bioRxiv preprint doi: https://doi.org/10.1101/036129; this version posted January 7, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

PEDLA: predicting enhancers with a deep learning-based algorithmic framework

Feng Liu[1], Hao Li[1], Chao Ren[1], Xiaochen Bo[1*], Wenjie Shu[1*]

[1]Department of Biotechnology, Beijing Institute of Radiation Medicine, Beijing 100850, China

[*]Corresponding author. To whom correspondence should be addressed. Tel & Fax: +86 10 68210077 66932211; Email: shuwj@bmi.ac.cn. Correspondence may also be addressed to: boxc@bmi.ac.cn.

bioRxiv preprint doi: https://doi.org/10.1101/036129; this version posted January 7, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Abstract

Transcriptional enhancers are non-coding segments of DNA that play a central role in the spatiotemporal regulation of gene expression programs. However, systematically and precisely predicting enhancers on a genome-wide scale remain a major challenge. Although existing methods have achieved some success in enhancer prediction, they still suffer from a limited number of training samples, a simplicity of features, class-imbalanced data, and inconsistent performance across diverse cell types/tissues. Here, we developed a deep learning-based algorithmic framework named PEDLA (https://github.com/wenjiegroup/PEDLA), which can directly learn an enhancer predictor from massively heterogeneous data and generalize in ways that are mostly consistent across various cell types/tissues. We first trained PEDLA with 1,114-dimensional heterogeneous features in H1 cells, and we demonstrated that our PEDLA framework integrates diverse heterogeneous features and gives state-of-the-art performance relative to five existing methods for enhancer prediction. We further extended PEDLA to continuously learn from 22 training cell types/tissues, and the results showed that PEDLA manifested superior performance consistency in both training and independent test sets. On average, PEDLA achieved 95.0% accuracy and a 96.8% geometric mean (GM) across 22 training cell types/tissues, as well as 95.7% accuracy and a 96.8% GM across 20 independent test cell types/tissues. Together, our work illustrates the power of harnessing state-of-the-art deep learning techniques to consistently identify regulatory elements at a genome-wide scale from massively heterogeneous data across diverse cell types/tissues.

bioRxiv preprint doi: https://doi.org/10.1101/036129; this version posted January 7, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Introduction

Enhancers are distal *cis*-acting DNA regulatory elements that play key roles in controlling cell type-/tissue-specific gene expression (1-4). In higher eukaryotes, enhancers recruit transcription factors (TFs) and cofactors to orchestrate vital biological processes including development and differentiation (5,6), maintenance of cell identity (7-9), response to stimuli (10-12), and interactions with target genes through promoter-enhancer looping (13-16). Genetic variation or disruption in enhancers is closely associated with diseases and cancers (17,18). Although enhancers have long been recognized for their importance in gene regulation and disease, the absence of common sequence features, their distal location from target genes, and their high cell type/tissue specificity have made them challenging to systematically and precisely identify.

The first genome-scale efforts to identify enhancers were based on evolutionary sequence conservation (19-22) because regulatory sequences are likely to evolve under negative evolutionary selection (23-25). Owing to the remarkable advances in next-generation sequencing (NGS) technologies, several high-throughput experimental technologies have been developed to predict enhancers in a genome-wide manner. These approaches involve mapping the binding sites of enhancer indicative transcription factors (TFs) (26-29) or cofactors (30-32) by ChIP-Seq, identifying open chromatin with techniques such as DNase-Seq (33-35), and interrogating covalent histone modifications by ChIP-Seq (7,12,36,37). Given the

bioRxiv preprint doi: https://doi.org/10.1101/036129; this version posted January 7, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

diversity and complexity of these high-throughput data, several computational approaches using supervised or unsupervised machine-learning (ML) algorithms have been developed for genome-wide enhancer prediction. Chromia (38) uses a hidden Markov model (HMM) to predict regulatory elements, CSI-ANN (39) introduces an artificial neural network (ANN) approach, RFECS (40) uses random forests to discriminate enhancers from non-enhancers, and DELTA (41) uses an adaptive boosting (AdaBoost) approach. ChromaGenSVM (42), EnhancerFinder (43) and DEEP (44) are all based on support vector machine (SVM) classifiers. The unsupervised learning approaches, ChromHMM (45) and Segway (46), offer genomic segmentation and characterization based on an HMM and a dynamic Bayesian network, respectively.

Although these ML-based computational approaches have achieved remarkable advances in genome-wide identification of enhancers, there remain some critical issues to be settled. First, the current computational approaches are limited by a small number of training set samples; thus, they are unlikely to produce a useful general classifier. Second, most of the computational approaches use merely data derived from histone modification marks. Chromia (38), CSI-ANN (39), RFECS (40), DELTA (41), ChromaGenSVM (42), and DEEP-ENCODE (44) combine different numbers and combinations of histone modification marks to predict enhancers. These methods have achieved some success in identifying thousands of enhancers in the human genome. However, it remains a challenging task for these approaches to

bioRxiv preprint doi: https://doi.org/10.1101/036129; this version posted January 7, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

systematically and precisely predict enhancers, given the simplicity of features derived from only one type of data and the limited number of chromatin modifications that have been examined. Additionally, it is infeasible to fully capture an exhaustive catalogue of chromatin modifications at enhancers. EnhancerFinder (43) tackles this issue by simultaneously integrating diverse types of data instead of using a single type of data, thereby improving its developmental enhancer predictions. Third, the class imbalance problem in which the number of enhancer classes is much smaller than the number of non-enhancer classes is extremely common in predicting enhancers (39,40,42,44), and ignoring this problem will result in a bias towards the non-enhancer class. RFECS (40), DELTA (41) and DEEP (44) use ensemble techniques to train classifiers with unbalanced classes by combining a set of individual classifiers. Finally, and most importantly, consistent performance across various cell types/tissues is a key challenge in genome-wide enhancer predictions as enhancers are highly cell type-/tissue-specific. The current computational techniques are either trained in a new model for each new cell type/tissue or predict enhancers in new cell types/tissues based on the learned model derived from one cell type/tissue. Regarding the former strategy, this approach is time-consuming and sometimes infeasible given the data availability of training sets for each cell type/tissue, whereas the latter strategy relies on a universal model derived from single cell type/tissue data to effectively identify enhancers in all other cell types/tissues. DEEP-ENCODE (44) alleviates this problem by combining four cell-specific models containing 4,000 (4 × 1,000) classifiers in total, and it achieved performance consistency across two other

bioRxiv preprint doi: https://doi.org/10.1101/036129; this version posted January 7, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

cell lines. However, this is far from sufficient as there exist hundreds and thousands of cell lines and tissue types.

To address these fundamental challenges, we developed a deep learning-based algorithmic framework named PEDLA, which is capable of learning an enhancer predictor that integrates massively heterogeneous data and generalizes in ways that are mostly consistent across various cell types/tissues. We used 1,114-dimensional features in total derived from nine categories of heterogeneous data to predict enhancers in H1 cells. The results show that our PEDLA method has three key advantages: limitless training samples, integration from massively heterogeneous data, and the embedded capability of handling highly class-imbalanced data in an unbiased way. Due to these factors, PEDLA shows state-of-the-art performance relative to five existing methods. Furthermore, we extended PEDLA to continuously learn from 22 training cell types/tissues from the ENCODE Project (47) and the Roadmap Epigenomics Project (48). The learned optimal model manifested superior classification performance and significant performance consistency in 22 training cell types/tissues and another 20 independent cell types/tissues. Together, our results illustrate the power of harnessing state-of-the-art deep learning techniques to consistently identify regulatory elements on a genome-wide scale from massively heterogeneous data across diverse cell types/tissues.

bioRxiv preprint doi: https://doi.org/10.1101/036129; this version posted January 7, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Materials and Methods

Feature data

For the construction of PEDLA to predict enhancers in H1 cells, we used nine sets of features, including histone modifications, TFs and cofactors, chromatin accessibility, transcription, DNA methylation, CpG islands, evolutionary conservation, sequence signatures, and occupancy of TF binding sites (TFBS). Features derived from ChIP-Seq data included 27 histone modification marks and 15 TFs and cofactors analysed in the ENCODE Project (47) and the Roadmap Epigenomics Project (48). Attributes coming from chromatin accessibility and transcription were obtained from DNase-Seq and mRNA-Seq in the Roadmap Epigenomics Project (48). Characteristics derived from DNA methylation were obtained from RRBS (Reduced Representation Bisulfite-Seq) in the Roadmap Epigenomics Project (48). Features obtained from CpG islands and evolutionary conservation were taken from the CpG island track and the vertebrate phastCons46way track (49) in the UCSC Genome Browser (50), respectively. A total of 340 sequence characteristics were derived by estimating the frequency of all possible *k*-mers (k = 1, 2, 3, 4) in the sequence, and 726 attributes derived from occupancy of TFBS motifs were obtained by scanning TF motif occurrences within open chromatin in the human genome using FIMO (51) with a *p*-value $< 10^{-5}$. We extracted 726 TF motifs from Transfac (52), Jaspar (53) and UniProbe (54) and defined open chromatin in the human genome as the merged TFBS-clustered regions across hundreds of cell types (47), which covers nearly all accessible regions in the human genome. In total, 1,114-dimensional features were

bioRxiv preprint doi: https://doi.org/10.1101/036129; this version posted January 7, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

used to predict enhancers with PEDLA in H1 cells (Table S1).

