REVIEW ARTICLE

Noncoding Variants Functional Prioritization Methods Based on Predicted Regulatory Factor Binding Sites

Haoyue Fu^{1,*}, Lianping Yang^{1,2} and Xiangde Zhang¹¹College of Sciences, Northeastern University, Shenyang, China; ²University of Southern California, Dept. Biol. Sci., Program Mol & Computat Biol, USA

Abstract: Backgrounds: With the advent of the post genomic era, the research for the genetic mechanism of the diseases has found to be increasingly depended on the studies of the genes, the gene-networks and gene-protein interaction networks. To explore gene expression and regulation, the researchers have carried out many studies on transcription factors and their binding sites (TFBSs). Based on the large amount of transcription factor binding sites predicting values in the deep learning models, further computation and analysis have been done to reveal the relationship between the gene mutation and the occurrence of the disease. It has been demonstrated that based on the deep learning methods, the performances of the prediction for the functions of the noncoding variants are outperforming than those of the conventional methods. The research on the prediction for functions of Single Nucleotide Polymorphisms (SNPs) is expected to uncover the mechanism of the gene mutation affection on traits and diseases of human beings.

Results: We reviewed the conventional TFBSs identification methods from different perspectives. As for the deep learning methods to predict the TFBSs, we discussed the related problems, such as the raw data preprocessing, the structure design of the deep convolution neural network (CNN) and the model performance measure *et al.* And then we summarized the techniques that usually used in finding out the functional noncoding variants from de novo sequence.

Conclusion: Along with the rapid development of the high-throughout assays, more and more sample data and chromatin features would be conducive to improve the prediction accuracy of the deep convolution neural network for TFBSs identification. Meanwhile, getting more insights into the deep CNN framework itself has been proved useful for both the promotion on model performance and the development for more suitable design to sample data. Based on the feature values predicted by the deep CNN model, the prioritization model for functional noncoding variants would contribute to reveal the affection of gene mutation on the diseases.

Keywords: Transcription factor binding sites, Noncoding variants, Deep convolution neural network, Single nucleotide polymorphisms, Saturated mutagenesis.

1. INTRODUCTION

In molecular biology and genetics, transcription factors (TFs) refer to a series of proteins that can specifically bind DNA sequences. It is specifically binding that TFs regulate genetic information on the process of transcription from DNA to mRNA [1]. To explore gene expression and regulation, the researchers have carried out many studies on transcription factors and their binding sites [2]. By binding on the DNA sequence in Promotor or Enhancer, TFs can increase or decrease the level of the gene transcription [3]. TFs can regulate gene expression by multiple mechanisms [4]. When acting alone or coordinating with other protein complex, TFs promote or depress the binding of the RNA

polymerase to the DNA template [4-7]. Furthermore, TFs are the main regulators of the pluripotency of stem cells. Along with the elaboration to the functions of TFs, stem cell biology gets rapid development [8].

It was proved that a large amount of TF determined the fate of the cell by regulating the expression in specific cell during development process [9, 10]. With the development of the studies on TFBSs and the accumulation of the experimentally validated TFBSs, the data information on transcription regulation has been greatly enriched [11]. The complexities of these new models are increased, but more information on TFBSs is obtained [12].

At the mean time, the researches on multiple genome sequencing, ChIP-seq, the structures of chromatin and the integration of multiple information has promoted the accuracies of the TFBSs motif discovering algorithms [13].

*Address correspondence to this author at the College of Sciences, Northeastern University, Shenyang, China; E-mail: fuhaoyue_neu@sina.com

In this paper, firstly we reviewed a few conventional algorithms to identify TFBSs and listed the main models and representations relating to TFBSs. Furthermore, we summarized the newest deep learning methods to predict the TFBSs. Meanwhile, using the values predicted by these deep learning methods, the functional prioritization of the noncoding variants, such as SNPs, can be efficiently calculated. It has been demonstrated that based on the deep learning methods, the performances of the prediction for the function of the noncoding variants are outperforming than those of conventional methods. The research on the prediction for functions of SNPs is expected to uncover the mechanism of the gene mutation affection on traits and diseases of human beings. In the last part of the paper, several models to prioritize the SNVs functions were summed up.

Single Nucleotide Polymorphisms (SNPs) refers to single nucleotide alteration in the DNA sequence which causes the diversity of genomes [14]. SNPs may appear in coding region, noncoding region or intergenic region of genes. The coding region SNPs may influence the amino acid sequence of a protein and then have an impact on the expression of the gene. The noncoding region SNPs may also affect gene expression by gene splicing, binding with transcription factors, mRNA degradation or other mechanisms [15, 16]. SNP is the main formation of DNA sequence mutation of human genomes and is connected with both the variation of organism traits and the production of disease [17]. The difference of SNP distribution characteristic causes the variety on ethnic origin, genetic disease susceptibility, physiological characteristics and appearance of human beings [18-20]. SNP is also the genomic level reflection of the above difference. So, the study on SNP is meaningful to both the health career and the evolution of human beings [21]. The majority of the disease-related SNPs are located in noncoding region whose working mechanisms are unclear. It is speculated that by specifically binding the TF or influencing on gene splicing, SNPs alter the gene expression, since a causal mutation of gene may affect the normal function of gene products [22]. It is discovered that the gene expression pattern mainly depends on regulatory factors, such as promoters and enhancers, which integrate various signal factors and transcription factors to regulate the expression of the genes [23]. Although great progress has been made, little is known about what is the mechanism on how the majority of functional mutations alter the gene expression.

To solve this issue, studies on prioritization of SNVs (Single Nucleotide Variants) have come into being in recent years, which use the machine learning methods to score the SNPs. Based on the pre-processed feature values, the prioritization pipeline can compute the score of a given SNP's functionality. The method funseq2 [24] even predicts the driver variants of the cancers based on the prioritization. The later prioritization studies have used the features which are the outputs of CNN models. Both the superiority of the performance and the characteristics of just relying on SNP's sequence information make the latter studies deserve further research in the future.

2. CONVENTIONAL METHODS TO IDENTIFY TFBS

2.1. Representation for the Pattern of TFBSs

Recognition of transcription factor binding sites (TFBSs) is to search for DNA sequence fragments similar in both

functionality and pattern. Since the binding sites of the same transcription factor are highly conserved, the patterns "Consensus" [25] and "PSSM" (position-specific scoring matrix) [26] have been devised to describe the characteristic of the TFBSs. On each position of a "Consensus" sequence there is a nucleotide that occurs most frequently in an alignment. Besides, owing to the specificity of the TFBSs, the sequence fragments that are different from the "Consensus" in individual positions might be recognized as the TFBS which are referred as instances. For instance, it is discovered that there are close frequencies of two or three different nucleotides on the same sequence position. In this case, since the "Consensus" cannot express the conservative of the TFBS, "Degenerate consensus" [27] has been usually used as the representation of a TFBS. Modeling the TFBS by statistical methods can flexibly and powerfully describe the nucleotides frequencies on each position of the site. The most-often used statistical model is the PFM (Position Frequency Matrix). However, taking into account the nucleotides bias in the constitution of the sequence, the PFM was commonly transformed into the PWM (Position Weight Matrix).

2.2. Methods Based on PWM Motif

The PWM methods went through a series of developments. The earlier methods used various strategies and approaches to find out the PWMs of the binding sites based on the binding sequences for a given transcription factor. But there were some disadvantages of the methods. The first was that they cannot represent the dependencies between the nucleotides in a binding site. The second was that they did not consider the relationship of separated binding sites of the same transcription factor. The later improvement on the PWM methods focused on the first one and a number of methods with dinucleotide model [28, 29] had been put forward. And then, the k-mer methods [30, 31] were developed which had made up for these two disadvantages. The test shows that the k-mer methods were the best ones among all the conventional methods.

