Biologically-relevant transfer learning improves transcription factor binding prediction

Gherman Novakovsky [*1] **Manu Saraswat** [*1] **Oriol Fornes** [*1] **Sara Mostafavi** [1 2 3] **Wyeth W. Wasserman** [1]

Abstract

Identifying the genome-wide locations where transcription factors (TF) bind is key to understanding gene regulation. Provided that there is sufficient data for training, deep learning models such as convolutional neural networks (CNN) are powerful tools for TF binding prediction. However, the amount of available binding data (i.e. ChIP-seq) for many TFs is sparse. Transfer learning has been shown to reduce the amount of data needed for training and improve model performance in different biological tasks, including TF binding prediction. Here, we present a comprehensive analysis of transfer learning for TF binding prediction. Using a state-of-the-art CNN architecture, we show that biologically-relevant prior knowledge, specifically of TFs with the same DNA-binding motifs, correlated binding, or functionally associated with the target TF, improves transfer learning. Moreover, we demonstrate that with transfer learning, we can train good performing models (Matthews correlation coefficient greater than 0.5) from less than 500 ChIP-seq peaks. Finally, using model interpretation techniques, we observe that the mechanism of transfer learning involves refining the CNN filters learnt during the pre-training step to resemble the binding motif(s) of the target TF. Our results confirm that transfer learning is a powerful technique to improve TF binding prediction.

[*]Equal contribution [1]Centre for Molecular Medicine and Therapeutics, Department of Medical Genetics, BC Children's Hospital Research Institute, University of British Columbia, Vancouver, BC V5Z 4H4, Canada [2]Department of Statistics, University of British Columbia, Vancouver, BC [3]Canadian Institute for Advanced Research, CIFAR AI Chair, and Child and Brain Development, Toronto, Canada, M5G 1M1 . Correspondence to: Oriol Fornes <oriol@cmmt.ubc.ca>, Wyeth W. Wasserman <wyeth@cmmt.ubc.ca>.

Preliminary work. Under review at the ICML 2020 Workshop on Computational Biology (WCB). Do not distribute.

1. Introduction

Transcription factors (TFs) are the main regulators of gene expression at the transcriptional level(Lambert et al., 2018). TFs bind to specific genomic locations known as TF binding sites (TFBSs) through the recognition of degenerate motifs (Badis et al., 2009) approximately 10 base pairs (bp) in length(Stewart et al., 2012). Due to their central role in gene regulation, disrupted TFs and TFBSs have been associated with many disorders(Lee et al., 2020b; Mathelier et al., 2015), including cancer(Khurana et al., 2016). Therefore, delineating the genome-wide locations to which TFs bind would help to understand how genes are regulated in health and disease. Chromatin immunoprecipitation followed by sequencing (ChIP-seq) is an experimental assay that enables the identification of TF-bound regions in vivo at a resolution of a few hundred bp(Johnson et al., 2007). These regions, known as ChIP-seq peaks, are expected to encompass the TFBSs. The ReMap database provides access to thousands of uniformly reprocessed ChIP-seq datasets(Cheneby et al., 2018; 2020). It stores millions of ChIP-seq peaks related to the binding of approximately 800 human TFs in 602 different cell and tissue types. Based on ReMap, the UniBind database provides reliable TFBS predictions within the ChIP-seq peaks of 231 human TFs in 315 different cell and tissue types (Gheorghe et al., 2019). UniBind TFBS predictions are based on four different computational models, including position weight matrices (PWMs).

Despite efforts by public consortia such as ENCODE(Dunham et al., 2012) to delineate the binding of each TF in the human genome, the task, if feasible, is far from complete. For instance, about 40% of human TFs have not been profiled by ChIP-seq, and only a few, such as CTCF, have been profiled extensively. In this scenario, there is a need for computational methods capable of predicting TF binding with high precision to complement experimental data. Driven by advances in the field of deep learning, computational prediction of TF binding has improved dramatically(Koo & Ploenzke, 2020). In recent years, several deep learning approaches have emerged that exploit convolutional neural networks (CNNs) for TF binding prediction(Alipanahi et al., 2015; Avsec et al., 2020; Lan et al., 2019; Quang & Xie, 2016; 2019; Wang et al.,

