UNIVERSITY OF CALIFORNIA SAN DIEGO

Mining Biological Data with Machine Learning to Obtain New Insights

A dissertation submitted in partial satisfaction of the

requirements for the degree Doctor of Philosophy

in

Chemistry

by

Chengyu Liu

Committee in charge:

Professor Wei Wang, Chair
Professor Elizabeth Komives
Professor Michael Gilson
Professor Yoav Freund
Professor Cornelis Murre
Professor Yuen-Zhou Joel

2022

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

University of California San Diego

2022

DEDICATION

I would like to dedicate this thesis to my parents who supported me through my scientific adventure.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

First, I am deeply thankful to my PhD advisor, Prof. Wei Wang, for his support and guidance during my PhD. Together, we have done a lot of projects in different areas of biological data mining and resulted in 6 manuscripts. Second, I would like to express my sincere gratitude to my committee members Prof. Elizabeth Komives, Prof. Michael Gilson, Prof. Yoav Freund, Prof. Cornelis Murre and Prof. Yuen-Zhou Joel for their helpful suggestions and advice for my research projects and continuous support for my scientific adventure. Third, I would like to thank my collaborators Dr. Yuchen Liu, Dr. Hsien-Da Huang and Dr. Richard Ainsworth for their hard work and brilliant ideas that lead to the fruition of circular RNA and information flow analysis projects. Last but not the least, I would also like to thank all the members of the Wang lab, from whom I have received inspiration and support.

The Introduction is, in part, based on the material currently being prepared for submission as “Examination of Hi-C Network Properties under Local and Global Centrality Measures” Ainsworth, R., Liu, C. & Wang, W. The Introduction is also, in part, based on material as it appears as "Contextual regression: an accurate and conveniently interpretable nonlinear model for mining discovery from scientific data" in *arxiv*, 2017. Liu, C., Wang, W. The Introduction is also, in part, based on material as it appears as "Biogenesis mechanisms of circular RNA can be categorized through feature extraction of a machine learning model" in *Bioinformatics*, 2019. Liu, C., Liu, Y.-C., Huang, H.-D. & Wang, W. The Introduction is also, in part, based on the material currently being prepared for submission as “Neural network facilitated ab initio derivation of linear formula: A case study on formulating the relationship

between DNA motifs and gene expression" Liu, C., Wang, W. The dissertation author was the primary investigator and author of these papers

Chapter 1, in full, is currently being prepared for submission of the material as "Examination of Hi-C Network Properties under Local and Global Centrality Measures" Ainsworth, R., Liu, C. & Wang, W. The dissertation author was a co-primary investigator and author of this paper.

Chapter 2, in full, is a reformatted reprint of the material as it appears as "Contextual regression: an accurate and conveniently interpretable nonlinear model for mining discovery from scientific data" in *arxiv*, 2017. Liu, C., Wang, W. The dissertation author was the primary investigator and author of this paper.

Chapter 3, in full, is a reformatted reprint of the material as it appears as "Biogenesis mechanisms of circular RNA can be categorized through feature extraction of a machine learning model" in *Bioinformatics*, 2019. Liu, C., Liu, Y.-C., Huang, H.-D. & Wang, W. The dissertation author was a co-primary investigator and author of this paper.

Chapter 4, in full, is currently being prepared for submission of the material as "Neural network facilitated ab initio derivation of linear formula: A case study on formulating the relationship between DNA motifs and gene expression" Liu, C., Wang, W. The dissertation author was the primary investigator and author of this paper.

VITA

2014	Bachelor of Arts, Washington University in Saint Louis
2017	Master of Science, University of California San Diego
2022	Doctor of Philosophy, University of California San Diego

Publications:

Ainsworth, R.*, **Liu, C.*** & Wang, W. Examination of Hi-C Network Properties under Local and Global Centrality Measures (in preparation)

Liu, C., Wang, W. Contextual regression: an accurate and conveniently interpretable nonlinear model for mining discovery from scientific data. arXiv preprint arXiv:171010728. 2017

Liu, C.*, Liu, Y.-C*., Huang, H.-D. & Wang, W. Biogenesis mechanisms of circular RNA can be categorized through feature extraction of a machine learning model. *Bioinformatics* 35, 4867–4870 (2019).

Liu, C., Wang, W. Neural network facilitated ab initio derivation of linear formula: A case study on formulating the relationship between DNA motifs and gene expression (in preparation)

Wang Z., Liu, C., Wang J., **Liu, C.**, Wang, M., Ngo, V., Wang, W. Predicting regional somatic mutation rates using DNA motifs (in preparation)

Wang, M., Zhang, K., Ngo, V., **Liu, C.**, Fan, S., Whitaker, JW., Chen, Y., Ai, R., Chen, Z., Wang, J, Zheng, L., Wang, W. Identification of DNA motifs that regulate DNA methylation. Nucleic Acids Res. 2019; 47(13):6753–68

*** co-first authors**

ABSTRACT OF THE DISSERTATION

Mining Biological Data with Machine Learning to Obtain New Insights

by

Chengyu Liu

Doctor of Philosophy in Chemistry

University of California San Diego, 2022

Professor Wei Wang, Chair

Molecular biology is a highly complicated subject due to the complexity of cells and organisms as systems. To decode them, recent efforts are dedicated to generating data at an increasingly large scale. Luckily, the emergence of two major methods in bioinformatics, network analysis and machine learning, have provided us with tools to digest these data into biological insights. During my PhD, I have improved and developed methods in network analysis and machine learning to make them more efficient and generate human understandable insights. In chapter I, colleague and I developed a new version of information flow analysis algorithm with time complexity of $O(n^2m)$ instead of $O(n^4)$. We also developed a GPU version of this algorithm that can achieve a 17.4x speed up over the original implementation on a Hi-C network of a relatively small chromosome. This would allow us to apply information flow analysis on much larger biological networks. In chapter II, I developed contextual regression, a deep neural network framework that can achieve state of the art prediction accuracy while generating human interpretable outputs. We tested this method on a simulated biological signal dataset and found our method not only has good prediction performance, but also is able to uncover the ground truth model even under a noise level of 80%. To test this method on real world data, we also applied this method to analyze the effect of histone marks on open chromatin. The model not only outperforms previous models in terms of prediction, but also uncover histone mark patterns that are associated with open chromatin formation. In chapter III, colleague and I applied the contextual regression framework on circular RNA data. Combining this new method and the circNet database, we uncovered 7 types of circular RNA genesis mechanisms. This discovery supports the hypothesis that multiple biogenesis mechanisms co-exist for different subsets of human circRNAs. In chapter IV, I developed a new contextual

regression architecture and applied it to the task of predicting RNA expression level from gene sequences. The new architecture not only achieved state of the art accuracy, but also learned important motifs that can affect gene expression. I then developed a selection process and generated a linear formula of 300 sequence motifs that can predict expression level as accurately as deep neural network models. The most influential motifs among these motifs are deeply involved in development, regulation and stimulus response. When we applied this linear formula to different cell lines, we found the parameters of this formula are associated with biological properties such as epigenetic influence through development, cell lineages and difference of expression patterns among cell lines. We anticipate that these studies have created tools towards the challenge of mining biological insights from increasing amounts of biological data.

INTRODUCTION

Decades have passed since the human genome has been sequenced[1,2] and most of the proteins in human cells are discovered[3,4], however, many mysteries of how human body functions are still present. In an attempt to uncover them, large biological datasets are proposed and generated. The Encyclopedia of DNA Elements (ENCODE) project[5] has produced 5992 experimental datasets, covering RNA transcription, chromatin structure and modification, DNA methylation, chromatin looping, and occupancy by transcription factors and RNA-binding proteins of various cells in both human and mice. The Roadmap project[6] has generated epigenomic data of 111 human cell lines. Other databases for circular RNA[7], DNA motifs[8,9] and protein interactions[10,11] are also generated.

With the growing number of databases comes a growing need for analysis methods and tools. Recently, network analysis[12,13] and deep learning[14–16] have emerged as effective tools to digest these datasets. Network analysis can be applied to study biological networks such as transcription factor networks[17] and protein-protein interactions[18]. Deep learning can be used to predict various biological properties such as open chromatin[19], histone marks[15], protein binding[14] and gene expression levels[16]. However, some of the major methods in these two domains have room for improvement. In network analysis, information flow analysis[12] has shown its potential to find globally important nodes in the network. But due to its heavy computing cost, it is limited to be used on small networks of a few thousand nodes. Deep learning models[14–16] can achieve state of the art prediction accuracy on biological datasets. However, unlike simple models such as linear regression[20], they are highly uninterpretable due to their complexity. This attribute

makes it very hard to derive biological insights from trained deep learning models. Thus we have devoted our efforts to improve the computing speed of information flow analysis and the interpretability of deep neural networks.

In Chapter 1, colleague and I show that an alternative version of information flow analysis[12] can reduce its time complexity from $O(n^4)$ to $O(n^2m)$ while generating the same outputs. On top of that, we developed a GPU implementation of this algorithm to further speed up the computing process.

Information flow analysis[12] is an effective way for finding important nodes in a network. Unlike degree, which only measures local centrality, information flow score can evaluate the global importance of a node in the network. This algorithm can be applied not only to protein-protein interaction networks, but also a promising method for analyzing Hi-C interaction networks. However, this algorithm has an unaffordable computational cost. It has a time complexity of $O(n^4)$ which limits its application to small size networks with less than a thousand nodes. A Hi-C network with resolution at 5kbp has tens of thousands of nodes. Thus speeding it up is essential for this application.

In Chapter 2, I develop a novel neural network framework called contextual regression that can make predictions as accurate as a deep neural network while generating interpretable outputs for mining biological insights. I tested this framework on simulated biological signal data and found that it can discover the ground truth model even under high noise levels. I also applied it to find important histone marks for open chromatin formation. The model not only achieved state of the art accuracy on the task of predicting open chromatin from histone marks, but also

identified H3K4me3 and H3K18ac as open chromatin associated motifs which are not well known previously.

Deep neural networks have emerged as the model with the highest accuracy in many tasks[14,15]. However, their black box nature has limited our ability to derive insight from their parameters. Interpretation methods such as LIME[21], DeepLift[22] and SHAP[23] have been developed to identify important features that affect model decision, but all of them are limited to be applied to existing models. They not only take extra time to generate interpretation, but also not allow users to apply regularization on the model to enforce restraint on interpretation such as sparsity. Thus a deep neural network framework that generates human interpretable results is a better solution in this aspect.

In Chapter 3, colleague and I developed a contextual regression model[19] and applied it to study the mechanism of circular RNA genesis. The model we developed achieved 72.6% prediction accuracy and 0.8 AUC on the dataset. It also discovered that human circular RNAs can be divided into 7 subgroups with main sequence features: RNA editing sites, simple repeat sequences, self-chains, RNA binding protein binding sites and CpG islands. This supports the hypothesis that multiple genesis mechanisms coexist in human circular RNA.

Circular RNAs play a very important role in the cell. They can function as miRNA sponge[24] and regulate transcription of parental genes by interacting with Pol II[25]. They are found to be important for various biological processes such as brain function[26] and fetal development[27]. Circular RNAs are generated through the process of exon circulation which is a process hypothesized to compete with pre-mRNA splicing[28]. Previously, researchers have suggested that

circular RNAs can be divided into multiple subsets with different mechanisms of genesis and these mechanisms are associated with sequence features[25,29–34]. The CircNet[35] database combined with the data mining capability of contextual regression[19] can help us shed light on the major factors of circular RNA genesis.

In Chapter 4, I developed a motif finding contextual regression model and applied it to find the important sequence motifs for gene expression. This model not only achieved the same accuracy as state of the art models[16], but also allowed us to derive a linear formula of gene expression level based on sequence motif counts. We then selected the top 300 motifs for predicting gene expression level and trained linear formulas of gene expression on 154 cell lines[5,6]. Among the high weight motifs, We found that their model parameters can accurately characterize the similarity between cells and their parameter variance across cell lines are correlated with the function of their associated genes.

Mathematical models are highly useful tools for understanding biological data. By training a model to predict a biological phenomenon from biological properties, the model allows us to identify the contribution of each property to the phenomenon. However, with the rise of deep neural network[36], the human interpretability of the models is falling[37] despite improving prediction accuracy. This is especially a problem in scientific research since we need to understand the underlying mechanism rather than only have a prediction model that works. But not all deep neural networks are uninterpretable. Material science, for instance, uses neural network potential (NNP) models[38,39] to predict system energy. This model generates the energy of each bond and atom in the system and then sums them up to get the result. In this way, the total

energy can be expressed by a summation formula of energies outputted by the neural network and the role of each atom and bond can be inferred from their individual energy value. Similarly, a model that can express biological phenomena as a linear formula of sequence motifs will also provide us with insights into the underlying biology.

References

1. Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., Funke, R., Gage, D., Harris, K., Heaford, A., Howland, J., Kann, L., Lehoczky, J., LeVine, R., McEwan, P., McKernan, K., Meldrim, J., Mesirov, J. P., Miranda, C., Morris, W., Naylor, J., Raymond, C., Rosetti, M., Santos, R., Sheridan, A., Sougnez, C., Stange-Thomann, Y., Stojanovic, N., Subramanian, A., Wyman, D., Rogers, J., Sulston, J., Ainscough, R., Beck, S., Bentley, D., Burton, J., Clee, C., Carter, N., Coulson, A., Deadman, R., Deloukas, P., Dunham, A., Dunham, I., Durbin, R., French, L., Grafham, D., Gregory, S., Hubbard, T., Humphray, S., Hunt, A., Jones, M., Lloyd, C., McMurray, A., Matthews, L., Mercer, S., Milne, S., Mullikin, J. C., Mungall, A., Plumb, R., Ross, M., Shownkeen, R., Sims, S., Waterston, R. H., Wilson, R. K., Hillier, L. W., McPherson, J. D., Marra, M. A., Mardis, E. R., Fulton, L. A., Chinwalla, A. T., Pepin, K. H., Gish, W. R., Chissoe, S. L., Wendl, M. C., Delehaunty, K. D., Miner, T. L., Delehaunty, A., Kramer, J. B., Cook, L. L., Fulton, R. S., Johnson, D. L., Minx, P. J., Clifton, S. W., Hawkins, T., Branscomb, E., Predki, P., Richardson, P., Wenning, S., Slezak, T., Doggett, N., Cheng, J. F., Olsen, A., Lucas, S., Elkin, C., Uberbacher, E., Frazier, M., Gibbs, R. A., Muzny, D. M., Scherer, S. E., Bouck, J. B., Sodergren, E. J., Worley, K. C., Rives, C. M., Gorrell, J. H., Metzker, M. L., Naylor, S. L., Kucherlapati, R. S., Nelson, D. L., Weinstock, G. M., Sakaki, Y., Fujiyama, A., Hattori, M., Yada, T., Toyoda, A., Itoh, T., Kawagoe, C., Watanabe, H., Totoki, Y., Taylor, T., Weissenbach, J., Heilig, R., Saurin, W., Artiguenave, F., Brottier, P., Bruls, T., Pelletier, E., Robert, C., Wincker, P., Smith, D. R.,

Doucette-Stamm, L., Rubenfield, M., Weinstock, K., Lee, H. M., Dubois, J., Rosenthal, A., Platzer, M., Nyakatura, G., Taudien, S., Rump, A., Yang, H., Yu, J., Wang, J., Huang, G., Gu, J., Hood, L., Rowen, L., Madan, A., Qin, S., Davis, R. W., Federspiel, N. A., Abola, A. P., Proctor, M. J., Myers, R. M., Schmutz, J., Dickson, M., Grimwood, J., Cox, D. R., Olson, M. V., Kaul, R., Raymond, C., Shimizu, N., Kawasaki, K., Minoshima, S., Evans, G. A., Athanasiou, M., Schultz, R., Roe, B. A., Chen, F., Pan, H., Ramser, J., Lehrach, H., Reinhardt, R., McCombie, W. R., de la Bastide, M., Dedhia, N., Blöcker, H., Hornischer, K., Nordsiek, G., Agarwala, R., Aravind, L., Bailey, J. A., Bateman, A., Batzoglou, S., Birney, E., Bork, P., Brown, D. G., Burge, C. B., Cerutti, L., Chen, H. C., Church, D., Clamp, M., Copley, R. R., Doerks, T., Eddy, S. R., Eichler, E. E., Furey, T. S., Galagan, J., Gilbert, J. G., Harmon, C., Hayashizaki, Y., Haussler, D., Hermjakob, H., Hokamp, K., Jang, W., Johnson, L. S., Jones, T. A., Kasif, S., Kaspryzk, A., Kennedy, S., Kent, W. J., Kitts, P., Koonin, E. V., Korf, I., Kulp, D., Lancet, D., Lowe, T. M., McLysaght, A., Mikkelsen, T., Moran, J. V., Mulder, N., Pollara, V. J., Ponting, C. P., Schuler, G., Schultz, J., Slater, G., Smit, A. F., Stupka, E., Szustakowki, J., Thierry-Mieg, D., Thierry-Mieg, J., Wagner, L., Wallis, J., Wheeler, R., Williams, A., Wolf, Y. I., Wolfe, K. H., Yang, S. P., Yeh, R. F., Collins, F., Guyer, M. S., Peterson, J., Felsenfeld, A., Wetterstrand, K. A., Patrinos, A., Morgan, M. J., de Jong, P., Catanese, J. J., Osoegawa, K., Shizuya, H., Choi, S., Chen, Y. J., Szustakowki, J. & International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature* **409,** 860–921 (2001).

2. Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., Smith, H. O., Yandell, M., Evans, C. A., Holt, R. A., Gocayne, J. D., Amanatides, P., Ballew, R. M.,

Huson, D. H., Wortman, J. R., Zhang, Q., Kodira, C. D., Zheng, X. H., Chen, L., Skupski, M., Subramanian, G., Thomas, P. D., Zhang, J., Gabor Miklos, G. L., Nelson, C., Broder, S., Clark, A. G., Nadeau, J., McKusick, V. A., Zinder, N., Levine, A. J., Roberts, R. J., Simon, M., Slayman, C., Hunkapiller, M., Bolanos, R., Delcher, A., Dew, I., Fasulo, D., Flanigan, M., Florea, L., Halpern, A., Hannenhalli, S., Kravitz, S., Levy, S., Mobarry, C., Reinert, K., Remington, K., Abu-Threideh, J., Beasley, E., Biddick, K., Bonazzi, V., Brandon, R., Cargill, M., Chandramouliswaran, I., Charlab, R., Chaturvedi, K., Deng, Z., Di Francesco, V., Dunn, P., Eilbeck, K., Evangelista, C., Gabrielian, A. E., Gan, W., Ge, W., Gong, F., Gu, Z., Guan, P., Heiman, T. J., Higgins, M. E., Ji, R. R., Ke, Z., Ketchum, K. A., Lai, Z., Lei, Y., Li, Z., Li, J., Liang, Y., Lin, X., Lu, F., Merkulov, G. V., Milshina, N., Moore, H. M., Naik, A. K., Narayan, V. A., Neelam, B., Nusskern, D., Rusch, D. B., Salzberg, S., Shao, W., Shue, B., Sun, J., Wang, Z., Wang, A., Wang, X., Wang, J., Wei, M., Wides, R., Xiao, C., Yan, C., Yao, A., Ye, J., Zhan, M., Zhang, W., Zhang, H., Zhao, Q., Zheng, L., Zhong, F., Zhong, W., Zhu, S., Zhao, S., Gilbert, D., Baumhueter, S., Spier, G., Carter, C., Cravchik, A., Woodage, T., Ali, F., An, H., Awe, A., Baldwin, D., Baden, H., Barnstead, M., Barrow, I., Beeson, K., Busam, D., Carver, A., Center, A., Cheng, M. L., Curry, L., Danaher, S., Davenport, L., Desilets, R., Dietz, S., Dodson, K., Doup, L., Ferriera, S., Garg, N., Gluecksmann, A., Hart, B., Haynes, J., Haynes, C., Heiner, C., Hladun, S., Hostin, D., Houck, J., Howland, T., Ibegwam, C., Johnson, J., Kalush, F., Kline, L., Koduru, S., Love, A., Mann, F., May, D., McCawley, S., McIntosh, T., McMullen, I., Moy, M., Moy, L., Murphy, B., Nelson, K., Pfannkoch, C., Pratts, E., Puri, V., Qureshi, H., Reardon, M., Rodriguez, R., Rogers, Y. H., Romblad, D., Ruhfel, B., Scott, R., Sitter, C., Smallwood, M.,

Stewart, E., Strong, R., Suh, E., Thomas, R., Tint, N. N., Tse, S., Vech, C., Wang, G., Wetter, J., Williams, S., Williams, M., Windsor, S., Winn-Deen, E., Wolfe, K., Zaveri, J., Zaveri, K., Abril, J. F., Guigó, R., Campbell, M. J., Sjolander, K. V., Karlak, B., Kejariwal, A., Mi, H., Lazareva, B., Hatton, T., Narechania, A., Diemer, K., Muruganujan, A., Guo, N., Sato, S., Bafna, V., Istrail, S., Lippert, R., Schwartz, R., Walenz, B., Yooseph, S., Allen, D., Basu, A., Baxendale, J., Blick, L., Caminha, M., Carnes-Stine, J., Caulk, P., Chiang, Y. H., Coyne, M., Dahlke, C., Deslattes Mays, A., Dombroski, M., Donnelly, M., Ely, D., Esparham, S., Fosler, C., Gire, H., Glanowski, S., Glasser, K., Glodek, A., Gorokhov, M., Graham, K., Gropman, B., Harris, M., Heil, J., Henderson, S., Hoover, J., Jennings, D., Jordan, C., Jordan, J., Kasha, J., Kagan, L., Kraft, C., Levitsky, A., Lewis, M., Liu, X., Lopez, J., Ma, D., Majoros, W., McDaniel, J., Murphy, S., Newman, M., Nguyen, T., Nguyen, N., Nodell, M., Pan, S., Peck, J., Peterson, M., Rowe, W., Sanders, R., Scott, J., Simpson, M., Smith, T., Sprague, A., Stockwell, T., Turner, R., Venter, E., Wang, M., Wen, M., Wu, D., Wu, M., Xia, A., Zandieh, A. & Zhu, X. The sequence of the human genome. *Science* **291,** 1304–1351 (2001).

3. Adhikari, S., Nice, E. C., Deutsch, E. W., Lane, L., Omenn, G. S., Pennington, S. R., Paik, Y.-K., Overall, C. M., Corrales, F. J., Cristea, I. M., Van Eyk, J. E., Uhlén, M., Lindskog, C., Chan, D. W., Bairoch, A., Waddington, J. C., Justice, J. L., LaBaer, J., Rodriguez, H., He, F., Kostrzewa, M., Ping, P., Gundry, R. L., Stewart, P., Srivastava, S., Srivastava, S., Nogueira, F. C. S., Domont, G. B., Vandenbrouck, Y., Lam, M. P. Y., Wennersten, S., Vizcaino, J. A., Wilkins, M., Schwenk, J. M., Lundberg, E., Bandeira, N., Marko-Varga, G., Weintraub, S. T., Pineau, C., Kusebauch, U., Moritz, R. L., Ahn, S. B., Palmblad, M.,

Snyder, M. P., Aebersold, R. & Baker, M. S. A high-stringency blueprint of the human proteome. *Nat. Commun.* **11,** 5301 (2020).

4. Legrain, P., Aebersold, R., Archakov, A., Bairoch, A., Bala, K., Beretta, L., Bergeron, J., Borchers, C. H., Corthals, G. L., Costello, C. E., Deutsch, E. W., Domon, B., Hancock, W., He, F., Hochstrasser, D., Marko-Varga, G., Salekdeh, G. H., Sechi, S., Snyder, M., Srivastava, S., Uhlén, M., Wu, C. H., Yamamoto, T., Paik, Y.-K. & Omenn, G. S. The human proteome project: current state and future direction. *Mol. Cell. Proteomics* **10,** M111.009993 (2011).
5. ENCODE Project Consortium, Moore, J. E., Purcaro, M. J., Pratt, H. E., Epstein, C. B., Shoresh, N., Adrian, J., Kawli, T., Davis, C. A., Dobin, A., Kaul, R., Halow, J., Van Nostrand, E. L., Freese, P., Gorkin, D. U., Shen, Y., He, Y., Mackiewicz, M., Pauli-Behn, F., Williams, B. A., Mortazavi, A., Keller, C. A., Zhang, X.-O., Elhajjajy, S. I., Huey, J., Dickel, D. E., Snetkova, V., Wei, X., Wang, X., Rivera-Mulia, J. C., Rozowsky, J., Zhang, J., Chhetri, S. B., Zhang, J., Victorsen, A., White, K. P., Visel, A., Yeo, G. W., Burge, C. B., Lécuyer, E., Gilbert, D. M., Dekker, J., Rinn, J., Mendenhall, E. M., Ecker, J. R., Kellis, M., Klein, R. J., Noble, W. S., Kundaje, A., Guigó, R., Farnham, P. J., Cherry, J. M., Myers, R. M., Ren, B., Graveley, B. R., Gerstein, M. B., Pennacchio, L. A., Snyder, M. P., Bernstein, B. E., Wold, B., Hardison, R. C., Gingeras, T. R., Stamatoyannopoulos, J. A. & Weng, Z. Expanded encyclopaedias of DNA elements in the human and mouse genomes. *Nature* **583,** 699–710 (2020).
6. Roadmap Epigenomics Consortium, Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., Heravi-Moussavi, A., Kheradpour, P., Zhang, Z., Wang, J., Ziller, M. J., Amin, V.,

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

7. Chen, Y., Yao, L., Tang, Y., Jhong, J.-H., Wan, J., Chang, J., Cui, S., Luo, Y., Cai, X., Li, W., Chen, Q., Huang, H.-Y., Wang, Z., Chen, W., Chang, T.-H., Wei, F., Lee, T.-Y. & Huang, H.-D. CircNet 2.0: an updated database for exploring circular RNA regulatory networks in cancers. *Nucleic Acids Res.* **50,** D93–D101 (2022).

8. Weirauch, M. T., Yang, A., Albu, M., Cote, A. G., Montenegro-Montero, A., Drewe, P., Najafabadi, H. S., Lambert, S. A., Mann, I., Cook, K., Zheng, H., Goity, A., van Bakel, H., Lozano, J.-C., Galli, M., Lewsey, M. G., Huang, E., Mukherjee, T., Chen, X., Reece-Hoyes, J. S., Govindarajan, S., Shaulsky, G., Walhout, A. J. M., Bouget, F.-Y., Ratsch, G.,

Larrondo, L. F., Ecker, J. R. & Hughes, T. R. Determination and inference of eukaryotic transcription factor sequence specificity. *Cell* **158,** 1431–1443 (2014).

9. Castro-Mondragon, J. A., Riudavets-Puig, R., Rauluseviciute, I., Lemma, R. B., Turchi, L., Blanc-Mathieu, R., Lucas, J., Boddie, P., Khan, A., Manosalva Pérez, N., Fornes, O., Leung, T. Y., Aguirre, A., Hammal, F., Schmelter, D., Baranasic, D., Ballester, B., Sandelin, A., Lenhard, B., Vandepoele, K., Wasserman, W. W., Parcy, F. & Mathelier, A. JASPAR 2022: the 9th release of the open-access database of transcription factor binding profiles. *Nucleic Acids Res.* **50,** D165–D173 (2022).
10. Szklarczyk, D., Gable, A. L., Nastou, K. C., Lyon, D., Kirsch, R., Pyysalo, S., Doncheva, N. T., Legeay, M., Fang, T., Bork, P., Jensen, L. J. & von Mering, C. The STRING database in 2021: customizable protein–protein networks, and functional characterization of user-uploaded gene/measurement sets. *Nucleic Acids Res.* **49,** D605–D612 (2020).
11. Lehne, B. & Schlitt, T. Protein-protein interaction databases: keeping up with growing interactomes. *Hum. Genomics* **3,** 291–297 (2009).
12. Missiuro, P. V., Liu, K., Zou, L., Ross, B. C., Zhao, G., Liu, J. S. & Ge, H. Information flow analysis of interactome networks. *PLoS Comput. Biol.* **5,** e1000350 (2009).
13. Zhang, K., Wang, M., Zhao, Y. & Wang, W. Taiji: System-level identification of key transcription factors reveals transcriptional waves in mouse embryonic development. *Sci Adv* **5,** eaav3262 (2019).
14. Alipanahi, B., Delong, A., Weirauch, M. T. & Frey, B. J. Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nat. Biotechnol.* **33,** 831–838 (2015).

15. Zhou, J. & Troyanskaya, O. G. Predicting effects of noncoding variants with deep learning–based sequence model. *Nat. Methods* **12,** 931–934 (2015).

16. Agarwal, V. & Shendure, J. Predicting mRNA Abundance Directly from Genomic Sequence Using Deep Convolutional Neural Networks. *Cell Rep.* **31,** 107663 (2020).

17. Yu, B., Zhang, K., Milner, J. J., Toma, C., Chen, R., Scott-Browne, J. P., Pereira, R. M., Crotty, S., Chang, J. T., Pipkin, M. E., Wang, W. & Goldrath, A. W. Epigenetic landscapes reveal transcription factors that regulate CD8+ T cell differentiation. *Nat. Immunol.* **18,** 573–582 (2017).

18. Raman, K. Construction and analysis of protein–protein interaction networks. *Automated Experimentation* **2,** 2 (2010).

19. Liu, C. & Wang, W. Contextual Regression: An Accurate and Conveniently Interpretable Nonlinear Model for Mining Discovery from Scientific Data. doi:10.1101/210997

20. Yan, X. & Su, X. *Linear Regression Analysis: Theory and Computing*. (World Scientific, 2009).

21. Ribeiro, M. T., Singh, S. & Guestrin, C. 'why should I trust you?': Explaining the predictions of any classifier. *arXiv [cs.LG]* (2016). doi:10.1145/2939672.2939778

22. Shrikumar, A., Greenside, P., Shcherbina, A. & Kundaje, A. Not Just a Black Box: Learning Important Features Through Propagating Activation Differences. *arXiv [cs.LG]* (2016). at <http://arxiv.org/abs/1605.01713>

23. Lundberg, S. & Lee, S.-I. A unified approach to interpreting model predictions. *arXiv [cs.AI]* (2017). at <http://arxiv.org/abs/1705.07874>

24. Panda, A. C. Circular RNAs Act as miRNA Sponges. *Adv. Exp. Med. Biol.* **1087,** 67–79

(2018).

25. Zhang, Y., Zhang, X.-O., Chen, T., Xiang, J.-F., Yin, Q.-F., Xing, Y.-H., Zhu, S., Yang, L. & Chen, L.-L. Circular intronic long noncoding RNAs. *Mol. Cell* **51,** 792–806 (2013).
26. Rybak-Wolf, A., Stottmeister, C., Glažar, P., Jens, M., Pino, N., Giusti, S., Hanan, M., Behm, M., Bartok, O., Ashwal-Fluss, R., Herzog, M., Schreyer, L., Papavasileiou, P., Ivanov, A., Öhman, M., Refojo, D., Kadener, S. & Rajewsky, N. Circular RNAs in the Mammalian Brain Are Highly Abundant, Conserved, and Dynamically Expressed. *Mol. Cell* **58,** 870–885 (2015).
27. Szabo, L., Morey, R., Palpant, N. J., Wang, P. L., Afari, N., Jiang, C., Parast, M. M., Murry, C. E., Laurent, L. C. & Salzman, J. Erratum to: Statistically based splicing detection reveals neural enrichment and tissue-specific induction of circular RNA during human fetal development. *Genome Biol.* **17,** 263 (2016).
28. Ashwal-Fluss, R., Meyer, M., Pamudurti, N. R., Ivanov, A., Bartok, O., Hanan, M., Evantal, N., Memczak, S., Rajewsky, N. & Kadener, S. circRNA biogenesis competes with pre-mRNA splicing. *Mol. Cell* **56,** 55–66 (2014).
29. Zhang, X.-O., Wang, H.-B., Zhang, Y., Lu, X., Chen, L.-L. & Yang, L. Complementary sequence-mediated exon circularization. *Cell* **159,** 134–147 (2014).
30. Jeck, W. R., Sorrentino, J. A., Wang, K., Slevin, M. K., Burd, C. E., Liu, J., Marzluff, W. F. & Sharpless, N. E. Circular RNAs are abundant, conserved, and associated with ALU repeats. *RNA* **19,** 141–157 (2013).
31. Liang, D. & Wilusz, J. E. Short intronic repeat sequences facilitate circular RNA production. *Genes Dev.* **28,** 2233–2247 (2014).

32. Ivanov, A., Memczak, S., Wyler, E., Torti, F., Porath, H. T., Orejuela, M. R., Piechotta, M., Levanon, E. Y., Landthaler, M., Dieterich, C. & Rajewsky, N. Analysis of Intron Sequences Reveals Hallmarks of Circular RNA Biogenesis in Animals. *Cell Reports* **10,** 170–177 (2015).

33. Conn, S. J., Pillman, K. A., Toubia, J., Conn, V. M., Salmanidis, M., Phillips, C. A., Roslan, S., Schreiber, A. W., Gregory, P. A. & Goodall, G. J. The RNA binding protein quaking regulates formation of circRNAs. *Cell* **160,** 1125–1134 (2015).

34. Li, B., Zhang, X.-Q., Liu, S.-R., Liu, S., Sun, W.-J., Lin, Q., Luo, Y.-X., Zhou, K.-R., Zhang, C.-M., Tan, Y.-Y., Yang, J.-H. & Qu, L.-H. Discovering the Interactions between Circular RNAs and RNA-binding Proteins from CLIP-seq Data using circScan. *bioRxiv* 115980 (2017). doi:10.1101/115980

35. Liu, Y.-C., Li, J.-R., Sun, C.-H., Andrews, E., Chao, R.-F., Lin, F.-M., Weng, S.-L., Hsu, S.-D., Huang, C.-C., Cheng, C., Liu, C.-C. & Huang, H.-D. CircNet: a database of circular RNAs derived from transcriptome sequencing data. *Nucleic Acids Res.* **44,** D209–15 (2016).

36. Lecun, Y., Bottou, L., Bengio, Y. & Haffner, P. Gradient-based learning applied to document recognition. *Proc. IEEE* **86,** 2278–2324 (1998).

37. Zhang, Y., Tino, P., Leonardis, A. & Tang, K. A Survey on Neural Network Interpretability. *IEEE Transactions on Emerging Topics in Computational Intelligence* **5,** 726–742 (2021).

38. Takamoto, S., Izumi, S. & Li, J. TeaNet: Universal neural network interatomic potential inspired by iterative electronic relaxations. *Comput. Mater. Sci.* **207,** 111280 (2022).

39. Xie, T. & Grossman, J. C. Crystal Graph Convolutional Neural Networks for an Accurate and Interpretable Prediction of Material Properties. *Phys. Rev. Lett.* **120,** 145301 (2018).

CHAPTER 1: A LOWER COMPLEXITY INFORMATION FLOW ANALYSIS ALGORITHM FOR NETWORK CENTRALITY ANALYSIS

1.1 Abstract

Information flow analysis is an effective way to identify and characterize important nodes in a network. This method can be applied not only to analyze interactions between genes and proteins, but also interaction networks in general. However, the method used to solve the linear system in the algorithm scales very fast ($O(n^4)$) as the size of the system increases. In this paper, we present an optimized algorithm that can solve this kind of system under $O(n^2m)$ time complexity while having a reasonable time constant and using twice the amount of memory as the input data.

1.2 Introduction

Information flow analysis, developed by Dr. Missiuro and her colleagues[1] offers a better way to extract information of gene interactions than betweenness and shortest path methods. The most computationally intense problem involved in this analysis can be described as below:

"Given an n x n symmetric matrix A, for i = 1 to n, zero the row i and column i in A, except A[i,i] = 1 (isolate the node from the circuit). Then, for j = i+1 to n, solve for X on the linear system: AX=Y, in which Y has 1 for element j and 0 for all other elements."

In the paper, Dr. Missiuro and her colleagues offer a solution as below (which I paraphrased to make later explanations easier):

1. assemble matrix A based on the conductance matrix
2. for i =1...N:
    a. remove the row and column i of A to get A^{**}
    b. decompose A^{**} into L and U matrices ($O(n^3)$)
    c. for j=(i+1)...N:
        i. assemble vector Y to have all 0s except 1 in the j^{th} entry
        ii. solve X: $X=U^{-1}(L^{-1}Y)$
        iii. compute the absolute sum of currents for each node

LU decomposition is an $O(n^3)$ operation. Running it for n times will result in an $O(n^4)$ runtime that scales up very fast. Even though many of the genetic interactome network matrices can be sparse, LU decomposition has a "fill-in" effect which causes the resulting L and U matrices to be dense. Previous research has developed strategies such as using pivoting matrices or hypergraph[2] to reduce this effect, which is not the focus here. Besides, making the algorithm asymptotically faster is meaningful despite the density of the matrices.

Here, we have developed an algorithm that can solve the problem in $O(n^2m)$, where m is the number of branches in the network. They are also more memory efficient than the original algorithm (In dense circumstances, A, A^{**}, L and U can take as much as 3 times the memory of the input) and have a reasonable time constant.

1.3 Results

Basic Implementation: We first states two formula and two facts that we will use a lot in constructing this algorithm:

1. If v is an 1xN vertical unit vector with 1 as its i^{th} element, M is a NxN matrix, then $M \cdot v$ is equivalent to extracting the i^{th} column of M and $v^T \cdot M$ is equivalent to extracting the i^{th} row of M.
2. Sherman-Morrison formula[3]: If a matrix W = uv^T, where u or v is a unit vector and both of them are vertical vectors. Then $(A+W)^{-1} = A^{-1} - \frac{A^{-1}uv^TA^{-1}}{1+v^TA^{-1}u}$.
3. Since A is a symmetric matrix and A^{**} is equal to A with i^{th} row and column removed, A^{-1}, A^{**} and thus A^{**-1} are symmetric as well.
4. Row operations are generally preferred over column operations due to cache locality (for instance, extracting rows is more efficient than extracting columns).

Given that A is a symmetric matrix, we use LDL^T decomposition instead of LU which is faster and more numerically stable. Then, we use backward elimination on L and L^T to compute A^{-1}.

For voltage X on each node, we solve $A^{**}X=Y$, which is essentially $X=A^{**-1}Y$. Since Y is a unit vector with 1 in position j and A^{**} is symmetric, $X=A^{**-1}Y=j^{th}$ column of A^{**-1} = j^{th} row of A^{**-1} ($A^{**-1}[j]$).

Thus, we only need to calculate A^{**-1} from A^{-1} to get X. Suppose matrix ϵ is all 0, except the i^{th} row is the minus of i^{th} row of A and $\epsilon[i,i] = \frac{1}{2}(1 - A[i,i])$. This can be done by applying the Sherman-Morrison formula to A^{-1} twice (Figure. 1.1), once with ϵ^T and once with ϵ.

Here, we will perturb A^{-1} with ϵ^T first. Suppose A^{*-1} is the matrix after the first perturbation, $\epsilon^T = uv^T$, v is the unit vector, then we have:

$$A^{*-1} = (A+\epsilon^T)^{-1} = A^{-1} - \frac{A^{-1}uv^TA^{-1}}{1+v^TA^{-1}u} \quad (1)$$

$$A^{**-1} = (A^*+\epsilon)^{-1} = A^{*-1} - \frac{A^{*-1}vu^TA^{*-1}}{1+u^TA^{*-1}v} \quad (2)$$

Let $u^TA^{-1}v = \lambda$ and $u^TA^{-1}u = \mu$, since A^{-1} is symmetric and v is a unit vector, $u^TA^{-1}v = v^TA^{-1}u$. Put formula (1) into formula (2) we have:

$$u^TA^{*-1}v$$

$$= u^TA^{-1}v - \frac{u^TA^{-1}uv^TA^{-1}v}{1+v^TA^{-1}u}$$

$$= \lambda - \frac{\mu A^{-1}[i,i]}{1+\lambda} \quad (3)$$

$$A^{*-1}vu^TA^{*-1}$$

$$= (A^{-1} - \frac{A^{-1}uv^TA^{-1}}{1+v^TA^{-1}u})\, vu^T\, (A^{-1} - \frac{A^{-1}uv^TA^{-1}}{1+v^TA^{-1}u})$$

$$= A^{-1}vu^{T}A^{-1} + \frac{A^{-1}uv^{T}A^{-1}vu^{T}A^{-1}uv^{T}A^{-1}}{(1+\lambda)^{2}} -$$

$$\frac{A^{-1}uv^{T}A^{-1}vu^{T}A^{-1}}{1+\lambda} - \frac{A^{-1}vu^{T}A^{-1}uv^{T}A^{-1}}{1+\lambda}$$

$$= A^{-1}v(A^{-1}u)^{T} + \frac{\mu A^{-1}[i,i]}{(1+\lambda)^{2}}A^{-1}u(A^{-1}v)^{T}$$

$$- \frac{A^{-1}[i,i]}{1+\lambda}A^{-1}u(A^{-1}u)^{T} - \frac{\mu}{1+\lambda}A^{-1}v(A^{-1}v)^{T} \quad (4)$$

Put (1), (3) and (4) into (2):

$$A^{**-1} = A^{-1} + aA^{-1}v(A^{-1}u)^{T} + bA^{-1}u(A^{-1}v)^{T}$$

$$+ cA^{-1}u(A^{-1}u)^{T} + dA^{-1}v(A^{-1}v)^{T} \quad (5)$$

The constants are replaced by letters a, b, c and d for simpler expression. This is an ugly expression, but deceptively simple to be calculated by a computer. We will mention in the later section that, if we remove more rows and columns from the original matrix, we can easily construct a polynomial to represent its inverse matrix by putting formula (5) into itself recursively. Thus far, we have come up with our basic algorithm implementation:

1. assemble matrix A based on the conductance matrix $O(n^2)$
2. decompose A into LDL^T $O(n^3)$
3. calculate A^{-1} using backward elimination on L and L^T (D is diagonal, so it's simple) $O(n^3)$

4. for i =1...N: $O(n^2m)$

   a. ϵ^T = -A[i] except $\epsilon^T[i]=\frac{1}{2}(1 - A[i,i])$($\epsilon^T$and ϵ are matrices, we represent it as array for simplicity)

   b. u = $\epsilon^T$and v = [0,0,...1,...,0] all zero except 1 in the i^{th} position

   c. $A^{-1}v = A^{-1}[i]$

   d. calculate $A^{-1}u$ $O(n^2)$

   e. calculate $\lambda= u^T \cdot A^{-1}v$ O(n)

   f. calculate $\mu= u^T \cdot A^{-1}u$ O(n)

   g. for j = i+1...N: O(nm)

      i. $X = A^{**-1}[j] = A^{-1}[j] + aA^{-1}v[j] \cdot A^{-1}u + bA^{-1}u[j] \cdot A^{-1}v + cA^{-1}u[j] \cdot A^{-1}u + dA^{-1}v[j] \cdot A^{-1}v$ (only calculate the row we need on the fly, both save memory and better for cache locality) O(n)

      ii. compute the absolute sum of currents for each node O(m)

Computing the absolute sum of current for each node requires O(m), in which, m is the number of branches in the circuit. Since in most cases, m > n, this leads to a $O(n^2m)$ total time complexity. The arrays only use memory of level O(n) which is negligible compared to the $O(n^2)$ usage of the input. Also, the matrix L can be discarded after A^{-1} is acquired. So the maximum memory usage of the algorithm is about twice as the input.