Data normalization for diverse data types

All 1,114 features were processed and normalized at a 200-bp resolution using a unified RPKM-like measure (reads per kilobase per million) (55). The ChIP-Seq mapping reads for the histone modifications, TFs and cofactors and the corresponding inputs were binned into 200-bp intervals, and the RPKM value of each bin was calculated. The DNase-Seq and RNA-Seq mapping reads were also binned into 200 bp intervals, and the RPKM value of each bin was calculated. For two or more replicates of these sequencing data, the RPKM value for each bin was averaged to obtain a single RPKM value. The signals of RRBS, the CpG island, evolutionary conservation, *k*-mers (k = 1, 2, 3, 4) sequence characteristics, and occupancies of TFBS motifs in each bin were normalized against the total signal of the whole genome and the length of the bin, similar to the RPKM calculation. Finally, the 1,114 feature signal values were scaled to the range of [0,1].

Construction of the positive and negative sets

Considering that there are no 'gold standard enhancers' that have been experimentally verified across various cell types/tissues, we constructed the enhancer set (positive set) using histone marker H3K27ac, which is viewed as an active enhancer hallmark (9,56-58) and is ubiquitously available across dozens of cell types/tissues. To

bioRxiv preprint doi: https://doi.org/10.1101/036129; this version posted January 7, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

construct the positive set, we started from the candidate H3K27ac peaks called by MACS (59) with a *p*-value < 10^{-9} for each replicate independently. Then, we selected the candidate H3K27ac peaks that were consistently identified in multiple replications. Next, we strictly used the highly reliable H3K27ac peaks, which met the following criteria as described previously (10): (1) H3K27ac peaks that were located within 1 kb of all annotated transcription start sites (TSSs) of protein-coding genes annotated by GENCODE (V15) (60) were removed. (2) H3K27ac peaks that had a 5'-sequenced expressed sequence tag (EST) with a 5' end within 2 kb of the H3K27ac peak and spanning an annotated TSS were removed on the basis of annotations of ESTs from the UCSC Genome Browser spliced EST track. (3) H3K27ac peaks that had significant enrichment of both H3K4me1 and H3K4me3 in a 2-kb window centred on the peak were removed. (4) H3K27ac peaks with significant enrichment of H3K4me3 within a 2 kb window centred on the peak were also removed. (5) An H3K4me1 peak had to be present within 2 kb of the H3K27ac peak. (6) The H3K27ac peak had to be at least 2 kb away from all rRNA genes.

Non-enhancers (the negative data set) contained an equal number of promoters and *x* times the number of random genomic loci not annotated as promoters or enhancers. Promoters are defined as 2-kb regions centred on TSSs of protein-coding genes, and random genomic loci were generated with an equivalent length distribution as enhancer regions. For each learning set, we set $x = 9$ to maintain a ratio of 1:10 between positive and negative samples, which is consistent with previous studies

bioRxiv preprint doi: https://doi.org/10.1101/036129; this version posted January 7, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

(39,40,42,44). In the analysis of class imbalance problems, we changed x from 1 to 9.

For enhancer and non-enhancer sets of each cell type/tissue, all the 200-bp intervals that overlapped with any enhancers or non-enhancers were compiled to be the final training samples for the development and assessment of our PEDLA. The average length of all enhancer and non-enhancer regions was approximately 2 kb. Therefore, we obtained approximately $(1+1+x)\times n\times 2{,}000/200 = 10n\times(2+x)$ training samples for n experimentally determined enhancers in the positive set. For H1 cells, we had 5,870 enhancers, and the number of training samples in the H1 cell line was 811,036.

Training the PEDLA for enhancer predictions

Enhancer prediction with PEDLA, which follows our recently reported method (61), involved two stages, namely training and prediction procedures (Fig. S1).

1. Training procedure

1) Read the normalized features and annotated labels as input data.
2) Estimate the initial state probability vector π and state transition probability matrix A of HMM.
3) Pre-train the DNN layer-by-layer in an unsupervised fashion, then use the pre-training as initialization to fine-tune the weight matrix W of the DNN using mini-batch gradient descent in a supervised manner.
4) Estimate the prior probability vector $P = \{p(q_t = S_j)\} = \{n(S_j)/n\}$, where $n(S_j)$ is the number of samples labelled with state S_j, and n is the total number of all

bioRxiv preprint doi: https://doi.org/10.1101/036129; this version posted January 7, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

samples.

2. Prediction procedure

1) Calculate the output of the DNN to estimate the posterior probability of each state based on the weight matrix W.
2) Divide the posterior probability by the prior probability P to obtain the observation probability distribution B of HMM.
3) Use the Viterbi algorithm to predict the final state of unlabelled data based on the triple (π, A, B).

Performance assessment of PEDLA

To assess and compare the performance of PEDLA with that of the existing methods, we computed the following three performance indicators used in our recent study (61).

$$Accuracy = (TP + TN)/(TP + TN + FN + FP) \quad (1)$$

where TP indicates true positives, FP indicates false positives, TN indicates true negatives, FN indicates false negatives, and

$$GM = \sqrt{Sensitivity \times Specificity} \quad (2)$$

where GM is the geometric mean of sensitivity and specificity, Sensitivity = TP/(TP+FN) and Specificity = TN/(TN+FP), and

$$F1score = 2/\left(\frac{1}{Recall} + \frac{1}{Precision}\right) \quad (3)$$

where the F1 score is the harmonic mean of recall and precision, Recall = Sensitivity and Precision = TP/(TP+FP).

bioRxiv preprint doi: https://doi.org/10.1101/036129; this version posted January 7, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

For PEDLA and other supervised methods, we used 5-fold cross-validation to obtain unbiased estimates of the three performance indicators. For the unsupervised methods, ChromHMM and Segway, we obtained these three performance indicators by comparing their enhancer predictions with the labels of the positive and negative sets.

Selecting optimal structure of PEDLA

To determine the optimal structure of hidden layers for PEDLA, we tuned the structure of hidden layers by fixing the number of hidden layers to 1 or 2 and the number of hidden units in each hidden layer to 10, 50, 100 or 500. We made enhancer predictions using PEDLA for diverse structures that are composites of various hidden layers and hidden units of each layer. We computed the three performance indicators using 5-fold cross-validation and compared them between the structure models of PEDLA to select the optimal structure model of PEDLA in terms of accuracy (Table S2).

Validation of enhancer predictions

To validate the enhancers predicted by PEDLA and other existing methods, we calculated the percentage of the predicted enhancers that overlapped with experimental data on enhancer markers (the validation rate) or TSSs of protein-coding genes (the misclassification rate). The enhancer markers included distal DNase-I hypersensitivity sites (DHSs), binding sites of p300 and TFs such as NANOG, OCT4

bioRxiv preprint doi: https://doi.org/10.1101/036129; this version posted January 7, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

and SOX2. The DHSs were derived from an ENCODE study (http://ftp.ebi.ac.uk/pub/databases/ensembl/encode/integration_data_jan2011/byDataType/openchrom/jan2011/fdrPeaks/) (47), and p300, NANOG, OCT4 and SOX2 binding sites were obtained from a recent study (40). TSS annotations were extracted from the GENCODE annotations (V15) (60). Predicted enhancers overlapping with a window of -100 to +100 bp centred at a distal enhancer marker that was greater than 5 kb away from the nearest TSS were classified as ‘‘validated’’, whereas predicted enhancers located within 2.5 kb of a TSS were classified as ‘‘misclassified’’.

Results

Prediction of enhancers using PEDLA with heterogeneous signatures

To systematically and precisely predict enhancers on a genome-wide scale, we developed a deep learning-based algorithm framework named PEDLA (see Materials and Methods, Fig. S1). Briefly, we first used nine categories of heterogeneous data in H1 cells that serve as discriminating features to identify enhancers, including histone modifications (ChIP-Seq), TFs and cofactors (ChIP-Seq), chromatin accessibility (DNase-Seq), transcription (RNA-Seq), DNA methylation (RRBS), CpG islands, evolutionary conservation, sequence signatures, and occupancy of TFBSs. In total, 1,114-dimensional features were used in the training and testing of PEDLA (Table S1). Then, we constructed the enhancer class (positive class set) containing 5,870 enhancer regions (Table S3) based on H3K27ac peaks in the H1 cell line as no 'gold

bioRxiv preprint doi: https://doi.org/10.1101/036129; this version posted January 7, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

standard' set of enhancers has been experimentally verified across various cell types/tissues. Next, we maintained a ratio of 1:10 between positive and negative samples by choosing an equal number of promoters and nine times the number of random background regions that were not annotated as promoters or enhancers as the non-enhancer class (negative class set). Finally, we optimized the structure of PEDLA by selecting the most optimal model, which consists of two hidden layers with each layer including 500 hidden units, in terms of accuracy using 5-fold cross-validation in both the training set and test set (Fig. 1A, Table S2). On average, PEDLA achieved 97.7% accuracy, a 97.2% GM (96.6% sensitivity and 97.9% specificity) and an 88.7% F1-score (96.6% recall and 82.1% precision) for the training set, as well as 97.7% accuracy, a 97.0% GM (96.2% sensitivity and 97.8% specificity) and an 88.3% F1-score (96.2% recall and 81.7% precision) for the independent test set. These results demonstrated the superior performance and outstanding robustness of PEDLA in both the training and test sets, suggesting that PEDLA has great representational power for heterogeneous data and excellent ability to solve the overfitting problem.

