GRAM: A generalized model to predict the cell-specific molecular effect of non-coding variants lie inside functional element

Abstract:

Identification and prioritization of function-associated variants become an increasing demand as next-generation sequencing data rapidly grows and accumulated. Current computational methods are developed to predict deleterious and disease-associated variants, not designed to predict cell-specific molecular phenotypes of these variants (i.e., their effects on gene expression regulation). In this paper, we proposed GRAM, a generalized model to predict molecular phenotype of non-coding variants in a cell-specific manner. We defined TF binding waiting-time features (TFT) to reflect the cell-type specific TF binding and dynamics. We first found that TF binding features are the most predictive features, while evolutionary conservation doesn't show indispensable contribution to molecular effect by employing comprehensive feature selection framework. Using in vitro SELEX TF binding features alone can achieve similar prediction power as using the TF binding features from ChIP-Seq. We then integrate with in-vitro TF binding features instead of those inferred from spotty covered ChIP-Seq data, and TFT features extracted from RNA-Seq to generalize our model to all other cell lines. In the multi-phase classification model, the AUROC reaches 0.728 and outperforms all the state-of-the-art tools. Finally, GRAM has been assessed in MCF7 and K562 cell lines, resulting in high predictive performance.

Introduction

Next-generation sequencing technologies enable high-throughput whole genome sequencing and exomes sequencing[1]. Many disease-associated mutations[2] and the vast majority of common single nucleotide variants have been identified in the human population [3, 4]. Genome-wide association studies(GWAS) have characterized many disease-associated variants. These variants mostly lie outside protein-coding regions, [5], emphasizing the importance of the function of regulatory elements in the human genome. This also drives an urgent need to develop high-throughput methods to sift through this deluge of sequence data to quickly determine the functional relevance of each noncoding variant[6].

It has been shown that only a fraction of noncoding variants are functional, and among the functional variants, the majority show only modest effects[7]. Therefore, highly quantitative assays are needed to study large number of variants. Luciferase assay is originally used to measure the regulatory effects of functional elements [8]. By comparing the difference of the assay output with and without the mutation, we can estimate the experimental molecular effect of non-coding variants lying in a functional element. By means of high throughput microarray and NGS methods, massively parallel reporter assay (MPRA) has extended the scales to the genome-wide level [9-14]. Recent, In Ke Tewhey et al.'s recent work, they have demonstrated the capability of MPRA to identify the causal variants that directly modulate gene expression. This study reports 842 variants (emVARs) showing significantly different expression modulation

Shaoke Lou 5/21/2018 9:46 AM
Deleted: High throughput reporter assays, like massively parallel reporter assay (MPRA) are successful in identifying functional elements in the whole genome. These MPRA datasets can be integrated with other next generation sequencing (NGS) data like ChIP-Seq to learn a knowledge model and predict molecular effect of variants. However, due to the heterogeneity of data sources and unbalanced data availability, most of TFs have ChIP-Seq experiments in only one or few cell lines, which make it difficult to build a model to estimate the molecular effect of variants within a functional element by considering these cell-line specific features. ... [1]

Shaoke Lou 5/21/2018 10:02 AM
Deleted: study the biological significance of

Shaoke Lou 5/21/2018 10:02 AM
Deleted: effect

Shaoke Lou 5/21/2018 10:04 AM
Deleted: , which can

Shaoke Lou 5/21/2018 10:04 AM
Deleted: not only

Shaoke Lou 5/21/2018 10:04 AM
Deleted: but loci specific information and is also easily calculated from RNA-Seq data

Shaoke Lou 5/24/2018 10:44 AM
Deleted: g

Shaoke Lou 5/24/2018 10:26 AM
Deleted: ,

TX Lee 5/21/2018 10:46 PM
Deleted: the understanding

TX Lee 5/21/2018 10:46 PM
Deleted: of

TX Lee 5/21/2018 11:54 PM
Deleted: is

TX Lee 5/22/2018 12:05 AM
Deleted: to be developed

TX Lee 5/22/2018 12:11 AM
Deleted: examine a

Shaoke Lou 5/24/2018 10:26 AM
Deleted: readout

TX Lee 5/22/2018 12:12 AM
Deleted: of the elements

TX Lee 5/22/2018 12:12 AM
Deleted: lie

Shaoke Lou 5/21/2018 10:08 AM
Deleted: Ryan's cell paper

effect and also provides a high-quality data source has been providing for computational modeling [15, 16].

There is also an increasing need for computational methods to effectively predict the molecular effect of variants and provide a better understanding of the underlying biology of these results. A host of approaches have been developed to address the problem of variant prioritization from different perspectives. According to the target of their predictions, there are mainly two categories: 1) disease-causing effect predictions: like Deepsea [17], GWAVA [18] and CADD[19] , try to prioritize causal disease variants and distinguish them from benign ones; 2) fitness consequence prioritization. fitCons and LINSIGHT \cite Huang, Gulko et al.) attempt to identify the variants on evolutionary fitness. Some tools, like Funseq2, may not belong to one particular category because of the integration of a comprehensive data context and unsupervised scoring system [6]. These computational methods are developed to predict and prioritize deleterious and disease-associated variants, but not designed to predict specific molecular phenotypes of these variants (i.e., their effects on activities of functional elements). Most importantly, none of the above tools takes into account cell specificity in their models. One reason may be because some cell-specific features derived from ChIP-Seq data are only available in a few cell lines, which is a major hindrance to the generalization of a model.

In this paper, we approach data mining methods from a new perspective to bridge the gap between the genotype and molecular phenotype. We developed GRAM: a generalized model predict the cell-specific molecular phenotype of non-coding variants. We define a TF waiting time (TFT) feature by considering the TF expression value and binding strength on a specific loci. We performed unsupervised and model-based feature selection and revealed TF binding score derived from in-vitro assay can achieve similar performance as the features from in-vivo ChIP-Seq experiments. We then built GRAM: a multi-stage classifier to account for various kinds of output from different experimental assay platforms. GRAM can achieve the highest performance compared with the state-of-the-art models according to the performance on the dataset from R. Tewhey et al.'s Cell paper. Finally, we assess our model using two sets of independent data in different cell lines: MPRA data in K562 and luciferase assay in MCF7, resulting in high predictive performance.

Results

Flowchart

In this study, as described in Fig1a, we firstly collected dataset from paper[15], which is the largest dataset so far for estimation of expression modulation differences between wild-type and mutants in GM12878 cell line. In his paper, he performed a large scale MPRA experiment, and provide a high-quality dataset contains 4xxx SNVs (3222 after filtered) with logSkew value, which measures the log fold change of the expression modulating differences between wild-type

Shaoke Lou 5/24/2018 10:24 AM
Deleted: F

TX Lee 5/22/2018 12:19 AM
Deleted: of important

TX Lee 5/22/2018 12:20 AM
Deleted: , are considering

Shaoke Lou 5/21/2018 10:09 AM
Deleted: tissue

Shaoke Lou 5/21/2018 10:10 AM
Deleted: Although

Shaoke Lou 5/21/2018 10:10 AM
Deleted: they may incorporate

Shaoke Lou 5/21/2018 10:11 AM
Deleted: line

Shaoke Lou 5/21/2018 10:10 AM
Deleted: or build a cell line specific model,

Shaoke Lou 5/21/2018 10:11 AM
Deleted: it is very limited because the feature it replys on is only

