Discovering differential genome sequence activity with interpretable and efficient deep learning

Abstract

Discovering sequence features that differentially direct cells to alternate fates is key to understanding both cellular development and the consequences of disease related mutations. Here we present a new method that efficiently learns sequence features that can predict cell state-specific chromatin accessibility in a framework that is readily accessible to non-experts.

Main

Neural networks have been increasingly helpful in understanding and predicting cellular function from genomic sequence in tasks such as transcription factor binding[1,2], chromatin accessibility[3,4], and histone modification[5]. Previously, we showed that DeepAccess, an ensemble of neural networks trained to predict chromatin accessibility, is able to identify *de novo* DNA sequence motifs that influence local cell type-specific accessibility in the context of a massively parallel reporter assay[6]. Here, we introduce an approach called the Expected Pattern Effect, which can be used to predict the effect of DNA sequence patterns such as known transcription factor motifs, combinations of motifs, or transcription factor spacing on the chromatin state of specific cell states. We also introduce a method to compare Expected Pattern Effects between two conditions or cell states, which we call the Differential Expected Pattern Effect.

Recovering interpretable biological hypothesizes from trained deep neural network models has been a challenge given the complexity of the models. Contemporary approaches for interpreting deep learning models of genomic sequence features utilize four main approaches: 1) interpretation of model first layer as position weight matrices[3,4,7,8], 2) model saliency and motif generation from input sequences[9,10], 3) saturating in silico mutagenesis[2,5,11], and 4) model response to in silico "designed" sequences[3,12–15].

While other deep learning interpretation methods provide means to identify DNA sequence features associated with a single condition, here we present the first method of identifying DNA sequence features that significantly differentiate between cell types or conditions. Such significant sequence features can suggest hypotheses that relate specific DNA binding factors to cell state. Thus these factors can be candidates for follow up studies on their effectiveness in directed differentiation protocols designed to produce specific cells types.

In our work we build upon the use of "designed sequences" to create a computational method to measure the learned differential effect of a transcription factor motif or other DNA sequence pattern, such as a combination of transcription factors, on the output of a DeepAccess model (Figure 1A). Many deep learning models are multi-task models that simultaneously predict more

than one function. For example, they may simultaneously predict the binding of multiple transcription factors or the chromatin accessibility of multiple cell types. The output of these models is often a probability. Thus, for a binary measurement, y, over N training examples, we can estimate the overall expected probability of y to be:

$$\mathbb{E}[y] = \frac{1}{N}\sum_{n=1}^{N} 1[y_n == 1]$$

Since we intend to compare the effect of genomic sequence features with imbalanced class data, we have to account for the relative effect of average class prediction probability. Thus, we take into account the average class prediction probability in our formalization of the Expected Pattern Effect, α, of a genomic pattern, p, such as a transcription factor motif. Given a model that predicts function from sequence $\mathcal{F} : x \rightarrow y$, where x is a sequence of nucleotides and y represents a corresponding binary functional measurement, the effect of pattern, p, can be estimated by the impact of the insertion of p into the sequence x to make the new sequence x^p, both of which are within the data distribution D, and is given by the ratio of expectations:

$$\alpha^p = \frac{\mathbb{E}_{x^p \sim D}[y|x^p]}{\mathbb{E}_{x \sim D}[y|x]}$$

Here we use a multi-task DeepAccess model $f_c : x \rightarrow y_c$, were y_c denotes the model predictions on functional task, c. DeepAccess can be trained and thus predict on one or more classes c. We select M sequences to approximate D and estimate α_c^p according to:

$$\alpha_c^p = \frac{1}{M}\sum_{m=1}^{M} \frac{f_c({x_m}^p)}{f_c(x_m)}$$

We define the significance of the Expected Pattern Effect (EPE) as the signed-rank test statistic of the ratios:

$$\mathrm{W} = \frac{1}{M}\sum_{m=1}^{M} sign\left(log_2\left(\frac{f_c({x_m}^p)}{f_c(x_m)}\right)\right) R_m$$

where R_m represents the rank of the absolute log ratio of x_m^p to x_m.