Ability to handle class-imbalanced data in an unbiased manner

To resolve the class imbalance problem, our PEDLA method has an embedded mechanism in which the prior probability $\boldsymbol{P}$ of each class, which reflects the imbalance of the class data, is directly estimated from the training data. Dividing the posterior probability by the prior probability P can fundamentally eliminate class-imbalanced influence (see Materials and Methods, Fig. S1). Thus, for PEDLA, we

bioRxiv preprint doi: https://doi.org/10.1101/036129; this version posted January 7, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

need to train only a single model instead of combining multiple trained classifiers as ensemble approaches do, which generally results in a bloated structure model. To assess the performance of PEDLA in resolving this issue, we maintained the number of enhancers, promoters and random regions not annotated as promoters or enhancers in the training set of the H1 cell line at 1 : 1 : x (x = 1, 2, …, 9), such that the ratio between positive and negative samples was 1 : (1 + x). Three performance indicators, namely sensitivity, specificity and GM, which were reported to be suitable for assessing highly imbalanced data sets, were measured using 5-fold cross-validation based on the optimal model of PEDLA with all 1,114-dimensional features (Fig. 1C-D). For comparison, we additionally developed a classic DNN algorithm and assessed the performance of the classic DNN in resolving this challenge by measuring the three performance indicators (Fig. S1). We found that for PEDLA, all three performance indicators were unbiased for different degrees of imbalance of the class-imbalanced data in both the training and test data. Notably, the performance of PEDLA even showed a certain degree of improvement with increasing imbalance. In contrast, for the classic DNN, the sensitivity and GM performance indicators decreased rapidly as the specificity increased slightly with increasing imbalance of the class data. The higher specificity and the lower sensitivity achieved by the classic DNN for highly class-imbalanced data suggested that the classic DNN tended to be biased towards the majority class, which is similar to the individual SVM for imbalanced data (62). These results indicate that PEDLA manifested greater capability in resolving the class imbalance problem.

bioRxiv preprint doi: https://doi.org/10.1101/036129; this version posted January 7, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Validation of predicted enhancers

To validate the predicted enhancers using PEDLA, we calculated the validation rate as the percentage of predicted enhancers overlapping with experimental data of enhancer markers, including distal DHSs, p300 binding sites and a few sequence-specific TFs known to function in H1 cells, such as NANOG, OCT4 and SOX2 (see Materials and Methods). Additionally, we computed the misclassification rate as the percentage of predicted enhancers overlapping with annotated promoters (see Materials and Methods). In human H1 cells, we identified 20,689 enhancers using the trained PEDLA with optimal structure, and we found that 42.3%, 16.8%, and 32.4% of these predicted enhancers overlapped with distal DHSs and binding sites of p300 and TFs, respectively. With 10,000 iterations of random shuffling of the enhancer predictions in the H1 genome, we found the average validation rate of distal DHS, p300 and TF to be 13.0%, 1.1% and 4.2%, respectively, and the actual validation rates were highly significant according to a one-sided *t*-test (*p*-values = 0) (Fig. 1B). Additionally, we found that 6.6% of enhancer predictions overlapped with TSSs annotated by GENCODE (V15) (60), indicating a low misclassification rate. These results demonstrate that PEDLA can accurately predict putative enhancers based on the nine categories of heterogeneous data.

bioRxiv preprint doi: https://doi.org/10.1101/036129; this version posted January 7, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Performance comparisons of PEDLA with existing methods

Next, we compared our enhancer predictions with those identified by five state-of-the-art methods, including the supervised learning methods CSI-ANN (39), RFECS (40), and DELTA (41) and the unsupervised learning methods ChromHMM (45) and Segway (46), in H1 cells. The optimal model of CSI-ANN was obtained based on 39 histone modifications from CD4+ T cells (39), the best model of RFECS was derived from the profiles of 24 histone modifications in H1 and IMR90 cells (40), whereas the optimal model of DELTA was trained on the optimal subset of histone modifications derived from H1 and CD4+ T cells (41). To make a fair comparison of performance of our method with those of the three supervised approaches, we applied PEDLA and the three supervised learning methods to a common histone modification dataset in H1 cells. For the common dataset, we used H3K4me1, H3K4me3 and H3K27ac as the minimal chromatin marks for enhancer prediction as these three histone markers were previously used as the minimal subset of histone modifications to predict enhancers by both CSI-ANN (39) and RFECS (40) and were also ranked within the top 4 in terms of variable importance in DELTA in H1 cells (41). Using these three histone markers in H1 cells, RFECS, CSI-ANN, and DELTA made 75,084, 30,173, and 112,044 enhancer predictions, respectively, based on the optimal parameters and default threshold used in their studies. PEDLA yielded 22,691 enhancers using the optimal structure of a hidden layer with 50 hidden units determined by 5-fold cross-validation. For ChromHMM and Segway, the enhancer predictions can be directly downloaded and extracted from the ENCODE page

bioRxiv preprint doi: https://doi.org/10.1101/036129; this version posted January 7, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

(http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeAwgSegmentation/), and obtained 26,869 and 131,698 predictions, respectively.

We compared PEDLA with the five existing methods based on five performance indicators: accuracy; GM; F1-score; validation rates of distal DHSs, p300, and TFs; and misclassification rate. The comparative analysis across these methods is summarized in Table 1. Based on the former four performance indicators, PEDLA consistently performed better than all other methods. Based on misclassification rate, DELTA and RFECS shared the best results, followed by ChromHMM, PEDLA, Segway and CSI-ANN. Furthermore, comparing PEDLA of the optimal model based on the three minimal chromatin marks with that based on the full 1,114 features, we found that PEDLA of full features consistently performed better than PEDLA of the three minimal chromatin marks for all five performance indicators. This result suggests that approaches integrating diverse types of data give more complete and precise representation of enhancers and thus can significantly improve enhancer prediction compared to methods using only a single type of data.

Because the different performance indicators illustrate the distinct advantages and disadvantages of these studied methods, we ranked their performance according to the five metrics. In total, we performed seven different tests, including seven methods and five performance indicators. According to a previously published method (63), we averaged the ranked positions of each of the seven methods used in comparison in all

bioRxiv preprint doi: https://doi.org/10.1101/036129; this version posted January 7, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

seven tests. Table 2 demonstrates the overall score and average rank position of each of the methods; a lower average rank indicates better performance. This analysis revealed that across the different performance tests, the PEDLA with both full signature sets and three minimal chromatin marks shared the best results, followed by CSI-ANN, ChromHMM, RFECS, DELTA and Segway. Together, our results convincingly demonstrate that PEDLA integrates diverse heterogeneous features and gives superior performance relative to the existing methods for enhancer predictions.

Prediction of enhancers with PEDLA in multiple human cell types

Predicting enhancers across more human cell types/tissues can help us better understand the underlying regulatory mechanisms of these elements. Currently, most computational approaches use individual models trained on data from one cell type/tissue to predict enhancers in multiple other cell types/tissues. Because enhancers are highly cell type-/tissue-specific, it is infeasible that a model derived from single cell type/tissue data can effectively predict enhancers in all the other types/tissues. To resolve this challenge, DEEP-ENCODE (44) combines four ensemble models that are trained on four cell lines and contain 4,000 (4 × 1000) classifiers in total to predict enhancers in multiple cell types, and the test results in two independent cell lines suggested that DEEP-ENCODE showed better generalization capabilities. However, to achieve consistent performance across various cell types/tissues, DEEP-ENCODE has to combine many cell type-/tissue-specific models, which makes the structure of DEEP-ENCODE more bloated.

bioRxiv preprint doi: https://doi.org/10.1101/036129; this version posted January 7, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Inspired by the core idea of deep learning using unsupervised pre-training followed by supervised fine-tuning, we developed a continuous PEDLA for enhancer prediction across diverse human cells and tissues to resolve the above-mentioned issues. Figure 2 shows a diagram of the steps of training PEDLA in multiple cell types/tissues. Initially, PEDLA was trained based on data derived from any cell type/tissue. Then, PEDLA was trained for a subsequent cell type/tissue iteratively, using the trained model of the previous cell type/tissue as initialization. For each training iteration, PEDLA learned cell type-/tissue-specific enhancer information from the corresponding cell type/tissue and "evolved" continuously. Finally, PEDLA learned enhancer information derived from all the cell types/tissues used in the training, and it showed consistent performance in predicting enhancers across diverse cell types/tissues in the human genome.

To explore the effectiveness of PEDLA to predict enhancers in multiple cell types/tissues, we selected 22 cell types/tissues from the ENCODE Project (47) and the Roadmap Epigenomics Project (48) as the training cell set, and we chose another 20 independent cell types/tissues that were not used in the training procedure as the test cell set (Table S4). The training and test procedures of PEDLA are much more time-consuming when more features are used. Thus, we used the 8-dimensional features, including H3K4me1, H3K4me3, H3K9me3, H3K27me3, H3K36me3, H3K27ac, input of histone modification and evolutionary conservation as the common features

bioRxiv preprint doi: https://doi.org/10.1101/036129; this version posted January 7, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

to train PEDLA across multiple cell types/tissues. The selection of these features is based on the consideration of both the data availability across tens of cell types/tissues and the variable importance of all features in enhancer predictions (Table S5).

PEDLA was trained based on the enhancer and non-enhancer classes that were constructed independently for each of the 22 cell types/tissues in the training cell sets using the optimal model of a hidden layer with 50 hidden units (Fig. 3A). The whole training procedure was repeated with 50 random permutations of the 22 training cell types/tissues (50 out of $P(22, 22) = 22!$), and each random order was repeated four times with random permutations of training samples for each cell type or tissue. In total, the training of PEDLA was repeated independently 200 times on the 22 training cell types/tissues. During each iterative step of the training procedures, we used the optimal model of PEDLA to predict enhancers in the 22 training cell types/tissues and to simultaneously assess performance on another 20 independent test cell types/tissues (Fig. 3B-C). We recorded the three performance indicators (accuracy, GM, and F1-score) across these cell types/tissues, which made it intuitive to interrogate the dynamic performance of PEDLA with the increasing number of cell types/tissues that it had been trained on. We found that the mean values of the performance indicators increased and that the variances of the performance indicators decreased with an increasing number of training cell types/tissues, suggesting that increased performance and increased consistency were obtained for of our PEDLA enhancer predictions, respectively. Furthermore, the three performance indicators increased

bioRxiv preprint doi: https://doi.org/10.1101/036129; this version posted January 7, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

rapidly when the second cell type/tissue was added into the training, indicating that the performance and consistency of PEDLA were greatly improved compared to the performance and consistency of the PEDLA that was trained on only one cell type/tissue. After finishing the training of the five cells/tissues, the performance indicators remained unchanged, suggesting that the performance and the consistency of PEDLA were stable at this stage.

We use the best-trained model of PEDLA that had been completed using the training on all 22 training cell types/tissues to further assess its generalizability using the three performance indicators (Fig. 4; Table S5-6). On average, PEDLA achieved 95.0% accuracy, a 96.8% GM (99.0% sensitivity and 94.6% specificity) and a 78.7% F1-score (99.0% recall and 65.4% precision) for 22 training cell types/tissues, as well as 95.7% accuracy, a 96.8% GM (98.2% sensitivity and 95.3% specificity) and an 81.0% F1-score (98.2% recall and 69.1% precision) for 20 independent test cell types/tissues. This result suggests that PEDLA shows superior performance and consistency across various cell types/tissues and generalizes well to both the training and test type/tissue sets. To further verify the performance consistency of PEDLA across diverse cell types/tissues, we assessed the performance of PEDLA for predicting specific and common enhancers across the cell types/tissues in both the training set and the test set (Fig. 4; Table S5-6). In the 22 training cell types/tissues sets, PEDLA achieved 94.9% accuracy, a 96.5% GM (98.5% sensitivity and 94.6% specificity) and a 74.4% F1-score (98.5% recall and 60.0% precision) for cell type-/tissue-specific enhancers, and

bioRxiv preprint doi: https://doi.org/10.1101/036129; this version posted January 7, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

it achieved 95.1% accuracy, a 97.0% GM (99.4% sensitivity and 94.6% specificity) and an 81.5% F1-score (99.4% recall and 69.1% precision) for common enhancers. Additionally, for both cell type-/tissue-specific and common enhancers, PEDLA achieved consistent performance within 20 independent test cell types/tissues sets. These results demonstrate that PEDLA has the properties of invariability (cell type-/tissue-common) and variability (cell type-/tissue-specificity) in enhancer prediction, further suggesting its superior performance consistency across various cell types/tissues.

Discussion

In this study, we developed a deep learning-based algorithmic framework named PEDLA (https://github.com/wenjiegroup/PEDLA) to systematically and precisely predict enhancers on a genome-wide scale using heterogeneous types of data. PEDLA has three outstanding characteristics that make it ideal for achieving state-of-the-art performance relative to five existing methods for enhancer predictions. First, PEDLA is capable of learning an enhancer predictor based on massively heterogeneous data to fully capture the universal patterns of enhancers, which makes its enhancer predictions more comprehensive and accurate. In H1 cells, we trained PEDLA using 1,114-dimensional heterogeneous features and 811,036 training samples and demonstrated that PEDLA has three key advantages: limitless training samples, integration of massively heterogeneous data, and an embedded ability to handle

bioRxiv preprint doi: https://doi.org/10.1101/036129; this version posted January 7, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

highly class-imbalanced data in an unbiased way (Fig. 1). All these merits make PEDLA a general and robust deep learning-based framework for enhancer predictions.

Second, PEDLA can generalize enhancer predictions in ways that are mostly consistent across various cell types/tissues. We extended PEDLA to continuously learn from multiple cell types/tissues, which is motivated by the core idea of deep learning that uses unsupervised pre-training as initialization of the subsequent supervised fine-tuning. The whole training procedure of PEDLA in multiple cell types/tissues consists of initial training and iterative training (Fig. 2). The initial training procedure adopts a traditional deep learning strategy, whereas the iterative training procedure uses the trained optimal model parameters of a previous cell type/tissue as initialization for the supervised fine-tuning of a subsequent cell type/tissue iteratively. We applied PEDLA to continuously train on 22 training cell types/tissues, and we found that the learned optimal model manifested superior predicting performance and significant performance consistency in 22 training cell types/tissues and another 20 independent cell types/tissues (Figs. 3-4).

Third, PEDLA is capable of extending to input data of any type and to predictions of any type of functional element/domain. Our PEDLA used nine categories of heterogeneous data in H1 cells as discriminating features to identify enhancers. In fact, any type of data that consists of real values can be served as input features to PEDLA. Additionally, the numbers of input features can range from a few to millions.

bioRxiv preprint doi: https://doi.org/10.1101/036129; this version posted January 7, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Furthermore, PEDLA can be trained on training sets of any number of cell types/tissues but will not result in bloated structures. Most importantly, provided with prepared suitable feature data and constructed accurate positive and negative sets for the task, PEDLA can be extended with almost no modifications to identify any types of functional elements/domains, including promoters, insulators, and repressors.

Although PEDLA achieved state-of-the-art performance in enhancer prediction, there is still room for further improvement. First, we carefully selected 1,114-dimensional features derived from nine categories of heterogeneous data in H1 cells; however, it remains challenging to select the comprehensive and discriminating features to identify enhancers; thus, further optimization of the feature set must be continued. Additionally, all these features used by PDELA are two-dimensional (2D). It is expected that integrating three-dimensional (3D) structure information derived from 5C, Hi-C, ChIA-PET and Capture-C will greatly improve the prediction of enhancers, given that 3D chromatin structure is increasingly considered an important regulator of gene expression and that enhancers regulate target genes through promoter-enhancer looping (64,65). A recent study combined Hi-C data with phylogenetic correlations to predict the target genes of distal regulatory elements, such as enhancers, repressors, and insulators (66). Another group demonstrated that modelling Hi-C data with their computational method, called graph-based regularization (GBR), greatly improved the prediction of replication and topological domains (67). These studies inspired us to further improve PEDLA through the incorporation of 3D structure information.

bioRxiv preprint doi: https://doi.org/10.1101/036129; this version posted January 7, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Second, due to the large number of input features and massive training samples, the PEDLA architecture has thousands of weights and millions of connections between units. Whereas training such large networks is time-consuming, progress in algorithm parallelization, distributed computing and GPU acceleration have reduced the training times to a few hours. Third, like many other leading approaches in machine learning, PEDLA based on deep learning is the most data-hungry, requiring thousands and even millions of labelled data to achieve state-of-the-art performance. Thus, reducing the reliance on large amounts of labelled data will greatly improve the applicability of PEDLA. Recently, Brenden M. Lake, *et al.* (68) presented a Bayesian program learning (BPL) framework that can learn a large class of visual concepts from only a single example and manifest creative generalization abilities that are mostly indistinguishable from human behaviour. A novel hybrid system combining deep learning and BPL may therefore provide an efficient strategy to dramatically reduce the reliance on large amounts of labelled data. Finally, PEDLA can be directly extended to predict any known functional elements/domains in the genome but it cannot be used to discover completely novel regulatory elements because PEDLA is supervised in nature. Human learning is largely unsupervised because we discover the world by observing it, not by being told what it is in advance (69). To achieve human-level learning, it becomes far more important to explore the potential computational model based on deep learning for unsupervised learning in the longer term. Taken together, our work illustrates the power of harnessing state-of-the-art deep learning techniques to consistently identify regulatory elements from massively heterogeneous

bioRxiv preprint doi: https://doi.org/10.1101/036129; this version posted January 7, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

data across diverse cell types/tissues.

Availability

Our DNN-HMM method and other three algorithms can be freely accessed at https://github.com/wenjiegroup/PEDLA.

Acknowledgements

We wish to thank the ENCODE Project and Roadmap Epigenomics Project for making their data publicly available.

Funding

This work was supported by grants from the Major Research plan of the National Natural Science Foundation of China (No. U1435222), the Program of International S&T Cooperation (No. 2014DFB30020) and the National High Technology Research and Development Program of China (No. 2015AA020108).

Conflict of interest statement. None declared.

bioRxiv preprint doi: https://doi.org/10.1101/036129; this version posted January 7, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Author Contributions

W.S. conceived the project. W.S. and X.B. designed all experiments. WS drafted the manuscript. FL wrote the programs, analyzed the results. HL and CR and helped in analysis and discussion, gave useful comments. All authors read and approved the final manuscript.

Supplementary Information

Supplemental Information includes two figures, seven tables, and source code of PEDLA algorithm can be found with this article online.