TX Lee 5/22/2018 12:20 AM
Deleted: will limite

Shaoke Lou 5/21/2018 10:12 AM
Deleted: effect

Shaoke Lou 5/21/2018 10:13 AM
Deleted: taking advantage of widely available RNA-Seq data

Shaoke Lou 5/21/2018 10:14 AM
Deleted: Besides the basic regression model, the core of GRAM is

Shaoke Lou 5/21/2018 10:16 AM
Deleted: , which includes a novel set of cell-specific effect TFT features and in vitro transcription factor binding features

Shaoke Lou 5/21/2018 10:17 AM
Deleted: Ryan's cell

and mutant alleles. Features are extracted according to R. Tewhey et al.'s Cell paper, such as cell-specific ChIP-Seq peaks and CAGE peaks, along with the knowledge from the other variants prioritization studies including evolutionary features and motif binding features. We performe a comprehensive feature selection framework, including unsupervised and model-based methods to identify the top impacting factors that affect the regulatory activity of element. Based on feature selection, we found the in vitro TF binding preferences that don't rely on cell line ChIP-Seq data is highly predictive in the model, which can repel the limitations from the spotty available ChIP-Seq data. We further define a TF binding waiting-time feature (TFT) using RNA-Seq data and combine with TF binding preference feature to build a multi-stage classification model. In the end, independent datasets from luciferase experiments and MPRA are then used to evaluate the model.

Exploration of conservation and transcription factor binding features

Evolutionary conservation is associated with deleterious fitness consequence and widely used in non-coding variant's prioritization algorithms, such as phyloP[20] and Phastcons [21] in LINSIGHT[22] and CADD, GERP in Funseq2. We performed comparative analyses for these three conservation features across different datasets. (Fig 1b), PhastCons and PhyloP pattern of emVar and non-emVar are less conserved than HGMD variants and similar to non-HGMD variants, which was thought to be a benign variant. GERP score show similar pattern but more centered in emVAR and non-EmVar compared to other datasets, with slightly larger values for emVAR. Since no different patterns were found between emVar and non-emVAR, we further discovered the correlation between logskew and conservation scores is low and the explained variance very close to 0 for all three features, which indicate these conservation scores standalone have no or minor contributions to molecular phenotype.

Transcription factor binding can link the molecular effect of noncoding variants to a cascade of regulatory network, which is thought to be an important contributing factor to the variants' regulatory effect (cadd, funseq, deepsea and deepbind). In R. Tewhey et al.'s Cell paper, they found the log skew positively associates with TF binding scores. To thoroughly look into the effect of TF binding, we tested all xxx TF motif break events and peaks overlapping with the SNVs in the dataset. Two sets of variants: emVAR and non-emVAR, were annotated and analyzed by Funseq2 [6]. The enrichment of transcription factor binding motifs in both sets, defined as ones with lowest p-values according to the hypergeometric distribution test, are shown in a bottom-up increasing order in Figures 1c, respectively. It was observed that emVAR set has more TF binding events compared with non-emVAR set. The top highly enriched TFs in emVAR are: xxxxx, . Besides the TF binding enrichment, we also further look at the motif break scores for these TFs, especially top enriched ones. The largest differential scores correspond to AP1 and EP300 motifs. In addition, for a smaller subset of motifs with lowest p-values, differences between the distribution of the binding alternative and reference genotypes in emVar is larger than that in the Non-emVar dataset for almost all motifs (Figure 1d), with the largest difference observed for AP1 and smallest for SMARC. According to the comparison, the emVAR

Shaoke Lou 5/21/2018 10:37 AM
Deleted: Since the logSkew value is continuous and supposed to contain more information than discretized classes, it is very straightforward to build a regression model. The model uses the logSkew value as target and

Shaoke Lou 5/21/2018 10:38 AM
Deleted: f

Shaoke Lou 5/21/2018 10:26 AM
Deleted:

Shaoke Lou 5/21/2018 10:26 AM
Deleted: Ryan's cell paper

Shaoke Lou 5/21/2018 11:01 AM
Formatted: Font color: Auto

Shaoke Lou 5/21/2018 11:01 AM
Formatted: Font color: Auto

Shaoke Lou 5/21/2018 10:38 AM
Deleted: , as predictors.

Shaoke Lou 5/21/2018 10:36 AM
Formatted: Font color: Orange

Shaoke Lou 5/21/2018 10:36 AM
Deleted: then employ

TX Lee 5/22/2018 12:25 AM
Comment [1]: May need a more "formal" expression?

TX Lee 5/22/2018 12:25 AM
Deleted: s

Shaoke Lou 5/21/2018 10:34 AM
Deleted: However, because the difference in the assay-based experiment is tha ... [2]

TX Lee 5/22/2018 12:28 AM
Deleted: found

Shaoke Lou 5/21/2018 10:40 AM
Deleted:

Shaoke Lou 5/21/2018 10:40 AM
Deleted: effect

TX Lee 5/22/2018 12:29 AM
Deleted: factor

TX Lee 5/22/2018 12:31 AM
Deleted: for

Shaoke Lou 5/21/2018 10:26 AM
Deleted: Ryan's paper

TX Lee 5/22/2018 12:32 AM
Deleted: s'

TX Lee 5/22/2018 12:37 AM
Deleted: with

TX Lee 5/22/2018 12:33 AM
Deleted: TFs

TX Lee 5/22/2018 12:35 AM
Deleted: differences between

set tends to have not only more TF binding events, but also larger binding alteration compared with non-emVAR set.

Model-based feature selection

In R. Tewhey et al.'s Cell paper, they found histone mark and CAGE highly enriched in emVAR regions, which indicates these features are potentially useful to predict the expression modulating effect. In addition, we also combine evolutionary feature and motif binding changes in our model-based feature selections. We collected 1678 training set from GM12878 cell line by removing variants that have no overlapping with any ChIP-seq peaks and incorporating features related to the CAGE, TFs, histone marks, and DNase I hypersensitivity sites. A comprehensive feature selection framework to select impactful features, is shown in Fig 2.

The most important features according to Lasso regression are TF binding features, and GERP scores just show very insignificant contributions (Fig2b). We then prioritized these features across models with different feature selection methods: Lasso, ridge, linear regression, stability selection [24](with five ƛstability values), random forest, mutual information, and Pearson correlation with the target variable. The 20 most important features (out of 515) w.r.t. mean importance across all methods is shown in decreasing order in Fig 2c. Expectedly, applying various methods on data with multiple dimensions leads to relatively varied results with regards to the importance of each feature across the method spectrum. Both ChIP-Seq and SELEX deepbind features show higher importance, with the top two being GM12878 ChIP-Seq features (SP1 And BCL3), which is cell line specific, then followed by some SELEX features starting with ETV1 and ETP63.

After considering feature importance values as per different criteria, we assess the performance difference between cell specific TF binding features (ChIP-Seq based) and non-specific ones (SELEX based) using SVR (support vector regressor), Lasso, and Random forest regression models. Interestingly, the incorporation of DeepBind ChIP-Seq derived features, which are cell-specific, does not boost the accuracy significantly for all three models. MSE values of both models, with and without DeepBind ChIP-Seq features, are shown in Fig 2d. Results suggest that we can reliably deploy the model trained on cell-line-independent Deepbind SELEX features (GRAM cell-line independent feature). Thus, we can rely on cell line independent features only to build a generalized model since not all the cell lines have available TF ChIP-Seq experiment for training of ChIP-Seq Deepbind binding model.

