DeepCAGE: Incorporating Transcription Factors in Genome-wide Prediction of Chromatin Accessibility

Qiao Liu [1,2], Kui Hua [1], Xuegong Zhang[1], Wing Hung Wong [2,*], Rui Jiang [1,*]

[1] Ministry of Education Key Laboratory of Bioinformatics, Research Department of Bioinformatics at the Beijing National Research Center for Information Science and Technology, Center for Synthetic and Systems Biology, Department of Automation, Tsinghua University, Beijing 100084, China

[2] Department of Statistics, Stanford University, Stanford, CA 94305, USA

[*] Corresponding authors.

E-mail: whwong@stanford.edu (Wong WH), ruijiang@tsinghua.edu.cn (Jiang R)

Running title: *Liu et al / DeepCAGE Predicts Chromatin Accessibility*

Count of words: 5747; **References**: 41; **Table**: 0; **Figures**: 6; **Supplementary Tables**: 7; **Supplementary Figures**: 5; **Count of letters for title**: 89; **Count of key words**: 4; **Abstract words**: 227

Abstract

Although computational approaches have been complementing high-throughput biological experiments for the identification of functional regions in the human genome, it remains a great challenge to systematically decipher interactions between transcription factors and regulatory elements to achieve interpretable annotations of chromatin accessibility across diverse cellular contexts. Towards this problem, we propose DeepCAGE, a deep learning framework that integrates sequence information and binding status of transcription factors, for the accurate prediction of chromatin accessible regions at a genome-wide scale in a variety of cell types. DeepCAGE takes advantage of a densely connected deep convolutional neural network architecture to automatically learn sequence signatures of known chromatin accessible regions, and then incorporates such features with expression levels and binding activities of human core transcription factors to predict novel chromatin accessible regions. In a series of systematic comparisons with existing methods, DeepCAGE exhibits superior performance in not only the classification but also the regression of chromatin accessibility signals. In detailed analysis of transcription factor activities, DeepCAGE successfully extracts novel binding motifs and measures the contribution of a transcription factor to the regulation with respect to a specific locus in a certain cell type. When applied to whole-genome sequencing data analysis, our method successfully prioritizes putative deleterious variants underlying a human complex trait, and thus provides insights into the understanding of disease-associated genetic variants. DeepCAGE can be downloaded from https://github.com/kimmo1019/DeepCAGE.

KEYWORDS: Chromatin accessibility; Deep learning; Transcription factor; Gene expression

Introduction

One of the fundamental questions in functional genomics is how activities of genes are spatially and temporally controlled through interactive effects of transcription factors (TFs) and regulatory elements (REs) such as promoters, enhancers and silencers. These regulatory elements, as short regions of non-coding DNA sequence, are known to typically reside in chromatin accessible regions and be bound by a set of TFs to carry out regulatory functionality in a manner specific to cellular contexts [1]. Therefore, the exploration of a landscape of chromatin accessible regions across major cell types will greatly facilitate the deciphering of gene regulatory mechanisms and further provide insights into cell differentiation, tissue homeostasis, and disease development [2].

Recent advances in deep sequencing techniques have enabled genome-wide assays of chromatin accessibility. For example, DNase-seq utilizes the DNase I enzyme to digest DNA sequences and identify DNase I-hypersensitive regions that are largely chromatin accessible [3]. ATAC-seq uses the Tn5 transposase to integrate primer DNA sequences into cleaved fragments that mainly come from chromatin accessible regions [4]. With the accomplishment of the ENCODE [5] and Roadmap [6] projects, these techniques have been successfully applied to the establishment of the chromatin accessibility landscape for dozens of cell lines across several species. The accumulation of these data provides an unprecedented opportunity for deepening our understanding of both gene regulation and occurrence of diseases [7,8,9].

However, due to limitations such as experimental cost, it is still impractical to further extend the landscape to cover all possible cell types, with the consideration of the huge variability in biological cellular contexts such as cell differentiation, environmental stimuli, and other factors. Towards this concern, computational approaches have been proposed to predict chromatin states by using such information as DNA sequence, gene expression, and other type of data [10-19]. For example, Kelley et al. proposed a deep convolutional neural network model called Basset to predict chromatin accessible regions purely relying on one-hot encoded DNA sequences [12]. Liu et al. developed a

hybrid deep learning model for integrating multiple forms of sequence representations to achieve high prediction performance [14]. Quang et al. used a hybrid convolutional and recurrent neural network for predicting chromatin signals [18]. However, a model purely relying on sequence data can hardly be generalized to make predictions across different cell types as sequence itself is not cell-type-specific. To overcome this limitation, Zhou et al. proposed a regression model called BIRD that utilized only gene expression data to predict chromatin accessible regions [13]. Nevertheless, with the complete removal of sequence data, the scope of application of this method is limited, because the availability of gene expression is not as wide as sequence data. With the above understanding, Nair et al. proposed a deep residual neural network [20] model called ChromDragoNN to combine both sequence and expression data towards the prediction of chromatin accessibility [21]. However, sequence signatures and expression features are combined by simple concatenation in this method. This formulation, though simple in computation, lacks interpretability and is not consistent with existing biological knowledge.

With the above understanding, we propose a method called DeepCAGE, that is, a **Deep** densely connected convolutional network for predicting **C**hromatin **A**ccessibility by incorporating **G**ene **E**xpression and binding status of transcription factors. Unlike BIRD and ChromDragoNN that take full expression data as predictors, our method carefully considers the binding status of chromatin-binding factors (e.g., transcription factors), based on the biological understanding that chromatin accessibility is largely determined by chromatin-binding factors that have access to DNA [2]. In a series of systematic evaluation, DeepCAGE achieves state-of-the-art performance in not only the classification of chromatin accessible status, but also the regression of DNase-seq signals. To make DeepCAGE more understandable, we propose a strategy for visualizing the weights in the first convolutional layer. Interestingly, many known motifs were successfully recovered by DeepCAGE. In the downstream application to

99 whole-genome sequencing (WGS) data analysis, DeepCAGE effectively prioritizes
100 deleterious variants for the prediction and interpretation of complex phenotypes.

101 # Method

102 ## Overview of DeepCAGE

103 DeepCAGE was designed based on the premise that binding status and gene expression
104 of transcription factors could complement sequence data towards the precise prediction
105 of chromatin accessibility. With this understanding, we designed DeepCAGE as a
106 hybrid neural network that consisted of a convolutional module for sequence data and
107 a feedforward module for chromatin accessibility prediction (**Figure 1**). Briefly, we
108 applied the one-hot encoding to the input sequence data, fed the encoded data to a
109 densely connected convolutional neural network (DenseNet), and took the output as the
110 sequence feature. For binding status, we scanned the input sequence for potential
111 binding sites for a set of 402 human transcription factors by using non-redundant motifs
112 in the HOCOMOCO database [22] with the tool Homer [23]. We then selected the
113 maximum score of reported binding sites for each transcription factor to obtain a vector
114 of 402 dimensions as the motif feature. For gene expression, we focused on log-
115 transformed TPM values of the 402 transcription factors and obtained a vector of 402
116 dimensions after quantile normalization as the expression feature. With these data, we
117 combined the two vectors of motif and expression features by taking the element-wise
118 product, and we concatenated the result to the sequence feature to obtain the hybrid
119 feature, which went through a feedforward neural network with a fully connected
120 hidden layer and an output layer for either classification or regression. We presented
121 detailed hyperparameters of the hybrid network in (Table S1).

122 ## Densely connected convolutional network

123 DeepCAGE extracts sequence features by using an architecture called the densely
124 connected convolutional network (DenseNet), which has the advantage of alleviating
125 the vanishing-gradient problem and strengthening the feature propagation [24]. As
126 shown in Figure 1, there are three dense blocks in our model. Each block includes five

127 convolutional layers, and each layer connects to every other layer in a feedforward
128 fashion. A convolutional layer consists of two consecutive small kernels of size 1×1
129 and 3×1, where the former aims at reducing the concatenated channels to a fixed
130 number, and the later acts as the traditional convolution. A transition module is
131 presented before a dense block for feature extracting and dimensionality reduction. An
132 input sequence is first extended to a fixed length of 1000 base pairs (bps) centered at
133 the midpoint of the sequence and then converted to a 1000×4 binary matrix by using
134 the one-hot encoding. The matrix is then fed to the first transition module that contains
135 a convolutional layer and a max-pooling layer. The convolutional layer has 160 kernels
136 of size 4×15 for extracting low-level features and detecting DNA binding motifs,
137 while the max-pooling layer is present for finding the most significant activation signal
138 in a given sliding window of each kernel. Similar settings are used for the other two
139 transition modules for extracting high-level features and dimensionality reduction.
140 Rectified linear units (ReLU) are used after each convolution operation for keeping
141 positive activations and setting negative activation values to zeros. Batch normalization
142 [25] and dropout [26] strategies are used after each ReLU function for reducing internal
143 covariate shift and avoiding overfitting, respectively. For DeepCAGE regression model,
144 there are two major differences from the classification model. First, the output layer
145 directly uses a linear transformation instead of a sigmoid function. Second, the mean
146 square error (MSE) instead of the cross-entropy is used as the loss function.

