Principled Methods and Models for Deep Learning Based Functional Genomics

by

Konstantin Krismer

Submitted to the Department of Biological Engineering
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in Biological Engineering

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

September 2021

Author . . . .
Department of Biological Engineering
July 26, 2021

Certified by . . . .
David K. Gifford
Professor of Electrical Engineering and Computer Science
Professor of Biological Engineering
Thesis Supervisor

Accepted by . . . .
Katharina Ribbeck
Professor of Biological Engineering
Graduate Program Chair, Department of Biological Engineering

Principled Methods and Models for Deep Learning Based Functional Genomics

by

Konstantin Krismer

Submitted to the Department of Biological Engineering
on July 26, 2021, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy in Biological Engineering

Abstract

Many advances in functional genomics and in biology more broadly can be attributed to the rise of massively parallel sequencing technology and its derivatives. As the volume of sequencing and other high-throughput experimental data increases exponentially, so does the need for computational methods to analyze and condense these vast amounts of data, and to help explain the underlying phenomena. In this thesis, I describe five projects that introduce novel techniques and methods in functional genomics.

The first project introduces a simulation-based framework to investigate neural network architectures that are trained on biological sequence data, as is common in functional genomics. The second project describes a two-pronged approach to study the determinants of cell type-specific chromatin accessibility, with an ensemble of neural networks trained on DNase-seq data to predict chromatin accessibility, and MIAA, the multiplexed integrated accessibility assay, to validate, experimentally, these *in silico* predictions. The third project presents a method to identify long-range genomic interactions from ChIA-PET and HiChIP data. Enabled by this work, the fourth project aims to provide a means to identify reproducible long-range genomic interactions. We continue the analysis of long-range interactions in the fifth project by performing co-enrichment analysis of transcription factor sequence motifs.

Collectively, these methods provide new approaches to a range of problems in functional genomics, from finding appropriate neural network architectures for sequence-based prediction tasks to uncovering patterns in long-range genomic interactions.

Thesis Supervisor: David K. Gifford
Title: Professor of Electrical Engineering and Computer Science
Professor of Biological Engineering

Acknowledgments

The experiences I made during my years at MIT were almost entirely positive, which is a sentiment not commonly expressed about graduate school. I am grateful to the many people who contributed, as mentors, lab mates, collaborators, and friends, to my intellectual and personal growth at MIT.

First, I thank my advisor, David Gifford, for gently guiding me through my doctorate. Thank you for meeting with me over 170 times during my PhD, and for starting most of these one-on-one meetings by asking "How can I help you today?". Thank you for making me feel like a valued member of your lab, for being a constant source of ideas and insights, for leaving space for scientific and career exploration, and for creating such a welcoming work environment.

I am also grateful to my thesis committee, Douglas Lauffenburger and Phillip Sharp. I enjoyed every minute of the meetings we had together and I always left the meeting room feeling excited for research and energized for the steps ahead. Thank you for your insights and feedback, both scientific and worldly. I heeded Phil's advice to visit Baxter State Park, and I clambered up Mount Katahdin—an uplifting memory that I will cherish. I'd like to add that, despite being luminaries in their fields, meetings with Doug and Phil were surprisingly easy to schedule, which I recognize is an indication of their commitment as thesis committee members.

I am deeply grateful for the group of students and post-docs who called the Gifford lab their academic home during the time I was there. We should pat each other on the back for maintaining a virtually conflict-free work environment all those years. Has Dave's calm, courteous demeanor influenced all of us? I was lucky to join the lab at around the same time as Jennifer Hammelman, with whom I had the pleasure to share room 32-G538 for the pre-pandemic years, and to collaborate on numerous projects, including three of the five projects presented in this thesis. I was equally lucky to have overlapped with Yuchun Guo, who was a senior post-doc at the time I joined the lab. Yuchun took me under his wing when I joined the lab, "all fresh-faced and new and not knowing anything", to borrow a line from Robert Lopez's Avenue Q.

At this point I introduce Sachit Saksena and Siddhartha Jain as protagonists in the Gifford lab chapter of my life. Both were great conversationalists and trusted companions on trips to the Muddy Charles Pub and beyond. Thanks also to Brandon Carter, Zheng Dai, Alexander Dimitrakakis, Logan Engstrom, Benjamin Holmes, Nathan Hunt, Jon Krog, Bianca Lepe, Ge (Saber) Liu, Wilson Louie, Jonas Mueller, Hyunjin Park, Max Shen, Tahin Syed, Ziheng (Tony) Wang, Hui Ting Grace Yeo, and Haoyang Zeng for countless inspiring conversations. I would also like to thank Linda Lynch for her unbureaucratic and swift help with conference registrations, flights, and countless other tasks.

I must thank our collaborators across the river at Harvard Medical School. The Sherwood lab provided the invaluable perspective and expertise of experimental biology, perfectly complementing the methods of the purely computational Gifford lab. Working with Richard Sherwood and Budhaditya Banerjee was a joy.

I am also grateful to the scientific mentors who set me on this path. Gerald Lirk's exciting biology lectures first lured me into bioinformatics and computational biology back in 2010. Karin Pröll and Stephan Dreiseitl encouraged me to come to the United States as an undergraduate visiting student. Jonathan Rameseder introduced me to life at MIT, and Michael Yaffe made sure I kept coming back.

My time at MIT would have been much less inspiring without my fellow students and friends in the BE-2016 cohort. I am proud to be part of such an exceptional group of people.

I am grateful for the friendships I made during my time in Boston, both on and off campus. Special thanks to my flatmates, Daniel Anderson and Joao Cavalcanti, who were reliable companions throughout my time at MIT; to the community around Paul Niksch, Özlem Niksch, and Jason Wheeler for being the right amount of crazy; to Nicola Martino and friends for generously hosting so many dinner parties; to Bert van de Kooij for Friday night beers; to Daniel Winklehner and Jette Lengefeld for countless Friday lunches; to Jürgen Cito and all my fellow Austrians in Boston; and to Filippos Sotiropoulos, my pan-European friend. My life is richer for knowing all of you.

Thank you, Robert Rotstein, for being the most generous landlord on this side of the Atlantic.

Dear Anne, thank you for putting up with me. As you know very well, I don't wear my heart on my sleeve, which must be frustrating at times. I am so grateful I got to experience it all with you. Also, thanks for being okay with eating Indian food so often.

I also thank Mama und Papa for being great parents. I am immensely grateful for the worry-free childhood you provided for me and for supporting all of my adventures, academic and otherwise. A big shout-out also to the other half of the home team, Antonia Krismer and Elias Rabl. You were fantastic hosts and intimidatingly professional cooks who put my pathetic student meals to shame.

Lastly, I would like to thank my brain, which seems to be wired in such a way that allows me to surf through life on a wave of endorphins. – *No problem, man!*

This doctoral thesis has been examined by a Committee of the Department of Biological Engineering as follows:

Douglas A. Lauffenburger . . . .
Thesis Committee Chair
Professor of Biological Engineering
Professor of Chemical Engineering
Professor of Biology
Massachusetts Institute of Technology

David K. Gifford . . . .
Thesis Supervisor
Professor of Electrical Engineering and Computer Science
Professor of Biological Engineering
Massachusetts Institute of Technology

Phillip A. Sharp . . . .
Thesis Committee Member
Institute Professor of Biology
Massachusetts Institute of Technology

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Chromatin accessibility

Chromatin structure is critical for the regulation of DNA-dependent processes such as transcription [34], replication [21], recombination [22], and DNA damage repair [20, 24], and thus is an upstream regulator of most biological functions, including gene expression patterns that shape cell identity. Furthermore, the dysregulation of chromatin accessibility is an underappreciated factor in cancer initiation and progression [2, 29], and a better understanding of the cell type and cell state specific rules of chromatin accessibility is needed in order to develop more accurate models of cell differentiation and identity, and diseases associated with its dysregulation.

We define chromatin accessibility as a measure of the relative depletion of local nucleosome contact with genomic DNA [35]. Nucleosomes are the basic units of chromatin architecture, which consist of DNA wrapped around eight histone proteins. Regions of *open* or *accessible* chromatin are nucleosome-depleted, and transcriptionally active [8], as well as commonly associated with active enhancers [32], while regions of *closed* chromatin are nucleosome-enriched and inaccessible to most transcription factors [12].

The chromatin state is modulated by post-translational modifications of histone proteins, which alter the charge of the histones and thereby strengthens or weakens interactions between histones and the negatively charged DNA. Collectively, these histone modifications constitute the so-called histone code. Histone modifications are handed down to mitotic daughter cells and are sometimes even maintained through meiosis [13].

Common post-translational modifications of histones include acetylation and deacetylation of lysine residues by histone acetyltransferases (HATs) and histone deacetylases (HDACs), respectively. HATs remove the positive charge on the histones through the enzymatic addition of acetyl groups. As a consequence, the histone packing decreases, which makes the chromatin more accessible.

Chromatin accessibility is also influenced by methylation. Histone methyltransferases (HMTs) add one or more methyl groups to lysine and arginine residues of histones, which increases their hydrophobicity. This modification has the opposite effect; methylated histones are more tightly packed than their unmethylated counterparts, effectively decreasing chromatin accessibility. The methyl groups are removed by histone demethylases (HDMs).

While the roles of other forms of histone modifications, such as phosphorylation, ubiquitylation, and sumoylation, are less well understood, their effect on chromatin accessibility has been shown in numerous studies [1, 28, 25, 27].

The histone modifiers themselves are not site-specific, but are directed to their site of action by sequence-specific transcription factors. This mechanism has been shown for HATs [14], HDACs [17], HMTs [31], and HDMs [26].

We use the term *grammar*, or cell type-specific grammar of chromatin accessibility, to refer to a set of probabilistic and spatial rules of sequence motifs that explains the differences between chromatin accessibility profiles of various cell types. These rules describe patterns of sequence motifs that are associated with open chromatin regions in a specific cell type or cell state. They may include combinations of sequence motifs, spacing or orientation constraints between sequence motifs. An example of such a rule would be the following: sequence motif A is found between 10 and 50 bp upstream of sequence motif B in open chromatin regions in cell state Y, where the sequence motifs might correspond to transcription factor binding sites. Spatial relationships between transcription factor binding sites are biologically relevant, as there exist both homotypic [7] as well as heterotypic clusters of transcription factors [15] and the combinatorial interactions between them have been shown to be important [30].

Chapter 2 introduces a flexible simulator that makes it possible to specify a set of probabilistic rules describing the hypothesized rules of chromatin accessibility and other biological phenomena, and subsequently synthesize sequence data that adheres to these rules. In chapter 3 we present a neural network-based approach to identify determinants of cell type-specific chromatin accessibility and how to validate them experimentally.

1.2 Long-range genomic interactions

Another layer in the many-layered regulation of gene expression are physical, three-dimensional chromatin interactions [3, 36]. These interactions can occur between genomic regions that are millions of base pairs apart from each other. Together, they form a functionally meaningful, higher-order organization of the genome [4].

Sequencing-based assays such as ChIA-PET [6], HiChIP [23], and Hi-C [19] have been used to discover numerous examples of long-range interactions with functional consequences, ranging from interactions mediated by structure-defining architectural proteins [33, 10, 5, 11] to enhancer-promoter interactions [16, 18, 37].

Chapters 4 to 6 introduce computational methods to detect long-range genomic interactions, to assess their reproducibility, and to uncover pairs of transcription factors that interact with each other throughout the genome.

1.3 Thesis outline

The following five chapters describe five different projects that were previously published (chapters 3 - 5) or are currently under review (chapters 2 and 6).

In chapter 2 we introduce *seqgra*, a simulation-based framework to investigate neural network architectures that are trained on biological sequence data, as is common in functional genomics. Chapter 3 describes a two-pronged approach to study the determinants of cell type-specific chromatin accessibility, with an ensemble of neural networks trained on DNase-seq data to predict chromatin accessibility, and MIAA, the multiplexed integrated accessibility assay, to experimentally validate these *in silico* predictions. In chapter 4 we present *CID*, a method to identify long-range genomic interactions from ChIA-PET and HiChIP data. Chapter 5 picks up where the previous chapter left off by introducing *IDR2D*, a method which represents a means to identify reproducible long-range genomic interactions. And lastly, we present *spatzie* in chapter 6, which continues the analysis of long-range interactions by performing co-enrichment analysis of transcription factor sequence motifs.

1.4 Collaborators

The work presented in this thesis would not have been possible without the many collaborators that I was fortunate enough to work with. The experimental arm of the MIAA project from chapter 3 was carried out by Budhaditya Banerjee and Richard I. Sherwood of the Sherwood Lab at Harvard Medical School. The method introduced in chapter 4 benefited greatly from the conceptual input of Michael Closser and Hynek Wichterle from the Wichterle Lab at Columbia University. Furthermore, all projects presented here are a result of close collaboration and countless conversations with members of the Gifford Lab, specifically Jennifer Hammelman and Yuchun Guo.

1.5 Availability

The software that is described in this thesis is freely available and licensed under permissive open source licenses. The method of chapter 2 was packaged as a pip-installable Python package and is part of the Python Package Index. Documentation can be found at `https://kkrismer.github.io/seqgra` and the source code is hosted on GitHub and available at `https://github.com/gifford-lab/seqgra`. The source code for the model from chapter 3 is also available on GitHub at `https://github.com/gifford-lab/DeepAccess`. CID, the method from chapter 4 is part of the larger

GEM [9] Java package and can be downloaded from http://groups.csail.mit.edu/cgs/gem/cid. The source code is available at https://github.com/gifford-lab/GEM3. IDR2D from chapter 5 and spatzie from chapter 6 are both available as an R/Bioconductor packages and part of their functionality is also offered online at https://idr2d.mit.edu and https://spatzie.mit.edu, respectively. The source code repositories for both methods are also available on GitHub at https://github.com/gifford-lab/idr2d and https://github.com/gifford-lab/spatzie.

Bibliography

[1] P. Cheung, C. D. Allis, and P. Sassone-Corsi. Signaling to chromatin through histone modifications. *Cell*, 103(2):263–271, Oct 2000.

[2] M. R. Corces, J. M. Granja, S. Shams, B. H. Louie, J. A. Seoane, W. Zhou, T. C. Silva, C. Groeneveld, C. K. Wong, S. W. Cho, A. T. Satpathy, M. R. Mumbach, K. A. Hoadley, A. G. Robertson, N. C. Sheffield, I. Felau, M. A. A. Castro, B. P. Berman, L. M. Staudt, J. C. Zenklusen, P. W. Laird, C. Curtis, W. J. Greenleaf, H. Y. Chang, R. Akbani, C. C. Benz, E. A. Boyle, B. M. Broom, A. D. Cherniack, B. Craft, J. A. Demchok, A. S. Doane, O. Elemento, M. L. Ferguson, M. J. Goldman, D. N. Hayes, J. He, T. Hinoue, M. Imielinski, S. J. M. Jones, A. Kemal, T. A. Knijnenburg, A. Korkut, D. C. Lin, Y. Liu, M. K. A. Mensah, G. B. Mills, V. P. Reuter, A. Schultz, H. Shen, J. P. Smith, R. Tarnuzzer, S. Trefflich, Z. Wang, J. N. Weinstein, L. C. Westlake, J. Xu, L. Yang, C. Yau, Y. Zhao, and J. Zhu. The chromatin accessibility landscape of primary human cancers. *Science*, 362(6413), 10 2018.

[3] J. Dekker. Gene regulation in the third dimension. *Science*, 319(5871):1793–1794, Mar 2008.

[4] J. Dekker and T. Misteli. Long-Range Chromatin Interactions. *Cold Spring Harb Perspect Biol*, 7(10):a019356, Oct 2015.

[5] J. M. Dowen, Z. P. Fan, D. Hnisz, G. Ren, B. J. Abraham, L. N. Zhang, A. S. Weintraub, J. Schujiers, T. I. Lee, K. Zhao, and R. A. Young. Control of cell

identity genes occurs in insulated neighborhoods in mammalian chromosomes. *Cell*, 159(2):374–387, Oct 2014.

[6] M. J. Fullwood and Y. Ruan. ChIP-based methods for the identification of long-range chromatin interactions. *J. Cell. Biochem.*, 107(1):30–39, May 2009.

[7] V. Gotea, A. Visel, J. M. Westlund, M. A. Nobrega, L. A. Pennacchio, and I. Ovcharenko. Homotypic clusters of transcription factor binding sites are a key component of human promoters and enhancers. *Genome Res.*, 20(5):565–577, May 2010.

[8] S. I. Grewal and D. Moazed. Heterochromatin and epigenetic control of gene expression. *Science*, 301(5634):798–802, Aug 2003.

[9] Y. Guo, S. Mahony, and D. K. Gifford. High resolution genome wide binding event finding and motif discovery reveals transcription factor spatial binding constraints. *PLoS Comput Biol*, 8(8):e1002638, 2012.

[10] L. Handoko, H. Xu, G. Li, C. Y. Ngan, E. Chew, M. Schnapp, C. W. Lee, C. Ye, J. L. Ping, F. Mulawadi, E. Wong, J. Sheng, Y. Zhang, T. Poh, C. S. Chan, G. Kunarso, A. Shahab, G. Bourque, V. Cacheux-Rataboul, W. K. Sung, Y. Ruan, and C. L. Wei. CTCF-mediated functional chromatin interactome in pluripotent cells. *Nat. Genet.*, 43(7):630–638, Jun 2011.

[11] N. Heidari, D. H. Phanstiel, C. He, F. Grubert, F. Jahanbani, M. Kasowski, M. Q. Zhang, and M. P. Snyder. Genome-wide map of regulatory interactions in the human genome. *Genome Res.*, 24(12):1905–1917, Dec 2014.

[12] K. L. Huisinga, B. Brower-Toland, and S. C. Elgin. The contradictory definitions of heterochromatin: transcription and silencing. *Chromosoma*, 115(2):110–122, Apr 2006.

[13] A. Imhof. Epigenetic regulators and histone modification. *Brief Funct Genomic Proteomic*, 5(3):222–227, Sep 2006.

[14] A. Imhof and A. P. Wolffe. Transcription: gene control by targeted histone acetylation. *Curr. Biol.*, 8(12):R422–424, Jun 1998.

[15] M. Kato, N. Hata, N. Banerjee, B. Futcher, and M. Q. Zhang. Identifying combinatorial regulation of transcription factors and binding motifs. *Genome Biol.*, 5(8):R56, 2004.

[16] K. R. Kieffer-Kwon, Z. Tang, E. Mathe, J. Qian, M. H. Sung, G. Li, W. Resch, S. Baek, N. Pruett, L. Grøntved, L. Vian, S. Nelson, H. Zare, O. Hakim, D. Reyon, A. Yamane, H. Nakahashi, A. L. Kovalchuk, J. Zou, J. K. Joung, V. Sartorelli, C. L. Wei, X. Ruan, G. L. Hager, Y. Ruan, and R. Casellas. Interactome maps of mouse gene regulatory domains reveal basic principles of transcriptional regulation. *Cell*, 155(7):1507–1520, Dec 2013.

[17] A. Kiermaier and M. Eilers. Transcriptional control: calling in histone deacetylase. *Curr. Biol.*, 7(8):R505–507, Aug 1997.

[18] G. Li, X. Ruan, R. K. Auerbach, K. S. Sandhu, M. Zheng, P. Wang, H. M. Poh, Y. Goh, J. Lim, J. Zhang, H. S. Sim, S. Q. Peh, F. H. Mulawadi, C. T. Ong, Y. L. Orlov, S. Hong, Z. Zhang, S. Landt, D. Raha, G. Euskirchen, C. L. Wei, W. Ge, H. Wang, C. Davis, K. I. Fisher-Aylor, A. Mortazavi, M. Gerstein, T. Gingeras, B. Wold, Y. Sun, M. J. Fullwood, E. Cheung, E. Liu, W. K. Sung, M. Snyder, and Y. Ruan. Extensive promoter-centered chromatin interactions provide a topological basis for transcription regulation. *Cell*, 148(1-2):84–98, Jan 2012.

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

[20] M. S. Luijsterburg and H. van Attikum. Chromatin and the DNA damage response: the cancer connection. *Mol Oncol*, 5(4):349–367, Aug 2011.

[21] H. K. MacAlpine, R. Gordan, S. K. Powell, A. J. Hartemink, and D. M. MacAlpine. Drosophila ORC localizes to open chromatin and marks sites of cohesin complex loading. *Genome Res.*, 20(2):201–211, Feb 2010.

[22] S. Maezawa, M. Yukawa, K. G. Alavattam, A. Barski, and S. H. Namekawa. Dynamic reorganization of open chromatin underlies diverse transcriptomes during spermatogenesis. *Nucleic Acids Res.*, 46(2):593–608, Jan 2018.

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

[24] P. Polak, M. S. Lawrence, E. Haugen, N. Stoletzki, P. Stojanov, R. E. Thurman, L. A. Garraway, S. Mirkin, G. Getz, J. A. Stamatoyannopoulos, and S. R. Sunyaev. Reduced local mutation density in regulatory DNA of cancer genomes is linked to DNA repair. *Nat. Biotechnol.*, 32(1):71–75, Jan 2014.

[25] K. Robzyk, J. Recht, and M. A. Osley. Rad6-dependent ubiquitination of histone H2B in yeast. *Science*, 287(5452):501–504, Jan 2000.

[26] Y. Shi, F. Lan, C. Matson, P. Mulligan, J. R. Whetstine, P. A. Cole, R. A. Casero, and Y. Shi. Histone demethylation mediated by the nuclear amine oxidase homolog LSD1. *Cell*, 119(7):941–953, Dec 2004.

[27] Y. Shiio and R. N. Eisenman. Histone sumoylation is associated with transcriptional repression. *Proc. Natl. Acad. Sci. U.S.A.*, 100(23):13225–13230, Nov 2003.

[28] A. Shilatifard. Chromatin modifications by methylation and ubiquitination: implications in the regulation of gene expression. *Annu. Rev. Biochem.*, 75:243–269, 2006.

[29] J. M. Simon, K. E. Hacker, D. Singh, A. R. Brannon, J. S. Parker, M. Weiser, T. H. Ho, P. F. Kuan, E. Jonasch, T. S. Furey, J. F. Prins, J. D. Lieb, W. K. Rathmell, and I. J. Davis. Variation in chromatin accessibility in human kidney cancer links H3K36 methyltransferase loss with widespread RNA processing defects. *Genome Res.*, 24(2):241–250, Feb 2014.

[30] R. P. Smith, L. Taher, R. P. Patwardhan, M. J. Kim, F. Inoue, J. Shendure, I. Ovcharenko, and N. Ahituv. Massively parallel decoding of mammalian regulatory sequences supports a flexible organizational model. *Nat. Genet.*, 45(9):1021–1028, Sep 2013.

[31] A. W. Snowden, P. D. Gregory, C. C. Case, and C. O. Pabo. Gene-specific targeting of H3K9 methylation is sufficient for initiating repression in vivo. *Curr. Biol.*, 12(24):2159–2166, Dec 2002.

[32] S. Spicuglia and L. Vanhille. Chromatin signatures of active enhancers. *Nucleus*, 3(2):126–131, Mar 2012.

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

[34] S. Thomas, X. Y. Li, P. J. Sabo, R. Sandstrom, R. E. Thurman, T. K. Canfield, E. Giste, W. Fisher, A. Hammonds, S. E. Celniker, M. D. Biggin, and J. A. Stamatoyannopoulos. Dynamic reprogramming of chromatin accessibility during Drosophila embryo development. *Genome Biol.*, 12(5):R43, 2011.

[35] R. E. Thurman, E. Rynes, R. Humbert, J. Vierstra, M. T. Maurano, E. Haugen, N. C. Sheffield, A. B. Stergachis, H. Wang, B. Vernot, K. Garg, S. John, R. Sandstrom, D. Bates, L. Boatman, T. K. Canfield, M. Diegel, D. Dunn, A. K. Ebersol,

T. Frum, E. Giste, A. K. Johnson, E. M. Johnson, T. Kutyavin, B. Lajoie, B. K. Lee, K. Lee, D. London, D. Lotakis, S. Neph, F. Neri, E. D. Nguyen, H. Qu, A. P. Reynolds, V. Roach, A. Safi, M. E. Sanchez, A. Sanyal, A. Shafer, J. M. Simon, L. Song, S. Vong, M. Weaver, Y. Yan, Z. Zhang, Z. Zhang, B. Lenhard, M. Tewari, M. O. Dorschner, R. S. Hansen, P. A. Navas, G. Stamatoyannopoulos, V. R. Iyer, J. D. Lieb, S. R. Sunyaev, J. M. Akey, P. J. Sabo, R. Kaul, T. S. Furey, J. Dekker, G. E. Crawford, and J. A. Stamatoyannopoulos. The accessible chromatin landscape of the human genome. *Nature*, 489(7414):75–82, Sep 2012.

[36] M. Yu and B. Ren. The Three-Dimensional Organization of Mammalian Genomes. *Annu. Rev. Cell Dev. Biol.*, 33:265–289, 10 2017.

[37] Y. Zhang, C. H. Wong, R. Y. Birnbaum, G. Li, R. Favaro, C. Y. Ngan, J. Lim, E. Tai, H. M. Poh, E. Wong, F. H. Mulawadi, W. K. Sung, S. Nicolis, N. Ahituv, Y. Ruan, and C. L. Wei. Chromatin connectivity maps reveal dynamic promoter-enhancer long-range associations. *Nature*, 504(7479):306–310, Dec 2013.

Chapter 2

seqgra: Principled Selection of Neural Network Architectures for Genomics Prediction Tasks

The contents of this chapter have been submitted for publication and a preprint is available at:

seqgra: Principled Selection of Neural Network Architectures for Genomics Prediction Tasks
Konstantin Krismer, Jennifer Hammelman, David K. Gifford
bioRxiv 2021.06.14.448415
DOI: 10.1101/2021.06.14.448415

Supplementary information can be found in Appendix A.

2.1 Abstract

Sequence models based on deep neural networks have achieved state-of-the-art performance on regulatory genomics prediction tasks, such as chromatin accessibility and transcription factor binding. But despite their high accuracy, their contributions to a mechanistic understanding of the biology of regulatory elements is often hindered by the complexity of the predictive model and thus poor interpretability of its decision boundaries. To address this, we introduce seqgra, a deep learning pipeline that incorporates the rule-based simulation of biological sequence data and the training and evaluation of models, whose decision boundaries mirror the rules from the simulation process. The method can be used to (1) generate data under the assumption of a hypothesized model of genome regulation, (2) identify neural network architectures capable of recovering the rules of said model, and (3) analyze a model's predictive performance as a function of training set size, noise level, and the complexity of the rules behind the simulated data.

2.2 Introduction

Over the last five to ten years, neural networks were successfully applied to make large gains on a wide range of tasks in such diverse fields as computer vision, computer audition, natural language processing, and robotics. While the structure and the semantics of the data used to train and evaluate neural networks can be vastly different, the core learning algorithms are almost always the same and the neural network architectures are often composed of similar building blocks. This is also true for the field of genomics, and computational biology as a whole, where deep neural networks are trained on data that are obtained experimentally using functional genomics assays such as DNase-seq [5], ATAC-seq [6], and ChIP-seq. Motivated by their success, architectural building blocks commonly seen in these networks, such as convolutional layers, recurrent layers, batch normalization, drop-out, and skip connections [15, 22, 31, 20], have been imported from computer vision and other fields. This cross-fertilization between fields and the general applicability of the building blocks of deep learning has more recently been seen in the adoption of transformer-based architectures for image classification tasks in computer vision and protein prediction tasks in biology. However, most data sets used to train supervised deep learning models in biology are different from data sets in computer vision and natural language processing in two ways. (1) Biological problems contain noisy input and noisy labels in that not only is there substantial intra-class variability and noise in the input, e.g., images labeled as *cat* contain cats that vary in terms of breed, color, position, pose, etc., but also a significant fraction of examples are mislabeled, i.e., images labeled as *cat* are empty or contain dogs. This is rare in computer vision data sets, but common in data sets derived from functional genomics assays. (2) Feature attribution or other model explanation methods are not human-interpretable. We understand images of cats in the sense that we know which parts of the image contain information that is relevant for the classification (because they belong to the cat) and which parts are irrelevant (because they belong to the background). This intuitive understanding is necessary when attribution methods such as saliency maps are applied to assess a model's ability to base predictions on relevant parts of the input. In biology, examples often include DNA sequence windows of various widths, most commonly 1000 base pairs (bp), which, unlike images of cats, are not *human-readable.* This biology-specific issue of inherently opaque examples exacerbates the general interpretability issue of deep neural networks, whereas the lack of high quality data sets contributes to the reproducibility crisis and makes it more difficult to compare architectures, as they are often only evaluated on a custom data set.

The method introduced here, *seqgra*, attempts to improve the process by which neural network architectures are chosen for specific genomics prediction tasks and provides a framework to evaluate model interpretation methods. Its fully reproducible pipeline provides a means to (1) simulate data based on a pre-defined set of probabilistic rules, (2) create and train models based on a precise description of their architecture, loss, optimizer, and training process, and (3) evaluate the trained models using conventional test set metrics as well as an array of feature attribution methods. These feature attribution methods in combination with simulated data and thus perfect ground truth enable an analysis of the model's decision boundaries and how well they capture the underlying rules of the data generation process from step 1. Utilizing this framework, models are not only evaluated based on their predictive performance, but also on the ability to recover the vocabulary (e.g., specific transcription factor binding site motifs) and grammar (e.g., spacing constraints between interacting transcription factors) of the data set, while assigning little weight to confounding factors and idiosyncratic noise.

Efforts in this area include Kipoi [1], a repository for trained genomics models, and Selene [8], a framework for biological sequence based deep learning models that supports training of PyTorch models, model evaluation with conventional test set metrics (ROC and precision-recall curves), and variant effect prediction and *in silico* mutagenesis of trained models. To our knowledge none of the existing methods offer functionality for simulating data using a general framework of probabilistic rules, nor do they incorporate feature attribution methods.

Furthermore, this simulation-based framework can also serve as a testbed for hy-

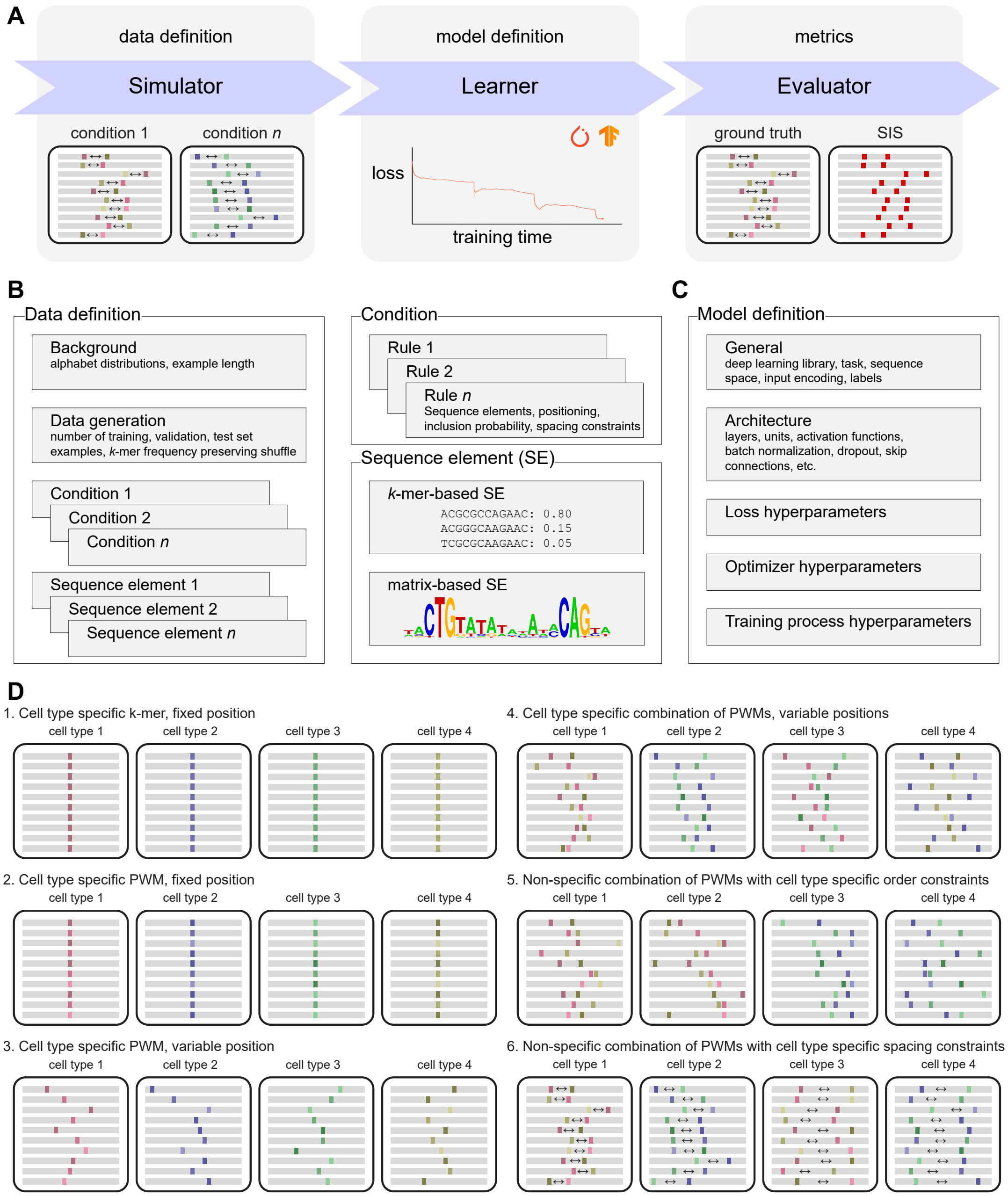

A
data definition
Simulator
condition 1
condition *n*
model definition
Learner
loss
training time
metrics
Evaluator
ground truth
SIS
B
Data definition
Background
alphabet distributions, example length
Data generation
number of training, validation, test set examples, *k*-mer frequency preserving shuffle
Condition 1
Condition 2
Condition *n*
Sequence element 1
Sequence element 2
Sequence element *n*
Condition
Rule 1
Rule 2
Rule *n*
Sequence elements, positioning, inclusion probability, spacing constraints
Sequence element (SE)
k-mer-based SE
ACGCGCCAGAAC: 0.80
ACGGGCAAGAAC: 0.15
TCGCGCAAGAAC: 0.05
matrix-based SE
C
Model definition
General
deep learning library, task, sequence space, input encoding, labels
Architecture
layers, units, activation functions, batch normalization, dropout, skip connections, etc.
Loss hyperparameters
Optimizer hyperparameters
Training process hyperparameters
D
1. Cell type specific k-mer, fixed position
2. Cell type specific PWM, fixed position
3. Cell type specific PWM, variable position
4. Cell type specific combination of PWMs, variable positions
5. Non-specific combination of PWMs with cell type specific order constraints
6. Non-specific combination of PWMs with cell type specific spacing constraints
cell type 1
cell type 2
cell type 3
cell type 4

Figure 2-1: A framework for simulation-based evaluation of neural network architectures. (figure caption continued on next page)

Figure 2-1: A framework for simulation-based evaluation of neural network architectures. (**A**) Schematic of the three main components: First, a simulator generates synthetic data according to the rules and specifications defined in the data definition file. Second, a learner creates a neural network model whose architecture and hyperparameters are specified in the model definition file, and trains it on the synthetic data from step 1. And third, the trained model is evaluated in terms of predictive performance and its ability to recover the rules specified in the data definition file. (**B**) The data definition specifies the basic properties of the synthetic data, including the alphabet (e.g., DNA, RNA, protein) and its distribution, as well as condition-specific rules (the *grammar*), which determine how information about the label y is encoded in the input x. (**C**) The model definition contains all information required to create and train the model. (**D**) A schematic of six simulated toy data sets for multi-class classification, where the classes y correspond to cell types and the input x are sequence windows (depicted as gray bars) that encode information about the class y at certain positions in x (colored areas). The rules that determine how this information is encoded range from basic (cell type specific k-mer at fixed position) to complex (non-specific combinations of position weight matrices with cell type specific spacing constraints).

potheses about biological phenomena or as a means to investigate the strengths and weaknesses of various feature attribution methods across different neural network architectures that are trained on data sets with varying degrees of complexity. In the former use case, the hypothesis is encoded in the rules of the simulation process to identify an appropriate neural network architecture, which is subsequently trained and evaluated on experimental data. The performance of this simulation-vetted architecture on experimental data serves as an indication of the validity of the hypothesis and its underlying assumptions about the biological phenomenon.

2.3 Materials and Methods

2.3.1 Position probability matrices and position weight matrices

We use position probability matrices (PPM) with a DNA alphabet ($\Sigma = \{\mathrm{A}, \mathrm{C}, \mathrm{G}, \mathrm{T}\}$) to represent sequence motifs:

$$\overbrace{\left(\begin{array}{ccccc} & \mathrm{A} & \mathrm{C} & \mathrm{G} & \mathrm{T} \\ 1 & y_{1,\mathrm{A}} & y_{1,\mathrm{C}} & y_{1,\mathrm{G}} & y_{1,\mathrm{T}} \\ 2 & y_{2,\mathrm{A}} & y_{2,\mathrm{C}} & y_{2,\mathrm{G}} & y_{2,\mathrm{T}} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ n & y_{n,\mathrm{A}} & y_{n,\mathrm{C}} & y_{n,\mathrm{G}} & y_{n,\mathrm{T}} \end{array}\right)}^{\text{PPM}} \tag{2.1}$$

As the name suggests, each cell of a PPM is a probability, the probability of observing a particular nucleotide at a particular position, and each row sums to one, i.e., at each position one of the four nucleotides must be present. We use the notation $\mathrm{PPM}_k(i, j)$ to access the probability of observing the jth nucleotide at the ith position in a specific PPM_k.

These PPMs usually describe experimentally obtained estimates of transcription factor binding sites, but may also describe artificially constructed sequence motifs.

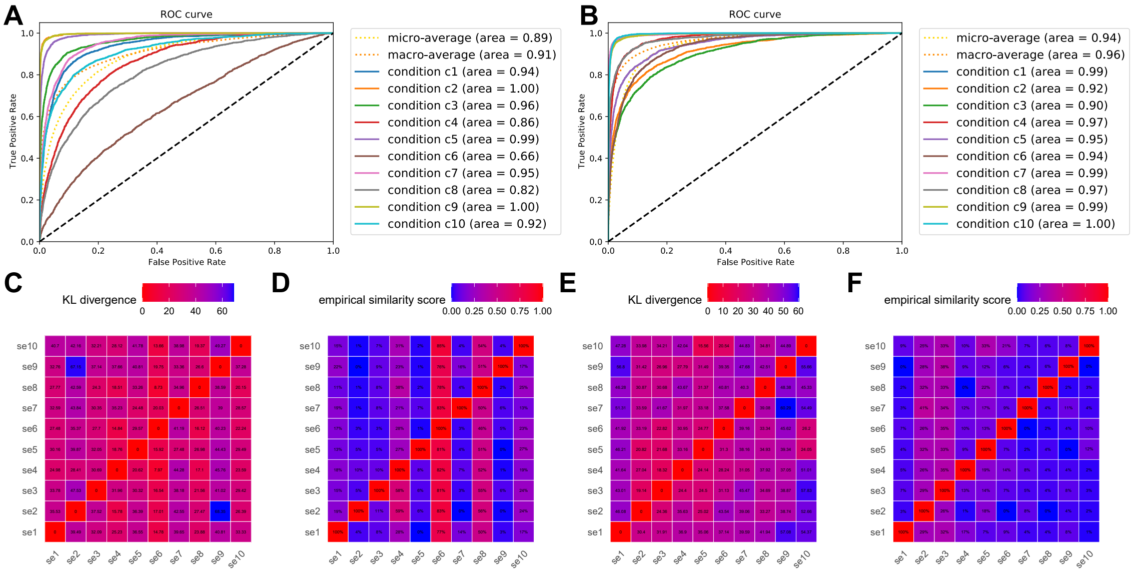


Figure 2-2: Selection of sequence motifs for simulation grammars. (**A**) ROC curve of Bayes Optimal Classifier on multi-class classification task with 10 classes, prior to filtering out ambiguous sequence motifs. (**B**) Same as panel A, after ambiguous sequence motifs were removed. (**C**) KL divergence matrix of 10 sequence motifs, prior to filtering. (**D**) Empirical similarity score matrix of 10 sequence motifs, prior to filtering. (**E**) Same as panel C, after removing ambiguous motifs. (**F**) Same as panel D, after removing ambiguous motifs.

To calculate the likelihood of a sequence given a PPM, we first convert the PPM to a position weight matrix (PWM) by transforming the elements of the PPM to log likelihoods,

$$y'_{i,j} = \log_2 \frac{y_{i,j}}{p_j}, \tag{2.2}$$

using background sequence probabilities $\boldsymbol{p}$, which are described in section 2.3.6. The *score* of a particular position in a DNA sequence is then calculated by adding the value of the observed nucleotide at each position in the PWM.

2.3.2 Motif information content

To calculate the information content of a sequence motif represented as a PPM, we first calculate $\mathrm{U}(i)$, the uncertainty at position i as follows:

$$\mathrm{U}(i) = -\sum_{j\in\Sigma} \mathrm{PPM}(i,j) \times \log_2(\mathrm{PPM}(i,j)). \tag{2.3}$$

The information content at position i is then defined as follows

$$\mathrm{IC}(i) = t - \mathrm{U}(i), \tag{2.4}$$

where $t = \log_2(|\Sigma|)$, the total information content per position in bits. In order to obtain MIC, the information content of the entire motif, we add up the individual positions:

$$\mathrm{MIC} = \sum_{i=1}^{n} \mathrm{IC}(i), \tag{2.5}$$

where n is the motif width in nucleotides (nt), see matrix in 2.1.

2.3.3 Relative entropy between motif and background distribution

The information content of a motif is a special case of the relative entropy of a motif where background probabilities $\boldsymbol{p}$ are uniform. Relative entropy, also known as KL divergence, between a motif and the background distribution is calculated per position, similarly to IC:

$$D_{\mathrm{KL}}(i) = \sum_{j \in \Sigma} \mathrm{PPM}(i,j) \times \log_2 \left(\frac{\mathrm{PPM}(i,j)}{p_j} \right), \tag{2.6}$$

and then summed over positions to obtain the Motif Relative Entropy,

$$\mathrm{MRE} = \sum_{i=1}^{n} D_{\mathrm{KL}}(i). \tag{2.7}$$

2.3.4 Relative entropy between two motifs

While the relative entropy between a particular motif, PPM_1, and the background distribution is a way to gauge the learnability of a grammar where the presence of PPM_1 carries information, the relative entropy between two motifs, PPM_1 and PPM_2, is equally useful to assess the learnability of grammars with multiple, semantically distinct sequence elements.

By slightly adjusting the D_{KL} from above, we calculate the KL divergence of position i between two motifs as follows:

$$\begin{gathered} D_{\mathrm{KL}}(\mathrm{PPM}_1, \mathrm{PPM}_2, i) = \\ \sum_{x \in \Sigma} \mathrm{PPM}_1(i,x) \times \log_2 \left(\frac{\mathrm{PPM}_1(i,x)}{\mathrm{PPM}_2(i,x)} \right). \end{gathered} \tag{2.8}$$

The motif pair relative entropy of PPM_1 relative to PPM_2 is then defined as

$$\mathrm{MPRE}(\mathrm{PPM}_1, \mathrm{PPM}_2) = \sum_{i=1}^{n} D_{\mathrm{KL}}(\mathrm{PPM}_1, \mathrm{PPM}_2, i). \tag{2.9}$$

To calculate the MPRE between motifs of unequal width, we pad the shorter motifs with *neutral* positions using background probabilities.

Another issue with equation 2.9 is that it does not capture highly similar but shifted motifs. PPM_1 might be equivalent to PPM_2 shifted by one position and thus considered highly similar, but $\mathrm{MPRE}(\mathrm{PPM}_1, \mathrm{PPM}_2)$ in its current form does not reflect this. To resolve this, we calculate $\mathrm{MPRE}(\mathrm{PPM}_1, \mathrm{PPM}_2)$ for several alignments of PPM_1 and PPM_2 and take the minimum.

2.3.5 Empirical similarity score between two motifs

The empirical similarity score (ESS) between PPM_1 and PPM_2 is another way to assess the similarity between two motifs and thus the difficulty to distinguish between them. $\mathrm{ESS}(\mathrm{PPM}_1, \mathrm{PPM}_2)$ is calculated by generating k (in this work, $k = 100$) instances of motif 2, flanked on both sides by background sequences of length n_1, where n_1 is the width of PPM_1. All positions of these k sequences are then scored by PWM_1 (the position weight matrix of PPM_1), and the highest score per sequence is returned. $\mathrm{ESS}(\mathrm{PPM}_1, \mathrm{PPM}_2)$ is then the mean of these k scores. ESS motif matrix plots (Figure 2-2 and Supplementary Figure A-3) depict adjusted empirical similarity scores, which are shifted by ESS_0 if $\mathrm{ESS}_0 < 0$, where $\mathrm{ESS}_0 = \min_j \mathrm{ESS}(\mathrm{PPM}_i, \mathrm{PPM}_j)$, and normalized such that the self similarity score $\mathrm{ESS}(\mathrm{PPM}_i, \mathrm{PPM}_i) = 1.0$.

Both MPRE and ESS are asymmetric, i.e., $\mathrm{ESS}(\mathrm{PPM}_1, \mathrm{PPM}_2) \neq \mathrm{ESS}(\mathrm{PPM}_2, \mathrm{PPM}_1)$.

2.3.6 Alphabet distribution for grammars

For all grammars discussed in this paper, we used the natural nucleotide distribution of the human genome, 29.565 % adenine (A), 20.435 % cytosine (C), 20.435 % guanine (G), and 29.565 % thymine (T) [21].

2.3.7 Motif database

We used HOMER motifs for all grammar sequence elements that were based on transcription factor binding site motifs. These motifs were obtained by analyzing data from publicly available ChIP-seq experiments [11].

2.3.8 Feature importance evaluators

While conventional test set metrics, such as ROC curves and precision-recall curves, assess model performance based on a set of examples (e.g., the test set), feature importance evaluators quantify the contribution of each input feature to the model's prediction. In the context of seqgra, feature importance evaluators are used to assess what we call grammar or vocabulary recovery, the degree to which a model was able to align its decision boundaries with the rules of the grammar that was used to simulate the data it was trained on. This is possible because for simulated data we not only know the ground truth label for each example, but also which positions are part of the background and thus contain no information about the class label, and which positions were altered by a grammar rule and thus do contain information about the class label. These position-level annotations (*background positions*, *grammar positions*) are provided for all simulated examples.

More formally, feature importance evaluators take a model $f(x)$, a target y and an example x_i of width n, and return z, an n-dimensional vector that contains the attribution value (also known as importance, relevance, contribution) of each input position to model $f(x)$ predicting target y. Please note that n is the sequence length of the example, not the number of features. For instance, if the input to the model is a 150 nt DNA sequence, x_i is a 150 by 4 matrix (one-hot encoded), containing 600 features, but its width $n = 150$. Feature attribution values in seqgra are grouped and reported at the position level, not the input feature level.

Attribution values are visualized with so-called grammar agreement plots, which are heatmaps depicting attributions and position-level annotations of several examples. The plots encode the attribution values in the color luminosity, where lighter colors indicate low values (low feature importance) and dark colors indicate high values (high feature importance). The position-level annotations are encoded in the color hue, with grammar positions in green and background positions in red.

2.3.9 Gradient-based feature importance evaluators

This large class of feature importance evaluators (FIEs) uses backpropagation to calculate the partial derivatives of the output, $f_y(x)$, with respect to the input, x_i. seqgra includes seven gradient-based feature importance evaluators off-the-shelf, whose implementations are based on code by Yulong Wang [28].

The most basic FIE, **raw gradient** [24], just returns the gradient with respect to the input example x_i:

$$z_{\text{RG}} = \frac{\partial f_y(x)}{\partial x_i}, \tag{2.10}$$

or short $\nabla f_y(x_i)$, where $f_j(\cdot)$ is the activation of the target neuron in the output layer, e.g., class j for multi-class classification tasks.