We compare the performances of models to predict molecular phenotype by using SVR and randomForest on different features sets, including GRAM all TF features, GRAM cell independent features, and the features from disease-association prediction tools: CADD, Funseq2, DeepSEA, GWAVA, LINSIGHT, Eigen decomposition, PCA, and Eigen.PC.phr. As shown in Fig 2e, the model with GRAM features lead to best models with the lowest mean squared error. As for other methods, results show that DeepSEA features result from the third best set of models (SVR and RF). It is indicates that the identification of disease assciation

TX Lee 5/22/2018 12:37 AM
Deleted: not only …ends to have not ... [3]

Shaoke Lou 5/21/2018 10:41 AM
Deleted: To learn the underlying patterns of variant modulated expression, we trained a host of machine learning models using a combination of epigenetic and evolutionary features. We firstly build a regression model to predict the log-skew difference in expression modulation fold change between wild-type and mutant alleles and then formulate a generalized model which classifies the variant effect as two experimental effect class: expression modulating (emVar, label 1) or non-modulating (nonEmVar, label 0).

Shaoke Lou 5/21/2018 10:24 AM
Deleted: Directly predict the expression modulating changes-logSkew

Shaoke Lou 5/21/2018 10:26 AM
Deleted: …R. Tewhey et al.'s Cell ... [4]

TX Lee 5/22/2018 12:38 AM
Deleted: without

Shaoke Lou 5/21/2018 10:42 AM
Deleted: A schematic representation of the regression task is shown …n Fig 2a ... [5]

Shaoke Lou 5/24/2018 9:53 AM
Deleted: We firstly learned a Lasso regression model with 10-fold cross-validation. The fine-tuning of ƛ, the penalty parameter in the cost function of this model, is determined according to the mean-squared error (MSE) values are shown in Fig 2b, with the best performance log(ƛ) ≅ -5. The R-square for prediction is 0.29 and 0.39 with TF binding features and with all features ... [6]

TX Lee 5/22/2018 12:43 AM
Comment [2]: Need clarification (cell-type binding specificity?)

TX Lee 5/22/2018 12:45 AM
Deleted: of

Shaoke Lou 5/21/2018 10:44 AM
Deleted: line …pecific TF binding fea ... [7]

TX Lee 5/22/2018 12:48 AM
Deleted: that can be used to infer

Shaoke Lou 5/21/2018 10:48 AM
Deleted: then …ompare the performa ... [8]

TX Lee 5/22/2018 12:50 AM
Deleted: output generated by

Shaoke Lou 5/21/2018 10:49 AM
Deleted: each of…CADD, Funseq2, ... [9]

TX Lee 5/22/2018 12:49 AM
Comment [3]: Figure 2e has not been presented so I don't quite understar ... [10]

variants is not equavalent to the prediction of the functional modulating variants. Hence, the tools that are built to predict the phenotyic consequences doesn't work very well in predicting molecular effects.

Build a generalized model by multi-phase learning

The mode-based feature selection illustrates the capability of prediction of the molecular effect of variants by the comprehensive integration of useful features. However, other than estimating the log skew value based on reads count as in MPRA, other different types of assay, such as Luciferase assay, GFP assay, and Lenti-virus based platforms, use fluorescence readouts instead, and apply different statistical methods or cutoff to determine the effects of the variants. Though different platforms may have consistent results [25, 26] , because of the varied raw measurements and analysis methods, interpretation of the outputs of these assays would be difficult. We need to define a unified prediction target that can be used for comparison cross different types of assays, with varied definition of molecular effect variants; In addition, we also constructed cell-specific features by extracting information from gene expression profile available in most of cell lines and tissues instead of ChIP-Seq assays whose availability is limited.

For phase one, we predict whether an element has regulatory activity. Using the Deepbind TF binding features as predictors, whether the element is functional element (emVAR) as the target, a randomForest classifier was trained to predict regulatory activity. The 10-fold cross validation demonstrate an exemplary performance with AUROC =0.938 and AUPRC = 0.924. The log odds based on the probabilities are highly correlated with actual logskew (with Pearson cor=0.5581, figure not shown).

In phase two, the cell-specific effect is considered by integrating TF gene expression profile. Log odds is calculated according to the the categorical table of the MPRA reads count for wild-type and mutant insertions and their backgrounds. The standard deviation of log odds (Vodds) represents the reliability of Chi-squared test. By comparing principal component loading of the Vodds from three cell lines: GM12878, GM19239, and HepG2, we found two GM cell lines are closer with each other than with HepG2 (fig 3d), which indicate the cell-type specificity of Vodds. Comparing emVAR with non-emVAR variants, the higher Vodds group tends to contain more non-emVar (Chi-square test p-value: 0.0002021), which indicates the emVAR class tends to have lower Vodds. From these results we could define the cell-specific classes (CS): high and low Vodds classes by top and bottom quartile value of Vodds, and then use TF binding score and TFT features to predict CS classes (fig3e). The 10-fold cross validation shows TF binding score can predict the CS target with AUC 0.80, and TFT features can achieve an AUC (0.65) which is also higher than a random effect (fig 3g-h).

The final phase is to predict whether the variants have significant expression modulating effect. The output from phase one and two are fed into a LASSO model, the emVar and non-emVar labels are used as the target. The AUROC of 10-fold cross-validation for the optimal model is 0.728 and AUPRC is 0.505, which is higher than the state-of-the-art for the study using the

Deleted: This comparison does not i ... [11]
Shaoke Lou 5/21/2018 10:52 AM
Deleted: regression …ode-based fe ... [12]
TX Lee 5/22/2018 12:52 AM
Deleted: instead of…ther than estim ... [13]
Shaoke Lou 5/21/2018 2:36 PM
Deleted: decide …he effects of the v ... [14]
TX Lee 5/22/2018 12:54 AM
Deleted: outcome as the
Shaoke Lou 5/21/2018 10:55 AM
Deleted: target …hat can be used fo ... [15]
TX Lee 5/22/2018 12:55 AM
Deleted: which and a classification n ... [16]
Shaoke Lou 5/21/2018 10:57 AM
Deleted: experimental …ffect variant ... [17]
TX Lee 5/22/2018 12:56 AM
Deleted: that
Shaoke Lou 5/21/2018 10:59 AM
Deleted: that are more easily obtain ... [18]
TX Lee 5/22/2018 12:56 AM
Deleted: which is only not available i ... [19]
Shaoke Lou 5/21/2018 11:00 AM
Deleted: and tissue… Gene expres ... [20]
TX Lee 5/22/2018 12:56 AM
Deleted: the P
Shaoke Lou 5/24/2018 10:07 AM
Deleted: will …redict whether an ele ... [21]
TX Lee 5/22/2018 12:58 AM
Deleted: a
Shaoke Lou 5/24/2018 10:05 AM
Deleted: enhancer-likeness
TX Lee 5/22/2018 12:59 AM
Deleted: .
TX Lee 5/22/2018 12:59 AM
Deleted: For
Shaoke Lou 5/21/2018 2:40 PM
Deleted: we want to consider
TX Lee 5/22/2018 12:59 AM
Deleted: in the study
Shaoke Lou 5/24/2018 10:09 AM
Deleted: The effect can reflect two ty ... [22]
TX Lee 5/22/2018 12:59 AM
Deleted: t
Shaoke Lou 5/21/2018 11:06 AM
Deleted: variance or …tandard devia ... [23]
TX Lee 5/22/2018 1:01 AM
Deleted: For Comparing the
Shaoke Lou 5/21/2018 11:11 AM
Deleted: variance
TX Lee 5/22/2018 1:01 AM
Deleted: have
Shaoke Lou 5/21/2018 2:42 PM
Deleted: .
TX Lee 5/22/2018 1:01 AM
Comment [4]: Redundant? (or move ... [24]
Shaoke Lou 5/21/2018 11:12 AM
Deleted: variances
TX Lee 5/22/2018 1:02 AM
Deleted: Then…rom these results w ... [25]
Shaoke Lou 5/21/2018 11:13 AM
... [26]
Shaoke Lou 5/21/2018 2:43 PM

same dataset (AUROC: 0.684, AUPRC: 0.478) [27]. For a generalized model, we redo phase one and two on the same dataset by excluding GRAM cell specific features that from ChIP-Seq model, which is not available for many other cell type or tissues, and keep all the other features as the optimal model, we get the model with AUROC = 0.674 and AUPRC = 0.452.