147 **Data processing**

148 DNase-seq bam files and narrow peaks across 55 human cell types were downloaded
149 from the ENCODE project [5] (Table S2-S3). The human hg19 reference genome was
150 divided into non-overlapping regions (locus) of 200 bps. Considering that a cell type
151 may have multiple DNase-seq replicates, a locus is regarded as chromatin accessible if
152 it overlaps with narrow peak regions of at least half of the replicates and inaccessible
153 otherwise (Figure S1). For the classification design, a binary label y_{lk} is assigned to
154 locus l, representing whether it is accessible in cell type k. For the regression design,

155 bam files of multiple replicates for a cell type are pooled, and the raw read count, n_{lk},
156 is obtained for locus l in cell type k. To eliminate the effect of sequencing depths, the
157 normalized read count, $\tilde{n}_{lk} = Nn_{lk}/N_k$, is calculated, where N_k denotes the total
158 number of pooled reads for cell type k and $N = \min\{N_k\}$ the minimal number of
159 pooled reads across all cell types. The normalized read counts are further log
160 transformed after adding a pseudocount of one. The transformed data represent the level
161 of chromatin accessibility and are then used as the response variable in the regression
162 model.

163 RNA-seq data across the same 55 human cell types were also downloaded from the
164 ENCODE project (Table S4). Gene expression data (TPM) of the 402 core human
165 transcription factors were extracted. After further log transformation and quantile
166 normalization based on TPM values, the normalized expression within each cell type
167 was averaged across multiple replicates and the mean expression profile of each cell
168 type was finally used.

169 Whole-genome sequencing (WGS) data and RNA-seq profiles of GTEx muscle
170 tissues were downloaded from dbGaP (accession number phs000424.v7.p2). Matching
171 these two types of data, a total of 491 donors were selected for downstream analysis
172 (Table S5). For each of these donors, RNA-seq data was processed in the same way as
173 ENCODE data, and WGS data was filtered by excluding all indels and rare variants
174 whose minor allele frequencies were less than or equal to 5 across all donors.

175 **Model evaluation**

176 Cell-type-level five-fold cross-validation experiments are designed for evaluating our
177 method. In each fold, the 55 cell types are partitioned into a training set with 44 cell
178 types and a test set with the remaining 11 cell types (Table S6-S7). Putative known
179 accessible loci are identified as genomic regions (loci) that are chromatin accessible in
180 at least two cell types in the training set. Putative novel accessible loci are identified as
181 genomic regions that are accessible in at least two test cell types and are not present in
182 the training data.

Cell-type-wise and locus-wise metrics are defined to evaluate our method from different perspectives (Figure S2). Cell-type-wise metrics are calculated within a test cell type across genomic regions to provide high level assessment of a method. Locus-wise metrics are calculated based on a genomic region across cell types to give detailed analysis of the performance. These metrics provide a comprehensive and systematic evaluation of our method in both the classification and the regression design.

Let $\boldsymbol{Y}_{L\times K}$ and $\widehat{\boldsymbol{Y}}_{L\times K}$ be the true label matrix and predicted matrix, where L denotes the number of putative loci and K that of cell types. In the classification design, y_{lk} and $\hat{y}_{lk}$ denote the true binary label and predicted probability of chromatin accessible status for locus l in cell type k, respectively. In this situation, the cell-type-wise auPR (area under the precision-recall curve) for cell type k is calculated based on $\boldsymbol{y}_{*k} = (y_{1k}, y_{2k}, \dots, y_{Lk})$ and $\widehat{\boldsymbol{y}}_{*k} = (\hat{y}_{1k}, \hat{y}_{2k}, \dots, \hat{y}_{Lk})$ as follows. Given a threshold t for a cell type k, the precision is defined as the number of correct predictions ($\sum_l y_{lk} I(\hat{y}_{lk} > t)$) over the number of all predictions ($\sum_l I(\hat{y}_{lk} > t)$), and the recall is defined as the number of correct predictions over the number of true accessible loci ($\sum_l y_{lk}$), where $I(x)$ is an indicator function that is equal to 1 if x is true and 0 otherwise. Varying the threshold from 0 to 1 and calculating the precision and recall at each threshold value, the precision-recall curve can be drawn, and the area under this curve can be obtained. The locus-wise auPR for locus l is calculated based on $\boldsymbol{y}_{l*} = (y_{l1}, y_{l2}, \dots, y_{lK})$ and $\widehat{\boldsymbol{y}}_{l*} = (\hat{y}_{l1}, \hat{y}_{l2}, \dots, \hat{y}_{lK})$ in a similar way.

In the regression design, y_{lk} and $\hat{y}_{lk}$ denote the true and predicted DNase-seq signal for locus l in cell type k, respectively. In this situation, the cell-type-wise PCC for cell type k is calculated as the Pearson correlation coefficient of $\boldsymbol{y}_{*k}$ and $\widehat{\boldsymbol{y}}_{*k}$, and the locus-wise PCC is calculated based on $\boldsymbol{y}_{l*}$ and $\widehat{\boldsymbol{y}}_{l*}$ in a similar way. The prediction squared error (PSR), which considers both cell-type-wise prediction and locus-wise prediction, is calculated as $\mathrm{PSR} = \sum_k \sum_l (y_{lk} - \hat{y}_{lk})^2 / \sum_k \sum_l (y_{lk} - \bar{y}_{*k})^2$, where $\bar{y}_{*k} = \sum_l y_{lk} / L$ is the mean of $\boldsymbol{y}_{*k}$.

Two statistics, cell range and cell variability, are introduced to describe the activity of a locus based on the true DNase-seq signal across test cell types. The cell range of locus l is calculated by $\max(\boldsymbol{y}_{l*}) - \min(\boldsymbol{y}_{l*})$, and the cell variability of locus l is defined by the standard deviation of $\boldsymbol{y}_{l*}$.

Baseline methods

Basset [12], DeepSEA [10], and DanQ [18] are three representative neural network models that take as input only DNA sequences. BIRD [13] is a regression model that takes as input only gene expression data. ChromDragoNN [21] is a neural network-based model that takes as input both DNA sequence and gene expression. Our method and ChromDragoNN have the following major differences. First, design principles of these two methods are notably different. ChromDragoNN predicts chromatin accessibility through direct concatenating of DNA sequence and expression data of all genes. DeepCAGE explains chromatin accessibility with DNA sequence and binding status of transcription factors. Therefore, DeepCAGE tries to interpret chromatin accessibility in a more natural way, since chromatin accessibility is believed to be largely determined by the occupancy and topological organization of nucleosomes as well as chromatin-binding factors [2]. Second, network architectures of these two methods are different. ChromDragoNN uses a ResNet to extract sequence features, while DeepCAGE uses a DenseNet that is a relatively new architecture and has also been experimentally validated to outperform ResNet in many tasks [20]. Third, inputs of these two methods are also different. ChromDragoNN requires DNA sequence and expression data of all genes, while DeepCAGE takes as input DNA sequence and expression data of 402 human core transcription factors. Motif binding profiles of these transcription factors can be annotated with existing motif database, which can be precomputed without additional experimental cost.

Gradient importance score

DeepCAGE takes advantage of the gradient importance score (GIS) to prioritize transcription factors given a pair of cell type and a genomic locus. Briefly, a locus is

extended to a 200kb genomic region centered at the midpoint of the locus. Then, the average absolute gradient of predicted accessibility within the extended region with respect to the expression of a transcription factor is calculated as

$$GIS_{ki} = \frac{1}{|\mathcal{L}|}\sum_{l\in\mathcal{L}}\left|\frac{\partial \hat{y}_{lk}}{\partial g_{ki}}\right|,$$

where, $\hat{y}_{lk}$ denotes the predicted accessibility of locus l in cell type k, g_{ki} the expression of transcription factor i in cell type k, and $\mathcal{L}$ the set of putative regulatory elements that contains all accessible loci within the extended region. The gradient importance score gives an intuition of which transcription factors play an important role in a specific cell type.

Motif analysis

The weights of the kernels from the first convolutional layer are converted into position weight matrices (PWMs) by counting subsequence occurrences in a set of input sequences that activate a kernel at a threshold value. All subsequences with activation value that greater than the threshold of a kernel are pooled together and aligned. The PWMs are then composed of the frequencies of the 4 nucleotides (A, C, G, and T) at each position. A subsequence at position i is regarded as activated if

$$\sum_{m=0}^{M-1}\sum_{n=0}^{N-1} w_{m,n}^{k} x_{i+m,n}^{j} > \alpha \cdot MAV^{k},$$

where $M \times N$ denotes the size of the kernels (4×15 in the first convolutional layer). MAV^{k} denotes the maximal activation value of kernel k and is represented as

$$MAV^{k} = \max_{i,j}\left(\sum_{m=0}^{M-1}\sum_{n=0}^{N-1} w_{m,n}^{k} x_{i+m,n}^{j}\right).$$

α is the control coefficient with the default value of 0.7 in all experiments.