References

1. Bulger, M. and Groudine, M. (2010) Enhancers: the abundance and function of regulatory sequences beyond promoters. *Dev. Biol.*, **339**, 250-257.
2. Ong, C.T. and Corces, V.G. (2011) Enhancer function: new insights into the regulation of tissue-specific gene expression. *Nat. Rev. Genet.*, **12**, 283-293.
3. Calo, E. and Wysocka, J. (2013) Modification of enhancer chromatin: what, how, and why? *Mol. Cell*, **49**, 825-837.
4. Bulger, M. and Groudine, M. (2011) Functional and mechanistic diversity of distal transcription enhancers. *Cell*, **144**, 327-339.
5. Bonn, S., Zinzen, R.P., Girardot, C., Gustafson, E.H., Perez-Gonzalez, A., Delhomme, N., Ghavi-Helm, Y., Wilczynski, B., Riddell, A. and Furlong, E.E. (2012) Tissue-specific analysis of chromatin state identifies temporal signatures of enhancer activity during embryonic development. *Nat. Genet.*, **44**, 148-156.
6. Levine, M. (2010) Transcriptional enhancers in animal development and evolution. *Curr. Biol.*, **20**, R754-763.
7. Heintzman, N.D., Hon, G.C., Hawkins, R.D., Kheradpour, P., Stark, A., Harp, L.F., Ye, Z., Lee, L.K., Stuart, R.K., Ching, C.W. *et al.* (2009) Histone modifications at human enhancers reflect global cell-type-specific gene expression. *Nature*, **459**, 108-112.

bioRxiv preprint doi: https://doi.org/10.1101/036129; this version posted January 7, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

8. Whyte, W.A., Orlando, D.A., Hnisz, D., Abraham, B.J., Lin, C.Y., Kagey, M.H., Rahl, P.B., Lee, T.I. and Young, R.A. (2013) Master transcription factors and mediator establish super-enhancers at key cell identity genes. *Cell*, **153**, 307-319.
9. Hnisz, D., Abraham, B.J., Lee, T.I., Lau, A., Saint-Andre, V., Sigova, A.A., Hoke, H.A. and Young, R.A. (2013) Super-enhancers in the control of cell identity and disease. *Cell*, **155**, 934-947.
10. Kim, T.K., Hemberg, M., Gray, J.M., Costa, A.M., Bear, D.M., Wu, J., Harmin, D.A., Laptewicz, M., Barbara-Haley, K., Kuersten, S. *et al.* (2010) Widespread transcription at neuronal activity-regulated enhancers. *Nature*, **465**, 182-187.
11. Jin, F., Li, Y., Ren, B. and Natarajan, R. (2011) PU.1 and C/EBP(alpha) synergistically program distinct response to NF-kappaB activation through establishing monocyte specific enhancers. *Proc. Natl. Acad. Sci. U. S. A.*, **108**, 5290-5295.
12. Heintzman, N.D., Stuart, R.K., Hon, G., Fu, Y., Ching, C.W., Hawkins, R.D., Barrera, L.O., Van Calcar, S., Qu, C., Ching, K.A. *et al.* (2007) Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nat. Genet.*, **39**, 311-318.
13. Sexton, T. and Cavalli, G. (2015) The role of chromosome domains in shaping the functional genome. *Cell*, **160**, 1049-1059.
14. He, B., Chen, C., Teng, L. and Tan, K. (2014) Global view of enhancer-promoter interactome in human cells. *Proc. Natl. Acad. Sci. U. S. A.*, **111**, E2191-2199.
15. Jin, F., Li, Y., Dixon, J.R., Selvaraj, S., Ye, Z., Lee, A.Y., Yen, C.A., Schmitt, A.D., Espinoza, C.A. and Ren, B. (2013) A high-resolution map of the three-dimensional chromatin interactome in human cells. *Nature*, **503**, 290-294.
16. Dekker, J., Marti-Renom, M.A. and Mirny, L.A. (2013) Exploring the three-dimensional organization of genomes: interpreting chromatin interaction data. *Nat. Rev. Genet.*, **14**, 390-403.
17. Mansour, M.R., Abraham, B.J., Anders, L., Berezovskaya, A., Gutierrez, A., Durbin, A.D., Etchin, J., Lawton, L., Sallan, S.E., Silverman, L.B. *et al.* (2014) Oncogene regulation. An oncogenic super-enhancer formed through somatic mutation of a noncoding intergenic element. *Science (New York, N.Y.)*, **346**, 1373-1377.
18. Loven, J., Hoke, H.A., Lin, C.Y., Lau, A., Orlando, D.A., Vakoc, C.R., Bradner, J.E., Lee, T.I. and Young, R.A. (2013) Selective inhibition of tumor oncogenes by disruption of super-enhancers. *Cell*, **153**, 320-334.
19. Prabhakar, S., Poulin, F., Shoukry, M., Afzal, V., Rubin, E.M., Couronne, O. and Pennacchio, L.A. (2006) Close sequence comparisons are sufficient to identify human cis-regulatory elements. *Genome Res.*, **16**, 855-863.
20. Pennacchio, L.A., Ahituv, N., Moses, A.M., Prabhakar, S., Nobrega, M.A., Shoukry, M., Minovitsky, S., Dubchak, I., Holt, A., Lewis, K.D. *et al.* (2006) In vivo enhancer analysis of human conserved non-coding sequences. *Nature*,

bioRxiv preprint doi: https://doi.org/10.1101/036129; this version posted January 7, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

444, 499-502.

21. Visel, A., Prabhakar, S., Akiyama, J.A., Shoukry, M., Lewis, K.D., Holt, A., Plajzer-Frick, I., Afzal, V., Rubin, E.M. and Pennacchio, L.A. (2008) Ultraconservation identifies a small subset of extremely constrained developmental enhancers. *Nat. Genet.*, **40**, 158-160.
22. Visel, A., Blow, M.J., Li, Z., Zhang, T., Akiyama, J.A., Holt, A., Plajzer-Frick, I., Shoukry, M., Wright, C., Chen, F. *et al.* (2009) ChIP-seq accurately predicts tissue-specific activity of enhancers. *Nature*, **457**, 854-858.
23. Nobrega, M.A., Ovcharenko, I., Afzal, V. and Rubin, E.M. (2003) Scanning human gene deserts for long-range enhancers. *Science (New York, N.Y.)*, **302**, 413.
24. Visel, A., Rubin, E.M. and Pennacchio, L.A. (2009) Genomic views of distant-acting enhancers. *Nature*, **461**, 199-205.
25. Loots, G.G., Locksley, R.M., Blankespoor, C.M., Wang, Z.E., Miller, W., Rubin, E.M. and Frazer, K.A. (2000) Identification of a coordinate regulator of interleukins 4, 13, and 5 by cross-species sequence comparisons. *Science (New York, N.Y.)*, **288**, 136-140.
26. Cheng, C., Alexander, R., Min, R., Leng, J., Yip, K.Y., Rozowsky, J., Yan, K.K., Dong, X., Djebali, S., Ruan, Y. *et al.* (2012) Understanding transcriptional regulation by integrative analysis of transcription factor binding data. *Genome Res.*, **22**, 1658-1667.
27. He, A., Kong, S.W., Ma, Q. and Pu, W.T. (2011) Co-occupancy by multiple cardiac transcription factors identifies transcriptional enhancers active in heart. *Proc. Natl. Acad. Sci. U. S. A.*, **108**, 5632-5637.
28. Chen, X., Xu, H., Yuan, P., Fang, F., Huss, M., Vega, V.B., Wong, E., Orlov, Y.L., Zhang, W., Jiang, J. *et al.* (2008) Integration of external signaling pathways with the core transcriptional network in embryonic stem cells. *Cell*, **133**, 1106-1117.
29. Heintzman, N.D. and Ren, B. (2009) Finding distal regulatory elements in the human genome. *Curr. Opin. Genet. Dev.*, **19**, 541-549.
30. May, D., Blow, M.J., Kaplan, T., McCulley, D.J., Jensen, B.C., Akiyama, J.A., Holt, A., Plajzer-Frick, I., Shoukry, M., Wright, C. *et al.* (2012) Large-scale discovery of enhancers from human heart tissue. *Nat. Genet.*, **44**, 89-93.
31. Ghisletti, S., Barozzi, I., Mietton, F., Polletti, S., De Santa, F., Venturini, E., Gregory, L., Lonie, L., Chew, A., Wei, C.L. *et al.* (2010) Identification and characterization of enhancers controlling the inflammatory gene expression program in macrophages. *Immunity*, **32**, 317-328.
32. Blow, M.J., McCulley, D.J., Li, Z., Zhang, T., Akiyama, J.A., Holt, A., Plajzer-Frick, I., Shoukry, M., Wright, C., Chen, F. *et al.* (2010) ChIP-Seq identification of weakly conserved heart enhancers. *Nat. Genet.*, **42**, 806-810.
33. Boyle, A.P., Song, L., Lee, B.K., London, D., Keefe, D., Birney, E., Iyer, V.R., Crawford, G.E. and Furey, T.S. (2011) High-resolution genome-wide in vivo footprinting of diverse transcription factors in human cells. *Genome Res.*, **21**, 456-464.

bioRxiv preprint doi: https://doi.org/10.1101/036129; this version posted January 7, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

