507 (YRI-3), GM18508 (YRI-4), GM18516 (YRI-5), GM18522 (YRI-6), GM19099 (YRI-7), GM19141 (YRI-8), GM19204 (YRI-10), GM19238 (YRI-11), GM19239 (YRI-12), GM19240 (YRI-13), HG00731 (PUR-1), HG00732 (PUR-2), HG00512 (CHS-1), HG00513 (CHS-2). We note that NA19239 (YRI-12) replicate 1 and NA19240 (YRI-13) replicate 2 were excluded because the BNBC algorithm requires

multiple samples from a given experimental batch to estimate batch effect parameters.

Identifying biological variability in Hi-C contact matrices To identify contacts with significant levels of between-individual variability we employed the following procedure, which mimics the analysis for INS, DI, FIRE and PC1, on contact matrices normalized by BNBC. For each contact matrix cell (representing loci separated by less than 28 Mb, this is a subset of the matrix cells normalized by BNBC) we used a linear model with individual modeled as a fixed factor, note we have 2 growth replicates for almost every individual. We used a parametric likelihood ratio test (equivalent to an F-test) to test whether there was significant between-individual variation. We used the IHW framework with the distance between anchor bins as informative covariate, to increase power and estimate false discovery rate. We used a FDR of 10% as significance threshold, resulting in 115,817 contact matrix cells with significant biological variability across the autosomes. To estimate effect size (depicted in **Figures 2.2a, 2.3a** and **Figure S2.5**) we used a linear mixed effect model with individual as random effect, to decompose the variance into between-individual variability (biological) and within-individual variability (technical). As the measure of biological variability in these figures, we used the estimated biological variance. For this analysis, all 16 samples we normalized using BNBC were used.

Correlation with other datasets To examine correlation between 3D genome organization and other genome features, we reidentified variable regions with the same pipeline mentioned above using only individuals of which data is available for other genome features, and then computed Spearman correlation coefficient between 3D genome metrics (DI, INS, PC1, and FIRE) and other genome features (RNA-seq, ChIP-seq, and DNase-seq) for each 40Kb bin that is variable. Signals for each 40Kb bins were calculated

by averaging signals for the bin. Specifically, signals for ChIP-seq were the average signal of all peaks within the bin, signals for RNA-seq were the average FPKM of all genes in the bin, and DNase signals were simply average signal for each base pair in the bin. In some cases, several consecutive bins were identified as variable. We only kept the bin with strongest signal for other genome features among consecutive bins. To generate random backgrounds, we permuted individual labels for the same set of bins and recomputed Spearman correlation coefficient. 10,000 such permutations were used to calculate the statistical significance of departure from the null hypothesis in which the median value of true correlation values and permuted correlation values are equal. Similar analysis was performed for variable matrix cells with the following modifications. First, we used the variable matrix cells in the preceding section. Second, to correlate matrix-cell-level contacts with bin-level DNase and ChIP-seq signals, anchor bins of variable matrix cells were used. Since each anchor bin may belong to more than one matrix cells, we only used each bin once and selected the one with the highest Spearman correlation coefficient. Exactly same approach was performed during permutation to ensure a fair comparison.

Identification of QTLs

Testable bins To identify testable bins for FIRE-QTL, DI-QTL, INS-QTL and C-QTL searches, we began with 72,036 autosomal 40Kb bins based on reference genome hg19. We eliminated “unreliable” bins with effective length, GC content, or mappability equal to zero, resulting in 66,597 bins remaining. We further removed any 40Kb bins within 200Kb of an unreliable bin, resulting in 64,337 40Kb bins. We also removed bins covering the chr6 MHC locus (hg19: chr6:28,477,797-33,448,354, which is extremely polymorphic and

may lead to complex mapping artifacts that are difficult to correct. To eliminate false signals in Hi-C data that could arise from large structural variations (SVs), we obtained SVs from the 1000 Genomes consortium³⁴ (ftp://ftp-trace.ncbi.nih.gov/1000genomes/ftp/phase3/integrated_sv_map/ALL.wgs.integrated_sv_map_v2.20130502.svs.genotypes.vcf.gz) and removed bins which overlap one or more structural variants previously annotated in these individuals (N=123,015 SVs), or within 200Kb of large structural variations (>10Kb, N=1,253 SVs). These filtering steps yielded a set of 51,511 testable bins, which represent a common starting point for FIRE-QTL, DI-QTL, INS-QTL and C-QTL searches as described below.

Testable SNPs We began with a list of 15,765,667 variants among all 20 LCL individuals. We kept 14,177,284 variants among 11 unrelated YRI individuals, removed all indels, HindIII site polymorphisms, multi-allelic SNPs, and SNPs with minor allele frequency (MAF) < 5%. We also required that remaining SNPs were within the 51,511 testable bins described above, and that both alleles were present in at least 2 individuals in the discovery set individuals. (N=4,132,791 SNPs remaining). Finally, where multiple SNPs in the same bin were in perfect LD among 11 unrelated YRI individuals, we selected one with the smallest genomic position (to avoid the introduction of a random selection that would not be perfectly reproducible), ultimately yielding 1,304,404 potentially testable SNPs that served as a common input set to all QTL searches.

Power Calculations To explore the power of our approach and data, we performed a Monte Carlo-based power calculation. Specifically, we varied four variables: (1) the minor allele frequency of a variant; (2) the effect size of genotype (a fixed effect); (3) the variability between subjects (a random effect); (4) the variability of the residuals. For

contact QTLs, we also varied the mean of the Hi-C contact frequency in question. For analyses reported, we fixed the number of replicates-per-subject to be 2 (consistent with our study design). We explored a variety of settings for these parameters to assess power as each variable changes. Each setting tested was chosen to reflect the distribution of observed values in our real Hi-C data. For each configuration of parameters, we performed the following simulation: We simulated genotypes by randomly sampling a set of alleles (one allele per subject) from a binomial distribution parameterized by the number of subjects and the MAF; we repeated this process twice and create per-subject genotypes by adding the results of the sampling of alleles. We simulated per-subject random effects, and per-sample residuals. To obtain a given sample's simulated Hi-C contact matrix value, we added the mean Hi-C contact matrix value to that sample's simulated genotype (multiplied by the pre-specified effect size), the specific subject's random intercept and the sample's random residual. After performing this for all samples, we then fitted the same LMM model used in our QTL search. We repeated this simulation and model fitting process 1,000 times and computed power as the fraction of times the null hypothesis that the effect of genotype is equal to 0 is rejected at a nominal p-value of 0.05.

FIRE, DI, and INS QTL searches. We limited our FIRE QTL search to the subset of testable bins that were called as FIRE in at least one YRI LCL (N=5,822 FIRE test bins), and the subset of testable SNPs therein (N=128,137 FIRE test SNPs). For the INS-QTL search, we examined 328,530 test SNPs with 12,976 variable INS bins. For the DI-QTL search, we examined 181,950 test SNPs with 7,590 variable DI bins. For the DI-QTL search, we further classified each DI bin based on which whether it showed stronger

upstream or downstream bias, because we saw a Simpson's paradox when we considered them together. For each test SNP, we identified the 40Kb bin it belongs to, and fitted a linear mixed effect model, using FIRE, DI (200Kb window), or INS score (200Kb window) in each biological replicate as the response variable and genotype of that testable SNP as the explanatory variable. Since two biological replicates from the same individuals are correlated included an individual-specific random effect to account for within-individual correlation. We used the R package "nlme" and R function "gls" to fit the linear mixed effect model. The quantile-quantile plots (QQplot) showed only minor genomic inflation (median p-value = 0.4821, lambda = 1.0864 for FIRE-QTLs; median p-value = 0.4864, lambda = 1.0649 for upstream-biased DI-QTLs; median p-value = 0.4828, lambda = 1.0826 for downstream-biased DI-QTLs; median p-value = 0.4865, lambda = 1.0646 for INS-QTLs). The linear mixed effect model identified 476, 315, 315, and 1,092 SNPs with false discovery rate (FDR) less than 0.20 for FIRE, upstream-biased DI, downstream-biased DI and INS, respectively. When more than one SNP in the same bin was identified, we selected the SNP with lowest p-value among them to be included in the final QTL sets. After this filtering, we ended up with 387 candidate FIRE-QTLs, 268 candidate upstream-biased DI-QTLs, 277 downstream-biased DI-QTLs, and 911 candidate INS-QTLs. As a control for each of these QTL searches, we randomly shuffled the score in question (i.e. FIRE, DI, or INS) among all 11 YRI individuals and performed QTL searches on this permuted data. In each of these tests, we found no SNPs associated with the permuted scores at $FDR < 0.20$.

C-QTL search To find QTLs affecting Hi-C contact strength we first identified 115,187 Hi-C contact matrix cells exhibiting substantial biological variability as described

above, and constrained our QTL search to these cells. We then intersected these contact cells with 1,304,404 testable SNPs by requiring a SNP to sit in one anchor bin of one of these variable matrix cells. We also filtered out matrix cells to ensure both anchor bins of the matrix cell are among 51,511 testable bins. In total, we obtained 3,109,039 tests involving 687,655 SNPs and 54,880 matrix cells on all 22 autosomes. For each test, we used the BNBC normalized data described above, but used only the 11 unrelated YRI individuals with genotypes available and fit a linear mixed effect model in which genotype is a fixed effect and subject is a random intercept. We then used “lmerTest” package in R to estimate p-values for the fixed effect of genotype⁶². We used the IHW framework to estimate FDR, with the distance between anchor bins as an informative covariate, and call any matrix cell with $FDR < 0.2$ as significant. We further filtered significant tests by selecting the most significant SNPs per matrix cell and kept the leftmost SNPs among SNPs in perfect LD in two anchor bins of the matrix cell. After filtering, we ended up with 463 tests involving 345 SNPs and 463 matrix cells. To make the aggregate contact plots in **Figure 2.4g**, we recoded the genotypes based on the direction of effect such that 0, 1, 2 refer to the genotypes containing 0, 1 or 2 alleles associated with the increased phenotype, respectively. Next, to avoid aggregating the same submatrix multiple times, we filtered by 1) selecting only the most significant matrix cell associated with each QTL, 2) selecting only the most significant QTL associated with each anchor bin (in some cases the same bin anchors multiple matrix cells associated with different QTL SNPs). This filtering left 165 unique matrix cell QTL interactions for plotting. For each matrix cell, we then extracted a submatrix including 25 bins upstream and 25 bins downstream. Submatrices with missing values were discarded. For each QTL, we then calculated the

mean submatrix values for each genotype, and then subtracted submatrices to calculate the difference in interaction frequency between the 1 and 0 genotypes, and between the 1 and 2 genotypes. These differences were then averaged across QTLs and plotted in **Figure 2.4g**.

Validation of QTLs in additional individuals Our validation set included six unrelated individuals not included in the discovery set: NA12878, NA19240, HG00512, HG00513, HG00731 and HG00732. For each QTL, we collected the genotype among six additional individuals, and the corresponding FIRE, DI, or INS scores. Note that a small fraction of QTLs have missing genotypes in these six individuals (coded as “-1”), and these missing data points were eliminated from validation analysis. We examined the distributions of scores for each genotype. For each QTL type (i.e. FIRE, DI, or INS), we found that the same direction of effect observed in the discovery set is observed on average in the validation set. To assess the significance of this observation, we approximated the null expectation as follows. For FIRE-QTLs, for example, we started from all 128,137 FIRE test SNPs and 5,822 FIRE test bins. Note that in our discovery set, we identified 387 FIRE-QTLs, each in a different 40Kb bin. To create a random control SNP group, we first randomly selected 387 40Kb bins from all 5,822 FIRE test bins. Next, within each select bin, we randomly selected one SNP, and combined all these 387 selected SNPs into a control SNP group. We then tested their SNP effect on the six additional individuals. We repeated such sampling with replacement 1,000 times, to create a null distribution of positive and negative SNP effect, respectively. We performed the same type of permutations for DI, INS. Similar analysis was performed for C-QTLs with a few modifications. First, we only used replicates from NA19420, HG00512,

HG00513, HG00731 and HG00732 as explained above. Second, 1,000 random permutations were performed by sampling matrix cells instead of bins. Third, we used values of biological replicates separately instead of as merged data because the BNBC normalization is performed at the level of replicates.

Examining epigenetic variation at FIRE, DI, INS, and C-QTLs To examine epigenetic variation at 3D genome QTLs, we re-analyzed DNase-seq data from 59 LCLs, histone modification ChIP-Seq data (H3K27ac, H3K4me1 and H3K4me3) for 65 LCLs, and CTCF ChIP-seq data from 11 LCLs. These data were re-mapped using the WASP pipeline to control for allelic mapping artifacts and calculating the signal in 40kb bins as described. We examined the effect of genotype at FIRE, DI, INS or C-QTLs on DNase-seq and ChIP-seq signal by linear regression. As a control, we randomly selected the matched number of SNPs with the same approach described above and re-did such validation analysis. We repeated such random sample 1,000 times to create the empirical null distribution of no genetic effect. For C-QTLs, we used the sum of epigenetic features in two anchor bins to calculate correlation with contact frequency.

Nominal fraction analyses

Comparing between 3D chromatin QTL types To compare between different 3D chromatin QTLs, we took the raw test results for each QTL set and projected other 3D QTLs into the test results. For example, in **Figure 2.4j** we selected subset of SNPs that are DI-QTLs and plotted them (dark green dots) using p-values from FIRE-QTLs along with all tested in the FIRE-QTL search (black dots). We also used all tested SNPs in the DI-QTL search (light green dots) as a control set. To assign significance to the overlap, we compared the fraction of SNPs with nominal significance ($p\text{-value} < 0.05$) in each set:

1) DI-QTL tested SNPs that were not significant QTLs, and 2) DI-QTLs. We calculated p-values for this comparison by Chi-square test. To rule out the effect of sampling bias when selecting a small number of SNPs, we also performed permutation. In each permutation, we randomly selected the same number of SNPs as the real QTL set (from the full set of tested SNPs) and calculated the fraction with nominal significance. We then computed bootstrap p-values using 10,000 such permutations under the null hypothesis that the fraction of nominal significance is the same between QTLs and random selected SNPs. For C-QTLs, one SNP may be tested against multiple matrix cells, so we only keep the most significant p-value for each SNP to avoid biases towards SNPs with multiple tests.

Comparing 3D chromatin QTLs to other molQTLs Similar approaches were used to assess overlap between 3D chromatin QTLs and other molQTLs. We obtained full test results (all tested SNPs with the p-values) from previous molQTL studies and projected 3D chromatin QTLs into those test results. We then calculated fraction of nominal significance and used chi-square test to evaluate significance between 3D-QTLs and non-3D-QTLs. Similarly, we performed bootstrap to estimate significance empirically. One modification is that we extended our QTL sets by incorporating all SNPs in perfect LD with the same 40Kb bin because we may not use the same tagging SNP in our study as used in other studies. To ensure a fair comparison, we performed the same extension for the control sets of all tested SNPs.

Comparing 3D chromatin QTLs to GWAS Comparison with the GWAS results was performed in the same manner as described above for other molQTLs. Instead of test results for other molQTLs, we used summary statistics from previous GWAS.

FISH

Cell preparation for FISH Approximately 100,000 cells were adhered to center of PDL-coated coverslips (Neuvitro, GG-22-15-PDL) by placing 100 uL of cells at 1×10^6 cells/mL. Cells on coverslips were incubated for an hour at 37°C, carefully washed with PBS, and fixed with 4% paraformaldehyde in 1X PBS for 10 mins. PFA was quenched with 0.1 M Tris-Cl, pH 7.4 for 10 mins, washed with PBS, and stored in 1X PBS at 4°C for up to 1 month.

BAC probe labeling and preparation All BAC clones were ordered from the BACPAC Resource Center at the Children's Hospital Oakland Research Institute: "U" probe is RP11-74P5, "C" probe is RP11-337N12, and "D" probe is RP11-248M23. BAC DNAs were labeled with either Chromatide Alexa Fluor 488-5 dUTP (Invitrogen, C-11397) or Alexa Fluor 647-aha-dUTP (Invitrogen, A32764) using nick-translation kit (Roche, 10976776001), and incubated in 15°C for 4 hours. The nick-translation reaction was deactivated using 1 uL of 0.5 M EDTA, pH 8.0 and heated for 10 mins at 65°C. The probes were then purified using illustra ProbeQuant G-50 Micro Columns (GE Healthcare, 28903408) and eluted to a concentration of 20 ng/uL. Probes were mixed with Human Cot-1 DNA (Invitrogen, 15279011) and salmon sperm (Invitrogen, 15632011), and precipitated with 1/10th volume of 3M sodium acetate, pH 5.2 and 2.5 volume of absolute ethanol for at least 2 hours at -20°C. Probes were then spun down, washed with cold 70% ethanol, resuspended in formamide and 40% dextran sulfate in 8X SSC, and incubated at 55°C.

Hybridization Cells on coverslips were blocked with 5% BSA and 0.1% triton-X 100 in PBS for 30 mins at 37°C, and washed twice with 0.1% triton-X 100 in PBS for 10 mins each with gentle agitation at room temperature. Cells were permeabilized with 0.1%

saponin and 0.1% triton-X 100 in PBS for 10 mins at room temperature. Next, they were incubated in 20% glycerol in PBS for 20 mins, freeze-thawed three times with liquid nitrogen, and incubated in 0.1M hydrogen chloride at room temperature for 30 mins. Cells were further blocked for 1 hour at 37°C in 3% BSA and 100 ug/mL RNase A in PBS. Cells were permeabilized again with 0.5% saponin and 0.5% triton-X 100 in PBS for 30 mins at room temperature. Lastly, they were rinsed with 1X PBS and washed with 2X SSC for 5 mins. For hybridization of probes, the prepared probes were denatured at 73°C for 5 mins in water bath. Cells were denatured in a two-step process in a 73°C water bath: 2.5 mins in 70% formamide in 2X SSC and 1 min in 50% formamide in 2X SSC. Denatured probes were transferred onto microscope slides, and coverslips were placed on top with cell-side facing down. The coverslips were sealed with rubber cement and incubated overnight at 37°C in a dark, humid chamber. Next day, coverslips were carefully removed and transferred onto a 6-well plate. Cells were washed at 37°C with gentle agitation, twice with 50% formamide in 2X SSC for 15 mins and three times with 2X SSC for 5 mins. The cells were then stained with DAPI (Invitrogen, D1306), mounted on microscope slides with ProLong Gold Antifade Mountant (Invitrogen, P36930), sealed with nail polish, and imaged.

Microscope and analysis Images were acquired with DeltaVision RT Deconvolution Microscope at UC San Diego's department of neuroscience (acquired with award NS047101). Captured images were processed using the TANGO72 plugin in ImageJ for quantitative analysis. Each FISH experiment contained two probes labeled with different color dyes (either U-C or C-D). We limited our analysis to nuclei containing 2 labeled foci for each color (4 total foci), allowing us to more confidently distinguish foci

in cis from those in trans. Distances were measured from the center of one color focus to the center of the closest focus of the other color.

Re-analysis of public datasets

Analysis of ChIP-seq data from Kasowski et al and McVicker et al Raw fastq files were downloaded from SRA database for each experiment (SRP030041 and SRP026077, respectively). Reads were aligned to hg19 reference genome using BWA MEM (Kasowski) or BWA ALN v0.7.8 (McVicker) with WASP pipeline⁴⁴ to eliminate allelic mapping bias. Only reads with high mapping quality (>10) were kept. PCR duplicates were removed using Picard tools v1.131 (<http://broadinstitute.github.io/picard>). MACS2⁶³ v2.2.1 was then used to call peaks using corresponding input files. For CTCF and SA1, default parameters were used for MACS2. For H3K27ac, H3K4me1, and H3K4me3, peak calling was done using “--nomodel” parameter because we do not expect sharp peaks for histone modifications. For H3K27me3 and H3K36me3, peak calling was done using “--nomodel --broad” parameter. Bigwig files were generated by MACS2 using fold enrichment for viewing in genome browser. All Kasowski data were processed in pair-end mode and both replicates were merged for analysis. All McVicker data were processed in single-end mode, and the pooled input data were used for all samples because there are no individual input files. To compute signals in peaks, we used a set of merged peaks across all individuals for each mark.

Analysis of RNA-seq data from Kasowski et al Raw fastq files were downloaded from SRA database (SRP030041). Reads were aligned to hg19 reference genome using STAR⁶⁴ v2.4.2a with the WASP pipeline in pair-end mode to eliminate allelic mapping bias. Gencode v24 annotation was used to construct STAR index and computing FPKM.

Only uniquely mapped reads were kept. Cufflinks⁶⁵ v2.2.1 was applied to compute FPKM values. Both replicates were merged for analysis.

Analysis of DNase-seq data from Degner et al Raw fastq files were downloaded from SRA database for each experiment (SRP007821). Reads were aligned to hg19 reference genome using BWA ALN with the WASP pipeline in single-end mode to eliminate allelic mapping bias. Only reads with high mapping quality (>10) were kept. PCR duplicates were removed using Picard tools. Bigwig files were generated using makeUCSCfile commands in homer tools⁶⁶ v4.9.1.

Analysis of ChIP-seq data from Ding et al Raw fastq files were downloaded from SRA database for each experiment (SRP004714). Reads were aligned to hg19 reference genome using BWA MEM v0.7.8 with the WASP pipeline to eliminate allelic mapping bias. Only reads with high mapping quality (>10) were kept. PCR duplicates were removed using Picard tools. We performed quality control for CTCF ChIP-seq data by FRIP (Fraction of Reads In Peaks) and used datasets with FRIP > 10. Bigwig files were generated using bamCoverage commands in deepTools⁶⁷ v2.3.3. To compute signals in peaks, we used the merged CTCF peaks from Kasowski data.

Analysis of ChIP-seq data from Grubert et al Bigwig files and peaks for H3K27ac, H3K4me1 and H3K4me3 were downloaded from GEO database (GSE62742). Peaks for each mark were merged and then used to compute the averaged signal.