The absolute gradient method or **saliency** is defined as

$$z_{\text{S}} = |\nabla f_y(x_i)|, \tag{2.11}$$

where $|x|$ applies the element-wise absolute value operation to vector x.

Gradient-x-input [2] (gradient times input) is defined as

$$z_{\text{GI}} = x_i \nabla f_y(x_i). \tag{2.12}$$

Integrated Gradients [27] takes the average of multiple (here, $K = 100$) gradients evaluated along the linear path from the baseline x_0 (which in seqgra is the zero vector) to the input example x_i. The method is defined as

$$z_{\text{IG}} = \frac{1}{K} \sum_{k}^{K} \nabla f_y \left(\frac{k}{K} x_i \right). \tag{2.13}$$

seqgra also supports gradient-based methods that alter the way the gradient is obtained using backpropagation, namely **Guided Backpropagation** [25], **Deconvolution** [29], and **DeepLIFT** [23]. The details of these methods are beyond the scope of this work.

2.3.10 Model-agnostic feature importance evaluators

Model-agnostic FIEs do not require access to the gradients and make no assumptions about the structure of the model, hence the name. They rely solely on the ability to evaluate $f_y(x)$, for various altered versions of x.

Sufficient Input Subsets (SIS) [7] is a perturbation-based method that identifies subsets of input features that are sufficient to keep $f_y(x) > \tau$, i.e., if all other features are masked, the class prediction does not change (is still above some threshold τ). Unlike gradient-based FIEs, which return a real-valued vector of feature attributions, SIS returns a binary vector, indicating for each feature whether it is part of a sufficient input subset or not.

2.3.11 Hardware infrastructure

Models presented in this paper were trained on three compute nodes with a total of 6 CPUs (2x Intel Xeon E5-2630 v4, 2x Intel Xeon Gold 6138, 2x Intel Xeon Gold 6240), 26 GPUs (8x NVIDIA GeForce GTX 1080 Ti with 11 GB GDDR5X, 10x NVIDIA GeForce RTX 2080 Ti with 11 GB GDDR6, and 8x NVIDIA Titan RTX with 24 GB GDDR6), and a total of 833 GB of main memory. The total GPU time (for training and evaluation) was roughly 12 GPU months.

2.3.12 Software infrastructure

All seqgra data presented in this paper was obtained on machines running Ubuntu 18.04.3 LTS, CUDA 10.1, cuDNN 7.6.5, Python 3.8, NumPy 1.19.2, TensorFlow 2.2.0, PyTorch 1.7.0, and R 4.0.

2.4 Results

2.4.1 seqgra provides a reproducible, simulation-based framework for neural network architecture evaluation

The method we describe in this paper (seqgra) generates synthetic biological sequence data according to predefined probabilistic rules in order to either (1) evaluate neural network architectures trained on these data sets, or (2) test whether the assumptions about the underlying biological phenomenon that the probabilistic rules of the simulation process are based on, accurately reflect experimentally obtained data. In the

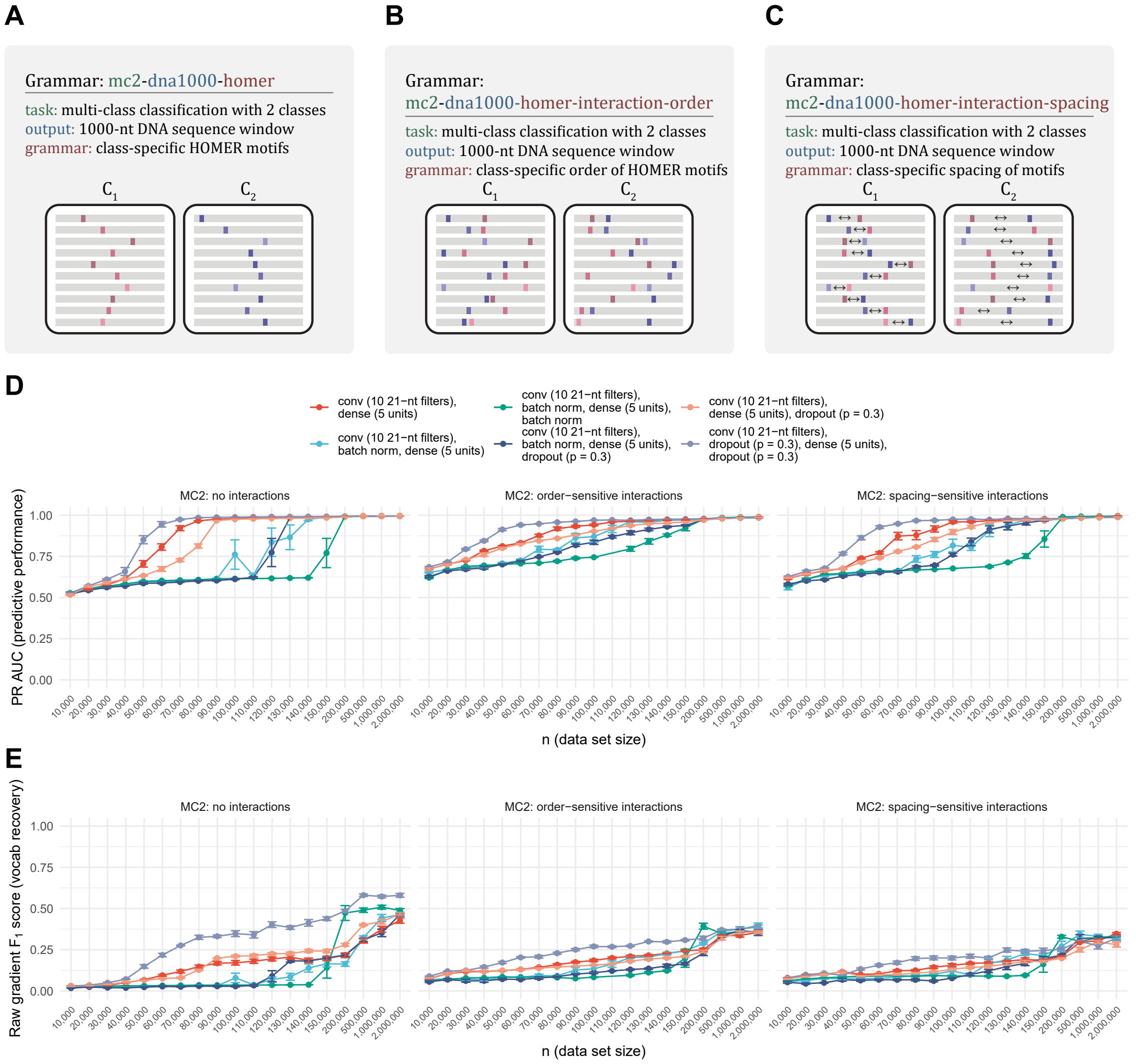


Figure 2-3: seqgra-enabled ablation analysis reveals most efficient neural network architecture. (**A**) Schematic of binary classification grammar using class-specific HOMER motifs as sequence elements. (**B**) Schematic of binary classification grammar using class-specific order of HOMER motifs. (**C**) Schematic of binary classification grammar using class-specific spacing of HOMER motifs. (**D**) Predictive performance of six neural network architectures with and without batch normalization and dropout. (**E**) Vocabulary recovery of six neural network architectures with and without batch normalization and dropout.

former scenario, the result would be a neural network architecture that—when trained on data sets generated from a similar set of rules—has high predictive performance and decision boundaries that closely reflect those set of generative rules. The goal of the latter approach is to arrive at a concise set of probabilistic rules that approximates the biological process in question, and a neural network architecture whose high performance on simulated data is recapitulated when trained on experimental data.

A data set in the context of seqgra, whether obtained by simulation or experiment, is always divided into three subsets, training set, validation set, and test set. Each of the subsets comprises a number of supervised examples, which are (x, y, a)-triplets. Here, the input variable x is a biological sequence (DNA, RNA, protein) of fixed or variable length, also referred to as sequence window or features; y is the target variable, the *condition* this example belongs to (e.g., cell type), which is either a mutually exclusive *class* or a non-mutually exclusive *label*, for multi-class classification tasks or multi-label classification tasks, respectively; and a is the positional annotation of the example, denoting for each position in x whether it is part of the *grammar* or part of the *background*. Grammar positions contain information related to y and are therefore important for classification, whereas background positions do not and are thus irrelevant for classification.

The core functionality of seqgra can be broken down into three components: (1) Simulator, (2) Learner, and (3) Evaluator. Each component corresponds to a distinct step in the pipeline depicted in Figure 2-1A.

In step 1, the simulator generates a synthetic data set according to the specifications laid out in the data definition (see Figure 2-1B), a document that contains a precise description of the generated data, from the background nucleotide distribution to the set of probabilistic rules that determines how information about the condition y (label, class) is encoded in the sequence window x. This set of probabilistic rules is also referred to as *grammar* or sequence grammar throughout this manuscript (hence the name *seq-gra*), and although related to formal grammars, seqgra's probabilistic rules are not expressed as and not equivalent to production rules in the context of

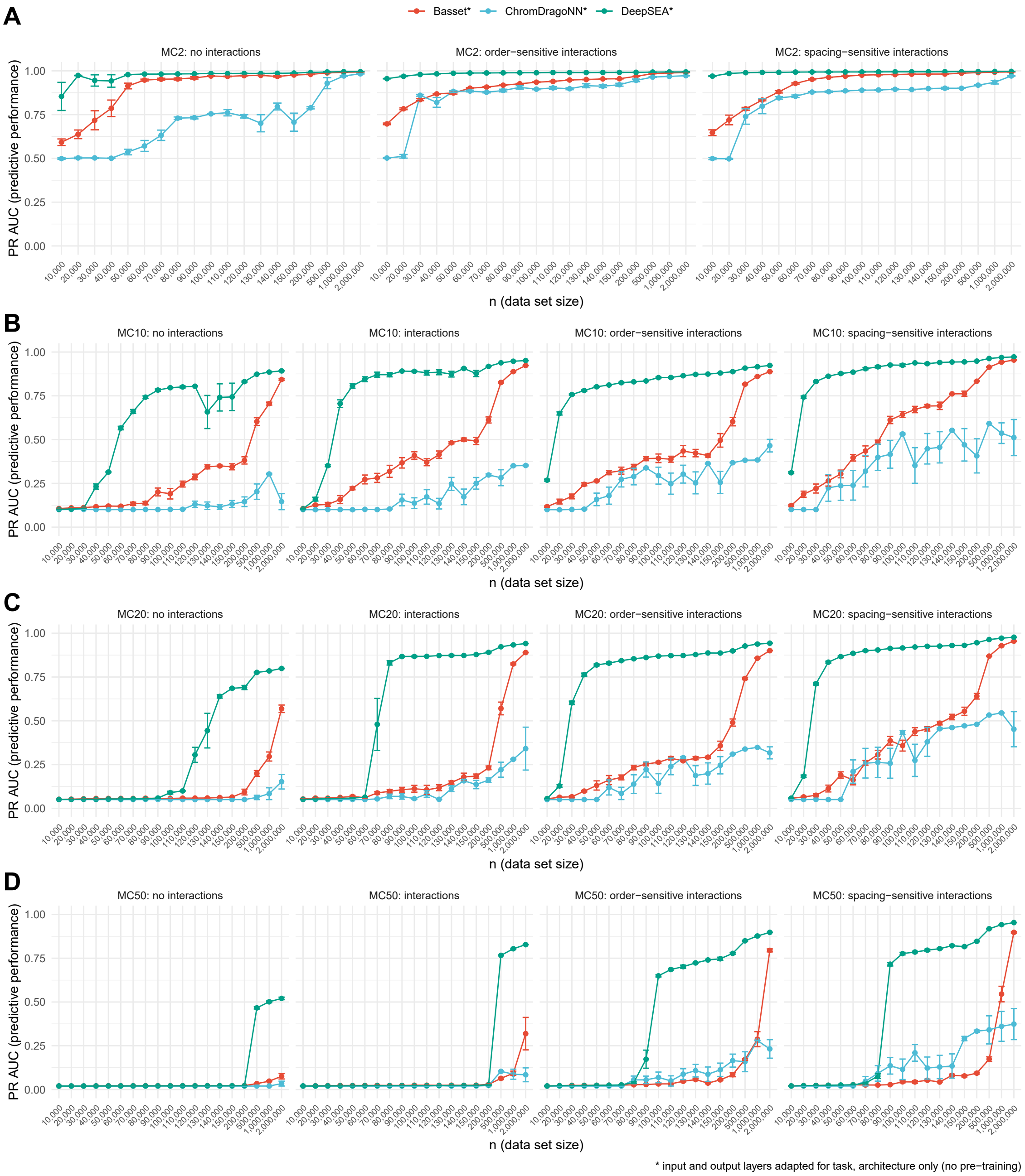


Figure 2-4: Comparison of neural network architectures Basset, ChromDragoNN, and DeepSEA. (figure caption continued on next page)

Figure 2-4: Comparison of neural network architectures Basset, ChromDragoNN, and DeepSEA. (**A**) Predictive performance on binary classification tasks of grammars with class-specific HOMER motifs (left), class-specific order of HOMER motifs (middle), and class-specific spacing of HOMER motifs. All architectures were trained on data sets ranging in size from 10,000 examples to 2,000,000 examples. Error bars are standard errors of five models trained on the same grammar, using five different simulation seeds. (**B**) Same as panel A, for multi-class classification tasks with 10 classes. The second plot from the left shows the predictive performance of models trained on data sets with class-specific interactions of HOMER motifs. (**C**) Same as panel B, for multi-class classification tasks with 20 classes. (**D**) Same as panel B, for multi-class classification tasks with 50 classes.

formal language theory.

Schematic depictions of six toy data sets, generated from probabilistic rules of varying complexity, are shown in Figure 2-1D. In each case, the data set contains examples belonging to one of four classes and the probabilistic rules determine how information about the class y (in this case, the cell type) is encoded in the sequence window x. The ability to recover this relationship during training is imperative for the model's predictive performance. The sequence windows of the examples are shown as gray bars with colored spots, where background positions are shown in gray and grammar positions are shown in color. In the first example, each of the four cell types can easily be identified by the presence of a class-specific k-mer at the center of the sequence window, a relationship that, unsurprisingly, can be learned perfectly (i.e., close to an ROC AUC of 1.0) and efficiently (i.e., with few training examples) by most neural network architectures. Since a set of rules as simple as the one used in example 1 will almost always be an inadequate description of any biological process, seqgra allows for various ways to increase the complexity. Example 2 represents a small step up in complexity by replacing the fixed, class-specific k-mer with a class-specific position weight matrix (PWM), which is a common representation of naturally occurring sequence elements, such as binding sites for a transcription factors. Another small step up in complexity is example 3, where the PWM is placed randomly within in sequence window. In example 4 none of the PWMs is class-specific, only a combination of PWMs. Rules like these could be used to model cell type specific chromatin accessibility that is dependent on the interaction between transcription factors. Examples 5 and 6 encode class information in the relative position of PWMs instead of their presence or absence, with example data set 5 using class-specific order constraints and example data set 6 class-specific spacing constraints.

Once the synthetic data set is generated, it is used by the learner component in step 2 to train a neural network model. It is important to note that the learner only has access to x and y of the (x, y, a) example triplets, and the positional annotations a are only utilized in step 3. Analogous to the role of the data definition for the simulator in step 1, the model definition (see Figure 2-1C) serves as a blueprint for the learner by providing a precise description of the neural network architecture, the loss function, the optimizer, and hyperparameters of the training process, and thus ensuring a reproducible model creation, training, and serving process for both PyTorch and TensorFlow models.

In step 3, the fully trained model from step 2 is then evaluated with the help of an array of conventional test set metrics and feature importance evaluators, such as Integrated Gradients [27] and Sufficient Input Subsets [7].

As a means to illustrate the various inputs and output of this pipeline, we prepared the results of a single seqgra analysis in Supplementary Figure A-2. For this example, we used a simple grammar, similar to the one described in example 1 of Figure 2-1D, but instead of always inserting the class-specific k-mer, we use different insertion probabilities for each class, ranging from 100 % present in examples of class 1, C_1, to 80 % present in C_2, 60 % present in C_3, 40 % present in C_4, 20 % present in C_5, 10 % present in C_6, 5 % present in C_7, and only present in 1 % of C_8 examples. We chose a neural network architecture with two hidden layers, a convolutional layer, followed by a fully connected layer (Supplementary Figure A-2A). After the simulation process finished, diagnostic plots were generated, depicting a heatmap of grammar positions for all examples per class (Supplementary Figure A-2B). These so-called positional grammar probabilities (i.e., the probability for a specific position to be a grammar position), depicted in the heatmap correspond to the insertion probabilities of the grammar, as expected. Furthermore, the class-specific ROC curves in Supplementary Figure A-2C show that the chosen neural network architecture was optimal in terms of predictive performance, with true positive rates of 1.0, 0.8, 0.6, 0.4, 0.2, 0.1, 0.05, and 0.01 (at the zero false positive level) for the classes C_1 to C_8, which are the theoretical upper limits given the insertion probabilities of the underlying grammar. This is also reflected in the precision-recall curves in Supplementary Figure A-2D. In panels E to G we show the results of the feature importance evaluators raw gradient, absolute gradient, and Sufficient Input Subsets (see sections 2.3.9 and 2.3.10 for details). These heatmaps show whether the model's predictions were based on relevant (i.e., grammar) positions and are therefore an indication of the model's ability to recover the underlying grammar of the data set. All three methods suggest high grammar recovery (many dark green positions, few dark red positions).

Supplementary Figure A-2 covered the results obtained from a single seqgra call, evaluating one neural network model trained on one synthetic data set, but most seqgra analyses compare various different architectures across a range of data sets (of different grammar complexities and sizes). For these situations, we provide a suite of convenient commands that streamline these analyses and provide a schematic description of their inputs and outputs in Supplementary Figure A-1.

2.4.2 Selection of unambiguous set of HOMER sequence motifs

In order to generate synthetic data sets that are closer to experimentally obtained data sets, we replaced the artificially constructed k-mers used in the insertion probability grammar of Supplementary Figure A-1 with transcription factor binding site motifs which were obtained from ChIP-seq assays and curated by HOMER [11]. However, before a collection of experimentally obtained motifs can be used effectively as sequence elements in grammars, degenerate motifs must be excluded. These include motifs with low information content and highly similar motif pairs. If these motifs are used as sequence elements that encode information about the condition y, but either cannot be differentiated from the background distribution or motifs specific to one condition are highly similar to motifs specific to another condition, the conditions are rendered inseparable and learning becomes impossible. This scenario is shown in Figure 2-2A, which depicts the test set ROC curves of a Bayes Optimal Classifier (BOC) for 10 classes of a data set generated by a grammar using 10 randomly selected HOMER motifs as class-specific sequence elements. BOCs in the context of seqgra are used to determine whether the conditions of a grammar are separable in principle, i.e., regardless of data set size and neural network architecture. Instead of neural network models whose weights are adjusted during training, the BOC has access to the data definition and uses the rules and sequence elements specified there directly to classify the examples. If the predictive performance of the BOC is low, as is the case with conditions C_6, C_8, and C_4 shown in Figure 2-2A, the rules associated with those conditions are not specific enough to differentiate between them. And since the rules in this case place a supposedly condition-specific sequence element at a random position in the sequence window, the only explanation is that these sequence elements are either indistinguishable from background or indistinguishable from each other. The latter is shown in the matrices in Figure 2-2C and Figure 2-2D, which identify the corresponding sequence elements SE_6, SE_8, and SE_4 as most similar to other sequence elements, i.e., lowest KL divergence and highest empirical similarity score, respectively (for details, see sections 2.3.4 and 2.3.5).

Figure 2-2B shows BOC performance after the most ambiguous motifs were removed, and the corresponding KL divergence and empirical similarity score matrices are shown in Figure 2-2E and F. A collection of experimentally derived sequence motifs will never be completely orthogonal, but the degree of dissimilarity between these 10 were deemed sufficient and all subsequent multi-class classification grammars with 10 classes used these 10 motifs. Supplementary Figure A-3 shows the same selection process for a collection of 100 HOMER motifs. All HOMER motifs used in this study are listed in Supplementary Table A.1, together with a IUPAC notation of the motif, the motif information content (see section 2.3.2) and the KL divergence between the motif and the background distribution (see section 2.3.3). Motifs used for binary classification tasks are listed in Supplementary Table A.2, those for multi-class classification tasks with 10, 20, and 50 classes are listed in Supplementary Tables A.3, A.4, and A.5, respectively.

2.4.3 seqgra-enabled ablation analysis reveals most efficient neural network architecture

Ablation, a technique widely used in neuroscience to determine the functions of brain regions by removing them one by one, has been used similarly to identify the relevant components of an artificial neural network [19, 17]. We performed an ablation analysis to determine the effects of dropout [26] and batch normalization [14] on the predictive performance and grammar recovery of a basic neural network architecture with two hidden layers, a convolutional layer with 10 21-nt wide filters, followed by a dense layer with 5 hidden units, and dropout or batch normalization operations after each layer. Models were trained on binary classification data sets generated by grammars using class-specific HOMER motifs (see schematic in Figure 2-3A), class-specific order of HOMER motifs (Figure 2-3B), and class-specific spacing constraints between HOMER motifs (Figure 2-3C). Test set precision-recall curve AUCs are shown for all models across all grammars in Figure 2-3D. Unsurprisingly, the predictive performance of all architectures increases with data set size, and all architectures approach a PR AUC of 1.0 for sufficiently large data sets. But this analysis reveals a striking difference between the neural network architectures in terms of their efficiency, i.e., how many training examples are required to reach an AUC of approximately 1.0. On the grammars tested here, batch normalization had a negative effect on efficiency, requiring up to 100,000 examples more to converge than architectures without the operation. The architecture with dropout after each hidden layer was the most efficient and highest performing, both in terms of predictive performance and grammar recovery (i.e., the model's propensity to classify examples based on grammar positions) as shown in Figure 2-3E.

2.4.4 DeepSEA dominates comparison of popular genomics deep learning architectures

Furthermore, we compared three popular neural network architectures used in the field of genomics, Basset [15], ChromDragoNN [20], and DeepSEA [31]. All three architectures were devised with functional genomics data sets in mind and were originally trained on multi-label classification data sets obtained from numerous DNase-seq assays, with ChromDragoNN also utilizing RNA-seq and DeepSEA ChIP-seq data. With over 4 million (Basset), over 6 million (DeepSEA), and over 20 million (ChromDragoNN) trainable parameters, all three can be considered high-capacity models. The three architectures make use of commonly used building blocks such as convolutional, followed by dense layers (all three), max pooling and dropout operations (all three), ReLU activation functions (all three), batch normalization (Basset and ChromDragoNN), and skip connections (ChromDragoNN). Input and output layers were adjusted to fit the prediction task and architectures were trained on simulated data sets from scratch without pre-training on their original data sets.

We used the area under the micro-averaged precision-recall curve to evaluate the test set predictive performance on four multi-class classification tasks (with 2, 10, 20, and 50 classes) and three or four grammars each, with a sequence window of 1000 nucleotides. The results are shown in Figure 2-4A for binary classification, and Figures 2-4B, 2-4C, and 2-4D for multi-class classification with 10, 20, and 50 classes, respectively. The HOMER motifs used by the grammars presented here are listed in Supplementary Tables A.2-A.5. Each panel contains precision-recall AUCs of models trained on data sets generated by one grammar, using 5 different random seeds for simulation (error bars) and 19 different data set sizes. The DeepSEA architecture exhibited an at times substantially higher predictive performance than Basset and ChromDragoNN and was the highest performing architecture on all tested data sets.

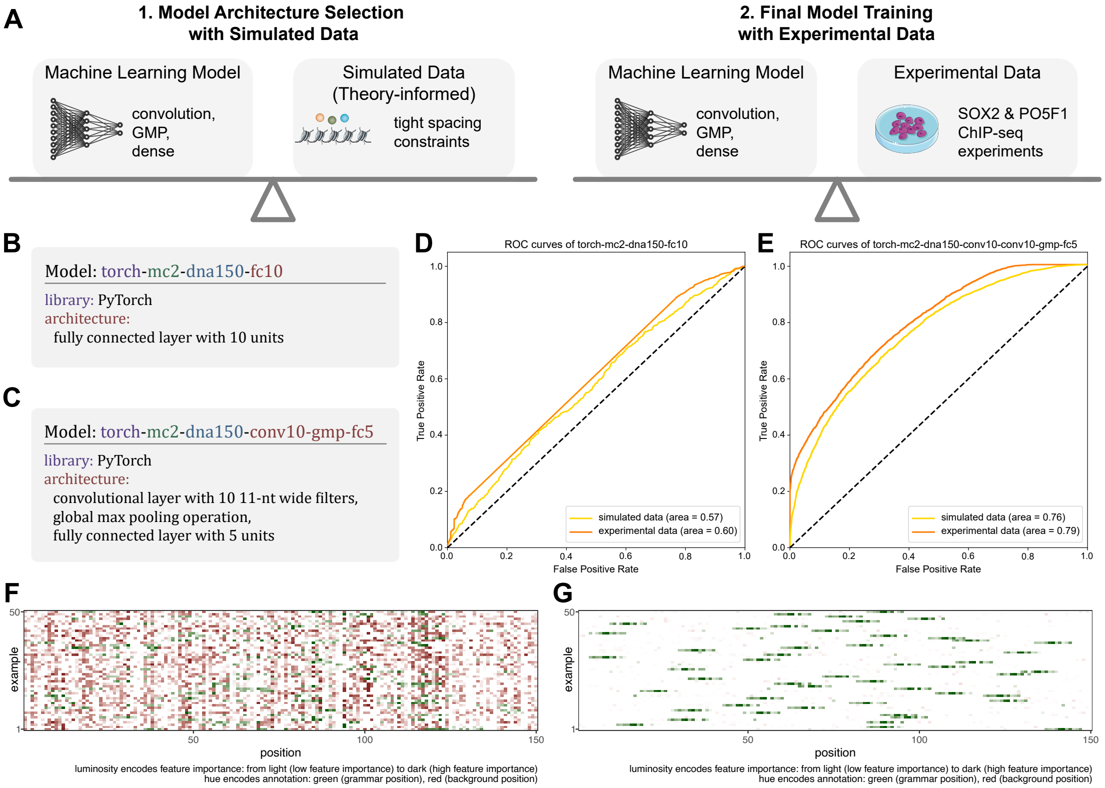


Figure 2-5: Predictive performance and grammar recovery of various model architectures on simulated and experimental data. (**A**) Schematic of model selection process: first, identify suitable model architectures on simulated data; second, train models with simulation-vetted architectures on experimental data. (**B**) Naive neural network architecture with fully connected layer. (**C**) Grammar-informed neural network architecture with convolutional layer, global max pooling, and fully connected layer. (**D**) Predictive performance of naive architecture, trained and evaluated on simulated and experimental data. (**E**) Predictive performance of grammar-informed architecture, trained and evaluated on simulated and experimental data. (**F**) Grammar agreement plot (Integrated Gradients) of naive architecture, trained on experimental data. (**G**) Grammar agreement plot (Integrated Gradients) of grammar-informed architecture, trained on experimental data.

While DeepSEA is the preferred architecture on data sets derived from the grammars we tested, this is not necessarily true for data sets with other grammars or experimentally obtained data. Interestingly, we observed that high capacity architectures such as those tested here perform better on data sets generated by grammars that include interactions, specifically interactions that encode the class label in the order or spacing of the interacting sequence elements. This is not the case for small-scale architectures with less than 100,000 trainable parameters, which, as expected, do better on grammars without interactions, where the class label is encoded in the presence of class-specific sequence elements.

2.4.5 High predictive performance of simulation-vetted neural network architecture recapitulated with ChIP-seq data

In this section we address the question of whether neural network architectures that perform well on simulated data also succeed on data obtained experimentally. We decided to model the well-known hetero-dimeric pair of transcription factors SOX2 and POU5F1, whose spacing constraints were previously characterized [9, 10]. To that end, we used the HOMER motifs `SOX2_HUMAN.H11MO.0.A` and `PO5F1_HUMAN.H11MO.1.A` as sequence elements in the data definition. We also included spacing constraints (0-3 bp between SOX2 and PO5F1 motifs). Figure 2-5A shows a schematic depiction of the analysis.

The experimental data set was based on two ChIP-seq assays, which targeted the two transcription factors. The preprocessed data was obtained from the Cistrome Data Browser [18], specifically the data associated with GEO IDs `GSM1701825` for SOX2 and `GSM1705258` for POU5F1.

We evaluated the same neural network architectures on both the simulated and the experimental data sets. The architecture described in Figure 2-5B with one fully connected layer (not counting the output layer) is an example of an architecture that does not assume any structure in the input. It is a naive architecture in the sense that it was constructed without any knowledge about the grammar that was used to simulate the data. The architecture described in Figure 2-5C, on the contrary, makes assumptions about the data that are in agreement with the grammar, such as a 1D spatial structure with information encoded in 11-nt long code words (enough to cover the SOX2-POU5F1 interaction), whose position in the sequence window is irrelevant.

As expected, the test set predictive performance of the naive architecture (Figure 2-5D) was significantly lower than the grammar-informed architecture (Figure 2-5E). Furthermore, the performance on the simulated data proved to be a good predictor for the performance on the experimental data (Figure 2-5D and E).

The agreement between feature importance and the grammar positions, a proxy for a model's ability to recover the SOX2 and POU5F1 motifs, is shown in Figure 2-5F for the naive architecture and in Figure 2-5G for the grammar-informed architecture. The grammar-informed model's predictions were based almost exclusively on grammar positions (positions that contained SOX2 and POU5F1 motifs), whereas this was not the case for the naive model. Both panels were created with the Integrated Gradients feature importance evaluator.

2.5 Discussion

In this paper we introduced seqgra, a deep learning infrastructure method for genomics. It is intended to streamline the development of deep learning models for biological sequence-based prediction tasks, by providing a reproducible unified framework for (1) flexible, rule-based synthetic data generation; (2) model training; and (3) model evaluation with conventional test set metrics and feature attribution methods. This three-step pipeline supports data sets obtained by simulation and experiment, models implemented in PyTorch and TensorFlow, and numerous gradient-based feature attribution methods as well as Sufficient Input Subsets, a model-agnostic feature attribution method, in addition to conventional ROC and precision-recall curves for model evaluation. Our method greatly simplifies an array of commonly performed diagnostics and performance assessments of deep learning models, such as ablation analysis, estimated data set size requirements, and tolerated noise thresholds. The simulator and the language of the probabilistic rules are flexible enough to span multi-class and multi-label classification tasks with any number of classes or labels, DNA or amino acid sequence windows of variable or fixed length, class-dependent background distributions, sequence elements defined as position weight matrices or list of k-mers with associated probabilities, and interactions between sequence elements with associated order or spacing constraints.

Moreover, the controlled environment of data simulation and reproducible model training, serving, and evaluation makes seqgra a suitable testbed for feature attribution and interpretability methods and their interdependencies with neural network architectures and the complexity level of the training data. The framework can even be used to perform extensive comparisons between deep learning libraries, which are rarely done (see Supplementary Figures A-7 and A-8) or identify undocumented behavior of the deep learning technology stack, such as an unusual training instability caused by a random seed of zero on some grammar-architecture combinations, which is reproducible and occurs in both PyTorch and TensorFlow (see Supplementary Figures A-4-A-6).

To avoid confusion, we would like to point out that seqgra is not a neural architecture search technique in the sense that it will not propose suitable neural network architectures for a particular data set. The model definition is an input, not an output of the seqgra pipeline. However, seqgra can be used in conjunction with neural architecture search, such as AMBER [30], a neural architecture search method for architectures aimed at genomics prediction tasks, or general hyperparameter optimization methods, such as Hyperband [16].

One caveat of all simulation-based approaches is the inevitable gap between simulated and real-world data sets, in the sense that the former is always a simplified approximation of the latter. Thus insights gained from simulated data might not carry over to the experimental world. In fact, to a certain degree this will always be the case. However, while high-performing neural network architectures on simulated data might not perform as highly on experimental data, the opposite is rarely the case, i.e., low-performing architectures in simulation are unlikely to improve when trained on noisier and/or smaller experimental data sets.

While the intricacies of noisy and biased high-throughput genomics experiments make for highly complex and poorly understood data sets, training highly complex alchemy-like [12] deep neural networks on them contributes little to a mechanistic understanding of the biological processes that are at work underneath and might worsen the reproducibility crisis in both machine learning [13] and biology [4, 3]. Simulated data, however, is perfectly understood, its noise levels controlled and any biases artificially introduced and accounted for, which makes it an excellent environment for model evaluation. With seqgra, the clean room of simulated data and a precise description of the patterns in the data (i.e., the probabilistic rules in the data definition) on the one end is paired with an array of feature attribution methods on the other, to answer questions that are often impossible to answer with poorly understood genomics data. One such question is whether the predictions of the model are based on those parts of the input that are in fact relevant for the phenomenon that is predicted, or, to put it another way, whether the model was able to recover the underlying rules of the data set.

Availability

The source code of the seqgra package is hosted on GitHub (`https://github.com/gifford-lab/seqgra`) and licensed under the MIT license. seqgra is part of the Python Package Index PyPI and can be installed using pip, the Python package installer. Extensive documentation can be found at `https://kkrismer.github.io/seqgra`.

Bibliography

[1] Z. Avsec, R. Kreuzhuber, J. Israeli, N. Xu, J. Cheng, A. Shrikumar, A. Banerjee, D. S. Kim, T. Beier, L. Urban, A. Kundaje, O. Stegle, and J. Gagneur. The Kipoi repository accelerates community exchange and reuse of predictive models for genomics. *Nat Biotechnol*, 37(6):592–600, 06 2019.

[2] David Baehrens, Timon Schroeter, Stefan Harmeling, Motoaki Kawanabe, Katja Hansen, and Klaus-Robert Müller. How to explain individual classification decisions. *Journal of Machine Learning Research*, 11:1803–1831, 08 2010.

[3] M. Baker. 1,500 scientists lift the lid on reproducibility. *Nature*, 533(7604):452–454, 05 2016.

[4] C. G. Begley and L. M. Ellis. Drug development: Raise standards for preclinical cancer research. *Nature*, 483(7391):531–533, Mar 2012.

[5] A. P. Boyle, S. Davis, H. P. Shulha, P. Meltzer, E. H. Margulies, Z. Weng, T. S. Furey, and G. E. Crawford. High-resolution mapping and characterization of open chromatin across the genome. *Cell*, 132(2):311–322, Jan 2008.

[6] J. D. Buenrostro, P. G. Giresi, L. C. Zaba, H. Y. Chang, and W. J. Greenleaf. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat Methods*, 10(12):1213–1218, Dec 2013.

[7] Brandon Carter, Jonas Mueller, Siddhartha Jain, and David Gifford. What made you do this? understanding black-box decisions with sufficient input subsets. *Proceedings of Machine Learning Research*, 89:567–576, 16–18 Apr 2019.

[8] K. M. Chen, E. M. Cofer, J. Zhou, and O. G. Troyanskaya. Selene: a PyTorch-based deep learning library for sequence data. *Nat Methods*, 16(4):315–318, 04 2019.

[9] J. L. Chew, Y. H. Loh, W. Zhang, X. Chen, W. L. Tam, L. S. Yeap, P. Li, Y. S. Ang, B. Lim, P. Robson, and H. H. Ng. Reciprocal transcriptional regulation of Pou5f1 and Sox2 via the Oct4/Sox2 complex in embryonic stem cells. *Mol Cell Biol*, 25(14):6031–6046, Jul 2005.

[10] Y. Guo, S. Mahony, and D. K. Gifford. High resolution genome wide binding event finding and motif discovery reveals transcription factor spatial binding constraints. *PLoS Comput Biol*, 8(8):e1002638, 2012.

[11] S. Heinz, C. Benner, N. Spann, E. Bertolino, Y. C. Lin, P. Laslo, J. X. Cheng, C. Murre, H. Singh, and C. K. Glass. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol Cell*, 38(4):576–589, May 2010.

[12] M. Hutson. Ai researchers allege that machine learning is alchemy. *Science*, 2018.

[13] Matthew Hutson. Artificial intelligence faces reproducibility crisis. *Science*, 359(6377):725–726, 2018.

[14] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research - ICML'15*, pages 448–456. PMLR, 07 2015.

[15] D. R. Kelley, J. Snoek, and J. L. Rinn. Basset: learning the regulatory code of the accessible genome with deep convolutional neural networks. *Genome Res.*, 26(7):990–999, 07 2016.

[16] Lisha Li, Kevin G. Jamieson, Giulia DeSalvo, Afshin Rostamizadeh, and Ameet Talwalkar. Efficient hyperparameter optimization and infinitely many armed bandits. *CoRR*, abs/1603.06560, 2016.

[17] Peter Lillian, Richard Meyes, and Tobias Meisen. Ablation of a robot's brain: Neural networks under a knife. *CoRR*, abs/1812.05687, 2018.

[18] S. Mei, Q. Qin, Q. Wu, H. Sun, R. Zheng, C. Zang, M. Zhu, J. Wu, X. Shi, L. Taing, T. Liu, M. Brown, C. A. Meyer, and X. S. Liu. Cistrome Data Browser: a data portal for ChIP-Seq and chromatin accessibility data in human and mouse. *Nucleic Acids Res*, 45(D1):D658–D662, 01 2017.

[19] Richard Meyes, Melanie Lu, Constantin Waubert de Puiseau, and Tobias Meisen. Ablation studies in artificial neural networks. *CoRR*, abs/1901.08644, 2019.

[20] S. Nair, D. S. Kim, J. Perricone, and A. Kundaje. Integrating regulatory DNA sequence and gene expression to predict genome-wide chromatin accessibility across cellular contexts. *Bioinformatics*, 35(14):i108–i116, 07 2019.

[21] A. Piovesan, M. C. Pelleri, F. Antonaros, P. Strippoli, M. Caracausi, and L. Vitale. On the length, weight and GC content of the human genome. *BMC Res Notes*, 12(1):106, Feb 2019.

[22] D. Quang and X. Xie. DanQ: a hybrid convolutional and recurrent deep neural network for quantifying the function of DNA sequences. *Nucleic Acids Res.*, 44(11):e107, 06 2016.

[23] Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Learning important features through propagating activation differences. *PMLR*, 70:3145–3153, 2017.

[24] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. In Yoshua Bengio and Yann LeCun, editors, *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Workshop Track Proceedings*, 2014.

[25] Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin A. Riedmiller. Striving for simplicity: The all convolutional net. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Workshop Track Proceedings*, 2015.

[26] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.*, 15(1):1929–1958, 1 2014.

[27] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pages 3319–3328. PMLR, 2017.

[28] Yulong Wang. Pytorch-visual-attribution. `https://github.com/yulongwang12/visual-attribution`, 2018.

[29] Matthew D. Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In David J. Fleet, Tomás Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I*, volume 8689 of *Lecture Notes in Computer Science*, pages 818–833. Springer, 2014.

[30] Z. Zhang, C. Y. Park, C. L. Theesfeld, and O. G. Troyanskaya. An automated framework for efficiently designing deep convolutional neural networks in genomics. *Nat Mach Intell*, 3:392–400, 2021.

[31] J. Zhou and O. G. Troyanskaya. Predicting effects of noncoding variants with deep learning-based sequence model. *Nat. Methods*, 12(10):931–934, Oct 2015.

Chapter 3

Identification of determinants of differential chromatin accessibility through a massively parallel genome-integrated reporter assay

The contents of this chapter were published as:

Identification of determinants of differential chromatin accessibility through a massively parallel genome-integrated reporter assay

Jennifer Hammelman, Konstantin Krismer, Budhaditya Banerjee, David K. Gifford, and Richard I. Sherwood

Genome Research, Volume 30, Issue 10, October 2020, pp 1468-1480

DOI: 10.1101/gr.263228.120 | PMID: 32973041

Supplementary information can be found in Appendix B.

3.1 Abstract

A key mechanism in cellular regulation is the ability of the transcriptional machinery to physically access DNA. Transcription factors interact with DNA to alter the accessibility of chromatin, which enables changes to gene expression during development or disease or as a response to environmental stimuli. However, the regulation of DNA accessibility via the recruitment of transcription factors is difficult to study in the context of the native genome because every genomic site is distinct in multiple ways. Here we introduce the multiplexed integrated accessibility assay (MIAA), an assay that measures chromatin accessibility of synthetic oligonucleotide sequence libraries integrated into a controlled genomic context with low native accessibility. We apply MIAA to measure the effects of sequence motifs on cell type-specific accessibility between mouse embryonic stem cells and embryonic stem cell-derived definitive endoderm cells, screening 7905 distinct DNA sequences. MIAA recapitulates differential accessibility patterns of 100-nt sequences derived from natively differential genomic regions, identifying E-box motifs common to epithelial-mesenchymal transition driver transcription factors in stem cell-specific accessible regions that become repressed in endoderm. We show that a single binding motif for a key regulatory transcription factor is sufficient to open chromatin, and classify sets of stem cell-specific, endoderm-specific, and shared accessibility-modifying transcription factor motifs. We also show that overexpression of two definitive endoderm transcription factors, *T* and *Foxa2*, results in changes to accessibility in DNA sequences containing their respective DNA-binding motifs and identify preferential motif arrangements that influence accessibility.

3.2 Introduction

Genomic DNA acts as an instruction book for the cellular machinery to carry out functional processes such as RNA production [44, 33] and DNA repair [4]. Some regions of the genome are constitutively used across all cell types for shared housekeeping processes [5, 24], whereas other regions are required only in specific cell types [57, 26]. One key mechanism used to control which regulatory regions are active is the physical accessibility of chromatin. Because many transcription factors are incapable of binding in inaccessible or "closed" chromatin, the regulation of chromatin accessibility ensures such transcription factors do not bind to extraneous or deleterious locations in the genome.

Transcription factors that interact with closed chromatin are thought to establish the accessibility of cell type-specific regions and initiate cell state change in differentiation [44, 50], cancer [9, 10], and environmental responses [41, 27] and allow "settler" transcription factors to bind and activate previously inactive genes. Massively parallel reporter assays (MPRAs) [20, 58] have been developed to measure the change to gene expression from the action of promoters [34, 14] or enhancers [32, 40, 23, 48, 29, 28] and thus can be used to probe the regulatory code. MPRAs allow for studies into the combinatorial logic of transcription factor action, such as whether specific combinations of transcription factor binding sites must be colocalized for proper gene expression [48, 12, 59]. However, MPRAs do not measure changes to chromatin accessibility and thus cannot disentangle gene regulation by transcription factors that depend upon changes in local accessibility.

Previous work has indicated specific transcription factor motifs and logic governing chromatin accessibility [30, 56, 6], but such effects are difficult to study in a native genomic context, in which motifs are not independent of nonlocal sequence effects. Recent approaches have extended MPRAs to measure nucleosome occupancy via bisulfite treatment [25] or MNase-seq [61] in yeast. However, bisulfite sequencing requires constrained library design to ensure sufficient CpG sites that act as a substrate for bisulfite conversion, and MNase-seq requires measurement over multiple

MNase concentrations to fully measure accessibility [42]. Restriction enzyme strategies have been used to measure nucleosome occupancy and accessibility in yeast [38] and mouse hepatocyte [7] and stem cells [49], and recently, adenine methyltransferase has been used to map nucleosome positioning in human cell lines [1, 52]. Here, we aim to develop an assay that takes advantage of adenine methyltransferase and restriction enzyme digestion for measuring the local DNA accessibility of genomically integrated large-scale reporter libraries, and probe the regulatory sequence determinants driving differential chromatin accessibility between stem cells and definitive endoderm.

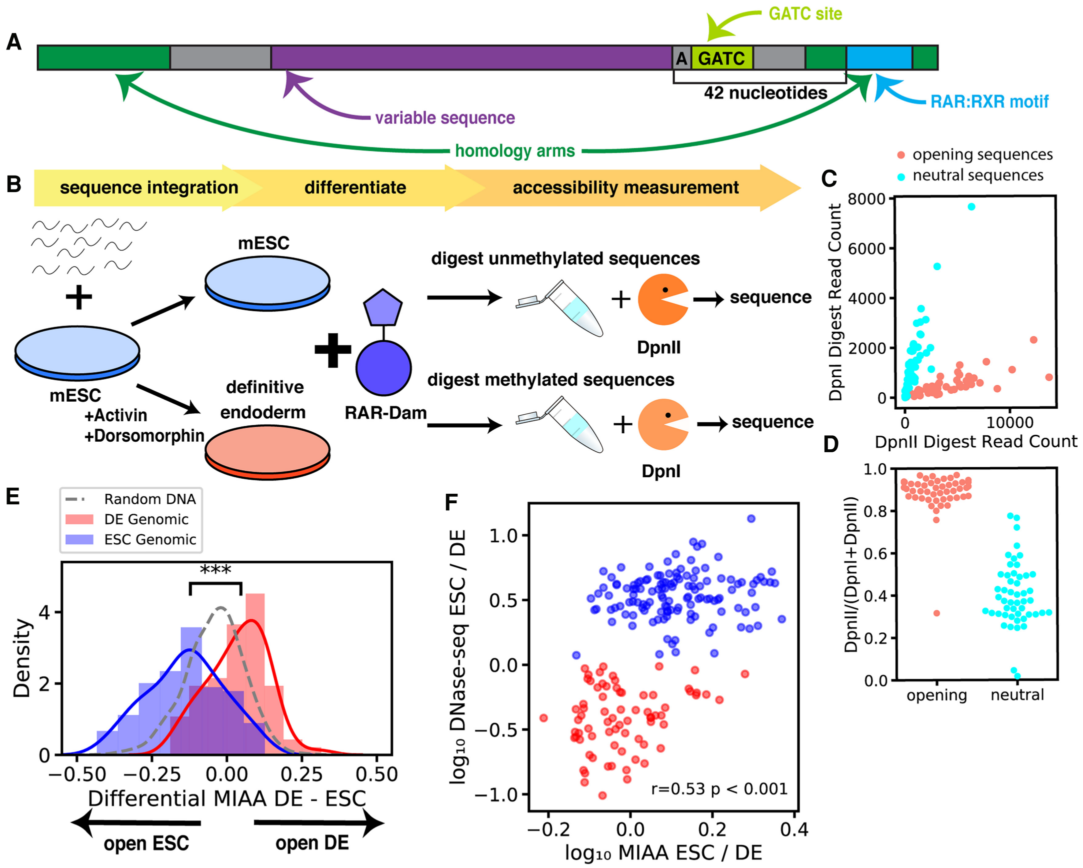


Figure 3-1: Multiplexed integrated accessibility assay (MIAA) measures local DNA accessibility of synthesized oligonucleotide DNA sequence libraries. (**A**) The MIAA library sequence construct contains a variable DNA sequence, homology arms for CRISPR-mediated HDR integration at a specific genomic locus that includes a binding site for retinoic acid receptor 42 nt downstream from the variable DNA sequence, and GATC site for DNA adenine methylase (Dam) methylation 1 nt downstream from the variable DNA sequence. (**B**) DNA sequences of 150 nt are integrated into ESCs at a designated genomic locus. ESCs are split, and half are differentiated into DE cells. Retinoic acid receptor fused to hyperactivated Dam enzyme results in methylation of DNA sequences that open DNA. DNA is extracted, and half is exposed to DpnII, which cleaves unmethylated sequences, whereas half is exposed to DpnI, which cleaves methylated sequences. Sequences are PCR amplified and sequenced. (figure caption continued on next page)

Figure 3-1: Multiplexed integrated accessibility assay (MIAA) measures local DNA accessibility of synthesized oligonucleotide DNA sequence libraries. (**C**) DpnI and DpnII read counts measured from a single DE replicate show difference between designed chromatin opening and neutral DNA sequences. (**D**) Proportion of DpnII read counts measured from a single ESC replicate gives estimate of MIAA openness. (**E**) Genomic sequences are differentially DE accessible or ESC accessible as reported by difference between MIAA Dpn proportion in definitive endoderm compared with ESCs with randomly shuffled DNA control sequences (significance computed by Wilcoxon rank-sum test). (**F**) Differential accessibility as measured by log change in normalized DNase-seq reads and MIAA methylation proportion shows correlation between native differential accessibility and MIAA accessibility. The correlation reported is the Pearson's correlation coefficient (r).