We then compute the Differential Expected Pattern Effect (DEPE) between two classes c_1 and c_2 based upon the average $\log_2$ ratio of predictions:

$$\mathrm{DEPE}(\alpha_{c_1}^p, \alpha_{c_2}^p) = \frac{1}{M}\sum_{m=1}^{M} \log_2\left(\frac{f_{c_1}({x_m}^p)}{f_{c_1}(x_m)}\right) - \log_2\left(\frac{f_{c_2}({x_m}^p)}{f_{c_2}(x_m)}\right)$$

In addition to computing the magnitude of the DEPE between two classes, we compute the significance of this effect:

$$\mathrm{W} = \frac{1}{M}\sum_{m=1}^{M} sign\left(\log_2\left(\frac{f_{c_1}(x_m{}^p)}{f_{c_1}(x_m)}\right) - \log_2\left(\frac{f_{c_2}(x_m{}^p)}{f_{c_2}(x_m)}\right)\right) R_m$$

where R_m represents the rank of the absolute log ratio of pattern effect of the pattern, *p,* between two classes c_1 and c_2, on a sequence *m.*

Our goal is to apply the DEPE to discover cell type-specific transcription factor activity from chromatin accessibility, but to first validate that DEPE can correctly identify differential transcription factor activity, we apply the DeepAccess architecture to predict DNA binding from ChIP-seq data. In contrast to chromatin accessibility where multiple or specific combinations of transcription factors may be responsible for activity, ChIP-seq should produce sequences enriched by the binding of a single DNA binding protein. Therefore, if a DeepAccess model has successfully learned differential DNA sequence patterns between two ChIP-seq experiments, DEPE should identify the DNA binding motifs of those transcription factors.

We evaluated the effectiveness of the DEPE by evaluating how well it could identify sequence motifs that were responsible for differential ChIP-seq binding. We trained a single DeepAccess model to predict the DNA sequences bound by six ChIP-seq experiments for the transcription factors ELF1, MEF2A, CTCF, USF1, TCF12, and NRF1 in mouse erythroleukemia (MEL) cells from ENCODE[16]. The trained DeepAccess model takes as input a DNA sequence and outputs the predicted binding of six transcription factors to the input sequence (six outputs). The DeepAccess model was trained with the 10,000 most significantly bound genomic regions from GPS[17,18] binding event calls. Each region is classified as being bound by either neither, one, or multiple transcription factors based on whether it overlaps a GPS binding event. After successfully training the DeepAccess model, we ran our method for DEPE to extract differential patterns which result in significant differential binding predictions, using the HOCOMOCOv11[19] database of transcription motifs as our patterns. The transcription factor motifs matching the transcription factor ChIP are found to have significant DEPE (Figure 1B-D), except in the case of differential motif discovery between NRF1 and TCF12, where the TCF12 motif was not discovered as a differential pattern between NRF1 and TCF12 binding (Figure 1D). Surprisingly, another motif (EVI1) had a strong DEPE on TCF12 binding, suggesting that TCF12 may be binding indirectly through partnering with EVI1 or another dispersed zinc finger protein in MEL cells. Despite the large overlap in binding sites between NRF1 and TCF12, DEPE is able to identify the NRF1 motif as a differential pattern between NRF1 and TCF12 binding, indicating that our approach is able to identify differential motifs even when two classes share a majority of genomic regions (Figure 1D).

