SeqEnhDL: sequence-based classification of cell type-specific enhancers using deep learning models

Yupeng Wang (✉ ywang@bdxconsult.com)
BDX Research & Consulting LLC https://orcid.org/0000-0002-3002-8069

Rosario Jaime-Lara
National Institutes of Health

Abhrarup Roy
National Institutes of Health

Ying Sun
BDX Research & Consulting LLC

Xinyue Liu
BDX Research & Consulting LLC

Paule Joseph
National Institutes of Health

Research note

Keywords: DNA, RNN, CNN, DanQ, SeqEnhDL

Posted Date: October 21st, 2020

DOI: https://doi.org/10.21203/rs.3.rs-94396/v1

Version of Record: A version of this preprint was published on March 19th, 2021. See the published version at https://doi.org/10.1186/s13104-021-05518-7.

Abstract

Objective

Computational identification of cell type-specific regulatory elements on a genome-wide scale is very challenging.

Results

We propose SeqEnhDL, a deep learning framework for classifying cell type-specific enhancers based on sequence features. DNA sequences of “strong enhancer” chromatin states in nine cell types from the ENCODE project were retrieved to build and test enhancer classifiers. For any DNA sequence, sequential *k*-mer (*k*=5, 7, 9 and 11) fold changes relative to randomly selected non-coding sequences were used as features for deep learning models. Three deep learning models were implemented, including multi-layer perceptron (MLP), Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN). All models in SeqEnhDL outperform state-of-the-art enhancer classifiers including gkm-SVM and DanQ, with regard to distinguishing cell type-specific enhancers from randomly selected non-coding sequences. Moreover, SeqEnhDL is able to directly discriminate enhancers from different cell types, which has not been achieved by other enhancer classifiers. Our analysis suggests that both enhancers and their tissue-specificity can be accurately identified according to their sequence features. SeqEnhDL is publicly available at https://github.com/wyp1125/SeqEnhDL.

Introduction

Cell type-specific enhancers, *cis*-regulatory elements that up-regulate gene transcription in a cell type, play a key role in determining the regulatory landscape of the human genome (1). Enhancers are commonly located in the introns and immediately upstream of the transcription start site (TSS) of their target genes, but they are also known to populate gene deserts (2), reside in introns of neighboring genes (3) and co-localize with coding exons (4). Enhancer mutations are often associated with diseases (5–7). Accurate prediction of enhancers from DNA sequences is the basis of assessing whether mutation(s) could disrupt an enhancer’s activity, a type of mechanism for genetic diseases.

Predicting enhancers based on transcription factor binding sites (TFBS) was proposed because TFBS tend to be conserved over vertebrate evolution (8–10). To ameliorate the uncertainty problem in conservation and TFBS information, direct sequence features such as *k*-mers were used to model enhancer prediction (11, 12). These early studies did not achieve high prediction accuracy nor were they able to distinguish enhancers of different cell types.

With wide application of ChIP-seq technologies, enhancers were frequently profiled on a genome-wide scale (13). The ENCODE project produced genome-wide profiles of various epigenetic marks for multiple human cell types (14). By applying a hidden Markov model (i.e. ChromHMM) to these epigenetic marks, the sequence of the human genome has been binned into more than ten chromatin states, including enhancers (15, 16). The “strong enhancer” state, shown to be associated with increased gene expression, provides genome-wide positioning of active enhancers for a cell type (15). Although the availability of these datasets renders positioning of enhancers unnecessary, the sequence structures of enhancers, especially their subtle differences among cell types, can be useful in understanding cell type-specific gene regulation and should be explored.

The effectiveness of enhancer classifiers is influenced by proper generation of negative sequences. Negative sequences should contain similar basic sequence features with enhancers such as length distributions, GC and repeat contents (12, 17, 18); otherwise, enhancer classifiers may learn different nucleotide compositions rather than occurrences of key DNA motifs. Although there are many published studies regarding sequence-based enhancer prediction, it is still unknown whether these enhancer classifiers can distinguish enhancers from different cell types or tissues.