The PWM model (Table 1) was proposed as early as 1982 [32]. There were two aspects that need to be specified of a PWM model. The first was the representation of the model. In the earliest PWM model, the weights of the site positions were regarded as the neutrons in a perception [32]; after that, the information contents were used as the measure that analog the frequencies in a PWM model [33]. The energy model emerged to enrich the PWM representations [34]. The other aspect of the PWM model was on how to devise the learning algorithm. EM algorithm (Expectation Maximization algorithm) [35] was a relatively authoritative one in early days. Later, an extended EM algorithm, the MEME (Multiple EM for Motif Elicitation) algorithm [36, 37], was presented to handle the co-regulated sites. Other algorithms using Gibbs sampling [38-42] were also proved to be effective to discover the PWM for a binding site.

2.3. Various Algorithms and Techniques Applying for TFBSs Prediction

Besides the non-exact PWM model, a class of exact algorithms have been developed to describe the motifs of the binding sites. Both the suffix tree model [43, 44] and projec-

Table 1. Conventional methods for identifying TFBSs.

Training Model Data	EM, Gibbs Simple	Alignment Algorithm	k-mer	HMM	Suffix Tree	Deep Learning
Known binding site	[35-42]	—	—	[125, 126]	[43, 44]	—
ChIP-seq	[55-57]	—	[30, 31]	—	—	[76, 77]
Phylogenetic footprinting	[51, 52]	[53, 54]	—	—	—	—
DNase I & histone-mark	—	—	—	[62-69]	—	[76]
PBM	—	—	—	—	—	[77]

Note: different lines represent methods with different training data; different columns represent methods with different models.

tion [45, 46] model are used as data structures to look for high quality motifs. Meanwhile, the GA (Genetic Algorithm) [47, 48] is applied in the PWM model to help it improve performance and Reid *et al.* [49] have combined the GA algorithm with the suffix tree model to make optimizations and advancements. Based on gradually developed Phylogenetic footprinting [50-54] and ChIP-seq techniques [55-57] lots of related methods for discovering TFBSs have been put forward. Among all the conventional methods, k-mer feature extraction method [30, 31] is acknowledged as the best one in performance. The features in the k-mer method are the frequencies of k-mer subsequences. Both the reduced requirement of training data and the binary feature lag without the threshold values are the superiority of the k-mer to the PWMs methods.

2.4. Integrating Chromatin Features

Now that the recognition effect solely based on the PWM model is often not ideal and it is certified that the bindings of transcription factor have a close relation to histone modification and DNase I hypersensitive site, the methods for identifying TFBSs motif have been enriched by association these chromatin features [58-61]. Many computational methods are proposed which combined with one or more epigenetic data for improving identification performance [62-69].

For instance, Whittington *et al.*'s study [68] has shown that filtering through histone H3K34Me3 threshold can improve binding sites identification accuracies of many transcription factors. Gusmao *et al.* [70] show that among many epigenetic data, DNase I sensitivity is the most significant factor to the effect of identification for TFBSs. DNase I splits DNA according to chromatin states. Compared with nucleotides that are not bound, those binding ones would be degraded less frequently. According to this, the possibility of nucleotides being bound can be inferred by the frequency of their degradation.

3. DEEP LEARNING METHODS TO PREDICT FOR TFBS

3.1. The Next-generation Sequencing Technique

Inferring chromatin effects from de novo sequences is a pressing need to decode genetic information to explain traits

or disease mechanisms of human beings. High-throughout experimental technologies on chromatin profiling provide huge amount of data, which require algorithms that can leverage and harness such big data. One of the most important next-generation sequencing techniques, Chromatin Immunoprecipitation-sequencing (ChIP-seq) [71-75] is an effective method for detecting DNA segments interaction with either transcription factors or histone modifications in genome-wide. With the development of the next generation sequencing technology represented by ChIP-seq, people can easily and cheaply obtain various information of whole-genome level, which contributes to a substantial proportion of biological research, thus establishing data analysis science. For example, an ordinary ChIP-seq experiment presents tens of thousands sequences which would be used altogether in the following computational analysis. As we all know, deep learning excels at big data processing, and there are several methods to identify protein-binding specificity (or transcription factor motif) based on these ChIP-seq data using deep learning model. Owing to this, there is a variety of the raw experiments data, so various data preparation methods are required to process training data.

3.2. Preprocessing the Sample Dataset

DeepSEA [76] is an outstanding method both to identify TFBSs and to prioritize the noncoding variants based on a deep convolution neural network. In its data preprocessing, DeepSEA [76] separates the whole-genome sequences into tens of millions 200-bp subsequences. For each chromatin feature of each 200-bp subsequences, DeepSEA [76] assigns 1 to the corresponding chromatin feature label if more than half of the 200-bp appear in the peak regions of the corresponding chromatin feature profiling and otherwise 0. For these tens of millions labeled subsequences, DeepSEA screens those which have more than one TF-binding feature positive. Therefore, altogether almost 17 percent of the whole-genome is left for the training, validation, and evaluating data of DeepSEA. It is worth noting that DeepSEA does not deliberately produce negative sample. This is due to the framework of DeepSEA which provides a sharing multi-task device that simultaneously predicts multiple chromatin features. The label vector contains many 1 and 0 components, which makes the extra negative samples unnecessary.

Nevertheless, another deep learning method DeepBind [77] turns to shuffled negative samples for help. One of the targets of DeepBind is to identify transcription factor binding sites using ChIP-seq data of ENCODE database [78-82]. DeepBind firstly filters and gains 506 TF binding profile datasets after removing some biases. In each profile dataset file, the sequences of the peaks are used as positive samples of DeepBind training, and the dinucleotide shuffles are used as the negative ones.

Thus, by comparing the pre-processing methods of training data using by different deep learning methods, we see that the data preprocessing methods are flexible with a variety even if their raw data are derived from the same ENCODE database. It is notable that DeepBind [77] can harness several different kinds of source data. PBM [83, 84] is another database that is used to verify DeepBind performance. Furthermore, using the specific PBM data, Dream5 [85] project becomes a benchmark platform for measuring the specificity of sequence functional elements and DeepBind [77] and has been proved to be the best method in all the participant teams. PBM gives the intensity of each microarray probe, which represents the specificity of the sequence. That is to say, using the PBM data, the machine learning model needs to predict an intensity value, while a model using ChIP-seq data needs to predict a class label. Because of the characteristic of the sample data, there is no need for the negative samples in the preprocess data of PBM.

3.3. Deep Convolution Neural Network for Identifying the TFBSs

The latest research findings indicate that the technique of deep learning can more efficiently discover the regulatory code of DNA sequence than all previous methods. The deep learning methods used here are those deep convolution neural networks. Compared to those previous motif discovery algorithms, such as PWM model [26, 86-89], dinucleotide model [28, 29, 90-92] and k-mer model [30, 31], the deep convolution neural networks make it possible to extract the long-range dependencies along the sequence.

Convolution-pooling-connection framework is the common character of the convolution neural networks applying in DNA sequence functional motif discovery. Convolution layer is used to identify local sequence code and the deeper is the convolution, the longer-range is the functional code of the sequence, and specifically, the matrixes in the first convolution layer is corresponding to the PWMs of the TFBSs; the rectified function substitutes minus to 0. The variation in the resolution of the sequence mainly relies on the pooling layer which alleviates the local position difference by maximizing operation. Finally, the connection layer refers to full connection which is a general neural setting.