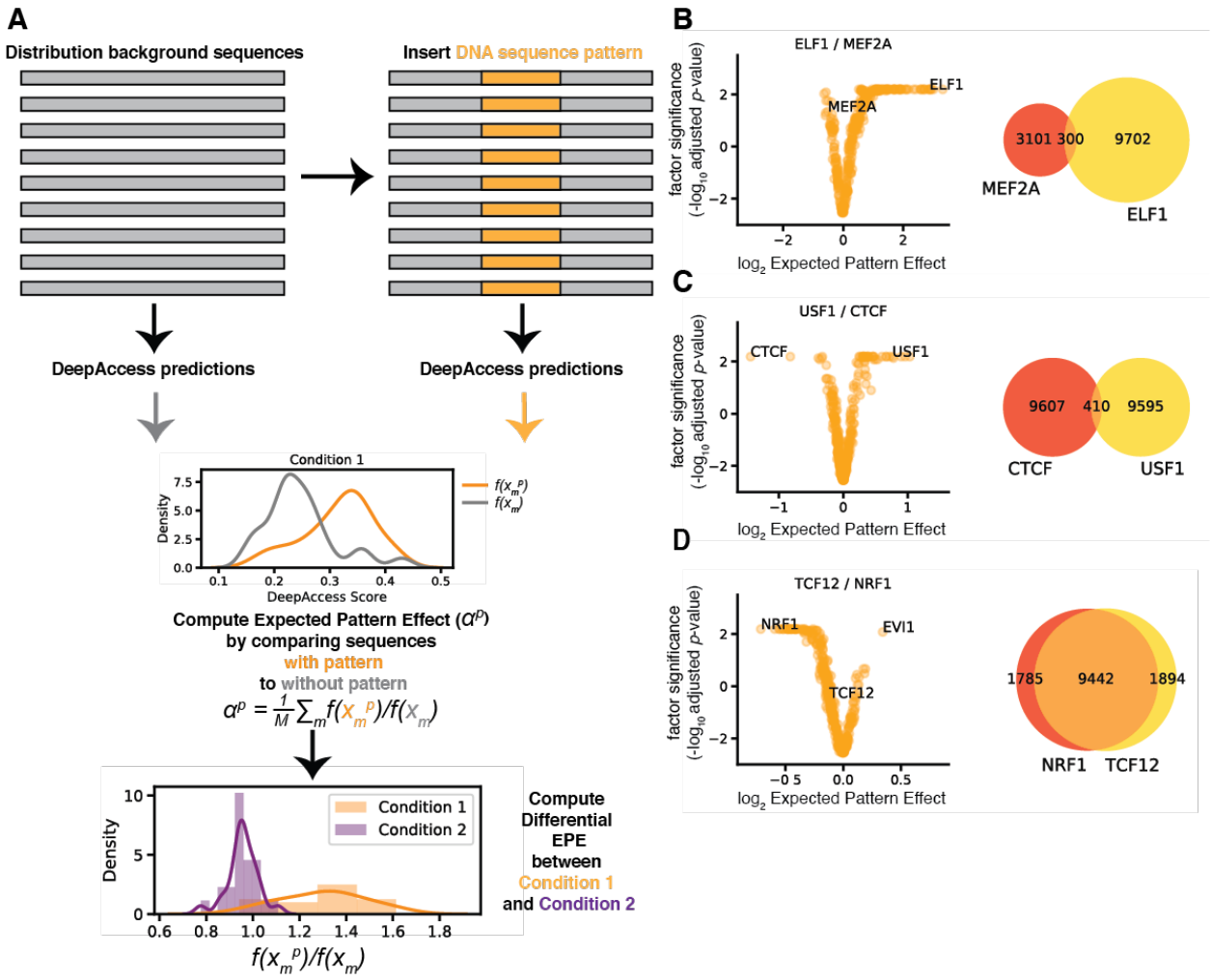

Figure 1. Differential Expected Pattern Effect identifies condition-specific transcription factor activity from DeepAccess models. A) Computation of Expected Pattern Effect and Differential Expected Pattern Effect investigates effects of transcription factor motifs on condition-specific accessibility. DeepAccess trained on mouse ChIP-seq data identifies correct motifs driving differential transcription factor activity for B) ELF1 and MEF2A, C) CTCF and USF1, D) NRF1 and TCF12. Venn diagrams show number and overlap of ChIP-seq binding events used as training and testing data for DeepAccess.

Another problem with deep learning methods is that they contain a large number of free parameters that need to be learned from data. With small data sets, large models may be overfit to training data, which results in models learning spurious and non-generalizable set of features and making less accurate predictions on test examples. Transfer learning is one solution to reducing the required volume of training data, where a model trained for a first task is used to help prime the training of a model for a second new task. In our implementation of transfer learning, we directly use the parameters learned for a first model to initialize a related second model to reduce the training data necessary for good performance of the second model.

We have increased the utility of our method by providing an executable version of DeepAccess for the Linux operating system that includes a pretrained model trained on ATAC-seq from ten

mouse cell types (Figure 2A). This approach reduces the time required to train new tasks and allows models to be trained with smaller training datasets by using transfer learning to adapt our pre-trained DeepAccess model to new problems. Using Keras (v2.3.1) with a Tensorflow (v1.15.0) backend for neural network deployment, the DeepAccess package automatically parallelizes to take advantage of multiple cores on a CPU. The DeepAccess system takes as input multiple bed files (Figure 2B), each labeled by experimental condition (i.e., cell type, condition, genomic signal), or fasta files with labels for each fasta input sequence (Figure 2C). The DeepAccess system then splits the data into training and test data, performs training with early stopping criteria based on a held-out validation set of the training data (Figure 2D). Performance metrics and predictions are reported for training and test data (Figure 2E). Finally, we include our methods for interpretation with testing EPE and Differential EPE between conditions (Figure 2F) and differential importance of individual nucleotides (Figure 2G). Additionally, the DeepAccess system is also able to make predictions on new input DNA sequences (Figure 2H).