However, since the elements in the A diagonal A[i,i] = - sum(A[i]), A is a singular matrix and thus does not have an inverse matrix, we have to get A^{**-1} in a roundabout way: first we calculate $A^{\#-1}$. $A^{\#}$ is the same as A except $A^{\#}[1,1] = 1$, which is still symmetric. Then, we perturb

$A^{\#-1}$ first with matrix $\epsilon_1^T = u_1 v_1^T$, $u_1 = [p,0,0,...,1,...,0]$, in which $p = \frac{1}{2}(A[1,1] - 1)$ and the second "1" lies at position i, $v_1 = [1,0,...,0]$ and then with ϵ_1. After this step, we still have a symmetric matrix and can use row operation instead of column operation. Then we can do the two step perturbation as we did before (see Figure. 1.2).

Let's call the matrix after first round of perturbation $A^{@@-1}$, we have:

$$
\begin{aligned}
&A^{@@-1} \\
&= A^{\#-1} + a' A^{\#-1} v_1 (A^{\#-1} u_1)^T + b' A^{\#-1} u_1 (A^{\#-1} v_1)^T \\
&+ c' A^{\#-1} u_1 (A^{\#-1} u_1)^T + d' A^{\#-1} v_1 (A^{\#-1} v_1)^T \\
&= A^{\#-1} + [(a'p + c'p^2 + b'p + d') A^{\#-1}[1] + (c'p + b') A^{\#-1}[i]] A^{\#-1}[1]^T \\
&+ [(a' + c'p) A^{\#-1}[1] + c' A^{\#-1}[i]] A^{\#-1}[i]^T \quad (6)
\end{aligned}
$$

In this expression:

$$\lambda_1 = pA^{\#-1}[1,1] + A^{\#-1}[1,i] \quad (7)$$

$$\mu_1 = p^2 A^{\#-1}[1,1] + 2pA^{\#-1}[1,i] + A^{\#-1}[i,i] \quad (8)$$

$$A^{\#-1} v_1 = A^{\#-1}[1] \quad (9)$$

$$A^{\#-1} u_1 = pA^{\#-1}[1] + A^{\#-1}[i] \quad (10)$$

Using formula (6), we can get the second round of perturbation, for example:

$$\lambda_2$$

$$= u_2^T A^{@@-1} v_2$$

$$= u_2^T A^{\#-1}[i] + u_2^T[(a'p+c'p^2+b'p+d')A^{\#-1}[1] + (c'p+b')A^{\#-1}[i]]A^{\#-1}[1,i]$$

$$+ u_2^T[(a'+c'p)A^{\#-1}[1] + c'A^{\#-1}[i]]A^{\#-1}[i,i] \quad (11)$$

The other formula can be acquired in the same way. Since the extra calculations related to (6) are in the O(n) level (inside the i loop). All these operations do not increase the constant of the highest order term in time complexity.

Algorithm Speed Testing: A speed performance test was applied to the Hi-C network data[4] on the smallest chromosome, chromosome 22 at 10Kbp resolution for GM12878_10Kbp (10Kbp Hi-C resolution in GM12878, number of nodes = 3,447 and number of edges = 92,026). The CPU implementation of our algorithm shows a speed up of 2.22 fold (see Figure. 1.3). The GPU implementation shows a 17.40 fold speed up compared to the original algorithm and makes the application of the algorithm to large chromosomes such as chromosome 2 for a 5Kbp resolution (GM12878_5Kbp, number of nodes = 47,103 and number of edges = 1,414,675) affordable. Indeed, the computation on chromosome 2 for GM12878_5Kbp that contains the most Hi-C interactions among all the chromosomes takes ~20 days to finish with our GPU implementation which would be impossible to be calculated using the original algorithm. Our new algorithm and the GPU implementation allow us to apply the information flow analysis to all the chromosomes in all of our cell-lines of analysis.

1.4 Discussion

We derived an algorithm that can solve the information flow analysis problem under $O(n^2m)$ time. In this algorithm, computing the currents between nodes dominates the time complexity. m can be in the order $O(n^2)$ for dense matrices. If we can find another property that describes the information flow without calculating absolute current flow between nodes, this analysis can be even more efficient.

1.5 Acknowledgments

We thank Dr. J. Andrew McCammon and his lab for providing us with GPUs for testing GPU implementation of our algorithm

Chapter 1, in full, is part of the unpublished material as Ainsworth, R., Liu, C. & Wang, W. Examination of Hi-C Network Properties under Local and Global Centrality Measures. The dissertation author was the primary investigator and author of this paper.

1.6 Author Contributions

This study was conceived and designed by A.R., L.C. and W.W.; algorithm developed by L.C.; data processed by A.R.; data analyzed by A.R. and L.C.; Manuscript written by A.R., L.C. and W.W. with input from all authors.

1.7 Figures

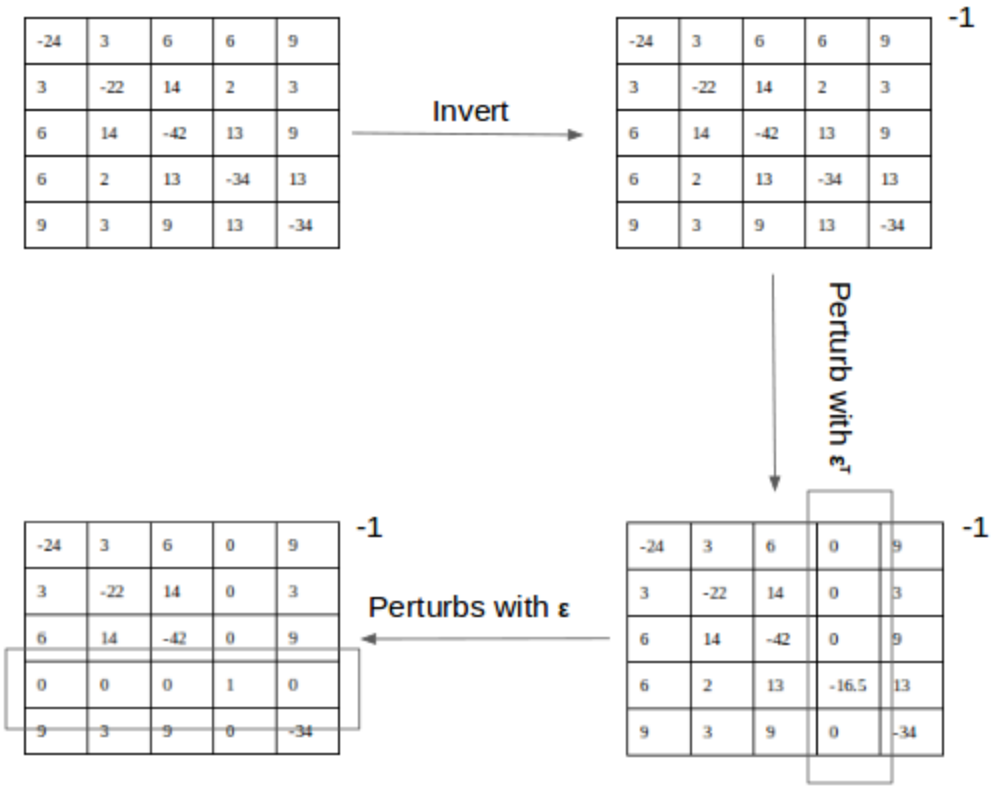


Figure. 1.1. obtaining the desired inverted matrix by perturbing the original inverted matrix twice using the Sherman-Morrison formula

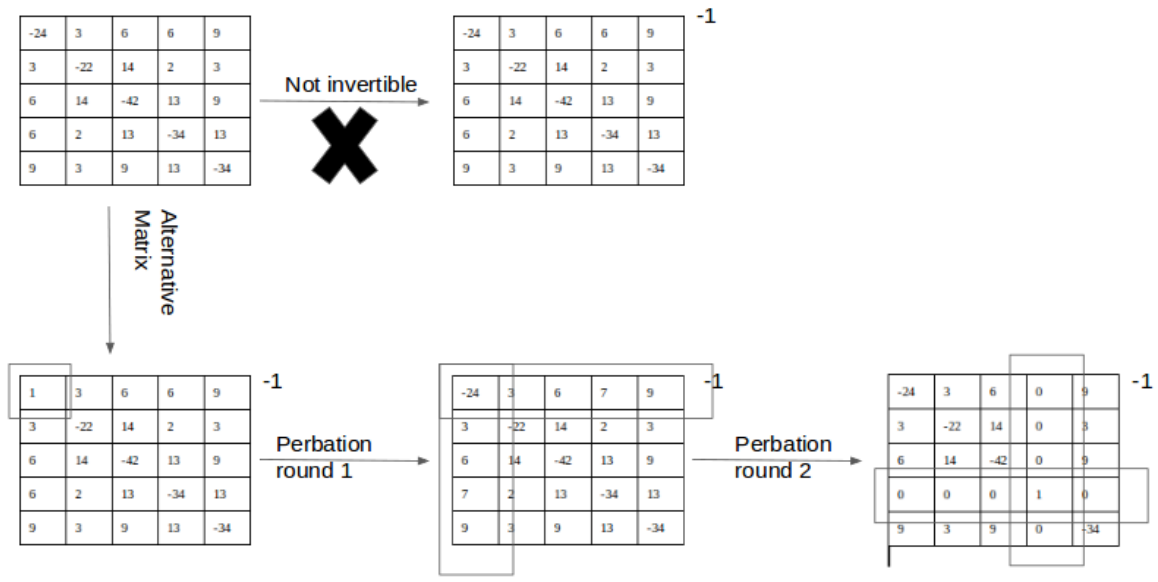


Figure. 1.2. since A is not invertible, we use a roundabout approach to acquire the matrix we need

Algorithm	Runtime (s)	Speed Compared to Original
Original Algorithm (CPU, 4 cores)	20976.2	1.00x
New Algorithm (CPU, 4 cores)	9433.4	2.22x
New Algorithm (GPU, GTX580x1)	1205.8	17.40x

Figure. 1.3. speed comparison of three information flow implementations on GM12878_10Kbp Chromosome 22 Hi-C network data

1.8 Supplementary Methods

Speed Benchmarking. The matlab version of the information flow algorithms are tested on intel i-5 quad core. The CUDA version of the algorithm is tested on a GTX-580 GPU. Each algorithm is run 5 times and the average run time is recorded.

Hi-C Testing Data. The Hi-C network data for testing are obtained from [4]. Only edges with $\log_e$ (p-value) < -15 are kept.

1.9 References

1. Missiuro, P. V., Liu, K., Zou, L., Ross, B. C., Zhao, G., Liu, J. S. & Ge, H. Information flow analysis of interactome networks. *PLoS Comput. Biol.* **5,** e1000350 (2009).
2. Oguz Kaya, Enver Kayaaslan, Bora Uçar, Iain S. Duff. Fill-in reduction in sparse matrix factorizations using hypergraphs. INRIA. 2014
3. Jack Sherman, Winifred J. Morrison. Adjustment of an Inverse Matrix Corresponding to a Change in One Element of a Given Matrix, Ann. Math. Statist. 21(1), 124-127, (1950)
4. Rao, S. S. P., Huntley, M. H., Durand, N. C., Stamenova, E. K., Bochkov, I. D., Robinson, J. T., Sanborn, A. L., Machol, I., Omer, A. D., Lander, E. S. & Aiden, E. L. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. Cell 159, 1665–1680 (2014).

CHAPTER 2: CONTEXTUAL REGRESSION: AN ACCURATE AND CONVENIENTLY INTERPRETABLE NONLINEAR MODEL FOR MINING DISCOVERY FROM SCIENTIFIC DATA

2.1 Abstract

Machine learning algorithms such as linear regression, SVM and neural network have played an increasingly important role in the process of scientific discovery. However, none of them is both interpretable and accurate on nonlinear datasets. Here we present contextual regression, a method that combines these two desirable properties together using a hybrid architecture of neural network embedding and dot product layer. We have demonstrated its high prediction accuracy and sensitivity through the task of predictive feature selection on a simulated dataset and the application of predicting open chromatin sites in the human genome. On the simulated data, our method achieved high fidelity recovery of feature contributions under random noise levels up to ±200%. On the open chromatin dataset, the application of our method not only outperformed the state-of-the-art method in terms of accuracy, but also unveiled two previously unfound open chromatin related histone marks. Our method fills in the gap of accurate and interpretable nonlinear modelling in scientific data mining tasks.

2.2 Introduction

Predictive models are important tools for data analysis. Linear models, such as linear[1] and logistic regression[2], can provide adequate prediction accuracy while their linear formulation makes them suitable for inferring relationship between prediction target and features. Recently, complex nonlinear models have gained increasing popularity as an alternative to linear models. Methods such as kernel SVM[3], neural network[4] and decision trees[5] can achieve better prediction performance than the linear models especially on nonlinear datasets. In biology, for instance, deep convolutional neural network (DCNN) methods such as DeepBind[6] and DeepSea[7] have been developed for scanning the genome for regions of interest with state-of-the-art accuracy.

This accuracy improvement, however, also comes with a cost: the parameters of a nonlinear model are hardly human interpretable. This not only makes model improvement and debugging difficult but also halts extracting knowledge from these models and validating their findings, which is particularly important when controversial conclusions can be reached from the prediction results[8,9]. While examining the filters in a DCNN[10,11] can offer some insight of the important feature combinations, it cannot provide quantification of the feature contributions: in the later layers, the features are processed through multiple rounds of matrix multiplication, addition and neuron activation which makes their contributions to the output intractable. Besides, this approach cannot be applied to other neural networks or machine learning methods such as feedforward neural network (FNN)[12] or Long Short Term Memory (LSTM)[13] neural network. To fill this gap, methods such as DeepLift[14], LIME[15] and SHAP[16] have been developed for quantifying feature contributions in general machine learning models. All of them are based on a similar overall strategy. When using these methods, the user first needs to choose a reference

instance from the dataset with the target value of interest, e.g. a genomic region with enhancer label of value 1 when predicting enhancers. Then the program will generate permuted data points by making small changes to the input features and from the magnitude of the output change caused by the input change, an interpreting model that describes the feature contribution to the prediction result can be generated for that data point. This approach, however, has several practical problems. All of them are based on finding the contribution of features near selected data samples that the users are interested in. This approach, however, has several practical problems. First, since interpretation is based on each user's selected reference data point, it is inherently a "local" model and thus can be overfitted to that point. Second, the training and interpretation processes are decoupled, which does not allow users to apply regularization on the interpretation to enforce restraints such as sparsity. Third, the user needs to be familiar with the dataset in order to pick representative reference points, which is not an easy task particularly in scientific data when (i) the prediction target value is a continuous value other than binary; (ii) the dataset contains natural noise that causes the target value to deviate from its real value; (iii) the data are of large quantity and diversity.

We have developed a new method called contextual regression to address these challenges. It can quantify feature contributions during training without user intervention while preserving the accuracy achieved by a complex nonlinear model. The idea of this method is to train a pseudo-linear model, which, instead of generating a single linear model for all the data points, generates different linear models at different data points (see below for mathematical details and intuitive explanation). In each of these linear models, the importance of each input feature is represented by its weight in the model, called contextual weight. The contextual weight

is derived from a linear transformation of the outputs from the non-linear model that can achieve accurate prediction such as the bidirectional LSTM model in the first panel of Figure 2.1ST. The non-linear model and the linear transformation are collectively called embedding function through which each feature vector is mapped to a corresponding linear model that can predict the target value most accurately (Figure 1). In principle, values of the elements in the input feature vector describe its context and the embedding serves as a classifier of the context. It generates a continuous value vector as the class of the context which is similar to the mechanism of attention[17] and word2vec[18]. Therefore, we call this class of the context "context weight" and this method "contextual regression." In this way, the contribution of each feature can be inferred from the statistics of the context weights on the dataset. This model allows us to not only fit an interpretation that is generalized to the entire dataset but also classify data into subtypes based on feature contributions to the prediction decision, both of which cannot be done with previous methods and highly desirable for scientific data mining. We have demonstrated its accuracy and high interpretability through quantifying feature contributions to prediction accuracy in a simulated dataset with known "ground truth" and an application to the open chromatin prediction.

2.3 Results

The contextual regression method. As shown in Figure 1, our model aims to learn a pseudo-linear model that approximates the function:

$$y = f(x_1, x_2, x_3, ..., x_n) \approx g(c(x) \cdot x + b) = g(\sum_{i=1}^{n} c_i(x_1, x_2, x_3, ..., x_n) \cdot x_i + b) \quad (1)$$

where y is the prediction target, x_i are the features, b is a constant, g is the regression function (for example, it is an identity function for linear regression and is a logistic function for logistic regression), and c is the embedding function, which calculates the context weight from the feature vector x as the linear model for the prediction of y at that data point.

An intuitive understanding of the contextual regression can be described as follows: the neural network learns to predict the target value from the features through adjusting its parameters to approximate a function that gives the lowest overall difference between the real and predicted values. Compared with linear regression, neural network can work for any smooth and compact function according to the universal approximation theorem[12]. Our interpretation method utilizes this property and, instead of letting the neural network learn a function that maps features to target values, our method let the neural network learn a function (c(x) in equation 1) that maps features (x) to local linear models (output of the function c(x)) that best predict the target value (y) from the features. The contribution of each feature at different data sub-classes on the dataset can be found by clustering the linear models (Context Weights in figure 1 and c(x) in the above equation, if one wants to find subclasses based on feature weight) or the features weighted by the linear model (if one wants to find subclasses based on feature contribution). Thus, rather than trying to fit all the data with a single linear model, our method learns many

linear models and applies different ones at different data points according to their values. As illustrated in Figure 2.1, the embedding model (whose parameters are obtained through training to maximize the prediction of y) takes a data point as its input, and outputs a linear model (the context weight). Then the data point is dot-producted with the linear model to output the prediction. Since the linear models are generated by a non-linear embedding model, this method can handle nonlinear relationship to make accurate predictions while still producing a human-readable linear relationship between features and the prediction target.

The local linear interpreter approach is also adapted by other neural network interpretation algorithms such as LIME, SHAP and in a similar form by DeepLift. However, these algorithms can only generate interpretation for one instance at a time. This property causes them to be inefficient time-wise for application on large datasets, unable to apply regularization on interpretation and unable to separate data into subclasses. Contextual regression, on the other hand, can generate interpretation for all data instances based on the statistics of the whole data set on the fly and enable the division of data into subclasses for further analysis.

One potential caveat, a universal problem in predictive modelling, is the existence of alternative models on the dataset, i.e., very different model parameters can yield similar prediction accuracy. Since a descent direction for y is not necessarily a descent direction for each individual c_i (see Supplementary Materials for detail), this scenario can possibly happen. We have developed the utopia penalty technique to address it. This technique reveals equivalent or interchangeable features by adding a term to the cost function that penalizes unbalanced context weights. The setup and an application example of it can be found in the Supplementary Materials.

We evaluated the performance of the proposed method using a simulated dataset and a DNaseI hypersensitivity (DHS) dataset[19]. In both cases, the features have distal relationship and thus we used the bidirectional-Long Short Term Memory (LSTM) neural network, a specialized neural network model for distally related data, in our embedding function.

Evaluating the Contextual Regression model on simulated data. The simulated dataset includes an artificial "ground truth" in its feature-target relationship that we wish to find. We sampled features x_i from an exponential distribution to (1) reduce the possibility of alternative models such that we can test the rule extraction ability of our method, and (2) imitate the scenario that the feature values differ in order of magnitude such as the sequencing read count in biological data. In the simulated data, we consider 50 features x_i that are in sequential order. Each element c_i of the context weight vector C is determined by the value of x_i and its neighbours. The "ground truth" context weights c_i and then the target value y are calculated from x_i using the formula below. The context weight formula is chosen such that it is composed of common functions observed in natural data (linear, square and square root) and includes relationship with the neighboring elements. More details about the setup can be found in the Supplementary Materials.

$$y = \sum_{i=1}^{n} c_i(x_1, x_2, x_3, \ldots, x_n) \cdot x_i \quad (2)$$

$$c_i(x_1, x_2, x_3, \ldots, x_n) = \frac{1}{1000}\left[\sum_{j=i-1}^{1} 0.3^{i-j} u(x_j) + \sum_{k=i+1}^{n} 0.3^{k-i} u(x_k) + x_i\right] \quad (3)$$

$$u(x_k) = \sqrt{500x_k} + \frac{x_k^2}{500} \quad (4)$$

Our model was tested under 5 noise levels from ±0% to ±80%. When adding the noise, a random number inside the noise range was sampled uniformly (for instance, between -80% to +80% for noise level ±80%) and this fraction was added or subtracted from the target value y. The model was trained on 70% of the randomly selected data (training set) and then its fitness was tested on the remaining 30% data (testing set). We used root mean square error (RMSE) as the measure of prediction accuracy. Cosine distance measures the angle difference between two vectors, and thus is a good measure of vector similarity. Therefore, for the measure of rule extraction fidelity, we used root mean square cosine distance (RMSCD), the cosine distance version of RMSE, between the "ground truth" weights and the weights produced by our embedding network at their corresponding data points. Since the target values contain noise, it is impossible for the model to achieve 100% accuracy. We thus compared RMSE with the expected error under the corresponding noise level, which was calculated by integrating the percent error in the error range. The details of this calculation can be found in the Supplementary Materials.

Under all noise levels, our model achieved high performance in both prediction accuracy and rule extraction fidelity on the testing set (Figure 2.2 and 2.1S), with RMSE less than 1% away from the expected error and RMSCD below 0.03 (Table 2.1S). We also visually inspected (Figure 2.2) some randomly selected "ground truth" context weights (green) versus context weights (blue) produced by the embedding network under noise level ±0% and ±80% (Figure 2.2). They are indeed visually similar which agrees with the overall low RMSCD values between the "ground truth" and network output context weights.

In certain scientific experiments, the signal to noise ratio can be extremely low. This is especially the case in some experiments of physics[20,21], social sciences[22,23] and biological sciences[24–26]. Thus we also tested our model under noise level of 200% (signal to noise ratio 1:2), 500% (signal to noise ratio 1:5) and 1000% (signal to noise ratio 1:10). We found that under all three noise levels, the prediction errors still approached the expected error very fast (Figure 2.2S and Table 2.2S). However, the RMSCD is only stably decreasing under noise level of 200% and 500%. Under 1000% noise, the RMSCD value starts to slightly increase and fluctuate when the training time gets longer. This is confirmed in our visual inspection of the sample data points (Figure 2.3S), where the sampled extracted weights still resemble the corresponding "ground truth" weights under 200% and 500% noise but distorted under 1000% noise.

Overall, the high prediction accuracy (relative error < 0.9% in all noise levels) and high rule extraction fidelity (RMSCD < 0.05 in all noise levels smaller than 1000%) on the simulated dataset support that our algorithm is highly reliable even under high noise level. These results also suggest that our method can be applied to datasets with missing information, as long as the effect of missing information on the prediction target is similar to a random noise of uniform distribution.

Evaluating the Contextual Regression model on DNaseI hypersensitivity data. To examine the method in real experimental data, we applied it to predict DNaseI hypersensitivity sites (DHSs) using histone modifications in the H1 and GM12878 cell lines. We binned all the histone modification ChIP-seq and DHS-seq data at 200bp. To consider distal relationship, we included signals from 10kbp upstream to 10kbp downstream of the location to be predicted, i.e.

one hundred 200bp bins in total. In the H1 cells, there were 27 histone marks and thus the number of features reached 100 * 27 = 2700, which is too many for modelling. Hence, we arranged the features into a 100 by 27 matrix and factorized it into the tensor products of two vectors:

$$C = F \otimes D$$

where C is the 100 * 27 combination feature matrix, F is the histone feature vector of length 27, corresponding to the number of histone marks, and D is the distal feature vector of length 100, corresponding to the number of bins. This approach greatly reduced the number of parameters in the model. For fast evaluation of model performance, we further simplified and speeded up the model by forcing the vector D to be the same for all the regions and applying a Lasso[27] penalty on F.

We compared contextual regression with linear regression and Bidirectional-LSTM with the same set up in the contextual regression embedding function (See Supplementary Materials for details). The task for comparison is the prediction of DHS confidence values in cell line H1 using the 27 histone marks ChIP-seq data in the same cell line (Figure 2.3 and 2.4S). The models were trained on 70% of the randomly selected genomic regions and tested on the rest 30%. The contextual regression obviously outperformed linear regression, showing its better ability to capture nonlinear relationship between histone marks and open chromatin. At the same time, its performance is as good as the LSTM benchmark, showing that our interpretation method preserves the accuracy of a complex nonlinear model in this task.

Since applying log to the histone mark data improves the result slightly, we used this technique in all the studies in this section. We then assessed the performance of contextual

regression on predicting RNA-seq and DHS data, which we selected because of their biological relevance[28,29], in cell lines H1 and GM12878 by comparing to ChromImpute[30], the state-of-the-art imputation method based on random forest algorithm. We ran both methods in three rounds of cross validations: (1) training on chromosome (chr) 1-6, test on chr 7-X, (2) training on chr 7-13, testing on chr 1-6+14-X and (3) training on chr 14-X, testing on chr 1-13. The models were trained using only the data from the same cell line. Contextual regression outperformed ChromImpute on both DHS and RNA-seq predictions (Figure 2.3 and 2.5S). Note that all the predictions were made using only the data from a single cell line rather than borrowing information from many other cell lines that may be similar to the cell line of interest.

We also performed several other evaluations of contextual regression (See Supplementary Materials for detail) on: high contribution feature validity, context weight assignment consistency and whether the target can be predicted well using only the high contribution features. As expected, the high contribution histone marks for both RNA-seq and DHS concord with previous research. Our model also assigned consistent weights to the features and made accurate predictions using only the high contribution features. These results have demonstrated that our method is able to achieve better performance than the state-of-the-art models and also robustly select important features.

Discovering Important Histone Mark Patterns in the Open Chromatin Regions of H1. The above model considered the histone modification signals of neighbor loci of the DHSs but did not quantify their contribution. To identify the significant contributions from loci other than the DHS peak site (i.e. distal contribution), we further modified the model so that it can

search for and reveal both distal and local histone mark patterns in the open chromatin regions in the H1 human embryonic stem cells. We made three changes to the model and training process: 1. we removed the restraint on distal feature vector D so that our model produces different D from different input feature vectors, 2. we did not apply log to make the result closer to the representation of the original magnitude, and 3. we applied a softmax[31] constraint on D rather than the Lasso penalty on F to force the distal weights to be concentrated on a small number of locations. To make sure the patterns extracted are valid, we checked the prediction accuracy on the training and testing set and, indeed, the model performed equally well (Table 2.3S).

When analyzing the histone mark contribution patterns, we extracted the correctly predicted peaks (logpval > 3 for both experimental and predicted) from both training and testing set. We applied element-wise multiplication to the context weights (100 * 27) with the magnitude of the corresponding input (also 100 * 27), normalized them into vectors (of dimension 100 * 27 = 2700) of length 1, applied PCA (of dimension 2700), picked the top 100 PCs (which captured > 90% variance in all cell lines) and then clustered it with K-means clustering[32]. With k = 20, the clusters are well separated indicated by the great difference between in and across cluster distance from the cluster center (Figure 2.6S).

In the cluster centers, we observed 4 types of major patterns: 1. the central dominant histone marks (with small contribution from other marks), 2. the spread histone marks, 3. the central dominant H2A.Z, and 4. the H3K27me3 pattern. Type 1 pattern (Figure 2.7S) has the peak contribution exactly from the center with a radius of 300bp by H3K4me1/2/3 and H3K27ac. This type is composed of majority of the clusters (0, 1, 2, 3, 5, 6, 7, 8, 9, 12, 13, 17, 19). Type 2 patterns (Figure 2.8S) have long range contributions (> 300bp) besides the central

dominant peaks by H3K4me1/2/3 and H3K27ac, including clusters (11, 14, 15, 16). Type 3 pattern (Figure 2.9S) has central dominant (< 300 bp) contribution from H2A.Z, including cluster 10 and 18. Type 4 pattern (Figure 2.10S, 2.12S) has the most significant contribution from H3K27me3 and ancillary contribution from H2A.Z, H4K8ac, H3K18ac and the H3K4me1/2/3. The effect of H3K27me3 is long range in this pattern and almost covers the whole 20kbp region. This type compose of cluster 4, a total of 3959 (4.2%) DNH peaks (logpval > 3 for both experimental and predicted). Many of the important distal locations found by our method are not overlapped with the DHS peaks (for instance, the H3K18ac peaks in Figure 2.4 and 2.13S), which cannot be found with previous methods.

Most of the open chromatin contributors found by our method, H3K27ac[33,34], H2A.Z[35,36], H3K4me1/2/3[37], H4K20me1[38] and H4K8ac[39], have been repeatedly reported by previous research. However, H3K27me3 correlation with open chromatin is only mentioned in several papers[40,41]. To ensure this pattern is not caused by artifact, we visually inspected several regions of type 4 pattern. All these regions have strong H3K27me3 signal that covers the whole region (Figure 2.10S). This is consistent with the contribution pattern (Figure 2.12S) discovered by our algorithm. We also found that the regions are enriched with transcription start sites (TSSs): about 64.4% of the regions in this cluster contain TSSs within 3kbp and 78.2% within 5kbp (Table 2.4S). Although visual inspection shows overall high concordance between open chromatin peaks with H3K4me2/3, there are regions that the DHS and H3K4me2/3 do not overlap well and H3K27me3 shows strong signals in the broad neighbor region. The most significant contribution from H3K27me3 to DHS peaks is surprising but consistent with the previous work[42] which reports the bivalent domains that are stem cell specific: regions contain H3K4me3 and

H3K27me3 sites of size 1kbp-18kbp and overlap with genes and TSSs. When applying our method to the DHS data in other cell-lines (H9, IMR90, GM12878, HUVEC) to search for cell line specific patterns, we found that pattern type 1-3 appear in all 5 cell-lines, however, type 4, the H3K27me3 pattern, is only found in H1 and H9, both are embryonic stem cells. This observation again coincides with the previous finding that the correlation between H3K27me3 and open chromatin is highly cell specific[40] and mostly found in embryo cells[42,43].

DNA motifs associated with open chromatin regions predictive by different histone marks in H1 and H9. We next investigated whether there are specific DNA sequence motifs associated with each pattern. We compared the appearance frequency of motifs in each cluster and in all the clusters. The p-value in the comparison was calculated using One-Proportion Z-test. We limited our discussion only to cell-lines H1 and H9, which has the most diverse histone marks.

The cluster 4 and 16 of H1 and cluster 0, 4, 5, 8, 15, 16 and 19 of H9 contain motifs that have frequency difference above 10%. H9-8 and H9-19 are classic H3K27ac patterns. Both clusters have very similar motif enrichment in top 7 (Table 2.5S). Their enriched motifs agree except for NKX21 and PO3F1 (the consensus motif sequence 5'-ATTTGCAT-3'). Among the motifs enriched in both clusters, POU5F1 (Oct4, the consensus motif sequence 5'-ATTTGCAT-3, a different PWM from the PO3F1 motif') and SOX2 form a complex and regulate the expression of embryonic genes[44]. POU2F1/2 (most preferable motif sequence 5'-ATTTGCAT-3') activate the genes that codes for H2B proteins[45]. POU3F2 (binds to two half-sites, 5'-GCAT-3' and 5'-TAAT-3') is a protein involved in differentiation[46]. We also examined the H3K27ac dominant

clusters in H1 (cluster 1 and 6) and found that many of the above motifs are also significantly enriched (Figure 2.11S and Table 2.6S). Thus a consistent motif enrichment pattern is confirmed for H3K27ac dominant open chromatin regions.

H1-cluster 4 and H9-cluster 4 are both H3K27me3 patterns (Figure 2.12S). Their motif enrichment profile are also very similar. Specifically, in H9-cluster 4, the frequency enrichment that is highly significant are shown in Table 2.7S. In this cluster, MBD2 (binds to Methyl-CpG) is the most enriched motif (62.9% overall -> 89.7% in cluster 4) and it plays the role of a transcription repressor that can recruit deacetylase and methyltransferase[46–50]. Another class of motifs that are enriched are the E2 family factors (E2F1-4, most preferable motif sequence 5'-TTTC[CG]CGC-3') which are very important for the cell cycles[51,52].

More interestingly, in H9 we found two marks, H3K18ac and H3K23me3, strongly associated with open chromatin that has not been reported. H9-cluster 15 and 16 are H3K18ac dominant clusters (Figure 2.4). The major contribution of H3K18ac is around 200bp up or downstream. The correlation of this mark with open chromatin is previously undiscovered and our algorithm only detected it in H9 among the 5 cell lines we studied. They compose 9.7% of the total predictable peak regions. To confirm the finding, we manually inspected some of the regions and H3K18ac does show up strongly around peaks of its prediction, although in some cases, the other marks also co-appear at the neighboring locations (Figure 2.4 and 2.13S). The motifs in the regions of cluster 15 and 16 show a unique enrichment profile (Figure 2.4). Among the enriched motifs, USF1,2 (binds to E-box, with most preferable motif sequence 5'-CACGTG-3') are associated with gene activation[53]. MYC family (MYC, MYCN, MAX and MLXPL) can form dimers and binds to E-box that regulates cell proliferation and

differentiation[54] while PLAL1 also facilitates the same function[55]. ZIC3 (most preferable motif sequence 5'-GGGTGGTC-3') works on early body axis formation[56]. All these factors fit the stem cell nature of the H9 cell line.

H9-cluster 5 is H3K23me2 dominant (Figure 2.5), which is another previously undiscovered and H9 unique pattern. The major contribution of H3K23me2 is around 200bp downstream (the LSTM considers orientation). It compose 5.6% of the total predictable peak regions. However, the visual correlation of H3K23me2 with open chromatin is not as strong as the dominant marks in aforementioned clusters and co-appears with other marks in the open chromatin regions (Figure 2.5 and 2.14S). This lower correlation is also found statistically by our model, which shows less percentage of contribution from H3K23me2 (max value 43.8%) than the dominant marks in other clusters (for instance, max value of H3K18ac is 112%, > 100% since there are negative contributions from some other marks). The signature motif profile of H3K23me2 is very similar to the one of H3K27me3, with high MBD2 and E2 family enrichment. The largest motif frequency difference (Figure 2.5) is that the H3K23me2 regions has CEBPD motif frequency close to background level compared to the H3K27me3 cluster (81.2% vs 65.0%). CEBPB and CEBPG motif frequencies also show large differences between these two clusters. Function-wise, CEBPD is associated with immune response, immune cell activation and differentiation[57]. Other motifs that have significantly different profile among these two clusters is the E2 family. One possibility is that H3K23me2 is important for the stem cell development at a certain stage which is the reason why it is only enriched in certain open chromatin regions in H9.

Furthermore, H1-cluster 16 and H9-cluster 0 (Figure 2.15S) are the spread type which include contributions from many histone marks while none of them are dominant (Figure 2.16S). These two clusters show very similar motif enrichment patterns: strong contribution from MBD2 (+18%) and E2 (+5%-10%) families and high CEBPD frequency (more than 80% in both clusters). We hypothesize that this cluster is intermediate in its motif and histone mark patterns between the H3K27me3 and H3K23me2 clusters, which suggests that these three clusters are open chromatins in repressive states at different levels.

The above observations demonstrate that the open chromatin patterns captured by our method not only accord with visual inspection but are enriched in unique motifs as well. This shows the existence of diverse histone marks and motif patterns that can relate to open chromatin via different mechanisms. Two of the patterns, H3K18ac dominant and H3K23me2 dominant have not been found in previous literature, which shows the sensitivity of our method in the task of important feature detection.

2.4 Discussion

In this paper, we described contextual regression, a generalized nonlinear model that is human interpretable. This method performs well in terms of prediction accuracy and important feature extraction in both simulated and experimental datasets. It can be easily applied to construct accurate, interpretable models for datasets in biology and other fields to discover predictive patterns with high sensitivity. On epigenetic datasets, our method has greatly outperformed the state-of-the-art algorithm in the task of predicting DHS and RNAseq using only histone modification data from the same cell-line. Our method revealed new features, including H3K27me3, H3K18ac and H3K23me2, associated with open chromatin that exhibit signature patterns in both significant histone marks and DNA sequence motifs. Importantly, our method also discovered important distal locations that are not overlapped with the prediction target sites (DHS peaks), which are difficult to find using previous algorithms. All these results prove the validity of our algorithm and the contextual regression provides a general framework that is applicable to any task of data mining, such as classification of circular RNA and identification of important contributing genomic features[58].

Our work here shows how to construct a type of machine learning model that can achieve the accuracy of a deep neural network while also providing interpretable results on very large datasets. Recently, scientific data are generated at an increasing speed and we believe our method will be highly valuable for the task of mining discoveries in various fields (especially biology). The code and tutorial are available at http://wanglab.ucsd.edu/star/ContextualRegression.

2.5 Acknowledgments

We would like to thank Hossein Hosseini and his fellows from Dr. Radha Poovendran's lab in University of Washington, Seattle for providing us with their implementation of LeNet-5 for study and comparison.

Funding: This work is partially supported by NIH (R01HG009626).

Chapter 2, in full, is a reformatted reprint of the material as it appears in Liu, C., Wang, W. Contextual regression: an accurate and conveniently interpretable nonlinear model for mining discovery from scientific data. arXiv preprint arXiv:171010728. 2017. The dissertation author was a primary investigator and author of this paper.

2.6 Author Contributions

This study was conceived and designed by L.C. and W.W.; algorithm developed, data processed and analyzed by L.C.; Manuscript written by L.C. and W.W. with input from all authors.

2.7 Figures

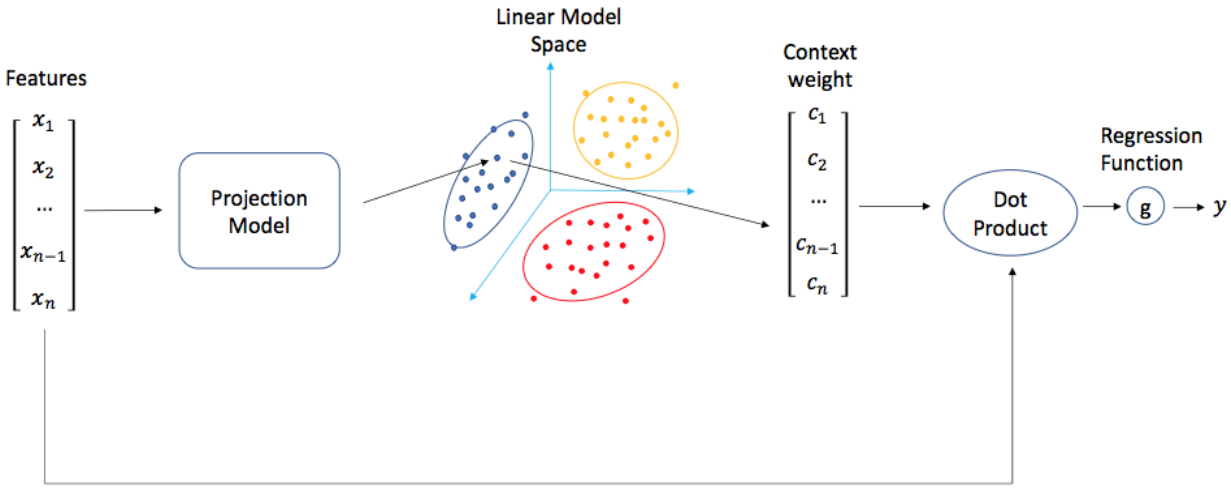


Figure 2.1. A graphic demonstration of the contextual regression model. The projection model (neural network) produces a vector output from a vector input which is used to weight each feature. The weighted features are then summed to generate prediction results. The neural network is trained on these prediction results iteratively so that it can generate weights that make better predictions. At the same time, the weighted features can be analyzed for importance. Different colors of points in the linear model space is an illustration that they can be classified into different subtypes.

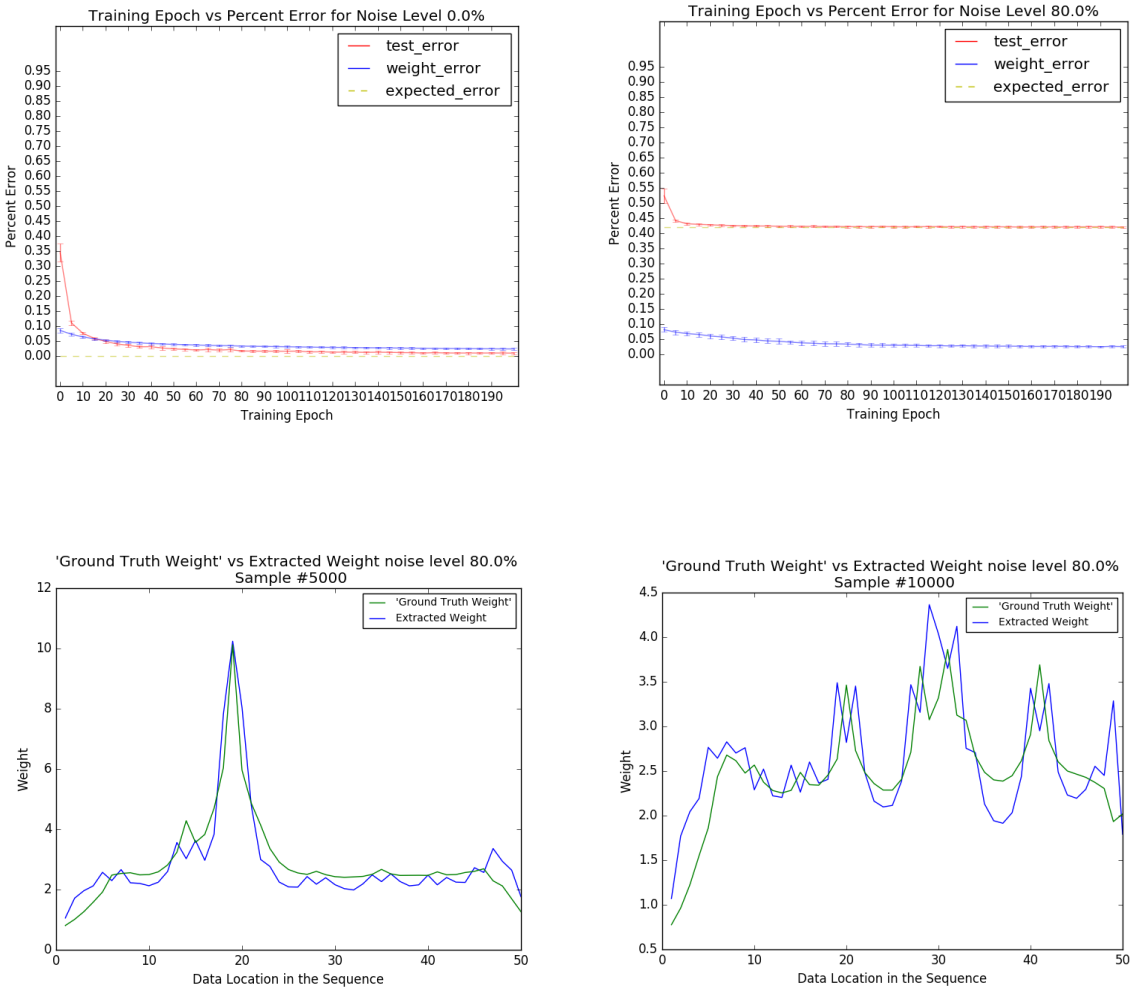


Figure 2.2. Performance assessment of contextual regression on simulated dataset- Top: Contextual regression performance on the testing set under noise level 0% and 80%. Height of the error bar is 1 standard deviation of RMSE or RMSCD among 20 runs. In each epoch (training cycle), the model was trained on 10% of the randomly selected data from the training set and tested on the whole testing set. We chose 10% of the data as one epoch, other than 100% that is ordinarily used, to show the evolution of error. Middle: Sample plot of "ground truth" context weight (green) vs context weight (blue) calculated by the embedding net in the contextual regression model for ±80% noise dataset. Bottom: Same comparison plots for ±0% noise dataset.