Motifs are identified using the tool TomTom (v4.12.0) [27] with the E-value threshold 0.05 and are compared to known motifs in the JASPAR database (version 2018) [28]. Besides, the information content of recovered motifs is calculated based on the information entropy, as $\mathrm{IC} = \sum_{i,j}(p_{ij}\log_2 p_{ij} - b_i \log_2 b_i)$, where p_{ij} is the element in PWM matrix, i and j the nucleotide type and position, respectively, and b_i (default value=0.25) the background frequency of nucleotide i.

265 **Phenotype prediction**

266 A linear regression model with l_1 penalty is adopted to predict the heights of GTEx
267 donors using the deleterious scores of variants, as

268 $$h = \alpha_0 + \sum_{k=1}^{K} \alpha_k \Delta O_k,$$

269 where h is the height of a GTEx donor, and ΔO_k denotes the deleterious score of
270 variant k calculated using DeepCAGE. The coefficient of the l_1 penalty is set to 0.5.
271 A ten-fold cross validation experiment is used in validation, and the average coefficient
272 of determinant (R^2) is used for evaluating how much variance in the phenotype can be
273 explained.

274 ## Results

275 **DeepCAGE accurately predicts binary chromatin accessibility status**

276 We first evaluated the performance of DeepCAGE in predicting whether an input DNA
277 sequence is chromatin accessible or not. To achieve this objective, we downloaded
278 paired DNase-seq and RNA-seq data across 55 cell types from the ENCODE project
279 [5] and conducted a five-fold cross-validation experiment at the cell type level. In each
280 fold of the validation, we partitioned the data into a training set of 44 cell types and a
281 test set of the remaining 11 cell types. We then defined putative known accessible loci
282 as genomic regions that are chromatin accessible in at least two cell types in the training
283 data. For each cell type, we further identified a positive set as putative loci that were
284 accessible in the cell type and a negative set as those inaccessible. After that, we trained
285 our model on the training data and classified positive loci against negative ones for each
286 test cell types. Finally, we calculated a criterion called the cell-type-wise auPR (see
287 Methods) to evaluate the performance of a classification method.

288 We compared the performance of DeepCAGE with four existing methods, including
289 Basset [12], DeepSEA [10], DanQ [18], and ChromDragoNN [21] in the above cross-
290 validation experiment. Results (**Figure 2**A) show that DeepCAGE achieves the highest
291 performance among all the methods with the mean cell-type-wise auPR of 0.418,
292 compared to 0.166 of Basset, 0.195 of DeepSEA, 0.188 of DanQ, and 0.319 of

293 ChromDragoNN. Particularly, DeepCAGE outperforms sequence-based methods by a
294 large margin, suggesting that these methods may fail in capturing cell-type-specific
295 information. Further analysis shows that the proportion of positive loci is in general
296 small in a cell type and exhibits large variation (ranging from 2.6% to 29%), suggesting
297 the ability of our method in dealing with unbalanced data.

298 We then took one step further to access the ability of our method in predicting novel
299 chromatin accessible loci. In each fold of the validation experiment, we identified
300 putative novel accessible loci as genomic regions that were accessible in at least two
301 test cell types and were not present in the training data, and we applied the trained
302 model to predict whether these loci were accessible or not in a test cell type. Results,
303 as shown in Figure 2A, also suggest the superiority of DeepCAGE with a mean cell-
304 type-wise auPR of 0.181, compared to 0.107 of Basset, 0.104 of DeepSEA, 0.110 of
305 DanQ, and 0.151 of ChromDragoNN.

306 We finally analyzed how the cell type specificity of accessible regions affects the
307 prediction performance of our method. To achieve this objective, we divided the
308 putative known accessible loci into three groups based on the proportion of cell types
309 in which a locus is accessible. We then evaluated the cross-validation results using a
310 criterion called the locus-wise auPR that evaluated the prediction performance of a
311 method on an accessible locus across cell types (see Methods). Results show that for
312 locus accessible in 10% cell types or less, DeepCAGE achieves a mean locus-wise
313 auPR of 0.578, and this criterion increases when a locus is accessible in more cell types
314 (Figure 2B). These results suggest that the cell type specificity is likely a factor that
315 affects prediction performance of a method.

316 **DeepCAGE recovers continuous degree of chromatin accessibility**

317 In the above classification experiments, we only consider the binary accessible status
318 of a genomic region in a specific cell type. In the real situation, however, the
319 accessibility of a genomic region given by a DNase-seq experiment is in continuous
320 form. Considering this situation, we further proposed a DeepCAGE regression model

321 to predict for a DNA region the degree of chromatin accessibility, which is defined as
322 the normalized average count of raw reads that fall into the corresponding region.
323 With the same cross-validation settings as in the above section, we compared the
324 performance of DeepCAGE to two baseline methods, BIRD [13] and ChromDragoNN
325 [21], and we assessed regression results in term of two criteria, the cell-type-wise
326 Pearson's correlation coefficient (PCC) and prediction squared error (PSE) (see
327 Methods). Results show that DeepCAGE achieves a mean cell-type-wise PCC of 0.785,
328 compared to 0.637 for BIRD and 0.735 for ChromDragoNN (**Figure 3**A). Further
329 analysis shows that in 18.2% of the test cell types, DeepCAGE achieves a cell-type-
330 wise PCC of 0.85 or higher. In two cell types, DeepCAGE even achieves a cell-type-
331 wise PCC of 0.9 or higher (see examples in Figure 3B). DeepCAGE also achieves the
332 minimal PSE (0.420), outperforming the two baseline methods (0.770 for BIRD and
333 0.570 for ChromDragoNN) by a quite large margin (Figure 3C).
334 We then explored the performance of DeepCAGE for putative accessible loci with
335 different cell type specificity by introducing two statistics, cell range and cell variability,
336 to describe the activity dynamics of a genomic region based on the true DNase-seq
337 signal cross cell types (see Methods). We divided known and novel accessible loci into
338 three groups (low, median and high) according to the 1/3 and 2/3 quantiles of these
339 statistics. Results show that DeepCAGE has high performance for accessible loci with
340 medium cell range and cell variability (Figure 3D), consistent with the results in BIRD
341 [13]. Briefly, DeepCAGE achieves a median locus-wise PCC (see Methods) of 0.512
342 for known accessible loci with medium cell range, compared to 0.435 and 0.399 for
343 loci with low and high cell ranges, respectively. When using the statistic of cell
344 variability, DeepCAGE achieves a median locus-wise PCC of 0.384, 0.514, and 0.448
345 for known accessible loci with low, median and high cell variabilities, respectively. The
346 results are similar for novel accessible loci, except that values of the criteria are slightly
347 low. We further divided known accessible loci into five groups based on the number of
348 cell types in which a locus is accessible. Results (Figure 3E) show that the performance

349 of DeepCAGE varies a lot for loci accessible in different numbers of cell types. Briefly,
350 the performance is high for loci accessible in medium proportion of cell types and low
351 for those accessible in only a small proportion of cell types.

352 Finally, we visualized in Figure 3F both the true (green) and predicted (yellow)
353 DNase-seq signal of a sample genomic region across three test cell types (GM12878,
354 HepG2, and H1-hESC) in the UCSC genome browser [29]. In contrast, we also
355 provided as a reference the mean signal (red) that was calculated by taking the average
356 DNase-seq signals across all training cell types. Obviously, DeepCAGE well
357 distinguishes the difference of DNase-seq signals among the three test cell types while
358 the mean signal fails.

359 **Model ablation analysis of DeepCAGE**

360 We studied the contributions of gene expression and binding scores of transcription
361 factors to the performance of our method. Taking the DeepCAGE regression model as
362 an example, by discarding gene expression data, the median cell-type-wise PCC
363 decreased by 13.1% (Figure S3, p-value=6.53×10^{-11} in a one-sided paired-sample
364 Wilcoxon signed-rank test). By removing binding scores, the median cell-type-wise
365 PCC decreased by 3.6% (Figure S3, p-value=3.78×10^{-4} in a one-sided paired-sample
366 Wilcoxon signed-rank test). These results suggest that gene expression data could
367 significantly help improve the performance of DeepCAGE in cross-cell-type prediction,
368 while binding scores slightly increase the performance. One potential reason behind
369 this observation is that a large proportion of DNA sequence motifs have already been
370 learned in the convolution layers of the neural network, and thus the binding scores
371 only provided complementary information regarding DNA sequence features.