Our packaging of the DeepAccess system avoids typical installation problems with the evolution of Python packages by packaging all dependencies including the Python interpreter within our software. We also avoid using package managers such as pip and conda which are present with frameworks such as Selene[20], Kipoi[21], and pyyster[22], as problems with incompatibility can still arise with typical package managers as a consequence of python, package manager, and operating system updates.

We note that Expected Pattern Effect resembles other methods for extracting pattern effects from deep neural networks[3,13], except that we compute Expected Pattern Effect such that we can compare pattern effects between conditions, which is important for identification of cell type-specific or condition-specific sequence features.

While DeepAccess was developed specifically for identifying cell type-specific sequence features from chromatin accessibility, Differential Expected Pattern Effect can be used to discover condition-specific sequence features from many types of experimental genome-wide sequencing data. We expect that DeepAccess will provide a useful approach for discovering sequence features that drive differential cell fates, and that the system we provide will make this capability accessible to a wide audience.

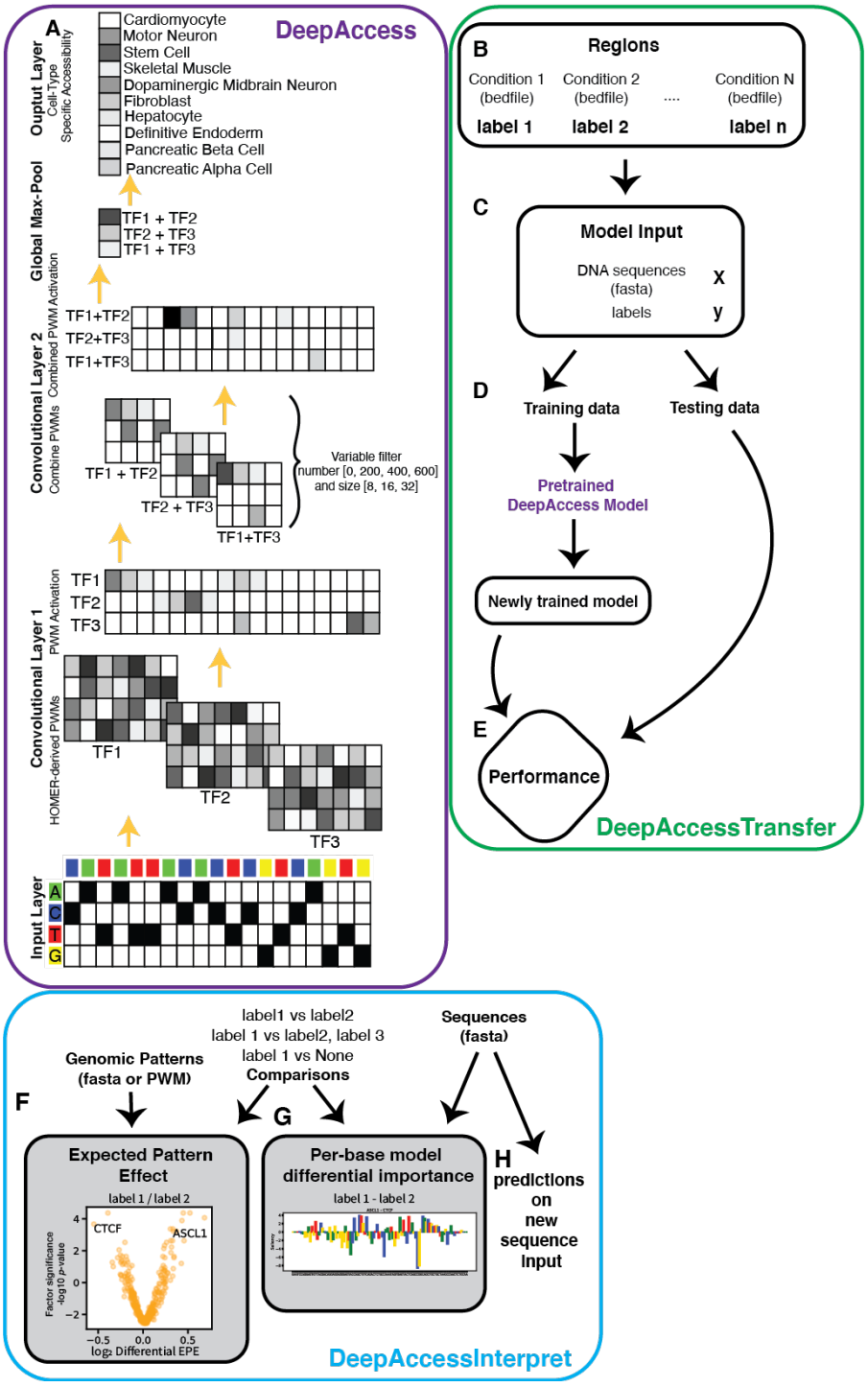

Figure 2. Accessible and interpretable deep learning with DeepAccess. A) Core DeepAccess model that acts as a starting model is trained on ATAC-seq data from 10 mouse cell types. The DeepAccess system takes as input either B) genomic regions from multiple conditions with labels, or C) DNA sequences as fasta files and labels. D) Data is split into training and testing, and pretrained model is retrained to new task. E) Model performance is computed. Finally, results of newly trained model can be interpreted using F) ExpectedPatternEffect and DifferentialExpectedPatternEffect to evaluate activity of genomic patterns for a given condition

or between conditions, G) per-base model differential importance scores between conditions and H) predictions on new input sequences.