3.3 Materials and Methods

3.3.1 DNA sequence library design

All oligonucleotide libraries were ordered from Twist Biosciences. Variable DNA sequences (70-100 nt depending on library) are flanked by 25-nt primer sequences containing a GATC site and homology arms for CRISPR integration. We identified six native genomic sequences of size 100 nt from a pilot experiment that did not drive differential accessibility with MIAA but varied in GC-content. We randomly perturbed these native sequences three times each to obtain a total of 24 neutral sequence backgrounds. For our first experiment, we took each background and inserted either one motif seven times (positions 2, 16, 30, 44, 58, 72, 86) or two motifs in which motif 1 is inserted four times (positions 2, 30, 58, 86) and motif 2 is inserted three times (positions 16, 44, 72). For our second experiment, we limited ourselves to nine backgrounds that we expected to have high reproducibility to the set of 24. In this experiment, we tested sequences of size 70 nt. By using the consensus sequences of known ES key TFs (POU5F1, SOX2, KLF4) or DE key TFs (FOXA2, GATA4, SOX17), we inserted one, two, or three motifs into each sequence. We tested homotypic DNA sequences consisting of one unique motif, as well as heterotypic DNA sequences enumerating all possible motif orders. Consensus motifs for key developmental transcription factors are listed in Supplemental Table S3. Additional hypotheses were tested within MIAA libraries that were not described in this paper. The DNA sequences that were used in this paper are denoted by a column within the Supplemental Data.

3.3.2 DNA sequence library integration

Electroporations were performed in two to four biological replicates into p2L RAR-DamA126 ESCs (for cell line construction and RARg-DamN126A-V5His construct sequence, see Supplemental Methods). Cells were grown for 5-8 d after electroporation to obtain adequate quantities for doxycycline treatment. When indicated, cells were differentiated to DE before doxycycline treatment.

3.3.3 High-throughput sequencing

After DpnI/II digestion, fragments are amplified with three steps of PCR. First, PCR primers to sequence outside the homology arms such that only sequences that are properly integrated at the desired locus and that have not been cleaved by the DpnI/II enzyme are amplified (13 cycles). The second PCR step and third PCR steps further amplify sequences and add adaptors for Illumina sequencing. For primer information and further details, see Supplemental Methods. Samples were sequenced on an Illumina NextSeq 550 instrument at the Harvard Medical School Biopolymers Facility or the MIT BioMicro Center.

3.3.4 DNA sequence library processing

Reads were mapped to library DNA sequences by taking the reverse complement to the raw read, in which the first N nucleotides (between 70 and 100 based on the size of the designed sequence) are the designed variable DNA sequence. Perfect matches were counted using a custom R script (Supplemental Code). Reads were normalized to reads per million over the total number of reads in the digest. DNA sequences were kept if they had a threshold number of total normalized reads over all replicates, based on the observation of high standard deviation at low total read counts. The threshold was selected based on visual inspection and can be found in the Supplemental Code. Once reads were normalized and high variability DNA sequences filtered, MIAA accessibility was computed as a proportion of DpnI/II read counts DpnII/(DpnI + DpnII).

3.3.5 DeepAccess model and motif importance

We obtain DNase-seq regions using the 100 nt centered at the MACS2 narrow peak call. Accessibility prediction is treated as a multitask classification problem, in which each genomic sequence (100 bp) is associated with a two-dimensional bit vector representing whether the sequence is open in each cell type (ESC and DE cell). We trained an ensemble of 10 convolutional neural networks. For specific details on network architecture, see Supplemental Methods. The fully connected output layer present in all neural network architectures contains two neurons with a sigmoid activation function that returns a value between zero and one, which represents the probability of the predicted DNA "openness" in each of the two cell types. DeepAccess is trained on a balanced data set with 400,000 sequences across four possible classification scenarios of a sequence (1) open in endoderm cells and closed in ESCs, (2) open in ESCs and closed in endoderm cells, (3) open in both cell types, or (4) closed in both cell types. A test set of 22,357 sequences is held out for performance evaluation.

We extracted motifs from DeepAccess by applying smoothed gradient ascent to score each nucleotide in the 100-nt DNA sequence by its importance for predicting the output [46, 47] and multiplied times the input (a one-hot encoding of the DNA sequence) because gradients will assign nonzero values to DNA characters not present in the sequence. To obtain sequence importance for features that drive accessibility differentially between DE cells and ESCs, we set the gradient loss to the difference between the predicted accessibility of two cell types. We then selected windows of size 10 with the highest ensemble weighted average saliency over a set of 5000 training sequences and used those as the DeepAccess-derived motifs. We also extracted the top motifs with the highest increase in saliency of differential accessibility between the CNN without trainable hidden layers and the CNNs with hidden layers, which represent motifs that gain importance from the CNNs that learn relationships between motifs.

3.4 Results

3.4.1 Multiplexed integrated accessibility assay measures local accessibility of integrated DNA sequences

In previous work, we used a DNase I cleavage assay, SLOT, to measure chromatin accessibility of a set of DNA sequences integrated into a defined genomic locus [18]. Although SLOT was able to determine the relative accessibility of classes of DNA sequences, it had poor resolution to measure accessibility of individual DNA sequences, because of the low cleavage probability of DNase I at enzyme concentrations capable of discriminating levels of chromatin accessibility. We hypothesized that we could measure changes in DNA accessibility with higher sensitivity by observing the chromatin accessibility-dependent methylation of *Escherichia coli* adenine DNA adenine methylase (Dam) to the locus, given the high efficiency and stability of Dam methylation in cells [55] and the known propensity of Dam to methylate more frequently in accessible chromatin [55, 54, 1, 52]. We further hypothesized that fusing Dam to retinoic acid receptor-gamma (RAR) would enhance the differential methylation of this RAR-Dam fusion protein at genomic loci with RAR binding motifs, and we make use of a mutant version of Dam methyltransferase shown to display increased signal-to-noise over wild-type Dam [55, 54].

We designed a library consisting of 150-nt synthesized oligonucleotides that consist of a 100-nt variable DNA sequence surrounded by a fixed sequence that allows for PCR amplification and contains an Illumina sequencing adapter and a Dam recognition sequence (GATC) (Fig. 3-1A). For integration, we chose a genomic locus with minimal prior DNase I accessibility proximal to a RAR binding site. To allow inducible expression of RAR-Dam, we integrated a single copy of RAR-Dam with a doxycycline-sensitive promoter into a fixed genomic locus using Cre/LoxP recombination into a mouse embryonic stem cell (mESC) line with constitutive rtTA expression [31].

After DNA sequence integration into the mESC cell line, we induce the expression of RAR-Dam and, after 24 h, collect genomic DNA (Fig. 3-1B). DNA sequences that increase chromatin accessibility should increase adenine methylation of the DNA sequence's GATC site, owing to the combined effect of the preference of Dam methylase to methylate in accessible chromatin, and increased local RAR binding, owing to increased chromatin accessibility. Purified genomic DNA is split it into two pools; one pool is exposed to the restriction enzyme DpnI and the other pool to DpnII, which preferentially cleave methylated and unmethylated GATC sites, respectively. From each pool, we then amplify DNA sequences using a three-step PCR amplification process (Supplemental Fig. B-1). First, DNA sequences are amplified by primers outside of the homology arms to ensure only correctly integrated DNA sequences are amplified. Only undigested DNA sequences will be amplified at this step owing to the site of the GATC site of restriction enzyme cleavage between the PCR primers. Then, two additional PCR steps are used to further amplify DNA sequences and add Illumina sequencing adapters for high-throughput sequencing. If a DNA sequence is more accessible, it will have fewer read counts in the DpnI digested pool and more read counts in the DpnII digested pool (Fig. 3-1C). The proportion of DpnII to DpnI sequencing counts, therefore, represents the impact of that DNA sequence on local DNA accessibility (Fig. 3-1D). We designate this high-throughput genomically integrated assay of chromatin accessibility the multiplexed integrated accessibility assay (MIAA).

Because our particular interest is in changes to accessibility during differentiation, we differentiated mESCs into definitive endoderm (DE) cells using a well-established differentiation protocol shown to yield >90% DE [45] before RAR-Dam induction.

We tested a library of 5978 DNA sequences in eight biological replicates (four replicates at sequence integration, each split into two replicates before RAR-Dam activation) for stem cells (ESCs) and four biological replicates (two replicates at sequence integration, each split into two replicates before RAR-Dam activation) for DE cells. To gauge the reliability of MIAA, we included sets of positive and negative control DNA sequences used in our previous work that maximally pack 100-nt variable sequences with DNA sequence motifs shown to have an opening or neutral effect on chromatin by a k-mer model trained on DNase-seq [18]. From MIAA measurements, we found that the Hashimoto et al. positive control DNA sequences yielded significantly higher Dam methylation than the negative control DNA sequences (Fig. 3-1C,D), with 81%-99% of positive control DNA sequences yielding higher methylation than the average negative control DNA sequence in each replicate ($P < 0.001$ by Wilcoxon rank-sum test for all replicates). We found in comparing control sequences with GC-content in the range of 30%-50%, MIAA replicates had 96%-100 of positive control DNA sequences yielding higher methylation than the average negative control DNA sequence, whereas SLOT had 4.5%-13.6% of positive control DNA sequences yielding higher methylation than the average negative control sequence (Supplemental Fig. S2), suggesting that MIAA provides a marked improvement over SLOT in the measurement of accessibility differences of single DNA sequences in the context of large libraries. Biological replicates of MIAA were also well correlated (Pearson's $r = 0.5$-0.79) (Supplemental Fig. S3).

We note that negative control (accessibility neutral) DNA sequences are still methylated at a rate of 20%-50%. In line with this result, we found $\approx 20\%$ RAR-Dam methylation in two known native genomic inaccessible chromatin loci as measured by qPCR, compared with 85%-95% methylation at known RAR binding sites (Supplemental Fig. S2). We do not know if this means that RAR-Dam can methylate $\approx 20\%$ of inaccessible chromatin while it is tightly wound or if the methylation is happening during cell cycle phases when chromatin is accessible. We also found that retinoic acid binding sites within our sequence appeared to have no impact on MIAA results (Supplemental Fig. S4), suggesting that linking RAR to Dam is unlikely to confound our aim of measuring chromatin accessibility.

We separately designed a pilot experiment of 2000 DNA sequences to determine whether MIAA could measure differential chromatin accessibility. First, we ran KMAC, a method for de novo motif enrichment [17], on differentially accessible DNase-seq regions using the top 10,000 peaks that were differentially accessible (defined by peak overlap) in DE-accessible or ESC-accessible genomic regions measuring motif enrichment relative to a background the top 10,000 of genomic regions that are DNase accessible in both ESCs and DE.We used a similar methodology to Hashimoto et al. (2016) to maximally pack oligonucleotides with DNA sequence motifs, by starting from a single motif and extending the designed sequence with the highest scoring KMAC motif that overlapped the previous motif by four bases. Our data show that that MIAAwas able to separate DNA sequences that were designed to open chromatin in DE cells from those that were designed to open chromatin in ESCs (Supplemental Fig. S5).

We then asked whether MIAA could measure differential accessibility of native genomic sequences. To help identify 100-nt native genomic sequences that were differentially accessible between DE cells and ESCs, we developed a deep learning model trained to predict DNase-accessible regions from underlying DNA sequence and cell type-specific DNase-seq training data. This method, which we call DeepAccess, trains an ensemble of 10 convolutional neural networks on DNase-seq data from ESCs and DE cells to predict whether a 100-nt genomic region is accessible or inaccessible in both cell types that had good performance on held-out genomic regions (for details, see Methods; Supplemental Fig. S6). We tested 213 native DNA sequences that DeepAccess predicted would be differentially accessible between ESCs and DE cells with MIAA, and found that as a group these DNA sequences showed differential accessibility between ESCs and DE cells (Fig. 3-1E) with a per-sequence effect size that correlates with differential accessibility measured by DNase-seq (Pearson's $r = 0.53$; $P < 0.001$) (Fig. 3-1F). Although statistically significant as a group, only 78% of the native genomic DNA sequences recapitulated the differential accessibility of the native loci from which they were derived by having both higher DNase-seq read counts and greater MIAA-measured accessibility in one cell type over the other. These

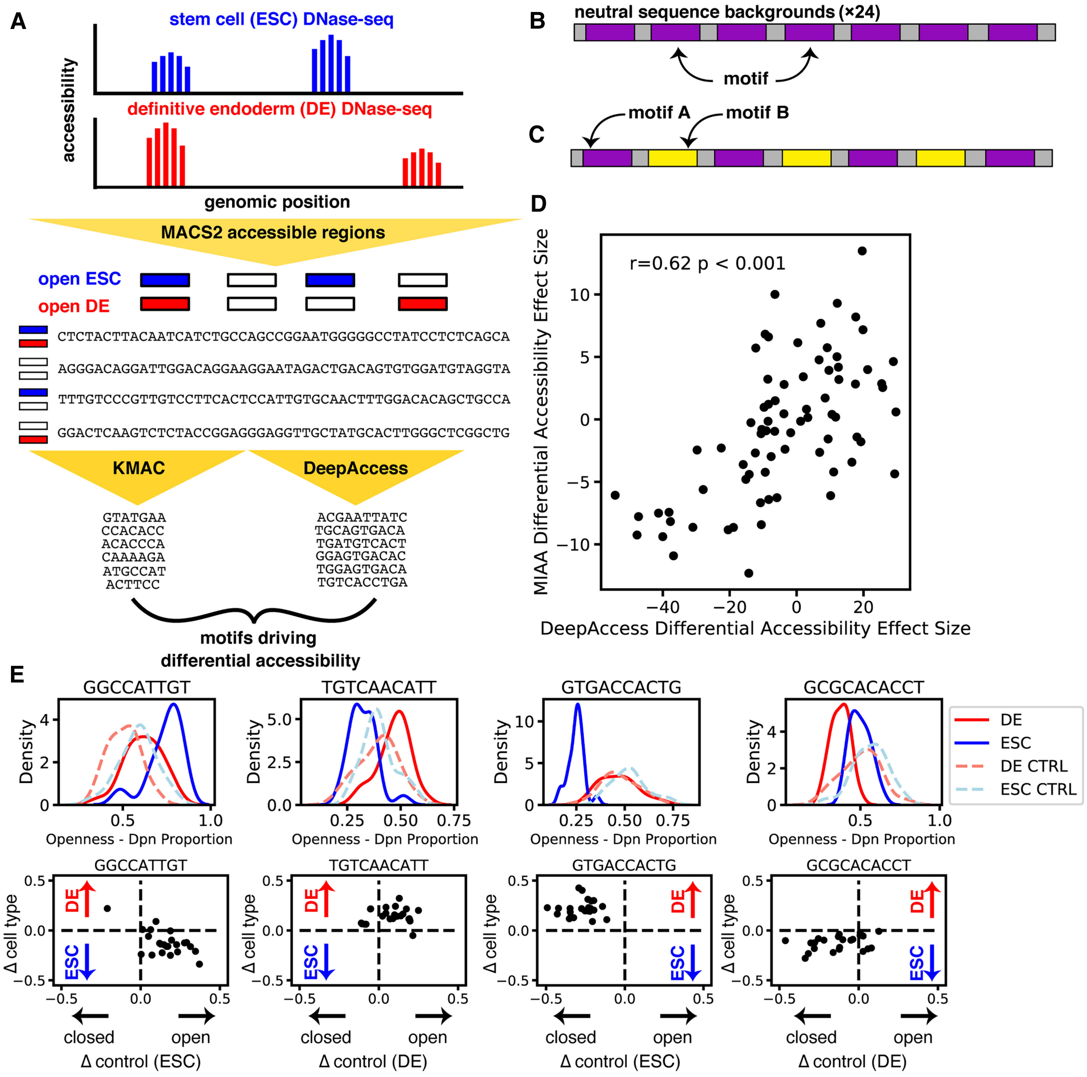


Figure 3-2: Differentially accessible motif generation from DNase-seq data validated by MIAA. (**A**) DNase-seq accessible regions called with MACS2 and 100-nt sequences extracted centered at narrow peak. KMAC and DeepAccess were applied to extract significant motifs potentially driving differential accessibility between ESCs and endoderm. (**B**) DNA sequences were designed using seven instances of each motif at the same locations in each DNA sequence inserted into 24 100-nt neutral sequence backgrounds, as well as pairs of motifs (**C**). (**D**) Predictions from DeepAccess for differential accessibility replicate experimental results (effect size by paired t-test between ESC and DE measurements). The correlation reported is the Pearson correlation coefficient (r). (figure caption continued on next page)

Figure 3-2: Differentially accessible motif generation from DNase-seq data validated by MIAA. (**E**) Motif sequences show differential accessibility via opening ESC, opening endoderm, closing ESC, and closing endoderm (left to right). (Top row) Distribution of MIAA-measured accessibility in ESCs and DE cells for KMAC- or DeepAccess-generated motif, tested over 24 neutral sequence backgrounds and randomly shuffled DNA controls (CTRL). (Bottom row) Measurements for a particular DeepAccess or KMAC motif. Each dot represents a single neutral background. The y-axis is the difference between endoderm and ESC accessibility, and the x-axis is the difference between each DNA sequence and its shuffled control. The cell type in which control measurement is made is in parentheses.

100-nt endogenous sequences were selected by DeepAccess from DNase-seq accessible regions that can be kilobases in length, so we hypothesize that sequences for which we did not observe differential accessibility may not contain all of the binding elements controlling accessibility of the native locus or may rely on either local or distal interactions with chromatin that were not recapitulated at our genomic integration site. The observed correlation in differential accessibility between DNase-seq and MIAA suggests that a 100-bp sequence transplanted into a specified locus can retain a substantial amount of the information required to encode a particular level of chromatin accessibility (Fig. 3-1F).

We also included in our library a randomly shuffled nucleotide counterpart for each DNA sequence in order to account for any potential effects of nucleotide composition. We found that most native genomic sequences that were more accessible in ESCs than in DE cells had similar accessibility in ESCs compared with randomly shuffled DNA controls but had lower accessibility in endoderm compared with shuffled control DNA sequences (Supplemental Fig. S7). We hypothesized that these DNA sequences contain motifs that result in decreases in accessibility in DE cells. We performed motif enrichment (for details, see Supplemental Methods) on these DNA sequences and found that 98% (compared with 0% of endoderm native sequences) contained a match to the ZEB2 motif (Supplemental Fig. S7), a known transcriptional repressor that has been implicated in early gastrulation by repression of CDH1 (also known as E-cadherin) [2], suggesting that the DeepAccess-selected ESC sequences were selected based on an endoderm-specific repressor of chromatin accessibility. In contrast, none of our DeepAccess-selected native genomic sequences contained motifs for the known ESC reprogramming factors POU5F1, SOX2, or KLF4 [50], which we would expect to increase chromatin accessibility in ESCs.

To investigate why DeepAccess chose ESC native genomic sequences that contain ZEB2 motifs over known reprogramming factors, we compared DeepAccess-predicted differential accessibility for ChIP-seq sites for the known pluripotency factors POU5F1, SOX2, and KLF4, which contained their DNA-binding motifs along with ZEB2 genomic motif instances, and found that although the knownpioneer transcription factor motifs had positive effects on ESC accessibility, ZEB2 motifs had the strongest predicted effect on differential accessibility by the presence of the motif causing a decrease in predicted accessibility in DE cells (Supplemental Fig. S7). ZEB2 binding sites were also enriched in ESC-specific genomic accessible regions with 14% containing a ZEB2 motif relative to 9% in endoderm-specific accessible regions ($P < 0.001$ by hypergeometric test). In comparison, 12% of genomic ESC-specific accessible regions contained a SOX2 motif, 6% contained a POU5F1 motif, and 6% contained a KLF4 motif. KEGG biological pathway analysis of ZEB2 motif sites in ESC-accessible regions showed an enrichment of motif sites proximal to genes regulating pluripotency of ESCs ($P < 0.001$), including the key pluripotency regulators KLF4, SOX2, and NANOG, a finding that is consistent with a model of ZEB2 repression of pluripotency during DE differentiation [53]. The ZEB2 motif is similar to motifs of other E-box epithelial-mesenchymal transition driver transcription factors such as ZEB1, SNAI1, SNAI2, and TWIST1 [51], all of which are expressed during ESC differentiation to endoderm. We note that subsequent MIAA libraries described in this paper show that DNA sequences containing POU5F1, SOX2, and KLF4 motifs do yield ESC-enriched accessibility. Overall, we find that 100-nt DNA sequences extracted from genomic regions with differential chromatin accessibility recapitulate this differential accessibility when transplanted to a fixed chromatin locus.

3.4.2 DNase-seq analysis identifies motifs driving cell type-specific accessibility

We then hypothesized that we could identify and confirm with MIAA motifs that control chromatin accessibility in a cell type-specific manner through a set of synthetic, designed DNA sequences. By using cell type-specific DNase-seq data, we extracted short (8- to 12-nt) DNA sequence motifs that we hypothesized would cause differential accessibility using two methods (Fig. 3-2A). First, we used the motifs that were derived from de novo motif discovery by running KMAC on ESC differentially accessible and DE differentially accessible genomic regions. Second, we used DeepAccess to obtain hypotheses about which motifs were most responsible for differential accessibility between DE cells and ESCs (for details, see Supplemental Methods). Unlike KMAC's pure enrichment approach, DeepAccess is able to learn nonlinear relationships between sequence motifs for predicting accessibility. From our set of motif hypotheses from both methods, we designed synthetic DNA sequences with either seven instances of one motif (Fig. 3-2B), which we call *motif sequences*, or two different motifs (Fig. 3-2C), which we call *motif pair sequences*, inserted into 24 fixed sequence backgrounds of varied GC-content. Fixed background sequences were previously measured to have a neutral impact on cell type-specific accessibility with MIAA (see Methods for details). We chose to pack each DNA sequence with the maximum number of motifs (54%-84% of the positions in each DNA sequence are part of a motif) while leaving space for sequence variation. For each DNA sequence, we also included a control in which the nucleotides are randomly shuffled to observe the influence of nucleotide content alone.

To determine whether DeepAccess was able to predict the effects of motif sequences or motif pair sequences, we compared the DeepAccess-predicted effect size of each motif or motif pair on differential accessibility to the equivalent MIAA measurement. We found that DeepAccess results are correlated (Pearson's $r = 0.62$; $P < 0.001$) with

MIAA-measured differential accessibility (Fig. 3-2D). However, we found that DeepAccess failed to perform well in predicting paired effects between DNA sequences and shuffled controls (ESC Pearson's $r = 0.24$; DE Pearson's $r = 0.42$) (Supplemental Fig. S8), which we hypothesize is the result of overconfidence of neural networks on out-of-distribution inputs [37, 35], because the network had not seen the shuffled control DNA sequences during training. We tested for statistically significant differential accessibility of our motifs and motif pairs by first performing paired tests between MIAA openness in ESCs and DE cells and then performing paired tests between DNA sequences and shuffled controls under a Benjamini-Hochberg multiple hypothesis correction at a false-discovery rate of 0.05 (for details, see Supplemental Methods). Out of 38 tested motif sequences, 20 induced differential accessibility, and out of 38 motif pair sequences, 26 induced differential accessibility. We also found these results to be largely consistent across a secondary closed integration locus (Supplemental Fig. S9). Thus, MIAA was able to confirm that motifs identified using DeepAccess are able to result in observable changes to accessibility both between cell types and compared with shuffled control sequences (Fig. 3-2E).

Out of the 46 motif or motif pair sequences that induced differential accessibility across cell types and were compared with shuffled control sequences as measured by MIAA, DeepAccess predicted the correct direction of differential accessibility between the two cell types in 76% (35/46) of cases (Supplemental Table S1). In comparing results from DeepAccess to KMAC, we found only 32% (8/25) of our KMAC motifs or motif pairs were differentially accessible compared with 74.5% (38/51) of DeepAccess (Supplemental Table S1), indicating our DeepAccess approach was successful in identifying motifs driving differential accessibility.

3.4.3 GC-content and transcription factor binding motifs control accessibility

We noticed previously that the positive control DNA sequences from the Hashimoto et al. (2016) library had higher GC-content than the negative control DNA sequences.

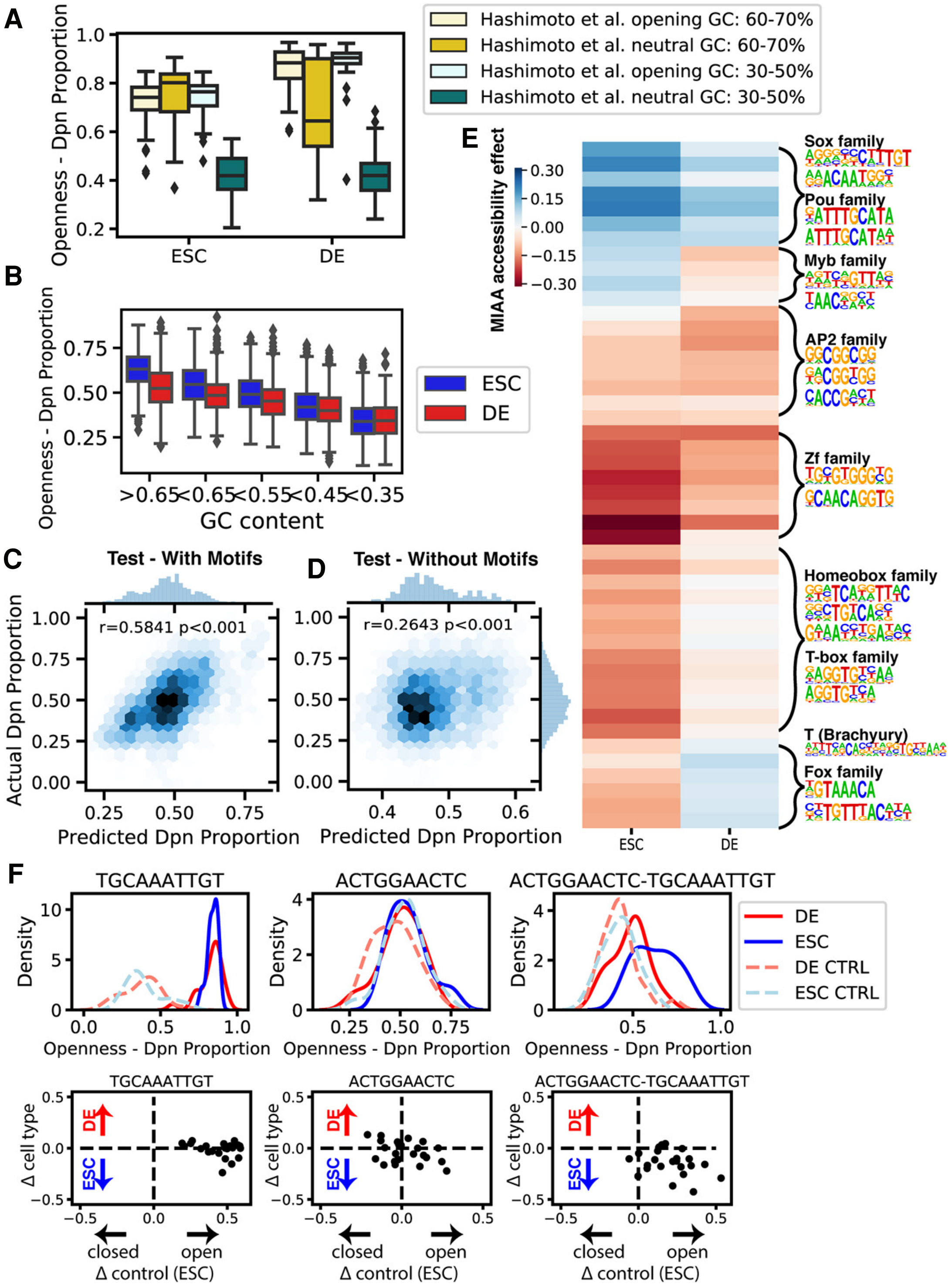


Figure 3-3: MIAA identifies global influence of GC-content and differentially accessible motifs. (figure caption continued on next page)

Figure 3-3: MIAA identifies global influence of GC-content and differentially accessible motifs. (**A**) GC-content observed to be correlated with accessibility in both stem and endoderm cells from positive (Hashimoto et al. opening) and negative (Hashimoto et al. neutral) control sequences. (**B**) GC-content correlated with accessibility in random DNA sequences. A regression model was trained on MIAA Dpn proportions with GC-content, replicate, and cell type-specific effects of 20 motifs and 26 motif pairs as features, and predicts well on (**C**) held-out test data (n = 4404) and performs significantly better than (**D**) a model trained without motif variables (adjusted R-squared motif model = 0.398; adjusted R-squared no motif model = 0.095). The correlation reported is the Pearson correlation coefficient (r). (**E**) Regression weights of individual motifs and motif pairs in stem and DE cells. Hierarchical clustering of regression weights followed by motif enrichment recovers clusters representing cell type-specific transcription factor DNA-binding motifs. (**F**) Example of individual motifs (left, middle) that alone do not result in differentially open chromatin but result in differentially open chromatin ESCs in combination (right). (Top row) Distribution of MIAA-measured accessibility in ESCs and DE cells for KMAC- or DeepAccess-generated motif, tested over 24 neutral sequence backgrounds and randomly shuffled DNA controls (CTRL). (Bottom row) Measurements for a particular DeepAccess or KMAC motif, in which each dot represents a single neutral background. The y-axis is the difference between endodermand ESC accessibility, and the x-axis is the difference between each DNA sequence and its shuffled control in ESCs.

To clarify the role of GC-content in driving accessibility, we selected a total of 200 positive and negative control DNA sequences from the Hashimoto et al. (2016) library, which were designed to include a string of motifs that were predicted by a model trained on DNase-seq to have a positive or neutral impact on accessibility [18]. We selected positive and negative controls with either high GC-content (60%-70%) or low GC-content (30%-50%). We found that in both cell types, positive control DNA sequences drove uniformly and equivalently high accessibility regardless of GC-content (Fig. 3-3A), suggesting that motifs associated with accessible regions can increase accessibility independently of GC-content. However, in endoderm, positive control DNA sequences for both GC-content bins had increased accessibility compared with negative control DNA sequences with matched GC-content ($P < 0.001$ by Wilcoxon rank-sum test), whereas in ESCs, only the low GC-content bin had differential accessibility between negative and positive controls ($P < 0.001$ by Wilcoxon rank-sum test) (Fig. 3-3A) because of high accessibility among high-GC neutral DNA sequences. GC-content was positively correlated with accessibility in both ESCs and DE cells among both sets of control DNA sequences (ESC Pearson's $r = 0.476$; DE Pearson's $r = 0.357$), suggesting that GC-content is a contributor to MIAA-measured accessibility alongside motif composition. DeepAccess-predicted accessibility was consistent with MIAA, indicating these effects were to be expected from observations on DNase-seq (Supplemental Fig. S10).

Because this result could be an effect of sequence motifs included in the high-GC-content negative control DNA sequences, we then examined the nucleotide-shuffled DNA sequences that we designed to act as controls for motif activity to see if the effect of GC-content on MIAA accessibility held in random DNA. We found that the GC-content of randomly shuffled sequences correlated with MIAA accessibility in both cell types (Fig. 3-3B). We also found that accessibility was significantly higher ($P <$ 0.001 by one-tailed Wilcoxon signed-rank test) in ESCs compared with endoderm cells across all GC-content bins, except in DNA sequences with <35% GC-content (N= 372). Altogether, these results indicate that GC-content alone is a sufficient DNA signal to drive accessibility in both ESCs and endoderm as measured by MIAA and also to drive accessibility differences between these two cell contexts through its heightened impact in ESCs.

Consistent with previous research that suggests a relationship between GC-rich regions and accessibility [39, 57, 42], we found that the top 5000 DE cell-specific regions and the top 5000 ESC-specific regions from DNase-seq have higher GC-content than randomly sampled DNase-inaccessible regions (Supplemental Fig. S10). We then set out to examine the impact each motif or motif pair sequence derived from our DeepAccess- and KMAC-derived hypotheses beyond the confounding effects of GC-content. We trained a linear regression model to predict MIAA Dpn ratios from GC-content, experimental replicate, and cell type-specific effects for all DNA sequences containing differential motifs or motif pairs. This linear model had good performance on training (Pearson's $r = 0.6335$) and held-out test data (Pearson's $r = 0.5841$) (Fig. 3-3C; for details, see Supplemental Methods) and significantly improved from regression models that did not include motif effects (adjusted R-squared motif model = 0.398; adjusted R-squared no motif model = 0.095) (Fig. 3-3D), reinforcing the salient effects of transcription factor binding motifs in controlling accessibility.

We next sought to determine which transcription factor binding motifs most strongly drove differential accessibility between ESCs and endoderm. Because KMAC and DeepAccess identified sequence motifs and motif pairs that could represent the same transcription factor binding site, we clustered the regression weights to identify clusters of motifs and motif pairs representing similar influences on MIAA-measured accessibility (Fig. 3-3E). We then ran motif discovery on the designed DNA sequences in each cluster to obtain transcription factor candidates (for details, see Supplemental Methods). We identified motifs for known transcription factors such as Pou and Sox motifs as ESC-enriched and motifs for T-box and Fox factors as enriched in DE cells. The regression weights for these differential accessibility-driving motifs were robust, showing high consistency between models trained on biological replicates (Pearson's $r = 0.963$) (Supplemental Fig. S11), indicating that although MIAA correlation at the level of individual DNA sequences is modest, our estimation of motif-level effects is highly reproducible. We also identified motif pair sequences that show interesting nonlinear activity with respect to differential accessibility compared to their motif sequence effects alone (Fig. 3-3F). In sum, MIAA data enable de novo discovery of features such as GC-content and transcription factor motifs that govern differential chromatin accessibility and validate predictions of motifs impacting differential chromatin accessibility made by DeepAccess.

3.4.4 Overexpression of DE transcription factors *T* and *Foxa2* increase accessibility of DNA sequences with their DNA-binding motifs

We then hypothesized we could connect our discovered motifs to transcription factors driving differential accessibility by ectopically expressing transcription factors known to bind to certain enriched motifs. We overexpressed the transcription factors *T* or *Foxa2* in ESCs and measured the accessibility of our DNA sequence library with MIAA (Fig. 3-4A). We trained a joint regression model to predict condition-specific accessibility with data from four conditions: ESCs, DE cells, ESCs with *Foxa2* overexpression, and ESCs with *T* overexpression (Supplemental Fig. S12; for details, see Supplemental Methods). We then selected the motifs with the greatest positive difference in regression weights between the overexpressed *T* (ESC + *T*) and the ESC conditions. We found that *T* overexpression increases MIAA accessibility most strongly in DNA sequences with a motif pair that partially matches the motif of a *T* homodimer with two motifs in a minus/plus orientation and is significantly enriched over other dimer orientations in *T* ChIPseq peaks ($P < 0.001$ by χ^2 test) (Supplemental Fig. S13). The second strongest motif is also significantly enriched in *T* binding in mouse DE as measured by ChIP-seq ($P < 0.05$ under Benjamini-Hochberg multiple hypothesis correction) (Fig. 3-4B). Overall, only 6/76 motifs or motif pairs showed a significant increase in ESC accessibility upon *T* overexpression (for details, see Supplemental Methods), supporting that *T* binding is capable of increasing accessibility specifically at motif-containing DNA sequences in a fixed chromatin context.

Similarly, we examined the motifs with the greatest increase in accessibility upon

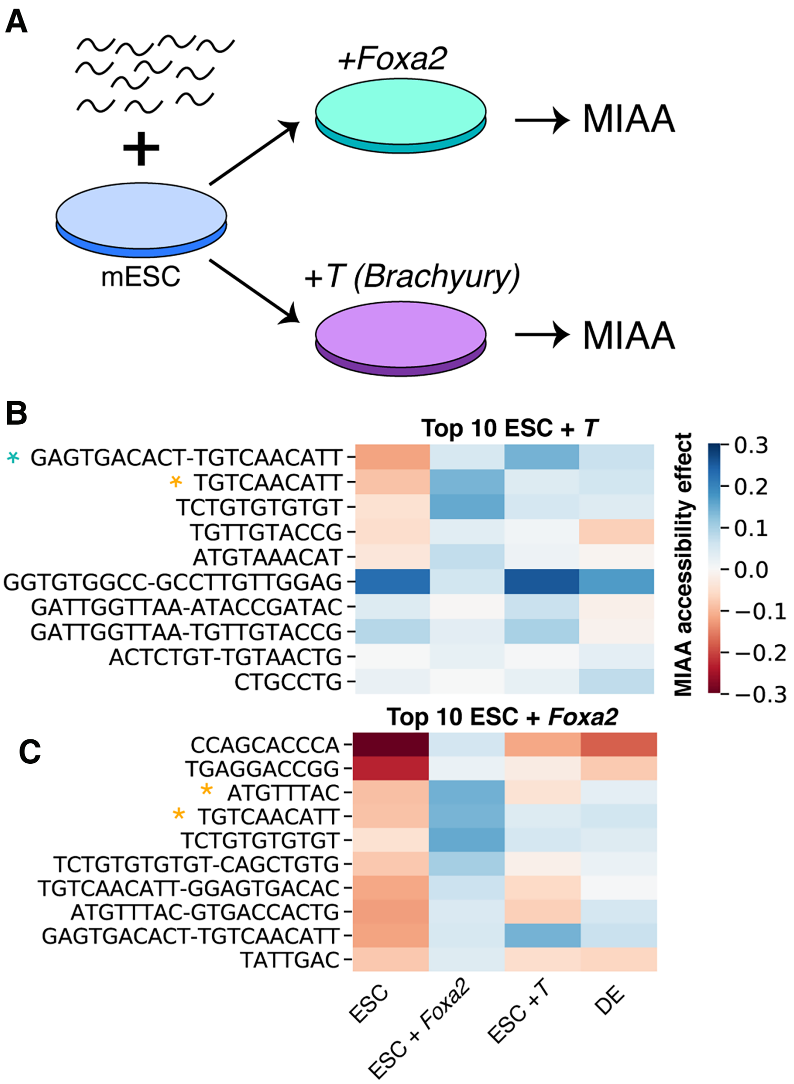


Figure 3-4: Overexpression of DE lineage-defining transcription factors results in changes to certain motifs representing DNA binding. (**A**) Synthetic DNA sequence library is integrated into ESCs, and *Foxa2* and *T* are overexpressed. (**B**) Regression weight heatmap of top motifs and motif pairs that increase accessibility under *T* overexpression compared with ESCs. Blue star indicates motif visually matches T homodimer in ± orientation that is enriched in ChIP-seq peaks. Yellow star indicates motif is statistically enriched in ChIP-seq peaks of T binding in mouse definitive endoderm cells ($P < 0.05$ HOMER motif enrichment with Benjamini-Hochberg correction). (figure caption continued on next page)

Figure 3-4: Overexpression of DE lineage-defining transcription factors results in changes to certain motifs representing DNA binding. (**C**) Regression weight heatmap of top motifs and motif pairs that increase accessibility under *Foxa2* overexpression compared to ESCs. Star indicates motif is statistically enriched in FOXA2 ChIP-seq peaks in mouse DE cells ($P < 0.05$ HOMER motif enrichment with Benjamini-Hochberg correction).

Foxa2 overexpression and found that the third and fourth top motifs were enriched in sequences from FOXA2 ChIP-seq peaks ($P < 0.05$ under Benjamini-Hochberg multiple hypothesis correction) (Fig. 3-4C). *Foxa2* overexpression results in more substantial changes in ESC motif accessibility profiles than *T* overexpression (Supplemental Fig. S14), which is consistent with data showing that *Foxa2* overexpression also results in more changes to gene expression (Supplemental Fig. S15), and therefore may lead to secondary chromatin accessibility changes unrelated to the FOXA2 motif. Both *T* and *Foxa2* overexpression resulted in increased accessibility at a TGTCAACATT motif, which is likely because it contains sequences capable of binding both factors and is consequently enriched in both T and FOXA2 ChIP-seq. We also found that both *Foxa2* and *T* overexpression resulted in chromatin accessibility changes that brought cells closer to the MIAA profile of DE cells (Supplemental Fig. S14). Thus, overexpression of individual transcription factors is capable of increasing the chromatin accessibility of a specific cohort of motif-containing sequences in a controlled chromatin context, providing evidence that binding of these factors leads to increased chromatin accessibility.

3.4.5 Exploration of ordering of ESC and endoderm key transcription factors uncovers subtle TF-TF interactions

Finally, we used MIAA to explore interactions between motifs that are difficult to measure from observational approaches such as DNase-seq because of the lack of suitably controlled genomic motif arrangements. To probe interaction effects over a

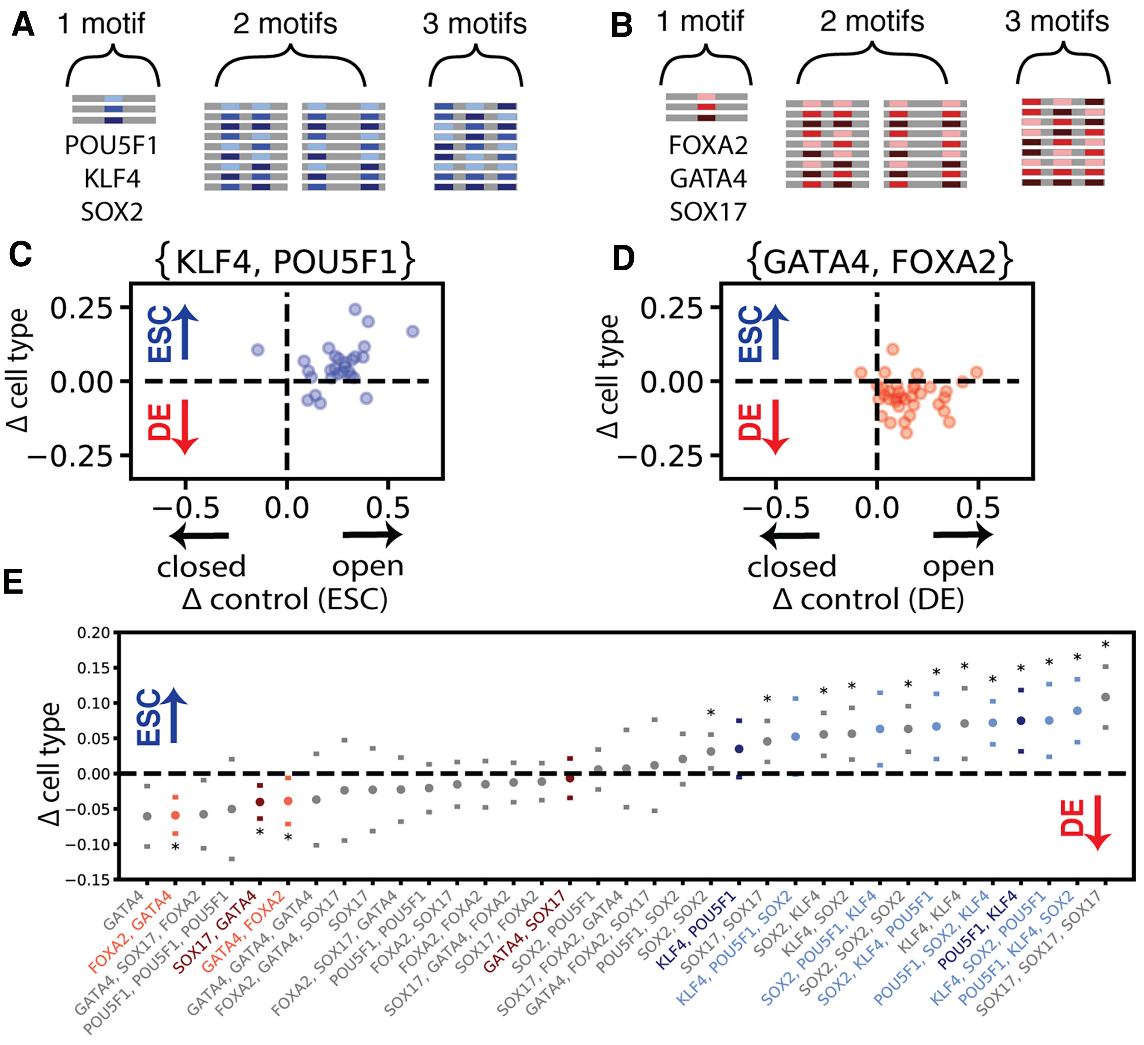


Figure 3-5: Lineage transcription factor motifs impact chromatin accessibility with preferential spatial ordering. (**A**) DNA sequence construction from the ESC key transcription factors POU5F1, SOX2, and KLF4. (**B**) DNA sequence construction from the DE key transcription factors GATA4, SOX17, and FOXA2. (**C**) Each dot represents a single neutral DNA background sequence that contains one instance of a POU5F1 motif and one instance of a KLF4 motif (two total motif instances per DNA sequence). On the y-axis is the difference between endoderm and ESC accessibility, and on the x-axis is the difference between each DNA sequence and its shuffled control in ESCs. (figure caption continued on next page)

Figure 3-5: Lineage transcription factor motifs impact chromatin accessibility with preferential spatial ordering. (**D**) Each dot represents a single neutral DNA background sequence that contains one instance of a GATA4 motif and one instance of a FOXA2 motif (two total motif instances per DNA sequence). On the y-axis is the difference between endoderm and ESC accessibility, and on the x-axis is the difference between each DNA sequence and its shuffled control in DE cells. (**E**) All motif orderings that had significant accessibility relative to random shuffled DNA controls, ranked by mean differential accessibility. Transcription factor pairs with significant changes in accessibility owing to transcription factor order are colored. Transcription factor orders with significant differential accessibility between DE cells and ESCs are starred (significance computed by paired t-test and Wilcoxon signed-rank with Benjamini-Hochberg correction at FDR < 0.05).

constrained set of known transcription factors, we designed a new library from the consensus binding motifs of the ESC lineage-defining transcription factors POU5F1, SOX2, and KLF4 (Fig. 3-5A) and the DE transcription factors FOXA2, SOX17, and GATA4 (Fig. 3-5B). We tested homotypic DNA sequences with one, two, or three instances of a motif and heterotypic DNA sequences with combinations of motifs with every possible motif ordering (in a single orientation).

We found that single motif instances were able to significantly increase accessibility compared with shuffled DNA sequences for 2/6 transcription factors (SOX17 and GATA4) but were rarely able to make DNA significantly differentially accessible (Supplemental Fig. S16). We note that the consensus motifs for SOX17 and SOX2 are highly similar, sharing a common sequence (CATTGTTT), so it is likely that both Sox factors and possibly others bind to both motifs tested. In contrast, in our DNA sequences containing two motif instances, 17/18 significantly increased accessibility compared with shuffled DNA sequences in at least one cell type (Supplemental Fig. S17), indicating that MIAA is capable of reliably detecting accessibility changes resulting from a minimum of two motif instances and that all six motifs open chromatin in at least one cell type. We then tested for differential accessibility with 6-nt versus 20-nt distance between motifs, which we selected based on literature supporting preferential distances between SOX2 and POU5F1 and between KLF4 and POU5F1 [16], and we found that none were significantly sensitive to spacing under multiple hypothesis correction. We found that overall the measured accessibility impact of these motifs did not match well with the expression of the canonical transcription factors that are expected to bind these motifs, suggesting that the MIAA assay measures more than the relative expression of specific transcription factors (Supplemental Fig. S18).

We then examined all homotypic and heterotypic conformations with one, two, or three motif instances for induction of accessibility and differential accessibility. Overall, we found that 35/42 conformations significantly increased accessibility compared with shuffled versions in at least one cell type, and 15 out of 42 motif conformations were statistically significant for differential accessibility induction after multiple hypothesis correction (Fig. 3-5E). Of these 15 conformations inducing differential accessibility, 10 are heterotypic, with POU5F1-KLF4 combinations and POU5F1-KLF4-SOX2 combinations preferentially driving accessibility in ESCs (Fig. 3-5C,E; Supplemental Fig. S19) and FOXA2-GATA4 and SOX17-GATA4 combinations driving endoderm accessibility (Fig. 3-5D,E; Supplemental Fig. S19).