2.6 Figures

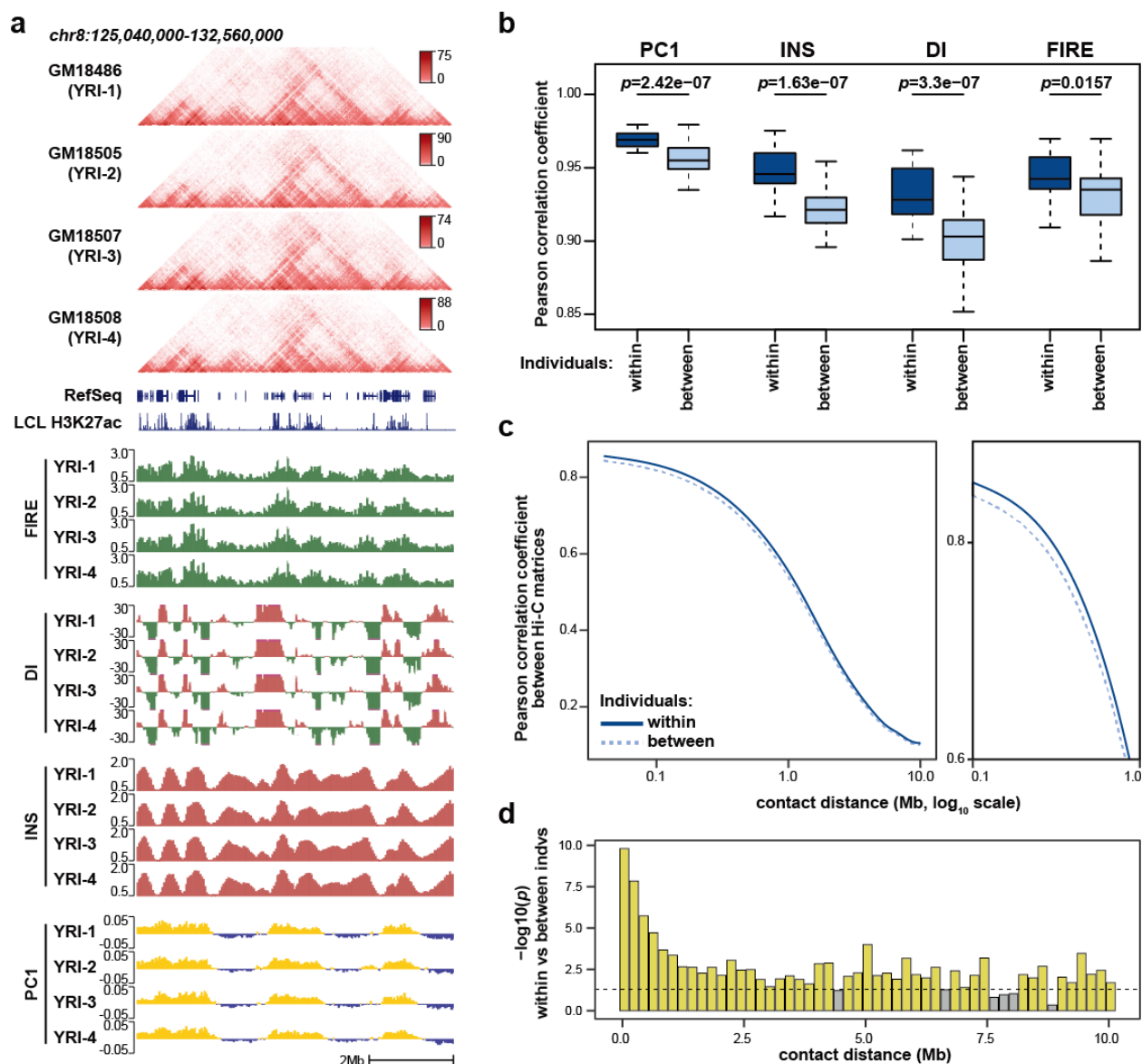


Figure 2.1. Biological variability in multiple aspects of 3D chromatin. (a) Browser view to illustrate the Hi-C-derived molecular phenotypes examined here: contact matrices, FIRE, DI, INS, and PC1 (*chr8:125,040,000-132,560,000*; hg19). (b) Boxplots show correlation between biological replicates from the same cell line (Individuals = “within”, N = 20), and between replicates from different cell lines (Individuals = “between”, N = 760). (c) The Pearson correlation coefficient between quantile normalized Hi-C matrix replicates from the same cell line or different cell lines is plotted as a function of genomic distance between anchor bins. (d) Significance of the difference between the “within” and “between” values in (c) was calculated at multiple points along the distance-correlation curve by two-sided Wilcoxon rank sum test.

Figure 2.2. Variable regions of 3D chromatin conformation. (a) Example of a variable region (chr15:93,040,000-100,560,000; hg19). (b) The number of testable bins and significantly variable regions for each 3D chromatin phenotype examined here. (c) Significance of pairwise overlap between different sets of variable regions. (d) Boxplots showing the distance between indicated probe sets in four different LCLs. (e) Representative images of nuclei corresponding to panel (d). (f) Blue line shows the fraction of variable matrix cells distributed across a range of interaction distances. Black shows the fraction of all matrix cells distributed across the same range of interaction distances. (g) Top panel shows the percentage of variable matrix cell anchor bins that overlap variable DI, FIRE, INS, or PC1 regions, respectively.

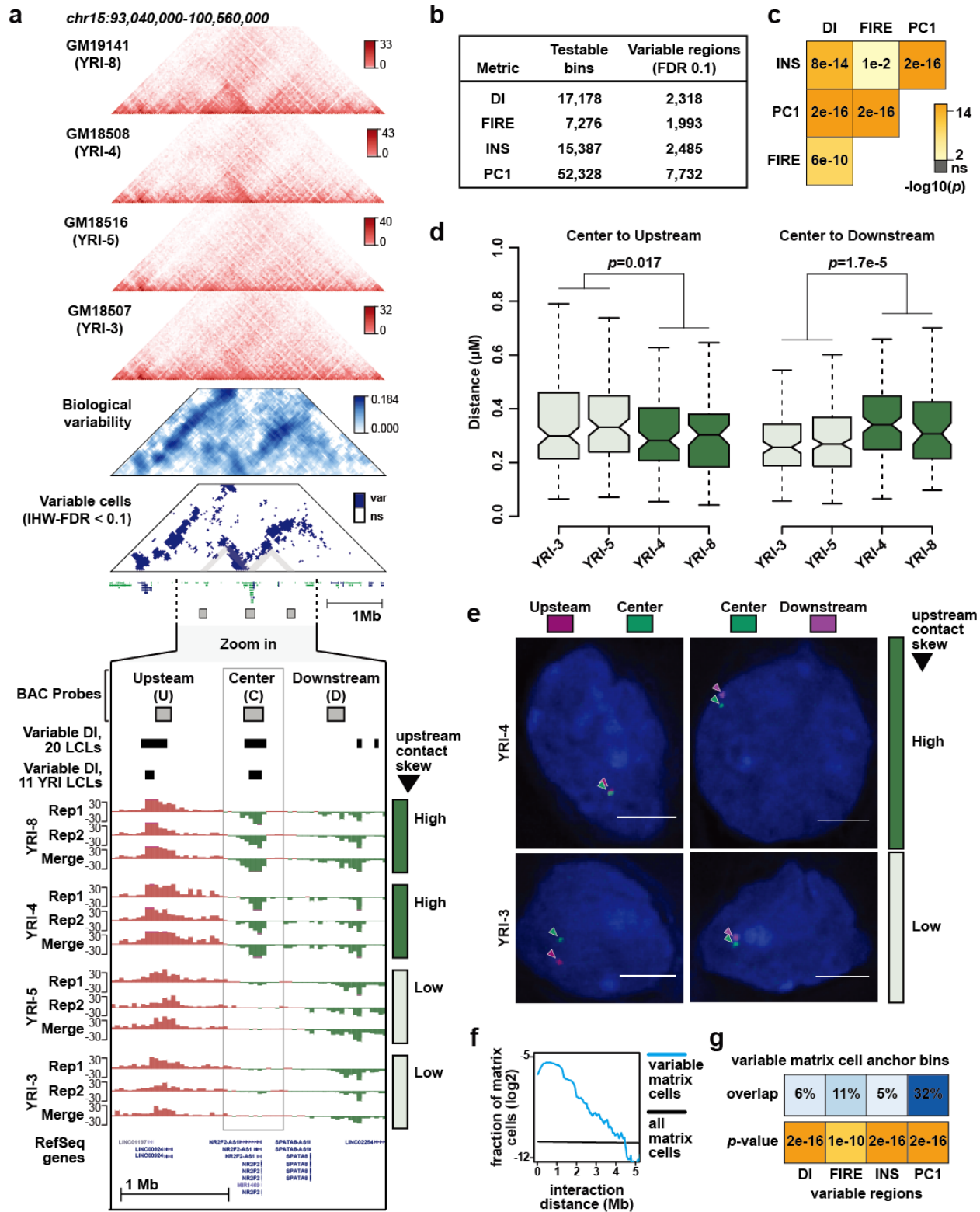


Figure 2.3. Coordinated variation of the 3D genome, epigenome, and transcriptome.

(a) Example of a variable region where 3D chromatin phenotypes are correlated with epigenomic and transcriptomic phenotypes (chr6:126,280,000-131,280,000; hg19). (b) Density plots show the distribution of Spearman correlation coefficients at variable regions between the epigenomic or transcriptomic phenotype indicated in the top margin of panel and the 3D chromatin phenotype indicated in the right margin of panel. (c) Heatmap showing Spearman correlation coefficients between PC1 and multiple epigenomic/transcriptomic phenotypes, arranged by k-means clustering (k=4). (d) Similar to (c), showing correlations with FIRE at N=132 variable FIRE regions. (e) Similar to (c), showing correlations with DI N=265 variable DI regions. (f) Similar to (c), showing correlations with INS at N=154 variable INS regions.

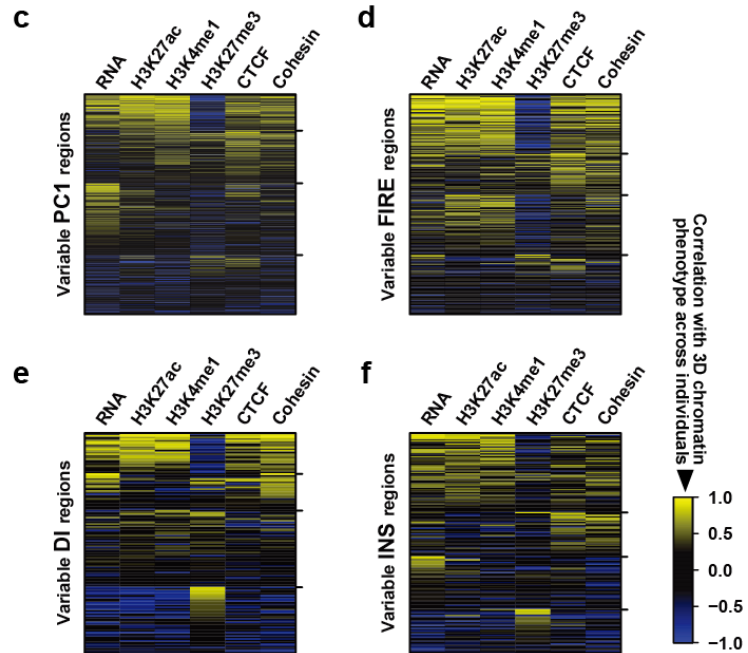
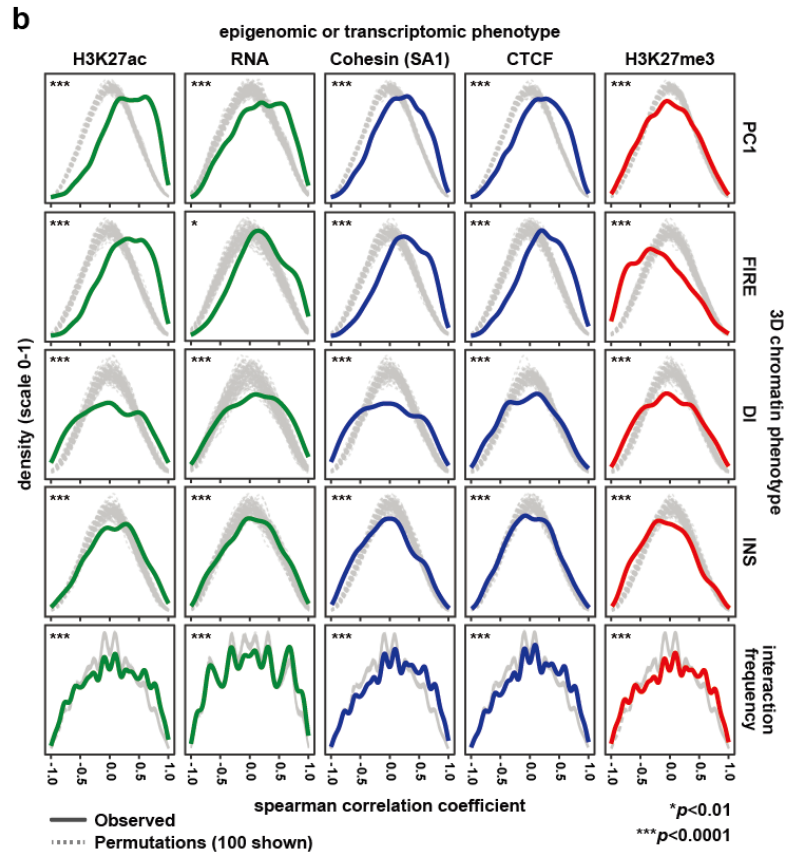
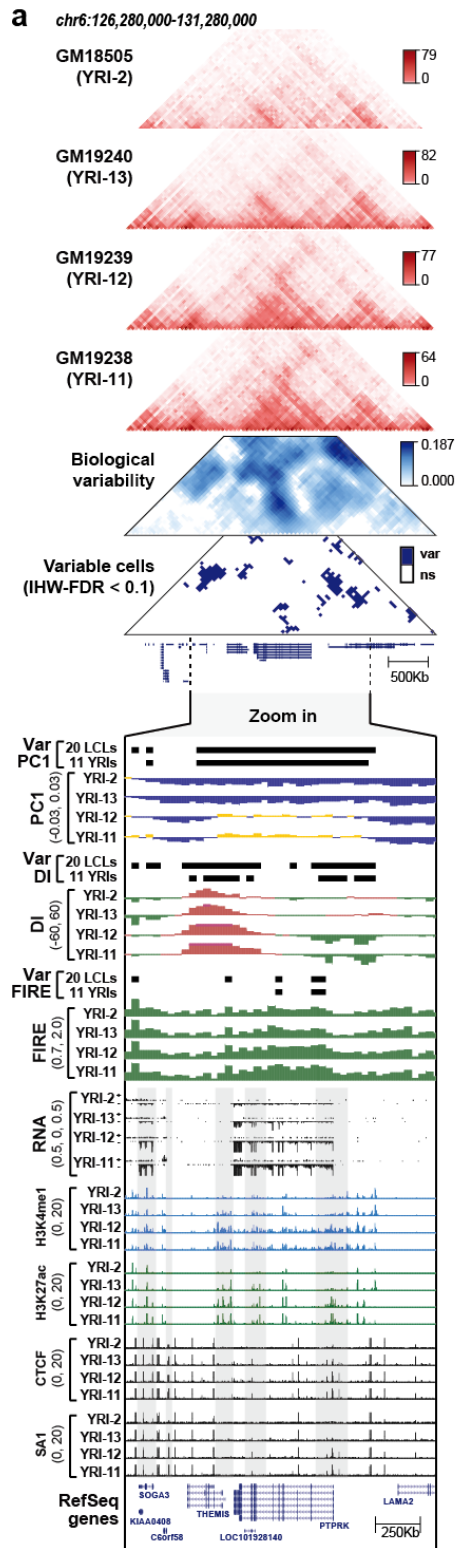


Figure 2.4. A genetic contribution to variations in 3D chromatin conformation. (a) A graphic representation of the CTCF Position Weight Matrix (PWM) is shown. (b) Boxplot shows the distribution of interaction frequencies at loops with exactly one anchor containing a CTCF motif disrupting SNP (N=138), separated according to genotype (c) Aggregate contact map shows the average difference in interaction frequency per loop between SS and SW genotypes (top; N=117 SNPs), and between SW and WW genotypes (bottom; N=31 SNPs). (d) Histogram shows the allelic imbalance in reads connecting loop anchors on the S vs W haplotypes in WS heterozygotes (N=135 loops). (e) Line plots show the genotype-dependent signal of FIRE-QTL, INS-QTL and DI-QTL using 11 independent YRI individuals. (f) For C-QTLs, an aggregate contact plot analogous to panel c is used to show the average difference in BNBC corrected interaction frequency (“ $\Delta \log(\text{norm contacts})$ ”) between the high and medium contact genotypes (top; N=138 interactions), and between the genotypes medium and low genotypes (bottom; N=94 interactions). (g) Boxplots show the genotype-dependent signal at QTLs using additional 6 individuals as a validation set. (h) Results of permutation test to evaluate the statistical significance of results in (g). (i) Line plot shows the fraction of foreground SNPs with nominal significance in the background association study (“nominal fraction”). (j) QQ Plot shows FIRE-QTL search results, including all SNPs tested for FIRE association (black points, N=128,137), and several subsets as follows: DI-QTL tested (light green, N=46,784), INS-QTLs tested (light red; N=6,238), C-QTL tested (light blue; N=69,847), DI-QTLs (dark green, N=152), INS-QTLs (dark red, N=60), C-QTLs (dark blue, N=53).

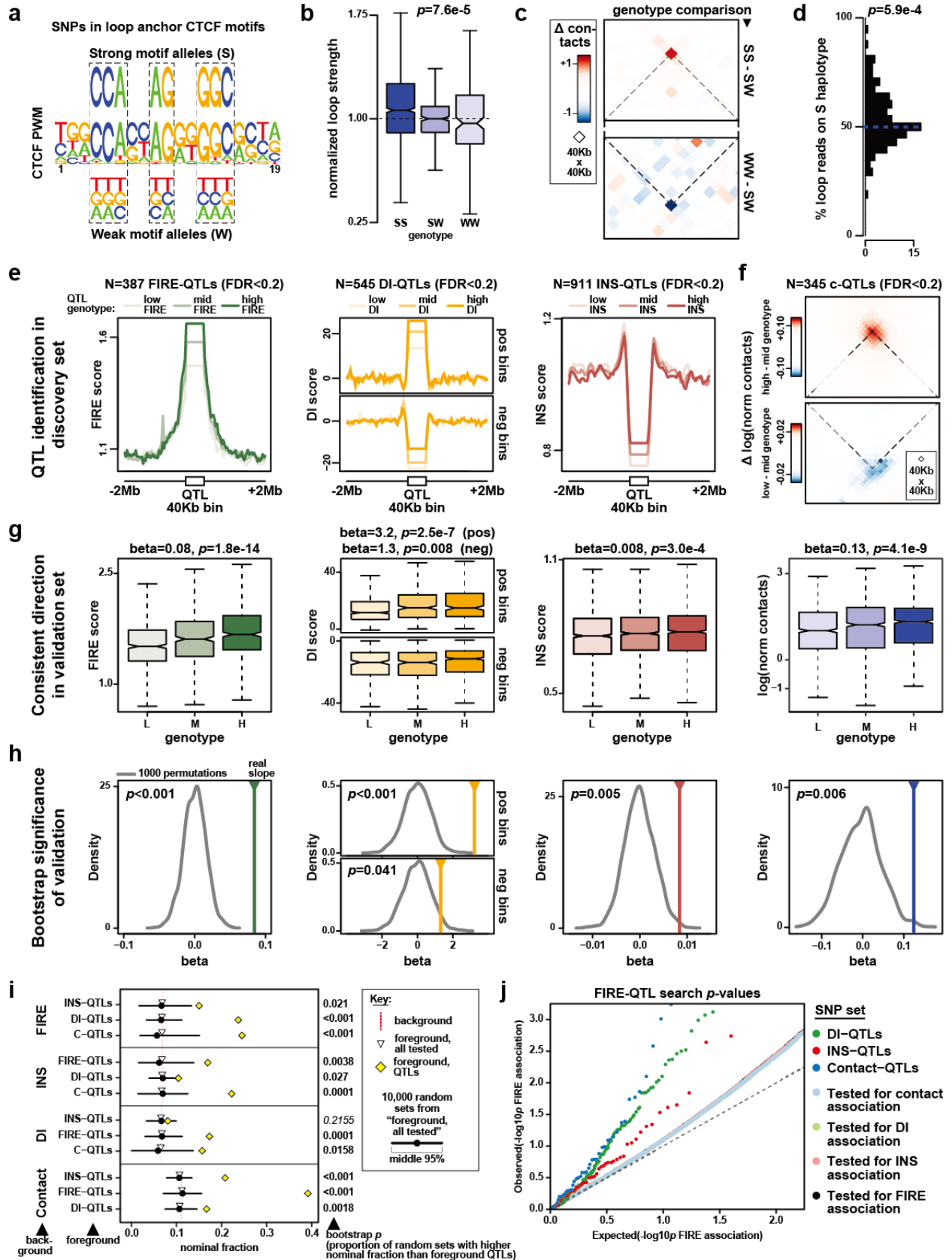
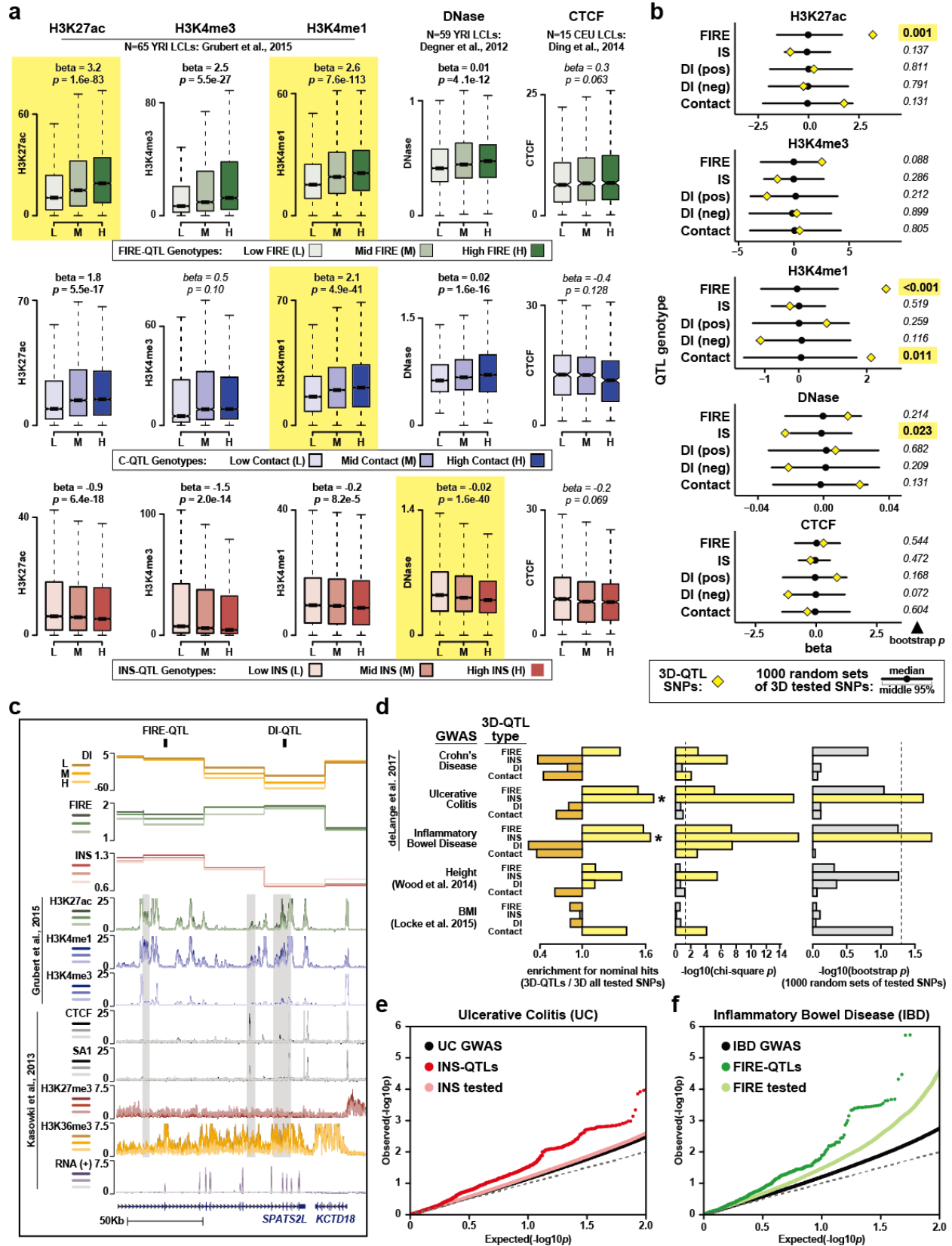


Figure 2.5. Contribution of 3D chromatin QTLs to other molecular and organismal phenotypes. (a) Boxplots show signal for epigenetic phenotypes separated by genotype at FIRE-QTLs (top row), C-QTLs (middle row), and INS-QTLs (bottom row). (b) Line plots shows beta values of linear relationships between QTL genotypes as indicated to the left and epigenetic phenotype indicated above each subpanel. (c) Genome browser view (chr2:201,222,342-201,386,844; hg19) showing examples of a DI-QTL (chr2:201333312) and FIRE-QTL (chr2:201254049). (d) Left subpanel shows the enrichment values for 3D QTL SNPs with nominal significance in the indicated GWAS study calculated as follows: (fraction of indicated 3D QTL SNPs with nominal significance in the indicated GWAS) / (fraction of SNPs tested in the indicated 3D QTL search with nominal significance in the indicated GWAS). (e) QQ plot shows the results of UC GWAS with all tested SNPs shows as black points, and two subsets as follows: SNPs also tested in our INS-QTL search (light red), and SNP called as INS-QTLs or in perfect LD with INS-QTLs in the same 40Kb bin (dark red). (f) QQ plot shows the results of IBD GWAS with all tested SNPs shows as black points, and two subsets as follows: SNPs also tested in our FIRE-QTL search (light green), and SNP called as FIRE-QTLs or in perfect LD with FIRE-QTLs in the same 40Kb bin (dark green).