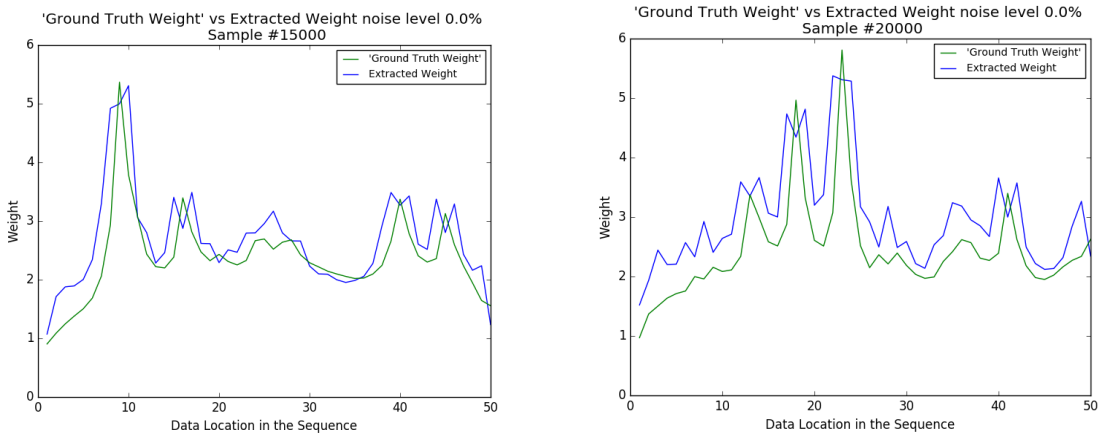


Figure 2.2. Performance assessment of contextual regression on simulated dataset- Top: Contextual regression performance on the testing set under noise level 0% and 80%. Height of the error bar is 1 standard deviation of RMSE or RMSCD among 20 runs. In each epoch (training cycle), the model was trained on 10% of the randomly selected data from the training set and tested on the whole testing set. We chose 10% of the data as one epoch, other than 100% that is ordinarily used, to show the evolution of error. Middle: Sample plot of "ground truth" context weight (green) vs context weight (blue) calculated by the embedding net in the contextual regression model for ±80% noise dataset. Bottom: Same comparison plots for ±0% noise dataset.

DHS Prediction Task	Pearson Correlation	Match1	Catch1obs	Catch1imp
Linear Regression	0.586	44.9%	75.3%	79.1%
Lasso Regression	0.614	44.3%	76.0%	78.6%
Lasso Regression (with log input)	0.537	47.4%	78.3%	82.0%
Contextual Regression	0.800	60.5%	90.5%	92.2%
Contextual Regression (with log input)	0.825	61.1%	90.0%	89.8%
LSTM Benchmark (with log input)	0.817	60.9%	89.2%	89.5%

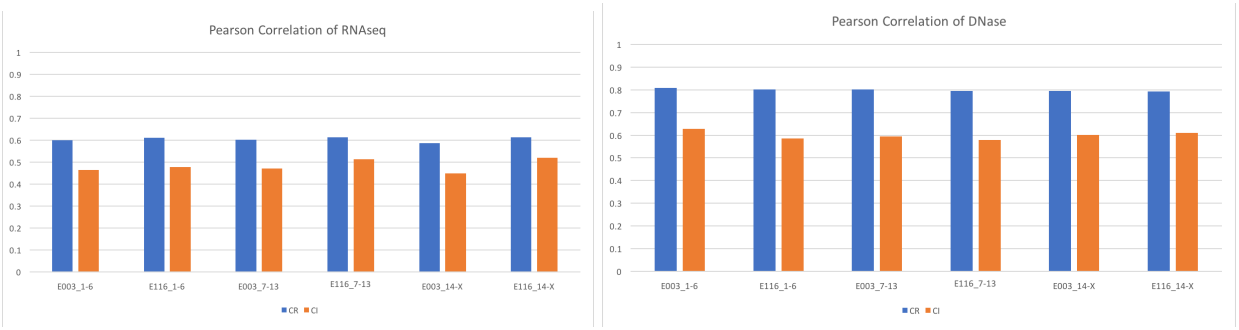


Figure 2.3. Prediction performance comparison of contextual regression with linear regression, LSTM model and ChromImpute- Top: Performance comparison of contextual regression with various linear regression and LSTM models on DHSs dataset in cell-line H1. We used four imputation quality evaluation metrics[30]: pearson correlation, match1 score (percent of 99 percentile experimental values in 99 percentile predicted values), Catch1obs (percent of 99 percentile experimental values in 95 percentile predicted values) and Catch1imp (percent of 99 percentile predicted values in 95 percentile experimental values). Bottom: Comparison of prediction accuracy (Pearson Correlation) between contextual regression (CR) and ChromImpute (CI).

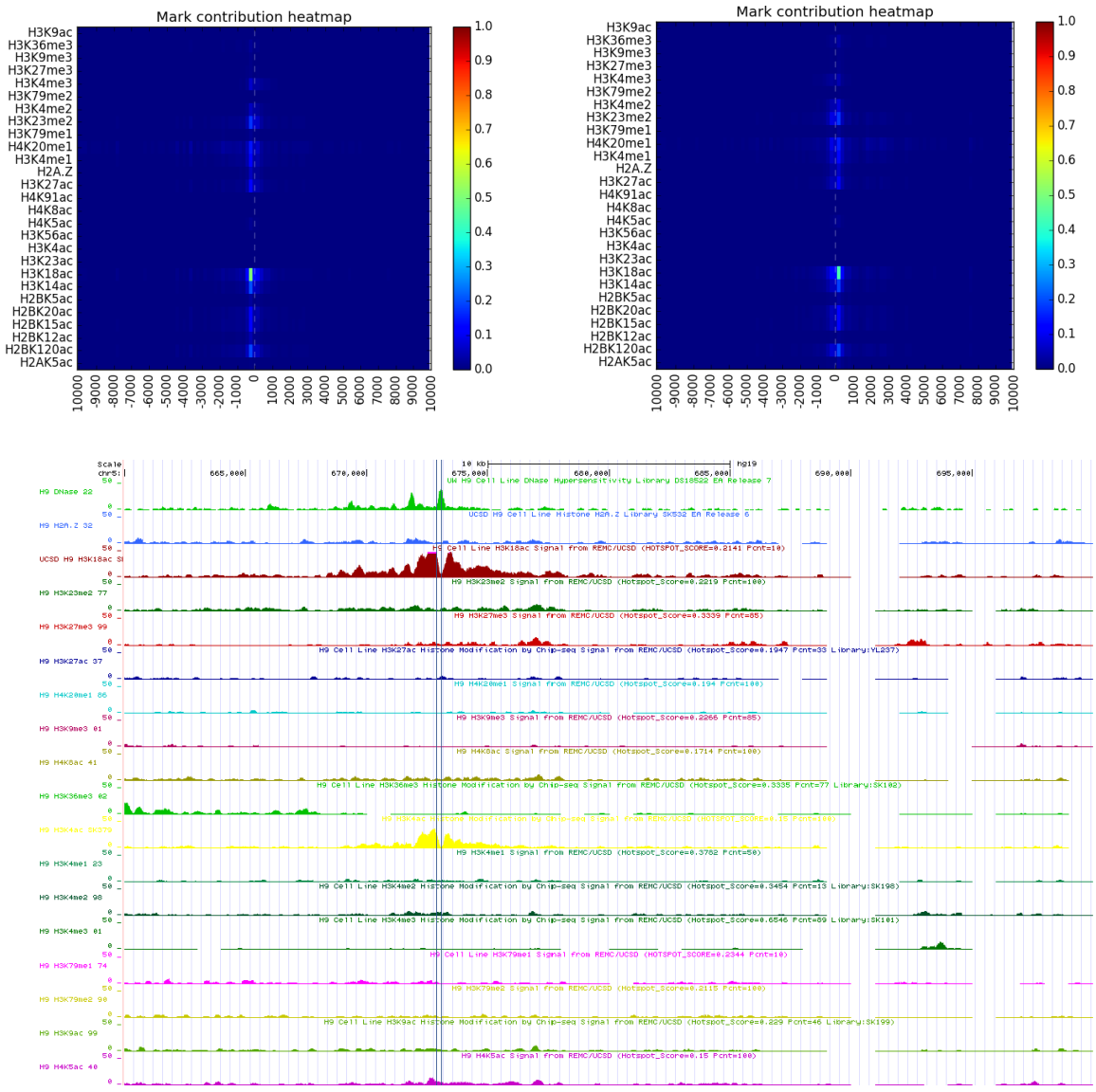
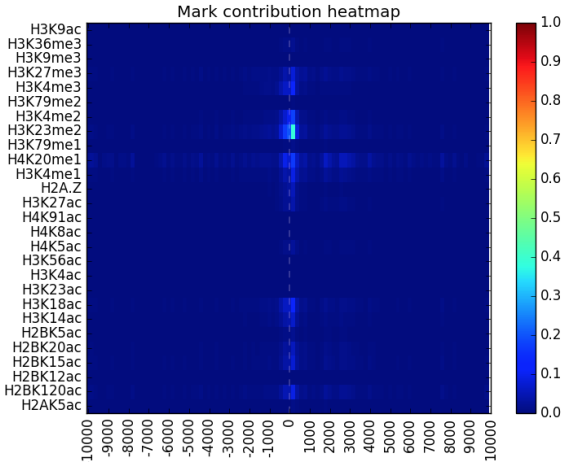


Figure 2.4. Study of H3K18ac pattern- Top: Mark contribution patterns of cluster 15 and 16 in H9 (H3K18ac pattern). Middle: Example genome browser view of the cluster 15 (H3K18ac dominant cluster). Bottom: Motif enrichment profile of cluster 15 in H9 (H3K18ac pattern).

H9-Cluster 15	Frequency Diff	p-value	Frequency in Cluster	Background Frequency
USF2	13.1%	2.03e-93	70.8%	57.7%
MYC	12.9%	4.28e-88	60.3%	47.4%
MAX	12.1%	1.71e-77	61.9%	49.9%
MLXPL	11.5%	1.34e-76	75.5%	64.1%
PLAL1	11.3%	2.70e-84	82.4%	71.1%
MYCN	11.1%	6.18e-70	48.6%	37.5%
ZIC3	10.4%	1.88e-65	77.1%	66.7%
USF1	9.9%	2.31e-53	53.2%	43.3%

Figure 2.4. Study of H3K18ac pattern- Top: Mark contribution patterns of cluster 15 and 16 in H9 (H3K18ac pattern). Middle: Example genome browser view of the cluster 15 (H3K18ac dominant cluster). Bottom: Motif enrichment profile of cluster 15 in H9 (H3K18ac pattern).



	Cluster 5	Cluster 4	BG
CEBPD	- 4.1%	- 20.3%	85.3%
CEBPG	- 5.3%	-16.7%	79.7%
CEBPB	- 5.6%	-16.7%	78.5%
E2F1	+9.4%	+25.6%	57.6%
E2F4	+7.9%	+24.3%	39.2%
E2F3	+4.7%	+21.5%	15.6%
E2F2	+3.8%	+15.0%	16.5%

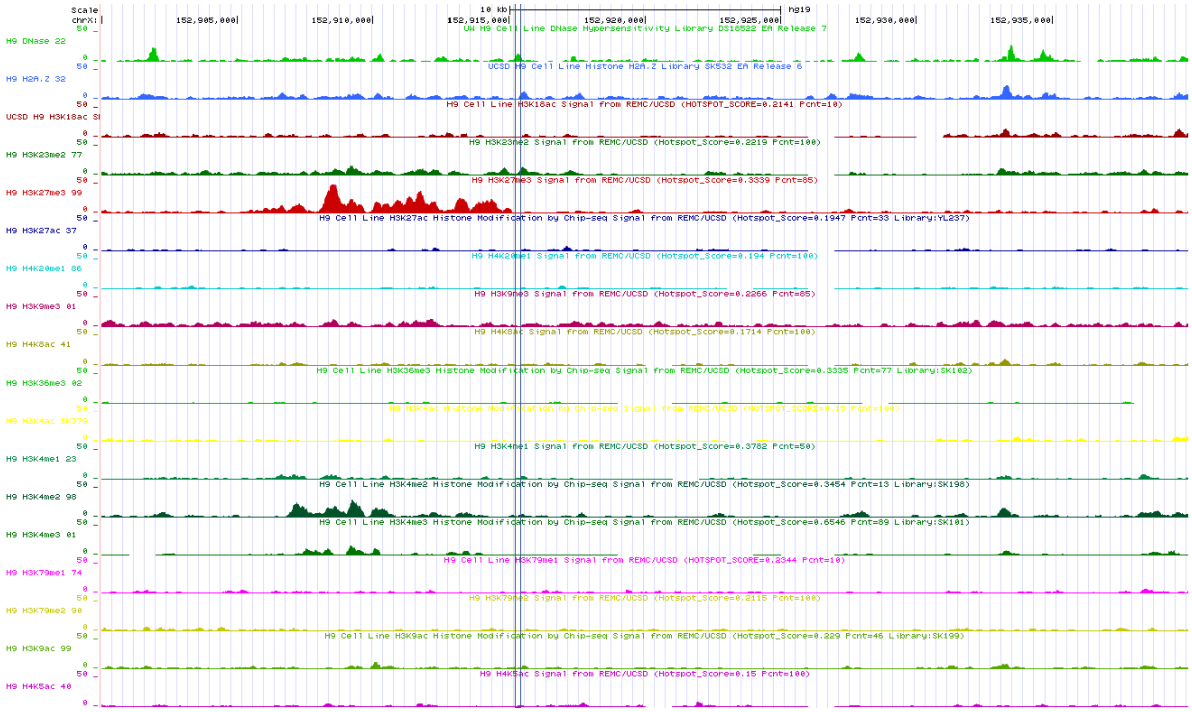


Figure 2.5. Study of H3K23me2 pattern- Top Left: Mark contribution pattern of cluster 5 H9 (H3K23me2 pattern), Top Right: Motif enrichment profile comparison between cluster 4 (H3K27me3 dominant) and 5 (H3K23me2 dominant) in H9. The 2nd and 3rd column are the motif frequency difference from the background and the 4th column is the BG (background) frequency, Bottom: Example Genome browser view of the cluster 5 (H3K23me2 pattern)

2.8 Supplementary Methods

Setup of the Neural Networks used in This Article. The network structure we used for the three tasks are shown as followed:

In the graphs, features is the input vector and y is the prediction target, context is the linear model produced by the embedding network from the input vector, LT represents linear transformation which maps a vector x to Ax+b where A is an matrix and b is an vector, FNN represents Feedforward neural network (here, a fully connected layer) which maps a vector x to s(Cx+d) where C is an matrix, d is an vector and s is the activation function (we used tanh), DP represents the dot product of the two input vectors. O_t is the output of the Bidirectional LSTM at the distal location t, after processing $X_{1..t}$ (forward) and $X_{n..t+1}$ (backward) in sequence. In LT and FNN, matrices A, C and vectors b, d are automatically learned by the model to maximize the prediction accuracy.

The Bidirectional LSTM is implemented with the tensorflow static_bidirectional_rnn function using layer size of 2. The batch size is set to 10. RMSD between the real and predicted data is used as the cost function in training, alone with lasso or utopia penalty as mentioned in the main text. The ADAM optimization algorithm[59] is used to train the model, learning rate is set to 1e-4, maximum gradient norm to 5. Hidden size is set to 80 in the ground truth test, 30 in the prediction accuracy comparison task and 40 in the discovering task, the hidden sizes are chosen to be approximated to the number of features in each task.

In the task of DHS and RNAseq prediction, there are only a small number of high confidence DHS and RNAseq regions. Thus from the training set, we randomly removed 70% of

the data between 60 and 90 percentile, 80% if the data between 30 and 60 percentile and 90% of the data below 30 percentile to balance the dataset.

Code Download and a Tutorial on Converting Your Own Predictive Model to Contextual Regression. Most neural network models with a vector output can be converted into contextual regression for feature analysis. For detail and sample code, please go to: https://github.com/HomoSapienLCY/Contextual_Regression

Setup of the Ground Truth Test (in section: Evaluating the Contextual Regression model on simulated data). The input of the Neural Network is a sequence of features of length 50 and the output is a float value y. Each feature ($x_1 \ldots x_{50}$) are sampled from the set (1, 3, 9, 27, 81) with probability (70%, 20%, 7%, 2.5%, 0.5%) and a random fraction below 20% is added or subtracted from the value to increase data diversity. The "Ground Truth weight" c_i and the expected output y are calculated using the function in the main text. After y is generated, a random noise between ±n% (n ranges from 0 to 1000 as mentioned in the main text) is added to y to simulated a noisy data collection environment.

In each run a total of 100000 data points are created, 70% are used for training and 30% are used for testing. A total of 20 simulations are run. The difference between learned weight and expected weight are measured by root mean square cosine distance (RMSCD) with the formula below:

$$\text{RMSCD} = \sqrt{\frac{1}{N}\sum_{i=1}^{N} CD(w_i, u_i)}$$

$$CD(w_i, u_i) = 1 - \left| \frac{w_i \cdot u_i}{\|w_i\|_2 \|u_i\|_2} \right|$$

Calculation of the Expected Error in the Simulated Data (in section: Evaluating the Contextual Regression model on simulated data): Under noise level of a, the error is uniformly sampled from the interval (-a, a). At a certain data point, suppose the error rate sampled is x% ∈ (-a, a). Then the target value after the addition of noise will be (1+x%) of the target value before the addition of noise, which is an x% different. Thus the expected error in RMSD will be the following integral:

$$\sqrt{\frac{\int_{-a}^{a} x\%^2 dx}{\int_{-a}^{a} (1+x\%)^2 dx}} = \sqrt{\frac{2a^3}{2a^3+6a}} = \sqrt{\frac{1}{1+\frac{3}{a^2}}}$$

Source of Data. Data of DHS, RNAseq and ChIP-seq are downloaded from the roadmap website (http://www.roadmapepigenomics.org/). The ENCODE blacklis[60] is used to filter the data. The motif scanning is made using the FIMO function of MEME with HOCOMOCOv9.meme database. The transcription site and coding sequence statistics is done with GENCODE Gene V27lift37 file downloaded from the Table Browser

(http://genome.ucsc.edu/cgi-bin/hgTables). The motif appearance frequency (MAF) is defined as shown below:

$$\text{MAF} = \frac{Number\ of\ 40kbp\ regions\ that\ contain\ the\ motif}{Total\ number\ of\ regions}$$

Comparison of Contextual Regression with Linear Regression, Bidirectional-LSTM and ChromImpute (in section: Evaluating the Contextual Regression model on DNaseI hypersensitivity data). The linear regression is trained with the same setting and input features as contextual regression using tensorflow. The only difference is the replacement of contextual regression model with a linear regression model.

The architecture of the bidirectional-LSTM network is as illustrated in the previous section of the supplement material (Setup of the Neural Networks used in This Article). Its parameter setting is the same as the setting of contextual regression.

The ChromImpute model is run using the same setup as shown in the ChromImpute website (http://www.biolchem.ucla.edu/labs/ernst/ChromImpute/). The prediction is first done with 25bp resolution as what ChromImpute is designed for, and then the blacklisted regions are removed and the neighboring bins are averaged to yield data of 200bp resolution. The data used for ChromImpute benchmarking are downloaded directly from the processed ChromImpute data folder (https://personal.broadinstitute.org/jernst/roadmapconverted/CONVERTEDDATA).

K-mean Clustering (in section: Discovering Important Histone Mark Patterns in the Open Chromatin Regions of H1). K-mean clustering is used to classify the pattern of histone marks around open chromatin regions. This task is carried out using the KMeans function in the python sklearn package with K = 20 and fixed random state = 0. Before clustering, we applied PCA on the weighted contribution of histone marks at each data point and the top 100 principal components are kept for the clustering step.

In and Across Cluster Distance Calculation (in section: Discovering Important Histone Mark Patterns in the Open Chromatin Regions of H1). All the vectors are first normalized to a length of 1. In cluster distance is the average squared RMSD between each member in the cluster and the cluster center (mean of the member vectors). Across cluster distance of cluster i and j is the average squared RMSD between each member in the cluster i and the cluster center of j.

Feature Contribution (in section: Supplementary Information - Model Performance Check). Feature contribution C_N^W is calculated using the formula below, this formula combines the context weight assigned by the embedding function and the feature magnitude:

$$C_N^W = \frac{C^W}{||C^W||_2}$$

$$C^W = C \odot H$$

$$H_i = h \cdot D$$

In the formula, C is the context weight of each histone mark, h is the histone mark signal strength vector with each element the signal strength at the corresponding location, D is the distal vector and $\odot$ represents element-wise multiplication.

2.10 Supplementary Information

Model Performance Check (in section: Evaluating the Contextual Regression model on DNaseI hypersensitivity data). A good model for data interpretation needs to be accurate but also stable in terms of the result it produces. We assessed the feature selection robustness of contextual regression during 5 runs on the H1 DHS and RNA-seq dataset. In each run, the whole dataset was divided randomly into training and testing sets of 7:3 ratio. During the similarity measurement, we calculated the average feature contribution of each histone mark in 3 data classes (detail of the calculation can be found in the previous section: Feature Contribution) : (1) the whole dataset, (2) the peak regions, defined by p-value > 3, and (3) the flat regions, defined by p-value < 1. We used cosine similarity, the cosine value of the angle between a pair of vectors, as the measure of similarity between weight vectors.

The similarity of distal weight is above 0.9 for both DHS-seq and RNAseq. The similarity of feature weights on the whole dataset and in the flat regions are relatively low, which is not surprising since most of the regions have low value signals and thus there are diverse sets of weight assignment that can predict them to be flat. In the regions of our interest, the peak regions, we observed highly consistent feature weights. For DHS, the average cosine similarity between feature weights is 0.991 and for RNAseq, the average cosine similarity between feature weights is 0.960 (Table 2.1ST).

Among the top 10 features (Table 2.2ST) with the largest feature contribution in DHS prediction, 5 of them (H3K27me3, H3K27ac, H3K9me3, H3K4me1 H3K36me3) are core marks[61]. Besides, H2A.Z[35,62], H4K20me1[63] and H4K8ac[39] have been found to be highly associated with open chromatin formation. In RNAseq, on the other hand, the transcription mark

H3K36me3[64,65] is ranked number 1 as expected. H2A.Z, H4K20me1, H3K4me1-3 and H3K27ac[66,34] are also selected in the top 10, which confirmed that marks highly ranked by our model match the result of analysis in the aforementioned papers. In open chromatin, however, H3K4me2-3 are not important contributors. We found that these two marks have relatively low average signal strength in the DNase peak regions (contain location with logpval > 3) after we applied log to the marks (Table 2.3ST). This may explain why they do not have a high importance ranking in our model: the Lasso penalty will prefer features with higher magnitude to reduce the weight penalty and thus give less emphasis to the features that have small magnitudes.

To check the predictivity of the marks selected by our model, we predict DNase and RNAseq using only the top 10 and top 18 features and found that the prediction accuracy is largely preserved (Table 2.4ST).

Possibility of Alternative Model (in section: The contextual regression method). Consider the case where regression function g is the identity function. Then the formula for the loss function of y and c_i are:

$$L_y = \left|\left| y - \hat{y} \right|\right|_2 = \left|\left| \sum_i (c_i - \hat{c}_i)\, x_i \right|\right|_2$$

$$L_{c_i} = \left|\left| c_i - \hat{c}_i \right|\right|_2$$

Where y and c_i are the prediction target and the i^{th} element of the context weight vector, and $\hat{y}$ *and* $\hat{c}_i$ are their predicted counterparts generated by the contextual regression model. Their

derivative to a model parameter θ_j are:

$$\frac{\partial L_y}{\partial \theta_j} = \frac{\partial L_y}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial \theta_j} = \frac{\partial L_y}{\partial \hat{y}} \sum_i \frac{\partial \hat{c}_i}{\partial \theta_j} x_i$$

$$\frac{\partial L_{c_i}}{\partial \theta_j} = \frac{\partial L_{c_i}}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial \theta_j} = \frac{\partial L_{c_i}}{\partial \hat{y}} \sum_i \frac{\partial \hat{c}_i}{\partial \theta_j} x_i$$

Thus $\frac{\partial L_{c_i}}{\partial \theta_j}$ will have the same sign as $\frac{\partial L_y}{\partial \theta_j}$ only if $\frac{\partial L_y}{\partial \hat{y}}$ and $\frac{\partial L_{c_i}}{\partial \hat{y}}$ has the same sign.

This is not true in general unless we optimize each $\hat{c}_i$ separately, in which case $L_y = L_{c_i}|x_i|$. This shows that when we move L_y in its descent direction, each individual L_{c_i} may not be moved in its descent direction as well. Therefore, when L_y is at its local minimum, it is not necessarily true that L_{c_i} is at its local minimum as well.

Utopia Penalty Technique (in section: The contextual regression method). We have shown that contextual regression is highly effective in terms of extracting important features. However, due to the high flexibility of neural network, it may only find one of the few alternative models that can give similar prediction results. Sometimes the features are not completely informationally independent, especially when people are putting in every features that possibly have an effect. This problem is even more cumbersome to solve for other neural network interpretation methods which have interpretation processes that are decoupled from the

training processes.

To mitigate this problem, methods such as PCA[67], gaussian elimination[68] can be pre-apply to simplify the data. Reducing the number of neurons or applying regularization on the weights in the network to shrink the possible model space are also good options. Besides these traditional approaches, we would like to introduce a more straight-forward way through the mean of modifying the cost function which allows users to monitor the model lively during training.

In our previous section, we applied the Lasso penalty to make the weight sparse. This approach will "deprive" weight from the "less effective" features, which also have predictive power but can be replaced by the combination of other features that require smaller weight assignment. We can counteract this effect by an opposite term that penalizes unbalance feature weights.

One possible term of such is the cosine distance between the context weight vector and a "utopia" vector u which has value 1 for all of its entries. This leads to a cost function as followed:

$$cost = \left|\left|y - \hat{y}\right|\right|_2 + L + B$$

$$L = \lambda \left|\left|w\right|\right|_1$$

$$B = \mu \left|\left|CD(w, u)\right|\right|_1$$

$$CD(w, u) = 1 - \left|\frac{w \cdot u}{||w||_2 ||u||_2}\right|$$

$$u = [1, 1, .., 1]$$

Where w is the context weight, λ is the Lasso penalty parameter, and CD stands for cosine distance. In practice, the user can set their minimum accuracy tolerance and gradually tune up the utopia penalty parameter μ until the accuracy is below the tolerance threshold.

As a demonstration, we performed this strategy on H1 DNase data. We started with λ = 1e-3 and μ = 0, increment μ from 1e-5 to 1e-1 in an exponential step size. Each run lasted for 150 epochs. The accuracy of the model with μ = 0 is 60.4% Match1 at epoch 150, thus we set the accuracy tolerance to be 59.4% Match1. We found that a utopia penalty parameter between 1e-5 and 1e-3 gives us Match1 above the tolerance. Under these levels of utopia penalty, we can see the contribution of the highest mark, H2A.Z decreases and the contribution of other marks increases in tune with the increment of utopia penalty strength (Figure 2.2ST). This illustrates that tuning up the utopia penalty is an effective and straightforward way to obtain a diverse set of alternative models.

2.11 Supplementary Figures And Tables

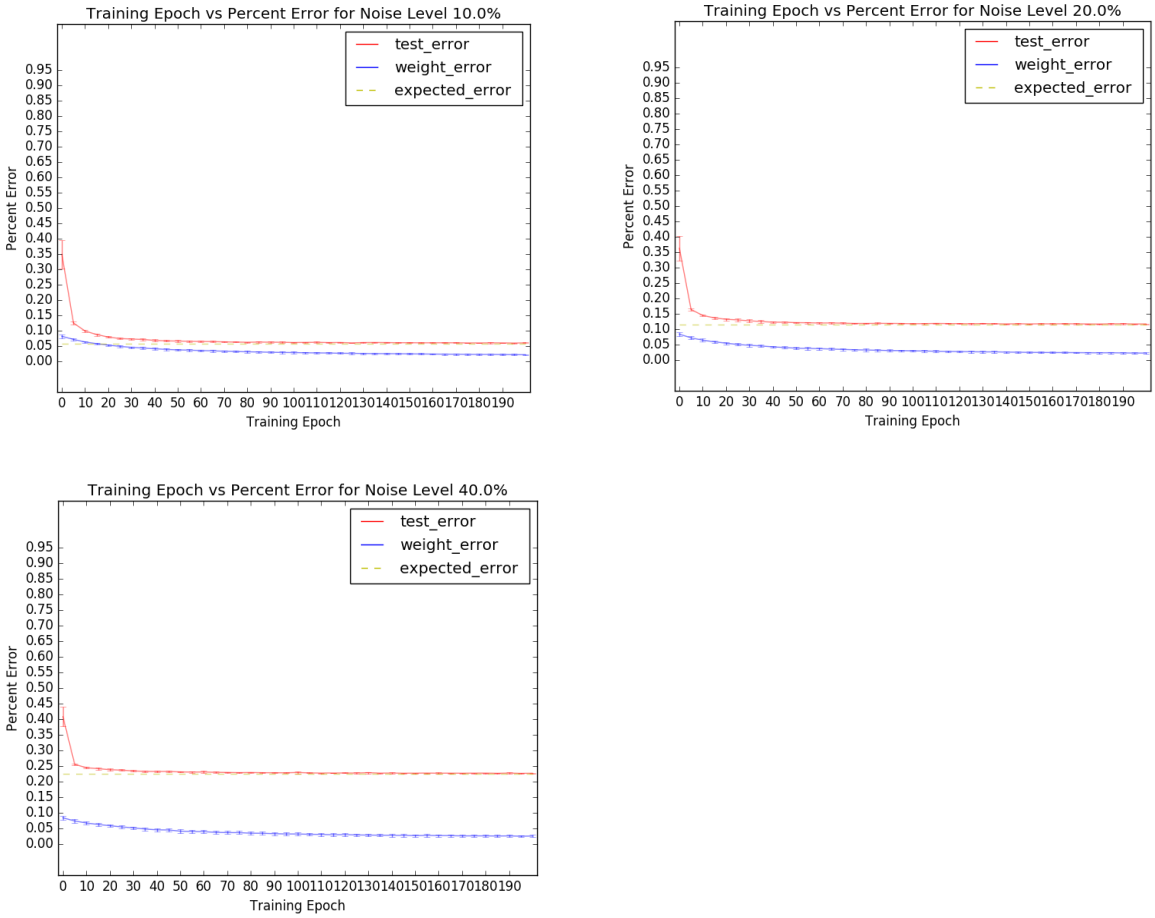


Figure 2.1S. Contextual regression performance on the simulated datasets under noise level 10%, 20% and 40%. Height of the error bar is 1 standard deviation of RMSE or RMSCD among 20 runs. In each epoch (training cycle), the model is trained on 10% of the randomly selected data from the training set and tested on the whole testing set.

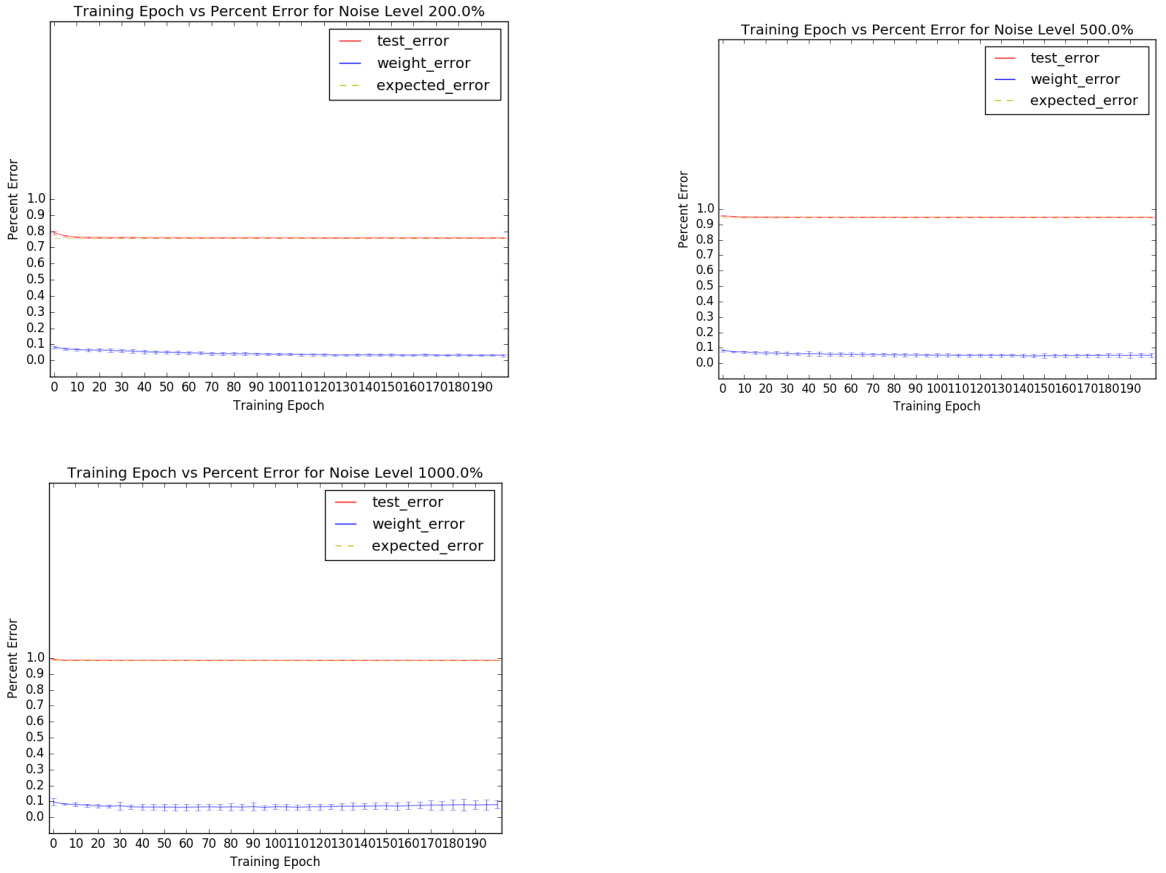


Figure 2.2S. Prediction and rule extraction accuracy of contextual regression on the simulated datasets under very high noise level.

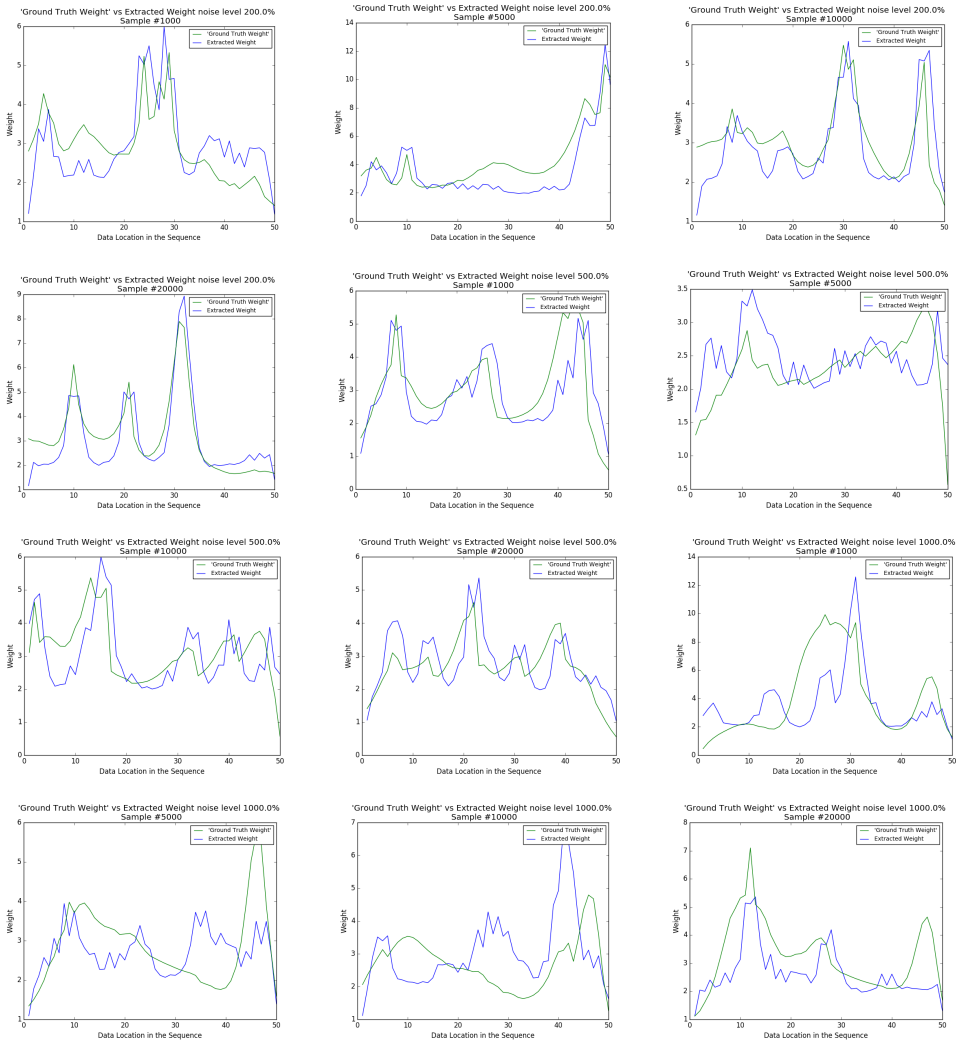


Figure 2.3S. Sample plot of "ground truth" context weight (green) vs context weight (blue) calculated by the embedding net in the contextual regression model under very high noise level.

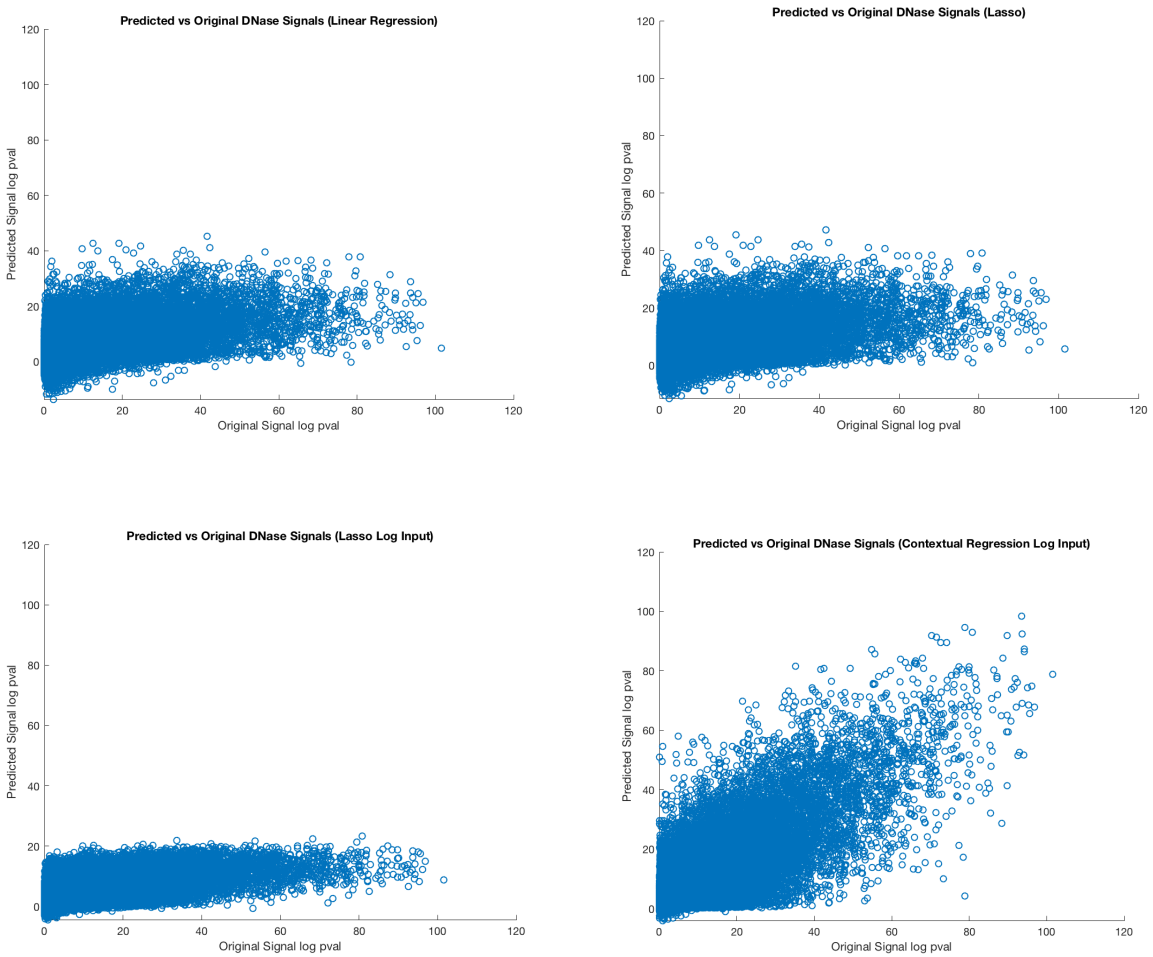


Figure 2.4S. Plot of original signal vs predicted signal by Linear Regression (top left), Lasso Regression (top right), Lasso Regression with log input (bottom left) and contextual regression with log input (bottom right) on DHSs in cell-line H1.

Figure 2.5S. Comparison of prediction accuracy (Match1, Catch1Obs, Catch1Imp) between contextual regression (CR) and ChromImpute (CI).

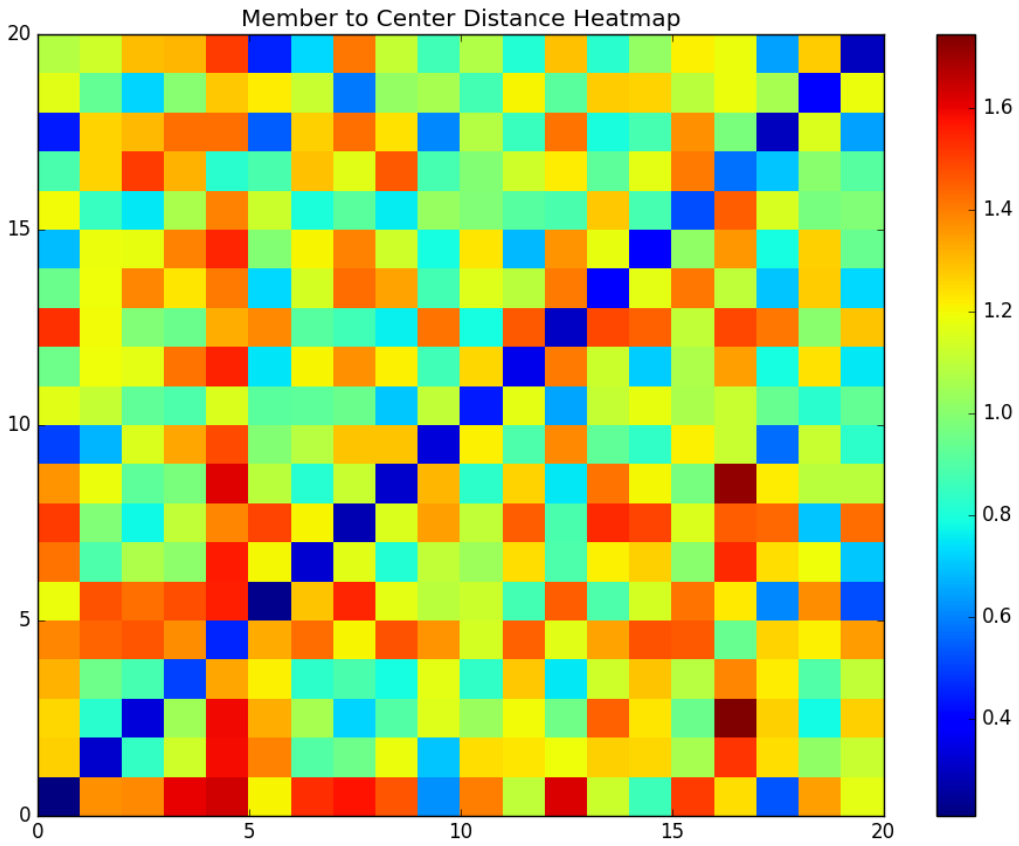


Figure 2.6S. Average in (on the main diagonal) and across (off the main diagonal) cluster distance from each cluster center

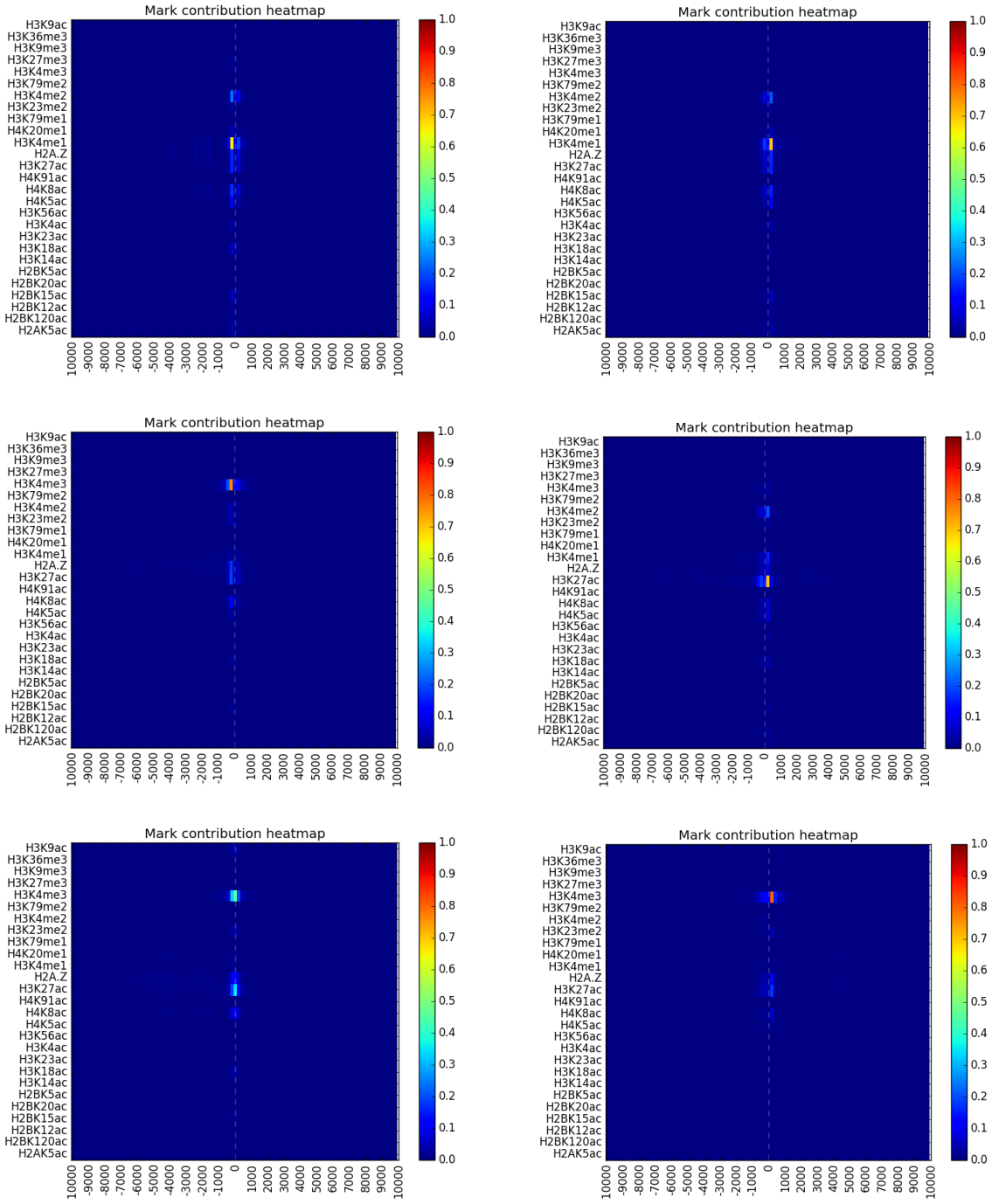


Figure 2.7S. Examples of the Type 1 (central dominant histone marks) cluster mark contribution patterns

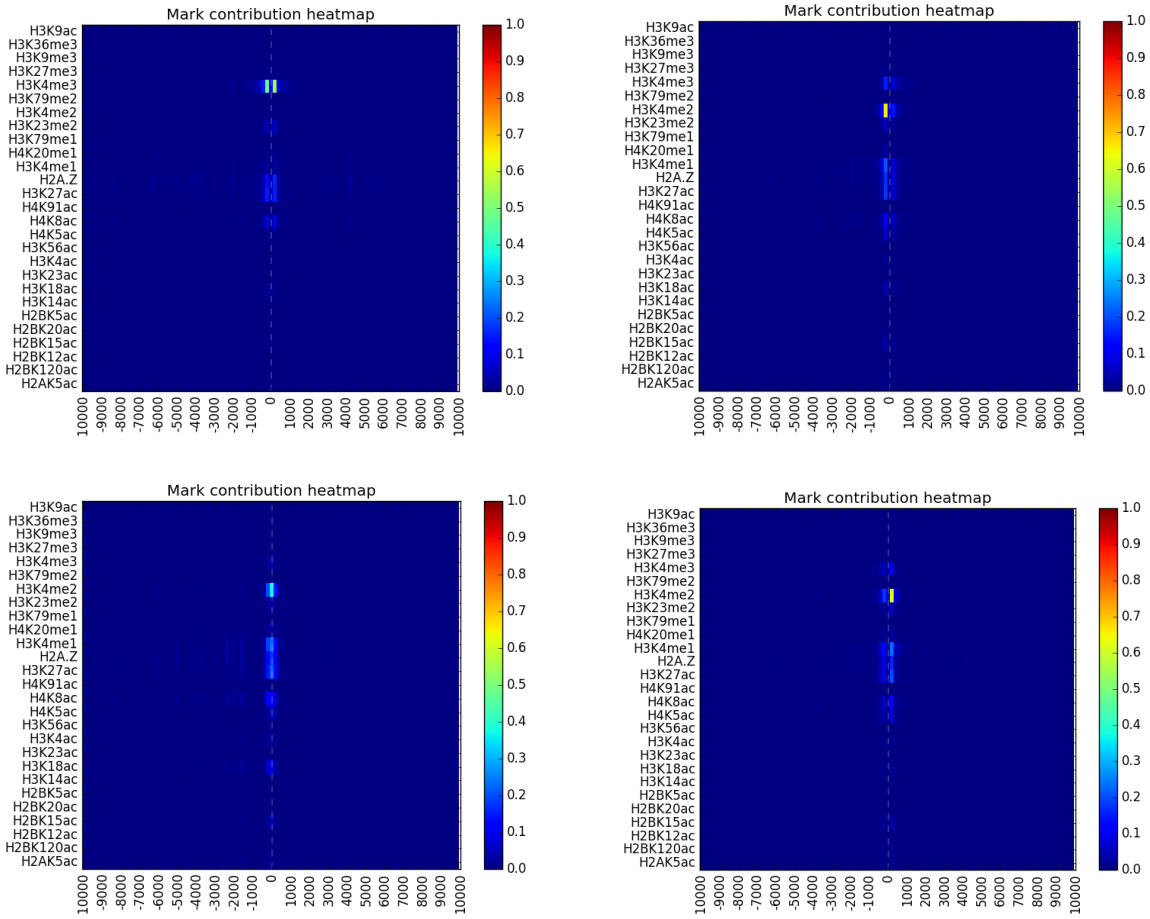


Figure 2.7S. Examples of the Type 1 (central dominant histone marks) cluster mark contribution patterns

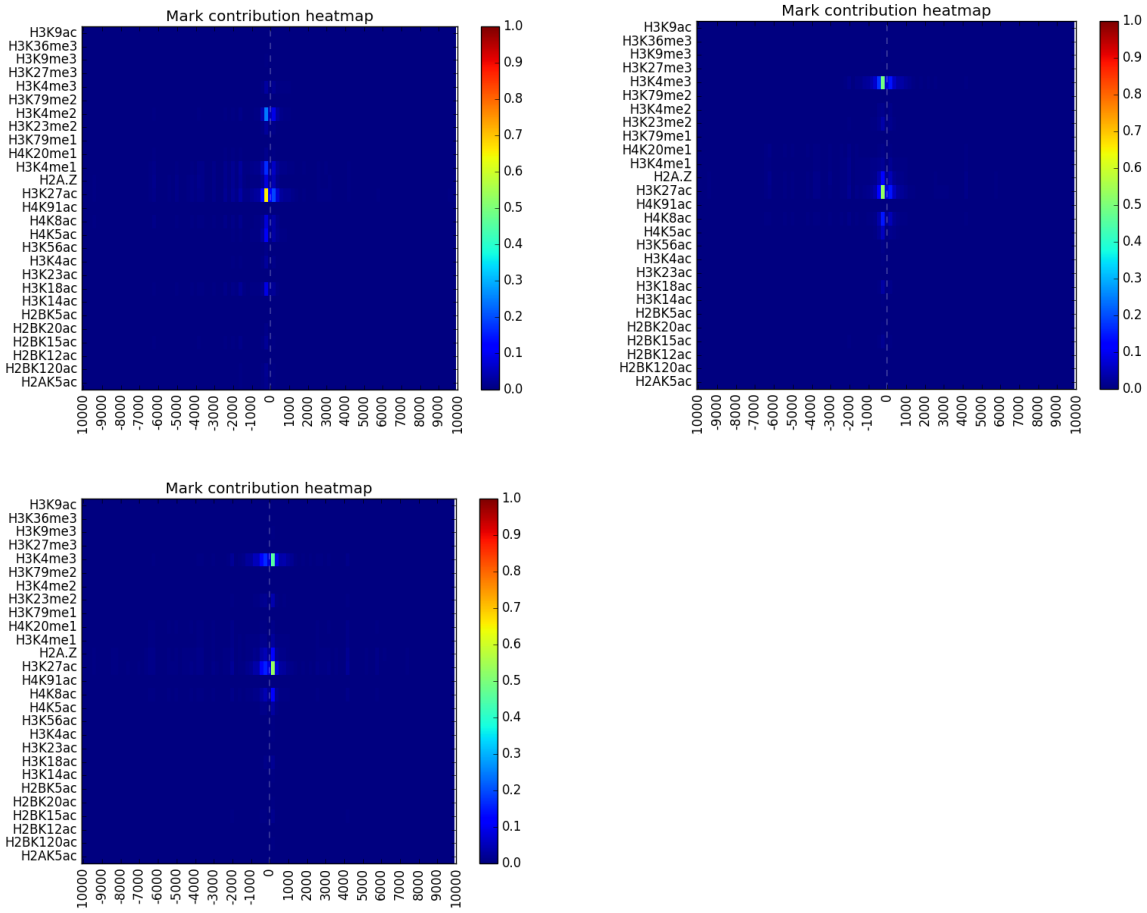


Figure 2.7S. Examples of the Type 1 (central dominant histone marks) cluster mark contribution patterns

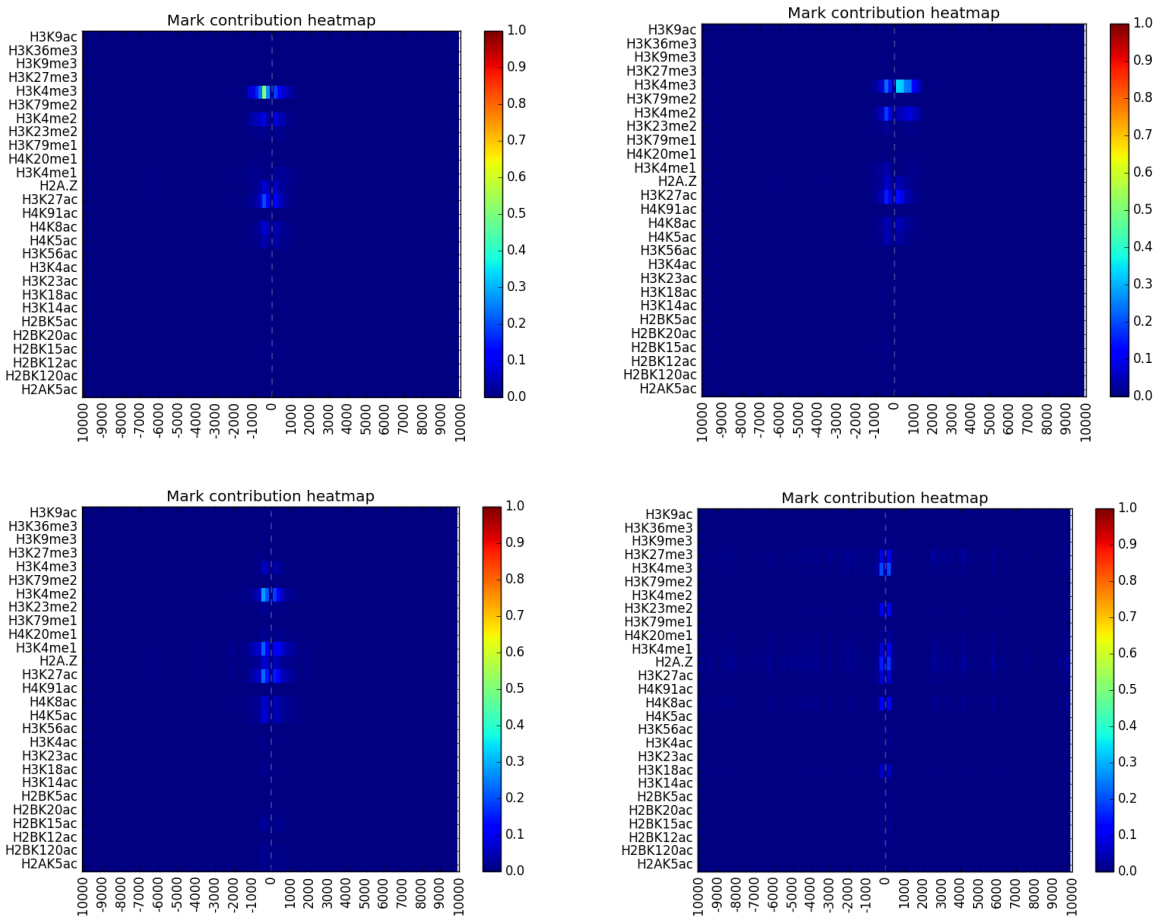


Figure 2.8S. Type 2 (spread histone marks) cluster mark contribution patterns

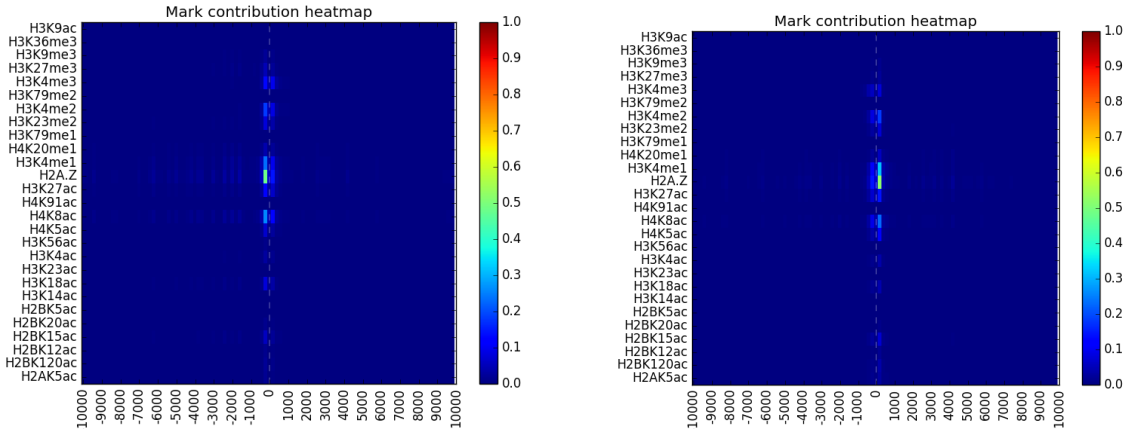


Figure 2.9S. Type 3 cluster mark (central dominant H2A.Z) contribution patterns

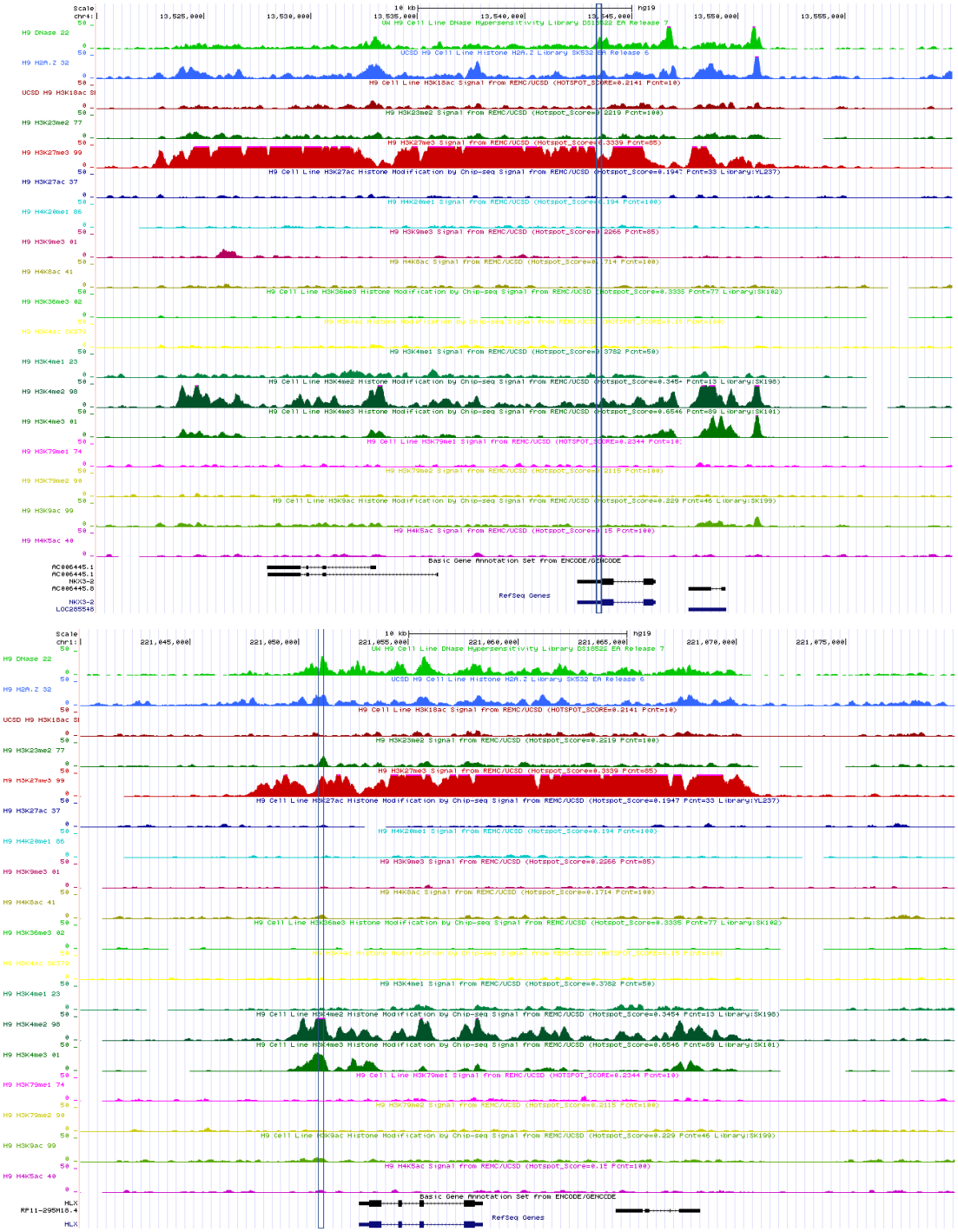


Figure 2.10S. Example genome browser views of cluster 4 in H1 cell line (H3K27me3 cluster), the blue square is the location of the peak in the cluster detected by our method. Most of these regions include a transcription start site (TSS) in their neighborhood.

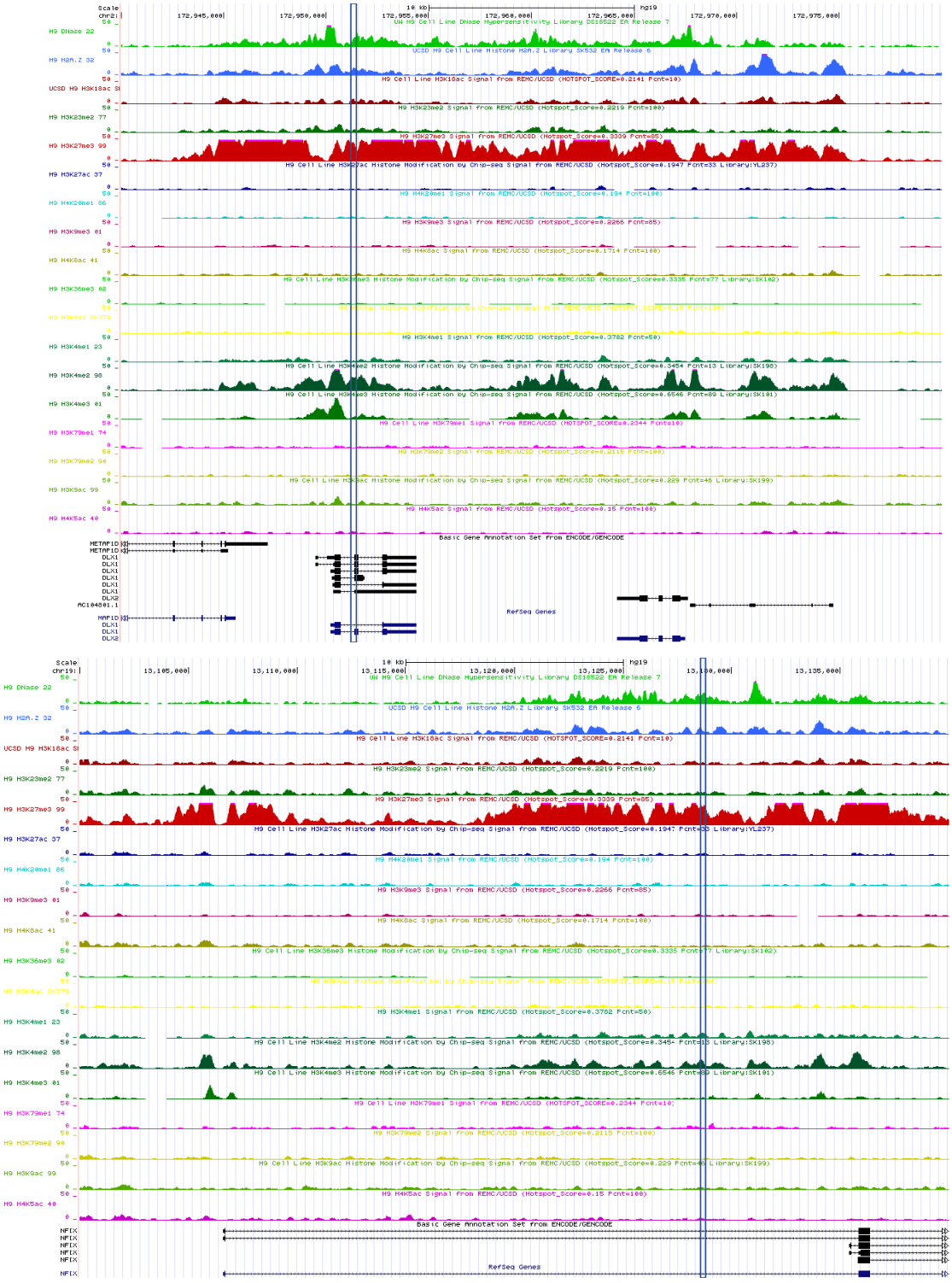


Figure 2.10S. Example genome browser views of cluster 4 in H1 cell line (H3K27me3 cluster), the blue square is the location of the peak in the cluster detected by our method. Most of these regions include a transcription start site (TSS) in their neighborhood.

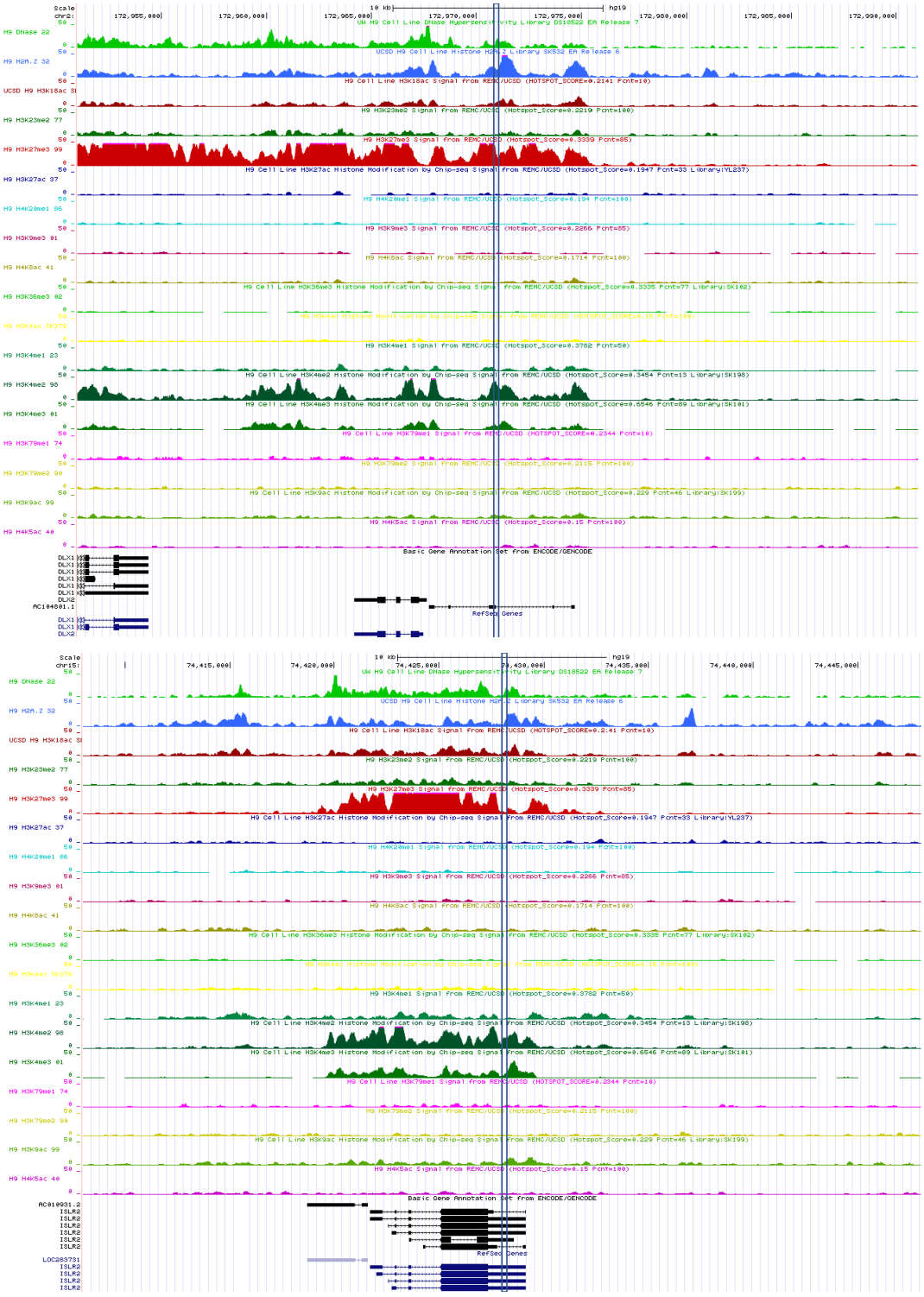


Figure 2.10S. Example genome browser views of cluster 4 in H1 cell line (H3K27me3 cluster), the blue square is the location of the peak in the cluster detected by our method. Most of these regions include a transcription start site (TSS) in their neighborhood.

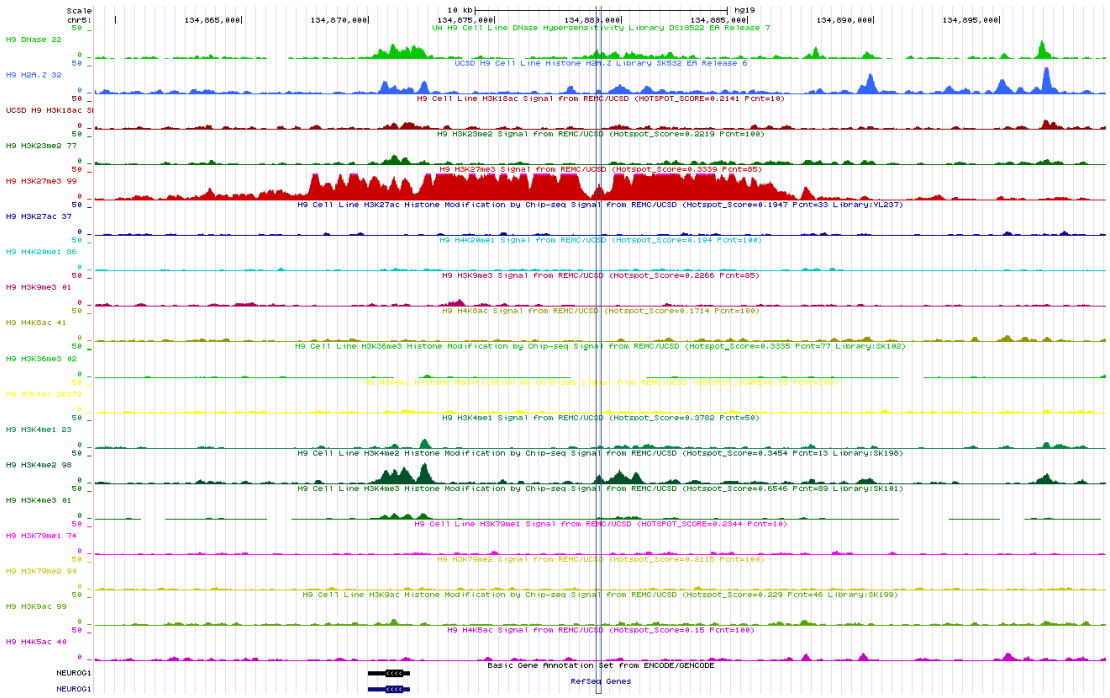


Figure 2.10S. Example genome browser views of cluster 4 in H1 cell line (H3K27me3 cluster), the blue square is the location of the peak in the cluster detected by our method. Most of these regions include a transcription start site (TSS) in their neighborhood.

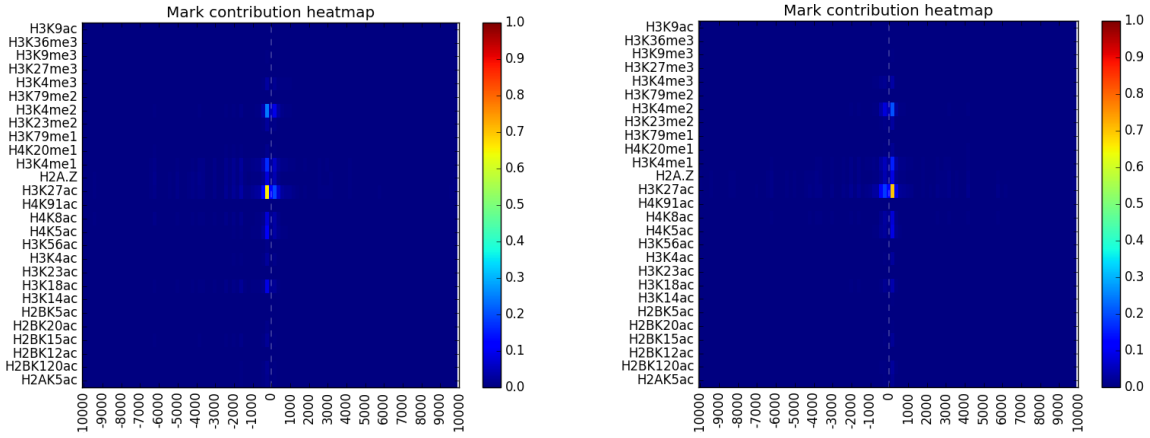


Figure 2.11S. Mark contribution patterns of cluster 1 and 6 in H1 cell line, H3K27ac dominant clusters

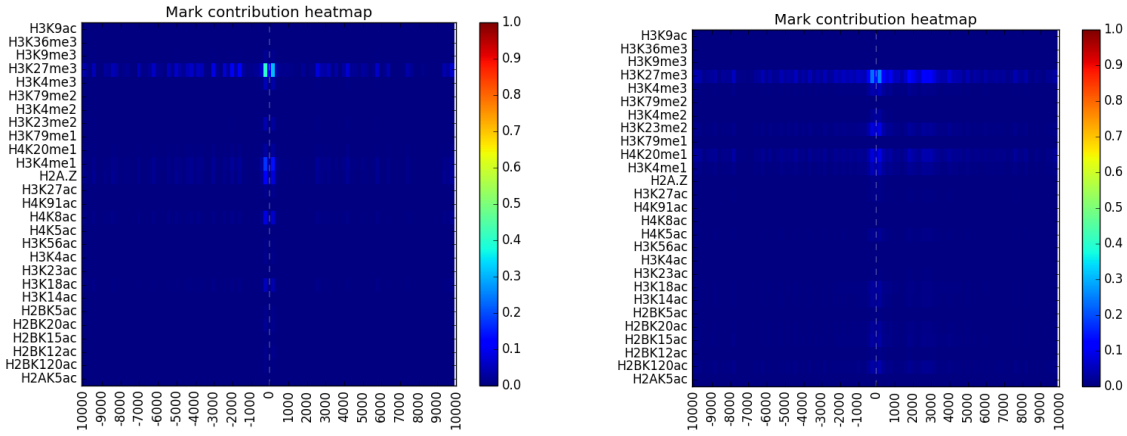


Figure 2.12S. Mark contribution pattern of cluster 4 in H1 (left) and H9 (right) cell lines, both are H3K27me3 patterns

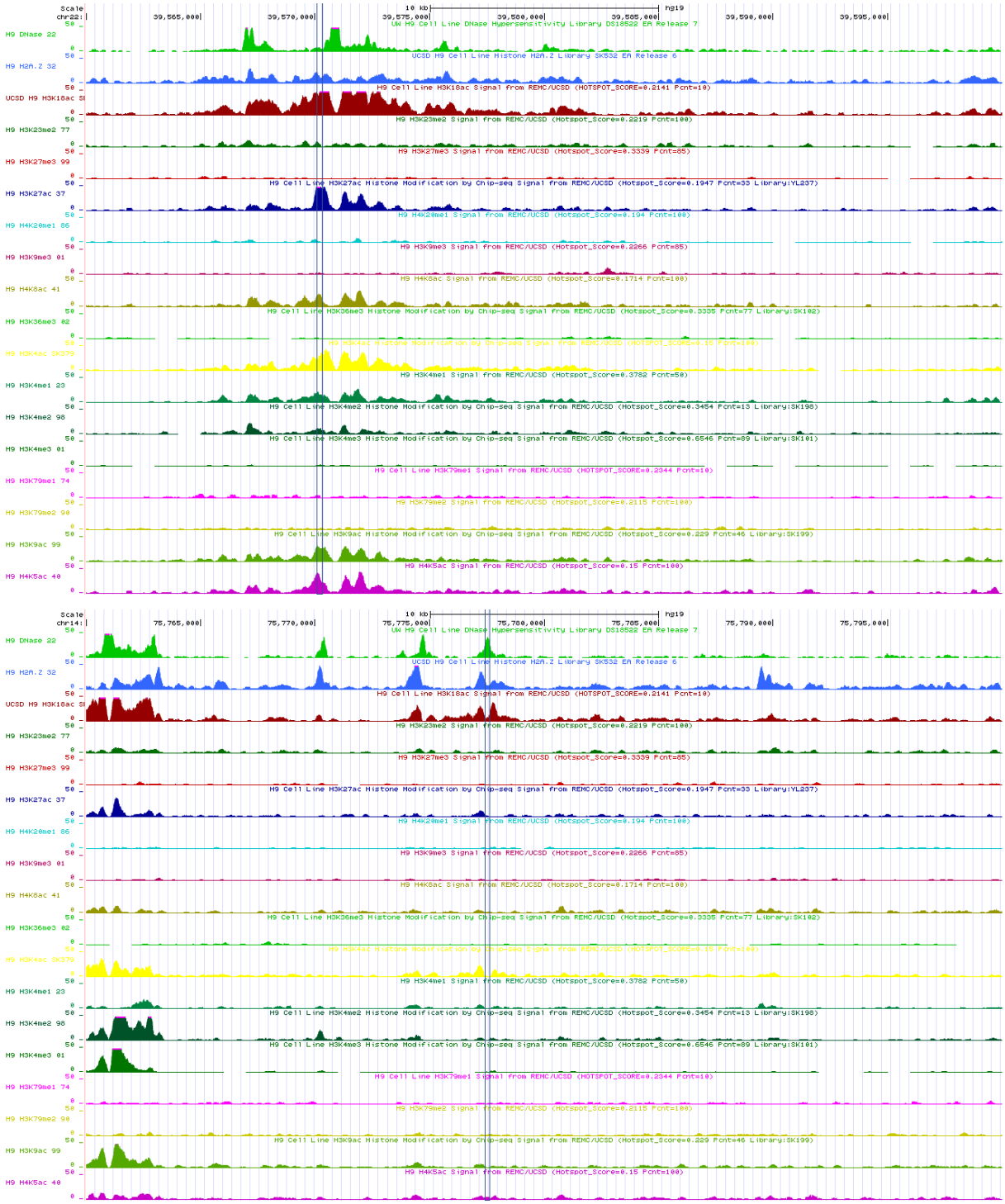


Figure 2.13S. Example genome browser views of the cluster 15 (H3K18ac dominant cluster) in H9 cell line

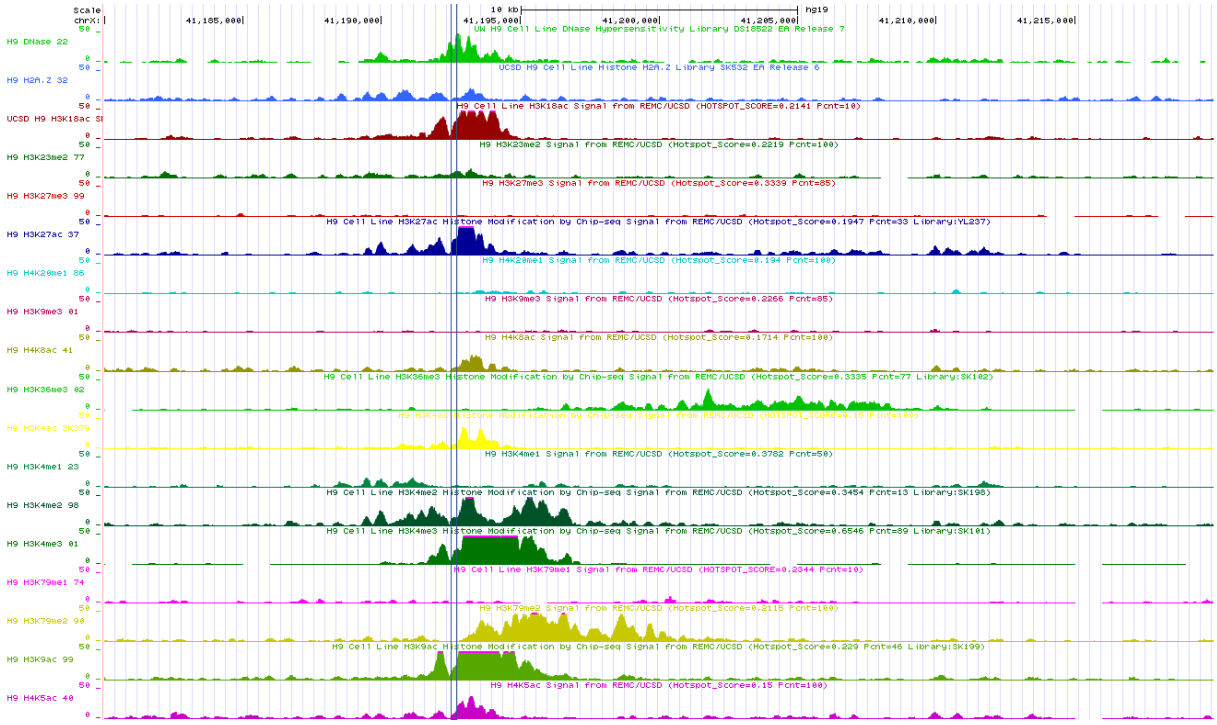


Figure 2.13S. Example genome browser views of the cluster 15 (H3K18ac dominant cluster) in H9 cell line

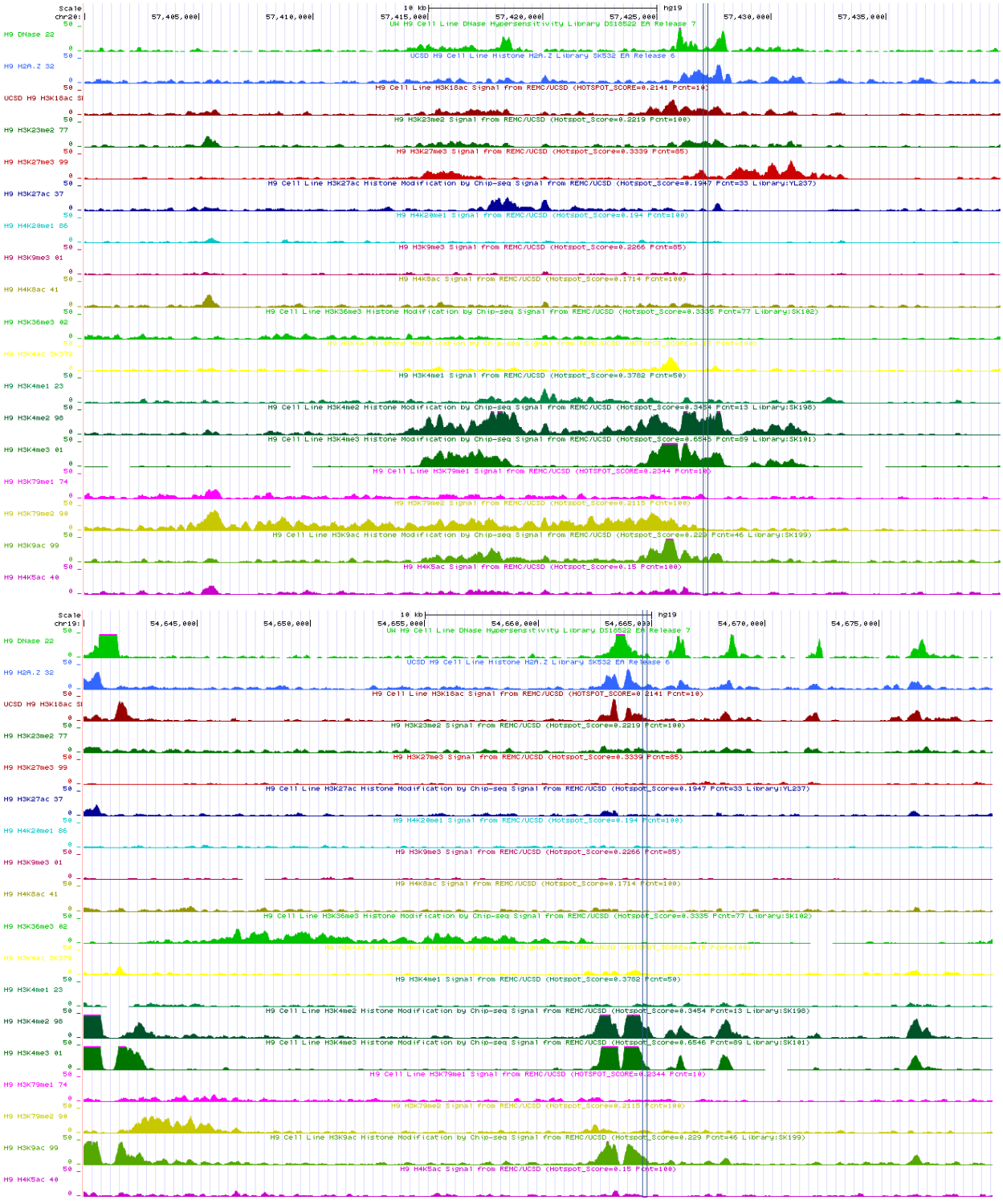


Figure 2.14S. Example Genome browser views of the cluster 5 (H3K23me2 pattern) in H9 cell line

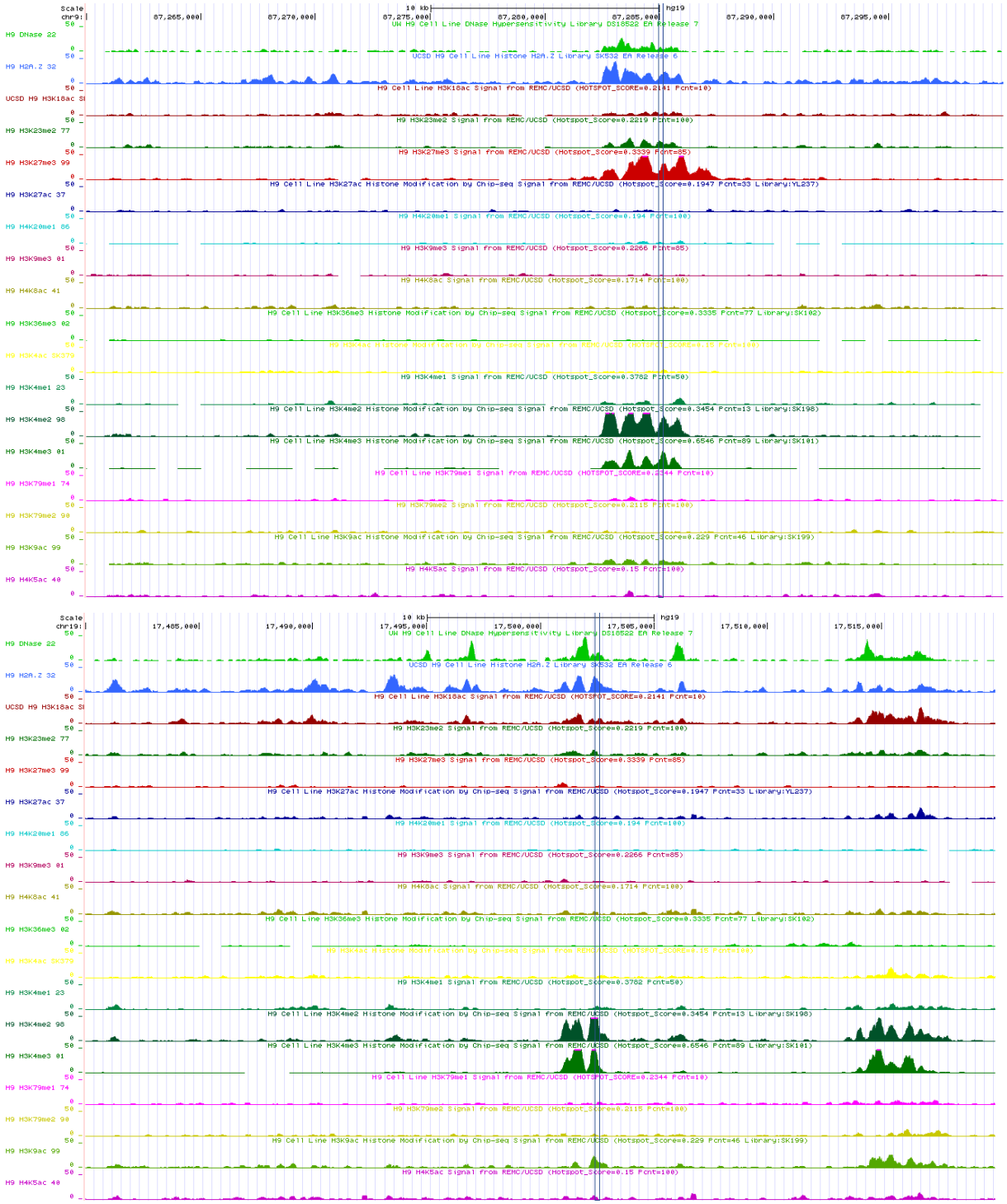


Figure 2.14S. Example Genome browser views of the cluster 5 (H3K23me2 pattern) in H9 cell line

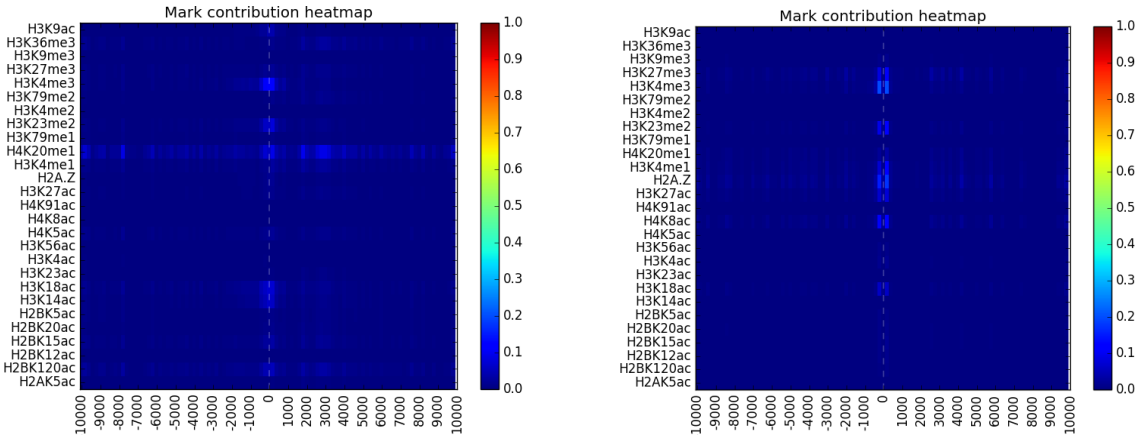


Figure 2.15S. Mark contribution patterns of cluster 16 in H1 and cluster 0 in H9 (spread type)

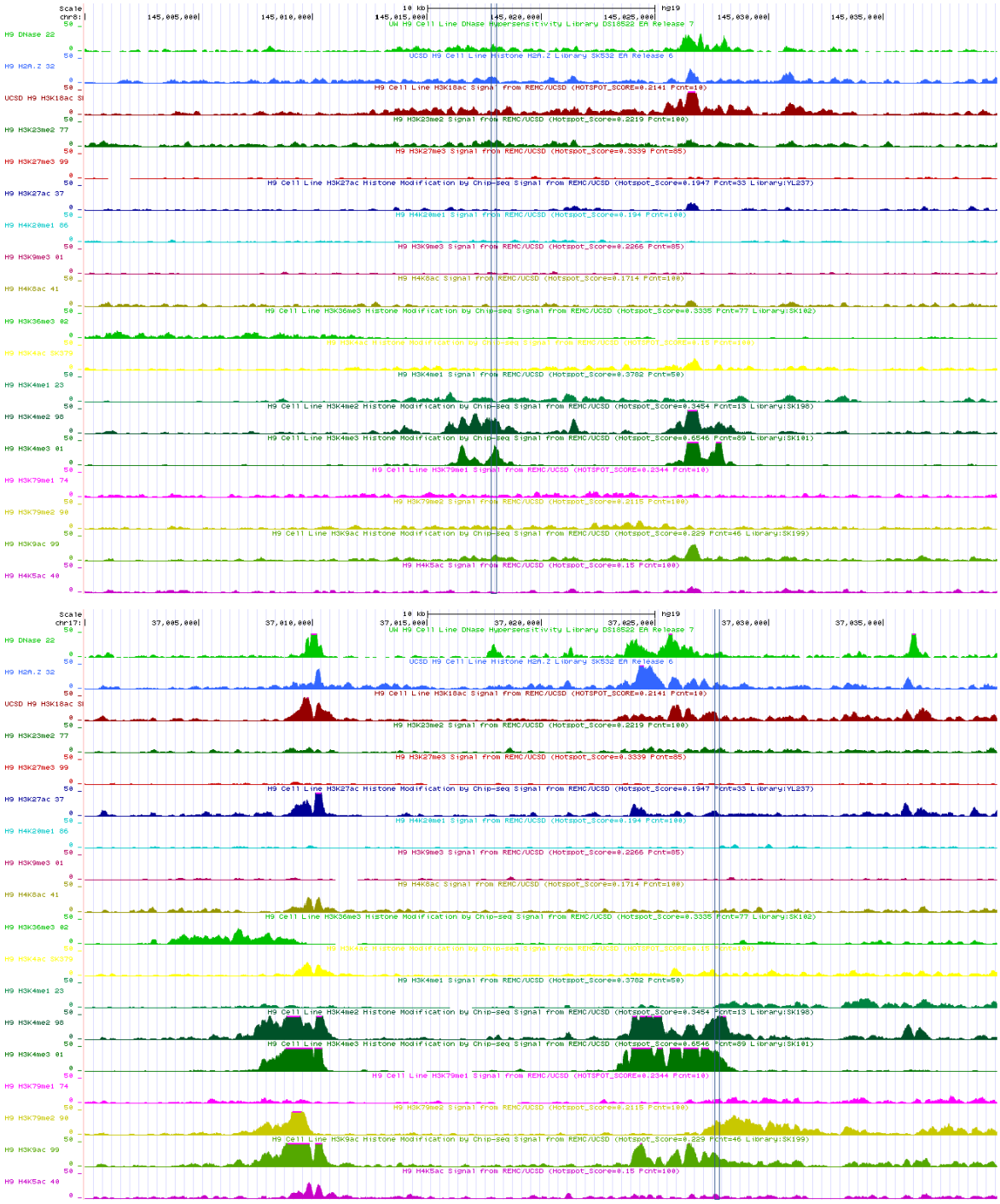


Figure 2.16S. Example genome browser views of the cluster 0 in H9 (spread type)

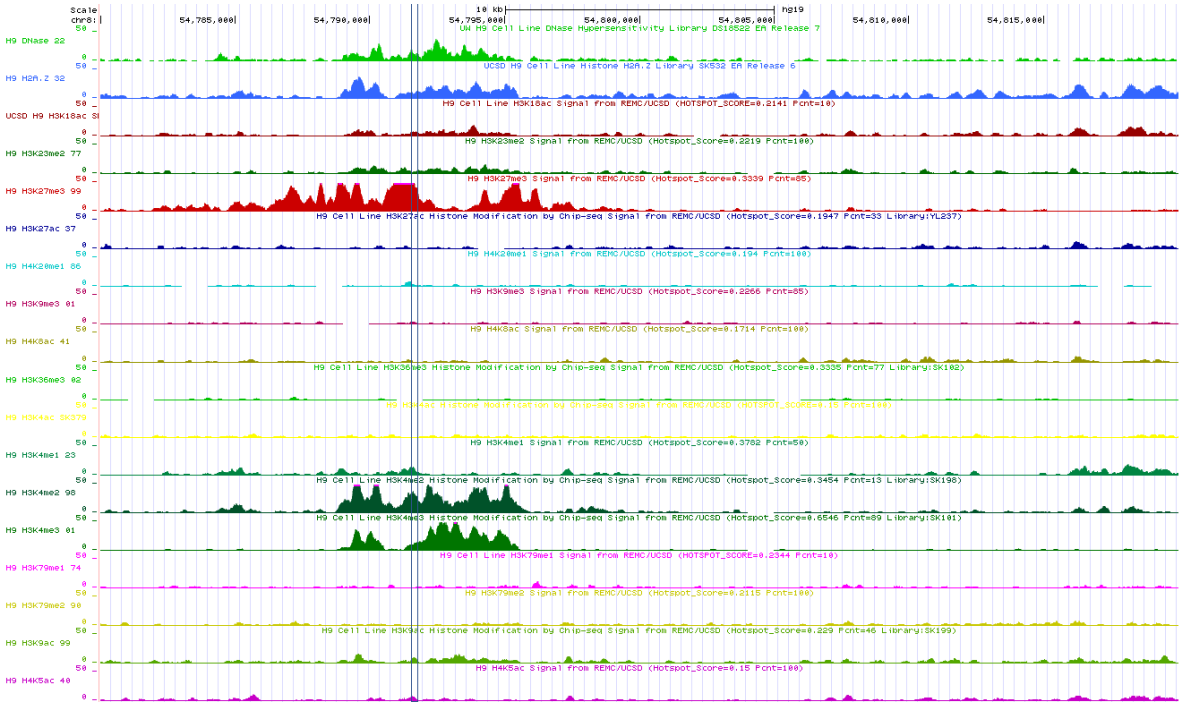
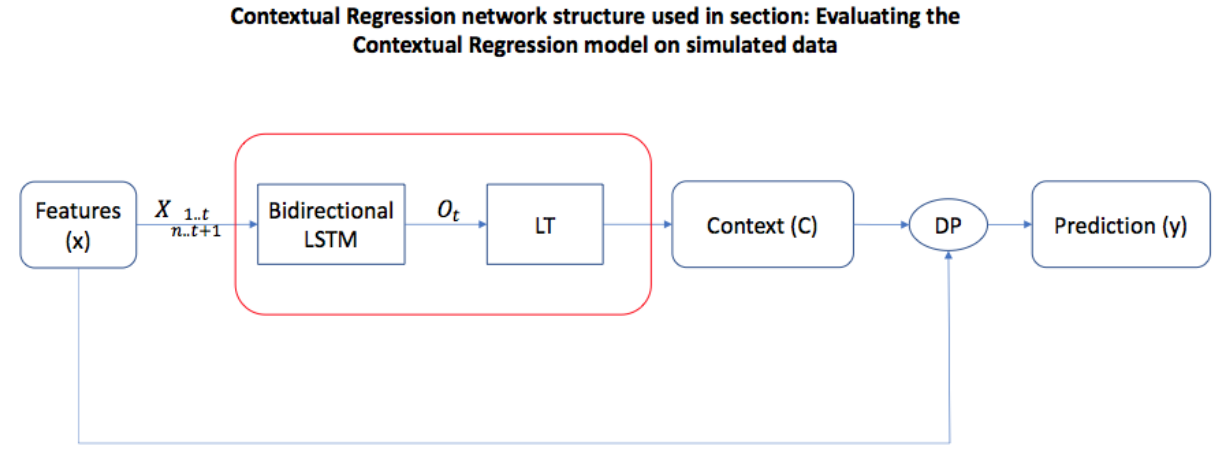


Figure 2.16S. Example genome browser views of the cluster 0 in H9 (spread type)



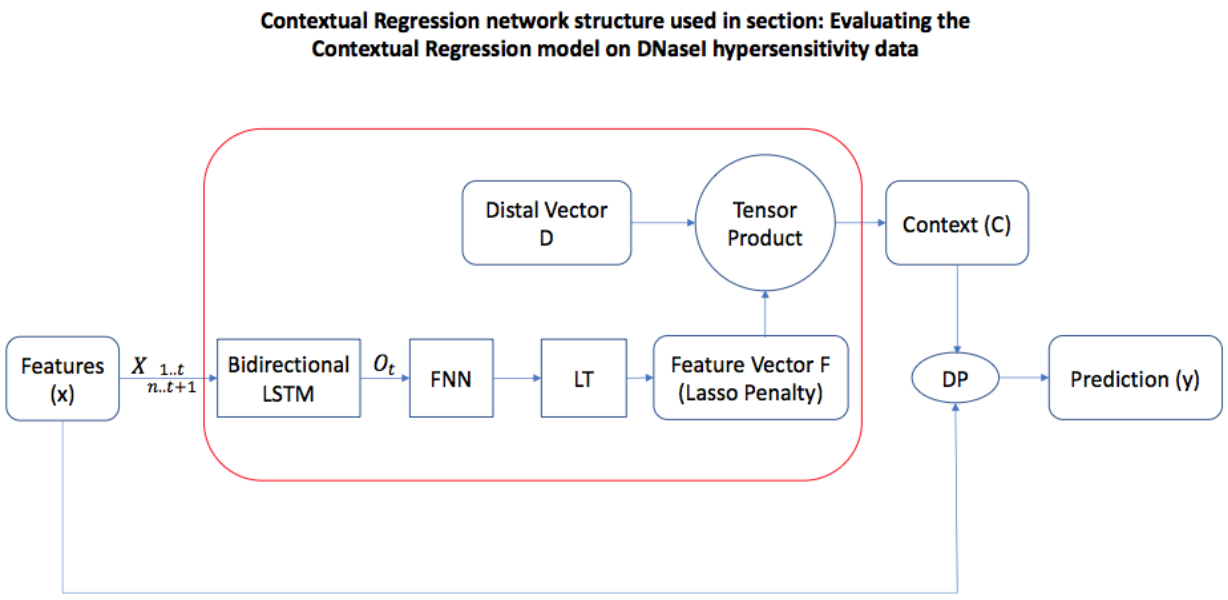


Figure 2.1ST. Architecture of the contextual regression models used in the main article, the embedding network in each model is marked by red squares.

Neural Network structure used in section as performance benchmark: Evaluating the Contextual Regression model on DNaseI hypersensitivity data

Contextual Regression network structure used in section: Discovering Important Histone Mark Patterns in the Open Chromatin Regions

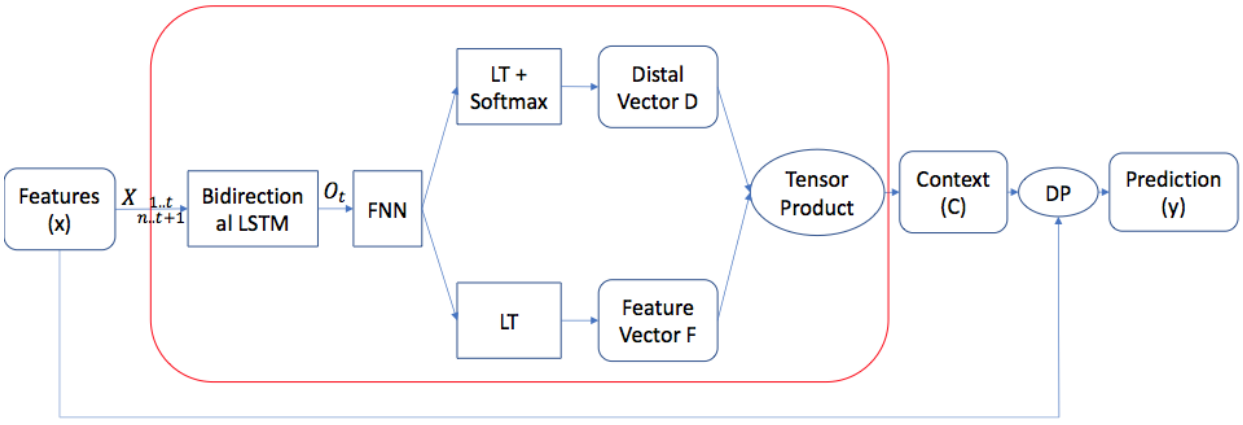


Figure 2.1ST. Architecture of the contextual regression models used in the main article, the embedding network in each model is marked by red squares.

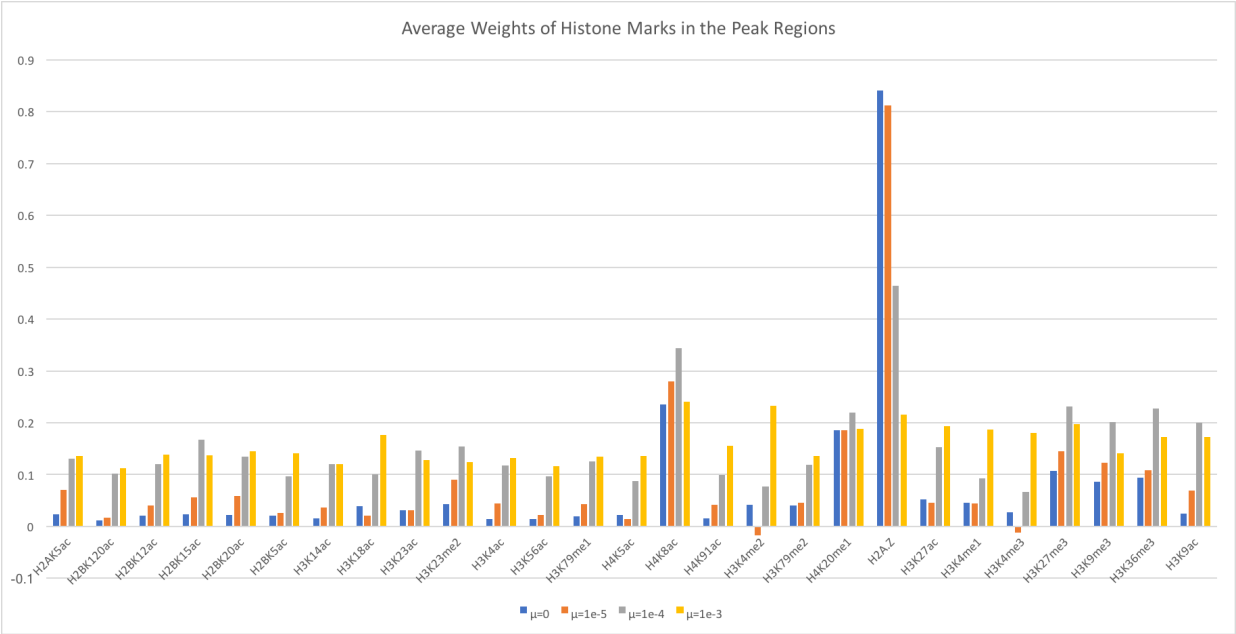


Figure 2.2ST. Average normalized contribution of histone marks to open chromatin in the peak regions under 4 difference value of utopia penalty parameter (μ)

Table 2.1S. Prediction and rule extraction accuracy of contextual regression on a simulated dataset under different levels of noises

Testing set performance	noise level ±0%	noise level ±10%	noise level ±20%	noise level ±40%	noise level ±80%
Average relative prediction error (expected error)	0.9% (0.0%)	6.0% (5.7%)	11.7% (11.5%)	22.7% (22.5%)	42.0% (41.9%)
Average RMSCD	0.024	0.022	0.022	0.026	0.026

Table 2.2S. Prediction and rule extraction accuracy of contextual regression on a simulated dataset under very high noise levels

Testing set performance	noise level ±200%	noise level ±500%	noise level ±1000%
Average relative prediction error (expected error)	75.9% (75.6%)	94.7% (94.5%)	98.5% (98.5%)
Average RMSCD	0.032	0.050	0.080

Table 2.3S. Prediction accuracy of the model used for histone mark pattern discovery

	Pearson Correlation	Match 1	Catch 1 Obs	Catch 1 Imp
Training Set	0.883	68.3%	93.9%	97.0%
Testing Set	0.810	60.9%	89.3%	91.0%

Table 2.4S. Percentage of neighborhoods of regions in H1-cluster 4 (H3K27me3 pattern) that contain TX (transcription) and CDS (coding sequence) start and end sites

H1-Cluster 4	TX start	TX end	CDS start	CDS end
in 1 kbp	38.4%	16.0%	33.1%	21.7%
in 3 kbp	64.4%	44.0%	61.2%	47.5%
in 5 kbp	78.3%	60.3%	76.5%	64.2%
in 7 kbp	84.6%	69.4%	83.5%	70.7%
in 10 kbp	92.3%	77.8%	91.3%	77.9%

Table 2.5S. Motif enrichment profile of the two H3K27ac dominant clusters in H9 (cluster 8 and 19)

H9-Cluster 19	Frequency Diff	p-value	Frequency in Cluster	Background Frequency
PO5F1	15.7%	2.17e-119	83.0%	67.3%
SOX2	14.9%	2.29e-103	79.4%	64.5%
PO2F1/2	14.3%	1.56e-86	63.2%	49.0%
NFAC2	14.1%	9.35e-85	66.9%	52.8%
PO3F2	13.3%	1.05e-76	67.6%	54.3%
NKX31	13.1%	5.92e-78	74.8%	61.6%
PIT1	12.6%	6.23e-72	73.4%	60.8%

H9-Cluster 8	Frequency Diff	p-value	Frequency in Cluster	Background Frequency
PO5F1	15.7%	3.16e-119	83.0%	67.3%
SOX2	14.6%	3.21e-99	79.1%	64,5%
PO2F1/2	13.1%	2.51e-73	62.1%	49.0%
PIT1	13.0%	1.55e-76	73.8%	60.8%
PO3F1	12.4%	8.89e-66	63.0%	50.7%
PO3F2	12.1%	2.45e-63	66.4%	54.3%
NFAC2	12.0%	6.17e-62	64.8%	52.8%

Table 2.6S. Motif enrichment profile of cluster 1 and 6 in H1, H3K27ac dominant clusters

H1-Cluster 1	Frequency Diff	p-value	Frequency in Cluster	Background Frequency
SOX2	7.6%	1.94e-29	71.0%	63.4%
PO5F1	7.0%	4.83e-26	72.9%	65.9%
PO2F1/2	6.0%	7.25e-18	53.8%	47.8%
SOX9	5.3%	1.08e-14	62.5%	57.2%
PO3F1	5.0%	4.53e-13	54.6%	49.6%

H1-Cluster 6	Frequency Diff	p-value	Frequency in Cluster	Background Frequency
SOX2	7.8%	4.59e-30	71.3%	63.4%
PO5F1	7.5%	3.07e-28	73.4%	65.9%
PO2F1/2	5.1%	8.28e-13	52.9%	47.8%
SOX9	5.3%	4.05e-14	62.5%	57.2%
PO3F1	4.5%	1.66e-10	54.1%	49.6%

Table 2.7S. Motif enrichment profile of cluster 4 in H9 (H3K27me3 pattern), pval of 0 values are caused by number underflow since the pval is smaller than machine accuracy

H9-Cluster 4	Frequency Diff	p-value	Frequency in Cluster	Background Frequency
MBD2	26.8%	0.0	89.7%	62.9%
E2F1	25.6%	0.0	83.2%	57.6%
E2F4	24.3%	0.0	63.5%	39.2%
NRF1	23.8%	0.0	75.9%	52.0%
E2F3	21.5%	0.0	37.1%	15.6%
E2F2	15.0%	1.75e-258	31.5%	16.5%

Table 2.1ST. The average cosine similarity of feature contribution vectors among 5 runs

Weight_similarity	DHS	RNAseq
Feature weight (whole dataset)	0.262	0.017
Feature weight (flat regions)	0.310	0.025
Feature weight (peak regions)	0.991	0.960
Distal weight	0.963	0.918

Table 2.2ST. Feature mark contribution for DNase and RNAseq prediction task, ranked from high to low

Avg_weight	DNase
H2A.Z	0.874
H4K8ac	0.186
H4K20me1	0.162
H3K27me3	0.075
H3K36me3	0.058
H3K27ac	0.043
H3K9me3	0.030
H3K4me1	0.026
H3K18ac	0.024
H4K91ac	0.024
H3K9ac	0.022
H3K79me1	0.020
H3K4ac	0.019
H3K79me2	0.019
H2AK5ac	0.019
H2BK12ac	0.018
H3K23ac	0.017
H3K14ac	0.014
H4K5ac	0.014
H2BK15ac	0.013
H2BK5ac	0.013
H2BK20ac	0.011
H2BK120ac	0.010
H3K4me2	0.008
H3K56ac	0.007
H3K4me3	0.005
H3K23me2	0.003

Avg_weight	RNAseq
H3K36me3	0.732
H4K8ac	0.387
H2A.Z	0.218
H4K20me1	0.176
H3K4me3	0.082
H3K4me2	0.073
H3K9ac	0.059
H3K27ac	0.055
H3K4me1	0.049
H3K18ac	0.033
H2BK15ac	0.027
H4K91ac	0.025
H2BK12ac	0.023
H3K23ac	0.023
H2AK5ac	0.021
H4K5ac	0.021
H2BK20ac	0.021
H2BK5ac	0.021
H3K79me2	0.018
H3K23me2	0.017
H3K79me1	0.017
H3K9me3	0.013
H3K56ac	0.013
H2BK120ac	0.012
H3K14ac	0.008
H3K4ac	0.007
H3K27me3	0.004

Table 2.3ST. Average signal strength of histone marks in the DNase peak region (contain location with logpval > 3)

Mark in Region	Avg signal
H2A.Z	0.436
H4K8ac	0.431
H4K20me1	0.371
H3K27ac	0.363
H3K18ac	0.351
H3K36me3	0.342
H3K4me1	0.327
H3K9me3	0.321
H3K9ac	0.315
H3K4me2	0.31
H3K27me3	0.31
H4K91ac	0.292
H3K79me2	0.292
H2BK20ac	0.287
H2BK15ac	0.283
H3K79me1	0.282
H2BK5ac	0.28
H2BK12ac	0.278
H3K4ac	0.276
H2AK5ac	0.273
H3K14ac	0.273
H3K23ac	0.269
H4K5ac	0.267
H2BK120ac	0.254
H3K23me2	0.251
H3K56ac	0.238
H3K4me3	0.237

Table 2.4ST. Prediction accuracy of contextual regression using selected histone marks

	Correlation	Match1	Catch1obs	Catch1imp
Dnase_all	0.817	60.6%	89.9%	89.7%
Dnase_top_18	0.783	56.6%	86.2%	88.1%
Dnase_top_10	0.745	53.0%	81.8%	85.8%
RNAseq_all	0.622	38.4%	77.2%	76.8%
RNAseq_top_18	0.649	35.1%	75.1%	75.4%
RNAseq_top_10	0.587	34.1%	72.8%	73.1%

2.12 References

1. Seber, G. A. F. & Lee, A. J. *Linear Regression Analysis*. (John Wiley & Sons, 2012).
2. Hosmer, D. W., Jr., Lemeshow, S. & Sturdivant, R. X. *Applied Logistic Regression*. (John Wiley & Sons, 2013).
3. Support Vector Machines and Regularization Networks. in *Template Matching Techniques in Computer Vision* 237–262
4. LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521,** 436–444 (2015).
5. Rokach, L. & Maimon, O. *Data Mining with Decision Trees: Theory and Applications*. (World Scientific, 2014).
6. Alipanahi, B., Delong, A., Weirauch, M. T. & Frey, B. J. Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nat. Biotechnol.* **33,** 831–838 (2015).
7. Zhou, J. & Troyanskaya, O. G. Predicting effects of noncoding variants with deep learning-based sequence model. *Nat. Methods* **12,** 931–934 (2015).
8. Wu, X., Zhang - arXiv preprint arXiv:1611.04135, X. & 2016. Automated Inference on Criminality using Face Images. *arxiv.org* (1611).
9. Kosinski, M. & Wang, Y. Deep neural networks are more accurate than humans at detecting sexual orientation from facial images. (2017).
10. Zeiler, M. D., Fergus - European conference on computer vision, R. & 2014. Visualizing and understanding convolutional networks. *Springer* (2014).
11. Zhang, Q., Wu, Y. N., Zhu - arXiv preprint arXiv:1710.00935, S. C. & 2017. Interpretable Convolutional Neural Networks. *arxiv.org* (1710).

12. Cybenko, G. Approximation by superpositions of a sigmoidal function. *Math. Control Signals Systems* **5,** 455–455 (1992).

13. Hochreiter, S. & Schmidhuber, J. *Long Short Term Memory*. (1995).

14. Shrikumar, A., Greenside, P., Shcherbina, A. & Kundaje, A. Not Just a Black Box: Learning Important Features Through Propagating Activation Differences. *arXiv [cs.LG]* (2016).

15. Ribeiro, M., Singh, S. & Guestrin, C. 'Why Should I Trust You?': Explaining the Predictions of Any Classifier. in *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations* (2016). doi:10.18653/v1/n16-3020

16. Lundberg, S. & Lee, S.-I. A unified approach to interpreting model predictions. *arXiv [cs.AI]* (2017). at <http://arxiv.org/abs/1705.07874>

17. Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R. & Bengio, Y. Show, Attend and Tell: Neural Image Caption Generation with Visual Attention. in *International Conference on Machine Learning* 2048–2057 (2015).

18. Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S. & Dean, J. Distributed Representations of Words and Phrases and their Compositionality. in *Advances in Neural Information Processing Systems 26* (eds. Burges, C. J. C., Bottou, L., Welling, M., Ghahramani, Z. & Weinberger, K. Q.) 3111–3119 (Curran Associates, Inc., 2013).

19. Roadmap Epigenomics Consortium, Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., Heravi-Moussavi, A., Kheradpour, P., Zhang, Z., Wang, J., Ziller, M. J., Amin, V., Whitaker, J. W., Schultz, M. D., Ward, L. D., Sarkar, A., Quon, G., Sandstrom, R. S., Eaton, M. L., Wu, Y.-C., Pfenning, A. R., Wang, X., Claussnitzer, M., Liu, Y., Coarfa, C., Harris,

R. A., Shoresh, N., Epstein, C. B., Gjoneska, E., Leung, D., Xie, W., Hawkins, R. D., Lister, R., Hong, C., Gascard, P., Mungall, A. J., Moore, R., Chuah, E., Tam, A., Canfield, T. K., Hansen, R. S., Kaul, R., Sabo, P. J., Bansal, M. S., Carles, A., Dixon, J. R., Farh, K.-H., Feizi, S., Karlic, R., Kim, A.-R., Kulkarni, A., Li, D., Lowdon, R., Elliott, G., Mercer, T. R., Neph, S. J., Onuchic, V., Polak, P., Rajagopal, N., Ray, P., Sallari, R. C., Siebenthall, K. T., Sinnott-Armstrong, N. A., Stevens, M., Thurman, R. E., Wu, J., Zhang, B., Zhou, X., Beaudet, A. E., Boyer, L. A., De Jager, P. L., Farnham, P. J., Fisher, S. J., Haussler, D., Jones, S. J. M., Li, W., Marra, M. A., McManus, M. T., Sunyaev, S., Thomson, J. A., Tlsty, T. D., Tsai, L.-H., Wang, W., Waterland, R. A., Zhang, M. Q., Chadwick, L. H., Bernstein, B. E., Costello, J. F., Ecker, J. R., Hirst, M., Meissner, A., Milosavljevic, A., Ren, B., Stamatoyannopoulos, J. A., Wang, T. & Kellis, M. Integrative analysis of 111 reference human epigenomes. *Nature* **518,** 317–330 (2015).

20. Mastoridis, T., Baudrenghien, P., Butterworth, A., Molendijk, J., Rivetta, C. & Fox, J. D. Radio frequency noise effects on the CERN Large Hadron Collider beam diffusion. *Physical Review Special Topics - Accelerators and Beams* **14,** (2011).
21. Zakareishvili, T. Muon Signals at a Low Signal-to-Noise Ratio Environment. (2017).
22. Bollen, K., Cacioppo, J. T., Kaplan, R. M. & Krosnick, J. A. Social, behavioral, and economic sciences perspectives on robust and reliable science: Report of the Subcommittee on Replicability in Science, Advisory …. *from the National Science ...* (2015).
23. Collaboration, O. S. Estimating the reproducibility of psychological science. *Science* **349,** aac4716 (2015).
24. Reiter, M., Kirchner, B., Müller, H., Holzhauer, C., Mann, W. & Pfaffl, M. W.

Quantification noise in single cell experiments. *Nucleic Acids Res.* **39,** e124 (2011).

25. Brennecke, P., Anders, S., Kim, J. K., Kołodziejczyk, A. A., Zhang, X., Proserpio, V., Baying, B., Benes, V., Teichmann, S. A., Marioni, J. C. & Heisler, M. G. Accounting for technical noise in single-cell RNA-seq experiments. *Nat. Methods* **10,** 1093–1095 (2013).
26. Grün, D., Kester, L. & van Oudenaarden, A. Validation of noise models for single-cell transcriptomics. *Nat. Methods* **11,** 637–640 (2014).
27. Tibshirani, R. Regression shrinkage and selection via the lasso: a retrospective. *J. R. Stat. Soc. Series B Stat. Methodol.* **73,** 273–282 (2011).
28. Hashimoto, T., Sherwood, R. I., Kang, D. D., Rajagopal, N., Barkal, A. A., Zeng, H., Emons, B. J. M., Srinivasan, S., Jaakkola, T. & Gifford, D. K. A synergistic DNA logic predicts genome-wide chromatin accessibility. *Genome Res.* **26,** 1430–1440 (2016).
29. Villani, A.-C., Satija, R., Reynolds, G., Sarkizova, S., Shekhar, K., Fletcher, J., Griesbeck, M., Butler, A., Zheng, S., Lazo, S., Jardine, L., Dixon, D., Stephenson, E., Nilsson, E., Grundberg, I., McDonald, D., Filby, A., Li, W., De Jager, P. L., Rozenblatt-Rosen, O., Lane, A. A., Haniffa, M., Regev, A. & Hacohen, N. Single-cell RNA-seq reveals new types of human blood dendritic cells, monocytes, and progenitors. *Science* **356,** (2017).
30. Ernst, J. & Kellis, M. Large-scale imputation of epigenomic datasets for systematic annotation of diverse human tissues. *Nat. Biotechnol.* **33,** 364–376 (2015).
31. Bishop, C. M. *Pattern Recognition and Machine Learning*. (2013).
32. Bernstein, I. H., Garbin, C. P. & Teng, G. K. Classification Methods—Part 2. Methods of Assignment. in *Applied Multivariate Analysis* 276–314 (1988).
33. Karlić, R., Chung, H.-R., Lasserre, J., Vlahoviček, K. & Vingron, M. Histone modification

levels are predictive for gene expression. *Proc. Natl. Acad. Sci. U. S. A.* **107,** 2926–2931 (2010).

34. Rintisch, C., Heinig, M., Bauerfeind, A., Schafer, S., Mieth, C., Patone, G., Hummel, O., Chen, W., Cook, S., Cuppen, E., Colomé-Tatché, M., Johannes, F., Jansen, R. C., Neil, H., Werner, M., Pravenec, M., Vingron, M. & Hubner, N. Natural variation of histone modification and its impact on gene expression in the rat genome. *Genome Res.* **24,** 942–953 (2014).

35. Xu, Y., Ayrapetov, M. K., Xu, C., Gursoy-Yuzugullu, O., Hu, Y. & Price, B. D. Histone H2A.Z Controls a Critical Chromatin Remodeling Step Required for DNA Double-Strand Break Repair. *Mol. Cell* **48,** 723–733 (2012).

36. Lombardi, L. *Maintenance of Open Chromatin States by Histone H3 Eviction and H2A.Z*. (2011).

37. Bannister, A. J. & Kouzarides, T. Regulation of chromatin by histone modifications. *Cell Res.* **21,** 381–395 (2011).

38. Huang, J., Marco, E., Pinello, L. & Yuan, G.-C. Predicting chromatin organization using histone marks. *Genome Biol.* **16,** 162 (2015).

39. Gomez, N. C., Hepperla, A. J., Dumitru, R., Simon, J. M., Fang, F. & Davis, I. J. Widespread Chromatin Accessibility at Repetitive Elements Links Stem Cells with Human Cancer. *Cell Rep.* **17,** 1607–1620 (2016).

40. Shu, W., Chen, H., Bo, X. & Wang, S. Genome-wide analysis of the relationships between DNaseI HS, histone modifications and gene expression reveals distinct modes of chromatin domains. *Nucleic Acids Res.* **39,** 7428–7443 (2011).

41. Zhang, W., Zhang, T., Wu, Y. & Jiang, J. Open Chromatin in Plant Genomes. *Cytogenet. Genome Res.* **143,** 18–27 (2014).

42. Bernstein, B. E., Mikkelsen, T. S., Xie, X., Kamal, M., Huebert, D. J., Cuff, J., Fry, B., Meissner, A., Wernig, M., Plath, K., Jaenisch, R., Wagschal, A., Feil, R., Schreiber, S. L. & Lander, E. S. A bivalent chromatin structure marks key developmental genes in embryonic stem cells. *Cell* **125,** 315–326 (2006).

43. Chen, T. & Dent, S. Y. R. Chromatin modifiers and remodellers: regulators of cellular differentiation. *Nat. Rev. Genet.* **15,** 93–106 (2013).

44. Takahashi, K., Tanabe, K., Ohnuki, M., Narita, M., Ichisaka, T., Tomoda, K. & Yamanaka, S. Induction of pluripotent stem cells from adult human fibroblasts by defined factors. *Cell* **131,** 861–872 (2007).

45. Segil, N., Roberts, S. B. & Heintz, N. Mitotic phosphorylation of the Oct-1 homeodomain and regulation of Oct-1 DNA binding activity. *Science* **254,** 1814–1816 (1991).

46. Wong, W. T., Matrone, G., Tian, X., Tomoiaga, S. A., Au, K. F., Meng, S., Yamazoe, S., Sieveking, D., Chen, K., Burns, D. M., Chen, J. K., Blau, H. M. & Cooke, J. P. Discovery of novel determinants of endothelial lineage using chimeric heterokaryons. *Elife* **6,** (2017).

47. Hendrich, B. & Bird, A. Identification and Characterization of a Family of Mammalian Methyl-CpG Binding Proteins. *Mol. Cell. Biol.* **18,** 6538–6547 (1998).

48. Ng, H. H., Zhang, Y., Hendrich, B., Johnson, C. A., Turner, B. M., Erdjument-Bromage, H., Tempst, P., Reinberg, D. & Bird, A. MBD2 is a transcriptional repressor belonging to the MeCP1 histone deacetylase complex. *Nat. Genet.* **23,** 58–61 (1999).

49. Fujita, H., Fujii, R., Aratani, S., Amano, T., Fukamizu, A. & Nakajima, T. Antithetic effects

of MBD2a on gene regulation. *Mol. Cell. Biol.* **23,** 2645–2657 (2003).

50. Cramer, J. M., Scarsdale, J. N., Walavalkar, N. M., Buchwald, W. A., Ginder, G. D. & Williams, D. C., Jr. Probing the dynamic distribution of bound states for methylcytosine-binding domains on DNA. *J. Biol. Chem.* **289,** 1294–1302 (2014).
51. Gaubatz, S., Lindeman, G. J., Ishida, S., Jakoi, L., Nevins, J. R., Livingston, D. M. & Rempel, R. E. E2F4 and E2F5 Play an Essential Role in Pocket Protein–Mediated G1 Control. *Mol. Cell* **6,** 729–735 (2000).
52. Tsai, S.-Y., Opavsky, R., Sharma, N., Wu, L., Naidu, S., Nolan, E., Feria-Arias, E., Timmers, C., Opavska, J., de Bruin, A., Chong, J.-L., Trikha, P., Fernandez, S. A., Stromberg, P., Rosol, T. J. & Leone, G. Mouse development with a single E2F activator. *Nature* **454,** 1137–1141 (2008).
53. Viollet, B., Lefrançois-Martinez, A.-M., Henrion, A., Kahn, A., Raymondjean, M. & Martinez, A. Immunochemical characterization and transacting properties of upstream stimulatory factor isoforms. *J. Biol. Chem.* **271,** 1405–1415 (1996).
54. Meyer, N. & Penn, L. Z. Reflecting on 25 years with MYC. *Nat. Rev. Cancer* **8,** 976–990 (2008).
55. Chou, W.-Y., Ho, C.-L., Tseng, M.-L., Liu, S.-T., Yen, L.-C. & Huang, S.-M. Human Spot 14 protein is a p53-dependent transcriptional coactivator via the recruitment of thyroid receptor and Zac1. *Int. J. Biochem. Cell Biol.* **40,** 1826–1834 (2008).
56. Zhu, L., Zhou, G., Poole, S. & Belmont, J. W. Characterization of the interactions of human ZIC3 mutants with GLI3. *Hum. Mutat.* **29,** 99–105 (2008).
57. Tsukada, J., Yoshida, Y., Kominato, Y. & Auron, P. E. The CCAAT/enhancer (C/EBP)

family of basic-leucine zipper (bZIP) transcription factors is a multifaceted highly-regulated system for gene regulation. *Cytokine* **54,** 6–19 (2011).

58. Liu, C., Liu, Y.-C., Huang, H.-D. & Wang, W. Biogenesis mechanisms of circular RNA can be categorized through feature extraction of a machine learning model. *Bioinformatics* **35**, 4867–4870 (2019).
59. Kingma, D., Ba - arXiv preprint arXiv:1412.6980, J. & 2014. Adam: A method for stochastic optimization. *arxiv.org* (1412).
60. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489,** 57–74 (2012).
61. Bannister, A. J. & Kouzarides, T. Regulation of chromatin by histone modifications. *Cell Res.* **21,** 381–395 (2011).
62. Lombardi, L. *Maintenance of Open Chromatin States by Histone H3 Eviction and H2A.Z.* (2011).
63. Huang, J., Marco, E., Pinello, L. & Yuan, G.-C. Predicting chromatin organization using histone marks. *Genome Biol.* **16,** 162 (2015).
64. High-Resolution Profiling of Histone Methylations in the Human Genome. *Cell* **129,** 823–837 (2007).
65. Schwartz, S., Meshorer, E. & Ast, G. Chromatin organization marks exon-intron structure. *Nat. Struct. Mol. Biol.* **16,** 990–995 (2009).
66. Karlić, R., Chung, H.-R., Lasserre, J., Vlahoviček, K. & Vingron, M. Histone modification levels are predictive for gene expression. *Proc. Natl. Acad. Sci. U. S. A.* **107,** 2926–2931 (2010).

67. Pearson, K. *On Lines and Planes of Closest Fit to Systems of Points in Space*. (1901).

68. Atkinson, K. E. *AN INTRODUCTION TO NUMERICAL ANALYSIS, 2ND ED*. (John Wiley & Sons, 2008).

CHAPTER 3: BIOGENESIS MECHANISMS OF CIRCULAR RNA CAN BE CATEGORIZED THROUGH FEATURE EXTRACTION OF A MACHINE LEARNING MODEL

3.1 Abstract

Motivation: In recent years, multiple circular RNAs biogenesis mechanisms have been discovered. While each reported mechanism has been experimentally verified in different circular RNAs, no single biogenesis mechanism has been proposed that can universally explain the biogenesis of all tens of thousands of discovered circular RNAs. Under the hypothesis that human circular RNAs can be categorized according to different biogenesis mechanisms, we designed a contextual regression model trained to predict the formation of circular RNA from a random genomic locus on human genome, with potential biogenesis factors of circular RNA as the features of the training data.

Results: After achieving high prediction accuracy, we found through the feature extraction technique that the examined human circular RNAs can be categorized into seven subgroups, according to the presence of the following sequence features: RNA editing sites, simple repeat sequences, self-chains, RNA binding protein binding sites and CpG islands within the flanking regions of the circular RNA back-spliced junction sites. These results support all of the previously reported biogenesis mechanisms of circRNA and solidify the idea that multiple biogenesis mechanisms co-exist for different subset of human circRNAs. Furthermore, we uncover a potential new links between circRNA biogenesis and flanking CpG island. We have

also identified RNA binding proteins putatively correlated with circRNA biogenesis.

3.2 Introduction

Circular RNAs (circRNAs) represent an emerging type of regulatory RNA with 3' and 5' ends covalently join together as "back-spliced junctions". Circular RNAs have been reported to have multiple functions. Beside serving as miRNA sponges, many circRNAs are important for brain function, synaptic plasticity[1] and fetal development[2]. Cell free circRNAs are found stable in saliva[3] as well as exosomes[4], which makes circRNA a promising diagnostic biomarker. Most circRNAs originated from circularization of coding gene exons, which leads to the hypothesis that circRNA biogenesis competes with pre-mRNA splicing[5]. Literature evidence suggests that the biogenesis of each circRNA subset may likely be regulated by different mechanisms[6-12], which supports the existence of multiple subclasses of circRNAs and each with specific roles.

In this study, we aim to decipher the relationship between the sequence features and circRNA subclasses. To this end, we have developed a computational model to predict whether a genomic locus would generate circRNA based on the presence of the following sequence features within the flanking regions of the circular RNA back-spliced junction sites: CpG islands, enhancers, RNA Binding Protein (RBP) binding sites, simple repeats, RNA editing sites and DNA self-chains. These features were selected based on the hypothesized circRNA biogenesis mechanisms[6-12] as well as sequence features that can potentially participate in biogenesis of non-coding RNAs including CpG islands[13] and enhancers regions[14-15]. We have designed a contextual regression model[16] that successfully distinguishes the circRNA back-spliced junction sites defined by transcriptome sequencing from randomly selected human genome loci in an average 72.6 % accuracy. Using the feature extraction technique in the contextual regression

model[17], we found that the examined 21,427 circRNAs can be categorized into seven groups based on the biogenesis contributing factors. Our analysis supports that multiple biogenesis mechanisms co-exist for different subset of human circRNAs. In particular, we found 79 RNA binding proteins were identified to be significantly correlated to circRNA biogenesis. Interestingly, we uncovered a potential new link between circRNA biogenesis and flanking CpG islands, which suggests the potential correlation between DNA methylation and circRNA biogenesis.

3.3 Methods

The data analysis process of this research is summarized in Figure 3.1S. First, 55,689 human circRNAs back-spliced junction sites were collected from the database CircNet[18] as positive training data, and equal amount of randomly selected locus on HG19 human genome as negative training data. These junction sites and randomly selected locus were then divided into training and testing sets (in a ratio of 7:3) for the contextual regression network model designed to predict whether a randomly selected locus from the human genome would generate circRNA. The features of the training set include whether CpG islands, enhancer regions, RNA Binding Protein (RBP) binding sites, simple repeats, A-to-I RNA editing sites (RNA editing sites for short in the later text) and DNA self-chains present in the upstream and downstream region of the selected locus. After reaching optimum average accuracy, through the application of feature extraction techniques[17], we successfully found that the examined 21,427 circRNAs can be categorized into seven groups based on the presumed biogenesis contributing factors.

The back-spliced junction sites in the positive training set were selected from the database CircNet[18]. The data includes reported human back-spliced junction sites were collected from 22 recent studies[3,7-8,11,19-32] and 465 human transcriptome sequencing datasets were collected from NCBI Sequence Read Archive[33]. The back-spliced junction sites in each RNA-seq sample were identified using a circRNA discovery pipeline referred as find_circ[24,34-35]. The criteria defined in the pipeline hence the detected junction sites met the same standards as those in the previous reports, as described in the Memczak et al. 2013 study[24] was applied. Adhering to the suggestion of recent year comparison study[34], we selected the back-spliced junction sites based

on the number of previous peer review reports in which the back-spliced junction sites were reported and the number of samples among the 465 collected samples in which the back-spliced junction sites were found meeting the criteria defined in find_circ[24,34-35]. Only the circRNAs with the sum of these two numbers greater than 3 were considered as positive training data in this study. Locus of the CpG islands, enhancer regions, simple repeats, RNA editing sites and DNA self-chains on HG19 human genome was collected from UCSC Genome Browser[36]. While the locus of the RBP binding sites was collected from the Ray et al. 2013 study[37].

3.4 Results

Using the feature extraction technique in the contextual regression model[16], we found that the examined 21,427 circRNAs can be categorized into seven groups based on the biogenesis contributing factors. Our analysis supports that multiple biogenesis mechanisms co-exist for different subset of human circRNAs. In particular, we found 79 RNA binding proteins were identified to be significantly correlated to circRNA biogenesis. Interestingly, we uncovered a potential new link between circRNA biogenesis and flanking CpG islands, which suggests the potential correlation between DNA methylation and circRNA biogenesis.

An interpretable neural network model to predict circRNAs. We implemented a contextual regression model to predict circRNA using these features. Instead of letting the neural network learn a function that maps features to target values, our method let the neural network learn a function that maps features to local linear models that best predict the target value from the features, thus generating a model that can both achieve state-of-the-art accuracy like a deep neural network while giving human interpretable quantification of feature contribution (Figure 3.1 A). As summarized in Figure 3.1S C, in this contextual regression model, we used a residual neural network[38] that is composed of 3 layers of FNN (feed forward neural network, Figure 3.1 B) as the embedding function to generate the linear models. Batch normalization[39] is applied in the input layer to reduce the variance between input batches and make the training process more stable. The output of the embedding function is then dot-producted with the features and fed into a logistic function to output the prediction result. The FNN model is implemented as an operation that maps a vector x to s(Ax+b) where A is an matrix, b is an vector and s is the

activation function. Both A and b are first initialized and then trained with tensorflow AdamOptimizer[40]. As for the parameter setting, all three layers of the FNN have 10 hidden units and tanh as their activation function, batch size is set to 50, max gradient norm to 10, learning rate to 0.0001. The weight matrix of each neural network is initialized with the tensorflow tf.truncated_normal function with standard deviation of 0.05 to prevent vanishing gradient problem. The bias term in each layer is initialized all to value 0. We used cross-entropy as the loss function during training. To make the feature contribution easily interpretable, we applied a Lasso penalty in the form of L1 regularization on the context weight with penalty coefficient 0.0001.

Prediction and the feature selection results. As summarized in Table 3.1S, the designed contextual regression model successfully distinguish circRNA back-spliced junction sites defined by transcriptome sequencing data from randomly selected human genome loci, in an average 72.6% accuracy and the area under curve of ROC curve 0.801 (Figure 3.2S). To extract the informative features, we selected the run with the highest accuracy (Run #5) from 10 runs. The difference between the accuracy on the training and testing sets were negligible (72.5% vs 72.8%), which suggested no overfitting. Therefore, we pooled together the training and testing sets in the feature analysis. We selected the most confidently predicted 21,427 circRNAs that have confident scores > 0.7 to evaluate the contribution of each feature. The weighted feature contribution (WFC) was then obtained from the model output. The features for each circRNA genesis mechanism (except “flanking short ALU repeats” (both_have_Alu) and “binding sites of single RBP” (if_same_RBP) since they only have 1 value for the whole region) were pooled into short (within 1,000bp radius from the circRNA) and long range (1,000bp to 2,000bp from the

circRNA) by summing the weighted feature contribution in each category for better data visualization. This resulted in 14 total feature contributions for each circular RNA data point (CpG short range, CpG long range, enhancer short range, enhancer long range, RBP short range, RBP long range, repeats short range, repeats long range, RNA editing short range, RNA editing long range, self-chain short range, self-chain long range, if head and tail both have Alu, if head and tail have the same RBP) that could be treated as a vector with 14 elements. Then, each of these vectors was normalized to the unit length and PCA was applied to the whole data point collection. The top 10 principal components were extracted which explained 98.9% of the total variance. Then a K-mean clustering was applied to separate the processed data into subclasses. Multiple values of k were experimented and k=7 was chosen for being the largest k that ensured all cosine similarities between cluster centers are less than 0.5. The significantly contributing features were plotted in the heat map format and confirmed in the original feature data.

As a result, we found that these circRNAs could be clustered into 7 different categories according to their biogenesis factors (Figure 3.2A). To validate that the features with high contribution scores are enriched in each cluster, we calculated the percent enrichment of each feature in each cluster (Figure 3.2B). The enrichment plot supported our clustering result.

In cluster 0, RNA editing sites occurrence within 1000 nucleotides upstream or downstream of the circRNA locus was considered to be the most important factor for the biogenesis of 4,576 circRNAs. For the other 4,585 circRNAs in cluster 1, the appearance of RNA editing site within range of 1000 to 2000 nucleotides upstream or downstream of circRNA was considered as the most important biogenesis factor. Similarly, in cluster 2, existence of RNA

binding protein binding sites within short, which means within 1000 nucleotides upstream or downstream of the circRNA locus, was considered as the most important biogenesis factor for the 7,503 circRNAs. For the 2,342 circRNAs in the cluster 4, occurrence in the long range, which means within range of 1000 to 2000 nucleotides upstream or downstream of the circRNA locus, of RNA binding protein binding sites was suggested as the main biogenesis factor. Short repeat sequence flanking circRNA locus was clustered as the main factor for the 1,344 circRNAs in cluster 3. Biogenesis of 620 circRNAs was linked to flanking CpG islands, while 457 circRNAs' biogenesis mechanism appeared to relate to the flanking DNA self-chains. The amount of these seven clusters is summarized in Figure 3.3S. Through examining the input data that forms cluster 2 and cluster 4, which were associated with RNA binding protein (RBP), we identified 79 RBPs presumably participate in the biogenesis of these 9,845 circRNAs. These 79 RBPs (available in Additional file 4) appear both in the short range (within 1000 nucleotides) in cluster 2 and long range (from 1000 nucleotides to 2000 nucleotides) in cluster 4 consistently. Among these RBPs, binding motifs of SRSF1, PUM1, SF3B4, ELAVL1, HNRNPA1, CSDA, SNRPA, RBFOX1 and PABPC1 were found appear in upstream or downstream of thousands of circRNAs in both clusters, as summarized in Table 3.2S.

Hence based on the result of this study, circRNAs can be categorized into multiple different subclasses, each with specific different function and biogenesis mechanism. Results of this research support all of previous discoveries and solidify the idea that multiple biogenesis mechanisms co-exist for different subsets of human circRNAs. Results of this study can also contribute to the advance of circRNA detection algorithms. Current mainstream research methods adopted to discover circRNAs are based on detection of back-spliced junction sites

spanning reads within transcriptome sequencing data. However, this kind of approach tends to have a high false positive rate. Sensitivity of the junction site detection is also limited by sequencing depth. Had a clear concept of circRNA biogenesis mechanism is available, an improved algorithm ruling out sequence base false positive might be able to be developed.

3.5 Acknowledgments

We would like to thank Dr. Shu Chien for his helpful comments and discussions.

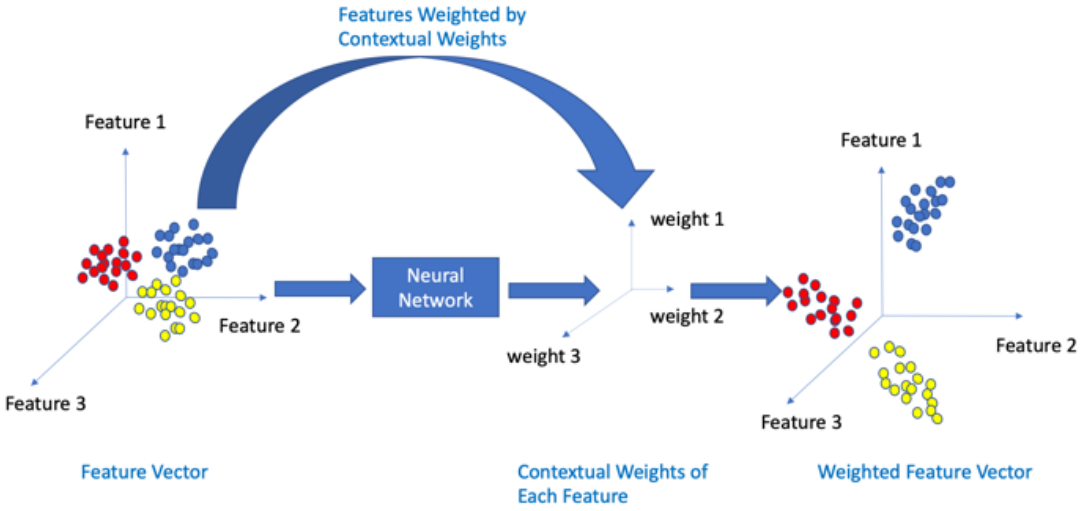
Funding: This work has been supported by the NIH R01 grant (HG009626)

Chapter 3, in full, is a reformatted reprint of the material as it appears in Liu, C., Liu, Y.-C., Huang, H.-D. & Wang, W. Biogenesis mechanisms of circular RNA can be categorized through feature extraction of a machine learning model. *Bioinformatics* **35**, 4867–4870 (2019). The dissertation author was a co-primary investigator and author of this paper.

3.6 Author Contributions

This study was conceived and designed by L.C., L.Y., H.H. and W.W.; data processed by L.Y.; algorithm developed by L.C.; data analyzed by L.Y. and L.C. Manuscript written by L.C., L.Y., H.H. and W.W. with input from all authors.

3.7 Figures



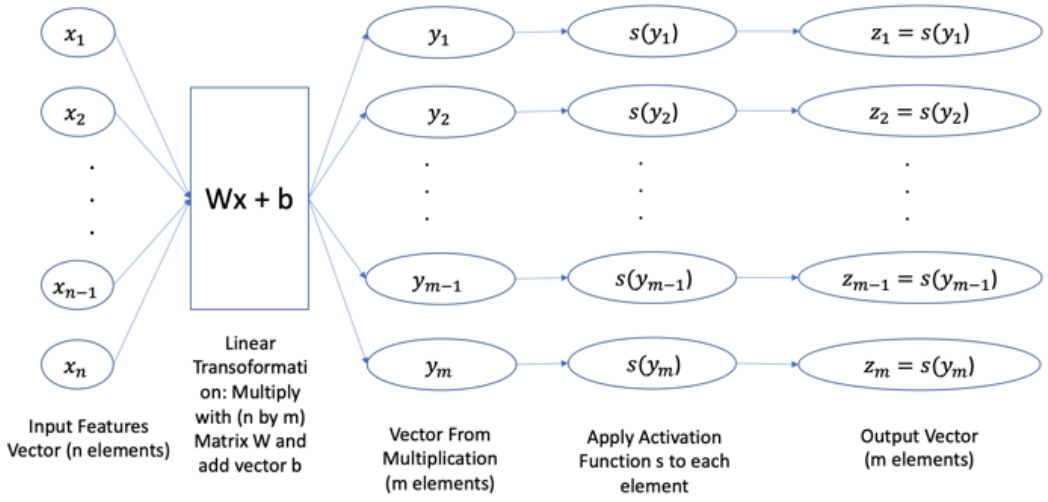


Figure 3.1. Illustration of Contextual Regression. A. graphic illustration of the mechanism of contextual regression: the features are inputted into a neural network that generates a contextual weight for each feature which represents the importance of the features. Then, the features are then weighted by the corresponding weights to make an easier separation of samples. In classification or regression tasks, the weighted features are then summed to yield the prediction. B. A graphic demonstration of the FNN parts in the contextual regression model.

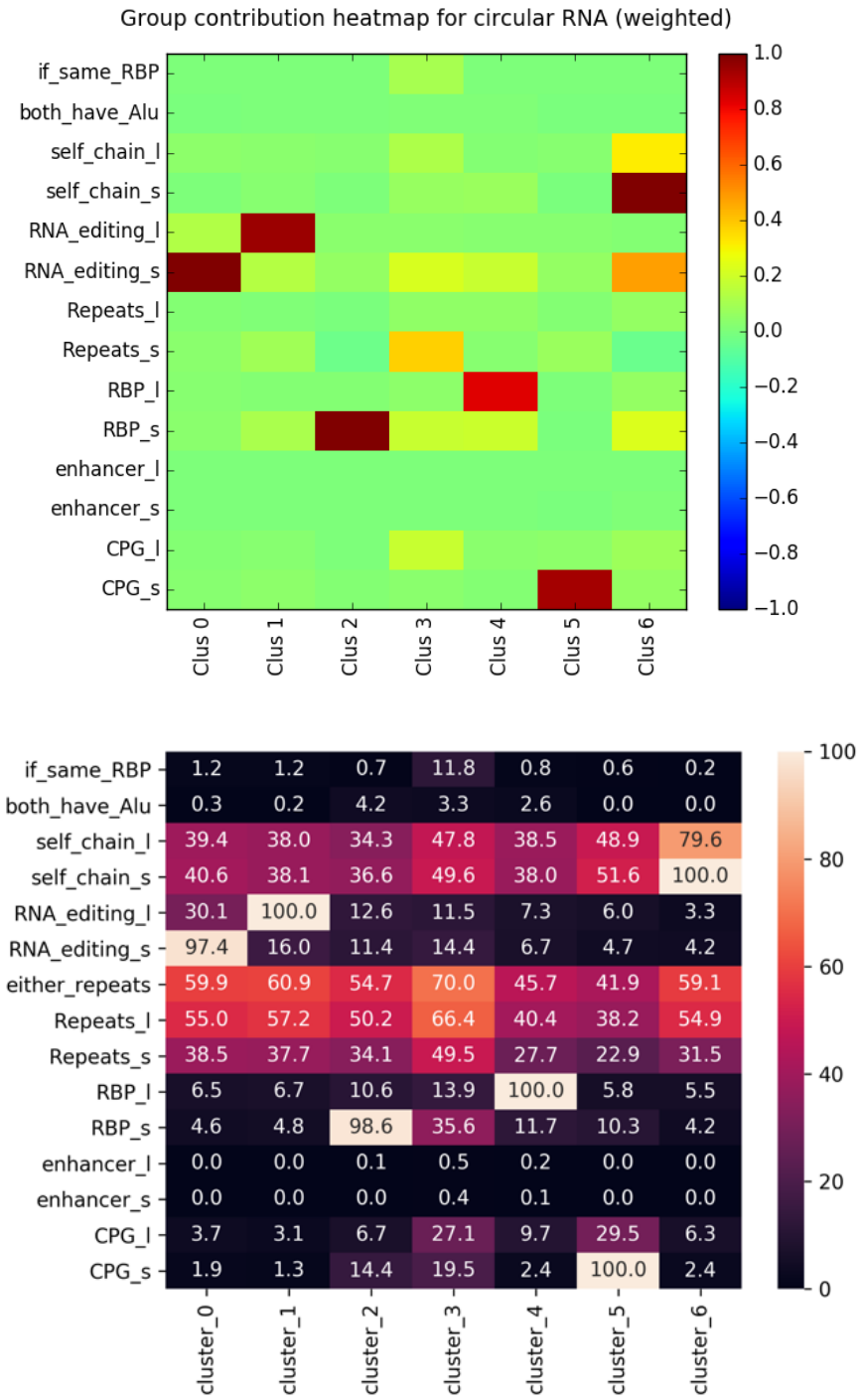


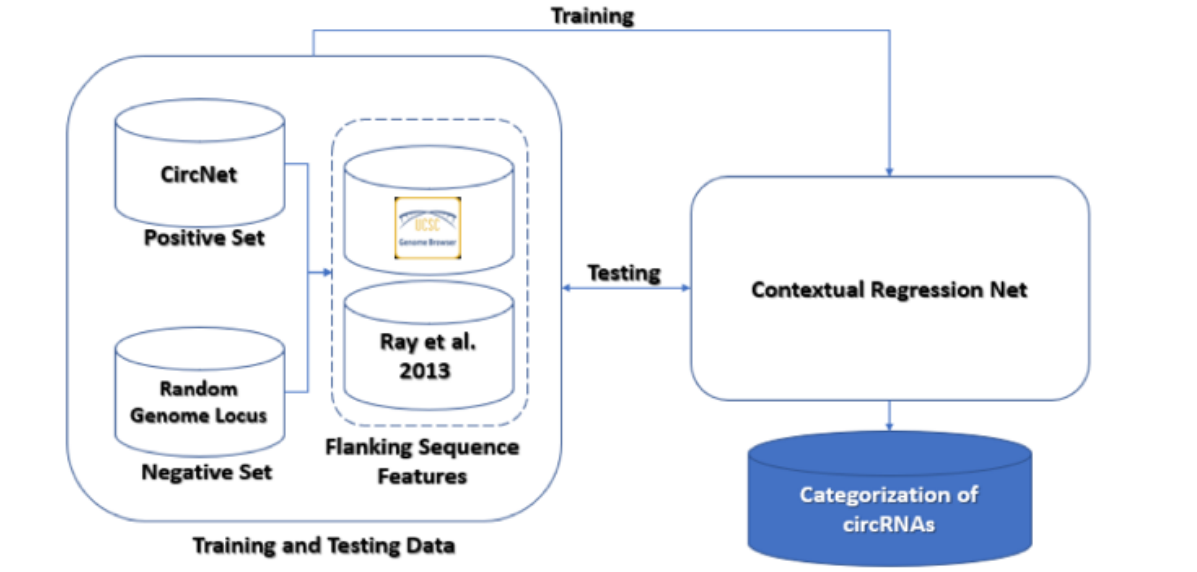
Figure 3.2. Prediction and feature collection result. The result of the feature collection is summarized in this figure. **A** Through the result we found that these circRNAs can be clustered into 7 different categories according to their biogenesis factors. The long range features are marked with "_l" and the short range ones are marked with "_s". **B** Percentage of members in each cluster that contains each of the features.

3.8 Supplementary Methods

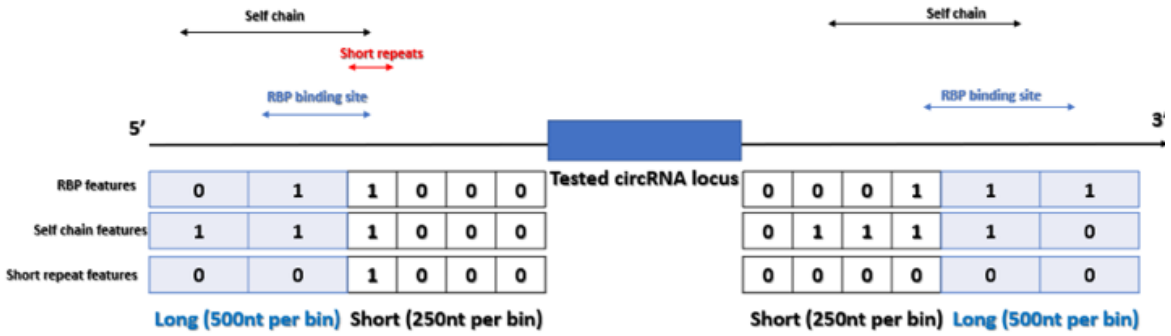
Transcriptome sequencing evidence based circRNA and sequence features selection. We first developed a computational method (scripts and tutorial are available at https://github.com/chl556/Contextual_Regression_for_CircRNA) to predict circRNAs using various sequence features (Figure 3.1S). We collected 55,689 transcriptome sequencing evidence based circRNAs and the same number of randomly selected loci in the hg19 human genome. As illustrated in Figure 3.1S B, the flanking regions of the selected sites were divided into short-range (from 0 to 1000 nucleotide upstream and downstream of the site) and long-range (from 1000 nucleotides to 2000 nucleotides upstream and downstream of the site) regions. Appearance of a specific feature within these regions was represented through bins with value of 0 or 1. The short-range regions were divided into 8 bins for every 250 nucleotides, while the long-range regions were divided into 4 bins for every 500 nucleotides. We considered 6 features related to circRNA biogenesis, including RNA Binding Protein (RBP) binding sites, simple repeats, RNA editing sites, DNA self-chains, CpG islands and enhancers. The first four features were selected based on the hypothesized circular biogenesis mechanisms in the previous studies[6-12]. CpG islands[13] and enhancers regions[14-15] are related to non-coding RNAs in general and thus considered as potential regulators for circRNA biogenesis. DNA self-chains are annotated regions where DNA is complementally aligned which are extracted from UCSC genome browser. A-to-I RNA editing sites are extracted from public available annotation[42]. Our list of RNA binding protein binding sites are extracted from available data[37]. The occurrence of these 6 features in the flanking regions of a selected genomic locus was represented by 72 bins. Furthermore, flanking short ALU repeats[8,43] and binding sites of single RBP such as QKI[11] are

hypothesized to present in flanking regions of circRNA to facilitate the biogenesis. Hence, in addition to the 72 bins, we added another 2 bins to represent whether both the upstream and downstream regions contain ALU repeats and whether the upstream or downstream regions contain binding sites of one particular RBP (QKI) binding sites.

3.9 Supplementary Figures

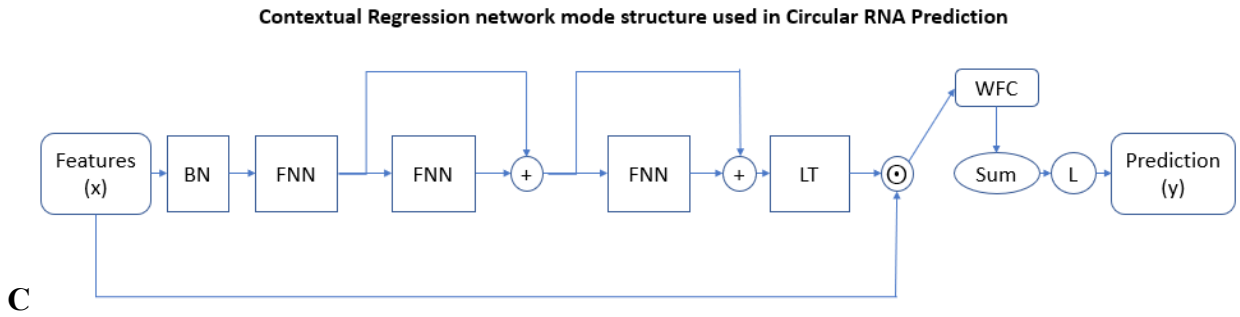


A



B

Figure 3.1S. Overview of the data analysis process. A. The general data analysis process is illustrated in this picture. Human circRNAs were collected from the database CircNet as positive training data, and equal amounts of random locus as negative training data. B. Flanking regions of the selected sites were divided into long and short range. Appearance of certain motifs within these regions was represented through bins with values of 0 or 1. C. With the selected features added, the data was used to train the contextual regression Network.



C

Figure 3.1S. Overview of the data analysis process. A. The general data analysis process is illustrated in this picture. Human circRNAs were collected from the database CircNet as positive training data, and equal amounts of random locus as negative training data. B. Flanking regions of the selected sites were divided into long and short range. Appearance of certain motifs within these regions was represented through bins with values of 0 or 1. C. With the selected features added, the data was used to train the contextual regression Network.

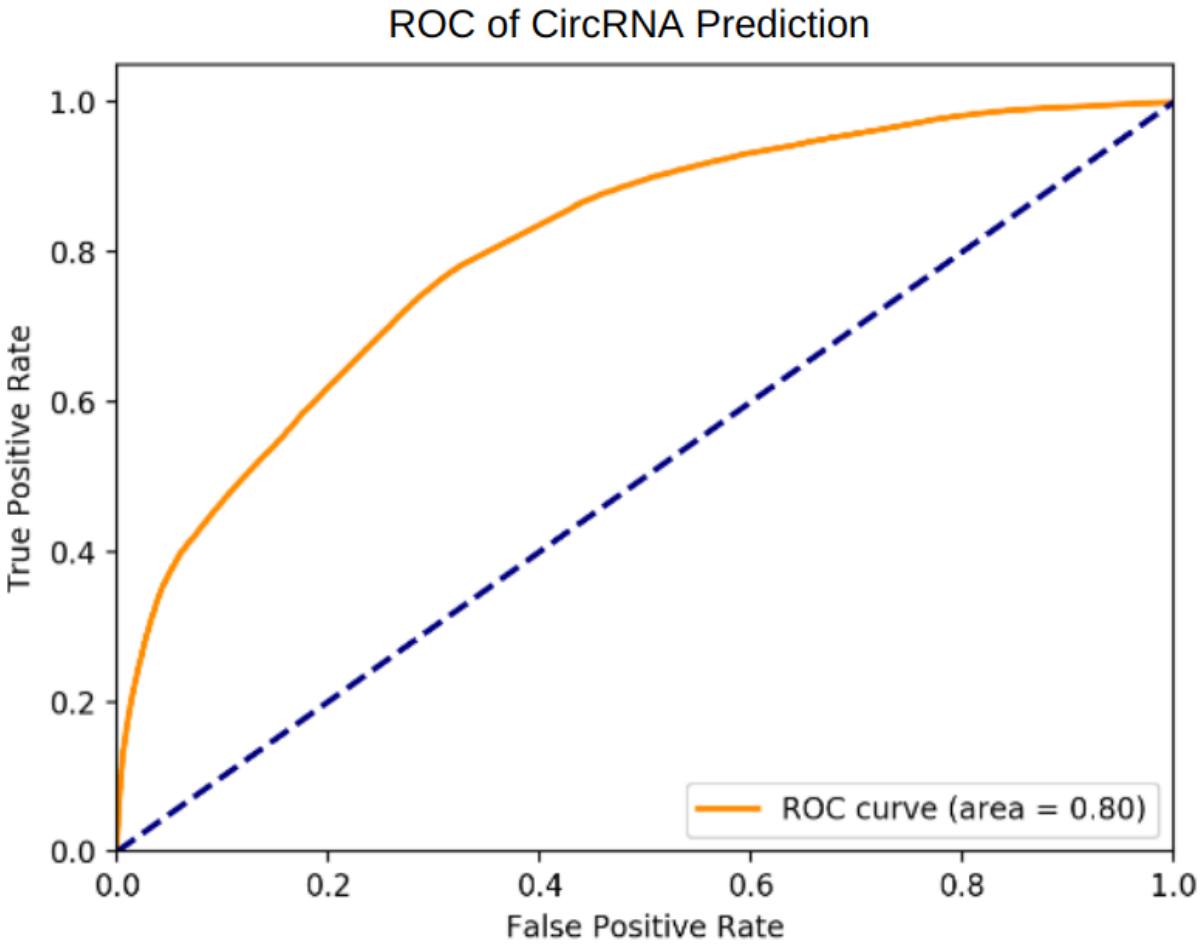


Figure 3.2S. ROC curve of circular RNA prediction. The area under curve of ROC curve is 0.81.

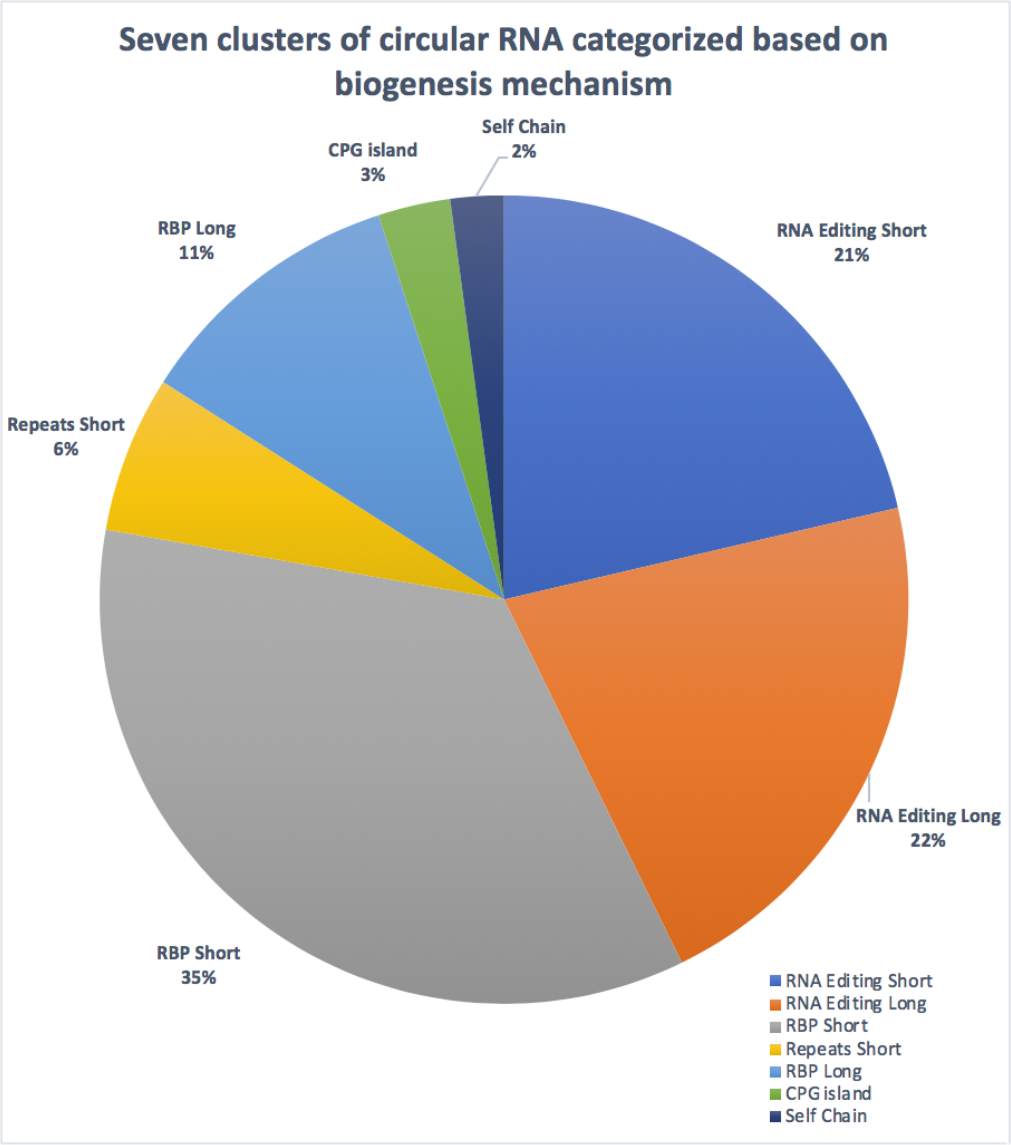


Figure 3.3S. Human circRNAs categorized based on biogenesis mechanism. The distribution of the seven clusters of circRNA is illustrated in this figure. Around 43% of the circRNA biogenesis was contributed to RNA editing domain, while 46% of other circRNA was contributed to flanking RBP binding sites.

Table 3.1S. Summary of the accuracy. Our contextual regression model achieved an average accuracy of 72.6% out of 10 runs, with an average AUC of ROC by 0.801.

Run	*Accuracy*	*AUC*
1	72.7%	0.801
2	72.6%	0.801
3	72.7%	0.800
4	72.6%	0.802
5	72.8%	0.802
6	72.5%	0.799
7	72.6%	0.801
8	72.7%	0.801
9	72.6%	0.803
10	72.6%	0.801
Avg	72.6%	0.801

Table 3.2S. Nine RBP found contribute to biogenesis of thousands of circRNAs. Binding motifs of SRSF1, PUM1, SF3B4, ELAVL1, HNRNPA1, CSDA, SNRPA, RBFOX1 and PABPC1 were found appear in upstream or downstream of thousands of circRNAs in both long and short range. The number in this table represents the amount of circRNA where the RBP binding motif appears in the flanking region.

RBP	*Short Upstream*	*Short Downstream*	*Long Upstream*	*Long Downstream*
PUM1	1353	1227	561	492
RBFOX1	1260	862	421	446
SRSF1	1149	817	668	654
ELAVL1	1004	927	364	371
CSDA	975	796	355	328
SF3B4	795	577	274	278
HNRNPA1	784	604	285	272
SNRPA	763	537	269	269
PABPC1	704	487	289	244

RBP	*PUM1*	*RBFOX 1*	*SRSF1*	*ELAVL1*	*CSDA*	*SF3B4*	*HNRNPA1*	*SNRPA*	*PABPC1*
Inside CircRN A	3814	2831	3906	2507	2949	2031	1287	1699	2150

Table 3.3S. Computational Time of Contextual Regression on Data of Different Scale. The computation is done on a machine with Intel Core m3 and 8G DDR3 memory. Results show that the time it takes to run contextual regression scales linearly with data size.

Run	Time Taken (110K data)	Time Taken (220K data)	Time Taken 440K (data)
1	897 seconds	1797 seconds	3566 seconds
2	908 seconds	1789 seconds	3572 seconds
3	922 seconds	1784 seconds	3595 seconds
4	905 seconds	1745 seconds	3489 seconds
5	910 seconds	1700 seconds	3508 seconds
Avg	908 seconds	1763 seconds	3546 seconds

Table 3.4S. Enrichment of Features in the Negative Set

	CPG_s	CPG_l	enhancer_s	enhancer_l	RBP_s	RBP_l	Both_have_Alu
Negative Set	3.59%	3.62%	0.27%	0.27%	2.44%	2.42%	0.09%

	Repeats_s	Repeats_l	RNA_editing_s	RNA_editing_l	self_chain_s	self_chain_l	If_same_RBP
Negative Set	72.13%	72.54%	4.20%	4.25%	43.57%	43.57%	0.42%

3.10 References

1. Rybak-Wolf, A., Stottmeister, C., Glažar, P., Jens, M., Pino, N., Giusti, S., Hanan, M., Behm, M., Bartok, O., Ashwal-Fluss, R., Herzog, M., Schreyer, L., Papavasileiou, P., Ivanov, A., Öhman, M., Refojo, D., Kadener, S. & Rajewsky, N. Circular RNAs in the Mammalian Brain Are Highly Abundant, Conserved, and Dynamically Expressed. *Mol. Cell* **58**, 870–885 (2015).
2. Szabo, L., Morey, R., Palpant, N. J., Wang, P. L., Afari, N., Jiang, C., Parast, M. M., Murry, C. E., Laurent, L. C. & Salzman, J. Statistically based spicing detection reveals neural enrichment and tissue-specific induction of circular RNA during human fetal development. *Genome Biol.* **16**, 126–126 (2015).
3. Bahn, J. H., Zhang, Q., Li, F., Chan, T.-M., Lin, X., Kim, Y., Wong, D. T. W. & Xiao, X. The landscape of microRNA, Piwi-interacting RNA, and circular RNA in human saliva. *Clin. Chem.* **61**, 221–230 (2015).
4. Li, Y., Zheng, Q., Bao, C., Li, S., Guo, W., Zhao, J., Chen, D., Gu, J., He, X. & Huang, S. Circular RNA is enriched and stable in exosomes: a promising biomarker for cancer diagnosis. *Cell Res.* **25**, 981–984 (2015).
5. Ashwal-Fluss, R., Meyer, M., Pamudurti, N. R., Ivanov, A., Bartok, O., Hanan, M., Evantal, N., Memczak, S., Rajewsky, N. & Kadener, S. circRNA biogenesis competes with pre-mRNA splicing. *Mol. Cell* **56**, 55–66 (2014).
6. Zhang, X.-O., Wang, H.-B., Zhang, Y., Lu, X., Chen, L.-L. & Yang, L. Complementary sequence-mediated exon circularization. *Cell* **159**, 134–147 (2014).
7. Zhang, Y., Zhang, X.-O., Chen, T., Xiang, J.-F., Yin, Q.-F., Xing, Y.-H., Zhu, S., Yang, L. &

Chen, L.-L. Circular intronic long noncoding RNAs. *Mol. Cell* **51**, 792–806 (2013).

8. Jeck, W. R., Sorrentino, J. A., Wang, K., Slevin, M. K., Burd, C. E., Liu, J., Marzluff, W. F. & Sharpless, N. E. Circular RNAs are abundant, conserved, and associated with ALU repeats. *RNA* **19**, 141–157 (2013).
9. Liang, D. & Wilusz, J. E. Short intronic repeat sequences facilitate circular RNA production. *Genes Dev.* **28**, 2233–2247 (2014).
10. Ivanov, A., Memczak, S., Wyler, E., Torti, F., Porath, H. T., Orejuela, M. R., Piechotta, M., Levanon, E. Y., Landthaler, M., Dieterich, C. & Rajewsky, N. Analysis of Intron Sequences Reveals Hallmarks of Circular RNA Biogenesis in Animals. *Cell Rep.* **10**, 170–177 (2015).
11. Conn, S. J., Pillman, K. A., Toubia, J., Conn, V. M., Salmanidis, M., Phillips, C. A., Roslan, S., Schreiber, A. W., Gregory, P. A. & Goodall, G. J. The RNA binding protein quaking regulates formation of circRNAs. *Cell* **160**, 1125–1134 (2015).
12. Li, B., Zhang, X.-Q., Liu, S.-R., Liu, S., Sun, W.-J., Lin, Q., Luo, Y.-X., Zhou, K.-R., Zhang, C.-M., Tan, Y.-Y., Yang, J.-H. & Qu, L.-H. Discovering the Interactions between Circular RNAs and RNA-binding Proteins from CLIP-seq Data using circScan. *bioRxiv* 115980 (2017). doi:10.1101/115980
13. Lai, F. & Shiekhattar, R. Where long noncoding RNAs meet DNA methylation. *Cell Res.* **24**, 263–264 (2014).
14. Chen, H., Du, G., Song, X. & Li, L. Non-coding Transcripts from Enhancers: New Insights into Enhancer Activity and Gene Expression Regulation. *Genomics Proteomics Bioinformatics* **15**, 201–207 (2017).
15. Lam, M. T. Y., Li, W., Rosenfeld, M. G. & Glass, C. K. Enhancer RNAs and regulated

transcriptional programs. *Trends Biochem. Sci.* **39**, 170–182 (2014).

16. Liu, C. & Wang, W. Contextual Regression: An Accurate and Conveniently Interpretable Nonlinear Model for Mining Discovery from Scientific Data. *arXiv preprint arXiv:1710.10728* (2017).
17. Khalid, S., Khalil, T. & Nasreen, S. *A survey of feature selection and feature extraction techniques in machine learning*. (IEEE, 2014).
18. Liu, Y.-C., Li, J.-R., Sun, C.-H., Andrews, E., Chao, R.-F., Lin, F.-M., Weng, S.-L., Hsu, S.-D., Huang, C.-C., Cheng, C., Liu, C.-C. & Huang, H.-D. CircNet: a database of circular RNAs derived from transcriptome sequencing data. *Nucleic Acids Res.* **44**, D209–D215 (2016).
19. Salzman, J., Gawad, C., Wang, P. L., Lacayo, N. & Brown, P. O. Circular RNAs are the predominant transcript isoform from hundreds of human genes in diverse cell types. *PLoS One* **7**, e30733 (2012).
20. Salzman, J., Chen, R. E., Olsen, M. N., Wang, P. L. & Brown, P. O. Cell-type specific features of circular RNA expression. *PLoS Genet.* **9**, e1003777 (2013).
21. Gao, Y., Wang, J. & Zhao, F. CIRI: an efficient and unbiased algorithm for de novo circular RNA identification. *Genome Biol.* **16**, 1 (2015).
22. Boeckel, J.-N., Jaé, N., Heumüller, A. W., Chen, W., Boon, R. A., Stellos, K., Zeiher, A. M., John, D., Uchida, S. & Dimmeler, S. Identification and characterization of hypoxia-regulated endothelial circular RNA. *Circ. Res.* **117**, 884–890 (2015).
23. Bachmayr-Heyda, Reiner & Auer. Correlation of circular RNA abundance with proliferation-exemplified with colorectal and ovarian cancer, idiopathic lung fibrosis, and

normal human tissues. *Sci. Rep.* **5**, 8057 (2015).

24. Memczak S, Papavasileiou P, Peters O, Rajewsky N. Identification and Characterization of Circular RNAs As a New Class of Putative Biomarkers in Human Blood. *PLoS One* **10**, e0141214 (2015)
25. Alhasan, A. A., Izuogu, O. G., Al-Balool, H. H., Steyn, J. S., Evans, A., Colzani, M., Ghevaert, C., Mountford, J. C., Marenah, L., Elliott, D. J., Santibanez-Koref, M. & Jackson, M. S. Circular RNA enrichment in platelets is a signature of transcriptome degradation. *Blood* **127**, e1–e11 (2016).
26. Cheng, J., Metge, F. & Dieterich, C. Specific identification and quantification of circular RNAs from sequencing data. *Bioinformatics* **32**, 1094–1096 (2016).
27. Zhang, X.-O., Dong, R., Zhang, Y., Zhang, J.-L., Luo, Z., Zhang, J., Chen, L.-L. & Yang, L. Diverse alternative back-splicing and alternative splicing landscape of circular RNAs. *Genome Res.* **26**, 1277–1287 (2016).
28. Song, X., Zhang, N., Han, P., Moon, B.-S., Lai, R. K., Wang, K. & Lu, W. Circular RNA profile in gliomas revealed by identification tool UROBORUS. *Nucleic Acids Res.* **44**, e87 (2016).
29. Dang, Y., Yan, L., Hu, B., Fan, X., Ren, Y., Li, R., Lian, Y., Yan, J., Li, Q., Zhang, Y., Li, M., Ren, X., Huang, J., Wu, Y., Liu, P., Wen, L., Zhang, C., Huang, Y., Tang, F. & Qiao, J. Tracing the expression of circular RNAs in human pre-implantation embryos. *Genome Biol.* **17**, 1 (2016).
30. Zheng, Q., Bao, C., Guo, W., Li, S., Chen, J., Chen, B., Luo, Y., Lyu, D., Li, Y., Shi, G., Liang, L., Gu, J., He, X. & Huang, S. Circular RNA profiling reveals an abundant

circHIPK3 that regulates cell growth by sponging multiple miRNAs. *Nat. Commun.* **7**, (2016).

31. Guo, J. U., Agarwal, V., Guo, H. & Bartel, D. P. Expanded identification and characterization of mammalian circular RNAs. *Genome Biol.* **15**, 10.1186 (2014).
32. Kelly, S., Greenman, C., Cook, P. R. & Papantonis, A. Exon Skipping Is Correlated with Exon Circularization. *J. Mol. Biol.* **427**, 2414–2417 (2015).
33. Leinonen, R., Sugawara, H., Shumway, M. & International Nucleotide Sequence Database, Collaboration. The sequence read archive. *Nucleic Acids Res.* **39**, D19–21 (2011).
34. Hansen, T. B., Venø, M. T., Damgaard, C. K. & Kjems, J. Comparison of circular RNA prediction tools. *Nucleic Acids Res.* **44**, e58–e58 (2016).
35. Glažar, P., Papavasileiou, P. & Rajewsky, N. circBase: a database for circular RNAs. *RNA* **20**, 1666–1670 (2014).
36. Casper, J., Zweig, A. S., Villarreal, C., Tyner, C., Speir, M. L., Rosenbloom, K. R., Raney, B. J., Lee, C. M., Lee, B. T., Karolchik, D., Hinrichs, A. S., Haeussler, M., Guruvadoo, L., Navarro Gonzalez, J., Gibson, D., Fiddes, I. T., Eisenhart, C., Diekhans, M., Clawson, H., Barber, G. P., Armstrong, J., Haussler, D., Kuhn, R. M. & Kent, W. J. The UCSC Genome Browser database: 2018 update. *Nucleic Acids Res.* **46**, D762–D769 (2017).
37. Ray, D., Kazan, H., Cook, K. B., Weirauch, M. T., Najafabadi, H. S., Li, X., Gueroussov, S., Albu, M., Zheng, H., Yang, A., Na, H., Irimia, M., Matzat, L. H., Dale, R. K., Smith, S. A., Yarosh, C. A., Kelly, S. M., Nabet, B., Mecenas, D., Li, W., Laishram, R. S., Qiao, M., Lipshitz, H. D., Piano, F., Corbett, A. H., Carstens, R. P., Frey, B. J., Anderson, R. A., Lynch, K. W., Penalva, L. O. F., Lei, E. P., Fraser, A. G., Blencowe, B. J., Morris, Q. D. &

Hughes, T. R. A compendium of RNA-binding motifs for decoding gene regulation. *Nature* 499, 172–177 (2013)

38. He, K., Zhang, X., Ren, S. & Sun, J. *Deep residual learning for image recognition*. (2016).
39. Ioffe, S. & Szegedy, C. *Batch normalization: Accelerating deep network training by reducing internal covariate shift*. (2015).
40. Kingma, D. P. & Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).
41. Liang, D. & Wilusz, J. E. Short intronic repeat sequences facilitate circular RNA production. *Genes Dev.* **28**, 2233–2247 (2014).
42. Ernesto Picardi, Anna Maria D'Erchia, Claudio Lo Giudice, Graziano Pesole, REDIportal: a comprehensive database of A-to-I RNA editing events in humans, *Nucleic Acids Research*, **45**, D750–D757 (2017)
43. Chen, L.-L. & Yang, L. Regulation of circRNA biogenesis. *RNA Biol.* **12**, 381–388 (2015).

CHAPTER 4: NEURAL NETWORK FACILITATED AB INITIO DERIVATION OF LINEAR FORMULA: A CASE STUDY ON FORMULATING THE RELATIONSHIP BETWEEN DNA MOTIFS AND GENE EXPRESSION

4.1 Abstract

Developing models with high interpretability and even deriving formulas to quantify relationships between biological data is an emerging need. We proposed a framework for ab initio derivation of sequence motifs and linear formula using a new approach based on the interpretable neural network model called contextual regression model. We showed that this linear model could predict gene expression levels using promoter sequences with a performance comparable to deep neural network models. We uncovered a list of 300 motifs with important regulatory roles on gene expression and showed that they also had significant contributions to cell-type specific gene expression in 154 diverse cell types. This work illustrates the possibility of deriving formulas to represent biology laws that may not be easily elucidated. (https://github.com/Wang-lab-UCSD/Motif_Finding_Contextual_Regression)

4.2 Introduction

Deep neural networks have been widely adopted in modeling biological data and they can achieve state-of-the-art prediction accuracy in different datasets such as gene expression, open chromatin and histone marks[1–7]. For example, convolutional neural network (CNN)[8] is suitable for processing sequence data as the filters in the CNN, which CNN uses to find sequence patterns for prediction, can represent DNA motifs and other functional patterns. A limitation of the highly complex and non-linear nature of neural network models is their low interpretability[9]. For instance, as the data go through multiple layers of matrix multiplication and non-linear activation functions in CNN, the contribution of each input feature to prediction result is hardly tractable and the filters also unlikely have equal contribution to the predictions. To address this challenge, methods such as LIME[10], SHAP[11], Contextual Regression (CR)[6] and DeepLift[12] have been developed to identify features important for the prediction performance. Such analyses are powerful and they can provide informative insights, such as which base-pairs are important in a sequence prediction task[4].

Despite these progresses, a tempting goal not yet achieved for building interpretable models is to derive formulas, which is intuitive for interpretation and human comprehension, for analytical quantification of the relationship between biological data such as DNA sequence and gene expression. A classic approach to derive formulas from data is symbolic regression[13] that searches combinations of predefined operations and functions, such as addition, subtraction, multiplication, division and exponent, on input variables to predict output variables. For example, symbolic regression[13] and, more recently, symbolic regression combined with neural networks[14] have been shown to successfully recover physics laws from data. These studies are

inspiring to investigate the possibility of deriving biology laws. However, it is imaginable that living systems are more complex than the physical subjects and in many cases defining appropriate input variables predictive of a target observation is a great challenge; in fact the complexity of the DNA sequence and the exponentially increasing searching space of operation combinations have hindered broad application of similar approaches in interpreting biological data.

In this study, we propose a framework to derive the simplest formulas, linear formulas, to model biological relationships. As many physics laws, such as Newton's second law F=ma and mass-energy equivalence $E=mC^2$, are in linear formulas, it is reasonable to argue some biology laws may also be linear [15]or can be well approximated by linear relationship[16–18]. We conducted a proof-of-concept study on predicting gene expression using promoter sequences. Promoters are crucial for regulating gene transcription[19]. DNA motifs, particularly those in the core promoter regions[20], that are important for transcriptional regulation have been investigated from molecular, genetic and functional aspects[21–36]. Many studies have shown that promoter sequences, genomic or synthetic ones, are predictive of gene expression and linear regression models of using sequence features in the promoters have reasonable performance while deep neural networks can achieve better predictions[5]. Such a linear relationship allows testing of our approach on (1) ab initio derivation of the input sequence features (i.e. DNA motifs) and (2) building a linear model on these derived features to achieve prediction performance comparable to deep neural networks.

4.3 Results

Ab initio derivation of linear model of sequence motifs to predict gene expression levels. A roadblock towards deriving a linear model to predict gene expression from promoter sequences is to uncover appropriate input features. DNA motifs represent genomic segments recognized by TFs and specific TFs binding to a particular promoter is the key of regulating transcription. Therefore, a linear model of DNA motifs to predict gene expression would naturally quantify the contribution of motifs and reveal the important transcriptional regulators. Our approach started with using a CNN model based on the contextual regression framework[6] that can learn expression related motifs and predict expression level by generating local linear models. The motifs and their occurrence counts were directly learned by this contextual regression model, and then the motif occurrence counts were used to fit a global linear model to predict RNA abundance achieving a performance comparable to pure neural network models.

The contextual regression model aims to simultaneously optimize prediction accuracy and assess the contribution of each input feature to the prediction. It is based on the strength of deep neural networks on end-to-end learning[37] process with automatic feature engineering (Figure 4.1SA). The model consists of two modules: motif count generating module and local linear model generating module. The architecture is demonstrated as shown below using motifs of 3-mers for illustrative purpose (in actuality, we used motifs of 8-mers in the prediction): The sequence is first input into a layer of CNN filters, each filter represents a motif of n-mers (8-mers were used in our model). The activation of each filter in a genomic location represents that the corresponding motif is found there. Then the output is fed into two branches: the first branch applies a sigmoid function (to binarize the output) and sum across the sequence to generate the

count of each motif, which we call motif finding branch. The second branch is input into a convolutional neural network that generates the contextual weights (local linear model), which we call contextual weight branch. The outputs from the two branches are then dot-producted (equivalent to applying the local linear model) to yield the prediction of expression level.

This way the CNN was trained on the promoter sequences and gene expression data to find local linear models for individual data points so that the two fitting problems were solved simultaneously: accurate prediction of the output (i.e. gene expression) and assessing the contribution of individual features represented by the contextual weights. We downloaded the gene expression data, which is the median expression level of 56 cell types, from ref. [5]. To assess the model performance, we randomly divided the data 10 times to generate 10 datasets for accuracy evaluation.

Human promoters include sequences upstream and downstream of transcription start site (TSS). Different lengths of sequences have been used in the literature, ranging widely from (-1000bp, +500bp) to (-7000bp, +3000bp). We thus first aimed to identify a reasonably short length of sequences around TSS that can predict gene expression with satisfactory performance, because relative short sequences would reduce the possibility of overfitting and alternative models, making it more straightforward to compare the contextual regression model with the pure neural network models. We assessed the prediction accuracy of our model on three different sizes around TSS: (-100bp, +100bp), (-200bp, +200bp) and (-1000bp, 1000bp). We found the performance (Pearson correlation R between predicted and experimental values) of (-200bp, +200bp) (0.675±0.018) was very similar to (-1000bp, 1000bp) (0.682±0.024) and yet sufficiently higher than (-100bp, +100bp) (0.648±0.019) (Figure 4.1B). For a comparison, a slightly higher

Pearson correlation could be achieved on a much longer region of (-7000bp, +3500bp) around TSS either using a deep neural network model in ref. [5] (Pearson R=0.71) or a contextual regression model with the same model structure using (-1000bp, 1000bp) (Pearson R=0.68, Figure 4.1A). It is obvious that the longer the promoter region, the more information obtained for more accurate prediction of gene expression. For our purpose of demonstrating the feasibility of deriving linear formulas, we chose (-200bp, +200bp) around TSS to train the model and the follow up analyses in the rest of the paper.

Then we performed the second step to generate the global linear model. A global linear model was fit for each of the 10 runs on the 10 random partitions of the data (Figure 4.1C). The global linear models show slightly worse but still highly compatible performance (R=0.65±0.02) compared to the contextual regression model (R=0.68±0.02) and the Xpresso model structure (R=0.67±0.02), suggesting that the RNA expression level can be predicted using linear models of sequence motifs (Figure 4.1D).

To investigate whether the contextual regression model helped to uncover appropriate features, we developed a benchmark neural network model which predicts gene expression solely using motif counts (called motif finding NN): this model is the motif finding branch of the contextual regression model connected to a 3-layer feed forward neural network model (Figure 4.1SB). This model achieved a PearsonR=0.52±0.02 which is not as good as contextual regression. This is probably due to the fact that the contextual weight branch of the contextual regression model takes in information from the whole sequence, which facilitates learning. We also built a global linear model to predict gene expression based on the activation count of the filters (i.e. equivalent to the motif counts) in the Xpresso model (i.e. the output of the first layer

activated by ReLU summed over each (-200bp, +200bp) region) and the average Pearson R was 0.41±0.03 on the test set. This result indicated that simply replacing neural network with linear regression does not necessarily work and the contextual regression framework facilitates learning input features that are able to fit a linear model for prediction.

Uncovering a list of promoter motifs important for regulating gene expression. DNA motifs in the human promoters have been well analyzed but new motifs and new insights keep emerging from recent studies[27,29,30,38]. Leveraging the power of the highly interpretable linear model, we set to define a list of promoter motifs that are most important for regulating gene expression. A technical hurdle to overcome is that there exist alternative models to achieve the same prediction accuracy. A common strategy is to combine those models, which can not only unify the models but also increase model adaptivity to different datasets. Thus we created a combined model from the models we generated in the 10 runs. Furthermore, some de novo motifs may match with the known motifs. To reduce the effort of distinguishing known and de novo motifs in the follow up analysis, we chose to include the known motifs in the core vertebrate JASPAR database (a total of 831 motifs)[39] in addition to the 640 de novo motifs found by the 10 contextual regression (CR) models (Figure 4.2A). Then we selected the top motifs based on their predictive power on gene expression. We chose the regression tree[40] as our selection model that uses entropy to represent feature significance. We recursively remove 100 motifs in each round until the number of motifs reach our target value.

We compared the prediction accuracy of linear models using the top selected motifs with a baseline linear regression model directly trained on the JASPAR motifs and the neural network

models to predict gene expression. We used the default division of train, validation and test set from ref.[5] for a fair comparison. Among all the models, the one using the top 300 motifs selected from JASPAR+CR Found motifs achieved the best performance on the test set while being linearly interpretable and simpler (Pearson R=0.70), same as the two neural network models and much better than the baseline linear model (R=0.60). On the test set, the top 300 motifs showed better performance than using all the motifs which is likely due to removing low information motifs that can fit to noise on the training set. Note that the top-300-motif model had the same performance on training, testing and whole dataset fitting, while the top-200-motif model, despite achieving close accuracy on the test set, performed slightly worse on training and whole dataset, suggesting some underfitting. Taken together, we chose the linear top-300-motif model for the follow up analysis.

We noticed that the most high-weight motifs (194 out of 300, mean absolute weight 1.53) came from contextual regression (in fact all top 20 positive and negative weighted motifs were from contextual regression) rather than from the JASPAR database (106 out of 300, mean absolute weight 0.012), suggesting that the de novo CR motifs are stronger features for expression prediction than the known motifs. This is not surprising since the motifs from contextual regression are directly trained to predict gene expression.

We next examined the spatial distribution of the motifs with the top 20 positive and negative weights in the promoters. The distribution is calculated by averaging the output from the motif_exists module (Figure 4.1SA) of our model over all sequences in the training, validation and testing set. These motifs exhibit a strong change of location preference near the TSS: for instance, motif R1M4 has an increasing presence when approaching TSS. Near TSS it

has a strong preference for specific positions (large fluctuation of distribution). Another example is motif R2M6, which has an even distribution before and after TSS and has a strong preference for specific positions near TSS (Figure 4.2B). This suggests certain binding spots near TSS are preferred.

The de novo motifs discovered by contextual regression model are highly effective on predicting transcription levels ande most of them do not match with the known motifs in the JASPAR[39] or CIS-BP[41] database. To infer their cellular function, for each of the top 20 positive and top 20 negative weight motifs, we extracted the top 500 genes with the highest total motif count and performed GO term enrichment analysis[42–44] on their biological process and molecular function. We found three main themes of GO terms associated with these motifs (Figure 4.2B): (1) development (for instance, nervous system development (GO:0007399), multicellular organism development (GO:0007275) and system development (GO:0048731)), (2) signaling and regulation (for instance, regulation of signaling (GO:0023051), regulation of cellular process (GO:0050794) and regulation of signal transduction (GO:0009966)) and (3) stimulus-response (for instance, detection of chemical stimulus (GO:0009593), detection of chemical stimulus involved in sensory perception (GO:0050907) and G protein-coupled receptor signaling pathway (GO:0007186)). The association of the high weighted motifs with these housekeeping functions indicated their importance of regulating expression levels of the related genes.

Predicting gene expression in diverse cell types. We next examined whether the top 300 motifs define a parts list of TF binding motifs in the promoters and whether the linear relationship remains valid on predicting cell-type specific gene expressions. We collected the

RNA-seq data of 154 cell types/tissues[45] from Epigenomic Roadmap Project[46] and ENCODE[47]. We trained and tested the model using the same division of train, validate and test set as in the previous section. Our model achieved an average Pearson Rof 0.61 on the test set for all cell types/tissues. Among the 154 cell lines, 98% (151) of them had a Pearson R above 0.5. Considering that no epigenetic state or enhancer regulation was included in the model, such a performance indicates that the 300 motifs in the promoters play a major role of regulating cell-type specific gene expression. The 3 cell lines that had Pearson R below 0.5 are all adult cell lines (R.adt.sigmoid.colon, R.adt.small.intestine, E.adt.right.lobe.of.liver). Indeed, the average Pearson correlation for the adult cell types/tissues was 0.59, which is significantly lower than 0.61 (p value=0.022) for children and 0.64 (p value=5.6e-7) for the embryonic cells/tissues (Figure 4.3B). As gene expressions of the cell types/tissues in the later stages of development are generally more affected by epigenetic modifications and enhancers, this observation highlights the importance to consider information other than promoter sequences in the prediction model.

We next investigated how the top 20 positive and negative weighted motifs selected from predicting the median gene expression values in the previous section are functioning in regulating cell-type specific expressions. We performed hierarchical clustering on the cells/tissues based on their linear model weights of the 40 motifs (Figure 4.3A). We observed that the cell types/tissues with the same origin tend to be clustered together, such as embryo muscular cells, immune cells and cells from the digestive tract.

Obviously, some motifs are universally important in the 154 cells/tissues while some are more specific. We calculated the normalized standard deviation (NSD, standard deviation divided by absolute mean value) of linear model weights for each motif. We classified NSD into

3 levels of low (<0.5), medium (0.5, 1) and high (>1) (Figure 3.B). Among the 20 positive weighted motifs, 60% had a low NSD, 30% a medium NSD and 10% a high NSD. Among the 20 negative weighted motifs, 65% had a low NSD, 15% a medium NSD and 20% a high NSD. For both negative and positive weight motifs, most of them (25 out of 40) had a low NSD, suggesting their broad importance in regulating gene expression across diverse cell/tissue types. Interestingly, 13 out of the 15 medium and high NSD motifs had top GO-terms in regulation and stimulus-response pathways. A possible explanation is that regulation and stimulus-response genes have a high variability of expression level in different cell types and thus are regulated by their associated motifs in very different ways.

4.4 Discussion

Here we have conducted a proof-of-concept study to demonstrate a new strategy to derive linear formulas to quantify the relationship between biological data. We showed that a linear model of DNA motifs could predict the median gene expression levels with a performance comparable to deep neural networks. Conventional approaches to interpret a neural network are to identify the important features and the contribution of each feature to prediction is evaluated by the reduction of prediction performance if a feature is removed. While these approaches are powerful, they do not quantify the relationship between individual features and the predicted variable(s) with an explicit linear regression model. Our study provides a possible solution by combining neural networks with linear models to create an interpretable and high performance model to quantitatively assess the individual DNA motifs' regulation on gene expression.

Linear relationship is the simplest formula and likely easiest to achieve during evolution of the biological systems. One key step of deriving the linear formula is to find appropriate combinations of the input features, which has not been well explored by deep learning approaches. A conventional neural network model using DNA motifs to predict gene expression only achieved a Pearson R=0.52. In comparison, the contextual regression model uncovered features capable of predicting outputs in a linear model with a Pearson R=0.65, comparable to pure deep learning models directly trained from the promoter sequences (Pearson R=0.67). This observation shows the unique advantage of contextual regression on simultaneously optimizing prediction accuracy and assessing the feature contribution, which enforces direct learning of the input features (motifs) fit to a linear formula for predicting outputs (gene expressions).

The strategy we developed here can be readily applied to deriving formulas to quantify the relationship between DNA sequences and different biological functions such as histone modifications, TF binding and open chromatin. This study exemplifies a possible next step of building interpretable models leading to derivation of analytical formulas to decipher the information encoded in the human genome and explain biology that has not been achieved by conventional deep learning models. Additional possible extension of this approach includes conducting symbolic regression on the identified features (motifs) from the CR model to derive non-linear formulas.

Using this approach, we have uncovered 300 motifs that are predictive of the median gene expressions with a linear model achieving similar accuracy as a deep neural network model. The weights of the motifs in the linear model naturally quantify their importance in regulating gene expression. Such a parts list of regulatory motifs in the promoters include a majority of new motifs that are more predictive of gene expression than the known motifs, suggesting unknown regulatory mechanisms implemented in the promoters. Furthermore, we showed these 300 motifs are predictive of cell-type specific gene expression particularly in the embryo cell types/tissues while performing less well on adult cell types/tissues, indicating the increasing importance of enhancers and epigenetic modification during development on controlling gene expressions.

4.5 Acknowledgments

Funding: This work was partially supported by NIH (R01HG009626 to W.W.).

Chapter 4, in full, is currently being prepared for submission of the material as Liu, C., Wang, W. Neural network facilitated ab initio derivation of linear formula: A case study on formulating the relationship between DNA motifs and gene expression. The dissertation author was a primary investigator and author of this paper.

4.6 Author Contributions

This study was conceived and designed by L.C. and W.W.; algorithm developed, data processed and analyzed by L.C.; Manuscript written by L.C. and W.W. with input from all authors.

4.7 Figures

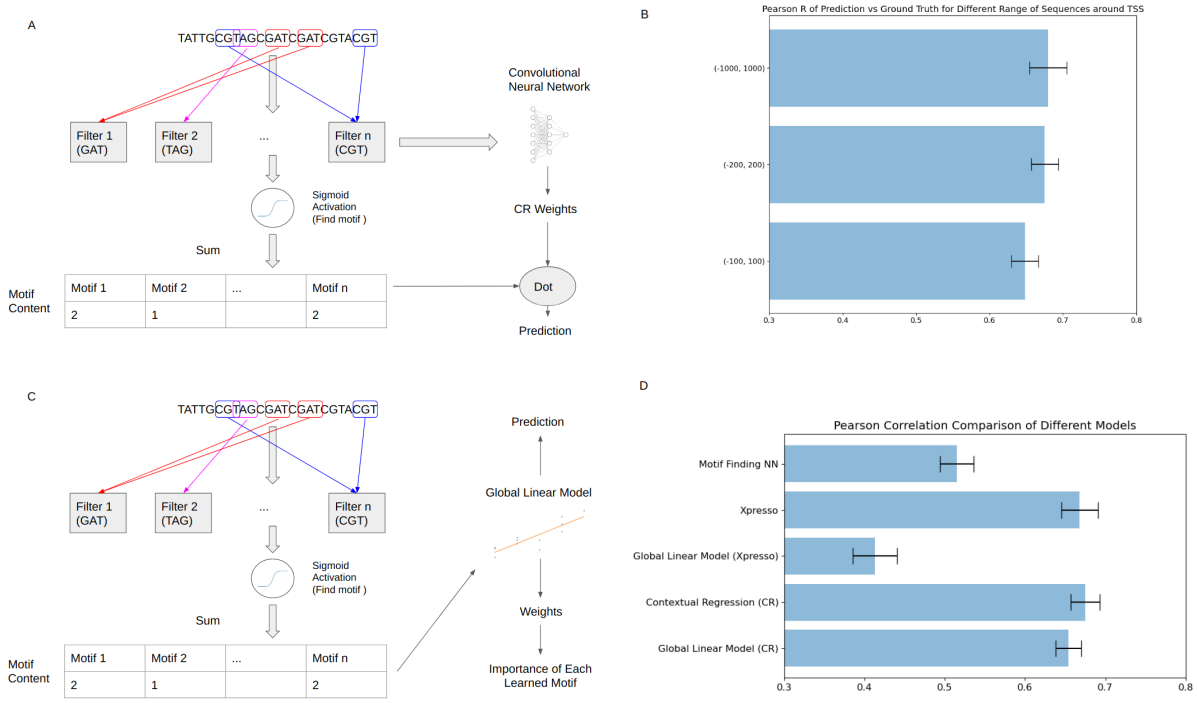


Figure 4.1. A. A new contextual regression model to predict expressions from promoter sequences. B. Accuracy of the contextual regression model using sequences of different sizes around TSS. C. Building a global linear model from a contextual regression output. D. Accuracy of different models on using (-200, 200) sequence around TSS.

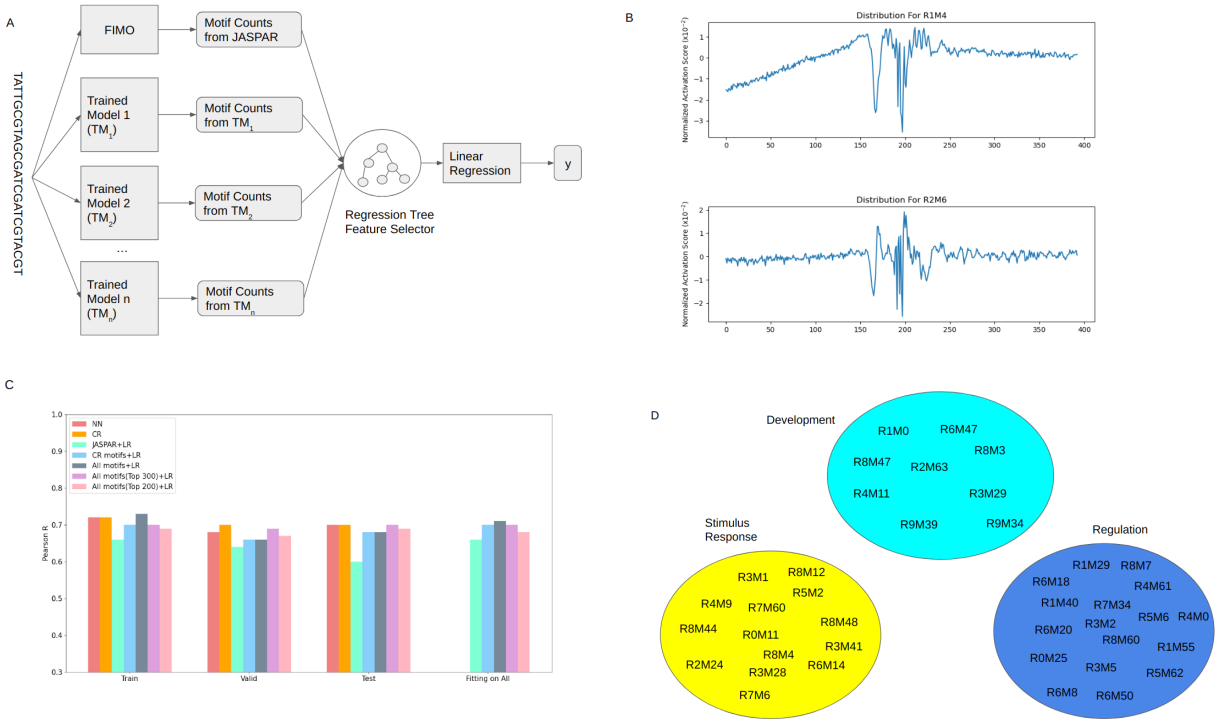


Figure 4.2. A. Selection of 300 promoter motifs important for predicting gene expression. B. Examples of the two typical distributions of motifs around TSS (R1M4 and R2M6) C. Accuracy comparison of different models on train, valid, test and all set D. The top 20 positive and negative weighted motifs grouped by their associated GO term categories.

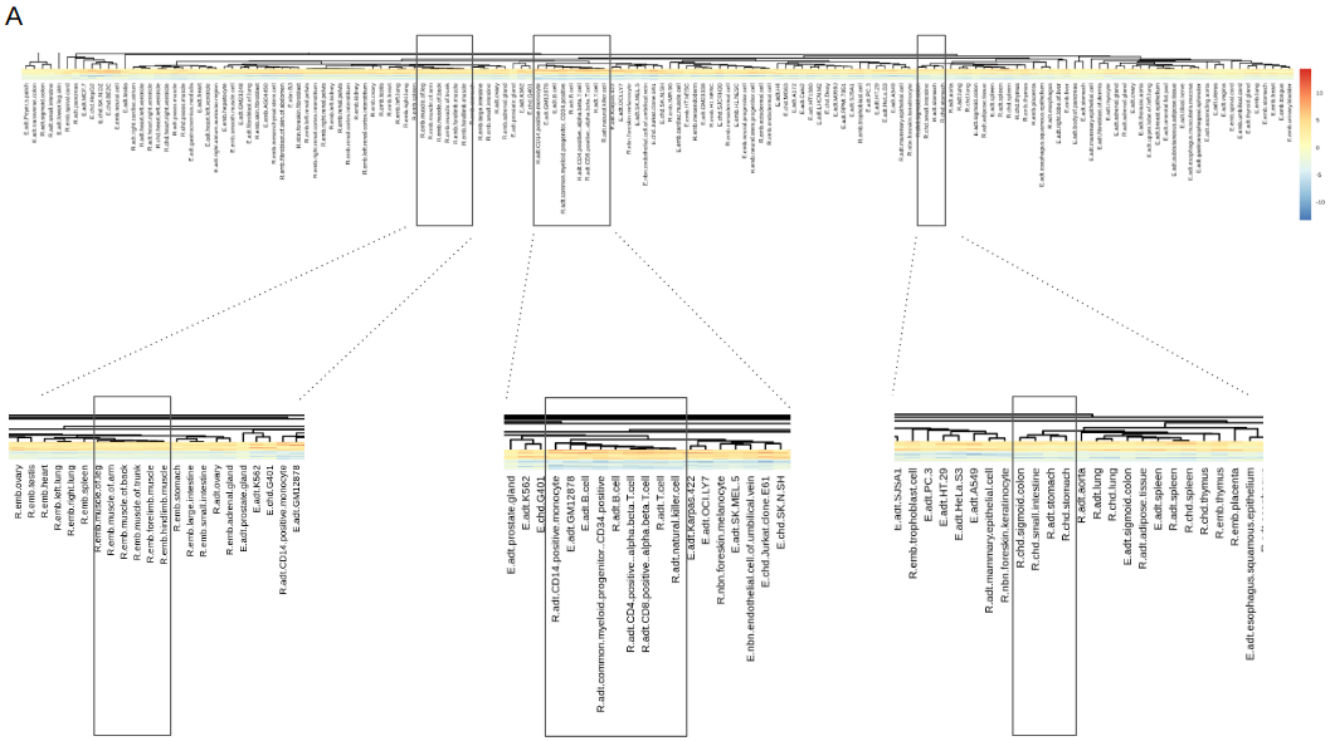


Figure 4.3. A. Hierarchical clustering of cell/tissue types by the linear model weights of the selected 300 motifs. Noticeable clusters of similar cell/tissue types are shown in amplification. B. Mean value of Pearson correlation, number of samples for the four different developmental stages and their one-sided t test p value vs the adult cells/tissues. C. The linear model weights of the 40 motifs in different cell/tissue types. The NSD level of each motif is marked below the motif sequence: blue represents low NSD, orange represents medium NSD and red represents high NSD.

B

	Mean Pearson R	Number of Samples	t-test p value vs adult
Adult	0.593	77	N/A
Child	0.620	18	0.02
New Born	0.648	4	N/A
Embryo	0.636	54	5.6e-7

Figure 4.3. A. Hierarchical clustering of cell/tissue types by the linear model weights of the selected 300 motifs. Noticeable clusters of similar cell/tissue types are shown in amplification. B. Mean value of Pearson correlation, number of samples for the four different developmental stages and their one-sided t test p value vs the adult cells/tissues. C. The linear model weights of the 40 motifs in different cell/tissue types. The NSD level of each motif is marked below the motif sequence: blue represents low NSD, orange represents medium NSD and red represents high NSD.

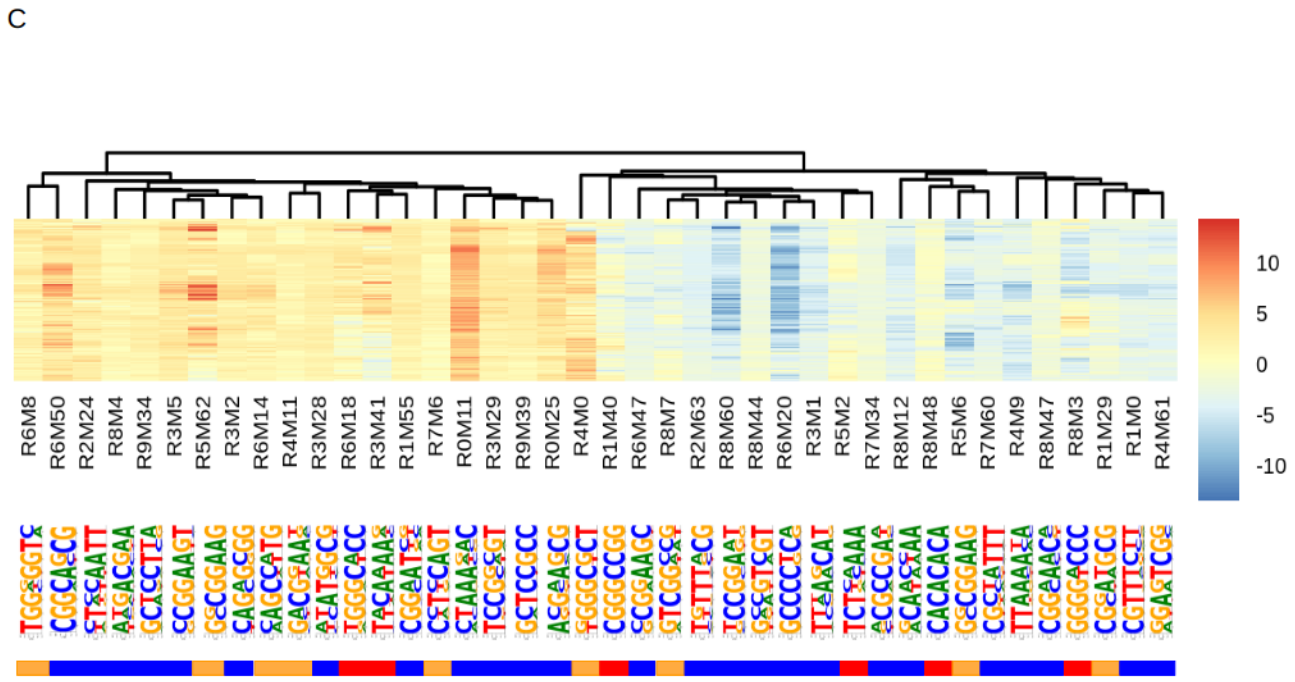


Figure 4.3. A. Hierarchical clustering of cell/tissue types by the linear model weights of the selected 300 motifs. Noticeable clusters of similar cell/tissue types are shown in amplification. B. Mean value of Pearson correlation, number of samples for the four different developmental stages and their one-sided t test p value vs the adult cells/tissues. C. The linear model weights of the 40 motifs in different cell/tissue types. The NSD level of each motif is marked below the motif sequence: blue represents low NSD, orange represents medium NSD and red represents high NSD.

4.8 Supplementary Figures And Table

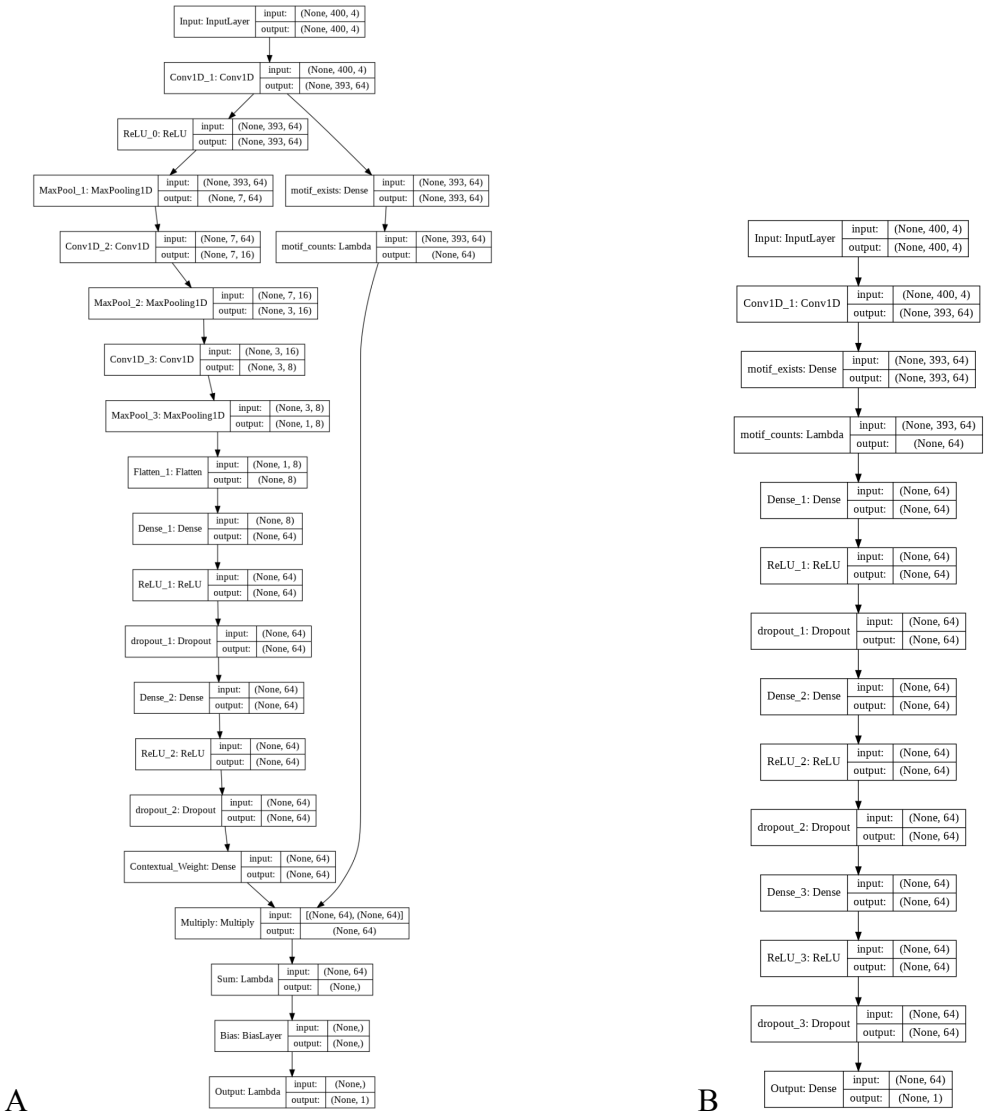


Figure 4.1S. A. Structure of the contextual regression model, the left branch is the contextual weight branch and the right branch is the motif finding branch. B. Structure of the motif finding NN

Table 4.1S. Top 20 negative and positive weighted motifs, with their PWM plot, distribution (average by base pair) around TSS and the top 5 associated GO terms. (Continued)

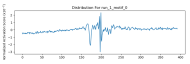
Negative Weight Motif Symbol	Motif PWM	Distribution	Biological Process (Top 5 FDR)	Molecular Function (Top 5 FDR)
R1M0			nervous system development (GO:0007399), negative regulation of cellular process (GO:0048523), regulation of cell junction assembly (GO:1901888), multicellular organism development (GO:0007275), carboxylic acid metabolic process (GO:0019752)	N/A

Table 4.1S. Top 20 negative and positive weighted motifs, with their PWM plot, distribution (average by base pair) around TSS and the top 5 associated GO terms. (Continued)

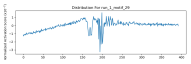
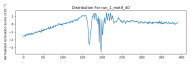
R1M29			regulation of signaling (GO:0023051), regulation of cell communication (GO:0010646), regulation of cellular process (GO:0050794), signaling (GO:0023052), regulation of signal transduction (GO:0009966)	nucleoside-triphosphatase regulator activity (GO:0060589), GTPase regulator activity (GO:0030695), protein serine/threonine/tyrosine kinase activity (GO:0004712), protein serine/threonine/tyrosine kinase activity (GO:0004712), guanyl-nucleotide exchange factor activity (GO:0005085)
R1M40			biological regulation (GO:0065007), regulation of cellular process (GO:0050794), negative regulation of cellular process (GO:0048523), regulation of molecular function (GO:0065009), regulation of signaling (GO:0023051)	binding (GO:0005488), protein binding (GO:0005515), nucleoside-triphosphatase regulator activity (GO:0060589), GTPase regulator activity (GO:0030695), molecular_function (GO:0003674)

Table 4.1S. Top 20 negative and positive weighted motifs, with their PWM plot, distribution (average by base pair) around TSS and the top 5 associated GO terms. (Continued)

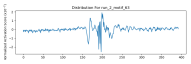
R2M63			nervous system development (GO:0007399), synapse organization (GO:0050808), system development (GO:0048731), multicellular organismal process (GO:0032501), multicellular organism development (GO:0007275)	N/A

Table 4.1S. Top 20 negative and positive weighted motifs, with their PWM plot, distribution (average by base pair) around TSS and the top 5 associated GO terms. (Continued)

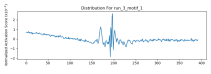
R3M1			detection of chemical stimulus (GO:0009593), detection of chemical stimulus involved in sensory perception (GO:0050907), detection of chemical stimulus involved in sensory perception of smell (GO:0050911), sensory perception of smell (GO:0007608), sensory perception of chemical stimulus (GO:0007606)	olfactory receptor activity (GO:0004984), G protein-coupled receptor activity (GO:0004930), transmembrane signaling receptor activity (GO:0004888), signaling receptor activity (GO:0038023), molecular transducer activity (GO:0060089)

Table 4.1S. Top 20 negative and positive weighted motifs, with their PWM plot, distribution (average by base pair) around TSS and the top 5 associated GO terms. (Continued)

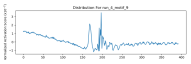
R4M9			detection of chemical stimulus (GO:0009593), detection of chemical stimulus involved in sensory perception (GO:0050907), G protein-coupled receptor signaling pathway (GO:0007186), sensory perception of chemical stimulus (GO:0007606), detection of chemical stimulus involved in sensory perception of smell (GO:0050911)	G protein-coupled receptor activity (GO:0004930), olfactory receptor activity (GO:0004984), transmembrane signaling receptor activity (GO:0004888), signaling receptor activity (GO:0038023), molecular transducer activity (GO:0060089)

Table 4.1S. Top 20 negative and positive weighted motifs, with their PWM plot, distribution (average by base pair) around TSS and the top 5 associated GO terms. (Continued)

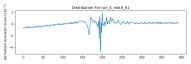
R4M61			regulation of transcription DNA-templated (GO:0006355) , regulation of developmental process (GO:0050793), regulation of cell communication (GO:0010646), biological regulation (GO:0065007), regulation of Wnt signaling pathway (GO:0030111)	binding (GO:0005488), molecular_function (GO:0003674), nucleoside-triphosphatase regulator activity (GO:0060589), GTPase regulator activity (GO:0030695), enzyme activator activity (GO:0008047)

Table 4.1S. Top 20 negative and positive weighted motifs, with their PWM plot, distribution (average by base pair) around TSS and the top 5 associated GO terms. (Continued)

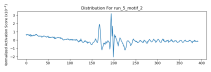
R5M2			detection of chemical stimulus involved in sensory perception (GO:0050907), detection of chemical stimulus (GO:0009593), detection of chemical stimulus involved in sensory perception of smell (GO:0050911), sensory perception of smell (GO:0007608), sensory perception of chemical stimulus (GO:0007606)	olfactory receptor activity (GO:0004984), G protein-coupled receptor activity (GO:0004930), transmembrane signaling receptor activity (GO:0004888), signaling receptor activity (GO:0038023), molecular transducer activity (GO:0060089)

Table 4.1S. Top 20 negative and positive weighted motifs, with their PWM plot, distribution (average by base pair) around TSS and the top 5 associated GO terms. (Continued)

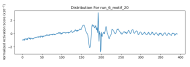
R5M6			regulation of cell communication (GO:0010646), regulation of biological process (GO:0050789), regulation of cellular process (GO:0050794), regulation of signaling (GO:0023051), biological regulation (GO:0065007)	transcription coregulator activity (GO:0003712), guanyl-nucleotide exchange factor activity (GO:0005085), nucleoside-triphosphatase regulator activity (GO:0060589), GTPase regulator activity (GO:0030695), protein binding (GO:0005515)
R6M20			regulation of signaling (GO:0023051), regulation of cell communication (GO:0010646), regulation of signal transduction (GO:0009966), regulation of response to stimulus (GO:0048583), phosphorylation (GO:0016310)	protein serine/threonine/tyrosine kinase activity (GO:0004712), protein kinase activity (GO:0004672), phosphotransferase activity alcohol group as acceptor (GO:0016773), kinase activity (GO:0016301), transferase activity transferring phosphorus-containing groups (GO:0016772)

Table 4.1S. Top 20 negative and positive weighted motifs, with their PWM plot, distribution (average by base pair) around TSS and the top 5 associated GO terms. (Continued)

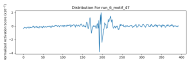
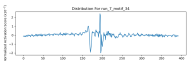
R6M47			system development (GO:0048731), developmental process (GO:0032502), multicellular organism development (GO:0007275), anatomical structure development (GO:0048856), regulation of biological process (GO:0050789)	protein tyrosine kinase binding (GO:1990782), protein binding (GO:0005515), binding (GO:0005488), olfactory receptor activity (GO:0004984), sequence-specific DNA binding (GO:0043565)
R7M34			regulation of signaling (GO:0023051), regulation of cell communication (GO:0010646), regulation of signal transduction (GO:0009966), regulation of response to stimulus (GO:0048583), multicellular organism development (GO:0007275)	protein binding (GO:0005515), signaling receptor binding (GO:0005102), binding (GO:0005488), olfactory receptor activity (GO:0004984), signaling receptor regulator activity (GO:0030545)

Table 4.1S. Top 20 negative and positive weighted motifs, with their PWM plot, distribution (average by base pair) around TSS and the top 5 associated GO terms. (Continued)

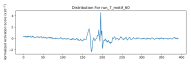
R7M60			detection of chemical stimulus involved in sensory perception (GO:0050907), detection of chemical stimulus involved in sensory perception of smell (GO:0050911), sensory perception of smell (GO:0007608), detection of chemical stimulus (GO:0009593), sensory perception of chemical stimulus (GO:0007606)	olfactory receptor activity (GO:0004984), G protein-coupled receptor activity (GO:0004930), transmembrane signaling receptor activity (GO:0004888), signaling receptor activity (GO:0038023), molecular transducer activity (GO:0060089)

Table 4.1S. Top 20 negative and positive weighted motifs, with their PWM plot, distribution (average by base pair) around TSS and the top 5 associated GO terms. (Continued)

R8M3			system development (GO:0048731), regulation of molecular function (GO:0065009), protein modification process (GO:0036211), biological regulation (GO:0065007), cellular protein modification process (GO:0006464)	catalytic activity acting on a protein (GO:0140096), binding (GO:0005488), molecular_function (GO:0003674), protein binding (GO:0005515), ion binding (GO:0043167)
R8M7			signaling (GO:0023052), regulation of signaling (GO:0023051), regulation of cell communication (GO:0010646), regulation of molecular function (GO:0065009), system development (GO:0048731)	nucleoside-triphosphatase regulator activity (GO:0060589), GTPase regulator activity (GO:0030695), protein kinase activity (GO:0004672), GTPase activator activity (GO:0005096), protein serine/threonine/tyrosine kinase activity (GO:0004712)

Table 4.1S. Top 20 negative and positive weighted motifs, with their PWM plot, distribution (average by base pair) around TSS and the top 5 associated GO terms. (Continued)

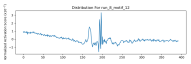
R8M12			G protein-coupled receptor signaling pathway (GO:0007186), detection of chemical stimulus involved in sensory perception (GO:0050907), detection of chemical stimulus (GO:0009593), sensory perception of chemical stimulus (GO:0007606), detection of stimulus involved in sensory perception (GO:0050906)	G protein-coupled receptor activity (GO:0004930), olfactory receptor activity (GO:0004984), transmembrane signaling receptor activity (GO:0004888), signaling receptor activity (GO:0038023), molecular transducer activity (GO:0060089)

Table 4.1S. Top 20 negative and positive weighted motifs, with their PWM plot, distribution (average by base pair) around TSS and the top 5 associated GO terms. (Continued)

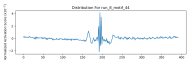
R8M44			sensory perception of chemical stimulus (GO:0007606), detection of chemical stimulus involved in sensory perception of smell (GO:0050911), detection of chemical stimulus involved in sensory perception (GO:0050907), detection of chemical stimulus (GO:0009593), sensory perception of smell (GO:0007608)	olfactory receptor activity (GO:0004984), G protein-coupled receptor activity (GO:0004930), signaling receptor activity (GO:0038023), molecular transducer activity (GO:0060089), transmembrane signaling receptor activity (GO:0004888)

Table 4.1S. Top 20 negative and positive weighted motifs, with their PWM plot, distribution (average by base pair) around TSS and the top 5 associated GO terms. (Continued)

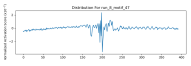
R8M47			multicellular organism development (GO:0007275), regulation of cell communication (GO:0010646), regulation of cellular process (GO:0050794), regulation of developmental process (GO:0050793), regulation of signaling (GO:0023051)	transcription regulator activity (GO:0140110), protein binding (GO:0005515), adenyl nucleotide binding (GO:0030554), binding (GO:0005488), olfactory receptor activity (GO:0004984)

Table 4.1S. Top 20 negative and positive weighted motifs, with their PWM plot, distribution (average by base pair) around TSS and the top 5 associated GO terms. (Continued)

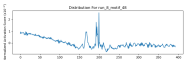
R8M48			detection of chemical stimulus involved in sensory perception (GO:0050907), detection of chemical stimulus (GO:0009593), detection of chemical stimulus involved in sensory perception of smell (GO:0050911), sensory perception of smell (GO:0007608), sensory perception of chemical stimulus (GO:0007606)	olfactory receptor activity (GO:0004984), G protein-coupled receptor activity (GO:0004930), transmembrane signaling receptor activity (GO:0004888), signaling receptor activity (GO:0038023), molecular transducer activity (GO:0060089)

Table 4.1S. Top 20 negative and positive weighted motifs, with their PWM plot, distribution (average by base pair) around TSS and the top 5 associated GO terms. (Continued)

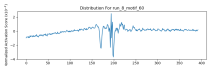
R8M60			negative regulation of cellular process (GO:0048523), regulation of molecular function (GO:0065009), regulation of cellular process (GO:0050794), negative regulation of biological process (GO:0048519), regulation of signal transduction (GO:0009966)	protein binding (GO:0005515), binding (GO:0005488), molecular_function (GO:0003674), regulation of signaling (GO:0023051), protein modification process (GO:0036211)

Table 4.1S. Top 20 negative and positive weighted motifs, with their PWM plot, distribution (average by base pair) around TSS and the top 5 associated GO terms. (Continued)

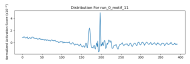
Positive Weight Motif Symbol	Motif PWM	Distribution	Biological Process	Molecular Function
R0M11			G protein-coupled receptor signaling pathway (GO:0007186), detection of chemical stimulus involved in sensory perception (GO:0050907), detection of chemical stimulus (GO:0009593), sensory perception of chemical stimulus (GO:0007606), detection of stimulus involved in sensory perception (GO:0050906)	G protein-coupled receptor activity (GO:0004930), transmembrane signaling receptor activity (GO:0004888), olfactory receptor activity (GO:0004984), signaling receptor activity (GO:0038023), molecular transducer activity (GO:0060089)

Table 4.1S. Top 20 negative and positive weighted motifs, with their PWM plot, distribution (average by base pair) around TSS and the top 5 associated GO terms. (Continued)

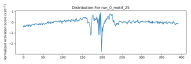
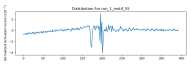
R0M25			regulation of cellular process (GO:0050794), regulation of developmental process (GO:0050793), regulation of biological process (GO:0050789), regulation of signaling (GO:0023051), regulation of cell communication (GO:0010646)	transcription regulator activity (GO:0140110), olfactory receptor activity (GO:0004984), nucleoside-triphosphatase regulator activity (GO:0060589), GTPase regulator activity (GO:0030695), guanyl-nucleotide exchange factor activity (GO:0005085)
R1M55			regulation of signaling (GO:0023051), regulation of cell communication (GO:0010646), regulation of cellular process (GO:0050794), regulation of signal transduction (GO:0009966), regulation of biological process (GO:0050789)	nucleoside-triphosphatase regulator activity (GO:0060589), GTPase regulator activity (GO:0030695), ion binding (GO:0043167), protein serine/threonine/tyrosine kinase activity (GO:0004712), guanyl-nucleotide exchange factor activity (GO:0005085)

Table 4.1S. Top 20 negative and positive weighted motifs, with their PWM plot, distribution (average by base pair) around TSS and the top 5 associated GO terms. (Continued)

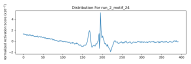
R2M24			detection of chemical stimulus (GO:0009593), detection of chemical stimulus involved in sensory perception (GO:0050907), detection of chemical stimulus involved in sensory perception of smell (GO:0050911), sensory perception of smell (GO:0007608), sensory perception of chemical stimulus (GO:0007606)	olfactory receptor activity (GO:0004984), G protein-coupled receptor activity (GO:0004930), transmembrane signaling receptor activity (GO:0004888), signaling receptor activity (GO:0038023), molecular transducer activity (GO:0060089)

Table 4.1S. Top 20 negative and positive weighted motifs, with their PWM plot, distribution (average by base pair) around TSS and the top 5 associated GO terms. (Continued)

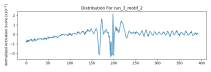
R3M2			regulation of developmental process (GO:0050793), regulation of cell communication (GO:0010646), regulation of signaling (GO:0023051), system development (GO:0048731), regulation of signal transduction (GO:0009966)	protein binding (GO:0005515), binding (GO:0005488), molecular_function (GO:0003674), transcription coregulator activity (GO:0003712), protein domain specific binding (GO:0019904)
R3M5			biological regulation (GO:0065007), regulation of signaling (GO:0023051), regulation of cell communication (GO:0010646), regulation of cellular process (GO:0050794), regulation of biological process (GO:0050789)	binding (GO:0005488), nucleoside-triphosphatase regulator activity (GO:0060589), protein serine/threonine/tyrosine kinase activity (GO:0004712), phosphotransferase activity alcohol group as acceptor (GO:0016773), protein binding (GO:0005515)

Table 4.1S. Top 20 negative and positive weighted motifs, with their PWM plot, distribution (average by base pair) around TSS and the top 5 associated GO terms. (Continued)

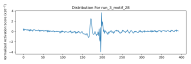
R3M28			detection of chemical stimulus (GO:0009593), detection of chemical stimulus involved in sensory perception (GO:0050907), sensory perception of chemical stimulus (GO:0007606), detection of chemical stimulus involved in sensory perception of smell (GO:0050911), detection of stimulus involved in sensory perception (GO:0050906)	olfactory receptor activity (GO:0004984), G protein-coupled receptor activity (GO:0004930), transmembrane signaling receptor activity (GO:0004888), signaling receptor activity (GO:0038023), molecular transducer activity (GO:0060089)

Table 4.1S. Top 20 negative and positive weighted motifs, with their PWM plot, distribution (average by base pair) around TSS and the top 5 associated GO terms. (Continued)

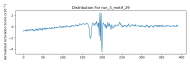
R3M29			tube development (GO:0035295), multicellular organism development (GO:0007275), system development (GO:0048731), tube morphogenesis (GO:0035239), regulation of cellular process (GO:0050794)	catalytic activity acting on a protein (GO:0140096) , growth factor binding (GO:0019838), binding (GO:0005488), metal ion binding (GO:0046872), cation binding (GO:0043169)

Table 4.1S. Top 20 negative and positive weighted motifs, with their PWM plot, distribution (average by base pair) around TSS and the top 5 associated GO terms. (Continued)

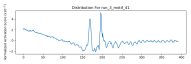
R3M41			detection of chemical stimulus (GO:0009593), detection of chemical stimulus involved in sensory perception (GO:0050907), sensory perception of chemical stimulus (GO:0007606), detection of stimulus (GO:0051606), detection of stimulus involved in sensory perception (GO:0050906)	G protein-coupled receptor activity (GO:0004930), olfactory receptor activity (GO:0004984), transmembrane signaling receptor activity (GO:0004888), signaling receptor activity (GO:0038023), molecular transducer activity (GO:0060089)

Table 4.1S. Top 20 negative and positive weighted motifs, with their PWM plot, distribution (average by base pair) around TSS and the top 5 associated GO terms. (Continued)

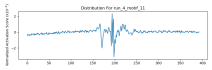
R4M0			biological regulation (GO:0065007), regulation of cellular process (GO:0050794), biological_process (GO:0008150), regulation of molecular function (GO:0065009), regulation of biological process (GO:0050789)	binding (GO:0005488), protein binding (GO:0005515), molecular_function (GO:0003674), protein kinase activity (GO:0004672), modification-dependent protein binding (GO:0140030)
R4M11			system development (GO:0048731), multicellular organism development (GO:0007275), developmental process (GO:0032502), anatomical structure development (GO:0048856), regulation of cellular process (GO:0050794)	enzyme regulator activity (GO:0030234), protein binding (GO:0005515), GTPase activator activity (GO:0005096), nucleoside-triphosphatase regulator activity (GO:0060589), olfactory receptor activity (GO:0004984)

Table 4.1S. Top 20 negative and positive weighted motifs, with their PWM plot, distribution (average by base pair) around TSS and the top 5 associated GO terms. (Continued)

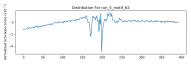
R5M62			regulation of molecular function (GO:0065009), system development (GO:0048731), regulation of cellular process (GO:0050794), biological regulation (GO:0065007), regulation of developmental process (GO:0050793)	protein binding (GO:0005515), binding (GO:0005488), molecular_function (GO:0003674), nucleoside-triphosphatase regulator activity (GO:0060589), protein phosphorylated amino acid binding (GO:0045309)
R6M8			biological regulation (GO:0065007), biological_process (GO:0008150), regulation of signaling (GO:0023051), regulation of cell communication (GO:0010646), regulation of cellular process (GO:0050794)	protein binding (GO:0005515), binding (GO:0005488), molecular_function (GO:0003674), metal ion binding (GO:0046872), catalytic activity acting on a protein (GO:0140096)

Table 4.1S. Top 20 negative and positive weighted motifs, with their PWM plot, distribution (average by base pair) around TSS and the top 5 associated GO terms. (Continued)

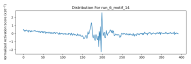
R6M14			G protein-coupled receptor signaling pathway (GO:0007186), detection of chemical stimulus involved in sensory perception (GO:0050907), detection of chemical stimulus (GO:0009593), sensory perception of chemical stimulus (GO:0007606), detection of stimulus involved in sensory perception (GO:0050906)	G protein-coupled receptor activity (GO:0004930), olfactory receptor activity (GO:0004984), transmembrane signaling receptor activity (GO:0004888), signaling receptor activity (GO:0038023), molecular transducer activity (GO:0060089)

Table 4.1S. Top 20 negative and positive weighted motifs, with their PWM plot, distribution (average by base pair) around TSS and the top 5 associated GO terms. (Continued)

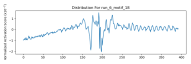
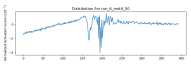
R6M18			regulation of molecular function (GO:0065009), system development (GO:0048731), nervous system development (GO:0007399), biological regulation (GO:0065007), negative regulation of cellular process (GO:0048523)	nucleoside-triphosphatase regulator activity (GO:0060589), protein binding (GO:0005515), protein kinase activity (GO:0004672), binding (GO:0005488), molecular_function (GO:0003674)
R6M50			regulation of signaling (GO:0023051), regulation of cell communication (GO:0010646), regulation of cellular process (GO:0050794), signaling (GO:0023052), regulation of signal transduction (GO:0009966)	protein serine/threonine/tyrosine kinase activity (GO:0004712), protein serine kinase activity (GO:0106310), protein kinase activity (GO:0004672), phosphotransferase activity alcohol group as acceptor (GO:0016773) , molecular_function (GO:0003674)

Table 4.1S. Top 20 negative and positive weighted motifs, with their PWM plot, distribution (average by base pair) around TSS and the top 5 associated GO terms. (Continued)

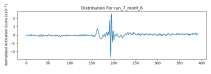
R7M6			sensory perception of chemical stimulus (GO:0007606), detection of chemical stimulus involved in sensory perception of smell (GO:0050911), detection of chemical stimulus involved in sensory perception (GO:0050907), detection of chemical stimulus (GO:0009593), sensory perception of smell (GO:0007608)	olfactory receptor activity (GO:0004984), signaling receptor activity (GO:0038023), molecular transducer activity (GO:0060089), transmembrane signaling receptor activity (GO:0004888), G protein-coupled receptor activity (GO:0004930)

Table 4.1S. Top 20 negative and positive weighted motifs, with their PWM plot, distribution (average by base pair) around TSS and the top 5 associated GO terms. (Continued)

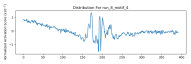
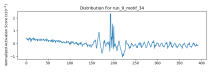
R8M4			detection of chemical stimulus involved in sensory perception (GO:0050907), detection of chemical stimulus (GO:0009593), G protein-coupled receptor signaling pathway (GO:0007186), sensory perception of chemical stimulus (GO:0007606), detection of stimulus involved in sensory perception (GO:0050906)	G protein-coupled receptor activity (GO:0004930), olfactory receptor activity (GO:0004984), transmembrane signaling receptor activity (GO:0004888), signaling receptor activity (GO:0038023), molecular transducer activity (GO:0060089)

Table 4.1S. Top 20 negative and positive weighted motifs, with their PWM plot, distribution (average by base pair) around TSS and the top 5 associated GO terms. (Continued)

R9M34			multicellular organismal process (GO:0032501), nervous system development (GO:0007399), multicellular organism development (GO:0007275), developmental process (GO:0032502), system process (GO:0003008)	olfactory receptor activity (GO:0004984), signaling receptor activity (GO:0038023), transmembrane signaling receptor activity (GO:0004888), molecular transducer activity (GO:0060089), DNA-binding transcription activator activity (GO:0001216)
R9M39			multicellular organism development (GO:0007275), system development (GO:0048731), nervous system development (GO:0007399), anatomical structure development (GO:0048856), regulation of signaling (GO:0023051)	binding (GO:0005488), kinase activity (GO:0016301), transferase activity transferring phosphorus-containing groups (GO:0016772), protein kinase activity (GO:0004672), phosphotransferase activity alcohol group as acceptor (GO:0016773)

4.9 References

1. Zhou, J. & Troyanskaya, O. G. Predicting effects of noncoding variants with deep learning–based sequence model. *Nat. Methods* **12,** 931–934 (2015).
2. Avsec, Ž., Agarwal, V., Visentin, D., Ledsam, J. R., Grabska-Barwinska, A., Taylor, K. R., Assael, Y., Jumper, J., Kohli, P. & Kelley, D. R. Effective gene expression prediction from sequence by integrating long-range interactions. *Nat. Methods* **18,** 1196–1203 (2021).
3. Kelley, D. R., Reshef, Y. A., Bileschi, M., Belanger, D., McLean, C. Y. & Snoek, J. Sequential regulatory activity prediction across chromosomes with convolutional neural networks. *Genome Res.* **28,** 739–750 (2018).
4. Avsec, Ž., Weilert, M., Shrikumar, A., Krueger, S., Alexandari, A., Dalal, K., Fropf, R., McAnany, C., Gagneur, J., Kundaje, A. & Zeitlinger, J. Base-resolution models of transcription-factor binding reveal soft motif syntax. *Nat. Genet.* **53,** 354–366 (2021).
5. Agarwal, V. & Shendure, J. Predicting mRNA Abundance Directly from Genomic Sequence Using Deep Convolutional Neural Networks. *Cell Rep.* **31,** 107663 (2020).
6. Liu, C. & Wang, W. Contextual Regression: An Accurate and Conveniently Interpretable Nonlinear Model for Mining Discovery from Scientific Data. doi:10.1101/210997
7. Alipanahi, B., Delong, A., Weirauch, M. T. & Frey, B. J. Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nat. Biotechnol.* **33,** 831–838 (2015).
8. Lecun, Y., Bottou, L., Bengio, Y. & Haffner, P. Gradient-based learning applied to document recognition. *Proc. IEEE* **86,** 2278–2324 (1998).
9. Zhang, Y., Tino, P., Leonardis, A. & Tang, K. A Survey on Neural Network Interpretability.

IEEE Transactions on Emerging Topics in Computational Intelligence **5,** 726–742 (2021).

10. Ribeiro, M. T., Singh, S. & Guestrin, C. 'why should I trust you?': Explaining the predictions of any classifier. *arXiv [cs.LG]* (2016). doi:10.1145/2939672.2939778
11. Lundberg, S. & Lee, S.-I. A unified approach to interpreting model predictions. *arXiv [cs.AI]* (2017). at <http://arxiv.org/abs/1705.07874>
12. Shrikumar, A., Greenside, P., Shcherbina, A. & Kundaje, A. Not Just a Black Box: Learning Important Features Through Propagating Activation Differences. *arXiv [cs.LG]* (2016). at <http://arxiv.org/abs/1605.01713>
13. Schmidt, M. & Lipson, H. Distilling free-form natural laws from experimental data. *Science* **324,** 81–85 (2009).
14. Udrescu, S.-M. & Tegmark, M. AI Feynman: A physics-inspired method for symbolic regression. *Sci Adv* **6,** eaay2631 (2020).
15. Basan, M., Hui, S., Okano, H., Zhang, Z., Shen, Y., Williamson, J. R. & Hwa, T. Overflow metabolism in Escherichia coli results from efficient proteome allocation. *Nature* **528,** 99–104 (2015).
16. Mori, M., Schink, S., Erickson, D. W., Gerland, U. & Hwa, T. Quantifying the benefit of a proteome reserve in fluctuating environments. *Nature Communications* **8,** (2017).
17. Dai, X., Zhu, M., Warren, M., Balakrishnan, R., Patsalo, V., Okano, H., Williamson, J. R., Fredrick, K., Wang, Y.-P. & Hwa, T. Reduction of translating ribosomes enables Escherichia coli to maintain elongation rates during slow growth. *Nat Microbiol* **2,** 16231 (2016).
18. You, C., Okano, H., Hui, S., Zhang, Z., Kim, M., Gunderson, C. W., Wang, Y.-P., Lenz, P., Yan, D. & Hwa, T. Coordination of bacterial proteome with metabolism by cyclic AMP

signalling. *Nature* **500,** 301–306 (2013).

19. Haberle, V. & Stark, A. Eukaryotic core promoters and the functional basis of transcription initiation. *Nat. Rev. Mol. Cell Biol.* **19,** 621–637 (2018).
20. Roy, A. L. & Singer, D. S. Core promoters in transcription: old problem, new insights. *Trends Biochem. Sci.* **40,** 165–171 (2015).
21. Landolin, J. M., Johnson, D. S., Trinklein, N. D., Aldred, S. F., Medina, C., Shulha, H., Weng, Z. & Myers, R. M. Sequence features that drive human promoter function and tissue specificity. *Genome Res.* **20,** 890–898 (2010).
22. Sopko, R., Huang, D., Preston, N., Chua, G., Papp, B., Kafadar, K., Snyder, M., Oliver, S. G., Cyert, M., Hughes, T. R., Boone, C. & Andrews, B. Mapping pathways and phenotypes by systematic gene overexpression. *Mol. Cell* **21,** 319–330 (2006).
23. Pesole, Régnier, Simonis & Sinha. Assessing computational tools for the discovery of transcription factor binding sites. *Nature* at <https://idp.nature.com/authorize/casa?redirect_uri=https://www.nature.com/articles/nbt1053&casa_token=Kz505hGUbGUAAAAA:BRp0dTSDgwpNPe9Lq3uuHfkFqyyomomqEvo6qBkoYQ6ivwCdhiJEOV1WvfSl8K8Aa5Lx25-LHMqxHnRleg>
24. Carey, L. B., van Dijk, D., Sloot, P. M. A., Kaandorp, J. A. & Segal, E. Promoter sequence determines the relationship between expression level and noise. *PLoS Biol.* **11,** e1001528 (2013).
25. Mogno, I., Kwasnieski, J. C. & Cohen, B. A. Massively parallel synthetic promoter assays reveal the in vivo effects of binding site variants. *Genome Res.* **23,** 1908–1915 (2013).
26. Aysha, J., Noman, M., Wang, F., Liu, W., Zhou, Y., Li, H. & Li, X. Synthetic Promoters:

Designing the cis Regulatory Modules for Controlled Gene Expression. *Mol. Biotechnol.* **60,** 608–620 (2018).

27. Vaishnav, E. D., de Boer, C. G., Molinet, J., Yassour, M., Fan, L., Adiconis, X., Thompson, D. A., Levin, J. Z., Cubillos, F. A. & Regev, A. The evolution, evolvability and engineering of gene regulatory DNA. *Nature* **603,** 455–463 (2022).
28. Weingarten-Gabbay, S. & Segal, E. The grammar of transcriptional regulation. *Hum. Genet.* **133,** 701–711 (2014).
29. de Boer, C. G., Vaishnav, E. D., Sadeh, R., Abeyta, E. L., Friedman, N. & Regev, A. Deciphering eukaryotic gene-regulatory logic with 100 million random promoters. *Nat. Biotechnol.* **38,** 56–65 (2020).
30. Ngoc, L. V., Huang, C. Y., Cassidy, C. J., Medrano, C. & Kadonaga, J. T. Identification of the human DPR core promoter element using machine learning. *Nature* **585,** 459–463 (2020).
31. Rochman, M., Aviv, M., Glaser, G. & Muskhelishvili, G. Promoter protection by a transcription factor acting as a local topological homeostat. *EMBO Rep.* **3,** 355–360 (2002).
32. Ward, L. D. & Bussemaker, H. J. Predicting functional transcription factor binding through alignment-free and affinity-based analysis of orthologous promoter sequences. *Bioinformatics* **24,** i165–71 (2008).
33. Haas, Pagie, Sluimer & Bussemaker. Genome-wide mapping of autonomous promoter activity in human cells. *Nature* at <https://idp.nature.com/authorize/casa?redirect_uri=https://www.nature.com/articles/nbt.3754&casa_token=aVJHSitOFegAAAAA:whK2TddANGE6cMwNO8nCXYBbnxk7i7kjZdsz

FA8HJW0b9PUoIZKHQPUXuhjQZs6rEi_8-VlpRiEnJjVw8w>

34. Marchal, Huang, Mordelet & Hartemink. Inferring gene expression from ribosomal promoter sequences, a crowdsourcing approach. *Genome* at <https://genome.cshlp.org/content/23/11/1928.short>

35. Bergman, D. T., Jones, T. R., Liu, V., Ray, J., Jagoda, E., Siraj, L., Kang, H. Y., Nasser, J., Kane, M., Rios, A., Nguyen, T. H., Grossman, S. R., Fulco, C. P., Lander, E. S. & Engreitz, J. M. Compatibility rules of human enhancer and promoter sequences. *Nature* **607,** 176–184 (2022).

36. van Arensbergen, J., FitzPatrick, V. D., de Haas, M., Pagie, L., Sluimer, J., Bussemaker, H. J. & van Steensel, B. Genome-wide mapping of autonomous promoter activity in human cells. *Nat. Biotechnol.* **35,** 145–153 (2017).

37. Schmidhuber, J. Deep learning in neural networks: an overview. *Neural Netw.* **61,** 85–117 (2015).

38. de Jongh, R. P. H., van Dijk, A. D. J., Julsing, M. K., Schaap, P. J. & de Ridder, D. Designing Eukaryotic Gene Expression Regulation Using Machine Learning. *Trends Biotechnol.* **38,** 191–201 (2020).

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

40. Gordon, A. D., Breiman, L., Friedman, J. H., Olshen, R. A. & Stone, C. J. Classification and Regression Trees. *Biometrics* **40,** 874 (1984).

41. Weirauch, M. T., Yang, A., Albu, M., Cote, A. G., Montenegro-Montero, A., Drewe, P., Najafabadi, H. S., Lambert, S. A., Mann, I., Cook, K., Zheng, H., Goity, A., van Bakel, H., Lozano, J.-C., Galli, M., Lewsey, M. G., Huang, E., Mukherjee, T., Chen, X., Reece-Hoyes, J. S., Govindarajan, S., Shaulsky, G., Walhout, A. J. M., Bouget, F.-Y., Ratsch, G., Larrondo, L. F., Ecker, J. R. & Hughes, T. R. Determination and inference of eukaryotic transcription factor sequence specificity. *Cell* **158,** 1431–1443 (2014).

42. Mi, H., Muruganujan, A., Casagrande, J. T. & Thomas, P. D. Large-scale gene function analysis with the PANTHER classification system. *Nat. Protoc.* **8,** 1551–1566 (2013).

43. Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M. & Sherlock, G. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* **25,** 25–29 (2000).

44. The Gene Ontology resource: enriching a GOld mine. *Nucleic Acids Res.* **49,** D325–D334 (2021).

45. Wang, J., Liu, C., Chen, Y. & Wang, W. Taiji-reprogram: a framework to uncover cell-type specific regulators and predict cellular reprogramming cocktails. *NAR Genom Bioinform* **3,** lqab100 (2021).

46. Roadmap Epigenomics Consortium, Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., Heravi-Moussavi, A., Kheradpour, P., Zhang, Z., Wang, J., Ziller, M. J., Amin, V.,

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

47. ENCODE Project Consortium, Moore, J. E., Purcaro, M. J., Pratt, H. E., Epstein, C. B., Shoresh, N., Adrian, J., Kawli, T., Davis, C. A., Dobin, A., Kaul, R., Halow, J., Van Nostrand, E. L., Freese, P., Gorkin, D. U., Shen, Y., He, Y., Mackiewicz, M., Pauli-Behn, F., Williams, B. A., Mortazavi, A., Keller, C. A., Zhang, X.-O., Elhajjajy, S. I., Huey, J., Dickel, D. E., Snetkova, V., Wei, X., Wang, X., Rivera-Mulia, J. C., Rozowsky, J., Zhang, J., Chhetri, S. B., Zhang, J., Victorsen, A., White, K. P., Visel, A., Yeo, G. W., Burge, C. B., Lécuyer, E., Gilbert, D. M., Dekker, J., Rinn, J., Mendenhall, E. M., Ecker, J. R., Kellis, M., Klein, R. J., Noble, W. S., Kundaje, A., Guigó, R., Farnham, P. J., Cherry, J. M., Myers, R.

M., Ren, B., Graveley, B. R., Gerstein, M. B., Pennacchio, L. A., Snyder, M. P., Bernstein, B. E., Wold, B., Hardison, R. C., Gingeras, T. R., Stamatoyannopoulos, J. A. & Weng, Z. Expanded encyclopaedias of DNA elements in the human and mouse genomes. *Nature* **583,** 699–710 (2020).