Methods

ChIP-seq read processing and binding event calling

Reads were trimmed for adaptors and low-quality positions using Trimgalore (Cutadapt v0.6.2) and aligned to the mouse genome (mm10) with bwa mem (v0.7.1.7) with default parameters. Duplicates were removed with samtools (v1.7.2) markdup, and ChIP binding events were called with GPS (v3.4) with default parameters.

MEF2A	ENCODE	ENCFF433FQJ; ENCFF423RCQ
ELF1	ENCODE	ENCFF186NAS; ENCFF592ERV
USF1	ENCODE	ENCFF996MWJ; ENCFF550MZG
TCF12	ENCODE	ENCFF880XUQ; ENCFF432YTR
NRF1	ENCODE	ENCFF845NFO; ENCFF205UPE
CTCF	ENCODE	ENCFF001NRQ; ENCFF001NRO
IGG	ENCODE	ENCFF001NCH; ENCFF001NCF

Table S1. ChIP-seq data used to test DeepAccess Differential Expected Pattern Effects.

DeepAccess ChIP-seq training

We used as input to DeepAccess 200bp regions centered at the GPS-called ChIP-seq binding event. DeepAccess then uses these regions to create a training data set of 87,114 100bp sequences which can be labeled as binding events for any combination of the 6 tested transcription factors, including regions where none of the 6 transcription factors are predicted to bind. We define a region as a ChIP binding event for a transcription factor if more that 50% of the 100bp region overlaps the 200bp region centered at a GPS binding event. The DeepAccess model was then trained for 3 epochs with default parameters.

Pretrained DeepAccess Model

Methods for DeepAccess and differential nucleotide importance with DeepAccess were previously described in Hammelman et al., 2020. Briefly, we trained DeepAccess on ATAC-seq data from ten cell types: stem cell, fibroblast, hepatocyte, endoderm, beta pancreatic cell, alpha pancreatic cell, cardiomyocyte, skeletal muscle, dopaminergic midbrain neuron, and spinal motor neuron using binary cross-entropy loss (multi-task classification) with 4,220,507 genomic regions for training: 3,220,507 regions were open in at least 1 cell type, and 1,000,000 regions were closed in all cell types (randomly sampled from the genome). Chromosome 18 and chromosome 19 were held out for validation and testing, respectively. To define training regions