2018; Zheng et al., 2020; Zhou & Troyanskaya, 2015). A limitation of deep learning models, including CNNs, is the availability of sufficient training data. The amount of ChIP-seq data available for some TFs is very small. For example, 381 (47.6%) of the human TFs in ReMap have been profiled in only one cell or tissue type, while 134 (16.7%) have less than 1,000 ChIP-seq peaks annotated. Transfer learning—the use of knowledge acquired while solving a task, to solve a different but related task—can reduce the amount of data required for training(Thrun & Pratt, 1998). Transfer learning has been successful in different biological tasks such as reconstructing gene regulatory networks(Mignone et al., 2020; Yang et al., 2019), denoising single-cell data(Wang et al., 2019a;b), or predicting both chromatin interactions(Schwessinger et al., 2019) and TF-bound regions(Lan et al., 2019; Zheng et al., 2020). Here, we perform an in-depth analysis of transfer learning for TF binding prediction using a popular CNN architecture for inferring chromatin accessibility(Kelley et al., 2016; Maslova et al., 2019) adapted to predict TF binding events. We begin by combining TF binding data from ReMap and UniBind to build a sparse matrix summarizing the binding of TFs to DNase I hypersensitive sites(DHSs) (Lee et al., 2020a; Thurman et al., 2012) in a cell and tissue type agnostic manner. Next, we train a multi-task model to predict the binding of 50 TFs and use the learnt weights to initialize models for individual TFs (i.e. transfer learning). We show that transfer learning significantly improves model performance even for TFs with small datasets. Moreover, we see that the benefit of transfer learning is greater when the multi-model is trained on biologically-relevant TFs; specifically, TFs with the same DNA-binding motifs, correlated binding, or functionally associated with the target TF. Finally, we apply model interpretation techniques in an attempt to decipher the mechanism of transfer learning. We observe that the filters learnt by the first layer of the multi-model during the pre-training step are refined to resemble the motif of the target TF. To the best of our knowledge, this is the first comprehensive study of its kind.

2. Methods

2.1. TF binding matrix

As source of accessible regions, we used DHSs from the UCSC Genome Browser track of ENCODE DNase I hypersensitivity peak clusters(Lee et al., 2020a; Thurman et al., 2012). DHSs were resized to 200 bp by extending the center of each cluster 100 bp in each direction using bedtools slop(Quinlan & Hall, 2010). As sources of TF binding features, we used ChIP-seq peak summits from ReMap 2018(Cheneby et al., 2018) and PWM-based TFBS predictions from UniBind(Gheorghe et al., 2019). All data were matched by cell and tissue type.

Next, we built a sparse matrix by aggregating the binding of 163 TFs to 1,817,918 accessible regions in a cell and tissue type agnostic manner, with rows representing TFs and columns regions. Each cell in the matrix takes one of three values: "1", if the region is bound by the TF; "0", if the region is not bound by the TF; or "null" indicating that the binding of the TF to the region cannot be resolved. A region is defined as bound by a TF if it is accessible and overlaps with binding features of the TF from ReMap and UniBind in at least one matched sample. Instead, if the region is accessible but does not overlap with binding features of the TF in any matched sample, it is defined as not bound. Other possibilities, for instance, if the region is not accessible in any matched samples with the TF, or if it is accessible but only overlaps with ReMap or UniBind binding features of the TF (not both), are defined as not resolved (i.e. there is not enough evidence to support whether the region is bound or not by the TF).

2.2. Model architecture and training

We adapted the CNN architecture from Basset(Kelley et al., 2016) and AI-TAC(Maslova et al., 2019) to predict TF binding events: three convolutional layers, each followed by batch normalization, ReLU activation function and max pooling, followed by three fully connected layers, two hidden and the last one yielding 1, 5 or 50 outputs.

To train multi-task models (i.e. multi-models), we re-define the TF binding matrix to make it less sparse. Specifically, accessible regions overlapping with either ReMap ChIP-seq peaks or UniBind TFBSs of a TF are defined as unbound (i.e. 0) rather than unresolved (i.e. null). We select a slice of the matrix such that it contains the maximum number of resolved regions for all TFs in the multi-model, and split it into training (80%), validation (10%), and testing (10%). We apply one-hot encoding to convert nucleotides into 4-element vectors as in Basset. Regions with one or more Ns are ignored. The model is trained with Adam optimizer(Kingma & Ba, 2017) on both strands of each region. We set the learning rate to 0.003, the batch size to 100, and use an early stopping criteria to avoid overfitting (i.e. when the model performance on the validation set does not improve). Single-task models (i.e. individual models) are trained in a similar way, but using the original sparse matrix and ignoring unresolved regions.