The sequence structures of enhancers may not be linear or additive. In fact, there could be complex grammar or semantics among different DNA elements that compose an enhancer (19, 20). In recent years, deep learning technologies have gained greater popularity than conventional machine learning methods, and have been adapted in biomedical research to address complex research questions (21–29). Thus, deep learing can be more powerful in classifying enhancers. In this study, we propose SeqEnhDL, a deep learning framework for classification of cell type-specific enhancers based on sequence features. The effectiveness and advantages of SeqEnhDL are demonstrated based on the chromatin state segmentation data of nine cell types from the ENCODE project (14).

Methods

Genome annotations

The sequences and transcripts of the human genome (hg19) were obtained from UCSC genome browser. The “knowngene” dataset was used to guide masking exons. Chromatin state annotations of gm12878, H1hesc, hepg2, Hmec, Hsmm, Huvec, K562, Nhek and Nhlf cell types generated by ChromHMM (Broad version) were obtained from the ENCODE project. The data had a total of 15 chromatin states. 4_Strong_Enhancer and 5_Strong_Enhancer states were used as enhancers in this study.

Feature extraction

We masked exons and repetitive sequences of the human genome prior to retrieving DNA sequences for building enhancer classifiers. We divided all initial enhancers into 200 bp enhancer units and treated each enhancer unit as an enhancer. The number of enhancers in each cell type is shown in Table S1. Each enhancer was a positive sequence. Both control and negative sequences were generated according to the GC contents of positive sequences. Control sequences, with a size of 3 folds of the positive sequences, were used to compute the background distributions of *k*-mers. Negative sequences, with the same size as the positive sequence set, were used as the negative set for training and testing enhancer classifiers. Fold changes of *k*-mer (k = 5,7,9 and 11) frequencies (a pseudo count of 1 was added to both the denominator and numerator) were computed between the positive set and the control set and were used as feature dictionaries. Then, each 200 bp sequence in the positive and negative sets was coded using the fold change of *k*-mers at each nucleotide position.

Construction of machine learning models

Positive sequences for any model were initially split into training and testing sets according to the 80:20 rule. Negative sequences were split according to the divisions of their corresponding positive sequences. For deep learning models, the initial training set was further divided into 70% training and 30% validation sets. Deep learning models were built based on the training and validation sets, and performance was assessed based on the testing set. Accuracies were defined as the proportion of correct classifications of the testing sequences. AUCs were computed based on the prediction scores on the testing sequences. Five-fold cross-validation was employed to generate reliable estimates of accuracies and AUCs (average from five runs).

Deep learning models

Deep learning models were generated using the Tensorflow and Keras software, available from Python3.7. Parameters were chosen heuristically and consistently to reach fair comparisons among different machine learning models and different cell types. Parameters included batch size: 512; learning rate: 0.001; epochs: 20; optimizer: Adam; loss: categorical_crossentropy. The best model during the 20 epochs was saved and used for prediction on the testing data. Structures of deep learning models are displayed in Table S2.

Computational resource

All programs of this study were executed on the NIH Biowulf linux cluster. Tensorflow, Keras and required Python libraries were pre-configured under Python 3.7 on Biowulf. Deep learning jobs were executed on GPU nodes.

Results

The SeqEnhDL framework

The SeqEnhDL framework is depicted in Fig. 1. The framework started from “bed” files containing the chromosomal positions of a large number (e.g. >1000) of enhancers. The DNA sequences of enhancers were retrieved from the human genome where exon and repetitive sequences were masked. Then, these DNA sequences were divided into individual enhancers with a fixed length of 200 bp, which makes features more standardized and comparable. Enhancer sequences composed the positive sequences. Control sequences for computing *k*-mer fold changes, and negative sequences for testing enhancer classifiers, were randomly selected from the genome where exon, repetitive and enhancer sequences were masked, according to the GC contents of enhancer sequences. *K*-mer (k = 5, 7, 9 and 11) fold changes between enhancer and control sequences were computed and used for generating the feature at each nucleotide position of any positive or negative sequence. We chose odd *k*-mers because fold changes of different *k*-mers can be aligned at their central nucleotide position. Any dataset for building an deep learning enhancer classifier should be devided into training, validation and testing data, in a cross-valiation mode. Three deep learning models-MLP, CNN and RNN were built. Multilayer perceptrons are fully connected networks. CNN takes advantage of the hierarchical pattern in data and assemble more complex patterns using smaller and simpler patterns. RNN makes use of sequential information among features. Particularly, bidirectional long short-term memory (LSTM) RNN which can learn long-term dependencies was adopted.

