Discovering differential genome sequence activity with interpretable and efficient deep learning

Jennifer Hammelman[1,2] and David K. Gifford[2,3]

1. Computational and Systems Biology, MIT, Cambridge, MA, USA.
2. Computer Science and Artificial Intelligence Laboratory, MIT, Cambridge, MA, USA
3. Department of Biological Engineering, MIT, Cambridge, MA, USA

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. We introduce Expected Pattern Effect and Differential Expected Pattern Effect, two black-box methods that can interpret genome regulatory sequences for cell type-specific or condition specific patterns. We show that these methods identify relevant transcription factor motifs and spacings that are predictive of cell state-specific chromatin accessibility. Finally, we integrate these methods into framework that is readily accessible to non-experts and available for download as a binary or installed via PyPI or bioconda at https://cgs.csail.mit.edu/deepaccess-package/.

Author Summary

Within the genome are the instructions to build all the cell types that make up the human body. However, understanding these instructions and how and when these instructions go wrong in cancer or genetically inherited disease is an open problem. Deep neural networks provide powerful models to learn the relationship between DNA sequence and functional consequence across many different cell types, such as whether a particular stretch of DNA is accessible and genes in that region can be expressed or is inaccessible and therefore genes are inactive. Despite these advances, a major setback in deep learning is that it is challenging to understand what patterns of DNA sequences a deep learning model has learned to associate with a particular genomic function, whether these patterns are significant, and how to determine whether these patterns are specific to a particular cell type or are general "housekeeping" patterns that function across many cell types. We introduce Expected Pattern Effect and Differential Expected Pattern Effect, two methods which allow us to evaluate the significance of particular patterns of DNA sequence features on models trained to predict function across multiple cell types, and apply this to problems of transcription factor binding and DNA accessibility across multiple cell types.

Introduction

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 (EPE) between two conditions or cell states, which we call the Differential Expected Pattern Effect (DEPE). While the scientific community has yet to come to a consensus on a precise definition of cell state or cell type, for the purposes of this work we define cell type as a shared characteristic gene expression profile. For example, motor neuron cells are defined by characteristic expression of the motor neuron and pancreas homeobox gene Hb9 (7). A cell state is similarly defined through gene expression response, but refers to a shared change in gene expression as a response to external stimuli. We focus on cell type in this work, but the methods we introduce are also readily applicable to analysis of cell state.

Results

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 three main approaches (Fig 1A):

1. Interpretation of model convolutional filters as position weight matrices (3,4,8,9).
2. Interpretation of model and individual sequences for important nucleotides using model saliency and motif generation from input sequences (10–13) or saturating in silico mutagenesis (2,5,12).
3. Model response to in silico “designed” sequences (3,14–17).

67 Methods in class 1 can generate novel potentially important motifs from the trained weights of
68 convolutional filters, but filters have been known to represent partial motifs (9) and additional
69 analysis is required to interpret the significance of these motifs on cell type-specific activity.
70 Methods in class 2 such as DeepSHAP (10,18) and Enhanced Integrated Gradients (13) provide
71 means to identify DNA sequence features associated with a single condition and differentially
72 between two conditions (Fig 1A), but these methods function to identify DNA sequences
73 features within an individual input sequence, so the importance of these sequence features is
74 dependent on the surrounding nucleotides within in the input DNA sequence and do not
75 represent the overall expected effect of a particular pattern such as a transcription factor motif on
76 cell type-specific accessibility. In the case of models operating on DNA sequence, we have
77 knowledge *a priori* that particular combinations of features (patterns) that could have important
78 effects, such as transcription factor motifs. Interpretation methods make use of our *a priori*
79 knowledge about the patterns of features such as DNA motifs and relationships between patterns
80 such as spacing to investigate what features the model has learned are overall most important
81 through evaluating model response to "designed" DNA sequences *in silico*. However, these
82 approaches have yet to explore how to assign the significance of these effects and how to
83 interpret whether these effects are differential between two conditions, such as cell types, when
84 applied to multi-task classification models.
85
86 Here we present a method of assigning both effect scores and significance values to DNA
87 sequence feature patterns of interest, such as transcription factor motifs or grammars such as
88 particular transcription factor motifs and motif spacings, that differentiate between cell types or
89 conditions from a multi-task neural network (Fig 1B). Significant differential 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 contrast to methods that analyze a single input sequence or a model, 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 on the output of a DeepAccess model (Fig 1A). These patterns can include a combination of transcription factor motifs. 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 (Fig 1B). 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:

$$\alpha^p = \mathbb{E}_{x \sim D}\left[\frac{\mathcal{F}(x^p)}{\mathcal{F}(x)}\right]$$

While EPE is a general method that can be applied to any multi-task classification model, here we apply it to investigate cell type-specific chromatin accessibility using a multi-task DeepAccess model $f_c : x \rightarrow y_c$, to approximate $\mathcal{F}$, 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:

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

where R_m represents the rank of the absolute log ratio of sequences with the pattern insertion, x_m^p, to sequences without the pattern insertion, x_m.

132 We then compute the Differential Expected Pattern Effect between two classes c_1 and c_2 based

133 upon the average $\log_2$ ratio of predictions:

134

135 $$\mathrm{DEPE}(\alpha_{c_1}^{p}, \alpha_{c_2}^{p}) = \frac{1}{M}\sum_{m=1}^{M} \log_2\left(\frac{f_{c_1}({x_m}^{p})}{f_{c_1}(x_m)}\right) - \log_2\left(\frac{f_{c_2}({x_m}^{p})}{f_{c_2}(x_m)}\right)$$

136

137 In addition to computing the magnitude of the DEPE between two classes, we compute the

138 significance of this effect:

139

140 $$\mathrm{W} = \frac{1}{M}\sum_{m=1}^{M} \mathrm{sign}\left(\log_2\left(\frac{f_{c_1}({x_m}^{p})}{f_{c_1}(x_m)}\right) - \log_2\left(\frac{f_{c_2}({x_m}^{p})}{f_{c_2}(x_m)}\right)\right) R_m$$

141

142 where R_m represents the rank of the absolute log ratio of the expected pattern effect of the pattern,

143 *p,* differentially between two classes c_1 and c_2, on a sequence *m.*

144

145 Our goal is to apply the DEPE to discover cell type-specific transcription factor activity from

146 chromatin accessibility, but to first validate that DEPE can correctly identify differential

147 transcription factor activity, we apply the DeepAccess architecture to predict DNA binding from

148 ChIP-seq data. In contrast to chromatin accessibility where multiple or specific combinations of

149 transcription factors may be responsible for activity, ChIP-seq should produce sequences

150 enriched by the binding of a single DNA binding protein. Therefore, if a DeepAccess model has

151 successfully learned differential DNA sequence patterns between two ChIP-seq experiments,