2.7 Supplemental Figures

Figure S2.1. Hi-C derived molecular phenotypes measured across 20 LCLs. (a) Hi-C contact matrices show for all 20 LCLs. (b) Same region as above, but showing PC1 and FIRE values. ChIP-seq data for several histone modifications, CTCF, and Cohesin subunit SA1 are shown for one LCL (YRI-13, GM19240) as a reference for the epigenomic landscape. (c) Same region as above, but showing DI and INS values. (d) Bar plots show the percentage of super-enhancers (left) or typical enhancers (right) in GM12878⁵⁹ that overlap with 6,980 LCL FIRE bins (called as FIRE in at least one individual in our dataset) and 6,980 random 40kb bins. (e) Biological Process Gene Ontology terms associated with genes proximate to FIRE regions as defined by GREAT.

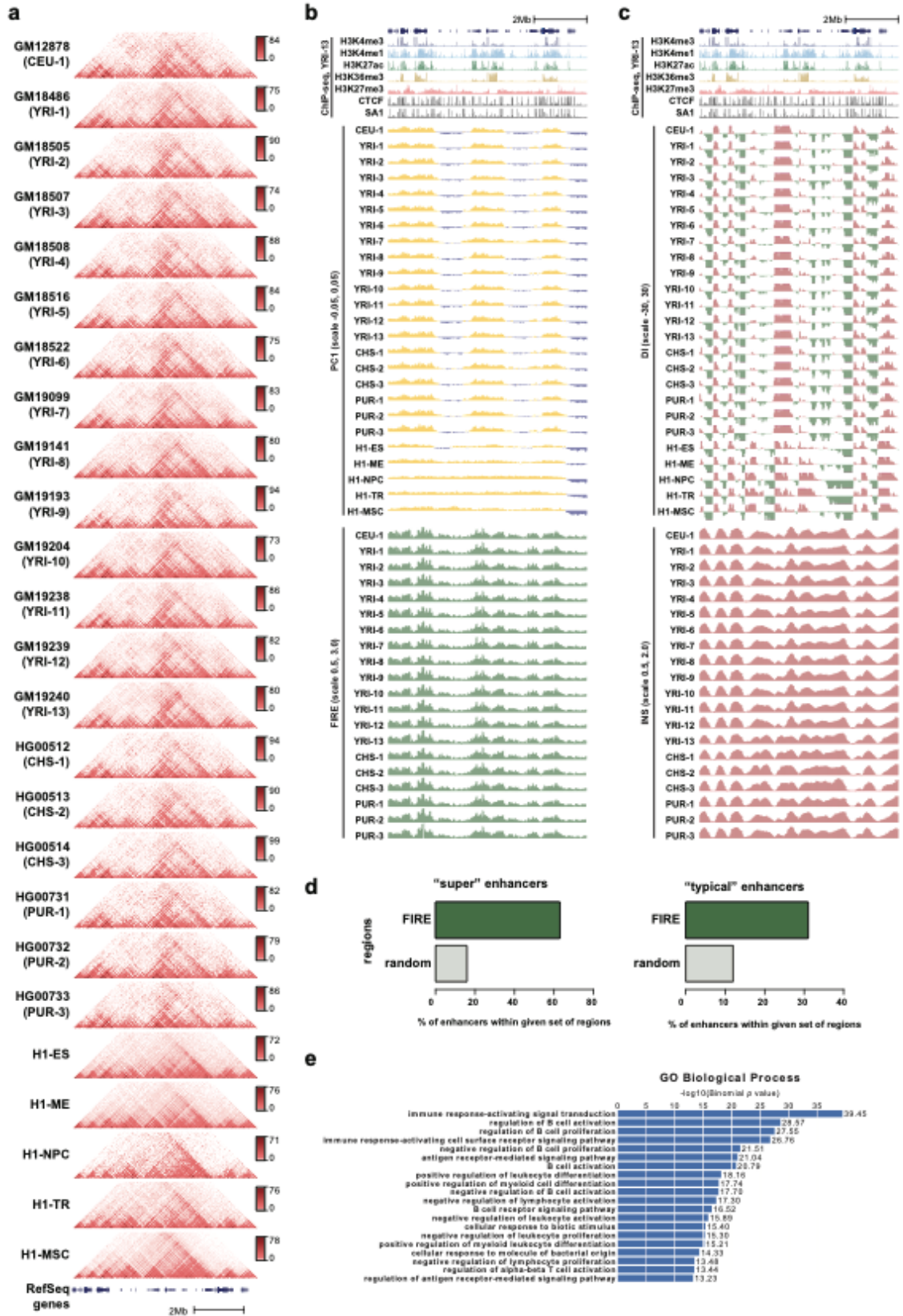


Figure S2.2. FIRE measures density of local interactions. Illustrative example showing that overall density of Hi-C reads (all reads irrespective of location of interacting partner, all *cis* interactions, or all *trans* interactions) is highly consistent across the genome. Top panel (a) shows the long arm of chr14 (chr14:24,406,737-104,693,368; hg19). Bottom panel (b) is a zoomed-in view of region boxed by dotted lines above (chr14:58,000,000-63,500,000; hg19).

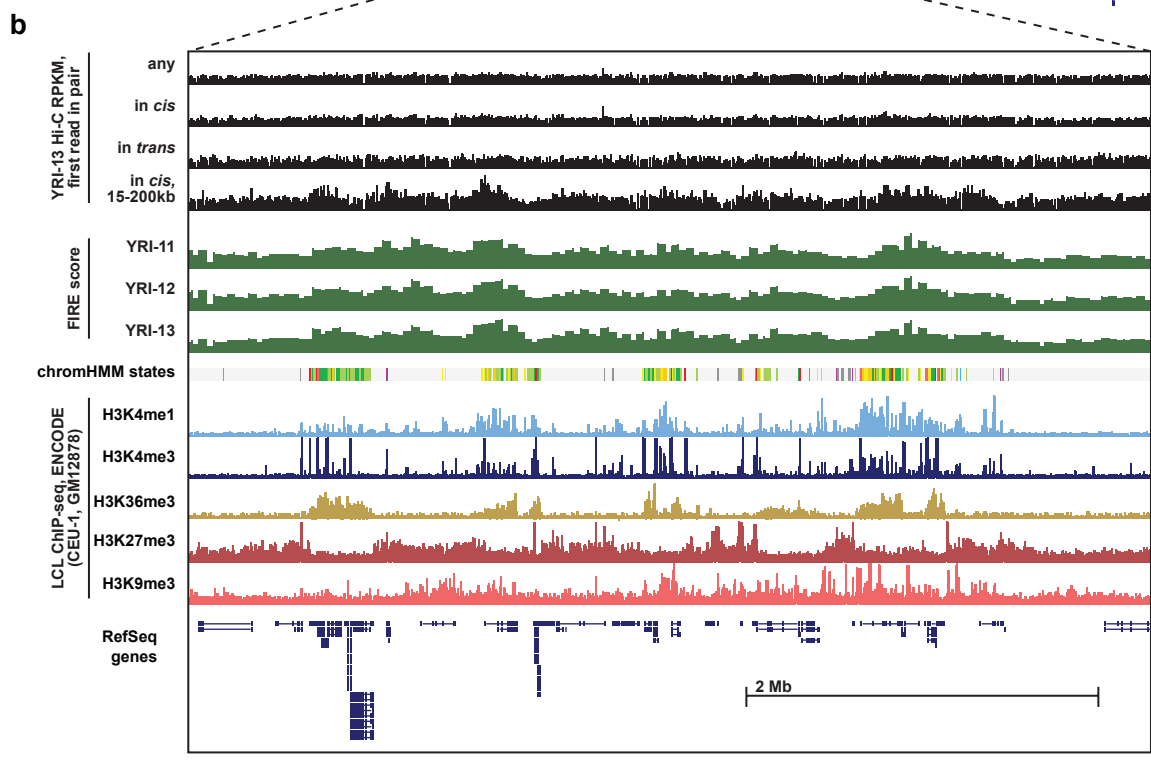
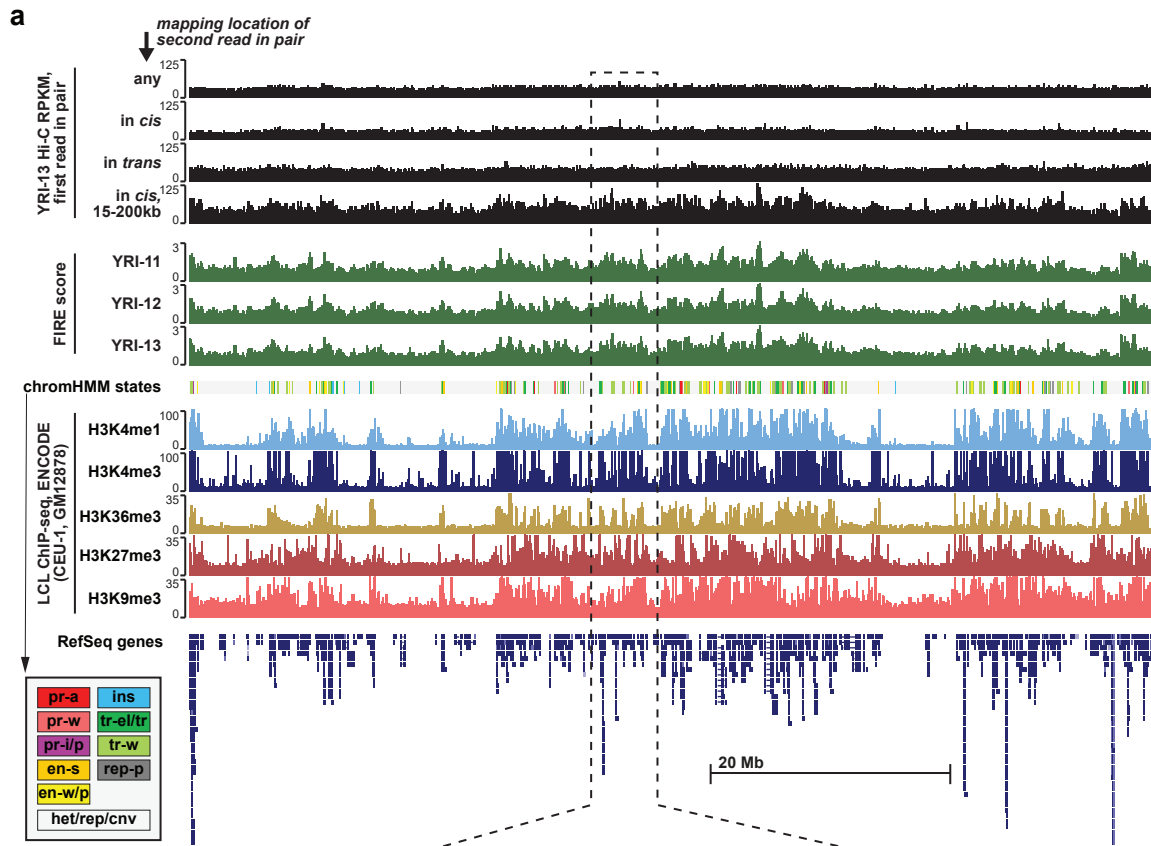


Figure S2.3. Aggregate looping interactions in each sample. Aggregate plots show the interaction frequencies at GM12878 HiCCUPS loops from Rao et al 2014 in each sample examined here.

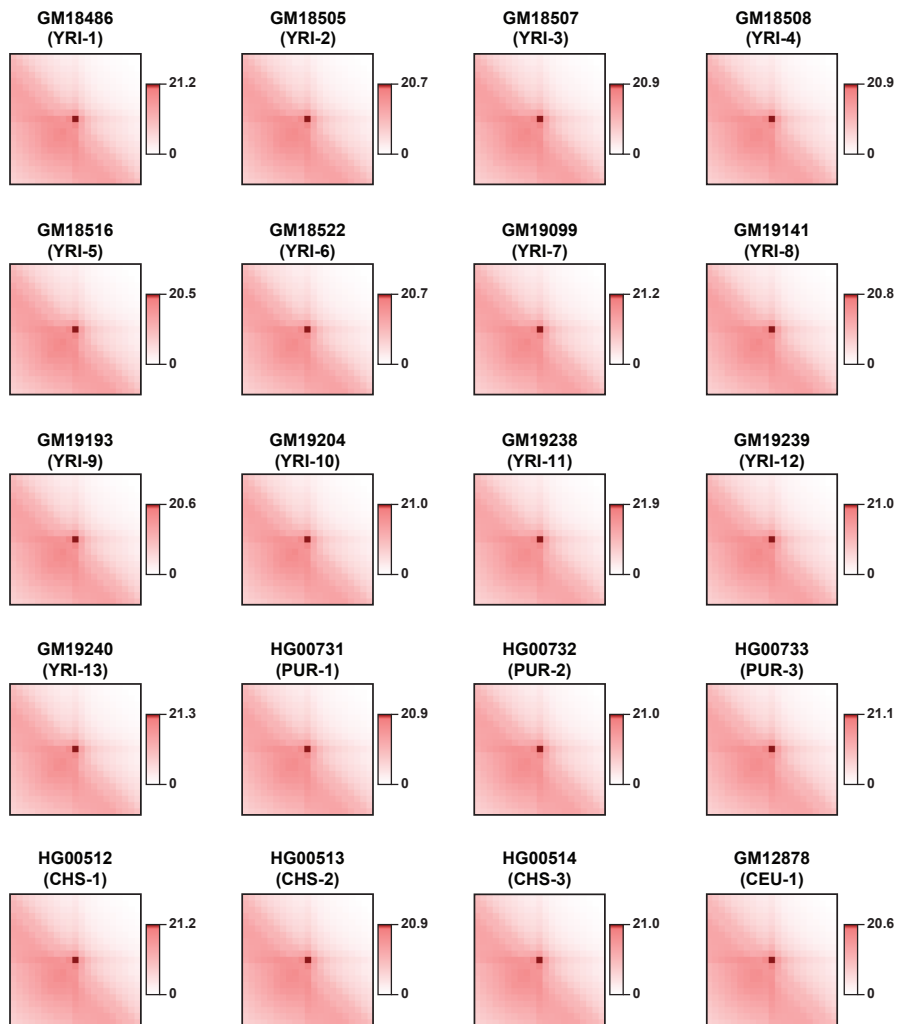


Figure S2.4. 3D chromatin variation among 20 LCLs and H1-derived lineages. (a) Graphical representation of the shuffling scheme used to assess biological variability in **Figure 2.1b-d**, and here in panels b-e. (b)-(e) Boxplots show Pearson correlation coefficient between biological replicates from the same cell line (Replicates = “True”), and between replicates from difference cell lines (Replicates = “Shuff”; short for “shuffled”). (f) Dendrograms from hierarchical clustering of 40 Hi-C replicates based on one of four Hi-C-derived phenotypes, as indicated above each dendrogram (DI, PC1, INS, or FIRE). (g) Principal Component Analysis of 20 LCLs using one of four Hi-C-derived phenotypes, as indicated above each plot.

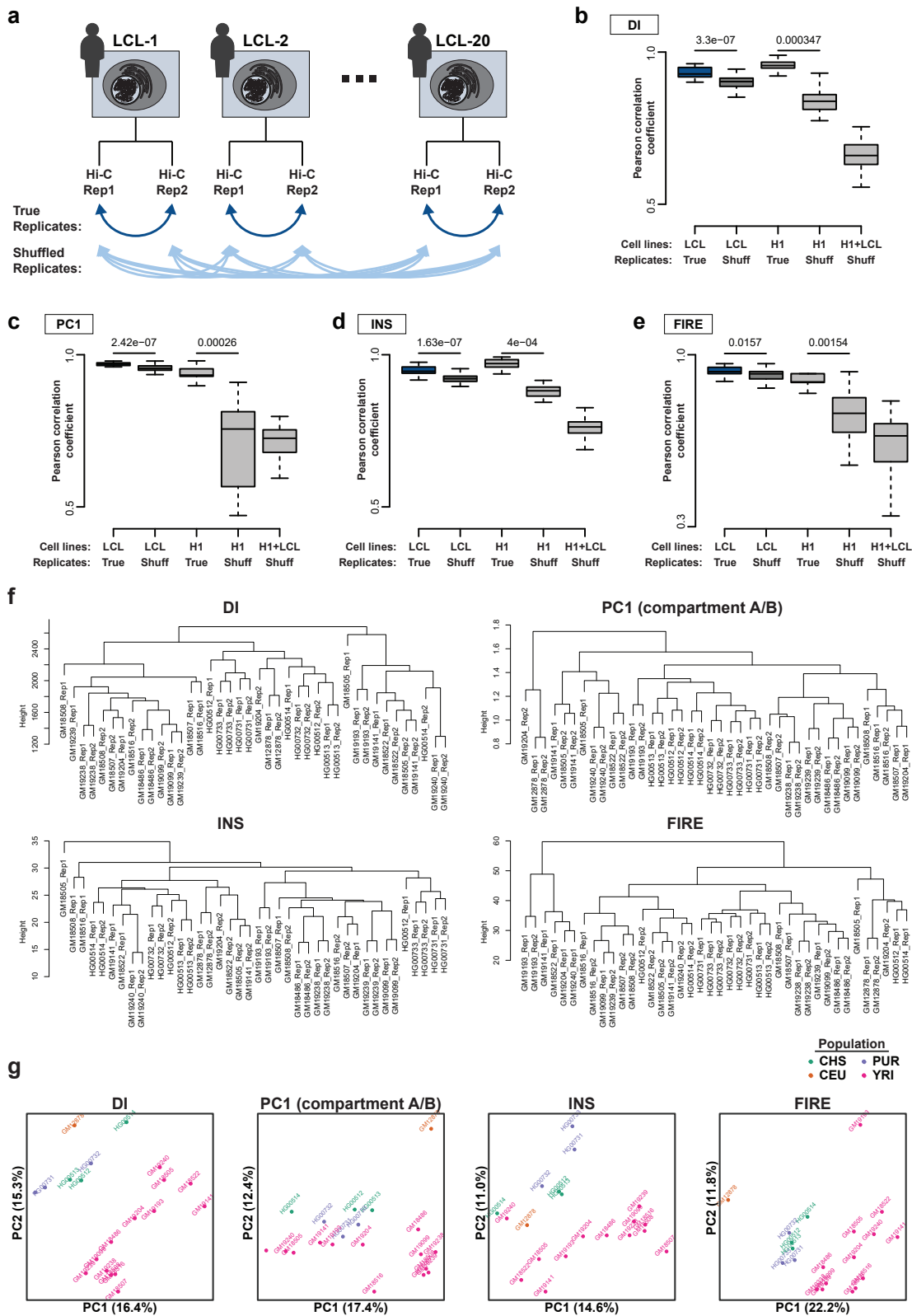
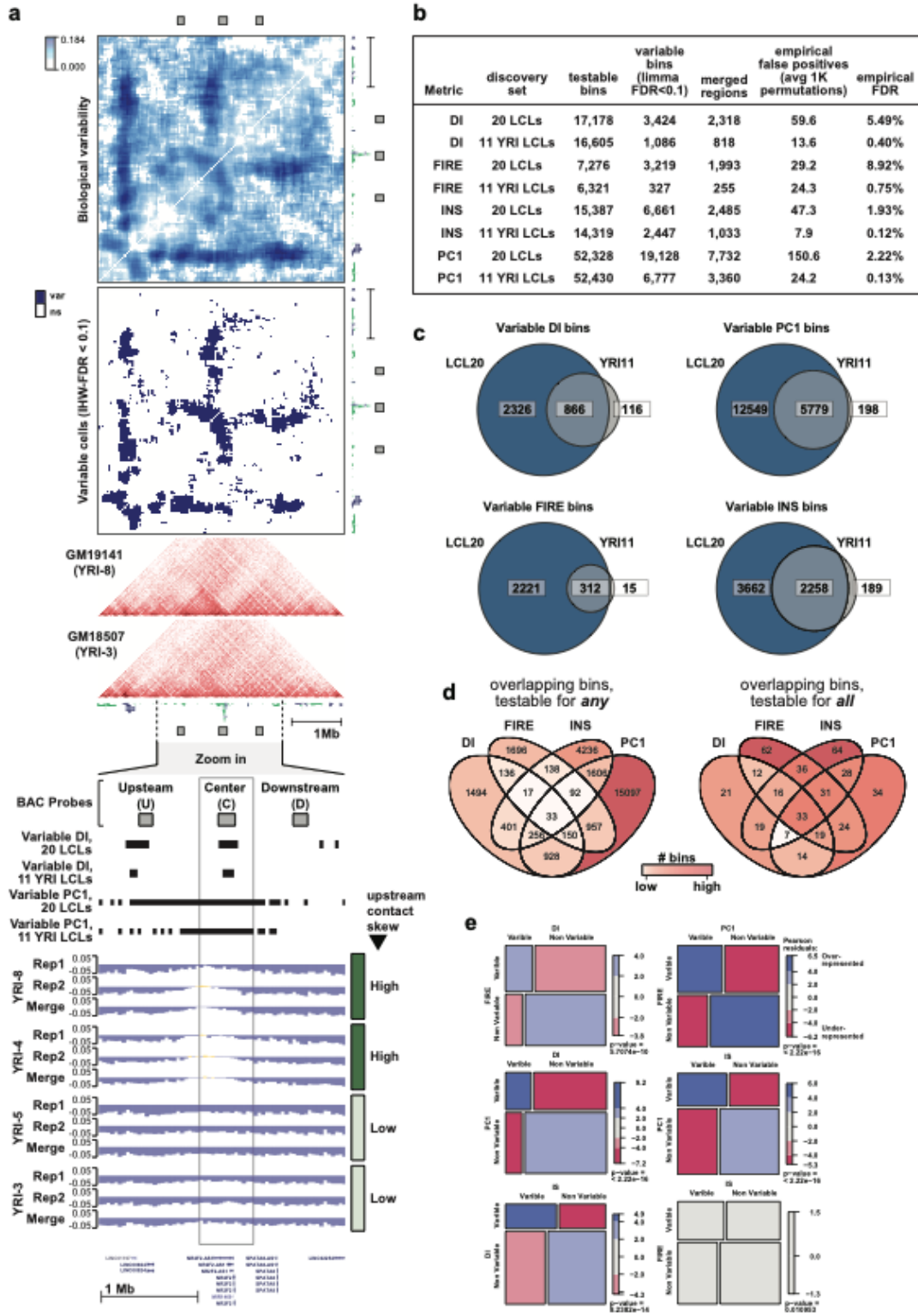


Figure S2.5. Characterization of variable regions of 3D chromatin conformation. (a) Same region as in **Figure 2.2a** (chr15:94,280,000-99,280,000), but showing reproducible variation in PC1, and full square matrices for contact matrix variability as opposed to the half-matrices shown in 2a. (b) Similar to **Figure 2.2b**, but with additional data columns. (c) Venn diagrams showing the overlap of variable regions identified using either all 20 LCLs (“LCL20”) or only the 11 unrelated YRI LCLs (“YRI11”). (d) Venn diagrams showing the number of variable bins for each phenotype or combination of phenotypes. (e) Mosaic plots show the significance of overlaps between variable regions in a pairwise fashion.