213 for DeepAccess, we generate 100bp genomic windows across the entire mouse genome. We
214 define a region as accessible in a given cell type if more that 50% of the 100bp region overlaps a
215 MACS2[23] accessible region from that cell type. The pretrained DeepAccess model is available as
216 a part of the DeepAccessTransfer package at http://cgs.csail.mit.edu/DeepAccessTransfer/
217
218 For ATAC-seq processing, reads were trimmed for adaptors and low-quality positions using
219 Trimgalore (Cutadapt v0.6.2). Reads were aligned to the mouse genome (mm10) with bwa mem
220 (v0.7.1.7) with default parameters. Duplicates were removed with samtools (v1.7.2) markdup
221 and properly paired mapped reads were filtered. Accessible regions were called using MACS2
222 (v2.2.7.1) with the parameters -f BAMPE -g mm -p 0.01 --shift -36 --extsize 73 --nomodel --
223 keep-dup all --call-summits. Accessible regions that overlapped genome blacklist regions were
224 excluded from downstream analysis.
225

Stem cell ATAC-seq	[24]	GSE64825
Fibroblast ATAC-seq	[25]	GSE66534
Skeletal Muscle ATAC-seq	[26]	GSE123879
Cardiomyocyte ATAC-seq	[27]	GSE95763
Endoderm ATAC-seq	[28]	GSM3223249
Hepatocyte ATAC-seq	ENCODE	ENCSR609OHJ
Pancreatic alpha/beta cell ATAC-seq	[29]	GSE99954
Dopaminergic Midbrain Neuron ATAC-seq	[30]	GSE122450
Spinal Motor Neuron ATAC-seq	Unpublished data	

226 **Table S2.** ATAC-seq data used for pre-trained DeepAccess model.
227
228 **Selection of Synthetic Distribution of Sequences**
229 For all experiments we used as our synthetic distribution sequences 24 randomly selected
230 genomic locations that were not overlapping ATAC-seq accessible regions in all ten mouse cell
231 types. We provide these regions with the DeepAccess executable as default backgrounds for new
232 experiments.
233

234 DeepAccess Executable

235 DeepAccessAccess executable was generated with pyinstaller (v4.1).

Data availability

No new data were produced by this study. Previously published ChIP-seq and ATAC-seq data are reported in Table S1 and Table S2 of methods.

Code availability

Binaries for DeepAccess training and interpretation with ExpectedPatternEffect and differential saliency is available http://cgs.csail.mit.edu/DeepAccessTransfer/. Source code is available at https://github.com/gifford-lab/DeepAccessTransfer

Acknowledgements

We thank members of the Gifford lab and Wichterle lab for helpful discussions and contributing data. We gratefully acknowledge funding from 1RO1HG008363 (D.K.G.), 1R01HG008754 (D.K.G.), 1R01NS109217 (D.K.G.), and National Science Foundation Graduate Research Fellowship (1122374) (J.H.).

.

Contributions

Conceptualization, J.H. and D.K.G.; Methodology, J.H.; Software, J.H.; Formal Analysis, J.H.; Investigation, J.H.; Resources, D.K.G.; Data Curation, J.H.; Writing - Original Draft, J.H.; Writing - Review & Editing, J.H., and D.K.G.; Visualization, J.H.; Supervision, D.K.G.; Funding Acquisition, J.H. and D.K.G.

Corresponding Author

Correspondence to David Gifford

Ethics Declarations

The authors declare no competing interests. The software described here is available under a MIT license and is free for non-commercial and commercial use.

References

1. Zeng, H., Edwards, M. D., Liu, G. & Gifford, D. K. Convolutional neural network architectures for predicting DNA-protein binding. *Bioinformatics* **32**, i121–i127 (2016).
2. 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 (2015).
3. Kelley, D. R., Snoek, J. & Rinn, J. L. Basset: Learning the regulatory code of the accessible genome with deep convolutional neural networks. *Genome Res.* **26**, 990–999 (2016).
4. Quang, D. & Xie, X. DanQ: a hybrid convolutional and recurrent deep neural network for quantifying the function of DNA sequences. *Nucleic Acids Res.* **44**, e107–e107 (2016).