372 Besides, to demonstrate the superiority of the network architecture used by
373 DeepCAGE, we additionally conducted the following two experiments. First, we
374 replaced the DenseNet with a ResNet which had the same number of layers as the
375 number of dense blocks and the same hidden nodes in the convolutional layers. Results
376 show that DenseNet leads to 6.4% increment in performance over ResNet in terms of

377 the median cell-type-wise PCC (Figure S4, *p*-value=3.15×10^{-6} in a one-sided paired-
378 sample Wilcoxon signed-rank test). Second, we explored the influence of two key
379 hyperparameters (the number of residual blocks and the convolutional layers within a
380 residual block) on the performance of ChromDragoNN. It is noted that a deeper model
381 architecture does not help improve the performance significantly (Figure S5).

382 **Gradient importance score helps prioritize cell-type related transcription factors**

383 We proposed a strategy for prioritizing cell-type related transcription factors according
384 to the absolute gradient of the predicted accessibility with respect to the expression of
385 a transcription factor. Taking the K562 cell line as an example, we calculated the
386 average gradient importance score (GIS) of all transcription factors from all putative
387 loci within up-streaming 100k bp to down-streaming 100k bp of a tumor suppressor
388 gene *TP53*, which was shown to have a key role in myeloid blast transformation [30].
389 The average gradient importance score across cell types and transcription factors with
390 respect to the transcription start site (TSS) of this gene was shown in **Figure 4**A. The
391 402 human core transcription factors were then prioritized by the average GIS score in
392 K562 cell line (Figure 4B). Interestingly, many top-ranked transcription factors were
393 related to functions in leukemia cell validated by literatures. For example, *EGR*-1
394 (rank^{1st}) was involved in regulating PMA-induced megakaryocytic differentiation of
395 K562 cell line [31]; the inhibition of *E2F7* (rank^{3rd}) might lead to a reduction of
396 miRNAs involved in leukemic cell lines [32]; the expression of *JunB* (rank^{5th}) was
397 inactivated by methylation in chronic myeloid leukemia [33]. The GO terms enriched
398 by the top 5% prioritized transcription factor coding genes also included biological
399 processes of leukocyte differentiation and hematopoietic development (Figure 4C). To
400 sum up, the GIS score gives us an intuitive interpretation of which transcription factor
401 may play an important role in predicting the chromatin accessibility given a specific
402 cell type and a genomic region.

403 **DeepCAGE automatically learns binding motifs of transcription factors**

404 In order to make DeepCAGE more understandable, we explored the features that were
405 automatically learned by DeepCAGE by investigating the weights of the 160 kernels in
406 the first convolutional layer. Briefly, we converted the weights into position weight
407 matrices (PWMs) (see Methods) and then compared them with known motifs in the
408 JASPAR database [28]. We found that 48 (30%) of the kernels could match known
409 motifs at the E-value threshold of 0.05. Among the matched kernels, 25 (52%) had at
410 least one matched core human transcription factor used in DeepCAGE model. We then
411 calculated the information content (see Methods), set the weights of each kernel to zeros,
412 and denoted the decrease in the cell-type-wise PCC as the influence score for each
413 kernel. We showed several learned unmatched motifs that have a high influence score
414 (**Figure 5**A) and illustrated a few examples of learned motifs that could match known
415 motifs in JASPAR database (Figure 5B). These results demonstrate that DeepCAGE
416 can not only help us find potential binding motifs, but also has the potential to guide
417 the finding of novel motifs which are not discovered by experiments yet.

418 **DeepCAGE prioritizes putative deleterious variants in WGS**

419 We applied DeepCAGE to whole-genome sequencing (WGS) data analysis and
420 demonstrated how our method could benefit the detection of individual-specific
421 deleterious variants in regulatory elements that potentially influenced phenotype. The
422 principle was to quantify the degree that a genetic variant affected the chromatin
423 accessibility of a nearby genomic region and then prioritize variants accordingly. As
424 shown in **Figure 6**A, for an individual, we fed the individual genome and the reference
425 genome separately to the trained DeepCAGE regression model and calculated
426 prediction scores for each of them. We then took the absolute log fold change of these
427 two scores as a measure of the change in chromatin accessibility. For a variant, we
428 defined its individual-level deleterious score by the change of chromatin accessibility
429 of a 200bp genomic region around. Finally, we obtained the cohort-level deleterious
430 score for a variant by applying the above procedure to all individuals in a cohort who
431 contain the variant and then averaged individual-level deleterious scores for the variant.

432 Note that we also took as input the expression profile of transcription factors in the
433 muscle tissue and only consider WGS variant with the minor allele frequency larger
434 than 5.

435 We downloaded WGS data of 491 donors with the height phenotype from the
436 database of Genotypes and Phenotypes (dbGap) of GTEx project (Table S5). We
437 collected 3290 risk single nucleic polymorphisms (SNPs) that were associated with
438 height by a large-scale genome-wide association study [34]. For each risk SNP, we
439 defined a risk region as a 200Kb genomic region centered at the SNP. We then ranked
440 SNPs within a risk region according to their cohort-level deleterious scores obtained
441 from donors (Figure 6B). As an illustration, we examined the risk region around a risk
442 SNP (rs5742714) in the promoter region of *IGF1*, a well-known growth factor [35].
443 The top-ranked variants within this risk region showed an obviously higher absolute
444 log fold change of average height than the bottom-ranked variants (Figure 6C).

445 We then quantitatively explored how much variance of the height phenotype can be
446 explained by the deleterious scores of risk variants. To achieve this objective, we
447 proposed a linear regression model with l_1 penalty, which took deleterious scores of a
448 set of variants as predictors and the height phenotype as the response variable (see
449 methods). Results show that the 1,103,572 WGS variants within the 3290 risk regions
450 together interpreted 2.49% of the height variance. Furthermore, the variants ranked
451 among top 10% according to their deleterious scores in each risk region together can
452 interpret 2.11% of the height variance. These results suggest that the small portion of
453 variants prioritized by our method already contained most information that is helpful in
454 the explanation of the phenotype. We also noticed that the bottom ranked 10% variants,
455 on the contrary, failed to interpret the height phenotype (Figure 6D). To conclude,
456 DeepCAGE is capable of giving a fine mapping of putative risk genetic variants and
457 prioritizing WGS variants that might be associated with a specific phenotype.

458

459 **Discussion**

In this paper, we have introduced a deep learning framework called DeepCAGE towards genome-wide prediction of chromatin accessibility. A hallmark of our method is the incorporation of sequence data and binding status of transcription factors into a unified deep neural network. With these two types of information complementing each other, our method overcomes limitations of existing approaches and demonstrates state-of-the-art performance in not only classification but also regression of chromatin accessibility signals. Our method provides insights into functional genomics in two aspects. First, the gradient importance score can give us an intuitional measurement of the contribution of a transcription factor to the regulation with respect to a specific locus in a certain cell type. Second, the visualization of convolutional kernels demonstrates that features automatically extracted by our method are not only consistent with existing knowledge, but also containing potentially novel binding motifs of transcription factors. Such interpretability of our model will benefit the dissection of regulatory landscape under a variety of cell conditions. Our method also provides the possibility of interpreting and prioritizing putative deleterious variants in genetic studies. Such ability in explaining complex traits can further be explored to promote the understanding of disease-associated genetic variants.

Certainly, our model can be further improved from the following aspects. First, currently we ignore the expression of genes that direct the synthesis of proteins other than transcription factors. However, it has been shown that proteins such as chromatin regulators, a class of enzymes with specialized function domains, can shape and maintain the epigenetic state in a cell context-dependent fashion [36], and thus can also provide information for inferring chromatin accessible state. How to incorporate information of these chromatin regulators into our model is one of the directions in our future work. Second, our model currently identifies chromatin accessible regions in a cell-type-specific manner, but cannot further distinguish the specific type of potential regulatory elements in these regions. With the accumulation of annotations regarding *cis*-regulatory elements such as enhancers [37] and silencers [38], as well as

computational methods for predicting interactions between these elements [39-41], it is expected that our framework can further be extended to uncover the comprehensive relationship between different types of genomic regulatory elements and the genome-wide transcriptomic profile.

Code availability

DeepCAGE is freely available at https://github.com/kimmo1019/DeepCAGE with step-by-step instructions.

CRediT author statement

Qiao Liu: Conceptualization, Software, Formal analysis, Writing - Original Draft, Writing - Review & Editing, Visualization. **Kui Hua:** Writing - Review & Editing. **Xuegong Zhang:** Supervision. **Wing Hung Wong:** Conceptualization, Investigation, Supervision, Writing - Review & Editing, Funding acquisition. **Rui Jiang:** Conceptualization, Investigation, Supervision, Writing - Review & Editing, Funding acquisition. All authors read and approved the final manuscript.

Competing interests

The authors have declared no competing interests.

Acknowledgments

This work has been partly supported by the National Natural Science Foundation of China (Grant Nos. 61721003, 61873141, 61573207), the National Key R&D Program of China (Grant Nov. 2018YFC0910404), and the Tsinghua-Fuzhou Institute for Data Technology; This work was also supported by National Institutes of Health grants (Grant Nov. P50HG007735, R01HG010359). The Genotype-Tissue Expression project was supported by the Common Fund of the Office of the Director of the National

Institutes of Health. We thank Mengmeng Wu, Zhana Duren and Fengling Chen for their helpful discussions.

Authors' ORCID IDs

0000-0002-9781-3360 (Qiao Liu)

0000-0003-2228-7025 (Kui Kua)

0000-0002-9684-5643 (Xuegong Zhang)

0000-0001-7466-2339 (Wing Hung Wong)

0000-0002-7533-3753 (Rui Jiang)

References

[1] Kellis M, Wold B, Snyder MP, Bernstein BE, Kundaje A, Marinov GK, et al. Defining functional DNA elements in the human genome. Proc Natl Acad Sci USA 2014;111:6131-8.

[2] Klemm SL, Shipony Z, Greenleaf WJ. Chromatin accessibility and the regulatory epigenome. Nat Rev Genet 2019;20:207-20.

[3] Crawford GE, Holt IE, Whittle J, Webb BD, Tai D, Davis S, et al. Genome-wide mapping of DNase hypersensitive sites using massively parallel signature sequencing (MPSS). Genome Res 2006;16:123-31.

[4] Buenrostro JD, Giresi PG, Zaba LC, Chang HY, Greenleaf WJ. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. Nat Methods 2013;10:1213.

[5] Consortium EP. An integrated encyclopedia of DNA elements in the human genome. Nature 2012;489:57-74.

[6] Roadmap Epigenomics C, Kundaje A, Meuleman W, Ernst J, Bilenky M, Yen A, et al. Integrative analysis of 111 reference human epigenomes. Nature 2015;518:317-30.

[7] Corces MR, Granja JM, Shams S, Louie BH, Seoane JA, Zhou W, et al. The chromatin accessibility landscape of primary human cancers. Science 2018;362.

[8] Trevino AE, Sinnott-Armstrong N, Andersen J, Yoon S-J, Huber N, Pritchard JK, et al. Chromatin accessibility dynamics in a model of human forebrain development. Science 2020;367.

[9] Song S, Cui H, Chen S, Liu Q, Jiang R. EpiFIT: functional interpretation of transcription factors based on combination of sequence and epigenetic information. Quant Biol 2019;7:233-43.

[10] Zhou J, Troyanskaya OG. Predicting effects of noncoding variants with deep learning-based sequence model. Nat Methods 2015;12:931-4.

[11] Liu Q, Gan M, Jiang R. A sequence-based method to predict the impact of regulatory variants using random forest. BMC Syst Biol 2017;11:7.
[12] Kelley DR, Snoek J, Rinn JL. Basset: learning the regulatory code of the accessible genome with deep convolutional neural networks. Genome Res 2016;26:990-9.
[13] Zhou W, Sherwood B, Ji Z, Xue Y, Du F, Bai J, et al. Genome-wide prediction of DNase I hypersensitivity using gene expression. Nat Commun 2017;8:1038.
[14] Liu Q, Xia F, Yin Q, Jiang R. Chromatin accessibility prediction via a hybrid deep convolutional neural network. Bioinformatics 2018;34:732-8.
[15] Min X, Zeng W, Chen N, Chen T, Jiang R. Chromatin accessibility prediction via convolutional long short-term memory networks with k-mer embedding. Bioinformatics 2017;33:i92-i101.
[16] Xu C, Liu Q, Zhou J, Xie M, Feng J, Jiang T. Quantifying functional impact of non-coding variants with multi-task Bayesian neural network. Bioinformatics 2020;36:1397-404.
[17] Yin Q, Wu M, Liu Q, Lv H, Jiang R. DeepHistone: a deep learning approach to predicting histone modifications. BMC genomics 2019;20:11-23.
[18] Quang D, Xie X. DanQ: a hybrid convolutional and recurrent deep neural network for quantifying the function of DNA sequences. Nucleic Acids Res 2016;44:e107-e.
[19] Ding K, Liu Q, Lee E, Zhou M, Lu A, Zhang S. Feature-Enhanced Graph Networks for Genetic Mutational Prediction Using Histopathological Images in Colon Cancer. International Conference on Medical Image Computing and Computer-Assisted Intervention 2020:294-304.
[20] He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. Proceedings of the IEEE conference on computer vision and pattern recognition 2016:770-8.
[21] Nair S, Kim DS, Perricone J, Kundaje A. Integrating regulatory DNA sequence and gene expression to predict genome-wide chromatin accessibility across cellular contexts. Bioinformatics 2019;35:i108-i16.
[22] Kulakovskiy IV, Vorontsov IE, Yevshin IS, Soboleva AV, Kasianov AS, Ashoor H, et al. HOCOMOCO: expansion and enhancement of the collection of transcription factor binding sites models. Nucleic Acids Res 2015;44:D116-D25.
[23] Heinz S, Benner C, Spann N, Bertolino E, Lin YC, Laslo P, et al. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. Molecular cell 2010;38:576-89.
[24] Huang G, Liu Z, Weinberger KQ, van der Maaten L. Densely connected convolutional networks. Proceedings of the IEEE conference on computer vision and pattern recognition 2017;1:3.
[25] Ioffe S, Szegedy C (2015), 'Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift', in Francis B., David B. Eds., *Proceedings of the 32nd International Conference on Machine Learning*, PMLR, Proceedings of Machine Learning Research, pp. 448--56.

[26] Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R. Dropout: A simple way to prevent neural networks from overfitting. The Journal of Machine Learning Research 2014;15:1929-58.
[27] Gupta S, Stamatoyannopoulos JA, Bailey TL, Noble WS. Quantifying similarity between motifs. Genome Biol 2007;8:R24.
[28] Khan A, Fornes O, Stigliani A, Gheorghe M, Castro-Mondragon JA, van der Lee R, et al. JASPAR 2018: update of the open-access database of transcription factor binding profiles and its web framework. Nucleic Acids Res 2017;46:D260-D6.
[29] Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, et al. The human genome browser at UCSC. Genome Res 2002;12:996-1006.
[30] Law JC, Ritke MK, Yalowich JC, Leder GH, Ferrell RE. Mutational inactivation of the p53 gene in the human erythroid leukemic K562 cell line. Leukemia research 1993;17:1045-50.
[31] Cheng T, Wang Y, Dai W. Transcription factor egr-1 is involved in phorbol 12-myristate 13-acetate-induced megakaryocytic differentiation of K562 cells. J Biol Chem 1994;269:30848-53.
[32] Gabra MM, Salmena L. microRNAs and Acute Myeloid Leukemia chemoresistance: a mechanistic overview. Frontiers in oncology 2017;7.
[33] Yang M-Y, Liu T-C, Chang J-G, Lin P-M, Lin S-F. JunB gene expression is inactivated by methylation in chronic myeloid leukemia. Blood 2003;101:3205-11.
[34] Yengo L, Sidorenko J, Kemper KE, Zheng Z, Wood AR, Weedon MN, et al. Meta-analysis of genome-wide association studies for height and body mass index in~ 700000 individuals of European ancestry. Hum Mol Genet 2018;27:3641-9.
[35] Becker NS, Verdu P, Georges M, Duquesnoy P, Froment A, Amselem S, et al. The role of GHR and IGF1 genes in the genetic determination of African pygmies' short stature. Eur J Hum Genet 2013;21:653-8.
[36] Chen T, Dent SY. Chromatin modifiers and remodellers: regulators of cellular differentiation. Nat Rev Genet 2014;15:93.
[37] Khan A, Zhang X. dbSUPER: a database of super-enhancers in mouse and human genome. Nucleic Acids Res 2016;44:D164-D71.
[38] Zeng W, Chen S, Cui X, Chen X, Gao Z, Jiang R. SilencerDB: a comprehensive database of silencers. Nucleic Acids Res 2021;49:D221-D8.
[39] Li W, Wong WH, Jiang R. DeepTACT: predicting 3D chromatin contacts via bootstrapping deep learning. Nucleic Acids Res 2019;47:e60-e.
[40] Liu Q, Lv H, Jiang R. hicGAN infers super resolution Hi-C data with generative adversarial networks. Bioinformatics 2019;35:i99-i107.
[41] Singh S, Yang Y, Poczos B, Ma J. Predicting enhancer-promoter interaction from genomic sequence with deep neural networks. Quant Bio 2019;7:122-37.

632 # Figure legends

633 **Figure 1 Overview of the DeepCAGE model**

634 The sequence of the input DNA region is converted to a one-hot matrix and goes
635 through a densely connected deep convolutional neural network (DenseNet) to extract
636 sequence features. Normalized expression levels of the 402 human transcription factors
637 and the corresponding motif binding scores are combined by using an element-wise
638 product and then concatenated with sequence features. The combined features are
639 finally fed to a feedforward neural network for chromatin accessibility prediction.

640 **Figure 2 Performance of the DeepCAGE classification model**

641 **A.** DeepCAGE achieves the highest cell-type-wise auPR for both known accessible loci
642 and novel accessible loci compared to baseline methods (Basset, DeepSEA, DanQ and
643 ChromDragoNN). **B.** The performance of DeepCAGE for loci with different activities
644 across test cell types.

645 **Figure 3 Performance of the DeepCAGE regression model**

646 **A.** DeepCAGE predicts DNase-seq signals in five test cell types. **B.** Cell-type-wise PCC
647 for three different methods across test cell types. * Two-sided paired-sample Wilcoxon
648 signed-rank test p-value=3.37×10^{-5}. **C.** Prediction square error (PSR) for three
649 different methods. **D.** Locus-wise PCC achieved by DeepCAGE with respect to two
650 statistics with both known accessible loci and novel accessible loci. **E.** Locus-wise PCC
651 achieved by DeepCAGE considering the number of accessible cell types under known
652 accessible loci. **F.** An example of true (green) and predicted (yellow) DNase-seq signal
653 of three test cell types under a same genomic region (chr1:42.83-42.93M). Mean signal
654 (red) denotes the average DNase-seq signal across training cell types.

655

656

657 **Figure 4 Gradient importance score helps identify important transcription**
658 **factors**
659 **A**. The GIS heatmap of the 402 human core transcription factors across 55 cell types.
660 **B.** The gradient importance score of the 20 top-ranked transcription factors in the K562
661 cell type. **C.** Enriched GO terms by top-ranked transcription factors in the K562 cell
662 type.

663 **Figure 5 DeepCAGE recovers both known and novel motifs**
664 **A.** Green dots and yellow dots represent known and novel motifs recovered by
665 DeepCAGE, respectively. The x-axis describes the information content (see Methods)
666 and the y-axis denotes the influence score. **B.** Matched motifs with an E-value threshold
667 0.05 in the format of sequence logos (above: known motif from the JASPAR database,
668 below: motif learned by DeepCAGE).

669 **Figure 6 DeepCAGE helps prioritize and interpret WGS variants**
670 **A.** The deleterious score is calculated by the absolute value of log change fold of
671 predicted chromatin accessibility of the reference and WGS genomes. **B.** WGS variants
672 within a risk region were ranked by averaging deleterious scores across donors
673 containing the variant. **C.** The absolute log change fold of average height with respect
674 to top-K and bottom-K ranked variants (K=20,40,80) around a height-associated gene.
675 * indicates p-value<0.05. **D.** Predicting phenotype height with deleterious scores with
676 all variants, top-ranked variants and bottom-ranked variants.
677

678 # Supplementary material

679 **Figure S1 Identification of putative accessible loci**

680 Putative accessible loci were determined by DNase-seq peaks according to the
681 following three steps. Step 1: The human reference genome was divided into non-
682 overlapping regions of 200 bps. Step 2: Regions is kept if it is contained in half of
683 replicated of a cell type. Step 3: Putative accessible loci were determined by collecting
684 regions that appear in at least two cell types.

685 **Figure S2 Definition of cell type-wise and locus-wise criteria**

686 In the cell-type-wise evaluation, the area under the precision-recall curve (auPR) and
687 Pearson's correlation were calculated based on rows of both label matrix and predicted
688 matrix. In the locus-wise evaluation, the Pearson's correlation coefficient was
689 calculated based on columns of both label matrix and predicted matrix.

690 **Figure S3 Model ablation analysis of DeepCAGE**

691 **A.** Four models were designed by considering different inputs. **B.** By removing both
692 expression and motif scores of transcription factors (model 1), the median cell type-
693 wise PCC decreases from 0.795 to 0.660. If only gene expression (model 2) or motif
694 scores (model 3) are discarded, the median cell type-wise PCC decreases to 0.664 and
695 0.759, respectively.

696 **Figure S4 Ablation study for model architecture of DeepCAGE**

697 **A.** We implemented DeepCAGE model with two different architectures (DenseNet and
698 ResNet). Note that we used a ResNet with three layers (equal to the number of dense
699 blocks) and the number of hidden nodes and the convolutional kernel size is the same
700 as convolutional layers in DenseNet. **B.** DeepCAGE with DenseNet architecture
701 achieves a median cell-type-wise PCC of 0.795 while DeepCAGE with ResNet
702 architecture achieves an average cell-type-wise PCC of 0.731.

703 **Figure S5 The cross-cell-type prediction performance of ChromDragoNN with**
704 **different hyperparameter settings**

705 **A.** The cell-type-wise PCC of ChromDragoNN with different number of convolutional
706 layers in each residual block. **B.** The cell-type-wise PCC of ChromDragoNN with
707 different number of residual blocks.
708

709 **Table S1 Hyperparameters of the DeepCAGE model**

710 **Table S2 The information of Dnase-seq peaks across 55 cell types from the**
711 **ENCODE project**

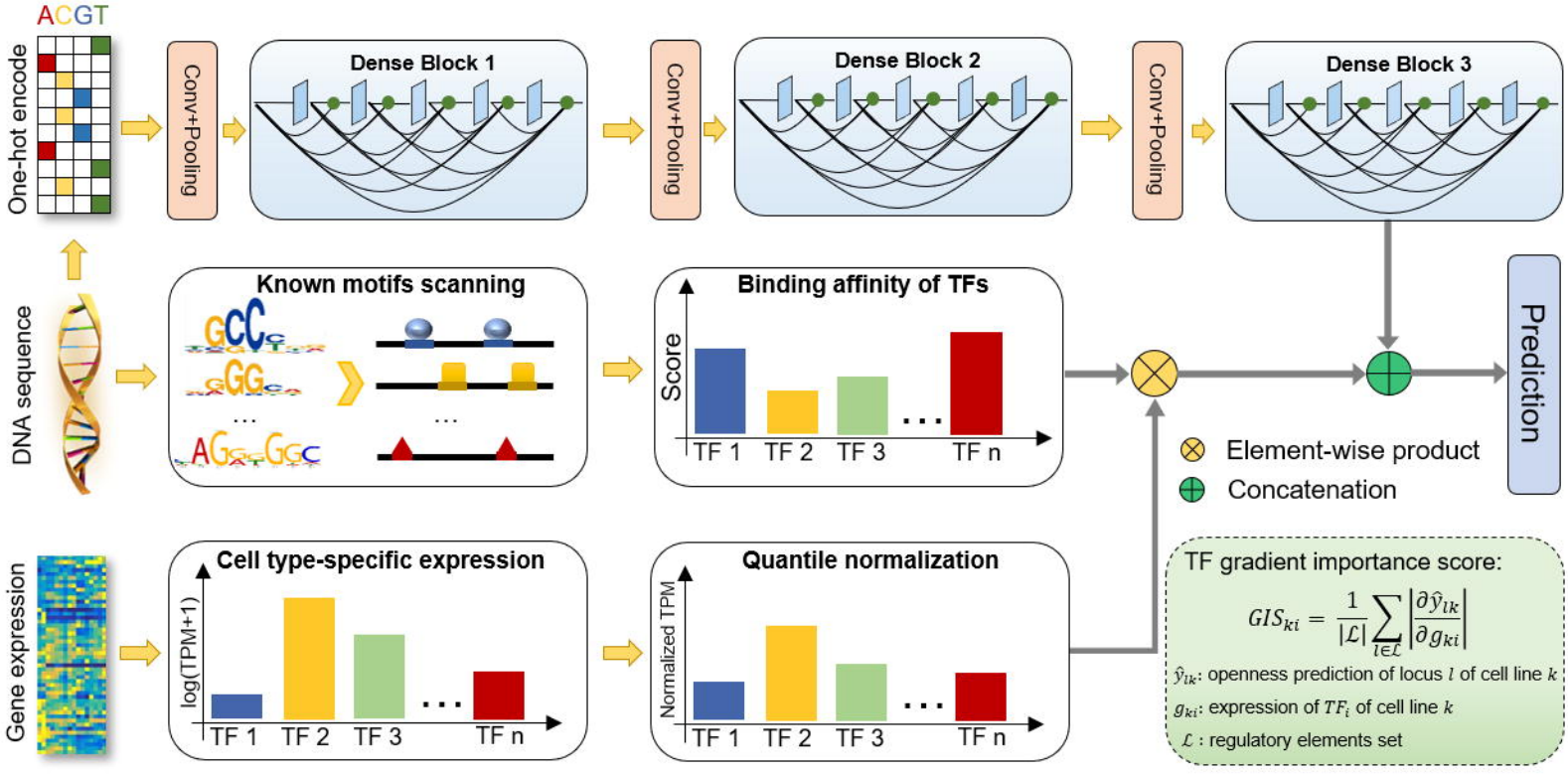
712 **Table S3 The information of Dnase-seq bam file across 55 cell types from the**
713 **ENCODE project**

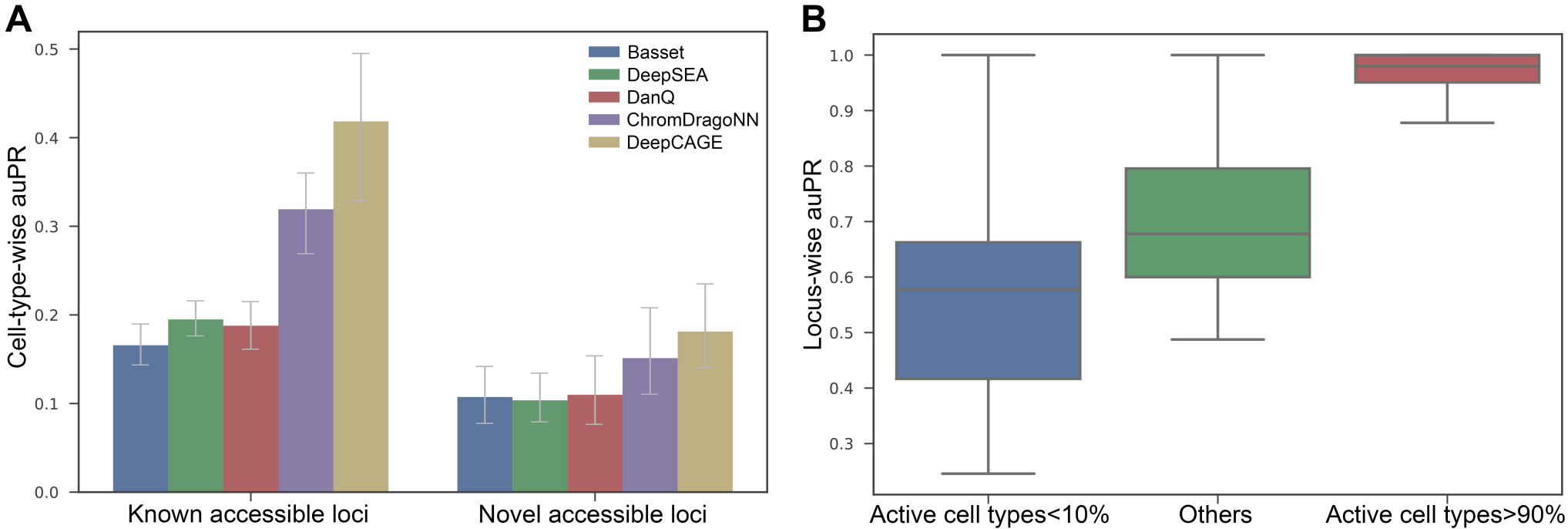
714 **Table S4 The information of RNA-seq data across 55 cell types from the**
715 **ENCODE project**

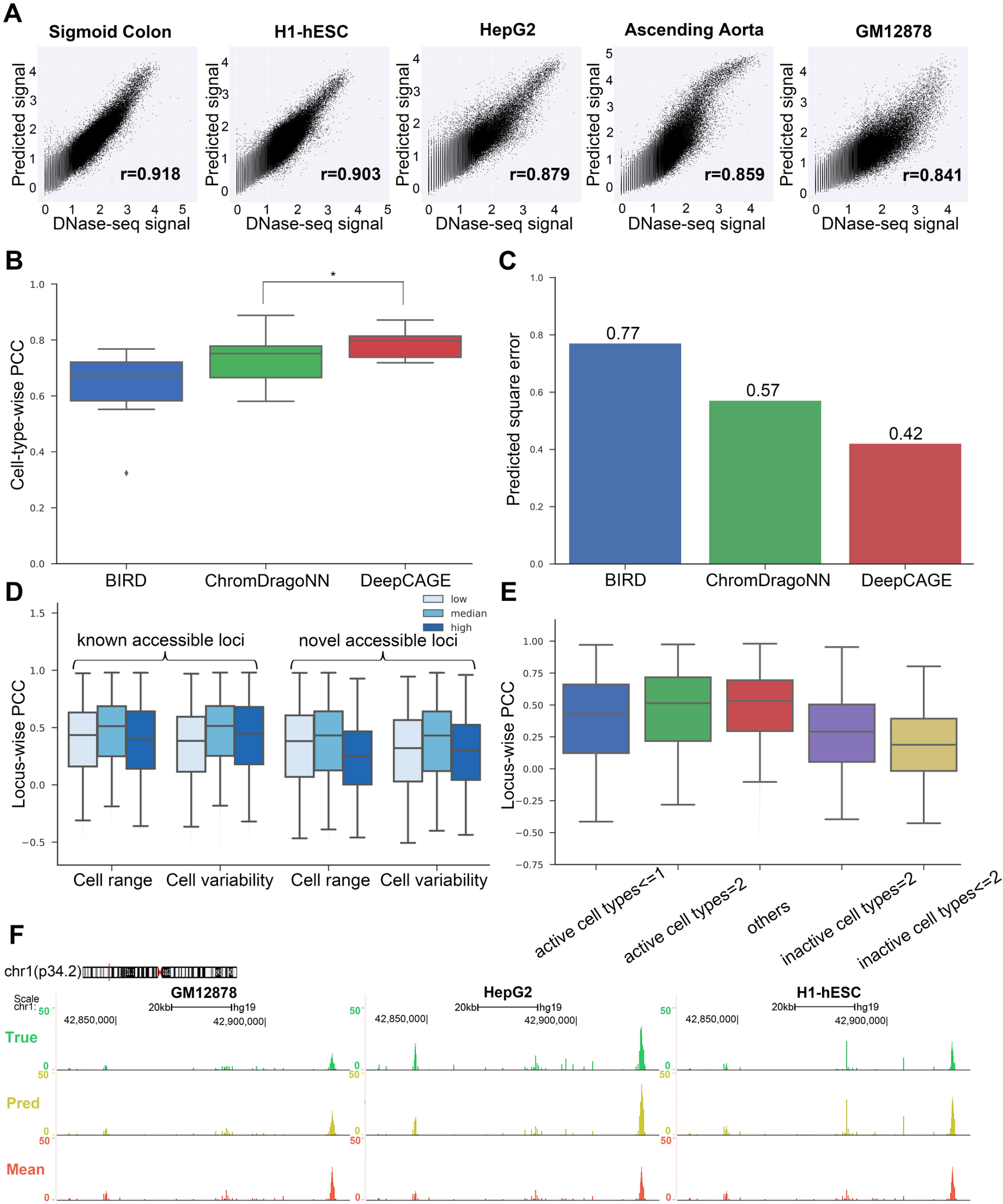
716 **Table S5 The information of GTEx data collected from 491 donors**

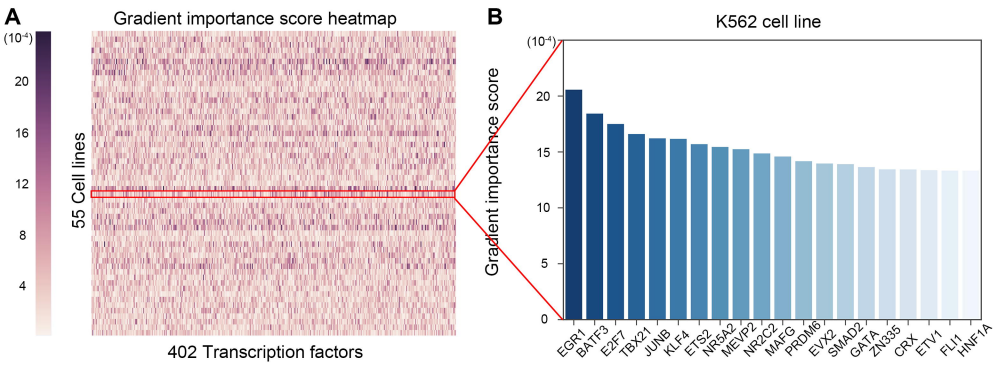
717 **Table S6 Information of the 55 cell types used in this study**

718 **Table S7 Data partition in the five-fold cross-validation experiment**

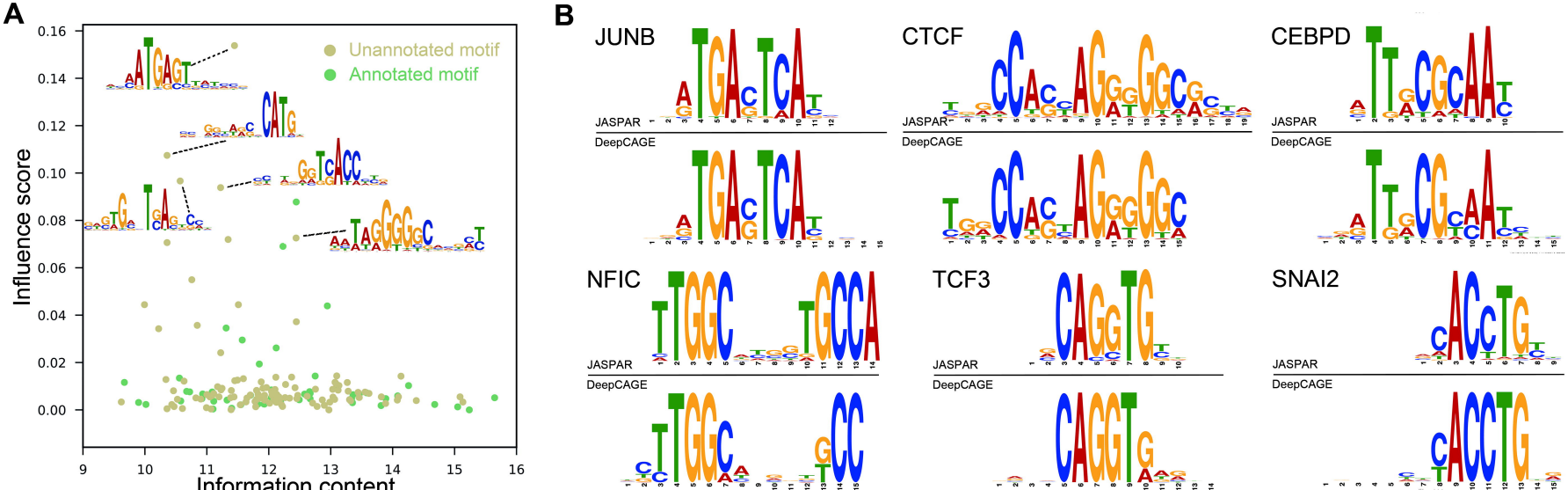

ACGT
One-hot encode
Conv+Pooling
Dense Block 1
Conv+Pooling
Dense Block 2
Conv+Pooling
Dense Block 3
DNA sequence
Known motifs scanning
Binding affinity of TFs
Score
TF 1
TF 2
TF 3
TF n
Prediction
Element-wise product
Concatenation
Gene expression
Cell type-specific expression
log(TPM+1)
TF 1
TF 2
TF 3
TF n
Quantile normalization
Normalized TPM
TF 1
TF 2
TF 3
TF n
TF gradient importance score:
$GIS_{ki} = \frac{1}{|\mathcal{L}|}\sum_{l\in\mathcal{L}}\left|\frac{\partial \hat{y}_{lk}}{\partial g_{ki}}\right|$
$\hat{y}_{lk}$: openness prediction of locus l of cell line k
g_{ki}: expression of TF_i of cell line k
$\mathcal{L}$: regulatory elements set

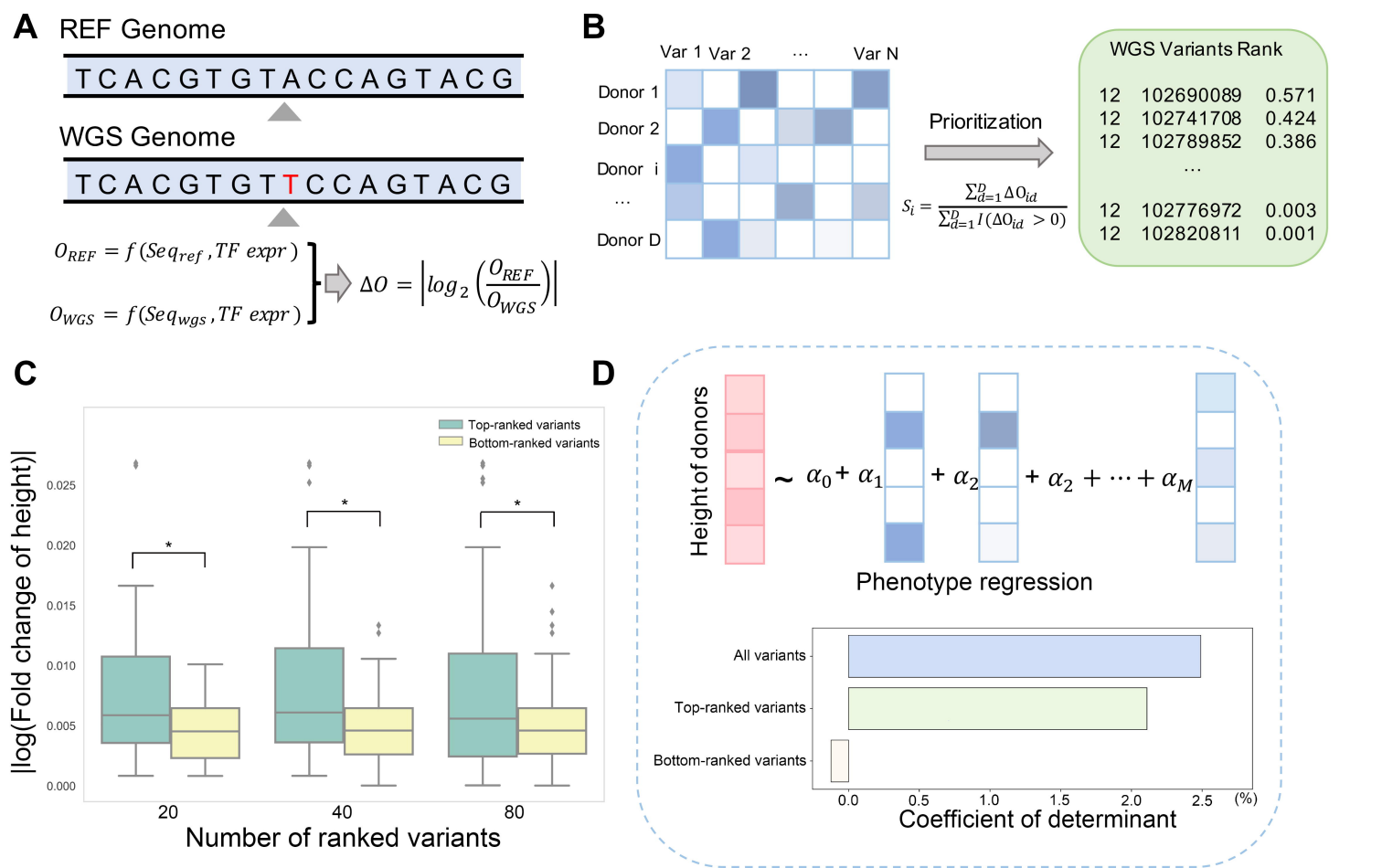

A
Cell-type-wise auPR
0.0
0.1
0.2
0.3
0.4
0.5
Basset
DeepSEA
DanQ
ChromDragoNN
DeepCAGE
Known accessible loci
Novel accessible loci
B
Locus-wise auPR
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
Active cell types<10%
Others
Active cell types>90%


A
Sigmoid Colon
H1-hESC
HepG2
Ascending Aorta
GM12878
Predicted signal
DNase-seq signal
r=0.918
r=0.903
r=0.879
r=0.859
r=0.841
B
Cell-type-wise PCC
BIRD
ChromDragoNN
DeepCAGE
C
Predicted square error
0.77
0.57
0.42
D
Locus-wise PCC
low
median
high
known accessible loci
novel accessible loci
Cell range
Cell variability
E
active cell types<=1
active cell types=2
others
inactive cell types=2
inactive cell types<=2
F
chr1(p34.2)
GM12878
HepG2
H1-hESC
Scale
chr1:
20kb
hg19
42,850,000
42,900,000
True
Pred
Mean



GO term ID	Biological process	P-value
GO:0030097	hemopoiesis	2.45e-06
GO:0048534	hematopoietic or lymphoid organ development	3.69e-06
GO:0002520	immune system development	4.78e-06
GO:0002521	leukocyte differentiation	9.99e-06
GO:0043537	negative regulation of blood vessel endothelial cell migration	1.13e-04


A
Unannotated motif
Annotated motif
Influence score
Information content
0.16
0.14
0.12
0.10
0.08
0.06
0.04
0.02
0.00
9
10
11
12
13
14
15
16
B
JUNB
CTCF
CEBPD
NFIC
TCF3
SNAI2
JASPAR
DeepCAGE


A
REF Genome
TCACGTGTACCAGTACG
WGS Genome
TCACGTGTTCCAGTACG
$O_{REF} = f(Seq_{ref}, TF\ expr)$
$O_{WGS} = f(Seq_{wgs}, TF\ expr)$
$\Delta O = \left|log_2\left(\frac{O_{REF}}{O_{WGS}}\right)\right|$
B
Var 1 Var 2 ... Var N
Donor 1
Donor 2
Donor i
...
Donor D
Prioritization
$S_i = \frac{\sum_{d=1}^{D} \Delta O_{id}}{\sum_{d=1}^{D} I(\Delta O_{id} > 0)}$
WGS Variants Rank
12 102690089 0.571
12 102741708 0.424
12 102789852 0.386
...
12 102776972 0.003
12 102820811 0.001
C
Top-ranked variants
Bottom-ranked variants
|log(Fold change of height)|
0.025
0.020
0.015
0.010
0.005
0.000
*
*
*
20
40
80
Number of ranked variants
D
Height of donors
$\sim \alpha_0 + \alpha_1 + \alpha_2 + \alpha_2 + \cdots + \alpha_M$
Phenotype regression
All variants
Top-ranked variants
Bottom-ranked variants
0.0
0.5
1.0
1.5
2.0
2.5
(%)
Coefficient of determinant