In several cases, homotypic motif arrays showed accessibility patterns inconsistent with the expression of their expected transcription factors. For example, homotypic SOX17 motifs drive ESC-enriched accessibility, and homotypic FOXA2 motifs drive accessibility equivalently in ESCs and endoderm in contrast to the endoderm-specific expression of both transcription factors. Though we chose canonical motifs for factors well known in the literature to be associated with ESCs and endoderm, motifs are often shared by multiple members of a transcription factor family. In fact, it has been shown that FOXD3 binds in ESCs to motifs that will eventually become occupied by FOXA2 in endoderm [60]. This same effect likely holds for SOX2 and SOX17 as well given the similarity of their motifs.

In addition, we observed several instances of heterotypic combinations of transcription factor motifs in which order (whether a transcription factor motif was closer to the 5′ or 3′ end of the designed ssDNA sequence) had an impact on accessibility. For ESC factor binding motifs, we found that one ordering of POU5F1 and KLF4 more strongly differentially opens chromatin, whereas the other opens chromatin equivalently in both cell types (Fig 3-5E). We also found four out of six sequences that contained all three ESC reprogramming motifs were differentially accessible, and the order of these motifs had an impact on the level of differential accessibility (Fig. 3-5E).

Among endoderm factor motif combinations, we found that particular FOXA2 and GATA4 and SOX17 and GATA4 (Fig. 3-5E) orientations promoted more differential accessibility. Previous studies have implicated GATA4 and FOXA2 as accessibility-enhancing transcription factors [8, 44] and have shown that their interaction can drive accessibility changes during endoderm differentiation [6]. The motif arrangements that produced the most differential MIAA accessibility were also most often enriched in the genome (Supplemental Fig. S20). Because such native genomic instances are rare and confounded by other differences, MIAA provides a more controlled approach to identifying motif arrangements with differential activity.

3.5 Discussion

The MIAA is a new assay for measuring changes in chromatin accessibility caused by short DNA sequences integrated into a fixed locus in the genome. Most prior approaches to understanding the control of chromatin accessibility have used correlative approaches that identify genomic DNA sequences that tend to coincide with accessible chromatin in a particular cell type [19, 43, 9, 13, 56] or leverage natively occurring SNPs to identify "DNase-QTLs" for which the single nucleotide change correlates with a change in chromatin accessibility [11, 15], revealing motifs whose disruption is enriched in such variants. MIAA enables screening of an arbitrarily large and diverse library of sequences for their impact on chromatin accessibility. The MIAA assay measures the differential accessibility induced by designed oligonucleotide libraries through the preference for RAR-Dam to bind and methylate accessible DNA. MIAA can measure the relative effects on local chromatin accessibility of many sequences in parallel in a fixed genomic context. This has enabled us to identify candidate accessibility modifiers such as transcription factor binding sites and cooperative interactions between such sites. Notably, because MIAA lacks the ability to measure exact nucleosome positions, it is not suitable to identify classically defined pioneer factors that must be shown to bind to nucleosomal DNA and move or evacuate nucleosomes.

We applied MIAA to study the effects of motifs on differential accessibility between ESC and DE cell states using a number of distinct experimental designs. Through the use of native genomic 100-nt DNA sequences transplanted to a fixed locus, we were able to recapitulate the differential accessibility from native DNase-seq (Pearson's $r = 0.53$; $P < 0.001$), which we believe can be partially attributed to the use of DeepAccess to scan for highly differential native sequences that are more likely to be causal for specifying differential chromatin accessibility. Through examination of randomly shuffled control DNA sequences, we identify a distinction between how a set of natively ESC-specific and endoderm-specific sequences achieved differential accessibility. The natively endoderm-accessible sequences opened chromatin more in endoderm than in ESCs and more than their shuffled versions on average, suggesting the presence of binding sites for endoderm-specific accessibility-promoting transcription factors. On an individual level, only a subset of sequences act in this way, suggesting that a 100-nt DNA sequence does not always fully recapitulate the chromatin accessibility status of native regulatory elements, which often span over a kilobase. This may be caused by the absence in MIAA of specific sequence elements outside the 100-nt sequence that either contribute to or interact with the 100-nt sequence in its native locus.

We found a distinct pattern in the natively ESC-accessible sequences. In this cohort of sequences, MIAA accessibility was higher in ESCs than in endoderm as expected; however, there was no difference between the ESC accessibility of the DNA sequences and their shuffled counterparts. Instead, the accessibility in endoderm was reduced compared with shuffled controls, suggesting that differential accessibility of these sequences was primarily achieved through binding sites that depress accessibility in endoderm. This result indicates that, for the integration locus used in this work, MIAA is capable of measuring sequence-dependent increases and decreases in accessibility. We found suggestive evidence that E-box binding sites used by epithelial-mesenchymal transition driver transcription factors such as Zeb factors maybe responsible for this repression, as such binding sites were found in 98% of the DeepAccess-proposed ESC-enriched native genomic sequences and none of the endoderm-enriched native genomic sequences. Because the native genomic sites were selected by DeepAccess based on predicted optimal differential accessibility modeled from DNase-seq regions, it is striking to have detected such a consistent difference in the mechanism of achieving differential accessibility, and it will be intriguing to explore a larger cohort of cell type-specific sequences to determine which mechanism is more common. It is important to note that DeepAccess results will be specific to the cell types that are being compared, which may also explain why DeepAccess did not strongly identify the key ESC transcription factors. We note that our subsequent exploration of POU5F1, SOX2, and KLF4 motif combinations identified a number of designs that consistently yielded ESC-enriched accessibility compared with scrambled versions, indicating that ESCs are also capable of achieving sequence-specific increases in chromatin accessibility.

To identify causal motifs and transcription factors involved in mediating differential chromatin accessibility, we then focused on exploring DNA sequences containing various combinations of sequence motifs. We show that, independently of binding motifs, higher GC-content increases accessibility. In MIAA, we can confirm this to occur in the absence of transcription factor binding motifs because of our use of shuffled versions of each designed DNA sequence. Although it is formally possible that this GC effect is an artifact of the use of Dam methylase, we show that native genomic accessible regions also show elevated GC-content, and it has been reported that transcription factors and DNase I hypersensitive regions are also enriched in GC-rich regions [57].

In spite of its importance, predicting MIAA chromatin accessibility of held-out DNA sequences purely based on GC-content yields poor results, whereas much better results are achieved by accounting for binding motifs. Of the motifs that can be confidently matched to known transcription factor families, our results are consistent with the action of known tissue-specific pioneer factors including SOX2 and POU5F1 in ESCs and GATA4 and FOXA2 in the endoderm [8, 21, 50]. We confirm the role of FOXA2 and T in endoderm-specific chromatin opening by showing that overexpression of these DE transcription factors in ESCs can increase MIAA-measured accessibility significantly in DNA sequences with DNA-binding motifs recognized by these factors. We found that our method of aggregating motif measurements over multiple sequence backgrounds resulted in highly reproducible estimates of motif effects over biological replicates ($r = 0.963$), highlighting the power of MIAA to identify accessibility-altering motifs.

We then designed a library using consensus motifs of several key transcription factors in all possible combinations and orderings, from which we provide evidence that a single binding site is sufficient to increase chromatin accessibility and as few as two binding sites are sufficient to induce differential accessibility between two cell types. These results suggest for the first time that individual transcription factor binding events in the absence of DNA-binding cofactors are capable of altering chromatin accessibility in mammalian cells.

We also found that for motifs known to bind to both ESC and DE transcription factors, motif order has a subtle effect on accessibility, which provides support for specific transcription factor interactions driving accessibility change. This result illustrates the complexity of differential accessibility induction, which cannot simply be distilled to the presence of consensus motifs for differentially expressed transcription factors. In addition to the reuse of genomic motifs by different members of the same transcription factor family in different cell states [60], certain transcription factors such as those in the Sox and Pou family can show profoundly distinct binding to specific dimeric motifs that differ in subtle ways [3]. MIAA offers an exciting new way to explore subtleties that influence transcription factor binding logic such as motif ordering, spacing, and dimeric motifs in a controlled genomic setting.

We observed subtle effects of motif order on differential accessibility in our library using consensus motifs of lineage transcription factors, and observed strong changes in accessibility by a motif pair matching a T dimer when T was overexpressed, suggesting that MIAA has the capacity to measure the effects of transcription factor interactions on accessibility. Predicting differential accessibility from DNA sequence has been a much more difficult task than predicting cell type-consistent accessibility [18, 22, 36], and one possible reason is that more conditional logic is used. The ability of MIAA to obtain sensitive measurements of the effects of specific motif combinations on differential accessibility by exhausting all possible combinations of motifs in a controlled fashion makes MIAA a valuable tool in training accurate predictive models of chromatin accessibility. There are many directions for future work, including a deeper examination of the impact of genomic integration site on local DNA accessibility as well as a further investigation into features such as motif spacing, which are likely to impact transcription factor interaction logic. MIAA may also find an important use in classifying the large collection of SNPs that may impact chromatin accessibility [11]. Another possible application of MIAA is to understand chromatin accessibility during differentiation by taking measurements at multiple timepoints to discover novel transcription factor regulatory logic, such as switching of binding partners, in developmentally relevant cell types.

Availability

All raw and processed sequencing data generated in this study have been submitted to the NCBI Gene Expression Omnibus (GEO; `https://www.ncbi.nlm.nih.gov/geo/`) under accession number GSE145920. Prefiltered unnormalized MIAA read counts are available as Supplemental Data. Accession numbers for previously published DNase-seq, ChIP-seq, and RNA-seq data that were used in this study are listed in Supplemental Table S2. Code for DeepAccess accessibility prediction and motif extraction is available at GitHub (`https://github.com/gifford-lab/DeepAccess`). Code for MIAA library processing and producing manuscript figures is available at GitHub (`https://github.com/gifford-lab/MIAA-analysis`).

Bibliography

[1] Nour J Abdulhay, Colin P McNally, Laura J Hsieh, Sivakanthan Kasinathan, Aidan Keith, Laurel S Estes, Mehran Karimzadeh, Jason G Underwood, Hani Goodarzi, Geeta J Narlikar, and Vijay Ramani. Massively multiplex single-molecule oligonucleosome footprinting. *bioRxiv*, 2020.

[2] H. Acloque, O. H. Ocaña, D. Abad, C. D. Stern, and M. A. Nieto. Snail2 and Zeb2 repress P-cadherin to define embryonic territories in the chick embryo. *Development*, 144(4):649–656, 02 2017.

[3] I. Aksoy, R. Jauch, J. Chen, M. Dyla, U. Divakar, G. K. Bogu, R. Teo, C. K. Leng Ng, W. Herath, S. Lili, A. P. Hutchins, P. Robson, P. R. Kolatkar, and L. W. Stanton. Oct4 switches partnering from Sox2 to Sox17 to reinterpret the enhancer code and specify endoderm. *EMBO J*, 32(7):938–953, Apr 2013.

[4] A. R. Ball and K. Yokomori. Damage site chromatin: open or closed? *Curr Opin Cell Biol*, 23(3):277–283, Jun 2011.

[5] B. R. Cairns. The logic of chromatin architecture and remodelling at promoters. *Nature*, 461(7261):193–198, Sep 2009.

[6] F. M. Cernilogar, S. Hasenöder, Z. Wang, K. Scheibner, I. Burtscher, M. Sterr, P. Smialowski, S. Groh, I. M. Evenroed, G. D. Gilfillan, H. Lickert, and G. Schotta. Pre-marked chromatin and transcription factor co-binding shape the pioneering activity of Foxa2. *Nucleic Acids Res*, 47(17):9069–9086, 09 2019.

[7] R. V. Chereji, P. R. Eriksson, J. Ocampo, H. K. Prajapati, and D. J. Clark. Accessibility of promoter DNA is not the primary determinant of chromatin-mediated gene regulation. *Genome Res*, 29(12):1985–1995, 12 2019.

[8] L. A. Cirillo, F. R. Lin, I. Cuesta, D. Friedman, M. Jarnik, and K. S. Zaret. Opening of compacted chromatin by early developmental transcription factors HNF3 (FoxA) and GATA-4. *Mol Cell*, 9(2):279–289, Feb 2002.

[9] M. R. Corces, J. D. Buenrostro, B. Wu, P. G. Greenside, S. M. Chan, J. L. Koenig, M. P. Snyder, J. K. Pritchard, A. Kundaje, W. J. Greenleaf, R. Majeti, and H. Y. Chang. Lineage-specific and single-cell chromatin accessibility charts human hematopoiesis and leukemia evolution. *Nat Genet*, 48(10):1193–1203, 10 2016.

[10] M. R. Corces, J. M. Granja, S. Shams, B. H. Louie, J. A. Seoane, W. Zhou, T. C. Silva, C. Groeneveld, C. K. Wong, S. W. Cho, A. T. Satpathy, M. R. Mumbach, K. A. Hoadley, A. G. Robertson, N. C. Sheffield, I. Felau, M. A. A. Castro, B. P. Berman, L. M. Staudt, J. C. Zenklusen, P. W. Laird, C. Curtis, W. J. Greenleaf, H. Y. Chang, R. Akbani, C. C. Benz, E. A. Boyle, B. M. Broom, A. D. Cherniack, B. Craft, J. A. Demchok, A. S. Doane, O. Elemento, M. L. Ferguson, M. J. Goldman, D. N. Hayes, J. He, T. Hinoue, M. Imielinski, S. J. M. Jones, A. Kemal, T. A. Knijnenburg, A. Korkut, D. C. Lin, Y. Liu, M. K. A. Mensah, G. B. Mills, V. P. Reuter, A. Schultz, H. Shen, J. P. Smith, R. Tarnuzzer, S. Trefflich, Z. Wang, J. N. Weinstein, L. C. Westlake, J. Xu, L. Yang, C. Yau, Y. Zhao, and J. Zhu. The chromatin accessibility landscape of primary human cancers. *Science*, 362(6413), 10 2018.

[11] J. F. Degner, A. A. Pai, R. Pique-Regi, J. B. Veyrieras, D. J. Gaffney, J. K. Pickrell, S. De Leon, K. Michelini, N. Lewellen, G. E. Crawford, M. Stephens, Y. Gilad, and J. K. Pritchard. DNase I sensitivity QTLs are a major determinant of human expression variation. *Nature*, 482(7385):390–394, Feb 2012.

[12] C. Fiore and B. A. Cohen. Interactions between pluripotency factors specify cis-regulation in embryonic stem cells. *Genome Res*, 26(6):778–786, 06 2016.

[13] L. T. Gray, Z. Yao, T. N. Nguyen, T. K. Kim, H. Zeng, and B. Tasic. Layer-specific chromatin accessibility landscapes reveal regulatory networks in adult mouse visual cortex. *Elife*, 6, 01 2017.

[14] S. R. Grossman, X. Zhang, L. Wang, J. Engreitz, A. Melnikov, P. Rogov, R. Tewhey, A. Isakova, B. Deplancke, B. E. Bernstein, T. S. Mikkelsen, and E. S. Lander. Systematic dissection of genomic features determining transcription factor binding and enhancer function. *Proc Natl Acad Sci U S A*, 114(7):E1291–E1300, 02 2017.

[15] F. Grubert, J. B. Zaugg, M. Kasowski, O. Ursu, D. V. Spacek, A. R. Martin, P. Greenside, R. Srivas, D. H. Phanstiel, A. Pekowska, N. Heidari, G. Euskirchen, W. Huber, J. K. Pritchard, C. D. Bustamante, L. M. Steinmetz, A. Kundaje, and M. Snyder. Genetic Control of Chromatin States in Humans Involves Local and Distal Chromosomal Interactions. *Cell*, 162(5):1051–1065, Aug 2015.

[16] Y. Guo, S. Mahony, and D. K. Gifford. High resolution genome wide binding event finding and motif discovery reveals transcription factor spatial binding constraints. *PLoS Comput Biol*, 8(8):e1002638, 2012.

[17] Y. Guo, K. Tian, H. Zeng, X. Guo, and D. K. Gifford. A novel k-mer set memory (KSM) motif representation improves regulatory variant prediction. *Genome Res*, 28(6):891–900, 06 2018.

[18] T. Hashimoto, R. I. Sherwood, D. D. Kang, N. Rajagopal, A. A. Barkal, H. Zeng, B. J. Emons, S. Srinivasan, T. Jaakkola, and D. K. Gifford. A synergistic DNA logic predicts genome-wide chromatin accessibility. *Genome Res*, 26(10):1430–1440, 10 2016.

[19] S. Heinz, C. Benner, N. Spann, E. Bertolino, Y. C. Lin, P. Laslo, J. X. Cheng, C. Murre, H. Singh, and C. K. Glass. Simple combinations of lineage-determining

transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol Cell*, 38(4):576–589, May 2010.

[20] F. Inoue and N. Ahituv. Decoding enhancers using massively parallel reporter assays. *Genomics*, 106(3):159–164, Sep 2015.

[21] M. Iwafuchi-Doi and K. S. Zaret. Pioneer transcription factors in cell reprogramming. *Genes Dev*, 28(24):2679–2692, Dec 2014.

[22] D. R. Kelley, J. Snoek, and J. L. Rinn. Basset: learning the regulatory code of the accessible genome with deep convolutional neural networks. *Genome Res*, 26(7):990–999, 07 2016.

[23] P. Kheradpour, J. Ernst, A. Melnikov, P. Rogov, L. Wang, X. Zhang, J. Alston, T. S. Mikkelsen, and M. Kellis. Systematic dissection of regulatory motifs in 2000 predicted human enhancers using a massively parallel reporter assay. *Genome Res*, 23(5):800–811, May 2013.

[24] S. L. Klemm, Z. Shipony, and W. J. Greenleaf. Chromatin accessibility and the regulatory epigenome. *Nat Rev Genet*, 20(4):207–220, 04 2019.

[25] M. Levo, T. Avnit-Sagi, M. Lotan-Pompan, Y. Kalma, A. Weinberger, Z. Yakhini, and E. Segal. Systematic Investigation of Transcription Factor Activity in the Context of Chromatin Using Massively Parallel Binding and Expression Assays. *Mol Cell*, 65(4):604–617, Feb 2017.

[26] C. Liu, M. Wang, X. Wei, L. Wu, J. Xu, X. Dai, J. Xia, M. Cheng, Y. Yuan, P. Zhang, J. Li, T. Feng, A. Chen, W. Zhang, F. Chen, Z. Shang, X. Zhang, B. A. Peters, and L. Liu. An ATAC-seq atlas of chromatin accessibility in mouse tissues. *Sci Data*, 6(1):65, May 2019.

[27] J. Lämke and I. Bäurle. Epigenetic and chromatin-based mechanisms in environmental stress adaptation and stress memory in plants. *Genome Biol*, 18(1):124, 06 2017.

[28] B. B. Maricque, H. G. Chaudhari, and B. A. Cohen. A massively parallel reporter assay dissects the influence of chromatin structure on cis-regulatory activity. *Nat Biotechnol*, Nov 2018.

[29] B. B. Maricque, J. D. Dougherty, and B. A. Cohen. A genome-integrated massively parallel reporter assay reveals DNA sequence determinants of cis-regulatory activity in neural cells. *Nucleic Acids Res*, 45(4):e16, 02 2017.

[30] E. O. Mazzoni, S. Mahony, M. Closser, C. A. Morrison, S. Nedelec, D. J. Williams, D. An, D. K. Gifford, and H. Wichterle. Synergistic binding of transcription factors to cell-specific enhancers programs motor neuron identity. *Nat Neurosci*, 16(9):1219–1227, Sep 2013.

[31] E. O. Mazzoni, S. Mahony, M. Iacovino, C. A. Morrison, G. Mountoufaris, M. Closser, W. A. Whyte, R. A. Young, M. Kyba, D. K. Gifford, and H. Wichterle. Embryonic stem cell-based mapping of developmental transcriptional programs. *Nat Methods*, 8(12):1056–1058, Nov 2011.

[32] A. Melnikov, A. Murugan, X. Zhang, T. Tesileanu, L. Wang, P. Rogov, S. Feizi, A. Gnirke, C. G. Callan, J. B. Kinney, M. Kellis, E. S. Lander, and T. S. Mikkelsen. Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay. *Nat Biotechnol*, 30(3):271–277, Feb 2012.

[33] K. Miyamoto, K. T. Nguyen, G. E. Allen, J. Jullien, D. Kumar, T. Otani, C. R. Bradshaw, F. J. Livesey, M. Kellis, and J. B. Gurdon. Chromatin Accessibility Impacts Transcriptional Reprogramming in Oocytes. *Cell Rep*, 24(2):304–311, 07 2018.

[34] I. Mogno, J. C. Kwasnieski, and B. A. Cohen. Massively parallel synthetic promoter assays reveal the in vivo effects of binding site variants. *Genome Res*, 23(11):1908–1915, Nov 2013.

[35] S. Moosavi-Dezfooli, A. Fawzi, and P. Frossard. Deepfool: A simple and accurate method to fool deep neural networks. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2574–2582, 2016.

[36] S. Nair, D. S. Kim, J. Perricone, and A. Kundaje. Integrating regulatory DNA sequence and gene expression to predict genome-wide chromatin accessibility across cellular contexts. *Bioinformatics*, 35(14):i108–i116, 07 2019.

[37] A. Nguyen, J. Yosinski, and J. Clune. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 427–436, 2015.

[38] E. Oberbeckmann, M. Wolff, N. Krietenstein, M. Heron, J. L. Ellins, A. Schmid, S. Krebs, H. Blum, U. Gerland, and P. Korber. Absolute nucleosome occupancy map for the Saccharomyces cerevisiae genome. *Genome Res*, 29(12):1996–2009, 12 2019.

[39] S. C. Parker, E. H. Margulies, and T. D. Tullius. The relationship between fine scale DNA structure, GC content, and functional elements in 1% of the human genome. *Genome Inform*, 20:199–211, 2008.

[40] R. P. Patwardhan, J. B. Hiatt, D. M. Witten, M. J. Kim, R. P. Smith, D. May, C. Lee, J. M. Andrie, S. I. Lee, G. M. Cooper, N. Ahituv, L. A. Pennacchio, and J. Shendure. Massively parallel functional dissection of mammalian enhancers in vivo. *Nat Biotechnol*, 30(3):265–270, Feb 2012.

[41] S. Schick, D. Fournier, S. Thakurela, S. K. Sahu, A. Garding, and V. K. Tiwari. Dynamics of chromatin accessibility and epigenetic state in response to UV damage. *J Cell Sci*, 128(23):4380–4394, Dec 2015.

[42] U. Schwartz, A. Németh, S. Diermeier, J. H. Exler, S. Hansch, R. Maldonado, L. Heizinger, R. Merkl, and G. Längst. Characterizing the nuclease accessibility of DNA in human cells to map higher order structures of chromatin. *Nucleic Acids Res*, 47(3):1239–1254, 02 2019.

[43] M. Setty and C. S. Leslie. SeqGL Identifies Context-Dependent Binding Signals in Genome-Wide Regulatory Element Maps. *PLoS Comput Biol*, 11(5):e1004271, May 2015.

[44] R. I. Sherwood, T. Hashimoto, C. W. O'Donnell, S. Lewis, A. A. Barkal, J. P. van Hoff, V. Karun, T. Jaakkola, and D. K. Gifford. Discovery of directional and nondirectional pioneer transcription factors by modeling DNase profile magnitude and shape. *Nat Biotechnol*, 32(2):171–178, Feb 2014.

[45] R. I. Sherwood, R. Maehr, E. O. Mazzoni, and D. A. Melton. Wnt signaling specifies and patterns intestinal endoderm. *Mech Dev*, 128(7-10):387–400, 2011.

[46] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps, 2014.

[47] Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda Viégas, and Martin Wattenberg. Smoothgrad: removing noise by adding noise, 2017.

[48] R. P. Smith, L. Taher, R. P. Patwardhan, M. J. Kim, F. Inoue, J. Shendure, I. Ovcharenko, and N. Ahituv. Massively parallel decoding of mammalian regulatory sequences supports a flexible organizational model. *Nat Genet*, 45(9):1021–1028, Sep 2013.

[49] Can Sönmezer, Rozemarijn Kleinendorst, Dilek Imanci, Laura Villacorta, Dirk Schübeler, Vladimir Benes, and Arnaud R Krebs. Single molecule occupancy patterns of transcription factors reveal determinants of cooperative binding in vivo. *bioRxiv*, 2020.

[50] A. Soufi, M. F. Garcia, A. Jaroszewicz, N. Osman, M. Pellegrini, and K. S. Zaret. Pioneer transcription factors target partial DNA motifs on nucleosomes to initiate reprogramming. *Cell*, 161(3):555–568, Apr 2015.

[51] M. P. Stemmler, R. L. Eccles, S. Brabletz, and T. Brabletz. Non-redundant functions of EMT transcription factors. *Nat Cell Biol*, 21(1):102–112, 01 2019.

[52] A. B. Stergachis, B. M. Debo, E. Haugen, L. S. Churchman, and J. A. Stamatoyannopoulos. Single-molecule regulatory architectures captured by chromatin fiber sequencing. *Science*, 368(6498):1449–1454, 06 2020.

[53] A. Stryjewska, R. Dries, T. Pieters, G. Verstappen, A. Conidi, K. Coddens, A. Francis, L. Umans, W. F. van IJcken, G. Berx, L. A. van Grunsven, F. G. Grosveld, S. Goossens, J. J. Haigh, and D. Huylebroeck. Zeb2 Regulates Cell Fate at the Exit from Epiblast State in Mouse Embryonic Stem Cells. *Stem Cells*, 35(3):611–625, 03 2017.

[54] T. Szczesnik, J. W. K. Ho, and R. Sherwood. Dam mutants provide improved sensitivity and spatial resolution for profiling transcription factor binding. *Epigenetics Chromatin*, 12(1):36, 06 2019.

[55] B. van Steensel and S. Henikoff. Identification of in vivo DNA targets of chromatin proteins using tethered dam methyltransferase. *Nat Biotechnol*, 18(4):424–428, Apr 2000.

[56] S. Velasco, M. M. Ibrahim, A. Kakumanu, G. Garipler, B. Aydin, M. A. Al-Sayegh, A. Hirsekorn, F. Abdul-Rahman, R. Satija, U. Ohler, S. Mahony, and E. O. Mazzoni. A Multi-step Transcriptional and Chromatin State Cascade Underlies Motor Neuron Programming from Embryonic Stem Cells. *Cell Stem Cell*, 20(2):205–217, 02 2017.

[57] J. Wang, J. Zhuang, S. Iyer, X. Lin, T. W. Whitfield, M. C. Greven, B. G. Pierce, X. Dong, A. Kundaje, Y. Cheng, O. J. Rando, E. Birney, R. M. Myers, W. S. Noble, M. Snyder, and Z. Weng. Sequence features and chromatin structure around the genomic regions bound by 119 human transcription factors. *Genome Res*, 22(9):1798–1812, Sep 2012.

[58] M. A. White. Understanding how cis-regulatory function is encoded in DNA sequence using massively parallel reporter assays and designed sequences. *Genomics*, 106(3):165–170, Sep 2015.

[59] M. A. White, J. C. Kwasnieski, C. A. Myers, S. Q. Shen, J. C. Corbo, and B. A. Cohen. A Simple Grammar Defines Activating and Repressing cis-Regulatory Elements in Photoreceptors. *Cell Rep*, 17(5):1247–1254, 10 2016.

[60] J. Xu, J. A. Watts, S. D. Pope, P. Gadue, M. Kamps, K. Plath, K. S. Zaret, and S. T. Smale. Transcriptional competence and the active marking of tissue-specific enhancers by defined transcription factors in embryonic and induced pluripotent stem cells. *Genes Dev*, 23(24):2824–2838, Dec 2009.

[61] C. Yan, H. Chen, and L. Bai. Systematic Study of Nucleosome-Displacing Factors in Budding Yeast. *Mol Cell*, 71(2):294–305, 07 2018.

Chapter 4

High resolution discovery of chromatin interactions

The contents of this chapter were published as:

High resolution discovery of chromatin interactions

Yuchun Guo, Konstantin Krismer, Michael Closser, Hynek Wichterle, David K. Gifford

Nucleic Acids Research, Volume 47, Issue 6, 08 April 2019, Page e35

DOI: 10.1093/nar/gkz051 | PMID: 30953075

Supplementary information can be found in Appendix C.

4.1 Abstract

Chromatin interaction analysis by paired-end tag sequencing (ChIA-PET) is a method for the genome-wide *de novo* discovery of chromatin interactions. Existing computational methods typically fail to detect weak or dynamic interactions because they use a peak-calling step that ignores paired-end linkage information. We have developed a novel computational method called Chromatin Interaction Discovery (CID) to overcome this limitation with an unbiased clustering approach for interaction discovery. CID outperforms existing chromatin interaction detection methods with improved sensitivity, replicate consistency, and concordance with other chromatin interaction datasets. In addition, CID also outperforms other methods in discovering chromatin interactions from HiChIP data. We expect that the CID method will be valuable in characterizing 3D chromatin interactions and in understanding the functional consequences of disease-associated distal genetic variations.

4.2 Introduction

Physical three-dimensional (3D) chromatin interactions between regulatory genomic elements play an important role in regulating gene expression [2, 31]. For example, the creation of chromatin interactions between the promoters and locus control regions of the β-globin gene is sufficient to trigger transcriptional activation, indicating that chromatin looping causally underlies gene regulation [4].

Chromatin interaction analysis by paired-end tag sequencing (ChIA-PET) is a technology for the genome-wide de novo detection of chromatin interactions mediated by a specific protein factor [8]. In ChIA-PET, crosslinked chromatin is sonicated and then immunoprecipitated by antibodies that bind to a protein of interest, followed by proximity ligation, and sequencing [8]. The paired-end tags (PETs) are then mapped to the genome to identify the two genomic locations that interact with each other. Therefore, similar to Hi-C data [19], the ChIA-PET interactions are represented by a pair of genomic locations that interact with each other. By focusing on the chromatin interactions associated with a specific protein, ChIA-PET is capable of generating high-resolution ($\approx$ 100 bp) genome-wide chromatin interaction maps of functional elements [29]. The ChIA-PET method has been used to detect structures defined by architectural proteins, including CTCF [29, 10] and cohesin [5, 13], detect enhancer-promoter interactions associated with RNAPII [14, 18, 32], and detect interactions involving other transcription factors [8, 30]. In addition, multiple studies have applied the ChIA-PET method to link distal genetic variants to their target genes and to study the structural and functional consequences of non-coding genetic variations [29, 9].

To gain biological insight from ChIA-PET data, computational analysis pipelines and statistical models have been developed [12, 17, 16, 23, 24]. Typically, analysis pipelines start with data pre-processing that includes linker filtering and linker removal. The resulting PETs are then mapped to the genome and duplicated PETs are removed. To detect chromatin interactions, a peak-calling step [17, 16, 24] is usually used to define peak regions enriched with reads as interaction anchors, and then groups of PETs linking two peak regions are considered as candidate interactions.

Finally, the number of PETs supporting a candidate interaction is used to compute the statistical significance of the interaction.

Existing chromatin interaction methods based on peak-calling [17, 16, 24] lose information at the peak-calling step by ignoring the paired-end linkage information that is indicative of chromatin interactions. For example, for an RNAPII ChIA-PET dataset that aims to detect promoter-enhancer interactions, the RNAPII signal enrichment at certain weak or dynamic enhancers may not be strong enough to be detected as a peak by the peak-calling algorithm. Thus, interactions involving weak enhancers typically will not be detected, even though there may be a sufficient number of PETs linking these enhancers to other genomic elements in the raw data. In addition, for interactions with detected anchors, the PET count quantification may be inaccurate because some nearby PETs may fall outside of the peak region boundaries. Thus, peak-calling-based approaches limit the detection of candidate interactions and can inaccurately quantify the PET count support.

We developed a novel computational method called chromatin interaction discovery (CID) that uses an unbiased clustering approach to detect chromatin interactions to address the shortcomings of peak-calling-based methods. We show that CID can be applied to both ChIA-PET and HiChIP data and that CID outperforms existing peak-calling-based methods in terms of sensitivity, replicate consistency, and concordance with other chromatin interaction datasets.

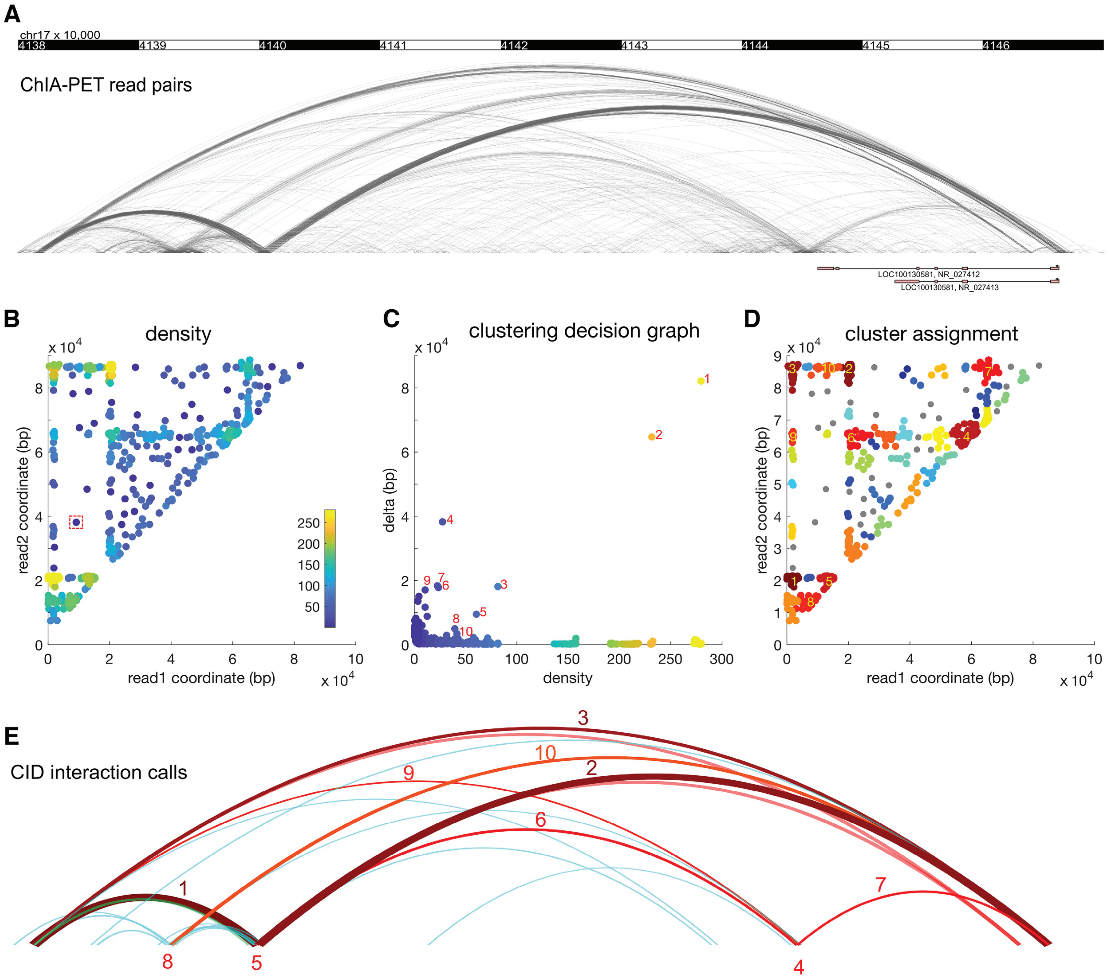


Figure 4-1: CID uses density-based clustering to discover chromatin interactions. (**A**) ChIA-PET interactions can be discovered as groups of dense arcs connecting two genomic regions. Each arc is a PET. (**B**) The PETs plotted on a two-dimensional map using the genomic coordinates of the two reads. Each point is a PET. The colors represent the density values, defined as the number of PETs in the neighborhood. The red dashed square represents the size of the neighborhood. (**C**) The clustering decision graph. Each point is a PET. The points with high density and high delta values are selected as cluster centers. For simplicity, only large clusters are labelled. (**D**) The read pairs are assigned to the nearest cluster centers. The clusters are labeled as in (C). (**E**) The clusters are visualized as arcs. The clusters are labeled as in (C) and (D).

4.3 Materials and Methods

4.3.1 Segmentation of PETs

First, CID groups all the single-end reads that are within 5000 bp of each other into non-overlapping regions. The maximum DNA fragment size in the ChIA-PET protocol is estimated to be ≈ 5000 bp [17]. Therefore, two groups of reads that are > 5000 bp apart are expected to belong to independent interaction anchor regions. Next, for each region, we group PETs whose left reads map to the region into groups where the right reads of the PETs map to independent anchor regions that are at least 5000 bp from each other. We then further split the PET groups if the left reads of the PETs in a group can be split into independent anchor regions. This process iterates until the PET groups cannot be further split. The result of this segmentation step is that millions of PETs are split into small non-overlapping groups that typically contain $< 10,000$ PETs.

4.3.2 ChIA-PET and HiChIP datasets

ChIA-PET datasets (17 datasets associated with protein factors such as POL2RA, CTCF and RAD21) from various cell types [13, 18, 9] (Supplementary Table C.1) were downloaded from the ENCODE Project portal (`https://www.encodeproject.org/`). FASTQ files of both biological replicates were pre-processed and aligned to the hg19 genome using the Mango pipeline [24]. The fastq and pre-processed SMC1A HiChIP data from GM12878 cells [22] were downloaded from NCBI GEO portal (GSE80820). BEDPE files from ChIA-PET and HiChIP datasets are used as inputs to CID.

4.3.3 Mango and ChIA-PET2 pipelines

Mango (version 1.2.1) [24] was downloaded from `https://github.com/dphansti/mango`. Additionally, we installed the dependencies R (version 3.4.4), bedtools (version 2.26.0), macs2 (version 2.1.1.20160309), and bowtie-align (version 1.2). Mango was executed with the default parameters and the flags verboseoutput and reportallpairs were set. For data sets that were generated with the ChIA-PET Tn5 tagmentation protocol, additional parameters recommended by the author were used: -keepempty TRUE -maxlength 1000 -shortreads FALSE.

The BEDPE files generated by Mango after step 3 were also used by the ChIA-PET2 and CID pipelines in order to examine the differences in the subsequent peak calling and interaction calling steps.

ChIA-PET2 (version 0.9.2) [16] was obtained from `https://github.com/GuipengLi/ChIA-PET2`. The default setting for all parameters were used, except that the starting step was set to 4 to start the analysis from Mango-derived BEDPE files.

4.3.4 hichipper pipeline

The HiChIP raw fastq files were initially processed with HiC-Pro [28] (`https://github.com/nservant/HiC-Pro`) using default settings except specifying MboI instead of HindIII digestion. Subsequently, hichipper [15] (`https://github.com/aryeelab/hichipper`) was used to analyze the HiC-Pro output, specifying EACH,ALL as the peaks option and providing the MboI BED file for restriction fragments.

4.3.5 Replicate consistency analysis

For each dataset, we counted the number of interactions that are present in both replicates. Jaccard coefficients are then calculated by dividing the intersection of interactions in replicates 1 and 2 by the union of interactions in both replicates. Interactions in replicates 1 and 2 were considered identical, if both interaction anchors overlapped between replicates or the gap between them was $<$1000 bp.

In situations where the ranking of interactions mattered (e.g. fraction of replicated interactions in the top n interactions), interactions were sorted in ascending order of their false discovery rate (FDR) and posterior probability (if there were tied FDR values).

4.3.6 Functional annotation of interaction calls

The GENCODE 19 gene annotation [11] was used to generate the promoter annotations. Each transcription start site (TSS) is expanded to 2.5 kb up/downstream to define a promoter. We used ChIP-seq peak calls of H3K27ac histone modification, which associates with active enhancers, as the enhancer annotations. The set of broad peak calls of H3K27ac ChIP-seq data from K562 cells was downloaded from ENCODE project website (accession ENCFF931VAQ). For the interaction calls from all the methods, a call is considered annotated as an enhancer-promoter interaction if one anchor region of the interaction overlaps with a promoter annotation and the other anchor overlaps an enhancer annotation.

4.3.7 Hi-C loop overlap analysis

Hi-C loop calls for GM12878 and K562 cells [25] were downloaded from NCBI GEO portal (GSE63525, combined primary and replicate samples). The HICCUPS loop calls from SMC1A HiChIP data were downloaded from Mumbach et al. [22]. The overlap between Hi-C loops and ChIA-PET and HiChIP interaction calls were computed using pairToPair in bedtools with parameters '-slop 1000 -type both -is'.

4.3.8 5C interaction overlap analysis

5C interaction calls for K562 cells were downloaded from the original study [27]. The overlap between 5C interactions (tested and positive) and ChIA-PET interaction calls were computed using pairToPair in bedtools with parameters '-slop 1000 -type both -is'.

4.3.9 Software availability

The CID software was implemented in Java. Information on CID is at (`http://giffordlab.mit.edu/cid/`). The source code and license is at (`https://github.com/gifford-lab/GEM3`)

4.4 Results

4.4.1 Chromatin interaction discovery (CID)

CID discovers chromatin interactions using a density-based clustering method [26] to cluster proximal PETs into interactions. CID continuously resolves anchors and thus is more flexible than peak-calling-based methods that can only discover interactions between statically identified peak regions. Once CID identifies candidate interactions, it then applies the MICC statistical model [12] to compute the statistical significance of the interactions.

CID first filters out PETs that are shorter than 5000 bp because they are likely to be self-ligation PETs. CID then efficiently clusters ChIA-PET data by segmenting the total set of PETs into independent groups of proximally located PETs (see section 4.3.1). CID then clusters each group of PETs.

A PET i is represented as a two-dimensional vector $[C_{i,L}, C_{i,R}]$, where $C_{i,L}$ and $C_{i,R}$ are the genomic coordinates of the center of the left and right reads of PET i, respectively, $C_{i,L} < C_{i,R}$. The distance between two PETs is quantified as the Chebyshev distance [1] calculated from the read coordinates of the PETs:

$$\text{Distance}\left(\text{PET}_i, \text{PET}_j\right) = \max\left(\left|C_{i,L} - C_{j,L}\right|, \left|C_{i,R} - C_{j,R}\right|\right) \tag{4.1}$$

We employ a density-based clustering method [26] that finds cluster centers that are characterized by a higher density than their neighbors and by a relatively large distance from points with higher densities. For ChIA-PET data (Figure 4-1A), the density of a PET is defined as the number of neighboring PETs within a certain cutoff distance to the PET. The densities of all PETs can be visualized by plotting each PET as a point (Ci,L,Ci,R) on a 2D space (Figure 4-1B). A high-density group of points on the plot suggests potential chromatin interactions between two genomic regions. After computing the density values, a delta value of a PET is defined as its distance to the nearest PET that has a higher density. For the PET with the highest density, delta is defined as the largest distance between any pair of PETs [26]. By requiring

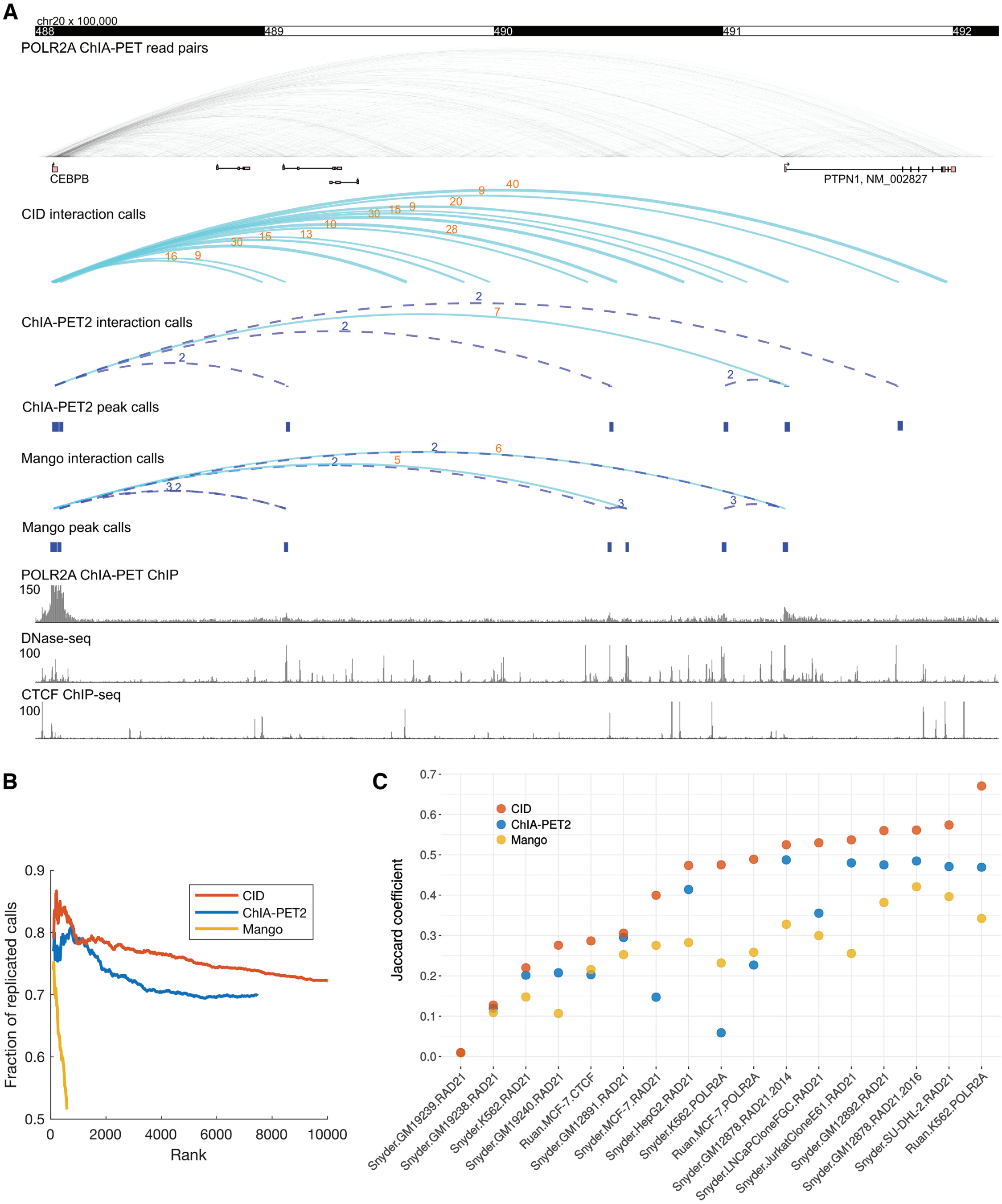


Figure 4-2: CID is more sensitive and consistent at discovering ChIA-PET interactions than peak-calling-based methods. (figure caption continued on next page)

Figure 4-2: CID is more sensitive and consistent at discovering ChIA-PET interactions than peak-calling-based methods. (**A**) Comparison of interactions called by CID, ChIA-PET2, and Mango in the CEBPB locus using POLR2A ChIA-PET data from K562 cells. The ChIA-PET2 and Mango interaction calls are based on peak calls (shown as blue rectangles) from the same ChIA-PET data by treating PETs as single-end reads (shown as the ChIA-PET ChIP track). The PET counts of the interactions are represented as the numeric values above the arcs. For CID, only significant interactions with >8 PETs are shown. Dashed-line arcs represents insignificant candidate interactions. (**B**) Interaction calls of CID are more consistent across replicates than those of ChIA-PET2 and Mango. Accumulative fractions of replicated interaction calls are computed using top ranking interactions at increasing ranks. For CID, only top 10 000 calls are shown. (**C**) Interaction calls of CID are more replicable than those of ChIA-PET2 and Mango across a large set of ChIA-PET data. For each dataset, same number of top-ranking calls in replicates are used to compute Jaccard coefficient for all three methods.

cluster centers have high delta values, the clustering method prevents too many points in a high-density region from being called as cluster centers [26]. The cluster centers are then the PETs with both high density and high delta values, as visualized in a clustering decision graph (Figure 4-1C). Following the clustering method [26], PETs are ranked by the product of their density and delta values and a PET is assigned to the same cluster as its nearest neighbor of higher density (Figure 4-1D). After cluster assignment, singleton clusters are interpreted as noise and are not considered as candidate interactions. Because some PET clusters may be close to each other, a post-processing step merges nearby PET clusters. The PET clusters that contain at least two PETs are then proposed as candidate interactions (Figure 4-1E).