Evaluation of the performance of SeqEnhDL

Discriminating enhancers of a single cell type/tissue from randomly selected sequences have been studied before and provided the foundation for evaluating the performance of SeqEnhDL. We retrieved DNA sequences located within the "strong enhancers" chromatin states of nine cell types from the ENCODE project (14). The performance of SeqEnhDL was evaluated in terms of accuracy and area under the curve (AUC) for distinguishing enhancers in each cell type. State-of-the-art methods were selected for comparison with SeqEhnDL. gkm-SVM (12, 17) was chosen for comparison because it uses *k*-mer information to predict enhancers. DanQ (28) was chosen for comparison because it is an RNN-based tool for predicting the functions of noncoding sequences. The performance of DanQ on each cell type was represented by the highest statistics among predictions on 919 ChIP/DNase-seq marks. When different tools were executed, five-fold cross-validation was employed in order to generate reliable performance measures. Comparisons of performances among different tools (Fig. 2) show that all the three models of SeqEnhDL greatly outperform gkm-SVM and DanQ. The accuracies of SeqEnhDL range from 0.961 to 0.999, suggesting that enhancers can be accurately identified on different cell types. Comparison of Receiver Operating Characteristic (ROC) curves on the hepg2 cell type (Figure S1) re-conformed that SeqEnhDL performed better. Of note we also ran a recent approach based on ensemble of deep RNNs (29) for a comparison. However, its accuracies and AUCs were around 0.5 (Table S3), indicating that this compared approach was ineffective on the datasets of this study.

To further demonstrate that deep learning models are better than conventional machine learning models based on the same features, we flattened the *k*-mer features and built enhancer classifiers based on six conventional machine learning models. Note that for each cell type 2000 postive and negative sequences were randomly selected and repeated 10 times in order to ensure training of each deep learning and conventional machine learning model could be finished within one hour. Figure S2 shows that accuracies of SeqEnhCNN and SeqEnhRNN are consistently higher than conventional machine learning models, and SeqEnhMLP is among the second tier in most cell types. These analyses collectively suggest that enhancers present in a single cell type can be accurately identified based on sequence features by SeqEnhDL, and SeqEnhDL greatly outperforms existing methods by better discriminating enhancers from randomly selected sequences.

SeqEnhDL can discriminate enhancers' cell types based on DNA sequences

Successful machine learning models for distinguishing enhancers from different cell types must learn cell type-specific sequence structures such as domains, motifs and their interactions. Previous enhancer classifiers were not examined regarding this capacity. Some may be adapted for distinguishing enhancers from different cell types by treating one cell type as the negative group. We applied gkm-SVM and SeqEnhDL to distinguish enhancers from different cell types. For each pair of cell types, we switched the assignments of positive and negative groups and computed the average accuracy and AUC. The accuracies and AUCs for all pairs of cell types are displayed in Fig. 3. The accuracies and AUCs of gkm-SVM for all pairs of cell types are around 0.5, indicating that gkm-SVM failed to capture tissue-specificities of enhancers. In contrast, all models of SeqEnhDL generated high accuracies (e.g. >0.9) and AUCs (e.g. >0.95) in most cell type combinations, indicating that SeqEnhDL is able to identify tissue-specificity. This analysis suggests that SeqEnhDL can learn distinct sequence features related to tissue-specificities and discriminate enhancers from different cell types.