2.3. Transfer learning

We used a two-step transfer learning approach consisting of pre-training and fine-tuning. In the pre-training step, we train a multi-model that predicts the binding of 5 or 50 TFs, depending on the case. Then, we initialize an individual model that predicts the binding of a single TF (i.e. target) by transferring all of the layers learned by the multi-model,

except the output layer. In the fine-tuning step, we reduce the learning rate to train the model of the target TF.

2.4. Results

2.4.1. Transfer learning improves TF binding prediction

From the total of 163 TFs present in the TF binding matrix, we selected the top 50 based on their number of resolved regions. We trained a multi-model to predict the binding of these 50 TFs, as well as 50 individual models with and without transfer learning from the multi-model. To allow for a fair comparison, the multi-model and the individual models were trained on different regions. Moreover, both types of individual models of a TF relied on the same regions for training, validation and testing. Transfer learning improved model performance for all TFs (Figure 1A). The improvement was inversely correlated with the number of 1s in the training data size (i.e. TFs with fewer bound regions benefited the most).

Next, we wondered what is the minimum dataset size required to achieve good model performance with transfer learning. We focused on SPI1. We trained models with and without transfer learning 5 times by downsampling 1,000 and 500 bound/unbound regions of SPI1 from the TF binding matrix at random, while accounting for %GC content. Surprisingly, models trained with transfer learning achieved a good performance (Matthews correlation coefficient $\geq$ 0.5) even when trained on 500 bound regions (Figure 1B). A similar trend was observed for other TFs (results not shown).

2.4.2. Biologically-relevant prior knowledge improves transfer learning

It has been shown that TFs from the same structural family share similar DNA-binding mechanisms(Badis et al., 2009) (i.e. binding modes). For instance, members of the T-box family of TFs bind to the consensus DNA sequence TCACACCT(Wilson & Conlon, 2002). We wondered whether binding mode information could explain the observed differences in transfer learning performance for different TFs.

To answer this question, we trained individual models, with and without transfer learning, to predict the binding of 98 additional TFs from the TF binding matrix (15 TFs were discarded due to their small number of bound regions in the matrix). The 50 TFs from the multi-model are represented in JASPAR(Fornes et al., 2020) by 34 unique binding modes; the remaining 98 TFs are represented by one of these 34 or a different binding mode. We observed that TFs represented by one of the 34 binding modes from the multi-model benefited from transfer learning significantly more (Welch t-test, p-value = 0.01; Figure 2A).

Next, we wondered if other biologically-relevant prior

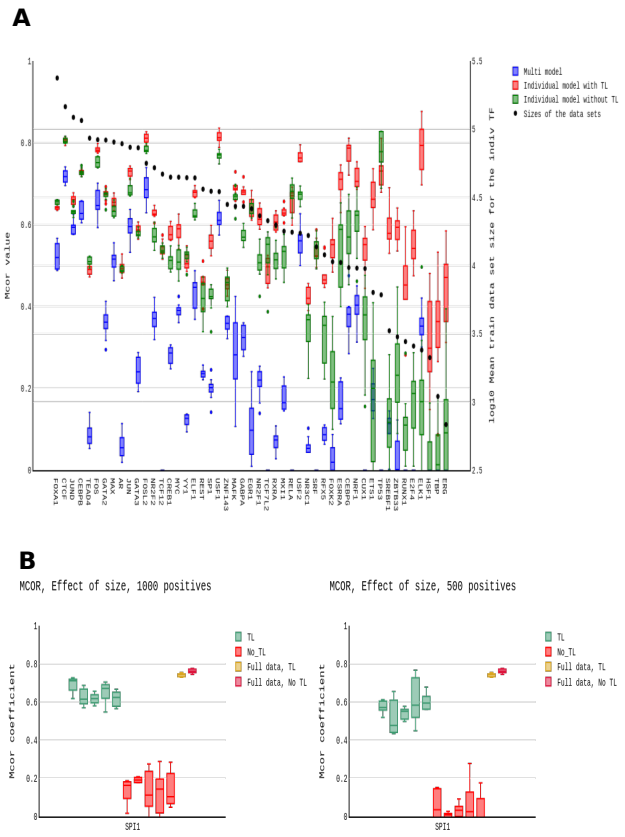

Figure 1. A) Performance of individual models and multi-model on the 50 TFs used to train the multi-model. Every individual model was either trained from scratch or using pre-initialized weights (i.e. TL) from the multi-model. B) Performance of SPI1 on 1000 and 500 training examples with and without TL. For comparison, performance by training on entire SPI1 dataset is also shown.