CID then applies the MICC statistical model [12] to compute the statistical significance of the candidate interactions. MICC applies a Bayesian mixture model to systematically separate true interactions from random ligation and random collision noise and computes false discovery rates (FDRs) for the candidate interactions [12]. The cutoff criteria for significant interactions are (i) FDR ≤ 0.05 and (ii) PET count > 3. In principle, CID can use the MICC, Mango, or ChiaSig [23] models to compute statistical significance of the discovered interactions. We chose the MICC model because it has been shown to be more sensitive than the Mango model [12].

4.4.2 CID is more sensitive at discovering ChIA-PET interactions than peak-calling-based methods

We compared CID with two peak-calling-based ChIA-PET analysis methods, ChIA-PET2 [16] and Mango [24], and found that CID is more sensitive than these methods at chromatin interaction discovery. We tested these three methods on a widely used dataset, POL2RA ChIA-PET from K562 cells [18]. We first studied the chromatin interactions called by three methods in the 400 kb genomic region downstream of the CEBPB gene. The pre-determined peak regions called by ChIA-PET2 and Mango limit the interactions that can be discovered. In contrast, CID uses an unbiased approach and discovers a substantial number of interactions that are missed by ChIA-

PET2 and Mango (Figure 4-2A). The missed interactions are between the CEBPB promoter and non-promoter regions that have weak enrichment of reads and are not called as peaks by ChIA-PET2 or Mango. In addition, peak-calling-based methods only count the PETs that are within the peak regions and miss nearby PETs that just fall out of the peak region boundaries. In contrast, CID's clustering approach includes all the neighboring PETs. Indeed, the PET count in the CID called interactions are higher than the same interactions called by the other methods (Figure 4-2A). The accurate quantification of PET counts for interactions is important for the subsequent test of their statistical significance. Many of the candidate interactions called by ChIA-PET2 and Mango contain too few PETs to reach statistical significance, yet the interactions called by CID across the same anchor regions are statistically significant because their PET counts are higher (Figure 4-2A). We further compared the significant interactions in the CEBPB locus between two biological replicates and found that CID called 11 replicable interactions that contain at least 9 PETs. In contrast, there are only one replicable ChIA-PET2 interaction and zero replicable Mango interactions in this region (Supplementary Figure C-1). Across the whole genome, CID discovers more interactions than ChIA-PET2 and Mango (Figure 4-2B, Supplementary Figure S2).

Next, we investigated whether the interactions discovered by CID are functionally relevant. For the K562 POL2RA ChIA-PET data, we overlapped the interaction calls by all three methods with the annotations of enhancers (E, H3K27ac ChIP-seq peaks in K562 cells) and promoters (P, 2.5kb up/downstream of annotated TSS in GENCODE 19). An interaction is annotated as a candidate enhancer-promoter interaction if one of its anchor regions overlaps with at least one promoter or enhancer annotation and the other anchor region overlaps with at least one annotation of the opposite type (E-P or P-E). More than 80% of CID calls are annotated as candidate enhancer-promoter interactions, at the similar percentage of overlaps of calls from ChIA-PET2 and Mango (Supplementary Figure S3). Furthermore, high-ranking CID calls overlap with annotations at a higher percentage than calls from the other methods. These results suggest that CID calls reveal chromatin interactions with relevant biological function.

4.4.3 CID is more consistent at discovering ChIA-PET interactions than peak-calling-based methods

In addition, CID interaction calls are more consistent across biological replicates than those of ChIA-PET2 and Mango. For each method, we compared the interactions called from biological replicates and computed the accumulated fraction of replicated calls with increasing number of top-ranking calls. For the K562 POL2RA ChIA-PET dataset, the interaction calls of CID are more replicable than those of ChIA-PET2 and Mango (Figure 4-2B). We further compared the replicate consistency of the three methods across a large set of replicated ChIA-PET datasets from the ENCODE project [6], which assay interactions mediated by factors such as POL2RA, CTCF and RAD21 (a cohesin subunit), across multiple cell types. Because the numbers of interactions called by the three methods are different, for each dataset, we took the same number of top-ranking interaction calls and computed the Jaccard coefficient between the two replicates. We found that CID has higher Jaccard coefficients than ChIA-PET2 and Mango across all 17 datasets we tested (Figure 4-2C). Across all these ChIA-PET datasets, CID is not only more sensitive but also more consistent in discovering chromatin interactions than ChIA-PET2 and Mango (Supplementary Figure S2, Supplementary Table S2). We also computed the interaction length distribution and anchor width distribution of interaction calls from Mango, ChIA-PET2, and CID for all 17 ChIA-PET data sets (Supplementary Figures S4 and S5). The interaction length distributions are similar among the tested methods. In contrast, the anchor width distribution of CID differs from other methods because CID called anchors are defined by clustered PETs instead of the peaks determined based on single-end read enrichment.

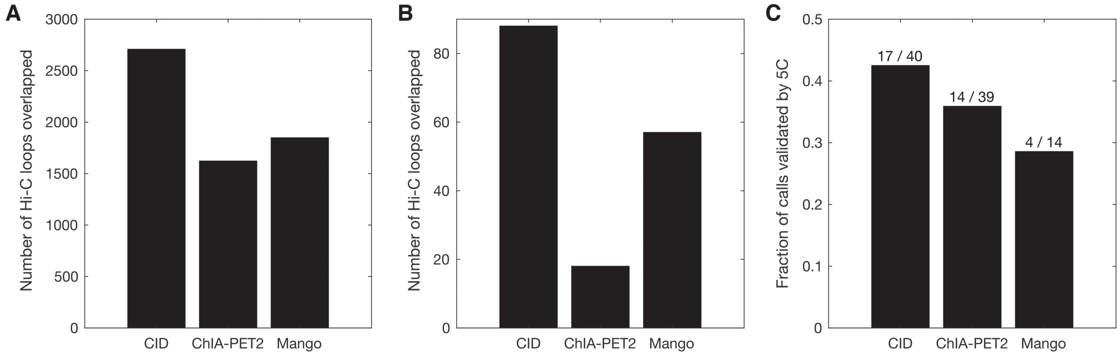


Figure 4-3: Interactions called by CID are more concordant with Hi-C and 5C data than interactions called by ChIA-PET2 and Mango. (**A**) Number of Hi-C loops in GM12878 cells that overlapped with top 5530 interactions called by three methods from RAD21 ChIA-PET data in GM12878 cells. (**B**) Number of Hi-C loops in K562 cells that overlapped with top 613 interactions called by three methods from POLR2A ChIA-PET data in K562 cells. (**C**) Fraction of interactions called by three methods from POLR2A ChIA-PET data in K562 cells that are validated by 5C interactions in K562 cells. The values above the bars show the number of 5C interactions tested positive and the number of 5C interactions tested, respectively.

4.4.4 Interactions called by CID are more concordant with Hi-C and 5C data than interactions called by other methods

To compare the accuracy and biological relevance of interactions detected by CID and other methods, we intersected the interaction calls with the chromatin loop calls from deeply sequenced Hi-C data [25]. We first tested the concordance between RAD21 ChIA-PET interactions calls and the Hi-C loop calls in GM12878 cells. Because CID, ChIA-PET2, and Mango called different numbers of significant interactions, we focused on comparing the top 5530 interactions called by the three methods. We found that interactions called by CID overlap with more Hi-C loops than those called by ChIA-PET2 and Mango. The number of Hi-C loops overlapped with interactions called by CID, ChIA-PET2, and Mango are 2708, 1622 and 1848, respectively (Figure 4-3A). Similarly, for K562 cells, POL2RA interactions called by CID overlap with more Hi-C loops than those called by ChIA-PET2 and Mango. The number of Hi-C loops overlapped with the top 631 interactions called by CID, ChIA-PET2, and Mango are 88, 18 and 57, respectively (Figure 4-3B). In addition, the number of Hi-C loops overlapped with the top 7498 interactions called by CID and ChIA-PET2 are 396 and 83, respectively (Supplementary Figure S6).

We also compared the significant interactions from the three methods with 3C-Carbon Copy (5C) data mapped as part of the ENCODE project across 1% of the genome [27]. For the K562 POL2RA ChIA-PET interactions called by the three methods, we compared the fraction of the interactions that are validated by 5C interactions in K562 cells. Out of 39 interactions tested by 5C that overlap the 7498 significant interactions called by ChIA-PET2, 14 were tested positive by 5C. Out of 14 interactions tested by 5C that overlap the 631 significant interactions called by Mango, 4 were tested positive by 5C. In comparison, 40 interactions tested by 5C overlap with the top 7498 significant interactions called by CID, 17 were tested positive by 5C. The fraction of positive 5C interactions are higher for CID than for ChIA-PET2 and Mango (Figure 4-3C).

Taken together, these results show that the interactions called by CID are more

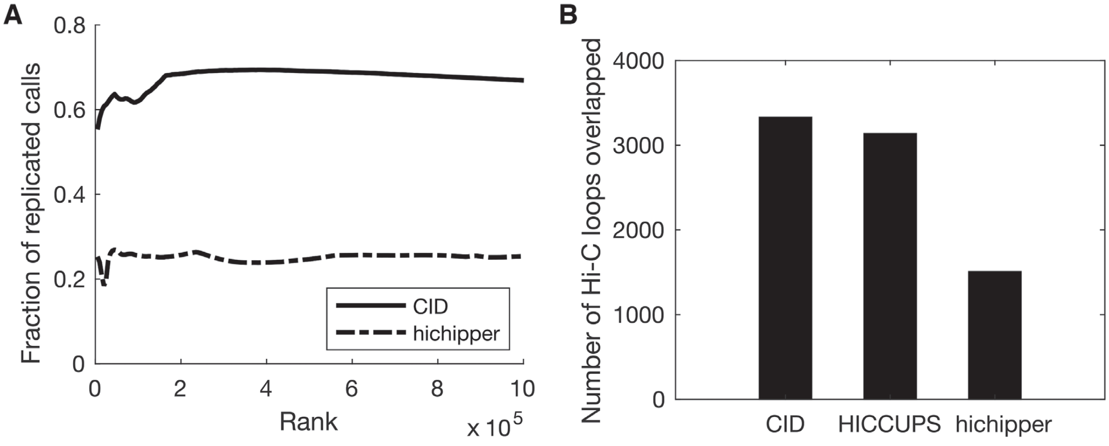


Figure 4-4: CID outperforms other methods in detecting chromatin interactions from HiChIP data. (**A**) Interaction calls of CID are more consistent across replicates than those of hichipper. Accumulative fractions of replicated interaction calls are computed using top ranking interactions at increasing ranks. Top 100 000 calls are shown. (**B**) Number of Hi-C loops in GM12878 cells that overlapped with top 10 255 interactions called by CID, HICCUPS, and hichipper from SMC1A HiChIP data in GM12878 cells.

concordant with Hi-C and 5C data than interactions called by other methods, suggesting that the interactions discovered by CID are more accurate and biologically relevant.

4.4.5 CID outperforms other methods in detecting chromatin interactions from HiChIP data

CID can also be applied to HiChIP [22] data for discovering chromatin interactions. HiChIP is a recently introduced method that is similar to ChIA-PET. It is an attractive alternative to ChIA-PET because it requires substantially fewer cells and a simpler protocol [22]. We applied CID to a cohesin-associated HiChIP dataset [22] and found that the interactions discovered by CID are similar to a cohesin-associated ChIA-PET dataset [9] in terms of replicate consistency (Supplementary Figure S7).

We then compared the results with those from hichipper [15], a peak-calling-based method for analyzing HiChIP data. We found that the interaction calls of CID are more consistent across two replicates than those of hichipper (Figure 4-4A). In addition, we overlapped CID, hichipper, and HICCUPS calls [22] from the same SMC1A HiChIP data with the Hi-C loops from GM12878 cells [25]. Because HICCUPS only called 10255 significant interactions from the HiChIP data, we focused on comparing the top 10255 interactions called by the three methods. We found that interactions called by CID overlap with slightly more Hi-C loops than those called by HICCUPS, and significantly more than those called by hichpper. The number of Hi-C loops overlapped with interactions called by CID, HICCUPS, and hichipper are 3331, 3137 and 1507, respectively (Figure 4-4B). We note that HICCUPS is the same software that called the loops from Hi-C data [25]. These results show that CID can also be used to detect chromatin interactions from HiChIP data.

4.5 Discussion

We have demonstrated that CID is more sensitive in discovering chromatin interactions from ChIA-PET data than existing peak-calling-based methods. In addition, the interactions discovered by CID are more consistent across biological replicates and more concordant with other types of chromatin interaction data than those discovered by existing methods. We anticipate the improved accuracy and reliability of CID will be important for elucidating the mechanisms of 3D genome folding and long-range gene regulation.

We have also shown that CID can be used to detect chromatin interactions from HiChIP data. A recent study [15] showed that correction of the cut site bias of the restriction enzyme improves the detection of interaction anchors. Future development of CID with HiChIP-specific modeling of the cut site bias may further improve the detection of interactions from HiChIP data.

Cell-type-specific gene expression is often regulated by distal enhancers, and these enhancers are often enriched with disease-associated variants [7, 21]. However, linking disease-associated non-coding variants to their affected genes in disease relevant tissues has been challenging due to the scarcity of long-range interaction data. With large scale on-going efforts such as the ENCODE project [6] and the 4D Nucleosome project [3], high resolution chromatin interaction mapping from a wider range of tissues and cells will become available in the near future. We expect that the CID method will be valuable in characterizing 3D chromatin interactions and in understanding the functional consequences of disease-associated distal genetic variations [20].

Bibliography

[1] James Abello, Panos M. Pardalos, and Mauricio G. C. Resende. *Handbook of Massive Data Sets*, volume 4. Springer, Boston, MA, 2002.

[2] J. Dekker. Gene regulation in the third dimension. *Science*, 319(5871):1793–1794, Mar 2008.

[3] J. Dekker, A. S. Belmont, M. Guttman, V. O. Leshyk, J. T. Lis, S. Lomvardas, L. A. Mirny, C. C. O'Shea, P. J. Park, B. Ren, J. C. R. Politz, J. Shendure, and S. Zhong. The 4D nucleome project. *Nature*, 549(7671):219–226, 09 2017.

[4] W. Deng, J. Lee, H. Wang, J. Miller, A. Reik, P. D. Gregory, A. Dean, and G. A. Blobel. Controlling long-range genomic interactions at a native locus by targeted tethering of a looping factor. *Cell*, 149(6):1233–1244, Jun 2012.

[5] J. M. Dowen, Z. P. Fan, D. Hnisz, G. Ren, B. J. Abraham, L. N. Zhang, A. S. Weintraub, J. Schujiers, T. I. Lee, K. Zhao, and R. A. Young. Control of cell identity genes occurs in insulated neighborhoods in mammalian chromosomes. *Cell*, 159(2):374–387, Oct 2014.

[6] I. Dunham, A. Kundaje, S. F. Aldred, P. J. Collins, C. A. Davis, F. Doyle, C. B. Epstein, S. Frietze, J. Harrow, R. Kaul, J. Khatun, B. R. Lajoie, S. G. Landt, B. K. Lee, F. Pauli, K. R. et al. Rosenbloom, and E. Birney. An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489(7414):57–74, Sep 2012.

[7] K. K. Farh, A. Marson, J. Zhu, M. Kleinewietfeld, W. J. Housley, S. Beik, N. Shoresh, H. Whitton, R. J. Ryan, A. A. Shishkin, M. Hatan, M. J. Carrasco-Alfonso, D. Mayer, C. J. Luckey, N. A. Patsopoulos, P. L. De Jager, V. K. Kuchroo, C. B. Epstein, M. J. Daly, D. A. Hafler, and B. E. Bernstein. Genetic and epigenetic fine mapping of causal autoimmune disease variants. *Nature*, 518(7539):337–343, Feb 2015.

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

[9] F. Grubert, J. B. Zaugg, M. Kasowski, O. Ursu, D. V. Spacek, A. R. Martin, P. Greenside, R. Srivas, D. H. Phanstiel, A. Pekowska, N. Heidari, G. Euskirchen, W. Huber, J. K. Pritchard, C. D. Bustamante, L. M. Steinmetz, A. Kundaje, and M. Snyder. Genetic Control of Chromatin States in Humans Involves Local and Distal Chromosomal Interactions. *Cell*, 162(5):1051–1065, Aug 2015.

[10] L. Handoko, H. Xu, G. Li, C. Y. Ngan, E. Chew, M. Schnapp, C. W. Lee, C. Ye, J. L. Ping, F. Mulawadi, E. Wong, J. Sheng, Y. Zhang, T. Poh, C. S. Chan, G. Kunarso, A. Shahab, G. Bourque, V. Cacheux-Rataboul, W. K. Sung, Y. Ruan, and C. L. Wei. CTCF-mediated functional chromatin interactome in pluripotent cells. *Nat. Genet.*, 43(7):630–638, Jun 2011.

[11] J. Harrow, A. Frankish, J. M. Gonzalez, E. Tapanari, M. Diekhans, F. Kokocinski, B. L. Aken, D. Barrell, A. Zadissa, S. Searle, I. Barnes, A. Bignell, V. Boychenko, T. Hunt, M. Kay, G. Mukherjee, J. Rajan, G. Despacio-Reyes, G. Saunders, C. Steward, R. Harte, M. Lin, C. Howald, A. Tanzer, T. Derrien, J. Chrast,

N. Walters, S. Balasubramanian, B. Pei, M. Tress, J. M. Rodriguez, I. Ezkurdia, J. van Baren, M. Brent, D. Haussler, M. Kellis, A. Valencia, A. Reymond, M. Gerstein, R. Guigó, and T. J. Hubbard. GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res.*, 22(9):1760–1774, Sep 2012.

[12] C. He, M. Q. Zhang, and X. Wang. MICC: an R package for identifying chromatin interactions from ChIA-PET data. *Bioinformatics*, 31(23):3832–3834, Dec 2015.

[13] N. Heidari, D. H. Phanstiel, C. He, F. Grubert, F. Jahanbani, M. Kasowski, M. Q. Zhang, and M. P. Snyder. Genome-wide map of regulatory interactions in the human genome. *Genome Res.*, 24(12):1905–1917, Dec 2014.

[14] K. R. Kieffer-Kwon, Z. Tang, E. Mathe, J. Qian, M. H. Sung, G. Li, W. Resch, S. Baek, N. Pruett, L. Grøntved, L. Vian, S. Nelson, H. Zare, O. Hakim, D. Reyon, A. Yamane, H. Nakahashi, A. L. Kovalchuk, J. Zou, J. K. Joung, V. Sartorelli, C. L. Wei, X. Ruan, G. L. Hager, Y. Ruan, and R. Casellas. Interactome maps of mouse gene regulatory domains reveal basic principles of transcriptional regulation. *Cell*, 155(7):1507–1520, Dec 2013.

[15] C. A. Lareau and M. J. Aryee. hichipper: a preprocessing pipeline for calling DNA loops from HiChIP data. *Nat. Methods*, 15(3):155–156, 02 2018.

[16] G. Li, Y. Chen, M. P. Snyder, and M. Q. Zhang. ChIA-PET2: a versatile and flexible pipeline for ChIA-PET data analysis. *Nucleic Acids Res.*, 45(1):e4, 01 2017.

[17] G. Li, M. J. Fullwood, H. Xu, F. H. Mulawadi, S. Velkov, V. Vega, P. N. Ariyaratne, Y. B. Mohamed, H. S. Ooi, C. Tennakoon, C. L. Wei, Y. Ruan, and W. K. Sung. ChIA-PET tool for comprehensive chromatin interaction analysis with paired-end tag sequencing. *Genome Biol.*, 11(2):R22, 2010.

[18] G. Li, X. Ruan, R. K. Auerbach, K. S. Sandhu, M. Zheng, P. Wang, H. M. Poh, Y. Goh, J. Lim, J. Zhang, H. S. Sim, S. Q. Peh, F. H. Mulawadi, C. T.

Ong, Y. L. Orlov, S. Hong, Z. Zhang, S. Landt, D. Raha, G. Euskirchen, C. L. Wei, W. Ge, H. Wang, C. Davis, K. I. Fisher-Aylor, A. Mortazavi, M. Gerstein, T. Gingeras, B. Wold, Y. Sun, M. J. Fullwood, E. Cheung, E. Liu, W. K. Sung, M. Snyder, and Y. Ruan. Extensive promoter-centered chromatin interactions provide a topological basis for transcription regulation. *Cell*, 148(1-2):84–98, Jan 2012.

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

[20] J. MacArthur, E. Bowler, M. Cerezo, L. Gil, P. Hall, E. Hastings, H. Junkins, A. McMahon, A. Milano, J. Morales, Z. M. Pendlington, D. Welter, T. Burdett, L. Hindorff, P. Flicek, F. Cunningham, and H. Parkinson. The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). *Nucleic Acids Res.*, 45(D1):D896–D901, 01 2017.

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

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

[23] J. Paulsen, E. A. Rødland, L. Holden, M. Holden, and E. Hovig. A statistical model of ChIA-PET data for accurate detection of chromatin 3D interactions. *Nucleic Acids Res.*, 42(18):e143, Oct 2014.

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

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

[26] A. Rodriguez and A. Laio. Machine learning. Clustering by fast search and find of density peaks. *Science*, 344(6191):1492–1496, Jun 2014.

[27] A. Sanyal, B. R. Lajoie, G. Jain, and J. Dekker. The long-range interaction landscape of gene promoters. *Nature*, 489(7414):109–113, Sep 2012.

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

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

[30] A. S. Weintraub, C. H. Li, A. V. Zamudio, A. A. Sigova, N. M. Hannett, D. S. Day, B. J. Abraham, M. A. Cohen, B. Nabet, D. L. Buckley, Y. E. Guo, D. Hnisz, R. Jaenisch, J. E. Bradner, N. S. Gray, and R. A. Young. YY1 Is a Structural Regulator of Enhancer-Promoter Loops. *Cell*, 171(7):1573–1588, Dec 2017.

[31] M. Yu and B. Ren. The Three-Dimensional Organization of Mammalian Genomes. *Annu. Rev. Cell Dev. Biol.*, 33:265–289, 10 2017.

[32] Y. Zhang, C. H. Wong, R. Y. Birnbaum, G. Li, R. Favaro, C. Y. Ngan, J. Lim, E. Tai, H. M. Poh, E. Wong, F. H. Mulawadi, W. K. Sung, S. Nicolis, N. Ahituv, Y. Ruan, and C. L. Wei. Chromatin connectivity maps reveal dynamic promoter-enhancer long-range associations. *Nature*, 504(7479):306–310, Dec 2013.

Chapter 5

IDR2D identifies reproducible genomic interactions

The contents of this chapter were published as:

IDR2D identifies reproducible genomic interactions

Konstantin Krismer, Yuchun Guo, David K. Gifford

Nucleic Acids Research, Volume 48, Issue 6, 06 April 2020, Page e31

DOI: 10.1093/nar/gkaa030 | PMID: 32009147

Supplementary information can be found in Appendix D.

5.1 Abstract

Chromatin interaction data from protocols such as ChIA-PET, HiChIP, and Hi-C provide valuable insights into genome organization and gene regulation, but can include spurious interactions that do not reflect underlying genome biology. We introduce an extension of the Irreproducible Discovery Rate (IDR) method called IDR2D that identifies replicable interactions shared by chromatin interaction experiments. IDR2D provides a principled set of interactions and eliminates artifacts from single experiments. The method is available as a Bioconductor package for the R community, as well as an online service at https://idr2d.mit.edu.

5.2 Introduction

The Irreproducible Discovery Rate [15] (IDR) method identifies a robust set of findings that comprise the signal component shared by two replicate experiments. The IDR method is akin to the false discovery rate (FDR) in multiple hypothesis testing, but instead of alleviating alpha error accumulation within one replicate, IDR quantifies the reproducibility of findings using a copula mixture model with one reproducible and one irreproducible component. A finding's IDR is the probability it belongs to the irreproducible component. This permits findings that are likely in the irreproducible noise component to be eliminated for subsequent analyses. Assessing the IDR of genomic findings has been adopted by ENCODE [10], and is recommended for all ChIP-seq experiments with replicates [1]. IDR has also been used in numerous projects outside of ENCODE [4, 2, 24, 28, 17].

Chromatin interaction experiments such as ChIA-PET [6], HiChIP [18], and Hi-C [16] provide important chromatin structure and gene regulation information, but the complexity of their results and the sampling noise present in their protocols makes the principled analysis of resulting data important. Single replicate false discovery rate (FDR) methods are often used to identify chromatin interactions, but questions can remain about the veracity of the interactions identified as significant as they may not be observed in replicate experiments.

Here we generalize IDR from one dimensional analysis, performed on a single genome coordinate, to the analysis of interactions that are identified in two dimensions by a pair of genome coordinates. We call this extended method Irreproducible Discovery Rate for Two Dimensions (IDR2D) and it can be readily applied to any experimental data type that produces two-dimensional genomic results that admit appropriate distance metrics. We demonstrate the application of IDR2D to data from ChIA-PET, HiChIP, and Hi-C experiments.

Like IDR, IDR2D independently ranks the findings from each replicate. This ranking reflects the confidence of the finding, with high-confidence interactions at the top and low-confidence interactions at the bottom of the list. In a subsequent step, corresponding interactions between replicates are identified. A genomic interaction from replicate 1 is said to correspond to an interaction in replicate 2, if both their interaction anchors overlap (see Figure 5-1C). After corresponding interactions are identified and ambiguous mappings of interactions between replicates are resolved (see equation 5.1 and Figure 5-1D), IDRs are computed for each replicated interaction.

If interaction $x_{i,1}$ in replicate 1 overlaps with more than one interaction in replicate 2, the ambiguous mapping is resolved by choosing $x_{*,2}$ in the following way:

$$x_{*,2} = \underset{x_{j,2} \in \Omega_{x_{i,1}}}{\operatorname{argmin}} f(x_{i,1}, x_{j,2}), \tag{5.1}$$

where $\Omega_{x_{i,1}}$ is the set of interactions in replicate 2 that overlaps with the interaction $x_{i,1}$ in replicate 1, and $f(\cdot, \cdot)$ is the *ambiguity resolution value* (ARV) between an interaction in replicate 1 and an overlapping interaction in replicate 2. Depending on the ambiguity resolution method, this value corresponds to the genomic distance between anchor midpoints (see 1. in Figure 5-1D), the additive inverse of the relative anchor overlap (see 2. in Figure 5-1D), or the sum of the interaction ranks, where more significant interactions have lower ranks.

IDR2D is used as the final step in chromatin interaction data workflows (see Figure 5-1A). The input to IDR2D are BEDPE formatted files of genomic interactions, where each genomic interaction has a score associated with it. This score is used to rank the interactions and can be probability-based, such as the scores from MICC-based methods [8, 14, 7], or based on a heuristic. For Hi-C experiments, IDR2D supports the .hic file format from the Juicer / Straw pipeline and the .matrix/.bed file formats from the Hi-C Pro pipeline. Figure 5-1B breaks the IDR2D procedure into five steps.

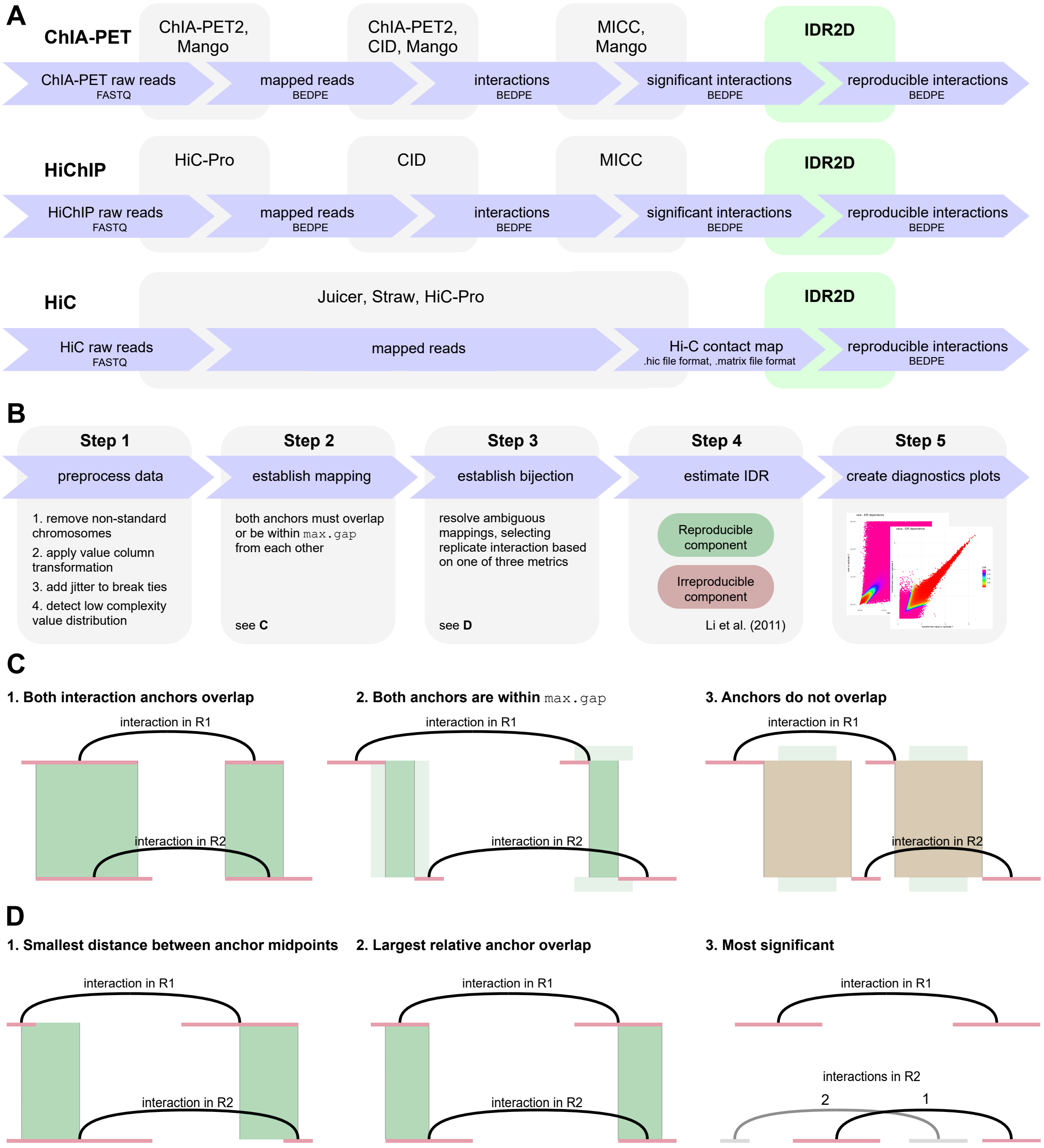


Figure 5-1: IDR2D identifies reproducible genomic interactions. (figure caption continued on next page)

Figure 5-1: IDR2D identifies reproducible genomic interactions. (**A**) IDR2D is a potential post-processing step in the data analysis pipelines for ChIA-PET, HiChIP, and Hi-C experiments that were done in replicate. It is compatible with a range of different interaction callers, such as ChIA-PET2, Mango, and CID. (**B**) This schematic depicts the five steps of the IDR2D procedure. In step 1, the data is prepared for IDR analysis, which includes the removal of interactions on non-standard chromosomes and a suitable transformation of the value column, which will be the basis of the ranking. In step 2, interactions in replicate 1 that overlap interactions in replicate 2 are identified, and in step 3 a one-to-one correspondence between overlapping interactions is established by resolving ambiguous cases. After this unambiguous mapping is established, in step 4 the irreproducible discovery rates are estimated for each interaction pair. Lastly, diagnostics plots are created in step 5. (**C**) An interaction in replicate 1 (R1) is assigned to all interactions in R2 for which both interaction anchors overlap or are within `maximum gap` of each other. (**D**) If more than one interaction in R2 overlaps with an interaction in R1, there are three methods to resolve this ambiguous mapping: select the interaction in R2 with (1) the smallest distance between the anchor midpoints (width of the green bars), (2) the largest relative overlap (width of the green bars divided by the sum of the anchor widths), or (3) the lowest rank sum of the interactions, which prioritizes more significant interactions.

5.3 Materials and Methods

5.3.1 IDR

IDR2D extends the reference implementation of IDR in R by Qunhua Li [15]. All datasets were analyzed with the IDR2D package using default parameters. We used *overlap* as ambiguity resolution method and allowed no gaps between overlapping interactions (`max.gap = 0L`). The applied value transformations were dependent on the interaction calling method. The results were not sensitive to the initial values of the mean, standard deviation, correlation coefficient, or proportion of the reproducible component.

5.3.2 ChIA-PET datasets

We used 17 ChIA-PET datasets associated with protein factors that include POL2RA, CTCF and RAD21 from selected cell types (Supplementary Table D.1). The datasets were downloaded from the ENCODE Project portal (`https://www.encodeproject.org/`). All FASTQ files of both biological replicates were pre-processed and aligned to the hg19 genome assembly using steps 1, 2 and 3 in the ChIA-PET data analysis software *Mango*.

5.3.3 HiChIP datasets

The FASTQ SMC1A HiChIP data from GM12878 cells [18] were downloaded from the NCBI GEO portal (GSE80820). Raw read files were analyzed with HiC-Pro [22], and interactions were subsequently called by CID and *hichipper* [12].

5.3.4 Hi-C datasets and subsampling procedure

Preprocessed contact matrix files in *.hic* format were downloaded from the NCBI GEO portal (Supplementary Table D.1) and parsed with *Straw*, a data extraction API for *.hic* files [5].

FASTQ files for Hi-C datasets from ENCODE were processed with the HiC-Pro pipeline [22] using default parameters for HindIII digested DNA. Contact matrices were normalized with the ICE procedure [9].

Subsampling of Hi-C contact matrices was performed using uniform sampling of individual reads without replacement.

5.3.5 Mango pipeline

Mango 1.2.1 [19] was downloaded from `https://github.com/dphansti/mango`. Additionally, we installed the dependencies *R* 3.4.4, *bedtools* 2.26.0, *macs2* (version 2.1.1.20160309) [29], and *bowtie-align* 1.2 [11]. Mango was executed with the default parameters and the flags `verboseoutput` and `reportallpairs` were set. For datasets that were generated with the ChIA-PET Tn5 tagmentation protocol, additional parameters recommended by the author were used: `-keepempty TRUE -maxlength 1000 -shortreads FALSE`. For subsequent IDR2D analyses, we used the `P` column in the Mango output files to establish the ranking of interactions. This column contains unadjusted p-values, which were transformed using the `log.additive.inverse` transformation to match the IDR semantics of the value column, where interactions with larger values are more likely to be true interactions.

The BEDPE files generated by Mango after step 3 were also used by the ChIA-PET2 and CID pipelines.

5.3.6 ChIA-PET2 pipeline

ChIA-PET2 0.9.2 [14] was obtained from `https://github.com/GuipengLi/ChIA-PET2`. The default setting was used for all parameters, except that the starting step was set to 4 to start the analysis from Mango-derived BEDPE files. The ranking for the IDR2D analysis was established by the untransformed `-log10(1-PostProb)` column, which is an output of *MICC* [8], a Bayesian mixture model used internally by ChIA-PET2 and *CID*.

5.3.7 CID pipeline

CID 1.0 [7] is part of the Java genomics software package *GEM* 3.4, which was downloaded from `https://groups.csail.mit.edu/cgs/gem/versions.html`. We used the default CID parameters. Before running MICC, we filtered all interactions that were supported by only one PET read. Same as with ChIA-PET2, we used the untransformed `-log10(1-PostProb)` column to rank interactions in IDR2D.

5.3.8 Package and web development

The R package development process was supported using *devtools*. We used *roxygen2* for inline function documentation, and *knitr* and *R Markdown* for package vignettes. With the R package we provide a platform-independent implementation of the methods introduced in this paper. The Hi-C analysis part of the package requires the Python package *hic-straw* [5], which is a data extraction API for Hi-C contact maps.

The website was developed in R with the reactive web application framework *shiny* from RStudio. The components of the graphical user interface were provided by shiny and *shinyBS*, which serve as an R wrapper for the components of the Bootstrap front-end web development framework. The analysis job queue of the website uses an *SQLite* database.

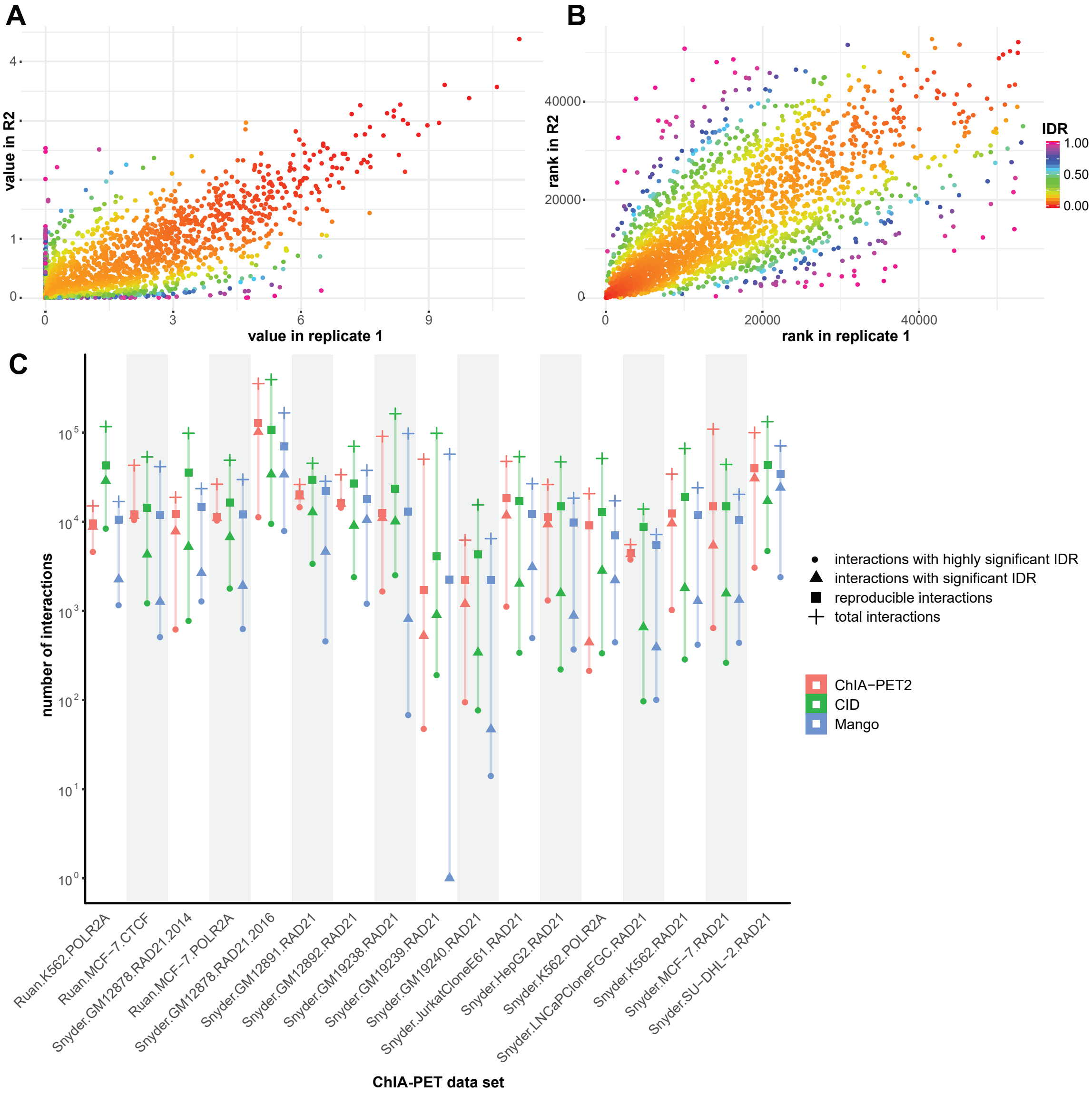


Figure 5-2: IDR2D analysis of 17 replicated ChIA-PET experiments identifies reproducible components. (figure caption continued on next page)

Figure 5-2: IDR2D analysis of 17 replicated ChIA-PET experiments identifies reproducible components. (**A**) Diagnostic scatterplot of IDRs of genomic interactions called by CID from replicated ChIA-PET experiments targeting RAD21 in HepG2 cells. Plotted are replicated interactions with their estimated IDR (color) and their scores in the two replicates (position). As expected, interactions with low IDRs that have a low probability of belonging to the irreproducible component, are along the diagonal (similar scores in both replicates). (**B**) Similar to panel **A**, but plots interaction ranks instead of scores (higher score results in lower rank). (**C**) A comparison of ChIA-PET interaction callers ChIA-PET2, CID, and Mango across 17 ChIA-PET experiments. Significant IDR < 0.05, highly significant IDR < 0.01, *total interactions* is the number of interactions in replicate 1.

5.4 Results

5.4.1 IDR2D identifies reproducible components of ENCODE ChIA-PET experiments

To assess the performance and utility of IDR2D we analyzed the read data of 17 ChIA-PET experiments that had replicates (see Supplementary Table D.1). Mango was used for data preprocessing such as linker removal, read mapping (via bowtie), and peak calling (via macs2). We called interactions with three different methods (ChIA-PET2, CID, and Mango) and then used IDR2D to identify reproducible interactions across replicates. The number of identified interactions varies greatly between the three interaction callers, with on average CID identifying the most, and Mango the fewest interactions (see Figure 5-2C and Supplementary Tables D.2, D.3, and D.4). As a result, the overall reproducibility of interactions is also dependent on the interaction caller. For example, the ChIA-PET experiments `Snyder.GM19239.RAD21` and `Snyder.GM19240.RAD21` show poor reproducibility across all three interaction calling methods. By identifying the reproducible component within each of the replicated experiments, IDR2D helps to assess the overall reproducibility of each experiment, as

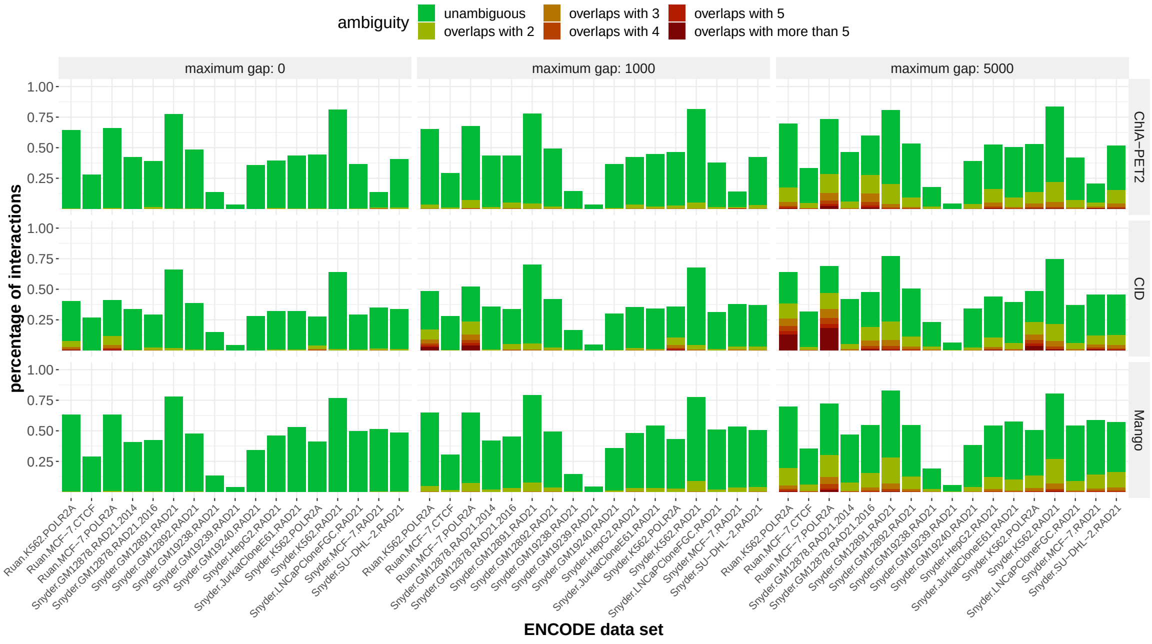


Figure 5-3: Mappings of genomic interactions between replicated ChIA-PET experiments are predominantly unambiguous. The great majority of interactions in replicate 1 that overlapped with interactions in replicate 2 only overlapped with one interaction, leading to an unambiguous assignment of corresponding replicated interactions (green bars). Unsurprisingly, the number of ambiguous mappings (interactions in replicate 1 that overlap with more than one interaction in replicate 2) increases when the maximum acceptable gap is increased, the tolerated distance between anchors to still be considered overlapping.

well as the reproducibility of individual findings, which in turn informs the conclusions drawn from the data. In addition, it can be used to help qualify new experimental protocols for consistent results. Venn diagrams depicting the overlap of identified interactions between ChIA-PET2, CID, and Mango are shown in Supplementary Figure D-1.

Furthermore, we used IDR2D to analyze experimental results from replicated HiChIP (see Supplementary Tables D.5 and D.6). Similar to ChIA-PET, IDR2D can identify reproducible HiChIP interactions and expose poorly replicated experiments, which is valuable information for subsequent analysis steps.

5.4.2 Mappings of genomic interactions between replicated ChIA-PET experiments are predominantly unambiguous

The great majority of interactions in replicate 1 that overlapped with interactions in replicate 2 overlapped with only one interaction, leading to an unambiguous assignment of corresponding replicated interactions (see Figure 5-3). Unsurprisingly, the number of ambiguous mappings (interactions in replicate 1 that overlap with more than one interaction in replicate 2) increases when the *maximum gap* is increased, the tolerated distance between anchors that are considered to overlap. On average, only 2.66% of interactions are ambiguous in the case of zero *max gap*, whereas this number increases to 8.00% and 24.11% with maximum gaps of 1000 and 5000 bp, respectively.

There are more ambiguous mappings between replicated interactions that were called with CID (14.73% for CID, 9.90% for ChIA-PET2, and 10.14% for Mango). We expect this is because (1) CID on average calls significantly more interactions than ChIA-PET2 and Mango, and (2) interactions called with CID exhibit a wider range of anchor lengths, and longer anchors naturally increase the probability of overlap.

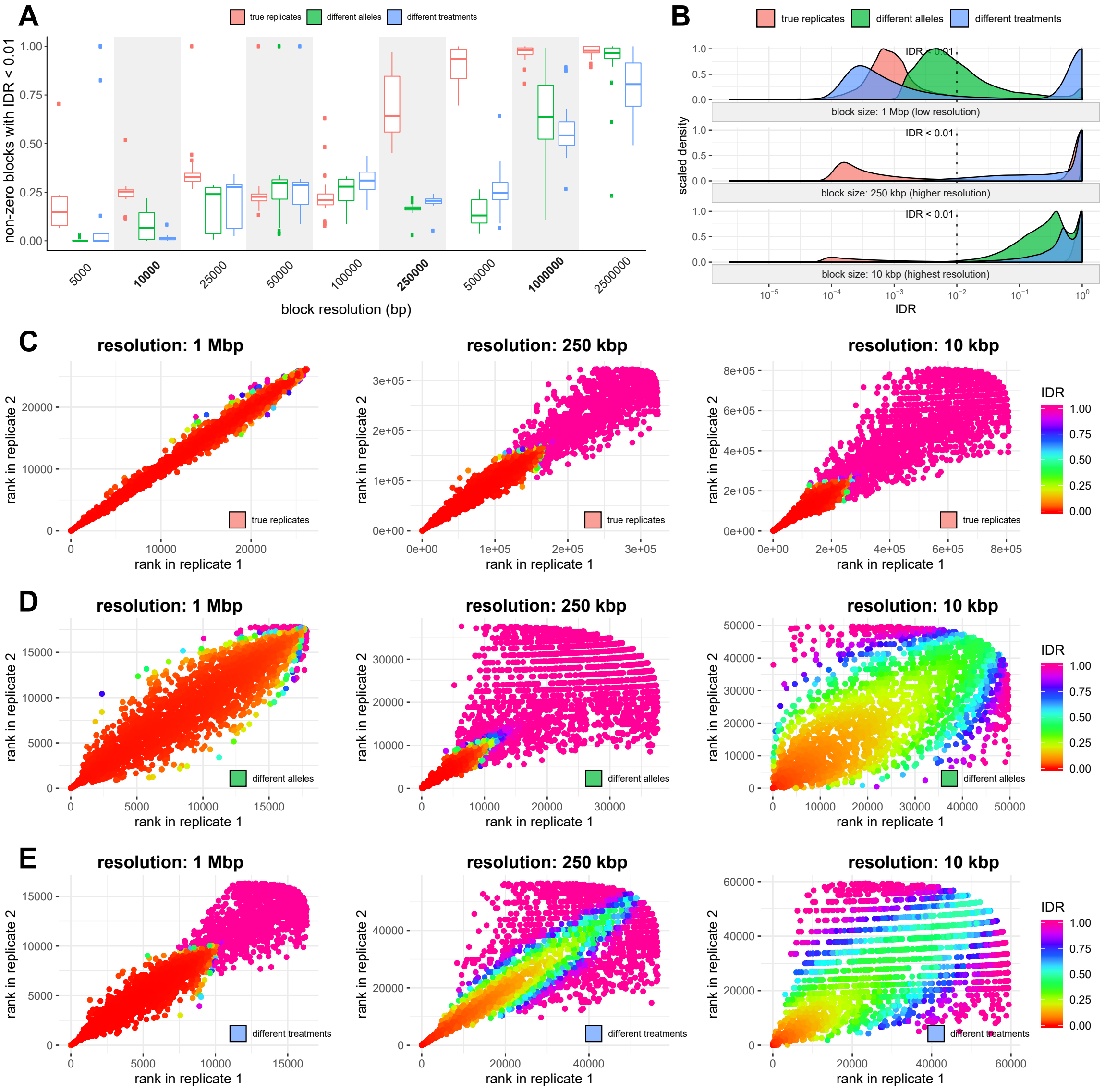


Figure 5-4: Reproducibility analysis of Hi-C experiments. (figure caption continued on next page)

Figure 5-4: Reproducibility analysis of Hi-C experiments. (**A**) Summary of IDR2D results on individual chromosomes of three pairs of Hi-C experiments, True replicate Hi-C experiments (Lieberman.GM12878) are compared to IDR2D analysis of Hi-C experiments of different alleles (Lieberman.Patski) and different treatments (Skok.NSD2). (**B**) Histograms of the IDR distribution of IDR values for all blocks of chromosome 1 for the three pairs of Hi-C experiments. (**C**) Scatterplots of block ranks of chromosome 1 of the two Hi-C replicate experiments, colored by IDR. (**D**) Analagous to **C**, for Hi-C experiments of paternal and maternal alleles. (**E**) Analagous to **C**, for Hi-C experiments before and after overexpression of NSD2. Axis scales are not fixed between scatterplots.

5.4.3 Assessing reproducibility of Hi-C experiments

When analyzing pairs of Hi-C experiments with IDR2D, blocks from Hi-C contact matrices are used as observations. The resolution of contact matrix values typically ranges between 5 kbp (kilo base pairs) to 2.5 Mbp blocks. With the fixed grid of contact map observations, finding corresponding observations in the second replicate is straightforward. Each block in replicate 1 is simply matched with the block spanning the same genomic regions in replicate 2. Blocks are subsequently ranked by their read counts and analyzed using the same procedure that was used for ChIA-PET and HiChIP data.

In addition to computing IDR values, IDR2D produces diagnostic plots that help interpret the overall reproducibility of a pair of Hi-C experiments, as well as identify reproducible parts of Hi-C contact matrices for a more focused, downstream analysis.

In Figure 5-4 we show IDR2D results for three pairs of Hi-C experiments. The first pair of Hi-C experiments consists of true replicate experiments in GM12878 cells [21]. The second pair of experiments were obtained in phased murine embryonic kidney fibroblasts, where allele-specific Hi-C reads [21] were available (*different alleles* in Figure 5-4) [3], and the third pair of Hi-C experiments were obtained before and after the overexpression of NSD2 (different treatments) [13]. GEO identifiers of all data sets are listed in the Supplementary Table D.1 and detailed results in Supplementary Table S7. Figure 5-4A gives an overview of all data sets and all resolutions, showing that, as expected, the reproducibility is highest between true replicates, and in general higher at lower resolutions (larger blocks). Figure 5-4B depicts the distribution of IDR values for chromosome 1 of each of the Hi-C pairs at block resolutions of 1 Mbp, 250 kbp, and 10 kbp. The largest fraction of reproducible blocks is found between replicated experiments. Figures 5-4C-E are scatterplots of interaction pairs (corresponding blocks in the contact matrices) of the two experiments, where the color denotes the IDR value of the interaction pair. Given a Hi-C experiment with a fixed sequencing depth, the higher the resolution of the Hi-C analysis the less reproducible the individual interactions will be as a consequence of sampling noise.

Not all Hi-C interaction pairs that lie on the diagonal and have similar ranks in both replicates are deemed reproducible by IDR2D. For example, see the upper right panel of Figure 5-4C. This lack of reproducibility is intended and is justified by taking into account the poor reproducibility of other interaction pairs with similar ranks. Hi-C contacts close to the diagonal can be found irreproducible when they are in rank neighborhoods of irreproducibility. We note that while experiment level methods may find a Hi-C experiment reproducible, IDR2D may find a specific interaction irreproducible because it is in a rank neighborhood that is not reproducibly ordered. IDR2D may require increased sequencing depth to consistently rank interactions to ascertain reproducibility of such interactions.

IDR2D is largely insensitive to sequencing depth when it is sufficient to recover contacts, thus reproducible pairs of experiments are identified as such even if their sequencing depths are different. Reproducibility as measured by IDR2D only starts to degrade significantly at extremely low coverage, with only very few reads (single to low double digits) per block. Subsampling experiments were performed to illustrate this behavior (see Supplementary Figure D-2).

5.5 Discussion

The appropriate choice of significance values for the computation of interaction ranks depends on the method used to identify contacts. As a general rule, larger values should reflect higher confidence and there should be as few ties as possible. IDR2D operates on the ranks of the interactions in both replicates and therefore is invariant to order-preserving transformations of the original significance values. If p-values are used as significance values for interactions, the additive inverse or the log additive inverse of uncorrected p-values is recommended. Unadjusted p-values are preferred over p-values adjusted for multiple hypothesis testing, because uncorrected p-values reduce rank ties.

Other methods assess the overall reproducibility of genome interaction experiments but do not characterize the reproducability of each reported contact. Such methods include HiCRep [27], HiC-spector [26], and GenomeDISCO [23]. GenomeDISCO also supports experimental data from ChIA-PET and HiChIP. HiCRep calculates a score of experiment reproducibility between contact matrices based on aggregated stratified Pearson correlation coefficients, while HiC-spector determines experiment reproducibility by comparing the eigenvectors of a spectral decomposition of the contact maps, and GenomeDISCO's concordance score is based on random walks on a graph representation of contact maps. These methods have in common that they assess the overall reproducibility of replicated experiments with a global measure of similarity between contact matrices. IDR2D provides a measure of reproducibility for each reported contact and then summarizes these findings to characterize experiment reproducability (see Supplementary Figure D-3). IDR2D's fine-grained analysis of reproducibility identifies contacts that are invariant across experimental replicates and those that are not, which is a unique capability. Thus, IDR2D is intended to complement, rather than replace previous Hi-C reproducibility assessment methods.

IDR2D, and the methods mentioned above, are limited to comparisons of two replicates at a time. If more replicates are available, multiple pairwise analysis can be performed and the results combined.

While IDR2D is a compatible post-processing step for the tested interaction callers, it cannot recover true interactions that were discarded by the interaction caller and therefore the identified set of reproducible interactions is always limited by the sensitivity of the caller.

IDR2D can potentially support single-cell or single-molecule chromatin interaction data obtained by methods such as Sci-Hi-C [20] and ChIA-Drop [30]. However, the sparsity of interaction data from single cells will necessitate data imputation or cell clustering as preprocessing steps to IDR2D, similar to strategies applied to single-cell ATAC-seq data [25].

IDR2D offers a complementary way to evaluate the results of chromatin interaction experiments for significance, and provides a foundation for subsequent analysis such as enhancer-gene mapping that incorporates the important concept of experimental replicability.

Availability

The implementation of IDR2D facilitates workflow integration with other data analysis pipelines, and is also web-accessible at `https://idr2d.mit.edu`. IDR2D is implemented in R and bundled as an R/Bioconductor package (idr2d), supporting observations with both one-dimensional and two-dimensional genomic coordinates. The IDR2D website implementation offers a number of ways to transform the scores to match IDR requirements, and to map interactions between replicates. The source code of the R package is hosted on GitHub (`https://github.com/gifford-lab/idr2d`).

Bibliography

[1] T. Bailey, P. Krajewski, I. Ladunga, C. Lefebvre, Q. Li, T. Liu, P. Madrigal, C. Taslim, and J. Zhang. Practical guidelines for the comprehensive analysis of ChIP-seq data. *PLoS Comput. Biol.*, 9(11):e1003326, 2013.

[2] X. Chen, D. Iliopoulos, Q. Zhang, Q. Tang, M. B. Greenblatt, M. Hatziapostolou, E. Lim, W. L. Tam, M. Ni, Y. Chen, J. Mai, H. Shen, D. Z. Hu, S. Adoro, B. Hu, M. Song, C. Tan, M. D. Landis, M. Ferrari, S. J. Shin, M. Brown, J. C. Chang, X. S. Liu, and L. H. Glimcher. XBP1 promotes triple-negative breast cancer by controlling the HIF1-alpha pathway. *Nature*, 508(7494):103–107, Apr 2014.

[3] E. M. Darrow, M. H. Huntley, O. Dudchenko, E. K. Stamenova, N. C. Durand, Z. Sun, S. C. Huang, A. L. Sanborn, I. Machol, M. Shamim, A. P. Seberg, E. S. Lander, B. P. Chadwick, and E. L. Aiden. Deletion of DXZ4 on the human inactive X chromosome alters higher-order genome architecture. *Proc. Natl. Acad. Sci. U.S.A.*, 113(31):E4504–4512, 08 2016.

[4] S. Djebali, C. A. Davis, A. Merkel, A. Dobin, T. Lassmann, A. Mortazavi, A. Tanzer, J. Lagarde, W. Lin, F. Schlesinger, C. Xue, G. K. Marinov, J. Khatun, B. A. Williams, C. Zaleski, J. Rozowsky, M. Roder, F. Kokocinski, R. F. Abdelhamid, T. Alioto, I. Antoshechkin, M. T. Baer, N. S. Bar, P. Batut, K. Bell, I. Bell, S. Chakrabortty, X. Chen, J. Chrast, J. Curado, T. Derrien, J. Drenkow, E. Dumais, J. Dumais, R. Duttagupta, E. Falconnet, M. Fastuca, K. Fejes-Toth, P. Ferreira, S. Foissac, M. J. Fullwood, H. Gao, D. Gonzalez, A. Gordon, H. Gunawardena, C. Howald, S. Jha, R. Johnson, P. Kapranov, B. King, C. Kingswood, O. J. Luo, E. Park, K. Persaud, J. B. Preall, P. Ribeca, B. Risk, D. Robyr,

M. Sammeth, L. Schaffer, L. H. See, A. Shahab, J. Skancke, A. M. Suzuki, H. Takahashi, H. Tilgner, D. Trout, N. Walters, H. Wang, J. Wrobel, Y. Yu, X. Ruan, Y. Hayashizaki, J. Harrow, M. Gerstein, T. Hubbard, A. Reymond, S. E. Antonarakis, G. Hannon, M. C. Giddings, Y. Ruan, B. Wold, P. Carninci, R. Guigo, and T. R. Gingeras. Landscape of transcription in human cells. *Nature*, 489(7414):101–108, Sep 2012.

[5] N. C. Durand, J. T. Robinson, M. S. Shamim, I. Machol, J. P. Mesirov, E. S. Lander, and E. L. Aiden. Juicebox Provides a Visualization System for Hi-C Contact Maps with Unlimited Zoom. *Cell Syst*, 3(1):99–101, 07 2016.

[6] M. J. Fullwood and Y. Ruan. ChIP-based methods for the identification of long-range chromatin interactions. *J. Cell. Biochem.*, 107(1):30–39, May 2009.

[7] Y. Guo, K. Krismer, M. Closser, H. Wichterle, and D. K. Gifford. High resolution discovery of chromatin interactions. *Nucleic Acids Res.*, 47(6):e35, 04 2019.

[8] C. He, M. Q. Zhang, and X. Wang. MICC: an R package for identifying chromatin interactions from ChIA-PET data. *Bioinformatics*, 31(23):3832–3834, Dec 2015.

[9] M. Imakaev, G. Fudenberg, R. P. McCord, N. Naumova, A. Goloborodko, B. R. Lajoie, J. Dekker, and L. A. Mirny. Iterative correction of Hi-C data reveals hallmarks of chromosome organization. *Nat. Methods*, 9(10):999–1003, Oct 2012.

[10] Stephen G. Landt, Georgi K. Marinov, Anshul Kundaje, Pouya Kheradpour, Florencia Pauli, Serafim Batzoglou, Bradley E. Bernstein, Peter Bickel, James B. Brown, Philip Cayting, Yiwen Chen, Gilberto DeSalvo, Charles Epstein, Katherine I. Fisher-Aylor, Ghia Euskirchen, Mark Gerstein, Jason Gertz, Alexander J. Hartemink, Michael M. Hoffman, Vishwanath R. Iyer, Youngsook L. Jung, Subhradip Karmakar, Manolis Kellis, Peter V. Kharchenko, Qunhua Li, Tao Liu, X. Shirley Liu, Lijia Ma, Aleksandar Milosavljevic, Richard M. Myers, Peter J. Park, Michael J. Pazin, Marc D. Perry, Debasish Raha, Timothy E. Reddy, Joel Rozowsky, Noam Shoresh, Arend Sidow, Matthew Slattery, John A. Stamatoyannopoulos, Michael Y. Tolstorukov, Kevin P. White, Simon Xi, Peggy J.

Farnham, Jason D. Lieb, Barbara J. Wold, and Michael Snyder. Chip-seq guidelines and practices of the encode and modencode consortia. *Genome Research*, 22(9):1813–1831, 2012.

[11] B. Langmead, C. Trapnell, M. Pop, and S. L. Salzberg. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.*, 10(3):R25, 2009.

[12] C. A. Lareau and M. J. Aryee. hichipper: a preprocessing pipeline for calling DNA loops from HiChIP data. *Nat. Methods*, 15(3):155–156, 02 2018.

[13] P. Lhoumaud, S. Badri, J. Rodriguez-Hernaez, T. Sakellaropoulos, G. Sethia, A. Kloetgen, M. Cornwell, S. Bhattacharyya, F. Ay, R. Bonneau, A. Tsirigos, and J. A. Skok. NSD2 overexpression drives clustered chromatin and transcriptional changes in a subset of insulated domains. *Nat Commun*, 10(1):4843, Oct 2019.

[14] G. Li, Y. Chen, M. P. Snyder, and M. Q. Zhang. ChIA-PET2: a versatile and flexible pipeline for ChIA-PET data analysis. *Nucleic Acids Res.*, 45(1):e4, 01 2017.

[15] Qunhua Li, James B. Brown, Haiyan Huang, and Peter J. Bickel. Measuring reproducibility of high-throughput experiments. *Ann. Appl. Stat.*, 5(3):1752–1779, 09 2011.

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

[17] Pedro Madrigal. Cexor: an r/bioconductor package to uncover high-resolution protein-dna interactions in chip-exo replicates. *EMBnet.journal*, 21(0):837, 2015.

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

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

[20] V. Ramani, X. Deng, R. Qiu, C. Lee, C. M. Disteche, W. S. Noble, J. Shendure, and Z. Duan. Sci-Hi-C: A single-cell Hi-C method for mapping 3D genome organization in large number of single cells. *Methods*, 170:61–68, Jan 2020.

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

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

[23] O. Ursu, N. Boley, M. Taranova, Y. X. R. Wang, G. G. Yardimci, W. Stafford Noble, and A. Kundaje. GenomeDISCO: a concordance score for chromosome conformation capture experiments using random walks on contact map graphs. *Bioinformatics*, 34(16):2701–2707, 08 2018.

[24] H. Wang, M. T. Maurano, H. Qu, K. E. Varley, J. Gertz, F. Pauli, K. Lee, T. Canfield, M. Weaver, R. Sandstrom, R. E. Thurman, R. Kaul, R. M. Myers, and J. A. Stamatoyannopoulos. Widespread plasticity in CTCF occupancy linked to DNA methylation. *Genome Res.*, 22(9):1680–1688, Sep 2012.

[25] L. Xiong, K. Xu, K. Tian, Y. Shao, L. Tang, G. Gao, M. Zhang, T. Jiang, and Q. C. Zhang. SCALE method for single-cell ATAC-seq analysis via latent feature extraction. *Nat Commun*, 10(1):4576, Oct 2019.

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

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

[28] F. Zanconato, M. Forcato, G. Battilana, L. Azzolin, E. Quaranta, B. Bodega, A. Rosato, S. Bicciato, M. Cordenonsi, and S. Piccolo. Genome-wide association between YAP/TAZ/TEAD and AP-1 at enhancers drives oncogenic growth. *Nat. Cell Biol.*, 17(9):1218–1227, Sep 2015.

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

[30] M. Zheng, S. Z. Tian, D. Capurso, M. Kim, R. Maurya, B. Lee, E. Piecuch, L. Gong, J. J. Zhu, Z. Li, C. H. Wong, C. Y. Ngan, P. Wang, X. Ruan, C. L. Wei, and Y. Ruan. Multiplex chromatin interactions with single-molecule precision. *Nature*, 566(7745):558–562, 02 2019.

Chapter 6

spatzie: An R package for identifying significant transcription factor motif co-enrichment from enhancer-promoter interactions

The contents of this chapter have been submitted for publication and a preprint is available at:

spatzie: An R package for identifying significant transcription factor motif co-enrichment from enhancer-promoter interactions
Jennifer Hammelman [*], Konstantin Krismer [*], David K. Gifford
bioRxiv 2021.05.25.445606
DOI: 10.1101/2021.05.25.445606

Supplementary information can be found in Appendix E.

[*] These authors contributed equally to this work.

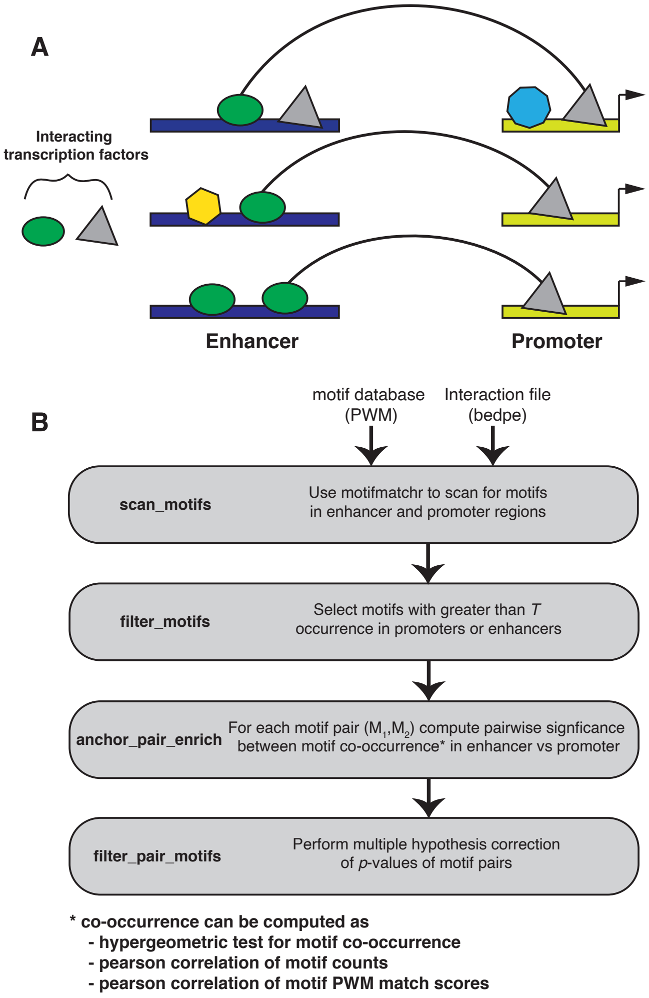
6.1 Abstract

Genomic interactions provide important context to our understanding of the state of the genome. One question is whether specific transcription factor interactions give rise to genome organization. We introduce *spatzie*, an R package and a website that implements statistical tests for significant transcription factor motif cooperativity between enhancer-promoter interactions. We conducted controlled experiments under realistic simulated data from ChIP-seq to confirm spatzie is capable of discovering co-enriched motif interactions even in noisy conditions. We then use spatzie to investigate cell type specific transcription factor cooperativity within recent human ChIA-PET enhancer-promoter interaction data. The method is available online at `https://spatzie.mit.edu`.

6.2 Introduction

Genome organization plays an important role in the function of the genome in development [19, 6, 21] and disease [5]. Specific transcription factor cooperation is a potential explanation for the cell type specificity of genomic interactions, especially those that tether enhancers to promoters. Recent methods seek to detect such transcription factor cooperativity by generating models to predict enhancer-promoter interactions and measuring the importance of model features [33, 30]. However, these methods can be difficult to interpret either due to the complexity of model choice or the use of shrinkage techniques that could eliminate correlated features.

Here we introduce *spatzie*, an R/Bioconductor package named after the German diminutive for sparrow and inspired by the long-range geographical patterns of their songs [32], a reference to the long-range genomic interactions of transcription factor cooperativity. Within spatzie we implement a collection of statistical methods to identify transcription factor co-enrichment in experimental data obtained by protein-centric chromatin conformation methods such as ChIA-PET [10] and HiChIP [28]. We demonstrate the utility of spatzie by discovering the co-enrichment of transcription factor binding motifs simulated from ChIP-seq data. Furthermore, we apply spatzie to investigate cell type-specific interactions from RAD21-targeted ChIA-PET experiments across 24 human cell lines.


A
Interacting transcription factors
Enhancer
Promoter
B
motif database (PWM)
Interaction file (bedpe)
scan_motifs
Use motifmatchr to scan for motifs in enhancer and promoter regions
filter_motifs
Select motifs with greater than *T* occurrence in promoters or enhancers
anchor_pair_enrich
For each motif pair (M_1,M_2) compute pairwise signficance between motif co-occurrence* in enhancer vs promoter
filter_pair_motifs
Perform multiple hypothesis correction of *p*-values of motif pairs

*** co-occurrence can be computed as**

- **hypergeometric test for motif co-occurrence**
- **pearson correlation of motif counts**
- **pearson correlation of motif PWM match scores**

Figure 6-1: spatzie identifies motif pairs underlying enhancer-promoter interactions using co-occurrence and correlation statistics. **(A)** spatzie is designed to identify transcription factors which are facilitating interactions between enhancers and promoters based on detecting co-enrichment relationships between the presence of DNA-binding motifs in enhancer-promoter pairs. **(B)** Given input of a database of transcription factor DNA-binding motifs and a set of enhancer-promoter interactions, scan interactions for motifs, then limit analysis by filtering to motifs that are frequently present within the interactions of interest. Next we compute pairwise significance of motif co-occurrence in the enhancer and promoter data. Finally, we filter motif pairs that significantly co-occur under multiple hypothesis correction.

6.3 Materials and Methods

6.3.1 ChIP-seq data for simulated co-enrichment

We simulate a cooperative relationship where binding of USF1 at promoters is co-dependent on binding of ELF1 at enhancers. Raw ChIP-seq data for USF1 and ELF1 from MEL mouse cells was downloaded from ENCODE (Supplementary Table E.1). Reads were trimmed for adaptors and low-quality positions using Trimgalore (Cutadapt v0.6.2) [26] and aligned to the mouse genome (mm10) with bwa mem (v0.7.1.7) [23] with default parameters. Duplicates were removed with samtools (v1.7.2) [24] markdup, and ChIP binding events were called with GPS (v3.4) [12] with default parameters.

6.3.2 ChIA-PET datasets

ChIA-PET interaction data was downloaded as processed files from Grubert et al. 2020 (see section E.3). Raw data for all experiments is accessible from GEO (Supplementary Table E.1).

6.3.3 Genomic annotations

For simulated data, mm10 promoter annotations were downloaded from the UCSC browser using the R package *GenomicFeatures* [22]. For human RAD21 ChIA-PET, hg19 promoter ensemble gene annotations were also downloaded from the UCSC browser using the same method. Within interaction data, regions that were within 2.5 kb of a promoter were classified as promoter regions. All other regions were classified as gene-distal enhancer regions.

6.3.4 Statistical cooperativity calculation with spatzie

We implement three methods to measure the relationship between transcription factor binding motifs in promoter and enhancer regions of genomic interactions. Each method takes as input two vectors, $\boldsymbol{x} = (x_1, x_2, \ldots, x_{n-1}, x_n)$ and $\boldsymbol{y} = (y_1, y_2, \ldots, y_{n-1}, y_n)$, where n is the number of enhancer-promoter interactions. x_i is a set that contains all PWM scores for motif a in the promoter region of interaction i. y_i, in contrast, contains those scores for motif b in the enhancer region of interaction i.

Score-based correlation coefficient

We assume motif scores follow a normal distribution and are independent between enhancers and promoters. We can therefore compute how correlated scores of any two transcription factor motifs are between enhancer and promoter regions using Pearson's product-moment correlation coefficient:

$$r = \frac{\sum(x'_i - \bar{x}')(y'_i - \bar{y}')}{\sqrt{\sum(x'_i - \bar{x}')^2 \sum(y'_i - \bar{y}')^2}},$$

where the input vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ from above are transformed to vectors $\boldsymbol{x'}$ and $\boldsymbol{y'}$ by replacing the set of scores with the maximum score for each region:

$$x'_i = \max x_i$$

x'_i is then the maximum motif score of motif a in the promoter region of interaction i, y'_i is the maximum motif score of motif b in the enhancer region of interaction i, and $\bar{x}'$ and $\bar{y}'$ are the sample means.

Significance is then computed by transforming the correlation coefficient r to test statistic t, which is Student t-distributed with $n - 2$ degrees of freedom.

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

All p-values are calculated as one-tailed p-values of the probability that scores are greater than or equal to r.

Count-based correlation coefficient

Instead of calculating the correlation of motif scores directly, the count-based correlation metric first tallies the number of instances of a given motif within an enhancer or a promoter region, which are defined as all positions in those regions with motif score p-values of less than $5 * 10^{-5}$, which tends to work well for human and mouse motifs [31, 18]. Formally, the input vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ are transformed to vectors $\boldsymbol{x''}$ and $\boldsymbol{y''}$ by replacing the set of scores with the cardinality of the set:

$$x''_i = |x_i|$$

And analogous for y''_i. Finally, the correlation coefficient r between $\boldsymbol{x''}$ and $\boldsymbol{y''}$ and its associated significance are calculated as described above.

Instance co-occurrence

Instance co-occurrence (or match association) uses the presence or absence of a motif within an enhancer or promoter to determine a statistically significant association, thus $\boldsymbol{x'''}$ and $\boldsymbol{y'''}$ are defined by:

$$x'''_i = \mathbb{1}_{x''_i > 0}$$

The significance of instance co-occurrence is determined by the hypergeometric test:

$$p = \sum_{k=I_{ab}}^{P_a} \frac{\binom{P_a}{k}\binom{n-P_a}{E_b-k}}{\binom{n}{E_b}},$$

where I_{ab} is the number of interactions that contain a match for motif a in the promoter and motif b in the enhancer, P_a is the number of promoters that contain motif a ($P_a = \sum_i^n x'''_i$), E_b is the number of enhancers that contain motif b ($E_b = \sum_i^n y'''_i$), and n is the total number of interactions, which is equal to the number of promoters and to the number of enhancers.

Multiple hypothesis testing

While the R package spatzie supports several methods to adjust p-values, three to control the family-wise error rate or FWER (Holm's method [16], Hochberg's method [15], Bonferroni's method [7]) and two to control the false discovery rate or FDR (Benjamini and Hochberg's method [1], and Benjamini and Yekutieli's method [2]), all p-values presented in this work were corrected using the method of Benjamini and Hochberg.

6.4 Results

6.4.1 spatzie tests transcription factor motifs for co-enrichment in enhancer-promoter interactions

The goal of spatzie is to identify pairs of transcription factor motifs which have a relationship such that the presence of motif A in an enhancer is associated with the presence of motif B in the promoter, indicating these transcription factors may be cooperating to drive enhancer-promoter interactions (Figure 6-1A). Given an input of interacting genomic loci, we select only those interactions which contain one locus that is gene-distal, which we label enhancer, and one locus overlapping a gene transcription start site, which we label promoter. Then, we scan these regions for transcription factor motifs using a database of DNA-binding motifs identified by ChIP-seq experiments, such as HOCOMOCO [20], HOMER [14], or JASPAR [9]. In order to limit hypothesis testing, spatzie provides a function to filter transcription factor motifs to those present in some threshold number of interactions. After filtering, we test transcription factor motifs pairwise for co-enrichment between enhancer and promoter pairs. Since the relationship between the DNA-binding motif and transcription factor activity is complex, spatzie provides three possibilities: 1) the strength of the transcription factor motif match (i.e., the PWM score), 2) the number of motif sites within the sequence, and 3) the presence or absence of motif sites. These definitions result in three different statistical tests for co-enrichment: the significance tests of 1) score-based or 2) count-based correlation coefficients, and 3) the hypergeometric test for co-occurrence over-representation (Figure 6-1B). Finally, we adjust the significance of these co-enrichment scores to account for multiple hypothesis testing and report significant transcription factor pairs.

6.4.2 spatzie identifies co-enrichment from simulated data

We validate spatzie by simulating a co-enrichment relationship where binding of USF1 at promoters is co-dependent on binding of ELF1 at enhancers. Using ELF1 and USF1 ChIP-seq data from ENCODE, we aligned and called binding events with GPS [12]. We then filtered USF1 binding sites to those that overlapped annotated promoters and filtered ELF1 binding events to any event that did not overlap a promoter. Then, we matched the most significant USF1 promoter event to the most significant ELF1 enhancer event, thus creating a simulated enhancer-promoter interaction data set where the strongest USF1 promoter events are matched with ELF1 enhancer events. We found that the three described methods (score-based correlation, count-based correlation, and motif presence/absence association) all result in significant co-occurrence between the USF1 motif and the ELF1 motif (Figure 6-2A-C). The DNA binding motifs of all other transcription factors with significant co-enrichment are highly similar to USF1 and ELF1 (Supplementary Figure E-1).

We also test that spatzie performs under noisy experimental conditions. We take the top N percent of enhancer interactions and randomly permute them such that they are paired with new promoters and then run spatzie using score-based correlation, count-based correlation, and motif presence/absence association (Figure 6-2D-F). We find that all methods collapse after 75% and 100% of the enhancers have been randomly permuted, indicating that co-occurrence is a result of the strength of the co-occurrence of the DNA binding motifs underlying the simulated enhancer-promoter interaction data.

6.4.3 spatzie identifies germ layer and tissue-specific enhancer-promoter transcription factor interactions

Finally, we investigate enhancer-promoter interactions from RAD21 ChIA-PET of 24 human cell lines from ENCODE [11]. Based on the most significant co-enrichment scores on simulated data coming from the score correlation method, we chose to use score correlation to investigate transcription factor motif co-occurrence in enhancer-

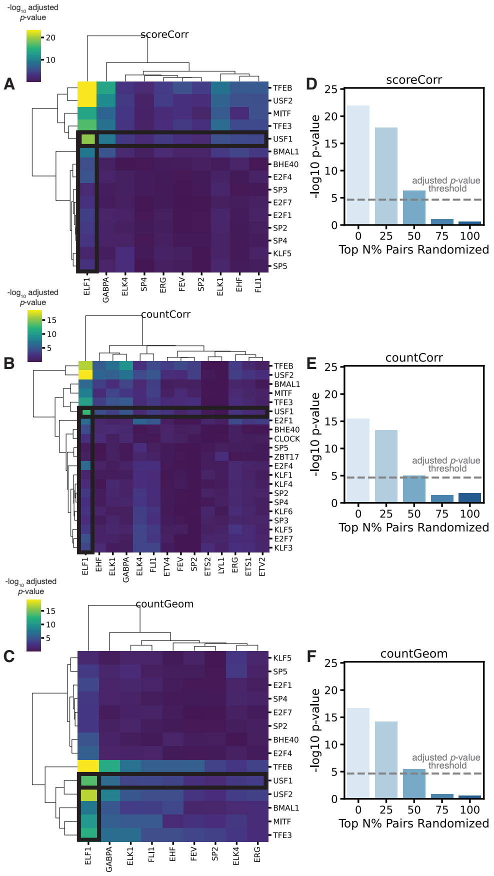


Figure 6-2: spatzie validates co-enrichment of ELF1 and USF1 on simulated enhancer-promoter interaction data. **(A)** spatzie cooperativity estimation computed using correlation of motif scores shows significant enhancer-promoter interactions for USF1 and ELF1 motifs. **(B)** spatzie cooperativity estimation computed using correlation of counts shows strongest enrichment between USF1 and ELF1 motifs. **(C)** spatzie cooperativity estimation computed using motif instance co-occurrence shows significant co-enrichment of USF2 and ELF1 motifs. Adjusted p-values were corrected with the Benjamini-Hochberg procedure. Randomization experiments where top $N\%$ of enhancer events are randomly permuted shows shrinking significance of co-enrichment under noisy data for **(D)** score correlation, **(E)** count correlation, and **(F)** hypergeometric co-enrichment. Dashed line represents the significance threshold at $p < 0.05$ under Bonferroni correction.

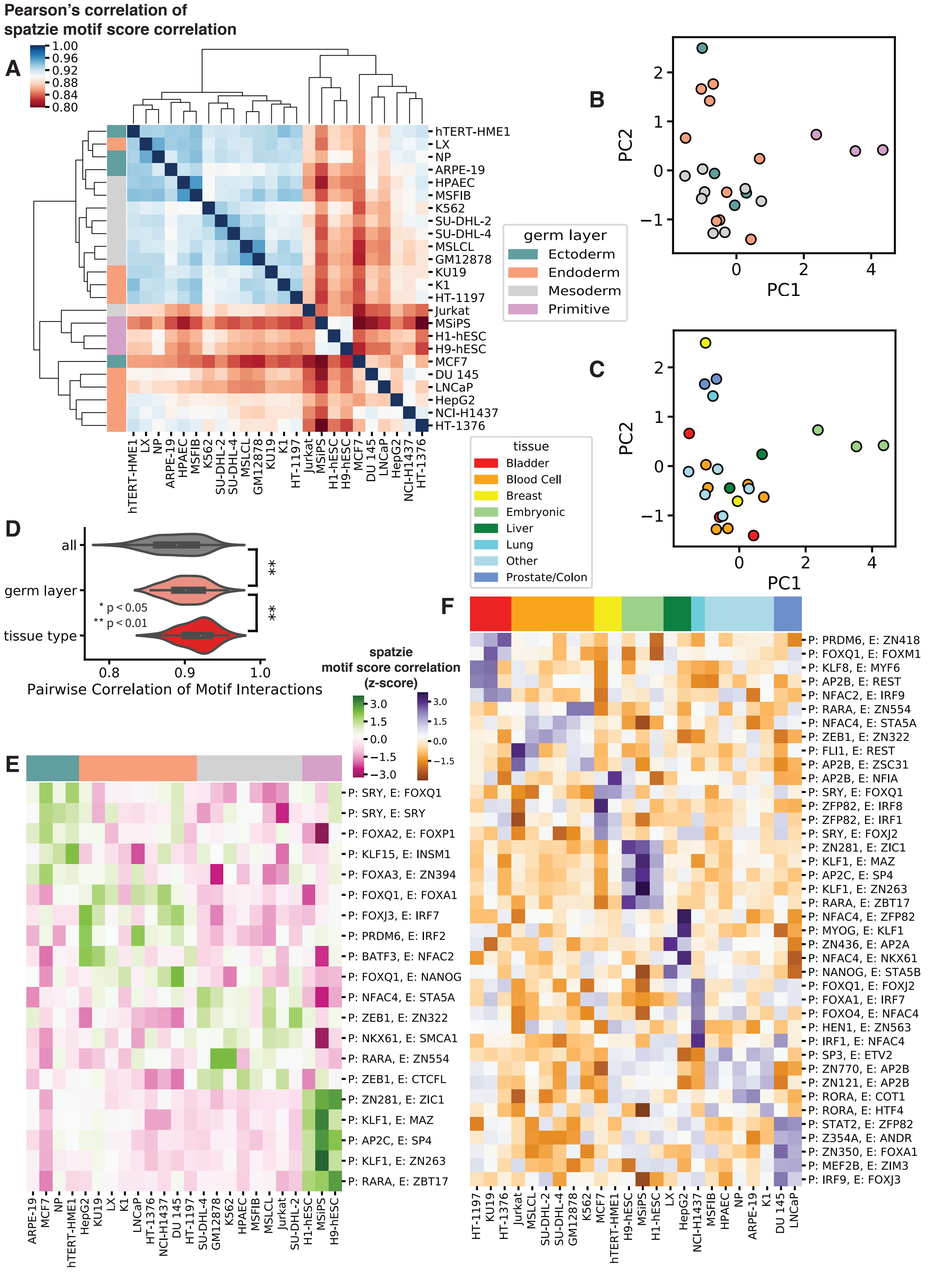


Figure 6-3: spatzie identifies transcription factor cooperation underlying interactions that are germ line and tissue-specific. (figure caption continued on next page)

Figure 6-3: spatzie identifies transcription factor cooperation underlying interactions that are germ line and tissue-specific. **(A)** Pearson correlation of pairwise interactions shows similarity between related tissues. PCA on spatzie discovered transcription factor interactions shows **(B)** germ layer **(C)** and tissue type clustering. **(D)** Pairwise correlation of spatzie transcription factor motif interactions shows increasing relatedness of germ layer and tissue type. Significance computed by Wilcoxon rank-sum test. **(E)** Extraction of germ layer-specific transcription factor motif interactions include relevant lineage-determining transcription factor families, such as Fox in endoderm and ectoderm. Spatzie correlation scores are z-scores normalized by row. **(F)** Extraction of tissue-specific transcription factor motif interactions include potential transcription factor trade-offs at the promoter and enhancer that may mediate tissue-specific enhancer-promoter interactions. Spatzie correlation scores are z-scores normalized by row.

promoter interactions. After evaluating with spatzie the score correlation of 50,286 enhancer-promoter interactions that were present in at least one cell type, we find interactions cluster by tissue and germ line (Figure 6-3A-C), and that correlation of discovered motif pairs increases among cell types from the same germ layer and tissue (Figure 6-3D). While previous work has shown that cohesin-mediated genomic interactions are similarly stratified by germ layer and tissue [11], our analysis with spatzie shows that there is sufficient information within the co-enrichment of motifs underlying enhancer and promoter interactions to reproduce biologically meaningful germ layer and tissue layer organization. We found the tissue-level correlation between spatzie-discovered co-enriched motifs was reproducible with interaction calls using CID (Supplementary Figure E-2), which was previously shown to recover more reproducible interactions from ChIA-PET data [13]. We then extracted enhancer-promoter interactions that had the highest germ layer-specific expression and found that these include transcription factors such as Fox family members, which have a known role in endoderm development [8, 4] and ectoderm development [27, 29], and Nfatc4 in mesoderm interactions, which is a known T cell [25] and myogenic [17, 3] differentiation factor. (Figure 6-3E). Similarly, we examined tissue-level specific interactions (Figure 6-3F) and found instances that were cell type specific, which is indicative of a potential trade-off in partnering of Sta5 at the enhancer with either Nfatc4 or Nanog at the promoter in blood or in liver tissues, respectively, or of AP2B at the promoter with Nfia or Zsc31 at the enhancers in breast or blood tissues, respectively (Figure 6-3F).

6.5 Discussion

Overall, spatzie contributes to a growing field of tools for the analysis of enhancer-promoter interaction data by providing a collection of statistical tests to identify transcription factor motif co-enrichment. While other methods such as PEP-Motif [33] and the graphical lasso approach taken in Pliner et al. [30] may identify such co-enrichment relationships, they spend computational power to identify motifs that predict the activity of enhancers and promoters independent of their interactions, whereas spatzie focuses exclusively on identifying motifs which share a co-enrichment relationship between enhancer-promoter interactions. We validate spatzie on experimental data where we use real ChIP-seq data to simulate enhancer-promoter interactions between USF1 binding at promoters and ELF1 binding at enhancers. We show that spatzie's three modes of motif pair relationship (motif score correlation, motif count correlation, and instance co-occurrence) all successfully identify USF1:ELF1 motif co-occurrence relationships even under noisy conditions, with motif score correlation achieving the most robust results. We also apply spatzie to data from 24 human cell lines and are able to show transcription factor co-enrichments that are discovered by spatzie cluster at germ-layer and tissue level, indicating these co-enrichment relationships are related to the organization of these cell types by germ layer and tissue. Furthermore, the identified germ layer and tissue-specific transcription factor interactions contain lineage-determining transcription factors, indicating that transcription factor co-enrichment between enhancers and promoters contains transcription factors that are known to play a role in differentiation and may point to their function as players in the structural organization of the genome. One concern with motif co-enrichment approaches is that the primary effects discovered can be attributed to the activity of cell type-specific transcription factors without a cooperative relationship. However, as evidence to the contrary we find examples of tissues that share enhancer or promoter motifs with different partners, indicating spatzie is identifying co-enrichment beyond general transcription factor activity. This combined with evidence that spatzie does not discover co-enrichment when we entirely randomize the relationship between binding for simulated interactions from USF1 and ELF1 ChIP-seq suggests that spatzie effects are not dominated by the general over-enrichment of motifs, but instead are based on a dependent relationship between a pair of transcription factors underlying enhancer-promoter interactions. In sum, we hope spatzie provides biological insight into the cell type-specific rules of transcription factor co-operativity underlying enhancer-promoter interactions.

Availability

The functionality of spatzie is bundled as an R package with the same name. The source code of the R package is hosted on GitHub (`https://github.com/gifford-lab/spatzie`). The core functionality is also available online at `https://spatzie.mit.edu`, which includes enhancer-promoter motif co-enrichment analysis with HOCO-MOCO or user-defined motifs on interactions data mapped to either hg38, hg19, mm9, or mm10.

Bibliography

[1] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):289–300, 1995.

[2] Yoav Benjamini and Daniel Yekutieli. The control of the false discovery rate in multiple testing under dependency. *Ann. Statist.*, 29(4):1165–1188, 08 2001.

[3] Paul B Bushdid, Hanna Osinska, Ronald R Waclaw, Jeffery D Molkentin, and Katherine E Yutzey. NFATc3 and NFATc4 are required for cardiac development and mitochondrial function. *Circulation research*, 92(12):1305–1313, 2003.

[4] Filippo M Cernilogar, Stefan Hasenoder, Zeyang Wang, Katharina Scheibner, Ingo Burtscher, Michael Sterr, Pawel Smialowski, Sophia Groh, Ida M Evenroed, Gregor D Gilfillan, Heiko Lickert, and Gunnar Schotta. Pre-marked chromatin and transcription factor co-binding shape the pioneering activity of Foxa2. *Nucleic Acids Research*, 47(17):9069–9086, sep 2019.

[5] A. Chakraborty and F. Ay. The role of 3D genome organization in disease: From compartments to single nucleotides. *Semin Cell Dev Biol*, 90:104–113, 06 2019.

[6] J. Dekker. Gene regulation in the third dimension. *Science*, 319(5871):1793–1794, Mar 2008.

[7] Jean Dunn and Olive Jean Dunn. Multiple comparisons among means. *American Statistical Association*, pages 52–64, 1961.

[8] Tiago Faial, Andreia S Bernardo, Sasha Mendjan, Evangelia Diamanti, Daniel Ortmann, George E Gentsch, Victoria L Mascetti, Matthew W B Trotter, James C Smith, and Roger A Pedersen. Brachyury and SMAD signalling collaboratively orchestrate distinct mesoderm and endoderm gene regulatory networks in differentiating human embryonic stem cells. *Development*, 142(12):2121–2135, 2015.

[9] O. Fornes, J. A. Castro-Mondragon, A. Khan, R. van der Lee, X. Zhang, P. A. Richmond, B. P. Modi, S. Correard, M. Gheorghe, D. Baranašić, W. Santana-Garcia, G. Tan, J. Chèneby, B. Ballester, F. Parcy, A. Sandelin, B. Lenhard, W. W. Wasserman, and A. Mathelier. JASPAR 2020: update of the open-access database of transcription factor binding profiles. *Nucleic Acids Res*, 48(D1):D87–D92, 01 2020.

[10] M. J. Fullwood and Y. Ruan. ChIP-based methods for the identification of long-range chromatin interactions. *J. Cell. Biochem.*, 107(1):30–39, May 2009.

[11] Fabian Grubert, Rohith Srivas, Damek V Spacek, Maya Kasowski, Mariana Ruiz-Velasco, Nasa Sinnott-Armstrong, Peyton Greenside, Anil Narasimha, Qing Liu, Benjamin Geller, Akshay Sanghi, Michael Kulik, Silin Sa, Marlene Rabinovitch, Anshul Kundaje, Stephen Dalton, Judith B Zaugg, and Michael Snyder. Landscape of cohesin-mediated chromatin loops in the human genome. *Nature*, 583(7818):737–743, 2020.

[12] Y. Guo, G. Papachristoudis, R. C. Altshuler, G. K. Gerber, T. S. Jaakkola, D. K. Gifford, and S. Mahony. Discovering homotypic binding events at high spatial resolution. *Bioinformatics*, 26(24):3028–3034, Dec 2010.

[13] Yuchun Guo, Konstantin Krismer, Michael Closser, Hynek Wichterle, and David K Gifford. High resolution discovery of chromatin interactions. *Nucleic acids research*, 47(6):e35–e35, 2019.

[14] S. Heinz, C. Benner, N. Spann, E. Bertolino, Y. C. Lin, P. Laslo, J. X. Cheng, C. Murre, H. Singh, and C. K. Glass. Simple combinations of lineage-determining

transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol Cell*, 38(4):576–589, May 2010.

[15] Yosef Hochberg. A sharper bonferroni procedure for multiple tests of significance. *Biometrika*, 75(4):800–802, 1988.

[16] S. Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6:65–70, 1979.

[17] V. Horsley and G. K. Pavlath. NFAT: ubiquitous regulator of cell differentiation and adaptation. *J Cell Biol*, 156(5):771–774, Mar 2002.

[18] J. Korhonen, P. Martinmäki, C. Pizzi, P. Rastas, and E. Ukkonen. MOODS: fast search for position weight matrix matches in DNA sequences. *Bioinformatics*, 25(23):3181–3182, Dec 2009.

[19] S. T. Kosak and M. Groudine. Form follows function: The genomic organization of cellular differentiation. *Genes Dev*, 18(12):1371–1384, Jun 2004.

[20] Ivan V Kulakovskiy, Ilya E Vorontsov, Ivan S Yevshin, Ruslan N Sharipov, Alla D Fedorova, Eugene I Rumynskiy, Yulia A Medvedeva, Arturo Magana-Mora, Vladimir B Bajic, Dmitry A Papatsenko, Fedor A Kolpakov, and Vsevolod J Makeev. HOCOMOCO: towards a complete collection of transcription factor binding models for human and mouse via large-scale ChIP-Seq analysis. *Nucleic Acids Research*, 46(D1):D252–D259, 2018.