152 DEPE should identify the DNA binding motifs of those transcription factors.

153

154 We evaluated the effectiveness of the DEPE by evaluating how well it could identify sequence
155 motifs that were responsible for differential ChIP-seq binding. We trained a single DeepAccess
156 model to predict the DNA sequences bound by six ChIP-seq experiments for the transcription
157 factors ELF1, MEF2A, CTCF, USF1, TCF12, and NRF1 in mouse erythroleukemia (MEL) cells
158 from ENCODE (19). The trained DeepAccess model takes as input a DNA sequence and outputs
159 the predicted binding of six transcription factors to the input sequence (six outputs). The
160 DeepAccess model was trained with the 10,000 most significantly bound genomic regions from
161 GPS (20,21) binding event calls. Each region is classified as being bound by either neither, one,
162 or multiple transcription factors based on whether it overlaps a GPS binding event. After
163 successfully training the DeepAccess model, we ran our method for DEPE to extract differential
164 patterns which result in significant differential binding predictions, using the HOCOMOCOv11
165 (22) database of transcription motifs as our patterns. The transcription factor motifs matching the
166 transcription factor ChIP are found to have significant DEPE (Fig 2), except in the case of
167 differential motif discovery between NRF1 and TCF12, where the TCF12 motif was not
168 discovered as a differential pattern between NRF1 and TCF12 binding (Fig 2C). The DMRT1
169 and PO5F1 motifs had the strongest DEPE on TCF12 binding relative to NRF1, suggesting that
170 TCF12 may be binding indirectly through partnering with other proteins, including a
171 transcription factor recognizing ACAAT DNA motif which is shared by DMRT1 and PO5F1 in
172 MEL cells. Despite the large overlap in binding sites between NRF1 and TCF12, DEPE is able to
173 identify the NRF1 motif as a differential pattern between NRF1 and TCF12 binding, indicating
174 that our approach is able to identify differential motifs even when two classes share a majority of
175 genomic regions (Fig 2C; Fig S1). The number of motifs that are identified as significantly
176 differential is asymmetric for comparisons between ELF1 and MEF2A (Fig 2A) and USF1 and

177 CTCF (Fig 2B). This can be explained by the large size of the bHLH-ZIP (USF1) and Ets
178 (ELF1) transcription factor families, each encompassing greater than 20 transcription factors
179 with highly similar binding motifs within the HOCOMOCOv11 transcription factor database
180 (22). However, the same is not true for NRF1 which also has asymmetric differential
181 transcription factor motifs when DEPE is computed between NRF1 binding and TCF12 binding.
182 It appears that NRF1 binding sites may frequently contain binding motifs for other transcription
183 factors such as Klf/Sp factors, and Gata which may indicate cooperative binding relationship.
184 This in combination with the large number overlap of Nrf1 and Tcf12 binding sites suggests
185 Nrf1 and Tcf12 binding may be regulated by logic beyond the presence or absence of their DNA
186 binding motifs. We also show that under class imbalanced data, EPE and DEPE generate robust
187 estimates of transcription factor effects in contrast to Global Importance Analysis (15), which
188 takes a related approach but uses a difference instead of a ratio to estimate the effect of a pattern
189 (Fig S2).
190
191 We then used EPE and Differential EPE to investigate chromatin accessibility. First, we
192 hypothesized that we can use EPE to investigate preferential spacing of Oct4 and Sox2 motifs
193 within accessible DNA in stem cells, as these two transcription factors are known to
194 cooperatively bind. We tested the EPEs for multiple spacings of Oct4 and Sox2 motifs (Fig 3A)
195 and find the DeepAccess model predicts the highest Expected Pattern Effect from the spacing of
196 Pou5f1 and Sox2 that is consistent with the spacing between the DNA motifs of Pou5f1 and
197 Sox2 co-binding events from ChIP-seq binding data (Fig 3B). We then examined Differential
198 EPE for all transcription factor motifs in the HOCOMOCOv11 database between stem cells and
199 motor neurons (Fig 3C) and find homeobox transcription factors such as Hoxc9, Meis1, and

200 Hoxa9 are differentially important for motor neuron accessibility consistent with the important
201 role that Hox genes play in motor neuron development (23,24), whereas transcription factor
202 motifs for Maz and Sp5 are differentially important in stem cells, with Sp5 and Maz sharing a
203 similar DNA binding motif and Sp5 playing an established role downstream of the WNT
204 pathway to maintain self-renewal in stem cells (25,26). We found that EPE and DEPEs for
205 transcription factors is generally not affected by the position of insertion within the designed
206 sequences, except when the site of insertion is at the very beginning of the DNA sequence (Fig
207 S3). Thus, we used the center of the DNA sequence as our insertion site for all subsequent
208 experiments. Additionally, we examined how selection of background sequences to approximate
209 D affects EPE and DEPE, and found that EPEs are smaller when distribution-approximating
210 sequences are selected from accessible or inaccessible parts of the genome (Fig S4) but are more
211 or less invariant to specific annotations (i.e. promoters) compared to randomly selected DNA.
212 Since we are most interested in motifs driving accessibility, we use randomly sampled closed
213 DNA as our distribution sequences. We also examined Differential EPE between liver and
214 fibroblast, identifying transcription factor motifs Hnf1a, FoxD3, and Hnf1B as being
215 differentially important to liver (Fig 3D), consistent with the role of hepatocyte nuclear factors
216 (Hnfs), including Fox, in liver differentiation (27–29). Finally, we compute the Differential EPE
217 for each of the ten cell types relative to the other nine cell types, for all 356 mouse core
218 HOCOMOCOv11 transcription factor motifs (Fig 3E), and find that Differential EPE of
219 transcription factors identifies transcription factors that are known lineage regulators, such as
220 LIM homeobox transcription factors in dopaminergic neurons [DN] (30,31) and motor neurons
221 [MN] (32–34), as well as Hox transcription factors which are specific to spinal motor neurons.
222 JUN/FOS/AP-1 family members c-Jun, c-Fos, and Fosl1are likely to play a role in fibroblast

223 [FIBRO] identity as knockdown of these transcription factors facilitated reprogramming of
224 fibroblast to stem cells (35) and over-expression led to reduced reprogramming efficiency to
225 stem cell state and the maintenance of accessibility of fibroblast-specific enhancers (36).
226 Additionally, we find Fox transcription factors in both pancreatic [aPN, bPN] and liver cells
227 [LIV] which is consistent with the role of Fox in pan endoderm differentiation and liver and
228 pancreatic cell differentiation (28,37–39). Finally, hierarchical clustering using the Differential
229 EPE for significant transcription factors showed relationships between cell types such as alpha
230 pancreatic and beta pancreatic cells, and motor neuron and dopaminergic neurons which share
231 more similar Differential EPEs (Fig 3E).
232
233 We have increased the utility of our method by providing an executable version of DeepAccess
234 for the Linux operating system as well as a python package available via PyPI and bioconda (Fig
235 4A). Using Keras (v2.4.3) with a Tensorflow (v2.4) backend for neural network deployment, the
236 DeepAccess package automatically parallelizes to take advantage of multiple cores on a CPU or
237 GPU. The DeepAccess system takes as input multiple bed files (Fig 4B), each labeled by
238 experimental condition (i.e., cell type, condition, genomic signal), or fasta files with labels for
239 each fasta input sequence (Fig 4C). The DeepAccess system then splits the data into training and
240 test data, performs training with early stopping criteria based on a held-out validation set of the
241 training data (Fig 4D). Performance metrics and predictions are reported for training and test
242 data (Fig 4E). Finally, we include our methods for interpretation with testing EPE and
243 Differential EPE between conditions (Fig 4F) and differential importance of individual
244 nucleotides (Fig 4G). Additionally, the DeepAccess system is also able to make predictions on
245 new input DNA sequences (Fig 4H).