5. Zhou, J. & Troyanskaya, O. G. Predicting effects of noncoding variants with deep learning-based sequence model. *Nat. Methods* **12**, 931–934 (2015).
6. Hammelman, J., Krismer, K., Banerjee, B., Gifford, D. K. & Sherwood, R. I. Identification of determinants of differential chromatin accessibility through a massively parallel genome-integrated reporter assay. *Genome Res.* **30**, (2020).
7. Zeng, H. & Gifford, D. K. Predicting the impact of non-coding variants on DNA methylation. *Nucleic Acids Res.* **45**, e99–e99 (2017).
8. Koo, P. K. & Eddy, S. R. Representation learning of genomic sequence motifs with convolutional neural networks. *PLoS Comput. Biol.* **15**, e1007560 (2019).
9. Shrikumar, A., Greenside, P. & Kundaje, A. Learning Important Features Through Propagating Activation Differences. *arxiv* (2017).
10. Shrikumar, A. *et al.* Technical Note on Transcription Factor Motif Discovery from Importance Scores (TF-MoDISco) version 0.5. 1.1. *arXiv Prepr. arXiv1811.00416* (2018).
11. Phuycharoen, M. *et al.* Uncovering tissue-specific binding features from differential deep learning. *Nucleic Acids Res.* **48**, e27–e27 (2020).
12. Koo, P. K., Anand, P., Paul, S. B. & Eddy, S. R. Inferring Sequence-Structure Preferences of RNA-Binding Proteins with Convolutional Residual Networks. *bioRxiv* 418459 (2018).
13. Koo, P. K., Ploenzke, M., Anand, P., Paul, S. & Majdandzic, A. Global Importance Analysis: A Method to Quantify Importance of Genomic Features in Deep Neural Networks. *bioRxiv* (2020).
14. Kim, D. *et al.* The dynamic, combinatorial cis-regulatory lexicon of epidermal differentiation. *bioRxiv* 2020.10.16.342857 (2020) doi:10.1101/2020.10.16.342857.
15. Avsec, Ž. *et al.* Base-resolution models of transcription-factor binding reveal soft motif syntax. *Nat. Genet.* (2021) doi:10.1038/s41588-021-00782-6.
16. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, (2012).
17. Guo, Y. *et al.* Discovering homotypic binding events at high spatial resolution. *Bioinformatics* **26**, 3028–3034 (2010).
18. Guo, Y., Mahony, S. & Gifford, D. K. High resolution genome wide binding event finding and motif discovery reveals transcription factor spatial binding constraints. *PLoS Comput Biol* **8**, (2012).
19. Kulakovskiy, I. V *et al.* HOCOMOCO: towards a complete collection of transcription factor binding models for human and mouse via large-scale ChIP-Seq analysis. *Nucleic Acids Res.* **46**, D252–D259 (2018).
20. Chen, K. M., Cofer, E. M., Zhou, J. & Troyanskaya, O. G. Selene: a

PyTorch-based deep learning library for sequence data. *Nat. Methods* **16**, 315–318 (2019).

21. Avsec, Ž. *et al.* The Kipoi repository accelerates community exchange and reuse of predictive models for genomics. *Nat. Biotechnol.* **37**, 592–600 (2019).
22. Budach, S. & Marsico, A. pysster: classification of biological sequences by learning sequence and structure motifs with convolutional neural networks. *Bioinformatics* **34**, 3035–3037 (2018).
23. Zhang, Y. *et al.* Model-based analysis of ChIP-Seq (MACS). *Genome Biol.* **9**, R137 (2008).
24. De Dieuleveult, M. *et al.* Genome-wide nucleosome specificity and function of chromatin remodellers in ES cells. *Nature* **530**, 113–116 (2016).
25. Cheloufi, S. *et al.* The histone chaperone CAF-1 safeguards somatic cell identity. *Nature* **528**, 218–224 (2015).
26. Ramachandran, K. *et al.* Dynamic enhancers control skeletal muscle identity and reprogramming. *PLoS Biol.* **17**, e3000467 (2019).
27. Quaife-Ryan, G. A. *et al.* Multicellular transcriptional analysis of mammalian heart regeneration. *Circulation* **136**, 1123–1139 (2017).
28. Cernilogar, F. M. *et al.* Pre-marked chromatin and transcription factor co-binding shape the pioneering activity of Foxa2. *Nucleic Acids Res.* **47**, 9069–9086 (2019).
29. Lawlor, N., Youn, A., Kursawe, R., Ucar, D. & Stitzel, M. L. Alpha TC1 and Beta-TC-6 genomic profiling uncovers both shared and distinct transcriptional regulatory features with their primary islet counterparts. *Sci. Rep.* **7**, 1–14 (2017).
30. McClymont, S. A. *et al.* Parkinson-associated SNCA enhancer variants revealed by open chromatin in mouse dopamine neurons. *Am. J. Hum. Genet.* **103**, 874–892 (2018).