[21] C. Lanctôt, T. Cheutin, M. Cremer, G. Cavalli, and T. Cremer. Dynamic genome architecture in the nuclear space: regulation of gene expression in three dimensions. *Nat Rev Genet*, 8(2):104–115, Feb 2007.

[22] M. Lawrence, W. Huber, H. Pagès, P. Aboyoun, M. Carlson, R. Gentleman, M. T. Morgan, and V. J. Carey. Software for computing and annotating genomic ranges. *PLoS Comput Biol*, 9(8):e1003118, 2013.

[23] Heng Li. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv preprint arXiv:1303.3997*, 2013.

[24] Heng Li, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, and Richard Durbin. The sequence alignment/map format and SAMtools. *Bioinformatics*, 25(16):2078–2079, 2009.

[25] Fernando Macian. NFAT proteins: key regulators of T-cell development and function. *Nature Reviews Immunology*, 5(6):472–484, 2005.

[26] Marcel Martin. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet. journal*, 17(1):10–12, 2011.

[27] A Paula Monaghan, Klaus H Kaestner, Evelyn Grau, and G Schutz. Postimplantation expression patterns indicate a role for the mouse forkhead/HNF-3 alpha, beta and gamma genes in determination of the definitive endoderm, chordamesoderm and neuroectoderm. *Development*, 119(3):567–578, 1993.

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

[29] Karen M Neilson, Steven L Klein, Pallavi Mhaske, Kathy Mood, Ira O Daar, and Sally A Moody. Specific domains of FoxD4/5 activate and repress neural transcription factor genes to control the progression of immature neural ectoderm to differentiating neural plate. *Developmental Biology*, 365(2):363–375, 2012.

[30] Hannah A Pliner, Jonathan S Packer, José L McFaline-Figueroa, Darren A Cusanovich, Riza M Daza, Delasa Aghamirzaie, Sanjay Srivatsan, Xiaojie Qiu, Dana Jackson, and Anna Minkina. Cicero predicts cis-regulatory DNA interactions from single-cell chromatin accessibility data. *Molecular cell*, 71(5):858–871, 2018.

[31] A. N. Schep, B. Wu, J. D. Buenrostro, and W. J. Greenleaf. chromVAR: inferring transcription-factor-associated accessibility from single-cell epigenomic data. *Nat Methods*, 14(10):975–978, Oct 2017.

[32] Abigail M. Searfoss, Wan chun Liu, and Nicole Creanza. Geographically well-distributed citizen science data reveals range-wide variation in the chipping sparrow's simple song. *Animal Behaviour*, 161:63 – 76, 2020.

[33] Yang Yang, Ruochi Zhang, Shashank Singh, and Jian Ma. Exploiting sequence-based features for predicting enhancer–promoter interactions. *Bioinformatics*, 33(14):i252–i260, 2017.

Chapter 7

Conclusions

This thesis has introduced several machine learning and bioinformatics methods for the analysis of high-throughput experimental data in functional genomics.

Chapter 2 described a framework for evaluating neural network architectures by combining rule-based simulation of biological sequence data and feature attribution methods, and thereby providing an avenue for investigating the model's ability to recover the underlying rules. In chapter 3 we used an ensemble of deep neural networks to predict cell type-specific chromatin accessibility. These *in silico* predictions were subsequently validated experimentally. The remaining three chapters presented novel computational methods for the analysis of long-range genomic interaction data from assays such as ChIA-PET [4], HiChIP [9], and Hi-C [8]. While chapter 4 focused on the identification of such genomic interactions, chapter 5 presented a method to assess their reproducibility. Chapter 6 concluded the analysis of long-range genomic interactions by uncovering co-enriched pairs of transcription factor sequence motifs in enhancer-promoter interactions.

7.1 Deep learning in biology

In recent years the field of computational biology has seen a proliferation of deep learning models, which were trained on a variety of biological sequence-based data types, such as data sets derived from ChIP-seq, DNase-seq, ATAC-seq [2], and RNA-seq experiments. The architectural building blocks of the neural networks that were trained on biological sequence data were in large part adapted from fields pioneering the application of neural networks, most noteably computer vision and natural language processing. Such building blocks include convolutional layers [1, 15, 6], recurrent layers [11, 12], dilated convolutional layers [5], skip connections [10], and, more recently, graph-convolutional layers [13] and transformer-based architectures [3, 7]. Despite the apparent diversity of neural network architectures in genomics and the various quantities and annotations they are trained to predict, they are similar in the sense that they are high-capacity models with millions of trainable parameters. This overparameterization oftentimes helps predictive performance, even on unseen data (i.e., low generalization error), for reasons that are not entirely understood [14]. However, large overparameterized models are difficult to interpret, with complicated decision boundaries and built-in redundancy. While trading interpretability for predictive performance can be beneficial in situations where prediction accuracy is the primary objective, with models trained on biological sequences the prediction is often secondary to a succinct explanation of the rules behind the predictions. Ultimately, the goal is to arrive at a mechanistic understanding of the underlying biological phenomena that are modeled by neural networks. As described in chapter 2, we hope a simulation-based approach, when paired with neural network interpretability methods, will be a valuable tool on the way to more informative models.

Bibliography

[1] B. Alipanahi, A. Delong, M. T. Weirauch, and B. J. Frey. Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nat Biotechnol*, 33(8):831–838, Aug 2015.

[2] J. D. Buenrostro, P. G. Giresi, L. C. Zaba, H. Y. Chang, and W. J. Greenleaf. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat Methods*, 10(12):1213–1218, Dec 2013.

[3] Jim Clauwaert and Willem Waegeman. Novel transformer networks for improved sequence labeling in genomics. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, pages 1–1, 2020.

[4] M. J. Fullwood and Y. Ruan. ChIP-based methods for the identification of long-range chromatin interactions. *J. Cell. Biochem.*, 107(1):30–39, May 2009.

[5] D. R. Kelley, Y. A. Reshef, M. Bileschi, D. Belanger, C. Y. McLean, and J. Snoek. Sequential regulatory activity prediction across chromosomes with convolutional neural networks. *Genome Res*, 28(5):739–750, 05 2018.

[6] D. R. Kelley, J. Snoek, and J. L. Rinn. Basset: learning the regulatory code of the accessible genome with deep convolutional neural networks. *Genome Res.*, 26(7):990–999, 07 2016.

[7] N. Q. K. Le, Q. T. Ho, T. T. Nguyen, and Y. Y. Ou. A transformer architecture based on BERT and 2D convolutional neural network to identify DNA enhancers from sequence information. *Brief Bioinform*, Feb 2021.

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

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

[10] S. Nair, D. S. Kim, J. Perricone, and A. Kundaje. Integrating regulatory DNA sequence and gene expression to predict genome-wide chromatin accessibility across cellular contexts. *Bioinformatics*, 35(14):i108–i116, 07 2019.

[11] D. Quang and X. Xie. DanQ: a hybrid convolutional and recurrent deep neural network for quantifying the function of DNA sequences. *Nucleic Acids Res.*, 44(11):e107, 06 2016.

[12] D. Quang and X. Xie. FactorNet: A deep learning framework for predicting cell type specific transcription factor binding from nucleotide-resolution sequential data. *Methods*, 166:40–47, 08 2019.

[13] Sungmin Rhee, Seokjun Seo, and Sun Kim. Hybrid approach of relation network and localized graph convolutional filtering for breast cancer subtype classification. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, pages 3527–3534. International Joint Conferences on Artificial Intelligence Organization, 7 2018.

[14] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning (still) requires rethinking generalization. *Commun. ACM*, 64(3):107–115, February 2021.

[15] J. Zhou and O. G. Troyanskaya. Predicting effects of noncoding variants with deep learning-based sequence model. *Nat. Methods*, 12(10):931–934, Oct 2015.

Appendix A

Supplementary information for *seqgra: Principled Selection of Neural Network Architectures for Genomics Prediction Tasks*

Supplementary information can also be found on the bioRxiv website: `https://doi.org/10.1101/2021.06.14.448415`.

A.1 Supplementary Figures

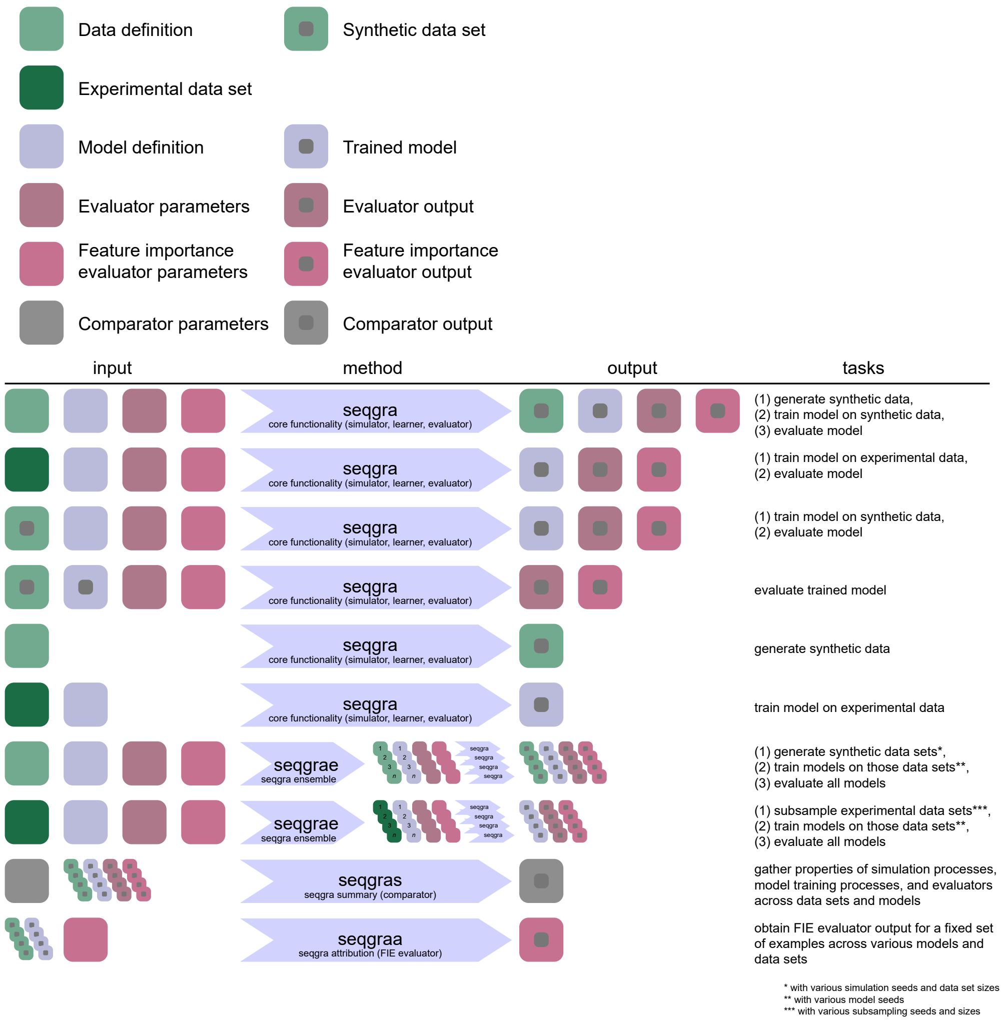

Data definition
Synthetic data set
Experimental data set
Model definition
Trained model
Evaluator parameters
Evaluator output
Feature importance evaluator parameters
Feature importance evaluator output
Comparator parameters
Comparator output
input
method
output
tasks
seqgra
core functionality (simulator, learner, evaluator)
(1) generate synthetic data,
(2) train model on synthetic data,
(3) evaluate model
seqgra
core functionality (simulator, learner, evaluator)
(1) train model on experimental data,
(2) evaluate model
seqgra
core functionality (simulator, learner, evaluator)
(1) train model on synthetic data,
(2) evaluate model
seqgra
core functionality (simulator, learner, evaluator)
evaluate trained model
seqgra
core functionality (simulator, learner, evaluator)
generate synthetic data
seqgra
core functionality (simulator, learner, evaluator)
train model on experimental data
seqgrae
seqgra ensemble
1
2
3
n
seqgra
seqgra
seqgra
seqgra
(1) generate synthetic data sets*,
(2) train models on those data sets**,
(3) evaluate all models
seqgrae
seqgra ensemble
1
2
3
n
seqgra
seqgra
seqgra
seqgra
(1) subsample experimental data sets***,
(2) train models on those data sets**,
(3) evaluate all models
seqgras
seqgra summary (comparator)
gather properties of simulation processes, model training processes, and evaluators across data sets and models
seqgraa
seqgra attribution (FIE evaluator)
obtain FIE evaluator output for a fixed set of examples across various models and data sets
* with various simulation seeds and data set sizes
** with various model seeds
*** with various subsampling seeds and sizes

Figure A-1: Schematic of common use cases for seqgra command line interface.

Figure A-1: Schematic of common use cases for seqgra command line interface. The seqgra package contains four commands, `seqgra`, `seqgrae`, `seqgras`, and `seqgraa`. `seqgra` contains the core functionality of (1) generating synthetic data using the *Simulator* component, (2) training models on either synthetic or experimental data using various *Learner* components, and (3) evaluating the model using various *Evaluator* components. `seqgrae`, short for seqgra ensemble, is a convenient way to generate multiple synthetic data sets with various data set sizes and simulation seeds and train models on them using a range of model seeds. `seqgras`, short for seqgra summary, is a tool to gather properties and metrics across a number of data sets and models and compare them using *Comparator* components. `seqgraa`, short for seqgra attribution, runs feature importance evaluators on a number of trained models, using the same set of examples each time.

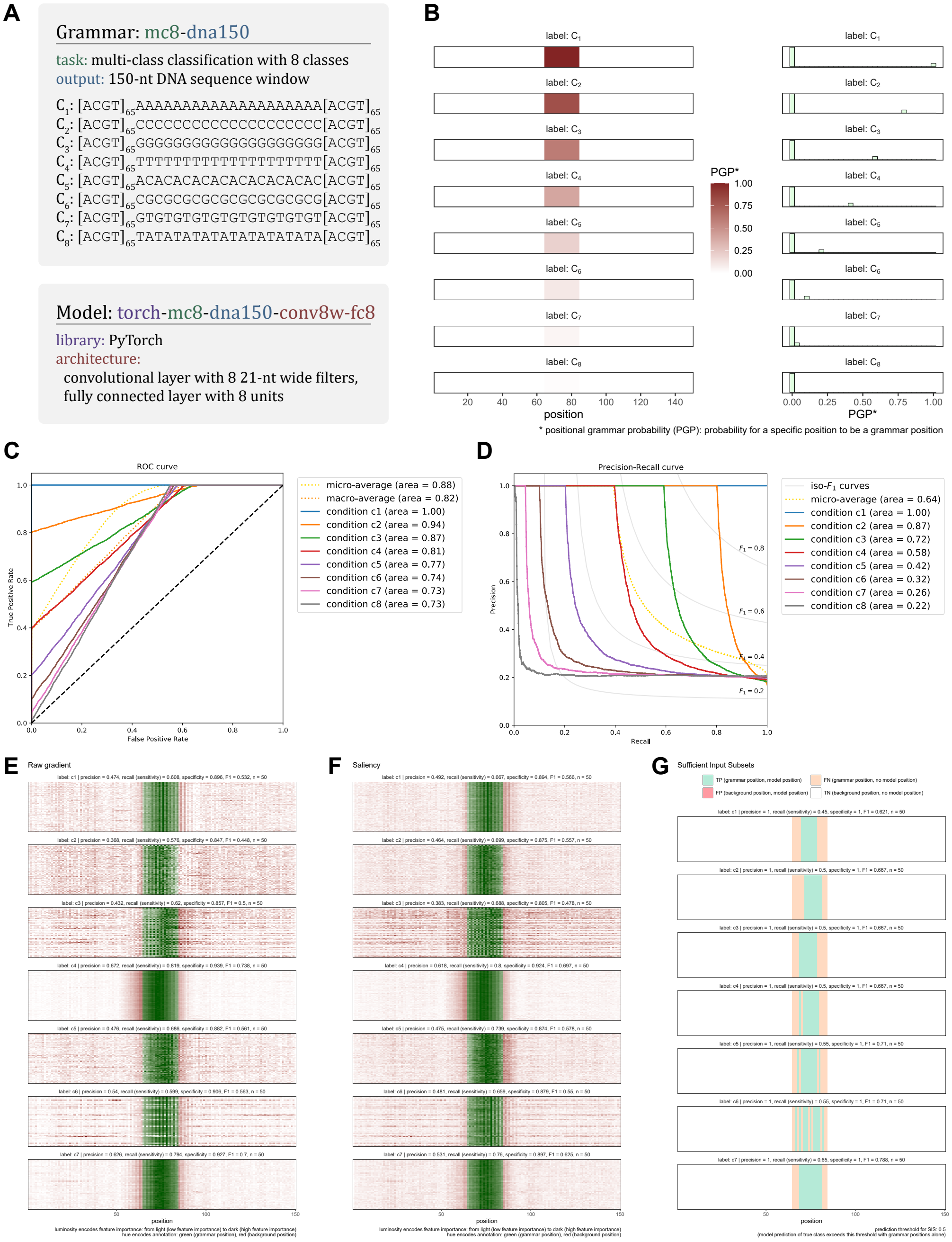


Figure A-2: Insertion probability test. (**A**) Grammar and model description. (**B**) Grammar position heatmap (on the left) depicting the probability of grammar annotation for all positions (1 - 150) and all classes (C_1 to C_8). (**C**) Test set ROC curve of classifier trained on synthetic data depicts class-specific true positive rates that mirror insertion probabilities, as expected.

Figure A-2: Insertion probability test. (**D**) Test set PR curve of same classifier, class-specific curves mirror insertion probabilities. (**E**) Raw gradient feature importance for classes C_1 to C_7 (classifier did not correctly predict class C_8). The x-axis is the position in the sequence window, the y-axis are randomly drawn examples with that class label. Dark green areas are grammar positions with high feature importance (desired), dark red areas are background positions with high feature importance (undesired). (**F**) Same as panel E, for absolute gradient (saliency) feature importance. (**G**) Same as panel E, for Sufficient Input Subsets (SIS) feature importance. The only difference is that SIS is an inherently discrete measure of feature importance, either positions are part of a sufficient input subset or not.

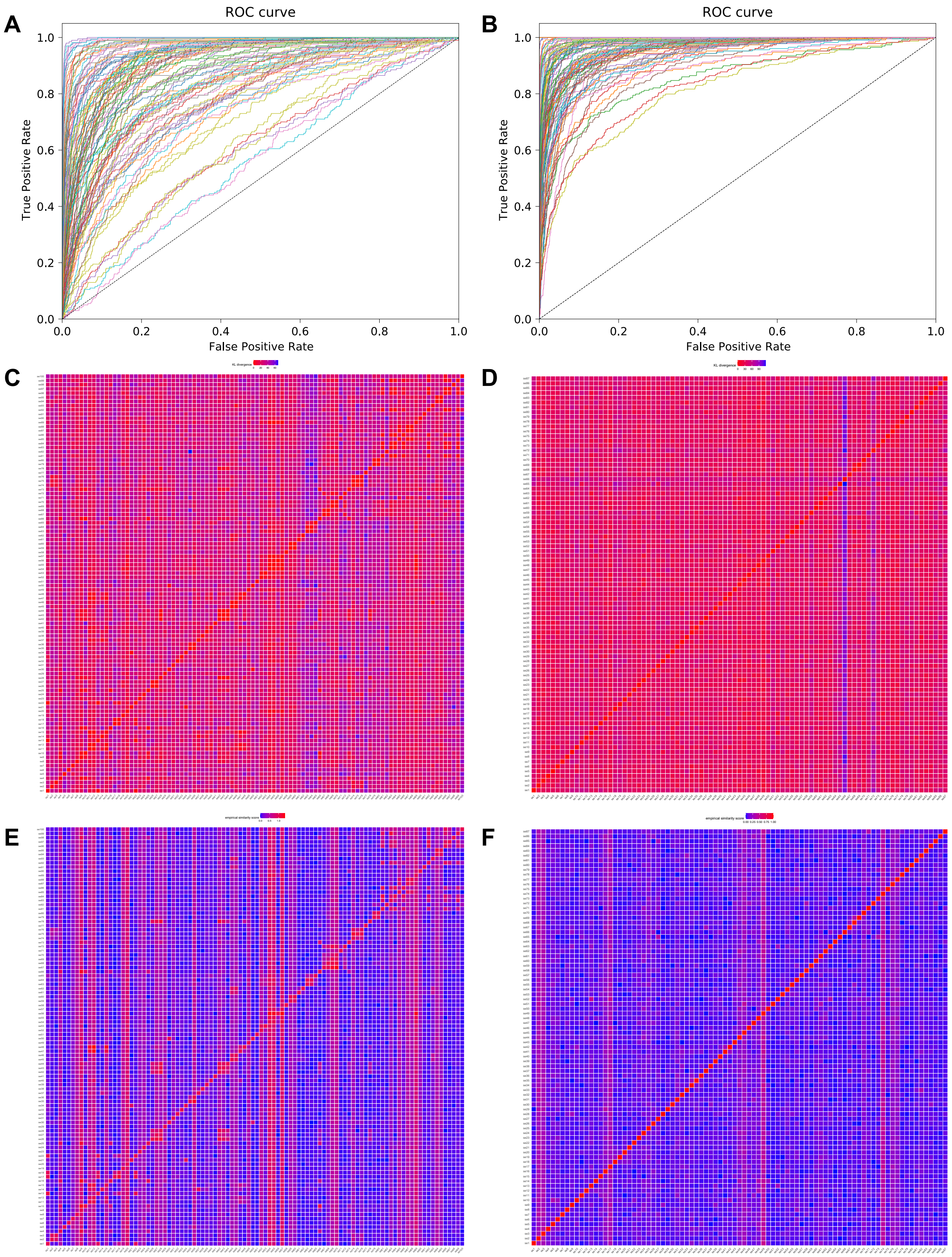


Figure A-3: Selection of sequence motifs for MC100 simulation grammars. (**A**) ROC curve of Bayes Optimal Classifier on multi-class classification task with 100 classes, prior to filtering out ambiguous sequence motifs.

Figure A-3: Selection of sequence motifs for MC100 simulation grammars. (**B**) Same as panel A, after ambiguous sequence motifs were removed. (**C**) KL divergence matrix of 100 sequence motifs, prior to filtering. (**D**) Same as panel C, after removing ambiguous motifs. (**E**) Empirical similarity score matrix of 100 sequence motifs, prior to filtering. (**F**) Same as panel D, after removing ambiguous motifs.

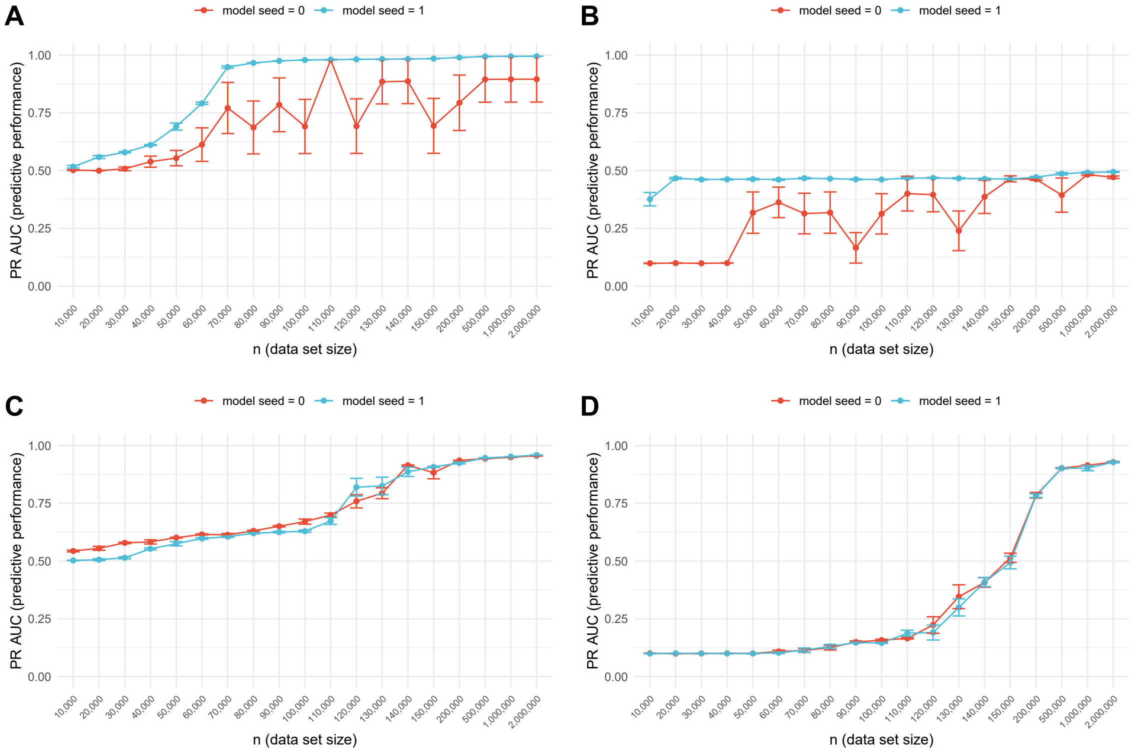


Figure A-4: PyTorch and TensorFlow affected by random seed induced instability. (**A**) Shown are test set PR AUCs of a PyTorch neural network architecture with two hidden layers, a convolutional layer with 10 21-nt wide filters followed by a dense layer with 5 units, trained on binary classification data sets using HOMER motifs without interactions. This PyTorch neural network architecture exhibits an unusual variability in PR AUC when trained with a random seed of zero. (**B**) Shown are test set PR AUCs of a TensorFlow neural network architecture with a convolutional layer with 10 21-nt wide filters, a global max pooling operation, and a dense layer with 10 units, trained on multi-class classification data sets with 10 classes using HOMER motifs with spacing-sensitive interactions. This TensorFlow neural network architecture exhibits an unusual variability in PR AUC when trained with a random seed of zero. (**C**) Unlike panel A, this PyTorch neural network architecture (convolutional layer with 10 11-nt wide filters followed by a dense layer with 5 units), which was trained on the same data sets as the one in panel A, does not exhibit unusually high variability of PR AUCs when trained with a random seed of zero.

Figure A-4: PyTorch and TensorFlow affected by random seed induced instability. (**D**) Similarly, this TensorFlow neural network architecture (convolutional layer with 10 21-nt wide filters followed by a dense layer with 10 units), which was trained on the same data sets as the one in panel B, also does not exhibit unusually high variability of PR AUCs when trained with a random seed of zero.

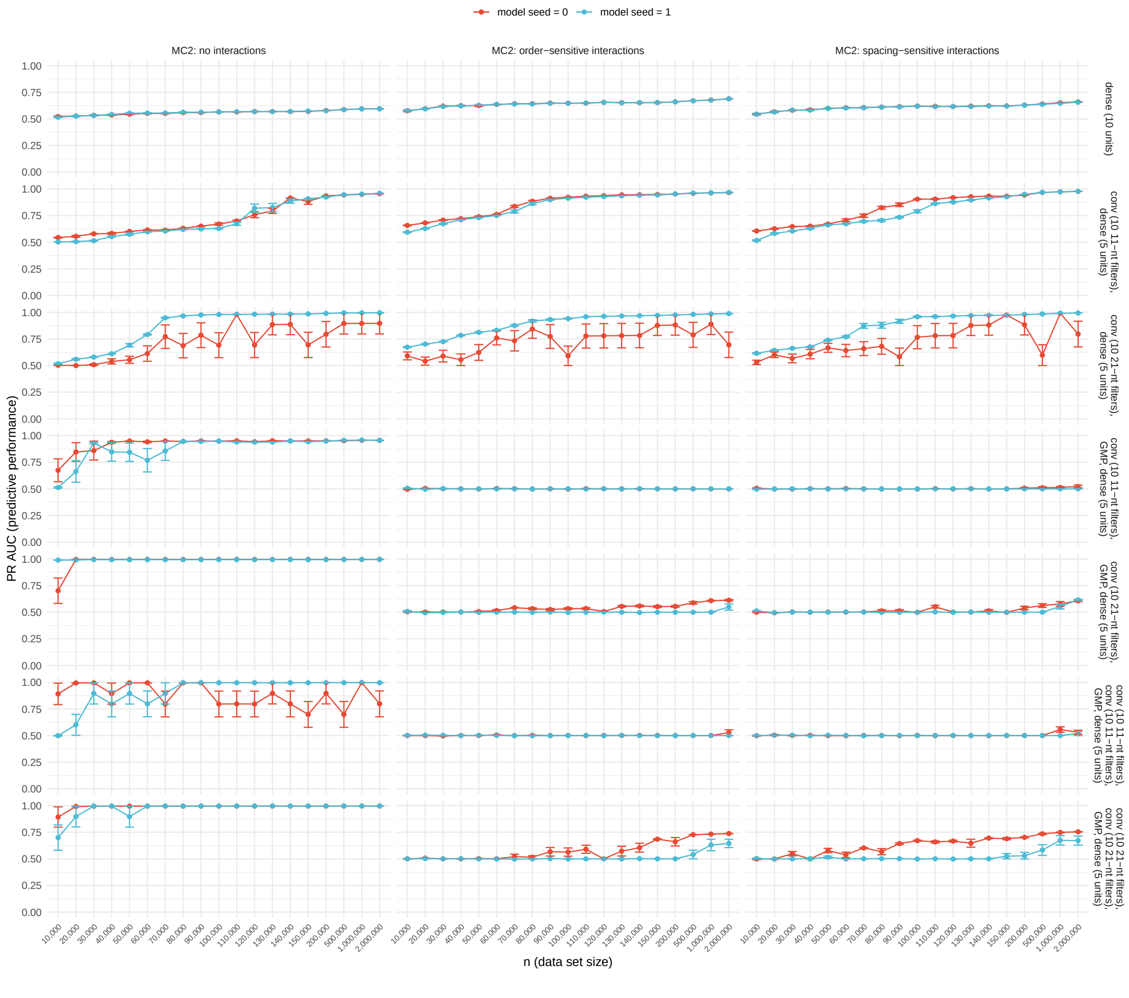


Figure A-5: PyTorch models trained with random seed 0 suffer from grammar-dependent and architecture-dependent instability. Not all grammar-architecture combinations are affected.

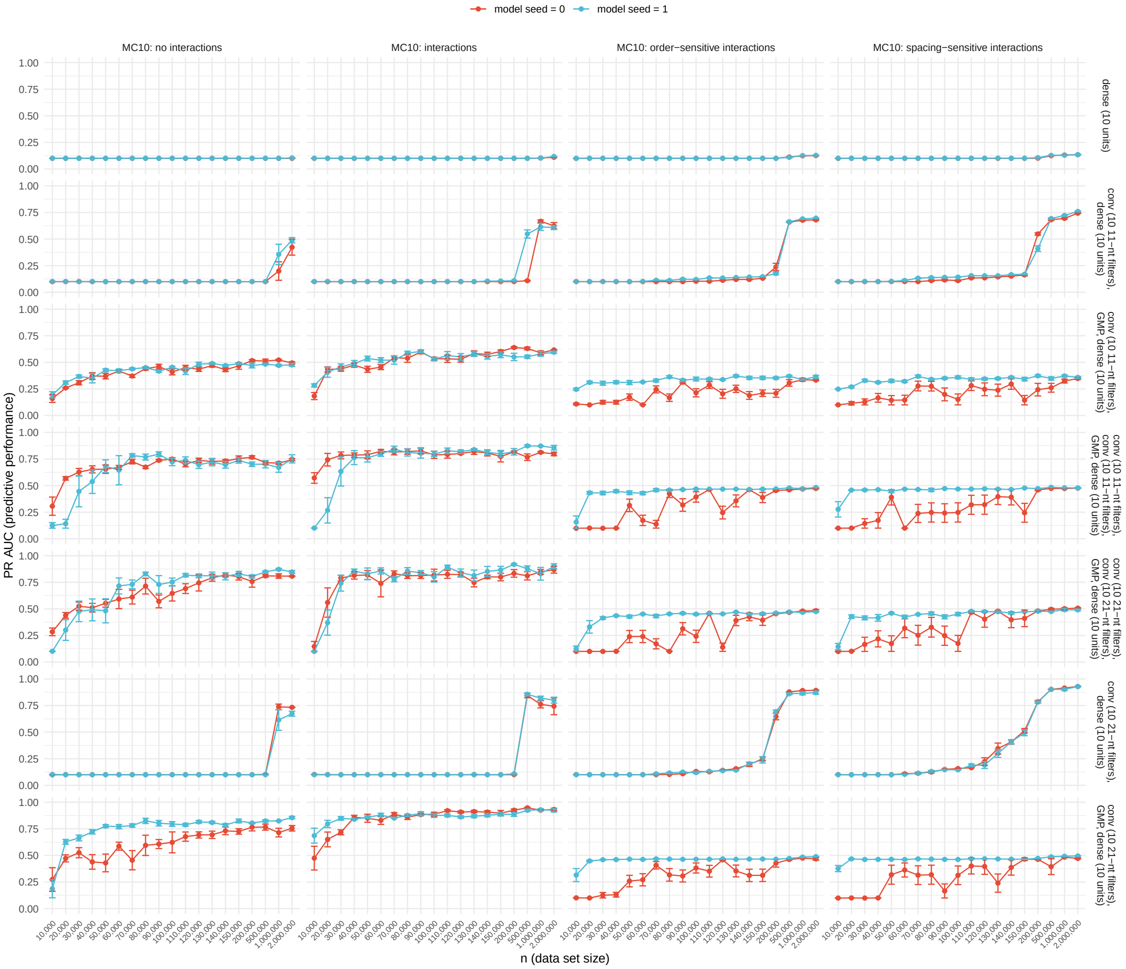


Figure A-6: TensorFlow models trained with random seed 0 suffer from grammar-dependent and architecture-dependent instability. Not all grammar-architecture combinations are affected.

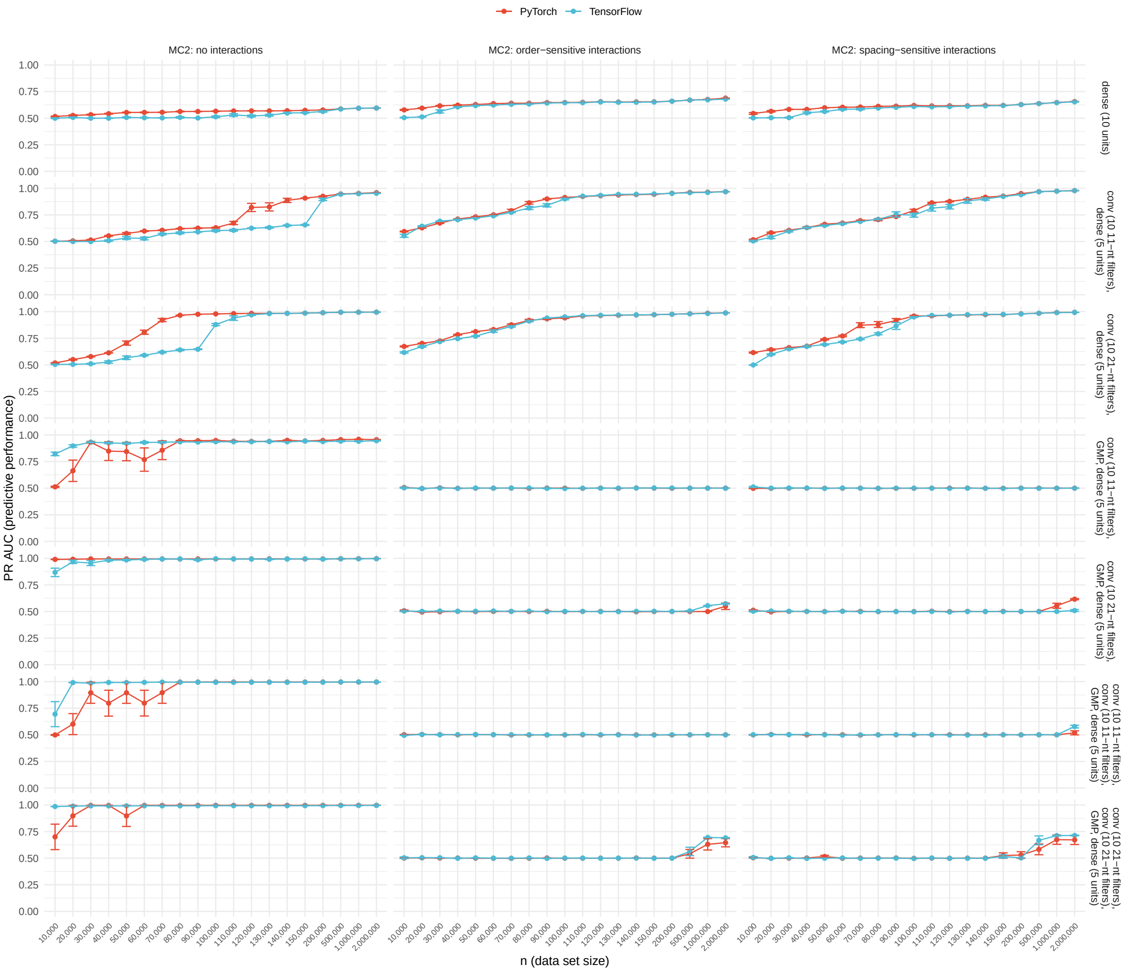


Figure A-7: Comparison of PyTorch and TensorFlow architectures trained on binary classification data sets. When comparing an equivalent neural network architecture, trained on the same data set, test set PR AUCs between models implemented and trained with deep learning libraries PyTorch and TensorFlow are similar. Shown here are comparisons across three grammars, 19 data set sizes, and seven neural network architectures.

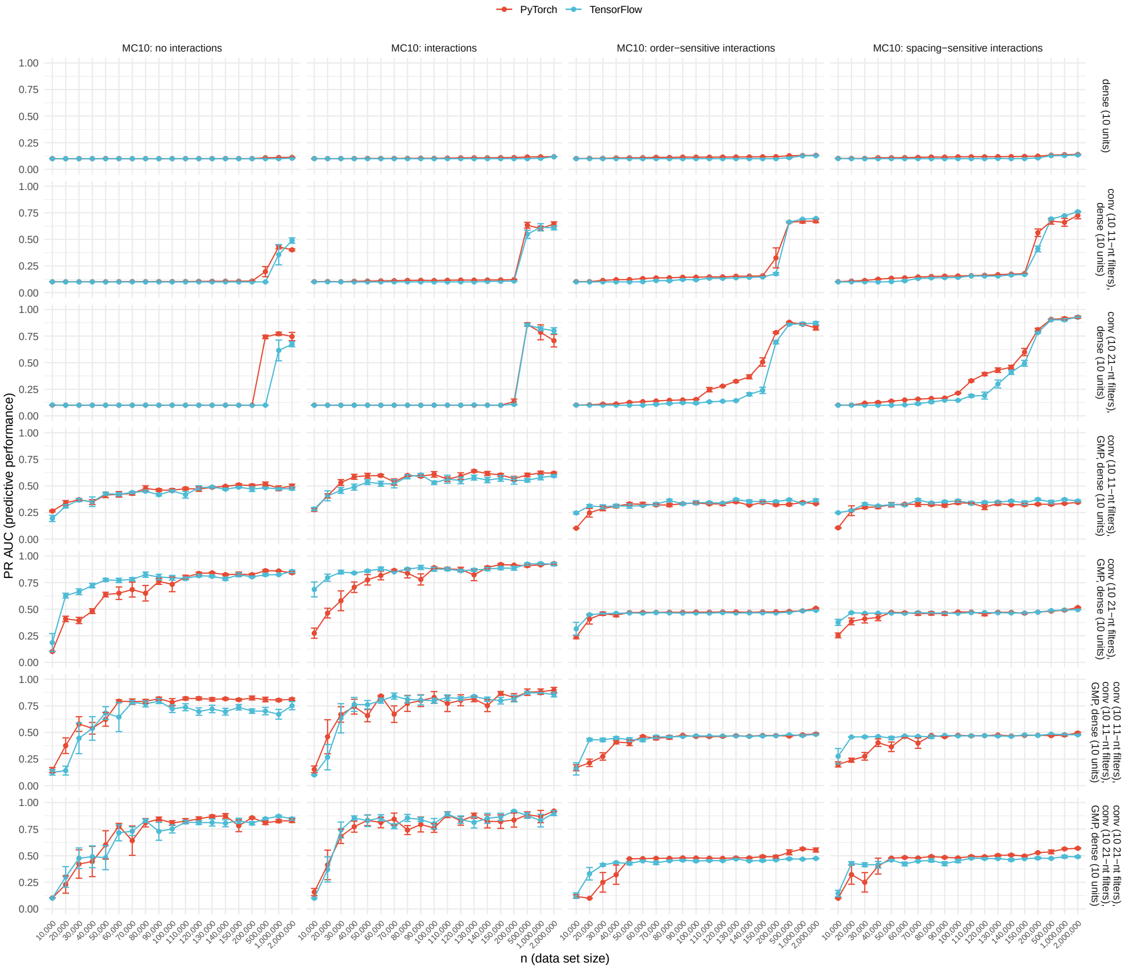


Figure A-8: Comparison of PyTorch and TensorFlow architectures trained on multi-class classification data sets with 10 classes. When comparing an equivalent neural network architecture, trained on the same data set, test set PR AUCs between models implemented and trained with deep learning libraries PyTorch and TensorFlow are similar. Shown here are comparisons across four grammars, 19 data set sizes, and seven neural network architectures.

A.2 Supplementary Tables

ID	Motif	IUPAC notation	Width (in nt)	MIC	$D_{\mathrm{KL}}(\cdot)$
se1	FOXA1:AR	AGTAAACAAAAAAGAACANA	20	18.9	17.4
se2	Bcl11a	TYTGACCASWRG	12	11.3	11.8
	MC2 grammar motifs above				
se3	Brachyury	ANTTMRCASBNNNGTGYKAAN	21	11.5	11.6
se4	CEBP:CEBP	NTNATGCAAYMNNHTGMAAY	20	14.8	14.3
se5	Chop	ATTGCATCAT	10	13.2	12.7
se6	CHR	CGGTTTCAAA	10	12.6	11.8
se7	CTCF-SatelliteElement	TGCAGTTCCAANAGTGGCCA	20	18.8	19.6
se8	Mouse Recombination Hotspot	ACTYKNATTCGTGNTACTTC	20	15.3	14.9
se9	RAR:RXR	RGGTCADNNAGAGGTCAV	18	16.3	17.3
se10	DUX	BCWGATTCAATCAAN	15	17.9	16.9
	MC10 grammar motifs above				
se11	E2F7	VDTTTCCCGCCA	12	13.4	14.6
se12	EBNA1	GGYAGCAYDTGCTDCCCNNN	20	18.1	19.2
se13	ERE	AAGGTCACNGTGACC	15	14.3	15.2
se14	ETS:E-box	AGGAAACAGCTG	12	17.3	17.6
se15	EWS:ERG-fusion	ATTTCCTGTN	10	13.7	13.5
se16	Foxh1	NNTGTGGATTSS	12	11.3	11.1
se17	FXR	AGGTCANTGACCTN	14	12.3	13.2
se18	GATA3	AGATGKDGAGATAAG	15	17.3	16.5
se19	GATA3	AGATSTNDNNDSAGATAASN	20	16.9	16.3
se20	GATA	NAGATWNBNATCTNN	15	14.0	13.3
	MC20 grammar motifs above				
se21	GATA:SCL	CGGCTGCNGNNNNCAGATAA	20	15.4	15.9
se22	Gfi1b	AAATCACTGC	10	13.9	13.8
se23	GRHL2	AAACYKGTTWDACMRGTTTB	20	13.5	13.4
se24	Hand2	TGACANARRCCAGRC	15	13.2	13.6
se25	HINFP	TWVGGTCCGC	10	11.7	13.2
se26	HOXB13	TTTTATKRGG	10	13.5	12.6
se27	LRF	AAGACCCYYN	10	11.2	12.5
se28	LXRE	GGGTTACTANAGGTCA	16	17.5	17.9
se29	NF1:FOXA1	NNTGTTTATTTTGGCA	16	17.3	16.7
se30	NFAT:AP1	GAATGGAAAAAATGAGTCAT	20	15.5	15.1
se31	NFAT	ATTTTCCATT	10	13.1	12.5
se32	NFY	AGCCAATCGG	10	13.3	13.8
se33	Nur77	TGACCTTTNCNT	12	15.1	14.8
se34	Oct2	ATATGCAAAT	10	15.3	14.1
se35	Oct4:Sox17	CCATTGTATGCAAAT	15	15.9	15.0
se36	OCT4-SOX2-TCF-NANOG	ATTTGCATAACAATG	15	16.4	14.9
se37	p53	ACATGCCCGGGCAT	14	16.7	18.2
se38	PAX3:FKHR-fusion	ACCGTGACTAATTNN	15	14.6	14.1

Continued on next page

Table A.1 – *Continued from previous page*

ID	Motif	IUPAC notation	Width (in nt)	MIC	$D_{\mathrm{KL}}(\cdot)$
se39	PAX5	GTCACGCTCNCTGA	14	15.1	16.3
se40	PAX6	NGTGTTCAVTSAAGCGKAAA	20	13.9	14.3
se41	Pax7	NTAATTDGCYAATTANNWWD	20	16.0	13.9
se42	Pax7	TAATCAATTA	10	16.3	14.6
se43	Pax8	GTCATGCHTGRCTGS	15	13.4	14.6
se44	Pitx1:Ebox	YTAATTRAWWCCAGATGT	18	12.7	11.8
se45	PRDM10	TGGTACATTCCA	12	11.9	12.3
se46	PRDM14	AGGTCTCTAACC	12	13.7	14.0
se47	PRDM15	YCCDNTCCAGGTTTT	15	13.2	13.7
se48	PRDM9	ADGGYAGYAGCATCT	15	12.8	13.1
se49	PSE	WAVTCACCMTAASYDAAAAG	20	10.6	10.3
se50	RBPJ:Ebox	GGGRAARRGRMCAGMTG	17	14.3	15.2

Table A.1: Homer transcription factor motifs for multi-class classification tasks with 2, 10, 20, and 50 classes (MC2-MC50): These motifs are used for grammars without interactions, with interactions, with interactions with order constraints, and with interactions with spacing constraints. The columns (from left to right) contain the seqgra-internal sequence element ID, the motif name (name of the transcription factor or complex), a summary of the motif in IUPAC notation, the width of the motif in nucleotides, the motif information content (MIC), and the KL divergence between the motif and the background sequence (using the human genomic nucleotide distribution).

ID	Motif	min $\boldsymbol{D}_{\mathrm{KL}}(\cdot,\cdot)$	max ESS$(\cdot,\cdot)$
se1	FOXA1:AR	46.1	0 %
se2	Bcl11a	30.4	33 %

Table A.2: Homer transcription factor motifs for binary classification tasks (MC2): These motifs are used for MC2 grammars without interactions, with interactions, with interactions with order constraints, and with interactions with spacing constraints. The columns (from left to right) contain the seqgra-internal sequence element ID, the motif name (name of the transcription factor or complex), the minimum KL divergence between the motif and the other motif in this grammar, and the maximum adjusted empirical similarity score (ESS) between the motif and the other motif in this grammar.

ID	Motif	min $\boldsymbol{D}_{\mathrm{KL}}(\cdot,\cdot)$	max ESS$(\cdot,\cdot)$
se1	FOXA1:AR	41.6	9 %
se2	Bcl11a	19.1	41 %
se3	Brachyury	18.3	38 %
se4	CEBP:CEBP	24.4	17 %
se5	Chop	15.6	33 %
se6	CHR	20.5	21 %
se7	CTCF-SatelliteElement	31.1	8 %
se8	Mouse Recombination Hotspot	33.3	6 %
se9	RAR:RXR	37.0	11 %
se10	DUX	24.1	12 %

Table A.3: Homer transcription factor motifs for multi-class classification tasks with 10 classes (MC10): These motifs are used for MC10 grammars without interactions, with interactions, with interactions with order constraints, and with interactions with spacing constraints. The columns (from left to right) contain the seqgra-internal sequence element ID, the motif name (name of the transcription factor or complex), the minimum KL divergence between the motif and the other 9 motifs in this grammar, and the maximum adjusted empirical similarity score (ESS) between the motif and the other 9 motif in this grammar.