In the different methods, the frameworks of the deep learning have a slight difference. For example, in DeepBind, instead of a single sequence input, a batch of input sequences are processed simultaneously; in Basset [93], the input of a sequence is simultaneously computed in altogether 164 cell-lines backgrounds; in DeepSEA, the multiple predictors are constructed into a framework, which implements the sharing of model sequence code. The comparison of the frameworks of the deep learning methods is shown in (Fig. 1).

3.4. Performance Measure

In term of the performances of the methods, the authoritative measure is ROC (Receiver Operating Characteristic) curve or the corresponding AUC (Area Under Curve) area. After being trained, for a given input, the deep learning classifier would compute a value and there need a threshold to classify the value into a class. Note that the FP(false positive) and TP(true positive) would change with the threshold changing. The ROC curve is generated by showing the pairs of the true positive rate and the corresponding false positive rate. As for the Dream5 PBM data platform, what is needed to predict is the intensity not the class label, so the threshold value is useless and it is the Pearson correlation not the AUC that is used to measure the performance of the predictor.

3.5. Deep convolution neural network improvement

A modified CNN framework DeepMotif [94] is presented in a recent paper, which makes comparison with the DeepBind in performance. The modification mainly lies in two ways: 1. the highway is substituted for the original conventional full connected-layer; 2. A better visualization motif map is given. There is also another method DanQ [95] that revises the DeepSEA into a hybrid model integrating both the CNN and the RNN. DanQ uses the same training data as DeepSEA and demonstrates that DanQ outperforms DeepSEA.

It should also be noted that the parameters of a CNN would affect the prediction performance. Zeng *et al.* [96] make experiments with various parameters and draw the conclusion that the improved CNN model excels in the performance by effectively adjusting the parameters, such as the number of the convolution layers, the number of the convolution filters and the length of the convolution filter window *et al.*

Along with the rapid development of the high-throughout assays, more and more positive sample data would be ready for training. More than this, more and more chromatin features would be conducive to improve CNN model prediction accuracy. What is more, getting more insights into the CNN framework itself has been proved useful for the promotion on model performance [97].

4. EXPLORING THE RELATIONSHIP BETWEEN THE SNVs AND THE DISEASE

With the advent of the post genomic era, the researches for the genetic mechanism of the diseases have increasingly depended on the studies on the genes, the gene-networks and gene-protein interaction networks [98-100]. Based on the large amount of TFBSs predicting values in the deep learning models, further computation and analysis have been done to reveal the relationship between the gene mutation and the occurrence of the disease, whereas the saturated mutagenesis is a necessary tool by which the influence of a mutation on each single position can be quantitatively measured.

4.1. Saturated Mutagenesis

Mutagenesis employs the physical, chemical or biological factors to induce mutation of the organism to occur, and breeds the new cultivars from the variant offspring based on

Fig. (1). Comparison framework on deep convolution neural network. It is obvious that the common to all framework is the convolution layer and the max pool layer. The differences are: **(a)** framework in DeepSEA [76] shares weights between multiple chromatin feature predictors; **(b)** framework in Basset shares weights between multiple cell type predictors **(c)** framework in DeepBind shares weights among batch of input sequences; **(d)** instead of fully connected layer in DeepSEA, framework in DanQ sets Recurrent and Dense layer **(e)** in framework in DeepMotif, a highway MLP layer takes the place of fully connected layer.

Table 2. Lists of methods for SNP functional prioritization.

Method	Features Used	Classifier
FunSeq2 [24]	In functional annotations In sensitive regions In ultra-sensitive regions Motif-breaking score (PWM changes) Motif-gaining score (PWM changes) Network centrality score GERP score In ultra-conserved elements In HOT regions In regulatory elements associated with genes Recurrent in multiple samples	Weighted scoring scheme
CADD [123]	conservation metrics; regulatory information; transcript information; protein-level scores (altogether 63 distinct annotations)	SVM model
GWAVA [124]	Open chromatin. Transcription factor binding. Histone modifications. RNA polymerase binding. CpG islands. Genome segmentation. Conservation. Human variation. Genic context. Sequence context.	Modified random forest algorithm
DeepSEA [76]	Evolutionary conservation scores Absolute difference features (919 features) Relative difference features (919 features)	Boosted logistic regression classifier
DeepFind [77]	~600 DeepBind TF predictor values for the wild type and mutant sequences (altogether ~1,200 features)	Neural network

breeding objective [101]. Mutagenesis technology can obtain excellent mutant in a short time, which has advantages over hybridization. Quantitative analysis of mutagenesis has been used in large amounts in bioinformatics. The impact of a gene locus on a phenotype is called “effect size” [102], which can be used to measure the effect caused by a single gene mutation on the phenotype. However, it is discovered that the effect size, the extent to which the allelic variation gives rise to the phenotypic change, is often low. Apart from deriving the gene mechanism of the disease phenotype [103, 104], mutagenesis can be applied for quantifying the specificity of each nucleotide in a sequence of DNA [105]. Saturated mutagenesis [106] is a procedure in which all possible nucleotides on a specific position are to be substituted in a given sequence, which makes possible for quantifying the specificity of a certain sequence on a single-nucleotide resolution.

4.2. Prediction Causal SNVs

The destination of the research on chromatin effects is to find the mechanism by which the traits or the diseases are

regulated [107]. In order to achieve that, accurate prediction for such chromatin effects is the first step. The focus study of the relationship between the SNP and the diseases is helpful to accomplish the goal. Both DeepSEA and DeepBind have made a lot of work in the establishment of the diseases dependency on the SNPs.

Using the results from the saturated mutagenesis, both DeepSEA [76] and DeepBind [77] demonstrate they have effectively exploited the effects of the SNVs on certain disorders, which have been verified in the previous experiments [108-119].

DeepBind have especially devised a visualized method---mutant map to intuitively display the SNPs’ influence on each position. For example, a mutant map of the saturated mutagenesis in [77] visually shows that the mutant C->D would decrease the transcription factor SP1 binding in the promoter of LDL-R, while the mutant G->C would increase the binding, while the C->G and C->T mutants have been recorded in Human Gene Mutation Database(HGMD) [77, 120].

4.3. Functional SNVs Prioritization

Besides individual cases, both DeepSEA and DeepBind have developed the model to predict the quantitative effect of SNVs caused by mutations of single nucleotide. It is believed that the genetic diseases are the result of DNA mutations, and the SNP is the dominant form of DNA sequence mutation in human genome. For example, the SNP may cause cancers, infectious diseases, autoimmune disorders, neurological illnesses, sickle-cell anaemias, thalassaemias and cystic fibrosis. Studies show that those disease-related SNPs usually locate in noncoding region and the mechanism of the SNP still needs further research. In the recent years, a great progress has been made on the research in the area. Researchers find that the majority of the trait-related DNA mutations change not the gene itself but the regular factors which control the gene expressions. However, surprisingly little is known for us exactly how most of the functional mutations modify the expression of the gene in detail. With the development of DNA sequencing and the decline of the cost of sequencing, researches on the SNP are rapidly expanding, which has done the full preparation for big data analysis. DeepSEA [76] is the first algorithm which large-scaled predicts the functional prioritization of SNPs from de novo sequences. Based on the chromatin feature values predicted by DeepSEA [76] deep learning model, altogether 1842 features are computed, which are generated by reference allele's 919 features and its corresponding alternative allele's 919 ones. The label of the sample is binary indicator of whether it is a real SNP. The machine learning model is Boosted logistic regression classifier [121]. The model is verified on three different databases. The negative sample is selected from sequences distant from the positive ones. As mentioned above, a classification model needs both positive samples and negative samples, hence it is inevitable to design proper negative samples in order to get better prediction performance and the authors of DeepSEA has testified the different performance with different selection of negative samples.