Model validation using cross-cellline and cross-assay datasets

The generalized model has been trained on Gm12878 MPRA dataset. To evaluate the performance of the model on the experimental results of other cell lines on different platforms. We collect nano luciferase assay data from MCF7 and MPRA assay data from K562[10]. 8 potential regulatory elements from MCF7 cell line have been experimentally tested, each one with a mutation as described in our study cite[ENCODEC]. We predict the regulatory activity for both wild-type and mutant alleles, and expression modulating differences between wide-type and mutant. For regulatory activity, the predicted probability to be an active regulator is positively correlated with luciferase assay fold change. The results are perfectly predicted (AUROC=1) for different luciferase fold change cutoffs from 1.2 – 2 that is used to define an active enhancer (fig5a). For the prediction of molecular effect, the significant differences between mutant and wild-type is defined by using absolute log2(fold change) cutoff. The predicted probability also showed a positive correlation with absolute log2 fold change. The AUROC value range from 0.7 to 0.9 given the absolute log2 cutoff from 0.5 to 1.5, which corresponding the fold change cut off from [1.414, 4] or [-4, -1.414]. For MPRA data in K562 cell line, we tested 2400 elements in 149bps with a variant centered in the inserted fragment. The AUC for regulatory activity is up to 0.68 as we decrease the cutoff of qvalue to 10^-9 and the molecular effect prediction also reach up to more than 0.8 if using a more stringent qvalue cutoff(10^-5). This indicates our model performs very well on the testing luciferase assay and MPRA dataset from a different cell lines even though they use different measurements.

Discussion

There is an increasing number of computational methods that can prioritize non-coding variants, as well as high-throughput whole-genome sequencing data that become the primary technique for identifying disease-associated variants. But it still lack a tool that can estimate the molecular effect of variant in a cell-specific manner. In this paper, we performed a thorough analysis of effect modeling on molecular effect of an SNV, trained both regression and classification models using MPRA data from Gm12878 cell lines. By taking advantage of the non-cell-specific SELEX TF binding feature, and easily obtained cell-specific TF expression data, we built a generalized model that can be potentially applied to any cell lines and tissues, and predict the significant expression modulation changes for different types of experiment assay. Experimental validation using luciferase assay on MCF7 cell lines, and MPRA assay on K562 to further verified the generality and robustness of the model.

In model-based featur selection, we tested features that may be associated with the molecular effect. In spite of the biological insight evolutionary features provide, Lasso regression indicates

Shaoke Lou 5/21/2018 2:43 PM
Deleted: Deepbind

Shaoke Lou 5/24/2018 9:52 AM
Formatted: Font:Bold

Shaoke Lou 5/24/2018 9:52 AM
Formatted: Font:Bold, Font color: Auto

Shaoke Lou 5/24/2018 9:52 AM
Formatted: Font:Bold

TX Lee 5/22/2018 1:03 AM
Deleted: at

Shaoke Lou 5/21/2018 2:45 PM
Deleted: We select

Shaoke Lou 5/21/2018 2:46 PM
Deleted: hich is

Shaoke Lou 5/21/2018 11:51 AM
Deleted: the

Shaoke Lou 5/21/2018 11:50 AM
Deleted: regression model

Shaoke Lou 5/21/2018 11:51 AM
Deleted: experimental

that they do not rank high in significance when predicting the molecular effect. The Histone Mark and CAGE features are chosen because of enrichment analysis between emVAR and non-emVAR, however, how these features work still unknown because no-chromatin context will be retained once the elements are inserted into a plasmid. The dataset of Histone Mark and CAGE is not always available for other cell lines, which will limit the application of the model. While the transcription factor binding is more biologically relevant, and the availability of in vitro SELEX model can help to expand the model to other cell type and tissues. Cell-specific ChIP-Seq-based TF binding features might help improve predictions but only to a limited extent, our models show that generalizability can be obtained using non-cell-specific SELEX TF binding features without a significant reduction in predictive performance.

In the cell-specific effect prediction, TF binding is still the most important factor, but TF waiting-time feature (TFT) also associate with cell-specific effect. The TFT feature is defined as a re-ordered TF expression matrix according to its binding strength(rank in its binding preference), which is inspired by the study of TF binding waiting time[28]. The waiting time of TF binding reflects the dynamics for TF bound and unbound to chromatin, and is thought to be related to TF binding free energy. The TF binding free energy can further be expressed as a function of the binding scores. Hence, TFT feature links the TF expression with binding waiting time and can unveal the dynamics TF cell specific effect. In our study, we simply use the quantile of binding preference in each TF's binding distribution to re-order the expression level and make the expression vector represent the binding order of TF and might affected by the noise of dataset and accuracy of TF binding preferences. However, our results indeed showed that the TFT feature has an association with the cell-specificity effect.

Though our model achieves so far the best performance, we recognize that dataset selection may introduce systematic bias because the SNVs we used in our model are only very small fraction of all non-coding variants but the regulatory effect of SNVs is very diverse, which will result in the overfitting of our model. However, our experimental validation has been performed on both small scale luciferase assay and high throughput MPRA data, our model shows high predictive performance in these blinded dataset. We will release our code publically, hope the community can help us improve and refine our model.

We aim to better understand the underlying patterns of variant modulation expression and considered cell specificity issues closely, having additional dataset generated from multiple cell line experiments would be quite helpful to derive more comprehensive conclusions. We will further expand this analysis contingent on the availability of data. In addition, continuous work on re-defining expression modulation remains an open question with large room for investigation

Methods

Dataset

Shaoke Lou 5/21/2018 11:54 AM
Deleted: however, features from a re-ordered TF expression matrix can also be problematic for some worse cases.

Shaoke Lou 5/21/2018 11:54 AM
Deleted: idea to

Shaoke Lou 5/21/2018 11:54 AM
Deleted: or

Shaoke Lou 5/21/2018 12:01 PM
Deleted:

Shaoke Lou 5/21/2018 11:53 AM
Deleted: ,

Shaoke Lou 5/21/2018 11:55 AM
Deleted: which

Shaoke Lou 5/21/2018 11:55 AM
Deleted: is further related to

Shaoke Lou 5/21/2018 12:02 PM
Deleted: just

Shaoke Lou 5/21/2018 12:00 PM
Deleted: re-ordered expression matrix

The data was downloaded from R. Tewhey et al.'s Cell paper [[remove, but keep only the citation]]. From about 79K tested elements, we only keep xxx variants that have at least either wild type or mutant elements show regulatory activity. We only keep the SNV with its logskew value and the logskew with maximum absolute value will be used if a SNV has been tested in two insertion directions in plasmid. Finally, we have 3222 SNVs tested in GM cell line in the our dataset. Each SNVs region is extended to both direction by 74bp, in total in 149bp. Another dataset from Ulirsch 2016 [10], there are 2756 variants tested in K562 cell line.