ID	Motif	min $D_{\mathrm{KL}}(\cdot,\cdot)$	max ESS$(\cdot,\cdot)$
se1	FOXA1:AR	39.0	10 %
se2	Bcl11a	18.0	40 %
se3	Brachyury	18.3	35 %
se4	CEBP:CEBP	24.4	13 %
se5	Chop	15.6	32 %
se6	CHR	20.5	19 %
se7	CTCF-SatelliteElement	31.1	10 %
se8	Mouse Recombination Hotspot	24.3	9 %
se9	RAR:RXR	35.1	9 %
se10	DUX	24.1	14 %
se11	E2F7	20.1	15 %
se12	EBNA1	35.0	13 %
se13	ERE	16.8	18 %
se14	ETS:E-box	33.1	17 %
se15	EWS:ERG-fusion	20.5	20 %
se16	Foxh1	27.9	30 %
se17	FXR	16.8	35 %
se18	GATA3	30.7	21 %
se19	GATA3	31.2	20 %
se20	GATA	31.7	15 %

Table A.4: Homer transcription factor motifs for multi-class classification tasks with 20 classes (MC20): These motifs are used for MC20 grammars without interactions, with interactions, with interactions with order constraints, and with interactions with spacing constraints. The columns (from left to right) contain the seqgra-internal sequence element ID, the motif name (name of the transcription factor or complex), the minimum KL divergence between the motif and the other 19 motifs in this grammar, and the maximum adjusted empirical similarity score (ESS) between the motif and the other 19 motif in this grammar.

ID	Motif	min $D_{\mathrm{KL}}(\cdot,\cdot)$	max ESS$(\cdot,\cdot)$
se1	FOXA1:AR	32.7	15 %
se2	Bcl11a	12	45 %
se3	Brachyury	18.3	39 %
se4	CEBP:CEBP	23.0	18 %
se5	Chop	15.6	35 %
se6	CHR	17.6	26 %
se7	CTCF-SatelliteElement	31.1	9 %
se8	Mouse Recombination Hotspot	24.2	10 %
se9	RAR:RXR	22.3	28 %
se10	DUX	24.1	23 %
se11	E2F7	20.1	15 %
se12	EBNA1	35.0	12 %
se13	ERE	16.8	20 %
se14	ETS:E-box	26.9	16 %
se15	EWS:ERG-fusion	17.9	18 %
se16	Foxh1	20.7	36 %
se17	FXR	13.2	38 %
se18	GATA3	28.1	18 %
se19	GATA3	31.2	20 %
se20	GATA	27.4	20 %
se21	GATA:SCL	19.5	14 %
se22	Gfi1b	23.1	14 %
se23	GRHL2	21.2	19 %
se24	Hand2	15.2	30 %
se25	HINFP	25.6	23 %
se26	HOXB13	25.6	14 %
se27	LRF	23.2	20 %
se28	LXRE	20.0	28 %

Continued on next page

Table A.5 – *Continued from previous page*

ID	**Motif**	**min $D_{\mathrm{KL}}(\cdot,\cdot)$**	**max ESS$(\cdot,\cdot)$**
se29	NF1:FOXA1	29.1	13 %
se30	NFAT:AP1	19.1	17 %
se31	NFAT	17.2	20 %
se32	NFY	19.8	25 %
se33	Nur77	27.6	18 %
se34	Oct2	17.1	29 %
se35	Oct4:Sox17	22.6	15 %
se36	OCT4-SOX2-TCF-NANOG	17.9	14 %
se37	p53	27.3	10 %
se38	PAX3:FKHR-fusion	19.7	17 %
se39	PAX5	20.8	13 %
se40	PAX6	23.6	14 %
se41	Pax7	21.7	23 %
se42	Pax7	20.7	25 %
se43	Pax8	25.4	18 %
se44	Pitx1:Ebox	14.1	33 %
se45	PRDM10	14.5	33 %
se46	PRDM14	18.4	17 %
se47	PRDM15	23.0	18 %
se48	PRDM9	17.9	34 %
se49	PSE	15.0	44 %
se50	RBPJ:Ebox	17.8	24 %

Continued on next page

Table A.5 – *Continued from previous page*

ID	**Motif**	**min $\boldsymbol{D}_{\mathrm{KL}}(\cdot,\cdot)$**	**max ESS$(\cdot,\cdot)$**

Table A.5: Homer transcription factor motifs for multi-class classification tasks with 50 classes (MC50): These motifs are used for MC50 grammars without interactions, with interactions, with interactions with order constraints, and with interactions with spacing constraints. The columns (from left to right) contain the seqgra-internal sequence element ID, the motif name (name of the transcription factor or complex), the minimum KL divergence between the motif and the other 49 motifs in this grammar, and the maximum adjusted empirical similarity score (ESS) between the motif and the other 49 motif in this grammar.

Appendix B

Supplementary information for *Identification of determinants of differential chromatin accessibility through a massively parallel genome-integrated reporter assay*

Additional supplementary information can be found on the Genome Research website: `https://doi.org/10.1101/gr.263228.120`.

B.1 Supplementary Figures

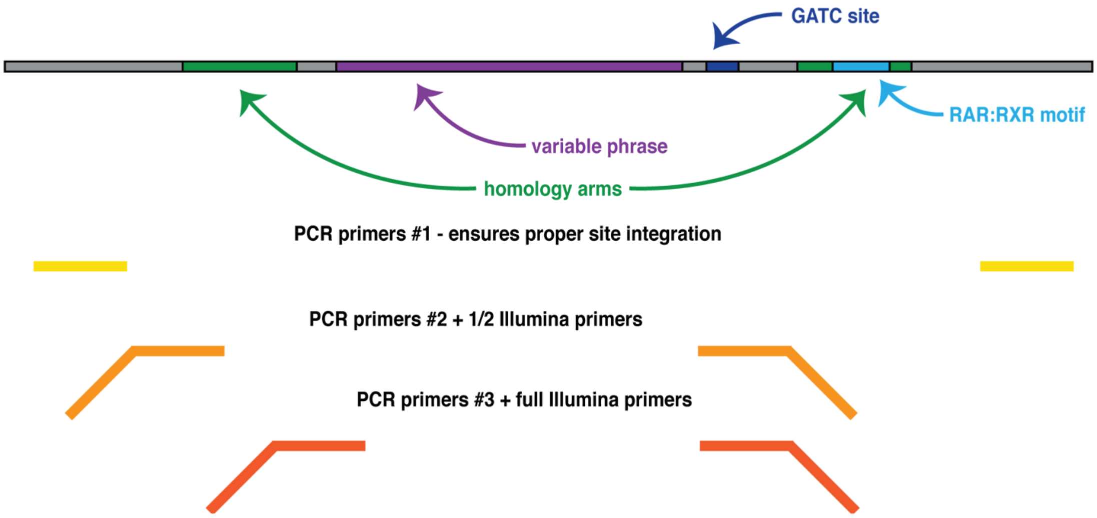

GATC site
RAR:RXR motif
variable phrase
homology arms
PCR primers #1 - ensures proper site integration
PCR primers #2 + 1/2 Illumina primers
PCR primers #3 + full Illumina primers

Figure B-1: MIAA PCR steps to select for proper integration and uncleaved library phrases. First set of PCR primers are designed to select for sequences that have been integrated at the specific genomic locus and are uncleaved by DpnI/II enzyme. Second set and third round of PCR primers enrich for add Illumina PE primers.

Appendix C

Supplementary information for *High resolution discovery of chromatin interactions*

Additional supplementary information can be found on the Nucleic Acids Research website:
`https://doi.org/10.1093/nar/gkz051`.

C.1 Supplementary Figures

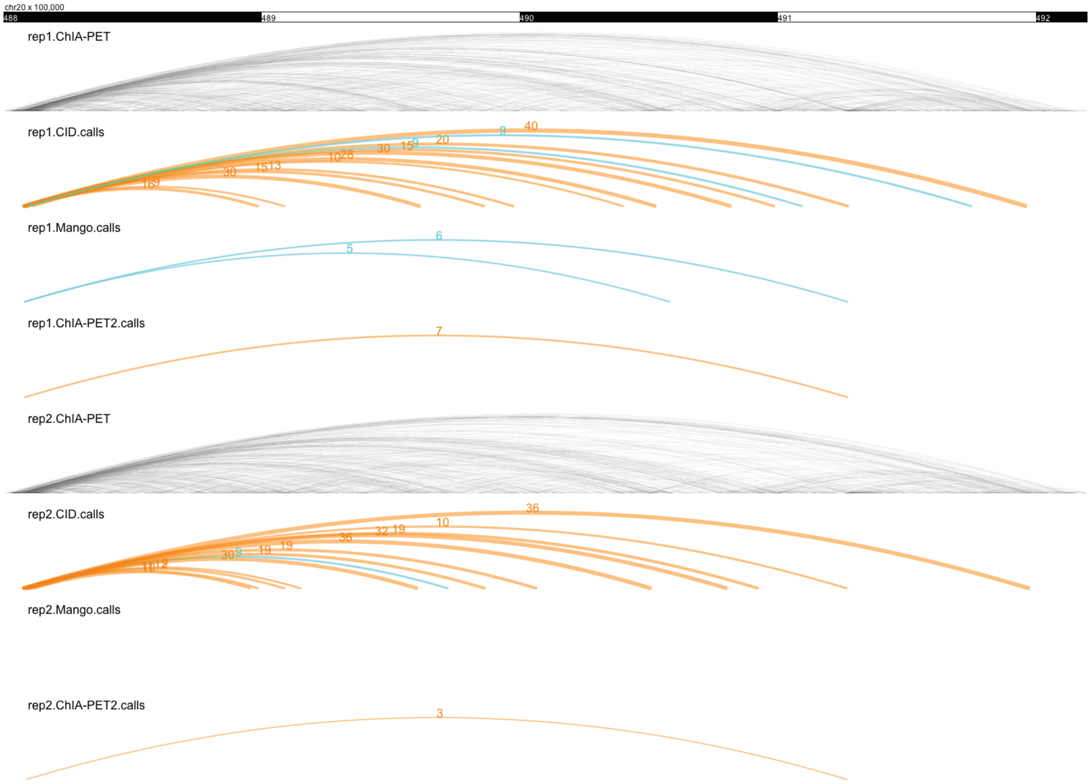


Figure C-1: CID is more sensitive and consistent at discovering ChIA-PET interactions than peak-calling-based methods. Comparison of interactions called by CID, ChIA-PET2, and Mango in the CEBPB locus using two POLR2A ChIA-PET replicates from K562 cells. The PET counts of the interactions are represented as the numeric values above the arcs. Arcs in orange represent significant interactions that are replicable across biological replicates. Arcs in cyan represent significant interactions that are not replicable across biological replicates.

C.2 Supplementary Tables

dataset label	Target	Cell line	GEO/ENCODE identifier
Ruan.K562.POLR2A	POLR2A	K562	ENCSR000BZY
Ruan.MCF-7.POLR2A	POLR2A	MCF-7	ENCSR000CAA
Snyder.K562.POLR2A	POLR2A	K562	ENCSR000FDC
Ruan.MCF-7.CTCF	CTCF	MCF-7	ENCSR000CAD
Snyder.GM12878.RAD21.2014	RAD21	GM12878	ENCSR752QCX
Snyder.GM12878.RAD21.2016	RAD21	GM12878	ENCSR981FNA
Snyder.GM12891.RAD21	RAD21	GM12891	ENCSR299VMZ
Snyder.GM12892.RAD21	RAD21	GM12892	ENCSR033GUP
Snyder.GM19238.RAD21	RAD21	GM19238	ENCSR527RXH
Snyder.GM19239.RAD21	RAD21	GM19239	ENCSR479MTN
Snyder.GM19240.RAD21	RAD21	GM19240	ENCSR312TUD
Snyder.HepG2.RAD21	RAD21	HepG2	ENCSR014ZXR
Snyder.JurkatCloneE61.RAD21	RAD21	Jurkat clone E61	ENCSR465NNU
Snyder.K562.RAD21	RAD21	K562	ENCSR000FDB
Snyder.LNCaPCloneFGC.RAD21	RAD21	LNCaP clone FGC	ENCSR011ITK
Snyder.MCF-7.RAD21	RAD21	MCF-7	ENCSR716WZI
Snyder.SU-DHL-2.RAD21	RAD21	SU-DHL-2	ENCSR466AXT
Chang.GM12878.SMC1A.HiChIP	SMC1A	GM12878	GSE80820

Table C.1: ChIP-seq and ChIA-PET datasets used for CID analyses.

Appendix D

Supplementary information for *IDR2D identifies reproducible genomic interactions*

Additional supplementary information can be found on the Nucleic Acids Research website:
`https://doi.org/10.1093/nar/gkaa030`.

D.1 Supplementary Figures

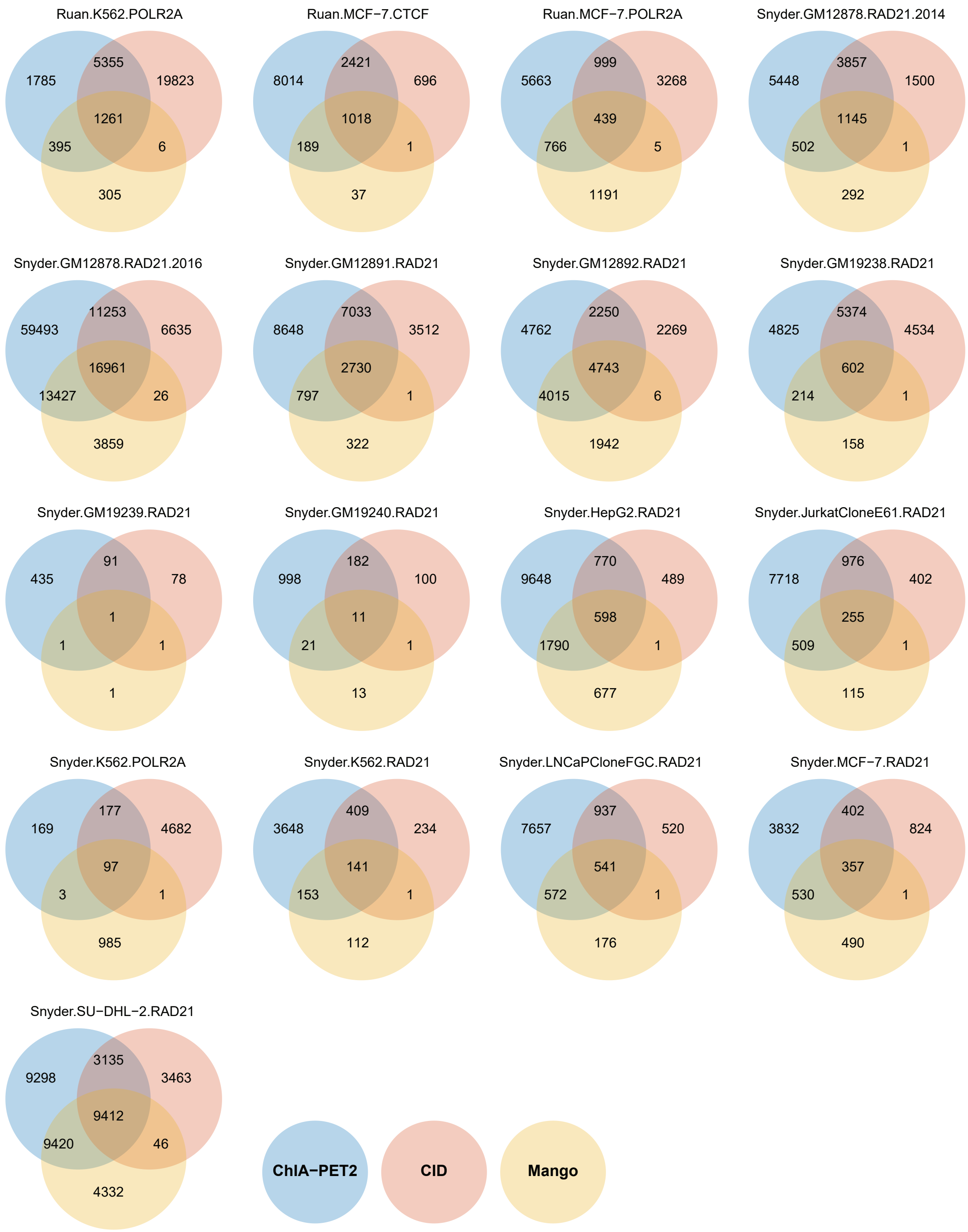


Figure D-1: Overlap between reproducible interactions identified by ChIA-PET2, CID, and Mango, and IDR2D. Each Venn diagram shows the overlap between interactions called by ChIA-PET2, CID, and Mango that had an IDR of less than 0.05 of one of 17 replicated ENCODE ChIA-PET experiments.

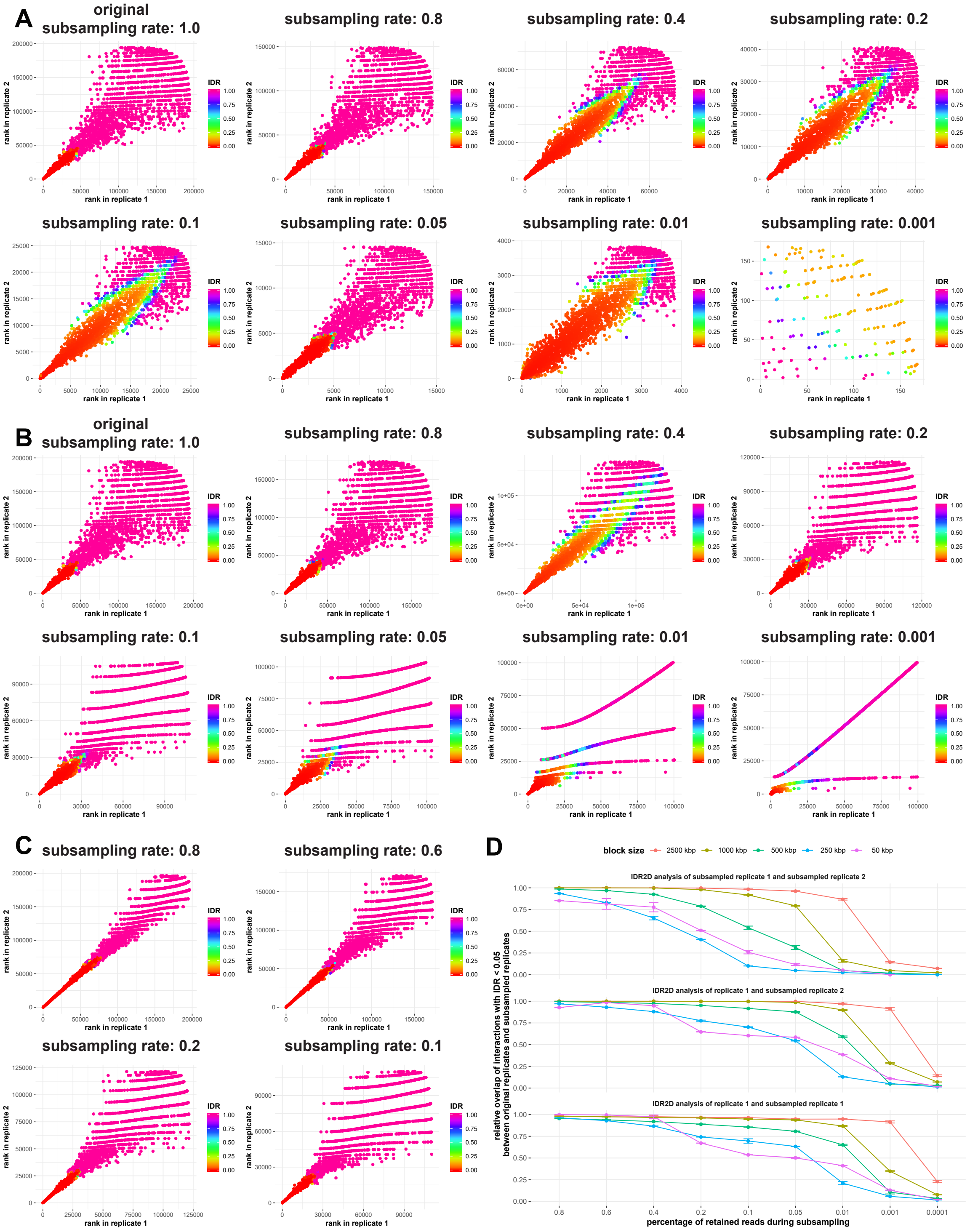


Figure D-2: Influence of sequencing depth on IDR2D analysis. All panels are based on IDR2D analysis of subsampled replicates of GSE63525, chromosome 16. Panels **A** - **C** use block sizes of 50 kbp.

Figure D-2: Influence of sequencing depth on IDR2D analysis. (**A**) Rank scatterplots of IDR2D analysis of subsampled replicate 1 and subsampled replicate 2 at read retention rates of 1.0 (all reads retained) to 0.001 (99.9% of reads removed). (**B**) Rank scatterplots of IDR2D analysis of original replicate 1 and subsampled replicate 2. (**C**) Rank scatterplots of IDR2D analysis of original replicate 1 and subsampled replicate 1. (**D**) Relative overlap of interactions with IDR < 0.05 between subsampled and original IDR2D analysis at various read retention rates and block sizes.

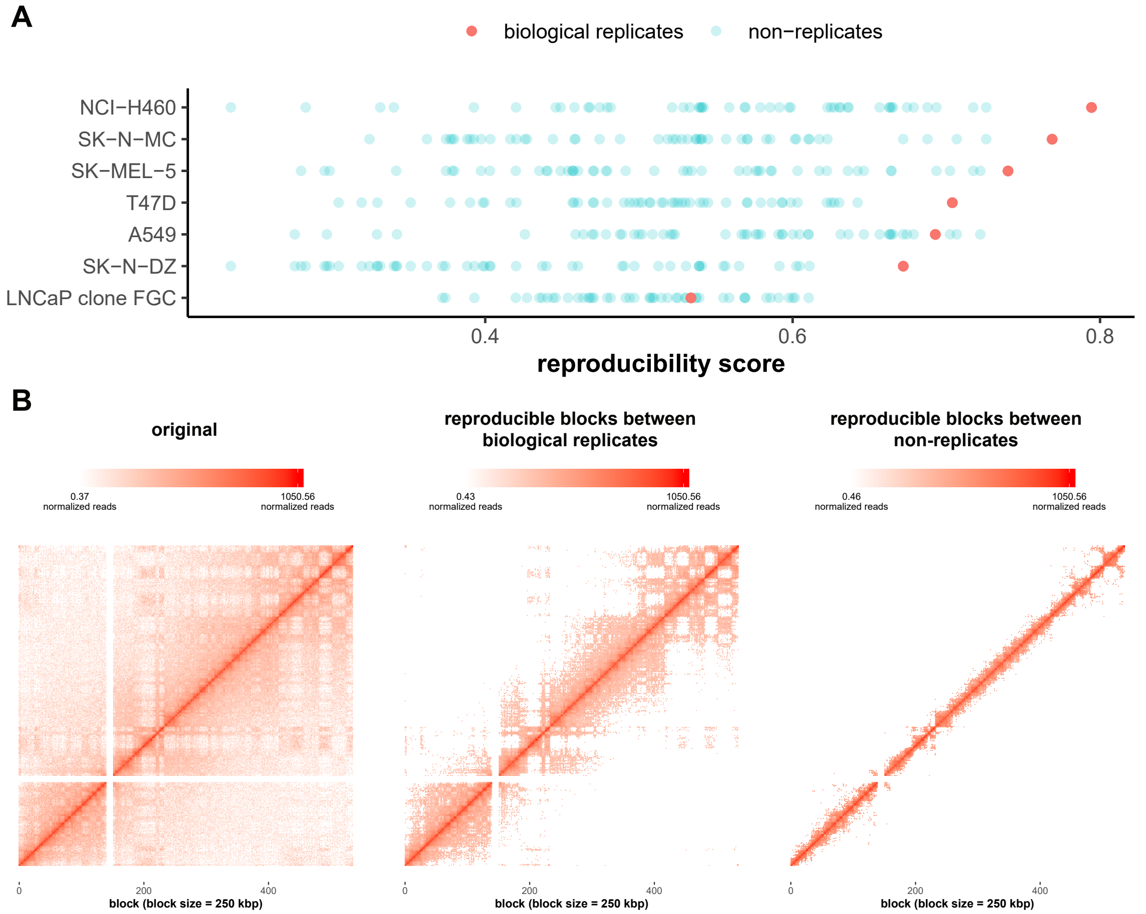


Figure D-3: Hi-C reproducibility between replicates and non-replicates. (**A**) Relative reproducibility scores of IDR2D analyses between biological replicates and non-replicates. Scores were calculated based on the number of highly reproducible blocks (IDR < 0.01) in the contact matrix, summarizing results of all chromosomes and averaged over various resolutions (block sizes of 2.5 Mbp, 1 Mbp, 500 kbp, 250 kbp, 100 kbp, 50 kbp, 25 kbp, 10 kbp, and 5 kbp). (**B**) Visualizations of contact maps of chromosome 12 showing all blocks (left), blocks with IDR < 0.05 between two biological replicates in cell line NCI-H460 (middle), and blocks with IDR < 0.05 between two non-replicates (right).

D.2 Supplementary Tables

ChIA-PET dataset identifier	**Target**	**Cell line**	**ENCODE identifier**
Ruan.K562.POLR2A	POLR2A	K562	ENCSR000BZY
Ruan.MCF-7.POLR2A	POLR2A	MCF-7	ENCSR000CAA
Snyder.K562.POLR2A	POLR2A	K562	ENCSR000FDC
Ruan.MCF-7.CTCF	CTCF	MCF-7	ENCSR000CAD
Snyder.GM12878.RAD21.2014	RAD21	GM12878	ENCSR752QCX
Snyder.GM12878.RAD21.2016	RAD21	GM12878	ENCSR981FNA
Snyder.GM12891.RAD21	RAD21	GM12891	ENCSR299VMZ
Snyder.GM12892.RAD21	RAD21	GM12892	ENCSR033GUP
Snyder.GM19238.RAD21	RAD21	GM19238	ENCSR527RXH
Snyder.GM19239.RAD21	RAD21	GM19239	ENCSR479MTN
Snyder.GM19240.RAD21	RAD21	GM19240	ENCSR312TUD
Snyder.HepG2.RAD21	RAD21	HepG2	ENCSR014ZXR
Snyder.JurkatCloneE61.RAD21	RAD21	Jurkat clone E61	ENCSR465NNU
Snyder.K562.RAD21	RAD21	K562	ENCSR000FDB
Snyder.LNCaPCloneFGC.RAD21	RAD21	LNCaP clone FGC	ENCSR011ITK
Snyder.MCF-7.RAD21	RAD21	MCF-7	ENCSR716WZI
Snyder.SU-DHL-2.RAD21	RAD21	SU-DHL-2	ENCSR466AXT
HiChIP dataset identifier	**Target**	**Cell line**	**GEO identifier**
Chang.GM12878.H3K27ac	H3K27ac	GM12878	GSE101498
Chang.K562.H3K27ac	H3K27ac	K562	GSE101498
Chang.mES.H3K27ac	H3K27ac	mES	GSE101498
Chang.MyLa.H3K27ac	H3K27ac	MyLa	GSE101498
Chang.Naive.CTCF	CTCF	Naive	GSE101498
Chang.Naive.H3K27ac	H3K27ac	Naive	GSE101498
Chang.Th17.H3K27ac	H3K27ac	Th17	GSE101498
Chang.Treg.H3K27ac	H3K27ac	Treg	GSE101498
Flynn.GM12878.Smc1a	Smc1a	GM12878	GSE80820
HiC dataset identifier		**Cell line**	**GEO identifier**
Lieberman.GM12878		GM12878	GSE63525
Lieberman.Patski		Patski	GSE71831
Skok.NSD2		multiple myeloma	GSE131651
HiC dataset identifier and cell line			**ENCODE identifier**
A549			ENCSR444WCZ
SK-N-DZ			ENCSR105KFX
SK-MEL-5			ENCSR312KHQ
LNCaP clone FGC			ENCSR346DCU
NCI-H460			ENCSR489OCU
T47D			ENCSR549MGQ
SK-N-MC			ENCSR834DXR

Table D.1: ChIA-PET, HiChIP, and HiC datasets used for IDR2D analyses.

Dataset identifier	Total int.	Rep. int.	IDR < 0.05	IDR < 0.01
Ruan.K562.POLR2A	15,029	9594	8816	4919
Ruan.MCF-7.POLR2A	23,540	14,718	2714	1251
Snyder.K562.POLR2A	20,683	9142	446	220
Ruan.MCF-7.CTCF	42,958	12,042	11,664	10,503
Snyder.GM12878.RAD21.2014	26,376	11,176	10,977	10,191
Snyder.GM12878.RAD21.2016	354,536	127,172	101,330	11,164
Snyder.GM12891.RAD21	26,180	20,144	19,245	14,628
Snyder.GM12892.RAD21	33,731	16,253	15,906	14,401
Snyder.GM19238.RAD21	90,887	12,471	11,031	1676
Snyder.GM19239.RAD21	50,387	1712	526	47
Snyder.GM19240.RAD21	6225	2226	1212	94
Snyder.HepG2.RAD21	47,544	18,308	12,819	1300
Snyder.JurkatCloneE61.RAD21	18,408	9745	889	367
Snyder.K562.RAD21	5540	4470	4352	3782
Snyder.LNCaPCloneFGC.RAD21	34,260	12,456	9718	1056
Snyder.MCF-7.RAD21	109,485	14,793	5117	639
Snyder.SU-DHL-2.RAD21	99,864	39,891	31,437	3266

Table D.2: IDR2D analysis of ChIA-PET interactions called by ChIA-PET2. Columns are (1) *total number of interactions in replicate 1*, (2) *number of reproducible interactions*, (3) *number of reproducible interactions with IDR* < 0.05, and (4) *number of reproducible interactions with IDR* < 0.01.

Dataset identifier	Total int.	Rep. int.	IDR < 0.05	IDR < 0.01
Ruan.K562.POLR2A	116,932	40,693	26,958	8698
Ruan.MCF-7.POLR2A	98,484	33,147	5064	681
Snyder.K562.POLR2A	51,323	12,352	6300	1513
Ruan.MCF-7.CTCF	53,351	14,321	4163	1162
Snyder.GM12878.RAD21.2014	49,102	16,309	6556	1733
Snyder.GM12878.RAD21.2016	394,045	105,704	36,662	10,363
Snyder.GM12891.RAD21	45,295	29,509	13,398	3633
Snyder.GM12892.RAD21	70,222	26,637	9594	2457
Snyder.GM19238.RAD21	162,936	23,444	10,573	2745
Snyder.GM19239.RAD21	98,449	4074	169	22
Snyder.GM19240.RAD21	15,473	4325	295	65
Snyder.HepG2.RAD21	53,725	16,886	1876	267
Snyder.JurkatCloneE61.RAD21	46,965	14,837	1642	228
Snyder.K562.RAD21	13,891	8791	790	104
Snyder.LNCaPCloneFGC.RAD21	66,356	19,021	2013	299
Snyder.MCF-7.RAD21	43,930	14,754	1641	262
Snyder.SU-DHL-2.RAD21	132,610	43,197	17,460	4765

Table D.3: IDR2D analysis of ChIA-PET interactions called by CID. Columns are (1) *total number of interactions in replicate 1*, (2) *number of reproducible interactions*, (3) *number of reproducible interactions with IDR* < 0.05, and (4) *number of reproducible interactions with IDR* < 0.01.

Dataset identifier	**Total int.**	**Rep. int.**	**IDR < 0.05**	**IDR < 0.01**
Ruan.K562.POLR2A	16,847	10,581	2273	1149
Ruan.MCF-7.POLR2A	23,540	14,718	2714	1251
Snyder.K562.POLR2A	17,191	7051	2419	477
Ruan.MCF-7.CTCF	41,542	11,975	1274	502
Snyder.GM12878.RAD21.2014	29,770	12,057	1993	619
Snyder.GM12878.RAD21.2016	166,543	69,897	36,597	8039
Snyder.GM12891.RAD21	28,426	21,983	3976	338
Snyder.GM12892.RAD21	37,730	17,900	11,050	1182
Snyder.GM19238.RAD21	97,679	13,084	1038	109
Snyder.GM19239.RAD21	57,254	2231	1	0
Snyder.GM19240.RAD21	6462	2219	47	14
Snyder.HepG2.RAD21	26,747	12,285	3091	494
Snyder.JurkatCloneE61.RAD21	18,408	9745	889	367
Snyder.K562.RAD21	7203	5474	412	98
Snyder.LNCaPCloneFGC.RAD21	24,065	11,992	1308	420
Snyder.MCF-7.RAD21	20,277	10,412	1453	511
Snyder.SU-DHL-2.RAD21	71,043	34,226	24,909	2552

Table D.4: IDR2D analysis of ChIA-PET interactions called by Mango. Columns are (1) *total number of interactions in replicate 1*, (2) *number of reproducible interactions*, (3) *number of reproducible interactions with IDR* < 0.05, and (4) *number of reproducible interactions with IDR* < 0.01.

Dataset identifier	**Total int.**	**Rep. int.**	**IDR < 0.05**	**IDR < 0.01**
Chang.GM12878.H3K27ac	8,170,291	4,076,304	2,133,248	408,385
Chang.K562.H3K27ac	6,059,672	2,724,718	991,394	295,193
Chang.mES.H3K27ac	4,997,523	2,593,482	1,902,221	118,607
Chang.MyLa.H3K27ac	6,010,073	2,745,081	1,052,481	345,503
Chang.Naive.CTCF	1,368,563	52,845	6885	1607
Chang.Naive.H3K27ac	69,861	49,823	13,504	1821
Chang.Th17.H3K27ac	2,135,349	1,418,902	889,736	188,608
Chang.Treg.H3K27ac	365,827	253,529	105,227	15,856
Flynn.GM12878.Smc1a	6,036,994	2,232,884	369,344	4166

Table D.5: IDR2D analysis of HiChIP interactions called by CID. Columns are (1) *total number of interactions in replicate 1*, (2) *number of reproducible interactions*, (3) *number of reproducible interactions with IDR* < 0.05, and (4) *number of reproducible interactions with IDR* < 0.01.

Dataset identifier	Total int.	Rep. int.	IDR < 0.05	IDR < 0.01
Flynn.GM12878.Smc1a	2,312,685	1,126,785	71,850	2892

Table D.6: IDR2D analysis of HiChIP interactions called by hichipper. Columns are (1) *total number of interactions in replicate 1*, (2) *number of reproducible interactions*, (3) *number of reproducible interactions with IDR* < 0.05, and (4) *number of reproducible interactions with IDR* < 0.01.

D.3 Supplementary Methods

D.3.1 IDR2D procedure

We first define $\Omega_{x_{i,1}}$ as the set of interactions in replicate 2 that overlaps with the interaction $x_{i,1}$ in replicate 1. Two interactions $x_{i,1}$ and $x_{j,2}$ are overlapping, if both of their interaction anchors are overlapping or are less than a predefined maximum gap away from each other.

If interaction $x_{i,1}$ overlaps with more than one interaction in replicate 2, the ambiguous mapping is resolved by choosing $x_{*,2}$ in the following way:

$$x_{*,2} = \underset{x_{j,2} \in \Omega_{x_{i,1}}}{\operatorname{argmin}} f(x_{i,1}, x_{j,2}), \tag{D.1}$$

where $f(\cdot, \cdot)$ is the *ambiguity resolution value* (ARV) between an interaction in replicate 1 and an overlapping interaction in replicate 2. The three supported ambiguity resolution methods are depicted in figure 1D of the main manuscript.

To estimate irreproducible discovery rates, IDR2D uses a two-component copula mixture model that is identical to the model described for the original IDR method, with one component for irreproducible interaction pairs and another one for reproducible interaction pairs. After a bijective mapping between interactions in replicate 1 and interactions in replicate 2 is established, the posterior probability that an interaction pair i denoted as $(x_{i,1}, x_{i,2})$, with significance values from replicates 1 and 2 belongs to the irreproducible component is defined as

$$\Pr((x_{i,1}, x_{i,2}); \theta_{\text{irrep}}, \theta_{\text{rep}}) = \frac{C_{\text{irrep}}}{C_{\text{irrep}} + C_{\text{reproducible}}} \tag{D.2}$$

$$C_{\text{irrep}} = \pi_0 h_0 \left(G^{-1}\left(F_1\left(x_{i,1}\right)\right), G^{-1}\left(F_2\left(x_{i,2}\right)\right)\right) \tag{D.3}$$

$$C_{\text{rep}} = \pi_1 h_1 \left(G^{-1}\left(F_1\left(x_{i,1}\right)\right), G^{-1}\left(F_2\left(x_{i,2}\right)\right)\right) \tag{D.4}$$

$$h_0 \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\right) \tag{D.5}$$

$$h_1 \sim \mathcal{N}\left(\begin{pmatrix} \mu_1 \\ \mu_1 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \rho_1\sigma_1^2 \\ \rho_1\sigma_1^2 & \sigma_1^2 \end{pmatrix}\right), \tag{D.6}$$

where $\theta_{\text{irrep}} = \pi_0$ and $\theta_{\text{rep}} = (\pi_1, \mu_1, \sigma_1^2, \rho_1)$, the estimated parameters of the two components (see Li et al. [1] section 2.2.3 for details on the estimation procedure). For a definition of G, and F_1 and F_2, the marginal distributions of the coordinates in the two replicates, see Li et al. [1] section 2.2.2 for details.

Bibliography

[1] Qunhua Li, James B. Brown, Haiyan Huang, and Peter J. Bickel. Measuring reproducibility of high-throughput experiments. *Ann. Appl. Stat.*, 5(3):1752–1779, 09 2011.

Appendix E

Supplementary information for *spatzie: An R package for identifying significant transcription factor motif co-enrichment from enhancer-promoter interactions*

Supplementary information can also be found on the bioRxiv website:
`https://doi.org/10.1101/2021.05.25.445606`.

E.1 Supplementary Figures

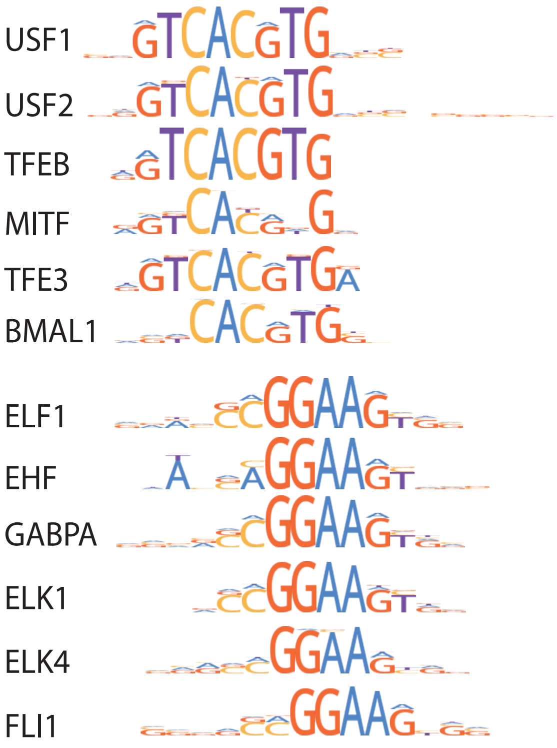


Figure E-1: Transcription factor motif position weight matrices that are discovered to have significant co-enrichment from USF1 and ELF1 simulated enhancer:promoter interaction task show high motif similarity to USF1 and ELF1 motifs.

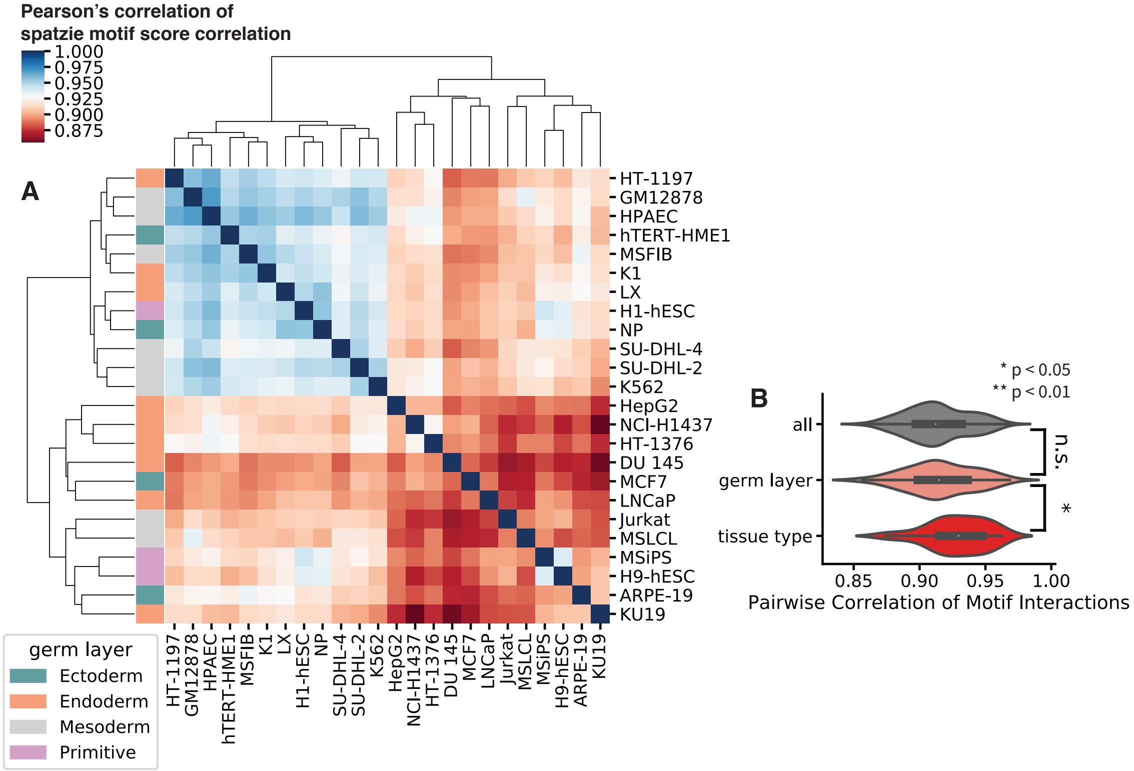


Figure E-2: **(A)** Pearson correlation of spatzie motif score of CID discovered enhancer:promoter interactions for RAD21 ChIA-PET of 24 human cell types. **(B)** Pairwise correlation of spatzie co-enriched motifs from CID enhancer:promoter interactions are higher among samples that are from the same tissue.

E.2 Supplementary Tables

dataset label	Target	Cell line	GEO/ENCODE identifier
USF1 rep1	ChIP	MEL	ENCFF996MWJ
USF1 rep2	ChIP	MEL	ENCFF550MZG
ELF1 rep1	ChIP	MEL	ENCFF186NAS
ELF1 rep2	ChIP	MEL	ENCFF592ERV
ARPE	RAD21	ARPE-19	GSE134745
DU145	RAD21	DU-145	GSE134745
MSFIB	RAD21	fibroblast	GSE134745
GM12878	RAD21	GM12878	GSE134745
H1	RAD21	H1	GSE134745
H9	RAD21	H9	GSE134745
LX	RAD21	hepatocyte	GSE134745
HepG2	RAD21	HepG2	GSE134745
HT1197	RAD21	HT-1197	GSE134745
HT1376	RAD21	HT-1376	GSE134745
HMTERT	RAD21	hTERT-HME1	GSE134745
Jurkat	RAD21	Jurkat-Clone-E6-1	GSE134745
AKTHY	RAD21	K1	GSE134745
K562	RAD21	K562	GSE134745
KU19	RAD21	KU-19-19	GSE134745
LNCAP	RAD21	LNCAP	GSE134745
MCF7	RAD21	MCF-7	GSE134745
MSIPS	RAD21	MSiPS	GSE134745
MSLCL	RAD21	MSLCL	GSE134745
H1437	RAD21	NCI-H1437	GSE134745
NP	RAD21	neural progenitor cells	GSE134745
ECS	RAD21	pulmonary artery endothelial cells	GSE134745
DHL2	RAD21	SU-DHL-2	GSE134745
DHL4	RAD21	SU-DHL-4	GSE134745

Table E.1: ChIP-seq and ChIA-PET datasets used for spatzie analyses.

E.3 Supplementary Methods

For analysis of human RAD21 ChIA-PET data for co-enrichment of motifs underlying enhancer:promoter interactions, we applied spatzie to both the interactions that were provided by Grubert et al. using a custom method for generating a unified interaction set from many ChIA-PET samples and then using PET support to call cell type-specific interaction events as well as a unified interaction set that was generated by calling interactions with CID [1].

E.3.1 Calling interaction events with CID

For each ChIA-PET experiment, we first used Mango 1.2.1 [3] (downloaded from `https://github.com/dphansti/mango`) to remove linker sequences and reads potentially due to polymerase chain reaction duplication, and aligned the raw reads using bowtie 1.2.3 (steps 1 - 3 in Mango pipeline). Mango was executed with the *reportallpairs* flag and the recommended parameter settings for the ChIA-PET Tn5 tagmentation protocol: *-keepempty TRUE -maxlength 1000 -shortreads FALSE*. CID (downloaded from `https://groups.csail.mit.edu/cgs/gem/versions.html`) was then run on the BEDPE file produced by Mango, using default parameters. Lastly, we used MICC [2] to assess the significance of all interactions identified by CID that are supported by more than one PET read.

E.3.2 Generating a unified CID interaction set

For the CID interaction set, we first ran CID independently on all 48 samples (24 cell types; 2 replicates for each cell type). We then took the union of interactions called in all 48 samples. To merge overlapping interactions, we independently merged overlapping anchor1 and anchor2. We then removed interactions containing anchors that were larger than 20kb in size or interactions where as an artifact of merging, anchor1 and anchor2 overlapped. This resulted in a set of 71,643 joint CID interactions.

E.3.3 Calling CID enhancer:promoter interactions for each cell type

We then obtained PET support from each of the 48 samples for the set of joint CID interactions and based on the correlation between replicates removed two samples (H1 rep2 and pulomonary artery endothelial cell rep2) that did not appear to have consistent correlation with their other cell type replicate and cell types within the same tissue. To obtain the interactions for each cell type from our joint CID interactions set, we looked at PET support and replicate correlation. For samples with two replicates, we called an interaction event within that cell type if it had greater than 1 PET in both replicates. For samples with one replicate, (H1 and pulomonary artery endothelial cell), we called an interaction event if the MICC FDR value of the event was less than 0.05. Interaction events were then filtered to those where one anchor was within 2.5kb of a promoter and the other was promoter distal (not within 2.5kb of a promoter). The selection criteria for calling cell type-specific allowed us to obtain a sufficient number of enhancer:promoter interaction events to run spatzie for co-motif enrichment (range 2,178-14,113 enhancer:promoter interactions per cell type) and were consistent with the range in the number of enhancer:promoter events discovered in Grubert et al. (range 1,476-11,850 enhancer:promoter interactions per cell type).

E.3.4 Analysis of spatzie results for CID and Grubert et al. interactions

After co-enrichment of motifs at enhancer:promoter interactions was computed with spatzie, we analyzed the correlation between co-enrichment scores. For Grubert et al. interactions, we used only motif pairs where co-enrichment was significant under multiple hypothesis correction in at least one cell type. For CID interactions, we used all motif pairs as limiting to interactions that were significant under multiple hypothesis correction resulted in less meaningful clustering of motif co-enrichment scores by tissue type.

Bibliography

[1] Y. Guo, K. Krismer, M. Closser, H. Wichterle, and D. K. Gifford. High resolution discovery of chromatin interactions. *Nucleic Acids Res*, 47(6):e35, 04 2019.

[2] C. He, M. Q. Zhang, and X. Wang. MICC: an R package for identifying chromatin interactions from ChIA-PET data. *Bioinformatics*, 31(23):3832–3834, Dec 2015.

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