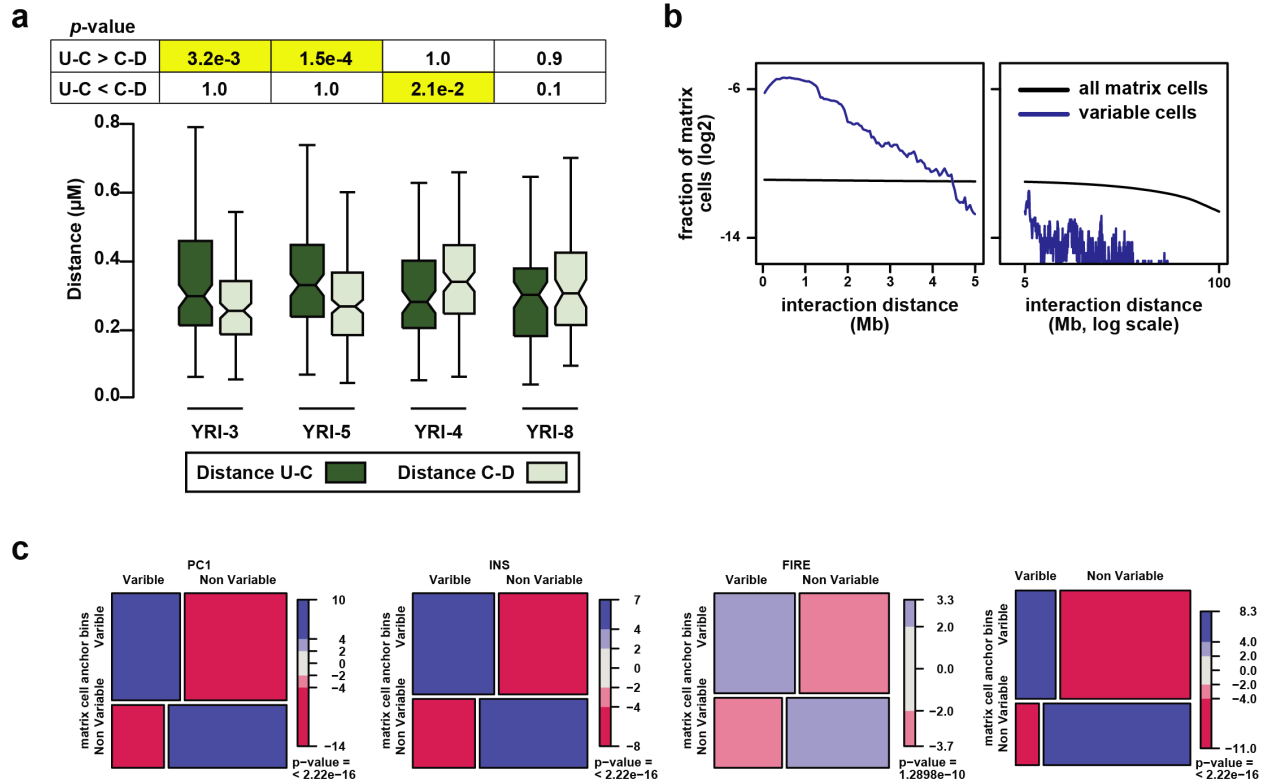
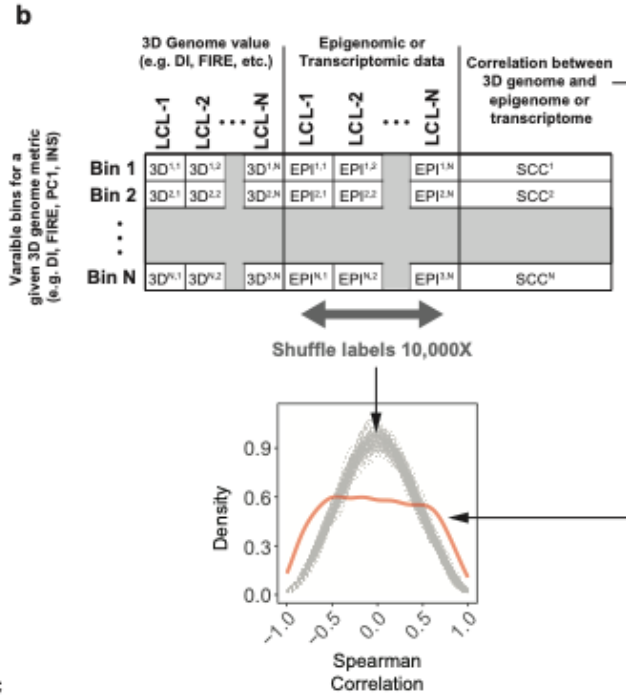
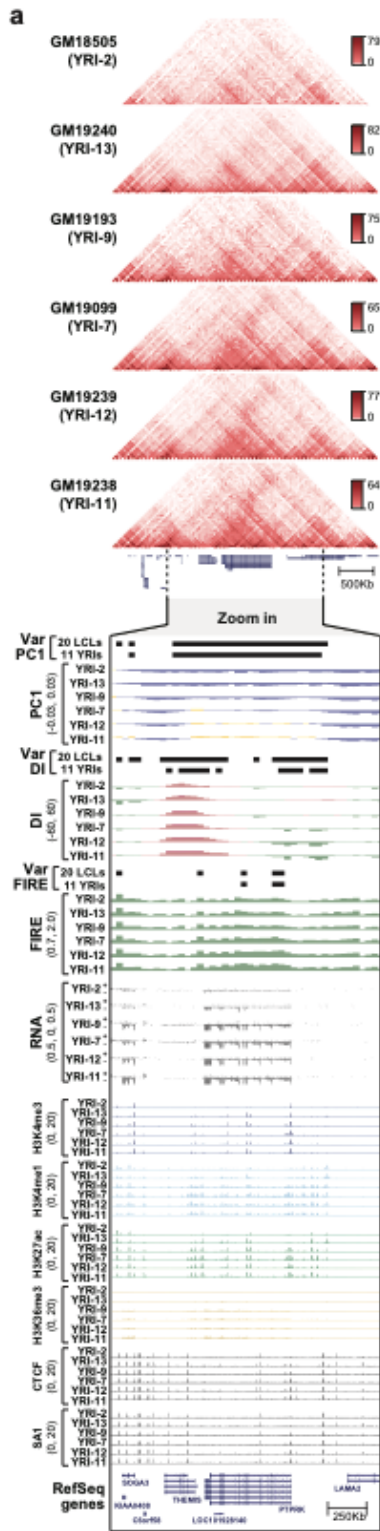


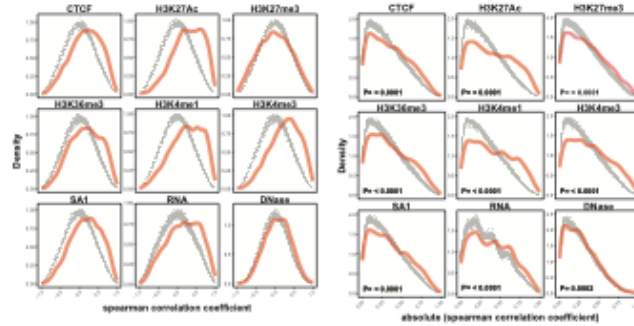
Figure S2.6. Additional characterization of variable regions of 3D chromatin conformation. (a) Same underlying FISH data as in **Figure 2.2e**, but here comparing the distance between U and C probes to the distance between C and D probes within the same LCL. (b) As in **Figure 2.2d**, blue line shows the fraction of variable matrix cells distributed across a range of interaction distances. (c) Mosaic plots show the significance of overlap between variable regions and anchor bins of variable matrix cells.

Figure S2.7. Coordinated variation between 3D chromatin conformation and multiple molecular phenotypes. (a) Same region as in **Figure 2.3a** (chr6:126,280,000-131,280,000; hg19), but showing additional individuals and additional data types as indicated. (b) Representation of permutation scheme used to calculate P-values in panels c as well as in **Figure 2.3b** and **Figures S2.8 and S2.9**. (c) Density plots in the top left quadrant show Spearman correlation coefficients (SCC) between PC1 and molecular phenotypes as indicated in the top margin of panel. Bottom left quadrant shows SCC like above, but using variable regions called in only 11 individuals. Bottom right quadrant shows SCC using ChIP-seq data from McVicker *et al.*

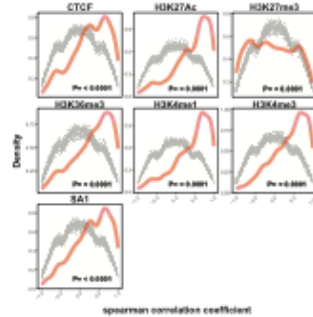


c
Type of variable regions: PC1

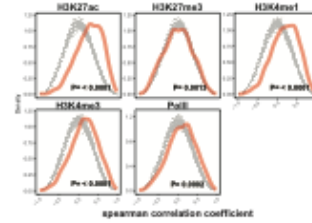
Individuals in discovery set: all 20
Public data set for comparison: CHIP-seq & RNA-seq from Kasowski et al; DNase-seq from Degner et al.



YRI 11
CHIP-seq from Kasowski et al

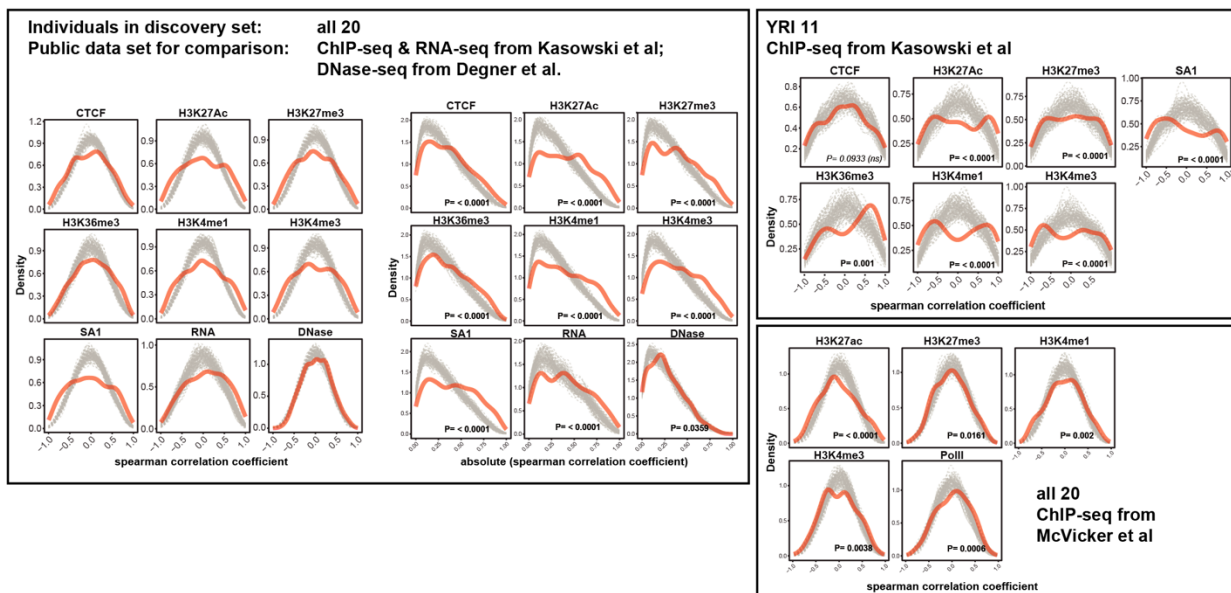


all 20
CHIP-seq from McVicker et al



a

Type of variable regions: DI



b

Type of variable regions: INS

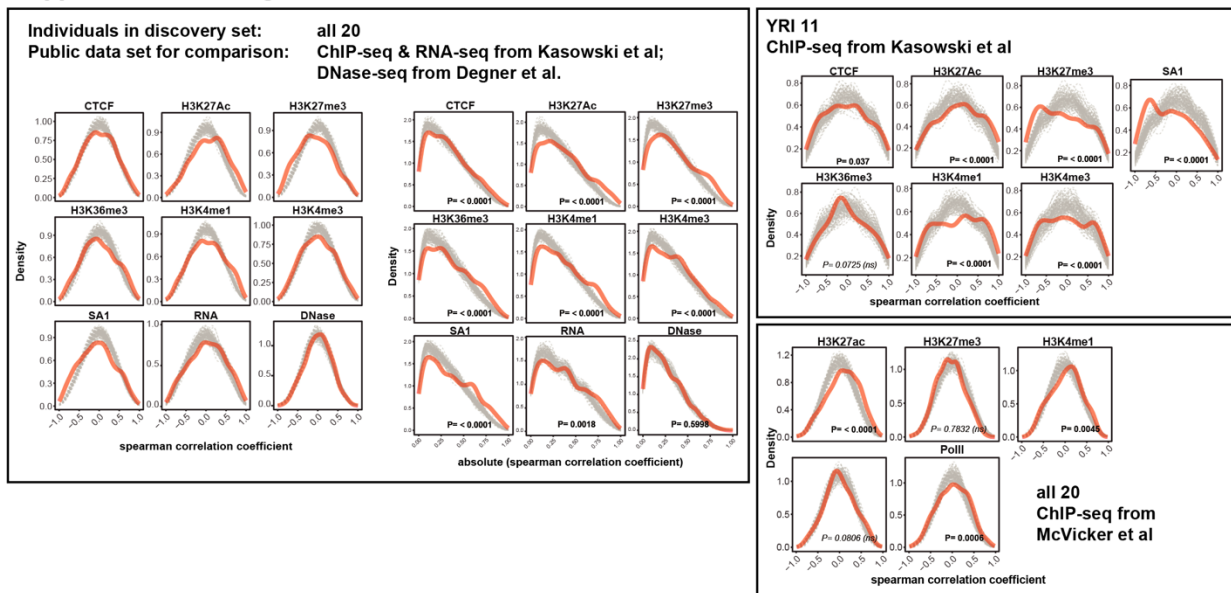
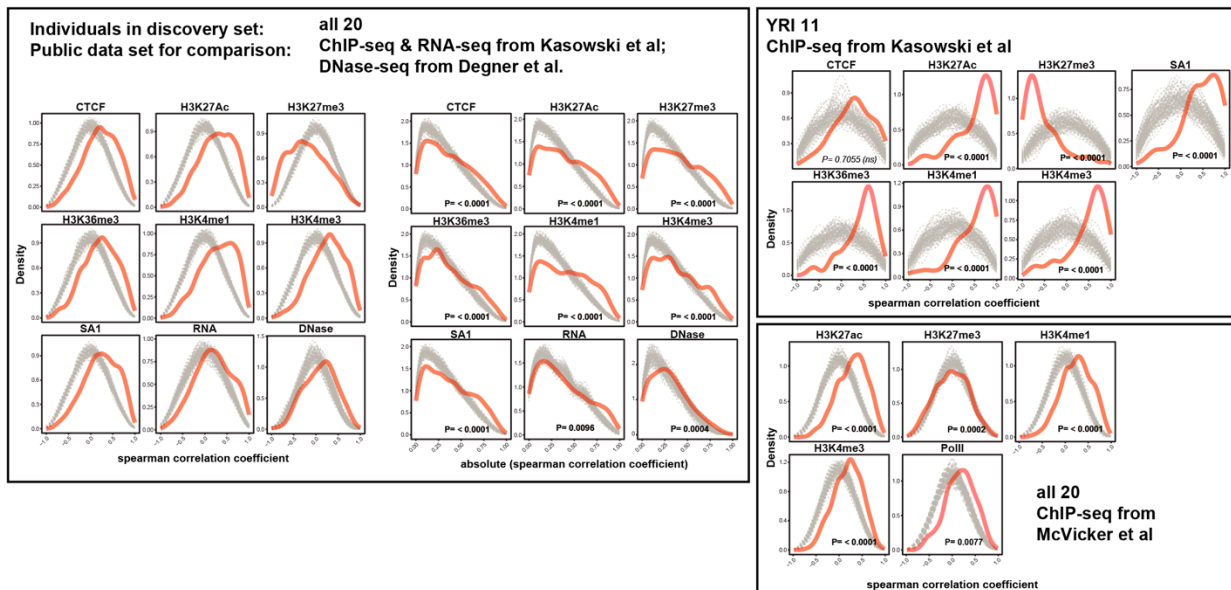


Figure S2.8. Correlations between DI, INS and multiple molecular phenotypes. Similar schema to **Figure 2.7c**, but focusing on DI in (a), and INS in (b).

a

Type of variable regions: FIRE



b

Type of variable regions: contact matrix

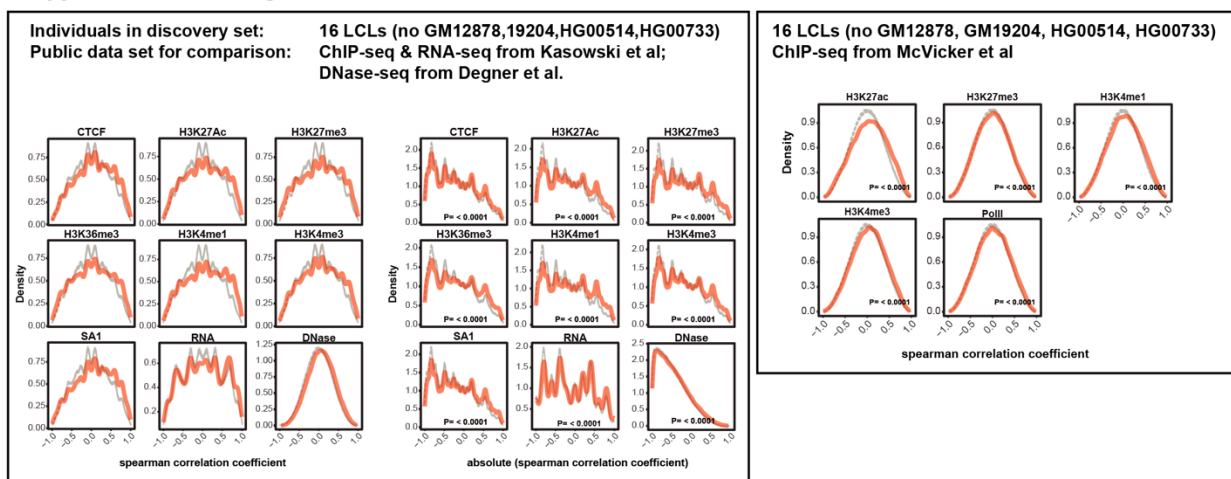


Figure S2.9. Correlations between FIRE, interaction frequency and multiple molecular phenotypes. Similar schema to Figure 2.7c, but focusing on FIRE in (a), and contact frequency (examining the anchor bins of variable matrix cells) in (b).

Figure S2.10. 3D chromatin QTLs. (a) QQ plots for each QTL search. In each QQ plot, the X-axis is the $-\log_{10}$ theoretical quantiles calculated from the uniform distribution. The Y-axis is the $-\log_{10}$ p -value calculated from linear mixed effects model for each type of QTL search. (b) Genotype trend for bins with positive DI (left), negative DI (right), and all QTLs (right). (c) Number of direct overlaps between QTL sets. (d) Similar schema to **Figure 2.4e**, but showing FIRE, INS, DI score (indicated on the Y axis) as a function of genotype for each QTL set as indicated above each column.

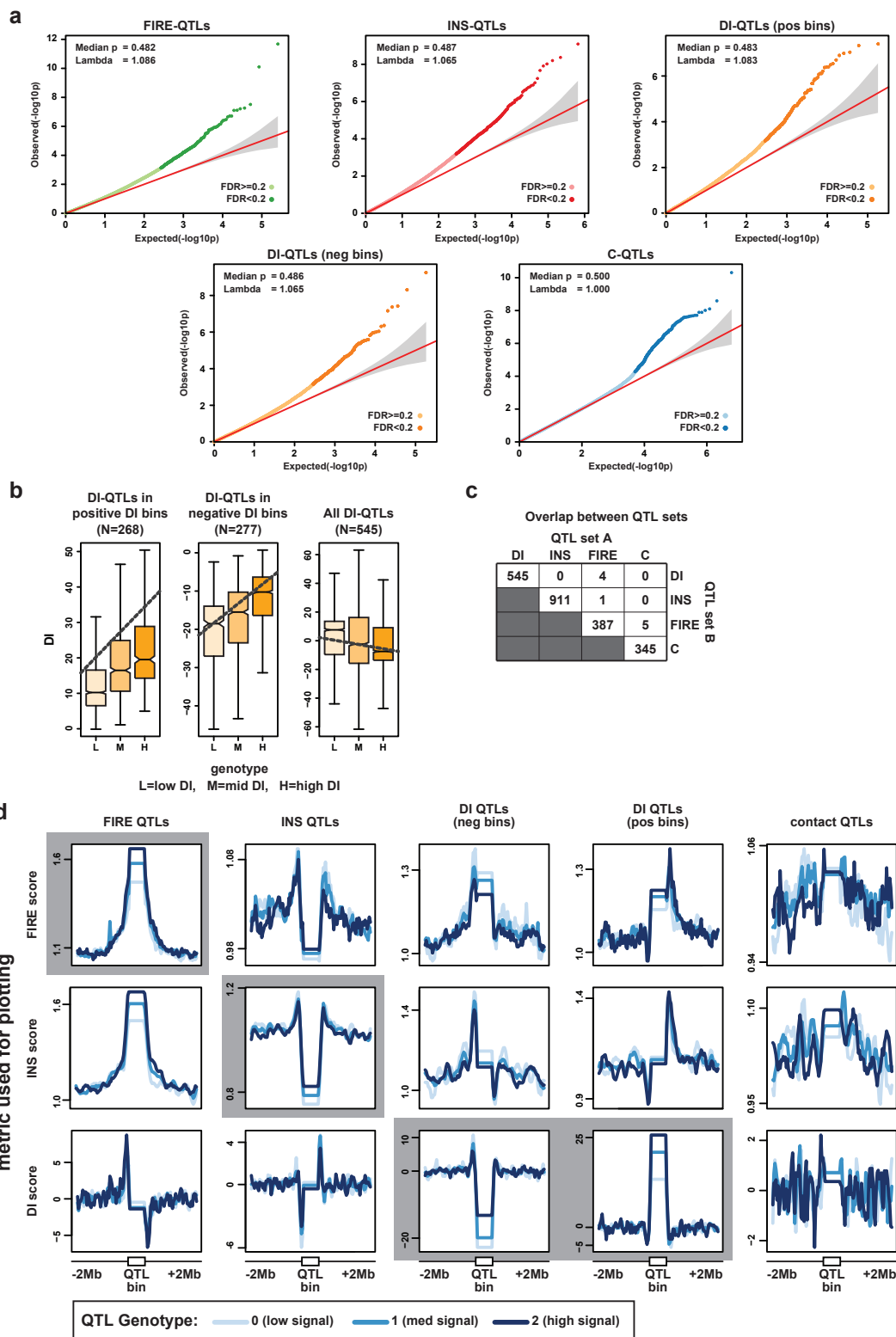
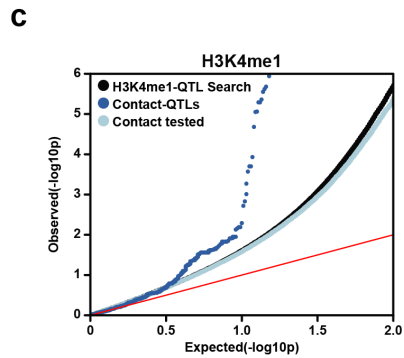
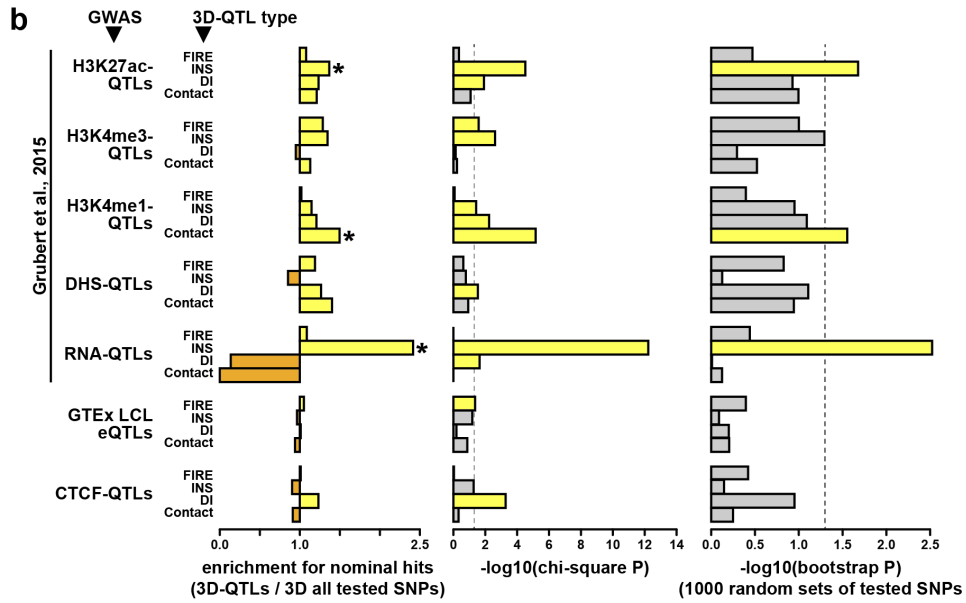
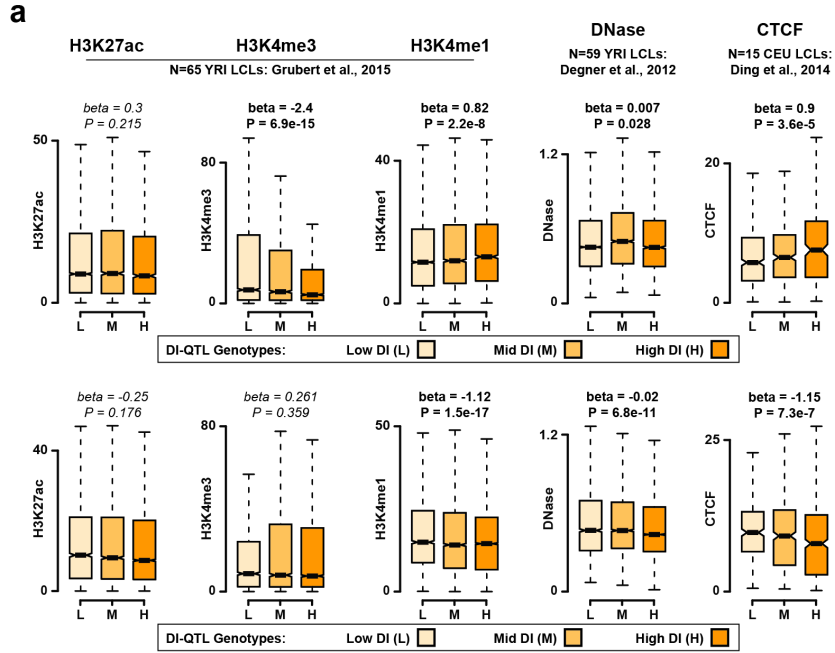


Figure S2.11. Influence of 3D chromatin QTLs on epigenomic and disease phenotypes. ((a) Similar schema to Figure 2.5a, but showing DI-QTLs in positive bins (top) and negative DI bins (bottom). (b) Left subpanel shows the enrichment for 3D QTL SNPs with nominal significance in the indicated epigenetic or eQTL study calculated as follows: (fraction of indicated 3D QTL SNPs with nominal significance in the indicated molQTL study) / (fraction of SNPs tested in the indicated 3D QTL search with nominal significance in the indicated molQTL study). Right panel shows the proportion of 1,000 random subsets selected from the tested SNPs with enrichment values higher than the indicated true QTL set. (c) QQ plot shows the results of H3K4me1 QTL search from Grubert et al., with all tested SNPs shown as black points, and two subsets as follows: SNPs also tested in our C-QTL search (light blue), and SNP called as C-QTLs or in perfect LD with C-QTLs in the same 40Kb bin (dark blue).



2.8 Acknowledgements

Chapter 2, in full, is a manuscript submitted as "Common DNA sequence variation influences 3-dimensional conformation of the human genome". David U. Gorkin, Yunjiang Qiu, Ming Hu, Kipper Fletez-Brant, Tristin Liu, Anthony D. Schmitt, Amina Noor, Joshua Chiou, Kyle J Gaulton, Jonathan Sebat, Yun Li, Kasper D. Hansen, and Bing Ren. The dissertation author was the primary investigator and author of this paper.

2.9 References

1. Gorkin, D. U., Leung, D. & Ren, B. The 3D Genome in Transcriptional Regulation and Pluripotency. *Cell Stem Cell* 14, 762–775 (2014).
2. Dekker, J. & Mirny, L. The 3D Genome as Moderator of Chromosomal Communication. *Cell* 164, 1110–1121 (2016).
3. Bouwman, B. A. M. & De Laat, W. Getting the genome in shape: the formation of loops, domains and compartments. *Genome Biol* 16, 154–9 (2015).
4. Pope, B. D., Ryba, T., Dileep, V., Yue, F., Wu, W., Denas, O., Vera, D. L., Wang, Y., Hansen, R. S., Canfield, T. K., Thurman, R. E., Cheng, Y., Gülsoy, G., Dennis, J. H., Snyder, M. P., Stamatoyannopoulos, J. A., Taylor, J., Hardison, R. C., Kahveci, T., Ren, B. & Gilbert, D. M. Topologically associating domains are stable units of replication-timing regulation. *Nature* 515, 402–405 (2015).
5. Dileep, V., Ay, F., Sima, J., Vera, D. L., Noble, W. S. & Gilbert, D. M. Topologically associating domains and their long-range contacts are established during early G1 coincident with the establishment of the replication-timing program. *Genome Res.* 25, 1104–1113 (2015).
6. Engreitz, J. M., Pandya-Jones, A., McDonel, P., Shishkin, A., Sirokman, K., Surka, C., Kadri, S., Xing, J., Goren, A., Lander, E. S., Plath, K. & Guttman, M. The Xist lncRNA exploits three-dimensional genome architecture to spread across the X chromosome. *Science* 341, 1237973–1237973 (2013).
7. Crane, E., Bian, Q., McCord, R. P., Lajoie, B. R., Wheeler, B. S., Ralston, E. J., Uzawa, S., Dekker, J. & Meyer, B. J. Condensin-driven remodelling of X chromosome topology during dosage compensation. *Nature* 523, 240–244 (2015).
8. Giorgetti, L., Lajoie, B. R., Carter, A. C., Attia, M., Zhan, Y., Xu, J., Chen, C.-J., Kaplan, N., Chang, H. Y., Heard, E. & Dekker, J. Structural organization of the inactive X chromosome in the mouse. *Nature* 535, 575–579 (2016).
9. More than just a focus: The chromatin response to DNA damage and its role in genome integrity maintenance. *Nat Cell Biol* 13, 1161–1169 (2011).
10. Organizing DNA repair in the nucleus: DSBs hit the road. *Curr. Opin. Cell Biol.* 46, 1–8 (2017).
11. Dekker, J., Rippe, K., Dekker, M. & Kleckner, N. Capturing chromosome conformation. *Science* 295, 1306–1311 (2002).
12. Lieberman-Aiden, E., van Berkum, N. L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., Amit, I., Lajoie, B. R., Sabo, P. J., Dorschner, M. O., Sandstrom, R., Bernstein, B., Bender, M. A., Groudine, M., Gnirke, A., Stamatoyannopoulos, J.,

- Mirny, L. A., Lander, E. S. & Dekker, J. Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome. *Science* 326, 289–293 (2009).
13. Rao, S. S. P., Huntley, M. H., Durand, N. C., Stamenova, E. K., Bochkov, I. D., Robinson, J. T., Sanborn, A. L., Machol, I., Omer, A. D., Lander, E. S. & Aiden, E. L. A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping. *Cell* 159, 1665–1680 (2014).
 14. The hierarchy of the 3D genome. *Molecular Cell* 49, 773–782 (2013).
 15. Dixon, J. R., Selvaraj, S., Yue, F., Kim, A., Li, Y., Shen, Y., Hu, M., Liu, J. S. & Ren, B. Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* 485, 376–380 (2012).
 16. Dixon, J. R., Gorkin, D. U. & Ren, B. Chromatin Domains: The Unit of Chromosome Organization. *Molecular Cell* 62, 668–680 (2016).
 17. 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. & Heard, E. Spatial partitioning of the regulatory landscape of the X-inactivation centre. *Nature* 485, 381–385 (2012).
 18. Schmitt, A. D., Hu, M., Jung, I., Xu, Z., Qiu, Y., Tan, C. L., Li, Y., Lin, S., Lin, Y., Barr, C. L. & Ren, B. A Compendium of Chromatin Contact Maps Reveals Spatially Active Regions in the Human Genome. *CellReports* 17, 2042–2059 (2016).
 19. Yan, J., Chen, S.-A. A., Local, A., Liu, T., Qiu, Y., Dorighi, K. M., Preissl, S., Rivera, C. M., Wang, C., Ye, Z., Ge, K., Hu, M., Wysocka, J. & Ren, B. Histone H3 lysine 4 monomethylation modulates long-range chromatin interactions at enhancers. *Cell Res.* 28, 387–387 (2018).
 20. Dixon, J. R., Jung, I., Selvaraj, S., Shen, Y., Antosiewicz-Bourget, J. E., Lee, A. Y., Ye, Z., Kim, A., Rajagopal, N., Xie, W., Diao, Y., Liang, J., Zhao, H., Lobanenko, V. V., Ecker, J. R., Thomson, J. A. & Ren, B. Chromatin architecture reorganization during stem cell differentiation. *Nature* 518, 331–336 (2015).
 21. Naumova, N., Imaev, M., Fudenberg, G., Zhan, Y., Lajoie, B. R., Mirny, L. A. & Dekker, J. Organization of the mitotic chromosome. *Science* 342, 948–953 (2013).
 22. Chromatin extrusion explains key features of loop and domain formation in wild-type and engineered genomes. *Proc. Natl. Acad. Sci. U.S.A.* 112, E6456–65 (2015).
 23. Jin, F., Li, Y., Dixon, J. R., Selvaraj, S., Ye, Z., Lee, A. Y., Yen, C.-A., Schmitt, A. D., Espinoza, C. A. & Ren, B. A high-resolution map of the three-dimensional chromatin interactome in human cells. *Nature* 503, 290–294 (2013).
 24. Darrow, E. M., Huntley, M. H., Dudchenko, O., Stamenova, E. K., Durand, N. C.,

- Sun, Z., Huang, S.-C., Sanborn, A. L., Machol, I., Shamim, M., Seberg, A. P., Lander, E. S., Chadwick, B. P. & Aiden, E. L. Deletion of DXZ4 on the human inactive X chromosome alters higher-order genome architecture. *Proc Natl Acad Sci USA* 113, E4504–E4512 (2016).
25. Cohesin Loss Eliminates All Loop Domains. *Cell* 171, 305–320.e24 (2017).
 26. Subtle changes in chromatin loop contact propensity are associated with differential gene regulation and expression. *Nature Communications* 10, 1054 (2019).
 27. Krijger, P. H. L. & De Laat, W. Regulation of disease-associated gene expression in the 3D genome. *Nat. Rev. Mol. Cell Biol.* 17, 771–782 (2016).
 28. 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. & Mundlos, S. Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions. *Cell* 161, 1012–1025 (2015).
 29. Franke, M., Ibrahim, D. M., Andrey, G., Schwarzer, W., Heinrich, V., Schöpflin, R., Kraft, K., Kempfer, R., Jerković, I., Chan, W.-L., Spielmann, M., Timmermann, B., Wittler, L., Kurth, I., Cambiaso, P., Zuffardi, O., Houge, G., Lambie, L., Brancati, F., Pombo, A., Vingron, M., Spitz, F. & Mundlos, S. Formation of new chromatin domains determines pathogenicity of genomic duplications. *Nature* 538, 265–269 (2016).
 30. Activation of proto-oncogenes by disruption of chromosome neighborhoods. *Science* 351, 1454–1458 (2016).
 31. Insulator dysfunction and oncogene activation in IDH mutant gliomas. *Nature* 529, 110–114 (2016).
 32. Obesity-associated variants within FTO form long-range functional connections with IRX3. *Nature* 507, 371–375 (2014).
 33. HERC2 rs12913832 modulates human pigmentation by attenuating chromatin-loop formation between a long-range enhancer and the OCA2 promoter. *Genome Res.* 22, 446–455 (2012).
 34. 1000 Genomes Project Consortium, Abecasis, G. R., Auton, A., Brooks, L. D., DePristo, M. A., Durbin, R. M., Handsaker, R. E., Kang, H. M., Marth, G. T. & McVean, G. A. An integrated map of genetic variation from 1,092 human genomes. *Nature* 491, 56–65 (2012).
 35. Kilpinen, H., Waszak, S. M., Gschwind, A. R., Raghav, S. K., Witwicki, R. M., Orioli, A., Migliavacca, E., Wiederkehr, M., Gutierrez-Arcelus, M., Panousis, N. I., Yurovsky, A., Lappalainen, T., Romano-Palumbo, L., Planchon, A., Bielser, D., Bryois, J., Padioleau, I., Udin, G., Thurnheer, S., Hacker, D., Core, L. J., Lis, J. T., Hernandez,

- N., Reymond, A., Deplancke, B. & Dermitzakis, E. T. Coordinated effects of sequence variation on DNA binding, chromatin structure, and transcription. *Science* 342, 744–747 (2013).
36. Lappalainen, T., Sammeth, M., Friedländer, M. R., 't Hoen, P. A. C., Monlong, J., Rivas, M. A., González-Porta, M., Kurbatova, N., Griebel, T., Ferreira, P. G., Barann, M., Wieland, T., Greger, L., van Iterson, M., Almlöf, J., Ribeca, P., Pulyakhina, I., Esser, D., Giger, T., Tikhonov, A., Sultan, M., Bertier, G., MacArthur, D. G., Lek, M., Lizano, E., Buermans, H. P. J., Padioleau, I., Schwarzmayr, T., Karlberg, O., Ongen, H., Kilpinen, H., Beltran, S., Gut, M., Kahlem, K., Amstislavskiy, V., Stegle, O., Pirinen, M., Montgomery, S. B., Donnelly, P., McCarthy, M. I., Flicek, P., Strom, T. M., Geuvadis Consortium, Lehrach, H., Schreiber, S., Sudbrak, R., Carracedo, A., Antonarakis, S. E., Häslér, R., Syvänen, A.-C., van Ommen, G.-J., Brazma, A., Meitinger, T., Rosenstiel, P., Guigo, R., Gut, I. G., Estivill, X. & Dermitzakis, E. T. Transcriptome and genome sequencing uncovers functional variation in humans. *Nature* 501, 506–511 (2013).
 37. 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. & Pritchard, J. K. DNase I sensitivity QTLs are a major determinant of human expression variation. *Nature* 482, 390–394 (2012).
 38. Grubert, F., Zaugg, J. B., Kasowski, M., Ursu, O., Spacek, D. V., Martin, A. R., Greenside, P., Srivas, R., Phanstiel, D. H., Pekowska, A., Heidari, N., Euskirchen, G., Huber, W., Pritchard, J. K., Bustamante, C. D., Steinmetz, L. M., Kundaje, A. & Snyder, M. Genetic Control of Chromatin States in Humans Involves Local and Distal Chromosomal Interactions. *Cell* 162, 1051–1065 (2015).
 39. McVicker, G., van de Geijn, B., Degner, J. F., Cain, C. E., Banovich, N. E., Raj, A., Lewellen, N., Myrthil, M., Gilad, Y. & Pritchard, J. K. Identification of genetic variants that affect histone modifications in human cells. *Science* 342, 747–749 (2013).
 40. Methylation QTLs are associated with coordinated changes in transcription factor binding, histone modifications, and gene expression levels. *PLoS Genet.* 10, e1004663 (2014).
 41. Kasowski, M., Kyriazopoulou-Panagiotopoulou, S., Grubert, F., Zaugg, J. B., Kundaje, A., Liu, Y., Boyle, A. P., Zhang, Q. C., Zakharia, F., Spacek, D. V., Li, J., Xie, D., Olarerin-George, A., Steinmetz, L. M., Hogenesch, J. B., Kellis, M., Batzoglou, S. & Snyder, M. Extensive variation in chromatin states across humans. *Science* 342, 750–752 (2013).
 42. Selvaraj, S., Dixon, J., Bansal, V. & Ren, B. Whole-genome haplotype reconstruction using proximity-ligation and shotgun sequencing. *Nat Biotechnol* 31, 1111–1118 (2013).
 43. The International HapMap Project. *Nature* 426, 789–796 (2003).

44. van de Geijn, B., McVicker, G., Gilad, Y. & Pritchard, J. K. WASP: allele-specific software for robust molecular quantitative trait locus discovery. *Nat. Methods* 12, 1061–1063 (2015).
45. Population differences in the rate of proliferation of international HapMap cell lines. *Am. J. Hum. Genet.* 87, 829–833 (2010).
46. Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W. & Smyth, G. K. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research* 43, e47–e47 (2015).
47. Dixon, J. R., Xu, J., Dileep, V., Zhan, Y., Song, F., Le, V. T., x00131, G. G. X. R. Y. X. M., Chakraborty, A., Bann, D. V., Wang, Y., Clark, R., Zhang, L., Yang, H., Liu, T., Iyyanki, S., An, L., Pool, C., Sasaki, T., Rivera-Mulia, J. C., Ozadam, H., Lajoie, B. R., Kaul, R., Buckley, M., Lee, K., Diegel, M., Pezic, D., Ernst, C., Hadjur, S., Odom, D. T., Stamatoyannopoulos, J. A., Broach, J. R., Hardison, R. C., Ay, F., Noble, W. S., Dekker, J., Gilbert, D. M. & Yue, F. Integrative detection and analysis of structural variation in cancer genomes. *Nat. Genet.* 4, 1–16 (2018).
48. Data-driven hypothesis weighting increases detection power in genome-scale multiple testing. *Nat Meth* 13, 577–580 (2016).
49. Zuin, J., Dixon, J. R., van der Reijden, M. I. J. A., Ye, Z., Kolovos, P., Brouwer, R. W. W., van de Corput, M. P. C., van de Werken, H. J. G., Knoch, T. A., van IJcken, W. F. J., Grosveld, F. G., Ren, B. & Wendt, K. S. Cohesin and CTCF differentially affect chromatin architecture and gene expression in human cells. *Proc. Natl. Acad. Sci. U.S.A.* 111, 996–1001 (2014).
50. Sofueva, S., Yaffe, E., Chan, W. C., Georgopoulou, D., Vietri Rudan, M., Rudan, M. V., Mira-Bontenbal, H., Bontenbal, H. M., Pollard, S. M., Schroth, G. P., Tanay, A. & Hadjur, S. Cohesin-mediated interactions organize chromosomal domain architecture. *The EMBO Journal* 32, 3119–3129 (2013).
51. CTCF-Mediated Human 3D Genome Architecture Reveals Chromatin Topology for Transcription. *Cell* 163, 1611–1627 (2015).
52. RNA splicing is a primary link between genetic variation and disease. *Science* 352, 600–604 (2016).
53. Ding, Z., Ni, Y., Timmer, S. W., Lee, B.-K., Battenhouse, A., Louzada, S., Yang, F., Dunham, I., Crawford, G. E., Lieb, J. D., Durbin, R., Iyer, V. R. & Birney, E. Quantitative genetics of CTCF binding reveal local sequence effects and different modes of X-chromosome association. *PLoS Genet.* 10, e1004798 (2014).
54. The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Research* 47, D1005–D1012 (2019).

55. de Lange, K. M., Moutsianas, L., Lee, J. C., Lamb, C. A., Luo, Y., Kennedy, N. A., Jostins, L., Rice, D. L., Gutierrez-Achury, J., Ji, S.-G., Heap, G., Nimmo, E. R., Edwards, C., Henderson, P., Mowat, C., Sanderson, J., Satsangi, J., Simmons, A., Wilson, D. C., Tremelling, M., Hart, A., Mathew, C. G., Newman, W. G., Parkes, M., Lees, C. W., Uhlig, H., Hawkey, C., Prescott, N. J., Ahmad, T., Mansfield, J. C., Anderson, C. A. & Barrett, J. C. Genome-wide association study implicates immune activation of multiple integrin genes in inflammatory bowel disease. *Nat. Genet.* 49, 256–261 (2017).
56. Wood, A. R., Esko, T., Yang, J., Vedantam, S., Pers, T. H., Gustafsson, S., Chu, A. Y., Estrada, K., Luan, J., Kutalik, Z., Amin, N., Buchkovich, M. L., Croteau-Chonka, D. C., Day, F. R., Duan, Y., Fall, T., Fehrmann, R., Ferreira, T., Jackson, A. U., Karjalainen, J., Lo, K. S., Locke, A. E., Mägi, R., Mihailov, E., Porcu, E., Randall, J. C., Scherag, A., Vinkhuyzen, A. A. E., Westra, H.-J., Winkler, T. W., Workalemahu, T., Zhao, J. H., Absher, D., Albrecht, E., Anderson, D., Baron, J., Beekman, M., Demirkan, A., Ehret, G. B., Feenstra, B., Feitosa, M. F., Fischer, K., Fraser, R. M., Goel, A., Gong, J., Justice, A. E., Kanoni, S., Kleber, M. E., Kristiansson, K., Lim, U., Lotay, V., Lui, J. C., Mangino, M., Mateo Leach, I., Medina-Gomez, C., Nalls, M. A., Nyholt, D. R., Palmer, C. D., Pasko, D., Pechlivanis, S., Prokopenko, I., Ried, J. S., Ripke, S., Shungin, D., Stančáková, A., Strawbridge, R. J., Sung, Y. J., Tanaka, T., Teumer, A., Trompet, S., van der Laan, S. W., van Setten, J., Van Vliet-Ostaptchouk, J. V., Wang, Z., Yengo, L., Zhang, W., Afzal, U., Arnlöv, J., Arscott, G. M., Bandinelli, S., Barrett, A., Bellis, C., Bennett, A. J., Berne, C., Blüher, M., Bolton, J. L., Böttcher, Y., Boyd, H. A., Bruinenberg, M., Buckley, B. M., Buyske, S., Caspersen, I. H., Chines, P. S., Clarke, R., Claudi-Boehm, S., Cooper, M., Daw, E. W., De Jong, P. A., Deelen, J., Delgado, G., Denny, J. C., Dhonukshe-Rutten, R., Dimitriou, M., Doney, A. S. F., Dörr, M., Eklund, N., Eury, E., Folkersen, L., Garcia, M. E., Geller, F., Giedraitis, V., Go, A. S., Grallert, H., Grammer, T. B., Gräßler, J., Grönberg, H., de Groot, L. C. P. G. M., Groves, C. J., Haessler, J., Hall, P., Haller, T., Hallmans, G., Hannemann, A., Hartman, C. A., Hassinen, M., Hayward, C., Heard-Costa, N. L., Helmer, Q., Hemani, G., Henders, A. K., Hillege, H. L., Hlatky, M. A., Hoffmann, W., Hoffmann, P., Holmen, O., Houwing-Duistermaat, J. J., Illig, T., Isaacs, A., James, A. L., Jeff, J., Johansen, B., Johansson, Å., Jolley, J., Juliusdottir, T., Junttila, J., Kho, A. N., Kinnunen, L., Klopp, N., Kocher, T., Kratzer, W., Lichtner, P., Lind, L., Lindstrom, J., Lobbens, S., Lorentzon, M., Lu, Y., Lyssenko, V., Magnusson, P. K. E., Mahajan, A., Maillard, M., McArdle, W. L., McKenzie, C. A., McLachlan, S., McLaren, P. J., Menni, C., Merger, S., Milani, L., Moayyeri, A., Monda, K. L., Morken, M. A., Müller, G., Müller-Nurasyid, M., Musk, A. W., Narisu, N., Nauck, M., Nolte, I. M., Nöthen, M. M., Oozageer, L., Pilz, S., Rayner, N. W., Renstrom, F., Robertson, N. R., Rose, L. M., Roussel, R., Sanna, S., Scharnagl, H., Scholtens, S., Schumacher, F. R., Schunkert, H., Scott, R. A., Sehmi, J., Seufferlein, T., Shi, J., Silventoinen, K., Smit, J. H., Smith, A. V., Smolonska, J., Stanton, A. V., Stirrups, K., Stott, D. J., Stringham, H. M., Sundström, J., Swertz, M. A., Syvänen, A.-C., Tayo, B. O., Thorleifsson, G., Tyrer, J. P., van Dijk, S., van Schoor, N. M., van der Velde, N., van Heemst, D., van Oort, F. V. A., Vermeulen, S. H., Verweij, N., Vonk, J. M., Waite, L. L., Waldenberger, M., Wennauer, R., Wilkens, L. R., Willenborg, C.,

Wilsgaard, T., Wojczynski, M. K., Wong, A., Wright, A. F., Zhang, Q., Arveiler, D., Bakker, S. J. L., Beilby, J., Bergman, R. N., Bergmann, S., Biffar, R., Blangero, J., Boomsma, D. I., Bornstein, S. R., Bovet, P., Brambilla, P., Brown, M. J., Campbell, H., Caulfield, M. J., Chakravarti, A., Collins, R., Collins, F. S., Crawford, D. C., Cupples, L. A., Danesh, J., de Faire, U., Ruyter, den, H. M., Erbel, R., Erdmann, J., Eriksson, J. G., Farrall, M., Ferrannini, E., Ferrières, J., Ford, I., Forouhi, N. G., Forrester, T., Gansevoort, R. T., Gejman, P. V., Gieger, C., Golay, A., Gottesman, O., Gudnason, V., Gyllensten, U., Haas, D. W., Hall, A. S., Harris, T. B., Hattersley, A. T., Heath, A. C., Hengstenberg, C., Hicks, A. A., Hindorff, L. A., Hingorani, A. D., Hofman, A., Hovingh, G. K., Humphries, S. E., Hunt, S. C., Hyppönen, E., Jacobs, K. B., Jarvelin, M.-R., Jousilahti, P., Jula, A. M., Kaprio, J., Kastelein, J. J. P., Kayser, M., Kee, F., Keinänen-Kiukaanniemi, S. M., Kiemenev, L. A., Kooner, J. S., Kooperberg, C., Koskinen, S., Kovacs, P., Kraja, A. T., Kumari, M., Kuusisto, J., Lakka, T. A., Langenberg, C., Le Marchand, L., Lehtimäki, T., Lupoli, S., Madden, P. A. F., Männistö, S., Manunta, P., Marette, A., Matise, T. C., McKnight, B., Meitinger, T., Moll, F. L., Montgomery, G. W., Morris, A. D., Morris, A. P., Murray, J. C., Nelis, M., Ohlsson, C., Oldehinkel, A. J., Ong, K. K., Ouwehand, W. H., Pasterkamp, G., Peters, A., Pramstaller, P. P., Price, J. F., Qi, L., Raitakari, O. T., Rankinen, T., Rao, D. C., Rice, T. K., Ritchie, M., Rudan, I., Salomaa, V., Samani, N. J., Saramies, J., Sarzynski, M. A., Schwarz, P. E. H., Sebert, S., Sever, P., Shuldiner, A. R., Sinisalo, J., Steinthorsdottir, V., Stolk, R. P., Tardif, J.-C., Tonjes, A., Tremblay, A., Tremoli, E., Virtamo, J., Vohl, M.-C., Electronic Medical Records and Genomics (eMEMERGE) Consortium, MIGen Consortium, PAGEGE Consortium, LifeLines Cohort Study, Amouyel, P., Asselbergs, F. W., Assimes, T. L., Bochud, M., Boehm, B. O., Boerwinkle, E., Bottinger, E. P., Bouchard, C., Cauchi, S., Chambers, J. C., Chanock, S. J., Cooper, R. S., de Bakker, P. I. W., Dedoussis, G., Ferrucci, L., Franks, P. W., Froguel, P., Groop, L. C., Haiman, C. A., Hamsten, A., Hayes, M. G., Hui, J., Hunter, D. J., Hveem, K., Jukema, J. W., Kaplan, R. C., Kivimäki, M., Kuh, D., Laakso, M., Liu, Y., Martin, N. G., März, W., Melbye, M., Moebus, S., Munroe, P. B., Njølstad, I., Oostra, B. A., Palmer, C. N. A., Pedersen, N. L., Perola, M., Perusse, L., Peters, U., Powell, J. E., Power, C., Quertermous, T., Rauramaa, R., Reinmaa, E., Ridker, P. M., Rivadeneira, F., Rotter, J. I., Saaristo, T. E., Saleheen, D., Schlessinger, D., Slagboom, P. E., Snieder, H., Spector, T. D., Strauch, K., Stumvoll, M., Tuomilehto, J., Uusitupa, M., van der Harst, P., Völzke, H., Walker, M., Wareham, N. J., Watkins, H., Wichmann, H.-E., Wilson, J. F., Zanen, P., Deloukas, P., Heid, I. M., Lindgren, C. M., Mohlke, K. L., Speliotes, E. K., Thorsteinsdottir, U., Barroso, I., Fox, C. S., North, K. E., Strachan, D. P., Beckmann, J. S., Berndt, S. I., Boehnke, M., Borecki, I. B., McCarthy, M. I., Metspalu, A., Stefansson, K., Uitterlinden, A. G., van Duijn, C. M., Franke, L., Willer, C. J., Price, A. L., Lettre, G., Loos, R. J. F., Weedon, M. N., Ingelsson, E., O'Connell, J. R., Abecasis, G. R., Chasman, D. I., Goddard, M. E., Visscher, P. M., Hirschhorn, J. N. & Frayling, T. M. Defining the role of common variation in the genomic and biological architecture of adult human height. *Nat. Genet.* 46, 1173–1186 (2014).

57. Genetic studies of body mass index yield new insights for obesity biology. *Nature* 518, 197–206 (2015).

58. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv q-bio.GN, (2013).
59. Hu, M., Deng, K., Selvaraj, S., Qin, Z., Ren, B. & Liu, J. S. HiCNorm: removing biases in Hi-C data via Poisson regression. *Bioinformatics* 28, 3131–3133 (2012).
60. McLean, C. Y., Bristor, D., Hiller, M., Clarke, S. L., Schaar, B. T., Lowe, C. B., Wenger, A. M. & Bejerano, G. GREAT improves functional interpretation of cis-regulatory regions. *Nat Biotechnol* 28, 495–501 (2010).
61. Johnson, W. E., Li, C. & Rabinovic, A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* 8, 118–127 (2007).
62. Kuznetsova, A., Brockhoff, P. B. & Christensen, R. H. B. ImerTest Package: Tests in Linear Mixed Effects Models. *J STAT SOFTW* 82, (2017).
63. Zhang, Y., Liu, T., Meyer, C. A., Eeckhoute, J., Johnson, D. S., Bernstein, B. E., Nusbaum, C., Myers, R. M., Brown, M., Li, W. & Liu, X. S. Model-based Analysis of ChIP-Seq (MACS). *Genome Biol* 9, 1–9 (2008).
64. Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M. & Gingeras, T. R. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* 29, 15–21 (2013).
65. Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D. R., Pimentel, H., Salzberg, S. L., Rinn, J. L. & Pachter, L. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat Protoc* 7, 562–578 (2012).
66. Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y. C., Laslo, P., Cheng, J. X., Murre, C., Singh, H. & Glass, C. K. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Molecular Cell* 38, 576–589 (2010).
67. Ramírez, F., Dündar, F., Diehl, S., acids, B. G. N.2014. deepTools: a flexible platform for exploring deep-sequencing data. academic.oup.com

Chapter 3: Dynamic 3D chromatin organization during differentiation of human embryonic stem cells to pancreatic progenitor cells

3.1 Abstract

Cellular differentiation is characterized by changes in gene expression, which in turn are facilitated by remodeling of chromatin structure. There is emerging evidence that the three-dimensional (3D) architecture of genomes contributes to the regulation of cell-specific transcription programs, yet how chromatin reorganizes and how this change contributes to dynamic gene expression during cell differentiation are still poorly understood. Here we mapped the 3D architecture of genomes using *in situ* Hi-C at defined stages during differentiation of human embryonic stem cells to pancreatic progenitor cells, and in human cadaveric islets. We observed dynamic 3D genome organization at multiple levels, including compartments, topological associated domains (TADs), and chromatin loops. Specifically, we identified 15,448 chromatin loops of which 5,452 are dynamic during differentiation. The anchors of dynamic chromatin loops are enriched with developmental stage-specific gene expression, chromatin modifications, and transcription factor (TF) binding. Furthermore, we identified chromatin interaction hubs that interact with multiple regions through chromatin loops. We found that a large number of hubs are stage-specific and harbor key developmental genes. Chromatin hubs are enriched for super-enhancers, active gene expression as well as binding of lineage-determining TFs, suggesting essential roles for hubs in the establishment of stage-specific transcriptional programs. Altogether, our study revealed dynamic chromatin organization during developmental lineage progression and provided insight into how

higher-order chromatin structure contributes to transcriptional regulation and lineage specification.

3.2 Introduction

The human genome is highly organized at multiple levels, ranging from chromosome territories that can be seen under microscopes¹ to interactions between individual loci that can only be seen by molecular techniques^{2,3}. Technologies based on chromatin conformation capture, including 5C⁴, ChIA-PET⁵, and Hi-C⁶ have provided important insights into the general principles of chromatin organization. At megabase scales, the genome is organized into TADs^{7,8}. Genomic loci within the same TADs tend to interact more with each other than loci that located in different TADs. At a finer resolution, chromatin loops are evidenced by the spatial proximity of two distal genomic elements, for example, enhancers and promoters^{2,3}. Despite the observation of chromatin loops, the mechanisms underlying loop formation remain unclear. While emerging evidence supports that chromatin loops are formed by CCCTC-binding factor (CTCF) and cohesin through loop extrusion⁹, transcription factors other than CTCF/cohesion have also been found to be associated with chromatin loops¹⁰⁻¹², suggesting diverse mechanisms for loop formation.

At all levels, 3D genome organization appears to be associated with multiple cellular processes and genome functions, in particular, transcriptional regulation². Proper folding of the genome is important for normal development and cellular differentiation. Disruption of chromatin loops linking enhancers and promoters could lead to dysregulation of gene expression and disease¹³. Thus, it is crucial to understand chromatin conformation and its relationship to transcriptional regulation in various contexts. Though chromatin conformation has been investigated in multiple studies, including different cell-type and human tissues^{14,15}, how chromatin organizations change

during cellular differentiation is still poorly understood. Previous studies either lacked the sequencing depth to achieve fine resolutions that enable detection of chromatin loops or focused on targeted regions and failed to provide a comprehensive picture^{10,11,16}.

Here, we performed *in situ* Hi-C with deep sequencing to chart chromatin conformation during cellular differentiation in the context of pancreatic development. We showed that stage-specific chromatin loops are formed during differentiation and are associated with developmental gene expression and histone modification. Importantly, we found that TFs responsible for lineage specification are enriched at stage-specific loop anchors. Moreover, we identified chromatin hubs that link multiple regions through chromatin loops and showed chromatin hubs are enriched for super-enhancers, active gene expression as well as binding of lineage-determining TFs. Taken together, our study revealed the putative roles of pioneer TFs in loop formation and shed lights into how chromatin conformation contributes to transcriptional regulation during differentiation.

3.3 Results

Mapping chromatin conformation during pancreatic differentiation

To study the dynamics of chromatin conformation during cell differentiation, we took advantage of the stepwise differentiation of human embryonic stem cells (hESC) towards pancreatic lineages^{17,18}. Specifically, we mapped chromatin interactions at four defined stages including hESCs, definitive endoderm (DE), primitive gut tube (GT), and pancreatic progenitor (PP) using *in situ* Hi-C (**Figure 3.1a**). To achieve high resolution, we sequenced each library to high depth (~1 billion intra-chromosomal long-range contacts per sample) in two biological replicates, allowing us to characterize not only larger-scale features of genome organization but also interactions between *cis*-regulatory elements. In addition to four defined stages, we also collected Hi-C data using the same protocol for human cadaveric islets from three donors and combined them with four stages for downstream analysis.

Complementing the Hi-C data, we have also generated ATAC-seq for the same set of samples and ChIP-seq for CTCF, Rad21, and important lineage-specification TFs including FOXA1, FOXA2, GATA4, GATA6, and PDX1 at the stages where the TF is expressed (**Figure 3.1a**). Also, we collected datasets that are publicly available for the same lineages, including ChIP-seq for H3K27ac, H3K4me3, H3K4me1, H3K27me3 to profile histone modification and RNA-seq for gene expression^{17,18}.

For each sample, we defined TADs using previously described approaches based on direction index (DI)⁷ and called chromatin loops adapting the previous method based on the local background¹⁹. On average, we identified ~5,000 chromatin loops per sample (FDR < 0.01; **Figure S3.1a**). Notably, there are dramatically more chromatin loops in the

DE stage than other stages, consistent with the previous observation that the DE cell genome harbors the largest number of active enhancers¹⁸. Two lines of evidence support the high confidence of chromatin loops. First, chromatin loops that were identified in each stage also show high reproducibility with ~60% chromatin loops called in both replicates (**Figure S3.1b**). Second, chromatin loops called in each stage show clear enrichment using aggregate peak analysis (APA) (**Figure 3.1c**). In total, 15,448 distinct chromatin loops were identified across all stages.

Linking enhancers to target genes via chromatin loops

Enhancers are distal cis-regulatory elements that regulate cell-type-specific gene expression. We have previously characterized enhancers during the same differentiation lineages and identified a set of stage-specific enhancers¹⁸. While enhancers play crucial roles in shaping the transcriptional profile, it remains a challenge to identify target genes for enhancers as they are often far away from the target genes. We hypothesized that chromatin loops could be used to link enhancers to target promoters as elements at two loop anchors are close in three-dimensional spaces.

To test if chromatin loops identified using Hi-C data could be used to assign enhancer to their target genes, we employed two independent approaches. First, we took advantage of our time-course data and reasoned that enhancers and promoters that are linked might change activity accordingly during the differentiation. As we expected, enhancer-promoter pairs that are linked by chromatin loops are significantly more likely to be correlated in terms of activity across stages comparing to random pairs (**Figure S3.1d**). Second, we compared enhancer-promoter pairs predicted by spatial proximity to enhancer-promoter pairs suggested by genetic evidence, in particular expression

quantitative trait loci (eQTL). We found that enhancers harboring pancreas eQTLs are significantly more likely to be in proximity with the associated genes in human islet samples (**Figure 3.1c**), suggesting enhancer-promoter pairs linked by chromatin loops are also consistent with genetic-based evidence.

Rewiring of chromatin loops during pancreatic differentiation

Cell differentiation is a highly regulated process accompanied by reshaping of transcriptional profiles. Specifically, previous reports observed that key developmental genes are expressed at certain stages¹⁷. For instance, FOXA2 plays crucial roles in endoderm development, and its activation at the DE stage is essential for proper lineage specification (**Figure 3.1b**). In addition to FOXA2 promoter, several enhancers surrounding FOXA2 gene are also activated, indicating a reshaping of the enhancer landscape as well. While this observation is consistent with previous studies that revealed a systematic change of transcription and enhancer profiles during pancreatic differentiation^{17,18}, it is not known how chromatin interactions are rewired.

To systematically chart the changes in chromatin interactions during the developmental time course, we focused on the 15,448 chromatin loops identified across the time-course and examined dynamic changes of those loops. Interestingly, we observed several chromatin loops anchoring at the FOXA2 promoter that appeared at the DE stage when the FOXA2 gene is activated (**Figure 3.1b**), consistent with a potential role of chromatin loops in the activation of FOXA2 gene during the time course. After systematic characterization, we identified 5,452 chromatin loops that are dynamic during the differentiation while remaining 9,996 chromatin loops are static (**Figure 3.1d**).

Interestingly, we also found dynamic loops tend to have slightly smaller size compared to static loops (**Figure 3.1e**).

Stage-specific chromatin loops are associated with developmental gene expression programs

Given the 5,452 dynamic chromatin loops, we next asked how those loops change during the time course. K-means clustering revealed five distinct clusters of dynamic chromatin loops (**Figure 3.2a**). The majority of dynamic loops are found to peak at one stage. For example, loops in cluster 2 (C2) are at their maximum strength at the DE stage. The only exception is that cluster 3 (C3) loops, which correspond to the GT stage, display strong signals in multiple stages likely because GT is not a well-defined biological stage but an intermediate between DE and PP.

We next sought to test if stage-specific chromatin loops are associated with changes in transcription. We performed gene ontology (GO) analysis for each cluster of loops to identify genes that are enriched at or near the anchors of the chromatin loops in each cluster. Surprisingly, we found that different groups of developmental genes are enriched for each cluster of loops. For instance, loops in the DE cluster are enriched for genes involved in endoderm development (**Figure 3.2b**). Similarly, genes enriched at cluster 4 (C4) loop anchors are involved in such processes as pancreatic development. These results suggest that stage-specific chromatin loops may contribute to regulation of stage-specific gene expression programs.

Chromatin loops are believed to help regulate gene expression by bringing distal enhancer to the proximity of promoters. To further investigate how stage-specific chromatin loops coincide with stage-specific transcriptional profiles, we examined several

developmental regulator genes. In each case, dynamic chromatin loops are accompanied by gene activation at the same stage. For example, ONECUT1, an important regulator in pancreatic progenitor, is specifically expressed at PP stage, and a PP-specific chromatin loop links ONECUT1 to a PP-specific enhancer as marked by ATAC and H3K27ac. We also observed switching of enhancers for the same gene at different stages. For instance, GATA6 is expressed across multiple stages during differentiation, and it is linked to distinct putative enhancers at different stages by loops activated at different stages (**Figure 3.2c**). These results further support the potential roles of chromatin loops in activating stage-specific genes.

While we observed enrichment of specific gene sets in distinct clusters of loops, the relationship between gene expression levels and loop strength is not well characterized. To address this question, we examined gene expression changes across stages for each cluster of loops and observed a similar pattern of transcription levels as loop strength. For instance, the C2 loops are strongest at the DE stage, and genes near loop anchors in the C2 cluster are most strongly expressed at the DE stage (**Figure 3.2d**). This suggests that chromatin loops may contribute to the stage-specific expression of developmental genes. We also examined the relationship between chromatin loops and other genomic features. Interestingly, we observed similar patterns for histone marks H3K27ac, H3K4me1, as well as RNA expression level (**Figure 3.2d**). Taken together, rewiring of chromatin loops is accompanied by changes in both transcription and histone modification.

Lineage-specification TFs contribute to the formation of stage-specific loops

Given the observation that stage-specific chromatin loops are associated with stage-specific gene expression profiles, we next investigated the potential regulators of those dynamic chromatin loops. As TFs have been shown to be crucial in driving cell differentiation, we wondered if TFs also help establish dynamic chromatin loops. To identify potential TFs that drive the formation of dynamic chromatin loops, we performed motif enrichment analysis for loop anchors of each cluster. For these analyses, we tested ATAC peaks within loop anchors, because loop anchors are 10Kb size and thus not suitable for motif search. Surprisingly, we observed significant enrichment of known pioneer TFs within the loop anchors. For example, FOXA1/2 are known to be pioneer TFs for the DE stage, and FOXA motifs are highly over-presented in the C2 loops (**Figure 3.3a**). PDX1 is TF responsible for PE formation, and we observed significant enrichment of PDX1 motif in the C4 loops (**Figure 3.3a**).