Discussion

When we built enhancer classifiers, we separated control and negative sequences, which can significantly reduce the chances of overfitting, which is a common problem of machine learning models. We used all enhancers' sequences to compute *k*-mer fold changes. Although theoretically enhancers can be divided into two subsets (one for computing *k*-mer fold changes and the other for testing enhancer classifiers), it is has practical limitations because longer *k*-mers are very important for composing enhancers and may occur only few times in a cell type.

We succesfully applied SeqEnhDL to discriminate enhancers from two cell types. gkm-SVM failed to distinguish enhancers from different cell types, indicating that most (if not all) previous *k*-mer based models tend to learn the common features of enhancers rather than tissue-specific motif structures. This successful application suggests that tissue/cell type-specific gene regulation could be better understood based on machine learning of high-level enhancers' structures.

Conclusion

We propose SeqEnhDL, a feature extraction and deep learning framework for classifying cell type-specific enhancers based on sequence features. A variety of analyses were performed to demonstrate that SeqEnhDL outperforms existing enhancer classifiers. We further proved that SeqEnhDL can be used to discriminate enhancers from different cell types.

Limitation

The training dataset of this study from ChromHMM may be highly noisy. The primary goal of this study is to demonstrate the effectiveness of SeqEnhDL. SeqEnhDL is expected to perform better on cleaner datasets.

Declarations

Availability

Genome sequences and annotations were downloaded from UCSC genome browser (http://genome.ucsc.edu/). Chromatin state segmentation data were downloaded from the ENCODE project (http://hgdownload.soe.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeBroadHmm/). Programs of this study were written in Perl, Python and R. All source code is freely available at https://github.com/wyp1125/SeqEnhDL. Testing datasets of SeqEnhDL for reproduction purpose are available at http://www.bdxconsult.com/SeqEnhDL.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

Not applicable.

Funding

PVJ is supported by the National Institute of Nursing Research [1ZIANR000035-01]. PVJ is also supported by the Office of Workforce Diversity, National Institutes of Health and the Rockefeller University Heilbrunn Nurse Scholar Award[no award number]. RBJL received Intramural Research Training Awards, Office of Intramural Training & Education, National Institutes of Health, Department of Health and Human Services[no award number]. Funding for open access charge: National Institutes of Health.

Authors’ contributions

Y.W. and P.V.J conceived the study. Y.W., Y.S. and X.L. made the programs. Y.W., P.V.J, R.J.L. and A.R. performed the analyses. Y.W. and P.V.J wrote the manuscript.

Acknowledgment

The authors would like to thank Dr. Joan Austin for her comments and editorial assistance.

References

1. Maston GA, Evans SK, Green MR. Transcriptional regulatory elements in the human genome. Annu Rev Genom Hum Genet. 2006;7:29–59.
2. Nobrega MA, Ovcharenko I, Afzal V, Rubin EM. Scanning human gene deserts for long-range enhancers. Science. 2003;302:413.
3. Ott CJ, Suszko M, Blackledge NP, Wright JE, Crawford GE, Harris A. A complex intronic enhancer regulates expression of the CFTR gene by direct interaction with the promoter. J Cell Mol Med.

2009;13:680–92.