Shaoke Lou 5/21/2018 10:26 AM
Deleted:

Shaoke Lou 5/21/2018 10:26 AM
Deleted: Ryan cell paper

Feature extraction:

GERP feature was extracted using Funseq2 annotation pipeline, which search the region of element over the whole genome GERP score file and get average score.

The Histone modification, CAGE and ChIP-Seq peaks were overlapped to SNV element regions. It will be set as 1 if overlap with any peaks or set as 0. The motif break and motif gain score was calculated using Funseq2. We also calculated the motif score using Deepbind [29] with both the SELEX and ChIP-Seq motif model. The SELEX motif model are based on in vitro binding assay: systematic evolution of ligands by exponential enrichment, but ChIP-Seq models are inferred using sequence from the transcription factor binding site from different cell lines. There are total 515 motif models were calculated (table s1: tbls1.deepbind.list.txt) .

Model-based feature selection

Shaoke Lou 5/24/2018 9:56 AM
Deleted: *Regression*

the log skew of the SNV are used as target (y) and the GERP, histone modification ChIP-Seq feature group (11), transcription factor ChIP-seq feature group(16), CAGE feature group(5) and motif feature, a linear regression model was trained, the L1-norm was used as regularization term to avoid overfitting. The 10-fold cross-validation was used to select suitable scale factor (lambda) for L1-norm.

$$\min_{(\beta_0,\beta)\in R^{p+1}} \frac{1}{2N}\sum_{i=1}^{N}(y_i-\beta_0-x_i^T\beta)^2+\lambda||\beta||_1,$$

We firstly learned a LASSO regression model with 10-fold cross-validation. The fine-tuning of ƛ, the penalty parameter in the cost function of this model, is determined according to the mean-squared error (MSE) values with the best performance log(ƛ) ≅ -5. The R-square for prediction is 0.29 and 0.39 with TF binding features and with all features respectively

we also compare SVR and random Forest regressor on the same dataset.

To compare the importance of features, we compared different metrics, which including stability selection [24], LASSO 10-fold cross-validation, pearson correlation, linear regression, randomForest regression, feature elimination, Ridge, normalized mutual information. The features importance for each selection methods are scaled to [0, 1] and take the mean of all the selection methods to represent the overall ranking.

The logskew shows large kurtosis than expected normal distribution, the model was biased by the large amount centered data, the extreme logskew value will not be learned. we then applied adaboost with 10-fold cross-validation to enable the extreme-value sensitive classification. Meanwhile the adaboost model with in vitro motif (SELEX) feature and chip-seq motif binding feature are compared.

We compare our models' MSE with CADD, Eigen, LINSIGHT, Funseq2, GAWVA, DeepSea. The GM12878 specific model and generalized non-cell specific model was tested using both support vector regression and random forest regression, which consider all deepbind feature and SELEX-based features respectively. For the other variants prioritization tools, we take the output of these methods, and then use the same SVR and RandomForest to train and predict logskew value.

GRAM:multistage generalized model

Shaoke Lou 5/24/2018 9:56 AM
Deleted: *Classification:*

We first define the "emVar" as positive and "non-emVar" as negative classes following cell paper standard. There has 3222 data records, including xxx positive and xxx negative dataset.

We build a three phase model. Firstly, we will predict the element regulatory (enhancer) activity for wild type and mutant respectively and then predict cell specific effect model. The features include deepbind TF binding score from above and cell specific TF binding waiting time (TFT) feature.

An element inserted into plasmid with or without mutation is defined as a functional element if the fold change between the element with the control is larger than a statistically significant cutoff. For example, for MPRA study, the statistical test based on DESeq2 will indicate whether it is significantly changed; while for Luciferase assay, a testing element that has the fold change compared to control (eGFP) greater than 1.5 or 2 will be thought as regulatory element.

The regulatory activity class are defined based on the fold change of either wild-type or mutant readout compared with the control. The element with at least 2 fold changes will be defined as positive regulator, while the elements with at most xxx fold change is the negative set.

The effect can reflect two types of biological meaning: cell type specificity for the same loci between different cell lines and tissues, which can be naturally captured by gene expression profile; and loci specificity among different genomic positions in the same cell line or tissue, which is denoted by TF binding preference and TF's expression.

The cell specific effect model is approximated by the standard deviation of log(odds) given 2x2 categorical table (n1,n2,n3,n4 for the average reads count) for the association between the SNV type ("wild type", and "mutant") and assay type("experimental" and "control"). The standard

deviation of log(odds) is calculated by sqrt(1/n1 + 1/n2 +1/n3 + 1/n4). The Transcription factor binding and its expression level is biologically associated with the effect. We define the two classes using the top and bottom quartile standard deviation.

The quantile of distribution for each deepbind model was calculated based on the TF scores of 3222 SNVs. The order of TF expression is defined by the order of TF score's quantile in each model, then the expression rank matrix was generated by this new order.

Given 258 Deepbind SELEX model score S for 3222 SNV, Sm,n is the score for nth model of m-th SNV. Then we generate a ranking matrix R using column-based rank, R'm,n denote the rank for nth model of m-th SNV in the nth model score of all 3222 SNV, For TF with multiple binding models, we take top-rank for each TF to generate a TF-based mxn' R' matrix, where n' is the number of unique TF in SELEX model.

For each SNV, the R'm: {1,.., n'} (n' is the number of unique TFs) is then used to generate a new ranked TF vector TR{1_r,…, n'_r} , which is ordered by the R'm. TFexpression value E {1,…,n'} is re-ordered according to new TF E'm{1,.., n'}. This E' vector indicate the relationship between expression level and binding preference on each SNV.

The predict probability to be active element from the first step is then used to calculate: log2(P_mut/(1-P_mut) /(P_ref/(1-P_ref))).

The last step is to predict whether there is significant change of regulatory activity between wild-type and mutant element using predicted prob odds and cell-specific effect by.

Experiment validation on MCF7 cell line

We introduced mutations into cloned non-coding elements by site-directed mutagenesis, following published procedures (Wei et al., 2014) in general. Briefly, a pair of mutagenesis primers was designed for each mutation with a webtool, PrimerDIY (primer.yulab.org). We set up mutagenesis PCR reactions with the entry clone plasmids carrying wild-type non-coding elements and their corresponding mutagenesis primer pairs. The PCR products were then digested with DpnI (New England BioLabs) and transformed into TOP10 chemically competent E. coli (Invitrogen) by heatshock. The transformed bacteria were recovered in SOC medium for 1h at 37°C, spread on LB agar plates supplemented with spectinomycin, and incubated at 37°C overnight. We randomly picked colonies yielded from the transformation and confirmed the success of mutagenesis by Sanger sequencing.