Because the presence of motif does not necessarily mean that a TF actually binds a given site at a specific stage, we further validated results of motif enrichment using TF ChIP-seq data which could measure TF binding directly. In particular, we tested if selected TFs are more likely to bind to loops for one cluster using static loops as the background. We first tested CTCF and cohesin, which are essential to establish chromatin loops. Surprisingly, CTCF and cohesin are not enriched at anchors of C2 cluster loops comparing to static loops (**Figure 3.3b**). However, FOXA1 and FOXA2, as suggested by previous motif enrichment, are significantly enriched in C2 loops comparing to static loops (**Figure 3.3b**). This suggests that FOXA1 and FOXA2 may contribute to the formation of C2 loops in particular.

Chromatin loops form interaction hubs

Despite individual dynamic chromatin loops, we also often noticed several interconnected loops at key developmental genes such as FOXA2 (**Figure 3.1b**). We hypothesized that chromatin loops might form interaction hubs via individual chromatin loops. To identify chromatin interaction hubs, we built chromatin interaction networks using chromatin loops, where nodes are 10Kb bins that are loop anchors and edges are chromatin loops (**Figure 3.4a**). After building the network, we used Kleinberg's authority score to calculate hubness scores for each node (**Figure 3.4a**). In general, loop anchors with more interactions and stronger interactions are more likely to have a higher hubness score. Hubness scores are highly reproducible between biological replicates (**Figure S3.2a**). In sum, we identified hundreds of interaction hubs at each stage (**Figure S3.2b**).

K-means clustering of hubness scores for all loop anchors again revealed stage-specific patterns (**Figure 3.4b**), where most genomic regions are only interaction hubs at one or two stages, suggesting that hubs are relatively stage-specific. In addition, stage-specific chromatin hubs often harbor key developmental genes. For instance, the key regulator and marker for DE stage SOX17 are located in a DE-specific hub (**Figure 3.4b**). We then tested if interaction hubs are also associated with transcriptional regulation. Several lines of evidence suggest interaction hubs play key roles in driving stage-specific transcription profiles. First, interaction hubs are significantly enriched for enhancer and super-enhancers (**Figure 3.4c; Figure S3.2c**). Second, there are significantly more TF binding sites for lineage-determining TFs in interaction hubs comparing to regions that are not hubs (**Figure 3.4d; Figure S3.2d**). Third, genomic regions with transcription start site (TSS) are more likely to be chromatin hubs (**Figure 3.4e**) and more active promoters are located in interaction hubs (**Figure 3.4f**). Taken together, evidence suggests that

chromatin interaction hubs are not only centered for chromatin interactions but also centers for transcriptional regulation.

3.4 Discussion

Cellular differentiation is a complex and highly regulated process involving the rewiring of transcriptional regulatory networks and reorganization of chromatin architecture. Here we characterized the dynamic chromatin interactions during pancreatic differentiation using a human embryonic stem cell model system and profiled comprehensively the dynamic landscape of chromatin organization at high resolution. In particular, we identified thousands of dynamic chromatin loops that are associated with transcription changes. These loops could be used to link distal enhancer to target genes and help reveal mechanisms of stage-specific gene expression.

We also revealed that lineage-determining TFs could work as potential regulators of stage-specific chromatin loops. Previous reports found TFs can help establish stage- or cell-type-specific transcriptional programs by activating enhancers¹⁸. Our results provide a novel mechanism for so-called “pioneer” TFs to regulate gene expression repertoire. Moreover, while chromatin loops have been shown to be important in gene expression regulation, the formation of chromatin loops is still not well-understood. Our data also point to novel TFs that may help establish chromatin loops apart from CTCF and cohesin. While more studies are needed to reveal mechanisms of loop formation fully, there are likely to be two classes of chromatin loops as proposed previously^{16,20}. One class is static loops that are regulated by CTCF, and another class is cell-type-specific, which are likely to be regulated by lineage-determining TFs.

Finally, we showed that chromatin interaction hubs are formed by multiple chromatin loops and are hubs for transcriptional regulation as well. Those hubs harbor not only essential developmental genes but also TF binding sites and enhancers. We

reason that activated regions in the genome are linked together to form hubs. However, how hubs are formed and whether hubs are essential for transcriptional programs need more investigation.

In summary, our study provides a comprehensive map of chromatin organization during cell differentiation. With more understanding of chromatin organization and its role in gene regulation, we are in a great position to fully understand the mechanisms underlying transcriptional regulation.

3.5 Methods

Cell Culture

Cythera49 stem cells were cultured as previously described²¹. Briefly, approximately 5.5 million cells per well were seeded into 6 well plates, with full media changes occurring at day 0, 2, 5, and 8, and half media changes occurring on days 1, 3, 4, 6, 7, and 9. The DE timepoint was collected on day 2, the GT timepoint was collected on day 5, the FG timepoint was collected on day 7, and the PE timepoint was collected on day 10.

H1 stem cells were cultured as described previously²². 5.5 million cells per well were seeded onto into 6 well plates, with full media changes occurring on each day of culture. The DE timepoint was collected on day 3, the GT timepoint was collected on day 6, the FG timepoint was collected on day 8, and the PE timepoint was collected on day 11.

in situ Hi-C Experiments

The *in situ* Hi-C was performed according to a previously described protocol¹⁹ with slight modifications. Briefly, the human islets were washed with cold PBS and cut into small pieces. For cells, the cells were trypsinized and washed with PBS. The chromatin was cross-linked with 1% formaldehyde (Sigma) at ambient temperature for 10 min and quenched with 125mM glycine for 5 min. PBS washed tissue was homogenized with loose fitting douncer for 30 strokes before centrifugation to isolate the nuclei.

Nuclei were isolated and directly applied for digestion using 4 cutter restriction enzyme MboI (NEB) at 37 °C o/n. The single strand overhang was filled with biotinylated-14-ATP (Life Tech.) using Klenow DNA polymerase (NEB). Different from tradition Hi-C,

with *in situ* protocol, the ligation was performed when the nuclear membrane was still intact. DNA was ligated for 4h at 16 °C using T4 ligase (NEB). Protein was degraded by proteinase K (NEB) treatment at 55 °C for 30 min. The crosslinking was reversed with 500 mM of NaCl and heated at 68 °C o/n. DNA was purified and sonicated to 300-700 bp small fragments. Biotinylated DNA was selected with Dynabeads My One T1 Streptavidin beads (Life Tech.). The sequencing library was prepared on beads, and intensive wash was performed between different reactions. Libraries were checked with Agilent TapeStation and quantified using Qubit (Life Tech.). Libraries were sequenced with illumina HiSeq 4000 100 cycles of paired-end reads.

ChIP-seq Experiments

For ChIP-seq, the ChIP-IT High-Sensitivity kit (Active Motif) was used according to the manufacturer's instructions. Briefly, aggregates containing approximately 10^7 cells were fixed for 15 min in an 11.1% formaldehyde solution, chromatin was extracted by lysing cells in a Dounce homogenizer followed by shearing via sonication in a Bioruptor®Plus (Diagenode), on high for 3x 5 min (30 sec on, 30 sec off). For immunoprecipitation, 10-30 µg of the sheared chromatin was incubated with 4 µg primary antibody ON at 4°C on an end-to-end rotator, followed by incubation with Protein G agarose beads for 3 h at 4°C on the rotator. Reversal of crosslinks and DNA purification were performed according to the ChIP-IT High-Sensitivity instructions with incubation at 65°C for 2 h. DNA libraries were constructed using KAPA DNA Library Preparation Kits for Illumina® (Kapa Biosystems) and library sequencing was performed using a HiSeq 4000 System (Illumina®) with single-end reads of 50 bp in the Institute for Genomic Medicine (IGM) core research facility at the University of California at San Diego (UCSD).

Hi-C data processing

Hi-C data were processed as previously described with some modifications¹⁴. Read pairs were aligned to the hg19 reference genome separately using BWA-MEM²³ with default parameters. Specifically, chimeric reads were processed to keep only the 5' position and reads with low mapping quality (<10) were filtered out. Read pairs were then paired using custom scripts. Picard tools were then used to remove PCR duplicates. Bam files with alignments were further processed into text format as required by Juicebox tools²⁴. Juicebox tools were then applied to generate hic files containing normalized contact matrices. All downstream analysis was based on 10Kb resolution KR normalized matrices.

Chromatin loops were identified by comparing each pixel with its local background, as described previously¹⁹ with some modifications. Specifically, we only compared the donut region around the pixel to model the expected count. Briefly, the KR-normalized contact matrices at 10Kb resolution were used as input for loop calling. For each pixel, distance-corrected contact frequencies were calculated for each surrounding bin and the average of all surrounding bins. The expected counts were then transformed to raw counts by multiplying the counts with the raw-to-KR normalization factor. Then we calculated the probability of observing raw expected counts using Poisson distribution. All pixels with p-value < 0.01 and distance less than 10Kb were selected as candidate pixels. Candidate pixels were then filtered to remove pixels without any neighboring candidate pixels since they are likely false positives. Finally, pixels within 20Kb of each other were collapsed and only the most significant pixel was selected. The collapsed pixels with p-value $< 1e-5$ were used as the final list of chromatin loops.

ChIP-seq and ATAC-seq data processing

Reads were aligned using BWA MEM with either single-end or pair-end model to the hg19 reference genome. Reads with low mapping quality ($\text{mapq} < 10$) were filtered out, and PCR duplicates were removed using Picard tool (<http://broadinstitute.github.io/picard/>). MACS2²⁵ were then applied to call peaks and generate signal tracks to view in the genome browser.

RNA-seq data processing

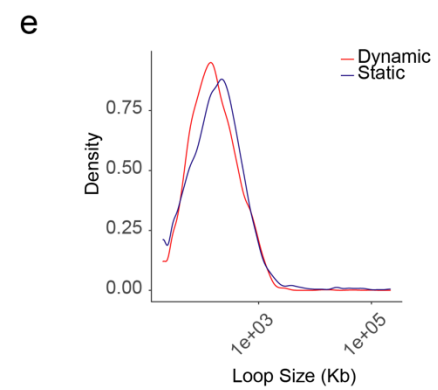
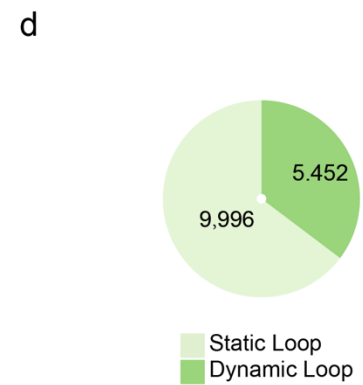
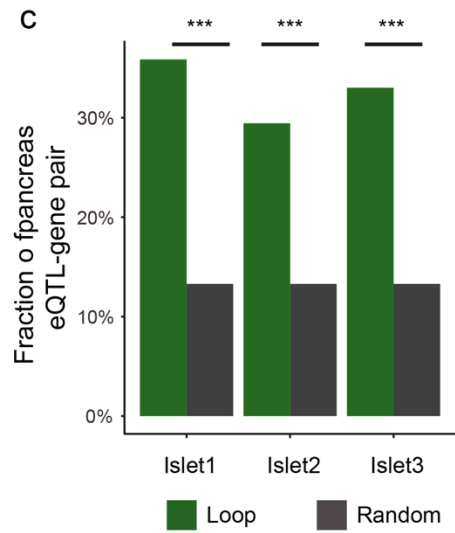
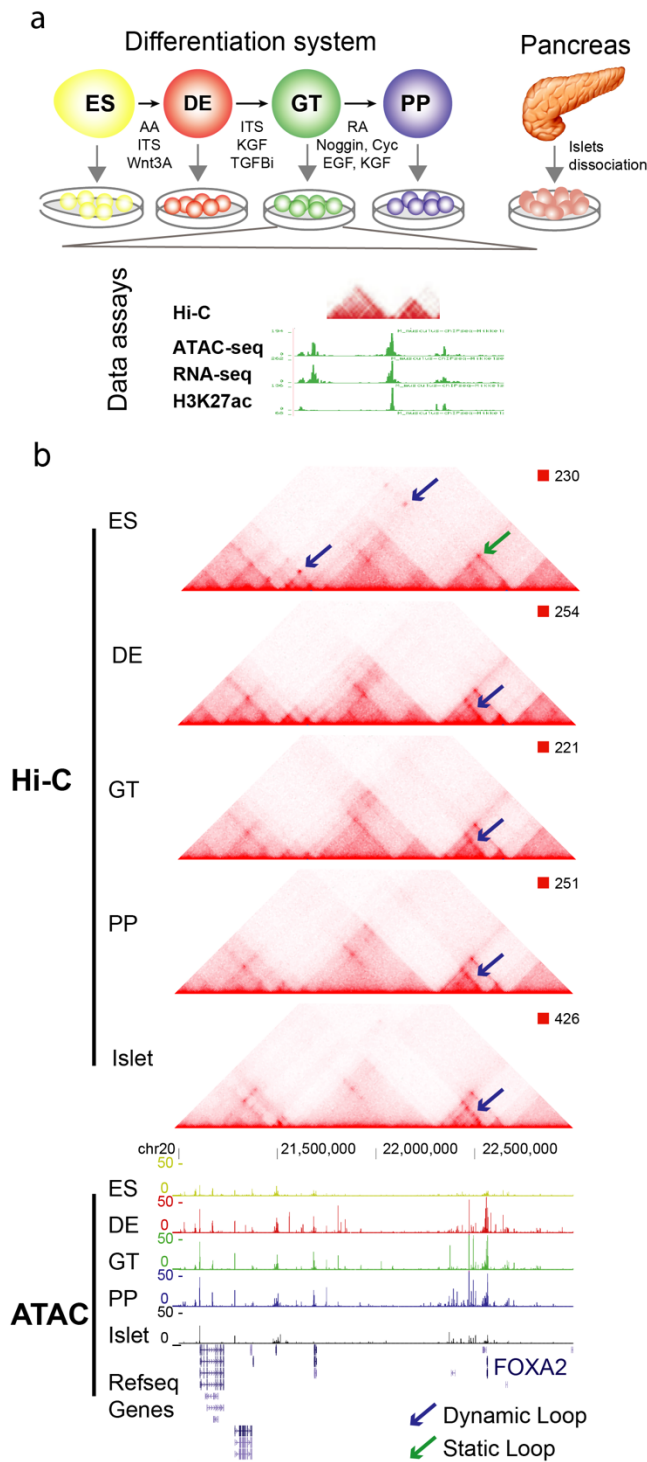
Reads were aligned to the hg19 reference genome using STAR 2.4.2a²⁶ with default parameters in the pair-end model. Only uniquely aligned reads were kept for further analysis. Cufflinks 2.2.1²⁷ was used to compute FPKM for each gene.

Identification of chromatin interaction hubs

A 3D interaction network was built for each chromosome by connecting 10kb regions based on their Hi-C contact values of chromatin loops, following normalization of Hi-C contact values for each stage in a 0-1 range. Nodes corresponding to 10kb regions were linked if their pairwise normalized contact score exceeded 0.1. For each stage, a measure of the centrality of each node was computed as the Kleinberg's authority score, hereafter referred to as hubness score, which takes into account both the degree of the node and the strength of edges connections. A random distribution of hubness scores was obtained for each stage by randomly shuffling network edges, with the number of permutations set to 1000. For each score, p-values were estimated as a cumulative probability from the corresponding null distribution. Hubs were selected with p-value < 0.05.

3.6 Figures

Figure 3.1. Characterization of three-dimension chromatin organization during pancreatic differentiation. (a) Schematic of the dataset generated. (b) Heatmap showing chromatin interactions and chromatin loops (top) and genome browser shots showing ATAC-seq signal for the same region (bottom). Dynamic and static chromatin loops are marked in blue and green arrows, respectively. (c) Pancreas eQTL and egene pairs are enriched in islet chromatin loops. *** $p < 0.001$ (d) Pie chart for dynamic and static chromatin loops. (e) Density plot shows sizes of dynamic and static chromatin loops.



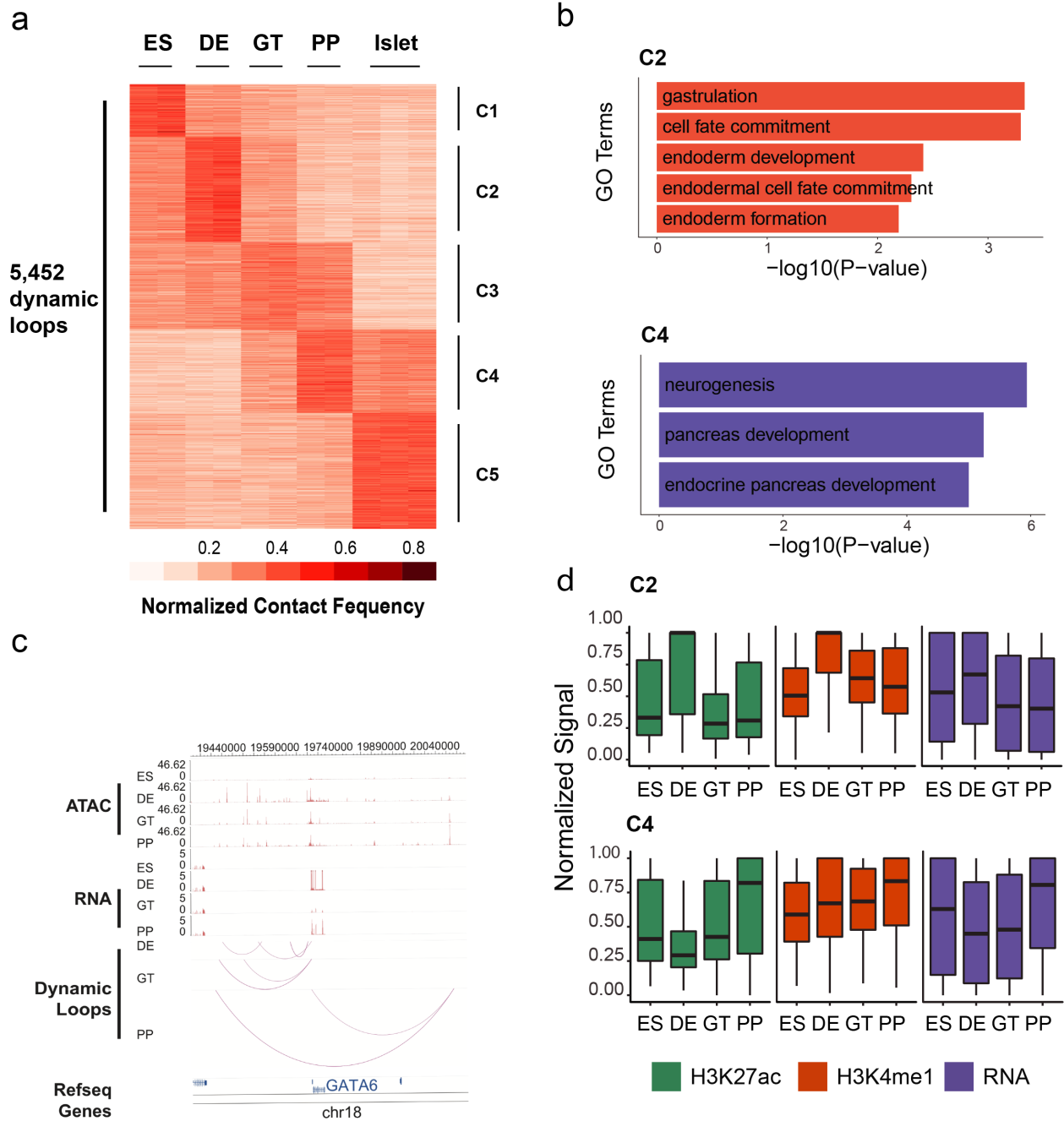









Figure 3.2. Dynamic chromatin loops are associated with stage-specific transcription regulation. (a) Heatmap showing k-means clustering of dynamic loops. (b) Selected enriched GO terms for C2 and C4 cluster loops. (c) Genome browser shot for GATA6 region, highlighting distinct enhancers for GATA6 at different stages. (d) Boxplot of histone modification and gene expression levels stratified by loop clusters.

a

C2		
TF name	Motif	P-value
GATA		1E-3706
SOX2		1E-923
FOXA		1E-453
C4		
TF name	Motif	P-value
FOXA		1E-1322
GATA		1E-1112
HNF6		1E-1007
PDX1		1E-554

b

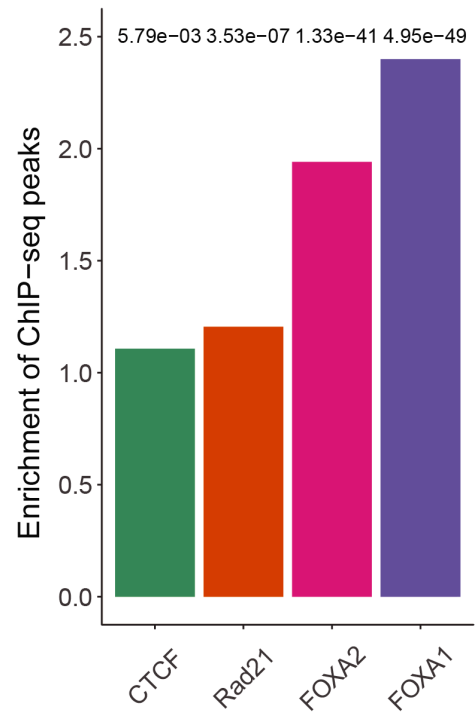
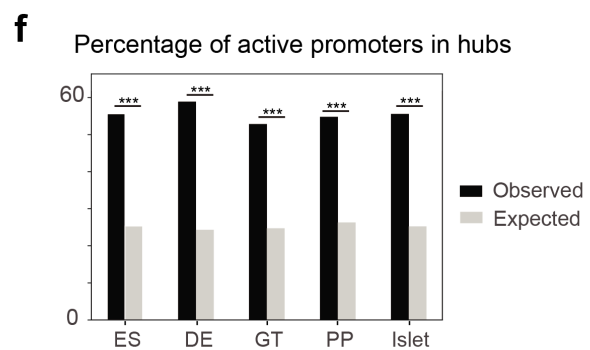
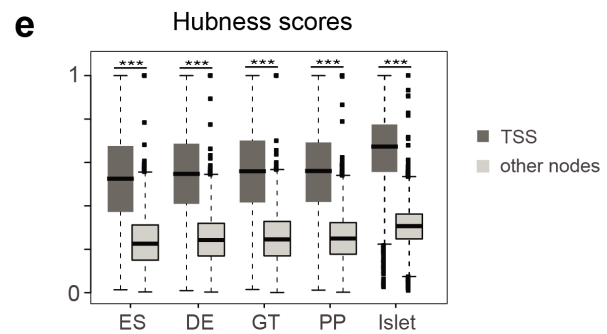
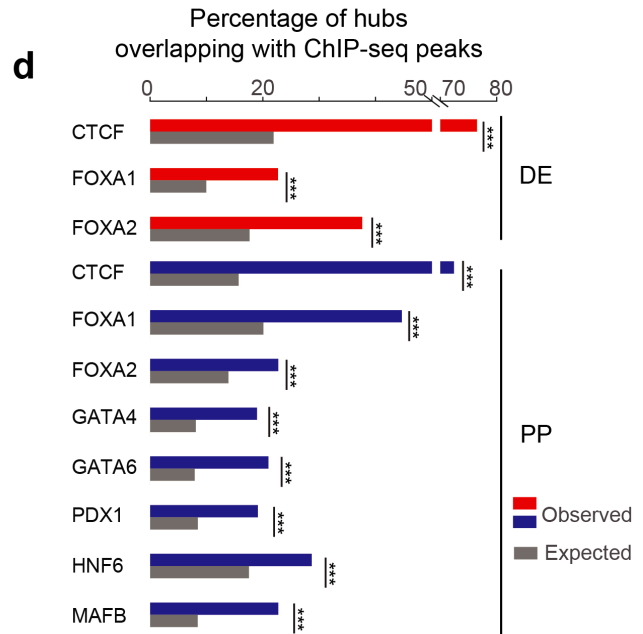
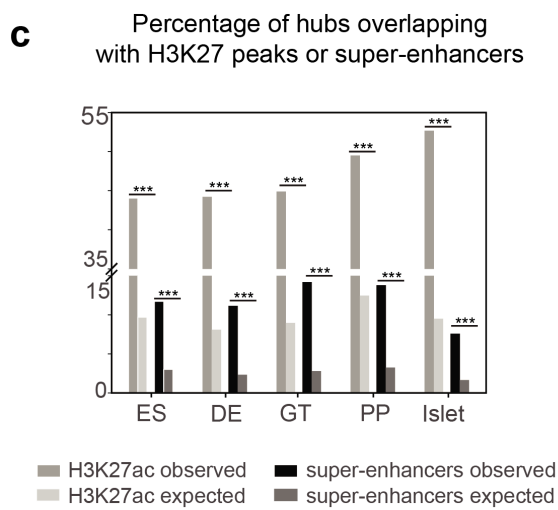
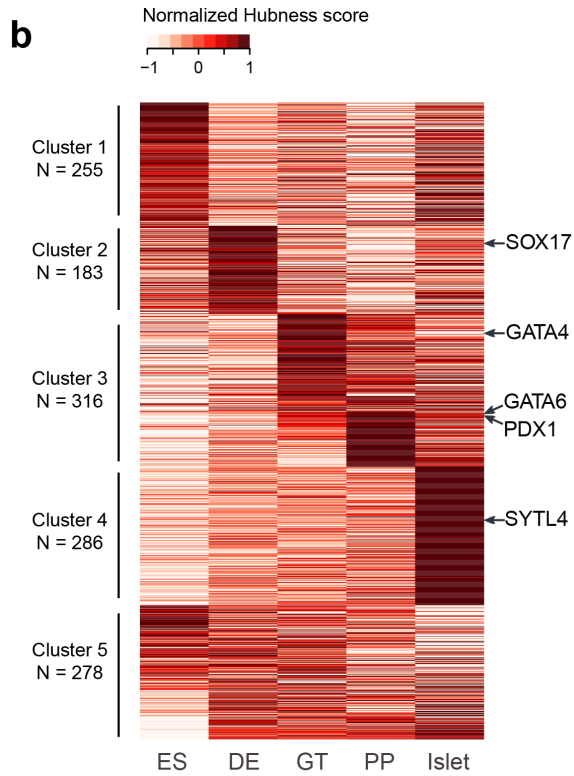
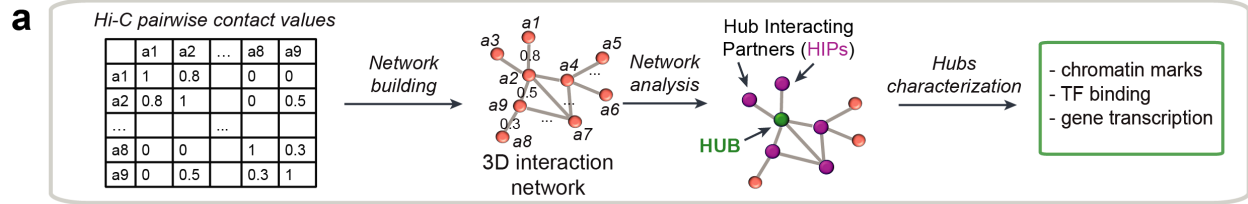


Figure 3.3. Dynamic chromatin loops are associated with lineage-determining TFs. (a) Enriched motifs for C2 and C4 cluster loops. (b) Barplot showing enrichment of TF binding peaks comparing C2 cluster loop with static loops.

Figure 3.4. Identification and characterization of hubs in 3D interaction networks.

(a) Workflow to generate 3D interaction networks and characterize the most connected nodes (hubs). (b) Heatmap showing normalized hubness scores for each stage. (c) Percentage of hubs overlapping with H3K27ac peaks or with super-enhancers, compared to expected overlaps. (d) Percentage of hubs overlapping with TF binding sites in DE and PP, compared to expected overlaps. (e) Distribution of hubness scores for hubs regions that include a TSS, compared to other nodes. (f) Percentage of active promoters in hubs of each stage.



3.7 Supplemental Figures

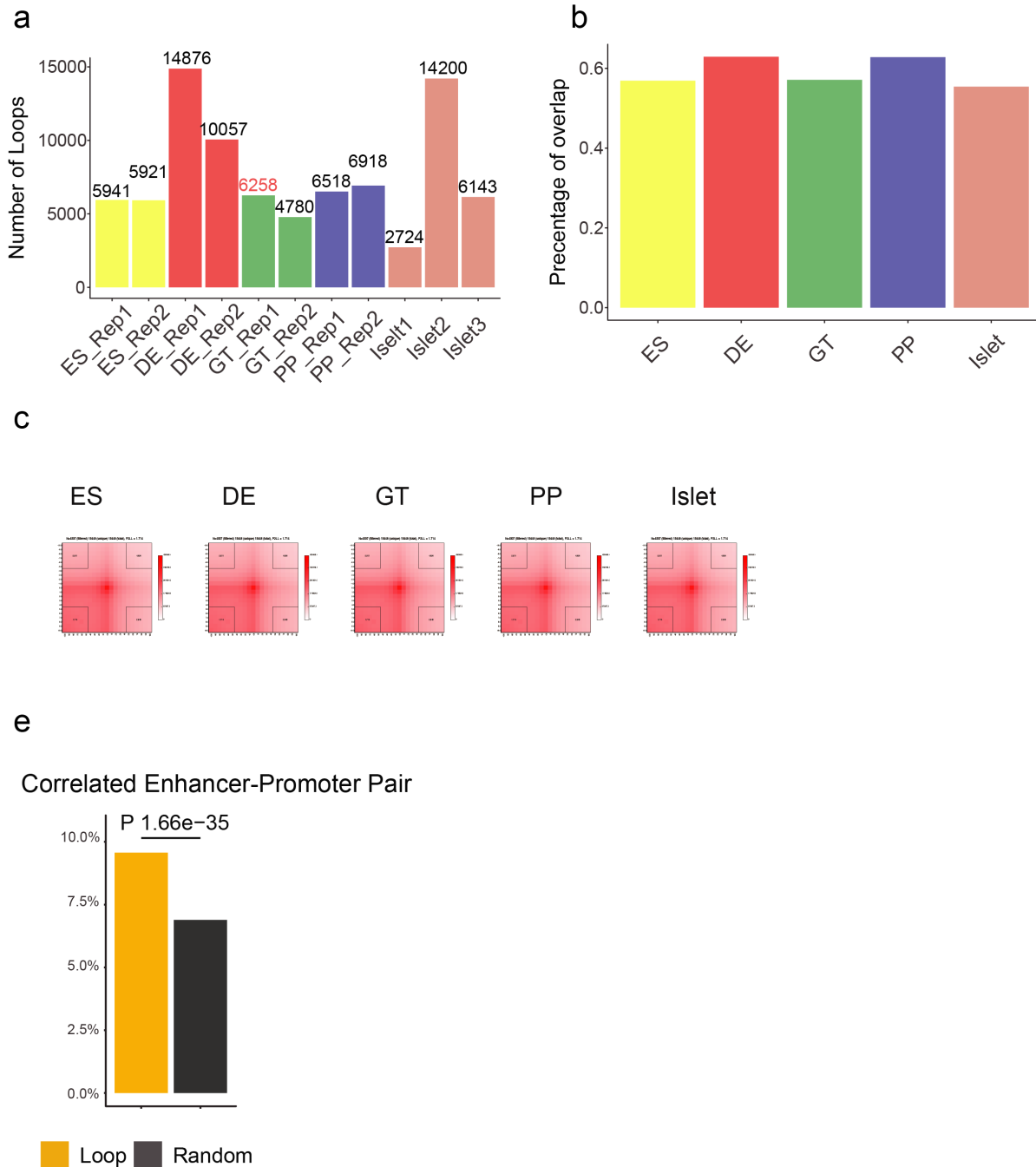
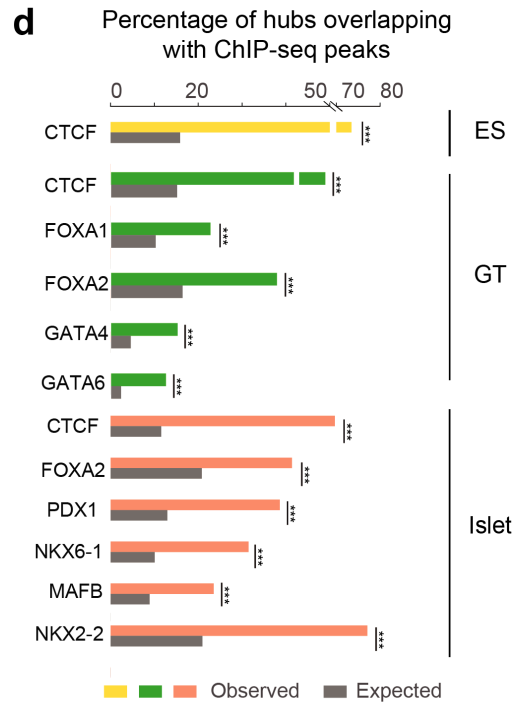
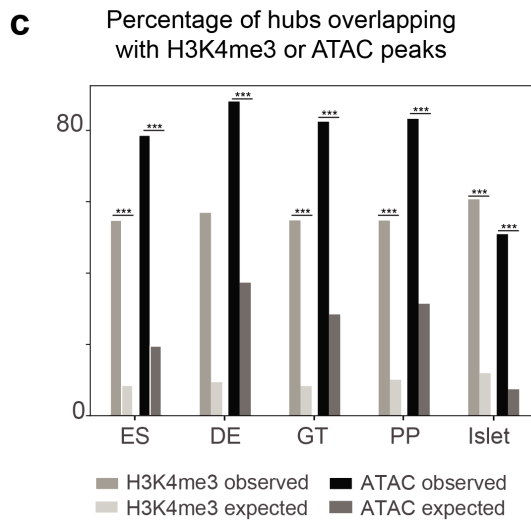
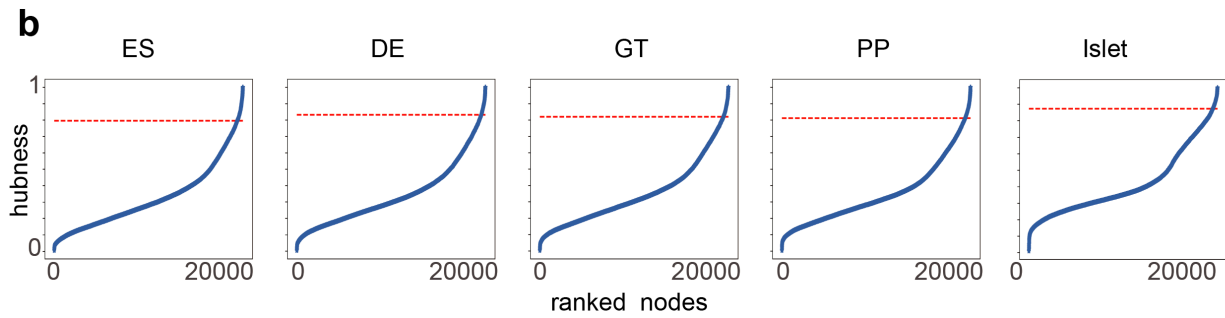
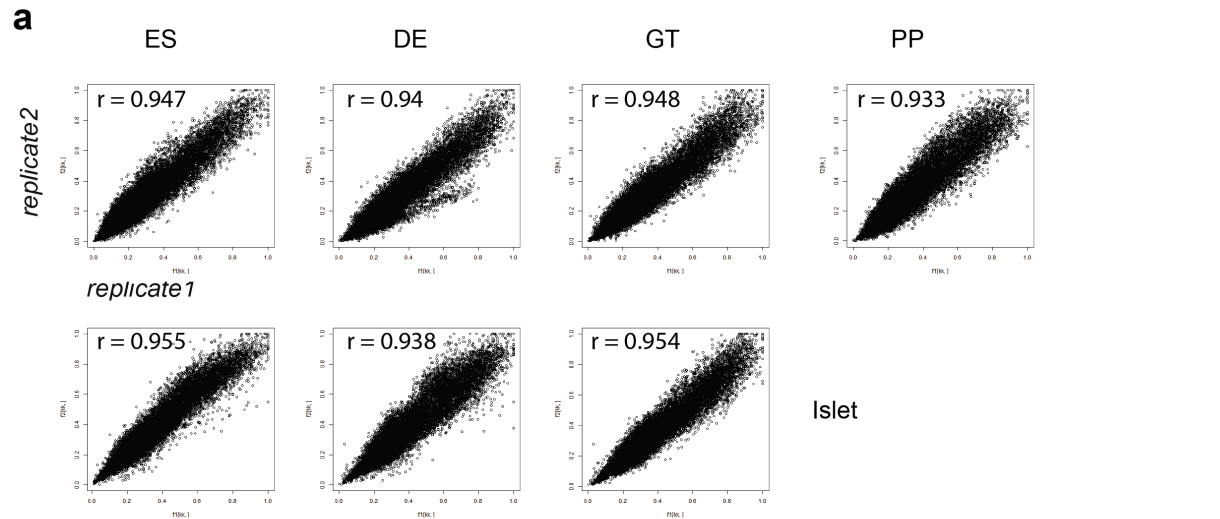


Figure S3.1. Characterization of three-dimension chromatin organization during pancreatic differentiation. (a) The number of chromatin loops identified in each biological replicate. (b) Percentage of overlapped chromatin loops between biological replicates for each stage. (c) Heatmap showing aggregate peak analysis of chromatin loops for each stage. (d) Barplot showing fraction of enhancer-promoter pairs that are correlated in terms of activity.

Figure S3.2. Quality control and characterization of hubs. (a) Correlation of hubness scores between replicates of Hi-C data. (b) Hubness scores for ranked network nodes, highlighting selected hubs corresponding to highest-scoring nodes. (c) Percentage of hubs overlapping with H3K4me3 peaks or with ATAC-seq peaks, compared to expected overlaps. (d) Percentage of hubs overlapping with TF binding sites in ES, GT or Islet, compared to expected overlaps.



3.8 Acknowledgments

Chapter 3, in full, is a manuscript in preparation as “Dynamic 3D chromatin organization during differentiation of human embryonic stem cells to pancreatic progenitor cells”. Yunjiang Qiu, Francesca Mulas, Ryan J Geusz, Jian Yan, Nick Vinckier, Allen Wang, Maike Sander, and Bing Ren. The dissertation author is the primary investigator and author of this paper.

3.9 References

1. Bolzer, A., Kreth, G., Solovei, I., Koehler, D., Saracoglu, K., Fauth, C., Müller, S., Eils, R., Cremer, C., Speicher, M. R. & Cremer, T. Three-dimensional maps of all chromosomes in human male fibroblast nuclei and prometaphase rosettes. *PLoS Biol.* 3, e157 (2005).
2. Gorkin, D. U., Leung, D. & Ren, B. The 3D Genome in Transcriptional Regulation and Pluripotency. *Cell Stem Cell* 14, 762–775 (2014).
3. Jin, F., Li, Y., Dixon, J. R., Selvaraj, S., Ye, Z., Lee, A. Y., Yen, C.-A., Schmitt, A. D., Espinoza, C. A. & Ren, B. A high-resolution map of the three-dimensional chromatin interactome in human cells. *Nature* 503, 290–294 (2013).
4. 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. & Dekker, J. Chromosome Conformation Capture Carbon Copy (5C): a massively parallel solution for mapping interactions between genomic elements. *Genome Res.* 16, 1299–1309 (2006).
5. Zhang, Y., Wong, C.-H., Birnbaum, R. Y., Li, G., Favaro, R., Ngan, C. Y., Lim, J., Tai, E., Poh, H. M., Wong, E., Mulawadi, F. H., Sung, W.-K., Nicolis, S., Ahituv, N., Ruan, Y. & Wei, C.-L. Chromatin connectivity maps reveal dynamic promoter–enhancer long-range associations. *Nature* 504, 306–310 (2013).
6. 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. & Dekker, J. Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome. *Science* 326, 289–293 (2009).
7. Dixon, J. R., Selvaraj, S., Yue, F., Kim, A., Li, Y., Shen, Y., Hu, M., Liu, J. S. & Ren, B. Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* 485, 376–380 (2012).
8. 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. & Heard, E. Spatial partitioning of the regulatory landscape of the X-inactivation centre. *Nature* 485, 381–385 (2012).
9. Nuebler, J., Fudenberg, G., Imakaev, M., Abdennur, N. & Mirny, L. A. Chromatin organization by an interplay of loop extrusion and compartmental segregation. *Proc Natl Acad Sci USA* 44, 201717730–10 (2018).
10. Phanstiel, D. H., Van Bortle, K., Spacek, D., Hess, G. T., Shamim, M. S., Machol, I., Love, M. I., Aiden, E. L., Bassik, M. C. & Snyder, M. P. Static and Dynamic DNA

Loops form AP-1-Bound Activation Hubs during Macrophage Development. *Molecular Cell* 1–26 (2017). doi:10.1016/j.molcel.2017.08.006

11. Beagan, J. A., Duong, M. T., Titus, K. R., Zhou, L., Cao, Z., Ma, J., Lachanski, C. V., Gillis, D. R. & Phillips-Cremins, J. E. YY1 and CTCF orchestrate a 3D chromatin looping switch during early neural lineage commitment. *Genome Res.* 27, gr.215160.116–1152 (2017).
12. Rubin, A. J., Barajas, B. C., Furlan-Magaril, M., Lopez-Pajares, V., Mumbach, M. R., Howard, I., Kim, D. S., Boxer, L. D., Cairns, J., Spivakov, M., Wingett, S. W., Shi, M., Zhao, Z., Greenleaf, W. J., Kundaje, A., Snyder, M., Chang, H. Y., Fraser, P. & Khavari, P. A. Lineage-specific dynamic and pre-established enhancer–promoter contacts cooperate in terminal differentiation. *Nat. Genet.* 512, 96–12 (2017).
13. Krijger, P. H. L. & de Laat, W. Regulation of disease-associated gene expression in the 3D genome. *Nat. Rev. Mol. Cell Biol.* 17, 771–782 (2016).
14. Dixon, J. R., Jung, I., Selvaraj, S., Shen, Y., Antosiewicz-Bourget, J. E., Lee, A. Y., Ye, Z., Kim, A., Rajagopal, N., Xie, W., Diao, Y., Liang, J., Zhao, H., Lobanenkov, V. V., Ecker, J. R., Thomson, J. A. & Ren, B. Chromatin architecture reorganization during stem cell differentiation. *Nature* 518, 331–336 (2015).
15. Schmitt, A. D., Hu, M., Jung, I., Xu, Z., Qiu, Y., Tan, C. L., Li, Y., Lin, S., Lin, Y., Barr, C. L. & Ren, B. A Compendium of Chromatin Contact Maps Reveals Spatially Active Regions in the Human Genome. *CellReports* 17, 2042–2059 (2016).
16. Phillips-Cremins, J. E., Sauria, M. E. G., Sanyal, A., Gerasimova, T. I., Lajoie, B. R., Bell, J. S. K., Ong, C.-T., Hookway, T. A., Guo, C., Sun, Y., Bland, M. J., Wagstaff, W., Dalton, S., McDevitt, T. C., Sen, R., Dekker, J., Taylor, J. & Corces, V. G. Architectural protein subclasses shape 3D organization of genomes during lineage commitment. *Cell* 153, 1281–1295 (2013).
17. Xie, R., Everett, L. J., Lim, H.-W., Patel, N. A., Schug, J., Kroon, E., Kelly, O. G., Wang, A., D'Amour, K. A., Robins, A. J., Won, K.-J., Kaestner, K. H. & Sander, M. Dynamic chromatin remodeling mediated by polycomb proteins orchestrates pancreatic differentiation of human embryonic stem cells. *Cell Stem Cell* 12, 224–237 (2013).
18. Wang, A., Yue, F., Li, Y., Xie, R., Harper, T., Patel, N. A., Muth, K., Palmer, J., Qiu, Y., Wang, J., Lam, D. K., Raum, J. C., Stoffers, D. A., Ren, B. & Sander, M. Epigenetic priming of enhancers predicts developmental competence of hESC-derived endodermal lineage intermediates. 16, 386–399 (2015).
19. Rao, S. S. P., Huntley, M. H., Durand, N. C., Stamenova, E. K., Bochkov, I. D., Robinson, J. T., Sanborn, A. L., Machol, I., Omer, A. D., Lander, E. S. & Aiden, E. L. A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping. *Cell* 159, 1665–1680 (2014).

20. Kagey, M. H., Newman, J. J., Bilodeau, S., Zhan, Y., Orlando, D. A., van Berkum, N. L., Ebmeier, C. C., Goossens, J., Rahl, P. B., Levine, S. S., Taatjes, D. J., Dekker, J. & Young, R. A. Mediator and cohesin connect gene expression and chromatin architecture. *Nature* 467, 430–435 (2010).
21. Kroon, E., Martinson, L. A., Kadoya, K., Bang, A. G., Kelly, O. G., Eliazar, S., Young, H., Richardson, M., Smart, N. G., Cunningham, J., Agulnick, A. D., D'Amour, K. A., Carpenter, M. K. & Baetge, E. E. Pancreatic endoderm derived from human embryonic stem cells generates glucose-responsive insulin-secreting cells in vivo. *Nat Biotechnol* 26, 443–452 (2008).
22. Rezanian, A., Bruin, J. E., Arora, P., Rubin, A., Batushansky, I., Asadi, A., O'Dwyer, S., Quiskamp, N., Mojibian, M., Albrecht, T., Yang, Y. H. C., Johnson, J. D. & Kieffer, T. J. Reversal of diabetes with insulin-producing cells derived *in vitro* from human pluripotent stem cells. *Nature Publishing Group* 32, 1121–1133 (2014).
23. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv q-bio.GN*, (2013).
24. Durand, N. C., Robinson, J. T., Shamim, M. S., Machol, I., Mesirov, J. P., Lander, E. S. & Aiden, E. L. Juicebox Provides a Visualization System for Hi-C Contact Maps with Unlimited Zoom. *Cell Systems* 3, 99–101 (2016).
25. Zhang, Y., Liu, T., Meyer, C. A., Eeckhoute, J., Johnson, D. S., Bernstein, B. E., Nusbaum, C., Myers, R. M., Brown, M., Li, W. & Liu, X. S. Model-based Analysis of ChIP-Seq (MACS). *Genome Biol* 9, 1–9 (2008).
26. Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M. & Gingeras, T. R. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* 29, 15–21 (2013).
27. Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D. R., Pimentel, H., Salzberg, S. L., Rinn, J. L. & Pachter, L. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat Protoc* 7, 562–578 (2012).