4. Birnbaum RY, Clowney EJ, Agamy O, Kim MJ, Zhao J, Yamanaka T, Pappalardo Z, Clarke SL, Wenger AM, Nguyen L, et al. Coding exons function as tissue-specific enhancers of nearby genes. Genome research. 2012;22:1059–68.
5. Weedon MN, Cebola I, Patch AM, Flanagan SE, De Franco E, Caswell R, Rodriguez-Segui SA, Shaw-Smith C, Cho CH, Allen HL, et al. Recessive mutations in a distal PTF1A enhancer cause isolated pancreatic agenesis. Nat Genet. 2014;46:61–4.
6. Emison ES, McCallion AS, Kashuk CS, Bush RT, Grice E, Lin S, Portnoy ME, Cutler DJ, Green ED, Chakravarti A. A common sex-dependent mutation in a RET enhancer underlies Hirschsprung disease risk. Nature. 2005;434:857–63.
7. Pasquali L, Gaulton KJ, Rodriguez-Segui SA, Mularoni L, Miguel-Escalada I, Akerman I, Tena JJ, Moran I, Gomez-Marin C, van de Bunt M, et al. Pancreatic islet enhancer clusters enriched in type 2 diabetes risk-associated variants. Nat Genet. 2014;46:136–43.
8. Taher L, McGaughey DM, Maragh S, Aneas I, Bessling SL, Miller W, Nobrega MA, McCallion AS, Ovcharenko I. Genome-wide identification of conserved regulatory function in diverged sequences. Genome research. 2011;21:1139–49.
9. Narlikar L, Sakabe NJ, Blanski AA, Arimura FE, Westlund JM, Nobrega MA, Ovcharenko I. Genome-wide discovery of human heart enhancers. Genome research. 2010;20:381–92.
10. Burzynski GM, Reed X, Taher L, Stine ZE, Matsui T, Ovcharenko I, McCallion AS. Systematic elucidation and in vivo validation of sequences enriched in hindbrain transcriptional control. Genome research. 2012;22:2278–89.
11. Lee D, Karchin R, Beer MA. Discriminative prediction of mammalian enhancers from DNA sequence. Genome research. 2011;21:2167–80.
12. Ghandi M, Lee D, Mohammad-Noori M, Beer MA. Enhanced regulatory sequence prediction using gapped k-mer features. PLoS Comput Biol. 2014;10:e1003711.
13. Visel A, Blow MJ, Li Z, Zhang T, Akiyama JA, Holt A, Plajzer-Frick I, Shoukry M, Wright C, Chen F, et al. ChIP-seq accurately predicts tissue-specific activity of enhancers. Nature. 2009;457:854–8.
14. Consortium EP. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012;489:57–74.
15. Ernst J, Kheradpour P, Mikkelsen TS, Shoresh N, Ward LD, Epstein CB, Zhang X, Wang L, Issner R, Coyne M, et al. Mapping and analysis of chromatin state dynamics in nine human cell types. Nature. 2011;473:43–9.
16. Hoffman MM, Ernst J, Wilder SP, Kundaje A, Harris RS, Libbrecht M, Giardine B, Ellenbogen PM, Bilmes JA, Birney E, et al. Integrative annotation of chromatin elements from ENCODE data. Nucleic acids research. 2013;41:827–41.
17. Ghandi M, Mohammad-Noori M, Ghareghani N, Lee D, Garraway L, Beer MA. gkmSVM: an R package for gapped-kmer SVM. Bioinformatics. 2016;32:2205–7.

18. Singh AP, Mishra S, Jabin S. Sequence based prediction of enhancer regions from DNA random walk. Scientific reports. 2018;8:15912.
19. Wang X, Lin P, Ho JWK. Discovery of cell-type specific DNA motif grammar in cis-regulatory elements using random Forest. BMC Genomics. 2018;19:929.
20. Weingarten-Gabbay S, Segal E. The grammar of transcriptional regulation. Human genetics. 2014;133:701–11.
21. Angermueller C, Parnamaa T, Parts L, Stegle O. Deep learning for computational biology. Molecular systems biology. 2016;12:878.
22. Quang D, Chen Y, Xie X. DANN: a deep learning approach for annotating the pathogenicity of genetic variants. Bioinformatics. 2015;31:761–3.
23. Zhou J, Troyanskaya OG. Predicting effects of noncoding variants with deep learning-based sequence model. Nature methods. 2015;12:931–4.
24. Alipanahi B, Delong A, Weirauch MT, Frey BJ. Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. Nature biotechnology. 2015;33:831–8.
25. Liu F, Li H, Ren C, Bo X, Shu W. PEDLA: predicting enhancers with a deep learning-based algorithmic framework. Scientific reports. 2016;6:28517.
26. Min X, Zeng W, Chen S, Chen N, Chen T, Jiang R. Predicting enhancers with deep convolutional neural networks. BMC Bioinform. 2017;18:478.
27. Kleftogiannis D, Kalnis P, Bajic VB. DEEP: a general computational framework for predicting enhancers. Nucleic acids research. 2015;43:e6.
28. Quang D, Xie X. DanQ: a hybrid convolutional and recurrent deep neural network for quantifying the function of DNA sequences. Nucleic acids research. 2016;44:e107.
29. Tan KK, Le NQK, Yeh HY, Chua MCH. (2019) Ensemble of Deep Recurrent Neural Networks for Identifying Enhancers via Dinucleotide Physicochemical Properties. *Cells*, 8.