Model validation using MPRA data from K562 cell line

Shaoke Lou 5/24/2018 9:57 AM
Formatted: Font:Italic

References:

1. Snyder M, Du J, Gerstein M. Personal genome sequencing: current approaches and challenges. Genes Dev. 2010;24(5):423-31. Epub 2010/03/03. doi: 10.1101/gad.1864110. PubMed PMID: 20194435; PubMed Central PMCID: PMCPMC2827837.
2. Stenson PD, Mort M, Ball EV, Howells K, Phillips AD, Thomas NS, et al. The Human Gene Mutation Database: 2008 update. Genome Med. 2009;1(1):13. Epub 2009/04/08. doi: gm13 [pii]
10.1186/gm13. PubMed PMID: 19348700; PubMed Central PMCID: PMC2651586.
3. The 1000 Genomes Project Consortium. An integrated map of genetic variation from 1,092 human genomes. Nature. 2012;491(7422):56-65. Epub 2012/11/07. doi: nature11632 [pii] 10.1038/nature11632. PubMed PMID: 23128226; PubMed Central PMCID: PMC3498066.
4. Fu W, O'Connor TD, Jun G, Kang HM, Abecasis G, Leal SM, et al. Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. Nature. 2013;493(7431):216-20. Epub 2012/12/04. doi: nature11690 [pii]
10.1038/nature11690. PubMed PMID: 23201682; PubMed Central PMCID: PMC3676746.
5. Maurano MT, Humbert R, Rynes E, Thurman RE, Haugen E, Wang H, et al. Systematic localization of common disease-associated variation in regulatory DNA. Science. 2012;337(6099):1190-5. doi: 10.1126/science.1222794. PubMed PMID: 22955828; PubMed Central PMCID: PMCPMC3771521.
6. Fu Y, Liu Z, Lou S, Bedford J, Mu X, Yip KY, et al. FunSeq2: A framework for prioritizing noncoding regulatory variants in cancer. Genome Biol. 2014;15:480. doi: 10.1186/s13059-014-0480-5. PubMed PMID: 25273974.
7. Patwardhan RP, Hiatt JB, Witten DM, Kim MJ, Smith RP, May D, et al. Massively parallel functional dissection of mammalian enhancers in vivo. Nat Biotechnol. 2012;30(3):265-70. doi: 10.1038/nbt.2136. PubMed PMID: 22371081; PubMed Central PMCID: PMCPMC3402344.
8. Smale ST. Luciferase assay. Cold Spring Harb Protoc. 2010;2010(5):pdb prot5421. Epub 2010/05/05. doi: 10.1101/pdb.prot5421. PubMed PMID: 20439408.
9. Grossman SR, Zhang X, Wang L, Engreitz J, Melnikov A, Rogov P, et al. Systematic dissection of genomic features determining transcription factor binding and enhancer function. doi: 10.1073/pnas.1621150114.
10. Ulirsch JC, Nandakumar SK, Wang L, Giani FC, Zhang X, Rogov P, et al. Systematic functional dissection of common genetic variation affecting red blood cell traits. Cell. 2016;165:1530-45. doi: 10.1016/j.cell.2016.04.048. PubMed PMID: 27259154.
11. Kheradpour P, Ernst J, Melnikov A, Rogov P, Wang L, Zhang X, et al. Systematic dissection of regulatory motifs in 2000 predicted human enhancers using a massively parallel reporter assay. Genome research. 2013;23:800-11. doi: 10.1101/gr.144899.112. PubMed PMID: 23512712.
12. Inoue F, Ahituv N. Decoding enhancers using massively parallel reporter assays. Genomics. 2015;106:159-64. doi: 10.1016/j.ygeno.2015.06.005.
13. Melnikov A, Murugan A, Zhang X, Tesileanu T, Wang L, Rogov P, et al. Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay. Nature Biotechnology. 2012;30:271-7. doi: 10.1038/nbt.2137.
14. Ernst J, Melnikov A, Zhang X, Wang L, Rogov P, Mikkelsen TS, et al. Genome-scale high-resolution mapping of activating and repressive nucleotides in regulatory regions. Nature Biotechnology. 2016;34:1180-90. doi: 10.1038/nbt.3678.
15. Tewhey R, Kotliar D, Park DS, Lander ES, Schaffner SF, Sabeti PC. Direct Identification of Hundreds of Expression- Modulating Variants using a Multiplexed Reporter Assay. Cell. 2016;165:1519-29. doi: 10.1016/j.cell.2016.04.027.

16. Zeng H, Edwards MD, Guo Y, Gifford DK. Accurate eQTL prioritization with an ensemble-based framework. doi: 10.1101/069757.
17. Zhou J, Troyanskaya OG. Predicting effects of noncoding variants with deep learning-based sequence model. Nature methods. 2015;12:931-4. doi: 10.1038/nmeth.3547. PubMed PMID: 26301843.
18. Ritchie GRS, Dunham I, Zeggini E, Flicek P. functional annotation of noncoding sequence variants. Nature methods. 2014;11:294-6. doi: 10.1038/nmeth.2832.
19. Kircher M, Witten DM, Jain P, O'roak BJ, Cooper GM, Shendure J. A general framework for estimating the relative pathogenicity of human genetic variants. Nature Genetics. 2014;46:310-5. doi: 10.1038/ng.2892. PubMed PMID: 24487276.
20. Cooper GM, Stone EA, Asimenos G, Program NCS, Green ED, Batzoglou S, et al. Distribution and intensity of constraint in mammalian genomic sequence. Genome Res. 2005;15(7):901-13. Epub 2005/06/21. doi: 10.1101/gr.3577405. PubMed PMID: 15965027; PubMed Central PMCID: PMCPMC1172034.
21. Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, et al. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res. 2005;15(8):1034-50. Epub 2005/07/19. doi: 10.1101/gr.3715005. PubMed PMID: 16024819; PubMed Central PMCID: PMCPMC1182216.
22. Huang Y-F, Gulko B, Siepel A. Fast, scalable prediction of deleterious noncoding variants from functional and population genomic data. Nature Genetics. 2017. doi: 10.1038/ng.3810.
23. Bannister AJ, Kouzarides T. Regulation of chromatin by histone modifications. Cell research. 2011;21:381-95. doi: 10.1038/cr.2011.22. PubMed PMID: 21321607.
24. Meinshausen N, Bühlmann P. Stability selection. Journal of the Royal Statistical Society: Series B (Statistical Methodology). 2010;72(4):417-73. doi: doi:10.1111/j.1467-9868.2010.00740.x.
25. Vesuna F, Winnard P, Raman V, Raman V. Enhanced green fluorescent protein as an alternative control reporter to Renilla luciferase. Analytical biochemistry. 2005;342:345-7. doi: 10.1016/j.ab.2005.04.047. PubMed PMID: 15950916.
26. Hall MP, Unch J, Binkowski BF, Valley MP, Butler BL, Wood MG, et al. Engineered Luciferase Reporter from a Deep Sea Shrimp Utilizing a Novel Imidazopyrazinone Substrate. ACS Chemical Biology. 2012;7:1848-57. doi: 10.1021/cb3002478. PubMed PMID: 22894855.
27. Guo Y, Tian K, Zeng H, Guo X, Gifford DK. A novel k-mer set memory (KSM) motif representation improves regulatory variant prediction. 2017. doi: 10.1101/130815.
28. Zabet NR, Adryan B. A comprehensive computational model of facilitated diffusion in prokaryotes. Bioinformatics. 2012;28(11):1517-24. Epub 2012/04/12. doi: 10.1093/bioinformatics/bts178. PubMed PMID: 22492644; PubMed Central PMCID: PMCPMC3356843.
29. Alipanahi B, Delong A, Weirauch MT, Frey BJ. Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. Nat Biotechnol. 2015;33(8):831-8. Epub 2015/07/28. doi: 10.1038/nbt.3300. PubMed PMID: 26213851.