34. Crawford, G.E., Holt, I.E., Whittle, J., Webb, B.D., Tai, D., Davis, S., Margulies, E.H., Chen, Y., Bernat, J.A., Ginsburg, D. *et al.* (2006) Genome-wide mapping of DNase hypersensitive sites using massively parallel signature sequencing (MPSS). *Genome Res.*, **16**, 123-131.
35. Boyle, A.P., Davis, S., Shulha, H.P., Meltzer, P., Margulies, E.H., Weng, Z., Furey, T.S. and Crawford, G.E. (2008) High-resolution mapping and characterization of open chromatin across the genome. *Cell*, **132**, 311-322.
36. Barski, A., Cuddapah, S., Cui, K., Roh, T.Y., Schones, D.E., Wang, Z., Wei, G., Chepelev, I. and Zhao, K. (2007) High-resolution profiling of histone methylations in the human genome. *Cell*, **129**, 823-837.
37. Wang, Z., Zang, C., Rosenfeld, J.A., Schones, D.E., Barski, A., Cuddapah, S., Cui, K., Roh, T.Y., Peng, W., Zhang, M.Q. *et al.* (2008) Combinatorial patterns of histone acetylations and methylations in the human genome. *Nat. Genet.*, **40**, 897-903.
38. Won, K.J., Chepelev, I., Ren, B. and Wang, W. (2008) Prediction of regulatory elements in mammalian genomes using chromatin signatures. *BMC bioinformatics*, **9**, 547.
39. Firpi, H.A., Ucar, D. and Tan, K. (2010) Discover regulatory DNA elements using chromatin signatures and artificial neural network. *Bioinformatics (Oxford, England)*, **26**, 1579-1586.
40. Rajagopal, N., Xie, W., Li, Y., Wagner, U., Wang, W., Stamatoyannopoulos, J., Ernst, J., Kellis, M. and Ren, B. (2013) RFECS: a random-forest based algorithm for enhancer identification from chromatin state. *PLoS Comput. Biol.*, **9**, e1002968.
41. Lu, Y., Qu, W., Shan, G. and Zhang, C. (2015) DELTA: A Distal Enhancer Locating Tool Based on AdaBoost Algorithm and Shape Features of Chromatin Modifications. *PloS one*, **10**, e0130622.
42. Fernandez, M. and Miranda-Saavedra, D. (2012) Genome-wide enhancer prediction from epigenetic signatures using genetic algorithm-optimized support vector machines. *Nucleic Acids Res.*, **40**, e77.
43. Erwin, G.D., Oksenberg, N., Truty, R.M., Kostka, D., Murphy, K.K., Ahituv, N., Pollard, K.S. and Capra, J.A. (2014) Integrating diverse datasets improves developmental enhancer prediction. *PLoS Comput. Biol.*, **10**, e1003677.
44. Kleftogiannis, D., Kalnis, P. and Bajic, V.B. (2015) DEEP: a general computational framework for predicting enhancers. *Nucleic Acids Res.*, **43**, e6.
45. Ernst, J. and Kellis, M. (2012) ChromHMM: automating chromatin-state discovery and characterization. *Nat. Methods*, **9**, 215-216.
46. Hoffman, M.M., Buske, O.J., Wang, J., Weng, Z., Bilmes, J.A. and Noble, W.S. (2012) Unsupervised pattern discovery in human chromatin structure through genomic segmentation. *Nat. Methods*, **9**, 473-476.
47. ENCODE Project Consortium. (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature*, **489**, 57-74.
48. Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., Heravi-Moussavi, A., Kheradpour, P., Zhang, Z., Wang, J., Ziller, M.J. *et al.* (2015) Integrative

bioRxiv preprint doi: https://doi.org/10.1101/036129; this version posted January 7, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

analysis of 111 reference human epigenomes. *Nature*, **518**, 317-330.

49. Siepel, A., Bejerano, G., Pedersen, J.S., Hinrichs, A.S., Hou, M., Rosenbloom, K., Clawson, H., Spieth, J., Hillier, L.W., Richards, S. *et al.* (2005) Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.*, **15**, 1034-1050.
50. Goldman, M., Craft, B., Swatloski, T., Cline, M., Morozova, O., Diekhans, M., Haussler, D. and Zhu, J. (2015) The UCSC cancer genomics browser: update 2015. *Nucleic Acids Res.*, **43**, D812-817.
51. Grant, C.E., Bailey, T.L. and Noble, W.S. (2011) FIMO: scanning for occurrences of a given motif. *Bioinformatics (Oxford, England)*, **27**, 1017-1018.
52. Matys, V., Kel-Margoulis, O.V., Fricke, E., Liebich, I., Land, S., Barre-Dirrie, A., Reuter, I., Chekmenev, D., Krull, M., Hornischer, K. *et al.* (2006) TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. *Nucleic Acids Res.*, **34**, D108-110.
53. Mathelier, A., Fornes, O., Arenillas, D.J., Chen, C.Y., Denay, G., Lee, J., Shi, W., Shyr, C., Tan, G., Worsley-Hunt, R. *et al.* (2015) JASPAR 2016: a major expansion and update of the open-access database of transcription factor binding profiles. *Nucleic Acids Res.*, **44**, D110-D115.
54. Hume, M.A., Barrera, L.A., Gisselbrecht, S.S. and Bulyk, M.L. (2015) UniPROBE, update 2015: new tools and content for the online database of protein-binding microarray data on protein-DNA interactions. *Nucleic Acids Res.*, **43**, D117-122.
55. Hawkins, R.D., Hon, G.C., Lee, L.K., Ngo, Q., Lister, R., Pelizzola, M., Edsall, L.E., Kuan, S., Luu, Y., Klugman, S. *et al.* (2010) Distinct epigenomic landscapes of pluripotent and lineage-committed human cells. *Cell stem cell*, **6**, 479-491.
56. Natoli, G. and Andrau, J.C. (2012) Noncoding transcription at enhancers: general principles and functional models. *Annu. Rev. Genet.*, **46**, 1-19.
57. Zhu, Y., Sun, L., Chen, Z., Whitaker, J.W., Wang, T. and Wang, W. (2013) Predicting enhancer transcription and activity from chromatin modifications. *Nucleic Acids Res.*, **41**, 10032-10043.
58. Creyghton, M.P., Cheng, A.W., Welstead, G.G., Kooistra, T., Carey, B.W., Steine, E.J., Hanna, J., Lodato, M.A., Frampton, G.M., Sharp, P.A. *et al.* (2010) Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proc. Natl. Acad. Sci. U. S. A.*, **107**, 21931-21936.
59. Zhang, Y., Liu, T., Meyer, C.A., Eeckhoute, J., Johnson, D.S., Bernstein, B.E., Nusbaum, C., Myers, R.M., Brown, M., Li, W. *et al.* (2008) Model-based analysis of ChIP-Seq (MACS). *Genome Biol.*, **9**, R137.
60. Harrow, J., Frankish, A., Gonzalez, J.M., Tapanari, E., Diekhans, M., Kokocinski, F., Aken, B.L., Barrell, D., Zadissa, A., Searle, S. *et al.* (2012) GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res.*, **22**, 1760-1774.
61. Liu, F., Ren, C., Li, H., Zhou, P., Bo, X. and Shu, W. (2015) De novo

bioRxiv preprint doi: https://doi.org/10.1101/036129; this version posted January 7, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

identification of replication-timing domains in the human genome by deep learning. *Bioinformatics (Oxford, England)*, 1-9.

62. Batuwita, R. and Palade, V. (2013) In He, H. and Ma, Y. (eds.), *Imbalanced Learning: Foundations, Algorithms, and Applications*. John Wiley & Sons, Inc., Hoboken, NJ, pp. 83-99.
63. Bajic, V.B. (2000) Comparing the success of different prediction software in sequence analysis: a review. *Brief Bioinform.*, **1**, 214-228.
64. Gorkin, D.U., Leung, D. and Ren, B. (2014) The 3D genome in transcriptional regulation and pluripotency. *Cell stem cell*, **14**, 762-775.
65. Smallwood, A. and Ren, B. (2013) Genome organization and long-range regulation of gene expression by enhancers. *Curr. Opin. Cell Biol.*, **25**, 387-394.
66. Lu, Y., Zhou, Y. and Tian, W. (2013) Combining Hi-C data with phylogenetic correlation to predict the target genes of distal regulatory elements in human genome. *Nucleic Acids Res.*, **41**, 10391-10402.
67. Libbrecht, M.W., Ay, F., Hoffman, M.M., Gilbert, D.M., Bilmes, J.A. and Noble, W.S. (2015) Joint annotation of chromatin state and chromatin conformation reveals relationships among domain types and identifies domains of cell-type-specific expression. *Genome Res.*, **25**, 544-557.
68. Lake, B.M., Salakhutdinov, R. and Tenenbaum, J.B. (2015) Human-level concept learning through probabilistic program induction. *Science (New York, N.Y.)*, **350**, 1332-1338.
69. LeCun, Y., Bengio, Y. and Hinton, G. (2015) Deep learning. *Nature*, **521**, 436-444.

bioRxiv preprint doi: https://doi.org/10.1101/036129; this version posted January 7, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Figure legends

Figure 1. Enhancer predictions using PEDLA with heterogeneous signatures and class-imbalanced data in H1 cells.

(A) Selecting the optimal structure of PEDLA for the purpose of enhancer prediction using 1,114 heterogeneous signatures. Three performance indicators, accuracy, GM and F1-score, were measured using 5-fold cross-validation in both a training set and test set. The error bar indicates the mean and standard deviation of the performance indicator.

(B) Validation of enhancer predictions by distal DHSs, binding sites of p300 and TFs (NANOG, OCT4 and SOX2) using the trained PEDLA with optimal structure. The bar shows the actual validation rate, whereas the error bar shows the mean and standard deviation of validation rates for 10,000 randomly shuffled predictions.

(C-D) Capability of handling class-imbalanced data unbiasedly. Three performance indicators, sensitivity, specificity and GM, were measured for the training set (C) and test set (D) using 5-fold cross-validation based on the optimal structure of PEDLA with all 1,114-dimensional features. The number of enhancers, promoters and random regions not annotated as promoters or enhancers were maintained at 1 : 1 : x (x = 1, 2, …, 9), such that the ratio between positive and negative samples was 1 : (1 + x).

Figure 2. Training procedure of PEDLA in multiple human cell types.

(A) Schematic diagram showing the framework of training PEDLA for the purpose of

bioRxiv preprint doi: https://doi.org/10.1101/036129; this version posted January 7, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

enhancer prediction in multiple human cells/tissues. The whole training procedure comprises initial training and iterative training.

(B) The pseudocode shows the detailed steps of training PEDLA in multiple human cells/tissues.

Figure 3. Evaluation of enhancer predictions using PEDLA in multiple human cells/tissues.

(A) Evaluation of enhancer prediction using PEDLA in 22 training cell types/tissues. The whole training procedure was repeated with 50 random orders of the 22 training cell types/tissues, and each random order was repeated four times with random permutations of training samples for each cell type/tissue. In total, the training of PEDLA was repeated 200 times on the 22 training cell types/tissues, independently. For each repeat, the trained optimally model of PEDLA was saved for later evaluation for each of the 22 training cell types/tissues. Thus, $22 \times 200 = 4{,}400$ optimal models were generated. $M_j^i (1 \leq j \leq 22, 1 \leq i \leq 200)$ denotes the optimal model that finished training on the *j*-th training cell type/tissue in the *i*-th run of the 200 independent runs.

(B-C) Performance evaluations of PEDLA in the training cell set and the independent test cell set along the training route. Three performance indicators, accuracy, GM, and F1-score, were assessed for the PEDLA with the trained optimal model in the 22 training cell types/tissues and 20 test cell types/tissues, independently. (B) For a fixed j of the X-axis, all 200 optimal models $M_j^i (1 \leq i \leq 200)$ were used to assess the performance indicators on the 22 training cell types/tissues. The red line represents

bioRxiv preprint doi: https://doi.org/10.1101/036129; this version posted January 7, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

the mean of the total $200 \times 22 = 4{,}400$ values of each performance indicator, and the light blue colour band indicates the 10th and 90th percentiles. (C) For a fixed j of the X-axis, all 200 optimal models $M_j^i (1 \leq i \leq 200)$ were used to assess the performance indicators on the 20 test cell types/tissues. The red line represents the mean of the total $200 \times 20 = 4{,}400$ values of each performance indicator, and the light blue colour band indicates the 10th and 90th percentiles.

Figure 4. Performance assessment of PEDLA with the best-trained model in multiple human cells/tissues.

(A-B) Performance assessment of PEDLA with the best-trained model in the training cell set and test cell set. Three performance indicators, accuracy, GM, and F1-score, were assessed for the PEDLA with the best trained model in the 22 training cell types/tissues (A) and 20 test cell types/tissues (B), independently. The best-trained model was the only one selected in terms of performance from the 200 optimal models $M_{22}^i (1 \leq i \leq 200)$ that finished training on 22 training cell types/tissues. All enhancers were classified as "specific" or "common" based on the number of cell types/tissues in which the enhancers occurred. An enhancer that occurred in not more than 4 cell types/tissues was termed specific; otherwise, it was considered common.

bioRxiv preprint doi: https://doi.org/10.1101/036129; this version posted January 7, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Tables

Table 1. Comparison of the performance of PEDLA with that of existing methods

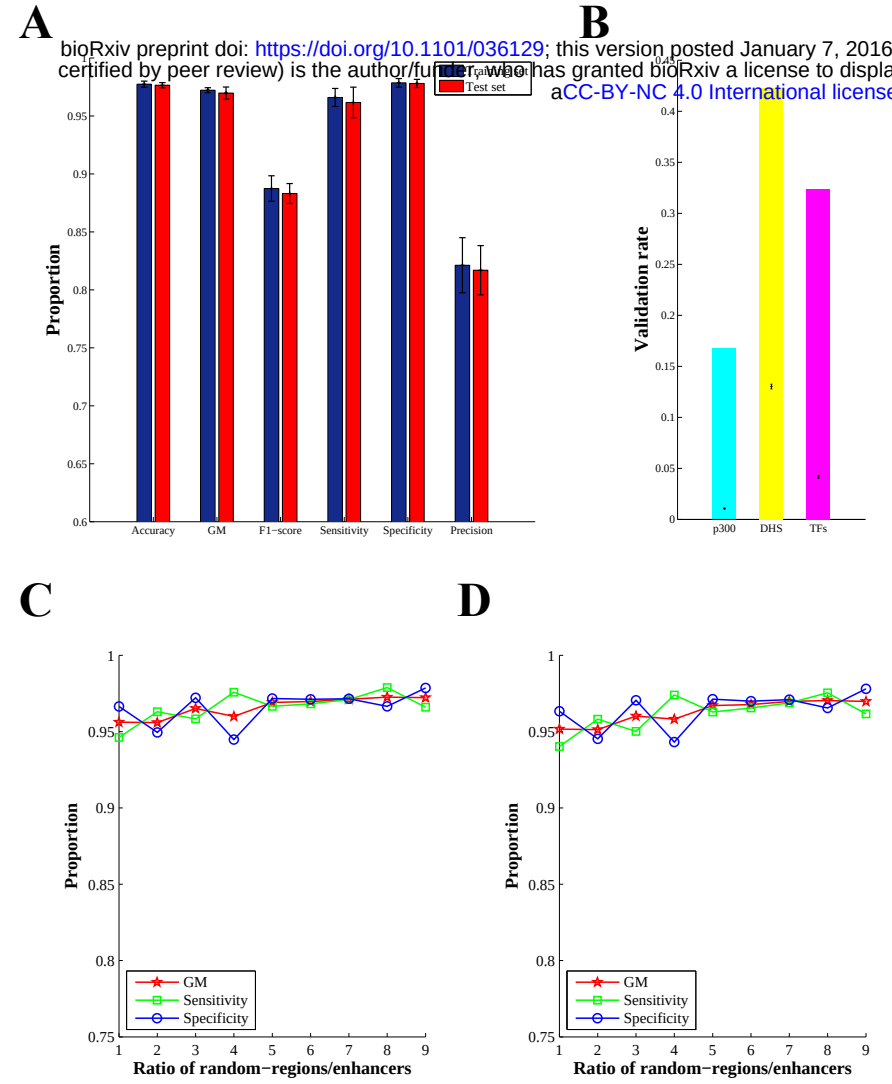
			PEDLA	RFECS	CSI-ANN	DELTA	ChromHMM	Segway	PEDLA (all features)
Number of prediction			22691	75084	30173	112044	26869	131698	20689
Performance metrics	**Accuracy**		96.30%	93.67%	95.58%	87.78%	94.03%	91.01%	97.65%
	GM		96.02%	79.26%	80.34%	81.29%	61.30%	35.71%	96.97%
	F1-score		83.01%	71.71%	73.06%	60.40%	53.74%	20.90%	88.31%
	Validation rate	**DHS**	40.68%	31.85%	30.65%	12.25%	38.86%	40.61%	42.29%
		P300	15.25%	7.26%	10.83%	1.57%	9.89%	3.52%	16.82%
		TFs	28.89%	17.71%	19.72%	5.75%	19.14%	6.42%	32.37%
	Misclassification rate		7.53%	3.09%	16.46%	3.01%	6.42%	14.53%	6.59%

bioRxiv preprint doi: https://doi.org/10.1101/036129; this version posted January 7, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Table 2. Relative ranking of PEDLA and existing methods based on the results of our comparison study

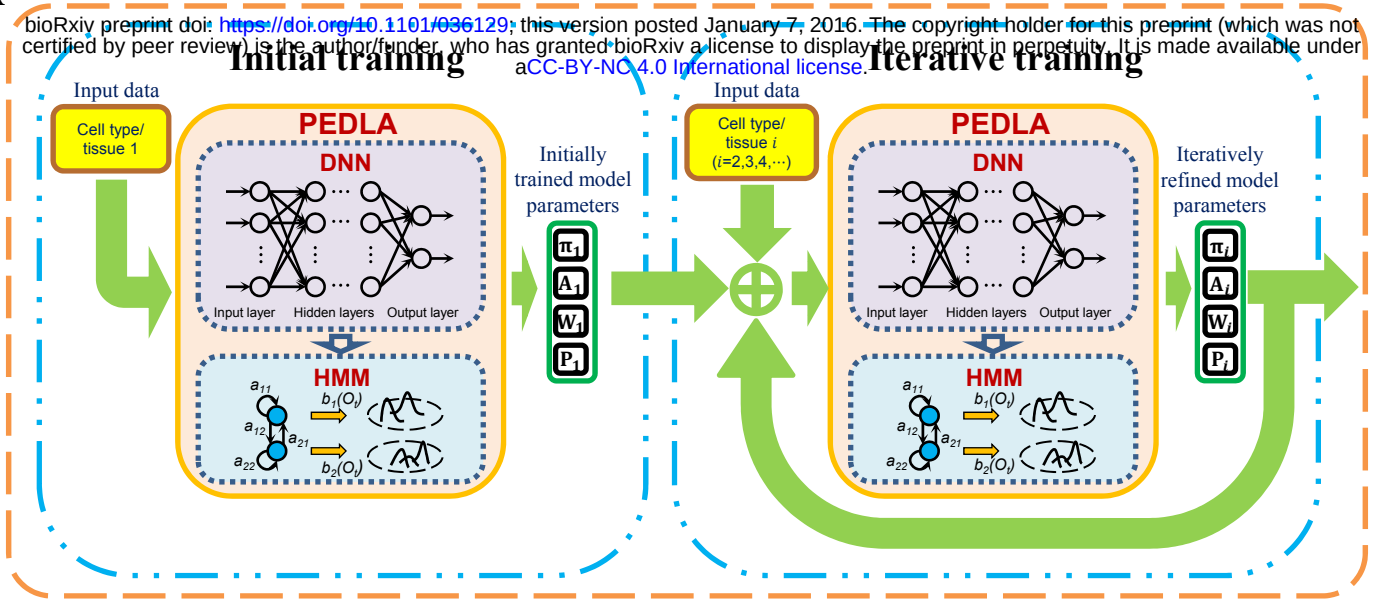
Performance metrics		PEDLA	RFECS	CSI-ANN	DELTA	ChromHMM	Segway	PEDLA (all features)
Accuracy		2	5	3	7	4	6	1
GM		2	5	4	3	6	7	1
F1-score		2	4	3	5	6	7	1
Validation rate	DHS	2	5	6	7	4	3	1
	P300	2	5	3	7	4	6	1
	TFs	2	5	3	7	4	6	1
Misclassification rate		5	2	7	1	3	6	4
OVERALL RANKING		**2nd (17), 2.43**	**4th (31), 4.43**	**3rd (29), 4.14**	**5th (37) 5.29**	**4th (31), 4.43**	**6th (41), 5.86**	**1st (10), 1.43**

bioRxiv preprint doi: https://doi.org/10.1101/036129; this version posted January 7, 2016
certified by peer review) is the author/funder, who has granted bioRxiv a license to displa
aCC-BY-NC 4.0 International license



bioRxiv preprint doi: https://doi.org/10.1101/036129; this version posted January 7, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

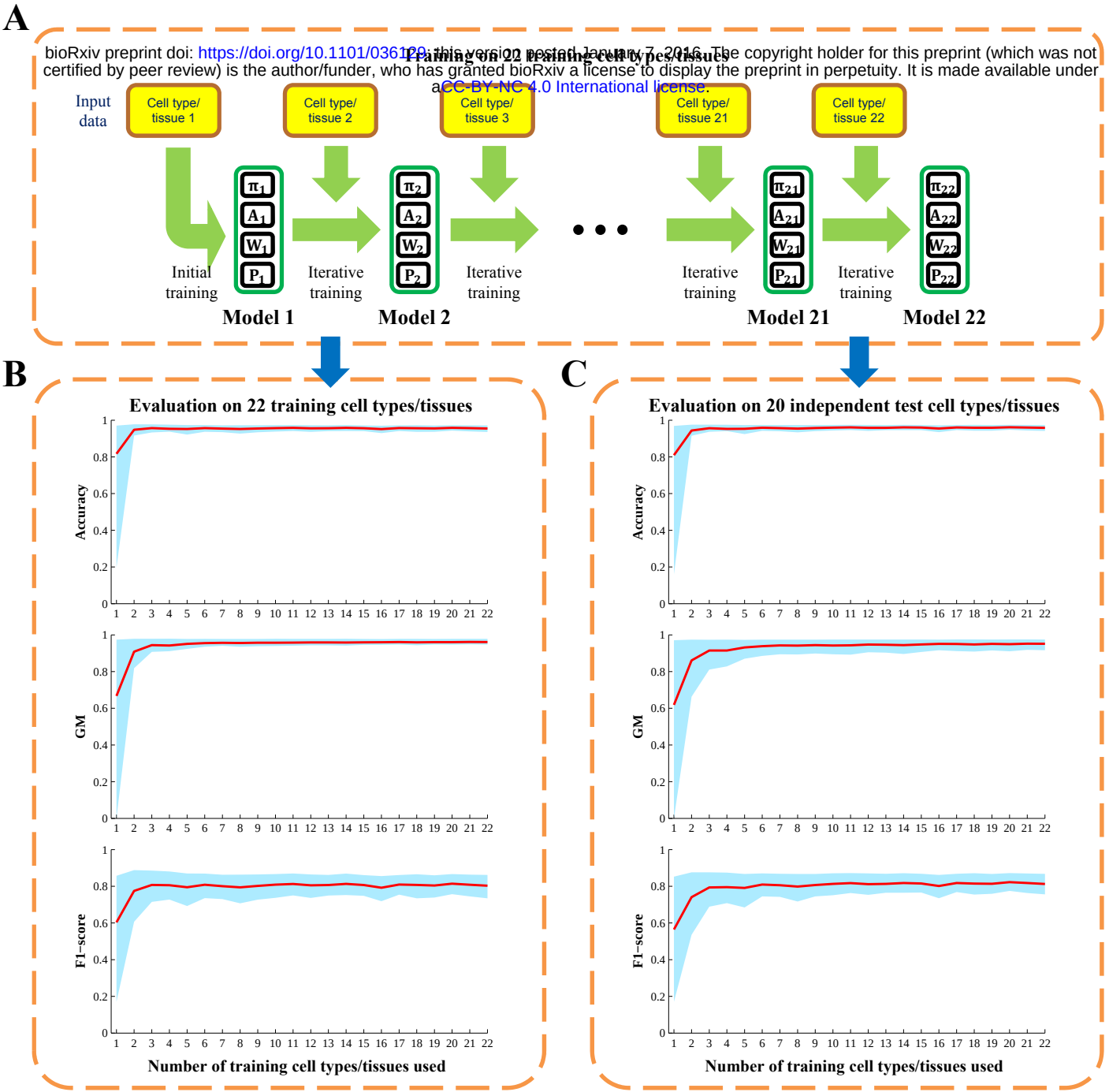
A



B

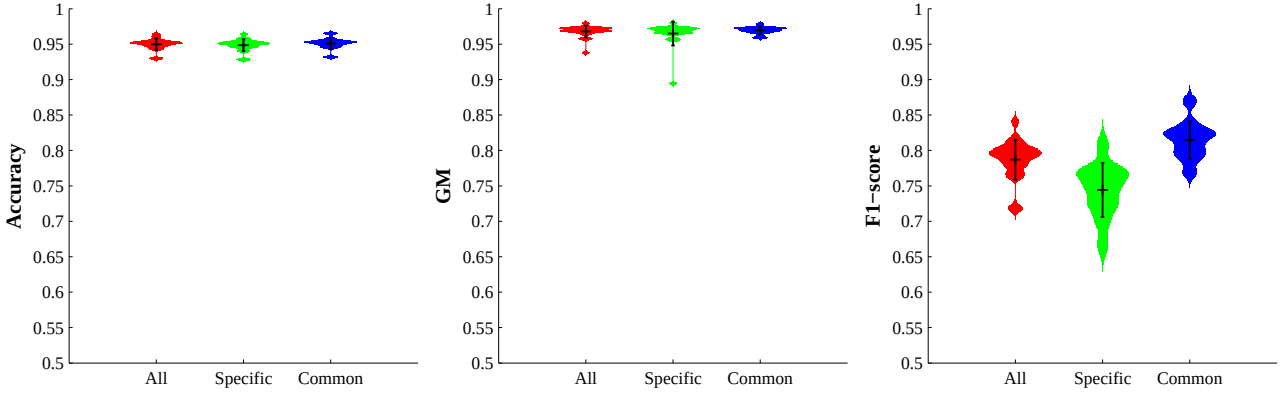
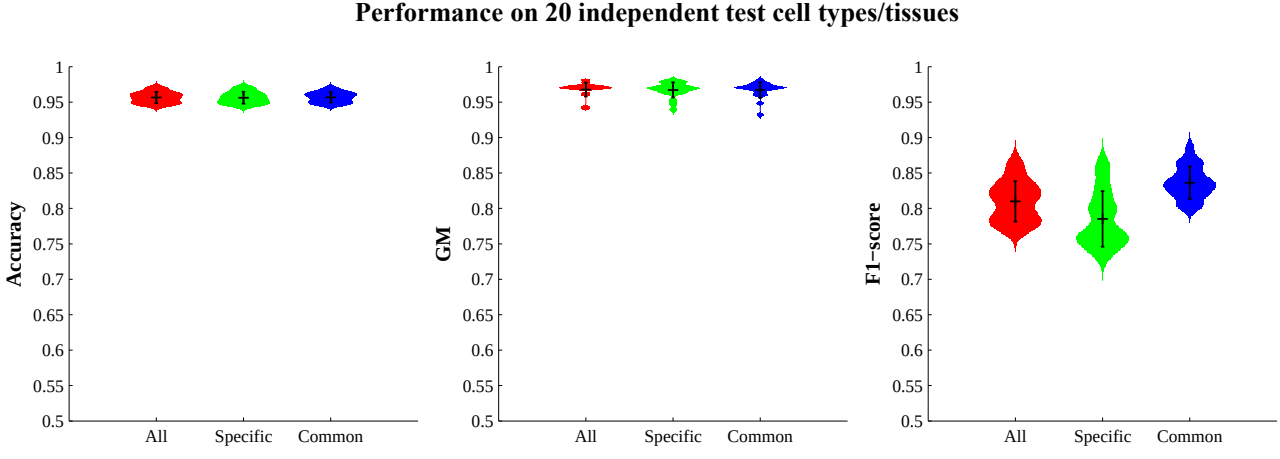
1. Initial training
   1) Read the normalized features and annotated labels of cell type/tissue 1.
   2) Estimate the initial state probability vector $\boldsymbol{\pi_1}$ and state transition probability matrix $\boldsymbol{A_1}$ of HMM.
   3) Pre-train DNN layer by layer unsupervisedly, and then use the pre-training as initialization to fine-tune the weight matrix $\boldsymbol{W_1}$ of DNN using mini-batch gradient descent in a supervised manner.
   4) Estimate the prior probability vector $\boldsymbol{P_1}$.
2. Iterative training
   1) Read the normalized features and annotated labels of cell type/tissue i ($i = 2,3,4,\cdots$).
   2) Fetch the trained model parameters $(\boldsymbol{\pi_{i-1}}, \boldsymbol{A_{i-1}}, \boldsymbol{W_{i-1}}, \boldsymbol{P_{i-1}})$ of a previous cell type/tissue as initialization.
   3) Update the initial state probability vector $\boldsymbol{\pi_i}$ and state transition probability matrix $\boldsymbol{A_i}$ of HMM.
   4) Use $\boldsymbol{W_{i-1}}$ as initialization to refine the weight matrix $\boldsymbol{W_i}$ of DNN using mini-batch gradient descent in a supervised manner.
   5) Re-estimate the prior probability vector $\boldsymbol{P_i}$.
   6) Go to step 1 of the iterative training procedure and set $i = i + 1$ if there is input data of another new cell type/tissue; Otherwise output the final optimal model parameters that have been trained on all cell types/tissues and exit.

bioRxiv preprint doi: https://doi.org/10.1101/036169; this version posted January 7, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-4.0 International license.



bioRxiv preprint doi: https://doi.org/10.1101/036129; this version posted January 7, 2016. The copyright holder for this preprint (which was n
certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available und
aCC-BY-NC 4.0 International license.

A



B