knowledge, such as from cooperative TFs (i.e. cofactors), which we define as TFs whose binding is positively correlated with the target TF, or functional partners from STRING (Szklarczyk et al., 2019) also had a positive effect on transfer learning. We focused on 5 TFs from the multi-model represented by different binding modes: HNFA4, JUND, MAX, SP1 and SPI1. For each TF, we trained 5 multi-models: 1) on 5 TFs with the same binding mode as the target TF; 2) on 5 co-factors with binding modes different from that of the target TF; 3) on 5 functional partners of the target TF from STRING; 4) on 5 TFs represented by the same binding mode but whose binding is not correlated with the target TF; and 5) on 5 randomly selected TFs with binding modes different from that of the target TF. Moreover, we trained 5 more multi-models in which one of the 5 TFs was replaced by the target TF. These multi-models

were used for transfer learning. Except for SP1, transfer learning from multi-models of TFs represented by the same binding mode as the target TF (i.e. 1) as well as co-factors (i.e. 2) performed well regardless of the presence of the target TF in the multi-model (Figure 2B). Interestingly, for SP1, transfer learning from multi-models trained with co-factors and functional partners from STRING (i.e. 2 and 3) performed better than the multi-models trained with other Krüppel-like factors(Swamynathan, 2010) (i.e. 1 and 4), which performed as badly as random (i.e. 5).

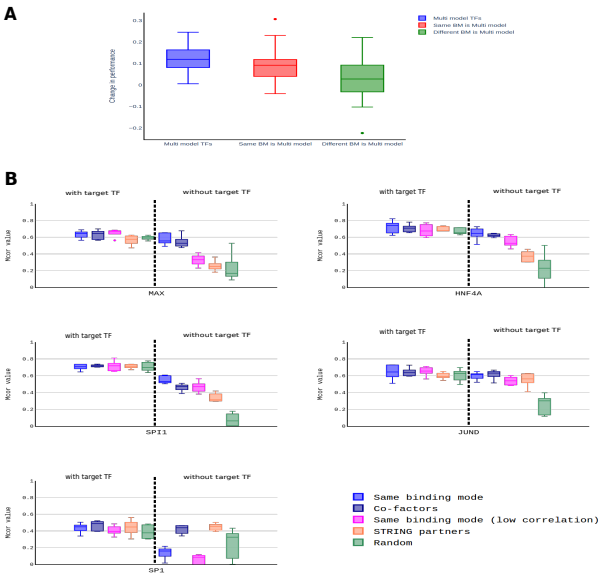

Figure 2. A) Effect of binding modes on TL performance improvement for TFs subsampled to 1000 positive/negative examples. Presence of a TF with the same binding mode as the target TF, results in better performance. B) Results of different TL initialisation. The presence of the target TF during the pre-training step provides good results on the test set. In the absence of the target TF, pre-training with the same binding mode TFs or cofactors gives better performance.

2.4.3. Transfer Learning Interpretation

To understand how transfer learning works, we converted the filters from the first layer of the multi-model into PWMs and compared them to TF binding profiles from the JASPAR database using TOMTOM(Gupta et al., 2007) (Figure 3A). As expected, more than half had significant similarities to known motifs. Next, we wondered if this interpretation technique could reveal the mechanisms of transfer learning. We focused on HNF4A. Briefly, we initialized an individual model for HNF4A with transfer learning from the original multi-model trained on 50 TFs. Before the fine-tuning step, the filters from the first layer of both the individual and the multi-model are identical. After the fine-tuning step, we found that some of the filters, which in the multi-model resembled a different TF or did not match any TF motif at all, became similar to the target TF (Figure 3B). Noteworthy, after transfer learning, we observed a large number of filters begin to resemble the target TF, compared to training the model from scratch without transfer learning (results not shown). Taken together, our findings suggest that the mechanism by which transfer learning improves model performance may be that pre-trained layers provide a good starting point for the new model to learn the relevant motifs.