DeepBind has developed a special modular DeepFind for predicting whether a SNV is causal. In the respect of the classifier, DeepFind adopts a neural network model which takes 1192 features as input which are generated from a wild-type sequence and its corresponding mutant one. Since it is a binary classifier, the positive samples and the negative ones are indispensable. DeepFind mainly carries out the experiment on CADD [122] simulator, using the 'simulated' SUVs as positive samples and the 'observed' ones as negative.

It is because of the classifier provided by DeepSEA [76] or DeepBind [77] that makes it possible to discriminate whether a noncoding variation (or a SNP) sequence is a functional one (or a disease-related one), which provides the basis for further research on the genetic derivation of the diseases (Table 2).

CONFLICT OF INTEREST

This work is supported by the Fundamental Research Funds for the Central Universities of China (Grant No. N120305005 and No. N130305006) and by the National Science Foundation of China (Grant No.31301086).

ACKNOWLEDGEMENTS

We would like to express our gratitude to all those who helped us during the writing of this paper. HF thanks Xiaojun Lu and Qingsong Tang for useful discussions and advices.

REFERENCES

- [1] Latchman, D.S. Transcription factors: An overview. *Intl. J. Biochem. Cell Biol.*, **1998**, 29(12), 1305-1312.
- [2] Pennacchio, L.A.; Rubin, E.M. Genomic strategies to identify mammalian regulatory sequences. *Nat. Rev. Genet.*, **2001**, 2(2), 100-9.
- [3] Hill, C.S.; Treisman, R. Transcriptional regulation by extracellular signals: mechanisms and specificity. *Cell*, **1995**, 80(2), 199-211.
- [4] Gill, G. Regulation of the initiation of eukaryotic transcription. *Essays In Biochem.*, **2001**, 37(37), 33-43.
- [5] Roeder, R.G. The role of general initiation factors in transcription by RNA polymerase II. *Trends Biochem. Sci.*, **1996**, 21(9), 327-335.
- [6] Nikolov, D.B.; Burley, S.K. RNA polymerase II transcription initiation: A structural view. *Proc. Natl. Acad. Sci.*, **1997**, 94(1), 15-22.
- [7] Leetong, I.; Young, R.A. Transcription of Eukaryotic Protein-Coding Genes. *Ann. Rev. Genet.*, **2000**, 34(1), 77-137.
- [8] Nichols, J.; Zevnik, B.; Anastassiadis, K.; Niwa, H.; Klewe-Nebenius, D.; Chambers, I.; Scholer, H.; Smith, A. Formation of pluripotent stem cells in the mammalian embryo depends on the POU transcription factor Oct 4. *Cell*, **1998**, 95(3), 379-391.
- [9] Lefebvre, V.; Dumitriu, B.; Penzo-Méndez, A.; Han, Y.; Pallavi, B. Control of cell fate and differentiation by Sry-related high-mobility-group box (Sox) transcription factors. *Intl. J. Biochem. Cell Biol.*, **2007**, 39(12), 2195-2214.
- [10] Pei, D. Regulation of pluripotency and reprogramming by transcription factors. *J. Biol. Chem.*, **2009**, 284(6), 3365-9.
- [11] Zambelli, F.; Pesole, G.; Pavesi, G. Motif discovery and transcription factor binding sites before and after the next-generation sequencing era. *Briefings Bioinform.*, **2012**, 14(2), 225-237.
- [12] Bulyk, M.L. Computational prediction of transcription-factor binding site location// Introduction to Lie algebras and representation theory /. Springer-Verlag, **1972**, 201.
- [13] Weirauch, M.T.; Cote, A.; Annala, M.; Zhao, Y.; Riley, T.R.; Saez-Rodriguez, J.; Cokelaer, T.; Vedenko, A.; Talukder, S.; Bussemaker, H.J.; Morris, Q.D.; Bulyk, M.L.; Stolovitzky, G.; Hughes, T.R. Evaluation of methods for modeling transcription-factor sequence specificity. *Nat. Biotechnol.*, **2013**, 31(2), 126-34.
- [14] Sachidanandam, R.; Weissman, D.; Schmidt, S.C. Kakol, J.M.; Stein, L.D.; Marth, G.; Sherry, S.; Mullikin, J.C.; Mortimore, B.J.; Willey, D.L.; Hunt, S.E.; Cole, C.G.; Coggill, P.C.; Rice, C.M.; Ning, Z.; Rogers, J.; Bentley, D.R.; Kwok, P.; Mardis, E.R.; Yeh, R.T.; Schultz, B.; Cook, L.; Davenport, R.; Dante, M.; Fulton, L.; Hillier, L.; Waterston, R.H.; McPherson, J.D. A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature*, **2001**, 409(6822), 928-933.
- [15] Genome, T.H.; Wang, D.G.; Ai, E. Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in. *Science*, **2007**, 280(5366), 1077-1082.
- [16] Cargill, M.; Altshule, D.; Ireland, J.; Sklar, P.; Ardlie, K.; Patil, N.; Lane, C.R.; Lim1, E.P.; Kalyanaraman, N.; Nemesh, J.; Ziaugra, L.; Friedland, L.; Rolfe, A.; Warrington, J.; Lipshutz, R.; Daley, G.Q.; Lander, E.S. Characterization of single-nucleotide polymorphisms in coding regions of human genes. *Nat. Genet.*, **1999**, 22(3), 231-238.
- [17] Syvänen, A.C. Accessing genetic variation: genotyping single nucleotide polymorphisms. *Nat. Rev. Genet.*, **2001**, 2(12), 930-942.
- [18] Li, Z.; Zou, L.J.; Rong, H.; Li, B. Association of single-nucleotide polymorphisms in toll-like receptor 5 gene with rheumatic heart disease in Chinese Han population. *Intl. J. Cardiol.*, **2010**, 145(1), 129-130.
- [19] Ignatovica, V.; Latkovskis, G.; Peculis, R.; Megnis, K.; Schioth, H. B.; Vaivade, I.; Fridmanis, D.; Pirags, V.; Erglis, A.; Klovins, J. Single nucleotide polymorphisms of the purinergic 1 receptor are not associated with myocardial infarction in a Latvian population. *Mol. Biol. Rep.*, **2012**, 39(2), 1917-1925.
- [20] Brumfield, R. T.; Beerli, P.; Nickerson, D.A.; Edwards, S.V. The

- utility of single nucleotide polymorphisms in inferences of population history. *Trends Ecol. Evol.*, **2003**, 18(5), 249-256.
- [21] Wakeley, J.; Nielsen, R.; Liu-Cordero, S.N.; Ardlie, K. The discovery of single-nucleotide polymorphisms and inferences about human demographic history. *Am. J. Hum. Genet.*, **2001**, 69(6), 1332-1347.
- [22] Butter, F.; Davison, L.; Viturawong, T.; Scheibe, M.; Vermeulen, M.; Todd, J.A. Proteome-wide analysis of disease-associated SNPs that show allele-specific transcription factor binding. *Plos Genetics*, **2012**, 8(9), 2364-2366.
- [23] He, X.; Sinha, S. ChIPs and regulatory bits. *Nature Biotechnology*, **2010**, 28(2), 142-143.
- [24] Fu, Y.; Liu, Z.; Lou, S.; Bedford, J.; Mu, X.J.; Yip, K.Y.; Khurana, E.; Gerstein, M. FunSeq2: A framework for prioritizing noncoding regulatory variants in cancer. *Genome Biol.*, **2014**, 15(10), 1-15. (FunSeq2)
- [25] Schneider, T.D. Consensus sequence Zen. *Appl. Bioinformat.*, **2002**, 1(3), 111-119.
- [26] Stormo, G.D. DNA binding sites: representation and discovery. *Bioinformatics*, **2000**, 16(1), 16-23.
- [27] Rose, T.M.; Schultz, E.R.; Henikoff, J.G.; Pietrokovski, S.; McCallum, C.M.; Henikoff, S. Consensus-degenerate hybrid oligonucleotide primers for amplification of distantly related sequences. *Nucleic Acids Res.*, **1998**, 26(7), 1628-1635.
- [28] Yang, K.L.; Qiang, X.U. Recognition of the transcription factor binding sites in *Saccharomyces cerevisiae* genome based on dinucleotides position weight matrix. *Life Sci. Res.*, **2008**, 12(2), 115-120.
- [29] Kulakovskiy I.V.; Levitsky, V.; Oshchepkov, D.; Bryzgalov, L.; Vorontsov, I.; Makeev, V. From binding motifs in CHIP-SEQ data to improved models of transcription factor binding sites. *J. Bioinform. Comput. Biol.*, **2013**, 11(1), 1340004.
- [30] Lee, D.; Karchin, R.; Beer, M.A. Discriminative prediction of mammalian enhancers from DNA sequence. *Genome Res.*, **2011**, 21(12), 2167-2180.
- [31] Ghandi, M.; Lee, D.; Mohammadnoori, M.; Beer, M.A. Enhanced regulatory sequence prediction using gapped k-mer features. *Plos Computat. Biol.*, **2014**, 10(7), e1003711-e1003711.
- [32] Stormo, G.D.; Schneider, T.D.; Gold, L.; Ehrenfeucht, A. Use of the 'Perceptron' algorithm to distinguish translational initiation sites in *E. coli*. [M]// Bibliography of low energy electron and photon cross section data (January 1975 through December 1977) / U.S. Dept. of Commerce, National Bureau of Standards: for sale by the Supt. of Docs. U.S. Govt. Print. Off. **1979**, 2997-3011.
- [33] Schneider, T.D.; Stormo, G.D.; Gold, L.; Ehrenfeucht, A. Information content of binding sites on nucleotide sequences. *J. Mol. Biol.*, **1986**, 188(3), 415-431.
- [34] Heumann, J.M.; Lapedes, A.S.; Stormo, G.D. Neural networks for determining protein specificity and multiple alignment of binding sites. *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, **1994**, 188-94.
- [35] Lawrence, C. E.; Reilly, A.A. An expectation maximization (EM) algorithm for the identification and characterization of common sites in unaligned biopolymer sequences. *Proteins: Struct. Funct. Bioinf.*, **1990**, 7(1), 41-51.
- [36] Bailey, T.L.; Elkan, C. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. International Conference on Intelligent Systems for Molecular Biology. *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, **1994**, 28-36.
- [37] Bailey, T.L.; Elkan, C. The value of prior knowledge in discovering motifs with MEME. *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, **1995**, 21-29.
- [38] Lawrence, C.E.; Altschul, S.F.; Boguski, M.S.; Liu, J.S.; Neuwald, A.F.; Wootton, J.C. Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. *Science*, **1993**, 262(5131), 208-214.
- [39] Neuwald, A.F.; Liu, J.S.; Lawrence, C.E. Gibbs motif sampling: Detection of bacterial outer membrane protein repeats. *Protein Sci.*, **1995**, 4(8), 1618-1632.
- [40] Hughes, J.D.; Estep, P.W.; Tavazoie, S.; Church, G.M. Computational identification of cis-regulatory elements associated with groups of functionally related genes in *Saccharomyces cerevisiae*. *J. Mol. Biol.*, **2000**, 296(5), 1205-1214.
- [41] Workman, C.T.; Stormo, G.; D. ANN-SPEC: A method for discovering transcription factor binding sites with improved specificity. *Pac. Symp. Biocomput.*, **2000**, 5, 467-78.
- [42] Liu, X.; Brutlag, D.L.; Liu, J.S. Bioproscpector: Discovering conserved DNA motifs in upstream regulatory regions of co-expressed genes. *Pac. Symp. Biocomput.*, **2000**, 41(1), 127-138.
- [43] Marsan, L.; Sagot, M.F. Algorithms for extracting structured motifs using a suffix tree with an application to promoter and regulatory site consensus identification. *J. Computat. Biol.*, **2000**, 7(7), 345-362.
- [44] Pavesi, G.; Mauri, G.; Pesole, G. An algorithm for finding signals of unknown length in DNA sequences. *Bioinformatics*, **2001**, 17 (Suppl), 207-214(8).
- [45] Buhler, J.; Tompa, M. Finding motifs using random projections. *J. Computat. Biol. A J. Computat. Mol. Cell Biol.*, **2002**, 9(2), 225-42.
- [46] Raphael, B.; Liu, L.T.; Varghese, G. A uniform projection method for motif discovery in DNA sequences. *IEEE/ACM Transact. Computat. Biol. Bioinformat.*, **2004**, 1(2), 91-94.
- [47] Wei, Z.; Jensen, S.T. GAME: detecting cis-regulatory elements using a genetic algorithm. *Bioinformatics*, **2006**, 22(13), 1577-1584.
- [48] Li, L. GADEM: a genetic algorithm guided formation of spaced dyads coupled with an EM algorithm for motif discovery. *J. Computat. Mol. Cell Biol.*, **2009**, 16(2), 317-329.
- [49] Reid, J.E.; Wernisch, L. STEME: efficient EM to find motifs in large data sets. *Nucleic Acids Res.*, **2011**, 39(18), 729-738.
- [50] Mccue, L.A.; Thompson, W.; Carmack, C. S.; Ryan, M.P.; Liu, J.S.; Derbyshire, V.; Lawrence, C.E. Phylogenetic footprinting of transcription factor binding sites in proteobacterial genomes. *Nucleic Acids Res.*, **2001**, 29(3), 774-82.
- [51] Sinha, S.; Blanchette, M.; Tompa, M. PhyME: A probabilistic algorithm for finding motifs in sets of orthologous sequences. *BMC Bioinformat.*, **2004**, 5(43), 1-17.
- [52] Siddharthan, R.; Siggia, E.D.; Nimwegen, E.V. PhyloGibbs: A gibbs sampling motif finder that incorporates phylogeny. *Plos Computat. Biol.*, **2005**, 1(7), e67.
- [53] Wang, T.; Stormo, G.D. Combining phylogenetic data with co-regulated genes to identify regulatory motifs. *Bioinformatics*, **2003**, 19(18), 2369-2380.
- [54] Moses, A.M.; Chiang, D.Y.; Pollard, D.A.; Iyer, V. N.; Eisen, M.B. MONKEY: identifying conserved transcription-factor binding sites in multiple alignments using a binding site-specific evolutionary model. *Genome Biol.*, **2004**, 5(12), 60-60.
- [55] Machanick, P.; Bailey, T.L. MEME-ChIP: motif analysis of large DNA datasets. *Bioinformatics*, **2011**, 27(12), 1696-1697.
- [56] Hu, M.; Yu, J.; Taylor, J.M.G.; Chinnaiyan, A.M.; Qin, Z.S. On the detection and refinement of transcription factor binding sites using ChIP-Seq data. *Nucleic Acids Res.*, **2010**, 38(7), 2154-2167.
- [57] Bailey, T.L. DREME: motif discovery in transcription factor ChIP-seq data. *Bioinformatics*, **2011**, 27(12), 1653-1659.
- [58] Kodama, Y.; Nagaya, S.; Shinmyo, A.; Kato, K. Mapping and characterization of DNase I hypersensitive sites in Arabidopsis chromatin. *Plant Cell Physiol.*, **2007**, 48(3), 459-470.
- [59] Boyle, A.P.; Davis, S.; Shulha, H.P.; Meltzer, P.; Margulies, E.H.; Weng, Z.; Furey, T.S.; Crawford, G.E. High-Resolution Mapping and Characterization of Open Chromatin across the Genome. *Cell*, **2008**, 132(2), 311-322.
- [60] Sheffield, N.C.; Thurman, R.E.; Song, L.; Safi, A.; Stamatoyannopoulos, J.A.; Lenhard, B.; Crawford, G.E.; Furey, T.S. Patterns of regulatory activity across diverse human cell types predict tissue identity, transcription factor binding, and long-range interactions. *Genome Res.*, **2013**, 23(5), 777-788.
- [61] Mercer, T.R.; Edwards, S.L.; Clark, M.B.; Neph, S.J.; Wang, H.; Stergachis, A.B.; John, S.; Sandstrom, R.; Li, G.; Sandhu, K.S.; Nielsen, Y. R.L.K.; Mattick, J.S.; Stamatoyannopoulos, J. A. DNase I-hypersensitive exons colocalize with promoters and distal regulatory elements. *Nat. Genet.*, **2013**, 45(8), 852-859.
- [62] Boyle, A. P.; Song, L.; Lee, B. K.; London, D.; Keefe, D.; Birney, E.; Iyer, V. R.; Crawford, G. E.; Furey, T. S. High-resolution genome-wide *in vivo* footprinting of diverse transcription factors in human cells. *Genome Res.*, **2011**, 21(3), 456-464.
- [63] Cuellar-Partida, G.; Buske, F.A.; McLeay, R.C.; Whittington, T.; Noble, W.S.; Bailey, T.L. Epigenetic priors for identifying active transcription factor binding sites. *Bioinformatics*, **2012**, 28(1), 56-62.
- [64] Gusmão, E.G.; Dieterich, C.; Costa, I.G. Prediction of transcription factor binding sites by integrating DNase digestion and histone modification. *Brazilian Symposium on Bioinformat.*, **2012**, 109-119.
- [65] Natarajan, A.; Yardimci, G. G.; Sheffield, N. C.; Frazer, K. A.

- Predicting cell-type-specific gene expression from regions of open chromatin. *Genome Res.*, **2012**, 22(9), 1711-1722.
- [66] Neph, S.; Vierstra, J.; Stergachis, A.B.; Reynolds, A.P.; Haugen, E.; Vernot, B.; Thurman, R.E.; John, S.; Sandstrom, R.; Johnson, A.K.; Maurano, M.T.; Humbert, R.; Rynes, E.; Wang, H.; Vong, S.; Lee, K.; Bates, D.; Diegel, M.; Roach, V.; Dunn, D.; Neri, J.; Schafer, A.; Hansen, R. S.; Kutayavin, T.; Giste, E.; Weaver, M.; Canfield, T.; Sabo, P.; Zhang, M.; Balasundaram, G.; Byron, R.; MacCoss, M.J.; Akey, J.M.; Bender, M.A.; Groudine, M.; Kaul, R.; Stamatoiyannopoulos, J.A. An expansive human regulatory lexicon encoded in transcription factor footprints. *Nature*, **2012**, 489(7414), 83-90.
- [67] Pique-Regi, R.; Degner, J.F.; Pai, A.A.; Gaffney, D.J.; Gilad, Y.; Pritchard, J. K. Accurate inference of transcription factor binding from DNA sequence and chromatin accessibility data. *Genome Res.*, **2011**, 21(3), 447-455.
- [68] Whittington, T.; Perkins, A.C.; Bailey, A.T.L. High-throughput chromatin information enables accurate tissue-specific prediction of transcription factor binding sites. *Nucleic Acids Res.*, **2009**, 37(1), 14-25.
- [69] Won, K.J.; Ren, B.; Wang, W. Genome-wide prediction of transcription factor binding sites using an integrated model. *Genome Biol.*, **2010**, 11(1), 79-82.
- [70] Gusmao, E.G.; Dieterich, C.; Zenke, M.; Costa, I.G. Detection of active transcription factor binding sites with the combination of DNase hypersensitivity and histone modifications. *Bioinformatics*, **2014**, 30(22), 3143-3151.
- [71] Schuster, S.C. Next-generation sequencing transforms today's biology. *Nat. Methods*, **2008**, 5(1), 16-18.
- [72] Shendure, J.; Ji, H. Next-generation DNA sequencing. *Nature Biotechnology*, **2008**, 26(10), 1135-1145.
- [73] Metzker, M.L. Sequencing technologies - the next generation. *Nat Rev Genetics*, **2010**, 11(1), 1-13.
- [74] Massie, C.E.; Mills, I.G. ChIPping away at gene regulation. *Embo Reports*, **2008**, 9(4), 337-343.
- [75] Park, P.J. ChIP-seq: advantages and challenges of a maturing technology. *Nat. Rev. Genet.*, **2009**, 10(10), 669-680.
- [76] Zhou, J.; Troyanskaya, O.G. Predicting effects of noncoding variants with deep learning-based sequence model. *Nat. Methods*, **2015**, 12(10), 931-934.
- [77] Alipanahi, B.; Delong, A.; Weirauch, M.T.; Frey, B.J. Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nat. Biotechnol.*, **2015**, 33(8), 831-8.
- [78] Gerstein, M.B.; Kundaje, A.; Hariharan, M.; Landt, S.G.; Yan, K.; Cheng, C.; Mu, X.J.; Khurana, E.; Rozowsky, J.; Alexander, R.; Min, R.; Alves, P.; Abyzov, A.; Addleman, N.; Bhardwaj, N.; Boyle, A.P.; Cayting, P.; Charos, A.; Chen, D.Z.; Cheng, Y.; Clarke, D.; Eastman, C.; Euskirchen, G.; Fietze, S.; Fu, Y.; Gertz, J.; Grubert, F.; Harmanci, A.; Jain, P.; Kasowski, M.; Lacroute, P.; Leng, J.J.; Lian, J.; Monahan, H.; O'Geen, H.; Ouyang, Z.; Partridge, E.C.; Patacsil, D.; Pauli, F.; Raha, D.; Ramirez, L.; Reddy, T.E.; Reed, B.; Shi, M.; Slifer, T.; Wang, J.; Wu, L.; Yang, X.; Yip, Y.K.; Zilberman-Schapira, G.; Batzoglou, S.; Sidow, A.; Farnham, P.J.; Myers, R.M.; Weissman, S.M.; Snyder, M.; Architecture of the human regulatory network derived from ENCODE data. *Nature*, **2012**, 489(7414), 91-100.
- [79] Rosenbloom, K.R.; Dreszer, T.R.; Pheasant, M.; Barber, G.P.; Meyer, L.R.; Pohl, A.; Raney, B.J.; Wang, T.; Hinrichs, A.S.; Zweig, A.S.; Fujita, P.A.; Learned, K.; Rhead, B.; Smith, K.E.; Kuhn, R.M.; Karolchik, D.; Haussler, D.; Kent, W.J. ENCODE whole-genome data in the UCSC Genome Browser. *Nucleic Acids Res.*, **2010**, 38(Database issue), D620-D625.
- [80] Maher, B. ENCODE: The human encyclopaedia. *Nature*, **2012**, 489(7414), 46-8.
- [81] Consortium, T.E.P. A User's Guide to the Encyclopedia of DNA Elements (ENCODE) *Plos Biol.*, **2011**, 9(4), e1001046.
- [82] Thurman, R.E.; Rynes, E.; Humbert, R.; Vierstra, J.; Maurano, M.T.; Haugen, E.; Sheffield, N.C.; Stergachis, A.B.; Wang, H.; Vernot, B.; Garg, K.; John, S.; Sandstrom, R.; Bates, D.; Boatman, L.; Canfield, T.K.; Diegel, M.; Dunn, D.; Ebersol, A.K.; Frum, T.; Giste, E.; Johnson, A.K.; Johnson, E.M.; Kutayavin, T.; Lajoie, B.; Lee, B.K.; Lee, K.; London, D.; Lotakis, D.; Neph, S.; Neri, F.; Nguyen, E.D.; Qu, H.; Reynolds, A.P.; Roach, V.; Safi, A.; Sanchez, M.E.; Sanyal, A.; Shafer, A.; Simon, J.M.; Song, L.; Vong, S.; Weaver, M.; Yan, Y.; Zhang, Z.; Zhang, Z.; Lenhard, B.; Tewari, M.; Dorschner, M.O.; Hansen, R.S.; Navas, P.A.; Stamatoiyannopoulos, G.; Iyer, V.R.; Lieb, J.D.; Sunyaev, S.R.; Akey, J.M.; Sabo, P.J.; Kaul, R.; Furey, T.S.; Dekker, J.; Crawford, G.E.; Stamatoiyannopoulos, J.A. The accessible chromatin landscape of the human genome. *Nature*, **2012**, 489(7414), 75-82.
- [83] Berger, M.F.; Bulyk, M.L. Protein binding microarrays (PBMs) for rapid, high-throughput characterization of the sequence specificities of DNA binding proteins. *J. Aoac Intl.*, **2006**, 79(4), 848-852.
- [84] Badis, G.; Berger, M.F.; Philippakis, A.A.; Talukder, S.; Gehrke, A.R.; Jaeger, S.A.; Chan, E.T.; Metzler, G.; Vedenko, A.; Chen, X.; Kuznetsov, H.; Wang, C.F.; Coburn, D.; Newburger, D.E.; Morris, Q.; Hughes, T.R.; Bulyk, M.L. Diversity and complexity in DNA recognition by transcription factors. *Science*, **2009**, 324(5935), 1720-1723.
- [85] The DREAM5 Project Available: <http://dreamchallenges.org/>, **2017**.
- [86] Maurer-Stroh, S.; Debulpae, M.; Kuemmerer, N.; Lopez de la, P. M.; Martins, I.C.; Reumers, J.; Morris, K.L.; Copland, A.; Serpell, L.; Serrano, L.; Schymkowitz, J.W.; Rousseau, F. Corrigendum: Exploring the sequence determinants of amyloid structure using position-specific scoring matrices. *Nat. Methods*, **2010**, 7(3), 237-242.
- [87] Tan, S.H.; Hugo, W.; Sung, W.K.; Ng, S.K. A correlated motif approach for finding short linear motifs from protein interaction networks. *BMC Bioinformatics*, **2006**, 7(1), 1-16.
- [88] Liu, X.S. An algorithm for finding protein-DNA binding sites with applications to chromatin-immunoprecipitation microarray experiments. *Nat. Biotechnol.*, **2002**, 20(8), 835-839.
- [89] Stoyan, G.; Boyle, A.P.; Karthik, J.; Ding, X.; Sayan, M.; Uwe, O. Evidence-ranked motif identification. *Genome Biol.*, **2010**, 11(2), 1-17.
- [90] Zhou, Q.; Liu, J.S. Modeling within-motif dependence for transcription factor binding site predictions. *Bioinformatics*, **2004**, 20(6), 909-916.
- [91] Hu, M.; Yu, J.; Taylor, J.M.G.; Chinnaiyan, A. M.; Qin, Z.S. On the detection and refinement of transcription factor binding sites using ChIP-Seq data. *Nucleic Acids Res.*, **2010**, 38(7), 2154-67.
- [92] Siddharthan, R. Dinucleotide Weight Matrices for Predicting Transcription Factor Binding Sites: Generalizing the Position Weight Matrix. *Plos One*, **2010**, 5(3), e9722.
- [93] Kelley, D. R.; Snoek, J.; Rinn, J. Basset: Learning the regulatory code of the accessible genome with deep convolutional neural networks. *Genome Res.*, **2016**.
- [94] Lanchantin, J.; Singh, R.; Lin, Z.; Qi, Y. Deep motif: Visualizing genomic sequence classifications. *ICLR workshops*, **2016**.
- [95] Quang, D.; Xie, X. DanQ: a hybrid convolutional and recurrent deep neural network for quantifying the function of DNA sequences. *Nucleic Acids Res.*, **2016**, 44(11), e107-e107.
- [96] Zeng, H.; Edwards, M.D.; Ge, L.; Gifford, D.K. Convolutional neural network architectures for predicting DNA-protein binding. *Bioinformatics*, **2016**, 32(12), i121-i127.
- [97] Engelhardt, B.E.; Brown, C.D. Diving deeper to predict noncoding sequence function. *Nat. Methods*, **2015**, 12(10), 925-926.
- [98] Zou, Q.; Li, J.; Hong, Q.; Lin, Z.; Shi, H.; Wu, Y.; Ju, Y. Prediction of microRNA-disease associations based on social network analysis methods. *BioMed Res. Intl.*, **2015**, 810514.
- [99] Li, P.; Guo, M.; Wang, C.; Liu, X.; Zou, Q. An overview of SNP interactions in genome-wide association studies. *Briefings in Funct. Genom.*, **2015**, 14(2), 143-155.
- [100] Zou, Q.; Li, J.; Wang, C.; Zeng, X. Approaches for recognition disease genes based on Network. *BioMed Res. Intl.*, **2014**, 416323.
- [101] Beale, G. The discovery of mustard gas mutagenesis by Auerbach and Robson in 1941. *Genetics*, **1993**, 134(2), 393-399.
- [102] Park, J. H.; Wacholder, S.; Gail, M. H.; Peters, U.; Jacobs, K.B.; Chanock, S.J.; Chatterjee, N. Estimation of effect size distribution from genome-wide association studies and implications for future discoveries. *Nat. Genet.*, **2010**, 42(7), 570-575.
- [103] Wells, R. D. Non-B DNA conformations, mutagenesis and disease. *Trends Biochem. Sci.*, **2007**, 32(6), 271-278.
- [104] Krawczak, M.; Cooper, D.N. Gene deletions causing human genetic disease: mechanisms of mutagenesis and the role of the local DNA sequence environment. *Hum. Genet.*, **1991**, 86(5), 425-441.
- [105] Cavallius, J.; Merrick, W. C. Site-directed mutagenesis of yeast eEF1A. Viable mutants with altered nucleotide specificity. *J. Biol. Chem.*, **1998**, 273(44), 28752-28758.
- [106] Ruff, A. J.; Kardashliev, T.; Dennig, A.; Schwaneberg, U. The Sequence Saturation Mutagenesis (SeSaM) Method. *Methods Mol.*