246 The DeepAccess system, including methods for estimating EPE and Differential EPE, is
247 available as a PyPI package and on bioconda. We also provide a binary of the DeepAccess
248 system which avoids typical installation problems with the evolution of Python packages by
249 packaging all dependencies including the Python interpreter within our software. Even when
250 using package managers such as pip and conda which are present with frameworks such as
251 Selene (40), Kipoi (41), and pyyster (42), problems with incompatibility can still arise with
252 typical package managers as a consequence of python, package manager, and operating system
253 updates.
254

255 **Discussion**

256 While previous work has explored the contributions of neural network architecture, training
257 methodology, and model interpretation to learn the relationship between DNA sequence and
258 differential DNA binding (12), these neural networks were trained to only predict cell type-
259 specific binding. In contrast, DEPE can be used to interpret DNA sequence features influencing
260 differential accessibility in cases where models have been trained to predict accessibility across
261 tens, hundreds, or thousands of unique cell types. In these cases, explicitly training models on
262 each differential comparison between cell types would be computational expensive and time-
263 intensive. We also note that Expected Pattern Effect resembles other methods for extracting
264 pattern effects from deep neural networks (3,15), except that we compute Expected Pattern
265 Effect to permit the comparison of pattern effects between conditions, which is important for
266 identification of cell type-specific or condition-specific sequence features and show that our use
267 of ratio to compare effects is more robust to analysis of cell type-specific transcription factor
268 activity under class imbalance.

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.

A

Input	Method	Requires Access to Weights?	Output
Model single input DNA sequence	Saliency	Yes	Per-nucleotide importance
	Enhanced Integrated Gradients		Per-nucleotide importance & significance
	DeepShap		Per-nucleotide importance
	Saturating in silico mutagenesis	No	Per-nucleotide importance
Model	Motifs from convolutional filters	Yes	PWM with unknown relationship to class in multitask/multiclass models
Model motif or grammar	Global Importance Analysis	No	Scores of motif or grammar
	Expected Pattern Effect		Scores & significance of motif or grammar
	Differential Expected Pattern Effect		Scores & significance of motif or grammar between classes

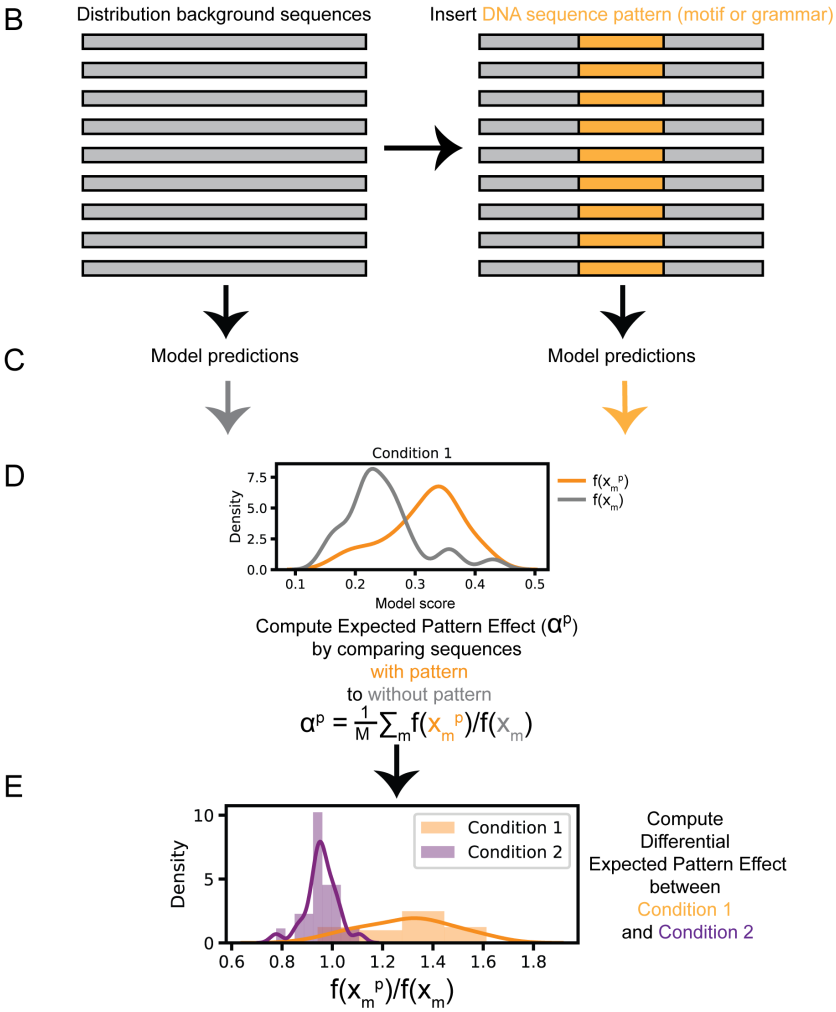

Fig 1. Expected Pattern Effect and Differential Expected Pattern Effect determine the significance of feature patterns such as transcription factor motifs or transcription factor grammar for multitask models.

(A) Comparison of interpretation methods shows 3 categories of interpretation methods for deep learning models based on input. Some methods operate on an individual sequence input, model pair, others turn learned convolutional filters into DNA sequence motif position weight matrices and then must assign class relevance post-hoc, or methods Global Importance Analysis and Expected Pattern Effect score the overall class importance of DNA sequence patterns such as transcription factor motifs or grammar such as combinations or spacing of transcription factor motifs. (B) Computation of Expected Pattern Effect and Differential Expected Pattern Effect starts by selection of a sequences that represent a "background" from the natural distribution of sequences. A pattern of interest either a transcription factor or combination or spacing of transcription factors is artificially inserted into each background sequence. (C) Model predicts functional signal for "background" sequences and sequences with inserted transcription factor motif. In our case we use a DeepAccess model, but principles for computing EPE and DEPE can be applied to any multitask classification model. (D) Average ratio between scores of background sequences with pattern of interest (in orange) and without pattern of interest – i.e. background sequences (in grey) determines the Expected Pattern Effect. (E) Differential Expected Pattern Effect can be computed by comparing Expected Pattern Effect between different classes – i.e. conditions or cell types.

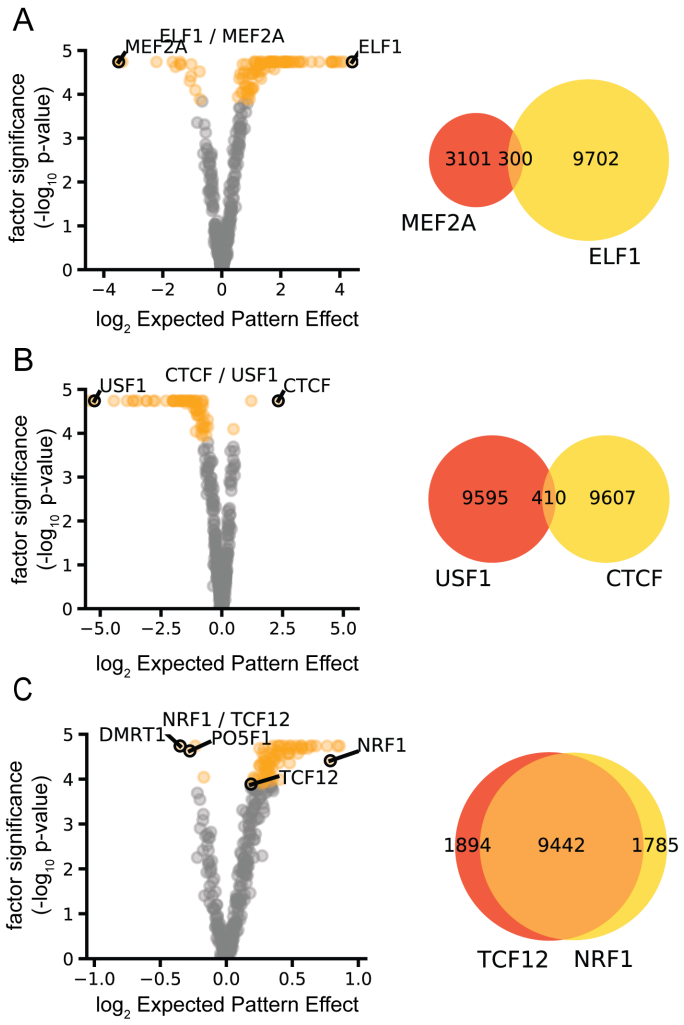

Fig 2. Differential Expected Pattern Effect identifies condition-specific transcription factor activity from DeepAccess models.
DeepAccess trained on mouse ChIP-seq data identifies correct motifs driving differential transcription factor activity for (A) ELF1 and MEF2A, (B) CTCF and USF1, (C) NRF1 and TCF12. In the scatter plot each point represents one of 356 HOCOMOCOv11 transcription factor motifs. Scatter plot y-axis is unadjusted rank ratio significance, x-axis is the Differential Expected Pattern Effect. Orange points are significant transcription factor motifs ($p < 0.05$) under Bonferroni multiple hypothesis correction. Venn diagrams show number and overlap of ChIP-seq binding events used as training and testing data for DeepAccess.

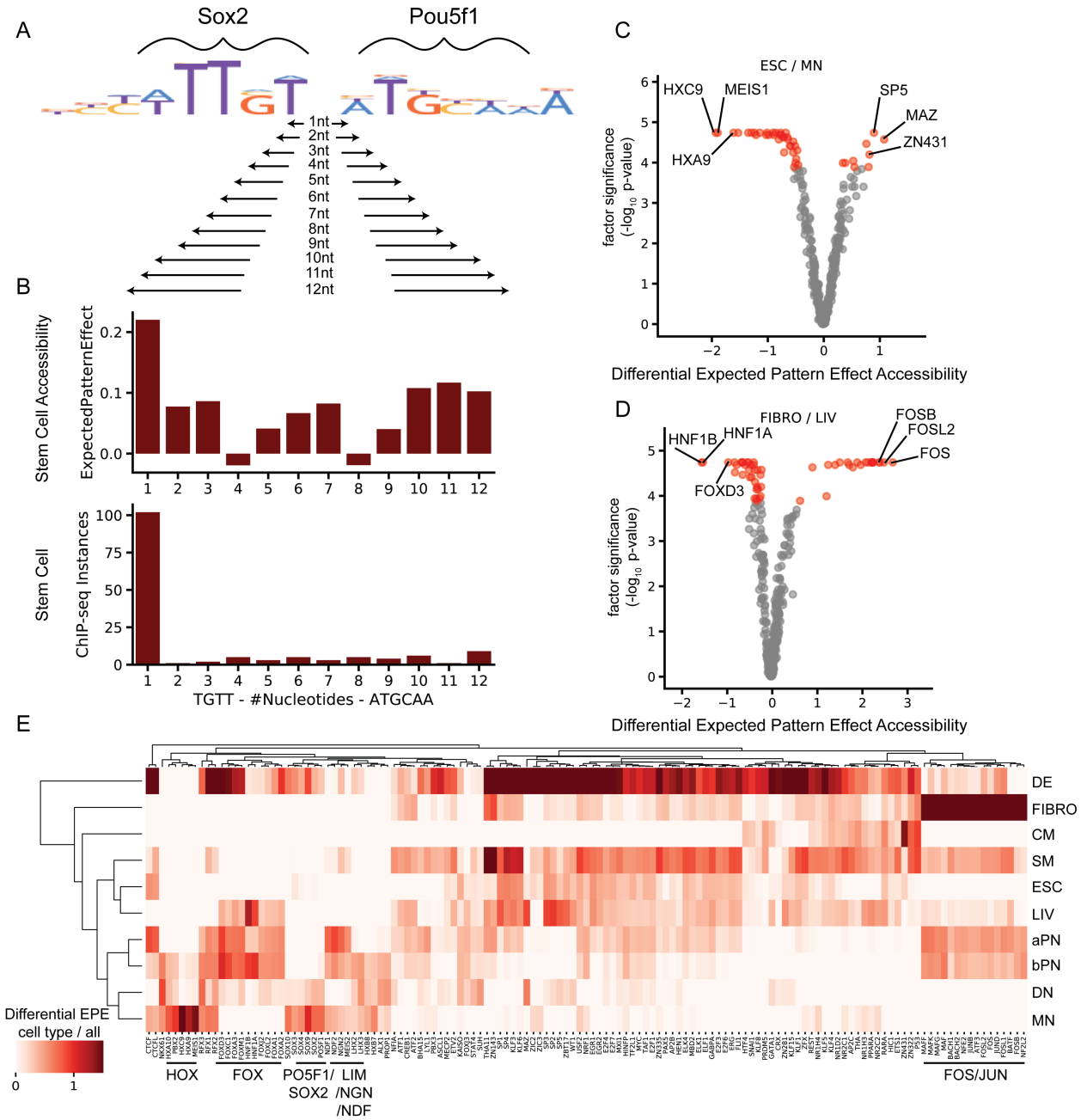

Fig 3. Expected Pattern Effect and Differential Expected Pattern Effect identify preferential spacing and cell type-specific transcription factor activity from DeepAccess models trained on chromatin accessibility data from ten cell types.
(A) Testing all possible spacing between Sox2 transcription factor and Pou5f1 (Oct4) transcription factor motifs input as patterns to compute Expected Pattern Effects from a DeepAccess model. (B) DeepAccess Expected Pattern Effect of Sox2 and Pou5f1spacing on predicted stem cell accessibility (top) is consistent with preferential spacing of Sox2 and Pou5f1 motifs from ChIP-seq binding data of Sox2 and Pou5f1 co-binding events (C) Differential EPE of transcription factor motifs between stem cells (ESC) and motor neurons (MN) identifies known MN-specific transcription factors such as Hox TFs and Meis. (D) Differential EPE of transcription factor motifs between fibroblasts (FIBRO) and liver cells (LIV) identifies liver-determining factors such as Hnf1 and Fox transcription factors. Scatter plot y-axis is

321 unadjusted rank ratio significance, x-axis is the Differential Expected Pattern Effect ($\log_2$ fold change in
322 Expected Pattern Effect between cell types), with each point representing one of 356 HOCOMOCOv11
323 transcription factor motifs. Red points are significant ($p < 0.05$) transcription factor motifs under
324 Bonferroni multiple hypothesis correction. (E) Transcription factors showing significant cell type-specific
325 EPEs for each of the ten cell types relative to all other cell types. Heatmap intensity indicates the $\log_2$ fold
326 change in Expected Pattern Effect between each cell type and the average of all other cell types.
327

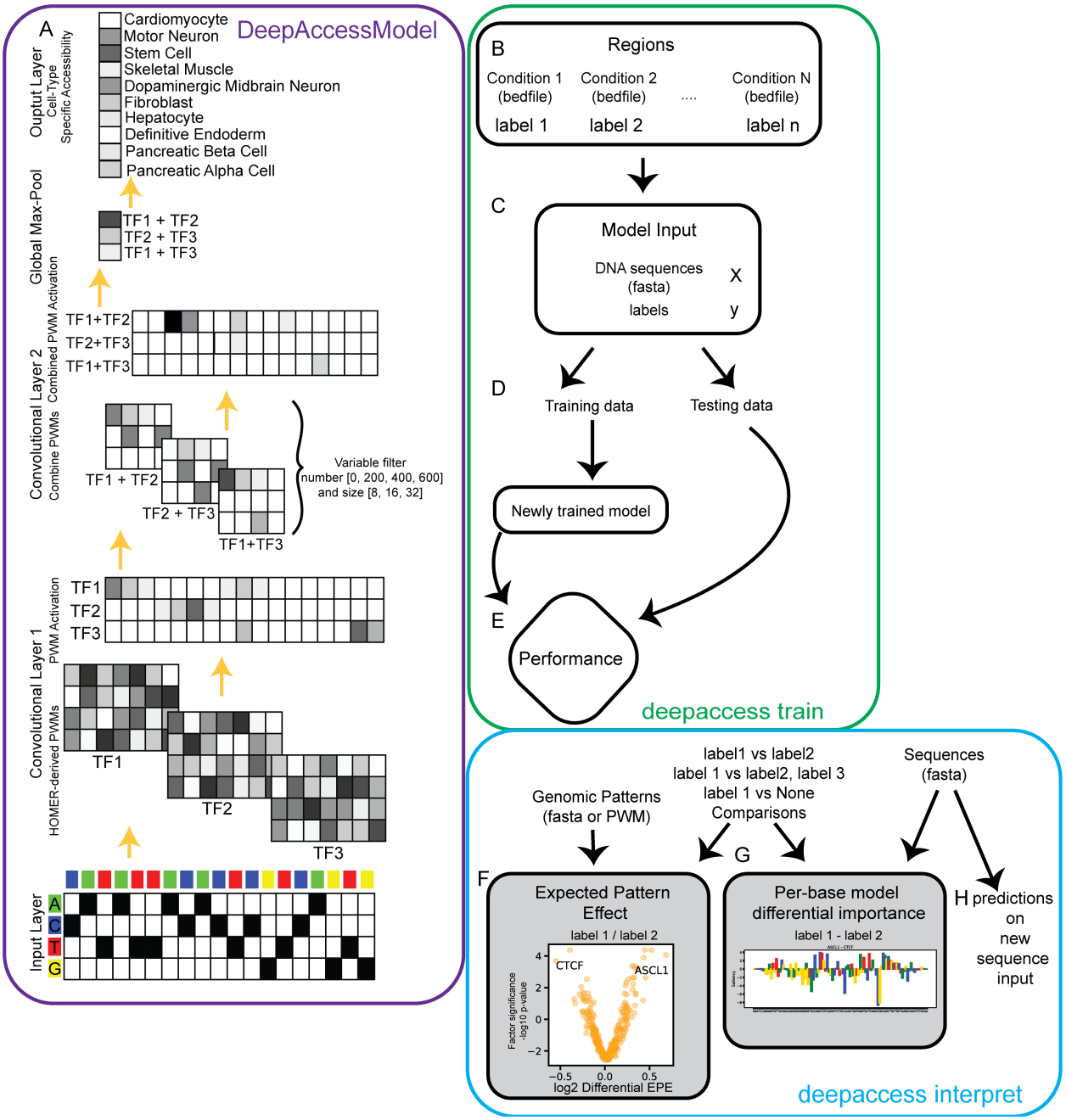

Fig 4. Accessible and interpretable deep learning with DeepAccess.
(A) Core DeepAccess model example trained to predict accessibility from 100nt DNA sequence using ATAC-seq data from 10 mouse cell types. Input layer is one-hot encoding of DNA sequence. Convolutional layers scan for transcription factor motifs using position weight matrices derived from the HOMER database. Ensemble of models with variable filter number and size in the second convolutional layer combine transcription factor motifs. Global max pooling layer condenses results into maximally activating position for convolutional filter. Output is the multi-task accessibility which can be interpreted as the probability that an input DNA sequence will be accessible in each cell type, state, or condition. 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 model is trained with early stopping criteria based on a validation subset of ATAC-seq data. (E) Model performance is computed. Finally, results of newly trained model can be interpreted using (F) ExpectedPatternEffect and

341 DifferentialExpectedPatternEffect to evaluate activity of genomic patterns for a given condition or
342 between conditions, (G) per-base model differential importance scores between conditions and (H)
343 predictions on new input sequences.
344

345 # Methods

346

347 **ChIP-seq / ChIP-nexus read processing and binding event calling**

348 Reads were trimmed for adaptors and low-quality positions using Trimgalore (Cutadapt v0.6.2)
349 (43) and aligned to the mouse genome (mm10) with bwa mem (v0.7.1.7) (44) with default
350 parameters. Duplicates were removed with samtools (v1.7.2) (45) markdup, and ChIP binding
351 events were called with GPS (v3.4) (20,21) with default parameters.

352 **DeepAccess ChIP-seq training**

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

360

361 **DeepAccess Model trained on ATAC-seq for 10 mouse cell types**

362 Methods for DeepAccess and differential nucleotide importance with DeepAccess were
363 previously described in Hammelman et al., 2020. Briefly, we trained DeepAccess on ATAC-seq
364 data from ten cell types: stem cell, fibroblast, hepatocyte, endoderm, beta pancreatic cell, alpha
365 pancreatic cell, cardiomyocyte, skeletal muscle, dopaminergic midbrain neuron, and spinal motor

366 neuron using binary cross-entropy loss (multi-task classification) with 4,220,507 genomic
367 regions for training: 3,220,507 regions were open in at least 1 cell type, and 1,000,000 regions
368 were closed in all cell types (randomly sampled from the genome). Chromosome 18 and
369 chromosome 19 were held out for validation and testing, respectively. To define training regions
370 for DeepAccess, we generate 100bp genomic windows across the entire mouse genome. We
371 define a region as accessible in a given cell type if more that 50% of the 100bp region overlaps a
372 MACS2 (46) accessible region from that cell type. The pretrained DeepAccess model is
373 available as a Zenodo record at https://zenodo.org/record/4908895#.YL6YpR0pDfY.
374
375 For ATAC-seq processing, reads were trimmed for adaptors and low-quality positions using
376 Trimgalore (Cutadapt v0.6.2) (43). Reads were aligned to the mouse genome (mm10) with bwa
377 mem (v0.7.1.7) (44) with default parameters. Duplicates were removed with samtools (v1.7.2)
378 markdup (45) and properly paired mapped reads were filtered. Accessible regions were called
379 using MACS2 (v2.2.7.1) (46) with the parameters -f BAMPE -g mm -p 0.01 --shift -36 --extsize
380 73 --nomodel --keep-dup all --call-summits. Accessible regions that overlapped Encode blacklist
381 regions ENCSR535HHO were excluded from downstream analysis.
382
383 **Preferential spacing of Oct4 and Sox2**
384 ChIP-nexus binding data for Oct4, Sox2, and patch cap control (17)(17) were downloaded as
385 fastqs from the Nucleotide Read Archive (see Table S1 for accession numbers). ChIP-seq
386 processing was performed as described (see ChIP-seq / ChIP-nexus read processing and binding
387 event calling). After, all GPS events where Sox2 or Oct4 binding sites were less than 50nt were
388 considered as close co-binding events. using bedtools (v2.29.2) (53) intersect. Events were

389 scanned for the first instances of TTGT (Sox2 DNA sequence motif) and ATGCAA (Oct4 DNA
390 sequence motif), and the distance between the last nucleotide of the Sox2 DNA sequence motif
391 and the first nucleotide of the Oct4 DNA sequence motif was defined as our distance between TF
392 sites.
393
394 **Selection of Synthetic Distribution of Sequences**
395 For all experiments we used as our synthetic distribution sequences 24 randomly selected
396 genomic locations that were not overlapping ATAC-seq accessible regions in all ten mouse cell
397 types. We provide these regions with the DeepAccess executable as default backgrounds for new
398 experiments.
399

400 DeepAccess Executable

401 DeepAccess executable was generated with pyinstaller (v4.1).
402

403 Data availability

404 No new data were produced by this study. Previously published ChIP-seq and ATAC-seq data
405 are reported in Table S1 and Table S2 of methods.
406

407 Code availability

408 Binaries for DeepAccess training and interpretation with ExpectedPatternEffect and differential
409 saliency is available http://cgs.csail.mit.edu/deepaccess-package/. Source code is available at
410 https://github.com/gifford-lab/deepaccess-package
411

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. and D.K.G.; 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 MD, Liu G, Gifford DK. Convolutional neural network architectures for predicting DNA-protein binding. Bioinformatics [Internet].

434 2016 Jun 15;32(12):i121–7. Available from:
435 http://dx.doi.org/10.1093/bioinformatics/btw255
436 2. Alipanahi B, Delong A, Weirauch MT, Frey BJ. Predicting the sequence
437 specificities of DNA- and RNA-binding proteins by deep learning. Nat
438 Biotechnol [Internet]. 2015 Jul 27;33:831. Available from:
439 http://dx.doi.org/10.1038/nbt.3300
440 3. Kelley DR, Snoek J, Rinn JL. Basset: Learning the regulatory code of the
441 accessible genome with deep convolutional neural networks. Genome Res.
442 2016;26(7):990–9.
443 4. Quang D, Xie X. DanQ: a hybrid convolutional and recurrent deep neural
444 network for quantifying the function of DNA sequences. Nucleic Acids Res
445 [Internet]. 2016 Apr 15;44(11):e107–e107. Available from:
446 https://doi.org/10.1093/nar/gkw226
447 5. Zhou J, Troyanskaya OG. Predicting effects of noncoding variants with deep
448 learning-based sequence model. Nat Methods [Internet]. 2015
449 Oct;12(10):931–4. Available from: http://dx.doi.org/10.1038/nmeth.3547
450 6. Hammelman J, Krismer K, Banerjee B, Gifford DK, Sherwood RI.
451 Identification of determinants of differential chromatin accessibility through
452 a massively parallel genome-integrated reporter assay. Genome Res.
453 2020;30(10).

454 7. Arber S, Han B, Mendelsohn M, Smith M, Jessell TM, Sockanathan S.
455 Requirement for the Homeobox Gene Hb9 in the Consolidation of Motor
456 Neuron Identity. Neuron [Internet]. 1999;23(4):659–74. Available from:
457 https://www.sciencedirect.com/science/article/pii/S089662730180026X
458 8. Zeng H, Gifford DK. Predicting the impact of non-coding variants on DNA
459 methylation. Nucleic Acids Res. 2017;45(11):e99–e99.
460 9. Koo PK, Eddy SR. Representation learning of genomic sequence motifs with
461 convolutional neural networks. PLoS Comput Biol. 2019;15(12):e1007560.
462 10. Shrikumar A, Greenside P, Kundaje A. Learning Important Features Through
463 Propagating Activation Differences. arxiv [Internet]. 2017; Available from:
464 http://arxiv.org/abs/1704.02685
465 11. Shrikumar A, Tian K, Avsec Ž, Shcherbina A, Banerjee A, Sharmin M, et al.
466 Technical Note on Transcription Factor Motif Discovery from Importance
467 Scores (TF-MoDISco) version 0.5. 1.1. arXiv Prepr arXiv181100416. 2018;
468 12. Phuycharoen M, Zarrineh P, Bridoux L, Amin S, Losa M, Chen K, et al.
469 Uncovering tissue-specific binding features from differential deep learning.
470 Nucleic Acids Res [Internet]. 2020 Mar 18;48(5):e27–e27. Available from:
471 https://doi.org/10.1093/nar/gkaa009
472 13. Jha A, K. Aicher J, R. Gazzara M, Singh D, Barash Y. Enhanced Integrated
473 Gradients: improving interpretability of deep learning models using splicing

474 codes as a case study. Genome Biol [Internet]. 2020;21(1):149. Available
475 from: c
476 14. Koo PK, Anand P, Paul SB, Eddy SR. Inferring Sequence-Structure
477 Preferences of RNA-Binding Proteins with Convolutional Residual
478 Networks. bioRxiv. 2018;418459.
479 15. Koo PK, Majdandzic A, Ploenzke M, Anand P, Paul SB. Global importance
480 analysis: An interpretability method to quantify importance of genomic
481 features in deep neural networks. PLOS Comput Biol [Internet]. 2021 May
482 13;17(5):e1008925. Available from:
483 https://doi.org/10.1371/journal.pcbi.1008925
484 16. Kim D, Risca V, Reynolds D, Chappell J, Rubin A, Jung N, et al. The
485 dynamic, combinatorial cis-regulatory lexicon of epidermal differentiation.
486 bioRxiv [Internet]. 2020 Jan 1;2020.10.16.342857. Available from:
487 http://biorxiv.org/content/early/2020/10/18/2020.10.16.342857.abstract
488 17. Avsec Ž, Weilert M, Shrikumar A, Krueger S, Alexandari A, Dalal K, et al.
489 Base-resolution models of transcription-factor binding reveal soft motif
490 syntax. Nat Genet [Internet]. 2021; Available from:
491 https://doi.org/10.1038/s41588-021-00782-6
492 18. Lundberg S, Lee S-I. A unified approach to interpreting model predictions.
493 arXiv Prepr arXiv170507874. 2017;

494 19. An integrated encyclopedia of DNA elements in the human genome. Nature
495 [Internet]. 2012;489. Available from: http://dx.doi.org/10.1038/nature11247
496 20. Guo Y, Papachristoudis G, Altshuler RC, Gerber GK, Jaakkola TS, Gifford
497 DK, et al. Discovering homotypic binding events at high spatial resolution.
498 Bioinformatics. 2010;26(24):3028–34.
499 21. Guo Y, Mahony S, Gifford DK. High resolution genome wide binding event
500 finding and motif discovery reveals transcription factor spatial binding
501 constraints. PLoS Comput Biol [Internet]. 2012;8. Available from:
502 http://dx.doi.org/10.1371/journal.pcbi.1002638
503 22. Kulakovskiy I V, Vorontsov IE, Yevshin IS, Sharipov RN, Fedorova AD,
504 Rumynskiy EI, et al. HOCOMOCO: towards a complete collection of
505 transcription factor binding models for human and mouse via large-scale
506 ChIP-Seq analysis. Nucleic Acids Res [Internet]. 2018 Jan 4;46(D1):D252–9.
507 Available from: https://doi.org/10.1093/nar/gkx1106
508 23. Catela C, Shin MM, Lee DH, Liu J-P, Dasen JS. Hox Proteins Coordinate
509 Motor Neuron Differentiation and Connectivity Programs through Ret/Gfrα
510 Genes. Cell Rep [Internet]. 2016;14(8):1901–15. Available from:
511 https://www.sciencedirect.com/science/article/pii/S2211124716300602
512 24. Dasen JS, Tice BC, Brenner-Morton S, Jessell TM. A Hox Regulatory
513 Network Establishes Motor Neuron Pool Identity and Target-Muscle

514 Connectivity. Cell [Internet]. 2005;123(3):477–91. Available from:
515 https://www.sciencedirect.com/science/article/pii/S0092867405009219
516 25. Huggins IJ, Bos T, Gaylord O, Jessen C, Lonquich B, Puranen A, et al. The
517 WNT target SP5 negatively regulates WNT transcriptional programs in
518 human pluripotent stem cells. Nat Commun [Internet]. 2017;8(1):1034.
519 Available from: https://doi.org/10.1038/s41467-017-01203-1
520 26. Ye S, Zhang D, Cheng F, Wilson D, Mackay J, He K, et al. Wnt/β-catenin
521 and LIF–Stat3 signaling pathways converge on Sp5 to promote mouse
522 embryonic stem cell self-renewal. J Cell Sci [Internet]. 2016 Jan
523 15;129(2):269–76. Available from: https://doi.org/10.1242/jcs.177675
524 27. Simeonov KP, Uppal H. Direct Reprogramming of Human Fibroblasts to
525 Hepatocyte-Like Cells by Synthetic Modified mRNAs. PLoS One [Internet].
526 2014 Jun 25;9(6):e100134. Available from:
527 https://doi.org/10.1371/journal.pone.0100134
528 28. Le lay J, Kaestner KH. The Fox Genes in the Liver: From Organogenesis to
529 Functional Integration. Physiol Rev [Internet]. 2010 Jan 1;90(1):1–22.
530 Available from: https://doi.org/10.1152/physrev.00018.2009
531 29. Genga RMJ, Kernfeld EM, Parsi KM, Parsons TJ, Ziller MJ, Maehr R.
532 Single-Cell RNA-Sequencing-Based CRISPRi Screening Resolves Molecular
533 Drivers of Early Human Endoderm Development. Cell Rep [Internet].

534 2019;27(3):708-718.e10. Available from:
535 http://www.sciencedirect.com/science/article/pii/S2211124719304061
536 30. Smidt MP, Asbreuk CHJ, Cox JJ, Chen H, Johnson RL, Burbach JPH. A
537 second independent pathway for development of mesencephalic
538 dopaminergic neurons requires Lmx1b. Nat Neurosci. 2000;3(4):337–41.
539 31. Nakatani T, Kumai M, Mizuhara E, Minaki Y, Ono Y. Lmx1a and Lmx1b
540 cooperate with Foxa2 to coordinate the specification of dopaminergic
541 neurons and control of floor plate cell differentiation in the developing
542 mesencephalon. Dev Biol [Internet]. 2010;339(1):101–13. Available from:
543 https://www.sciencedirect.com/science/article/pii/S0012160609014225
544 32. Mazzoni EO, Mahony S, Closser M, Morrison CA, Nedelec S, Williams DJ,
545 et al. Synergistic binding of transcription factors to cell-specific enhancers
546 programs motor neuron identity. Nat Neurosci [Internet]. 2013;16(9):1219–
547 27. Available from: https://doi.org/10.1038/nn.3467
548 33. Thaler JP, Lee S-K, Jurata LW, Gill GN, Pfaff SL. LIM factor Lhx3
549 contributes to the specification of motor neuron and interneuron identity
550 through cell-type-specific protein-protein interactions. Cell.
551 2002;110(2):237–49.
552 34. Velasco S, Ibrahim MM, Kakumanu A, Ohler U, Mahony S, Mazzoni EO. A
553 Multi-step Transcriptional and Chromatin State Cascade Underlies Motor

554 Neuron Programming from Embryonic Stem Cells Article A Multi-step
555 Transcriptional and Chromatin State Cascade Underlies Motor Neuron
556 Programming from Embryonic Stem Cells. Cell Stem Cell. 2017;205–17.
557 35. Knaupp AS, Buckberry S, Pflueger J, Lim SM, Ford E, Larcombe MR, et al.
558 Transient and Permanent Reconfiguration of Chromatin and Transcription
559 Factor Occupancy Drive Reprogramming. Cell Stem Cell [Internet].
560 2017;21(6):834-845.e6. Available from:
561 https://www.sciencedirect.com/science/article/pii/S1934590917304599
562 36. Li D, Liu J, Yang X, Zhou C, Guo J, Wu C, et al. Chromatin accessibility
563 dynamics during iPSC reprogramming. Cell Stem Cell. 2017;21(6):819–33.
564 37. Cernilogar FM, Hasenoder S, Wang Z, Scheibner K, Burtscher I, Sterr M, et
565 al. Pre-marked chromatin and transcription factor co-binding shape the
566 pioneering activity of Foxa2. Nucleic Acids Res. 2019 Sep;47(17):9069–86.
567 38. Spaeth JM, Hunter CS, Bonatakis L, Guo M, French CA, Slack I, et al. The
568 FOXP1, FOXP2 and FOXP4 transcription factors are required for islet alpha
569 cell proliferation and function in mice. Diabetologia [Internet]. 2015/05/29.
570 2015 Aug;58(8):1836–44. Available from:
571 https://pubmed.ncbi.nlm.nih.gov/26021489
572 39. Bastidas-Ponce A, Roscioni SS, Burtscher I, Bader E, Sterr M, Bakhti M, et
573 al. Foxa2 and Pdx1 cooperatively regulate postnatal maturation of pancreatic

β-cells. Mol Metab [Internet]. 2017 Mar 25;6(6):524–34. Available from: https://pubmed.ncbi.nlm.nih.gov/28580283

40. Chen KM, Cofer EM, Zhou J, Troyanskaya OG. Selene: a PyTorch-based deep learning library for sequence data. Nat Methods. 2019;16(4):315–8.

41. Avsec Ž, Kreuzhuber R, Israeli J, Xu N, Cheng J, Shrikumar A, et al. The Kipoi repository accelerates community exchange and reuse of predictive models for genomics. Nat Biotechnol. 2019;37(6):592–600.

42. Budach S, Marsico A. pysster: classification of biological sequences by learning sequence and structure motifs with convolutional neural networks. Bioinformatics. 2018;34(17):3035–7.

43. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet J. 2011;17(1):10–2.

44. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv Prepr arXiv13033997. 2013;

45. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The sequence alignment/map format and SAMtools. Bioinformatics. 2009;25(16):2078–9.

46. Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, et al. Model-based analysis of ChIP-Seq (MACS). Genome Biol. 2008;9(9):R137.

47. De Dieuleveult M, Yen K, Hmitou I, Depaux A, Boussouar F, Dargham DB,

594 et al. Genome-wide nucleosome specificity and function of chromatin
595 remodellers in ES cells. Nature. 2016;530(7588):113–6.
596 48. Cheloufi S, Elling U, Hopfgartner B, Jung YL, Murn J, Ninova M, et al. The
597 histone chaperone CAF-1 safeguards somatic cell identity. Nature.
598 2015;528(7581):218–24.
599 49. Ramachandran K, Senagolage MD, Sommars MA, Futtner CR, Omura Y,
600 Allred AL, et al. Dynamic enhancers control skeletal muscle identity and
601 reprogramming. PLoS Biol. 2019;17(10):e3000467.
602 50. Quaife-Ryan GA, Sim CB, Ziemann M, Kaspi A, Rafehi H, Ramialison M, et
603 al. Multicellular transcriptional analysis of mammalian heart regeneration.
604 Circulation. 2017;136(12):1123–39.
605 51. Lawlor N, Youn A, Kursawe R, Ucar D, Stitzel ML. Alpha TC1 and Beta-
606 TC-6 genomic profiling uncovers both shared and distinct transcriptional
607 regulatory features with their primary islet counterparts. Sci Rep.
608 2017;7(1):1–14.
609 52. McClymont SA, Hook PW, Soto AI, Reed X, Law WD, Kerans SJ, et al.
610 Parkinson-associated SNCA enhancer variants revealed by open chromatin in
611 mouse dopamine neurons. Am J Hum Genet. 2018;103(6):874–92.
612 53. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing
613 genomic features. Bioinformatics. 2010;26(6):841–2.

614

615 **Supplemental Legends**

616 Table S1. ChIP-seq data used to test DeepAccess Differential Expected Pattern Effects and for
617 preferential spacing of Oct4 / Sox2 in stem cell accessible DNA.
618
619
620 Table S2. ATAC-seq data used for pre-trained DeepAccess model.
621
622 Fig S1. All pairwise differential Expected Pattern Effects from ChIP-seq shows correct TF
623 motifs are identified in most comparative analyses.
624
625 Fig S2. Stability of estimated motif effects with Global Importance Analysis (subtraction) and
626 Expected Pattern Effect (ratio) shows EPE more accurate on multi-task data sets with class
627 imbalance. Scatter plots of GIA (A-C) or EPE/DEPE (D-F) of effect sizes of 356
628 HOCOMOCOv11 mouse motifs on predictions of CTCF binding from a trained DeepAccess
629 model on balanced class data (x-axis) compared to 2x, 10x, and 30x more CTCF binding sites :
630 USF1 binding sites (y-axis) (A) GIA of CTCF binding (B) GIA of USF1 binding (C) GIA of
631 CTCF vs USF1 differential binding (D) EPE of CTCF binding (E) EPE of USF1 binding (F)
632 DEPE of CTCF vs USF1 binding.
633
634 Fig S3. Effect of motif insertion position on EPE. (A) EPE of Hoxc9 motif on EPE from
635 DeepAccess model trained to predict motor neuron chromatin accessibility shows EPE of Hoxc9
636 motifs are low for positions in the beginning of the DNA sequence and high for positions at the
637 middle and end of the DNA sequence. (B) Scatter plot of EPEs for 356 HOCOMOCOv11 mouse
638 transcription factor motifs when insertion is at position 0 compared to position 50 indicate early
639 insertion site in DNA sequence results in different EPE values for transcription factors, with
640 some having lower EPEs and others having higher EPEs. (C) Scatter plot of EPEs for 356
641 HOCOMOCOv11 mouse transcription factor motifs when insertion is at position 70 compared to
642 position 50 indicate early insertion site in DNA sequence results in similar EPE values for all
643 transcription factors. (D) Heatmap of Pearson's r of EPEs for DeepAccess motor neuron
644 accessibility for insertions at all starting positions within the DNA sequence shows most
645 insertion positions result in highly similar EPEs except for early insertion positions. (E) Heatmap
646 of Pearson's r of DEPEs for DeepAccess motor neuron vs fibroblast differential accessibility for
647 insertions at all starting positions within the DNA sequence shows most insertion positions result
648 in highly similar EPEs except for early insertion positions. (F) Row z-score normalized EPEs for
649 DeepAccess motor neuron accessibility at each position show most motifs have higher EPEs
650 when inserted at a position in the center of the DNA sequence. Some motifs (like Sp motifs)
651 have higher EPEs at the beginning of the DNA sequence.
652
653 Fig S4. Choice of M sequences approximating sequence distribution affects EPE scores. (A)
654 Estimates of EPE for 356 HOCOMOCOv11 transcription factor motifs on predicted motor
655 neuron accessibility from a learned DeepAccess model. M sequences are selected randomly from
656 regions closed in all 10 cell types, open in all 10 cell types, promoters which are closed in all 10
657 cell types, and promoters which are open in all 10 cell types. Shade indicates distribution size

(M=10, M=25, M=100). Boxplot box indicates median and quartiles and whiskers extend to 1.5 times the inter-quartile range. Dots are outliers outside 1.5 times the inter-quartile range. (B) Estimates of significance of EPE for size and type of sequence used to estimate total sequence distribution. (C) Comparison of EPEs for 356 HOCOMOCOv11 transcription factor motifs for default (randomly selected 24 closed) compared to randomly selected closed (row 1), randomly selected open (row 2), promoter closed (row 3), and promoter open (row 4) shows while values of EPE dependent on selection of sequence distribution, rank remains relatively invariant.