Figures

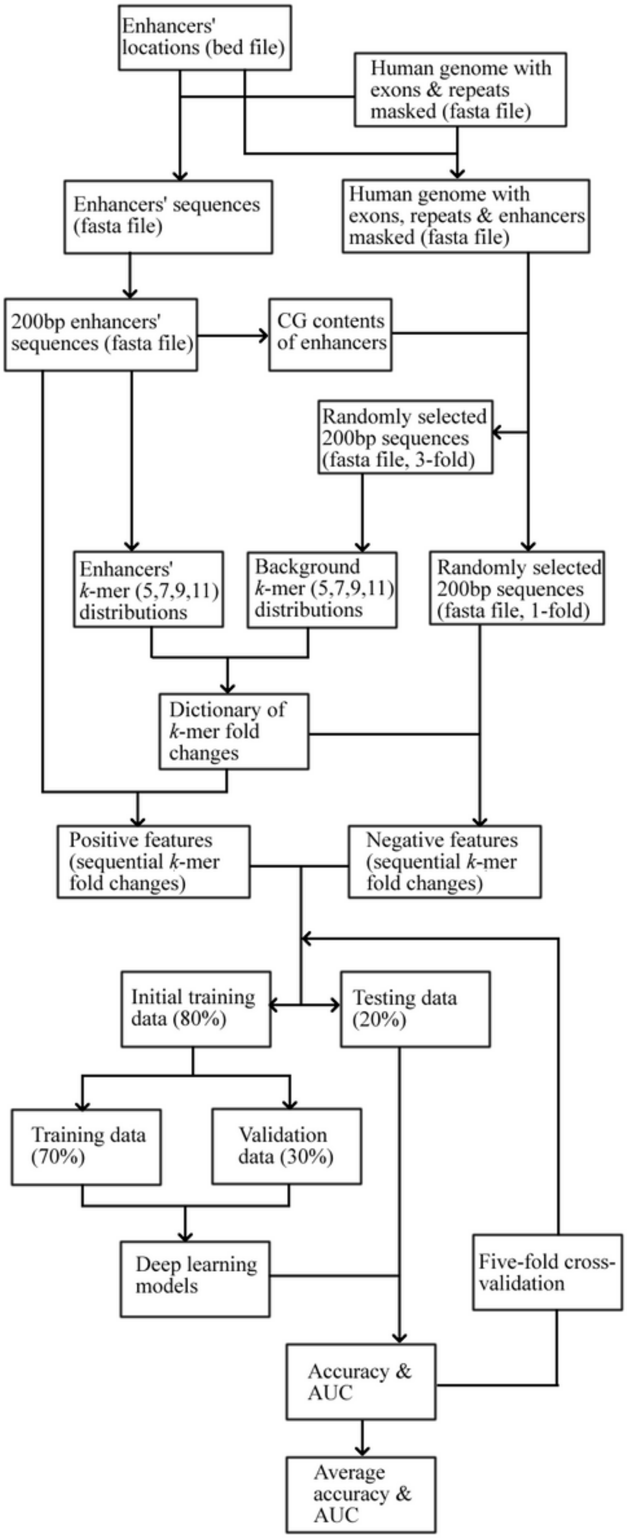
Enhancers' locations (bed file)
Human genome with exons & repeats masked (fasta file)
Enhancers' sequences (fasta file)
Human genome with exons, repeats & enhancers masked (fasta file)
200bp enhancers' sequences (fasta file)
CG contents of enhancers
Randomly selected 200bp sequences (fasta file, 3-fold)
Enhancers' k-mer (5,7,9,11) distributions
Background k-mer (5,7,9,11) distributions
Randomly selected 200bp sequences (fasta file, 1-fold)
Dictionary of k-mer fold changes
Positive features (sequential k-mer fold changes)
Negative features (sequential k-mer fold changes)
Initial training data (80%)
Testing data (20%)
Training data (70%)
Validation data (30%)
Deep learning models
Five-fold cross-validation
Accuracy & AUC
Average accuracy & AUC

Figure 1

Flowchart of the SeqEnhDL framework.