- Biol.*, **2014**, 1179, 45-68.
- [107] Mehta, G.; Jalan, R.; Mookerjee, R.P. Cracking the ENCODE: From transcription to therapeutics. *Hepatology*, **2013**, 57(6), 2532–2535.
- [108] Bonifer, C.; Cockerill, P.N. Transcriptional and epigenetic mechanisms regulating normal and aberrant blood cell development. *Bulletin of the School of Oriental & African Studies*, **2009**, 72(1), 191-192.
- [109] Weedon, M. N.; Cebola, I.; Patch, A. M.; Flanagan, S.E.; De Franco, E.; Caswell, R.; Rodríguez-Seguí, S.A.; Shaw-Smith, C.; Cho, C.H.; Lango, Allen H.; Houghton, J.A.; Roth, C.L.; Chen, R.; Hussain, K.; Marsh, P.; Vallier, L.; Murray, A.; International Pancreatic Agnesis Consortium.; Ellard, S.; Ferrer, J.; Hattersley, A.T. Recessive mutations in a distal PTF1A enhancer cause isolated pancreatic agenesis. *Nat. Genet.*, **2013**, 46(1), 61-4.
- [110] Stenson, P.D.; Mort, M.; Ball, E.V.; Shaw, K.; Phillips, A.; Cooper, D.N. The Human Gene Mutation Database: building a comprehensive mutation repository for clinical and molecular genetics, diagnostic testing and personalized genomic medicine. *Hum. Genet.*, **2014**, 133(1), 1-9.
- [111] Castro-Orós, I.D.; Pampín, S.; Bolado-Carrancio, A.; De Cubas, A.; Palacios, L.; Plana, N.; Puzo, J.; Martorell, E.; Stef, M.; Masana, L.; Civeira, F.; Rodríguez-Rey, J.C.; Pocovi, M. Functional analysis of LDLR promoter and 5' UTR mutations in subjects with clinical diagnosis of familial hypercholesterolemia. *Hum. Mutat.*, **2011**, 32(8), 868-72.
- [112] Pomerantz, M.M.; Ahmadiyeh, N.; Jia, L.; Herman, P.; Verzi, M.P.; Doddapaneni, H.; Beckwith, C.A.; Chan, J.A.; Hills, A.; Davis, M.; Yao, K.; Kehoe, S.M.; Lenz, H.J.; Haiman, C.A.; Yan, C.; Henderson, B.E.; Frenkel, B.; Barretina, J.; Bass, A.; Taberero, J.; Baselga, J.; Regan, M.M.; Manak, J.R.; Shivdasani, R.; Coetzee, G.A.; Freedman, M.L. The 8q24 cancer risk variant rs6983267 shows long-range interaction with MYC in colorectal cancer. *Nat. Genet.*, **2009**, 41(8), 882-884.
- [113] De Gobbi, M.; Viprasak, V.; Hughes, J.R.; Fisher, C.; Buckle, V.J.; Ayyub, H.; Gibbons, R.J.; Vernimmen, D.; Yoshinaga, Y.; de Jong, P.; Cheng, J.F.; Rubin, E.M.; Wood, W.G.; Bowden, D.; Higgs, D.R. A regulatory SNP causes a human genetic disease by creating a new transcriptional promoter. *Science*, **2006**, 312(5777), 1215-1217.
- [114] Kyrölahti, A.; Rämö, M.; Tamminen, M.; Unkila-Kallio, L.; Butzow, R.; Leminen, A.; Nemer, M.; Rahman, N.; Huhtaniemi, I.; Heikinheimo, M. GATA-4 regulates Bcl-2 expression in ovarian granulosa cell tumors. *Endocrinology*, **2008**, 149(11), 5635-42.
- [115] Forbes, S. A.; Bindal, N.; Bamford, S.; Cole, C.; Kok, C.Y.; Beare, D.; Jia, M.; Shepherd, R.; Leung, K.; Menzies, A.; Teague, J.W.; Campbell, P.J.; Stratton, M.R.; Futreal, P.A. COSMIC: mining complete cancer genomes in the catalogue of somatic mutations in cancer. *Nucleic Acids Res.*, **2011**, 39(2), D945–D950.
- [116] Bae, B.I.; Tietjen, I.; Atabay, K.D.; Evrony, G.D.; Johnson, M.B.; Asare, E.; Wang, P.P.; Murayama, A.Y.; Im, K.; Lisgo, S.N.; Overman, L.; Šestan, N.; Chang, B.S.; Barkovich, A.J.; Grant, P.E.; Topçu, M.; Politsky, J.; Okano, H.; Piao, X.; Walsh, C.A. Evolutionarily dynamic alternative splicing of GPR56 regulates regional cerebral cortical patterning. *Science*, **2011**, 343(6172), 764-768.
- [117] Bell, R.J.; Rube, H.T.; Kreig, A.; Mancini, A.; Fouse, S.D.; Nagarajan, R.P.; Choi, S.; Hong, C.; He, D.; Pekmezci, M.; Wiencke, J.K.; Wensch, M.R.; Chang, S.M.; Walsh, K.M.; Myong, S.; Song, J.S.; Costello, J.F. Cancer. The transcription factor GABP selectively binds and activates the mutant TERT promoter in cancer. *Science*, **2015**, 348(6238), 1036-1039.
- [118] Horn, S.; Figl, A.; Rachakonda, P. S.; Fischer, C.; Sucker, A.; Gast, A.; Kadel, S.; Moll, I.; Nagore, E.; Hemminki, K.; Schadendorf, D.; Kumar, R. TERT promoter mutations in familial and sporadic melanoma. *Science*, **2013**, 339(6122), 959-961.
- [119] Huang, F.W.; Hodis, E.; Xu, M.J.; Kryukov, G.V.; Chin, L.; Garraway, L.A. Highly recurrent TERT promoter mutations in human melanoma. *Science*, **2013**, 339(6122), 957-959.
- [120] Stenson, P.D.; Mort, M.; Ball, E.V.; Shaw, K.; Phillips, A.; Cooper, D.N. The human gene mutation database: building a comprehensive mutation repository for clinical and molecular genetics, diagnostic testing and personalized genomic medicine. *Hum. Genet.*, **2014**, 133, 1-9.
- [121] Chen, T.; Guestrin, C. XGBoost: A Scalable Tree Boosting System. In *22nd SIGKDD Conference on Knowledge Discovery and Data Mining*, **2016**.
- [122] Kircher, M.; Witten, D.M.; Jain, P.; O'Roak, B.J.; Cooper, G.M.; Shendure, J. A general framework for estimating the relative pathogenicity of human genetic variants. *Nat. Genet.*, **2014**, 46(3), 310-5.
- [123] Vincent, A.; Audo, I.; Tavares, E.; Maynes, J.T.; Tumber, A.; Wright, T.; Li, S.; Michiels, C.; GNB3 Consortium.; Condroyer, C.; MacDonald, H.; Verdet, R.; Sahel, J.A.; Hamel, C.P.; Zeitz, C.; Héon, E. Biallelic Mutations in GNB3, Cause a Unique Form of Autosomal-Recessive Congenital Stationary Night Blindness. *Am. J. Hum. Genet.*, **2016**, 98(5), 1011.
- [124] Ritchie, G. R.; Dunham, I.; Zeggini, E.; Flicek, P. Functional annotation of noncoding sequence variants. *Nat. Methods*, **2014**, 11(3), 294-296.
- [125] Song, J.; Liu, C.; Song, Y.; Qu, J.; Hura, G. S. Alignment of Multiple Proteins with an Ensemble of Hidden Markov Models. *ICMLA*, **2007**, 60–71.
- [126] Song, Y.; Wang, C.; Qu, J. A parameterized algorithm for predicting transcription factor binding sites. *Intelligent Comput. Bioinform.*, **2014**, 339-350.