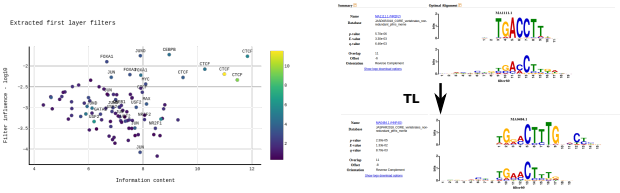

Figure 3. A) The multimodel for 50 TFs learns motif representation of the target TFs in its first layer; color scale shows the statistical significance of the resemblance to JASPAR motif database motif instances (-log10(q.value)). B) Fine-tuning of the multi-model filters on a new target task results in the refining of certain filters, which at the end better resemble the target TF binding site.

References

Alipanahi, B., Delong, A., Weirauch, M. T., and Frey, B. J. Predicting the sequence specificities of dna- and rna-binding proteins by deep learning. *Nature Biotechnology*, 33(88):831–838, Aug 2015. ISSN 1546-1696. doi: 10.1038/nbt.3300.

Avsec, , Weilert, M., Shrikumar, A., Krueger, S., Alexandari, A., Dalal, K., Fropf, R., McAnany, C., Gagneur, J., Kundaje, A., and et al. Deep learning at base-resolution reveals cis-regulatory motif syntax. *bioRxiv*, pp. 737981, Mar 2020. doi: 10.1101/737981.

Badis, G., Berger, M. F., Philippakis, A. A., Talukder, S., Gehrke, A. R., Jaeger, S. A., Chan, E. T., Metzler, G., Vedenko, A., Chen, X., et al. Diversity and complexity in dna recognition by transcription factors. *Science*, 324 (5935):1720–1723, 2009.

Cheneby, J., Gheorghe, M., Artufel, M., Mathelier, A., and Ballester, B. Remap 2018: an updated atlas of regulatory regions from an integrative analysis of dna-binding chip-seq experiments. *Nucleic Acids Research*, 46(D1):D267–D275, Jan 2018. ISSN 0305-1048. doi: 10.1093/nar/gkx1092.

Cheneby, J., Menetrier, Z., Mestdagh, M., Rosnet, T., Douida, A., Rhalloussi, W., Bergon, A., Lopez, F., and

Ballester, B. Remap 2020: a database of regulatory regions from an integrative analysis of human and arabidopsis dna-binding sequencing experiments. *Nucleic Acids Research*, 48(D1):D180–D188, Jan 2020. ISSN 0305-1048. doi: 10.1093/nar/gkz945.

Dunham, I., Kundaje, A., Aldred, S. F., Collins, P. J., Davis, C. A., Doyle, F., Epstein, C. B., Frietze, S., Harrow, J., Kaul, R., and et al. An integrated encyclopedia of dna elements in the human genome. *Nature*, 489 (74147414):57–74, Sep 2012. ISSN 1476-4687. doi: 10.1038/nature11247.

Fornes, O., Castro-Mondragon, J. A., Khan, A., van der Lee, R., Zhang, X., Richmond, P. A., Modi, B. P., Correard, S., Gheorghe, M., Baranašić, D., et al. Jaspar 2020: update of the open-access database of transcription factor binding profiles. *Nucleic acids research*, 48(D1):D87–D92, 2020.

Gheorghe, M., Sandve, G. K., Khan, A., Cheneby, J., Ballester, B., and Mathelier, A. A map of direct tf–dna interactions in the human genome. *Nucleic Acids Research*, 47(4):e21–e21, Feb 2019. ISSN 0305-1048. doi: 10.1093/nar/gky1210.

Gupta, S., Stamatoyannopoulos, J. A., Bailey, T. L., and Noble, W. S. Quantifying similarity between motifs. *Genome Biology*, 8(2):R24, Feb 2007. ISSN 1474-760X. doi: 10.1186/gb-2007-8-2-r24.

Johnson, D. S., Mortazavi, A., Myers, R. M., and Wold, B. Genome-wide mapping of in vivo protein-dna interactions. *Science*, 316(5830):1497–1502, Jun 2007. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.1141319.

Kelley, D. R., Snoek, J., and Rinn, J. L. Basset: learning the regulatory code of the accessible genome with deep convolutional neural networks. *Genome Research*, 26(7): 990–999, Jan 2016. ISSN 1088-9051, 1549-5469. doi: 10.1101/gr.200535.115. Company: Cold Spring Harbor Laboratory PressDistributor: Cold Spring Harbor Laboratory PressInstitution: Cold Spring Harbor Laboratory PressLabel: Cold Spring Harbor Laboratory Presspublisher: Cold Spring Harbor LabPMID: 27197224.

Khurana, E., Fu, Y., Chakravarty, D., Demichelis, F., Rubin, M. A., and Gerstein, M. Role of non-coding sequence variants in cancer. *Nature Reviews Genetics*, 17(22): 93–108, Feb 2016. ISSN 1471-0064. doi: 10.1038/nrg.2015.17.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. *arXiv:1412.6980 [cs]*, Jan 2017. URL http://arxiv.org/abs/1412.6980. arXiv: 1412.6980.

Koo, P. K. and Ploenzke, M. Deep learning for inferring transcription factor binding sites. *Current Opinion in Systems Biology*, Jun 2020. ISSN 2452-3100. doi: 10.1016/j.coisb.2020.04.001. URL http://www.sciencedirect.com/science/article/pii/S2452310020300032.

Lambert, S. A., Jolma, A., Campitelli, L. F., Das, P. K., Yin, Y., Albu, M., Chen, X., Taipale, J., Hughes, T. R., and Weirauch, M. T. The human transcription factors. *Cell*, 172(4):650–665, Feb 2018. ISSN 0092-8674. doi: 10.1016/j.cell.2018.01.029.

Lan, G., Zhou, J., Xu, R., Lu, Q., and Wang, H. Cross-cell-type prediction of tf-binding site by integrating convolutional neural network and adversarial network. *International Journal of Molecular Sciences*, 20(1414):3425, Jan 2019. doi: 10.3390/ijms20143425.

Langley, P. Crafting papers on machine learning. In Langley, P. (ed.), *Proceedings of the 17th International Conference on Machine Learning (ICML 2000)*, pp. 1207–1216, Stanford, CA, 2000. Morgan Kaufmann.

Lee, C. M., Barber, G. P., Casper, J., Clawson, H., Diekhans, M., Gonzalez, J. N., Hinrichs, A. S., Lee, B. T., Nassar, L. R., Powell, C. C., and et al. Ucsc genome browser enters 20th year. *Nucleic Acids Research*, 48 (D1):D756–D761, Jan 2020a. ISSN 0305-1048. doi: 10.1093/nar/gkz1012.

Lee, R. v. d., Correard, S., and Wasserman, W. W. Deregulated regulators: Disease-causing cis variants in transcription factor genes. *Trends in Genetics*, 36(7):523–539, Jul 2020b. ISSN 0168-9525. doi: 10.1016/j.tig.2020.04.006.

Maslova, A., Ramirez, R. N., Ma, K., Schmutz, H., Wang, C., Fox, C., Ng, B., Benoist, C., Mostafavi, S., and Project, t. I. G. Learning immune cell differentiation. *bioRxiv*, pp. 2019.12.21.885814, Dec 2019. doi: 10.1101/2019.12.21.885814.

Mathelier, A., Shi, W., and Wasserman, W. W. Identification of altered cis-regulatory elements in human disease. *Trends in Genetics*, 31(2):67–76, Feb 2015. ISSN 0168-9525. doi: 10.1016/j.tig.2014.12.003.

Mignone, P., Pio, G., DElia, D., and Ceci, M. Exploiting transfer learning for the reconstruction of the human gene regulatory network. *Bioinformatics*, 36(5):1553–1561, Mar 2020. ISSN 1367-4803. doi: 10.1093/bioinformatics/btz781.

Quang, D. and Xie, X. Danq: a hybrid convolutional and recurrent deep neural network for quantifying the function of dna sequences. *Nucleic Acids Research*, 44(11):e107–e107, Jun 2016. ISSN 0305-1048. doi: 10.1093/nar/gkw226.

Quang, D. and Xie, X. Factornet: A deep learning framework for predicting cell type specific transcription factor binding from nucleotide-resolution sequential data. *Methods*, 166:40–47, Aug 2019. ISSN 1046-2023. doi: 10.1016/j.ymeth.2019.03.020.

Quinlan, A. R. and Hall, I. M. Bedtools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26(6):841–842, Mar 2010. ISSN 1367-4803. doi: 10.1093/bioinformatics/btq033.

Schwessinger, R., Gosden, M., Downes, D., Brown, R., Telenius, J., Teh, Y. W., Lunter, G., and Hughes, J. R. Deepc: Predicting chromatin interactions using megabase scaled deep neural networks and transfer learning. *bioRxiv*, pp. 724005, Aug 2019. doi: 10.1101/724005.

Stewart, A. J., Hannenhalli, S., and Plotkin, J. B. Why transcription factor binding sites are ten nucleotides long. *Genetics*, 192(3):973–985, Nov 2012. ISSN 0016-6731. doi: 10.1534/genetics.112.143370.

Swamynathan, S. K. Krüppel-like factors: Three fingers in control. *Human Genomics*, 4(4):263, Apr 2010. ISSN 1479-7364. doi: 10.1186/1479-7364-4-4-263.

Szklarczyk, D., Gable, A. L., Lyon, D., Junge, A., Wyder, S., Huerta-Cepas, J., Simonovic, M., Doncheva, N. T., Morris, J. H., Bork, P., and et al. String v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Research*, 47(D1):D607–D613, Jan 2019. ISSN 0305-1048. doi: 10.1093/nar/gky1131.

Thrun, S. and Pratt, L. *Learning to Learn: Introduction and Overview*, pp. 3–17. Springer US, 1998. ISBN 978-1-4615-5529-2. doi: 10.1007/978-1-4615-5529-2_1. URL https://doi.org/10.1007/978-1-4615-5529-2_1.

Thurman, R. E., Rynes, E., Humbert, R., Vierstra, J., Maurano, M. T., Haugen, E., Sheffield, N. C., Stergachis, A. B., Wang, H., Vernot, B., and et al. The accessible chromatin landscape of the human genome. *Nature*, 489 (74147414):75–82, Sep 2012. ISSN 1476-4687. doi: 10.1038/nature11232.

Wang, J., Agarwal, D., Huang, M., Hu, G., Zhou, Z., Ye, C., and Zhang, N. R. Data denoising with transfer learning in single-cell transcriptomics. *Nature Methods*, 16(99): 875–878, Sep 2019a. ISSN 1548-7105. doi: 10.1038/s41592-019-0537-1.

Wang, M., Tai, C., E, W., and Wei, L. Define: deep convolutional neural networks accurately quantify intensities of transcription factor-dna binding and facilitate evaluation of functional non-coding variants. *Nucleic Acids Research*, 46(11):e69–e69, Jun 2018. ISSN 0305-1048. doi: 10.1093/nar/gky215.

Wang, T., Johnson, T. S., Shao, W., Lu, Z., Helm, B. R., Zhang, J., and Huang, K. Bermuda: a novel deep transfer learning method for single-cell rna sequencing batch correction reveals hidden high-resolution cellular subtypes. *Genome Biology*, 20(1):165, Aug 2019b. ISSN 1474-760X. doi: 10.1186/s13059-019-1764-6.

Wilson, V. and Conlon, F. L. The t-box family. *Genome Biology*, 3(6):reviews3008.1, May 2002. ISSN 1474-760X. doi: 10.1186/gb-2002-3-6-reviews3008.

Yang, Y., Fang, Q., and Shen, H.-B. Predicting gene regulatory interactions based on spatial gene expression data and deep learning. *PLOS Computational Biology*, 15(9):e1007324, Sep 2019. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1007324.

Zheng, A., Lamkin, M., Wu, C., Su, H., and Gymrek, M. Deep neural networks identify context-specific determinants of transcription factor binding affinity. *bioRxiv*, pp. 2020.02.26.965343, Feb 2020. doi: 10.1101/2020.02.26.965343.

Zhou, J. and Troyanskaya, O. G. Predicting effects of noncoding variants with deep learning–based sequence model. *Nature Methods*, 12(1010):931–934, Oct 2015. ISSN 1548-7105. doi: 10.1038/nmeth.3547.