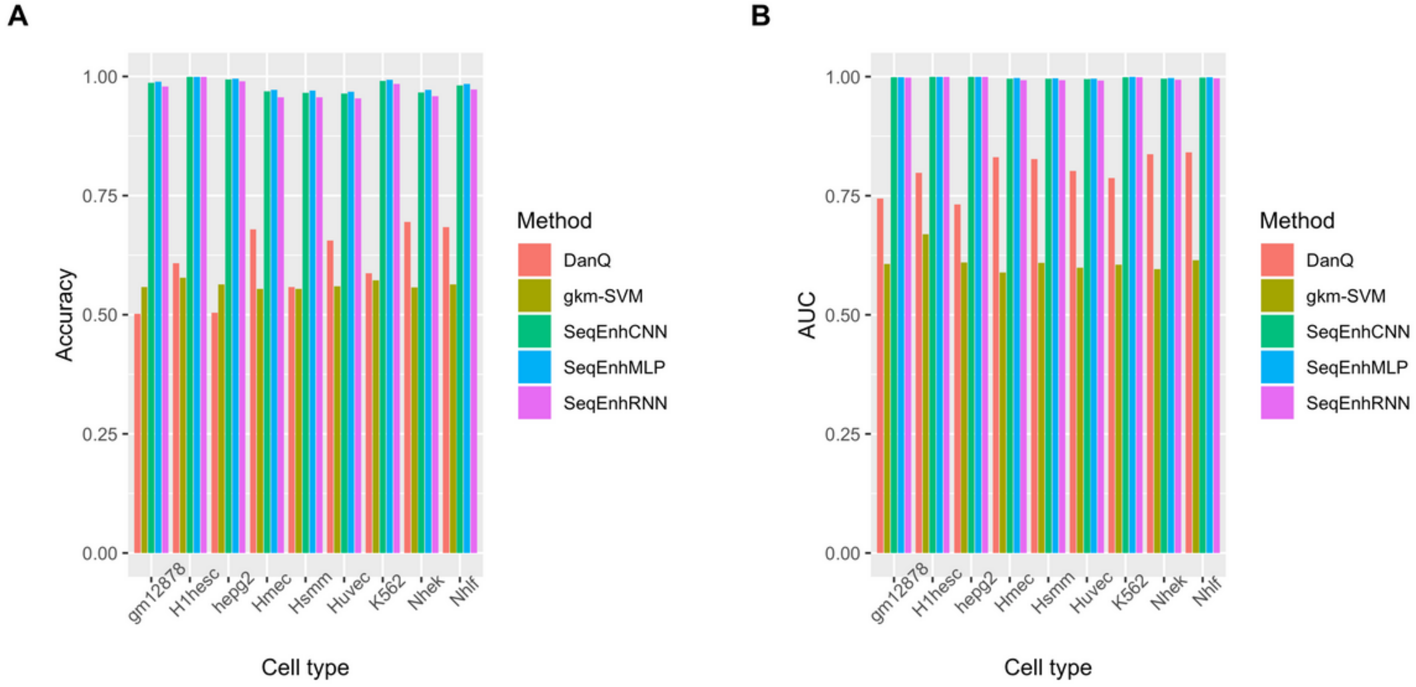

Figure 2

Comparison among different enhancer classifiers with regard to distinguishing cell type-specific enhancers from randomly selected non-coding sequences. (A) Comparison of accuracies. (B) Comparison of AUCs.