Figures:

Figure 1 (a) flowchart of our study. (b) Conservation scores (c) MOTIFBR - motif-based - P-value (bottom- sorted up increasing order) (d) motif score changes between wild-type and mutant allele.

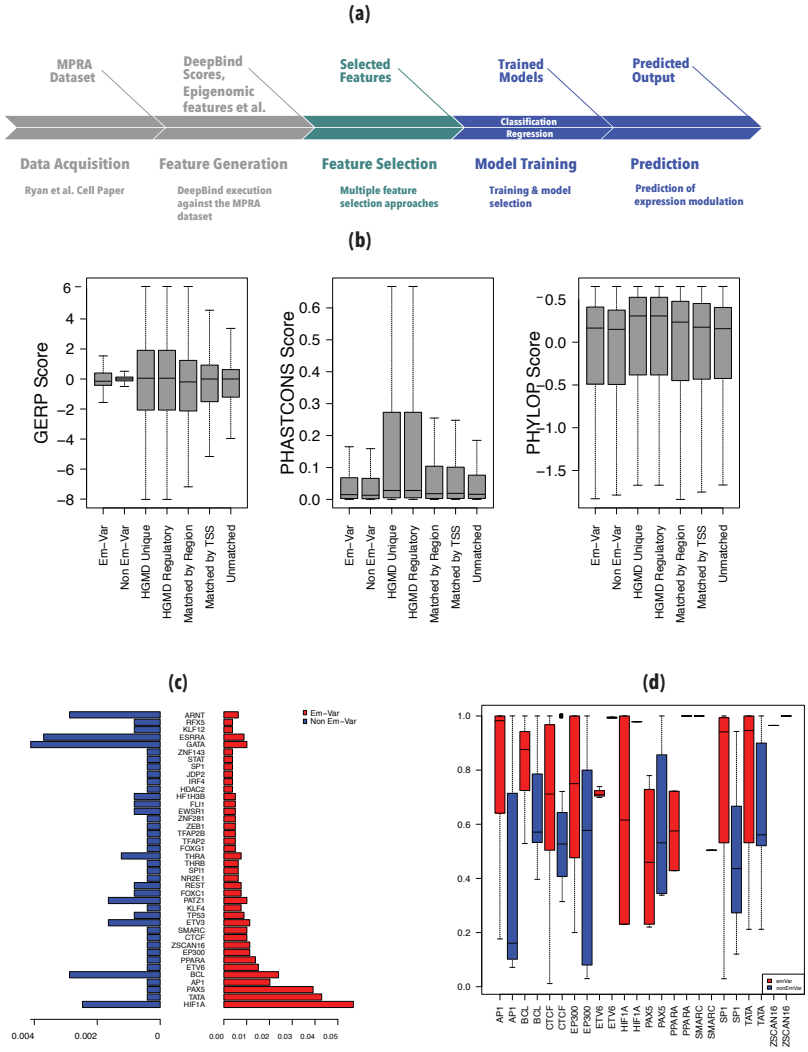
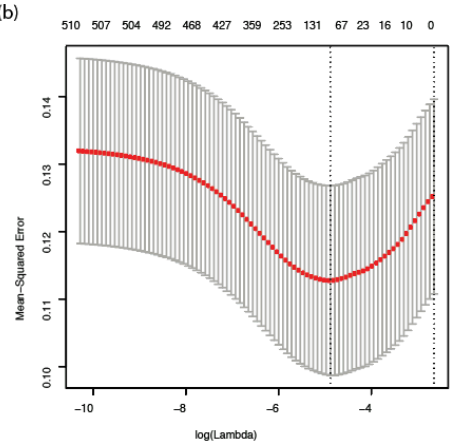

(a)
MPRA Dataset
DeepBind Scores, Epigenomic features et al.
Selected Features
Trained Models
Predicted Output
Classification
Regression
Data Acquisition
Feature Generation
Feature Selection
Model Training
Prediction
Ryan et al. Cell Paper
DeepBind execution against the MPRA dataset
Multiple feature selection approaches
Training & model selection
Prediction of expression modulation
(b)
GERP Score
PHASTCONS Score
PHYLOP Score
Em−Var
Non Em−Var
HGMD Unique
HGMD Regulatory
Matched by Region
Matched by TSS
Unmatched
(c)
Em−Var
Non Em−Var
(d)

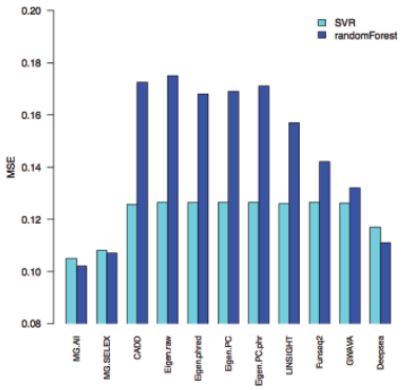
Fig2. Regression model to predict logSkew. (a): diagram of features in regression model (b) Lasso regression with 10-fold cross-validation (c) feature selection for Deepbind motif scores, identify cell-line specific feature from top ranking list. (d) comparison the performance of cell-line specific ChIP-Seq TF binding scores with SELEX TF binding scores. (e):Compare with the the-state-of-the art, we use their direct output as features, then train 10-fold cross-validation model using svr and random forest to compare with our model.

(a)

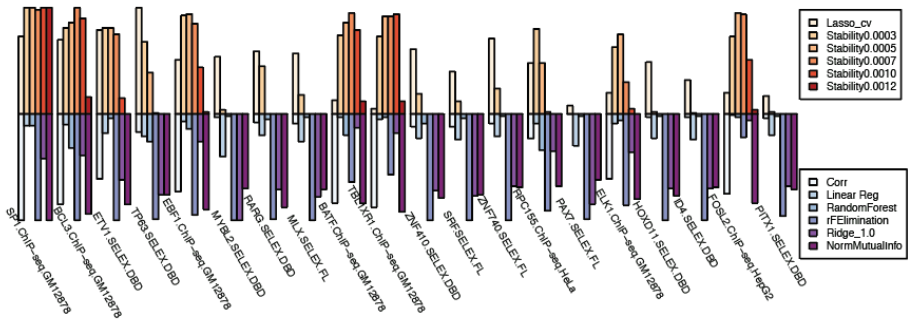
(b)



(e)



(e)



(d)

MSE	Lasso 1se	SVM	RandomForest
ChIPseq + SELEX	0.106	0.105	0.102
SELEX	0.111	0.108	0.107

Fig3 multi-stage classification model. (a) the diagram of multi-phase model: before predict the molecular effect, the regulatory activity and cell specificity is predicted. (b) ROC curve for regulatory activity prediction. (c) PRC curve for regulatory activity prediction, (d) the principal component analysis for Vodds, the loadings for PC1 and PC2 are shown. (e) The high and low variable cell specificity class are defined by the top and bottom quantile. (f) The prediction of cell specificity variation prediction using TF binding feature and TFT features.

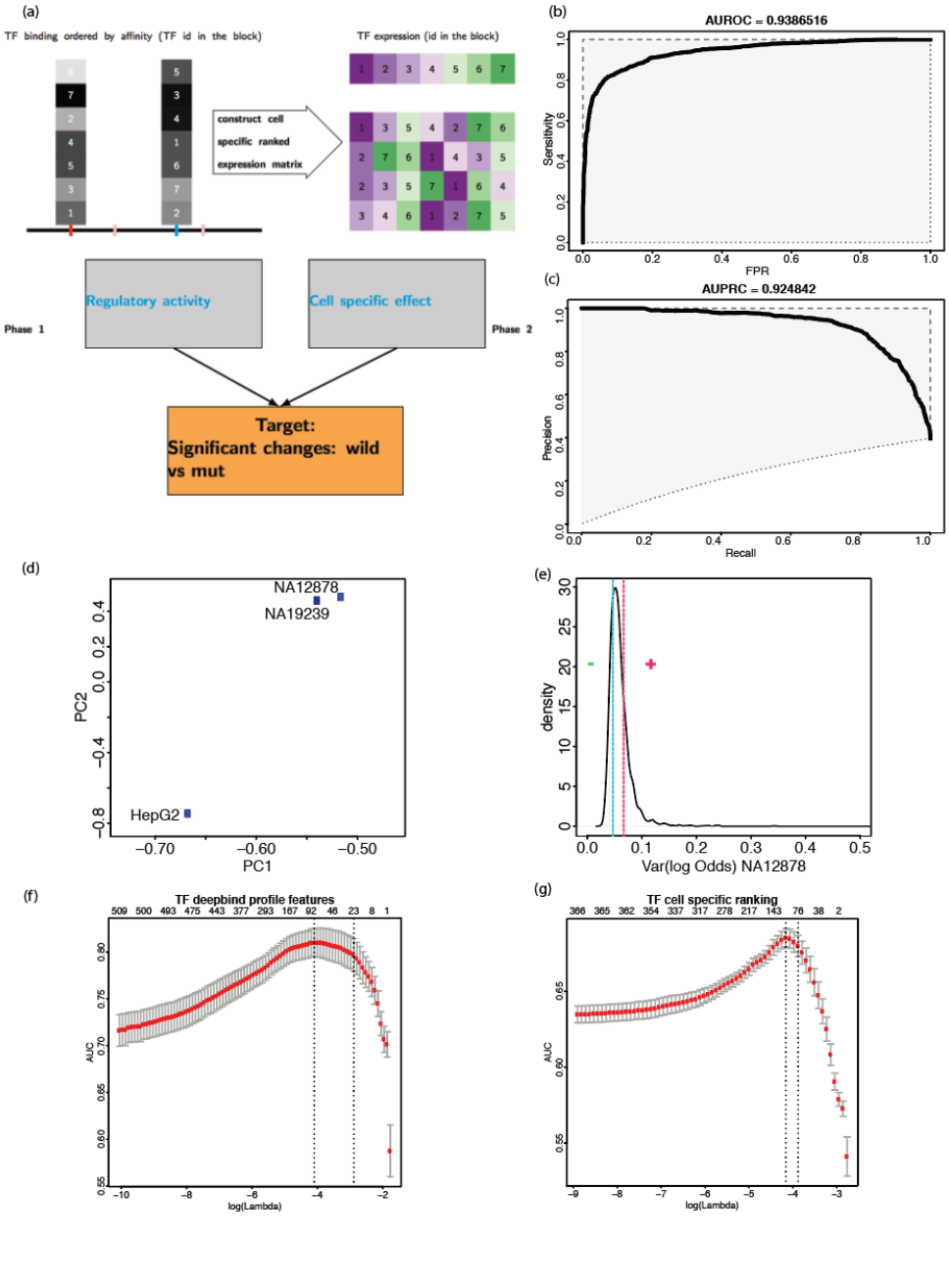

(a)
TF binding ordered by affinity (TF id in the block)
TF expression (id in the block)
construct cell specific ranked expression matrix
Regulatory activity
Cell specific effect
Phase 1
Phase 2
Target: Significant changes: wild vs mut
(b)
AUROC = 0.9386516
Sensitivity
FPR
(c)
AUPRC = 0.924842
Precision
Recall
(d)
NA12878
NA19239
HepG2
PC2
PC1
(e)
density
Var(log Odds) NA12878
(f)
TF deepbind profile features
AUC
log(Lambda)
(g)
TF cell specific ranking
AUC
log(Lambda)

Fig4. Performance of classification model. A,B ROC and PRC for model including tissue-specific ChIP-Seq Deepbind scores, C, D ROC and PRC for generalized model

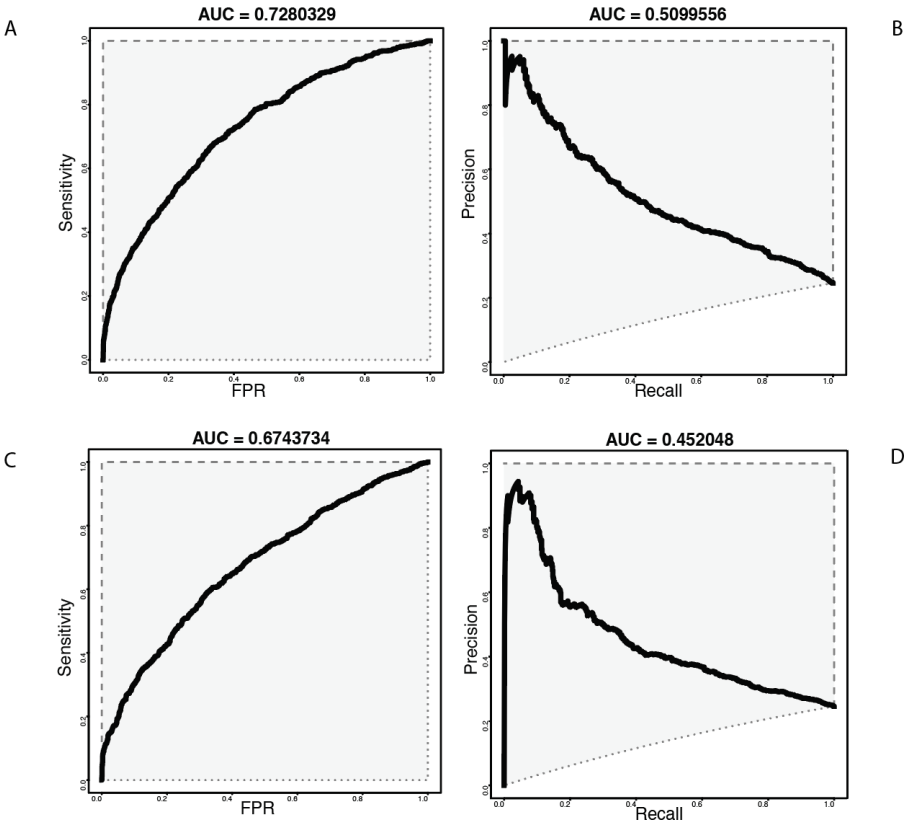


Fig5 (a) enhancer-likeness prediction. x-axis: fold change from experiment, the vertical dot lines represent the cut off (1.5, or 2) to determine positive (enhancer) and negative, the horizontal dot line is predicted probability cutoff (0.5). (b): predicted probability for emVar and non-emVAR versus absolute log2 odds from luciferase assay. (c): the AUROC value versus the different absolute log2 odds cutoff [0.5, 2.0]; (d) testing on a independent dataset K562 MPRA dataset