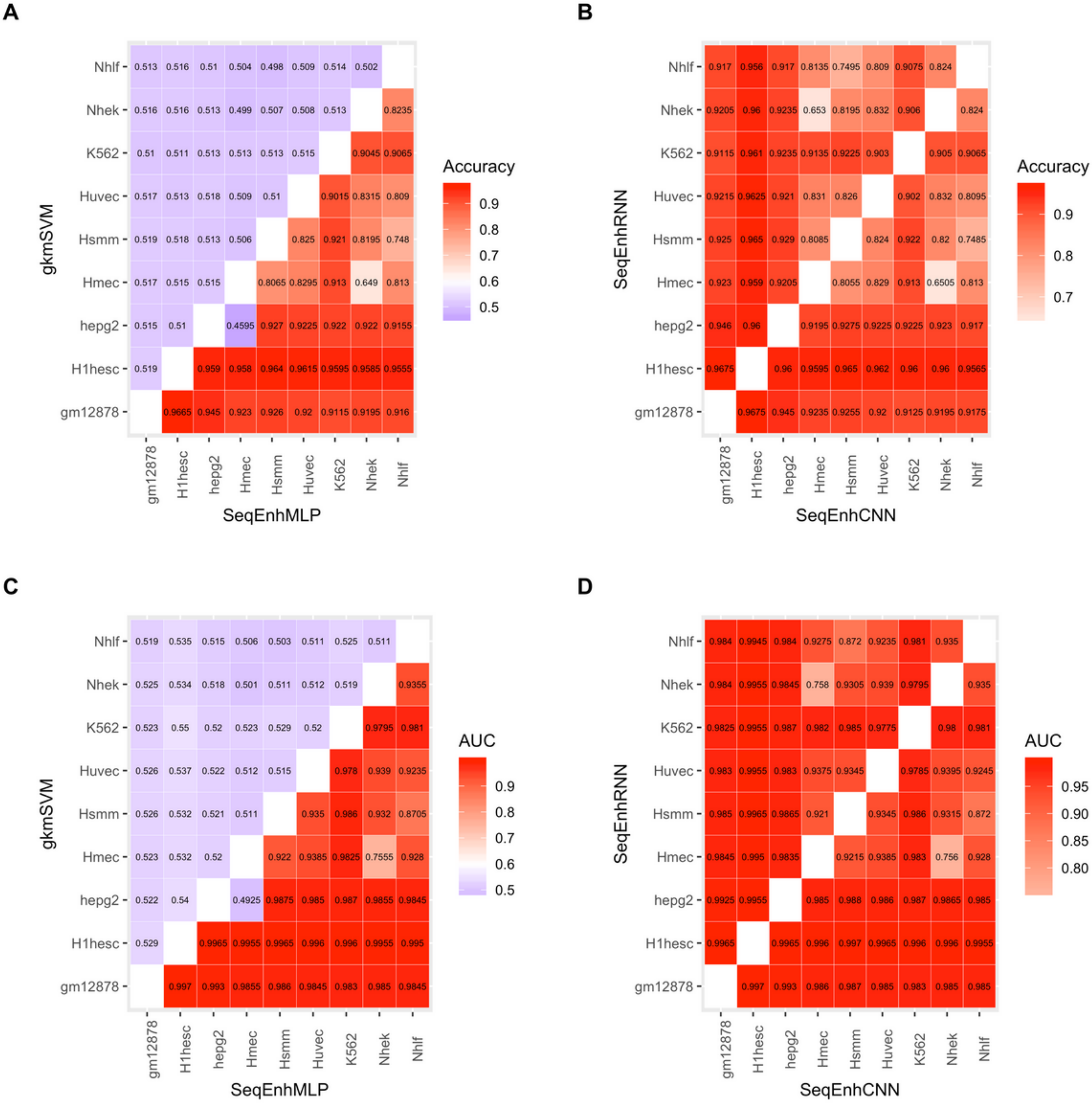

Figure 3

Comparison between gkm-SVM and SeqEnhDL with regard to discriminating enhancers from two cell types. (A) and (B) Comparison of accuracies. (C) and (D) Comparison of AUCs.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementalfileBMCResearchNotes.docx